BIOINFORMATICS OF GENOME REGULATION AND SYSTEMS BIOLOGY

EDITED BY: Yuriy L. Orlov and Ancha Baranova

PUBLISHED IN: Frontiers in Genetics and Frontiers in Plant Science

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88966-014-8
DOI 10.3389/978-2-88966-014-8

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

BIOINFORMATICS OF GENOME REGULATION AND SYSTEMS BIOLOGY

Topic Editors:
Yuriy L. Orlov, First Moscow State Medical University, Russia
Ancha Baranova, George Mason University, United States

Citation: Orlov, Y. L., Baranova, A., eds. (2020). Bioinformatics of Genome Regulation and Systems Biology. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88966-014-8

Table of Contents

frontiers in Genetics

EDITORIAL
published: 28 July 2020
doi: 10.3389/fgene.2020.00625

Editorial: Bioinformatics of Genome Regulation and Systems Biology

Yuriy L. Orlov[1,2,3]* *and Ancha V. Baranova*[4]*

[1] Institute of Digital Medicine, I.M.Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russia, [2] Life Sciences Department, Novosibirsk State University, Novosibirsk, Russia, [3] Agrobiotechnology Department, Agrarian and Technological Institute, Peoples' Friendship University of Russia, Moscow, Russia, [4] School of Systems Biology, George Mason University, Fairfax, VA, United States

Keywords: genomics, bioinformactics, systems biology, plant science, gene expression, special issue

Editorial on the Research Topic

Bioinformatics of Genome Regulation and Systems Biology

This Research Topic presents the studies in the field of computational genomics. These papers were discussed at BGRS\SB-2018 (Bioinformatics of Genome Regulation and Structure Systems Biology) multi-conference, along with the hybrid wet-lab/computational genetics studies focused on genome-wide gene expression regulation. The BGRS is the major event in the computational biology field, which has been held in Novosibirsk, Russia biannually since 1998. The main conference is typically followed by a series of special post-conference journal issues covering contemporary computational genetics and genomics applications (Orlov et al., 2016, 2019a; Tatarinova et al., 2019). First Special Issues covering BGRS\SB conference were presented in the *Journal of Bioinformatics and Computational Biology* in 2012 (Kolchanov and Orlov, 2013; Orlov et al., 2015, 2019b) and other platforms (Chen et al., 2017; Baranova et al., 2019; Orlov, 2019; Medical Genetics and Bioinformatics special issue). Starting in 2018, extended discussion of the conference materials in genetics and genomics is being presented in *Frontiers in Genetics*.

In this Research Topic, we arranged the papers by areas of applications—clinical bioinformatics and human genome studies are followed by the plant genetics and then by systems biology applications.

Bah et al. comprehensively reviewed genomics tools and databases allowing us to dissect the pathophysiology of bacterial and parasitic infection, spanning the species from *Mycobacterium tuberculosis* to *Plasmodium falciparum*. These databases provide the data and tools for in-depth investigations of disease outbreaks and pathophysiological mechanisms, genomic variation and co-evolution of hosts and pathogens, diagnostic markers and vaccine targets, with special attention to the contributions of genomics and bioinformatics to the management of both common and neglected tropical diseases, including tuberculosis, dengue fever, malaria, and filariasis.

The TCGA (The Cancer Genome Atlas) database was mined from an entirely new technical viewpoint of developing reference genes with stable mRNA levels for quantitative PCR studies of cancer cells (Krasnov et al.). A scoring system for the assessment of gene expression stability allowed authors to highlight previously untried reference gene candidates, specific to each cancer type, along with several more "universal," pan-cancer reference gene candidates, namely *SF3A1*, *CIAO1*, and *SFRS4*. The application on colon adenocarcinoma was presented in Fedorova et al. (2019), another work in the frames of BGRS SB conference series.

The study by Ivanov et al. highlighted methodological problems for an up-and-coming biomarker mining technique, a sequencing of cell-free DNA (cfDNA) in human plasma. As fragmentation patterns of cfDNA are far from being random due to nucleosome patterns reflecting tissue-specific epigenetic signatures, these patterns may be used for guiding the design of amplicon-based NGS panels. Therefore, the sensitivity of mutation detection in liquid biopsy samples may be much improved, allowing for a lessening of the amount of body fluids collected from patients.

OPEN ACCESS

Edited and reviewed by:
Richard D. Emes,
University of Nottingham,
United Kingdom

***Correspondence:**
Yuriy L. Orlov
orlov@d-health.institute
Ancha V. Baranova
abaranov@gmu.edu

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *17 May 2020*
Accepted: *26 May 2020*
Published: *28 July 2020*

Citation:
Orlov YL and Baranova AV (2020) Editorial: Bioinformatics of Genome Regulation and Systems Biology. Front. Genet. 11:625. doi: 10.3389/fgene.2020.00625

Khatun et al. work in the medical bioinformatics field; they have developed a computational tool PreAIP (Predictor of Anti-Inflammatory Peptides), aimed at augmenting the search for novel biologics. Integrative analysis of stomach carcinoma samples by pairing DNA methylation patterns with gene regulatory network topology was presented in Wu et al.. The authors showed conservation of epigenetic patterns across various stages of this important type of human malignancies.

Gene expression regulation at genome level is important in evolution and adaptation studies (Ponomarenko et al., 2017; Igoshin et al.). Igoshin et al. looked into the adaptation of humans to cold climate. They have concentrated on the *TRPM8* gene, which encodes for a cold-sensing ion channel. In a population data set, they found a very promising single nucleotide polymorphism rs7577262 with a signature of selective sweep. Chadaeva et al. employed bioinformatics to discern behavioral pattern in mice and identify variants contributing to the dominance and the subordination traits continuing bioinformatics behavior studies in laboratory animals (Bragin et al., 2017). Using the prediction on-line tool SNP_TATA_Comparator (Ponomarenko et al., 2017) a set of candidate SNP markers contributing to the dominance and the subordination were uncovered. The studies using same SNP analysis tool were continued in Oshchepkov et al. (2019) and Ponomarenko et al. (2020).

Zverkov et al. considered a problem of genome reduction in primitive parasites. Among the two groups of microscopic parasitic invertebrates, the Dicyemida, and Orthonectida, overall morphological organization is much simplified, with tissues and organs almost absent. In these species, homeodomain transcription factors, G-protein-coupled receptors, and many other protein families have undergone a massive reduction. Interestingly, it seems that the dramatic simplification of body plans in dicyemids and orthonectids has evolved independently.

Das et al. discuss the application of ancestry informative markers (AIMs), previously developed for the inference of genomic ancestry in humans (Das and Upadhyai, 2018), for the delineation of gorilla lineages. Three of the four AIMs-determining approaches were successful for gorilla species (Das et al.).

The next group of papers in the Research Topic highlight the findings in genome regulation related to plants genetics. Kovalev et al. developed a computer pipeline and a machine learning classifier of deleterious coding mutations in agricultural plants, with the performance exceeding that of the popular PolyPhen-2 tool. The novel tool will improve the annotation of genes located in QTL and GWAS hit regions. This work was initially discussed at BGRS\SB-2018 plant biology session as well (Orlov et al., 2019c).

Zhang et al. studied abiotic stress in a model of *Populus euphratica* and its sister species *P. pruinosa*, differing by their adaptability to the content of salt in the soil. The authors performed transcriptome analyses of three seed germination phases from both of the species of desert poplar, and presented their findings in a form of a database suitable for use by poplar breeders. Wang et al. also studied *Populus euphratica*, in this case to infer genetics mechanisms of crossover Interference. Four-point linkage analysis allowed them to show the distribution of the crossover interference through the entire genome of this tree, uniquely suited for survival in saline deserts.

The following work by Khassanova et al. continues the line of studies of salinity resistance by exploring expression profiles in the chickpea (*Cicer arietinum* L.). They have tested six accessions of Chickpea ecotypes, all selected from field trials, for tolerance to abiotic stresses, found the involvement of *CaRabC* gene and developed markers for genotyping chickpea germplasm. Gene expression patterns in bread wheat exposed to drought were studied in Zotova et al.. The authors' team had identified general transcription repressor *TaDr1*, a part of *TaDr1*, *TaDr1A* and *TaDr1B* gene set, with drought-dependent variable expression. It seems that the general transcription repressor *TaDr1* controls expression of *TaVrn1* and *TaFT1* and, consequently, flowering time. These finding have direct implications for plant productivity in the dry environment.

Flowering time in plants is important agricultural feature determined by genetics and environment. Gursky et al. dissected the core genetic regulatory network canalizing the flowering signals to the decision to flower. While discovered and extensively studied in the model plant *Arabidopsis thaliana,* the flowering model may hold in other species (Kozlov et al., 2019). When the authors built a model gene network in chickpea (*Cicer arietinum*), activation from the *FLOWERING LOCUS T* gene or its homologs to the flowering decision led to a high expression of the meristem identity genes, including *AP1*. Different levels of activation from *AP1* may explain the differences observed in the expression of the two homologs of the repressor gene *TFL1* in species compared. Zhao et al. worked on tea plant (*Camellia sinensis*). In this plant, the development of new sprouts directly affects the yield and quality of the tea leaves, by affecting the content of catechins, theanine, and caffeine. Using High-Performance Liquid Chromatography-Mass Spectrometry, authors showed that conserved miRNA are playing a role in primary metabolism of a tea plant during sprouting. Li et al. presented their study of the chloroplast genomes of *Vicia sepium*, an important wild resource plant suitable for cultivation in extreme cold and dry conditions. The authors have compared a new complete chloroplast genome of *V. sepium* with the chloroplast genomes from related genera belonging to tribe Fabeae, then reconstructed the evolutionary history of the chloroplast genomes in these species.

Orlov M. et al. have studied promoters of *Mycoplasma gallisepticum,* an intracellular parasite affecting the respiratory tract of poultry, and found that the vlhA promoters differ by carrying a variable GAA repeats region upstream of transcription start site. These data have implications for the studies of the phase variation in *M. gallisepticum*. The computer technique of such promoter studies were continued in Orlov and Sorokin (2020).

Liu et al. presented their study of gender differences in solitary parasitoid species *Brachymeria lasus*, which has been evaluated as a potential candidate for release to control the

gypsy moth, *Lymantria dispar*, a pest of worldwide importance. Work by Qin et al. considers the polyploidy problem in vertebrates. They have analyzed genome organization in the autotetraploid of the red crucian carp (*Carassius auratus* red var.). The loss of chromosomal loci, base variations in non-transcribed spacer, and array recombination of repeat units have been detected.

Overall, we are proud of the Research Topic at Frontiers in Genetics we collated. We hope that you will find this paper collection a stimulating reading, and will consider coming to the next BGRS\SB conferences in Novosibirsk, Russia as well as read next "Bioinformatics of Genome Regulation" Research Topic in *Frontiers* (https://www.frontiersin.org/research-topics/14266/bioinformatics-of-genome-regulation).

AUTHOR CONTRIBUTIONS

YO and AB organized the Research Topic as guest editors, supervised the reviewing of the manuscript, and wrote this Editorial paper. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The guest editors are grateful to the authors contributing to this special issue papers collection and thank the reviewers who helped improve the manuscripts. The publication has been prepared with the support of the RUDN University Program 5-100.

REFERENCES

Baranova, A. V., Klimontov, V. V., Letyagin, A. Y., and Orlov, Y. L. (2019). Medical genomics research at BGRS-2018. *BMC Med. Genomics* 12(Suppl. 2):36. doi: 10.1186/s12920-019-0480-0

Bragin, A. O., Saik, O. V., Chadaeva, I. V., Demenkov, P. S., Markel, A. L., Orlov, Y. L., et al. (2017). Role of apoptosis genes in aggression revealed using combined analysis of ANDSystem gene networks, expression and genomic data in grey rats with aggressive behavior (In Russian). *Vavilov J. Genet. Breed.* 21, 911–919. doi: 10.18699/VJ17.312

Chen, M., Harrison, A., Shanahan, H., and Orlov, Y. (2017). Biological big bytes: integrative analysis of large biological datasets. *J. Integr. Bioinform.* 14:20170052. doi: 10.1515/jib-2017-0052

Das, R., and Upadhyai, P. (2018). An ancestry informative marker set which recapitulates the known fine structure of populations in South Asia. *Genome Biol. Evol.* 10, 2408–2416. doi: 10.1093/gbe/evy182

Fedorova, M. S., Krasnov, G. S., Lukyanova, E. N., Zaretsky, A. R., Dmitriev, A. A., Melnikova, N. V., et al. (2019). The CIMP-high phenotype is associated with energy metabolism alterations in colon adenocarcinoma. *BMC Med. Genet.* 20(Suppl. 1):52. doi: 10.1186/s12881-019-0771-5

Kolchanov, N. A., and Orlov, Y. L. (2013). Introductory note for BGRS-2012 special issue. *J. Bioinform. Comput. Biol.* 11:1302001. doi: 10.1142/S0219720013020010

Kozlov, K., Singh, A., Berger, J., Bishop-von Wettberg, E., Kahraman, A., Aydogan, A., et al. (2019). Non-linear regression models for time to flowering in wild chickpea combine genetic and climatic factors. *BMC Plant Biol.* 19(Suppl. 2):94. doi: 10.1186/s12870-019-1685-2

Orlov, M. A., and Sorokin, A. A. (2020). DNA sequence, physics, and promoter function: Analysis of high-throughput data On T7 promoter variants activity. *J. Bioinform. Comput. Biol.* 18:2040001. doi: 10.1142/S0219720020400016

Orlov, Y. L. (2019). 5-th International Scientific Conference of "Plant Genetics, Genomics, Bioinformatics, and Biotechnology" (24—29 June 2019, Novosibirsk, Russia). *J. Food Qual. Hazards Control* 6, 41–41. doi: 10.18502/jfqhc.6.1.458

Orlov, Y. L., Baranova, A. V., and Markel, A. L. (2016). Computational models in genetics at BGRS SB-2016: introductory note. *BMC Genet.* 17(Suppl. 3):155. doi: 10.1186/s12863-016-0465-3

Orlov, Y. L., Hofestädt, R., and Baranova, A. V. (2019a). Systems biology research at BGRS-2018. *BMC Syst. Biol.* 13(Suppl. 1):21 doi: 10.1186/s12918-019-0685-z

Orlov, Y. L., Hofestädt, R., and Tatarinova, T. V. (2019b). Bioinformatics research at BGRS SB-2018. *J. Bioinform. Comput. Biol.* 17:1902001. doi: 10.1142/S0219720019020013

Orlov, Y. L., Hofestädt, R. M., and Kolchanov, N. A. (2015). Introductory note for BGRS SB-2014 special issue. *J. Bioinform. Comput. Biol.* 13:1502001. doi: 10.1142/S0219720015020011

Orlov, Y. L., Salina, E. A., Eslami, G., and Kochetov, A. V. (2019c). Plant biology research at BGRS-2018. *BMC Plant Biol.* 19 (Suppl. 1):56. doi: 10.1186/s12870-019-1634-0

Oshchepkov, D., Ponomarenko, M., Klimova, N., Chadaeva, I., Bragin, A., Sharypova, E., et al. A (2019). Rat model of human behavior provides evidence of natural selection against underexpression of aggressiveness-related genes in humans. *Front. Genet.* 10:1267. doi: 10.3389/fgene.2019.01267

Ponomarenko, M., Rasskazov, D., Chadaeva, I., Sharypova, E., Drachkova, I., Oshchepkov, D., et al. (2020). Candidate SNP Markers of atherogenesis significantly shifting the affinity of TATA-binding protein for human gene promoters show stabilizing natural selection as a sum of neutral drift accelerating atherogenesis and directional natural selection slowing it. *Int. J. Mol. Sci.* 21:1045. doi: 10.3390/ijms21031045

Ponomarenko, M., Rasskazov, D., Chadaeva, I., Sharypova, E., Ponomarenko, P., Arkova, O., et al. (2017). SNP_TATA_Comparator: genomewide landmarks for preventive personalized medicine. *Front. Biosci.* 9, 276–306. doi: 10.2741/s488

Tatarinova, T. V., Chen, M., and Orlov, Y. L. (2019). Bioinformatics research at BGRS-2018. *BMC Bioinformatics* 20(Suppl. 1):33. doi: 10.1186/s12859-018-2566-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 November 2018
doi: 10.3389/fgene.2018.00547

Dynamical Modeling of the Core Gene Network Controlling Flowering Suggests Cumulative Activation From the *FLOWERING LOCUS T* Gene Homologs in Chickpea

Vitaly V. Gursky[1,2], *Konstantin N. Kozlov*[2], *Sergey V. Nuzhdin*[2,3] *and Maria G. Samsonova*[2]*

[1] Theoretical Department, Ioffe Institute, Saint Petersburg, Russia, [2] Systems Biology and Bioinformatics Laboratory, Peter the Great Saint Petersburg Polytechnic University, Saint Petersburg, Russia, [3] Molecular and Computational Biology, University of Southern California, Los Angeles, CA, United States

OPEN ACCESS

Edited by:
Yuriy L. Orlov,
Institute of Cytology and Genetics (RAS), Russia

Reviewed by:
Inna N. Lavrik,
Medizinische Fakultät, Universitätsklinikum Magdeburg, Germany
Filippo Geraci,
Consiglio Nazionale Delle Ricerche (CNR), Italy

***Correspondence:**
Maria G. Samsonova
m.g.samsonova@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: 25 August 2018
Accepted: 26 October 2018
Published: 20 November 2018

Citation:
Gursky VV, Kozlov KN, Nuzhdin SV and Samsonova MG (2018) Dynamical Modeling of the Core Gene Network Controlling Flowering Suggests Cumulative Activation From the FLOWERING LOCUS T Gene Homologs in Chickpea. Front. Genet. 9:547. doi: 10.3389/fgene.2018.00547

Initiation of flowering moves plants from vegetative to reproductive development. The time when this transition happens (flowering time), an important indicator of productivity, depends on both endogenous and environmental factors. The core genetic regulatory network canalizing the flowering signals to the decision to flower has been studied extensively in the model plant *Arabidopsis thaliana* and has been shown to preserve its main regulatory blocks in other species. It integrates activation from the *FLOWERING LOCUS T* (*FT*) gene or its homologs to the flowering decision expressed as high expression of the meristem identity genes, including *AP1*. We elaborated a dynamical model of this flowering gene regulatory network and applied it to the previously published expression data from two cultivars of domesticated chickpea (*Cicer arietinum*), obtained for two photoperiod durations. Due to a large number of free parameters in the model, we used an ensemble approach analyzing the model solutions at many parameter sets that provide equally good fit to data. Testing several alternative hypotheses about regulatory roles of the five *FT* homologs present in chickpea revealed no preference in segregating individual *FT* copies as singled-out activators with their own regulatory parameters, thus favoring the hypothesis that the five genes possess similar regulatory properties and provide cumulative activation in the network. The analysis reveals that different levels of activation from *AP1* can explain a small difference observed in the expression of the two homologs of the repressor gene *TFL1*. Finally, the model predicts highly reduced activation between *LFY* and *AP1*, thus suggesting that this regulatory block is not conserved in chickpea and needs other mechanisms. Overall, this study provides the first attempt to quantitatively test the flowering time gene network in chickpea based on data-driven modeling.

Keywords: chickpea, flowering time, FT genes, ICCV 96029, CDC Frontier, dynamical model

INTRODUCTION

The depleted genetic diversity of many domesticated agriculturally important plants is a common problem for breeders, providing an obstacle in developing new forms with desired features. One such feature important for domesticated chickpea (*Cicer arietinum*) is early flowering time, which enforces more rapid transition from vegetative to reproductive growth. Due to high sensitivity of chickpea to ascochyta blight, it is essential to reduce the full plant cycle, from sowing to maturation, in order to fit it to relatively short growing seasons having dry weather and, hence, low disease pressure (Kumar and Abbo, 2001). These growing seasons are quite short in major chickpea growing regions, pushing breeders to developing chickpea lines with early flowering time. Thus, it is important to identify key genes regulating floral transition and quantitatively understand the behavior of the flowering time gene network.

The floral transition has been intensively studied in model organisms, such as Arabidopsis (*Arabidopsis thaliana*) (Srikanth and Schmid, 2011; Andrés and Coupland, 2012), and in other plants, including important crops and legumes (Kumar and Abbo, 2001; Dong et al., 2012; Shrestha et al., 2014; Blümel et al., 2015; Peng et al., 2015; Weller and Ortega, 2015; Zhang et al., 2016; Ridge et al., 2017). Flowering starts in response to various environmental signals, including photoperiod and vernalization, and endogenous signals, such as autonomous and circadian clock, and molecular pathways have been identified conducting these signals to the core gene network that integrates them into a binary decision to flower. Despite the high complexity of these pathways and many unknown regulators, it has been shown that key genes regulating the process are conserved between species. In particular, the flowering signals lead to the elevated expression of the floral pathway integrator gene *FLOWERING LOCUS T* (*FT*), or its homologs, in the leaves (Kardailsky et al., 1999; Kobayashi et al., 1999; Pin and Nilsson, 2012; Jaeger et al., 2013).

In Arabidopsis, the understanding of the core gene network integrating the flowering signals transmitted via the expression of *FT* has evolved to the general scheme illustrated in **Figure 1A** (Jaeger et al., 2013). FT is a mobile factor transported from the leaves to the apical meristem, where it forms the complex with the transcription factor FD. This complex activates the meristem identity genes *LEAFY* (*LFY*) and *APETALA1* (*AP1*), which also activate each other. The expression of *AP1* activates genes controlling flower development and thus can be considered as the output of the network specifying the floral transition (Kaufmann et al., 2010). In order to keep the center of the shoot apical meristem in a vegetative state, the key floral repressor *TERMINAL FLOWER1* (*TFL1*) inhibits expression of *LFY* and *AP1* in this region. The resulting gene interaction graph takes the form shown in **Figure 1B**, incorporating evidence for some additional interactions: TFL1 acts as a repressor in the complex with FD, LFY activates *FD*, and AP1 represses *TFL1*. As many genes are omitted, each node in the graph in fact represents a group of genes (Jaeger et al., 2013).

The knowledge about the regulatory interactions between the genes from **Figure 1** has been obtained via extensive genetic studies, and it provides a unique opportunity for computational

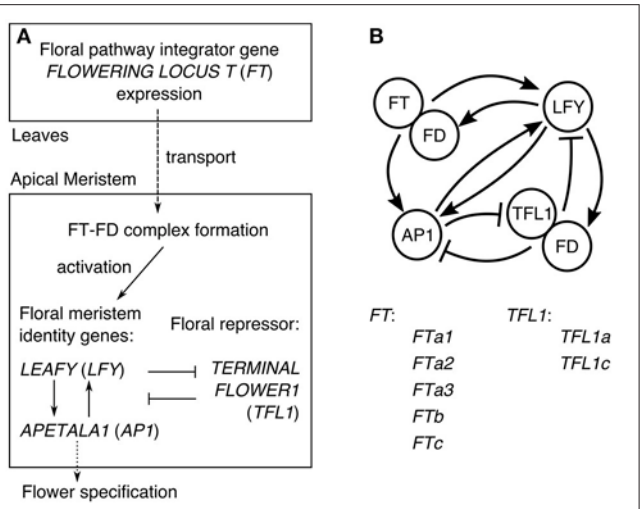


FIGURE 1 | The core gene network controlling floral transition. **(A)** The general scheme of processes involved in floral transition. **(B)** The graph of the regulatory interactions proposed for Arabidopsis, and the list of the *FT* and *TFL1* homologs in chickpea considered in our model. The interaction graph was adopted from (Jaeger et al., 2013).

modeling of this gene regulatory network, when experimental data on the system behavior is available. The modeling allows to gain mechanistic insights into specific properties of the floral transition system and produce testable predictions. Jaeger et al. (2013) elaborated a dynamical model of the core network from **Figure 1** based on the data on the flowering time for a set of the wild type and mutant Arabidopsis genotypes. They showed that the floral transition dynamics can be explained by splitting the network into several feedback and forward loops, each bearing a clear functional role (Pullen et al., 2013). Leal Valentim et al. (2015) studied a similar gene network, particularly considering that the complex TF-FD activates *LFY* via the intermediate transcription factors SOC1 and AGL24. They measured expression dynamics of all genes involved and used this data to calibrate a dynamical model. Using this data-driven approach, they tested various hypotheses about regulation of *LFY* by SOC1 and AGL24 and showed that perturbations can spread through the network in a nonlinear way.

A possibility to extend these results to chickpea depends on what we know about the inflorescence genes in this species. We concentrate on two chickpea cultivars in this study, CDC Frontier and ICCV 96029. CDC Frontier is a photoperiod-sensitive kabuli chickpea cultivar developed at the University of Saskatchewan (Warkentin et al., 2005), exhibiting relatively late flowering (Daba et al., 2016; Ridge et al., 2017). The reference genome sequence was obtained for this cultivar (Varshney et al., 2013). ICCV 96029 is a photoperiod-insensitive desi chickpea cultivar developed by the International Crops Research Institute for the Semi-Arid Tropics, India, representing the earliest flowering chickpea cultivar currently known. Quantitative trait loci associated with early flowering were investigated, and it was shown that a single recessive allele with some additional modifiers confer early flowering of ICCV 96029 (Kumar and van Rheenen, 2000; Gaur

et al., 2015; Upadhyaya et al., 2015; Mallikarjuna et al., 2017). Ridge et al. (2017) provided evidence that a mutation in an ortholog of the key circadian gene *ELF3* can be associated with earliness in ICCV 96029 under short day growth conditions, but their analysis of the expression of clock genes in ICCV 96029 did not reveal any clear differences for this cultivar.

In contrast to the single *FT* gene in Arabidopsis, Ridge et al. (2017) identified five *FT* homologs in chickpea: *FTa1*, *FTa2*, *FTa3*, *FTb*, and *FTc*, named according to affiliation with one of the three clades (*FTa*, *FTb*, and *FTc*). They also found two chickpea orthologs of *TFL1* (*TFL1a* and *TFL1c*). Furthermore, Ridge et al. (2017) measured the expression dynamics of the homologs of all genes from the core gene network for CDC Frontier and ICCV 96029 under two growth conditions (short day, SD, and long day, LD) and identified specific differences in expression between these genotypes. In particular, they noted that the up-regulation of *FT* and *AP1* expression was synchronous with floral bud initiation, thus confirming that regulation of floral transition in chickpea occurs via the *FT* gene family.

We aimed to investigate a possibility to extend the core gene network from **Figure 1** to chickpea. Assuming this network is conserved, we developed a dynamical model of gene expression and applied it to the previously published expression time series (Ridge et al., 2017). We used the resultant model to dissect interactions in which targets were found insensitive to regulator action. This points to chickpea specific deviations in regulation of floral transition. We also studied if the *TFL1* homologs are mutually distinguishable in the context of the model. Finally, we tested several hypotheses about how the *FT*-like genes combine in their activation of the meristem identity genes.

RESULTS

Model

We modeled the flowering time gene network shown in **Figure 1**. We formulated the model in terms of the ordinary differential equations in which the change rates of gene product concentrations are regulated by the activators and inhibitors via the Hill-type regulation functions (the model equations (1–5) are described in details in section Materials and Methods). The formulation of the model equations depends on how we combine the activation from the *FT*-like genes. The baseline model (model, or hypothesis, *H0*) assumes that the five *FT* homologs are mutually indistinguishable in their activation of the meristem identity genes (*LFY* and *AP1*). In this model, FD forms the complex with the total FT concentration equal to the sum of the protein concentrations from each *FT* homolog. The activation of *LFY* by the FT-FD complex is characterized in the model equations by the regulation function containing the following regulatory parameters: one Michaelis–Menten constant (K_8), one Hill parameter (n_8), and one maximal synthesis rate (v_8) (see equation (6) in section Materials and Methods), and a similar set of regulatory parameters quantify the activation of *AP1* by the total FT concentration. An alternative model (*H1*) assumes that only one of the five FT's is enough to activate transition to flowering, so the concentration of only that FT participates in the complex FT-FD and activates *LFY* and *AP1* (see equation (7) in section Materials and Methods for the case of *LFY* activation). In another alternative model (*H2*), we tried to distinguish a single *FT* gene from the other four assuming that this singled-out gene has the regulatory parameters distinct from the rest of the *FT* genes, while these FT's still activate cumulatively (like in model *H0*). The activation from the singled-out *FT* gene and the activation from the total concentration of the rest of the *FT* genes are represented in the model by two distinct regulation functions (see equation (8) in section Materials and Methods for the case of *LFY* activation). Models *H1* and *H2* have five possible versions, where each version is associated with one *FT* homolog separated from the other *FT*-like genes. We tested only four of them, excluding *FTa3* from the analysis due to its very low expression in both growth conditions.

We applied the models to describe the previously published dynamic expression data for all genes from the core network measured in two chickpea cultivars, ICCV 96029 and CDC Frontier (Ridge et al., 2017). We failed to find a good model solution for the expression data from CDC Frontier (the best solution is shown in **Supplementary Figure 1**; we also discuss possible reasons in Discussion). Therefore, the rest of the paper describes modeling results for ICCV 96029.

Parameter Estimation and Model Solutions for ICCV 96029

Models *H0* and *H1* have the same number of free parameters ($k = 31$), and model *H2* has six parameters more ($k = 37$). We estimated values of these parameters by minimizing the weighted sum of squared residuals quantifying the difference between the model solution and the ICCV 96029 data for the two growth conditions (SD and LD) simultaneously (section Materials and Methods). The data comprised expression levels of five genes (*TFL1a*, *TFL1c*, *FD*, *LFY*, and *AP1*) in ICCV 96029 on 7 days under SD and LD, with the total number of data points equal to $m = 70$. After estimating the parameter values, we applied the Akaike information criterion corrected for small data samples for model comparison, as described further in the text.

As k was relatively large, we refused to estimate the parameter values by fitting the model to the data from one condition (either LD or SD) and testing on the data from the other condition. In that case, the number of parameters k in model *H2* would exceed the number of data points ($m = 35$ in LD or SD) and k in other model versions would be close to m, and that would complicate the application of the Akaike information criterion for model comparison. As a control, we performed the fitting to the LD data and tested on the SD data in model *H0* and made sure that the corresponding solutions were qualitatively similar to the two-conditions fitting results (**Supplementary Figure 2**).

We further circumvented an overfitting potential of the two-conditions fitting applying the ensemble approach in the analysis of model behavior (Samee et al., 2015). In this approach, all sets of parameter values and solutions resulted from the fitting procedure were considered as equally suited for biological conclusions, and the conclusions were derived based on the analysis of the whole ensemble of the solutions and optimized parameter values.

The parameter optimization under hypothesis *H0* resulted in the model solutions of very similar quality (**Figure 2**; distributions of the estimated parameter values are shown in **Supplementary Figure 3**). The model correctly reproduces the main characteristics of the data. The dynamic increase of LFY and AP1 concentrations can be explained by activation from the rising expression of the *FT* genes. LFY activates *FD*, resulting in the dynamic increase of its expression. Finally, the floral repressors TFL1a and TFL1c decrease in time due to repression by AP1.

Reduced *LFY* and *AP1* Activation

The solution in **Figure 2** shows somewhat insufficient expression levels of both *LFY* under SD and *AP1* under LD. The analysis of the expression data reveals that LFY behaves rather counterintuitively under SD as compared with LD and differs in this behavior from AP1. Namely, *LFY* is down-regulated in LD compared to SD, despite the increased activation from the raising expression of the *FT* genes in LD compared to SD, and this holds both for ICCV 96029 and CDC Frontier (**Figure 3**). In contrast, the integral expression of *AP1* increases from SD to LD in accordance with the rising activation from FT. This anticorrelation between LFY and its sole activators (FT and AP1) observed in the data hampers the model in finding a better solution.

We analyzed how LFY and other transcription factors are involved in their regulations in the model for ICCV 96029 by plotting average values of the Hill functions which implement in the model equations each regulatory interaction from the gene network (**Figure 4**). An active regulation tends to keep the Hill function value between 0 and 1, while the limit values (0 or 1) evidence that the interaction between genes is saturated, with no sensitivity to specific expression levels of the regulators. This type of saturation occurs for activation of *LFY* by AP1, with the corresponding Hill function values pushed to zero. Activation of *AP1* by LFY is also characterized by the Hill function values close to zero, but the analysis of the Jacobian values of the right-hand side of the model equations for this regulation still shows relatively high LFY influence on *AP1* (**Supplementary Figure 4**). Another saturated regulation involving LFY is activation of *FD*. At the same time, *LFY* is sensitive to its repressors (the complexes TFL1a-FD and TFL1c-FD), in contrast to the

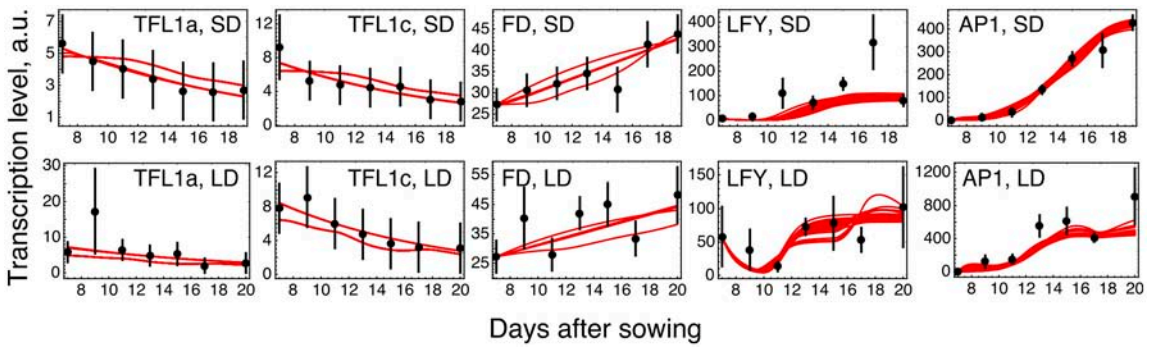


FIGURE 2 | Model *H0* solutions for ICCV 96029 under two growing conditions. The model solutions (red curves) corresponding to all parameter sets found by optimization are shown for five flowering time genes and for the short day (SD, upper panels) and long day (LD, lower panels) conditions. The black dots and error ranges are the mean expression data and standard deviation, respectively, taken from (Ridge et al., 2017).

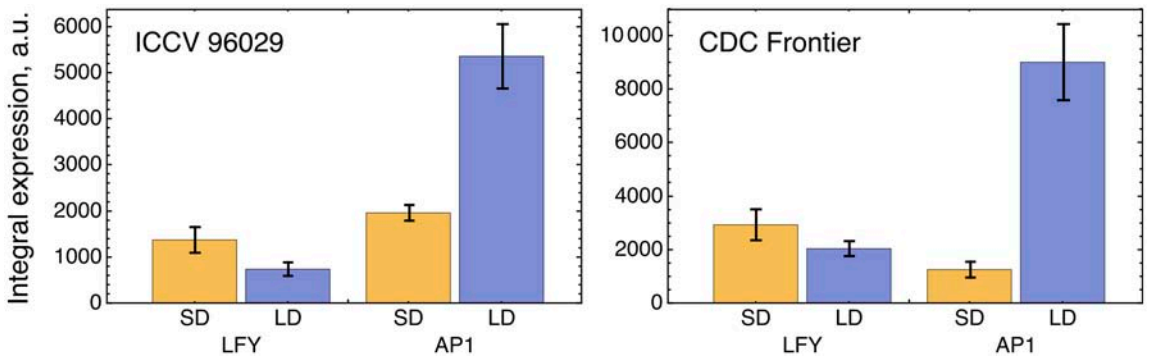


FIGURE 3 | Integral expression levels of *LFY* and *AP1* under two growth conditions in two cultivars, based on the data from (Ridge et al., 2017). At each time where data was available, 100 expression values were sampled from the normal distribution with the mean and s.d. presented at this temporal point in the data. These values then were interpolated across time, producing a set of 100 expression dynamics, and these dynamics were integrated over time. The chart and error bars show means and standard deviations, respectively, over this set of the integral values.

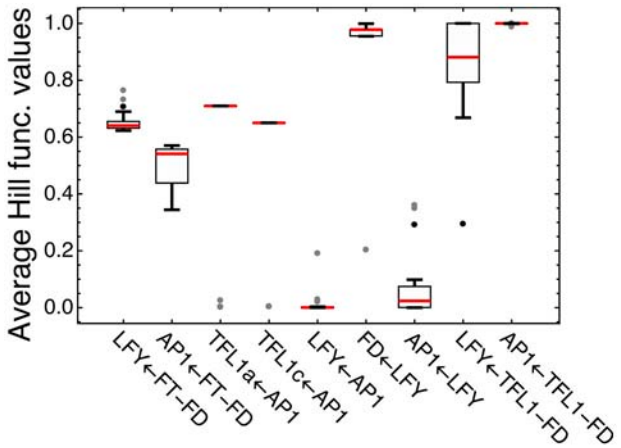


FIGURE 4 | Average values of the regulation functions characterizing regulation in the model for ICCV 96029. For each set of the optimized parameter values, the averaged values of the regulation functions from the model equations were obtained by integrating these functions over time under SD and LD and dividing by the integration time interval; the figure shows box plots of the distributions of these values over all sets of the optimized parameter values. The type of regulation corresponding to each regulation function is shown on the horizontal axis, where arrow indicates the direction of the regulation. FT-FD and TFL1-FD denote the complexes of FD with all FT and two TFL1 proteins, respectively. Dots show outliers.

saturated repression of *AP1* by these complexes (**Figure 4**). Overall, this analysis of the model and expression data suggests that there are regulators of *LFY* missing in the core gene network under study.

Figure 4 shows four regulations characterized by the average Hill function values that are considerably far from the saturation limits: activation of *LFY* and *AP1* by FT and repression of *TFL1a* and *TFL1c* by AP1. This fact allows us to use the model for testing various alternative hypotheses about these regulations.

Difference in *TFL1a* and *TFL1c* Expression can be Explained by Different Regulation by AP1

We tested a hypothesis that a small difference in *TFL1a* and *TFL1c* expression observed in the data (**Figure 5**) can be explained by different regulation by AP1. Because of this difference in the expression, we included TFL1a and TFL1c in the model as two distinct dynamical variables whose dynamics are under control of the following four parameters per factor (equations (1–2) in section Materials and Methods): maximal expression rate v_i, dissociation constant K_i, cooperativity parameter n_i, and degradation rate λ_i ($i = 1,2$). If the model fitting produced no significant difference in these parameters between TFL1a and TFL1c, there would be no means to distinguish between these factors in the model and we would have to consider a single dynamical variable TFL1 = TFL1a + TFL1c instead. If the difference in parameter values exists, there is an interesting question about whether this difference can be explained by different regulation from AP1. If AP1 is indeed involved, a statistically significant difference should exist between values of the regulatory parameters K_1 and K_2 and/or between values of n_1 and n_2, because these parameters are associated with repression of *TFL1a* and *TFL1c* by AP1. A possible difference in v_i and/or λ_i should be attributed to other, AP1 independent, factors.

The optimized parameter values for TFL1a and TFL1c form two clearly separated clusters, which correspond to the main box ("main cluster") and the outliers ("outlying cluster") in the AP1→TFL1a and AP1→TFL1c parts of **Figure 4**, and it is already seen in this figure that the regulation by AP1 differs between the analyzed target genes within the main cluster. The Hill exponents n_i are the same in the main cluster for both TFL1a and TFL1c ($n_i = 1$, $i = 1,2$), but we see the significant difference in K_i values in this cluster: $K_1 = 561.14 \pm 0.13$ (TFL1a) and $K_2 = 401.14 \pm 0.08$ (TFL1c) (p-value $= 2 \times 10^{-9}$). Therefore, the model suggests different regulatory properties of AP1 in its action on the genes *TFL1a* and *TFL1c*, linked to possible different association kinetics to their promoters. The outlying cluster is characterized by a small influence of AP1 and contain only from 5 to 6 parameter sets with very similar K_i and n_i values, so we consider this cluster as not relevant.

Model Suggests Cumulative Activation by the *FT* Homologs

We tested whether an individual *FT* gene stands out against the other *FT* homologs by fitting the three versions of the model (models *H0*, *H1*, and *H2*) described above and in Materials and Methods, with subsequent comparison of their fitting quality. We considered only four of the five *FT* genes in the tests excluding *FTa3*, since its expression was small relative to the other ones (**Figure 6A**).

We first checked if a single *FT* gene can provide the full activation from the *FT* gene family in the network, thus serving as a unique transmitter of the flowering signal (model *H1*). Under this assumption, we replaced the sum of FT concentrations in the model equations by the concentration of one of the four FT's and fitted each resulted version of the model to the expression data for ICCV 96029. For each tested *FT* gene, model *H1* demonstrated worse fitting quality as compared to the baseline model with the cumulative activation from all *FT* genes (model *H0*) (**Figure 6B**; p-value $= 3 \times 10^{-7}$ for FTa1 as the sole activator; 7×10^{-9}, FTa2; 2×10^{-5}, FTb; 10^{-4}, FTc). Breaking the cost function into the separate SD- and LD-related components reveals that all versions of model *H1* have worse quality in description of the LD data and all except the *FTa2*- and *FTc1*-related models have worse description of the SD data (**Supplementary Figure 5**). Since models *H0* and *H1* have the same number of parameters, neither of them is prone to overfitting to a larger extent than the other one, and, hence, we can conclude about better relevance of model *H0* based on the fitting quality comparison and without applying additional quality measures.

As several *FT* genes are required for better description of the expression data, a question yet remains about whether different FT's activate the meristem identity genes differently in terms of their regulatory parameters. We implemented this possibility in model *H2* by singling an FT out from the other four and adding a new regulation function to the model equations representing the

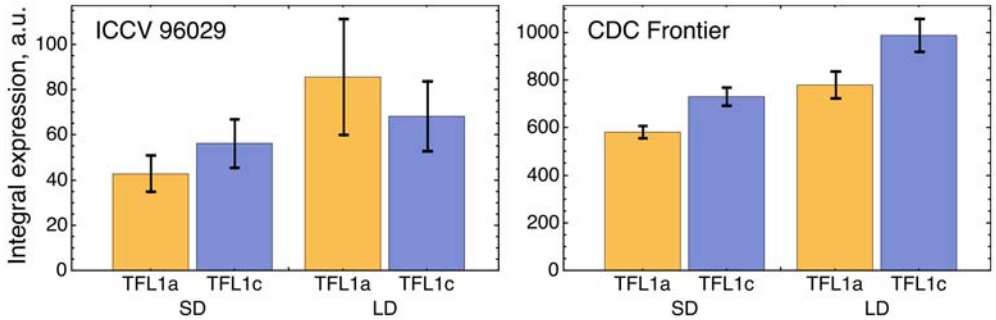


FIGURE 5 | Integral expression levels of *TFL1a* and *TFL1c* under two growth conditions in two cultivars, based on the data from (Ridge et al., 2017). The integral expression levels were calculated as described in **Figure 3**.

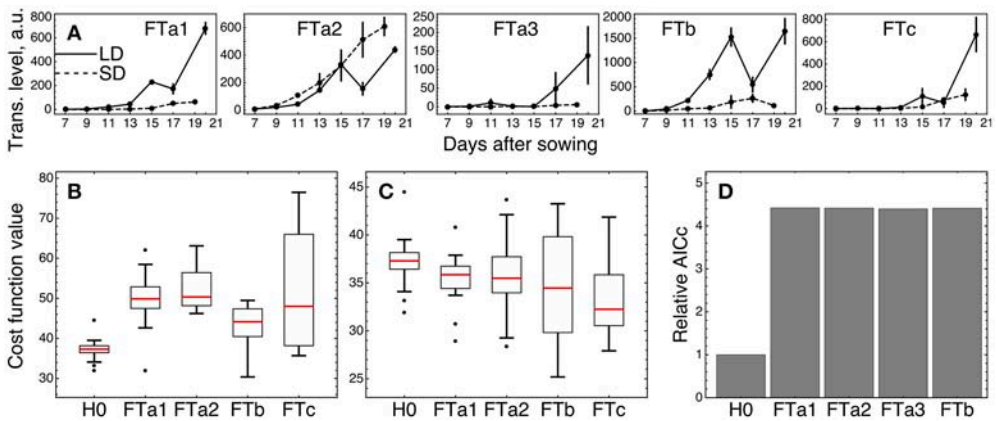


FIGURE 6 | Testing alternative hypotheses on regulation by the *FT* genes in ICCV 96029. **(A)** Expression data of the *FT* genes in ICCV 96029 under SD and LD; reproduced from (Ridge et al., 2017). **(B)** Values of the cost function (weighted residual sum of squares; equation (9) in Materials and Methods) quantifying the goodness of fit for model *H0* and four versions of model *H1*, for all optimized parameter sets. Each version of model *H1* is marked on the bottom of the panel by the name of the *FT* gene participating as a sole FT activator in the model. **(C)** The same as in **(B)**, but for model *H2*. Each version of model *H2* is marked on the bottom of the panel by the name of the *FT* gene singled out in the model equations from the other *FT* genes. **(D)** Akaike information criterion corrected for small data samples (AICc; equation (10) in Material and Methods) for *H0* and four versions of model *H2*, marked as in **(C)**. The relative values of AICc normalized to the *H0* value are shown. The use of a more conventional form of AICc yields a similar figure (**Supplementary Figure 8** and **Supplementary Text**).

activating action of this FT with its own regulatory parameters (v, K, and n), while preserving in the equations the activation from the sum of the other FT concentrations. Model *H2* exhibited a better fitting quality than *H0* for the singled-out genes *FTa1* (p-value = 0.005) and *FTc* (p-value = 0.0004), with no improvement for the other two *FT* genes (p-value = 0.09 for the singled-out *FTa2* and 0.12 for *FTb*) (**Figure 6C**). Both *FTa1*- and *FTc*-related models *H2* demonstrate better fit to the LD-data, with no significant improvements in fits to the SD-data (**Supplementary Figure 6**).

We can try to find features in the expression of *FTa2* and *FTb* that can be attributed to their worse individual performance in the model. **Figure 6A** shows that the expression dynamics of *FTa2* is almost identical under SD and LD for a long time and becomes down-regulated under LD at later days, in contrary to the behavior of all other FT's and to the up-regulation of *AP1* in LD (**Figure 3**). At the other extreme, the up-regulation of *FTb* in LD is the strongest among the *FT* genes, and this raise in expression might be too large to represent the difference between SD and LD adequately. However, model *H1* with FTb as the only FT activator performs best among all *FT* genes on average (**Figure 6B**), and both *FTb*-related models (*H1* and *H2*) provide the lowest cost function values among all models, including *H0* (see the minimal cost values in **Figures 6B,C**), which hints at possible importance of this gene.

The observed better performance of models *H2* with the singled-out genes *FTa1* and *FTc* can be related to overfitting, since model *H2* has six parameters more than the baseline model *H0*. We controlled this by evaluating the Akaike information criterion corrected for small data samples (AICc; equation (10) in section Materials and Methods), which assesses the quality of a model applied to a data by combining the fitting quality of the model and its complexity in terms of the number of free parameters. Smaller values of this measure correspond to better models. AICc evaluation reveals that its value for each version of model *H2* is more than four times larger than for model

H0 (**Figure 6D**), which suggests that the complexity added to model *H2* is not justified by the resulted improvement in fitting. Therefore, we conclude that the model with the cumulative activation from all *FT* genes (model *H0*) is the most relevant for the given expression data.

DISCUSSION

We presented a computational model of the core gene network controlling the floral transition and investigated its ability to describe the expression data in two chickpea cultivars. We were able to find good model solutions for ICCV 96029, which suggests a general conservation of the core gene network from **Figure 1** in this chickpea cultivar. On the other hand, the modeling results were negative for CDC Frontier. A possible reason for this could be related to the specific choice of the modeling formalism. This explanation does not seem likely, since the modeling formalism is quite general and has been successfully applied to the same gene network in Arabidopsis (Leal Valentim et al., 2015). Another explanation which we find more probable is that this gene network is more perturbed in CDC Frontier than in ICCV 96029.

Several key differences between CDC Frontier and ICCV 96029 were reported based on the analysis of the expression data (Ridge et al., 2017): ICCV 96029 exhibits much earlier and much stronger up-regulation of the expression of *AP1*, according to the earlier appearance of visible floral buds as compared to CDC Frontier. The floral repressors *TFL1a* and *TFL1c* have lower expression levels in ICCV 96029 than in CDC Frontier, also in accordance with the early flowering of the former. On the other hand, the differences in expression of *FD* and *LFY* are not as visible between the cultivars.

The expression levels of the *FT* genes in the data are significantly different for the two cultivars, and the total FT concentration in CDC Frontier can be estimated as close to the background levels (**Figure 7A**). This can partially explain why the model is not feasible for the expression data from CDC Frontier. Such small FT levels could possibly be related to the observed fact that the first floral buds, appeared in CDC Frontier at 31 days after sowing in SD and at 32 days in LD, were abortive, although the low expression of some of these genes persisted for much longer time (Ridge et al., 2017). Furthermore, investigation of the autocorrelation functions of the *FT* expression time series reveals very different patterns in the FT signals between the cultivars (**Figure 7B**), and these patterns are translated to the rest of the core network genes almost without changes (**Figure 7C**). It is interesting to note a periodic signal in the FT dynamics in CDC Frontier with a period of two days, although this signal can yet be an experimental artifact related to low expression levels.

Another important difference between the cultivars that we see in the data and that might contribute to the difference in the modeling results concerns the dependence between concentrations of TFL1a/TFL1c and LFY/AP1. TFL1a and TFL1c repress *LFY* and *AP1*, and AP1 represses the *TFL1*-like genes (Ratcliffe et al., 1999; Kaufmann et al., 2010). Therefore, we should expect that these two groups of transcripts should avoid coexistence in the data and, hence, exhibit a negative correlation over time. We do see this correlation in the data from ICCV 96029, but not from CDC Frontier (**Table 1**). Moreover, **Table 1** shows that these mutual repressors tend to show a positive correlation in the CDC Frontier data. Regardless of whether this inconsistency in the CDC Frontier data should be attributed to an artifact or it hints at alternative regulations between the *TFL1*-like genes and the inflorescence identity genes in this cultivar, this property evidently impedes the modeling success under given assumptions.

It has been shown that *LFY* is involved in positive regulation of *AP1* and is positively regulated by AP1 in Arabidopsis (Wagner et al., 1999; Jaeger et al., 2013; Leal Valentim et al., 2015). Our modeling results suggest that some additional factors should exist providing insufficient activation of these genes in the model for chickpea. The counterintuitive increase in the integral expression of *LFY* under SD as compared with LD, contrary to the decreasing activation from the *FT*-like genes, may indicate that additional activators of *LFY* participate under SD and compensate the missing activation. We believe that the absence of such factors in the core gene network considered in our model and, as a consequence, the inability to properly handle the LD

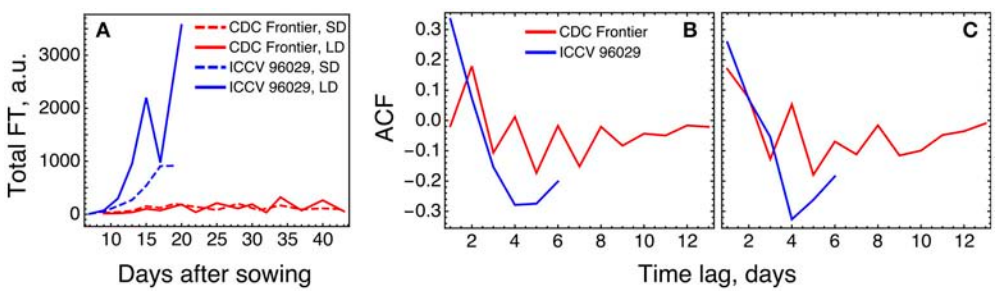


FIGURE 7 | Difference in FT behavior between ICCV 96029 and CDC Frontier, based on the expression data from (Ridge et al., 2017). **(A)** The dynamics of the sum of concentrations of all five FT transcripts, for the two cultivars and two growth conditions. Developing floral buds were first detected at 15 days (under SD) and 13 days (LD) in ICCV 96029 and at 31 days (SD) and 32 days (LD) in CDC Frontier (Ridge et al., 2017). **(B)** Autocorrelation function (ACF) for the expression data time series of the *FT* genes. ACF estimates similarity (correlation) between data points as a function of the time lag between them. For each time lag value, an ACF value was calculated for the expression time series for each *FT* gene and growth condition (SD and LD), and then an average ACF was calculated over the *FT* genes and conditions. **(C)** The same as in **(B)** but for the expression dynamics of the genes *TFL1a*, *TFL1b*, *FD*, *LFY*, and *AP1*.

TABLE 1 | Correlations between the expression dynamics of TFL1a/TFL1c and LFY/AP1 in the data from (Ridge et al., 2017).

	ICCV 96029		CDC Frontier	
	SD	**LD**	**SD**	**LD**
TFL1a vs. LFY	−0.89 ($P < 0.01$)*	−0.57 ($P = 0.10$)	0.80 ($P < 0.01$)*	0.14 ($P = 0.36$)
TFL1a vs. AP1	−0.89 ($P = 0.01$)*	−0.64 ($P = 0.07$)	0.18 ($P = 0.31$)	−0.14 ($P = 0.41$)
TFL1c vs. LFY	−0.61 ($P = 0.10$)	−0.64 ($P < 0.01$)*	0.81 ($P < 0.01$)*	0.33 ($P = 0.13$)
TFL1c vs. AP1	−0.96 ($P < 0.01$)*	−0.86 ($P < 0.01$)*	0.22 ($P = 0.20$)	0.13 ($P = 0.28$)

The Spearman rank correlation coefficient ρ was calculated for each cultivar (CDC Frontier and ICCV 96029) and growth condition (SD and LD). The p-values (P) were calculated by one-tailed permutation test, and the p-values below 0.05 are marked with asterisk.

vs. SD changes in expression is the reason why AP1 is almost excluded as an activator of *LFY* in the model solutions. In other words, this allows for the hypothesis that the *LFY-AP1* regulation module is not conserved in chickpea. However, we should also consider the possibility that the LD vs. SD increase in expression of *LFY* is due to insufficient quality of the data. Future work, both modeling and experimental, should clarify this point.

Since ICCV 96029 is day length neutral and floral transition is conferred via the *FT* genes, we might expect no difference in *FT* expression between SD and LD treatments in this cultivar. However, the expression data by Ridge et al. (2017) shows an essential difference in expression of these genes (**Figures 6**, **7A**), and it is important that this difference is transferred to the SD/LD difference in expression of *AP1* (**Figure 3**), so that the key gene specifying flower meristem identity exhibits sensitivity to photoperiod according to the data. This expression data was collected from the plants with first visible floral buds appeared at 15 days after sowing in SD and 13 days in LD (Ridge et al., 2017), thus providing the two days difference in floral bud initiation time between SD and LD. This two days difference diverges from previous measurements showing no difference in this time in ICCV 96029 (19 days from seeding ± 0.0) (Daba et al., 2016), but it qualitatively matches with the observed difference in expression.

Irrespective of whether this match is confident or not, the observed raise in expression of the *FT* genes and *AP1* in LD suggests that some compensatory mechanisms, or missing repressors, should exist diminishing the influence of that extra expression on the time to flower. It is reasonable to presume that these mechanisms should operate in the post-inductive phase of flower development, as they take the increased expression of floral meristem identity genes as the input. However, this conjecture is not in correspondence with the previously observed fact that ICCV 96029 does not exhibit photoperiod sensitivity on any of the pre-, inductive, or post-inductive phases of flower development (Daba et al., 2016). We believe this expression-based photoperiod sensitivity effect in ICCV 96029 is a fascinating subject for further studies.

An important difference of legumes and other species from Arabidopsis is in multiple orthologs of the inflorescence genes, such as *FT*, that present in a single copy in Arabidopsis (Pin and Nilsson, 2012). The regulatory roles of individual copies can sometimes be separated from the others; for example, *FTb* has been shown to have the leading role in pea (Hecht et al., 2011). The main purpose of our modeling approach was to infer possible differences in regulatory roles or other properties associated with the five *FT* homologs and two *TFL1* homologs in chickpea (Ridge et al., 2017). It is important that the model and expression data in principle allow to perform such inference, as the fitting results reveal that both *FT*- and *TFL1*-like genes are involved in active regulations.

AP1 was shown to repress *TFL1*-like genes (Liljegren et al., 1999; Kaufmann et al., 2010; Jaeger et al., 2013), and we found that this repression can be different for *TFL1a* and *TFL1c* in chickpea. As this difference concerns only the values of the equilibrium dissociation constant *K*, we can suggest that AP1 has different binding properties to the promoters of *TFL1a* and *TFL1c*.

Visual comparison between the expression of the five *FT*-like genes in ICCV 96029 does not help in differentiating their regulatory properties. Our modeling results support the cumulative activation model, in which all FT proteins have very similar regulatory properties and activation of the meristem identity genes occurs via the total FT concentration. Analyzing their expression data, Ridge et al. pointed at *FTb* as particularly important for induction of flowering (Ridge et al., 2017). However, this gene becomes indistinguishable from the others if we put it in the modeling context. The ensemble of model fits in which this gene is singled out does not improve the model, and we get the same conclusions using the Akaike information criterion to assess the relative performance of the model. On the other hand, we found that singling *FTb* out produced the lowest values of the minimal cost in all types of the computational experiments, suggesting that its potential of being the leading *FT* activator is not exhausted and is not seen only due to possible imperfections of the model and/or data.

As any modeling approach, our model has limitations. Perhaps the most important one concerns the large number of free parameters. We tackled this inevitable problem by utilizing the ensemble approach in the analysis of the model behavior (Samee et al., 2015). Despite the existing interdependence between the model parameters, the optimized parameter values led to the set of very similar solutions for ICCV 96029. We drew any conclusions only based on the average over the ensemble of the optimized parameter values, thus utilizing the "wisdom of the crowd" principle. We note that, for example, both the model with the single *FTb* and the model with the singled-out *FTb* provide the minimal costs among all alternative models, while they do not

perform better on average. Even with the given number of free parameters, the model was not able to reproduce the expression data from CDC Frontier, which, in particular, indicates that we cannot fit any data. Therefore, we believe that the ensemble approach increases the confidence of our results.

MATERIALS AND METHODS

Model Equations

We model the expression of *TFL1a*, *TFL1c*, *FD*, *LFY*, and *AP1* with the following set of differential equations:

$$\frac{du_{TFL1a}}{dt} = v_1 \frac{{K_1}^{n1}}{{K_1}^{n1}+{u_{AP1}}^{n1}} - \lambda_1 u_{TFL1a}\,, \tag{1}$$

$$\frac{du_{TFL1c}}{dt} = v_2 \frac{{K_2}^{n2}}{{K_2}^{n2}+{u_{AP1}}^{n2}} - \lambda_2 u_{TFL1c}\,, \tag{2}$$

$$\frac{du_{FD}}{dt} = v_3 \frac{{u_{LFY}}^{n3}}{{K_3}^{n3}+{u_{LFY}}^{n3}} - \lambda_3 u_{FD}\,, \tag{3}$$

$$\frac{du_{LFY}}{dt} = \left(v_4 \frac{{u_{AP1}}^{n4}}{{K_4}^{n4}+{u_{AP1}}^{n4}} + f_{FT\to LFY}(t)\right) \times \left(\frac{{K_5}^{n5}}{{K_5}^{n5}+[u_{FD}(u_{TFL1a}+u_{TFL1c})]^{n5}}\right) - \lambda_4 u_{LFY}, \tag{4}$$

$$\frac{du_{AP1}}{dt} = \left(v_5 \frac{{u_{LFY}}^{n6}}{{K_6}^{n6}+{u_{LFY}}^{n6}} + f_{FT\to AP1}(t)\right) \times \left(\frac{{K_7}^{n7}}{{K_7}^{n7}+[u_{FD}(u_{TFL1a}+u_{TFL1c})]^{n7}}\right) - \lambda_5 u_{AP1}, \tag{5}$$

where u's describe the protein concentrations, v_i are the maximal protein synthesis rates, K_i are the Michaelis–Menten constants (which can be seen as the equilibrium dissociation constants for the regulators binding the target gene promoters in the case of a direct transcriptional regulation), ni are the Hill constants (accounting for the cooperative effects), and λ_i are the protein degradation constants. We do not model translation explicitly, but instead assume that protein concentrations are proportional to mRNA concentrations for simplicity.

The specific form of the equations is chosen according to the regulatory graph in **Figure 1** and can be read as follows. The last terms on the right-hand side of all the equations represent degradation of each protein. The first term on the right-hand side of equation (1) is the regulation function describing repression of *TFL1a* by AP1. The same regulation function but with different parameters describes repression of *TFL1c* by AP1 in equation (2). The first term on the right-hand side of equation (3) represents activation of *FD* by LFY. The first brackets in equation (4) contains the sum of the activating inputs to *LFY* expression from AP1 (the first term in the sum) and the FT homologs (the function $f_{FT\to LFY}(t)$, described below). This input is multiplied by the regulation function in the second brackets of this equation, representing repression of *LFY* by the FD-TFL1 complex. This repression is represented under the assumption that TFL1a and TFL1c have equivalent regulatory properties, and the concentration of the complex is proportional to the product of the FD concentration (u_{FD}) and the total concentration of TFL1a and TFL1c ($u_{TFL1a}+u_{TFL1c}$). The first brackets in equation (5) contains the sum of the activating inputs to *AP1* expression from LFY (the first term in the sum) and the FT homologs (the function $f_{FT\to AP1}(t)$, described below). This input is multiplied by the regulation function in the second brackets of this equation, representing repression of *AP1* by the FD-TFL1 complex.

We test three alternative hypotheses (*H0*, *H1*, and *H2*) about functions $f_{FT\to LFY}$ and $f_{FT\to AP1}$. Under the null hypothesis *H0*, we assume regulatory equivalence of the five *FT* homologs, so the total concentration of all FT proteins forms the complex with FD and activate *LFY* and *AP1* with a single Michaelis–Menten constant and a single Hill constant, according to the following expression:

$$\boldsymbol{H0}:\ f_{FT\to LFY}(t) = v_6 \frac{\left[u_{FD}\sum_{i=1}^{5} u_i(t-\tau)\right]^{n8}}{{K_8}^{n8}+\left[u_{FD}\sum_{i=1}^{5} u_i(t-\tau)\right]^{n8}}, \tag{6}$$

$$u_1 = u_{FTa1},\ \ u_2 = u_{FTa2},\ \ u_3 = u_{FTa3},\ \ u_4 = u_{FTb},\ \ u_5 = u_{FTc}\,,$$

and a similar expression for the function $f_{FT\to AP1}$ with the *AP1*-related constants v_7, K_9, and $n9$. The FT concentrations in equation (6) are calculated with a time delay τ, which is taken to transport FT from the leaves to the apical meristem.

In the hypothesis *H1*, we assume that a single *FT* gene (with index k) is capable to fully represent the FT-mediated activation of *LFY* and *AP1*:

$$\boldsymbol{H1}:\ f_{FT\to LFY}(t) = v_6 \frac{[u_{FD}u_k(t-\tau)]^{n8}}{{K_8}^{n8}+[u_{FD}u_k(t-\tau)]^{n8}}\,, \tag{7}$$

and a similar expression for the function $f_{FT\to AP1}$ with the same u_k and with the *AP1*-related constants v_7, K_9, and $n9$.

Under the hypothesis *H2*, we assume that a member u_k of the FT family is distinguishable from the rest four members of the family in terms of regulation of *LFY* and *AP1*, so that we can separate it into a distinct regulation function with its own regulatory constants as follows:

$$\boldsymbol{H2}:\ f_{FT\to LFY}(t) = v_6 \frac{\left[u_{FD}\sum_{i\neq k}^{4} u_i(t-\tau)\right]^{n8}}{{K_8}^{n8}+\left[u_{FD}\sum_{i\neq k}^{4} u_i(t-\tau)\right]^{n8}} + v_7 \frac{[u_{FD}u_k(t-\tau)]^{n9}}{{K_9}^{n9}+[u_{FD}u_k(t-\tau)]^{n9}}\,, \tag{8}$$

and a similar expression for the function $f_{FT\to AP1}$ with the *AP1*-related constants v_8, v_9 K_{10}, K_{11}, $n10$, and $n11$. The first

term in equation (8) describes the cumulative activation from four FT proteins distinct from the FT protein with index k, whose activating input is represented by the second term in this equation. Depending on which gene of the *FT* family is singled out in the described way, we have five possible forms of $f_{FT\to LFY}$ and $f_{FT\to AP1}$ to test under hypothesis *H2*.

We solved numerically equations (1–5) replacing the concentrations of all regulators in the right-hand side of the equations with their expression data values interpolated in time. This effectively splits the model into four independent parts which do not contain common parameters: single equations for TFL1a, TFL1c, and FD, and the system of two equations for LFY and AP1 sharing the common parameter τ. The initial conditions for all proteins except TFL1a and TFL1c were equal to the value of each transcript at the first available day from the expression data (Ridge et al., 2017). Setting the initial conditions for TFL1a and TFL1c in the same way led to undesirable artifacts in the solutions resulted from the fitting procedure (**Supplementary Figure 7**); therefore, the initial conditions for these proteins were set to zero at $t = 0$, and the functions in the right-hand side of the model equations were obtained by interpolating the data values back to zero concentrations at $t = 0$. Numerical solution was obtained using either the *ode23s* solver in Octave or the *NDSolve* function in Wolfram Mathematica.

Parameter Estimation

The model contains 31 free parameters (7 v_i's, 9 K_i's, 9 ni's, 5 λ_i's, and τ) under hypothesis *H0* and in each version of the model under hypothesis *H1*, and there are six more parameters in *H2*. For the ICCV 96029 cultivar, the parameter values were found by minimizing the following weighted residual sum of squares (*wRSS*):

$$wRSS = \sum_{g=1}^{5}\sum_{k=1}^{T} \frac{\left(u_g\left(t_k\right) - u_g^{dat}\left(t_k\right)\right)^2}{{\sigma_{g,k}}^2}, \qquad (9)$$

in which the difference between the model solution u_g for genes g and the data u_g^{dat} is summed over all genes and over T times at which the data is available; $\sigma_{g,k}$ is the standard deviation of the data for gene g and time t_k. For fits to the CDC Frontier data, *wRSS* was additionally complemented with a penalty term equal to the covariance between the model solution and data.

The model fitting was performed either to the LD data only (and the SD data was used for testing) or to the joint LD and SD data, in which case *wRSS* from equation (9) should be calculated for the two growth conditions and summed. In the case of the LD fits, there were 35 data points in total for ICCV and 75 data points for CDC Frontier. In the case of fits to the joint SD and LD data, there were 70 and 145 data points for ICCV and CDC Frontier, respectively. The expression data for the five genes under modeling and the five *FT* homologs in chickpea was obtained from **Figure 5** of the paper by Ridge et al. (2017). The figure was digitized by the web-based tool WebPlotDigitizer (Rohatgi, 2018; the extracted expression data is available at https://zenodo.org, DOI:10.5281/zenodo.1451748). The cost functional was minimized by the differential evolution, which is a global parameter search method, using either a wolframscript program utilizing *NMinimize* function in Wolfram Mathematica or an entirely parallelized version of the method implemented in the DEEP software (Kozlov et al., 2016).

We assessed the quality of the alternative models *H0–H2* using the Akaike information criterion adjusted for small data samples:

$$AICc = 2k - 2\log\hat{L} + \frac{2k^2+2k}{m-k-1}, \qquad (10)$$

where k is the number of parameters in a model, m is the number of data points used for model fitting, and $\hat{L}$ is the maximum value of the likelihood function. In our case, $2log\ \hat{L} = -wRSS_{\min}$ — the minimal value of the *wRSS* functional from equation (9) estimated from the set of model fits (see **Supplementary Text** for derivation of $\hat{L}$). We also used a classical likelihood function appearing in least squares fitting.

AUTHOR CONTRIBUTIONS

MS and SN conceived and coordinated the project. VG and KK conducted the computational experiments. VG analyzed and summarized the results and wrote the first draft of the manuscript. All the authors participated in finalizing the manuscript.

FUNDING

The work was supported by the Russian Science Foundation, grant 16-16-00007.

ACKNOWLEDGMENTS

We thank Stephen Ridge for valuable discussions about expression data and Sergey Rukolaine for helpful advices on model inference.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00547/full#supplementary-material

REFERENCES

Andrés, F., and Coupland, G. (2012). The genetic basis of flowering responses to seasonal cues. *Nat. Rev. Genet.* 13, 627–639. doi: 10.1038/nrg3291

Blümel, M., Dally, N., and Jung, C. (2015). Flowering time regulation in crops—what did we learn from Arabidopsis? *Curr. Opin. Biotechnol.* 32, 121–129. doi: 10.1016/j.copbio.2014.11.023

Daba, K., Warkentin, T. D., Bueckert, R., Todd, C. D., and Tar'an, B. (2016). Determination of photoperiod-sensitive phase in chickpea (cicer arietinum L.). *Front. Plant Sci.* 7:478. doi: 10.3389/fpls.2016.00478

Dong, Z., Danilevskaya, O., Abadie, T., Messina, C., Coles, N., and Cooper, M. (2012). A Gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. *PLoS ONE* 7:e43450doi: 10.1371/journal.pone.0043450

Gaur, P. M., Samineni, S., Tripathi, S., Varshney, R. K., and Gowda, C. L. L. (2015). Allelic relationships of flowering time genes in chickpea. *Euphytica* 203, 295–308. doi: 10.1007/s10681-014-1261-7

Hecht, V., Laurie, R. E., Vander Schoor, J. K., Ridge, S., Knowles, C. L., Liew, L. C., et al. (2011). The pea GIGAS gene is a FLOWERING LOCUS T homolog necessary for graft-transmissible specification of flowering but not for responsiveness to photoperiod. *Plant Cell* 23, 147–161. doi: 10.1105/tpc.110.081042

Jaeger, K. E., Pullen, N., Lamzin, S., Morris, R. J., and Wigge, P. A. (2013). Interlocking feedback loops govern the dynamic behavior of the floral transition in Arabidopsis. *Plant Cell* 25, 820–833. doi: 10.1105/tpc.113.109355

Kardailsky, I., Shukla, V. K., Ahn, J. H., Dagenais, N., Christensen, S. K., Nguyen, J. T., et al. (1999). Activation tagging of the floral inducer FT. *Science* 286, 1962–1965

Kaufmann, K., Wellmer, F., Muiño, J. M., Ferrier, T., Wuest, S. E., Kumar, V., et al. (2010). Orchestration of floral initiation by APETALA1. *Science* 328, 85–89. doi: 10.1126/science.1185244

Kobayashi, Y., Kaya, H., Goto, K., Iwabuchi, M., and Araki, T. (1999). A pair of related genes with antagonistic roles in mediating flowering signals. *Science* 286, 1960–1962.

Kozlov, K. N., Samsonov, A. M., and Samsonova, M. G. (2016). A software for parameter optimization with differential evolution entirely parallel method. *PeerJ Comp Sci.* 2, e74–e20. doi: 10.7717/peerj-cs.74

Kumar, J., and Abbo, S. (2001). Genetics of flowering time in chickpea and its bearing on productivity in semiarid environments. *Adv. Agronomy* 72, 107–138. doi: 10.1016/S0065-2113(01)72012-3

Kumar, J., and van Rheenen, H. A. (2000). A major gene for time of flowering in chickpea. *J. Hered.* 91, 67–68. doi: 10.1093/jhered/91.1.67

Leal Valentim, F., Mourik, S. V., Pos,é, D., Kim, M. C., Schmid, M., van Ham, R. C., et al. (2015). A quantitative and dynamic model of the arabidopsis flowering time gene regulatory network. *PLoS ONE* 10:e0116973. doi: 10.1371/journal.pone.0116973

Liljegren, S. J., Gustafson-Brown, C., Pinyopich, A., Ditta, G. S., and Yanofsky, M. F. (1999). Interactions among APETALA1, LEAFY, and TERMINAL FLOWER1 specify meristem fate. *Plant Cell* 11, 1007–1018.

Mallikarjuna, B. P., Samineni, S., Thudi, M., Sajja, S. B., Khan, A. W., Patil, A., et al. (2017). Molecular mapping of flowering time major genes and QTLs in chickpea (Cicer arietinum L.). *Front. Plant Sci.* 8:1140. doi: 10.3389/fpls.2017.01140

Peng, F. Y., Hu, Z., and Yang, R. C. (2015). Genome-Wide comparative analysis of flowering-related genes in arabidopsis, wheat, and barley. *Int. J. Plant Genomics* 2015, 874361–874317. doi: 10.1155/2015/874361

Pin, P. A., and Nilsson, O. (2012). The multifaceted roles of Flowering Locus T in plant development. *Plant Cell Environ.* 35, 1742–1755. doi: 10.1111/j.1365-3040.2012.02558.x

Pullen, N., Jaeger, K. E., Wigge, P. A., and Morris, R. J. (2013). Simple network motifs can capture key characteristics of the floral transition in Arabidopsis. *Plant Signal. Behav.* 8:e26149. doi: 10.4161/psb.26149

Ratcliffe, O. J., Bradley, D. J., and Coen, E. S. (1999). Separation of shoot and floral identity in Arabidopsis. *Development* 126, 1109–1120.

Ridge, S., Deokar, A., Lee, R., Daba, K., Macknight, R. C., Weller, J. L., et al. (2017). The chickpea Early flowering 1 (Efl1) locus is an ortholog of arabidopsis ELF3. *Plant Physiol.* 175, 802–815. doi: 10.1104/pp.17.00082

Rohatgi, A. (2018). *WebPlotDigitizer [Internet]. Version 4.1. Austin, Texas (USA).* (Accessed Apr 11, 2018). Available online at: https://automeris.io/WebPlotDigitizer

Samee, M. A. H., Lim, B., Samper, N., Lu, H., Rushlow, C. A., Jiménez, G., et al. (2015). A systematic ensemble approach to thermodynamic modeling of gene expression from sequence data. *Cell Syst* 1, 396–407. doi: 10.1016/j.cels.2015.12.002

Shrestha, R., Gómez-Ariza, J., Brambilla, V., and Fornara, F. (2014). Molecular control of seasonal flowering in rice, arabidopsis and temperate cereals. *Ann. Bot.* 114, 1445–1458. doi: 10.1093/aob/mcu032

Srikanth, A., and Schmid, M. (2011). Regulation of flowering time: all roads lead to Rome. *Cell. Mol. Life Sci.* 68, 2013–2037. doi: 10.1007/s00018-011-0673-y

Upadhyaya, H. D., Bajaj, D., Das, S., Saxena, M. S., Badoni, S., Kumar, V., et al. (2015). A genome-scale integrated approach aids in genetic dissection of complex flowering time trait in chickpea. *Plant Mol. Biol.* 89, 403–420. doi: 10.1007/s11103-015-0377-z

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., et al. (2013). Draft genome sequence of chickpea (Cicer arietinum) provides a resource for trait improvement. *Nat. Biotechnol.* 31, 240–246. doi: 10.1038/nbt.2491

Wagner, D., Sablowski, R. W., and Meyerowitz, E. M. (1999). Transcriptional activation of apetala1 by leafy. *Science* 285, 582–584.

Warkentin, T., Banniza, S., and Vandenberg, A. (2005). CDC Frontier kabuli chickpea. *Can. J. Plant Sci.* 85, 909–910. doi: 10.4141/P04-185

Weller, J. L., and Ortega, R. (2015). Genetic control of flowering time in legumes. *Front. Plant Sci.* 6:207. doi: 10.3389/fpls.2015.00207

Zhang, X., Zhai, H., Wang, Y., Tian, X., Zhang, Y., Wu, H., et al. (2016). Functional conservation and diversification of the soybean maturity gene E1 and its homologs in legumes. *Sci. Rep.* 6:29548. doi: 10.1038/srep29548

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 21 November 2018
doi: 10.3389/fgene.2018.00569

Comparative Analysis of *Mycoplasma gallisepticum* vlhA Promoters

Mikhail Orlov[1], Irina Garanina[2]*, Gleb Y. Fisunov[2] and Anatoly Sorokin[1]*

[1] Institute of Cell Biophysics, Russian Academy of Sciences, Pushchino, Russia, [2] Federal Research and Clinical Center of Physical-Chemical Medicine, Federal Medical-Biological Agency, Moscow, Russia

Mycoplasma gallisepticum is an intracellular parasite affecting respiratory tract of poultry that belongs to class Mollicutes. *M. gallisepticum* features numerous variable lipoprotein hemagglutinin genes (vlhA) that play a role in immune escape. The vlhA promoters have a set of distinct properties in comparison to promoters of the other genes. The vlhA promoters carry a variable GAA repeats region at approximately 40 nts upstream of transcription start site. The promoters have been considered active only in the presence of exactly 12 GAA repeats. The mechanisms of vlhA expression regulation and GAA number variation are not described. Here we tried to understand these mechanisms using different computational methods. We conducted a comparative analysis among several *M. gallisepticum* strains. Nucleotide sequences analysis showed the presence of highly conserved regions flanking repeated trinucleotides that are not linked to GAA number variation. VlhA genes with 12 GAA repeats and their orthologs in 12 *M. gallisepticum* strains are more conserved than other vlhA genes and have narrower GAA number distribution. We conducted comparative analysis of physicochemical profiles of *M. gallisepticum* vlhA and sigma-70 promoters. Stress-induced duplex destabilization (SIDD) profiles showed that sigma-70 group is characterized by the common to prokaryotic promoters sharp maxima while vlhA promoters are hardly destabilized with the region between GAA repeats and transcription start site having zero opening probability. Electrostatic potential profiles of vlhA promoters indicate the presence of the distinct patterns that appear to govern initial stages of specific DNA-protein recognition. Open state dynamics profiles of vlhA demonstrate the pattern that might facilitate transcription bubble formation. Obtained data could be the basis for experimental identification of mechanisms of phase variation in *M. gallisepticum*.

Keywords: *Mycoplasma gallisepticum*, promoter, transcription regulation, DNA physics, vlhA

OPEN ACCESS

Edited by:
Yuriy L. Orlov,
Russian Academy of Sciences, Russia

Reviewed by:
Mikhail P. Ponomarenko,
Russian Academy of Sciences, Russia
Enrique Medina-Acosta,
Universidade Estadual do Norte Fluminense Darcy Ribeiro, Brazil

****Correspondence:***
Mikhail Orlov
orlovmikhailanat@gmail.com
Irina Garanina
irinagaranina24@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *08 August 2018*
Accepted: *06 November 2018*
Published: *21 November 2018*

Citation:
Orlov M, Garanina I, Fisunov GY and Sorokin A (2018) Comparative Analysis of Mycoplasma gallisepticum vlhA Promoters. Front. Genet. 9:569. doi: 10.3389/fgene.2018.00569

INTRODUCTION

Mycoplasmas are genome-reduced bacteria without a cell wall and with a parasitic lifestyle. Mycoplasmas parasitize diverse animal and plant species and humans. Like other intracellular parasites, they need to adapt to the host's immune system. One of main mechanisms Mycoplasmas employ is changing the repertoire of surface lipoproteins (phase variation) (Rosengarten and Wise, 1990). Other pathogenic bacteria, including *Haemophilus*, *Chlamydia*, and *Streptococcus* species,

also use phase variation to escape of host defense mechanisms (Noormohammadi, 2007). Phase variation in Mycoplasmas can occur spontaneously or due to an immune attack, it is important for persistence and survival of Mycoplasmas in a host (Markham et al., 1998; Glew et al., 2000; Ma et al., 2015; Czurda et al., 2017; Chopra-Dewasthaly et al., 2017). Numerous mechanisms of phase variation are described for Mycoplasmas (Citti et al., 2010). Usually, the mechanisms of variation are species-specific and occur in one species or closely related Mycoplasmas. They include DNA slippage, site-specific recombination, reciprocal recombination, and gene conversion (Citti et al., 2010). However, the phase variation system of *Mycoplasma gallisepticum* is unique, and has not been described so far. Therefore, studying phase variation genes can reveal novel mechanisms of gene expression regulation in bacteria.

Mycoplasma gallisepticum is a major bacterial pathogen inducing widespread respiratory disease in poultry and wild birds, which leads to significant economic losses throughout the world (Bencina, 2002). Phase variation of *M. gallisepticum* includes the switching on variable lipoprotein and hemagglutinin (vlhA) gene expression (Markham et al., 1992). The exact function of vlhA proteins is still unknown. They involve in haemagglutination (Bencina, 2002; Noormohammadi, 2007), based on data obtained on avian Mycoplasmas it can be assumed that vlhA proteins participate in host cell adhesion and invasion (May et al., 2014; Matyushkina et al., 2016; Hegde et al., 2018). VlhA genes are organized into 3–5 cassettes, uniting ten genes per cassette (Baseggio et al., 1996). The promoter structure of these genes is significantly different from the promoters of the other *M. gallisepticum* genes. VlhA genes lack conserved sigma-70 promoter sequence and often have GTG start codon (Markham et al., 1994). They are proposed to employ an alternative sigma factor binding GCGAAAAT sequence (Fisunov et al., 2016). Long regions of GAA repeats are located upstream of vlhA genes (Markham et al., 1994). In general, the GAA repeats can be considered as short-sequence repeats (SSRs). SSRs were found in all eukaryotic and many prokaryotic genomes (Mrázek et al., 2007; Avvaru et al., 2017). In bacteria, SSRs were identified in genes coding for bacterial virulence factors including lipopolysaccharide-modifying enzymes or adhesins (Mrázek, 2006; Wei et al., 2015). So, SSRs provide genetic and, therefore, phenotypic variability. Changes in number of repeated units and/or in the repeat unit itself may arise from recombination processes or polymerase errors including slipped-strand mispairing (SSM), either solely or in combination with DNA repair deficiencies (van Belkum et al., 1998; Rocha, 2003; Torres-Cruz and van der Woude, 2003).

First experiments showed that *M. gallisepticum* express only one vlhA family member at a time and expression depends on the presence of exactly 12 GAA trinucleotide repeats upstream of the gene (Glew et al., 1995, 1998; Liu et al., 2002). Recently it was shown that expression of the gene preceded by 12 GAA exceeds the other vlhA genes, but the other genes with a different number of repeats are also expressed and some of them are expressed at a high level (Matyushkina et al., 2016; Pflaum et al., 2016; Butenko et al., 2017). *In vivo* experiments showed the non-stochastic character of vlhA switching during infection, vlhA expression pattern changes during infection progression and differs between strains (Pflaum et al., 2016, 2018). So, vlhA expression is determined by GAA repeats, but probably the additional expression control mechanisms exist. An interesting question here is how the cell defines what promoter needs to be activated. One explanation here is the existence of hemagglutinin activator protein (HAP) recognizing 12-GAA repeats (Liu et al., 2002).

Another question is the mechanism of GAA repeat variation in *M. gallisepticum*. It would be interesting to find out how many repeats changes at a time, whether the change depends on the number of repeats of a given gene, or on the sequences surrounding the GAA repeats and their physicochemical properties. In the present study we used computational methods to analyze genomes of several *M. gallisepticum* strains and shed light to the mechanism of phase variation and vlhA expression control. For this purpose, we used comparative bioinformatics analysis of sequences of vlhA promoters and genes. We assumed that a nonstandard structure of vlhA promoters may be related to the physicochemical properties of their sequences, using computational methods we predicted these properties on the DNA of vlhA promoters and compared them with the corresponding properties of experimentally obtained sigma-70 promoters of *M. gallisepticum S6*.

MATERIALS AND METHODS

Bioinformatics Analysis of (GAA)n and vlhA Genes

We used 12 complete genomes of *M. gallisepticum* strains isolated from chickens and house finches of various levels of virulence available for download in June 2018 in the GenBank database (Papazisi et al., 2003; Szczepanek et al., 2010; Fisunov et al., 2011; Tulman et al., 2012; Fleming-Davies et al., 2018). List of the genomes and their characteristics (size, GC content, and number of genes) are provided in **Supplementary Table S1**. We obtained sequences of vlhA promoters of all 12 strains to study GAA number variation. For comparison of physicochemical properties, we retrieved sequences of sigma-70 promoters of S6 strain. The exact coordinates of the transcription start sites of *M. gallisepticum* S6 were obtained from our published work there 5′-end enriched RNA-seq sequencing was conducted (Mazin et al., 2014).

The GAA repeats were defined as 4–27 non-interspaced trinucleotides repeated in a row. A smaller number of the repeats appeared to be non-specific; no 28 or more repeats were detected. We proposed that for the possible GAA recognizing protein the length of GAA tract should be more important than the substitutions in one repeat inside the (GAA)n. So, we considered units with substitutions inside the (GAA)n as intact units and shortened the (GAA)n to the units with at least one substitution if it was at the end of the (GAA)n. We did not detect GAA tracts containing more than two damaged GAA inside the tract. For sequencing retrieval and GAA counting we used Python 2.7 custom script.

To analyze GAA number variation we classified vlhA genes into orthologous groups. Not all vlhA have clear annotation, most are annotated as hypothetical proteins. Since we are interested only in vlhA under the control of (GAA)n containing promoters, to find all vlhA genes we first mapped GAA repeats and then found corresponding vlhA genes. Several times we observed short GAA repeat in coding regions of other genes or GAA that not connected with vlhA, this cases we corrected manually. ProteinOrtho program (version V5.16) was used to computing orthologous vlhA proteins (Lechner et al., 2011). Parameters identity =70% and minimum coverage of best blast alignments =50% were used. Fisher exact test was performed using fisher.test() function in R with two.sided alternative hypothesis.

To reconstruct the phylogenetic tree of vlhA genes for **Figure 3** we obtained consensus sequences of orthologous clusters applying Biopython command dumb_consensus() to orthologous group alignments (Cock et al., 2009). VlhA proteins and their consensus sequences we aligned by T-coffee program implemented in JalView software (version 2.10.5) with default parameters (Waterhouse et al., 2009; Di Tommaso et al., 2011). Phylogenetic tree of consensus sequences was constructed by Phylogeny.fr tool where the method of maximum-likelihood is implemented (Dereeper et al., 2008). The histogram of GAA number and distributions were constructed in R.

Analysis of (GAA)n Flanks

For analysis of (GAA)n flanking regions, we extracted 50 nucleotide sequences upstream and downstream of the (GAA)n. We aligned upstream and downstream flanks independently by T-Coffee program implemented in JalView software (version 2.10.5) with default parameters (Waterhouse et al., 2009; Di Tommaso et al., 2011) and merged corresponding aligned flanks using Biopython Python 2.7 library (Cock et al., 2009). See flanks alignment in **Supplementary Materials**. WebLogo was used for sequence logos construction (Crooks et al., 2004).

To compare (GAA)n flanking sequences between 12-GAA and the other vlhA genes we used a non-linear algorithm of dimension reduction t-SNE (t-Distributed Stochastic Neighbor Embedding). t-SNE allows a visualization a high-dimensional data to see high-dimensional objects in two- or three-dimensional space. t-SNE visualizes the data in compact and clear view and has advantages over other dimension reduction methods, like PCA (van der Maaten and Hinton, 2008). Alignment was transformed into the table presenting nucleotides and gaps with numbers, columns correspond to positions in alignment, rows to individual genes. We employed PCA algorithm with default parameters and t-SNE algorithm with perplexity parameter 30 implemented in sklearn Python 2.7 library (Pedregosa et al., 2011).

Calculation of Physicochemical Properties of Promoters

Stress-induced duplex destabilization (SIDD) is a theoretical method developed to analyze denaturation in superhelical DNA of a specified sequence (Benham, 1990). SIDD profile analysis predicts the DNA positions where the DNA duplex becomes susceptible to separation when under superhelical stress (Benham, 1990). SIDD calculation was carried out as implemented by its authors (Zhabinskaya et al., 2015). The conformational and thermodynamic parameters were derived from the endonuclease digestion experiments on superhelical DNA (Kowalski et al., 1988; Benham, 1992). Theoretical calculations using these parameters were consistent with experimental data (Benham, 1992).

For SIDD calculations 1000 nts-long intervals with transcription start site (TSS) at the center were considered, usage long DNA regions take into account broader genomic context. We filtered nucleotide sequences containing more than one promoter. SIDD profiles were obtained by means of perl script. SIDD calculation was performed using default settings (superhelicity level 0.06, energy threshold 12, and ionic strength 0.01). Temperature value was equal to the average chicken body temperature (314 K). The difference between SIDD profile maximum values was tested by the non-parametric Mann–Whitney U test implemented in R using wilcox.test() function with parameter paired =FALSE.

Distribution of electrostatic potential is DNA duplex feature that contributes to the initial stages of DNA–protein interactions (Jones et al., 2003). The DNA characteristic profiles were obtained using method suitable for genome-wide application (Polozov et al., 1999). The approach is based on Coulomb formula and allows to analyze electrostatic profiles of promoters within the electrostatic map of a whole genome DNA. It is widely used in studies concerning electrostatic patterns of bacterial and phage promoters (Polozov et al., 1999; Kamzolova et al., 2005, 2006, 2009; Sorokin et al., 2006; Osypov et al., 2010). Finally, DNA open states dynamical properties, including their activation energy (E0) and size (d). These are believed to affect transcription bubble formation and introduce additional to the encoded by steady-state DNA properties information. The used model equation was derived from the sine-Gordon equation by adding two additional terms which more accurately take into account heterogeneous nature of the DNA sequence. The profiles were shown to be in agreement with the function of the corresponding DNA regions: promoters are evolving open states with most ease, while terminator are likely to stop the transcription bubble (Grinevich et al., 2015). Therefore, SIDD profiles were obtained by means of perl script, electrostatic profiles was calculated using the algorithm implemented in R, and the dynamical properties of DNA open states were obtained using the algorithm implemented in Matlab 9.2.

RESULTS

VlhA Promoters Share Conserved GAA-Flanking Sequences Irrespective of GAA Units Number

Comparative analysis of GAA repeats number for vlhA genes of different strains was conducted to identify possible patterns of variation. All vlhA genes from 12 strains were clustered

into orthologous groups according to the sequence similarity. Previous studies revealed that activation of vlhA transcription occurs if 12 GAA repeats are present within the promoter. Flanking regions of the GAA repeats were also found to be essential for vlhA expression (Liu et al., 2000). Here we analyzed conservation of GAA flanks among different *M. gallisepticum* strains and vlhA orthologous groups to identify the mechanism of vlhA expression activation. For each vlhA gene sequences upstream and downstream of (GAA)n were obtained. Totally 368 promoters were taken into analysis. GAA tracts were defined as repeat regions containing 4 or more GAA trinucleotides without substitutions at the ends of the (GAA)n. The logos build demonstrate conserved sequences both upstream and downstream of GAA repeats (**Figure 1**). The conservation level varies among positions of the motifs. We searched for similar sequences in nucleotide collection at NCBI blast by blastn program and did not find any matches in other species. So, these sequences show no sequence homology with sequenced genomes and appear to be identified in *M. gallisepticum* genome only. The sequences comprise neither repetitive sequences nor palindromes that often are present in regulatory motifs.

We compared flanking sequences of 12-GAA tracts with other vlhA promoters. First, we looked over logos of 12-GAA and non-12-GAA flanks (**Figure 1**). No traceable distinction was found between the two groups. To more precise comparison we visualized sequences in three-dimensional space using t-SNE method (**Figure 2**). This method shows sequences similarity as a distance in two- or three-dimensional space. No clustering of promoters with 12 GAA was identified by t-SNE and by similar method PCA (**Supplementary Figure S1**). So, analysis of GAA flanking regions revealed conserved positions around GAA tract and did not show correlations between 12-GAA units in (GAA)n and sequence of (GAA)n flanks.

To consider in more detail the flanking sequences, we constructed their alignments and phylogenetic trees for genes belonging to the same orthologous groups. In the article we describe two representative examples of trees (**Figure 3**) and the alignments of flanks of orthologous groups (**Supplementary Materials**). The identity level between vlhA proteins of these two orthologous groups is higher than 90% for all protein pairs. The first tree represents the tree of the merged flanks of (GAA)n for the orthologous cluster containing 4 genes with 12-GAA repeats. This is the largest orthologous group, containing proteins represented in all strains. The alignment and tree show that the sequences are conservative within the groups of strains isolated from different species: strains F, S6, Rlow, and Rhigh were obtained from chickens, the remaining strains from house finches. Genomes of finch strains have almost identical genome sequences with a low number of substitutions, but the difference exists (Tulman et al., 2012; Kristensen et al., 2017). Chicken strains are less similar to each other than strains from finches according to data from the ATGC database (Tulman et al., 2012). That is, in this case, one would expect slight differences between the (GAA)n flanks of individual strains, but the sequences for the orthologous group are completely identical within two groups. It is interesting that the flanks and the corresponding genes are located in different vlhA cassettes, the genes from chicken strains are located in the first cassette, and finch genes are located in the third and fourth cassettes. So, the moving to other cassette did not affect sequences of (GAA)n flanks. The orthologous group includes 4 genes with 12-GAA repeats, no differences between them and other genes are noticeable. We observed that the number of repeats within the orthologs cluster varied, while sequences of repeats were conservative. This suggests that the change in the number of GAA repeats does not depend on the sequences flanking them. **Figure 3B** shows the tree of another orthologous group, which also contains 12-GAA repeat genes. The tree confirms the lack of connection between the number of repeats and the sequence of flanks. These flanking sequences are less conservative among themselves than sequences of the first group. Thus, analysis of trees and alignments of particular orthologous groups showed no connections between (GAA)n number and their flanking sequences.

Number of GAA Repeats Varies Among Orthologs vlhA and Different Strains of *M. gallisepticum*

Comparative analysis of GAA repeats a number of vlhA genes from different strains were conducted to identify possible patterns of GAA number variation. All vlhA genes from 12 strains were clustered into orthologous groups (**Figure 4A**). The distribution of GAA tract lengths shows that the majority of values reside within a narrow range of 6-12 repeats. We divided vlhA orthologous clusters into two groups: the one containing 12 repeats at least in one strain and the one including the rest. The distribution within 12-GAA containing group is even narrower varying from 8 to 12 repeats. This may indicate that GAA number changes by an increase/decrease of a small number of repeats.

The number of 12-GAA promoters varies across the strains from zero to three per genome. We found the positive correlation between gene conservation level and the presence of 12-GAA repeats within an ortholog cluster. Genes with 12 repeats are more frequently occur in full ortholog clusters comprising to genes that are represented in all strains (Fisher exact test p-value =0.0248).

The number of repeats varies within one genome as well as within one orthologous cluster. We analyzed the distribution of GAA repeats number among the strains and orthologs clusters (**Figures 4B,C**). The data shows that the prevalent GAA repeats number is 8 and frequency decreases as the number of repeats increases. Genes with 12 GAA repeats follow the common trend and have no exceptional frequencies. Comparison of dispersion in repeats number among the strains and ortholog clusters showed that the number of repeats is more conserved within one strain than within one ortholog cluster. The majority of the strains tend to follow this trend, except for S6 strain which exhibited the most versatile repeat number. Certain ortholog clusters are more conserved than others which may indicate differences in VlhA expression among strains. Therefore, analysis of GAA repeats number did not reveal any traceable patterns in the distribution of repeats. We suggest that alike patterns might be established after considering a bigger set of strains.

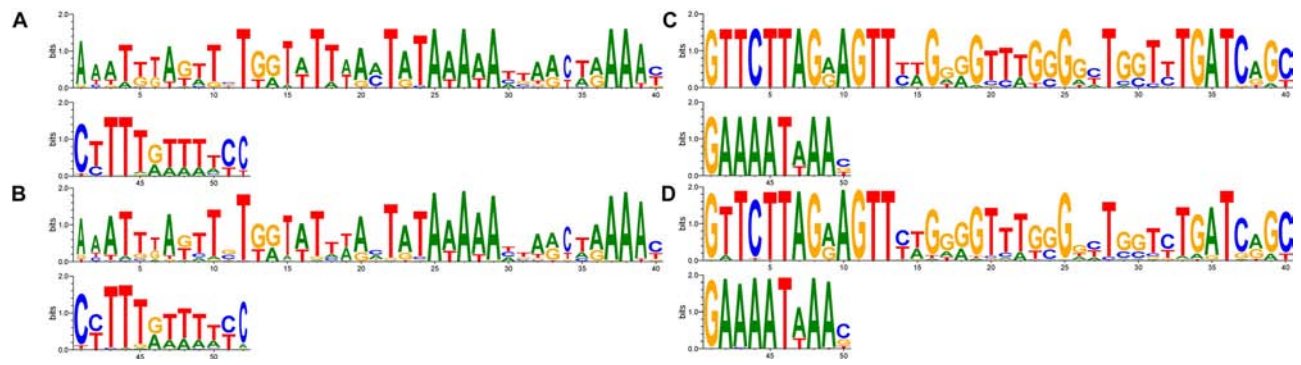


FIGURE 1 | The motif of (GAA)n flanking sequences in vlhA promoters. Logos show identity of motifs for promoters with different GAA number. Sequences 50 bp length were aligned by T-coffee program, gaps included in the alignment. Logo constructed by WebLogo 3.6.0; **(A,B)** logos for upstream flanks, **(C,D)** logos for downstream flanks; **(A,C)** logos of 22 sequences of 12-GAA promoters; **(B,D)** logos of 344 non-12-GAA promoters.

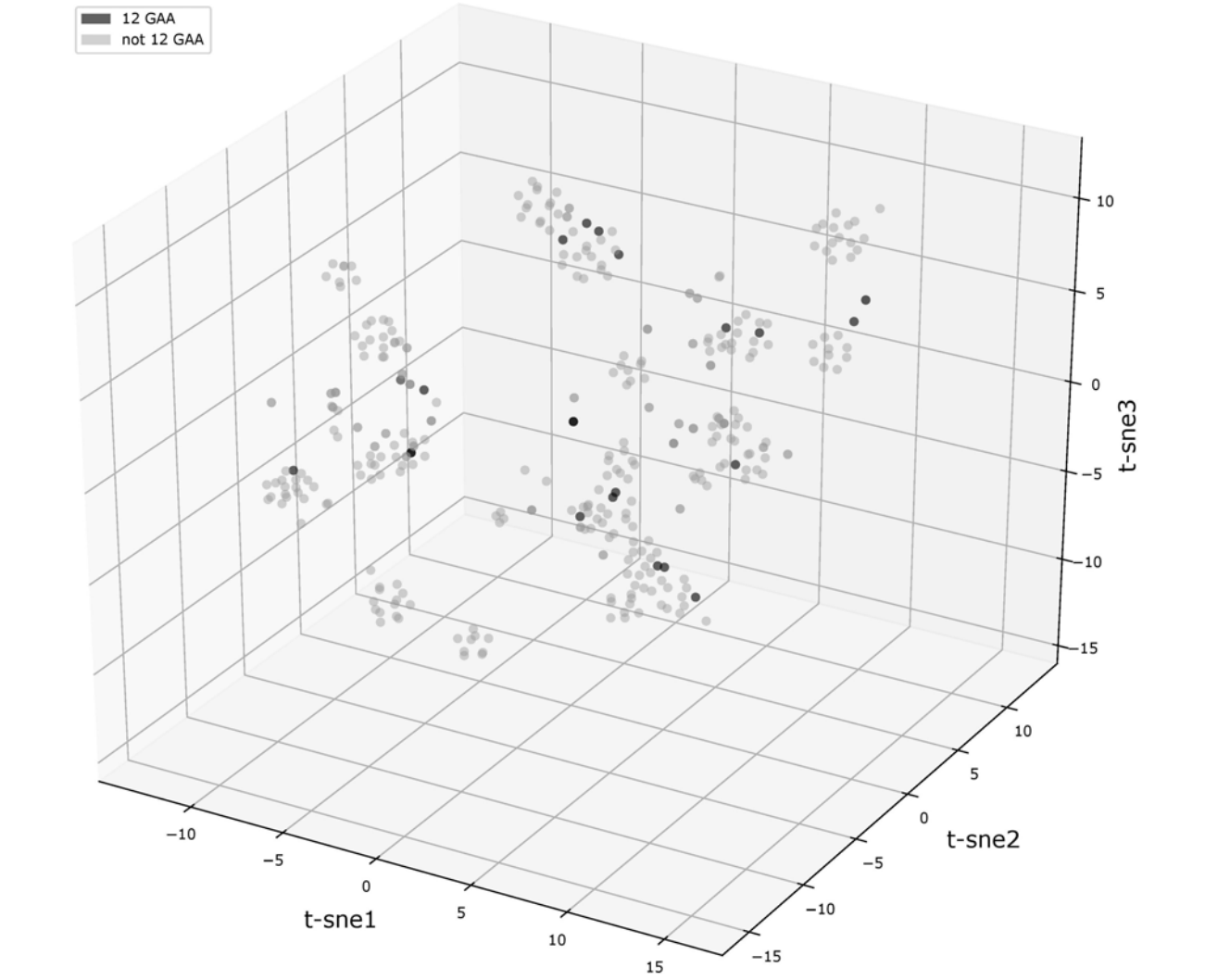


FIGURE 2 | t-SNE analysis plot of (GAA)n flanking motifs. Points represent individual vlhA genes of all analyzed strains, the analysis made on concatenated left and right (GAA)n flanking sequences. Black points show 12-GAA promoters. In analysis was used t-SNE algorithm implemented in sklearn Python library with the parameter of perplexity =30.

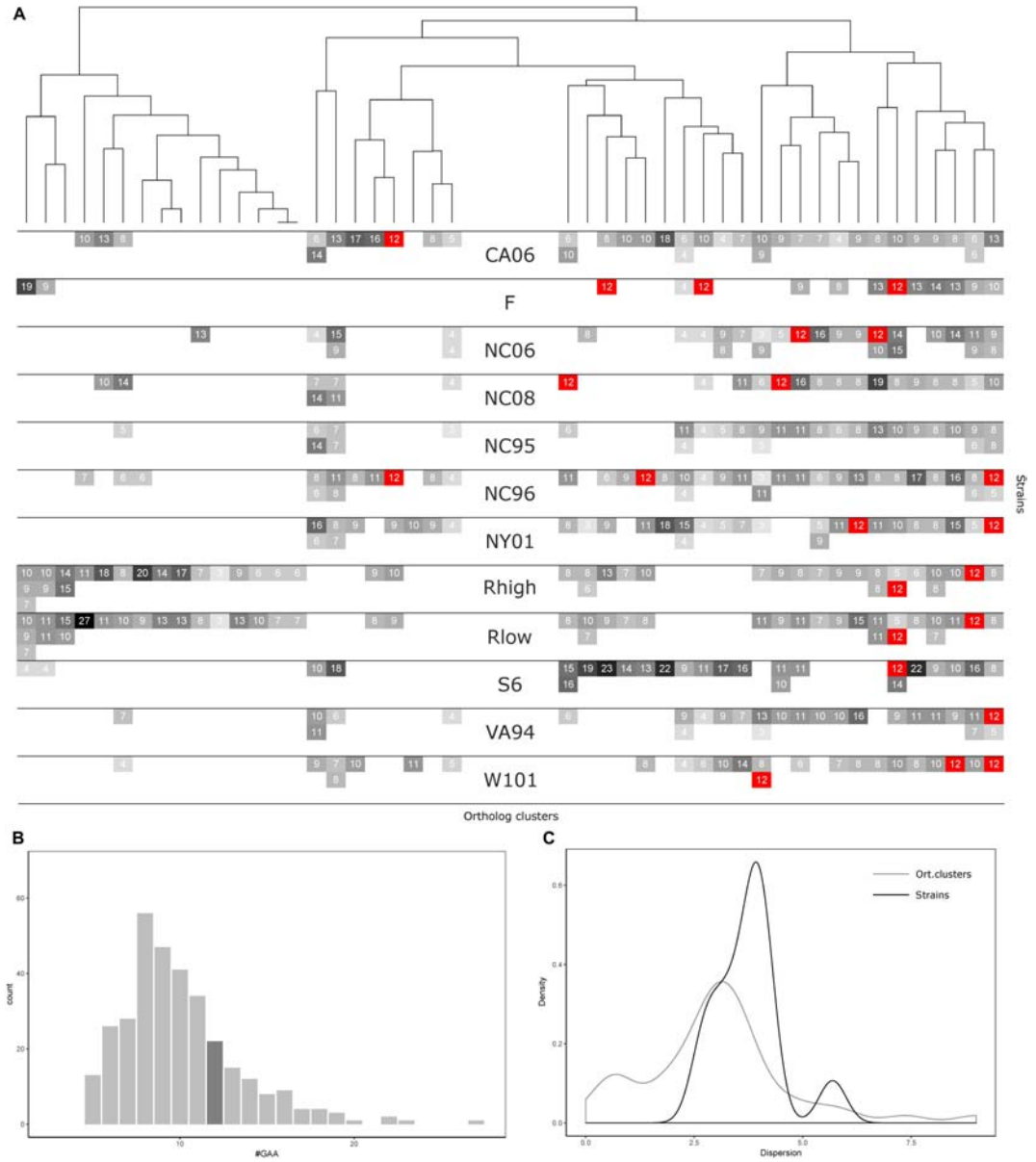


FIGURE 3 | GAA repeats number statistics for 12 *Mycoplasma gallisepticum* strains and vlhA orthologous clusters. **(A)** Heatmap showing number of GAA repeats for each vlhA promoter. The number of repeats is indicated by colors, 12-GAA repeats are shown with red. One strain corresponds to three rows (three is the maximum numbers of vlhA paralogs observed for a strain). Names of the strains are shown in the heatmap center. Orthologous clusters correspond to columns. The tree was constructed by Phylogeny.fr software based on T-coffee protein alignment of consensus sequences of orthologous groups using the maximum-likelihood method for phylogeny reconstruction. **(B)** Histogram of the number of GAA repeats. The dark gray bar shows 12-GAA promoters. **(C)** Distribution of dispersion of GAA repeats number among strains and orthologous clusters.

VlhA Promoters Have Lowest Opening Probability Under Superhelical Stress (SIDD Profiles) While Non-vlhA Promoters Are Highly Destabilized

In order to describe the possible role of physicochemical interactions in phase variation of *M. gallisepticum* several DNA properties of promoter regions were obtained in the form of profiles. SIDD as a DNA parameter shows a robust correlation with various regulatory DNA loci including promoters, replication origins, etc. The promoters of E. coli can be classified into SIDD-dependent and SIDD-independent groups according to their SIDD profile, which seems to correlate with their functional specialization (Wang and Benham, 2006). In the present article we analyzed SIDD profiles for vlhA promoters from various *M. gallisepticum* strains as well as, for standard sigma-70 promoters experimentally identified in S6 strain (Mazin et al., 2014). Promoters of both type

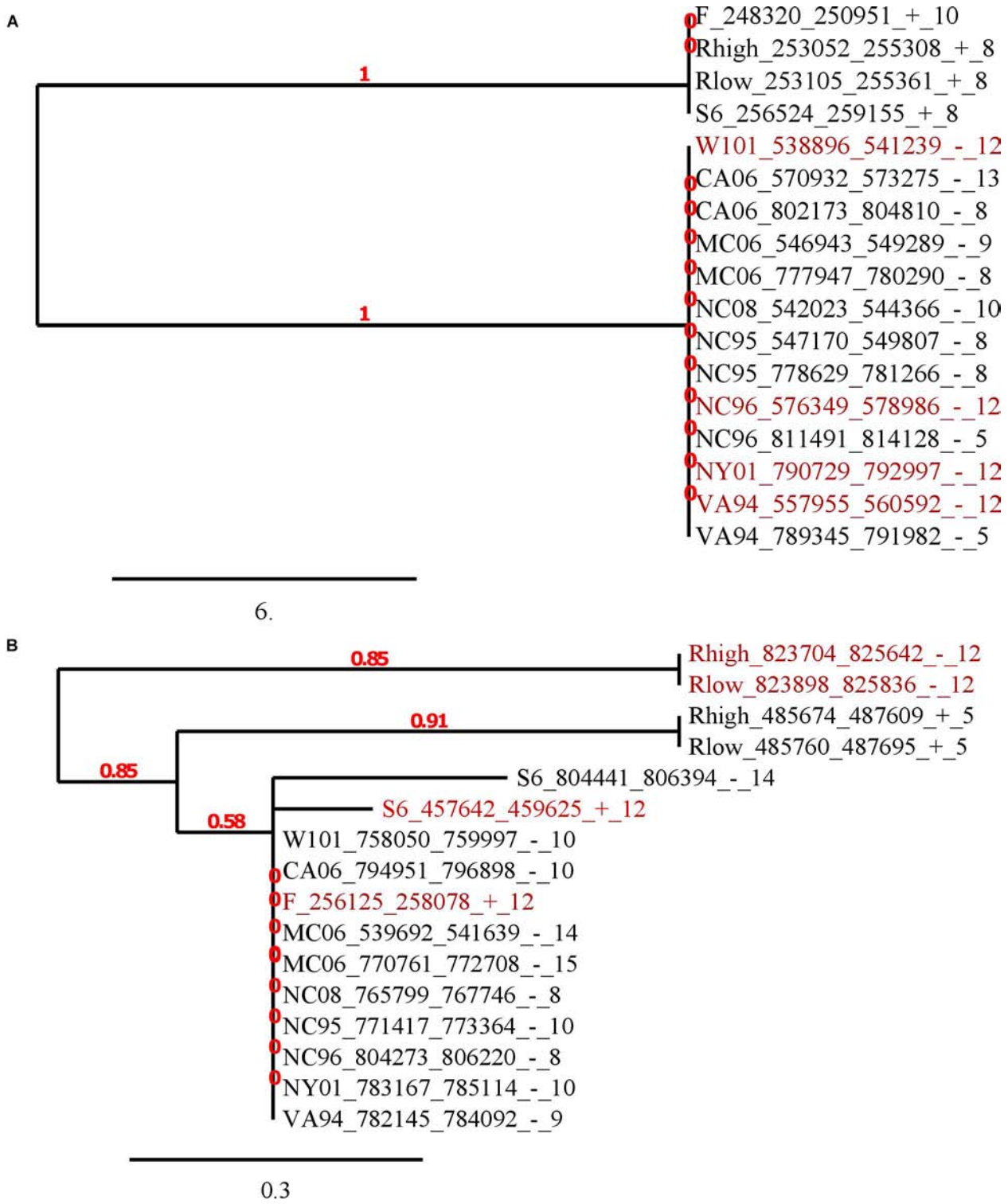


FIGURE 4 | Phylogenetic trees of (GAA)n flanking sequences. The trees were constructed by Phylogeny.fr software based on T-coffee nucleotide alignments of 50 bp flanks. The maximum-likelihood method was applied for phylogeny reconstruction. Labels of sequences consist of strain short name, corresponding vlhA gene position (start, end, strand separated by '_') and GAA units number separated by '_'. Red color shows 12-GAA promoters. The scale bar shows 0.02 changes. Red numbers on branches display branch support values. **(A)** The tree of the biggest orthologous group that is depicted in the last column in **Figure 3A**; **(B)** the tree of another orthologous group, consisting of four 12-GAA genes. The group is depicted in 41 column in **Figure 3A**.

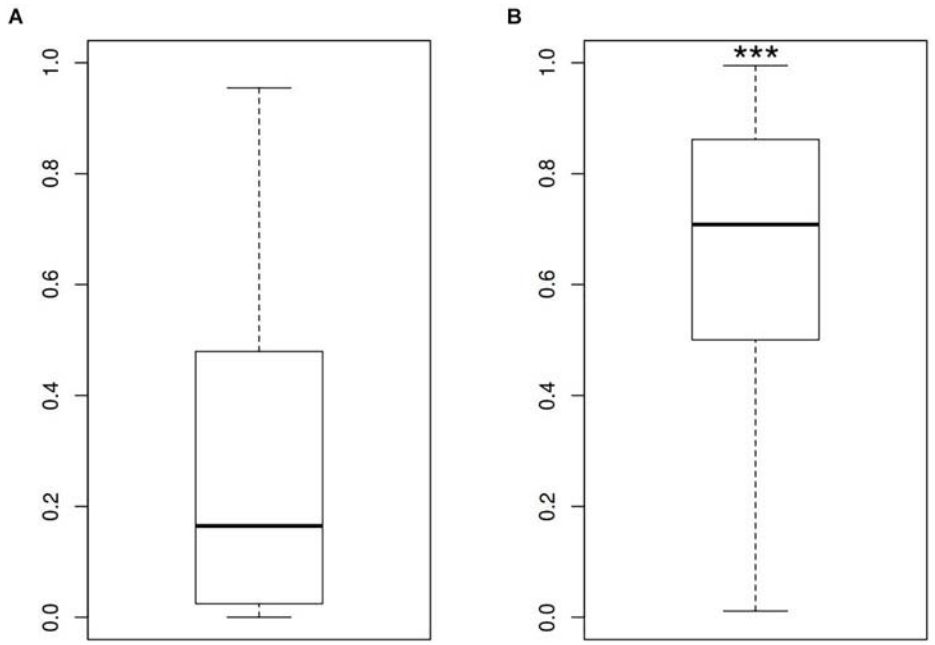


FIGURE 5 | SIDD profile maximum values for **(A)** vlhA promoters of 12 *M. gallisepticum* strains and **(B)** sigma 70 promoters of S6 strain. ***Mann–Whitney test p-value <0.05.

feature same GC-content of 0.3, which is the average GC-content of *M. gallisepticum* genome. Sigma-70 promoters are substantially more destabilized with the profile maxima located in the vicinity of TSS, while vlhA promoters did not incline to melt under the considered conditions (**Figure 5**). Peaks of vlhA promoters' profiles do not overlap TSS region with the sequence adjacent to GAA repeats having zero melting probability. At the same time, the majority of sigma-70 promoters demonstrate sharp maxima in the upstream region [−100; −50] nts (Mann-Whitney test p-value <0.05) (**Figure 6**). The fact to some extent supports the notion that there is no direct correlation between SIDD profiles and GC-content of a DNA segment.

Dynamical Properties of DNA Open States and Electrostatic Potential Profiles of vlhA Promoters Show Distinct Patterns

Dynamics of DNA open states was shown to be important for transcription bubble formation (Grinevich et al., 2015). The lower the open states activation energy, the more the DNA duplex is prone to open thus facilitating transcription initiation. Open states activation energy profiles, as well as the size of open states profiles, were calculated for vlhA and sigma-70 promoters. We identified that the transition of vlhA promoters to an open state occurred more efficiently in the region downstream TSS. The activation energy for the promoter group in the interval [−70; 20] nts appeared to have a decreasing slope which starts at the right GAA repeats boundary. It may seem tempting to suggest that the slope facilitate the directed movement of RNA-polymerase along the promoter. At that, no traceable patterns were detected for sigma-70 promoters (**Figures 7**,**8**).

Distribution of electrostatic potential (EP) around DNA duplex is a physical property that could be recognized by other molecules at a distance and prior to their direct interaction. It appears to be crucial at the initial stages of promoter recognition by RNA-polymerase (Polozov et al., 1999). Promoters of vlhA genes show characteristic EP pattern with the peak at about 30 nt after TSS. Neither visual assessment nor clusterization revealed traceable patterns for sigma-70 promoters profiles (**Figure 9**).

DISCUSSION

The promoters of vlhA genes feature a remarkable mechanism of transcriptional regulation. It includes two functional components: transcriptional activation at 12-GAA containing promoters and variation of GAA repeats number. In the article we have analyzed conservation, GAA number distribution, and physicochemical properties of vlhA promoters in *M. gallisepticum*. We proposed that physicochemical properties of promoters including SIDD, DNA open states dynamics, and electrostatic potential could be connected to the vlhA genes expression regulation.

We demonstrated that the GAA repeats in vlhA promoters are flanked by highly conserved sequences with distinct structure. Altogether the regulatory region takes more than 50 nt. Sequences of such length are generally too large for binding a typical bacterial transcription factor (Rodionov, 2007). Regulatory sequences of this length are unique in bacteria.

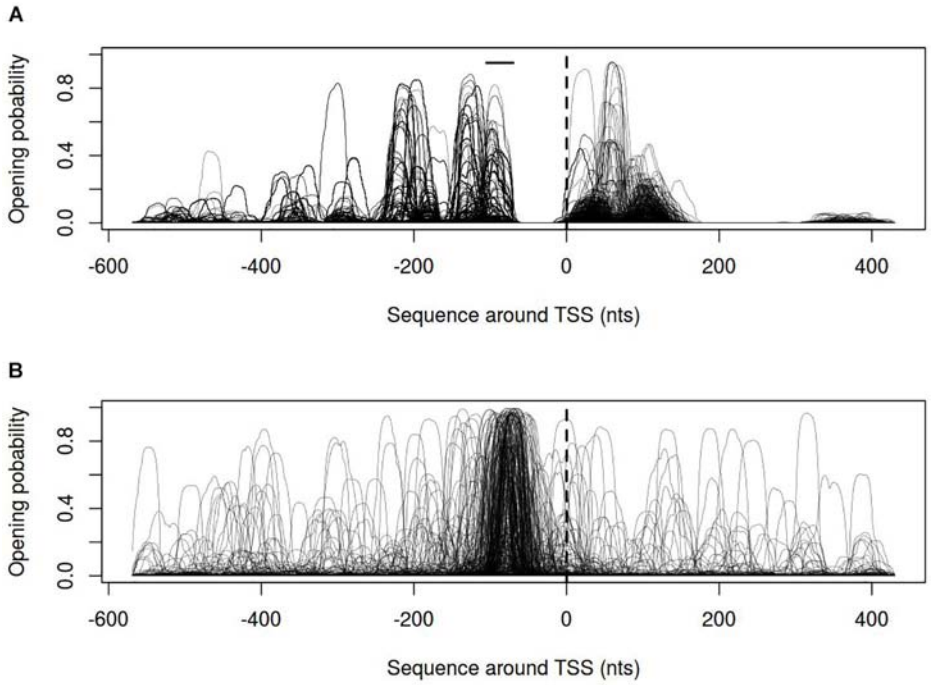


FIGURE 6 | Opening probability (SIDD) profiles for **(A)** vlhA promoters of 12 *M. gallisepticum* strains; **(B)** sigma 70 promoters of S6 strain. Dashed line denotes transcription start site; solid horizontal line – approximate GAA repeats location.

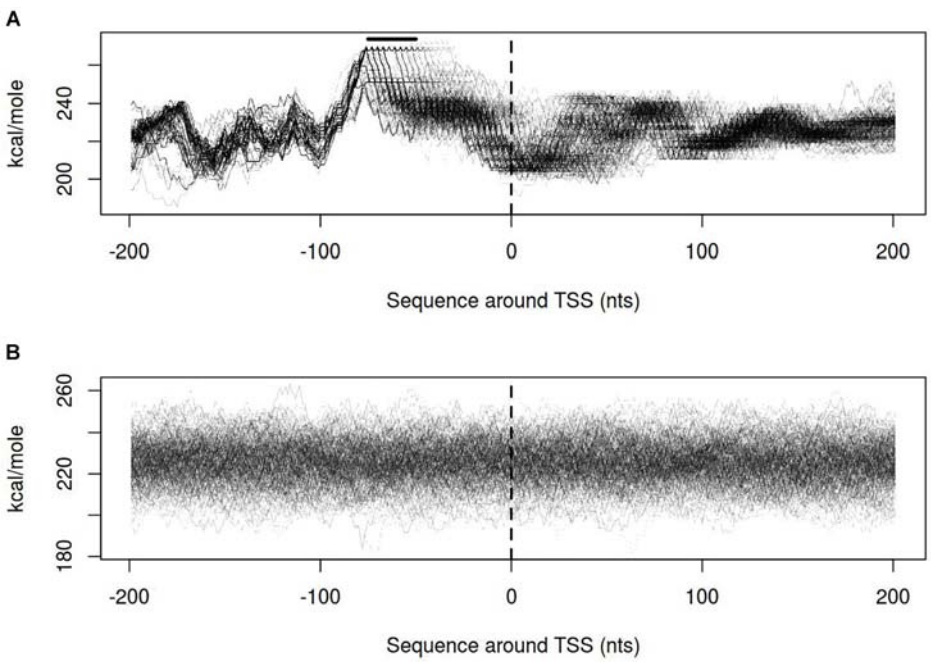


FIGURE 7 | Open states activation energy profiles for **(A)** vlhA promoters of 12 *M. gallisepticum* strains; **(B)** sigma 70 promoters of S6 strain. Dashed line denotes transcription start site; solid horizontal line – approximate GAA repeats location.

It is possible that *M. gallisepticum* has unique DNA binding proteins with the unknown spatial structure of the DNA binding region that standard annotation programs cannot identify. The hypothesis is supported by the fact that Mycoplasmas have a large number of orphan genes with unknown functions (Tatarinova et al., 2016).

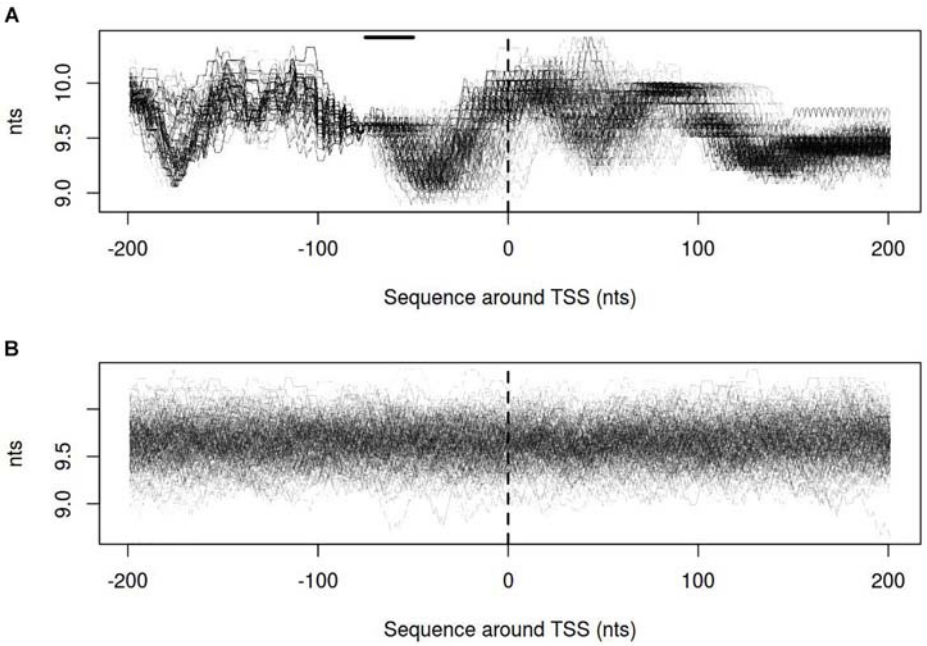


FIGURE 8 | Size of open states profiles for **(A)** vlhA promoters of 12 *M. gallisepticum* strains; **(B)** sigma 70 promoters of S6 strain. Dashed line denotes transcription start site; solid horizontal line – approximate GAA repeats location.

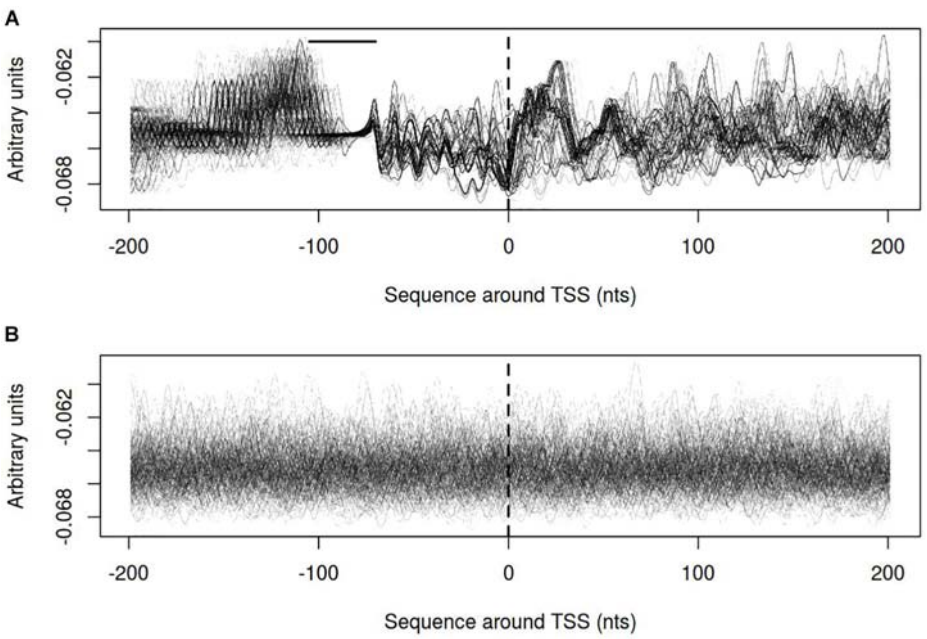


FIGURE 9 | Electrostatic potential (EP) profiles for **(A)** vlhA promoters of 12 *M. gallisepticum* strains; **(B)** sigma 70 promoters of S6 strain. Dashed line denotes transcription start site; solid horizontal line – approximate GAA repeats location.

Most of the analyzed strains are isolated from wild birds and are pathogenic for the host. We observed 12-GAA vlhA genes occur more than one time in the genome. Obtained data implies that the presence of a single 12-GAA vlhA gene is not the only possible combination enabling pathogenicity manifestation. Closely related strains Rlow and Rhigh demonstrate similar distributions of 12-GAA genes but have distinct virulence potential (Szczepanek et al., 2010). Vaccine strains F with a low level of pathogenicity have the maximum number of genes with 12-GAA repeats and lacks numerous vlhA genes. One can

speculate the inability of proper vlhA switching may result in a decrease of pathogenicity.

We identified that the distribution of GAA number resides within narrow borders of 8–12 repeats only in case orthologous clusters with at least one 12-GAA promoter were considered. We hypothesize that there is a "working range" of GAA repeats within which the number can iterate while having a considerable chance to get back to 12. Promoters that occasionally go out of range are not functional, while they still may remain conserved. The corresponding genes will never be activated again. The orthologous clusters lacking 12-GAA promoters are distributed in considerably fewer strains which corroborates with the idea that they lost function and represent a decaying group of vlhA.

Calculation of physical properties of vlhA promoters and sigma-70 promoters of S6 strain allowed to identify distinct patterns in open states dynamics and electrostatic potential profiles. We hypothesize that the former could facilitate transcription bubble formation thus stimulating processive transcription, while the latter could contribute to the initial stage of DNA-protein recognition. By contrast, SIDD profiles of vlhA promoters are hardly destabilized and have zero opening probability near TSS while sigma-70 promoters have overall high destabilization levels with maxima associated with TSS position. It corroborates with the idea that an alternative sigma-factor rather than sigma-70 is utilized for transcription of vlhA. One can speculate that zero open probability of vlhA promoters under superhelical stress reflects that fact that these loci are wrapped around activator complex, e.g., are at a high local degree of negative supercoiling. At the same time, improper transcription should not be facilitated from vlhA promoters since their −10 boxes show a substantial degree of similarity with those of sigma-70.

CONCLUSION

Analysis of promoters of vlhA indicates the presence of conserved sequences upstream and downstream to GAA repeats. Sequences of (GAA)n flanks are not connected with the number of GAA repeats. The distribution of (GAA)n length among the strains of *M. gallisepticum* shows a preferred range within which this number iterates: 6–12 repeats. Distribution of GAA units number varies among strains and orthologous groups. VlhA orthologous groups having at least one 12-GAA gene in the group have a narrower distribution of GAA number with values within the range 8–12 and are more conserved among strains than other orthologous groups. As compared to sigma-70 promoters of *M. gallisepticum* promoters of vlhA feature distinct and characteristic profiles of physical properties including opening probability under superhelical stress, open state activation energy, and electrostatic potential.

DATA AVAILABILITY

The datasets analyzed and scripts for this study can be found in the https://github.com/FVortex/Orlov_et_al._Frontiers_in_Genetics_Mycoplasma_gallispeticum_script.

AUTHOR CONTRIBUTIONS

IG contributed in analysis of genomes, GAA repeats, cauterization, and writing the manuscript. MO contributed in analysis of physicochemical properties and writing the manuscript. GF and AS wrote the manuscript.

FUNDING

This work was funded by the Russian Science Foundation grant 14-24-00159 "Systems research of minimal cell on a Mycoplasma gallisepticum model".

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00569/full#supplementary-material

TABLE S1 | Description of strains used in the study.

TABLE S2 | Data on vlhA genes, sequences of genes, their promoters and GAA repeats.

REFERENCES

Avvaru, A. K., Saxena, S., Sowpati, D. T., and Mishra, R. K. (2017). MSDB: a comprehensive database of simple sequence repeats. *Genome Biol. Evol.* 9, 1797–1802. doi: 10.1093/gbe/evx132

Baseggio, N., Glew, M. D., Markham, P. F., Whithear, K. G., and Browning, G. F. (1996). Size and genomic location of the pMGA multigene family of *Mycoplasma gallisepticum*. *Microbiology* 142(Pt 6), 1429–1435. doi: 10.1099/13500872-142-6-1429

Bencina, D. (2002). Haemagglutinins of pathogenic avian mycoplasmas. *Avian Pathol. J.* 31, 535–547. doi: 10.1080/0307945021000024526

Benham, C. J. (1990). Theoretical analysis of heteropolymeric transitions in superhelical DNA molecules of specified sequence. *J. Chem. Phys.* 92, 6294–6305. doi: 10.1063/1.458353

Benham, C. J. (1992). Energetics of the strand separation transition in superhelical DNA. *J. Mol. Biol.* 225, 835–847. doi: 10.1016/0022-2836(92)90404-8

Butenko, I., Vanyushkina, A., Pobeguts, O., Matyushkina, D., Kovalchuk, S., Gorbachev, A., et al. (2017). Response induced in *Mycoplasma gallisepticum* under heat shock might be relevant to infection process. *Sci. Rep.* 7:11330. doi: 10.1038/s41598-017-092377

Chopra-Dewasthaly, R., Spergser, J., Zimmermann, M., Citti, C., Jechlinger, W., and Rosengarten, R. (2017). Vpma phase variation is important for survival and persistence of *Mycoplasma agalactiae* in the immunocompetent host. *PLoS Pathog.* 13:e1006656. doi: 10.1371/journal.ppat.1006656

Citti, C., Nouvel, L.-X., and Baranowski, E. (2010). Phase and antigenic variation in mycoplasmas. *Future Microbiol.* 5, 1073–1085. doi: 10.2217/fmb.10.71

Cock, P., Antao, T., Chang, J., Chapman, B., Cox, S., Dalke, A., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004

Czurda, S., Hegde, S. M., Rosengarten, R., and Chopra-Dewasthaly, R. (2017). Xer1-independent mechanisms of Vpma phase variation in *Mycoplasma agalactiae* are triggered by Vpma-specific antibodies. *Int. J. Med. Microbiol.* 307, 443–451. doi: 10.1016/j.ijmm.2017.10.005

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., et al. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.* 36, W465–W469. doi: 10.1093/nar/gkn180

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., et al. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* 39, W13–W17. doi: 10.1093/nar/gkr245

Fisunov, G. Y., Alexeev, D. G., Bazaleev, N. A., Ladygina, V. G., Galyamina, M. A., Kondratov, I. G., et al. (2011). Core proteome of the minimal cell: comparative proteomics of three mollicute species. *PLoS One* 6:e21964. doi: 10.1371/journal.pone.0021964

Fisunov, G. Y., Garanina, I. A., Evsyutina, D. V., Semashko, T. A., Nikitina, A. S., and Govorun, V. M. (2016). Reconstruction of transcription control networks in Mollicutes by high-throughput identification of promoters. *Front. Microbiol.* 7:1977. doi: 10.3389/fmicb.2016.01977

Fleming-Davies, A. E., Williams, P. D., Dhondt, A. A., Dobson, A. P., Hochachka, W. M., Leon, A. E., et al. (2018). Incomplete host immunity favors the evolution of virulence in an emergent pathogen. *Science* 359, 1030–1033. doi: 10.1126/science.aao2140

Glew, M. D., Baseggio, N., Markham, P. F., Browning, G. F., and Walker, I. D. (1998). Expression of the pMGA genes of *Mycoplasma gallisepticum* is controlled by variation in the GAA trinucleotide repeat lengths within the 5' noncoding regions. *Infect. Immun.* 66, 5833–5841.

Glew, M. D., Browning, G. F., Markham, P. F., and Walker, I. D. (2000). pMGA phenotypic variation in *Mycoplasma gallisepticum* occurs in vivo and is mediated by trinucleotide repeat length variation. *Infect. Immun.* 68, 6027–6033. doi: 10.1128/IAI.68.10.6027-6033.2000

Glew, M. D., Markham, P. F., Browning, G. F., and Walker, I. D. (1995). Expression studies on four members of the pMGA multigene family in *Mycoplasma gallisepticum* S6. *Microbiology* 141(Pt 11), 3005–3014. doi: 10.1099/13500872-141-11-3005

Grinevich, A. A., Ryasik, A. A., and Yakushevich, L. V. (2015). Trajectories of DNA bubbles. *Chaos Solitons Fractals* 75, 62–75. doi: 10.1016/j.chaos.2015.02.009

Hegde, S., Zimmermann, M., Rosengarten, R., and Chopra-Dewasthaly, R. (2018). Novel role of Vpmas as major adhesins of *Mycoplasma agalactiae* mediating differential cell adhesion and invasion of Vpma expression variants. *Int. J. Med. Microbiol.* 308, 263–270. doi: 10.1016/j.ijmm.2017.11.010

Jones, S., Shanahan, H. P., Berman, H. M., and Thornton, J. M. (2003). Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Res.* 31, 7189–7198. doi: 10.1093/nar/gkg922

Kamzolova, S. G., Sorokin, A. A., Beskaravainy, P. M., and Osypov, A. A. (2006). "Comparative analysis of electrostatic patterns for promoter and nonpromoter DNA in *E. Coli* genome," in *Bioinformatics of Genome Regulation and Structure II*, eds N. Kolchanov, R. Hofestaedt, and L. Milanesi (Boston, MA: Springer), 67–74. doi: 10.1007/0-387-29455-4_7

Kamzolova, S. G., Sorokin, A. A., Dzhelyadin, T. D., Beskaravainy, P. M., and Osypov, A. A. (2005). Electrostatic potentials of *E. coli* genome DNA. *J. Biomol. Struct. Dyn.* 23, 341–345.

Kamzolova, S. G., Sorokin, A. A., Osipov, A. A., and Beskaravainyi, P. M. (2009). [Electrostatic map of bacteriophage T7 genome. Comparative analysis of electrostatic properties of sigma70-specific T7 DNA promoters recognized by RNA-polymerase of *Escherichia coli*]. *Biofizika* 54, 975–983.

Kowalski, D., Natale, D. A., and Eddy, M. J. (1988). Stable DNA unwinding, not "breathing," accounts for single-strand-specific nuclease hypersensitivity of specific A+T-rich sequences. *Proc. Natl. Acad. Sci. U.S.A.* 85, 9464–9468. doi: 10.1073/pnas.85.24.9464

Kristensen, D. M., Wolf, Y. I., and Koonin, E. V. (2017). ATGC database and ATGC-COGs: an updated resource for micro- and macro-evolutionary studies of prokaryotic genomes and protein family annotation. *Nucleic Acids Res.* 45, D210–D218. doi: 10.1093/nar/gkw934

Lechner, M., Findeiss, S., Steiner, L., Marz, M., Stadler, P. F., and Prohaska, S. J. (2011). Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* 12:124. doi: 10.1186/1471-2105-12-124

Liu, L., Dybvig, K., Panangala, V. S., van Santen, V. L., and French, C. T. (2000). GAA trinucleotide repeat region regulates M9/pMGA gene expression in *Mycoplasma gallisepticum*. *Infect. Immun.* 68, 871–876. doi: 10.1128/IAI.68.2.871-876.2000

Liu, L., Panangala, V. S., and Dybvig, K. (2002). Trinucleotide GAA repeats dictate pMGA gene expression in *Mycoplasma gallisepticum* by affecting spacing between flanking regions. *J. Bacteriol.* 184, 1335–1339. doi: 10.1128/JB.184.5.1335-1339.2002

Ma, L., Jensen, J. S., Mancuso, M., Myers, L., and Martin, D. H. (2015). Kinetics of genetic variation of the *Mycoplasma genitalium* MG192 gene in experimentally infected chimpanzees. *Infect. Immun.* 84, 747–753. doi: 10.1128/IAI.01162-15

Markham, P. F., Glew, M. D., Brandon, M. R., Walker, I. D., and Whithear, K. G. (1992). Characterization of a major hemagglutinin protein from *Mycoplasma gallisepticum*. *Infect. Immun.* 60, 3885–3891.

Markham, P. F., Glew, M. D., Browning, G. F., Whithear, K. G., and Walker, I. D. (1998). Expression of two members of the pMGA gene family of *Mycoplasma gallisepticum* oscillates and is influenced by pMGA-specific antibodies. *Infect. Immun.* 66, 2845–2853.

Markham, P. F., Glew, M. D., Sykes, J. E., Bowden, T. R., Pollocks, T. D., Browning, G. F., et al. (1994). The organisation of the multigene family which encodes the major cell surface protein, pMGA, of *Mycoplasma gallisepticum*. *FEBS Lett.* 352, 347–352. doi: 10.1016/0014-5793(94)00991-0

Matyushkina, D., Pobeguts, O., Butenko, I., Vanyushkina, A., Anikanov, N., Bukato, O., et al. (2016). Phase transition of the bacterium upon invasion of a host cell as a mechanism of adaptation: a *Mycoplasma gallisepticum* model. *Sci. Rep.* 6:35959. doi: 10.1038/srep35959

May, M., Dunne, D. W., and Brown, D. R. (2014). A sialoreceptor binding motif in the *Mycoplasma synoviae* adhesin VlhA. *PLoS One* 9:e110360. doi: 10.1371/journal.pone.0110360

Mazin, P. V., Fisunov, G. Y., Gorbachev, A. Y., Kapitskaya, K. Y., Altukhov, I. A., Semashko, T. A., et al. (2014). Transcriptome analysis reveals novel regulatory mechanisms in a genome-reduced bacterium. *Nucleic Acids Res.* 42, 13254–13268. doi: 10.1093/nar/gku976

Mrázek, J. (2006). Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. *Mol. Biol. Evol.* 23, 1370–1385. doi: 10.1093/molbev/msk023

Mrázek, J., Guo, X., and Shah, A. (2007). Simple sequence repeats in prokaryotic genomes. *Proc. Natl. Acad. Sci. U.S.A.* 104, 8472–8477. doi: 10.1073/pnas.0702412104

Noormohammadi, A. H. (2007). Role of phenotypic diversity in pathogenesis of avian mycoplasmosis. *Avian Pathol. J.* 36, 439–444. doi: 10.1080/03079450701687078

Osypov, A. A., Krutinin, G. G., and Kamzolova, S. G. (2010). Deppdb–DNA electrostatic potential properties database: electrostatic properties of genome DNA. *J. Bioinform. Comput. Biol.* 8, 413–425. doi: 10.1142/S0219720010004811

Papazisi, L., Gorton, T. S., Kutish, G., Markham, P. F., Browning, G. F., Nguyen, D. K., et al. (2003). The complete genome sequence of the avian pathogen *Mycoplasma gallisepticum* strain R(low). *Microbiology* 149, 2307–2316. doi: 10.1099/mic.0.26427-0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *JMLR* 12, 2825–2830.

Pflaum, K., Tulman, E. R., Beaudet, J., Canter, J., and Geary, S. J. (2018). Variable lipoprotein Hemagglutinin A (vlhA) gene expression in variant *Mycoplasma gallisepticum* strains *in vivo*. *Infect. Immun.* 86, e524–e518. doi: 10.1128/IAI.00524-18

Pflaum, K., Tulman, E. R., Beaudet, J., Liao, X., and Geary, S. J. (2016). Global changes in *Mycoplasma gallisepticum* phase-variable lipoprotein gene vlhA expression during *In Vivo* infection of the natural chicken host. *Infect. Immun.* 84, 351–355. doi: 10.1128/IAI.01092-15

Polozov, R. V., Dzhelyadin, T. R., Sorokin, A. A., Ivanova, N. N., Sivozhelezov, V. S., and Kamzolova, S. G. (1999). Electrostatic potentials of DNA. Comparative analysis of promoter and nonpromoter nucleotide sequences. *J. Biomol. Struct. Dyn.* 16, 1135–1143. doi: 10.1080/07391102.1999.10508322

Rocha, E. P. C. (2003). An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. *Genome Res.* 13, 1123–1132. doi: 10.1101/gr.966203

Rodionov, D. A. (2007). Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. *Chem. Rev.* 107, 3467–3497. doi: 10.1021/cr068309

Rosengarten, R., and Wise, K. S. (1990). Phenotypic switching in mycoplasmas: phase variation of diverse surface lipoproteins. *Science* 247, 315–318. doi: 10.1126/science.1688663

Sorokin, A. A., Osypov, A. A., Dzhelyadin, T. R., Beskaravainy, P. M., and Kamzolova, S. G. (2006). Electrostatic properties of promoter recognized by *E. coli* RNA polymerase Esigma70. *J. Bioinform. Comput. Biol.* 4, 455–467. doi: 10.1142/S0219720006002077

Szczepanek, S. M., Tulman, E. R., Gorton, T. S., Liao, X., Lu, Z., Zinski, J., et al. (2010). Comparative genomic analyses of attenuated strains of *Mycoplasma gallisepticum*. *Infect. Immun.* 78, 1760–1771. doi: 10.1128/IAI.01172-09

Tatarinova, T. V., Lysnyansky, I., Nikolsky, Y. V., and Bolshoy, A. (2016). The mysterious orphans of Mycoplasmataceae. *Biol. Direct* 11:2. doi: 10.1186/s13062-015-0104-3

Torres-Cruz, J., and van der Woude, M. W. (2003). Slipped-strand mispairing can function as a phase variation mechanism in *Escherichia coli*. *J. Bacteriol.* 185, 6990–6994. doi: 10.1128/JB.185.23.6990-6994.2003

Tulman, E. R., Liao, X., Szczepanek, S. M., Ley, D. H., Kutish, G. F., and Geary, S. J. (2012). Extensive variation in surface lipoprotein gene content and genomic changes associated with virulence during evolution of a novel North American house finch epizootic strain of *Mycoplasma gallisepticum*. *Microbiology* 158, 2073–2088. doi: 10.1099/mic.0.058560-0

van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H. (1998). Short-sequence DNA repeats in prokaryotic genomes. *Microbiol. Mol. Biol. Rev.* 62, 275–293.

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Wang, H., and Benham, C. J. (2006). Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. *BMC Bioinformatics* 7:248. doi: 10.1186/1471-2105-7-248

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Wei, W., Davis, R. E., Suo, X., and Zhao, Y. (2015). Occurrence, distribution and possible functional roles of simple sequence repeats in phytoplasma genomes. *Int. J. Syst. Evol. Microbiol.* 65, 2748–2760. doi: 10.1099/ijs.0.000273

Zhabinskaya, D., Madden, S., and Benham, C. J. (2015). SIST: stress-induced structural transitions in superhelical DNA. *Bioinformatics* 31, 421–422. doi: 10.1093/bioinformatics/btu657

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

REVIEW
published: 27 November 2018
doi: 10.3389/fgene.2018.00575

Highlights on the Application of Genomics and Bioinformatics in the Fight Against Infectious Diseases: Challenges and Opportunities in Africa

Saikou Y. Bah[1,2], Collins Misita Morang'a[1†], Jonas A. Kengne-Ouafo[1†], Lucas Amenga–Etego[1] and Gordon A. Awandare[1]**

[1] West African Centre for Cell Biology of Infectious Pathogens, University of Ghana, Accra, Ghana, [2] Vaccine and Immunity Theme, MRC Unit The Gambia at London School of Hygiene & Tropical Medicine, Banjul, Gambia

OPEN ACCESS

Edited by:
Ancha Baranova, George Mason University, United States

Reviewed by:
David John Studholme, University of Exeter, United Kingdom
Sandeep Kumar Dhanda, La Jolla Institute for Allergy and Immunology (LJI), United States

***Correspondence:**
Saikou Y. Bah
sbah@ug.edu.gh;
sabah@mrc.gm
Gordon A. Awandare
gawandare@ug.edu.gh

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *18 June 2018*
Accepted: *08 November 2018*
Published: *27 November 2018*

Citation:
Bah SY, Morang'a CM, Kengne-Ouafo JA, Amenga–Etego L and Awandare GA (2018) Highlights on the Application of Genomics and Bioinformatics in the Fight Against Infectious Diseases: Challenges and Opportunities in Africa. Front. Genet. 9:575. doi: 10.3389/fgene.2018.00575

Genomics and bioinformatics are increasingly contributing to our understanding of infectious diseases caused by bacterial pathogens such as *Mycobacterium tuberculosis* and parasites such as *Plasmodium falciparum*. This ranges from investigations of disease outbreaks and pathogenesis, host and pathogen genomic variation, and host immune evasion mechanisms to identification of potential diagnostic markers and vaccine targets. High throughput genomics data generated from pathogens and animal models can be combined with host genomics and patients' health records to give advice on treatment options as well as potential drug and vaccine interactions. However, despite accounting for the highest burden of infectious diseases, Africa has the lowest research output on infectious disease genomics. Here we review the contributions of genomics and bioinformatics to the management of infectious diseases of serious public health concern in Africa including tuberculosis (TB), dengue fever, malaria and filariasis. Furthermore, we discuss how genomics and bioinformatics can be applied to identify drug and vaccine targets. We conclude by identifying challenges to genomics research in Africa and highlighting how these can be overcome where possible.

Keywords: bioinformatics, genomics, infectious diseases, antimicrobial resistant, diagnosis

INTRODUCTION: OMICS AND BIOINFORMATICS IN INFECTIOUS DISEASES

Genomics and bioinformatics have contributed immensely to our understanding of infectious diseases: from disease pathogenesis, mechanisms and the spread of antimicrobial resistance, to host immune responses. Herein, we review some of the major contributions of genomics and bioinformatics in infectious disease research using examples of three diseases that account for large proportions of morbidity and mortality as well as a neglected tropical disease. Specifically, we review *M. tuberculosis*, which causes TB, a disease responsible for approximately two million deaths globally per year. Dengue virus (DENV) causes Dengue fever, which is a re-emerging mosquito borne viral disease, responsible for more than 350 million cases annually (WHO, 2017; World Health Organization Western Pacific Region, 2018). *Plasmodium falciparum* causes malaria,

a parasitic disease that accounts for the highest morbidity and mortality in Sub-Saharan Africa, especially in children under five and pregnant women (WHO, 2018b), and Filariasis, which is a neglected tropical disease. **Figure 1** shows a circular wheel of genomics/bioinformatics as can be applied in infectious diseases as discussed herein, ranging from understanding host and pathogen genome biology to genome-wide association studies (GWAS) as well as the identification of drug targets and drug resistance surveillance to patient management. This encompasses molecular techniques, bioinformatics and clinical applications (**Figure 1**). We also highlight the application of genomics and bioinformatics to the identification of vaccine targets and drug discovery. We conclude by highlighting some challenges of conducting bioinformatics research in resource-limited countries in sub-Saharan Africa.

OMICS OF TUBERCULOSIS PATHOGENS AND HOST RESPONSES

Tuberculosis caused by members of the *M. tuberculosis* complex is a leading cause of death, with about 9 million cases and two million deaths per year globally (WHO, 2018a). The mycobacterial genome was first sequenced in 1998 and many more *M. tuberculosis* genomes have since been sequenced (Cole et al., 1998; Guerra-Assunção et al., 2015; Yun et al., 2016). These genomes provide great avenues for the genomic characterization, development of improved diagnostic tools, drug susceptibility testing, and molecular epidemiology of circulating mycobacterial strains. Host-pathogen genomics and transcriptomics have over the past decade enhanced our understanding of human-mycobacterium interactions and in the identification of potential diagnostic and prognostic markers (Anderson et al., 2014; Maertzdorf et al., 2015).

An understanding of the *M. tuberculosis* genome biology is invaluable in the control of TB. The *M. tuberculosis* genome is GC rich and consists of about 4000 genes and, unlike other bacteria, a large proportion of its genome encodes proteins and enzymes involved in lipogenesis and lipolysis (Cole et al., 1998), reflecting its thick lipid cell wall. TB control is hampered by antimycobacterial resistance, multidrug resistance (MDR) and, recently, extensively drug resistant (XDR) mycobacterial strains (Leisching et al., 2016). Genomics analysis has immensely contributed to the identification of drug resistance-conferring mutations and surveillance (Köser et al., 2013). Whole genome analyses have demonstrated that mycobacterial drug resistance is largely attributed to single nucleotide polymorphisms (SNPs); for example, rifampicin (RIF) resistance arises from mutations in the *rpoB* gene and mutations in the *katG* and *inhA* lead to isoniazid resistance (da Silva et al., 2011). Newly characterized genetic mutations in *M. tuberculosis* genomes have also been shown to play key roles in the emergence of antimycobacterial drug resistance (Sun et al., 2012). Analyses of 161 drug resistant *M. tuberculosis* genomes identified 72 genes, 28 intergenic regions and 21 SNPs with strong and consistent associations with drug resistance (Zhang et al., 2013). Genomic analysis has also identified lineage mutation rate differences and predicted the emergence of antimycobacterial resistance (Ford et al., 2013). A retrospective analysis of thousands of *M. tuberculosis* genomes collected from African and European patients identified 120 resistance-determining mutations for first and second line antimycobacterial drugs, which could be valuable in developing new assays for drug susceptibility testing (Walker et al., 2015). Furthermore, genomics through the use of GWAS has been used to identify novel mutations associated with resistance to cycloserine, ethionamide, and para-aminosalicylic acid, suggesting the involvement of efflux pump in the emergence of resistance (Coll et al., 2018). A number of genomics-based tools have been developed to detect drug resistance including Mykrobe Predictor, PhyResSE, and TB-Profiler, which are easy to use by researchers with no bioinformatics expertise and can predict drug resistance within minutes after obtaining sequences (Bradley et al., 2015; Coll et al., 2015; Feuerriegel et al., 2015). Mykrobe Predictor has a sensitivity and specificity of 82.6 and 98.5%, respectively (Bradley et al., 2015). TB-Profiler was developed using a mutation library consisting of 1,325 mutations in different genes associated with drug resistance in 15 anti-tuberculosis drugs and had more than 75% sensitivity as well as more than 90% specificity for all drugs tested (Coll et al., 2015). A recent study evaluating the performance of these tools showed that their sensitivity ranges from 74 to 80% along with a specificity of more than 95% (van Beek et al., 2018). However, there is still a need for optimization of analysis pipelines to make them applicable in field settings where the disease burden is usually the highest.

Genomics analysis has also been used to determine the evolutionary history and spread of mycobacterial strains such as the Beijing strain, demonstrating its spread from the Far East (Merker et al., 2015). An investigation of *M. tuberculosis* transmission dynamics is important in monitoring outbreak; Mehaffy et al. (2014) demonstrated that whole genome analysis can be used to monitor infections to decipher transmission dynamics. Furthermore, genomics has also been applied to decipher transmission dynamics of *M. tuberculosis* in Vietnam, suggesting that SNPs in ESX-5 type VII secreted protein EsxW could potentially contribute to enhancing transmission (Holt et al., 2018). Furthermore, genomics has been applied to investigate TB outbreaks, genotyping of the outbreak associated lineages, and their evolution during the outbreak (Jamieson et al., 2014; Stucki et al., 2015). Indeed, analysis tools have been developed for the prediction of *M. tuberculosis* spoligotypes from raw sequence reads, and in combination with other analysis tools also determine antibiotic resistance as well as transmission dynamics (Coll et al., 2012; Bradley et al., 2015). Some genomics methods can also be employed to identify mixed infections as well as infections with a single strain and have recently been applied to clinical isolates from Malawi (Sobkowiak et al., 2018).

Genome-wide association study (GWAS) has also been used to identify candidate gene variants associated with susceptibility to active tuberculosis. GWAS analyses in African patients from Ghana, Gambia, Uganda and Tanzania identified TB disease-associated SNPs located on three chromosomal loci: 18q11, 11p13, and 5q33 (Thye et al., 2010, 2012; Sobota et al., 2016). Similarly, GWAS studies have also been done in Europe

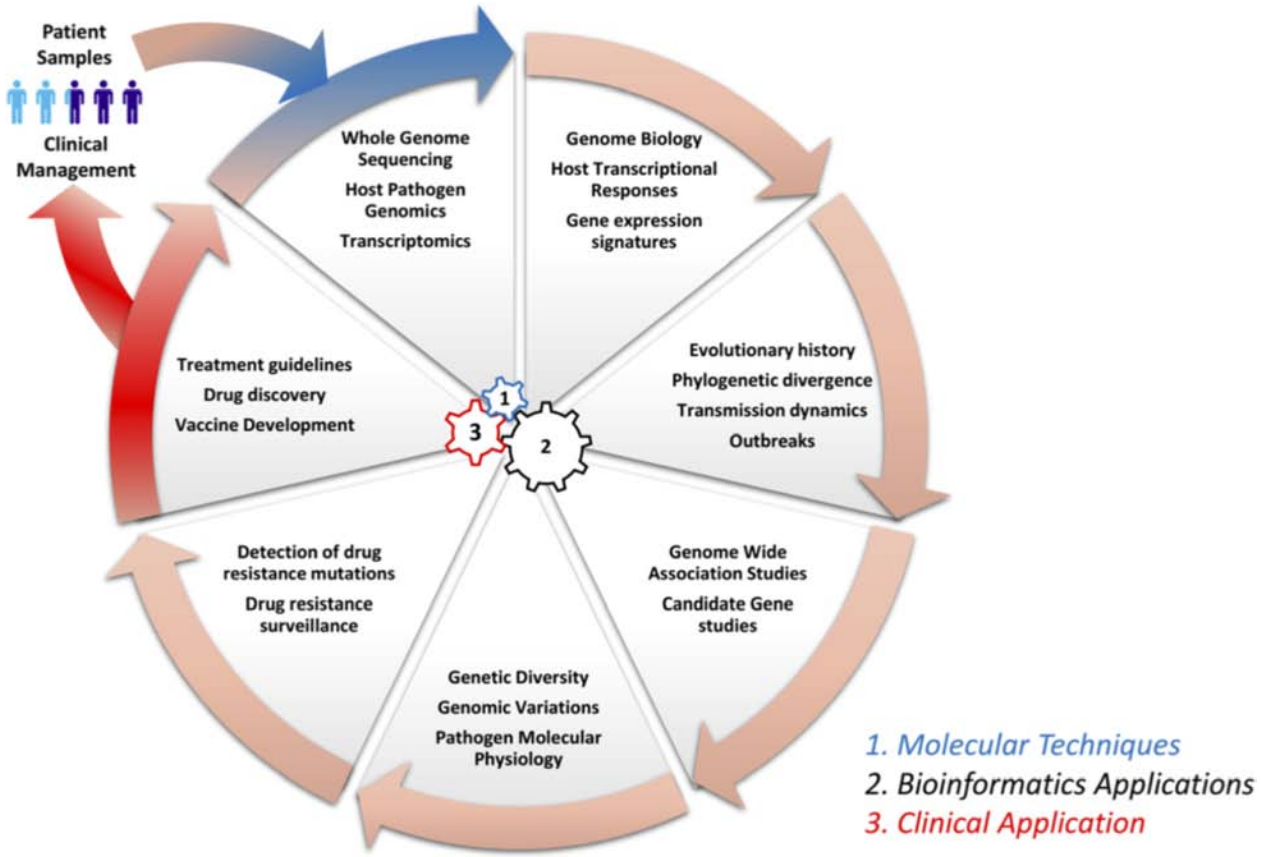


FIGURE 1 | Circle of infection disease genomics/bioinformatics. This shows a representation of genomic laboratory technologies, bioinformatics analysis, and potential applications. This includes: (1) molecular techniques such as whole genome sequencing by methods like Illumina to generate sequence reads, which are needed for the (2) bioinformatics application to study host systemic responses, pathogen genomics, and transmission dynamics. Further, bioinformatics can be applied to determine genetic diversity, investigation of drug resistance mechanisms and surveillance, and the identification of vaccine targets in systems vaccinology. Finally (3), all this information can be integrated to define treatment guidelines and patient management.

identifying SNPs in the *ASAP1* gene on chromosome 8q24 and in a genomic region in which class II human leucocyte antigen (HLA II) is encoded (Curtis et al., 2015; Sveinbjornsson et al., 2016). Recently, a GWAS study in a Han Chinese population also found SNPs in mitofusin-2 (*MFN2*), regulator of G protein signaling 12 (*RGS12*) and HLA II beta chain to be associated with active TB (Qi et al., 2017). This highlights that host genetics play significant roles in susceptibility to active TB and may explain why some individuals remain latently infected while some develop active TB despite having similar exposure levels. Furthermore, based on host genetic variants, GWAS analysis could be applied to identify latently infected individuals who are at a high risk of developing active TB for preventative interventions. Once validated, identified SNPs can be used to develop point of care diagnostics to identify high risk people for mass preventative treatment.

Host transcriptomics are increasingly being used to understand systemic responses to infections and to identify diagnostic and prognostic markers. Mistry et al. (2007) were among the first to use microarray technology to study host systemic response to TB, identifying a nine gene-signature with potential for TB diagnosis. Jacobsen et al. (2007) applied microarray analysis to investigate the host pathway biology and potential diagnostic biomarkers. Analyzing peripheral blood mononuclear cells (PBMCs), they found a monocyte-derived gene expression signature identifying CD64, lactoferrin and Ras-Associated GTPase-33A as potential diagnostic biomarkers, which were further validated in another independent study population in South Africa (Maertzdorf et al., 2011). Applying gene set enrichment analysis to microarray gene expression identified metabolic pathways such as insulin metabolism, immune cell differentiation and inflammation in TB (Lesho et al., 2011). A neutrophil-driven interferon signature consisting both type I and type II interferon during TB was also identified using microarray analysis (Berry et al., 2010). The type I interferon pathway was also observed by Ottenhoff et al. (2012) identifying IL15RA, UBE2L6, and GBP4 as the main molecules involved. A 393-transcript signature for active TB and an 86-transcript signature with a potential for distinguishing TB from other inflammatory diseases were also identified (Berry et al., 2010). In addition, a biosignature consisting of 27 transcript signatures to distinguish active from latent TB and 44 transcript signatures to

distinguish active TB from other diseases were recently identified (Kaforou et al., 2013). Microarrays have also been used to demonstrate that host transcriptional responses to *M. africanum* and *M. tuberculosis* differ following treatment (Tientcheu et al., 2015), which could be important in the management of patients infected with the different mycobacterial strains. Furthermore, host gene expression has also been used to monitor treatment responses and predict treatment outcome, which will be valuable in testing new drug regimens and new antimycobacterial drugs (Thompson et al., 2017). These studies prove the potential of host genomics in providing a better understanding of disease pathophysiology, prognosis and host pathway biology in response to an infectious agent.

In addition, arrays have also been applied to childhood TB, to identify signatures for active tuberculosis and a signature that distinguishes active tuberculosis from other diseases in sub-Saharan Africa (Anderson et al., 2014). Similarly, a 9-gene signature was also identified in Warao Amerindian children, further highlighting the potential of using host biomarkers for TB diagnosis (Verhagen et al., 2013). Host transcriptional analysis is moving from array-based technologies to RNA sequencing and has been applied to 16 gene signatures that identified people with a high risk of developing TB 2 years before diagnosis in sub-Saharan Africa (Zak et al., 2016). However, it is noteworthy that identified biosignatures have a variable number of genes, from about 10 to more than 100, and there is very little overlap between some signatures. It will be valuable to conduct a meta-analysis of available datasets to increase statistical power and identify high confidence signatures across studies regardless of circulating pathogens and local environmental factors. In doing such analysis, confounders due to technologies, age and circulating endemic pathogens can be accounted for to give a strong as well as diagnostic and prognostic signature. These studies highlight the potential application of genomics and bioinformatics to interrogate host response for the diagnosis and prognosis of TB, which will contribute immensely to curbing TB morbidity and mortality.

DENGUE VIRUS RESEARCH IN THE ERA OF BIOINFORMATICS

Dengue virus (DENV) is a pathogenic single-stranded RNA virus that belongs to the flavivirus genus, which comprises other known pathogenic viruses such as West Nile, yellow fever, Japanese encephalitis, St. Louis encephalitis, tick-borne encephalitis, Omsk hemorrhagic fever and Zika virus (Gould and Solomon, 2008). The re-emergence, evolution, diversity and geographic distribution of flaviviruses make them interesting pathogens (Moureau et al., 2015). Phylogenetic analysis of divergence times suggests that flaviviruses originated from a common ancestor (100,000 years ago) and later split into mosquito and tick borne flaviviruses about 40,000 years ago (Holbrook, 2017). Approximately 40% of the world population is at risk of DENV infection with more than 350 million cases reported annually.

Illumina SNPs genotyping and SNPs identified through whole genome analysis have been used in case-control GWAS statistical analysis to identify SNPs that predispose or confer protection against DENV infection (de Carvalho et al., 2017). The DENV shock syndrome (DSS) has been shown in a GWAS analysis of SNPs in a cohort of 2008 pediatric cases to have a strong association ($P < 0.5 \times 10^{-8}$) with the human major histocompatibility complex (*MHC*) (rs3132468) on chromosome 6 and phospholipase C (rs3740360 and rs3765524) on chromosome 10 (Khor et al., 2011). Dang et al. replicated the study in 917 Thai children with DSS and confirmed that alleles rs3132468 [MHC I chain related protein A (*MICB*)] and rs3765524 [phospholipase C epsilon 1 (*PLCEI*)] predispose Southeast Asians to DSS (Dang et al., 2014). In contrast, Whitehorn et al. (2013) genotyped 3,961 confirmed cases and 5,968 controls and found that rs3132468 *MICB* and rs3740360 alleles *PLCEI* were associated with less severe phenotypes of DENV infection in both infants and adults. This implies that the effect of these SNPs could be population-specific. Other candidate genes include dendritic cell-specific intracellular adhesion molecule (*ICAM*)*-3* grabbing non-integrin (*DC-SIGN*), C-Type Lectin Domain Containing 5A (*CLEC5A*), immunoglobulin gamma constant fragment receptor (*FCGRIIA*), Toll-Like receptors (*TLRs*), Tumor necrosis Factor (*TNF*), Interferons (*IFNs*), 2′-5′-oligoadenylate synthase (*OASs*), Janus Kinase (*JAK*), Stimulator of Interferon Genes (*STING*), cytokines, chemokines, *ICAM-1* and tryptase 1 proteases (de Carvalho et al., 2017).

Whole genome sequencing (WGS) and phylogenetic methods have been used to investigate DENV outbreaks. Faria et al. (2017) analyzing 92 viral genomes from DENV patients during the 2012 outbreak in Rio de Janeiro, found that at least two thirds of infections went unnoticed and their analysis highlighted the scale of the epidemic spread of DENV after the outbreak. Ahn et al. (2015) investigated the genetic variations in 8,826 nucleotide sequences of whole-genome DENV virus, and demonstrated that there was a distinctive genetic pattern between the four DENV subtypes across different regions (American, Oceanian, Asian, and Africa).

Analyses of envelope encoding nucleotide sequences from India have shown a shift from DENV subtype III to subtype IV, suggesting some level of positive selection (Manakkadan et al., 2013). These phylodynamic methods, which indicate evolutionary process or patterns of genetic diversity of the DENV virus, have also been reconciled with the virus epidemiology so as to decrease the variation between the two methods that are mainly used to study the population dynamics or viral behaviors (Pybus et al., 2012; Rasmussen et al., 2014). Due to the importance of genomics and bioinformatics in viral research, a range of tools has been developed to analyze viral genomes and make inferences (Stamatakis, 2014; Brody et al., 2017).

The use of RNA folding, structural predictions and functional studies has shown that genetic variation of the DENV occurs in nature due to high rates of recombination and error-prone RNA polymerases. A deleterious DENV genome was first shown by Aaskov et al. (2006) whereby a stop codon in the envelope coding region resulted in a defective DENV. Li et al. (2011)

also discovered defective interfering viral particles by analyzing short fragments of DENV, suggesting that they may be part of a broader disease attenuating process mediated by the deleterious virus and the defective interfering particles are important in viral replication, thereby enhancing the overall transmission capability of DENV (Li and Aaskov, 2014). Structural RNA predictions have implicated other elements in modulating replication of the virus, such as the downstream cyclization sequence (Friebe et al., 2012), cis-acting elements occurring in the capsid coding region (de Borba et al., 2015), and elements in the promoter Stem Loop A (SLA) and non-structural protein 5 (NS5) regions (Gebhard et al., 2011).

Understanding intra- and inter-host genetic diversity was previously mired with experimental and analytical methods that did not fully account for errors in viral amplifications. Thai et al. (2012) used various statistical approaches to correct for the artefactual mutations resulting from PCR amplifications and sanger sequencing, and showed that the genetic diversity index (Pi) of the DENV was low, ranging from 0 to 0.0013. This suggested sequence conservation, but they were able to show mixed infections and phylogenetically distinct DENV lineages present within the same host. Furthermore, genome-wide scans for patterns of intra-host diversity in DENV identified variants between genes suggesting significant differences in intra-host diversity of the virus in the Nicaraguan population (Parameswaran et al., 2012). Functional annotation of the variants showed the impact of viral mutations on protein function, which strongly suggested purifying selection across transmission events.

Deep sequencing, RNA structural analysis and fitness evaluation have been used to determine processes that DENV employs for host specialization (mosquito or human) using RNA elements in the 3′-UTR (Villordo et al., 2015). A host adaptable stem loop structure was found to be duplicated, which DENV uses to accumulate mutations that are beneficial in one host and deleterious in another host, but the duplication confers a robust mechanism during host switching (Villordo et al., 2015). Recently, Waman et al. (2016) used population genetics methods to compute the genotype diversity and evolution of 990 DENV genomes, and revealed that the DENV-2 population is subdivided into 15 lineages. Their study also indicated the presence of intra-genotype diversity and that the population structure of DENV-2 is spatiotemporal, shaped by episodic positive selection and viral recombination (Waman et al., 2016). The application of genomics and bioinformatics in the study of DENV shows the complexity of the virus biology, which can be exploited in target identification for drug discovery and vaccine development (Guy et al., 2016; Low et al., 2017).

PROGRESS IN MALARIA GENOMICS

Malaria incidence and mortality rates decreased by 21 and 29%, respectively, between 2010 and 2015 (WHO, 2018b). The genetic landscape of *P. falciparum,* the main cause of malaria, is increasingly being unraveled by using deep sequencing to identify polymorphisms and structural and copy number variations, which are fundamental for parasite evolution (Kwiatkowski, 2015). Sequencing consortia such as the MalariaGEN improve our understanding of genomics of both the Anopheles vector and the plasmodium species[1]. A recent study on genotyping accuracy using deep sequencing of *Plasmodium* parental generations and their progenies revealed that polymorphism frequencies can be used as markers of high recombination rates (Miles et al., 2016), which is an important contributor to enhancing immune evasion and drug resistance. Using whole genome deep sequencing and micro-array analysis, a study observed 18 deletions on regions encoding multigene families that are associated with immune evasion (Bopp et al., 2013). The authors showed the presence of chromosomal crossovers in six of the deletions and were able to estimate mutation rates of *P. falciparum* (Bopp et al., 2013).

Bioinformatics has contributed to our understanding of resistant mechanisms to previous drugs such as chloroquine and the emerging resistance to artemisinin-based combination therapies (ACT). Robinson et al. deployed next generation sequencing to investigate multi-clonality, population genetics and drug-resistant genotypes (Robinson et al., 2011). More recently, WGS was used to discover that mutations in the Kelch propeller domain (K-13) are associated with ACT resistance in Cambodia (Ariey et al., 2014; Straimer et al., 2015). Profiling of the drug resistance genes [*P. falciparum* chloroquine resistance transporter (*pfcrt)*, *P. falciparum* multidrug resistance (*pfmdr1*), *P. falciparum* dihydrofolate reductase (*dhfr*) and *P. falciparum* dihydropteroate synthetase (*dhps*), and *P. falciparum* Kelch protein 13 (*pfk13*)] was done using Illumina next generation sequencing and demonstrated that the resistance-associated K-13 variants were largely absent in Africa (MalariaGEN Plasmodium falciparum Community Project, 2016; Nag et al., 2017).

Furthermore, bioinformatics tools have been used to demonstrate multi-locus linkage disequilibrium and local diversity, recent selection through integrated haplotype scores, regional gene flow and allele frequency differentiations (Duffy et al., 2017). Intra-host diversity can now be statistically characterized using the F*ws* metrics because sequencing platforms are able to generate read count data. Auburn et al. characterized within host diversity in 64 samples from West Africa, capturing a multiplicity of infections, number of clone ratios, clonal variation and within-host diversity (Auburn et al., 2012). Bioinformatics analysis of deep sequencing revealed large-scale genetic variations in *P. falciparum* (86158 SNPs), and genome wide allelic frequencies, population structure, linkage disequilibrium and intra-host diversity (Manske et al., 2012). The genetic diversity of *P. falciparum* is dependent on directional and balancing selection, whereby drug pressure and host immunity are the major selective agents, respectively (Mobegi et al., 2014; Duffy et al., 2015).

Genomics has been used to discover novel malaria resistance loci in humans, which provide 33% protection from severe malaria (Malaria Genomic Epidemiology Network, 2015). In Ghana, GWAS identified two unknown genetic loci associated with severe malaria: 1q32 within the ATPase Plasma Membrane Ca^{2+} Transporting 4 (*ATP2B4*) gene and the 16q22.2 linked

[1]https://www.malariagen.net/

to a tight junction protein known as *MARVELD3* (Timmann et al., 2012). Most recently, GWAS was used in a longitudinal surveillance to detect K-13 signatures, which led to the identification of a Kelch variant that is suggested to be a potential modulator of artemisinin resistance (Cerqueira et al., 2017).

The *Plasmodium* pathophysiology is increasingly being explored using transcriptomics and proteomics. Bioinformatics and statistical models have been used to describe the genome-wide translational dynamics of *P. falciparum*, showing that parasite transcription and translation are tightly coupled presenting a broad and high resolution of parasite gene expression profiles (Caro et al., 2014). ChIP-Seq and RNA sequencing have been used for polysome profiling to understand the regulation of *Plasmodium* gene expression in humans. Bunnik et al. (2013) observed a delay in peak polysomal transcript abundance for several genes as compared to the mRNA fraction, which they reported to be alternative polysomal mRNA splicing events of non-coding transcripts.

DNA microarray technologies had been used to describe the gene expression patterns of *P. falciparum* during the intra-erythrocytic stage (Bozdech et al., 2003), gametocyte (Young et al., 2005), sporozoite (Siau et al., 2008), liver stage (Tarun et al., 2008), and even between three different strains (Llinás et al., 2006). Recently, microarrays have been used to characterize parasite transcriptomes during cerebral and asymptomatic malaria, which revealed some differentially expressed genes encoding proteins involved in protein trafficking, Maurer's cleft proteins, transcriptional factor proteins and several hypothetical proteins (Almelli et al., 2014). RNA sequencing has also been used to describe *P. falciparum* expression profiles at different time points and has found novel gene transcripts, alternative splicing events and predicted untranslated regions of some genes providing further information on the parasite biology (Otto et al., 2010). Yamagishi et al. (2014) simultaneously analyzed the human host and the parasite transcriptomes using RNA sequencing, and showed that several human and parasite genes such as Toll-like receptor 2 and TIR domain-containing adapter molecule 2 (*TICAM2*) correlated with clinical symptoms. RNA sequencing has also been employed to study the transcriptome of *P. vivax,* which revealed a hotspot of *vir* genes on chromosome 2, new gene transcripts and the presence of species-specific genes (Zhu et al., 2016). It would be valuable to compare this data with similar data from other related *Plasmodium* species to identify species-specific transcriptomes. Analyzing the transcriptome of Chloroquine sensitive and resistant parasites identified 89 upregulated genes and 227 downregulated genes that were associated with resistance (Antony et al., 2016). These differentially expressed genes are involved in immune evasion mechanisms, pathogenesis, and various host-parasite interactions and could be targeted for drug and vaccine development.

Currently, single-cell RNA sequencing is revolutionizing the study of cell-to-cell heterogeneity. For example, the use of this method led to the discovery of novel variations in the expression of specific gene families that are involved in host-parasite interactions among asexual populations (Reid et al., 2018). Altogether, these studies demonstrate the profound impact of malaria parasite transcriptomics and genomics on our understanding of the parasite (Lee et al., 2017), and identify possible candidate targets for drugs, vaccines and diagnostics (Ludin et al., 2012; Hoo et al., 2016).

GENOMICS RESEARCH IN FILARIASIS

Filariasis is a neglected chronic disease caused by tissue-dwelling nematodes (filariae) with onchocerciasis and lymphatic filariasis (LF), causing significant health concerns with a disease burden approaching 86 million cumulatively (WHO/Department of Control of Neglected Tropical Diseases, 2016). Onchocerciasis is caused by *Onchocerca volvulus* while LF is caused by three different parasites, namely *Wuchereria bancrofti*, *Brugia malayi*, and *Brugia timori* (Taylor et al., 2010). Elimination of filariasis is challenging because of the unavailability of sensitive diagnostic tools, lack of appropriate treatments and inadequate control measures in resource limited countries.

The *W. bancrofti* and *O. volvulus* genomes have been sequenced, providing opportunities for further genomic analyses (Desjardins et al., 2013; Cotton et al., 2016). Bioinformatics revealed the presence of gene coding for host immune system regulators such as human-like autoantigens as well as serine and cysteine protease inhibitors (Molehin et al., 2012; Cotton et al., 2016).

Molecular studies coupled with computational analyses have demonstrated an association between human host factors and filariasis clinical manifestations. LF infections have been shown to cluster in some families using pedigree studies (Cuenco et al., 2004; Chesnais et al., 2016). These studies show that genetic factors are involved in the regulation of LF infections and affect both the presence and intensity of microfilariae. However, a GWAS would be more comprehensive to demonstrate this genetic susceptibility to LF as has been the case for a tropical lymphedema (Podoconiosis) of non-filarial origin (Tekola Ayele et al., 2012). It is worth mentioning that lymphedema, or elephantiasis, is one of the main features of LF and normally occurs as a result of a compromised lymphatic system (Addiss, 2010). As opposed to LF, which is infectious, Podoconiosis is a non-communicable disease caused by soil particles such as aluminum and silica predominant in volcanic regions (Price, 1976; Davey et al., 2007). A comparative genomics-based study of LF would help to better understand these clinical manifestations.

Most of the pathological features of LF are associated with human-immunogenetics (Taylor, 2003; Junpee et al., 2010), which has been investigated using genomics and bioinformatics. Gene candidate-based genomics studies carried out in Thailand revealed that polymorphisms in the *TLR-2* gene (−196 to −173 deletion, +597 T > C and +1350 T > C) have a strong linkage disequilibrium and were associated with increased risk of asymptomatic LF (Junpee et al., 2010). In a functional study, individuals with the −196 to −173 deletion were found to have significantly low transcription levels compared to those with the wild-type gene (Junpee et al., 2010). Further analyses showed strong association of a mutation (M196A) in human tumor necrosis factors (TNF) receptor-II with hydrocele development,

while the A288S mutation of endothelin-1 (ET-1) correlated with low ET-1 1 plasma levels and elephantiasis (Panda et al., 2011).

Population genetics is very important for assessing and understanding the epidemiology and transmission dynamics of filarial diseases (Small et al., 2016; Doyle et al., 2017). Population genomics of *O. volvulus* samples collected from different geographical zones – West Africa (WA), Uganda and Ecuador – demonstrated some level of population structure between WA and other populations (Choi et al., 2016). Furthermore, phylogenetic signals indicative of gene flow and genetic admixture between WA forest and savanna populations were identified. These signals could serve as markers to delineate forest from savanna populations and/or sort out admixed populations (Choi et al., 2016). A study using both nuclear and mitochondrial sequences identified regions in the *W. bancrofti* genome that exhibited an arrangement which was consistent with both balancing and directional selection (Small et al., 2016).

The control of filariasis in general is difficult due to the complex parasite life cycle. In an attempt to demystify the complex life cycle of the parasite, RNA sequencing has been used to investigate gene expression profiles of different developmental stages of *Brugia malayi* (Choi et al., 2011). Transcriptomics analyses revealed stage-specific gene expression correlating with stage-specific pathway activation. Upregulated proteins included cathepsin L and Z-like cysteine proteases that were previously demonstrated to be essential for larva molting in *O. volvulus* (Lustigman et al., 2004) and cuticle and eggshell remodeling in filarial nematodes in general (Guiliano et al., 2004). Another study using a filarial microarray chip composed of 18,104 gene probes revealed that gene expression in *B. malayi* infective larvae (L3s) depends on environmental factors (Li et al., 2009). The gene expression patterns in irradiated L3s, laboratory-adapted L3s and those collected from mosquitoes were found to be different. Gene Ontology analyses showed that upregulated genes in laboratory-adapted and mosquito-derived L3s were mostly involved in growth and invasion, whereas those in irradiated L3s were enriched with immunogenic proteins and proteins involved in radiation repair (Li et al., 2009). Such high throughput genomics analysis is important for understanding the biology/development, invasion, and immune evasion mechanisms of the parasite and could help improve disease control measures (Choi et al., 2011).

Mass drug treatment with Ivermectin (IVM) or Mectizan® and Albendazole is the main strategy for filariasis control in Africa and has been going on for decades (Amazigo, 2008). However, cases of drug resistance have been reported and genomic methods are increasingly being used to investigate mechanisms of resistance. Genotyping and sequencing studies have shown an association between SNPs in some *O. volvulus* genes (P-glycoprotein-like protein, β-tubulin) and the development of resistance (Nana-Djeunga et al., 2012; Osei-Atweneboana et al., 2012). P-glycoprotein was recently demonstrated to be associated with resistance to IVM in a horse filarial species (cyathostomins) with transcript levels measured by RNA-Seq and confirmed by RT q-PCR found to be significantly higher in the resistant compared to sensitive worm population (Peachey et al., 2017). Moreover, GWAS demonstrated that reduced sensitivity of *O. volvulus* to IVM is accounted for by genetic drift and soft selective sweeps. Pooled next generation sequencing of *O. volvulus* worms collected from Ghana and Cameroon repeatedly treated with IVM and phenotypically characterized into poor responder (PR) and good responder (GR) parasites identified genetic variants that considerably delineate GR and PR parasites. One of these variants (SNP, OM1b_7179218) was common in both Cameroon and Ghana worm populations, whereas the others were country-specific (Nana-Djeunga et al., 2014; Doyle et al., 2017). These variants were found to be grouped in quantitative trait loci (QTLs) in which published genes associated with IVM resistance were scarcely found. Gene Ontology[2] analysis revealed that genes found in those QTLs regions were linked to pathways involved in neurotransmission, development, and stress responses (Harris et al., 2004; Doyle et al., 2017). The involvement of neurotransmission is a promising finding here because one of the main targets of IVM is a ligand-gated channel at neuromuscular junctions (Cully et al., 1994).

The molecular mechanism of Ivermectin is not clearly understood and has been investigated using bioinformatics approaches. RNA-Seq analyses of ivermectin-challenged *B. malayi* adult female worms revealed that genes involved in cell division (meiosis) and oxidative phosphorylation were drastically downregulated as early as 24 h post-exposure (Ballesteros et al., 2016). A similar study in which the worms were instead challenged with flubendazole (FLBZ), a potential macrofilaricide, demonstrated the effect of FLBZ on embryogenesis and cuticle integrity (O'Neill et al., 2016a). Expression of cuticle-related genes and those involved in mitosis or meiosis were notably affected by the treatment. These studies further elucidate the drug-induced inhibition of embryogenesis and microfilarial release from the female worm uterus during larval development as previously demonstrated (O'Neill et al., 2015, 2016b). Knowledge of this mechanism could help in drug repurposing whereby drugs known to have a similar mode of action or mechanism, but are used for the treatment of other parasitic diseases, could be tested for their efficacy on filarial parasites.

APPLICATION OF OMICS TO VACCINE TARGET IDENTIFICATION AND DRUG DISCOVERY

The availability of whole genome sequences of both the host and pathogens in different databases such as GenBank[3] (Benson et al., 2004), EuPathDB ([4]formerly ApiDB), WormBase[5], Virus Pathogen Database and Analysis Resource (ViPR) has led to tremendous advances in the search for new drug and vaccine targets (Yan et al., 2015; Xia, 2017). This enables high throughput *in silico* screening for the identification of vaccine and drug targets, thus focusing expensive laboratory screening on selected high affinity targets. Though not yet fully implemented in Africa,

[2]http://geneontology.org/
[3]http://www.ncbi.nlm.nih.gov
[4]http://EuPathDB.org
[5]http://www.wormbase.org

omics technologies and bioinformatics analyses have aided significantly in the generation of new knowledge toward drug and vaccine target discovery (Yan et al., 2015; Xia, 2017). Genomic, transcriptomic and proteomic analyses of pathogens such as *filariasis* parasites have identified new potential biomarkers that can be invaluable in diagnostics, vaccine and drug development (Armstrong et al., 2016; Bennuru et al., 2017). Kumar et al. (2007), using genome wide *C. elegans* RNA-interference data as proxy, identified a set of 3,059 essential genes in the *B. malayi* genome, from which 589 were characterized as potential drug targets. The prioritization algorithm helps in the prediction of the efficacy, selectivity and tractability of each target.

Phylogenomic analyses across *Plasmodium* spp. and comparative genomic studies in humans have led to the identification of new drug targets in *P. falciparum*. Identification of essential genes (targets) responsive to specific inhibitors led to the discovery of 40 potential drug targets, which includes known ones such as calcium dependent protein kinase and previously unknown ones such as phosphoisomerase and carboxylase (Ludin et al., 2012). Comparing the transcriptomes of six *Plasmodium* spp. during blood stage infection revealed about 800 genes that have similar expression patterns across species, among which 240 were demonstrated to be druggable by online drug target prioritization databases (Hoo et al., 2016). Similarly, genomic and transcriptomic analyses have been carried out with other pathogens with encouraging results in fungi (Kaltdorf et al., 2016), bacteria (Turab Naqvi et al., 2017), and viruses (Dapat and Oshitani, 2016).

In vaccine target identification, pathogen genomes are being scanned in a bid to identify genes encoding proteins or molecules with vaccine candidate properties such as low antigenic variation, polymorphism, and immunogenicity (Masignani et al., 2002; De Groot et al., 2008). Despite the success of whole-organism vaccines such as those for polio, whole-organism vaccines for pathogens such as *Plasmodium spp.*, *Mycobacterium spp.* and HIV remain a challenge (Doolan et al., 2014; Proietti and Doolan, 2015). Genomics offers a potential way around this challenge through the discovery of immunogenic antigens using whole-genome scans (Doolan et al., 2014; Proietti and Doolan, 2015). Here, omics techniques and bioinformatics tools are used to determine genes or proteins that are involved in the virulence of the pathogen and pathogenesis of the disease by comparing, for example, attenuated and pathogenic disease agents. Algorithms can be used to predict T cell epitopes or regions with high affinity within HLA molecules in translated peptides found in databases (Grubaugh et al., 2013; Davies et al., 2015) in order to inform the choice of the right antigens for vaccine design. Omics technologies have been reviewed in the context of vaccine target identification by He (2012).

Most of the tools used for epitope identification rely on statistics and machine learning. Some of them include servers to predict MHC-binding, peptides namely RANKPEP (Reche et al., 2004), which uses Position Specific Scoring Matrices (PSSMs), and nHLAPred[6] (Bhasin and Raghava, 2007), based on Artificial Neural Networks (ANNs) and quantitative matrices among others. Some severs are specific for B-cell epitope prediction, such as Bcepred[7] (Saha and Raghava, 2004), ABCpred[8] (Saha and Raghava, 2006), and BepiPred[9] (Jespersen et al., 2017). These tools work based on the physicochemical properties and location of the peptides. They function alongside epitope-containing databases such as Swiss-Prot, SYFPEITHI, and IEDB (Fleri et al., 2017). The list of tools, methods and databases mentioned here is not exhaustive, however, they have been extensively reviewed elsewhere (Soria-Guerra et al., 2015).

Nowadays, due to advances in the fields of computer sciences, genomics, proteomics, bioinformatics and management of patients' health records, etc., there seems to be a paradigm shift from generalized medicine to personalized therapy (Sorber et al., 2017). For example, many drugs are metabolized by cytochrome P450 enzymes with drug action depending on the expressed gene variant (BlueCross and BlueShield Association, 2004; Daly et al., 2006). Moreover, malaria patients with glucose-6-phosphate (G6p) deficiency have been reported with severe complications such as cardiotoxicity and acute hemolytic anemia following treatment with quinidine gluconate (Damhoff et al., 2014). These complications have been described as a consequence of inherited (X-linked trait) mutations in the *g6p* gene (Luzzatto and Seneca, 2014). These mutations do not cause the complete loss of the G6P enzyme but instead affect its stability and level in red blood cells (Luzatto et al., 2001). In the same line rifampicin, which is the drug of choice for TB treatment, is transported after administration by a human anion transporter encoded by the *SLCO1B1* gene. Studies have shown that mutations in the *SLCO1B1* gene, namely rs11045819 and rs4149032, are associated with decreased RIF plasma levels in South-African populations (Weiner et al., 2010; Chigutsa et al., 2011; Gengiah et al., 2014). However, this finding could not be replicated in Malawian and South Indian populations, implying that this could be population-specific (Ramesh et al., 2016; Sloan et al., 2017). These show, in a nutshell, the implication of genomics and bioinformatics in drug discovery and precision therapy (Hamburg and Collins, 2010; Rabbani et al., 2016).

CHALLENGES AND OPPORTUNITIES IN CONDUCTING OMICS AND BIOINFORMATICS STUDIES IN AFRICA

Bioinformatics is increasingly becoming an important cornerstone in contemporary research on infectious diseases (Mulder et al., 2017), where Africa has the highest morbidity and mortality but less genomics research output compared to other regions of the world (Fatumo et al., 2014; Karikari, 2015). This slow pace of genomics research output is due to several challenges in omics and bioinformatics research facilities in Africa; three of the major ones are briefly discussed.

[6]http://www.imtech.res.in/raghava/nhlapred/

[7]http://www.imtech.res.in/raghava/bcepred/

[8]http://www.imtech.res.in/raghava/abcpred/

[9]http://www.cbs.dtu.dk/services/BepiPred/

Inadequate Infrastructure

Bioinformatics and genomics analysis require powerful computers and a reliable source of electricity for large data storage and high throughput analyses (H3Africa Consortium et al., 2014). With the exception of some South African universities, most sub-Saharan African universities lack high performance computing facilities (Karikari et al., 2015; Mulder et al., 2016). There is also a limitation of high-speed internet for sharing data and accessing bioinformatics databases and repositories (Fatumo et al., 2014; Karikari, 2015). This hinders the application of cloud-based web services which could have circumvented the need for local high-performance computing facilities (Navale and Bourne, 2018). Furthermore, few research institutions in Africa have sequencing facilities and therefore resort to sequencing abroad through collaborations. Such collaborations often result in a loss of ownership of the data and resulting publications usually have the external collaborators as lead and correspondence authors. Notable efforts being made to bridge this infrastructural gap include the installation of high-performance computers (HPCs) at The Developing Excellence in Leadership and Genetics Training for Malaria Elimination in sub-Saharan Africa (DELGEME) at the University of Science Technique and Technologies of Bamako, Mali, the West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), University of Ghana and the Medical Research Council Unit, The Gambia at the London School of Hygiene and Tropical Medicine, to support storage and high throughput analyses of genomic data. These HPC facilities are complemented by NGS sequencing facilities at WACCBIP and MRC in addition to some institutions in East Africa such the International Livestock Research Institute (ILRI-Kenya). This infrastructural development, and pressure from initiatives such as Human Heredity and Health in Africa (H3Africa), will hopefully serve as a springboard for Africa to increase her involvement in the study design, sample collection, analysis and ownership of data rather than just collecting samples for international collaborators.

Lack of Training Opportunities and Well-Structured Bioinformatics Courses

Until the recent introduction of bioinformatics training courses by H3ABioNet, there were limited bioinformatics training courses in Africa. Such training programs were mostly short courses organized by local bioinformaticians with support from experts in the field across Africa and other external collaborators (Gurwitz et al., 2017). Very few African universities have structured bioinformatics courses, most of these universities are South African, while some are North African and few are in sub-Saharan Africa (Bishop et al., 2015). The DELGEME, through funding from the Wellcome Trust, is also providing funding for Master of Science courses in bioinformatics, which are mostly done in South Africa. The other form of bioinformatics training is through local capacity building, which institutions organize for staff with support usually through North-South collaborations and transfer of expertise. However, the downside of short courses is that there is no mentorship beyond the course, which hinders consolidation of the knowledge gained. In addition to these, some organizations working predominantly on crop production, such as the International Institute of Tropical Agriculture Bioscience Center[10] and Consultative Group on International Agricultural Research institute[11], offer short bioinformatics training opportunities to African scholars. Sometimes some students from Africa get training from European universities, but the challenge is that most of the trainees do not come back to join local institutions because of poor infrastructures. Furthermore, there is a disconnect between biologists and other scientific disciplines such as computer science, statistics and mathematics in most African universities. This affects multidisciplinary research, which is crucial in modern-day infectious disease research. Ultimately, the lack of well-structured bioinformatics curricula hampers the development and maintenance of highly needed experts in the field in Africa, since they often move to Europe and North America for better career prospects.

Limited Research Funding

A major challenge to research on the African continent is the lack of funding for biomedical research. Current research is mainly funded from international donors, with limited or no funding from national governments and African regional bodies such as the African Union (Hamburg and Collins, 2010; Karikari, 2015). However, a few countries such as South Africa, through the South Africa's National Research Foundation and Medical Research Council, do provide funding for genomics research projects (Karikari et al., 2015). Until the initiation of H3Africa, through funding from the National Institute of Health (United States) and the Wellcome Trust (United Kingdom), there was limited to no funding for genomics and bioinformatics in Africa (Adoga et al., 2014; Mulder et al., 2017).

CONCLUSION AND PERSPECTIVE

Herein we highlight how genomics and bioinformatics has contributed to our understanding of infectious diseases of significant health concern, ranging from bacterial and viral to parasitic infections, as well as their applications to drug and vaccine target identification. This ranges from understanding pathogenesis, host systemic responses and host-pathogen interactions to identification of prognostic and diagnostic markers. However, in Africa, despite the high morbidity and mortality due to infectious diseases, there is limited expertise in the field of bioinformatics and hence limited bioinformatics research output in terms of publications. Thus, there is a need to strengthen training and capacity building in bioinformatics in Africa to improve infectious disease genomics and host-pathogen genomics on the continent. This can be achieved through the establishment of well-structured courses, mentorship for junior

[10] http://bioscience.iita.org/index.php/en/services/bioinformatics
[11] https://www.cgiar.org/

and trainee bioinformaticians and better career prospects to maintain trained bioinformaticians on the continent.

AUTHOR CONTRIBUTIONS

All authors listed contributed substantially to the intellectual, writing and editing, and approved the manuscript for publication.

FUNDING

All authors were supported by a DELTAS Africa grant (DEL-15-007: GA). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency), with funding from the Wellcome Trust (107755/Z/15/Z: GA) and the United Kingdom Government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the United Kingdom Government.

ACKNOWLEDGMENTS

We are grateful to WACCBIP for providing us with the funding and conducive environment to do quality research.

REFERENCES

Aaskov, J., Buzacott, K., Thu, H. M., Lowry, K., and Holmes, E. C. (2006). Long-term transmission of defective RNA viruses in humans and aedes mosquitoes. *Science* 311, 236–238.

Addiss, D. G. (2010). Global elimination of lymphatic filariasis: addressing the public health problem. *PLoS Negl. Trop. Dis.* 4:e741. doi: 10.1371/journal.pntd.0000741

Adoga, M. P., Fatumo, S. A., and Agwale, S. M. (2014). H3Africa: a tipping point for a revolution in bioinformatics, genomics and health research in Africa. *Source Code Biol. Med. BioMed Central* 9:10. doi: 10.1186/1751-0473-9-10

Ahn, I., Jang, J.-H., Han, Y., and Tung, T. Q. (2015). "Phylogenetic analysis of dengue viruses using bioinformatics techniques," in *Paper Presented at the Bioscience and Medical Research 2015*, Cambridge, MA, 12–18.

Almelli, T., Nuel, G., Bischoff, E., Aubouy, A., Elati, M., Wang, C. W., et al. (2014). Differences in gene transcriptomic pattern of *Plasmodium falciparum* in children with cerebral malaria and asymptomatic carriers. *PLoS One* 9:e114401. doi: 10.1371/journal.pone.0114401

Amazigo, U. (2008). The african programme for onchocerciasis control (APOC). *Ann. Trop. Med. Parasitol.* 102(Suppl. 1), 19–22.

Anderson, S. T., Kaforou, M., Brent, A. J., Wright, V. J., Banwell, C. M., Chagaluka, G., et al. (2014). Diagnosis of childhood tuberculosis and host RNA expression in Africa. *N. Engl. J. Med.* 370, 1712–1723. doi: 10.1056/NEJMoa1303657

Antony, H. A., Pathak, V., Parija, S. C., Ghosh, K., and Bhattacherjee, A. (2016). Transcriptomic analysis of chloroquine-sensitive and chloroquine-resistant strains of *Plasmodium falciparum*: toward malaria diagnostics and therapeutics for global health. *OMICS* 20, 424–432. doi: 10.1089/omi.2016.0058

Ariey, F., Witkowski, B., Amaratunga, C., Beghain, J., Langlois, A.-C., Khim, N., et al. (2014). A molecular marker of artemisinin-resistant *Plasmodium falciparum* malaria. *Nature* 505, 50–55. doi: 10.1038/nature12876

Armstrong, S. D., Xia, D., Bah, G. S., Krishna, R., Ngangyung, H. F., LaCourse, E. J., et al. (2016). Stage-specific proteomes from *Onchocerca ochengi*, sister species of the human river blindness parasite, uncover adaptations to a nodular lifestyle. *Mol. Cell. Proteom.* 15, 2554–2575. doi: 10.1074/mcp.M115.055640

Auburn, S., Campino, S., Miotto, O., Djimde, A. A., Zongo, I., Manske, M., et al. (2012). Characterization of within-host *Plasmodium falciparum* diversity using next-generation sequence data. *PLoS One* 7:e32891. doi: 10.1371/journal.pone.0032891

Ballesteros, C., Tritten, L., O'Neill, M., Burkman, E., Zaky, W. I., Xia, J., et al. (2016). The effects of ivermectin on *Brugia malayi* females in vitro: a transcriptomic approach. *PLoS Negl. Trop. Dis*. 10:e0004929. doi: 10.1371/journal.pntd.0004929

Bennuru, S., O'Connell, E. M., Drame, P. M., and Nutman, T. B. (2017). Mining filarial genomes for diagnostic and therapeutic targets. *Trends Parasitol*. 34, 80–90. doi: 10.1016/j.pt.2017.09.003

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2004). GenBank. *Nucleic Acids Res.* 33, D34–D38.

Berry, M. P. R., Graham, C. M., McNab, F. W., Xu, Z., Bloch, S. A. A., Oni, T., et al. (2010). An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 466, 973–977. doi: 10.1038/nature09247

Bhasin, M., and Raghava, G. P. S. (2007). A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J. Biosci.* 32, 31–42.

Bishop, Ö. T., Adebiyi, E. F., Alzohairy, A. M., Everett, D., Ghedira, K., Ghouila, A., et al. (2015). Bioinformatics education-perspectives and challenges out of Africa. *Brief. Bioinform*. 16, 355–364. doi: 10.1093/bib/bbu022

BlueCross and BlueShield Association (2004). Special report: genotyping for cytochrome P450 polymorphisms to determine drug-metabolizer status. *Technol. Eval. Cent. Assess. Program Exec. Summ*. 19, 1–2.

Bopp, S. E. R., Manary, M. J., Bright, A. T., Johnston, G. L., Dharia, N. V., Luna, F. L., et al. (2013). Mitotic evolution of *Plasmodium falciparum* shows a stable core genome but recombination in antigen families. *PLoS Genet.* 9:e1003293. doi: 10.1371/journal.pgen.1003293

Bozdech, Z., Llinás, M., Pulliam, B. L., Wong, E. D., Zhu, J., and DeRisi, J. L. (2003). The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biol.* 1:e5. doi: 10.1371/journal.pbio.0000005

Bradley, P., Gordon, N. C., Walker, T. M., Dunn, L., Heys, S., Huang, B., et al. (2015). Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat. Commun*. 6:10063. doi: 10.1038/ncomms10063

Brody, T., Yavatkar, A. S., Park, D. S., Kuzin, A., Ross, J., and Odenwald, W. F. (2017). Flavivirus and filovirus evoprinters: new alignment tools for the comparative analysis of viral evolution. *PLoS Negl. Trop. Dis.* 11:e0005673. doi: 10.1371/journal.pntd.0005673

Bunnik, E. M., Chung, D.-W. D., Hamilton, M., Ponts, N., Saraf, A., Prudhomme, J., et al. (2013). Polysome profiling reveals translational control of gene expression in the human malaria parasite *Plasmodium falciparum*. *Genome Biol.* 14:R128. doi: 10.1186/gb-2013-14-11-r128

Caro, F., Ahyong, V., Betegon, M., and DeRisi, J. L. (2014). Genome-wide regulatory dynamics of translation in the *Plasmodium falciparum* asexual blood stages. *eLife* 3:e04106. doi: 10.7554/eLife.04106

Cerqueira, G. C., Cheeseman, I. H., Schaffner, S. F., Nair, S., McDew-White, M., Phyo, A. P., et al. (2017). Longitudinal genomic surveillance of *Plasmodium falciparum* malaria parasites reveals complex genomic architecture of emerging artemisinin resistance. *Genome Biol.* 18:78. doi: 10.1186/s13059-017-1204-4

Chesnais, C. B., Sabbagh, A., Pion, S. D., Missamou, F., Garcia, A., and Boussinesq, M. (2016). Familial aggregation and heritability of *Wuchereria bancrofti* infection. *J. Infect. Dis.* 214, 587–594. doi: 10.1093/infdis/jiw212

Chigutsa, E., Visser, M. E., Swart, E. C., Denti, P., Pushpakom, S., Egan, D., et al. (2011). The SLCO1B1 rs4149032 polymorphism is highly prevalent in South Africans and is associated with reduced rifampin concentrations: dosing

implications. *Antimicrob. Agents Chemother.* 55, 4122–4127. doi: 10.1128/AAC.01833-10

Choi, Y.-J., Ghedin, E., Berriman, M., McQuillan, J., Holroyd, N., Mayhew, G. F., et al. (2011). A deep sequencing approach to comparatively analyze the transcriptome of lifecycle stages of the filarial worm, *Brugia malayi*. *PLoS Negl. Trop. Dis.* 5:e1409. doi: 10.1371/journal.pntd.0001409

Choi, Y.-J., Tyagi, R., McNulty, S. N., Rosa, B. A., Ozersky, P., Martin, J., et al. (2016). Genomic diversity in *Onchocerca volvulus* and its *Wolbachia* endosymbiont. *Nat. Microbiol.* 2:16207. doi: 10.1038/nmicrobiol.2016.207

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 393, 537–544.

Coll, F., Mallard, K., Preston, M. D., Bentley, S., Parkhill, J., McNerney, R., et al. (2012). SpolPred: rapid and accurate prediction of *Mycobacterium tuberculosis* spoligotypes from short genomic sequences. *Bioinformatics* 28, 2991–2993. doi: 10.1093/bioinformatics/bts544

Coll, F., McNerney, R., Preston, M. D., Guerra-Assunção, J. A., Warry, A., Hill-Cawthorne, G., et al. (2015). Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. *Genome Med.* 7:51. doi: 10.1186/s13073-015-0164-0

Coll, F., Phelan, J., Hill-Cawthorne, G. A., Nair, M. B., Mallard, K., Ali, S., et al. (2018). Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat. Genet.* 50, 307–316. doi: 10.1038/s41588-017-0029-0

Cotton, J. A., Bennuru, S., Grote, A., Harsha, B., Tracey, A., Beech, R., et al. (2016). The genome of *Onchocerca volvulus*, agent of river blindness. *Nat. Microbiol.* 2:16216. doi: 10.1038/nmicrobiol.2016.216

Cuenco, K. T., Halloran, M. E., Louis-Charles, J., and Lammie, P. J. (2004). A family study of lymphedema of the leg in a lymphatic filariasis-endemic area. *Am. J. Trop. Med. Hyg.* 70, 180–184.

Cully, D. F., Vassilatis, D. K., Liu, K. K., Paress, P. S., Van der Ploeg, L. H. T., Schaeffer, J. M., et al. (1994). Cloning of an avermectin-sensitive glutamate-gated chloride channel from *Caenorhabditis elegans*. *Nature* 371, 707–711.

Curtis, J., Luo, Y., Zenner, H. L., Cuchet-Lourenço, D., Wu, C., Lo, K., et al. (2015). Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. *Nat. Genet.* 47, 523–527. doi: 10.1038/ng.3248

da Silva, P. E. A., Palomino, J. C., Almeida Da Silva, P. E. A., and Palomino, J. C. (2011). Molecular basis and mechanisms of drug resistance in *Mycobacterium tuberculosis*: classical and new drugs. *J. Antimicrob. Chemother.* 66, 1417–1430. doi: 10.1093/jac/dkr173

Daly, A. K., King, B. P., and Leathart, J. B. S. (2006). Genotyping for cytochrome P450 polymorphisms. *Methods Mol. Biol.* 320, 193–207.

Damhoff, H. N., Kuhn, R. J., and Stadler, L. P. (2014). Severe malaria complicated by G6PD deficiency in a pediatric tanzanian immigrant. *J. Pediatr. Pharmacol. Ther.* 19, 325–334. doi: 10.5863/1551-6776-19.4.325

Dang, T. N., Naka, I., Sa-Ngasang, A., Anantapreecha, S., Chanama, S., Wichukchinda, N., et al. (2014). A replication study confirms the association of GWAS-identified SNPs at MICB and PLCE1in Thai patients with dengue shock syndrome. *BMC Med. Genet.* 15:58. doi: 10.1186/1471-2350-15-58

Dapat, C., and Oshitani, H. (2016). Novel insights into human respiratory syncytial virus-host factor interactions through integrated proteomics and transcriptomics analysis. *Expert Rev. Anti Infect. Ther.* 14, 285–297. doi: 10.1586/14787210.2016.1141676

Davey, G., Tekola, F., and Newport, M. J. (2007). Podoconiosis: non-infectious geochemical elephantiasis. *Trans. R. Soc. Trop. Med. Hyg.* 101, 1175–1180. doi: 10.1016/j.trstmh.2007.08.013

Davies, D. H., Duffy, P., Bodmer, J.-L., Felgner, P. L., and Doolan, D. L. (2015). Large screen approaches to identify novel malaria vaccine candidates. *Vaccine* 33, 7496–7505. doi: 10.1016/j.vaccine.2015.09.059

de Borba, L., Villordo, S. M., Iglesias, N. G., Filomatori, C. V., Gebhard, L. G., and Gamarnik, A. V. (2015). Overlapping local and long-range RNA-RNA interactions modulate dengue virus genome cyclization and replication. *J. Virol.* 89, 3430–3437. doi: 10.1128/JVI.02677-14

de Carvalho, C. X., Cardoso, C. C., de Souza Kehdy, F., Pacheco, A. G., and Moraes, M. O. (2017). Host genetics and dengue fever. *Infect. Genet. Evol.* 56, 99–110. doi: 10.1016/j.meegid.2017.11.009

De Groot, A. S., McMurry, J., and Moise, L. (2008). Prediction of immunogenicity: in silico paradigms, ex vivo and in vivo correlates. *Curr. Opin. Pharmacol.* 8, 620–626. doi: 10.1016/j.coph.2008.08.002

Desjardins, C. A., Cerqueira, G. C., Goldberg, J. M., Dunning Hotopp, J. C., Haas, B. J., Zucker, J., et al. (2013). Genomics of Loa loa, a Wolbachia-free filarial parasite of humans. *Nat. Genet.* 45, 495–500. doi: 10.1038/ng.2585

Doolan, D. L., Apte, S. H., and Proietti, C. (2014). Genome-based vaccine design: the promise for malaria and other infectious diseases. *Int. J. Parasitol.* 44, 901–913. doi: 10.1016/j.ijpara.2014.07.010

Doyle, S. R., Bourguinat, C., Nana-Djeunga, H. C., Kengne-Ouafo, J. A., Pion, S. D. S., Bopda, J., et al. (2017). Genome-wide analysis of ivermectin response by *Onchocerca volvulus* reveals that genetic drift and soft selective sweeps contribute to loss of drug sensitivity. *PLoS Negl. Trop. Dis.* 11:e0005816. doi: 10.1371/journal.pntd.0005816

Duffy, C. W., Assefa, S. A., Abugri, J., Amoako, N., Owusu-Agyei, S., Anyorigiya, T., et al. (2015). Comparison of genomic signatures of selection on *Plasmodium falciparum* between different regions of a country with high malaria endemicity. *BMC Genomics* 16:527. doi: 10.1186/s12864-015-1746-3

Duffy, C. W., Ba, H., Assefa, S., Ahouidi, A. D., Deh, Y. B., Tandia, A., et al. (2017). Population genetic structure and adaptation of malaria parasites on the edge of endemic distribution. *Mol. Ecol.* 26, 2880–2894. doi: 10.1111/mec.14066

Faria, N. R., da Costa, A. C., Lourenço, J., Loureiro, P., Lopes, M. E., Ribeiro, R., et al. (2017). Genomic and epidemiological characterisation of a dengue virus outbreak among blood donors in Brazil. *Sci. Rep.* 7:15216. doi: 10.1038/s41598-017-15152-8

Fatumo, S. A., Adoga, M. P., Ojo, O. O., Oluwagbemi, O., Adeoye, T., Ewejobi, I., et al. (2014). Computational biology and bioinformatics in nigeria. *PLoS Comput. Biol.* 10:e1003516. doi: 10.1371/journal.pcbi.1003516

Feuerriegel, S., Schleusener, V., Beckert, P., Kohl, T. A., Miotto, P., Cirillo, D. M., et al. (2015). PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J. Clin. Microbiol.* 53, 1908–1914. doi: 10.1128/JCM.00025-15

Fleri, W., Paul, S., Dhanda, S. K., Mahajan, S., Xu, X., Peters, B., et al. (2017). The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Front. Immunol.* 8:278. doi: 10.3389/fimmu.2017.00278

Ford, C. B., Shah, R. R., Maeda, M. K., Gagneux, S., Murray, M. B., Cohen, T., et al. (2013). *Mycobacterium tuberculosis* mutation rate estimates from different lineages predict substantial differences in the emergence of drug-resistant tuberculosis. *Nat. Genet.* 45, 784–790. doi: 10.1038/ng.2656

Friebe, P., Peña, J., Pohl, M. O. F., and Harris, E. (2012). Composition of the sequence downstream of the dengue virus 5′ cyclization sequence (dCS) affects viral RNA replication. *Virology* 422, 346–356. doi: 10.1016/j.virol.2011.10.025

Gebhard, L. G., Filomatori, C. V., and Gamarnik, A. V. (2011). Functional RNA elements in the dengue virus genome. *Viruses* 3, 1739–1756. doi: 10.3390/v3091739

Gengiah, T., Botha, J., Soowamber, D., and Naidoo, K., Abdool Karim, S. S. (2014). Low rifampicin concentrations in tuberculosis patients with HIV infection. *J. Infect. Dev. Ctries.* 8, 987–993. doi: 10.3855/jidc.4696

Gould, E., and Solomon, T. (2008). Pathogenic flaviviruses. *Lancet* 371, 500–509.

Grubaugh, D., Flechtner, J. B., and Higgins, D. E. (2013). Proteins as T cell antigens: methods for high-throughput identification. *Vaccine* 31, 3805–3810. doi: 10.1016/j.vaccine.2013.06.046

Guerra-Assunção, J., Crampin, A., Houben, R., Mzembe, T., Mallard, K., Coll, F., et al. (2015). Large-scale whole genome sequencing of *M. tuberculosis* provides insights into transmission in a high prevalence area. *eLife* 4:e05166. doi: 10.7554/eLife.05166

Guiliano, D. B., Hong, X., McKerrow, J. H., Blaxter, M. L., Oksov, Y., Liu, J., et al. (2004). A gene family of cathepsin L-like proteases of filarial nematodes are associated with larval molting and cuticle and eggshell remodeling. *Mol. Biochem. Parasitol.* 136, 227–242.

Gurwitz, K. T., Aron, S., Panji, S., Maslamoney, S., Fernandes, P. L., Judge, D. P., et al. (2017). Designing a course model for distance-based online bioinformatics training in Africa: the H3ABioNet experience. *PLoS Comput. Biol.* 13:e1005715. doi: 10.1371/journal.pcbi.1005715

Guy, B., Lang, J., Saville, M., and Jackson, N. (2016). Vaccination against dengue: challenges and current developments. *Annu. Rev. Med*. 67, 387–404. doi: 10.1146/annurev-med-091014-090848

H3Africa Consortium, Rotimi, C., Abayomi, A., Abimiku, A., Adabayeri, V. M., Adebamowo, C., et al. (2014). Research capacity. Enabling the genomic revolution in Africa. *Science* 344, 1346–1348.

Hamburg, M. A., and Collins, F. S. (2010). The path to personalized medicine. *N. Engl. J. Med*. 363, 301–304.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. 32, D258–D261.

He, Y. (2012). Omics-based systems vaccinology for vaccine target identification. *Drug Dev. Res*. 73, 559–568.

Holbrook, M. R. (2017). Historical perspectives on flavivirus research. *Viruses* 9, 1–19.

Holt, K. E., McAdam, P., Thai, P. V. K., Thuong, N. T. T., Ha, D. T. M., Lan, N. N., et al. (2018). Frequent transmission of the *Mycobacterium tuberculosis* Beijing lineage and positive selection for the EsxW Beijing variant in Vietnam. *Nat. Genet*. 50, 849–856. doi: 10.1038/s41588-018-0117-9

Hoo, R., Zhu, L., Amaladoss, A., Mok, S., Natalang, O., Lapp, S. A., et al. (2016). Integrated analysis of the *Plasmodium* species transcriptome. *EBioMedicine* 7, 255–266. doi: 10.1016/j.ebiom.2016.04.011

Jacobsen, M., Repsilber, D., Gutschmidt, A., Neher, A., Feldmann, K., Mollenkopf, H. J., et al. (2007). Candidate biomarkers for discrimination between infection and disease caused by *Mycobacterium tuberculosis*. *J. Mol. Med.* 85, 613–621.

Jamieson, F. B., Teatero, S., Guthrie, J. L., Neemuchwala, A., Fittipaldi, N., and Mehaffy, C. (2014). Whole-genome sequencing of the *Mycobacterium tuberculosis* manila Sublineage results in less clustering and better resolution than Mycobacterial interspersed repetitive-unit-variable-number tandem-repeat (MIRU-VNTR) typing and Spoligotyping. *J. Clin. Microbiol*. 52, 3795–3798. doi: 10.1128/JCM.01726-14

Jespersen, M. C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* 45, W24–W29. doi: 10.1093/nar/gkx346

Junpee, A., Tencomnao, T., Sanprasert, V., and Nuchprayoon, S. (2010). Association between Toll-like receptor 2 (TLR2) polymorphisms and asymptomatic bancroftian filariasis. *Parasitol. Res.* 107, 807–816. doi: 10.1007/s00436-010-1932-9

Kaforou, M., Wright, V. J., Oni, T., French, N., Anderson, S. T., Bangani, N., et al. (2013). Detection of tuberculosis in HIV-infected and -uninfected African adults using whole blood RNA expression signatures: a case-control study. *PLoS Med.* 10:e1001538. doi: 10.1371/journal.pmed.1001538

Kaltdorf, M., Srivastava, M., Gupta, S. K., Liang, C., Binder, J., Dietl, A.-M., et al. (2016). Systematic identification of anti-fungal drug targets by a metabolic network approach. *Front. Mol. Biosci.* 3:22. doi: 10.3389/fmolb.2016.00022

Karikari, T. K. (2015). Bioinformatics in africa: the rise of ghana? *PLoS Comput. Biol.* 11:e1004308. doi: 10.1371/journal.pcbi.1004308

Karikari, T. K., Quansah, E., and Mohamed, W. M. Y. (2015). Developing expertise in bioinformatics for biomedical research in Africa. *Appl. Transl. Genomics* 6, 31–34.

Khor, C. C., Chau, T. N., Pang, J., Davila, S., Long, H. T., Ong, R. T., et al. (2011). Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. *Nat. Genet*. 43, 1139–1141. doi: 10.1038/ng.960

Köser, C. U., Bryant, J. M., Becq, J., Török, M. E., Ellington, M. J., Marti-Renom, M. A., et al. (2013). Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. *N. Engl. J. Med*. 369, 290–292.

Kumar, S., Chaudhary, K., Foster, J. M., Novelli, J. F., Zhang, Y., Wang, S., et al. (2007). Mining predicted essential genes of *Brugia malayi* for nematode drug targets. *PLoS One* 2:e1189. doi: 10.1371/journal.pone.0001189

Kwiatkowski, D. (2015). Malaria genomics: tracking a diverse and evolving parasite population. *Int. Health* 7, 82–84. doi: 10.1093/inthealth/ihv007

Lee, H. J., Walther, M., Georgiadou, A., Nwakanma, D., Stewart, L. B., Levin, M., et al. (2017). Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria. *Sci. Transl. Med*. 10:eaar3619. doi: 10.1126/scitranslmed.aar3619

Leisching, G., Pietersen, R.-D., Mpongoshe, V., van Heerden, C., van Helden, P., Wiid, I., et al. (2016). The host response to a clinical MDR mycobacterial strain cultured in a detergent-free environment: a global transcriptomics approach. *PLoS One* 11:e0153079. doi: 10.1371/journal.pone.0153079

Lesho, E., Forestiero, F. J., Hirata, M. H., Hirata, R. D., Cecon, L., Melo, F. F., et al. (2011). Transcriptional responses of host peripheral blood cells to tuberculosis infection. *Tuberculosis* 91, 390–399. doi: 10.1016/j.tube.2011.07.002

Li, B.-W., Rush, A. C., Mitreva, M., Yin, Y., Spiro, D., Ghedin, E., et al. (2009). Transcriptomes and pathways associated with infectivity, survival and immunogenicity in *Brugia malayi* L3. *BMC Genomics* 10:267. doi: 10.1186/1471-2164-10-267

Li, D., and Aaskov, J. (2014). Sub-genomic RNA of defective interfering (D.I.) dengue viral particles is replicated in the same manner as full length genomes. *Virology* 468–470, 248–255. doi: 10.1016/j.virol.2014.08.013

Li, D., Lott, W. B., Lowry, K., Jones, A., Thu, H. M., and Aaskov, J. (2011). Defective interfering viral particles in acute dengue infections. *PLoS One* 6:e19447. doi: 10.1371/journal.pone.0019447

Llinás, M., Bozdech, Z., Wong, E. D., Adai, A. T., and DeRisi, J. L. (2006). Comparative whole genome transcriptome analysis of three *Plasmodium falciparum* strains. *Nucleic Acids Res.* 34, 1166–1173.

Low, J. G. H., Ooi, E. E., and Vasudevan, S. G. (2017). Current status of dengue therapeutics research and development. *J. Infect. Dis.* 215(Suppl. 2), S96–S102. doi: 10.1093/infdis/jiw423

Ludin, P., Woodcroft, B., Ralph, S. A., and Mäser, P. (2012). In silico prediction of antimalarial drug target candidates. *Int. J. Parasitol.* 2(Suppl. C), 191–199. doi: 10.1016/j.ijpddr.2012.07.002

Lustigman, S., Zhang, J., Liu, J., Oksov, Y., and Hashmi, S. (2004). RNA interference targeting cathepsin L and Z-like cysteine proteases of *Onchocerca volvulus* confirmed their essential function during L3 molting. *Mol. Biochem. Parasitol.* 138, 165–170.

Luzatto, L., Mehta, A., and Vulliamy, T. (2001). "Glucose 6-phosphate dehydrogenase deficiency," in *The Metabolic and Molecular Bases of Inherited Disease,* 8th Edn, eds C. R. Scriver, A. L. Beaudet, W. S. Sly, and D. Valle (New York, NY: McGraw-Hill), 4517–4553.

Luzzatto, L., and Seneca, E. (2014). G6PD deficiency: a classic example of pharmacogenetics with on-going clinical implications. *Br. J. Haematol.* 164, 469–480. doi: 10.1111/bjh.12665

Maertzdorf, J., McEwen, G., Weiner, J. III, Tian, S., Lader, E., Schriek, U., et al. (2015). Concise gene signature for point-of-care classification of tuberculosis. *EMBO Mol. Med*. 8, 86–95. doi: 10.15252/emmm.201505790

Maertzdorf, J., Ota, M., Repsilber, D., Mollenkopf, H. J., Weiner, J., Hill, P. C., et al. (2011). Functional correlations of pathogenesis-driven gene expression signatures in tuberculosis. *PLoS One* 6:e26938. doi: 10.1371/journal.pone.0026938

Malaria Genomic Epidemiology Network (2015). A novel locus of resistance to severe malaria in a region of ancient balancing selection. *Nature* 526, 253–257. doi: 10.1038/nature15390

MalariaGEN Plasmodium falciparum Community Project (2016). Genomic epidemiology of artemisinin resistant malaria. *eLife* 5:e08714. doi: 10.7554/eLife.08714

Manakkadan, A., Joseph, I., Prasanna, R. R., Kunju, R. I., Kailas, L., and Sreekumar, E. (2013). Lineage shift in Indian strains of Dengue virus serotype-3 (Genotype III), evidenced by detection of lineage IV strains in clinical cases from Kerala. *Virol. J.* 10:37. doi: 10.1186/1743-422X-10-37

Manske, M., Miotto, O., Campino, S., Auburn, S., Almagro-Garcia, J., Maslen, G., et al. (2012). Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature* 487, 375–379. doi: 10.1038/nature11174

Masignani, V., Rappuoli, R., and Pizza, M. (2002). Reverse vaccinology: a genome-based approach for vaccine development. *Expert Opin. Biol. Ther*. 2, 895–905.

Mehaffy, C., Guthrie, J. L., Alexander, D. C., Stuart, R., Rea, E., and Jamieson, F. B. (2014). Marked microevolution of a unique *Mycobacterium tuberculosis* strain in 17 years of ongoing transmission in a high risk population. *PLoS One* 9:e112928. doi: 10.1371/journal.pone.0112928

Merker, M., Blin, C., Mona, S., Duforet-Frebourg, N., Lecher, S., Willery, E., et al. (2015). Evolutionary history and global spread of the *Mycobacterium tuberculosis* Beijing lineage. *Nat. Genet*. 47, 242–249. doi: 10.1038/ng.3195

Miles, A., Iqbal, Z., Vauterin, P., Pearson, R., Campino, S., Theron, M., et al. (2016). Indels, structural variation, and recombination drive genomic diversity in *Plasmodium falciparum*. *Genome Res.* 26, 1288–1299. doi: 10.1101/gr.203711.115

Mistry, R., Cliff, J. M., Clayton, C. L., Beyers, N., Mohamed, Y. S., Wilson, P. A., et al. (2007). Gene-expression patterns in whole blood identify subjects at risk for recurrent tuberculosis. *J. Infect. Dis.* 195, 357–365.

Mobegi, V. A., Duffy, C. W., Amambua-Ngwa, A., Loua, K. M., Laman, E., Nwakanma, D. C., et al. (2014). genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in west african populations of differing infection endemicity. *Mol. Biol. Evol.* 31, 1490–1499. doi: 10.1093/molbev/msu106

Molehin, A. J., Gobert, G. N., and McManus, D. P. (2012). Serine protease inhibitors of parasitic helminths. *Parasitology* 139, 681–695. doi: 10.1017/S0031182011002435

Moureau, G., Cook, S., Lemey, P., Nougairede, A., Forrester, N. L., Khasnatinov, M., et al. (2015). New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences. *PLoS One* 10:e0117849. doi: 10.1371/journal.pone.0117849

Mulder, N. J., Adebiyi, E., Adebiyi, M., Adeyemi, S., Ahmed, A., Ahmed, R., et al. (2017). Development of bioinformatics infrastructure for genomics research. *Glob. Heart* 12, 91–98. doi: 10.1016/j.gheart.2017.01.005

Mulder, N. J., Christoffels, A., de Oliveira, T., Gamieldien, J., Hazelhurst, S., Joubert, F., et al. (2016). The development of computational biology in south africa: successes achieved and lessons learnt. *PLoS Comput. Biol.* 12:e1004395. doi: 10.1371/journal.pcbi.1004395

Nag, S., Dalgaard, M. D., Kofoed, P.-E., Ursing, J., Crespo, M., Andersen, L. O., et al. (2017). High throughput resistance profiling of *Plasmodium falciparum* infections based on custom dual indexing and Illumina next generation sequencing-technology. *Sci. Rep.* 7:2398. doi: 10.1038/s41598-017-02724-x

Nana-Djeunga, H., Bourguinat, C., Pion, S. D., Kamgno, J., Gardon, J., and Njiokou, F. (2012). Single nucleotide polymorphisms in beta-tubulin selected in *Onchocerca volvulus* following repeated ivermectin treatment: possible indication of resistance selection. *Mol. Biochem. Parasitol.* 185, 10–18. doi: 10.1016/j.molbiopara.2012.05.005

Nana-Djeunga, H. C., Bourguinat, C., Pion, S. D., Bopda, J., Kengne-Ouafo, J. A., and Njiokou, F. (2014). Reproductive status of *Onchocerca volvulus* after ivermectin treatment in an ivermectin-naïve and a frequently treated population from cameroon. *PLoS Negl. Trop. Dis.* 8:e2824. doi: 10.1371/journal.pntd.0002824

Navale, V., and Bourne, P. E. (2018). Cloud computing applications for biomedical science: a perspective. *PLoS Comput. Biol.* 14:e1006144. doi: 10.1371/journal.pcbi.1006144

O'Neill, M., Ballesteros, C., Tritten, L., Burkman, E., Zaky, W. I., Xia, J., et al. (2016a). Profiling the macrofilaricidal effects of flubendazole on adult female *Brugia malayi* using RNAseq. *Int. J. Parasitol. Drugs Drug Resist.* 6, 288–296. doi: 10.1016/j.ijpddr.2016.09.005

O'Neill, M., Mansour, A., DiCosty, U., Geary, J., Dzimianski, M., McCall, S. D., et al. (2016b). An in vitro/in vivo model to analyze the effects of flubendazole exposure on adult female *Brugia malayi*. *PLoS Negl. Trop. Dis.* 10:e0004698. doi: 10.1371/journal.pntd.0004698

O'Neill, M., Geary, J. F., Agnew, D. W., Mackenzie, C. D., and Geary, T. G. (2015). In vitro flubendazole-induced damage to vital tissues in adult females of the filarial nematode *Brugia malayi*. *Int. J. Parasitol. Drugs Drug Resist.* 5, 135–140. doi: 10.1016/j.ijpddr.2015.06.002

Osei-Atweneboana, M. Y., Boakye, D. A., Awadzi, K., Gyapong, J. O., and Prichard, R. K. (2012). Genotypic analysis of beta-tubulin in *Onchocerca volvulus* from communities and individuals showing poor parasitological response to ivermectin treatment. *Int. J. Parasitol. Drugs Drug Resist.* 2, 20–28. doi: 10.1016/j.ijpddr.2012.01.005

Ottenhoff, T. H. M., Dass, R. H., Yang, N., Zhang, M. M., Wong, H. E. E., Sahiratmadja, E., et al. (2012). Genome-wide expression profiling identifies type 1 interferon response pathways in active tuberculosis. *PLoS One* 7:e45839. doi: 10.1371/journal.pone.0045839

Otto, T. D., Wilinski, D., Assefa, S., Keane, T. M., Sarry, L. R., Böhme, U., et al. (2010). New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq. *Mol. Microbiol.* 76, 12–24. doi: 10.1111/j.1365-2958.2009.07026.x

Panda, A. K., Sahoo, P. K., Kerketta, A. S., Kar, S. K., Ravindran, B., and Satapathy, A. K. (2011). Human lymphatic filariasis: genetic polymorphism of endothelin-1 and tumor necrosis factor receptor II correlates with development of chronic disease. *J. Infect. Dis.* 204, 315–322. doi: 10.1093/infdis/jir258

Parameswaran, P., Charlebois, P., Tellez, Y., Nunez, A., Ryan, E. M., Malboeuf, C. M., et al. (2012). Genome-wide patterns of intrahuman dengue virus diversity reveal associations with viral phylogenetic clade and interhost diversity. *J. Virol.* 86, 8546–8558. doi: 10.1128/JVI.00736-12

Peachey, L. E., Pinchbeck, G. L., Matthews, J. B., Burden, F. A., Lespine, A., von Samson-Himmelstjerna, G., et al. (2017). P-glycoproteins play a role in ivermectin resistance in cyathostomins. *Int. J. Parasitol. Drugs Drug Resist.* 7, 388–398. doi: 10.1016/j.ijpddr.2017.10.006

Price, E. W. (1976). The association of endemic elephantiasis of the lower legs in East Africa with soil derived from volcanic rocks. *Trans. R. Soc. Trop. Med. Hyg.* 70, 288–295.

Proietti, C., and Doolan, D. L. (2015). The case for a rational genome-based vaccine against malaria. *Front. Microbiol.* 5:741. doi: 10.3389/fmicb.2014.00741

Pybus, O. G., Suchard, M. A., Lemey, P., Bernardin, F. J., Rambaut, A., Crawford, F. W., et al. (2012). Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. U.S.A.* 109, 15066–15071. doi: 10.1073/pnas.1206598109

Qi, H., Zhang, Y.-B., Sun, L., Chen, C., Xu, B., Xu, F., et al. (2017). Discovery of susceptibility loci associated with tuberculosis in Han Chinese. *Hum. Mol. Genet.* 26, 4752–4763. doi: 10.1093/hmg/ddx365

Rabbani, B., Nakaoka, H., Akhondzadeh, S., Tekin, M., and Mahdieh, N. (2016). Next generation sequencing: implications in personalized medicine and pharmacogenomics. *Mol. Biosyst.* 12, 1818–1830.

Ramesh, K., Hemanth Kumar, A. K., Kannan, T., Vijayalakshmi, R., Sudha, V., Manohar Nesakumar, S., et al. (2016). SLCO1B1 gene polymorphisms do not influence plasma rifampicin concentrations in a South Indian population. *Int. J. Tuberc. Lung Dis.* 20, 1231–1235. doi: 10.5588/ijtld.15.1007

Rasmussen, D. A., Boni, M. F., and Koelle, K. (2014). Reconciling phylodynamics with epidemiology: the case of dengue virus in southern vietnam. *Mol. Biol. Evol.* 31, 258–271. doi: 10.1093/molbev/mst203

Reche, P., Glutting, J.-P., Zhang, H., and Reinherz, E. (2004). Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles. *Immunogenetics* 56, 405–419.

Reid, A. J., Talman, A. M., Bennett, H. M., Gomes, A. R., Sanders, M. J., Illingworth, C. J. R., et al. (2018). Single-cell RNA-seq reveals hidden transcriptional variation in malaria parasites. *Elife* 7:e33105. doi: 10.7554/eLife.33105

Robinson, T., Campino, S. G., Auburn, S., Assefa, S. A., Polley, S. D., Manske, M., et al. (2011). Drug-resistant genotypes and multi-clonality in *Plasmodium falciparum* analysed by direct genome sequencing from peripheral blood of malaria patients. *PLoS One* 6:e23204. doi: 10.1371/journal.pone.0023204

Saha, S., and Raghava, G. P. S. (2006). Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* 65, 40–48.

Saha, S., and Raghava, G. P. S. (2004). "BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties," in *Proceedings of the 3rd International Conference Artificial Immune Systems, ICARIS 2004*, eds G. Nicosia, V. Cutello, P. J. Bentley, and J. Timis (Berlin: Springer).

Siau, A., Silvie, O., Franetich, J. F., Yalaoui, S., Marinach, C., Hannoun, L., et al. (2008). Temperature shift and host cell contact up-regulate sporozoite expression of *Plasmodium falciparum* genes involved in hepatocyte infection. *PLoS Pathog.* 4:e1000121. doi: 10.1371/journal.ppat.1000121

Sloan, D. J., McCallum, A. D., Schipani, A., Egan, D., Mwandumba, H. C., Ward, S. A., et al. (2017). Genetic determinants of the pharmacokinetic variability of rifampin in malawian adults with pulmonary tuberculosis. *Antimicrob. Agents Chemother.* 61:e00210-17. doi: 10.1128/AAC.00210-17

Small, S. T., Reimer, L. J., Tisch, D. J., King, C. L., Christensen, B. M., Siba, P. M., et al. (2016). Population genomics of the filarial nematode parasite *Wuchereria*

bancrofti from mosquitoes. *Mol. Ecol.* 25, 1465–1477. doi: 10.1111/mec.13574

Sobkowiak, B., Glynn, J. R., Houben, R. M. G. J., Mallard, K., Phelan, J. E., Guerra-Assunção, J. A., et al. (2018). Identifying mixed *Mycobacterium tuberculosis* infections from whole genome sequence data. *BMC Genomics* 19:613. doi: 10.1186/s12864-018-4988-z

Sobota, R. S., Stein, C. M., Kodaman, N., Scheinfeldt, L. B., Maro, I., Wieland-Alter, W., et al. (2016). A locus at 5q33.3 confers resistance to tuberculosis in highly susceptible individuals. *Am. J. Hum. Genet*. 98, 514–524. doi: 10.1016/j.ajhg.2016.01.015

Sorber, L., Zwaenepoel, K., Deschoolmeester, V., Van Schil, P. E. Y., Van Meerbeeck, J., Lardon, F., et al. (2017). Circulating cell-free nucleic acids and platelets as a liquid biopsy in the provision of personalized therapy for lung cancer patients. *Lung Cancer* 107, 100–107. doi: 10.1016/j.lungcan.2016.04.026

Soria-Guerra, R. E., Nieto-Gomez, R., Govea-Alonso, D. O., and Rosales-Mendoza, S. (2015). An overview of bioinformatics tools for epitope prediction: implications on vaccine development. *J. Biomed. Inform.* 53, 405–414. doi: 10.1016/j.jbi.2014.11.003

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Straimer, J., Gnädig, N. F., Witkowski, B., Amaratunga, C., Duru, V., Ramadani, A. P., et al. (2015). Drug resistance. K13-propeller mutations confer artemisinin resistance in *Plasmodium falciparum* clinical isolates. *Science* 347, 428–431. doi: 10.1126/science.1260867

Stucki, D., Ballif, M., Bodmer, T., Coscolla, M., Maurer, A. M., Droz, S., et al. (2015). Tracking a tuberculosis outbreak over 21 years: strain-specific single-nucleotide polymorphism typing combined with targeted whole-genome sequencing. *J. Infect. Dis*. 211, 1306–1316. doi: 10.1093/infdis/jiu601

Sun, G., Luo, T., Yang, C., Dong, X., Li, J., Zhu, Y., et al. (2012). Dynamic population changes in *mycobacterium tuberculosis* during acquisition and fixation of drug resistance in patients. *J. Infect. Dis*. 206, 1724–1733. doi: 10.1093/infdis/jis601

Sveinbjornsson, G., Gudbjartsson, D. F., Halldorsson, B. V., Kristinsson, K. G., Gottfredsson, M., Barrett, J. C., et al. (2016). HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. *Nat. Genet*. 48, 318–322. doi: 10.1038/ng.3498

Tarun, A. S., Peng, X., Dumpit, R. F., Ogata, Y., Silva-Rivera, H., Camargo, N., et al. (2008). A combined transcriptome and proteome survey of malaria parasite liver stages. *Proc. Natl. Acad. Sci. U.S.A.* 105, 305–310. doi: 10.1073/pnas.0710780104

Taylor, M. J. (2003). Wolbachia in the inflammatory pathogenesis of human filariasis. *Ann. N. Y. Acad. Sci.* 990, 444–449.

Taylor, M. J., Hoerauf, A., and Bockarie, M. (2010). Lymphatic filariasis and onchocerciasis. *Lancet* 376, 1175–1185.

Tekola Ayele, F., Adeyemo, A., Finan, C., Hailu, E., Sinnott, P., Burlinson, N. D., et al. (2012). HLA class II locus and susceptibility to podoconiosis. *N. Engl. J. Med.* 366, 1200–1208. doi: 10.1056/NEJMoa1108448

Thai, K. T. D., Henn, M. R., Zody, M. C., Tricou, V., Nguyet, N. M., Charlebois, P., et al. (2012). High-resolution analysis of intrahost genetic diversity in dengue virus serotype 1 infection identifies mixed infections. *J. Virol.* 86, 835–843. doi: 10.1128/JVI.05985-11

Thompson, E. G., Du, Y., Malherbe, S. T., Shankar, S., Braun, J., Valvo, J., et al. (2017). Host blood RNA signatures predict the outcome of tuberculosis treatment. *Tuberculosis* 107, 48–58. doi: 10.1016/j.tube.2017.08.004

Thye, T., Owusu-Dabo, E., Vannberg, F. O., van Crevel, R., Curtis, J., Sahiratmadja, E., et al. (2012). Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat. Genet*. 44, 257–259. doi: 10.1038/ng.1080

Thye, T., Vannberg, F. O., Wong, S. H., Owusu-Dabo, E., Osei, I., Gyapong, J., et al. (2010). Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet*. 42, 739–741. doi: 10.1038/ng.639

Tientcheu, L. D., Maertzdorf, J., Weiner, J., Adetifa, I. M., Mollenkopf, H.-J., Sutherland, J. S., et al. (2015). Differential transcriptomic and metabolic profiles of *M. africanum*- and *M. tuberculosis*-infected patients after, but not before, drug treatment. *Genes Immun.* 16, 347–355. doi: 10.1038/gene.2015.21

Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., et al. (2012). Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature* 489, 443–446. doi: 10.1038/nature11334

Turab Naqvi, A. A., Rahman, S., Rubi., Zeya, F., Kumar, K., Choudhary, H., et al. (2017). Genome analysis of Chlamydia trachomatis for functional characterization of hypothetical proteins to discover novel drug targets. *Int. J. Biol. Macromol.* 96, 234–240. doi: 10.1016/j.ijbiomac.2016.12.045

van Beek, J., Haanperä, M., Smit, P. W., Mentula, S., and Soini, H. (2018). Evaluation of whole genome sequencing and software tools for drug susceptibility testing of *Mycobacterium tuberculosis*. *Clin. Microbiol. Infecti*. doi: 10.1016/j.cmi.2018.03.041 [Epub ahead of print].

Verhagen, L. M., Zomer, A., Maes, M., Villalba, J. A., del Nogal, B., Eleveld, M., et al. (2013). A predictive signature gene set for discriminating active from latent tuberculosis in Warao Amerindian children. *BMC Genomics* 14:74. doi: 10.1186/1471-2164-14-74

Villordo, S. M., Filomatori, C. V., Sánchez-Vargas, I., Blair, C. D., and Gamarnik, A. V. (2015). Dengue virus RNA structure specialization facilitates host adaptation. *PLoS Pathog.*11:e1004604. doi: 10.1371/journal.ppat.1004604

Walker, T. M., Kohl, T. A., Omar, S. V., Hedge, J., Del Ojo Elias, C., Bradley, P., et al. (2015). Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study. *Lancet Infect. Dis*. 15, 1193–1202.

Waman, V. P., Kolekar, P., Ramtirthkar, M. R., Kale, M. M., and Kulkarni-Kale, U. (2016). Analysis of genotype diversity and evolution of Dengue virus serotype 2 using complete genomes. *PeerJ* 4:e2326. doi: 10.7717/peerj.2326

Weiner, M., Peloquin, C., Burman, W., Luo, C.-C., Engle, M., Prihoda, T. J., et al. (2010). Effects of tuberculosis, race, and human gene SLCO1B1 polymorphisms on rifampin concentrations. *Antimicrob. Agents Chemother.* 54, 4192–4200. doi: 10.1128/AAC.00353-10

Whitehorn, J., Chau, T. N. B., Nguyet, N. M., Kien, D. T. H., Quyen, N. T. H., Trung, D. T., et al. (2013). Genetic variants of MICB and PLCE1 and associations with non-severe dengue. *PLoS One* 8:e59067. doi: 10.1371/journal.pone.0059067

WHO (2017). *WHO Dengue and Severe Dengue*. Geneva: World Health Organization.

WHO (2018a). *Global Tuberculosis Report 2018*. Geneva: World Health Organization.

WHO (2018b). *World Malaria Day 2018: Ready to Beat Malaria*. Geneva: World Health Organization, 1–3.

WHO/Department of Control of Neglected Tropical Diseases (2016). Global programme to eliminate lymphatic filariasis: progress report, 2015. *Wkly. Epidemiol. Rec.* 39, 441–460.

World Health Organization Western Pacific Region (2018). Dengue Situation Update Number 500: Update on the Dengue situation in the Western Pacific Region (Northern Hemisphere)., (500), pp. 1–5*.

Xia, X. (2017). Bioinformatics and drug discovery. *Curr. Top. Med. Chem.* 17, 1709–1726.

Yamagishi, J., Natori, A., Tolba, M. E. M., Mongan, A. E., Sugimoto, C., Katayama, T., et al. (2014). Interactive transcriptome analysis of malaria patients and infecting *Plasmodium falciparum*. *Genome Res.* 24, 1433–1444. doi: 10.1101/gr.158980.113

Yan, S.-K., Liu, R.-H., Jin, H.-Z., Liu, X.-R., Ye, J., Shan, L., et al. (2015). 'Omics' in pharmaceutical research: overview, applications, challenges, and future perspectives. *Chin. J. Nat. Med.* 13, 3–21.

Young, J. A., Fivelman, Q. L., Blair, P. L., de la Vega, P., Le Roch, K. G., Zhou, Y., et al. (2005). The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification. *Mol. Biochem. Parasitol.* 143, 67–79.

Yun, M.-R., Han, S. J., Yoo, W. G., Kwon, T., Lee, S., Lee, J. S., et al. (2016). Draft genome sequence of *Mycobacterium tuberculosis* KT-0204, isolated in South Korea. *Genome Announc.* 4:e01519-15. doi: 10.1128/genomeA.01519-15

Zak, D. E., Penn-Nicholson, A., Scriba, T. J., Thompson, E., Suliman, S., Amon, L. M., et al. (2016). A blood RNA signature for tuberculosis disease risk: a prospective cohort study. *Lancet* 387, 2312–2322. doi: 10.1016/S0140-6736(15)01316-1

Zhang, H., Li, D., Zhao, L., Fleming, J., Lin, N., Wang, T., et al. (2013). Genome sequencing of 161 *Mycobacterium tuberculosis* isolates from China identifies genes and intergenic regions associated with drug resistance. *Nat. Genet.* 45, 1255–1260. doi: 10.1038/ng.2735

Zhu, L., Mok, S., Imwong, M., Jaidee, A., Russell, B., Nosten, F., et al. (2016). New insights into the *Plasmodium vivax* transcriptome using RNA-Seq. *Sci. Rep.* 6:20498. doi: 10.1038/srep20498

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 28 November 2018
doi: 10.3389/fpls.2018.01734

A Pipeline for Classifying Deleterious Coding Mutations in Agricultural Plants

Maxim S. Kovalev[1], Anna A. Igolkina[1], Maria G. Samsonova[1]* and Sergey V. Nuzhdin[1,2]*

[1] *Department of Applied Mathematics, Peter the Great St.Petersburg Polytechnic University, St. Petersburg, Russia,*
[2] *Program Molecular & Computational Biology, Dornsife College of Letters Arts and Science, University of Southern California, Los Angeles, CA, United States*

The impact of deleterious variation on both plant fitness and crop productivity is not completely understood and is a hot topic of debates. The deleterious mutations in plants have been solely predicted using sequence conservation methods rather than function-based classifiers due to lack of well-annotated mutational datasets in these organisms. Here, we developed a machine learning classifier based on a dataset of deleterious and neutral mutations in *Arabidopsis thaliana* by extracting 18 informative features that discriminate deleterious mutations from neutral, including 9 novel features not used in previous studies. We examined linear SVM, Gaussian SVM, and Random Forest classifiers, with the latter performing best. Random Forest classifiers exhibited a markedly higher accuracy than the popular PolyPhen-2 tool in the *Arabidopsis* dataset. Additionally, we tested whether the Random Forest, trained on the *Arabidopsis* dataset, accurately predicts deleterious mutations in *Orýza sativa* and *Pisum sativum* and observed satisfactory levels of performance accuracy (87% and 93%, respectively) higher than obtained by the PolyPhen-2. Application of Transfer learning in classifiers did not improve their performance. To additionally test the performance of the Random Forest classifier across different angiosperm species, we applied it to annotate deleterious mutations in *Cicer arietinum* and validated them using population frequency data. Overall, we devised a classifier with the potential to improve the annotation of putative functional mutations in QTL and GWAS hit regions, as well as for the evolutionary analysis of proliferation of deleterious mutations during plant domestication; thus optimizing breeding improvement and development of new cultivars.

Keywords: deleterious mutation, random forest (bagging) and machine learning, *Orýza*, *Pisum*, *Cicer*

OPEN ACCESS

Edited by:
Yuriy L. Orlov,
Institute of Cytology and Genetics (RAS), Russia

Reviewed by:
Vasily Ramensky,
Moscow Institute of Physics and Technology, Russia
Konstantin Vladimirovich Gunbin,
Institute of Cytology and Genetics (RAS), Russia

****Correspondence:***
Anna A. Igolkina
igolkinaanna11@gmail.com
Maria G. Samsonova
m.g.samsonova@gmail.com

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Plant Science

Received: *18 September 2018*
Accepted: *08 November 2018*
Published: *28 November 2018*

Citation:
Kovalev MS, Igolkina AA, Samsonova MG and Nuzhdin SV (2018) A Pipeline for Classifying Deleterious Coding Mutations in Agricultural Plants. Front. Plant Sci. 9:1734. doi: 10.3389/fpls.2018.01734

INTRODUCTION

New mutations continuously arise in populations. Some of them are neutral, but many are deleterious (Grossman et al., 2010). Under most circumstances, natural selection is effective in maintaining strong deleterious mutations at low level, however mildly deleterious variants may reach considerable frequency in populations due to hitchhiking and population bottlenecks. Deleterious variants may affect phenotypic traits and decrease organismal fitness. Quite the opposite, in maize intermediate and weakly deleterious alleles are involved in heterosis

(Yang et al., 2017). In human rare, deleterious SNPs are associated with common diseases and cancer (Taylor et al., 2015). Therefore, it is no wonder that estimation of the deleterious mutations prevalence in different species is a topic of vivid interests.

Theoretical predictions place the fraction of deleterious mutations in barley, soybean, rice, maize, and *Arabidopsis* genomes from 20% to 40% approximately (Günther and Schmid, 2010; Mezmouk and Ross-Ibarra, 2014; Kono et al., 2016). Deleterious alleles are usually at low frequency, an observation that is in agreement with the action of weak purifying selection. The prevalence of deleterious alleles differs between wild species, landraces, and elite cultivars. Using rice sequences Günther and Schmid (2010) found fewer deleterious substitutions in the wild than in cultivated rice. In comparisons with traditional landraces, elite maize inbreds show an increase in the proportion of deleterious variants fixed within the population, but the much smaller proportion of segregating deleterious variants (Yang et al., 2017). This is explained by bottlenecks during modern breeding that results in fixation of the majority of mutations, therefore reducing a fraction of segregating variation.

The issue of deleterious variation in plant genotypes is particularly essential for crop improvement, because crop productivity may be reduced due to a persistence of deleterious variants at a moderate frequency. Indeed Yang et al. (2017) found that deleterious variants may contribute substantially to variation in fitness-related quantitative traits in maize and that incorporation of information about deleterious mutations may improve existing genomic prediction frameworks.

NGS technologies open a way to annotate the functional effect of individual SNPs. As the regulatory code responsible for gene activity still remains a puzzle, only genetic variants in the coding regions are considered. The general belief is that non-synonymous substitutions may change protein structure and therefore many of them should have the deleterious effect on protein function, which in turn manifests as biochemical or morphological mutations. The methods for prediction of deleterious effects of non-synonymous substitutions in proteins could be subdivided into two groups. The first group methods exploit sequence conservation and are based on the assumptions that SNPs in evolutionarily conserved regions are likely to be deleterious. Some of them like SIFT use simple cut-off to discriminate deleterious variants from neutral (Sim et al., 2012), while other like MAPP (Stone and Sidow, 2005) and GERP+++ (Davydov et al., 2010) employ phylogenetic information in addition.

The machine learning algorithms lay the foundation of the second group methods. Of these the most widely used is PolyPhen-2 (Adzhubei et al., 2010). This method employs the rigorously annotated datasets of human disease-causing mutations for training that preconditions its high predictive accuracy. As a machine learning method PolyPhen-2 consists of three steps: firstly a set of features that characterize a mutation was extracted using sequence characteristics, multiple alignment scores, and information about the 3D structure of the resulting protein. At the next steps, training and cross-validation were performed followed by classification with a naïve Bayes approach. It should be noted, that being trained on human data, PolyPhen-2 is sometimes applied to predict deleterious mutations in other species. There is, however, little consensus about the eligibility of such a direct knowledge transfer. Indeed, it is known that alleles annotated as deleterious in humans at about 15% of cases correspond to normal alleles in other mammals (Kondrashov et al., 2002). It appears from this that to achieve more accurate predictions training might have to be separately executed species by species. However, for many species, information required for classifier training might be substantially more limited than for humans. Accordingly, the question arises whether it is possible to use the information obtained for one species for the search for harmful mutations in another, perhaps phylogenetically close, species.

This question has long been discussed in machine learning in the following formulation: how to transfer knowledge from one object to another, considered to be close (in the sense of data sampling distribution), to solve a specific problem (whether classification or regression). A set of methods that provide the methodology for solving such problems is denoted Transfer Learning (TL). These methods have found broad application in many practical problems. For instance, Lagunas and Garces (2017) classify the painted images of various objects using their naturalistic form (photos). Closer to home, Transfer Learning was used for evaluating the quality of protein models (Hurtado et al., 2018), the localization of proteins in the cell based on ontology databases (Mei et al., 2011) and the search for associations between the genome and the phenotype (Petegrosso et al., 2018).

Up to now, most publications predicting deleterious mutations in plants use sequence conservation methods that is mostly due to lack of well-annotated datasets of deleterious and neutral mutations in these organisms. However, recently, Kono et al. (2016) have assembled a validated database of 2,910 function-altering mutations in *Arabidopsis* that opens the way for development of machine learning methods specifically tailored for plants. Here, we developed the Random Forest classifier that being tested on two plant species – *Orýza sativa* and *Pisum sativum* – for which the sufficient number of neutral and functional mutations are known – showed substantially better performance than PolyPhen-2. We also attempted to improve our classifier using the approaches of Transfer learning, as this technique could provide knowledge transfer from one species for which a lot of information is available to a close species with limited information. Finally, we validate this classifier using population data on single nucleotide allele frequency available for *Cicer arietinum* (Plekhanova et al., 2017). We believe our classifier will be helpful in plant research for prioritizing mutations in QTL and GWAS support intervals for functional validation, for developing GRN-based models to solve the genotype-to-phenotype problem, as well as for improvement of breeding programs.

MATERIALS AND METHODS

Arabidopsis Training Database

The list of amino acid substitutions in *Arabidopsis thaliana* proteins was obtained from the database created by Kono et al. (2018). The database consists of 13,707 replacements available, of them 4,409 were labeled mutations in 994 proteins: 2,894 deleterious and 1,515 neutral. The protein sequences were downloaded from "The *Arabidopsis* Information Resource."

Oryza sativa and *Pisum sativum* Test Datasets

The sets of deleterious mutations in rice (*O. sativa*) and pea (*P. sativum*) were extracted from the UniProt mutation database (The UniProt Consortium, 2017). To construct a set of neutral mutations in rice and pea BLASTp program (Altschul et al., 1997) was used to align each protein sequence against SwissProt sequence database (Bairoch, 1996) and proteins with more than 95% identity to a query sequence were selected. At the next step, the selected sequences were multiply aligned with Clustal Omega (Sievers and Higgins, 2014) and a set of neutral mutations was generated under the following rule. We consider amino acid substitutions without any known phenotype, not present in a continuous block of substituted residues (i.e., are isolated) and independent (i.e., there were no other substitutions in the same sequences of alignment). This rule makes it possible to avoid the phenomenon of correlated mutational behavior between columns in multiple sequence alignment (Kowarsch et al., 2010). Besides we consider only alignment columns that have no more than one substitution. To balance the datasets, neutral mutations were randomly downsampled so that their number was equal to the number of deleterious mutations. Overall, the dataset for rice contained 764 mutations in 400 proteins (by 382 deleterious and neutral); the pea dataset contained 136 mutations in 60 proteins (by 68 deleterious and neutral).

Cicer arietinum Target Dataset

433 *Cicer arietimum* landraces from N. I. Vavilov All-Russian Institute for Genetic Resources (VIR collection) were genotyped by GBS sequencing and variants were called and filtered following standard criteria; overall 56855 SNPs were identified (Plekhanova et al., 2017). Identification of SNPs in protein coding regions and classification of those into synonymous and non-synonymous classes was done with SnpEff tools (Cingolani et al., 2012): 3023 synonymous and 3467 non-synonymous replacements were determines within 2569 proteins.

Classifier Features

The set of classification features was aggregated by different methods. To extract a set of features characterizing substitutions, the PolyPhen-2 web service (Adzhubei et al., 2010) was used. Additional servers and sources of information were also involved, such as the PfamScan (Finn et al., 2014) and the PCI-SS (Green et al., 2009). The former was used to check whether the amino acid substitution locates within a protein domain of the Pfam database. Features obtained with the latter service incorporate information about the secondary structure of the protein in the loci of the substitution. Since information about the three-dimensional structure of a target protein is not always known, these features played the role of alternative structural characteristics. PCI-SS server indicates a protein secondary structure – α-helix, β-sheet, or non-regular structure – which contains the substitution of interest, and also provides three quantitative characteristics about the structural state of the target amino acid in the protein based on the mean-square error between the models considered in the PCI-SS algorithms. To evaluate the physicochemical nature of amino acid substitutions, several measures were used: the Grantham distance (Grantham, 1974), the Sneath index (Sneath, 1966), the Epstein's coefficient of difference (Epstein, 1967), and the Miyata distance (Miyata et al., 1979). The quantitative evaluation of the amino acid substitution by the matrix of BLOSUM62 substitutions was added as an extra feature (Henikoff and Henikoff, 1992).

Two additional features have been constructed that take into account the amino acid context around the mutation position. The first feature was defined as the mean distance over the Grantham matrix between the wild-type amino acid in the mutation position and each of the two neighboring amino acids. The second feature was calculated in the same way but considering two amino acids from a mutant position at a distance of one. The construction of these features was based on the following hypothesis: if the amino acids that are very different in their physicochemical properties are next to each other, this is most likely justified by the constraints on functions to be performed. Therefore, the more physicochemical differences are in the amino acid position from its context, the more likely it is for the mutation in the position of this amino acid to be harmful.

Classifiers

To solve the classification problem of mutations to deleterious versus neutral, three classifiers were tested: Support Vector Machines with a linear kernel (Linear SVM), Support Vector Machines with a Gaussian kernel (Gaussian SVM) (Cristianini and Shawe-Taylor, 2000), and Random Forest (RF) (Breiman, 2001). The Linear SVM method is based on the search for a separating hyperplane with the maximum gap between the data. To use a non-linear separation of classes, the Gaussian SVM was examined; it utilizes the Gaussian kernel instead of the scalar product in the Linear SVM (Cristianini and Shawe-Taylor, 2000). The RF uses the ideas of bagging, or Bootstrap Aggregating (a composition of independent classifiers, in this case, of decision trees) and the method of random subspaces (description of objects using subspaces of the feature space) (Breiman, 2001).

The choice of hyperparameter values for classifiers was carried out on the *Arabidopsis* dataset. For each classifier, the traditional procedure – grid search with fivefold cross-validation – was performed to find the optimal values of hyperparameters. These values are usually selected as the values that provide the highest cross-validation score that leads to the preventing of overfitting. Further, the optimal hyperparameters were utilized while classifiers' training. One might see that the overfitting effect

was not observed (**Supplementary Figure S1**). Cross-validation was performed with tools from the scikit-learn Python module[1].

The accuracy was chosen as the characteristic by which the best values of hyperparameters were selected, as calculated by the following formula: Accuracy = (TP + TN)/N, where N is the sample size for which the classification was made, and TP and TN are the numbers of correctly defined deleterious mutations and neutral ones, respectively. To select the best classifier, the data for *A. thaliana* were divided into training and validation sets (3409 and 1000 samples, respectively). Classifiers were first trained, and then the classification on the validation set was performed. We used Linear SVM, Gaussian SVM, and RF methods from scikit-learn Python module (see footnote 1); the pipeline for tuning, training and testing the classifiers is available at the GitHub repository https://github.com/kovmax/DelMut.

Transfer Learning

The transfer learning (TL) is a machine learning technique that improves a model trained on the target data by transfer knowledge from the related and usually larger source data (Pan and Yang, 2010). In our study, we applied TL for training classifiers to predict deleterious mutations in rice and pea datasets (target data) based on the knowledge about deleterious mutations in *A. thaliana* dataset (source data). We examined the Transductive Transfer Learning which assumes that the source data is labeled (classes of samples are known) but the target data is not and, accordingly, labels for the target data were not used until final validation of the predictions. To implement Transductive TL we assign a weight (W) for each sample from the source data, which inversely depends on the distance in the feature space from this sample to the mean of the target data domain:

$$\mathrm{W} = \exp\left(-||x_i^S - m^t||^2\right)$$

where x_i^S is i-th sample from the source data, m^t represents mean values of the target dataset features (Pan and Yang, 2010; Lapin et al., 2014). The Transductive TL classifier predicts classes of the target dataset and learns on the weighted source data: the closer a sample form the source data to the target dataset, the more significant it is for training. We applied the Transductive TL technique to Linear SVM, Gaussian SVM, and RF classifiers with hyperparameter values estimated for these classifiers without TL. Methods were implemented with tools of scikit-learn Python module (see footnote 1); all datasets and scripts are available at the GitHub repository https://github.com/kovmax/DelMut.

RESULTS

Feature Extraction

To develop a method for predicting damaging missense mutations in plants we use machine learning approach and three annotated datasets of non-synonymous deleterious and neutral mutations in *A. thaliana*, *O. sativa*, and *P. sativum* (see Materials and Methods). The method employs classification algorithms and therefore we need to characterize the datasets with a set of features able to discriminate classes. In total, 18 features were selected characterizing the impact of substitution of the wild-type allele by mutant allele on protein sequence and structure. As **Figure 1** shows the distributions of all the features differ between subsets of neutral and deleterious mutations in *A. thaliana* that points on their utility for discrimination between these subsets.

Best Classifier for the *Arabidopsis thaliana* Dataset

The dataset was divided into training and test samples. The test sample was randomly determined, containing 357 neutral and 643 deleterious mutations, and was used to compare the accuracy of the predictions of the four classifiers (PolyPhen-2, Linear SVM, Gaussian SVM, and Random Forest). The results (see **Table 1**) showed that all the classifiers — Linear SVM, Gaussian SVM, and Random Forest — were more accurate than Polyphen-2, and the most accurate one was Random Forest, it had the highest accuracy and AUC values (ROC-curves are presented in **Supplementary Figures S2–S4**) and the lowest False Negative and False Positive Rates.

Classification of *Orýza sativa* and *Pisum sativum* With and Without Transfer Learning

Each classifier was trained on *Arabidopsis* training samples and applied for prediction in two settings: direct prediction or prediction additionally involving Transfer Learning. Since there is an element of randomization in the Random Forest classification method, estimates for this method were obtained by choosing the best prediction of 300 trained classifiers (**Figure 2**). By comparing the predicted and annotated class values for the rice and pea mutations, we concluded that the best of the proposed classifiers is Random Forest without the addition of Transfer Learning (**Table 2**). Predictions of PolyPhen-2 were better only by the criterion False Positive rate, but by the criterion False Negative Rate was significantly underperforming. Overall the Random Forest classifier makes fewer errors in the predictions of a truly deleterious mutation. The prediction of classifiers in the modes without and with Transfer Learning did not exhibit significant differences. Moreover, for the best Random Forest classifier the mode with Transfer Learning turned out to be less accurate.

Classification of Non-synonymous Mutations in *Cicer arietinum*

To test whether or not our classifiers reasonably perform across different angiosperm species, we chose to annotate deleterious mutations in chickpea, *C. arietinum*. Classification has been pursued with both PolyPhen-2 and the Random Forest classifier demonstrated the best discriminating ability on rice and pea datasets (see **Figure 2**). One may observe (**Table 3**) that there is a general correspondence between annotations, with 1923 designated as neutral and 851 as deleterious by both classifiers. However, there were also appreciable differences, as may be

[1]http://scikit-learn.org

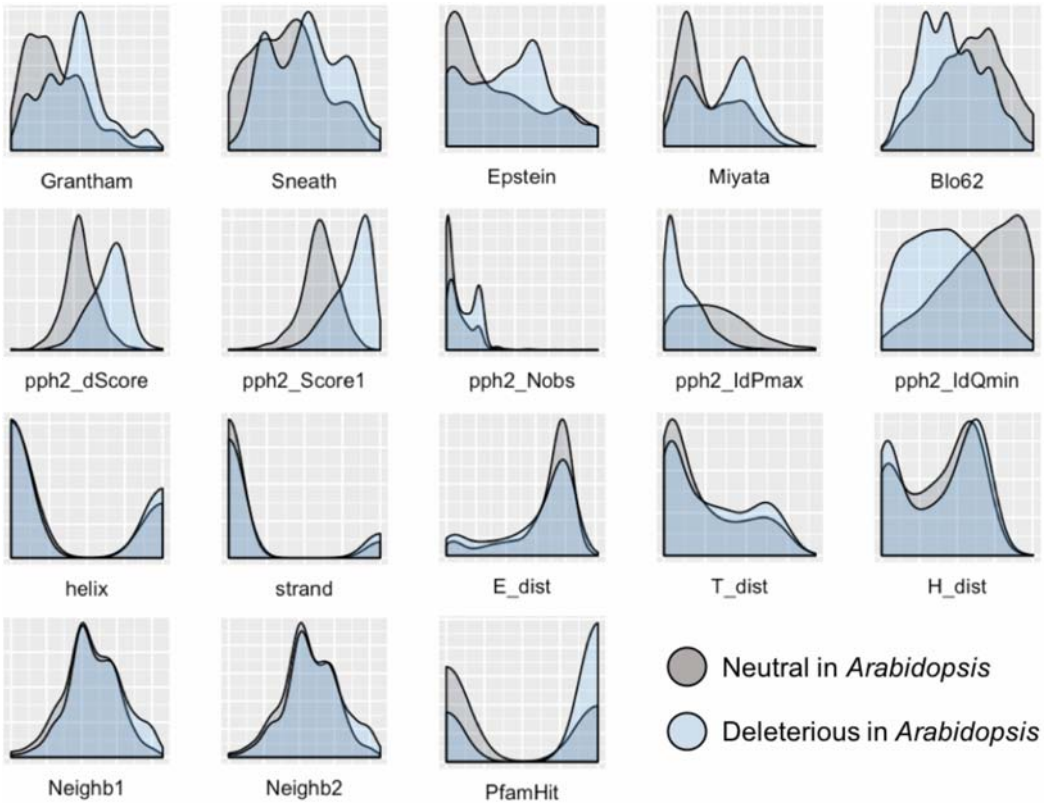


FIGURE 1 | Distribution of features used to characterize the impact of amino acid substitutions in protein sequence for subsets of neutral and deleterious mutations in *Arabidopsis thaliana*. The first row of features – Grantham, Sneath, Epstein, Miyata, and Blo62 (BLOSUM62) – represents distributions of substitution scores based on five corresponding distance matrices. The second row represents the scores obtained with the PolyPhen-2 service: pph2_Score1 and pph2_dScore reflect PSIC scores; pph2_IdPmax, pph2_IdQmin, and pph2_Nobs represent specific features based on the multiple protein alignments. The third row contains features of the secondary protein structure: two features of belonging to helix or strand (helix, strand), and three scores obtained with PCI-SS service (E_dist, T_dist, H_dist). The last row includes two features of the amino acid context around the substitution of interest (Neighb1, Neighb2) and belonging to known Pfam domains (PfamHit). The detailed explanation of features are presented in the **Supplementary Table S1**.

TABLE 1 | Performance of four classifiers: PolyPhen2, Linear SVM, Gaussian SVM and Random Forest on the *Arabidopsis thaliana* dataset.

		PolyPhen-2 (PPh2)		Linear SVM (lSVM)		Gaussian SVM (gSVM)		Random Forest (RF)	
		Neutral	**Deleterious**	**Neutral**	**Deleterious**	**Neutral**	**Deleterious**	**Neutral**	**Deleterious**
Actual classes	**Neutral**	293	64	296	61	301	56	306	51
	Deleterious	100	543	70	573	74	569	60	583
Accuracy		0.836		0.869		0.870		0.889	
False Positive Rate (FPR)		0.179		0.171		0.157		0.143	
False Negative Rate (FNR)		0.156		0.109		0.115		0.093	
Sensitivity		0.844		0.891		0.885		0.907	
Specificity		0.821		0.829		0.843		0.857	
AUC		0.907		0.937		0.935		0.952	

observed by alternative classifications for 517 mutations. Overall, concordance between two classification results was 84.3%.

Due to the lack of annotated missense mutations in chickpea only circumstantial evidence could be used to demonstrate the validity of predictions in this species. To this end, we analyzed the population frequencies of classified polymorphisms in the dataset of 433 chickpea accessions (see Material and Methods). We have calculated the frequencies of synonymous (that are mostly neutral), predicted neutral and predicted deleterious mutations. Due to a large number of missed data, only those genome positions that were called in at least 300 accessions were retained for analysis. Overall, there were 1028 non-synonymous (672 neutral and 356 deleterious) and 901 synonymous polymorphisms (**Table 4**).

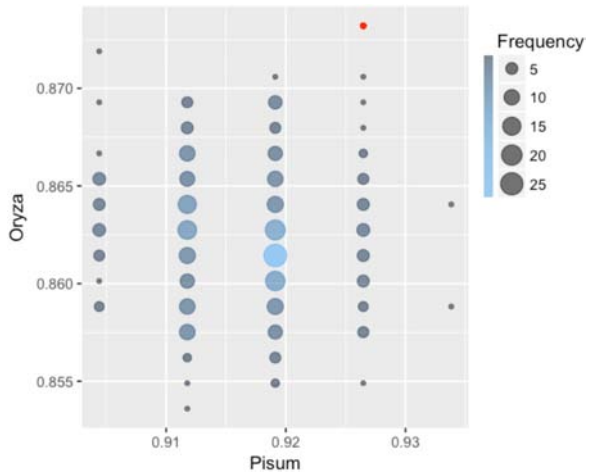


FIGURE 2 | Classification accuracy of 300 Random Forest classifiers learned on the *Arabidopsis thaliana* dataset and applied to classify mutations in pea and rice. Some of the 300 classifiers demonstrated the same values of accuracy on both *Orýza sativa* and *Pisum sativum*. Size and color of circles show frequencies of the classifiers with the same performance. The accuracy value for the best classifier is emphasized with red color.

TABLE 3 | Comparison of the number of deleterious and neutral mutation predicted by PolyPhen-2 and Random Forest classifier in *Cicer arietinum*.

		Random forest	
		Neutral	**Deleterious**
PolyPhen-2	**Neutral**	1923	239
	Deleterious	278	851

TABLE 4 | Mean ffrequencies of non-synonymous deleterious and neutral mutations, as well as synonymous mutations in chickpea dataset.

	Mean frequency
Deleterious	0.050
Neutral	0.097
Synonymous	0.109

TABLE 5 | Results of the Wilcoxon rank sum test for mutation frequencies comparison.

	Neutral	**Synonymous**
Deleterious	0.036 (<0.05)	0.003 (<0.05)
Neutral		0.279 (>0.05)

Applying the Wilcoxon rank sum test with continuity correction, we showed that there was no statistically significant difference between frequencies of neutral and synonymous substitutions; however, the frequency of deleterious mutations is statistically significantly lower than the frequency of mutations from other classes (one sided test, $P < 0.05$) (**Table 5**). These results are fully consistent with previous studies on deleterious mutations in other species (Günther and Schmid, 2010; Mezmouk and Ross-Ibarra, 2014) and could be explained by the action of weak purifying selection that sweeps deleterious mutations away. We conclude that our classifier appears to be working across a broad range of angiosperm species.

DISCUSSION

Here we aimed to develop a classifier specifically tailored for plant datasets that classifies coding non-synonymous mutations into neutral versus functionally deleterious. We have trained the Random Forest classifier in the deleterious mutations in *A. thaliana* using 18 selected features and accomplished a substantially better performance than PolyPhen-2 for two plant species – *O. sativa* and *P. sativum* – for which the sufficient number of neutral and functional mutations is known. The accuracy of our classifier based on Random Forest approach versus PolyPhen-2 was 87% versus 81% for rice and 93% versus 90% for pea. The new classifier also exhibited the superior balance of type I versus type II errors.

We also attempted to improve our classifier using the approaches of Transfer Learning (TL). This has been justified by the following considerations. The task of calling mutation as neutral and deleterious can be set as a classification problem and solved by various methods of machine learning. In mammals, it appeared that the same nucleotide might be

TABLE 2 | Testing classifiers learned on *Arabidopsis* dataset to discriminate deleterious and neutral mutations in rice and pea.

	Orýza sativa				***Pisum sativum***			
	Accuracy	**FPR**	**FNR**	**AUC**	**Accuracy**	**FPR**	**FNR**	**AUC**
PPh2	0.814	0.102	0.270	0.855	0.897	0.044	0.162	0.975
lSVM	0.848	0.144	0.160	0.918	0.912	0.103	0.074	0.971
gSVM	0.842	0.164	0.152	0.890	0.912	0.088	0.088	0.955
RF	0.873	0.115	0.139	0.928	0.926	0.074	0.074	0.981
lSVM + TL	0.848	0.144	0.160	0.918	0.912	0.103	0.074	0.971
gSVM + TL	0.803	0.285	0.110	0.902	0.904	0.147	0.044	0.960
RF + TL	0.861	0.128	0.149	0.926	0.919	0.088	0.074	0.979

PPh2, PolyPhen-2; lSVM, linear SVM; gSVM, Gaussian SVM; RF, random forest; TL, transfer learning.

deleterious in one species but neutral in another (Kondrashov et al., 2002). Accordingly, training might have to be separately executed species by species. TL appears to be a suitable methodology to implement species-specific training as it could provide knowledge transfer from one species for which à lot of information is available to a close species with limited information. However, here we failed to improve the classifier performance with TL. In fact, the performance of our best Random Forest-based classifier dropped between 1% and 2% for both species, *O. sativa* and *P. sativum*. The reason why TL does not improve classifier performance is not clear. There might be unknown technical reasons, but also some biological considerations. It is known, for instance, that alleles annotated as deleterious in humans at about 15% of cases correspond to normal alleles in other mammals (Kondrashov et al., 2002). Which is to say, as GRNs and proteins diverge between species, the functional importance of different amino acids may also diverge. This might partially be explained by a highly epistatic landscape of amino acid substitutions, as best documented for green fluorescence protein (Sarkisyan et al., 2016). When species with diverged GRNs and proteins mate, their progeny suffer from F1 incompatibility and F2 hybrid breakdown because of epistatic incompatibilities (Turelli and Orr, 2000; Rieseberg and Willis, 2007; Coyne, 2016). It is rather interesting to note that the hybrids between different angiosperm species are much more frequently viable, even at higher phylogenetic distances, than mammals are. In fact, rather than suffering from incompatibilities, plant hybrids may exhibit remarkable hybrid vigor (Garcia et al., 2008; Charlesworth and Willis, 2009) raising a question whether the patterns of GRN and protein divergence in plants are functionally equivalent to those in mammals. It might imply that amino acids substitutions in plant proteins and GRNs are less epistatic, which is to say whether an amino acid substitution is deleterious or not could only weekly change between angiosperm species, unlike mammals. If so, then TL should result in substantial improvements when applied to mammals but not angiosperms. Of course, at this moment, this consideration is nothing more than speculation, but the one deserving attention and specially designed analysis to try the TL methodology in mammals.

While somewhat disappointing, that the classifier works well for different species without the need for species-specific learning also has positive aspects – the classifier does not have to be retrained before applying across angiosperms. To test whether our classifier would work with a new species, we utilized the data on population polymorphisms available for *C. arietinum*. Our hypothesis was that if we annotate these chickpea polymorphisms the population frequency of neutral non-synonymous positions would be identical to the frequencies of synonymous mutations, while the frequencies of functional (i.e., mostly deleterious) mutations would be significantly lower, as these mutations are actively removed by natural selection. This hypothesis was strongly supported, thus the use of our classifier is justified for a broad use with flowering plants.

Overall, our advances open the path to multiple future directions of research. For instance, it would be interesting to infer how different are domesticated plants from their wild progenitors at the genomic level? While it might be assumed that only a few loci contribute to the process of domestication (Gross and Olsen, 2010), domestication can also indirectly affect the entire genome by interfering with natural selection. First, there is strong selection fixing segregating and novel functional alleles. Second, there is an extensive relaxation of natural selection on characters that are important in the wild but not in cultivation, including due to population size reduction. The selective spread of beneficial mutations but also a consequent build-up of deleterious mutations (especially closely linked to selective sweeps) have been well-documented in plants, including rice (Günther and Schmid, 2010) and maize (Pyhäjärvi et al., 2013). However, whether deleterious mutation build-up is a minor nuisance or a major drag on yield remains incompletely understood, and can now be researched. This will help to understand whether 'cleaning out' such adverse mutations, for instance with CRISPR-based tools, might contribute to substantial gains in yield. Further, it opens the way to prioritizing these mutations for being edited out – perhaps of substantial value to the workflow in future agricultural advances.

AUTHOR CONTRIBUTIONS

MK and AI have contributed equally to this work. MS and SN supervised the study.

FUNDING

This work was supported by RSF (Russian Science Foundation) Grant No. 16-16-00007.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2018.01734/full#supplementary-material

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* 7, 248–249. doi: 10.1038/nmeth0410-248

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Bairoch, A. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* 24, 21–25. doi: 10.1093/nar/24.1.21

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Charlesworth, D., and Willis, J. H. (2009). The genetics of inbreeding depression. *Nat. Rev. Genet.* 10, 783–796. doi: 10.1038/nrg2664

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide

polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6, 80–92. doi: 10.4161/fly.19695

Coyne, J. A. (2016). Theodosius Dobzhansky on Hybrid Sterility and Speciation. *Genetics* 202, 5–7. doi: 10.1534/genetics.115.184770

Cristianini, N., and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge: Cambridge University Press, doi: 10.1017/CBO9780511801389

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput. Biol.* 6:e1001025. doi: 10.1371/journal.pcbi.1001025

Epstein, C. J. (1967). Non-randomness of Ammo-acid Changes in the Evolution of Homologous Proteins. *Nature* 215, 355–359. doi: 10.1038/215355a0

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Garcia, A. A. F., Wang, S., Melchinger, A. E., and Zeng, Z.-B. (2008). Quantitative Trait Loci Mapping and The Genetic Basis of Heterosis in Maize and Rice. *Genetics* 180, 1707–1724. doi: 10.1534/genetics.107.082867

Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. *Science* 185, 862–864.

Green, J. R., Korenberg, M. J., and Aboul-Magd, M. O. (2009). PCI-SS: MISO dynamic nonlinear protein secondary structure prediction. *BMC Bioinformatics* 10:222. doi: 10.1186/1471-2105-10-222

Gross, B. L., and Olsen, K. M. (2010). Genetic perspectives on crop domestication. *Trends Plant Sci.* 15, 529–537. doi: 10.1016/j.tplants.2010.05.008

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., et al. (2010). A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science* 327, 883–886. doi: 10.1126/science.1183863

Günther, T., and Schmid, K. J. (2010). Deleterious amino acid polymorphisms in Arabidopsis thaliana and rice. *Theor. Appl. Genet.* 121, 157–168. doi: 10.1007/s00122-010-1299-4

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* 89, 10915–10919. doi: 10.1073/pnas.89.22.10915

Hurtado, D. M., Uziela, K., and Elofsson, A. (2018). A. Deep transfer learning in the assessment of the quality of protein models. arXiv:1804.06281 [Preprint].

Kondrashov, A. S., Sunyaev, S., and Kondrashov, F. A. (2002). Dobzhansky–Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* 99, 14878–14883. doi: 10.1073/pnas.232565499

Kono, T. J. Y., Fu, F., Mohammadi, M., Hoffman, P. J., Liu, C., Stupar, R. M., et al. (2016). The Role of Deleterious Substitutions in Crop Genomes. *Mol. Biol. Evol.* 33, 2307–2317. doi: 10.1093/molbev/msw102

Kono, T. J. Y., Lei, L., Shih, C.-H., Hoffman, P. J., Morrell, P. L., and Fay, J. C. (2018). Comparative genomics approaches accurately predict deleterious variants in plants. *G3 (Bethesda)* 8, 3321–3329. doi: 10.1534/g3.118.200563

Kowarsch, A., Fuchs, A., Frishman, D., and Pagel, P. (2010). Correlated Mutations: A Hallmark of Phenotypic Amino Acid Substitutions. *PLoS Comput. Biol.* 6:e1000923. doi: 10.1371/journal.pcbi.1000923

Lagunas, M., and Garces, E. (2017). *Transfer Learning for Illustration Classification*. Zaragoza: Spanish Computer Graphics Conference (CEIG).

Lapin, M., Hein, M., and Schiele, B. (2014). Learning using privileged information: SVM+ and weighted SVM. *Neural Netw.* 53, 95–108. doi: 10.1016/j.neunet.2014.02.002

Mei, S., Fei, W., and Zhou, S. (2011). Gene ontology based transfer learning for protein subcellular localization. *BMC Bioinformatics* 12:44. doi: 10.1186/1471-2105-12-44

Mezmouk, S., and Ross-Ibarra, J. (2014). The Pattern and Distribution of Deleterious Mutations in Maize. *G3* 4, 163–171. doi: 10.1534/g3.113.008870

Miyata, T., Miyazawa, S., and Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.* 12, 219–236. doi: 10.1007/BF01732340

Pan, S. J., and Yang, Q. (2010). A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Petegrosso, R., Park, S., Hwang, T. H., and Kuang, R. (2018). Systems biology Transfer learning across ontologies for phenome – genome association prediction. *Bioinformatics* 33, 529–536. doi: 10.1093/bioinformatics/btw649

Plekhanova, E., Vishnyakova, M. A., Bulyntsev, S., Chang, P. L., Carrasquilla-Garcia, N., Negash, K., et al. (2017). Genomic and phenotypic analysis of Vavilov's historic landraces reveals the impact of environment and genomic islands of agronomic traits. *Sci. Rep.* 7, 4816. doi: 10.1038/s41598-017-05087-5

Pyhäjärvi, T., Hufford, M. B., Mezmouk, S., and Ross-Ibarra, J. (2013). Complex Patterns of Local Adaptation in Teosinte. *Genome Biol. Evol.* 5, 1594–1609. doi: 10.1093/gbe/evt109

Rieseberg, L. H., and Willis, J. H. (2007). Plant Speciation. *Science* 317, 910–914. doi: 10.1126/science.1137729

Sarkisyan, K. S., Bolotin, D. A., Meer, M. V., Usmanova, D. R., Mishin, A. S., Sharonov, G. V., et al. (2016). Local fitness landscape of the green fluorescent protein. *Nature* 533, 397–401. doi: 10.1038/nature17995

Sievers, F., and Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. *Methods Mol. Biol.* 1079, 105–116. doi: 10.1007/978-1-62703-646-7-6

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* 40, W452–W457. doi: 10.1093/nar/gks539

Sneath, P. H. (1966). Relations between chemical structure and biological activity in peptides. *J. Theor. Biol.* 12, 157–195.

Stone, E. A., and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* 15, 978–986. doi: 10.1101/gr.3804205

Taylor, J. C., Martin, H. C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., et al. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* 47, 717–726. doi: 10.1038/ng.3304

The UniProt Consortium. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45, D158–D169. doi: 10.1093/nar/gkw1099

Turelli, M., and Orr, H. A. (2000). Dominance, epistasis and the genetics of postzygotic isolation. *Genetics* 154, 1663–1679.

Yang, J., Mezmouk, S., Baumgarten, A., Buckler, E. S., Guill, K. E., McMullen, M. D., et al. (2017). Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *PLoS Genet.* 13:e1007019. doi: 10.1371/journal.pgen.1007019

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 04 February 2019
doi: 10.3389/fgene.2018.00711

Characterization of DNA Methylation Associated Gene Regulatory Networks During Stomach Cancer Progression

Jun Wu[1], Yunzhao Gu[2], Yawen Xiao[3], Chao Xia[2], Hua Li[2], Yani Kang[2], Jielin Sun[4], Zhifeng Shao[2], Zongli Lin[5] and Xiaodong Zhao[4]**

[1] School of Life Sciences, East China Normal University, Shanghai, China, [2] Bio-ID Center, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China, [3] Department of Automation, Shanghai Jiao Tong University, Shanghai, China, [4] Shanghai Center for Systems Biomedicine, Shanghai Jiao Tong University, Shanghai, China, [5] Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, United States

OPEN ACCESS

Edited by:
Yuriy L. Orlov, Institute of Cytology and Genetics (RAS), Russia

Reviewed by:
Sheng Liu, Indiana University, United States
Anna Kudryavtseva, Engelhardt Institute of Molecular Biology (RAS), Russia
Leonid Olegovich Bryzgalov, Independent Researcher, Novosibirsk, Russia

***Correspondence:**
Zongli Lin
zl5y@virginia.edu
Xiaodong Zhao
xiaodongzhao@sjtu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *16 October 2018*
Accepted: *18 December 2018*
Published: *04 February 2019*

Citation:
Wu J, Gu Y, Xiao Y, Xia C, Li H, Kang Y, Sun J, Shao Z, Lin Z and Zhao X (2019) Characterization of DNA Methylation Associated Gene Regulatory Networks During Stomach Cancer Progression. Front. Genet. 9:711. doi: 10.3389/fgene.2018.00711

DNA methylation plays a critical role in tumorigenesis through regulating oncogene activation and tumor suppressor gene silencing. Although extensively analyzed, the implication of DNA methylation in gene regulatory network is less characterized. To address this issue, in this study we performed an integrative analysis on the alteration of DNA methylation patterns and the dynamics of gene regulatory network topology across distinct stages of stomach cancer. We found the global DNA methylation patterns in different stages are generally conserved, whereas some significantly differentially methylated genes were exclusively observed in the early stage of stomach cancer. Integrative analysis of DNA methylation and network topology alteration yielded several genes which have been reported to be involved in the progression of stomach cancer, such as *IGF2*, *ERBB2*, *GSTP1*, *MYH11*, *TMEM59,* and *SST*. Finally, we demonstrated that inhibition of *SST* promotes cell proliferation, suggesting that DNA methylation-associated *SST* suppression possibly contributes to the gastric cancer progression. Taken together, our study suggests the DNA methylation-associated regulatory network analysis could be used for identifying cancer-related genes. This strategy can facilitate the understanding of gene regulatory network in cancer biology and provide a new insight into the study of DNA methylation at system level.

Keywords: DNA methylation, gene regulation network, stomach cancer, tumor stages, system level

INTRODUCTION

DNA methylation plays a critical role in tumorigenesis through regulating oncogene activation and tumor suppressor gene silencing (He et al., 2008), and has raised extensive attention in the past decade. It has been shown that tumor initiation and development are associated with aberrant DNA methylation patterns, as documented in stomach cancer development (Tahara and Arisawa, 2015; Yamamoto et al., 2016). Aberrant DNA methylation pattern is the hallmark in the cancer genome (Baylin et al., 2000; Bergman and Cedar, 2013) and is involved in malignant progression (Jones et al., 2013). Although critically involved in malignancy, the implication of DNA methylation in tumorigenesis at system level is less characterized.

The gene regulatory network based analysis is regarded as a powerful way to understand the mechanism of tumorigenesis at system level (Kreeger and Lauffenburger, 2010), and various robust machine learning methods based gene regulatory network inference algorithms were proposed for such analysis (Haury et al., 2012; Slawek and Arodz, 2013; Wu et al., 2016). On the other hand, the rapid development of deep sequencing technologies promotes the generation of a tremendous amount of sequencing data, and an increasing number of network-based methods have been recently applied to understand the molecular mechanism of tumor formation and progression (Anglani et al., 2014; Yang et al., 2014; Bicker et al., 2015).

To further investigate the role of DNA methylation in tumorigenesis at system level, in this study we analyzed the DNA methylation-associated the topology dynamics of gene regulatory network in stomach cancer. We observed that although the DNA methylation patterns are generally conserved, the locus-specific DNA methylation patterns can be identified, especially in the early stage. Comparison of the topology of gene regulatory networks derived from different stages yielded several genes, such as *IGF2*, *ERBB2*, *GSTP1*, *MYH11*, *TMEM59*, and *SST*, of which the regulatory relationship is found to be most severely disrupted. To evaluate the biological relevance, we performed siRNA assay against *SST* in gastric epithelial cell line GES-1 and found that down-regulation of *SST* significantly promotes gastric cell proliferation. Collectively, these results suggest that the integrative analysis of DNA methylation and gene regulatory network across different stages of stomach cancer would be used to identify genes involved in stomach cancer initiation and development, and provides a new insight into the understanding of DNA methylation in carcinogenesis at system level.

RESULTS

Probe-Gene Pairs Assignment

The DNA methylation datasets downloaded from the Cancer Genome Altas (TCGA) data portal were generated using two Illumina Infinium DNA methylation bead arrays (HM27 and HM450). Considering the incompleteness of DNA methylation data, we focused our study on the probes located in the gene promoter regions. Technically, more than one probes were generally designed for a given gene promoter region and it remains unclear which probe-hit methylated region actually affect the expression of the target gene. To address this issue, the distance and correlation criteria were used to assign the proper probes to a gene (See Materials and Methods for further details).

It has been well recognized that DNA hyper-methylation at the promoter region is associated with gene suppression (Bell et al., 2011; Jones, 2012). Due to the unavailability of DNA methylation data and the matched RNA-seq data in normal tissues, we examined the correlation between the pair of the expression level and the DNA methylation level of probes located in the promoter region of a given gene in each tumor stage. Not surprisingly, we observed that negatively correlated pairs outnumber the positive correlated ones (**Figure 1A**). Particularly, in the significantly correlated pairs we found that almost all probe-gene pairs were negatively correlated (**Figure 1B**). The probe-gene pair was assigned if the DNA methylation level of the probe and expression level of a gene are significantly negatively correlated in one of the four tumor stages. With these criteria, 10,777 probe-gene pairs, which consist of 9,830 probes and 7,546 genes, were defined and then used for the downstream analysis.

Global Conserved and Locus Specific DNA Methylation Patterns Across Different Stomach Cancer Stages

With the selected probe-gene pairs, we firstly examined the global methylation patterns across all stomach cancer stages and the normal samples. We classified the probes into unmethylated, hemi-methylated and fully methylated groups using the approach similar to Lokk et al. (2012). To determine proper thresholds, we examined the distributions of the methylation level in all five phenotypes (**Figure 2A**). We found that the distributions of the methylation level in all five phenotypes are very similar. More than half of the probes were unmethylated and only about 15% probes were fully methylated in all samples. The dynamics in the methylation patterns across the five phenotypes was also analyzed. We found that the conservation between every two phenotypes was higher than 80% (**Figure 2B**), indicating that the DNA methylation patterns are globally conserved across all the five phenotypes. Additionally, we found that DNA methylation patterns are relatively more conserved in tumor stages.

Although the overall patterns are considerably conserved, the phenotype-specific methylation presumably plays an important role in initiation and progress of stomach cancer. To test this presumption, we examined the presence of both the unmethylated and fully methylated probe-linked genes in the five phenotypes. Interestingly, we found that both the unmethylated and fully methylated probe-linked genes in normal samples were significantly more than those in tumor samples (**Figure 3**). We next performed gene ontology (GO) analysis of these genes with DAVID (Huang et al., 2009a,b). The results showed that the fully methylated probe-linked genes in normal samples were enriched in the GO items of defense response to bacterium and innate immune response (**Supplementary Table S1**), including *LPO* and *S100A8* which have been reported to be activated in the *H. pylori*-infected gastric mucosa (Semper et al., 2014; Zhuang et al., 2015).

To further understand the biological relevance of the DNA methylation in different stages of stomach cancer, we compared the samples in stages I–IV with the normal samples and identified the significantly differentially methylated probes. We found 1,059, 716, 673 and 635 genes linked to significantly differentially methylated genes in stages I–IV samples, respectively. The top 20 significantly differentially methylated probe linked genes with largest positive and negative mean differences were shown in **Figure 4**, in which we found that several oncogenes and tumor suppressor genes were at the top of the lists (positive and negative directions, respectively) in all four tumor stages, including *ITGA4*, *FGF2*, *FLI1*, *EGFR*, *ERBB2*, VIM, and *DAPK1*. *ITGA4* encodes a member of the integrin alpha chain family that may play a role in cell motility and migration, and the promoter

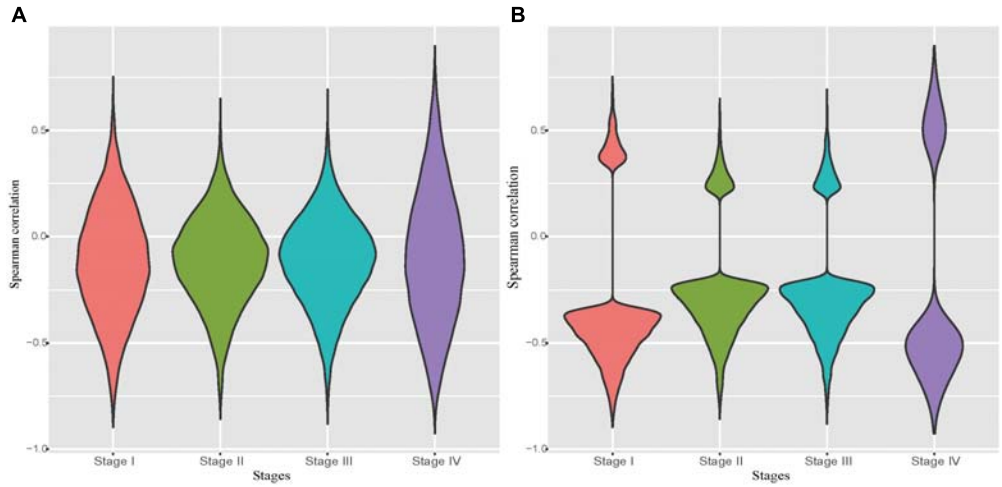


FIGURE 1 | Distribution of correlations between the probe methylation level and the expression of target genes. **(A)**: Distribution of spearman correlation of all potential probe-gene pairs in the four stomach cancer stages. **(B)**: Distribution of spearman correlation of all significantly correlated potential probe-gene pairs in the four stomach cancer stages.

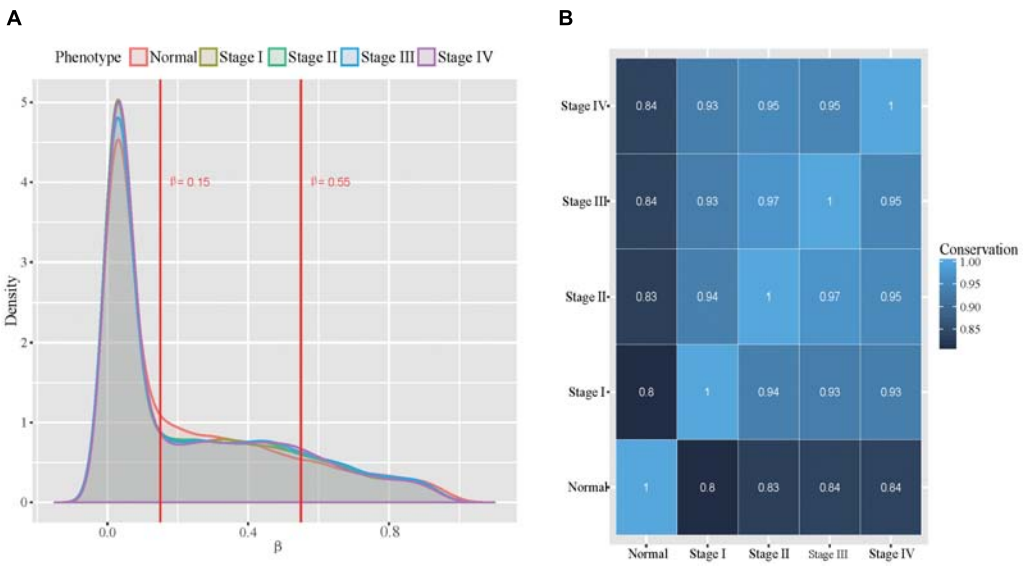


FIGURE 2 | Global view of methylation patterns in all the five types. **(A)**: The distribution of methylation level across all the five phenotypes, where the two red lines represent the thresholds used for dividing the probes into three groups. **(B)**: The conservation between every two phenotypes.

of *ITGA4* was reported to be hyper-methylated in various cancers, such as colorectal cancer (Gerecke et al., 2015), breast cancer (Lian et al., 2012) and gastric cancer (Kim et al., 2009). *DAPK1*, a positive mediator of gamma-interferon induced programmed cell death, was reported to be fully hypo-methylated or up-regulated in several types of cancer, including fistula associated mucinous type anal adenocarcinoma (Sen et al., 2010), nasopharyngeal carcinoma (Luo et al., 2011) and gastric cancer (Zhang et al., 2006).

The Venn diagram of genes with significantly differentially methylation was shown in **Figure 5**. We found that most genes were shared by stages II – IV except in stage I. The GO analysis (**Supplementary Table S2**) shows that the commonly hyper-methylated probe linked genes are mainly involved in carcinogenesis related biological processes, such as cell motion, cell death and cell migration. While the commonly hypo-methylated probe linked genes are mainly involved in development and differentiation biological processes (**Supplementary Table S3**). We also found some genes exclusively present in stage I, suggesting that they are presumably associated with the early stage of stomach cancer. The GO analysis results revealed that both the specifically hyper-methylated genes and the specifically hypo-methylated genes are involved in cell adhesion and

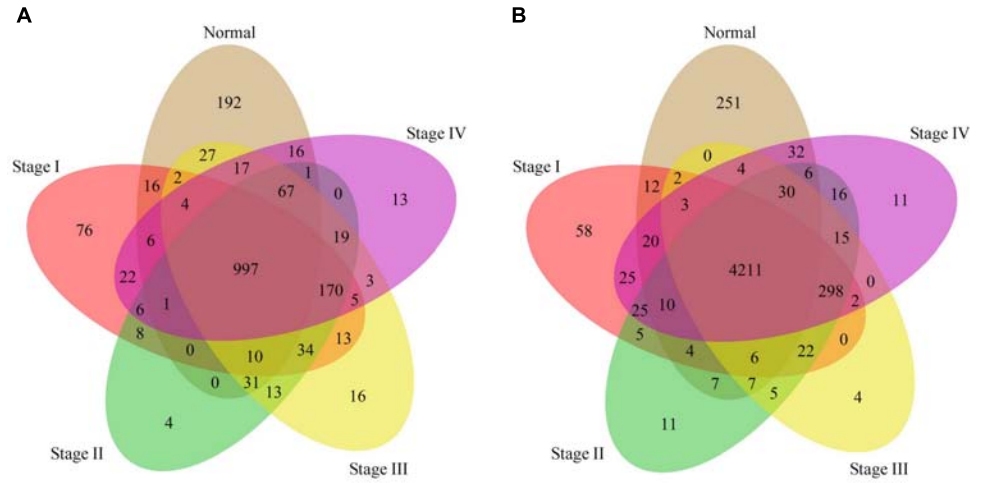


FIGURE 3 | Venn diagrams of genes linked to the fully and unmethylated probes. **(A)**: The Venn diagram of fully methylated probe linked genes with respect to the five phenotypes. **(B)**: The Venn diagram of unmethylated probe linked genes with respect to the five phenotypes.

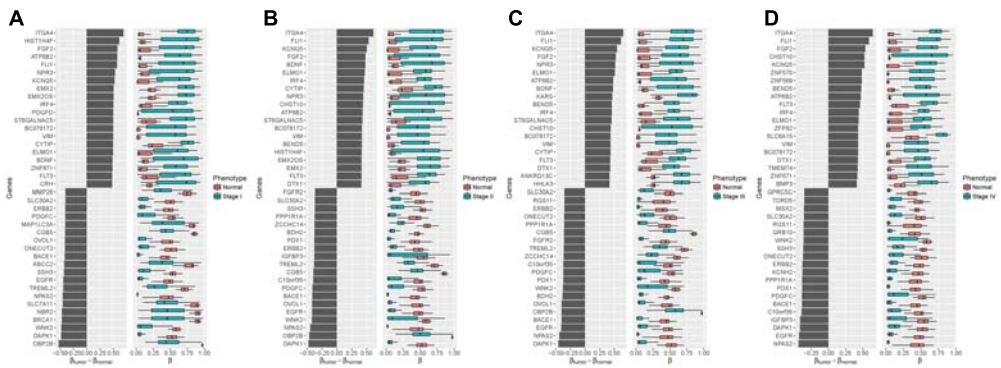


FIGURE 4 | Differential methylation analysis between four tumor stages and the normal phenotype **(A)**: Stage I vs. Normal; **(B)**: Stage II vs. Normal; **(C)**: Stage III vs. Normal; **(D)**: Stage IV vs. Normal. Left: Mean difference between the methylation level in the tumor samples and the normal samples. Right: Distributions of methylation level, with black vertical lines showing medians. Top 20 of the largest positive and negative mean differences with an adjusted *p*-value less than 0.05 are shown.

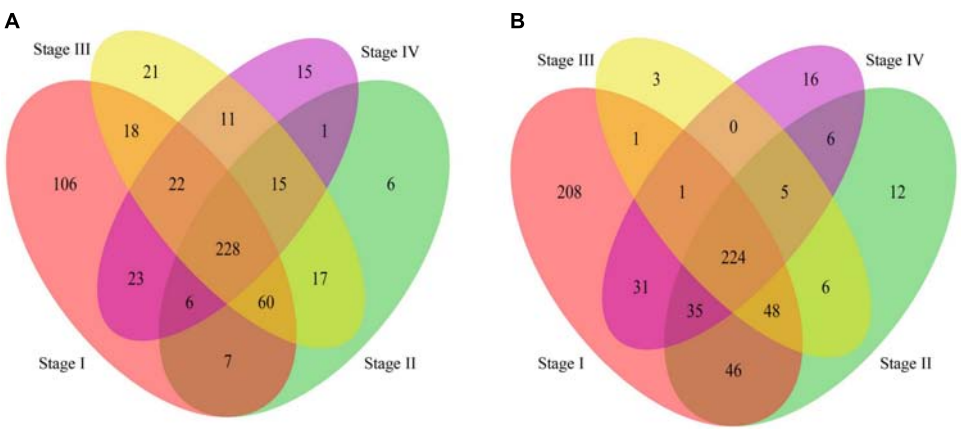


FIGURE 5 | Venn diagram of genes linked to the differentially methylated probes in stage I to IV compared to the normal phenotype. **(A)**: The Venn diagram of genes linked to the hyper-methylated probes. **(B)**: The Venn diagram of genes linked to the hypo-methylated probes.

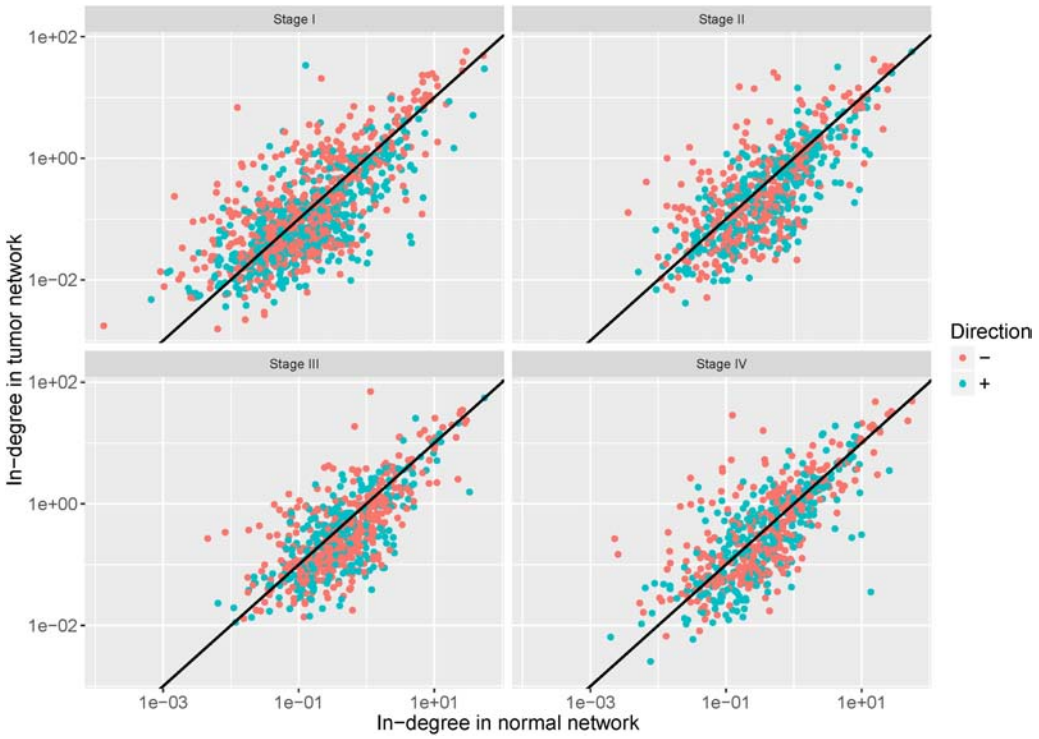


FIGURE 6 | In-degree of each target gene in each network pair. The red dots represent the retained genes that satisfy the assumption that hyper-methylation may cause loss of regulation and hypo-methylation may cause its gain. The blue dots represent genes discarded in the further analysis.

transmembrane transport. The difference is that the genes linked to the specifically hyper-methylated probes are particularly involved in eating behavior and positive regulation of appetite (**Supplementary Table S4**), while the genes linked to the specifically hypo-methylated probes are particularly involved in immune response, response to bacterium and negative regulation of Wnt signaling pathway (**Supplementary Table S5**).

Regulation Gain or Loss Induced by DNA Methylation Alteration

DNA methylation is one of the key epigenetic mechanisms involved in regulation of gene expression. To further understand the role of DNA methylation alteration during the stomach cancer development, we constructed a DNA methylation associated gene regulatory network for each phenotype and analyzed the topology differences among these networks.

To examine the regulation alteration affected by the DNA methylation changes, we screened the target genes based on the assumption that the hyper-methylation leads to the reduction of affinity between the TFs and the binding regions and then may cause the loss of regulation while the hypo-methylation causes its gain (Yao et al., 2016). We calculated in-degree for each target gene and the genes with in-degree increase linked to hypo-methylated probes (in-degree decrease genes linked to hyper-methylated probes) were retained. The in-degree of each target gene in each network pair were shown in **Figure 6**. After filtering, 57%, 52%, 59%, and 54% of target genes were retained in stages I–IV, respectively.

To further investigate the regulation alteration in four tumor stages compared to the normal phenotype, we constructed the differential regulatory networks by subtracting the normal weight matrix from the tumor weight matrixes. The regulation relationship with the absolute weight difference ranking top 1,000 was regarded as true alterations. Finally, for each tumor stage we obtained a differential regulatory network consisting of 1,000 edges that point to 172, 172, 189, and 176 target genes in the four tumor stages. The numbers of edges pertaining to gain or loss of regulation were listed in **Table 1**, in which we observed that the gain number is larger than the loss number in each of the four tumor networks.

For the differential regulatory network in stages I–IV, we ranked the target genes according to the number of gained or lost regulation, respectively. We found several genes were at the top in all the tumor stages. The top 10 target genes (listed in

TABLE 1 | Numbers of gain and loss of regulation in each of the four tumor related networks.

	Stage I	Stage II	Stage III	Stage IV
Loss	308	408	464	419
Gain	692	592	536	581

Supplementary Table S6) with the largest number of regulation alteration were shown in **Figure 7**. In these subgraphs we found that *IGF2*, *ERBB2*, and *GSTP1* rank top in the largest number of regulation gained in all the four differential regulatory networks, and *MYH11*, *SST*, and *TMEM59* rank top in the largest number of regulation lost in all the four differential regulatory networks. *IGF2* is an imprinting gene and plays an essential role in the embryonic development. However, activation of *IGF2* stimulates the proliferation of tumor cells and prevents damaged cells from being destroyed. It was reported that overexpression of *IGF2* plays an important role in carcinogenesis of diffuse type gastric cancer (Wu et al., 1997). *MYH11* belongs to a group of proteins called myosins, which are involved in cell movement and the transport of material within and between cells. It was reported that *MYH11* is not expressed in gastric cancer cell lines (Saeki et al., 2015) and down-regulated *MYH11* correlates with poor prognosis in stage II and stage III colorectal cancer (Wang et al., 2014). These results indicate that the methylation-mediated network analysis facilitates the identification of the key genes involved in tumorigenesis.

To evaluate the authenticity of the genes identified through our network analysis, we performed a siRNA assay against *SST* in gastric epithelial cell line GES-1. Comparing with the control, we found that *SST* suppression results in an increase of cells in S and G2/M phases and the decrease of cells in the G0/G1 phase (**Figure 8**), indicating that *SST* down-regulation promotes cell proliferation. From the results, we found that inhibition of *SST* promotes cell proliferation, which suggests that DNA methylation-associated *SST* suppression possibly contributes to the gastric cancer progression.

DISCUSSION

It has been recognized that aberrant DNA methylation play an import role in tumorigenesis. However, the implication of DNA methylation in gene regulatory network is less characterized. Thus, we performed an integrative analysis of DNA methylation and gene regulatory network with the RNA-seq and DNA methylation data to understand the role of DNA methylation change in the gene regulatory network alteration across different stomach cancer stages.

We first assigned a gene with appropriate probes according to both the location information and correlation relationship. We found that the DNA methylation pattern was global conserved across all phenotypes except some locus specific DNA methylation patterns in the normal phenotype. The differential methylation analysis was also performed to identify the significantly differentially methylated genes in each tumor stage samples. Interestingly, we found more specific alterations in the stage I phenotype compared to the other tumor stages and the GO analysis results showed that these genes are particularly involved in the biological processes closely related to the cancer initiation.

To identify the gene regulation alteration affected by the DNA methylation change, we constructed a DNA methylation associated gene regulatory network in each phenotype and subtracted the normal network from the four tumor networks, respectively. The differential network analysis results showed that the number of regulations gained was larger than that of regulations lost in each of the four tumor networks. We ranked the target genes according to the number of altered regulations

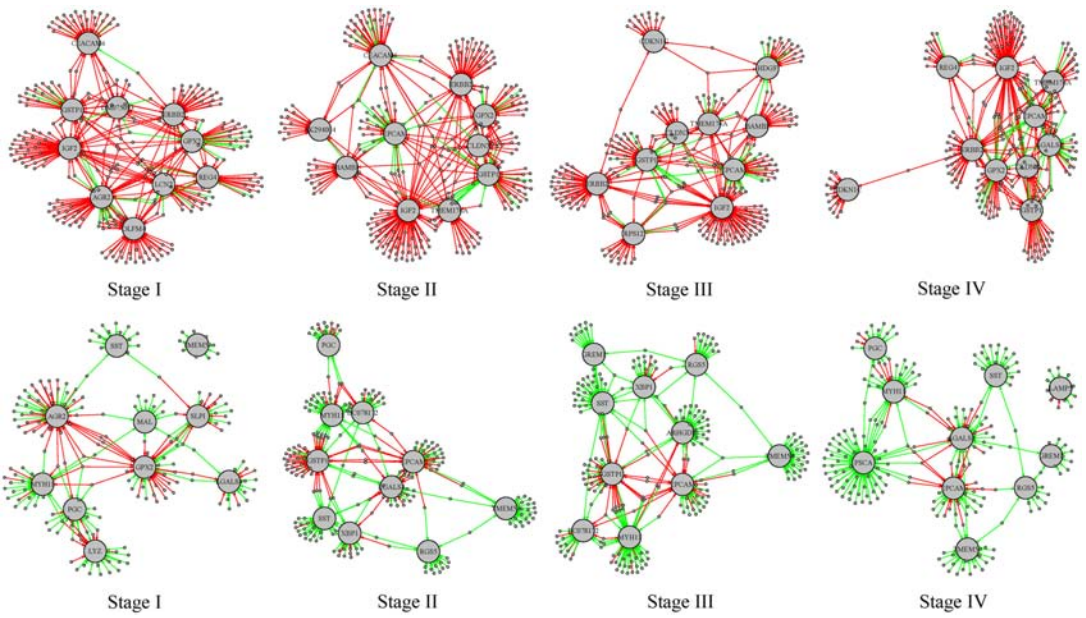


FIGURE 7 | Subgraphs involving the top 10 target genes with the largest number of regulations gained or lost stages I–IV. The red edges represent the regulations gained in the tumor phenotype and the green edges represent regulations lost in the tumor phenotype. The larger gray nodes are target genes and the smaller gray dots are transcription factors involved. The top 4 subgraphs are regulation relationships involving the top 10 target genes with the largest number of regulations gained; the bottom 4 subgraphs are regulation relationships involving the top 10 target genes with the largest number of regulations lost.

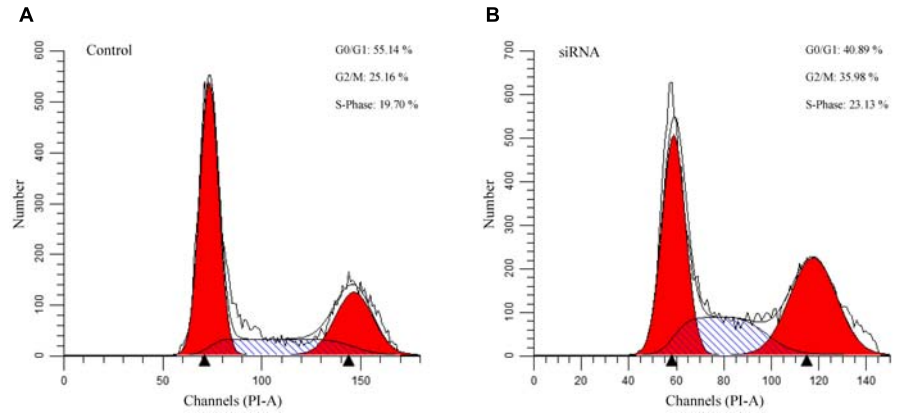


FIGURE 8 | Flow cytometry analysis of the *SST* knockdown gastric cells. **(A)**: Cell cycle analysis of control siRNA GES-1. **(B)**: Cell cycle analysis of *SST* knockdown siRNA GES-1.

and obtained several genes that rank top in all the tumor stages. For example, *IFG2*, *ERBB2*, and *GSTP1* ranked top in the largest number of regulation gain and *MYH11*, *TMEM59*, and *SST* ranked top with the largest number of regulations loss. To examine the biological relevance of the genes identified, we selected *SST* for functional evaluation. We found that inhibition of *SST* can significantly promote cell proliferation, which suggests that down-regulation of *SST* is involved in stomach cancer progression.

In brief, our study demonstrated that integrative analysis of the regulatory network and DNA methylation allows identifying cancer-related gene. The strategy proposed here provides new insight into understanding of the role of DNA methylation in disease at system level.

MATERIALS AND METHODS

Data Collection and Differentially Methylated Sites Identification

The DNA methylation data, gene expression data and clinical data were downloaded from TCGA data portal. The DNA methylation data consist of 302 samples, which were generated using two Illumina Infinium DNA methylation bead arrays, HumanMethylation27 (HM7) and HumanMethylation450 (HM450). The HM27 array contains 27,578 probes that target CpG sites located in proximity to the transcription start sites and the HM450 array contains 482,421 probes that target CpG sites throughout the genome. For ease of description, in the following sections of this article we used probes to represent the corresponding CpG sites.

As neither the HM27 nor the HM45 data contains enough samples for analysis for each phenotype, we only took probes located in gene promoters into account even though the DNA methylation of transcriptional enhancers was also reported to be closely associated with carcinogenesis (Aran and Hellman, 2013). We adopted the strategy mentioned in a previous report (Bass et al., 2014) to preprocess the DNA methylation. Briefly, the probes shared by both the HM27 and HM450 platforms were selected, and the probes that overlap with SNPs, repeat and have any "NA"-masked data points were removed. The probes that hit X and Y chromosomes were also removed. After that we obtained 19,736 probes for further analysis. The gene expression data of 272 samples and 26,540 genes were generated using RNA-seq. The DNA methylation samples and the gene expression samples were further divided into five phenotypes, which are normal and tumor stages I–IV, according to the clinical data. Sample numbers for all phenotype are listed in **Table 2**.

As we did not expect all cases to be from a single molecular subtype, and we sought to identify methylation changes within cases from the same molecular subtype. To identify the significantly differentially methylated probes, we excluded the 10% of samples with the lowest methylation and 10% samples with the highest methylation for each probe and the Wilcoxon Rank Sum test was used to measure the significance. Probes with a BH-adjusted p-value less than 0.05 and an absolute methylation difference greater than 0.2 were regarded as significantly differentially methylated.

Assigning DNA Methylation Sites to the Target Gene

In general more than one DNA methylation probes of the DNA array were designed for a given gene promoter region. Thus, it remains unclear which probes actually affect the expression of the target gene. To address this issue, we used two criteria to

TABLE 2 | Number of samples in each phenotype for the RNA-seq and DNA methylation data.

	Normal	Stage I	Stage II	Stage III	Stage IV
RNA-seq	29	35	93	92	23
DNA methylation	27	37	102	111	25
Matched	0	35	93	92	23

assign the DNA methylation probes for each gene. We initially assigned a probe to a gene if the probe located in the promoter region of the gene. The promoter region of a gene is defined as ±2 kb region around the transcription start site of the gene. The relationship between a probe and a gene is then confirmed with the aid of gene expression based on the evidence that DNA methylation can repress the transcription when it occurs in the promoter region. The samples with matched gene expression data and methylation data were used for the analysis. For each candidate, we tested the significance of the correlation between the DNA methylation level of the probe and expression level of the gene. The Spearman's coefficient was used as the measure of correlation. The correlation significance was obtained with *t*-test and the t statistic was calculated as:

$$t = \frac{r\sqrt{n-2}}{1-r^2},$$

where r is the correlation between the methylation and gene expression and n is the number of samples. The probe-gene pairs were finally confirmed if the BH-adjusted p-value is less than 0.05 and the correlation less than zero.

DNA Methylation Associated Gene Regulatory Network Construction

To construct the DNA methylation associated gene regulatory network, the potential TFs which maybe bind to the DNA methylated regions should be identified. We first obtained JASPAR-2014 motif position weight matrices (PWMs) and ENCODE motif PWMs from the R package motifDb and 2,182 motif PWMs were used for further analysis (ENCODE Project Consortium, 2004; Mathelier et al., 2014). The potential TFs bound to each target gene were predicted according to sequence affinity. We used FIMO (Grant et al., 2011) to scan a ±100 bps sequence around each probe in search for instances of the selected PWMs. A TF was regarded a potential regulator of a probe-linked genes if the p-value of its motif is less than $1E$-4. However, a high sequence affinity just indicated that the TF has a high opportunity to bind to the regulatory region. It was unclear whether the gene relate to the regulatory element is actually bound by the TF.

To measure the confidence of such regulation relationship, we assigned a weight to the edge outgoing from a potential TF to the target gene using our previously proposed gene regulatory network inference method (Wu et al., 2016) with the RNA-seq data. Briefly, we assumed that the expression level of target gene can be formulated by an unknown function of the expression of TFs. We first solved the individual regression problem with the guided regularized random forest algorithm, and then a q-norm normalization was employed to reduce the bias among different regression results and the final results were obtained through refining the previous results according to the sparsity property of large scale gene regulatory networks.

RNA Interference and Cell Cycle Analysis

RNA interference assays were performed as reported previously. SiRNAs for *SST*, or negative control, were synthesized by Shanghai GenePharma Co., Ltd. Cells were transfected with *SST* siRNA or control siRNA using Lipofectamint^{TM} 2000 Transfection Reagent (11668027, Invitrogen) according to the manufacturer's protocol. To measure the efficacy of the gene knockdown, the quantitative real-time reverse transcription polymerase chain reaction (RT-qPCR) was used. Total RNA was extracted using TRIzol Reagent (15596-018, Invitrogen) and resuspended in RNase free water. Reverse transcription of 1 μg RNA was performed using the oligo-dT primer and SuperScrip®III Reverse Transcriptase (18080-044, Invitrogen) according to the manufacturer's protocol. Expression levels were determined by real-time PCR using ABI step one plus (Applied Biosystems, United States). β-actin was used as a control gene for normalization. The relative level of mRNA was calculated as $2^{-\Delta\Delta Ct}$ (means ± SEM, n = 3). The *SST*-targeting siRNA, primer sequences and the RT-qPCR results were provided in **Supplementary Table S7**.

AUTHOR CONTRIBUTIONS

XZ and JW conceived and designed the project. JW wrote the manuscript. YG and YK performed the experiments. JW, YX, CX, and HL performed the analysis and interpretation of data. JS, XZ, ZL, and ZS made a substantial contributions to the design and revisions of the manuscript. All authors have read and approved the final version of the manuscript.

FUNDING

This work was partially funded by the National Natural Science Foundation of China (31671299, 81720108017, and 31801118), the Medicine and Engineering cooperation project of Shanghai Jiao Tong University (YG2017ZD15 and YG2015MS33), the Development Program for Basic Research of China (2014YQ09070904), and the Shanghai Science and Technology Committee Program (17JC1400804).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2018.00711/full#supplementary-material

REFERENCES

Anglani, R., Creanza, T. M., Liuzzi, V. C., Piepoli, A., Panza, A., Andriulli, A., et al. (2014). Loss of connectivity in cancer co-expression networks. *PLoS One* 9:e87075. doi: 10.1371/journal.pone.0087075

Aran, D., and Hellman, A. (2013). DNA methylation of transcriptional enhancers and cancer predisposition. *Cell* 154, 11–13. doi: 10.1016/j.cell.2013.06.018

Bass, A. J., Thorsson, V., Shmulevich, I., Reynolds, S. M., Miller, M., Bernard, B., et al. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202–209. doi: 10.1038/nature13480

Baylin, S. B., Belinsky, S. A., and Herman, J. G. (2000). Aberrant methylation of gene promoters in cancer – Concepts, misconcepts, and promise. *J. Natl. Cancer Inst.* 92, 1460–1461. doi: 10.1093/jnci/92.18.1460

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., et al. (2011). DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* 12:R10. doi: 10.1186/gb-2011-12-1-r10

Bergman, Y., and Cedar, H. (2013). DNA methylation dynamics in health and disease. *Nat. Struct. Mol. Biol.* 20, 274–281. doi: 10.1038/nsmb.2518

Bicker, A., Brahmer, A. M., Meller, S., Kristiansen, G., Gorr, T. A., and Hankeln, T. (2015). The distinct gene regulatory network of myoglobin in prostate and breast cancer. *PLoS One* 10:e0142662. doi: 10.1371/journal.pone.0142662

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia of DNA elements) Project. *Science* 306, 636–640. doi: 10.1126/science.1105136

Gerecke, C., Scholtka, B., Lowenstein, Y., Fait, I., Gottschalk, U., Rogoll, D., et al. (2015). Hypermethylation of ITGA4, TFPI2 and VIMENTIN promoters is increased in inflamed colon tissue: putative risk markers for colitis-associated cancer. *J. Cancer Res. Clin.* 141, 2097–2107. doi: 10.1007/s00432-015-1972-8

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27, 1017–1018. doi: 10.1093/bioinformatics/btr064

Haury, A. C., Mordelet, F., Vera-Licona, P., and Vert, J. P. (2012). TIGRESS: trustful inference of gene regulation using stability selection. *BMC Syst. Biol.* 6:145. doi: 10.1186/1752-0509-6-145

He, X. M., Chang, S. H., Zhang, J. J., Zhao, Q., Xiang, H., Kusonmano, K., et al. (2008). MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res.* 36, D836–D841.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Jones, A., Teschendorff, A. E., Li, Q. X., Hayward, J. D., Kannan, A., Mould, T., et al. (2013). Role of DNA Methylation and epigenetic silencing of HAND2 in endometrial cancer development. *PLoS Med.* 10:e1001551. doi: 10.1371/journal.pmed.1001551

Jones, P. A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 13, 484–492. doi: 10.1038/nrg3230

Kim, J. H., Jung, E. J., Lee, H. S., Kim, M. A., and Kim, W. H. (2009). Comparative analysis of DNA methylation between primary and metastatic gastric carcinoma. *Oncol. Rep.* 21, 1251–1259. doi: 10.3892/or_00000348

Kreeger, P. K., and Lauffenburger, D. A. (2010). Cancer systems biology: a network modeling perspective. *Carcinogenesis* 31, 2–8. doi: 10.1093/carcin/bgp261

Lian, Z. Q., Wang, Q., Li, W. P., Zhang, A. Q., and Wu, L. (2012). Screening of significantly hypermethylated genes in breast cancer using microarray-based methylated-CpG island recovery assay and identification of their expression levels. *Int. J. Oncol.* 41, 629–638. doi: 10.3892/ijo.2012.1464

Lokk, K., Vooder, T., Kolde, R., Välk, K., Võsa, U., Roosipuu, R., et al. (2012). Methylation markers of early-stage non-small cell lung cancer. *PLoS One* 7:e39813. doi: 10.1371/journal.pone.0039813

Luo, X. J., Li, L. L., Deng, Q. P., Yu, X. F., Yang, L. F., Luo, F. J., et al. (2011). Grifolin, a potent antitumour natural product upregulates death-associated protein kinase 1 DAPK1 via p53 in nasopharyngeal carcinoma cells. *Eur. J. Cancer* 47, 316–325. doi: 10.1016/j.ejca.2010.09.021

Mathelier, A., Zhao, X. B., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., et al. (2014). JASPAR: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 42, D142–D147. doi: 10.1093/nar/gkt997

Saeki, N., Komatsuzaki, R., Chiwaki, F., Yanagihara, K., and Sasaki, H. (2015). A GSDMB enhancer-driven HSV thymidine kinase-expressing vector for controlling occult peritoneal dissemination of gastric cancer cells. *BMC Cancer* 15:439. doi: 10.1186/s12885-015-1436-1

Semper, R. P., Mejias-Luque, R., Gross, C., Anderl, F., Müller, A., Vieth, M., et al. (2014). *Helicobacter pylori*-Induced IL-1 beta secretion in innate immune cells is regulated by the NLRP3 inflammasome and requires the Cag pathogenicity island. *J. Immunol.* 193, 3566–3576. doi: 10.4049/jimmunol.1400362

Sen, M., Ozdemir, O., Turan, M., Arici, S., Yildiz, F., Koksal, B., et al. (2010). Epigenetic inactivation of tumor suppressor SFRP2 and point mutation in KRAS proto-oncogene in fistula - associated mucinous type anal adenocarcinoma: report of two cases. *Intern. Med.* 49, 1637–1640. doi: 10.2169/internalmedicine.49.3249

Slawek, J., and Arodz, T. (2013). ENNET: inferring large gene regulatory networks from expression data using gradient boosting. *BMC Syst. Biol.* 7:106. doi: 10.1186/1752-0509-7-106

Tahara, T., and Arisawa, T. (2015). DNA methylation as a molecular biomarker in gastric cancer. *Epigenomics* 7, 475–486. doi: 10.2217/epi.15.4

Wang, R. J., Wu, P., Cai, G. X., Wang, Z. M., Xu, Y., Peng, J. J., et al. (2014). Down-regulated MYH11 expression correlates with poor prognosis in stage II and III colorectal cancer. *Asian Pac. J. Cancer Prev.* 15, 7223–7228. doi: 10.7314/APJCP.2014.15.17.7223

Wu, J., Zhao, X., Lin, Z., and Shao, Z. (2016). Large scale gene regulatory network inference with a multi-level strategy. *Mol. Biosyst.* 12, 588–597. doi: 10.1039/c5mb00560d

Wu, M. S., Wang, H. P., Lin, C. C., Sheu, J. C., Shun, C. T., Lee, W. J., et al. (1997). Loss of imprinting and overexpression of IGF2 gene in gastric adenocarcinoma. *Cancer Lett.* 120, 9–14. doi: 10.1016/S0304-3835(97)00279-6

Yamamoto, H., Yoshida, Y., Morita, R., Oikawa, R., Maehata, T., Watanabe, Y., et al. (2016). Methylation analysis of gastric juice-derived exosomal DNA is useful for early detection of gastric cancer. *Gastroenterology* 150, S871–S871. doi: 10.1038/ctg.2016.40

Yang, Y., Han, L., Yuan, Y., Li, J., Hei, N., and Liang, H. (2014). Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* 5:3231. doi: 10.1038/ncomms4231

Yao, L. J., Shen, H., Laird, P., Farnham, P. J., and Berman, B. P. (2016). Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Clin. Cancer Res.* 16:105. doi: 10.1186/s13059-015-0668-3

Zhang, X. T., Yashiro, M., Ren, J., and Hirakawa, K. (2006). Histone deacetylase inhibitor, trichostatin A, increases the chemosensitivity of anticancer drugs in gastric cancer cell lines. *Oncol. Rep.* 16, 563–568. doi: 10.3892/or.16.3.563

Zhuang, Y., Cheng, P., Liu, X. F., Peng, L. S., Li, B. S., Wang, T. T., et al. (2015). A pro-inflammatory role for Th22 cells in *Helicobacter pylori*-associated gastritis. *Gut* 64, 1368–1378. doi: 10.1136/gutjnl-2014-307020

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 07 February 2019
doi: 10.3389/fgene.2019.00040

Intracellular Vesicle Trafficking Genes, *RabC*-GTP, Are Highly Expressed Under Salinity and Rapid Dehydration but Down-Regulated by Drought in Leaves of Chickpea (*Cicer arietinum* L.)

*Gulmira Khassanova[1], Akhylbek Kurishbayev[1], Satyvaldy Jatayev[1], Askar Zhubatkanov[1], Aybek Zhumalin[1], Arysgul Turbekova[1], Bekzak Amantaev[1], Sergiy Lopato[2], Carly Schramm[2], Colin Jenkins[2], Kathleen Soole[2], Peter Langridge[3,4] and Yuri Shavrukov[2]**

[1] Faculty of Agronomy, S. Seifullin Kazakh AgroTechnical University, Astana, Kazakhstan, [2] Biological Sciences, College of Science and Engineering, Flinders University, Bedford Park, SA, Australia, [3] School of Agriculture, Food and Wine, University of Adelaide, Adelaide, SA, Australia, [4] Wheat Initiative, Julius-Kühn-Institute, Berlin, Germany

OPEN ACCESS

Edited by:
Yuriy L. Orlov,
Russian Academy of Sciences, Russia

Reviewed by:
Awais Rasheed,
International Maize and Wheat Improvement Center, Mexico
Mehar Hasan Asif,
National Botanical Research Institute (CSIR), India

***Correspondence:**
Yuri Shavrukov
yuri.shavrukov@flinders.edu.au

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *25 October 2018*
Accepted: *18 January 2019*
Published: *07 February 2019*

Citation:
Khassanova G, Kurishbayev A, Jatayev S, Zhubatkanov A, Zhumalin A, Turbekova A, Amantaev B, Lopato S, Schramm C, Jenkins C, Soole K, Langridge P and Shavrukov Y (2019) Intracellular Vesicle Trafficking Genes, RabC-GTP, Are Highly Expressed Under Salinity and Rapid Dehydration but Down-Regulated by Drought in Leaves of Chickpea (Cicer arietinum L.). Front. Genet. 10:40. doi: 10.3389/fgene.2019.00040

Intracellular vesicle trafficking genes, *Rab*, encoding small GTP binding proteins, have been well studied in medical research, but there is little information concerning these proteins in plants. Some sub-families of the *Rab* genes have not yet been characterized in plants, such as *RabC* – otherwise known as *Rab18* in yeast and animals. Our study aimed to identify all *CaRab* gene sequences in chickpea (*Cicer arietinum* L.) using bioinformatics approaches, with a particular focus on the *CaRabC* gene sub-family since it featured in an SNP database. Five isoforms of the *CaRabC* gene were identified and studied: *CaRabC-1a*, *-1b*, *-1c*, *-2a* and *-2a**. Six accessions of both Desi and Kabuli ecotypes, selected from field trials, were tested for tolerance to abiotic stresses, including salinity, drought and rapid dehydration and compared to plant growth under control conditions. Expression analysis of total and individual *CaRabC* isoforms in leaves of control plants revealed a very high level of expression, with the greatest contribution made by *CaRabC-1c*. Salinity stress (150 mM NaCl, 12 days in soil) caused a 2-3-fold increased expression of total *CaRabC* compared to controls, with the highest expression in isoforms *CaRabC-1c*, *-2a** and *-1a*. Significantly decreased expression of all five isoforms of *CaRabC* was observed under drought (12 days withheld water) compared to controls. In contrast, both total *CaRabC* and the *CaRabC-1a* isoform showed very high expression (up-to eight-fold) in detached leaves over 6 h of dehydration. The results suggest that the *CaRabC* gene is involved in plant growth and response to abiotic stresses. It was highly expressed in leaves of non-stressed plants and was down-regulated after drought, but salinity and rapid dehydration caused up-regulation to high and very high levels, respectively. The isoforms of *CaRabC* were differentially expressed, with the highest levels recorded for *CaRabC-1c* in controls and under salinity stress,

and for *CaRabC-1a* – in rapidly dehydrated leaves. Genotypic variation in *CaRabC-1a*, comprising eleven SNPs, was found through sequencing of the local chickpea cultivar Yubileiny and germplasm ICC7255 in comparison to the two fully sequenced reference accessions, ICC4958 and Frontier. Amplifluor-like markers based on one of the identified SNPs in *CaRabC-1a* were designed and successfully used for genotyping chickpea germplasm.

Keywords: abiotic stresses, Amplifluor-like SNP markers, bioinformatics, *CaRab* gene, differential gene expression, gene isoforms

INTRODUCTION

Plant genomes include a superfamily of genes that encode small GTP-binding proteins (Guanosine triphosphatases) that are classified into four groups: *Arf*, *Rab*, *Ran* and *Rho*; and an additional *Ras*-GTP gene group is found only in yeast and animals (Ma, 2007). Small GTP-binding proteins were first described in medical research, where the term "*Ras*" stemmed from their association with rat sarcoma (Chang et al., 1982; Bishop, 1985; Chavrier et al., 1990). The remaining three-letter names are not related in structure or function to the genes but rather refer to their product or some other feature (Coffin et al., 1981). Small GTP-binding proteins are known to be involved in a diverse range of activities in eukaryotes that are vital for growth, development and repair; from cytoskeletal organization, vacuolar storage and signaling, to modulation of gene expression (Takai et al., 2001). The mechanism for the regulation of GTP-binding proteins is conserved in all organisms and involves cycling between active (GTP-bound) and inactive (GDP-bound) forms, so they are often described as "molecular switches" that are turned "on" or "off" via the hydrolysis of GTP (Marshall, 1993). Activation requires the dissociation of GDP, which can be either stimulated by a regulatory factor named GEP (GDP/GTP Exchange Protein) or inhibited by GDI (GDP Dissociation Inhibitor; Takai et al., 2001; Liu et al., 2015; Martín-Davison et al., 2017).

Rab proteins, encoded by *Rab*-GTP genes, are normally prenylated at their carboxyl terminus. The hydrophobic prenyl-groups facilitate attachment to membranes and are therefore integral to the biological role performed by Rab proteins in vesicle trafficking via endocytic and exocytic pathways between the endoplasmic reticulum, Golgi membrane network, endosome, plasma membrane and all intracellular membranes (Alory and Balch, 2003). Rab proteins are highly conserved across kingdoms, from yeast to animals and plants (Haubruck et al., 1987; Marcote et al., 2000), but are most often present as a small family of highly similar genes. They are divided into either nine (Ma, 2007) or 18 clades (Agarwal et al., 2009) based on their structure, with only eight clades represented in plants. Historically, different nomenclatures were adopted for identification of *Rab* genes in plants compared to animals. For example, in plants, eight capitalized letters from A to H were used in the names of *Rab* genes, while the numbers 1 to 11 were applied in human, animal and yeast research. In the absence of a universal system of nomenclature for *Rab* genes and their proteins, a list of all known genes and their respective identifiers for both nomenclatures is given later in the text.

The genes for Rab GTP-binding proteins should not be confused with the similarly named *Dehydrin* genes in plants, which are also known as *RAB*, meaning "Responsive to ABA" (Abscisic acid). *Dehydrins* encode proteins belonging to the large but very different group of Late embryogenesis abundant proteins, LEA (Hundertmark and Hincha, 2008). For example, *AtRAB18* (or *AtRab18*) was described and studied in *Arabidopsis thaliana* in response to various abiotic stresses and ABA treatment (Lång and Palva, 1992; Rushton et al., 2012; Hernández-Sánchez et al., 2017). Despite the identical name, this gene is neither structurally nor functionally related to the *Rab*-GTP genes, and care must be taken to clearly distinguish between the two. The mixing of these two different types of genes is unfortunately apparent in recent publications. For example, Jiang et al. (2017) studied the correctly designated *TaRab18* (=*TaRabC1*) gene in response to stripe rust in bread wheat, but this gene was incorrectly compared with *RAB18* (Responsive to ABA) in *Arabidopsis*, rice and maize. As a result, the Authors wrongly cited work by Lång and Palva (1992) and others on the *Dehydrin AtRab18* to support their findings on the sensitivity of *TaRab18* (=*TaRabC1*) to ABA.

In plants, *Rab*-GTP genes are reportedly involved in multiple physiological processes (Borg et al., 1997; Rehman and Sansebastiano, 2014; He et al., 2018; Lawson et al., 2018) and are often highly expressed in response to biotic and abiotic stresses (Marcote et al., 2000; Stenmark and Olkkonen, 2001; Zerial and McBride, 2001; Rutherford and Moore, 2002; Ma, 2007; Woollard and Moore, 2008; Agarwal et al., 2009). However, despite the numerous links, little is known about the precise molecular mechanisms underlying their involvement in plant stress responses.

One of the first studies to report a link between Rab protein and abiotic stress was a report by O'Mahony and Oliver (1999) who found increased transcript levels of the *Rab2* gene (otherwise known as *RabB*) in the desiccation-tolerant grass *Sporobolus stapfianus* in response to dehydration, but decreased transcript levels after rehydration. This suggested the involvement of *SsRab2* in both the short-term response and later recovery from desiccation. SsRab2 was found to share 90% similarity to Rab2 proteins found in rice, maize, *Arabidopsis*, *Lotus japonicus* and soybean (O'Mahony and Oliver, 1999). Since that time, links to various stresses have been established for genes encoding Rab proteins in numerous plants, and especially in species with high abiotic stress-tolerance such as *Lilium formolongi* – *LfRabB*

(Howlader et al., 2017), poplar – *PtRabE1b* (Zhang et al., 2018), and *Mesembryanthemum crystallinum* – *McRab5b* (=*McRabF*) (Bolte et al., 2000). Interestingly, many plant species were studied for *RabG* genes and their corresponding proteins including the halophyte species, *Aeluropus lagopoides* – *AlRab7* (=*AlRabG*) (Rajan et al., 2015) and food grain crop, *Pennisetum glaucum* – *PgRab7* (=*PgRabG*) (Agarwal et al., 2008), as well as more stress susceptible crops such as rice, *Oryza sativa* – *OsRab7* (=*OsRabG*) (Nahm et al., 2003) and peanut, *Arachis hypogaea* – *AhRabG* (Sui et al., 2017), and the model species *A. thaliana* – *AtRab7* (=*AtRabG*) (Mazel et al., 2004). A comprehensive analysis of all *MpRab* genes was reported for the liverwort, *Marchantia polymorpha* (Minamino et al., 2018).

Rab transcripts are often found to show different responses to abiotic stresses. For example, in rice, dehydration triggered a strong increase in *OsRab7* (=*OsRabG*) transcript after 4 h and then a decrease after 10 h. However, no significant changes were found in response to cold or salinity stress (Nahm et al., 2003). Similarly, in the halophytic grass *A. lagopoides*, *AlRab7* (=*AlRabG*) was upregulated by dehydration, but salinity stress caused no significant increase in transcript levels (Rajan et al., 2015). In another halophyte, *M. crystallinum*, expression of *McRab5b* (=*McRabF*) was higher after 2 h and continued to rise over 3 days of very strong salt stress (400 mM NaCl), but wilting or osmotic stress triggered no change in expression (Bolte et al., 2000). These differences obviously reflect various roles of the intracellular membrane system to abiotic stresses and may provide the key to uncovering the precise molecular mechanisms underlying differential plant susceptibility or tolerance to an environmental stress.

A number of studies have used a transgenic approach to shed light on the mechanisms explaining the link between Rab proteins and plant stress and to explore how Rab proteins could play a role in the breeding of more stress-tolerant crops. For example, Mazel et al. (2004) constitutively overexpressed *AtRabG3e* (=*AtRab7*) in *Arabidopsis*. The transgenic plants accumulated more sodium in vacuoles and showed greater tolerance to salinity and osmotic stress. Evidence was also found for increased endocytosis in roots and leaves and entry of Reactive oxygen species into the cell to trigger signaling and subsequent activation of stress tolerance mechanisms (Mazel et al., 2004; Baral et al., 2015). *AhRabG*, *OsRab7 (=OsRabG)* and *OsRab11* (=*OsRabA*) were also overexpressed in transgenic peanut and rice, respectively, producing plants that showed relatively higher salinity tolerance compared to wild-type plants (Peng et al., 2014; Sui et al., 2017; Chen and Heo, 2018). In transgenic peanut plants, of 132 genes differentially expressed, most were identified as transcription factors (TF) relating to salinity tolerance (Sui et al., 2017).

The aim of this study was to identify and analyze a possible candidate gene involved in the tolerance to drought, salinity and rapid dehydration in chickpea, *C. arietinum*, a species that is becoming increasingly popular as a cash crop in agricultural areas with the requirements for moderate tolerance to high temperatures, drought and salinity stress during the growing season. A candidate gene *CaRabC1*, belonging to the family of *Rab*-GTP genes, was identified from an SNP database using bioinformatic and molecular genetic analyses. Currently, the only report concerning chickpea *Rab*-GTP genes was published by Muñoz et al. (2001), who identified a Rab-specific GDI in chickpea seedlings showing 96% homology to *MsRab11f* (=*MsRabG*), a GDI in *Medicago truncatula* (Yaneva and Niehaus, 2005). Our study therefore represents the first report of the *Rab*-GTP family of genes in *C. arietinum*. We present the results of bioinformatic analyses of the identified genes and tests conducted to assess the expression of all isoforms of the *CaRabC* gene family in response to salinity, drought and rapid dehydration in selected chickpea genotypes. Amplifluor-like markers based on one of the identified SNPs in *CaRabC-1a* were used for genotyping of chickpea germplasm.

MATERIALS AND METHODS

Plant Material

A germplasm collection comprising 250 chickpea (*C. arietinum* L) samples from the ICRISAT Reference set plus local accessions were tested over 3 years in field trials in Northern and Central Kazakhstan. Six accessions were selected during field trials for further molecular analyses, as listed in **Table 1**. The first accession, cv. Yubileiny, originated from Krasnokutskaya Breeding Station, in the Saratov region (Russia), and is used as a Standard for local field trials with chickpea accessions. The remaining five chickpea lines were selected from the original 230 collected in the ICRISAT Reference set, to represent diverse gene-pool sources.

Identification of the Gene of Interest Using Bioinformatics and Molecular Phylogenetic Comparative Analysis

Bioinformatics and systems biology methods were applied in this study to identify a target gene or "Gene of Interest" (GoI) with a potential role in tolerance to abiotic stresses in chickpea. Initially, the SNP database for *C. arietinum* L.[1] was used to search and select one suitable SNP with a short fragment of sequence for further study. The full-length nucleotide sequence of the GoI and its corresponding polypeptide sequence was retrieved from the same database and used for both BLASTN and BLASTP in NCBI and in GenomeNet Database Resources, hosted by Kyoto University, Japan[2]. All chickpea gene sequences with KEGG and NCBI identification and the encoded proteins were downloaded from GenomeNet and NCBI databases, while chromosome locations were checked using LIS, Legume Information System database[3]. The *A. thaliana* genes displaying the highest level of similarity to each GoI within the gene family were identified using alignments from the same database.

Multiple sequence alignments of nucleotide sequences for the *Rab* family of genes were conducted in CLUSTALW

[1]https://www.ncbi.nlm.nih.gov/snp

[2]https://www.genome.jp/tools/blast

[3]https://legumeinfo.org/organism/Cicer/arietinum

TABLE 1 | List and short description of six selected chickpea germplasm accessions used for molecular analyses.

Code	Name	Cultivar/Landrace/Line	Ecotype	Origin
Yub	Yubileiny	Cultivar	Kabuli	Russia
ICC-4841	P6615	Landrace	Kabuli	Morocco
ICC-7255	NEC1628; SN8	Landrace	Kabuli	India
ICC-1392	P1240; 141-1	Landrace	Desi	India
ICC-4918		Elite line	Desi	India
ICC-12726	RFA100-3	Landrace	Desi	Ethiopia

using GenomeNet Database Resources[4], while CLC Main Workbench software[5] was used for protein amino acid sequence alignment.

The molecular dendrogram was constructed using BLASTP at GenomeNet Database Resources (See footnote 2) with the function of ETE3 v3.0.0b32 (Huerta-Cepas et al., 2016) and MAFFT v6.861b applied using the default options (Katoh and Standley, 2013). The FastTree v2.1.8 program with default parameters was used for phylogenetic tree preparation (Price et al., 2009).

Abiotic Stress Treatments: Salinity, Drought and Rapid Dehydration

Three experiments applying abiotic stress treatments (salinity, slow drought and rapid dehydration) were carried out in chickpea for RT-qPCR analyses using the same conditions as described earlier in our publication for wheat (Zotova et al., 2018). The size of containers used, number of plants, soil type and growth conditions were all as described and no artificial inoculation of rhizobium was applied.

For salt stress, twenty-four uniform seedlings in each of six accessions were grown for one month in two separate containers. On "Day 0," three plants from each accession (three biological replicates) were randomly selected from each container, before the salt stress was applied. The two youngest fully developed leaves were collected from each selected plant into a 10-ml plastic tube and immediately frozen in liquid nitrogen and stored at –80°C until RNA extraction. Subsequently, 200 ml of 150 mM NaCl was applied to the container, covering the entire soil surface but avoiding any direct contact with the plants. The NaCl treatments were applied four times, on every third day following Day 0 (over 12 days in total) in treatment containers, while the same volume of tap-water without NaCl was used under the same schedule in the control containers. No solution was lost through drainage from any container. No supplementary $CaCl_2$ was added despite the recommended requirements in experiments with hydroponics. This is because the soil mix used contained sufficient available calcium and no symptoms of Ca deficiency were apparent in the treated plants. After 12 days, as on Day 0, the two youngest fully developed leaves were collected from each of three plants both in salt treatments and controls. Leaf samples were immediately frozen in liquid nitrogen and stored at –80°C for RNA extraction.

Experiments with slowly droughted plants and rapid dehydration of detached leaves were carried out using exactly the same schedule as described in Experiments 1 and 2 in our previous paper on wheat (Zotova et al., 2018).

RNA Extraction, cDNA Synthesis and qPCR Analysis

Frozen leaf samples were ground as described below for DNA extraction. TRIzol-like reagent was used for RNA extraction following the protocol described by Shavrukov et al. (2013) and all other steps for RNA extraction and cDNA synthesis were as described previously (Zotova et al., 2018). The procedures included DNase treatment (Qiagen, Germany), and the use of a MoMLV Reverse Transcriptase kit (Biolabmix, Novosibirsk, Russia). All cDNA samples were checked for quality control using PCR and yielded bands of the expected size on agarose gels.

Diluted (1:2) cDNA samples were used for qPCR analyses using either a QuantStudio-7 Real-Time PCR instrument (Thermo Fisher Scientific, United States) at S. Seifullin Kazakh AgroTechnical University, Astana, Kazakhstan, or Real-Time qPCR system, Model CFX96 (BioRad, Gladesville, NSW, Australia) at Flinders University, Australia. The qPCR protocol was similar in both instruments as published earlier (Shavrukov et al., 2016), wherein the total volume of 10 μl q-PCR reactions included either 5 μl of 2xBiomaster HS-qPCR SYBR Blue (Biolabmix, Novosibirsk, Russia) for experiments in Kazakhstan or 5 μl of 2xKAPA SYBR FAST (KAPA Biosystems, United States) for experiments in Australia, 4 μl of diluted cDNA, and 1 μl of two gene-specific primers (3 μM of each primer) (**Supplementary material 1**). Expression data for the target genes were calculated relative to the average expression of the two reference genes: *CAC*, Clathrin adaptor complexes, medium subunit (Reddy et al., 2016) and *GAPDH*, Glyceraldehyde-3-phosphate dehydrogenase (Garg et al., 2010) (**Supplementary material 1**). At least three biological and two technical replicates were used in each qPCR experiment.

DNA Extraction, Sequencing and SNP Identification

Plants were grown in control (non-stressed) conditions in containers with soil as described above. Five uniform one month-old individual plants were selected from each accession and five leaves were collected and bulked for leaf samples. Leaf samples

[4]https://www.genome.jp/tools-bin/clustalw

[5]https://www.qiagenbioinformatics.com/products/clc-main-workbench

frozen in liquid nitrogen were ground in 10-ml tubes with two 9-mm stainless ball bearings using a Vortex mixer. DNA was extracted from the bulked leaf samples with phenol-chloroform as described in our earlier papers (Shavrukov et al., 2016; Zotova et al., 2018). One microliter of DNA was checked on a 0.8% agarose gel to assess quality, and concentration was measured by Nano-Drop (Thermo Fisher Scientific, United States).

To identify SNPs in the GoI and compare them with annotated accessions in databases, primers were designed in exon regions flanking introns (**Supplementary material 1**). PCR was performed in 60 μl volume reactions containing 8 μl of template DNA adjusted to 20 ng/μl, and with the following components in the final concentrations listed: 1xPCR Buffer, 2.2 mM $MgCl_2$, 0.2 mM each of dNTPs, 0.25 μM of each primer and 4.0 units of Taq-DNA polymerase in each reaction (Maxima, Thermo Fisher Scientific, United States). PCR was conducted on a SimpliAmp Thermal Cycler (Thermo Fisher Scientific, United States), using a program recommended by the Taq-polymerase manufacturer, with the following steps: initial denaturation, 95°C, 4 min; 35 cycles of 95°C for 20 s, 55°C for 20 s, 72°C for 1 min, and final extension, 72°C for 5 min. Single bands of the expected size were confirmed after visualization of 5 μl of the PCR product in 1% agarose gel. The remaining PCR reaction volume (55 μl) was purified using FavorPrep PCR Purification kit (Favorgene Biotec Corp., Taiwan) following the Manufacturer's protocol. The concentrations of purified PCR products were measured using NanoDrop (Thermo Fisher Scientific, United States) and later used as a template (100 ng) in a sequencing reaction with the Beckman Coulter Sequencing kit, following the Manufacturer's protocol. Sanger sequencing and analysis of results were performed on a Beckman Coulter Genetic Analysis System, Model CEQ 8000 (Beckman Coulter, United States) following the Manufacturer's protocol and software at S. Seifullin Kazakh AgroTechnical University, Astana (Kazakhstan). The identified SNPs were used to design allele-specific primers that were applied in Amplifluor-like SNP analysis. Two fully sequenced chickpea accessions, ICC4958 of the Desi ecotype, and Frontier of the Kabuli ecotype, were used as the reference genomes[6].

SNP Amplifluor Analysis

Amplifluor-like SNP analysis was carried out using a QuantStudio-7 Real-Time PCR instrument (Thermo Fisher Scientific, United States) as described previously (Jatayev et al., 2017; Zotova et al., 2018) with the following modifications: Each reaction contained 3 μl of template DNA adjusted to 20 ng/μl, 5 μl of Hot-Start 2xBioMaster (MH020-400, Biolabmix, Novosibirsk, Russia[7]) with all other components as recommended by the manufacturers, including $MgCl_2$ (2.0 mM). One microliter of a mixture of two fluorescently labeled Universal probes was added (0.25 μM each) and 1 μl of allele-specific primer mix (0.15 μM of each of two forward primers and 0.78 μM of the common reverse primer). Four microliter of Low ROX (Thermo Fisher Scientific, United States) was added as a passive reference label to the Master-mix as prescribed for the qPCR instrument. Assays were performed in 96-well microplates. Sequences of the Universal probes and primers as well as the sizes of amplicons are presented in **Supplementary material 1**.

PCR was conducted using a program adjusted from those published earlier (Jatayev et al., 2017; Zotova et al., 2018): initial denaturation, 95°C, 2 min; 14 "doubled" cycles of 95°C for 10 s, 60°C for 10 s, 72°C for 20 s, 95°C for 10 s, 55°C for 20 s and 72°C for 50 s; with recording of allele-specific fluorescence after each cycle. Genotyping by SNP calling was determined automatically by the instrument software, but each SNP result was also checked manually using amplification curves and final allele discrimination. Experiments were repeated twice over different days, where two technical replicates confirmed the confidence of SNP calls.

Statistical Analysis

IBM SPSS Statistical software was applied to calculate means, standard errors, and to estimate the probabilities for significance using ANOVA tests.

RESULTS

Bioinformatics and Comparative Phylogenetic Analysis

During the initial screening of SNP No. 2103, rs853191221 [*C. arietinum*] within the chickpea SNP database (**Supplementary material 2**), NCBI BLAST analysis revealed the closest nucleotide accession to be XM_012715175.1, encoding a Ras-related protein in *C. arietinum* with the corresponding *RabC1*-like gene (LOC101496214, transcript variant X2, mRNA). We designated this gene as the isoform *CaRabC-1a*.

To identify the full list of all members of *CaRab* genes in chickpea, bioinformatics approaches were used to search and analyze annotated sequences and whole genome sequences available in public databases using comparisons to the reference genome of *A. thaliana*. As a result, eight sub-families of *CaRab* gene were identified, with 54 isoforms. The corresponding accession IDs for the genes and proteins, as well as references to *Arabidopsis* genes with the highest level of similarity are shown in **Table 2**.

The sequences of all 54 isoforms of *CaRab* genes identified in chickpea were used to construct a phylogenetic tree (**Figure 1**). Eight distinct clades were identified in the molecular dendrogram, and the letter corresponding to each sub-family name is used to distinguish the corresponding clade. The biggest and most diverse was Clade A, the *CaRabA* gene sub-family while Clades B and F contained only two accessions each. Clades D, G, H and F are molecularly similar, but most distanced from other sub-families of the *CaRab* gene. The sub-family *CaRabC* contained five isoforms with the closest sub-families being *CaRabD* and *CaRabE* (**Figure 1**).

Protein sequence analysis of five isoforms from sub-family *CaRabC* (**Figure 2**) showed distinct separation of CaRabC-1 from CaRabC-2. The closest molecular similarity was found between CaRabC-1b and CaRabC-1c with the next greatest similarity

[6]http://www.cicer.info/databases.php

[7]http://biolabmix.ru/en/products

TABLE 2 | The eight identified sub-families (*RabA – RabH*) of the chickpea *CaRab*, with 54 isoforms and their corresponding accession ID listed for genes and proteins as well as reference to closest genes in *Arabidopsis*.

Clades of *Rab* genes in plants	Group of *Rab* genes in mammals	Chromosome in chickpea	KEGG ID/NCBI Gene ID in chickpea	NCBI protein ID in chickpea	ID of closest gene in *Arabidopsis thaliana*
RabA-1b	*Rab11*	Ca1	101493366	XP_004489015	At1g16920
*RabA-1b**		Unknown	101507798	XP_004514628	At1g16920
RabA-1c		Ca6	101498968	XP_004505464	At5g45750
RabA-1d		Ca1	101504539	XP_004487583	At4g18800
RabA-1f		Ca6	101497770	XP_004505114	At5g60860
*RabA-1f**		Ca1	101489774	XP_004485855	At5g60860
RabA-1g		Ca1	101509653	XP_004487187	At3g15060
*RabA-1g**		Ca6	101492905	XP_004505524	At3g15060
RabA-2a		Ca1	101495904	XP_004485429	At1g09630
RabA-2b		Ca6	101503536	XP_004503210	At1g07410
*RabA-2b**		Ca2	101512788	XP_004490850	At1g07410
RabA-2d		Ca6	101512110	XP_004507156	At5g59150
RabA-3		Ca3	101490262	XP_004494844	At1g01200
*RabA-3**		Ca7	101511394	XP_004510635	At1g01200
RabA-4a		Ca5	101489316	XP_004502410	At5g65270
*RabA-4a**		Ca6	101515580	XP_004504100	At5g65270
RabA-4c		Unknown	101491858	XP_004516148	At5g47960
RabA-4d		Ca3	101508321	XP_004494010	At3g12160
*RabA-4d**		Ca4	101501691	XP_004497938	At3g12160
RabA-5a		Ca4	101498836	XP_004495529	At5g47520
RabA-5b		Ca3	101509902	XP_004492211	At3g07410
RabA-5e		Unknown	101508123	XP_004515580	At1g05810
RabA-6a		Ca8	101513235	XP_004512724	At1g73640
*RabA-6a**		Ca6	101500012	XP_004503285	At1g73640
RabB-1b	*Rab2*	Ca2	101515168	XP_004489550	At4g35860
RabB-1c		Ca1	101501307	XP_004486381	At4g17170
*RabB-1c**		Ca7	101496042	XP_004510833	At4g17170
RabC-1a	***Rab18***	**Ca4**	**101496214**	**XP_004498372**	**At1g43890**
RabC-1b		**Ca5**	**101488438**	**XP_004502943**	**At1g43890**
RabC-1c		**Ca6**	**101490080**	**XP_004503936**	**At1g43890**
RabC-2a		**Unknown**	**101497183**	**XP_004515929**	**At5g03530**
RabC-2a*		**Ca4**	**101498490**	**XP_004496130**	**At5g03530**
RabD-1	*Rab1*	Ca1	101495577	XP_004485428	At3g11730
RabD-2a		Ca3	101506934	XP_004492924	At1g02130
*RabD-2a**		Unknown	101514122	XP_004515343	At1g02130
RabD-2c		Ca7	101496365	NP_001265926	At4g17530
RabE-1a	*Rab8*	Ca4	101497052	XP_004495000	At3g53610
*RabE-1a**		Ca1	101515594	XP_004487032	At5g59840
RabE-1b		Ca3	101504780	XP_004494002	At3g53610
RabE-1c		Ca4	101491866	XP_004497298	At3g46060
*RabE-1c**		Ca6	101506447	XP_004505885	At3g46060
RabF-1	*Rab5*	Ca4	101504907	XP_004496241	At3g54840
RabF-2b		Ca6	105851137	XP_012572119	At4g19640
RabG-3a	*Rab7*	Ca4	101504052	XP_004496410	At4g09720
RabG-3b		Ca2	101498677	XP_004489906	At1g22740
*RabG-3b**		Ca6	101514161	XP_004504012	At1g22740
RabG-3c		Ca6	101510738	XP_012573047	At3g16100
RabG-3d		Ca1	101509216	XP_004486740	At1g52280
RabG-3e		Ca6	101513957	XP_012573072	At1g49300
RabG-3f		Ca6	101496247	XP_004507422	At3g18820

(Continued)

TABLE 2 | Continued

Clades of *Rab* genes in plants	Group of *Rab* genes in mammals	Chromosome in chickpea	KEGG ID/NCBI Gene ID in chickpea	NCBI protein ID in chickpea	ID of closest gene in *Arabidopsis thaliana*
RabH-1d	*Rab6*	Ca7	101507228	XP_004508989	At2g44610
*RabH-1d**		Ca8	101496604	XP_004511760	At2g22290
RabH-1e		Ca5	101492440	XP_004502420	At5g10260
*RabH-1e**		Ca6	101507522	XP_004504075	At5g10260

The sub-family CaRabC studied in this paper is indicated in bold type.

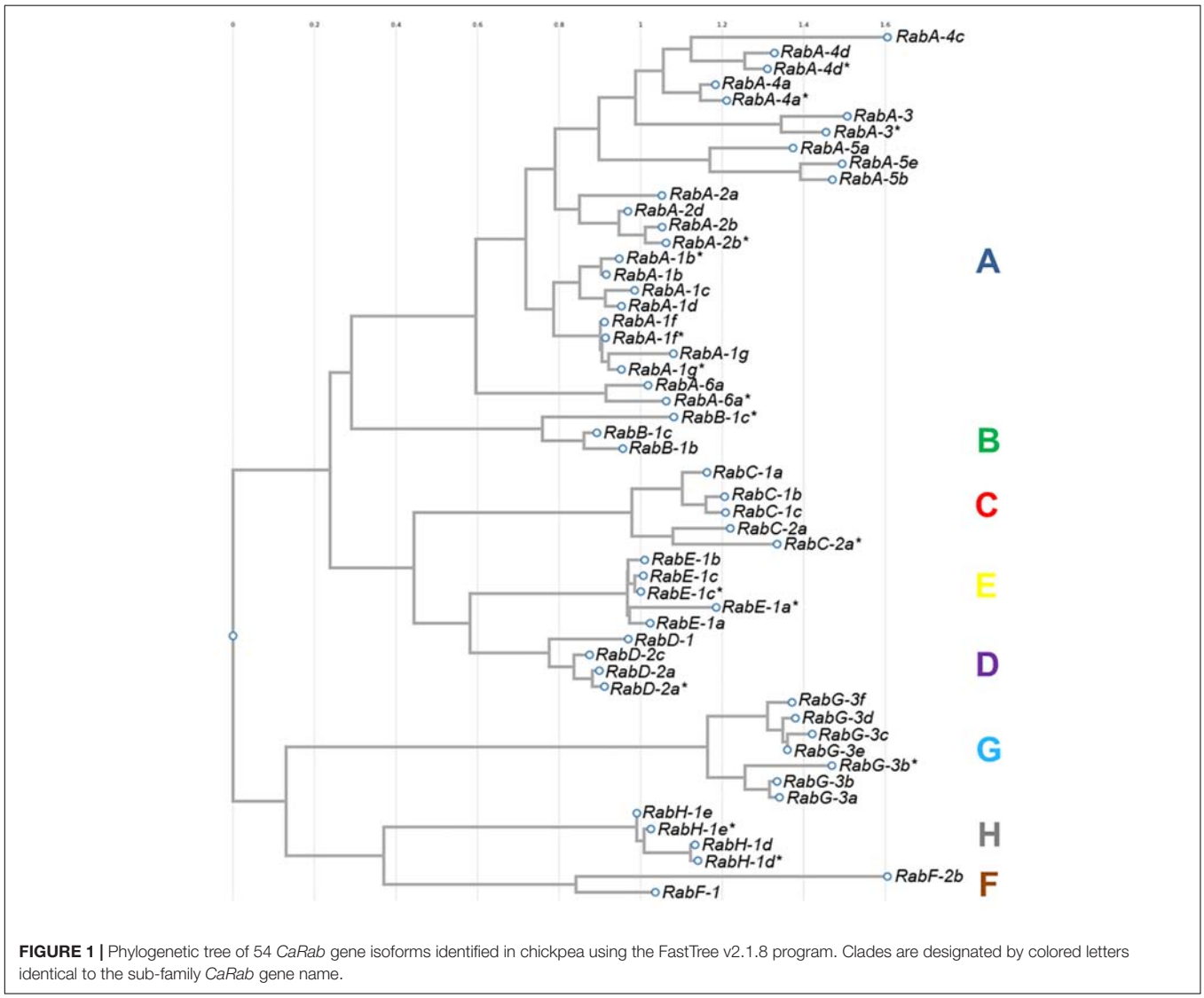


FIGURE 1 | Phylogenetic tree of 54 *CaRab* gene isoforms identified in chickpea using the FastTree v2.1.8 program. Clades are designated by colored letters identical to the sub-family *CaRab* gene name.

shared with CaRabC-1a, while CaRabC-2a and CaRabC-2a* were the most diverged from all others (**Figure 2**).

RT-qPCR and Gene Expression Analysis

Primers for RT-qPCR analysis were designed based on the alignment and comparison of CDS sequences of five identified *CaRabC* isoforms listed in **Table 2**. To estimate the total expression level of all five *CaRabC* genes combined, common primers with degenerative nucleotides were designed based on the longest consensus regions in the alignments. In addition, 3'-ends of gene-specific primers were designed for specific SNPs to maximize the specificity of qPCR analysis for each of the five isoforms of *CaRabC* gene (**Supplementary material 3**).

Initially, the expression level of *CaRabC* gene was determined in control plants grown under favorable conditions for all isoforms combined, as well as for each of them separately (**Figure 3A**). All six studied chickpea accessions, 3 Kabuli and

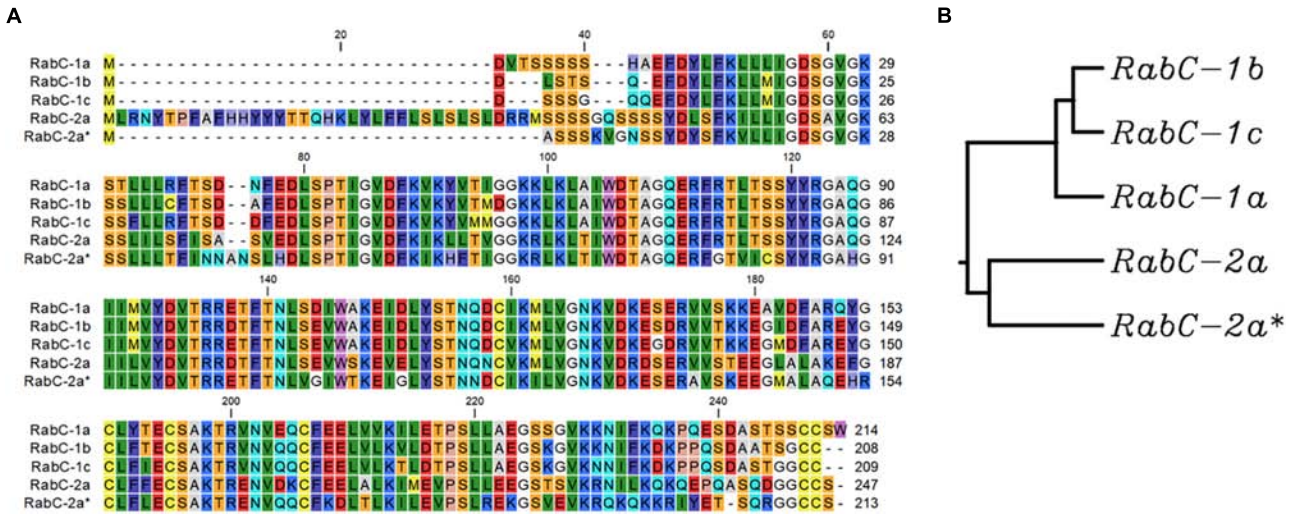


FIGURE 2 | A comparison of amino acid sequences **(A)**, and Rooted UPGMA phylogenetic tree with branch length **(B)** of the five isoforms of CaRabC proteins identified in chickpea. Multiple sequence alignment conducted presented using CLC Main Workbench software.

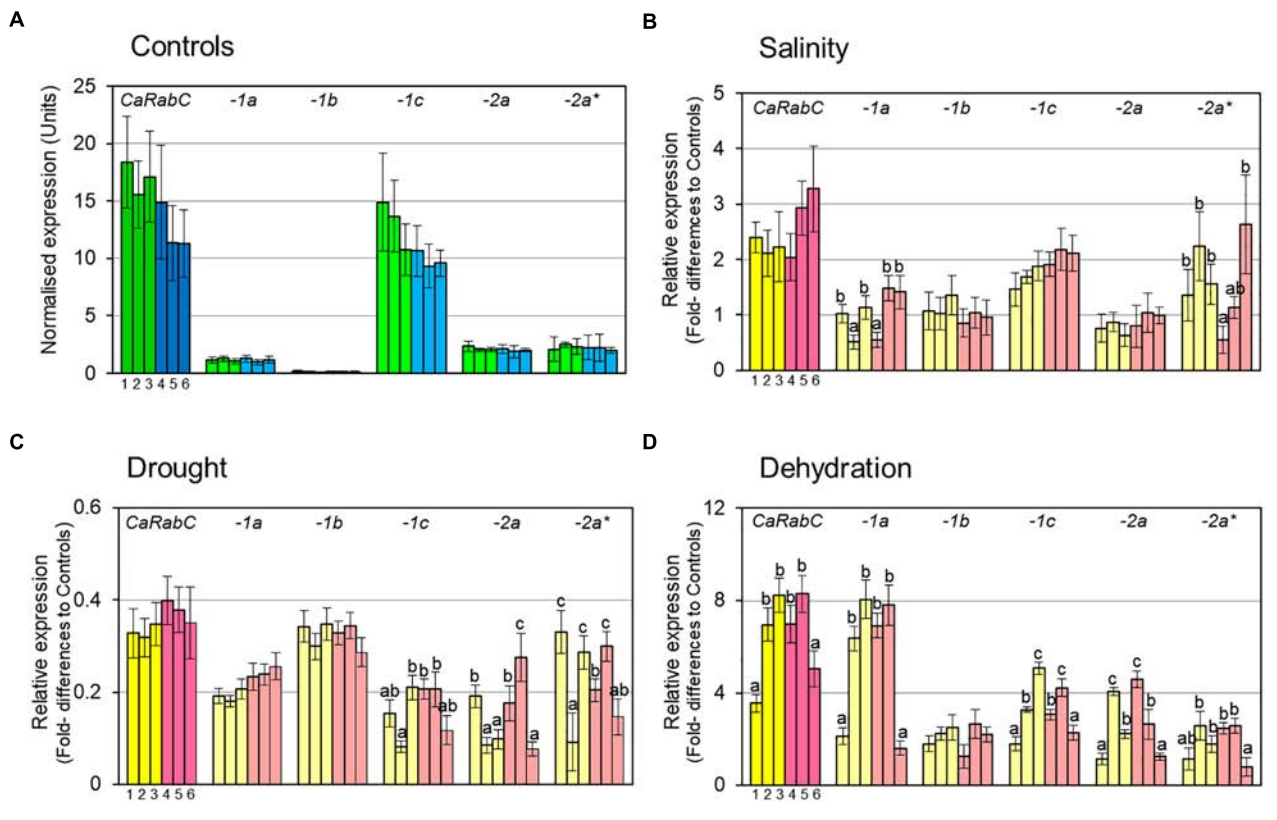


FIGURE 3 | RT-qPCR analysis of *CaRabC* gene family expression in chickpea leaves: **(A)** In favorable, non-stressed conditions (Controls) for 3 Kabuli and 3 Desi (dark green and dark blue, respectively); and the relative gene expression compared to Controls under: **(B)** Gradual salt stress application, 150 mM NaCl, 7 days; **(C)** Slowly developing drought in pots with soil, 12 days; and **(D)** Rapid dehydration of detached leaves, 6 h, room temperature. All isoforms of *CaRabC* gene combined (darker colors) and the five separate isoforms (lighter colors) of the *CaRabC-1a*, *-1b*, *-1c*, *-2a* and *-2a** (for corresponding gene family) were analyzed separately. Each set contained six chickpea accessions, including three Kabuli ecotypes, shown in yellow (**1**, Yubileiny; **2**, ICC-7255; and **3**, ICC-4841), and three Desi ecotypes, shown in pink (**4**, ICC-1392; **5**, ICC-4918; and **6**, ICC-12726). Data were normalized using an average for two reference genes, calculated with ANOVA, and are presented as means for three biological and two technical replicates ± SE, shown as error bars. Significant differences (at least for $P > 0.95$) for each gene isoform and within each set of chickpea accessions are shown by different letters according to ANOVA tests.

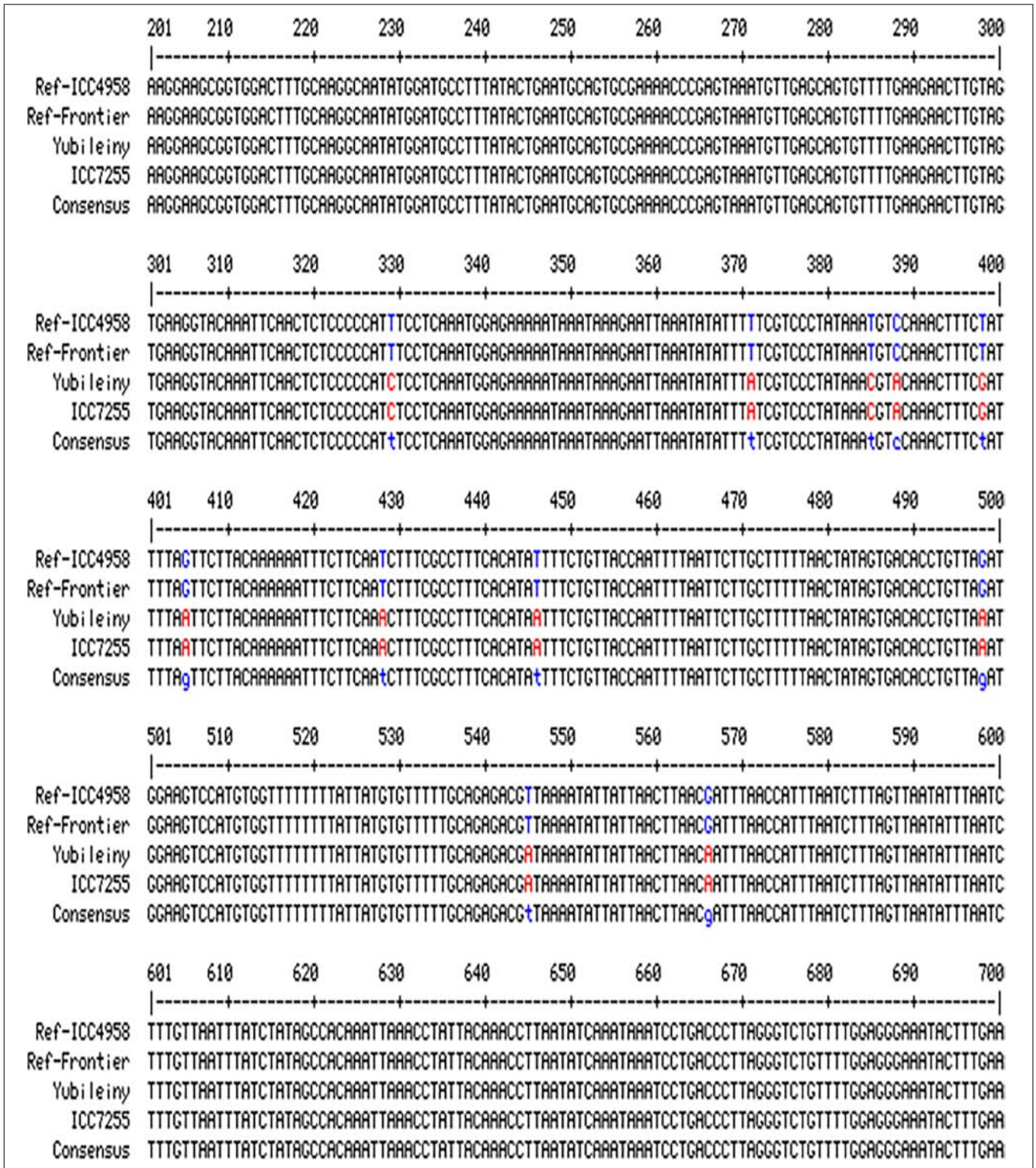


FIGURE 4 | Partial alignments of *CaRabC-1a* (LOC101496214) amplicons. The alignment was produced with primers F5&R5 from the studied chickpea germplasm, Yubileiny and ICC7255, compared to two fully sequenced reference accessions, ICC4958 and Frontier. The studied amplicons were located on Chromosome 4, position 27,819,854–27,818,706 (Reverse order), in the reference accession ICC4958 (Desi ecotype), while the amplicon position in the second reference accession Frontier (Kabuli ecotype), was at 38,617,439–38,616,292 (Reverse order), also on Chromosome 4. Eleven identified SNPs are shown in red for the two studied accessions against those indicated in blue for the two reference accessions.

3 Desi (dark green and dark blue, respectively, in **Figure 3A**), showed a very high level of total *CaRabC* gene expression, ranging from 11.2 to 18.4 relative expression units, with non-significant differences among the six studied genotypes. The expression level of a single isoform of *CaRabC-1c* had maximal (63–88%) contribution in the *CaRabC* gene expression in total. Two isoforms, *CaRabC-2a* and *-2a**, both showed very similar levels of 1.9–2.5 expression units. A level of around 1 expression unit was observed in the isoform *CaRabC-1a*, similar to the average level for the two reference genes used in this study. An extremely low level of expression (approximately 10-fold lower than both reference genes) was shown for the last isoform *CaRabC-1b* (**Figure 3A**).

For salinity stress (**Figure 3B**), a high level of expression of the total *CaRabC* gene family was observed with 2-3.3-fold higher expression relative to Controls, but no significant differences were found within each set of six studied accessions due to relatively wide variability between replicates. In all studied genotypes, the isoform *CaRabC-1c* made the highest contribution to the gene expression (around 1.5–2-fold above the Controls). Only two accessions, No. 2 (ICC-7255, Kabuli) and No. 6 (ICC-12726, Desi), showed a higher level of *CaRabC-2a** isoform expression (2.2- and 2.6-fold, respectively) but these data were quite variable. Significant genetic variation was found for expression levels of *CaRabC-1a* and *CaRabC-2a**. Expression levels of two isoforms, *CaRabC-1b* and *CaRabC-2a*, did not differ from Controls (**Figure 3B**).

A different expression pattern for the *CaRabC* gene family was found for the drought experiment, where total expression was down-regulated by 0.3–0.4-fold compared to Controls (**Figure 3C**). The highest contribution to gene expression was made by the isoform *CaRabC-1b*. There was no significant genetic variation for *CaRabC-1a* and *CaRabC-1b* among the studied germplasm while the other three isoforms were quite variable (**Figure 3C**).

In contrast, rapid dehydration of detached leaves resulted in an up-to 8-fold increase of expression for the total *CaRabC* gene family expression, as well as isoform *CaRabC-1a,* compared to controls (**Figure 3D**). With the exception of *CaRab1b*, significant genetic variation was observed among the studied chickpea accessions for all other isoform expression profiles.

Amplicon Sequencing Showed an SNP in the Candidate Gene CaRabC-1a

The initial SNP discovered was annotated at position 516 from the start-codon in the identified CDS, LOC101496214, based on the reverse-compliment order in the SNP-containing fragment. The full nucleotide sequence of the accession and position of this initial SNP is presented in **Supplementary material 2**.

To check for the presence/absence of the initial SNP in the studied chickpea accessions, several pairs of primers were designed flanking the SNP. The most successful primer pair, F5&R5, amplified a fragment of 1148 bp. A fragment of the alignment showing polymorphic amplicons from the germplasm

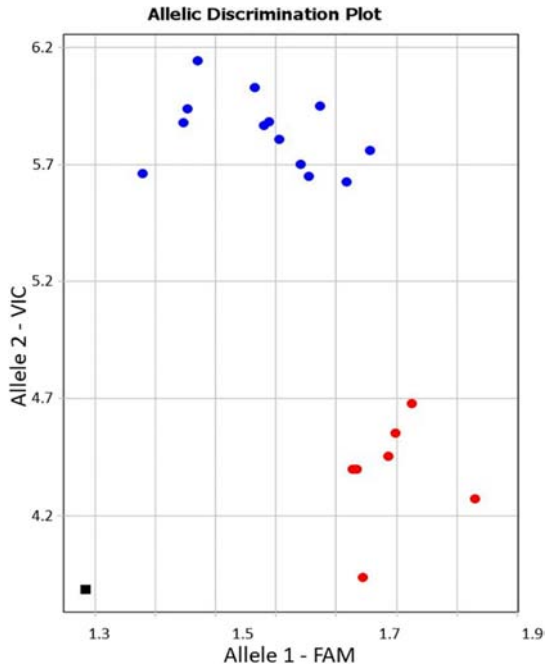


FIGURE 5 | An example of allele discrimination in the chickpea germplasm collection using the Amplifluor-like SNP marker KATU-C22. *X*- and *Y*-axes show Relative amplification units, ΔRn, for FAM and VIC fluorescence signals, respectively. Red dots represent homozygote (*aa*) genotypes with allele 1 (FAM), and blue dots represent homozygote (*bb*) genotypes for allele 2 (VIC) identified with automatic SNP calling. The black square shows the no template control (NTC) using water instead of template DNA.

sequences compared to two fully sequenced reference chickpea accessions (ICC4958, Desi ecotype and Frontier, Kabuli ecotype) in *CaRabC-1a* is presented in **Figure 4**. The sequencing of the amplified fragments revealed the presence of 11 new SNPs in two chickpea accessions, Yubileiny and ICC7255, both Kabuli ecotypes (**Table 1**), compared to the two reference accessions. All 11 identified SNPs recorded high scores, and clear nucleotide peaks at the SNP positions were assessed manually. Interestingly, the initial SNP recorded in the database was monomorphic among the two reference accessions and two genotypes sequenced in our study.

SNP Screening in CaRabC-1a Using Amplifluor-Like Markers

Allele-specific primers, KATU-C22-F&R, were designed for one of the selected SNPs from the 11 identified in the studied fragment of isoform *CaRabC-1a* to use with Amplifluor-like genotyping analysis. Details on the design of primers and positions of the studied SNPs are presented in **Supplementary material 4**. The example in **Figure 5** shows allele discrimination using Amplifluor-like SNP marker KATU-C22, where allele 1 (FAM) has been identified in chickpea accessions with SNP genotypes similar to reference accessions ICC4958 and Frontier but allele 2 (VIC) was found in germplasm similar to Yubileiny and ICC7255 (**Figure 5**).

DISCUSSION

Rab-GTP proteins are well known in oncology studies in human and animals, but in plants there is increasing evidence that they play a central role in the tolerance to abiotic and biotic stresses. Nevertheless, it appears that the mechanism of membrane trafficking with which they are associated is similar in cells of both humans and plants. Most *Rab* genes of the eight clades represented in the molecular phylogenetic tree in plants, have similar corresponding groups of genes in human and other animal genomes. A greater or lesser diversity of isoforms for each clade of *Rab* genes just reflects the differing outcomes of evolution in the plant and animal kingdoms.

In plants, the most studied groups of *Rab* genes are from Clades G and H, where multiple vacuolar trafficking pathway components were demonstrated (Vernoud et al., 2003; Peng et al., 2014; Uemura and Ueda, 2014; Brillada and Rojas-Pierce, 2017). These types of *Rab* genes encode proteins that have been associated with a response to salinity and osmotic stresses, and are thought to associate with pre-vacuolar vesicles. Thus, Rab proteins may enhance relocation of Na^+ ions to the vacuole, after they reach a toxic level in the cytoplasm of cells. Whilst there has been less attention placed on other groups of *Rab* genes, including the diverse Clade A with its many isoforms and the non-diverse Clade B with only two gene members, there is practically nothing known about Clade C of *Rab* in plants (Vernoud et al., 2003; Jha et al., 2014; Rehman and Sansebastiano, 2014; Lawson et al., 2018). Despite the strong similarity between *A. thaliana* and *C. arietinum*, our bioinformatic results show significant differences in the number of *Rab* isoforms in most clades.

In the work described here, 54 isoforms of *CaRab* genes were identified in chickpea, indicating an evolutionary reorganization when compared to *A. thaliana*, where 57 *AtRab* isoforms have been identified (Vernoud et al., 2003). Clade C in the chickpea dendrogram has not been previously identified, described or studied, and contains the five isoforms: *CaRabC-1a*, *-1b*, *-1c*, *-2a* and *-2a**. The first three isoforms show similarity to *AtRabC-1* (At1g43890, **Table 1**) while the latter two isoforms in chickpea were similar to another single isoform *AtRabC-2a* (At5g03530). The isoform *AtRabC-2b* (At3g09910), listed in a comprehensive analysis of the *Rab* genes in *A. thaliana* (Vernoud et al., 2003), has no ortholog in the *C. arietinum* genome. To avoid any misunderstanding with the classification of *CaRabC-2a* and *-2a** isoforms, we have used an asterisk instead of another letter, to indicate its very similar polypeptide structure.

Following the bioinformatics study, the expression analyses of total *CaRabC* for all five isoforms revealed high levels of expression of the gene family in leaves of non-stressed young chickpea plants compared to two reference genes (**Figure 3A**). More importantly, a single isoform, *CaRabC-1c*, made the major contribution to the gene expression, indicating a very active role of this isoform in chickpea plant development under non-stressed conditions. In the absence of other reports comparing expression of individual and combined (bulk) isoforms of *Rab* genes in plants, our conclusions await further verification and discussion.

Under salt stress, the dominance of the *CaRabC-1c* isoform in expression profiles was not as pronounced as under control conditions and was more comparable to other isoforms in some of the studied chickpea accessions, particularly *CaRabC-1a* and *CaRabC-2a**. Therefore, at least three isoforms of *CaRabC* were salinity-responsive and the two latter ones were strongly genotype-dependent (**Figure 3B**).

An unexpected result was found in the comparison of *CaRabC* gene expression in response to slowly progressing drought of whole plants and rapid dehydration of detached leaves. Only a few reports have described expression of different genes in parallel experiments with drought and dehydration. For example, a peroxisomal isoform of *APX*, Ascorbate peroxidase, was down-regulated under strong drought but up-regulated in desiccated leaves in a cultivar of cowpea, *Vigna unguiculata* (D'Arcy-Lameta et al., 2006). Similar results were reported for two genes associated with loss of water during slow drought progression compared to rapid dehydration of barley leaves: *HvMT2,* a metallothionein-like protein, and *2HvLHCB*, Chlorophyll *a-b* binding protein of LHCII type III (Gürel et al., 2016). Therefore, there are examples of genes related to drought and dehydration that can be down- and up-regulated, in several plant species. However, our results show for the first time that all isoforms of *CaRabC* were strongly down-regulated under the slowly developing drought, but very strongly up-regulated in rapidly dehydrated leaves (**Figures 3C,D**).

Amplifluor-like SNP markers and other molecular markers are very helpful in identifying genetic polymorphisms in diverse germplasm accessions. In the current study, the molecular marker KATU-C22 was useful for genotyping one isoform *CaRabC-1a* (**Figure 5**). This allows for tracking of the different variants of this gene and the possibility of linking variants with an associated phenotype. Additional markers are now needed for all other isoforms of *CaRabC* and other GoI, but this will require further investment in sequencing in the future. It also may be worth looking for SNPs in the upstream promoter regions of the gene family, since this could explain the variation in expression between the genotypes.

CaRabC is just one sub-family from a large *CaRab* gene family involved in controlling cell membrane trafficking, and like the other *Rab* genes investigated to date (reviewed in Flowers et al., 2018), it is responsive and potentially associated with the adaptation of plants to abiotic stresses. For comparison, in the bacteria *Salmonella*, the Rab18 protein (related to RabC in plants) is actively involved in endocytosis and is localized in the early endocytic compartment of cells (Hashim et al., 2000). In plants, there is increasing evidence for the role of endocytosis under salinity and osmotic stress (Martín-Davison et al., 2017). The implications of increased endocytosis during these stresses would be a reduction in total plasma-membrane area, thereby limiting water loss from the cell through a decrease in the number of aquaporins. Additionally, it may represent a mechanism to obtain Na^+ ions directly from outside the cell for accumulation in the vacuole, thus keeping the cytoplasmic level of Na^+ low (Baral et al., 2015). In future work, we hope to explore the role of *CaRabC* on endocytosis and Na^+ compartmentalization. There has been very little work published to date concerning

endocytosis and extended drought. The different responses shown in the changes in expression observed in this study between salinity and dehydration (both components of osmotic stress), is intriguing and probably indicative of the underlying biological role of RabC proteins themselves.

Further research is required in several selected chickpea accessions to assess tolerance to salinity, drought and rapid dehydration. This would allow us to explore possible associations between sequence variants and levels of stress tolerance. The genotype-dependent role of each isoform of *CaRabC* as well as other genes from the gene family will be studied, and we plan to carry out these experiments in the near future. These new experiments should elaborate on the mechanism and clarify the suggested roles of these proteins in cell polarization and recycling to the plasma membrane, as suggested by Vernoud et al. (2003) and Rutherford and Moore (2002), respectively. Hopefully, our study of *CaRabC* extends the knowledge of *Rab* gene family structure and function in plants.

AUTHOR CONTRIBUTIONS

GK conducted the experiments with chickpea germplasm and the genotyping with Amplifluor-like SNP analysis. AK and SJ supervised the experiments and interpreted the results. AsZ conducted the experiments with plant stresses and sampling. AyZ carried out sequencing. AT worked with plants in the field trial. BA coordinated the experiments in the field. SL analyzed gene sequences in databases and wrote the corresponding sections. CS analyzed the results, and revised and edited the manuscript. CJ analyzed the qRT-PCR data and revised the corresponding section. KS coordinated the qRT-PCR study and revised other sections. PL supervised the study and revised the final version of the manuscript. YS coordinated all experiments and wrote the first version of the manuscript.

FUNDING

This study was supported by the Ministry of Education and Science (Kazakhstan), Research Program BR05236500 (SJ).

ACKNOWLEDGMENTS

We would like to thank the staff and students of S. Seifullin Kazakh AgroTechnical University, Astana (Kazakhstan) and Flinders University, SA (Australia) for their support in this research and help with critical comments to the manuscript. The results of this study were presented at the International Conference 'Bioinformatics and Computational Biology', August 2018, Novosibirsk, Russia. The Authors acknowledge the Organizing Committee for their support in the presentation and publication of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00040/full#supplementary-material

REFERENCES

Agarwal, P., Reddy, M. K., Sopory, S. K., and Agarwal, P. K. (2009). Plant Rabs: characterization, functional diversity, and role in stress tolerance. *Plant Mol. Biol. Rep.* 27, 417–430. doi: 10.1007/s11105-009-0100-9

Agarwal, P. K., Agarwal, P., Jain, P., Jha, B., Reddy, M. K., and Sopory, S. K. (2008). Constitutive overexpression of a stress-inducible small GTP-binding protein PgRab7 from *Pennisetum glaucum* enhances abiotic stress tolerance in transgenic tobacco. *Plant Cell Rep.* 27, 105–115. doi: 10.1007/s00299-007-0446-0

Alory, C., and Balch, W. E. (2003). Molecular evolution of the Rab-escort-protein/guanine-nucleotide-dissociation-inhibitor superfamily. *Mol. Biol. Cell* 14, 3857–3867. doi: 10.1091/E03-04-0227

Baral, A., Shruthi, K. S., and Mathew, M. K. (2015). Vesicular trafficking and salinity responses in plants. *IUBMB Life* 67, 677–686. doi: 10.1002/iub.1425

Bishop, J. M. (1985). Viral oncogenes. *Cell* 42, 23–38. doi: 10.1016/S0092-8674(85)80098-2

Bolte, S., Schiene, K., and Dietz, K. J. (2000). Characterization of a small GTP-binding protein of the Rab5 family in *Mesembryanthemum crystallinum* with increased level of expression during early salt stress. *Plant Mol. Biol.* 42, 923–936. doi: 10.1023/A:1006449715236

Borg, S., Brandstrup, B., Jensen, T. J., and Poulsen, C. (1997). Identification of new protein species among 33 different small GTP-binding proteins encoded by cDNAs from *Lotus japonicus*, and expression of corresponding mRNAs in developing root nodules. *Plant J.* 11, 237–250. doi: 10.1046/j.1365-313X.1997.11020237.x

Brillada, C., and Rojas-Pierce, M. (2017). Vacuolar trafficking and biogenesis: a maturation in the field. *Curr. Opin. Plant Biol.* 40, 77–81. doi: 10.1016/j.pbi.2017.08.005

Chang, E. H., Gonda, M. A., Ellis, R. W., Scolnick, E. M., and Lowy, D. R. (1982). Human genome contains four genes homologous to transforming genes of harvey and kirsten murine sarcoma viruses (molecular cloning/restriction endonuclease mapping/heteroduplex analysis/intervening sequences/gene family). *Proc. Natl. Acad. Sci. U.S.A.* 79, 4848–4852. doi: 10.1073/pnas.79.16.4848

Chavrier, P., Parton, R. G., Hauri, H. P., Simons, K., and Zerial, M. (1990). Localization of low molecular weight GTP binding proteins to exocytic and endocytic compartments. *Cell* 62, 317–329. doi: 10.1016/0092-8674(90)90369-P

Chen, C., and Heo, J. B. (2018). Overexpression of constitutively active OsRab11 in plants enhances tolerance to high salinity levels. *J. Plant Biol.* 61, 169–176. doi: 10.1007/s12374-018-0048-0

Coffin, J. M., Varmus, H. E., Bishop, J. M., Essex, M., Hardy, W. D., Martin, J. G. S., et al. (1981). Proposal for naming host cell-derived inserts in retrovirus genomes. *J. Virol.* 40, 953–957.

D'Arcy-Lameta, A., Ferrari-Iliou, R., Contour-Ansel, D., Pham-Thi, A. T., and Zuily-Fodil, Y. (2006). Isolation and characterization of four ascorbate peroxidase cDNAs responsive to water deficit in cowpea leaves. *Ann. Bot.* 97, 133–140. doi: 10.1093/aob/mcj010

Flowers, T. J., Glenn, E. P., and Volkov, V. (2018). Could vesicular transport of Na^+ and Cl^- be a feature of salt tolerance in halophytes? *Ann. Bot.* doi: 10.1093/aob/mcy164 [Epub ahead of print].

Garg, R., Sahoo, A., Tyagi, A. K., and Jain, M. (2010). Validation of internal control genes for quantitate gene expression studies in chickpea (*Cicer arietinum* L.). *Biochem. Biophys. Res. Commun.* 396, 283–288. doi: 10.1016/j.bbrc.2010.04.079

Gürel, F., Öztürk, N. Z., Yörük, E., Uçarlı, C., and Poyraz, N. (2016). Comparison of expression patterns of selected drought-responsive genes in barley (*Hordeum vulgare* L.) under shock-dehydration and slow drought treatments. *Plant Growth Regul.* 80, 183–193. doi: 10.1007/s10725-016-0156-0

Hashim, S., Mukherjee, K., Raje, M., Basu, S. K., and Mukhopadhyay, A. (2000). Live *Salmonella* modulate expression of Rab proteins to persist in a specialized compartment and escape transport to lysosomes. *J. Biol. Chem.* 275, 16281–16288. doi: 10.1074/jbc.275.21.16281

Haubruck, H., Disela, C., Wagner, P., and Gallwitz, D. (1987). The *ras*-related *ypt* protein is an ubiquitous eukaryotic protein: isolation and sequence analysis of mouse cDNA clones highly homologous to the yeast *YPTJ* gene. *EMBO J.* 6, 4049–4053. doi: 10.1002/j.1460-2075.1987.tb02750.x

He, M., Lan, M., Zhang, B., Zhou, Y., Wang, Y., Zhu, L. A., et al. (2018). Rab-H1b is essential for trafficking of cellulose synthase and for hypocotyl growth in *Arabidopsis thaliana*. *J. Integr. Plant Biol.* 60, 1051–1069. doi: 10.1111/jipb.12694

Hernández-Sánchez, I. E., Maruri-López, I., Graether, S. P., and Jiménez-Bremont, J. F. (2017). *In vivo* evidence for homo- and heterodimeric interactions of *Arabidopsis thaliana* dehydrins AtCOR47, AtERD10, and AtRAB18. *Sci. Rep.* 7:17036. doi: 10.1038/s41598-017-15986-2

Howlader, J., Park, J. I., Robin, A. H. K., Sumi, K. R., and Nou, I. S. (2017). Identification, characterization and expression profiling of stress-related genes in easter lily (*Lilium formolongi*). *Genes* 8:172. doi: 10.3390/genes8070172

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* 33, 1635–1638. doi: 10.1093/molbev/msw046

Hundertmark, M., and Hincha, D. K. (2008). LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics* 9:118. doi: 10.1186/1471-2164-9-118

Jatayev, S., Kurishbayev, A., Zotova, L., Khasanova, G., Serikbay, D., Zhubatkanov, A., et al. (2017). Advantages of amplifluor-like SNP markers over KASP in plant genotyping. *BMC Plant Biol.* 17:254. doi: 10.1186/s12870-017-1197-x

Jha, Y., Sablok, G., Subbarao, N., Sudhakar, R., Fazil, M. H. U. T., Subramanian, R. B., et al. (2014). Bacterial-induced expression of RAB18 protein in *Oryza sativa* salinity stress and insights into molecular interaction with *GTP* ligand. *J. Mol. Recognit.* 27, 521–527. doi: 10.1002/jmr.2371

Jiang, Z., Wang, H., Zhang, G., Zhao, R., Bie, T., Zhang, R., et al. (2017). Characterization of a small GTP-binding protein gene *TaRab18* from wheat involved in the stripe rust resistance. *Plant Physiol. Biochem.* 113, 40–50. doi: 10.1016/j.plaphy.2017.01.025

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Lång, V., and Palva, T. (1992). The expression of a *rab*-related gene, *rab18*, is induced by abscisic acid during the cold acclimation process of *Arabidopsis thaliana* (L.) *Heynh*. *Plant Mol. Biol.* 20, 951–962. doi: 10.1007/BF00027165

Lawson, T., Mayes, S., Lycett, G. W., and Chin, C. F. (2018). Plant Rabs and the role in fruit ripening. *Biotechnol. Genet. Eng. Rev.* 34, 181–197. doi: 10.1080/02648725.2018.1482092

Liu, Z., Luo, C., Li, L., Dong, L., and Can, V. (2015). Isolation, characterization and expression analysis of the GDP dissociation inhibitor protein gene *MiRab-GDI* from *Mangifera indica* L. *Sci. Hortic.* 185, 14–21. doi: 10.1016/j.scienta.2015.01.008

Ma, Q. H. (2007). Small GTP-binding proteins and their functions in plants. *J. Plant Growth Regul.* 26, 369–388. doi: 10.1007/s00344-007-9022-7

Marcote, M. J., Gu, F., Gruenberg, J., and Aniento, F. (2000). Membrane transport in the endocytic pathway: animal versus plant cells. *Protoplasma* 210, 123–132. doi: 10.1007/BF01276852

Marshall, C. J. (1993). Protein prenylation: a mediator of protein-protein interactions. *Science* 259, 1865–1866. doi: 10.1126/science.8456312

Martín-Davison, A. S., Pérez-Díaz, R., Soto, F., Madrid-Espinoza, J., González-Villanueva, E., Pizarro, L., et al. (2017). Involvement of SchRabGDI1 from *Solanum chilense* in endocytic trafficking and tolerance to salt stress. *Plant Sci.* 263, 1–11. doi: 10.1016/j.plantsci.2017.06.007

Mazel, A., Leshem, Y., Tiwari, B. S., and Levine, A. (2004). Induction of salt and osmotic stress tolerance by overexpression of an intracellular vesicle trafficking protein AtRab7 (AtRabG3e). *Plant Physiol.* 134, 118–128. doi: 10.1104/pp.103.025379

Minamino, N., Kanazawa, T., Era, A., Ebine, K., Nakano, A., and Ueda, T. (2018). RAB GTPases in the basal land plant *Marchantia polymorpha*. *Plant Cell Physiol.* 59, 845–856. doi: 10.1093/pcp/pcy027

Muñoz, F. J., Esteban, R., Labrador, E., and Dopico, B. (2001). Expression of a novel chickpea Rab-GDI cDNA mainly in seedlings. *Plant Physiol. Biochem.* 39, 363–366. doi: 10.1016/S0981-9428(01)01249-9

Nahm, M. Y., Kim, S. W., Yun, D., Lee, S. Y., Cho, M. J., and Bahk, J. D. (2003). Molecular and biochemical analyses of *OsRab7*, a rice *Rab7* homolog. *Plant Cell Physiol.* 44, 1341–1349. doi: 10.1093/pcp/pcg163

O'Mahony, P. J., and Oliver, M. J. (1999). Characterization of a desiccation-responsive small GTP-binding protein (Rab2) from the desiccation-tolerant grass *Sporobolus stapfianus*. *Plant Mol. Biol.* 39, 809–821. doi: 10.1023/A:1006183431854

Peng, X., Ding, X., Chang, T., Wang, Z., Liu, R., Zeng, X., et al. (2014). Overexpression of a vesicle trafficking gene, *OsRab7*, enhances salt tolerance in rice. *Sci. World J.* 214:483526. doi: 10.1155/2014/483526

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* 26, 1641–1650. doi: 10.1093/molbev/msp077

Rajan, N., Agarwal, P., Patel, K., Sanadhya, P., Khedia, J., and Agarwal, P. K. (2015). Molecular characterization and identification of target protein of an important vesicle trafficking gene *AlRab7* from a salt excreting halophyte *Aeluropus lagopoides*. *DNA Cell Biol.* 34, 83–91. doi: 10.1089/dna.2014.2592

Reddy, D. S., Bhatnagar-Mathur, P., Reddy, P. S., Cindhuri, K. S., Ganesh, A. S., and Sharma, K. K. (2016). Identification and validation of reference genes and their impact on normalized gene expression studies across cultivated and wild *Cicer* species. *PLoS One* 11:e0148451. doi: 10.1371/journal.pone.0148451

Rehman, R. U., and Sansebastiano, D. G. P. (2014). "Plant Rab GTPases in membrane trafficking and signalling," in *Plant Signaling: Understanding the Molecular Crosstalk*, eds K. R. Hakeem, R. U. Rehman, and I. Tahir (New Delhi: Springer), 51–73. doi: 10.1007/978-81-322-1542-4-3

Rushton, D. L., Tripathi, P., Rabara, R. C., Lin, J., Ringler, P., Boken, A. K., et al. (2012). WRKY transcription factors: key components in abscisic acid signalling. *Plant Biotechnol. J.* 10, 2–11. doi: 10.1111/j.1467-7652.2011.00634.x

Rutherford, S., and Moore, I. (2002). The *Arabidopsis* Rab GTPase family: another enigma variation. *Curr. Opin. Plant Biol.* 5, 518–528. doi: 10.1016/S1369-5266(02)00307-2

Shavrukov, Y., Bovill, J., Afzal, I., Hayes, J. E., Roy, S. J., Tester, M., et al. (2013). *HVP10* encoding V-PPase is a prime candidate for the barley *HvNax3* sodium exclusion gene: evidence from fine mapping and expression analysis. *Planta* 237, 1111–1122. doi: 10.1007/s00425-012-1827-3

Shavrukov, Y., Zhumalin, A., Serikbay, D., Botayeva, M., Otemisova, A., Absattarova, A., et al. (2016). Expression level of the DREB2-type gene, identified with Amplifluor SNP markers, correlates with performance, and tolerance to dehydration in bread wheat cultivars from Northern Kazakhstan. *Front. Plant Sci.* 7:1736. doi: 10.3389/fpls.2016.01736

Stenmark, H., and Olkkonen, V. M. (2001). The rab GTPase family. *Genome Biol.* 2:3007. doi: 10.1186/gb-2001-2-5-reviews3007

Sui, J. M., Li, G., Chen, G. X., Yu, M. Y., Ding, S. T., Wang, J. S., et al. (2017). Digital expression analysis of the genes associated with salinity resistance after overexpression of a stress-responsive small GTP-binding RabG protein in peanut. *Genet. Mol. Res.* 16:gmr16019432. doi: 10.4238/gmr16019432

Takai, Y., Sasaki, T., and Matozaki, T. (2001). Small GTP-binding proteins. *Physiol. Rev.* 81, 153–208. doi: 10.1152/physrev.2001.81.1.153

Uemura, T., and Ueda, T. (2014). Plant vacuolar trafficking driven by RAB and SNARE proteins. *Curr. Opin. Plant Biol.* 22, 116–121. doi: 10.1016/j.pbi.2014.10.002

Vernoud, V., Horton, A. C., Yang, Z., and Nielsen, E. (2003). Analysis of the small GTPase gene superfamily of *Arabidopsis*. *Plant Physiol.* 131, 1191–1208. doi: 10.1104/pp.013052

Woollard, A. A. D., and Moore, I. (2008). The functions of Rab GTPases in plant membrane traffic. *Curr. Opin. Plant Biol.* 11, 610–619. doi: 10.1016/j.pbi.2008.09.010

Yaneva, I., and Niehaus, K. (2005). Molecular cloning and characterisation of a Rab-binding GDP-dissociation inhibitor from *Medicago truncatula*. *Plant Physiol. Biochem.* 43, 203–212. doi: 10.1016/j.plaphy.2005.01.019

Zerial, M., and McBride, H. (2001). RAB proteins as membrane organizers. *Nat. Rev. Mol. Cell Biol.* 2, 107–119. doi: 10.1038/35052055

Zhang, J., Li, Y., Liu, B., Wang, L., Zhang, L., Hu, J., et al. (2018). Characterization of the *Populus Rab* family genes and the function of *PtRabE1b* in salt tolerance. *BMC Plant Biol.* 18:124. doi: 10.1186/s12870-018-1342-1

Zotova, L., Kurishbayev, A., Jatayev, S., Khassanova, G., Zhubatkanov, A., Serikbay, D., et al. (2018). Genes encoding transcription factors TaDREB5 and TaNFYC-A7 are differentially expressed in leaves of bread wheat in response to drought, dehydration and ABA. *Front. Plant Sci.* 9:1441. doi: 10.3389/fpls.2018.01441

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 February 2019
doi: 10.3389/fgene.2019.00043

Using Ancestry Informative Markers (AIMs) to Detect Fine Structures Within Gorilla Populations

Ranajit Das[1]*, *Ria Roy*[2] *and Neha Venkatesh*[3]

[1] Manipal Centre for Natural Sciences, Manipal Academy of Higher Education, Manipal, India, [2] Department of Biotechnology Engineering, Sahrdaya College of Engineering and Technology, Kodakara, India, [3] Department of Genetics, University of Mysore, Mysore, India

The knowledge of ancestral origin is monumental in conservation of endangered animals since it can aid in preservation of population level genetic integrity and prevent inbreeding among related individuals. Despite maintenance of studbook, the biogeographical affiliation of most captive gorillas is largely unknown, which has constrained management of captive gorillas aiming at maximizing genetic diversity at the population level. In recent years, ancestry informative markers (AIMs) has been successfully employed for the inference of genomic ancestry in a wide range of studies in evolutionary genetics, biomedical research, genetic stock identification, and introgression analysis and forensic analyses. In this study, we sought to derive the AIMs yielding the most cohesive and faithful understanding of biogeographical affiliation of query gorillas. To this end, we compared three commonly used AIMs-determining methods namely, Infocalc, F_{ST}, and Smart Principal Component Analysis (SmartPCA) with ADMIXTURE, using gorilla genome data available through Great Ape Genome Project database. Our findings suggest that the SNPs that were detected by at least three of the four AIMs-determining approaches ($N = 1{,}531$), is likely most suitable for delineation of gorilla AIMs. It recapitulated the finer structure within western lowland gorilla genomes with high degree of precision. We further have validated the robustness of our results using a randomized negative control containing the same number of SNPs. To the best of our knowledge, this is the first report of an AIMs panel for gorillas that may aid in developing cost-effective resources for large-scale demographic analyses, and greatly help in conservation of this charismatic mega-fauna.

Keywords: ancestry informative marker (AIM), gorilla ancestry, conservation genetic management, admixture, informativeness of SNPs

OPEN ACCESS

Edited by:
Yuriy L. Orlov,
Russian Academy of Sciences, Russia

Reviewed by:
GaneshPrasad Arun ArunKumar,
SASTRA University, India
Luciana Werneck Zuccherato,
Universidade Federal de Minas Gerais, Brazil

***Correspondence:**
Ranajit Das
ranajit.das@manipal.edu

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *03 September 2018*
Accepted: *21 January 2019*
Published: *08 February 2019*

Citation:
Das R, Roy R and Venkatesh N (2019) Using Ancestry Informative Markers (AIMs) to Detect Fine Structures Within Gorilla Populations. Front. Genet. 10:43. doi: 10.3389/fgene.2019.00043

BACKGROUND

Effective conservation of endangered animals with unknown ancestral origin entails delineation of the biogeographic affinities of their ancestors in order to facilitate preservation of the population level integrity of genomic signal. The knowledge of ancestral origin could be particularly relevant for planned re-introduction of animals to wild habitats and management of captive breeding programs in order to avoid inbreeding depression.

Gorillas, the largest living ape, were pronounced as critically endangered by IUCN Red List in 2007 (Walsh et al., 2008). Since the gorilla population is rapidly dwindling in the wild as a

result of severe habitat encroachment and the illegal bushmeat trade, effective management of captive breeding programs has become monumental in order to both increase their numbers and to protect them from inbreeding. Overall 283 wild gorillas were imported to North America till 1970s, which subsequently stopped owing to the introduction of Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) in 1975 (Nsubuga et al., 2010). It is noteworthy that despite maintenance of studbooks, insufficient information is available pertaining to the biogeographic origin of the majority of captive gorillas in the USA (Wharton, 2009) and that has likely constrained proper management of captive gorillas pertaining to maximizing genetic diversity at the population level. Proper knowledge of ancestry is of great importance in captive breeding programs of gorillas in order to avoid inbreeding depression and at the same time to conserve the genomic integrity of the native gorilla populations.

While whole genome approaches can efficiently resolve the biogeographical affiliation of gorillas by measuring genomic ancestry and level of admixture occurring among various gorilla populations, it is not cost-effective and dependent on the quality of DNA samples such that lower DNA quality (such as DNA extracted through non-invasive techniques) can hamper genome re-sequencing methods to a considerable extent. An alternative cost-effective strategy to whole genome approaches could be estimation of genomic ancestry using a handful of highly informative Single Nucleotide Polymorphisms (SNPs) which may range from a few hundreds to a few thousands. These highly informative SNPs that exhibit large differences in allele frequencies between ancestral populations are commonly referred to as Ancestry Informative Markers (AIMs) (Rosenberg et al., 2003; Shriver et al., 2003; Nassir et al., 2009).

Over the years AIMs panels have been successfully used for inferring biogeographical ancestry of humans (Rosenberg et al., 2003; Shriver et al., 2003; Kosoy et al., 2009; Nassir et al., 2009; Kidd et al., 2011; Tandon et al., 2011; Galanter et al., 2012; Huckins et al., 2014; Vongpaisarnsin et al., 2015), detection of illegal trade and translocation of wild animals (Frantz et al., 2006), food forensics (Wilkinson et al., 2012), genetic stock identification and introgression analysis (Munoz et al., 2015), forensic analysis (Phillips et al., 2016) to name a few. Recently, 9,000 genetic markers have been identified which are unique to a specific subspecies of chimpanzee and gorilla, and around 40,000 markers have been detected that are specific to each hominoid species or lineage (Hormozdiari et al., 2013).

In this study, we have compared three strategies previously used for AIMs determination, namely Infocalc algorithm (Paschou et al., 2007; Kosoy et al., 2009), Wright's F_{ST} (Tian et al., 2007; Kidd et al., 2011; Nievergelt et al., 2013), Smart Principal Component Analysis (SmartPCA) (Patterson et al., 2006) with a novel ADMIXTURE based approach (Alexander et al., 2009) to interrogate previously published whole genome data of 31 gorillas available in Great Ape Genome Project (GAGP) (Prado-Martinez et al., 2013) corresponding to two subspecies of western gorillas (*Gorilla gorilla*), namely western lowland gorilla (*Gorilla gorilla gorilla*) and Cross River gorilla (*Gorilla gorilla dielhi*), as well as the eastern lowland gorilla (*Gorilla beringei graueri*), to delineate an AIMs panel that can reproducibly capture the genomic ancestry of gorillas at the population level and aid in identification of gorillas at the individual level.

We performed our analysis in three steps. In the first step we evaluated the performance of the four AIMs determining approaches (Wright's F_{ST}, Infocalc, SmartPCA and ADMIXTURE) by comparing them with complete SNP sets (CSS). Subsequently, we developed a consensus dataset, incorporating the SNPs that are common to at least three of the four AIMs-determining strategies. Finally, we developed a negative control dataset (randomly chosen SNPs from CSS) containing the same number of SNPs as the consensus dataset and re-evaluated the performance of the consensus dataset and four AIMs determining approaches. The consideration of the consensus SNPs as the AIMs panel for gorilla was robust since it balanced out the limitations of each individual AIMs determining method and at the same time recapitulated the ancestry information of query gorillas with high precision.

METHODS

Dataset

The dataset employed in this study comprised of 31 gorilla genomes available in GAGP, which overall sequenced 79 great ape individuals to a mean coverage of 25X in an Illumina HiSeq 2000 platform (Prado-Martinez et al., 2013; Das and Upadhayai, 2018): western lowland gorilla (*Gorilla gorilla gorilla*, $N = 27$), eastern lowland gorilla (*Gorilla beringei graueri,* $N = 3$), and Cross River gorilla (*Gorilla gorilla dielhi*, $N = 1$). As indicated previously (Prado-Martinez et al., 2013; Das and Upadhayai, 2018) the western lowland gorilla genomes employed in this study belong to three distinct wild populations: Cameroonian, Congolese, and Equatorial Guinean. The biogeographical origin of the gorilla genomes as mentioned in the Studbook and that predicted through Geographical Population Structure (GPS) algorithm is mentioned in **Supplemental Table 1**. The same dataset comprised of 354,080 markers that has been used recently for tracing ancestry of gorillas (Das and Upadhayai, 2018) was used in this study.

Population Clustering and Admixture Analysis Employing the CSS

Principal component analysis (PCA) was performed in PLINK v1.9 using - -pca command. The ancestry of the gorilla genomes was estimated using unsupervised clustering as implemented in ADMIXTURE v1.3 (Alexander et al., 2009). Similar to our recent study (Das and Upadhayai, 2018), we chose $K = 3$ for all downstream analysis to differentiate the western gorilla genomes into the Congolese and Cameroonian clusters and detection of AIMs for identification of genomic ancestry of gorillas at the population level. PCA and Admixture plots were generated in R v3.2.3.

Determination of AIMs

In order to deduce the SNP markers that are able to infer the genomic ancestry of gorilla samples with accuracy comparable

to that of the CSS of 354,080 SNPs, we evaluated four AIMs determining approaches enumerated below.

1. Infocalc

The first method employed was the Infocalc algorithm (Rosenberg et al., 2003), implemented in Infocalc v1.1, which determines the amount of information multiallelic markers provide regarding an individual's ancestry by calculating the informativeness (I) of each marker individually. Infocalc determines I based on the mathematical expression described previously (Rosenberg et al., 2003):

$$I = \sum_{j=1}^{N} \left(-p_j \log p_j + \sum_{i=1}^{K} \frac{p_{ij}}{K} \log p_{ij} \right)$$

Where, p_j is the mean frequency of allele j over all populations, p_{ij} is the relative frequency of allele j in population i and K is the total number of populations.

We selected the top 10,000 most informative markers from the Infocalc v1.1 output file. Infocalc v1.1 compatible files were generated by using - -structure modifier to the PLINK v1.9 command line. The top 10,000 most informative markers were selected based on the informativeness defining column (I_n) of the output file (**Supplemental Figure S1**).

2. Wright's F_{ST}

F_{ST} (Sewall Wright, 2006) measures the degree of differentiation among populations likely arising due to genetic structure within them. Given a set of populations, PLINK estimated the fixation indices (F_{ST}) separately for all 354,080 markers under evaluation in this study using - -*Fst* command. The Family ID (FID) was used as the indicator of the geographical affinity of the gorilla genomes to different wild populations as mentioned previously (Prado-Martinez et al., 2013) and/or estimated through our recent biogeographical analysis (Das and Upadhayai, 2018).

The 10,000 SNPs with highest F_{ST} values were selected for subsequent analyses (**Supplemental Figure S2**).

3. ADMIXTURE

Analyzing the ADMIXTURE output file with SNP information (P file) for K of 3, we identified 10,662 SNPs with high K (column to column) variance (≥ 0.15).

4. SmartPCA

In order to determine the most informative markers, SNP weightings for each principal component (PC) were calculated using the "SmartPCA" algorithm implemented in EIG v7.2.1 (Patterson et al., 2006; Price et al., 2006). SmartPCA, which is especially designed for analysis of genomic data, employs PCA to determine whether the test samples come from one homogenous population or there is any signature of population structure and outputs principal components (eigenvectors) and eigenvalues. In addition to these two files SmartPCA generates a "snpwt" file, depicting the weight of all 354,080 markers for each principal component.

The 10,000 SNPs with the highest "weights" for the first principal component (PC1) was selected for subsequent analyses (**Supplemental Figure S3**).

Estimation of Candidate AIMs Panels

To determine the optimal AIMs-determining strategy for gorilla genomes, we first compared the datasets comprising of the top 10,000 SNPs generated through F_{ST}, Infocalc, and SmartPCA with 10,662 SNPs detected through ADMIXTURE both qualitatively (via Admixture analysis and PCA) and quantitatively (by computing the Euclidean distances between the admixture components of the query datasets and the CSS).

Further we developed a consensus dataset, containing SNPs that are common to the four AIMs determining strategies (F_{ST}, Infocalc, Admixture, and SmartPCA-based). Here, we note that only 37 SNPs were found to be common to all four approaches evaluated in this study, which was insufficient to recapitulate intraspecific ancestry information of the query gorillas (data not shown). So, in order to generate a consensus SNP panel that is likely to be sufficient to detect the fine structure within western gorilla populations, we developed a dataset comprising of 1,531 SNPs that were common to at least three of the four AIMs-determining methods (**Supplemental Figure S4**). Finally, to adjudge the predictive accuracy of the candidate AIMs datasets, we developed a negative control dataset by randomly sampling 1,531 SNPs from CSS and compared this with those comprising of the top 1,531 SNPs extracted through F_{ST}, Infocalc, Admixture, SmartPCA-based methods and the consensus.

RESULTS

ADMIXTURE Analyses

Qualitative Analysis

The ancestry of 31 gorilla genomes was estimated using unsupervised clustering as implemented in ADMIXTURE v1.3 (Alexander et al., 2009). For CSS, at $K = 3$ the eastern lowland gorillas were homogeneously assigned to a unique cluster (blue) while most western gorillas appeared to be a genomic admixture of Cameroonian (green) and Congolese (red) components in varying proportions (**Figure 1A**, **Supplemental Figure S5A**). While the entire genome of Akiba-Beri, Choomba, Paki, Oko, Kolo and Amani is consisted of the Cameroonian admixture component, Katie (B650) and Katie (KB4986) also appeared to be pure-bred and their genome is entirely composed of the Congolese admixture component.

At $K = 3$, the dataset comprising of the top 10,000 Inforcalc SNPs (Infocalc-10,000) performed the best by successfully and precisely capturing the population structure of gorilla genomes as depicted by the CSS. It homogenously assigned Akiba-Beri, Choomba, Paki, Oko, Kolo and Amani to Cameroon and the Katies (B650 and KB4986) to Congo. Further, similar to the CSS, this dataset revealed fractions of eastern lowland ancestry (blue) in Kokomo, Mimi, Delphi, Coco, Carolyn, and Porta. However, unlike the CSS, Infocalc-10,000 revealed minor fractions of (<1%) eastern lowland ancestry in Kowali and Azizi (**Figure 1B**, **Supplemental Figure S5B**).

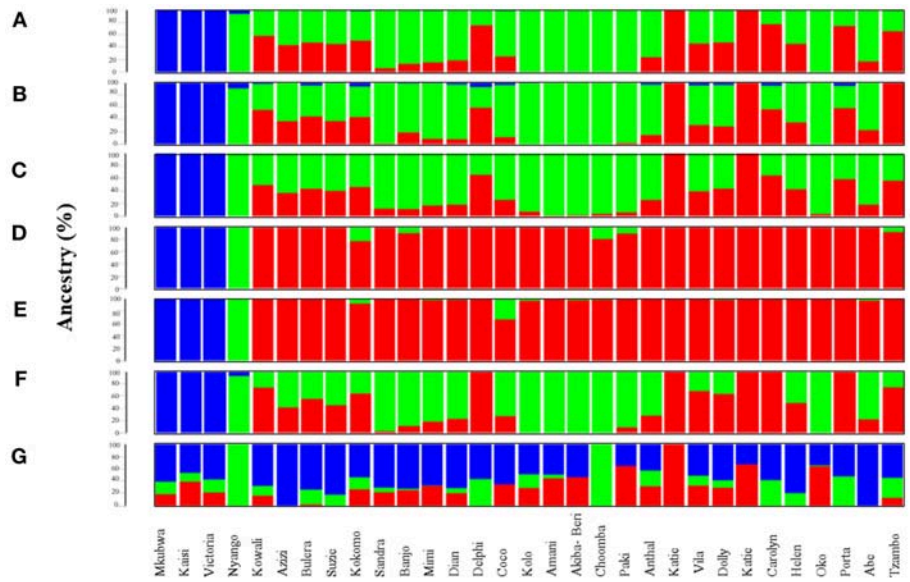


FIGURE 1 | Admixture analysis of data subsets generated through top 1,531 most informative SNPs detected by various AIMs-determining strategies. Admixture plots showing the ancestry components of gorilla genomes. **(A)** Admixture analysis of the CSS (354,080 SNPs); **(B)** Admixture analysis of Infocalc-1,531 dataset; **(C)** Admixture analysis of Admixture-1,531 dataset; **(D)** Admixture analysis of SmartPCA-1,531 dataset; **(E)** Admixture analysis of F_{ST}-1,531 dataset; **(F)** Admixture analysis of Consensus-1,531 dataset; and **(G)** Admixture analysis of Random-1,531 dataset. Admixture proportions were generated through an unsupervised admixture analysis at $K = 3$ using ADMIXTURE v1.3 and plotted in R v3.2.3. Each individual is represented by a vertical line partitioned into colored segments whose lengths are proportional to the contributions of the ancestral components to the genome of the individual. Blue represents eastern lowland ancestry component while green and red represent Cameroonian and Congolese ancestral components, respectively.

The dataset comprising of the top 10,662 Admixture SNPs (Admixture-10,000) appeared to be the second best. In concordance with CSS, Admixture-10,000 homogenously assigned Akiba-Beri, Choomba, Oko and Amani to Cameroon and the Katies (B650 and KB4986) to Congo. However, unlike the CSS, this dataset depicted ∼2, 3, and 4% Congolese ancestral component in the cross river gorilla Nyango, Kolo and Paki, respectively, and eastern lowland ancestral component in Helen and Anthal, which can be attributed to the likely loss of resolution (**Supplemental Figure S5C**).

The remaining two datasets, comprising of 10,000 SNPs generated using SmartPCA and F_{ST}-based approaches (SmartPCA-10,000 and F_{ST}-10,000, respectively), performed moderately. While SmartPCA-10,000 successfully homogenously assigned Akiba-Beri, Choomba, Paki, Oko, Kolo and Amani to Cameroon and the Katies (B650 and KB4986) to Congo, it additionally assigned Delphi, Carolyn and Porta homogenously to Congo and thus failed to capture their discernible proportions of Cameroonian ancestry (**Supplemental Figure S5D**). Among the four approaches, F_{ST}-10,000 performed the worst. In addition to incorrectly assigning Delphi, Carolyn and Porta homogenously to Congo, F_{ST}-10,000 revealed Congolese ancestry in Kolo, Akiba-Beri and Paki, which were otherwise homogenously assigned to Cameroon by all AIMs-determining approaches (**Supplemental Figure S5E**).

Among datasets comprising of top 1,531 SNPs deduced via F_{ST}, Infocalc, Admixture, and SmartPCA, the 1,531 SNPs derived using Infocalc (Infocalc-1,531) was superior to the rest and most comparable to the CSS in recapitulating the population structure for query gorillas (**Figure 1B**). This was closely followed by a panel of 1,531 SNPs generated as a consensus of at least three of the four AIMs-determining strategies (Consensus-1,531) (**Figure 1F**), and that were detected using Admixture (Admixture-1,531) (**Figure 1C**). Here we note that among all 1,531 datasets, only Consensus-1,531 and Infocalc-1,531 were the only two who could capture the eastern lowland ancestry in the cross river gorilla, Nyango, as revealed by the CSS. In contrast, the SNP panel inferred using SmartPCA (SmartPCA-1,531) and F_{ST} (F_{ST}-1,531) completely failed to capture the population structure revealed by the CSS (**Figures 1D,E**). Finally, the negative control dataset comprising of 1,531 random SNPs (Random-1,531) was expectedly unsuccessful in capturing the ancestry information of the query gorillas, underscoring the superiority of the AIMs over randomly selected markers in delineating ancestry information (**Figure 1G**).

Quantitative Analysis

For comparing the test datasets quantitatively, we computed Euclidean distances between the three admixture components (eastern lowland, Cameroonian and Congolese) of all datasets and the CSS. The shortest mean Euclidean distance ($\mu = 0.022$) was found between Admixture-10,000 and the CSS, closely followed by Infocalc-10,000 and the CSS ($\mu = 0.064$) (**Figure 2**). Among other 10,000 SNP panels, the longest Euclidean distance was found between the CSS and F_{ST}-10,000, followed by the CSS and SmartPCA-10,000 (0.154 and 0.108, respectively).

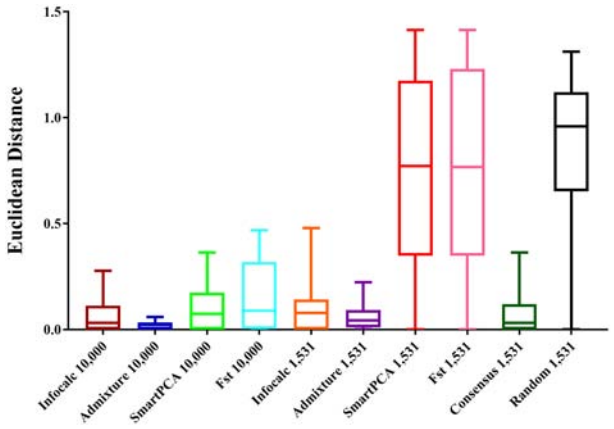


FIGURE 2 | Box and whisker plots comparing the Euclidean distances between the admixture proportions of the gorilla genomes obtained from the CSS and those obtained from the reduced datasets. The number of SNPs the datasets are comprised of is mentioned in their nomenclature. The random dataset was comprised of 1,531 randomly generated SNPs from the CSS and the Consensus-1,531 dataset comprised of 1,531 SNPs that were detected by at least three of the four AIMs-determining approaches.

Among the 1,531 panels, the shortest distance was revealed between Admixture-1,531 and the CSS ($\mu = 0.059$). Consensus-1,531 appeared as the second most sensitive approach ($\mu = 0.087$), closely followed by Infocalc-1,531 ($\mu = 0.095$). All three aforesaid 1,531 panels highly significantly outperformed all the remaining datasets including the random dataset (Tukey's *post hoc* test; p-value < 0.0001). Congruent with our results from qualitative analyses in their inability to capture the accurate population structure for query gorilla genomes, the SmartPCA and F_{ST}-based datasets appeared to be the farthest from the CSS ($\mu = 0.75$ in both cases) and performed similar to the Random-1,531 dataset (Tukey's *post hoc* test; p-value $= 0.94$ and 0.95, respectively). Here further we note that, although Admixture-1,531 had the shortest mean Euclidean distance from the CSS, its performance was statistically very similar to Consensus-1,531 and Infocalc-1,531 (Tukey's *post hoc* test; p-value $= 0.99$).

Overall, our result indicates that while Infocalc-1,531 turned out to be the best method in qualitative ADMIXTURE analysis, Admixture-1,531 was superior to all other approaches in the quantitative analysis. However, in both cases, Consensus-1,531 was a close second and its performance was statistically similar to the other two. Additionally, Consensus-1,531 had discernibly smaller median Euclidean distance from the CSS (0.032) compared to both Infocalc-1,531 (0.078) and Admixture-1,531 (0.043) which further advocates for its candidacy to be considered as the AIMs panel for the gorillas.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was performed in PLINK v1.9 and the top two PCs were plotted in R v3.2.3. The PCA results for the CSS was in coherence with previous observations of an eastern gorilla-western gorilla contrast along the horizontal principal component (PC1) and vertical delineation (PC2) among western gorilla genomes (Prado-Martinez et al., 2013; Das and Upadhayai, 2018)(**Figure 3A**, **Supplemental Figure S6A**). Further, as observed previously, two distinct clusters were found among western gorillas along PC1: one predominantly composed of Cameroonian gorillas and the other predominantly of Congolese gorillas. Also, as found previously, Coco, the only Equatorial Guinea gorilla employed in our study clustered with the Cameroonian gorillas owing to its genomic proximity to the latter (Das and Upadhayai, 2018).

Similar to ADMIXTURE analysis, Infocalc-10,000 (**Supplemental Figure S6B**) and Admixture-10,000 (**Supplemental Figure S6C**) best replicated the population clusters depicted by CSS-based dataset (**Supplemental Figure S6A**) with high precision. Both datasets successfully recapitulated the overlap of some of the Cameroonian and Congolese gorillas at the center of PC2 and the genomic proximity of the cross river gorilla Nyango to Cameroonian gorillas. Among the remaining datasets, SmartPCA-10,000 could recapitulate the overlap of Cameroonian and Congolese gorillas along PC2, but it failed to recapture the high genomic proximity of Nyango with Cameroonian gorillas as depicted by the CSS (**Supplemental Figure S6D**). Finally, F_{ST}-10,000 portrayed two distinct clusters of Cameroonian and Congolese gorillas and failed to replicate the overlap of some of the Cameroonian and Congolese gorillas at the center of the vertical principal component (PC2) (**Supplemental Figure S6E**).

Among the 1,531 SNP panels, Infocalc-1,531 was superior to all other AIMs-determining strategies in replicating the population structure of query gorillas depicted by the CSS (**Figure 3B**). Coherent with the ADMIXTURE analysis, Consensus-1,531 turned out to be the second best (**Figure 3F**), followed by Admixture-1,531 (**Figure 3C**). Among the remaining datasets, SmartPCA-1,531 and F_{ST}-1,531 performed discernibly worse and completely failed to depict any contrast among the western gorilla genomes along PC2 (**Figures 3D,E**). Finally, in concordance with the ADMIXTURE analysis, Random-1,531 was completely unsuccessful in capturing population structure of all query gorillas, such that it even failed to depict the eastern gorilla-western gorilla contrast along the horizontal principal component (PC1) (**Figure 3G**). The failure of the random dataset once again underscored the superiority of the AIMs over randomly selected markers in portraying population structure of query genomes.

Taking together all analyses, our study revealed that while Infocalc performed better than other approaches in qualitative analysis, the Admixture-based approach turned out to be the best in the quantitative analysis. This indicates that no single AIMs-determining strategy may be sufficient to recapitulate the ancestry information of gorillas. So, we propose that Consensus-1,531 which performed consistently well in both qualitative and quantitative analysis (ranked 2nd in both) should be elucidated as the AIMs panel for the gorillas as it emerged as the smallest set of SNPs that delineates the ancestry information and population structure of gorillas with optimum precision. Further, we have generated a set of 262 most informative SNPs from the 1,531 AIMs panel, which can be detected through common genotyping

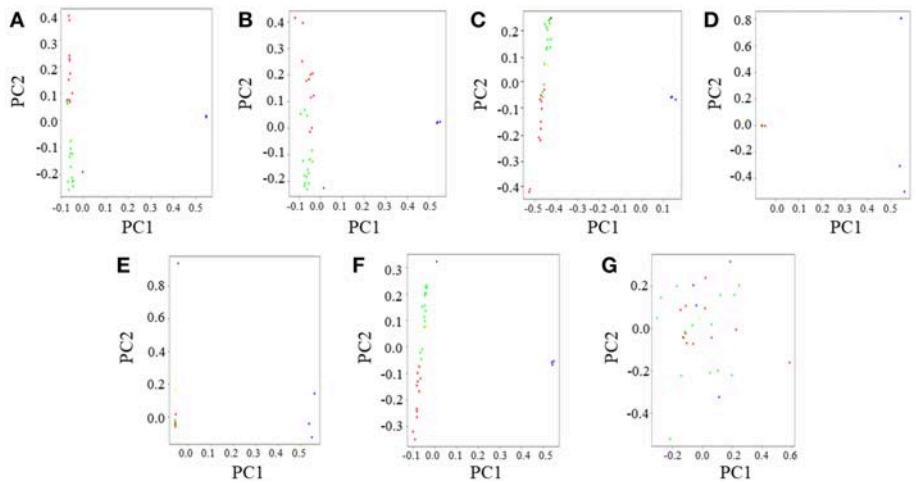


FIGURE 3 | Principal Component Analysis (PCA) of gorilla genomes. PCA plots showing genetic differentiation among query gorilla genomes. The data subsets were generated using top 1,531 most informative SNPs detected through various AIMs determining approaches. **(A)** PCA of the CSS (354,080 SNPs); Here, the X-axis (PC1) explained 45% variance while the Y-axis (PC2) explained 23% variance of the data. **(B)** PCA of Infocalc-1,531; In this case, the X-axis (PC1) explained 45% variance while the Y-axis (PC2) explained only 21% variance of the data. **(C)** PCA of Admixture-1,531; In this case, the X-axis (PC1) explained 68% variance while the Y-axis (PC2) explained 22% variance of the data. **(D)** PCA of SmartPCA-1,531; Here, the X-axis (PC1) explained 88% variance while the Y-axis (PC2) explained only 6% variance of the data. **(E)** PCA of F_{ST}-1,531; In this case, the X-axis (PC1) explained 85% variance while the Y-axis (PC2) explained only 5% variance of the data. **(F)** PCA of Consensus-1,531; In this case, the X-axis (PC1) explained 82% variance while the Y-axis (PC2) explained 10% variance of the data. **(G)** PCA of Random-1,531; Here, the X-axis (PC1) explained 28% variance while the Y-axis (PC2) explained 24% variance of the data. Notable populations are marked with circles such that the blue circles represent eastern lowland gorillas; brown represents the cross river gorilla; and green, red and yellow represents western lowland gorillas of Cameroonian, Congolese and Equatorial Guinean ancestry, respectively. In all cases, PCA was performed in PLINK v1.9 and the top four principal components (PCs) were extracted. Top two PCs (PC1 and PC2), explaining the highest variance of the data were plotted in R v3.2.3.

techniques and are powerful enough to detect fine structure within gorilla populations (**Supplemental Table 2**).

DISCUSSION

Over the years, Gorillas, with dwindling population size and increasingly reduced and restricted distribution in the wild, are faced with serious threats for their survival. As a consequence, conservation of wild as well as captive gorillas and preservation of unique gorilla gene pools has garnered a lot of attention in recent years. The gorilla breeding programs that affords to increase genetic diversity in order to avoid inbreeding depression, have been restricted by insufficient information about the ancestry of the gorillas (Wharton, 2009; Nsubuga et al., 2010; Simons et al., 2012; Prado-Martinez et al., 2013). Hence, the determination of the biogeographical affiliation of gorillas can be invaluable to foster their population level (intra-specific) management and preservation of unique gorilla gene pools.

In this study we sought to compare three strategies previously used for AIMs determination, namely Infocalc algorithm (Paschou et al., 2007; Kosoy et al., 2009), Wright's F_{ST} (Tian et al., 2007; Kidd et al., 2011; Nievergelt et al., 2013), and Smart Principal Component Analysis (SmartPCA) (Patterson et al., 2006) with a novel ADMIXTURE based approach (Alexander et al., 2009) to delineate an AIMs panel that can reproducibly capture the genomic ancestry of gorillas at the population level and aid in identification of gorillas at the individual level. To this end, we developed the first AIMs panel for gorillas containing 1,531 SNPs that were common to at least three out of four AIMs-determining approaches. Our results indicate that this AIMs panel can recapitulate the ancestry information of query gorillas with high precision and can help in population level identification of gorillas, which can be monumental in the preservation of unique gorilla gene pools and selection of individuals for captive breeding program.

Our AIMs panel (Consensus-1,531) consisted of 1,531 SNPs, generated as a consensus of at least three of the four aforesaid AIMs-determining strategies and thus likely balanced out the limitations of each individual approach (Wilkinson et al., 2011). Here we note that out of 1,531 SNPs, 1,359 SNPs were common among F_{ST}, ADMIXTURE and SmartPCA and were not detected by the Infocalc based method (**Figure 2**). The great extent of overlap of top-ranked AIMs of the aforementioned strategies indicates that these three strategies essentially captured the same information regarding the ancestry of query gorillas. Further, while the two worst performing approaches-SmartPCA and F_{ST} revealed the highest number of overlapping SNPs (>26%), Infocalc generated the highest number of exclusive SNPs (94%), followed by ADMIXTURE (66%). These results indicates a likely relationship between the exclusiveness of a SNP and its ability to recapture the ancestry information.

Overall, our qualitative and quantitative analyses concur that Consensus-1,531 could recapitulate the ancestry information of query gorillas with high precision. While Consensus-1,531 had the shortest median Euclidean distance from the CSS (0.032), it appeared as the second most sensitive approach in terms of the

mean Euclidean distance from the same ($\mu = 0.087$) indicating its high precision of recapitulating the ancestral information depicted by the whole dataset. Further, quantitative assessment reflected that the performance of Consensus-1,531 was indistinct from the larger 10,000 SNP based datasets (p-value > 0.99) and had the highest number of individuals ($N = 9$) with zero Euclidean distances from the CSS. However, we note that while Consensus-1,531 successfully replicated the ancestry information of most query gorillas employed in this study, it failed to capture the Cameroonian ancestry component for Carolyn, Delphi and Porta and homogenously assigned them to Congo (**Figure 1**) and thus appeared to be the second-most sensitive in the qualitative assessment, falling short of the number matched Infocalc derived panel.

Amidst the remaining approaches, we note that F_{ST} was the poorest in capturing fine-scale population structure of query gorillas, closely followed by the SmartPCA based approach (**Figures 1–3**), suggesting the ineffectiveness of these two strategies in recapitulating the ancestral history of gorillas. We further note that most AIMs determining approaches employed in this study (except F_{ST}, and SmartPCA) and their consensus appeared to be superior to the randomly selected markers in capturing the population structure delineated by the CSS (**Figures 1–3**), advocating the usefulness of AIMs in tracing biogeographical origin of organisms over randomized SNPs.

Here we note that the goal of this study was to develop AIMs that can be used to tell apart various populations within western lowland gorilla (below subspecies level). Eastern and western lowland gorillas are considered to be different species and are genetically so distinct from each other that they can be differentiated through most markers present in the complete SNP set (CSS). Despite our restriction in terms of sample size and data availability, since most gorilla genomes used in this study belong to various western gorilla populations (27 out of 31), our results should reflect our intended outcome of deducing AIMs that can differentiate western gorillas below subspecies level.

The quest of developing an AIMs panel for gorillas is not new. A previous study has developed polymorphic MEIs, including those that can be considered ancestry-informative markers and MEIs corresponding to regions of incomplete lineage sorting (ILS) (Hormozdiari et al., 2013). However, to the best of our knowledge, this is the first study to have developed an AIMs panel for gorillas, which can recapitulate their ancestry information with high precision. With limited availability of funding, the conservation geneticists need to draw a balance between the costs of genotyping multiple loci and the inadequacy of information when limited number of loci are genotyped. Comprised of only 1,531 SNPs, the gorilla AIMs panel described here, can become a likely cost-effective solution to this problem. Our AIMs panel can resolve the ancestry information of gorillas with highest resolution power and can detect fine structures within gorilla populations below subspecies level at a highly affordable cost.

CONCLUSIONS

Effective conservation of gorilla populations requires the delineation of their ancestry information to facilitate preservation of the population level integrity of genomic signal and avoidance of inbreeding depression. To this end, we have developed an AIMs panel comprising of 1,531 SNPs that can recapitulate the ancestry information of gorillas with high precision. Our AIMs panel can afford a cost-effective solution to whole genome sequencing and/or large-scale genotyping of gorillas for large-scale biogeographic analysis and conservation genetics studies.

To the best of our knowledge this is the first AIMs panel developed for gorillas that can bolster their efficient management and aid in the conservation of their genetic integrity.

AUTHOR CONTRIBUTIONS

RD has conceived the idea of the project, written the manuscript and helped in the analysis. RR and NV performed all the analysis.

FUNDING

This work was supported by Manipal Academy of Higher Education, Manipal, India.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00043/full#supplementary-material

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Das, R., and Upadhayai, P. (2018). Application of the Geographic Population Structure (GPS) algorithm for biogeographical analyses of non-human individuals: a case study of wild and captive gorillas. *BMC Bioinform.* 18:109. doi: 10.1186/s12863-017-0579-2

Frantz, A. C., Pourtois, J. T., Heuertz, M., Schley, L., Flamand, M. C., Krier, A., et al. (2006). Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. *Mol Ecol.* 15, 3191–3203. doi: 10.1111/j.1365-294X.2006.03022.x

Galanter, J. M., Fernandez-Lopez, J. C., Gignoux, C. R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., et al. (2012). Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genet.* 8:e1002554. doi: 10.1371/journal.pgen.1002554

Hormozdiari, F., Konkel, M. K., Prado-Martinez, J., Chiatante, G., Herraez, I. H., Walker, J. A., et al. (2013). Rates and patterns of great ape retrotransposition. *Proc. Natl. Acad. Sci. U.S.A.* 110, 13457–13462. doi: 10.1073/pnas.1310914110

Huckins, L. M., Boraska, V., Franklin, C. S., Floyd, J. A., Southam, L., Gcan, W., et al. (2014). Using ancestry-informative markers to identify fine structure across 15 populations of European origin. *Eur. J. Hum. Genet.* 22, 1190–1200. doi: 10.1038/ejhg.2014.1

Kidd, J. R., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, F. M., and Kidd, K. K. (2011). Analyses of a set of 128 ancestry informative

single-nucleotide polymorphisms in a global set of 119 population samples. *Investig. Genet.* 2:1. doi: 10.1186/2041-2223-2-1

Kosoy, R., Nassir, R., Tian, C., White, P. A., Butler, L. M., Silva, G., et al. (2009). Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum. Mutate.* 30, 69–78. doi: 10.1002/humu.20822

Munoz, I., Henriques, D., Johnston, J. S., Chavez-Galarza, J., Kryger, P., and Pinto, M. A. (2015). Reduced SNP panels for genetic identification and introgression analysis in the dark honey bee (*Apis mellifera mellifera*). *PLoS ONE* 10:e0124365. doi: 10.1371/journal.pone.0124365

Nassir, R., Kosoy, R., Tian, C., White, P. A., Butler, L. M., Silva, G., et al. (2009). An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genet.* 10:39. doi: 10.1186/1471-2156-10-39

Nievergelt, C. M., Maihofer, A. X., Shekhtman, T., Libiger, O., Wang, X., Kidd, K. K., et al. (2013). Inference of human continental origin and admixture proportions using a highly discriminative ancestry informative 41-SNP panel. *Investig. Genet.* 4:13. doi: 10.1186/2041-2223-4-13

Nsubuga, A. M., Holzman, J., Chemnick, L. G., and Ryder, O. A. (2010). The cryptic genetic structure of the North American captive gorilla population. *Conserv. Genet.* 11, 161–172. doi: 10.1007/s10592-009-0015-x

Paschou, P., Ziv, E., Burchard, E. G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. W., et al. (2007). PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.* 3, 1672–1686. doi: 10.1371/journal.pgen.0030160

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2:e190. doi: 10.1371/journal.pgen.0020190

Phillips, C., Santos, C., Fondevila, M., Carracedo, A., and Lareu, M. V. (2016). Inference of ancestry in forensic analysis I: autosomal ancestry-informative marker sets. *Methods Mol. Biol.* 1420, 233–253. doi: 10.1007/978-1-4939-3597-0_18

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., et al. (2013). Great ape genetic diversity and population history. *Nature* 499, 471–475. doi: 10.1038/nature12228

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Rosenberg, N. A., Li, L. M., Ward, R., and Pritchard, J. K. (2003). Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* 73, 1402–1422. doi: 10.1086/380416

Sewall Wright (2006). *Evolutiona and the Genetics of populations: The Theory of Gene Frequencies*. Chicago; IL: The University of Chicago Press. 1969.

Shriver, M. D., Parra, E. J., Dios, S., Bonilla, C., Norton, H., and Jovel, C., et al. (2003). Skin pigmentation, biogeographical ancestry and admixture mapping. *Hum. Genet.* 112, 387–399. doi: 10.1007/s00439-002-0896-y

Simons, N. D., Wagner, R. S., and Lorenz, J. G. (2012). Genetic diversity of North American captive-born gorillas (*Gorilla gorilla gorilla*). *Ecol. Evol.* 3, 80–88. doi: 10.1002/ece3.422

Tandon, A., Patterson, N., and Reich, D. (2011). Ancestry informative marker panels for African Americans based on subsets of commercially available SNP arrays. *Genet. Epidemiol.* 35, 80–83. doi: 10.1002/gepi.20550

Tian, C., Hinds, D. A., Shigeta, R., Adler, S. G., Lee, A., Pahl, M. V., et al. (2007). A genomewide single-nucleotide-polymorphism panel for Mexican American admixture mapping. *Am. J. Hum. Genet.* 80, 1014–1023. doi: 10.1086/513522

Vongpaisarnsin, K., Listman, J. B., Malison, R. T., and Gelernter, J. (2015). Ancestry informative markers for distinguishing between Thai populations based on genome-wide association datasets. *Leg. Med.* 17, 245–250. doi: 10.1016/j.legalmed.2015.02.004

Walsh, P. D., Tutin, C. E. G., Oates, J. F., Baillie, J. E. M., Maisels, F., Stokes, E. J., et al. (2008). Gorilla gorilla. *IUCN 2009 IUCN Red List of Threatened Species Version 20092*.

Wharton, D. (2009). *North American Regional Western Lowland Gorilla Studbook*. Chicago, IL: Chicago Zoological Society.

Wilkinson, S., Archibald, A. L., Haley, C. S., Megens, H. J., Crooijmans, R. P., Groenen, M. A., et al. (2012). Development of a genetic tool for product regulation in the diverse British pig breed market. *BMC Genom.* 13:580. doi: 10.1186/1471-2164-13-580

Wilkinson, S., Wiener, P., Archibald, A. L., Law, A., Schnabel, R. D., McKay, S. D., et al. (2011). Evaluation of approaches for identifying population informative markers from high density SNP chips. *BMC Genet.* 12:45. doi: 10.1186/1471-2156-12-45

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 08 February 2019
doi: 10.3389/fgene.2019.00063

The General Transcription Repressor *TaDr1* Is Co-expressed With *TaVrn1* and *TaFT1* in Bread Wheat Under Drought

*Lyudmila Zotova[1], Akhylbek Kurishbayev[1], Satyvaldy Jatayev[1], Nikolay P. Goncharov[2], Nazgul Shamambayeva[1], Azamat Kashapov[1], Arystan Nuralov[1], Ainur Otemissova[1], Sergey Sereda[3], Vladimir Shvidchenko[1], Sergiy Lopato[4], Carly Schramm[4], Colin Jenkins[4], Kathleen Soole[4], Peter Langridge[5,6] and Yuri Shavrukov[4]**

[1] Faculty of Agronomy, S.Seifullin Kazakh AgroTechnical University, Astana, Kazakhstan, [2] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia, [3] A.F.Khristenko Karaganda Agricultural Experimental Station, Karaganda, Kazakhstan, [4] Biological Sciences, College of Science and Engineering, Flinders University, Bedford Park, SA, Australia, [5] School of Agriculture, Food and Wine, University of Adelaide, Adelaide, SA, Australia, [6] Wheat Initiative, Julius Kühn-Institut, Berlin, Germany

OPEN ACCESS

Edited by:
Yuriy L. Orlov, Russian Academy of Sciences, Russia

Reviewed by:
Yin-Gang Hu, Northwest A&F University, China
Sintho Wahyuning Ardie, Bogor Agricultural University, Indonesia

***Correspondence:**
Yuri Shavrukov
yuri.shavrukov@flinders.edu.au

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *13 November 2018*
Accepted: *24 January 2019*
Published: *08 February 2019*

Citation:
Zotova L, Kurishbayev A, Jatayev S, Goncharov NP, Shamambayeva N, Kashapov A, Nuralov A, Otemissova A, Sereda S, Shvidchenko V, Lopato S, Schramm C, Jenkins C, Soole K, Langridge P and Shavrukov Y (2019) The General Transcription Repressor TaDr1 Is Co-expressed With TaVrn1 and TaFT1 in Bread Wheat Under Drought. Front. Genet. 10:63. doi: 10.3389/fgene.2019.00063

The general transcription repressor, *TaDr1* gene, was identified during screening of a wheat SNP database using the Amplifluor-like SNP marker KATU-W62. Together with two genes described earlier, *TaDr1A* and *TaDr1B*, they represent a set of three homeologous genes in the wheat genome. Under drought, the total expression profiles of all three genes varied between different bread wheat cultivars. Plants of four high-yielding cultivars exposed to drought showed a 2.0–2.4-fold increase in *TaDr1* expression compared to controls. Less strong, but significant 1.3–1.8-fold up-regulation of the *TaDr1* transcript levels was observed in four low-yielding cultivars. *TaVrn1* and *TaFT1*, which controls the transition to flowering, revealed similar profiles of expression as *TaDr1*. Expression levels of all three genes were in good correlation with grain yields of evaluated cultivars growing in the field under water-limited conditions. The results could indicate the involvement of all three genes in the same regulatory pathway, where the general transcription repressor *TaDr1* may control expression of *TaVrn1* and *TaFT1* and, consequently, flowering time. The strength of these genes expression can lead to phenological changes that affect plant productivity and hence explain differences in the adaptation of the examined wheat cultivars to the dry environment of Northern and Central Kazakhstan. The Amplifluor-like SNP marker KATU-W62 used in this work can be applied to the identification of wheat cultivars differing in alleles at the *TaDr1* locus and in screening hybrids.

Keywords: Amplifluor-like SNP marker, bioinformatics, drought, general repressor of transcription, *TaDr1*, *TaFT1*, *TaVrn1*

INTRODUCTION

Amongst the many types of abiotic stresses, drought or water limitation is one of the most important challenges for native plants and crops. There are several genetic and breeding strategies aimed at improving tolerance to drought in crops (Reviewed in: Ingram and Bartels, 1996; Yordanov et al., 2000; Tuberosa and Salvi, 2006; Valliyodan and Nguyen, 2006; Shanker et al., 2014;

Berger et al., 2016; Kaur and Asthir, 2017). One potential approach is the modulation of flowering time, where wheat plants grow faster and complete their life-cycles a few days earlier, therefore minimizing interruption from oncoming, terminal drought (Reviewed in: Shavrukov et al., 2017). Genetic polymorphism and the introgression of novel alleles from wheat progenitors, relatives and wild species from the genus *Triticum* is a very powerful tool to enrich the genome of modern cultivars (Reviewed in: Arzani and Ashraf, 2017; Mwadzingeni et al., 2017; Wang et al., 2018).

Molecular markers are used widely for the identification of novel and existing alleles, and to track specific alelles in elite wheat breeding lines and introgression from landraces or wild species. Analysis of SNP (Single nucleotide polymorphism) is a rapidly developing technology with a diverse range of methods and applications (Reviewed in: Schramm et al., 2019). Amplifluor SNP markers are well-established and have been successfully applied in the recent genotyping of candidate genes for various plant species (Absattar et al., 2018; Yerzhebayeva et al., 2018; Khassanova et al., 2019). This includes research in bread wheat, where alleles of candidate genes for drought tolerance, *TaDREB5* and *TaNFYC-A7*, were identified using Amplifluor SNP markers. These genes demonstrate differential expression in high- and low-yielding wheat cultivars from Kazakhstan under a progressive drought and rapid dehydration (Shavrukov et al., 2016b; Zotova et al., 2018). In other studies, over-expression of transcription factors, *TaNFYA-B1* and *TaNF-YB3;l* showed increased yield and nitrogen uptake, and quicker root development and improved tolerance to drought than controls, respectively (Qu et al., 2015; Yang et al., 2017). Similarly, the rice genes *OsNF-YA7* and *OsNF-YB1* were reported to be responsive to drought. Over-expression of *OsNF-YA7* increased drought tolerance in transgenic rice plants (Lee et al., 2015), and *OsNF-YB1* controls grain filling, resulting in improved yield (Xu et al., 2016).

Transcription factor (TF) Nuclear Factor Y (NF-Y) is a synonym of CCAAT Binding Factor (CBF) and Heme Activator Protein (HAP). Three subunits (A, B, and C) usually function in a single protein complex of NF-Y, and each of the three components is essential for binding to *cis*-elements in the promoter regions of target genes (Siefers et al., 2009; Petroni et al., 2012). In plants, the functions of NF-Y proteins are quite diverse, but, for the purposes of this paper, we will focus on just three: (1) regulation of flowering time; (2) response to abiotic stress, particularly drought; and (3) overall productivity in different plants (Gusmaroli et al., 2001; Nelson et al., 2007; Petroni et al., 2012; Kuromori et al., 2014; Swain et al., 2017; Zhao et al., 2017) including bread wheat (Qu et al., 2015; Yadav et al., 2015; Zotova et al., 2018).

In *Arabidopsis*, the C subunits of NF-Y factor, AtNF-YC3, AtNF-YC4, and AtNF-YC9, are involved in the regulation of photoperiod-mediated flowering time through the GA signaling pathway by binding to RGA (Repressor of ga1-3) and RGL2 (RGA-like2) proteins (Hou et al., 2014; Liu et al., 2016). Over-expression of many individual NF-YC subunits (such as NF-YC1, NF-YC2, NF-YC3, NF-YC4, and NF-YC9) alters flowering time. Individual subunits of the NF-Y complex can affect the transcript levels of *Flowering locus T* (*FT*). This gene encodes the protein that is the key integrator in the flowering time pathway, and up- or down-regulation of FT in interaction with the NF-Y complex, leads to either early or late flowering in *Arabidopsis* (Kumimoto et al., 2010; Cao et al., 2014; Hou et al., 2014; Xu et al., 2016).

The flowering time trait has a complicated, multi-level control. Transcriptional up-regulation of two genes, *Vrn* (Vernalisation) and *FT*, is strongly required for the transition from the vegetative to reproductive stage, largely determining time to flowering (Reviewed in: Greenup et al., 2009; Jung and Müller, 2009; Yan, 2009; Jarillo and Piòeiro, 2011; Song et al., 2013; Milec et al., 2014; Blümel et al., 2015). In wheat, one of the most important crops, the genetic control of the flowering time trait has been extensively studied (Reviewed in: Li and Dubcovsky, 2008; Craufurd and Wheeler, 2009; Distelfeld et al., 2009; Campoli and Korff, 2014; Kamran et al., 2014). The main regulatory control of flowering time in wheat is through the up-regulation of *TaFT1* – *TaVrn3* and *TaVrn1* genes (Li and Dubcovsky, 2008; Distelfeld et al., 2009).

Interestingly, flowering time is controlled not only by genes during ontogenesis, but is strongly impacted by abiotic stresses (Reviewed in: Kazan and Lyons, 2016; Takeno, 2016). Plants of various species have been reported to alter their development and flowering time in response to different types of abiotic stresses, ranging from osmotic stress in *Arabidopsis* (Chen et al., 2007), to soil pH in a native population of *Corydalis sheareri*, Papaveraceae (Huang et al., 2017). However, drought has been shown to be one of the major abiotic factors affecting development of flowering in various plant species such as tea, *Camellia sinensis* (Sharma and Kumar, 2005), litchi, *Litchi chinensis* (Shen et al., 2016) and lemon (Li et al., 2017). The genetic control of reproductive development and time to flowering in response to various abiotic stresses are well studied in cereals (Gol et al., 2017), where the influence of cold (Li et al., 2018) and drought (Pinto et al., 2010; Gudys et al., 2018) in particular, affect grain yields. Early flowering as a drought escape strategy in wheat and other species and was reviewed recently (Shavrukov et al., 2017).

In bread wheat, the *TaVrn1* gene was mapped to the long arm of chromosome 5A, tightly linked with the *Q* gene controlling spike morphology (Kato et al., 1998). The *Q* gene belongs to the large *AP2/ERF* family of TF (Konopatskaia et al., 2016), which includes *DREB* genes responsive to drought and dehydration, and reports have shown that the *Q* gene is also regulated by drought (Gürsoy et al., 2012). Therefore, flowering time and spike morphology seem to have a shared regulatory framework with *TaVrn1* and *Q* genes, and a strong response to drought.

The gene sequence and structure of the general repressor of transcription, *Dr1* (alternative name – *NC2β*), is conserved among various eukaryotes. It operates as a heterodimeric complex with the product of another gene, *DrAP1* (alternative name – *NC2α*), and strongly represses the transcriptional activity of RNA polymerase II and III, but not RNA polymerase I (Kim et al., 1997). Originally, Dr1/DrAp1 was identified in human cells as an unknown factor that was able to inhibit TBP-dependent basal transcription *in vitro* (Inostroza et al., 1992). Mammalian DrAp1 itself cannot repress transcription and therefore it is considered as an enhancer of Dr1 repression activity (Mermelstein et al., 1996; Kim et al., 1997; Yeung et al., 1997). In *Drosophila*, Dr1/DrAp1 represses the transcription

from TATA-containing promoters and activates the transcription from promoters without TATA-boxes (Willy et al., 2000).

In plants, *Dr1* was originally discovered in *Arabidopsis* (Kuromori and Yamamoto, 1994). Later, the rice *OsDr1* and *OsDrAp1* genes were cloned, and formation of the heterodimeric complex, interaction of the protein complex with DNA, and repressive activities of the subunits and protein complex were characterized using the Y2H system, *in vitro* methods, and a transient expression assay (Song et al., 2002). These authors demonstrated several differences between the properties of Dr1 and DrAp1 in mammals and rice. Firstly, the plant DrAp1 protein was found to be larger than the mammalian and yeast proteins, and both plant Dr1 and DrAp1 contained a greater number of domains/motifs than their mammalian counterparts. Secondly, OsDrAp1 alone showed stronger repression activity than OsDr1, therefore in plants, OsDr1 most likely plays the co-repressor role and enhances the activity of OsDrAp1 (Song et al., 2002).This differs from mammals and yeast, where Dr1 is the repressor and DrAp1 plays the role of a regulatory subunit (Inostroza et al., 1992; Kim et al., 1997; Prelich, 1997).

Two homologs *Dr1* genes from bread wheat, *TaDr1A* and *TaDr1B*, were identified and their expression patterns were reported in different wheat tissues under control and drought conditions (Stephenson et al., 2007). Transcripts of both *TaDr1* homologs were abundant in all tested plant tissues and strongly up-regulated in leaves under drought.

In yeast, a 71% similarity between Dr1 and CBF-A (=NF-YB) was reported (Sinha et al., 1996). In bread wheat, TaDr1 and TaDr2 proteins (accessions AF464903 and BT009234, respectively), showed a "high degree of similarity" with TaNF-YB3 amino acid residues (Stephenson et al., 2007). Therefore, the authors suggested that the Dr1/DrAp1 complex could, potentially, inhibit transcription by acting as antagonist to all or to particular NF-YB and NF-YC subunits, thus preventing subunit association and subsequent binding of the activation NF-Y complex (Stephenson et al., 2007). This could be a possible mechanism to explain *TaDr*-mediated global repression of transcription.

The aims of this work were: (1) to compare flowering time and time to grain maturity of high-yielding and low-yielding wheat cultivars from Kazakhstan; (2) to analyze the genetic polymorphism of the *TaDr1* gene in eight selected bread wheat cultivars, and in an F_3 segregating population 18-6 originating from a complex interspecies hybridisation; (3) to study *TaDr1, TaVrn1* and *TaFT1* gene expression in response to drought in leaves of selected wheat cultivars; and (4) to assess the co-expression of *TaDr1, TaVrn1,* and *TaFT1* genes and grain yields of wheat cultivars in the dry conditions of Northern and Central Kazakhstan.

MATERIALS AND METHODS

Plant Material, Conditions of Plant Growth and Drought Application

Eight wheat cultivars, representing two groups with contrasting yields were selected from local varieties tested in field trials, based on their grain yields under the dry conditions in Northern Kazakhstan (current study) and Central Kazakhstan, described earlier by Shavrukov et al. (2016b). Descriptions of plant materials and all experiments were as reported earlier (Zotova et al., 2018). These descriptions included: seeds obtained, conditions of plant growth in the research field in Central Kazakhstan and the controlled conditions in the "Phytotron" experiments on gradual drought using plants in soil-filled containers over 12 days (Experiment 1) (Zotova et al., 2018).

A small outdoor trial was conducted in the research field of S.Seifullin Kazakh AgroTechnical University, Astana in Northern Kazakhstan in the dry season of 2017. Total rainfall was 107 mm during the vegetative growth period, lower than the average of 166 mm that was observed over many years in this region, and a 3°C higher than average temperature for August (20.3°C compared to the average, 17.3°C) was recorded that year. Two-row plots were sown, 1 m in length with 5 cm between plants in rows and 20 cm between rows, and four randomized replicates were used. The number of days between sowing and first flowering of 50% of plants in each plot was counted as "Days to flowering" (DF), while "Days to maturity" (DM) was recorded when all plants in each plot reached the ripening stage. Grain yield was measured for each plot and re-calculated in "g/m^2" with statistical treatment as described below.

A complex interspecific cross [♀ *Triticum spelta*, k-53660 × ♂ (*T. aestivum*, Novosibirskaya 67 / T. *dicoccum*, k-25516)] was produced by one of the authors, Nikolay Goncharov, at the Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk (Russia). F_3 plants from the hybridisation were grown in pots with soil in a "Phytotron" with controlled conditions as mentioned above.

Identification of the "Gene of Interest" Using Bioinformatics and Molecular Phylogenetic Comparative Analysis

The cereals SNP database[1] was used to search and select a single target gene or "Gene of Interest" (GoI) for further research. BLAST analysis of the genetic fragments containing a SNP was applied to identify the full-length GoI using the Nucleotide collection of bread wheat in the NCBI database[2].

Bioinformatics and systems biology methods were applied in this study to identify the full-length nucleotide sequence of the GoI, *TaDr1*, and its corresponding polypeptide sequence was used for both BLASTN and BLASTP in NCBI and in GenomeNet Database Resources, Kyoto University, Japan[3]. All wheat gene sequences with KEGG identification and their encoded proteins were retrieved from GenomeNet databases. Multiple sequence alignments of nucleotide sequences for the *TaDr1A* and *TaDr1B* genes were conducted in CLUSTALW using the CLC Main Workbench software[4].

[1]http://www.cerealsdb.uk.net/cerealgenomics/CerealsDB

[2]https://blast.ncbi.nlm.nih.gov

[3]https://www.genome.jp/tools/blast

[4]https://www.qiagenbioinformatics.com/products/clc-main-workbench

Chromosome locations of all *TaDr1* homeologous genes in the wheat genome were found using BLAST analysis with high confidence annotated genes of the IWGSC database at the Gramene web-site[5].

The molecular dendrogram of polypeptides of *TaDr1* from bread wheat and other monocot plants was constructed using SplitsTree4 program[6] (Huson and Bryant, 2006), with Phylogram Splits and Tree Selector option.

DNA Extraction and SNP Amplifluor Analysis

Plants were grown in control (non-stressed) conditions in containers with soil as described above. Five uniform, 1 month-old individual plants were selected from each accession and five leaves were collected and bulked for leaf samples. Leaf samples frozen in liquid nitrogen were ground in 10-ml tubes with two 9-mm stainless ball bearings using a Vortex mixer. DNA was extracted from the bulked leaves with phenol-chloroform as described in our earlier papers (Shavrukov et al., 2016b; Zotova et al., 2018). 1 μl of DNA was loaded on a 0.8% agarose gel to assess quality, and concentration was measured by Nano-Drop (ThermoFisher, United States).

Amplifluor-like SNP analysis was carried out using a QuantStudio-7 Real-Time PCR instrument (ThermoFisher Scientific, United States) as described previously (Jatayev et al., 2017; Zotova et al., 2018) with the following adjustment for wheat genotyping. Each reaction contained 3 μl of template DNA adjusted to 20 ng/μl, 5 μl of Hot-Start 2xBioMaster (MH020-400, Biolabmix, Novosibirsk, Russia[7]) with all other components as recommended by the manufacturers, including $MgCl_2$ (2.0 mM). One μl of the two fluorescently labeled Universal probes was added (0.125 μM each) and 1 μl of allele-specific primer mix (0.075 μM of each of two forward primers and 0.39 μM of the common reverse primer). 4 μl of Low ROX (ThermoFisher, United States) was added as a passive reference label to the Master-mix as prescribed for the qPCR instrument. Assays were performed in 96-well microplates. The annotated SNP sites were used to design allele-specific primers. Sequences of the Universal probes and primers and sizes of amplicons generated are presented in **Supplementary Material 1**.

PCR was conducted using a program adjusted from those published earlier (Jatayev et al., 2017; Zotova et al., 2018): initial denaturation, 95°C, 2 min; 20 "doubled" cycles of 95°C for 10 s, 60°C for 10 s, 72°C for 20 s, 95°C for 10 s, 55°C for 20 s and 72°C for 50 s; with recording of Allele-specific fluorescence after each cycle. Genotyping by SNP calling was determined automatically by the instrument software, but each SNP result was also checked manually using amplification curves and final allele discrimination. Experiments were repeated twice over different days, where two technical replicates confirmed the confidence of SNP calls.

RNA Extraction, cDNA Synthesis and qPCR Analysis

Plants were grown in the controlled conditions of a "Phytotron" at S.Seifullin Kazakh AgroTechnical University, Astana, Kazakhstan, as described earlier in Experiment 1 (Zotova et al., 2018). In brief, for mild drought stress with 1-month old plants, watering was withdrawn in one of two soil-filled containers for 12 days until wilted leaves were observed. Control plants in similar containers were watered continuously. Five individual plants were used for each cultivar in drought-affected and well-watered containers. All leaves were collected from each plant in plastic tubes as separate biological replicates, frozen immediately in liquid nitrogen and kept at −80°C until RNA extraction. Three samples were used for RNA extraction in each cultivar and treatment, while two additional samples were used as replacements in case of failed extraction or poor RNA quality.

Frozen leaf samples were ground as described above for DNA extraction. TRIzol-like reagent was used for RNA extraction following the protocol described by Shavrukov et al. (2013) and all other steps for RNA extraction and cDNA synthesis were as described previously (Zotova et al., 2018) including DNase treatment (Qiagen, Germany), and the use of a MoMLV Reverse Transcriptase kit (Biolabmix, Novosibirsk, Russia). The quality of all cDNA samples was confirmed by PCR with products of the expected size.

Samples of cDNA diluted with water (1:2) were used for qPCR analyses using both a QuantStudio-7 Real-Time PCR instrument (ThermoFisher Scientific, United States) at Kazakh AgroTechnical University, Astana, Kazakhstan, and Real-Time qPCR system, Model CFX96 (BioRad, Gladesville, NSW, Australia) at Flinders University, Australia. Similar qPCR protocols were used in both instruments, as described earlier (Shavrukov et al., 2016b). Differences between protocols were: the total volume of 10 μl q-PCR reactions included either 5 μl of 2xBiomaster HS-qPCR SYBR Blue (Biolabmix, Novosibirsk, Russia) for experiments in Kazakhstan or 5 μl of 2xKAPA SYBR FAST (KAPA Biosystems, United States) for experiments in Australia, 4 μl of diluted cDNA, and 1 μl of two gene-specific primers (3 μM of each primer) (**Supplementary Material 2**). Expression data for the target genes were calculated relative to the average expression of the two reference genes: *Ta22845*, ATP-dependent 26S proteasome and *Ta54825*, actin (Paolacci et al., 2009). At least three biological and two technical replicates were used in each qPCR experiment.

Statistical Analysis

IBM SPSS Statistical software was used to calculate and analyze means and standard error using ANOVA, to estimate the probabilities for significance using Student's *t*-test. A correlation analysis was performed using Tests of Between-Subjects Effects (IBM SPSS, Statistics Desktop 25.0.0.0).

[5]http://www.gramene.org

[6]http://www.splitstree.org

[7]http://biolabmix.ru/en/products

RESULTS

Phenological Characteristics and Grain Yield of Studied Wheat Cultivars

To assess the relative grain yield performance of the bread wheat cultivars in the dry conditions of Northern and Central Kazakhstan, eight wheat cultivars were selected from our previously published paper (Shavrukov et al., 2016b), and tested in the field during the dry season of 2017. The group of four cultivars (1. Aktyubinka; 2. Albidum 188; 3. Altayskaya 110; and 4. Saratovskaya 60) performed as expected, confirming their high-yielding status, which was significantly higher than the group with low-yield (5. Vera; 6. Volgouralskaya; 7. Yugo-Vostochnaya 2; and 8. Zhenis) (**Table 1**).

The superior high-yielding cultivar Aktyubinka (240 g/m^2) had the shortest DF (39 days) and so earliest start to flowering, while its DM was about average for this group (66 days). In contrast, the lowest-yield cultivar, Yugo-Vostochnaya 2, with more than two-fold lower grain yield than Aktyubinka, started flowering after a 3 day delay (42 days) but was only 1 day shorter in DM (65 days) compared to Aktyubinka. On average, the four high-yielding cultivars started flowering a significant 2.5 days earlier compared to the low-yielding group, while a less pronounced and insignificant difference (1.8 days) was found in DM between the two groups of cultivars (**Table 1**).

Genotyping of Wheat Accessions for the *TaDr1* Gene Using an Amplifluor SNP Marker

During screening of annotated SNPs in bread wheat, the contig BC000036325 was identified for the drought-responsive candidate gene (*TaDr1*) using the publicly available database Cereal DB (see text footenote 1). The SNP marker KATU-W62 was developed to target the annotated SNP [W = A/T] in the 3′-UTR (untranslated region) based on the sequence of BC000036325. Both selected wheat cultivars and the segregating population 18-6 showed genetic polymorphism, with the more common allele being the nucleotide "A" and rarer allele "T" at the SNP position (**Figure 1**).

Genotyping of plants from the eight studied cultivars using the Amplifluor SNP marker KATU-W62 revealed clear discrimination of homozygote genotypes "*aa*" in all four high-yielding cultivars (1–4) while low-yielding cultivars (5–8) were characterized by a mixture of "*bb*" (**5.** Vera; and **7.** Yugo-Vostochnaya 2) and "*ab*" (**6.** Volgouralskaya; and **8.** Zhenis) genotypes (**Figure 1A**). At this stage, it remains unclear whether the "*ab*" genotypes of cultivars Volgouralskaya and Zhenis belong to true heterozygotes, a mixture of several genotypes or both cases together.

Segregation of genotypes for the SNP marker KATU-W62 was observed in the F_3 population 18-6 (**Figure 1B**) originating from the complex cross, where the favorable allele "*a*" was inherited from the paternal line. The analysis of the entire hybrid population is still ongoing and will include progeny analyses in the next generation.

TABLE 1 | Phenological characteristics of eight wheat cultivars grown in the Akmola region, Northern Kazakhstan, in the dry season of 2017.

Group	Cultivar	Days to flowering	Days to maturity	Grain yield (g/m^2)
High-yield	Aktyubinka	39	66	240 ± 14^a
	Albidum 188	42	66	165 ± 11^b
	Altayskaya 110	42	68	155 ± 10^b
	Soratovskaya 60	40	66	162 ± 10^b
Average of the high-yielding group		40.8 ± 0.9*	66.5 ± 0.6	180.5 ± 23.0*
Low-yield	Vera	43	67	129 ± 9^c
	Volgouralskaya	43	74	122 ± 9^c
	Yugo–Vostoch. 2	42	65	112 ± 8^c
	Zhenis	45	67	129 ± 7^c
Average of the low-yielding group		43.3 ± 0.7*	68.3 ± 2.3	123.0 ± 4.3*

Number of Days to flowering (DF) was counted when 50% of plants in the plot started flowering, while number of Days to maturity (DM) was recorded once all plants in each plot reached the ripening stage. Grain yield was calculated in g/m^2, as average of four replicates ± SE. Different letters in superscripts and asterisks (*) indicate significant differences ($p < 0.05$) using ANOVA.

Bioinformatic Characterisation of the *TaDr1* Candidate Gene and Protein

BLASTN results at NCBI[8] for bread wheat gene sequences revealed two accessions, BT009234 for *TaDr1B*, and AF464903 for *TaDr1A*, published and described earlier (Stephenson et al., 2007), with 96% identity in both genes, and covering 96% and 89% of the sequences, respectfully.

Genomic DNA analysis using high confidence genes annotated by the IWGSC database revealed that *TaDr1A* and *TaDr1B* are located on homeologous chromosomes 3A and 3D, in the positions 689,352,814-689,357,320 and 552,949,442-552,953,939, on the forward strands of the physical map, respectively. These genes, TraesCS3A02G450700 and TraesCS3D02G443500, contained five exons, produced 1,536 and 1,565 bp long transcripts which encoded 291 and 298 amino acid long proteins, respectively. The sequence of contig BC000036325, which contained the identified SNP, had the highest level of identity (99.7%) with the gene TraesCS3B02G487800, located in the position 733,818,973-733,823,767, on the forward strand of the physical map of the homeologous chromosome 3B. The gene presented in the BC000036325 contig also contained five exons, transcribed a single 1,317 bp long transcript and encoded a 296 amino acid long protein. Therefore, the two annotated genes *TaDr1A* and *TaDr1B*, and the BC000036325 contig from the SNP database, together represent the three homeologous genes of *TaDr1* in wheat genomes A, D and B, respectively.

The protein encoded by BC000036325 shared 99.3% and 85.% identity with TaDr1B and TaDr1A, respectively, while a low similarity score and only 18.9% identity was found compared to TaNF-YB3, accession BT009265 (**Figure 2**). This result shows that accession BC000036325 from the B genome used in this work has much stronger similarity to TaDr1B and to the corresponding gene *TaDr1B* from the D genome of wheat.

[8]https://www.ncbi.nlm.nih.gov

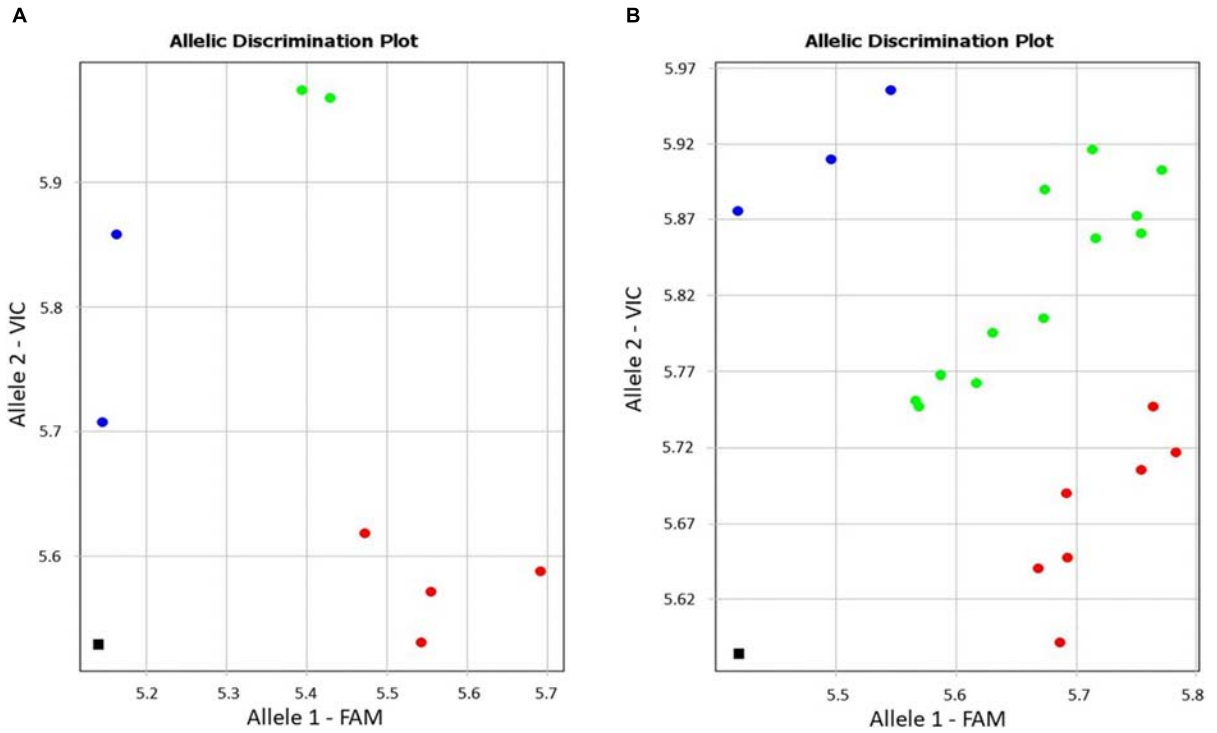


FIGURE 1 | Allele discrimination in eight wheat cultivars **(A)** and in the segregating population 18-6 **(B)** using the Amplifluor-like SNP marker KATU-W62. *X*- and *Y*-axes show relative amplification units, ΔR_n, for FAM and VIC fluorescence signals, respectively. Red dots represent homozygote (*aa*) genotypes with allele 1 (FAM) associated with the high yielding cultivars, blue dots represent homozygote (*bb*) genotypes for allele 2 (VIC), and green dots represent heterozygote (*ab*) or mixed genotypes identified with automatic SNP calling. The black squares show the no template control (NTC) using water instead of template DNA.

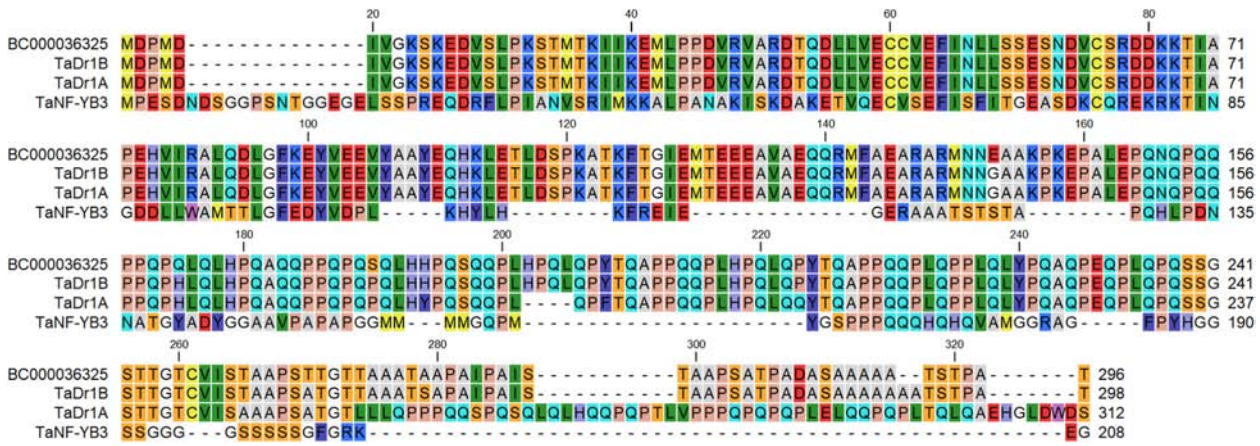


FIGURE 2 | BLASTP protein comparison of the annotated sequence BC000036325 (http://www.cerealsdb.uk.net) with two forms of the general repressor of transcription, TaDr1B (BT009234) and TaDr1A (AF464903), and the TF TaNF-YB3 (BT009265), presented using CLC Main Workbench software.

Molecular Dendrogram of the *TaDr1* Gene

The phylogenetic tree was constructed based on a BLASTX search for molecular similarity for the *TaDr1* protein (BC000036325) in cereal plant species and a group of TFs TaNF-YB for the comparison from NCBI Database. The sequences of all Dr1 proteins are distinct from all TaNF-YB TFs. Among Dr1 sequences, bread wheat (*Triticum aestivum*) and the diploid progenitor of A genome (*T. urartu*) form the first sub-clade; and cultivated rice (*Oryza sativa*) and closely related native grass from tropical Africa (*O. brachyantha*) are isolated in the second sub-clade. All other cereal species are joined together in the third sub-clade including sorghum (*Sorghum bicolor*), maize (*Zea mays*), foxtail millet (*Setaria italica*), and Hall's panicgrass (*Panicum hallii*) (**Figure 3**).

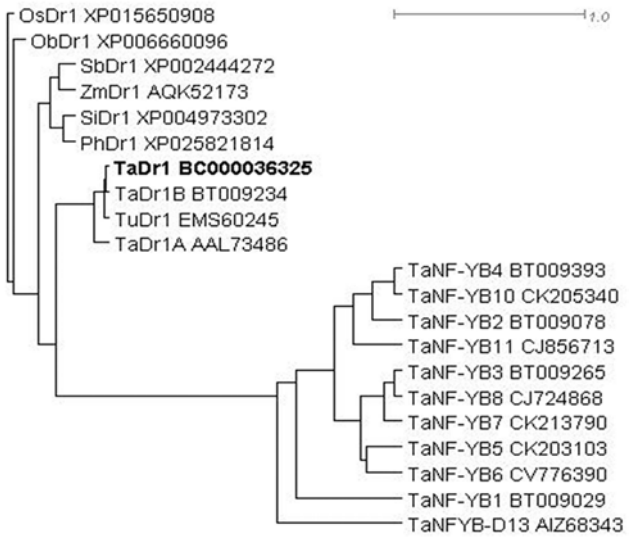


FIGURE 3 | Molecular phylogenetic tree of proteins encoded by *Dr1* genes in monocot plants with the comparison to peptide sequences of TaNF-YB TFs in wheat. Rooted BioNJ dendrogram was generated by program SplitsTree4 (Huson and Bryant, 2006; http://www.splitstree.org). Scale bar shows uncorrected P genetic distance equivalent to 1.0. Accession sequences were retrieved from NCBI database. Plant species are coded as follows: Os, *Oryza sativa*; Ob, *O. brachyantha*; Sb, *Sorghum bicolor*; Zm, *Zea mays*; Si, *Setaria italic*; Ph, *Panicum hallii*; Ta, *Triticum aestivum*; Tu, *T. urartu*. The studying *TaDr1* accession is indicated in Bold.

Expression Analysis of the *TaDr1* in Leaves of Control Plants and Plants Exposed to Drought

Expression profiles for *TaDr1* were recorded as the total of all three homeologous genes, *TaDr1A*, *TaDr1B* and BC000036325 using primers designed for the most conserved regions of these genes. Reference genes used in this study were stable across all genotypes in control and treatment conditions (**Figure 4A**). In plants exposed to drought, our results revealed significant up-regulation of *TaDr1* in all eight studied wheat cultivars (**Figure 4B**). Four high-yielding cultivars increased production of *TaDr1* transcripts 2–2.4 fold, while expression levels in plants of low-yielding cultivars were also increased compared to controls but not as strongly as in plants of high-yielding cultivars (**Figure 4B**).

Both flowering time regulators, *TaVrn1* and *TaFT1*, showed drought responsive expression similar to the expression of *TaDr1*. High-yielding cultivars (1–4) had higher expression levels of *TaVrn1* and *TaFT1* than low-yielding cultivars (5–8), although differences for some cultivars were not significant. These results show genotype-dependent co-expression following the same trend in all three studied genes, *TaDr1*, *TaVrn1*, and *TaFT1*, in leaves of plants grown under drought (**Figures 4B–D**).

Statistical analysis using Tests of Between-Subjects Effects for the gene expressions presented in **Figures 4B–D** shows a very low correlation between groups of high-yielding cultivars (1–4) and low-yielding cultivars (5–8), with R^2 = 0.081, 0.123 and 0.118, respectively. In contrast, strong correlations (R^2 = 0.897 and R^2 = 0.957) were found between cultivars within each group, 1–4 and 5–8, for the three studied genes *TaDr1*, *TaVrn1*, and *TaFT1*, respectively (**Table 2**).

DISCUSSION

Flowering time is a very important trait in wheat, and it was documented that earlier flowering by just a few days can increase the likelihood that plants can minimize the impact of terminal drought and ultimately improve their yield performance compared to wheat genotypes with later flowering times (Reviewed in: Shavrukov et al., 2017). Terminal or late season drought is the most common form of drought stress under most wheat production environments. In the current work, we compared the flowering time of four high-yielding and four low-yielding wheat cultivars and the expression of some genes related to flowering time. In a population of Recombinant breeding lines of durum wheat (*Triticum durum* Desf.) in diverse environments with drought, one QTL for heading date was identified in Chromosome 2A. However, this QTL had minimal or no effect on grain yield (Maccaferri et al., 2008). Different results were reported concerning early heading in synthetic bread wheat lines that correlated with higher grain yield under dry conditions compared to controls (Inagaki et al., 2007). The authors concluded that genes from the D genome could make an important contribution to the correlation in bread wheat, which is absent in tetraploid durum wheat.

The *TaDr1* gene was selected from a SNP database for genetic polymorphism analysis using molecular markers. This gene encodes a protein belonging to the group of general transcription repressors and is an important part of the plant regulatory system.

Two of the three homologous genes, *TaDr1A* and *TaDr1B*, were identified earlier in wheat (Stephenson et al., 2007), and a third *TaDr1* gene with the temporary name of contig BC000036325 identified in the current study, were localized in A, D and B genomes of bread wheat. Alignment of TaDr1 proteins with TaNF-YB3 reveals a high level of identity in the histone fold domain responsible for protein-protein and protein-DNA interactions (**Figure 2**). This result is in agreement with the previously published statement about the "high degree of similarity between TaDr1A, TaDr1B and TaNF-YB subunit members" (Stephenson et al., 2007).

The expression analysis of all three homeologous genes of *TaDr1* comprised an important part of the study of gene function, as published by Stephenson et al. (2007). However, analysis of the primer design for qPCR analysis of the genes, *TaDr1A* and *TaDr1B*, in Stephenson et al. (2007) did not reveal sufficient discrimination between these genes (**Supplementary Material 2**). One pair of primers published by Stephenson et al. (2007) was based on BT009234 and targeted the *TaDr1B* sequence for qPCR analysis, but it shows full consensus between the two genes, with no mismatches (indicated in green, **Supplementary Material 2**). Therefore, the use of these primers gave total (combined) expression for both genes, *TaDr1A* and *TaDr1B*. The second pair of primers, used and reported by Stephenson et al. (2007), was based on AF464903, where the reverse primer was again designed in the conserved region which is identical in both genes. Only a single nucleotide insertion and one SNP were

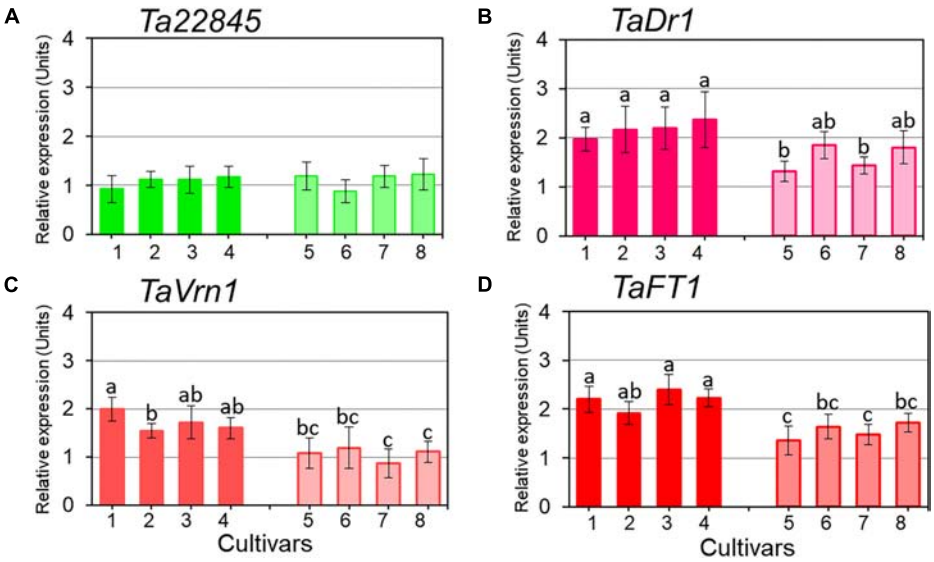


FIGURE 4 | Expression of the reference gene *Ta22845* (ATP-dependent 26S proteasome, regulatory subunit) and target genes, *TaDr1*, *TaVrn1*, and *TaFT1*, in leaves of eight wheat cultivars in response to drought. The expression levels of *Ta22845* **(A)**, *TaDr1* **(B)**, *TaVrn1* **(C)**, and *TaFT1* **(D)** were calculated under drought relative to the corresponding controls in well-watered conditions. Eight wheat cultivars were studied, high-yielding are shown as darker boxes (**1.** Aktyubinka; **2.** Albidum 188; **3.** Altayskaya 110; and **4.** Saratovskaya 60), and the four low-yielding cultivars are shown as framed light filled boxes (**5.** Vera; **6.** Volgouralskaya; **7.** Yugo-Vostochnaya 2; and **8.** Zhenis). With the exception of Panel **A**, expression data were normalized using the averages of two reference genes, *Ta22845* and *Ta54825* (Actin), and presented as the average ± SE of three biological and two technical replicates for each genotype, experiment and treatment. Different letters above the bars indicate significant differences ($p < 0.05$) within each experiment calculated using ANOVA.

found in the sequence of the TaDr1A-Fd primer (indicated in pink, **Supplementary Material 2**). We estimate that it contributes about 90–95% of the studied *TaDr1A* isoform specificity, so in the results presented by Stephenson et al. (2007), *TaDr1B* was over-estimated and represented the total expression of both genes combined, *TaDr1A* and *TaDr1B* (*TaDr1*).

In this context, we similarly measured total expression of all three homeologous genes *TaDr1* with qPCR primers based on the sequence BC000036325. Two mismatches at the 5′-end of the reverse primer (indicated in blue, **Supplementary Material 2**) can affect the specificity of the amplified mRNA of both genes, *TaDr1A* and *TaDr1B*, but only at an equal rate due to perfect consensus between AF464903 and BT009234 sequences in the primer-binding region.

In this work, the associations of an individual GoI with complex traits, such as flowering time and performance under drought, were studied in bread wheat cultivars. The regulatory gene, *TaDr1*, is clearly involved in the plant's response to drought and its expression pattern correlates with the expression patterns of two other regulatory genes, *TaVrn1* and *TaFT1*, which are well-known regulators of flowering time. The existence of small differences in flowering time between high- and low-yielding wheat cultivars under moderate drought was also demonstrated.

In addition, over-expression of regulatory transgenes, *TaNF-YB4*, *TaDREB3*, or *TaSHN1*, as was shown in our earlier papers, activated sets of downstream genes and this led to significantly improved drought tolerance and/or increased grain yield of transgenic wheat plants (Yadav et al., 2015; Shavrukov et al., 2016a; Bi et al., 2018). These results confirm the relevance of the "single-gene for single-trait" approach in studying complex regulatory gene networks, such as, for instance, the response of bread wheat under limited water conditions.

The eight local wheat cultivars from Kazakhstan used in our study were separated into two groups representing high- and low-yielding varieties in the dry conditions of Northern and Central Kazakhstan, as discussed in our previous paper (Shavrukov et al., 2016b) and confirmed in the current study (**Table 1**). Under drought, the two groups of wheat cultivars showed quite variable expression profiles of *TaDr1,* with 2–2.4-fold and 1.3–1.8-fold higher expression of *TaDr1* in the first and second groups of cultivars, respectively (**Figure 4B**). The expression of *TaDr1,* identified as *TaDr1B* in cv. Babax (Stephenson et al., 2007), was reported to be about 2.3-fold above the level of controls, which is close to the highest level of the first group of wheat cultivars in the current study.

TABLE 2 | Correlation analysis between groups of high-yielding and low-yielding cultivars for expression of the three genes, *TaDr1*, *TaVrn1*, and *TaFT1* (right column), and between cultivars within each group (bottom row).

	High-yielding cultivars	Low-yielding cultivars	R^2
TaDr1	2.17 ± 0.08	1.60 ± 0.15	0.081
TaVrn1	1.72 ± 0.10	1.06 ± 0.08	0.123
TaFT1	2.18 ± 0.10	1.55 ± 0.10	0.118
R^2	0.897	0.957	

Data represent the average of the relative expression units for four cultivars, with three biological replicates in each (n = 12) ± SE, extracted from ***Figure 4****. The R^2 correlation coefficient was calculated using Tests of Between-Subjects.*

Our results indicate that the expression of *TaDr1* is dependent on wheat genotype. Four high-yielding cultivars showed very high expression of *TaDr1,* while gene expression was moderate in all four low-yielding cultivars compared to controls under drought treatment.

The two TFs, *TaVrn1* and *TaFT1*, are well studied and are known to strongly regulate the flowering time trait in wheat. Abiotic stresses, such as drought, can affect plant growth and development including flowering. In our recent paper, we reported that the *TaNFYC-A7* gene was differentially expressed under drought in the same cultivars studied here (Zotova et al., 2018). It is suggested that the TaDr1 protein could bind one or both of the TaNF-YB and TaNF-YC type subunits and consequently prevent their interactions or binding to the third subunit, TaNF-YA. It can therefore act as a repressor of the trimeric NF-Y transcription factor. We can extend this hypothesis and speculate that TaNF-Y, which is affected (deactivated) by TaDr1, can release the activity of *TaVrn1* and *TaFT1* promoters. This in turn leads to earlier flowering and ultimately improved performance of wheat genotypes grown in the dry environment of Northern and Central Kazakhstan. The proposed signaling pathway from *TaDr1* to *TaVrn1* and *TaFT1* is supported by the three genes' co-expression results in the current study in wheat plants under drought. High expression of *TaDr1* was accompanied by significant up-regulation of *TaVrn1* and *TaFT1* transcripts. In experiments with drought stress, co-expression patterns in *TaDr1*, *TaVrn1*, and *TaFT1* were genotype-dependent and highly correlated, being much stronger in the four high-yielding wheat cultivars and less pronounced, but still significant, in the four low-yielding cultivars. Further strong evidence will be required to support or reject this hypothesis, including direct "protein-protein" interactions in the studied wheat genotypes.

The application of the Amplifluor-like SNP marker, KATU-W62, like other molecular markers, is a helpful tool for wheat genotyping of both modern cultivars and interspecific hybrids with wild relatives or species related to the genus *Triticum*. In this study, we were able to show that the markers can be deployed in tracking the different alleles in an F_3 population resulting from a complex cross. This population will be used to assess the value of the marker in screening for enhanced drought tolerance under production conditions in Northern Kazakhstan. If our hypothesis is correct, we expect lines carrying the "*a*" allele to perform better under drought, with the strongest improvement shown for homozygotes "*aa*" in the presented study.

Identification of the *TaDr1* alleles can result in a better understanding of genetic polymorphism in the control of down-stream genes, like *TaVrn1* and *TaFT1*, which regulate vernalisation and flowering time. Together with the *Q* gene, the combined regulatory system can change the reproductive architecture of wheat plants and improve their tolerance to abiotic stresses, primarily drought.

AUTHOR CONTRIBUTIONS

LZ conducted the experiments with eight wheat cultivars and the genotyping with Amplifluor-like SNP analysis. AkK and SJ supervised experiments and interpreted results. NG supervised works with vernalisation and flowering time genes, and analysis of interspecific hybrid. NS, AzK, and AN conducted experiments with plant stresses and sampling. AO carried out work and analysis of interspecific hybrid. SS worked with plants in the field trial. VS coordinated experiments in the field. SL analyzed gene sequences in databases and wrote the corresponding section. CS analyzed results, and revised and edited the manuscript. CJ analyzed qRT-PCR data and revised the corresponding section. KS coordinated the qRT-PCR study and revised other sections. PL supervised the study and revised the final version of the manuscript. YS coordinated all experiments and wrote the first version of the manuscript.

FUNDING

This research has been supported by the Ministry of Education and Science, Kazakhstan, Research Program BR05236500 (SJ). The study for genes affecting plant architectonics was funded by the Russian Science Foundation (Russia), grant no. 16-16-10021 (NG). Preliminary evaluation of growth habit phenotypes (spring vs. winter) of varieties from the collection of the accessions was carried out within the framework of the Budget project no. 0324-2018-0018 (NG).

ACKNOWLEDGMENTS

We want to thank the staff and students of S.Seifullin Kazakh AgroTechnical University, Astana (Kazakhstan), Flinders University of South Australia, SA (Australia), and Institute of Cytology and Genetics, Russian Academy of Sciences, Novosibirsk (Russia) for their support in this research and help with critical comments to the manuscript. The results of this study were presented at the International Conference 'Bioinformatics and Computational Biology', August 2018, Novosibirsk, Russia. The authors acknowledge the Organizing Committee for their support in the presentation and publication of this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00063/full#supplementary-material

REFERENCES

Absattar, T., Absattarova, A., Fillipova, N., Otemissova, A., and Shavrukov, Y. (2018). Diversity array technology (DArT) 56K analysis, confirmed by SNP markers, distinguishes one crested wheatgrass *Agropyron* species from two others found in Kazakhstan. *Mol. Breed.* 38:37. doi: 10.1007/s11032-018-0792-3

Arzani, A., and Ashraf, M. (2017). Cultivated ancient wheats (*Triticum* spp.): a potential source of health-beneficial food products. *Compr. Rev. Food Sci. Food Saf.* 16, 477–488. doi: 10.1111/1541-4337.12262

Berger, J., Palta, J., and Vadez, V. (2016). Review: an integrated framework for crop adaptation to dry environments: responses to transient and terminal drought. *Plant Sci.* 253, 58–67. doi: 10.1016/j.plantsci.2016.09.007

Bi, H., Shi, J., Kovalchuk, N., Luang, S., Bazanova, N., Chirkova, L., et al. (2018). Overexpression of the *TaSHN1* transcription factor in bread wheat leads to leaf surface modifications, improved drought tolerance and no yield penalty under controlled growth conditions. *Plant Cell Environ.* 41, 2549–2566. doi: 10.1111/pce.13339

Blümel, M., Dall, N., and Jung, C. (2015). Flowering time regulation in crops - what did we learn from Arabidopsis? *Curr. Opin. Biotechnol.* 32, 121–129. doi: 10.1016/j.copbio.2014.11.023

Campoli, C., and von Korff, M. (2014). Genetic control of reproductive development in temperate cereals. *Adv. Bot. Res.* 72, 131–158. doi: 10.1016/B978-0-12-417162-6.00005-5

Cao, S., Kumimoto, R. W., Gnesutta, N., Calogero, A. M., Mantovani, R., and Holt, B. F. III. (2014). A distal CCAAT/NUCLEAR FACTOR Y complex promotes chromatin loop in gat the FLOWERING LOCUS T promoter and regulates the timing of flowering in *Arabidopsis*. *Plant Cell* 26, 1009–1017. doi: 10.1105/tpc.113.120352

Chen, N. Z., Zhang, X. Q., Wei, P. C., Chen, Q. J., Ren, F., Chen, J., et al. (2007). AtHAP3b plays a crucial role in the regulation of flowering time in *Arabidopsis* during osmotic stress. *J. Biochem. Mol. Biol.* 40, 1083–1089. doi: 10.5483/BMBRep.2007.40.6.1083

Craufurd, P. Q., and Wheeler, T. R. (2009). Climate change and the flowering time of annual crops. *J. Exp. Bot.* 60, 2529–2539. doi: 10.1093/jxb/erp196

Distelfeld, A., Li, C., and Dubcovsky, J. (2009). Regulation of flowering in temperate cereals. *Curr. Opin. Plant Biol.* 12, 178–184. doi: 10.1016/j.pbi.2008.12.010

Gol, L., Tomé, F., and von Korff, M. (2017). Floral transitions in wheat and barley: interactions between photoperiod, abiotic stresses, and nutrient status. *J. Exp. Bot.* 68, 1399–1410. doi: 10.1093/jxb/erx055

Greenup, A., Peacock, W. J., Dennis, E. S., and Trevaskis, B. (2009). The molecular biology of seasonal flowering-responses in *Arabidopsis* and the cereals. *Ann. Bot.* 103, 1165–1172. doi: 10.1093/aob/mcp063

Gudys, K., Guzy-Wrobelska, J., Janiak, A., Dziurka, M. A., Ostrowska, A., Hura, K., et al. (2018). Prioritization of candidate genes in QTL regions for physiological and biochemical traits underlying drought response in barley (*Hordeum vulgare* L.). *Front. Plant Sci.* 9:769. doi: 10.3389/fpls.2018.00769

Gürsoy, M., Balkan, A., and Ulukan, H. (2012). Ecophysiological responses to stresses in plants: a general approach. *Pak. J. Biol. Sci.* 15, 506–516. doi: 10.3923/pjbs.2012.506.516

Gusmaroli, G., Tonelli, C., and Mantovani, R. (2001). Regulation of the CCAAT-binding NF-Y subunits in *Arabidopsis thaliana*. *Gene* 264, 173–185. doi: 10.1016/S0378-1119(01)00323-7

Hou, X., Zhou, J., Liu, C., Liu, L., Shen, L., and Yu, H. (2014). Nuclear factor Y-mediated H3K27me3 demethylation of the *SOC1* locus orchestrates flowering responses of *Arabidopsis*. *Nat. Commun.* 5:4601. doi: 10.1038/ncomms5601

Huang, Z., Lu, Q., and Chen, Y. (2017). Comparative study on reproductive success of *Corydalis sheareri* (Papaveraceae) between alkaline limestone soil and red soil habitats in a karst area. *Biodivers. Sci.* 25, 972–980. doi: 10.17520/biods.2017163

Huson, D. H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23, 254–267. doi: 10.1093/molbev/msj030

Inagaki, M. N., Valkoun, J., and Nachit, M. M. (2007). Effect of soil water deficit on grain yield in synthetic bread wheat derivatives. *Cereal Res. Commun.* 35, 1603–1608. doi: 10.1556/crc.35.2007.4.7

Ingram, J., and Bartels, D. (1996). The molecular basis of dehydration tolerance in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 47, 377–403. doi: 10.1146/annurev.arplant.47.1.377

Inostroza, J. A., Mermelstein, F. H., Ha, I., Lane, W. S., and Reinberg, D. (1992). Dr1, a TATA-binding protein-associated phosphoprotein and inhibitor of class II gene transcription. *Cell* 70, 477–489. doi: 10.1016/0092-8674(92)90172-9

Jarillo, J. A., and Piòeiro, M. (2011). Timing is everything in plant development. The central role of floral repressors. *Plant Sci.* 181, 364–378. doi: 10.1016/j.plantsci.2011.06.011

Jatayev, S., Kurishbaev, A., Zotova, L., Khasanova, G., Serikbay, D., Zhubatkanov, A., et al. (2017). Advantages of amplifluor-like SNP markers over KASP in plant genotyping. *BMC Plant Biol.* 17:254. doi: 10.1186/s12870-017-1197-x

Jung, C., and Müller, A. E. (2009). Flowering time control and applications in plant breeding. *Trends Plant Sci.* 14, 563–573. doi: 10.1016/j.tplants.2009.07.005

Kamran, A., Iqbal, M., and Spaner, D. (2014). Flowering time in wheat (*Triticum aestivum* L.): a key factor for global adaptability. *Euphytica* 197, 1–26. doi: 10.1007/s10681-014-1075-7

Kato, K., Miura, H., Akiyama, M., Kuroshima, M., and Sawada, S. (1998). RFLP mapping of the three major genes, *Vrn1, Q* and *B1*, on the long arm of chromosome 5A of wheat. *Euphytica* 101, 91–95. doi: 10.1023/A:1018372231063

Kaur, G., and Asthir, B. (2017). Molecular responses to drought stress in plants. *Biol. Plant* 61, 201–209. doi: 10.1007/s10535-016-0700-9

Kazan, K., and Lyons, R. (2016). The link between flowering time and stress tolerance. *J. Exp. Bot.* 67, 47–60. doi: 10.1093/jxb/erv441

Khassanova, G., Kurishbaev, A., Jatayev, S., Zhubatkanov, A., Zhumalin, A., Turbekova, A., et al. (2019). Intracellular vesicle trafficking genes, *RabC*-GTP, are highly expressed under salinity and rapid dehydration but down-regulated by drought in leaves of chickpea (*Cicer arietinum* L.). *Front. Genet.* 10:40. doi: 10.3389/fgene.2019.00040

Kim, S., Na, J. G., Hampsey, M., and Reinberg, D. (1997). The Dr1/DRAP1 heterodimer is a global repressor of transcription *in vivo*. *Proc. Natl. Acad. Sci. U. S. A.* 94, 820–825. doi: 10.1073/pnas.94.3.820

Konopatskaia, I., Vavilova, V., Blinov, A., and Goncharov, N. P. (2016). Spike morphology genes in wheat species (*Triticum* L.). *Proc. Latvian Acad. Sci. Sec. B.* 70, 345–355. doi: 10.1515/prolas-2016-0053

Kumimoto, R. W., Zhang, Y., Siefers, N., and Holt, B. F. I. I. I. (2010). NF-YC3, NF-YC4 and NF-YC9 are required for CONSTANS-mediated, photoperiod-dependent flowering in *Arabidopsis thaliana*. *Plant J.* 63, 379–391. doi: 10.1111/j.1365-313X.2010.04247.x

Kuromori, T., Mizoi, J., Umezawa, T., Yamaguchi-Shinozaki, K., and Shinozaki, K. (2014). "Drought stress signaling network," in *Molecular Biology. The Plant Sciences 2*, ed. S. H. Howell (New York: Springer).

Kuromori, T., and Yamamoto, M. (1994). Cloning of cDNAs from *Arabidopsis thaliana* that encode putative protein phosphatase 2C and a human Dr1-like protein by transformation of a fission yeast mutant. *Nucleic Acids Res.* 22, 5296–5301. doi: 10.1093/nar/22.24.5296

Lee, D. K., Kim, H. I., Jang, G., Chung, P. J., Jeong, J. S., Kim, Y. S., et al. (2015). The NF-YA transcription factor *OsNF-YA7* confers drought stress tolerance of rice in an abscisic acid independent manner. *Plant Sci.* 241, 199–210. doi: 10.1016/j.plantsci.2015.10.006

Li, C., and Dubcovsky, J. (2008). Wheat FT protein regulates *VRN1* transcription through interactions with FDL2. *Plant J.* 55, 543–554. doi: 10.1111/j.1365-313X.2008.03526.x

Li, J. X., Hou, X. J., Zhu, J., Zhou, J. J., Huang, H. B., Yue, J. Q., et al. (2017). Identification of genes associated with lemon floral transition and flower development during floral inductive water deficits: a hypothetical model. *Front. Plant Sci.* 8:1013. doi: 10.3389/fpls.2017.01013

Li, Q., Byrns, B., Badawi, M. A., Diallo, A. B., Danyluk, J., Sarhan, F., et al. (2018). Transcriptomic insights into phenological development and cold tolerance of wheat grown in the field. *Plant Physiol.* 176, 2376–2394. doi: 10.1104/pp.17.01311

Liu, X., Hu, P., Huang, M., Tang, Y., Li, Y., Li, L., et al. (2016). The NF-YC-RGL2 module integrates GA and ABA signalling to regulate seed germination in *Arabidopsis*. *Nat. Commun.* 7:12768. doi: 10.1038/ncomms12768

Maccaferri, M., Sanguineti, M. C., Corneti, S., Ortega, J. L. A., Salem, M. B., Bort, J., et al. (2008). Quantitative trait loci for grain yield and adaptation of durum wheat (*Triticum durum* Desf.) across a wide range of water availability. *Genetics* 178, 489–511. doi: 10.1534/genetics.107.077297

Mermelstein, F., Yeung, K., Cao, J., Inostroza, J. A., Erdjument-Bromage, H., Eagelson, K., et al. (1996). Requirement of a corepressor for Dr1-mediated repression of transcription. *Genes Dev.* 10, 1033–1048. doi: 10.1101/gad.10.8.1033

Milec, Z., Valárik, M., Bartoš, J., and Šafáø, J. (2014). Can a late bloomer become an early bird? Tools for flowering time adjustment. *Biotechnol. Adv.* 32, 200–214. doi: 10.1016/j.biotechadv.2013.09.008

Mwadzingeni, L., Figlan, S., Shimelis, H., Mondal, S., and Tsilo, T. J. (2017). Genetic resources and breeding methodologies for improving drought tolerance in wheat. *J. Crop Improv.* 31, 648–672. doi: 10.1080/15427528.2017.1345816

Nelson, D. E., Repetti, P. P., Adams, T. R., Creelman, R. A., Wu, J., Warner, D. C., et al. (2007). Plant nuclear factor Y (NF-Y) B subunits confer drought tolerance

and lead to improved corn yields on water-limited acres. *Proc. Natl. Acad. Sci. U. S. A.* 104, 16450–16455. doi: 10.1073/pnas.0707193104

Paolacci, A. R., Oronzo, A. T., Porceddu, E., and Ciaffi, M. (2009). Identification and validation of reference genes for quantitative RT-PCR normalization in wheat. *BMC Mol. Biol.* 10:11. doi: 10.1186/1471-2199-10-11

Petroni, K., Kumimoto, R. W., Gnesutta, N., Calvenzani, V., Fornari, M., Tonelli, C., et al. (2012). The promiscuous life of plant NUCLEAR FACTOR Y transcription factors. *Plant Cell* 24, 4777–4792. doi: 10.1105/tpc.112.105734

Pinto, R. S., Reynolds, M. P., Mathews, K. L., McIntyre, C. L., Olivares-Villegas, J. J., and Chapman, S. C. (2010). Heat and drought adaptive QTL in a wheat population designed to minimize confounding agronomic effects. *Theor. Appl. Genet.* 121, 1001–1021. doi: 10.1007/s00122-010-1351-4

Prelich, G. (1997). *Saccharomyces cerevisiae BUR6* encodes a DRAP1/NC2α homolog that has both positive and negative roles in transcription in vivo. *Mol. Cell. Biol.* 17, 2057–2065. doi: 10.1128/MCB.17.4.2057

Qu, B., He, X., Wang, J., Zhao, Y., Teng, W., Shao, A., et al. (2015). A wheat CCAAT box-binding transcription factor increases the grain yield of wheat with less fertilizer input. *Plant Physiol.* 167, 411–423. doi: 10.1104/pp.114.246959

Schramm, C., Kurishbaev, A., Jatayev, S., Anderson, P., and Shavrukov, Y. (2019). "Development of single nucleotide polymorphism (SNP) markers for cereal breeding and crop research: current methods and future prospects," in *Advances in Crop Breeding Techniques*, eds F. Ordon and W. Friedt (Cambridge: Burleigh Dodds Science Publishing).

Shanker, A. K., Maheswari, M., Yadav, S. K., Desai, S., Bhanu, D., Attal, N. B., et al. (2014). Drought stress responses in crops. *Funct. Integr. Genomics* 14, 11–22. doi: 10.1007/s10142-013-0356-x

Sharma, P., and Kumar, S. (2005). Differential display-mediated identification of three drought-responsive expressed sequence tags in tea [*Camellia sinensis* (L.) O. Kuntze]. *J. Biosci.* 30, 231–235. doi: 10.1007/BF02703703

Shavrukov, Y., Baho, M., Lopato, S., and Langridge, P. (2016a). The *TaDREB3* transgene transferred by conventional crossings to different genetic backgrounds of bread wheat improves drought tolerance. *Plant Biotechnol. J.* 14, 313–322. doi: 10.1111/pbi.12385

Shavrukov, Y., Zhumalin, A., Serikbay, D., Botayeva, M., Otemisova, A., Absattarova, A., et al. (2016b). Expression level of the DREB2-type gene, identified with amplifluor SNP markers, correlates with performance, and tolerance to dehydration in bread wheat cultivars from Northern Kazakhstan. *Front. Plant Sci.* 7:1736. doi: 10.3389/fpls.2016.01736

Shavrukov, Y., Bovill, J., Afzal, I., Hayes, J. E., Roy, S. J., Tester, M., et al. (2013). *HVP10* encoding V-PPase is a prime candidate for the barley *HvNax3* sodium exclusion gene: evidence from fine mapping and expression analysis. *Planta* 237, 1111–1122. doi: 10.1007/s00425-012-1827-3

Shavrukov, Y., Kurishbaev, A., Jatayev, S., Shvidchenko, V., Zotova, L., Koekemoer, F., et al. (2017). Early flowering as a drought escape mechanism in plants: how can it aid wheat production? *Front. Plant Sci.* 8:1950. doi: 10.3389/fpls.2017.01950

Shen, J., Xiao, Q., Qiu, H., Chen, C., and Chen, H. (2016). Integrative effect of drought and low temperature on litchi (*Litchi chinensis* Sonn.) floral initiation revealed by dynamic genome-wide transcriptome analysis. *Sci. Rep.* 6:32005. doi: 10.1038/srep32005

Siefers, N., Dang, K. K., Kumimoto, R. W., Bynum, W. E. T., Tayrose, G., and Holt, B. F. I. I. I. (2009). Tissue-specific expression patterns of *Arabidopsis* NF-Y transcription factors suggest potential for extensive combinatorial complexity. *Plant Physiol.* 149, 625–641. doi: 10.1104/pp.108.130591

Sinha, S., Kim, I. S., Sohn, K. Y., de Crombrugghe, B., and Maity, S. N. (1996). Three classes of mutations in the A subunit of the CCAAT binding factor CBF delineate functional domains involved in the three-step assembly of the CBF-DNA complex. *Mol. Cell. Biol.* 16, 328–337. doi: 10.1128/MCB.16.1.328

Song, W., Solimeo, H., Rupert, R. A., Yadav, N. S., and Zhu, Q. (2002). Functional dissection of a rice Dr1/DrAp1 transcriptional repression complex. *Plant Cell* 14, 181–195. doi: 10.1105/tpc.010320

Song, Y. H., Ito, S., and Imaizumi, T. (2013). Flowering time regulation: photoperiod- and temperature-sensing in leaves. *Trends Plant Sci.* 18, 575–583. doi: 10.1016/j.tplants.2013.05.003

Stephenson, T. J., McIntyre, C. L., Collet, C., and Xue, G. P. (2007). Genome-wide identification and expression analysis of the NF-Y family of transcription factors in *Triticum aestivum*. *Plant Mol. Biol.* 65, 77–92. doi: 10.1007/s11103-007-9200-9

Swain, S., Myers, Z. A., Siriwardana, C. L., and Holt, B. F. III. (2017). The multifaceted roles of NUCLEAR FACTOR-Y in *Arabidopsis thaliana* development and stress responses. *Biochim. Biophys. Acta Gene Regul. Mech.* 1860, 636–644. doi: 10.1016/j.bbagrm.2016.10.012

Takeno, K. (2016). Stress-induced flowering: the third category of flowering response. *J. Exp. Bot.* 67, 4925–4934. doi: 10.1093/jxb/erw272

Tuberosa, R., and Salvi, S. (2006). Genomics-based approaches to improve drought tolerance of crops. *Trends Plant Sci.* 11, 405–411. doi: 10.1016/j.tplants.2006.06.003

Valliyodan, B., and Nguyen, H. T. (2006). Understanding regulatory networks and engineering for enhanced drought tolerance in plants. *Curr. Opin. Plant Biol.* 9, 189–195. doi: 10.1016/j.pbi.2006.01.019

Wang, M., Wang, S., Liang, Z., Shi, W., Gao, C., and Xia, G. (2018). From genetic stock to genome editing: gene exploitation in wheat. *Trends Biotechnol.* 36, 160–172. doi: 10.1016/j.tibtech.2017.10.002

Willy, P. J., Kobayashi, R., and Kadonaga, J. T. (2000). A basal transcription factor that activates or represses transcription. *Science* 290, 982–984. doi: 10.1126/science.290.5493.982

Xu, J. J., Zhang, X. F., and Xue, H. W. (2016). Rice aleurone layer specific OsNF-YB1 regulates grain filling and endosperm development by interacting with an ERF transcription factor. *J. Exp. Bot.* 67, 6399–6411. doi: 10.1093/jxb/erw409

Yadav, D., Shavrukov, Y., Bazanova, N., Chirkova, L., Borisjuk, N., Kovalchuk, N., et al. (2015). Constitutive overexpression of the *TaNF-YB4* gene in transgenic wheat significantly improves grain yield. *J. Exp. Bot.* 66, 6635–6650. doi: 10.1093/jxb/erv370

Yan, L. (2009). "The flowering pathway in wheat," in *Wheat Science and Trade*, ed. B. F. Carver (Oxford: Wiley-Blackwell), 57–72. doi: 10.1002/9780813818832.ch3

Yang, M., Zhao, Y., Shi, S., Du, X., Gu, J., and Xiao, K. (2017). Wheat nuclear factor Y (NF-Y) B subfamily gene *TaNF-YB3;l* confers critical drought tolerance through modulation of the ABA-associated signaling pathway. *Plant Cell Tiss. Organ Cult.* 128, 97–111. doi: 10.1007/s11240-016-1088-0

Yerzhebayeva, R., Abekova, A., Konysbekov, K., Bastaubayeva, S., Kabdrakhmanova, A., Absattarova, A., et al. (2018). Two sugar beet chitinase genes, *BvSP2* and *BvSE2*, analysed with SNP amplifluor-like markers, are highly expressed after Fusarium root rot inoculations and field susceptibility trial. *PeerJ* 6:e5127. doi: 10.7717/peerj.5127

Yeung, K., Kim, S., and Reinberg, A. (1997). Functional dissection of a human Dr1-DrAp1 repressor complex. *Mol. Cell. Biol.* 17, 36–45. doi: 10.1128/MCB.17.1.36

Yordanov, I., Velikova, V., and Tsonev, T. (2000). Plant responses to drought, acclimation, and stress tolerance. *Photosynthetica* 38, 171–186. doi: 10.1023/A:1007201411474

Zhao, H., Wu, D., Kong, F., Lin, K., Zhang, H., and Li, G. (2017). The *Arabidopsis thaliana* nuclear factor Y transcription factors. *Front. Plant Sci.* 7:2045. doi: 10.3389/fpls.2016.02045

Zotova, L., Kurishbayev, A., Jatayev, S., Khassanova, G., Zhubatkanov, A., Serikbay, D., et al. (2018). Genes encoding transcription factors TaDREB5 and TaNFYC-A7 are differentially expressed in leaves of bread wheat in response to drought, dehydration and ABA. *Front. Plant Sci.* 9:1441. doi: 10.3389/fpls.2018.01441

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 February 2019
doi: 10.3389/fgene.2019.00073

Natural Selection Equally Supports the Human Tendencies in Subordination and Domination: A Genome-Wide Study With *in silico* Confirmation and *in vivo* Validation in Mice

Irina Chadaeva[1,2], Petr Ponomarenko[3], Dmitry Rasskazov[2], Ekaterina Sharypova[2], Elena Kashina[2], Maxim Kleshchev[1], Mikhail Ponomarenko[1,2], Vladimir Naumenko[1], Ludmila Savinkova[2], Nikolay Kolchanov[1,2], Ludmila Osadchuk[1] and Alexandr Osadchuk[1]*

[1] Institute of Cytology and Genetics, Siberian Branch of Russian Academy of Sciences, Novosibirsk, Russia, [2] Novosibirsk State University, Novosibirsk, Russia, [3] University of La Verne, La Verne, CA, United States

OPEN ACCESS

Edited by:
Ancha Baranova, George Mason University, United States

Reviewed by:
Dusanka Savic Pavicevic, University of Belgrade, Serbia
Harinder Singh, J. Craig Venter Institute, United States

***Correspondence:**
Mikhail Ponomarenko
pon@bionet.nsc.ru

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *09 August 2018*
Accepted: *28 January 2019*
Published: *20 February 2019*

Citation:
Chadaeva I, Ponomarenko P, Rasskazov D, Sharypova E, Kashina E, Kleshchev M, Ponomarenko M, Naumenko V, Savinkova L, Kolchanov N, Osadchuk L and Osadchuk A (2019) Natural Selection Equally Supports the Human Tendencies in Subordination and Domination: A Genome-Wide Study With in silico Confirmation and in vivo Validation in Mice. Front. Genet. 10:73. doi: 10.3389/fgene.2019.00073

We proposed the following heuristic decision-making rule: "IF {an excess of a protein relating to the nervous system is an experimentally known physiological marker of low pain sensitivity, fast postinjury recovery, or aggressive, risk/novelty-seeking, anesthetic-like, or similar agonistic-intolerant behavior} AND IF {a single nucleotide polymorphism (SNP) causes overexpression of the gene encoding this protein} THEN {this SNP can be a SNP marker of the tendency in dominance} WHILE {underexpression corresponds to subordination} AND *vice versa*." Using this decision-making rule, we analyzed 231 human genes of neuropeptidergic, non-neuropeptidergic, and neurotrophinergic systems that encode neurotrophic and growth factors, interleukins, neurotransmitters, receptors, transporters, and enzymes. These proteins are known as key factors of human social behavior. We analyzed all the 5,052 SNPs within the 70 bp promoter region upstream of the position where the protein-coding transcript starts, which were retrieved from databases Ensembl and dbSNP using our previously created public Web service SNP_TATA_Comparator (http://beehive.bionet.nsc.ru/cgi-bin/mgs/tatascan/start.pl). This definition of the promoter region includes all TATA-binding protein (TBP)-binding sites. A total of 556 and 552 candidate SNP markers contributing to the dominance and the subordination, respectively, were uncovered. On this basis, we determined that 231 human genes under study are subject to natural selection against underexpression (significance $p < 0.0005$), which equally supports the human tendencies in domination and subordination such as the norm of a reaction (plasticity) of the human social hierarchy. These findings explain vertical transmission of domination and subordination traits previously observed in rodent models. Thus, the results of this study equally support both sides of the century-old unsettled scientific debate on

whether both aggressiveness and the social hierarchy among humans are inherited (as suggested by Freud and Lorenz) or are due to non-genetic social education, when the children are influenced by older individuals across generations (as proposed by Berkowitz and Fromm).

Keywords: gene, promoter, TBP, TATA-box, SNP, expression change, social hierarchy, candidate SNP marker

INTRODUCTION

Social dominance-subordination hierarchy is a set of structured relationships between individuals. These relationship ensure coexistence of individuals by reducing mutual aggression and increasing order in the competition for limited environmental resources as well as elevating their reproductive potential (Hinde, 1970; Rowell, 1974). In animals, such intraspecies hierarchy is a result of agonistic aggressive behavior defined by ethologists as an innate form of action to protect oneself, shelter, progeny, and territory (Lorenz, 2002). Artificial selection of animals for either aggressiveness (Kulikov et al., 2016) or domestication (Belyaev, 1979) has demonstrated the contribution of genetic factors to the phenotypic manifestation of aggressiveness (Ehrman and Parsons, 1981; Moore, 2013). Finally, a genome-wide search for genetic factors of both fear and aggressive behaviors has been conducted on model animals, e.g., in canines, which were artificially selected for both domestication and agonistic behavior (Zapata et al., 2016).

In humans, the reference genome (Colonna et al., 2014) and the full set of single-nucleotide polymorphisms (SNPs) available in the public databases Ensembl (Zerbino et al., 2015) and dbSNP (Sherry et al., 2001). In humans, genetic polymorphism exemplifies the results of natural selection rather than artificial one Dobzhansky (1963) concluded: "man is genetically specialized to be unspecialized," meaning that human behavioral tolerance to social and environmental challenges is broad. The recent genome-wide comparison between humans and apes (Gunbin et al., 2018) indicated that the origin of human species coincided with a reliable increase in the plasticity of the transcription regulation of neuronal genes, while in apes the regulatory plasticity of these genes reduced. This observation points at the action of destabilizing (disruptive) natural selection rather than directional or stabilizing natural selection (Belyaev, 1979). Notably, comprehensive multifactorial regression analysis of healthy young athletes (i.e., boxers, kick boxers, and karate fighters) revealed a significant positive correlation between their aggression and anxiety rates, which helps to achieve top combat levels owing to the prevention of injuries under extreme conditions in the arena (Tiric-Campara et al., 2012). Finally, there is the century-old unsettled scientific dispute where one side – e.g., Freud (1921, 1930) and Lorenz (1964, 2002) – explains both human aggressiveness and social hierarchy as a consequence of their genetic predisposition, while the other side – e.g., Fromm (1941, 1973), Berkowitz (1962, 1993), and Skinner (Rogers and Skinner, 1956; Skinner, 1981) – explains this by the continuous non-genetic social education which continues from childhood to the oldest age (Markel, 2016).

Notably, the social dominance-subordination hierarchy in social species (e.g., humans) limits the permissible aggression range, which is under pressure of natural selection as a norm of a reaction (plasticity) to aggressive behavior (Eldakar and Gallup, 2011). Conditions, quality, and the lifespan of an individual depend on his\her rank within the social hierarchy (Michopoulos et al., 2012). In murine micropopulations as combinations of inbred and hybrid individuals, manifestation of the social dominance phenotype reliably depends on some behavioral features taken together with a genotype (Serova et al., 1991). As for human aggressiveness as a target of some antipsychotic drugs [e.g., olanzapine (Ellingrod et al., 2005)], there are a number of biomedical SNP markers that represent statistically significant differences between the reference human genome and the individual genome of patients having either a certain psychiatric disease or resistance/susceptibility to certain treatments of this disease.

Each discovery of the SNP markers associated with the human phenotypic traits had been a unique success in the pregenomic era, whereas now, this task is one of the major aims of the largest scientific project: "1000 genomes" (Colonna et al., 2014). The main results of this project are publicly available within two regularly synchronized and updated databases Ensembl (Zerbino et al., 2015), which is the reference human genome consisting of the most frequent (ancestral) nucleotides at each DNA position, and dbSNP (Sherry et al., 2001) as the human variome containing all the carefully verified SNPs. Now these databases contain a carefully curated extract that summarizes information on more than 10000 individual human genomes and more than 100 million SNPs (Telenti et al., 2016). As for all the 8.58 billion possible human whole-genome SNPs, creation of a relevant database, dbWGFP, was already reported (Wu et al., 2016); this database is designed to compile all the available information about each of these SNPs to use it in the nearest future to handle the requests from the people who want to sequence their own individual genome and, then, get his/her individual benefits from it.

Because biomedical SNP markers may be used for diagnosis and selection of treatments for humans, there is only one acceptable approach to identify them: that is, to estimate the statistical significance of differences in the prevalence of a given SNP in the representative cohorts of individuals with the phenotypic trait of interest (Varzari et al., 2018). It is unlikely that this extremely time-consuming and expensive procedure is applicable to each of the 8.58 billion possible human SNPs (Abbas et al., 2006). Moreover, both Haldane's dilemma (Haldane, 1957) and Kimura's theory of neutral evolution (Kimura, 1968) predict neutrality of the absolute majority of human SNPs.

These neutral SNPs should be discarded by computer-based calculations in order to reduce the total cost of biomedical SNP markers. Currently, there are many public Web services (e.g., Bendl et al., 2016), predicting candidate SNP markers and eliminating the most probable neutral SNPs while taking into account various similarity measures for genome-wide data during infections (Leschner et al., 2012) or diseases (Hu et al., 2013) as well as after treatment (Hein and Graver, 2013) and in health (Ni et al., 2012). The accuracy of these similarity-based predictions increases with the increase in diversity of available genome-wide data, in agreement with our predictions (Ponomarenko et al., 1999) based on Central Limit Theorem.

The best accuracy of these bioinformatics predictions corresponds to SNPs in the protein-coding regions owing to their reliable manifestation as protein damage, whereas in the case of SNPs in the regulatory regions of genes, none of the proteins is damaged (Amberger et al., 2015). Notably, the 70 bp promoter regions in front of the transcription start sites (TSSs) contain the majority of the clinically verified regulatory SNP markers (Ponomarenko et al., 2013) due to the TATA-binding protein (TBP)-binding site (e.g., TATA-box), which is obligatory for the primary initiation of gene transcription (Martianov et al., 2002). Finally, Mogno et al. (2010) experimentally found that the increase in TBP-binding affinity for the TBP-binding sites altered by SNPs causes overexpression of the appropriate genes whereas underexpression corresponds to a decrease in the affinity.

In our previous works, we created a public Web service SNP_TATA_Comparator (see text footnote 1) (Ponomarenko et al., 2015) for selecting the statistically significant SNP-caused alterations in TBP's affinity for the promoter regions 70 bp upstream of the protein-coding TSSs. This Web service is based on our three-step model of the TBP–promoter binding to each other (Ponomarenko et al., 2008), namely: (i) TBP slides along DNA ↔ (ii) TBP stops at a putative TBP-binding site ↔ (iii) the TBP–promoter complex is fixed by the DNA bending at a right angle, as was experimentally discovered (Delgadillo et al., 2009). Using SNP_TATA_Comparator, we predicted candidate SNP markers – within TBP-binding sites of the human gene promoters – associated with obesity, chronopathology, aggressiveness, and autoimmune and Alzheimer's diseases (for review, see Ponomarenko P. et al., 2017). Recently, we preliminarily studied (Chadaeva et al., 2017) the possibility to predict candidate SNP markers for social hierarchy using a short representative set of 21 human genes homologous to the animal genes encoding the known physiological markers of aggressiveness, which represent nervous, endocrine, immune, respiratory, vascular, muscular, and other systems of the human body.

In this work, due to our observation (Bragin et al., 2006) of domination of adult male BALB/cLac mice over CBA/Lac mice, we made a genome-wide prediction for the human tendencies dominance and subordination within the framework of the neuropeptidergic, non-neuropeptidergic, and neurotrophinergic systems and verified it using a mouse model of human inheritance. We discuss how our results fit both genetic (e.g., Freud and Lorenz) and non-genetic (e.g., Berkowitz and Fromm) irreconcilable sides of the century-old scientific debate about the origin of both aggressiveness and social hierarchy in humans.

[1]http://beehive.bionet.nsc.ru/cgi-bin/mgs/tatascan/start.pl

MATERIALS AND METHODS

Animals

This study was carried out in accordance with the recommendations of Directive 2010/63/EU of the European Parliament and of the Council of September 22, 2010, on the protection of animals used for scientific purposes. Manipulations of animals and experimental procedures were performed in compliance with the international rules according to the "Guidelines for the care and use of mammals in neuroscience and behavioral research"[1]. The research protocol was approved by the Interinstitutional Commission on Bioethics at the ICG SB RAS, 10 Lavrentyev Avenue, Novosibirsk, Russia.

Analysis of the inheritance of agonistic behavior indicators and social dominance levels was conducted on 230 adult male mice that are diallelic crosses of a set of five maternal inbred mouse strains (i.e., PT, DD, YT, A/He, and C57BL/6J) with two analytic inbred paternal strains (BALB/cLac and CBA/Lac) of the murine tendencies in dominance and subordination, respectively, as determined experimentally previously (Bragin et al., 2006).

All the mice were maintained under standard conditions of a conventional animal facility of the ICG SB RAS.

Identification of Inheritance of the Mouse Tendencies in Dominance and Subordination

One can see all the 230 diallelic crosses in **Table 1**, where five rows and two columns present F1 males. In each row of this table, there are descendants of mothers of the same inbred strain. Thus, the maternal non-genetic (pre- and postnatal) and cytoplasmic effects are the same for males of the same row of this table. To exclude non-genetic paternal postnatal effects on offspring, pregnant female mice were isolated from male mice.

We made up groups of F1 hybrid male mice with the minimal society size, namely: two males each: one from each column

[1]https://grants.nih.gov/grants/olaw/National_Academies_Guidelines_for_Use_and_Care.pdf

TABLE 1 | The experimental design for identification of inheritance of the murine tendencies in dominance and subordination.

Paternal genotype Maternal genotype	BALB/cLac	CBA/Lac
PT	PT × BALB/cLac (31)	PT × CBA/Lac (31)
C57Bl/6J	C57Bl/6J × BALB/cLac (20)	C57Bl/6J × CBA/Lac (20)
YT	YT × BALB/cLac (21)	YT × CBA/Lac (21)
DD	DD × BALB/cLac (20)	DD × CBA/Lac (20)
A/He	A/He × BALB/cLac (23)	A/He × CBA/Lac (23)

The number of male mice for each of the 10 F1 hybrids is indicated in parentheses.

of the same row of **Table 1**. In each pair, both male mice had identical age, weight, and body size, but visually differed from each other in color. This approach allowed us to estimate the influence of the paternal genotype on the social dominance level of the appropriate F1 crosses.

A total of 115 experimental pairs (230 F1 hybrids) were distributed into five groups, corresponding to the maternal inbred strains (see **Table 1**). For each mouse male pair tested, we performed 14 observations (20 min each) during 5 days. Each observation was recorded using a video camera in automatic mode with a fixed period. Next, we analyzed these video recordings using the protocols of software The Observer XT 7.0 (version: 7.0, Noldus Information Technology, license No. OB070-03670). This way, we identified the social rank for each male within the appropriate pair according to asymmetry in agonistic behavior, in particular, by means of attacks and submissive poses as described in the Supplementary Experiment (**Supplementary File S6**).

The Basic Decision-Making Rule

Both domesticated and laboratory animals are artificially selected using the known target traits (Belyaev, 1979; Kulikov et al., 2016), which can help in any computer-based genome-wide analysis of these animals (e.g., Zapata et al., 2016) in contrast to the human genome, which is the result of natural selection in favor of unknown unspecializing target traits (Dobzhansky, 1963). Hence, on the basis of our preliminary work (Chadaeva et al., 2017), we proposed the following heuristic decision-making rule: "IF {an excess of a protein relating to the nervous system is an experimentally known physiological marker of low pain sensitivity, fast post-injury recovery, or aggressive, fearless, impulsive, anxious, exploratory, risk/novelty-seeking, anesthetic-like, or similar agonistic-intolerant behavior} AND IF {a given SNP can cause overexpression of a gene encoding this protein} THEN {this SNP can be a SNP marker of predisposition to social dominance} WHILE {the underexpression corresponds to subordination} AND *vice versa*." This whole study is devoted to evaluation of this decision-making rule.

DNA Sequences

Using the aforementioned basic decision-making rule (see subsection "The Basic Decision-Making Rule"), we analyzed all the 5052 SNPs retrieved from the dbSNP database (build 150, Sherry et al., 2001), which are found within the 70 bp promoter regions upstream of the protein-coding transcripts of all the 231 human genes of the neuropeptidergic, non-neuropeptidergic, and neurotrophinergic systems retrieved from database Ensembl (GRCh38/hg38 assembly, Zerbino et al., 2015), which are listed in the alphabetic order in the first columns of **Supplementary Tables S1–S3**, respectively (hereinafter: see **Supplementary Files S1–S3**, respectively). These genes encode proteins that are known as key factors altering human social behavior, namely, neurotrophic and growth factors, interleukins, neurotransmitters, receptors, transporters, and enzymes.

Using our public Web service SNP_TATA_Comparator (Ponomarenko et al., 2015), we compared the DNA sequences of the ancestral (wt) and minor (min) alleles of SNPs of the 70 bp promoter region of these genes. We applied it together with the public Web service UCSC Genome Browser (Haeussler et al., 2015) and two public databases dbSNP (Sherry et al., 2001) and ClinVar (Landrum et al., 2014), as described in the Supplementary Web-service (**Supplementary File S5**). As a result, we obtained two pairs of $(-\ln(K_D^{(wt)}) \pm \delta_{(wt)})$ and $(-\ln(K_D^{(min)}) \pm \delta_{(min)})$ values of TBP affinity for these alleles of the promoter being studied according to contextual, conformational, and physicochemical changes in its B-helical DNA under the influence of a given SNP, as described in the Supplementary Method (**Supplementary File S4**). Next, we calculated Fisher's Z-score as follows: $Z = \mathrm{abs}[\ln(K_D^{(min)}/K_D^{(wt)})]/[\delta^2_{(min)}+\delta^2_{(wt)}]^{1/2}$, and in turn found the *p*-value of statistical significance of this score using package R (Waardenberg et al., 2015).

Finally, using this *p*-value, we discarded all the SNPs the effects of which were estimated as insignificant; otherwise, using decisions on the SNP-caused significant increase and decrease of the binding affinity of TBP for the analyzed promoters, we predicted the candidate SNP markers for over- or underexpression of the appropriate genes, respectively, as demonstrated experimentally (Mogno et al., 2010). Readers can find all our predictions within the columns "K_D, nM, prediction" of **Supplementary Tables S1–S3**. Their subcolumns "wt" and "min" contain K_D values of TBP's binding affinity for the ancestral and minor alleles of the appropriate promoters, respectively. Furthermore, subcolumns "Δ" and "α" correspond to the human gene expression alterations and their statistical significance levels α, which are equal to $(1 - p)$. In addition, subcolumn "ρ" presents a heuristic rank of our predictions varying in alphabetical order from the "best" (A) to the "worst" (E). Finally, **Table 2** contains total numbers of our predictions (N_{RES}) as well as the numbers of the candidate SNP markers for either overexpression ($N_>$) or underexpression ($N_<$) of the human genes, as predicted by this work.

The Keyword Search in the PubMed Database

For each candidate SNP marker predicted, we manually performed a two-step keyword search in the PubMed database (Lu, 2011) as shown in **Figure 1**.

As presented in this figure, we handled each candidate SNP marker independently of the others, one by one. First of all, we checked whether the SNP in question was annotated by database ClinVar (Landrum et al., 2014) as depicted in **Supplementary Figure S1C** (hereinafter: see **Supplementary File S5** "Supplementary Web service") and boldfaced in both the first and third rightmost columns of **Supplementary Tables S1–S3**.

When this database associated the SNP under study with the human diseases, we manually carried out a primary keyword search for the literature data on the known physiological marker of these diseases, which corresponds to the gene expression alteration predicted for this SNP as described elsewhere

TABLE 2 | Predictions of candidate SNP markers that can statistically significantly alter the TATA-binding protein (TBP)-binding sites of the human gene promoters of all the protein-coding transcripts relating to neuropeptidergic, non-neuropeptidergic, and neurotrophinergic systems.

Data studied: GRCh38, dbSNP 150			Result	H_0: social status equivalence			H_0: neutral natural selection		
Human body systems	N_{GENE}	N_{SNP}	N_{RES}	$N_{\uparrow}$	$N_{\downarrow}$	$P(N_{\uparrow} \equiv N_{\downarrow} \equiv N_{RES}/2)$	$N_{>}$	$N_{<}$	$P(N_{<} \equiv 4N_{>} \equiv 4N_{RES}/5)$
Genome-wide estimate (1000 Genomes Project Consortium et al., 2012)	10^4	10^5	1000				200	800	> 0.52
Clinical SNP markers of hereditary diseases within the TBP-binding sites (Ponomarenko et al., 2015)	33	203	51				14	37	> 0.93
Candidate SNP markers within the TBP-binding sites of promoters of reproductivity-related genes (Chadaeva et al., 2018)	22	129	24				19	5	< 0.000001
Candidate SNP markers within the TBP-binding sites of promoters of familial Alzheimer's disease-related genes (Ponomarenko P. et al., 2017)	5	143	28				16	12	< 0.000025
Candidate SNP markers within the TBP-binding sites of promoters of circadian clock core genes (Ponomarenko et al., 2016)	16	162	52				39	13	< 0.000001
All: a representative set of genes (Chadaeva et al., 2017)	21	381	92	45	47	>0.9	66	26	< 0.000001
Neuropeptidergic	27	395	97	51	46	>0.6	66	31	< 0.000001
Non-neuropeptidergic	109	2226	505	240	265	>0.2	342	163	< 0.000001
Neurotrophinergic	95	2431	506	265	241	>0.3	346	160	< 0.000001
TOTAL	231	5052	1108	556	552	>0.9	754	354	< 0.000001

N_{GENE} and N_{SNP}, total numbers of the human genes and their SNPs (single nucleotide polymorphisms) within the 70 bp promoter region for the protein-coding transcripts, respectively, in this study; N_{RES}, the total number of the candidate SNP markers predicted in this work that can increase ($N_{>}$) or decrease ($N_{<}$) the TATA-binding protein (TBP) binding affinity for these promoters and, correspondingly, the expression of these genes; $N_{\uparrow}$ and $N_{\downarrow}$, the total numbers of the candidate SNP markers for the human tendencies in dominance and subordination, respectively; $P(H_0)$, the estimate of a probability for the acceptance of this H_0 hypothesis, according to the binomial distribution.

(Lu, 2011). **Figure 1** depicts this procedure as two boxes consisting of dashed lines. In the case of a successful finding of such a publication, the clinical data taken from database ClinVar (Landrum et al., 2014) indicated the adequacy of our predictions for the SNP under consideration. These confirmations of our predictions are *italicized* in both the first and third rightmost column of **Supplementary Tables S1–S3**.

Finally, two dotted boxes in **Figure 1** depict our secondary keyword search for the known physiological markers for pain sensitivity, postinjury repair efficiency, or agonistic behavior, which correspond to underexpression of the human gene containing this SNP. This way, we tested the basic decision-making rule of this work (hereinafter: see subsection "The Basic Decision-Making Rule" "Basic decision-making rule"). As the main bioinformatic results, we predicted the candidate SNP markers for the human tendencies in dominance and subordination, which are in both the first and third rightmost column of **Supplementary Tables S1–S3**. **Table 2** contains the total number of these candidate SNP markers ($N_{\uparrow}$ and $N_{\downarrow}$, respectively).

The section "References" lists the articles cited in **Supplementary Tables S1–S3** and in section "Supplementary Method."

Statistical Analysis

We analyzed dichotomies via the equiprobable binomial distribution and χ^2 criteria taken from the standard statistical package Statistica (StatSoft^{TM}, Tulsa, United States).

In the genome-wide study *in silico*, using only Fisher's *Z*-score test, we predicted the candidate SNP markers, the numbers of which for the human gene overexpression and underexpression were compared with one another using the binomial distribution as well as in the case of the human tendencies in dominance and subordination.

During *in vivo* validation in mice, by means of the χ^2 criterion, we compared the actual numbers of dominants and subordinates among male mice, which were the F1 hybrids of crossing females from inbred strains of an unknown tendency in social hierarchy with males from two inbred strains BALB\cLac and CBA\Lac of the previously experimentally identified tendencies in dominance and subordination, respectively (Bragin et al., 2006).

RESULTS AND DISCUSSION

Our analysis of 5052 SNPs of the TBP-binding regions of 231 human neuron-related genes uncovered 1108 candidate

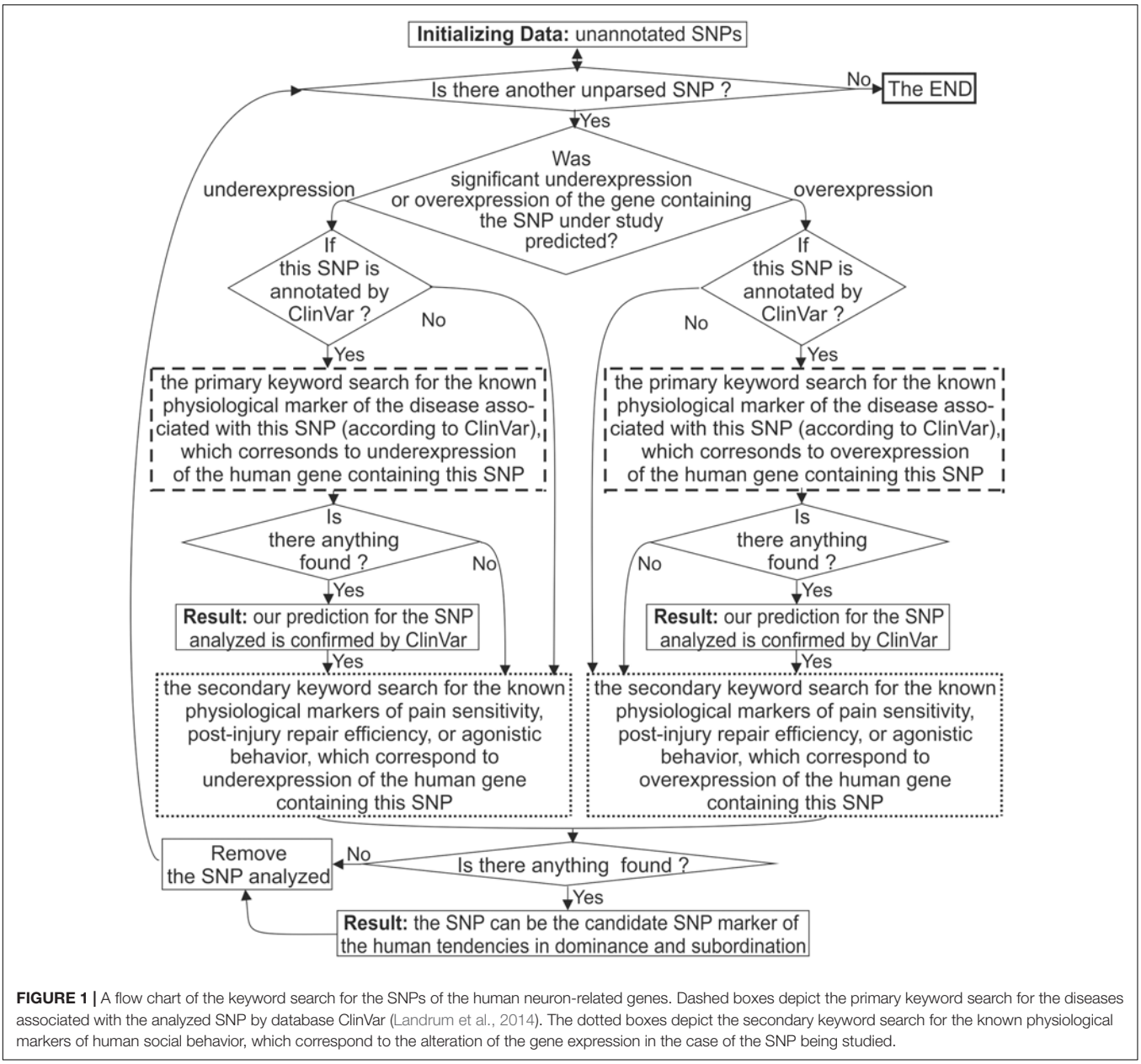


FIGURE 1 | A flow chart of the keyword search for the SNPs of the human neuron-related genes. Dashed boxes depict the primary keyword search for the diseases associated with the analyzed SNP by database ClinVar (Landrum et al., 2014). The dotted boxes depict the secondary keyword search for the known physiological markers of human social behavior, which correspond to the alteration of the gene expression in the case of the SNP being studied.

SNP markers for the human tendencies in dominance and subordination (**Table 2**). These predictions are shown in **Supplementary Tables S1–S3** and exemplified in **Figures 2**, **3** and **Supplementary Figure S1**. For 36 of the 231 genes (16%), namely: *ADRA1B*, *ADRA2A*, *ADRA2B*, *ADRB1*, *AVP*, *AVPR1Â*, *CHRNB2*, *CNR2*, *FGF15*, *FGF16*, *FGF2*, *FGF23*, *FGF7*, *FIGF*, *FLT3*, *GABARAPL3*, *GABRA3*, *GABRA4*, *GABRQ*, *GMFA*, *GRIA3*, *GRIK4*, *GRIN2B*, *GRM6*, *IGF2R*, *IL27RA*, *KDR*, *LIF*, *MANF*, *MAOA*, *MAOB*, *NGF*, *OXT*, *TACR3*, *TGFBRAP1*, and *VEGFC*, no candidate SNP markers were found (data not shown). Let us focus our analysis of our results on the candidate SNP markers that have independent clinical information within database ClinVar (Landrum et al., 2014) to both verify and discuss their relevance to the human genes under study.

Candidate SNP Markers Near TBP-Binding Sites in the Promoter of the Human Genes Encoding Neuropeptidergic-System-Related Proteins (e.g., Neurotransmitters)

We applied our experimentally verified public Web service (Ponomarenko et al., 2015) to analyze 395 SNPs in 70 bp proximal promoter regions of 27 human genes encoding neuropeptidergic-system–related proteins, namely: arginine vasopressin receptors (*AVPR*s), C-X-C motif chemokine receptors (*CXCR*s); neuropeptide Y and its receptors (*NPY*s), opioid growth factor receptor (*OGFR*), opioid receptors (*OPR*s), oxytocin and its receptor (*OXT*s), prodynorphin

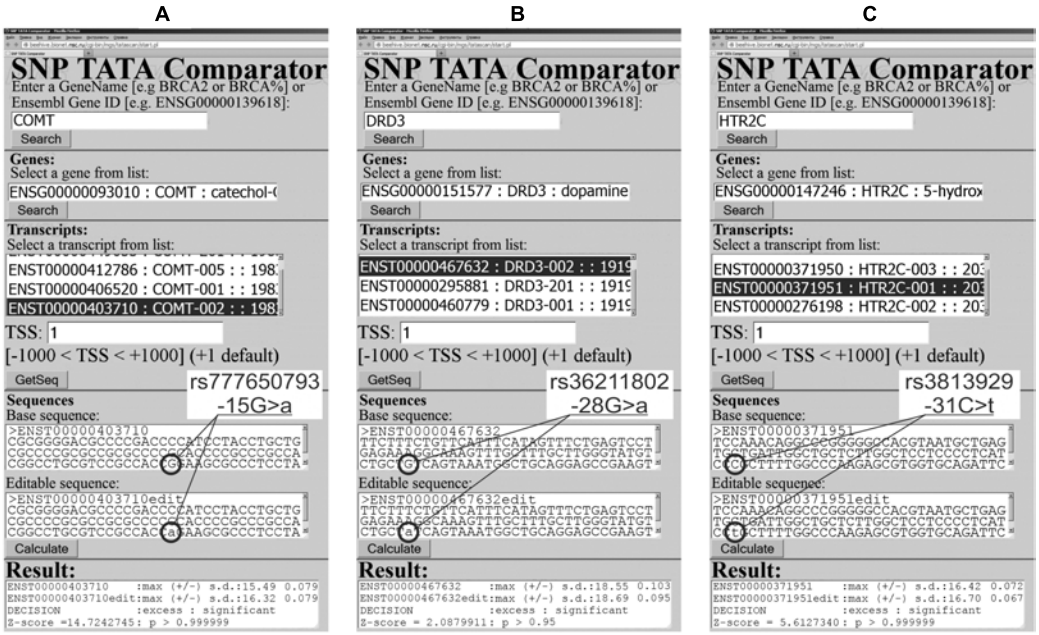


FIGURE 2 | Examples of our predictions in this work in the case of the human genes encoding neuropeptidergic-system-unrelated proteins. **(A)** rs777650793; **(B)** rs36211802; and **(C)** rs3813929.

(*PDYN*), proenkephalin (*PENK*), prepronociceptin (*PNOC*), proopiomelanocortin (*POMC*), and tachykinins together with their precursors and receptors (*TAC*s). The results obtained can be found in **Supplementary Table S1**.

The human *PDYN* gene, i.e., the opioid polypeptide hormone prodynorphin, which is a basic building block of endogenous opioid neuropeptides, so-called endorphins, that can inhibit the pain signals peripherally and cause a feeling of euphoria (when acting in the brain) as neurotransmitters of happiness and joy. SNP rs886056538 of this gene's promoter was annotated within database ClinVar (Landrum et al., 2014), where it is associated with spinocerebellar ataxia as shown in **Supplementary Figure S1C**. **Supplementary Figure S1D** illustrates our prediction for this SNP, which is the line "Decision: excess significant" accompanied by the line "*Z*-score = 2.51, $p > 0.95$" within the textbox "Result." This outcome means that this SNP can statistically significantly cause overexpression of this gene. Our primary keyword search (hereinafter: two dashed boxes in **Figure 1**) produced an original experiment (Smeets et al., 2015) involving a mouse model of the human diseases, which has identified the prodynorphin excess as a physiological marker for spinocerebellar ataxia. As one can see, these *in vivo* experimental data independently support our prediction for SNP rs886056538 (**Supplementary Figure S1**). This observation indicates the suitability of our Web service (Ponomarenko et al., 2015) for computer-based analysis of the human genes encoding neuropeptidergic-system–related proteins as *italicized* in **Supplementary Table S1**.

After this validation, we manually conducted our secondary keyword search (hereinafter: two dotted boxes in **Figure 1**) and found the original experiment (Szklarczyk et al., 2012) in a mouse model of human behavior, which associated the prodynorphin excess with reduced conditioned fear. Using our basic decision-making rule within the limitations of the above experimental model of human behavior (Szklarczyk et al., 2012), we predicted that the analyzed SNP rs886056538 can be a candidate SNP marker for the human tendency in dominance (**Supplementary Table S1**).

Near this clinically characterized SNP marker, we found two unannotated SNPs (rs371345545 and rs557431815), which can also cause overexpression of the human *PDYN* gene (hereinafter: according to our predictions shown in **Supplementary Tables S1–S3**). That is why we suggest them as two candidate SNP markers of the same genetic tendencies, namely: spinocerebellar ataxia with limitations (Smeets et al., 2015) and social dominance within the framework of the model (Szklarczyk et al., 2012) as presented in **Supplementary Table S1**.

This way, we predicted 66 and 31 candidate SNP markers for excess and deficiency of the proteins of the human neuropeptidergic system, respectively, which are also 51 and 46 candidate SNP markers predicted by this work for the human tendencies in dominance and subordination (**Table 2** and **Supplementary Table S1**). First of all, readers can see that the numbers of the candidate SNP markers predicted for the human tendencies in dominance and subordination markers are not statistically significantly different from one another according to equiprobable binomial distribution criterion ($P(N_{\uparrow} \equiv N_{\downarrow} \equiv N_{RES}/2) > 0.6$). This finding is in agreement with our preliminary estimate (Chadaeva et al., 2017), namely: $P(N_{\uparrow} \equiv N_{\downarrow} \equiv N_{RES}/2) > 0.9$.

On the contrary, the numbers of the candidate SNP markers predicted for excess and deficiency of the proteins of the human

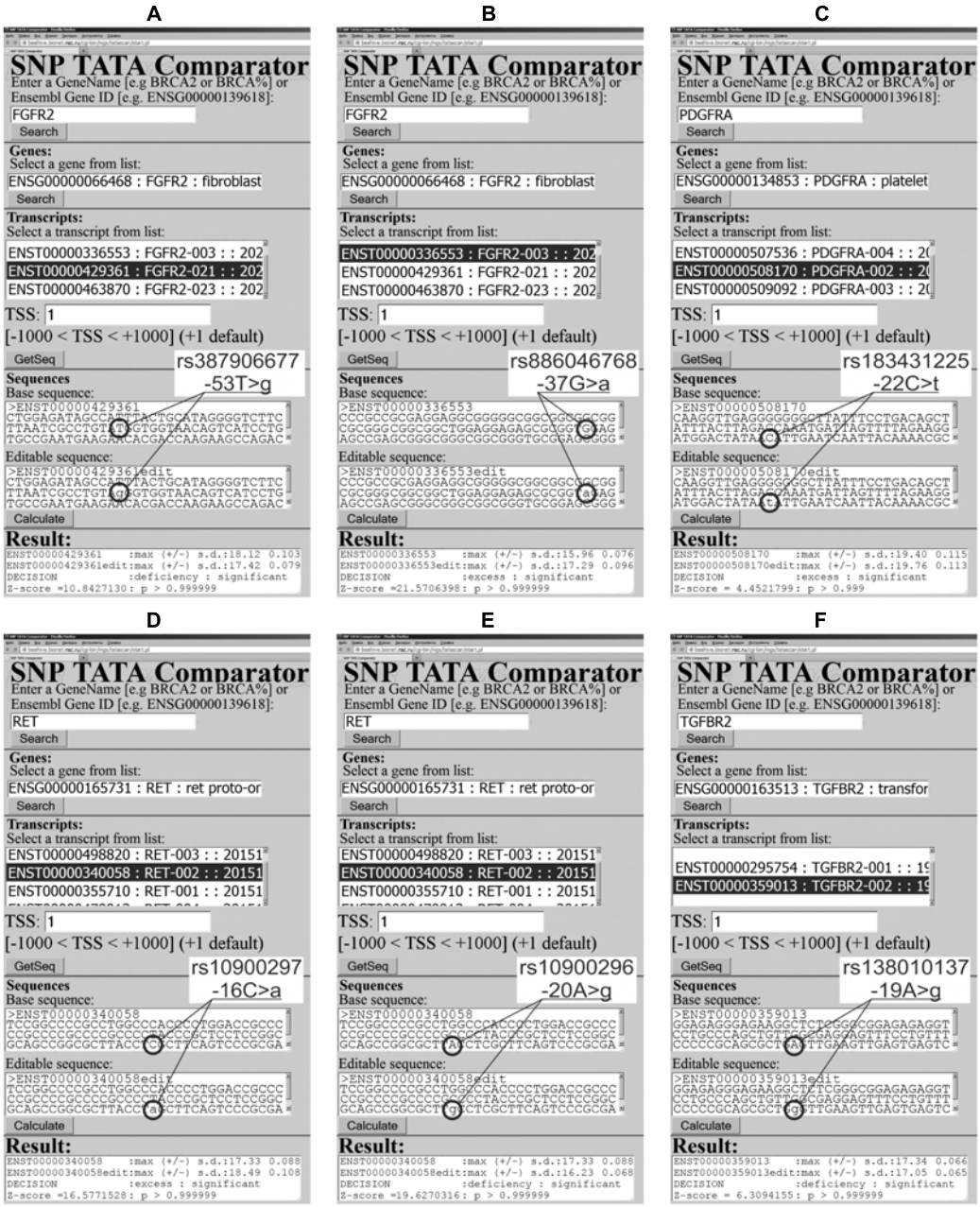


FIGURE 3 | Examples of our predictions in this work in the case of human genes encoding neurotrophinergic-system-related proteins. **(A)** rs387906677; **(B)** rs886046768; **(C)** rs183431225; **(D)** rs10900297; **(E)** rs10900296; and **(F)** rs138010137.

neuropeptidergic system are significantly different from one another according to the equiprobable binomial distribution criterion ($P(N_{>} \equiv N_{<} \equiv N_{RES}/2) < 0.0005$) in line with our preliminary observations (Chadaeva et al., 2017), as presented in **Table 2**: $N_{>} = 66$, $N_{<} = 26$ ($P(N_{>} \equiv N_{<} \equiv N_{RES}/2) < 0.0005$). According to a number of studies, various molecular phenomena can shift frequencies of mutations – e.g., influence of the nucleotide context on the occurrence and repair of pre-mutational damage to genomic DNA, gene conversion, pleiotropic and epistatic effects – Kasowski et al. (2010) first noticed that SNPs decreasing the protein–DNA affinity are much more frequent than SNPs increasing this affinity within the human genome. Next, the authors of ref. (1000 Genomes Project Consortium et al., 2012) quantitatively characterized this mutational shift, namely: there are ∼800 SNPs damaging the transcription factor binding sites and ∼200 SNPs improving these sites per random individual human genome as shown in **Table 2**. According to Haldane's dilemma (Haldane, 1957) and neutral evolution theory (Kimura, 1968), this genome-wide estimate can correspond to the neutral mutational drift as a

norm. Indeed, we observed 37 clinically proven SNP markers of the human hereditary diseases, which decrease the TBP–promoter affinity, and 14 such SNP markers increasing this affinity (Ponomarenko et al., 2015) in agreement with the above-mentioned genome-wide estimate (**Table 2**). This pattern matches the commonly accepted opinion on these diseases as a genetic load of the neutral mutational drift in the norm.

Nevertheless, in the case of human reproductive potential, which is considered the target of natural selection, we observed a diametrically opposite pattern, namely: five candidate SNP markers were decreasing the TBP–promoter affinity and 19 candidate SNP markers were increasing this affinity (Chadaeva et al., 2018). Besides, we found (Ponomarenko P. et al., 2017) only a minority (12 of 28) of candidate SNP markers of familial Alzheimer's disease that can decrease the TBP–promoter affinity; this finding is consistent with natural selection for its very slow pathogenesis, whose clinical manifestation is observed only at the age of over 65 (**Table 2**). In addition, in the case of core genes of the circadian clock (Ponomarenko et al., 2016), which are naturally selected for continuous coordination between the functioning of systems of the human body and daily fluctuations of the environment, we found 13 candidate SNP markers that can decrease the TBP–promoter affinity and 39 candidate SNP markers increasing this affinity (**Table 2**).

Looking through **Table 2**, we noticed that our predictions for the neuropeptidergic gene system are more similar to those for natural selection cases than to those for neutral drift within the normal range. That is why here we predict that the human genes encoding neuropeptidergic-system-related proteins are under natural selection pressure, which equally supports the human tendencies in subordination and domination, as was preliminarily estimated elsewhere (Chadaeva et al., 2017). This way, we followed the semicentennial bioinformatic tradition to compare the actual frequencies of natural mutations within their various dichotomies [e.g., transitions versus transversions (Kimura, 1980) as well as synonymous versus non-synonymous changes (Li et al., 1985)].

Candidate SNP Markers Near TBP-Binding Sites in the Promoter of the Human Genes Encoding Proteins Related to the Non-neuropeptidergic System (e.g., Receptors)

Using our public Web service (Ponomarenko et al., 2015), we analyzed 2226 SNPs located within the TBP-binding regions of 109 human genes encoding proteins that are related to the non-neuropeptidergic system, e.g., adenosine receptors (*ADOR*s), adrenoceptors (*ADR*s), muscarinic cholinergic receptors (*CHRM*s), nicotinic cholinergic receptors (*CHRN*s), central cannabinoid receptor 1 (*CNR1*), catechol-O-methyltransferase (*COMT*), dopamine D receptors (*DRD*s), GABA type A receptor-associated proteins (*GABARAP*s), γ-aminobutyric acid type B receptor subunits (*GABBR*s), γ-aminobutyric acid − type A receptor subunits (*GABR*s), G protein–coupled receptors (*GRP*s), glutamate ionotropic receptor AMPA–type subunits (*GRIA*s), glutamate ionotropic receptor NMDA-type subunits (*GRIN*s), glutamate metabotropic receptors (*GRM*s), 5-hydroxytryptamine (serotonin) receptors (*HTR*s), dopamine transporter DAT (*SLC6A3*), Na^+/Cl^--dependent serotonin transporter SERT (*SLC6A4*), tyrosine hydroxylase (*TH*), and tryptophan hydroxylase 2 (*TPH2*). **Table 2** and **Supplementary Table S2** list the results.

The human *COMT* gene for catechol-O-methyltransferase has, in its promoter, a clinically annotated SNP, rs777650793, whose association with human cardiovascular disease was documented by database ClinVar (Landrum et al., 2014). **Figure 2A** presents our prediction for this SNP, which is an excess of this protein. As a non-statistical validation of this prediction, we manually performed our primary keyword search, which resulted in an experimental study (He et al., 2011) on a rat model of human pathologies, which has identified *COMT* overexpression as a physiological marker of cerebral vasospasm. This correspondence between our prediction (**Figure 2A**) and these experimental data (He et al., 2011) can support the suitability of the results of our Web service (Ponomarenko et al., 2015) in the case of a study of the human non-neuropeptidergic system as *italicized* in **Supplementary Table S2**.

As for our secondary keyword search, it resulted in an *in vivo* experiment in a rat model of human behavior (Wilhelm et al., 2013), where a catechol-O-methyltransferase excess was a physiological marker of depression. Within the framework of the behavioral animal model (Wilhelm et al., 2013), we predicted the candidate SNP marker of the human tendency in subordination (**Supplementary Table S2**).

The human *DRD3* gene (dopamine receptor D3) carries SNP rs36211802 annotated by database ClinVar (Landrum et al., 2014), which associates it with hereditary essential tremor. This SNP can cause an excess of this receptor, according to our prediction given in **Figure 2B**. We validated this prediction by our primary keyword search, which found the original experimental data (Kosmowska et al., 2016) on resistance to the high-dose DRD3-agonist treatment of tremor in a laboratory rat model of this human pathology as *italicized* in **Supplementary Table S2**.

In addition, our secondary search revealed (**Supplementary Table S2**) that a DRD3 excess reduced both motor activity and behavioral motivation in a mouse model of human motor activity (Ikeda et al., 2013). This finding allows us to predict rs36211802 as a candidate SNP marker of the human tendency in subordination (**Supplementary Table S2**).

The human *HTR2C* gene encodes 5-hydroxytryptamine (serotonin) receptor 2C and carries SNP rs3813929, manifestation of which is an abnormal response to olanzapine (antipsychotic) according to database ClinVar (Landrum et al., 2014). For this SNP, we predict an excess of this serotonin receptor as shown in **Figure 2C**. Our primary keyword search pointed to the clinical data (Ellingrod et al., 2005) on an HTR2C excess caused by this SNP, whose manifestation is a resistance to olanzapine-caused increase in body mass. It is noteworthy that Tecott et al. (1995) reported that knockout mice ($5HT2C^{(-)/(-)}$) are obese, whereas Stahl (1998) observed eating behavior downregulation with a 5HT2C level increase. With this in mind, our prediction of the rs3813929-related 5HT2C excess

(**Figure 2C**) fits the clinical observation of the rs3813929-related resistance to olanzapine-caused increase in body mass (Ellingrod et al., 2005). This agreement between our prediction shown in **Figure 2C** and the clinical observations (Tecott et al., 1995; Stahl, 1998; Ellingrod et al., 2005) is consistent with our verification of our predictions of this type by electrophoretic mobility shift assays (EMSAs) under equilibrium (Savinkova et al., 2013) and non-equilibrium conditions (Drachkova et al., 2014) *in vitro*. Besides, this result is in agreement with our verification of our predictions on this subject using biosensor ProteON^{TM} (Bio-Rad Lab, United States) (Drachkova et al., 2012) and stopped-flow spectrometer SX.20 (Applied Photophysics, United Kingdom) (Arkova et al., 2014, 2017) in real-time mode. In addition, it fits our verification of our analogous predictions using human cell lines transfected with the pGL 4.10 vector (Promega, United States) (for a review, Ponomarenko M. et al., 2017). Finally, it is in line with our verification of our predictions on this subject using independent data from 60 experiments (for a review, see Ponomarenko et al., 2010) and by means of 43 known clinical SNP markers of human diseases (Ponomarenko et al., 2009) and 38 known genetic SNP markers of the breeding traits of animals and plants (Suslov et al., 2010). All these verification data can be a reason for the applicability of our Web-service (Ponomarenko et al., 2015) when the human genes relating to the non-neuropeptidergic system are studied, as *italicized* in **Supplementary Table S2**.

Our secondary keyword search yielded empirical data on two laboratory rat strains, which were bred for 60 generations for the presence and absence of high levels of stress-evoked aggression toward humans (Popova et al., 2010). According to these data, increases in both mRNA and protein levels were seen in the brains of non-aggressive rats in comparison with the aggressive ones (**Supplementary Table S2**). On this basis, we propose the candidate SNP marker for human tendency in subordination (**Supplementary Table S2**).

In total, we predicted 342 and 163 candidate SNP markers that can increase and decrease, respectively, the expression of the human proteins related to the non-neuropeptidergic system. Besides, these 505 predictions can be clustered as 240 and 265 candidate SNP markers for the human tendencies in dominance and subordination (**Table 2** and **Supplementary Table S2**). As readers can see in **Table 2**, these results are again consistent with our preliminary estimates (Chadaeva et al., 2017) that natural selection equally supports the human tendencies in dominance and subordination.

Candidate SNP Markers Near TBP-Binding Sites in the Promoter of the Human Genes Encoding Neurotrophinergic-System-Related Proteins (e.g., Growth Factors, Receptors)

We applied our public Web service (Ponomarenko et al., 2015) to study 2431 SNPs in 70 bp regions in front of the TSSs of 95 human genes encoding neurotrophinergic-system–related proteins, namely, adenylate cyclase-activating polypeptide 1 and its receptor (*ADCYAP1*s), artemin (*ARTN*), brain-derived neurotrophic factor (*BDNF*), cerebral dopamine neurotrophic factor (*CDNF*), ciliary neurotrophic factor (*CNTF*), fibroblast growth factors and their receptors (*FGF*s), Fms-related tyrosine kinases and their ligand (*FLT*s), glial-cell-derived neurotrophic factor (*GDNF*), GDNF family receptors (*GFR*s), glia maturation factors (*GMF*s), insulin like growth factors and their receptors (*IGF*s), interleukins as well as their receptors and signal transducers (*IL*s), leukemia-inhibitory factor (IL6-family cytokine) and its receptor (*LIF*s), nerve growth factor and its receptor (*NGF*s), neuregulins (*NRG*s), neuropilins (*NRP*s), neurturin (NRTN), neurotrophins (*NTF*s), neurotrophic receptor tyrosine kinases (*NTRK*s), oncostatin M and its receptor (*OSM*s), platelet-derived growth factor subunits and receptors (*PDGF*s), placental growth factor (PGF), persephin (PSPN), Ret receptor tyrosine kinase (*RET*), transforming growth factors β, its receptors and associated protein 1 (*TGFB*s), and vascular endothelial growth factors (*VEGF*s). We show our results in **Table 2** and **Supplementary Table S3**.

The human *FGFR2* gene (fibroblast growth factor receptor 2) contains two SNPs rs387906677 and rs886046768, which were clinically detected in patients with bent bone dysplasia syndrome and craniosynostosis, respectively, as documented by database ClinVar (Landrum et al., 2014). Readers can see in **Figures 2B**, **3A** how we predicted the FGFR2 deficiency in the case of rs387906677, whereas rs886046768 corresponds to an FGFR2 excess.

At first, our primary keyword search revealed an experimental report (Merrill et al., 2012) on a mouse model of human embryonic development, which linked bent bone dysplasia with reduced levels of FGFR2. Next, in the same way, we found the original experiment (Mansukhani et al., 2000) on mouse osteoblast cell culture *ex vivo* that points to FGFR2 as an inducer of apoptosis in these cells and an inhibitor of their differentiation, hyperactivity of which causes craniosynostosis-linked alterations in cell culture. As depicted in the figures, these independent findings confirm the validity of our predictions (**Figures 3A,B**) in the case of the neurotrophinergic system analysis, as *italicized* in **Supplementary Table S3**.

After this validation, our secondary keyword search yielded an article (Meyer et al., 2012) on FGFR2 deficiency as a physiological marker of delayed post-injury skin wound healing. Analogously, we found a biomedical paper (Baatar et al., 2002) on the injections of recombinant human FGFR2 around ulcers, which have accelerated ulcer healing in rats as an animal model of the human pathologies. On this basis, we predicted rs387906677 and rs886046768 as candidate SNP markers of the human tendencies in subordination and dominance, respectively (**Supplementary Table S3**).

The human *PDGFRA* gene encodes platelet-derived growth factor receptor α and contains SNP rs183431225 annotated by database ClinVar (Landrum et al., 2014) in connection with both idiopathic hypereosinophilic syndrome and gastrointestinal stromal tumor. **Figure 3C** presents our prediction for this SNP: overexpression of this receptor. Our primary keyword

search revealed two biomedical papers, one of which (Score et al., 2006) reports the PDGFRA excess as a marker of patients with hypereosinophilia, and another one (Hayashi et al., 2015) reveals reduced proliferation of gastrointestinal stromal tumor cells under the influence of a selective inhibitor of PDGFRA. Thus, these independent literature data support applicability of our predictions to the study of human genes encoding neurotrophinergic-system–related proteins as *italicized* in **Supplementary Table S3**.

Then, we did our secondary keyword search and found a mouse model of human behavior indicating that the PDGFRA overexpression causes oligodendrocyte-associated nociceptive hypersensitivity to neuropathic pain (Shi et al., 2016). That is why we assumed that rs183431225 is a candidate SNP marker of the human tendency in subordination (**Supplementary Table S3**).

The human *RET* gene codes for the Ret proto-oncogene, where two SNPs (rs10900297 and rs10900296) have been associated with three human diseases (renal adysplasia, Hirschsprung disease, and pheochromocytoma) as documented in database ClinVar (Landrum et al., 2014). As readers can see in **Figures 3D,E**, our predictions for these SNPs surprisingly correspond to over- and underexpression of this gene. Nevertheless, using our primary keyword search, we learned that both an excess (Sarin et al., 2014) and deficit (Bridgewater et al., 2008) of RET are known as physiological markers of renal adysplasia. In addition, both overexpression (Ishii et al., 2013) and underexpression (Zhan et al., 1999) of the *RET* gene can contribute to the pathogenesis of Hirschsprung disease. Finally, both increased (Huang et al., 2003) and decreased (Moore and Zaahl, 2012) levels of this proto-oncogene are often seen in pheochromocytoma. Thus, the above publications additionally validate our results (**Figures 3D,E**) as *italicized* in **Supplementary Table S3**.

Accordingly, we conducted a secondary keyword search and thus selected two animal models of human behavior. The rat model (Wang et al., 2017) associated the RET excess with hypersensitivity to neuropathic pain. In the mouse model (Golden et al., 2010), the RET deficit reduced epidermal innervation. Within the limitations of these models, we predicted two candidate SNP markers (rs10900297 and rs10900296) of the human tendency in subordination (**Supplementary Table S3**).

The human *TGFBR2* gene (transforming growth factor β receptor 2) contains SNP rs138010137, which occurs in patients with thoracic aortic aneurysm as documented in database ClinVar (Landrum et al., 2014). According to our prediction illustrated in **Figure 3F**, this SNP can reduce levels of receptor TGFBR2 in humans. Using a primary keyword search, we found an original work about the TGFBR2-deficient aortic aneurysm and aortic dissection as the specific forms of these pathologies (Angelov et al., 2017). As one can see, this is one more argument in favor of the applicability of our Web service (Ponomarenko et al., 2015) to research on the human genes related to the neurotrophinergic system as *italicized* in **Supplementary Table S3**.

Next, our secondary keyword search yielded a transgenic mouse model of human health (Martinez-Ferrer et al., 2010), in which the TGFBR2 deficit accelerates healing, closure, and resurfacing of skin wounds. For this reason, we suggest rs138010137 as a candidate SNP marker of the human tendency in dominance (**Supplementary Table S3**).

Summarizing all the above, we can see 506 candidate SNP markers predicted by this work in the case of human genes encoding the neurotrophinergic-system-related proteins (**Table 2** and **Supplementary Table S3**). These predictions can be grouped into 346 and 160 candidate SNP markers of the excess and deficiency of these proteins, respectively, as well as into 265 and 241 candidate SNP markers of the human tendencies in dominance and subordination (**Table 2**). Notably, the first of these dichotomies of SNPs in the human genome is statistically significantly uneven, whereas the second one is uniform. This is one more actual piece of evidence for the pressure of natural selection on the human neuron-specific genes, which equally supports the human tendencies in dominance and subordination, in agreement with our preliminary estimates (Chadaeva et al., 2017) as well as with all the other predictions of this work.

In silico Validation of All the Genome-Wide Predictions Made in This Work

Altogether, we analyzed 5052 SNPs within all the TBP-binding regions of all the promoters in front of all the protein-coding transcripts of all the 231 known human neuron-specific genes and selected 1108 candidate SNP markers that can significantly affect the affinity of TBP for these promoters (22%) as shown in the bottom row of **Table 2**. This result of our exhaustive whole-genome analysis of three systems of the human body (neuropeptidergic, non-neuropeptidergic, and neurotrophinergic) is consistent with both Haldane's dilemma (Haldane, 1957) and Kimura's neutral evolution theory (Kimura, 1968). Our *in silico* fivefold reduction in the number of unannotated SNPs for their subsequent *in vivo* studies is in line with the current need for reducing the cost of both experimental and clinical searches for valuable SNP markers in the human genome by trial and error through preliminary computer analysis of the known SNPs (Deplancke et al., 2016).

With this in mind, we selected all the 10 among 1,108 candidate SNP markers predicted in this work (**Figures 2**, **3** and **Supplementary Figure S1**), which are currently linked to the human diseases by public database ClinVar (Landrum et al., 2014). As described above, we non-statistically validated this set of our selected predictions by our primary keyword search in the public PubMed database (Lu, 2011). Essentially, this match between our 10 selected predictions and the found literature data is statistically significant at the level of $\alpha < 0.001$ according to the criterion of the equiprobable binomial distribution.

It is important to note that most of the candidate SNP markers that were marked in database ClinVar (Landrum et al., 2014) had a "Clinically insignificant" label because the number of patients with these candidate SNP markers varied from one to six, whereas for clinical significance it is necessary to use cohorts of several hundred patients. This observation

supports subsequent verification (using clinical protocols) of the candidate SNP markers predicted by this work. In this way, genotyping for the elite combat athletes in addition to the widely used textual psychological questionnaires for them (Tiric-Campara et al., 2012) could enrich personalized sports medicine.

In addition, we used the semicentennial bioinformatic tradition of comparing the actual frequencies of mutations for their various dichotomies [transitions versus transversions (Kimura, 1980), synonymous versus non-synonymous changes (Li et al., 1985), etc.]. To this end, we grouped all the 1108 predictions into 754 and 354 candidate SNP markers for the increase and decrease in the TBP binding affinity for promoters of the human neuron-related proteins, respectively (**Table 2**: N_{RES}, $N_{>}$ and $N_{<}$). This dichotomy contradicts the binomial distribution of the whole-genome ratio 4:1 of the SNPs reducing versus SNPs increasing affinity of the transcription factors for the human gene promoters (1000 Genomes Project Consortium et al., 2012) as neutral drift according to Haldane's dilemma (Haldane, 1957) and neutral evolution theory (Kimura, 1968), **Table 2**: $p(N_{<} = 4N_{>} = 4N_{RES}/5) < 0.000001$. This significant contradiction means the adaptive pressure of natural selection on the human neuron-specific genes is in line with the commonly accepted opinion about the adaptive role of both the nervous system and social behavior in the course of human origin and evolution. That is one more evolutionary argument for the reliability of our predictions made in this work.

Finally, by the same reasoning, we grouped all the 1,108 predictions into 556 and 552 candidate SNP markers for the human tendencies in dominance and subordination, respectively (**Table 2**: N_{RES}, $N_{\uparrow}$, and $N_{\downarrow}$). In contrast to the above dichotomy, this one corresponds to the highly probable H_0 hypothesis about the equiprobable binomial distribution of these candidate SNP markers for human social hierarchy [**Table 2**: $p(H_0: N_{\uparrow} = N_{\downarrow} = N_{RES}/2) > 0.9$]. This correspondence means that the pressure of natural selection proven above equally supports the human tendencies in dominance and subordination.

Notably, so that natural selection can control the human tendencies in dominance and subordination, it is necessary that this human tendencies can be inherited from generation to generation from parents to offspring. That is why, we *in vivo* validated our *in silico* predictions of this work in a mouse model of human inheritance as described below.

In vivo Validation of Our Predictions Using a Mouse Model of Human Inheritance

Each public Web service addresses a specific sort of regulatory SNP analysis (e.g., Bendl et al., 2016), and each has its specific advantages and disadvantages. Therefore, a comparison between the particular predictions and experimental data as an independent commonly accepted uniform platform (rather than between predictions of various Web services) needs to be a necessary step for prediction of candidate SNP markers *in silico* (Yoo et al., 2015; Ponomarenko M. et al., 2017). Keeping this in mind, we *in vivo* validated our *in silico* predictions on the equal natural-selection support of the human tendencies in dominance and subordination using a mouse model of human inheritance as described in the section "Materials and Methods." The obtained results are given in **Figure 4** and **Table 3**.

Figure 4 indicates that we completely reproduced the temporal pattern of both formation and maintenance of the social hierarchy in mouse pairs by means of both the number and duration of attacks and submissive poses.

As one can see in the first row "PT" of this table, 21 of 31 mouse males of the F1 hybrids carrying the PT × BALB\cLac genotype dominated over the male F1 hybrids of the PT × CBA/Lac genotype, and 10 mouse males of the PT × CBA/Lac genotype were dominant in the remaining pairs of the same combination. This actual difference between the F1 male hybrids PT × BALB/cLac and PT × CBA\Lac is characterized by the χ^2-score equal to 3.9, which is statistically significant at the level of $\alpha < 0.05$. In addition, we observed the same significant dominance of the BALB/cLac-related F1 hybrids over the CBA/Lac-related ones, when the maternal inbred strains were DD and YP (**Table 3**). In addition, in the cases of maternal inbred strains C57BL/6J and A/He, we found only a tendency for the same dominance, which was insignificant, possibly because of the insufficient number of the appropriate mouse male pairs studied regarding these maternal genotypes. Finally, the last row of **Table 3** represents the final result: the statistically significant majority of 79 among 115 BALB/cLac-related male hybrids achieved their dominant social status within their pairs with the CBA\Lac-related males of the same maternal inbred strains. This finding means that this mouse model of human inheritance reveals an ability of the tendencies in dominance and subordination to be inherited from generation to generation from parents to offspring and, therefore, to be an object of natural selection. This is the main genetic *in vivo* argument in favor of the reliability of our *in silico* predictions in this work.

Finally, looking through **Figure 4**, one can see that, in contrast to the first day of microsocial observation of a pair of adult male mice, which was characterized by numerous and lasting attacks of one mouse on the other, by the end of the second day a social hierarchy is established, with rare short-term ritualized attacks of dominant and/or ritualized submissive poses of a

TABLE 3 | The results of identification of inheritance of the murine tendencies in dominance and subordination.

Paternal genotype / Maternal genotype	BALB/cLac	CBA/Lac	χ^2 criterion	
			χ^2	Significance, α
PT	**21**	**10**	**3.90**	**0.05**
DD	**16**	**4**	**7.20**	**0.01**
C57BL/6J	13	7	1.80	> 0.1
YT	**16**	**5**	**5.76**	**0.025**
A/He	13	10	0.39	> 0.5
TOTAL	**79**	**36**	**16.08**	**0.001**

The number of male mice – that dominated over their neighbors within the framework of their pair – is indicated; statistically significant results are boldfaced.

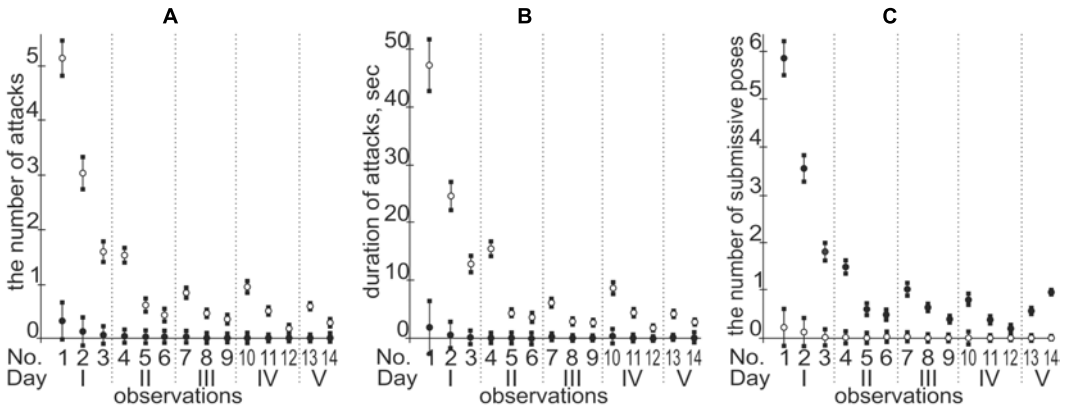


FIGURE 4 | A temporal pattern of both formation and maintenance of the social hierarchy in mouse pairs. Legend: ○ and ●, dominant and subordinate male mice, respectively; **(A)** the number of attacks; **(B)** duration of attacks (second); **(C)** the number of submissive poses; the circle and error bar denote the arithmetic mean and SD for 115 observations, respectively.

subordinate without any injuries and dangers for their lives and health (Lorenz, 2002). This is the main ecological benefit of establishing and maintaining social hierarchy, as a result of which natural selection equally supports the human tendencies for both dominance and subordination.

CONCLUSION

In this work, we analyzed only how SNPs can alter TBP's binding affinity for the human gene promoters, whereas more than 2500 human DNA-binding proteins are already known (Babu et al., 2004). Consequently, now there is a huge variety of Web services for studying the effects of SNPs on the binding affinity of the human gene promoters for these proteins and the respective phenotypic manifestations (e.g., Bendl et al., 2016). Their use can significantly expand the research capabilities in comparison with the use of our Web service alone (Ponomarenko et al., 2015).

The main finding of this work is that natural selection equally supports the human tendencies in dominance and subordination, which can be inherited from parents to offspring. The results of current study could be seen as an argument in favor of the genetic side within the century-old irreconcilable scientific debate on the nature of both aggressiveness and social hierarchy in humans [e.g., Freud (1921, 1930) and Lorenz (1964, 2002)]. Nevertheless, in the case of a random individual, these human tendencies can define the possible ranges (plasticity) of his/her aggressiveness and social rank rather that their actual levels, which depend on his/her continuous non-genetic social education from childhood to the oldest age (Markel, 2016). Certainly, this one is an argument in favor the other (non-genetic) side of the debate in question [e.g., Fromm (1941, 1973), Berkowitz (1962, 1993), Skinner (Rogers and Skinner, 1956; Skinner, 1981)]. According to recent reports on epigenetics (e.g., Merkulov et al., 2017), various stressors may cause epigenetic reprogramming of the individual genome and, in this way, modulate the actual levels of both individual aggressiveness and social status. Moreover, this reprogrammed pattern of the human genome is inherited from parents to offspring across at least two generations. Definitely, this notion equally supports both sides of the above debate as does our main finding in this work.

Finally, there are social mechanisms of transfer of the hierarchy status from parents to their offspring, previously described in macaques (Prud'Homme and Chapais, 1993), deer (Dusek et al., 2007), and hyenas (Engh et al., 2000). Clearly, the real effects of inherited genotypes on the human social hierarchy are much more complex, diverse, richer, brighter, and more interesting than our maximally simplified decision-making rule (see subsection "The Basic Decision-Making Rule" "Basic decision-making rule"). Nevertheless, at least a somewhat valid decision-making rule is necessary for application of the bioinformatic calculations to the genome-wide analysis *in silico*. In any case, as a computer-based prediction, each candidate SNP marker of the human tendencies in dominance and subordination predicted by this work should be experimentally verified in the studies of large human cohorts.

AUTHOR CONTRIBUTIONS

NK contributed to concept. DR and PP contributed to software. IC contributed to data compilation. ES and LS contributed to data analysis. MK, EK, LO, and AO performed the *in vivo* experiment. MP wrote the manuscript. VN performed the revised manuscript study design.

FUNDING

Manuscript writing was supported by the Russian Ministry of Science and Education within the 5-100 Excellence Program (for MP). The software development was supported by the project #0324-2019-0040 from the Russian Government Budget (for DR). The concept was supported by the integration project #0324-2018-0021 from the Presidium of the Siberian Branch of the Russian Academy of Sciences (for NK). The data

compilation was supported by project #18-34-00496 from the Russian Foundation for Basic Research (for IC). The data analysis was supported by project # 0324-2019-0042 from Russian Government Budget (for ES and LS). The study design was supported by project #16-54-12016 from the Russian Foundation for Basic Research (for VN). The *in vivo* experiment on animals was supported by a publicly funded project #0324-2019-0041 from Russian Government Budget (for MK, EK, LO, and AO) and implemented using the equipment of the Center for Genetic Resources of Laboratory Animals at ICG SB RAS, supported by the Russian Ministry of Education and Science (unique identifier of the project RFMEFI62117X0015).

ACKNOWLEDGMENTS

We are grateful to Shevchuk Editing[2] (Brooklyn, NY, United States) for English translation and editing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00073/full#supplementary-material

[2] http://www.shevchuk-editing.com

REFERENCES

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., et al. (2012). An integrated map of genetic variation from 1.092 human genomes. *Nature* 491, 56–65. doi: 10.1038/nature11632

Abbas, A., Lechevrel, M., and Sichel, F. (2006). Identification of new single nucleotid polymorphisms (SNP) in alcohol dehydrogenase class IV ADH7 gene within a French population. *Arch. Toxicol.* 80, 201–205. doi: 10.1007/s00204-005-0031-7

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: online mendelian inheritance in man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798. doi: 10.1093/nar/gku1205

Angelov, S. N., Hu, J. H., Wei, H., Airhart, N., Shi, M., and Dichek, D. A. (2017). TGF-β (Transforming Growth Factor-β) signaling protects the thoracic and abdominal aorta from angiotensin II-induced pathology by distinct mechanisms. *Arterioscler. Thromb. Vasc. Biol.* 37, 2102–2113. doi: 10.1161/atvbaha.117.309401

Arkova, O., Kuznetsov, N., Fedorova, O., and Savinkova, L. (2017). A real-time study of the interaction of TBP with a TATA box-containing duplex identical to an ancestral or minor allele of human gene LEP or TPI. *J. Biomol. Struct. Dyn.* 35, 3070–3081. doi: 10.1080/07391102.2016.1241190

Arkova, O. V., Kuznetsov, N. A., Fedorova, O. S., Kolchanov, N. A., and Savinkova, L. K. (2014). Real-time interaction between TBP and the TATA box of the human triosephosphate isomerase gene promoter in the norm and pathology. *Acta Naturae* 6, 36–40.

Baatar, D., Kawanaka, H., Szabo, I. L., Pai, R., Jones, M. K., Kitano, S., et al. (2002). Esophageal ulceration activates keratinocyte growth factor and its receptor in rats: implications for ulcer healing. *Gastroenterology* 122, 458–468. doi: 10.1053/gast.2002.31004

Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.* 14, 283–291. doi: 10.1016/j.sbi.2004.05.004

Belyaev, D. K. (1979). The Wilhelmine E. Key 1978 invitational lecture. Destabilizing selection as a factor in domestication. *J. Hered.* 70, 301–308. doi: 10.1093/oxfordjournals.jhered.a109263

Bendl, J., Musil, M., Stourac, J., Zendulka, J., Damborsky, J., and Brezovsky, J. (2016). PredictSNP2: a unified platform for accurately evaluating SNP effects by exploiting the different characteristics of variants in distinct genomic regions. *PLoS Comput. Biol.* 12:e1004962. doi: 10.1371/journal.pcbi.1004962

Berkowitz, L. (1962). *Aggression: A Social Psychological Analysis*. New York, NY: McGraw-Hill.

Berkowitz, L. (1993). *Aggression: Its Causes, Consequences, and Control*. Philadelphia, PA: Temple University Press.

Bragin, A. V., Osadchuk, L. V., and Osadchuk, A. V. (2006). The experimental model of establishment and maintenance of social hierarchy in laboratory mice. *Zh. Vyssh. Nerv. Deiat. Im. I P Pavlova* 56, 412–419.

Bridgewater, D., Cox, B., Cain, J., Lau, A., Athaide, V., Gill, P. S., et al. (2008). Canonical WNT/beta-catenin signaling is required for ureteric branching. *Dev. Biol.* 317, 83–94. doi: 10.1016/j.ydbio.2008.02.010

Chadaeva, I., Rasskazov, D., Sharypova, E., Savinkova, L., Ponomarenko, P., and Ponomarenko, M. (2017). Candidate SNP markers of social dominance, which may affect the affinity of the TATA-binding protein for human gene promoters. *Russ. J. Genet. Appl. Res.* 7, 523–537. doi: 10.1134/S2079059717050045

Chadaeva, I. V., Ponomarenko, P. M., Rasskazov, D. A., Sharypova, E. B., Kashina, E. V., Zhechev, D. A., et al. (2018). Candidate SNP markers of reproductive potential are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *BMC Genomics* 19(Suppl. 3). doi: 10.1186/s12864-018-4478-3

Colonna, V., Ayub, Q., Chen, Y., Pagani, L., Luisi, P., Pybus, M., et al. (2014). Human genomic regions with exceptionally high levels of population differentiation identified from 911 whole-genome sequences. *Genome Biol.* 15:R88. doi: 10.1186/gb-2014-15-6-r88

Delgadillo, R. F., Whittington, J. E., Parkhurst, L. K., and Parkhurst, L. J. (2009). The TATA-binding protein core domain in solution variably bends TATA sequences via a three-step binding mechanism. *Biochemistry* 48, 1801–1809. doi: 10.1021/bi8018724

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The genetics of transcription factor DNA binding variation. *Cell* 166, 538–554. doi: 10.1016/j.cell.2016.07.012

Dobzhansky, T. (1963). "Genetic entities in hominid evolution," in *Classification and Human Evolution*, ed. S. L. Washburn (Chicago, IL: Aldine), 347–362.

Drachkova, I., Savinkova, L., Arshinova, T., Ponomarenko, M., Peltek, S., and Kolchanov, N. (2014). The mechanism by which TATA-box polymorphisms associated with human hereditary diseases influence interactions with the ÒÀÒÀ-binding protein. *Hum. Mutat.* 35, 601–608. doi: 10.1002/humu.22535

Drachkova, I. A., Shekhovtsov, S. V., Peltek, S. E., Ponomarenko, P. M., Arshinova, T. V., Ponomarenko, M. P., et al. (2012). Surface plasmon resonance study of the interaction between the human TATA-box binding protein and the TATA element of the NOS2A gene promoter. *Vavilovskii Zhurnal Genet. Selektsii* 16, 391–396.

Dusek, A., Bartos, L., and Svecova, L. (2007). The effect of a mother's rank on her offspring's pre-weaning rank in farmed red deer. *Appl. Anim. Behav. Sci.* 103, 146–155. doi: 10.1016/j.applanim.2006.03.020

Ehrman, L., and Parsons, P. A. (1981). *Behavior Genetics and Evolution*. New York, NY: Mc Graw-Hill.

Eldakar, O. T., and Gallup, A. C. (2011). The group-level consequences of sexual conflict in multigroup populations. *PLoS One* 6:e26451. doi: 10.1371/journal.pone.0026451

Ellingrod, V. L., Perry, P. J., Ringold, J. C., Lund, B. C., Bever-Stille, K., Fleming, F., et al. (2005). Weight gain associated with the -759C/T polymorphism of the 5HT2C receptor and olanzapine. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* 134B, 76–78. doi: 10.1002/ajmg.b.20169

Engh, A. L., Esch, K., Smale, L., and Holekamp, K. E. (2000). Mechanisms of maternal rank "inheritance" in the spotted hyaena, *Crocuta crocuta*. *Anim. Behav.* 60, 323–332. doi: 10.1006/anbe.2000.1502

Freud, S. (1921). *Massenpsychologie und Ich-Analyse*. Vienna: International Psychoanalytic Publishing House.

Freud, S. (1930). *Civilization and its Discontents*. London: Hogarth.

Fromm, E. (1941). *Escape from Freedom*. New York, NY: Rinehart & Co.

Fromm, E. (1973). *The Anatomy of Human Destructiveness*. New York, NY: Holt, Rinehart and Winston.

Golden, J. P., Hoshi, M., Nassar, M. A., Enomoto, H., Wood, J. N., Milbrandt, J., et al. (2010). RET signaling is required for survival and normal function of nonpeptidergic nociceptors. *J. Neurosci.* 30, 3983–3994. doi: 10.1523/jneurosci.5930-09.2010

Gunbin, K. V., Ponomarenko, M. P., Suslov, V. V., Gusev, F., Fedonin, G. G., and Rogaev, E. I. (2018). Evolution of brain active gene promoters in human lineage towards the increased plasticity of gene regulation. *Mol. Neurobiol.* 55, 1871–1904. doi: 10.1007/s12035-017-0427-4

Haeussler, M., Raney, B. J., Hinrichs, A. S., Clawson, H., Zweig, A. S., Karolchik, D., et al. (2015). Navigating protected genomics data with UCSC genome browser in a box. *Bioinformatics* 31, 764–766. doi: 10.1093/bioinformatics/btu712

Haldane, J. B. S. (1957). The cost of natural selection. *J. Genet.* 55, 511–524. doi: 10.1007/bf02984069

Hayashi, Y., Bardsley, M. R., Toyomasu, Y., Milosavljevic, S., Gajdos, G. B., Choi, K. M., et al. (2015). Platelet-derived growth factor receptor-α regulates proliferation of gastrointestinal stromal tumor cells with mutations in KIT by stabilizing ETV1. *Gastroenterology* 149, 420–432.e16. doi: 10.1053/j.gastro.2015.04.006

He, Z., Sun, X., Guo, Z., and Zhang, J. H. (2011). Expression and role of COMT in a rat subarachnoid hemorrhage model. *Acta Neurochir. Suppl.* 110, 181–187. doi: 10.1007/978-3-7091-0353-1_32

Hein, M., and Graver, S. (2013). Tumor cell response to bevacizumab single agent therapy in vitro. *Cancer Cell Int.* 13:94. doi: 10.1186/1475-2867-13-94

Hinde, R. A. (1970). *Animal Behaviour*. New York, NY: McGraw-Hill.

Hu, J., Locasale, J. W., Bielas, J. H., O'Sullivan, J., Sheahan, K., Cantley, L. C., et al. (2013). Heterogeneity of tumor-induced gene expression changes in the human metabolic network. *Nat. Biotechnol.* 31, 522–529. doi: 10.1038/nbt.2530

Huang, S. C., Torres-Cruz, J., Pack, S. D., Koch, C. A., Vortmeyer, A. O., Mannan, P., et al. (2003). Amplification and overexpression of mutant RET in multiple endocrine neoplasia type 2-associated medullary thyroid carcinoma. *J. Clin. Endocrinol. Metab.* 88, 459–463. doi: 10.1210/jc.2002-021254

Ikeda, E., Matsunaga, N., Kakimoto, K., Hamamura, K., Hayashi, A., Koyanagi, S., et al. (2013). Molecular mechanism regulating 24-hour rhythm of dopamine D3 receptor expression in mouse ventral striatum. *Mol. Pharmacol.* 83, 959–967. doi: 10.1124/mol.112.083535

Ishii, K., Doi, T., Inoue, K., Okawada, M., Lane, G. J., Yamataka, A., et al. (2013). Correlation between multiple RET mutations and severity of Hirschsprung's disease. *Pediatr. Surg. Int.* 29, 157–163. doi: 10.1007/s00383-012-3196-1

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., et al. (2010). Variation in transcription factor binding among humans. *Science* 328, 232–235. doi: 10.1126/science.1183621

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* 217, 624–626. doi: 10.1038/217624a0

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/bf01731581

Kosmowska, B., Wardas, J., Głowacka, U., Ananthan, S., and Ossowska, K. (2016). Pramipexole at a low dose induces beneficial effect in the harmaline-induced model of essential tremor in rats. *CNS Neurosci. Ther.* 22, 53–62. doi: 10.1111/cns.12467

Kulikov, A. V., Bazhenova, E. Y., Kulikova, E. A., Fursenko, D. V., Trapezova, L. I., Terenina, E. E., et al. (2016). Interplay between aggression, brain monoamines and fur color mutation in the American mink. *Genes Brain Behav.* 15, 733–740. doi: 10.1111/gbb.12313

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985. doi: 10.1093/nar/gkt1113

Leschner, S., Deyneko, I. V., Lienenklaus, S., Wolf, K., Bloecker, H., Bumann, D., et al. (2012). Identification of tumor-specific *Salmonella typhimurium* promoters and their regulatory logic. *Nucleic Acids Res.* 40, 2984–2994. doi: 10.1093/nar/gkr1041

Li, W. H., Wu, C. I., and Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2, 150–174. doi: 10.1093/oxfordjournals.molbev.a040343

Lorenz, K. (1964). "Ritualized aggression," in *The Natural History of Aggression*, eds J. D. Carthy and F. J. Ebling (New York, NY: Academic Press), 39–50.

Lorenz, K. (2002). *On Aggression*. Hove: Psychology Press.

Lu, Z. (2011). PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* 2011:baq036. doi: 10.1093/database/baq036

Mansukhani, A., Bellosta, P., Sahni, M., and Basilico, C. (2000). Signaling by fibroblast growth factors (FGF) and fibroblast growth factor receptor 2 (FGFR2)-activating mutations blocks mineralization and induces apoptosis in osteoblasts. *J. Cell Biol.* 149, 1297–1308. doi: 10.1083/jcb.149.6.1297

Markel, A. L. (2016). Biosocial base of aggressiveness and aggressive behavior. *Zh. Vyssh. Nerv. Deiat. Im. I P Pavlova* 66, 669–681. doi: 10.7868/S0044467716060071

Martianov, I., Viville, S., and Davidson, I. (2002). RNA polymerase II transcription in murine cells lacking the TATA binding protein. *Science* 298, 1036–1039. doi: 10.1126/science.1076327

Martinez-Ferrer, M., Afshar-Sherif, A. R., Uwamariya, C., de Crombrugghe, B., Davidson, J. M., and Bhowmick, N. A. (2010). Dermal transforming growth factor-beta responsiveness mediates wound contraction and epithelial closure. *Am. J. Pathol.* 176, 98–107. doi: 10.2353/ajpath.2010.090283

Merkulov, V. M., Merkulova, T. I., and Bondar, N. P. (2017). Mechanisms of brain glucocorticoid resistance in stress-induced psychopathologies. *Biochemistry* 82, 351–365. doi: 10.1134/S0006297917030142

Merrill, A. E., Sarukhanov, A., Krejci, P., Idoni, B., Camacho, N., Estrada, K. D., et al. (2012). Bent bone dysplasia-FGFR2 type, a distinct skeletal disorder, has deficient canonical FGF signaling. *Am. J. Hum. Genet.* 90, 550–557. doi: 10.1016/j.ajhg.2012.02.005

Meyer, M., Muller, A. K., Yang, J., Moik, D., Ponzio, G., Ornitz, D. M., et al. (2012). FGF receptors 1 and 2 are key regulators of keratinocyte migration in vitro and in wounded skin. *J. Cell Sci.* 125, 5690–5701. doi: 10.1242/jcs.108167

Michopoulos, V., Higgins, M., Toufexis, D., and Wilson, M. E. (2012). Social subordination produces distinct stress-related phenotypes in female rhesus monkeys. *Psychoneuroendocrinology* 37, 1071–1085. doi: 10.1016/j.psyneuen.2011.12.004

Mogno, I., Vallania, F., Mitra, R. D., and Cohen, B. A. (2010). TATA is a modular component of synthetic promoters. *Genome Res.* 20, 1391–1397. doi: 10.1101/gr.106732.110

Moore, A. J. (2013). Genetic influences on social dominance: cow wars. *Heredity* 110, 1–2. doi: 10.1038/hdy.2012.85

Moore, S. W., and Zaahl, M. (2012). The Hirschsprung's-multiple endocrine neoplasia connection. *Clinics* 67, 63–67. doi: 10.6061/clinics/2012(Sup01)12

Ni, Y., Hall, A. W., Battenhouse, A., and Iyer, V. R. (2012). Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data. *BMC Genet.* 13:46. doi: 10.1186/1471-2156-13-46

Ponomarenko, M., Mironova, V., Gunbin, K., and Savinkova, L. (2013). "Hogness box," in *Brenner's Encyclopedia of Genetics*, 2nd Edn, Vol. 3, eds S. Maloy and K. Hughes (San Diego, CA: Academic Press), 491–494. doi: 10.1016/B978-0-12-374984-0.00720-8

Ponomarenko, M., Rasskazov, D., Arkova, O., Ponomarenko, P., Suslov, V., Savinkova, L., et al. (2015). How to use SNP_TATA_Comparator to find a significant change in gene expression caused by the regulatory SNP of this gene's promoter via a change in affinity of the TATA-binding protein for this promoter. *Biomed Res. Int.* 2015:359835. doi: 10.1155/2015/359835

Ponomarenko, M., Rasskazov, D., Chadaeva, I., Sharypova, E., Ponomarenko, P., Arkova, O., et al. (2017). SNP_TATA_Comparator: genomewide landmarks for preventive personalized medicine. *Front. Biosci.* 9, 276–306. doi: 10.2741/s488

Ponomarenko, M. P., Ponomarenko, J. V., Frolov, A. S., Podkolodnaya, O. A., Vorobyev, D. G., Kolchanov, N. A., et al. (1999). Oligonucleotide frequency matrices addressed to recognizing functional DNA sites. *Bioinformatics* 15, 631–643. doi: 10.1093/bioinformatics/15.7.631

Ponomarenko, P., Chadaeva, I., Rasskazov, D. A., Sharypova, E., Kashina, E. V., Drachkova, I., et al. (2017). Candidate SNP markers of familial and sporadic Alzheimer's diseases are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Front. Aging Neurosci.* 9:231. doi: 10.3389/fnagi.2017.00231

Ponomarenko, P., Rasskazov, D., Suslov, V., Sharypova, E., Savinkova, L., Podkolodnaya, O., et al. (2016). Candidate SNP markers of chronopathologies are predicted by a significant change in the affinity of TATA-binding protein for human gene promoters. *Biomed Res. Int.* 2016:8642703. doi: 10.1155/2016/8642703

Ponomarenko, P. M., Ponomarenko, M. P., Drachkova, I. A., Lysova, M. V., Arshinova, T. V., Savinkova, L. K., et al. (2009). Prediction of the affinity of the TATA-binding protein to TATA boxes with single nucleotide polymorphisms. *Mol. Biol.* 43, 472–479. doi: 10.1134/S0026893309030157

Ponomarenko, P. M., Savinkova, L. K., Drachkova, I. A., Lysova, M. V., Arshinova, T. V., Ponomarenko, M. P., et al. (2008). A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism. *Dokl. Biochem. Biophys.* 419, 88–92. doi: 10.1134/S1607672908020117

Ponomarenko, P. M., Suslov, V. V., Savinkova, L. K., Ponomarenko, M. P., and Kolchanov, N. A. (2010). A precise equilibrium equation for four steps of binding between TBP and TATA-box allows for the prediction of phenotypical expression upon mutation. *Biophysics* 55, 358–369. doi: 10.1134/S0006350910030036

Popova, N. K., Naumenko, V. S., Kozhemyakina, R. V., and Plyusnina, I. Z. (2010). Functional characteristics of serotonin 5-HT2A and 5-HT2C receptors in the brain and the expression of the 5-HT2A and 5-HT2C receptor genes in aggressive and non-aggressive rats. *Neurosci. Behav. Physiol.* 40, 357–361. doi: 10.1007/s11055-010-9264-x

Prud'Homme, J., and Chapais, B. (1993). Aggressive interventions and matrilineal dominance relations in semifree-ranging Barbary macaques (*Macaca sylvanus*). *Primates* 34, 271–283. doi: 10.1007/BF02382621

Rogers, C. R., and Skinner, B. F. (1956). Some issues concerning the control of human behavior: a symposium. *Science* 124, 1057–1066. doi: 10.1126/science.124.3231.1057

Rowell, T. E. (1974). The concept of social dominance. *Behav. Neural Biol.* 11, 131–154. doi: 10.1016/S0091-6773(74)90289-2

Sarin, S., Boivin, F., Li, A., Lim, J., Svajger, B., Rosenblum, N. D., et al. (2014). β-Catenin overexpression in the metanephric mesenchyme leads to renal dysplasia genesis via cell-autonomous and non-cell-autonomous mechanisms. *Am. J. Pathol.* 184, 1395–1410. doi: 10.1016/j.ajpath.2014.01.018

Savinkova, L., Drachkova, I., Arshinova, T., Ponomarenko, P., Ponomarenko, M., and Kolchanov, N. (2013). An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. *PLoS One* 8:e54626. doi: 10.1371/journal.pone.0054626

Score, J., Curtis, C., Waghorn, K., Stalder, M., Jotterand, M., Grand, F. H., et al. (2006). Identification of a novel imatinib responsive KIF5B-PDGFRA fusion gene following screening for PDGFRA overexpression in patients with hypereosinophilia. *Leukemia* 20, 827–832. doi: 10.1038/sj.leu.2404154

Serova, L. I., Kozlova, O. N., and Naumenko, E. V. (1991). The significance of the genotype and some individual behavioral features for the manifestation of the dominant phenotype in mice in micropopulations. *Zh. Vyssh. Nerv. Deyat. Im. I P Pavlova* 41, 79–84.

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., et al. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311. doi: 10.1093/nar/29.1.308

Shi, Y., Shu, J., Liang, Z., Yuan, S., and Tang, S. J. (2016). EXPRESS: oligodendrocytes in HIV-associated pain pathogenesis. *Mol. Pain* 12:1744806916656845. doi: 10.1177/1744806916656845

Skinner, B. F. (1981). Selection by consequences. *Science* 213, 501–504. doi: 10.1126/science.7244649

Smeets, C. J., Jezierska, J., Watanabe, H., Duarri, A., Fokkens, M. R., Meijer, M., et al. (2015). Elevated mutant dynorphin A causes Purkinje cell loss and motor dysfunction in spinocerebellar ataxia type 23. *Brain* 138, 2537–2552. doi: 10.1093/brain/awv195

Stahl, S. M. (1998). Neuropharmacology of obesity: my receptors made me eat it. *J. Clin. Psychiatry* 59, 447–448. doi: 10.4088/JCP.v59n0901

Suslov, V. V., Ponomarenko, P. M., Ponomarenko, M. P., Drachkova, I. A., Arshinova, T. V., Savinkova, L. K., et al. (2010). TATA box polymorphisms in genes of commercial and laboratory animals and plants associated with selectively valuable traits. *Russ. J. Genet.* 46, 394–403. doi: 10.1134/S1022795410040022

Szklarczyk, K., Korostynski, M., Golda, S., Solecki, W., and Przewlocki, R. (2012). Genotype-dependent consequences of traumatic stress in four inbred mouse strains. *Genes Brain Behav.* 11, 977–985. doi: 10.1111/j.1601-183X.2012.00850.x

Tecott, L. H., Sun, L. M., Akana, S. F., Strack, A. M., Lowenstein, D. H., Dallman, M. F., et al. (1995). Eating disorder and epilepsy in mice lacking 5-HT2C serotonin receptors. *Nature* 374, 542–546. doi: 10.1038/374542a0

Telenti, A., Pierce, L. C., Biggs, W. H., di Iulio, J., Wong, E. H., Fabani, M. M., et al. (2016). Deep sequencing of 10,000 human genomes. *Proc. Natl. Acad. Sci. U.S.A.* 113, 11901–11906. doi: 10.1073/pnas.1613365113

Tiric-Campara, M., Tupkovic, E., Mazalovic, E., Karalic, E., Biscevic, M., Djelilovic-Vranic, J., et al. (2012). Correlation of aggressiveness and anxiety in fighting sports. *Med. Arh.* 66, 116–121. doi: 10.5455/medarh.2012.66.116-121

Varzari, A., Tudor, E., Bodrug, N., Corloteanu, A., Axentii, E., and Deyneko, I. V. (2018). Age-specific association of CCL5 gene polymorphism with pulmonary tuberculosis: a case-control study. *Genet. Test. Mol. Biomarkers* 22, 281–287. doi: 10.1089/gtmb.2017.0250

Waardenberg, A. J., Basset, S. D., Bouveret, R., and Harvey, R. P. (2015). CompGO: an R package for comparing and visualizing Gene Ontology enrichment differences between DNA binding experiments. *BMC Bioinformatics* 16:275. doi: 10.1186/s12859-015-0701-2

Wang, H. J., Song, G., Liang, J., Gao, Y. Y., and Wang, C. J. (2017). Involvement of integrin β1/FAK signaling in the analgesic effects induced by glial cell line-derived neurotrophic factor in neuropathic pain. *Brain Res. Bull.* 135, 149–156. doi: 10.1016/j.brainresbull.2017.10.008

Wilhelm, C. J., Choi, D., Huckans, M., Manthe, L., and Loftis, J. M. (2013). Adipocytokine signaling is altered in Flinders sensitive line rats, and adiponectin correlates in humans with some symptoms of depression. *Pharmacol. Biochem. Behav.* 103, 643–651. doi: 10.1016/j.pbb.2012.11.001

Wu, J., Wu, M., Li, L., Liu, Z., Zeng, W., and Jiang, R. (2016). dbWGFP: a database and web server of human whole-genome single nucleotide variants and their functional predictions. *Database* 2016:baw024. doi: 10.1093/database/baw024

Yoo, S. S., Jin, C., Jung, D. K., Choi, Y. Y., Choi, J. E., Lee, W. K., et al. (2015). Putative functional variants of XRCC1 identified by RegulomeDB were not associated with lung cancer risk in a Korean population. *Cancer Genet.* 208, 19–24. doi: 10.1016/j.cancergen.2014.11.004

Zapata, I., Serpell, J. A., and Alvarez, C. E. (2016). Genetic mapping of canine fear and aggression. *BMC Genomics* 17:572. doi: 10.1186/s12864-016-2936-3

Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T., and Flicek, P. R. (2015). The Ensembl regulatory build. *Genome Biol.* 16:56. doi: 10.1186/s13059-015-0621-5

Zhan, J., Xiu, Y., Gu, J., Fang, Z., and Hu, X. L. (1999). Expression of RET proto-oncogene and GDNF deficit in Hirschsprung's disease. *J. Pediatr. Surg.* 34, 1606–1609. doi: 10.1016/s0022-3468(99)90626-7

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 01 March 2019
doi: 10.3389/fgene.2019.00097

Pan-Cancer Analysis of TCGA Data Revealed Promising Reference Genes for qPCR Normalization

George S. Krasnov, Anna V. Kudryavtseva, Anastasiya V. Snezhkina, Valentina A. Lakunina, Artemy D. Beniaminov, Nataliya V. Melnikova and Alexey A. Dmitriev**

Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia

OPEN ACCESS

Edited by:
Yuriy L. Orlov, Institute of Cytology and Genetics (RAS), Russia

Reviewed by:
Alexey V. Pindyurin, Institute of Molecular and Cellular Biology (RAS), Russia
Vladimir Kiselev, Wellcome Trust Sanger Institute (WT), United Kingdom
Shengjie Yang, NorthShore University HealthSystem, United States

***Correspondence:**
George S. Krasnov
gskrasnov@mail.ru
Alexey A. Dmitriev
alex_245@mail.ru

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *31 October 2018*
Accepted: *29 January 2019*
Published: *01 March 2019*

Citation:
Krasnov GS, Kudryavtseva AV, Snezhkina AV, Lakunina VA, Beniaminov AD, Melnikova NV and Dmitriev AA (2019) Pan-Cancer Analysis of TCGA Data Revealed Promising Reference Genes for qPCR Normalization. Front. Genet. 10:97. doi: 10.3389/fgene.2019.00097

Quantitative PCR (qPCR) remains the most widely used technique for gene expression evaluation. Obtaining reliable data using this method requires reference genes (RGs) with stable mRNA level under experimental conditions. This issue is especially crucial in cancer studies because each tumor has a unique molecular portrait. The Cancer Genome Atlas (TCGA) project provides RNA-Seq data for thousands of samples corresponding to dozens of cancers and presents the basis for assessment of the suitability of genes as reference ones for qPCR data normalization. Using TCGA RNA-Seq data and previously developed CrossHub tool, we evaluated mRNA level of 32 traditionally used RGs in 12 cancer types, including those of lung, breast, prostate, kidney, and colon. We developed an 11-component scoring system for the assessment of gene expression stability. Among the 32 genes, *PUM1* was one of the most stably expressed in the majority of examined cancers, whereas *GAPDH*, which is widely used as a RG, showed significant mRNA level alterations in more than a half of cases. For each of 12 cancer types, we suggested a pair of genes that are the most suitable for use as reference ones. These genes are characterized by high expression stability and absence of correlation between their mRNA levels. Next, the scoring system was expanded with several features of a gene: mutation rate, number of transcript isoforms and pseudogenes, participation in cancer-related processes on the basis of Gene Ontology, and mentions in PubMed-indexed articles. All the genes covered by RNA-Seq data in TCGA were analyzed using the expanded scoring system that allowed us to reveal novel promising RGs for each examined cancer type and identify several "universal" pan-cancer RG candidates, including *SF3A1*, *CIAO1*, and *SFRS4*. The choice of RGs is the basis for precise gene expression evaluation by qPCR. Here, we suggested optimal pairs of traditionally used RGs for 12 cancer types and identified novel promising RGs that demonstrate high expression stability and other features of reliable and convenient RGs (high expression level, low mutation rate, non-involvement in cancer-related processes, single transcript isoform, and absence of pseudogenes).

Keywords: cancer, gene expression, reference genes, quantitative PCR, data normalization, RNA-Seq, TCGA, CrossHub

INTRODUCTION

Quantitative PCR (qPCR) is the most widely used technique for quantification of gene expression. qPCR is rapid, has a very high dynamic range of mRNA level quantification and provides a measurement of even small gene expression alterations in a large number of samples. The most common and convenient approach for qPCR data normalization assumes mRNA quantification of a reference gene (RG) with stable expression level between the samples under study (Huggett et al., 2005). It is a bottleneck of qPCR, and the reliability of qPCR results strongly depends on the selection of appropriate RGs. This issue becomes more acute when it comes to assessing the moderate changes in the mRNA level of target genes (<2-fold).

The problem of selecting appropriate RGs is especially crucial in cancer studies because of the presence of several molecular subtypes within a histological type and, moreover, a unique molecular portrait of each tumor (Janssens et al., 2004). Despite the fact that almost 30 years have passed since the moment when the issue of picking appropriate RGs had arisen, there is still no consensus (Janssens et al., 2004; Rubie et al., 2005; Gur-Dedeoglu et al., 2009; Ibusuki et al., 2013; Zhao et al., 2018). Many studies indicate that most frequently used RGs (*GAPDH*, *ACTB*, *B2M*, *etc.*) have a wide but limited field of applicability: they should not be illegibly used for a wide spectrum of diseases or stress conditions (Barber et al., 2005; Rubie et al., 2005; Kozera and Rapacz, 2013; Chapman and Waldenstrom, 2015). To increase the reliability of qPCR data, one should use at least two or more RGs that are not co-expressed with each other (Chapman and Waldenstrom, 2015). The most rigorous approach is to analyze a panel of 5–20 RGs and choose those with the most stable expression for a current study. Several tools have been developed for these purposes: geNorm (Vandesompele et al., 2002), NormFinder (Andersen et al., 2004), BestKeeper (Pfaffl et al., 2004). However, the vast part of researchers do not perform the analysis of RG suitability and just rely on the existing literature data concerning the object of study (Chapman and Waldenstrom, 2015).

Whole-transcriptomic data allow us to look at the problem from the other side. RNA-Seq opens up great opportunities for a complex expression analysis and identifying trends in the mRNA level changes of groups of genes between the samples. RNA-Seq data are free of bias that comes from the instability of RG expression. The most common RNA-Seq data normalization strategy is based on the assumption that the mRNA level of the majority of genes is stable. This method is implemented in popular RNA-Seq differential expression analysis packages, including edgeR [trimmed mean of M-values method, TMM; Robinson et al., 2010], DESeq2 (Love et al., 2014), and others. There are other normalization strategies: by total read count, by upper quartile or median values, FPKM/RPKM, TPM, "remove unwanted variation" (RUV) (Risso et al., 2014); as well as machine-learning approaches: RNA-Seq by Expectation-Maximization (RSEM) (Li and Dewey, 2011) and Sailfish (Patro et al., 2014). Despite the diversity of the methods, in most cases, they give rather similar results, which differ by 20–30%, with the exception of some cases when the expression of half or more of genes is changed significantly (Dillies et al., 2013; Li et al., 2015; Zyprych-Walczak et al., 2015; Evans et al., 2018).

Analysis of highly representative RNA-Seq and microarray datasets is very attractive in terms of the identifying stably expressed RGs for human (Popovici et al., 2009; Tilli et al., 2016; Chen et al., 2017; Chim et al., 2017; Hoang et al., 2017) or other organisms (Alexander et al., 2012; Carmona et al., 2017; Zhou et al., 2017). This approach is valuable for the detection of novel housekeeping gene candidates with constitutively stable mRNA level.

In 2016, Tilli et al. suggested a strategy including the large-scale screening of potential RGs from RNA-Seq data with further validation by qPCR and applied it for breast cancer (Tilli et al., 2016). The authors analyzed datasets of The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) and found that several non-traditional RGs, *CCSER2, SYMPK, ANKRD17*, as well as known RG *PUM1* demonstrated the least expression variability in breast cancer samples and normal tissues (Tilli et al., 2016). The similar approach was realized by Chen et al. for the identification of reference mRNA and miRNA suitable for human esophageal squamous cell carcinoma studies (Chen et al., 2017). It allowed authors to identify non-standard RG candidates—*DDX5*, *LAPTM4A*, *P4HB*, and *RHOA*.

TCGA is the largest resource in the field of cancer biology that is aimed at the discovery of the molecular features of various cancer types (https://cancergenome.nih.gov/). TCGA database includes genomic, transcriptomic, and epigenetic data for 33 human cancer types represented with more than 11,000 individual samples. In the present work, we analyzed TCGA transcriptome sequencing data in order to evaluate the expression stability of widely used RGs and identify novel RG candidates in 12 most common cancer types. The use of representative TCGA sample sets allows us to pay extra attention to the overall stability of mRNA level and presence of outliers, the cases of dramatic expression "blow up" or falling down in single samples. Besides the data on mRNA level, we took into account if this is a well-studied gene or not (by evaluating the number of mentions in PubMed-indexed titles/abstracts), if a gene is involved in cancer-associated biological processes like cell cycle, differentiation, and adhesion (using Gene Ontology). Additionally, we evaluated if a gene is highly mutated (using TCGA data on somatic mutations) that indicates its implication in cancer. Also, we tried to minimize the number of pseudogenes and alternatively spliced transcripts in order to improve usability: the presence of pseudogenes makes it difficult to pick up cDNA-specific primer pairs, and the presence of alternative transcripts complicates the expression analysis and may lead to flawed results. We integrated all the parameters listed above into a single scoring system. Finally, we looked for genes that demonstrate cross-tissue expression stability and may represent "universal" pan-cancer RGs.

MATERIALS AND METHODS

In the present work, we focused on TCGA data for 12 cancer types for which RNA-Seq data were available for representative

sample sets: at least 100 tumor (T) and 20 normal (N) tissue samples. The data were processed with a modified version of CrossHub (Krasnov et al., 2016), a tool for the multi-way analysis of TCGA transcriptomic and genomic data. Read counts data were downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/) and normalized using the TMM method and then recalculated for 1 million library size. The derived CPM (read counts per million) values were used as a measure of mRNA level of a gene for further expression stability analysis.

In order to assess gene expression stability, we developed a scoring system, which included several components (S_i) responsible for T-N expression level difference, expression level stability within pools of N and T samples, and correlations of mRNA level with clinical and pathological characteristics [disease stage, TNM (tumor, node, metastasis) classification, follow-up status]. Each scoring component S_i takes values from 0 to 100. All S_i are taken with different weights (W_i), which reflect the importance of component. Overall expression scoring S^{Exp} is calculated as follows:

$$S^{\mathrm{Exp}} = \left(\prod_{i=1}^{N} (S_i + \mathrm{CA}_i)^{\mathrm{W}_i} \right)^{1/\sum_{i=1}^{N} W_i}$$

where:

- CA_i is a constant summand, which is used to mitigate the impact of zero values of S_i;
- W_i is weight of a component S_i ($i = 1 \ldots \mathrm{N}$);
- N is a number of components, $N = 48$.

Values of these parameters are presented in **Table 1**.

Each individual component S_i is calculated with a common parametric formula:

$$S_i = \frac{100}{1 + \mathrm{Sq} \times \left(\frac{\max(x - \mathrm{IV}; 0)}{\mathrm{IP} - \mathrm{IV}} \right)^{\mathrm{CS}}}$$

This formula provides a (1-sigma)-like function with a customizable inflection point, tilt, and region of maximal score values. The function takes values from 0 to 100. Here:

- x is a variable to be scored (see **Table 1**).
- IV is an "ideal value." All cases with $x \leq$ IV would produce the maximum score (100). For example, S_{DP}, the component responsible for T-N expression level fold change (see **Table 1**) would be equal to 100 for any $\log_2\mathrm{FC_P}$ between -0.05 and $+0.05$ since IV = 0.05.
- IP is an "inflection point." In this point, there is the maximum decrease rate of S_i. When x is equal to IP and Sq = 1 (Sq takes these value for most S_i), the scoring component $S_i = 50$. Ideally, IP value should reflect the marginally acceptable value of x. For example, the relative standard deviation of gene expression (RelSD) values from 0 to 0.25 are appropriate, but RelSD = 0.4 … 0.5 is almost unacceptable. For the corresponding component (S_{EStD}), we chose IV = 0.1 and IP = 0.3 (see **Table 1**).
- CS is a "curve slope." The greater CS value, the stronger S_i decrease rate. Higher CS values should be assigned to more important scoring components.
- Sq is a "Squeeze," this is an auxiliary parameter. For most scoring components, it is equal to 1.

All scoring components S_i and parameters (IV, IP, CS, Sq) are presented in **Table 1**. The derived scoring functions are shown in **Figure 1**.

Two components, S_{DP} and S_{DL}, are responsible for T-N expression level difference. This is the major factor of RG suitability. S_{DP} is calculated for pooled, and S_{DL}–for paired samples. Hence, we applied the strongest scoring parameters (IV = 0.05, IP = 0.25, CS = 2.5) and assigned high weight (W = 4) for these two components. S_{DP} (or S_{DL}) would be equal to 50 if the absolute value of average $\log_2\mathrm{FC_P}$ (or $\log_2\mathrm{FC_L}$) is equal to IP = 0.25, i.e., fold change between tumor and normal is about 20%. We chose IV = 0.05–0.1 for all the components that are responsible for expression level (S_{DP}, S_{DL}, S_{DoO}, S_{DoU}, S_{DLc}, S_{EStD}, S_{EoH}, S_{EoL}). This means that 5–10% mRNA level changes are ignored.

S_{DP} and S_{DL} are calculated using the trimmed means of either CPM (pooled sample) or $\log_2\mathrm{FC_L}$ (paired samples). Only values from 10 to 90th percentiles are included. To take into account T-N expression outliers, we added two other scorings, S_{DoO} and S_{DoU}, that are responsible for the upper and lower deciles of $\log_2\mathrm{FC_L}$. For these components, we assigned increased IP value (IP = 0.7) since it is expected that $\mathrm{Abs}[\mathrm{Average}(\log_2\mathrm{FC_L})_{90-100}]$ calculated for 90–100th percentiles of $\log_2\mathrm{FC_L}$ will be much greater than such value calculated for 10–90th percentiles.

S_{EStD}, S_{EoH}, S_{EoL} are responsible for evaluating expression stability within pools of normal and tumor samples. S_{EStD} scores trimmed standard deviation of CPM values (10–90th percentiles), and S_{EoH} (or S_{EoL}) is responsible for outliers with high (or low) mRNA level (in terms of CPM). Additionally, we included scoring for average expression level (S_{EA}) and set high weight ($W = 6$) for this component in order to completely exclude genes with low mRNA level from the analysis.

Finally, we added scorings for correlations between gene expression and six clinical and pathological characteristics: pathologic TNM classification (separately for T, N, and M indexes), pathologic stage, follow-up person neoplasm cancer status and follow-up treatment success status. S_{Cr} is the component responsible for Spearman's correlation coefficient, and S_{Cp}–for correlation p-value. IV values were chosen in such a way that cases with $p > 0.25$ and $|r_s| < 0.1$ have score equals to 100. In total, each of these two components is taken 18 times: 6 clinical characteristics are analyzed for associations with CPM in tumor samples, CPM in normal samples and T-N expression fold change (paired samples). Hence, we assigned low weights—$W =$ 0.2 and 0.3 for S_{Cr} and S_{Cp}, respectively.

Besides stable and high enough expression level, an appropriate RG should also demonstrate a low mutation rate, single transcript isoform, and absence of pseudogenes in order to avoid problems with PCR priming and ensure the

TABLE 1 | Components of the scoring function.

Component	Factor	Variable (x = …)*	IV	IP	CS	Sq	CA	W	Number of times applied
EXPRESSION SCORING									
S_{DP}	T-N expression level difference (pooled samples)	Abs $(\log_2 FC_P)_{10-90}$	0.05	0.25	2.5	1	0	4	1 (all samples)
S_{DL}	T-N expression level difference (paired samples)	Abs (Average$(\log_2 FC_L)_{10-90}$)							1 (paired samples)
S_{DoO}	T-N expression level difference: outliers, overexpression	Abs (Average$(\log_2 FC_L)_{90-100}$)	0.1	0.7	2.5	1	10	1	1 (paired samples)
S_{DoU}	T-N expression level difference: outliers, underexpression	Abs (Average$(\log_2 FC_L)_{0-10}$)							1 (paired samples)
S_{DLc}	Cumulative T-N expression difference among paired samples	Average (Abs$(\log_2 FC_L)_{10-90}$)	0.1	0.5	2.5	1	5	2	1 (paired samples)
S_{EStD}	Expression level stability: standard deviation	StDev $(CPM)_{10-90}$/Average $(CPM)_{10-90}$	0.1	0.3	2	1	5	1.5	2 (all samples: normal and tumor)
S_{EoH}	Expression level stability: outliers (high expression)	$\log_2$ (Average$(CPM)_{90-100}$/Average $(CPM)_{10-90}$)	0.1	0.7	2.5	1	5	0.75	2 (all samples: normal and tumor)
S_{EoL}	Expression level stability: outliers (low expression)	$\log_2$ (Average$(CPM)_{10-90}$/Average $(CPM)_{0-10}$)							2 (all samples: normal and tumor)
S_{EA}	Average expression level	$1/\log_2 (CPM)_{10-90}$	0.07	0.15	3	1	0	6	1 (all tumor samples)
S_{Cp}	Correlations of expression with clinical parameters (*p*-values)	-$\log_2$ (*p*-value)	2	4	3	0.3	5	0.3	18 (3 × 6; 3: CPM_{10-90} all tumor samples, CPM_{10-90} all normal samples, $(\log_2 FC_L)_{10-90}$; 6: pathologic TNM classification, pathologic stage, follow-up—person neoplasm cancer status, follow-up—treatment success)
S_{Cr}	Correlations of expression with clinical parameters (r_s)	Abs (r_s)	0.1	0.25	2.5	0.3	5	0.2	18 (the same as above)
"ANTI-SCORINGS"									
S^{Mut}	Percentile of mutation rate		75	95	4	1			
$S^{Isoforms}$	Number of transcript isoforms		1	3	2	0.4			
$S^{Pseudogenes}$	Number of pseudogenes		0	2	2	0.4			

**Percentiles, which were taken into calculation, are indicated as a subscript.*

IV, ideal value; IP, inflection point; CS, curve slope; Sq, "squeeze"; CA, constant add; W, weight; Abs (…), absolute value; Average (…), mean value; CPM, counts per million, gene expression level; FC_P, ratio of the average CPM in a pool of tumor samples to the average CPM in a pool of normal samples; FC_L, ratio of CPM values between tumor and matched normal tissue (per each paired sample); StDev (…), standard deviation; r_s, Spearman's correlation coefficient.

rigorous evaluation of mRNA level. The mutation rate of a gene was assessed using TCGA data on somatic mutations. The number of transcript isoforms (per gene) was obtained from the Ensembl human genome annotation (hg38, release 88). The number of pseudogenes (per gene) was derived from psiCube (Sisu et al., 2014). Therefore, we extended the scoring system with three additional components, "anti-scorings" (**Table 1** and **Figure 1**). The resulting score S^{Final} is calculated as follows:

$$S^{Final} = S^{Exp} \cdot S^{Mut} \cdot S^{Isoforms} \cdot S^{Pseudogenes}$$

Next, we tried to find RGs that are stably expressed across multiple tissues and cancer types. For this purpose, we calculated the pan-cancer score as follows:

$$S^{Final}_{Pan-cancer} = S^{Exp\&Mut}_{Pan-cancer} \cdot S^{Isoforms} \cdot S^{Pseudogenes}$$

where:

$$S^{Exp\&Mut}_{Pan-cancer} = \left(\frac{\sum_{j=1}^{M} \left(S_j^{Exp} \cdot S_j^{Mut} + CA \right)^k}{M} \right)^{1/k}$$

where $M = 12$ (a number of cancer types analyzed); $k = -0.4$ (negative k value implies that the pan-cancer score is a harmonic mean of individual scores); $CA = 12$ (a constant add).

Finally, we assessed the involvement of a gene in cancer-related processes on the basis of Gene Ontology (GO; The Gene Ontology, 2017) data and mentions in the articles indexed by PubMed (titles and abstracts).

A RG should not be involved in cellular processes that are frequently altered in cancer. A penalty system based on GO data was developed. We evaluated the involvement of

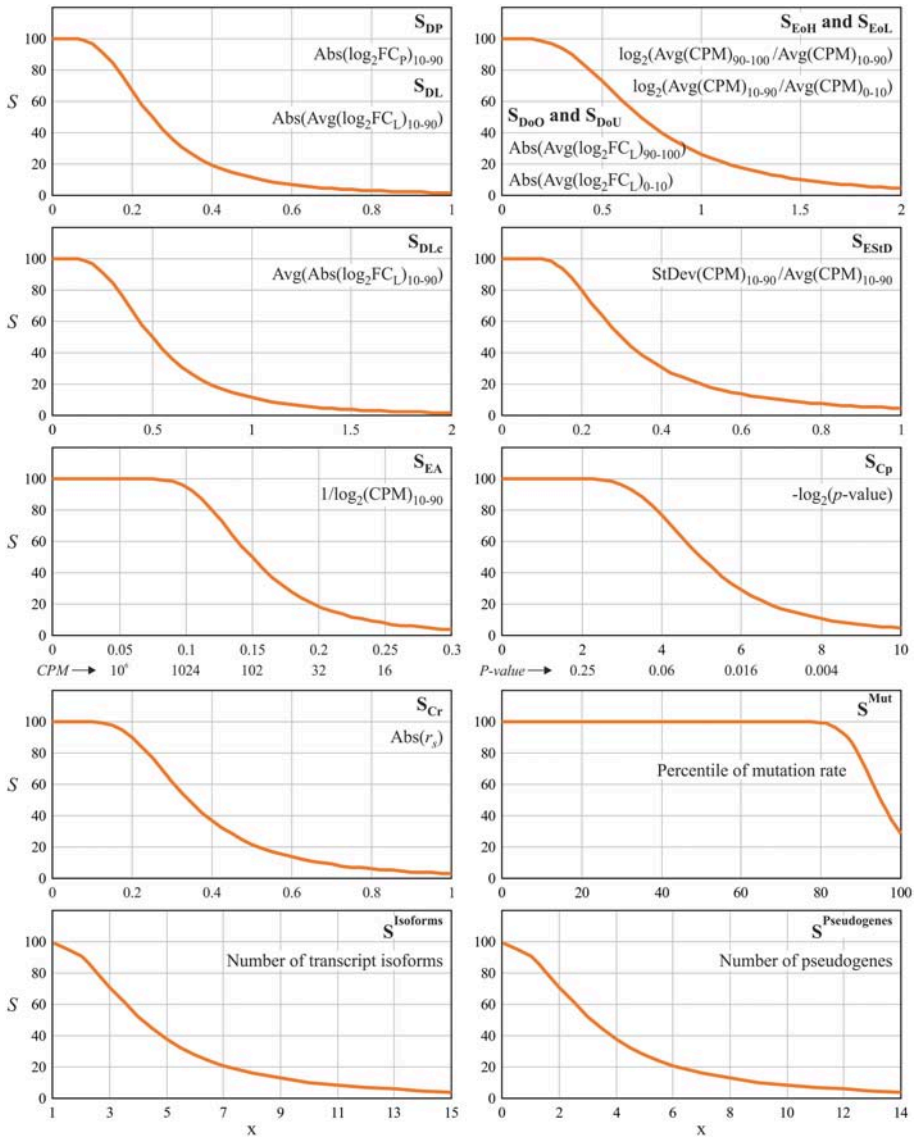


FIGURE 1 | Scoring functions used for evaluation of gene suitability for qPCR data normalization. Percentiles, which were taken into calculation, are indicated as a subscript. Abs(…), absolute value; Avg(…), mean value; CPM, counts per million, gene expression level; FC_P, ratio of the average CPM in a pool of tumor samples to the average CPM in a pool of normal samples; FC_L, ratio of CPM values between tumor and matched normal tissue (per each paired sample); StDev(…), standard deviation; r_S, Spearman's correlation coefficient.

a gene in 6 cancer-associated biological processes: cell cycle, differentiation, stress response, immune response, angiogenesis, adhesion, and cell communication. The relation of a gene to each of these processes was followed by the assignment of penalty points (from 2 to 5). Finally, these points were summed up. According to this system, a gene is penalized (1) with 5 points if its GO annotation contains at least one keyword related to cell cycle process: *cell cycle, cell division, cell growth, cell proliferation, apoptosis, apoptotic process, cell death, MAPK cascade, tumor, oncogenic, apoptotic*; (2) with 4 points if GO annotation contains a keyword related to cell differentiation: *cell differentiation, epithelial to mesenchymal transition, mesenchymal to epithelial transition, stem cell, fetal, embryonic, embryonal, embryo, gastrulation, tissue development, cellular developmental process, organ development*; (3) with 3 points for stress response related processes: *response to stress, DNA damage, DNA repair*; (4) with 2 points for inflammation and immune response: *inflammation, inflammatory, immune response, T cell activation, macrophage activation, antigen*; (5) with 2 points for angiogenesis: *angiogenesis*; (6) with 2 points for intercellular interactions: *cell communication, cell-cell signaling, cell adhesion, cell motility, cell migration*. Thus, a gene may have a maximum of $5 + 4 + 3 + 2 + 2 + 2 = 18$ penalty points.

The more accurately the gene is annotated, the more likely it is to find one of the keywords in its annotation. Therefore, GO

penalty should be normalized taking into account the number of assigned GO terms for the gene. On the other hand, the better the gene is annotated, the more extensively it is studied, and such genes represent more attractive candidates. In order to keep a balance between these two factors, we introduced normalization coefficient evaluated as the total number of GO terms (assigned for the gene) to the power of 0.3. If a gene lacked sufficient GO annotation (<3 GO terms), we assigned it 10 penalty points.

The number of PubMed-indexed articles with the mention of a gene name or its aliases was evaluated to assesses how well a gene is studied. Next, within this pool of gene-related publications, the number of cancer-related articles was also evaluated. One of the following words should be present in an article title to be treated as cancer-related: *cancer, tumor, *carcinoma, sarcoma, glioma, glioblastoma*, and other keywords.

The described components (GO and Pubmed) were not included in the main scoring and were only used for manual exclusion of cancer-associated genes. Besides, functional annotations from RefSeqGene (https://www.ncbi.nlm.nih.gov/refseq/rsg/) were added to each gene.

When revealing optimal RG pairs for each of examined cancer types, we paid special attention to the co-expression of RG candidates to avoid genes with a pronounced correlation between their mRNA levels. To implement the scoring system, we modified our previously developed CrossHub tool (the updated version can be downloaded at https://sourceforge.net/projects/crosshub/).

RESULTS

We performed the analysis of 12 cancer types from the TCGA project that have RNA-Seq data for representative sample sets: 285-1095 tumor and 19-113 matched normal tissues. These are: breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), kidney renal cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), prostate adenocarcinoma (PRAD), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), liver hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), and bladder urothelial carcinoma (BLCA). For the remaining TCGA cancer types, RNA-Seq data were available only for a few normal tissue samples, and this makes it impossible to use such datasets for the discovery of reliable RGs.

First, we assessed the expression stability of a set of 32 frequently used RGs in 12 selected cancer types: *ACTB*, *ALAS1*, *B2M*, *CDKN1A*, *G6PD*, *GAPDH*, *GUSB*, *HBB*, *HMBS*, *HPRT1*, *HSP90AB1*, *IPO8*, *LDHA*, *NONO*, *PGK1*, *POP4*, *PPIA*, *PPIH*, *PSMC4*, *PUM1*, *RPL13A*, *RPL30*, *RPLP0*, *RPS17*, *RPS18*, *SDHA*, *TBP*, *TFRC*, *UBC*, *YWHAZ*, *TUBB*, *RPN1*. This set of 32 RGs was composed of commercially available RG panels: Roche "Human Reference Gene Panel, 384" (Switzerland), TATAA "Reference Gene Panel Human" (Sweden), and Bio-Rad "Reference Genes H384" (USA). In total, 31 unique genes are included in the panels, plus we added the *RPN1* gene, which was identified by us earlier as a reliable RG for lung, kidney, and colorectal cancers (Krasnov et al., 2011; Fedorova et al., 2015). Expression stability scores were calculated for each gene in each examined cancer type. The results for the top 5 genes are presented in **Table 2** and full data—in **Supplementary Table 1**. In almost each cancer type, there were 1–10 genes with expression score about 70 or more (with a theoretical maximum of 100), which can be considered as moderately high score value. PRAD and THCA demonstrated the highest number of genes with stable mRNA level-10 and 7, respectively. Only in BCLA, all the genes had scores below 70, possibly because of potential bias due to a small number of matched normal tissues (19—the smallest number among the cancer types examined). The cross-tissue analysis of 12 cancer types revealed that the most stably expressed genes were: *PUM1* ($S^{Exp} = 70$), *IPO8* ($S^{Exp} = 61$), *UBC* ($S^{Exp} = 60$), *ACTB* ($S^{Exp} = 55$), and *RPN1* ($S^{Exp} = 54$). *GAPDH*, one of the most frequently used RGs, showed one of the least stability of mRNA level—position 25 out of 32 ($S^{Exp} = 32$). According to the obtained results, *GAPDH* can be reasonably applied as a RG only in prostate and stomach adenocarcinomas. *RPN1* gene suggested by us demonstrated high expression stability score in lung, renal, colon, liver, thyroid, and prostate cancers.

Next, for each of 12 cancer types, we searched for a pair of the most suitable RGs focusing on S^{Exp} values and correlation between mRNA levels of genes in a pair. As a result, we revealed 12 optimal pairs of RGs with S^{Exp} above 65 for each gene and absence of co-expression (**Table 2** and **Supplementary Table 1**). *PUM1* came into the pair of RGs for 9 out of 12 cancer types.

It should be noted that genes with high S^{Exp} values may be inconvenient in practice because of the presence of numerous pseudogenes, alternatively spliced transcripts or a high mutation rate. Among the traditionally used RGs with high expression scores, only 3 genes met the requirements—*PUM1*, *IPO8*, and *RPN1*. These genes have no pseudogenes, one (*RPN1*), or two (*PUM1* and *IPO8*) transcript isoforms, and relatively low mutation rate in examined cancer types.

Using the expanded scoring system (**Figure 2**), in which 3 "anti-scorings" counting mutation rate, number of transcript isoforms and pseudogenes were included, we analyzed a complete list of human genes in order to reveal the most prominent pan-cancer RG candidates (**Supplementary Table 2**). Top 10 pan-cancer RG candidates included *MBTPS1*, *HNRNPA0*, *SF3A1*, *SF3B2*, *GGNBP2*, *HNRNPUL2*, *SFRS3*, *RTF1*, *CIAO1*, *TM9SF3*. All these genes had stable and high enough mRNA level and low mutation rate in most of 12 cancer types, only one annotated transcript isoform and no pseudogenes. Taking into account PubMed article search, GO annotations, and RefSeqGene information, we selected three most promising RG candidates—*SF3A1*, *CIAO1*, and *SFRS4*.

DISCUSSION

The use of inappropriate RGs leads to unreliable data and nullifies potentially high accuracy of a qPCR technique in the evaluation of differential gene expression. The search for a RG with a stable mRNA level under experimental conditions represents a separate object of research and is rarely performed during the original

TABLE 2 | Top 5 traditionally used reference genes with the highest expression scores in 12 cancer types.

Cancer type	1		2		3		4		5	
	Gene	S^{Exp}	Gene	S^{Exp}	Gene	S^{Exp}	Gene	S^{Exp}	Gene	S^{Exp}
BRCA	***UBC***	82.1	***PUM1***	75.7	*IPO8*	71.8	*RPLP0*	69.8	*RPS18*	66.2
LUAD	***UBC***	79.8	*ACTB*	76.4	***PUM1***	69.6	*RPN1*	67.9	*RPL13A*	65.5
LUSC	***UBC***	81.4	*IPO8*	72.9	*ACTB*	71.4	***PUM1***	70.7	*RPL13A*	66.3
KIRC	***NONO***	82.6	*HSP90AB1*	73.2	***RPN1***	69.7	*YWHAZ*	68.7	*PSMC4*	64.7
KIRP	***PUM1***	70.3	***PSMC4***	66.0	*PGK1*	63.2	*ALAS1*	61.7	*IPO8*	61.1
PRAD	***SDHA***	80.8	*YWHAZ*	78.4	*PSMC4*	76.2	***PUM1***	76.1	*UBC*	75.8
COAD	***PUM1***	76.9	*GUSB*	73.4	***UBC***	72.8	*ACTB*	72.0	*IPO8*	71.6
HNSC	***RPL30***	73.4	***PUM1***	72.7	*IPO8*	68.1	*ACTB*	64.2	*PSMC4*	63.1
LIHC	***RPN1***	82.3	***ACTB***	80.9	*UBC*	78.4	*PUM1*	65.7	*RPS17*	56.4
STAD	***IPO8***	71.7	***RPL30***	71.0	*GAPDH*	69.7	*RPLP0*	68.7	*PUM1*	68.1
THCA	***RPN1***	84.4	*HSP90AB1*	84.3	***PUM1***	80.0	*TUBB*	79.2	*YWHAZ*	76.0
BLCA	***SDHA***	66.3	***PUM1***	65.9	*HSP90AB1*	63.3	*RPL30*	62.2	*RPS17*	61.2
Cross-tissue	*PUM1*	70.1	*IPO8*	60.8	*UBC*	59.8	*ACTB*	54.7	*RPN1*	54.3

Optimal pairs of reference genes for each cancer type are shown in bold.

studies. RNA-Seq data of TCGA project offer a great opportunity for evaluating gene expression stability. Using our CrossHub tool, we developed a complex scoring system that allowed us to assess the suitability of 32 traditionally used RGs for qPCR data normalization in 12 cancer types characterized by high morbidity and mortality rates. The alterations of mRNA level were shown for a number of these genes, including the most frequently used *GAPDH*, in examined cancer types. The analysis across 12 cancer types revealed that *PUM1* and *IPO8* genes demonstrate the most stable expression among the 32 genes.

PUM1 (Pumilio RNA Binding Family Member 1) serves as a translational regulator of specific mRNAs by binding to their 3'-UTRs. It may be involved in translational regulation of embryogenesis, cell development, and differentiation. There are several functions that call into question its applicability as a RG. After growth factor stimulation, PUM1 binds to 3'-UTR of *CDKN1B/p27* tumor suppressor, inhibits its expression and promotes a rapid entry to the cell cycle (Kedde et al., 2010). PUM1 is capable of repressing many mitotic, DNA repair, and DNA replication factors (Lee et al., 2016). Moreover, some authors reported that *PUM1* promotes ovarian cancer proliferation, migration, and invasion (Guan et al., 2018). However, *PUM1* is identified as one of the most stably expressed genes in uterine cervical cancer (Tan et al., 2017), endometrial carcinoma (Ayakannu et al., 2015), gallbladder (Yu et al., 2015), leiomyoma (Almeida et al., 2014), breast (Ibusuki et al., 2013; Kilic et al., 2014), and non-small cell lung (Soes et al., 2013) cancers. This gene has only 2 transcript isoforms and no pseudogenes that makes it even more attractive for use as a reference one.

Recently, Tilli et al. performed a screening of breast cancer RNA-Seq datasets from the International Cancer Genome Consortium (ICGC), GEO, and TCGA repositories. Authors found that *PUM1*, along with "novel" RGs - *CCSER2*, *SYMPK*, and *ANKRD17*, had the most stable mRNA level (Tilli et al., 2016). This agrees with previous qPCR analyses of RG expression stability in breast carcinomas (Ibusuki et al., 2013; Kilic et al., 2014).

IPO8 (*importin 8*), which has 2 transcript isoforms and no pseudogenes, is the second in the cross-tissue stability list, but its mRNA level is much less stable than that of *PUM1* according to TCGA data. IPO8 mediates nuclear import of proteins with a classical nuclear localization signal. Previously, *IPO8* was found to be suitable for data normalization in endometrial (Ayakannu et al., 2015) and ovarian carcinomas (Kolkova et al., 2013), colon adenocarcinoma cell lines (Krzystek-Korpacka et al., 2016), non-small cell lung cancer (Soes et al., 2013), and other tissues and diseases: brain edema (Du et al., 2017), heart cavities (Molina et al., 2018), T cells, and neutrophils (Ledderose et al., 2011).

The *RPN1* gene (0 pseudogenes, 1 transcript isoform), which was previously suggested by us for normalization of qPCR data in LUAD, LUSC, KIRC, KIRP, and COAD (Krasnov et al., 2011; Fedorova et al., 2015), demonstrate stable expression in these cancer types as well as in PRAD, LIHC, and THCA.

The majority of the remaining genes from the set of 32 genes, even if they demonstrate stable mRNA level in certain cancer types, have many pseudogenes or high mutation rate (for example, *UBC* is above the 99th percentile in BRCA). The presence of pseudogenes is a weakness of such widely used RGs as *GAPDH* and *ACTB* (67 and 64, respectively) (Sun et al., 2012), or genes encoding ribosomal proteins, including *RPL13A* and *RPS17* (Tonner et al., 2012).

Next, we tried to find out novel reliable and convenient RGs suitable for most cancer types. As it was described above, for this purpose, we evaluated expression and mutation scorings for each examined cancer type, calculated pan-cancer scoring values given the "anti-scorings" for the number of transcript isoforms and pseudogenes, and selected the promising candidates taking into account information on functions of the genes and their involvement in carcinogenesis.

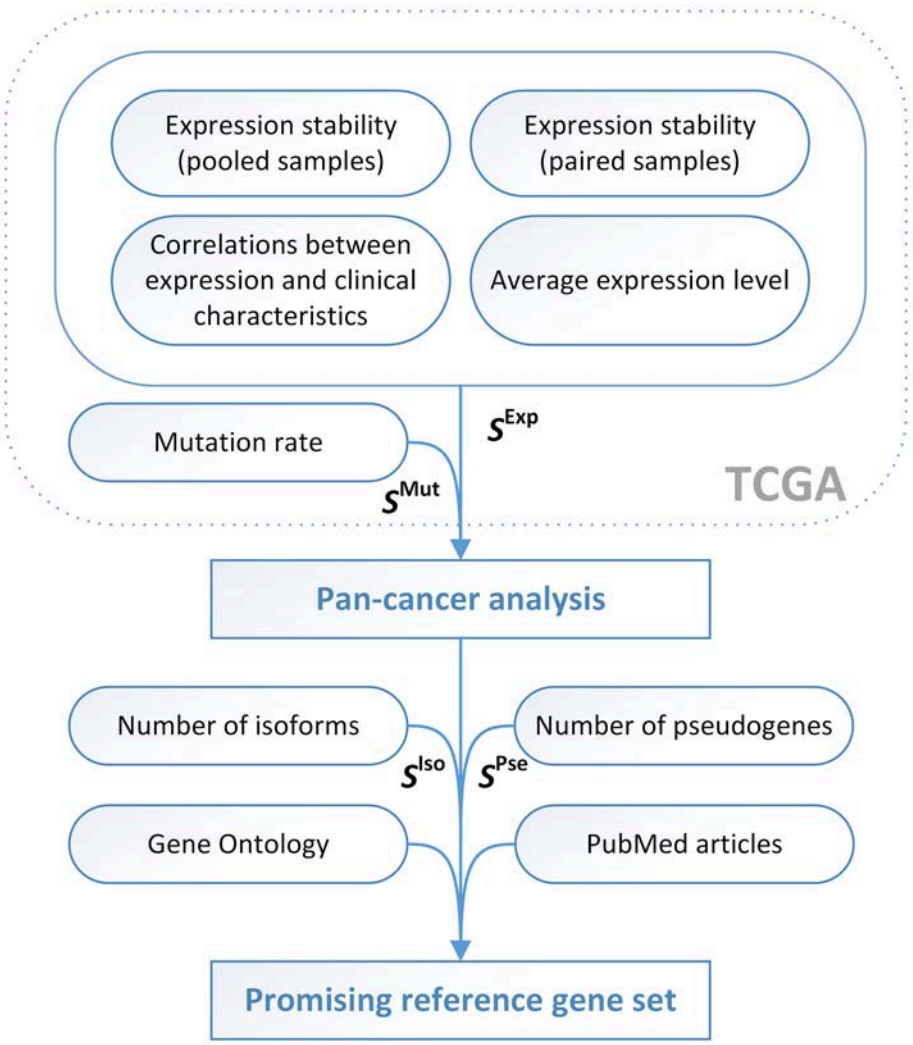


FIGURE 2 | The pipeline for identification of promising reference genes for qPCR data normalization in cancer studies. $S^{Iso} - S^{Isoforms}$, $S^{Pse} - S^{Pseudogenes}$.

Along with *SFRS4* (number 13 in the top list of "universal" reference genes), three genes that participate in pre-mRNA splicing and processing pathways (*SF3A1, SF3B2*, and *SFRS3*) are present in the top 10 of promising pan-cancer RGs. The splicing machinery (namely spliceosome) is the largest molecular machine so far described. It is composed of five small nuclear ribonucleoproteins (snRNPs U1, U2, U4, U5, and U6) and more than 100 different polypeptides (Ghigna et al., 2008). Aberrant splicing in cancer provides a way to generate alternatively spliced transcripts encoding proteins with distinct functions (Ghigna et al., 2008). There are at least two ways resulting in splicing aberrations in cancer: mutations in the affected genes, e.g., in their splice sites (*cis*-effect), and altered expression and/or activity of the elements of splicing machinery (*trans*-effect). Some of the splicing factors are known to be deregulated in cancer, by means of mRNA level alterations, mutations or posttranslational modifications (Stickeler et al., 1999; Blaustein et al., 2005; Ghigna et al., 2008). On the other hand, some of the splicing factors are considered as potential RGs. This may be explained by the complexity of the splicing machinery and various roles of its elements (David and Manley, 2010).

SF3A1 and *SF3B2* encode the subunits of splicing factors 3a and 3b. These two splicing factors together with 12S RNA unit form the U2 small nuclear ribonucleoproteins complex, which binds pre-mRNA upstream of the intron's branch site and may anchor the U2 snRNP to the pre-mRNA (Will et al., 2002). *SF3A1* is considered as a RG in sarcoma (Aggerholm-Pedersen et al., 2014), its expression was found to be stable in breast cancer (Maltseva et al., 2013), colorectal adenocarcinoma Caco-2 cells under exposure to food products (Vreeburg et al., 2011), white blood cells under treatment with growth hormone (Castigliego et al., 2010), bovine blastocysts produced by different methods (Luchsinger et al., 2014), bovine granulosa cells of dominant follicles during follicular growth and aging (Khan et al., 2016).

Considering the other splicing machinery gene, *SFRS4* (*serine and arginine rich splicing factor 4*), some authors earlier demonstrated that its mRNA level is stable in hepatocellular carcinoma (HCC) cell lines (Liu et al., 2017) and patients with alcoholic liver disease (Boujedidi et al., 2012). *SFRS4* remains stably expressed in hepatitis C virus-induced HCC, whereas *ACTB* and *GAPDH* are significantly deregulated (Waxman and Wurmbach, 2007).

CIAO1 (number 9 in the top list) is a key component of the cytosolic iron-sulfur protein assembly (CIA) complex. This is a multiprotein complex that mediates the incorporation of iron-sulfur cluster into extramitochondrial Fe/S proteins (provided by GeneCards; Stelzer et al., 2016). *CIAO1* was not previously described as a RG. Till now, there is only one article describing the possible role of the encoded protein in cancer development, namely interacting with the tumor suppressor protein WD40 (Johnstone et al., 1998). Besides this, there is almost no data on the association of this gene with cancer.

CONCLUSIONS

To reveal reliable RGs for qPCR data normalization, a comprehensive analysis of TCGA data was performed. We took into account expression stability, average mRNA level, expression correlation with clinical and pathological characteristics, number of pseudogenes and transcript isoforms, mutation rate, GO terms, and mentions of a gene in titles/abstracts of articles from PubMed. The most reliable pairs of traditionally used RGs were suggested for each of 12 examined cancer types, as well as unsuitability of some frequently used RGs was shown. Pan-cancer analysis revealed promising RG candidates with stable and sufficiently high expression level and low mutation rate across 12 cancer types. Besides, these genes have only one known transcript isoform and no pseudogenes.

DATA AVAILABILITY

All datasets generated for this study are included in the manuscript and/or the supplementary files.

AUTHOR CONTRIBUTIONS

GK, AK, NM, and AD conceived and designed the work. GK, AK, AS, VL, AB, NM, and AD performed data analysis. GK and AD wrote the manuscript. All authors agreed with the final version of the manuscript and all aspects of the work.

FUNDING

This work was financially supported by the Russian Science Foundation, grant 17-74-20064.

ACKNOWLEDGMENTS

This work was performed using the equipment of Genome center of Engelhardt Institute of Molecular Biology (http://www.eimb.ru/rus/ckp/ccu_genome_c.php).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00097/full#supplementary-material

REFERENCES

Aggerholm-Pedersen, N., Safwat, A., Baerentzen, S., Nordsmark, M., Nielsen, O. S., Alsner, J., et al. (2014). The importance of reference gene analysis of formalin-fixed, paraffin-embedded samples from sarcoma patients-an often underestimated problem. *Transl. Oncol.* 7, 687–693. doi: 10.1016/j.tranon.2014.09.012

Alexander, H., Jenkins, B. D., Rynearson, T. A., Saito, M. A., Mercier, M. L., and Dyhrman, S. T. (2012). Identifying reference genes with stable expression from high throughput sequence data. *Front. Microbiol.* 3:385. doi: 10.3389/fmicb.2012.00385

Almeida, T. A., Quispe-Ricalde, A., Montes de Oca, F., Foronda, P., and Hernandez, M. M. (2014). A high-throughput open-array qPCR gene panel to identify housekeeping genes suitable for myometrium and leiomyoma expression analysis. *Gynecol. Oncol.* 134, 138–143. doi: 10.1016/j.ygyno.2014.04.012

Andersen, C. L., Jensen, J. L., and Orntoft, T. F. (2004). Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res.* 64, 5245–5250. doi: 10.1158/0008-5472.CAN-04-0496

Ayakannu, T., Taylor, A. H., Willets, J. M., Brown, L., Lambert, D. G., McDonald, J., et al. (2015). Validation of endogenous control reference genes for normalizing gene expression studies in endometrial carcinoma. *Mol. Hum. Reprod.* 21, 723–735. doi: 10.1093/molehr/gav033

Barber, R. D., Harmer, D. W., Coleman, R. A., and Clark, B. J. (2005). GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiol. Genomics* 21, 389–395. doi: 10.1152/physiolgenomics.00025.2005

Blaustein, M., Pelisch, F., Tanos, T., Munoz, M. J., Wengier, D., Quadrana, L., et al. (2005). Concerted regulation of nuclear and cytoplasmic activities of SR proteins by AKT. *Nat. Struct. Mol. Biol.* 12, 1037–1044. doi: 10.1038/nsmb1020

Boujedidi, H., Bouchet-Delbos, L., Cassard-Doulcier, A. M., Njike-Nakseu, M., Maitre, S., Prevot, S., et al. (2012). Housekeeping gene variability in the liver of alcoholic patients. *Alcohol. Clin. Exp. Res.* 36, 258–266. doi: 10.1111/j.1530-0277.2011.01627.x

Carmona, R., Arroyo, M., Jimenez-Quesada, M. J., Seoane, P., Zafra, A., Larrosa, R., et al. (2017). Automated identification of reference genes based on RNA-seq data. *Biomed. Eng. Online* 16(Suppl. 1): 65. doi: 10.1186/s12938-017-0356-5

Castigliego, L., Armani, A., Li, X., Grifoni, G., Gianfaldoni, D., and Guidi, A. (2010). Selecting reference genes in the white blood cells of buffalos treated with recombinant growth hormone. *Anal. Biochem.* 403, 120–122. doi: 10.1016/j.ab.2010.04.001

Chapman, J. R., and Waldenstrom, J. (2015). With reference to reference genes: a systematic review of endogenous controls in gene expression studies. *PLoS ONE* 10:e0141853. doi: 10.1371/journal.pone.0141853

Chen, L., Jin, Y., Wang, L., Sun, F., Yang, X., Shi, M., et al. (2017). Identification of reference genes and miRNAs for qRT-PCR in human esophageal squamous cell carcinoma. *Med. Oncol.* 34:2. doi: 10.1007/s12032-016-0860-7

Chim, S. S. C., Wong, K. K. W., Chung, C. Y. L., Lam, S. K. W., Kwok, J. S. L., Lai, C. Y., et al. (2017). Systematic selection of reference genes for the normalization of circulating RNA transcripts in pregnant women based on RNA-Seq data. *Int. J. Mol. Sci.* 18:E1709. doi: 10.3390/ijms18081709

David, C. J., and Manley, J. L. (2010). Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged. *Genes Dev.* 24, 2343–2364. doi: 10.1101/gad.1973010

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinformatics* 14, 671–683. doi: 10.1093/bib/bbs046

Du, Y., Xu, J. T., Jin, H. N., Zhao, R., Zhao, D., Du, S. H., et al. (2017). Increased cerebral expressions of MMPs, CLDN5, OCLN, ZO1 and AQPs are associated with brain edema following fatal heat stroke. *Sci. Rep.* 7:1691. doi: 10.1038/s41598-017-01923-w

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinformatics* 19, 776–792. doi: 10.1093/bib/bbx008

Fedorova, M. S., Kudryavtseva, A. V., Lakunina, V. A., Snezhkina, A. V., Volchenko, N. N., Slavnova, E. N., et al. (2015). Downregulation of OGDHL expression is associated with promoter hypermethylation in colorectal cancer. *Mol. Biol.* 49, 608–617. doi: 10.1134/S0026893315040044

Ghigna, C., Valacca, C., and Biamonti, G. (2008). Alternative splicing and tumor progression. *Curr. Genomics* 9, 556–570. doi: 10.2174/138920208786847971

Guan, X., Chen, S., Liu, Y., Wang, L. L., Zhao, Y., and Zong, Z. H. (2018). PUM1 promotes ovarian cancer proliferation, migration and invasion. *Biochem. Biophys. Res. Commun.* 497, 313–318. doi: 10.1016/j.bbrc.2018.02.078

Gur-Dedeoglu, B., Konu, O., Bozkurt, B., Ergul, G., Seckin, S., and Yulug, I. G. (2009). Identification of endogenous reference genes for qRT-PCR analysis in normal matched breast tumor tissues. *Oncol. Res.* 17, 353–365. doi: 10.3727/096504009788428460

Hoang, V. L. T., Tom, L. N., Quek, X. C., Tan, J. M., Payne, E. J., Lin, L. L., et al. (2017). RNA-seq reveals more consistent reference genes for gene expression studies in human non-melanoma skin cancers. *PeerJ* 5:e3631. doi: 10.7717/peerj.3631

Huggett, J., Dheda, K., Bustin, S., and Zumla, A. (2005). Real-time RT-PCR normalisation; strategies and considerations. *Genes Immun.* 6, 279–284. doi: 10.1038/sj.gene.6364190

Ibusuki, M., Fu, P., Yamamoto, S., Fujiwara, S., Yamamoto, Y., Honda, Y., et al. (2013). Establishment of a standardized gene-expression analysis system using formalin-fixed, paraffin-embedded, breast cancer specimens. *Breast Cancer* 20, 159–166. doi: 10.1007/s12282-011-0318-x

Janssens, N., Janicot, M., Perera, T., and Bakker, A. (2004). Housekeeping genes as internal standards in cancer research. *Mol. Diagn.* 8, 107–113. doi: 10.1007/BF03260053

Johnstone, R. W., Wang, J., Tommerup, N., Vissing, H., Roberts, T., and Shi, Y. (1998). Ciao 1 is a novel WD40 protein that interacts with the tumor suppressor protein WT1. *J. Biol. Chem.* 273, 10880–10887. doi: 10.1074/jbc.273.18.10880

Kedde, M., van Kouwenhove, M., Zwart, W., Oude Vrielink, J. A., Elkon, R., and Agami, R. (2010). A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nat. Cell Biol.* 12, 1014–1020. doi: 10.1038/ncb2105

Khan, M. I., Dias, F. C., Dufort, I., Misra, V., Sirard, M. A., and Singh, J. (2016). Stable reference genes in granulosa cells of bovine dominant follicles during follicular growth, FSH stimulation and maternal aging. *Reprod. Fertil. Dev.* 28, 795–805. doi: 10.1071/RD14089

Kilic, Y., Celebiler, A. C., and Sakizli, M. (2014). Selecting housekeeping genes as references for the normalization of quantitative PCR data in breast cancer. *Clin. Transl. Oncol.* 16, 184–190. doi: 10.1007/s12094-013-1058-5

Kolkova, Z., Arakelyan, A., Casslen, B., Hansson, S., and Kriegova, E. (2013). Normalizing to GADPH jeopardises correct quantification of gene expression in ovarian tumours - IPO8 and RPL4 are reliable reference genes. *J. Ovarian Res.* 6:60. doi: 10.1186/1757-2215-6-60

Kozera, B., and Rapacz, M. (2013). Reference genes in real-time PCR. *J. Appl. Genet.* 54, 391–406. doi: 10.1007/s13353-013-0173-x

Krasnov, G. S., Dmitriev, A. A., Melnikova, N. V., Zaretsky, A. R., Nasedkina, T. V., Zasedatelev, A. S., et al. (2016). CrossHub: a tool for multi-way analysis of The Cancer Genome Atlas (TCGA) in the context of gene expression regulation mechanisms. *Nucleic Acids Res.* 44:e62. doi: 10.1093/nar/gkv1478

Krasnov, G. S., Oparina, N. Y., Dmitriev, A. A., Kudryavtseva, A. V., Anedchenko, E. A., Kondrat'eva, T. T., et al. (2011). RPN1, a new reference gene for quantitative data normalization in lung and kidney cancer. *Mol. Biol.* 45, 211–220. doi: 10.1134/S0026893311020129

Krzystek-Korpacka, M., Hotowy, K., Czapinska, E., Podkowik, M., Bania, J., Gamian, A., et al. (2016). Serum availability affects expression of common house-keeping genes in colon adenocarcinoma cell lines: implications for quantitative real-time PCR studies. *Cytotechnology* 68, 2503–2517. doi: 10.1007/s10616-016-9971-4

Ledderose, C., Heyn, J., Limbeck, E., and Kreth, S. (2011). Selection of reliable reference genes for quantitative real-time PCR in human T cells and neutrophils. *BMC Res. Notes* 4:427. doi: 10.1186/1756-0500-4-427

Lee, S., Kopp, F., Chang, T. C., Sataluri, A., Chen, B., Sivakumar, S., et al. (2016). Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* 164, 69–80. doi: 10.1016/j.cell.2015.12.017

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Li, P., Piao, Y., Shon, H. S., and Ryu, K. H. (2015). Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data. *BMC Bioinformatics* 16:347. doi: 10.1186/s12859-015-0778-7

Liu, Y., Qin, Z., Cai, L., Zou, L., Zhao, J., and Zhong, F. (2017). Selection of internal references for qRT-PCR assays of human hepatocellular carcinoma cell lines. *Biosci. Rep.* 37:BSR20171281. doi: 10.1042/BSR20171281

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Luchsinger, C., Arias, M. E., Vargas, T., Paredes, M., Sanchez, R., and Felmer, R. (2014). Stability of reference genes for normalization of reverse transcription quantitative real-time PCR (RT-qPCR) data in bovine blastocysts produced by IVF, ICSI and SCNT. *Zygote* 22, 505–512. doi: 10.1017/S0967199413000099

Maltseva, D. V., Khaustova, N. A., Fedotov, N. N., Matveeva, E. O., Lebedev, A. E., Shkurnikov, M. U., et al. (2013). High-throughput identification of reference genes for research and clinical RT-qPCR analysis of breast cancer samples. *J. Clin. Bioinforma.* 3:13. doi: 10.1186/2043-9113-3-13

Molina, C. E., Jacquet, E., Ponien, P., Munoz-Guijosa, C., Baczko, I., Maier, L. S., et al. (2018). Identification of optimal reference genes for transcriptomic analyses in normal and diseased human heart. *Cardiovasc. Res.* 114, 247–258. doi: 10.1093/cvr/cvx182

Patro, R., Mount, S. M., and Kingsford, C. (2014). Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.* 32, 462–464. doi: 10.1038/nbt.2862

Pfaffl, M. W., Tichopad, A., Prgomet, C., and Neuvians, T. P. (2004). Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: bestKeeper–excel-based tool using pair-wise correlations. *Biotechnol. Lett.* 26, 509–515. doi: 10.1023/B:BILE.0000019559.84305.47

Popovici, V., Goldstein, D. R., Antonov, J., Jaggi, R., Delorenzi, M., and Wirapati, P. (2009). Selecting control genes for RT-QPCR using public microarray data. *BMC Bioinformatics* 10:42. doi: 10.1186/1471-2105-10-42

Risso, D., Ngai, J., Speed, T. P., and Dudoit, S. (2014). Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* 32, 896–902. doi: 10.1038/nbt.2931

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi: 10.1093/bioinformatics/btp616

Rubie, C., Kempf, K., Hans, J., Su, T., Tilton, B., Georg, T., et al. (2005). Housekeeping gene variability in normal and cancerous colorectal, pancreatic, esophageal, gastric and hepatic tissues. *Mol. Cell. Probes* 19, 101–109. doi: 10.1016/j.mcp.2004.10.001

Sisu, C., Pei, B., Leng, J., Frankish, A., Zhang, Y., Balasubramanian, S., et al. (2014). Comparative analysis of pseudogenes across three phyla. *Proc. Natl. Acad. Sci. U.S.A.* 111, 13361–13366. doi: 10.1073/pnas.1407293111

Soes, S., Sorensen, B. S., Alsner, J., Overgaard, J., Hager, H., Hansen, L. L., et al. (2013). Identification of accurate reference genes for RT-qPCR analysis of formalin-fixed paraffin-embedded tissue from primary non-small cell lung cancers and brain and lymph node metastases. *Lung Cancer* 81, 180–186. doi: 10.1016/j.lungcan.2013.04.007

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016). The genecards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinformatics* 54:1 30 31–31 30 33. doi: 10.1002/cpbi.5

Stickeler, E., Kittrell, F., Medina, D., and Berget, S. M. (1999). Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis. *Oncogene* 18, 3574–3582. doi: 10.1038/sj.onc.1202671

Sun, Y., Li, Y., Luo, D., and Liao, D. J. (2012). Pseudogenes as weaknesses of ACTB (Actb) and GAPDH (Gapdh) used as reference genes in reverse transcription and polymerase chain reactions. *PLoS ONE* 7:e41659. doi: 10.1371/journal.pone.0041659

Tan, S. C., Ismail, M. P., Duski, D. R., Othman, N. H., Bhavaraju, V. M., and Ankathil, R. (2017). Identification of optimal reference genes for normalization of RT-qPCR data in cancerous and non-cancerous tissues of human uterine cervix. *Cancer Invest.* 35, 163–173. doi: 10.1080/07357907.2017.1278767

The Gene Ontology, C. (2017). Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res.* 45, D331–D338. doi: 10.1093/nar/gkw1108

Tilli, T. M., Castro Cda, S., Tuszynski, J. A., and Carels, N. (2016). A strategy to identify housekeeping genes suitable for analysis in breast cancer diseases. *BMC Genomics* 17:639. doi: 10.1186/s12864-016-2946-1

Tonner, P., Srinivasasainagendra, V., Zhang, S., and Zhi, D. (2012). Detecting transcription of ribosomal protein pseudogenes in diverse human tissues from RNA-seq data. *BMC Genomics* 13:412. doi: 10.1186/1471-2164-13-412

Vandesompele, J., De Preter, K., Pattyn, F., Poppe, B., Van Roy, N., De Paepe, A., et al. (2002). Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 3:RESEARCH0034. doi: 10.1186/gb-2002-3-7-research0034

Vreeburg, R. A., Bastiaan-Net, S., and Mes, J. J. (2011). Normalization genes for quantitative RT-PCR in differentiated Caco-2 cells used for food exposure studies. *Food Funct.* 2, 124–129. doi: 10.1039/C0FO00068J

Waxman, S., and Wurmbach, E. (2007). De-regulation of common housekeeping genes in hepatocellular carcinoma. *BMC Genomics* 8:243. doi: 10.1186/1471-2164-8-243

Will, C. L., Urlaub, H., Achsel, T., Gentzel, M., Wilm, M., and Luhrmann, R. (2002). Characterization of novel SF3b and 17S U2 snRNP proteins, including a human Prp5p homologue and an SF3b DEAD-box protein. *EMBO J.* 21, 4978–4988. doi: 10.1093/emboj/cdf480

Yu, S., Yang, Q., Yang, J. H., Du, Z., and Zhang, G. (2015). Identification of suitable reference genes for investigating gene expression in human gallbladder carcinoma using reverse transcription quantitative polymerase chain reaction. *Mol. Med. Rep.* 11, 2967–2974. doi: 10.3892/mmr.2014.3008

Zhao, H., Ma, T. F., Lin, J., Liu, L. L., Sun, W. J., Guo, L. X., et al. (2018). Identification of valid reference genes for mRNA and microRNA normalisation in prostate cancer cell lines. *Sci. Rep.* 8:1949. doi: 10.1038/s41598-018-19458-z

Zhou, Z., Cong, P., Tian, Y., and Zhu, Y. (2017). Using RNA-seq data to select reference genes for normalizing gene expression in apple roots. *PLoS ONE* 12:e0185288. doi: 10.1371/journal.pone.0185288

Zyprych-Walczak, J., Szabelska, A., Handschuh, L., Gorczak, K., Klamecka, K., Figlerowicz, M., et al. (2015). The impact of normalization methods on RNA-Seq data analysis. *Biomed Res. Int.* 2015:621690. doi: 10.1155/2015/621690

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

TECHNOLOGY REPORT
published: 05 March 2019
doi: 10.3389/fgene.2019.00129

PreAIP: Computational Prediction of Anti-inflammatory Peptides by Integrating Multiple Complementary Features

Mst. Shamima Khatun[1†], *Md. Mehedi Hasan*[1†] *and Hiroyuki Kurata*[1,2*]

[1] *Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Fukuoka, Japan,* [2] *Biomedical Informatics R&D Center, Kyushu Institute of Technology, Fukuoka, Japan*

Numerous inflammatory diseases and autoimmune disorders by therapeutic peptides have received substantial consideration; however, the exploration of anti-inflammatory peptides via biological experiments is often a time-consuming and expensive task. The development of novel *in silico* predictors is desired to classify potential anti-inflammatory peptides prior to *in vitro* investigation. Herein, an accurate predictor, called PreAIP (Predictor of Anti-Inflammatory Peptides) was developed by integrating multiple complementary features. We systematically investigated different types of features including primary sequence, evolutionary and structural information through a random forest classifier. The final PreAIP model achieved an AUC value of 0.833 in the training dataset via 10-fold cross-validation test, which was better than that of existing models. Moreover, we assessed the performance of the PreAIP with an AUC value of 0.840 on a test dataset to demonstrate that the proposed method outperformed the two existing methods. These results indicated that the PreAIP is an accurate predictor for identifying AIPs and contributes to the development of AIPs therapeutics and biomedical research. The curated datasets and the PreAIP are freely available at http://kurata14.bio.kyutech.ac.jp/PreAIP/.

Keywords: inflammatory disease, anti-inflammatory peptides prediction, feature encoding, feature selection, random forest

OPEN ACCESS

Edited by:
Yuriy L. Orlov, Institute of Cytology and Genetics (RAS), Russia

Reviewed by:
Deepak Singla, Punjab Agricultural University, India
Hifzur Rahman Ansari, King Abdullah International Medical Research Center KAIMRC, Saudi Arabia

***Correspondence:**
Hiroyuki Kurata
kurata@bio.kyutech.ac.jp

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *25 July 2018*
Accepted: *06 February 2019*
Published: *05 March 2019*

Citation:
Khatun MS, Hasan MM and Kurata H (2019) PreAIP: Computational Prediction of Anti-inflammatory Peptides by Integrating Multiple Complementary Features. Front. Genet. 10:129. doi: 10.3389/fgene.2019.00129

INTRODUCTION

Inflammation responses occur under the normal conditions when tissues are damaged by bacteria, toxins, trauma, heat, or any other reason (Ferrero-Miliani et al., 2007). These responses cause chronic autoimmune and inflammation disorders, including neurodegenerative disease, asthma, psoriasis, cancer, rheumatoid arthritis, diabetes, and multiple sclerosis (Zouki et al., 2000; Steinman et al., 2012; Tabas and Glass, 2013; Patterson et al., 2014; Hernández-Flórez and Valor, 2016). Numerous inflammation mechanisms are crucial for the upkeep of the state of tolerance (Miele et al., 1988; Corrigan et al., 2015). Numerous endogenous peptides recognized through inflammatory reactions function as anti-inflammatory agents can be employed by new therapies for autoimmune and inflammatory illnesses (Gonzalez-Rey et al., 2007; Delgado and Ganea, 2008). The immunotherapeutic aptitude of these anti-inflammatory peptides (AIPs) has various clinical applications such as generation of regulatory T cells and inhibition of antigen-specific T(H)1-driven responses (Delgado and Ganea, 2008). Moreover, certain synthetic AIPs act as effective

therapeutic agents for autoimmune and inflammatory disorders (Zhao et al., 2016). For instance, chronic adenoidal direction of human amyloid-β peptide causes an Alzheimer's disease. Mice models result in compact deposition of amyloid-β peptides, which is a pathological marker of Alzheimer's disease, astrocytosis, microgliosis, and neuritic dystrophy in the brain (Boismenu et al., 2002; Gonzalez et al., 2005; Kempuraj et al., 2017). The present therapy for autoimmune and inflammatory disorders involves the use of non-specific anti-inflammatory drugs and other immunosuppressants, which are frequently related to different side effects, such as initiation of a higher possibility of infectious diseases and ineffectiveness alongside inflammatory disorders (Tabas and Glass, 2013).

Notwithstanding the increasing number of experimentally examined AIPs *in vivo*, the molecular mechanism of AIP specificity remains largely unknown. On the other hand, large-scale experimental analysis of AIPs is time-consuming, laborious, and expensive. An alternative, computational approach that provides an accurate and reliable prediction of AIPs is required to complement the experimental efforts and to access the prompt identification of potential AIPs prior to their synthesis. To date, two *in silico* methods have been proposed to predict AIPs (Gupta et al., 2017; Manavalan et al., 2018). In 2017 Gupta et al. employed hybrid features with a support vector machine (SVM) classifier to develop the AntiInflam predictor (Gupta et al., 2017). Manavalan et al. developed the AlPpred predictor by using the primary sequence encoding features with a random forest (RF) classifier (Manavalan et al., 2018). These two methods used the primary sequence feature information without considering any evolutionary or structural features.

Nonetheless, the performance of the abovementioned existing predictors is not sufficient and remains to be improved. In this study, we have developed an accurate predictor named PreAIP (Predictor of Anti-Inflammatory Peptides) by integrating multiple complementary. We investigated different types sequence features including the primary sequence, evolutionary, and structural through a RF classifier. The PreAIP achieved higher performance on both the training and test datasets than the existing methods. In addition, we obtained valuable insights into the essential sequence patterns of AIPs.

MATERIALS AND METHODS

Dataset Collection

To construct the PreAIP, we collected training and test datasets from a recently published article of the AIPpred (Manavalan et al., 2018) and the IEDB database (Vita et al., 2019). A peptide was considered as anti-inflammatory (positive sample) if the anti-inflammatory cytokines of peptides induce any one of IL-10, IL-4, IL-13, IL-22, TGFb, and IFN-a/b in T-cell analyses of mouse and human (Marie et al., 1996; Jin et al., 2014). Meanwhile, the linear peptides for anti-inflammatory cytokines were considered non-AIPs (i.e., negative samples). To solve the overfitting problem of the prediction model, CD-HIT was employed with a sequence identity threshold of 0.8 (Huang et al., 2010). After eliminating redundant peptides, the same training and test samples were retrieved from the AIPpred predictor (Manavalan et al., 2018). More reliable performance would be achieved by using a more stringent criterion of 0.3 or 0.4, as executed in (Hasan et al., 2016, 2017a). However, this study did not use such a stringent criterion, because the length of the currently available AIPs is between 4 and 25. If we apply a stringent criterion of <0.8, the number of the available AIPs is greatly reduced so that we cannot retrieve the datasets employed by the previous predictor (Manavalan et al., 2018). The collected training dataset results in 1,258 positive and 1,887 negative samples, and the test dataset contains 420 positive and 629 negative samples. All of curated datasets are included in our web server.

Computational Framework

An overall computational framework of the proposed PreAIP is shown in **Figure 1**. After collecting the positive and negative AIPs from the AIPpred server (Manavalan et al., 2018), their sequence datasets were transformed into the primary sequence, evolutionary and structural features. We considered polypeptides with 1 to 25 natural amino acids. When the peptide contains less than 25 residues, our scheme provides gaps (-) to the missing residues to compensate a peptide length of 25. To encode the primary sequence features, we employed two encoding methods of the composition of *k*-spaced amino acid pairs (KSAAP) and AAindex properties. An evolutionary feature was encoded by using the position specific encoding matrix, i.e., profile-based composition *k*-space of amino acid pair (pKSAAP). The structural feature (SF) was encoded by using SPIDER2 (Yang et al., 2017) and PEP2D (http://crdd.osdd.net/raghava/pep2d/) bioinformatics tools. The resulting five types of descriptors were independently put into RF models to produce five consecutive, independent RF prediction scores. Those RF scores were linearly combined using the weight coefficients to obtain the final prediction score. A web server was developed to implement the PreAIP.

Feature Encoding

The PreAIP was constructed based on a binary classification problem (positive AIPs and negative-AIPs) through RF algorithms. The extraction of a set of relevant features is a crucial step to present a classifier. To keep the generated feature vectors, a high-quality peptide encoding method is necessary. As a substitute of the simple binary representation, we adopted five types of complicated feature encoding methods: AAindex, KSAAP, SPIDER2, PEP2D, and pKSAAP, which are briefly described in the following subsections.

Amino Acid Index Properties

Numerical physicochemical properties of amino acids exist in the AAindex database (version 9.1) (Kawashima et al., 2008). After assessing different types of AAindex indices, we selected 8 types of high indices (HI) and ordered them from HI1 to HI8 (**Table S1**). In a peptide sequence with length L, a ($L \times 20$) feature vector was generated through the AAindex encoding.

KSAAP Encoding

The KSAAP encoding descriptor is widely used in bioinformatics research (Carugo, 2013; Hasan et al., 2018a,b). The procedure of

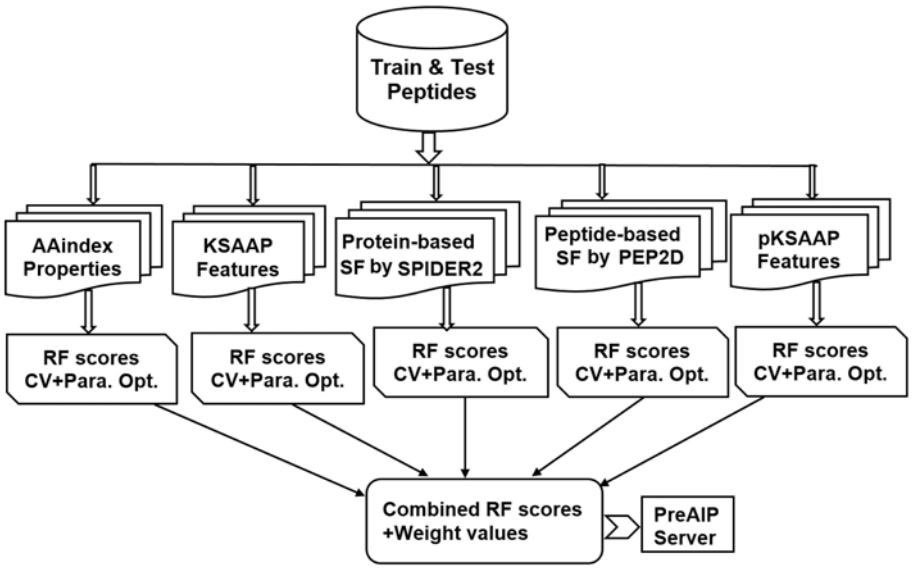


FIGURE 1 | Computational framework of PreAIP.

KSAAP is briefly described as follows. Peptide sequences contain (20×20) types of amino acid pairs (i.e., AA, AC, AD, …, YY)$_{400}$ for every single k, where k denotes the space between two amino acids. The optimal k_{max} was set to 0–4 to generate $(20 \times 20 \times 5) = 2{,}000$ dimensional feature vectors for each corresponding peptide sequence. Details of the KSAAP encoding method are described elsewhere (Hasan et al., 2015).

Structural Features

Protein-Based SF

The protein-based SF features are generated by the SPIDER2 software that is widely used in bioinformatics research (Yang et al., 2017; López et al., 2018). Three types of features were generated by SPIDER2: accessible surface area (ASA), backbone torsion angles (BTA), and secondary structure (SS). The BTA generated 4-type feature vectors of phi, psi, theta and tau. The SS generated 3-type feature vectors of helix, strand and coil. Totally, 8-type feature vectors were generated SPIDER2. For each peptide sequence, $(L \times 8)$ dimensional feature vectors were generated, where L was the length of a given AIP.

Peptide-Based SF

We employed PEP2D to generate a peptide structure prediction feature (http://crdd.osdd.net/raghava/pep2d/). The PEP2D generated three types of probability scores: Helix Prob, Sheet Prob, and Coil Prob. For each peptide sequence, $(L \times 3)$ dimensional feature vectors were generated, where L was the length of a given AIP.

pKSAAP Encoding

In protein or peptide sequence analysis, the PSSM provides useful evolutionary information. This matrix measures the replacement probability of each residue in a protein with all the residues of the genomic code. The *PSSM profile* was created by using PSI-BLAST (version of 2.2.26+) against the whole Swiss-Prot NR90 database (version of December 2010) with two default parameters, an e-value cutoff of 1.0×10^{-4} and an iteration number of 3 (Hasan et al., 2015). Then, we extracted the feature vectors using the given peptide sequences. After generating the PSSM profile, we generated possible k-space pair composition from the PSSM, i.e., pKSAAP, in the same manner as the previous study of protein pupylation site prediction (Hasan et al., 2015). When an optimal k-space was between 0 and 4, a $(5 \times 20 \times 20 = 2{,}000)$ dimensional feature vector was generated.

Moreover, we utilized a similarity-search-based tool of BLAST (version of ncbi-blast-2.2.25+) (Altschul et al., 1997; Bhasin and Raghava, 2004) to investigate whether a query peptide belongs to AIPs or not. The BLASTP with an e-value of 1.0×10^{-2} was used for the whole Swiss-Prot NR90 database (version of December 2010).

Feature Selection

To find the top ranking features for predicting AIPs, a well-established, supervised method for feature dimensionality reduction, Information Gain (IG) (Azhagusundari and Thanamani, 2013; Huang, 2015; Manavalan et al., 2018), was used through a WEKA package (Frank et al., 2004). A large value of the IG indicates that the corresponding residues have a great impact on prediction performance. The IG processes the decrease in entropy when given information is used to group values of an alternative (class) feature. The entropy of feature U is defined as

$$H(U) = -\sum_i P(u_i) log_2 (P(u_i)) \quad (1)$$

where u_i is a set of values of U and $P(u_i)$ is the prior probability of u_i. Conditional entropy $H(U/V)$, given another feature V, is defined as

$$H(U|V) = -\sum_j P(v_j) \sum_i P(u_i|v_j) \log_2(P(u_i|v_j)) \quad (2)$$

where $P(u_i \mid v_j)$ is the posterior probability of U given by the value $v_j$of V. The IG is defined as the decreased entropy calculated by subtracting the conditional entropy of U given by V from the entropy of U, as follows.

$$IG(U|V) = H(U) - H(U|V) \quad (3)$$

Random Forest

The RF is a supervised machine learning algorithm (Breiman, 2001) and is widely used for various biological problems (Manavalan et al., 2017, 2018; Bhadra et al., 2018; Hasan and Kurata, 2018). In brief, the following steps are carried to construct n trees of the RF model. Initially, to obtain a new dataset, N samples are obtained from the training set by random selection with replacement procedures. To get n different datasets this procedure is repeated n times and n decision trees are built based on the n datasets. In this assembling process, for K input features, k ($k << K$) features are selected randomly, where k is the constant during construction of the RF. To split the node, a *gini* impurity criterion is used from the given features. To grow completely, each decision tree is grown without pruning. Afterward getting n decision trees, the class with the most votes is the final prediction (Breiman, 2001). An R package was implemented to train the proposed model (https://cran.r-project.org/web/packages/randomForest/). We set n to 1000 through the 10-fold cross-validation (CV) test, which is large enough to gain stable prediction.

Other Machine Learning Algorithms

The performance of the RF was characterized in comparison to three commonly used machine learning algorithms: Naive Bayes (NB) (Lowd, 2005), SVM (Hearst, 1998), and artificial neural network (ANN) (Michalski et al., 2013). We used the NB and ANN algorithms of the WEKA software (Frank et al., 2004) and the SVM algorithm with a kernel radial basis function (RBF) of the LIBSVM package (https://www.csie.ntu.edu.tw/~cjlin/libsvm/). In the NB algorithm, we set batch size to 1,000 through the 10-fold CV via the WEKA software. For the ANN algorithm, we considered "MultilayerPerceptron –L 0.3 –M 0.2 –N 500 –V 0 S 0 –E 20 –H a" via the WEKA software. To optimize the parameters of the SVM model, the cost and gamma functions were set to 8 and 0.03125 for KSAAP, respectively, via the LIBSVM package. Similarly, the cost and gamma functions were set to 2 and 0.0123 for AAindex, 32 and 0.0625 for pKSAAP, 16 and 0.125 for SPIDER2, and 8 and 0.015625 for PEP2D.

Combined Method

To make an efficient and robust prediction model, optimization of incorporative feature methods is generally essential. We linearly combined the RF scores of the five encoding methods: AAindex, KSAAP, SPIDER2, PEP2D, and pKSAAP, using the following formula (Hasan et al., 2017b):

$$\begin{aligned}\text{Combined} = {} & w_1 \times \text{SPIDER2} + w_2 \times \text{PEP2D} + w_3 \times \text{KSAAP} \\ & + w_4 \times \text{AAindex} + w_5 \times \text{pKSAAP} \end{aligned} \quad (4)$$

where w_1, w_2, w_3, w_4, and w_5 are the weight coefficients indicating the strength of the five descriptors; the sum of w_1, w_2, w_3, w_4, and w_5 is 1. We adjusted each weight from 0 to 1 with an interval of 0.05. When w_1, w_2, w_3, w_4, and w_5 were 0.00, 0.00, 0.15, 0.25, and 0.6, respectively, the AUC value on the CV of training dataset was maximal. Therefore, the linear combination of the three successive RF models of KSAAP, AAindex, and pKSAAP was actually "Combined."

Performance Assessment

To investigate the performance of the PreAIP, the threshold-dependent and threshold-independent indices were measured. Using the threshold-dependent indices, four widely used statistical measures denoted as accuracy (Ac) specificity (Sp), sensitivity (Sn), and Matthews correlation coefficient (MCC), respectively, were considered. The four outcomes are presented in the following formulas,

$$\text{Ac} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (5)$$

$$\text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (6)$$

$$\text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (7)$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TN} + \text{FN}) \times (\text{TP} + \text{FP}) \times (\text{TN} + \text{FP}) \times (\text{TP} + \text{FN})}} \quad (8)$$

where TP exemplifies the number of correctly predicted positive samples; TN the number of correctly predicted negative samples; FP the number of incorrectly predicted positive samples, and FN the number of incorrectly predicted negative samples. Furthermore, we used the receiver operating characteristics (ROC) curve (Sn vs. 1-Sp plot) to evaluate the area under the ROC curve (AUC) of the threshold-independent parameter (Centor, 1991; Gribskov and Robinson, 1996).

Since the balance between the correctly predicted AIPs and non-AIPs is critically responsible for accurate prediction, Sp and Sn are intuitive, intelligible measures. Typically, high Sp decreases Sn. In this study, the prediction performance of the PreAIP for the training dataset was evaluated with a stepwise change in Sp. We calculated Sn, Ac, and MCC at high (0.903), moderate (0.801) and low (0.709) levels of Sp. These three levels of Sp were given by setting the high (0.468), moderate (0.388), and low (0.342) thresholds of the RF score. In the same manner, we measured the performance of the individual encoding scheme of KSAAP, AAindex, SPIDER2, PEP2D, and pKSAAP at each level of Sp. When the same threshold values of the RF score were applied to prediction of the test dataset, the high, moderate

and low levels of Sp were calculated as 0.871, 0.747, and 0.636, respectively.

To assess the performance of the PreAIP using the measures of Ac, Sp, Sn, MCC, and AUC, a 10-fold CV test was used. For the 10-fold CV, original training samples were randomly and equally picked up into 10 subclasses. Among 10 subclasses, one subclass was singled out as the test sample, and the remaining 9 subclasses were considered as the training sample. Then we computed all performance measures for each predictor. We repeated this procedure 10 times by changing the training and test samples. Eventually, we calculated the average value of each performance measure for each predictor.

RESULTS AND DISCUSSION

Sequence Preference Analysis of AIPs

To investigate the amino acid preference of positive and negative AIPs, we performed sequence compositional preference analysis using the amino acids from the 1 to 15 N-terminal residues of training sets. The length of the AIPs ranged between 4 and 25 amino acid residues in this study. The average length of AIPs was 15 amino acids. Since Ialenti et al. suggested that the AIP activity is located in the N-terminal region of the molecule (Ialenti et al., 2001), we investigated the 1 to 15 N-terminal amino acids by the sequence compositional preference analysis. A non-existing residue was coded by "O" to fill the corresponding position of the AIPs.

At first, we submitted the 1 to 15 N-terminal amino acids of positive and negative AIPs to the sample logo online server (http://www.twosamplelogo.org/) to generate the sequence logo representations (**Figure 2**). The height for each amino acid was in proportion to the percentage of positive (over-represented) or negative (under-represented) peptides. The logos were scaled according to their statistical significance threshold of $p < 0.05$ by Welch's t-test. Leucine (L) at positions 5, 7, 10, 11, and 15, cysteine (C) at position 7 and 10, isoleucine (I) at positions 2 and 7, arginine (R) at position 5, phenylalanine (F) at position 8, and lysine (K) at position 15 were significantly overrepresented compared with other amino acids, while aspartic acid (D) at positions 4, 5, 10, 13, and 15, threonine (T) at positions 3 and 7, valine (V) at position15 were significantly underrepresented. In addition, tyrosine (Y) at positions 4 and 5 was overrepresented, while Y at positions 5 and 10 underrepresented. These results suggested that positive and negative AIPs are significantly different.

Secondly, we examined the evolutionary conservation features of the PreAIP using the average PSSM value (APV) for each amino acid within 1 to15 N-terminal amino acids of AIPs. The evolutionary conservation information of APV of both the positive and negative AIPs is illustrated in **Figure 3**. Some of amino acid positions of positive and negative AIPs showed significantly different scores. Furthermore, a nonparametric Kruskal–Wallis (KW) test was used to examine whether positive and negative AIPs were significantly dissimilar. The p-values were calculated and corrected by the Bonferroni test (**Table S2**).

Thirdly, we examined the AAindex encoding features of PreAIP. Eight types of informative amino acid indices were

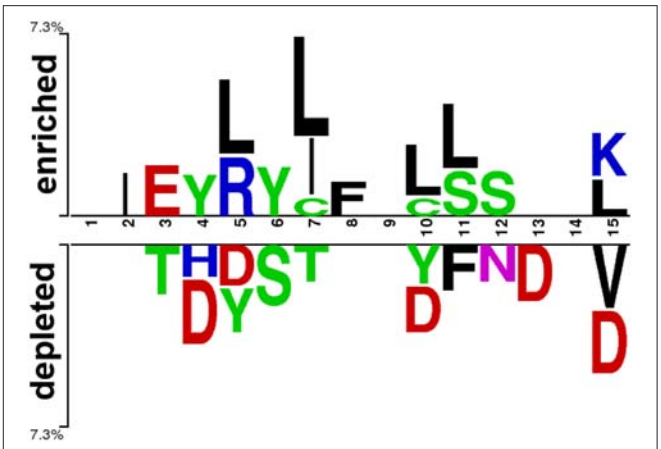


FIGURE 2 | Sequence logo representation of positive and negative AIPs. The upper portion (enriched) is represented by positive AIPs, while lower portion (depleted) negative AIPs. The statistically significant local sequence within the N-terminal 15-residues of AIPs was plotted with $p < 0.05$ by Welch's t-test.

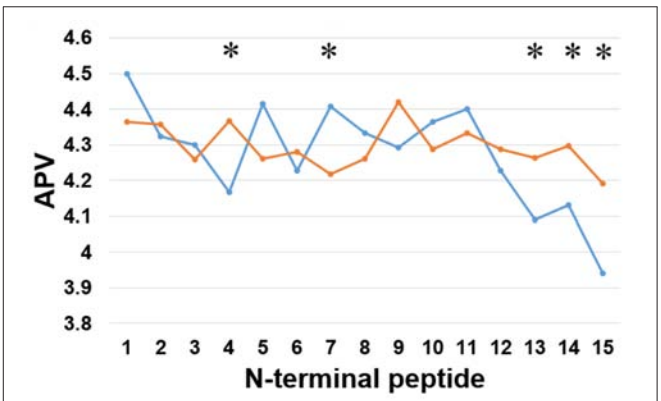


FIGURE 3 | Comparison of evolutionary information of positive and negative AIPs. Blue lines represent the positive AIP, while orange lines the negative AIPs. "*" represents that the APV is statistically different between both the AIPs, with $p < 0.05$ by the KW test.

used and named HI1 to HI8 as the input feature vectors from the AAindex database. We examined these HI amino acid properties of both the positive and negative AIPs. As illustrated in **Figure 4**, the average values of the eight indices were renamed as AVHI1 to AVHI8. These indices represented the amino acid compositions of intracellular proteins. Some of the AIPs had distinct amino acid compositions in the eight high-quality amino acid indices between two samples of AIPs (**Figure 4**). The KW test was used to examine whether two samples of AIPs were significantly dissimilar with respect to the eight HI properties. The p-values were calculated and corrected by the Bonferroni test (**Table S3**). Significantly different AAindex values with p-value <0.05 appeared at some positions of AIPs, as marked with "*" in **Figure 4**.

Finally, we examined the difference in 8 types of SFs by SPIDER2 between the positive and negative AIPs, as shown in

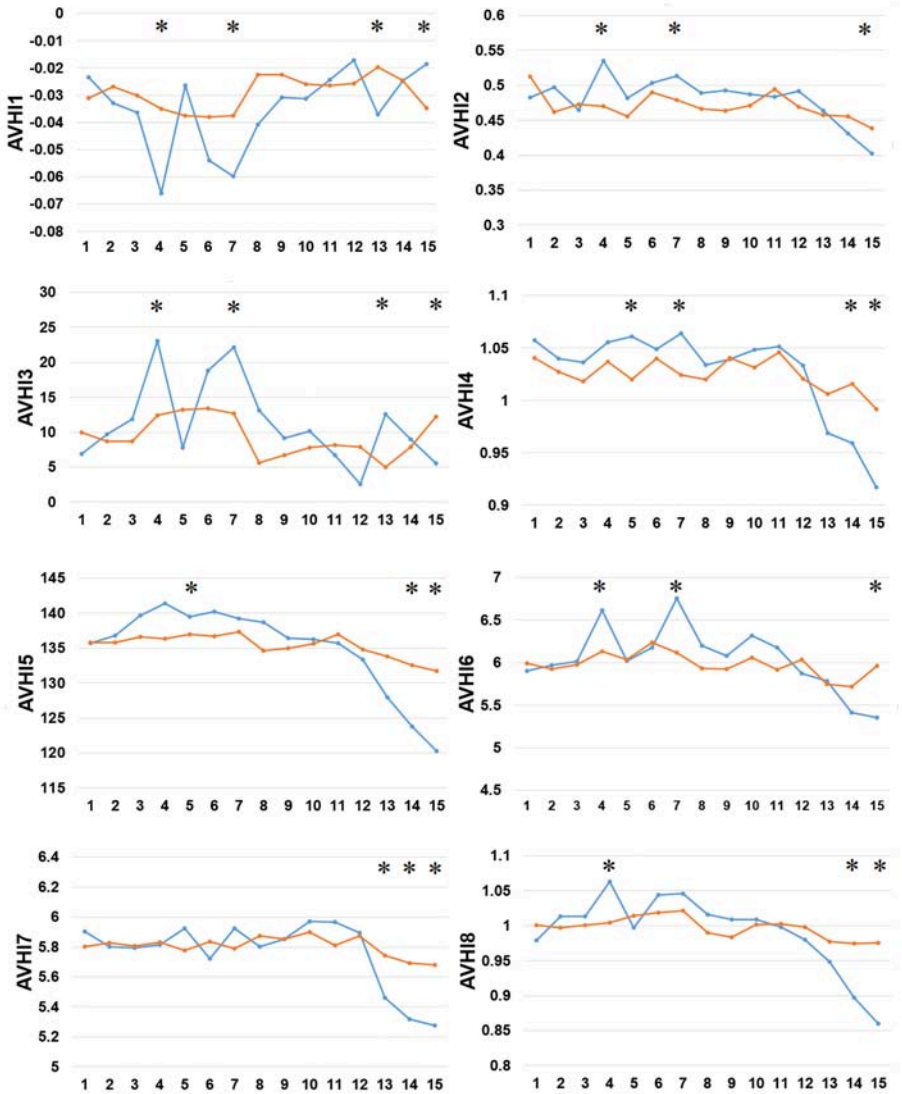


FIGURE 4 | Comparison of eight high-quality amino acid indices between two samples of AIPs. The eight high-quality amino acid indices from HI1 to HI8 are placed at the centers of eight amino acid index clusters, which indicate high residue propensities of AAindex. The row represents the N-terminal peptide, while the blue lines signify the positive AIP and the orange lines the negative AIPs. "*" represents that the amino acid indices are statistically different between both the samples with $p < 0.05$ by the KW test.

Figure 5. We calculated the average value of 8 types of SFs for SPIDER2: ASA, phi, psi, theta, tau, coil, stand, and helix of both the positive and negative AIPs. The average features were represented as AVAS, AVPhi, AVPsi, AVThe, AVTau, AVCoil, AVSta, and AVHel (**Figure 5**). We plotted these average values of SFs with respect to the 1–15 N-terminal AIPs. Distinguished differences were observed between the positive and negative samples of AIPs. The KW test was employed to examine whether two sample of AIPs were significantly dissimilar among the eight SFs. The *p*-values were calculated and corrected by the Bonferroni test (**Table S4**). Significantly different SFs were perceived at some positions of AIPs, with a *p*-value <0.05, as indicated with "*" in **Figure 5**.

The above analysis of residue preference between the positive and negative AIPs suggested that the combination of the primary sequence, evolutionary, and structural amino acid occurrences achieves a precise prediction.

Overall Prediction Performance of PreAIP

The selected five descriptors (AAindex, KSAAP, SPIDER2, PEP2D, and pKSAAP) were separately used for prediction of AIPs. Optimization of multiple encoded features is generally essential in the training model to reduce dimensionality while retaining the significant feature. To achieve this, we performed multiple rounds of experiments to select appropriate feature vectors using the IG feature selection via 10-fold CV test on training set; however, it turned out that the IG feature selection did not improve prediction performance. Thus, the IG feature was used to collect significant features and for interpreting a superiority of KSAAP encoding.

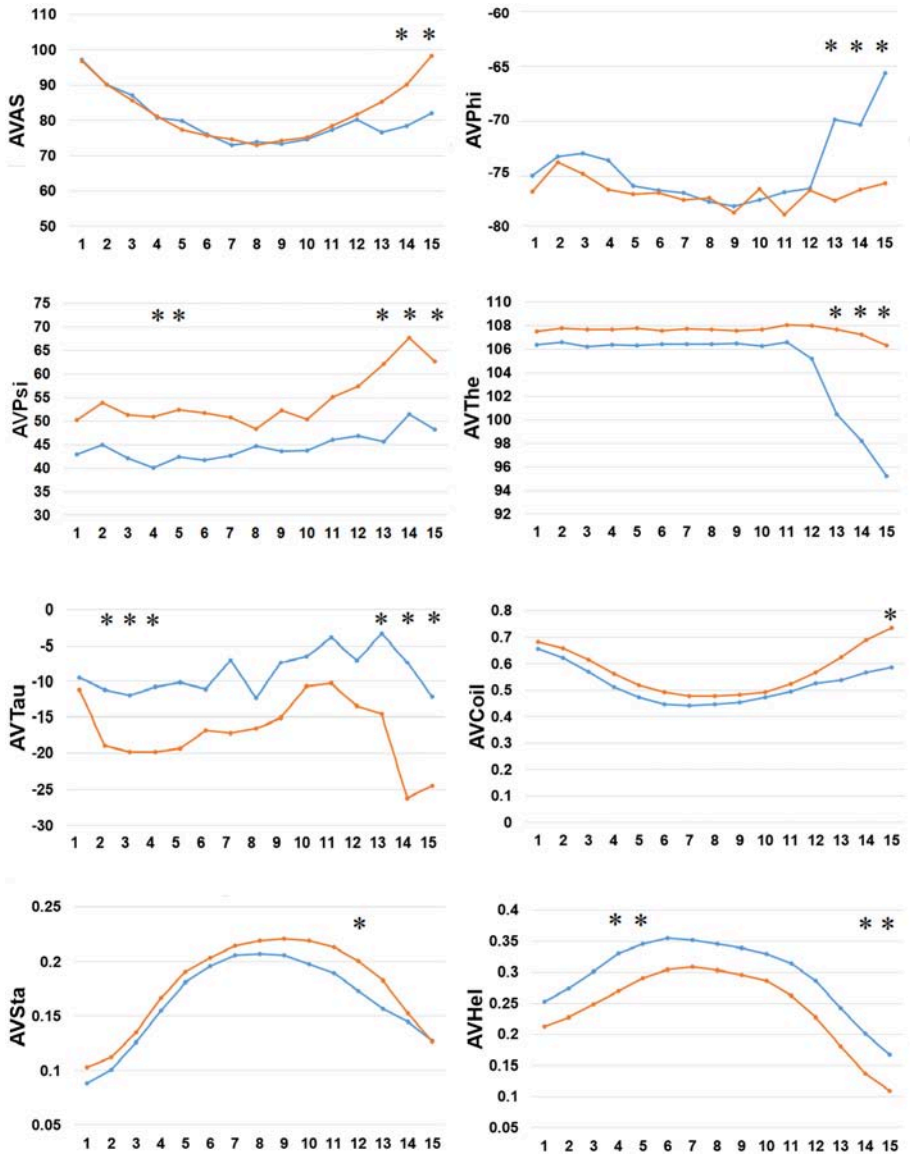


FIGURE 5 | Comparison of 8 types of the SFs by SPIDER2 between positive and negative AIPs. The row represents the N-terminal peptide, while the blue lines signify the positive AIPs and the orange lines the negative AIPs. "*" represents that the SFs are statistically different between both the samples with $p < 0.05$ by the KW test.

We accessed the performances of the training model of five successive encoding methods of AAindex, KSAAP, SPIDER2, PEP2D, and pKSAAP through a 10-fold CV test using the RF classifier. The prediction results by each of five encoding features and the "Combined features" are shown in **Figure 6A**. The AUCs of AAindex, KSAAP, SPIDER2, PEP2D, and pKSAAP were 0.774, 0.813, 0.739, 0.734, and 0.789, respectively. The KSAAP performed best for the 5 single encoding approaches in terms of Sn, MCC and AUC (**Table 1**). The "Combined features" (PreAIP) showed better performance with an AUC of 0.833 than any other single feature. It is noted that "Combined features" means a linear combination of the RF scores (Materials and Methods). Moreover, the PreAIP presented the highest AUC value (0.840) in the test dataset (**Figure 6B**). The performance of PreAIP was effective and reasonable for all the tested cases (**Figure 6**) and was best in the AIP prediction.

To present the known AIPs in the training dataset, we used BLAST to search the (weak) homologs, and ranked them to obtain the best hit e-value (Bhasin and Raghava, 2004). Total 256 positive and 397 negative hits were found out of 1,258 positive and 1,887 negative samples by BLASTP with an e-value of 1.0×10^{-2}. The reduced numbers of the samples may be due to the peptide length of 5–25. Then, we measured the BLAST performances through 10-fold CV test. The prediction performances of Sp, Sn, Ac, MCC, and AUC were 0.752, 0.269, 0.563, 0.159, and 0.632, respectively, which were lower than those

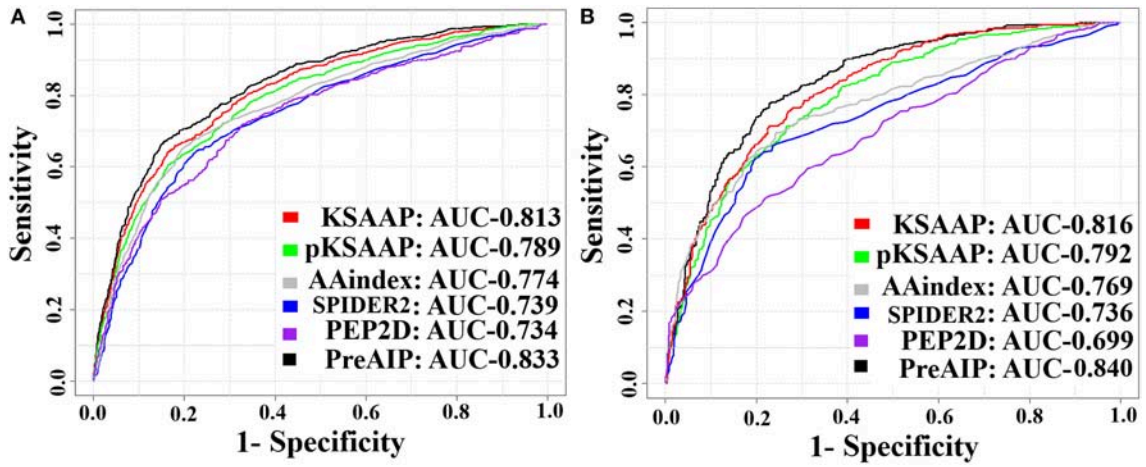


FIGURE 6 | ROC curves of the various prediction models. **(A)** 10-fold CV test on a training dataset and **(B)** test dataset. The PreAIP combined the KSAAP, pKSAAP, and AAindex methods. High AUC values show accurate performance.

TABLE 1 | AUC values for prediction performance of the training dataset by 10-fold CV test.

Methods	Sp	Sn	Ac	MCC	AUC	*p*-value
pKSAAP	0.798	0.647	0.738	0.450	0.789	0.017
AAindex	0.795	0.644	0.735	0.448	0.774	0.012
SPIDER2	0.765	0.434	0.633	0.235	0.739	0.004
PEP2D	0.769	0.411	0.629	0.219	0.734	0.004
KSAAP	0.805	0.656	0.745	0.463	0.813	0.118
PreAIP*	0.806	0.709	0.767	0.508	0.833	

**PreAIP is the linear combination of the RF scores estimated by SPIDER2, PEP2D, KSAAP, AAindex, and pKSAAP encoding schemes and their weight coefficients are 0.00, 0.00, 0.15, 0.25, and 0.6, respectively. A p-value was computed based on the final model of AUC values by using a Wilcoxson matched-pair signed test.*

by the other sequence encoding-based models. Therefore, we did not consider BLAST for final prediction.

In addition, we found that KSAAP performed best for all the five single encoding methods. To investigate the most significant residue of the KSAAP method, the top 20 amino acid pairs of AIPs were examined through the IG feature selection. The top 20 significant residue pair scores and their corresponding positions are listed in **Table S5**. These significant features are also presented using a radar diagram (**Figure 7A**). For example, the feature sequence motif "L×L," which is represented by 1-spaced residue pair of "LL," is the most important residue pair, where "×" stands for any amino acid. The feature "L×××L" represented the second enriched motif surrounding positive samples of AIPs. Similarly, the feature "LL," which represents a 0-spaced residue pair of "LL," is important and enriched in the negative samples AIPs. Similarly, to keep other *k*-space amino acid pairs from KSAAP, the same exemplification was employed. Residue preference analysis demonstrated that "L," "Y," "C," "D," and "I" residues frequently appear for AIPs (**Figures 2**, **7A**). These residues are expected to play a key role in the recognition of AIPs. To characterize the top 20 KSAAP-specific features, we compared the numbers of positive and negative AIPs. **Figure 7B** showed the top 20 average value of feature scores (AVFS) by the IG. The average of top 20 features was significantly different between two samples of AIPs with $p < 0.05$, suggesting the effectiveness of the KSAAP encoding. The significant residue pair scores are listed in **Table S5**, which provides some insights into the sequence patterns of the AIPs. They deserve further experimental validation.

Comparison of PreAIP With Existing Predictors Using Test Dataset

We evaluated the performances of PreAIP along with that of existing predictors on the test dataset. We submitted the test set to the AIPpred (Manavalan et al., 2018) and AntiInflam (Gupta et al., 2017) servers to assess the performance. It is noted that AntiInflam server provides different thresholds values. We used two threshold values of −0.3 and 0.5 and renamed as less accurate (LA) and more accurate (MA) models (Gupta et al., 2017), respectively. The AIPpred represents the state-of-the-art predictor available. The average performances of the LA, MA, AIPpred, and PreAIP are illustrated in the **Table 2**. The LA showed the highest Sp (0.892) with the lowest Sn (0.258), MCC (0.197), and AUC (0.647) for all the predictors. The PreAIP with the high threshold presented much higher Sn (0.618) Ac (0.770), MCC (0.512), and AUC (0.840) than LA, while it provided Sp (0.871) comparable to LA. The PreAIP with the low threshold showed the highest Sn (0.863), while keeping Sp, Ac, MCC, and AUC at a high level. While the AIPpred presented considerably high values to all the measures of Sp, Sn, Ac, MCC, and AUC, the PreAIP with the moderate threshold outperformed the AIPpred, presenting well-balanced, high prediction performances. The PreAIP performance improvement was found distinct on the test dataset by the Wilcoxson matched-pair signed test, demonstrating its ability to predict unseen peptides.

Comparison of PreAIP With AIPpred Using Training Dataset

We compared the performance of the proposed PreAIP with the AIPpred using the same training dataset. In this study, the same dataset as the AIPpred set was used to make a fair comparison

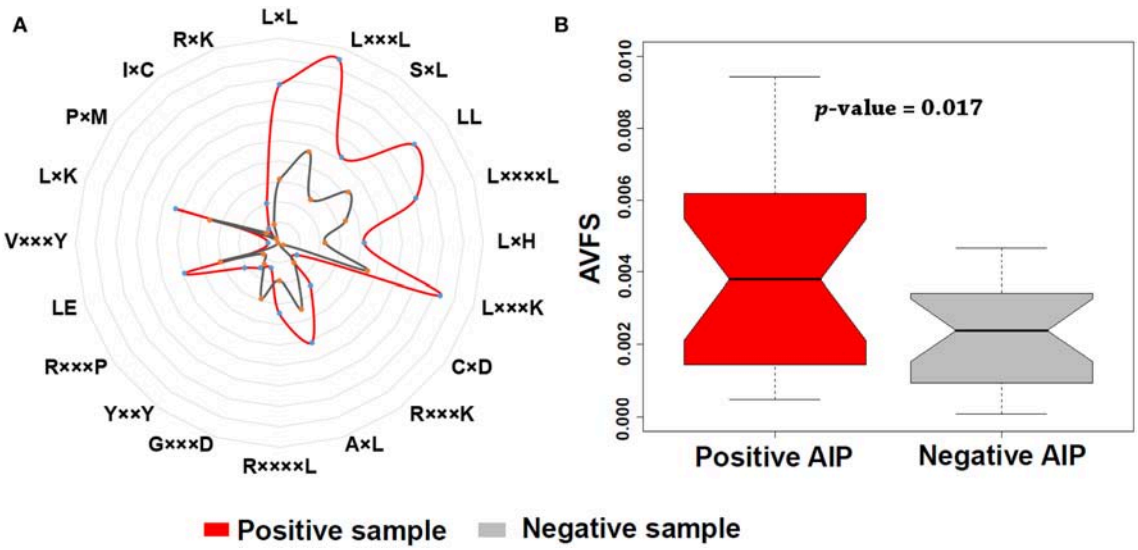


FIGURE 7 | Top 20 amino acid pairs selected by the IG feature of the KSAAP method. **(A)** The radar diagram is represented by the composition of each amino acid pair whose length is proportional to the composition of KSAAP features. **(B)** Box plot shows the top 20 average value of feature scores (AVFS) by the IG. Red color denotes the positive AIPs, while gray color denotes the negative AIPs. The *p*-value is computed by two-sample *t*-test.

TABLE 2 | Performance comparison with exiting predictors using test dataset.

Predictor	Threshold	Sp	Sn	Ac	MCC	AUC	*p*-value
AntiInflam (LA)	−0.3	0.892	0.258	0.638	0.197	0.647	<0.001
AntiInflam (MA)	0.5	0.417	0.786	0.565	0.210	0.706	<0.001
AIPpred	Server	0.746	0.741	0.744	0.479	0.813	0.039
PreAIP	High	0.871	0.618	0.770	0.512	0.840	
	Moderate	0.747	0.784	0.762	0.522	0.840	
	Low	0.636	0.863	0.727	0.492	0.840	

A p-value was computed based on AUC values by using a Wilcoxson matched-pair signed test and p < 0.05 indicates a statistically significant difference between the proposed PreAIP and each selected method. The performances of AntiInflam LA and MA methods were computed using default threshold (server) values of −0.3 and 0.5, respectively. The AIPpred threshold was the same as given by its server.

TABLE 3 | Performance comparison of PreAIP with AIPpred using training dataset.

Methods	Threshold	Sp	Sn	Ac	MCC	AUC	*p*-value
AIPpred	Default given in the server	0.711	0.758	0.730	0.460	0.801	0.034
PreAIP	High	0.903	0.632	0.795	0.566	0.833	
	Moderate	0.801	0.719	0.768	0.520	0.833	
	Low	0.709	0.784	0.739	0.484	0.833	

A p-value was computed based on AUC values by using a Wilcoxson matched-pair signed test and p < 0.05 indicates a statistically significant difference between the proposed PreAIP and AIPpred.

for prediction performance of AIPs. As shown in **Table 3**, the PreAIP achieved a better performance than the AIPpred in terms of Ac, Sp, Sn, MCC, and AUC. The AUC value was nearly 3% higher than the AIPpred predictor. The PreAIP performance (AUC) improvement over the AIPpred was demonstrated on the training set by the Wilcoxson matched-pair signed test (**Table 3**).

Comparison of Different Machine Learning Algorithms

The performance of the RF was compared to the three widely used machine learning algorithms, NB, SVM, and ANN by using the same training datasets and features, as shown in **Table 4**. The AUC values of the prediction by the five algorithms were calculated by a 10-fold CV test, while using the SPIDER2, PEP2D, AAindex, KSAAP, and pKSAAP encodings and their combined method. The RF provided higher AUC than any other algorithms for all the encoding methods and their combined method.

The Effect of Peptide Redundancy on the Predictive Model

The peptide redundancy may lead to the overestimation on the predictive performance. Therefore, we performed the CD-HIT with 60% identity cutoff at the peptide level (Huang et al., 2010). After removing the 60% sequence redundancy, we re-assembled a training dataset that contained 1,098 positive and 1,226 negative samples, and the test dataset that contained 308 positive and 275 negative samples. While the overall performance (AUC = 0.821) of the PreAIP by the 10-fold CV test decreased slightly (**Table S6**), the PreAIP could still achieve the best performance on the independent testing dataset (**Figure S1**). The PreAIP achieved 6 and 8% higher AUC values than the AntiInflam and the AIPpred, respectively, demonstrating that the PreAIP with the 60% peptide redundancy removal provides a stable or competitive performance compared with the other predictors, as well as the 80% peptide redundancy removal.

TABLE 4 | AUC values of AIP prediction by different machine learning algorithms based on a 10-fold CV test.

Algorithms	SPIDER2	PEP2D	AAindex	KSAAP	pKSAAP	Combined
RF	0.739	0.734	0.774	0.813	0.789	0.833
NB	0.659	0.655	0.707	0.729	0.717	0.736
SVM	0.698	0.677	0.738	0.766	0.749	0.779
ANN	0.662	0.649	0.716	0.741	0.736	0.753

"Combined" indicates that the performance of the optimized combined features. The combined score of RF was given as the sum of the five SPIDER2, PEP2D, AAindex, KSAAP, and pKSAAP features with weight values of 0.00, 0.00, 0.15, 0.25, and 0.6 respectively. In the same way, the weight values of NB, SVM, and ANN were given as (0.00, 0.00, 0.10, 0.35, and 0.55), (0.00, 0.00, 0.22, 0.45, and 0.33), and (0.00, 0.00, 0.18, 0.5, and 0.32), respectively.

Advantages of PreAIP

In theoretical viewpoints, comparison of the proposed PreAIP with existing predictors is summarized: (1) The PreAIP investigated the primary sequence, physicochemical properties, structural, and evolutionary features, although the AIPpred and AntiInflam predictors used only primary sequence encoding method. For instance, in AntiInflam method (Gupta et al., 2017), studied hybrid features based on primary sequence encoding schemes such as amino acid composition (AAC), dipeptide composition (DPC), and tripeptide composition with SVM algorithm. The AIPpred (Manavalan et al., 2018) studied individual composition (AAC, AAindex, DPC, and chain-transition-composition) through multiple machine learning algorithms. (2) Since existing prediction tools did not control the Sp level, users cannot understand which AIP is highly positive or negative from their servers. On the other hand, the PreAIP controlled Sp at high, moderate and low levels by changing the threshold of the RF scores, based on 10-fold CV test results. A limitation of the PreAIP is that the employed dataset is still small, but we believe that the dataset will grow to enable intensive identification of AIPs. In addition, the calculation speed remains to be improved. The processing time of the PreAIP was <3 min for one peptide sequence, where the generation of PSSM profiles requires a long time.

Server of PreAIP

A web server of the PreAIP has been developed and publically accessible at http://kurata14.bio.kyutech.ac.jp/PreAIP/. The web application was implemented by programming languages of Java scripts, Perl, R, CGI scripts, PHP, and HTML. After submitting a query sequence to the server, it generates consecutive feature vectors. Then, the server optimizes the performances through RFs. After completing the submission job, the server returns the result in the output webpage which consists of the job ID and probability scores of the predicted AIPs in a tabular form. A user gets a job ID like "2018032900067" and can save this ID for a future query. The server stores this job ID for one month. The input peptide sequence must be in the FASTA format. Each of the 20 types of standard amino acids must be written as one uppercase letter. See the test example on the server. The length of AIP sequence was limited from 1 to 25. If users submit 200 amino acids, the PreAIP takes first 1–25 residues to analyze. When the peptide contains less than 25 residues, the PreAIP provides gaps (–) to the missing residues to compensate a peptide length of 25.

CONCLUSIONS

We have designed an accurate and efficient computational predictor for identifying potential AIPs. It outperforms the existing methods and is effective in understanding some mechanisms of AIP identification. An IG-based feature selection method was carried out to suggest sequence motifs of AIPs from KSAAP encoding. A user-friendly web-server was developed and freely available for academic users.

AUTHOR CONTRIBUTIONS

MK, MH, and HK conceived and designed the study. MK and MH collected data and performed the analyses. MH, MK, and HK wrote the manuscript. All authors discussed the prediction results and commented on the manuscript.

ACKNOWLEDGMENTS

This work was supported by the Grant-in-Aid for Challenging Exploratory Research with JSPS KAKENHI Grant Number 17K20009. This research is partially supported by the developing key technologies for discovering and manufacturing pharmaceuticals used for next-generation treatments and diagnoses both from the Ministry of Economy, Trade and Industry, Japan (METI) and from Japan Agency for Medical Research and Development (AMED).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00129/full#supplementary-material

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Azhagusundari, B., and Thanamani, A. S. (2013). Feature selection based on information gain. *Int. J. Innov. Technol. Explor. Eng.* 2, 2278–3075.

Bhadra, P., Yan, J., Li, J., Fong, S., and Siu, S. W. I. (2018). AmPEP: sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci. Rep.* 8:1697. doi: 10.1038/s41598-018-19752-w

Bhasin, M., and Raghava, G. P. (2004). GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.* 32, W383– W389. doi: 10.1093/nar/gkh416

Boismenu, R., Chen, Y., Chou, K., El-Sheikh, A., and Buelow, R. (2002). Orally administered RDP58 reduces the severity of dextran sodium sulphate induced colitis. *Ann. Rheum. Dis.* 61(Suppl. 2), 19–24. doi: 10.1136/ard.61.suppl_2.ii19

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Carugo, O. (2013). Frequency of dipeptides and antidipeptides. *Comput. Struct. Biotechnol. J.* 8:e201308001. doi: 10.5936/csbj.201308001

Centor, R. M. (1991). Signal detectability - the use of roc curves and their analyses. *Med. Decis. Making* 11, 102–106. doi: 10.1177/0272989X9101100205

Corrigan, M., Hirschfield, G. M., Oo, Y. H., and Adams, D. H. (2015). Autoimmune hepatitis: an approach to disease understanding and management. *Br. Med. Bull.* 114, 181–191. doi: 10.1093/bmb/ldv021

Delgado, M., and Ganea, D. (2008). Anti-inflammatory neuropeptides: a new class of endogenous immunoregulatory agents. *Brain Behav. Immun.* 22, 1146–1151. doi: 10.1016/j.bbi.2008.06.001

Ferrero-Miliani, L., Nielsen, O. H., Andersen, P. S., and Girardin, S. E. (2007). Chronic inflammation: importance of NOD2 and NALP3 in interleukin-1beta generation. *Clin. Exp. Immunol.* 147, 227–235. doi: 10.1111/j.1365-2249.2006.03261.x

Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics* 20, 2479–2481. doi: 10.1093/bioinformatics/bth261

Gonzalez, R. R., Fong, T., Belmar, N., Saban, M., Felsen, D., and Te, A. (2005). Modulating bladder neuro-inflammation: RDP58, a novel anti-inflammatory peptide, decreases inflammation and nerve growth factor production in experimental cystitis. *J. Urol.* 173, 630–634. doi: 10.1097/01.ju.0000143192.68223.f7

Gonzalez-Rey, E., Anderson, P., and Delgado, M. (2007). Emerging roles of vasoactive intestinal peptide: a new approach for autoimmune therapy. *Ann. Rheum. Dis.* 66(Suppl 3), 70–76. doi: 10.1136/ard.2007.078519

Gribskov, M., and Robinson, N. L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.* 20, 25–33. doi: 10.1016/S0097-8485(96)80004-0

Gupta, S., Sharma, A. K., Shastri, V., Madhu, M. K., and Sharma, V. K. (2017). Prediction of anti-inflammatory proteins/peptides: an insilico approach. *J. Transl. Med.* 15:7. doi: 10.1186/s12967-016-1103-6

Hasan, M. M., Guo, D., and Kurata, H. (2017a). Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information. *Mol. Biosyst.* 13, 2545–2550. doi: 10.1039/C7MB00491E

Hasan, M. M., Khatun, M. S., and Kurata, H. (2018a). A comprehensive review of *in silico* analysis for protein S-sulfenylation sites. *Protein Pept. Lett.* 25, 815–821. doi: 10.2174/0929866525666180905110619

Hasan, M. M., Khatun, M. S., Mollah, M. N. H., Cao, Y., and Guo, D. (2017b). A systematic identification of species-specific protein succinylation sites using joint element features information. *Int. J. Nanomed.* 12, 6303–6315. doi: 10.2147/IJN.S140875

Hasan, M. M., Khatun, M. S., Mollah, M. N. H., Yong, C., and Dianjing G. (2018b). NTyroSite: Computational identification of protein nitrotyrosine sites using sequence evolutionary features. *Molecules* 23:1667. doi: 10.3390/molecules23071667

Hasan, M. M., and Kurata, H. (2018). GPSuc:global prediction of generic and species-specific succinylation sites by aggregating multiple sequence features. *PLoS ONE* 13:e0200283. doi: 10.1371/journal.pone.0200283

Hasan, M. M., Yang, S., Zhou, Y., and Mollah, M. N. (2016). SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. *Mol. Biosyst.* 12, 786–795. doi: 10.1039/C5MB00853K

Hasan, M. M., Zhou, Y., Lu, X., Li, J., Song, J., and Zhang, Z. (2015). Computational identification of protein pupylation sites by using profile-based composition of k-spaced amino acid pairs. *PLoS ONE* 10:e0129635. doi: 10.1371/journal.pone.0129635

Hearst, M. A. (1998). Support vector machines. *IEEE Intell. Syst.* 18–28. doi: 10.1109/5254.708428

Hernández-Flórez, D., and Valor, L. (2016). Protein-kinase inhibitors: a new treatment pathway for autoimmune and inflammatory diseases? *Reumatol. Clin.* 12, 91–99. doi: 10.1016/j.reuma.2015.06.004

Huang, S. H. (2015). Supervised feature selection: a tutorial. *Artif. Intell. Res.* 4:6. doi: 10.5430/air.v4n2p22

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Ialenti, A., Santagada, V., Caliendo, G., Severino, B., Fiorino, F., Maffia, P., et al. (2001). Synthesis of novel anti-inflammatory peptides derived from the amino-acid sequence of the bioactive protein SV-IV. *Eur. J. Biochem.* 268, 3399–3406. doi: 10.1046/j.1432-1327.2001.02236.x

Jin, Y., Wi, H. J., Choi, M. H., Hong, S. T., and Bae, Y. M. (2014). Regulation of anti-inflammatory cytokines IL-10 and TGF-beta in mouse dendritic cells through treatment with *Clonorchis sinensis* crude antigen. *Exp. Mol. Med.* 46:e74. doi: 10.1038/emm.2013.144

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205. doi: 10.1093/nar/gkm998

Kempuraj, D., Selvakumar, G. P., Thangavel, R., Ahmed, M. E., Zaheer, S., Raikwar, S. P., et al. (2017). Mast cell activation in brain injury, stress, and post-traumatic stress disorder and alzheimer's disease pathogenesis. *Front. Neurosci.* 11:703. doi: 10.3389/fnins.2017.00703

López, Y., Sharma, A., Dehzangi, A., Lal, S. P., Taherzadeh, G., Sattar, A., et al. (2018). Success: evolutionary and structural properties of amino acids prove effective for succinylation site prediction. *BMC Genom.* 19:923. doi: 10.1186/s12864-017-4336-8

Lowd, D. (2005). "Naive Bayes models for probability estimation," in *05 Proceedings of the 22nd International Conference on Machine Learning* (New York, NY), 529–536. doi: 10.1145/1102351.1102418

Manavalan, B., Basith, S., Shin, T. H., Choi, S., Kim, M. O., and Lee, G. (2017). MLACP: machine-learning-based prediction of anticancer peptides. *Oncotarget* 8, 77121–77136. doi: 10.18632/oncotarget.20365

Manavalan, B., Shin, T. H., Kim, M. O., and Lee, G. (2018). AIPpred: sequence-based prediction of anti-inflammatory peptides using random forest. *Front. Pharmacol.* 9:276. doi: 10.3389/fphar.2018.00276

Marie, C., Pitton, C., Fitting, C., and Cavaillon, J. M. (1996). Regulation by anti-inflammatory cytokines (IL-4, IL-10, IL-13, TGFbeta)of interleukin-8 production by LPS- and/ or TNFalpha-activated human polymorphonuclear cells. *Med. Inflamm.* 5, 334–340. doi: 10.1155/S0962935196000488

Michalski, R. S., Carbonell J. G., Mitchell T. M. (2013). *Machine Learning: An Artificial Intelligence Approach.* Berlin; Heidelberg: Springer-Verlag. doi: 10.1007/978-3-662-12405-5

Miele, L., Cordella-Miele, E., Facchiano, A., and Mukherjee, A. B. (1988). Novel anti-inflammatory peptides from the region of highest similarity between uteroglobin and lipocortin I. *Nature* 335, 726–730. doi: 10.1038/335726a0

Patterson, H., Nibbs, R., Mcinnes, I., and Siebert, S. (2014). Protein kinase inhibitors in the treatment of inflammatory and autoimmune diseases. *Clin. Exp. Immunol.* 176, 1–10. doi: 10.1111/cei.12248

Steinman, L., Merrill, J. T., Mcinnes, I. B., and Peakman, M. (2012). Optimization of current and future therapy for autoimmune diseases. *Nat. Med.* 18, 59–65. doi: 10.1038/nm.2625

Tabas, I., and Glass, C. K. (2013). Anti-inflammatory therapy in chronic disease: challenges and opportunities. *Science* 339, 166–172. doi: 10.1126/science.1230720

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2019). The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* 47:D339–D343. doi: 10.1093/nar/gky1006

Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., et al. (2017). SPIDER2: a package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. *Methods Mol. Biol.* 1484, 55–63. doi: 10.1007/978-1-4939-6406-2_6

Zhao, L., Wang, X., Zhang, X. L., and Xie, Q. F. (2016). Purification and identification of anti-inflammatory peptides derived from simulated gastrointestinal digests of velvet antler protein (Cervus elaphus Linnaeus). *J. Food Drug. Anal.* 24, 376–384. doi: 10.1016/j.jfda.2015.10.003

Zouki, C., Ouellet, S., and Filep, J. G. (2000). The anti-inflammatory peptides, antiflammins, regulate the expression of adhesion molecules on human leukocytes and prevent neutrophil adhesion to endothelial cells. *FASEB J.* 14, 572–580. doi: 10.1096/fasebj.14.3.572

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 05 March 2019
doi: 10.3389/fgene.2019.00172

Sexual Transcription Differences in *Brachymeria lasus* (Hymenoptera: Chalcididae), a Pupal Parasitoid Species of *Lymantria dispar* (Lepidoptera: Lymantriidae)

Peng-Cheng Liu[1,2], *Shuo Tian*[1,2] *and De-Jun Hao*[1,2]*

[1] *Co-Innovation Center for the Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing, China,*
[2] *The College of Forestry, Nanjing Forestry University, Nanjing, China*

Sex differences in gene expression have been extensively documented, but little is known about these differences in parasitoid species that are widely applied to control pests. *Brachymeria lasus* is a solitary parasitoid species and has been evaluated as a potential candidate for release to control *Lymantria dispar*. In this study, gender differences in *B. lasus* were investigated using Illumina-based transcriptomic analysis. The resulting 37,453 unigene annotations provided a large amount of useful data for molecular studies of *B. lasus*. A total of 1416 differentially expressed genes were identified between females and males, and the majority of the sex-biased genes were female biased. Gene Ontology (GO) and Pathway enrichment analyses showed that (1) the functional categories DNA replication, fatty acid biosynthesis, and metabolism were enhanced in females and that (2) the only pathway enriched in males was phototransduction, while the GO subcategories enriched in males were those involved in membrane and ion transport. In addition, thirteen genes involving transient receptor potential (TRP) channels were annotated in *B. lasus*. We further explored and discussed the functions of TRPs in sensory signaling of light and temperature. In general, this study provides new molecular insights into the biological and sexually dimorphic traits of parasitoids, which may improve the application of these insects to the biological control of pests.

Keywords: sexually dimorphic, *Brachymeria lasus*, transcriptomic analysis, sex determination, venom protein, transient receptor potential channels

OPEN ACCESS

Edited by:
Ancha Baranova,
George Mason University,
United States

Reviewed by:
Juan Pedro M. Camacho,
University of Granada, Spain
Nakatada Wachi,
University of the Ryukyus, Japan

***Correspondence:**
De-Jun Hao
djhao@njfu.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *12 October 2018*
Accepted: *18 February 2019*
Published: *05 March 2019*

Citation:
Liu P-C, Tian S and Hao D-J (2019) Sexual Transcription Differences in Brachymeria lasus (Hymenoptera: Chalcididae), a Pupal Parasitoid Species of Lymantria dispar (Lepidoptera: Lymantriidae). Front. Genet. 10:172. doi: 10.3389/fgene.2019.00172

INTRODUCTION

Parasitoids are animals that parasitize other organisms (Godfray, 1994). All invertebrate life stages, such as egg, larva or nymph, pupa and adult, can be attacked by oviposition on or in the host or by depositing a larva on or near a host (Boulton et al., 2015). Based on the number of offspring reared in a host, parasitoid wasps are classified as solitary (one parasitoid per host), quasi-gregarious (one parasitoid per host, but hosts are spatially clumped, such as a clutch of eggs on a leaf), or gregarious (multiple parasitoids per host). The vast majority of parasitoids are solitary wasps (Mayhew, 1998).

Parasitoids can also be classified as koinobionts (in which hosts continue to develop and grow to some extent) or idiobionts (in which hosts do not grow further). Parasitoid wasps are haplodiploid: males develop from unfertilized eggs and are haploid, while females develop from fertilized eggs and are diploid (Cook, 1993; Heimpel and de Boer, 2008). Parasitoid species (e.g., *Sclerodermus harmandi*, *Trichogramma*) are important insects and have been extensively applied to reduce the population size of pest species (Hassan, 1993; Li, 1994; Terayama, 1999; Zhishan et al., 2003; Parra and Zucchi, 2004; Lim et al., 2006). In addition to having important applications, parasitoid and mutualistic Chalcidoidea, such as jewel (*Nasonia vitripennis)* and fig (*Pleistodontes froggatti*) wasps, have been important study models of behavioral ecology and evolutionary biology for such traits as their sexual dimorphism in longevity, body size, and dispersal (Hamilton, 1967; Charnov, 1982; Yan et al., 1989; Godfray, 1994).

Animals from a broad range of taxa show sex differences, which include behavioral (Breedlove, 1992), physiological (Bardin and Catterall, 1981), and morphological dimorphisms (Darwin, 1871). It is often assumed that the majority of sexually dimorphic traits arise from differences in the expression of genes present in both sexes (Connallon and Knowles, 2005; Rinn and Snyder, 2005). Sex-biased gene expression has been documented in brown algae (Lipinska et al., 2015), birds (Pointer et al., 2013), nematodes (Albritton et al., 2014), *Daphnia pulex* (Eads et al., 2007), and multiple insect species, including *Drosophila* (Jin et al., 2001; Arbeitman et al., 2002; Ranz et al., 2003; Chang et al., 2011), *Anopheles gambiae* (Hahn and Lanzaro, 2005; Marinotti et al., 2006; Baker et al., 2011), *Tribolium castaneum* (Prince et al., 2010), vespid wasps (Hunt and Goodisman, 2010), and *Bemisia tabaci* (Wen et al., 2014). However, few studies of sex differences in gene expression have been done in Hymenoptera insects, and these studies have focussed mainly on social species (e.g., honeybee; Cameron et al., 2013) and model organisms of parasitoids, e.g., jewel wasp *N. vitripennis* (Wang et al., 2015), which is a classic gregarious species. Most species of parasitoid wasps are thought of as solitary species (Mayhew, 1998), but their sexual transcription differences have not been addressed.

Gypsy moth, *Lymantria dispar* is a worldwide pest, and its pupal stage can be parasitized by *Brachymeria lasus*. *B. lasus* is a solitary parasitoid species and has been evaluated as a potential candidate for release to control *L. dispar* (Simser and Coppel, 1980), *Homona magnanima* (Mao and Kunimi, 1991) and *Sylepta derogate* (Kang et al., 2006). In addition, *B. lasus* has a wide host range, including many Lepidoptera species (e.g., *Mythimna separata*, *Hyphantria cunea*, and *Cnaphalocrocis medinalis*) (Habu, 1960). Male and female *B. lasus* differ in many important biological traits, including longevity (Mao and Kunimi, 1994b); development time in the egg, larval and pupal stages (Mao and Kunimi, 1994a); secondary symbionts; and body size (Yan et al., 1989). As *B. lasus* is a classic solitary species with many documented sex differences, though not yet at the gene expression level, it was used as the experimental material in this study. To reveal *B. lasus* sex differences at the transcriptional level, we carried out an Illumina-based transcriptomic analysis. This study attempted to provide comprehensive insight into the sexually dimorphic traits of parasitoid wasps at the transcriptome level to improve our understanding of other biological traits with the aim of improving the application of parasitoids to the biological control of pest species.

MATERIALS AND METHODS

Insect Cultures

In northern China, in addition to *L. dispar*, *B. lasus* is also an important pupal parasitoid of *H. cunea*, for which the parasitism ratio is approximately 1.06–3.39% in the field (Yang et al., 2001). To acquire *B. lasus* adults, we collected the pupae of *H. cunea*, which may be parasitized by *B. lasus* and other parasitoid species (e.g., *Coccygomimus disparis* Viereck; *Chouioia cunea* Yang) from a field in Xuzhou City, Jiangsu Province, China, in March 2016. After collection, we isolated the pupae individually in polyethylene tubes (height: 7.5 cm; diameter: 1 cm) whose openings were covered by a cotton ball and incubated them at a temperature of 28 ± 0.5°C, a relative humidity (RH) of 70 ± 5% and a photoperiod of 14 L:10 D. We observed and selected *B. lasus* after adult eclosion.

Transcriptomic Analyses

For the transcriptomic experiment, only 1-day-old *B. lasus* adults were selected, and the sex was determined under a microscope (Leica M205A, Germany). Then, five adults of the same sex were pooled into a plastic tube (1.5 ml), snap-frozen in liquid nitrogen, and transferred to a –80°C freezer for long-term storage. RNA from each sample group (whole bodies of female and male adults) was extracted with TRIzol reagent (Invitrogen; United States). Each group had three replicates. The quality of the isolated RNA was assessed using a NanoDrop (Thermo Fisher Scientific NanoDrop 2000, United States), and the A260/280 values were all above 2.0. A total of 3 μg total RNA from each sample was converted into cDNA using the NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, United States). In total, six cDNA libraries were constructed and subsequently sequenced with the Illumina HiSeq 2000 platform by Beijing Biomarker Technologies Co., Ltd, resulting in raw reads. Raw sequence data generated were deposited into Sequence Read Archive (SRA) database of NCBI with the accession no. PRJNA513855. Clean reads were obtained by removing reads containing adapter, poly-N reads and low-quality reads from the raw data using FASTX-Toolkit[1], and these clean reads were used for further analysis. Then, transcriptome assembly was performed using Trinity (v2.5.1) with the default parameters (Grabherr et al., 2011). For functional annotation, pooled assembled unigenes were searched using BLASTX (v2.2.31) against five public databases, Clusters of Orthologous Groups (COG), Swiss-Prot, NCBI non-redundant protein sequences (nr), KEGG Ortholog database (KO) and GO, with an *E*-value cutoff of 10^{-5}. Using our assembled transcriptome as a reference, we identified putative genes expressed in males and females by RSEM (Li and Dewey, 2011),

[1]http://hannonlab.cshl.edu/fastx_toolkit/

using the reads per kb per million reads (RPKM) method. Genes with at least 2-fold changes (i.e., $\log_2|FC| \geq 1$) and a false discovery rate [FDR] < 0.01 as found by DESeq R package (1.10.1) were considered differentially expressed. The GOseq R package (Young et al., 2010) and KOBAS software (Mao et al., 2005) were used to implement the statistical enrichment of differentially expressed genes (DEGs) in the GO and KEGG pathways, respectively, and an adjusted Q-value <0.05 was chosen as the significance cutoff.

Validation by mRNA Expression and Behavior

Based on transcriptomic data, a gene of transient receptor potential (*trp*) involved in the phototransduction pathway enriched only in males (ko: 04745; **Supplementary Figure S1-d**), *trp* (Leung et al., 2000), was down-regulated in females, which may lead to a reduction in light response (Leung et al., 2000; Popescu et al., 2006). Therefore, we checked this result at the mRNA expression and behavioral levels.

Quantitative Real-Time PCR (qRT-PCR) Analysis

Total RNA was extracted from the whole bodies of five female or five male adults reared on the pupae of *H. cunea* using TRIzol (Invitrogen; United States) according to the manufacturer's protocols, then resuspended in nuclease-free water. Finally, the RNA concentration was measured using a NanoDrop (Thermo Fisher Scientific NanoDrop 2000; United States). Each group have four replicates. Approximately 0.5 mg of total RNA was used as template to synthesize the first-strand cDNA using a PrimeScript RT Reagent Kit (TaKaRa; Japan) following the manufacturer's protocols. The resultant cDNA was diluted to 0.1 mg/ml for further qRT-PCR analysis (ABI StepOne Plus; United States) using SYBR Green Real-Time PCR Master Mix (TaKaRa; Japan). Primers (**Supplementary Table S1**) for *trp* gene were designed using Primer Express 2.0 software. The cycling parameters were 95°C for 30 s followed by 40 cycles of 95°C for 5 s and 62°C for 34 s, ending with a melting curve analysis (65 to 95°C in increments of 0.5°C every 5 s) to check for nonspecific product amplification. Relative gene expression was calculated by the $2^{-\Delta\Delta Ct}$ method using the housekeeping gene GAPDH as a reference to eliminate sample-to-sample variations in the initial cDNA samples.

Phototaxis Assays

A glass Y-maze (main arm: 12 cm; two side arms: 5 cm; inner diameter: 1.5 cm; angle between two side arms: 75°) was used for phototaxis assays in a completely dark room (<10 lux, measured by illuminometer, LX-9621, China) at a temperature of 22–26°C. One 1-day-old *B. lasus* adult (female or male) began the trial in a tube at the base of the apparatus and faced a choice between two tubes, one of which was dark and the other of which was lighted with a 40-watt bulb (approximately 600 lux). After 1 min, the choice was recorded. The sample sizes of the male and female groups were 18 and 24, respectively. After each test, the Y-maze was washed and dried, and the two side arms were changed for the new test.

Statistical Analysis

Prior to analysis, the raw data were tested for normality and homogeneity of variances with the Kolmogorov-Smirnov test and Levene's test, respectively, and the data were log-transformed if necessary. The qRT-PCR data comparing gene expression in females and males were analyzed with the independent *t*-test. In phototaxis assays, the preferences for light and dark were analyzed using sign tests, and the differences in female and male phototaxis were analyzed by the chi-square test. All analyses were performed using SPSS v.20 (IBM SPSS, Armonk, NY, United States).

RESULTS AND DISCUSSION

Sexual dimorphism is the condition where the two sexes of the same species exhibit different characteristics (e.g., size, color, behavior) beyond the differences in their sexual organs (Bonduriansky, 2007). Most sexually dimorphic traits are often assumed to arise from differences in the expression of genes present in both sexes (Connallon and Knowles, 2005; Rinn and Snyder, 2005). To reveal *B. lasus* sex differences at the transcriptional level, we carried out an Illumina-based transcriptomic analysis.

Transcriptome Sequencing, Read Assembly and Annotation

All high-quality reads (101,945,678) from the six samples were pooled and assembled by using Trinity with the default parameters, and a total of 254,656 transcripts with lengths longer than 200 bp were generated. The N50 size was 2706 bp with 57,605 sequences longer than 1 kb. We chose the longest isoform of each gene to construct the unigene set. After isoforms were considered, these assembled transcripts were predicted to be produced from a total of 164,709 unigenes. The N50 size of the unigenes was approximately 814 bp, and their mean length was 572.08 bp (**Supplementary Table S2**). For annotation, the pooled assembled unigenes were searched using blastx against five public databases with an *E*-value cutoff of 10^{-5}. A total of 37,453 unigenes were successfully annotated, as shown in **Table 1**, including 17,248 genes in GO, 13,491 genes in COG, 35,427 genes in nr, 18,195 genes in Swiss-Prot, and 15,133 genes in KEGG.

In the GO analysis, 17,248 unigenes were successfully annotated and classified into three major GO categories: molecular function (MF), cell component (CC), and biological

TABLE 1 | Annotation of a pooled assembly including both male and female *B. lasus* transcriptomes.

Annotation database	Annotated unigenes	Number of DEGs
COG	13,491	420
GO	17,248	442
KEGG	15,133	396
Swiss-Prot	18,195	613
nr	35,427	1024
Total	37,453	1416

processes (BP), then assigned to 56 subcategories based on GO level 2. The dominant subcategories for the classified genes were catalytic activity and binding for the MF category; cell and cell part for the CC category; and metabolic process, cellular process, and single-organism process for the BP category (**Supplementary Table S3**). A total of 15,133 KEGG-annotated unigenes were classified into 190 pathways (>10 associated unigenes). Among these pathways, the ten most highly represented were ribosome, carbon metabolism, protein processing in endoplasmic reticulum, oxidative phosphorylation, biosynthesis of amino acids, spliceosome, RNA transport, purine metabolism, peroxisome, and ubiquitin mediated proteolysis (**Supplementary Table S4**).

Sex-Biased Genes

Although in most species the male and female genomes differ by a few genes located on sex-specific chromosomes (such as the Y chromosome of mammals), the vast majority of sexually dimorphic traits result from the differential expression of genes that are present in both sexes (Connallon and Knowles, 2005; Rinn and Snyder, 2005; Ellegren and Parsch, 2007), and this is especially true in hymenopteran insects. Because sex determination in hymenopteran species is haplodiploid, females and males are nearly identical genetically (Ellegren and Parsch, 2007). Such DEGs include those that are expressed exclusively in one sex (sex-specific expression) and those that are expressed in both sexes but at a higher level in one sex (sex-biased expression). These sex-biased genes can be further separated into male-biased and female-biased genes, depending on which sex shows higher expression. Genes with equal expression in the two sexes are referred to as unbiased (Ellegren and Parsch, 2007).

Using our assembled transcriptome as a reference, we identified putative genes expressed in males and females using the RPKM method, and genes with at least 2-fold changes and FDR < 0.01 were defined as DEGs. By comparing female and male transcriptomes, 1416 DEGs were found in *B. lasus*, of which 442 genes were annotated in GO, 420 in COG, 1024 in nr, 613 in Swiss-Prot, and 396 in KEGG (**Table 1**). Among these DEGs, 986 were up-regulated in females and 430 were up-regulated in males (**Supplementary Table S5**).

GO Enrichment Analyses

In the GO enrichment analyses, 12 and five subcategories were enriched in females and males, respectively. In females, the enriched subcategories were microtubule cytoskeleton, cytoskeletal part, MCM complex, nucleus, protein complex, kinesin complex, and nucleosome for the CC category; DNA replication initiation, cell division and protein phosphorylation for the BP category; and alpha-1,4-glucosidase activity and zinc ion binding for the MF category (**Figure 1A**). These results showed that, consistent with the results in flies, mosquitoes, and *Daphnia* (Ranz et al., 2003; Hahn and Lanzaro, 2005; Eads et al., 2007), including Hymenoptera insects of *Nasonia* (Wang et al., 2015), most categories were related to DNA replication, which are probably expressed to produce eggs in females (Spradling, 1993; Parisi et al., 2004). The over-representation of transcripts from genes required for DNA replication may be required for nurse cell polyploidization or for the rapid division of embryonic cells, which rely on maternally deposited gene products (Spradling, 1993; Parisi et al., 2004).

In males, the enriched subcategories were integral component of membrane, cell junction, and postsynaptic membrane for the CC category; ion transport for the BP category; and potassium channel activity for the MF category (**Figure 1B**), consistent with a study in *D. melanogaster* (Parisi et al., 2004), which may be mainly related to spermatogenesis (Fuller, 1993). For example, the enriched subcategories associated with membranes were likely due to the requirements of the sperm axoneme structure (Parisi et al., 2004). However, in parasitoids of *N. vitripennis* species, highly enriched subcategories in males are related to sex-pheromone synthetic enzymes (Wang et al., 2015). Those differences might be likely to contribute by their difference in sexual maturity period. Sexual maturity in many gregarious and quasi-gregarious males (e.g., *N. vitripennis*) happens before eclosion, and these males can immediately mate with females after eclosion and near the emergence site (Boulton et al., 2015), while solitary *B. lasus* have mating ability for some days after eclosion (Yan et al., 1989).

KEGG Pathway Enrichment Analyses

Consistent with the results of GO enrichment in females, pathway enrichment tests revealed that DNA replication (ko: 03030; **Supplementary Figure S1-a**) was enriched in *B. lasus* females. The functional categories enriched in females also included fatty acid biosynthesis (ko: 00061; **Supplementary Figure S1-b**) and metabolism (ko01212; **Supplementary Figure S1-c**). The fatty acid synthase gene (FASN), which encoded the enzyme catalyzing fatty acid synthesis (Jayakumar et al., 1994, 1995; Persson et al., 2008), was probably crucial for egg yolk production and thus female fecundity. In some insects, for example yellow fever mosquito *Aedes aegypti*, brown planthopper *Nilaparvata lugens*) (Alabaster et al., 2011; Li et al., 2016), when *FAS* expression decreases in females, the number of oviposited eggs significantly decreases.

We found that only the phototransduction-fly pathway (ko: 04745; **Supplementary Figure S1-d**) was enriched in males, which is associated with perception of light signals (Leung et al., 2000). Its potential functions are discussed below.

Annotated Genes Involved in Venom Proteins

In terms of biological control, parasitoid species have been extensively applied for reducing pest species population sizes (Hassan, 1993; Li, 1994; Terayama, 1999; Zhishan et al., 2003; Parra and Zucchi, 2004; Lim et al., 2006) because parasitoids can propagate on or in other arthropods. The venom of parasitoid wasps, which is injected into a host by females before or at oviposition, is important for the successful development of the progeny. Parasitoid venoms have diverse physiological effects on hosts, including developmental arrest; alteration in growth and physiology; suppression of immune responses; induction of paralysis, oncosis, or apoptosis; and alteration of host behavior (Edwards et al., 2006; Price et al., 2009; Tian et al., 2010; Kryukova et al., 2011). In total, three female-biased

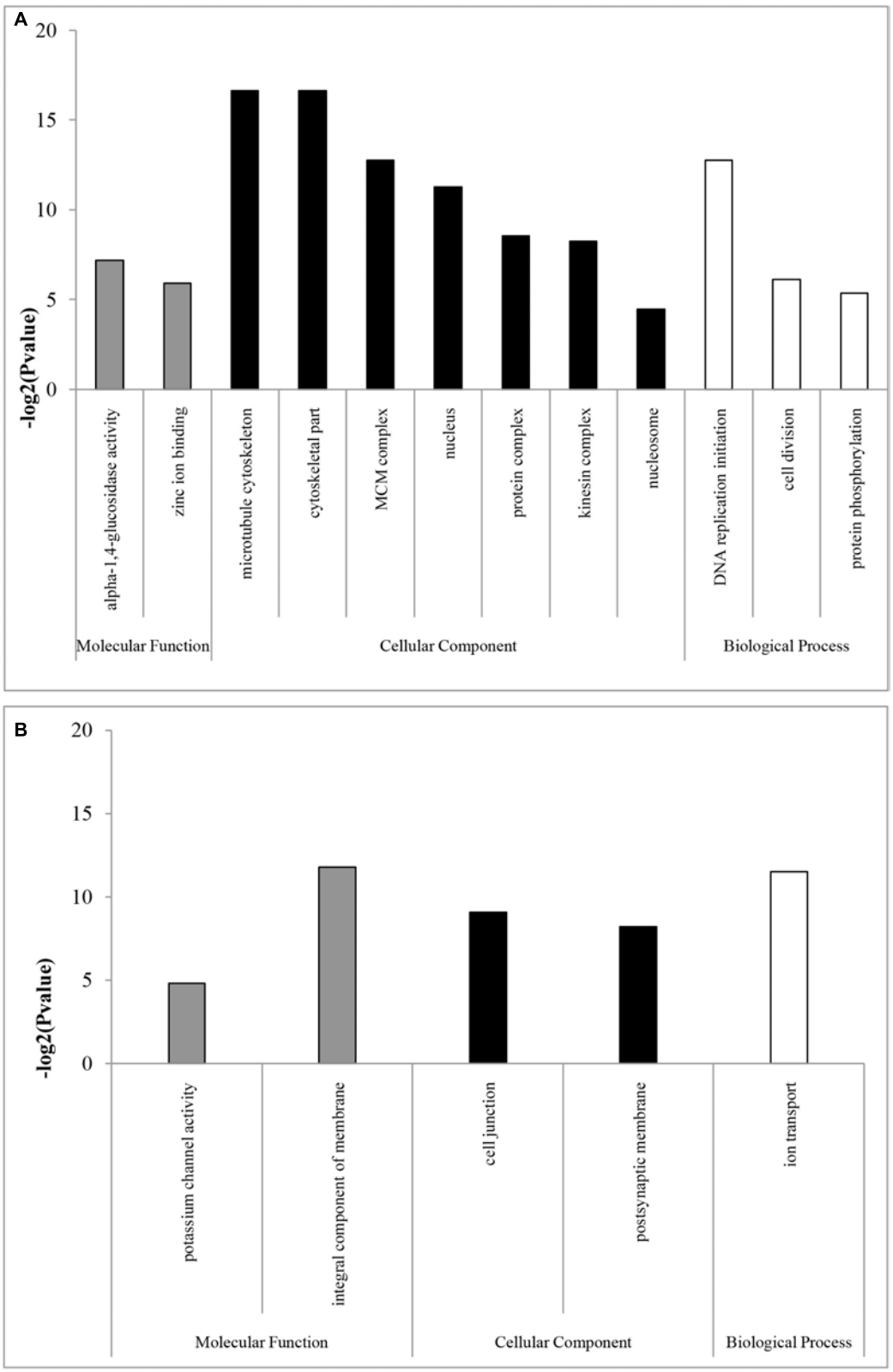


FIGURE 1 | GO enrichment analysis of **(A)** female- and **(B)** male-biased genes. GOSeq explicitly takes into account gene selection bias due to differences in gene length and thus the numbers of overlapping sequencing reads. GOSeq was used for the GO enrichment analysis, and an adjusted Q-value <0.05 was chosen as the significance cutoff.

TABLE 2 | TRP channel genes in the *B. lasus* transcriptome.

Gene name	Subfamily	*Drosophila* orthologue name	Function in *Drosophila*	Comparative analyses with RNAseq data
c103240.graph_c0	TRPC	*trp*	phototransduction	up
c107438.graph_c0		*trp* gamma	phototransduction	normal
c107438.graph_c1		*trp* gamma	phototransduction	normal
c87378.graph_c0		*trp* gamma	phototransduction	normal
c107458.graph_c0	TRPM	*trpm*	unknown	normal
c107458.graph_c1		*trpm*	unknown	normal
c103139.graph_c0	TRPA	pyrexia	geotaxis	normal
c106854.graph_c0		pyrexia	geotaxis	normal
c107721.graph_c1		pyrexia	geotaxis	normal
c108434.graph_c0		pyrexia	geotaxis	normal
c89491.graph_c0		pyrexia	geotaxis	up
c106747.graph_c0		painless	nociception	normal
c108178.graph_c0	TRPML	*trpml*	TRPML	normal

genes (c100635.graph_c0, c101314.graph_c0, c101670.graph_c0; **Supplementary Table S5**) in this study were annotated for venom proteins, which were related to known insect venoms from *N. vitripennis* and belonged to previously known insect venom families, such as serine proteases (Graaf et al., 2010; Werren et al., 2010). Despite the large diversity of parasitoid wasp species, only a small number of venom proteins have been described from wasps. A wealth of unexplored biomolecules is present in parasitoid venoms; these proteins are of value in basic evolutionary studies, venom biology, host-parasite interactions, and the study of the evolution of life strategies, and they may potentially contain components that could be used in biological control and pharmacology (Moreau and Asgari, 2015).

Annotation of Genes in the TRP Channel Family and Function Validation

Transient Receptor Potential channels are cation channels that are mainly considered as unique polymodal cell sensors; TRPs can be subdivided into six main subfamilies: the TRPC (canonical), TRPV (vanilloid), TRPM (melastatin), TRPP (polycystin), TRPML (mucolipin), and TRPA (ankyrin) groups (Gees et al., 2010). Functionally, TRP channels cause cell depolarization when activated, which may trigger many voltage-dependent ion channels. Upon stimulation, Ca^{2+}-permeable TRP channels generate changes in the intracellular Ca^{2+} concentration,

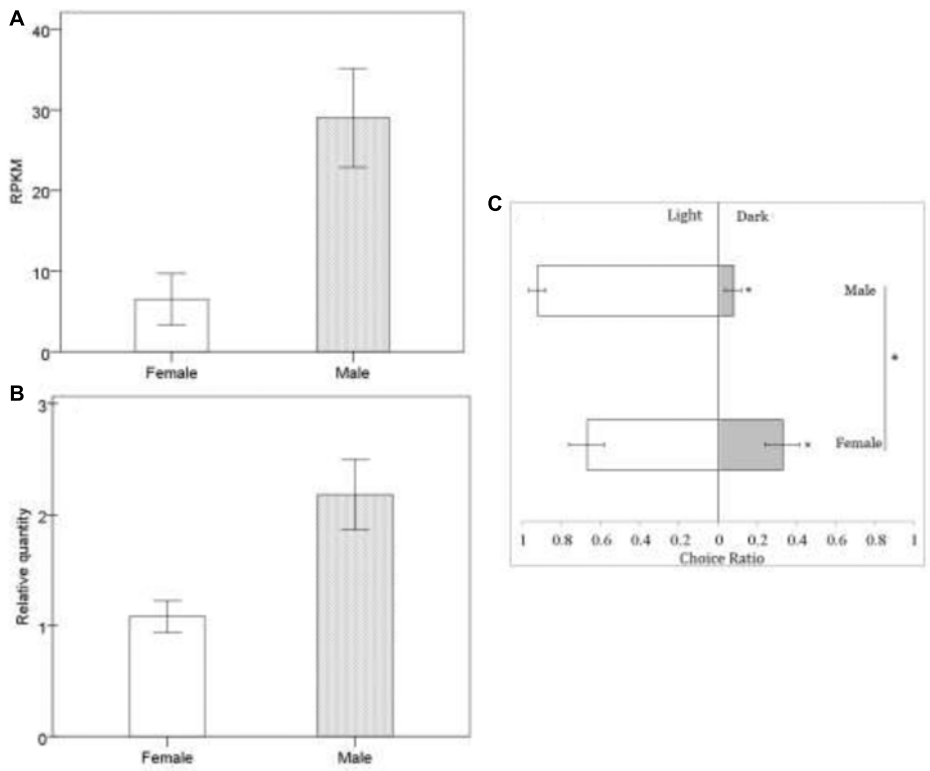


FIGURE 2 | Sexual difference in response to light at mRNA level **(A,B)** and behavioral level **(C)**. In transcriptomic data, we identified putative genes expressed using the reads per kb per million reads (RPKM) method. Quantitative real-time PCR (qRT-PCR) analysis was used to calculated the relative gene expression to further check the transcriptomic data, in which the differences in female and male were analyzed by the independent *t*-test. There was a highly significant correlation co-efficient of 0.885 between transcriptomic data and qRT-PCR data. Behavioral responses of *Brachymeria lasus* adults to dark or light were tested with phototaxis assays. The differences in female (n = 24) and male (n = 18) phototaxis were analyzed by the chi-square test. *indicates $p < 0.05$. The error bars indicate standard errors.

$[Ca^{2+}]_i$, due to Ca^{2+} entry via the plasma membrane. However, evidence is increasing that TRP channels are also located in intracellular organelles and serve as intracellular Ca^{2+} release channels (Berridge et al., 2000; Bootman et al., 2001; Gees et al., 2010). TRP channels in *Drosophila* are involved in the perception of sensory signals such as light, temperature, humidity, pheromones, sound, and touch (Lin et al., 2005). In our study, we found 13 TRP channel genes in *B. lasus*; *Nasonia* and honey bee contain 12 and 11 genes, respectively, indicating that the number of *trp* channels seems to be well conserved in Hymenoptera (Werren et al., 2010). Of the TRP channel genes in *B. lasus*, most belong to two subfamilies: TRPC and TRPA (**Table 2**).

In *Drosophila*, TRPC plays an important role in the perception of light signals, i.e., the phototransduction pathway (Leung et al., 2000) (ko: 04745; **Supplementary Figure S1-d**), which was enriched in *B. lasus* male adults. In *Drosophila*, a number of genes in the visual signal transduction pathway have been characterized, with functions including rhodopsin activation, phosphoinoside signaling, and the opening of TRP and TRPL channels (Wolff and Ready, 1993; Zuker, 1996; Leung et al., 2000; Wang and Montell, 2007). Our transcriptional analyses (**Figure 2A**: FDR < 0.01, $\log_2$ FC $= 1.62$) and q-PCR results (**Figure 2B**: $t = -3.169$, $df = 6$, $p = 0.019$), showed that the gene corresponding to *trp* (c103240.graph_c0) was more highly expressed in *B. lasus* males, consistent with the phototaxis test. Although both females and males tended to move toward light (**Figure 2C**: female, $Z = -1.34$, $p < 0.05$; male, $Z = -1.6$, $p < 0.05$), the tendency to prefer light was significantly influenced by sex in adults (**Figure 2C**. $\chi^2 = 4.17$, $df = 1$, $p < 0.05$), males more preferring to move to light. This result is similar to the results of research on *trp* mutants in *Drosophila*, which had altered phenotypes, including a reduction in light response (Leung et al., 2000; Popescu et al., 2006). Female reduction in light response might be due to their long periods living in the dark to search for hosts and lay offspring into them, as most host species (e.g., pupae of *L. dispar* or *H. cunea*) hide in dark environments, such as the litter horizon (Yan et al., 1989; Yang et al., 2001). Surprisingly, five members of the TRPA subfamily, which is involved in sensing environmental temperature, were annotated in our study. Animals must maintain thermal homeostasis and avoid prolonged contact with harmfully hot or cold objects (Caterina, 2007; Karashima et al., 2009). Unlike most parasitoid species, which overwinter in their hosts as eggs or larvae, *B. lasus* lives through the winter in its adult stage (Yan et al., 1989). Thus, TRPA may be essential for *B. lasus* adults, allowing them to sense harmful cold during winter. In addition, intraspecific aggregations in *B. lasus* have been observed in previous research, and an active component that elicited the aggregation response was isolated and identified as 3-hexanone (Mohamed and Coppel, 1987). The effects of aggregation behavior include mating, host attack, defense, and thermoregulation, and in this species, a previous study suggested that aggregation resulted from an increase in reproductive success by increasing the probability of mate location, as well as offering the possibility of mate choice (Mohamed and Coppel, 1987). However, combining the above results, adults may also aggregate at a site for purposes of thermoregulation, especially in winter, in response to cold. Further studies are required to elucidate the nature of this cue.

CONCLUSION

Brachymeria lasus is a solitary parasitoid species and has been evaluated as a potential candidate for release to control *L. dispar*. Whereas previous studies have focussed on the application of parasitoids and their sex differences in phenotypes, this study focussed mainly on sex differences in gene expression. *Brachymeria lasus* as a representative of solitary species was studied, which enriched our understanding of sexual transcription differences in parasitoid wasps, especially solitary species. Here, we performed transcriptome assembly using the Trinity program, which provided a large amount of useful information for molecular studies of *B. lasus*, including venom protein and perception of sensory signals. In addition to sex-biased genes, epigenetic processes, such as DNA methylation, are known to play important roles in differentiating phenotype and have been widely studied in Hymenopteran insects, for example, female morphs (queens and workers) in the honeybee, *Apis mellifera* (Kucharski et al., 2008; Lyko et al., 2010), although these processes do not appear to be in *Nasonia* (Wang et al., 2015). More future research will be conducted to better understand the molecular mechanisms underlying the biological traits of sex differences in *B. lasus* and to better apply this parasitoid to the biological control of pests.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://dataview.ncbi.nlm.nih.gov/object/PRJNA513855.

ETHICS STATEMENT

There was no requirement to seek ethical approval to carry out the work described above. However, the use of insects in the above experiments was kept to a minimum.

AUTHOR CONTRIBUTIONS

P-CL conceived and designed the experiments. P-CL and ST performed the experiments. P-CL and D-JH wrote the manuscript. All the authors reviewed the manuscript.

FUNDING

A project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD). This work was also supported by the Doctorate Fellowship Foundation of Nanjing Forestry University and the Natural Science Foundation of Jiangsu Province (BK20131421).

ACKNOWLEDGMENTS

We gratefully acknowledge undergraduates Ju Luo, Min Li, and Chenxi Zhao of the Nanjing Forestry University for their assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00172/full#supplementary-material

REFERENCES

Alabaster, A., Isoe, J., Zhou, G., Lee, A., Murphy, A., Day, W. A., et al. (2011). Deficiencies in acetyl-CoA carboxylase and fatty acid synthase 1 differentially affect eggshell formation and blood meal digestion in *Aedes aegypti*. *Insect Biochem. Mol. Biol.* 41, 946–955. doi: 10.1016/j.ibmb.2011.09.004

Albritton, S. E., Kranz, A. L., Rao, P., Kramer, M., Dieterich, C., and Ercan, S. (2014). Sex-biased gene expression and evolution of the x chromosome in nematodes. *Genetics* 197, 865–883. doi: 10.1534/genetics.114.163311

Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., et al. (2002). Gene expression during the life cycle of *Drosophila melanogaster*. *Science* 297, 2270–2275. doi: 10.1126/science.1072152

Baker, D. A., Nolan, T., Fischer, B., Pinder, A., Crisanti, A., and Russell, S. (2011). A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector. *Anopheles gambiae*. *BMC Genomics* 12:296. doi: 10.1186/1471-2164-12-296

Bardin, C. W., and Catterall, J. F. (1981). Testosterone: a major determinant of extragenital sexual dimorphism. *Science* 211, 1285–1294. doi: 10.1126/science.7010603

Berridge, M. J., Lipp, P., and Bootman, M. D. (2000). The versatility and universality of calcium signalling. *Nat. Rev. Mol. Cell Biol.* 1, 11–21. doi: 10.1038/35036035

Bonduriansky, R. (2007). The evolution of condition-dependent sexual dimorphism. *Am. Nat.* 169, 9–19.

Bootman, M. D., Collins, T. J., Peppiatt, C. M., Prothero, L. S., MacKenzie, L., De Smet, P., et al. (2001). Calcium signalling-an overview. *Semin. Cell Dev. Biol.* 12, 3–10. doi: 10.1006/scdb.2000.0211

Boulton, R. A., Collins, L. A., and Shuker, D. M. (2015). Beyond sex allocation: the role of mating systems in sexual selection in parasitoid wasps. *Biol. Rev.* 90, 599–627. doi: 10.1111/brv.12126

Breedlove, S. M. (1992). Sexual dimorphism in the vertebrate nervous-system. *J. Neurosci.* 12, 4133–4142. doi: 10.1523/JNEUROSCI.12-11-04133.1992

Cameron, R. C., Duncan, E. J., and Dearden, P. K. (2013). Biased gene expression in early honeybee larval development. *BMC Genomics* 14:903. doi: 10.1186/1471-2164-14-903

Caterina, M. J. (2007). Transient receptor potential ion channels as participants in thermosensation and thermoregulation. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 292, R64–R76. doi: 10.1152/ajpregu.00446.2006

Chang, P. L., Dunham, J. P., Nuzhdin, S. V., and Arbeitman, M. N. (2011). Somatic sex specific transcriptome differences in Drosophila revealed by whole transcriptome sequencing. *BMC Genomics* 12:364. doi: 10.1186/1471-2164-12-364

Charnov, E. L. (1982). *The Theory of Sex Allocation*. Princeton: Princeton University Press.

Connallon, T., and Knowles, L. L. (2005). Intergenomic conflict revealed by patterns of sex-biased gene expression. *Trends Genet.* 21, 495–499. doi: 10.1016/j.tig.2005.07.006

Cook, J. M. (1993). Sex determination in the hymenoptera-a review of models and evidence. *Heredity* 71, 421–435. doi: 10.1038/hdy.1993.157

Darwin, C. R. (1871). *The Descent of Man, and Selection in Relation to Sex*, 2nd Edn. London: John Murray.

Eads, B. D., Colbourne, J. K., Bohuski, E., and Andrews, J. (2007). Profiling sex-biased gene expression during parthenogenetic reproduction in *Daphnia pulex*. *BMC Genomics* 8:464. doi: 10.1186/1471-2164-8-464

Edwards, J. P., Bell, H. A., Audsley, N., Marris, G. C., Kirkbride-Smith, A., Bryning, G., et al. (2006). The ectoparasitic wasp *Eldophus pennicornis* (Hymenoptera: Eulophiclae) uses instar-specific endocrine disruption strategies to suppress the development of its host *Lacanobia oleracea* (Lepidoptera: Noctuidae). *J. Insect Physiol.* 52, 1153–1162. doi: 10.1016/j.jinsphys.2006.08.003

Ellegren, H., and Parsch, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev. Genet.* 8:689. doi: 10.1038/nrg2167

Fuller, M. T. (1993). "Spermatogenesis," in *The Development of Drosophila*, eds M. Bate and A. Martinez-Arias (Cold Sping Harbor, NY: Cold Sping Harbor Laboratory Press), 71–148.

Gees, M., Colsoul, B., and Nilius, B. (2010). The role of transient receptor potential cation channels in Ca2+ Signaling. *Cold Spring Harb. Perspect. Biol.* 2:a003962. doi: 10.1101/cshperspect.a003962

Godfray, H. C. J. (1994). *Parasitoids: Behavioural and Evolutionary Ecology*. Princeton: Princeton University Press.

Graaf, D. C. D., Aerts, M., Brunain, M., Desjardins, C. A., Jacobs, F. J., Werren, J. H., et al. (2010). Insights into the venom composition of the ectoparasitoid wasp *Nasonia vitripennis* from bioinformatic and proteomic studies. (special issue: the *Nasonia* genome.). *Insect Mol. Biol.* 19, 11–26. doi: 10.1111/j.1365-2583.2009.00914.x

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Habu, A. (1960). A revision of the Chalcididae (Hymenoptera) of Japan, with descriptions of sixteen new species. *Bull. Natl. Inst. Agric. Sci.* 11, 131–363.

Hahn, M. W., and Lanzaro, G. C. (2005). Female-biased gene expression in the malaria mosquito *Anopheles gambiae*. *Curr. Biol.* 15, 192–193. doi: 10.1016/j.cub.2005.03.005

Hamilton, W. D. (1967). Extraordinary sex ratios. *Science* 156, 477–488. doi: 10.1126/science.156.3774.477

Hassan, S. A. (1993). The mass rearing and utilization of *Trichogramma* to control lepidopterous pests: achievements and outlook. *Pest Manage. Sci.* 37, 387–391. doi: 10.1002/ps.2780370412

Heimpel, G. E., and de Boer, J. G. (2008). Sex determination in the Hymenoptera. *Ann. Rev. Entomol.* 53, 209–230. doi: 10.1146/annurev.ento.53.103106.093441

Hunt, B. G., and Goodisman, M. A. (2010). Evolutionary variation in gene expression is associated with dimorphism in eusocial *vespid wasps*. *Insect Mol. Biol.* 19, 641–652. doi: 10.1111/j.1365-2583.2010.01021.x

Jayakumar, A., Chirala, S. S., Chinault, A. C., Baldini, A., Abu-Elheiga, L., and Wakil, S. J. (1994). Isolation and chromosomal mapping of genomic clones encoding the human fatty acid synthase gene. *Genomics* 23, 420–424. doi: 10.1006/geno.1994.1518

Jayakumar, A., Tai, M. H., Huang, W. Y., Al-Feel, W., Hsu, M., Abu-Elheiga, L., et al. (1995). Human fatty acid synthase: properties and molecular cloning. *Proc. Natl. Acad. Sci. U.S.A.* 92, 8695–8699. doi: 10.1073/pnas.92.19.8695

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passadorgurgel, G., and Gibson, G. (2001). The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.* 29:389. doi: 10.1038/ng766

Kang, X. X., Chen, J., Wang, C. C., and Yang, Y. Z. (2006). Identification and behaviors of parasitoids of *Sylepta derogata* in the Yangtze River and Huihe Valley. *Chin. Bull. Entomol.* 35, 241–245.

Karashima, Y., Talavera, K., Everaerts, W., Janssens, A., Kwan, K. Y., Vennekens, R., et al. (2009). Trpa1 acts as a cold sensor in vitro and in vivo. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1273–1278. doi: 10.1073/pnas.0808487106

Kryukova, N., Dubovskiy, I., Chertkova, E., Vorontsova, Y., Slepneva, I., and Glupov, V. (2011). The effect of Habrobracon hebetor venom on the activity of the prophenoloxidase system, the generation of reactive oxygen species and encapsulation in the haemolymph of *Galleria mellonella* larvae. *J. Insect Physiol.* 57, 769–800. doi: 10.1016/j.jinsphys.2011.03.008

Kucharski, R., Maleszka, J., Foret, S., and Maleszka, R. (2008). Nutritional control of reproductive status in honeybees via DNA methylation. *Science* 319, 1827–1830. doi: 10.1126/science.1153069

Leung, H. T., Geng, C., and Pak, W. L. (2000). Phenotypes of trpl mutants and interactions between the transient receptor potential (TRP) and TRP-like

channels in Drosophila. *J. Neurosci.* 20, 6797–6803. doi: 10.1523/JNEUROSCI.20-18-06797.2000

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12:323. doi: 10.1186/1471-2105-12-323

Li, L. (1994). "Worldwide use of Trichogramma for biological control on different crops: a survey," in *Biological Control with Egg Parasitoids*, eds E. Wajnberg and S. A. Hassan (Wallingford: Cab International).

Li, L., Jiang, Y., Liu, Z., You, L., Wu, Y., Xu, B., et al. (2016). Jinggangmycin increases fecundity of the brown planthopper, *Nilaparvata lugens* (Stål) via fatty acid synthase gene expression. *J. Proteomics* 130, 140–149. doi: 10.1016/j.jprot.2015.09.022

Lim, J. O., Lyu, D. P., Choi, G. S., Jeong, Y. J., Shin, S. C., and Lee, S. H. (2006). A taxonomie note on *Sclerodermas harmandi*, ectoparasite of stem and wood boring insect larvae (Hymenoptera: Chrysidoidea'-Bethylidae) in South Korea. *J. Asia Pac. Entomol.* 9, 115–119. doi: 10.1016/S1226-8615(08)60282-4

Lin, H., Mann, K. J., Starostina, E., Kinser, R. D., and Pikielny, C. W. (2005). A Drosophila DEG/ENaC channel subunit is required for male response to female pheromones. *Proc. Natl. Acad. Sci. U.S.A.* 102, 12831–12836. doi: 10.1073/pnas.0506420102

Lipinska, A., Cormier, A., Luthringer, R., Peters, A. F., Corre, E., Gachon, C. M., et al. (2015). Sexual dimorphism and the evolution of sex-biased gene expression in the brown alga ectocarpus. *Mol. Biol. Evol.* 32, 1581–1597. doi: 10.1093/molbev/msv049

Lyko, F., Foret, S., Kucharski, R., Wolf, S., Falckenhayn, C., and Maleszka, R. (2010). The honey bee epigenomes: differential methylation of brain DNA in queens and workers. *PLoS Biol.* 8:e1000506. doi: 10.1371/journal.pbio.1000506

Mao, H., and Kunimi, Y. (1991). Pupal mortality of the oriental tea tortrix, *Homona magnanima* Diakonoff (Lepidoptera: Tortricidae), caused by parasitoids and pathogens. *Jpn. J. Appl. Entomol. Zool.* 35, 241–245. doi: 10.1303/jjaez.35.241

Mao, H., and Kunimi, Y. (1994a). Effects of temperature on the development and parasitism of *Brachymeria lasus*, a pupal parasitoid of *Homona magnanima*. *Entomol. Exp. Appl.* 71, 87–90. doi: 10.1111/j.1570-7458.1994.tb01773.x

Mao, H., and Kunimi, Y. (1994b). Longevity and fecundity of *Brachymeria lasus* (Walker) (Hymenoptera: Chalcididae), a pupal parasitoid of the Oriental tea tortrix, *Homona magnanima* Diakonoff (Lepidoptera: Tortricidae) under laboratory conditions. *Appl. Entomol. Zool.* 29, 237–243. doi: 10.1303/aez.k29.237

Mao, X., Cai, T., Olyarchuk, J. G., and Wei, L. (2005). Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 21, 3787–3793. doi: 10.1093/bioinformatics/bti430

Marinotti, O., Calvo, E., Nguyen, Q. K., Dissanayake, S., Ribeiro, J. M., and James, A. A. (2006). Genome-wide analysis of gene expression in adult *Anopheles gambiae*. *Insect Mol. Biol.* 15, 1–12. doi: 10.1111/j.1365-2583.2006.00610.x

Mayhew, P. J. (1998). The life-histories of parasitoid wasps developing in small gregarious broods. *Neth. J. Zool.* 48, 225–240. doi: 10.1163/156854298X00084

Mohamed, M. A., and Coppel, H. C. (1987). Pheromonal basis for aggregation behavior of parasitoids of the gypsy moth: *Brachymeria intermedia*, (Nees) and *Brachymeria lasus*, (Walker) (Hymenoptera: Chalcididae). *J. Chem. Ecol.* 13, 1385–1393. doi: 10.1007/BF01012285

Moreau, S. J. M., and Asgari, S. (2015). Venom proteins from parasitoid wasps and their biological functions. *Toxins* 7, 2385–2412. doi: 10.3390/toxins7072385

Parisi, M., Nuttall, R., Edwards, P., Minor, J., Naiman, D., Lü, J., et al. (2004). A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. *Genome Biol.* 5:R40. doi: 10.1186/gb-2004-5-6-r40

Parra, J. R. P., and Zucchi, A. R. (2004). Trichogramma in Brazil: feasibility of use after twenty years of research. *Neotrop. Entomol.* 33, 271–281. doi: 10.1590/S1519-566X2004000300001

Persson, B., Bray, J. E., Bruford, E., Dellaporta, S. L., Favia, A. D., Duarte, R. G., et al. (2008). The sdr (short-chain dehydrogenase/reductase and related enzymes) nomenclature initiative. *Chem. Biol. Interact.* 178, 94–98. doi: 10.1016/j.cbi.2008.10.040

Pointer, M. A., Harrison, P. W., Wright, A. E., and Mank, J. E. (2013). Masculinization of gene expression is associated with exaggeration of male sexual dimorphism. *PLoS Genet.* 9:e1003697. doi: 10.1371/journal.pgen.1003697

Popescu, D. C., Ham, A. J., and Shieh, B. H. (2006). Scaffolding protein INAD regulates deactivation of vision by promoting phosphorylation of transient receptor potential by eye protein kinase C in *Drosophila*. *J. Neurosci.* 26, 8570–8577. doi: 10.1523/JNEUROSCI.1478-06.2006

Price, D., Bell, H., Hinchliffe, G., Fitches, E., Weaver, R., and Gatehouse, J. A. (2009). Venom metalloproteinase from the parasitic wasp *Eulophus pennicornis* is toxic towards its host, tomato moth (*Lacanobia oleracae*). *Insect Mol. Biol.* 18, 195–202. doi: 10.1111/j.1365-2583.2009.00864.x

Prince, E. G., Kirkland, D., and Demuth, J. P. (2010). Hyperexpression of the X chromosome in both sexes results in extensive female bias of X-linked genes in the flour beetle. *Genome Biol. Evol.* 2, 336–346. doi: 10.1093/gbe/evq024

Ranz, J., Castillo-Davis, C., Meiklejohn, C., and Hartl, D. (2003). Sex-dependent gene expression and evolution of the Drosophila transcriptome. *Science* 300, 1742–1745. doi: 10.1126/science.1085881

Rinn, J. L., and Snyder, M. (2005). Sexual dimorphism in mammalian gene expression. *Trends Genet.* 21, 298–305. doi: 10.1016/j.tig.2005.03.005

Simser, D. H., and Coppel, H. C. (1980). Female-produced sex pheromone in *Brachymeria lasus* and *B. intermedia* (Hym.: Chalcididae). *BioControl* 25, 373–380.

Spradling, A. C. (1993). "Developmental genetics of oogenesis," in *The Development of Drosophila*, eds M. Bate and A. Martinez-Arias (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press), 1–70.

Terayama, M. (1999). "Description of new species of the family Bethylidae from the Ryukyus, and taxonomic notes on the Japanese species of the genus Sclerodermus," in *Identification guide to the Aculeata of the Nansei Islands*, eds Y. Seiki, S. Ikudome, and M. Terayama (Sapporo: Hokkaido University Press).

Tian, C., Wang, L., Ye, G., and Zhu, S. (2010). Inhibition of melanization by a Nasonia defensin-like peptide: implications for host immune suppression. *J. Insect Physiol.* 56, 1857–1862. doi: 10.1016/j.jinsphys.2010.08.004

Wang, T., and Montell, C. (2007). Phototransduction and retinal degeneration in *Drosophila*. *Pflügers Arch. Eur. J. Physiol.* 454, 821–847. doi: 10.1007/s00424-007-0251-1

Wang, X., Werren, J. H., and Clark, A. G. (2015). Genetic and epigenetic architecture of sex-biased expression in the jewel wasps *Nasonia vitripennis* and *giraulti*. *Proc. Natl. Acad. Sci. U.S.A.* 112, E3545–E3554. doi: 10.1073/pnas.1510338112

Wen, X., Guo, L., Jiao, X., Yang, N., Xin, Y., Wu, Q., et al. (2014). Transcriptomic dissection of sexual differences in *Bemisia tabaci*, an invasive agricultural pest worldwide. *Sci. Rep.* 4:4088. doi: 10.1038/srep04088

Werren, J. H., Richards, S., Desjardins, C. A., Niehuis, O., Gadau, J., Colbourne, J. K., et al. (2010). Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* species. *Science* 327, 343–348. doi: 10.1126/science.1178028

Wolff, T., and Ready, D. (1993). "Pattern formation in the Drosophila retina," in *The Development of Drosophila melanogaster*, eds M. Bate and A. M. Arias (Plainview, NY: Cold Spring Harbor Lab. Press), 1277.

Yan, J. J., Xu, C. H., Li, G. W., Zhang, P. Y., Gao, W. C., Yao, D. F., et al. (1989). *Parasites and Predators of Forest Pest*. Beijing: China Forestry Publishing House.

Yang, X. Q., Wei, J. R., and Yang, Z. Q. (2001). A survey on insect natural enemies of Hyphantriacunea in Dalian district, Liaoning Province. *Chin. J. Biol. Control* 17, 40–42.

Young, M. D., Wakefield, M. J., Smyth, G. K., and Oshlack, A. (2010). Gene ontology analysis for rna-seq: accounting for selection bias. *Genome Biol.* 11:R14. doi: 10.1186/gb-2010-11-2-r14

Zhishan, W., Hopper, K. R., Ode, P. J., Fuester, R. W., Jia-Hua, C., and Heimpel, G. E. (2003). Complementary sex determination in Hymenopteran parasitoids and its implications for biological control. *Entomol. Sin.* 10, 81–93. doi: 10.1111/j.1744-7917.2003.tb00369.x

Zuker, C. S. (1996). The biology of vision in *Drosophila*. *Proc. Natl. Acad. Sci. U.S.A.* 93, 571–576. doi: 10.1073/pnas.93.2.571

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 12 March 2019
doi: 10.3389/fgene.2019.00194

Utility of cfDNA Fragmentation Patterns in Designing the Liquid Biopsy Profiling Panels to Improve Their Sensitivity

Maxim Ivanov[1], Polina Chernenko[2], Valery Breder[2], Konstantin Laktionov[2], Ekaterina Rozhavskaya[3,4], Sergey Musienko[3], Ancha Baranova[1,3,5,6] and Vladislav Mileyko[3]*

[1] Department of Biological and Medical Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Russia, [2] N.N. Blokhin Russian Cancer Research Center, Moscow, Russia, [3] Atlas Oncology Diagnostics, Ltd., Moscow, Russia, [4] Vavilov Institute of General Genetics, Moscow, Russia, [5] Research Centre for Medical Genetics, Moscow, Russia, [6] School of Systems Biology, George Mason University, Fairfax, VA, United States

OPEN ACCESS

Edited by:
Richard D. Emes, University of Nottingham, United Kingdom

Reviewed by:
Vladimir B. Teif, University of Essex, United Kingdom
Kuo-Ping Chiu, Academia Sinica, Taiwan
Tatiana V. Tatarinova, University of La Verne, United States

***Correspondence:**
Maxim Ivanov
maksim.v.ivanov@phystech.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *19 October 2018*
Accepted: *25 February 2019*
Published: *12 March 2019*

Citation:
Ivanov M, Chernenko P, Breder V, Laktionov K, Rozhavskaya E, Musienko S, Baranova A and Mileyko V (2019) Utility of cfDNA Fragmentation Patterns in Designing the Liquid Biopsy Profiling Panels to Improve Their Sensitivity. Front. Genet. 10:194. doi: 10.3389/fgene.2019.00194

Genotyping of cell-free DNA (cfDNA) in plasma samples has the potential to allow for a noninvasive assessment of tumor biology, avoiding the inherent shortcomings of tissue biopsy. Next generation sequencing (NGS), a leading technology for liquid biopsy analysis, continues to be hurdled with several major issues with cfDNA samples, including low cfDNA concentration and high fragmentation. In this study, by employing Ion Torrent PGM semiconductor technology, we performed a comparison between two multi-biomarker amplicon-based NGS panels characterized by a substantial difference in average amplicon length. In course of the analysis of the peripheral blood from 13 diagnostic non-small cell lung cancer patients, equivalence of two panels, in terms of overall diagnostic sensitivity and specificity was shown. A pairwise comparison of the allele frequencies for the same somatic variants obtained from the pairs of panel-specific amplicons, demonstrated an identical analytical sensitivity in range of 140 to 170 bp amplicons in size. Further regression analysis between amplicon length and its coverage, illustrated that NGS sequencing of plasma cfDNA equally tolerates amplicons with lengths in the range of 120 to 170 bp. To increase the sensitivity of mutation detection in cfDNA, we performed a computational analysis of the features associated with genome-wide nucleosome maps, evident from the data on the prevalence of cfDNA fragments of certain sizes and their fragmentation patterns. By leveraging the support vector machine-based machine learning approach, we showed that a combination of nucleosome map associated features with GC content, results in the increased accuracy of prediction of high inter-sample sequencing coverage variation (areas under the receiver operating curve: 0.75, 95% CI: 0.750–0.752 vs. 0.65, 95% CI: 0.63–0.67). Thus, nucleosome-guided fragmentation should be utilized as a guide to design amplicon-based NGS panels for the genotyping of cfDNA samples.

Keywords: NGS, cfDNA, liquid biopsy, cancer, DNA fragmentation, nucleosome, amplicon, primer design

INTRODUCTION

In an approach known as "liquid biopsy," cell-free DNA (cfDNA) which circulates in the plasma may be used for a diagnostic detection of tumor-specific mutations (Dawson et al., 2013; Pupilli et al., 2013; Xi et al., 2016). In the frame of the Lab-Developed Tests (LDT) paradigm, analysis of cfDNA has already gained approval for a number of common indications, including the detection of the resistance mutation T790M in the EGFR encoding gene (Malapelle et al., 2016), which commonly emerges in lung adenocarcinomas treated with tyrosine kinase inhibitors.

At their inception, cfDNA-based LDTs commonly exploited one or another conventional DNA analysis technique, including real-time PCR, droplet digital PCR and beads, emulsions, amplification, and magnetics (BEAM)ing digital PCR (Dawson et al., 2013; Oxnard et al., 2014; Siravegna et al., 2015; Thress et al., 2015; Sacher et al., 2016). Many studies showed that the concordance of liquid biopsy and tissue-based analysis is relatively high; nevertheless, these approaches are not free of limitations. Typically, PCR-based and hybridization-based cfDNA profiling techniques are developed to detect particular DNA variants, which most commonly underlie one or another previously described pathophysiological process. These and other variant-specific techniques are not suitable for the exploratory analysis of cfDNA, which is necessary for acquisition of knowledge concerning non-conventional, emerging resistance pathways, for co-detection of the mismatch repair phenotype, and for off-label prescribing of anticancer medications commonly required for personalized treatment of metastatic tumors (Tafe et al., 2015; Wei et al., 2016; Zehir et al., 2017). These limitations are readily surmounted by an advent of sequencing-based technologies, including whole exome sequencing or, more applicable to cfDNA analysis, amplicon-based panels, which are limited to their target genes, but are still exploration-permissive.

With reported sensitivity and a specificity of more than 80%, and 98 to 100%, respectively (Krishnamurthy et al., 2017), a next generation sequencing (NGS) analysis of cfDNA has already inserted itself into the ranks of the commonly used LDTs. Nevertheless, further improvement of the sensitivity in liquid biopsy-based tests is warranted. The most common way to improve sensitivity of the mutation detection in liquid biopsy samples, is to increase the coverage, which in turn leads to a substantial increase in the cost of an assay. Deep or ultradeep coverage is necessary in order to account for low concentrations of total cfDNA in plasma samples that are compounded by the dilution of tumor-specific cfDNA fragments, by substantial amounts of non-tumoral cfDNA fragments (Hellwig et al., 2018).

Another physical characteristic of cfDNA, the distribution of the sizes of its fragments, is relevant to the detection of DNA variants both by sequencing and by PCR. Recent whole-genome sequencing (WGS) studies of cfDNA demonstrated that the distribution of the sizes of plasma derived DNA fragments is far from the typical lognormal distribution that reflects the patterning of DNA in formalin fixed-paraffin-embedded samples or snap-frozen tissues. In fact, cfDNA exhibits a predominant peak at a fragment length of ∼167 bp accompanied by the second, significantly less pronounced extremum at around 350 bp (Ma et al., 2017). These observations mean that the majority of these fragments are suitable to assess the technique that relies on conventional lengths of PCR amplicons. It is of note that tumor-derived cfDNA fragments are even shorter than those that originate from healthy cells of the same origin (Jiang et al., 2015). In the domain of conventional systems for the detection of DNA variants, these characteristic of cfDNA have prompted the development of ultra-short amplicon PCR, which allows for the substantial increase of analytical and, as a consequence, diagnostic sensitivity of these assays.

Moreover, recent studies have shown that fragmentation pattern of cfDNA is not random. As cfDNA degradation is guided by nucleosome patterns defined by epigenetic regulation within particular loci (Ivanov et al., 2015), recurrent underrepresentation of some regions in cfDNA introduces systematic bias in the PCR based enrichment of target amplicons and undermine the sensitivity at a local scale.

In this study, we investigated the effect of the amplicon length on the diagnostic and analytical sensitivity of mutation detection, using two amplicon-based NGS panels with diverse amplicon lengths. We also describe ways to utilize the knowledge of cfDNA fragmentation patterns to increase the sensitivity of mutation detection in a liquid biopsy setting.

MATERIALS AND METHODS

Sample Collection

The sequencing was performed on cfDNA fragments extracted from previously collected plasma samples of 13 non-small cell lung cancer (NSCLC) patients, treated at the Blokhin Russian Cancer Research Centre in 2014 to 2015. For each patient, tumor tissue-based EGFR mutation status was assessed using the therascreen EGFR RGQ PCR Kit (Qiagen, Milan, Italy) according to the manufacturers protocol.

For nucleosome-guided cfDNA fragmentation pattern analysis we used publicly available, anonymized WGS data of cfDNA, described by Snyder et al. (2016) and included in dataset [PRJNA291063].

The present study was approved by the Atlas Biomed Internal Review Board. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

DNA Extraction and Sample Quality Control

For each NSCLC patient, a peripheral blood sample was collected into an EDTA-containing vacutainer tube (BD). Samples were fractionated into plasma and blood cells by centrifugation at 400 *g* for 15 min within 4 h after venipuncture, followed by a secondary spin at 1200 *g* for 20 min. Resultant plasma samples were frozen in aliquots and stored at −80°C until DNA isolation. Circulating DNA was extracted from 4 ml of plasma using the Blood Plasma DNA Isolation Kit (BioSilica Ltd., Russia) according to the manufacturer's instructions, eluted by 120 μl of nuclease-free water, mixed with 3 μl of glycogen (20 mg/ml, Fermentas, Lithuania), 1/10 volume of 50 mM triethylamine

and then precipitated with 5 volumes of acetone (Bryzgunova et al., 2011). After reconstitution in 30–50 μl of water, cfDNA concentrations were measured using the Qubit fluorometer.

Library Preparation and Quality Control

Sequencing libraries were prepared according to the manufacturer's protocol for Ion AmpliSeq Cancer Hotspot Panel (ITCHP2), designed to amplify 207 target regions across 50 cancer-related genes. Additionally, a custom panel namely Atlas Clinical Panel (AODCP), was designed to cover the following genes: EGFR, IDH2, NRAS, KIT, BRAF, TP53, PDGFRA, PTEN, IDH1, KRAS, PIK3CA, ERBB2, CTNNB1 (AODCP, 55 target regions). The custom panel was designed using the Ion AmpliSeq Designer server (pipeline version 5.2). The two panels had several loci in common, allowing for their comparison.

Sequencing and Data Analysis

Pooled libraries were sequenced employing Ion Torrent PGM, according to the manufacturers protocol. As low frequency mutant alleles were expected, initial analysis was performed using Ion Torrent Suite software (version 5.2.0) on low stringency settings. In order to exclude false negative single nucleotide variant (SNV) calls, concomitant Bowtie2-Strelka pipeline analysis was carried out. After aligning all reads to the genome (GRCh37) (Bowtie2 parameters: –rdg 5,2 –rfg 5,2 -N 1 -L 17), further off-target reads were removed, while the remaining reads were realigned on target sequences. Primer sequences were excluded from reads employing in-house software (Ivanov et al., 2018). Somatic variant calling was performed employing Strelka (maxInputDepth set to −1; indelMaxRefRepeat set to 6; indelMaxWindowFilteredBasecallFrac set to 0.4; indelMaxIntHpolLength set to 6; lower quality bound for SNV and indels set to 9 and 2, respectively). Variants supported with less than 20 reads in total were discarded. If less than four reads supported alternative allele, the variant was omitted. Mutation hotspots were defined as nucleotide variations identified in ten or more COSMIC (Forbes et al., 2010) samples. Detected variants located within mutation hotspots were supposed to be confidently somatic. Variants outside mutation hotspots with minor allele frequency in the general population, as defined by 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015), of 5% and more were supposed to be confidently germline. Further analysis was limited to confidently somatic and confidently germline variants. Preprocessed fastq files were additionally screened for mutation hotspots by inputting wild type and expected mutant reads into the Poisson distribution statistical model with complexity-dependent variable expectation probability of SNVs and indels. Somatic variant calls were verified manually, in the Tablet (version 1.16.09.06) read alignment visualization tool (Milne et al., 2010). Variant allele frequencies were quantified within raw read sets as a ratio of reads confirming the mutation to the total count of qualified reads covering the mutation site. Normalization of mutation allele frequencies to amplicon coverage was performed by bootstrapping. The genome variation analysis was limited to the nucleotide changes affecting the protein sequence, unless otherwise specified. Publicly available software and database versions used were Bowtie2 v. 2.1.0 (Langmead and Salzberg, 2012), Strelka v. 1.0.14 (Saunders et al., 2012), and SAMtools v. 0.1.19 (Li, 2011). COSMIC and dbSNP databases were assessed in December 2017.

GC content normalization for linear regression analysis was performed leveraging a simple adjustment according to the equation $\widetilde{r_i} = r_i {}^{m}/_{m_{\mathrm{GC}}}$, where r_i stands for the read count of the ith amplicon, m_{GC} is the median read count of all windows with the same GC content as the ith amplicon, and m is the overall median of all the amplicons. Deviation of coverage from the mean was performed for 5% GC content bands rather than percentages of 0, 1, 2, 3, . . ., 100%. Linear regression analysis was performed employing simple least square fitting.

Nucleosome-guided cfDNA fragmentation patterns were analyzed in publicly available sequences obtained from plasma samples pooled from an unknown number of healthy individuals (GSM1833219). The details of the DNA extraction, library preparation and sequencing are provided in Snyder et al. (2016). Briefly, cfDNA libraries underwent paired-end sequencing with Illumina sequence-by-synthesis technology generating reads of 101 bp in size. Importantly, at the library preparation stage, plasma DNA samples did not undergo fragmentation by sonication and, thus, original cfDNA molecules were preserved, granting the opportunity to investigate its fragmentation patterns. The fastq read sequences were aligned to the human genome (aforementioned reference build) with BWA-mem v. 0.7.12 (Li and Durbin, 2009). cfDNA fragment length may exceed sequencing read length, however, paired-end sequencing allows to capture both start and end positions of the fragment. Paired reads, thus, continued to represent WGS fragments. Nucleosome position stringencies were calculated essentially as described in Valouev et al., using the NuMap software with standard parameters. NuMap performs the nucleosome mapping based on the kernel smoothed reads count calculation (Valouev et al., 2011).

For ITCHP2 and AODCP panel amplicons, fragment counts were generated *in silico* after matching both primers with the fragment amplified and sequenced experimentally. To understand the patterns of amplicon coverage by experimentally observed fragments, the fragments were generated using paired reads, then further filtered by length to include only fragments in the range of 80 to 250 bp. Dinucleosome fragments were therefore excluded. To improve resolution, resulting fragments were trimmed by 40 bp around dyads to generate a set of equal-length fragments. For each sequenced nucleotide position, counts of overlapping fragments were recorded. Generated data were subjected to a lowpass filter with the square pulse kernel with the width of 21 base pairs, then resulting coverage plots were mapped to amplicons genome positions.

Statistical analysis was performed using R, version 3.2.3. For machine learning, we used the open source library Orange (Demsar et al., 2013). Five machine learning algorithms were evaluated to find the best model, demonstrating the highest prediction accuracy based on all descriptors [support vector machine (SVM), neural network, multiple linear regression, naïve Bayes, and random forest].

RESULTS

Sample Sequencing and Mutation Analysis

In this study, fourteen cfDNA samples collected from patients with NSCLC, were analyzed using the screening panels ITCHP2 and AODCP. The mean sequencing coverages across all experiments were set at 1150× for the AODCP panel and 802× for the ITCHP2 panel with corresponding medians of 1002× and 674×, respectively.

Variant detection results were completely concordant for two panels across 18 identified somatic mutations. Plasma variant detection results were concordant with baseline tissue analysis in 9 samples (69%). False negative samples were limited to the cases, characterized with low plasma DNA concentration (**Figure 1**). In addition to mutations identified by tissue analysis at baseline, namely, these in EGFR and RAS, the sequencing of 13 plasma cfDNA samples revealed five additional somatic missense mutations, including these in PIK3CA and TP53 genes (**Figure 1**).

Significance of Amplicon Length for Mutation Detection Sensitivity and Specificity

The average length of amplicons in panel AODCP was much shorter than that in panel ITCHP2 (**Figure 2A**), with median amplicon lengths to include primer sequences at 137 and 156 bp, respectively. Despite the difference in amplicon sizes, variant calling results obtained for each panel were completely concordant, with a total of 51 either somatic or germline variants detected. Therefore, diagnostic sensitivity and specificity of these two detection systems were the same at the study power.

In order to explore possible influences of the amplicon length on the limits of detection and, therefore, analytical sensitivity to the presence of the mutations in liquid biopsy, we performed a pairwise comparison of the frequencies for same mutated allele in reads obtained from pairs of panel-specific amplicons. For the synonymous germline variant, namely, EGFR p.Gln787= with the total of 15 alleles identified (1000 Genomes MAF 0.43), allele frequencies extracted from analysis of AODCP and ITCHP2 amplicons were equivalent (Wilcoxon signed rank test p-value = 0.88). On the other hand, analysis of somatic mutations, which are typically present in a relatively small fraction of the reads, showed Pearson's correlation coefficients of 0.88 (p-value = 0.02; Wilcoxon signed rank test p-value = 0.44) for point mutations in genes *EGFR*, *TP53*, and *PIK3CA*, and 0.95 for the deletions of the *EGFR* exon 19 (p-value = 0.001; Wilcoxon signed rank test p-value = 0.53) (**Figure 3**). Since *EGFR* deletions further reduce the length of amplified fragments by 15 or more bp, their presence should, at least in theory, increase analytical sensitivity of the detection system (**Figure 2B**). Notably, the geometric mean ratio of the allele frequency of the EGFR exon 19 deletions, detected by two panels, was 1.16 (95% CI, 0.72–1.88; p-value > 0.1). This indicates that the analytical sensitivity of this assay is unlikely to change even if the difference in the average sizes of amplicons would increase further.

Finally, we performed a regression analysis to estimate the relationship between amplicon length and its average coverage across samples for the ITCHP2 panel, representing a wider spectrum of amplicon lengths. After normalization on GC-content and overall sample read count, linear regression analysis employing the least squares fitting approach, demonstrated a negative slope with a Student t-test p-value of 0.0063. However, regression analysis across the set of amplicons with a length of 170 bp or less yielded a non-significant slope coefficient (p-value 0.69) (**Figures 4A,B**). Regression analysis between amplicon length and its coverage covariance demonstrated no significant correlation in any amplicon length range (data not shown). Considering that amplicons with a length of 120 or less comprises of only 5% of that set, this indicates that the NGS sequencing of plasma cfDNA equally tolerates amplicons with a length in the 120–170 bp range.

Nucleosome-Guided Pattern May Facilitate Primer Panel Design

According to the most commonly cited hypothesis, plasma cfDNA originates from apoptotic cells where genomic DNA is digested by a set of nucleases (Ma et al., 2017). Wrapping around nucleosomes protects some of the DNA fragments from digestion; that is why cfDNA fragments correspond primarily to the mononucleosome bound regions. Originally supported only by a unimodal distribution of cfDNA fragments sizes (Fan et al., 2008; Lo et al., 2010), this hypothesis has been recently validated in several studies (Chandrananda et al., 2015; Snyder et al., 2016; Ulz et al., 2016). In particular, employing whole exome sequencing of cfDNA fragments to infer the read depth coverage allowed the construction of 'plasma genome-wide nucleosome maps. Mapping the fragments covered by the ITCHP2 panel, to these nucleosome maps, showed that the positions of the ITCHP2 primers were selected in a non-optimal way with respect to the nucleosome positioning (p-value for nucleosome peaks and amplicons interception 0.36). An amplicon covering KRAS exon 4 serves as a good illustration for non-optimal selection of primers which fall in between two peaks (**Figure 5A**). Because of that, amounts of spanning cfDNA fragments are much lower than for the primers selected to amplify the fragment located within the same peak. A similar situation may be observed for the EGFR exon 21; shifting positions of the primers by the order of 100 nucleotides may result in an increase of the depth and the uniformity of the coverage, without compromising amplification of the clinically relevant, mutation-harboring locus.

At the next stage of analysis, we inquired whether efficiency of targeted resequencing of cfDNA samples depends on the pattern of DNA fragmentation. To perform this analysis, for all amplicons represented in the ITCHP2 panel, the fragmentation patterns were extracted from the repository of reads obtained after a shotgun sequencing of cfDNA fragments purified from the pool of plasma samples, of healthy individuals and from five individual patients with solid tumors (**Figure 5B**).

It is known that both the nucleosome positioning (Struhl and Segal, 2013), which, in turn, guides the fragmentation of cDNA (Ma et al., 2017), and the depth and the uniformity of

Patient Id	Plasma DNA concentration (ng/µl)	Tumor alteration	Plasma EGFR status			Additional mutations identifeid		
			overall	L858R	ex19del	Gene	mutation	allele frequency
PI10	NA	EGFR	negative					
PI84	15,6	EGFR	negative					
PI87	16,7	EGFR	positive		2,6%	PIK3CA	p.His1060Gln	23,60%
PI90	32,8	ALK	positive		10,4%			
PI92	38,5	EGFR	positive	7,7%	2,8%			
PI94	46,3	EGFR	positive	5,9%	0,8%	TP53	p.Arg248Trp	3,2%
PI89	59,1	EGFR	positive		1,2%			
PI93	62,4	EGFR	positive		13,7%			
PI91	64,9	EGFR	positive		90,6%	PIK3CA	p.Glu542Gln	20,00%
						TP53	p.Arg213Leu	34,20%
PI88	84,2	EGFR	positive		0,4%			
PI19	86,1	EGFR	positive		0,6%			
PI29	92,4	EGFR	positive	7,9%		TP53	p.Ser241Phe	3,30%
PI13	169,8	EGFR	positive		4,1%			

FIGURE 1 | Samples used for data analysis as well as mutations identified during NGS sequencing and allele frequencies thereof (plasma EGFR status). Mutations identified employing a conventional sequencing method indicated in the tumor alteration column while its match (green) or mismatch (red) with NGS results specified in plasma EGFR status column.

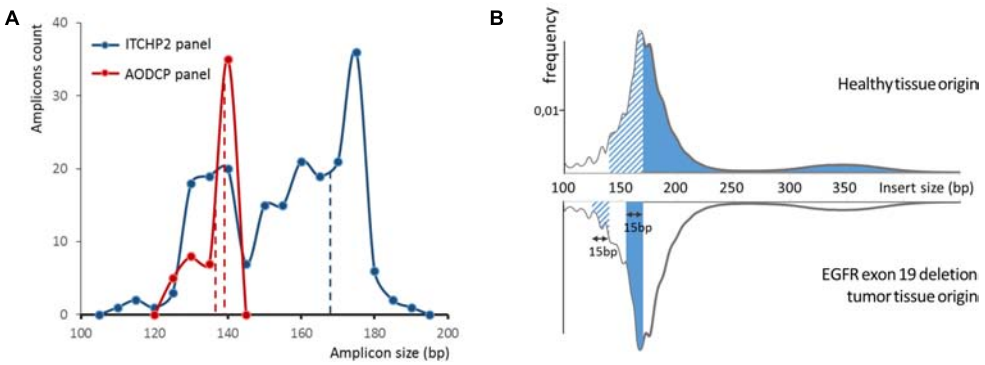


FIGURE 2 | (A) Ion torrent cancer hotspot 2 (ITCHP2) and custom AODCP primer panels amplicon length distribution. A constant window of 5 bp was used to discretize amplicon length. Dotted lines demonstrate length of amplicons, covering exon 19 of the EGFR. **(B)** cfDNA fragment length distribution influence available for the amplification DNA molecules in plasma and, thus, amplification effectiveness. Solid fill at the top panel demonstrates the spectrum of cfDNA fragments involved in EGFR exon 19 PCR amplification employing the ITCHP2 panel. Dashed fill demonstrates the extension of that spectrum in case the AODCP panel is used. Fills in the bottom panel demonstrate the spectrum extension for two panels, respectively, in case of the 15 bp exon 19 deletion mutation.

the coverage by sequencing reads (Benjamini and Speed, 2012), are influenced by the GC content. In the following analysis, we aimed at finding out whether any characteristic related to the fragmentation pattern of cfDNA within the locus of interest may influence the depth and the uniformity of coverage with amplification based sequencing reads.

For the ITCHP2 panel, each amplicon was matched to an individual nucleosome map and evaluated according to four features: *(i)* absolute count of experimentally observed continuous cfDNA fragments spanning the whole amplicon (Feature A), *(ii)* read signal amplitude within the amplicon (Feature B), *(iii)* read signal change at the boundaries of amplicon (Feature C), and *(iv)* read signal shape defined as the area between its linear approximation and itself (Feature D) (**Figure 5C**). Uniformity of the coverage was defined as a coefficient of inter-individual variation in read coverage between all cfDNA samples. To calculate the robustness of the nucleosome mapping, we assessed the inter-sample variance of the defined features calculated for each amplicon. Averaged coefficients of the variation of features D, B and C were at 390, 68, and 38%, respectively, pointing at significant inter-sample variation.

Further, we estimated the feature quality, employing the RReliefF method (Robnik-Sikonja and Kononenko, 2003) estimating how well their values distinguish between target variables that are near to each other. Despite previously demonstrated low robustness of the nucleosome associated features, the count of spanning fragments (Feature A) was ranked even higher than the GC content, while the other three features, B, C, and D, closely followed feature A and the GC content (**Figure 6A**). This finding indicates that uniformity of the locus coverage, with amplified sequencing reads, may depend on the underlying pattern of cfDNA fragmentation.

Univariate polynomial regression of the sequencing coverage depth and its coefficient of variation based on the GC content with second degree polynomial yielded coefficients of determination of 0.29 and 0.19, respectively. Furthermore,

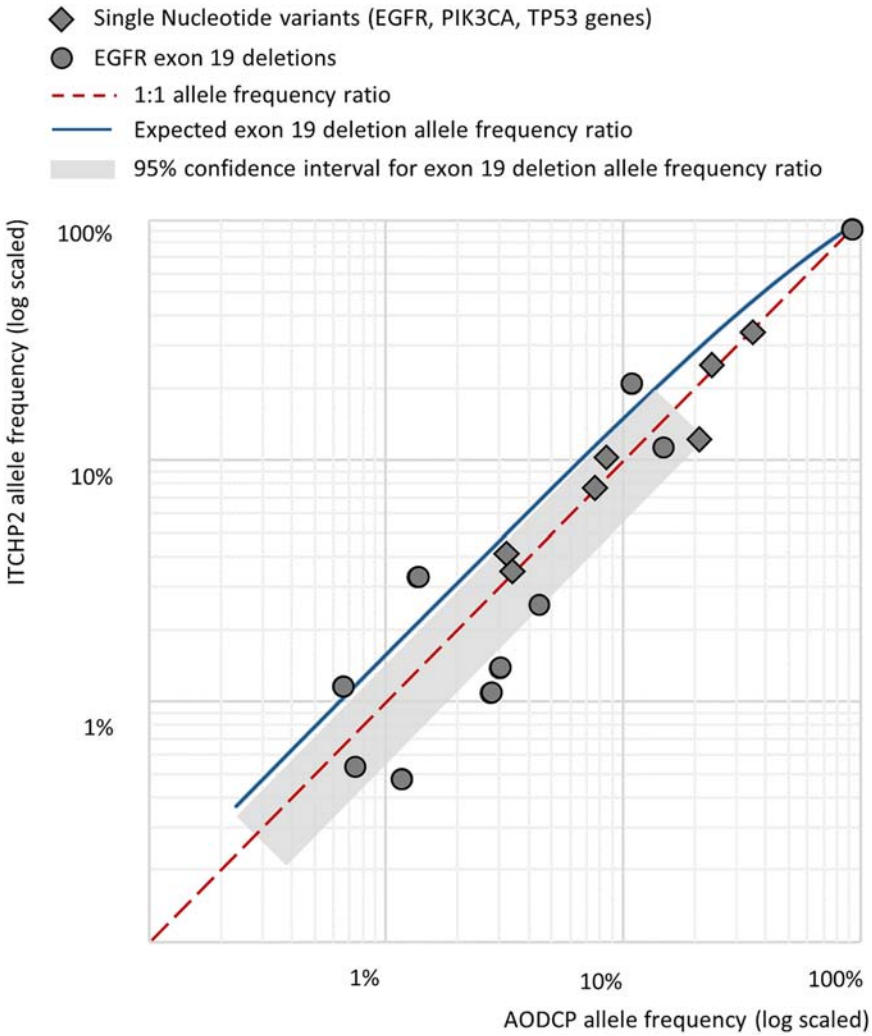
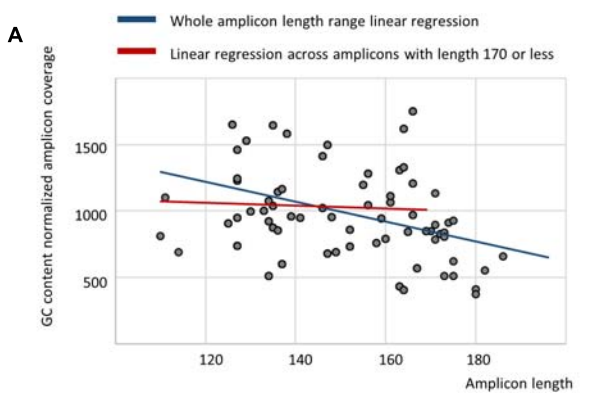


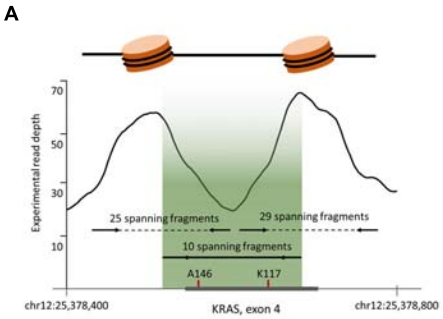
FIGURE 3 | Pairwise comparison of the frequencies for same mutated allele in reads obtained from pairs of panel-specific amplicons across detected somatic variants.



B

Length	Linear regression slope across amplicons with higher length	lower length
120	-7,5	
130	-7,0	22,6
140	-11,5	4
150	-16,3	0,7
160	-28,3	-1,3
170	-28,6	-1,1
180		-3,8

FIGURE 4 | The regression analysis demonstrates no relationship between amplicon length and its average coverage when restricting the analysis to amplicons with a length of 170 or less **(A)**, while, broken stick regression indicates that the system suffers a shock after 170 bp **(B)** (green fill indicates statistically significant slope coefficient with *p*-value of 0.05 or less).



B

Sample Name	Clinical Dx	Sex	Fragments sequenced (millions)
GSM1833219	Healthy, pooled plasma	mixed	1300
GSM1833259	DCIS	female	410
GSM1833261	COAD	female	860
GSM1833242	HCC	male	740
GSM1833240	SSLC	male	650
GSM1833245	LUSC	male	480

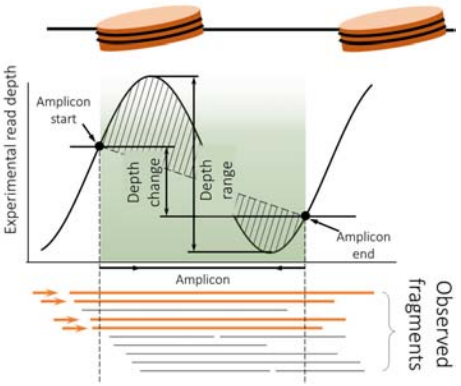


FIGURE 5 | Plasma cfDNA fragmentation pattern biases analytical characteristics of PCR-based somatic detection system. **(A)** ITCHP2 primer panel design mapped to the nucleosome guided cfDNA fragmentation pattern indicates a possible bias in amplification effectiveness. Experimental coverage was assessed based on the pair-end WGS plasma sequencing of healthy individuals (GSM1833219). Fragment counts were calculated as WGS captured and sequenced fragments (continued read pairs), completely covering the amplicon of interest. Localization of the ITCHP2 panel amplicon (solid line with arrows, indicating primers as well as lightened area) in local minimum would result in the lack of availability for the amplification of cfDNA fragments. Shifting amplicons (dotted lines) within clinically relevant mutations may result in an increase of available fragments and thus amplification effectiveness. A146 and K117 indicates clinically relevant KRAS mutation hotspots in the exon of interest. A similar situation can be observed for the EGFR L858R mutation and respective amplicon design – moving amplicons within clinically relevant mutation site mutations may result in increased coverage uniformity. **(B)** Publicly available cfDNA WGS sequencing data [PRJNA291063] was used to generate cfDNA fragmentation maps. **(C)**. In order to decipher the complex variable of the cfDNA fragmentation pattern and its mapping to amplicons positions and lengths, four features were introduced, namely, observed fragments (Feature A), depth range (Feature B), depth change (Feature C), and depth shape (Feature D).

GC content equal-frequency discretization (four groups) and analysis of variance of both dependent variables between groups, yielded a *p*-value of less than 1e-6. Thus, a strong non-linear correlation between the GC content, a sequencing coverage and its uniformity (**Figures 6C,D**) was detected. Despite significant linear correlation between counts of spanning fragments and the GC contents (**Figure 6B**), no similar relationship between this feature and sequencing coverage was seen (**Figures 7A,B**). In contrast, as for coverage uniformity, both spanning fragments, count and read depth coverage, shape the demonstrated correlation in relation to it (ANOVA test *p*-value of 0.037 and 0.013, respectively) (**Figures 7C,D**). No correlation was seen for depth change or depth range (data not shown).

Finally, we tested the performance of the SVM classifier for its prediction of coverage depth and coverage uniformity by either employing the GC content as a single feature or in a combination with all the other features analyzed above. Following 3-groups equal-frequency discretization, the target classes were defined as coverage depth in the lowest third tertile and coverage uniformity in the highest third tertile. For predicting the depth of coverage, GC content in combination with depth change (Feature C) were selected as features. To predict the uniformity of coverage, GC content in combination with the spanning fragment counts (Feature A) and read depth shape (Feature D) were selected as features. A radial basis function (RBF)-kernel utilizing SVM classifier was then applied, using threefold cross-validation. Performance of the SVM classifiers, built upon several features for predicting coverage uniformity, was better than that of the GC-content only classifiers (areas under the receiver operating curve (AUROCs) of 0.75, 95% CI: 0.750–0.752 vs. 0.65, 95% CI: 0.63–0.67; precision – 0.74 vs. 0.68). This indicates that non-GC content features may aid in the prediction of the amplicons with a high coverage variation across samples. For coverage depths, however, applying a similar strategy has not resulted in a significant improvement (AUROCs of 0.69, 95% CI: 0.68–0.70 vs. 0.70, 95% CI: 0.70–0.71; precision – 0.69 vs. 0.69) (**Figures 7E,F**).

DISCUSSION

The share of cfDNA fragments originating from tumor rather than normal tissues, may vary greatly among patients. In early-stage disease, the share could be as low as 0.01% of the total cfDNA (Thierry et al., 2017). Because of that, the issue of the

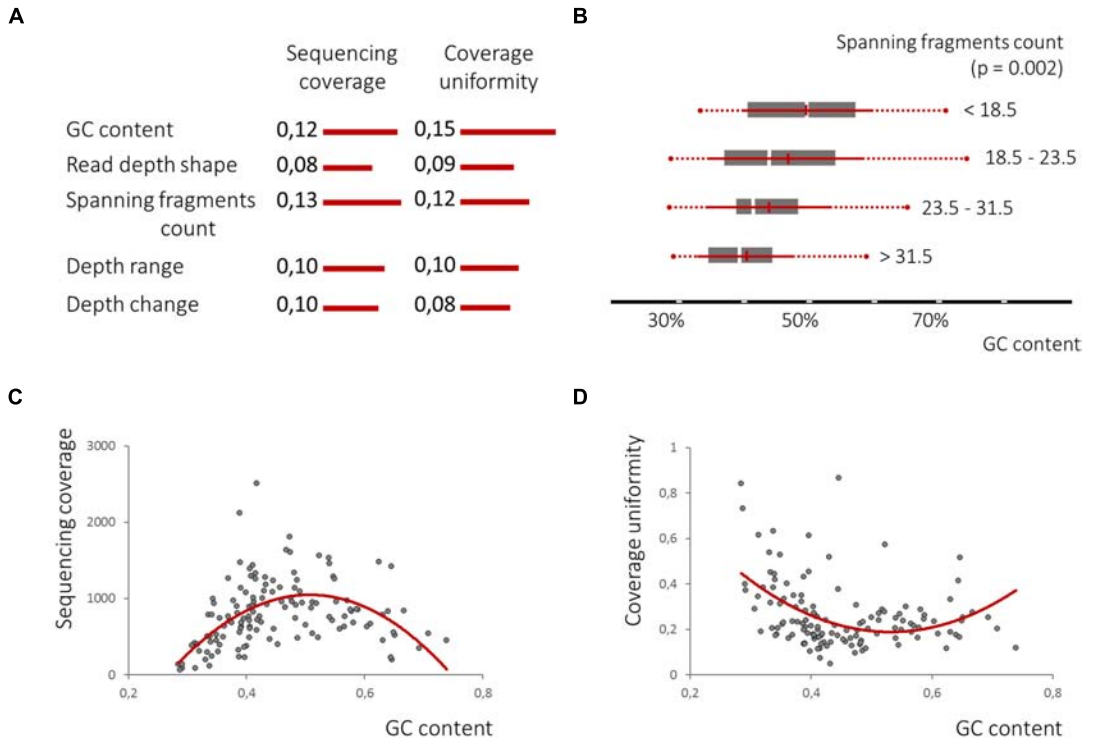


FIGURE 6 | The nucleosome-guided cfDNA fragmentation pattern influences amplicon mean coverage and its uniformity across samples. **(A)**. RReliefF ranking of the defined features in relation to amplicon mean coverage and its uniformity across samples listed in comparison with GC content ranks, previously shown to have strong non-linier correlation with both dependent variables **(C,D)**, demonstrates the significance of all four features for the prediction of target variables, though linear correlation between spanning fragment counts and GC content was observed **(B)**.

detection of low frequency mutant alleles, represents one of the biggest technical challenges to the development of diagnostic and prognostic assays involving the sequencing of cfDNA. In this study we examined various approaches to increase diagnostic and analytical sensitivity of the detection of somatic mutations in liquid biopsy samples.

In a heterogeneous cohort of patients, the liquid biopsy was performed at baseline, at disease progression and/or within the framework of disease monitoring. Overall diagnostic sensitivity of NGS to detect EGFR mutations in cfDNA was at 83%. Of note, when we limited the sample set to the plasma specimens with DNA concentration of 20 ng/ml and higher, the false negative rate was reduced from 17 to 0%. This observation points at low concentrations of cfDNA samples as a primary contributor to imperfect sensitivity of the liquid biopsy assays and at a necessity to either improve the recovery of tumor DNA fragments, or to require cfDNA profiling labs to introduce more stringent QC metrics, which may render many samples ineligible for downstream processing.

Sensitivity of cfDNA based mutation detection assays may be aided by an improvement of amplification efficiency. Plasma cfDNA is known to be highly fragmented (Fleischhacker et al., 2011; Klevebring et al., 2014; **Figure 2B**). Therefore, it is commonly recognized that an increase in length of PCR amplicons may result in the elimination of a majority of the extracted DNA fragments as possible templates. In this study we sought to dissect how much of the amplicon length influences the sensitivity of subsequent mutation detection. For this we performed, to the best of our knowledge, the first comparison of two amplicon based NGS panels characterized by a substantial difference in average amplicon length (**Figure 2A**). The comparison was performed in relation to the panels' diagnostic and analytical sensitivity. Surprisingly, the yield of both the germline and somatic mutations between two panels were completely concordant, pointing at an irrelevance of amplicon size of the specified short range to diagnostic sensitivity of resultant assays.

As a particular example defying "the shorter amplicon, the better amplification efficiency" logic, we dissected the detection of EGFR exon 19 deletion alleles by amplicons of 138 and 168 nt in length. Based on the area under the fragment length distribution curves (**Figure 2B**), mutant alleles should be amplified 1.45 times more efficiently than wild-type ones by the panel with larger amplicons, while the panel with shorter amplicons would be 1.04 times more efficient for mutant cfDNA fragments. Considering that tumor-derived cfDNA fragments are even shorter than normal tissue-derived ones (Jiang et al., 2018), these rates would increase to 1.84 and 1.16, respectively (**Figure 3**). This should result in approximately and increase of 1.6 times of the mutant allele frequencies detected with a larger-amplicon panel as compared to a smaller-amplicon panel. In our experiment, no statistically significant difference in mutant allele frequencies

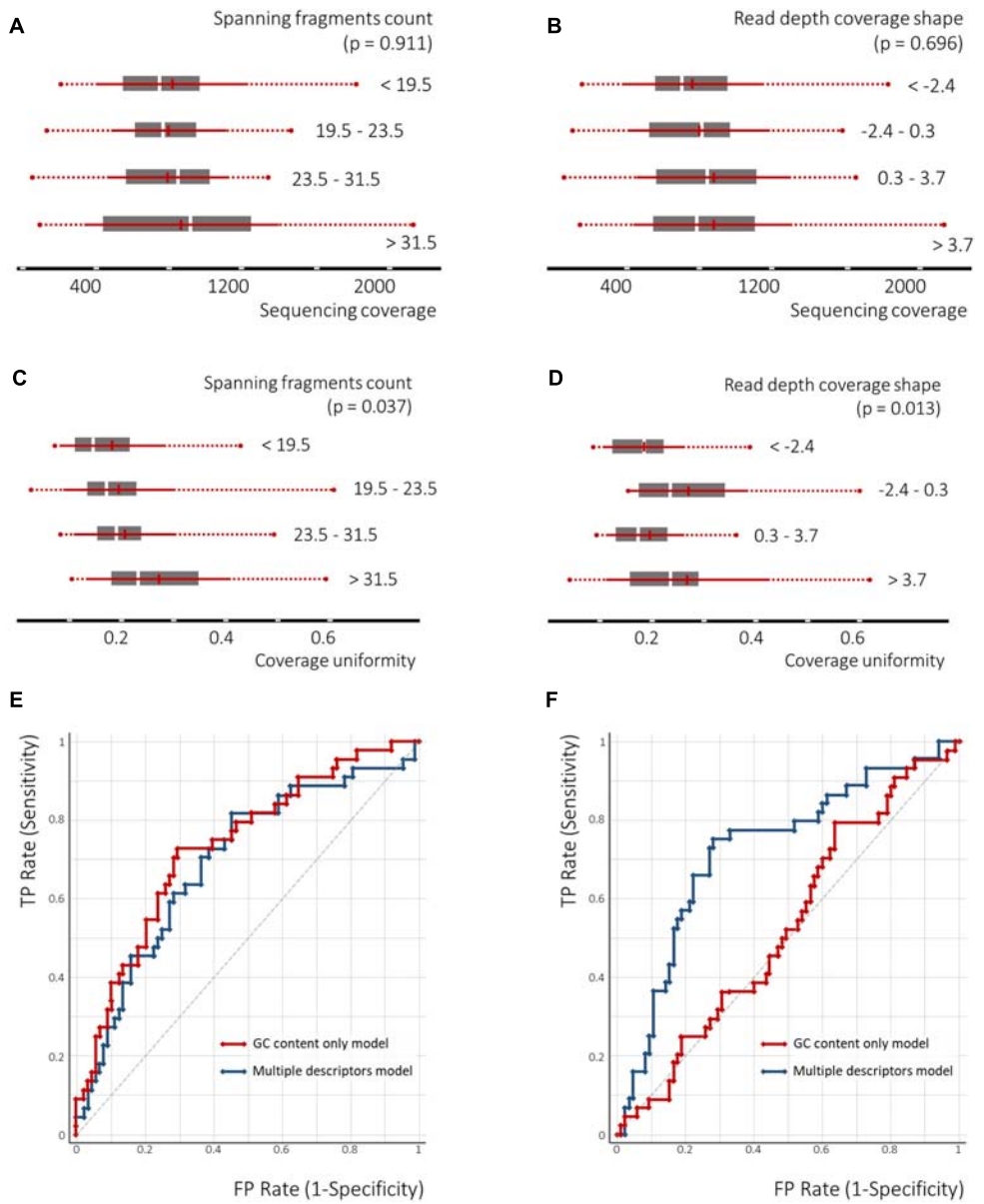


FIGURE 7 | The nucleosome-guided cfDNA fragmentation pattern facilitates the prediction of amplicon coverage uniformity across samples. The ANOVA test demonstrates dependency between spanning fragment counts and coverage uniformity as well as read depth overage shape and coverage uniformity **(C,D)**, though no similar relation to the mean amplicon coverage across the samples was observed **(A,B)**. The SVM classification model utilizing the GC content as a single feature, or in combination with the cfDNA fragmentation pattern defining features, demonstrates the significance of the latter to predict amplicon coverage uniformity across samples **(F)**, but not of the mean sequencing coverage across samples **(E)**.

was noted, with the observed trend being the opposite to what was expected, indicating that the size of the amplicons does not contribute to the analytical sensitivity of cfDNA assays.

Notably, our observations contradict some previous work (Chan et al., 2004; Koide et al., 2005), which show a length-dependent decrease in efficiency of amplification of cfDNA templates in up to a 250 nt fragment range, which corresponds to the mononucleosome fraction representing approximately 85% of all cfDNA fragments (**Figure 2B**). In these previous studies, the yield of DNA dropped by almost 30 and 60% when using amplicons with a size of 145 nt instead of 105 and 201 nt instead of 145 nt, while for amplicons with larger sizes no pronounced effect was observed. Furthermore, another study demonstrated that increases in the DNA yield may be observed at a lower amplicon size range: a direct digital PCR comparison of the 50 bp to the 84 bp amplicon resulted in significant favoring of the shorter amplicon (Koide et al., 2005; Sikora et al., 2010). It is important, however, to note that reported observations were obtained in course of analysis if cfDNA samples collected either from healthy individuals or in setting of prenatal diagnostics

aimed at amplifying fetal cfDNA and, therefore, cannot be directly projected onto the templates of tumor-derived cfDNA which is known for the shorter sizes of its fragments (Pinzani et al., 2011; Mouliere and Rosenfeld, 2015) and lower integrity (Underhill et al., 2016). The studies of cfDNA specimens collected from patients with tumors show that 60 bp fragments are almost five times more abundant than 150 bp ones, thus pointing at the necessity to use amplicons with sizes of 100 bp or lower (Mouliere et al., 2011).

Importantly, in many cases, reaping the benefit of shorter amplicon size may not be possible due to complications arising from the necessity of the precise positioning of the primers restricting optimization of their GC content, matching melting temperatures and preventing oligonucleotide dimerization. While designing PCR systems for select loci may be still possible, with EGFR analysis being the common example (Reckamp et al., 2016), the introduction of ultra-short amplicons into highly multiplexed systems aiming at a broader molecular profiling of human tumors, may not be feasible. Particular concerns about this multiplexing precluding approach to the amplicon design are owed to the recent observations of a wide mutational spectrum in the liquid biopsies of metastatic cancer patients and its relevance to possible inclusion in clinical trials (Rothé et al., 2014; Frenel et al., 2015). In light of an obvious necessity for multiplexing, the finding that varying amplicon sizes in a range from 140 up to 170 nt does not influence analytical sensitivity is significant, as it shifts the attention of panel designers from minimizing the length of the amplicons to optimizing compatibility of oligonucleotides.

Additionally, cfDNA as a template for a designed PCR-based assay may introduce a set of additional restraints. Both the prevalence of cfDNA fragments of certain sizes and the fragmentation patterns depend on the positioning of the nucleosomes within its tissue of origin. To describe this novel complex variable depicting nucleosome positioning, we introduced four features namely, a spanning fragment count, a read depth change, a read depth range and a read depth shape (**Figure 5C**), which collectively portray the coverage of select amplicon by experimentally obtained WGS reads. When read coverage maps of WGS-sequenced cfDNA fragments from pooled plasma of healthy patients were aligned to the amplicons employed for liquid biopsy analysis of patients with NSCLC, these four features were utilized to determine the extent of the influence of nucleosome positioning on two dependent variables: sequencing coverage and coverage uniformity. A SVM-based classifier demonstrated that combining the GC content with spanning fragment counts and read depth shape, results in an increased accuracy of prediction of both dependent variables. Therefore, this variable should be taken in consideration when designing PCR primer systems.

Nevertheless, the overall robustness of nucleosome positioning remains unclear. It is known that several regulatory events defining the gene expression require the strict positioning of nucleosomes; these events are typically associated with promoter regions (Hesson et al., 2014; Lövkvist et al., 2018). However, nucleosome positioning is not absolute, and even with major shifts in gene expression, some cells fail to change nucleosome configuration (Small et al., 2014), thus, indicating an underlying complexity of nucleosome positioning. Importantly, the majority of clinically relevant mutations are located within exons, which, according to the current view of cfDNA nucleosome maps, do not retain a strict pattern of cfDNA fragmentation. Therefore, nucleosome arranging within such exons may be variable, either between molecular subtypes of the same disease or even between normal tissue specimens. Nevertheless, despite a potential for low robustness, a substantial correlation observed between nucleosome maps revealed by unbiased read coverage in cfDNA from healthy patients, and the sequencing coverage and its uniformity in amplicons obtained in cfDNA of patients with NSCLC, indicates that the efficiency of amplification may be improved if the unbiased read coverages are taken into account.

In conclusion, low plasma cfDNA concentration remains the major factor that limits the sensitivity of liquid biopsy assays. Above we showed that the design of a highly multiplexed system equally tolerates amplicons in the range of 140–170 bp in size, thus allowing the shift of attention toward the melting temperature, GC clamps, cross homology and other controllable variables. We have also provided evidence that the nucleosome placement in the tissue of origin and the resultant genome-wide cfDNA fragmentation pattern, may be used as a guide for primer positioning to improve both the sequencing coverage and its uniformity.

DATA AVAILABILITY

The datasets generated for this study can be found in the Sequence Read Archive under accession number SRP167082 (https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP167082). The additional datasets analyzed for this study can be found in the Sequence Read Archive under accession number SRP061633.

AUTHOR CONTRIBUTIONS

VM, AB, MI, and SM designed the work. PC, KL, and VB collected the samples. ER performed the experiments. All authors participated in the interpretation of the results and in writing the article.

FUNDING

This study was supported by the Ministry of Science and Education, Russia (Project No. RFMEFI60714X0098).

ACKNOWLEDGMENTS

The authors wish to gratefully acknowledge technical support from laboratory of epigenetics, Medical Genetic Science Center RAMS. Special thanks to Drs. Strelnikov and Tanas for their technical assistance and helpful discussions of presented here results.

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Benjamini, Y., and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* 40:e72. doi: 10.1093/nar/gks001

Bryzgunova, O., Bondar, A., Morozkin, E., Mileyko, V., Vlassov, V., and Laktionov, P. (2011). A reliable method to concentrate circulating DNA. *Anal. Biochem.* 408, 354–356. doi: 10.1016/j.ab.2010.09.005

Chan, K. C., Zhang, J., Hui, A. B., Wong, N., Lau, T. K., Leung, T. N., et al. (2004). Size distributions of maternal and fetal DNA in maternal plasma. *Clin. Chem.* 50, 88–92. doi: 10.1373/clinchem.2003.024893

Chandrananda, D., Thorne, N. P., and Bahlo, M. (2015). High-resolution characterization of sequence signatures due to non-random cleavage of cell-free DNA. *BMC Med. Genomics* 8:29. doi: 10.1186/s12920-015-0107-z

Dawson, S. J., Tsui, D. W., Murtaza, M., Biggs, H., Rueda, O. M., Chin, S. F., et al. (2013). Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N. Engl. J. Med.* 368, 1199–1209. doi: 10.1056/NEJMoa1213261

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., et al. (2013). Orange: data mining toolbox in python. *J. Mach. Learn. Res.* 14, 2349–2353.

Fan, H. C., Blumenfeld, Y. J., Chitkara, U., Hudgins, L., and Quake, S. R. (2008). Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16266–16271. doi: 10.1073/pnas.0808319105

Fleischhacker, M., Schmidt, B., Weickmann, S., Fersching, D. M., Leszinski, G. S., Siegele, B., et al. (2011). Methods for isolation of cell-free plasma DNA strongly affect DNA yield. *Clin. Chim. Acta* 412, 2085–2088. doi: 10.1016/j.cca.2011.07.011

Forbes, S. A., Tang, G., Bindal, N., Bamford, S., Dawson, E., Cole, C., et al. (2010). COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer. *Nucleic Acids Res.* 38, D652–D657. doi: 10.1093/nar/gkp995

Frenel, J. S., Carreira, S., Goodall, J., Roda, D., Perez-Lopez, R., Tunariu, N., et al. (2015). Serial next-generation sequencing of circulating cell-free DNA evaluating tumor clone response to molecularly targeted drug administration. *Clin. Cancer Res.* 21, 4586–4596. doi: 10.1158/1078-0432.CCR-15-0584

Hellwig, S., Nix, D. A., Gligorich, K. M., O'Shea, J. M., Thomas, A., Fuertes, C. L., et al. (2018). Automated size selection for short cell-free DNA fragments enriches for circulating tumor DNA and improves error correction during next generation sequencing. *PLoS One* 13:e0197333. doi: 10.1371/journal.pone.0197333

Hesson, L. B., Sloane, M. A., Wong, J. W., Nunez, A. C., Srivastava, S., Ng, B., et al. (2014). Altered promoter nucleosome positioning is an early event in gene silencing. *Epigenetics* 9, 1422–1430. doi: 10.4161/15592294.2014.970077

Ivanov, M., Baranova, A., Butler, T., Spellman, P., and Mileyko, V. (2015). Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genomics* 16(Suppl. 13):S1. doi: 10.1186/1471-2164-16-S13-S1

Ivanov, M., Matsvay, A., Glazova, O., Krasovskiy, S., Usacheva, M., Amelina, E., et al. (2018). Targeted sequencing reveals complex, phenotype-correlated genotypes in cystic fibrosis. *BMC Med. Genomics* 11(Suppl. 1):13. doi: 10.1186/s12920-018-0328-z

Jiang, P., Chan, C. W., Chan, K. C., Cheng, S. H., Wong, J., Wong, V. W., et al. (2015). Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc. Natl. Acad. Sci. U.S.A.* 112, E1317–E1325. doi: 10.1073/pnas.1500076112

Jiang, P., Sun, K., Tong, Y. K., Cheng, S. H., Cheng, T. H. T., Heung, M. M. S., et al. (2018). Preferred end coordinates and somatic variants as signatures of circulating tumor DNA associated with hepatocellular carcinoma. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10925–E10933. doi: 10.1073/pnas.1814616115

Klevebring, D., Neiman, M., Sundling, S., Eriksson, L., Darai Ramqvist, E., Celebioglu, F., et al. (2014). Evaluation of exome sequencing to estimate tumor burden in plasma. *PLoS One* 9:e104417. doi: 10.1371/journal.pone.0104417

Koide, K., Sekizawa, A., Iwasaki, M., Matsuoka, R., Honma, S., Farina, A., et al. (2005). Fragmentation of cell-free fetal DNA in plasma and urine of pregnant women. *Prenat. Diagn.* 25, 604–607. doi: 10.1002/pd.1213

Krishnamurthy, N., Spencer, E., Torkamani, A., and Nicholson, L. (2017). Liquid biopsies for cancer: coming to a patient near you. *J. Clin. Med.* 6:E3. doi: 10.3390/jcm6010003

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi: 10.1038/nmeth.1923

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. doi: 10.1093/bioinformatics/btr509

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Lo, Y. M., Chan, K. C., Sun, H., Chen, E. Z., Jiang, P., Lun, F. M., et al. (2010). Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Sci. Transl. Med.* 2:61ra91. doi: 10.1126/scitranslmed.3001720

Lövkvist, C., Sneppen, K., and Haerter, J. O. (2018). Exploring the link between nucleosome occupancy and DNA methylation. *Front. Genet.* 8:232. doi: 10.3389/fgene.2017.00232

Ma, X., Zhu, L., Wu, X., Bao, H., Wang, X., Chang, Z., et al. (2017). Cell-Free DNA provides a good representation of the tumor genome despite its biased fragmentation patterns. *PLoS One* 12:e0169231. doi: 10.1371/journal.pone.0169231

Malapelle, U., Pisapia, P., Rocco, D., Smeraglio, R., di Spirito, M., Bellevicine, C., et al. (2016). Next generation sequencing techniques in liquid biopsy: focus on non-small cell lung cancer patients. *Transl. Lung Cancer Res.* 5, 505–510. doi: 10.21037/tlcr.2016.10.08

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F., et al. (2010). Tablet–next generation sequence assembly visualization. *Bioinformatics* 26, 401–402. doi: 10.1093/bioinformatics/btp666

Mouliere, F., Robert, B., Arnau Peyrotte, E., Del Rio, M., Ychou, M., Molina, F., et al. (2011). High fragmentation characterizes tumour-derived circulating DNA. *PLoS One* 6:e23418. doi: 10.1371/journal.pone.0023418

Mouliere, F., and Rosenfeld, N. (2015). Circulating tumor-derived DNA is shorter than somatic DNA in plasma. *Proc. Natl. Acad. Sci. U.S.A.* 112, 3178–3179. doi: 10.1073/pnas.1501321112

Oxnard, G. R., Paweletz, C. P., Kuang, Y., Mach, S. L., O'Connell, A., Messineo, M. M., et al. (2014). Noninvasive detection of response and resistance in EGFR-mutant lung cancer using quantitative next-generation genotyping of cell-free plasma DNA. *Clin. Cancer Res.* 20, 1698–1705. doi: 10.1158/1078-0432.CCR-13-2482

Pinzani, P., Salvianti, F., Zaccara, S., Massi, D., De Giorgi, V., Pazzagli, M., et al. (2011). Circulating cell-free DNA in plasma of melanoma patients: qualitative and quantitative considerations. *Clin. Chim. Acta* 412, 2141–2145. doi: 10.1016/j.cca.2011.07.027

Pupilli, C., Pinzani, P., Salvianti, F., Fibbi, B., Rossi, M., Petrone, L., et al. (2013). Circulating BRAFV600E in the diagnosis and follow-up of differentiated papillary thyroid carcinoma. *J. Clin. Endocrinol. Metab.* 98, 3359–3365. doi: 10.1210/jc.2013-1072

Reckamp, K. L., Melnikova, V. O., Karlovich, C., Sequist, L. V., Camidge, D. R., Wakelee, H., et al. (2016). A highly sensitive and quantitative test platform for detection of NSCLC EGFR mutations in urine and plasma. *J. Thorac. Oncol.* 11, 1690–1700. doi: 10.1016/j.jtho.2016.05.035

Robnik-Sikonja, M., and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* 53, 23–69. doi: 10.1023/A:1025667309714

Rothé, F., Laes, J. F., Lambrechts, D., Smeets, D., Vincent, D., Maetens, M., et al. (2014). Plasma circulating tumor DNA as an alternative to metastatic biopsies for mutational analysis in breast cancer. *Ann. Oncol.* 25, 1959–1965. doi: 10.1093/annonc/mdu288

Sacher, A. G., Paweletz, C., Dahlberg, S. E., Alden, R. S., O'Connell, A., Feeney, N., et al. (2016). Prospective validation of rapid plasma genotyping for the detection of EGFR and KRAS mutations in advanced lung cancer. *JAMA Oncol.* 2, 1014–1022. doi: 10.1001/jamaoncol.2016.0173

Saunders, C. T., Wong, W. S., Swamy, S., Becq, J., Murray, L. J., and Cheetham, R. K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 28, 1811–1817. doi: 10.1093/bioinformatics/bts271

Sikora, A., Zimmermann, B. G., Rusterholz, C., Birri, D., Kolla, V., Lapaire, O., et al. (2010). Detection of increased amounts of cell-free fetal DNA with short PCR amplicons. *Clin. Chem.* 56, 136–138. doi: 10.1373/clinchem.2009.132951

Siravegna, G., Mussolin, B., Buscarino, M., Corti, G., Cassingena, A., Crisafulli, G., et al. (2015). Clonal evolution and resistance to EGFR blockade in the blood of colorectal cancer patients. *Nat. Med.* 21, 795–801. doi: 10.1038/nm.3870

Small, E. C., Xi, L., Wang, J. P., Widom, J., and Licht, J. D. (2014). Single-cell nucleosome mapping reveals the molecular basis of gene expression heterogeneity. *Proc. Natl. Acad. Sci. U.S.A.* 111, 2462–2471. doi: 10.1073/pnas.1400517111

Snyder, M. W., Kircher, M., Hill, A. J., Daza, R. M., and Shendure, J. (2016). Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell* 164, 57–68. doi: 10.1016/j.cell.2015.11.050

Struhl, K., and Segal, E. (2013). Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* 20, 267–273. doi: 10.1038/nsmb.2506

Tafe, L. J., Gorlov, I. P., de Abreu, F. B., Lefferts, J. A., Liu, X., Pettus, J. R., et al. (2015). Implementation of a molecular tumor board: the impact on treatment decisions for 35 patients evaluated at dartmouth-hitchcock medical center. *Oncologist* 20, 1011–1018. doi: 10.1634/theoncologist.2015-0097

Thierry, A. R., Pastor, B., Jiang, Z. Q., Katsiampoura, A. D., Parseghian, C., Loree, J. M., et al. (2017). Circulating DNA demonstrates convergent evolution and common resistance mechanisms during treatment of colorectal cancer. *Clin. Cancer Res.* 23, 4578–4591. doi: 10.1158/1078-0432.CCR-17-0232

Thress, K. S., Brant, R., Carr, T. H., Dearden, S., Jenkins, S., Brown, H., et al. (2015). EGFR mutation detection in ctDNA from NSCLC patient plasma: a cross-platform comparison of leading technologies to support the clinical development of AZD9291. *Lung Cancer* 90, 509–515. doi: 10.1016/j.lungcan.2015.10.004

Ulz, P., Thallinger, G. G., Auer, M., Graf, R., Kashofer, K., Jahn, S. W., et al. (2016). Inferring expressed genes by whole-genome sequencing of plasma DNA. *Nat. Genet.* 48, 1273–1278. doi: 10.1038/ng.3648

Underhill, H. R., Kitzman, J. O., Hellwig, S., Welker, N. C., Daza, R., Baker, D. N., et al. (2016). Fragment length of circulating tumor DNA. *PLoS Genet.* 12:e1006162. doi: 10.1371/journal.pgen.1006162

Valouev, A., Johnson, S. M., Boyd, S. D., Smith, C. L., Fire, A. Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. *Nature* 474, 516–520. doi: 10.1038/nature10002

Wei, Z., Shah, N., Deng, C., Xiao, X., Zhong, T., and Li, X. (2016). Circulating DNA addresses cancer monitoring in non small cell lung cancer patients for detection and capturing the dynamic changes of the disease. *Springerplus* 5:531. doi: 10.1186/s40064-016-2141-5

Xi, L., Pham, T. H., Payabyab, E. C., Sherry, R. M., Rosenberg, S. A., and Raffeld, M. (2016). Circulating tumor DNA as an early indicator of response to T-cell transfer immunotherapy in metastatic melanoma. *Clin. Cancer Res.* 22, 5480–5486. doi: 10.1158/1078-0432.CCR-16-0613

Zehir, A., Benayed, R., Shah, R. H., Syed, A., Middha, S., Kim, H. R., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* 23, 703–713. doi: 10.1038/nm.4333

Conflict of Interest Statement: ER, SM, VM, and AB were employed by company Atlas Oncology Diagnostics, Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer, TT declared a past co-authorship with one of the authors AB to the handling Editor.

frontiers in Genetics

ORIGINAL RESEARCH
published: 25 March 2019
doi: 10.3389/fgene.2019.00231

Transcriptomic Analysis of Seed Germination Under Salt Stress in Two Desert Sister Species (*Populus euphratica* and *P. pruinosa*)

Caihua Zhang[1†], Wenchun Luo[1†], Yanda Li[2], Xu Zhang[1], Xiaotao Bai[1], Zhimin Niu[1], Xiao Zhang[1], Zhijun Li[3] and Dongshi Wan[1]*

[1] State Key Laboratory of Grassland Agro-Ecosystem, School of Life Sciences, Lanzhou University, Lanzhou, China, [2] Computer Science and Engineering Department, University of California, San Diego, La Jolla, CA, United States, [3] Xinjiang Production & Construction Corps, Key Laboratory of Protection and Utilization of Biological Resources in Tarim Basin, College of Life Sciences, Tarim University, Xinjiang, China

OPEN ACCESS

Edited by:
Ancha Baranova,
George Mason University,
United States

Reviewed by:
Petronia Carillo,
Università degli Studi della Campania Luigi Vanvitelli Caserta, Italy
Andrés A. Borges,
Spanish National Research Council (CSIC), Spain

***Correspondence:**
Dongshi Wan
wandsh@lzu.edu.cn

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 16 October 2018
Accepted: 04 March 2019
Published: 25 March 2019

Citation:
Zhang C, Luo W, Li Y, Zhang X, Bai X, Niu Z, Zhang X, Li Z and Wan D (2019) Transcriptomic Analysis of Seed Germination Under Salt Stress in Two Desert Sister Species (Populus euphratica and P. pruinosa). Front. Genet. 10:231. doi: 10.3389/fgene.2019.00231

As a major abiotic stress, soil salinity limits seed germination and plant growth, development and production. Seed germination is highly related not only to the seedlings survival rate but also subsequent vegetative growth. *Populus euphratica* and *P. pruinosa* are closely related species that show a distinguished adaptability to salinity stress. In this study, we performed an integrative transcriptome analyses of three seed germination phases from *P. euphratica* and *P. pruinosa* under salt stress. A two-dimensional data set of this study provides a comprehensive view of the dynamic biochemical processes that underpin seed germination and salt tolerance. Our analysis identified 12831 differentially expressed genes (DEGs) for seed germination processes and 8071 DEGs for salt tolerance in the two species. Furthermore, we identified the expression profiles and main pathways in each growth phase. For seed germination, a large number of DEGs, including those involved in energy production and hormonal regulation pathways, were transiently and specifically induced in the late phase. In the comparison of salt tolerance between the two species, the flavonoid and brassinosteroid pathways were significantly enriched. More specifically, in the flavonoid pathway, *FLS* and *F3′5′H* exhibited significant differential expression. In the brassinosteroid pathway, the expression levels of *DWF4*, *BR6OX2* and *ROT3* were notably higher in *P. pruinosa* than in *P. euphratica*. Our results describe transcript dynamics and highlight secondary metabolite pathways involved in the response to salt stress during the seed germination of two desert poplars.

Keywords: transcriptome, salt stress, seed germination, differentially expressed gene, desert poplar species

INTRODUCTION

Soil salinization is caused by many factors and conditions, such as unsuitable irrigation practices, irrigation with salinized water and seasonal effects (Ottow et al., 2005; Annunziata et al., 2017). As one of the most prominent abiotic stresses, salinity stress is considered the greatest threat to crop production and environmental conservation (Ottow et al., 2005; Arbona et al., 2013). Salinity stress leads to osmotic and ionic stress, which reduces cell and tissue expansion, and to ion excesses

that changes the osmotic potential of plant cells and induce nutritional imbalances (Munns, 2002), sequentially affecting plant growth, development and survival (Carillo et al., 2019). To solve the serious problem of soil salinization, various efforts have been made; these efforts mainly concentrate on enhancing the salt resistance of economically important salt-sensitive plants through traditional breeding and biotechnological approaches or the use of plants that naturally display high salt tolerance (Flowers, 2004).

Populus euphratica and its sister species *P. pruinosa* are naturally distributed in China's western desert region; due to their extraordinary adaptability to desert environments (Chen et al., 2002; Hukin et al., 2005), both species are also called desert poplars. The distinguished adaptability of these species provides beneficial ecological effects in northwest China. Currently, both poplars are considered important genetic resources in tree breeding and in research elucidating physiological and molecular mechanisms involving stress tolerance in trees (Tuskan et al., 2006; Wullschleger et al., 2013). As the genome data of *P. euphratica* becomes available (Ma et al., 2013), the resistance mechanism of both poplars have been revealed at multiple levels, e.g., a phylogenetic analysis shows that the two species diverged approximately 1–2 million years ago (Wang J. et al., 2011), and ancient polymorphisms contributed to their genomic divergence (Ma et al., 2018). In addition to leaf morphology and leaf trichome differences (Ma et al., 2016), both poplars occupy different ecological habitats. *P. pruinosa* prefers desert areas with high ground water levels, while *P. euphratica* can grow in desert areas where the groundwater levels are low (Ottow et al., 2005). These differences between the sister poplars result from differences in genetic mechanisms, such as the adaptive evolution of genes (Ma et al., 2013; Zhang et al., 2014) and gene expression divergences among orthologs (Qiu et al., 2011; Zhang et al., 2013).

Seed germination is constrained significantly by soil salinity (Kaya et al., 2003). Soil salinity creates an osmotic potential around the outside of seeds, resulting in decreased water uptake during germination and an increase in the excessive uptake of ions, which causes the toxic effects of Na^+ and Cl^- ions to seeds (Murillo-Amador et al., 2002; Khajeh-Hosseini et al., 2003). Therefore, salt stress can inhibit or delay seed germination (Almansouri et al., 2001). However, studies focusing on the genetic mechanism of seed germination under salt stress are limited.

Seed germination begins with imbibition and ends with the embryonic axis breaking through the seed coat (Bewley et al., 2012). Seed germination includes three phases. In phase I, the seed begins to expand, with a rapidly increasing water content. Then, the seed enters a plateau phase (phase II), in which the water uptake remains at a stable level. In phase III, the water uptake increases rapidly. Phase III ceases as the embryonic axis breaks through the seed coat, upon which seed germination is complete (Bewley, 1997). Energy production and respiration play important roles in the seed germination process. In the early stage, anaerobic respiration provides the main energy source, and then respiratory activity increases with oxygen uptake. Subsequently, plant hormones, such as gibberellins (GA), abscisic acid (ABA), brassinosteroids (BRs), ethylene, auxins, and cytokinins, are widely involved in determining the physiological state of a seed and regulating the germination process (Kucera et al., 2005; Holdsworth et al., 2008; Müller et al., 2009; North et al., 2010). Furthermore, numerous complex networks, including those related to gene expression and regulation commanded by various transcription factors (Chen et al., 2002), ion transporting processes, such as NHX (Na^+/H^+ antiporter), SOS (salt overly sensitive) (Zhu, 2001), and HKT (high-affinity K^+ transporter) (Ren et al., 2005) processes, and secondary metabolism are all involved in the response to salt stress (Bewley, 1982, 1997; Bewley and Black, 1984; Biligetu et al., 2011). Recently, transcriptomic analyses of several poplar species under various stresses have been extensively conducted (Chen et al., 2002, 2012; Brinker et al., 2010; Janz et al., 2010; Qiu et al., 2011; Wang J. et al., 2011; Ma et al., 2013; Ziemann et al., 2013). These data provide us with a basic understanding of seed germination. However, detailed transcriptomic dynamics and physiological mechanisms under salt stress during seed germination have not yet been revealed. Such an exploration might be useful to identify the genes that improve poplar salt tolerance by biotechnological manipulation. Moreover, most genes associated with seed germination are poorly understood due to the complexity of the germination process.

Here, we present a comprehensive transcriptome study encompassing the whole process of seed germination for two species under salt stress, which provides a valuable gene resource for genetic manipulation in poplar breeding.

MATERIALS AND METHODS

Plant Materials and Growth Conditions

We collected three replicate samples of the seeds of the two studied species from a total of 18 trees in the Tarim Basin (Xinjiang, China) and stored the seeds at 4°C. For germination, vigorous seeds were imbibed in distilled water (control), 0.2%, 0.4%, 0.6%, 0.8%, 1.0%, 1.2%, 1.4%, and 1.6% NaCl, and then germinated on wet filter paper in 9 cm diameter Petri dishes in a plant growth incubator (21°C 200 μmol $m^{-2}s^{-1}$, 16 h: 8 h light/dark photoperiod). The germination rate was measured using the Chinese national standard test (GB2772-1999). Each sample contained 50 seeds and had three replicates (Wang et al., 2013). The germinating seeds were scanned and photographed using a stereo microscope (Nikon SM Z1500, Japan) to record their morphology. The moisture content of the seed samples was measured in seeds oven dried at 75°C to a constant weight. The moisture content [g (g $FW)^{-1}$] was calculated as [(FW−DW)/FW].

The percentage of seeds with two cotyledons turning green or with emerging radicles (>1 mm) was considered the germination rate (Wang Y. et al., 2011). For the germination percentage, counts were made until no additional germination was observed for 72 h (Bradford, 1990). To elucidate the threshold salinity for the two species under the salt treatments, we measured relative indexes, including GR, RGP, GT, GI, K, and RSH (Imit et al., 2015). For RNA isolation, the seeds were imbibed in 1.0% NaCl (to expose them to salt stress) and then removed after 4, 12, 24,

48, and 72 h for RNA preparation. The control samples were collected from dry seeds (0 h). We rapidly transferred all the samples to storage at −80°C before RNA extraction.

Reactive Oxygen Species (ROS) Level and Enzyme Activity Determination

For germination, seeds were imbibed in 0%, 0.4%, 0.8% and 1.0% NaCl as described above for 24 h. The levels of ROS, superoxide dismutase (SOD) and catalase (CAT) were measured using the standard protocol for the toolkit from Suzhou Comin Biotechnology.

Determination of RNA Extraction and Quality

Using the CTAB procedure, we extracted and purified total RNA three times from each of the sample set (Chang et al., 1993). The A260/A280 ratios of all the RNA samples ranged from 1.9 to 2.0. We examined the integrity of all RNA samples by the Agilent 2100 Bioanalyzer, and all the RNA integrity number (RIN) values ranged from 7 to 10.

cDNA Library Construction and RNA Sequencing

Construction of the cDNA library and RNA sequencing were performed by BIOMARKER (Beijing, China) using the Illumina (San Diego, CA, United States) Genome Analyzer platform in accordance with the manufacturer's protocols. Paired-end sequencing was performed using a HiSeq 2500 (Illumina) platform with a read length of 125 bp.

Initial Mapping of Reads

We trimmed reads by removing adapter sequences, reads with too many (> 5%) unknown base calls (N), low-complexity sequences, and low-quality bases (i.e., sequences for which > 65% of the bases had a quality score ≤ 7). HISAT2 (Kim et al., 2015) was used to align all reads of the two species to the *P. euphratica* genome (Ma et al., 2013). Because the intrinsic divergence between the species could result in poor mapping, we did not map RNA-seq reads from the two species onto their own genomes. Next, StringTie (Pertea et al., 2015) created multiple isoforms of genes and estimated the gene expression levels (FPKM) (Trapnell et al., 2010) during assembly. To reduce the effects of background transcription, genes with FPKM ≥ 1 were used for the subsequent analysis. We calculated the Pearson correlation coefficient between biological replicates with R software using the expression data. The Pearson correlation calculated by R was used to evaluate repeatability between biological replicates.

Analysis of DEGs

We applied Ballgown (Frazee et al., 2015) to determine which transcripts were differentially expressed between two or more experiments, confirming their significance with an *F*-test. Ballgown allows both time-course and fixed-condition differential expression analyses. Therefore, two methods were employed to identify DEGs: (1) time as the main variable and species as the covariate; (2) species as the main variable and time as the covariate.

Hierarchical Clustering and Gene Co-expression Analysis

Using normalized $\log_2$ (FPKM+1) values, hierarchical clustering was completed with the pvclust package. Based on the normalized FPKM values, *K*-means clustering was performed by the *K*-Means/K-Medians Support Module (KMS) embedded in MEV 4.9[1].

Gene Functional Enrichment and qRT-PCR Analysis

GO and KEGG enrichment analyses of the two differently expressed transcript data sets were performed using a modified Chi-square test and Fisher's exact test in *R* (p-value < 0.01 and false discovery rate < 0.05). Transcription levels of genes were quantified with a MX3005P Real-Time PCR Detection System (Agilent) based on the 2(-delta C(T)) method (Livak and Schmittgen, 2001). The experiment was performed in a 20 μL volume reaction system containing 10 μL 2 × SYBR Premix ExTaq (TaKaRa) with the intercalating dye SYBR Green. All primers were designed using PRIMER5.0 software and are listed in **Supplementary Table S4**.

RESULTS

Physiological and Morphological Changes During Seed Germination

To evaluate the effect of salt stress on seed germination, the progress of seed germination has traditionally been divided into three phases based on seed water uptake during imbibition (Nonogaki et al., 2010). The first phase (phase I) occurs within the period of 0 h–36 h; the plateau phase (phase II) occurs within the period of 36 h–64 h; and phase III is continuous for 64 h–120 h during the transition to seedling growth (**Figure 1A**). We investigated the relationship between the germination rate and NaCl concentration. Seeds exhibiting high germination rates were selected and cultured in distilled water with a gradient of NaCl concentrations (0%, 0.2%, 0.4%, 0.6%, 0.8%, 1.0%, 1.2%, 1.4%, and 1.6%) (**Figure 1B**). With increasing NaCl concentration, seed germination was significantly inhibited. At different NaCl concentrations, the seed germination rates of *P. pruinosa* were higher than those of *P. euphratica*. The relative germination percentage of the two species exceeded 80% in the 0.4% NaCl solution, whereas the value approached zero in 2.4% NaCl. We hypothesized that when the relative germination percentages were 75%, 50%, and 25%, the corresponding salt concentrations could be considered suitable, critical and limiting for seed germination, respectively. In our study, for *P. euphratica*, the suitable, critical and limiting values were 0.602%, 1.161%, and 1.72%, respectively, whereas for *P. pruinosa*, these values were 0.599%, 1.179%, and 1.759%, respectively, suggesting that the

[1]http://www.tm4.org/mev

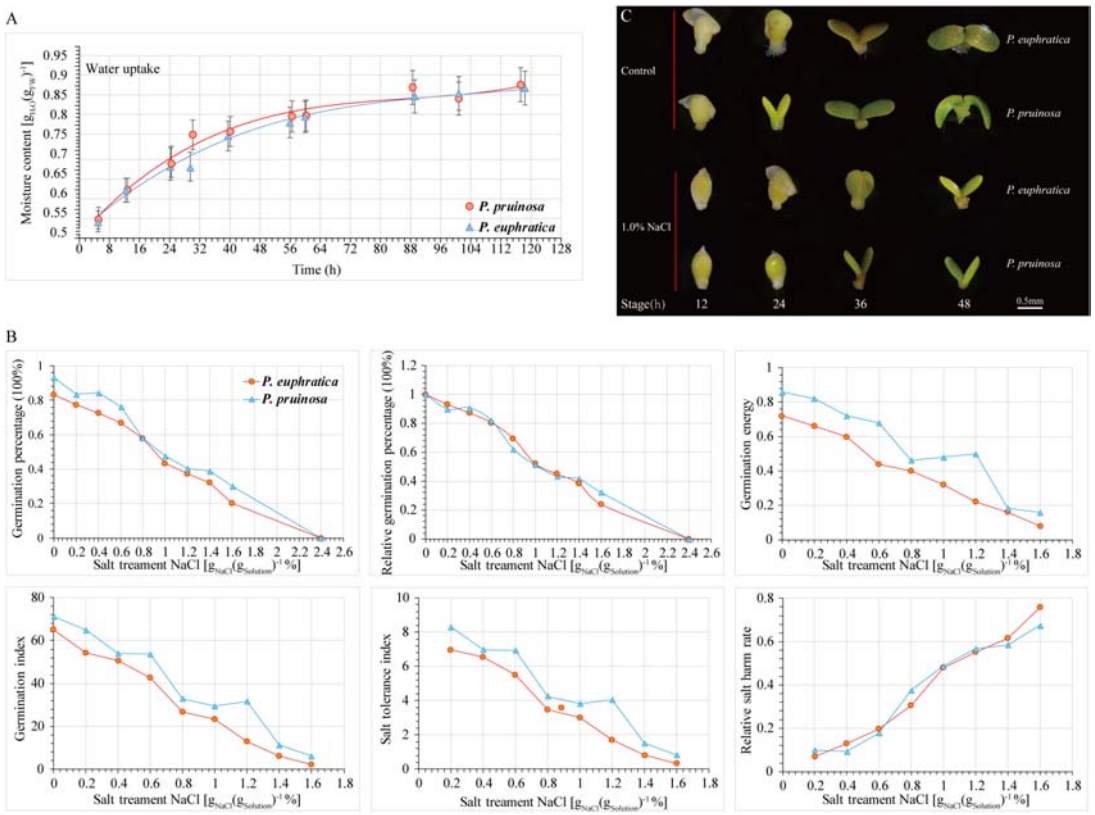


FIGURE 1 | Comparison of physiological indices for *P. euphratica* and *P. pruinosa* seeds under salt stress treatment. **(A)** Time series of changes in moisture content of *P. euphratica* and *P. pruinosa* seeds during germination. **(B)** Effect of different NaCl concentrations on 6 physiological indexes of *P. euphratica* and *P. pruinosa*. **(C)** Morphological changes in *P. euphratica* and *P. pruinosa* seeds from 12 h to 48 h under salt stress and control. Bar = 0.5 mm.

threshold salinity for the two species differed. We also measured the germination index, salt tolerance index, relative salt harm rate and germination energy (**Figure 1B**). The average germination percentage, subordinate function values, and threshold salinity for *P. pruinosa* were higher than those for *P. euphratica*. Based on the results, a 1.0% NaCl concentration was selected for the subsequent salt treatment. The seed phenotypes of the two species were observed at four time points (**Figure 1C**). In the controls, the radicle emergence was completed within 12 h, and the hypocotyl and cotyledons emerged from the seed coat by 24 h. The cotyledons started to open by 36 h and opened fully and turned green by 48 h. In contrast, under the salt treatment, the seeds were still in the imbibition stage at 12 h, the radicle emergence stage was completed by 24 h, and the subsequent stages were all delayed by 12 h.

RNA-Seq and Mapping of Illumina-Solexa Sequencing Reads

To systematically investigate the transcriptome dynamics of the two species' seeds during germination under salt stress, we obtained 36 transcriptome samples. After removing low-quality sequences and trimming adapter sequences, 3–6 GB 125-bp paired-end clean reads were generated from each library (**Supplementary Table S1**). Approximately 80% of the reads matched the genome (**Supplementary Table S2**). All the genes and transcripts were reassembled (**Table 1**).

In the detection of minor differential gene expression between time points and the two species, we used three biological replicates (**Supplementary Figures S1**, **S2**) to assess our data quality. The results showed that the expression values of biological replicates from the same samples were highly correlated (average $R^2 > 0.8$). Among the genes, FPKM values exceeding 73% ranged from 1 to 100 at each time point (**Supplementary Figure S3A**). We used the average RPKM of the biological replicates as the expression quantity. To examine the divergence in gene expression between the two species under salt stress in more detail, we performed a hierarchical clustering analysis for all the expressed genes from *P. euphratica* and *P. pruinosa* at each time point using bootstrapping

TABLE 1 | Numbers of assembled genes and transcriptions.

		0h	4h	12h	24h	48h	72h
P. euphratica	Trans Num	68202	73566	79203	69325	78689	81459
	Gene Num	28836	29992	30379	29901	30505	30985
P. pruinosa	Trans Num	85952	54047	65547	58952	76926	71959
	Gene Num	30272	26891	29152	27620	30126	29842

(**Supplementary Figure S3B**). The correlation dendrogram in **Supplementary Figure S3B** shows that samples collected at 0, 4 and 12 h clustered together, while those collected at 24, 48 and 72 h clustered into another group. This result indicates that one set of genes was activated during the early stress and germination stages, while there was another set of genes that was differentially expressed after 48 h. Therefore, based on a Spearman correlation analysis, the germinating seed samples from 0, 4 and 12 h were in the early phase, the seeds in the sample from 24 h were in the middle phase, and the seeds from 48 h to 72 h were in the late phase of the germination process.

Identification of DEGs, Temporal Expression Trends and GO Functional Enrichment

To identify global transcriptional changes that occurred during seed germination under salt stress, we confirmed the two data sets of DEGs, including 12831 DEGs and 19004 differentially expressed transcripts (DETs) for seed germination processes, and 8071 DEGs and 19000 DETs for salt tolerance, of two species. The DEGs were grouped into ten clusters (designated K1–K10) (**Supplementary Figure S4**) to examine the temporal expression trends of seed germination processes. To better understand the functions of the DEGs and obtain a view of functional transitions across time during seed germination in the two species, GO category enrichment analysis was performed (**Supplementary Figure S5**) to identify important events (biophysical, biochemical, and cellular processes) during seed germination.

According to the cluster analysis results, all the clusters of *P. pruinosa* and *P. euphratica* could be divided into early (0–12 h), middle (24 h), and late (48–72 h) phases (**Supplementary Figure S3B**). The early phase (represented by clusters K1 to K4) was strongly expressed at 0–12 h and gradually downregulated between 12 and 72 h in the two species. Based on the GO enrichment results, genes related to "adenyl nucleotide binding," "adenyl ribonucleotide binding," "purine ribonucleoside binding," and "purine nucleoside binding" were increasingly expressed after imbibition (**Supplementary Figure S5**). Second, some genes associated with "structural molecule activity," "structural constituent of cytoskeleton," "intracellular non-membrane-bounded organelle," "non-membrane-bounded organelle" and "cellular structure restoration" were enriched (**Supplementary Figure S5**). In addition, some genes associated with "ATP binding" were enriched (**Supplementary Figure S5**).

Genes in cluster K5 were highly expressed at 0 to 24 h and downregulated from 48 to 72 h. In the middle phase, the enriched genes included genes associated with "catalytic activity," "mitochondrial part," "nutrient reservoir activity," "electron transport chain," and "respiratory electron transport chain" (**Supplementary Figure S5**). Each of the five co-expression modules of the two species could be roughly categorized in the late (K6 to K10) phase. Transcripts of these modules were significantly upregulated during at least the last two time points. Many genes of this stage were typified by the enriched functions of "catabolic process," "generation of precursor metabolites and energy," "lipid metabolic process," "carbohydrate metabolic process," "hydrolase activity," and "catalytic activity" (**Supplementary Figure S5**). Moreover, some upregulated genes of this stage were associated with "cellular nitrogen compound biosynthetic process" and "NAD binding" (**Supplementary Figure S5**).

Functional Regulatory Network Analysis (KEGG Pathway Enrichment) of Seed Germination Process DEGs

To further elucidate the seed germination process DEGs associated with biochemical pathways, we performed a KEGG pathway enrichment analysis. A total of 3847 out of 12831 DEGs enriched 328 pathways, and 58 pathways were significantly (p-value $\leq$ 0.01) overrepresented during seed germination (**Supplementary Figure S6**).

The early phase was exemplified by an observed statistically significant enrichment of "ribosome," "proteasome," and "protein processing in endoplasmic reticulum" pathways (**Supplementary Figure S6**). The middle phase exhibited the enrichment of "flavonoid biosynthesis," "oxidative phosphorylation," "ribosome," "proteasome" and "spliceosome" pathways (**Supplementary Figure S6**). While many genes related to the metabolism of free amino acids were enriched in phase III (**Supplementary Figure S6**), most of the major pathways were enriched in the late phase, including "carbon metabolism," "glycolysis/gluconeogenesis," "starch and sucrose metabolism," "oxidative phosphorylation," "photosynthesis," "porphyrin and chlorophyll metabolism," and "carotenoid biosynthesis" (**Supplementary Figure S6**). "Oxidative phosphorylation" provides ATP for other metabolism pathways, such as mitochondrial repair and differentiation (Weitbrecht et al., 2011). The glyoxylate pathway contains a key step in the conversion of fatty acids to sucrose (Pritchard et al., 2002).

DEGs Related to Energy Production for Seed Germination Processes

During the preliminary phase, due to the inactivation of photosynthesis, the degradation of storage needed for energy production via processes such as glycolysis, the glyoxylate cycle, and the tricarboxylic acid (TCA) cycle, largely determines germination vigor. We defined the relative functional categories to be "carbon metabolism," "glyoxylate and dicarboxylate metabolism," "glycolysis/gluconeogenesis," and "starch and sucrose metabolism." Then, we identified the four major energy production processes, i.e., fermentation, the TCA cycle, glyoxylate and glycolysis, representing significantly overrepresented functional pathways, and we examined the expression patterns of the related DEGs (**Figure 2**). Here, ten gene families participating in the TCA cycle were differentially expressed over time in the two species. With respect to glycolysis, numerous gene families were upregulated, such as GALM, PFK, FBP, ALDO, GAPDH, and PK. In anaerobic respiration, three related gene families, PDC, ADH, and LDH, were all upregulated in the two species.

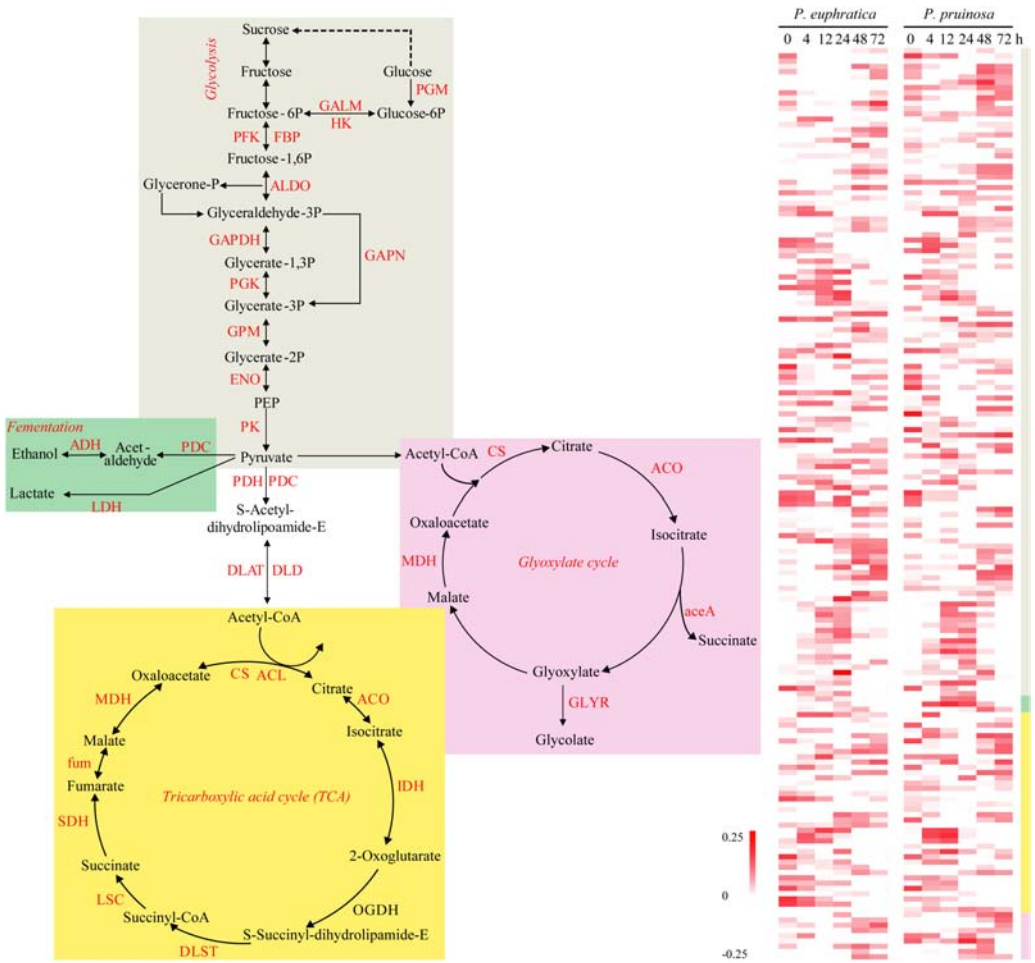


FIGURE 2 | The energy production during seed germination and its possible relationship to early transcriptome changes in *P. euphratica* and *P. pruinosa*. Glycolysis (gray region), TCA cycle (yellow region) and fermentation respiration (green region) are the common pathways associated with ATP production. Glyoxylate (pink region) is also an important energy source for the germinating seed via lipid metabolism. The colors behind the heat map represent expression patterns of the counterpart genes in the four pathways. For details of abbreviations, see **Supplementary Table S3**.

Hormonal Regulation of Seed Germination in *P. euphratica* and *P. pruinosa* Under Salt Stress

In our study, 100 genes associated with "plant hormone signal transduction" were differentially expressed over time in the two species. We identified the key hormone signal transduction genes and further compared the expression profiles of the multi-step signaling pathways of ABA, GA and ethylene (**Figure 3** and **Supplementary Figure S7**). Genes related to ABA signal transduction, e.g., PYL/PYR1, the negative regulator PP2C and the positive regulator SnRK2, exhibited similar expression patterns in the two poplars. The expression level of PP2C was high at 0 and 4 h but decreased after 12 h. In the GA signaling pathway, DEGs exhibited different regulatory expression patterns between the species during germination under salt stress. Specifically, the DELLA protein expression was upregulated from 0 to 12 h in *P. euphratica* but was continuously high level in *P. pruinosa*. GID1 was strongly upregulated during the middle and late phases of seed germination, while the expression of specific genes differed between the two species. Furthermore, most GA signal transcription-related genes were upregulated in the middle and late phases. We also identified the genes involved in ethylene signaling, as shown **Figure 3**. The expression pattern analysis indicated that most of the DEGs exhibited similar expression patterns in the two species for ETR and EIN3 (**Supplementary Figure S7**). CTR expression was upregulated in the late phase, while ETR was highly expressed after the early phase of germination in the two species.

Transcription Factors and Genes Involved in Salt Responses During Seed Germination

Numerous transcription factors that regulate the response to salt stress in desert poplars have been identified (Trapnell et al., 2010). Here, a total of 1582 and 1573 expressed transcripts were categorized as transcription factors in *P. euphratica* and

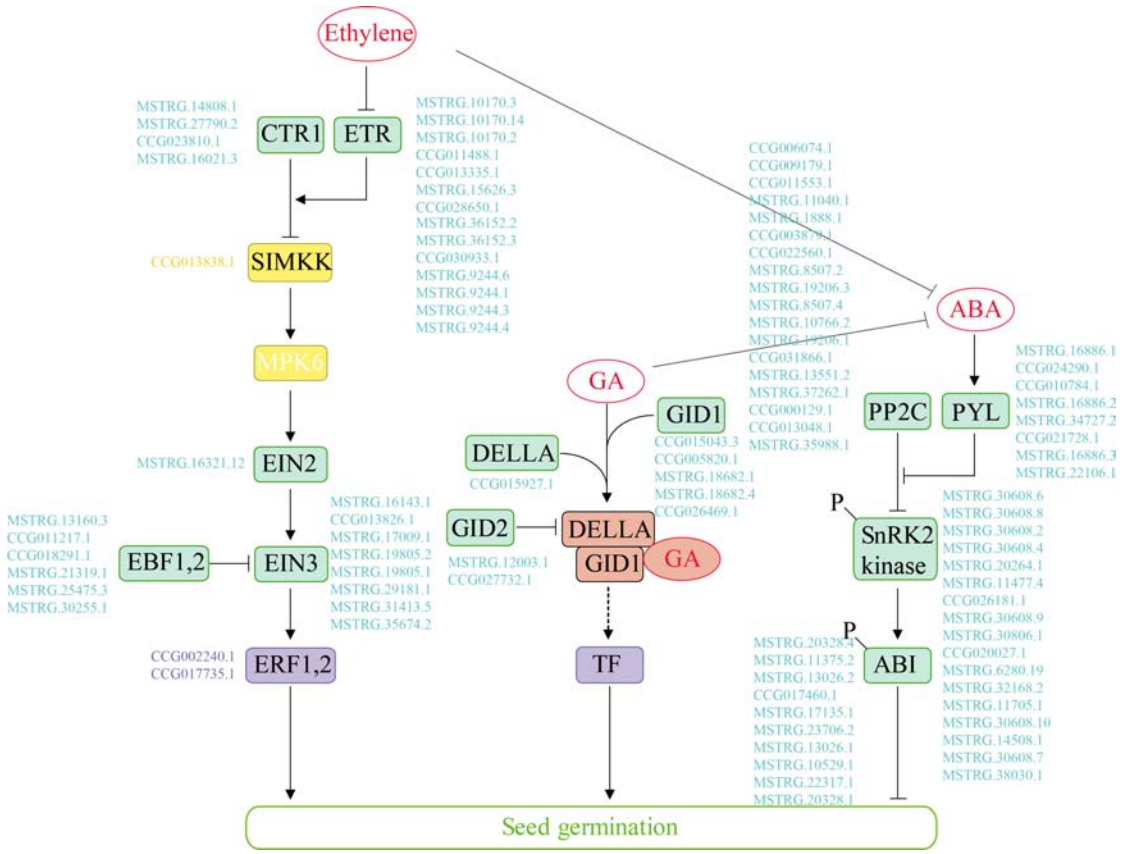


FIGURE 3 | The Ethylene, GA and ABA major metabolic pathways in the germinating seeds of *P. euphratica* and *P. pruinosa* under salt stress. Promotion or inhibition is indicated by thick arrows and blocks, respectively. Enzymes and genes are shown. A MAPK cascade (yellow box) consisted of SIMKK and MPK6. Examples of transcription factors (purple box) and other regulators (green box) contributing to these processes are presented. Genes are indicated with blue letters. The GA-GID1-DELLA complex (red box) is recognized by an F-box protein (GID2) for ubiquitination in the SCF complex (included in the dotted arrow). For details of abbreviations, see **Supplementary Table S3**.

P. pruinosa, respectively (**Figure 4A**). In total, 1480 transcription factors were expressed in both species (**Figure 4B**). Relatively few genes displayed species-specific expression. *MYBs*, *bZIPs*, *WRKY*, and *ERF*, as key response factors to abiotic stresses, were all induced by salt stress (**Figure 4B**), and the changes in their expression dynamics may reveal their critical functions in response to salt stress (Yamaguchi-Shinozaki and Shinozaki, 2005). Furthermore, some proteins regulating Na^+/H^+ transport and controlling ion homeostasis, such as NHXs, SOS1, SOS2, SOS3, and HKTs, were induced by salt stress (**Figure 4**). These results confirm that the genes related to ion transport and chloride channels play vital roles in maintaining and re-establishing homeostasis in the cytoplasm (Hasegawa et al., 2000; Wang et al., 2008; Sun et al., 2009; Ye et al., 2009; Qiu et al., 2011). *BCH1* and *ZEP*, which are involved in the biosynthesis of ABA, were highly upregulated in salt-stressed samples in the two species (**Figure 4B**). In addition, the expression of BADH and GolS, which are involved in critical solute biosynthesis processes that help plants maintain high osmotic pressure under salt stress. (Taji et al., 2002; Bartels and Sunkar, 2005), was induced by the salt treatment. Nevertheless, the expression patterns of genes responding to salt stress in *P. euphratica* were consistent with those in *P. pruinosa*, indicating there is extensive transcriptional consistency in the two species with respect to their responses to salt stress.

GO Functional Enrichment Between the Two Species Over the Time Series

The temporal expression trends of DEGs between the two species during germination were obviously different, suggesting that the two desert poplars might have evolved different gene expression patterns to adapt to different salty desert habitats. To obtain a better view of the functional differences between the species over the course of germination, GO enrichment analysis was employed, comparing the two species in two phases (the middle phase had only one DEG) (**Supplementary Figure S8**). The results indicated that in the early phase, 2766 DEGs were mainly enriched, and these DEGs were associated with the functional classifications "ribosomes," "amide biosynthetic process," "cellular macromolecule biosynthetic process," "protein activity," and "ATP binding." In the late phase, 5305 enriched DEGs

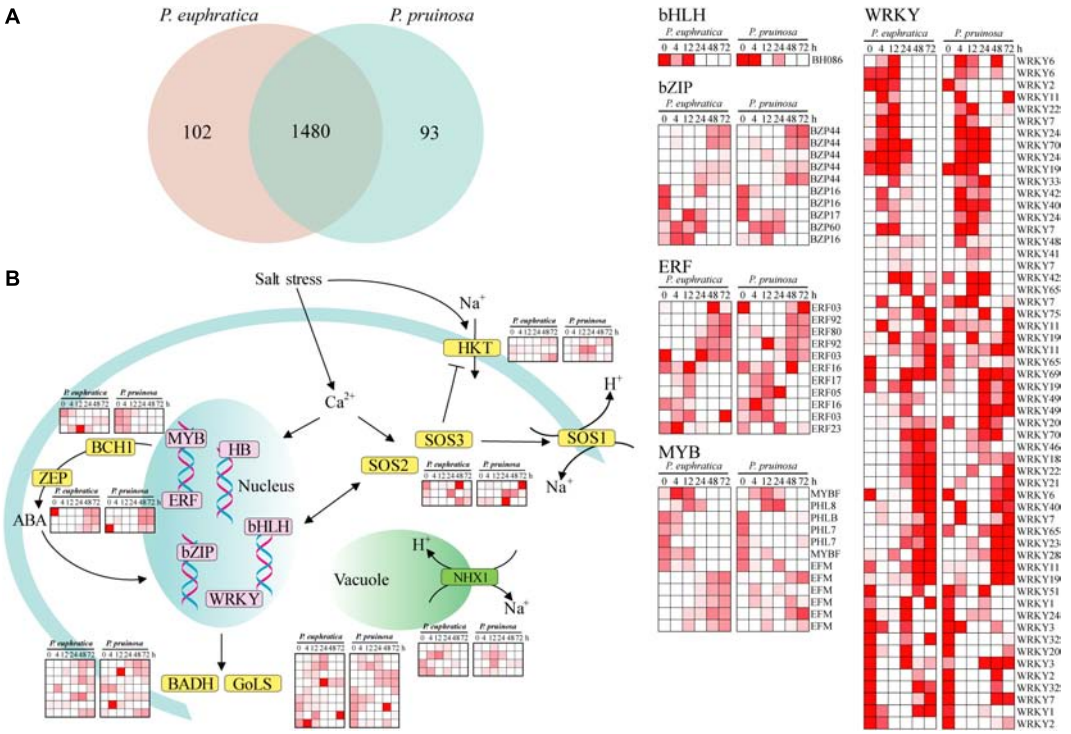


FIGURE 4 | The current known components and relationships of transcription factors related to salt response in *P. euphratica* and *P. pruinosa*. **(A)** Venn diagram showing overlaps between the transcription factors of *P. euphratica* and *P. pruinosa*. **(B)** The expanded and positively selected genes in the salt response pathways of *P. euphratica* and *P. pruinosa* (yellow). Expression of the differently expressed transcription factors of *P. euphratica* and *P. pruinosa* (pink). The heatmap was generated from hierarchical cluster analysis of genes.

were associated with "response to stress," "response to oxidative stress," "response to abiotic stress," "response to stimulus," "ATP metabolic process," "photosystem," "photosynthesis," "growth," "developmental process," "ion binding," "calcium ion binding," and "oxidoreductase activity."

KEGG Functional Enrichment in the Two Species Over the Time Series

To further elucidate the different enriched biochemical pathways, DEGs of the two species were mapped into 352 pathways, 11 of which were significantly (p-value $\leq$ 0.01) enriched, including "flavonoid biosynthesis," "stilbenoid, diarylheptanoid and gingerol biosynthesis," "brassinosteroid biosynthesis," "phenylpropanoid biosynthesis," "diterpenoid biosynthesis," and "monoterpenoid biosynthesis" (**Supplementary Figure S9**). The results indicate that many antioxidants, antioxidases and secondary metabolites are involved in the adaptation to salt stresses by these two species (Burritt and Mackenzie, 2003). The second metabolite in the flavonoid pathway plays vital roles in stress protection, but the biosynthesis of this metabolite is regulated by key enzymes (Winkel-Shirley, 2002). In this study, *PAL* was induced at 12 h in *P. euphratica* seeds and at 48 h in *P. pruinosa* seeds (**Figure 5**). CHS, whose five gene copies had different expression patterns between the two species, initiated flavonoid biosynthesis. Furthermore, the FLS expression in *P. pruinosa* was higher than that in *P. euphratica*. Specifically, FLS was highly expressed in the early phase in *P. euphratica* and was significantly and highly expressed during the seed germination process. In addition, the expression levels of F3′5′H and CHS in *P. pruinosa* were significantly higher than those in *P. euphratica* (**Figure 5**). F3H converts naringenin to dihydrokaempferol which is further converted to kaempferol and quercetin by FLS. The duplication of FLS may allow the ability to diversify the types and amounts of flavonols produced in different tissues and under different stresses (Winkel-Shirley, 2002).

Brassinosteroids are involved in an extensive range of effects, such as cell division, cell expansion, xylem differentiation and seed germination, in plants (Kagale et al., 2007). In the two species, six gene families related to brassinosteroid metabolism were enriched, including DET2, DWF4, BR6OX1, BRox2, ROT3 and BAS1. Among them, *DET2* and *BAS1* were highly expressed in *P. euphratica* and exhibited relatively low expression in *P. pruinosa*, while *ROT3* was highly expressed in *P. pruinosa* but was not detected in *P. euphratica* (**Figure 6**). Moreover, there were two copies of both *DWF4* and *BR6OX2*, and each copy exhibited a different expression pattern between the two poplars.

ROS Level and Enzyme Activity Determination

We measured ROS levels and related enzyme activities. The quantification assay indicated that more hydrogen peroxide

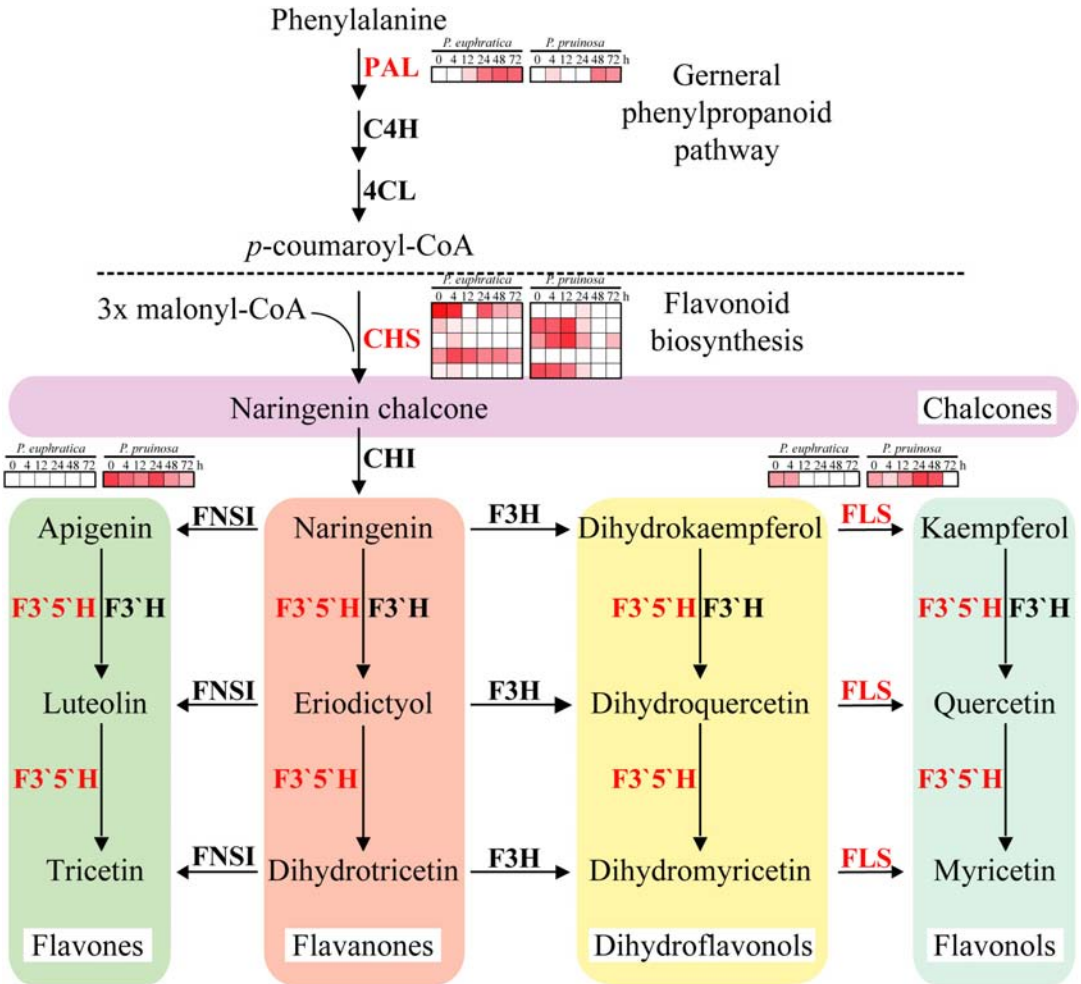


FIGURE 5 | Regulatory network of flavonoid biosynthesis underlying the co-regulated DEGs in *P. euphratica* and *P. pruinosa*. For details of abbreviations, see **Supplementary Table S3**.

accumulated in *P. euphratica* than in *P. pruinosa* under the various salt conditions, especially in 1.0% NaCl, where the levels were approximately 2-fold higher in *P. euphratica* than in *P. pruinosa* (**Figure 7**). Therefore, SOD activities were significantly higher in *P. pruinosa* than in *P. euphratica* after treatment with 0.4% NaCl solution. The CAT activities in the treatment with 1.0% NaCl solution demonstrated a similar pattern.

Verification of Expression Patterns by qRT-PCR

To validate the RNA-seq results, qRT-PCR analysis was conducted at different time points during seed germination in the two species (**Figure 8**). The results of genes studied by the RT-PCR analysis, including those in the plant hormone signal transduction (*PYL* and *GID1*), flavonoid biosynthesis (*PAL*) and brassinosteroid biosynthesis (*ROT3* and *DWF4*) pathways, were all similar to the RNA-seq results.

DISCUSSION

P. pruinosa Showed a Higher Salt Tolerance Than *P. euphratica* at the Three Seed Germination Stages

Populus euphratica and *P. pruinosa* diverged from a recent common ancestor between 1 and 2 million years ago (Wang J. et al., 2011) and exhibited different ecological adaptations to desert habitats. In this context, the two desert poplars have evolved different genetic strategies (Ma et al., 2013; Zhang et al., 2013). However, it is not known whether these genetic variations also underlie differences in seed germination.

In the present study, the rate of seed germination in *P. pruinosa* faster than that in *P. euphratica* during seed germination (**Figure 1C**). Based on the seed moisture content, the seed germination time courses for the two species, upon the transfer of seeds to water, can be divided into three phases, which agree with the three classical phases of seed

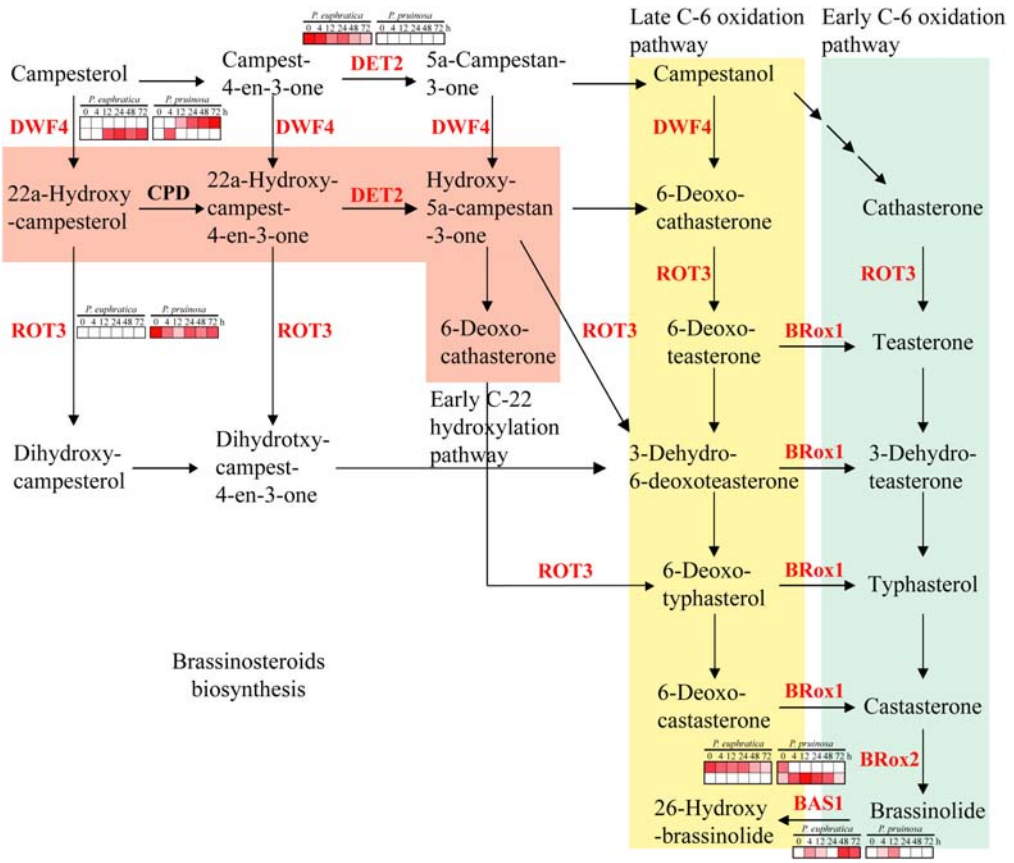


FIGURE 6 | Regulatory network of brassinosteroid biosynthesis underlying the co-regulated DEGs in *P. euphratica* and *P. pruinosa*. For details of abbreviations, see **Supplementary Table S3**.

germination (Nonogaki et al., 2010; **Figure 1A**). We also investigated the relationship between the germination rate and NaCl concentration. The average germination percentage, subordinate function values, and threshold salinity of *P. pruinosa* were higher than those of *P. euphratica*. Based on transcriptome analysis, approximately 80% of the reads matched with the genome, and the number of mapped genes in each library was 52%–70%, indicating that most genes were expressed in the seeds of the two species under salt stress. The correlation dendrogram is consistent with the separate phases classified by seed water uptake (**Supplementary Figure S3B**), suggesting that, in two stages (early and late phases), the sister species evolved divergent regulatory and metabolic pathways associated with seed germination in different salt habitats.

Biochemical Processes of Poplar Seeds Are Regulated by Highly Coordinated Transcript Dynamics

The expression data in the three seed germination phases showed a high reproducibility in both species, and each phase was clearly distinguished by expression dynamics. The DEGs induced early in seed germination (early phase) appeared to be associated with the repair of genetic materials, the cellular structure and the resumption of energy metabolism. During seed germination, the free amino acids involved in protein synthesis are provided by storage protein degradation induced by osmopriming in the first hours of imbibition (Wang et al., 2012). Accordingly, proteases are newly synthesized and accumulate during imbibition (Yang et al., 2007). Therefore, we speculate that in *P. euphratica* and *P. pruinosa*, amino acid biosynthesis genes are expressed after 48 h of the seed germination process, and their products provide for the synthesis and metabolism of *de novo* proteins in the growing embryo (Joosen et al., 2013). Thus, the stored proteins in seeds act not only as important sources of amino acids but also as a source of energy (Angelovici et al., 2011). The middle phase was associated with the active nutrient reservoir, amino acid metabolism and catalytic activity. In this stage, producing a redox state is likely a primary function of the fast recovery of cellular metabolism at the beginning of imbibition (Rosental et al., 2014). The functions of the enriched genes not only produce energy but also promote the activity of essential enzymes to support the completion of germination (Van Dongen et al., 2011). Moreover, flavonoids can induce a delay in the germination rate and play important roles in protection against diverse stresses (D'Auria and Gershenzon, 2005). Most flavonoid genes in this study were enhanced in *P. pruinosa* in the middle phase, indicating that the "flavonoid biosynthesis"

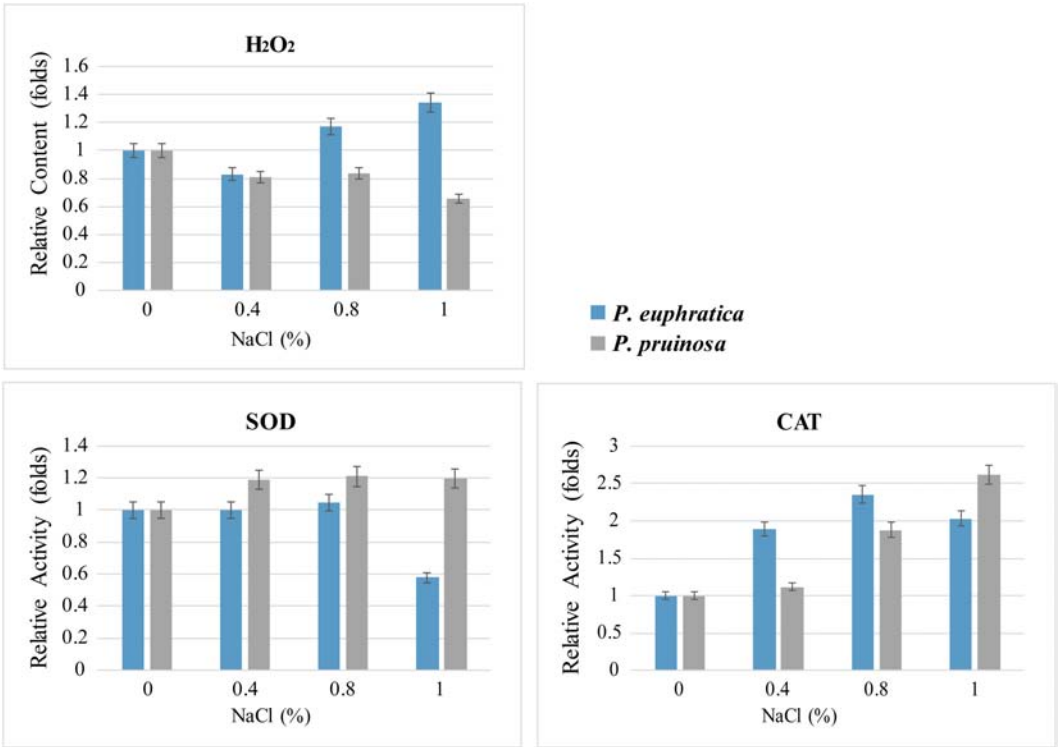


FIGURE 7 | Quantitative comparison of superoxide contents and antioxidant enzyme activities (SOD and CAT) for the two species exposed to salt stress. The changes in content and activity were analyzed with different salt treatments.

pathway might lead to a difference in the seed germination rate between *P. pruinosa* and *P. euphratica*.

Reserves used for the germination of seeds are primarily stored in the form of starch, lipids and proteins in the embryo or endosperm (Yang et al., 2009). Proteins related to hydrolase activity contribute to starch and protein degradation (Rosental et al., 2014), while catalytic activity proteins may increase enzyme activities or provide the energy required during seed germination. Nitrogen-containing compounds release seeds from dormancy, presumably leading to the oxidation of NADPH and therefore providing an increased carbon flow through the glycolytic and oxidative pentose phosphate pathways (PPP) (Roberts, 1964; Hendricks and Taylorson, 1974; Roberts and Lord, 1979; Cohn et al., 1983; Hilhorst and Karssen, 1989). NADP, as a coenzyme of glucose-6-phosphate dehydrogenase, plays a key role in linking the glycolysis pathway and PPP (Nonogaki et al., 2010). Here, we investigated the DEGs involved in fermentation, the TCA cycle, glyoxylate and glycolysis during seed germination. We found that energy production mainly occurred in the later phase and increased gradually. Interestingly, many genes of key enzymes in the TCA cycle were expressed during the early phase, which would lead to the accumulation of many key enzymes during early germination (Weitbrecht et al., 2011). In the late phase, some of the enriched genes were linked to "the photosystem II oxygen evolving complex," "photosystem," and "photosynthesis," suggesting that the seeds had already started to photosynthesize, contributing to the energy supply and powering the productivity of the seed (Ruuska et al., 2004; Goffman et al., 2005; Allen et al., 2009). Meanwhile, "glyoxylate metabolism" activity was enriched in the last phase of germination, which suggests that lipid metabolism is also an important energy source for seed germination. The results demonstrated that activation of energy metabolism during early germination is necessary for seed germination, however, energy production is more complex in the late phase than in earlier phases. In addition to energy metabolism, the expression of genes associated with detoxification were also involved in responses to salt stress in the two desert poplars (Zhu, 2001); these genes included genes associated with "glutathione metabolism," "flavonoid biosynthesis" and "cytochrome P450," which all are tightly correlated with seed tolerance to salt stress (Alscher, 1989). These observations suggest that *P. euphratica* and *P. pruinosa* quickly establish energetic and developmental balances under salt stress. Germination is actuated by a large number of cellular processes, such as transcription, translation, repair mechanisms, responses to various stresses, organelle reassembly and cellular structure reconstruction. All the processes are supported by metabolism for energy generation. Together, the above results indicate that the transition of primary biochemical processes over time during the seed germination of the two studied poplars is produced partly by highly coordinated transcript dynamics. As the two desert poplars have adapted to different salty desert habitats, these species may have developed different genetic pathways under salt stress during seed germination.

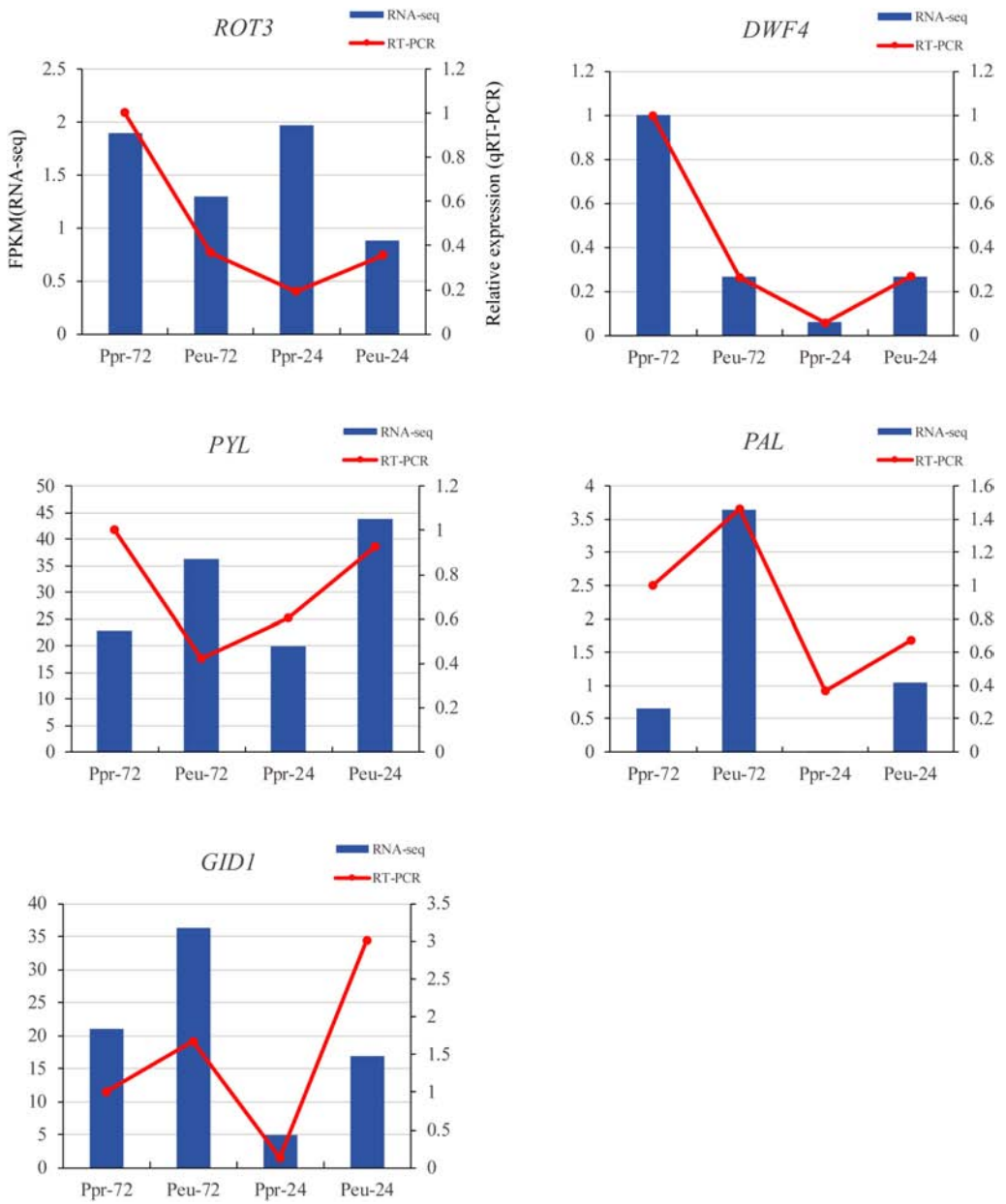


FIGURE 8 | qRT-PCR verification of five selected DEGs. We compared the RNA-seq data (blue bar) with qRT-PCR data (red lines). And we indicated the normalized expression level (FPKM) of RNA-seq on the y-axis to the left. The relative qRT-PCR expression level is shown on the y-axis to the right. *CYC063* was set as the internal control. Both methods agree, showing similar gene expression trends.

Hormonal Regulation Contributes to the Difference in Seed Germination Phases

Some genes were enriched in the "plant hormone signal transduction" functional category, which is key for physiological state determination and the regulation of seed germination, especially the GA-ABA balance (Meyer et al., 2009). ABA positively regulates the induction of dormancy and negatively regulates germination. Here, the genes related to ABA signal transduction exhibited similar expression patterns in the two poplars. For example, *PYL*/*PYR1*, which are considered ABA receptors, exhibited upregulated expression during the first stage, suggesting that the ABA content of the dry seeds was high and decreased during imbibition (Preston et al., 2009). In addition, the negative regulator PP2C has been found to be a major core component of ABA signaling; its expression level was high at 0 and 4 h but decreased after 12 h (Fujii and Zhu, 2009; Umezawa et al., 2009).

GAs play an important role in the promotion of germination and the release of dormancy (Kucera et al., 2005) by stimulating ABA degradation. Here, many DEGs associated with the GA

signaling pathway exhibited different expression patterns during germination under salt stress. Specifically, DELLA proteins belonging to the GRAS family were negatively regulated in the GA signaling pathway (Sun and Gubler, 2004) and upregulated from 0 to 12 h in *P. euphratica*, while they were continuously expressed at high levels in *P. pruinosa*. *GID1*, coding a soluble GA receptor, was strongly upregulated during the middle and late phases of seed germination. The GA protein can interact with DELLA when bioactive GAs are present (Ueguchi-Tanaka et al., 2007). Furthermore, most GA signaling transcription-related genes were upregulated in the middle and late phases, which corresponds to the results of a previous study showing that the GA content increased during germination in seeds during phase II.

Ethylene is implicated in the promotion of germination in many species. Here, we identified the DEGs involved in ethylene signaling in seed germination. We found that most of these DEGs, alongside *ETR* and *EIN3*, exhibited similar expression patterns in the two species (**Supplementary Figure S7**). In the absence of ethylene, ETR1 activates CTR1, which negatively regulates downstream signaling components and is inactive in the presence of ethylene. *CTR* expression was upregulated in the late phase of germination in the two species, while ETR was highly expressed after the early phase of germination. These proteins are regulated by ethylene levels during seed germination by the inactivation of a MAPK cascade comprising SIMKK and MPK6, which are positive regulators of the ethylene response pathway (Ouaked et al., 2003). EIN3 and EIN3-LIKE proteins bind to the promoter of the *ERF1* (ethylene responsive factor 1) gene and thereby confer a hierarchy of transcription factors involved in ethylene signaling (Lee and Kim, 2003). Most importantly, the expression patterns of DEGs related to the ethylene pathway were different between *P. euphratica* and *P. pruinosa*, indicating that ETR and EIN3 distinctly regulate ethylene signal transcription pathways during seed germination.

Overall, GAs increase and counteract ABA inhibition in the early and late phases of germination (North et al., 2010). Ethylene counteracts ABA inhibition by interfering with ABA signaling during the late phase of germination, while the ABA content is regulated by an equilibrium between the biosynthesis and catabolism of ABA (Nambara and Marion-Poll, 2005). Thus, many of the DEGs exhibited analogous expression patterns in the two species in the models for GA, ABA and ethylene in response to salinity stress but exhibited completely different expression patterns during seed germination.

The Fine Regulation of the Synthesis of Flavonoids and Brassinosteroids in Desert Poplars Contributes to Their Environmental Adaptation

Flavonoids have an extensive range of biological functions, including protecting plants under various stresses (Winkel-Shirley, 2002). Flavonoids are synthesized by the phenylpropanoid pathway and found in most seeds and grains; the major types of flavonoids in seeds are flavonols, anthocyanins, phlobaphenes, isoflavones and proanthocyanidins (Lepiniec et al., 2006). Several genes that encode key enzymes in the flavonoid biosynthetic pathway were expressed differently between the seeds of *P. euphratica* and *P. pruinosa* under salt stress (**Figure 5**). We suggest that the phenylpropanoid pathway, especially the flavonoid metabolism pathway, is widely involved in protection from salt stress in both desert poplars. In general, salt stress is often accompanied by an oxidative burst in plants. In this study, the hydrogen peroxide (H_2O_2) accumulation in *P. euphratica* was 2-fold higher than that in *P. pruinosa* under the various salt conditions, especially in 1.0% NaCl (**Figure 7**), suggesting that salt treatment might induce oxidative stress in the seeds of *P. euphratica*. The unavoidable accumulation of H_2O_2 and scavenging pathways activity should be maintained in balance, where H_2O_2 could either perform a signaling role or reach a nontoxic level in plants under salt stress conditions. To alleviate and eliminate highly reactive oxygen species, plants have evolved a battery of antioxidative mechanisms, and the antioxidant defense system includes hydrophilic and hydrophobic antioxidants and enzymes such as SOD and CAT (Shalata and Tal, 1998). SOD activities in *P. pruinosa* were significantly higher than those in *P. euphratica* when the seeds were exposed to concentrations of NaCl above 0.4%, while CAT activities in *P. pruinosa* were also higher than those in *P. pruinosa* when seeds were treated with 1.0% NaCl. Both antioxidases could play a crucial role in scavenging redundant ROS (H_2O_2) induced by salt stress. Altogether, a significant proportion of the antioxidants induced by salt stress were secondary metabolites, such as a vast amount of compounds primarily derived by the phenylpropanoid pathway (Dixon and Paiva, 1995).

Brassinosteroids are involved in a wide range of growth and development aspects in plants (Kagale et al., 2007). One of the most interesting influences of brassinosteroids is their ability to confer resistance to various abiotic stresses. Several brassinosteroid biosynthesis genes have been identified by molecular genetic analysis and reverse genetic analysis (Takahashi et al., 2005). Among the gene families enriched in the brassinosteroid pathway, *DET2* and *BAS1* were highly expressed in *P. euphratica* and exhibited relatively low expression in *P. pruinosa*, while *ROT3* was highly expressed in *P. pruinosa* but was not detected in *P. euphratica* (**Figures 6**, **8**). Moreover, *DWF4* and *BR6OX2* each contain two copies, and each copy exhibited a different expression pattern between the two poplars. The results suggest that the fine regulation of the synthesis of brassinosteroids in desert poplars contributes to their environmental adaptation.

CONCLUSION

In this study, a multidimensional transcriptome dataset allowed us to discern highly dynamic and coordinated gene expression, as well as functional and regulatory shifts exhibited by the germinating seeds of two species in response to continuous salinity stress. Based on these results, we conclude that the fine regulation of the synthesis of flavonoids and brassinosteroids in desert poplars contributes to their environmental adaptation.

DATA AVAILABILITY

The Illumina sequencing data sets are available at the NCBI Sequence Read Archive (SRA) database with the project accession number: PRJNA484685.

AUTHOR CONTRIBUTIONS

DW conceived and designed the experiments. CZ, WL, and YL conducted the bioinformatic work and wrote the manuscript. XuZ, XB, and ZN contributed to conducting experiments for physiology and transcript analysis. XiZ and ZL provided assistance in sample collection. All authors read, revised and approved the final manuscript.

FUNDING

This research was supported by the National Science Foundation of China (Nos. 31470620 and 31870580).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00231/full#supplementary-material

FIGURES S1, S2 | Reproducibility of each trio of biological replicates. The samples were collected at different time points, and total RNA isolation was used to construct RNA-seq libraries for them independently. FPKM values of all the genes expressed in at least one of the 36 sequenced samples are shown in scatter plots and were used as input for the Pearson product-moment correlation coefficient analysis. The correlations between the biological replicates were high in both two species [average $r = 0.945$ in *P. euphratica* (**Supplementary Figure S1**) and $r = 0.939$ in *P. pruinosa* (**Supplementary Figure S2**)].

FIGURE S3 | Number of genes expressed at each time point **(A)** and hierarchical clustering of six time points for *P. euphratica* and *P. pruinosa* **(B)**.

FIGURE S4 | Hierarchical clustering of six time points for *P. euphratica* and *P. pruinosa*. A to B, the expression patterns of co-expression modules of *P. euphratica* **(A)** and *P. pruinosa* **(B)**, ordered according to the sample time points of their peak expression. **(C)** The gene numbers and the expression fitted curves of all the modules in A and B. For each gene, the FPKM value normalized by the maximum value of all FPKM values of the gene over all time points is shown.

FIGURE S5 | GO function enrichment of the DEGs for seed germination processes.

FIGURE S6 | KEGG function enrichment of the DEGs for seed germination processes.

FIGURE S7 | The expression pattern of the hormone-related genes in the two poplars. Expression patterns of hormone-related genes. Normalized expression levels of genes related to ethylene, GA and ABA are shown.

FIGURE S8 | GO function enrichment of the DEGs for salt tolerance variety of the two species.

FIGURE S9 | KEGG function enrichment of the DEGs for salt tolerance variety of the two species.

TABLE S1 | Overview of the data size of all the samples of *P. euphratica* and *P. pruinosa*.

TABLE S2 | Summary of the illumine sequencing reads and the matches in the *P. euphratica* and *P. pruinosa*.

TABLE S3 | The details of abbreviations.

TABLE S4 | Primer used for real-time quantitative PCR in this study.

REFERENCES

Allen, D. K., Ohlrogge, J. B., and Shachar-Hill, Y. (2009). The role of light in soybean seed filling metabolism. *Plant J.* 58, 220–234. doi: 10.1111/j.1365-313X.2008.03771.x

Almansouri, M., Kinet, J.-M., and Lutts, S. (2001). Effect of salt and osmotic stresses on germination in durum wheat (*Triticum durum* Desf.). *Plant Soil* 231, 243–254. doi: 10.1023/A:1010378409663

Alscher, R. G. (1989). Biosynthesis and antioxidant function of glutathione in plants. *Physiol. Plant.* 77, 457–464. doi: 10.1111/j.1399-3054.1989.tb05667.x

Angelovici, R., Fait, A., Fernie, A. R., and Galili, G. (2011). A seed high-lysine trait is negatively associated with the TCA cycle and slows down *Arabidopsis* seed germination. *New Phytol.* 189, 148–159. doi: 10.1111/j.1469-8137.2010.03478.x

Annunziata, M. G., Ciarmiello, L. F., Woodrow, P., Maximova, E., Fuggi, A., and Carillo, P. (2017). Durum wheat roots adapt to salinity remodeling the cellular content of nitrogen metabolites and sucrose. *Front. Plant Sci.* 7:2035. doi: 10.3389/fpls.2016.02035

Arbona, V., Manzi, M., Ollas, C. D., and Gómez-Cadenas, A. (2013). Metabolomics as a tool to investigate abiotic stress tolerance in plants. *Int. J. Mol. Sci.* 14, 4885–4911. doi: 10.3390/ijms14034885

Bartels, D., and Sunkar, R. (2005). Drought and salt tolerance in plants. *Crit. Rev. Plant Sci.* 24, 23–58. doi: 10.1080/07352680590910410

Bewley, J. (1982). "Protein and nucleic acid synthesis during seed germination and early seedling growth," in *Nucleic Acids and Proteins in Plants 1. Encyclopedia of Plant Physiology*, eds D. Boulter and B. Parthier (New York, NY: Springer), 559–591.

Bewley, J., and Black, M. (1984). Physiology and biochemistry of seeds in relation to germination. *Plant Ecol.* 57, 71–74.

Bewley, J. D. (1997). Seed germination and dormancy. *Plant Cell* 96, 1055–1066. doi: 10.1105/tpc.9.7.1055

Bewley, J. D., Bradford, K., and Hilhorst, H. (2012). *Seeds: Physiology of Development, Germination and Dormancy*. New York, NY: Springer.

Biligetu, B., Schellenberg, M., Mcleod, J., Borges, E., Borges, R., Buckeridge, M., et al. (2011). Seeds: physiology of development and germination. *Alterações Fisiológicas E Bioquímicas Durante A* 39:26.

Bradford, K. J. (1990). A water relations analysis of seed germination rates. *Plant Physiol.* 94, 840–849. doi: 10.1104/pp.94.2.840

Brinker, M., Brosché, M., Vinocur, B., Abo-Ogiala, A., Fayyaz, P., Janz, D., et al. (2010). Linking the salt transcriptome with physiological responses of a salt-resistant *Populus* species as a strategy to identify genes important for stress acclimation. *Plant Physiol.* 154, 1697–1709. doi: 10.1104/pp.110.164152

Burritt, D. J., and Mackenzie, S. (2003). Antioxidant metabolism during acclimation of *Begonia× erythrophylla* to high light levels. *Ann. Bot.* 91, 783–794. doi: 10.1093/aob/mcg076

Carillo, P., Cirillo, C., De Micco, V., Arena, C., De Pascale, S., and Rouphael, Y. (2019). Morpho-anatomical, physiological and biochemical adaptive responses to saline water of *Bougainvillea spectabilis* Willd. trained to different canopy shapes. *Agric. Water Manage.* 212, 12–22. doi: 10.1016/j.agwat.2018.08.037

Chang, S., Puryear, J., and Cairney, J. (1993). A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* 11, 113–116. doi: 10.1007/BF02670468

Chen, S., Jiang, J., Li, H., and Liu, G. (2012). The salt-responsive transcriptome of *Populus simonii× Populus nigra* via DGE. *Gene* 504, 203–212. doi: 10.1016/j.gene.2012.05.023

Chen, S., Li, J., Fritz, E., Wang, S., and Hüttermann, A. (2002). Sodium and chloride distribution in roots and transport in three poplar genotypes under increasing NaCl stress. *For. Ecol. Manage.* 168, 217–230. doi: 10.1016/S0378-1127(01)00743-5

Cohn, M. A., Butera, D. L., and Hughes, J. A. (1983). Seed dormancy in red rice III. Response to nitrite, nitrate, and ammonium ions. *Plant Physiol.* 73, 381–384. doi: 10.1104/pp.73.2.381

D'Auria, J. C., and Gershenzon, J. (2005). The secondary metabolism of *Arabidopsis thaliana*: growing like a weed. *Curr. Opin. Plant Biol.* 8, 308–316. doi: 10.1016/j.pbi.2005.03.012

Dixon, R. A., and Paiva, N. L. (1995). Stress-induced phenylpropanoid metabolism. *Plant Cell* 7, 1085–1097. doi: 10.1105/tpc.7.7.1085

Flowers, T. (2004). Improving crop salt tolerance. *J. Exp. Bot.* 55, 307–319. doi: 10.1093/jxb/erh003

Frazee, A. C., Pertea, G., Jaffe, A. E., Langmead, B., Salzberg, S. L., and Leek, J. T. (2015). Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.* 33, 243–246. doi: 10.1038/nbt.3172

Fujii, H., and Zhu, J.-K. (2009). *Arabidopsis* mutant deficient in 3 abscisic acid-activated protein kinases reveals critical roles in growth, reproduction, and stress. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8380–8385. doi: 10.1073/pnas.0903144106

Goffman, F. D., Alonso, A. P., Schwender, J., Shachar-Hill, Y., and Ohlrogge, J. B. (2005). Light enables a very high efficiency of carbon storage in developing embryos of rapeseed. *Plant Physiol.* 138, 2269–2279. doi: 10.1104/pp.105.063628

Hasegawa, P. M., Bressan, R. A., Zhu, J. K., and Bohnert, H. J. (2000). Plant cellular and molecular responses to high salinity. *Annu. Rev. Plant Biol.* 51, 463–499. doi: 10.1146/annurev.arplant.51.1.463

Hendricks, S., and Taylorson, R. (1974). Promotion of seed germination by nitrate, nitrite, hydroxylamine, and ammonium salts. *Plant Physiol.* 54, 304–309. doi: 10.1104/pp.54.3.304

Hilhorst, H. W., and Karssen, C. M. (1989). Nitrate reductase independent stimulation of seed germination in *Sisymbrium officinale* L. (hedge mustard) by light and nitrate. *Ann. Bot.* 63, 131–137. doi: 10.1093/oxfordjournals.aob.a087715

Holdsworth, M. J., Bentsink, L., and Soppe, W. J. (2008). Molecular networks regulating *Arabidopsis* seed maturation, after-ripening, dormancy and germination. *New Phytol.* 179, 33–54. doi: 10.1111/j.1469-8137.2008.02437.x

Hukin, D., Cochard, H., Dreyer, E., Le Thiec, D., and Bogeat, M. B. (2005). Cavitation vulnerability in roots and shoots: does *Populus euphratica* Oliv., a poplar from arid areas of Central Asia, differ from other poplar species? *J. Exp. Bot.* 56, 2003–2010. doi: 10.1093/jxb/eri198

Imit, Y., Maimaiti, A., Taxi, Z., and Cyffka, B. (2015). Seed germination characteristics of *Populus euphratica* from different provenances under NaCl stess. *J. Northwest For. Univ.* 30, 88–94. doi: 10.3969/j.issn.1001-7461.2015.06.15

Janz, D., Behnke, K., Schnitzler, J.-P., Kanawati, B., Schmitt, P., and Polle, A. (2010). Pathway analysis of the transcriptome and metabolome of salt sensitive and tolerant poplar species reveals evolutionary adaption of stress tolerance mechanisms. *BMC Plant Biol.* 10:150. doi: 10.1186/1471-2229-10-150

Joosen, R. V. L., Arends, D., Li, Y., Willems, L. A., Keurentjes, J. J., Ligterink, W., et al. (2013). Identifying genotype-by-environment interactions in the metabolism of germinating *Arabidopsis* seeds using generalized genetical genomics. *Plant Physiol.* 162, 553–566. doi: 10.1104/pp.113.216176

Kagale, S., Divi, U. K., Krochko, J. E., Keller, W. A., and Krishna, P. (2007). Brassinosteroid confers tolerance in *Arabidopsis thaliana* and *Brassica napus* to a range of abiotic stresses. *Planta* 225, 353–364. doi: 10.1007/s00425-006-0361-6

Kaya, M. D., Ipek, A., and Öztürk, A. (2003). Effects of different soil salinity levels on germination and seedling growth of safflower (*Carthamus tinctorius* L.). *Turk. J. Agric. For.* 27, 221–227.

Khajeh-Hosseini, M., Powell, A., and Bingham, I. (2003). The interaction between salinity stress and seed vigour during germination of soyabean seeds. *Seed Sci. Technol.* 31, 715–725. doi: 10.15258/sst.2003.31.3.20

Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317

Kucera, B., Cohn, M. A., and Leubner, G. (2005). Plant hormone interactions during seed dormancy release and germination. *Seed Sci. Res.* 15, 281–307. doi: 10.1079/SSR2005218

Lee, J. H., and Kim, W. T. (2003). Molecular and biochemical characterization of VR-EILs encoding mung bean ETHYLENE INSENSITIVE3-LIKE proteins. *Plant Physiol.* 132, 1475–1488. doi: 10.1104/pp.103.022574

Lepiniec, L., Debeaujon, I., Routaboul, J. M., Baudry, A., Pourcel, L., Nesi, N., et al. (2006). Genetics and biochemistry of seed flavonoids. *Annu. Rev. Plant Biol.* 57, 405–430. doi: 10.1146/annurev.arplant.57.032905.105252

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2- ΔΔCT method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Ma, J., He, X., Bai, X., Niu, Z., Duan, B., Chen, N., et al. (2016). Genome-wide survey reveals transcriptional differences underlying the contrasting trichome phenotypes of two sister desert poplars. *Genes* 7:111. doi: 10.3390/genes7120111

Ma, T., Wang, J., Zhou, G., Yue, Z., Hu, Q., Chen, Y., et al. (2013). Genomic insights into salt adaptation in a desert poplar. *Nat. Commun.* 4:2797. doi: 10.1038/ncomms3797

Ma, T., Wang, K., Hu, Q., Xi, Z., Wan, D., Wang, Q., et al. (2018). Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. *Proc. Natl. Acad. Sci. U.S.A.* 115, E236–E243. doi: 10.1073/pnas.1713288114

Meyer, E. H., Tomaz, T., Carroll, A. J., Estavillo, G., Delannoy, E., Tanz, S. K., et al. (2009). Remodeled respiration in ndufs4 with low phosphorylation efficiency suppresses *Arabidopsis* germination and growth and alters control of metabolism at night. *Plant Physiol.* 151, 603–619. doi: 10.1104/pp.109.141770

Müller, K., Carstens, A. C., Linkies, A., Torres, M. A., and Leubner, G. (2009). The NADPH-oxidase AtrbohB plays a role in *Arabidopsis* seed after-ripening. *New Phytol.* 184, 885–897. doi: 10.1111/j.1469-8137.2009.03005.x

Munns, R. (2002). Comparative physiology of salt and water stress. *Plant Cell Environ.* 25, 239–250. doi: 10.1046/j.0016-8025.2001.00808.x

Murillo-Amador, B., López-Aguilar, R., Kaya, C., Larrinaga-Mayoral, J., and Flores-Hernández, A. (2002). Comparative effects of NaCl and polyethylene glycol on germination, emergence and seedling growth of cowpea. *J. Agron. Crop Sci.* 188, 235–247. doi: 10.1046/j.1439-037X.2002.00563.x

Nambara, E., and Marion-Poll, A. (2005). Abscisic acid biosynthesis and catabolism. *Annu. Rev. Plant Biol.* 56, 165–185. doi: 10.1146/annurev.arplant.56.032604.144046

Nonogaki, H., Bassel, G. W., and Bewley, J. D. (2010). Germination-still a mystery. *Plant Sci.* 179, 574–581. doi: 10.1016/j.plantsci.2010.02.010

North, H., Baud, S., Debeaujon, I., Dubos, C., Dubreucq, B., Grappin, P., et al. (2010). *Arabidopsis* seed secrets unravelled after a decade of genetic and omics-driven research. *Plant J.* 61, 971–981. doi: 10.1111/j.1365-313X.2009.04095.x

Ottow, E. A., Brinker, M., Teichmann, T., Fritz, E., Kaiser, W., Brosché, M., et al. (2005). *Populus euphratica* displays apoplastic sodium accumulation, osmotic adjustment by decreases in calcium and soluble carbohydrates, and develops leaf succulence under salt stress. *Plant Physiol.* 139, 1762–1772. doi: 10.1104/pp.105.069971

Ouaked, F., Rozhon, W., Lecourieux, D., and Hirt, H. (2003). A MAPK pathway mediates ethylene signaling in plants. *EMBO J.* 22, 1282–1288. doi: 10.1093/emboj/cdg131

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122

Preston, J., Tatematsu, K., Kanno, Y., Hobo, T., Kimura, M., Jikumaru, Y., et al. (2009). Temporal expression patterns of hormone metabolism genes during imbibition of *Arabidopsis thaliana* seeds: a comparative study on dormant and non-dormant accessions. *Plant Cell Physiol.* 50, 1786–1800. doi: 10.1093/pcp/pcp121

Pritchard, S. L., Charlton, W. L., Baker, A., and Graham, I. A. (2002). Germination and storage reserve mobilization are regulated independently in *Arabidopsis*. *Plant J.* 31, 639–647. doi: 10.1046/j.1365-313X.2002.01376.x

Qiu, Q., Ma, T., Hu, Q., Liu, B., Wu, Y., Zhou, H., et al. (2011). Genome-scale transcriptome analysis of the desert poplar, *Populus euphratica*. *Tree Physiol.* 31, 452–461. doi: 10.1093/treephys/tpr015

Ren, Z. H., Gao, J. P., Li, L. G., Cai, X. L., Huang, W., Chao, D. Y., et al. (2005). A rice quantitative trait locus for salt tolerance encodes a sodium transporter. *Nat. Genet.* 37, 1141–1146. doi: 10.1038/ng1643

Roberts, E. (1964). The distribution of Oxidation-reduction enzymes and the effects of respiratory inhibitors and oxidising agents on dormancy in rice seed. *Physiol. Plant.* 17, 14–29. doi: 10.1111/j.1399-3054.1964.tb09013.x

Roberts, L. M., and Lord, J. M. (1979). Developmental changes in the activity of messenger RNA isolated from germinating castor bean endosperm. *Plant Physiol.* 64, 630–634. doi: 10.1104/pp.64.4.630

Rosental, L., Nonogaki, H., and Fait, A. (2014). Activation and regulation of primary metabolism during seed germination. *Seed Sci. Res.* 24, 1–15. doi: 10.1017/S0960258513000391

Ruuska, S. A., Schwender, J., and Ohlrogge, J. B. (2004). The capacity of green oilseeds to utilize photosynthesis to drive biosynthetic processes. *Plant Physiol.* 136, 2700–2709. doi: 10.1104/pp.104.047977

Shalata, A., and Tal, M. (1998). The effect of salt stress on lipid peroxidation and antioxidants in the leaf of the cultivated tomato and its wild salt-tolerant relative *Lycopersicon pennellii*. *Physiol. Plant.* 104, 169–174. doi: 10.1034/j.1399-3054.1998.1040204.x

Sun, J., Chen, S., Dai, S., Wang, R., Li, N., Shen, X., et al. (2009). NaCl-induced alternations of cellular and tissue ion fluxes in roots of salt-resistant and salt-sensitive poplar species. *Plant Physiol.* 149, 1141–1153. doi: 10.1104/pp.108.129494

Sun, T. P., and Gubler, F. (2004). Molecular mechanism of gibberellin signaling in plants. *Annu. Rev. Plant Biol.* 55, 197–223. doi: 10.1146/annurev.arplant.55.031903.141753

Taji, T., Ohsumi, C., Iuchi, S., Seki, M., Kasuga, M., Kobayashi, M., et al. (2002). Important roles of drought- and cold-inducible genes for galactinol synthase in stress tolerance in *Arabidopsis thaliana*. *Plant J.* 29, 417–426. doi: 10.1046/j.0960-7412.2001.01227.x

Takahashi, N., Nakazawa, M., Shibata, K., Yokota, T., Ishikawa, A., Suzuki, K., et al. (2005). shk1-D, a dwarf *Arabidopsis* mutant caused by activation of the CYP72C1 gene, has altered brassinosteroid levels. *Plant J.* 42, 13–22. doi: 10.1111/j.1365-313X.2005.02357.x

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511–515. doi: 10.1038/nbt.1621

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Ueguchi-Tanaka, M., Nakajima, M., Motoyuki, A., and Matsuoka, M. (2007). Gibberellin receptor and its role in gibberellin signaling in plants. *Annu. Rev. Plant Biol.* 58, 183–198. doi: 10.1146/annurev.arplant.58.032806.103830

Umezawa, T., Sugiyama, N., Mizoguchi, M., Hayashi, S., Myouga, F., Yamaguchi, K., et al. (2009). Type 2C protein phosphatases directly regulate abscisic acid-activated protein kinases in *Arabidopsis*. *Proc. Natl. Acad. Sci. U.S.A.* 106, 17588–17593. doi: 10.1073/pnas.0907095106

Van Dongen, J. T., Gupta, K. J., Ramírez, S. J., Araújo, W. L., Nunes, A., and Fernie, A. R. (2011). Regulation of respiration in plants: a role for alternative metabolic pathways. *J. Plant Physiol.* 168, 1434–1443. doi: 10.1016/j.jplph.2010.11.004

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature* 456, 470–476. doi: 10.1038/nature07509

Wang, H., Han, L., and Jia, W. (2013). Response of seeds germination of *Populus euphratica* and *Populus pruinosa* to slat stress. *J. Desert Res.* 33, 743–750.

Wang, J., Wu, Y., Ren, G., Guo, Q., Liu, J., and Lascoux, M. (2011). Genetic differentiation and delimitation between ecologically diverged *Populus euphratica* and *P. pruinosa*. *PLoS One* 6:e26530. doi: 10.1371/journal.pone.0026530

Wang, Y., Li, L., Ye, T., Zhao, S., Liu, Z., Feng, Y. Q., et al. (2011). Cytokinin antagonizes ABA suppression to seed germination of *Arabidopsis* by downregulating ABI5 expression. *Plant J.* 68, 249–261. doi: 10.1111/j.1365-313X.2011.04683.x

Wang, W. Q., Møller, I. M., and Song, S. Q. (2012). Proteomic analysis of embryonic axis of *Pisum sativum* seeds during germination and identification of proteins associated with loss of desiccation tolerance. *J. Proteomics* 77, 68–86. doi: 10.1016/j.jprot.2012.07.005

Weitbrecht, K., Müller, K., and Leubner, G. (2011). First off the mark: early seed germination. *J. Exp. Bot.* 62, 3289–3309. doi: 10.1093/jxb/err030

Winkel-Shirley, B. (2002). Biosynthesis of flavonoids and effects of stress. *Curr. Opin. Plant Biol.* 5, 218–223. doi: 10.1016/S1369-5266(02)00256-X

Wullschleger, S. D., Weston, D., DiFazio, S. P., and Tuskan, G. A. (2013). Revisiting the sequencing of the first tree genome: *Populus trichocarpa*. *Tree Physiol.* 33, 357–364. doi: 10.1093/treephys/tps081

Yamaguchi-Shinozaki, K., and Shinozaki, K. (2005). Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. *Trends Plant Sci.* 10, 88–94. doi: 10.1016/j.tplants.2004.12.012

Yang, M. F., Liu, Y. J., Liu, Y., Chen, H., Chen, F., and Shen, S. H. (2009). Proteomic analysis of oil mobilization in seed germination and postgermination development of *Jatropha curcas*. *J. Proteome Res.* 8, 1441–1451. doi: 10.1021/pr800799s

Yang, P., Li, X., Wang, X., Chen, H., Chen, F., and Shen, S. (2007). Proteomic analysis of rice (*Oryza sativa*) seeds during germination. *Proteomics* 7, 3358–3368. doi: 10.1002/pmic.200700207

Ye, C. Y., Zhang, H. C., Chen, J. H., Xia, X. L., and Yin, W. L. (2009). Molecular characterization of putative vacuolar NHX-type Na+/H+ exchanger genes from the salt-resistant tree *Populus euphratica*. *Physiol. Plant.* 137, 166–174. doi: 10.1111/j.1399-3054.2009.01269.x

Zhang, J., Feng, J., Lu, J., Yang, Y., Zhang, X., Wan, D., et al. (2014). Transcriptome differences between two sister desert poplar species under salt stress. *BMC Genomics* 15:337. doi: 10.1186/1471-2164-15-337

Zhang, J., Xie, P., Lascoux, M., Meagher, T. R., and Liu, J. (2013). Rapidly evolving genes and stress adaptation of two desert poplars, *Populus euphratica* and *P. pruinosa*. *PLoS One* 8:e66370. doi: 10.1371/journal.pone.0066370

Zhu, J. K. (2001). Plant salt tolerance. *Trends Plant Sci.* 6, 66–71. doi: 10.1016/S1360-1385(00)01838-0

Ziemann, M., Kamboj, A., Hove, R. M., Loveridge, S., El, A., and Bhave, M. (2013). Analysis of the barley leaf transcriptome under salinity stress using mRNA-Seq. *Acta Physiol. Plant.* 35, 1915–1924. doi: 10.1016/j.crvi.2015.03.010

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 March 2019
doi: 10.3389/fgene.2019.00237

Conserved MicroRNA Act Boldly During Sprout Development and Quality Formation in Pingyang Tezaocha (*Camellia sinensis*)

Lei Zhao[1,2†], *Changsong Chen*[3], *Yu Wang*[1], *Jiazhi Shen*[4] *and Zhaotang Ding*[1*]

[1] Qingdao Key Laboratory of Genetic Improvement and Breeding in Horticultural Plants, College of Horticulture, Qingdao Agricultural University, Qingdao, China, [2] Department of Plant Science and Landscape Architecture, University of Maryland, College Park, MD, United States, [3] Tea Research Institute, Fujian Academy of Agricultural Sciences, Fu'an, China, [4] College of Horticulture, Nanjing Agricultural University, Nanjing, China

OPEN ACCESS

Edited by:
Yuriy L. Orlov, Institute of Cytology and Genetics (RAS), Russia

Reviewed by:
Lidiia Samarina, Russian Research Institute of Floriculture and Subtropical Crops (RRIFSC), Russia
Oksana Gennadèvna Belous, Russian Research Institute of Floriculture and Subtropical Crops (RRIFSC), Russia
Weiwei Wen, Huazhong Agricultural University, China

***Correspondence:**
Zhaotang Ding
dzttea@163.com
orcid.org/0000-0002-6814-3038

†*Lei Zhao*
orcid.org/0000-0003-1019-3814

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *15 October 2018*
Accepted: *04 March 2019*
Published: *28 March 2019*

Citation:
Zhao L, Chen C, Wang Y, Shen J and Ding Z (2019) Conserved MicroRNA Act Boldly During Sprout Development and Quality Formation in Pingyang Tezaocha (Camellia sinensis). Front. Genet. 10:237. doi: 10.3389/fgene.2019.00237

Tea tree [*Camellia sinensis* (L.) O. Kuntze] is an important leaf (sometimes tender stem)-using commercial plant with many medicinal uses. The development of newly sprouts would directly affect the yield and quality of tea product, especially significant for Pingyang Tezaocha (PYTZ) which takes up a large percent in the early spring tea market. MicroRNA (miRNA), particularly the conserved miRNAs, often position in the center of subtle and complex gene regulatory systems, precisely control the biological processes together with other factors in a spatio-temporal pattern. Here, quality-determined metabolites catechins, theanine and caffeine in PYTZ sprouts including buds (sBud), different development stages of leaves (sL1, sL2) and stems (sS1, sS2) were quantified. A total of 15 miRNA libraries of the same tissue with three repetitions for each were constructed to explore vital miRNAs during the biological processes of development and quality formation. We analyzed the whole miRNA profiles during the sprout development and defined conserved miRNA families in the tea plant. The differentially expressed miRNAs related to the expression profiles buds, leaves, and stems development stages were described. Twenty one miRNAs and eight miRNA-TF pairs that most likely to participate in regulating development, and at least two miRNA-TF-metabolite triplets that participate in both development and quality formation had been filtered. Our results indicated that conserved miRNA act boldly during important biological processes, they are (i) more likely to be linked with morphological function in primary metabolism during sprout development, and (ii) hold an important position in secondary metabolism during quality formation in tea plant, also (iii) coordinate with transcription factors in forming networks of complex multicellular organism regulation.

Keywords: conserved miRNA, sprouts development, quality formation, transcription factors, *Camellia sinensis* (L.) O. Kuntze

INTRODUCTION

Originally produced in China, green tea nowadays is the star among the top list beverages, attribute to its good taste, health benefits, and mysterious process, which brings considerable economic benefit in planting and exporting countries such as China, India, Kenya, and Sri Lanka. Based on reports from the China Tea Marketing Association 2017 (http://www.stats.gov.cn/),

approximately 10.3 million tons of fresh tea leaves were harvested to produce various tea products. As young leaves and tender stems from the tea tree are processed to prepare "tea", the developmental characters are supposed to have a direct and significant bearing on the yield and the quality of tea product.

MicroRNAs (miRNAs), are endogenous single-stranded non-coding small RNAs, that could both regulate their target messenger RNAs (mRNAs) at chromatin state and could also perfectly or imperfectly bind to their targets for further translation suppression by cleaving at some complementary site (Rubio-Somoza and Weigel, 2011; Zheng et al., 2015). The plant miRNA families were thus placed to be at the central position within gene expression programs, always with small numbers per cell and large amounts of transcripts (Voinnet, 2009), yet have powerful effect in developmental regulation, morphogenesis, stress responses (Axtell and Bowman, 2008; De Lima et al., 2012; Jones-Rhoades, 2012; Yang et al., 2013). Still, seldom gene regulation could be completed without consideration of the transcription factors. Many transcription factors in the plant kingdoms are highly conserved even stride over large evolutionary distances, and for some of them, they could still share similar developmental roles in diverse species (Zhao et al., 2013; Xu et al., 2016). Hypothesis demonstrated that the miRNA binding sites evolve faster than the transcription factor binding sites, as the ways to repress a gene is relatively much more than to activate one (Chen and Rajewsky, 2007).

However, some miRNA seems extremely well conserved (Lu et al., 2006). A large portion of the conserved miRNAs and their conventional target transcription factors as well as F-box proteins play pivotal roles in governing plastic behavior during development, such as phase change and plant architecture (Kidner, 2010; Rubio-Somoza and Weigel, 2011), making miRNA-TF mRNA pairs more fascinating. At least 7 kinds of miRNAs were widely reported to regulate in the three stages of leaf development and leaf morphology. At the initiation stage, a division of leaf primordia are commonly considered to be the key stage in the leaf development process, which comes from a group of cells localized on the flanks of the shoot apical meristem (SAM) loses their indeterminacy (Micol and Hake, 2003). During this stage, miR390/ARF pathway has been described in the regulation of leaf polarity (Braybrook and Kuhlemeier, 2010). miR165/166 regulates the leaf polarity by targeting the HD-ZIP genes and thus control the adaxial cell fate (Rubio-Somoza and Weigel, 2011; Sun, 2012). Recent discovery revealed that the leaf dorsoventral polarity (adaxial-abaxial) signals which may cause mechanical heterogeneity of the cell wall, is linking to the methyl-esterification of cell-wall pectins in tomato and *Arabidopsis* (Qi et al., 2017). The shape and architecture of leaf need the orchestration of auxin, *KNOX* genes and miRNA regulation. *KNOX* genes could be down-regulated by CUC transcriptional regulators, which are important for organ boundaries building (Takada and Tasaka, 2002; Chen, 2009), floral patterning, and leaf morphogenesis (Micol and Hake, 2003; Engstrom et al., 2004). NAC (NAM, CUC1/2-like) is one branch of CUC gene family regulated by miR164. MiR164/GOB (a CUC2 ortholog gene), well-studied in tomato, is necessary for controlling leaf polarity and determining the serration or smooth of the leaf boundaries (Berger et al., 2009). MiR319, encoding by three loci including miR-JAW (miR319a) in *Arabidopsis*, regulates five *TEOSINTE BRANCHED/CYCLOIDEA/PCF* (TCP) family members (Palatnik et al., 2003, 2007), which could also lead to the regulation of CUC genes. Overexpression of miR319 or loss function of these five *TCP* genes would result in crinkly leaves (Palatnik et al., 2003; Liu et al., 2018). TCP regulated growth and senescence via jasmonic acid synthesis pathway (Schommer et al., 2008). The cell number and cell size, which reported to be precisely spatial and temporal controlled (Usami et al., 2009), are mainly regulated by an SQUAMOSA PROMOTER BINDING PROTEIN PROTEIN-LIKE (SPL)-dependent pathway (Ferreira e Silva et al., 2014; Xu et al., 2016). In *Arabidopsis*, miR156 targets 11 of the 17 SPL genes, among which SPL3, 4, and 5 accelerates the juvenile-to-adult phase change, SPL9 and SPL15 regulate plastochron length (Wang et al., 2008; Wu et al., 2009; Xu et al., 2016). MiR396 plays an important role in plant leaf growth and development, most likely by repressing Growth-Regulating Factor (GRF) genes in *Arabidopsis*. Transgenic miR396-overexpressing plants have narrow-leaf phenotypes due to a reduction in cell number (Liu et al., 2009).

During the long cultivation history for more than 2,000 years in China, numerous elite tea varieties have been bred for different characteristics like early germination, high yield, good performance under environmental stress, and distinctive aroma or flavor. *Camellia sinensis* (L.) O. Kuntze "Pingyang Tezaocha" (PYTZ), an elite cultivar with short internodes selected in Zhejiang Province in the late century, is now popularized in the north tea area in China attributes to its high yield for about 3tons green tea products per hectare [Data from e-China tea from Tea Research Institute of China Academy of Agriculture Sciences AS (TRI, CAAS)] (http://www.e-chinatea.cn/other_shujuku.aspx) (Zhao et al., 2017). What's more, its early germination in April helps taking up a large percent in the early spring tea market annually (Yang, 2015). In tea plant, phenolic compounds is one of the most important secondary metabolites, accounting for 18% to 36% dry weight in the fresh leaves and tender stem (Jiang et al., 2013), is also the main flavor components and functional ingredients that had been intensely studied in the past decades for its effective and extensive pharmacological activities (Zhao et al., 2013). Accordingly, the accumulation of some phenolic-like nongalloylated catechins like epigallocatechin (EGC) and epicatechin (EC) (Zhao et al., 2017), quinic acid and flavonol glycosides are gradually increasing along with the developing stages (Jiang et al., 2013). What's more, some key enzyme genes involved in the biosynthetic pathway of phenolic compounds in different organs and leaves at different developmental stages also have the similar expression patterns, such like CsDHQ/DHS2 (DHQ/DHS, 3-dehydroquinate synthase), CsCHS1 (CHS, chalcone synthase), CsUGT78E1 (UGT, uridine diphosphate Glycosyltransferase), Cs4CL1 (4CL,4-coumaroyl-CoA ligase), CsF3′H1 (F3′H, flavonoid 3′-hydroxylase), and some TF genes like Sg4 of CsMYB family (Jiang et al., 2013; Li et al., 2017a), CsMYB5-1 and bHLH24-3 (Jiang et al., 2013). Beyond catechins in tea plant, theanine and caffeine are the other two characteristic constituents determine tea quality (Xia et al., 2017). No matter it is primary or secondary, there is no doubt that metabolisms

synchronize along with plant growth and development, are under the precise entire spatio-temporal network control (Chen and Rajewsky, 2007).

Here, the content of three kinds of main taste compounds catechins, theanine, and caffeine in different tissues of spring sprouts including bud, two stages of leaves, and stems of PYTZ were quantified by High-Performance Liquid Chromatography-Mass Spectrometry (HPLC-MS/MS). MiRNA libraries of the same tissues were constructed by Illumina HiSeq technology in order to explore how miRNA works between development and quality formation. Key miRNAs involved in regulating sprout development have been speculated based on computational expression. Conserved miRNA families in tea plant were obtained and mainly studied. To what extent the conserved miRNAs might be linked with morphogenesis function during sprout development was further investigated through other six morphologically-different tea cultivars. Regulations in metabolic pathways of conserved miRNA together with their target genes, especially transcription factor genes that would finally determine tea quality have been studied and discussed. The consistency of performance between development and quality need more cross understanding and balance in the subsequent process of screening tea cultivars.

MATERIALS AND METHODS

Plant Material and RNA Isolation

The four-year-old tea plant cultivar *Camellia sinensis* (L.) O. Kuntze "Pingyang Tezaocha" (PYTZ) were planted in the Germplasm of Qingdao Tea Repository at the Tea Research Institute located in Qingdao (35°N119°E, Qingdao city, China) under natural light condition. To ensure the successiveness of gene expression during the development of newborn branch, we collected the samples of bud, leaves, and stems orderly downwards from the top (**Figure 1A**) (Shen et al., 2019). For normalization, the buds about 3 cm long (sBud), and the first leaf below the bud (sL1) about 3.5cm long, the second leaf with higher maturity below the bud (sL2) 4.5cm long, the stem between the first leaf and the second leaf (sS1) with about 1.5cm long, and the more mature stem between the second leaf and the fish leaf (sS2) with 2.1cm long were measured and collected. For collecting samples of RNA, healthy buds, leaves and stems at different developmental stages were collected and frozen immediately in liquid nitrogen and stored in −80°C freezers before use (Fan et al., 2015). Three biological replicates were collected and pooled from at least five individuals, and each biological replicate contained more than five buds, leaves and stems. The total RNA for each sample was extracted using TRIzol reagent (Invitrogen, Burlington, ON, Canada). The quality, purity, concentration, and integrity of the total RNA was checked using 1% agarose gel electrophoresis, NanoDrop Photometer Spectrophotometer (IMPLEN, Westlake Village, CA, USA), Qubit RNA Assay Kit in Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA), and RNA Nano 6000 Assay Kit of the Bioanalyser 2100 system (Agilent Technologies, Santa Clara, CA, USA), respectively. RNA samples with a 260/280 ratio between 1.8 and 2.0, 260/230 ratio between 2.0 and 2.5, and RNA integrity number more than 8.0, were used for sequencing and quantitative PCR analysis described below.

Extraction and Quantification of Catechins, Caffeine, and Theanine

The extraction of catechins and caffeine followed by a previously described method with minor modifications (Jiang et al., 2013; Wang et al., 2018): 0.2 g of each fresh samples (sBud, sL1, sL2, sS1, and sS2) were ground in liquid nitrogen and extracted with an extraction solution (80% methanol and 20% water), followed by vortexing and sonicating for 30 min at a low temperature. Then, the samples were centrifuged at 3,500 g for 15 min, and the residues were re-extracted twice as mentioned above until the final volume of the pooled supernatants was 2 mL. The supernatants were then extracted three times with chloroform and three times with ethyl acetate. The pooled supernatant was concentrated to remove the ethyl acetate at a low temperature with a vacuum pump. Finally, the product was dissolved in 200 μL methanol for quantification. The theanine was extracted as reported by the method of Jeon et al. (2017) with some modifications. One gram of each finely ground sample was mixed with 100 mL boiling distilled water and brewing for 10 min (with the help of magnetic stirrer). All the obtained extract were filtered by 0.45 μ nylon membrane (after cooling down) and approximately 1 ml of the sample solution were centrifuged at 13,000 rpm for 10 min prior to HPLC analysis.

The isolation and detection of quality-related metabolites catechins, caffeine, and theanine in the sprouts of PYTZ were performed by high performance liquid chromatography-mass spectrometry (HPLC-MS/MS). HPLC analyses were performed on an Agilent 1298 LC system (Agilent, Santa Clara, CA, USA), and MS/MS detection was carried out using an Agilent 6460 Series Triple Quadrupole instrument (Agilent). Caffeine, theanine and six major tea standards: catechins, (-)-epigallocatechin gallate (EGCG), (-)-epigallocatechin (EGC), (-)-epicatechin gallate (ECG), (-)-epicatechin (EC), (-)-gallocatechin (GC), and (+)-catechin (C) were purchased from Sigma (St Louis, MO, USA). An Agilent 20RBAX RRHD Eclipse Plus C18 column (particle size: 1.8 mm, length: 100 mm, and internal diameter: 2.1 mm) was used at a flow rate of 1 mL min^{-1}. For catechins and caffeine, the mobile phase consisted of 0.4% acetic acid in water and 100% acetonitrile; and the gradient of latter increased linearly from 0 to 10% (v/v) within 5 min, and to 35% at 20 min, to 10% at 21 min, to 1% at 25 min. For theanine, the mobile phase consisted of HPLC water and acetonitrile; and the gradient of former remained at 100% within 10 min, and decreased linearly from 100% to 20% (v/v) to 12 min, and kept at 20% to 20 min, to 100% at 22 min, and kept at 100% to 40 min. Mass spectra were acquired simultaneously using electrospray ionization in the positive and negative ionization modes over the range of m/z 100 to 2000. A drying gas flow of 6 L min^{-1}, drying gas temperature of 350°C, nebulizer pressure of 45 psi, and capillary voltages of 3,500 V were used. The compounds were identified qualitatively using LC-MS by comparing the retention times (t_R), wavelengths of maximum absorbance (λ_{max}), protonated/deprotonated

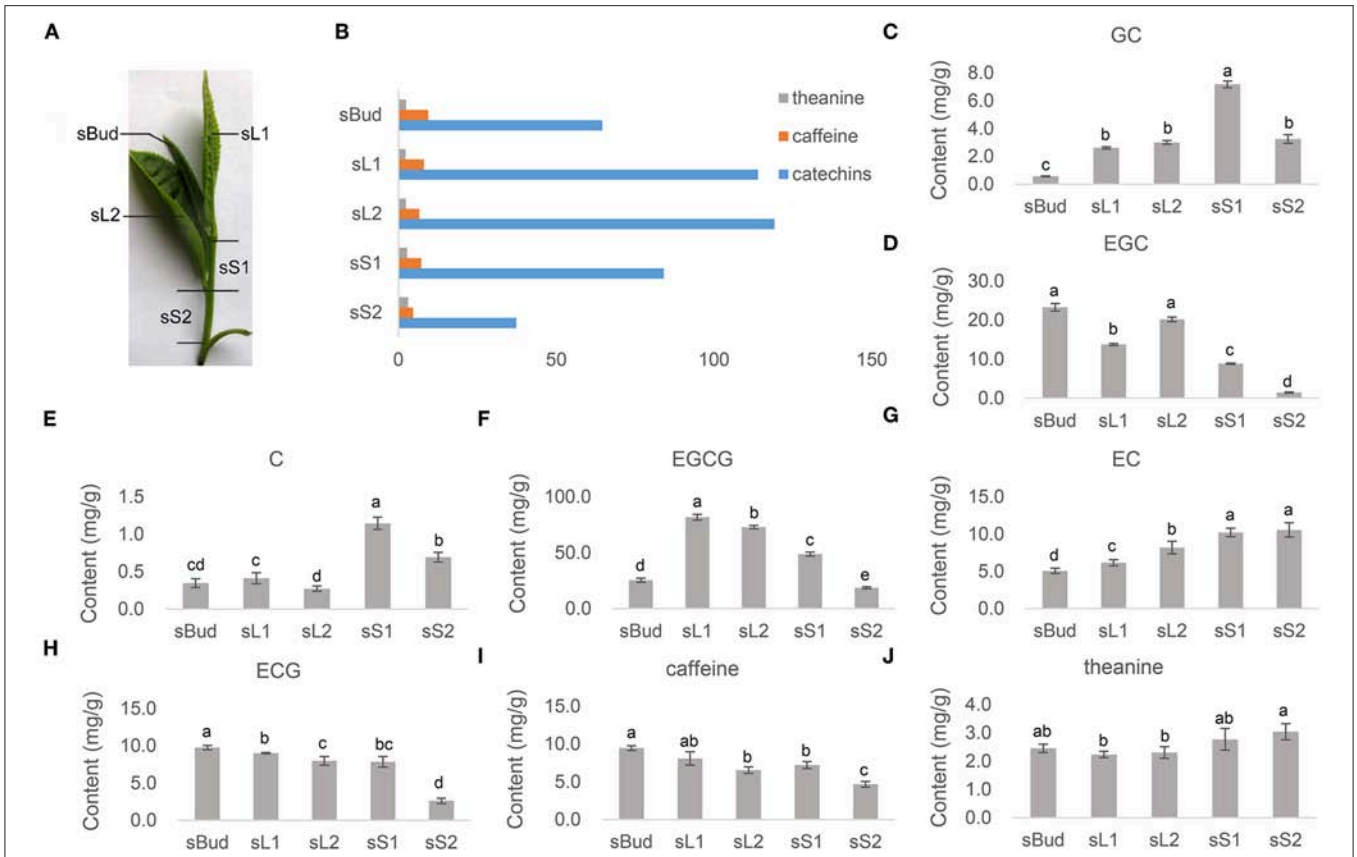


FIGURE 1 | The contents of characteristic metabolites. **(A)** Samples of bud, leaves, and stems were collected orderly downwards from the top, named by sBud, sL1, sL2, sS1, and sS2, from buds, leaves, and stems, correspondingly. **(B)** The content of total catechins, caffeine, and theanine. **(C–J)** The content of GC, EGC, C, EGCG, EC, ECG, caffeine, and theanine in each tissue detected by HPLC-MS/MS. Standard catechins were (-)-gallocatechin (GC), (-)-epigallocatechin (EGC), (+)-catechin (C), (-)-epigallocatechin gallate (EGCG), (-)-epicatechin (EC), and (-)-epicatechin gallate (ECG). (Above each column, the different low case letters mean significant difference, $p < 0.05$, while labeled with the same ones mean insignificance, $p > 0.05$. p-values were calculated with Student's t-test).

molecules ($[M+H]^+/[M-H]^-$), and major fragment ions with those of the authentic standards and published literature (Jiang et al., 2013; Jeon et al., 2017; Wang et al., 2018).

Library Construction and Small RNA Sequencing

For sRNA library construction, 3 μg of total RNA per sample was used for the RNA sample preparations. Sequencing libraries were generated using NEBNext Multiplex Small RNA Library Prep Set for Illumina (NEB, USA). The library preparations were sequenced on an Illumina HiSeq™ 2500 sequencer, by Gene Denovo Biotechnology Co. (Guangzhou, China). The generated 50 bp single-end reads were then filtered out the impure sequences (adaptor sequences and the low quality reads) and removed cellular structural RNAs such as rRNA, snoRNA, snRNA, and tRNA based on the alignment with small RNAs in GeneBank database (Release 209.0) and Rfam database (11.0). The clean reads were mapped to the tea tree genome without mismatch to analyze their expression and distribution (NCBI Sequence Read Archive Database under accession PRJNA381277). Tags that mapped to exons or introns and repeat sequences were also removed.

Identification of Known miRNAs and Novel miRNA

Since tea miRNA dataset was not included in the miRBase, the clean tags were subjected to a Blastn search against miRBase 21.0, to identify and annotate known miRNAs from all other plant miRNAs, allowing two mismatches. All the known miRNAs were further checked for the existence through 72 plant species, to figure out their conservative property. The unannotated tags were aligned with tea tree genome to identify novel miRNA candidates according to their genome positions and hairpin structures predicted by software Mireap (https://github.com/liqb/mireap, version 0.20).

miRNA Expression Profiles and Prediction of Target mRNAs

The expression levels of both known miRNA and novel miRNA from each sample were calculated and normalized to transcripts per million (TPM) (Wu et al., 2017b). The

formula is TPM= Actual miRNA counts / Total counts of clean tags*10^6. Meanwhile, the correlation coefficient between every two replicas was calculated to evaluate repeatability between samples. Differential expression analysis across samples was performed using the DEGseq (2010) R package. miRNAs with $p < 0.05$ and log2-fold change ≥ 2 in comparison were set as the threshold for significantly differentially expressed miRNAs (DEM). Candidate target genes were predicted by using software PatMatch (Version 1.2) blasting against tea tree genome, abiding by some rigorous parameters as follows: No more than four mismatches between sRNA/target (G-U bases count as 0.5 mismatches); For the miRNA/target duplex (5′ of miRNA), (a) no more than two adjacent mismatches, (b) no adjacent mismatches in positions 2–12, (c) no mismatches in positions 10–11, (d) no more than 2.5 mismatches in positions 1–12, and the minimum free energy (MFE) of the miRNA/target duplex should be no < 60% compared to the MFE of the miRNA bound to its perfect complement (Yan et al., 2005; Wu et al., 2017a).

Functional Enrichment Analysis of Target mRNAs

Gene Ontology (GO) enrichment analysis and KEGG pathway analysis were performed to the target mRNAs of DEM in order to comprehensively figure out their biological functions. All DEM target genes were mapped to GO terms in the Gene Ontology database (http://www.geneontology.org/), then the enriched significant GO terms (taking FDR ≤ 0.05 as a threshold, derived from calculated *p*-value) comparing to tea tree genome background were categorized into three levels, "biological process," "cellular component" and "molecular function". KEGG is the major public pathway-related database (Kanehisa et al., 2008) for further understand how genes interact with each other to play roles in certain biological functions. The calculating formula is the same as that in GO analysis. KEGG pathway enrichment analysis identify significantly enriched metabolic pathways or signal transduction pathways (Liu et al., 2014). Some online platforms or commercial services that based on same or different mathematical algorithms could help us to reconstruct gene networks, for example STRING (https://string-db.org/), Pathway Commons (https://www.pathwaycommons.org/) (Luna et al., 2016), ANDSystem (Ivanisenko et al., 2019), and so on (Saik et al., 2018). Functional enrichment of both target genes of miRNAs in single samples and DEM in a compare group were carried out in our analysis. Here, STRING was used to show the enrichment networks.

Trend analysis was aiming at the expression of all miRNAs performing in continuous tissues samples (Bud/L1/L2 and Bud/S1/S2) to cluster genes with similar expression patterns. Trend analysis was carried out by software Short Time-series Expression Miner (Ernst and Bar-Joseph, 2006) (under parameters -pro 20 -ratio 1.0). GO and KEGG pathway enrichment analysis was then be done to target genes of miRNAs in each trend, and the *p*-value was obtained by hypothesis testing. Those GO term and KEGG pathway were defined as significant ones satisfying Q value ≤ 0.05. Q value was that *p*-value corrected by FDR (Benjamin and Hochberg, 1995).

Quantitative PCR for miRNAs and mRNAs

The expression profiles of mature miRNAs and the potential target mRNAs were further validated by quantitative PCR. Synthesis of the first strand cDNA was performed with Mir-X™ miRNA First-Strand Synthesis Kit (Cat. No. 638313, Clontech Laboratories, Inc., CA, USA), with 5.8S rRNA served as an internal control. The first strand cDNA of mRNA were synthesized by using PrimeScript™ RT reagent Kit with gDNA Eraser (Perfect Real Time) (Code No. RR047A, Takara, Tokyo, Japan), with *glyceraldehyde-3-phosphate dehydrogenase* (*GAPDH*) gene for normalization. The primers of miRNAs and mRNA were listed in **Supplementary Table 1**. Quantitative PCR was carried out with SYBR *Premix Ex Taq*™ IIKit (Tli RNase H Plus) (Code No. RR820A, Takara, Tokyo, Japan), on a LightCycler 480 instrument (Roche Molecular Systems, Inc., Indianapolis, IN, USA). The amplification program of miRNA was performed under the following parameters: 95°C for 10 min, 40 cycles at 95°C for 15 s, 60°C for 1 min (Zheng et al., 2015). The amplification program of mRNA was performed at 94°C for 10 s, 58°C for 10 s and 72°C for 10 s (Li et al., 2017b). Triplicates of each reaction were performed, and 5.8S rRNA and *GAPDH* were used as endogenous control separately. CT values obtained through quantitative PCR were analyzed using $2^{-\Delta\Delta CT}$ methods to calculate relative fold change values.

In addition, fresh spring sprouts were plucked from seven tea varieties from one experimental tea garden of Tea Research Institute, Fujian Academy of Agricultural Sciences in Fu'an, China (27°10′N, 119°35′E): *Camellia sinensis* "Jinfenghuang" (JFH), *Camellia sinensis* "Pingyang Tezaocha" (PYTZ), *Camellia sinensis* "Zhengdayin" (ZDY), *Camellia sinensis* "Dayewulong" (DYWL), *Camellia sinensis* "Huangdan" (HD), *Camellia sinensis* "Jiukeng 6" (JK), *Camellia sinensis* "Queshe" (QS). The bud, leaves, and stems were also sampled at the same position mentioned above for quantitative PCR analysis. The result of relative expressions was presented as a heatmap by using TBtools (Chen et al., 2018).

Transcription Factor Prediction and miRNA-mRNA-Metabolite Network Construction

All the mRNA genes target by DEM predicted above were blasted against the transcription factor (TF) database from the plant (http://planttfdb.cbi.pku.edu.cn/, version 4.0) to annotate potential TFs. The resulting target TF genes were classified in each tissue for analysis. The resulting TF genes were further blasted against the tea tree genome (Xia et al., 2017) (NCBI Sequence Read Archive Database No. PRJNA381277). The methods for TF blasting and expression analysis were followed by Zhao et al. (2017). Different expression profiles were finally grouped into profile 1, 2, and so on. For the network analysis was based on Savoi's method (Savoi et al., 2016), the mRNA and metabolite association were obtained based on Pearson correlation coefficient between the contents of each metabolite and the expression levels of mRNA, and filtered the pairs when the absolute value of cor was larger than 0.9 and *p*-value was smaller than 0.05. The miRNA and mRNA association were

obtained by Spearman's correlation coefficient according to their expression levels calculated by TPM and filtered the pairs when cor-value was no larger than −0.5, and the *p*-value was smaller than 0.05. The network was visualized by Cytoscape (V3.6.0) (Praneenararat et al., 2012).

Availability of Supporting Data

Clean Illumina sequencing reads of 15 small RNA of PYTZ sprout have been deposited in the NCBI Sequence Read Archive Database under accession PRJNA510482.

RESULTS

The Contents of Quality-Related Metabolites in Tea Sprouts

The buds, the first leaves, the first stems, sometimes with the second leaves together, are the most common raw materials for producing green tea. The contents of the most quality-related metabolites in tea sprouts that contributing flavors and health-promoting functions were quantified by HPLC-MS/MS (**Figures 1B–J**). Among the three kinds of characteristic metabolites, the content of total catechin accounts for the most majority proportion, while distributing significantly different among tender leaves and stems (**Figure 1B**). The galloylated catechins such like EGCG (**Figure 1F**) and ECG (**Figure 1H**), take up 67.4% in total catechin concentration obtained by summation of the individual components ranged from 37.3 to 119.2 mg/g (**Figure 1B**). Both EGCG and ECG had a relatively low concentration in sS2. The level of GC (**Figure 1C**), C (**Figure 1E**), and EC (**Figure 1G**), were found to be low in bud and leaves, especially for C. The concentrations of caffeine ranging from 4.7 to 9.7 mg/g, accumulated to its highest levels in sBud and lowest in sS2 (**Figure 1I**). Theanine, however, showed its highest accumulation in stems in PYTZ sprouts (**Figure 1J**).

Overview of microRNA Profile and Its Mapping to Tea Tree Genome

To figure out what role the miRNA play during the formation of characteristic metabolites along with the development of the PYTZ sprout, 15 sRNA-Seq libraries including buds, leaves, and stems were separately sequenced on Illumina HiSeq^{TM} 2500 platform generating a total of 292,653,360 raw reads. After removing dirty reads containing adapters and low quality bases, in average, the clean tags of 14,080,519 for sBud (bud), 12,776,739 for sL1 (the 1st leaf), 11,830,623 for sL2 (the 2nd leaf), 11,484,341 for sS1 (the younger stem), and 9,822,915 for sS2 (the older stem) were retained. The filtering data of each procedure were listed in **Supplementary Table 2**. Most clean tags had the length of 21-24nt, in which the 24 nt sRNAs were the most abundant (**Figure 2A**). The proportion of different length tags has no obvious difference among the five sample groups, and generally showed the trends of increased and then decreased bounded by the 24 nt sRNAs. Notably, the number of 24 nt sRNAs in sS1 was the lowest, while in sL1 was the highest.

About 76.97% clean tags were perfectly mapped to tea tree genome (NCBI Sequence Read Archive Database No. PRJNA381277), which indicated a credible quality of sequencing, and the rates of a genomic match were similar across these samples. We removed the tags mapped to exons located in positive-sense strands which might be fragments from mRNA degradation. (For statistics of mapping to tea tree genome, see **Supplementary Table 3**). The tags mapped to repeat sequences were also excluded. The clean tags were then aligned with the Rfam database (11.0) and the percentage of annotation was summarized in **Supplementary Table 4**. The average of rRNA, snRNA, snoRNA, and tRNA in the samples took up of 18.11, 0.19, 0.56, and 1.01%.

Known miRNAs Identification and Novel microRNAs Prediction in PYTZ

To identify known miRNAs in tea, all the unannotated unique tags were blast-searched against plant miRNAs in miRBase (Release 21.0, June 2014). Overall, a total of 1,928,678 miRNA clean tags were identified from 15 libraries (**Table 1**), and 156 known miRNAs were identified (**Supplementary Table 5**). Most of the identified known miRNAs (81.97%) belonged to the 21 nt length miRNA families, the remaining ones belonged to 18–24 nt miRNAs families (**Supplementary Table 5**). For the 156 known miRNAs, 122 precursors and 99 kinds of characteristic hairpin structures of the known miRNAs were identified. The length of the precursors varied from 71 to 288 nt, with an average of 144 nt, and the average minimum free energy (MFE) was −57.23 kcal/mol, ranging from −22.3 to −90.2 kcal/mol (**Supplementary Table 6**). The abundance of miRNA including the novel miRNA showed a high distribution in leaf than in stem (ratio value in **Table 1**), with that in S1 the lowest.

For the four nucleic acids, the frequency of cytosine (C) (32.09%) and uracil (U) (29.65%) is higher than guanine (G) (19.26%) and adenosine (A) (19.00%). In the five samples, U had a high appearance at the 1st, 17th, 22th, and 23rd positions, with an average of 84.34, 61.80, 54.57, and 52.48%, respectively (**Supplementary Figure 1**). C occupied a very high percentage (87.44%) at 19th position. The analysis showed that A had a relatively high proportion at 9th position in the stem (sS1 and sS2) and 17th position in leaf (sL1 and sL2), in contrast, A was seldom present at 2nd, 13th, 18th, and 20th positions in the five tissues. For the first nucleotide bias analysis, U had the absolute predominance in miRNAs with the length of 20, 21, and 22 nt (**Supplementary Figure 2**).

The remaining reads which couldn't get mapped to known miRNAs were used to identify novel miRNAs. 384,768 novel miRNA tags were identified from 15 libraries (**Table 1**), and 1186 novel miRNA tags were identified by predicting the hairpin structures of their precursor sequences (**Supplementary Table 7**). The length of the novel miRNAs ranged from 18 to 27 nt, different with known miRNAs, the 22 nt length miRNA families were the most abundant (46.71%), followed by 21nt (41.23%). These novel miRNA were involved in 1130 hairpin miRNA precursors. The length of these precursors varied from 65 to 373 nt, with an average of 178 nt. The average minimum free energy (MFE) was −56.26 kcal/mol, ranging from −18.1 to −292.3 kcal/mol (**Supplementary Table 7**). The

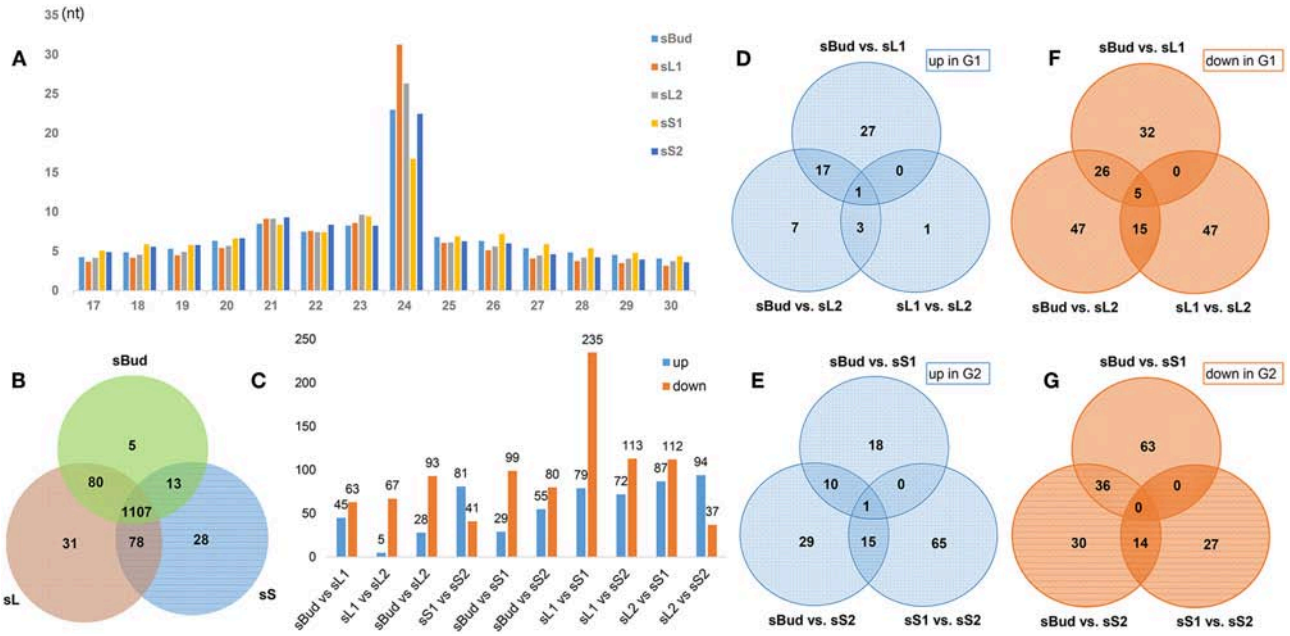


FIGURE 2 | Overview of PYTZ microRNA libraries. **(A)** Size-class distribution of sRNAs in the sBud, sL1, sL2, sS1, and sS2 libraries. Designations the same as on previous figure. **(B)** Number statics of tissue-specific miRNAs. **(C)** Number statics of differentially expressed miRNAs (DEM) between each two sample groups. **(D–G)** The number of up/down-regulated DEM between different tissues in G1/G2.

numbers of novel miRNAs were most in sL1 and lowest in sS1, the trend was the same with known miRNAs.

Whole miRNA Expression Characters in PYTZ

To figure out the whole miRNAs express patterns in the spring sprouts of PYTZ, we need to evaluate the reliability of parallel experimental results as well as operational stability. The expression level of all miRNAs including known miRNAs and novel miRNAs from 15 libraries was normalized to generate TPM, which further used to compute the related coefficients. The strong correlation between every two biological replicates for interlibrary of all five sample groups brought out that the sequencing results are highly reliable (**Supplementary Figure 3**). The correlations between sL1 and sL2, sS1, and sS2 were substantially higher than other inter-groups, suggesting closely associated integral processes in the separate development of leaf and stem.

To lock the target miRNAs which might be responsible for the tea shoot development, we firstly define sBud/ sL1/sL2 as Group 1 (G1), sBud/ sS1/sS2 as Group 2 (G2) to see the whole miRNA change characters. In G1, a total of 226 miRNAs showed different expression among the three tissues and classified into 8 profiles according to their trends. Overall, the down-expression trend miRNAs took up a larger percentage (69.5%), which belonged to profile 3 (75 miRNAs), profile 0 (67 miRNAs), and profile 1 (15 miRNAs). The up-expression trend miRNAs which had the most abundant expression in sL2 belonged to profile 6 (40 miRNAs), profile 7 (12 miRNAs), and profile 4 (1 miRNA). There were other 9 and 7 miRNAs which had the highest and lowest expression level in sL1, separately (**Figure 3A**).

In G2, a total of 273 miRNAs showed different expression among the three tissues and had been classified into 8 profiles according to their trends. Overall, 91 miRNAs showed the down-expression trend belonging to profile 3 (42 miRNAs), profile 1 (34 miRNAs), and profile 0 (15 miRNAs). 64 miRNAs showed the up-expression trend belonging to profile 6 (35 miRNAs), profile 4 (25 miRNAs), and profile 7 (4 miRNA). It is noteworthy that there were 95 miRNAs and 23 miRNAs showed the lowest and highest expression level in sS1, separately (**Figure 3B**).

Conserved miRNAs Families and Tissue-Specific miRNAs

The 156 known miRNAs belonging to 125 families, among which 27 families were well-conserved that present in more than 10 plant species out of 72 plant species (**Table 2**). miR156 was the most popular one, which was found in 51 plant species, followed by miR396 and miR166, which were conserved in 47 and 45 plant species, respectively.

The expression of miRNA usually tells more about the code of regulating new shoot elongation and development. In order to filter the tissue-specific miRNAs in PYTZ, we merged the DEM from sL1 and sL2 into sL, sS1 and sS2 into sS, and removed duplicates, separately. The Venn diagram (**Figure 2B**) showed that most miRNA were existed in all tissues (82.49%) or at least in one tissue, regardless of their relatively high or medium expression abundance. Interestingly, some miRNA could only

TABLE 1 | Numbers and ration of clean tags for conserved miRNA and novel miRNA.

Sample	Total	Known miRNA			Novel miRNA		
		Mirna num	Tags uniq	Tags total/ratio	Mirna num	Tags uniq	Tags total/ratio
sBud-1	13025496	90	2281	145547(1.12%)	986	1205	35371(0.27%)
sBud-2	13672142	87	2150	96863(0.71%)	780	1003	22599(0.17%)
sBud-3	15543919	86	2172	79234(0.51%)	669	876	23593(0.15%)
sL1-1	13217393	100	2571	237953(1.80%)	1000	1244	47097(0.36%)
sL1-2	12086276	98	3140	205017(1.70%)	1008	1276	36970(0.31%)
sL1-3	13026547	91	1967	127242(0.98%)	854	1066	30010(0.23%)
sL2-1	11146588	94	2523	151824(1.36%)	820	1037	26207(0.24%)
sL2-2	13556636	90	2518	162349(1.20%)	737	872	21904(0.16%)
sL2-3	10788646	90	2485	201288(1.87%)	783	1005	27762(0.26%)
sS1-1	10821241	73	1763	48215(0.45%)	508	693	11804(0.11%)
sS1-2	11025165	76	1753	48019(0.44%)	483	669	10145(0.09%)
sS1-3	12606617	81	1902	74064(0.59%)	659	852	18301(0.15%)
sS2-1	9125465	81	2079	114250(1.25%)	745	982	25010(0.27%)
sS2-2	11303334	90	2451	130149(1.15%)	847	1025	26705(0.24%)
sS2-3	9039947	79	2144	106664(1.18%)	781	945	21290(0.24%)
total	179985412	1306	33899	1928678(16.31%)	11660	14750	384768(3.25%)

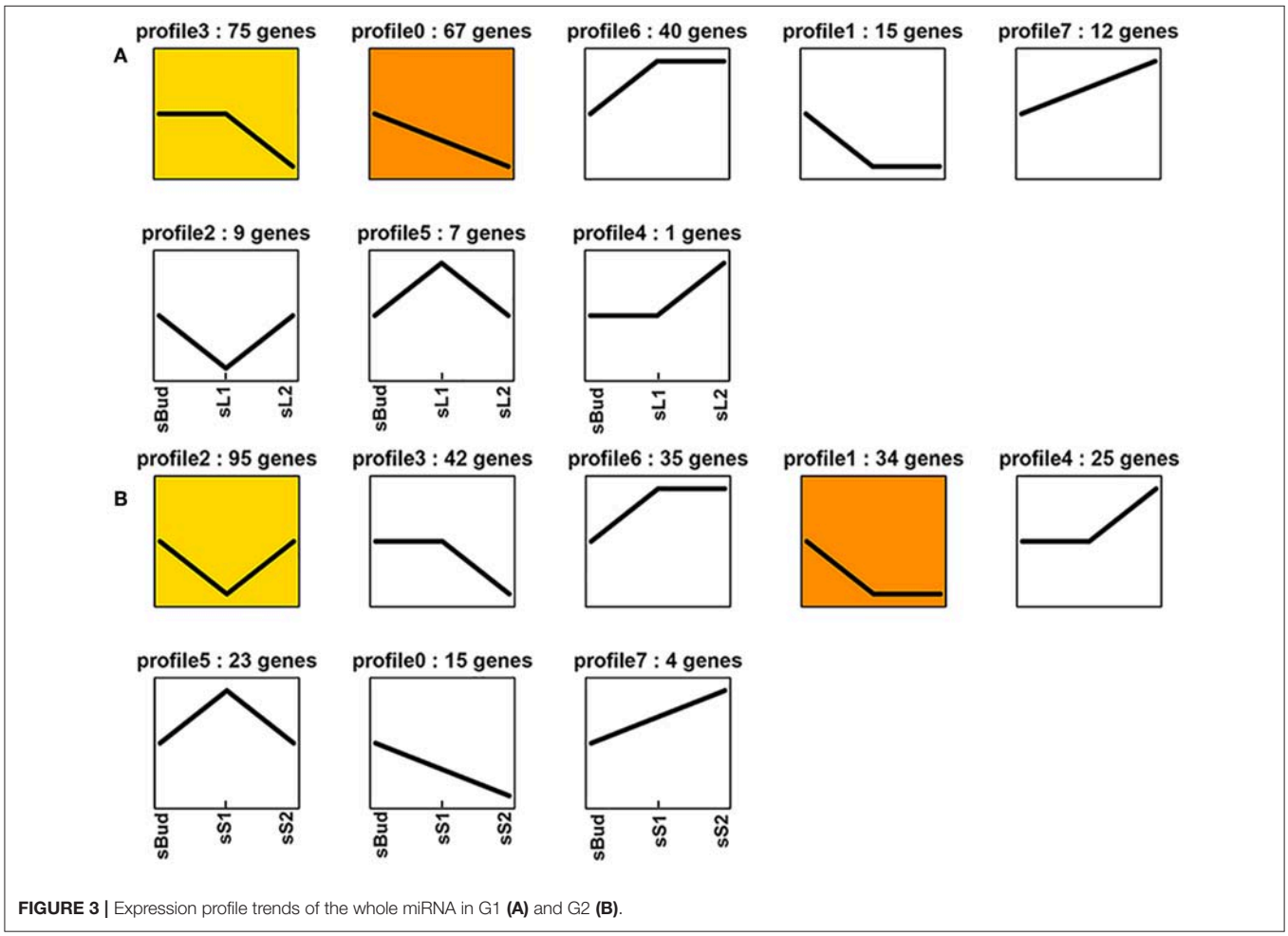


FIGURE 3 | Expression profile trends of the whole miRNA in G1 **(A)** and G2 **(B)**.

TABLE 2 | 27 conserved miRNA families from PYTZ searched in other 72 plant species.

Species	MIR 156	MIR 396	MIR 166	MIR 171	MIR 160	MIR 167	MIR 164	MIR 172	MIR 319	MIR 159	MIR 169	MIR 408	MIR 390	MIR 395	MIR 398	MIR 168	MIR 399	MIR 162	MIR 393	MIR 394	MIR 482	MIR 403	MIR 530	MIR 2111	MIR 477	MIR 535	MIR 2118
Brachypodium distachyon	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	−	−	−	+
Glycine max	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+
Arabidopsis thaliana	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	−	+	−	−	−
Oryza sativa	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	−	−	+	+
Arabidopsis lyrata	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	−	+	−	−	−
Malus domestica	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+
Prunus persica	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−
Vitis vinifera	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	−
Populus trichocarpa	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−
Solanum tuberosum	+	+	+	+	+	+	+	+	+	−	+	+	+	+	+	−	+	+	+	−	+	−	+	−	+	−	−
Citrus sinensis	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+	−	+	+	−
Cucumis melo	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	+	+	−	−
Manihot esculenta	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+	+	+	−
Amborella trichopoda	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+	+	−	−	+	+	−	−	−	+	+	+	−
Medicago truncatula	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	−	+	+	−	−	+
Nicotiana tabacum	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	+	+	−	−	−	+	−	−
Aegilops tauschii	+	+	+	+	+	+	+	+	+	−	+	+	+	+	+	+	+	−	+	+	−	−	−	−	−	−	+
Zea mays	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	+
Sorghum bicolor	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	−	−	−	−	+
Theobroma cacao	+	+	+	+	+	+	+	+	+	−	+	−	+	+	+	+	+	+	+	+	−	+	+	+	−	+	−
Carica papaya	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	+	+	−	−	−	−	+	+	−
Linum usitatissimum	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	−	−	−	−
Ricinus communis	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	−	−	+	−	−	−	+	−
Solanum lycopersicum	+	+	+	+	+	+	+	+	+	+	+	−	+	+	−	+	+	+	−	+	+	+	−	−	+	−	−
Triticum aestivum	+	+	−	+	+	+	+	−	+	+	+	+	−	+	+	−	+	−	−	−	−	−	+	−	−	−	−
Aquilegia caerulea	+	+	+	+	+	+	−	+	+	+	+	+	−	+	+	+	+	−	−	−	+	−	+	−	+	+	−
Brassica napus	+	+	+	+	+	+	+	+	−	+	+	−	+	+	−	+	+	+	+	+	−	+	−	+	−	−	−
Brassica rapa	+	+	−	+	+	+	+	+	+	+	−	+	+	+	+	+	−	+	−	−	−	+	−	+	−	−	−
Cynara cardunculus	+	+	−	+	+	+	+	+	+	−	+	+	+	+	+	+	+	−	+	+	−	−	−	−	−	−	−
Physcomitrella patens	+	−	+	+	+	+	−	−	+	−	−	+	+	+	−	−	−	−	−	−	−	−	−	−	+	+	−
Gossypium hirsutum	+	+	+	−	+	+	+	+	−	−	+	−	+	−	+	−	+	+	+	+	+	−	−	−	−	−	−
Hevea brasiliensis	+	+	+	−	−	−	−	−	+	+	−	+	−	−	+	−	−	−	−	−	+	−	−	−	−	−	+
Vigna unguiculata	+	−	−	−	+	−	+	+	+	−	+	+	−	+	−	+	+	+	−	−	+	−	−	−	−	−	+
Gossypium raimondii	−	−	+	−	−	+	+	+	−	−	−	−	−	−	+	−	+	−	−	−	+	−	+	−	+	−	−
Salvia sclarea	+	+	+	+	−	−	+	+	−	−	+	−	−	+	+	−	+	−	−	+	−	−	−	−	−	−	−
Pinus taeda	+	+	+	+	−	−	−	−	+	+	−	+	+	−	+	−	−	−	−	−	+	−	−	−	−	−	−

(Continued)

TABLE 2 | Continued

Species	MIR 156	MIR 396	MIR 166	MIR 171	MIR 160	MIR 167	MIR 164	MIR 172	MIR 319	MIR 159	MIR 169	MIR 408	MIR 390	MIR 395	MIR 398	MIR 168	MIR 399	MIR 162	MIR 393	MIR 394	MIR 482	MIR 403	MIR 530	MIR 2111	MIR 477	MIR 535	MIR 2118
Festuca arundinacea	+	+	+	+	+	−	+	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Saccharum sp.	+	+	+	−	−	+	−	−	−	+	+	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−
Arachis hypogaea	+	−	−	−	+	+	−	−	−	+	−	+	−	−	+	−	−	−	−	+	−	−	−	−	−	−	−
Helianthus tuberosus	+	−	−	+	+	−	−	−	−	+	−	−	−	−	−	−	−	+	+	−	−	+	+	−	−	−	−
Hordeum vulgare	+	−	+	+	−	−	−	−	−	+	+	−	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−
Lotus japonicus	−	+	−	+	−	+	−	+	−	−	−	+	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−
Pinus densata	−	+	+	+	−	−	−	−	−	+	+	−	+	−	−	−	−	+	−	−	+	−	−	−	−	−	−
Selaginella moellendorffii	+	+	+	+	+	−	−	−	+	+	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Digitalis purpurea	+	+	+	−	+	+	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Picea abies	−	+	+	−	+	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	+	−	−	−	−	+	−
Acacia auriculiformis	−	+	−	−	+	−	−	+	+	−	−	−	−	−	−	+	−	+	−	−	−	−	−	−	−	−	−
Citrus trifoliata	+	−	+	+	−	+	+	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Phaseolus vulgaris	−	−	+	−	−	−	−	−	+	+	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−	−	+
Saccharum officinarum	+	+	−	−	−	+	−	−	−	+	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−
Brassica oleracea	−	−	−	+	−	−	−	+	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−
Citrus clementina	−	+	−	+	−	+	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−
Citrus reticulata	−	−	+	+	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−
Cunninghamia lanceolata	−	−	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−
Glycine soja	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−
Helianthus annuus	+	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Helianthus paradoxus	+	−	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Helianthus petiolaris	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	+	−	−	−	−	−
Rehmannia glutinosa	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Acacia mangium	−	+	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Avicennia marina	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Bruguiera cylindrica	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Bruguiera gymnorhiza	+	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Helianthus argophyllus	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−
Helianthus ciliaris	+	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Panax ginseng	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	+
Chlamydomonas reinhardtii	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Elaeis guineensis	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Gossypium herbaceum	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Helianthus exilis	−	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Populus euphratica	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Triticum turgidum	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Sum	51	47	45	43	41	40	37	37	37	36	36	34	33	33	33	32	32	31	26	26	23	17	16	15	15	13	12

TABLE 3 | Tissue specific miRNAs of PYTZ.

Tissue specific	miRNAs of PYTZ
Bud	miR1128-x, miR781-y, miR8590-y, miR870-y, miR9759-y
Leaf	miR1042-x, miR1045-x, miR1057-y, miR1063-x, miR1515-x, miR164-y, miR1866-y, miR2083-y, miR2111-y, miR2275-y, miR393-x, miR5061-y, miR5181-y, miR5523-y, miR5538-x, miR5653-y, miR6173-y, miR6281-x, miR7528-y, miR7711-x, miR7725-x, miR845-z, miR858-y, miR8665-y, miR9569-x, novel-m0703-5p, novel-m0722-5p, novel-m0945-3p, novel-m0953-5p, novel-m1089-3p, novel-m1125-5p
Stem	miR1127-x, miR1865-x, miR2275-x, miR3512-y, miR3630-x, miR4388-y, miR474-x, miR474-y, miR5021-x, miR5385-x, miR5834-x, miR6485-x, miR7713-x, miR7717-x, miR7762-y, miR8007-y, miR8681-y, miR9863-x, novel-m0024-3p, novel-m0095-3p, novel-m0155-3p, novel-m0331-5p, novel-m0392-5p, novel-m0828-3p, novel-m0906-3p, novel-m1037-3p, novel-m1040-3p, novel-m1041-5p

have their expression in specific tissues. Five miRNAs were bud-specific that could only be expressed in the bud, 31 were leaf-specific, and 28 were stem-specific, which had been summarized in **Table 3**.

Differentially Expressed miRNAs (DEM)

The differentially expressed miRNAs (DEM) were pairwise compared among sBud, sL and sS, with their expression values higher than a 2-fold change and $p \leq 0.05$, aiming to find out key miRNAs during the development. The numbers of DEM between tissues were summarized in **Figures 2C–G**. It's worth noting that according to the developmental order, there was a sharp down-trend numbers of DEM (235) and 79 up-trend DEM in sL1 than in sS1 (referred to as sL1 vs. sS1), compared to 67 down-trend and 5 up-trend DEM in sL1 vs. sL2, reminding that there are distinct regulatory changes and thus a metabolites accumulate differences bounded between L1 and S1 (**Figure 2C**). DEM with ups and downs both in G1 (**Figures 2D,F**) and G2 (**Figures 2E,G**) were also classified. The continuously changing ones were seemed to be possibly interesting regulators, such as 5 DEM (miR390-x, novel-m0578-5p, novel-m0634-5p, novel-m0503-3p, novel-m0531-5p) that had downtrend expression in G1 (**Figure 2F**), miR396-x uptrend in G1 (**Figure 2D**) and novel-m0331-5p (**Figure 2E**) uptrend in G2.

GO Enrichment and KEGG Pathway Analyses of DEM

miRNA sequences were searched against tea tree genomic sequences using the plant miRNA potential target finder to predict target mRNAs. The annotation of the target unigene of DEMs was conducted based on GO enrichment and KEGG analyses. In this study, a total of 5501 potential unigenes were predicted to be targeted by 934 miRNAs, including 138 conserved and 796 novel miRNAs. Among the miRNA, miR5385-x targeted the most unigenes (770), followed by miR5658-x (533), and miR8577-x (223). 280 miRNAs targeted one unigene, while most miRNAs could target multiple sites. Similarly, one unigene was also targeted by several miRNAs, and there were 1351 unigenes could be regulated by more than one miRNA, 58 of which were targeted by no <10 conserved miRNA. (The complete list of target genes of all miRNA were listed in **Supplementary Table 8**).

Gene Ontology (GO) enrichment analysis offered a strictly defined concept to describe properties of the target genes and recognize the main biological functions in a dynamic-updated controlled vocabulary. Within biological process categories, represented GO terms associated with these target genes in all tissues (**Supplementary Figure 4**) were related to "metabolic process" the most, followed by "cellular process" and "single-organismal process". Within cellular components categories, the unigenes were similarly represented, mainly in "cell," "cell part," "membrane," "organelle" and their parts. Within the molecular function categories, the top two GO terms were "catalytic activity" and "binding".

Key DEM Involved in Growth and Development

We focused on the expressions of miRNAs to filter the possible ones that participate in growth and development, and we found that all of the reported growth and development associate miRNAs belong to the up- or down- trends pattern, except miR172. So we firstly narrowed down miRNAs with similar expression trends in the fore-mentioned G1 and G2 (in which the down-trends including profile 3, 0 and 1; the up-trends including profile 6, 7 and 4), and then further screened by GO and KEGG pathway analysis. Twenty one miRNAs, including 6 novel miRNAs were screened out to be potential developmentally important miRNAs in PYTZ (**Table 4**). Mature sequences of these miRNAs and their target genes in tea genome were also listed in **Table 4**. The heat map of the 21 miRNAs which represent their transcription levels calculated by TPM in the samples were displayed in **Figure 4**. Not each miRNA has the same expression pattern in G1 compared with that in G2: 3 miRNAs only changed in G1, with 1 uptrend and 2 downtrends; 4 miRNAs only changed in G2, with 2 uptrends and 2 downtrends; 11 miRNAs had the same trends in both G1 and G2, with 3 uptrends and 8 downtrends. Interestingly, miR319-y had a downtrend expression pattern in G1 and uptrend in G2. Some other miRNAs hadn't been included in these trends might not because of their expression trends or levels, but the difference of expressions among development stages was not significant ($P < 0.05$, log2-fold change ≥ 2).

Potential Transcription Factor Target Genes

As transcription factors were intensely studied in their numerous important roles during plant growth and development in many species (Ramachandran et al., 1994; Zhang et al., 2009; Chen et al., 2010), we are here supposed to analysis transcription factor genes for identifying key TFs performing this function. All the 5,501 predicted target genes were blasted against Plant Transcription Factor Database (http://planttfdb.cbi.pku.edu.cn/), resulting in a total of 46 kinds of transcription factors involving 352 mRNAs were detected (The full list of identified TFs was provided in

TABLE 4 | Potential developmentally important miRNAs in *Pingyang tezaocha*.

miRNA	Expression pattern	Mature miRNA sequence (5′-3′)	Target mRNA ID in tea genome
miR156-x	G2 up	CUGACAGAAGAGAGUGAGCAC	CSA007311 CSA009012 CSA011373 CSA013149 CSA017838 CSA019508 CSA020439 CSA023442 CSA031667
miR160-x	G1/G2 down	UGCCUGGCUCCCUGUAUGCCA	CSA021765 CSA026035 CSA026847
miR164-x	G1/G2 up	UGGAGAAGCAGGGCACGUGCA	CSA013362 CSA027185 CSA033493
miR165-y	G1 up	UCGGACCAGGCUUCAUCCCCU	CSA023057 CSA030874
miR166-x	G1/G2 down	GGAAUGUUGGCUGGCUCGAUG	CSA001544 CSA017082 CSA028252 CSA028940 CSA031362 CSA031363
miR166-y	-	UCGGACCAGGCUUCAUUCCCC	CSA030874
miR166-z	-	UCGGACCAGGCUUCAUUCCCU	CSA030874
miR319-x	G1down	GAGCUUCCUUCUGUCCACUU	CSA003974 CSA005021 CSA007451 CSA014848 CSA017867 CSA018024 CSA018464 CSA021711 CSA036427
miR319-y	G1down G2 up	UUGGACUGAAGGGAGCUCCCU	CSA002826 CSA013833 CSA031080 CSA036087
miR390-x	G1/G2down	AAGCUCAGGAGGGAUAGCGCC	CSA001776 CSA001956 CSA006028 CSA007412 CSA011969 CSA016385 CSA022179 CSA026282 CSA030630 CSA036196
miR396-x	G1/G2 up	UUCCACAGCUUUCUUGAACUU	CSA003330
miR396-y	G2 up	GCUCAAGAAAGCUGUGGGAAG	CSA008398
miR5083-y	G1/G2 down	CUACAAUUAUCUGAUCAAA	CSA036373
miR8175-y	G2 down	UCCCCGGCAACGGCGCCA	CSA013921 CSA018936
miR8577-x	G1 down	UGAGAUGAUGAUCAUGAU	CSA008171 CSA030921
novel-m0675-3p	G1/G2 down	GAAUAUGAUGAAUUUGAAUG	CSA006098
novel-m0297-3p	G1/G2 down	ACCCCUAACCCCAACACCCAAUC	CSA012411
novel-m0243-3p	G1/G2 down	GAUCAGGAUGAAGCAACAUU	CSA027591 CSA034664
novel-m0284-3p	G1/G2 down	UUGGGCUGGGCAGAAAUUGGGC	CSA016454
novel-m0800-3p	G1/G2 up	GUUCAGUGAAGCUGUGGAAAG	CSA010612 CSA028516
novel-m0187-3p	G2 down	AUUUCCCUUUCCAAAUUCCUU	CSA012066 CSA024775

The underlined genes are the targeted mRNA belonged to TF genes. The gray shaded ones are the genes have reciprocal expression profiles with responding miRNAs.

Supplementary Table 9). Then, types and numbers of TFs genes targeted by miRNAs in each tissue (sBud, sL1, sL2, sS1, and sS2) were analyzed and summarized in **Figure 5**. On the whole, the numbers of transcription factors genes targeted by miRNAs in sL2 were the most, and that in sS1 was the least. MYB showed the widest involvement, followed by HD-ZIP and bHLH (basic Helix-Loop-Helix) transcription factors. Some kinds of transcription factors couldn't be targeted in all tissues, such as that the CAMTA family of calmodulin binding TF genes in sS2 and GRF (Growth Regulating Factor) in sL1.

Quantitative PCR was further performed to validate the mentioned 21 growth and development associate miRNAs (**Figure 6**) and their predicted target mRNAs (**Figure 7**), among which 14 of them were TF genes (underlined in **Table 4**). Universally, plant miRNAs might be involved in many complicated and diverse functions in the complex regulatory networks, the fundamental role of miRNAs is to suppress the expression of target genes (Tang and Chu, 2017). Herein, we got eight miRNA-TF genes pairs with reciprocal expression profiles (gray shaded mRNAs in **Table 4**), and the complementary correspondence of miRNA toward the target sites were shown in **Supplementary Figure 5**. The miRNA-TF genes pairs were miR156-x-CSA011373 (SBP), miR156x-CSA019508 (SBP), miR156x-CSA023442 (SBP), miR156x-CSA031667 (SBP), miR165y-CSA023057 (Class III HD-Zip), miR165y-CSA030874 (HD-Zip), miR319y-CSA036087 (MYB), and miR8577x-CSA030921 (bHLH).

Expression Profiles of Reported Morphological miRNA in Different Tea Varieties

Originated in Yunnan and Tibet region of China, the tea tree has been evolved over thousands of years and now at least 246 cultivars have been selected breeding with significant differences in morphology and physiology. Such as the focused cultivar in this study, PYTZ, has oblong leaf shape, blunt tip, tight dentate margins, and shorter internode. These characteristics are one of the indicators of screening and distinguishing cultivars. Plant growth and development are accompanied by morphogenesis that some regulators including small RNAs and TFs may participate in both biological processes simultaneously. In order to do some basal research linking developmental associate miRNA toward morphology, the expression profiles of reported morphological miRNAs were performed in other six tea varieties with obvious differences in leaf morphology. The focused 21 miRNAs were again checked by quantitative PCR for their expression levels in several representative cultivars. Bud, the 1st leaf, the 2nd leaf, the younger stem, and the older stem were also sampled from each variety in each group for quantitative PCR analysis (**Figure 8**). For each miRNA, the relative expression in JFH was set as the reference so as to get better understanding of the expression levels among varieties. In general, most of the miRNAs had similar expression patterns in ZYD, PYTZ, DYWL, and QS, with high expression levels in the 1st leaf and

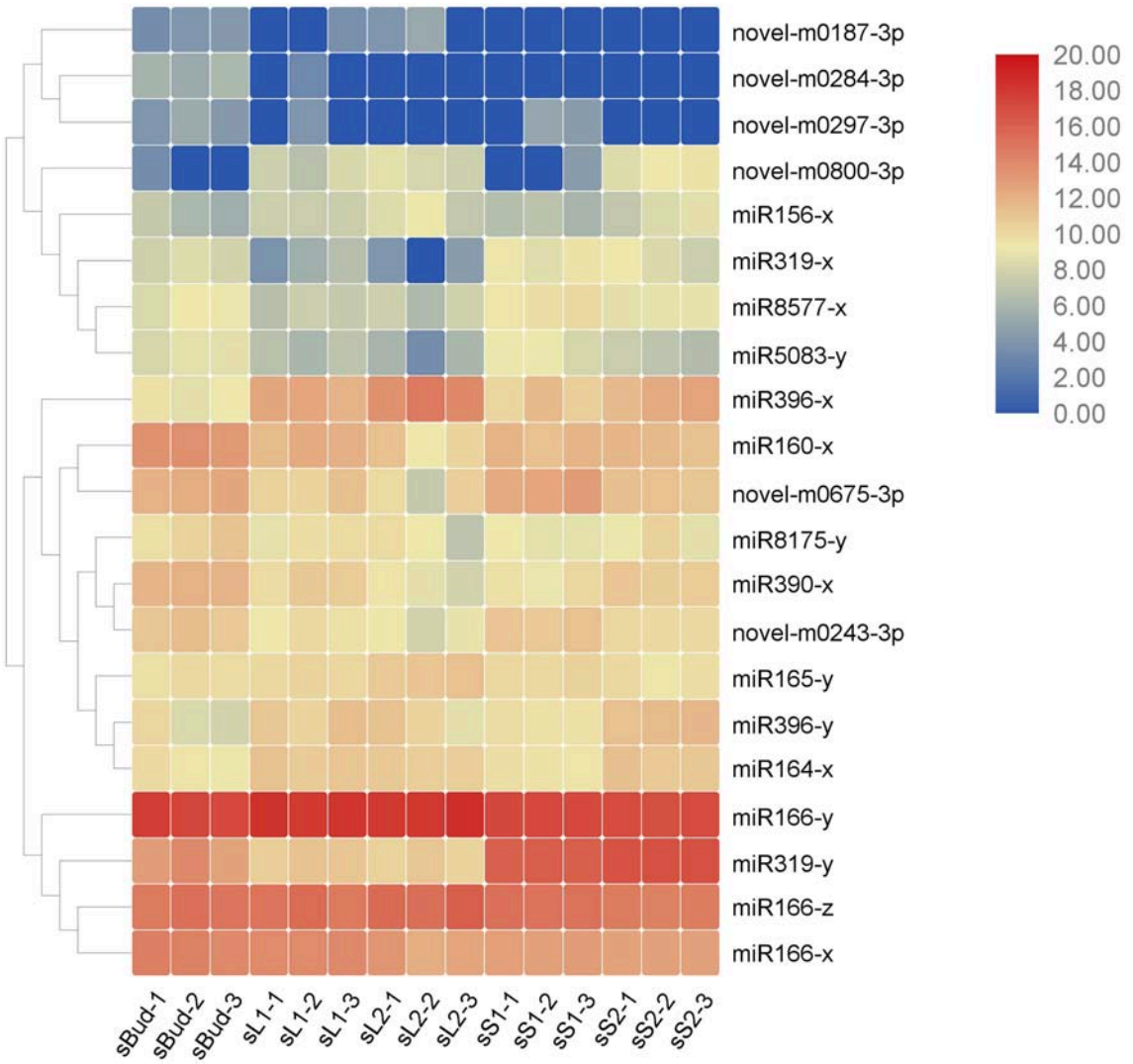


FIGURE 4 | Heat map of the 21 miRNAs by the value of TPM in the 15 samples. The three major clusters on the left showed evolutionary relationships of the miRNAs, with closely related miRNAs neighboring clustered. Similar expression patterns could be found within one cluster.

then the 2nd leaf. And for other miRNAs were likely to have high expression levels in the bud, such like miR160x in HD and miR156y in JFH.

Network Analysis on miRNA, Target mRNA, and Quality-Related Metabolites

Correlation analyses were conducted to figure out the extent of the 21 conserved miRNA mentioned above participated in the quality formation progress (**Figure 9**). Surprisingly, compared to catechins and caffeine, theanine had a quite strong relationship and relatively large numbers with target mRNAs, which could be definitely set as the center regulated metabolite, at least during the development of PYTZ sprouts. For the galloylated catechins, ECG was regulated by much more multiple mRNAs than EGCG. The kinds of mRNA participated in regulating the nongalloylated catechins were approximately equal. As for the conserved miRNA, miR8577-x, miR160-x, and novel-m0187-3p were the most social ones. However, the neighborhood relationships for miR156-x and miR319-y seemed rather simple.

DISCUSSION

Different miRNAs Were Involved in Different Tissues and Stages During the Sprout Development in PYTZ

How a plant builds leaves from a few cells that grow, divide, and differentiate to form into the complex organ has been well-studied, the same with the research of mechanical regulation (Braybrook and Kuhlemeier, 2010; Chen, 2012; Qi et al., 2017). miRNA, typically multigene families, allowing for subtlety and complexity of control in different regulatory processes, are

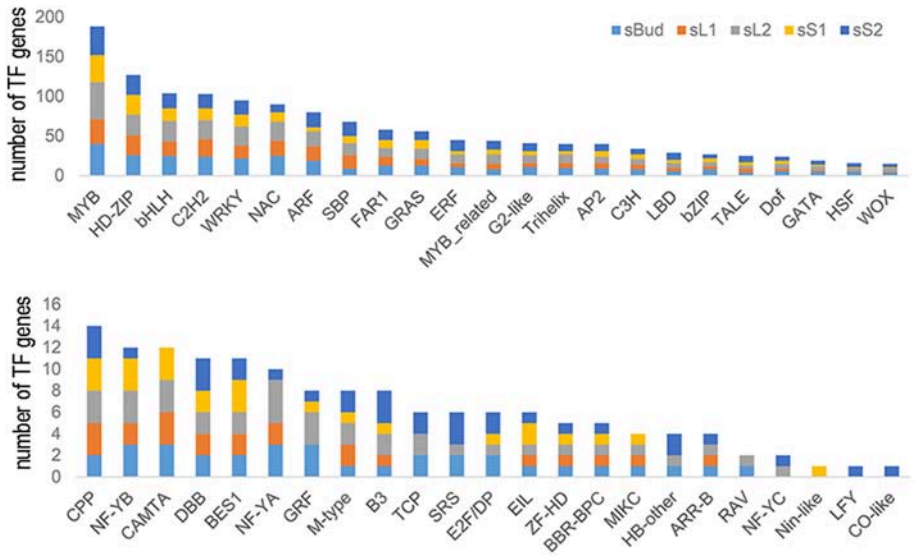


FIGURE 5 | Types and numbers of transcription factors (TFs) genes targeted by miRNAs in each tissue.

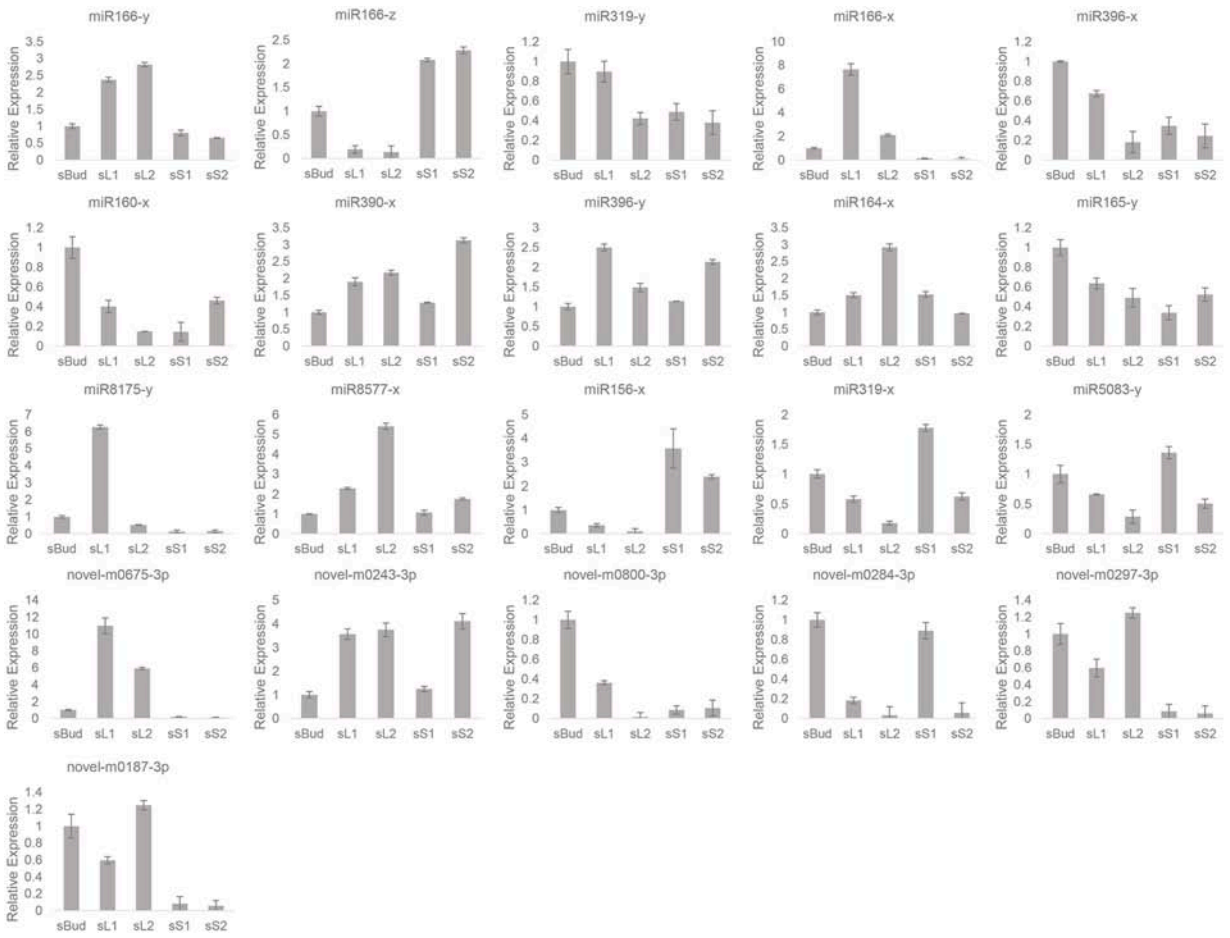


FIGURE 6 | Relative expressions of the 21 miRNA by quantitative PCR.

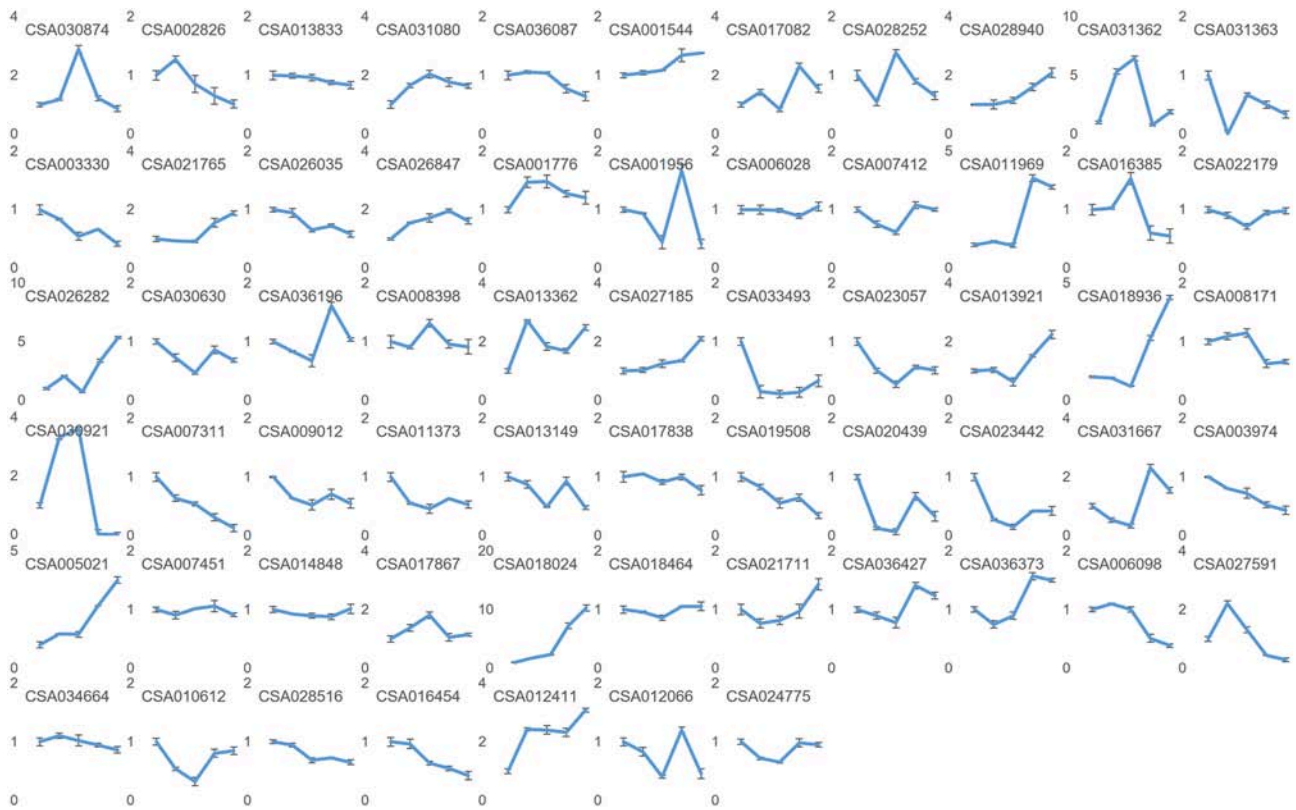


FIGURE 7 | Relative expressions of the predicted target mRNA genes by quantitative PCR. Nodes on axis X for each inset image from left to the right are: sBud, sL1, sL2, sS1, and sS2. Designations the same as on previous figure.

described as factors in many aspects of plant development (Kidner, 2010).

As tea is the leaf-using crash plant, the performance of leaf development has more practical significance. Thus, we took the perspective of looking at the function of miRNA in different tissues and development stages separately. In the comparison of G1 and G2 mentioned above, miRNAs that participate in leaf and stem development are different, hinting that different miRNAs need coordinate working in leaf and stem developmental process, separately. For example, both novel-m0155-3p and novel-m0331-5p were stem-specific miRNAs and have high expression levels in sS2 than in sS1, which had strong possibilities responsible for stem elongation. And likewise, one kind of miRNA may function differently in different tissues development. In the 78 miRNAs that both exist in leaf and stem (**Figure 2B**), 28 of them have different including miR5049y and other 27 novel miRNAs could be found both in leaf and stem, in which 23 of them share one similar expression pattern (**Supplementary Figure 6**) ($P < 0.05$, log2-fold change ≥ 1.5), 27 of them with the expression levels in sL1 higher than in sL2 (**Supplementary Figures 6A–C**), 24 of them with the expression levels in sS2 higher than in sS1 (**Supplementary Figures 6A,D**). The non-conserved miR5049, included in profile1 (**Supplementary Figure 6A**), had been reported to be drought stress response miRNA in the root of drought-tolerant cultivar wheat (Akdogan et al., 2016).

Interestingly, the number of down-trend miRNAs (157 from profile 3, profile 0 and profile 1) are much more than up-trend ones in G1 (53 from profile 6, profile 7 and profile 4) (**Figure 3A**) and in G2 (91 vs. 64) (**Figure 3B**), this is the same case for the abundance of these miRNAs. Discarding the ones with TPM lower than 100, the abundance of down-trend miRNAs occupied 79.04% and 54.30% in G1 and G2, separately (**Supplementary Table 10**). The percentage indicated that during the development, especially in leaf, mRNAs regulated by miRNA have a large percentage in increasing tendency. This trend is consistent with the expression levels from tea leaf transcriptome that 72% of genes were up-regulated in the second leaf stage compared to the first leaf stage (Guo et al., 2017). The coherence of expression levels of the regulator and the content of secondary metabolites is particularly impressive, a point we return to below.

Evolutionarily Conserved miRNAs Were Closely Connected to Morphogenesis Functions During the Sprout Development

miRNA is usually be concerned whether to be conserved or not, which typically depends on their degree of presentation in all or at least most of the species, and thus the division could be influenced by sampling and the phylogenetic diversity of available species that miRNAs have been characterized and annotated (Baldrich et al., 2018). In the 21 developmental associated

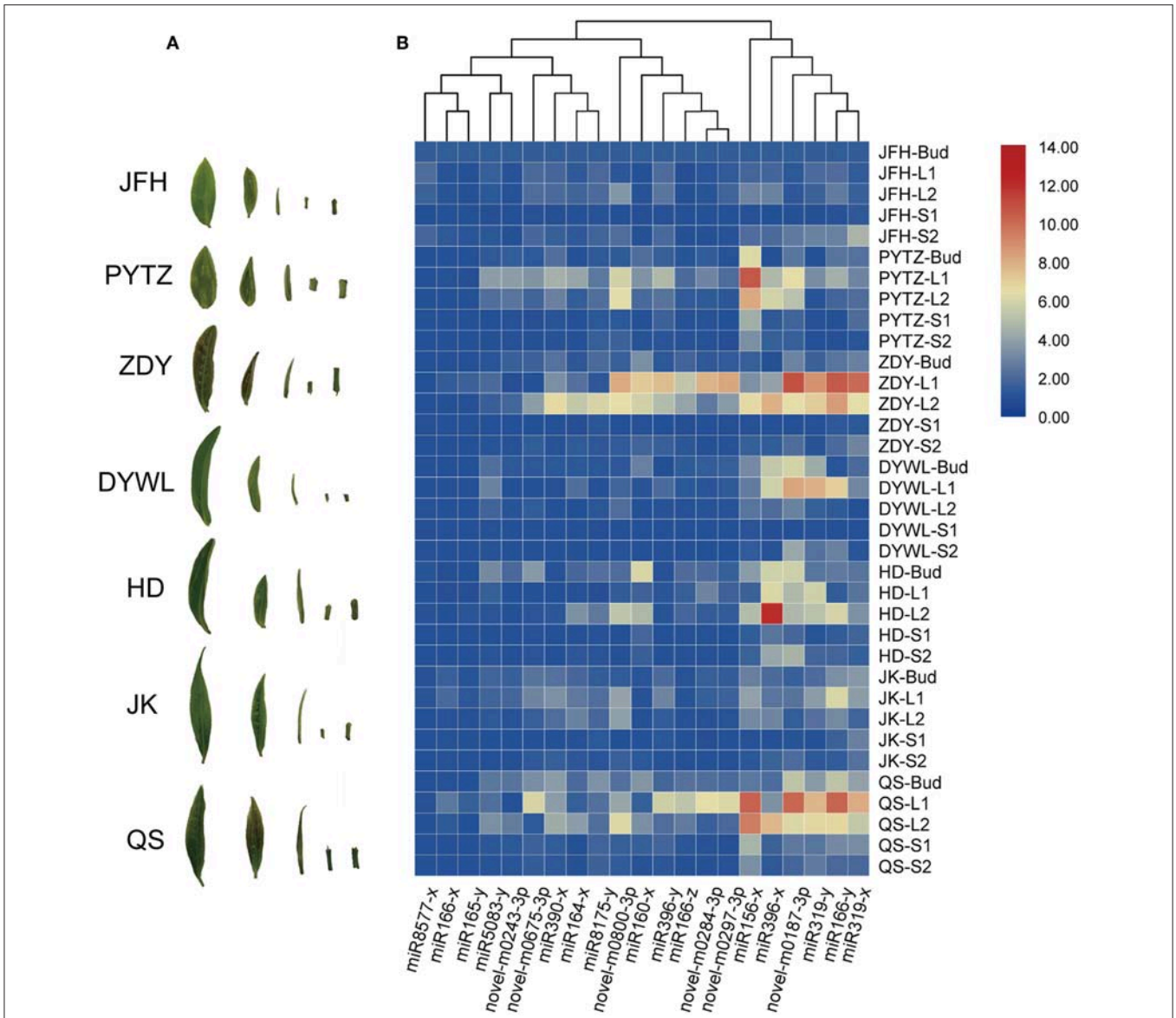


FIGURE 8 | Heat map of the relative expression levels of 21 miRNAs in different tissues by Real-time PCR. **(A)** Samples from tea tree cultivars corresponding to **(B)**. Heat map of the relative expression levels of 21 miRNAs in different tissues from the *Camellia sinensis* "Jinfenghuang" (JFH), *Camellia sinensis* "Pingyang Tezaocha" (PYTZ), *Camellia sinensis* "Zhengdayin" (ZDY), *Camellia sinensis* "Dayewulong" (DYWL), *Camellia sinensis* "Huangdan" (HD), *Camellia sinensis* "Jiukeng 6" (JK), *Camellia sinensis* "Queshe" (QS).

miRNAs filtered out in this study (**Table 4**), except for 6 novel miRNAs, 11 of the known miRNAs are conserved miRNAs that could be found in at least 33 plant species out of 72 plant species (**Table 2**). Notably, none of the miRNAs from tea could be found in *Chlamydomonas reinhardtii*, which is a conventionally model for a photosynthetic cell in studying photosynthesis (Funes et al., 2007), abiotic stress (Hema et al., 2007), circadian clock (Ral et al., 2006) and so on. This result to some extent is an echo of the conservation of miRNA within one kingdom, and no miRNA had been found conserved in green algae and land plants (Baldrich et al., 2018). Interestingly, there was no tea miRNA found in *Populus euphratica* either, which is an ideal model system of woody plants for research into the abiotic stress resistance (Li et al., 2009), such like drought (Li et al., 2011) and salt (Li et al., 2013). Previously studies on *Populus euphratica* had been reported that only 9 out of 21 miRNAs families (miR156; miR163; miR172; miR398; miR393; miR171; miR408; miR169; miR472) were conserved in other plants, with other 12 miRNA family candidates show none homologies in *Populus*, *Arabidopsis*, and *Oryza* (Li et al., 2009) and can thus be considered as quite ancient and independent evolution species.

The seven tea cultivars studied above were famous cultivars that frequently used in producing fermented or non-fermented tea in China. Typically in the tea processing industry, tea-processing suitability and tea quality are basically determined by the main characteristic metabolic compounds, which directly

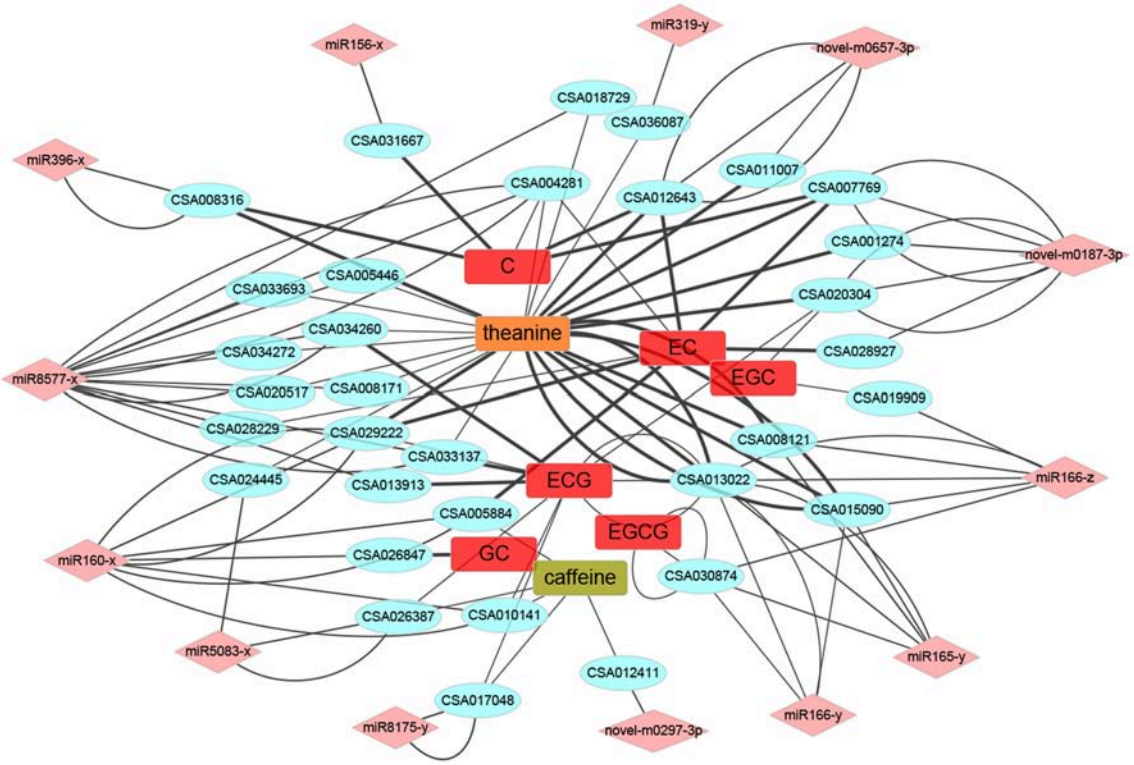


FIGURE 9 | The miRNA-mRNA-metabolites association networks. The round rectangle placed in the center were several metabolites, the ellipses at the interlayer were potential target mRNAs, and the diamond at the outermost layer were miRNAs. Line thickness represented the strength of the relationship.

linked up with the development and morphogenesis of the tea sprouts (Xia et al., 2017). In the 11 known conserved tea miRNA, miR156 was the most popular one, which was found in 51 plant species, followed by miR396 and miR166, which were conserved in 47 and 45 plant species, respectively. Similar with their high abundance (miR166, miR319, miR396, miR160, and miR390 were listed in the top five kinds of miRNAs), which means that conservation is not only represent low sequence variation across diverse plant species, but also to be the large and older miRNA families with abundant copy and target number (Chavez Montes et al., 2014), in order to grantee their tightly constrained roles in function and less gene loss in the regulatory network (Shi et al., 2017). Mature miRNAs in plant often have multiple target genes with similar complementary sequences, among which these evolutionarily conserved miRNAs and their predominantly target genes characteristically play essential roles in developmental regulation, morphogenesis, stress responses (Axtell and Bowman, 2008; Yang et al., 2013). Even more noteworthy is, the tissue-specific miRNAs (**Table 2**) may contribute to the development of the specific tissue, which doesn't mean they are the dominated ones and conserved ones, either. For example, miR319 had been reported to increase the number of longitudinal small veins thus might account for the leaf blade width (Yang et al., 2013) and miR159 was involved in stem elongation (Tsuji et al., 2006), but they are both conserved miRNAs and have expression in bud, leaves, and stems. Only miR164, miR393 and miR2111 were leaf-specific miRNAs and conserved miRNAs as well. Not all of the conserved miRNAs have similar expression patterns in the investigated cultivars (**Figure 8**), which might be the result of the flexible of the "fine-tuners," to enhance the ability of a fast response to evolution (Muleo, 2012). The conservation in sequence doesn't always represent functional conservation (Ason et al., 2006). Though the evidence of miRNA in the plant is less than that in the animal, it is widely accepted that plants miRNA genes are evolved independently as they do in the animal kingdoms. It is thus believed that the larger the miRNA family is, which means the more multiple paralogous copies of one miRNA in plants, brings more flexible of evolution rate and more possibility in diversification though, the more essential function in development it may be involved. More research combined the contents of metabolites with the function of conserved miRNAs in species-level phenotypic differences needs to be further studied.

Development Associate miRNAs Might Play Crucial Roles in the Quality Formation Together With Their Potential Target Transcription Factor Genes

Higher plants evolved precise and robust spatio-temporal patterns of gene regulatory systems, among which transcription factors and miRNAs are two of the best studied regulatory mechanisms separately at transcriptional and post-transcriptional level (Chen and Rajewsky, 2007). TFs and

TABLE 5 | GO pathway enrichment analysis to the target mRNA genes of DEM between each two samples from Bud, sL1, sL2, sS1, and sS2.

Characteristic pathway	Metabolic pathway	Annotated genes	Related mRNA	miRNA
Flavonoid pathway	Phenylalanine metabolism	Amidase	CSA014307	miR8577-x
		Omega-amidase	CSA016772	novel-m0263-3p
	Phenylpropanoid biosynthesis	Cinnamyl-alcohol dehydrogenase (CAD)	CSA019604	novel-m0018-3p, novel-m0119-3p, novel-m0279-3p, novel-m0498-3p, novel-m0608-3p, novel-m0674-3p, novel-m0888-3p, novel-m0904-5p, novel-m0977-3p, novel-m1032-3p, novel-m1080-3p, novel-m1106-3p
			CSA005610	miR1511-y, miR5139-x, miR8155-y, novel-m0999-3p
			CSA009196	miR6118-y
		Peroxidase	CSA026001	miR5658-x, novel-m0675-3p
		Beta-glucosidase	CSA014637	novel-m0349-3p, novel-m0360-3p, novel-m0428-3p
		Shikimate O-hydroxycinnamoyltransferase (SHT)	CSA011696	miR8577-x
	Flavonoid biosynthesis	Naringenin 3-dioxygenase (F3H)	CSA004930	miR8577-x
		Chalcone isomerase (CHI)	CSA006623	novel-m0466-3p, novel-m0809-3p
	Anthocyanin biosynthesis	Anthocyanidin 3-O-glucoside 5-O-glucosyltransferase (UGT75C1)	CSA036671	miR1865-x, novel-m0674-3p, novel-m1032-3p
		UGAT	CSA013643	miR4995-x, miR6483-y
Caffeine Pathway	Caffeine Metabolism	Urate oxidase (UOX)	CSA020658	novel-m0021-5p, novel-m0508-3p, novel-m0698-3p, novel-m0834-3p, novel-m0964-3p, novel-m0964-3p
		Xanthine dehydrogenase/oxidase	CSA006612	miR5059-x
Amino acid Pathway	Citrate cycle (TCA cycle)	Isocitrate dehydrogenase (NAD+)	CSA018911	miR5083-y
		Pyruvate dehydrogenase E2 component (dihydrolipoamide acetyltransferase)	CSA034852	miR8175-y
		2-oxoglutarate dehydrogenase E1 component	CSA032631	novel-m0381-3p
		2-oxoglutarate dehydrogenase E2 component (dihydrolipoamide succinyltransferase)	CSA019997	novel-m0184-5p
		ATP citrate (pro-S)-lyase	CSA033096 CSA027464 CSA034373	miR5059-x
		Succinyl-CoA synthetase beta subunit	CSA027414	novel-m0018-3p, novel-m0119-3p, novel-m0279-3p, novel-m0498-3p, novel-m0608-3p, novel-m0674-3p, novel-m0888-3p, novel-m0904-5p, novel-m0977-3p, novel-m1032-3p, novel-m1080-3p, novel-m1106-3p
		Aconitate hydratase	CSA024268 CSA025056 CSA020729	novel-m0945-3p
		Fumarate hydratase, class II	CSA004158	novel-m0574-5p
			CSA036624	miR5054-y
		Succinyl-CoA synthetase beta subunit	CSA027414	novel-m0018-3p, novel-m0119-3p, novel-m0279-3p, novel-m0498-3p, novel-m0608-3p, novel-m0674-3p, novel-m0888-3p, novel-m0904-5p, novel-m0977-3p, novel-m1032-3p, novel-m1080-3p, novel-m1106-3p
	Alanine, aspartate and glutamate metabolism	Glutamate dehydrogenase (NAD(P)+)	CSA030852 CSA002178 CSA019055	novel-m0541-3p
		Alanine transaminase	CSA034511	novel-m0410-3p, novel-m0435-3p, novel-m0549-3p, novel-m0553-5p, novel-m0554-5p, novel-m0752-5p, novel-m0769-3p, novel-m0863-3p, novel-m0938-3p, novel-m1097-3p, novel-m1123-5p
		4-aminobutyrate—pyruvate transaminase	CSA023854	novel-m0074-3p, novel-m0075-3p
		Calcium-binding protein CML	CSA036463	novel-m1108-5p
			CSA009111	
		Amidophosphoribosyl transferase	CSA032114	miR5059-x, miR6281-x
		3-deoxy-7-phosphoheptulonate synthase	CSA021087	novel-m1042-3p
			CSA011007	novel-m0657-3p, novel-m1042-3p
		Arogenate/prephenate dehydratase	CSA006958	miR5054-y, novel-m0388-3p

miRNAs generally do not work isolation, but instead, together with co-regulators in the same layer or not, to form large networks of cooperating and interacting in complex multicellular organisms (Dawid, 2006). But they are usually positioned at the center of regulating many aspects of developmental plasticity along with the life cycle (Rubio-Somoza and Weigel, 2011). We perform network and enrichment analysis to the 352 TF genes targeted by miRNA in STRING (Szklarczyk et al., 2017), and marked the developmental relative process (**Supplementary Figure 7**). Most of them were involved were clustered in some certain pathway in regulating gene expression and primary metabolic process, which could be easily understood that TFs were involved in the primary metabolic process because of the fundamental maintenance of living for plant themselves. In shrinking the research objectives, we further perform GO analysis to the DEM in G1 and G2 above, many pathways were enriched in developmental and morphological process (**Supplementary Figure 8**), especially in the G1 up trend expression profiles, which also confirmed us the effective way of filtering key miRNAs. The eight miRNA-TF mRNA pairs verified by their reciprocal expression relationships were much more likely to participate in development regulation, which doesn't imply only this eight pairs of miRNA-TFs were involved.

Metabolism was along with the development. It has been reported that plant miRNA were widely involved in quality formation regulation (Wu et al., 2014; Liu et al., 2017b). As the three kinds of characteristic metabolites which finally determine the quality of tea (Xia et al., 2017; Wei et al., 2018), catechins mainly confer astringent taste, theanine contributes to the umami and sweet tastes, and caffeine offers a bitter taste (Wei et al., 2018). We did GO pathway enrichment analysis to the target mRNA genes of DEM between every two samples from Bud, sL1, sL2, sS1, and sS2, with the evolved participants showed in **Table 5**. The correlation of the chemical analysis on catechins, theanine, caffeine and the soluble matter would finally affect the sensory evaluation of green tea taste. In our study, we found that theanine turned out to be even more active in the network (**Figure 9**). Target mRNAs which belonging to TF genes were further picked out and constructed a more metabolic directivity one due to their correlation (**Supplementary Figure 9**). TF genes like CSA013022 (HD ZIP) and CSA029222 (ARF) had a strong positive relationship with the biosynthesis of theanine, while the later also positively regulated EC, referring to miR165-y, miR166-z, and miR160-x, respectively. CSA031667, a TF gene belonging to SBP family, had a positive correlation with C, and be controlled by miR156-x. When taking the eight miRNA-TF pairs mentioned above into consideration together, there were at least two triplets that participate in both development and quality formation: miR156-x-CSA031667 (SBP)-C and miR319-y-CSA036087 (MYB)-theanine. Molecular mechanisms of sprout development and accumulation of metabolites would be gradually uncovered after the release of tea tree genome (Xia et al., 2017; Wei et al., 2018) and tea organic transcriptomes (Zheng et al., 2015, 2016; Guo et al., 2017; Liu et al., 2017a). More connections would be further studied toward small RNAs to improve breeding efficiency of developing better cultivars with higher quality.

DATA AVAILABILITY

The datasets generated for this study can be found in Sprouts development of tea plants, PRJNA510482.

ETHICS STATEMENT

The authors declare that we have complied with all relevant ethical regulations.

AUTHOR CONTRIBUTIONS

ZD conceived the study. LZ, CC, JS, and YW performed the experiment and analyzed the data. LZ wrote the paper. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Special Foundation for Distinguished Taishan Scholar of Shandong Province (Ts201712057), the Natural Science Foundation of China (31600557 and 31470027), Science and Technology Plan Projects in Colleges and Universities of Shandong Province (J15LF02), School Fund Project of Qingdao Agricultural University (631412), Qingdao Applied Basic Research Program (grant 15-9-1-45-jch). This work was also supported by China Scholarship Council (201708370012).

ACKNOWLEDGMENTS

We acknowledge Gene *de novo* Co., Ltd. and RuiBo Co., Ltd. at Guangzhou and for their assistance in original data processing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00237/full#supplementary-material

Supplementary Figure 1 | The nucleic acids frequency of known miRNAs at each position in PYTZ. sBud **(A)**, sL1 **(B)**, sL2 **(C)**, sS1 **(D)**, and sS2 **(E)**. The frequency of cytosine (C) (32.09%) and uracil (U) (29.65%) are higher than guanine (G) (19.26%) and adenosine (A) (19.00%). U had a high appearance at the 1st, 17th, 22th, and 23th positions, with an average of 84.34, 61.80, 54.57, and 52.48% respectively.

Supplementary Figure 2 | The nucleotide bias of 18nt-30nt length known miRNAs at the 1st position in PYTZ. sBud **(A)**, sL1 **(B)**, sL2 **(C)**, sS1 **(D)**, and sS2 **(E)**.

Supplementary Figure 3 | Related coefficients of the 15 miRNA libraries across the five sample groups with three replicates of each group.

Supplementary Figure 4 | Level 2 GO terms of all target genes from PTZY sprouts.

Supplementary Figure 5 | Complementary correspondence of developmental miRNA toward the target sites.

Supplementary Figure 6 | (A–E) The expression profiles of miRNA that exist both in sL (leaf) and sS (stem).

Supplementary Figure 7 | Relative process of the targeted TF genes by STRING analysis. (STRING: https://string-db.org/).

Supplementary Figure 8 | Level 3 GO analysis of DEM in G1 and G2.

Supplementary Figure 9 | The miRNA-TF-metabolites networks. The round rectangle placed in the center were metabolites, the ellipses at the interlayer were potential target mRNAs that belong to TF genes, and the diamond at the outermost layer were miRNAs. Line thickness represented the strength of the relationship. The red line represented the positive correlation efficient and the blue line meant negative.

Supplementary Table 1 | The primers sequences of miRNA and mRNA genes used in Real-time PCR.

Supplementary Table 2 | The filtering data of 15 sRNA-Seq libraries from *Camellia sinensis* cv. Pingyang Tezaocha.

Supplementary Table 3 | Statistics of sRNA-Seq libraries mapping to tea tree genome.

Supplementary Table 4 | Statistics of sRNA-Seq libraries mapping to Rfam.

Supplementary Table 5 | The lengths, sequences, and expressions of 156 known miRNAs.

Supplementary Table 6 | 122 precursors and 99 kinds of characteristic hairpin structures, with their length and energy information.

Supplementary Table 7 | 1186 novel miRNAs and their 1130 hairpin structures, with their length and energy information.

Supplementary Table 8 | The complete list of target genes of all miRNA.

Supplementary Table 9 | The annotation and the corresponding miRNAs of the total 352 target transcription factor genes.

Supplementary Table 10 | Down-trend and up-trend miRNAs in G1 and G2 with TPM more than 100.

REFERENCES

Akdogan, G., Tufekci, E. D., Uranbey, S., and Unver, T. (2016). miRNA-based drought regulation in wheat. *Funct. Integr. Genom.* 16, 221–233. doi: 10.1007/s10142-015-0452-1

Ason, B., Darnell, D. K., Wittbrodt, B., Berezikov, E., Kloosterman, W. P., Wittbrodt, J., et al. (2006). Differences in vertebrate microRNA expression. *Proc. Natl. Acad. Sci. U.S.A.* 103, 14385–14389. doi: 10.1073/pnas.0603529103

Axtell, M. J., and Bowman, J. L. (2008). Evolution of plant microRNAs and their targets. *Trends Plant Sci.* 13, 343–349. doi: 10.1016/j.tplants.2008.03.009

Baldrich, P., Beric, A., and Meyers, B. C. (2018). Despacito: the slow evolutionary changes in plant microRNAs. *Curr. Opin. Plant Biol.* 42, 16–22. doi: 10.1016/j.pbi.2018.01.007

Benjamin, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R S. Soc. Series B.* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berger, Y., Harpaz-Saad, S., Brand, A., Melnik, H., Sirding, N., Alvarez, J. P., et al. (2009). The NAC-domain transcription factor GOBLET specifies leaflet boundaries in compound tomato leaves. *Development* 136, 823–832. doi: 10.1242/dev.031625

Braybrook, S. A., and Kuhlemeier, C. (2010). How a plant builds leaves. *Plant Cell* 22, 1006–1018. doi: 10.1105/tpc.110.073924

Chavez Montes, R. A., De Fatima Rosas-Cardenas, F., De Paoli, E., Accerbi, M., Rymarquis, L. A., Mahalingam, G., et al. (2014). Sample sequencing of vascular plants demonstrates widespread conservation and divergence of microRNAs. *Nat. Commun.* 5:3722. doi: 10.1038/ncomms4722

Chen, C., Xia, R., Chen, H., and He, Y. (2018). TBtools, a Toolkit for Biologists integrating various HTS-data handling tools with a user-friendly interface. *bioRxiv [Preprint]*. doi: 10.1101/289660

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* 8, 93–103. doi: 10.1038/nrg1990

Chen, X. (2009). Small RNAs and their roles in plant development. *Annu. Rev. Cell Dev. Biol.* 35, 21–44. doi: 10.1146/annurev.cellbio.042308.113417

Chen, X. (2012). Small RNAs in development - insights from plants. *Curr. Opin. Genet. Dev.* 22, 361–367. doi: 10.1016/j.gde.2012.04.004

Chen, X., Zhang, Z., Liu, D., Zhang, K., Li, A., and Mao, L. (2010). SQUAMOSA promoter-binding protein-like transcription factors: star players for plant growth and development. *J. Integr. Plant Biol.* 52, 946–951. doi: 10.1111/j.1744-7909.2010.00987.x

Dawid, I. B. (2006). The Regulatory Genome, by Eric H. Davidson, Academic Press. *FASEB J.* 20, 2190–2191. doi: 10.1096/fj.06-1103ufm

De Lima, J. C., Loss-Morais, G., and Margis, R. (2012). MicroRNAs play critical roles during plant development and in response to abiotic stresses. *Genet. Mol. Biol.* 35, 1069–1077. doi: 10.1590/S1415-47572012000600023

Engstrom, E. M., Izhaki, A., and Bowman, J. L. (2004). Promoter bashing, microRNAs, and Knox genes. new insights, regulators, and targets-of-regulation in the establishment of lateral organ polarity in arabidopsis. *Plant Physiol.* 135, 685–694. doi: 10.1104/pp.104.040394

Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* 7:191. doi: 10.1186/1471-2105-7-191

Fan, Z., Li, J., Li, X., Wu, B., Wang, J., Liu, Z., et al. (2015). Genome-wide transcriptome profiling provides insights into floral bud development of summer-flowering Camellia azalea. *Sci. Rep.* 5:9729. doi: 10.1038/srep09729

Ferreira e Silva, G. F., Silva, E. M., Azevedo Mda, S., Guivin, M. A., Ramiro, D. A., Figueiredo, C. R., et al. (2014). microRNA156-targeted SPL/SBP box transcription factors regulate tomato ovary and fruit development. *Plant J.* 78, 604–618. doi: 10.1111/tpj.12493

Funes, S., Franzen, L. G., and Gonzalez-Halphen, D. (2007). Chlamydomonas reinhardtii: the model of choice to study mitochondria from unicellular photosynthetic organisms. *Methods Mol. Biol.* 372, 137–149. doi: 10.1007/978-1-59745-365-3_10

Guo, F., Guo, Y., Wang, P., Wang, Y., and Ni, D. (2017). Transcriptional profiling of catechins biosynthesis genes during tea plant leaf development. *Planta* 246, 1139–1152. doi: 10.1007/s00425-017-2760-2

Hema, R., Senthil-Kumar, M., Shivakumar, S., Chandrasekhara Reddy, P., and Udayakumar, M. (2007). Chlamydomonas reinhardtii, a model system for functional validation of abiotic stress responsive genes. *Planta* 226, 655–670. doi: 10.1007/s00425-007-0514-2

Ivanisenko, V. A., Demenkov, P. S., Ivanisenko, T. V., Mishchenko, E. L., and Saik, O. V. (2019). A new version of the ANDSystem tool for automatic extraction of knowledge from scientific publications with expanded functionality for reconstruction of associative gene networks by considering tissue-specific gene expression. *BMC Bioinformatics* 20:34. doi: 10.1186/s12859-018-2567-6

Jeon, D. B., Hong, Y. S., Lee, G. H., Park, Y. M., Lee, C. M., Nho, E. Y., et al. (2017). Determination of volatile organic compounds, catechins, caffeine and theanine in Jukro tea at three growth stages by chromatographic and spectrometric methods. *Food Chem.* 219, 443–452. doi: 10.1016/j.foodchem.2016.09.184

Jiang, X., Liu, Y., Li, W., Zhao, L., Meng, F., Wang, Y., et al. (2013). Tissue-specific, development-dependent phenolic compounds accumulation profile and gene expression pattern in tea plant [Camellia sinensis]. *PLoS ONE* 8:e62315. doi: 10.1371/journal.pone.0062315

Jones-Rhoades, M. W. (2012). Conservation and divergence in plant microRNAs. *Plant Mol. Biol.* 80, 3–16. doi: 10.1007/s11103-011-9829-2

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480–484. doi: 10.1093/nar/gkm882

Kidner, C. A. (2010). The many roles of small RNAs in leaf development. *J. Genet. Genomics* 37, 13–21. doi: 10.1016/S1673-8527(09)60021-7

Li, B., Qin, Y., Duan, H., Yin, W., and Xia, X. (2011). Genome-wide characterization of new and drought stress responsive microRNAs in Populus euphratica. *J. Exp. Bot.* 62, 3765–3779. doi: 10.1093/jxb/err051

Li, B. S., Duan, H., Li, J. G., Deng, X. W., Yin, W. L., and Xia, X. L. (2013). Global identification of miRNAs and targets in Populus euphratica under salt stress. *Plant Mol. Biol.* 81, 525–539. doi: 10.1007/s11103-013-0010-y

Li, B. S., Yin, W. L., and Xia, X. L. (2009). Identification of microRNAs and their targets from Populus euphratica. *Biochem. Biophys. Res. Commun.* 388, 272–277. doi: 10.1016/j.bbrc.2009.07.161

Li, M., Li, Y., Guo, L., Gong, N., Pang, Y., Jiang, W., et al. (2017a). Functional characterization of Tea (Camellia sinensis) MYB4a transcription factor using an integrative approach. *Front. Plant. Sci.* 8:943. doi: 10.3389/fpls.2017.00943

Li, S., Ying, Y., Secco, D., Wang, C., Narsai, R., Whelan, J., et al. (2017b). Molecular interaction between PHO2 and GIGANTEA reveals a new crosstalk between flowering time and phosphate homeostasis in Oryza sativa. *Plant Cell Environ.* 40, 1487–1499. doi: 10.1111/pce.12945

Liu, D., Song, Y., Chen, Z., and Yu, D. (2009). Ectopic expression of miR396 suppresses GRF target gene expression and alters leaf growth in Arabidopsis. *Physiol Plant.* 136, 223–236. doi: 10.1111/j.1399-3054.2009.01229.x

Liu, F., Wang, Y., Ding, Z., Zhao, L., Xiao, J., Wang, L., et al. (2017a). Transcriptomic analysis of flower development in tea (*Camellia sinensis* (L.)). *Gene* 631, 39–51. doi: 10.1016/j.gene.2017.08.013

Liu, H., Wang, Z. L., Tian, L. Q., Qin, Q. H., Wu, X. B., Yan, W. Y., et al. (2014). Transcriptome differences in the hypopharyngeal gland between Western Honeybees (Apis mellifera) and Eastern Honeybees (Apis cerana). *BMC Genomics* 15:744. doi: 10.1186/1471-2164-15-744

Liu, H., Yu, H., Tang, G., and Huang, T. (2018). Small but powerful: function of microRNAs in plant development. *Plant Cell Rep* 37, 515–528. doi: 10.1007/s00299-017-2246-5

Liu, Z., Zhang, Y., Ou, L., Kang, L., Liu, Y., Lv, J., et al. (2017b). Identification and characterization of novel microRNAs for fruit development and quality in hot pepper (*Capsicum annuum* L.). *Gene* 608, 66–72. doi: 10.1016/j.gene.2017.01.020

Lu, C., Kulkarni, K., Souret, F. F., Muthuvalliappan, R., Tej, S. S., Poethig, R. S., et al. (2006). MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res.* 16, 1276–1288. doi: 10.1101/gr.5530106

Luna, A., Babur, O., Aksoy, B. A., Demir, E., and Sander, C. (2016). PaxtoolsR: pathway analysis in R using pathway commons. *Bioinformatics* 32, 1262–1264. doi: 10.1093/bioinformatics/btv733

Micol, J. L., and Hake, S. (2003). The development of plant leaves. *Plant Physiol* 131, 389–394. doi: 10.1104/pp.015347

Muleo, A. G. M. C. M. M. E. F. M. S. R. (2012). "Epigenetic role in olive plant architecture," in *4th Next Generation Sequencing and Epigenomics Workshop, At Bari, 5-7 dicembre 2012, Volume: Proceedings of the 4th Next Generation Sequencing and Epigenomics Workshop* (Bari), 36–37.

Palatnik, J. F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J. C., et al. (2003). Control of leaf morphogenesis by microRNAs. *Nature* 425, 257–263. doi: 10.1038/nature01958

Palatnik, J. F., Wollmann, H., Schommer, C., Schwab, R., Boisbouvier, J., Rodriguez, R., et al. (2007). Sequence and expression differences underlie functional specialization of Arabidopsis microRNAs miR159 and miR319. *Dev. Cell* 13, 115–125. doi: 10.1016/j.devcel.2007.04.012

Praneenararat, T., Takagi, T., and Iwasaki, W. (2012). Integration of interactive, multi-scale network navigation approach with Cytoscape for functional genomics in the big data era. *BMC Genomics* 13 (Suppl 7):S24. doi: 10.1186/1471-2164-13-S7-S24

Qi, J., Wu, B., Feng, S., Lu, S., Guan, C., Zhang, X., et al. (2017). Mechanical regulation of organ asymmetry in leaves. *Nat. Plants* 3, 724–733. doi: 10.1038/s41477-017-0008-6

Ral, J. P., Colleoni, C., Wattebled, F., Dauvillee, D., Nempont, C., Deschamps, P., et al. (2006). Circadian clock regulation of starch metabolism establishes GBSSI as a major contributor to amylopectin synthesis in *Chlamydomonas reinhardtii*. *Plant Physiol.* 142, 305–317. doi: 10.1104/pp.106.081885

Ramachandran, S., Hiratsuka, K., and Chua, N. H. (1994). Transcription factors in plant growth and development. *Curr. Opin. Genet. Dev.* 4, 642–646. doi: 10.1016/0959-437X(94)90129-Q

Rubio-Somoza, I., and Weigel, D. (2011). MicroRNA networks and developmental plasticity in plants. *Trends Plant Sci.* 16, 258–264. doi: 10.1016/j.tplants.2011.03.001

Saik, O. V., Demenkov, P. S., Ivanisenko, T. V., Bragina, E. Y., Freidin, M. B., Goncharova, I. A., et al. (2018). Novel candidate genes important for asthma and hypertension comorbidity revealed from associative gene networks. *BMC Med. Genomics* 11:15. doi: 10.1186/s12920-018-0331-4

Savoi, S., Wong, D. C. J., Arapitsas, P., Miculan, M., Bucchetti, B., Peterlunger, E., et al. (2016). Transcriptome and metabolite profiling reveals that prolonged drought modulates the phenylpropanoid and terpenoid pathway in white grapes (*Vitis vinifera* L.). *BMC Plant Biol.* 16:67. doi: 10.1186/s12870-016-0760-1

Schommer, C., Palatnik, J. F., Aggarwal, P., Chetelat, A., Cubas, P., Farmer, E. E., et al. (2008). Control of jasmonate biosynthesis and senescence by miR319 targets. *PLoS Biol.* 6:e230. doi: 10.1371/journal.pbio.0060230

Shen, J., Wang, Y., Ding, Z., Ding, S., Wang, H., Bi, C., et al. (2019). Metabolic analyses reveal growth characteristics of young tea shoots in spring. *Sci. Hortic.* 246, 478–489. doi: 10.1016/j.scienta.2018.11.022

Shi, T., Wang, K., and Yang, P. (2017). The evolution of plant microRNAs: insights from a basal eudicot sacred lotus. *Plant J.* 89, 442–457. doi: 10.1111/tpj.13394

Sun, G. (2012). MicroRNAs and their diverse functions in plants. *Plant Mol Biol* 80, 17–36. doi: 10.1007/s11103-011-9817-6

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2017). The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

Takada, S., and Tasaka, M. (2002). Embryonic shoot apical meristem formation in higher plants. *J Plant Res.* 115, 411–417. doi: 10.1007/s10265-002-0061-7

Tang, J., and Chu, C. (2017). MicroRNAs in crop improvement: fine-tuners for complex traits. *Nat Plants* 3:17077. doi: 10.1038/nplants.2017.77

Tsuji, H., Aya, K., Ueguchi-Tanaka, M., Shimada, Y., Nakazono, M., Watanabe, R., et al. (2006). GAMYB controls different sets of genes and is differentially regulated by microRNA in aleurone cells and anthers. *Plant J.* 47, 427–444. doi: 10.1111/j.1365-313X.2006.02795.x

Usami, T., Horiguchi, G., Yano, S., and Tsukaya, H. (2009). The more and smaller cells mutants of Arabidopsis thaliana identify novel roles for SQUAMOSA PROMOTER BINDING PROTEIN-LIKE genes in the control of heteroblasty. *Development* 136, 955–964. doi: 10.1242/dev.028613

Voinnet, O. (2009). Origin, biogenesis, and activity of plant microRNAs. *Cell* 136, 669–687. doi: 10.1016/j.cell.2009.01.046

Wang, J. W., Schwab, R., Czech, B., Mica, E., and Weigel, D. (2008). Dual effects of miR156-targeted SPL genes and CYP78A5/KLUH on plastochron length and organ size in Arabidopsis thaliana. *Plant Cell* 20, 1231–1243. doi: 10.1105/tpc.108.058180

Wang, P., Zhang, L., Jiang, X., Dai, X., Xu, L., Li, T., et al. (2018). Evolutionary and functional characterization of leucoanthocyanidin reductases from Camellia sinensis. *Planta* 247, 139–154. doi: 10.1007/s00425-017-2771-z

Wei, C., Yang, H., Wang, S., Zhao, J., Liu, C., Gao, L., et al. (2018). Draft genome sequence of Camellia sinensis var. sinensis provides insights into the evolution of the tea genome and tea quality. *Proc. Natl. Acad. Sci. U.S.A.* 115, E4151–E4158. doi: 10.1073/pnas.1719622115

Wu, G., Park, M. Y., Conway, S. R., Wang, J. W., Weigel, D., and Poethig, R. S. (2009). The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis. *Cell* 138, 750–759. doi: 10.1016/j.cell.2009.06.031

Wu, J., Wang, D., Liu, Y., Wang, L., Qiao, X., and Zhang, S. (2014). Identification of miRNAs involved in pear fruit development and quality. *BMC Genomics* 15:953. doi: 10.1186/1471-2164-15-953

Wu, Y., Lv, W., Hu, L., Rao, W., Zeng, Y., Zhu, L., et al. (2017a). Identification and analysis of brown planthopper-responsive microRNAs in resistant and susceptible rice plants. *Sci. Rep.* 7:8712. doi: 10.1038/s41598-017-09143-y

Wu, Y., Yang, L., Yu, M., and Wang, J. (2017b). Identification and expression analysis of microRNAs during ovule development in rice (Oryza sativa) by deep sequencing. *Plant Cell Rep.* 36, 1815–1827. doi: 10.1007/s00299-017-2196-y

Xia, E. H., Zhang, H. B., Sheng, J., Li, K., Zhang, Q. J., Kim, C., et al. (2017). The tea tree genome provides insights into tea flavor and independent evolution of caffeine biosynthesis. *Mol. Plant* 10, 866–877. doi: 10.1016/j.molp.2017.04.002

Xu, M., Hu, T., Zhao, J., Park, M. Y., Earley, K. W., Wu, G., et al. (2016). Developmental Functions of miR156-Regulated Squamosa promoter binding protein-like (SPL) Genes in Arabidopsis thaliana. *PLoS Genet.* 12:e1006263. doi: 10.1371/journal.pgen.1006263

Yan, T., Yoo, D., Berardini, T. Z., Mueller, L. A., Weems, D. C., Weng, S., et al. (2005). PatMatch: a program for finding patterns in peptide and nucleotide sequences. *Nucleic Acids Res*. 33**,** W262–266. doi: 10.1093/nar/gki368

Yang, C., Li, D., Mao, D., Liu, X., Ji, C., Li, X., et al. (2013). Overexpression of microRNA319 impacts leaf morphogenesis and leads to enhanced cold tolerance in rice (Oryza sativa L.). *Plant Cell Environ*. 36**,** 2207–2218. doi: 10.1111/pce.12130

Yang, Y. L. (2015). *Chinese Clonal Tea Varieties*. Shanghai: Shanghai Science and Technology Press.

Zhang, L. Y., Bai, M. Y., Wu, J., Zhu, J. Y., Wang, H., Zhang, Z., et al. (2009). Antagonistic HLH/bHLH transcription factors mediate brassinosteroid regulation of cell elongation and plant development in rice and Arabidopsis. *Plant Cell* 21**,** 3767–3780. doi: 10.1105/tpc.109.070441

Zhao, L., Gao, L., Wang, H., Chen, X., Wang, Y., Yang, H., et al. (2013). The R2R3-MYB, bHLH, WD40, and related transcription factors in flavonoid biosynthesis. *Funct. Integr. Genomics* 13, 75–98. doi: 10.1007/s10142-012-0301-4

Zhao, L., Jiang, X. L., Qian, Y. M., Wang, P. Q., Xie, D. Y., Gao, L. P., et al. (2017). Metabolic characterization of the anthocyanidin reductase pathway involved in the biosynthesis of Flavan-3-ols in elite shuchazao tea (Camellia sinensis) cultivar in the field. *Molecules* 22:E2241. doi: 10.3390/molecules22122241

Zheng, C., Wang, Y., Ding, Z., and Zhao, L. (2016). Global transcriptional analysis reveals the complex relationship between tea quality, leaf senescence and the responses to cold-drought combined stress in camellia sinensis. *Front Plant Sci*. 7**:**1858. doi: 10.3389/fpls.2016.01858

Zheng, C., Zhao, L., Wang, Y., Shen, J., Zhang, Y., Jia, S., et al. (2015). Integrated RNA-Seq and sRNA-Seq analysis identifies chilling and freezing responsive key molecular players and pathways in tea plant (*Camellia sinensis*). *PLoS ONE* 10**:**e0125031. doi: 10.1371/journal.pone.0125031

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 May 2019
doi: 10.3389/fgene.2019.00440

The Genomic Landscape of Crossover Interference in the Desert Tree *Populus euphratica*

Ping Wang[1], *Libo Jiang*[1]*, *Meixia Ye*[1], *Xuli Zhu*[1] and *Rongling Wu*[1,2]

[1] Center for Computational Biology, College of Biological Sciences and Biotechnology, Beijing Forestry University, Beijing, China, [2] Center for Statistical Genetics, The Pennsylvania State University, Hershey, PA, United States

Crossover (CO) interference is a universal phenomenon by which the occurrence of one CO event inhibits the simultaneous occurrence of other COs along a chromosome. Because of its critical role in the evolution of genome structure and organization, the cytological and molecular mechanisms underlying CO interference have been extensively investigated. However, the genome-wide distribution of CO interference and its interplay with sex-, stress-, and age-induced differentiation remain poorly understood. Multi-point linkage analysis has proven to be a powerful tool for landscaping CO interference, especially within species for which CO mutants are rarely available. We implemented four-point linkage analysis to landscape a detailed picture of how CO interference is distributed through the entire genome of *Populus euphratica*, the only forest tree that can survive and grow in saline desert. We identified an extensive occurrence of CO interference, and found that its strength depends on the length of chromosomes and the genomic locations within the chromosome. We detected high-order CO interference, possibly suggesting a highly complex mechanism crucial for *P. euphratica* to grow, reproduce, and evolve in its harsh environment.

Keywords: euphrates poplar, genetic interference, mapping population, meiotic crossover, four-point analysis

OPEN ACCESS

Edited by:
Ancha Baranova, George Mason University, United States

Reviewed by:
Xiyin Wang, North China University of Science and Technology, China
Longjiang Fan, Zhejiang University, China

***Correspondence:**
Libo Jiang
libojiang@bjfu.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *11 October 2018*
Accepted: *29 April 2019*
Published: *15 May 2019*

Citation:
Wang P, Jiang L, Ye M, Zhu X and Wu R (2019) The Genomic Landscape of Crossover Interference in the Desert Tree Populus euphratica. Front. Genet. 10:440. doi: 10.3389/fgene.2019.00440

INTRODUCTION

Crossovers (COs) are recombination events involving a reciprocal exchange of genetic material. During meiotic prophase, COs are essential for the accurate segregation of homologous chromosomes (Hillers, 2004). In most organisms, the abundance and distribution of COs is highly regulated by universal mechanisms, referred to as CO interference or genetic interference. The fact that the presence of a CO interferes with the occurrence of other COs within the same chromosome has been confirmed. Due to such interferences, chiasmata are more evenly placed along chromosomes than previously expected (Hillers, 2004; Hultén, 2011). Moreover, CO interference is ubiquitous in eukaryotes and plays a crucial role in their evolution. However, our understanding of CO interference mechanisms and their distribution in biota remains very limited.

Sturtevant and Muller constructed a Drosophila genetic map and found that COs were more evenly spaced than would be expected from random placement (Lam et al., 2005). CO interference is widespread in most eukaryotes and can confer selectivity advantages. The extent of CO interference decreases with genetic distance between COs; however, given the same distance, it is stronger on the same chromosomal arm than on different arms (Berchowitz and Copenhaver, 2010). The variability of CO interference within a specific chromosome region is affected by the overall size and structure of the chromosome (Hillers, 2004), and CO interferences are regulated by

the anti-recombinase RTEL-1 protein in *Caenorhabditis elegant* (Youds et al., 2010). A reduction in CO interference can result from a lack of DNA-damage-response-kinase Tel1/ATM (Anderson et al., 2015). Links between CO interferences and sex differences (Jan et al., 2007; Szatkiewicz et al., 2013), stress-induced adaptation (Yant et al., 2013; Aggarwal et al., 2015), and aging (Campbell et al., 2015; Wang Z. et al., 2016) have been discovered, highlighting the multifaceted role of COs in mediating biological processes. As an evolutionary phenotype, CO interference varies with biotic and abiotic environmental parameters, such as sex, age, and stress. For example, in mice and cattle, interference is stronger in females than in males (Szatkiewicz et al., 2013; Wang Z. et al., 2016). However, the opposite is found in humans, where interference is stronger in males than in females, although this pattern varies by chromosome (Campbell et al., 2015).

Many methods have been used to study the mechanisms of CO interference, including the count-location model, the gamma model, and multi-point linkage analysis. Initially, CO interference was genetically defined and characterized by cytology, the location of protein complexes, and chromosomal CO events. Recent studies have explored the mechanistic basis of CO interference using cytogenetics and molecular methods, whereas more traditional interference studies use the coefficient of coincidence (CoC) between two disjoint intervals on a genetic map. The CoC is defined as the ratio of the observed frequency to the expected frequency, and represents all possible intervals of gametes with double CO for each pair (Waterworth, 2000). Traditional models of interference suggest that the occurrence of a CO produces signals or substances that prevent additional CO events and then spreads along the chromosome at a similar distance on both sides (Housworth and Stahl, 2003). The polymerization model states that early recombination events are distributed independently with each other and then have the same chance of initiating bidirectional aggregation events per unit of time (King and Mortimer, 1990).

More recently, many model and non-model systems have been developed to characterize the phenomenon of CO interference. CO interference has been investigated mainly by tracking DNA markers on a single chromosome of parents during a specific period under electron fluorescence microscopy. The gamma model has recently received attention and suggests that the shape parameter of the gamma distribution is an indicator for uniformity and an indirect indicator for interference (Lam et al., 2005). The mechanical stress model assumes that each CO event releases a specific distance of pressure along the chromosome to prevent the presence of nearby COs (Wang et al., 2015). At present, multi-point linkage analysis has been proven to be more advantageous in genetic distance estimation and gene ordering, and it is equipped with a strong ability to discern and quantify CO interferences.

Despite numerous theoretical and empirical studies, our understanding of how interference is distributed across genomes remains unclear (Housworth and Stahl, 2003). This can be attributed to a number of reasons. First, traditional genetic screens for mutations affecting interference require numerous meiotic progenies to include meiotic COs in multiple intervals along a chromosome (Berchowitz and Copenhaver, 2010). Second, most of the mutations that modify interference affect chromosomal proteins, which not only mediate interference but also play a role in CO formation (Joshi et al., 2009). Thus, genetic strategies that abolish mutation interference also reduce or eliminate CO events. Third, many mutants differ in their frequency of occurrence of CO in different loci and environments (Getz et al., 2008). Therefore, combining multi-point analysis and cytology tools, which are used widely for locating and sequencing genes, can increase the ability to detect interference (Broman and Weber, 2000). The multi-analytic statistical model, which is based on the linkage analysis method of genetic maps, can describe CO interference that take place not only between two adjacent chromosome intervals, but also in multiple consecutive intervals. Additionally, multi-point analysis provides a quantitative method to estimate CO interference (Zickler and Kleckner, 2016). In particular, by assessing the chromosomal distribution of CO interference, multi-point analysis can activate the use of linkage mapping as a routine genetic tool to investigate further dimensions of genomic structure and organization (Lu et al., 2004).

Populus euphratica is the only arbor species in arid-semiarid regions and plays an important role in maintaining the ecological balance in desert regions. The goals of this study were to identify the distribution of CO interference in *P. euphratica* at a whole-genome scale using multi-point analysis based on the full-sib family of *P. euphratica* and to study the relationship between the overall CO interference strength and length of the chromosome, as well as the region of the chromosome. Due to the impact of climate change and anthropogenic activities, the area of *P. euphratica* in northwest China has declined sharply and its ecological security and agricultural production are facing severe challenges (Qiu et al., 2011). By using four-point linkage analysis to analyze the CO interference of *P. euphratica*, we can describe its distribution within the genome in detail, which will provide a theoretical basis for the follow-up forest genetic research and molecular marker-assisted breeding. It is of great significance to understand the genetic diversity and evolutionary history of *P. euphratica* and to find their core germplasm resources.

MATERIALS AND METHODS

Plant Material and Genetic Linkage Map

One male and one female *P. euphratica* individual were randomly selected along the Tarim River in the Korla region of Xinjiang, China. The individuals were located 31 km from one another, ensuring a large genetic difference between them. Male and female flowering branches from the individuals were planted in an artificial climate chamber at Beijing Forestry University. After cultivation was completed, a series of experimental treatments, including dehydration, thinning, and freezing with liquid nitrogen, were performed on the selected materials. Finally, the F_1 progeny of 408 individuals were obtained. DNA was extracted using the TIANGEN plant genomic DNA extraction kit (Beijing, China). The quality of all samples was assessed and RAD technology was used for high-throughput DNA sequencing

(Conesa et al., 2005). The genetic map of *P. euphratica* was constructed from the resultant sequence data.

Multi-Point Linkage Analysis

A four-point analysis was developed so that four consecutive markers could be analyzed simultaneously (Wang J. et al., 2016). It beyond three-point analysis, can characterize crossover interference that takes place not only between two adjacent chromosomal intervals, but also over multiple successive intervals (We call the interference occurred in multiple marker intervals of more than three markers as high dimensional CO interference). We used the CoC to describe the ratio of the observed number of double recombinants to this expected number. As we have known, the recombination events occurring between different marker intervals are not independent. Thus, the extent to which this coefficient corresponds to the strength of CO interference.

In the full-sib family of *P. euphratica*, two heterozygous F_1 individuals, ABCD/abcd and ABCD/abcd, were crossed to produce a segregated F_2 population. Each F_1 parent produced 16 gametes, divided into eight types (**Table 1**). The frequencies of the gamete types are represented by g000,..., g111, where the subscripts represent the number of COs between a particular pair of tags. Based on the genetic map of *P. euphratica*, we grouped single-nucleotide polymorphism markers on 19 linkage groups with four markers in every group. The genotype frequencies of the gamete types were calculated by counting the number of genotypes within the 408 individuals of each group. The four consecutive markers (i.e., A-B-C-D) had six possible recombination moieties. From these gamete-type frequencies, we expressed the recombination fractions of each marker pair, denoted by r_{AB}, r_{BC}, r_{CD}, r_{AC}, r_{BD}, and r_{AD}, as follows:

$$\begin{aligned} r_{AB} &= g_{111} + g_{110} + g_{101} + g_{100} \\ r_{BC} &= g_{111} + g_{110} + g_{011} + g_{010} \\ r_{CD} &= g_{111} + g_{101} + g_{011} + g_{001} \\ r_{AC} &= g_{101} + g_{100} + g_{011} + g_{010} \\ r_{BD} &= g_{110} + g_{010} + g_{101} + g_{001} \\ r_{AD} &= g_{111} + g_{010} + g_{100} + g_{001} \end{aligned} \tag{1}$$

Denote the coefficients of coincidence (a measure of crossover interference) between double marker intervals A-B and B-C, double marker intervals B-C and C-D, double marker intervals A-B and C-D, and triple marker intervals A-B, B-C, and C-D by C_1, C_2, C_3, and C_4, respectively (Sun et al., 2017). Wang J. et al. (2016) formulated the relationship between different recombination fractions based on the CoC and derived a process to estimate and test each coefficient, as follows:

$$\begin{aligned} C_4 &= \frac{g_{111}}{r_{AB}r_{BC}r_{CD}} \\ C_1 &= \frac{g_{111} + g_{110}}{r_{AB}r_{BC}} \\ C_2 &= \frac{g_{111} + g_{011}}{r_{BC}r_{CD}} \\ C_3 &= \frac{g_{111} + g_{101}}{r_{AB}r_{CD}} \end{aligned} \tag{2}$$

providing a method to characterize the genomic distribution of CO interference along the chromosome.

For an F_2 offspring family of *P. euphratica*, two F_1 progenies crossed to produce 136 diploids, divided into 81 identifiable genotypes. This situation differs from the backcross population, which is more complex and requires the Expectation Maximization algorithm to be implemented (Dempster et al., 1977). **Table 2** provides the frequencies of these 81 genotypes, as well as the corresponding numbers. The frequencies of heterozygous genotypes are a mix of products of gamete-type frequencies (Wang J. et al., 2016). Subsequently, the *P. euphratica* data were analyzed by multi-point analysis to obtain the CoC values representing the CO interference strength. If the CoC value is 0, it indicates that interference is absent.

The Relationship Between Overall High Dimensional CO Interference Strength and Chromosome Length

Differences in CO interference strength are affected by the overall size of the chromosome (Albini, 2010). Through four-point linkage analysis, we obtained the recombination rate between four marker intervals on each linkage group and the corresponding CoC. To study the relationship between chromosome length and overall high-order CO interference strength, we assumed that the length of the linkage group on the genetic map was the length of the chromosome. Next, the distribution interval of high dimensional CO interference strength on the 19 chromosomes was characterized by a boxplot displaying the maximum, minimum, median, and upper and lower quartiles of the data. Due to different structural characteristics of chromosomes, there are many factors affecting the strength of CO interference; therefore, the mean of the CO interference strength on each chromosome was calculated

TABLE 1 | Gamete types and their frequencies at four ordered markers, A-B-C-D.

No.		1	2	3	4	5	6	7	8
Gamete type		**ABCD/abcd**	**ABCd/abcD**	**Abcd/abCD**	**ABcD/abCd**	**Abcd/aBCD**	**AbcD/aBCd**	**AbCD/aBcd**	**AbCd/aBcD**
Number of crossovers	A-B	0	0	0	0	1	1	1	1
	B-C	0	0	1	1	0	0	1	1
	C-D	0	1	0	1	0	1	0	1
Gamete type frequency		g000	g001	g010	g011	g100	g101	g110	g111

TABLE 2 | Four-marker genotype observations and expected frequencies composed of gamete-type frequencies produced by each parent in a full-sib population.

Genotype	g000	g001	g010	g011	g100	g101	g110	g111	Frequency	Observation
AABBCCDD	2	0	0	0	0	0	0	0	g_{000}^2	n_{2222}
AABBCCDd	1	1	0	0	0	0	0	0	$2g_{000}g_{001}$	n_{2221}
AABBCCdd	0	2	0	0	0	0	0	0	g_{001}^2	n_{2220}
AABBCcDD	1	0	0	1	0	0	0	0	$2g_{000}g_{011}$	n_{2212}
AABBCcDd	ϕ_3	$1-\phi_3$	ϕ_3	$1-\phi_3$	0	0	0	0	$2(g_{000}g_{010}+g_{001}g_{011})$	n_{2211}
AABBCcdd	0	1	1	0	0	0	0	0	$2g_{001}g_{010}$	n_{2210}
AABBccDD	0	0	0	2	0	0	0	0	g_{011}^2	n_{2202}
AABBccDd	0	0	1	1	0	0	0	0	$2g_{011}g_{010}$	n_{2201}
AABBccdd	0	0	2	0	0	0	0	0	g_{010}^2	n_{2200}
AABbCCDD	1	0	0	0	0	0	1	0	$2g_{000}g_{110}$	n_{2122}
AABbCCDd	ϕ_{10}	$1-\phi_{10}$	0	0	0	0	$1-\phi_{10}$	ϕ_{10}	$2(g_{000}g_{111}+g_{001}g_{110})$	n_{2121}
AABbCCdd	0	1	0	0	0	0	0	1	$2g_{001}g_{111}$	n_{2120}
AABbCcDD	ϕ_7	0	0	$1-\phi_7$	0	ϕ_7	$1-\phi_7$	0	$2(g_{000}g_{101}+g_{011}g_{110})$	n_{2112}
AABbCcDd	ϕ_6	ϕ_{16}	ϕ_{23}	ϕ_{27}	ϕ_6	ϕ_{16}	ϕ_{23}	ϕ_{27}	$2(g_{000}g_{100}+g_{001}g_{101}+g_{010}g_{110}+g_{011}g_{111})$	n_{2111}
AABbCcdd	0	ϕ_{15}	$1-\phi_{15}$	0	ϕ_{15}	0	0	$1-\phi_{15}$	$2(g_{001}g_{100}+g_{010}g_{111})$	n_{2110}
AABbccDD	0	0	0	1	0	1	0	0	$2g_{011}g_{101}$	n_{2102}
AABbccDd	0	0	ϕ_{22}	$1-\phi_{22}$	$1-\phi_{22}$	ϕ_{22}	0	0	$2(g_{010}g_{101}+g_{011}g_{100})$	n_{2101}
AABbccdd	0	0	1	0	1	0	0	0	$2g_{010}g_{100}$	n_{2100}
AAbbCCDD	0	0	0	0	0	0	2	0	g_{110}^2	n_{2022}
AAbbCCDd	0	0	0	0	0	0	1	1	$2g_{110}g_{111}$	n_{2021}
AAbbCCdd	0	0	0	0	0	0	0	2	g_{111}^2	n_{2020}
AAbbCcDD	0	0	0	0	0	1	1	0	$2g_{011}g_{101}$	n_{2012}
AAbbCcDd	0	0	0	0	ϕ_{30}	$1-\phi_{30}$	ϕ_{30}	$1-\phi_{30}$	$2(g_{110}g_{100}+g_{111}g_{101})$	n_{2011}
AAbbCcdd	0	0	0	0	1	0	0	1	$2g_{111}g_{100}$	n_{2010}
AAbbccDD	0	0	0	0	0	2	0	0	g_{101}^2	n_{2002}
AAbbccDd	0	0	0	0	1	1	0	0	$2g_{101}g_{100}$	n_{2001}
AAbbccdd	0	0	0	0	2	0	0	0	g_{100}^2	n_{2000}
AaBBCCDD	1	0	0	0	1	0	0	0	$2g_{000}g_{100}$	n_{1222}
AaBBCCDd	ϕ_8	$1-\phi_8$	0	$1-\phi_8$	0	ϕ_8	0	0	$2(g_{000}g_{101}+g_{001}g_{011})$	n_{1221}
AaBBCCdd	0	1	0	0	0	1	0	0	$2g_{001}g_{101}$	n_{1220}
AaBBCcDD	ϕ_{11}	0	0	$1-\phi_{11}$	$1-\phi_{11}$	0	0	ϕ_{11}	$2(g_{000}g_{111}+g_{011}g_{100})$	n_{1212}
AaBBCcDd	ϕ_9	ϕ_{18}	ϕ_{21}	ϕ_{26}	ϕ_{21}	ϕ_{26}	ϕ_9	ϕ_{18}	$2(g_{000}g_{110}+g_{001}g_{111}+g_{010}g_{100}+g_{011}g_{101})$	n_{1211}
AaBBCcdd	0	ϕ_{17}	$1-\phi_{17}$	0	0	$1-\phi_{17}$	ϕ_{17}	0	$2(g_{001}g_{110}+g_{010}g_{101})$	n_{1210}
AaBBccDD	0	0	0	1	0	0	0	1	$2(g_{010}g_{111}+g_{011}g_{110})$	n_{1202}
AaBBccDd	0	0	ϕ_{24}	$1-\phi_{24}$	0	0	$1-\phi_{24}$	ϕ_{24}	$2(g_{010}g_{111}+g_{011}g_{110})$	n_{1201}
AaBBccdd	0	0	1	0	0	0	1	0	$2g_{010}g_{110}$	n_{1200}
AaBbCCDD	ϕ_4	0	ϕ_4	0	$1-\phi_4$	0	$1-\phi_4$	0	$2(g_{000}g_{010}+g_{110}g_{100})$	n_{1122}
AaBbCCDd	ϕ_5	ϕ_{13}	ϕ_{13}	ϕ_5	ϕ_{31}	ϕ_{33}	ϕ_{33}	ϕ_{31}	$2(g_{000}g_{011}+g_{001}g_{010}+g_{110}g_{101}+g_{111}g_{100})$	n_{1121}
AaBbCCdd	0	ϕ_{14}	0	ϕ_{14}	0	$1-\phi_{14}$	0	$1-\phi_{14}$	$2(g_{001}g_{011}+g_{111}g_{101})$	n_{1120}
AaBbCcDD	ϕ_2	ϕ_2	ϕ_{20}	ϕ_{20}	ϕ_{29}	ϕ_{29}	ϕ_{35}	ϕ_{35}	$2(g_{000}g_{001}+g_{011}g_{010}+g_{110}g_{111}+g_{101}g_{100})$	n_{1112}
AaBbCcDd	$2\phi_1$	$2\phi_{12}$	$2\phi_{19}$	$2\phi_{25}$	$2\phi_{28}$	$2\phi_{32}$	$2\phi_{34}$	$2\phi_{36}$	$g_{000}^2+g_{001}^2+g_{010}^2+g_{100}^2+g_{101}^2+g_{110}^2+g_{011}^2+g_{111}^2)$	n_{1111}
AaBbCcdd	ϕ_2	ϕ_2	ϕ_{20}	ϕ_{20}	ϕ_{29}	ϕ_{29}	ϕ_{35}	ϕ_{35}	$2(g_{000}g_{001}+g_{011}g_{010}+g_{110}g_{111}+g_{101}g_{100})$	n_{1110}
AaBbccDD	0	ϕ_{14}	0	ϕ_{14}	0	$1-\phi_{14}$	0	$1-\phi_{14}$	$2(g_{001}g_{011}+g_{111}g_{101})$	n_{1102}
AaBbccDd	ϕ_5	ϕ_{13}	ϕ_{13}	ϕ_5	ϕ_{31}	ϕ_{33}	ϕ_{33}	ϕ_{31}	$2(g_{000}g_{011}+g_{001}g_{010}+g_{110}g_{101}+g_{111}g_{100})$	n_{1101}
AaBbccdd	ϕ_4	0	ϕ_4	0	$1-\phi_4$	0	$1-\phi_4$	0	$2(g_{000}g_{010}+g_{110}g_{100})$	n_{1100}
AabbCCDD	0	0	1	0	0	0	1	0	$2g_{010}g_{110}$	n_{1022}
AabbCCDd	0	0	ϕ_{24}	$1-\phi_{24}$	0	0	$1-\phi_{24}$	ϕ_{24}	$2(g_{010}g_{111}+g_{011}g_{110})$	n_{1021}
AabbCCdd	0	0	0	1	0	0	0	1	$2g_{011}g_{111}$	n_{1020}
AabbCcDD	0	ϕ_{17}	$1-\phi_{17}$	0	0	$1-\phi_{17}$	ϕ_{17}	0	$2g_{011}g_{111}$	n_{1012}
AabbCcDd	ϕ_9	ϕ_{18}	ϕ_{21}	ϕ_{26}	ϕ_{21}	ϕ_{26}	ϕ_9	ϕ_{18}	$2(g_{000}g_{110}+g_{001}g_{111}+g_{010}g_{100}+g_{011}g_{101})$	n_{1011}

(Continued)

TABLE 2 | Continued

Genotype	g000	g001	g010	g011	g100	g101	g110	g111	Frequency	Observation
AabbCcdd	ϕ_{11}	0	0	1-ϕ_{11}	1-ϕ_{11}	0	0	ϕ_{11}	$2(g_{000}g_{111}+g_{011}g_{100})$	n_{1010}
AabbccDD	0	1	0	0	0	1	0	0	$2g_{001}g_{101}$	n_{1002}
AabbccDd	ϕ_8	1-ϕ_8	0	1-ϕ_8	0	ϕ_8	0	0	$2(g_{000}g_{101}+g_{001}g_{011})$	n_{1001}
Aabbccdd	1	0	0	0	1	0	0	0	$2g_{000}g_{100}$	n_{1000}
aaBBCCDD	0	0	0	0	2	0	0	0	$g_{100}{}^2$	n_{0222}
aaBBCCDd	0	0	0	0	1	1	0	0	$2g_{101}g_{100}$	n_{0221}
aaBBCCdd	0	0	0	0	0	2	0	0	$g_{101}{}^2$	n_{0220}
aaBBCcDD	0	0	0	0	1	0	0	1	$2g_{111}g_{100}$	n_{0212}
aaBBCcDd	0	0	0	0	ϕ_{30}	1-ϕ_{30}	ϕ_{30}	1-ϕ_{30}	$2(g_{110}g_{100}+g_{111}g_{101})$	n_{0211}
aaBBCcdd	0	0	0	0	0	1	1	0	$2g_{011}g_{101}$	n_{0210}
aaBBccDD	0	0	0	0	0	0	0	2	$g_{111}{}^2$	n_{0202}
aaBBccDd	0	0	0	0	0	0	1	1	$2g_{110}g_{111}$	n_{0201}
aaBBccdd	0	0	0	0	0	0	2	0	$g_{110}{}^2$	n_{0200}
aaBbCCDD	0	0	1	0	1	0	0	0	$2g_{010}g_{100}$	n_{0122}
aaBbCCDd	0	0	ϕ_{22}	1-ϕ_{22}	1-ϕ_{22}	ϕ_{22}	0	0	$2(g_{010}g_{101}+g_{011}g_{100})$	n_{0121}
aaBbCCdd	0	0	0	1	0	1	0	0	$2g_{011}g_{101}$	n_{0120}
aaBbCcDD	0	ϕ_{15}	1-ϕ_{15}	0	ϕ_{15}	0	0	1-ϕ_{15}	$2(g_{001}g_{100}+g_{010}g_{111})$	n_{0122}
aaBbCcDd	ϕ_6	ϕ_{16}	ϕ_{23}	ϕ_{27}	ϕ_6	ϕ_{16}	ϕ_{23}	ϕ_{27}	$2(g_{000}g_{100}+g_{001}g_{101}+g_{010}g_{110}+g_{011}g_{111})$	n_{0111}
aaBbCcdd	ϕ_7	0	0	1-ϕ_7	0	ϕ_7	1-ϕ_7	0	$2(g_{000}g_{101}+g_{011}g_{110})$	n_{0110}
aaBbccDD	0	1	0	0	0	0	0	1	$2g_{001}g_{111}$	n_{0102}
aaBbccDd	ϕ_{10}	1-ϕ_{10}	0	0	0	0	1-ϕ_{10}	ϕ_{10}	$2(g_{000}g_{111}+g_{001}g_{110})$	n_{0101}
aaBbccdd	1	0	0	0	0	0	1	0	$2g_{000}g_{110}$	n_{0100}
aabbCCDD	0	0	2	0	0	0	0	0	$g_{010}{}^2$	n_{0022}
aabbCCDd	0	0	1	1	0	0	0	0	$2g_{011}g_{010}$	n_{0021}
aabbCCdd	0	0	0	2	0	0	0	0	$g_{011}{}^2$	n_{0020}
aabbCcDD	0	1	1	0	0	0	0	0	$2g_{001}g_{010}$	n_{0012}
aabbCcDd	ϕ_3	1-ϕ_3	ϕ_3	1-ϕ_3	0	0	0	0	$2(g_{000}g_{010}+g_{001}g_{011})$	n_{0011}
aabbCcdd	1	0	0	1	0	0	0	0	$2g_{000}g_{011}$	n_{0010}
aabbccDD	0	2	0	0	0	0	0	0	$g_{001}{}^2$	n_{0002}
aabbccDd	1	1	0	0	0	0	0	0	$2g_{000}g_{001}$	n_{0001}
aabbccdd	2	0	0	0	0	0	0	0	$g_{000}{}^2$	n_{0000}

ϕ refers to the ratio of the frequency of each gamete genotype to the corresponding genotype frequency.

to account for the relationship between chromosome size and overall CO interference strength. Due to the distribution of chromosome 1 deviates more from the distribution of other chromosomes, it was determined to be an outlier and was removed from the dataset. Subsequently, chromosomes 2, 3, 4, and 6 were fitted with a linear model (blue line), and the remaining chromosomes were fitted with a trend line (red line). Through the fitting curves, the distribution of the overall high dimensional CO interference strength on different chromosomes was observed.

Ratio Variance in High Dimensional CO Interference Strength Between Different Chromosome Regions

CO rates are closely related to chromosome region (Giraut et al., 2011), allowing for differences in CO interference strength in different regions to be explored. In this study, each chromosome was divided into three parts according to genetic distance uniformity, and the three sections were labeled NO.1, NO.2, and NO.3, respectively. The CO interference strength of each was subtracted separately. NO.1-NO.2, NO.2-NO.3, and NO.1-NO.3 indicate the difference ratio (sum of the difference value of each corresponding CO interference strength between intervals) of CO interference strength in the first (NO.1) and second (NO.2) parts, the second part and the third (NO.3) part, the first and third part, respectively. This allowed for differences in the distribution of CO interference strength between the regions of the chromosome to be seen.

To display the impact of the three regions (NO.1, NO.2, and NO.3) in the chromosome on the CO interference strength distribution, we employed δ to quantitatively evaluate the difference of the CO interference strength distribution in different sections of chromosome, which can be calculated by

$$\delta = \sum_{i=1}^{N} \left| p_i^1 - p_i^2 \right| \tag{3}$$

where N is the total number of intervals of the CO interference strength value, p_i^1 and p_i^2 represent the percentage of the ith interval in two different chromosome regions, respectively. We further derived the range of δ:

$$0 \leq \delta = \sum_{i=1}^{N} \left| p_i^1 - p_i^2 \right| \leq \sum_{i=1}^{N} p_i^1 + \sum_{i=1}^{N} p_i^2 = 2 \quad (4)$$

When the CO interference strength distributions in both regions 1 and 2 were the same, δ was equal to 0, whereas δ reached the maximum of 2 when there was no overlapping region between the CO interference strength distributions of two regions. In all other cases, δ is larger than 0 and smaller than 2. δ reflects the difference of two different CO interference strength distributions.

RESULTS

In this study, we first used a four-point linkage analysis model to quantitatively analyze the CO interference on a full-sib population of *P. euphratica*. The genetic map contained 8,305 markers on 19 linkage groups. The total genetic distance was 4574.89 cM for the entire genetic map, among which the shortest linkage group was linkage group 19 (LG19) with a genetic distance of 130.26 cM and the longest linkage group was LG1 with

TABLE 3 | Recombination rates corresponding to the first two groups in each linkage group and the coefficient of coincidence (CoC) of crossover interference strength.

r_{AB}	r_{BC}	r_{CD}	r_{AC}	r_{BD}	r_{AD}	C1	C2	C3	C4	lg	Marker 1	Marker 2	Marker 3	Marker 4
0.50	0.02	0.50	0.50	0.50	0.03	1.00	1.00	1.91	0.00	1	nn_np_92921	lm_ll_11315	lm_ll_4392	nn_np_8171
0.02	0.03	0.50	0.04	0.50	0.50	13.46	1.00	1.00	13.46	1	lm_ll_12452	lm_ll_7926	lm_ll_9277	nn_np_8634
0.02	0.04	0.02	0.04	0.04	0.04	11.53	15.37	7.69	177.29	2	lm_ll_10351	lm_ll_4075	lm_ll_11480	lm_ll_9220
0.02	0.03	0.01	0.02	0.03	0.02	21.12	20.52	11.15	98.29	2	lm_ll_9058	hk_hk_3051	hk_hk_1972	hk_hk_2298
0.04	0.03	0.03	0.05	0.02	0.04	9.03	19.87	0.00	0.00	3	lm_ll_11125	lm_ll_8568	lm_ll_8093	lm_ll_4728
0.05	0.02	0.03	0.05	0.03	0.04	10.94	16.30	8.30	71.58	3	lm_ll_11556	hk_hk_2534	hk_hk_3271	hk_hk_1110
0.01	0.01	0.05	0.01	0.05	0.05	54.89	6.12	3.50	0.01	4	hk_hk_2770	hk_hk_3044	hk_hk_1608	nn_np_5712
0.50	0.50	0.50	0.01	0.02	0.50	1.97	1.96	1.93	3.86	4	lm_ll_10753	nn_np_12108	lm_ll_8376	nn_np_12562
0.03	0.04	0.09	0.03	0.09	0.08	17.70	5.68	0.52	14.20	5	hk_hk_3094	hk_hk_2811	hk_hk_616	hk_hk_2812
0.02	0.02	0.04	0.03	0.05	0.06	11.40	7.24	3.95	96.10	5	hk_hk_2689	hk_hk_3110	hk_hk_3313	hk_hk_1097
0.01	0.01	0.01	0.02	0.01	0.01	21.62	48.28	0.00	0.00	6	hk_hk_1495	hk_hk_3053	hk_hk_2055	hk_hk_2954
0.01	0.01	0.02	0.01	0.01	0.02	44.45	37.22	9.80	873.87	6	hk_hk_2493	hk_hk_3029	hk_hk_3009	hk_hk_1331
0.05	0.02	0.06	0.05	0.06	0.01	9.32	8.49	12.63	0.00	7	hk_hk_2701	hk_hk_2144	hk_hk_2510	hk_hk_1471
0.03	0.08	0.02	0.08	0.09	0.10	6.61	2.55	4.50	57.32	7	hk_hk_2148	hk_hk_1178	hk_hk_2076	hk_hk_2322
0.03	0.03	0.03	0.03	0.02	0.02	15.60	23.40	0.00	0.00	8	lm_ll_12636	lm_ll_7930	lm_ll_10007	lm_ll_9837
0.02	0.02	0.02	0.01	0.03	0.02	46.77	6.82	0.00	0.01	8	lm_ll_3885	lm_ll_11525	lm_ll_9742	lm_ll_12405
0.02	0.01	0.02	0.02	0.02	0.03	31.41	20.19	0.00	0.02	9	nn_np_8717	nn_np_10414	nn_np_12617	nn_np_9786
0.02	0.04	0.01	0.03	0.05	0.03	20.74	2.42	22.34	0.00	9	hk_hk_3317	hk_hk_1551	hk_hk_2994	nn_np_11331
0.04	0.50	0.01	0.50	0.50	0.50	1.00	1.00	0.00	0.00	10	lm_ll_7856	lm_ll_5397	nn_np_8256	nn_np_7630
0.50	0.01	0.50	0.50	0.50	0.02	1.00	1.00	1.92	0.00	10	nn_np_11692	lm_ll_12780	lm_ll_6166	nn_np_5449
0.50	0.02	0.50	0.50	0.50	0.04	1.00	1.00	1.91	0.77	11	hk_hk_2928	hk_hk_3329	hk_hk_2833	hk_hk_2138
0.49	0.04	0.07	0.51	0.11	0.46	0.64	0.67	1.68	0.00	11	hk_hk_871	hk_hk_1707	hk_hk_1946	nn_np_7373
0.02	0.09	0.03	0.08	0.11	0.11	10.25	3.16	11.69	123.68	12	hk_hk_2068	hk_hk_2300	hk_hk_2504	hk_hk_682
0.12	0.01	0.01	0.11	0.01	0.11	6.08	32.58	2.77	0.00	12	hk_hk_2654	hk_hk_1675	hk_hk_2586	hk_hk_253
0.03	0.03	0.50	0.04	0.50	0.50	11.43	1.00	1.00	11.43	13	nn_np_5159	nn_np_5866	nn_np_4850	lm_ll_1263
0.50	0.50	0.50	0.01	0.01	0.50	1.98	1.98	1.96	3.93	13	nn_np_11032	lm_ll_9285	nn_np_10951	lm_ll_12427
0.03	0.06	0.02	0.07	0.05	0.06	5.50	13.21	0.00	0.00	14	lm_ll_4169	lm_ll_8617	lm_ll_11626	lm_ll_10819
0.06	0.01	0.07	0.07	0.07	0.13	0.00	7.58	0.00	0.00	14	lm_ll_12378	lm_ll_12011	lm_ll_6388	lm_ll_9230
0.50	0.03	0.04	0.50	0.02	0.50	1.00	20.20	1.00	20.20	15	nn_np_12714	lm_ll_5117	lm_ll_9900	lm_ll_10761
0.01	0.03	0.02	0.02	0.02	0.01	28.20	24.42	6.55	0.03	15	nn_np_10876	nn_np_6544	nn_np_10233	hk_hk_1031
0.04	0.03	0.50	0.03	0.50	0.50	16.13	1.00	1.00	16.13	16	lm_ll_7848	lm_ll_12916	lm_ll_12998	nn_np_10507
0.50	0.00	0.03	0.50	0.04	0.50	1.00	0.00	1.00	0.00	16	lm_ll_9352	nn_np_3807	nn_np_8544	nn_np_6630
0.50	0.50	0.01	0.01	0.50	0.02	1.98	1.00	1.00	1.60	17	nn_np_11597	lm_ll_8106	nn_np_7356	nn_np_10281
0.04	0.02	0.04	0.06	0.04	0.05	3.91	14.56	4.25	0.00	17	hk_hk_2057	lm_ll_12267	lm_ll_8687	lm_ll_2545
0.03	0.02	0.04	0.04	0.03	0.03	9.25	16.94	11.26	38.84	18	hk_hk_3194	hk_hk_2315	hk_hk_2540	hk_hk_2773
0.03	0.02	0.01	0.03	0.03	0.04	16.50	9.08	0.00	0.01	18	nn_np_10997	nn_np_3830	nn_np_10927	nn_np_12341
0.07	0.03	0.50	0.10	0.50	0.50	0.00	1.00	1.00	0.00	19	lm_ll_2856	lm_ll_12917	lm_ll_9631	nn_np_10203
0.50	0.50	0.50	0.02	0.02	0.50	1.96	1.97	1.92	3.85	19	nn_np_10340	lm_ll_10921	nn_np_9928	lm_ll_2421

a distance of 530.03 cM. The average distance of markers on each individual linkage group was 0.40–0.66 cM (Zhang et al., 2017).

The recombination rates r_{AB}, r_{BC}, r_{CD}, r_{AC}, r_{BD}, and r_{AD} and the corresponding C_1, C_2, C_3, and C_4 between every four consecutive markers were obtained by four-point linkage analysis (**Table 3**). According to the CoC (**Table 3**) and the genetic distance of each linkage group, we determined the CO interference between two adjacent intervals, the CO interference of one interval apart, and the high dimensional CO interference of triple marker intervals. CO interference is ubiquitous within a genome, exhibiting COs between two adjacent marker intervals distributed throughout the genome and varied with the length of the chromosome (**Figure 1A**), making the distribution of COs across each linkage group more even. However, the distribution of interference between two non-adjacent marker intervals occasionally occurs at lower frequencies and lower intensities than the adjacent intervals (**Figure 1B**). Interestingly, high dimensional CO interference was highly distributed across the 19 linkage groups and had a wide distribution within the genome (**Figure 1C**). By comparison, high dimensional CO interference with high-density distribution existed on linkage group 4 (LG4) and linkage group 5 (LG5), whereas the high-dimensional CO interference distribution density of linkage group 11 (LG11) was lower.

We plotted the first eight high-dimensional CO interference in the 19 linkage groups to visualize the distribution of high-dimensional CO interference on the eight linkage groups more directly (**Figure 2**). Although the chromosome length varied, higher-dimensional CO interferences were evenly distributed within each chromosome and the amplitudes were larger and denser than the other two genetic disturbances. Additionally, the location information of the markers where CO interference occurred could be seen (**Figure 2**). There was an obvious correlation between the density of high-dimensional CO interference and chromosome length, with different chromosome lengths resulting in different distributions of high-dimensional CO interference.

We analyzed the correlation between the genetic distance of chromosomes and overall high-dimensional CO interference strength. The median of the overall CO interference strength was concentrated between 0 and 1, and the interquartile range (IQR) was variable and dependent on chromosome length. The IQR of chromosome 5 was the longest, reaching 41.63 cM; the IQR of chromosome 11 was the shortest, about 1.74 cM; the other 17 chromosomes were similar to chromosome 1, which was about 16.94 cM (**Figure 3**). In other words, the overall strength of CO interference was related to the genetic distance of the chromosome (**Figure 4**). Chromosomes 2, 3, 4, and 6 were locally linearly fitted (blue line) with an adjusted R^2 of 0.71. Simultaneously, the other chromosomes were fitted (red line) with an adjusted R^2 of 0.85 (**Figure 4**). Although the two fitted curves had different slopes, they both increased with the length of the chromosome. These results suggest that the correlation between the genetic distance of chromosomes and the overall high-dimensional CO interference strength was significant.

We plotted the first three of the 19 chromosomes to visualize the distribution of high dimensional CO interference on different chromosome parts (NO.1, NO.2, and NO.3) (**Figure 5**). The CO interference strength of each chromosome part differed in terms of intensity interval. For example, on chromosome 1, there was no CO interference in the first part (interval of 60–80 cM), whereas chromosome 2 exhibited CO interference. Therefore, different intervals along the chromosome contained different strengths and distributions of CO interference.

The difference ratio was used to compare the differences among the three intervals on each chromosome and study the distribution of high dimensional CO interference strength in different regions of the chromosome. The difference ratios of NO.1-NO.2, NO.2-NO.3, and NO.1-NO.3 in each chromosome were 0.1429-0.9474, 0.0952-1.1250, and 0.2353-0.8750, respectively (**Figure 6**). Moreover, fluctuations of CO interference strength between the first region and the third region were small, whereas the CO interference strength between the second region and the third region fluctuated greatly (**Figure 6**). The high dimensional CO interference strength between the middle region and both side regions on the chromosome was

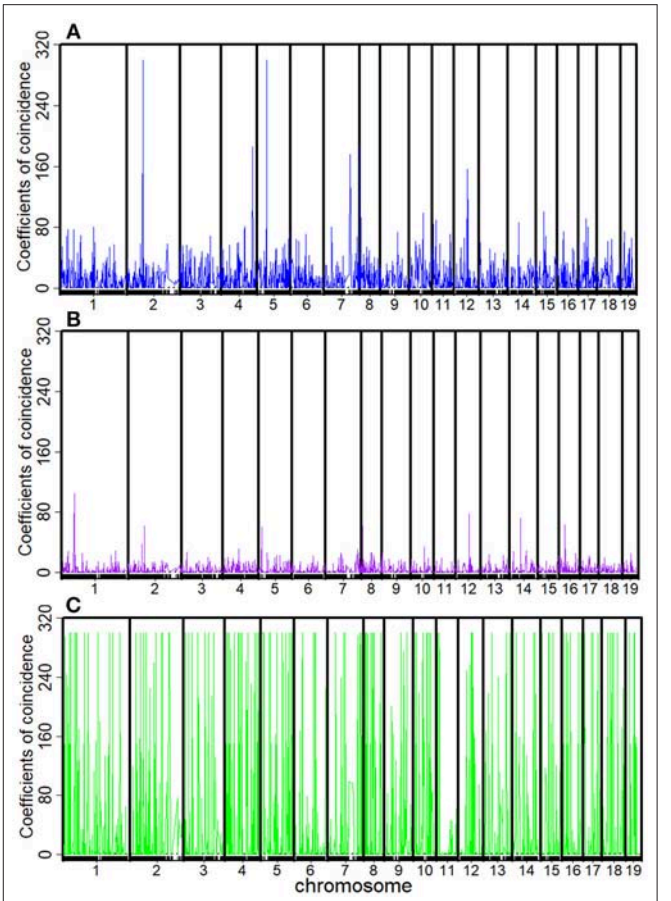


FIGURE 1 | Distribution of crossover interference within the *Populus euphratica* genome, composed of 19 chromosomes, estimated from a full-sib family of two different cultivars. **(A)** Crossover interference between two adjacent marker intervals (C_1 and C_2); **(B)** crossover interference between two non-adjacent marker intervals (C_3); **(C)** high-dimensional crossover interference over three successive marker intervals (C_4).

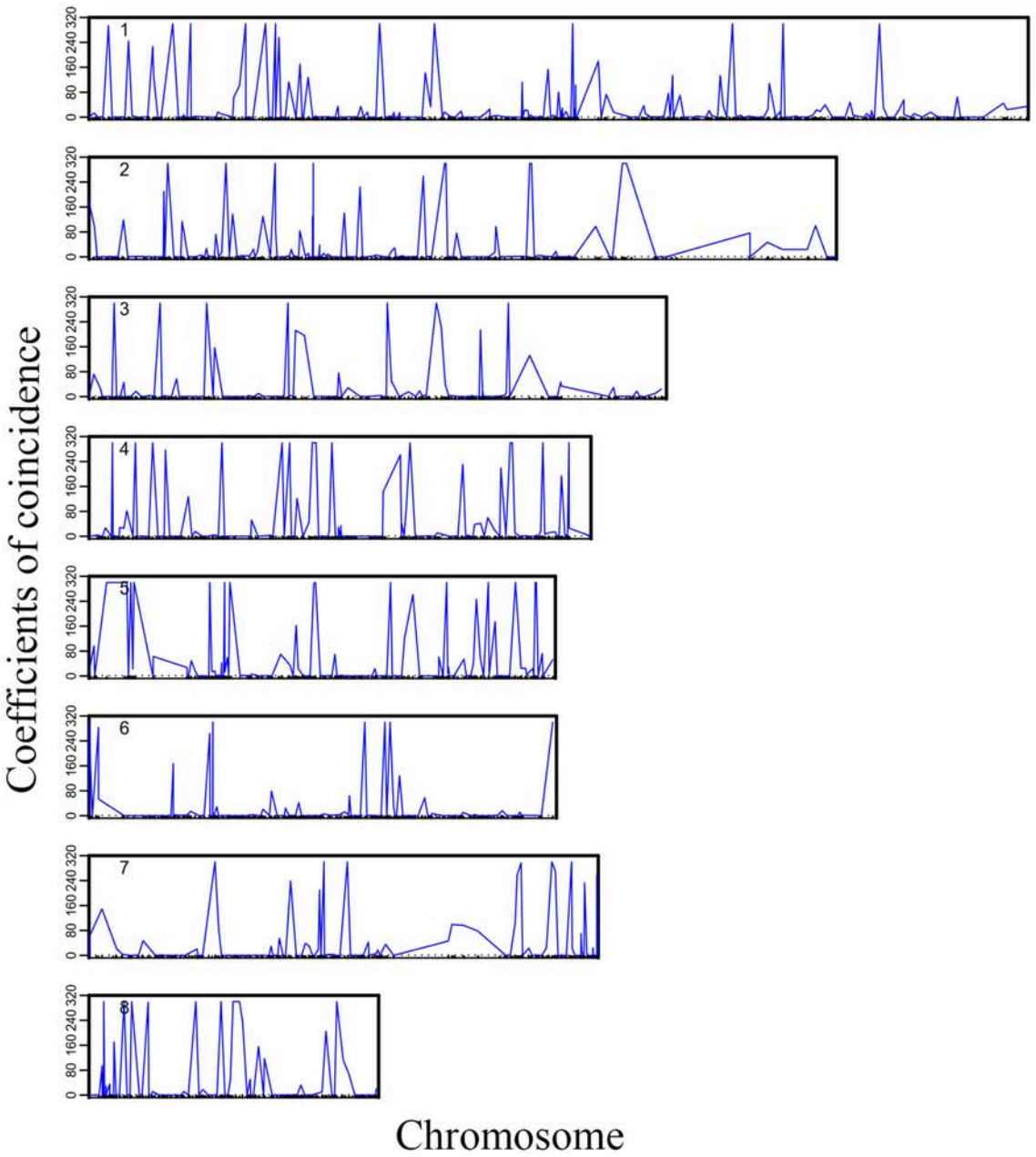


FIGURE 2 | Landscape of crossover interference along eight chromosomes within the *Populus euphratica* genome.

very different. Thus, the overall strength of high dimensional CO interference was not only related to the length of the chromosome, but also varied among chromosome regions.

DISCUSSION

The phenomenon of CO interference has been observed in most organisms. Within eukaryotes, interference may be quite long. For example, in the nematode *C. elegans*, interference can span a fusion chromosome of 50 Mb (Lian et al., 2008). The results of this study provide strong evidence for the existence of high-order CO interference. We assessed CO interference in the full-sib family of *P. euphratica* by mapping the distributions of CO interferences in different dimensions along 19 chromosomes. We observed that high-dimensional CO interference existed to varying degrees on all 19 chromosomes, and found that these high-dimensional interferences were even stronger than one- or two-dimensional CO interferences. The discovery of CO interference in the full-sib family of *P. euphratica* and the relationship between the strength of the overall CO interference

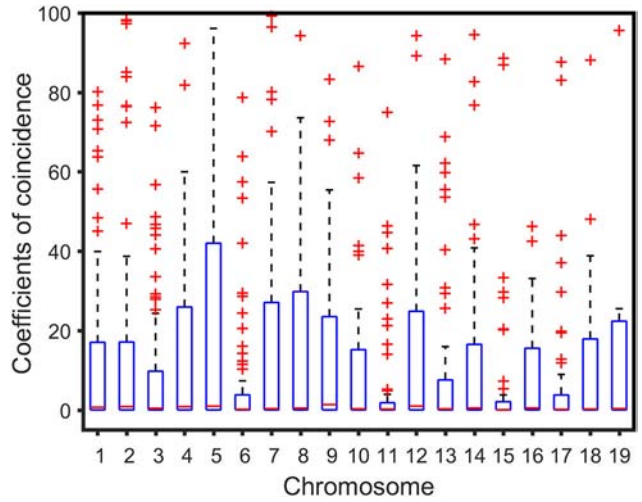


FIGURE 3 | Boxplot of crossover interference over the *Populus euphratica* genome composed of 19 chromosomes.

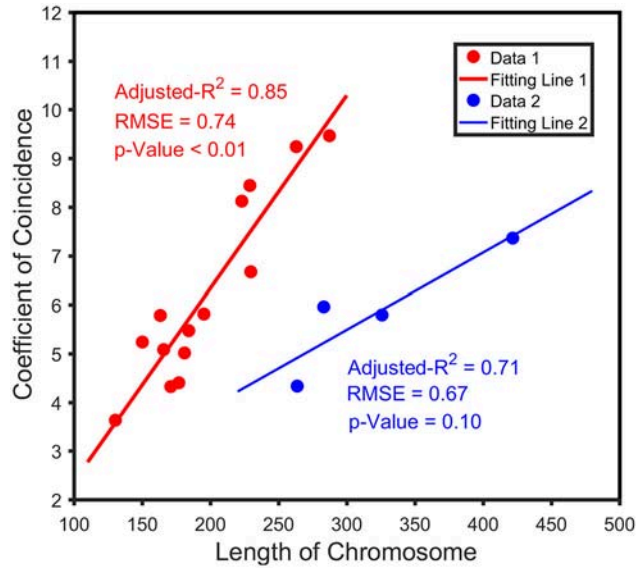


FIGURE 4 | Relationship between the length of the chromatid and the coefficient of coincidence (CoC). The blue line indicates the trend between crossover interference and the chromosome length of chromosomes 2, 3, 4, and 6; the red line indicates the trend of genetic interference of the remaining chromosomes with chromosome length.

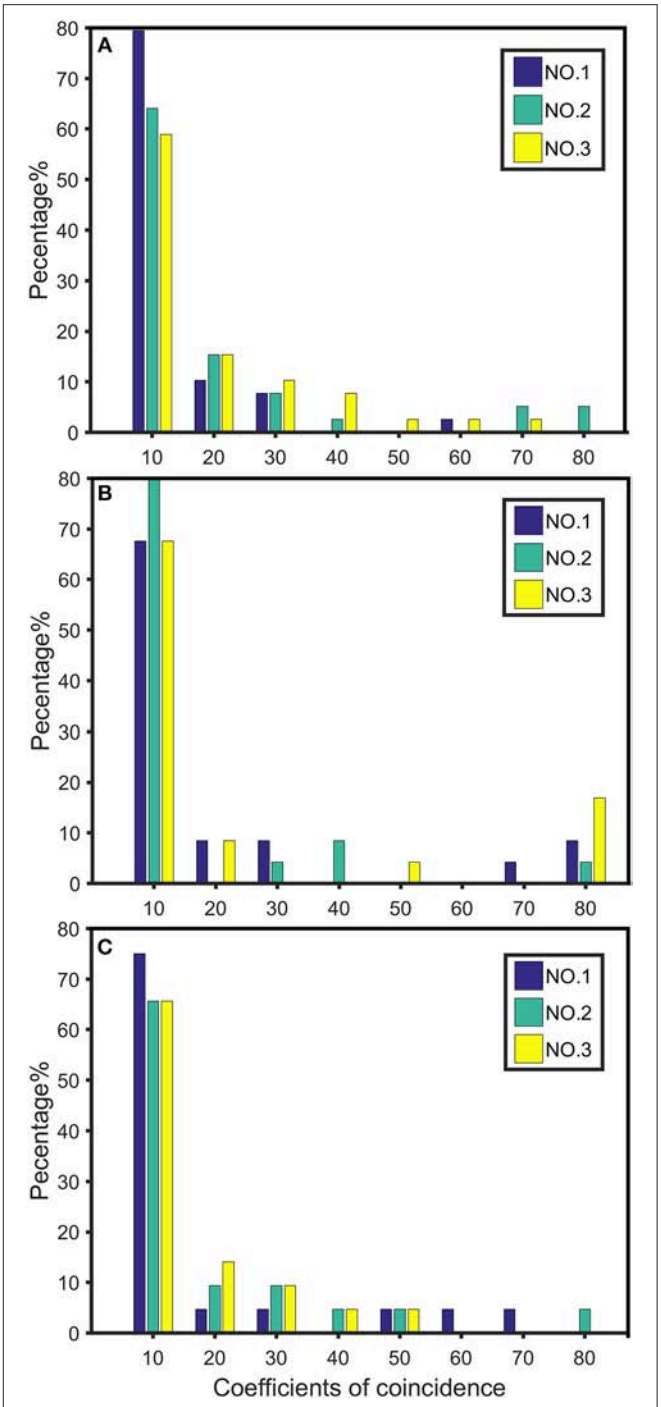


FIGURE 5 | Distribution of crossover interference strength on chromosomes 1 **(A)**, 2 **(B)**, and 3 **(C)**. NO.1 represents the distribution strength of the first third of the chromosome length; NO.2 represents the distribution strength of the middle third of the chromosome; NO.3 indicates the distribution strength of the last third of the chromosome.

and the chromosome structure can not only help identify and quantify CO interference in the entire genome, but also has the potential to impact further inference on the genome structure, organization, and evolution of *P. euphratica* populations.

We correlated the genetic length of the chromosome with the strength of the overall high-dimensional CO interference, and found that the mean of CO interference strength on each chromosome had a linear relationship with the genetic length of the chromosome. CO rates and chromosome lengths were previously found to be relevant in other eukaryotic species, including humans, mice, *Arabidopsis*, and zebrafish (Kleckner et al., 2003). In addition, CO interference affects the CO rate and is affected by the length of the chromosome. In some species, such as yeast, dogs, mice, and pigeons, small chromosomes often have a higher CO density (Froenicke et al., 2002; Basheva et al., 2008; Mancera et al., 2008). Surprisingly, the CO interference

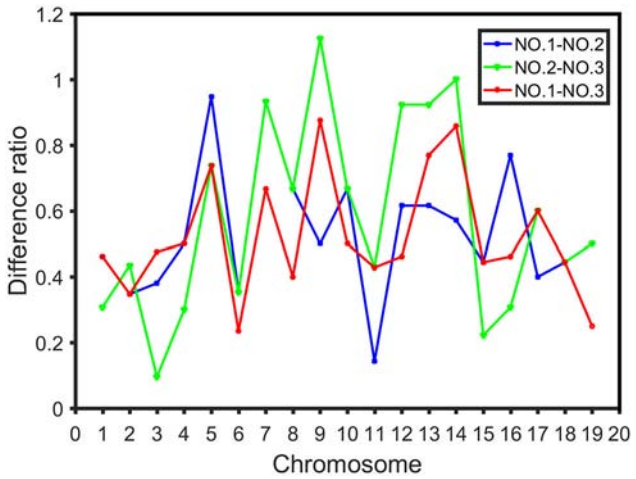


FIGURE 6 | Difference ratio in the distribution of CO interference between the three parts of the chromosome, where the blue line represents the difference ratio between the first part (NO.1) and the middle part (NO.2), the green line represents the difference ratio between the middle part (NO.2) and the third part (NO.3), and the red line indicates the difference ratio between the first part (NO.1) and the third part (NO.3).

strength in this study increased with chromosome length, with longer chromosomes containing a higher CO interference density and a correspondingly smaller CO density. This finding has far-reaching implications on biological evolution. Due to the existence of CO interference, the occurrence of CO events is regulated accordingly (Broman et al., 2002). The length of chromosomes indirectly affects the total strength of heritage interference, thereby affecting genetic diversity and having important implications for evolution.

According to previous studies, the occurrence of CO events is closely related to the center and terminal regions on chromosomes (Chelysheva et al., 2007). Meanwhile, CO interference has variable intensities and distributions in different regions of the chromosome. Moreover, CO interference can have different regulatory effects on a CO event in the corresponding region and exerts subtle influences on biological inheritance and evolution. We further studied the distribution and difference of CO interference between different regions on the chromosome, finding that the distribution of CO interference strength differed among regions. By defining the range of difference ratios, we found a difference in CO interference strength among chromosome regions. Studies of *Arabidopsis* chromosomes have shown that CO rates correlate with different genomic features associated with chromosome structure, such as the GC content and CpG ratio. Therefore, the differences in CO interference are also clearly related to these factors.

In this study, we used multi-point analysis methods to measure CO interference in the full-sib family of *P. euphratica*, extending from traditional linkage analysis to analyze multiple markers simultaneously. Previous studies have demonstrated that this method is a powerful tool for identifying and estimating CO interference (Wang J. et al., 2016). Accurate estimates of high-dimensional CO interference have significant implications in genomic research (Weeks et al., 1994). First, previous studies of interference in experimental organisms generally only involved adjacent interval groups, whereas multi-point analysis can not only accurately estimate the recombination rate between two adjacent markers, but also between multiple marker intervals and provide additional information about genomic structure and organization. Second, using this method, the strength and distribution of CO interferences between adjacent intervals along a chromosome can be estimated and the results can be used to study the relationship with the structure of the chromosome.

An increasing number of studies have investigated the phenomenon of CO interference. It has been found that CO interference is highly related to many evolutionary and developmental processes, such as gender differences, heterogeneity, senescence, and stress tolerance. The distribution of recombination achieved by CO interference can be determined by genetic background, gender, and many environmental factors, such as temperature and age. However, most genetic mapping studies have not considered CO interference. Regardless, multi-point analysis using genetic mapping has been used to estimate the degree of correlation between CO interference and evolution, and can capture this important phenomenon without extra cost. Similarly, Aggarwal et al. (2015) used multi-point analysis to determine the rules of recombinant frequency and CO interference in fruit flies that were targeted by dry, hypoxia, or high-oxygen tolerance. Here, we have expanded the research on CO interference, allowing for future studies to explore the molecular mechanism of CO in the *P. euphratica* genome through combination of multi-point analysis with cytology, clarify the development and evolution of COs, and investigate whether specific genes regulate CO interference.

AUTHOR CONTRIBUTIONS

PW performed data analysis. PW, MY, and XZ interpreted the result. RW and LJ conceived of the idea and designed the model. PW and LJ wrote the manuscript.

FUNDING

This work is supported by Fundamental Research Funds for the Central Universities (NO. BLYJ201605, NO. BLX201715, NO. 2015ZCQ-SW-06), grant 31700576 from National Natural Science Foundation of China, grant 31600536 from National Natural Science Foundation of China, grant 201404102 from the State Administration of Forestry of China, NSF/IOS award No. 0923975, and the Thousand-person Plan Award.

REFERENCES

Aggarwal, D. D., Rashkovetsky, E., Michalak, P., Cohen, I., Ronin, Y., Zhou, D., et al. (2015). Experimental evolution of recombination and crossover interference in Drosophila caused by directional selection for stress-related traits. *BMC. Biol.* 13:101. doi: 10.1186/s12915-015-0206-5

Albini, S. M. (2010). A karyotype of the *Arabidopsis thaliana* genome derived from synaptonemal complex analysis at prophase

I of meiosis. *Plant. J.* 5, 665–672. doi: 10.1111/j.1365-313X.1994.00665.x

Anderson, C. M., Oke, A., Yam, P., Zhuge, T., and Fung, J. C. (2015). Reduced crossover interference and increased ZMM-independent recombination in the absence of Tel1/ATM. *Plos. Genet*. 11:e1005478. doi: 10.1371/journal.pgen.1005478

Basheva, E. A., Bidau, C. J., and Borodin, P. M. (2008). General pattern of meiotic recombination in male dogs estimated by MLH1 and RAD51 immunolocalization. *Chromosome. Res*. 16:709. doi: 10.1007/s10577-008-1221-y

Berchowitz, L. E., and Copenhaver, G. P. (2010). Genetic interference: don't stand so close to me. *Curr. Genom*. 11, 91–102. doi: 10.2174/138920210790886835

Broman, K. W., Rowe, L. B., Churchill, G. A., and Paigen, K. (2002). Crossover interference in the mouse. *Genetics*. 160, 1123–1131.

Broman, K. W., and Weber, J. L. (2000). Characterization of human crossover interference. *Am. J. Hum. Genet*. 66, 1911–1926. doi: 10.1086/302923

Campbell, C. L., Furlotte, N. A., Eriksson, N., Hinds, D., and Auton, A. (2015). Escape from crossover interference increases with maternal age. *Nat. Commun*. 6:6260. doi: 10.1038/ncomms7260

Chelysheva, L., Gendrot, G., Vezon, D., Doutriaux, M. P., Mercier, R., and Grelon, M. (2007). Zip4/Spo22 is required for class I CO formation but not for synapsis completion in *Arabidopsis thaliana*. *PLoS. Genet*. 3:e83. doi: 10.1371/journal.pgen.0030083

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*. 39, 1–22. doi: 10.1111/j.2517-6161.1977.tb01600.x

Froenicke, L., Anderson, L. K., Wienberg, J., and Ashley, T. (2002). Male mouse recombination maps for each autosome identified by chromosome painting. *Am. J. Hum. Genet*. 71, 1353–1368. doi: 10.1086/344714

Getz, T. J., Banse, S. A., Young, L. S., Banse, A. V., Swanson, J., Wang, G. M., et al. (2008). Reduced mismatch repair of heteroduplexes reveals "non"-interfering crossing over in wild-type *Saccharomyces cerevisiae*. *Genetics*. 178, 1251–1269. doi: 10.1534/genetics.106.067603

Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O. C., and Mézard, C. (2011). Genome-wide crossover distribution in Arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. *PLoS. Genet*. 7:e1002354. doi: 10.1371/journal.pgen.1002354

Hillers, K. J. (2004). Crossover interference. *Curr. Biol*. 14, R1036–R1037. doi: 10.1016/j.cub.2004.11.038

Housworth, E. A., and Stahl, F. W. (2003). Crossover interference in humans. *Am. J. Hum. Genet*. 73, 188–197. doi: 10.1086/376610

Hultén, M. A. (2011). On the origin of crossover interference: a chromosome oscillatory movement (COM) model. *Mol. Cytogenet*. 4:10. doi: 10.1186/1755-8166-4-10

Jan, D., Raphaël, M., Liudmila, C., Aurélie, B., Matthieu, F., Olivier, M., et al. (2007). Sex-specific crossover distributions and variations in interference level along *Arabidopsis thaliana* chromosome 4. *PLoS. Genet*. 3:e106. doi: 10.1371/journal.pgen.0030106

Joshi, N., Barot, A., Jamison, C., and Börner, G. V. (2009). Pch2 links chromosome axis remodeling at future crossover sites and crossover distribution during yeast meiosis. *PLoS. Genet*. 5:e1000557. doi: 10.1371/journal.pgen.1000557

King, J. S., and Mortimer, R. K. (1990). A polymerization model of chiasma interference and corresponding computer simulation. *Genetics*. 126, 1127–1138.

Kleckner, N., Storlazzi, A., and Zickler, D. (2003). Coordinate variation in meiotic pachytene SC length and total crossover/chiasma frequency under conditions of constant DNA length. *Trends. Genet*. 19, 623–628. doi: 10.1016/j.tig.2003.09.004

Lam, S. Y., Horn, S. R., Radford, S. J., Housworth, E. A., Stahl, F. W., and Copenhaver, G. P. (2005). Crossover interference on nucleolus organizing region-bearing chromosomes in Arabidopsis. *Genetics* 170, 807–812. doi: 10.1534/genetics.104.040055

Lian, J., Yin, Y., Oliver-Bonet, M., Liehr, T., Ko, E., Turek, P., et al. (2008). Variation in crossover interference levels on individual chromosomes from human males. *Hum. Mol. Genet*. 17, 2583–2594. doi: 10.1093/hmg/ddn158

Lu, Q., Cui, Y., and Wu, R. (2004). A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full-sib family. *BMC Genet*. 5:20. doi: 10.1186/1471-2156-5-20

Mancera, E., Bourgon, R., Brozzi, A., Huber, W., and Steinmetz, L. M. (2008). High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature*. 454, 479–485. doi: 10.1038/nature07135

Qiu, Q., Ma, T., Hu, Q., Liu, B., Wu, Y., Zhou, H., et al. (2011). Genome-scale transcriptome analysis of the desert poplar, *Populus euphratica*. *Tree. Physiol*. 31, 452–461. doi: 10.1093/treephys/tpr015

Sun, L., Wang, J., Sang, M., Jiang, L., Zhao, B., Cheng, T., et al. (2017). Landscaping crossover interference across a genome. *Trends. Plant. Sci*. 22, 894–907. doi: 10.1016/j.tplants.2017.06.008

Szatkiewicz, J. P., Neale, B. M., O'Dushlaine, C., Fromer, M., Goldstein, J. I., Moran, J. L., et al. (2013). Detecting large copy number variants using exome genotyping arrays in a large Swedish schizophrenia sample. *Mol. Psychiatr*. 18, 1178–1184. doi: 10.1038/mp.2013.98

Wang, J., Sun, L., Jiang, L., Sang, M., Ye, M., Cheng, T., et al. (2016). A high-dimensional linkage analysis model for characterizing crossover interference. *Brief. Bioinform*. 18:382. doi: 10.1093/bib/bbw033

Wang, S., Zickler, D., Kleckner, N., and Zhang, L. (2015). Meiotic crossover patterns: obligatory crossover, interference and homeostasis in a single process. *Cell. Cycle*. 14, 305–314. doi: 10.4161/15384101.2014.991185

Wang, Z., Shen, B., Jiang, J., Li, J., and Ma, L. (2016). Effect of sex, age and genetics on crossover interference in cattle. *Sci. Rep*. 6:37698. doi: 10.1038/srep37698

Waterworth, D. (2000). Analysis of human genetic linkage. By J. O TT. Baltimore, London: Johns Hopkins University Press. 1999. Pp.382. *Ann. Hum. Genet*. 64, 89–92. doi: 10.1017/S0003480000227899

Weeks, D. E., Ott, J., and Lathrop, G. M. (1994). Detection of genetic interference: simulation studies and mouse data. *Genetics* 136, 1217–1226.

Yant, L., Hollister, J. D., Wright, K. M., Arnold, B. J., Higgins, J. D., Franklin, F. C. H., et al. (2013). Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. *Curr. Biol*. 23, 2151–2156. doi: 10.1016/j.cub.2013.08.059

Youds, J. L., Mets, D. G., Mcilwraith, M. J., Martin, J. S., Ward, J. D., Oneil, N. J., et al. (2010). RTEL-1 enforces meiotic crossover interference and homeostasis. *Science* 327, 1254–1258. doi: 10.1126/science.1183112

Zhang, M., Bo, W., Xu, F., Li, H., Ye, M., Jiang, L., et al. (2017). The genetic architecture of shoot-root covariation during seedling emergence of a desert tree, *Populus euphratica*. *Plant. J.* 90, 918–928. doi: 10.1111/tpj.13518

Zickler, D., and Kleckner, N. (2016). A few of our favorite things: pairing, the bouquet, crossover interference and evolution of meiosis. *Semin. Cell. Dev. Biol*. 54, 135–148. doi: 10.1016/j.semcdb.2016.02.024

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 May 2019
doi: 10.3389/fgene.2019.00437

Molecular Organization and Chromosomal Localization Analysis of *5S* rDNA Clusters in Autotetraploids Derived From *Carassius auratus* Red Var. (♀) × *Megalobrama amblycephala* (♂)

*QinBo Qin, QiWen Liu, ChongQing Wang, Liu Cao, YuWei Zhou, Huan Qin, Chun Zhao and ShaoJun Liu**

State Key Laboratory of Developmental Biology of Freshwater Fish, College of Life Sciences, Hunan Normal University, Changsha, China

OPEN ACCESS

Edited by:
Ancha Baranova, George Mason University, United States

Reviewed by:
Lenin Arias Rodriguez, Universidad Juárez Autónoma de Tabasco, Mexico
Jingou Tong, Institute of Hydrobiology (CAS), China

****Correspondence:***
ShaoJun Liu
lsj@hunnu.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *08 September 2018*
Accepted: *29 April 2019*
Published: *15 May 2019*

Citation:
Qin QB, Liu QW, Wang CQ, Cao L, Zhou YW, Qin H, Zhao C and Liu SJ (2019) Molecular Organization and Chromosomal Localization Analysis of 5S rDNA Clusters in Autotetraploids Derived From Carassius auratus Red Var. (♀) × Megalobrama amblycephala (♂). Front. Genet. 10:437. doi: 10.3389/fgene.2019.00437

The autotetraploid fish (4*n* = 200, RRRR) (abbreviated as 4*n*RR) resulted from the whole genome duplication of red crucian carp (*Carassius auratus* red var., 2*n* = 100, RR) (abbreviated as RCC). During investigation of the influence of polyploidization on organization and evolution of the multigene family of *5S* rDNA, molecular organization and chromosomal localization of the *5S* rDNA were characterized in autotetraploid fish. By sequence analysis of the coding region (*5S*) and adjacent non-transcribed spacer (NTS), three distinct *5S* rDNA units (type I: 203 bp; type II: 340 bp; and type III: 477bp) were identified and characterized in 4*n*RR. These *5S* rDNA units were inherited from their female parent (RCC), in which obvious base variations in NTS and array recombination of repeat units were found. Using fluorescence *in situ* hybridization employing different *5S* rDNA units as probes, these *5S* rDNA clusters were localized in chromosomes of 4*n*RR, respectively, and showed obvious loss of chromosomal loci (type I and type II). Our data revealed genetic variation of the *5S* rDNA multigene family in the genome of autopolyploid fish. Furthermore, results provided new insights into the evolutionary patterns of this vertebrate multigene family.

Keywords: autotetraploid line, distant hybridization, *5S* rDNA, FISH, chromosomal loci

INTRODUCTION

Polyploidy is a significant mode of speciation in eukaryotes (Mallet, 2007; Otto, 2007), especially in vertebrates. Ohno (1970) proposed the genome duplication hypothesis, in which two rounds of whole genome duplication occurred during early vertebrate evolution. Polyploids are generally divided in to categories depending on their chromosomal composition and their manner of formation. The autopolyploids have chromosome sets coming from the genome of one species (e.g., AAAA) and exhibit multivalent pairing during meiosis, while the allopolyploids result from the combination of sets of chromosomes from two or more

Abbreviations: 4*n*RB, allotetraploid hybrids; 4*n*RR, autotetraploid fish; BSB, blunt snout bream; FISH, fluorescence *in situ* hybridization; ICRs, internal control regions; NTS, non-transcribed spacer; PCR, polymerase chain reaction; RCC, red crucian carp.

different taxa (e.g., AABB) and predominantly form bivalent pairings (Comai, 2005). Notably, multivalent pairing may cause meiotic irregularities and result in reduced fertility compared with diploid progenitors (Jackson, 1982; Parisod et al., 2010). Thus, vertebrate autopolyploids are relatively rare compared with allopolyploids, and the influence of autopolyploidization on intragenomic variation is poorly understood.

5S ribosomal RNA (rRNA) is a component of the large ribosomal subunit in all ribosomes. In vertebrates, the *5S* ribosomal DNA (*5S* rDNA) is organized in tandem arrays with repeat units composed of a 120-bp coding sequence (*5S*) that encodes the *5S* rRNA and a highly variable non-transcribed spacer (NTS) (Korn and Brown, 1978; Nielsen et al., 1993; Hallenberg and Frederiksen, 2001; Pasolini et al., 2006). Molecular organization and chromosomal localization of the *5S* rDNA have been extensively characterized in bony fish (Iue et al., 1989; Rocco et al., 2005; Qin et al., 2010; Danillo et al., 2011). Polyploidization plays an important role in the evolution of fish. However, the features of the *5S* rDNA have been rarely reported in polyploid fish. Previously, we successfully obtained fertile allotetraploid hybrids (4*n* = 148, RRBB) (abbreviated as 4*n*RB) from the first generation of *Carassius auratus* red var. (2*n* = 100, RR) (♀) × *Megalobrama amblycephala* (2*n* = 48, BB) (♂) hybrids (Qin et al., 2014a). The abnormal chromosomal behavior of allotetraploid hybrids during meiosis leads to the formation of autodiploid sperm and autodiploid ova, and the fertilization of these ova by these sperm in turn produces autotetraploid fish (4*n*RR) (Qin et al., 2014b, 2015a). Autotetraploids produce diploid ova and diploid sperm and maintain the formation of the autotetraploid line (F_1–F_{10}), which could be used as a new model system for investigating the influence of polyploidy on the organization and evolution of the multigene family of *5S* rDNA. In this paper, molecular organization and chromosomal localization of the *5S* rDNA have been characterized in the autotetraploid and their parents (RCC). Obvious loss of chromosomal loci, base variations in NTS, and array recombination of repeat units have been found in the newly established autotetraploidy genomes. Our results extend the knowledge of the influence of polyploidy on the organization and evolution of *5S* rDNA of fish, and are also useful in clarifying aspects of vertebrate genome evolution.

MATERIALS AND METHODS

Source of Samples

All fish were cultured in ponds and fed with artificial feed at the Protection Station of Polyploidy Fish, Hunan Normal University. Fish treatments were carried out according to the regulations for protected wildlife and the Administration of Affairs Concerning Animal Experimentation, and approved by the Science and Technology Bureau of China. Approval from the Department of Wildlife Administration was not required for the experiments conducted in this paper. The fish were deeply anesthetized with 100 mg/L MS-222 (Sigma-Aldrich, St. Louis, MO, United States) before dissection.

Animals and Crosses

During the reproductive season (April to June) in 2012, the first generation (4*n*RB) of *C. auratus* red var. (♀) × *M. amblycephala* (♂) was produced. During the reproductive season (April to June) of 2014, the second generation (4*n*RR) was produced by self-crossing of 4*n*RB.

Preparation of Chromosome Spreads

Chromosome counts were performed using kidney tissue from 10 RCC and 10 4*n*RR. After culture for 1–3 days at a water temperature of 18–22°C, the samples were injected with concanavalin one to three times at a dose of 2–8 mg/g body weight. The interval between injections was 12–24 h. Six hours prior to dissection each sample was injected with colchicine at a dose of 2–4 mg/g body weight. The excised kidney tissue was ground in 0.9% NaCl, followed by hypotonic treatment with 0.075 M KCl at 37°C for 40–60 min and then fixed in 3:1 methanol–acetic acid with three changes. The cells were dropped onto cold, wet slides and stained for 30 min in 4% Giemsa. The shape and number of chromosomes were analyzed under a microscope. For each type of fish, 200 metaphase spreads (20 metaphase spreads from each sample) of chromosomes were analyzed. The preparations were examined under an oil lens at a magnification of 3330×.

PCR Amplification and Sequencing of *5S* rDNA Sequences

Total genomic DNA was isolated from peripheral blood cells according to the standard phenol: chloroform extraction procedure described by Sambrook et al. (1989). To acquire preliminary information on the organization of the *5S* rDNA repeat variants, and to test for the possible coexistence of different repeat units in the same array, DNA samples of 3 RCC and 3 4*n*RR were amplified with primers *5S*P1-*5S*P2R (5′-GCTATGCCCGATCTCGTCTGA-3′ and 5′-CAGGTTGGTAT GGCCGTAAGC-3′) and with primers *5S*NT1-*5S*NT2R (5′-GGCGAGTAGATTGGCTGAACA-3′ and 5′-CAATCTAATCGCCAGTACATTATAT-3′). The PCR reaction was performed in a volume of 25 μL with approximately 20 ng of genomic DNA, 1.5 mM of $MgCl_2$, 200 μM of each dNTP, 0.4 μM of each primer, and 1.25 U of Taq polymerase (Takara). The temperature profile was as follows: an initial denaturation step at 94°C for 5 min, followed by 30 cycles of 94°C for 30 s, 56°C for 30 s, and 72°C for 1 min, with a final extension step at 72°C for 10 min. Amplification products were separated on a 3.0% agarose gel using TBE buffer. The DNA fragments were purified using a gel extraction kit (Sangon) and ligated into pMD18-T (Takara). Plasmids were transformed into *Escherichia coli* DH5a, propagated, and then purified. The cloned DNA fragments were sequenced using an automated DNA sequencer (ABI PRISM 3730). Sequence homology and variation among the fragments amplified from 3 RCC and 3 4*n*RR were analyzed using ClustalW software[1].

[1]http://www.ebi.ac.uk/clustalw/intex.html

Fluorescence *in situ* Hybridization

The probes for fluorescence *in situ* hybridization (FISH) for the *5S* gene were constructed for RCC and amplified by PCR using the primers 5′-GCTATGCCCGATCTCGTCTGA-3′ and 5′-CAGGTTGGTATGGCCGTAAGC-3′. The FISH probes were produced by Dig-11-dUTP labeling (using a Nick Translation Kit, Roche, Germany) of purified PCR products. Purified PCR products of *5S* rDNA labeled with Dig-11-dUTP (Roche, Germany) were used as probes, and hybridization was performed according to the method described by Yi et al. (2003) with minor modifications. Purified PCR products of *5S* rDNA labeled with Dig-11-dUTP (Roche, Germany) were used as probes, and hybridization was performed according to the method described by Yi et al. (2003) with minor modifications. After treatment with 30 μg/ml RNase A in 2 × SSC for 30 min at 37°C, the slides with chromosome metaphase spreads were denatured in 70% deionized formamide/2 × SSC for 2 min at 70°C, dehydrated in a 70, 90, and 100% ethanol series for 5 min each (1 × SSC is 0.15 M NaCl/0.015 M sodium citrate, pH 7.6), and then air-dried. 4 μl of the hybridization mixture (approximately 100 ng of labeled probes, 50% formamide, 10 mg dextran sulfate/ml and 2 × SSC) was denatured for 10 min in boiling water, applied to the air-dried slides carrying denatured metaphase chromosomes under a 22 × 22 mm coverslip, and sealed with rubber cement. The slides were then put in a moist chamber and allowed to incubate overnight at 37°C.

Following overnight incubation, the coverslips were removed and the slides were rinsed at 43°C in: 2 × SSC with 50% formamide, twice, 15 min each; 2 × SSC, 5 min; 1 × SSC, 5 min, then air-dried. The spectrum signals were achieved by application of 8 μl of 5 μg/ml FITC-conjugated antidigoxigenin antibody from sheep (Roche, Germany) and a final incubation in the humidity chamber at 37°C. After a series of washes with TNT (containing 0.1 M Tris–HCl, 0.15 M NaCl, 0.05% Tween 20) at 43°C, the slides were mounted in antifade solution containing 2 μg/ml 4′, 6-diamidino-2-phenylindole (DAPI) for 5 min. Slides were viewed under a Leica inverted CW4000 microscope and a Leica LCS SP2 confocal image system (Leica, Germany). Metaphase spreads of chromosomes were analyzed in 10 RCC and 10 4*n*RR (20 metaphase spreads in each sample).

RESULTS

Molecular Organization of the *5S* rDNA Classes

Using the primers *5S*P1 and *5S*P2R, fragments of approximately 200, 340, and 500 bp were generated from RCC and 4*n*RR (**Figure 1A**). All fragments proved to be *5S* rDNA sequences, each included the 3′ end of the coding region (pos. 1-21), the whole NTS region, and a large 5′ portion of the coding region of the adjacent unit (pos. 22-120; see **Figure 1B**). In RCC, the three types of *5S* rDNA classes (designated type I: 203 bp; type II: 340 bp; and type III: 477 bp) were characterized by distinct NTS types (designated NTS-I, NTS-II and NTS-III for the 83-, 220-, and 357-bp sequences, respectively; **Figure 2**). 4*n*RR had three types of *5S* rDNA classes, which were completely inherited from RCC (type I, type II and type III; **Figure 2**). All *5S* rDNA sequences have been submitted to GenBank, and their accession numbers are listed in **Table 1**.

Comparison of the 120-bp coding region of *5S* rDNA with those of RCC and 4*n*RR revealed great similarity (**Figure 3**). Nucleotide variation was not detected among the internal control regions (ICRs, i.e., the promoters for transcription) in 4*n*RR (**Figure 3**). A comparison of NTS-I revealed six base substitutions among the sequences (**Figure 4A**). A comparison of NTS-II showed five base substitutions and a deletion-insertion at position -177 (**Figure 4B**). A comparison of NTS-III elements showed nine base substitutions and a deletion-insertion at position -164 (**Figure 4C**). The above results indicate that obvious nucleotide variations were found in NTS sequences of 4*n*RR. In addition, characterization of the NTS-up stream region showed

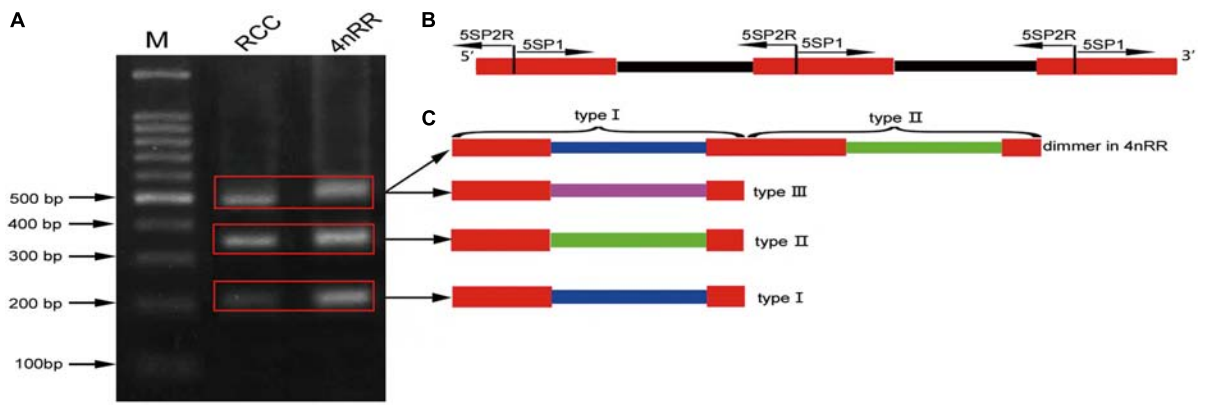


FIGURE 1 | DNA bands amplified from RCC and 4*n*RR. **(A)** Using the primer pair *5S*P1-*5S*P2R, DNA fragments were amplified from RCC and 4*n*RR. M, DNA ladder markers (100 bp increments); lane 1, three DNA fragments (approximately 200, 340, and 500 bp) from RCC; lane 2, three DNA fragments (approximately 200, 340, and 500 bp) from the 4*n*RR; **(B)** Arrangement of higher eukaryotic *5S* rRNA genes (red) intercalated with non-transcribed DNA segments (NTS; black); **(C)** Representative sequences of *5S* rDNA type I, II, and III from RCC and 4*n*RR; red indicate *5S* rRNA genes; blue, green and purple indicate distinct NTS sequences.

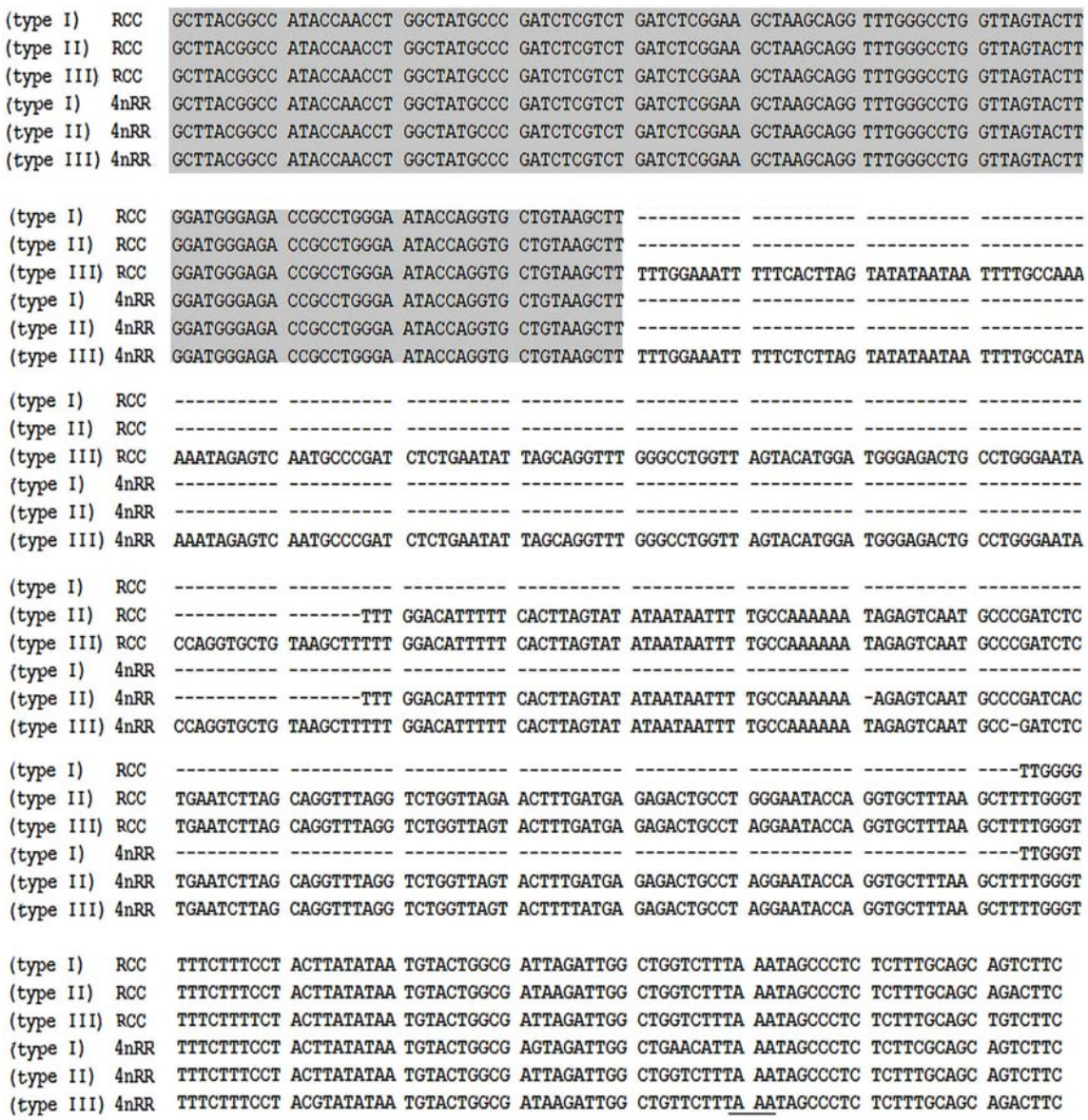


FIGURE 2 | Representative sequences of *5S* rDNA from RCC and 4*n*RR. Complete *5S* coding regions are shaded; the NTS upstream TATA elements are underlined.

that the TATA control element, the regulatory region for *5S* gene transcription, was identifiable in the NTS of RCC and 4*n*RR (at -29 in all NTS sequences, where it was modified to TAAA; **Figure 4**), suggesting that all sequences analyzed here were likely to correspond to functional genes.

Array Recombination of the *5S* rDNA Repeat Units

Thirty clones of the 500 bp fragment from 4*n*RR were analyzed, and the sequence analysis revealed that five clones were dimeric *5S* rDNA formed by *5S* rDNA type I (the 99 bp gene sequence, 83 bp of NTS, and 21 bp gene sequence) and *5S* rDNA type II (the 99 bp gene sequence, 220 bp of NTS, and 21 bp gene sequence) (**Figure 1C** and **Supplementary Figure 1**). To verify whether the different *5S* rDNA classes (type I and type II) were associated within the same tandem array, we designed

TABLE 1 | GenBank accession numbers of the *5S* rDNA sequences in RCC and 4*n*RR.

DNA fragments (bp)	GenBank accession numbers of the sequences	
	RCC	**4*n*RR**
203	GQ485555	MH44410
339, 340	GQ485556	MH44408
476, 477	GQ485557	MH44409

RCC, Carassius auratus red var.; 4nRR, autotetraploid fish.

FIGURE 3 | Comparison of *5S* coding regions from RCC and 4*n*RR. Internal control regions of the coding region are shaded.

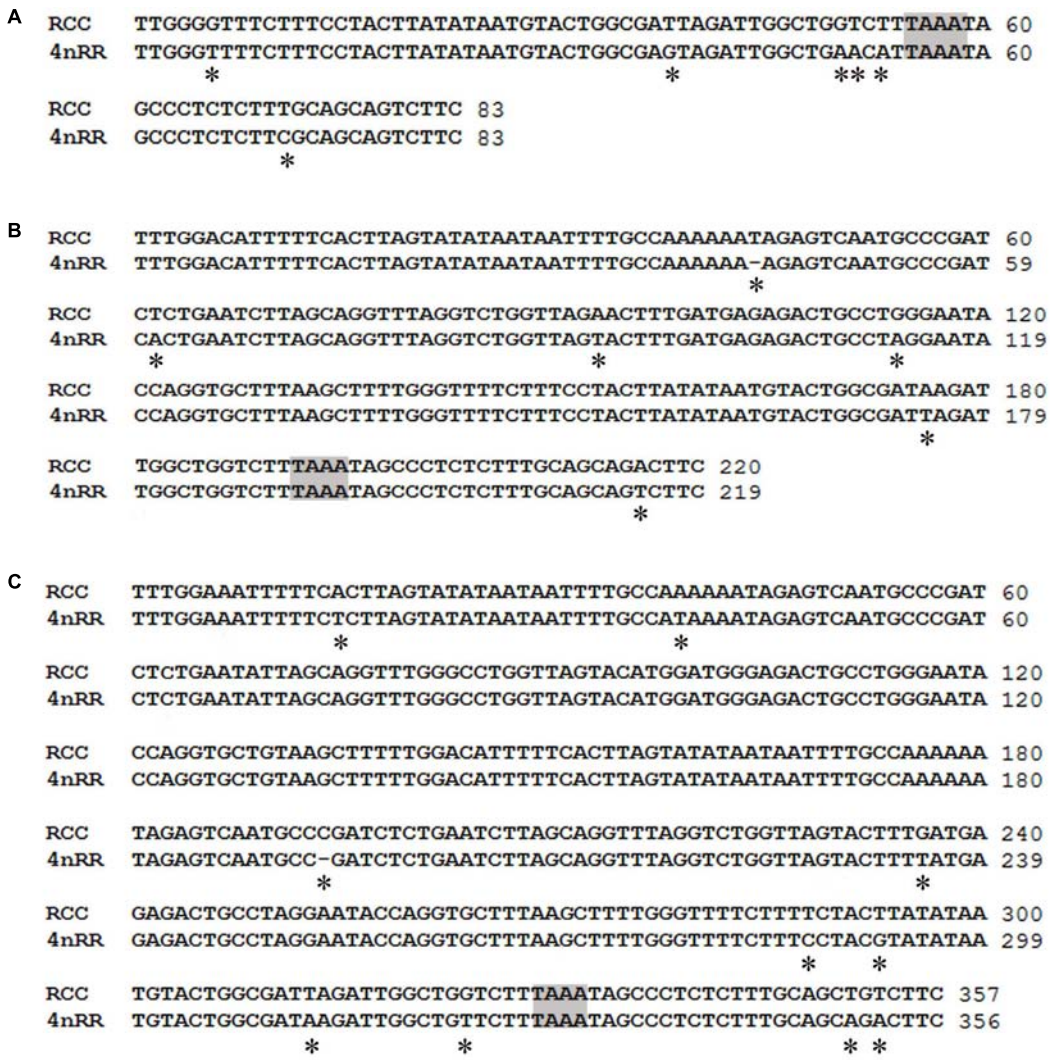


FIGURE 4 | Comparison of the NTS sequences from RCC and 4*n*RR. **(A)** NTS-I sequences from RCC and 4*n*RR; **(B)** NTS-II from RCC and 4*n*RR; **(C)** NTS-III from RCC and 4*n*RR. The NTS upstream TATA elements are shaded; asterisks mark variable sites in the NTS.

the primers *5S*NT1-*5S*NT2R. Using these primers, the PCR yielded a single band of 352 bp in 4*n*RR, but no band in RCC and 4*n*RB (**Figure 5**). Sequence analysis revealed that this fragment was formed by 72 bp of the type I (a 51 bp of the NTS and 21 bp gene sequence) and 280 bp of type II (the 99 bp of gene sequence and 181 bp of NTS) (**Figure 5** and

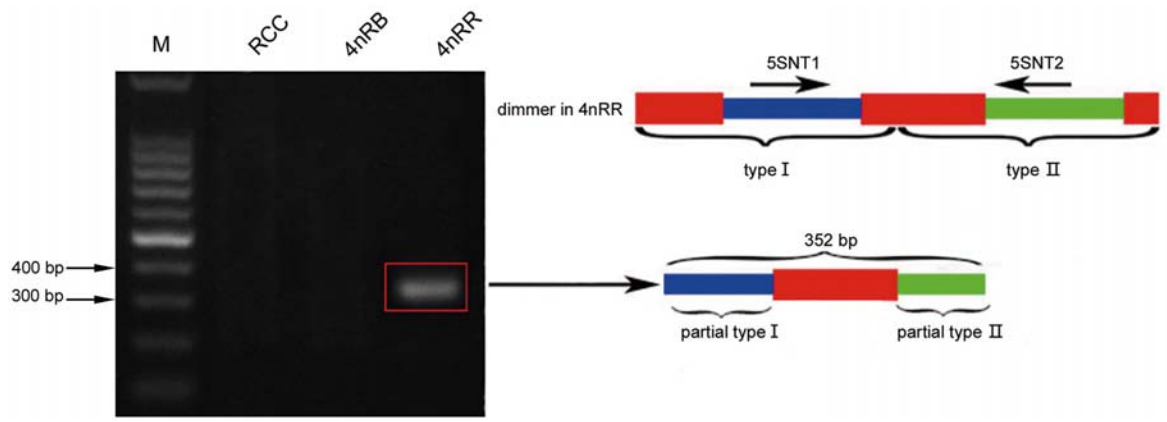


FIGURE 5 | DNA bands amplified from 4*n*RR. Using the primer pair *5S*NT1-*5S*NT2R, DNA fragments were only amplified from 4*n*RR (approximately 352 bp); M, DNA ladder markers (100 bp increments); red indicates the *5S* rRNA gene, blue indicates NTS of type I, and green indicates NTS of type II.

Supplementary Figure 2). The PCR amplification products of the two primers provided direct evidence to prove that the type I and type II repeats were associated within the same tandem array in 4*n*RR, suggesting that recombination of chromosomes occurred in the autotetraploid genome.

Chromosomal Loci of *5S* rDNA

The hybridization of type I *5S* rDNA probes showed eight *5S* gene loci in RCC chromosomal metaphases (**Figure 6A** and **Table 2**). Sixteen *5S* gene loci were expected in 4*n*RR chromosomal metaphases, but only twelve *5S* gene loci were found (**Figure 6B** and **Table 2**). Using type II *5S* rDNA as a probe, a pair of large *5S* gene loci was identified on homologous submetacentric chromosomes in RCC chromosomal metaphases, and a pair of small *5S* gene loci was localized on homologous subtelocentric chromosomes (**Figure 6C** and **Table 2**). In 4*n*RR chromosomal metaphases, a pair of large *5S* gene loci on a homologous submetacentric chromosome were found and other a pair of large *5S* gene loci on a homologous submetacentric chromosome were lost; two pairs of small *5S* gene loci was localized on homologous subtelocentric chromosomes (**Figure 6D** and **Table 2**). FISH hybridization of the type III *5S* rDNA probe to the RCC metaphase chromosomes yielded eight *5S* gene loci (**Figure 6E** and **Table 2**). As expected, sixteen *5S* gene loci were found in 4*n*RR chromosomal metaphases (**Figure 6F** and **Table 2**). The above results indicate that obvious loss of chromosomal loci occurred in 4*n*RR.

DISCUSSION

The evolution of *5S* rDNA is driven by birth-and-death processes with strongly purifying selection (Nei and Rooney, 2005; Pinhal et al., 2011; Vizoso et al., 2011), which can lead to the existence of different types of NTS (Pinhal et al., 2011). In teleosts, two distinct *5S* rDNA classes are characterized by distinct NTS types and base substitutions in the *5S* rRNA gene (Pendas et al., 1994; Moran et al., 1996; Martins et al., 2000; Wasko et al., 2001; Pinhal et al., 2009). Thus, possession of two *5S* rDNA classes seems to be a general trend for the organization of these sequences in the genomes of fish (Martins and Galetti, 2001). As ancient polyploidy fish, RCC have undergone an additional round of whole-genome duplication (Qin et al., 2016b). The origin of genic variants has been attributed to events of genome duplication followed by processes that result in the divergence of the duplicated sequences. Thus, RCC possess three distinct *5S* rDNA classes that are characterized by distinct types of NTS (Qin et al., 2010). In the current study, 4*n*RR derived from the distant hybridization of *C. auratus* red var. ($2n$ = 100, RR) (♀) × *M. amblycephala* ($2n$ = 48, BB) (♂), possess four sets of RCC-derived chromosomes and exhibit stability in chromosome number (or ploidy) over consecutive generations (F_1–F_{10}) (Qin et al., 2014b). 4*n*RR have three distinct *5S* rDNA classes that are completely inherited from RCC, but no new types of *5S* rDNA class were found, suggesting that divergence of the duplicated *5S* rDNA sequences were not fully formed in the early generations of the autotetraploid fish.

Because of incompatibility between parental chromosomes, allopolyploidization can increase genomic changes (Pontes et al., 2004). Our previous study revealed the influence of allopolyploidy on *5S* rDNA in fish, including parental genome specific loss, substitutions, and insertions-deletions in the NTS sequence (Qin et al., 2010, 2016a). Theoretically, homologous chromosomes should have high compatibility in autotetraploids. In this paper, however, obvious base variation and insertions-deletions of NTS were also observed in 4*n*RR, suggesting that autotetraploidization could lead to genetic variation in newly established autotetraploid genomes. Although there was genetic variation in NTS of *5S* rDNA, all sequences analyzed here were likely to correspond to functional genes, because they exhibited all the necessary features for correct gene expression: three ICRs (box A, internal element, and box C), a TATA control element, and a T-rich tail.

Autopolyploids are traditionally used to demonstrate multivalent pairing multivalent pairing during meiosis. However, the coexistence of four homologous chromosome sets does

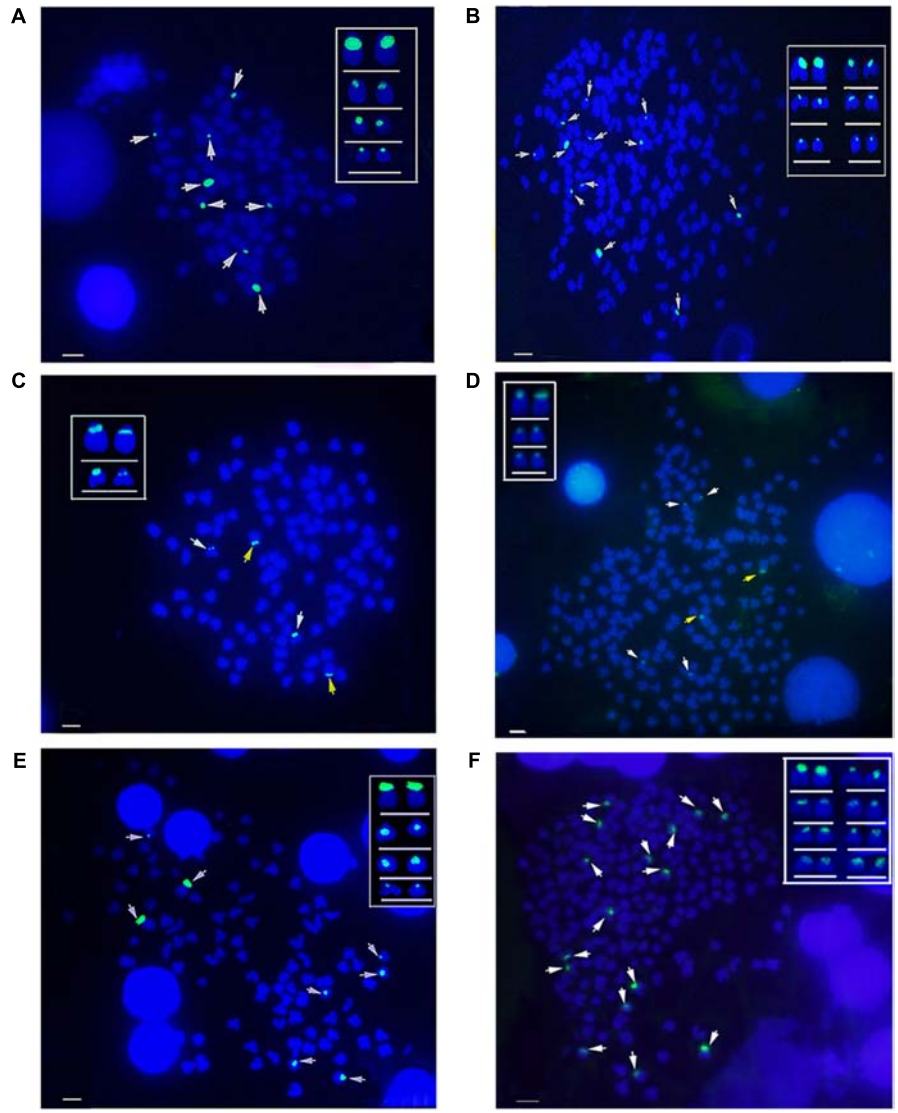


FIGURE 6 | Examination of hybridizing signals by FISH in RCC and 4*n*RR. **(A)** Type I as probe showed eight *5S* gene loci (white arrows) in RCC; **(B)** Type I as probe showed twelve *5S* gene loci of type I (white arrows) in 4*n*RR; **(C)** Type II as probe showed two big (yellow arrows) and two small *5S* gene loci (white arrows) in RCC; **(D)** Type II as probe showed two big (yellow arrows) and four small *5S* gene loci (white arrows) in 4*n*RR; **(E)** Type III as probe showed eight *5S* gene loci (white arrows) in RCC; **(F)** Type III as probe sixteen *5S* gene loci (white arrows) in 4*n*RR. Bar = 3 μm.

TABLE 2 | Examination of chromosome locus numbers in RCC and 4*n*RR.

Fish type	No. of fish	No. of metaphase	Type I	Type II		Type III
			No. of loci	No. of big loci	No. of small loci	No. of loci
RCC	10	200	8	2	2	8
4*n*RR	10	200	12	2	4	16

RCC, Carassius auratus red var.; 4nRR, autotetraploid fish.

not result in multivalent formation during meiosis in 4*n*RR, and diploid-like chromosome pairing was restored (Qin et al., 2019). The presence of two distinct *5S* rDNA sequence types organized in different chromosomal regions or even on different chromosomes has been described for several fish (Pendas et al., 1994; Moran et al., 1996; Sajdak et al., 1998; Martins et al., 2002; Rodrigues et al., 2012; Qin et al., 2015b). In the current study, the different *5S* rDNA classes (type I and type II) were associated within the same tandem array in 4*n*RR. In addition, type I and type II *5S* rDNA clusters were localized in the

chromosomes of 4*n*RR, and showed obvious loss of chromosomal loci. These findings are clear evidence that elimination of repetitive sequences and recombination of chromosomes occurred in newly established autotetraploid genomes. A positive linear relationship was found between increased bivalent pairing and elimination of specific, low-copy DNA sequences (Wendel, 2000). Thus, we speculate that the elimination of DNA sequences or recombination of chromosomes might generate immediate divergence between homologous chromosomes, providing a physical basis for diploid-like chromosome pairing in 4*n*RR.

ETHICS STATEMENT

Fish treatments were carried out according to the regulations for protected wildlife and the Administration of Affairs Concerning Animal Experimentation, and approved by the Science and Technology Bureau of China. Approval from the Department of Wildlife Administration was not required for the experiments conducted in this manuscript. The fish were deeply anesthetized with 100 mg/L MS-222 (Sigma-Aldrich, St. Louis, MO, United States) before dissection.

AUTHOR CONTRIBUTIONS

QQ and SL designed the experiments. QL, CW, LC, YZ, HQ, and CZ performed the experiments. QQ and QL performed the statistical analysis. QQ wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This research was financially supported by grants from the Natural Science Foundation of Hunan Province for Distinguished Young Scholars (Grant No. 2017JJ1022), the National Natural Science Foundation of China (Grant Nos. 31430088 and 31210103918), the Major Program of the Educational Commission of Hunan Province (Grant No. 17A133), the State Key Laboratory of Developmental Biology of Freshwater Fish, the Cooperative Innovation Center of Engineering and New Products for Developmental Biology of Hunan Province (20134486), the Earmarked Fund for China Agriculture Research System (CARS-45), and the Construction Project of Key Disciplines of Hunan Province and China.

ACKNOWLEDGMENTS

We would like to sincerely thank many researchers who help to complete this manuscript, including Drs. Yao Zhanzhou and Zhao Rurong.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00437/full#supplementary-material

REFERENCES

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.* 6, 836–846. doi: 10.1038/nrg1711

Danillo, P., Yoshimura, T. S., Araki, C. S., and Martins, C. (2011). The 5S rDNA family evolves through concerted and birth-and-death evolution in fish genomes: an example from freshwater stingrays. *BMC Evol. Biol.* 11:151. doi: 10.1186/1471-2148-11-151

Hallenberg, C., and Frederiksen, S. (2001). Effect of mutations in the upstream promoter on the transcription of human 5S rRNA genes. *Biochim. Biophys. Acta* 1520, 169–173. doi: 10.1016/S0167-4781(01)00264-0

Iue, S., Shostak, N. G., Kuprianova, N. S., Serenkova, T. I., Fel'Gengauér, P. E., Gimalov, F., et al. (1989). Intragenomic polymorphism of the primary structure of 5s rRNA gene variants of the loach (*misgurnus fossilis* l.). Determination of the transcriptional activity. *Mol. Biol.* 23, 1295–1308.

Jackson, R. C. (1982). Polyploidy and diploidy: new perspectives on chromosome pairing and its evolutionary implications. *Am. J. Bot.* 69, 1512–1523. doi: 10.1002/j.1537-2197.1982.tb13400.x

Korn, L. J., and Brown, D. D. (1978). Nucleotide sequence of *Xenopus borealis* oocyte 5S DNA: comparison of sequences that flank several related eucaryotic genes. *Cell* 15, 1145–1156. doi: 10.1016/0092-8674(78)90042-9

Mallet, J. (2007). Hybrid speciation. *Nature* 446, 279–283. doi: 10.1038/nature05706

Martins, C., and Galetti, P. M. Jr. (2001). Two 5S rDNA arrays in neotropical fish species: is it a general rule for fishes? *Genetica* 111, 439–446. doi: 10.1023/A:1013799516717

Martins, C., Wasko, A. P., Oliveira, C., Porto-Foresti, F., Parise-Maltempi, P. P., Wright, J. M., et al. (2002). Dynamics of 5S rDNA in the tilapia (*Oreochromis niloticus*) genome: repeat units, inverted sequences, pseudogenes and chromosome loci. *Cytogenet. Genome Res.* 98, 78–85. doi: 10.1159/000068542

Martins, C., Wasko, A. P., Oliveira, C., and Wright, J. M. (2000). Nucleotide sequence of 5s rDNA and localization of the ribosomal RNA genes to metaphase chromosomes of the tilapiine cichlid fish, oreochromis niloticus. *Hereditas* 133, 39–46. doi: 10.1111/j.1601-5223.2000.00039.x

Moran, P., Martinez, J. L., Garcia-Vazquez, E., and Pendas, A. M. (1996). Sex chromosome linkage of 5S rDNA in rainbow trout (*Oncorhynchus mykiss*). *Cytogenet. Genome Res.* 75, 145–150. doi: 10.1159/000134466

Nei, M., and Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.* 39, 121–152. doi: 10.1146/annurev.genet.39.073003.112240

Nielsen, J. N., Hallenberg, C., Frederiksen, S., Sørensen, P. D., and Lomholt, B. (1993). Transcription of human 5S rRNA genes is influenced by an upstream DNA sequence. *Nucleic Acids Res.* 21, 3631–3636. doi: 10.1093/nar/21.16.3631

Ohno, S. (1970). Evolution by gene duplication. *Am. J. Hum. Genet.* 23:541.

Otto, S. P. (2007). The evolutionary consequences of polyploidy. *Cell* 131, 452–462. doi: 10.1016/j.cell.2007.10.022

Parisod, C., Holderegger, R., and Brochmann, C. (2010). Evolutionary consequences of autopolyploidy. *New Phytol.* 186, 5–17. doi: 10.1111/j.1469-8137.2009.03142.x

Pasolini, P., Costagliola, D., Rocco, L., and Tinti, F. (2006). Molecular organization of 5S rDNAs in Rajidae (Chondrichthyes): structural features and evolution of piscine 5S rRNA genes and nontranscribed intergenic spacers. *J. Mol. Evol.* 62, 564–574. doi: 10.1007/s00239-005-0118-z

Pendas, A. M., Moran, P., Freije, J. P., and Garcia-Vazquez, E. (1994). Chromosomal mapping and nucleotide sequence of two tandem repeats of Atlantic salmon 5S rDNA. *Cytogenet. Cell Genet.* 67, 31–36. doi: 10.1159/000133792

Pinhal, D., Araki, C. S., Gadig, O. B. F., and Martins, C. (2009). Molecular organization of 5s rdna in sharks of the genus *Rhizoprionodon*: insights into the evolutionary dynamics of 5s rdna in vertebrate genomes. *Genet. Res.* 91, 61–72. doi: 10.1017/s0016672308009993

Pinhal, D., Yoshimura, T. S., Araki, C. S., and Martins, C. (2011). The 5s rDNA family evolves through concerted and birth-and-death evolution in fish genomes: an example from freshwater stingrays. *BMC Evol. Biol.* 11:151. doi: 10.1186/1471-2148-11-151

Pontes, O., Neves, N., Silva, M., Lewis, M. S., Madlung, A., Comai, L., et al. (2004). Chromosomal locus rearrangements are a rapid response to formation of the allotetraploid arabidopsis suecica genome. *Proc. Natl. Acad. Sci. U.S.A.* 101, 18240–18245. doi: 10.1073/pnas.0407258102

Qin, Q., Cao, L., Wang, Y., Ren, L., Liu, Q., Zhou, Y., et al. (2019). Rapid genomic and genetic changes in the first generation of autotetraploid lineages derived from distant hybridization of *Carassius auratus*, red var. (♀) × *Megalobrama amblycephala*, (♂). *Mar. Biotechnol.* 21, 139–149. doi: 10.1007/s10126-018-9859-8

Qin, Q., He, W., Liu, S., Wang, J., Xiao, J., and Liu, Y. (2010). Analysis of 5S rDNA organization and variation in polyploid hybrids from crosses of different fish subfamilies. *J. Exp. Zool. B Mol. Dev. Evol.* 314, 403–411. doi: 10.1002/jez.b.21346

Qin, Q., Lai, Z., Cao, L., Xiao, Q., Wang, Y., and Liu, S. (2016a). Rapid genomic changes in allopolyploids of *Carassius auratus* red var. *(♀) × Megalobrama amblycephala (♂)*. *Sci. Rep.* 6:34417. doi: 10.1038/srep34417

Qin, Q., Wang, J., Hu, M., Huang, S., and Liu, S. (2016b). Autotriploid origin of *Carassius auratus* as revealed by chromosomal locus analysis. *Sci. China Life Sci.* 59, 622–626. doi: 10.1007/s11427-016-5040-7

Qin, Q., Wang, J., Dai, J., Wang, Y., Liu, Y., and Liu, S. (2015a). Induced All-female autotriploidy in the allotetraploids of *Carassius auratus* red var. (♀) × *Megalobrama amblycephala* (♂). *Mar. Biotechnol.* 17, 604–612. doi: 10.1007/s10126-015-9647-7

Qin, Q., Wang, J., Wang, Y., Liu, Y., and Liu, S. (2015b). Organization and variation analysis of 5S rDNA in gynogenetic offspring of *Carassius auratus* red var. (♀) × *Megalobrama amblycephala* (♂). *BMC Genet.* 16:26. doi: 10.1186/s12863-015-0186-z

Qin, Q., Wang, Y., Wang, J., Dai, J., Liu, Y., and Liu, S. (2014a). Abnormal chromosome behavior during meiosis in the allotetraploid of *Carassius auratus* red var. (♀) × *Megalobrama amblycephala* (♂). *BMC Genet.* 15:95. doi: 10.1186/s12863-014-0095-6

Qin, Q., Wang, Y., Wang, J., Dai, J., Xiao, J., Hu, F., et al. (2014b). The autotetraploid fish derived from hybridization of *Carassius auratus* red var. *(Female) × Megalobrama amblycephala (Male)*. *Biol. Reprod.* 91:93. doi: 10.1095/biolreprod.114.122283

Rocco, L., Costagliola, D., Fiorillo, M., Tinti, F., and Stingo, V. (2005). Molecular and chromosomal analysis of ribosomal cistrons in two cartilaginous fish, *Taeniura lymma* and *Raja montagui* (Chondrichthyes, Batoidea). *Genetica* 123, 245–253. doi: 10.1007/s10709-004-2451-3

Rodrigues, D. S., Rivera, M., and Lourenço, L. B. (2012). Molecular organization and chromosomal localization of 5S rDNA in AmazonianEngystomops (Anura, Leiuperidae). *BMC Genet.* 13:17. doi: 10.1186/1471-2156-13-17

Sajdak, S. L., Reed, K. M., and Phillips, R. B. (1998). Intraindividual and interspecies variation in the 5S rDNA of coregonid fish. *J. Mol. Evol.* 46, 680–688. doi: 10.1007/PL00006348

Sambrook, J., Fritsch, E., and Maniatis, T. (1989). Molecular cloning: a laboratory manual+ cold spring harbor. *Immunology* 49:411.

Vizoso, M., Vierna, J., González-Tizón, A. M., and Martínez-Lage, A. (2011). The 5S rDNA gene family in mollusks: characterization of transcriptional regulatory regions, prediction of secondary structures, and long-term evolution, with special attention to Mytilidae mussels. *J. Hered.* 102, 433–447. doi: 10.1093/jhered/esr046

Wasko, A. P., Martins, C., Wright, J. M., and Galetti, P. M. Jr. (2001). Molecular organization of 5S rDNA in fishes of the genus Brycon. *Genome* 44, 893–902. doi: 10.1139/gen-44-5-893

Wendel, J. F. (2000). Genome evolution in polyploids. *Plant Mol. Biol.* 42, 225–249. doi: 10.1007/978-94-011-4221-2_12

Yi, M. S., Li, Y. Q., Liu, J. D., Zhou, L., Yu, Q. X., and Gui, J. F. (2003). Molecular cytogenetic detection of paternal chromosome fragments in allogynogenetic gibel carp,carassius auratus gibeliobloch. *Chromosom. Res.* 11, 665–671.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 24 May 2019
doi: 10.3389/fgene.2019.00443

Dicyemida and Orthonectida: Two Stories of Body Plan Simplification

*Oleg A. Zverkov[1], Kirill V. Mikhailov[1,2], Sergey V. Isaev[1,3], Leonid Y. Rusin[1,4], Olga V. Popova[2], Maria D. Logacheva[1,2,5], Alexey A. Penin[1,2], Leonid L. Moroz[6], Yuri V. Panchin[1,2], Vassily A. Lyubetsky[1] and Vladimir V. Aleoshin[1,2]**

[1] Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia, [2] A.N. Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia, [3] Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia, [4] Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia, [5] Skolkovo Institute of Science and Technology, Moscow, Russia, [6] Department of Neuroscience, McKnight Brain Institute, University of Florida, Gainesville, FL, United States

Two enigmatic groups of morphologically simple parasites of invertebrates, the Dicyemida (syn. Rhombozoa) and the Orthonectida, since the 19th century have been usually considered as two classes of the phylum Mesozoa. Early molecular evidence suggested their relationship within the Spiralia (=Lophotrochozoa), however, high rates of dicyemid and orthonectid sequence evolution led to contradicting phylogeny reconstructions. Genomic data for orthonectids revealed that they are highly simplified spiralians and possess a reduced set of genes involved in metazoan development and body patterning. Acquiring genomic data for dicyemids, however, remains a challenge due to complex genome rearrangements including chromatin diminution and generation of extrachromosomal circular DNAs, which are reported to occur during the development of somatic cells. We performed genomic sequencing of one species of *Dicyema*, and obtained transcriptomic data for two *Dicyema* spp. Homeodomain (homeobox) transcription factors, G-protein-coupled receptors, and many other protein families have undergone a massive reduction in dicyemids compared to other animals. There is also apparent reduction of the bilaterian gene complements encoding components of the neuromuscular systems. We constructed and analyzed a large dataset of predicted orthologous proteins from three species of *Dicyema* and a set of spiralian animals including the newly sequenced genome of the orthonectid *Intoshia linei*. Bayesian analyses recovered the orthonectid lineage within the Annelida. In contrast, dicyemids form a separate clade with weak affinity to the Rouphozoa (Platyhelminthes plus Gastrotricha) or (Entoprocta plus Cycliophora) suggesting that the historically proposed Mesozoa is a polyphyletic taxon. Thus, dramatic simplification of body plans in dicyemids and orthonectids, as well as their intricate life cycles that combine metagenesis and heterogony, evolved independently in these two lineages.

Keywords: Mesozoa, Dicyemida, Orthonectida, genome, mitochondrial DNA, phylogeny

OPEN ACCESS

Edited by:
Ancha Baranova, George Mason University, United States

Reviewed by:
Andreas Hejnol, University of Bergen, Norway
Denis Baurain, University of Liège, Belgium
Max Telford, University College London, United Kingdom

***Correspondence:**
Vladimir V. Aleoshin
aleshin@genebee.msu.su

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *03 November 2018*
Accepted: *29 April 2019*
Published: *24 May 2019*

Citation:
Zverkov OA, Mikhailov KV, Isaev SV, Rusin LY, Popova OV, Logacheva MD, Penin AA, Moroz LL, Panchin YV, Lyubetsky VA and Aleoshin VV (2019) Dicyemida and Orthonectida: Two Stories of Body Plan Simplification. Front. Genet. 10:443. doi: 10.3389/fgene.2019.00443

INTRODUCTION

In spite of more than one hundred years of studies, the evolutionary relationships of the Mesozoa are still elusive. The name of this taxon reflects the traditional view of mesozoans as organisms with intermediate organization between unicellular protozoans and multicellular metazoans (Van Beneden, 1876; Hyman, 1940). Indeed, the two groups of microscopic parasitic invertebrates,

the Dicyemida, and Orthonectida, display a remarkably simple morphological organization and a nearly complete absence of tissues and organs (Malakhov, 1990). Adult dicyemids inhabit the renal sacs of cephalopod mollusks and consist of just about 40 somatic cells, lack recognized muscular, nervous, sensory cells, and the organs typical for eumetazoans (Furuya et al., 2004). Dicyemids do not have a morphologically recognized basal membrane (Czaker, 2000), and never develop "true" tissues throughout their complex life cycle (Furuya and Tsuneki, 2003). The trophic stage of orthonectids is a syncytial plasmodium, which resides inside the invertebrate host and generates ephemeral ciliated organisms that exit the host for reproduction (Slyusarev, 2008). These organisms are composed of several hundred somatic cells without anatomically recognized digestive, circulatory, or excretory systems. Before the discovery of muscular and nervous systems in the swimming stages of orthonectids (Slyusarev and Starunov, 2015), they were thought to have a planula-like organization and were grouped with dicyemids in the Mesozoa as multicellular animals with an incredibly simple body plan, perhaps – the simplest among all Metazoa, and comparable to placozoans.

Intricate life cycles of dicyemids and orthonectids exhibit the alternation of asexual and sexual generations, termed metagenesis. Ameiotic generative cells (agametes) develop inside the dicyemid axial cell and later produce the next vermiform generation possessing gametic cells that undergo self-fertilization. In orthonectids, agametes develop inside the parasitic plasmodium and produce the free-living diecious (or hermaphroditic) generation (Cheng, 1986; Slyusarev, 2008). The phenomenon of successive sexual parthenogenetic and amphimictic generations is termed heterogony. In this sense, orthonectids and dicyemids as well as parasitic flatworms combine metagenesis and heterogony in their life cycles. Particularly, trematode sporocysts and rediae that parasitize gastropod mollusks produce the next generation from ameiotic generative cells (Dobrovolskij and Ataev, 2003; Ataev, 2017). Similarities in life cycles for long sustained the hypothesis about close relationships of dicyemids and orthonectids with digenetic trematodes. On the other hand, intracellular localization of generative cells relates dicyemids and orthonectids with myxozoans rather than trematodes. Such intricate combination of traits makes life strategies in dicyemids and orthonectids unique among animals.

The phylogenetic affinity of dicyemids and orthonectids has been called into question on the grounds of morphology (Kozloff, 1990; Brusca and Brusca, 2003; Ruppert et al., 2004). Molecular data conclusively demonstrated that both dicyemids and orthonectids are in fact bilaterians (Katayama et al., 1995; Hanelt et al., 1996; Pawlowski et al., 1996; Aruga et al., 2007) and belong to the diverse clade of Lophotrochozoa (=Spiralia) (Kobayashi et al., 1999, 2009; Petrov et al., 2010; Suzuki et al., 2010; Mikhailov et al., 2016; Lu et al., 2017; Schiffer et al., 2018), thus implying that their simple organization evolved as the result of their parasitic lifestyle.

In molecular phylogenetic analyses, dicyemid and orthonectid lineages display extremely high levels of divergence, and their exact placement among the spiralians remains ambiguous and potentially prone to long branch attraction artifacts. Complicating the matter is the uncertainty in relationships between other spiralian taxa, including the Annelida, Mollusca, Nemertea, Brachiopoda, Entoprocta, and Bryozoa (Kocot, 2016). Recent phylogenomic analyses lead to conflicting conclusions regarding the mesozoan phylogeny. Lu et al. (2017) using a dataset of 348 orthologs (58,124 alignment positions) from 23 spiralian species, including an orthonectid and a dicyemid, report the monophyly of the Mesozoa either as a sister group to the Rouphozoa (Platyhelminthes + Gastrotricha) or within the Gastrotricha. Alternatively, Schiffer et al. (2018) using a dataset of 469 orthologs (190,027 alignment positions) from 29 spiralian species, including an orthonectid and two dicyemids, conclude that Orthonectida and Dicyemida evolved independently within the Lophotrochozoa, with the orthonectids exhibiting clear affinity to annelids, and dicyemids occupying an isolated position within Lophotrochozoa. Here, we obtained transcriptomic and genomic data for dicyemid species to resolve this contradiction.

The dicyemid genome is distinguished by uncommon features, such as the genome rearrangements during the life cycle and generation of circular DNAs (Noto et al., 2003), including those that encode mitochondrial proteins and rRNAs (Watanabe et al., 1999; Catalano et al., 2015). It is not yet established if the mitochondrial protein-coding genes are encoded only by small circular DNA molecules (Watanabe et al., 1999) or whether they are produced during the dicyemid development from a precursor mitochondrial DNA with a more typical metazoan organization (Awata et al., 2006). Using high-throughput genomic sequencing we sought to find any properties of dicyemid sequences that would reveal their genome organization. We also estimated the extent of gene losses due to the simplification of dicyemid morphological organization, and analyzed whether losses in particular gene families and regulatory pathways are the same or different compared to an orthonectid *Intoshia linei*.

RESULTS AND DISCUSSION

Genomic Sequencing and Assembly of *Dicyema* sp.

Direct assembly of a dicyemid genome from whole DNA extracts using standard approaches is an extremelly challenging problem due to drastic genome rearrangements that occur in dicyemids during development. Previous studies have demonstrated that somatic cells of dicyemids undergo drastic genome rearrangements and chromatin elimination (Noto et al., 2003), and suggested that selective and whole genome amplification takes place at different stages of their development (Awata et al., 2006). Accordingly, the sequencing of whole DNA extracts from *Dicyema* sp. resulted in a highly fragmented assembly with uneven coverage and N50 of 942 bp, where the largest contig was only around 20 Kb. The total size of the assembly is 858 Mbp in nearly 1 million contigs over the length of 500 bp, and includes contaminating cephalopod sequences. Due to significant genetic difference between the dicyemid host *Enteroctopus dofleini* and the available genomic sequence of *Octopus bimaculoides*, the filtering of the assembly

was performed at the level of predicted gene products. Only predictions identifiable by hits against the InterPro database were retained for the subsequent comparative analyses and filtered from the cephalopod contamination using the best hit approach with BLAST searches against the NCBI nr database. Out of 38,410 predictions with InterPro hits, 43% were discarded as contamination, resulting in 21,842 putative dicyemid genes with 71% complete and 12% fragmented universal eukaryotic orthologs evaluated by BUSCO (**Table 1**). Similar values are obtained for gene predictions after normalizing on the number of BUSCOs found in at least one filtered transcriptome: 76% complete and 12% fragmented. The total percentage of BUSCOs recovered by at least one sequencing library, including genomic and transcriptomic filtered data, approaches values seen in typical metazoan genomes: 91% complete and 3% fragmented. For all analyses in Sections 2.4–2.11 we used original genomic data on *Dicyema* sp., and the three transcriptomes, including the two originally obtained and the one of *Dicyema japonicum* available from the published source (Lu et al., 2017).

The dicyemid genes display miniaturization of spliceosomal introns – the median length of introns is 27 bp, and approximately two thirds of predicted introns are under the length of 30 bp (**Figure 1**). This agrees with an earlier survey that revealed extreme intron shortening in a set of 40 genes from *D. japonicum* (Ogino et al., 2010). The estimated intron density in *Dicyema* sp. is 4.9 introns/gene for predictions with intact start and stop codons, which is also similar to the 5.3 introns/gene reported for *D. japonicum*. Similar value of intron density is seen in the genome of orthonectid *I. linei* (Mikhailov et al., 2016). Notably, the orthonectid genes also harbor short spliceosomal introns, but the majority of its introns are longer than 30 bp, and the median size is 57 bp, considerably exceeding the intron lengths observed in dicyemid genes.

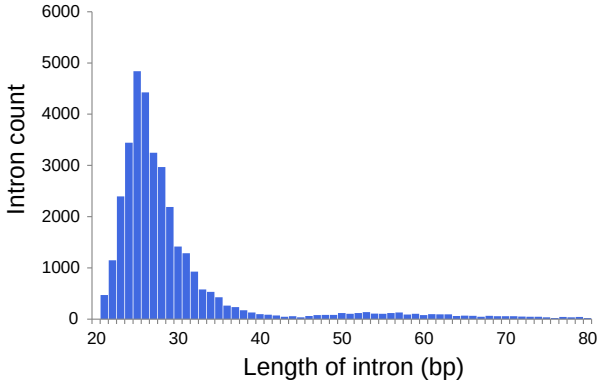


FIGURE 1 | Distribution of intron lengths in predicted genes of *Dicyema* sp.

"Circular" Contigs in Genomic Assembly of *Dicyema* sp.

Using the genomic assembly we have identified 24,065 "circular" contigs (see section "Materials and Methods"). The distribution of circular contig lengths in the assembly is multimodal (**Figure 2A**). The first abundant pool of sequences is formed from contigs less than 500 bp. The second pool, which includes sequences of a length over 500 bp, consists of 3,220 contigs with the median length of 702 bp. The properties of the sequences in this pool (such as length and abundance) are consistent with previous data of DNA gel electrophoresis, EM and PCR experiments (Noto et al., 2003), which supports the conjecture that these sequences are circular DNA rather than direct repeats. "Short" circles (up to 500 bp length) were shown to possess 38.1% low complexity regions, while "long" circles – only 2.9%. This observation might suggest that a fraction of predicted short circles represents direct repeats. Following this rationale, we considered the two sub-pools separately in analyses.

The lengths of sequences from the second pool of circular contigs are distributed non-uniformly which is particularly evident within the 600–800 bp range (**Figure 2B**). The average distance between two adjacent peaks of this distribution is 10.44, which closely corresponds to the number of base pairs in one turn of B-DNA. Multimodal distribution was also observed

TABLE 1 | Assembly statistics.

	Dicyema sp. genomic	*Dicyema* sp. 454	*Dicyema* sp.	*Dicyema japonicum*	*Dicyema* sp.genomic filtered*	*Dicyema* sp. 454 filtered**	*Dicyema* sp. filtered**	*Dicyema japonicum* filtered**
Assembly size (bp)	858,248,066	19,669,371	64,598,656	44,413,963	N/A	N/A	N/A	N/A
Contigs/transcripts (>500 bp)	939,453	22,115	52,176	29,091	N/A	N/A	N/A	N/A
Predicted genes/peptides	984,055	12,379	22,286	11,330	21,842	11,726	21,656	11,233
Complete BUSCOs, eukaryota_odb9	77.6%	65.0%	82.2%	85.1%	71.3%	62.0%	80.2%	84.2%
Complete and single-copy BUSCOs (S)	74.6%	63.0%	77.9%	82.5%	68.3%	60.7%	76.6%	81.2%
Complete and duplicated BUSCOs (D)	3.0%	2.0%	4.3%	2.6%	3.0%	1.3%	3.6%	3.0%
Fragmented BUSCOs (F)	14.2%	22.8%	9.2%	6.6%	11.6%	23.1%	10.2%	6.6%
Missing BUSCOs (M)	8.2%	12.2%	8.6%	8.3%	17.1%	14.9%	9.6%	9.2%

**Genomic predictions were filtered by retaining only hits to the InterPro database and cleaned from the cephalopod contamination with BLAST searches against the NCBI nr database.*

***Transcriptome assemblies were filtered with BLAST searches against the RefSeq database as detailed in Materials and Methods, Section "Assembly and Filtering of Dicyemid Sequences."*

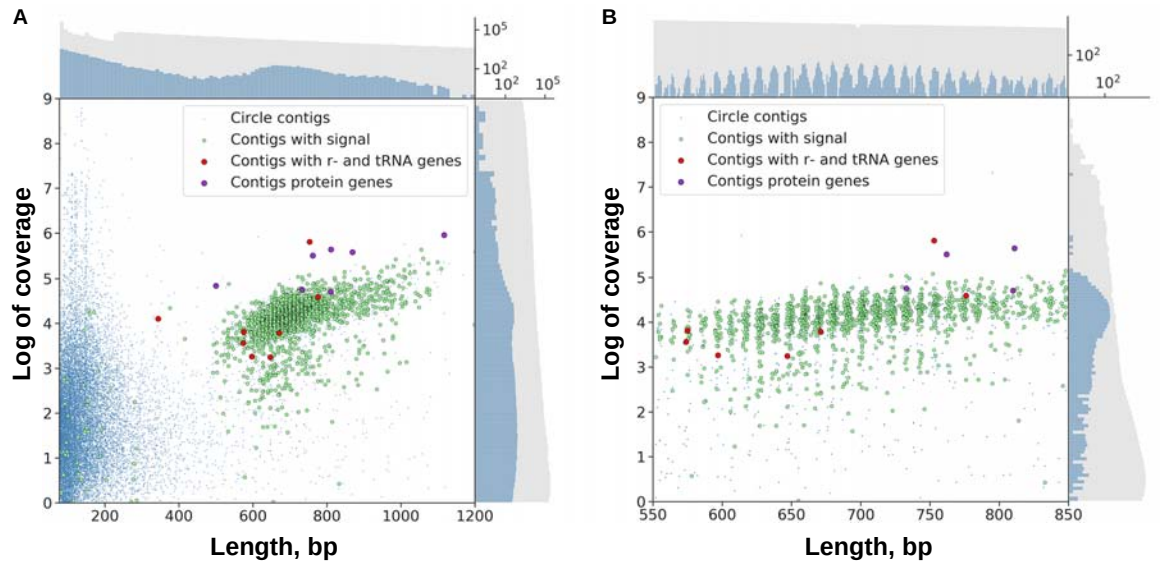


FIGURE 2 | Scatter plot of circular contigs in the *Dicyema* sp. genome. Different markers stand for circles with the specific signal and with mitochondrial genes (see Text for clarification). **(A)** The distributions for both length and coverage logarithm are non-uniform — there is a pool of "long" contigs with a high level of coverage in the set. **(B)** The distribution of "long" contigs with a high level of coverage is, in turn, non-uniform; the distance between two adjacent peaks is approximately 10.44 bp. On both scatterplots the upper axis shows the distribution of contig lengths, and the right axis shows the coverage. Blue dots correspond to circular contigs, and green dots are linear contigs. The scale is logarithmic. Both the coverage and length of circular contigs is obviously less uniform compared to those of linear contigs.

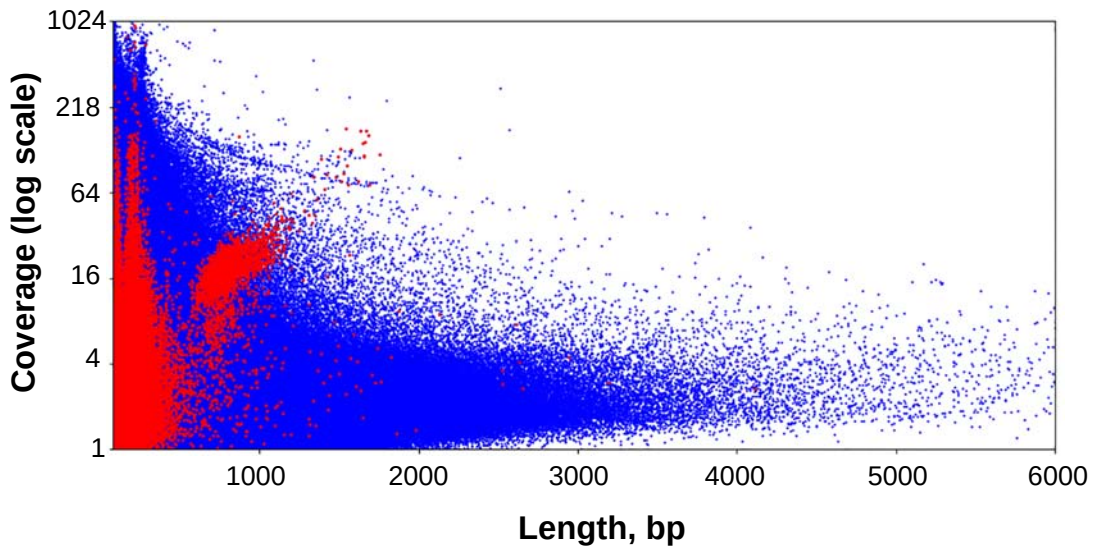


FIGURE 3 | Scatter plot of the entire set of *Dicyema* sp. genome assembly contigs. Blue dots are linear contigs, and red dots are circular ones.

(Kolmogorov–Smirnov test p-value is 0.999) when performing assembly with the varying k-mer size (55 or 77) and with another assembly method (**Supplementary Figure S1**). The presence of this pattern is unexpected, and presumably could be attributed to the greater stability of circles or tendency to circularize for molecules with an integer amount of turns of a relaxed form of DNA. A similar effect has also been observed in short (<200 bp) sequences as a result of rolling circle replication bias (Joffroy et al., 2018). This distribution can result from the random ligation of linear molecules cut from the genome as it leads to the reduction in DNA supercoiling. Alternatively, replicating mini-circular DNA molecules can be selected in length to reduce their supercoiling. **Figure 3** shows that the coverage value for "long" circular contigs is not lower than for linear ones, which casts doubt on the proposed diminution of circular molecules during ontogenesis.

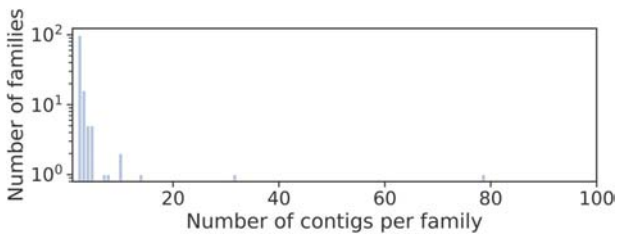


FIGURE 4 | Distribution of families of long circular contigs in *Dicyema* sp. vs. the number of members per family. To detect families, *blastn* searches of all long (>500 bp) circular contigs against themselves were conducted. The graph elements are contigs, and the edge weight is hits *E*-value. All edges with the weight greater than 1e-10 were deleted, and contigs in one connected component are considered a family.

Long circular contigs are predominantly not similar in nucleotide sequences. Only 15% of them have at least one fairly similar contig, and only three families of contigs unite more than 10 members (**Figure 4**).

Two independent motif detection methods (Bailey and Elkan, 1994; Rubanov et al., 2016) have been applied to the circles of length 600–800 bp with a coverage logarithm of over 3 (2,031 sequences). In 1,871 sequences (92.12% of sequences in the analysis) common motifs have been found (*E*-value: 4.8e-82, see **Figure 5A**).

At a p-value $< 10^{-5}$, the most common motif occurs on average once every 874 bp in "long" circles and every 21,863 bp throughout the entire assembly (statistical significance of the difference provided by the chi-squared criterion: p-value < 0.001). The search for highly conserved sequences in various subsets of genome sequences has demonstrated that less common motifs with high information content can also be found in circles (**Figures 5B–E**).

The search for conserved domains in circular contigs recovered only domains of mtDNA-encoded proteins (10 conserved domains, 13 contigs including paralogs). These sequences are presumably transcribed as they are also found in the RNA-seq data (*blastn* search, *E*-value < 1e-30).

Mitochondrial DNA of *Dicyema* sp.

Genomic data confirm the localization of mitochondrial genes of dicyemids on circular DNA molecules (Watanabe et al., 1999; Catalano et al., 2015). The search for mtDNA genes in the genomic data found 21 circles with length varying from 344 to 1605 bp. The following gene sequences were found: *cox1-3*, *cob*, *nad1-5*, *atp6*, *rrnL*, *rrnS*, *trnH*, *trnI*, *trnK*, *trnL1*, *trnN*, *trnP*, *trnQ*, *trnR*, *trnS2*, and *trnY* (**Figure 6**). In earlier studies the dicyemid mitochondrial contigs were found to carry either one protein coding gene (Watanabe et al., 1999) or a protein coding gene and a tRNA gene (Robertson et al., 2018). We found one circle that contains two genes – *cox2* and *rrnS*, and three circles that contain two tRNA genes each. Protein identity between mitochondrial predictions for *Dicyema* sp. and the earlier published *D. japonicum* (Robertson et al., 2018) varies from 39% (nad2) to 75% (cox1). The majority of mitochondrial genes can also be found in the transcriptomic data, except for *atp6* and *nad5*. The mtDNA circles also contain the motif described above (**Figure 5A**) (p-value $< 10^{-5}$).

The *nad2* and *atp6* genes were found in two different variants in the genomic data. Two paralogs of nad2 with lengths of 215 and 252 amino acids have 42% identity at the amino acid level. Two paralogs of atp6 with lengths of 117 and 149 amino acids have 89% identity at the amino acid level, and share two long deletions with other dicyemids. These deletions are specific for dicyemids and are not found in other taxa including Orthonectida. Both of dicyemid deletions are located outside of the transmembrane helices – the first one with the length of 16 amino acids is located in the region facing the mitochondrial matrix and the second one with the length of 17 amino acids is located in the region facing the intermembrane space, according to the alignment of atp6.

We predicted 11 mitochondrial tRNA genes in *Dicyema* sp. including two paralogs of glutamine tRNA gene (**Supplementary Figure S2**). Both dicyemid glutamine tRNAs have similar secondary structures and lack a T-arm. Dicyemid arginine tRNA also lacks a T-arm and lysine tRNA lacks a D-arm. Other mitochondrial tRNAs maintain the typical clover leaf structure, although several tRNA genes have single nucleotide insertions and/or non-complementary pairs in stems. Experimental evidence is needed to confirm all the predicted tRNA genes, as well as decisions whether numerous not listed tRNA-like sequences with p-value below the threshold are functional genes.

Read mapping to the genomic assembly revealed no read pairs that would facilitate mtDNA scaffolding. Whenever one read from a pair would map to the circular mitochondrial contig, the other would map to the same contig or have a sequence of low complexity. Thus, our genomic data fails to confirm the hypothesized existence of an unprocessed mtDNA precursor, which would generate the mtDNA circles (Awata et al., 2006).

The presence of common sequence motifs in circles with mtDNA genes and without them seems to be surprising. It can be interpreted as a consequence of a similar mechanism of generation and maintaining of circles irrespective of their function.

The partitioning of mtDNA into circular molecules is a rare feature for the animal mitochondrial genomes (Odintsova and Yurina, 2005; Burger et al., 2012; Kolesnikov and Gerasimov, 2012; Smith and Keeling, 2015; Lavrov and Pett, 2016, for review). In bilaterians, the mtDNA is fragmented into a large number of mini-chromosomes in the cyst-forming nematodes *Globodera* spp. (Armstrong et al., 2000; Gibson et al., 2007) and sucking lice (Shao et al., 2009). Notably, the mitochondrial DNAs from orthonectids *Intoshia linei*, *Intoshia variabili*, and *Rhopalura ophiocomae* retain typical structure for metazoans and encode the full set of mitochondrial genes on a single circular molecule (Robertson et al., 2018; Bondarenko et al., 2019). The reason why the mitochondrial *Dicyema* spp. genome is fragmented is unknown. Earlier, the fragmentation of the mitochondrial genome of sucking lice was considered (Shao et al., 2009) as an adaptation to the high rate of molecular evolution, which is

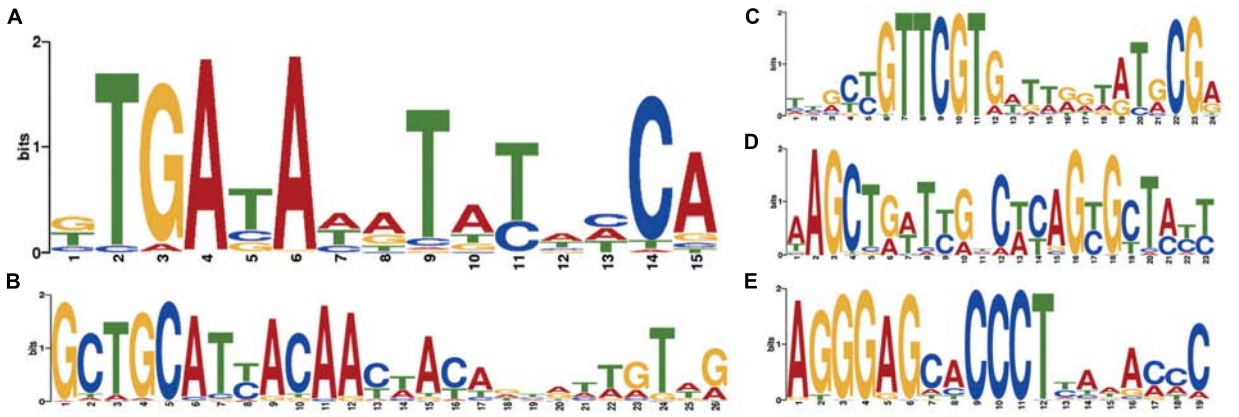


FIGURE 5 | Motifs specific to *Dicyema* sp. circular contigs. **(A)** Signal found in the majority of contigs. **(B–E)** Highly conserved signals found in a group of circular contigs of high length and coverage; **(B)** found in 541 contigs; **(C)** found in 284 contigs; **(D)** found in 234 contigs; **(E)** found in 438 contigs. All counts are provided at a p-value $< 10^{-4}$.

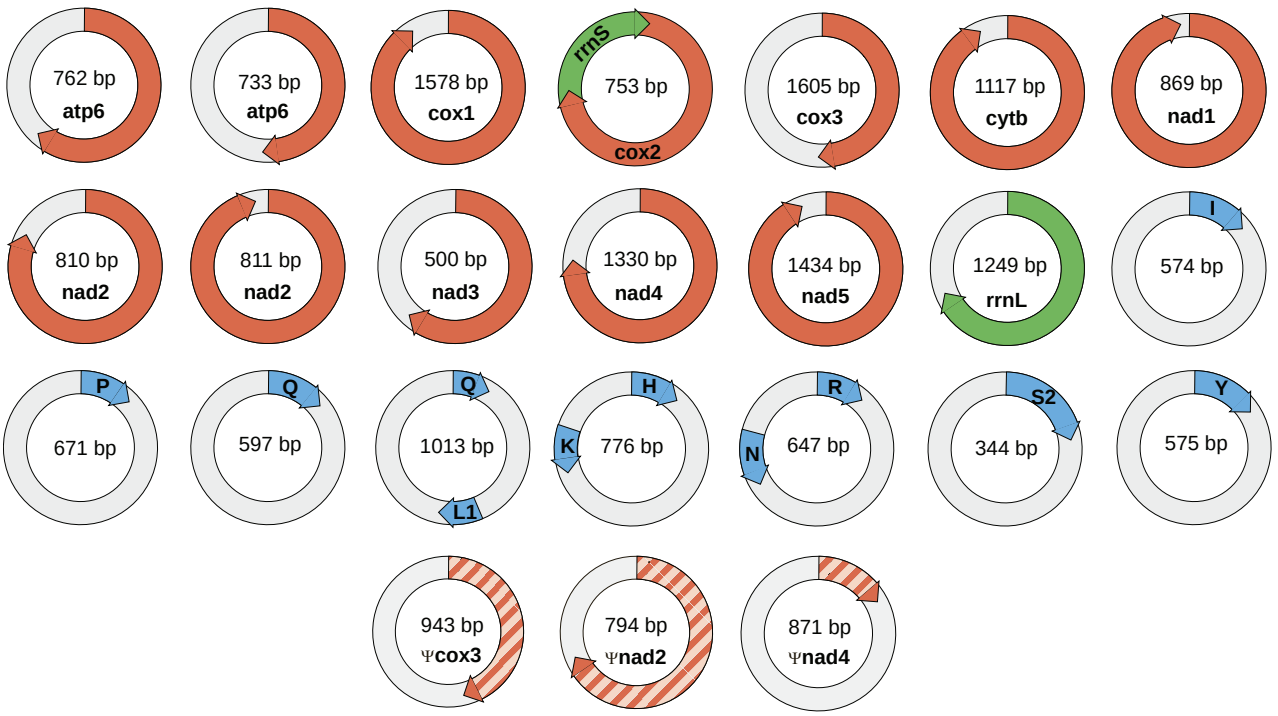


FIGURE 6 | Mini-circles coding mitochondrial genes in *Dicyema* sp. Lenght of each minicircle is marked in its center. Each minicircle has a coding region with gene name and transcription orientation, and a non-coding region (in light gray). Protein-coding genes marked in red, rRNA genes marked in green, and tRNA genes marked in blue. tRNA genes indicated by a one-letter amino acid code.

even more characteristic of *Dicyema* spp. It is possible that under conditions of high mutagenesis, a set of uncorrupted genes is easier to assemble from individual than concatenated molecules.

Analysis of mitochondrial DNA suggests an explanation of the multiple observed circular contigs. For searches with the *tblastx* algorithm we used proteins from the annotated mitochondrial contigs as the query and all 3,220 "long" circle contigs as the database. The searches returned many circular contigs that encode highly diverged genes *cox3*, *nad2*, and *nad4* (**Figure 7**). The *cox3* homolog is largely diverged, while *nad2* and *nad4* contain stop codons and frame shifts. These contigs therefore represent mitochondrial pseudogenes.

Previous publications and our new data confirm the presence of two unusual features of the *Dicyema* genome. First, mitochondrial genes in *Dicyema* are not located on a single long DNA molecule as in most animals, but are partitioned

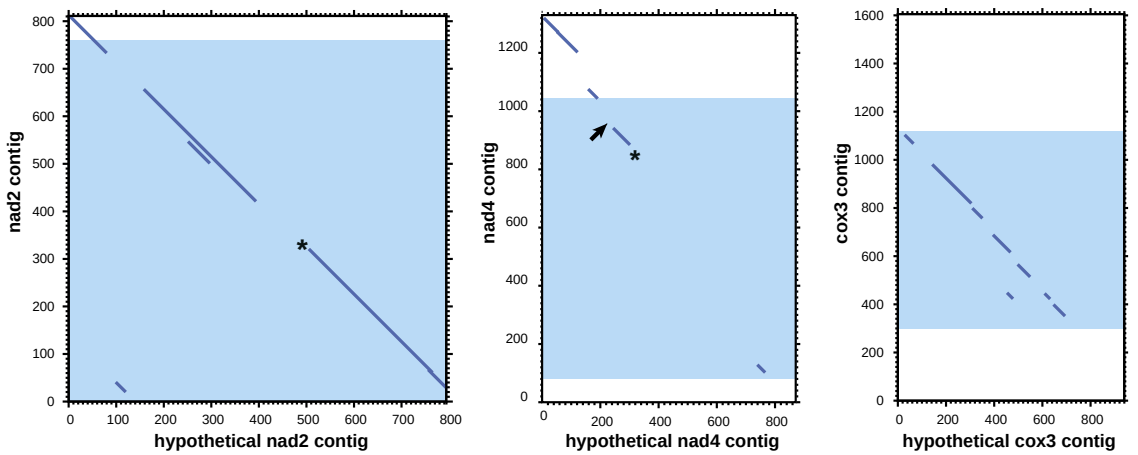


FIGURE 7 | Dotplots of local alignments of circular contigs against genes *nad2*, *nad4*, and *cox3*. Blue zones represent mitochondrial genes with the open frame. Asterisks mark homologs with stop codons, arrows mark frameshifts. Results obtained with LAST (Kiełbasa et al., 2011) (threshold *E*-value for local alignment 0.008).

into smaller circular molecules. The second interesting feature of *Dicyema* is the presence of thousands of non-coding circular DNA sequences. Both types of circular DNA molecules fall in a similar range of size and coverage in DNA assembly and bear a common set of similar 12–20 bp DNA patterns, which might be hypothetical signal sequences. We assume that all circular DNAs in *Dicyema* may have a common origin, although experimental evidence is necessary. We speculate that the presence of multiple mtDNA mini-rings instead of one long molecule might have produced serious problems in mitochondrial division. This requires special mechanisms to correct distribution of multiple minicircular DNA molecules upon mitochondrion division so that both descendants would obtain a complete set of genes. Specific signal patterns like the ones we observe could be used to support circular mtDNA duplication, their protection against elimination or their correct distribution between descendent mitochondria. When such mechanisms are established, it is possible that rings carrying mutated (pseudo)genes or other selfish non-coding DNA circular elements acquire similar signal sequences that ensure their preservation in a similar way as with parasitic mobile genetic elements.

Homeobox Transcription Factors

Homeodomain (homeobox) transcription factors are crucial regulators of animal development that play central roles in tissue differentiation and axial body patterning. Bilaterian genomes encode from over 300 to around 60 homeobox genes. The genome of orthonectid *I. linei* was found to possess one of the smallest repertoires of homeoboxes (Mikhailov et al., 2016), which matches the reduced complexity of their organization. To determine how the extreme simplification of body plan seen in dicyemids relates to their homeobox gene content we searched for these genes in the dicyemid genomic and transcriptomic data. For analyses with HMMER, we used gene predictions coming from the genomic assembly of *Dicyema* sp. (PRJNA527259; designated as *Dicyema* sp. 1) and transcriptome assemblies of *Dicyema* sp. (SRR827581; designated as *Dicyema* sp. 2) and *Dicyema japonicum* (DRR057371). HMMER searches using the homeodomain profile identified 38, 39, and 55 homeoboxes in the three dicyemids after filtering out contaminating cephalopod sequences. Phylogenetic inference suggests that dicyemid homeoboxes form up to 39 families, and each dicyemid was found to contain 31 or 34 families (**Table 2**). A high level of sequence divergence complicates classification of the dicyemid homeoboxes. Although most dicyemid sequences could be assigned to one of the homeobox classes (Zhong and Holland, 2011), their attribution to known families is inconclusive.

Phylogenetic inference with PRD class homeoboxes reveals six dicyemid families, including the previously identified orthologs of the Pax6 and Otx (Aruga et al., 2007; Kobayashi et al., 2009). Five dicyemid sequence groups were found among the LIM homeoboxes, two of which are grouped with the Lhx6/8 and Islet family sequences. An additional LIM class homeobox of the Lhx2/9 family was found in *D. japonicum*, but could neither be confirmed by data from the other dicyemids nor discarded as contamination. Another five dicyemid gene groups belong to the POU class homeoboxes, but branch outside of any known families. The dicyemids form at least 5 TALE class sequence groups, with one group branching within the Pbx family. An additional single member of the TALE Tgif family was found only in the genomic data. Four SINE class families were found among the dicyemid sequences, with one grouping with the Six3/6 family. The dicyemids also possess a zinc finger homeobox, and a group of Onecut family sequences, which can be subdivided into 2 dicyemid-specific families.

Reconstructions with the ANTP class homeoboxes recover 8 dicyemid sequence groups (**Figure 8**). Three of these groups fall within the central Hox sequences. One of the dicyemid central Hox groups corresponds to orthologs of DoxC – a dicyemid member of the spiralian Lox5 family (Kobayashi et al., 1999), which was shown to have an expression pattern consistent with defining anterior–posterior boundaries in the developing dicyemids (Aruga et al., 2007). The analysis suggests

TABLE 2 | The list of homeodomain transcription factors in three species of Dicyemida.

	Dicyema sp. 1	*Dicyema* sp. 2	*Dicyema japonicum*
Class ANTP	9	13	10
Subclass HOXL	5	8	7
Family Hox6-8 or 'central' Hox genes	3	3	3
Dicyemid 'central' Hox group 1 (DoxC)	1	1	1
Dicyemid 'central' Hox group 2 (DoxC paralog)	1	1	1
Dicyemid 'central' Hox group 3	1	1	1
Family Hox9-13(15) or 'posterior' Hox genes	0	0	1
Dicyemid HOXL group	1	4	2
Family Evx (even-skipped)	1	1	1
Subclass NKL	4	5	2
Family Dlx (distal-less)	1	1	0
Nk2 genes (families Nk2.1 and Nk2.2)	3	4	2
Dicyemid Nk2 group 1	1	2	1
Dicyemid Nk2 group 2	2	2	1
Other ANTP	0	0	1
Class PRD	5	6	6
Family Pax 4/6	1	1	1
Family Otx (orthodenticle)	1	1	1
Dicyemid PRD group 1	1	3	1
Dicyemid PRD group 2	1	0	1
Dicyemid PRD group 3	0	1	1
Dicyemid PRD group 4	1	0	1
Class POU	5	7	1
Dicyemid POU group 1	1	1	0
Dicyemid POU group 2	1	1	1
Dicyemid POU group 3	1	1	0
dicyemid POU group 4	1	1	0
Dicyemid POU group 5	1	3	0
Class LIM	4	9	4
Family Lhx6/8	1	2	0
Family Lhx2/9	0	0	1
Family Isl	1	1	2
Dicyemid LIM group 1	0	1	1
Dicyemid LIM group 2	1	4	0
Dicyemid LIM group 3	1	1	0
Class SINE	3	5	5
Family Six3/6	1	1	1
Dicyemid SINE group 1	1	1	2
Dicyemid SINE group 2	1	1	1
Dicyemid SINE group 3	0	2	1
Class CUT	2	2	2
Family Onecut	2	2	2
Dicyemid Onecut group 1	1	1	1
Dicyemid Onecut group 2	1	1	1
Class TALE	8	7	9
Family Pbx	1	1	1
Family Tgif	1	0	0
Dicyemid TALE group 1	3	1	2
Dicyemid TALE group 2	1	1	2
Dicyemid TALE group 3	1	2	1
Dicyemid TALE group 4	1	2	3
Class ZF	2	2	1
Family Zfhx	2	2	1

that dicyemids possess another member of the Lox5 family – all three dicyemids were found to encode a paralog of the DoxC. The paralog displays greater sequence divergence, but similar to other members of the family retains a Lox5-specific motif flanking the C-terminus of the homeodomain (de Rosa et al., 1999). The third dicyemid group within the central Hox sequences is found outside the Lox5 family and tends to group with the Lox2/Lox4 families, but beyond that does not lend itself to classification. A single posterior type Hox gene was found in *D. japonicum*, but once again could not be verified using other dicyemid sequences or rejected as contamination by BLAST searches. The dicyemid ANTP class homeoboxes also include members of the Evx, Dlx, and Nk2 families, and a conspicuous group of Hox-like sequences (**Figure 8**). Sequences within the dicyemid Hox-like assemblage share a common ancestor and retain a YPWM motif, which is essential for binding Hox cofactors (Prince et al., 2008), but this group is too divergent to be classified with any family, and is placed with the longest branch of Hox-like genes – the ParaHox Cdx family.

The survey of dicyemid genes suggests that overall they possess fewer homeoboxes than the orthonectid *I. linei* and their sequences are also markedly more diverged. Unlike the orthonectid, no ParaHox or anterior Hox families could be readily identified in the dicyemid data. Reduction of homeobox transcription factors in dicyemids is consistent with extreme simplification of their body plan. Unexpectedly, the dicyemids also experience several lineage-specific expansions of homeoboxes, notably the duplication of central Hox gene DoxC, which opposes the general trend of regulatory gene loss.

Basement Membrane

The basement membrane is a structure that enables the compartmentalization of cells to form tissues and organs. It is present in the majority of metazoans, with exception of sponges, placozoans, and acoelomorphs. The reported loss of a morphologically recognized basement membrane in dicyemids would indicate unprecedented simplification in this animal group. Even though this topic has been studied (Czaker, 2000), it is still unclear whether dicyemids have a basement membrane during any of their life cycle stages. The basement membrane consists of a set of "basement membrane toolkit" proteins, but the most important are collagen IV and laminin (Fidler et al., 2017). Both laminin and type IV collagen are multi-domain proteins that include specific domains (LamNT for

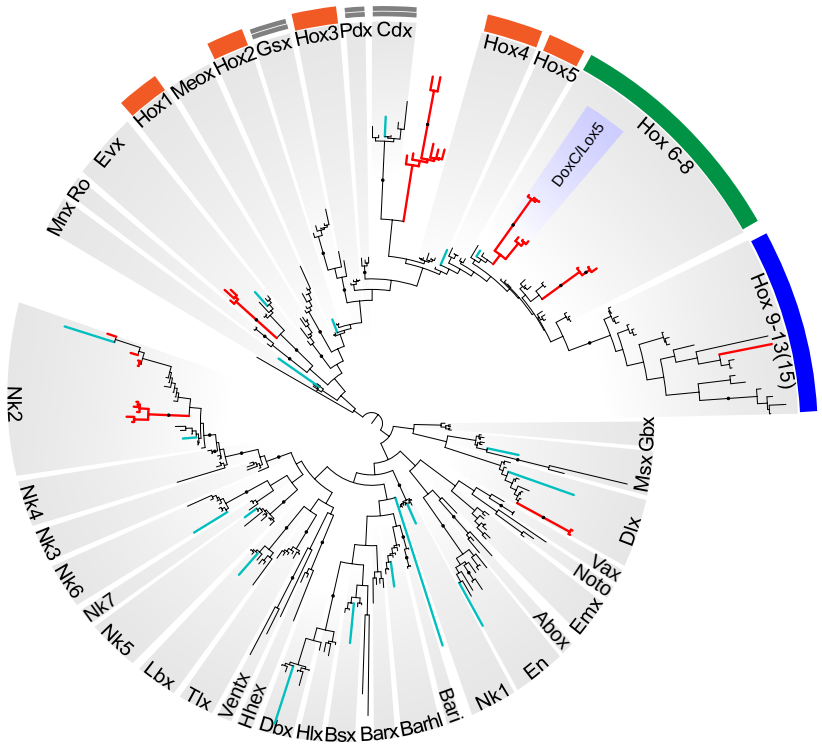


FIGURE 8 | Bayesian tree of the ANTP class homeodomain sequences from *Homo sapiens*, *Drosophila melanogaster*, *Capitella teleta*, *Octopus bimaculoides*, *Intoshia linei*, and three dicyemids: *Dicyema* sp. 1, *Dicyema* sp. 2, and *Dicyema japonicum*. The dicyemid sequences are given in red, and the orthonectid homeoboxes are labeled with teal color. The groups of anterior Hox genes (Hox1-5) are outlined in orange, the central Hox genes (Hox6-8) – in green, and the posterior Hox genes (Hox9-13) – in blue; ParaHox orthologs (Gsx, Pdx, and Cdx) are marked with a double line. The dicyemid DoxC/Lox5 genes are labeled inside the group of central Hox genes. Nodes with ≥0.95 posterior probability are marked with black dots.

laminin and C4 in the case of collagen type IV) and non-specific domains (EGF-like and other). The BLAST and Pfam searches showed that these domains of canonical molecules forming basement membrane are absent from the sequenced genome of *Dicyema* sp., therefore supporting the proposed secondary loss of this trait in dicyemids. The apparent absence of the recognized basement membrane is parallel with a reported loss of muscular and nervous systems in these animals. Indeed, in bilaterians, the basement membrane supports the maintainance of the muscular and nervous system architecture, their development and compartmentalization, and supporting growth factor signaling gradients among other functions.

The complete life cycle of dicyemids is not entirely understood, and more complex structures of transitional obscure life forms of these organisms are not excluded. An unknown stage can potentially exist between the infusorioform larvae that exits the host and the vermiform embryos found in cephalopods. The lack or reduced representation of genes encoding key elements of the basement membrane or other mediators of organ formation further supports the idea that dicyemids are secondarily simplified to an outstanding state.

Membrane Receptor Proteins

Cell surface membrane receptors act in cell signaling and allow communication between the cell and the extracellular space. Their diversity reflects the complexity of the organism and its ability to respond to different external signals. The number of genes encoding receptor proteins in dicyemids is exceptionally low. We found only two PF00001 domain hits corresponding to the 7 pass transmembrane receptor proteins of rhodopsin family in *Dicyema* sp. This family of G-protein-coupled receptors (GPCRs) is ubiquitously present and abundant in metazoans and contains tens to hundreds of members in different species. The minimum number of the rhodopsin family genes (six per genome) is detected in the sponge *Amphimedon queenslandica*; even the genome of the simplified orthonectid *I. linei* contains 32 genes of the rhodopsin family. The actual specificity of these GPCRs proteins is unknown, although their BLAST search shows best similarity to the rhodopsin family neuropeptide receptors from other animals. Four proteins from another GPCR 7 pass transmembrane receptor family – secretin family (PF00002) were predicted in the *Dicyema* sp. data. This is fewer than in most metazoans yet some flatworms have even fewer (Zamanian et al., 2011), and the Orthonectida have no such proteins. We found one putative metabotropic glutamate receptor with a PF00003 domain. Curiously, this metabotropic glutamate receptor also contains a (LIVBP)-like domain that is characteristic of ionotropic glutamate receptors. Two ionotropic glutamate receptors (iGluRs) that are ligand-gated ion channels activated by the neurotransmitter glutamate with Lig_chan

(PF00060) domain were identified in *Dicyema* sp. One of them with a PF10613 (Lig_chan-Glu_bd) and another with PF01094 (ANF_receptor). Thus, both distinct types of glutamate receptors (ionotropic and metabotropic types) are present in *Dicyema* sp. It is well known that glutamate is often associated with non-neuronal signaling and is highly abundant in some animals that lack nervous systems (such as sponges and *Trichoplax*). Previously, we reported that iGluRs are absent in the genome of orthonectid (Mikhailov et al., 2016). Glutamate receptors are also found in plants and many other eukaryotes outside Metazoa (Turano et al., 2001).

Another big group of ionotropic receptors is the Cys-loop ligand-gated ion channel superfamily that is composed of nicotinic acetylcholine, GABA-A, GABA-A-ρ, glycine, 5-HT3, and zinc-activated (ZAC) receptors. We found 8 genes for this superfamily in *Dicyema* sp., identified by the specific transmembrane region domain (PF02932) and the ligand binding domain (PF02931). All these receptors are predicted to be nicotinic acetylcholine-like receptors.

Ion Channels

Despite the reported absence of muscles and neurons, tetrameric ion channels that are often associated with cellular electrical excitability are present in *Dicyema* sp. in numbers similar to the orthonectid *I. linei* (33 and 36 sequences with PF00520, and 11 and 9 with PF07885 in *Dicyema* and the orthonectid, respectively). Although unlike Orthonectida no signatures for voltage-gated sodium ion channel (Na_trans_assoc PF06512) were detected in *Dicyema*, Pfam analysis (for Ca_chan_IQ PF08763) and reciprocal BLAST searches indicates the presence of voltage-gated calcium ion channels in this animal group. The presence of such channels together with tetrameric potassium ion channels implies that electrical excitability in the form of action potentials might exist in dicyemid cells. **Figure 9** provides a hypothetical schema of the intercellular communication and an analog of the neuromuscular junction in dicyemids. This structure may potentially assemble from key predicted proteins typical to many other metazoans.

Genes Encoding Putative Contractile/Muscular Elements

"True" muscle cells are absent in dicyemids and detection of the muscle-specific genes in these animals is interesting. Most of the core muscle proteins, including a type II myosin heavy chain (MyHC) motor protein was already present in unicellular eukaryotes before the origin of multicellular animals (Steinmetz et al., 2012). At the same time, the troponin complex appears to be a universal innovation of bilaterians. Troponin is a complex of three proteins (troponin C, troponin I, and troponin T). These proteins are detected in the dicyemid data by BLAST search, and the troponin domain PF00992 is found by Pfam search. The troponin complex is characteristic of skeletal and cardiac muscles, but not for smooth muscles. It appears that throughout the radical simplification in dicyemids that resulted in massive gene loss (including most of genes encoding the extracellular matrix ECM molecules) and in the absence of specialized muscle cells the troponins remain essential. The presence of troponins relates dicyemids to all other bilaterians with one remarkable exception – the orthonectid. In contrast to dicyemids and other bilaterians, the genome of orthonectid *I. linei* has no troponins despite having specialized muscles. Morphological data suggest that muscles in *I. linei* are similar to smooth muscles, so troponin was likely lost in *I. linei*, and its absence is a derived feature. At the same time another bilaterian hallmark – the myogenic regulatory factor (Myogenic Basic domain PF01586) – is present in the genome of *I. linei*, but was not detected in dicyemids. These findings support the mosaic evolution of many bilaterian traits, supporting the possibility of independent simplifications in these two parasitic lineages.

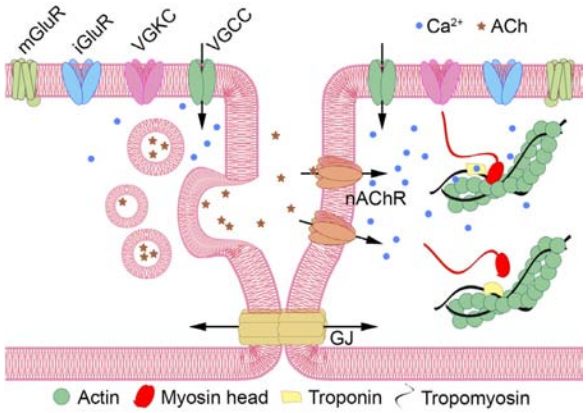


FIGURE 9 | Hypothetical schema of intercellular communication in Dicyemida. Dicyemids have no recognizable neurons and muscles, and yet they have key elements of the neuromuscular system. Metabotropic (mGluRs) or ionotropic glutamate receptors (iGluRs) activate the "presynaptic" cell **(left)**. Voltage-gated tetrameric calcium (VGCCs) and potassium (VGKCs) ion channels generate propagating action potentials. Ca^{++} (blue dots) enters the cytoplasm via VGCCs and triggers the vesicular acetylcholine (ACh, shown by asterisks) release. Activation of nicotinic acetylcholine-like receptors (nAChR) increase Ca^{++} level in the "postsynaptic muscular" cell **(right)** directly or by depolarization of the plasma membrane and VGCC opening. Ca^{++} promote contractile elements activation via the Ca^{++} dependent troponin-tropomyosin-actin-myosin mechanism. Additional interaction of the two cells can occur via gap junctions (GJ).

Gap Junctions and Adhesion Molecules

Gap junctions are a distinct type of intercellular communication channels. In Metazoa, the gap junction proteins belong to two unrelated families – connexins and pannexins (also known as innexins). The connexins are only found in chordates, while the pannexin family is widespread in invertebrates. The presence of gap junctions and innexin/pannexins in dicyemids was demonstrated earlier by transmission electron microscopy (TEM) (Furuya et al., 1997) and molecular cloning (Suzuki et al., 2010). BLAST and Pfam searches with our dicyemid data detected 21 hits with the innexin/pannexin-specific Pfam domain (PF00876) and no connexins. The number of dicyemid pannexins is similar to other invertebrates (25 in *Caenorhabditis elegans*, 13 in *Drosophila melanogaster*). It

appears that unlike the highly reduced chemical signaling, direct intercellular communication via gap junctions is conserved in dicyemids.

Other hallmarks of multicellularity – the adhesion molecules and adherens junctions are retained in dicyemids and were demonstrated in these organisms earlier by TEM (Furuya et al., 1997). The universal metazoan proteins *Integrin* alpha and *Integrin* beta are detected in dicyemids in single copies; immunoglobulin domain is present in 6 sequences and *Cadherin* in 18 copies.

Axon Guidance Molecules and Their Receptors

The simplicity of the nervous system in Orthonectida is associated with a reduction of genes encoding components of axon guidance and synapse formation (Mikhailov et al., 2016). Dicyemids are presumably entirely deprived of the nervous system and follow the same trend of gene loss. Both animal groups lack genes encoding semaphorins, important neuronal pathfinding signaling molecules, and their receptors (plexins). Genes potentially involved in the nervous system development, such as Netrin, Ephrins, and Ephrin receptors are present in Orthonectida but were not identified in Dicyemids. Interestingly, the fasciclin domain (PF02469) is absent in the genome of *I. linei*, but we found its three orthologs in *Dicyema*. Fasciclin (FAS1 domain) is a cell adhesion domain found in neural cell adhesion molecules involved in axonal guidance in insects (Grenningloh and Goodman, 1992).

Peroxisome

The proteins and Pfam domains specific to peroxisome organelles, found in most metazoans, are absent from the dicyemid data. The peroxisomal proteins PEX3, PEX10, PEX12, and PEX19, mandatory for peroxisome function are apparently missing. Failure to detect these genes unequivocally suggests the absence of the organelle. Eight Pfam domains (PF01756, PF04088, PF04614, PF04882, PF05648, PF07163, PF09262, and PF12634) linked to peroxisome in the GO database[1] were not detected in *Dicyema* spp. In this respect, dicyemids are similar to Orthonectida and parasitic flatworms (Tsai et al., 2013).

Phylogenetic Analyses

To clarify the relationships of the two mesozoan groups, Orthonectida and Dicyemida, we used the sequenced transcriptomes of two unidentified species of *Dicyema*. We included the gene predictions of the orthonectid *I. linei* (Mikhailov et al., 2016) in the set of orthologous genes based on the dataset published by Struck et al. (2014). Given the high uncertainty in phylogenetic affinities of orthonectids and dicyemids, we extended taxonomic sampling by adding 30 spiralian taxa from available transcriptomes (see section "Materials and Methods"). Although the data broadly covers the spiralian diversity, several taxonomic groups are still missing or underrepresented in the complement of sequenced genes. To minimize missing data, we merged closely related species within several operational taxonomic units (OTUs) and produced the final matrix with 73 OTUs (69 OTUs for spiralian species) and 87,610 aa positions from 452 individual protein alignments. The proportion of missing data in the concatenated alignment is 40%.

Highly divergent sequences of mesozoans pose a formidable challenge for inference methods due to the confounding effect of long branched taxa on phylogenetic reconstructions. A recognized approach to tackle the long branch attraction (LBA) problem is to use a site-heterogeneous model of sequence evolution (Rodríguez-Ezpeleta et al., 2007). In the Bayesian tree constructed with PhyloBayes (Lartillot et al., 2013) under the site-heterogeneous CAT-GTR model, the dicyemid and orthonectid lineages form the longest branches, yet they do not group thus contradicting monophyly of the Mesozoa (**Figure 10**). We recovered the position of the orthonectid *Intoshia* within annelids with the posterior probability of 1.0. Specifically, the orthonectid forms a branch of the Pleistoannelida that comprises the annelid groups Errantia and Sedentaria (Weigert et al., 2014), while *Owenia*, *Magelona*, Chaetopteridae, and *Phascolosoma* (Sipuncula) occupy more basal positions in the annelid subtree.

The same analysis placed the dicyemid lineage near the base of a group uniting the Rouphozoa (Platyhelminthes, Gastrotricha) and Entoprocta + Cycliophora. However, the position of dicyemids in Bayesian inference is unstable. In about one-third of trees dicyemids were recovered as a sister group to the clade uniting Annelida, Nemertea, Lophophorata (Brachiopoda + Phoronida + Bryozoa), and Mollusca. In about 10% of trees the dicyemids branch off at the base of this group plus (Platyhelminthes + Gastrotricha) plus (Entoprocta + Cycliophora) (**Figure 11**, green branch). The grouping of *Intoshia linei* and Pleistoannelida has been observed in all summed trees. However, the exact position of the orthonectids relative to pleistoannelids is less certain in our analyses. The basal placement of the orthonectids is observed in 50% of trees, and the orthonectids were recovered as a sister group of Sedentaria or Errantia in 38 and 11% of trees, respectively.

The consensus Bayesian tree was obtained from four independent chains. The majority of bipartitions are shared across chains, but convergence on a single topology was not observed. Topologies in each chain uniquely reflect the concurrent hypotheses of spiralian relationships (Kocot, 2016; Kocot et al., 2017). All four chains of our Bayesian analysis are consistent in several major assemblages, including the Rouphozoa (Platyhelminthes, Gastrotricha), Gnathifera (Gnathostomulida, Micrognathozoa, and Syndermata), the sister relationship of Rouphozoa and Entoprocta + Cycliophora, and the basal position of Gnathifera relative to other spiralians. Importantly, in all topologies, the orthonectid is nested within the Annelida, and the dicyemid lineage is inferred sister to the assemblage of Platyhelminthes, Gastrotricha, Entoprocta, and Cycliophora (to the inclusion of Bryozoa in some topologies, **Figure 11**).

Alternative groupings obtained in our analysis include the Lophophorata (Brachiopoda, Phoronida, and Bryozoa) versus

[1] http://geneontology.org/external2go/pfam2go

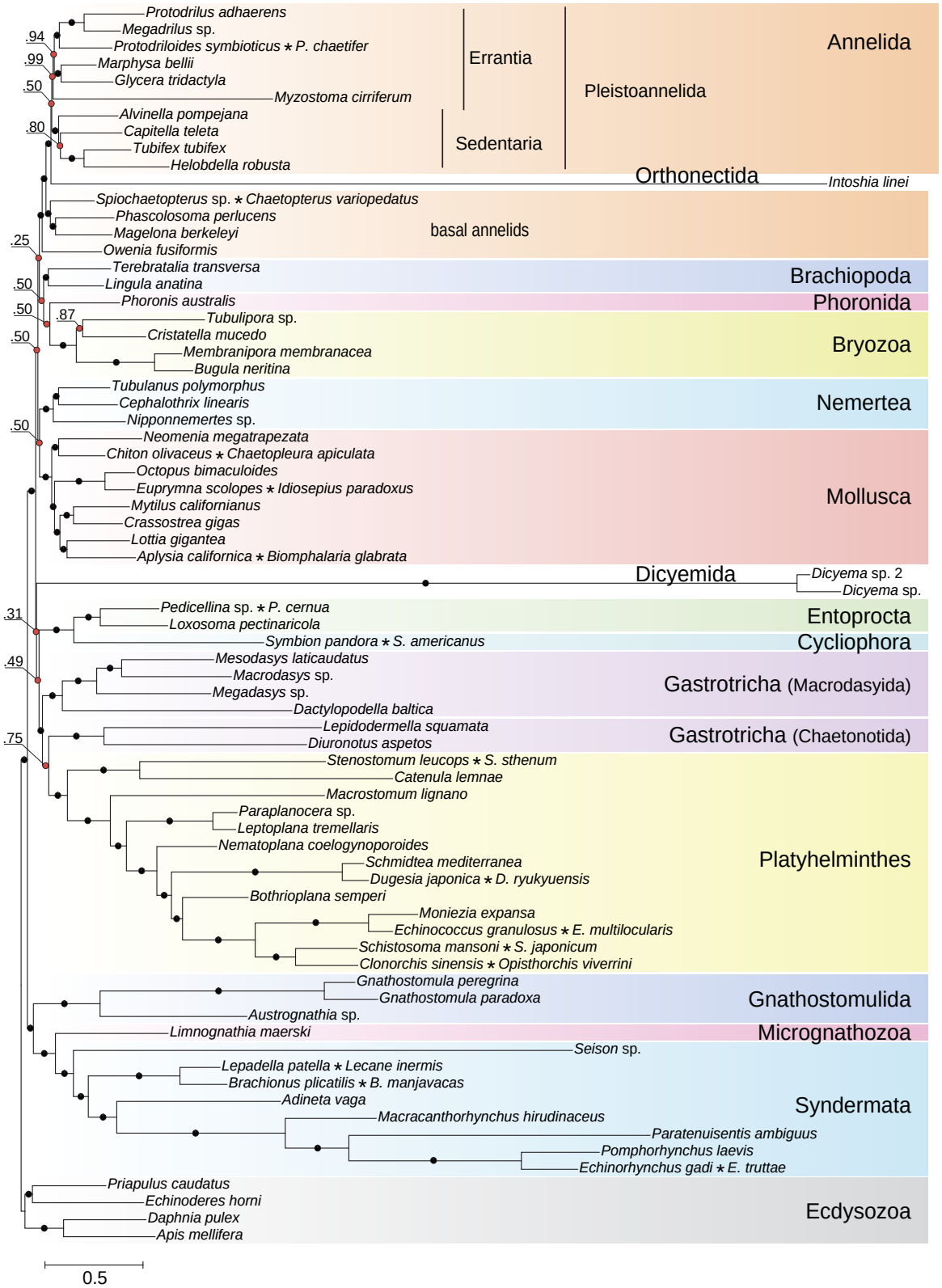


FIGURE 10 | Bayesian tree of Spiralia/Lophotrochozoa with the inclusion of Mesozoa. The consensus topology was constructed from four chains of a PhyloBayes run with the CAT + GTR + Γ4 evolutionary model. Nodes with posterior probabilities below 1.0 are marked with red dots, with those of 1.0 – with black dots. Chimeric operational taxonomic units include names of merged species signed with an asterisk. The tree is rooted with four ecdysozoan lineages.

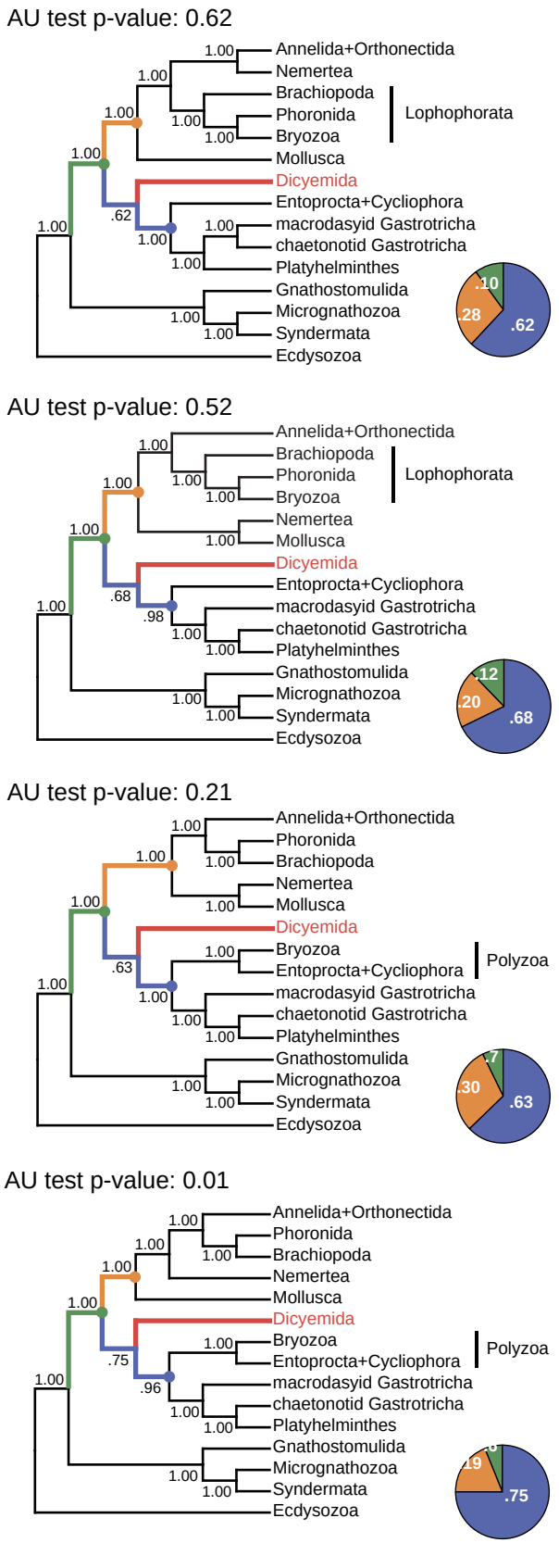


FIGURE 11 | Continued

FIGURE 11 | Tree topologies in the four chains of the PhyloBayes run. Each panel summarizes the topology obtained in a single chain of the analysis. The monophyly of almost all clades and all major spiralian phyla receives posterior probability of 1.0 in each chain (even if they differ between the chains). In contrast, the position of the dicyemid lineage receives moderate support in each chain. The pie charts reflect the portion of trees where the dicyemid lineage occupies one of the three observed positions in the cunsensus (represented with color). Topologies in each chain were compared with the approximately unbiased (AU) test using the "sitelogl" option of the PhyloBayes; AU test *p*-values are shown above each topology.

Polyzoa (Entoprocta, Cycliophora, and Bryozoa), and Vermizoa (Annelida, Nemertea) versus Nemertea + Mollusca (**Figure 11**). A comparison of topologies across chains based on site-wise likelihoods computed with PhyloBayes (the "sitelogl" option of the PhyloBayes readpb_mpi) under the CAT-GTR model and the approximately unbiased (AU) test (Shimodaira, 2002) show that the difference in likelihoods of topologies in chains 1–3 is not significant, but the topology likelihood in chain 4 is significantly lower (*p*-value = 0.01). Chain 2 (**Figure 11**) converges on a topology identical to the consensus four-chains topology (**Figure 10**) but its likelihood is lower than in chain 1 (non-significantly). Excluding chain 4 that failed the AU test and constructing the consensus with the three remaining chains does not affect the topology itself but only node supports due to eliminating the effects of non-monophyletic Lophophorata in chain 4 (**Supplementary Figure S3**).

The best scoring topology supports the monophyletic Lophophorata, the grouping of Annelida and Nemertea, and also the monophyly of macrodasyid and chaetonotid gastrotrichs, which frequently find themselves separate in our analyses (**Figure 10**). Maximum likelihood (ML) analyses of the same dataset with RAxML (Stamatakis, 2014) and IQ-TREE (Nguyen et al., 2015) produce a different view on the phylogeny of Mesozoa. The dicyemids and the orthonectid form a monophyletic group in ML trees with maximal support. In the RAxML analysis under the GTR model the monophyletic Mesozoa branch off with chaetonotid gastrotrichs, similarly to the result obtained in the recent phylogenomic analysis (Lu et al., 2017), but the support of the group is minimal (**Supplementary Figure S4**). In the IQ-TREE run under the C60 profile mixture model the monophyletic Mesozoa are found at the base of the Rouphozoa, again with weak support (64% of ultrafast bootstrap replicates) (see **Supplementary Figure S5**).

Although ML analyses show disagreement with the result of Bayesian inference, modeling of site-heterogeneity by the IQ-TREE profile mixture model does shed light on some spurious cases in spiralian relationships. The divergent annelid *Myzostoma* is correctly grouped with other annelids in the IQ-TREE analysis, in contrast with the RAxML tree where it forms a clade with long branches of the Rouphozoa, Gnathifera, and Mesozoa. The clustering of Rouphozoa and Gnathifera referred to as the Platyzoa, receives maximal support in the RAxML analysis but was previously shown to be artefactual (Struck et al., 2014; Laumer et al., 2015). This grouping is not inferred by both the IQ-TREE and Bayesian analyses.

In contrast with IQ-TREE and PhyloBayes, the RAxML tree supports monophyletic Polyzoa, a group uniting Entoprocta, Cycliophora, and Bryozoa, which was also suggested to be erroneous and caused by the compositional bias in amino acid sequences (Nesnidal et al., 2013).

To test for expected LBA effects, particularly to exclude the possibility of the orthonectid being attracted to annelids by the divergent *Myzostoma*, we conducted additional analyses excluding each of the long branched lineages. Additional datasets were generated by removing *Myzostoma*, *Myzostoma* and both dicyemids, *Myzostoma* and *Intoshia*. Bayesian analyses of additional datasets recovered placement of *Intoshia* within annelids in the absence of *Myzostoma* (**Supplementary Figures S6**, **S7**). The position of dicyemids is also unaffected by the exclusion of other long-branched taxa – the dicyemids occupy a basal position within the Lophotrochozoa after the divergence of Gnatifera in all analyses of the additional datasets (**Supplementary Figures S6**, **S8**).

We also tested our dataset for the effects of compositional heterogeneity by discarding highly heterogeneous alignments and utilizing the data recoding approach (Susko and Roger, 2007). Bayesian inference with a concatenate of 150 protein alignments retained after discarding highly compositionally heterogeneous alignments from the original dataset recovers the same groupings of the mesozoan taxa as the analysis of the full dataset. The orthonectid is nested within the annelid clade (1.0 posterior probability) and the dicyemids branch off at the base of the Rouphozoa + Entoprocta + Cycliophora clade with weak support (0.46 posterior probability) (**Supplementary Figure S9**). Similarly, inference with the Dayhoff-recoded alignment groups the orthonectid with annelids, while leaving the position of the dicyemids uncertain within the Lophotrochozoa (**Supplementary Figure S10**). Remarkably, the PhyloBayes run with recoded data shows adequate convergence between chains (*maxdiff* = 0.17) and infers the monophyletic Gastrotricha. Several conventional groupings, such as the Rouphozoa, are not recovered. Consistent with the proposed artefactual nature of the grouping of Bryozoa and Entoprocta due to compositional heterogeneity (Nesnidal et al., 2013), both test datasets support the monophyletic Lophophorata, whereas the alternative Polyzoa was frequently observed for the complete and non-recoded data. The lack of support for monophyletic Rouphozoa in the analysis with the recoded dataset was similarly obtained in a recent study of the spiralian phylogeny, which aimed at counteracting the impact of compositional heterogeneity (Marlétaz et al., 2019).

Schiffer et al. (2018) selected proteins that support annelid monophyly as an approach to verify the orthonectid position. We also selected 111 protein alignments that contain the annelid signal but with a different method, and used those for Bayesian inference with the PhyloBayes program. In contrast to other Bayesian runs, the consensus presents a stable topology (*maxdiff* value 0.15). In this tree, the orthonectid *I. linei* stabilizes inside the annelids [posterior probability (PP) 1.0], whereas the species of *Dicyema* are not attracted to annelids (**Figure 12**). The position of *Dicyema* remains uncertain within the lineage of long-branched taxa (Platyhelminthes, Gastrotricha, Entoprocta, Cycliophora). Lophophorata and Gastrotricha are reconstructed with PP 1.0 (as in case of the Dayhoff-recoded dataset mentioned above and the non-recoded full dataset in chain 1 that reaches the highest likelihood). Bayesian topologies obtained in chain 1 (**Figure 11**) and both the sub-sampled datasets of 111 proteins with the annelid signal and the 150 proteins with low compositional heterogeneity (**Supplementary Figure S9**) reconstruct the sister relationship of annelids with nemertines. Marlétaz et al. (2019) also report the grouping of annelids and nemertines, with the inclusion of Platyhelminthes.

The lack of convergence in most PhyloBayes analyses precludes strong assertions regarding problematic areas of the spiralian tree. Nevertheless, some clades are reconstructed consistently. We do not observe monophyly of the Mesozoa in any of the chains, in contrast to the recent study by Lu et al. (2017) and in agreement with the evidence from Schiffer et al. (2018), nor do we observe their direct relations with Platyhelminthes.

The orthonectid *I. linei* occupies a stable position within the annelid part of the tree. Its placement is among the major conflicts between ML and Bayesian topologies, which likely indicates the impact of a more complex CAT-GTR model in the presence of long branches of highly divergent lineages like orthonectids and dicyemids. Noteworthy, polyphyletic Mesozoa and the proposed affinity of orthonectids to annelids was also recovered in Schiffer et al. (2018) in Bayesian analyses of a dataset with the less extensive representation of the lophotrochozoan diversity.

The orthonectids share with annelids certain morphological features: the presence of microvillar cuticle, metameric muscles, gonochory (Slyusarev, 2008), and the dorsal ganglion in adult specimens (Slyusarev and Starunov, 2015). Cases of dramatic morphological reduction in annelids are known in archiannelids (Andrade et al., 2015) and lobatocerebrids (Laumer et al., 2015), and especially in dwarf males of the echiurid *Bonnelia*, dinophidid *Dinophilus gyrociliatus*, spionid *Scolelepis laonicola* (Vortsepneva et al., 2008), and siboglinids (Worsaae and Rouse, 2010). Adaptations of orthonectids that had led to the complete loss of coelomic cavity, gonad wall, chaetae, gastral system, nephridia, trochophore larva, and spiral cleavage further demonstrate the extent of morphological regress in the evolution of annelids.

The dicyemid lineage in our analyses exhibits affinity to the Rouphozoa clade, in congruence with Lu et al. (2017), with the intercalation of Entoprocta or Polyzoa, which were not included in analyses by Lu et al. (2017). Schiffer et al. (2018) report an uncertain position of dicyemids at the base of the Lophotrochozoa. We did not recover the position of the dicyemids as part of the Platyhelminthes. A previous analysis of innexin genes (Suzuki et al., 2010) also rejects the kinship of the dicyemids and Platyhelminthes. Furthermore, rhabditophoran platyhelminths are known to possess a unique non-standard mitochondrial genetic code, which was shown to be not the case for the studied mesozoans (Telford et al., 2000; Schiffer et al., 2018). Dicyemids were placed within Spiralia in various taxonomic contexts in molecular phylogenetic studies (Pawlowski et al., 1996; Petrov et al., 2010;

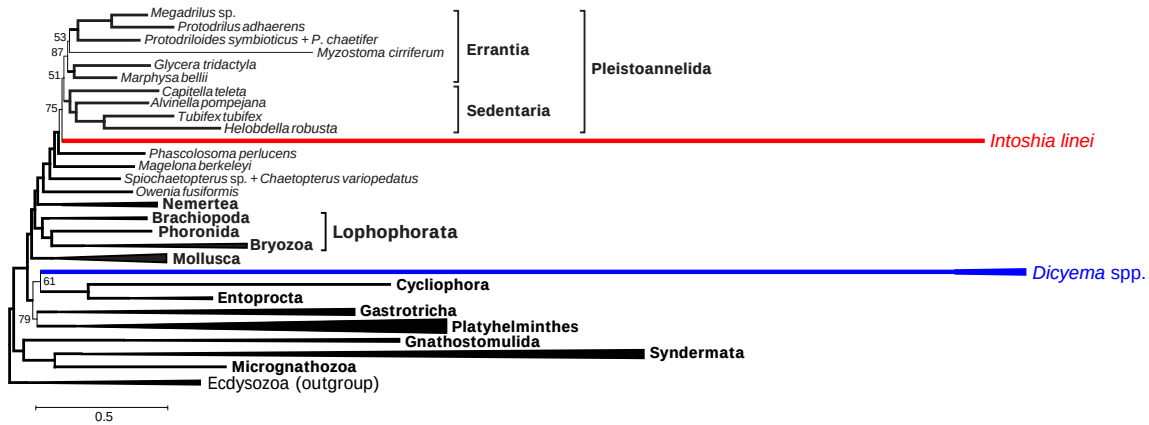


FIGURE 12 | PhyloBayes topology for proteins with the strong annelid signal (concatenate of 111 protein alignments that possess at least 3% positions with $q > 1/2$, 18686 alignment positions in total, CAT + GTR + Γ4 model, 50,000 chain steps, 50% burn-in). Only posterior probability values less than 0.95 are shown. Convergence value of *maxdiff* = 0.15.

Suzuki et al., 2010; Lu et al., 2017) but the interpretation of their body plan remains enigmatic. Being among the simplest known bilaterians, they yet possess multiciliated epithelia, which is not a primitive trait and suggests secondary evolutionary regress, and do not display evident synapomorphies with other animal phyla. Dicyemids might represent a relict lineage of lophotrochozoan animals with no direct relatives that had survived to the present days.

Conclusion

We confirm that orthonectids are extremely simplified annelids and do not form a monophyletic group with dicyemids. Mesozoa is a polyphyletic taxon. Dramatic simplification of their body plan, as well as the metagenetic life cycle, evolved independently in the two lineages. Many conserved bilaterian genes are absent in the genomes of Dicyemida and Orthonectida. At the same time, the pattern of their loss and presence is different, which supports the conclusion that these animal groups are not close relatives and have simplified independently. Analyses of genes related to the basement membrane, neuronal and muscular systems expose the extreme simplicity of dicyemids. Intriguingly, dicyemids lack muscle cells and the genetic factors of muscle cell differentiation but possess the troponin complex specific for striated muscles. Taken together with detection of a relatively big set of nicotinic acetylcholine receptors often associated with neuromuscular signaling and the presence of voltage-gated ion channels, this fact urges reevaluation of the traditional view that dicyemids completely lost the neuro-muscular system. Appealing is to experimentally check if some contractility and movements could be induced in dicyemids by signal molecules such as acetylcholine or glutamate, and for the presence of electrical excitability in the form of propagated calcium action potential in their cells. Small circular extrachromosomal molecules are present in total DNA extracts of dicyemids. Mitochondrial rRNA, tRNA, protein-coding genes and pseudogenes are located on circular molecules. There are short nucleotide sequence motifs confined specifically to circular DNAs in *Dicyema* sp.

MATERIALS AND METHODS

Biological Material, Genome and Transcriptome Sequencing

The original live material on *Dicyema* sp. 1 was collected at the Vostok marine biological station of the Institute for Marine Biology of the Russian Academy of Sciences (the Vostok Bay of the Sea of Japan, Vladivostok, Russia) from dissected kidneys of the giant Pacific octopus *E. dofleini*. Live dicyemids were rinsed individually in filtered marine water and fixed in the RNAlater stabilization solution (Ambion). Total DNA was isolated from tissue samples by Diatom DNA Prep (IsoGene). The sequencing of dicyemid genomic data was performed with an Illumina HiSeq2000 system, generating 140 million paired-end reads.

Total RNA was isolated by TRIzol kit (Invitrogen) and further used for ds cDNA synthesis using the SMART approach (Zhu et al., 2001). SMART-prepared amplified cDNA was then normalized using the DSN normalization method (Zhulidov et al., 2004). Normalization included cDNA denaturation/reassociation, treatment by the duplex-specific nuclease (Shagin et al., 2002), and PCR amplification of the normalized fraction (8 PCR cycles: 95°C for 7 s; 65°C for 20 s; 72°C for 3 min). Normalized cDNA libraries were sequenced using the Roche 454 sequencing technology, producing about 480,000 reads with an average length of 444.6 bases.

Specimens of *Dicyema* sp. 2 was collected at the Friday Harbor Laboratories (Friday Harbor, WA, United States) from circulatory system and kidneys of the octopus *E. dofleini*. All individual animals were washed 3–5 times in 0.2 μm filtered seawater. Then RNA was extracted from individual animals and processed as described elsewhere (Moroz and Kohn, 2013) for Illumina HiSeq 2000 sequencing.

The sequences are deposited in the NCBI: BioProject PRJNA527259 (*Dicyema* sp. 1) and SRA SRP021079 (*Dicyema* sp. 2).

Assembly and Filtering of Dicyemid Sequences

The reads obtained from the DNA library for *Dicyema* sp. 1 were trimmed for adapters with Trimmomatic (Bolger et al., 2014) and assembled by SPAdes (Nurk et al., 2013) using *k*-mer values of 21, 33, 55, and 77. We also performed genome assembly with the Newbler GS De Novo Assembler software (v. 2.9) (using 1/10 of all reads) as a control to our method of circular contigs identification. Gene prediction was performed with Augustus (Stanke and Waack, 2003) after constructing a training set of 200 dicyemid sequences identified in the genomic assembly. The predicted genes were queried against the InterPro database (Finn et al., 2017) with InterProScan (Jones et al., 2014) and genes with InterPro hits were screened for cephalopod sequences with BLAST (Altschul et al., 1997) searches against the NCBI *nr* database. Predictions producing best hit with cephalopod sequences were discarded from the gene set. Completeness estimates were performed with BUSCO (Waterhouse et al., 2017) using the eukaryota_odb9 ortholog set (Zdobnov et al., 2017). HMMER (Eddy, 2011) searches were carried out with Pfam (Finn et al., 2016) Homeodomain (PF00046) and Homeobox_KN (PF05920) profiles to identify homeobox transcription factors in the data. Phylogeny reconstructions for homeobox sequences were performed with IQ-TREE (Nguyen et al., 2015) using the LG + C20 + G4 model of sequence evolution or with PhyloBayes (Lartillot et al., 2013) using the LG +CAT + G4 model.

The reads obtained from the cDNA library for *Dicyema* sp. 1 were trimmed for adapters, non-coding RNA, low-quality and low-complexity sequences with the SeqClean software (Dana-Farber Cancer Institute[2]), and about 430,000 reads were retained. Data was further assembled with the original 454 Newbler GS De Novo Assembler software (v. 2.9) utilizing flowgram quality data and settings that maximize contig overlap. The "-urt" option was invoked to improve contigging in low depth portions of the assembly. Fusions of transcripts that can potentially occur with low-depth assembly extensions in densely packed genomes are subsequently eliminated in our experimental design by alignment filtering at the supermatrix construction step. The obtained assembly contained 19,641,638 bases, and 22,082 isotigs of average size 889 bases, N50 size of 1,081, and the largest isotig size of 9,199. Protein coding regions were predicted using TransDecoder (Haas et al., 2013) with settings to maximize the sensitivity of capturing ORFs regardless of the predicted coding likelihood score by accounting for homology to known proteins in the Pfam (Finn et al., 2014) and UniProtKB/Swiss-Prot (UniProt Consortium, 2015) curated databases. Coding region prediction with TransDecoder was set to the minimal predicted protein length of 80 aa. The predicted proteome contained 15,227 unique coding regions.

The second dicyemid transcriptome sequenced using the Illumina platform was assembled with Trinity (Grabherr et al., 2011). Before assembly the reads were processed with the SeqClean software, and the prediction of coding regions was performed by TransDecoder, similarly to the transcriptome of the first dicyemid.

The transcriptomes of dicyemids were derived from samples contaminated with their cephalopod host. Therefore, we paid special attention to avoid mixing dicyemid and cephalopod sequences in the phylogenetic analysis. The transcriptomes of dicyemids were first screened for cephalopod sequences by performing BLAST (Altschul et al., 1997) searches against the NCBI RefSeq database (O'Leary et al., 2016). Two dicyemid transcriptomes were processed independently. In the first step of decontamination we filtered out proteins having best hit in RefSeq belonging to prokaryotes (and having at least 50% identity). This lead to rejection of only 35 proteins for *Dicyema* sp. 1 and 363 proteins for *Dicyema* sp. 2. In the second step we removed all dubious proteins if their local alignment score with any cephalopod protein higher than in all the other considered species (with the same query protein). Sequences with best hits to cephalopods were discarded from the transcriptomes if the sequence identity exceeded 70%. For the third filtering step we queried the proteins of *O. bimaculoides* combined with several transcriptomes of *Octopus vulgaris* (NCBI BioProject PRJNA79361 and Sequence Read Archive entries SRR331946, SRR1507221) against a custom database containing 9 metazoan proteomes (4 molluscs, 2 annelids, a brachiopod *Lingula anatina*, an ecdysozoan *Limulus polyphemus*, and a deuterostome *Danio rerio*) and the dicyemid transcriptome, and inspected dicyemid sequences that produced hits with the highest match to the cephalopod queries among the 10 metazoans. All dubious sequences (hits with at least 80% identity) captured by this method were discarded from the dicyemid transcriptomes as potential cephalopod contamination.

Search for "Circular" Contigs, Signals, and Mitochondrial Sequences

The contigs constructed from shotgun fragments display special characteristics emerging from the genome assembly algorithms based on De Bruijn graph of *k*-mers. This approach results in "circular" contigs starting and ending with the same *k*-mer. After assembly, terminal repeats equal in length to the *k*-mer were cut off. Contigs analyzed in sections "Circular Contigs in Genomic Assembly of *Dicyema* sp." and "Mitochondrial DNA of *Dicyema* sp.," and NCBI submission data have been cleaned off the terminal repeats. In this study, a contig was considered "circular" if it had terminal direct repeats $\geq$ 77 nt in length (k77). The length distribution of contigs assembled by different methods (Newbler and SPAdes) was compared with the two-sample Kolmogorov–Smirnov test implemented in the SciPy package in Python 3. Here the null hypothesis is that contig lengths come from the same distribution. High *p*-values in this case reflect high probabilities of this hypothesis. Low complexity regions were detected with the DUST algorithm from the MEME Suite (Bailey and Elkan, 1994) with standard settings. MEME and ChIPMunk (Kulakovskiy et al., 2010) tools with the default parameters were applied to the task of finding

[2]https://sourceforge.net/projects/seqclean

specific motifs. The reverse lookup for the signal presence was done via FIMO (from the MEME Suite) with the *p*-value threshold of 10^{-4}. Moreover, highly conserved elements of circles in dicyemids were found utilizing the technique borrowed from Rubanov et al. (2016). The method identifies highly conserved DNA elements on the base of the identification of dense subgraphs in a specially built multipartite graph (whose parts correspond to genomes). Specifically, the algorithm does not rely on genome alignments, no pre-identified perfectly conserved elements; instead, it performs a fast search for pairs of words (in different genomes) of maximum length with the difference below the specified edit distance. Such pair defines an edge whose weight equals the maximum (or total) length of words assigned to its ends. The graph composed of these edges is then compacted by merging some of its edges and vertices. The dense subgraphs are identified by a cellular automaton-like algorithm; each subgraph defines a cluster composed of similar inextensible words from different genomes (Rubanov et al., 2016).

HMMER3 package (Eddy, 2011) along with the Pfam-A database were used to find the circles containing protein-coding sequences, whereas an additional verification step was performed in BLAST. The search itself was conducted through the database composed of six-frame translated circular sequences. The search for genes coding for mitochondrial proteins was conducted with BLAST using mitochondrial protein-coding gene sequences from flatworms as queries, MITOS (Bernt et al., 2013) and HMMER3 using HMM profiles from the Pfam-A database. Mitochondrial rrnS genes in dicyemids are highly diverged and poorly detected with BLAST. Their detection was conducted with HMMER3 with HMM profiles preliminarily generated from the set of 140 rrnS alignments from other organisms (140 species of bilaterians, cnidarians, and placozoans). All findings were verified using *blastp* or *blastn* with *nr* NCBI database. It was proposed that the dicyemid small mitochondrial circular DNA molecules are generated from the usual long multigene mitochondrial DNA (Awata et al., 2006). If such long mtDNA exists together with mitochondrial mini-circles we can expect the cases when one read from the sequencing library corresponds to a particular mitochondrial mini-circle while its pair read maps elsewhere. *Blastn* with minimal word size was used to map raw paired end reads to circular contigs coding for mitochondrial genes to search for hypothetical high-molecular-weight mtDNA. Reads pair analysis was conducted after that in order to find the reads whose pair does not map to the initial circular contig. Mitochondrial tRNA secondary structures were predicted using the MiTFi program (Jühling et al., 2012).

Taxonomic Expansion of Alignments

The starting set of orthologous genes used in this work is based on a dataset for phylogenetic reconstructions within Spiralia assembled by Struck et al. (2014) that was later expanded with sequences of orthonectid *I. linei* (Mikhailov et al., 2016). The base set of orthologs contained 469 alignments with a total of 62 spiralian species and four ecdysozoan species. To extend the taxonomic sampling of Spiralia and minimize the missing data in the dataset we obtained predicted proteins from several genomic projects accessible through public databases and collected transcriptomic data from the NCBI Sequence Read Archive. The annotations for the genomes of *Clonorchis sinensis*, *Echinococcus granulosus*, *L. anatina*, *O. bimaculoides*, *Priapulus caudatus* were obtained from the GenBank database, and the proteins of *Adineta vaga* were obtained from the Genoscope database. The NCBI Sequence Read Archive was used to extract raw sequence data of another 31 spiralian species (see **Supplementary Table S1**).

The assemblies of the SRA transcriptome data were performed with Trinity (Grabherr et al., 2011) after cleaning the reads with SeqClean (Dana-Farber Cancer Institute[3]) from adapter sequences using the UniVec_Core database[4] and filtering ribosomal RNA sequences using a database of eukaryotic rRNAs. The prediction of proteins in the assembled transcripts was performed with TransDecoder (Haas et al., 2013), which was assisted with searches against the Pfam (Finn et al., 2014) and UniProtKB/Swiss-Prot (UniProt Consortium, 2015) databases.

The addition of proteins from the newly assembled data to orthologous groups featured in the base set of alignments was performed using the procedure for mapping genes to existing orthologous groups (Fischer et al., 2011) of the OrthoMCL database (Chen et al., 2006). The genes from the initial dataset and novel transcriptomic and genomic data were assigned to orthologous groups of OrthoMCL-DB, and the genes within the same orthologous group were extracted and aligned together using MUSCLE (Edgar, 2004). When more than one sequence per organism was assigned to the same group of orthologs, only the sequence scoring highest against the orthologous group in the initial dataset was selected for the alignment.

Phylogenetic Analyses

The concatenation of individual gene alignments was performed with Scafos (Roure et al., 2007) using the option to construct chimeric sequences for several closely related taxa. The following 15 chimeric taxa were constructed for the analysis: *Aplysia californica* + *Biomphalaria glabrata*, *Brachionus plicatilis* + *B. manjavacas*, *Chiton olivaceus* + *Chaetopleura apiculata*, *Clonorchis sinensis* + *Opisthorchis viverrini*, *Dugesia japonica* + *Dugesia ryukyuensis*, *Echinococcus granulosus* + *Echinococcus multilocularis*, *Echinorhynchus gadi* + *Echinorhynchus truttae*, *Euprymna scolopes* + *Idiosepius paradoxus*, *Lepadella patella* + *Lecane inermis*, *Pedicellina* sp. + *P. cernua*, *Protodrilloides symbioticus* + *P. chaetifer*, *Schistosoma mansoni* + *S. japonicum*, *Spiochaetopterus* sp. + *Chaetopterus variopedatus*, *Stenostomum leucops* + *Stenostomum sthenum*, *Symbion pandora* + *S. americanus*. Another ten species that were present in the starting set of alignments were removed due to poor representation in the final alignment: *Alcyonidium diaphanum*, *Fasciola gigantica*, *Flustra foliacea*, *Lumbricus rubellus*, *Philodina roseola*, *Rotatoria rotatoria*, *Spirometra erinacei*, *Stylochoplana maculata*, *Taenia solium*, *Turbanella ambronensis*. The final number of operational taxonomic units featured in the analysis is 73. Before concatenation, the alignments were trimmed with TrimAl

[3]https://sourceforge.net/projects/seqclean

[4]http://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/

(Capella-Gutierrez et al., 2009) to remove poorly aligned regions. The trimming was performed with a gap threshold of 0.5 and a similarity threshold of 0.001. After the removal of invariant positions, the length of the concatenated alignment totaled 87,610 positions, with 40% missing data. Compositional heterogeneity in the alignment partitions (i.e., individual protein alignments after masking) was evaluated using the relative composition frequency variability (RCFV) metric (Zhong et al., 2011). The RCFV values were calculated using BaCoCa (Kuck and Struck, 2014). The low compositional heterogeneity dataset was generated by discarding 302 partitions (referred to in the paper simply as protein alignments) with RCFV value exceeding 0.115.

The phylogenetic reconstructions were performed with PhyloBayes-MPI 1.7 (Lartillot et al., 2013), RAxML (Stamatakis, 2014), and IQ-TREE (Nguyen et al., 2015). The RAxML analysis was carried out utilizing the complete analysis function (−f a) with 150 rapid bootstrap replicates and the PROTCATGTR model of evolution. The IQ-TREE analysis was performed using the LG + C60 + F + G4 evolutionary model, and node support was calculated using the ultrafast bootstrap approximation (Minh et al., 2013) with 1,000 replicates. The Bayesian inference with PhyloBayes was carried out using the CAT + GTR + Γ4 model, and the analyses were run with four chains. For the main dataset, the majority rule consensus tree was reconstructed after 30,000 cycles using one out of ten cycles with a 60% burn-in. PhyloBayes analyses of the additional datasets were conducted similarly to the main dataset; the consensus trees were reconstructed after 5,000 or 15,000 cycles with a 50% burn-in. Analysis of the recoded alignment was performed with PhyloBayes utilizing the recode option and the Dayhoff recoding scheme with six amino acid groups (Dayhoff et al., 1978). The convergence of the chains was assessed by comparing bipartitions using the *pbcomp* utility from the PhyloBayes package.

Comparison of topologies in the four chains of the Bayesian inference of the main dataset was performed using the CONSEL program (Shimodaira and Hasegawa, 2001) and the "sitelogl" option of the PhyloBayes readpb_mpi program. The site-specific marginal log likelihoods were computed for each chain across 10 data points sampled over 2,000 cycles after a 20,000 cycle burn-in.

Alignment partitions (i.e., individual protein alignments after masking) with the strong annelid signal were selected as follows. In a protein alignment we define two sets of sequences – G_1 (ingroup), and G_2 (outgroup). Only alignment positions containing no more than a half of missing data (gaps or X's) in each of the two sets are considered. For each such position i-value $q(i)$ is determined as the maximum of frequency differences of each amino acid in this position from G_1 and G_2. Missing data is ignored. Maximum $q(i)$ value is 1 when G_1 consists only of one character, and G_2 does not contain this character. Under any $q(i) > 1/2$ there exists a character $a(i)$ observed in more than a half of taxa from G_1 but much less frequently in G_2 [frequency difference is $q(i) > 1/2$]. In the phylogenetic context, when $G_1 + G_2$ constitute a monophyletic clade, and G_1 is a narrower natural clade, high $q(i)$ values can be interpreted as presence of a synapomorphy against G_2. Notably, in this analysis $q(i)$ values are used only to select partitions but not for alignment editing or positions removal. In our case of detecting the annelid signal, G_1 contained all annelids except the orthonectid, and G_2 – all non-annelid taxa except dicyemids in order to obtain $q(i)$ estimates unbiased with respect to the lineages under study.

AUTHOR CONTRIBUTIONS

OZ performed most of the computations, analyzed the data, and drafted the manuscript. KM, YP, SI, OP, and LR performed additional computations, analyzed the data, and wrote the manuscript. LR obtained original RNA-Seq data, assembled the transcriptome of *Dicyema* sp. 1. ML and AP obtained original DNA-Seq data. LM obtained original RNA-Seq data on *Dicyema* sp. 2. VL supervised the computational part of the work. VA designed and supervised the research. All authors read and approved the manuscript.

FUNDING

This research was performed at IITP RAS and supported by the Russian Science Foundation, project no. 14-50-00150. Sequencing of the *Dicyema* sp. 2 transcriptome was supported by the Government of the Russian Federation, grant #14.W03.31.0015. The phylogenetic analyses were supported by the Russian Foundation for Basic Research grant nos. 18-29-13014 and 18-29-13037. The computations were carried out on MVS-10P at Joint Supercomputer Center of the Russian Academy of Sciences (JSCC RAS).

ACKNOWLEDGMENTS

We thank V. P. Kuznetsov for graphic design in figures. We are deeply grateful to all reviewers for the productive dialogue that led to the enrichment of the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00443/full#supplementary-material

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Andrade, S. C., Novo, M., Kawauchi, G. Y., Worsaae, K., Pleijel, F., Giribet, G., et al. (2015). Articulating "Archiannelids": phylogenomics and annelid relationships, with emphasis on meiofaunal taxa. *Mol. Biol. Evol.* 32, 2860–2875. doi: 10.1093/molbev/msv157

Armstrong, M. R., Blok, V. C., and Phillips, M. S. (2000). A multipartite mitochondrial genome in the potato cyst nematode *Globodera pallida*. *Genetics* 154, 181–192.

Aruga, J., Odaka, Y. S., Kamiya, A., and Furuya, H. (2007). *Dicyema* Pax6 and Zic: tool-kit genes in a highly simplified

bilaterian. *BMC Evol. Biol.* 7:201. doi: 10.1186/1471-2148-7-201

Ataev, G. L. (2017). *Reproduction of Trematode Parthenites: An Overview of the Main Theories*. St. Petersburg: Nauka, 87.

Awata, H., Noto, T., and Endoh, H. (2006). Peculiar behavior of distinct chromosomal DNA elements during and after development in the dicyemid mesozoan *Dicyema japonicum*. *Chromosome Res.* 14, 817–830. doi: 10.1007/s10577-006-1084-z

Bailey, T. L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2, 28–36.

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., et al. (2013). MITOS: improved de novo metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.* 69, 313–319. doi: 10.1016/j.ympev.2012.08.023

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Bondarenko, N., Bondarenko, A., Starunov, V., and Slyusarev, G. (2019). Comparative analysis of the mitochondrial genomes of Orthonectida: insights into the evolution of an invertebrate parasite species. *Mol. Genet. Genomics* .

Brusca, R. C., and Brusca, G. J. (2003). *Invertebrates*. Sunderland, MA: Sinauer Associates.

Burger, G., Jackson, C. J., and Waller, R. F. (2012). "Unusual mitochondrial genomes and genes," in *Organelle Genetics*, ed. C. Bullerwell (Berlin: Springer), 41–77. doi: 10.1007/978-3-642-22380-8_3

Capella-Gutierrez, S., Silla-Martinez, J. M., and Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi: 10.1093/bioinformatics/btp348

Catalano, S. R., Whittington, I. D., Donnellan, S. C., Bertozzi, T., and Gillanders, B. M. (2015). First comparative insight into the architecture of COI mitochondrial minicircle molecules of dicyemids reveals marked inter-species variation. *Parasitology* 142, 1066–1079. doi: 10.1017/S0031182015000384

Chen, F., Mackey, A. J., Stoeckert, C. J. Jr., and Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34, D363–D368. doi: 10.1093/nar/gkj123

Cheng, T. C. (1986). *General Parasitology*, 2nd Edn. Cambridge, MA: Academic Press. 827.

Czaker, R. (2000). Extracellular matrix (ECM) components in a very primitive multicellular animal, the dicyemid mesozoan *Kantharella antarctica*. *Anat. Rec.* 259, 52–59. doi: 10.1002/(SICI)1097-0185(20000501)259:1<52::AID-AR6<3.0.CO;2-J

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). "A model of evolutionary change in proteins," in *Atlas of Protein Sequence and Structure*, Vol. 5, ed. M. O. Dayhoff (Washington, D.C: National Biomedical Research Foundation), 345–352.

de Rosa, R., Grenier, J. K., Andreeva, T., Cook, C. E., Adoutte, A., Akam, M., et al. (1999). Hox genes in brachiopods and priapulids and protostome evolution. *Nature* 399, 772–776. doi: 10.1038/21631

Dobrovolskij, A. A., and Ataev, G. L. (2003). "Thr nature of reproduction of digenea rediae and sprosycts," in *Taxonomy, Ecology and Evolution of Metazoan Parasites*, eds C. Combes, J. Jourdane, A. Ducreux-Modat, and J. R. Pagès (Perpignan: University of Perpignan Press), 273–290.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Fidler, A. L., Darris, C. E., Chetyrkin, S. V., Pedchenko, V. K., Boudko, S. P., Brown, K. L., et al. (2017). Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues. *Elife* 6:e24176. doi: 10.7554/eLife.24176

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., et al. (2017). InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* 45, D190–D199. doi: 10.1093/nar/gkw1107

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., et al. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* 44, D279–D285. doi: 10.1093/nar/gkv1344

Fischer, S., Brunk, B. P., Chen, F., Gao, X., Harb, O. S., Iodice, J. B., et al. (2011). Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. *Curr. Protoc. Bioinform.* 6, 11–19. doi: 10.1002/0471250953.bi0612s35

Furuya, H., Hochberg, F. G., and Tsuneki, K. (2004). Cell number and cellular composition in infusoriform larvae of dicyemid mesozoans (phylum Dicyemida). *Zool. Sci.* 20, 877–889. doi: 10.2108/zsj.21.877

Furuya, H., and Tsuneki, K. (2003). Biology of dicyemid mesozoans. *Zool. Sci.* 20, 519–532. doi: 10.2108/zsj.20.519

Furuya, H., Tsuneki, K., and Koshida, Y. (1997). Fine structure of dicyemid mesozoans, with special reference to cell junctions. *J. Morphol.* 231, 297–305. doi: 10.1002/(SICI)1097-4687(199703)231:3<297::AID-JMOR8<3.0.CO;2-8

Gibson, T., Blok, V. C., and Dowton, M. (2007). Sequence and characterization of six mitochondrial subgenomes from *Globodera rostochiensis*: multipartite structure is conserved among close Nematode relatives. *J. Mol. Evol.* 65, 308–315. doi: 10.1007/s00239-007-9007-y

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Grenningloh, G., and Goodman, C. S. (1992). Pathway recognition by neuronal growth cones: genetic analysis of neural cell adhesion molecules in *Drosophila*. *Curr. Opin. Neurobiol.* 2, 42–47. doi: 10.1016/0959-4388(92)90160-m

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Hanelt, B., Van Schyndel, D., Adema, C. M., Lewis, L. A., and Loker, E. S. (1996). The phylogenetic position of *Rhopalura ophiocomae* (Orthonectida) based on 18S ribosomal DNA sequence analysis. *Mol. Biol. Evol.* 13, 1187–1191. doi: 10.1093/oxfordjournals.molbev.a025683

Hyman, L. H. (1940). *The Invertebrates: Protozoa Through Ctenophra*, Vol. 1. New York, NY: McGraw Hill.

Joffroy, B., Uca, Y. O., Prešern, D., Doye, J. P. K., and Schmidt, T. L. (2018). Rolling circle amplification shows a sinusoidal template length-dependent amplification bias. *Nucleic Acids Res.* 46, 538–545. doi: 10.1093/nar/gkx1238

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi: 10.1093/bioinformatics/btu031

Jühling, F., Putz, J., Bernt, M., Donath, A., Middendorf, M., Florentz, C., et al. (2012). Improved systematic tRNA gene annotation allows new insights into the evolution of mitochondrial tRNA structures and into the mechanisms of mitochondrial genome rearrangements. *Nucleic Acids Res.* 40, 2833–2845. doi: 10.1093/nar/gkr1131

Katayama, T., Wada, H., Furuya, H., Satoh, N., and Yamamoto, M. (1995). Phylogenetic position of the dicyemid mesozoa inferred from 18S rDNA sequences. *Biol. Bull.* 189, 81–90. doi: 10.2307/1542458

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493. doi: 10.1101/gr.113985.110

Kobayashi, M., Furuya, H., and Holland, P. W. (1999). Dicyemids are higher animals. *Nature* 401:762. doi: 10.1038/44513

Kobayashi, M., Furuya, H., and Wada, H. (2009). Molecular markers comparing the extremely simple body plan of dicyemids to that of lophotrochozoans: insight from the expression patterns of *Hox*, *Otx*, and *brachyury*. *Evol. Dev.* 11, 582–589. doi: 10.1111/j.1525-142X.2009.00364.x

Kocot, K. M. (2016). On 20 years of Lophotrochozoa. *Org. Divers. Evol.* 16, 329–343. doi: 10.1007/s13127-015-0261-3

Kocot, K. M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., et al. (2017). Phylogenomics of Lophotrochozoa with consideration of systematic error. *Syst Biol.* 66, 256–282. doi: 10.1093/sysbio/syw079

Kolesnikov, A. A., and Gerasimov, E. S. (2012). Diversity of mitochondrial genome organization. *Biochemistry (Moscow)* 77, 1424–1435. doi: 10.1134/S0006297912130020

Kozloff, E. N. (1990). *Invertebrates*. Philadelphia: Saunders College Publishing.

Kuck, P., and Struck, T. H. (2014). BaCoCa–a heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. *Mol. Phylogenet. Evol.* 70, 94–98. doi: 10.1016/j.ympev.2013.09.011

Kulakovskiy, I. V., Boeva, V. A., Favorov, A. V., and Makeev, V. J. (2010). Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* 26, 2622–2623. doi: 10.1093/bioinformatics/btq488

Lartillot, N., Rodrigue, N., Stubbs, D., and Richer, J. (2013). PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* 62, 611–615. doi: 10.1093/sysbio/syt022

Laumer, C. E., Bekkouche, N., Kerbl, A., Goetz, F., Neves, R. C., Sorensen, M. V., et al. (2015). Spiralian phylogeny informs the evolution of microscopic lineages. *Curr. Biol.* 25, 2000–2006. doi: 10.1016/j.cub.2015.06.068

Lavrov, D. V., and Pett, W. (2016). Animal mitochondrial DNA as we do not know it: mt-genome organization and evolution in nonbilaterian lineages. *Genome Biol. Evol.* 8, 2896–2913. doi: 10.1093/gbe/evw195

Lu, T. M., Kanda, M., Satoh, N., and Furuya, H. (2017). The phylogenetic position of dicyemid mesozoans offers insights into spiralian evolution. *Zool. Lett.* 3:6. doi: 10.1186/s40851-017-0068-5

Malakhov, V. V. (1990). *Mysterious Group of Marine Invertebrates: Tricholpax, Orthonectida, Dicyemida, Porifera*. Moscow: Moscow University Press, 144.

Marlétaz, F., Peijnenburg, K. T. C. A., Goto, N., Satoh, N., and Rokhsar, D. S. (2019). A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. *Curr. Biol.* 29, 312–318. doi: 10.1016/j.cub.2018.11.042

Mikhailov, K. V., Slyusarev, G. S., Nikitin, M. A., Logacheva, M. D., Penin, A. A., Aleoshin, V. V., et al. (2016). The genome of *Intoshia linei* affirms orthonectids as highly simplified spiralians. *Curr. Biol.* 26, 1768–1774. doi: 10.1016/j.cub.2016.05.007

Minh, B. Q., Nguyen, M. A., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30, 1188–1195. doi: 10.1093/molbev/mst024

Moroz, L. L., and Kohn, A. B. (2013). Single-neuron transcriptome and methylome sequencing for epigenomic analysis of aging. *Methods Mol. Biol.* 1048, 323–352. doi: 10.1007/978-1-62703-556-9_21

Nesnidal, M. P., Helmkampf, M., Meyer, A., Witek, A., Bruchhaus, I., Ebersberger, I., et al. (2013). New phylogenomic data support the monophyly of lophophorata and an ectoproct-phoronid clade and indicate that polyzoa and kryptrochozoa are caused by systematic bias. *BMC Evol. Biol.* 13:253. doi: 10.1186/1471-2148-13-253

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi: 10.1093/molbev/msu300

Noto, T., Yazaki, K., and Endoh, H. (2003). Developmentally regulated extrachromosomal circular DNA formation in the mesozoan *Dicyema japonicum*. *Chromosoma* 111, 359–368. doi: 10.1007/s00412-002-0216-2

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., et al. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* 20, 714–737. doi: 10.1089/cmb.2013.0084

Odintsova, M. S., and Yurina, N. P. (2005). Genomics and evolution of cellular organelles. *Russ. J. Genet.* 41, 957–967. doi: 10.1007/s11177-005-0187-5

Ogino, K., Tsuneki, K., and Furuya, H. (2010). Unique genome of dicyemid mesozoan: highly shortened spliceosomal introns in conservative exon/intron structure. *Gene* 449, 70–76. doi: 10.1016/j.gene.2009.09.002

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44, D733–D745. doi: 10.1093/nar/gkv1189

Pawlowski, J., Montoya-Burgos, J. I., Fahrni, J. F., Wuest, J., and Zaninetti, L. (1996). Origin of the Mesozoa inferred from 18S rRNA gene sequences. *Mol. Biol. Evol.* 13, 1128–1132. doi: 10.1093/oxfordjournals.molbev.a025675

Petrov, N. B., Aleshin, V. V., Pegova, A. N., Ofitserov, M. V., and Slyusarev, G. S. (2010). New insight into the phylogeny of Mesozoa: evidence from the 18S and 28S rRNA genes. *Moscow Univers. Biol. Sci. Bull.* 65, 167–169. doi: 10.3103/S0096392510040127

Prince, F., Prince, F., Katsuyama, T., Oshima, Y., Plaza, S., Resendez-Perez, D., et al. (2008). The YPWM motif links antennapedia to the basal transcriptional machinery. *Development* 135, 1669–1679. doi: 10.1242/dev.018028

Robertson, H. E., Schiffer, P. H., and Telford, M. J. (2018). The mitochondrial genomes of the mesozoans *Intoshia linei*, *Dicyema* sp. and *Dicyema japonicum*. *Parasitol. Open* 4:e16. doi: 10.1017/pao.2018

Rodríguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B. F., and Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Syst. Biol.* 56, 389–399. doi: 10.1080/10635150701397643

Roure, B., Rodriguez-Ezpeleta, N., and Philippe, H. (2007). SCaFoS: a tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol. Biol.* 7(Suppl. 1):S2. doi: 10.1186/1471-2148-7-S1-S2

Rubanov, L. I., Seliverstov, A. V., Zverkov, O. A., and Lyubetsky, V. A. (2016). A method for identification of highly conserved elements and evolutionary analysis of superphylum Alveolata. *BMC Bioinformatics* 17:385. doi: 10.1186/s12859-016-1257-5

Ruppert, E. E., Fox, R. S., and Barnes, R. D. (2004). *Invertebrate Zoology: A Functional Evolutionary Approach*, Seventh Edn. Boston, MA: Brooks/Cole Thompson Learning.

Schiffer, P., Robertson, H., and Telford, M. J. (2018). Orthonectids are highly degenerate annelid worms. *Curr. Biol.* 28, 1970.e3–1974.e3. doi: 10.1016/j.cub.2018.04.088

Shagin, D. A., Rebrikov, D. V., Kozhemyako, V. B., Altshuler, I. M., Shcheglov, A. S., Zhulidov, P. A., et al. (2002). A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas. *Genome Res.* 12, 1935–1942. doi: 10.1101/gr.547002

Shao, R., Kirkness, F., and Barker, S. C. (2009). The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*. *Genome Res.* 19, 904–912. doi: 10.1101/gr.083188.108

Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* 51, 492–508. doi: 10.1080/10635150290069913

Shimodaira, H., and Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17, 1246–1247. doi: 10.1093/bioinformatics/17.12.1246

Slyusarev, G. S. (2008). Phylum Orthonectida: morphology, biology, and relationships to other multicellular animals. *Zh. Obshch. Biol.* 69, 403–427.

Slyusarev, G. S., and Starunov, V. V. (2015). The structure of the muscular and nervous systems of the female *Intoshia linei* (Orthonectida). *Org. Divers. Evol.* 16, 65–71. doi: 10.1007/s13127-015-0246-2

Smith, D. R., and Keeling, P. J. (2015). Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10177–10184. doi: 10.1073/pnas.1422049112

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(Suppl. 2), ii215–ii225. doi: 10.1093/bioinformatics/btg1080

Steinmetz, P. R., Kraus, J. E., Larroux, C., Hammel, J. U., Amon-Hassenzahl, A., Houliston, E., et al. (2012). Independent evolution of striated muscles in cnidarians and bilaterians. *Nature* 487, 231–234. doi: 10.1038/nature11180

Struck, T. H., Wey-Fabrizius, A. R., Golombek, A., Hering, L., Weigert, A., Bleidorn, C., et al. (2014). Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of spiralia. *Mol. Biol. Evol.* 31, 1833–1849. doi: 10.1093/molbev/msu143

Susko, E., and Roger, A. J. (2007). On reduced amino acid alphabets for phylogenetic inference. *Mol. Biol. Evol.* 24, 2139–2150. doi: 10.1093/molbev/msm144

Suzuki, T. G., Ogino, K., Tsuneki, K., and Furuya, H. (2010). Phylogenetic analysis of dicyemid mesozoans (phylum Dicyemida) from innexin amino acid sequences: dicyemids are not related to Platyhelminthes. *J. Parasitol.* 96, 614–625. doi: 10.1645/GE-2305.1

Telford, M. J., Herniou, E. A., Russell, R. B., and Littlewood, D. T. (2000). Changes in mitochondrial genetic codes as phylogenetic characters: two examples from the flatworms. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11359–11364. doi: 10.1073/pnas.97.21.11359

Tsai, I. J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sánchez-Flores, A., Brooks, K. L., et al. (2013). The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* 496, 57–63. doi: 10.1038/nature12031

Turano, F. J., Panta, G. R., Allard, M. W., and van Berkum, P. (2001). The putative glutamate receptors from plants are related to two superfamilies of animal neurotransmitter receptors via distinct evolutionary mechanisms. *Mol. Biol. Evol.* 18, 1417–1420. doi: 10.1093/oxfordjournals.molbev.a003926

UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212. doi: 10.1093/nar/gku989

Van Beneden, E. (1876). Recherches sur les dicyémides, survivants actuels d'un embranchement des mésozoaires. *Bull. Acad. Belg. Cl. Sci.* 41, 1160–1205.

Vortsepneva, E., Tzetlin, A., Purschke, G., Mugue, N., Haß-Cordes, E., and Zhadan, A. (2008). The parasitic polychaete known as *Asetocalamyzas laonicola* (Calamyzidae) is in fact the dwarf male of the spionid *Scolelepis laonicola* (comb. nov.). *Invert. Biol.* 127, 403–416. doi: 10.1111/j.1744-7410.2008.00137.x

Watanabe, K. I., Bessho, Y., Kawasaki, M., and Hori, H. (1999). Mitochondrial genes are found on minicircle DNA molecules in the mesozoan animal *Dicyema*. *J. Mol. Biol.* 286, 645–650. doi: 10.1006/jmbi.1998.2523

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., et al. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* doi: 10.1093/molbev/msx319 [Epub ahead of print].

Weigert, A., Helm, C., Meyer, M., Nickel, B., Arendt, D., Hausdorf, B., et al. (2014). Illuminating the base of the annelid tree using transcriptomics. *Mol. Biol. Evol.* 31, 1391–1401. doi: 10.1093/molbev/msu080

Worsaae, K., and Rouse, G. W. (2010). The simplicity of males: dwarf males of four species of *Osedax* (Siboglinidae, Annelida) investigated by confocal laser scanning microscopy. *J. Morphol.* 271, 127–142. doi: 10.1002/jmor.10786

Zamanian, M., Kimber, M. J., McVeigh, P., Carlson, S. A., Maule, A. G., and Day, T. A. (2011). The repertoire of G protein-coupled receptors in the human parasite Schistosoma mansoni and the model organism Schmidtea mediterranea. *BMC Genomics* 12:596. doi: 10.1186/1471-2164-12-596

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simão, F. A., Ioannidis, P., et al. (2017). OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* 45, D744–D749. doi: 10.1093/nar/gkw1119

Zhong, M., Hansen, B., Nesnidal, M., Golombek, A., Halanych, K. M., and Struck, T. H. (2011). Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. *BMC Evol. Biol.* 11:369. doi: 10.1186/1471-2148-11-369

Zhong, Y. F., and Holland, P. W. (2011). HomeoDB2: functional expansion of a comparative homeobox gene database for evolutionary developmental biology. *Evol. Dev.* 13, 567–568. doi: 10.1111/j.1525-142X.2011.00513.x

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., and Siebert, P. D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* 30, 892–897. doi: 10.2144/01304pf02

Zhulidov, P. A., Bogdanova, E. A., Shcheglov, A. S., Vagner, L. L., Khaspekov, G. L., Kozhemyako, et al. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* 32:e37. doi: 10.1093/nar/gnh031

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

PERSPECTIVE
published: 23 August 2019
doi: 10.3389/fgene.2019.00759

Searching for Signatures of Cold Climate Adaptation in *TRPM8* Gene in Populations of East Asian Ancestry

Alexander V. Igoshin[1], Konstantin V. Gunbin[2,3,4], Nikolay S. Yudin[3,5] and Mikhail I. Voevoda[6]*

[1] Sector of the Genetics of Industrial Microorganisms, The Federal Research Center Institute of Cytology and Genetics, The Siberian Branch, The Russian Academy of Sciences, Novosibirsk, Russia, [2] Center of Brain Neurobiology and Neurogenetics, The Federal Research Center Institute of Cytology and Genetics, The Siberian Branch, The Russian Academy of Sciences, Novosibirsk, Russia, [3] V. Zelman Institute for Medicine and Psychology Novosibirsk State University, Novosibirsk, Russia, [4] Center for Mitochondrial Functional Genomics, Institute of Living Systems, Immanuel Kant Baltic Federal University, Kaliningrad, Russia, [5] Laboratory of Livestock Molecular Genetics and Breeding, The Federal Research Center Institute of Cytology and Genetics, The Siberian Branch, The Russian Academy of Sciences, Novosibirsk, Russia, [6] Laboratory of Human Molecular Genetics, The Federal Research Center Institute of Cytology and Genetics, The Siberian Branch, The Russian Academy of Sciences, Novosibirsk, Russia

Dispersal of *Homo sapiens* across the globe during the last 200,000 years was accompanied by adaptation to local climatic conditions, with severe winter temperatures being probably one of the most significant selective forces. The *TRPM8* gene codes for a cold-sensing ion channel, and adaptation to low temperatures is the major determinant of its molecular evolution. Here, our aim was to search for signatures of cold climate adaptation in *TRPM8* gene using a combined data set of 19 populations of East Asian ancestry from the 1000 Genomes Project and Human Genome Diversity Project. As a result, out of a total of 60 markers under study, none showed significant association with the average winter temperatures at the locations of the studied populations considering the multiple testing thresholds. This might suggest that the principal mode of *TRPM8* evolution may be different from widespread models, where adaptive alleles are additive, dominant or recessive, at least in populations with the predominant East Asian component. For example, evolution by means of selectively preferable epistatic interactions among amino acids may have taken place. Despite the lack of strong signals of association, however, a very promising single nucleotide polymorphism (SNP) was found. The SNP rs7577262 is considered the best candidate based on its allelic correlations with winter temperatures, signatures of selective sweep and physiological evidences. The second top SNP, rs17862920, may participate in adaptation as well. Additionally, to assist in interpreting the nominal associations, the other markers reached, we performed SNP prioritization based on functional evidences found in literature and on evolutionary conservativeness.

Keywords: TRPM8, environmental correlation analysis, SNP, cold adaptation, East Asian ancestry

OPEN ACCESS

Edited by:
Ancha Baranova, George Mason University, United States

Reviewed by:
Margarida Matos, University of Lisbon, Portugal
Toni Gossmann, University of Sheffield, United Kingdom

****Correspondence:***
Alexander V. Igoshin
igoshin@bionet.nsc.ru

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *30 October 2018*
Accepted: *17 July 2019*
Published: *23 August 2019*

Citation:
Igoshin AV, Gunbin KV, Yudin NS and Voevoda MI (2019) Searching for Signatures of Cold Climate Adaptation in TRPM8 Gene in Populations of East Asian Ancestry. Front. Genet. 10:759. doi: 10.3389/fgene.2019.00759

INTRODUCTION

Recent paleoanthropological evidences show the presence of anatomically modern humans in Africa as early as 300 kya (Hublin et al., 2017), with the earliest known "Out of Africa" migration event dating back to 200 kya (Hershkovitz et al., 2018). Dispersal of *Homo sapiens* across the globe during the last 200,000 years was accompanied by adaptation to local environments. Spatial variations in selective

pressures have ultimately led to observable geographic distribution of many physiological and anatomical traits in present-day humans. For example, the low level of UV radiation at higher latitudes is now considered to be the major cause of evolution of depigmented skin (Jablonski and Chaplin, 2010).

Since the mid-2000s, there has been significant progress in genotyping technologies, followed by publicly available databases of human genetic variation. This circumstance helped the population geneticists to discover signatures of human local adaptation from genome-wide genotyping data. Human microevolution driven by the action of low temperatures has long been attracting attention of the scientific community. Now, a number of studies have been dedicated to this issue both at the genome level (e.g., Hancock et al., 2011a; Cardona et al., 2014; Valverde et al., 2015) and at the level of selected regions or genes (Hancock et al., 2008; Ohashi et al., 2011; Hancock et al., 2011b; Sazzini et al., 2014; Quagliarello et al., 2017).

Probably the best-known gene in terms of its possible role in adaptation to cold climate is *TRPM8* located on human chromosome 2. This gene codes for ion channel functioning as a thermal sensor, detecting temperatures in the range from 15 to 30°C (Fernández et al., 2011). There are evidences supporting its physiological role in thermoregulation, and in fact, TRPM8 is the only well-established cold receptor in mammals (Bautista et al., 2007; Colburn et al., 2007; Dhaka et al., 2007). Besides these, there are data on associations of its single nucleotide polymorphisms (SNPs) with sensitivity to cold (Kozyreva et al., 2011), the respiratory system response to cooling (Kozyreva et al., 2014), blood lipids, and anthropometric parameters in humans (Potapova et al., 2014). The *TRPM8* gene was suggested to underlie genetic adaptation to cold in ground squirrel and hamster (Matos-Cruz et al., 2017), sheep (Fariello et al., 2014; Liu et al., 2016), and humans (Cardona et al., 2014; Key et al., 2018). According to modern views, adaptation to low temperatures is the major determinant of *TRPM8* molecular evolution (Myers et al., 2009; Majhi et al., 2015).

To our knowledge, a study by Key and colleagues (2018) is the only one to use environmental data to search for signatures of cold climate adaptation in the *TRPM8* gene. The authors used latitudes and annual average temperatures at the locations of the populations of the Old World as predictors for SNP allele frequency distributions. They found evidences that SNP rs10166942 had undergone climate-mediated selection, which raised its derived allele frequency from south to north.

In our opinion, focusing on closely related populations is preferable to using large population sets for the following reasons. First, the ability to survive in a severely cold climate is supposed to be highly polygenic, as many biological processes like vasoconstriction, nonshivering thermogenesis, regulation of adipocyte differentiation, and thermoception are expected to be involved. It is known that the adaptation to cold can be associated with quite different genetic bases (Yudin et al., 2017). Because different branches of *Homo sapiens* are likely to have had distinct genetic background before and during the process of climate-driven selection, it is possible that in phylogenetically distant groups, adaptation may have recruited different genes. Second, even in the case when selection is acting on the same gene, variants involved in adaptation may differ in different branches. Our supposition is supported by the example of variants associated with lactase persistence. Thus, within European populations, the activity of the lactase enzyme in adulthood is connected with the C/T-13910 variant in the enhancer region of the *LCT* gene, whereas in sub-Saharan Africa, this trait is mainly correlated with the presence of the G/C-14010 mutation (Tishkoff et al., 2007). Therefore, it would be sensible to search for microevolution in clusters of related populations.

For the above reasons, the aim of this study was to search for signatures of adaptation to low temperatures in the *TRPM8* gene under various null hypotheses of population structure and dynamics using a combined data set of 19 populations of the East Asian ancestry from Human Genome Diversity Project (HGDP) and 1000 Genomes (1000G) Project with the assistance of environmental correlation analysis techniques. Locations of chosen populations are characterized by a large range of average winter temperatures (−37–+27°C), implying substantial differences in selection pressures.

MATERIALS AND METHODS

Genotypic and Environmental Data

In this study, we used genotypic data on 656 individuals from 19 HGDP (Cann et al., 2002) and 1000G Project (1000 Genomes Project Consortium et al., 2010) populations (**Supplementary Table 1**) having predominantly the East Asian genetic component. Data on SNPs belonging to *TRPM8* gene were obtained from NCBI dbSNP (https://www.ncbi.nlm.nih.gov/snp/), resulting in 60 polymorphic markers (minor allele frequency >0.01) being at the intersection of HGDP and 1000G sets. The missing genotypes in HGDP data were imputed using fastPHASE v.1.4.8 software (Scheet and Stephens, 2006) with default parameters. Genotypic information from HGDP and VCF formats was combined by using a self-made Python 3 script, so that inconsistency of DNA strands between databases (if that was the case) was resolved using 1000G VCF as a reference. Besides *TRPM8* SNPs, we used 5 Mb regions upstream and downstream this gene (1,309 markers at $r^2 < 0.7$) to infer the phylogenetic tree used in PGLS, to estimate the covariance matrix used in Bayenv2-BLM and Bayenv2-SRC, to correct for background levels of the population structure in LFMM and to make population inferences in BayScEnv more precise. Also, this set of SNPs was used to construct a null distribution for empirical p-value calculation. Information on latitudes and longitudes for 1000G populations was taken from Key et al. (2018). Latitudes and longitudes for HGDP populations were taken from Cann et al. (2002), and average winter temperature values were obtained from ClimateCharts.net database (https://climatecharts.net) using the corresponding coordinates. We believe the average winter temperature to be a more pertinent predictor for distribution of cold-adaptive alleles than the annual average temperature, as regions with a continental climate may have cold winters and hot summers.

Construction of a Phylogenetic Tree and Statistical Analysis

As our sample consisted of phylogenetically close populations, we first performed conventional Spearman's rank correlation test not accounting for the sample structure (Spearman, 1904) at the population (i.e., using allele frequencies) and individual's (i.e., using allele dosages) levels.

PGLS analysis (R package 'ape,' Paradis et al., 2004) was carried out at the individual's level using the simplest Brownian motion model. We selected this type of analysis because it is an opposite alternative (phyletic evolution) to the conventional Spearman's rank correlation test. The phylogenetic tree used in this test was reconstructed with IQ-TREE v. 1.5.5 subprogram ModelFinder (Kalyaanamoorthy et al., 2017) based on the best nucleotide substitution model.

Our primary aim was to test the association between the climatic factor and allele frequencies. For this purpose, we chose two independent approaches: the Bayesian linear model from Bayenv2 (further referred to as Bayenv2-BLM) software (Günther and Coop, 2013) and LFMM (Frichot et al., 2013), each building a regression model relating allele frequencies to environmental values. To minimize the problem of false-positive associations between allele frequencies and environment because of the population structure, the above methods take into account allele frequency correlations across populations while performing the analysis by various ways.

In addition, we used Spearman's rank correlation test from Bayenv2 (further referred to as Bayenv2-SRC) that uses allele frequencies standardized to have no covariance. It is less powerful than Bayenv2-BLM but more robust to outliers and can detect monotonic relationships.

BayScEnv test (de Villemereuil and Gaggiotti, 2015) was used as an alternative to Bayenv2-BLM and LFMM. This method assumes that all populations are independent and exchange genes through the limited migrant pool; it includes a locus-specific effect unrelated to the environmental variable, taking into consideration locus-specific deviations from a neutral model. BayScEnv software was also used to calculate F_{st} distances for each locus averaged over populations (**Supplementary Table 2**).

In addition to the correlation techniques, we tried the XP-CLR test (Chen et al., 2010) as a complementary approach. This test is designed for detecting selective sweeps on the basis of joint modeling of the multilocus allele frequency differentiation between two populations. The method does not require information on the ancestral/derived status at each SNP (Chen et al., 2010; Vatsiou et al., 2016).

For more details on phylogenetic tree construction and statistical analysis, please see the Supplementary Material.

SNP Prioritization

We prioritized SNPs based on three types of evidences found in literature, "association with trait relevant to survival in a cold climate," "evidences for cold-mediated selection," and "association with any other phenotype or risk." These categories were given weights 3, 2, and 1, respectively, and a score for each SNP was summarized (**Supplementary Table 3**). The key assumption behind this prioritization is that because of the pleiotropic nature of *TRPM8* gene, allelic substitutions having any functional manifestation may potentially have more chances of affecting survival in cold climate conditions than those not having any known effects.

Additionally, we obtained the PhyloP100way vertebrate conservation score for each SNP (**Supplementary Table 4**) from UCSC Genome Browser (Casper et al., 2018). Currently, it is commonly thought that the genetic drift plays a minor role in the evolution of conservative sites, and relatively rare allele replacements occurring therein are mostly driven by positive selection (Andolfatto, 2005; Cai et al., 2009; Halligan et al., 2011; Bazykin and Kondrashov, 2012). Therefore, significant allelic correlation with environmental gradient supported by a high conservation score promises to be the true sign of local adaptation.

RESULTS

Contrary to what we had expected, only three SNPs out of a total of 60 markers under study showed nominally significant association with the average winter temperatures at the locations of the studied populations by any two types of analysis (the results of tests carried out using default/recommended parameters are shown in **Table 1**; for full results,

TABLE 1 | Genic and upstream *TRPM8* variants showing nominally significant (in bold) association in at least two correlation tests with default/recommended parameters (K = 2 for LFMM and pi = 0.1/p = 0.5 for BayScEnv).

Methods and scores	rs7577262	rs17862920	rs6723922	rs11682848
SNP evidence score	**NA***	**5**	**1**	**1**
PhyloP100way conservation score	**−1.7724**	**−0.1171**	**−2.2644**	**−5.512**
Bayenv2-BLM empirical p-value	**0.0374**	0.06	0.064	**0.0069**
PGLS p-value	**0.0331**	**0.04**	0.101	0.121
Bayenv2-SRC empirical p-value	**0.0252**	**0.0328**	0.075	**0.0145**
SRC individual-based empirical p-value	**0.0267**	0.054	0.096	**0.0481**
SRC population-based empirical p-value	**0.042**	0.149	**0.0428**	0.102
LFMM p-value	**0.074**	0.127	**0.005**	**0.0013**
BayScEnv posterior error probability/q-value	0.97/0.829	0.984/0.903	0.993/0.957	0.795/0.525
BayScEnv empirical p-value	**0.042**	0.084	0.208	**0.0115**

**The SNP evidence score was not calculated for rs7577262 because of a high level of confidence in its adaptive role.*

see **Supplementary Tables 5 and 6**). When considering the multiple testing threshold, however, none of them is significant (adjusted p values not shown).

SNP rs11682848 was previously reported as associated with the prognosis of colorectal cancer (Walther, 2010). Interestingly, such connection of climate-associated loci with cancer has already been noticed by other researchers (Hancock et al., 2011a). Furthermore, it has been recently shown by combination of 247 genome-wide association studies that cold selected genes are enriched with cancer-associated genes (Voskarides, 2018). Ironically, SNP rs11682848 has the lowest conservation score among 60 markers under study. This means that either rs11682848 is being a false-positive finding or being linked to some functional variant.

SNP rs17862920 has evidences of associations with migraine susceptibility (Freilinger et al., 2012; Meng et al., 2018). The rs17862920-C allele predisposing to migraine is more prevalent in northern latitudes. Also, rs17862920 has been shown to be associated with sensing cold pain in Finnish and Norwegian individuals, with C allele carriers being more susceptible (Kaunisto et al., 2013). Migraine has been reported to be related to increased pain perception of nonnoxious cold temperatures (Burstein et al., 2000). Unlike rs11682848, PhyloP100way conservation score for rs17862920 is more promising and has a rank of 20/60 while still being negative. It is possible that allele substitution in rs17862920 has a functional effect. Thus, rs17862920 was predicted to regulate *TRPM8* transcription by TFsearch and GoldenPath in F-SNP bioinformatics tool (Ghosh et al., 2013).

As for rs6723922, this SNP is a genetic risk factor for severe cutaneous adverse drug reactions (Park et al., 2018). One could hypothesize that there is a certain mechanism underlying both the altered cold sensation and the increased cutaneous susceptibility to chemicals. It could be no surprise given that the TRPM8 channel is activated by a variety of chemical ligands (Beccari et al., 2017). Like rs11682848, SNP rs6723922 has a low conservation score (the rank of 55/60), implying conclusions for this marker similar to those for rs11682848.

The results of the XP-CLR test are more encouraging (**Supplementary Figure S1**). It appears that there is a pronounced trend for several pairs of populations to show the signature of a selective sweep 10 ± 6 Kb upstream from the *TRPM8* gene. The direction of selection in this region is seen when reversing tested and reference populations in pairs (e.g., compare "JPT vs. KHV" and "KHV vs. JPT"). The strongest XP-CLR peaks within this putative sweep are mainly located near rs10929317 and rs7577262 SNP loci. The former was removed from the analysis because of high LD with rs17862920: $r^2 = 0.966/D' = 0.995$ in East Asian populations (LDlink tool; Machiela and Chanock, 2015) and is therefore expected to be as significant as rs17862920. The latter was used in the control set of 1,309 markers. Surprisingly, this SNP demonstrates significant association with our climatic variable in almost all of the correlation tests (**Table 1**). Also, rs7577262 has been reported to be associated with susceptibility to migraine (Anttila et al., 2013) and blood pressure response to the cold pressor test (He et al., 2013).

DISCUSSION

A variety of facts have led us to think of *TRPM8* gene as being under intense positive selection. We expected that the large amount of SNPs in *TRPM8* would demonstrate strong signals because of being under selection immediately or being linked to some causal variants. However, this is not the case. Furthermore, SNPs detected do not pass the corrected threshold, considering multiple testing. Among possible explanations are the following hypotheses:

1. The predominant mode of *TRPM8* evolution may be different from the widespread models exploited by our tests, where adaptation is assumed to be mediated by selecting alleles with additive or at least recessive/dominant trait coding. However, it is suggested that epistatic interactions may play a role in the evolution of thermoTRP channels (Saito and Tominaga, 2017). Therefore, conformational epistasis-based evolution, where some epistatic interactions among amino acids are preferred, might have resulted in the inability of approaches we used to detect strong signatures of selection in the *TRPM8* gene.
2. The level of allele frequency variation (see **Supplementary Table 7** for minor allele frequency distributions) in the populations studied is not sufficient to robustly discriminate the loci under selection. Thus, the averaged F_{st} distances for populations used by Key et al. (2018) are much higher than those for our data set (**Supplementary Table 2**).
3. Only 5 out of 19 locations of populations from our data set have the average winter temperature below −10°C. It is possible that the underrepresentation of northern populations in this study might lead to insufficient signal strength. Further accumulation of open access data on genetic variation in the north would help in detecting loci under selection.

Despite the lack of strong signals of association, however, a very promising candidate SNP was found. SNP rs7577262 is 7.1 kb upstream of the transcription start site for *TRPM8* mRNA, implying its possible involvement in transcriptional regulation. In addition to correlations and signatures of sweep, physiological data contribute equally to the evidences in favor of selection acting on rs7577262. The rs7577262-G allele is associated with a higher blood pressure response to the cold pressor test (He et al., 2013). It is known that the blood pressure response to the cold pressor test primarily stems from alpha-adrenergically mediated peripheral vasoconstriction (Leppäluoto and Hassi, 1991; Larra et al., 2015), which is, in turn, one of the basic mechanisms of cold

adaptation (Daanen and Lichtenbelt, 2016). Given that this allele is more prevalent in northern latitudes (**Figure 1**), its adaptive role may be assumed.

Another SNP, rs17862920, is linked with rs7577262 ($r^2 = 0.59$ in East Asians). Probably, this is the reason for the correlation the former demonstrates. Both SNPs are risk loci for migraine. At the same time, it has been mentioned above that rs17862920 is associated with sensing cold pain in Finnish and Norwegian individuals, with C allele carriers (more prevalent in northern latitudes) being more susceptible. It can be assumed that both loci are independently involved in adaptation to low temperatures. In that case, however, the adaptive role of rs17862920-C allele is hard to explain. The possible mechanism of differential survival might be avoidance of potentially lethal hypothermia by those harboring C allele.

As for rs6723922 and rs11682848 loci, none of them shows any sign of selective sweep in the XPCLR test. Probably, those are false positives, or at least, linked to an unobserved variant under selection. It is also worth noting that SNP evidence scores for these loci are quite low.

In addition to the search for signatures of selection, we would like to note some details on the BayScEnv test not published anywhere (as far as we know).

Changing model parameters drastically affects the output in BayScEnv (see **Supplementary Table 6**). For example, significant results (q value <0.05) were obtained when using model parameters pi = 0.5/p = 0.1 (SNP rs11682848 being significant) or pi = 0.9/p = 0.1 (18/60 SNPs being significant). At the same time, empirical p values are more stable. Thus, we suggest using them in hypothesis-driven studies (with default model parameters) of local adaptation and choosing the significance threshold based on expert's opinions rather than relying on FDR outputs.

Counterintuitively, a reduction in the number of tests in BayScEnv does not lead to a greater number of statistically significant FDR outputs (the posterior error probability and the q value). Furthermore, in our case, given parameters pi = 0.9/p = 0.1, 18 out of a total of 60 SNPs reach significance level when analyzing 1,369 markers, whereas none is significant when using 60 SNPs. This discrepancy might be explained by less precise constructing a null model of population structure.

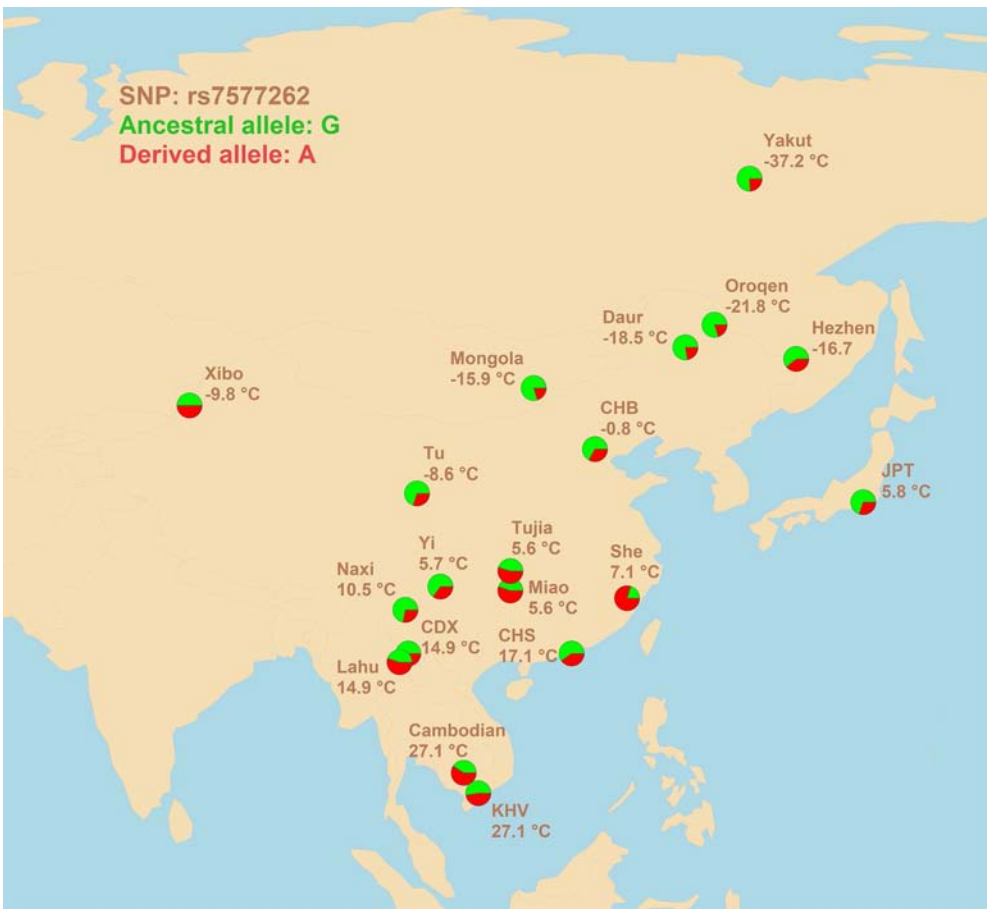


FIGURE 1 | Geographic distribution of allele frequencies for rs7577262 polymorphism in populations of East Asian ancestry. Average winter temperatures at the locations of the populations studied are shown. 1000 Genomes population: CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan; CDX, Chinese Dai in Xishuangbanna, China; CHS, Han Chinese South, China; KHV, Kinh in Ho Chi Minh City, Vietnam.

CONCLUSIONS

Several lines of evidence point to possible involvement of rs7577262 in cold adaptation. This SNP is considered the best candidate based on its allelic correlations with winter temperatures, signatures of selective sweep and physiological evidences. The second top SNP, rs17862920, may participate in adaptation as well. As for rs6723922 and rs11682848 loci, these appear to be false positives or at least linked to some unobserved selected variant.

AUTHOR CONTRIBUTIONS

NY and MV conceived the project. KG supervised the project. AI and KG processed and analyzed the data. AI, KG, and NY drafted the manuscript.

FUNDING

This study was supported by budget from project No. 0324-2019-0041 of the Federal Research Center «Institute of Cytology and Genetics» SB RAS (ICG SB RAS).

ACKNOWLEDGMENTS

The Common Use Center «Bioinformatics» (ICG SB RAS) is gratefully acknowledged for providing computer facilities.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00759/full#supplementary-material

REFERENCES

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., et al. (2010). A map of human genome variation from population-scale sequencing. *Nature* 467, 1061–1073. doi: 10.1038/nature09534

Andolfatto, P. (2005). Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437, 1149–1152. doi: 10.1038/nature04107

Anttila, V., Winsvold, B. S., Gormley, P., Kurth, T., Bettella, F., McMahon, G., et al. (2013). Genome-wide meta-analysis identifies new susceptibility loci for migraine. *Nat. Genet.* 45, 912–917. doi: 10.1038/ng.2676

Bautista, D. M., Siemens, J., Glazer, J. M., Tsuruda, P. R., Basbaum, A. I., Stucky, C. L., et al. (2007). The menthol receptor TRPM8 is the principal detector of environmental cold. *Nature* 448, 204–208. doi: 10.1038/nature05910

Bazykin, G. A., and Kondrashov, A. S. (2012). Major role of positive selection in the evolution of conservative segments of Drosophila proteins. *Proc. Biol. Sci.* 279, 3409–3417. doi: 10.1098/rspb.2012.0776

Beccari, A. R., Gemei, M., Lo Monte, M., Menegatti, N., Fanton, M., Pedretti, A., et al. (2017). Novel selective, potent naphthyl TRPM8 antagonists identified through a combined ligand- and structure-based virtual screening approach. *Sci. Rep.* 7, 10999. doi: 10.1038/s41598-017-11194-0

Burstein, R., Yarnitsky, D., Goor-Aryeh, I., Ransil, B. J., and Bajwa, Z. H. (2000). An association between migraine and cutaneous allodynia. *Ann. Neurol.* 47, 614–624. doi: 10.1002/1531-8249(200005)47:5<614::AID-ANA9>3.0.CO;2-N

Cai, J. J., Macpherson, J. M., Sella, G., and Petrov, D. A. (2009). Pervasive hitchhiking at coding and regulatory sites in humans. *PLoS Genet.* 5, e1000336. doi: 10.1371/journal.pgen.1000336

Cann, H. M., de Toma, C., Cazes, L., Legrand, M. F., Morel, V., Piouffre, L., et al. (2002). A human genome diversity cell line panel. *Science* 296, 261–262. doi: 10.1126/science.296.5566.261b

Cardona, A., Pagani, L., Antao, T., Lawson, D. J., Eichstaedt, C. A., Yngvadottir, B., et al. (2014). Genome-wide analysis of cold adaptation in indigenous Siberian populations. *PLoS One* 9, e98076. doi: 10.1371/journal.pone.0098076

Casper, J., Zweig, A. S., Villarreal, C., Tyner, C., Speir, M. L., Rosenbloom, K. R., et al. (2018). The UCSC genome browser database: 2018 update. *Nucleic Acids Res.* 46, D762–D769. doi: 10.1093/nar/gkx1020

Chasman, D. I., Schürks, M., Anttila, V., de Vries, B., Schminke, U., Launer, L. J., et al. (2011). Genome-wide association study reveals three susceptibility loci for common migraine in the general population. *Nat. Genet.* 43, 695–698. doi: 10.1038/ng.856

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.* 20, 393–402. doi: 10.1101/gr.100545.109

Colburn, R. W., Lubin, M. L., Stone, D. J., Jr., Wang, Y., Lawrence, D., D'Andrea, M. R., et al. (2007). Attenuated cold sensitivity in TRPM8 null mice. *Neuron* 54, 379–386. doi: 10.1016/j.neuron.2007.04.017

Daanen, H. A., and Lichtenbelt, W. D. (2016). Human whole body cold adaptation. *Temperature* 3, 104–118. doi: 10.1080/23328940.2015.1135688

de Villemereuil, P., and Gaggiotti, O. E. (2015). A new FST-based method to uncover local adaptation using environmental variables. *Methods Ecol. Evol.* 6, 1248–1258. doi: 10.1111/2041-210X.12418

Dhaka, A., Murray, A. N., Mathur, J., Earley, T. J., Petrus, M. J., and Patapoutian, A. (2007). TRPM8 is required for cold sensation in mice. *Neuron* 54, 371–378. doi: 10.1016/j.neuron.2007.02.024

Fariello, M. I., Servin, B., Tosser-Klopp, G., Rupp, R., Moreno, C., International Sheep Genomics Consortium, et al. (2014). Selection signatures in worldwide sheep populations. *PLoS One* 9, e103813. doi: 10.1371/journal.pone.0103813

Fernández, J. A., Skryma, R., Bidaux, G., Magleby, K. L., Scholfield, C. N., McGeown, J. G., et al. (2011). Voltage- and cold-dependent gating of single TRPM8 ion channels. *J. Gen. Physiol.* 137, 173–195. doi: 10.1085/jgp.201010498

Freilinger, T., Anttila, V., de Vries, B., Malik, R., Kallela, M., Terwindt, G. M., et al. (2012). Genome-wide association analysis identifies susceptibility loci for migraine without aura. *Nat. Genet.* 44, 777–782. doi: 10.1038/ng.2307

Frichot, E., Schoville, S. D., Bouchard, G., and François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* 30, 1687–1699. doi: 10.1093/molbev/mst063

Ghosh, J., Pradhan, S., and Mittal, B. (2013). Genome-wide-associated variants in migraine susceptibility: a replication study from North India. *Headache* 53, 1583–1594. doi: 10.1111/head.12240

Günther, T., and Coop, G. (2013). Robust identification of local adaptation from allele frequencies. *Genetics* 195, 205–220. doi: 10.1534/genetics.113.152462

Halligan, D. L., Oliver, F., Guthrie, J., Stemshorn, K. C., Harr, B., and Keightley, P. D. (2011). Positive and negative selection in murine ultra-conserved noncoding elements. *Mol. Biol. Evol.* 28, 2651–2660. doi: 10.1093/molbev/msr093

Hancock, A. M., Clark, V. J., Qian, Y., and Di Rienzo, A. (2011b). Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. *Mol. Biol. Evol.* 28, 601–614. doi: 10.1093/molbev/msq228

Hancock, A. M., Witonsky, D. B., Alkorta-Aranburu, G., Beall, C. M., Gebremedhin, A., Sukernik, R., et al. (2011a). Adaptations to climate-mediated selective pressures in humans. *PLoS Genet.* 7, e1001375. doi: 10.1371/journal.pgen.1001375

Hancock, A. M., Witonsky, D. B., Gordon, A. S., Eshel, G., Pritchard, J. K., Coop, G., et al. (2008). Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet.* 4, e32. doi: 10.1371/journal.pgen.0040032

He, J., Kelly, T. N., Zhao, Q., Li, H., Huang, J., Wang, L., et al. (2013). Genome-wide association study identifies 8 novel loci associated with blood pressure responses to interventions in Han Chinese. *Circ. Cardiovasc. Genet.* 6, 598–607. doi: 10.1161/CIRCGENETICS.113.000307

Hershkovitz, I., Weber, G. W., Quam, R., Duval, M., Grün, R., Kinsley, L., et al. (2018). The earliest modern humans outside Africa. *Science* 359, 456–459. doi: 10.1126/science.aap8369

Hublin, J. J., Ben-Ncer, A., Bailey, S. E., Freidline, S. E., Neubauer, S., Skinner, M. M., et al. (2017). New fossils from Jebel Irhoud, Morocco and the pan-African origin of homo sapiens. *Nature* 546, 289–292. doi: 10.1038/nature22336

Jablonski, N. G., and Chaplin, G. (2010). Human skin pigmentation as an adaptation to UV radiation. *Proc. Natl. Acad. Sci. U. S. A.* 107 Suppl 2, 8962–8968. doi: 10.1073/pnas.0914628107

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., and Jermiin, L. S. (2017). Model finder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi: 10.1038/nmeth.4285

Kaunisto, M., Holmström, E., Anttila, V., Kallela, M., Hämäläinen, E., Stubhaug, A. et al. (2013). TRPM8 variants that protect from migraine make individuals less susceptible to cold pain. In: *Presented at the 63rd Annual Meeting of The American Society of Human Genetics*. http://www.ashg.org/2013meeting/abstracts/fulltext/f130121596.htm.

Key, F. M., Abdul-Aziz, M. A., Mundry, R., Peter, B. M., Sekar, A., D'Amato, M., et al. (2018). Human local adaptation of the TRPM8 cold receptor along a latitudinal cline. *PLoS Genet.* 14, e1007298. doi: 10.1371/journal.pgen.1007298

Kozyreva, T. V., Tkachenko, E. Ya., Potapova, T. A., and Voevoda, M. I. (2014). Respiratory system response to cooling in subjects with single nucleotide polymorphism rs11562975 in gene of thermosensitive TRPM8 ion channel. *Fiziol Cheloveka* 40, 94–98. doi: 10.1134/S0362119714020108

Kozyreva, T. V., Tkachenko, E. Ya., Potapova, T. A., Romashchenko, A. G., and Voevoda, M. I. (2011). Relationship of single-nucleotide polymorphism rs11562975 in thermo-sensitive ion channel TRPM8 gene with human sensitivity to cold and menthol. *Fiziol Cheloveka* 37, 71–76. doi: 10.1134/S0362119711020101

Larra, M. F., Schilling, T. M., Röhrig, P., and Schächinger, H. (2015). Enhanced stress response by a bilateral feet compared to a unilateral hand Cold Pressor Test. *Stress* 18, 589–596. doi: 10.3109/10253890.2015.1053452

Leppäluoto, J., and Hassi, J. (1991). Human physiological adaptations to the arctic climate. *Arctic* 44, 139–145. doi: 10.14430/arctic1530

Liu, Z., Ji, Z., Wang, G., Chao, T., Hou, L., and Wang, J. (2016). Genome-wide analysis reveals signatures of selection for important traits in domestic sheep from different ecoregions. *BMC Genomics* 17, 863. doi: 10.1186/s12864-016-3212-2

Machiela, M. J., and Chanock, S. J. (2015). LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* 31, 3555–3557. doi: 10.1093/bioinformatics/btv402

Majhi, R. K., Saha, S., Kumar, A., Ghosh, A., Swain, N., Goswami, L., et al. (2015). Expression of temperature-sensitive ion channel TRPM8 in sperm cells correlates with vertebrate evolution. *PeerJ.* 3, e1310. doi: 10.7717/peerj.1310

Matos-Cruz, V., Schneider, E. R., Mastrotto, M., Merriman, D. K., Bagriantsev, S. N., and Gracheva, E. O. (2017). Molecular prerequisites for diminished cold sensitivity in ground squirrels and hamsters. *Cell Rep.* 21, 3329–3337. doi: 10.1016/j.celrep.2017.11.083

Meng, W., Adams, M. J., Hebert, H. L., Deary, I. J., McIntosh, A. M., and Smith, B. H. (2018). A genome-wide association study finds genetic associations with broadly-defined headache in UK Biobank (N = 223,773). *E Bio. Med.* 28, 180–186. doi: 10.1016/j.ebiom.2018.01.023

Myers, B. R., Sigal, Y. M., and Julius, D. (2009). Evolution of thermal response properties in a cold-activated TRP channel. *PLoS One* 4, e5741. doi: 10.1371/journal.pone.0005741

Ohashi, J., Naka, I., and Tsuchiya, N. (2011). The impact of natural selection on an ABCC11 SNP determining earwax type. *Mol. Biol. Evol.* 28, 849–857. doi: 10.1093/molbev/msq264

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in r language. *Bioinformatics* 20, 289–290. doi: 10.1093/bioinformatics/btg412

Park, H. W., Kim, S. H., Chang, Y. S., Kim, S. H., Jee, Y. K., Lee, A. Y., et al. (2018). The Fas signaling pathway is a common genetic risk factor for severe cutaneous drug adverse reactions across diverse drugs. *Allergy Asthma Immunol. Res.* 10, 555–561. doi: 10.4168/aair.2018.10.5.555

Potapova, T. A., Babenko, V. N., Kobzev, V. F., Romashchenko, A. G., Maksimov, V. N., and Voevoda, M. I. (2014). Associations of cold receptor TRPM8 gene single nucleotide polymorphism with blood lipids and anthropometric parameters in Russian population. *Bull. Exp. Biol. Med.* 157, 757–761. doi: 10.1007/s10517-014-2660-4

Quagliarello, A., De Fanti, S., Giuliani, C., Abondio, P., Serventi, P., Sarno, S., et al. (2017). Multiple selective events at the PRDM16 functional pathway shaped adaptation of western European populations to different climate conditions. *J. Anthropol. Sci.* 95, 235–247. doi: 10.4436/JASS.95011

Saito, S., and Tominaga, M. (2017). Functional diversity and evolutionary dynamics of thermoTRP channels. *Cell Calcium* 57, 214–221. doi: 10.1016/j.ceca.2014.12.001

Sazzini, M., Schiavo, G., De Fanti, S., Martelli, P. L., Casadio, R., and Luiselli, D. (2014). Searching for signatures of cold adaptations in modern and archaic humans: hints from the brown adipose tissue genes. *Heredity* 113, 259–267. doi: 10.1038/hdy.2014.24

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78, 629–644. doi: 10.1086/502802

Spearman, C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101. doi: 10.2307/1412159

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.* 39, 31–40. doi: 10.1038/ng1946

Valverde, G., Zhou, H., Lippold, S., de Filippo, C., Tang, K., LópezHerráez, D., et al. (2015). A novel candidate region for genetic adaptation to high altitude in Andean populations. *PLoS One* 10, e0125444. doi: 10.1371/journal.pone.0125444

Vatsiou, A. I., Bazin, E., and Gaggiotti, O. E. (2016). Detection of selective sweeps in structured populations: a comparison of recent methods. *Mol. Ecol.* 25, 89–103. doi: 10.1111/mec.13360

Voskarides, K. (2018). Combination of 247 genome-wide association studies reveals high cancer risk as a result of evolutionary adaptation. *Mol. Biol. Evol.* 35, 473–485. doi: 10.1093/molbev/msx305

Walther, A. (2010). Germline determinants of outcome and risk in colorectal cancer. Ph.D. thesis. London: University College London.

Yudin, N. S., Larkin, D. M., and Ignatieva, E. V. (2017). A compendium and functional characterization of mammalian genes involved in adaptation to Arctic or Antarctic environments. *BMC Genet.* 18 (Suppl 1), 111. doi: 10.1186/s12863-017-0580-9

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

ORIGINAL RESEARCH
published: 20 February 2020
doi: 10.3389/fgene.2020.00073

Initial Characterization of the Chloroplast Genome of *Vicia sepium*, an Important Wild Resource Plant, and Related Inferences About Its Evolution

Chaoyang Li[1], *Yunlin Zhao*[1], *Zhenggang Xu*[1,2*], *Guiyan Yang*[1], *Jiao Peng*[1] *and Xiaoyun Peng*[2]

[1] Hunan Research Center of Engineering Technology for Utilization of Environmental and Resources Plant, Central South University of Forestry and Technology, Changsha, China, [2] Hunan Urban and Rural Ecological Planning and Restoration Engineering Research Center, Hunan City University, Yiyang, China

OPEN ACCESS

Edited by:
Ancha Baranova,
George Mason University,
United States

Reviewed by:
Aleksandar M. Mikich,
Independent Researcher,
Novi Sad, Serbia
Tatiana V Tatarinova,
University of La Verne,
United States

***Correspondence:**
Zhenggang Xu
rssq198677@163.com

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: 08 October 2018
Accepted: 22 January 2020
Published: 20 February 2020

Citation:
Li C, Zhao Y, Xu Z, Yang G, Peng J and Peng X (2020) Initial Characterization of the Chloroplast Genome of Vicia sepium, an Important Wild Resource Plant, and Related Inferences About Its Evolution. Front. Genet. 11:73. doi: 10.3389/fgene.2020.00073

Lack of complete genomic information concerning *Vicia sepium* (Fabaceae: Fabeae) precludes investigations of evolution and populational diversity of this perennial high-protein forage plant suitable for cultivation in extreme conditions. Here, we present the complete and annotated chloroplast genome of this important wild resource plant. *V. sepium* chloroplast genome includes 76 protein-coding genes, 29 tRNA genes, 4 rRNA genes, and 1 pseudogene. Its 124,095 bp sequence has a loss of one inverted repeat (IR). The GC content of the whole genome, the protein-coding, intron, tRNA, rRNA, and intergenic spacer regions was 35.0%, 36.7%, 34.6%, 52.3%, 54.2%, and 29.2%, respectively. Comparative analyses with plastids from related genera belonging to Fabeae demonstrated that the greatest variation in the *V. sepium* genome length occurred in protein-coding regions. In these regions, some genes and introns were lost or gained; for example, *ycf4*, *clpP* intron, and *rpl16* intron deletions and *rpl20* and *ORF292* insertions were observed. Twelve highly divergent regions, 66 simple sequence repeats (SSRs) and 27 repeat sequences were also found in these regions. Detailed evolutionary rate analysis of protein-coding genes showed that *Vicia* species exhibit additional interesting characteristics including positive selection of *ccsA*, *clpP*, *rpl32*, *rpl33*, *rpoC1*, *rps15*, *rps2, rps4*, and *rps7*, and the evolutionary rates of *atpA*, *accD*, and *rps2* in *Vicia* are significantly accelerated. These genes are important candidate genes for understanding the evolutionary strategies of *Vicia* and other genera in Fabeae. The phylogenetic analysis showed that *Vicia* and *Lens* are included in the same clade and that *Vicia* is paraphyletic. These results provide evidence regarding the evolutionary history of the chloroplast genome.

Keywords: chloroplast genome, comparative analysis, phylogenetic analysis, positive selection, *Vicia sepium*

INTRODUCTION

Complete chloroplast sequences are indispensable for analyzing genome evolution and phylogenetics (Sabir et al., 2014; Moner et al., 2018). These sequences offer two advantages over genomic ones, namely, a high degree of conservation and a relatively compact gene alignment, resulting from symbiotic horizontal transfer (Timmis et al., 2004). In angiosperms, the chloroplast is a uniparentally inherited organelle. It originated from a cyanobacterium-like organism through an endosymbiosis event. Compared to the nuclear genome, chloroplast genomes, with a quadripartite circular structure, exhibit highly conserved sizes, structures and gene contents across photosynthetic plants (Wicke et al., 2011). Nuclear genomes are highly complex because of the high frequency of the loss and gain of genetic material at any time (Wolfe et al., 1987), making the identification of orthologous genes difficult. Evolutionary and phylogenetic analyses based on complete chloroplast sequences can provide more valuable information of a higher quality than that obtained by analysis of one or more gene loci (Martin et al., 2005). Complete chloroplast sequence datasets contain all site patterns (or all genes) for the reconstruction of evolutionary history. The comparison of complete genomes can reduce the sampling error inherent in analyses of only one or a few genes. That is not to say that we oppose the use of one or a few genes in evolutionary studies, but we instead suggest the investigation of conflicts between complete chloroplast genomes and analyses of one or a few genes that may indicate crucial evolutionary events. Another advantage of the chloroplast genome is that it contributes to structural diversity at low taxonomic levels and among basal lineages. Although genome organization is relatively well conserved in angiosperms, several types of structural diversity have been found. This structural diversity, including the loss of one copy of IRs, gene and intron gains or losses, large inversions, expansions, contractions and localized hypermutable phenomena, provides a powerful tool for evaluating genomic evolutionary history. For example, the loss of one IR is observed in the inverted-repeat-lacking clade (IRLC) (Sabir et al., 2014); the loss of *accD*, *psaI*, *ycf4*, *rpl33*, *clpP*, and *rps16* resulting in gene function loss is observed in various legume lineages; a 36-kb inversion is observed in the *Genistoid* clade; a 39-kb inversion is observed in *Robinia* (Keller et al., 2017); and hypermutation of *ycf4* is observed in *Lathyrus* (Magee et al., 2010). With the development of high-throughput sequencing, more than 800 complete chloroplast genomes have been made available in the National Center for Biotechnology Information (NCBI) database (Asaf et al., 2017a).

The Fabaceae family, especially the Papilionoideae subfamily, is considered a model system for understanding the mechanisms of chloroplast genome evolution due to the presence of major genome rearrangements in this group such as loss of one IR, gene and intron gains and losses, large inversions, expansions, contractions and localized hypermutable regions (Sabir et al., 2014; Keller et al., 2017). However, the mechanisms of these chloroplast genome rearrangements are not known (Sveinsson and Cronk, 2016). Some scholars believe that these genome rearrangements within the Fabaceae chloroplast genomes may be derived from the loss of one copy of IRs; however, *Medicago* and *Cicer* species, which exhibit the typical conserved quadripartite structure found in angiosperms (Jansen et al., 2005), also present extensive chloroplast genome rearrangements (Jansen et al., 2008; Sveinsson and Cronk, 2016). Therefore, further in-depth research on the mechanisms of chloroplast genome evolution is needed.

Previous research on Fabaceae chloroplast genomes demonstrated that the deletion or addition of genes and introns, inversions, repeats, and nucleotide variability can result in significant changes in genome length, GC content, and gene composition and orientation (Lei et al., 2016; Yin et al., 2017; Wang et al., 2018). In these genomes, coding regions are better conserved than intergenic spacer (IGS) regions (Sabir et al., 2014; Asaf et al., 2017b; Yin et al., 2018). However, it is unclear whether a consistent pattern in the genomic variation can be observed in species of the tribe Fabeae, which belong to Fabaceae. A possible explanation for these results may be the lack of complete genomic information for Fabeae. To date, 21 complete Fabeae chloroplast genomes have been sequenced (including 18 in the last four years), mainly from the genus *Lathyrus* (13) and a few from the genera *Lens* (1), *Pisum* (4) and *Vicia* (3). Another possible explanation is the structural diversity among Papilionoideae (Jansen et al., 2008; Sabir et al., 2014; Sveinsson and Cronk, 2016). For example, even within the same genus, the *Trifolium subterraneum* (Fabaceae) chloroplast genome exhibits 14-18 inversions, while there are only 3 inversions in *Trifolium grandiflorum* and *Trifolium aureum* (Sabir et al., 2014). Therefore, the study of the genomic variation and phylogeny of Fabeae species can provide a basis for understanding chloroplast genome evolution.

Vicia sepium (Bush vetch), belonging to the tribe Fabeae, is an important wild resource plant with a wide distribution area (Maxted, 1995), various flowering periods from May to November, abundant proteins, and suitability for cultivation in extreme cold and dry conditions (Maršalkienė, 2016) and can be used as a good potential perennial forage. Additionally, compared with other legumes, *V. sepium* provides herbage for a long period because of its perennial habit (Maršalkienė, 2016). This plant also produces extrafloral nectaries to attract ants, which act as plant defenders by preying on arthropod herbivores or interrupting their oviposition or feeding (Lenoir and Pihlgren, 2006). However, previous studies on *V. sepium* have mainly focused on the morphological characteristics (Maršalkienė, 2016) and classification (Schaefer et al., 2012; Jaaska, 2015) of this plant and the relationship between plants and insects (Kruess and Tscharntke, 2000; Lenoir and Pihlgren, 2006). Therefore, little is known regarding the nutrient content, genetic resources, and forage value of this species. As a result, no plant materials of *V. sepium* have been released for commercial production. However, another *Vicia* species, *Vicia sativa*, has been widely used as forage and for hay and silage production. A key difficulty in the use of *V. sativa* is the presence

of a neurotoxic compound in its seeds (Huang et al., 2017). Therefore, the expansion of forage resources based on *Vicia* species is necessary.

Another difficulty in the utilization of *V. sepium* is that the taxonomy of some taxa in Fabeae remains controversial (Schaefer et al., 2012; Jaaska, 2015; Iberite et al., 2017) because of the high morphological variability among species. Notably, some variation in morphological characteristics is genetically fixed. For example, Iberite's cultivation tests (Iberite et al., 2017) conducted in *V. sativa*, *Vicia barbazitae*, *Vicia grandiflora* and *V. sepium* showed that the characteristics of the leaf margins are maintained through successive generations. Recent molecular phylogenetic studies have focused on multitribe legumes or tribe level analyses of Fabeae (Schaefer et al., 2012). These studies have suggested that the taxonomy of some genera in Fabeae is not monophyletic. However, these phylogenetic studies did not use the complete chloroplast genome, instead using plastid DNA sequence data, such as the *matK*, *trnL*, *rbcL*, and nuclear ribosomal internal transcribed spacer (ITS) sequences. Therefore, it is necessary to acquire comprehensive knowledge regarding the organization and evolution of *V. sepium*.

Here, we present a new complete chloroplast genome of *V. sepium*, from the genus *Vicia*. We compare it with chloroplast genomes from related genera (*Lens*, *Pisum*, *Lathyrus*) belonging to tribe Fabeae. The aim of this work is to reveal the genome variation and phylogeny of Fabeae and the genus *Vicia* and to provide evidence regarding the history of chloroplast genome evolution.

MATERIALS AND METHODS

Plant Material

The sample was collected from the Dongting Lake region (28°48′ 46.06″N, 112°21′10.19″E) and stored at the Hunan Research Center of Engineering Technology for Utilization of Environmental and Resources Plant, China, under accession number 20170707JJ. Plant sampling was performed in areas that were not privately owned or protected in any way, and no specific permits were required for this study. We collected mature *V. sepium* leaves and placed them in a liquid nitrogen container. Leaf samples were stored at -80°C until sequencing. Extraction of total chloroplast DNA was carried out with the Plant Chloroplast Purification Kit and Column Plant DNA Extraction Kit (Beijing Baiaolaibo Technology, Co., Ltd., China). The chloroplast DNA of *V. sepium* was fragmented using a Covaris M220 (Covaris, USA) instrument. Whole-genome sequencing and paired-end (PE) library construction were performed according to the method described by Zhang et al. (2017). Raw data were obtained through next-generation sequencing with PE 150-bp reads. Then, N-containing sequences and adapter sequences were removed. Sequences with a Q value less than 20 or an average four-base mass of less than 20 were also removed. Finally, if the length of the reads was less than 50 nt, the reads were removed. All the above filtering steps were performed using Trimmomatic v 0.32 (Bolger et al., 2014), and clean data for subsequent analysis were obtained. Then, all high-quality paired reads were assembled into contigs by using SOAPdenovo2 (Luo et al., 2012) and scaffolded by using SSPACE (Boetzer et al., 2011) to obtain the whole-genome sequence. In this process, different K-mers were selected first for assembly, and the best K-mer, k=25, was chosen to obtain the assemblies. The above K parameter was determined on the basis of a K-mer curve and experience. Finally, one contig of 124,095 bp was obtained.

Genome Annotation and Sequence Architecture

Our previous study used the programs CpGAVAS (Liu et al., 2012) and DOGMA (Wyman et al., 2004) to annotate the complete chloroplast genome of *V. sepium* (Li et al., 2018). In this study, to study genomic evolution between *V. sepium* and its related species in Fabeae, the same *V. sepium* genome was annotated in Plann (Huang and Cronk, 2015) against the *V. sativa* genome (NC027155). Gene mapping and relative synonymous codon usage (RSCU) were performed in OGDRAW v1.2 (Lohse et al., 2013) and DAMBE6 (Xia, 2017) according to Dong's method (Dong et al., 2019).

SSRs and Repeated Sequences Analysis

We detected SSRs by referring to the method of Lei et al. (2016) using the MISA Perl Script (Thiel et al., 2003) with parameter settings of 8 for mono-, 4 for di- and tri-, and 3 for tetra-, penta- and hexa-nucleotide SSRs. Forward, palindromic, reverse, and complement sequences were identified as described by Cauz-Santos et al. (2017) using REPuter (Kurtz et al., 2001) with 90% or greater sequence identity and a length of 30 bp or longer. Tandem repeats were identified using Tandem Repeats Finder version 4.09 (Benson, 1999) with default parameters.

Comparative Analysis

Blast ring image generator (BRIG) (Alikhan et al., 2011) and mVISTA (Frazer et al., 2004) software were used to compare the complete chloroplast genome variation in all available Fabeae chloroplast genomes using *V. sepium* annotation as a reference. BRIG focus on protein coding segment variation and mVISTA align whole chloroplast genome without discrimination. All the species were included the following twenty-one Fabeae species and one Cicereae species (*Cicer arietinum*), listed with the corresponding GenBank accession numbers: *V. sepium*, *V. sativa* (NC027155), *V. faba* (KF042344), *Pisum abyssinicum* (NC037830), *Pisum sativum* (NC014057), *Pisum sativum* subsp. Elatius (NC039371), *Pisum fulvum* (NC036828), *Lens culinaris* (NC027152), *Lathyrus pubescens* (NC027079), *Lathyrus venosus* (NC027080), *Lathyrus palustris* (NC027078), *Lathyrus japonicus* (NC027075), *Lathyrus ochroleucus* (NC027077), *Lathyrus davidii* (NC027073), *Lathyrus littoralis* (NC027076), *Lathyrus inconspicuus* (NC027149), *Lathyrus graminifolius* (NC027074), *Lathyrus tingitanus* (NC027151), *Lathyrus clymenum* (NC027148), *Lathyrus sativus* (NC014063), *Lathyrus odoratus* (NC027150), and *C. arietinum* (NC011163). Genome rearrangement relative to *V. sepium* was performed in Mauve (Darling et al., 2004).

Phylogenetic Analysis

To determine the phylogenetic position of *V. sepium* within Fabeae, four datasets were used to construct the following phylogenetic trees for Fabeae: (I) the complete chloroplast genomes of 21 Fabeae species and *C. arietinum* (that is, the same 22 species in the comparative analysis); (II) the conserved chloroplast protein-coding sequences of 21 Fabeae species and *C. arietinum* (that is, the same 22 species in the comparative analysis); (III) the *rbcL* gene sequences of 50 Fabeae species, *Trifolium pretense* and *T. repens*; and (IV) the *matK* gene sequences of 62 Fabeae species, *T. pretense* and *T. repens*. The names of the species included in the four phylogenetic analyses can be found in **Table S1**.

Specifically, the conserved chloroplast protein-coding sequence of each species comprised 70 concatenated homologous genes shared among twenty-two related species. These genes were *atpA*, *atpB*, *atpE, atpF*, *atpH*, *atpI*, *ccsA*, *cemA*, *clpP*, *matK*, *ndhA*, *ndhB*, *ndhC*, *ndhD*, *ndhE*, *ndhF*, *ndhG*, *ndhH*, *ndhI*, *ndhJ*, *ndhK*, *petA*, *petB*, *petD*, *petG*, *petL*, *petN*, *psaA*, *psaB*, *psaC*, *psaJ*, *psbA*, *psbB*, *psbC*, *psbD*, *psbE*, *psbF*, *psbH*, *psbI*, *psbK*, *psbL*, *psbM*, *psbN*, *psbT*, *psbZ*, *rbcL*, *rpl14*, *rpl16*, *rpl2*, *rpl20*, *rpl23*, *rpl32*, *rpl33*, *rpl36*, *rpoA, rpoB*, *rpoC1*, *rpoC2*, *rps11*, *rps12*, *rps14*, *rps15*, *rps19*, *rps2*, *rps3*, *rps4*, *rps7*, *rps8*, *ycf1*, *ycf2*, and *ycf3*.

All datasets were aligned using MAFFT v7.380 (Katoh and Standley, 2013) under the FFT-NS-2 default setting. The alignments were used for phylogenetic analysis. All alignments were used to construct phylogenetic trees *via* the neighbor joining (NJ) method in MEGA7.0 (Kumar et al., 2016) under the default settings. Then, we obtained four NJ trees.

In addition, we used another method, the maximum likelihood (ML) method, to construct a phylogenetic tree based on conserved chloroplast protein-coding sequences. The aim of this work was to test the effects of different methods on the phylogenetic relationships of Fabeae species. First, we used MAFFT v7.380 to align twenty-two conserved chloroplast protein-coding sequences under the FFT-NS-2 default settings. Second, ModelTest was employed to find the best model in MEGA7.0. Finally, the tree was constructed using the ML method with the GTR+G+I model and 1,000 bootstrap replicates. *C. arietinum* was selected as the outgroup.

Evolutionary Rate Analysis

To determine the sequence divergence of the complete chloroplast genomes, the average pairwise sequence distances of twenty-one Fabeae species and one Cicereae species (that is, the same 22 species in the comparative analysis) were calculated. After alignment with MAFFT v7.380, the average pairwise sequence distances (K2P rate) of these species were presented according to Asaf's method using MEGA7 (Kimura, 1980; Asaf et al., 2017b).

Additionally, the synonymous (Ks) and nonsynonymous (Ka) nucleotide substitution rates as well as the Ka/Ks ratio were used to calculate the sequence divergence of other homologous protein-coding regions. All twenty-one available chloroplast genomes belonging to the genera *Vicia*, *Pisum*, *Lens*, and *Lathyrus* were selected for this analysis. These species were divided into two groups: (I) within *Vicia*: *V. sepium*, *V. sativa*, *V. faba*; (II) outside of *Vicia* (or other genera): *V. sepium*, *P. abyssinicum*, *P. sativum*, *P. sativum* subsp. Elatius, *P. fulvum*, *L. pubescens*, *L. venosus*, *L. palustris*, *L. japonicus*, *L. ochroleucus*, *L. davidii*, *L. littoralis*, *L. inconspicuus*, *L. graminifolius*, *L. tingitanus*, *L. clymenum*, *L. sativus*, and *L. odoratus*. A total of 71 homologous genes (**Table S2**) from these species were selected and examined separately. After aligning each gene using the ClustalW (Codons) program in MEGA7, the Ks, Ka, and Ka/Ks values between *V. sepium* and other species were determined according to Dong's method (Dong et al., 2019) with the program from the PAML package (Yang and Nielsen, 1998). The two independent-samples t-test was used to examine the significance of the sequence divergence between *Vicia* and other genera. The *p*-values were determined with Levene's test. If the Levene's test result was less than 0.05, we used the unequal variance as the *p*-value; if not, we used the equal variance as the *p*-value.

Once *Vicia* showed a significantly higher Ka/Ks ratio than the other genera, codon-based likelihood analysis based on the branch model test in CodeML from the PAML package was carried out to identify the lineages in Fabeae that exhibited significantly high evolutionary rates. This test employed the user-defined topology of Fabeae lineages with five other lineages: A0 (*Cicer*), A1 (*Pisum* and *Lathyrus*), A2 (*Lens* and *Vicia*), A3 (*Lens*), and A4 (*Vicia*). This topology was constructed based on the concatenated DNA sequences of *matK* and *rbcL* (**Figure S1**) using the ML method with the GTR+G50 model in MEGA7.0. The method was the same as that used for the phylogenetic analysis described previously. A one-ratio model (model = 0) and a two-ratio model (model =2) were used to calculate the Ka/Ks ratio for each branch. A one-ratio model, or null model (model = 0), is one in which all clades (or all lineages) exhibit the same Ka/Ks ratio. A two-ratio model, or alternative model (model = 2), is one in which one or more clades present different Ka/Ks ratios. The transition/transversion and Ka/Ks ratios were set as automatically estimated. Codon frequencies were set as the F3 × 4 method. The hypotheses of the two-ratio model are described in **Table S3**. The likelihood ratio test (LRT) was used to find the best model ($P < 0.05$) through comparison of two different models. From the best model, we could infer whether a homologous gene showed accelerated evolution in *Vicia*. In addition, all genes exhibiting accelerated evolution were compared with two genes showing nonaccelerated evolution (*matK* and *rbcL*), in two ways. First, we compared their synonymous and nonsynonymous nucleotide substitution rates in Ks trees and Ka trees. The branch lengths representing the substitutions per synonymous site or nonsynonymous site were determined from the best model. Second, we compared their amino acid sequence differences. Amino acid sequence alignment was performed in Jalview v2.10.5 (Waterhouse et al., 2009).

RESULTS

Chloroplast Genome Characteristics and Structure of *V. sepium*

The original image data obtained by next-generation sequencing technology was converted into the original sequenced reads by CASAVA base calling analysis to obtain raw reads (10,808,365) or raw data (3.24 gigabytes). A total of 7,696,368 clean reads (2.31 gigabytes of clean data) with an average length of 150 bp were obtained after the adapter sequences and low-quality reads were removed. A single long contig of 124,095 bp was assembled using clean data *via de novo* assembly, forming a loop representing the whole chloroplast genome sequence of *V. sepium*. The *V. sepium* chloroplast genome, under GenBank accession number NC039595, showed the loss of one IR and contained 76 protein-coding genes, 29 tRNA genes, four rRNA genes and one pseudogene (*rpl23* Ψ). In particular, one unannotated protein-coding gene, *ORF292*, was identified (**Table 1**). The gene map of these 110 genes was presented (**Figure 1**). Among these protein-coding genes, 9 genes (*ndhA*, *ndhB*, *rpl2*, *rpl16*, *petD*, *petB*, *atpF*, *rpoC1*, *clpP*) contained a single intron, while one gene, *ycf3*, contained two introns (**Table 2**). Additionally, four tRNA genes containing one intron were also identified as follows: $trnV^{UAC}$, $trnA^{UGC}$, $trnI^{GAU}$, and $trnL^{UAA}$. As observed in most legumes, the *infA*, *rpl22*, and *rps16* genes were lost (Lei et al., 2016). The overall GC content of the *V. sepium* chloroplast genome was 35.0%, whereas that of the protein-coding, intron, tRNA, rRNA and IGS regions was 36.7%, 34.6%, 52.3%, 54.2%, and 29.2%, respectively (**Table S4**). The RSCU result revealed that the *V. sepium* protein-coding sequences showed codon usage bias, with all preferred synonymous codons ending with A/T nucleotides and a high AT content at the 3rd codon positions (72.2%) (**Figure S2**, **Table S4**).

TABLE 1 | Genes predicted in the chloroplast genome of *V. sepium*.

Category	Group of genes	Names of genes
Self-replication	Large subunit of ribosomal proteins	*rpl2, rpl14, rpl16, rpl20, rpl23* [a], *rpl32, rpl33, rpl36*
	Small subunit of ribosomal proteins	*rps2, rps3, rps4, rps7, rps8, rps11, rps12* [b], *rps14, rps15, rps18, rps19*
	DNA dependent RNA polymerase	*rpoA, rpoB, rpoC1, rpoC2*
	rRNA genes	*rrn16S, rrn4.5S, rrn23S, rrn5S*
	tRNA genes	*trnA-TGC, trnC-GCA, trnD-GTC, trnE-TTC, trnF-GAA, trnG-TCC, trnH-GTG, trnI-CAT, trnI-GAT, trnK-TTT, trnL-CAA, trnL-TAA, trnL-TAG, trnM-CAT* [c], *trnMf-CAT, trnN-GTT, trnP-GGG, trnP-TGG, trnQ-TTG, trnR-ACG, trnR-TCT, trnS-GCT, trnS-GGA, trnS-TGA, trnT-GGT, trnV-TAC, trnW-CCA, trnY-GTA*
Photosynthesis	Photosystem I	*psaA, psaB, psaC, psaI, psaJ*
	Photosystem II	*psbA, psbB, psbC, psbD, psbE, psbF, psbH, psbI, psbJ, psbK, psbL, psbM, psbN, psbT, psbZ*
	NADH dehydrogenase	*ndhA, ndhB, ndhC, ndhD, ndhE, ndhF, ndhG, ndhH, ndhI, ndhJ, ndhK*
	Cytochrome b6/f complex	*petA, petB, petD, petG, petL, petN*
	ATP synthase	*atpA, atpB, atpE, atpF, atpH, atpI*
	Rubisco	*rbcL*
Other genes	Maturase	*matK*
	Protease	*clpP*
	Envelope membrane protein	*cemA*
	Subunit acetyl-CoA-carboxylase	*accD*
	C-type cytochrome synthesis gene	*ccsA*
Genes of unknow function	Conserved open reading frames	*ycf1, ycf2, ycf3, ycf4*

One open reading frame, ORF292, *could not be annotated.* [a]*pseudogene;* [b]*trans-splicing gene;* [c]*duplicated gene.*

SSRs and Repeats in *V. sepium*

We analyzed the presence of SSRs and repeats in *V. sepium*. SSRs, which are regarded as useful gene markers, exhibited a high mutation rate. In this study, a total of 201 SSRs were found in the chloroplast genome of *V. sepium* (**Figure 2**). A majority of the SSRs were composed of mono-nucleotide and di-nucleotide repeat motifs. The types of SSRs distributed within the chloroplast genome of *V. sepium* were characterized, revealing that the SSR motifs of mono-nucleotide repeats mainly consisted of A/T (98.5%) and that those of di-nucleotide repeats mainly consisted of AT/TA (86.8%). A total of 116 and 66 V. *sepium* SSRs were distributed in the IGS and CDS regions, respectively (**Figure 2**).

Repeat sequences are essential for genome rearrangements, phylogenetic construction (Cavalier-Smith, 2002) and indel, and substitution variation (Yi et al., 2013). Sixty-two repeats, including 46 forward repeats, 4 palindromic repeats, and 12 tandem repeats, were found in the chloroplast genome of *V. sepium*. The lengths of the palindromic repeats were 45, 50, 54, and 155 bp, and the lengths of the forward repeats and tandem repeats ranged from 45 to 222 bp and 32 to 229 bp, respectively (**Table S5**). In addition, the maximum number of repeats (n = 49) were located in IGS regions, followed by those in CDSs (n = 27) (**Table S5**). We also found that most of these repeats were located in the *psaB-rps14* (n = 20), *ycf1-trnN-GUU* (n = 10), *accD* (n = 6) and *rps14* (n = 5) regions.

Comparative Analyses of the Chloroplast Genomes of Fabeae Species

Twenty complete chloroplast genomes from within Fabeae were selected for comparison with *V. sepium*. One Cicereae species, *C. arietinum*, was set as the outgroup (**Table 3**). The changes in chloroplast genome length in these species ranged from 120, 289 bp (*L. odoratus*) to 126,421 bp (*L. pubescens*), and the greatest variation in length relative to *V. sepium* was 3.0% in the protein-coding region of *L. culinaris*, followed by the IGS region (2.8%) of *L. culinaris*. An average difference in length of only 0.1% was found in the tRNA and rRNA gene regions. Additionally, the GC content of the twenty-two complete chloroplast genomes ranged from 33.9% to 35.2%, exhibiting little change. After comparing

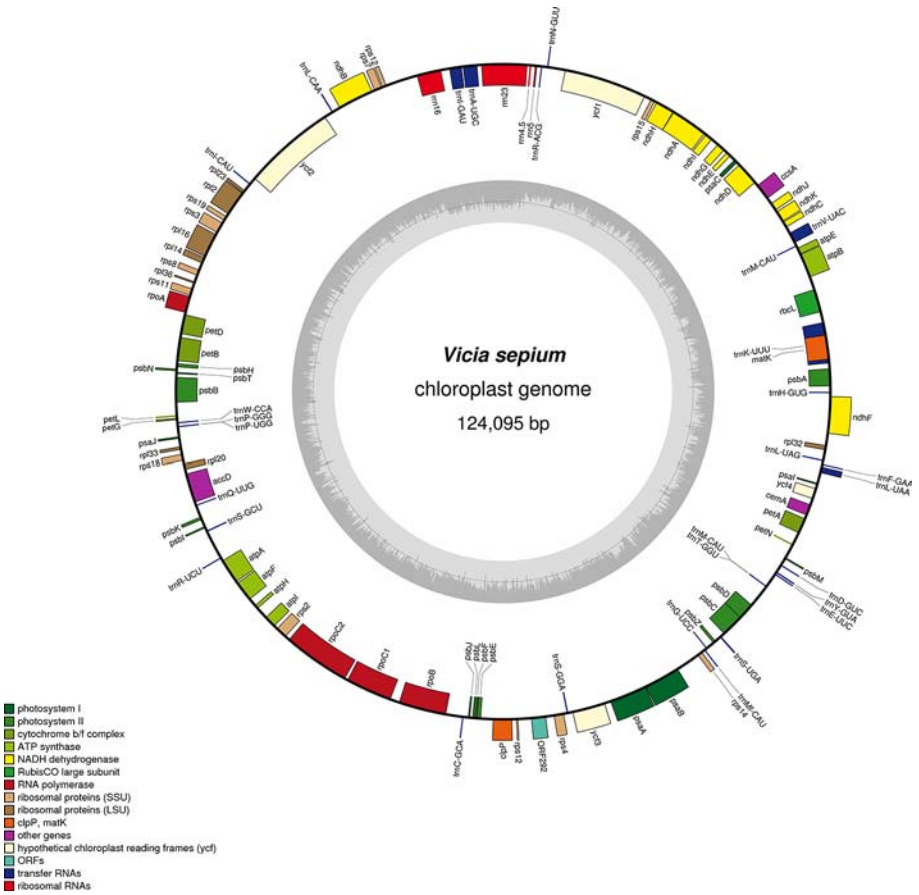


FIGURE 1 | Gene map of the complete chloroplast genome of *V. sepium*. Genes inside the circle are transcribed clockwise, and those outside are transcribed counterclockwise. The different colors of the blocks represent different functional groups. The darker gray color of the inner circle corresponds to the GC content, and the lighter gray color corresponds to the AT content.

TABLE 2 | Lengths of introns and exons of the split genes in the *V. sepium* complete chloroplast genome.

Gene name	Gene Location			Length (bp)				
	Strand	Start	End	Exon I	Intro I	Exon II	Intro II	Exon III
ndhA	-	17,922	20,213	552	1,200	540		
ndhB	+	39,164	41,349	720	674	792		
rpl2	+	49,205	50,732	393	700	435		
rpl16	+	52,173	53,655	9	1,072	402		
petD	-	57,360	58,556	9	714	474		
petB	-	58,753	60,207	6	804	645		
atpF	-	74,347	75,592	168	670	411		
rpoC1	-	83,263	86,132	435	791	1,644		
clpP	+	92,455	93,604	363	559	228		
ycf3	+	97,292	99,294	126	742	228	781	126
trnV-UAC	+	9,320	9,976	39	581	37		
trnA-UGC	-	32,593	33,473	38	808	35		
trnI-GAU	-	33,539	34,292	42	677	35		
trnL-UAA	+	119,177	119,535	37	272	50		

V. sepium genes with those of twenty-one other Fabeae species, we found an inserted gene that is a unique unannotated protein-coding gene, *ORF292*, between *rps12* and *rps4* in *V. sepium*. Moreover, the *rps12* to *rps4* region in *V. sativa* also contained an inserted duplicated *rpl20* gene (not mentioned in the table). From genome rearrangement, we can infer that inversion events may result in gene insertion (**Figure S3**). We also found a pseudogene, *rpl23*, in *V. sepium*, *V. sativa*, *P. abyssinicum*, *P. sativum*, *P. sativum* subsp. Elatius and *L. sativus*. By analyzing gene and intron losses, all twenty-two species lost the *infA*, *rpl22*, and *rps16* genes, similar to most of the IR-lacking species. *Ycf4* genes were found in only *V. sepium*, *V. faba*, *P. sativum*, and *L. sativus*. Moreover, one intron of the *clpP* and *rpl16* genes was lost in *L. graminifolius* and *V. faba*, respectively (**Table 3**).

The sequence identity of the chloroplast genomes of *V. sepium* and twenty-one other Fabaceae species was visualized (**Figures 3** and **S4**), and the results revealed that coding regions are more highly conserved than noncoding regions. Usually,

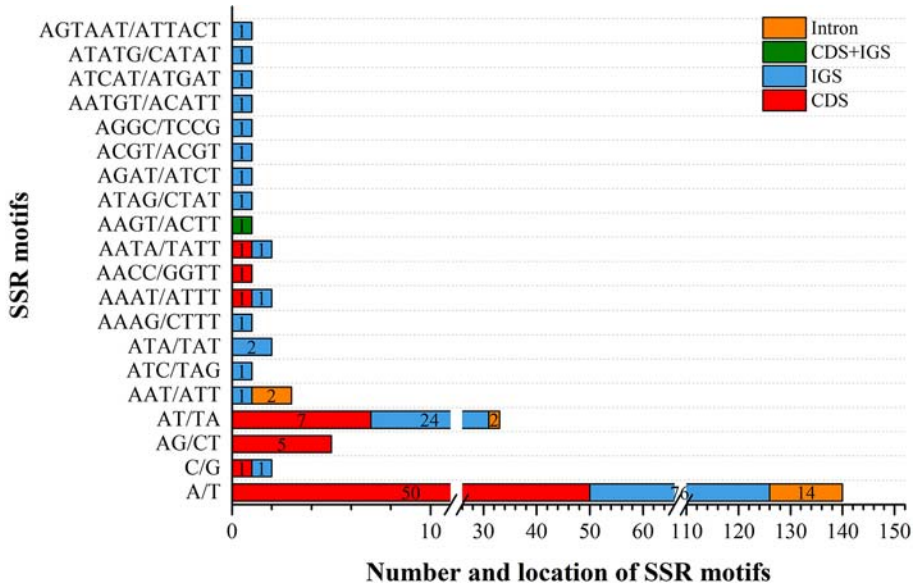


FIGURE 2 | The types and distribution of SSRs along the chloroplast genome of *V. sepium*. Different locations, including CDS, IGS, CDS and IGS, and intron regions, are represented as colored boxes.

regions with 50% or less sequence identity can be regarded as highly divergent regions. In coding regions, *ycf1*, *ycf2*, *rpl23*, *rps3*, *rps18*, *accD*, *rpoC1*, *clpP*, *ORF292*, *ycf4, psaI*, and *rpl32* contained relatively low identity regions. In addition, these highly divergent noncoding regions include *rps15-ycf1*, *ycf1-trnN-GUU*, *rrn16-rps12*, *ycf2-trnI-CAU*, *trnI-CAU -rpl23*, *rpl16* intron, *rpl14-rps8*, *rps8-rpl36*, *psbB-petL*, *accD-trnQ-UUG*, *trnQ-UUG-psbK*, *psbE-clpP*, *clpP-rps12*, *psaB-rps14*, *psbD-trnT-GUU*, *ycf4-psaI*, *psaI-trnL-UAA*, and *rpl32-ndhF* (**Figure 3** and **S4**).

Evolutionary Rate of Fabeae Species

The pairwise distances (K2P rates) of complete chloroplast genome sequences from twenty-one Fabeae species and one Cicereae species were calculated (**Table S6**). The results showed that the nucleotide variability rate ranged from 0.001 to 0.248 (*L. sativus* vs *C. arietinum*). Compared with *V. sepium*, the lowest K2P rate was 0.027 (*V. sativa*) while the highest K2P rate was found in *C. arietinum* (0.246) (**Table S6**). The mean K2P rate between *Pisum* and *V. sepium* was 0.217. The mean K2P rate between *Lathyrus* and *V. sepium* was 0.193. Specifically, the K2P rate between *V. faba* and *V. sepium* was 0.207, which was higher than the rate between *V. sepium* and some *Lathyrus* species. We hypothesized that *V. sepium* and *V. sativa* were located in the same clade and showed different evolutionary directions compared with *V. faba.*

Ka and Ks nucleotide substitutions within *Vicia* and outside of *Vicia* were calculated with *V. sepium* as the reference, as well as the Ka/Ks ratio (**Table S2**, **Figure 4**). The Ka/Ks ratio is an important parameter for determination of the selective constraint acting on each gene (Keller et al., 2017). Ka/Ks > 1 indicates that the gene was under positive selection, whereas Ka/Ks = 1 or <1 indicates genes under neutral selection or purifying selection (Kimura, 1980). The mean Ks between *V. sepium* and twenty Fabeae species ranged from 0.0058 (*petN*) to 0.2375 (*ycf1*), and the mean Ka ranged from 0 (*petG*, *psbF*) to 0.1846 (*clpP*) (**Table S2**). Within the genus *Vicia*, nine genes (*ccsA*, *clpP*, *rpl32*, *rpl33*, *rpoC1*, *rps15*, *rps2, rps4* and *rps7*) with a Ka/Ks ratio >1 (**Figure 4**) evolved under beneficial mutations, and 60 genes evolved under purifying selection, including sixteen genes that evolved almost neutrally, showing a ratio range of 0.5 to 1. Twelve conserved genes (*atpH*, *petG*, *petN*, *psaC*, *psbA*, *psbD*, *psbF*, *psbH*, *psbK*, *psbL*, *psbM* and *rpl36* with Ka/Ks = 0) presented a very strong purifying selective pressure. Comparison of sequence divergence between *Vicia* and other genera showed that the Ka/Ks ratios of the eight genes (*accD*, *atpA*, *matK*, *rpl32*, *rpl33*, *rps2*, *rps4*, *ycf1*) were significantly higher ($P < 0.05$) in *Vicia*, and among these genes, the ratios of *accD*, *atpA*, *rpl32*, *rps2* and *rps4* were extremely significantly higher ($P < 0.01$).

Codon-based likelihood analysis (**Table S3**; **Figure S1**) was performed to compare the Ka/Ks ratios of the *accD*, *atpA*, *rpl32*, *rps2*, and *rps4* genes across different Fabeae lineages. *C. arietinum* was set as the reference. The null model (H0) hypothesized that the A0 (*Cicer*), A1 (*Pisum* and *Lathyrus*), A2 (*Lens* and *Vicia*), A3 (*Lens*), and A4 (*Vicia*) clades exhibit the same Ka/Ks ratio. The alternative model hypothesized that one or more clades present different Ka/Ks ratios. By comparing the *p*-values of the two different models, the results demonstrated that the best models for *accD*, *atpA*, *rpl32*, *rps2*, and *rps4* are H2, H3, H0, H2, and H0, respectively (**Table S3**). A higher Ka/Ks ratio in a specific clade is considered to indicate accelerated evolution of the clade. The Ka/Ks ratios of *accD*, *atpA* and *rps2* in

TABLE 3 | Characteristics of twenty-one Fabeae species and *Cicer arietinum*.

Species	*V. sep*	*V. sat*	*V. fab*	*P. aby*	*P. sat*	*P. sat*^{stub}	*P. ful*	*L. cul*	*L. pub*	*L. ven*	*L. pal*	*L. jap*	*L. och*	*L. dav*	*L. lit*	*L. inc*	*L. gra*	*L. tin*	*L. cly*	*L. sat*	*L. odo*	*C. ari*
Length (bp)	12,4095	12,2467	123,722	12,2174	12,2169	12,1958	120,837	122,967	126,421	125,459	124,287	124,242	123,911	123,895	123,734	123,153	122,438	122,165	121,263	121,020	120,289	125,319
Genes	110	109	110	108	110	108	107	108	108	110	109	109	109	108	108	108	109	108	110	109	107	108
CDS genes*	77	76	76	75	76	75	73	74	74	75	75	75	75	74	74	74	75	74	75	75	73	75
tRNA genes	29	29	30	29	30	29	30	30	30	30	30	30	30	30	30	30	30	30	31	30	30	29
Introns**	14	14	14	14	15	14	15	14	15	15	15	15	15	14	15	15	14	14	15	15	15	15
CDS region	53.9%	52.9%	53.2%	53.9%	54.3%	53.9%	52.7%	50.9%	53.3%	52.2%	53.0%	52.9%	53.0%	52.6%	53.1%	54.0%	53.5%	53.5%	53.7%	55.3%	54.1%	52.3%
Intron region	8.9%	9.0%	8.8%	9.1%	9.7%	9.1%	9.2%	9.0%	8.8%	8.8%	8.8%	8.8%	8.8%	8.9%	8.9%	9.1%	8.5%	9.2%	9.3%	9.7%	9.2%	9.7%
IGS region	31.9%	32.7%	32.6%	31.5%	30.4%	31.5%	32.5%	34.7%	32.6%	33.7%	32.7%	32.8%	32.8%	33.1%	32.6%	31.4%	32.5%	31.8%	31.5%	29.4%	31.0%	32.6%
tRNA region	1.7%	1.7%	1.8%	1.8%	1.9%	1.8%	1.8%	1.7%	1.7%	1.8%	1.8%	1.8%	1.8%	1.8%	1.8%	1.8%	1.8%	1.8%	1.9%	1.9%	1.8%	1.8%
rRNA region	3.6%	3.7%	3.6%	3.7%	3.7%	3.7%	3.7%	3.7%	3.6%	3.6%	3.7%	3.7%	3.6%	3.7%	3.6%	3.7%	3.7%	3.7%	3.7%	3.8%	3.7%	3.6%
Genome GC	35.0%	35.2%	34.6%	34.8%	34.8%	34.8%	34.9%	34.4%	35.0%	34.9%	34.9%	34.9%	34.9%	34.9%	34.8%	34.7%	35.0%	34.9%	35.0%	35.1%	35.2%	33.9%
CDS GC	36.7%	36.9%	36.6%	36.8%	36.8%	36.8%	36.9%	36.6%	36.7%	36.7%	36.7%	36.7%	36.7%	36.8%	36.6%	36.6%	36.7%	36.7%	36.9%	36.9%	37.0%	36.3%
Intron GC	34.6%	34.7%	34.5%	34.4%	34.1%	34.4%	34.3%	34.1%	34.6%	34.7%	34.7%	34.7%	34.7%	34.7%	34.6%	34.1%	35.0%	34.3%	34.5%	34.2%	34.3%	33.5%
IGS GC	29.2%	29.4%	28.0%	28.2%	28.1%	28.2%	28.5%	28.4%	29.4%	29.1%	28.8%	28.9%	29.0%	28.9%	28.8%	28.4%	28.9%	28.8%	28.6%	28.6%	29.0%	27.0%
tRNA GC	52.3%	52.5%	52.4%	52.8%	52.7%	52.8%	52.7%	52.2%	52.4%	52.8%	52.7%	52.8%	52.8%	52.9%	52.8%	52.7%	52.8%	52.4%	52.6%	52.7%	52.6%	52.7%
rRNA GC	54.2%	54.2%	54.2%	54.0%	53.9%	54.0%	54.0%	53.8%	54.0%	54.2%	54.1%	54.1%	54.2%	53.9%	54.2%	54.1%	54.0%	54.0%	54.3%	53.6%	53.7%	54.2%
Gene losses		*ycf4*		*ycf4*		*ycf4*	*ycf4* *psaI* *accD*	*ycf4* *rps18* *psbZ* *lhbA*	*ycf4* *psaI*	*ycf4*	*ycf4*	*ycf4*	*ycf4*	*ycf4* *psaI*	*ycf4* *psbJ*	*ycf4* *psaI*	*ycf4*	*ycf4* *rps18*	*ycf4*	*psaI*	*ycf4* *rps18* *psaI*	*ycf4*
Gene gains	*ORF292*																					

*All of these species contain four rRNA genes except for L. ven (L. venosus). The full names of the twenty-two species are as follows: V. sep, V. sepium; V. sat, V. sativa; V. fab, V. faba; P. aby, P. abyssinicum; P. sat, P. sativum; P. satsub, P. sativum subsp. Elatius; P. ful, P. fulvum; L. cul, L. culinaris; L. pub, L. pubescens; L. ven, L. venosus; L. pal, L. palustris; L. jap, L. japonicus; L. och, L. ochroleucus; L. dav, L. davidii; L. lit, L. littoralis; L. inc, L. inconspicuus; L. gra, L. graminifolius; L. tin, L. tingitanus; L. cly, L. clymenum; L. sat, L.sativus; L. odo, L.odoratus; C. ari, C. arietinum. *pseudogenes: rpl23 in V. sepium, V. sativa, P. abyssinicum, P. sativum, P. sativum subsp. Elatius and L. sativus; ycf1 in L. culinaris; ycf4 in P. sativum. **intron gains: one intron added to tRNA-Gly (V. faba, P. sativum, L. sativus, V. faba, C. arietinum) and ycf2 (P. fulvum, L. pubescens, L. venosus, L. palustris, L. japonicus, L. ochroleucus, L. littoralis, L. inconspicuus, L. graminifolius, L. clymenum, L. odoratus). intron losses: one intron missing in clpP (L. graminifolius) and rpl16 (V. faba).*

the *Vicia* clade were higher than those in the *Cicer* clade, and the Ka/Ks ratios of *rpl32* and *rps4* were the same in the two clades. The results revealed that evolution rates increased in *atpA*, *rps2* and *accD* of the *Vicia* lineage but exhibited no change in *rpl32* or *rps4* (**Table S3**). The Ka/Ks ratios of *rps2* in the *Vicia* clade were higher than those in the *Pisum* and *Lathyrus* clade, but the *Vicia* clade presented lower Ka/Ks ratios in the *accD* and *atpA* genes. The results revealed that *rps2* exhibited a higher evolutionary rate in *Vicia*, while *atpA* and *accD* in *Pisum* and *Lathyrus* evolved much faster. We also compared the synonymous and nonsynonymous nucleotide substitution rates of genes that evolved rapidly (*accD*, *atpA*, and *rps2*) in different Fabeae lineages to the rates observed in genes that did not evolve rapidly (*rbcL* and *matK*) based on codon-based ML phylogenetic analysis. As shown in **Figure 5**, in the Ka and Ks trees, the substitutions per nonsynonymous site of *rps2* evolved much faster in *Vicia* than in other Fabeae species, but no similar acceleration was observed in *rbcL* and *matK*. In addition, all Fabeae lineages showed accelerated evolution in the *accD* gene for high synonymous and nonsynonymous nucleotide substitution rates compared to *rbcL* and *matK*. This result can supplement Magee's findings (Magee et al., 2010). We also detected amino acid differences in the *accD*, *atpA*, *rps2*, *matK*, and *rbcL* genes within and outside of *Vicia* by aligning the sequences from Fabeae species (**Figures S5**–**S9**). Notably, there is less amino acid sequence conservation in *accD* (83.03% identity between *Vicia* species) and *rps2* (91.98% identity between *Vicia* species) than in *matK* (94.23% identity between *Vicia* species) and *rbcL* (99.16% identity between *Vicia* species). The lengths of the amino acid sequences ranged from 165 to 1,141 in *accD*.

Phylogenetic Analysis of *V. sepium*

Considering the rather limited number of complete *Vicia* chloroplast genomes (only 3), it is difficult to determine whether *Vicia* is paraphyly. Therefore, in addition to the complete chloroplast genomes and conserved chloroplast protein-coding sequences, we constructed a phylogenetic tree of *Vicia* using two widely sequenced chloroplast genes, namely, *rbcL* and *matK*, to support our result. Detailed information regarding these four datasets can be found in **Table S1**. Upon comparing the four NJ trees, we found that *V. sepium*, *V. sativa*, and *V. faba* were located in the same evolutionary branch with support rates of 100% in the protein-coding sequence tree, 99% in the *matK* tree, and 49% in the *rbcL* tree. However, in the whole-genome tree, the result was different, with *V. sepium* and *V. sativa* located in the same clade and *V. faba* located in another clade. These results indicated that the evolutionary histories of *V. sepium* and *V. sativa* were similar but different from that of *V. faba* (**Figures S10**–**S12**). Both the *rbcL* and *matK* phylogenetic trees showed that *Vicia* species were included in different clades, which supports our hypothesis that *Vicia* is paraphyletic (**Figures S11** and **S12**).

Both the NJ and ML phylogenetic trees for homologous protein-coding sequences showed that *Vicia* and *Lens* were included in the same clade, together with *Pisum* and *Lathyrus* (**Figure 6**), but the ML tree presented a higher support rate for the *Vicia* and *Lens* clade than the NJ tree.

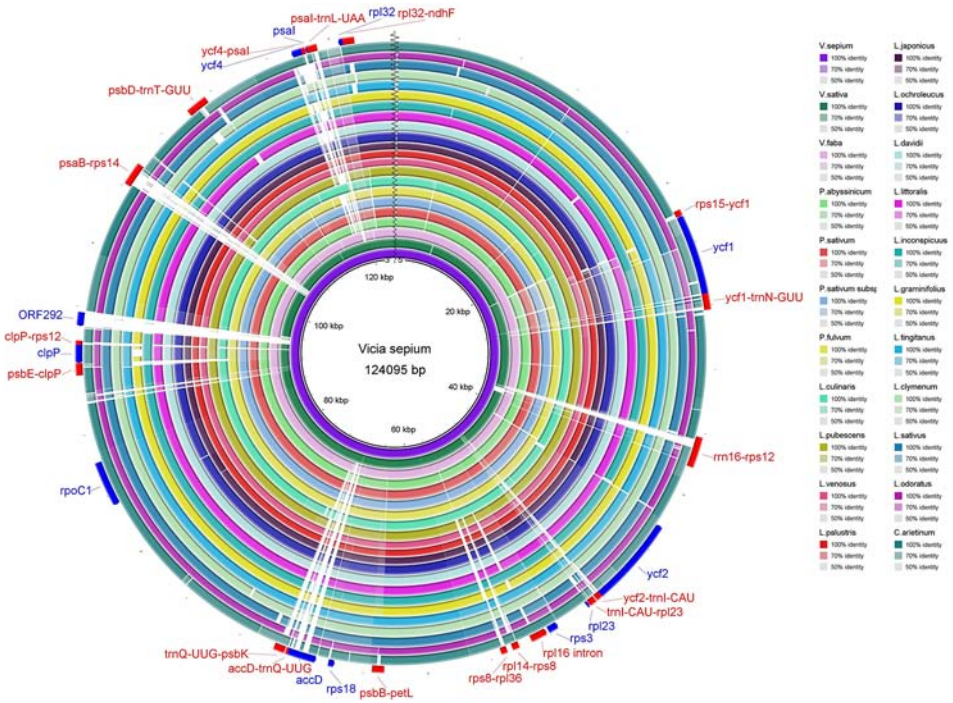


FIGURE 3 | The sequence identity of 22 Fabaceae species. The inner circle is the reference genome. Next circles represent the sequence identity between *V. sepium* and 21 other species. The outermost circle corresponds to the protein-coding genes and intergenic spacer regions. Genes with clockwise arrows represent reverse strands, while genes with the counterclockwise arrow represent forward strands.

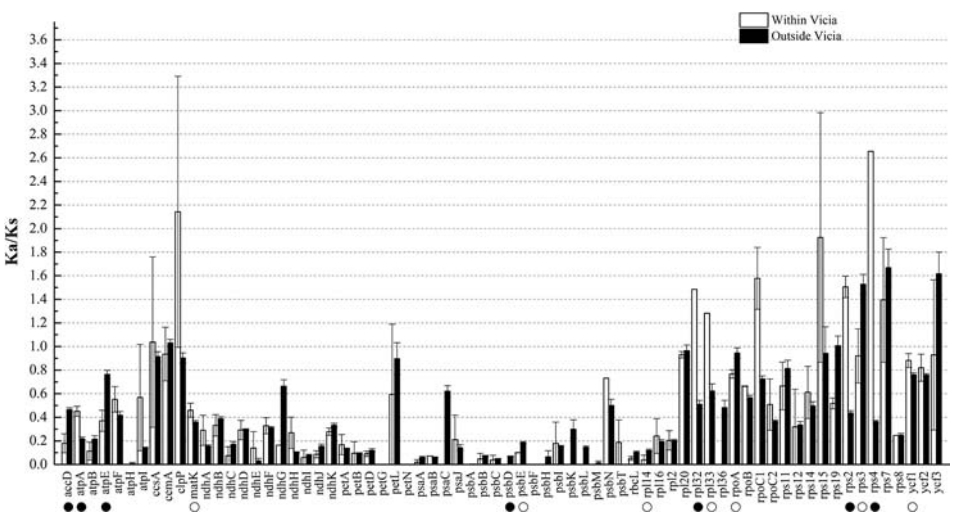


FIGURE 4 | The Ka/Ks ratios of homologous protein-coding genes within and outside of the genus *Vicia* with *V. sepium* as the reference. White boxes represent the mean Ka/Ks values within the genus *Vicia*, and black boxes indicate the mean Ka/Ks values outside of the genus *Vicia*. The data are the arithmetic mean ± SE. Symbols under the gene names indicate levels of statistical significance between the species within *Vicia* and the species outside of *Vicia*: no symbol, $P > 0.05$, blank circle, $P = 0.01$–0.05; black circle, $P < 0.01$. The X-axis denotes the homologous genes.

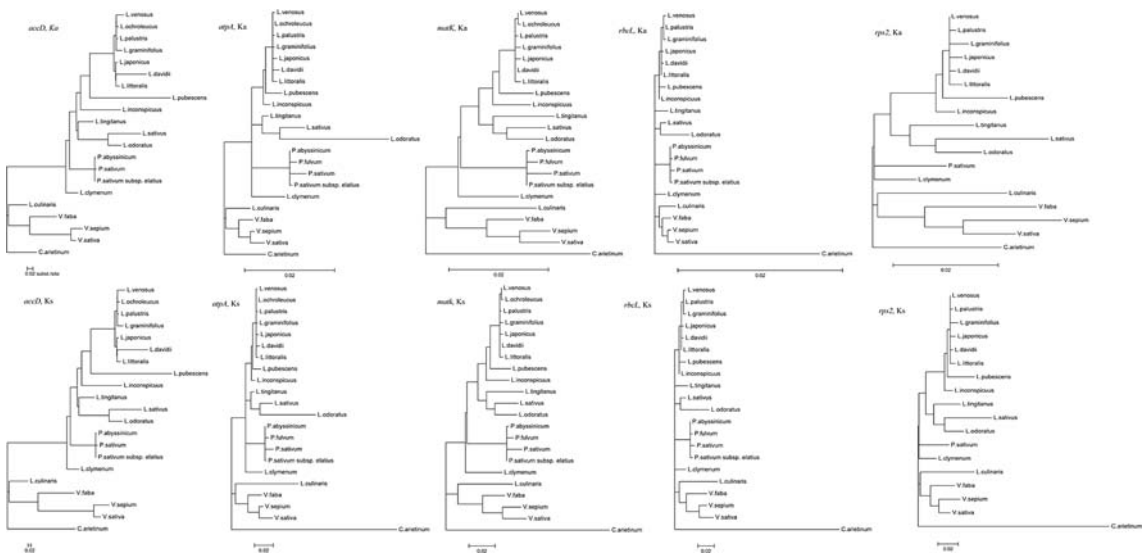


FIGURE 5 | Synonymous and nonsynonymous divergence in the Fabeae chloroplast genes. All tree topologies were completely constrained as described in the Methods section. All trees were drawn to the same scale representing the number of substitutions per synonymous or nonsynonymous site.

DISCUSSION

Beneficial Gene Mutations Observed in the Protein-Coding Regions

In our study, within genus *Vicia*, *ccsA*, *clpP*, *rpl32*, *rpl33*, *rpoC1*, *rps15*, *rps2*, *rps4*, and *rps7* showed positive selection, with a Ka/Ks ratio >1 (**Figure 4**). None of these genes are related to photosynthesis (*psa*, *psb*, *ndh*, *pet*, *atp*). In fact, genes related to photosynthesis were under less selection pressure than other types of genes (Du et al., 2016; Li et al., 2017; Gao et al., 2018). Such positive selection is also found in other species, as observed for two genes flanking *ycf4* (*accD* and *cemA*) in *Lathyrus* (Magee et al., 2010); *accD*, *ycf1*, and *atpA* in seed plants (Zheng et al., 2017); *rps14* in *Dodonaea viscosa* and *Sapindus mukorossi* (Saina et al., 2018); and the *atpF* gene in two deciduous *Quercus* species (Yin et al., 2018). In general, genes under selection pressures are mainly identified by comparing the synonymous and nonsynonymous nucleotide substitution rates in related species. Thus, genes under positive selection pressure in different lineages can be identified. However, the positive selection acting on genes in a specific lineage contrasts with the silent molecular clock hypothesis, according to which the point mutation rate in all regions of the same genome is almost constant (Ochman and Wilson, 1987). The factors causing a higher Ka/Ks ratio in some sequences than in the rest of the genome remain unclear. Here, we consider two explanations for this difference. One possible explanation for this phenomenon is that a greater number of nucleotide substitutions are associated with gene duplications and gene losses. Erixon found that positive selection acting on the *clpP* gene in various plant lineages is related to repeated duplication (Erixon and Oxelman, 2008). Magee showed that the Ka/Ks ratios of *cemA* and *accD* flanking *ycf4* are >1 in *Lathyrus*. This may occur because the increase in the nucleotide mutation rate near the hypermutational *ycf4* gene affects the purifying selection acting on the amino acid sequence (Magee et al., 2010). Another possibility is that differential selection may act on gene divergence. For example, research on oak species showed that the *atpF* gene is highly divergent (Ka/Ks > 1) in the comparation between deciduous oak and evergreen sclerophyllous oak because the former loses its leaves in the cold and drought seasons (Yin et al., 2018). Another study on seed plants suggested that genes affected by positive selection are always involved in plant adaptation, such as *accD*, *ycf1* and *atpA* (Zheng et al., 2017).

We also found that *atpA*, *accD*, and *rps2* of *Vicia* showed significantly accelerated evolution (**Figures 4**, **5**, **S5**–**S7**, **Table S3**). *Rps2*, encoding the ribosomal protein S2, is retained in almost all plants. The exceptions mainly occur in Apocynaceae. For example, in milkweeds, a 2.4-kb mitochondrial DNA sequence was horizontally transferred to the *rps2-rpoC2* plastid intergenic region, resulting in two pseudogenes, namely, *rps2* and *rpoC2*, contained in plastomes (Straub et al., 2013). However, such plastome insertion is rare. A relatively common type of evolution is the point mutation described in our study. For example, Ka and Ks rates are elevated in parasitic Scrophulariaceae and Orobanchaceae, which provide suitable material for studying the evolution of hemi- and holoparasitic plant lineages (dePamphilis et al., 1997). In *Gossypium*, the Yrp8 and Cys11 sites of *rps2* and the other nine genes are undergoing protein sequence evolution, which may aid the adaptation of cotton species to diverse environments (Wu et al., 2018). The accelerated evolution of *atpA* (participating in ATP synthesis) has also been found in other species, such as Dipsacales (Fan

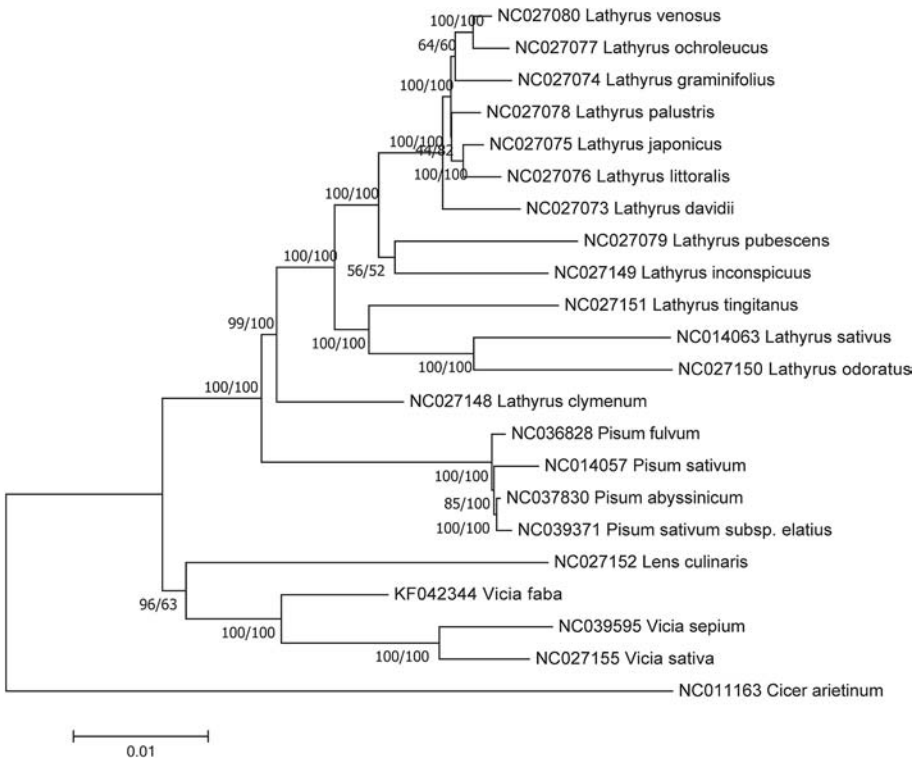


FIGURE 6 | Phylogenetic relationships based on the conserved chloroplast protein-coding sequences of 21 Fabeae species and *C. arietinum* with the maximum likelihood (ML) method and the neighbor joining (NJ) method. *C. arietinum* was selected as the outgroup. Numbers on the left and right side at the branches represent bootstrap values of the ML method and the NJ method respectively.

et al., 2018) and *Urophysa* (Xie et al., 2018) species. Consistent with our study, only one to three sites show positive selection. *AccD* is essential for plant leaf development and has been lost in some angiosperm lineages. It is believed that *accD* was functionally transferred to the nucleus (Magee et al., 2010; Sabir et al., 2014).

At present, *Vicia* is the only known legume genus in which so many genes show positive selection and accelerated evolution in the chloroplast genome. Therefore, a comprehensive understanding of the mechanism underlying the increased nucleotide substitution of homologous protein-coding genes is necessary, and *Vicia* species may be suitable model systems for such studies.

Genome Variation in the Chloroplast Genomes of *V. sepium*

To detect the genome variation in the chloroplast genome of *V. sepium*, we compared *V. sepium* with related genera in the tribe Fabeae. Our results revealed that the greatest variation in genome length relative to *V. sepium* was located in protein-coding regions (**Table 3**). This finding is consistent with Zheng's research (Zheng et al., 2017), showing that chloroplast gene length is an important factor affecting chloroplast genome size based on phylogenetic signals. The length variation of protein-coding regions may result from gene loss and gain or differences in the lengths of homologous genes. *Ycf4*, encoding a photosystem I assembly protein, is the most easily deleted gene in Fabeae species (**Table 3**). This result supports previous findings revealing that *ycf4* has been lost in many species of *Lathyrus* and *Pisum* due to its functional transfer to the nuclear genome (Magee et al., 2010). Furthermore, gene insertion events involving one new unannotated protein-coding gene, namely, *ORF292* (879 bp) and one duplicated gene, namely, *rpl20* (354 bp), were found in *V. sepium and V. sativa*, respectively. One pseudogene, *rpl23*, was identified in *V. sepium* and *V. sativa* (**Table 3**). This indicates that the evolutionary histories of *V. sepium* and *V. sativa* are similar and that *V. faba* may be located in a different evolutionary clade. In general, a chloroplast gene cannot be lost arbitrarily unless the function of the gene is transferred to the nuclear genome or replaced by that of a nuclear gene (Magee et al., 2010). Therefore, the mechanism of loss of the *rpl23* gene in *V. sepium* and *V. sativa* requires further in-depth research. In addition to gene loss, one intron was also missing in *clpP* (*L. graminifolius*) and *rpl16* (*V. faba*) (**Table 3**). The *clpP* gene normally contains two introns in angiosperms (Jansen et al., 2007; Jansen et al., 2008). Jansen determined that the IRLC lineage (in which Fabeae is included) has lost one intron of *clpP* (Jansen et al., 2008). However, the loss of two

introns observed in *clpP* is rare; Sabir's research (Sabir et al., 2014) on the IRLC lineage (in which Fabeae is included) showed that this phenomenon has only occurred in *Glycyrrhiza glabra*, and our findings are complementary to this previous work. *V. faba* was the only species found to have lost the intron of *rpl16* in the tribe Fabeae, and the *rpl16* intron shows high divergence in *Chusquea* (Kelchner and Clark, 1997), *Gleditsia* (Schnabel and Wendel, 1998), and Cacteae (Butterworth et al., 2002). This result indicates that different evolutionary clades exist in *Vicia*. In addition to gene loss and gain, differences in the lengths of homologous genes are also found in Fabeae species (ranging from 495 to 3,423, 36 to 537, and 3,879 to 5,403 in *accD*, *rps12* and *ycf1*, respectively). In seed plants, the length difference in *atpA*, *accD*, and *ycf1* is the main reason for chloroplast genome size variation (Zheng et al., 2017).

In addition to protein-coding region expansion and contraction in *V. sepium*, protein-coding sequence divergence also exists. In our study, the GC content of the chloroplast genome of *V. sepium* was found to be lower than that of other species, such as *Chikusichloa mutica* [tribe rice (Wu et al., 2017)], *Arabidopsis thaliana* [Brassicaceae (Asaf et al., 2017a)], and *Quercus aquifolioides* [Fagaceae (Yin et al., 2018)], which exhibit a conserved structure and evolution of the chloroplast genome (**Table S4**). Normally, a higher GC content indicates a more stable genome sequence (Wu et al., 2017). Therefore, to consider the genome variation in *V. sepium* protein-coding regions, we surveyed SSRs, repeat loci, highly divergent regions and pairwise sequence divergence. Many SSRs and repeat loci appeared in the protein-coding regions (CDSs) (**Table S5**, **Figure 2**). These results are consistent with previous reports on *Astragalus membranaceus* (Lei et al., 2016). Because of the slippage of DNA strands, SSRs, regarded as useful gene markers, present a high mutation rate (Huang et al., 2018). Repeated sequences are believed to result in aberrant replication and repair pathways (Sabir et al., 2014). The genes *ycf1*, *ycf2*, *rpl23*, rps3, *rpl18*, *accD*, *rpoC1*, *clpP*, *ORF292*, *ycf4, psaI*, and *rpl32* share relatively low identity (**Figures 3** and **S4**). *V. sepium* showed considerable differences from other Fabeae species (with the exception of *V. sativa*), even *V. faba*. Therefore, *Vicia* presents profound genome variation, which is significant for the evolutionary history of the chloroplast genome.

Evolution in *Vicia*

The phylogenetic analysis conducted with the conserved chloroplast protein-coding sequences of *rbcL* and *matK* showed that *Vicia* and *Lens* were included in the same clade (**Figures 6** and **S12**). This result is also supported by the synapomorphy that is observable in the currently available research. *Vicia* and *Lens* both produce the phytoalexin wyerone, which is not found in *Pisum* and *Lathyrus* (Schaefer et al., 2012), and show high average protein richness and *in vitro* protein digestibility (Pastor-Cavada et al., 2014). However, even within *Vicia*, different evolutionary directions can be found, resulting in the paraphyly of *Vicia*. For example, in our study, the pairwise distance between *V. sepium* and *V. sativa* was much greater than that between *V. sepium* and *V. faba* (**Table S6**). The former species also showed a gene insertion in the *rps12* to *rps4* region (**Figure S3**) and an accelerated evolutionary rate in *accD* (**Figure 5**). In addition to chloroplast genome characteristics, the life form, stylar characteristics, and chromosome numbers of these species support this result. Ancestral *Vicia* species originating from the Mediterranean shared an annual life form, a basic chromosome number of 2n=14 and evenly hairy styles. However, the recent evolutionary reconstruction of *Vicia* indicates that a perennial life form, a chromosome number of 2n=12 (or 10, 24, 28, 42) and adaxially/abaxially hairy styles have arisen in *Vicia* (Schaefer et al., 2012). In the comparison of *Vicia* species in our study, all of the species were found to produce adaxially hairy styles, but *V. sepium* has evolved a perennial life form, while *V. sativa* and *V. faba* share the same characteristic of an annual life form. Nevertheless, the evolution of the life form of *Vicia* verified that *V. sepium* and *V. sativa* had a shared evolutionary history. Therefore, we can infer from all of these results that *Vicia* species may adopt different evolutionary strategies and that the chloroplast genome provides ideal material for reconstructing the evolutionary history of *Vicia*.

In summary, a new chloroplast genomic resource for an important wild resource plant, *V. sepium*, is presented. This study fills the gap in *V. sepium* genomic resources and provides novel insights into evolutionary dynamics in a poorly studied *Vicia* clade. Our results reveal that *Vicia* species may have experienced many instances of positive selection in the chloroplast genome and accelerated evolution of protein-coding genes, which is rare, being found in only a few angiosperm species. Detailed surveys show that *V. sepium* presents profound genomic variation in terms of *ORF292* gene insertion, *rpl23* pseudogene detection, lower GC content, CDS length variation, and accelerated evolution of the *atpA*, *accD*, and *rps2* genes. Analysis of the phylogenetic relationships show that *Vicia* and *Lens* are included in the same clade and that the evolutionary direction of *V. sepium* and *V. sativa* is different from that of *V. faba*. Therefore, *Vicia* species may be a suitable model system for understanding the mechanisms of chloroplast genome evolution. This study is expected to attract researchers toward *Vicia* species, leading to the identification of further evidence regarding the evolutionary history of the chloroplast genome.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

CL, ZX, and GY conceived the study. All authors collected field samples. CL, JP, and XP analyzed the final data. YZ acquired

funds (2016NK2148, 2016TP2007) for this study. CL wrote the original manuscript, and all authors commented on an early draft of the manuscript.

FUNDING

This work was supported by the Major Science and Technology Program of Hunan Province (2017NK1014), Key Technology R&D Program of Hunan Province (2016NK2148, 2016TP2007, 2017TP2006), Forestry Science and Technology Project of Hunan Province (XLK201825, XLK201920) and Natural Science Foundation of Hunan Province (2019JJ50027).

ACKNOWLEDGMENTS

We would like to thank Wu Liang for providing insightful writing assistance. We would also like to thank the anonymous reviewers for their valuable comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00073/full#supplementary-material

FIGURE S1 | Topology of Fabeae lineages obtained from a concatenated data set consisting of *matK* and *rbcL*. *C. arietinum* was selected as the out group.

FIGURE S2 | Codon usage and relative synonymous codon usage (RSCU) of the *V. sepium* chloroplast genome. The color of the histogram corresponds to the color of the codon. The size of the histogram corresponds to the RSCU of the codon. The X-axis represents different amino acids and the associated codons.

FIGURE S3 | Genomic rearrangement of six Fabeae species relative to *V. sepium*. Locally collinear blocks (LCBs) are colored to indicate syntenic regions. Blocks below the center line indicate regions that align in the reverse complement (inverse) orientation. The small boxes below the LCBs of each chloroplast genome are represented as genes.

FIGURE S4 | Alignment visualization of twenty-two Fabaceae complete chloroplast genomes using *V. sepium* as a reference. The vertical scale indicates the percent identity, ranging from 50% to 100%. Arrows indicate the annotated genes and their transcriptional direction. The different colored boxes correspond to exons, tRNA or rRNA, and noncoding sequences (CNSs).

FIGURE S5 | Alignments of the *accD* protein sequences from Fabeae species.

FIGURE S6 | Alignments of the *atpA* protein sequences from Fabeae species.

FIGURE S7 | Alignments of the *rps2* protein sequences from Fabeae species.

FIGURE S8 | Alignments of the *matK* protein sequences from Fabeae species.

FIGURE S9 | Alignments of the *rbcL* protein sequences from Fabeae species.

FIGURE S10 | Phylogenetic relationships based on the complete chloroplast genomes of twenty-two related species obtained by the neighbor joining (NJ) method. C. arietinum was selected as the outgroup.

FIGURE S11 | Phylogenetic relationships based on *rbcL* gene sequences of 50 Fabeae species, *T. pretense* and *T. repens* obtained by the neighbor joining (NJ) method. *T. pretense* and *T. repens* were selected as the outgroup.

FIGURE S12 | Phylogenetic relationships based on *matK* gene sequences of 62 Fabeae species, *T. pretense* and *T. repens* obtained by the neighbor joining (NJ) method. *T. pretense* and *T. repens* were selected as the outgroup.

REFERENCES

Alikhan, N. F., Petty, N. K., Ben Zakour, N. L., and Beatson, S. A. (2011). BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics* 12, 402. doi: 10.1186/1471-2164-12-402

Asaf, S., Khan, A. L., Khan, M. A., Waqas, M., Kang, S. M., Yun, B. W., et al. (2017a). Chloroplast genomes of Arabidopsis halleri ssp gemmifera and Arabidopsis lyrata ssp petraea: structures and comparative analysis. *Sci. Rep.* 7, 7556. doi: 10.1038/s41598-017-07891-5

Asaf, S., Waqas, M., Khan, A. L., Khan, M. A., Kang, S. M., Imran, Q. M., et al. (2017b). The complete chloroplast genome of Wild Rice (Oryza minuta) and its comparison to related species. *Front. Plant Sci.* 8, 304. doi: 10.3389/fpls.2017.00304

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Butterworth, C. A., Cota-Sanchez, J. H., and Wallace, R. S. (2002). Molecular systematics of Tribe Cacteae (Cactaceae: Cactoideae): a phylogeny based on rpl16 intron sequence variation. *Syst. Bot.* 27, 257–270. doi: 10.1043/0363-6445-27.2.257

Cauz-Santos, L. A., Munhoz, C. F., Rodde, N., Cauet, S., Santos, A. A., Penha, H. A., et al. (2017). The chloroplast genome of passiflora edulis (Passifloraceae) assembled from long sequence reads: structural organization and phylogenomic studies in malpighiales. *Front. Plant Sci.* 8, 334. doi: 10.3389/fpls.2017.00334

Cavalier-Smith, T. (2002). Chloroplast evolution: secondary symbiogenesis and multiple losses. *Curr Biol.* 12, R62–R64. doi: 10.1016/S0960-9822(01)00675-3

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1403. doi: 10.1101/gr.2289704

dePamphilis, C. W., Young, N. D., and Wolfe, A. D. (1997). Evolution of plastid gene rps2 in a lineage of hemiparasitic and holoparasitic plants: many losses of photosynthesis and complex patterns of rate variation. *Proc. Natl. Acad. Sci. U. S. A.* 94, 7367–7372. doi: 10.1073/pnas.94.14.7367

Dong, M., Zhou, X., Ku, W., and Xui, Z. (2019). Detecting useful genetic markers and reconstructing the phylogeny of an important medicinal resource plant, Artemisia selengensis, based on chloroplast genomics. *PloS One* 14, e0211340. doi: 10.1371/journal.pone.0211340

Du, Q., Bi, G., Mao, Y., and Sui, Z. (2016). The complete chloroplast genome of Gracilariopsis lemaneiformis (Rhodophyta) gives new insight into the evolution of family Gracilariaceae. *J. Phycol.* 52, 441–450. doi: 10.1111/jpy.12406

Erixon, P., and Oxelman, B. (2008). Whole-gene positive selection, elevated synonymous substitution rates, duplication, and indel evolution of the chloroplast clpP1 gene. *PloS One* 3, e1386–e1386. doi: 10.1371/journal.pone.0001386

Fan, W. B., Wu, Y., Yang, J., Shahzad, K., and Li, Z. H. (2018). Comparative Chloroplast Genomics of Dipsacales Species: Insights Into Sequence Variation, Adaptive Evolution, and Phylogenetic Relationships. *Front. Plant Sci.* 9, 689–689. doi: 10.3389/fpls.2018.00689

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 32, W273–W279. doi: 10.1093/nar/gkh458

Gao, X., Zhang, X., Meng, H., Li, J., Zhang, D., and Liu, C. (2018). Comparative chloroplast genomes of Paris Sect. Marmorata: insights into repeat regions and evolutionary implications. *BMC Genomics* 19, 878. doi: 10.1186/s12864-018-5281-x

Huang, D. I., and Cronk, Q. C. B. (2015). Plann: a command-line application for annotating plastome sequences. *Appl. In Plant Sci.* 3, 1500026. doi: 10.3732/apps.1500026

Huang, Y. F., Gao, X. L., Nan, Z. B., and Zhang, Z. X. (2017). Potential value of the common vetch (Vicia sativa L.) as an animal feedstuff: a review. *J. Anim. Physiol. Anim. Nutrition* 101, 807–823. doi: 10.1111/jpn.12617

Huang, L. S., Sun, Y. Q., Jin, Y., Gao, Q., Hu, X. G., Gao, F. L., et al. (2018). Development of high transferability cpSSR markers for individual identification and genetic investigation in Cupressaceae species. *Ecol. Evol.* 8, 4967–4977. doi: 10.1002/ece3.4053

Iberite, M., Abbate, G., and Iamonico, D. (2017). Vicia incisa (Fabaceae): taxonomical and chorological notes. *Annali Di Bot.* 7, 57–65. doi: 10.4462/annbotrm-13842

Jaaska, V. (2015). Phylogenetic relationships among sections Vicia, Sepium and Lathyroides of Vicia subgenus Vicia: isozyme evidence. *Biochem. Syst. Ecol.* 62, 186–193. doi: 10.1016/j.bse.2015.08.002

Jansen, R. K., Raubeson, L. A., Boore, J. L., dePamphilis, C. W., Chumley, T. W., Haberle, R. C., et al. (2005). Methods for obtaining and analyzing whole chloroplast genome sequences. *Methods In Enzymol.* 395, 348–384. doi: 10.1016/s0076-6879(05)95020-9

Jansen, R. K., Cai, Z., Raubeson, L. A., Daniell, H., Depamphilis, C. W., Leebens-Mack, J., et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. *Proc. Natl. Acad. Sci. U. S. A.* 104, 19369–19374. doi: 10.1073/pnas.0709121104

Jansen, R. K., Wojciechowski, M. F., Sanniyasi, E., Lee, S.-B., and Daniell, H. (2008). Complete plastid genome sequence of the chickpea (Cicer arietinum) and the phylogenetic distribution of rps12 and clpP intron losses among legumes (Leguminosae). *Mol. Phylogenet. Evol.* 48, 1204–1217. doi: 10.1016/j.ympev.2008.06.013

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kelchner, S. A., and Clark, L. G. (1997). Molecular evolution and phylogenetic utility of the chloroplast rpl16 intron in Chusquea and the Bambusoideae (Poaceae). *Mol. Phylogenet. Evol.* 8, 385–397. doi: 10.1006/mpev.1997.0432

Keller, J., Rousseau-Gueutin, M., Martin, G. E., Morice, J., Boutte, J., Coissac, E., et al. (2017). The evolutionary fate of the chloroplast and nuclear rps16 genes as revealed through the sequencing and comparative analyses of four novel legume chloroplast genomes from Lupinus. *DNA Res.* 24, 343–358. doi: 10.1093/dnares/dsx006

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120. doi: 10.1007/bf01731581

Kruess, A., and Tscharntke, T. (2000). Species richness and parasitism in a fragmented landscape: experiments and field studies with insects on Vicia sepium. *Oecologia* 122, 129–137. doi: 10.1007/pl00008829

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi: 10.1093/nar/29.22.4633

Lei, W. J., Ni, D. P., Wang, Y. J., Shao, J. J., Wang, X. C., Yang, D., et al. (2016). Intraspecific and heteroplasmic variations, gene losses and inversions in the chloroplast genome of Astragalus membranaceus. *Sci. Rep.* 6, 21669. doi: 10.1038/srep21669

Lenoir, L., and Pihlgren, A. (2006). Effects of grazing and ant/beetle interaction on seed production in the legume Vicia sepium in a seminatural grassland. *Ecol. Entomol.* 31, 601–607. doi: 10.1111/j.1365-2311.2006.00818.x

Li, B., Lin, F., Huang, P., Guo, W., and Zheng, Y. (2017). Complete chloroplast genome sequence of decaisnea insignis: genome organization, genomic resources and comparative analysis. *Sci. Rep.* 7, 10073. doi: 10.1038/s41598-017-10409-8

Li, C., Zhao, Y., Huang, H., Ding, Y., Hu, Y., and Xu, Z. (2018). The complete chloroplast genome of an inverted-repeat-lacking species, Vicia sepium, and its phylogeny. *Mitochondrial DNA Part B-Resour.* 3, 137–138. doi: 10.1080/23802359.2018.1431071

Liu, C., Shi, L. C., Zhu, Y. J., Chen, H. M., Zhang, J. H., Lin, X. H., et al. (2012). CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics* 13, 715. doi: 10.1186/1471-2164-13-715

Lohse, M., Drechsel, O., Kahlau, S., and Bock, R. (2013). OrganellarGenomeDRAW-a suite of tools for generating physical maps of plastid and mitochondrial genomes and visualizing expression data sets. *Nucleic Acids Res.* 41, W575–W581. doi: 10.1093/nar/gkt289

Luo, R. B., Liu, B. H., Xie, Y. L., Li, Z. Y., Huang, W. H., Yuan, J. Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *GigaScience* 1, 18. doi: 10.1186/2047-217x-1-18

Magee, A. M., Aspinall, S., Rice, D. W., Cusack, B. P., Semon, M., Perry, A. S., et al. (2010). Localized hypermutation and associated gene losses in legume chloroplast genomes. *Genome Res.* 20, 1700–1710. doi: 10.1101/gr.111955.110

Maršalkienė, N. (2016). Flowering spectrum diversity of Vicia sepium. *Biologija* 62, 116–123. doi: 10.6001/biologija.v62i2.3337

Martin, W., Deusch, O., Stawski, N., Grunheit, N., and Goremykin, V. (2005). Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. *Trends Plant Scie.* 10, 203–209. doi: 10.1016/j.tplants.2005.03.007

Maxted, N. (1995). *An ecogeographical study of Vicia subgenus vicia* (Rome: International Plant Genetic Resources Institute).

Moner, A. M., Furtado, A., Chivers, I., Fox, G., Crayn, D., and Henry, R. J. (2018). Diversity and evolution of rice progenitors in Australia. *Ecol. Evol.* 8, 4360–4366. doi: 10.1002/ece3.3989

Ochman, H., and Wilson, A. C. (1987). Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* 26, 74–86. doi: 10.1007/bf02111283

Pastor-Cavada, E., Juan, R., Pastor, J. E., Alaiz, M., and Vioque, J. (2014). Protein and amino acid composition of select wild legume species of tribe Fabeae. *Food Chem.* 163, 97–102. doi: 10.1016/j.foodchem.2014.04.078

Sabir, J., Schwarz, E., Ellison, N., Zhang, J., Baeshen, N. A., Mutwakil, M., et al. (2014). Evolutionary and biotechnology implications of plastid genome variation in the inverted-repeat-lacking clade of legumes. *Plant Biotechnol. J.* 12, 743–754. doi: 10.1111/pbi.12179

Saina, J. K., Gichira, A. W., Li, Z. Z., Hu, G. W., Wang, Q. F., and Liao, K. (2018). The complete chloroplast genome sequence of Dodonaea viscosa: comparative and phylogenetic analyses. *Genetica* 146, 101–113. doi: 10.1007/s10709-017-0003-x

Schaefer, H., Hechenleitner, P., Santos-Guerra, A., de Sequeira, M. M., Pennington, R. T., Kenicer, G., et al. (2012). Systematics, biogeography, and character evolution of the legume tribe Fabeae with special focus on the middle-Atlantic island lineages. *BMC Evol. Biol.* 12, 250. doi: 10.1186/1471-2148-12-250

Schnabel, A., and Wendel, J. F. (1998). Cladistic biogeography of Gleditsia (Leguminosae) based on ndhF and rpl16 chloroplast gene sequences. *Am. J. Bot.* 85, 1753–1765. doi: 10.2307/2446510

Straub, S. C. K., Cronn, R. C., Edwards, C., Fishbein, M., and Liston, A. (2013). Horizontal Transfer of DNA from the Mitochondrial to the Plastid Genome and Its Subsequent Evolution in Milkweeds (Apocynaceae). *Genome Biol. Evol.* 5, 1872–1885. doi: 10.1093/gbe/evt140

Sveinsson, S., and Cronk, Q. (2016). Conserved gene clusters in the scrambled plastomes of IRLC legumes (Fabaceae: Trifolieae and Fabeae). *bioRxiv* 040188. doi: 10.1101/040188

Thiel, T., Michalek, W., Varshney, R. K., and Graner, A. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). TAG. Theoretical and applied genetics. *Theoretische und angewandte Genetik.* 106, 411–422. doi: 10.1111/j.1556-4029.2011.01810.x

Timmis, J. N., Ayliffe, M. A., Huang, C. Y., and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* 5, 123–135. doi: 10.1038/nrg1271

Wang, Y. H., Wicke, S., Wang, H., Jin, J. J., Chen, S. Y., Zhang, S. D., et al. (2018). Plastid Genome Evolution in the Early-Diverging Legume Subfamily Cercidoideae (Fabaceae). *Front. Plant Sci.* 9, 138. doi: 10.3389/fpls.2018.00138

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191. doi: 10.1093/bioinformatics/btp033

Wicke, S., Schneeweiss, G. M., dePamphilis, C. W., Muller, K. F., and Quandt, D. (2011). The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. *Plant Mol.Biol.* 76, 273–297. doi: 10.1007/s11103-011-9762-4

Wolfe, K. H., Li, W. H., and Sharp, P. M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proc. Natl. Acad. Sci. U. S. A.* 84, 9054–9058. doi: 10.1073/pnas.84.24.9054

Wu, Z. Q., Gu, C. H., Tembrock, L. R., Zhang, D., and Ge, S. (2017). Characterization of the whole chloroplast genome of Chikusichloa mutica and its comparison with other rice tribe (Oryzeae) species. *PLoS One* 12, e0177553. doi: 10.1371/journal.pone.0177553

Wu, Y., Liu, F., Yang, D.-G., Li, W., Zhou, X.-J., Pei, X.-Y., et al. (2018). Comparative chloroplast genomics of gossypium species: insights into repeat sequence variations and phylogeny. *Front. Plant Sci.* 9, 376. doi: 10.3389/fpls.2018.00376

Wyman, S. K., Jansen, R. K., and Boore, J. L. (2004). Automatic annotation of organellar genomes with DOGMA. *Bioinformatics* 20, 3252–3255. doi: 10.1093/bioinformatics/bth352

Xia, X. H. (2017). DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution. *J. Hered.* 108, 431–437. doi: 10.1093/jhered/esx033

Xie, D. F., Yu, Y., Deng, Y. Q., Li, J., Liu, H. Y., Zhou, S. D., et al. (2018). Comparative Analysis of the chloroplast genomes of the Chinese endemic genus urophysa and their contribution to chloroplast phylogeny and adaptive evolution. *Int. J. Mol. Sci.* 19, 1847. doi: 10.3390/ijms19071847

Yang, Z., and Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46, 409–418. doi: 10.1007/pl00006320

Yi, X., Gao, L., Wang, B., Su, Y. J., and Wang, T. (2013). The complete chloroplast genome sequence of Cephalotaxus oliveri (Cephalotaxaceae): evolutionary comparison of cephalotaxus chloroplast DNAs and Insights into the loss of inverted repeat copies in gymnosperms. *Genome Biol. Evol.* 5, 688–698. doi: 10.1093/gbe/evt042

Yin, D. M., Wang, Y., Zhang, X. G., Ma, X. L., He, X. Y., and Zhang, J. H. (2017). Development of chloroplast genome resources for peanut (Arachis hypogaea L.) and other species of Arachis. *Sci. Rep.* 7, 11649. doi: 10.1038/s41598-017-12026-x

Yin, K. Q., Zhang, Y., Li, Y. J., and Du, F. K. (2018). Different natural selection pressures on the atpF gene in evergreen sclerophyllous and deciduous oak species: evidence from comparative analysis of the complete chloroplast genome of quercus aquifolioides with other oak species. *Int. J. Mol. Sci.* 19, 1042. doi: 10.3390/ijms19041042

Zhang, W., Zhao, Y. L., Yang, G. Y., Tang, Y. C., and Xu, Z. G. (2017). Characterization of the complete chloroplast genome sequence of Camellia oleifera in Hainan, China. *Mitochondrial DNA Part B-Resour.* 2, 843–844. doi: 10.1080/23802359.2017.1407687

Zheng, X., Wang, J., Feng, L., Liu, S., Pang, H., Qi, L., et al. (2017). Inferring the evolutionary mechanism of the chloroplast genome size by comparing whole-chloroplast genome sequences in seed plants. *Sci. Rep.* 7. doi: 10.1038/s41598-017-01518-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership