OMICS TECHNOLOGIES TOWARD SYSTEMS BIOLOGY

EDITED BY: Fatemeh Maghuly, Gorji Marzban and Joanna Jankowicz-Cieslak
PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88971-568-8
DOI 10.3389/978-2-88971-568-8

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: frontiersin.org/about/contact

OMICS TECHNOLOGIES TOWARD SYSTEMS BIOLOGY

Topic Editors:
Fatemeh Maghuly, University of Natural Resources and Life Sciences Vienna, Austria
Gorji Marzban, University of Natural Resources and Life Sciences Vienna, Austria
Joanna Jankowicz-Cieslak, International Atomic Energy Agency, Austria

Citation: Maghuly, F., Marzban, G., Jankowicz-Cieslak, J., eds. (2022). Omics Technologies Toward Systems Biology. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88971-568-8

Table of Contents

frontiers in Genetics

EDITORIAL
published: 14 September 2021
doi: 10.3389/fgene.2021.756847

Editorial: Omics Technologies Toward Systems Biology

Fatemeh Maghuly and Gorji Marzban*

Department of Biotechnology, BOKU-VIBT, University of Natural Resources and Life Sciences Vienna, Vienna, Austria

Keywords: multi-omics, data integration, functional genomics, intrinsic and extrinsic stressors, living organisms

Editorial on the Research Topic

Omics Technologies Toward Systems Biology

By the end of Twenty century, analytical methodologies were enabled to explore thousands of biomolecules obtained from any given organisms. Moreover, numerous rising techniques qualified our laboratories to produce enormous amounts of data at different levels by several omics' approaches with exceptional precision, resulting in development of databases and resources (Kumar et al.). However, the dilemma of obtained data remains persistently elaborating information due to the high number and volume. Besides, identification of biomolecules without a comprehensive understanding of the complexity underlying the cellular mechanisms would not deliver much information. Therefore, systems biology employs multi-omics platforms and computational approaches for data integration in different contexts (Krassowski et al.).

Nevertheless, analysis of different types of biomolecules requires extraction protocols compatible with the analytical instrumentation. Therefore, to conduct multi-omics efficiency, aliquots of the same sample are required for different extraction procedures optimized for different biomolecules, thereby decreasing sample handling time and increasing throughput. Kang et al. adapted a biphasic fractionation to extract proteins, metabolites, and lipids from one single sample (3-in-1) for liquid chromatography-tandem mass spectrometry (LC-MS/MS). The results showed that their method has great value to multi-omics and systems biology toward understanding the cellular networks, traits and phenotypes.

On the other hand, the combination of multi-omics data provide useful insight into the flow of biological information at multiple levels, thus can help to elucidate the complex mechanisms controlling the biological condition. For example, Guo et al. combined transcriptome and metabolome data obtained from clonally propagated plants at four developmental stages and three different environments to identify the spatial-temporal variation of flavonoids biosynthesis in leaves of Ginkgo. They indicated that flavonoids content varied considerably at different developmental stages and environments. Therefore, they expect that the accurate selection of planting region(s) and optimization harvesting time would substantially improve the production and management of Gingko in an industrial manner.

In the frame of cell factories and selected targets, multi-omics and systems biology reflect a challenging area for the engineering of cellular metabolism and maximizing the production of valuable compounds through bioconversion. *In silico* experiments were shown to replace time and laborious processes to win information about the cell networks. Tafur Rangel et al. proposed that computational tools and metabolic modeling in combination with transcriptomics can accelerate the optimization of cell factories by identifying key metabolic engineering targets (genes/reactions) and not only by predicting mutants. However, It depends on the level of completeness and accuracy of the metabolic model, which could be improved by omics data.

OPEN ACCESS

Edited and reviewed by:
Kelvin Yuen Kwong Chan, Tsan Yuk Hospital, Hong Kong, SAR China

****Correspondence:***
Fatemeh Maghuly
fatemeh.maghuly@boku.ac.at

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

Received: *11 August 2021*
Accepted: *31 August 2021*
Published: *14 September 2021*

Citation:
Maghuly F and Marzban G (2021) Editorial: Omics Technologies Toward Systems Biology. Front. Genet. 12:756847. doi: 10.3389/fgene.2021.756847

Dictated by the rapidly growing worldwide human population, increased agricultural productivity is necessary to cope with the food demand. In this context, multi-omics technologies have helped plant biologists complete their understanding of plant metabolism by reconsidering and identifying novel pathways. Kumar et al. represented how this new knowledge can be utilized to develop improved cultivars by targeting metabolic pathways and use this information for re-domestication and *de novo* domestication of wild relatives.

Considering that combination of two or more omics data sets in data analysis, visualization and interpretation are essential to determine the mechanism of a biological process; Krassowski et al. provided an excellent overview of the current state of the field, inform on available reliable resources, discuss findable, accessible, interoperable, reusable research, and point to best practices in benchmarking. They also addressed challenges with biological complexity, acknowledged current tools limitations, and concluded future perspectives in this field.

Accordingly, systems biology tools made it possible to develop personalized medicine directly related to analyzing huge amounts of data delivered by high throughput technologies. Pires et al. described how to perform the translation from RNA-seq data into therapeutic targets. They present an online platform using the MEAN stack supported by a Galaxy pipeline for translating RNA-seq data into protein targets suitable for the chemotherapy of solid tumors.

To gain new insight into the evolution of extremophiles and the actual limits for life, in-depth knowledge of proteome-related alterations in cell physiology is crucial. Furthermore, in extreme environments, microbial extremophiles are of great interest to understand stress adaptation and survival mechanisms. Therefore, Tesei et al. pioneered the qualitative and quantitative proteomic analyses on the mycelia, a lack fungi, and supernatant of culture medium to show its ability to cope with microgravity, which has significance to exobiology and implications to planetary protection policies.

Given that acidification of arable lands is one of the biggest problems of modern agronomy, Szurman-Zubrzycka et al. studied the global transcriptome of root meristematic cells from barley grown at low pH treated with aluminium. They showed that low pH is a stress factor; however, aluminium causes more changes at the transcriptome level by long term stress. Thus, aluminium toxicity in acidic soils, resulting in inhibition of both elongation and division rates of root cells, consequently reducing water and nutrient uptake and finally reducing growth and yield.

Taking together, living organisms are innately exposed to a wide range of intrinsic and extrinsic sources, causing damage to the DNA, thereby promoting genomic instability. To escape the harmful effects, organisms harbor several DNA damage repair (DDR) pathways. Because plants are sessile, they have involved highly conserved DDR pathways that share several components with other organisms. In this manner, they maintain their genetic integrity and transfer their accurate genetic information to subsequent plant generations. Raina et al. summarized these complex mechanisms by which plants repair their DNA from severe exposure to both biotic and abiotic stresses and how lack of the DDR pathway affects various developmental stages.

In addition, the sequencing of several genomes based on comparative approaches and recent discoveries of Small Open Reading Frames (small ORFs/sORFs/smORFs) peptides has been recently described as essential players in biological processes and opened new avenues for smORF research, reported as potential non-functional or junk DNA. In this context, Guerra-Almeida and Nunes-da-Fonseca represent intriguing questions to debate further investigation and future perspectives for the non-functional smORF peptides.

On the other hand, an effective high throughput functional genomics tool for studying genes responsible for desired phenotypes is required to facilitate genome-wide investigations. TILLING (Targeting Induced Local Lesions IN Genomes) is a powerful reverse genetics method in plant functional genomics; however, one of the main challenges for a successful TILLING experiment is that currently available bioinformatic tools for variant detection are not designed to identify mutations with low frequencies or to perform sample identification from variants in overlapping pools. To overcome this shortage, Gil et al. developed, through the Next Generation Sequencing Experience Platform, two novel functionalities for TILLING: a TILLING experiment simulator and a TILLING detector. These new bioinformatic tools increase the precision of TILLING experiments, which is useful for implementing TILLING as a tool for functional genomics and breeding.

In this context, Guo et al. also identified loss-of-function mutations in blast susceptible genes through TILLING by sequencing. Furthermore, they suggested that identified mutants might also provide enhanced immunity with severe effects on protein function and resistance to wheat blast. Thus, the study provides a new strategy, novel resistant lines, and valuable gene resources to tackle disease-resistant wheat breeding.

We wish to thank all contributors to this special issue and hope that its appearance provides interest to users recent and novel research trends in the application of omics technologies.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Genetics

OPINION
published: 20 October 2020
doi: 10.3389/fgene.2020.574737

Small Open Reading Frames: How Important Are They for Molecular Evolution?

Diego Guerra-Almeida[1]* *and Rodrigo Nunes-da-Fonseca*[1,2]*

[1] *Institute of Biodiversity and Sustainability, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil,* [2] *National Institute of Science and Technology in Molecular Entomology, Rio de Janeiro, Brazil*

Keywords: small ORF, junk DNA, coding potential maturation, causal roles, selected effects, non-coding RNA, alternative ORF, pervasive translation

OPEN ACCESS

Edited by:
Fatemeh Maghuly, University of Natural Resources and Life Sciences Vienna, Austria

Reviewed by:
Benedikt Obermayer, Charité Medical University of Berlin, Germany
Mona Wu Orr, Amherst College, United States

***Correspondence:**
Diego Guerra-Almeida
diegoguerra@ufrj.br
Rodrigo Nunes-da-Fonseca
rnfonseca@macae.ufrj.br

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

Received: *21 June 2020*
Accepted: *25 August 2020*
Published: *20 October 2020*

Citation:
Guerra-Almeida D and Nunes-da-Fonseca R (2020) Small Open Reading Frames: How Important Are They for Molecular Evolution? Front. Genet. 11:574737. doi: 10.3389/fgene.2020.574737

INTRODUCTION

Small Open Reading Frames (small ORFs/sORFs/smORFs) are important sources of putative peptides previously dismissed as being non-functional or junk DNA, as determined by early gene prediction methods. In fact, smORFs of <100 codons are possible coding sequences but sufficiently small to occur very frequently and randomly in genomes; thus, the detection of their coding potential and functional assessment is similar to a walk in the dark. Furthermore, while dozens of smORF peptides have been recently described as essential players in biological processes, many are reported to be potential non-functional products of junk DNA under pervasive translation, leading to the question: from what perspective is this lack of function assessed? In this context, it was recently suggested that non-functional smORF peptides might play a major role during *de novo* protein coding gene birth, but the evolutionary mechanism is still unclear. Thus, the role of pervasive translation of smORFs in molecular evolution remains puzzling. Here, we present interesting questions for debate and further investigation about the perspective of non-functional smORF peptides as underappreciated hotspots of molecular evolution in eukaryotes.

SMALL OPEN READING FRAMES: A SUBTOPIC IN THE DISCUSSION OF JUNK DNA FUNCTION

With respect to the evolution of molecular function, part of the DNA elements accumulate mutations by genetic drift; thus, the evolution of these elements is non-adaptive and neutral (Ohta, 2002). In some cases, the amount of neutrally evolving elements in junk DNA are analogous to the items on a menu available to natural selection (Knibbe et al., 2007; Faulkner and Carninci, 2009; Lynch et al., 2011). Interestingly, it was reported by the ENCODE consortium (the Encyclopedia of DNA Elements) that most of the human junk DNA exhibits some type of biochemical activity (ENCODE Project Consortium, 2012), but lacking adaptive relevance and selective pressure (Doolittle, 2013; Graur et al., 2013). Importantly, junk DNA represents 75–90% of the human genome (Graur, 2017).

Part of the junk DNA menu is composed of neutrally evolving smORF peptides. For instance, thousands of non-coding RNAs are generated by the extensive transcription coverage on junk DNA (ENCODE Project Consortium, 2007). Increasing evidence shows that thousands of smORFs undergo pervasive translation in transcripts annotated as non-coding or in untranslated regions (UTR) of mRNAs (e.g., Aspden et al., 2014; Ingolia et al., 2014). Interestingly, non-coding RNAs and ORFs lacking homologs were reported to be candidates for *de novo* evolution of protein coding genes (Tautz and Domazet-Lošo, 2011). Moreover, it was recently suggested that neutrally evolving smORF peptides might play a major role in this process (Ruiz-Orera et al., 2018), but the

evolutionary mechanism remains to be determined (Ruiz-Orera et al., 2018; Singh and Wurtele, 2020). In this context, two previously proposed concepts used to discuss molecular function evolution are at the core of the junk DNA debate: "causal roles" and "selected effects" (Doolittle and Brunet, 2017), which will be discussed here in the context of smORFs and protein coding gene birth.

The "causal role" describes the activity performed by a neutrally evolving element by chance. For example, a hypothetical genomic sequence generated by a random nucleotide mutation to resemble a TATA box may be recognized and bound by transcription factors but does not trigger gene transcription (Griffiths, 2009; Graur et al., 2013). In other words, "causal roles" are non-adaptive phenotypes, their emergence is random, and they tend to rapidly disappear during evolution. On the other hand, "selected effects" describe the acquisition of adaptive phenotypes based on natural selection (Graur et al., 2013), such as canonical TATA boxes or ORFs that are translated into important proteins. In other words, "selected effects" are functionally relevant for cells.

Importantly, while natural selection drives adaptive evolution (selected effects), it is widely accepted that genetic drift drives junk DNA evolution, as well as the synonymous modifications in coding DNA sequences (CDS) and mutations in UTRs of mRNAs (Ridley, 2004).

DISCUSSION

Applying the aforementioned evidence and concepts, we discuss here a possible eukaryotic mechanism by which neutrally evolving smORFs advance proteome evolution and the evolutionary significance of smORFs.

Firstly, part of the roles performed by neutrally evolving smORF peptides possibly transit from "causal roles" to "selected effects" under environmental pressure, thereby exposing their neutral phenotypes to natural selection and triggering the evolution of new coding genes. Thus, when neutral smORF peptides are selected, they are no longer neutral (Ruiz-Orera et al., 2018). In other words, neutral smORF peptides may be special entrees on the junk DNA menu that are available for natural selection (**Figure 1A**).

Upon smORFs being selected for, they probably contain low adaptive relevance due to their non-coding transcript characteristics, such as low translation rate, lack of 3′-terminal processing and other suboptimal coding features (non-coding RNA features are reviewed in Quinn and Chang, 2016). This hypothesis is based on the fact that hundreds of smORFs are described as highly conserved but display low expression, low translation efficiency and are observed in transcripts with non-coding characteristics (Cabili et al., 2011; Aspden et al., 2014; Bazzini et al., 2014). However, the nearly neutral theory (Ohta, 2002) suggests that non-coding parts of fixed smORF transcripts are modified by random genetic drift, in some cases, producing small advantageous (or disadvantageous) adaptive effects throughout evolution; thus, we propose that, at a certain point, these modifications refine and elevate the coding potential of smORF transcripts and consequently enhance the adaptive relevance of their peptides, as seen in a large number of important smORF peptides recently discovered (e.g., Magny et al., 2013; Anderson et al., 2015; Lauressergues et al., 2015; Nelson et al., 2016; Pengpeng et al., 2017; Kim et al., 2018; Polycarpou-Schwarz et al., 2018; Chugunova et al., 2019; Tobias-Santos et al., 2019; Pang et al., 2020; Vassallo et al., 2020). Importantly, the acquisition of several optimal coding features might be favored after the smORF has been selected for, because modifications driven by genetic drift could be fixed by natural selection if they improve the translation efficiency of the newly selected smORF. Before the smORF has been selected for, eventual optimal coding features acquired in the nucleotide sequence could rapidly disappear during genetic drift evolution without fixation. Alternatively, nucleotide changes may negatively affect the coding potential and silence a gene. Optimal coding features include structural stabilization, emergence of Kozak consensus, internal ribosome entry sites (IRES), coverage by enhancers and, in some cases, the elongation of coding smORFs to enlarge the CDSs (**Figure 1B**). Recently, Couso and Patraquim (2017) proposed that at least a portion of functional smORFs are potential *de novo* precursors of large CDSs via a stop codon mutation pattern called "CDS elongation."

Considering the supposition that the action of evolution is gradual, we propose that the aforementioned process be called "coding potential maturation" (**Figure 1B**). For example, smORF translation is widely reported in transcripts with long non-coding RNA (lncRNA) characteristics (Crappé et al., 2013; Ingolia et al., 2014; Ji et al., 2015; Mackowiak et al., 2015; Li et al., 2018; Lu et al., 2019). These lncRNAs exhibit smORF conservation in divergent species, hinting at natural selection fixation and indicating coding immaturity.

Another potential pathway of coding gene generation occurs via alternative smORFs in UTRs or overlapping the reference CDS of canonical mRNAs. In this scenario, alternative smORFs undergo pervasive translation or the act of translation itself is important for cis-regulatory purposes (Vanderperre et al., 2013; Wu et al., 2020). If the "causal roles" performed by neutrally evolving smORF peptides become "selected effects," the alternative smORFs would generate independent gene units by retrotransposition, or they would be fixed as alternative smORFs in the original transcripts (**Figure 1B**). Hence, during retrotransposition events, at least a portion of the transcripts investigated on the basis of pseudogenization may, in fact, represent the maturation of new coding genes, as suggested by a report that pseudogenes can be translated into highly conserved smORF peptides (Ji et al., 2015).

smORFs might be sequence reservoirs potentially activated during the evolution of new phenotypic variations, especially during speciation. Importantly, speciation events have been associated with the evolution of new molecular phenotypes and new relationships with the environment (Bao et al., 2018). Thus, the amount of junk DNA and lncRNAs in cells deserves investigation not only as a random accumulation of sequences and translational noise but also as a repository of substrates to advance the evolution of new coding genes. Interestingly, polyploidization, or whole genome duplication

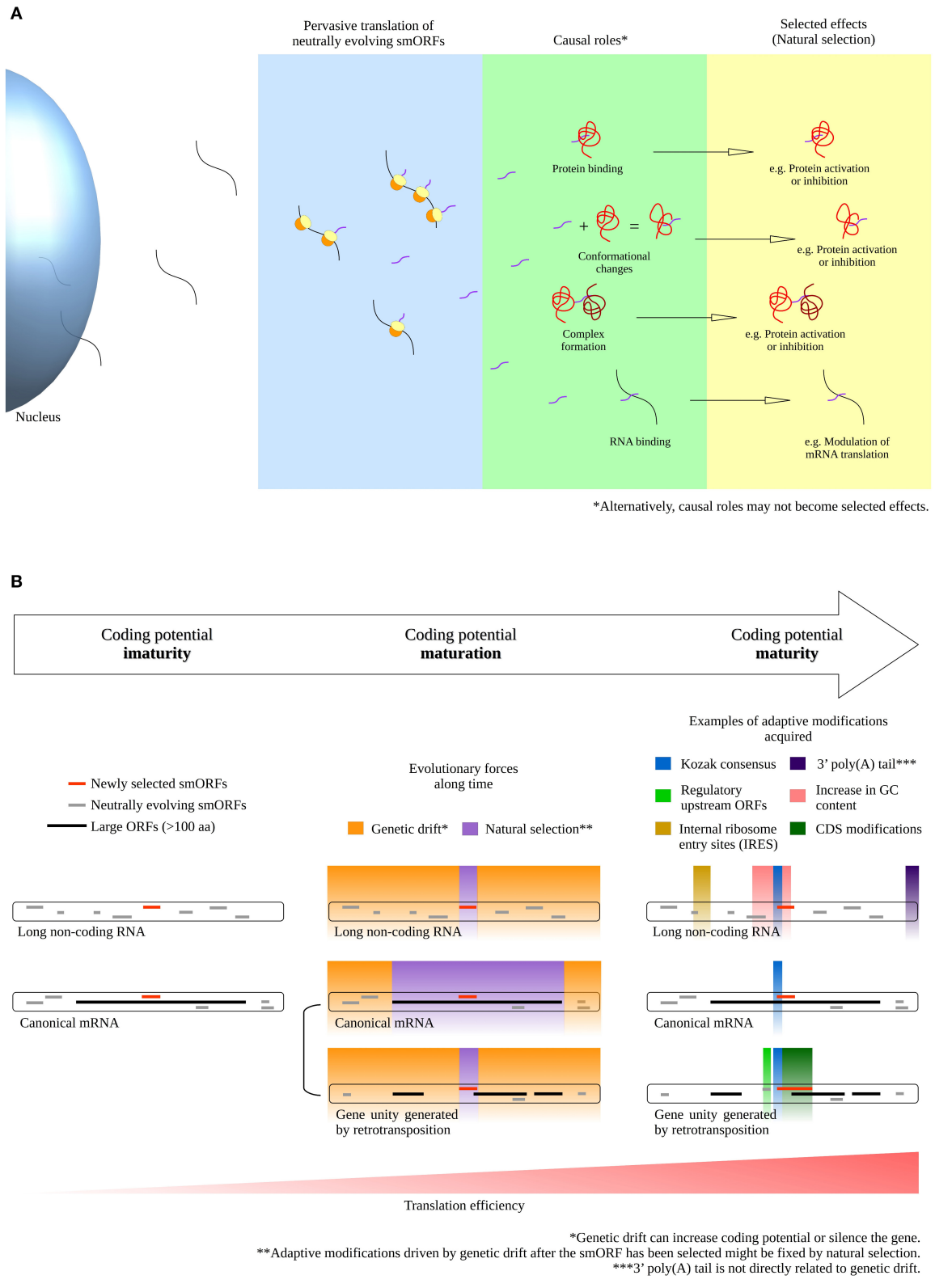

FIGURE 1 | Phenotype selection and coding potential maturation of smORF transcripts. **(A)** Transition of smORF peptides from "causal roles" to "selected effects" after pervasive translation events. Pervasive translation of neutrally evolving smORFs possibly advances proteome evolution by exposing neutral

(Continued)

FIGURE 1 | phenotypes to natural selection under environmental pressure. **(B)** Scheme for coding potential maturation, a hypothetical mechanism that increase the translation efficiency of a mRNA after a smORF has been selected for (selected effect) in a transcript with suboptimal coding features. On the left, coding potential immaturity; in the middle, coding potential maturation; on the right, coding potential maturity. During the coding potential immaturity phase, newly selected smORFs are observed in transcripts with suboptimal coding features, either in long non-coding RNAs or as alternative smORFs in canonical mRNAs. Although canonical mRNAs exhibit optimal coding features, alternative smORFs are usually secondarily or pervasively translated; thus, some alternative smORFs may reside in suboptimal coding regions. During the coding potential maturation phase, natural selection and genetic drift may act in different parts of a transcript. While natural selection acts by fixing the selected parts, genetic drift acts by changing the non-coding parts of a transcript, as postulated by the nearly neutral theory (Ohta, 2002). Natural selection promotes fine-tuned adjustments to the selected phenotypes, such as synonymous mutations and CDS modifications. Genetic drift can establish adaptive mutations in a transcript by evolving sequences that potentially increase smORF translation, such as the Kozak consensus, regulatory upstream ORFs, internal ribosome entry sites (IRES) and increases in GC content. Additionally, other adaptive modifications not directly related to sequence mutations in transcripts might increase smORF expression, such as the 5′ cap, 3′ poly(A) tail, cis-regulatory elements in the genome and, in the case of alternative smORFs, independent gene unit generation by retrotransposition. Importantly, the acquisition of optimal coding features might be favored after the smORF has been selected for, because modifications driven by genetic drift could be fixed by natural selection if they improve the translation efficiency of the newly selected smORF. Before the smORF has been selected for, eventual optimal coding features acquired could rapidly disappear during genetic drift evolution without fixation. Alternatively, mutations evolved by genetic drift can silence the gene. Finally, smORFs reach the coding potential maturity phase when optimal coding features are acquired and translation efficiency increases. Consequently, the translation rate of smORF peptides is largely increased upon completion of the described process, contributing to the establishment of molecular innovations and protein coding gene birth.

(WGD) events, have been correlated with an increase in the adaptive potential of cells and organisms exposed to stressful conditions (Van De Peer et al., 2017). Unfortunately, thus far, studies of WGD have neglected the role and retention of smORFs during evolution, probably due to methodological difficulties in smORF identification.

However, the sequencing of several genomes based on comparative approaches has recently opened new avenues for smORF research. For instance, recent evolutionary studies performed by our group on the smORFs in the *millepattes/tarsalless/polished rice* (*mlpt*) gene, the most well-known smORF-containing gene in insects (Savard et al., 2006; Kondo et al., 2007; Pueyo and Couso, 2008, 2011; Cao et al., 2017; Ray et al., 2019), showed that a new ∼80 amino acid smORF (smHemiptera) appeared during Hemiptera evolution (Tobias-Santos et al., 2019). Thus, this smORF in the polycistronic *mlpt* mRNA has been conserved for over 250 million years in the group, and it is not present in the genomes of other insect orders. We expect that new comparative analyses of genomes in the future will yield additional examples of order-specific smORFs, which might constitute an underappreciated reservoir of new genes and evolutionary innovations.

In summary, the study of smORFs has been considerably increasing during the last 5 years because of recent discoveries of important smORF peptides. Accordingly, the advent of ribosome profiling has allowed the discovery of many neutrally evolving and potentially non-functional smORFs undergoing pervasive translation, whose significance remains to be determined (Crappé et al., 2013; Aspden et al., 2014; Bazzini et al., 2014; Olexiouk et al., 2016). In this context, the intriguing question is posed: why would cells spend energy on transcription and translation of neutral and non-functional elements? There is probably more than one answer; however, considering the subjects discussed in this paper, we propose the following perspective: what if the pervasive translation of neutrally evolving smORF peptides composes an elegant mechanism to advance proteome evolution, especially during speciation events? If it does, then non-functional smORF peptides display an important function in an evolutionary sense. Based on this discussion, we suggest that the concept of functionality be revised in the context of smORFs.

AUTHOR CONTRIBUTIONS

DG-A and RN contributed equally to the writing of this manuscript. RN contributed to funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

RN was supported by CNPq (307952/2017-7 and 431354/2016-2) and FAPERJ (E-26/210-150/2016, E-26/203.298/2016, E-26/202.605/2019, and E-26/211.169/2019). DG-A was a master's student of PPG-PRODBIO-UFRJ/Macaé (CAPES scholarship).

REFERENCES

Anderson, D. M., Anderson, K. M., Chang, C. L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., et al. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell* 160, 595–606. doi: 10.1016/j.cell.2015.01.009

Aspden, J. L., Eyre-Walker, Y. C., Phillips, R. J., Amin, U., Mumtaz, M. A. S., Brocard, M., et al. (2014). Extensive translation of small open reading frames revealed by poly-ribo-seq. *eLife* 3:e03528. doi: 10.7554/eLife.03528

Bao, R., Dia, S. E., Issa, H. A., Alhusein, D., and Friedrich, M. (2018). Comparative evidence of an exceptional impact of gene duplication on the developmental evolution of *Drosophila* and the higher Diptera. *Front. Ecol. Evol.* 6:63. doi: 10.3389/fevo.2018.00063

Bazzini, A. A., Johnstone., T. G., Christiano, R., Mackowiak, S. D., Obermayer, B., Fleming, E. S., et al. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* 33, 981–993. doi: 10.1002/embj.201488411

Cabili, M., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs

reveals global properties and specific subclasses. *Genes Dev*. 25, 1915–1927. doi: 10.1101/gad.17446611

Cao, G., Gong, Y., Hu, X., Zhu, M., Liang, Z. C., Huang, L., et al. (2017). Identification of tarsal-less peptides from the silkworm *Bombyx mori*. *Appl. Microbiol. Biotechnol*. 102, 1809–1822. doi: 10.1007/s00253-017-8708-4

Chugunova, A., Loseva, E., Mazin, P., Mitina, A., Navalayeu, T., Bilan, D., et al. (2019). LINC00116 codes for a mitochondrial peptide linking respiration and lipid metabolism. *Proc. Natl. Acad. Sci. U.S.A*. 116, 4940–4945. doi: 10.1073/pnas.1809105116

Couso, J., and Patraquim, P. (2017). Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol*. 18, 575–589. doi: 10.1038/nrm.2017.58

Crappé, J., Criekinge, W. V., Trooskens, G., Hayakawa, E., Luyten, W., Baggerman, G., et al. (2013). Combining *in silico* prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics* 14:648. doi: 10.1186/1471-2164-14-648

Doolittle, W. F. (2013). Is junk DNA bunk? A critique of ENCODE. *Proc. Natl. Acad. Sci. U.S.A*. 110, 5294–5300. doi: 10.1073/pnas.1221376110

Doolittle, W. F., and Brunet, T. D. P. (2017). On causal roles and selected effects: our genome is mostly junk. *BMC Biol*. 15:116. doi: 10.1186/s12915-017-0460-9

ENCODE Project Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447, 799–816. doi: 10.1038/nature05874

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247

Faulkner, G. J., and Carninci, P. (2009). Altruistic functions for selfish DNA. *Cell Cycle* 8, 2895–900. doi: 10.4161/cc.8.18.9536

Graur, D. (2017). An upper limit on the functional fraction of the human genome. *Genome Biol. Evol*. 9, 1880–1885. doi: 10.1093/gbe/evx121

Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., Elhaik, E., et al. (2013). On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol*. 5, 578–590. doi: 10.1093/gbe/evt028

Griffiths, P. E. (2009). In what sense does "nothing make sense except in the light of evolution?" *Acta Biotheor*. 57:11. doi: 10.1007/s10441-008-9054-9

Ingolia, N. T., Brar, G. A., Stern-Ginossar, N., Harris, M. S., Talhouarne, G. J. S., Jackson, S. E., et al. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep*. 8, 1365–1379. doi: 10.1016/j.celrep.2014.07.045

Ji, Z., Song, R., Regev, A., and Struhl, K. (2015). Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* 4:e08890. doi: 10.7554/eLife.08890

Kim, K. H., Son, J. M., Benayoun, B. A., and Lee, C. (2018). The mitochondrial-encoded peptide MOTS-c translocates to the nucleus to regulate nuclear gene expression in response to metabolic stress. *Cell Metab*. 28, 516.e7–524.e7. doi: 10.1016/j.cmet.2018.06.008

Knibbe, C., Coulon, A., Mazet, O., Fayard, J. M., and Beslon, G. (2007). A long-term evolutionary pressure on the amount of noncoding DNA. *Mol. Biol. Evol*. 24, 2344–2353. doi: 10.1093/molbev/msm165

Kondo, T., Hashimoto, Y., Kato, K., Inagaki, S., Hayashi, S., and Kageyama, Y. (2007). Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat. Cell Biol*. 9, 660–665. doi: 10.1038/ncb1595

Lauressergues, D., Couzigou, J. M., Clemente, H. S., Martinez, Y., Dunand, C., Bécard, G., et al. (2015). Primary transcripts of microRNAs encode regulatory peptides. *Nature* 520, 90–93. doi: 10.1038/nature14346

Li, H., Xiao, L., Zhang, L., Wu, J., Wei, B., Sun, N., et al. (2018). FSPP: A tool for genome-wide prediction of smORF-encoded peptides and their functions. *Front. Genet*. 9:96. doi: 10.3389/fgene.2018.00096

Lu, S., Zhang, J., Lian, X., Sun, L., Meng, K., Chen, Y., et al. (2019). A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Res*. 47, 8111–8125. doi: 10.1093/nar/gkz646

Lynch, M., Bobay, L. M., Catania, F., Gout, J. F., and Rho, M. (2011). The repatterning of eukaryotic genomes by random genetic drift. *Annu. Rev. Genomics Hum. Genet*. 12, 347–366. doi: 10.1146/annurev-genom-082410-101412

Mackowiak, S. D., Zauber, H., Bielow, C., Thiel, D., Kutz, K., Calviello, L., et al. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol*. 16:179. doi: 10.1186/s13059-015-0742-x

Magny, E. G., Pueyo, J. I., Pearl, F. M. G., Cespedes, M. A., Niven, J. E., Bishop, S. A., et al. (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames. *Science* 341, 1116–1120. doi: 10.1126/science.1238802

Nelson, B. R., Makarewich, C. A., Anderson, D. M., Winders, B. R., Troupes, C. D., Wu, F., et al. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* 351, 271–275. doi: 10.1126/science.aad4076

Ohta, T. (2002). Near-neutrality in evolution of genes and gene regulation. *Proc. Natl. Acad. Sci. U.S.A*. 99, 16134-16137. doi: 10.1073/pnas.252626899

Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L., and Menschaert, G. (2016). sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res*. 44, D324–D329. doi: 10.1093/nar/gkv1175

Pang, Y., Liu, Z., Han, H., Wang, B., Li, W., Mao, C., et al. (2020). Peptide SMIM30 promotes HCC development by inducing SRC/YES1 membrane anchoring and MAPK pathway activation. *J. Hepatol*. doi: 10.1016/j.jhep.2020.05.028. [Epub ahead of print].

Pengpeng, B., Ramirez-Martinez, A., Li, H., Cannavino, J., McAnally, J. R., Shelton, J. M., et al. (2017). Control of muscle formation by the fusogenic micropeptide myomixer. *Science* 356, 323–327. doi: 10.1126/science.aam9361

Polycarpou-Schwarz, M., Groß, M., Mestdagh, P., Schott, J., Grund, S. E., Hildenbrand, C., et al. (2018). The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene* 37, 4750–4768. doi: 10.1038/s41388-018-0281-5

Pueyo, J. I., and Couso, J. P. (2008). The 11-aminoacid long tarsal-less peptides trigger a cell signal in *Drosophila* leg development. *Dev. Biol*. 324, 192–201. doi: 10.1016/j.ydbio.2008.08.025

Pueyo, J. I., and Couso, J. P. (2011). Tarsal-less peptides control Notch signalling through the shavenbaby transcription factor. *Dev. Biol*. 355, 183–1936. doi: 10.1016/j.ydbio.2011.03.033

Quinn, J. J., and Chang, H. Y. (2016). Unique features of long non-coding RNA biogenesis and function. *Nat. Rev. Genet*. 17, 47–62. doi: 10.1038/nrg.2015.10

Ray, S., Rosenberg, M. I., Chanut-Delalande, H., Decaras, A., Schwertner, B., Toubiana, W., et al. (2019). The mlpt/Ubr3/Svb module comprises an ancient developmental switch for embryonic patterning. *eLife* 8:e39748. doi: 10.7554/eLife.39748

Ridley, M. (2004). *Evolution, 3rd edn*. Oxford: Blackwell Pub.

Ruiz-Orera, J., Verdaguer-Grau, P., Villanueva-Cañas, J. L., Messeguer, X., and Alb,à, M. M. (2018). Translation of neutrally evolving peptides provides a basis for *de novo* gene evolution. *Nat. Ecol. Evol*. 2, 890–896. doi: 10.1038/s41559-018-0506-6

Savard, J., Marques-Souza, H., Aranda, M., and Tautz, D. (2006). A segmentation gene in *Tribolium* produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell* 126, 559–569. doi: 10.1016/j.cell.2006.05.053

Singh, U., and Wurtele, E. S. (2020). How new genes are born. *eLife* 2020:e55136. doi: 10.7554/eLife.55136

Tautz, D., and Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nat. Rev. Genet*. 12, 692–702. doi: 10.1038/nrg3053

Tobias-Santos, V., Guerra-Almeida, D., Mury, F., Ribeiro, L., Berni, M., Araujo, H., et al. (2019). Multiple roles of the polycistronic gene *Tarsal-Less*/*Mille-Pattes*/*Polished-Rice* during embryogenesis of the kissing bug *Rhodnius prolixus*. *Front. Ecol. Evol*. 7:379. doi: 10.3389/fevo.2019.00379

Van De Peer, Y., Mizrachi, E., and Marchal, K. (2017). The evolutionary significance of polyploidy. *Nat. Rev. Genet*. 18, 411–424. doi: 10.1038/nrg.2017.26

Vanderperre, B., Lucier, J. F., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., et al. (2013). Direct detection of alternative open reading

frames translation products in human significantly expands the proteome. *PLoS ONE* 8:e70698. doi: 10.1371/journal.pone.0070698

Vassallo, A., Palazzotto, E., Renzone, G., Botta, L., Faddetta, T., Scaloni, A., et al. (2020). The *Streptomyces coelicolor* small ORF *trpM* stimulates growth and morphological development and exerts opposite effects on actinorhodin and calcium-dependent antibiotic production. *Front. Microbiol.* 11:224. doi: 10.3389/fmicb.2020.00224

Wu, Q., Wright, M., Gogol, M. M., Bradford, W. D., Zhang, N., and Bazzini, A. (2020). Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J.* 39:e104763. doi: 10.15252/embj.2020104763

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: 27 November 2020
doi: 10.3389/fgene.2020.589326

Temporospatial Flavonoids Metabolism Variation in *Ginkgo biloba* Leaves

Ying Guo[1,2,3], Tongli Wang[3], Fang-Fang Fu[1,2], Yousry A. El-Kassaby[3] and Guibin Wang[1,2]**

[1] Co-Innovation Centre for Sustainable Forestry in Southern China, Nanjing Forestry University, Nanjing, China, [2] College of Forestry, Nanjing Forestry University, Nanjing, China, [3] Department of Forest & Conservation Sciences, Faculty of Forestry, The University of British Columbia, Vancouver, BC, Canada

Ginkgo (*Ginkgo biloba* L.) is a high-value medicinal tree species characterized by its flavonoids beneficial effects that are abundant in leaves. We performed a temporospatial comprehensive transcriptome and metabolome dynamics analyses of clonally propagated Ginkgo plants at four developmental stages (time: May to August) across three different environments (space) to unravel leaves flavonoids biosynthesis variation. Principal component analysis revealed clear gene expression separation across samples from different environments and leaf-developmental stages. We found that flavonoid-related metabolism was more active in the early stage of leaf development, and the content of total flavonoid glycosides and the expression of some genes in flavonoid biosynthesis pathway peaked in May. We also constructed a co-expression regulation network and identified eight *GbMYB*s and combining with other TF genes (3 *GbERF*s, 1 *GbbHLH,* and 1 *GbTrihelix*) positively regulated the expression of multiple structural genes in the flavonoid biosynthesis pathway. We found that part of these *GbTF*s (Gb_11316, Gb_32143, and Gb_00128) expressions was negatively correlated with mean minimum temperature and mean relative humidity, while positively correlated with sunshine duration. This study increased our understanding of the molecular mechanisms of flavonoids biosynthesis in Ginkgo leaves and provided insight into the proper production and management of Ginkgo commercial plantations.

Keywords: *Ginkgo biloba*, flavonoids biosynthesis, leaf development, transcriptome dynamics, temporospatial variation

OPEN ACCESS

Edited by:
Gorji Marzban,
University of Natural Resources and Life Sciences, Vienna, Austria

Reviewed by:
Biao Jin,
Yangzhou University, China
Feng Xu,
Yangtze University, China

****Correspondence:***
Guibin Wang
guibinwang99@163.com
Yousry A. El-Kassaby
y.el-kassaby@ubc.ca

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

Received: *30 July 2020*
Accepted: *05 November 2020*
Published: *27 November 2020*

Citation:
Guo Y, Wang T, Fu F-F, El-Kassaby YA and Wang G (2020) Temporospatial Flavonoids Metabolism Variation in Ginkgo biloba Leaves. Front. Genet. 11:589326. doi: 10.3389/fgene.2020.589326

INTRODUCTION

Ginkgo (*Ginkgo biloba* L.) leaves contain a variety of medicinal compounds, which have been used in healthcare and food industries. Flavonoids are the major bioactive ingredients in Ginkgo leaves, including flavonols, flavones, and anthocyanins (Meng et al., 2019). These molecules have been reported to have beneficial effects in preventing metabolic syndrome at different levels such as early stage Alzheimer's and cardiovascular diseases (Tian et al., 2017; Gruenwald et al., 2020). Flavonoids also act as growth regulators controlling single organ and whole plant development (Agati et al., 2012). Therefore, it is essential to understand the molecular mechanisms of flavonoids

accumulation during Ginkgo leaves development to ultimately improve the production and management of Ginkgo plantations.

Recently, considerable efforts have been dedicated to improving Ginkgo leaves flavonoids for commercial production. Studies showed that several agronomic measures could increase the flavonoids content, such as alternative partial root-zone irrigation (Wang et al., 2016), fertilization (Guo et al., 2016), and foliar fertilization (Wu et al., 2020). Treatments with salicylic acid, UV-B, and NaCl, all have shown a positive effect on increasing Ginkgo leaves flavonoids content (Ni et al., 2017, 2018; Zhao et al., 2020). More importantly, additional efforts have been directed at the molecular level to achieve the same objective. For example, transcriptome libraries have been constructed for various Ginkgo tissues (Ye et al., 2019) and leaves with different flavonoid contents (Wu et al., 2018), for improving the understanding of flavonoid biosynthesis. Another strategy for improving Ginkgo leaves metabolites yield is through genetic engineering; however, detailed information on gene expression profiling and transcriptional dynamics that regulate flavonoids accumulation is scarce.

Leaves undergo a series of developmental and physiological changes during their lifespans, involving complex, but highly regulated molecular processes to maximize fitness in a given ecological setting (Leopold, 1961; Fenner, 1998). It was found that leaves from the same Ginkgo tree could exhibit differences in flavonoids content at different developmental stages (young vs. mature leaves) (Guo et al., 2020). Additionally, Ginkgo leaves from plants growing at different elevations (different environments), and the same growing period also displayed substantial differences in their flavonoids accumulation (Zou et al., 2019). However, the current understanding of flavonoids accumulation regulation mechanism, which varies according to the development stage and geographical distribution, is limited. Transcriptomes can provide information regarding gene expression and regulation at specific developmental stages or under specific physiological conditions (Sato et al., 2011). Furthermore, integration of different-omics data, such as metabolome, will help elucidate the complex mechanism controlling flavonoid biosynthesis (Weckwerth, 2008).

Here, we conducted comprehensive temporospatial transcriptome and metabolome dynamics analyses of clonally propagated Ginkgo plants at four developmental stages (May to August) across three different environments (test-sites) to unravel leaves flavonoids biosynthesis spatial-temporal variation. The study-specific objectives are to: (1) quantify the transcriptional responses to spatial (environmental cues) and temporal (development stages) conditions; (2) explore the association between flavonoids accumulation and expression of flavonoid related structural genes; and (3) elucidate the regulatory network involved in gene expression associated with flavonoids biosynthesis. The broader aim of this work is intended to improve our understanding of the transcriptional dynamics that regulate flavonoids accumulation at the molecular level and provide insightful information for enhancing flavonoids content of Ginkgo leaves for the proper production and management of Ginkgo commercial plantations.

MATERIALS AND METHODS

Plant Materials and Sample Collection

Generally, the optimum age of flavonoids production in Ginkgo leaf-harvest plantations is trees under 5-year-old (Zou et al., 2019), thus older trees are considered suboptimum. Therefore, in the present study we utilized leaf samples collected from 2-year-old clonally propagated (grafted) Ginkgo trees. Trees are spatially replicated over three test sites (i.e., different environments) (**Table 1**). These sites are: (1) Yi Ning (YN), located in northwestern China (lat.: 43.41°N, long.: 81.11°E), characterized by a typical mid-temperate continental semi-arid climate; (2) Pi Zhou (PZ), located in central China (lat.: 34.21°N, long.: 117.58°E) characterized by a warm temperate monsoon climate; and (3) Qu Jing (QJ), located in southern China (lat.: 25.52°N, long.: 103.58°E), characterized by a subtropical plateau monsoon climate. In each site, the experiment is planted as a complete randomized block design with three blocks (replicates), each harboring 20 Ginkgo clones.

Samples were conducted between leaf expansion (May, after majority of leaves expansion) and leaf "commercial" ripening (August, before autumnal senescence). During this biological window, Ginkgo leaves are at their substance's peak activity and are easy to harvest and store (Ellnain-Wojtaszek et al., 2002). Leaves were collected on a clear day in the middle of each month (May to August) to represent four temporal leaf developmental stages. A single clone was randomly selected across the three blocks (i.e., 3 biological replications) and the collected leaf samples provided the material for the metabolomics and transcriptomics analyses. Each sampled tree was represented by three crown positions (top, middle, and bottom), each provided a single complete and healthy leaf. In total, 36 samples (4 development stages × 3 environments × 3 biological replicates) were used for the metabolomics and transcriptomics analyses. Collected leaves were immediately preserved in liquid nitrogen, freeze-dried, and kept at −80°C until further use. To measure the temporal variation in total flavonoid glycosides (TFG) content, a monthly time series sampling was conducted (at mid-month between May and August) on the PZ site and nine leaves were randomly collected from the 20 Ginkgo clones planted in each block. Additionally, to measure the spatial changes of TFG content across environments, nine leaves were randomly collected from the same 20 Ginkgo clones from the three blocks and the same sampling scheme was conducted across the three sites (sampling was conducted in mid-August). Leaves

TABLE 1 | Geographical distribution and climate factors [mean annual temperature (MAT), mean annual precipitation (MAP), and mean annual sunshine duration (MASD)] of the studied three test sites [Yi Ning (YN), Pi Zhou (PZ), and Qu Jing (QJ)].

Site	Latitude (°N)	Longitude (°E)	Altitude (m)	MAT (°C)	MAP (mm)	MASD (h)
YN	43.41	81.11	820	5.2	331	7.1
PZ	34.21	117.58	44	14.5	845	5.9
QJ	25.52	103.58	2,160	14.1	1,067	6.5

were oven-dried (70°C, 48 h), crushed, sieved through a 100-mesh sieve, and vacuum packed. All experiments were performed with three biological replicates.

Total Flavonoid Glycosides Measurement

Ginkgo leaves flavonoids were extracted following the Pharmacopoeia of the People's Republic of China (PPRC) procedures (Commision, 2010), and flavonoid glycosides content were determined by high-performance liquid chromatography (HPLC). In brief, approximately 0.5 g of oven-dried leaf powder per sample was immersed in petroleum ether and refluxed at 70°C for 2 h to remove impurities. Samples were then soaked in methanol and each sample's extract was evaporated on a rotary evaporator after refluxed at 70°C for 4 h. Subsequently, the pellet was washed with 25 mL of a 25% methanol-HCl (4:1, v/v) mixture and the eluent was collected and refluxed for 30 min. After cooling to room temperature, the eluent was brought to 50 mL with methanol, then used for determination by HPLC. HPLC (Waters 1525, United States) conditions were set as follows: the mobile phase was methanol and 0.4% H_3PO_4 solution (56:44, v/v) at 1.0 mL min^{-1}; the column temperature was 30°C; the detection was performed at 360 nm. Quercetin, kaempferol, and isorhamnetin were selected as standard substances following the supplier's specifications (Yuanye Biological Co., Shanghai, China).

Total flavonoid glycosides content = (quercetin + kaempferol + isorhamnetin) × 2.51 (Commision, 2010). Means and standard errors for each sample were calculated. Differences among samples were determined using one-way ANOVA and significant differences were detected (defined as $P < 0.05$) using the least significant difference (LSD) test.

Metabolomics Analysis

The supernatant extraction for each sample was performed as previously described (Guo et al., 2020). In summary, about 50 mg freeze-dried sample was put into an EP tube after grinding. After the addition of 1 mL of extract solvent (acetonitrile-methanol-water, 2:2:1, containing 0.1 mg L^{-1} lidocaine as an internal standard), the samples were swirled for 30 s, homogenized at 45 Hz for 4 min, and sonicated for 5 min in an ice water bath. The homogenate and sonicate circle was repeated three times, followed by incubation at −20°C for 1 h and centrifugation at 12,000 rpm and 4°C for 15 min. The resulting supernatants were transferred to LC-MS vials and stored at −80°C for later use. LC-MS/MS analyses were performed using an UHPLC system (1290, Agilent Technologies) with a UPLC HSS T3 column coupled to Q Exactive (Orbitrap MS, Thermo). The mobile phase A was 0.1% formic acid in water for positive, and 5 mmol/L ammonium acetate in water for negative, and the mobile phase B was acetonitrile. The elution gradient was set as follows: 0 min, 1% B; 1 min, 1% B; 8 min, 99% B; 10 min, 99% B; 10.1 min, 1% B; 12 min, 1% B [see **Supplementary Figure S1** for a UHPLC chromatogram of standards and samples from Yi Ning (YN) site at different sampling stages]. MS raw data files were converted to the mzML format using ProteoWizard, and processed by R package XCMS. OSI-SMMS (version 1.0, Dalian Chem Data Solution Information Technology Co. Ltd.) was used for peak annotation after data processing with an in-house MS/MS database. The metabolites were mapped to the Kyoto Encyclopedia of Genes and Genomics (KEGG) metabolic pathways to identify the substances in the related pathways of flavonoid biosynthesis (ko 00941- ko 00944).

Transcriptomics Analysis

Total RNA extraction, library preparation, and sequencing for each sample (36 libraries: four developmental stages in three different environments with three biological replicas) were performed as previously described (Guo et al., 2020). Total RNA was extracted from the freeze-dried samples using Trizol reagent kit (Invitrogen, Carlsbad, CA, United States) according to the manufacturer's protocol. RNA quality was assessed on an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, United States) and checked using RNase free agarose gel electrophoresis. Then the enriched mRNA was fragmented into short fragments using fragmentation buffer and reverse transcribed into cDNA with random primers. Second-strand cDNA were synthesized by DNA polymerase I, RNase H, dNTP and buffer. Then the cDNA fragments were purified with QiaQuick PCR extraction kit (Qiagen, Venlo, Netherlands), end repaired, poly (A) added, and ligated to Illumina sequencing adapters. The ligation products were size selected by agarose gel electrophoresis, PCR amplified, and sequenced using Illumina HiSeq2500. The Ginkgo Illumina raw sequencing data were submitted to the NCBI BioProject database under project number PRJNA657336.

An index of the reference genome was built, and paired-end clean reads were mapped to the Ginkgo's reference genome[1] using Hisat2. The mapped outputs were processed via StringTie software to obtain FPKM (fragment per kilobase of transcript per million mapped reads) for all the Ginkgo genes in each sample. Based on gene expression, principal component analysis (PCA) and hierarchical clustering analysis were performed with R packages, *gmodels* and *pheatmap*[2], which were also used to reveal the relationship among samples. The FPKM data were directly used to estimate the differential expression of genes (DEGs) between samples. FDR < 0.05 and $|\log2FC| > 1$ were used as thresholds to identify significant DEGs. The Short Time-series Expression Miner (STEM) software was used to obtain the temporal expression profile of DEGs. Subsequently, DEGs in enriched clustered profiles were used for KEGG pathway enrichment analysis (Q value ≤ 0.05) to assess metabolic pathways and related gene functions.

Weighted gene co-expression network analysis (WGCNA) was performed in the R environment. After filtering with the R package DCGL, a total of 23,182 genes (FPKM > 0) were reserved for subsequent analysis. The adjacency matrix between different genes was constructed with a threshold power of 10. A dynamic tree cut procedure (merge cut height = 0.70, min module size = 50) in R package WGCNA was used to identify similar modules in the hierarchical tree. The expression profile of module genes in each sample was displayed by module eigengene, which was defined as the first principal component of a given module.

[1] http://gigadb.org/dataset/100613

[2] https://cran.r-project.org/package=pheatmap

The Pearson correlations between the eigengenes of each module and the abundance of flavonoids were plotted by R package ggplot2. Subsequently, we identified the encoding transcription factor (TF) genes and the structural genes in the biosynthesis pathways of related flavonoids from the target modules. The gene regulatory networks were generated by Cytoscape software (Version 3.7.1).

The promoter region of TF genes was analyzed for presence of cis-acting regulatory elements by PlantCARE[3] and visualized by TBtools software. Additionally, to explore the regulatory effect of environmental factors on the expression of TF genes, Pearson's product-moment correlation analysis was conducted between TF genes expression and environmental factors during development (daily meteorological data for each area from May to August 2019[4]).

Quantitative Real-Time PCR (qRT-PCR) Analysis

Ten genes involved in flavonoid biosynthesis were randomly chosen for validation by qRT-PCR. According to the manufacturer's instructions, cDNA was obtained using MonScript RTIII All-in-One Mix with dsDNase kits (Monad, China) and qRT-PCR analysis was carried out using an Applied Biosystems^{TM} 7500 Real-Time PCR Systems (Monad, China). Primers used were designed in Primer Premier 5 (United States), and the primer sequences are provided in **Supplementary Table S2**. Glyceraldehyde 3- phosphate dehydrogenase (GAPDH, GenBank Accession No. L26924) gene was used as an internal standard. The relative transcript abundance was calculated using the $2^{-\Delta\Delta C_T}$ method (Livak and Schmittgen, 2001). August samples from YN, PZ, and QJ sites were used for qRT-PCR analysis. As designed, each sample included three biological replicates and three independent technical repetitions.

Statistical Analysis

All statistical analyses were conducted in R environment (R Core Team, 2019). Differences among TFGs content were determined using one-way analysis of variance (ANOVA) and significant differences were calculated using the least significant difference (LSD) test (defined as $P < 0.05$). The complex relationships between gene expression profiles were intuitively displayed by a PCA plot and a cluster heat map. Relationships between expression of structural genes and abundance of flavonoids were evaluated using the Pearson's product-moment correlation analysis ($P < 0.05$, significant correlation).

RESULTS

Changes in Total Flavonoid Glycosides Content

Temporarily, TFG content showed a declining trend with sampling time during the leaf development process. PZ site's TFG content time-course analysis showed the highest value occurred in May, followed by a drastic drop in June ($P < 0.05$) and a significant recovery in July (**Figure 1A**). Compared to May samples, the TFG content of June, July, and August samples decreased by 66.40, 15.98, and 21.50%, respectively (**Figure 1A**). Spatially, apparent differences in TFG content were observed across the three growing environments. Compared to August's PZ

[3]http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

[4]http://data.cma.cn/

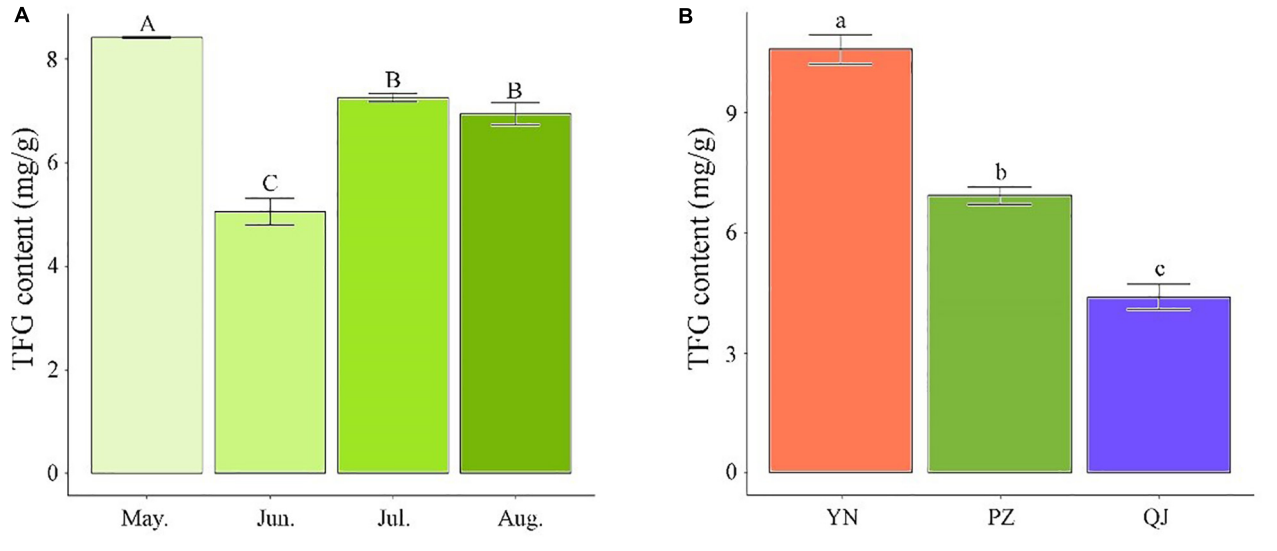

FIGURE 1 | Differences in total flavonoid glycosides (TFG) content among Ginkgo leaf samples. **(A)** Temporal differences in TFG content among developmental stages in PZ site. **(B)** Spatial differences in TFG content in August-samples across the three environments. Measurements were performed in triplicates. Error bars indicate standard deviations, and different capital and small letters represent a significant difference ($P < 0.05$) between developmental stages and environments, respectively.

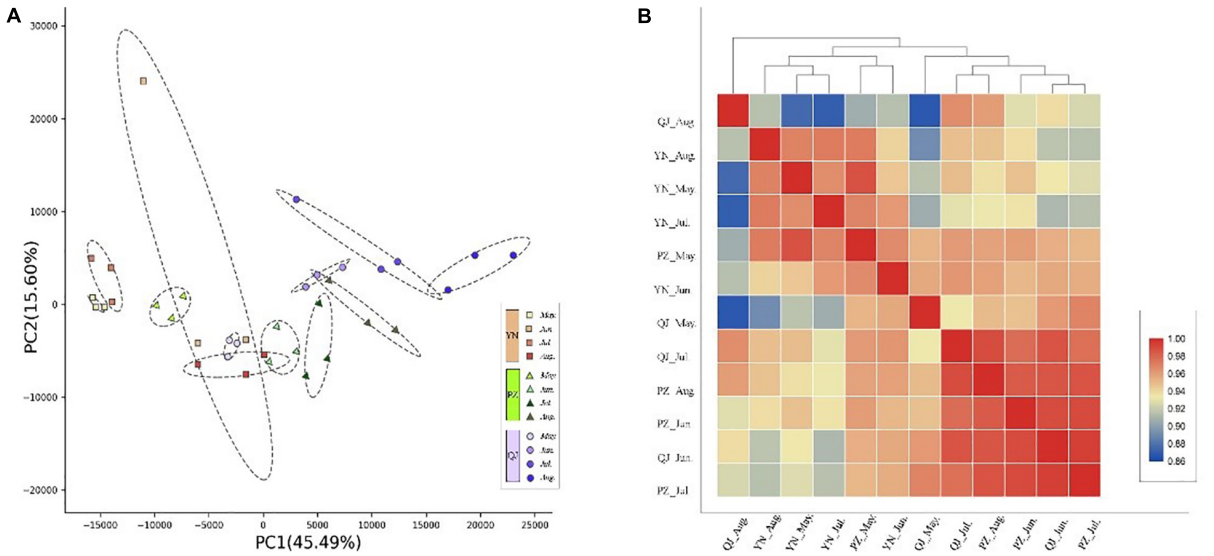

FIGURE 2 | Temporospatial differences in gene expression profile. **(A)** Principal component analysis (PCA) plot showing clustering of leaf transcriptomes at four developmental stages (temporal) under three different environments (spatial). Red squares, green triangles, and blue circles represent samples from YN, PZ, and QJ sites, respectively. The change in color from light to dark represents the four leaf development stages (May to August). **(B)** Pearson product-moment correlation coefficients and clusters of the RNA-seq data from leaf samples. The redder the rectangle, the stronger the correlation between the samples, whereas the bluer the rectangle, the weaker the correlation.

and QJ samples, the TFG content of YN was larger by 52.67 and 140.45%, respectively (**Figure 1B**).

Changes in Gene Expression Profile

Through transcriptional dynamics analysis, we identified approximately 2.0 billion clean reads from the 36 cDNA libraries that were mapped to the Ginkgo genome. The mapping rates of each library ranged from 91.14 to 95.39% (**Supplementary Table S1**). Spatially, PCA analysis results showed clear separation on the PC biplot, accounting for 61.1% of total gene expression variance in the data set (**Figure 2A**). Samples were spatially separated along the PC1 axis with YN, PZ, and QJ positioned on the left, middle, and right, respectively (**Figure 2A**). Temporally, within each site, the four developmental stage samples tended to follow the same left-to-right trend along PC1, while this trend did not persist for PC2 (**Figure 2A**). In the hierarchical clustering analysis, we did not detect any evidence of clustering among samples at either the different developmental stages (temporal) or at any given environment (spatial) (**Figure 2B**). Interestingly, PZ samples of the earlier stage (May) exhibited a closer correlation with the YN samples, whereas the PZ samples of a later stage (August) tended to correlate with the QJ samples, suggesting that major transcriptional program differences existed among development stages within each environment.

Differential Gene Expression During Leaf Development

At each leaf developmental stage, we identified different expression genes (DEGs) among samples from different environments (**Figure 3A**). We found more DEGs differences existed between QJ and YN samples (number of stage-specific genes varied from 644 to 3,318), while fewer DEGs differences between PZ and QJ samples (number of stage-specific genes varied from 74 to 1,097). The variable number of DEGs differences suggested that each stage of Ginkgo clones development had an independent strategy in response to their respective different environmental conditions.

To analyze the temporal expression pattern of DEGs, the 24,958 DEGs were further clustered by Short Time-series Expression Miner (STEM) software. There were 8 identifiable statistically significance ($P < 0.05$) temporal expression patterns, which were divided into 5 clusters containing a total of 1,908 DEGs (**Figure 3B**). The expression of DEGs contained in the 0 profile was gradually down-regulated during leaf development, while the temporal expression pattern of the 19 profile showed an opposite pattern. The KEGG pathway enrichment analysis of 1,908 DEGs revealed that 33 pathways were significantly enriched, including a large number of secondary metabolites, carbohydrate, and lipid metabolic pathways (**Figure 3C**). In particular, the phenylalanine and flavonoid biosynthesis pathways (ko 00940 and ko 00941) were enriched in several profiles. Therefore, these results suggested that the expression of some genes in the flavonoids-related biosynthesis pathways varied as a function of environmental factors (spatial) and developmental stages (temporal). The reliability of the RNA-seq results and the differentially expression analysis was further verified by qRT-PCR (**Supplementary Figure S2**).

Identification and Screening of Gene Co-expression Modules

Twelve modules were identified in a dendrogram comprising 105 – 3,908 genes, and each module harbored genes encoding

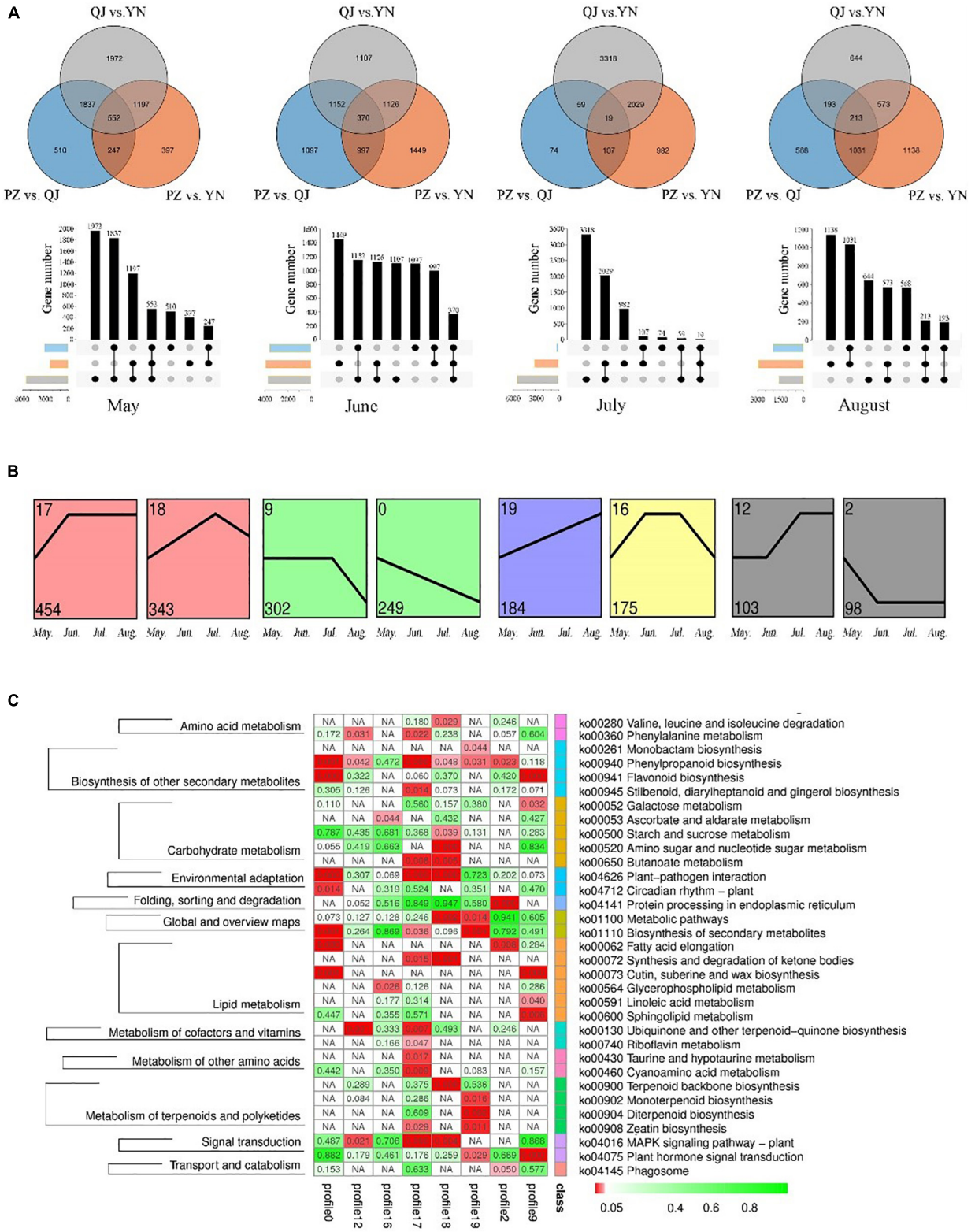

FIGURE 3 | Temporospatial expression pattern of differential expressed genes (DEGs). **(A)** Venn diagrams and column charts showing DEGs between samples from different growth environments (spatial) at four development stages (temporal). The gray circle/rectangle represents the difference between samples from YN and QJ; the blue one represents the difference between samples from PZ and QJ; the orange one represents the differences between the PZ and YN samples. **(B)** Profile blocks with a colored background are significant clusters of the $P \leq 0.05$, and the same color represents that the profiles are the same cluster. **(C)** An enriched KEGG map shows significant pathways among the genes of eight profiles. The red rectangle represents a significant enrichment pathway ($P < 0.05$).

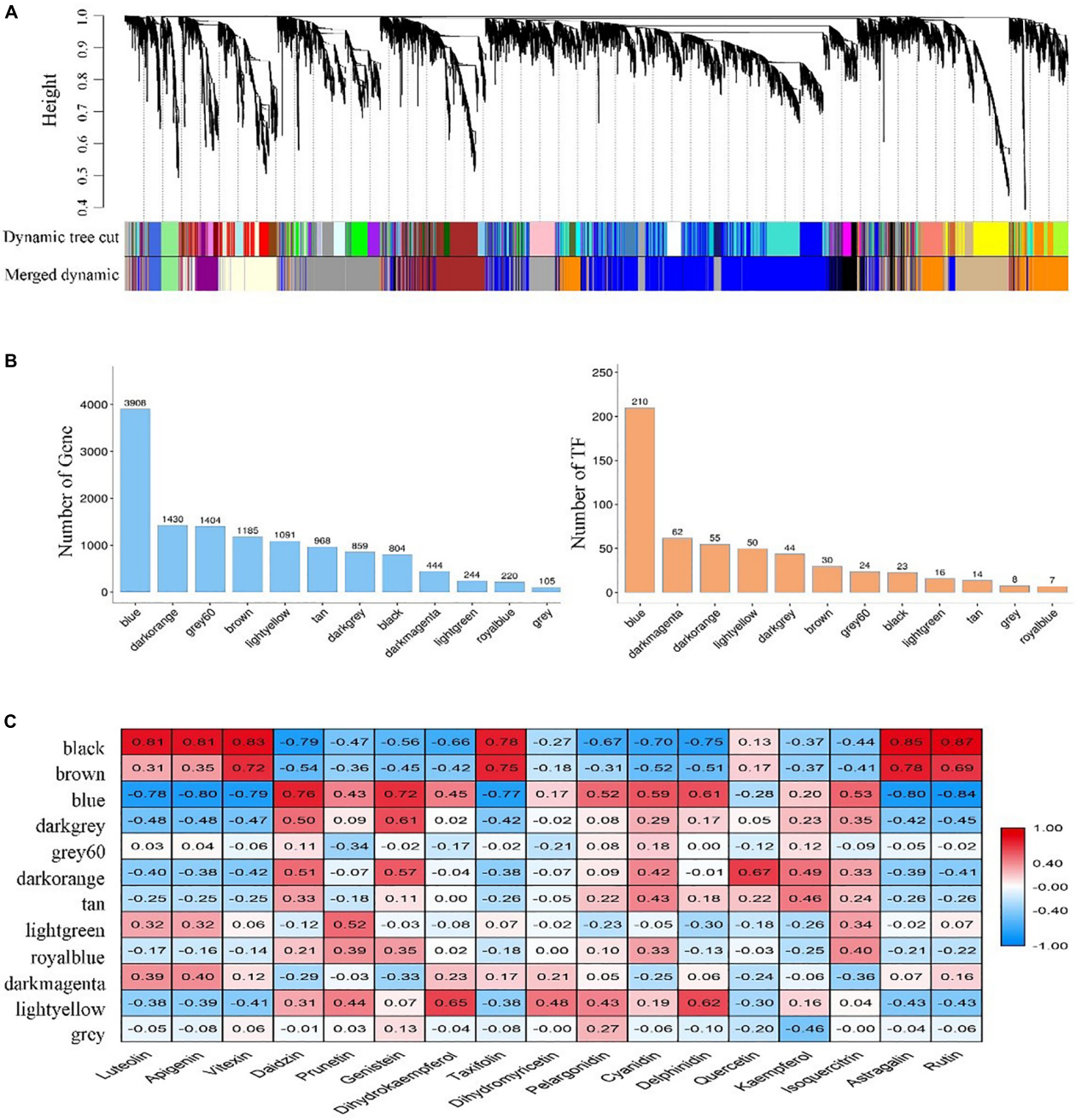

FIGURE 4 | Identification and screening of gene co-expression modules. **(A)** Dendrogram showing modules identified by the weighted gene co-expression network analysis (WGCNA) and clustering dendrogram of expressed genes; **(B)** The number of genes and transcription factors (TFs) contained in each module; **(C)** Correlation coefficient between the abundance of flavonoids and module eigengenes presented with a color scale with red and blue representing positive and negative correlations, respectively.

the number of transcription factors (TFs) varying from 7 to 210 (**Figures 4A,B**). In most modules, TF-encoding genes accounted for more than 5% of the total genes, indicating that the transcriptional activity was strictly regulated. Also, the eigengene of each module was associated with the abundance of 17 flavonoids revealed by Pearson product-moment correlation coefficient analysis. Remarkably, three modules (Black, Blue, and Brown) exhibited a strong correlation ($r > |0.7|$, $P < 0.05$) between gene expression and flavonoids accumulation (**Figure 4C**).

To better understand the function of genes in these three modules, we assigned the genes to KEGG terms. The top 20 enriched pathways in each module were revealed by bubble maps. The genes from Black and Blue modules were

significantly enriched in pathways related to translation, folding, sorting and degradation, signal translation, amino acid metabolism, and energy metabolism (**Supplementary Figures S3A,B**). In these pathways, some unigenes encoding glutathione S-transferase (GST), vacuolar sorting receptors (VSR), multi-antimicrobial extrusion protein (MATE) were found, which were thought to be involved in the transportation of flavonoids from cytosolic biosynthesis to their vacuolar accumulation (Petrussa et al., 2013). Notably, genes from the Brown module were significantly enriched in pathways related to secondary metabolites biosyntheses, such as flavonoid biosynthesis and phenylpropanoid biosynthesis (**Supplementary Figure S3C**). Interestingly, a large number of genes encoding key enzyme (flavonoids-related structural genes) had been identified in these pathways, including genes encoding phenylalanine ammonia-lyase (*PAL*), cinnamate 4-hydroxylase (*4CH*), chalcone synthase (*CHS*), chalcone isomerase (*CHI*), flavonol synthase (*FLS*), dihydroflavonol 4-reductase (*DFR*), anthocyanin synthase (*ANS*), anthocyanidin reductase (*ANR*), and UDP-glycosyltransferase (*UGT*).

Construction of Flavonoid-Related Gene Regulation Network

After screening the target modules, we constructed the biosynthesis pathways of six flavonoids (flavone, isoflavone, flavanonol, anthocyanins, flavonol, and flavonol glycoside) and identified the structural genes involved in these pathways from the Brown module (**Figure 5A**). The developmental stage specificity of the 12 flavonoids accumulation and the 15 structural genes expression was visualized (**Figure 5B**). We found that three flavones (luteolin, apigenin, and vitexin), one flavanonol (taxifolin), and two flavonol glycosides (astragalin and rutin) had the highest accumulation in May. In contrast, two isoflavones (daidzin and genistein), two anthocyanins (cyanidin and delphinidin), and one flavonol glycosides (isoquercitrin) had the lowest accumulation in leaves at the same developmental stage. Additionally, the accumulation of one flavonol (quercetin) was the highest in July (see **Supplementary Table S3** for quantitative values of the 12 identified flavonoids). The 15 structural genes identified in the Brown module had similar developmental expression patterns, with high expression in May and low expression in August. Correlation analysis of transcriptome and metabolome indicated that some structural genes were significantly correlated with specific flavonoids ($r > |0.6|$, $P < 0.05$). For example, quercetin content was positively correlated with the expression of a gene encoding *FLS* enzyme, while cyanidin content was negatively correlated with the expression of some genes encoding *ANS*, *DFR*, and *UGT* enzyme. Thus, these structural genes may play crucial roles in the accumulation of some flavonoids.

In Brown module, a total of 13 genes belonging to four transcription factor families were identified, including those encoding *MYB* (8 genes), *ERF* (3 genes), *bHLH* (1 gene), and *Trihelix* (1 gene), which may be involved in the regulation of flavonoids accumulation. To explore the regulatory effect of TFs on flavonoid biosynthesis, a co-expression regulation sub-network was established among TF genes and flavonoid-related structural genes according to the correlation analysis (**Figure 6A**). We observed that *GbMYB* (Gb_40628) had the highest connectivity and was closely associated with 10 structural genes. Additionally, we observed that a structural gene was regulated by multiple TFs simultaneously, such as *GbCHI* (Gb_21115) was positively correlated with five *GbMYB*s (Gb_11316, Gb_32143, Gb_33428, Gb_39081, and Gb_40628), three *GbERF*s (Gb_00128, Gb_26438, and Gb_37188), and one *GbTrihelix* (Gb_02053). Therefore, we suggested that these TFs participated in the regulation of gene expression in the flavonoid biosynthesis pathway.

Further, the correlation between genes encoding TFs and climate factors was analyzed (**Figure 6B**). We found that several *GbMYB*s (Gb_11316, Gb_26833, Gb_32143, Gb_33428, and Gb_40628), *GbERF*s (Gb_26438 and Gb_37188), *GbbHLH* (Gb_17233), and *GbTrihelix* (Gb_02053) were significantly and negatively ($r > |0.6|$, $P < 0.05$) correlated with mean minimum temperature (Tmin); three *GbMYB*s (Gb_11316, Gb_32143, and Gb_33428), one *GbERF* (Gb_00128), and one *GbTrihelix* (Gb_02053) had significant negative correlations with mean relative humidity (Hum); and two *GbMYB*s (Gb_11316 and Gb_32143) and one *GbERF* (Gb_00128) had significant positive correlations with sunshine duration (SD). These results suggested that the sunny environment was favorable to the expression of genes encoding TFs, while conversely the cold and humid environment was unfavorable to their expression. The promoter analysis also supported our hypothesis, as multiple light responsiveness (G-Box, Box 4, AE-box, I-box, L-box, Gap-box, Box II, and G-box) and low-temperature responsiveness (LTR) elements were found in the promoter regions of TF genes (**Figure 6C**).

DISCUSSION

Flavonoids represent one of the main classes of secondary metabolites that play an important role in plant defense against environmental stresses (e.g., temperature, precipitation, and light) (Akula and Ravishankar, 2011). Additionally, flavonoids extracted are also beneficial compounds for human health as cardioprotective, antihypertensive, and antioxidants (Fuchs et al., 2016). While studying the metabolic process of flavonoids in Ginkgo has been the subject of intense investigations (Wu et al., 2018; Meng et al., 2019; Guo et al., 2020), limited information is available at the genomic level. In the present study, we investigated the temporospatial (four leaves developmental stages and three contrasting test sites) transcriptome and metabolome dynamics biological processes to increase our understanding of Ginkgo's flavonoids regulatory networks and to provide additional information of the molecular mechanisms of flavonoids accumulation during its leaves development. We used clonally propagated plant material, so the observed differences are attributable to either time (four leaves developmental stages) or space (three contrasting test sites).

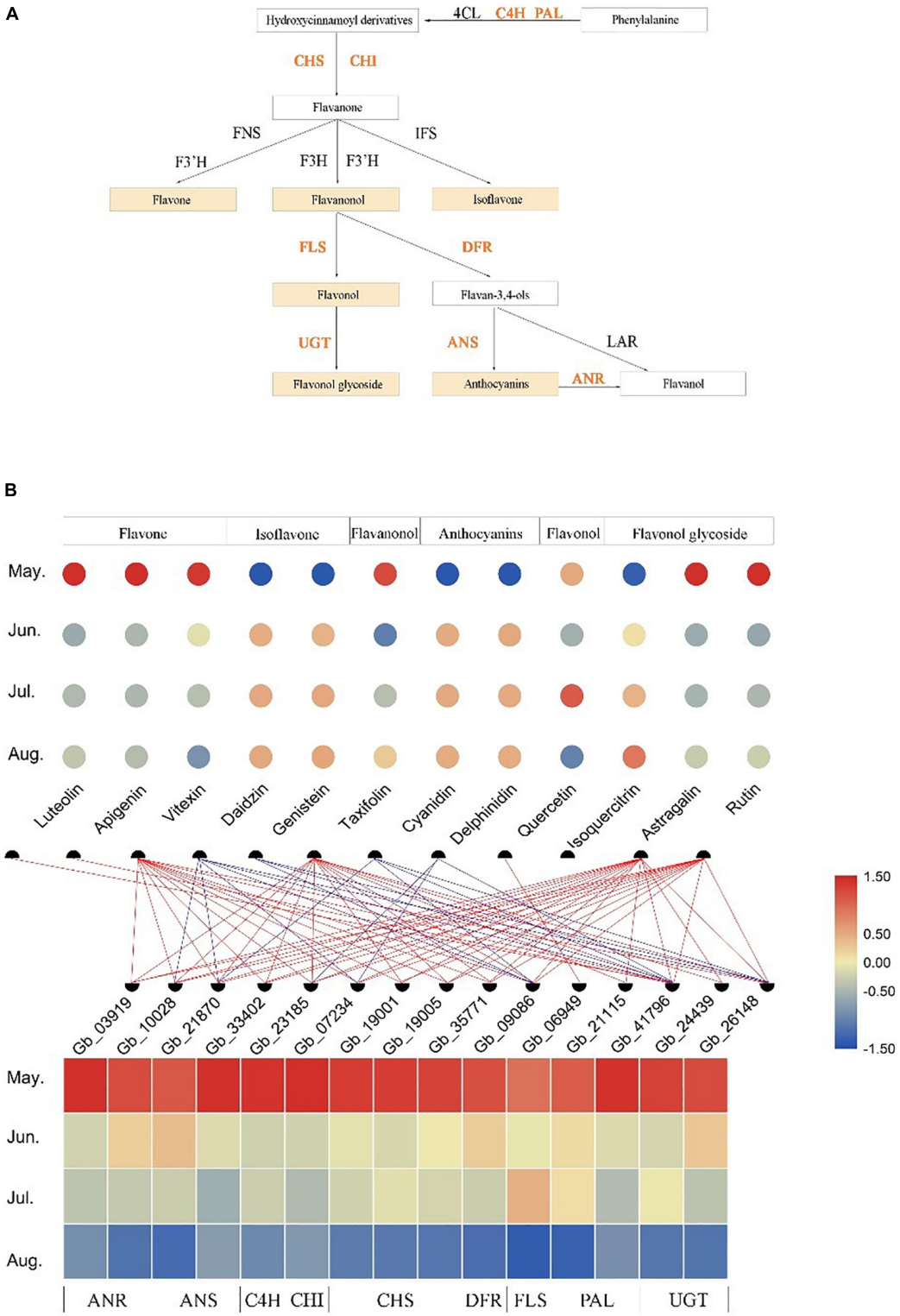

FIGURE 5 | Temporospatial expression patterns of flavonoid-related structural genes in different leaf developmental stages (temporal) and association analysis with the accumulation of several kinds of flavonoids. **(A)** The biosynthesis pathways of several kinds of flavonoids. **(B)** The accumulation patterns of 12 flavonoids and the expression patterns of 15 flavonoid-related structural genes in different growth stages and the association analysis between them. The change in color of circle/rectangle from red to blue represents a gradual decrease in the abundance of flavonoids/expression of structural genes. The red line represents the positive correlation between the expression of structural genes and the abundance of flavonoids, and the blue line represents the negative correlation between them ($r > |0.6|$, $P < 0.05$).

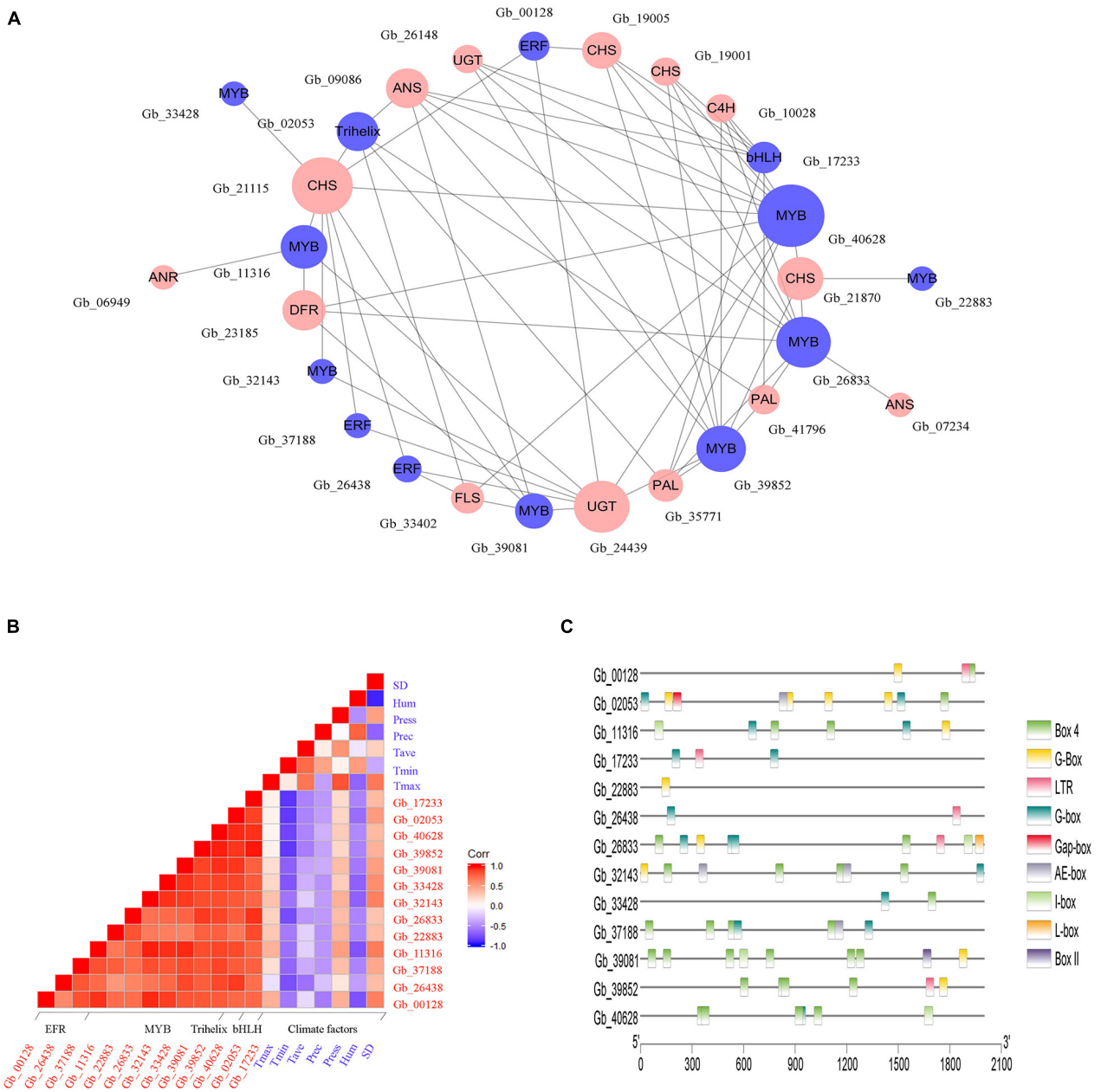

FIGURE 6 | Environmental factors may trigger the activation of TFs to control the expression of flavonoid-related structural genes. **(A)** The co-expression network between TFs (blue circles) and flavonoid-related structural genes (red circles). The circle size is positively correlated with the connectivity of genes in the regulatory network. **(B)** The correlation analysis between environmental factors and the expression of TF genes. The red and blue square indicate positive and negative correlations between specific rows and columns, respectively. Climate factors include sunshine duration (SD), relative humidity (Hum), air pressure (Press), average temperature (Tave), minimum temperature (Tmin), and maximum temperature (Tmax). **(C)** *Cis*-regulatory elements analysis of TF genes.

Temporospatial Gene Expression Profiles Variation

We used the RNA-seq approach to detect changes in the studied samples gene expression profiles and the principal component analysis, PCA-plot, clearly showed a spatial separation among samples growing at different environments, suggesting that gene expression at the transcriptome level is strongly influenced/modified by environmental conditions (**Figure 2A**). Observations confirming previously held view that transcriptional regulatory cascades may be key components of differential resilience shown by plants to changing environments (D'esposito et al., 2017). Furthermore, the four developmental stages also were separated (PCA-plot: **Figure 2A**), suggesting that the observed temporal differences provided a reasonable description of transcription activities during leaf development stages (Garg et al., 2017). Interestingly and more strikingly, the

transcriptome from different spatial (environmental) samples showed similar stage-specific expression patterns that were gradually separated in the same direction (PCA-1) during leaf development. These findings reflected the dynamic nature and flexibility of gene expression in response to internal (genetic) and external (environmental) cues at the transcription level during leaf development (Bar and Ori, 2014).

Flavonoids Metabolism Is Temporospatially Influenced

The observed differentially expressed genes (DEGs) among spatially different samples (environments) were identified and exhibited a collinear pattern with the environmental differences between sites. As the environmental differences between sites increase, this was accompanied by a concomitant increase in the differences of DFGs of their respective samples. For example, the difference between DEGs from YN and QJ sites was in line with the observed differences between these two sites environments (**Figure 3A**). Ginkgo may have developed a genetic control system as a survival strategy in response to different environments (Cho et al., 2018; Wang et al., 2020). Further, the eight temporal expression patterns (DEG sets) identified by STEM analysis (**Figure 3B**), contained genes significantly affected by environmental (spatial) and developmental (temporal) processes. The KEGG pathway enrichment analysis (**Figure 3C**) indicated that the expression of genes from Profile 0 presented a down-regulated trend with leaf development, and these genes were significantly enriched in both phenylpropanoid and flavonoid biosynthesis pathways (ko 00940 and ko 00941). These results indicated that a set of genes related to flavonoid biosyntheses, such as structural genes and TF genes, performed the stage-specific (temporally sequenced) function under external environmental stimuli. It has been reported that there is a rhythm, a time-distribution character, to the biosynthesis and metabolism of flavonoids in Ginkgo leaves (Cheng et al., 2012; Sati et al., 2013). Our results indicated that flavonoid-related metabolism was more active at the transcriptional level in the early stage of leaf development, consistent with previous studies (Yan et al., 2019; Zhu et al., 2020). These findings were also confirmed by the result of HPLC analysis; where we found that the content of TFG peaked in the early stage (**Figure 1A**). We also found that samples from YN and QJ sites had the greatest difference in TFG content (**Figure 1B**).

A Regulated Transcriptional Network for Flavonoids Biosynthesis

Previous studies have focused on the identification of flavonoid-related structural genes in Ginkgo leaves, such as genes encoding *PAL*, *C4H*, *4CL*, *CHS*, *CHI*, and *F3H* in early flavonoid biosynthesis pathway, and genes encoding *DFR*, *ANS*, and *ANR* in downstream steps of the pathway (Li et al., 2018; Wu et al., 2018). In the present study, an intensive association network was observed between the expression of 15 selected structural genes and abundance of flavonoids (**Figure 5B**), suggesting that these structural genes (*GbANR*, *GbANS*, *GbC4H*, *GbCHI*, *GbCHS*, *GbDFR*, *GbFLS*, *GbPAL*, and *GbUGT*) may play crucial roles in the accumulation of specific flavonoids. More specifically, we identified one gene (Gb_41796) encoding *PAL*, which is an upstream key enzyme and rate limiting of the flavonoids biosynthesis pathway (Wang et al., 2014), whose expression was positively correlated with the abundance of three flavones (luteolin, apigenin, and vitexin), but negatively correlated with the two isoflavones (daidzin and genistein). Additionally, we found two genes encoding *UGT* (Gb_24439 and Gb_26148) whose expression was significantly and positively correlated with the abundance of flavonoid glycosides (astragalin), consistent with and supporting previous studies (Cui et al., 2016; Zhu et al., 2020). The accumulation of anthocyanins has been reported to be positively correlated with the expression of *GbDFR*s (Ni et al., 2020), while we found the gene encoding *DFR* (Gb_09086) was negatively associated with cyanidin accumulation.

Flavonoids biosynthesis is mainly regulated by transcription factors at the transcription level (Xu et al., 2014; Cao et al., 2020). We constructed a co-expression regulation network among TF genes and flavonoid-related structural genes to explore their regulatory relationship (**Figure 6A**). We discovered eight *GbMYB*s that positively regulated the expression of multiple structural genes in the flavonoid biosynthesis pathway. *MYB* TFs represent one of the largest families of a transcription factor in plants, involving in the regulation of different biological processes (Dubos et al., 2010). The large number of *GbMYB*s in the Ginkgo genome indicated that each of them may involve unique functions. Meng et al. (2019) found that the *GbMYB5* was involved in the positive regulation of flavonoid biosynthesis, while Xu et al. (2014) suggested that the *GbMYBF2* was responsible for repressing flavonoid biosynthesis. *MYB* and *bHLH* can act individually or in concert with other TFs to regulate a series of structural genes involved in flavonoid metabolism (Terrier et al., 2009; Carletti et al., 2013). Similarly, this co-expression network showed positive correlations between *GbERF*, *GbbHLH*, and *GbTrihelix* and certain structural genes associated with flavonoids.

TFs are considered as the major regulators of gene expression in response to environmental changes. *MYB*, *ERF*, and *bHLH* have been shown to play important roles in regulating environmental stress responses (Nakashima et al., 2009; Agarwal and Jha, 2010). In this study, we found that some *GbTF*s expression was negatively correlated with mean minimum temperature but positively correlated with sunshine duration (**Figure 6B**). Meanwhile, we also identified abundant light responsiveness elements and LTRs elements in *GbTF*s promoter regions (**Figure 6C**). It has been confirmed in previous studies, anthocyanins accumulation in *Pinus contorta* seedlings grown under short sunlight was significantly lower than those growing in the long sunlight area; long-term light irradiation (16 h) on leaves of *Ipomoea batatas* generated a dramatic increase in flavonoids content (Camm et al., 1993; Carvalho et al., 2010). As the amount of sunlight increases, there is a concomitant rise in temperature, and the composition of flavonoids in *Ribes nigrum* has been found to be positively correlated with temperature (Zheng et al., 2012). Our findings further support that proper control of gene expression by TFs was essential for the flavonoids biosynthesis, which played an important role in response to environmental changes (López-Maury et al., 2008).

CONCLUSION

Our investigation of the temporospatial transcriptome and metabolome dynamics biological processes provided new insights into the biosynthesis of flavonoids in Ginkgo leaves. We indicated that flavonoids content varied greatly at different developmental stages (temporally) and in different growth environments (spatially). Therefore, the careful selection of planting region(s) and optimization of leaf harvesting time are expected to substantially improve the benefits of Ginkgo utilization as a non-timber forest product. We constructed a co-expression regulation network and identified 13 TF genes having crucial roles in controlling the transcriptomic regulation of flavonoids by activating the expression of multiple structural genes. These results provide candidate genes for future enhancement of flavonoids production by genetic strategies in Ginkgo. Furthermore, the large amount of data resources generated will serve as the foundation for a system biology approach to study the dynamics of leaf development and flavonoids accumulation in other plants.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA657336, SUB7912734.

AUTHOR CONTRIBUTIONS

GW and YE-K conceived the study. YG collected the field samples, analyzed the data, and drafted the manuscript. GW, YE-K, TW, and F-FF modified the manuscript. All the authors have approved the manuscript.

FUNDING

This research was funded by the National Natural Science Foundation of China (grant no. 31971689), the National Key Research and Development Program of China (2017YFD0600700), and the Doctorate Fellowship Foundation of Nanjing Forestry University.

ACKNOWLEDGMENTS

We are grateful to B. Ratcliffe for critical review and suggestions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.589326/full#supplementary-material

Supplementary Figure 1 | UHPLC chromatogram of standards and samples at different sampling stages. **(A)** Standards; **(B–E)** samples collected from the Yi Ning (YN) site in May, June, July, and August, respectively.

Supplementary Figure 2 | The qRT-PCR analysis results of key genes involved in the flavonoid biosynthesis. The FPKM **(A)** and relative expression levels **(B)** were presented in bar plot with the error bar by calculating the mean and stand deviation (SD) of three independent replicates.

Supplementary Figure 3 | Bubble maps show the top 20 significantly enriched KEGG pathways among the genes in three modules highly associated with flavonoid accumulation.

Supplementary Table 1 | Summary of the sequencing quality of 36 RNA libraries of Ginkgo leaves.

Supplementary Table 2 | Primers used for the qRT-PCR assay.

Supplementary Table 3 | Summary of the quantitative values (peak area) of the identified 12 flavonoids.

REFERENCES

Agarwal, P., and Jha, B. (2010). Transcription factors in plants and ABA dependent and independent abiotic stress signalling. *Biol. Plant.* 54, 201–212. doi: 10.1007/s10535-010-0038-7

Agati, G., Azzarello, E., Pollastri, S., and Tattini, M. (2012). Flavonoids as antioxidants in plants: location and functional significance. *Plant Sci.* 196, 67–76. doi: 10.1016/j.plantsci.2012.07.014

Akula, R., and Ravishankar, G. A. (2011). Influence of abiotic stress signals on secondary metabolites in plants. *Plant Signal. Behav.* 6, 1720–1731. doi: 10.4161/psb.6.11.17613

Bar, M., and Ori, N. (2014). Leaf development and morphogenesis. *Development* 141, 4219–4230. doi: 10.1242/dev.106195

Camm, E., Mccallum, J., Leaf, E., and Koupai-Abyazani, M. (1993). Cold-induced purpling of *Pinus contorta* seedlings depends on previous daylength treatment. *Plant Cell Environ.* 16, 761–764. doi: 10.1111/j.1365-3040.1993.tb00497.x

Cao, Y., Li, K., Li, Y., Zhao, X., and Wang, L. (2020). MYB transcription factors as regulators of secondary metabolism in plants. *Biology* 9:61. doi: 10.3390/biology9030061

Carletti, G., Lucini, L., Busconi, M., Marocco, A., and Bernardi, J. (2013). Insight into the role of anthocyanin biosynthesis-related genes in *Medicago truncatula* mutants impaired in pigmentation in leaves. *Plant Physiol. Biochem.* 70, 123–132. doi: 10.1016/j.plaphy.2013.05.030

Carvalho, I. S., Cavaco, T., Carvalho, L. M., and Duque, P. (2010). Effect of photoperiod on flavonoid pathway activity in sweet potato (*Ipomoea batatas* (L.) Lam.) leaves. *Food Chem.* 118, 384–390. doi: 10.1016/j.foodchem.2009.05.005

Cheng, S., Feng, X., Linling, L., Cheng, H., and Zhang, W. (2012). Seasonal pattern of flavonoid content and related enzyme activities in leaves of *Ginkgo biloba* L. *Notulae Botanicae Horti Agrobotanici Cluj-Napoca* 40, 98–106. doi: 10.15835/nbha4017262

Cho, S. M., Lee, H., Jo, H., Lee, H., Kang, Y., Park, H., et al. (2018). Comparative transcriptome analysis of field-and chamber-grown samples of *Colobanthus quitensis* (Kunth) Bartl, an Antarctic flowering plant. *Sci. Rep.* 8, 1–14. doi: 10.1038/s41598-018-29335-4

Commision, C. P. (2010). *Pharmacopoeia of the People's Republic of China*. Beijing: Chinese Medical Science and Technology Press.

Cui, L., Yao, S., Dai, X., Yin, Q., Liu, Y., Jiang, X., et al. (2016). Identification of UDP-glycosyltransferases involved in the biosynthesis of astringent taste compounds in tea (*Camellia sinensis*). *J. Exp. Bot.* 67, 2285–2297. doi: 10.1093/jxb/erw053

D'esposito, D., Ferriello, F., Dal Molin, A., Diretto, G., Sacco, A., Minio, A., et al. (2017). Unraveling the complexity of transcriptomic, metabolomic and quality

environmental response of tomato fruit. *BMC Plant Biology* 17:66. doi: 10.1186/s12870-017-1008-4

Dubos, C., Stracke, R., Grotewold, E., Weisshaar, B., Martin, C., and Lepiniec, L. (2010). MYB transcription factors in *Arabidopsis*. *Trends Plant Sci.* 15, 573–581. doi: 10.1016/j.tplants.2010.06.005

Ellnain-Wojtaszek, M., Kruczyński, Z., and Kasprzak, J. (2002). Variations in the free radical scavenging activity of *Ginkgo biloba* L. leaves in the period of complete development of green leaves to fall of yellow ones. *Food Chem.* 79, 79–84. doi: 10.1016/S0308-8146(02)00181-4

Fenner, M. (1998). The phenology of growth and reproduction in plants. *Perspect. Plant Ecol. Evol. Syst.* 1, 78–91. doi: 10.1078/1433-8319-00053

Fuchs, D., Nyakayiru, J., Draijer, R., Mulder, T. P., Hopman, M. T., Eijsvogels, T. M., et al. (2016). Impact of flavonoid-rich black tea and beetroot juice on postprandial peripheral vascular resistance and glucose homeostasis in obese, insulin-resistant men: a randomized controlled trial. *Nutr. Metab.* 13:34. doi: 10.1186/s12986-016-0094-x

Garg, R., Singh, V. K., Rajkumar, M. S., Kumar, V., and Jain, M. (2017). Global transcriptome and coexpression network analyses reveal cultivar-specific molecular signatures associated with seed development and seed size/weight determination in chickpea. *Plant J.* 91, 1088–1107. doi: 10.1111/tpj.13621

Gruenwald, J., Eckert, A., and Kressig, R. W. (2020). The effects of standardized *Ginkgo biloba* extracts (GBE) on subjective cognitive decline (SCD) in middle-aged adults: a review. *Adv. Aging Res.* 9, 45–65. doi: 10.4236/aar.2020.93005

Guo, J., Wu, Y., Wang, B., Lu, Y., Cao, F., and Wang, G. (2016). The effects of fertilization on the growth and physiological characteristics of *Ginkgo biloba* L. *Forests* 7:293. doi: 10.3390/f7120293

Guo, Y., Gao, C., Wang, M., Fu, F., El-Kassaby, Y. A., Wang, T., et al. (2020). Metabolome and transcriptome analyses reveal flavonoids biosynthesis differences in *Ginkgo biloba* associated with environmental conditions. *Ind. Crops Prod.* 158:112963. doi: 10.1016/j.indcrop.2020.112963

Leopold, A. (1961). Senescence in plant development. *Science* 134, 1727–1732.

Li, W., Yang, S., Lu, Z., He, Z., Ye, Y., Zhao, B., et al. (2018). Cytological, physiological, and transcriptomic analyses of golden leaf coloration in *Ginkgo biloba* L. *Hortic. Res.* 5, 1–14. doi: 10.1038/s41438-018-0015-4

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

López-Maury, L., Marguerat, S., and Bähler, J. (2008). Tuning gene expression to changing environments: from rapid responses to evolutionary adaptation. *Nat. Rev. Genet.* 9, 583–593. doi: 10.1038/nrg2398

Meng, J., Wang, B., He, G., Wang, Y., Tang, X., Wang, S., et al. (2019). Metabolomics integrated with transcriptomics reveals redirection of the phenylpropanoids Metabolic flux in *Ginkgo biloba*. *J. Agric. Food Chem.* 67, 3284–3291. doi: 10.1021/acs.jafc.8b06355

Nakashima, K., Ito, Y., and Yamaguchi-Shinozaki, K. (2009). Transcriptional regulatory networks in response to abiotic stresses in *Arabidopsis* and grasses. *Plant Physiol.* 149, 88–95. doi: 10.1104/pp.108.129791

Ni, J., Dong, L., Jiang, Z., Yang, X., Sun, Z., Li, J., et al. (2018). Salicylic acid-induced flavonoid accumulation in *Ginkgo biloba* leaves is dependent on red and far-red light. *Ind. Crops Product.* 118, 102–110. doi: 10.1016/j.indcrop.2018.03.044

Ni, J., Hao, J., Jiang, Z., Zhan, X., Dong, L., Yang, X., et al. (2017). NaCl induces flavonoid biosynthesis through a putative novel pathway in post-harvest Ginkgo leaves. *Front. Plant Sci.* 8:920. doi: 10.3389/fpls.2017.00920

Ni, J., Ruan, R., Wang, L., Jiang, Z., Gu, X., Chen, L., et al. (2020). Functional and correlation analyses of dihydroflavonol-4-reductase genes indicate their roles in regulating anthocyanin changes in *Ginkgo biloba*. *Ind. Crops Product.* 152:112546. doi: 10.1016/j.indcrop.2020.112546

Petrussa, E., Braidot, E., Zancani, M., Peresson, C., Bertolini, A., Patui, S., et al. (2013). Plant flavonoids-biosynthesis, transport and involvement in stress responses. *Int. J Mol. Sci.* 14, 14950–14973. doi: 10.3390/ijms140714950

R Core Team (2019). *R: A Language and Environment for Statistical Computing (version 3.5. 3)[Software]*. Vienna: R Foundation for Statistical Computing.

Sati, P., Pandey, A., Rawat, S., and Rani, A. (2013). Phytochemicals and antioxidants in leaf extracts of *Ginkgo biloba* with reference to location, seasonal variation and solvent system. *J. Pharm. Res.* 7, 804–809. doi: 10.1016/j.jopr.2013.09.001

Sato, Y., Antonio, B., Namiki, N., Motoyama, R., Sugimoto, K., Takehisa, H., et al. (2011). Field transcriptome revealed critical developmental and physiological transitions involved in the expression of growth potential in japonicarice. *BMC Plant Biol.* 11:10. doi: 10.1186/1471-2229-11-10

Terrier, N., Torregrosa, L., Ageorges, A., Vialet, S., Verries, C., Cheynier, V., et al. (2009). Ectopic expression of VvMybPA2 promotes proanthocyanidin biosynthesis in grapevine and suggests additional targets in the pathway. *Plant Physiol.* 149, 1028–1041. doi: 10.1104/pp.108.131862

Tian, J., Liu, Y., and Chen, K. (2017). *Ginkgo biloba* extract in vascular protection: molecular mechanisms and clinical applications. *Curr. Vasc. Pharmacol.* 15, 532–548. doi: 10.2174/1570161115666170713095545

Wang, G., Cao, F., Li, C., Guo, X., Wang, J., and Lu, Z. (2014). Temperature has more effects than soil moisture on biosynthesis of flavonoids in Ginkgo (*Ginkgo biloba* L.) leaves. *New For.* 45, 797–812. doi: 10.1007/s11056-014-9437-5

Wang, L., Cui, J., Jin, B., Zhao, J., Xu, H., Lu, Z., et al. (2020). Multifeature analyses of vascular cambial cells reveal longevity mechanisms in old *Ginkgo biloba* trees. *Proc. Natl. Acad. Sci. U.S.A.* 117, 2201–2210. doi: 10.1073/pnas.1916548117

Wang, L., Shi, H., Wu, J., and Cao, F. (2016). Alternative partial root-zone irrigation enhances leaf flavonoid accumulation and water use efficiency of *Ginkgo biloba*. *New For.* 47, 377–391. doi: 10.1007/s11056-015-9521-5

Weckwerth, W. (2008). Integration of metabolomics and proteomics in molecular plant physiology–coping with the complexity by data-dimensionality reduction. *Physiol. Plant.* 132, 176–189. doi: 10.1111/j.1399-3054.2007.01011.x

Wu, D., Feng, J., Lai, M., Ouyang, J., Liao, D., Yu, W., et al. (2020). Combined application of bud and leaf growth fertilizer improves leaf flavonoids yield of *Ginkgo biloba*. *Ind. Crops Product.* 150:112379. doi: 10.1016/j.indcrop.2020.112379

Wu, Y., Guo, J., Zhou, Q., Xin, Y., Wang, G., and Xu, L.-A. (2018). De novo transcriptome analysis revealed genes involved in flavonoid biosynthesis, transport and regulation in *Ginkgo biloba*. *Ind. Crops Product.* 124, 226–235. doi: 10.1016/j.indcrop.2018.07.060

Xu, F., Ning, Y., Zhang, W., Liao, Y., Li, L., Cheng, H., et al. (2014). An R2R3-MYB transcription factor as a negative regulator of the flavonoid biosynthesis pathway in *Ginkgo biloba*. *Funct. Integr. Genom.* 14, 177–189. doi: 10.1007/s10142-013-0352-1

Yan, J., Yu, L., He, L., Zhu, L., Xu, S., Wan, Y., et al. (2019). Comparative transcriptome analysis of celery leaf Blades identified an R2R3-MYB transcription factor that regulates apigenin metabolism. *J. Agric. Food Chem.* 67, 5265–5277. doi: 10.1021/acs.jafc.9b01052

Ye, J., Cheng, S., Zhou, X., Chen, Z., Kim, S. U., Tan, J., et al. (2019). A global survey of full-length transcriptome of *Ginkgo biloba* reveals transcript variants involved in flavonoid biosynthesis. *Ind. Crops Product.* 139:111547. doi: 10.1016/j.indcrop.2019.111547

Zhao, B., Wang, L., Pang, S., Jia, Z., Wang, L., Li, W., et al. (2020). UV-B promotes flavonoid synthesis in *Ginkgo biloba* leaves. *Ind. Crops Product.* 151:112483. doi: 10.1016/j.indcrop.2020.112483

Zheng, J., Yang, B., Ruusunen, V., Laaksonen, O., Tahvonen, R., Hellsten, J., et al. (2012). Compositional differences of phenolic compounds between black currant (*Ribes nigrum* L.) cultivars and their response to latitude and weather conditions. *J. Agric. Food Chem.* 60, 6581–6593. doi: 10.1021/jf3012739

Zhu, J., Xu, Q., Zhao, S., Xia, X., Yan, X., An, Y., et al. (2020). Comprehensive co-expression analysis provides novel insights into temporal variation of flavonoids in fresh leaves of the tea plant (*Camellia sinensis*). *Plant Sci.* 290:110306. doi: 10.1016/j.plantsci.2019.110306

Zou, K., Liu, X., Zhang, D., Yang, Q., Fu, S., Meng, D., et al. (2019). Flavonoid biosynthesis is likely more susceptible to elevation and tree age than other branch pathways involved in phenylpropanoid biosynthesis in *Ginkgo* leaves. *Front. Plant Sci.* 10:983. doi: 10.3389/fpls.2019.00983

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

REVIEW
published: 10 December 2020
doi: 10.3389/fgene.2020.610798

State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing

Michal Krassowski[1†], Vivek Das[2†], Sangram K. Sahu[3†] and Biswapriya B. Misra[4†]*

[1] Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, United Kingdom, [2] Novo Nordisk Research Center Seattle, Inc, Seattle, WA, United States, [3] Independent Researcher, Bengaluru, India, [4] Independent Researcher, Namburu, India

OPEN ACCESS

Edited by:
Fatemeh Maghuly,
University of Natural Resources and Life Sciences, Vienna, Austria

Reviewed by:
Heinz Himmelbauer,
University of Natural Resources and Life Sciences, Vienna, Austria
Subina Mehta,
University of Minnesota Twin Cities, United States
Wan M. Aizat,
National University of Malaysia, Malaysia

***Correspondence:**
Biswapriya B. Misra
bbmisraccb@gmail.com

†ORCID:
Michal Krassowski
orcid.org/0000-0002-9638-7785
Vivek Das
orcid.org/0000-0003-0614-0373
Sangram K. Sahu
orcid.org/0000-0001-5010-9539
Biswapriya B. Misra
orcid.org/0000-0003-2589-6539

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

Received: 27 September 2020
Accepted: 20 November 2020
Published: 10 December 2020

Citation:
Krassowski M, Das V, Sahu SK and Misra BB (2020) State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Front. Genet. 11:610798. doi: 10.3389/fgene.2020.610798

Multi-omics, variously called integrated omics, pan-omics, and trans-omics, aims to combine two or more omics data sets to aid in data analysis, visualization and interpretation to determine the mechanism of a biological process. Multi-omics efforts have taken center stage in biomedical research leading to the development of new insights into biological events and processes. However, the mushrooming of a myriad of tools, datasets, and approaches tends to inundate the literature and overwhelm researchers new to the field. The aims of this review are to provide an overview of the current state of the field, inform on available reliable resources, discuss the application of statistics and machine/deep learning in multi-omics analyses, discuss findable, accessible, interoperable, reusable (FAIR) research, and point to best practices in benchmarking. Thus, we provide guidance to interested users of the domain by addressing challenges of the underlying biology, giving an overview of the available toolset, addressing common pitfalls, and acknowledging current methods' limitations. We conclude with practical advice and recommendations on software engineering and reproducibility practices to share a comprehensive awareness with new researchers in multi-omics for end-to-end workflow.

Keywords: machine learning, benchmarking, FAIR, integrated omics, multi-omics, reproducibility, visualization, data heterogeneity

INTRODUCTION

In the last decade, the application of different individual omic studies (e.g., genomics, epigenomics, transcriptomics, proteomics, metagenomics) that aimed at understanding a particular problem in human disease (Karczewski and Snyder, 2018), agriculture (Ichihashi et al., 2020), plant science (Liu et al., 2016), microbiology (Quinn et al., 2016), and the environment have been successful to a great extent. These studies generate a plethora of data, which, with careful integration under a suitable statistical and mathematical framework, can help to solve broader queries pertaining to basic and applied areas of biology.

Abbreviations: AI, artificial intelligence; API, application programming interface; DL, deep learning; EDA, exploratory data analysis; FAIR, findable, accessible, interoperable, and reproducible; FDR, false discovery rate; GPU, graphics processing unit; KEGG, Kyoto Encyclopedia of Genes and Genomes; ML, machine learning; MOFA, multi-omics factor analysis; NGS, next generation sequencing; OR, odds ratio; PCA, principal component analysis; PMC, PubMed Central; QC, quality control; R, statistical programming language R; SNF, similarity network fusion; TCGA, The Cancer Genome Atlas.

More generally, performing multiple omics research often means having datasets with very different data modalities originating from varied assay types and increased dimensionality. In a multi-omics workflow (e.g., while profiling RNA, protein, or metabolites) the transcriptomics dataset, from RNA-seq efforts, can generate hundreds to thousands of transcripts (and the isoforms). In comparison, an individual researcher can only profile a few thousand proteins (and the proteoforms) or a few hundred identified metabolites (and features). Thus, the information burden from the transcriptome can easily overshadow the more actionable discoveries made from proteins or metabolites that are closer to the phenotype (Fiehn, 2002). This can add annotation bias and lead to enrichment of noise if robust integrative frameworks for data handling are not employed. Multi-omics aims to identify molecular markers associated with biological processes by revealing the regulatory units across diverse omics layers (e.g., obtained from DNA, RNA, proteins, metabolites, *etc.*). Multi-omics provides insights in understanding the mechanisms underlying biological processes and molecular functions, interactions and cellular fate, whether *in vivo* or *in vitro*, to reveal molecular phenotypes. Multi-omics can support discovery of predictive or prognostic biomarkers and/or potentially repurposed and novel drug targets in the era of precision medicine. Thus, the ultimate purpose of applied multi-omics is to increase the diagnostic yield for health, improve disease prognosis and produce improved agricultural outputs via robust understanding of genotype-to-phenotype relationship.

Figure 1 represents an artist's depiction of the complexity of multi-omics, a merger of omics-driven biology, data science, informatics and computational sciences. In spite of such challenges, the goal of multi-omics data is to support greater understanding of the overall biological process by bridging the gap of genotype-to-phenotype relationship.

We define multi-omics as three or more omic datasets coming from different layers of biological regulation – not necessarily within one level (exclusively derived from nucleic acid/DNA-derived, i.e., epigenomics, transcriptomics, and genomics). We have also not included proteogenomic that has immensely contributed to our improved understanding of protein sequences databases, gene annotations, gene models, and identification of peptides by interrogating genomics and transcriptomics while validating such protein data evidence using proteomics (Nesvizhskii, 2014). Further, this review does not discuss how other non-molecular data (i.e., phenotype data, clinical measures, imaging *etc.*) can be integrated with multiple omics datasets, as it entails a very different scope. While navigating this article, we recommend the readers consult **Box 1**, which contains the terms and concepts to support their understanding.

WHY IS MULTI-OMICS CHALLENGING?

Firstly, each individual omics analysis presents a multitude of challenges (Gomez-Cabrero et al., 2014; Misra et al., 2019). Multi-omics analysis inherits challenges from the single omics datasets, and confounds further analyses with other new challenges of the integration/fusion, clustering, visualization, and functional characterization (Pinu et al., 2019; Jamil et al., 2020). For instance, prior to integrating two or more omics, analysts or investigators can face challenges in terms of data harmonization (e.g., different data scaling, data normalization, and data transformation needs pertaining to individual omics dataset). Further, given dimensionality constraints posed while integrating large multiple omics data sets (e.g., a large population study with thousands of individual samples), the computational burden and storage space requirements can be limiting for a given study.

Even the identifiers (IDs) mapping – a prerequisite of some integration methods – is not an easy task when matching genes with associated transcripts or proteins (which is not a one-to-one correspondence), or a substantial challenge for other omics combinations, such as mapping genes to associated metabolites. Moreover, annotation of the omic entities (e.g., transcripts, proteins, and metabolites) with additional information, such as pathway membership and molecular characteristics, may require mapping IDs to various database systems (e.g., RefSeq or KEGG). Some of which may not cover all the omics of interest (e.g., metabolites are absent from RefSeq), while others may present outdated IDs due to delays after changes are made in the primary sources (e.g., KEGG GENE being based on RefSeq). The repertoire of identified and annotated molecules varies across omics, ranging from very good coverage of the genome, through a not-yet-complete picture of phosphoproteome and selective coverage of the metabolome. The challenges of metabolite identification may act as a bottleneck for advancement of the joint omics analyses. On the statistical side, unsupervised multi-omics methods can strengthen any signal, including systematic batch effect if present before quantitative measurements are taken, such as during sample acquisition, transport, processing logistics and operations. Failure to correct for such unwanted sources of technical variation, which may not be possible if the necessary information was not recorded during the sample handling steps, can misguide the overall integration process and impact the downstream interpretations and inferences (Kellman et al., 2020). **Figure 2** exemplifies the complexity of individual omics data heterogeneity and data sources in the multi-omics framework in a human focused, biomedical study. In the section below, we identify three of the major challenges and pitfalls that explain the above scenarios:

Data Wrangling

Also referred to as *"data munging,"* includes various levels of "transformation" and "mapping," is critical to the multi-omics field. Transformation is accomplished by data scaling, normalization, and imputation that help harmonize different omics data together. Category of "mapping" can be the process of harmonization of IDs across various omics data types or simply annotating data across available meta-data, a labor-intensive process that requires massive one-to-one or one-to-many relationship operations. Careful registration of samples and robust metadata recording tables, with involvement of data generation and analysis teams can help circumvent this challenge and mitigate errors.

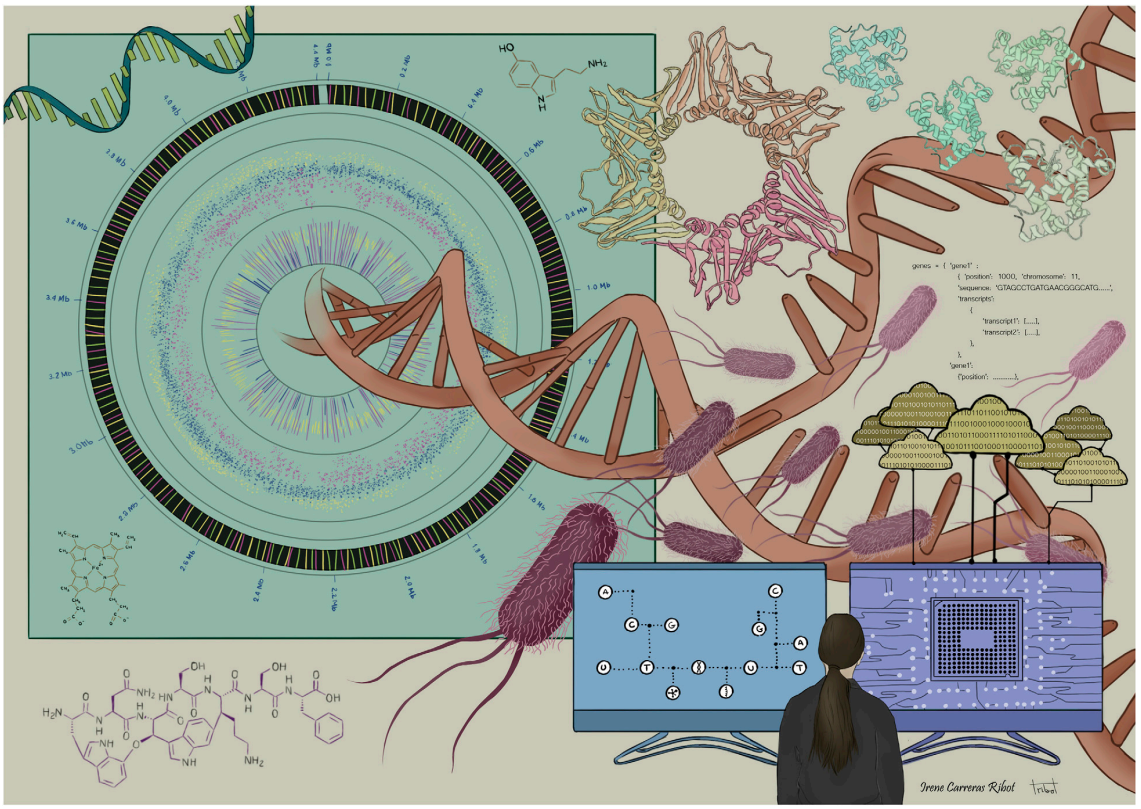

FIGURE 1 | The complexity of multi-omics: merger of omics-driven biology, data science, informatics, statistics, and computational sciences.

Data Heterogeneity

Data heterogeneity is often another bottleneck while dealing with multi-omics data as these are generated via varied technologies (i.e., consider sequencing versus mass-spectrometry, or microarray versus mass-spectrometry scenarios) and platforms (i.e., targeted versus untargeted, high resolution versus single cell). Pre-processing steps pertaining to individual datasets may not help overall, especially when democratizing them under a unified framework still remains challenging. However, some tools have led to improved handling, such as similarity network fusion (SNF) (Wang et al., 2014), mixOmics (Rohart et al., 2017), Multi-Omics Factor Analysis (MOFA) (Argelaguet et al., 2018), among others. Their utility depends on matrix factorization, network fusion, canonical correlation, factor analysis, and are used for downstream feature extraction and feature selection purposes for phenotypic prediction. Efforts have focused on dimension reduction (Meng et al., 2016), integration approaches while running into multicollinearity (Meng et al., 2014), and integration issues when dealing with multi-omics and non-omics data (López de Maturana et al., 2019) as explained below.

Dimension Reduction and Representation

Data representation, by means of dimensionality reduction that intends to project relationships of features (e.g., SNPs, transcripts, proteins, metabolites) across observations (e.g., samples, conditions, different omics layers) in a reduced space, is a common practice *a priori* in multi-omics efforts. Typically, following post-preprocessing after data normalization, data representation is applied to identify outliers, technical sources of variation – such as batch effects – and obvious biological patterns at each level of analysis – such as feature identification, extraction, and selection. This exercise aids in learning biological patterns and relationships of the data in bias identification and mitigation via appreciation of technical factors contributing to noise, adjusting them via batch effect correction, and identification of groups/sub-groups to confirm hypotheses of phenotypic conditions of interest in a given study. This is achieved by using clustering methods that are k-means, density-based, or graph-based, followed by generating visual representations using dimensionality reduction methods like principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP) to capture linear and non-linear relationships in the data. However, this approach is often challenging given the complexity of the analytical space and the study goals due to latent patterns encoded in input samples originating from different omics layers, technologies and platforms. Such complexities in representation can be attributed to the lack of optimal tunable algorithms both at mathematical and statistical levels. These challenges are well documented in bulk gene

BOX 1 | Terms, concepts, expressions, and definitions for clarity of readers foraying into multi-omics.

Terms, concepts, expressions	Definitions
Multi-omics/panomics/ integromics/integrated omics polyomics/transomics cross-omics	An approach aiming to improve the understanding of systems regulatory biology, molecular central dogma and genotype-phenotype relationship by combining 3 or more different omics data.
Multi-table, Multi-block	Terms focusing on the format of the data rather than its nature, popular in chemoinformatics (among other fields); can (but does not have to) imply a larger number of features than observations in the integrated tables/blocks.
Multi-view	Method often used in the field of ML for learning heterogeneity in the data and identification of patterns. By comparison to multiple cameras viewing an object from different angles, in omics context, the object can vary – whether it's "cell," "organism," or just "genome" viewed via different seq* techniques.
Multi-source	This term encompasses datasets that are derived from multiple sources of molecular assays. This terminology is used, for example by the joint and individual variation explained (JIVE) tool (O'Connell and Lock, 2016) during EDA.
Multi-modal	A term often used in omics in reference to multiple measurements methods done at molecular level to gain holistic insights of cellular machinery (e.g., one cell at a time). It is also popular in drug repositioning that involves integration of more nuanced *electronic health record* (EHR) data integration.
Central dogma of molecular biology	This is an explanation of the flow of genetic information within a biological system from DNA to RNA (transcription) to protein (translation) to metabolites (enzyme catalysis).
Machine learning (ML) method	Algorithm (a sequence of instructions) aimed at learning from data, with applications including exploration/dimensionality reduction (unsupervised methods, e.g., PCA, matrix factorization) and classification/prediction (supervised or semi-supervised methods)
Deep learning (DL) method	A subtype of ML using deep neural networks, composed of artificial neurons (signal aggregating or transforming units) arranged in layers; the depth of the DL refers to the number of "hidden" layers between the "input" (exclusive) and "output" layers (inclusive).
Fusion (Baldwin et al., 2020)	A specific type of integration that applies a uniform method in a scalable manner, to solve biological problems which the multi-omics measurements target.
Exploratory data analysis (EDA)	It is an approach that is heavily used in statistics, data science field during early data analysis steps often coupled with visualization.
Matrix factorization	A class of ML algorithms based on matrix decomposition, i.e., representation of a data matrix by two or more matrices (factors) that can be multiplied together to obtain the original matrix (or its approximation). It can be used for classification, prediction, or exploration.
Data heterogeneity	The data with a structural variation that can be explained by the composition of the analyzed dataset; encompasses both the clinical heterogeneity (e.g., presence of two groups with different genetic make-up due to ancestral differences, or different underlying etiologies of a disease) and technical heterogeneity (i.e., batch effects).
Meta-data	A table of organized information and instructions that helps to summarize the data properties in order to make it findable and usable for data analysis across same or multiple projects.
Git	A version-control system for tracking changes in source code and other documents during software development. Platforms such as Github and Gitlab are built on top of it.

expression studies that show that there is no single best latent dimensionality or compression algorithm for analyzing gene expression data (Way et al., 2020). Similarly, Hu and Greene (2018) proposed having a third-party evaluation by methods developers on unseen data while benchmarking autoencoder (unsupervised neural network) methods in single cell RNA-Seq (scRNA-seq) data for learning representations. These issues substantially change the results while interrogating high dimensional biological data. This problem is also applicable and extendable in multi-omics analytical space given the varied nature of data types in each omics layer with diverse biological modalities, such as while integrating single cell genome sequencing (genomics), RNA-Seq (transcriptomics), ATAC-Seq (epigenomics) and/or Bisulfite-Seq (epigenomics) together after pre-processing, batch-correction and normalization steps. Additionally, the data is also challenging to integrate as the relationship between multi-omics data layers can extend from one-to-one and one-to-many to many-to-many. This is also a very well-established concept in the Gene Regulatory Network (GRN) area of Systems Biology where gene-to-gene relationship establishments across various DNA, RNA, protein, metabolite, *etc*. are often better associated and represented using non-linear methods. Mutual Information (MI) based networks were found to perform better than other methods in such areas (Liu, 2017).

In **Figure 3**, we demonstrate a flow diagram to adhere to best practice guidelines in a multi-omics study for FAIR data sharing.

BEFORE YOU START: THE NEED FOR CONSULTATION AND PILOT DATA UPFRONT

Only a robust study design can lead to error-free execution of a multi-omics workflow. Though there are several proposed study design considerations and guidelines available for individual omics in genomics (Honaas et al., 2016) and metabolomics (Chu et al., 2019), such comprehensive guidelines are not developed for multi-omic studies to our knowledge. It is not surprising that the study design guidelines for individual omics vary in scope and coverage since each omic field faces different challenges and opportunities. Without proper experimental design, poorly planned multi-omics efforts lead to analytical complexity, non-informative inferencing, exclusion of tangible interpretations, overriding true biological signals, and eventually feed into the

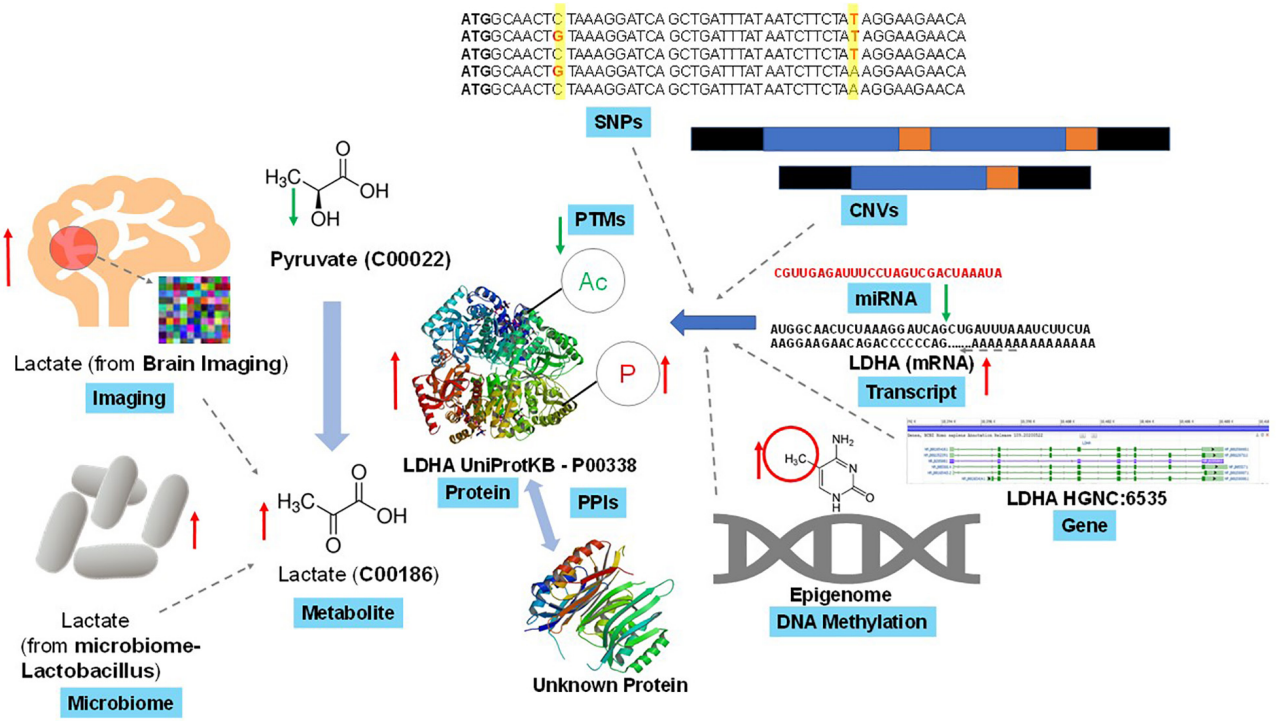

FIGURE 2 | Example of complexity and interconnectivity of omics data sources in a multi-omics framework. A simple cellular endogenous metabolite, lactate is biosynthesized enzymatically from pyruvate (another metabolite) with the help of lactate dehydrogenase (LDHA, a catalytic protein). In turn this LDHA can interact with several known and unknown proteins through protein-protein interactions to regulate its own function, and itself is subjected to diverse post-translational modifications (PTMs) that regulate its catalytic function. Lactate measurement through techniques such as *in vivo* brain imaging in human or other model animals can generate lactate's spatial distribution. Gut microbiome via Lactobacillus and other microbes can synthesize lactate and release into human physiological systems to contribute to lactate levels. Lactate biosynthesis regulation can be due to various levels of genetic (e.g., SNPs, CNV, etc.), transcriptomic, post-transcriptomic (e.g., miRNA) and/or epigenetics (e.g., DNA methylation) changes on the LDHA gene. Though this is one of the well-studied set of multi-omics interactions, but one can expect more complex and unknown interactions while integrating multi-omics datasets.

reproducibility crises plaguing high throughput omics domains. Some of the considerations needed to overcome these issues include: (a) careful assessment of statistical power and effect size appropriate to the experimental design, (b) identification of confounders (e.g., sex, age, input materials) inherent to the data, biases (e.g., replicates: biological and technical) and sources of variations (e.g., batch, analytical, unwanted) that are anticipated in the course of data generation, (c) quality assurance (QA) and quality control (QC) measures that are associated with individual omics data generation and analytical platforms and (d) cross-validation measures implemented in cases of unavoidable biases.

Sample Size and Statistical Power: Challenges and Opportunities

Different omics data require different numbers of samples to draw reliable conclusions. Reliability is dependent on false-discovery rate (FDR), which is influenced by the number of measured entities (i.e., transcripts/proteins/metabolites). Smaller omics data generation platforms such as microRNAs may need about 19 samples per experimental group to achieve a power of 0.8 at a fold change of 1.5 (Kok et al., 2018) with FDR < 0.1. Whereas, a set of 10,000 transcripts, each with at least 10 counts, would require a minimum of 35 samples per group for the same effect size at the same power and FDR control level, as calculated with *ssizeRNA* (Bi and Liu, 2016) using parameters pi0 = 0.8 disp = 0.1. The power calculation is not equally easy for each of the omics. While many tools were devised for transcriptomics/genomics power analysis, there are fewer dedicated tools available for metabolomics and proteomics studies. Only recently, a method to estimate an optimal sample size for multi-omic experiments was proposed (Tarazona et al., 2020) that addresses power calculation in multi-omics studies. This is one of the first comprehensive work that performed rigorous evaluations of relevant parameters across varied omic technologies (both sequencing and non-sequencing/i.e., mass-spectrometry based), built an open source tool (MultiPower[1]), that will enable future researchers to perform power and sample size estimation for their choice of multi-omics ED platforms while designing future studies and projects.

Further, pending cost-benefit tradeoff considerations, investigators typically decide on inclusion or exclusion of an individual omics experiment in a multi-omics setup. In certain cases, doubling the sample size is more informative than

[1]https://github.com/ConesaLab/MultiPower/

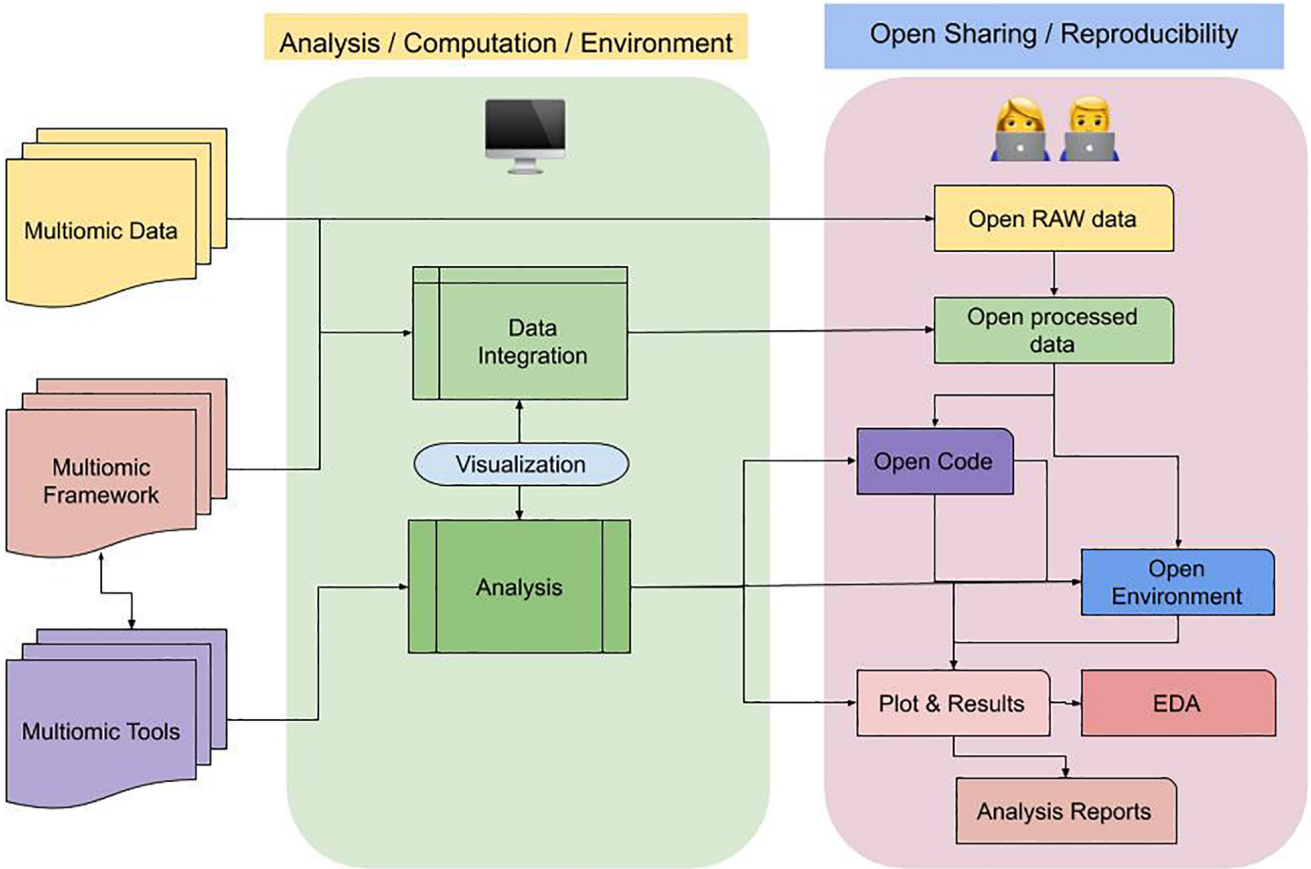

FIGURE 3 | Flow diagram of best practice guideline in a multi-omics study for FAIR sharing. A multi-omic study entails data varied assays/sources/omics type, that can be integrated using various framework and tools. This process (represented in block with light-green) can be computationally intensive. As a by-product we get processed data, which can be taken forward to do multiple steps involving exploration, inferencing and interpretations. Sharing both the data and code alongside compute environment allows interoperability and non-reinventing the wheel. Here (represented in third block with light-purple), describes the open sharing of different components in a multi-omic project, the connected blocks that can eventually generate reproducible results in forms of reports for users.

inclusion of an additional omics assay. For example, since small effects may not be clinically useful, increasing sample size may not be prudent when looking for biomarkers where assessing multi-omic panels may be more useful. When investigating disease subtypes, or patient stratification, a larger sample size may be desirable to achieve higher power in each of the subtypes. Subtyping of complex disease may benefit from diverse omics representation. Whereas, a study of biological mechanisms may benefit from related omics for a focused analysis of chosen omics types.

When planning a multi-omic analysis for a method that requires matched samples from all available omics datasets, the omic with the largest sample size requirement may dictate the need for such a large sample size across all analyzed omics. Here we provide two scenarios explaining the issue. For instance, in scenario one, when a study recruits 20 patients, collecting their biofluids for: genotyping, RNAseq, and metabolomics; and receives 19 genotypes, 18 transcriptomes and 17 metabolomes, one may incorrectly infer that the data is representative of 17 patients, but actually the failed samples (and QCs) originate from different patients across platforms. In reality, the experiment may result in only 14 patients with a complete set of measurements post QC across all omics. In scenario two, one study can recruit up to 100 patients but cannot afford to complete all three experiments on every patient. Hence, the researchers may decide to acquire data for 1000 genotypes, as those are affordable, and then split the transcriptome and metabolome equally to 70 a piece. This translates to matched samples for only 70 patients, thereby indicating missingness of data within and across omics layers. While the resulting missingness appears suboptimal, the integrative multi-omic design may allow researchers to decrease the sample size requirement; this is due to the increased potential of integrative analysis (Rappoport and Shamir, 2018). In this case, one can handle such sparsity by making a trade-off between genes (highly variable) given sample size is low or use sparsity methods in underlying available multi-omics frameworks. Moreover, the researchers may not consider each of the omics as equally important for their biological question and may be willing to focus on observations of larger effect sizes in an individual omics, which would drive

up the cost of the project. One of the recent works on such parameter harmonization and power size estimate in the realms of multi-omics is very well captured and addressed elsewhere (Tarazona et al., 2020).

Sample size is also an important consideration for multi-omic studies of rare diseases or difficult-to-access tissue, such as cerebrospinal fluid or endometrial tissue. These studies may struggle to recruit larger numbers of patients, exacerbating the disproportion between the number of samples and features. The early integration multi-omics strategies may be a good fit for such low sample-size experiments, as those allow to detect more subtle effects if consistently present across analyzed omics (Rappoport and Shamir, 2018). When choosing whether to include an additional omics layer, we advise a thorough examination of previous studies combining the omics intended for use, as the cost/benefit trade-off while including an additional omic layer may vary (information gain), the omic characteristics (e.g., signal/noise ratio) and the availability of validated computational methods for specific omics type or in combination.

Consulting Platform Experts and Incorporating Pilot Data

Given that the platform-specific characteristics–such as varying dispersion rates–require tailored solutions, researchers may require different parameters for RNAseq versus microarrays in transcriptomics, for liquid chromatography-mass spectrometry (LC-MS) versus aptamer-based proteomics or targeted versus untargeted metabolomics. Expert consultation is prudent before start of a pilot study to gauge the overall feasibility of the experiments and capabilities of the individual platforms in yielding optimal features (Tarazona et al., 2020), to design the final multi-omics study (note: the number of features or predictors in a given study is often denoted by 'p').

CURRENT STATE OF THE ART AND THE TOOLS

Multi-omics approaches can broadly be categorized as:

(a) Supervised – classification tasks that include discrete outcomes, such as disease/control status, and prediction tasks like that of continuous outcome, (e.g., survival, pain score).
(b) Exploratory – unsupervised clustering (e.g., disease subtype discovery) and relationship-based analysis (e.g., correlation/covariance and network models).

Even, over the past decade or so, a diverse array of multi-omics tools have been developed (Misra et al., 2019; Subramanian et al., 2020), some of which have gained popularity in recent years, including: mixOmics (Rohart et al., 2017), SNF (Wang et al., 2014), Paintomics (Hernández-de-Diego et al., 2018), 3Omics (Kuo et al., 2013), miodin (Ulfenborg, 2019), and MOFA (Argelaguet et al., 2018), as evident from the growing number of applications, user support requests, and citations. **Table 1** presents types of tools and resources which are useful for execution of a multi-omics workflow, together with the examples for each of the categories.

ADVANCES AND LIMITATIONS IN BENCHMARKING

The increasing reliance on computational methods necessitates systematic evaluation (benchmarking) of the omics data analysis tools and methods (Mangul et al., 2019). The key challenges in omics-scale benchmarking of computational tools, include: acquisition of "gold standard" datasets (providing unbiased ground truth), incorporating new methods for establishing benchmarks as they are published (continuous/extendible benchmarks), and ensuring reproducibility in the context of increasing complexity of the software involved (Mangul et al., 2019; Weber et al., 2019; Marx, 2020). Each of these challenges is amplified in the multi-omics field – matched omics measurements are more difficult to obtain, novel methods can rely on specific combinations of omics being available (limiting opportunities for extending previous benchmarks) and software requirements may increase in complexity as authors strive to combine results of multiple state-of-the-art single-omics tools for improved multi-omics performance.

Gold standard datasets that incorporate multiple omics and provide unbiased ground truth are a prerequisite for proper systematic evaluation of multi-omics methods. The Cancer Genome Atlas (TCGA), which includes genomic, epigenomic, transcriptomic, proteomic, and clinical data for 32 cancers (Blum et al., 2018), is a landmark dataset for multi-omics methods development. Our literature search reveals that references to TCGA are enriched in the multi-omics computational method articles compared to other article types (48.5% versus 19.7%, OR = 3.83, p-value = 4.5×10^{-07}, full-text analysis of the open-access PMC subset; see below for methods). While many other multi-omics datasets exist (e.g., for inflammatory bowel disease[2] or amyotrophic lateral sclerosis[3]); the community is yet to decide on a suitable "gold standard" across varied disease and tissue types, other than cancers. This process will require the expertise of domain-experts and characterization of statistical and technical properties of the datasets (e.g., presence of batch effects, analysis of confounders) (Marx, 2020).

A handful of notable multi-omics benchmarks are available, comparing: multi-omics and multi-view clustering algorithms (Rappoport and Shamir, 2018), multi-omics dimensionality reduction (Cantini et al., 2020) and multi-omics survival prediction methods (Herrmann et al., 2020). All three benchmarks were performed using the TCGA cancer data. While it is beneficial to use the same dataset for comparison, results obtained this way cannot be generalized beyond cancer biology, nor applied to the integration of other omics – such as metabolomics, or microbiome data – that are not included in the TCGA. With new multi-omic tools being developed, a comprehensive comparison against existing tools

[2]https://ibdmdb.org/
[3]http://data.answerals.org/

TABLE 1 | A complied list of various resources for supporting FAIR and interactive multi-omics study.

Serial No	Tools	Purpose	Link	References (if any)
Popular/Emerging Multi-omics Tools				
1	mixOmics	A tool with a framework that provides wide range of multivariate statistical methods for exploratory data analysis (EDA). This involves features identification, extraction and selection.	http://mixomics.org/	Rohart et al., 2017
2	MOFA	A probabilistic multi-omics factor analysis-based framework that involves EDA and data integration. (Unsupervised)	https://github.com/bioFAM/MOFA	Argelaguet et al., 2018
3	SNF	A multi-view network and fusion analysis framework for feature extraction, pairwise similarity, clustering, classification, etc.	https://cran.r-project.org/web/packages/SNFtool/index.html	Wang et al., 2014
4	miodin	A multi-level statistical framework involving vertical and horizontal integration of multi-omics data.	https://algoromics.gitlab.io/miodin/	Ulfenborg, 2019
5	Paintomics	A web-based systems biology tool for multi-omic integration and visualization across multi-species.	www.paintomics.org	Hernández-de-Diego et al., 2018
6	3Omics	A web-based application for integration and analysis of multi-omics data.	https://3omics.cmdm.tw/	Kuo et al., 2013
Data Sharing				
1	OmicsDI	An aggregated database facilitating the discovery of heterogenous published omics datasets across studies.	http://www.omicsdi.org	Perez-Riverol et al., 2017
2	Zenodo	A general-purpose open-access data, softwares, etc repository that allows user to obtain a citable DOI.	https://zenodo.org/	NA
3	OSF	An open platform to enable collaboration by registering research projects, materials, data and documentation.	https://osf.io/	NA
Code Sharing				
1	GitHub	A version-controlled code sharing and collaborative platform.	https://github.com/	NA
2	BitBucket		https://bitbucket.org/	NA
3	GitLab		https://about.gitlab.com/	NA
Workflow Sharing				
1	Common Workflow Language (CWL)	An open standard for describing analysis workflows which makes them portable and scalable across a variety of software and hardware environments.	https://www.commonwl.org/	Amstutz et al., 2016
2	Nextflow	An enterprise level workflow language for writing scalable and reproducible scientific pipelines.	https://www.nextflow.io/	Di Tommaso et al., 2017
3	Snakemake	A workflow language for writing scalable and reproducible scientific pipelines.	https://snakemake.readthedocs.io/en/stable/	Koster and Rahmann, 2012
Environment Sharing				
1	Conda	A package manager and computation environment management system.	https://docs.conda.io/en/latest/	NA
2	Bioconda	A channel for the conda package manager specializing in bioinformatics software.	https://bioconda.github.io/	Grüning et al., 2018
3	Docker	A container platform that provided OS-level virtualization for providing reproducible computation environment.	https://www.docker.com/	NA
4	BioContainers	A community-driven project that provides docker based containerized bioinformatics software.	https://biocontainers.pro/	da Veiga Leprevost et al., 2017
5	renv	A R-package that helps create reproducible environments for R-based projects.	https://rstudio.github.io/renv/	NA

(Continued)

TABLE 1 | Continued

Serial No	Tools	Purpose	Link	References (if any)
Data Visualization				
1	Shiny	A framework in R for doing GUI based interactive applications.	https://shiny.rstudio.com/	NA
2	Plotly	A cross language interactive plot library.	https://plotly.com/	NA
3	bokeh	A Python library for Interactive data visualization in browser.	https://bokeh.org/	NA
4	D3,js	A JavaScript library for producing dynamic, interactive data visualizations in web browsers.	https://d3js.org/	NA
5	Cytoscape	A platform for network data integration, analysis, and visualization.	https://cytoscape.org/	NA

is clearly missing, primarily attributable to limited availability of "gold standard" data sets. Other than the widely used multi-omics datasets from TCGA cancer patients, only limited studies incorporate simulated datasets, such as the R InterSIM package–which is also based on data dependence structure from the TCGA cancer studies.

Even the evaluation of a method on real-world data can be limited by the quality of the ground truth. One such scenario is the multiple multi-omic methods benchmarking against breast cancer subtypes that are primarily derived from a transcriptome based PAM50 signature (Bernard et al., 2009; Mathews et al., 2019). Such ground truth may favor the transcriptomic signal that could explain the limited perceived benefit of the multi-omics methods over single omics. Therefore, alternative strategies may be beneficial in the evaluation of subtypes derived by multi-omics methods (e.g., survival, drug response).

Given the limitations in the systematic characterization of multi-omic tools and methods, researchers need to choose tools that are either well benchmarked in appropriate scenarios and/or evidenced in multiple observational studies and systemically evaluated.

*FAIR*IFICATION OF MULTI-OMICS EFFORTS

Reproducing results in the multi-omics domain is understandably challenging because of the use of diverse data analysis methods, tools, and statistical processing, but as a research community we strive to make research efforts conform to findability, accessibility, interoperability, and reusability (FAIR) standards. Thus, the latest advancements in data sharing and environment replication can be leveraged to address this issue. In the following sections, we introduce means and approaches to share data, code, workflow, and environment while executing a multi-omics analysis to enhance the FAIRness (Wilkinson et al., 2016) which is suboptimal in the multi-omics field.

In order to determine the usage of multi-omics terms and their variants in the literature, to capture the trends in similar research domains, to identify their FAIRness in publications and the overrepresentation of research areas in them, we performed a systematic search (see **Figure 4**). We searched the PubMed database for articles pertaining to multi-omics on 25th July 2020, using fourteen terms (multi| pan| trans| poly| cross-omics, multi-table| source| view| modal| block omics, integrative omics, integrated omics and integromics) including plural/singular and hyphenated/unhyphenated variants and their combinations. The search was automated via Entrez E-utilities API and restricted to Text Words to avoid matching articles based on the affiliation of authors to commercial entities with such names. Further, the full text and additional metadata were retrieved from the PubMed Central (PMC) database for the open access subset of articles. Feature extraction was performed via n-gram matching against ClinVar (diseases and clinical findings) and NCBI Taxonomy (species) databases, while omics references annotation was based on regular expressions capturing phrases with suffix "-ome" or "-omic" (accounting for multi-omic phrases and plural variants). All disease and species matches were manually filtered down to exclude false or irrelevant matches and to merge plural forms. The article type was collated from five sources: (a) MeSH Publication Type as provided by PubMed, (b) community-maintained list of multi-omics software packages and methods available at https://github.com/mikelove/awesome-multi-omics [accessed on *2020-06-24*], (c and d) PMC-derived: Article Type and Subjects (journal-specific) and (e) manual annotation of articles published in Bioinformatics (Oxford, United Kingdom), due to lack of methods subject annotations in PMC data for this journal. The details and code are available in the online repository: https://github.com/krassowski/multi-omics-state-of-the-field.

The results of this systematic literature screen led to various interesting conclusions, as shown in **Figures 5A–E**. Primarily, our analysis revealed that multi-omic studies tend to focus on three layers of omics encompassing transcripts, genes, and proteins. This is followed by omics layers including metabolites and epigenetic modifications and combinations thereof (**Figure 5A**). A search of PubMed articles revealed that "multi-omics," as a terminology, is dominant over "integrated omics" and other omics-associated terms with an incremental trend since 2010 (**Figure 5B**). The search for "-ome" and "-omic" terms suggested that review articles tend to discuss the highest number of distinct omics, while computational methods articles appear to discuss the fewest, suggesting a potential disparity between the abilities

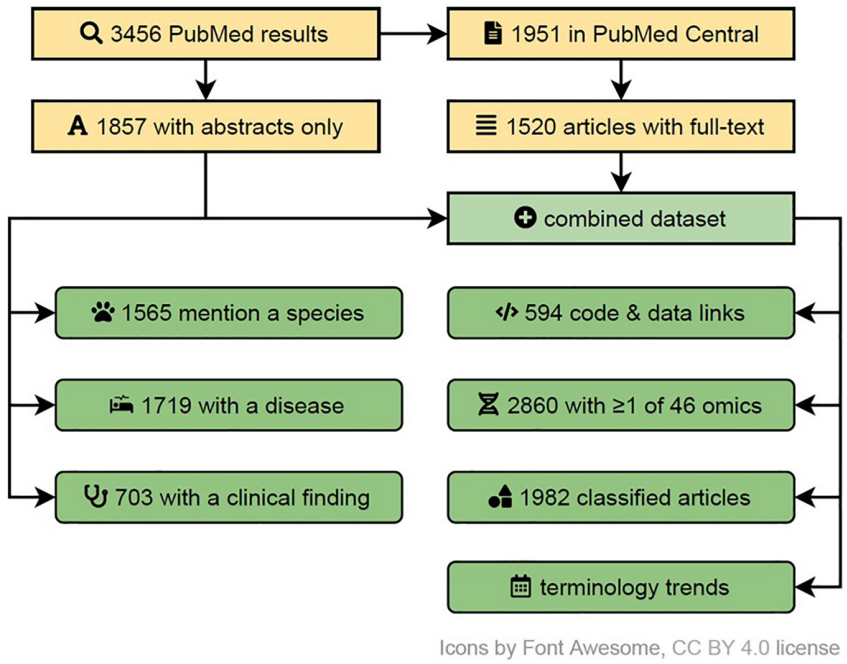

FIGURE 4 | A systematic flow diagram to screen multi-omics literature in PubMed indexes articles (up to July 2020). This flow diagram represents the various steps of inclusion and exclusion criteria used to identify varied characteristics and attributes associated with published multi-omics studies. A detailed self-explanatory method with reproducible code are available at https://github.com/krassowski/multi-omics-state-of-the-field.

of available computational tools and the ambitions and needs of the multi-omics community (**Figure 5C**). Of the disease terms, the multi-omic studies most frequently featured "cancer" and "carcinoma," while among the searched species "human" and "mice" dominated, indicating little representation of non-model species, organisms and biological systems. Articles mentioning "cancer" in title or abstract were overrepresented among the multi-omic articles when compared to other articles from the same time span, from the same journals and weighted by journal frequency in the multi-omics subset (22.7% vs. 7.5%, OR = 3.04, $p < 10^{-104}$) (**Figure 5D**). Toward FAIR sharing of data and code, "GitHub" appears to be the most popular platform, followed by "Bioconductor" and "Comprehensive *R* Archive Network (CRAN)," among many others (**Figure 5E**). Below we share few topics contributing to FAIR approaches:

Data Sharing

Different public databases are in place aiming to store and share specific kinds of omics data types as public repositories [e.g., genomics data in NCBI-SRA (Leinonen et al., 2011), GEO (Barrett et al., 2012) and EBI-ENA (European Bioinformatics Institute, 2016), proteomics data at PRIDE (Vizcaíno et al., 2016) and ProteomeXchange (Vizcaíno et al., 2014), or metabolomics data at MetaboLights (Haug et al., 2013), Metabolomics Workbench (Sud et al., 2016) and GNPS-MASSIVE (Wang et al., 2016)]. Only recently, have there been efforts to link these databases in a discoverable manner in the form of OmicsDI (Perez-Riverol et al., 2017). Mostly, raw sequences or very specific processed (count tables) data are being submitted to those databases, whereas, the intermediate outputs and analysis files are not shared, thus preventing reproducibility. The following resources can alleviate such scenarios: (a) Zenodo: allows users to upload raw data files, tables, figures and code. It supports code repositories, with GitHub integration, in addition to providing digital object identifiers (DOIs), and (b) OSF (Open Science Framework) (Foster and Deardorff, 2017): provides users with a platform where projects can be hosted with varied data types and file formats and contains a built in version control system. It also supports DOIs while promoting open source sharing that adheres with the FAIR guidelines.

However, adoption of such resources appears low in the multi-omics field as evident in our meta-analysis, with only 0.58% of publications (20 out of 3455 screened) linking to Dryad, OSF or Zenodo (**Figure 5E**).

Code Sharing

To enable FAIR sharing of code, a data analyst can explore one of the multiple venues available that publicly hosts codebases. These are: (a) GitHub, (b) Bitbucket, and (c) GitLab. All of these platforms use the Git system to provide version control. Also, native Markdown and Jupyter based notebooks render support for providing an exploratory data analysis (EDA) narrative alongside code and its output.

Workflow Sharing

As multi-omic analyses are often multi-step with each output being the input of another, in order to increase the efficiency workflows can be written with Domain Specific Languages (DSL)

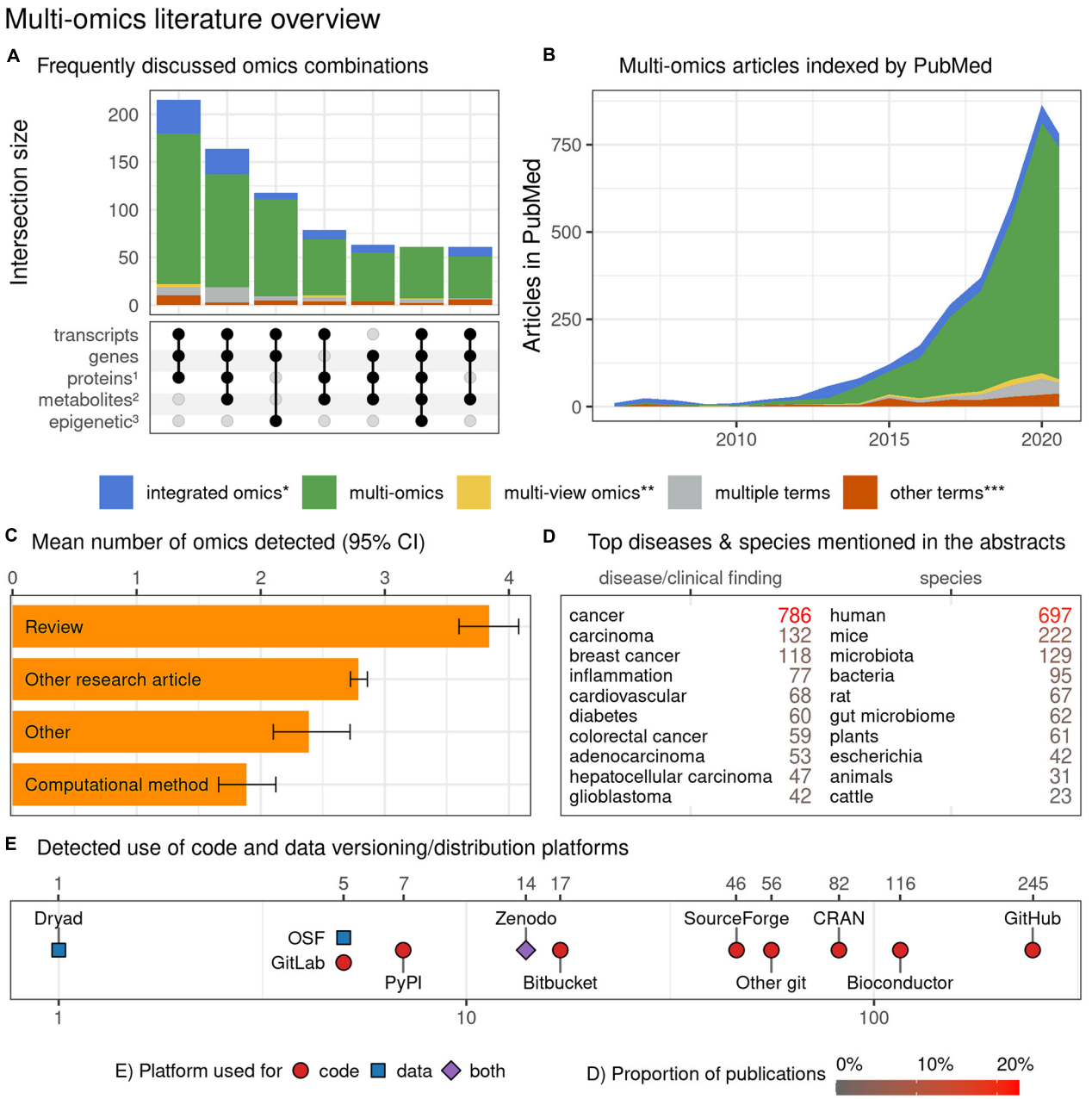

FIGURE 5 | Characterization of multi-omics literature based on a systematic screen of PubMed indexed articles (up to July 2020). **(A)** Combinations of omics (grouped by the characterized entities) commonly discussed occurring together in multi-omics articles (intersections with ≥ 3 omics and at least 50 papers). *The proteins group (1) also includes peptides; the metabolites group (2) includes other endogenous molecules; the epigenetic group (3) encompasses all epigenetic modifications.* **(B)** Trend plot representing the rapidly increasing number of multi-omics articles indexed in PubMed (also after adjusting for the number of articles published in matched journals – data not shown); the dip in 2020 can be attributed to indexing delay which was not accounted for in the current plot. **(C)** Distribution of article categories that mention different numbers of omics; while it is understandable that multi-omics "Review" category discusses many omics, the "Computational method" category articles appear to lag all other article category types. The detected number of omics may underestimate the actual numbers (due to the automated search strategy) but should put a useful lower bound on the number of omics discussed. Bootstrapped 95% confidence intervals around the mean are presented with the whiskers**. (D)** The number of articles mentioning the most popular clinical findings, disease terms (here screening is based on ClinVar diseases list) and species (based upon NCBI Taxonomy database). Both databases were manually filtered down to remove ambiguous terms and merge plural/singular forms. Only the abstracts were screened here**. (E)** The detected references to code, data versioning, distribution platforms and systems (links to repositories with deposited code/data); both the abstracts and full-texts (open-access subset, 44% of all articles) were screened. No manual curation to classify intent of the link inclusion (i.e., to share authors' code/data vs. to report the use of a dataset/tool) was undertaken. The details of the methods with reproducible code are available at github.com/krassowski/multi-omics-state-of-the-field. The comprehensive search terms (see the online repository for details) were collapsed into four categories; integrated omics (*) includes integromics and integrative omics, multi-view (**) includes multi-view| block| source| modal omics, other terms (***) include pan-, trans-, poly-, cross-omics.

such as: (a) Common Workflow Language (CWL) (Amstutz et al., 2016), (b) Nextflow (Di Tommaso et al., 2017), (c) Snakemake (Koster and Rahmann, 2012), and (d) Galaxy-workflows (McGowan et al., 2020).

Environment Sharing

The entire data analysis environment can be created and shared, saving time and aiding reproducibility (i.e., version control). Even accessing the intrinsic versioning information of each tool helps users in terms of interoperability, however, command line version handling parameters (e.g., −v/−V) are sometimes missing. The correction to a multi-omics clustering methods benchmark highlights the need for specifying the computational environment down to the processor architecture details (32 or 64 bit) (Rappoport and Shamir, 2019). As investigators attempt to build upon state-of-the-art implementations from various domains, like machine learning (ML), genetics, cell biology, the dependency on tools using different programming languages is incremental and some require a dedicated runtime environment (e.g., R and/or Python). Dedicated tools can help researchers who try to combine packages written in different languages in a single analysis workflow by allowing transparent data exchange and the use of interoperable functions across languages. One example of such a tool is the Python-R interface *rpy2* (rpy2, 2020), which found use in recent multi-omics tools (e.g., ReactomeGSA, Griss et al., 2020) and research scripts (Neyton et al., 2019). However, the use of multiple complex runtime environments can result in (version) conflicts if versions are not properly matched. This hinders the reuse of proposed tools and reproduction of published results. For example, each version of rpy2 requires a specific version of Python and R. The problem is not limited to Python → R workflow – the complimentary R → Python interface, *reticulate* (Reticulate, 2020) can be challenging to configure.

In order to ease the burden of interoperability and reproducibility that investigators often face while analyzing large multi-omics datasets with available algorithmic packages, several environment sharing avenues can be implemented, for example: (a) Conda (Conda, 2020): a cross-language tool repository and environment management system. With a shareable configuration (in yml format) file, an entire analysis environment can be re-installed in another system. Bioconda (Grüning et al., 2018) is a conda based project specifically designed for bioinformatic tools. (b) Docker (Docker, 2020): a ready to use lightweight portable virtual container, where an environment can be established, with all the required tools, for a particular analysis and shared. Specifically, bioinformatics tools such as Biocontainer (da Veiga Leprevost et al., 2017) are available. (c) Packrat (Packrat, 2020) (recently superseded by *renv*) and checkpoint: dependency management packages specific to R, which help to create isolated and portable R environments. *Checkpoint* facilitated one of the previous multi-omics benchmarking efforts (Herrmann et al., 2020).

Computational Power

Multi-omics analysis does not necessarily require high-performance computational resources, unless performing large scale consortia data extraction, transformation, load (ETL) tasks across a few hundred-thousand samples. However, some recent supervised multi-omics methods and packages can be computationally expensive given the amount of training that happens during the feature level analysis (e.g., Data Integration Analysis for Biomarker discovery using Latent Components (DIABLO), MOFA, *etc.*). Such bottlenecks can be overcome using a higher end central processing unit (CPU), high-performance computing cluster (HPC) and/or a cloud resource. The requirement of storing large downloaded files can be overcome using raw data streaming feature, however only a few tools support such feature.

Regulatory and Ethical, Legal, and Social Implications (ELSI) Issues

Additionally, multi-omics allows researchers to make more inferences on individuals in the event of a security incident, and labs/clinics that do translational research are often under regulatory compliances that restrict any data upload to any server for analysis when patient information is involved. There are multiple regulatory compliance-related restrictions spanning data security, ethical, personal information *etc.*, that can serve as bottleneck challenges. Alternatively, any researcher who develops a multi-omics tool for the community and makes it server/web/cloud-based should consider the needs of healthcare researchers who will often encounter restrictions when uploading such a dataset due to privacy concerns and other regulatory checks. In such cases, researchers can explore and take resources from non-open source enterprise level analytics platforms that can be either cloud-based or stand-alone if such enterprise platforms are Good Manufacturing Practices (GMP) certified, adhering to Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR). There can be additional regulatory compliances, given the data is produced by Clinical Laboratory Improvement Amendments (CLIA) certified entities. If all such regulatory compliances are in place, then patient data can be used in either a stand-alone third-party platform or uploaded in a web/cloud-based server for any analytics followed by inferencing under strict vigilance. For example, some commercial companies that have such cloud-based solutions include Amazon AWS, Google Cloud and MS Genomics (Microsoft Genomics, 2020). All of these platforms, together with other commercially available enterprise platforms like KNIME (KNIME4Bio | KNIME, 2020), can provide the necessary toolbox for multi-omics research and development.

APPLICATION OF MACHINE AND DEEP LEARNING (ML/DL) IN MULTI-OMICS

Over the years, machine learning (ML) and/or deep Learning (DL) have become increasingly popular in biomedical research due to their ability to perform unsupervised and supervised analyses using large datasets to provide logical or probabilistic inference. In the current data-driven era, apart from the large text mining exercises, pattern recognition and medical

imaging, ML/DL growth has contributed to analysis of large-scale high-dimensional data that are typically generated using high throughput omics assays. Their use and challenges in the multi-omics field are very well summarized in a recent review by Mirza et al. (2019) that discusses topics of integrative analysis encompassing dimensionality reduction/representation, data heterogeneity, data missingness, class imbalance and scalability issues. Other impressive applications of ML/DL are often encountered in regulatory genomics to study DNA-protein interactions and relationships. Some examples of relevant studies and models related to regulatory genomics approaches are available under http://kipoi.org/. However, much of ML/DL newer bioinformatics applications are developed in varied forms of supervised and unsupervised manner, such as specific neural networks models have been built for feature identification, extraction, and selection purposes. Some of these approaches in the DL space are discussed in the review by Ching et al. (2018). Such DL models have been extensively used for multi-omics integration purposes to predict better molecular signatures associated with improved patient survival and capture intricate relationship patterns for better clustering over conventional methods and drug response prediction. Such pattern extraction, selection and representation are often difficult to achieve solely by traditional linear modeling unless coupled with advanced non-linear models. Some methods and tools from the multivariate statistics/ML/DL area that have been developed for multi-omics integration include: (a) Multi-Omics Model and Analytics (MOMA), (b) Multiple Kernel Learning (MKL) (Wilson et al., 2019), (c) DIABLO (Singh et al., 2019), (d) a multi-omics late integration method (MOLI) (Sharifi-Noghabi et al., 2019), (e) multi-omics deep learning method (DCAP) (Chai et al., 2019), and (f) Multi-omics Autoencoder Integration (MAUI) (Ronen et al., 2019). Partly this can be attributed to the reasons described above and partly as described in the following paragraph.

Often simple models do not account for the principles of dynamics and kinetics that underlie a set of biological processes. Considering central dogma as the key hypothesis (Reinagel and Speth, 2016) of molecular life, for the entire process from replication through transcription to translation machineries that are at play, each of these biological processes (i.e., a disease) have pre and post events that are building more complex functions at each step adding up to the biological stochasticity. These stochastic events are often not well accounted for in simpler models as researchers tend to overgeneralize using mathematical modeling, calculus and/or statistics. Frequently, such strategies are not adopted in multi-omics experimental design and also, as datasets are not always longitudinal in nature, they can often lead to biases or ineffective generalization or approximation in multi-omics results. Another argument occurs when DNA and RNA are assumed as distinct genetic materials. DNA and RNA can work individually to bring about structural or functional protein consequences that lead to a phenotypic change. This was addressed to an extent by Koonin (2012), where central dogma is challenged by "*genetic assimilation of prion-dependent phenotypic heredity,*" and only a few phenotypes might fall under such categories and phenomena. This can be due to (a) genetic insults, like chromosomal instability and loss of function mutations that directly impact the translational process, (b) insults to RNA machinery without upstream DNA impact, while any abnormalities in the RNA phase impinges the translational events and (c) insults possibly seen in few systemic diseases where not everything is reliant on DNA or germline mutations, but rather due to abnormality in the underlying regulatory machineries during transcription or pre-translation stages. Such events can often be guided by upstream epigenetic insults like DNA methylation, histone modifications or even specific enhancer binding processes on a different gene promoter thus impacting overall transcription and translation, leading to a phenotype.

Even at the level of proteins, the regulome is often guided by protein-protein interactions, and those by kinases and phosphatases, are barely predictable from the genome. Similarly, regulations of metabolite levels (catabolic and anabolic processes leading to their levels in a given system) are not predictable from the enzyme levels, let alone their protein or DNA sequences. These kinds of upstream processes are often not well captured via omics technology, as our current models or frameworks are yet to be fully optimized and cannot generalize at such a level of non-linear system dynamic relationships that leads to specific phenotypic processes (Reinagel and Speth, 2016). Taken together, all of the above lead to the motivation of developing more advanced variants of ML/DL-based tools in biomedical research for multi-omics integration to improve understanding of genotype to phenotype relationships. However, these methods can be very computationally expensive and not robustly validated as they will be under continuous development.

DATA VISUALIZATION TOOLS

Visual representation is one of the most important ways of deriving interpretations and inferences with data in multi-omics. With the advent of high-dimensional data generation platforms, such as NGS technology and mass spectrometry, such representation has become very popular. Currently, there is a trend of developing dynamic web-based and stand-alone applications among the larger research community in diverse omics domains. These are often published alongside code for reproducibility of the results as an additional resource for other users in the research community to explore and for hypothesis development. Visualization avenues of multidimensional data in an interactive platform adheres to FAIR standards. The need for joint visualization of multiple omics datasets prompted the adoption of dashboarding applications, such as BioTools (Biotools, 2020) and WilsON (WIlsON, 2020). Dashboards display together multiple interactive panels with high-dimensional data and are available for the majority of data-exploration ecosystems (e.g., R, Python, Jupyter, Tableau). The interactive visualization tools and dashboards can be installed locally as stand-alone tools (e.g., in workstation/server) or can be completely web-browser based (e.g., launched locally from a server or a cloud-based platform).

Some of these popular tools that have found application in multi-omics are: (a) R-based Shiny (Shiny, 2020) apps. Numerous Shiny based apps help with exploratory data analysis

for testing of hypotheses, given the end-user is able to grasp the underlying statistical models/frameworks that perform a required task of a specialized biological query. Such shiny apps (Dwivedi and Kowalski, 2018; Kmezhoud/BioCancer, 2020; WIlsON, 2020) can be launched both locally on a computer, server or even hosted publicly catering to a larger community of researchers. Binder (Jupyter et al., 2018) allows researchers to quickly create the computational environment needed to interact with research code and data shared online. *Voilà* (Voilà, 2020) turns Jupyter notebooks into standalone web applications. (b) Plotly (Plotly, 2020) (multiple languages; both open source and commercial) includes several tools designed for using these resources either in a stand-alone manner or in conjunction with other available frameworks (Zeng et al., 2019). In a way similar to Shiny, it supports creation of complex dashboards when used with Python-oriented Plotly Dash. (c) Python-based tools with or without integration servers like bokeh (Bokeh, 2020) enables Python users to create interactive web-based applications for end-users with front-end. (d) Network and other advanced visualizations, including JavaScript-based libraries such as D3.js (data-driven documents) (D3.js, 2020), have functionality amenable for web-based network tools creation. Cytoscape (Otasek et al., 2019; Cytoscape, 2020), available both as a JavaScript library for online visualizations (Cytoscape.js) and stand-alone application for EDA, is a popular tool employed in the field of systems biology. Bacnet (BACnet Stack, 2020) is another available framework for developing custom multi-omics analysis websites including network and other advanced visualizations.

COMPUTATIONAL RESOURCES NEEDED FOR MULTI-OMICS ANALYSIS

In the following sections, we provide pointers for using computational resources and expertise needed for executing a multi-omics experiment.

Knowledge of Programming Languages and Frameworks

Provided below are a few programming languages that are relevant and applicable to experiments in multi-omics: (a) Bash scripting and Python are useful for basic data pre-processing and workflow organization, (b) C/C++/Java may be useful for development of performant methods and algorithms, (c) R and Python are *de facto* standard for statistical programming and data visualization in the omics context and (d) Shiny/Bokeh are visualization frameworks convenient for creating web-based interactive multi-omics functions.

Computational Infrastructure

We advise learning to handle a standard Linux distribution, enterprise-level or open-source cloud-based computational interface, such as Google Colab in order to run workflows/pipelines for EDA and launching softwares/tools for performing any integrative multi-omics/bioinformatics related tasks. These infrastructures can feed into varied analytical tasks, such as data wrangling, data integration, data analytics, data visualization, and functional analysis. Given such varied data intensive tasks are associated with multi-omics analysis, more often users need resources for stand-alone workstations with well powered Central Processing Unit (CPU), servers having Graphical Processing Unit (GPU) or high-end computing infrastructures with Tensor Processing Unit (TPU). The need of a GPU or TPU is however needed while running end-to-end ML/DL models with high-volume features and parameters.

Databases, Visualizers and Portals

Numerous portals, databases, and data-centric tools can be used for integrative multi-omics explorations. Examples of those are cBioportal (Gao et al., 2013) (Cancer Bioportal); Xena browser (Goldman et al., 2019) (UCSC Xena Browser is an online exploratory tool for analyzing public and private, multi-omic and clinical/phenotype datasets); ICGC Data portal (Zhang et al., 2011) (International Cancer Genome Consortium Data portal); ENCODE Data Portal (Davis et al., 2018) [The Encyclopedia of DNA Elements (ENCODE) is a public research project which aims to identify functional elements in the human genome]; FANTOM5 (functional annotation of the mammalian genome 5) (The FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014) and The Human Protein Atlas (HPA) (Thul et al., 2017). It is also important to gain basic knowledge of the underlying methods employed in these large databases by reading the associated manuscripts, frequently asked questions and tutorials/vignettes in order to gain substantial knowledge before using them for exploratory purposes.

FUTURE PROSPECTS AND CONCLUSION

Challenges abound – from dealing with biological complexity, to over-simplified models, to technological limitations associated with data generation, to organization of high throughput data for comprehensible visualization, to drawing meaningful conclusions. In this treatise, we did not cover the success achieved with multi-omics in various domains of microbial, plant, animal, and biomedical research in recent times to keep the scope focused and relevant to a diverse audience.

In this document, we have not touched upon several upcoming and exciting areas of multi-omics research as they are yet to mature. For instance, single-cell multi-omics are currently driven with efforts mostly at the genomic (single cell DNASeq), transcriptional (e.g., single cell/single nuclear) and epigenomic (single ATAC-Seq, single cell bisulfite sequencing) levels. They are currently in the early stages of inception and, as more promising works will ensue, researchers will reach precision with efficient capture of single cell proteomics and metabolomics. Currently, some early single cell proteomics work is emerging in the mass spectrometry driven omics area of proteomics (e.g., SCOPE2) (Specht et al., 2019). Prevailing challenges remain in terms of maximizing information from a single cell (Macaulay et al., 2017) using current proteomics and metabolomics strategies, where barely a handful of metabolites are captured (Nemes et al., 2012). However, there are already

some early exciting works of single cell multi-omics integration methods available that are upcoming in manifold [e.g., MAGAN (Amodio and Krishnaswamy, 2018), UnionCom (Cao et al., 2020) and non-manifold – such as LIGER (Welch et al., 2019) and MOFA+ (Argelaguet et al., 2020)]. Hopefully, these will be addressed and covered in future multi-omics efforts.

From collective experience and evidence, the key to effective exploratory data analysis, hypothesis generation and interpretations is reliant – to an extent – on understanding the underlying methods used to build or digest them and draw inferences. With more high dimensional biological data generation in various arms of biology, be it plant, microbial, developmental/disease biology, and future implementation of various multi-modal multi-omics, it will be more likely to observe growth of such ML/DL methods. Hence, the applied ML/DL community in the bioinformatics domain will have to generate models that are interoperable, stable, and well benchmarked at various regularizations (tunable) for users to derive robust reproducible results. Alternatively, such ML/DL developers and researchers can also clarify the uncertainty bounds associated with their tools for the user community. As a nascent field, there is a dearth of studies or benchmark tools and resources to direct an upcoming community, but this review serves as a guideline for future multi-omics researchers from a computational standpoint.

AUTHOR CONTRIBUTIONS

BBM and VD conceived the idea. MK performed the meta-analysis. MK, VD, SS, and BBM wrote the manuscript. MK and SS generated the tables and figures. All the authors have read, agreed to the content, and approved the submitted version of the manuscript.

ACKNOWLEDGMENTS

We would like to acknowledge the independent reviewers and the editor for their comments to help improve this manuscript. We would like to thank the developers and researchers of the multi-omics community who drive the field forward with their code, packages, tools, and resources, whether their work was discussed or not included in this review (due to space limitations or inadvertently). We also acknowledge a paid artist Ms. Irene Carreras Ribot for generating **Figure 1**.

REFERENCES

Amodio, M., and Krishnaswamy, S. (2018). "MAGAN: aligning biological manifolds," in *35th International Conference on Machine Learning ICML 2018*, Vol. 1, Stockholm, 327–335.

Amstutz, P., Chapman, B., Chilton, J., Heuer, M., and Stojanovic, E. (2016). *Common Workflow Language, v1.0 Common Workflow Language (CWL) Command Line Tool Description, v1.0*. doi: 10.6084/m9.figshare.3115156.v2

Argelaguet, R., Arnol, D., Bredikhin, D., Deloro, Y., Velten, B., Marioni, J. C., et al. (2020). MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* 21:111. doi: 10.1186/s13059-020-02015-1

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., et al. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* 14:e8124. doi: 10.15252/msb.20178124

BACnet Stack (2020). *BACnet Stack*. Available online at: https://github.com/bacnet-stack (accessed August 3, 2020).

Baldwin, E., Han, J., Luo, W., Zhou, J., An, L., Liu, J., et al. (2020). On fusion methods for knowledge discovery from multi-omics datasets. *Comput. Struct. Biotechnol. J.* 18, 509–5017. doi: 10.1016/j.csbj.2020.02.011

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bernard, P. S., Parker, J. S., Mullins, M., Cheung, M. C. U., Leung, S., Voduc, D., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27, 1160–1167. doi: 10.1200/JCO.2008.18.1370

Bi, R., and Liu, P. (2016). Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics* 17:146. doi: 10.1186/s12859-016-0994-9

Biotools (2020). *Biotools*. Available online at: https://www.biotools.fr/ (accessed August 2, 2020).

Blum, A., Wang, P., and Zenklusen, J. C. (2018). SnapShot: TCGA-analyzed tumors. *Cell* 173:530. doi: 10.1016/j.cell.2018.03.059

Bokeh (2020). *Bokeh*. Available online at: https://bokeh.org/ (accessed August 3, 2020).

Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., et al. (2020). Benchmarking joint multi-omics dimensionality reduction approaches for cancer study. *bioRxiv*. doi: 10.1101/2020.01.14.905760

Cao, K., Bai, X., Hong, Y., and Wan, L. (2020). Unsupervised topological alignment for single-cell multi-omics integration. *bioRxiv* [Preprint]. doi: 10.1101/2020.02.02.931394

Chai, H., Zhou, X., Cui, Z., Rao, J., Hu, Z., Lu, Y., et al. (2019). Integrating multi-omics data with deep learning for predicting cancer prognosis. *bioRxiv* [Preprint]. doi: 10.1101/807214

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* 15:20170387. doi: 10.1098/rsif.2017.0387

Chu, S., Huang, M., Kelly, R., Benedetti, E., Siddiqui, J., Zeleznik, O., et al. (2019). Integration of metabolomic and other omics data in population-based study designs: an epidemiological perspective. *Metabolites* 9:117. doi: 10.3390/metabo9060117

Conda (2020). *Conda*. Available online at: https://anaconda.org/anaconda/conda (accessed August 2, 2020).

Cytoscape (2020). *Cytoscape*. Available online at: https://cytoscape.org/ (accessed August 3, 2020).

D3.js (2020). *D3. js*. Available online at: https://d3js.org/ (accessed August 3, 2020).

da Veiga Leprevost, F., Grüning, B. A., Alves Aflitos, S., Röst, H. L., Uszkoreit, J., Barsnes, H., et al. (2017). BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* 33, 2580–2582. doi: 10.1093/bioinformatics/btx192

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., et al. (2018). The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 46, D794–D801. doi: 10.1093/nar/gkx1081

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* 35, 316–319. doi: 10.1038/nbt.3820

Docker (2020). *Docker*. Available online at: https://www.docker.com/ (accessed August 2, 2020).

Dwivedi, B., and Kowalski, J. (2018). shinyGISPA: a web application for characterizing phenotype by gene sets using multiple omics data combinations. *PLoS One* 13:e0192563. doi: 10.1371/journal.pone.0192563

European Bioinformatics Institute (2016). *European Nucleotide Archive*. Available online at: http://www.ebi.ac.uk/ena (accessed August 2, 2020).

Fiehn, O. (2002). Metabolomics – The link between genotypes and phenotypes. *Plant Mol. Biol.* 48, 155–171. doi: 10.1023/A:1013713905833

Foster, E. D., and Deardorff, A. (2017). Open science framework (OSF). *J. Med. Libr. Assoc.* 105, 203–206. doi: 10.5195/jmla.2017.88

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6:pl1. doi: 10.1126/scisignal.2004088

Goldman, M., Craft, B., Hastie, M., Repeèka, K., McDade, F., Kamath, A., et al. (2019). The UCSC Xena platform for public and private cancer genomics data visualization and interpretation. *bioRxiv* [Preprint]. doi: 10.1101/326470

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). Data integration in the era of omics: current and future challenges. *BMC Syst. Biol.* 8:I1. doi: 10.1186/1752-0509-8-S2-I1

Griss, J., Viteri, G., Sidiropoulos, K., Nguyen, V., Fabregat, A., and Hermjakob, H. (2020). ReactomeGSA – efficient multi-omics comparative pathway analysis. *bioRxiv* [Preprint]. doi: 10.1101/2020.04.16.044958

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., et al. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* 15, 475–476. doi: 10.1038/s41592-018-0046-7

Haug, K., Salek, R. M., Conesa, P., Hastings, J., de Matos, P., Rijnbeek, M., et al. (2013). MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* 41, D781–D786. doi: 10.1093/nar/gks1004

Hernández-de-Diego, R., Tarazona, S., Martínez-Mira, C., Balzano-Nogueira, L., Furió-Tarí, P., Pappas, G. J., et al. (2018). PaintOmics 3: a web resource for the pathway analysis and visualization of multi-omics data. *Nucleic Acids Res.* 46, W503–W509. doi: 10.1093/nar/gky466

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., and Boulesteix, A.-L. (2020). Large-scale benchmark study of survival prediction methods using multi-omics data. *arXiv* [Preprint], Available online at: http://arxiv.org/abs/2003.03621 (accessed August 2, 2020).

Honaas, L. A., Altman, N. S., and Krzywinski, M. (2016). Study Design for Sequencing Studies. *Methods Mol. Biol.* 1418, 39–66. doi: 10.1007/978-1-4939-3578-9_3

Hu, Q., and Greene, C. S. (2018). Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics. *bioRxiv* [Preprint]. doi: 10.1101/385534

Ichihashi, Y., Date, Y., Shino, A., Shimizu, T., Shibata, A., Kumaishi, K., et al. (2020). Multi-omics analysis on an agroecosystem reveals the significant role of organic nitrogen to increase agricultural crop yield. *Proc. Natl. Acad. Sci. U.S.A.* 117, 14552–14560. doi: 10.1073/pnas.1917259117

Jamil, I. N., Remali, J., Azizan, K. A., Nor Muhammad, N. A., Arita, M., Goh, H.-H., et al. (2020). Systematic multi-omics integration (MOI) approach in plant systems biology. *Front. Plant Sci.* 11:944. doi: 10.3389/fpls.2020.00944

Jupyter, P., Bussonnier, M., Forde, J., Freeman, J., Granger, B., Head, T., et al. (2018). "Binder 2.0," in *Proceedings of the 17th Python in Science Conference (SciPy)*, Austin, TX, 113–120. doi: 10.25080/majora-4af1f417-011

Karczewski, K. J., and Snyder, M. P. (2018). Integrative omics for health and disease. *Nat. Rev. Genet.* 19, 299–310. doi: 10.1038/nrg.2018.4

Kellman, B. P., Baghdassarian, H. M., Pramparo, T., Shamie, I., Gazestani, V., Begzati, A., et al. (2020). Multiple freeze-thaw cycles lead to a loss of consistency in poly(A)-enriched RNA 1 sequencing. *bioRxiv* [Preprint]. doi: 10.1101/2020.04.01.020792

Kmezhoud/BioCancer. (2020). *A Shiny App for Interactive Multi-OMICS Cancer Data Visualization and Analysis*. Available online at: https://github.com/kmezhoud/bioCancer (accessed October 18, 2020).

KNIME4Bio | KNIME (2020). *KNIME4Bio | KNIME*. Available online at: https://www.knime.com/community/knime4bio (accessed August 3, 2020).

Kok, M. G. M., de Ronde, M. W. J., Moerland, P. D., Ruijter, J. M., Creemers, E. E., and Pinto-Sietsma, S. J. (2018). Small sample sizes in high-throughput miRNA screens: a common pitfall for the identification of miRNA biomarkers. *Biomol. Detect. Quantif.* 15, 1–5. doi: 10.1016/j.bdq.2017.11.002

Koonin, E. V. (2012). Does the central dogma still stand? *Biol. Direct.* 7:27. doi: 10.1186/1745-6150-7-27

Koster, J., and Rahmann, S. (2012). Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi: 10.1093/bioinformatics/bts480

Kuo, T. C., Tian, T. F., and Tseng, Y. J. (2013). 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst. Biol.* 7:64. doi: 10.1186/1752-0509-7-64

Leinonen, R., Sugawara, H., and Shumway, M. (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi: 10.1093/nar/gkq1019

Liu, H., Wang, F., Xiao, Y., Tian, Z., Wen, W., Zhang, X., et al. (2016). MODEM: multi-omics data envelopment and mining in maize. *Database* 2016:baw117. doi: 10.1093/database/baw117

Liu, Z. P. (2017). Quantifying gene regulatory relationships with association measures: a comparative study. *Front. Genet.* 8:96. doi: 10.3389/fgene.2017.00096

López, de Maturana, E., Alonso, L., Alarcón, P., Martín-Antoniano, I. A., Pineda, S., et al. (2019). Challenges in the integration of omics and non-omics data. *Genes* 10:238. doi: 10.3390/genes10030238

Macaulay, I. C., Ponting, C. P., and Voet, T. (2017). Single-cell multiomics: multiple measurements from single cells. *Trends Genet.* 33, 155–168. doi: 10.1016/j.tig.2016.12.003

Mangul, S., Martin, L. S., Hill, B. L., Lam, A. K. M., Distler, M. G., Zelikovsky, A., et al. (2019). Systematic benchmarking of omics computational tools. *Nat. Commun.* 10, 1–11. doi: 10.1038/s41467-019-09406-4

Marx, V. (2020). Bench pressing with genomics benchmarkers. *Nat. Methods* 17, 255–258. doi: 10.1038/s41592-020-0768-1

Mathews, J. C., Nadeem, S., Levine, A. J., Pouryahya, M., Deasy, J. O., and Tannenbaum, A. (2019). Robust and interpretable PAM50 reclassification exhibits survival advantage for myoepithelial and immune phenotypes. *npj Breast Cancer* 5:30. doi: 10.1038/s41523-019-0124-8

McGowan, T., Johnson, J. E., Kumar, P., Sajulga, R., Mehta, S., Jagtap, P. D., et al. (2020). Multi-omics visualization platform: an extensible galaxy plug-in for multi-omics data visualization and exploration. *Gigascience* 9:giaa025. doi: 10.1093/gigascience/giaa025

Meng, C., Kuster, B., Culhane, A. C., and Gholami, A. M. (2014). A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 15:162. doi: 10.1186/1471-2105-15-162

Meng, C., Zeleznik, O. A., Thallinger, G. G., Kuster, B., Gholami, A. M., and Culhane, A. C. (2016). Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* 17, 628–641. doi: 10.1093/bib/bbv108

Microsoft Genomics (2020). *Microsoft Genomics*. Available online at: https://www.microsoft.com/en-us/genomics/ (accessed August 2, 2020).

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Genes* 10:87. doi: 10.3390/genes10020087

Misra, B. B., Langefeld, C., Olivier, M., and Cox, L. A. (2019). Integrated omics: tools, advances and future approaches. *J. Mol. Endocrinol.* 62, R21–R45.

Nemes, P., Knolhoff, A. M., Rubakhin, S. S., and Sweedler, J. V. (2012). Single-cell metabolomics: changes in the metabolome of freshly isolated and cultured neurons. *ACS Chem. Neurosci.* 3, 782–792. doi: 10.1021/cn300100u

Nesvizhskii, A. I. (2014). Proteogenomics: concepts, applications and computational strategies. *Nat. Methods* 11, 1114–1125. doi: 10.1038/nmeth.3144

Neyton, L., Zheng, X., Skouras, C., Wilson, A. B., Gutmann, M. U., Yau, C., et al. (2019). Multiomic definition of generalizable endotypes in human acute pancreatitis. *bioRxiv* [Preprint]. doi: 10.1101/539569

O'Connell, M. J., and Lock, E. F. (2016). R.JIVE for exploration of multi-source molecular data. *Bioinformatics* 32, 2877–2879. doi: 10.1093/bioinformatics/btw324

Otasek, D., Morris, J. H., Bouças, J., Pico, A. R., and Demchak, B. (2019). Cytoscape Automation: empowering workflow-based network analysis. *Genome Biol.* 20:185. doi: 10.1186/s13059-019-1758-4

Packrat (2020). *Packrat*. Available online at: https://rstudio.github.io/packrat/ (accessed August 2, 2020).

Perez-Riverol, Y., Bai, M., Da Veiga, Leprevost, F., Squizzato, S., Park, Y. M., et al. (2017). Discovering and linking public omics data sets using the omics discovery index. *Nat. Biotechnol.* 35, 406–409. doi: 10.1038/nbt.3790

Pinu, F. R., Beale, D. J., Paten, A. M., Kouremenos, K., Swarup, S., Schirra, H. J., et al. (2019). Systems biology and multi-omics integration: viewpoints from the metabolomics research community. *Metabolites* 9:76. doi: 10.3390/metabo9040076

Plotly (2020). *Plotly*. Available online at: https://plotly.com/ (accessed August 3, 2020).

Quinn, R. A., Navas-Molina, J. A., Hyde, E. R., Song, S. J., Vázquez-Baeza, Y., Humphrey, G., et al. (2016). From sample to multi-omics conclusions in under 48 hours. *mSystems* 1:e00038-16. doi: 10.1128/mSystems.00038-16

Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* 46, 10546–10562. doi: 10.1093/nar/gky889

Rappoport, N., and Shamir, R. (2019). Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.* 47, 1044–1044. doi: 10.1093/nar/gky1226

Reinagel, A., and Speth, E. B. (2016). Beyond the central dogma: model-based learning of how genes determine phenotypes. *CBE Life Sci. Educ.* 15:ar4. doi: 10.1187/cbe.15-04-0105

Reticulate (2020). *Reticulate*. Available online at: https://rstudio.github.io/reticulate/ (accessed August 2, 2020).

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* 13:e1005752. doi: 10.1371/journal.pcbi.1005752

Ronen, J., Hayat, S., and Akalin, A. (2019). Evaluation of colorectal cancer subtypes and cell lines using deep learning. *Life Sci. Alliance* 2:e201900517. doi: 10.26508/lsa.201900517

rpy2 (2020). *rpy2*. Available online at: https://pypi.org/project/rpy2/ (accessed August 2, 2020).

Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C., and Ester, M. (2019). MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* 35, i501–i509. doi: 10.1093/bioinformatics/btz318

Shiny (2020). *Shiny*. Available online at: https://shiny.rstudio.com/ (accessed August 2, 2020).

Singh, A., Shannon, C. P., Gautier, B., Rohart, F., Vacher, M., Tebbutt, S. J., et al. (2019). DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 35, 3055–3062. doi: 10.1093/bioinformatics/bty1054

Specht, H., Emmott, E., Petelski, A. A., Gray Huffman, R., Perlman, D. H., Serra, M., et al. (2019). Single-cell mass-spectrometry quantifies the emergence of macrophage heterogeneity. *bioRxiv* [Preprint]. doi: 10.1101/665307

Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinform. Biol. Insights* 14:117793221989905. doi: 10.1177/1177932219899051

Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., et al. (2016). Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* 44, D463–D470. doi: 10.1093/nar/gkv1042

Tarazona, S., Balzano-Nogueira, L., Gómez-Cabrero, D., Schmidt, A., Imhof, A., Hankemeier, T., et al. (2020). Harmonization of quality metrics and power calculation in multi-omic studies. *Nat. Commun.* 11;3092. doi: 10.1038/s41467-020-16937-8

The FANTOM Consortium and the RIKEN PMI and CLST (DGT)., Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., de Hoon, M. J., et al. (2014). A promoter-level mammalian expression atlas. *Nature* 507, 462–470. doi: 10.1038/nature13182

Thul, P. J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., et al. (2017). A subcellular map of the human proteome. *Science* 356:eaal3321. doi: 10.1126/science.aal3321

Ulfenborg, B. (2019). Vertical and horizontal integration of multi-omics data with miodin. *BMC Bioinformatics* 20:649. doi: 10.1186/s12859-019-3224-4

Vizcaíno, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., et al. (2016). 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* 44, D447–D456. doi: 10.1093/nar/gkv1145

Vizcaíno, J. A., Deutsch, E. W., Wang, R., Csordas, A., Reisinger, F., Ríos, D., et al. (2014). ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* 32, 223–226. doi: 10.1038/nbt.2839

Voilà (2020). *voilà*. Available online at: https://blog.jupyter.org/and-voilà-f6a2c08a4a93 (accessed August 3, 2020).

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Wang, M., Carver, J. J., Phelan, V. V., Sanchez, L. M., Garg, N., Peng, Y., et al. (2016). Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* 34, 828–837. doi: 10.1038/nbt.3597

Way, G. P., Zietz, M., Rubinetti, V., Himmelstein, D. S., and Greene, C. S. (2020). Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. *Genome Biol.* 21:109. doi: 10.1186/s13059-020-02021-3

Weber, L. M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P. P., et al. (2019). Essential guidelines for computational method benchmarking. *Genome Biol.* 20, 1–12. doi: 10.1186/s13059-019-1738-8

Welch, J. D., Kozareva, V., Ferreira, A., Vanderburg, C., Martin, C., and Macosko, E. Z. (2019). Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* 177, 1873.e–1887.e. doi: 10.1016/j.cell.2019.05.006

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1–9.

WIlsON (2020). WIlsON. Available online at: http://loosolab.mpi-bn.mpg.de/wilson/ (accessed August 2, 2020).

Wilson, C. M., Li, K., Yu, X., Kuan, P. F., and Wang, X. (2019). Multiple-kernel learning for genomic data mining and prediction. *BMC Bioinformatics* 20:426. doi: 10.1186/s12859-019-2992-1

Zeng, S., Lyu, Z., Narisetti, S. R. K., Xu, D., and Joshi, T. (2019). Knowledge base commons (KBCommons) v1.1: a universal framework for multi-omics data integration and biological discoveries. *BMC Genomics* 20:947. doi: 10.1186/s12864-019-6287-8

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., et al. (2011). International Cancer genome consortium data portal–a one-stop shop for cancer genomics data. *Database* 2011:bar026. doi: 10.1093/database/bar026

Conflict of Interest: VD currently works as a Post-Doctoral Researcher in Novo Nordisk Research Center Seattle, Inc. He did not receive any funding for this work. BBM works as a Computational Biologist in Enveda Therapeutics and did not receive any funding for this work. SS has no conflicts of interest. MK has no financial conflicts of interest, but he contributed to two of the discussed projects: rpy2 and Jupyter.

frontiers in Genetics

TECHNOLOGY AND CODE
published: 03 February 2021
doi: 10.3389/fgene.2021.624513

Accurate, Efficient and User-Friendly Mutation Calling and Sample Identification for TILLING Experiments

Juanita Gil[1†‡], Juan Sebastian Andrade-Martínez[2,3‡] and Jorge Duitama[1]*

[1] Systems and Computing Engineering Department, Universidad de Los Andes, Bogotá, Colombia, [2] Research Group on Computational Biology and Microbial Ecology, Department of Biological Sciences, Universidad de Los Andes, Bogotá, Colombia, [3] Max Planck Tandem Group in Computational Biology, Universidad de Los Andes, Bogotá, Colombia

OPEN ACCESS

Edited by:
Joanna Jankowicz-Cieslak, International Atomic Energy Agency, Austria

Reviewed by:
Katarzyna Gajek, University of Silesia, Poland
Prateek Gupta, Hebrew University of Jerusalem, Israel

***Correspondence:**
Jorge Duitama
ja.duitama@uniandes.edu.co

†Present address:
Juanita Gil
Department of Entomology and Plant Pathology, University of Arkansas, Fayetteville, AR, United States

‡*These authors share first authorship*

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

Received: *31 October 2020*
Accepted: *08 January 2021*
Published: *03 February 2021*

Citation:
Gil J, Andrade-Martínez JS and Duitama J (2021) Accurate, Efficient and User-Friendly Mutation Calling and Sample Identification for TILLING Experiments. Front. Genet. 12:624513. doi: 10.3389/fgene.2021.624513

TILLING (Targeting Induced Local Lesions IN Genomes) is a powerful reverse genetics method in plant functional genomics and breeding to identify mutagenized individuals with improved behavior for a trait of interest. Pooled high throughput sequencing (HTS) of the targeted genes allows efficient identification and sample assignment of variants within genes of interest in hundreds of individuals. Although TILLING has been used successfully in different crops and even applied to natural populations, one of the main issues for a successful TILLING experiment is that most currently available bioinformatics tools for variant detection are not designed to identify mutations with low frequencies in pooled samples or to perform sample identification from variants identified in overlapping pools. Our research group maintains the Next Generation Sequencing Experience Platform (NGSEP), an open source solution for analysis of HTS data. In this manuscript, we present three novel components within NGSEP to facilitate the design and analysis of TILLING experiments: a pooled variants detector, a sample identifier from variants detected in overlapping pools and a simulator of TILLING experiments. A new implementation of the NGSEP calling model for variant detection allows accurate detection of low frequency mutations within pools. The samples identifier implements the process to triangulate the mutations called within overlapping pools in order to assign mutations to single individuals whenever possible. Finally, we developed a complete simulator of TILLING experiments to enable benchmarking of different tools and to facilitate the design of experimental alternatives varying the number of pools and individuals per pool. Simulation experiments based on genes from the common bean genome indicate that NGSEP provides similar accuracy and better efficiency than other tools to perform pooled variants detection. To the best of our knowledge, NGSEP is currently the only tool that generates individual assignments of the mutations discovered from the pooled data. We expect that this development will be of great use for different groups implementing TILLING as an alternative for plant breeding and even to research groups performing pooled sequencing for other applications.

Keywords: software, mutagenesis, functional genomics, TILLING, variants detection

INTRODUCTION

Targeting Induced Local Lesions in Genomes (TILLING) is a powerful reverse genetics method used in plant sciences which allows the identification of point mutations or SNPs, introduced randomly throughout the whole genome by chemical mutagenesis (Missirian et al., 2011). In brief, TILLING consists of mutagenesis, DNA extraction and pooling of several individuals of a population, PCR amplification of regions of interest, and high-throughput mutation discovery in target genes (McCallum et al., 2000). Despite newer technologies being available for targeted modification of genes such as CRISPR-Cas, TILLING remains a useful and effective functional genomics tool for studying genes responsible for desired phenotypes because large populations can be screened for mutations before bringing plants to the field, thus reducing phenotyping costs, and it generates genome-wide mutations allowing to target in multiple genes at the same time (Irshad et al., 2020). With the advance in high-throughput sequencing technologies and their current lower costs, TILLING by Sequencing proves to be the best choice for the identification of mutations and the corresponding mutant individuals in pooled samples, and for linking the identified base pair changes with their impact on specific traits (Tsai et al., 2011).

The application of bioinformatic tools contributes to virtually all elements of the TILLING pipeline, including identification of the genes in the species of interest, amplicon design, and analysis of the effect of produced mutations in protein products (Kurowska et al., 2011). The biggest bioinformatic challenge in TILLING is variant calling in multidimensional experiments. In essence, an efficient pipeline for detection must not only call variants in each pool but triangulate the outputs per pool to identify true variants and determine the individual carrying each mutation based on the specific pooling design (Missirian et al., 2011). Moreover, mutations produced through TILLING are rare within the population. Hence, special efforts must be taken to distinguish true variants from noise (Missirian et al., 2011). While some of the available tools for variant calling are able to detect variants in pooled samples (Huang et al., 2015), they are not designed toward the posterior triangulation of the variants detected from each individual pool. Moreover, most tools require high coverages and high sequencing qualities to achieve good accuracy (Missirian et al., 2011). Accuracy and efficiency vary amply between software tools (Huang et al., 2015). As of today, the only available tool specifically designed for variant calling in TILLING experiments is CAMBa, which employs Bayesian statistics for yielding the most probable mutations in a TILLING experiment per individual (Missirian et al., 2011).

Since the advent of Next-Generation Sequencing (NGS), it has been proposed that TILLING procedures could eventually be carried out totally *in silico* (Wang et al., 2012; Chen et al., 2014). As mentioned above, tools have been developed in the past for *in silico* identification of candidate genes, such as CODDLE (Slota et al., 2017), for analysis of the effects of putative or detected mutations in TILLING populations such as PARSESNP (Taylor and Greene, 2003), SAS (Milburn et al., 1998), or SOPMA (Geourjon and Deleage, 1995), and for variant detection, such as CAMBa and our own implementation. Nonetheless, to the best of our knowledge there is no tool available for simulation of NGS pool-sequencing in the context of multidimensional TILLING experiments. This tool would be critical both for potential fully *in silico* TILLING experiments, as well as for guiding the design of *in vivo* procedures.

The development of new bioinformatic tools to increase the precision of TILLING experiments is crucial, especially when considering TILLING branching applications. Moreover, pooled sequencing for variant discovery is used in other protocols related to crop breeding and even in distant fields such as the study of rare human genetic diseases. Pooled sequencing followed by the identification of *de novo* variants has facilitated the typing of a larger number of donors for stem cell transplants at the same time, increasing the chance of finding a good match for recipient patients (Lange et al., 2014) and can also improve diagnostics rates of genetic disorders by increasing the number of probands tested at a time at a reduced cost (Dashnow et al., 2019). In the context of plant breeding, introducing natural or artificial allelic diversity in crops is widely used to develop new varieties with improved traits that meet the current global demands for food production. Some examples are kernel hardness in wheat (Ma et al., 2017), drought tolerance (Yu et al., 2012), and starch quality (Raja et al., 2017) in rice, seed weight in chickpea (Bajaj et al., 2016), and starch biosynthesis and herbicide tolerance in cassava (Duitama et al., 2017).

We have developed, through the Next Generation Sequencing Experience Platform (NGSEP), two new functionalities for TILLING analyses: a TILLING experiment simulator and a TILLING detector. The simulator is able to generate pool reads derived from any set of genomic sequences, creating an *in silico* population for the experiment with associated variants assigned to specific individuals. The detector leverages NGSEP variant detection to first call variants per pool, which are then triangulated to perform identification of the individuals associated with the discovered mutations.

RESULTS

Novel Functionalities for Simulation and Read Analysis in TILLING Experiments

In a TILLING experiment, a mutagenic agent is used to treat the seeds and induce random mutations across the entire genome of a particular organism. One of the most commonly used agents is ethyl methanesulfonate (EMS), which induces 2 to 10 mutations/Mb of diploid DNA (Henry et al., 2014). Mutagenized TILLING populations are analyzed for the identification of the mutations generated across the individuals of the population. If sequencing occurs after one round of selfing (usually called generation M2), about half of the mutations are heterozygous in the population. Although it is technically possible to sequence independently and call variants on each individual of the population, this procedure is not cost effective given that most of the individuals will not carry interesting mutations and promising individuals usually go over further rounds of selfing to stabilize the mutation and its potential phenotypic effect. Hence, the TILLING by sequencing

design suggests a tridimensional pooled strategy in which each individual is included in a unique combination of three different pools, one per dimension (**Figure 1**). Pools are then sequenced and mutations are identified in the pools. Taking into account the mutation rate, it is very unlikely that two individuals carry exactly the same mutation. Thus, individual assignments can be performed looking for mutations consistently called in three pools of different dimensions. This design allows to perform mutation detection and individual assignment for hundreds of individuals sequencing only the sum of the pools generated for each dimension.

One of the main aspects to take into account for a TILLING experiment is the design of the number of pools to include in each dimension and the number of individuals per pool. To provide a tool to explore *in silico* the behavior of a TILLING experiment in different scenarios, we developed a simulator of TILLING experiments based on a set of target regions from a reference genome, a given population size, a mutation rate, and a design of overlapping pools. Based on this information, the simulator follows *in silico* the steps of mutagenesis, sample pooling, and sequencing. Regions to be amplified for each individual in the population were created first as an exact copy of the reference and then mutations were assigned randomly to each individual at a random position and to a random base pair, distinct from the reference. According to the pooling design, each individual was assigned to a row, column and plate pool. Given the sensitivity limitation of the variant calling process of mutations occurring at low frequencies, the smaller the population and the number of samples per pool, the higher the probability of calling true mutations. Therefore, we proposed an experimental design of overlapping amplicons per target gene for a population of 288 individuals. By having pools of maximum 48 individuals (96 haplotypes) we reduced the noise caused by the simulated and expected sequencing errors.

Paired-end high throughput sequencing (HTS) reads were simulated for each pool from the *in silico* mutated amplicon sequences. Mimicking the actual sequencing process and the known error rate patterns of Illumina, a read was generated for each pool selecting a random amplicon within the pool. A forward and a reverse read of a given length were then simulated starting from each end of the selected amplicon. Given a minimum and maximum error rate, the simulator generates substitution errors at random according to a stepwise distribution which starts from the minimum error rate at the 5' end and ends with the maximum error rate at the 3' end of the read.

We tested the performance of our simulator by recording the time and memory spent during different simulations varying the number of individuals of the population, dimensions of the pooling design, read lengths, and sequencing depths (**Table 1**). In all cases, the simulator ran in less than 2 h, and in less than 1 h for all cases of 50X and 10X coverage. In general, time is affected mostly by the number of reads, with simulations of similar coverages running faster with longer read lengths. Memory requirements did not exceed 3 GB in any case. This factor was mainly determined by the size of the simulated population.

We also developed modifications of the core algorithm for variants detection available in NGSEP and a new functionality to perform the specific analysis of HTS data required by TILLING experiments. From the algorithmic perspective, the discovery of mutations within each pool is the most challenging part of the analysis because mutations are expected to be carried by one or at most two haplotypes within each pool. Hence, the variants discovery module should be able to separate true variants with allele frequencies of one divided by the number of haplotypes in a pool from sequencing errors. As detailed in the next section, we modified the Bayesian model implemented in NGSEP to identify mutations in these circumstances. Once mutations are identified, and taking into account the pooling strategy, the main outcome of a TILLING experiment should not only be the identification of true mutations but the identification of individuals carrying these mutations. Hence, we developed a module that receives the individual VCF files with variants called within each pool and a text file with the configuration of samples included in each pool, and performs the individual sample genotyping of the mutations (also called triangulation). Taking into account that each sample is included in a unique combination of pools, a variant is assigned to a sample if and only if it is called in all pools in which the sample was included. The triangulation module traverses in parallel the pool VCF files and, for each mutation identified in three pools of different dimensions, queries the pool configuration information to determine which individual is present in the three pools and assigns the mutation to such individual. The output of this process is a VCF file with one column per individual, which in simulation experiments can be directly compared with the VCF gold standard file produced by the simulator.

Variant Detection and Genotyping in Polyploid Individuals and Pools

We modified the core module of NGSEP (the variants detector) with two related goals: to improve the accuracy of variant calling in polyploid individuals and to allow identification of variants at different allele frequencies in pooled samples. First, sites in which at least one allele different from the reference is observed with a count at least $0.5/a$ were identified, where a is the total number of haplotypes in the sample. For a pool of n individuals with ploidy p, the total number of haplotypes a would correspond to n^*p. For each selected site, the algorithm calculates the conditional probability of the data assuming a homozygous genotype for the allele with the highest read count and the conditional probabilities of the data assuming each possible heterozygous allele dosage for the allele with the second read count, from $1/a$ to 0.5. Both the homozygous genotype and the heterozygous genotypes can be encoded as m copies of a major allele G_1 and $a–m$ copies of a minor allele G_2, where $m \geq 0.5^*a$. A value of $m = a$ would correspond to a homozygous genotype. Similar to the case of a single diploid individual, given a pileup position of the genome and the set R of reads spanning that position, the conditional probability of R given the genotype $G = {G_1}^m\ {G_2}^{a-m}$ can be calculated as the product of the conditional probability of each read $r \in R$ given G. Calling b the base pair of r spanning

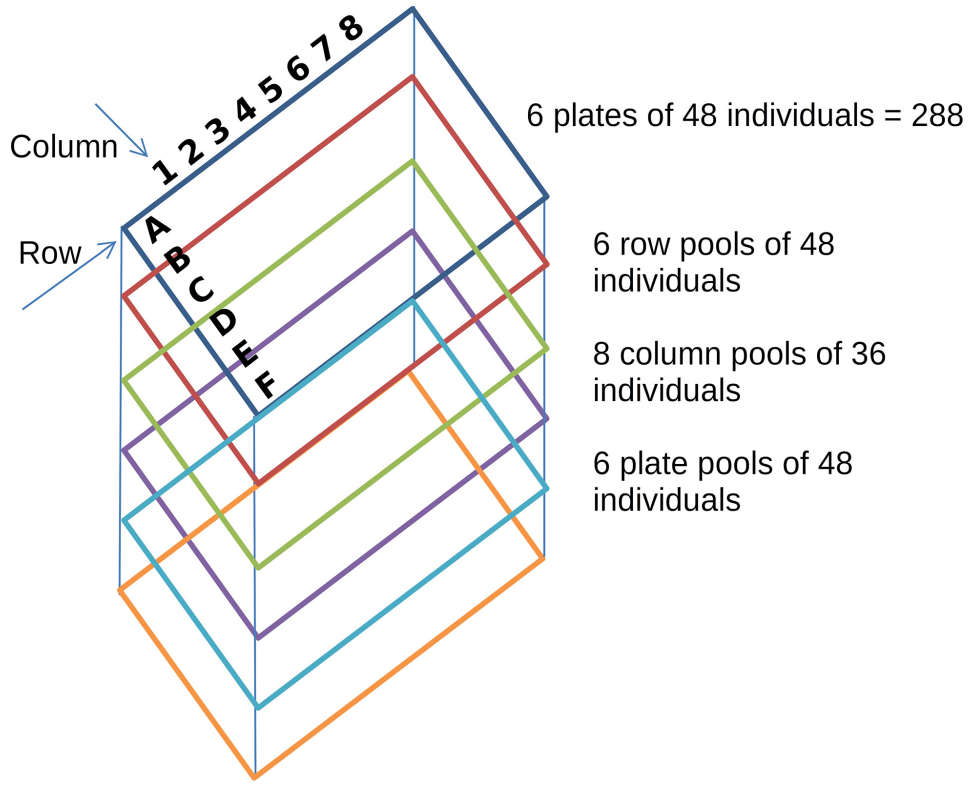

FIGURE 1 | Pooling strategy in a simulated TILLING experiment. Each colored rectangle represents a plate of 6 × 8 wells with 48 individuals; rows are represented by the letters A to F and columns by the numbers 1 to 8. The tridimensional pooling strategy results in 20 pools. Row pools consist of individuals of each single row of each plate (48 individuals). Column pools consist of individuals of each single column of each plate (36 individuals) and plate pools consist of the individuals of each plate.

TABLE 1 | Running times (in seconds) and memory requirements (in MB) of three different TILLING experiment simulations ran by the simulator: 8 × 12 row by column plates for 800 individuals, 8 × 8 plates for 800 individuals, and 6 × 8 plates for 288 individuals with a total of 300 mutations within the population.

Dimension (row × column × plate)	Population size	Read length (bp)	Time (s)			Memory (MB)		
			Depth			Depth		
			10X	50X	100X	10X	50X	100X
8 × 12 × 9	800	200	404.04	1933.18	3961.07	2215.70	2215.70	2211.86
		100	638.88	3315.90	6883.64	2215.70	2215.70	2211.84
8 × 8 × 13	800	200	281.81	1283.45	2660.51	2215.70	2215.70	2215.70
		100	498.09	2389.13	4422.11	2215.69	2215.70	2211.84
6 × 8 × 6	288	200	146.75	721.99	1325.39	2139.41	2139.42	2139.46
		100	245.22	1182.25	2509.16	2139.40	2139.42	2139.42

For each case, coverage of 10X, 50X, and 100X was tested, for 100 and 200 bp read length.

the analyzed position, e its error probability and $f = m/a$ the frequency of the major allele, the conditional probability $P(b|\ G)$ is given by this formula:

$$P\left(b|G_1^m G_2^{a-m}\right) = \begin{cases} 1-e, a = m \wedge b = G_1 \\ \frac{e}{3}, b \neq G_1 \wedge b \neq G_2 \\ f(1-e) + \frac{(1-f)e}{3}, a < m \wedge G_1 = b \\ (1-f)(1-e) + \frac{fe}{3}, a < m \wedge G_2 = b \end{cases}$$

Similar to the case with diploid individuals, a prior probability $P(G)$ can be calculated from previous knowledge on heterozygosity rate. We set a non-informed prior in our experiments with simulated and real data.

Comparison of Variant Calling Tools

Comparison of the performance of different variant calling tools was carried out based on the simulated sequences. As observed in the simulations and real data, the sequencing error rate

becomes the most critical factor to determine the number of total haplotypes (and by extension samples) that can be included within each pool to be able to separate true mutations from sequencing errors. Given an average sequencing error rate of 0.5% we could achieve good accuracy with up to 64 diploid individuals (128 haplotypes) per pool. Hence, we present here the results of the simulation experiment with 288 individuals arrayed in 6 plates of 6 × 8 rows by column set up (**Figure 1**).

The results of the variant detection step obtained with our algorithm were compared with the results obtained from other tools frequently used for variant detection, such as GATK haplotype caller (McKenna et al., 2010) and Freebayes (Garrison and Marth, 2012), as well as tools designed to identify low frequency variants like Lofreq (Wilm et al., 2012), or to identify variants in pools like CRISP (Bansal, 2010), and SNVer (Wei et al., 2011). Freebayes was the only tool that did not identify variants in any of the pools of the simulation experiments and was not considered for further comparisons. Given that the output VCF file generated by SNVer is outdated and could not be modified to run the comparisons, this tool was also discarded for comparison purposes. We were unable to run CAMBa (Missirian et al., 2011) by ourselves nor received a response after trying to contact the developers, so we omitted said tool.

We compared the tools in terms of their sensitivity, expressed as the number of true positives divided by the sum of true positive and false negative values. For each of the 20 pools, sensitivity was calculated and compared between the four selected variant callers and for each experiment varying the read depth (**Figure 2A**). We also tried to calculate specificity but it was 100% in all cases. CRISP consistently showed the lowest sensitivity among the tools and read depths. Lofreq showed improved performance with increasing read depth, showing the best results of all tools at a coverage of 100X, but the worst sensitivity at a coverage of 10X. GATK and NGSEP both showed consistent high sensitivities at all read depths. While GATK shows slightly higher sensitivities than NGSEP at 10X and 50X of coverage, NGSEP performs slightly better than GATK at 100X coverage. We also compared sensitivities in randomly selected pools by varying the total number of haplotypes or ploidy (**Figures 2B,C**). At low coverage (10X) Lofreq is the worst performing tool regardless of this number. However, it performs two times better calling variants in pools with less haplotypes. This is evident by the sensitivity drop from 57% in the pool with 72 haplotypes (36 individuals) to 23% in the pool with 96 haplotypes (48 individuals). However, this tool outperforms CRISP and GATK in these two particular pools selected for comparison at higher coverages (50X and 100X). Smaller variations in sensitivity were observed in the other three tools when comparing specific pools with two different ploidies at different read depth, with NGSEP and GATK showing the most similar sensitivity values between both samples across the coverage range.

We used our new functionality to identify the individuals carrying mutations in a simulated population and compared the ability of each variant caller to successfully call variants in overlapping pools. Sensitivity was determined as the total number of SNPs identified over the total number of SNPs that should have been detected corresponding to the simulated mutations (**Figure 3**). The sensitivity of Lofreq was zero at the lowest simulated sequence coverage. This means that although the tool is able to call variants in every pool as shown in **Figure 2**, those are not found in the three pools that overlap and the SNP cannot be assigned to any individual. Nevertheless, its performance improved with increasing coverage calling between 80 and 93% of the mutations that correspond to one single individual. CRISP showed the poorest performance among the four compared tools at read depths of 50X and 100X. It reached, however, a sensitivity equal to and above 80% at the two highest read depths, respectively. NGSEP and GATK are the best performing tools regardless of read depth. However, with increasing read depth both tools showed higher sensitivities, reaching 90.6 and 90.3% at 100X, respectively. Although sequencing depth improves the sensitivity of the tools, both NGSEP and GATK can detect around 75% of low frequency SNPs in a mutated population and those can be correctly assigned to mutated individuals at a read depth as low as 10X.

Running times spent by each tool were compared to further assess the performance of the variant calling process (**Figure 4**). NGSEP was the most efficient tool even at the highest sequencing read depth. GATK was the slowest of all tools taking up to 12 h (∼40,000 s) to call variants in 20 pool samples of a population of 288 individuals, while the other tools required a maximum of 1.5 h to perform the same job. NGSEP showed the most steady time performance over increasing read depths. Variant calling in the simulated experimental setup only took 6.2 min at 10X coverage, 12.3 min at 50X and 21.2 min at 100X when running NGSEP on a laptop.

Analysis of a Rice TILLING Population

We used the publicly available data of a sequencing experiment of 44 pools from a rice TILLING population comprising 768 individuals and 32 gene fragments that added up to 42,034 bp. Tsai et al., 2011 reported 122 mutations in overlapping pools detected with the tool CAMBa (Missirian et al., 2011) in this dataset. We identified 262 biallelic SNVs in those pools using NGSEP, 1,852 with GATK, 751 with Lofreq and 0 with CRISP. Despite having an acceptable to good performance on simulated data, none of the SNPs called by CRISP passed the quality filter using the rice population and the VCF files could not be used for the triangulation process in which mutations are assigned to individuals. We calculated exact genomic positions for all SNPs reported by Tsai et al., 2011 to assess if the SNPs detected by NGSEP, GATK, and Lofreq corresponded to the previously reported mutations including the same expected effect on the corresponding gene based on the annotation of the VCF files (**Supplementary Table 1**). The three variant calling tools used to test the new triangulation function of NGSEP identified more than 122 mutations. NGSEP reported 262 mutations (**Supplementary Table 2**), GATK 1,852 (**Supplementary Table 3**), and Lofreq 751 (**Supplementary Table 4**). We compared the results according to the number of variants detected by each tool (NGSEP, GATK, and Lofreq in this study and CAMBa in the previous study) and the type of variant (or predicted effect) on the sequenced gene (**Figure 5A**). The most common type of variant was intronic variants with

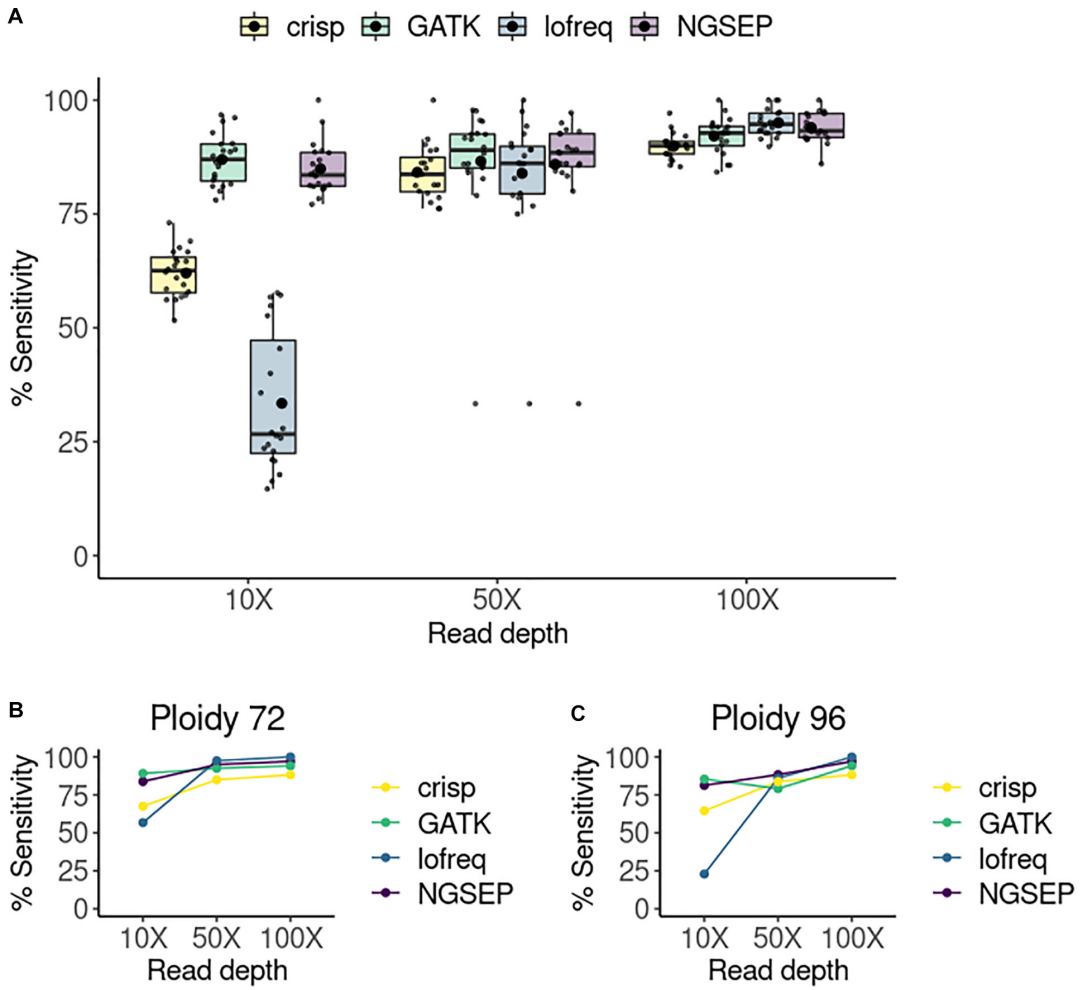

FIGURE 2 | Sensitivity of variant calling per pool. **(A)** Sensitivity of variant detection in a simulated TILLING population comprising 288 individuals sequenced in 20 pools. The small points represent each single pool. The box plots represent the median and first and third quartiles of the sensitivity in all pools. The bigger black point shows the average sensitivity of all pools per tool and read depth. **(B)** Sensitivity of variants detected in a randomly selected pool with 48 individuals (ploidy equal to 96 haplotypes) at three different read depths. **(C)** Sensitivity of variants detected in a randomly selected column pool with 36 individuals (ploidy equal to 72 haplotypes).

90 mutations being identified by NGSEP, 577 by GATK, 279 by Lofreq, and 40 by CAMBa. Missense and synonymous variants were the second and third most common type of variants identified by all tools. The least frequent type of variant was mutations leading to a stop codon. Regarding the type of SNVs identified in the mutated rice samples, the most common were *G* to *A* and *C* to *T* transitions according to the results obtained with NGSEP (32.4%) and CAMBa (66.9%). Conversely, AT to GC transitions were the most frequent mutation type for GATK (89.15%) and Lofreq (65.9%). These transitions were found only in 27.9% of the mutations reported by NGSEP and 19% of the mutations reported by CAMBa. In accordance with Tsai et al., 2011 *G* to *C* or *C* to *G* transversions were the least common (<5.2% of all mutations for all tools). Overall, there were more transitions than transversions (**Supplementary Table 5**). Although NGSEP reported the highest percentage of transversions (35.11%), looking at the number of pools where these transversions are called, we found that transversions were called in more pools than transitions (**Supplementary Table 2**). Applying a filter keeping only variants called in at most six pools, the percentage of GC > AT transitions increased to 47.13% for NGSEP and the percentage of transversions reduced to 16.09%. GATK and Lofreq also reduced the percentage of transversions after this filter but preserved the excess of AT > GC transitions, compared to the results originally reported using CAMBa.

In the previous study conducted by Tsai et al., 2011, the mutations were categorized into homozygous, heterozygous, implausible or false based on validation experiments, and not tested for mutations that were not validated. We used these categories to analyze how many of the mutations were found in exactly three, more than three or less than three pools within each category using NGSEP, GATK, and Lofreq (**Figure 5B**). We found that within the validated mutations more than 67% of them were assigned to exactly three pools by each tool,

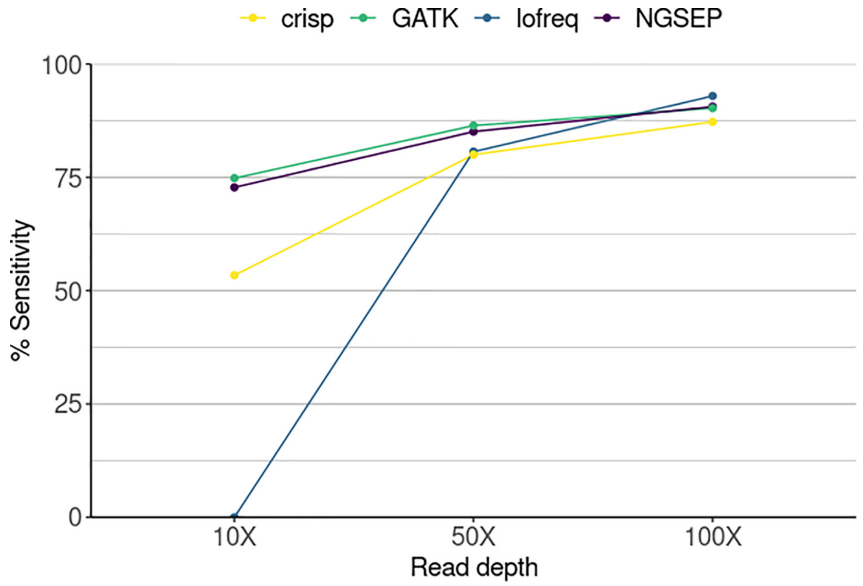

FIGURE 3 | Sensitivity in the identification of mutant individuals in a simulated TILLING population comprising 288 individuals sequenced in 20 pools and three individual experiments varying the sequencing read depth. Sensitivity corresponds here to the number of SNPs detected by each tool divided by the total number of SNPs in a gold standard.

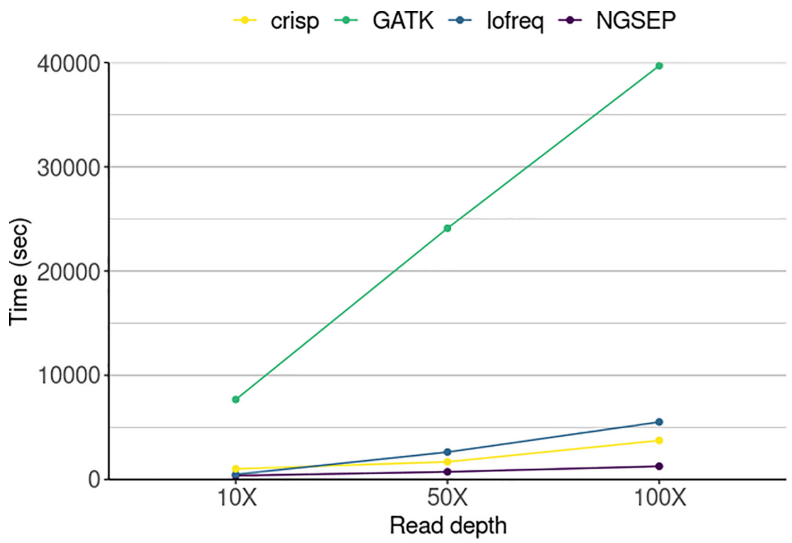

FIGURE 4 | Running time of four variant detector tools for a simulated TILLING population comprising 288 individuals sequenced in 20 pools in three different experiments varying sequencing read depth. The time represents the amount of seconds used by each tool to call variants in 20 pools.

which is the expectation for tridimensional pooling strategies. In contrast, within the category containing implausible or false positive mutations, less than 20% of the mutations were assigned to exactly three pools, and found principally in less than three pools (red bars in **Figure 5B**). Within the not tested category, 75, 71.9, and 68.8% of the mutations were assigned to exactly three pools for NGSEP, GATK and Lofreq, respectively, and could potentially be true mutations in the population. Considering the number of variants assigned to a number of pools different than three, GATK reported 12 validated variants in more than three pools, whereas this number was only three for NGSEP and Lofreq. Conversely, validated variants called in less than 3 pools were only 5 for GATK, whereas this number was 15 and 17 for NGSEP and Lofreq, respectively. This behavior is consistent for the non tested variants and reflects that GATK predicted a much larger overall number of mutations.

To further validate the performance of the new functionalities in NGSEP using real data, we selected two of the genes for which mutations have also been reported elsewhere and their effect has been described. The inositol kinase-like gene Os09g34300

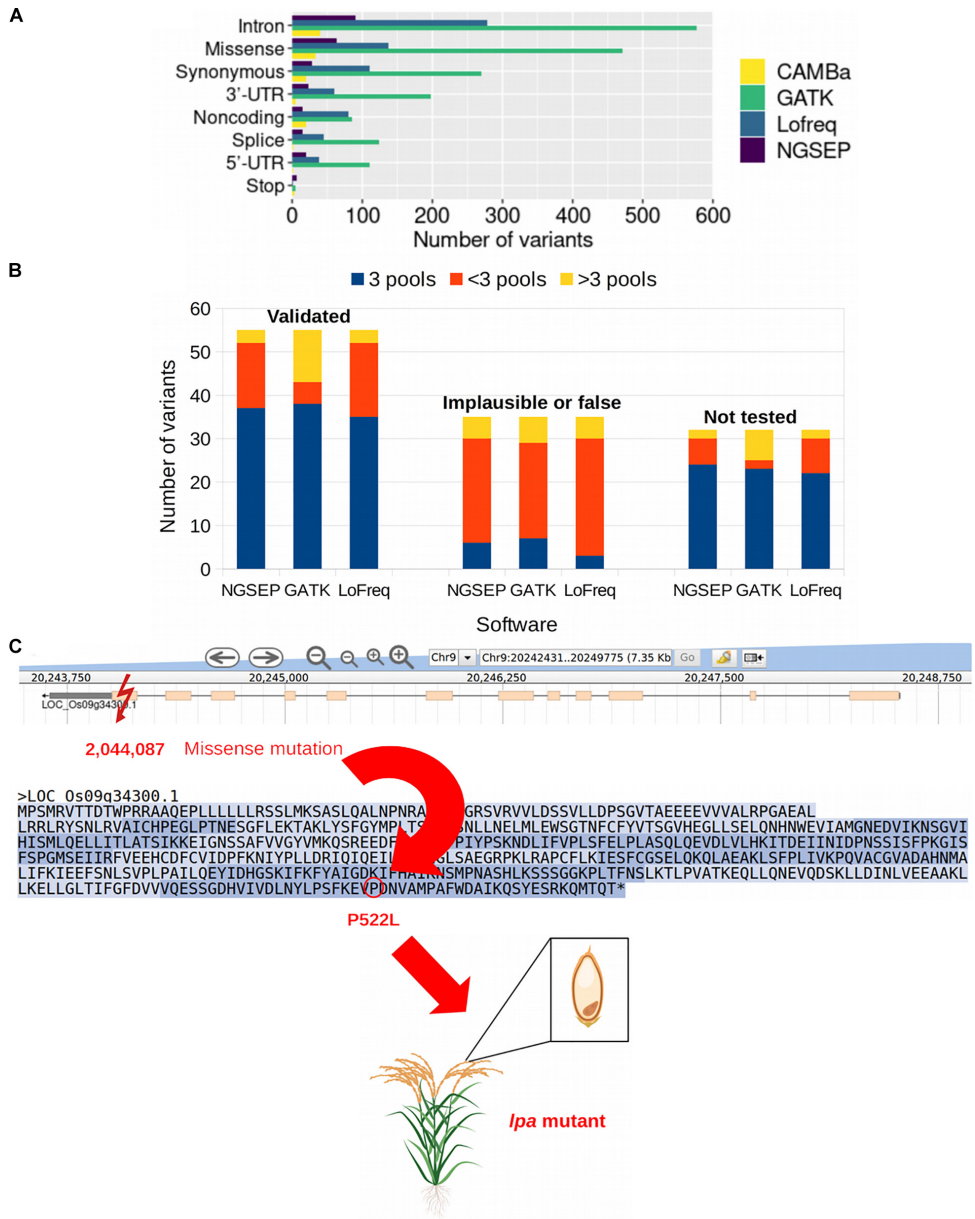

FIGURE 5 | Mutations in a rice TILLING population. **(A)** Number of variants identified in a rice TILLING population comprised by 768 individuals categorized by the type of variant. NGSEP, GATK, and Lofreq were the variant caller used in this study and variants reported by CAMBa were obtained from Tsai et al. (2011). **(B)** Mutations called in pools by three software, NGSEP, GATK, and Lofreq, classified into three categories based on results by Tsai et al. (2011): Validated, which are verified mutations; implausible or false, which are mutations found in non overlapping pools or are false positives; and not tested, which are not verified mutations. Within the category of validated mutations, and possibly also within the not tested category, most of the mutations are expected to be assigned to three pools (blue bars). In the implausible or false category most of the mutations should be found in <3 pools or >3 pools (red and yellow bars, respectively). **(C)** Example of a missense mutation in gene Os09g34300. The mutation at position 2,044,087 on chromosome 9 results in the amino acid substitution from proline to leucine amino acid at position 522 in the protein, leading to reduced phytic acid (lpa) content in the grains of the mutant plants with no negative effect on grain weight or delayed seedling growth (Kim and Tai, 2014).

and the multidrug resistance-associated gene Os03g04920 are both involved in the reduction of phytic acid (myo-inositol 1,2,3,4,5,6-hexakisphosphate) in rice seeds. Phytic acid is considered an antinutrient because humans and other non-ruminants are unable to efficiently digest it and it prevents the absorption of important micronutrients in their intestines (Perera et al., 2019). Four novel mutations obtained by TILLING were reported by Kim and Tai (2014) within these two genes, obtaining four low phytic acid (lpa) mutants, two of which were similar to wild-type plants in seed weight, germination, and seedling growth. One missense mutation in the last exon of the gene Os09g34300 leading to the amino acid change

P522L was identified by NGSEP, GATK, and Lofreq as well. This mutation was also reported and validated in (Tsai et al., 2011) and confirmed as a lpa mutant by (Kim and Tai, 2014) in the laboratory, being a promising line for breeding in rice programs aiming at developing varieties with improved nutritional quality. Within this same gene, an intronic and a splice variant not tested in the study of (Tsai et al., 2011) were also identified using the three variant callers. The position and effect of the mutation in the inositol kinase-like gene are schematically shown in **Figure 5C**. Both of the reported mutations by (Kim and Tai, 2014) within gene Os03g04920 that led to a lpa phenotype and were also included in the list of mutations by (Tsai et al., 2011) were identified by NGSEP, GATK and Lofreq as well. Furthermore, from the seven validated mutations within this gene, NGSEP, GATK, and LoFreq called five of them in exactly three pools. From the six non tested variants, NGSEP and GATK called five and Lofreq called four in exactly three pools.

DISCUSSION

Mutagenesis is a widely used experimental technique in functional genomics because it allows to generate genetic variability not present in natural populations in an unbiased manner. The TILLING experimental setup reduces the cost needed to identify mutations in candidate genes across populations developed through mutagenesis and to perform the identification of mutated individuals (Kurowska et al., 2011). Functional effects of identified mutations can then be further investigated using protein modeling or even through more directed approaches such as CRISPR, besides the observation of the expected phenotype. Although recent technologies have been developed for targeted genome editing, including CRISPR, the resulting organism is usually considered a genetically modified organism (GMO) presenting an important problem and limitation in plant breeding for the release of improved varieties into the markets of countries with strict regulations about GMOs. Mutagenesis, however, has been considered a safe method to rapidly induce genetic variation and develop improved varieties, that are not regulated by the GMO legislations (Holme et al., 2019). Moreover, since TILLING samples come from a population whose individuals (or their seeds) are readily available for the researcher (Wang et al., 2012), individual identification is particularly useful because it allows to perform validation of the potential phenotype differences generated by the identified mutations in the associated individuals, providing valuable information for future plant breeding.

In this work, we presented new functionalities of NGSEP to facilitate the data analysis steps required to obtain the expected information from a TILLING experiment. First, we developed an improved model to perform accurate variant identification in pooled samples, which is useful for different applications of HTS. Either in mutagenized or in natural populations, variants could be quickly identified by bulk sequencing of large numbers of individuals to avoid the costs of sample by sample barcoding and library preparation. Germplasm banks are using pooled sequencing to validate genetic stability of accessions avoiding the cost of individual sequencing of potential clones (see for example Rubinstein et al., 2019). Moreover, the same underlying model to perform pooled genotyping might be useful to perform individual genotyping in species with high ploidy such as sugar cane, where genotyping by sequencing is preferred over SNP arrays for variant detection (Manimekalai et al., 2020). We show through simulation experiments that our model has comparative accuracy and better efficiency compared to other solutions. NGSEP showed consistently high sensitivities (above 80%) across varying read depths (10X, 50X, and 100X) in variants called in individual pools in simulated data. GATK also showed good sensitivities in pool variant calling at different read depths. Conversely, Lofreq and CRISP only showed good performances (sensitivity $>$ 75%) at higher read depths (50X and 100X). Moreover, we show how different tools for regular variant identification should be adapted to increase sensitivity in pooled data. This is very important for mutagenesis experiments because identification of mutations present in only one haplotype of the pool is the most challenging case of pooled variant identification, with different variant callers achieving different performances depending on sequencing depth (Huang et al., 2015). Hence, researchers struggle trying to adapt individual genotyping tools to experimental setups including pooled sequencing.

Particularly for TILLING, we developed a functionality to perform individual assignment of variants from the information of individuals included in each pool. To the best of our knowledge, NGSEP is currently the only open source software able to perform this step of the analysis process. Moreover, both the variant identification and the individual assignment can be executed from the graphical interface of NGSEP. Finally, we also built a functionality to perform simulations of TILLING experiments. Besides being useful to perform benchmarking of current and future analysis pipelines developed by different research groups, the simulator can also be used to validate the effectiveness of different pool configurations to achieve the goals of the experiment, saving time and money in *in vivo* analyses. This is particularly important given that preparation of populations for TILLING analysis is a long and costly process (Wang et al., 2012), so *in silico* experiments can help to make large-scale TILLING procedures more cost-effective.

We analyzed a large rice mutant population for high throughput mutation identification using the approach of TILLING by sequencing and a tridimensional pooling strategy. Using the publicly available sequences from a previous study by Tsai et al. (2011) we compared the performance of four variant calling tools, NGSEP, GATK, Lofreq, and CRISP. CRISP was discarded from the comparisons using real data because the SNPs called in pools did not pass the quality filter. In the previous study, the authors reported 122 mutations in the population using the tool CAMBa, developed by the same group. After filtering by number of pools, with NGSEP we identified 87 mutations, which was the second closest result compared to the previous report. On the other hand, GATK and Lofreq identified 569 and 127 mutations, respectively. Considering the small fragment of the genome that was targeted during the TILLING experiment, the number of mutations reported by GATK would represent an unexpectedly high mutation rate for

the mutagenesis experiment (Till et al., 2007). Moreover, both GATK and Lofreq reported an excess of AT > GC transitions which was not observed in the results reported by NGSEP and CAMBa. The analyzed rice population was treated by EMS mutagenesis, which has a G-alkylating action favoring primarily GC to AT base pair transitions. This corresponds to the result obtained with NGSEP and the previous report from CAMBa. The raw output of NGSEP showed a large percentage of transversions (39.7%), while the other tools reported less than 22%. However, most of the transversions were easy to filter out because they are found in a large number of pools, which is not expected in a TILLING experiment due to the low probability of finding a given mutation in more than one individual. Possible explanations for these variants are natural variation between the parent and the reference genome or systematic errors producing consistent false positive calls among pools.

From the set of validated mutations of the study carried out by Tsai et al. (2011), NGSEP, GATK, and Lofreq detected 67.3, 69.1, and 63.6% of them. Assuming a 100% success rate in the verification experiment, the performance of these tools in terms of sensitivity is lower using real data than those obtained using the simulated data of an artificial mutant population. Nevertheless, considering that the tools called mutations in three overlapping pools in less than 20% of the cases of implausible or false positive variants (again assuming this classification is 100% accurate) and that up to 75% of the not tested mutations in the study by (Tsai et al., 2011) were called and identified in three overlapping pools, the three tools show promising results for the analysis of large TILLING populations. From these three tools, NGSEP is the only one that offers the functionality of identifying mutations in overlapping pools and assigning mutations to the corresponding mutant individual in the population. Regarding computational efficiency, NGSEP was the most efficient tool, calling variants in all 44 pools of the rice TILLING population comprising 768 individuals with average sequencing coverage per pool ranging from 300X to 31,500X with the computational resources of a laptop and in less than 2.5 h.

With an ever growing population, the demand for food is increasing around the world. However, to increase crop productivity in a timely manner as required by the necessity of meeting the current global demands, it is critical to explore all possible alternatives to develop plants that are higher yielding and more resilient to climatic changes and their associated problems such as the raise of different pests and diseases or variable abiotic stresses such as drought, higher temperatures, and flooding, among others. We expect that the new developments presented in this manuscript will be useful for researchers implementing TILLING and other experimental techniques for functional genomics and breeding.

METHODS

Software Development and Implementation Details

We implemented the TILLING simulator and the functionality to perform individual assignments of discovered variants (also called triangulation process) as new functionalities of NGSEP. This allows to have these new functionalities integrated in the same software solution implementing the variant discovery step. Hence, the software is implemented in Java, following an object oriented design. The algorithm to perform variant discovery in pools was implemented within the general functionalities of NGSEP to perform single sample and multisample variants discovery. The new developed algorithm is activated when the number of haplotypes in the pool is provided in the "ploidy" option of these two functionalities. The three functionalities, namely simulation, variant calling, and triangulation, can be executed either from the command line or from the graphical interface of NGSEP v4, built in JavaFX (manuscript in preparation). NGSEP is distributed as an open source software solution available in http://ngsep.sf.net.

In the simulation process, given the pool dimensions selected by the user, the simulator will assign pools to each individual distributing the samples in the plates (wells) from left to right and from the top to the bottom, starting from the first plate to the last one. Depending on the number of individuals and plate size, some pools might contain less individuals than those of a full plate. For example, if a 12 by 8 plate configuration is selected, and the number of individuals is set to 100, then the plate pool for the second plate will only have four individuals, since the remaining 96 are located in the first plate. This implies that different pools will have different numbers of samples and, therefore, different numbers of total haplotypes. Although large populations can be analyzed, pools containing more than 96 individuals should be avoided.

To simulate errors for each read, the range between the minimum and the maximum rates is split into n intervals, where n is the read length. For the nth base in each read, a random decimal within the nth interval is selected and used as the error probability. This number is converted to a quality score for the fastq file. With the decimal selected, a random integer between 0 and 1 is generated, and if it is smaller than the latter, a random base different from the correct one is placed in that position to simulate an error.

The simulator produces a series of files with a given prefix. The first one is a VCF file with the simulated mutations generated for each individual. This file serves as a gold-standard for benchmark experiments. The second is a csv file that indicates which row, column and plate pool is associated with each individual in the population. Two fastq files are generated for each pool according to the current standard for paired-end sequencing. Read ids include the associated individual from which it was obtained, the pool number, and a unique identifier.

Data Sets

Simulated TILLING Dataset

We tested the newly added functions to NGSEP using two datasets. The first one was derived from the simulator: we selected eight genes in common bean (*Phaseolus vulgaris L.*) that are considered to be important for agronomic traits in this crop such as seed color, resistance to herbicides and tolerance to drought, among others. For each gene, primers were designed using the

online tool primer3[1] to generate amplicons that ranged from 279 to 621 bp and covered all exons in each gene when possible. Overlapping amplicons were designed to improve coverage of the target regions. The simulation was run for a population of 288 diploid individuals in 6 × 8 plates, with a read length of 100 bp and coverage of either 10X, 50X, and 100X. The pool design and population size leads to a total of 48 individuals per row and plate pool and 36 individuals for the column pools (**Figure 1**).

Time and memory benchmarking of the simulator were performed by running other two simulations, along with the one mentioned above. For both of the other simulations we considered a population of 800 individuals with 300 mutations, one with an 8 × 8 plate design and another with an 8 × 12 design. Simulations were run on a Desktop Computer with an Intel Core i7-6700 CPU @ 3.40 GHz, 16 GB of memory and Windows 10 operating system. Times and memory usage were recorded in Java.

Rice Dataset

The raw sequencing reads of a TILLING experiment described in Tsai et al. (2011) were downloaded from the SRA NCBI database (BioSample: SAMN00715843) and mapped to the rice reference genome *Oryza sativa* v7.0. This experiment included 768 individuals sequenced in 44 pools with maximum 64 individuals per pool.

Read Mapping and Variant Calling

All reads in fastq format were mapped to the respective reference genome of the corresponding organism, *Phaseolus vulgaris* for the simulated data and *Oryza sativa* for public data, using NGSEP option ReadsAligner with following parameters modified from the default settings: -k 20 and -m 1. Obtained bam files were then sorted by coordinate using Picard 2.23.0[2]. Alignment rates of 100% were obtained for all mapped reads.

For benchmarking TILLING variant calling and triangulation through NGSEP, we tested a total of 5 additional variant callers in the same datasets: CRISP (Bansal, 2010), Lofreq (Wilm et al., 2012), Freebayes (Garrison and Marth, 2012), GATK (McKenna et al., 2010[3]), and SNVer (Wei et al., 2011). To the best of our knowledge, the only tool capable of identifying mutant individuals in overlapping pools is CAMBa (Missirian et al., 2011). We were unable either to run CAMBa by ourselves nor received a response after trying to contact the developers, so we omitted said tool.

For a fair comparison between tools we adjusted different parameters for variant calling as follows: CRISP, —use duplicates was set to 1 and —qvoffset to 33. For LoFreq, we used the —no-default-filter option and set -m to 20. For Freebayes, the —pooled-discrete option was used and —min-mapping-quality was set to 20. For GATK, we used the HaplotypeCaller algorithm with option —heterozygosity equal to 0.5 and option —max-reads-per-alignment-start set to 0. Finally, we ran SNVerPool with default parameters. SNVer and CRISP allowed us to specify and include the ploidy of each pool (either 72 or 96 depending on the specific pool) in the input file containing the names or paths to the bam files of each pool. Freebayes and GATK were run independently for each pool setting the ploidy to 72 for all column pools and to 96 for the row and plate pools. Lofreq is designed to call low frequency variants and does not have a ploidy option. Finally, for NGSEP, we ran the SingleSampleVariantsDetector functionality with options -h equal to 0.5, -maxAlnsPerStartPos set to 0, -maxBaseQS set to 100 (for real data this option was set to 30), and -psp. Ploidy was adjusted based on the specific pool as explained for the tools above. The commands and parameters are provided in **Supplementary File 1**.

Comparison of Variant Callers Performance

Performance of four of the variant callers was determined in terms of the time spent to call variants in all 20 pools of an artificial mutant population comprising 288 individuals. All tools were tested on a laptop with 4 GB memory, Intel Core i5-7200U CPU @ 2.50 GHz × 4 and Ubuntu 20.04.1 LTS as operating system.

Accuracy of four of the variant callers was determined in terms of the number of variants correctly called in each pool. First, the pool gold standard vcf was generated using the class TillingIndividualVCF2PoolVCF in NGSEP. Then, the function VCFFilter was used to generate the gold standard vcfs per pool using the options -saf to provide the pool ID to be filtered out each time and -fi to filter out sites in which only one allele was observed. Finally, the function VCFGoldStandardComparator was used to compare the vcfs obtained from the variant calling step with the gold standard for the same pool. The output of this comparison is a text file that includes the number of true positives (TP), false negatives (FN), and false positives (FP) detected after variant calling, among others. These values were used to calculate the sensitivity of each tool expressed as TP/(TP+FN).

Identification of Individuals Carrying the Mutations

With the exception of SNVer and CRISP, all variant callers generate a single VCF per pool. The VCFs from SNVer and Crisp include all the samples in one single file. The VCF file generated by lofreq is outdated and does not provide the genotypes per sample. We designed custom scripts to fix the output files of lofreq and crisp. Once fixed, the output VCFs obtained from CRISP were filtered using the option VCFFilter from NGSEP to generate individual VCFs per pool from the population VCF. The parameters used were -saf to provide the pool ID to be filtered out from the original VCF and -fir to remove sites in which only the reference allele was observed. The output VCF obtained from SNVer does not provide information about the observed allele frequencies per sample and could not be fixed to generate a file that could be filtered with NGSEP to generate the individual files. Once we had the VCF files per pool from each tool, we triangulated the output of each caller using the TillingPoolsIndividualGenotyper

[1] https://bioinfo.ut.ee/primer3/

[2] http://broadinstitute.github.io/picard/

[3] https://www.biorxiv.org/content/10.1101/201178v3

functionality. Briefly, the genotyper triangulates the calls of all possible trios of pools (overlapping pools) and then assigns mutations to each individual using the information of the row, column and plate pool to which every member of the population is associated, which was obtained from the simulation process. The output VCF was then compared to the gold standard VCF that contains the true mutations in each individual of the population using the option VCFComparator in NGSEP. Sensitivity was determined as the number of SNPs identified by each tool over the total number of SNPs in the individual gold standard VCF.

Analysis of TILLING Data

Variant calling in the rice TILLING population was performed using NGSEP SingleSampleVariantsDetector, the GATK HaplotypeCaller, Lofreq and CRISP modifying the same parameters as for the simulated data. Ploidy was adjusted according to the pooling strategy described in Tsai et al. (2011) for row, column and dimension (plate) pools varying from 96 to 128. Single vcfs per pool were subjected to the triangulation process using the functionality TillingPoolsIndividualGenotyper providing a pools descriptor file that we generated based on the size of the population (768 individuals) and sampling strategy used in their study. The final vcf was annotated using the function VCFAnnotate and filtered with the function VCFFilter in NGSEP to keep only biallelic SNVs. Summary statistics were calculated using the function VCFSummaryStats of NGSEP.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: Data from the rice population analyzed in this study is available at the sequence read archive (SRA) of NCBI with biosample accession number SAMN00715843.

AUTHOR CONTRIBUTIONS

JG, JA-M, and JD conceived the study. JA-M and JD developed the software components, JG and JA-M performed the simulation and comparison experiments and analyzed the rice population. All authors contributed to the manuscript and approved its final version.

FUNDING

This work was supported by internal funds of Universidad de los Andes through the FAPA initiative led by the Vice-presidency of Research and Knowledge Creation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.624513/full#supplementary-material

Supplementary Table 1 | Correspondance between variants detected by NGSEP, GATK, and Lofreq in the rice TILLING population and those reported by Tsai et al., 2011.

Supplementary Table 2 | Raw variants detected by NGSEP in a rice TILLING population.

Supplementary Table 3 | Raw variants detected by GATK in a rice TILLING population.

Supplementary Table 4 | Raw variants detected by Lofreq in a rice TILLING population.

Supplementary Table 5 | General summary of types of mutations detected in a rice TILLING population.

Supplementary File 1 | Details of parameters used to run each of the variant calling tools compared in this work.

REFERENCES

Bajaj, D., Srivastava, R., Nath, M., Tripathi, S., Bharadwaj, C., Upadhyaya, H. D., et al. (2016). EcoTILLING-based association mapping efficiently delineates functionally relevant natural allelic variants of candidate genes governing agronomic traits in chickpea. *Front. Plant Sci.* 7:450. doi: 10.3389/fpls.2016.00450

Bansal, V. (2010). A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics* 26, i318–i324. doi: 10.1093/bioinformatics/btq214

Chen, L., Hao, L., Parry, M. A., Phillips, A. L., and Hu, Y. G. (2014). Progress in TILLING as a tool for functional genomics and improvement of crops. *J. Integr. Plant Biol.* 56, 425–443. doi: 10.1111/jipb.12192

Dashnow, H., Bell, K. M., Stark, Z., Tan, T. Y., White, S. M., and Oshlack, A. (2019). *Pooled-parent Exome Sequencing to Prioritise De Novo Variants in Genetic Disease. BioRxiv [Preprint]*. Available online at: https://www.biorxiv.org/content/10.1101/601740v1.abstract (Accessed October 28, 2020)

Duitama, J., Kafuri, L., Tello, D., Leiva, A. M., Hofinger, B., Datta, S., et al. (2017). Deep assessment of genomic diversity in cassava for herbicide tolerance and starch biosynthesis. *Comput. Struct. Biotechnol. J.* 15, 185–194. doi: 10.1016/j.csbj.2017.01.002

Garrison, E., and Marth, G. (2012). *Haplotype-based Variant Detection from Short-read Sequencing. arXiv [Preprint]*. Available online at: https://arxiv.org/abs/1207.3907 (Accessed October 28, 2020).

Geourjon, C., and Deleage, G. (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinformatics* 11, 681–684. doi: 10.1093/bioinformatics/11.6.681

Henry, I. M., Nagalakshmi, U., Lieberman, M. C., Ngo, K. J., Krasileva, K. V., Vasquez-Gross, H., et al. (2014). Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. *Plant Cell* 26, 1382–1397. doi: 10.1105/tpc.113.121590

Holme, I. B., Gregersen, P. L., and Brinch-Pedersen, H. (2019). Induced genetic variation in crop plants by random or targeted mutagenesis: convergence and differences. *Front. Plant Sci.* 10:1468. doi: 10.3389/fpls.2019.01468

Huang, H. W., Mullikin, J. C., Hansen, N. F., and NISC Comparative Sequencing Program (2015). Evaluation of variant detection software for pooled next-generation sequence data. *BMC Bioinform.* 16:235. doi: 10.1186/s12859-015-0624-y

Irshad, A., Guo, H., Zhang, S., and Liu, L. (2020). TILLING in cereal crops for allele expansion and mutation detection by using modern sequencing technologies. *Agron. J.* 10:405. doi: 10.3390/agronomy10030405

Kim, S., and Tai, T. H. (2014). Identification of novel rice low phytic acid mutations via TILLING by sequencing. *Mol. Breed.* 34, 1717–1729. doi: 10.1007/s11032-014-0127-y

Kurowska, M., Daszkowska-Golec, A., Gruszka, D., Marzec, M., Szurman, M., Szarejko, I., et al. (2011). TILLING-a shortcut in functional genomics. *J. Appl. Genet.* 52:371. doi: 10.1007/s13353-011-0061-1

Lange, V., Böhme, I., Hofmann, J., Lang, K., Sauter, J., Schöne, B., et al. (2014). Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genom.* 15:63. doi: 10.1186/1471-2164-15-63

Ma, X., Sajjad, M., Wang, J., Yang, W., Sun, J., Li, X., et al. (2017). Diversity, distribution of Puroindoline genes and their effect on kernel hardness in a diverse panel of Chinese wheat germplasm. *BMC Plant Biol.* 17:158. doi: 10.1186/s12870-017-1101-8

Manimekalai, R., Suresh, G., Govinda Kurup, H., Athiappan, S., and Kandalam, M. (2020). Role of NGS and SNP genotyping methods in sugarcane improvement programs. *Crit. Rev. Biotechnol.* 40, 865–880.

McCallum, C. M., Comai, L., Greene, E. A., and Henikoff, S. (2000). Targeting induced local lesions in genomes (TILLING) for plant functional genomics. *Plant Physiol.* 123, 439–442. doi: 10.1104/pp.123.2.439

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome Res.* 20, 1297–1303. doi: 10.1101/gr.107524.110

Milburn, D., Laskowski, R. A., and Thornton, J. M. (1998). Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.* 11, 855–859. doi: 10.1093/protein/11.10.855

Missirian, V., Comai, L., and Filkov, V. (2011). Statistical mutation calling from sequenced overlapping DNA pools in TILLING experiments. *BMC Bioinform.* 12:287. doi: 10.1186/1471-2105-12-287

Perera, I., Fukushima, A., Akabane, T., Horiguchi, G., Seneweera, S., and Hirotsu, N. (2019). Expression regulation of myo-inositol 3-phosphate synthase 1 (INO1) in determination of phytic acid accumulation in rice grain. *Sci. Rep.* 9:14866. doi: 10.1038/s41598-019-51485-2

Raja, R. B., Agasimani, S., Jaiswal, S., Thiruvengadam, V., Sabariappan, R., Chibbar, R. N., et al. (2017). EcoTILLING by sequencing reveals polymorphisms in genes encoding starch synthases that are associated with low glycemic response in rice. *BMC Plant Biol.* 17:13. doi: 10.1186/s12870-016-0968-0

Rubinstein, M., Eshed, R., Rozen, A., Zviran, T., Kuhn, D. N., Irihimovitch, V., et al. (2019). Genetic diversity of avocado (Persea americana Mill.) germplasm using pooled sequencing. *BMC Genom.* 20:379. doi: 10.1186/s12864-019-5672-7

Slota, M., Maluszynski, M., and Szarejko, I. (2017). "Bioinformatics-based assessment of the relevance of candidate genes for mutation discovery," in *Biotechnologies for Plant Mutation Breeding*, eds J. Jankowicz-Cieslak, T. H. Tai, J. Kumlehn, and B. J. Till (Cham: Springer Nature), 263–280.

Taylor, N. E., and Greene, E. A. (2003). PARSESNP: a tool for the analysis of nucleotide polymorphisms. *Nucleic Acids Res.* 31, 3808–3811. doi: 10.1093/nar/gkg574

Till, B. J., Cooper, J., Tai, T. H., Colowit, P., Greene, E. A., Henikoff, S., et al. (2007). Discovery of chemically induced mutations in rice by TILLING. *BMC Plant Biol.* 7:19. doi: 10.1186/1471-2229-7-19

Tsai, H., Howell, T., Nitcher, R., Missirian, V., Watson, B., Ngo, K. J., et al. (2011). Discovery of rare mutations in populations: TILLING by sequencing. *Plant Physiol.* 156, 1257–1268. doi: 10.1104/pp.110.169748

Wang, T. L., Uauy, C., Robson, F., and Till, B. (2012). TILLING in extremis. *Plant Biotechnol. J.* 10, 761–772. doi: 10.1111/j.1467-7652.2012.00708.x

Wei, Z., Wang, W., Hu, P., Lyon, G. J., and Hakonarson, H. (2011). SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res.* 39:132. doi: 10.1093/nar/gkr599

Wilm, A., Aw, P. P. K., Bertrand, D., Yeo, G. H. T., Ong, S. H., Wong, C. H., et al. (2012). LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* 40, 11189–11201. doi: 10.1093/nar/gks918

Yu, S., Liao, F., Wang, F., Wen, W., Li, J., Mei, H., et al. (2012). Identification of rice transcription factors associated with drought tolerance using the ecotilling method. *PLoS One* 7:e30765. doi: 10.1371/journal.pone.0030765

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

TECHNOLOGY AND CODE
published: 18 February 2021
doi: 10.3389/fgene.2021.624259

Galaxy and MEAN Stack to Create a User-Friendly Workflow for the Rational Optimization of Cancer Chemotherapy

Jorge Guerra Pires[1†], Gilberto Ferreira da Silva[1†], Thomas Weyssow[2], Alessandra Jordano Conforte[1,3], Dante Pagnoncelli[4], Fabricio Alves Barbosa da Silva[3] and Nicolas Carels[1†]*

[1] Plataforma de Modelagem de Sistemas Biológicos, Center for Technology Development in Health (CDTS), Oswaldo Cruz Foundation (FIOCRUZ), Rio de Janeiro, Brazil, [2] Informatic Department, Free University of Brussels (ULB), Brussels, Belgium, [3] Laboratório de Modelagem Computacional de Sistemas Biológicos, Scientific Computing Program, FIOCRUZ, Rio de Janeiro, Brazil, [4] Instituto COI, Rio de Janeiro, Brazil

OPEN ACCESS

Edited by:
Fatemeh Maghuly,
University of Natural Resources and Life Sciences, Vienna, Austria

Reviewed by:
Julie Krainer,
Austrian Institute of Technology (AIT), Austria
Vishal Sarsani,
University of Massachusetts Amherst, United States

***Correspondence:**
Nicolas Carels
nicolas.carels@cdts.fiocruz.br;
nicolas.carels@gmail.com

[†] These authors have contributed equally to this work

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

Received: 31 October 2020
Accepted: 22 January 2021
Published: 18 February 2021

Citation:
Pires JG, Silva GF, Weyssow T, Conforte AJ, Pagnoncelli D, Silva FAB and Carels N (2021) Galaxy and MEAN Stack to Create a User-Friendly Workflow for the Rational Optimization of Cancer Chemotherapy. Front. Genet. 12:624259. doi: 10.3389/fgene.2021.624259

One aspect of personalized medicine is aiming at identifying specific targets for therapy considering the gene expression profile of each patient individually. The real-world implementation of this approach is better achieved by user-friendly bioinformatics systems for healthcare professionals. In this report, we present an online platform that endows users with an interface designed using MEAN stack supported by a Galaxy pipeline. This pipeline targets connection *hubs* in the subnetworks formed by the interactions between the proteins of genes that are up-regulated in tumors. This strategy has been proved to be suitable for the inhibition of tumor growth and metastasis *in vitro*. Therefore, Perl and Python scripts were enclosed in Galaxy for translating RNA-seq data into protein targets suitable for the chemotherapy of solid tumors. Consequently, we validated the process of target diagnosis by (i) reference to subnetwork entropy, (ii) the critical value of density probability of differential gene expression, and (iii) the inhibition of the most relevant targets according to TCGA and GDC data. Finally, the most relevant targets identified by the pipeline are stored in MongoDB and can be accessed through the aforementioned internet portal designed to be compatible with mobile or small devices through Angular libraries.

Keywords: systems biology, translational oncology, personalized medicine, Galaxy, MEAN stack, angular, protein–protein network, Shannon entropy

INTRODUCTION

The worldwide estimate of people diagnosed with cancer was 18.1 million in 2017[1] and it is predicted by the *World Health Organization* (WHO) to be 27 million new cases worldwide by 2030. On its own, breast cancer (BC) continues to be among the most frequent cancer around the world alongside the prostate one. Moreover, BC, alone accounts for almost 2.1 million new cases diagnosed annually worldwide, causing an estimate of 600,000 deaths every year (Bray et al., 2018). Because of these dire statistics, BC has received huge attention from both the academic and the

[1] https://ourworldindata.org/cancer

industry, which resulted in a large corpus of publication (culminating at 25,000 in 2019[2]) and publicly available datasets.

In addition, the well-known heterogeneity of breast cancer has justified the genomic study of tumors on a large scale in search for tumor subtypes that could allow a better understanding of the tumor biology and could serve as support for the establishment of genetic signatures, which, when validated in clinical trials, could pave the way for an increasingly specific and more precise treatment than the clinical parameters currently in use.

It is a more in-depth knowledge of tumor biology that has allowed for greater individualization of available treatments and has made it possible to overcome the relapse and resistance eventually observed with traditional treatments (Naito and Urasaki, 2018). In addition, clinical experience has shown that knowledge of the individual characteristics of each tumor may contribute to better therapeutic results with less toxicity.

According to the *one-size-fits-all* approach of chemotherapy, treatment should fit every individual of a population. As a consequence, it is intrinsically imprecise since it does not take into account the genetic peculiarities of each patient. Thus, a one-size-fits-all treatment approach does not work for everyone and may cause harmful side effects. By contrast, *personalized oncology*, which can be placed into a wider paradigm shift called *personalized medicine*, involves the tailoring of medical treatment to the individual characteristics or symptoms and responses of a patient during all stages of care.

The paradigm of one-size-fits-all treatment is now undergoing a shift toward personalized oncology with the identification of molecular pathways predicting both tumor biology as well as response to therapy. Most of those achievements have been inserted into mathematical and computational models by different groups, which can be used to test therapies and hypothesis; the one presented herein fall into this category.

A *new taxonomy* of disease based on molecular and environmental determinants rather than signs and symptoms has been proposed (Collins and Varmus, 2015). The paradigm revolution lies in the change from a clinician selecting a generic therapy on a heuristic basis to one based on molecular facts, a process called *evidence-based medicine* (Masic et al., 2008).

The tools of systems biology made it possible to analyze the huge amount of data delivered by high throughput technologies (broadly named Big Data, Willems et al., 2019). At the moment, the most common strategy for implementing high throughput technologies in oncology is to map mutations that promote suppressor and oncogenes (Guo et al., 2014; Campbell et al., 2020), which is a typical activity of *pharmacogenomics*. Briefly, pharmacogenomics aims at understanding why individuals respond differently to medicines on a genetic level. Consequently, it enables one to predict an individual's response to a drug according to genetic information and allows one to choose the most appropriate medication according to an individual's genetic composition. Furthermore, when the molecular diagnosis is performed, targeted therapy is designed for acting on specific molecular targets supposed to be relevant for the tumor under consideration (Wilsdon et al., 2018). Notwithstanding all the knowledge we have gathered so far, the relevance of a drug target is not obvious, and many criteria were pursuit in that quest (Catharina et al., 2018).

The development of personalized medicine is directly related to the availability of high-throughput technologies. High-throughput techniques, such as microarray, *RNA sequencing* (RNA-seq), and nanoString[3] are important tools for the characterization of tumors and their adjacent non-malignant tissues (Finak et al., 2006). Therefore, these techniques allow a better understanding of tumor biology (Carels et al., 2020). In particular, RNA-seq analysis through in silico methodologies demonstrated that each tumor is unique considering the protein profile of their up-regulated genes (Carels et al., 2015a).

Following the current state of the art, there are mainly two types of omics tests: (i) prognostic tests, which predicts a clinical outcome, and (ii) therapy guiding tests (theranostics), which enable the identification of patient subgroups with a similar response to a particular therapy (McShane and Polley, 2013). In this report, we focus on theranostics.

A variety of multigene assays are in clinical use or under investigation, which further defines the molecular characteristics of the cancers' dominant biologic pathways. Even if there has been a growing use of biomarkers in clinical trials, the use of single-marker and panel tests is still limited (Vuckovic et al., 2016). Gaining insight into the molecular composition of each tumor is recommended for eliminating the misuse of ineffective and potentially harmful drugs.

Mapping gene alterations by reference to the genome is generally performed to characterize indirect relationships between tumor development and indels, mutations, hyper- or hypo-methylation. By contrast, the description of transcriptome, proteome, or metabolome allows the characterization of a molecular phenotype. Interestingly, most *companion diagnostics* (CD) for cancer characterization on the market are based on mutation profiling. Accordingly, CDs are expected to guide the application of a specific therapy supposed to be efficacious for a given patient's condition (Verma, 2012). As a result, CDs allow the selection of a treatment that is more likely to be effective for each individual based on the genetic signatures of their tumors. Moreover, CDs are also developed for better predicting the patient response to a given treatment.

An approach based on molecular phenotyping recently proposed was the identification of the most relevant protein targets for specific therapeutic intervention in malignant BC cell lines (Carels et al., 2015a) based on the diagnosis of up-regulated interactome hubs. This strategy combined *protein-protein interactions* (PPI) and RNA-seq data for inferring (i) the topology of the signaling network of up-regulated genes in malignant cell lines and (ii) the most relevant protein targets therein. Hence, it has the benefit to allow the association of a drug to the entropy of a target and, additionally, to rank drugs according to their respective entropy by reference to their targets (Carels et al., 2015b).

[2]https://pubmed.ncbi.nlm.nih.gov/?term=breast+cancer

[3]https://www.nanostring.com

Three concepts were considered in the approach followed by Carels et al. (2015a): (i) A vertex with a high expression level is more influential than a vertex with a low expression level. (ii) A vertex with a high connectivity level (hub) is more influential than a vertex with a low connectivity level. (iii) A protein target must be expressed at a significantly higher level in tumor cells than in the cells used as a non-malignant reference to reduce harmful side effects to the patient after its inhibition. It is worth mentioning that each combination of targets that most closely satisfied these conditions was found to be specific for its respective malignant cell lines. These statements were validated *in vitro* on a BC model by Tilli et al. (2016). These authors showed that the inactivation, by *small interfering RNA* (siRNA), of the five top-ranked hubs of connection (top-5) identified for MDA-MB-231, a triple-negative cell line of invasive BC, resulted in a significant reduction of cell proliferation, colony formation, cell growth, cell migration, and cell invasion. Inhibition of these targets in other cell lines, such as MCF-7 (non-invasive malignant breast cell line) and MCF-10A (non-tumoral cell line used as a control), showed little or no effect, respectively. In addition, the effect of joint target inhibition was greater than the one expected from the sum of individual target inhibitions, which is in line with the buffer effect of regulatory pathway redundancy in malignant cells (Tilli et al., 2016).

The signaling network of a biological system is scale-free (Albert et al., 2000), which means that few proteins have high connectivity values and many proteins have low connectivity values. As proven mathematically, the inhibition of proteins with high connectivity values has a greater potential for signaling network disruption than randomly selected proteins (Albert et al., 2000). This evidence was proven in silico by Conforte et al. (2019) in the particular case of tumor signaling networks.

In terms of systems biology, the inhibitory activity of a drug may be modeled by the removal of its corresponding protein target from the signaling network to which it belongs (Carels et al., 2015b; Conforte et al., 2019). The impact of vertex removal from a network can be evaluated by the use of the Shannon entropy, which has been proposed as a network complexity measure and applied by many authors to determine a relationship between network entropy and tumor aggressiveness. Breitkreutz et al. (2012), for instance, inferred a negative correlation between the entropy of networks made of genes documented in the *Kyoto Encyclopedia of Genes and Genomes* (KEGG[4]) database considering cancer types and their respective 5-year survival. The existence of this negative correlation was demonstrated later on by Conforte et al. (2019) using RNA-seq data from bench experiments stored in *The Cancer Genome Atlas* (TCGA now hosted by the *Genomic Data Commons Data Portal* – GDC Data Portal[5]).

The Shannon entropy (H) is given by formula 1

$$H = -\sum_{k=1}^{n} p\left(k\right) log_2\left(p\left(k\right)\right) \quad (1)$$

where $p(k)$ is the probability that a vertex with a connectivity value k occurs in the analyzed network.

The process of multistep mining of high throughput data can be cumbersome to handle by humans and needs translation into machine language and automation (Deelman et al., 2009). Thus, according to the scientific challenge, we developed codes in Perl and Python. To deal with assembling a workflow based on *heterogeneous programming*, i.e., a workflow including more than one programming language, we chose Galaxy (Afgan et al., 2018) that fit this purpose.

Since we believe that a molecular phenotyping strategy is worthwhile for complementing the genotyping approach, we described in this report how to perform the translation from RNA-seq data into therapy targets based on the process described in more detail in Conforte et al. (2019). The most relevant targets stored in MongoDB can be accessed through an internet portal written in JavaScript using the software bundle called MEAN stack and portable to mobile and small devices through Angular Flex-Layout library and *Lazy loading*[6] strategies as described by Fain and Moiseev (2018) and Holmes and Herber (2019).

MATERIALS AND METHODS

Galaxy Pipeline

TCGA Data

The gene expression data were obtained as RNA-seq files from paired samples (control and tumor samples from the same patient) and downloaded from TCGA[7] in February 2016 and from the GDC Data Portal[8] in March 2020. The data selection followed two criteria: (i) for each cancer type, approximately 30 patients with paired samples were required to satisfy statistical significance; and (ii) the tumor samples had to be from a solid tumor. The data from TCGA and GDC are given in **Table 1**.

In TCGA, gene expression values were given for 20,532 genes referred to as GeneSymbol, calculated by *RNA-seq through*

[4]http://www.genome.jp/kegg

[5]https://portal.gdc.cancer.gov

[6]https://en.wikipedia.org/wiki/Lazy_loading. Accessed on 14/10/2020.

[7]https://cancergenome.nih.gov/

[8]https://portal.gdc.cancer.gov/

TABLE 1 | RSEM-UQ from paired tumor-stroma data retrieved from TCGA and FPKM-UQ from GDC.

Tumor type	Abbreviation	OS[1]	TCGA, *n*[2]	GDC, *n*
Stomach adenocarcinoma	STAD	38	32	27
Lung adenocarcinoma	LUAD	40	57	57
Lung squamous cell carcinoma	LUSC	47	50	48
Liver hepatocellular carcinoma	LIHC	49	49	50
Kidney renal clear cell carcinoma	KIRC	63	71	71
Kidney renal papillary cell carcinoma	KIRP	75	32	31
Breast cancer	BRCA	82	72	46
Thyroid cancer	THCA	93	57	56
Prostate cancer	PRAD	98	51	50

[1]*OS: 5-years overall survival taken from Liu et al. (2018) according to Conforte et al. (2019), %.* [2]*n: Sample size, number.*

expectation maximization (RSEM) (Mortazavi et al., 2008; Li and Dewey, 2011). Since they were normalized according to the upper quartile methods (formula 2) as reported in GDC documentation[9], we denoted them as RSEM-UQ. In the case of GDC, gene expression values were given for 60,483 sequences, calculated by FPKM and referred to as Ensembl accession number. As those values were also normalized by upper quartile, they were denoted, here, as FPKM-UQ. We considered RNA-seq from BRCA and LUAD as non-significant because of inconsistencies between *raw counts* file names, which led to a final sample of 16 and 17 for LUAD and BRCA, respectively. The 14,126 genes for which the equivalence between GeneSymbols and UniProtKB could be obtained went through further analysis.

$$N_{norm} = \frac{RC_g * 10^9}{RC_{g75} * L} \quad (2)$$

where:

RC_g: Number of reads mapped to the gene;

RC_{g75}: The 75th percentile read count value for genes in the sample;

L: Length of the coding sequence in base pairs.

ArrayEXPRESS Data

Fastq files from RNA-seq of tumor-stroma paired samples from 14 PRAD[10], and 18 *non-small cell lung cancer* (NSCLC)[11], were retrieved from ArrayEXPRESS[12]. These files were compared to the proteins of the EBI's interactome (see below) using BLASTx and processed through our pipeline to measure the average entropies of malignant up-regulated genes from both PRAD and NSCLC. The statistical significance of average entropy differences between PRAD and NSCLC was assessed through the Student's *t*-test using formula 3:

$$u_{obs} = \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{SCE_1}{n_1(n_1-1)} + \frac{SCE_2}{n_2(n_2-1)}}} \quad (3)$$

where:

$\bar{x}_i$: The average of sample i;

SCE_i: the sum of squared differences of sample i;

n_i: the size of sample i.

Because sample sizes of PRAD (n = 14) and NSCLC (n = 18) were less than n = 20, u_{obs} was compared to the theoretical value $t_{1-\alpha/2}$ of the Student's distribution using the k degree of freedom calculated according to formula 4 (Welch, 1949; Dagnelie, 1970):

$$k = \frac{\left[\frac{SCE_1}{n_1(n_1-1)} + \frac{SCE_2}{n_2(n_2-1)}\right]^2}{\frac{1}{n_1-1}\left[\frac{SCE_1}{n_1(n_1-1)}\right]^2 + \frac{1}{n_2-1}\left[\frac{SCE_2}{n_2(n_2-1)}\right]^2} \quad (4)$$

with n_1-1 < k < $n_1 + n_2$-2.

[9]https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/

[10]https://www.ebi.ac.uk/ena/data/view/PRJEB2449

[11]https://www.ebi.ac.uk/ena/data/view/PRJNA320473

[12]https://www.ebi.ac.uk/arrayexpress/

Identification of Hubs Among Genes Up-Regulated in Tumor Samples

To identify genes that were significantly differentially expressed in the tumor samples of patients, we subtracted gene expression values of control samples from their respective tumor paired samples. The resulting values were called differential gene expression. Negative differential gene expression values indicated higher gene expressions in control samples, while positive differential gene expression values indicated higher gene expressions in tumor samples.

The histogram of differential expression was normalized with the Python packages *scipy*. We used the probability density and cumulative distribution functions, respectively abbreviated as PDF and CDF, in the interval of differential gene expression from −20.000 to +20.000, to calculate the critical value corresponding to the one-tail cumulated probability p = 0.975, which corresponded to a p-value α = 0.025. We considered the genes as up-regulated when their differential expression was larger than the critical value corresponding to p = 0.975. The −20.000 to +20.000 range worked fine for the p-value and normalization conditions presented in this report. However, some normalization procedures flatten the probability distribution with Bayesian functions for variance minimization. Under these conditions, a p–value of 0.001 may represent a very large critical value of 80,000 or more, which would induce the scipy package to return "out of range." To beat this challenge, we introduced the possibility of tuning the −20.000 to +20.000 range to allow the user to try other normalization conditions together with more restrictive p-values. However, for coherence, all the data produced in this report were obtained with critical values in the −20.000 to +20.000 range.

In a subsequent step, the protein–protein interaction (PPI) subnetworks were inferred for the proteins identified as products of up-regulated genes. The subnetworks were obtained by comparing these gene lists with the human interactome.

The human interactome (151,631 interactions among 15,526 human proteins with UniProtKB accessions) was obtained from the intact-micluster.txt file (version updated December 2017) accessed on January 11, 2018[13].

We used the PPI subnetworks of up-regulated genes from each patient to identify each vertex (protein) degree through automated counting of their edges. These values were used to calculate the Shannon entropy of each PPI subnetwork as explained in the section "Shannon Entropy" below.

Shannon Entropy

The Shannon entropy was calculated with formula 1, where $p(k)$ is the probability of occurrence of a vertex with a rank order k (k edges) in the subnetwork considered. The subnetworks were generated automatically from gene lists found to be up-regulated in each patient.

Validation Process

The diagnosis of up-regulated genes with a higher vertex degree, which we considered as the most relevant target here, depends

[13]ftp://ftp.ebi.ac.uk/pub/databases/intact/current/psimitab/intact-micluster.txt

on how *fastq* and *raw count* files are processed. First, *fastq* reads need to be transformed into *raw counts* and, second, *raw counts* need to be normalized. For validating this process, we used the data of RSEM-UQ from TCGA as available in 2016 that we referenced to as TCGA RSEM-UQ below. When referring to the FPKM-UQ files from GDC accessed in March 2020, we denoted them as *GDC FPKM-UQ*. Since we had no access to the raw counts files of TCGA, we used the data from GDC. GDC provided the TCGA data in Bam format, *raw counts*, FPKM, and FPKM-UQ files. Since we knew the correlation between the entropy and the 5-years *overall survival* (OS) for nine cancer types as established from TCGA RSEM-UQ (Conforte et al., 2019), the validation challenge was (i) to normalize the GDC *raw counts* files (we characterized this step as $RPKM_{upper}$, see the description below) from tumors of the nine cancer types; (ii) to compare the $RPKM_{upper}$ normalization to the TCGA RSEM-UQ for critical value, number of up-regulated genes, and the correlation between entropy and 5-years OS as well as targets; (iii) to compare $RPKM_{upper}$, TCGA RSEM-UQ and GDC FPKM-UQ for critical value, number of up-regulated genes, the correlation between entropy and 5-years OS, and targets, and (iv) to optimize $RPKM_{upper}$ by log transformation for target selection given the maximization of the correlation coefficient of the relationship between entropy and 5-years OS. Having this process validated, it might be applied to any method of read counting from *fastq* file by read mapping. This process is summarized in **Figure 1**.

As TCGA, GDC uses the RSEM methodology to map reads to reference genes. Here, instead of using the human genome sequence GRCh38.d1.vd1[14], we used the proteins sequences from UniprotKB as a reference. Since only about 80% of the proteins from the EBI's interactome referenced by UniprotKB matched the *consensus coding sequences* (CCDS)[15] of Ensembl, we decided to map reads in *fastq* files directly with the proteins sequences of the intact-micluster interactome using BLASTx. Thus, in the first instance, the exercise of validation concerned the processing of *raw counts* into RPKM-UQ output.

For *raw count* normalization, we used a modified version of the RPKM formula (5):

$$RPKM = \frac{RC_g * 10^9}{RC_{pc} * L} \tag{5}$$

where:

RC_g: Number of reads mapped to the gene;

RC_{pc}: Number of reads mapped to all protein-coding genes;

L: Length of the coding sequence in base pairs.

RPKM is relative to the total number of reads, which is a linear expectation. Quantile normalization (Bolstad et al., 2003) forces the distribution of the normalized data to be the same for each sample by replacing each quantile with the average quantile across all samples. Instead, one may focus on a specific quantile. For instance, the upper quartile normalization (Bullard et al., 2010) divides each read count by the 75th percentile of the read counts in its sample. However, the gene frequency (y) according to the gene expression (x) follows a power law (the relationship of log(y) and log(x) is linear, data not shown) (see also Balwierz et al., 2009; Awazu et al., 2018). RPKM, as defined in formula 5, does not take the non-linearity associated to large expression level into account. By contrast, the *upper quartile* normalization enables us to take the non-linearity associated with extreme expression values into account. Formula 5 can be written as formula 6:

$$RPKM_{upper} = \frac{RC_g * 10^9}{L * (RC_{pc} - (\delta * RC_{pc}))} \tag{6}$$

where δ is a tuning factor.

For δ = 0, formula 6 is equivalent to RPKM (formula 5) and for δ = 0.25, it is equivalent to a *upper quartile* normalization. In this work, we used δ = 0.05 because it optimized the coefficient of correlation between entropy and 5-years OS.

It appeared that in addition to the TCGA RSEM-UQ (accessed in 2016), GDC (accessed in March 2020) implemented a correction for false positive minimization (Anders and Huber, 2010; Love et al., 2014; Holmes and Huber, 2019). The result of this minimization is a flatten power law of gene expression with an effect similar to that of formula (7):

$$LogNorm = C * x_i * (log_b \left(log_b(x_i + 1)\right) + 1) \tag{7}$$

where:

C: is a constant that was set to 20 to optimize the coefficient of correlation of the relationship between entropy and 5-years OS;

x_i: is the $RPKM_{upper}$ value of the i_{th} element;

b: is the base of the logarithm, which was set to 1.1.

As can be seen from formula 7, the FPKM-UQ output follows a *log-log* relationship except for the variance that is stabilized by a Bayesian process.

For assessing the efficiency of TCGA *raw counts* processing according to formula 6, we tabulated the sample size of subnetworks of up-regulated genes as well as the critical values obtained for PDF = 0.975. This process was performed by calculating $RPKM_{upper}$ on the *raw counts* available from GDC, and compared the critical values to those obtained from GDC FPKM and TCGA RSEM-UQ. We also compared the correlation between entropy and 5-years OS obtained with *raw counts* normalized with $RPKM_{upper}$ to that obtained by using the TCGA RSEM-UQ. Finally, we compared the most relevant targets obtained from both processes.

In the case of the GDC FPKM-UQ, one more step was necessary since the *raw counts* sequentially processed through formula 6 and 7 had to be compared to FPKM-UQ data available from the GDC portal. Again, we compared the performance of processing *raw counts* with formula 6 and 7 to GDC FPKM-UQ data considering (i) the critical values for PDF = 0.975, (ii) the subnetwork size of up-regulated genes, (iii) the correlation of entropy vs. 5-years OS, and (iv) the list of most relevant targets obtained through both processes.

Finally, we also compared the performance of sequentially processing *raw counts* through formula 6 and 8 (formula 8 is derived from Cloonan et al., 2008) by using the same measures as just described (i to iv). We applied this formula because we

[14]https://gdc.cancer.gov/about-data/data-harmonization-and-generation/gdc-reference-files

[15]https://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi

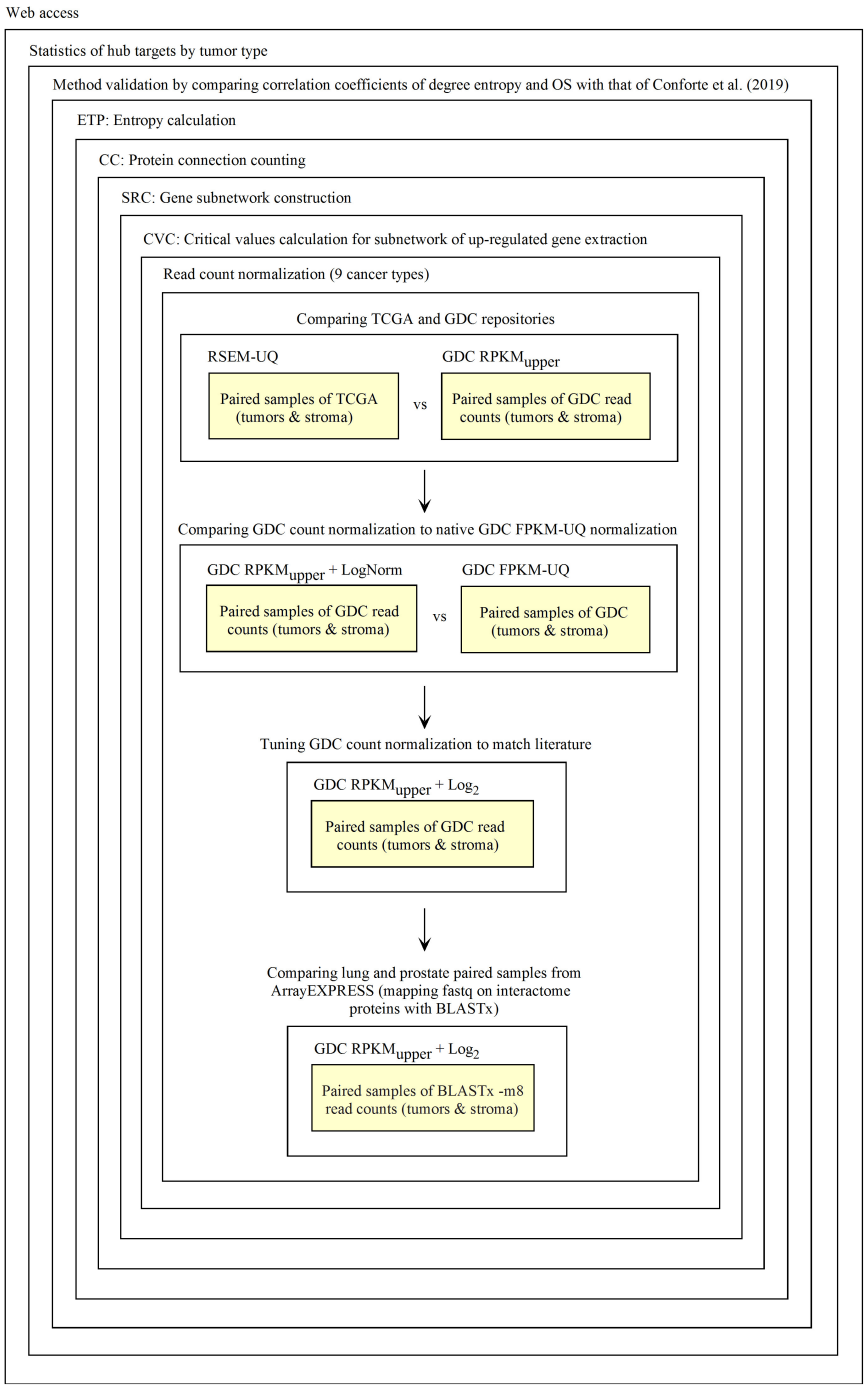

FIGURE 1 | Process of Galaxy workflow validation.

noticed that it optimized the coefficient of correlation of the relationship between entropy and 5-years OS.

$$Log2 = x_i \left(log_b(x_i + 1)\right) \tag{8}$$

where:

x_i: is the $RPKM_{upper}$ value of the i_{th} element;

b: is the base of the logarithm, which was set to 2.

Galaxy Scripts

Galaxy is a scientific open-source workflow platform that aims at helping users to perform repetitive and complex operations over large datasets. With Galaxy, users can visually create processing pipelines reproducing the data flow over programs and datasets that are viewed as interconnected box objects. Additionally, Galaxy is written in Python and JavaScript, but has an XML like

interface able to transfer the processing flux to other languages. Galaxy comes with a rather large initial set of tools that can be added to the desktop according to simulation demands. Internally, every Galaxy tool is made up of a XML file that describes its functionalities and interface. Once XML interfaces are programmed, Galaxy is very simple to operate in an object-oriented mode by linking input data with scripts together.

By means of a specific script (see below), Galaxy can store data in MongoDB, which is a non-relational object-oriented database (NoSQL) (Bradshaw et al., 2019). MongoDB can be accessed through Angular, which serves as a frontend framework for users (the physician or/and technician operating the system) (Fain and Moiseev, 2018).

As outlined in the introduction of this report, our Galaxy workflows are derived from the agglomeration of Perl scripts (except for CVC.py) that were written for previous reports (Carels et al., 2015a; Conforte et al., 2019). These tools are as follow:

(1) *Count Connections* (CC) counts the number of connections that each protein has with their neighbors in a subnetwork of up-regulated genes. CC is an intermediate step to compute the entropy.
(2) *Critical Value Calculation* (CVC) computes a critical value according to the normal distribution that fits the observed data and a probability level informed by the user. All genes with expression values above the critical value, used here as a threshold, are considered as up-regulated.
(3) *Differentially Expressed Genes List* (DEGL) computes de differential gene expression between RNA-seq data from tumoral and control samples (tumor minus control).
(4) *Entropy Calculation* (ETP) computes the Shannon entropy corresponding to a subnetwork. Here, we typically considered the subnetworks of genes that are up-regulated in tumors.
(5) *Translation of Gene Symbol into UniProt KB accession numbers* (GS2UP). Former TCGA data files identified genes by gene Symbol, while the interactome from EBI (the intact-micluster.txt file) uses UniProtKB accession numbers. GS2UP translates the gene symbols to UniProtKB accession numbers to build the subnetwork of up-regulated genes.
(6) *Translation from Ensembl into UniProt KB accession numbers* (Ensembl2UP). GDC data files identify genes by reference to Ensembl, while the interactome from EBI (the intact-micluster.txt file) uses UniProtKB accession numbers. Ensembl2UP translates the Ensembl to UniProtKB accession number to build the subnetwork of up-regulated genes.
(7) *Protein To Total Connections Sorted* (PTTCS) sorts the file of malignant up-regulated genes according to the level of connectivity found for their respective protein in descending order.
(8) Subnetwork Construction (SRC) computes a subnetwork of proteins based on a gene list by reference to the intaractome; here, the gene list is typically the list of up-regulated genes.
(9) *Reads Per Kilobase Million – Upper Normalization* ($RPKM_{upper}$) computes de normalization of RNA-seq data according to formula 6.
(10) *Double Logarithm Transformation* (LogNorm) computes de normalization of $RPKM_{upper}$ data according to formula 7.
(11) *Base 2 Logarithm Transformation* (Log2) computes de normalization of $RPKM_{upper}$ data according to formula 8.
(12) *PTTCS to MongoDB* (P2M) computes the data storage within MongoDB.

These tools can be downloaded from GitHub: https://github.com/BiologicalSystemModeling/Theranostics under the MIT License, however, the concept of theranostics based on this approach is under the regulation of intellectual property number BR1020150308191 for Brazil.

Pipeline Scaling

To investigate how the pipeline scales, we processed the GDC raw counts data using an AMD Ryzen 9 3900X (4.6 GHz) CPU with 20 threads dedicated to Galaxy and 64 GB RAM. First, we chose LUSC and PRAD tumors as representing high entropy (low OS) and low entropy (high OS) cancer types, respectively. In these two cases, we could exactly compare their scaling until 45 patients by increments of five. For STAD, LIHC, THCA, and KIRC, we measured the processing time for only two patient numbers (15 and 25). We also analyzed the statistical significance of the difference in processing speed observed for entropy and PTTCS pipeline for 25 patients with the Student's *t*-test. Considering the pipeline for hub diagnosis from BLASTx output, we only had access to a small number of patients, which limited the power of the experiment. We compared 3, 6, 9, 12 patients in PRAD and NSCLC from ArrayEXPRESS (see above).

Web Application

As outlined in the introduction, we aimed at releasing a tool based on a phenotyping approach for the rational therapy of cancer. At the moment, the current approach of cancer therapy is still largely based on mutation mapping (genotyping approach), but the potential benefits of integrating RNA-seq data must be considered and this is the purpose of this report.

When producing a bioinformatic application, it is necessary to validate it according to some objective criterion. As presented in the previous section, we chose degree entropy as such a criterion for the validation of the Galaxy pipeline. Galaxy enabled us to test the performance of several configurations for optimizing the correlation between the degree entropy of up-regulated subnetworks and the patient's 5-years overall survival.

However, a website is necessary to make this tool available to the medical community and its development makes part of another step of validation that is its acceptance by professionals. Below, we briefly describe the technologies that we used to build the web site and then described how we implemented them through forms for data submission.

MEAN Stack

Both MongoDB and Angular are part of the MEAN stack (MEAN for M of MongoDB, E of Express.js, A of Angular, and N of Node.js). The use of MongoDB with Node.js, its native driver, is facilitated by the Mongoose[16] library. Mongoose, amongst other benefits, allows (i) the use of JavaScript as a programming language, which save the need for database programming, (ii) the modeling of data before their saving into MongoDB, and (iii) the *horizontal scaling*[17], which means that one can expand storage capacity without the need of multiple structural changes. This last feature decreases the cost of prototyping and expansion. It also enables one to work with several database connections simultaneously.

Node.js is part of the MEAN stack that we used to build the backend of the web application; it is the server used to connect the database and the frontend. Essentially, Node.js is a framework that is used to create servers and has its own HTTP handler (Holmes and Herber, 2019), which eliminates the need of other intermediate libraries.

The MEAN also included Express.js, a JavaScript-based library whose purpose is to facilitate the exploration of the Node.js functionalities (e.g., creating routes).

In addition to JavaScript, Angular also allows programming in TypeScript, which includes the concept of *variable type* and a set of internal libraries (e.g., RxJS for asynchronous programming). Furthermore, Angular offers compatibility with many web development libraries, such as Bootstrap, jQuery, and Forms.

MEAN stack elements have JavaScript as a common programming language and *JavaScript Object Notation* (JSON) as a common file exchange format. Except for Angular which is a frontend technology, MongoDB, Express.js, and Node.js run on the server-side, as so they are generally classified as the 'backend' of a web application (Holmes and Herber, 2019).

Our web application has been deployed in a cloud environment using Heroku[18,19] by implementing the MEAN stack (Holmes and Herber, 2019). The version of Angular that we used here was CLI 8.3.23. In addition to those technologies, we were also using NPM libraries designed to support the MEAN stack. We used JavaScript for interfacing with MongoDB, Express.js, and Node.js as well as several free packages available in NPM to support these technologies[20]. For instance, we used *Visual Studio Code* (version 1.48) as a programming platform and *Avast Secure Browser* as a testing browser. Avast provides a built-in test system for small devices such as smartphones.

Angular

After compilation, Angular generates *Single Page Applications* (SPAs), which means that the code is sent to the browser at once when the user accesses the page for the first time. The main benefit of this approach is to create *dynamic pages*, improving the navigation experience to the frontend user. Angular speeds up the server–client communication by avoiding multiple client accesses and enabling complex calculations as well as data validations within the client browser. Moreover, the main difference of SPAs compared to a classic web application based on PHP (i.e., *static pages*) is that it does not load the page when one changes from page to page since all the code is already on the browser. Therefore, the main benefits of Angular are that (i) heavy calculations can be performed on the frontend side, which can alleviate the computing charge on the server; (ii) pre-validated data may be submitted to the server, avoiding the need for *back and forth* validation process; (iii) TypeScript (a superset of JavaScript) has the structures of a conventional programming language with powerful build-in libraries (e.g., RxJS), which enables the performance of scientific calculations on the frontend side if needed.

We also took advantage from the Angular library called *Angular Material*[21], which allows predefined functions such as forms and themes. Angular Material can be used either within the HTML language as predefined tags or within TypeScript for dynamic pages (e.g., for Reactive forms). We used Angular Material within TypeScript since it provides much more programming freedom, e.g., form validation.

Node.js

One of the key features of Node.js is that it allows the usage of JavaScript (or TypeScript) on the server-side. Until then, JavaScript was restricted to browsers and this progress has been possible due to the V8 Engine that compiles JavaScript code to native machine code at runtime. We used the NPM repository to install and manage all the Node.js (version 10.16.3) packages.

Node.js applications are *stateless*, which means that they do not keep information about the user stored locally and for that reason only require low amount of local RAM. Node.js applications are also single thread, which means that they do not stop the main thread as they result from users' interactions.

We chose the *JSON Web Token* (JWT) approach to save the user information temporally on the frontend. JWT is an encoded string used when the frontend communicates with the server. The benefits of JWT are (i) that it carries a server signature, which must match whenever the user tries to communicate with the server, and (ii) that an expiration date may be set, which implies token refreshing.

Express.js

Express.js is a library whose purpose is to facilitate the exploration of the Node.js functionalities (e.g., creating routes and servers). Here, we used Passport.js[22] together with Express.js (version 4.16.1) to build user sections as described by Holmes and Herber (2019).

MongoDB

MongoDB can be accessed through Angular using Node.js as server; Angular serves as a frontend framework for users (Fain and Moiseev, 2018). MongoDB is *horizontally*

[16] https://mongoosejs.com/docs/

[17] https://docs.mongodb.com/manual/sharding/

[18] https://www.heroku.com/

[19] http://teranostico.herokuapp.com/

[20] https://www.npmjs.com/package/repository

[21] https://material.angular.io/. Accessed on 14/10/2020.

[22] http://www.passportjs.org/. Accessed on 14/10/2020.

expandable[23], which enables to expand storage capability without extensive physical changes. This feature decreases the cost of prototyping and posterior expansion. Another interesting property of MongoDB is the *MongoDB Atlas*[24], which provides cloud storage.

The usage of MongoDB with Node.js is facilitated by the Mongoose[25] library. Mongoose, amongst other benefits, allows (i) the usage of JavaScript as a programming language, which saves the need for database programming, (ii) the modeling of data before their storage into MongoDB, and (iii) the easier exploration of the MongoDB horizontal scaling capability[26].

Angular Flex-Layout

According to Fain and Moiseev (2018), we used a single code to implement *Responsive Web Design* (RWD) to optimize maintenance costs. This strategy allows the user interface layout to change in response to the device screen size (desktop or cell phone). RWD allows the interface simplification on small devices by limiting the display of extra-small devices to key functions (see **Supplementary Figure 1** for screen size and Angular screen size settings).

We tested the responsiveness of our portal on a desktop computer using the built-in developer tool of Avast Secure Browser. We also tested it on the following devices: Moto G4, Galaxy S5, Pixel 2, Pixel 2 XL, iPhone 5/SE, iPhone 6/7/8, iPhone 6/7/8 Plus, iPhone X, iPad, iPad Pro. However, the Avast Secure Browser simulator does not necessarily consider the operating system, and it may give an unexpected display in uncommon devices.

Passport.js

For creating the user section, we used Passport.js[27]. Its main benefits are the possibility of (i) creating customized login system or use pre-defined ones, such as those of Facebook, for example; and (ii) using it with JWT tokens due to their built-in libraries that facilitate their use. To implement JWT within Passport.js, we used *express-jwt*[28], which allows the validation of JWT tokens, including expiration date and abnormal tokens.

Forms

The function of the patient main form is to collect and to store basic information regarding the patient and its tissue samples for genetic analysis. This information is necessary for the posterior retrieval from the system database of patients' medical records. Patient data are central to the system since they articulate genetic analyses with medical records that must be encrypted (e.g., patient name, mother's name, and patient id). The patient data collected through the main form of the frontend are stored together with genetic data from the backend within MongoDB.

The request for a genetic exam is of key importance when it comes to the service provided. When physicians send tumor samples, they will be asked to request their gene expression analysis and provide patient information as well as medical records (see **Supplementary Figure 2**).

The outcome form has such as (i) details of the treatment applied, (ii) treatment benefits, (iii) whether the gene expression-based recommendations were followed, and so forth (see **Supplementary Figure 3**). The outcome form is essential for establishing case statistics.

Angular provides two options when it comes to forms: *Template-Driven Forms* and *Reactive Forms*[29]; we used the latter. The main reason for this choice was that this option provides (i) a set of built-in routines for form validation, including error messages that can easily be shown on the frontend, and (ii) the possibility of building its own customized error handling routines. By error, we mean any input to the form fields that does not fit what is expected, e.g., e-mail out of the format or password that does not match. We were also using form validators that communicate with the server on the background side to check data consistency.

Additionally, we used FormBuilder[30] that is an Angular service used for the programming of Reactive form. With FormBuilder, one can construct JSON objects (our data format), validate the inputs of the forms individually or as a group, and other functionalities.

Encryption, Decryption, Hashing, and JWT Coding

Since we are dealing with potentially sensitive information, we followed standard practices to protect the information submitted to the system and stored on our database. In the current stage of development, we are using standard libraries, which can be replaced by more secure ones as soon as the platform scale up. In the current version, we are using three different approaches to protect information from potential unauthorized accesses: (i) encryption/decryption, (ii) hashing, and (iii) JWT (e.g., communication with API[31]). For encryption/decryption, we are using the library *CryptoJS*.[32] The 'secret' is kept on the server using a library known as *dotenv*[33], which is largely used to store sensitive information in Node.js applications. For hashing, we are using the library *bcrypt*[34] in the following configurations: *bcrypt.genSalt*(*10, callback*), the first argument is the size of the *salt* and the second is the function for hashing.

[23]https://docs.mongodb.com/manual/sharding/. Accessed on 14/10/2020.

[24]https://www.mongodb.com/cloud/atlas. Accessed on 14/10/2020.

[25]https://mongoosejs.com/docs/. Accessed on 14/10/2020

[26]https://docs.mongodb.com/manual/sharding/. Accessed on 14/10/2020.

[27]http://www.passportjs.org/. Accessed on 16/10/20.

[28]https://www.npmjs.com/package/express-jwt. Accessed on 16/10/20.

[29]Components in the Angular realm is a set of three files: CSS (appearance-related), TS (typescript, coding), and HTML (classical static page design file). A page is built from at least one component, which can independently interact with each one of the others (see Fain and Moiseev, 2018 for a more detailed discussion).

[30]https://angular.io/guide/reactive-forms

[31]Application Programming Interface. These routines are designed to access the database following some pre-defined rules such as token authentication.

[32]https://www.npmjs.com/package/crypto-js

[33]https://www.npmjs.com/package/dotenv

[34]https://www.npmjs.com/package/bcrypt

The code for the web site can be downloaded from GitHub: https://github.com/Teranostico under the MIT License.

RESULTS

Galaxy Pipeline

We validated and automated the process published by Conforte et al. (2019). Thus, one sought to reproduce the results obtained by Conforte et al. (2019) when the pipeline was fed with the same data (TCGA RSEM-UQ). We indeed succeeded to reproduce the correlation $r = -0.68$ between entropy and patient's 5-years OS for a probability of $p = 0.975$ in the determination of up-regulated genes, which allowed us to test whether the maximization of r really occurred for $p = 0.975$. To meet this challenge, we measured the correlation coefficient for $p = 0.97$ and $p = 0.98$, and found $r = -0.53$ and $r = -0.60$, respectively. The automated workflow is given in **Figure 2A**.

As shown in **Figure 2A**, the *input data collection* represents a collection of paired samples (tumors identified as 01A and control identified as 11A) with the same list of genes (identified by gene symbol) for each patient of the TCGA database. Following the processing flux, the gene symbols are transformed into UniprotKB accession numbers (GS2UP) to perform the subtraction of the control RNA-seq expression data from that of the tumor (DEGL). The calculation of the critical value that identifies up-regulated genes is performed by the Python script CVC. The critical value is calculated according to a probability level chosen by the user and is used by the script SRC for extracting the list of up-regulated genes. This list is used by the CC script for counting the connections at each vertex of the subnetwork of up-regulated genes. The connection count at each vertex is necessary for computing the Shannon entropy of the tumor subnetwork of up-regulated genes by the ETP script.

We validated the pipeline with the GDC *raw counts* comparing their $RPKM_{upper}$ to the TCGA RSEM-UQ (**Figure 2B** without the log transformation step). First, we computed the *raw counts* according to $RPKM_{upper}$ excluding BRCA and LUAD because of inconsistencies between file names available for FPKM-UQ and *raw counts*. In both BRCA and LUAD, cleaning samples for perfectly matched files led to sample size below $n = 20$, which may bias comparison (sample size is considered to be statistically trustworthy from at least $n = 30$ and needs correction below this threshold). When we compared the critical values for $p = 0.975$ considering the *raw counts* normalized with $RPKM_{upper}$ (**Table 2**, column GDC $RPKM_{upper}$), we found values similar to those obtained by processing TCGA RSEM-UQ data (**Table 2**, column TCGA RSEM-UQ).

We found that critical values for $p = 0.975$ of GDC FPKM-UQ were ∼5 times larger (**Table 2**, column GDC FPKM-UQ), on the average (**Figure 2B** without normalization and log transformation steps), than those of TCGA RSEM-UQ (**Table 2**, column TCGA RSEM-UQ and GDC $RPKM_{upper}$). This difference is due to the processing update performed during the data transfer from TCGA to GDC portal involving the flattening of the differential gene expression distribution.

When we successively computed GDC *raw counts* with formula 6 ($RPKM_{upper}$) and 7 (LogNorm), we found critical values for $p = 0.975$ (**Table 2**, column GDC $RPKM_{upper}$ + LogNorm) close to that of GDC FPKM-UQ (**Table 2**, column GDC FPKM-UQ), suggesting a similar behavior of differential gene expression flattening as the one applied by the GDC data processing (**Figure 2B**).

The comparison of the size of subnetworks of up-regulated genes in tumors is given in **Table 3**. The difference of subnetwork size between GDC FPKM-UQ and GDC $RPKM_{upper}$ + LogNorm samples, on one hand, and TCGA RSEM-UQ and GDC $RPKM_{upper}$ samples, on the other hand, raised the question of whether the large subnetwork size of GDC FPKM-UQ and GDC $RPKM_{upper}$ + LogNorm might be trusted.

The subnetwork sizes obtained by successively processing GDC *raw counts* with formula 6 and 8 (**Table 4**, column Node number) were smaller and more realistic, representing between ∼2% and ∼5% of the human proteome.

As explained above, we did not consider BRCA and LUAD for comparison between $RPKM_{upper}$ and FPKM-UQ. However, the FPKM-UQ correlation plot was similar to that of other authors (data not shown).

The features of the linear regression between the subnetwork entropies and the 5-years OS are given in **Table 5** for the different pipeline configurations tested here.

Interestingly all the combinations involving $RPKM_{upper}$ of **Table 5** resulted in a larger slope of the regression line; in other word, they resulted in an increased statistical significance of the regression line.

Compared to GDC $RPKM_{upper}$ (**Figure 3A**), the introduction of the LogNorm in the workflow of **Figure 2B** resulted in a systematic shift of entropies by as much as ∼1.5 bit toward larger values (in the range of 3.6–4.0 compared to 2.0–2.5 in Conforte et al., 2019), which denote a larger subnetwork of up-regulated genes with larger number of hubs as a consequence of the distribution flattening of differential gene expression. The correlation obtained by successively processing *raw counts* with $RPKM_{upper}$ and LogNorm (**Figure 3B**) was similar ($r = -0.86$ without BRCA and LUAD) to that obtained with GDC FPKM-UQ ($r = -0.76$ without BRCA and LUAD) (**Figure 3D**). Finally, it is the correlation obtained by successively processing *raw counts* with formula 6 and 8 (**Figure 3C**) that showed the best correlation coefficient and slope of the regression line ($r = -0.91$).

The effect of LogNorm on distribution flattening of differential gene expression when comparing $RPKM_{upper}$ to $RPKM_{upper}$ + LogNorm was similar to that observed when comparing TCGA RSEM-UQ (**Figure 4A**) to GDC FPKM-UQ (**Figure 4D**), respectively.

When we compared the correlation coefficient according to p for GDC FPKM-UQ data, we obtained $r = -0.758$, $r = 0.763$, and $r = 0.477$ for $p = 0.95$, $p = 0.98$, and $p = 0.99$, respectively. This result shows that the maximum of r was associated with $p = 0.98$, but the difference with $p = 0.975$ was only 0.002 units of the correlation coefficient, which confirmed that the peak around the maximum of r was less sharp for GDC FPKM-UQ than for TCGA RSEM-UQ since it spreads over $p = 0.95$ and $p = 0.98$.

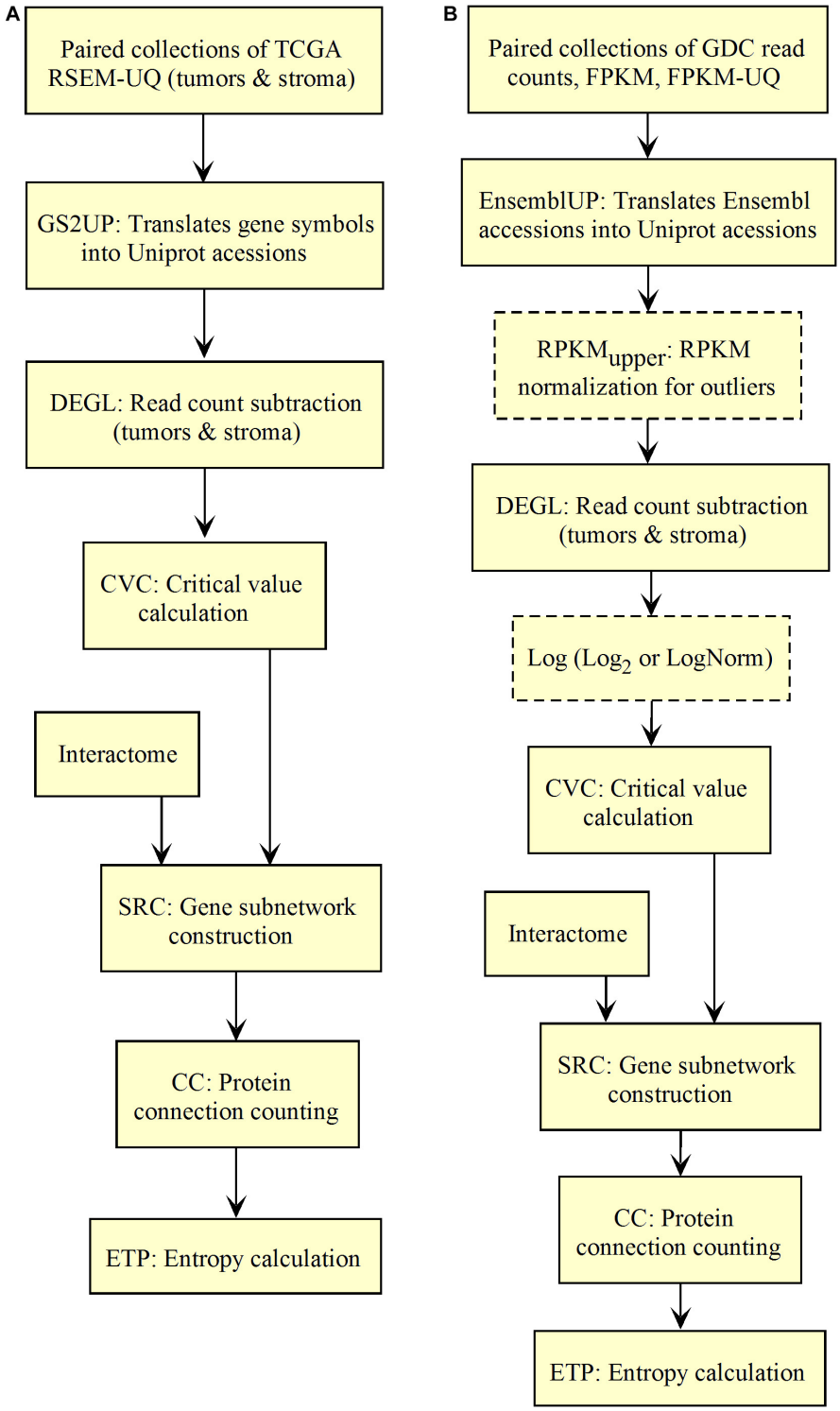

FIGURE 2 | Workflow for the validation of the correlation between the entropy of the subnetworks of up-regulated genes from different tumors and their respective 5-years OS. **(A)** TCGA. **(B)** GDC.

The flattening of the correlation peak according to the probability density appeared as a consequence of the probability density distribution shape. The distribution of FPKM-UQ values was flatter in GDC FPKM-UQ (**Figures 4D–F**) compared to TCGA RSEM-UQ (**Figures 4A–C**), which is reflected by larger critical values associated with GDC (**Table 2**). The validation of

TABLE 2 | Critical values of probability density for $p = 0.975$.

Cancer	GDC $RPKM_{upper}$		TCGA RSEM-UQ		GDC $RPKM_{upper}$ + LogNorm		GDC FPKM-UQ	
Type	Av.	StDev	Av.	StDev	Av.	StDev	Av.	StDev
PRAD	2661.73	498.89	2566.88	507.38	15558.79	1053.56	15809.77	779.91
LUAD	2897.95	437.50	3138.07	313.74	15720.85	1221.94	16340.59	860.48
LUSC	3532.06	426.30	3527.89	429.98	15775.55	857.31	16161.27	730.71
BRCA	3211.72	434.50	3024.87	465.83	15346.96	664.95	15923.64	682.80
KIRC	3133.16	236.39	3162.20	363.44	15820.34	742.76	16310.36	604.57
KIRP	3084.69	365.17	3089.64	390.28	15482.35	905.77	16165.59	597.30
THCA	2610.75	313.49	2590.59	406.12	14876.38	1089.35	15559.35	713.54
STAD	3330.13	444.58	3273.89	470.64	16511.00	865.11	16473.27	718.44
LIHC	3085.76	474.40	3409.36	468.48	16235.74	1087.23	15639.15	801.43
Average	3060.88	403.47	3139.90	420.79	15703.11	943.11	16042.55	721.02
St. Dev.	298.36	83.63	299.07	56.29	479.01	181.82	324.01	86.51

TABLE 3 | Size of subnetwork (vertex number) of genes up-regulated in tumors for a probability density of $p = 0.975$.

Cancer	GDC $RPKM_{upper}$		TCGA RSEM-UQ		GDC $RPKM_{upper}$ + LogNorm		GDC FPKM-UQ	
Type	Av.	StDv.	Av.	StDv	Av.	StDv.	Av.	StDv.
PRAD	269.19	62.01	254.20	40.66	5046.23	1209.45	4029.75	499.89
LUAD	290.35	58.66	276.35	49.07	4973.16	1203.25	4779.27	401.19
LUSC	345.21	48.50	317.12	48.33	5824.63	904.38	4981.60	460.49
BRCA	311.55	46.50	286.50	42.16	5305.85	1219.24	4816.61	361.22
KIRC	332.28	42.37	328.10	52.85	5117.83	881.77	4556.75	294.80
KIRP	313.48	49.64	303.22	41.14	4983.68	1136.93	4678.77	305.26
THCA	256.52	47.31	276.95	57.44	4016.13	948.02	4142.73	387.09
STAD	341.67	52.90	276.59	51.66	6773.41	928.62	4764.48	351.76
LIHC	352.74	68.99	256.24	85.31	7007.28	143.05	4522.08	400.83
Average	312.55	52.99	286.14	52.07	5449.80	1096.08	4585.78	384.72
St. Dev.	34.34	8.57	25.49	13.72	942.86	189.53	315.90	66.69

the mapping process of reads on the EBI interactome proteins needed similarity comparison of *fastq* files using BLASTx. We performed this validation by recycling the components of **Figure 2** for processing RNA-seq data as shown in **Figure 5**.

The workflow shown in **Figure 5** needed to be fed with BLASTx outputs. After mapping reads to their respective protein sequences in the interactome, both tumor and control *raw count* files were normalized ($UTCENG_{upper}$) according to their coding sequence size ($RPKM_{upper}$ step) and expression level using formula 2. The rest of the pipeline is as in **Figure 2** except for the last step of sorting by decreasing level of connection (PTTCS) and data storage in MongoDB (P2M).

The list of top-n connected up-regulated hubs is released as output data from the workflow, and stored in MongoDB (**Figure 5**) together with the patient's clinical data. These data can be formatted as a medical report by the JavaScript code within the web page according to the user request.

Considering the entropies of subnetworks of NSCLC up-regulated genes (x_1 = PRJNA320473) and PRAD (x_2 = PRJEB2449), the u_{obs} calculated with formula 3 with $\bar{x}_1 = 2.99475$ and $\bar{x}_2 = 1.66472$, respectively, as well as $SCE_1 = 10.31347$ and $SCE_2 = 6.70566$, respectively, was 5.00748. Since k was found to be 29.06411 ($\sim$29) for the sample sizes considered, the theoretical values of t for $p = 0.975$ and $p = 0.999$ were 2.045 and 3.396, respectively. Because $u_{obs} > t_{th}$, we rejected the null hypothesis of average equality for NSCLC and PRAD and concluded that the entropy of NSCLC was

TABLE 4 | Critical values of $RPKM_{upper}$ + Log2 for a probability density of $p = 0.975$ and vertex number of subnetworks of genes up-regulated in tumors.

Cancer	Critical value		Vertex number	
Type	Average	StDev	Average	StDev
PRAD	7359.58	1019.74	884.69	248.58
LUAD	7985.00	1105.05	946.58	219.40
LUSC	9325.45	1187.62	1244,35	232,65
BRCA	8398.81	1232.49	1087.20	240.74
KIRC	8335.51	561.98	1035.82	143.58
KIRP	8299.87	777.86	1014.35	206.89
THCA	7210.95	706.19	775.96	140.05
STAD	9173.28	1154.74	1264.93	249.58
LIHC	8398.81	1232.49	1235.50	294.36
Average	8276.36	997.58	1054.38	219.53
St. Dev.	707.07	251.53	171.09	50.27

TABLE 5 | The features of the linear regression between the entropy and the 5-years OS for $p = 0.975$.

Normalization method	Coef. Correl. (with BRCA + LUAD)	Coef. Correl. (without BRCA + LUAD)	Regression (without BRCA + LUAD)
GDC RPKM	–0.36	–0.55	–
GDC $RPKM_{upper}$	–0.68	–0.86 (**Figure 3A**)	$y = -0.0084x + 2.717$
GDC $RPKM_{upper}$ + LogNorm	–0.67	–0.85 (**Figure 3B**)	$y = -0.0090x + 4.473$
GDC $RPKM_{upper}$ + Log2	–0.69	–0.91 (**Figure 3C**)	$y = -0.0096x + 3.460$
GDC FPKM	–0.11	–0.13	–
TCGA FPKM-UQ*	–0.68	–0.64	$y = -0.004x + 2.507$
GDC FPKM-UQ	–0.71	–0.76 (**Figure 3D**)	$y = -0.0039x + 4.025$

*See *Conforte et al., 2019*.

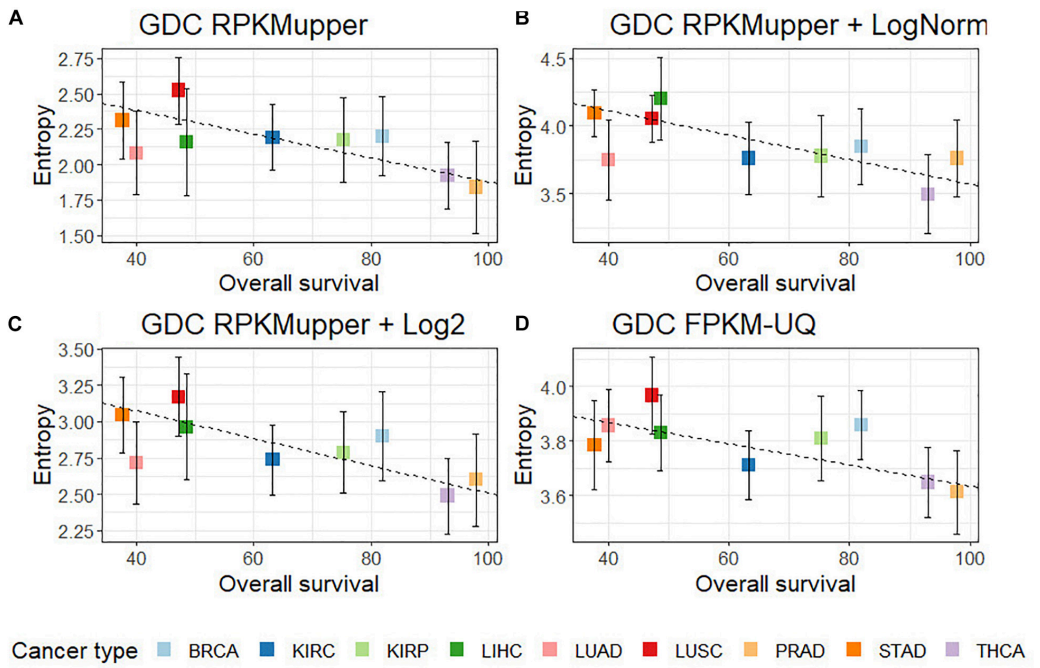

FIGURE 3 | Correlation of subnetwork entropies vs. 5-years OS for $p = 0.975$. **(A)** GDC $RPKM_{upper}$. **(B)** GDC $RPKM_{upper}$ + LogNorm. **(C)** GDC $RPKM_{upper}$ + Log2. **(D)** GDC FPKM-UQ.

significantly larger than that of PRAD. This result is in agreement with the negative correlations of **Figure 3** and validates the pipeline here presented.

As the methodology was validated, it could be used for the diagnosis of the top-n most connected proteins within the list of up-regulated genes in the tumor compared to the stroma. It is important to underline that the entropy was used only for the purpose of methodology validation.

A pipeline to identify the connection hubs is given in **Figures 6A,B**, where the purpose of PTTCS is to compare up-regulated genes to the list of vertex connections in the interactome to rank them in decreasing order of connection number in the output file. *A priori*, top-20 most connected proteins among the up-regulated genes of tumors should be enough to design a personalized treatment. However, this number depends on drug availability.

The comparison of the most relevant targets associated with the different normalization methods applied in this report is shown in **Table 6**. **Table 6** reports the number of tissues (# column) where the gene of a given protein (Acc column) was up-regulated among nine different tumors. For illustration, we only kept genes up-regulated in at least 70% of tumor samples of each cancer type (pink). The colors in the first column report for the targets that are common between different sections (A to E) of **Table 6** (turquoise is for the genes common to **Tables 6A–E**; blue is for the genes common to **Tables 6A,B,D,E**; yellow is for the genes common to **Tables 6A–C,E**; mallow is for the genes common to **Tables 6A,B,E**; and green is for the genes common to **Table 6D,E**).

Tables 6A–E show that the most relevant targets are largely shared among methods. In **Tables 6C,D**, target personalization according to the tumor was lower than in **Tables 6A,B,E**. Because of the larger average network size that it produced, the normalization with $RPKM_{upper}$ + Log2 (**Table 6E**) showed a larger targets number than TCGA RSEM-UQ and GDC $RPKM_{upper}$ (**Tables 6A,B**), similar to those of **Tables 6C,D**

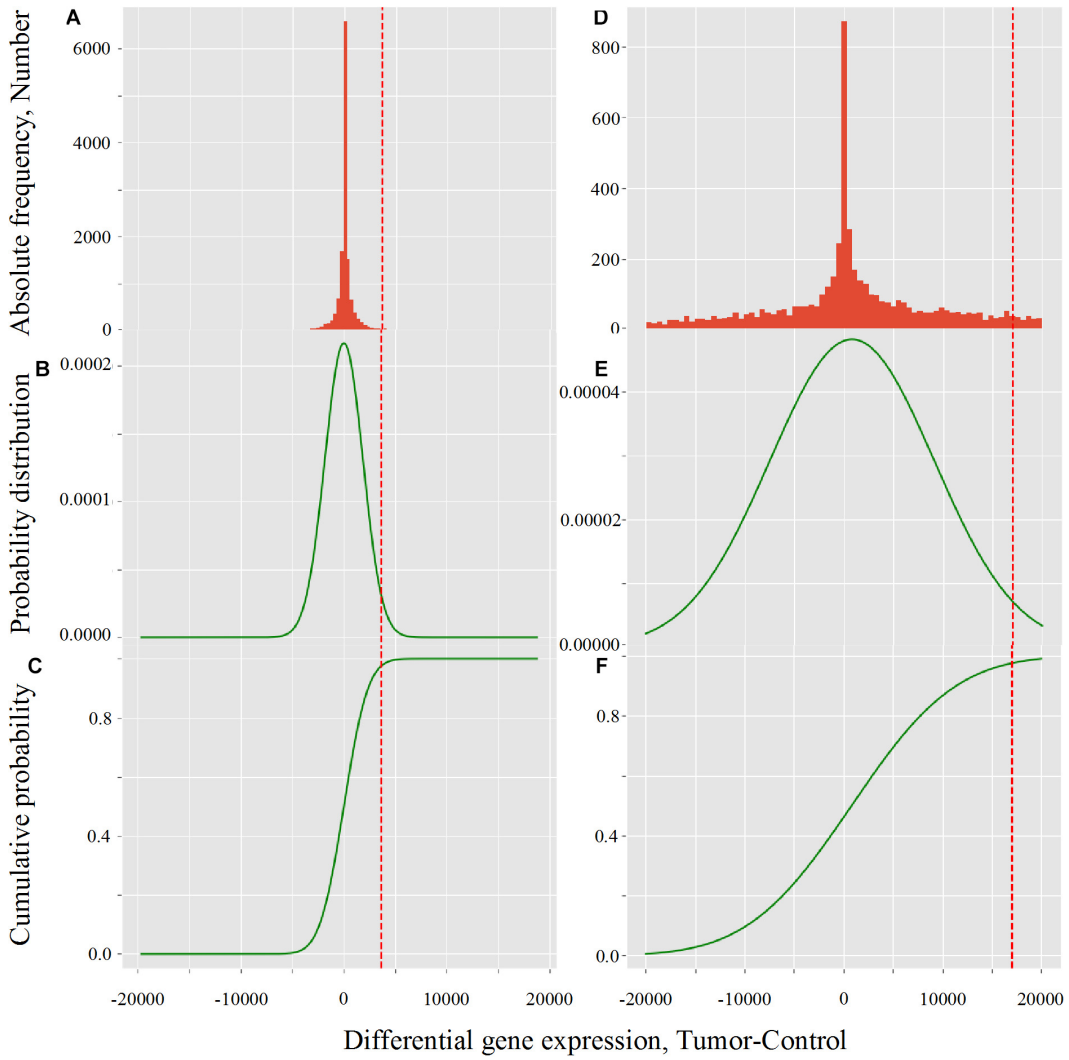

FIGURE 4 | Critical value calculation (red dot line) by CVC script in TCGA **(A–C)** and GDC **(D–F)** in LUSC (TCGA-22-4593 sample). **(A,D)** Histogram of observed differential gene expression distribution (tumor-control) of genes. **(B,E)** Function of density of probability. **(C,F)** Function of cumulated probability. The critical values were 3,633.8 and 17,042.9 for TCGA and GDC, respectively.

but with a larger level of tumor personalization. Because of the reasonable size of subnetworks and the best correlation relationship between entropy and the 5-years OS it produced, the successive processing through $RPKM_{upper}$ and Log2 normalization was considered here as the best compromise.

Scaling Analysis

The analysis of LUSC and PRAD over 45 patients showed that the scaling of pipeline processing is linear and perfectly predictable (**Supplementary Table 1** and **Figure 7A**). In addition, **Supplementary Table 1** shows that the entropy pipeline takes a systematically larger time to be completed for high entropy cancer types than for low entropy cancer types. This is also true for the hub diagnosis pipeline (PTTCS). A more careful analysis for 25 patients for LUSC, STAD, LIHC, on one hand, and PRAD, THCA, KIRC, on the other hand, showed that this assumption is statistically significant (**Figure 7B**). Considering the pipeline for hub diagnosis from BLASTx output, we found the time series 50, 94, 137, 187 and 53, 100, 145, 190, for PRAD and NSCLC, respectively. These differences were not significant, but suggest that this pipeline scales similarly to the PTTCS one.

Web Application

The web application implements the graphical interface that allows the user to interact with the forms and their respective accounts (i.e., private areas). As outline above, it is the server that runs Galaxy and hosts MongoDB that stores the up-regulated hubs and patient data introduced by the user, which are necessary to produce the medical record. The frontend includes a succession of forms for data introduction and a private area, which allow access to patient data whenever necessary with user's privileges.

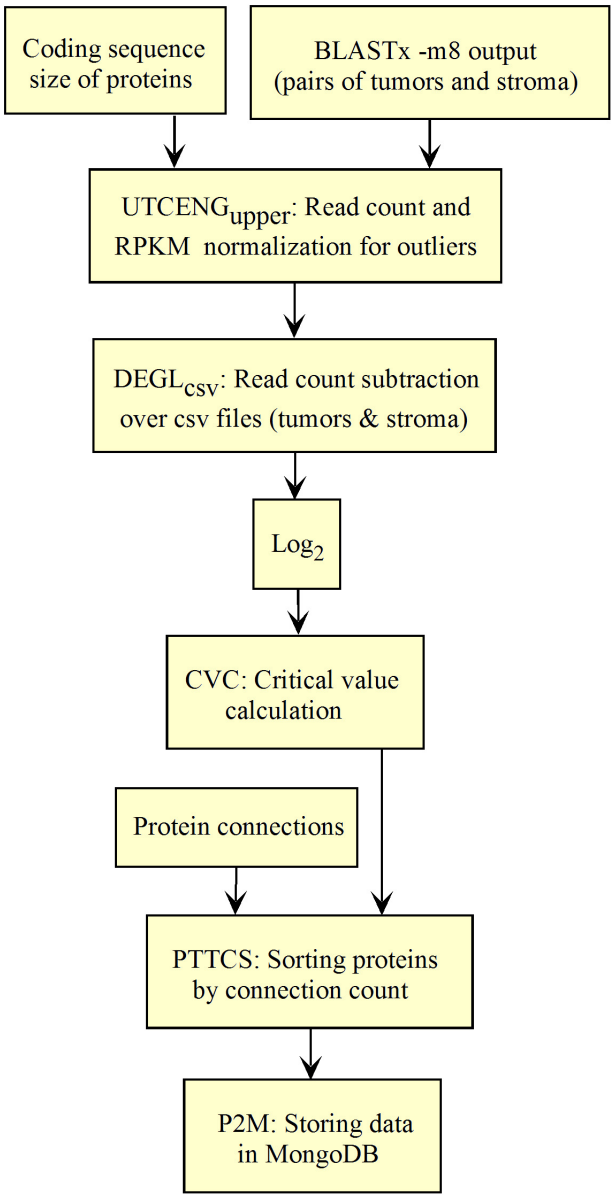

FIGURE 5 | Workflow for collections of BLASTx output processing.

User Private Area

The private area is the section accessed by the user after logging in (see the dashboard in **Figure 8**). The key advantages of a private area are that (i) the user may access their information any time, (ii) sections can be customized, with different levels of privileges, (iii) they can be customized according to business models (Blank and Dorf, 2012).

Dashboard

The dashboard (**Figure 8**) is the first page one sees when accessing the platform after login in from the *welcome* page. On the welcome page, users can register an account. The main goal of the dashboard is gathering all the essential information contained in the portal for the logged in user (e.g., forms to be submitted by users). Thus, users can either introduce the data of their patients or retrieve analysis reports, if they are physicians or administrate the platform, if they are system administrators.

We implemented a simplified version for small devices to fit their screen size and limit the system to the essential (**Figure 8A**). The user is informed when using the system on small devices, which is a benefit compared to Bootstrap. As a result of screen simplification, most of the information from the desktop version (**Figure 8B**) is omitted on small devices, which means that users must access the platform either from desktops or middle size devices (e.g., iPads) for a full-version.

Components

Components in Angular are a set of three types of files: CSS (appearance-related), TS (typescript, coding), and HTML (classic static page). A page is built from at least one component, which can independently interact with each other (Fain and Moiseev, 2018). From a software engineering viewpoint, this technology makes the pages more dynamic and faster, and its parts can be easily reused on other pages. The main components of the dashboard are the menu and central cards. The menu, located upward, displays basic and customized information eventually organized in options. The central cards, movable downward, display information and make them available as active links (e.g., a list of forms submitted by the user).

Protecting Confidential and Sensitive Information

Patients' data are confidential and require protection as stated by policies all over the world (e.g., *Health Insurance Portability and Accountability Act*, HIPAA for the United States). Thus, new users must first register and enter some basic information to gain access to the server.

Login

The *login card* (**Supplementary Figure 5**) is a standard login page. In the current frontend version, we are following a simple login system strategy. Essentially, the user must enter its e-mail and password as previously registered to log in and access the dashboard.

Since we are storing JWT locally, it is up to the user to decide when to log out. Normally, JWT expires after 15 min on a standard basis; we set the expiration time to 1 day. This approach avoids repeated login whenever the JWT expires.

Forms

In the current frontend version, we have two sets of forms: the patient main form (**Supplementary Figure 7**) and the outcome form (**Supplementary Figure 8**). The patient main form is expected to be sent alongside the patient samples, which is independent, while the outcome form is expected to be sent in case of death (for documentation).

All the information related to a patient is stored in different documents and is merged for display using a method called

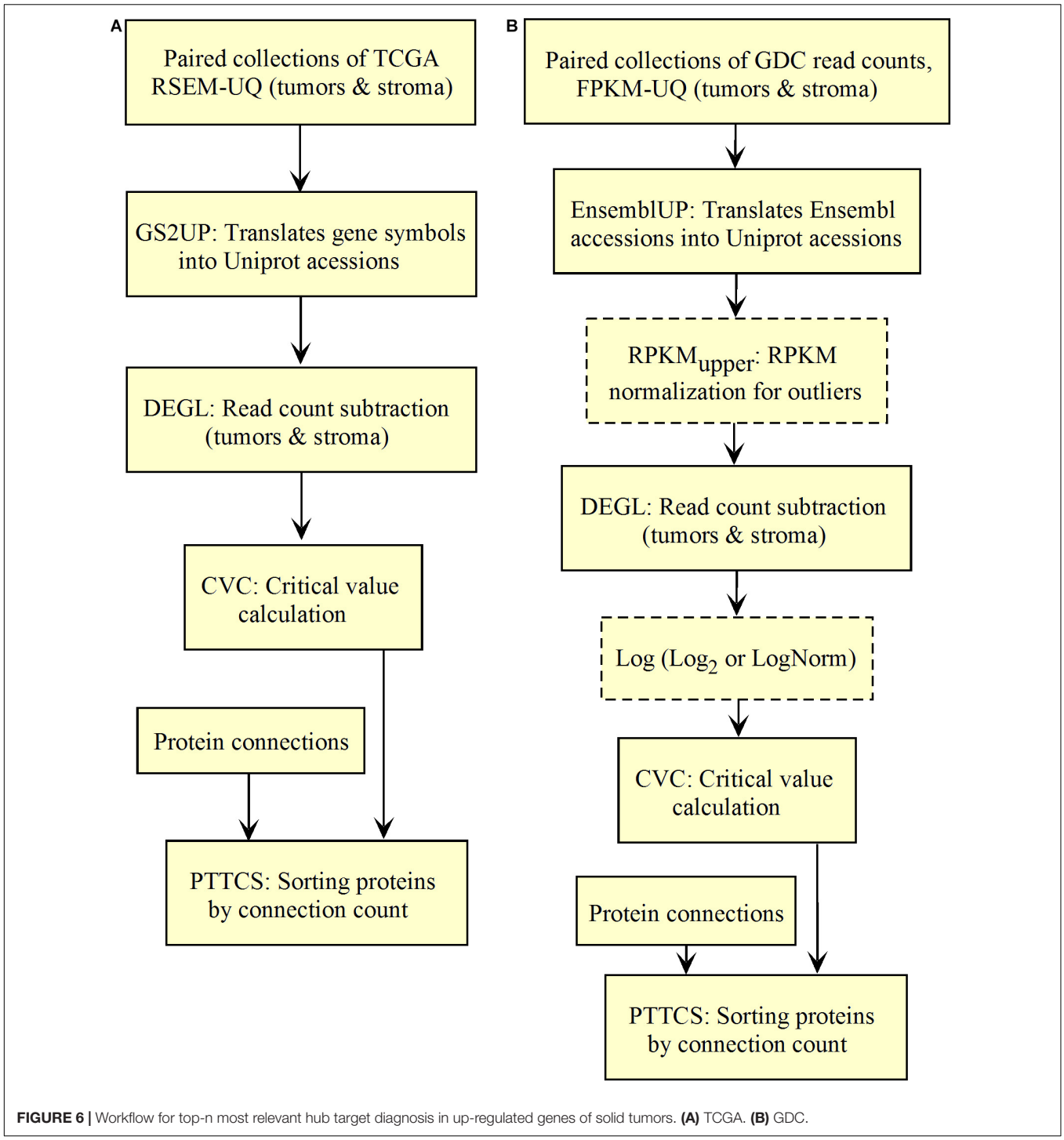

FIGURE 6 | Workflow for top-n most relevant hub target diagnosis in up-regulated genes of solid tumors. **(A)** TCGA. **(B)** GDC.

populate from Mongoose, which enable the information retrieval from other related documents.

Because of this design, we created a *header* form (**Figure 9**), whose function is to (i) collect encrypted patient id, (ii) provide a password for encryption (optional), and (iii) provide privacy-related options.

The form remains in contact with the server for validating information on the background, while the user is filling out the fields; most of the validations are done without communication with the server.

Sensitive information are entered on the first page and encrypted in a similar way to the data introduced through the header. Any form can be recovered from a list of links that are made available on the *movable card* on the dashboard.

Finally, a submission receipt is automatically generated upon form submission (see **Supplementary Figure 9**), which provides

TABLE 6A | Comparative pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to TCGA RSEM-UQ normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av	StDv
HSP90AB1	30.00	80.49	77.78	66.67	56.06	58.06	40.35	83.33	85.71	9	64.27	19.77
YWHAZ	72.00	85.37	93.33	88.89	33.33	29.03	28.07	56.67	36.73	9	58.16	27.26
FN1	26.00	56.10	44.44	77.78	87.88	51.61	85.96	53.33	32.65	9	57.31	22.30
ACTB	26.00	53.66	33.33	44.44	63.64	77.42	38.60	33.33	42.86	9	45.92	16.37
MYH9	14.00	43.90	37.78	33.33	48.48	25.81	43.86	80.00	53.06	9	42.25	18.55
VIM	32.00	12.20	11.11	11.11	93.94	96.77	47.37	23.33	28.57	9	39.60	33.73
RPL10	10.00	46.34	28.89	22.22	80.30	45.16	47.37	10.00	30.61	9	35.66	22.09
EEF1A1	12.00	21.95	24.44	22.22	68.18	22.58	78.95	16.67	22.45	9	32.16	23.93
PKM	NA	92.68	100.00	77.78	77.27	74.19	68.42	70.00	14.29	8	71.83	25.71
HSPA5	60.00	87.80	71.11	77.78	42.42	25.81	NA	43.33	48.98	8	57.15	20.79
HSPB1	NA	41.46	80.00	22.22	42.42	93.55	38.60	40.00	73.47	8	53.97	24.94
HSP90AA1	26.00	65.85	73.33	66.67	NA	48.39	17.54	70.00	57.14	8	53.12	20.97
CLTC	14.00	73.17	24.44	33.33	NA	45.16	12.28	26.67	28.57	8	32.20	19.57
SFN	NA	26.83	71.11	11.11	NA	16.13	NA	10.00	NA	5	27.04	25.52
LRRK2	NA	NA	NA	NA	36.36	54.84	71.93	NA	NA	3	54.38	17.79
VCAM1	NA	NA	NA	11.11	80.30	54.84	NA	NA	NA	3	48.75	35.00
EGLN3	NA	NA	26.67	NA	95.45	NA	NA	NA	NA	2	61.06	48.64
SYNPO	NA	NA	NA	NA	75.76	NA	10.53	NA	NA	2	43.14	46.13

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

TABLE 6B | Comparative pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to GDC $RPKM_{upper}$ normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av	StDv
PKM	16.67	91.23	100.00	87.5	78.873	77.42	80.36	92.59	60.00	9	76.07	25.05
FN1	33.33	70.18	50.00	100	90.141	51.61	87.50	59.26	62.00	9	67.11	21.79
YWHAZ	29.17	63.16	100.00	77.5	50.704	48.39	32.14	70.37	74.00	9	60.60	22.82
UBE2I	10.42	71.93	54.17	47.5	52.113	77.42	83.93	66.67	42.00	9	56.24	22.28
HSP90AB1	54.17	47.37	75.00	57.5	40.845	38.71	19.64	81.48	80.00	9	54.97	20.94
NPM1	60.42	47.37	43.75	50	73.239	35.48	26.79	25.93	58.00	9	46.77	15.79
CTNNB1	27.08	71.93	20.83	27.5	16.901	9.68	69.64	66.67	56.00	9	40.69	25.01
CDKN1A	12.50	52.63	16.67	10	50.704	51.61	71.43	11.11	26.00	9	33.63	23.08
ACTB	NA	36.84	37.50	75	61.972	77.42	33.93	44.44	80.00	8	55.89	19.86
HSP90AA1	27.08	29.82	72.92	70	NA	45.16	10.71	85.19	68.00	8	51.11	26.66
HSPB1	NA	21.05	77.08	40	32.394	77.42	16.07	29.63	64.00	8	44.71	24.70
RPL10	52.08	33.33	18.75	NA	78.873	32.26	28.57	14.81	44.00	8	37.84	20.55
VIM	NA	24.56	NA	12.5	92.958	96.77	35.71	11.11	16.00	7	41.37	37.50
SKP1	12.50	22.81	NA	45	16.901	NA	12.50	11.11	80.00	7	28.69	25.51
TSC22D1	16.67	36.84	NA	NA	12.676	9.68	69.64	NA	12.00	6	26.25	23.45
EGFR	NA	10.53	39.58	NA	74.648	19.35	NA	NA	NA	4	36.03	28.47

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

the user with the information necessary for future access to the forms submitted (see **Supplementary Figure 10**).

DISCUSSION

Galaxy Pipeline

In this report, we presented a workflow for processing RNA-seq data that allows the rational diagnosis of top connected hubs among genes that are up-regulated in tumors according to the non-tumoral peripheral area (stroma). The use of the stroma as a control to measure the malignant differential expression via RNA-seq has been recognized to be equivalent to the use of healthy tissues for this purpose (Finak et al., 2006). Of course, many factors may promote cancer such as chemicals, radiation as well as genetic defects in reparation and replication molecular machinery. To gain inside into such a complex problem as a molecular approach of cancer together with a still-evolving protocol of RNA-seq treatment regarding normalization procedure or error rate (Li et al., 2020), a robust measure was

TABLE 6C | Comparative pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to GDF FPKM-UQ normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av.	StDv
HSP90AB1	81.25	90.91	87.23	86.11	77.46	74.19	64.29	92.59	94.00	9	83.12	9.78
TP53	81.25	75.76	68.09	66.67	78.87	87.10	85.71	66.67	54.00	9	73.79	10.78
TRAF2	45.83	66.67	70.21	83.33	84.51	87.10	35.71	100.00	86.00	9	73.26	20.95
YWHAZ	39.58	81.82	100.00	83.33	66.20	80.65	60.71	66.67	74.00	9	72.55	17.05
PPP1CA	58.33	72.73	78.72	97.22	46.48	61.29	83.93	55.56	74.00	9	69.81	15.84
TRIM27	79.17	90.91	42.55	75.00	52.11	64.52	60.71	66.67	70.00	9	66.85	14.39
FN1	39.58	60.61	42.55	97.22	91.55	54.84	87.50	55.56	32.00	9	62.38	24.08
GRB2	35.42	39.39	36.17	72.22	66.20	93.55	64.29	66.67	66.00	9	59.99	19.39
MAPK6	85.42	69.70	93.62	66.67	45.07	45.16	42.86	37.04	48.00	9	59.28	20.37
GOLGA2	85.42	57.58	59.57	63.89	56.34	67.74	44.64	48.15	46.00	9	58.81	12.75
SNW1	45.83	72.73	68.09	52.78	50.70	58.06	75.00	44.44	58.00	9	58.40	11.28
VCAM1	25.00	72.73	44.68	47.22	88.73	67.74	26.79	66.67	32.00	9	52.40	22.59
CDC37	39.58	27.27	29.79	41.67	70.42	54.84	51.79	74.07	74.00	9	51.49	18.31
MYC	70.83	33.33	63.83	19.44	85.92	77.42	32.14	48.15	30.00	9	51.23	23.95
IKBKE	NA	81.82	76.60	66.67	45.07	80.65	64.29	62.96	58.00	8	67.01	12.44
OTUB1	35.42	69.70	91.49	80.56	21.13	NA	71.43	40.74	60.00	8	58.81	24.24
MDFI	45.83	84.85	87.23	25.00	16.90	NA	80.36	77.78	34.00	8	56.49	29.16
EGFR	27.08	39.39	78.72	NA	95.77	77.42	62.50	44.44	24.00	8	56.17	26.37
HSPB1	NA	45.45	42.55	50.00	40.85	64.52	23.21	37.04	72.00	8	46.95	15.43
YWHAB	22.92	39.39	40.43	77.78	NA	NA	39.29	62.96	72.00	7	50.68	20.30
MAP1LC3B	NA	27.27	NA	27.78	50.704	80.65	60.71	29.63	30.00	7	43.82	20.87
KDM1A	72.92	42.42	36.17	41.67	NA	NA	28.57	40.74	40.00	7	43.21	13.94
WDYHV1	41.67	60.61	65.96	77.78	NA	NA	NA	33.33	66.00	6	57.56	16.73
KSR1	NA	30.30	19.15		84.507	NA	NA	18.52	20.00	5	34.50	28.37

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

needed. We found this measure in the degree entropy. Entropy offers the benefit to be independent of sample size. In this report, we calibrated our approach by reference to OS, but after an optimization round for the treatment of RNA-seq data, other factors could be taken into account to understand how they interact with the signaling network complexity.

Normalization of raw read counts account for (i) within-sample effects induced by factors such as coding sequence size (Oshlack and Wakefield, 2009), GC-content (Risso et al., 2011), (ii) between-sample effects such as sequencing depth (total number of molecules sequenced) (Robinson and Oshlack, 2010), and (iii) batch effect (Tom et al., 2017). As underlined by Evans et al. (2018), "normalization methods perform poorly when their assumptions are violated." Thus, the exercise is to "select a normalization method with assumptions that are met and that produces a meaningful measure of expression for the given experiment."

Following these recommendations, we must first consider that the purpose of our approach is to list the top-n most relevant target among subnetworks of genes that are up-regulated in tumor samples compared to their controls. Consequently, the complexity of the up-regulated gene subnetwork must be coherent with the 5-years OS. Indeed, our supporting hypothesis is that the complexity of the malignant subnetwork or the number of times that the malignant subnetwork can reorganize itself after perturbation is in line with its information content, i.e., its Shannon entropy. This is the reason why it makes sense to optimize the normalization process for maximizing the coefficient of correlation between entropy and 5-years OS. We aimed to diagnose the subnetwork complexity because it is correlated to the 5-years OS and this is important for therapy's success (whatever being performed with drugs or biopharmaceuticals) in the context of a personalized approach of oncology.

The PDF and CDF functions of the Python's scipy package allowed the calculation of the critical values given the density of probability of non-differentially expressed genes. These distribution are rather similar regardless of the RNA-seq considered for a given normalization process. These genes are thousands while the up-regulated ones are hundreds, which makes critical value determined in this way rather precise and reproducible. Concerning the statistical significance of the method we applied, one has to say that we face a classification problem. In such circumstances, one usually looks for the optimization between false positive and false negative rates. However, when dealing with medical purpose, one has to look to bias the classification process toward the minimization of false-positive rate to reduce toxic drug collateral effects to patients that would derive from hubs still expressed at a significant level in the stroma (this consideration does not concern drug toxicity due to off-target effects). There is a compromise between minimizing the false positive rate and the availability of hub targets for

TABLE 6D | Pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to successive processing through $RPKM_{upper}$ and LogNorm normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av	StDv
HSP90AB1	83.33	73.68	87.50	65.00	76.06	74.19	42.86	96.30	98.00	9	77.44	16.92
TP53	79.17	84.21	66.67	65.00	74.65	83.87	73.21	77.78	66.00	9	74.51	7.42
YWHAZ	47.92	70.18	100.00	87.50	63.38	67.74	46.43	77.78	86.00	9	71.88	17.97
TRAF2	47.92	47.37	72.92	85.00	84.51	87.10	26.79	96.30	86.00	9	70.43	23.85
FN1	43.75	70.18	52.08	100.00	90.14	54.84	87.50	59.26	62.00	9	68.86	19.43
PPP1CA	56.25	68.42	79.17	95.00	46.48	51.61	67.86	70.37	76.00	9	67.91	14.97
GRB2	43.75	43.86	39.58	87.50	71.83	87.10	46.43	81.48	94.00	9	66.17	22.45
MAPK6	85.42	66.67	95.83	65.00	49.30	58.06	42.86	55.56	72.00	9	65.63	16.92
GOLGA2	87.50	61.40	70.83	65.00	54.93	51.61	37.50	62.96	80.00	9	63.53	15.00
TRIM27	79.17	68.42	56.25	62.50	54.93	45.16	53.57	77.78	68.00	9	62.86	11.47
SNW1	39.58	54.39	72.92	47.50	43.66	45.16	44.64	74.07	82.00	9	55.99	15.94
HSCB	56.25	56.14	68.75	47.50	74.65	29.03	51.79	59.26	58.00	9	55.71	12.95
VCAM1	31.25	47.37	47.92	45.00	87.32	67.74	30.36	81.48	46.00	9	53.83	20.48
CDC5L	43.75	43.86	68.75	52.50	53.52	48.39	16.07	77.78	78.00	9	53.62	19.49
MYC	72.92	40.35	66.67	10.00	85.92	77.42	30.36	55.56	36.00	9	52.80	25.20
OTUB1	37.50	54.39	83.33	77.50	15.49	22.58	62.50	40.74	54.00	9	49.78	23.03
IKBKE	10.42	75.44	60.42	72.50	32.39	64.52	76.79	22.22	22.00	9	48.52	26.46
REL	35.42	29.82	45.83	52.50	26.76	32.26	25.00	70.37	16.00	9	37.11	16.58
EGFR	35.42	57.89	79.17	NA	95.77	67.74	53.57	51.85	44.00	8	60.68	19.55
MDFI	47.92	80.70	87.50	30.00	15.49	NA	82.14	81.48	26.00	8	56.40	29.80
GABARAPL2	37.50	22.81	NA	20.00	11.27	25.81	30.36	59.26	80.00	8	35.87	22.86
YWHAB	20.83	31.58	22.92	75.00	NA	16.13	32.14	22.22	28.00	8	31.10	18.57
LRRK2	NA	49.12	NA	15.00	83.10	87.10	89.29	22.22	14.00	7	51.40	34.87
LNX1	50.00	61.40	66.67	42.50	NA	NA	48.21	29.63	44.00	7	48.92	12.32
MAP1LC3B	NA	35.09	NA	17.50	42.25	80.65	46.43	62.96	52.00	7	48.13	20.16

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

therapy. A larger p-value ($p > 0.025$) would release a larger list of up-regulated genes with more hub targets; a larger list of potential drugs for the case under consideration, but also a larger probability of toxic effects on the stroma. In contrast, lower p-value ($p < 0.025$) will minimize toxic effect of therapy to patient, but would also decrease the number of potential hubs for therapy. Of course, this consideration neglects the tissue specific expression of genes and a gene that is up-regulated in a tumor compared to its stroma could also be up-regulated in another tissue, on a normal basis. Here, we neglected this issue, but it is possible to preferentially target tumors through nanoparticle therapy or by local application.

As pointed out by Abbas-Aghababazadeh et al. (2018), it is possible that some of the estimated *latent factors* are not technical artifacts but rather represent true biological features reflected in the data. The correction of these latent factors may introduce unwanted biases. Here, we did not want to stabilize the subnetwork size variance (Smyth, 2004; Cloonan et al., 2008; Love et al., 2014; Holmes and Huber, 2019) because we believe that it is part of the challenge. One cannot exclude the possibility of network size varying among samples according to the specificities of genome deregulation proper to a given tumor. Despite commonalities that were recognized between tumors of the same cancer type, many features such as gene demethylation, copy numbers, somatic crossing over, and chromosome karyotype contribute to the specificity of the molecular phenotype of a tumor and it is the correct diagnosis of these specificities that can make the difference in terms of patient benefits (Duesberg et al., 2005; Ozery-Flato et al., 2011; Ogino et al., 2012; Grade et al., 2015; Bloomfield and Duesberg, 2016; Ye et al., 2018; Xia et al., 2019).

According to the considerations just outlined, the size of the malignant subnetwork is also important because it directly affects the number of targets available for therapy. The size of the malignant subnetworks also depends on the normalization process. There is a tradeoff between the size of the malignant subnetwork and the level of tumor personalization that is effectively reported by the top-n targets as a result of the normalization process. From our perspective, the normalization corresponding to GDC FPKM-UQ and $RPKM_{upper}$ + LogNorm generate subnetworks that are too large since they represent as much as 20% of the human proteome (>4,000 genes). By contrast, subnetworks produced by GDC $RPKM_{upper}$ + Log2 normalization account for between 2 and 5% of the human proteome, which seems to be more realistic (Danielsson et al., 2013; Malvia et al., 2019).

The target lists that we found with the various normalization methods presented here were consistent among one another and

TABLE 6E | Pattern of distribution for the most relevant targets among solid tumors of nine cancer types according to successive processing through $RPKM_{upper}$ and Log2 normalization.

Acc	PRAD	LUAD	LUSC	BRCA	KIRC	KIRP	THCA	STAD	LIHC	#	Av	StDv
HSP90AB1	77.08	66.67	85.42	94.00	64.79	64.52	35.71	96.30	94.00	9	75.39	19.73
YWHAZ	41.67	70.18	100.00	82.00	60.56	67.74	46.43	74.07	82.00	9	69.41	18.21
TP53	72.92	70.18	64.58	60.00	66.20	74.19	58.93	70.37	60.00	9	66.37	5.86
FN1	41.67	70.18	52.08	62.00	90.14	54.84	87.50	59.26	62.00	9	64.41	15.92
NPM1	72.92	68.42	72.92	60.00	85.92	41.94	51.79	59.26	60.00	9	63.68	12.98
YWHAG	52.08	40.35	87.50	68.00	52.11	80.65	10.71	59.26	68.00	9	57.63	22.91
CDC37	12.50	28.07	25.00	86.00	67.61	38.71	44.64	81.48	86.00	9	52.22	28.54
MAPK6	79.17	43.86	87.50	54.00	22.54	29.03	23.21	44.44	54.00	9	48.64	23.06
MYH9	39.58	40.35	37.50	62.00	43.66	12.90	23.21	81.48	62.00	9	44.74	20.98
PKM	14.58	73.68	41.67	22.00	47.89	32.26	78.57	51.85	22.00	9	42.72	22.68
HSPB1	NA	49.12	89.58	92.00	59.15	93.55	50.00	37.04	92.00	8	70.31	23.74
RPL10	66.67	56.14	45.83	74.00	88.73	54.84	44.64	NA	74.00	8	63.11	15.41
OTUB1	27.08	42.11	83.33	78.00	NA	16.13	48.21	66.67	78.00	8	54.94	25.35
YWHAB	14.58	35.09	39.58	84.00	NA	35.48	26.79	77.78	84.00	8	49.66	27.82
YBX1	31.25	21.05	66.67	56.00	49.30	41.94	NA	70.37	56.00	8	49.07	16.96
MYC	66.67	15.79	56.25	28.00	73.24	67.74	NA	40.74	28.00	8	47.05	21.80
EGFR	16.67	40.35	68.75	28.00	95.77	58.06	30.36	NA	28.00	8	45.75	26.53
CSNK2A1	20.83	21.05	85.42	30.00	11.27	41.94	NA	51.85	30.00	8	36.54	23.49
GRB2	NA	15.79	16.67	64.00	32.39	87.10	NA	48.15	64.00	7	46.87	26.75
TUBA1A	NA	47.37	10.42	18.00	59.15	54.84	71.43	NA	18.00	7	39.89	24.06
LRRK2	NA	38.60	NA	NA	57.75	67.74	76.79	NA	NA	4	60.22	16.38
VCAM1	NA	10.53	10.42	NA	84.51	61.29	NA	NA	NA	4	41.69	37.27
LZTS2	NA	36.84	NA	NA	NA	29.03	71.43	NA	NA	3	45.77	22.56
EGLN3	NA	NA	22.92	NA	83.10	NA	NA	NA	NA	2	53.01	42.56

The numbers in the table represent the proportion (%) of tumors of a given cancer type that showed the gene among the top-20 most connected proteins of the subnetwork of up-regulated genes. The pink color concerns up-regulated genes in at least 70% of tumor samples of each cancer type.

with that of Conforte et al. (2019). The normalization method corresponding to the best compromise according to subnetwork size, correlation, and the target list was $RPKM_{upper}$ + Log2, and it is that method that was, therefore, kept for new sample analyses.

To be coherent with former studies, we included LUAD and BRCA, however, these two cancer types discredited the analyses for two obvious reasons: (i) In the case of LUAD, the samples of *raw counts* did not match those of FPKM-UQ, which prohibit direct comparison between both datasets and raised the question why FPKM-UQ normalization was not performed on a large proportion of *raw counts* files and why, on the other hand, other samples were taken into account in the FPKM-UQ processing. This discrepancy may explain why LUAD does not match the regression line in GDC $RPKM_{upper}$ + Log2, while it does in GDC FPKM-UQ; (ii) In the case of BRCA, after filtrating samples for matching between *raw counts* and FPKM-UQ, the total sample size was less than 20, which is not sufficient for statistical significance given subtype heterogeneity. BRCA is composed of four subtypes whose 5-years OS varies between 70 and 82%. **Figure 3** shows that depending on sampling, BRCA could very well match the linear regression.

The relevance of inhibiting hub of connections has been proven mathematically by Albert et al. (2000) and its benefit for patients has been confirmed by Conforte et al. (2019) through Shannon entropy analysis. The negative correlation found between the subnetwork entropy and the 5-years OS is in agreement with the results obtained later on from the modeling of basins of attraction in BC with Hopfield network (Conforte et al., 2020). This study revealed that five tumor samples converged toward the basin of attraction associated with control samples instead of the tumor ones. Those samples were associated with a good prognosis, initial stages of tumor development, and four of them presented the smallest subnetwork entropy among the dataset of 70 tumor samples under study.

As the research concept has been validated through different approaches, the workflow presented here was built with the aim of automating the analysis, which will allow its translation to the medical context. With that concern, the larger time needed for entropy and PTTCS pipelines to be completed when analyzing high entropy cancer types compared to the processing time spent with low entropy cancer types suggests a positive relationship between subnetwork complexity and their processing time. If confirmed, this observation means that the computation model, presented here, reproduces a main biological feature of cancer that is the larger complexity associated with subnetwork of up-regulated genes in aggressive tumors. In any case, the difference in the processing time of the PTTCS pipeline for high and low entropy cancers was not large (∼50 s for 25 patients).

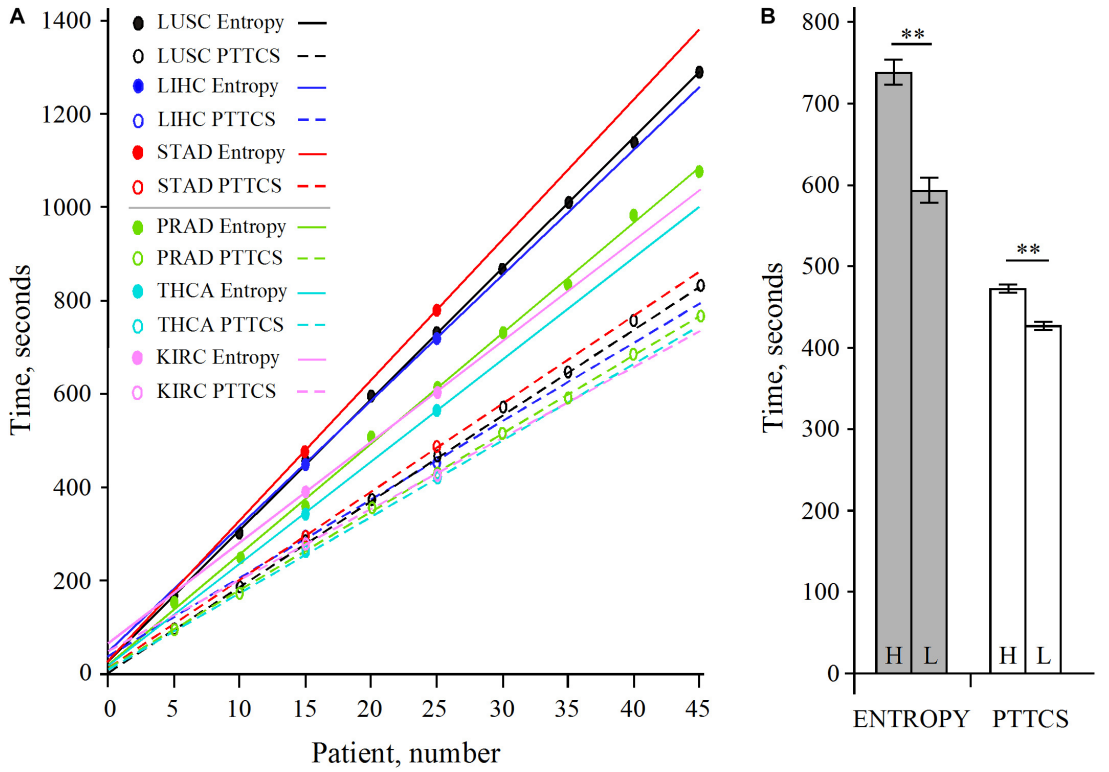

FIGURE 7 | Scaling of pipeline from **Figure 2B** (entropy) and **Figure 6B** (PTTCS) using GDC read counts (see data in **Supplementary Table 1**). **(A)** Linear scaling for 5–45 patients in LUSC and PRAD as well as 15 and 25 patients for STAD, LIHC, THCA, and KIRC. **(B)** Statistical analysis of scaling for high entropy (H) cancer (LUSC, LIHC, STAD) and low entropy (L) cancer (KIRC, PRAD, THCA) for entropy (gray) and PTTCS (white) pipelines. **Significant at $\alpha \leq 0.01$ for $k = 4$ degrees of freedom. The horizontal bars are for the standard error of the mean (SEM).

We believe that our strategy will contribute constructively to cancer treatment because the molecular phenotype of a cell is directly connected to its genetic alterations, which is not necessarily the case for genomic alterations. Genomic alterations allow a diagnosis based on probabilistic data obtained with large patient cohorts. By contrast, the molecular phenotype portraits the cell or the genomic disease and points to proteins that should be targeted in the first instance to disrupt malignant phenotypes while affecting the healthy one the least possible.

The phenotype approach also reflects which genes that malign cells most need to maintain themselves in the tissue given its selective constraints. In any pathogenic relationship, one distinguishes between *primary* and *secondary determinants* of the disease (Yoder, 1980). The primary determinants are those that make the relationship compatible (qualitative) and the secondary determinants are those that deal with its quantitative expression (virulence). Thus, the question to deal with, in the case of cancer, is to target primary determinants. When considering gene expression, one may reason that the heterogeneity is something related to secondary determinants (it is not because a cell is mutating that the new mutations are worse than the previous ones). Actually, it has been well described that a tumor developed by the accumulation of mutations in a small number of key oncogenes or suppressor genes in stem cells and that the probability of this event to occur is very low (Hornsby et al., 2007; Belikov, 2017). Thus, there is a difference between these primary mutations that allow the tumor to establish itself and the secondary ones that may affect its aggressiveness. On the same line, when one sequences the mRNA of a tumor area, one takes the gene expression profile of many cells into account. By consequence, secondary mutations promoting or inhibiting a given gene in different cell lineages inside the same tumor compensate themselves. By contrast, those genes that are key to maintain a malignant cell line will be positively selected to remain up-regulated in most cells and, therefore, if one detects a gene that is up-regulated in a tumor by comparing its expression level with the surrounding stroma, it means that it is essential for malign cell survival.

Considering the number of hubs to target, the results obtained by Conforte et al. (2019) suggest 3–10, on average. Other authors already suggested such complex mixes (Calzolari et al., 2007, 2008; Preissner et al., 2012; Hu et al., 2016; Antolin et al., 2016; Lu et al., 2017). Three to ten specific drugs may appear a small number to control such a complex disease as cancer, but the cell death induction may be explained by a cascading effect, which is larger when targeting hubs as suggested before (Carels et al., 2015a; Barabási, 2016; Tilli et al., 2016; Conforte et al., 2019). According to Conforte et al. (2019), this cascading effect would be inversely proportional to the tumor aggressiveness. The pitfall is that the number of specific drugs for hub targets

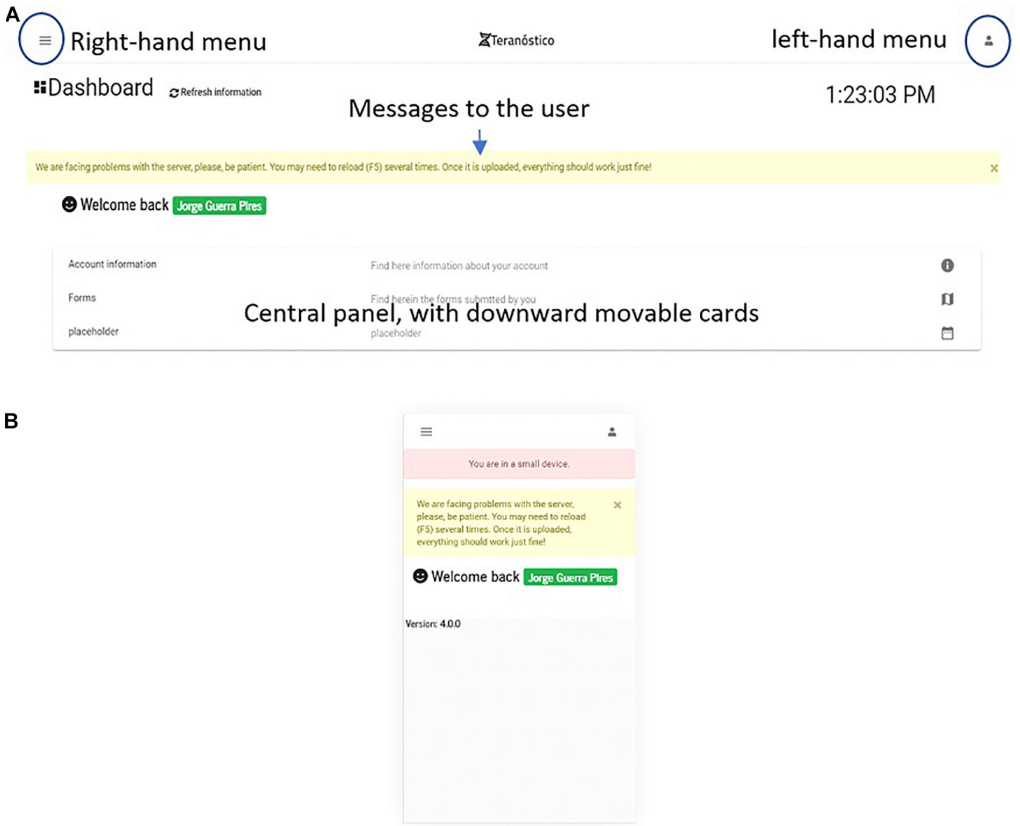

FIGURE 8 | Dashboard on **(A)** a desktop and **(B)** an extra small devices (Galaxy S5).

that are approved by FDA is still very small (Antolin et al., 2016). While new drugs and biopharmaceuticals or products of other strategies continuously appear, key targets remain the same. Some are highly personalized and often secondary while others are constant across tumor types or within a tumor type; these last targets play, in most likelihood, a primary role in the disease and it is essential to diagnose them (even if only for their prognostic value). In addition, nothing prohibits the combination of specific drugs with cytotoxic or hormonal treatments (Nikanjam et al., 2016). The idea is to improve as much as possible the rational drug use to maximize the patient benefit. Many patients are dying from the toxic collateral effect of the chemotherapy; it would be a great success if the use of specific drugs in a standard therapy protocol could enable to decrease the dose of cytotoxic drugs and improve the therapy acceptance by patients in some specific cases in the context of theranostics. For this kind of exercise, an automated pipeline is needed and a clinical trial testing the validity of hubs as potential molecular targets is urgent.

The replication number that can be done for RNA-seq is another limitation given the still high cost of this technology. Thus, analyses as the one described in our manuscript are expected to be done only once per time in a time series for each patient. According to Barabasi's theory (Barabási, 2016), hubs with the same connection rate are expected to have the same disarticulation effect on the signaling network. On a clinical basis, p-values (here critical value) may be adapted to the specific case of each patient. On the same line of reasoning, our methodology can be easily adapted taking into account more powerful bioinformatics tools and statistical analysis, but this issue is beyond the scope of this report. For such a methodology improvement, we believe that entropy is a good measure because it is universal, robust, and not dependent on sample size. Different combinations of normalization and statistical analyses as those reported by Li et al. (2020) can be compared in the same framework we presented here and in Conforte et al. (2019), by looking at how they may maximize the correlation coefficient of the negative relationship between entropy and OS. Of course, this depends on accepting the hypothesis that more aggressive tumors have more complex signaling networks, but again, this statement has been repeatedly claimed by several authors worldwide and along several years (Teschendorff and Severini, 2010; van Wieringen and van der Vaart, 2011; Breitkreutz et al., 2012; West et al., 2012; Banerji et al., 2015). If this hypothesis is true, the negative correlation between entropy and OS may serve as a calibration to study the optimization of RNA-seq methodologies and the influence of other factors in cancer development and dynamics.

Cancer is a genomic disease that affects DNA replication checkpoints through mutations of key oncogenes and suppressor genes (Lee and Muller, 2010). There are ten main hallmarks

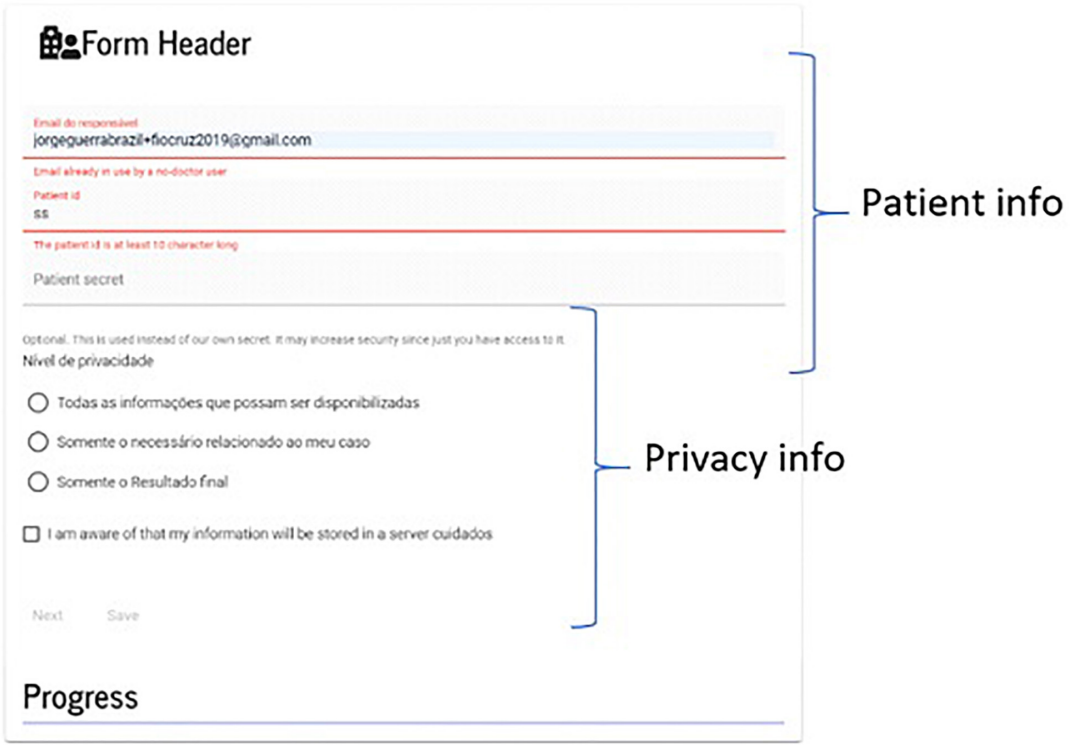

FIGURE 9 | Form header. A user is warned when leaving without saving the form. The patient password is encrypted and kept on the server using a specialized type of file; nonetheless, users can choose their own passwords.

for cancer from which uncontrolled division is the key one (Hanahan and Weinberg, 2011). When the disease is taken at a late stage, it may have spread in the body through metastasis and secondary tumors may have different molecular profiles. In such late tumor stages, an approach of cancer therapy only based on personalized oncology would in most likelihood be unsuccessful (Ashdown et al., 2015). However, specific drugs could increase the patient benefit by supplementing traditional therapies based on cytotoxic drugs. As a consequence, the maximum benefit of a personalized oncology approach of solid tumor therapy based on a molecular phenotype diagnosis is in the early stages of malign cell multiplication. Despite its limitations, the phenotype approach of molecular diagnosis proposed here is needed for rational drug (or biopharmaceutical) therapy to maximize patient benefit.

At the moment, the methodology and the web site that we described here can be assimilated to *laboratory developed tests* (LDT). It is notorious that LDT for being a type of *in vitro* diagnostic test designed, manufactured, and used within a single laboratory is poorly supported by oncologists (8%) and pathologists (12%) because of the legitimate fair of innovation. Biomarkers and CDs strongly depend on the regulation by official organizations for their acceptance by health decision-makers (Novartis, 2020). However, barriers by regulation are no reason to stop the innovation necessary for progress. Otherwise, regulation fails with its purpose of protecting lives (see Carels et al., 2020 for a review).

Web Application

System biology has gained considerable attention in medical sciences in the last decade thanks to the ever-increasing computer power. However, system biology models can be tricky to use or to interpret by non-experts in modeling. A recurrent question is how to integrate models into the physician daily lives such that they could best participate in their decision-making process. One potential solution, which seems to be the predominant one on the current state of the art, is by packing algorithms into software bundles and to make them available by user-friendly interfaces, such that little, or even no, expertise is required to use them. This is the paradigm we followed in this report.

The power and diversity of Angular programmed with TypeScript enable to expand the functionalities of the prototype proposed here in future versions, including the implementation of heavy calculations on the frontend side.

We chose MongoDB for storing genetic and medical records even if Galaxy has its own database system (postgreSQL). Our choice of MongoDB was motivated by the care of keeping coherence with MEAN stack, and also because of the power of MongoDB for Big Data storage. In addition, MongoDB is a non-relational database (NoSQL), which allows the storage of data in different formats within the same database.

Our implementation of online forms offers the possibility of creating new functions such as data validation. Data can be validated by comparing frontend to backend information through the database and making sure, for instance, that an

entered e-mail does not already belong to someone else already registered in the system.

Finally, one common concern on web-programming is to minimize client communication with the server to maximize performance. For such purpose, we implemented a process of form validation on the frontend side. Since we are using FormBuilder (see for more details Fain and Moiseev, 2018), there are a set of built-in validation routines, and the possibility to easily create customized validation, thus any specific demand concerning data validation can be handled on future versions using the current source code.

CONCLUSION

In a successive set of publications, we developed a rational methodology for the diagnosis of connection hubs among up-regulated genes of malignant subnetworks. This strategy is an application of graph theory, whose relevance has been mathematically proven by Albert et al. (2000). The inference of this theory into biological systems performed by Carels et al. (2015a) has been successfully validated on malignant cells by Tilli et al. (2016) and extended to tumor tissues by Conforte et al. (2019).

Here, in a translational oncology effort, we outlined a workflow that automated that research and allows its application to a large set of RNA-seq data to interact with public entities of the oncological sector, such as pharmaceutical companies, hospitals, diagnostic laboratories, public health care systems, and insurance groups around the world.

We belief that innovation in new translational solutions, like the one outlined here, is an imperative attribute of research centers; however, other agents such as (i) pharmaceutical companies may certainly help these initiatives with their experience concerning regulation, market barriers, financial support and (ii) startups whose processing speed and innovation potential were already well-documented (Blank and Dorf, 2012).

Herein, we aimed at transcending basic cancer inferences to bring a solution for clinical applications on a global scale.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://galaxy.cdts.fiocruz.br/.

AUTHOR CONTRIBUTIONS

JP contributed for the development of the web application and wrote the corresponding sections. GS built the galaxy environment. TW wrote the Phython script. AC did the R analysis and contribute to the pipeline logic. DP prepared the medical forms. FS contributed with manuscript writing. NC wrote the Perl scripts, set the pipeline logic up, and wrote and managed the manuscript writing. All the authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by a grant (E-26/290.077/2017 - 227190) to NC and a fellowship (E-26/260.046/2019 - 242550) to JP from Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.624259/full#supplementary-material

Supplementary Figure 1 | Adaptative display according to device screen size (source: Fain and Moiseev, 2018).

Supplementary Figure 2 | Flowchart of main form filling for exam request.

Supplementary Figure 3 | Flowchart of outcome form filling.

Supplementary Figure 4 | Authentication process. A *Guard* function double-check a user's requisition and if access conditions are met the user is allowed to see the content of a requested page (green arrows). By contrast, if something went wrong (e.g., token expired), the access is denied (red arrow).

Supplementary Figure 5 | Login card.

Supplementary Figure 6 | Example of main form options.

Supplementary Figure 7 | Example of outcome form being implemented.

Supplementary Figure 8 | The page #1 of the main form given as an example.

Supplementary Figure 9 | Receipt of main form submission.

Supplementary Figure 10 | Form after retrieval from the Dashboard (the entire form does not fit the page).

Supplementary Table 1 | Scaling of pipeline from **Figure 2B** (entropy) and **Figure 6B** (PTTCS) using GDC read counts (see **Figure 7**).

REFERENCES

Abbas-Aghababazadeh, F., Li, Q., and Fridley, B. L. (2018). Comparison of normalization approaches for gene expression studies completed with highthroughput sequencing. *PLoS One* 13:e0206312. doi: 10.1371/journal.pone.0206312

Afgan, E., Baker, D., Batut, B., Beek, M., Bouvier, D., Čech, M., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.* 46, W537–W544.

Albert, R., Jeong, H., and Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature* 406, 378–382. doi: 10.1038/35019019

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106.

Antolin, A. A., Workman, P., Mestres, J., and Al-Lazikani, B. (2016). Polypharmacology in precision oncology: current applications and future prospects. *Curr. Pharm. Des.* 22, 6935–6945. doi: 10.2174/1381612822666160923115828

Ashdown, M. L., Robinson, A. P., Yatomi-Clarke, S. L., Ashdown, M. L., Allison, A., Abbott, D., et al. (2015). Chemotherapy for late-stage cancer patients: meta-analysis of complete response rates. *F1000Res* 4:232. doi: 10.12688/f1000research.6760.1

Awazu, A., Tanabe, T., Kamitani, M., Tezuka, A., and Nagano, A. J. (2018). Broad distribution spectrum from gaussian to power law appears in stochastic

variations in RNA-seq data. *Sci. Rep.* 8:8339. doi: 10.1038/s41598-018-26735-4

Balwierz, P. J., Carninci, P., Daub, C. O., Kawai, J., Hayashizaki, Y., Van Belle, W., et al. (2009). Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol.* 10:R79.

Banerji, C. R. S., Severini, S., Caldas, C., and Teschendorff, A. E. (2015). Intratumour signalling entropy determines clinical outcome in breast and lung cancer. *PLoS Comput. Biol.* 11:e1004115. doi: 10.1371/journal.pcbi.1004115

Barabási, A.-L. (2016). *Network Science*. Cambridge: Cambridge University Press, 475.

Belikov, A. V. (2017). The number of key carcinogenic events can be predicted from cancer incidence. *Sci. Rep.* 7:12170. doi: 10.1038/s41598-017-12448-7

Blank, S., and Dorf, B. (2012). *The Startup Owner's Manual: The Step-By-Step Guide for Building a Great Company*. Bartlett: K & S Ranch.

Bloomfield, M., and Duesberg, P. (2016). Inherent variability of cancer-specific aneuploidy generates metastases. *Mol. Cytogenet.* 9:90. doi: 10.1186/s13039-016-0297-x

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003). A comparison of normalization methods for highdensity oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185–193. doi: 10.1093/bioinformatics/19.2.185

Bradshaw, S., Brazil, E., and Chodorow, K. (2019). *MongoDB: The Definitive Guide: Powerful and Scalable Data Storage*, 3rd Edn. Newton, MA: O'Reilly Media, Inc, 514.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Breitkreutz, D., Hlatky, L., Rietman, E., and Tuszynski, J. A. (2012). Molecular signaling network complexity is correlated with cancer patient survivability. *Proc. Natl. Acad. Sci. U S A.* 109, 9209–9212. doi: 10.1073/pnas.1201416109

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94. doi: 10.1186/1471-2105-11-94

Calzolari, D., Bruschi, S., Coquin, L., Schofield, J., Feala, J. D., Reed, J. C., et al. (2008). Search algorithms as a framework for the optimization of drug combinations. *PLoS Comput. Biol.* 4:e1000249. doi: 10.1371/journal.pcbi.1000249

Calzolari, D., Paternostro, G., Harrington, P. L., Piermarocchi, C., and Duxbury, P. M. (2007). Selective control of the apoptosis signaling network in heterogeneous cell populations. *PLoS One* 2:e547. doi: 10.1371/journal.pone.0000547

Campbell, P. J., Getz, G., Korbel, J. O., and The Icgc/Tcga Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature* 578, 82–93. doi: 10.1038/s41586-020-1969-6

Carels, N., Conforte, A. J., Lma, C. R., and da Silva, F. A. B. (2020). "Challenges for the optimization of drug therapy in the treatment of cancer," in *Computational Biology*, 1ed Edn, Vol. 32, eds F. A. B. da Silva, N. Carels, T. M. dos Santos, and F. J. P. Lopes (Cham: Springer International Publishing), 163–198. doi: 10.1007/978-3-030-51862-2_8

Carels, N., Tilli, T., and Tuszynski, J. A. (2015a). A computational strategy to select optimized protein targets for drug development toward the control of cancer diseases. *PLoS One* 10:e0115054. doi: 10.1371/journal.pone.0115054

Carels, N., Tilli, T. M., and Tuszynki, J. A. (2015b). Optimization of combination chemotherapy based on the calculation of network entropy for protein-protein interactions in breast cancer cell lines. *EPJ Nonlinear Biomed. Phys.* 3:6.

Catharina, L., de Menezes, M. A., and Carels, N. (2018). "System biology to access target relevance in the research and development of molecular inhibitors," in *Theoretical and Applied Aspects of System Biology. Computational Biology*, 1ed Edn, Vol. 27, eds F. A. B. da Silva, N. Carels, and F. Paes Silva Jr. (Cham: Springer International Publishing), 221–242. doi: 10.1007/978-3-319-74974-7_12

Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619. doi: 10.1038/nmeth.1223

Collins, F. S., and Varmus, H. (2015). A new initiative on precision medicine. *N. Engl. J. Med.* 372, 793–795. doi: 10.1056/nejmp1500523

Conforte, A. J., Alves, L. D., Coelho, F. C., Carels, N., and da Silva, F. A. B. (2020). Modeling basins of attraction for breast cancer using Hopfield networks. *Front. Genet.* 11:314. doi: 10.3389/fgene.2020.00314

Conforte, A. J., Tuszynski, J. A., da Silva, F. A. B., and Carels, N. (2019). Signaling complexity measured by shannon entropy and its application in personalized medicine. *Front. Genet.* 10:930. doi: 10.3389/fgene.2019.00930

Dagnelie, P. (1970). *Théorie et méthodes Statistiques: Applications Agronomiques Vol. 2. Les méthodes de l'inférence Statistique*. Gembloux: J. Duculot, 451.

Danielsson, F., Skogs, M., Huss, M., Rexhepaj, E., O'Hurley, G., Klevebring, D., et al. (2013). Majority of differentially expressed genes aredown-regulated during malignant transformationin a four-stage model. *Proc. Natl. Acad. Sci. U S A.* 110, 6853–6858. doi: 10.1073/pnas.1216436110

Deelman, E., Gannon, D., Shields, M., and Taylor, I. (2009). Workflows and e-Science: an overview of workflow system features and capabilities. *Future Generat. Comp. Systems* 25, 528–540. doi: 10.1016/j.future.2008.06.012

Duesberg, P., Li, R., Fabarius, A., and Hehlmann, R. (2005). Aneuploidy and cancer: from correlation to causation. *Cell. Oncol.* 27, 293–318.

Evans, C., Hardin, J., and Stoebel, D. M. (2018). Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief. Bioinform.* 19, 776–792. doi: 10.1093/bib/bbx008

Fain, Y., and Moiseev, A. (2018). *Angular Development with TypeScript, Second Edition*. Shelter Island, NY: Manning Publications, 560.

Finak, G., Sadekova, S., Pepin, F., Hallett, M., Meterissian, S., Halwani, F., et al. (2006). Gene expression signatures of morphologically normal breast tissue identify basal-like tumors. *Breast Cancer Res.* 8:R58.

Grade, M., Difilippantonio, M. J., and Camps, J. (2015). Patterns of chromosomal aberrations in solid tumors. *Recent Results Cancer Res.* 200, 115–142. doi: 10.1007/978-3-319-20291-4_6

Guo, X. E., Ngo, B., Modrek, A. S., and Lee, W.-H. (2014). Targeting tumor suppressor networks for cancer therapeutics. *Curr. Drug Targets* 15, 2–16. doi: 10.2174/1389450114666140106095151

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

Holmes, S., and Herber, C. (2019). *Getting MEAN with Mongo, Express, Angular, and Node*. Shelter Island, NY: Manning Publications.

Holmes, S., and Huber, W. (2019). *Modern Statistics for Modern Biology*. Cambridge: Cambridge University Press, 402.

Hornsby, C., Page, K. M., and Tomlinson, I. P. (2007). What can we learn from the population incidence of cancer? armitage and doll revisited. *Lancet Oncol.* 8, 1030–1038. doi: 10.1016/s1470-2045(07)70343-1

Hu, Q., Sun, W., Wang, C., and Gu, Z. (2016). Recent advances of cocktail chemotherapy by combination drug delivery systems. *Adv. Drug. Deliv. Rev.* 98, 19–34. doi: 10.1016/j.addr.2015.10.022

Lee, E. Y., and Muller, W. J. (2010). Oncogenes and tumor suppressor genes. *Cold Spring Harb. Perspect Biol.* 2:a003236. doi: 10.1101/cshperspect.a003236

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12:323. doi: 10.1186/1471-2105-12-323

Li, X., Cooper, N. G. F., O'Toole, T. E., and Rouchka, E. C. (2020). Choice of library size normalization and statistical methods for differential gene expression analysis in balanced two-group comparisons for RNA-seq studies. *BMC Genomics* 21:75. doi: 10.1186/s12864-020-6502-7

Liu, J., Lichtenberg, T., Hoadley, K., Poisson, L., Lazar, A., Cherniack, A., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416.

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.

Lu, D.-Y., Lu, T.-R., Yarla, N. S., Wu, H.-Y., Xu, B., Ding, J., et al. (2017). Drug combination in clinical cancer treatments. *Rev. Recent Clin. Trials* 12, 202–211.

Malvia, S., Bagadi, S. A. R., Pradhan, D., Chintamani, C., Bhatnagar, A., Arora, D., et al. (2019). Study of gene expression profiles of breast cancers in indian women. *Sci. Rep.* 9:10018. doi: 10.1038/s41598-019-46261-1

Masic, I., Miokovic, M., and Muhamedagic, B. (2008). Evidence based medicine – new approaches and challenges. *Acta Inform. Med.* 16, 219–225. doi: 10.5455/aim.2008.16.219-225

McShane, L. M., and Polley, M. Y. (2013). Development of omics-based clinical tests for prognosis and therapy selection: the challenge of achieving statistical

robustness and clinical utility. *Clin. Trials* 10, 653–665. doi: 10.1177/1740774513499458

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Naito, Y., and Urasaki, T. (2018). Precision medicine in breast cancer. review article. *Chin. Clin. Oncol.* 7:29. doi: 10.21037/cco.2018.06.04

Nikanjam, M., Liu, S., and Kurzrock, R. (2016). Dosing targeted and cytotoxic two-drug combinations: lessons learned from analysis of 24,326 patients reported 2010 through 2013. *Int. J. Cancer* 139, 2135–2141. doi: 10.1002/ijc.30262

Novartis (2020). *The Precision Oncology Annual Trend Report: Perspectives From Oncologists, Pathologists, and Payers. Sixth Edition. 48.* Available online at: https://www.hcp.novartis.com/globalassets/migration-root/hcp/care-management-new/assets/mmo-1224797-the-precision-oncology-annual-trend-report-6th-edition.pdf

Ogino, S., Fuchs, C. S., and Giovannucci, E. (2012). How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. *Expert. Rev. Mol. Diagn.* 12, 621–628. doi: 10.1586/erm.12.46

Oshlack, A., and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct* 4:14. doi: 10.1186/1745-6150-4-14

Ozery-Flato, M., Linhart, C., Trakhtenbrot, L., Izraeli, S., and Shamir, R. (2011). Large-scale analysis of chromosomal aberrations in cancer karyotypes reveals two distinct paths to aneuploidy. *Genome Biol.* 12:R61. doi: 10.1186/gb-2011-12-6-r61

Preissner, S., Dunkel, M., Hoffmann, M. F., Preissner, S. C., Genov, N., Rong, W. W., et al. (2012). Drug cocktail optimization in chemotherapy of cancer. *PLoS One* 7:e51020. doi: 10.1371/journal.pone.0051020

Risso, D., Schwartz, K., Sherlock, G., and Dudoit, S. (2011). GC-content normalization for RNA-Seq data. *BMC Bioinformatics* 12:480. doi: 10.1186/1471-2105-12-480

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11:R25.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* 2004:3.

Teschendorff, A. E., and Severini, S. (2010). Increased entropy of signal transduction in the cancer metastasis phenotype. *BMC Syst. Biol.* 4:104. doi: 10.1186/1752-0509-4-104

Tilli, T. M., Carels, N., Tuszynski, J. A., and Pasdar, M. (2016). Validation of a network-based strategy for the optimization of combinatorial target selection in breast cancer therapy: siRNA knockdown of network targets in MDA-MB-231 cells as an in vitro model for inhibition of tumor development. *Oncotarget* 7, 63189–63203. doi: 10.18632/oncotarget.11055

Tom, J. A., Reeder, J., Forrest, W. F., Graham, R. R., Hunkapiller, J., Behrenset, T. W., et al. (2017). Identifying and mitigating batch effects in whole genome sequencing data. *BMC Bioinformatics* 18:351. doi: 10.1186/s12859-017-1756-z

van Wieringen, W. N., and van der Vaart, A. W. (2011). Statistical analysis of the cancer cell's molecular entropy using high-throughput data. *Bioinformatics* 27, 556–563. doi: 10.1093/bioinformatics/btq704

Verma, M. (2012). Personalized medicine and cancer. *J. Pers. Med.* 2, 1–14. doi: 10.1016/j.pmu.2014.03.007

Vuckovic, N., Vuckovic, B. M., Liu, Y., and Paranjape, K. (2016). *Accelerating Clinical Genomics to Transform Cancer Care*. Santa Clara, CA: Intel.

Welch, B. L. (1949). Further note on Mrs Aspin's tables and on certain approximations to the tabled function. *Biometrika* 36, 293–296.

West, J., Bianconi, G., Severini, S., and Teschendorff, A. E. (2012). Differential network entropy reveals cancer system hallmarks. *Sci. Rep.* 2:802. doi: 10.1038/srep00802

Willems, S. M., Abeln, S., Feenstra, K. A., de Bree, R., van der Poel, E. F., de Jong, R. J. B., et al. (2019). The potential use of big data in oncology. *Oral Oncol.* 98, 8–12. doi: 10.1016/j.oraloncology.2019.09.003

Wilsdon, T., Barron, A., Edwards, G., and Lawlor, R. (2018). *The Benefits of Personalised Medicine to Patients, Society and Healthcare Systems*. Boston, MA: Charles River Associates.

Xia, Y., Fan, C., Hoadley, K. A., Parker, J. S., and Perou, C. M. (2019). Genetic determinants of the molecular portraits of epithelial cancers. *Nat. Commun.* 10:5666. doi: 10.1038/s41467-019-13588-2

Ye, C. J., Regan, S., Liu, G., Alemara, S., and Heng, H. H. (2018). Understanding aneuploidy in cancer through the lens of system inheritance, fuzzy inheritance and emergence of new genome systems. *Mol. Cytogenet.* 11:31. doi: 10.1186/s13039-018-0376-2

Yoder, O. C. (1980). Toxins in pathogenesis. *Annu. Rev. Phythopathol.* 18, 103–129.

Conflict of Interest: The intellectual property of this research is protected by the Brazilian patent number BR1020150308191.

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: 18 March 2021
doi: 10.3389/fgene.2021.638708

Effects of Simulated Microgravity on the Proteome and Secretome of the Polyextremotolerant Black Fungus *Knufia chersonesos*

Donatella Tesei[1,2], Abby J. Chiang[3], Markus Kalkum[3], Jason E. Stajich[4], Ganesh Babu Malli Mohan[5], Katja Sterflinger[6] and Kasthuri Venkateswaran[2]*

[1] Department of Biotechnology, University of Natural Resources and Life Sciences, Vienna, Austria, [2] Biotechnology and Planetary Protection Group, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, United States, [3] Department of Molecular Imaging and Therapy, Beckman Research Institute of City of Hope, Duarte, CA, United States, [4] Department of Microbiology and Plant Pathology, Institute of Integrative Genome Biology, University of California, Riverside, Riverside, CA, United States, [5] Department of Biotechnology, Centre for Research and Infectious Diseases, SASTRA Deemed University, Thanjavur, India, [6] Institute for Natural Sciences and Technology in the Arts, Academy of Fine Arts Vienna, Vienna, Austria

OPEN ACCESS

Edited by:
Joanna Jankowicz-Cieslak, International Atomic Energy Agency, Austria

Reviewed by:
Mateusz Molon, University of Rzeszow, Poland
Khaled Y. Kamal, Texas A&M University, United States
Ruth Bryan, Albert Einstein College of Medicine, United States

***Correspondence:**
Donatella Tesei
donatella.tesei@boku.ac.at

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

Received: *07 December 2020*
Accepted: *19 February 2021*
Published: *18 March 2021*

Citation:
Tesei D, Chiang AJ, Kalkum M, Stajich JE, Mohan GBM, Sterflinger K and Venkateswaran K (2021) Effects of Simulated Microgravity on the Proteome and Secretome of the Polyextremotolerant Black Fungus Knufia chersonesos. Front. Genet. 12:638708. doi: 10.3389/fgene.2021.638708

Black fungi are a group of melanotic microfungi characterized by remarkable polyextremotolerance. Due to a broad ecological plasticity and adaptations at the cellular level, it is predicted that they may survive in a variety of extreme environments, including harsh niches on Earth and Mars, and in outer space. However, the molecular mechanisms aiding survival, especially in space, are yet to be fully elucidated. Based on these premises, the rock-inhabiting black fungus *Knufia chersonesos* (Wt) and its non-melanized mutant (Mut) were exposed to simulated microgravity—one of the prevalent features characterizing space conditions—by growing the cultures in high-aspect-ratio vessels (HARVs). Qualitative and quantitative proteomic analyses were performed on the mycelia and supernatant of culture medium (secretome) to assess alterations in cell physiology in response to low-shear simulated microgravity (LSSMG) and to ultimately evaluate the role of cell-wall melanization in stress survival. Differential expression was observed for proteins involved in carbohydrate and lipid metabolic processes, transport, and ribosome biogenesis and translation via ribosomal translational machinery. However, no evidence of significant activation of stress components or starvation response was detected, except for the scytalone dehydratase, enzyme involved in the synthesis of dihydroxynaphthalene (DNH) melanin, which was found to be upregulated in the secretome of the wild type and downregulated in the mutant. Differences in protein modulation were observed between *K. chersonesos* Wt and Mut, with several proteins being downregulated under LSSMG in the Mut when compared to the Wt. Lastly, no major morphological alterations were observed following exposure to LSSMG. Similarly, the strains' survivability was not negatively affected. This study is the first to characterize the response to simulated microgravity in black fungi, which might have implications on future astrobiological missions.

Keywords: microgravity, black fungi, extremophiles, secretomics, proteomics, astrobiology, *Knufia chersonesos* (syn. *K. petricola*)

INTRODUCTION

Based on their nature as settlers in extreme environments, microbial extremophiles are of great interest to studies aiming to elucidate stress adaptation and survival mechanisms. Successful examples of extremophiles can be found in the fungal domain, alongside bacteria, and archaea, which for a long time were considered to be the sole colonizers of habitats previously considered uninhabitable. Some of these fungal extremophiles show even higher resistance than that of prokaryotes (Shtarkman et al., 2013; Aguilera and González-Toril, 2019). Black fungi in particular represent a group of highly melanized microfungi, whose ability to survive in a variety of extreme environments has in recent decades attracted increasing attention (Selbmann et al., 2005; Sterflinger, 2006; Gorbushina, 2007; Onofri et al., 2007). Desiccation, low nutrient availability, excessive radiation, extreme temperatures, salinity, and pH are some of the multiple sources of stress characterizing the habitats where these organisms have been shown to thrive (Gunde-Cimerman et al., 2000, 2003; Selbmann et al., 2008). The highest diversity of black fungi has especially been observed in rocky environments, ranging from mountain summits to the Atacama Desert and the Antarctic cold regions (Gonçalves et al., 2016; Ametrano et al., 2019). Rock represents a harsh habitat and a quite ancient niche for life, believed to reflect early Earth conditions, and thus considered to be a model for extraterrestrial life (Selbmann et al., 2014).

The discovery of melanized fungi in extreme environments has prompted researchers to investigate microbial physiology at the absolute edges of adaptability, aiming at a deeper understanding of the limits for life (Dadachova and Casadevall, 2008; Tesei et al., 2012; Zakharova et al., 2013). Other studies have focused on testing fungal survival in space conditions through simulations in ground-based facilities or in space missions that enabled the assessment of the habitability of extraterrestrial environments and, hence, the possibility of life beyond Earth (Scalzi et al., 2012; Onofri et al., 2019). Space and outer space conditions are by definition hostile, as they include enhanced irradiation, microgravity, and temperature extremes (Rabbow et al., 2012, 2017; Senatore et al., 2018). Nevertheless, the isolation of melanized fungi from spacecraft and space stations—e.g., the International Space Station (ISS)—has been reported frequently (Checinska et al., 2015; Checinska-Sielaff et al., 2019). In this respect, a number of studies have examined the molecular adaptations of ISS-isolated strains to space conditions, showing alterations in metabolome and proteome (Knox et al., 2016; Blachowicz et al., 2019b; Romsdahl et al., 2019). Investigations of microbial survival in space are therefore relevant also in the context of space missions, to prevent contaminations and to develop strategies to reduce hazard, especially in the case of opportunistic species (Urbaniak et al., 2019).

In black fungi, astrobiological studies have evaluated the potential effects of Mars or ISS conditions on fungal viability and metabolism. One investigation revealed that a rock isolate from Antarctica, *Cryomyces minteri*, could survive simulated Martian atmosphere and pressure, temperature fluctuations between −20 and 20°C, ultraviolet radiation, and vacuum (Onofri et al., 2008). Comparative 2D-PAGE proteomics was carried out for other rock-inhabiting fungi (RIF) (i.e., *Cryomyces antarcticus*, *Knufia perforans*, and *Exophiala jeanselmei*) exposed to these conditions and showed a decrease in protein complexity, followed by recovery of the metabolic activity after 1 week of exposure (Zakharova et al., 2014). Further, survivability of *C. antarcticus* in outer space was shown via colony counts following a 1.5-year-long exposure on board the EXPOSE-E facility of the ISS (Onofri et al., 2012). A more recent experiment of *C. antarcticus* exposure on rock analogs under space and simulated Martian conditions revealed only slight ultra-structural and molecular damage and pointed out the high stability of DNA within melanized cells (Pacelli et al., 2016). In other investigations, the resistance to acute ionizing radiations was demonstrated in RIF (Pacelli et al., 2017, 2018). Together, these studies have revealed the ability of rock-colonizing black fungi to endure space conditions; however, to date, reports that specifically evaluate the response to microgravity have not been produced.

Microgravity is an important factor influencing microbial life in space environments, a condition in which the gravity level is almost zero but not neutralized. Due to the technological and logistical hurdles linked to studies of microgravity in space, different methods have been developed to simulate microgravity and analyze microbial responses (Herranz et al., 2013). Accordingly, the term low-shear simulated microgravity (LSSMG) is used to describe the environmental condition created by these devices, resembling the low-shear effects of the fluid on the cells (Yamaguchi et al., 2014). Spaceflight and ground-based microgravity analog experiments have suggested that microgravity can affect microbial gene expression, cell morphology, physiology, and metabolism, also triggering increased virulence in pathogenic bacteria and fungi (Altenburg et al., 2008; Taylor, 2015; Sathishkumar et al., 2016; Huang et al., 2018). Although the effects of microgravity on microbes have been studied for several years, only the responses of a few typical prokaryotic and eukaryotic model organisms—e.g., *Escherichia coli*, *Candida albicans*, *Saccharomyces cerevisiae, Penicillium sp., Aspergillus sp.*—have hitherto been investigated (Huang et al., 2018). Hence, studying the reaction of black fungi to microgravity holds potential for the elucidation of the molecular basis of tolerance in extremotolerant and extremophilic fungi, and can also contribute to unearthing the biological uniqueness of these species and their adaptability to space conditions.

In the present study, the qualitative and quantitative proteomic characterization of a black fungus response to LSSMG was carried out for the first time. The rock-associated *Knufia chersonesos* (syn. *Knufia petricola*) was selected as the model organism due to its reported poikilo-tolerance, e.g., the ability to endure xeric conditions, desiccation, high UV-radiation and temperatures (Sterflinger et al., 2012) and to feed on alternative carbon sources like monoaromatic compounds (Nai et al., 2013) and synthetic polyesters (Tesei et al., 2020). *K. chersonesos* aptitude to withstand levels of gaseous ozone up to 11 ppm was also shown (Tesei et al., unpublished). Furthermore, being the only black fungus known to have a melanin-deficient spontaneous mutant (Tesei et al., 2017), *K. chersonesos* allows comparative studies attempting to evaluate the stress-protective

role of melanin. Cultures of *K. chersonesos* wild type and mutant were grown in HARVs and analyzed for changes at the proteome—whole-cell proteome (mycelia) and secretome (culture supernatant)—and at the morphological level, with an eye toward the impact of cell-wall melanization on the physiological response to the stress.

MATERIALS AND METHODS

Fungal Strains

The fungal strains used in this study included the non-pathogenic rock-inhabiting fungus *K. chersonesos* MA5789 wild type (Wt) and MA5790 mutant (Mut), both obtained from the ACBR fungal culture collection of the University of Natural Resources and Life Sciences, Vienna, Austria. The Wt, characterized by a highly melanized mycelium, was isolated from red sandstone in Ny London, Svalbard, Norway. The pink mutant, whose pigmentation is due to unmasking of carotenoids resulting from the lack of melanin, originated spontaneously under laboratory conditions (Tesei et al., 2017). *K. chersonesos* (syn. *K. petricola*) is an emerging model organism for analyses aiming at the elucidation of the rock-lifestyle and RIF physiology (Nai et al., 2013). Along with its ascertained thermo-, pH-, UV- and desiccation tolerance (Gorbushina et al., 2008; Sterflinger et al., 2012), the fungus was recently reported to have an aptitude for using synthetic polymers as an alternative carbon source (Tesei et al., 2020). These features and the ability to endure high levels of gaseous ozone (Tesei et al., unpublished), altogether make *K. chersonesos* particularly suited for astrobiology studies. Further, the availability of a mutant strain allows comparative studies attempting to evaluate the role of melanin in stress protection. Fungal cultures were maintained in flasks containing 50 mL of 2% malt extract broth (MEB, pH 5, 2% malt extract, 2% glucose, 1% peptone) at 21°C with shaking at 63 rpm (Innova, Eppendorf). Cultures at exponential phase were used for inoculation of media for all experiments.

Exposure to Low-Shear Simulated Microgravity (LSSMG)

Fungal pellets were obtained from 5-day-old cultures by centrifugation, washed in 1X phosphate-buffered saline (PBS; Thermo Fisher Scientific) and subsequently mildly ribolyzed (3 × 20", speed 4; MP Biomedicals) to separate cells. As clump-like growth is an inherent characteristic of black fungi, mechanical disruption of cell clusters is a prerequisite for both inoculation of cultures and cell counting (Voigt et al., 2020). Following the cell count (Neubauer), the cell number was adjusted to the initial concentration of 2×10^5 ml^{-1} with 10 mL fresh 2% MEB and used as seed culture. High-aspect-ratio vessels (HARVs; Synthecon Inc) were filled with the cell suspensions (10 mL) and rotated at 30 rpm (initial rotation rate) in the vertical axis to provide LSSMG conditions inside a chamber with 60% humidity and 22°C, for 7 days. Control runs were set up by rotating the bioreactors in the horizontal axis to provide normal gravity condition (1G) (Rosenzweig et al., 2010; Kim and Rhee, 2016). During cultivation, the rotation speed was adjusted to 43 rpm in order to keep the cells pellets orbiting within the vessel in continual fall and to prevent their contact with the vessel walls (**Figures 1A,B**). A total of two biological replicates were maintained throughout the experiments. An additional set of experiments identical to the conditions mentioned above was established for measurement of cell concentration and microscope observations. Cell concentration was assessed via hemocytometer count at different time points until completion of the experiment, using two biological replicates (i.e., biomass from two different vessels) for each experimental condition. Following removal of the vessel from the rotator base, samples of cells were collected under sterile conditions in a laminar flow hood using a luer-lock syringe and the syringe port in the culture vessel. The sampling was carried out one vessel at a time while the remaining vessels were kept rotating on the rotator base. Cell survivability at the end of the LSSMG exposure was assessed by enumeration of colony forming units (CFU) using ImageJ software (Schneider et al., 2012), according to Choudhry (2016).

Microscopy Studies

Scanning Electron Microscopy (SEM)

For SEM studies, samples from the two strains were collected from all the established cultures—LSSMG and normal gravity condition (1G)—at various timepoints, i.e., 1, 3, 5, and 7 days from the beginning of the cultivation. A 1:10 dilution of each sample was prepared using 1X phosphate-buffered saline (PBS; Thermo Fisher Scientific) and thereafter transferred onto electron microscopy coupons. For each sample, two biological replicates (and two technical replicates each i.e., aliquots of 20 μl) were prepared and let dry at 35°C before examination. A Sirion (FEI, Hillsboro, OR, United States) field-emission scanning electron microscope (FE-SEM) was used for examination of the fungal cells. No specimen preparation procedures, such as sample coating with a surface-conducting (carbon or metal) layer, were performed as they were not necessary when using the FE-SEM. The samples were analyzed using an acceleration voltage of 10–20 kV, beam current of 40–50 mA and positioning the detector 6–10 mm away from the coupon. Secondary electron images were acquired in the high-vacuum mode. For each sample coupon, a minimum of 10 fields was observed, and images were acquired with various magnification (100×–5,000×).

Fluorescence Microscopy

For the cell integrity and viability assay, wheat germ agglutinin (WGA) and propidium iodide (PI) in combination with SYTO 9 were used. The carbohydrate-binding lectin WGA has a known affinity for β-1,4-N-acetylglucosamine (GlcNAc) oligomers present in the fungal polysaccharide chitin. PI, a membrane impermeant dye that is generally excluded from viable cells, was applied together with the nuclear and chromosome counterstain SYTO 9 for a dead/live stain. 200 μl aliquots were collected from all the established cultures after 1, 3, 5, and 7 days from the beginning of the cultivation. Fungal cells were washed twice in 1X PBS and suspended in 500 μl. Following a 1:10 dilution, 100 μl cell suspension was incubated with 50 μl Alexa Fluor 350 conjugate of WGA (Thermo Fisher Scientific)

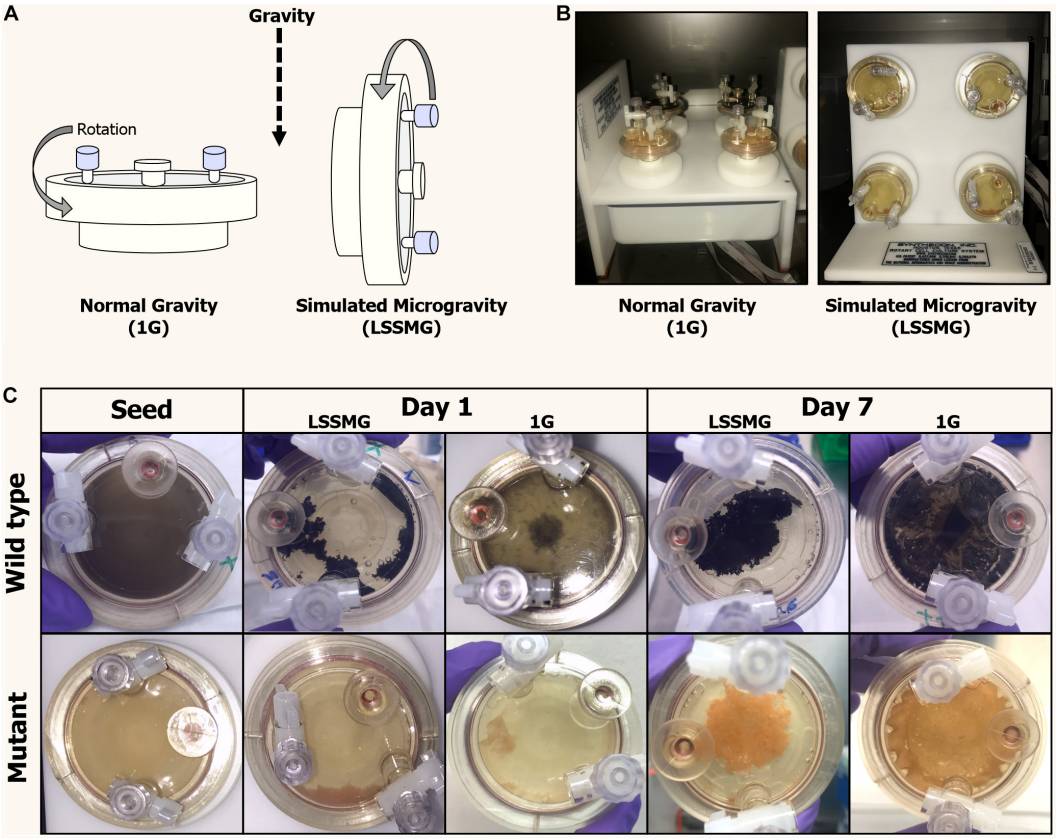

FIGURE 1 | Rotary Cell Culture System (RCCS^{TM}, Synthecon) and High Aspect Ratio Vessels (HARVs) used in the LSSMG experiments. **(A)** Schematic diagram of mechanical principle of the different RCCSs (design: Emily Klonicki). In normal gravity, the vessels (HARVs) rotate around a plane parallel to the gravity vector. Microgravity is simulated by rotating the samples around a plane perpendicular to the gravity vector. **(B)** The RCCS used to generate 1G condition (control) on ground (left); the RCCS used to generate LSSMG condition on ground (right); **(C)** Colony and cultural features of *Knufia chersonesos* MA 5789 (wild type, Wt) and MA5790 (mutant, Mut) under LSSMG and 1G on the first and last day of cultivation. The HARVs provide oxygenation via a flat, silicon rubber gas transfer membrane located at the base of the vessel and directly underneath the cultures.

and 15 µl of FungaLight containing equal amounts of PI and SYTO 9 solution (Thermo Fisher Scientific), for 15 min at room temperature in the dark. The chosen amounts of the fluorescent probes are justified by the thickening and the increase in melanization of the cell walls, which hindered the staining process. Cell suspensions were mounted over glass slides and analyzed under an Axioplane microscope equipped with an AxioCam camera (Carl Zeiss). For each sample, two technical replicates were prepared for observation.

Tandem Mass Tag (TMT)-Based Quantitative Shotgun Proteomics

Sample Preparation

The content of each vessel was spun at 7500 × g (5810, Eppendorf) and at 4°C for 15 min to separate fungal cell pellet (mycelium) from culture supernatant (secretome). Protease inhibitors (10 µl:1 mL v/v, Halt, Thermo Fisher Scientific) were thereafter added—and mixed with 1 mL 1 × PBS in the case of biomass samples—and samples were stored at −80°C prior to protein extraction. Proteins were extracted as previously described by Romsdahl et al. (2018) with some modifications. In brief, 1 mL lysis buffer consisting of 100 mM triethylammonium bicarbonate (TEAB) with 1:100 Halt Protease Inhibitor Cocktail (Thermo Fisher Scientific) was added to fungal cell pellets (i.e., ∼0.5 g) and culture supernatants (i.e., ∼0.5 mL), the latter previously concentrated in a SpeedVac system. Sample homogenization was achieved using a bead beater (Bertin) at 4°C (3 × 5500 rpm for 1 min., with 15 s. breaks in between) and following centrifugation at 17,000 × *g* and 4°C for 15 min, protein concentrations were determined by BCA assay (Thermo Fisher Scientific). The protein extracts were processed for a tandem mass tag (TMT) labeling as previously described (Romsdahl et al., 2018). For each sample, 250 µg proteins were precipitated in 20% TCA at 4°C, reduced by tris(2-carboxyethyl)phosphine (TCEP), alkylated with iodoacetamide (IAA), and digested with Trypsin/LysC (Promega, Madison, WI, United States) overnight at 37°C. Peptide quantitation was performed using the Pierce Quantitative Colorimetric Peptide Assay (Thermo Fisher Scientific). The proteomic and secretomic profiling of *K. chersonesos* Wt and Mut were carried out in two separate TMT LC/MS experiments. A total of 40 µg peptides

from each cell pellet sample or 13 μg peptides from each culture supernatant were labeled with the Thermo Scientific TMT10 plex (TMT^{10}) Isobaric Mass Tagging Kit according to the manufacturer protocol. Eight TMT tags were used to label samples from the same experimental set. The TMT^{10}-131 label was used as a reference that contained a pool of 5 μg of peptides from all samples. All nine labeled-peptide samples were combined into a single tube, mixed and fractionated using the Pierce High pH Reversed-Phase Peptide Fractionation Kit (Thermo Fisher Scientific). The fractionated samples were dried using a SpeedVac concentrator and dissolved in 1% formic acid prior to LC-MS/MS analysis.

LC-MS/MS Analysis

The samples were analyzed on an Orbitrap Fusion Lumos mass spectrometer with a Dionex UltiMate 3000 RSLCnano system, a 300 μm × 5 mm PepMap100 C18 precolumn, a 75 μm × 50 cm PepMap RSLC C18 analytical column, and an Easy-Spray ion source (Thermo Scientific). The column temperature was maintained at 45°C, and the peptides were eluted at a flow rate of 300 nL/min over a 110 min gradient, from 3 to 30% solvent B (100 min), 30 to 50% solvent B (5 min), 50 to 90% solvent B (1 min), 90% solvent B (1 min), and 90% to 3% solvent B (3 min). Solvent A was 0.1% formic acid in water and solvent B was 0.1% formic acid in acetonitrile. The full MS survey scan (400–1,600 m/z) was acquired in the orbitrap at a resolution of 240,000 and with an automatic gain control (AGC) target of 4×10^5. The maximum injection time for MS scans was 50 ms. Monoisotopic precursor ions were selected with charge states 2–7 within a ± 10 ppm mass window using a 60 s dynamic exclusion. The MS^2 scan (400–1,200 m/z) was performed using the linear ion trap with the CID collision energy set to 35%. The ion trap scan rate was set to "rapid," with an AGC target of 1×10^4, and a maximum injection time of 30 ms. Ten fragment ions from each MS^2 experiment were then simultaneously selected for an MS^3 experiment. The MS^3 scan (100–500 m/z) was performed to generate the TMT reporter ions in the orbitrap at a resolution of 30,000 and using HCD at a collision energy setting of 65%, an AGC target of 5×10^4, and a maximum injection time of 54 ms.

Quantitative Proteomics Analysis

All MS/MS spectra were analyzed using Proteome Discoverer (version 2.2.0.388, Thermo Fisher Scientific) with the Sequest-HT searching engines against an in-house annotated draft genome sequence of *K. chersonesos* MA5789 Wt (GCA_002319055.1, assembly ASM231905v1, NCBI) (Tesei et al., 2017) consisting of 9.818 predicted protein coding gene models. The genome was annotated using funannotate (v1.3.4) (doi: 10.5281/zenodo.1284502), which combined predictions from *ab initio* gene predictors with Augustus trained by BUSCO gene models (fungi_odb9) (Zdobnov et al., 2017) and GeneMark.hmm (Ter-Hovhannisyan et al., 2008) informed by protein evidence from Swiss-Prot (Boutet et al., 2007) together into composite gene models with EvidenceModeler (Haas et al., 2008). Functional predictions for genes was assigned by protein homology to Pfam (El-gebali et al., 2019), Swiss-Prot/UniProt (v 2018_05), and EggNog (v1.10) databases (Huerta-Cepas et al., 2019). The following parameters were selected for the search: A maximum of two missed cleavage sites, a minimum peptide length of six residues, 5 ppm tolerance for precursor ion masses, and 0.6 Da tolerance for fragment ion masses. The static modification settings included carbamidomethyl of cysteine residues, and dynamic modifications included oxidation of methionine, TMT modification of lysine ε-amino groups and peptide N-termini, and acetyl modification of protein N-terminus. A false discovery rate (FDR) of 1% for peptides and proteins was obtained using a target-decoy database search. The reporter ions integration tolerance was 0.5 Da while the co-isolation threshold was 75%. The average signal-to-noise threshold of all reporter peaks was greater than 10. The quantitative abundance of each protein was determined from the total intensity of the detected reporter ions. For statistical analysis, the sum of reporter ion intensities for each protein was log2 transformed, and the technical triplicate measurements for each protein were averaged. Only the proteins that were identified with at least one peptide and quantified in all technical (n = 3) and biological replicates (n = 2), were considered for the statistical analysis. Student's t-test was performed to identify differentially expressed proteins between each LSSMG-exposed and 1G-exposed group as well as between all groups. To compare all four experimental conditions, each condition was normalized to the reference channel, which contained equal amounts of peptides from each group. To evaluate changes between treatment (LSSMG) and control (1G), the protein abundance levels of each LSSMG-exposed sample were normalized to the 1G-exposed counterpart. Proteins with p-values of ≤0.05 were further evaluated for increased or decreased abundance using a cut-off value of ≥ ± 1.5-fold change (log2 fold change of ≥ ± 0.584).

Protein Identification and Bioinformatics Analysis

Protein identifications and functional insights were obtained from searching for sequence homologs using OmicsBox v. 1.4.11 (BioBam Bioinformatics S.L). To characterize proteins with respect to the biological process they are involved in, Gene Ontology (GO) terms were assigned to domains. The sequences were blasted (cloud BLASTP 2.10.0+, *E*-Value 1.0E-3, Filter: Fungi), and the blast hits were mapped and annotated with GOs using the GO database (Goa version 2019_11[1]; *E*-Value 1.0E-6, Filter GO by Taxonomy: taxa: 4751, Fungi) (Götz et al., 2008). GOs were additionally assigned using InterProScan (IPS) to retrieve domains/motif information in a sequence-wise manner, and EggNOG using precomputed EggNOG-based orthology assignments. Corresponding GOs were then transferred to the sequences and merged with already existing GO terms. The annotations were validated based on the True-Path-Rule by removing all redundant terms to a given sequence. A GO-Slim analysis was run to summarize the GO annotation using the *Aspergillus* slim. Existing GO terms were additionally mapped to enzymes codes, when possible. A pathway analysis was performed

[1]http://geneontology.org

to retrieve metabolic pathways based on the GO terms and the enzyme codes using the Load KEGG Pathway tool (Kanehisa and Goto, 2000). To elucidate the identity of the uncharacterized proteins, a search for homology was further performed in the UniProtKB database[2] (BLASTP parameters: E-Threshold: 10; matrix BLOSUM62). In cases of blast results where the most significant match was represented by an uncharacterized protein, the first match in the list of homologous proteins where a protein ID was available was considered. Information about the predicted protein localization was obtained using BUSCA[3], based on the identification of signal and transit peptides, GPI-anchors and alpha-helical and beta-stranded transmembrane domains (Savojardo et al., 2018). Protein-protein interaction analyses were performed using STRING v11.0 with high confidence (0.70) (Jensen et al., 2009), selecting the proteome of the black yeast *Exophiala dermatitidis* as reference database based on its phylogenetic proximity to *K. chersonesos* (Tesei et al., 2020).

RESULTS

Growth Behavior

Cell concentration in the LSSMG-exposed and unexposed samples was measured via hemocytometer at four different time points: (1) start (seed, day 0), (2) acceleration phase (in between lag and exponential phases, day 3), (3) exponential phase (day 5) and (4) at the end of the experiment (stationary phase, day 7). In *K. chersonesos* Wt_{LSSMG}, ∼23.5-fold increase in cell concentration was observed during acceleration phase when compared to the original inoculum (2×10^5 cells/mL). Such an increase during acceleration phase was also noticed in *K. chersonesos* Mut_{LSSMG} (10-fold), *K. chersonesos* Wt_{1G} (22-fold) and *K. chersonesos* Mut_{1G} (19-fold). During stationary phase the increase in cell concentration was 28-fold (*K. chersonesos* Wt_{LSSMG}), 30-fold (*K. chersonesos* Mut_{LSSMG}), 42-fold (*K. chersonesos* Wt_{1G}), and 36-fold (*K. chersonesos* Mut_{1G}) when compared to original inoculum. Among LSSMG and Earth gravity grown cultures, higher values were recorded at normal gravity condition. Overall, only ∼one log growth in 7 days might be due to the clumping nature of *Knufia* cells. Cell survivability in simulated microgravity was measured by CFU using ImageJ software (Schneider et al., 2012) according to established protocol (Choudhry, 2016) and were: *K. chersonesos* $Wt_{LSSMG} = 2.8 \times 10^6$ cell/mL, *K. chersonesos* Mut_{LSSMG} 2.5×10^6 cell/mL, *K. chersonesos* Wt_{1G} 2.6×10^6 cell/mL, *K. chersonesos* Mut_{1G} 2.3×10^6 cell/mL (**Supplementary Table 1**). Potential mechanical damages or cell lysis caused by cell separation via mild bead beating (i.e., ribolyzer, 3 × 20 s, speed 4) prior to cell count are to be excluded since optical microscope observations and tests involving repeated bead beater treatments indicated that it does not decrease cell viability. However, underestimation of cell survivability by CFU due to cell clumping, an inherent characteristic of black fungi, should be taken into account.

[2]http://www.uniprot.org/blast

[3]http://busca.biocomp.unibo.it

Under LSMMG conditions, fungal cells showed an increased clumping, resulting in the aggregation of cells to form a mycelial growth in the center of the vessel. However, growth morphology under Earth gravity resembled a biofilm of cells adhering to the vessel's oxygenation membrane (**Figure 1C**).

Microscopy

To evaluate the presence of LSSMG-dependent alterations of fungal cell morphology in the melanotic (Wt) and non-melanized spontaneous mutant (Mut) strain, we examined cellular differentiation by growing cells at low density. In order to analyze single cellular morphology of the Wt and Mut strain before LSSMG-treatment, exponentially growing fungal cells were spotted on top of aluminum coupons and visualized using FE-SEM. The observation of the control fungal cell morphology in non-coated FE-SEM specimens revealed that the Wt produced biconcave cells, pseudohyphae and hyphae, while the mutant showed smooth spherical cells, pseudohyphae and cell aggregates (**Figures 2**, **3**).

Low-shear simulated microgravity-exposed and 1G-exposed fungal cells were collected at various time points, as described in the section "Materials and Methods." A detailed examination of fungal cellular morphology in non-coated FE-SEM specimens revealed 1G and LSSMG Wt and mutant strains showing similar morphology at day 1 and 3. Remarkably, the 1G Wt hyphae formation was delayed until day 7, whereas LSSMG induced hyphae formation at day 5 (**Figure 2**). Notably, the mutant showed no evidence of hyphae formation in LSSMG condition, but the aggregation of cells was evident at day 3 and increased until day 7, similarly to what was observed in the 1G cells (**Figure 2**). Furthermore, a closer observation of Wt and mutant fungal cells in LSSMG and 1G condition at 2500× magnification clearly shows that the Wt cells form biconcave morphology compared to the mutant cells (**Figure 3**). LSSMG treatment led to no significant change in cell size relative to 1G conditions, in either strain (**Supplementary Table 2**). Similarly, images of PI-SYTO 9 and WGA staining generated by confocal microscopy did not reveal changes in cell viability and integrity in response to LSSMG exposure. Greatest fluorescence was seen with WGA at the beginning of the cultivation, although increasing concentrations of the dye were applied; this is most probably due to the thickening of the cell wall and increase of melanization, which naturally occurs in the fungus over the course of cultivation (**Supplementary Figure 1**). Collectively, these findings uncover that the LSSMG and 1G conditions did not influence the fungal cellular growth and morphology.

Overview of Whole-Cell Proteome and Secretome Analysis

The 16 samples, comprised of four experimental conditions with two biological replicates each for each of the analyzed strains, yielded 3777 proteins as detected by isobaric TMT labeling-based LC-MS/MS. Of these, 3177 were in the whole-cell proteome and 600 in the secreted fraction. Furthermore, 2602 proteins were unique to the proteome and 25 to the

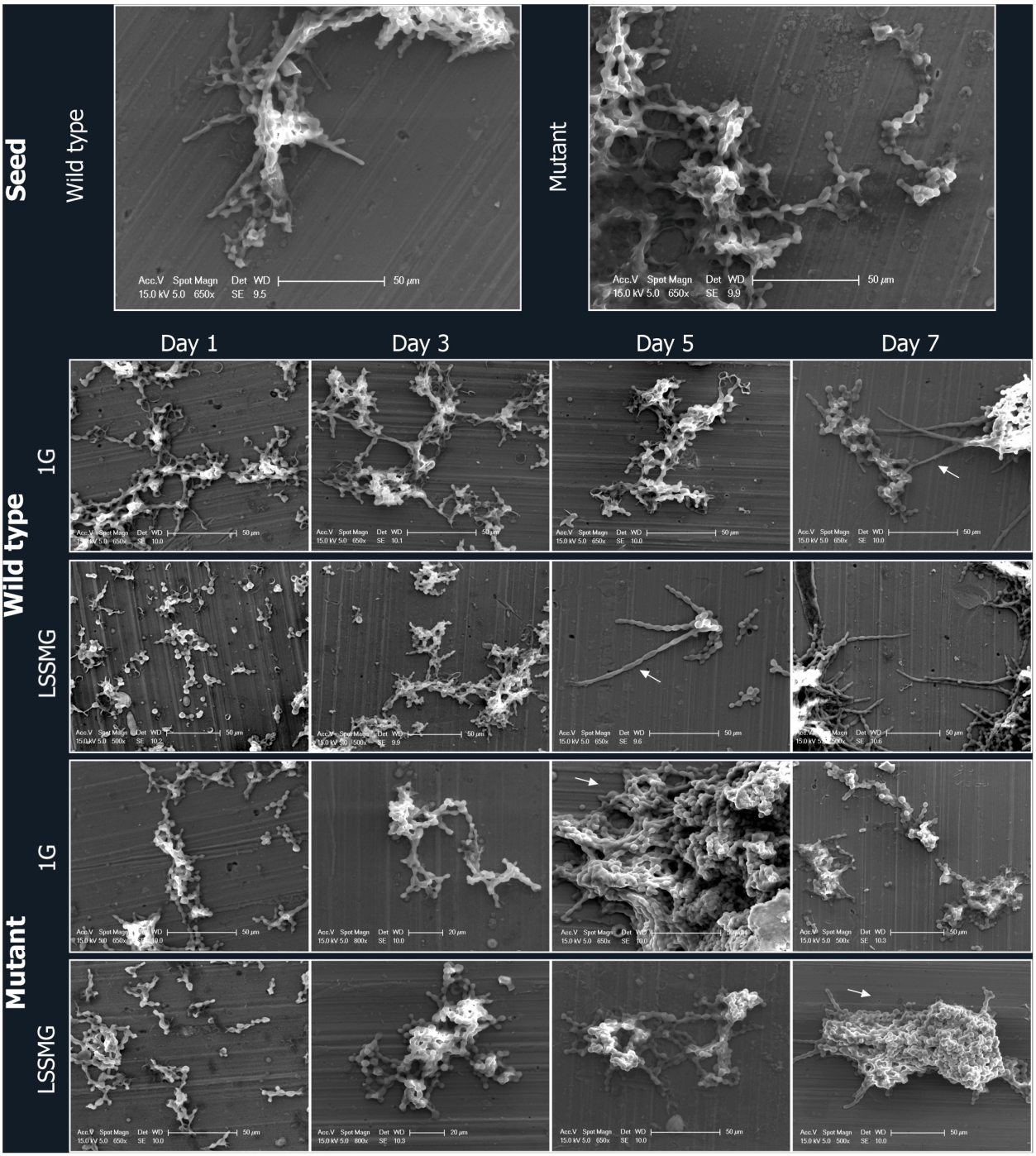

FIGURE 2 | Field emission scanning electron microscopy (FE-SEM) micrographs of *Knufia chersonesos* Wt and Mut on aluminum coupons during a 7-days exposure to LSSMG or 1G (control). Arrows indicate hyphal growth in the wild type and dense cell aggregation in the mutant strain.

secretome, while 575 were found to overlap (**Figure 4A**). A total of 3163 were identified by homology search. An overview of the biological and molecular functions of the detected proteins was obtained through an annotation statistics analysis. Distribution of the GO terms for all 3 categories (i.e., biological process BP, molecular function MF, and cellular component CC), with the highest number of associated protein sequences in the whole-cell proteome and the secretome, is displayed in **Supplementary Figures 2A,B**. Several GOs were common to proteome and secretome, whereas others were instead unique to each set of proteins. GOs for different types of cell metabolic processes were the most represented

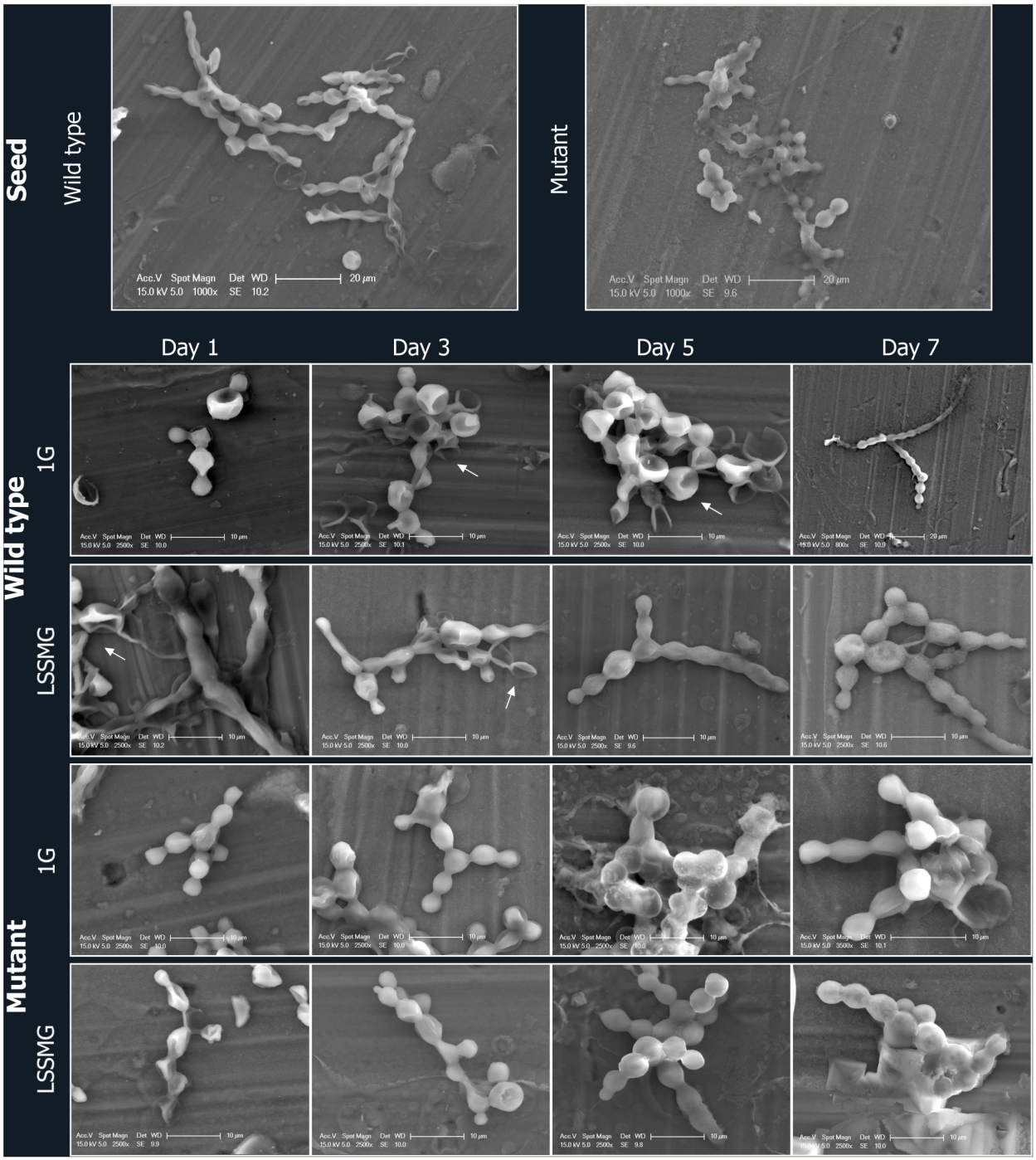

FIGURE 3 | FE-SEM micrographs of *Knufia chersonesos* Wt and Mut on aluminum coupons during a 7-days exposure to LSSMG or 1G (control). Details of cells morphology are shown; arrows indicate biconcave cell morphology.

in both secretory and intracellular proteins, whereas BPs such as response to chemical, cell cycle, and negative regulation of cellular processes were found exclusively within the latter. Conversely, proteins involved in cell communication and positive regulation of cellular process could solely be detected in the secreted fraction of the proteome. Proteins with hydrolase activity (MF GO:0016787) represented the most abundant group in the secretome and second most abundant in the whole-cell proteome, immediately followed by proteins with organic cyclic and heterocyclic compound binding, oxidoreductase and transferase activity. In both sets of samples, proteins with transcription and translation regulation activity were instead

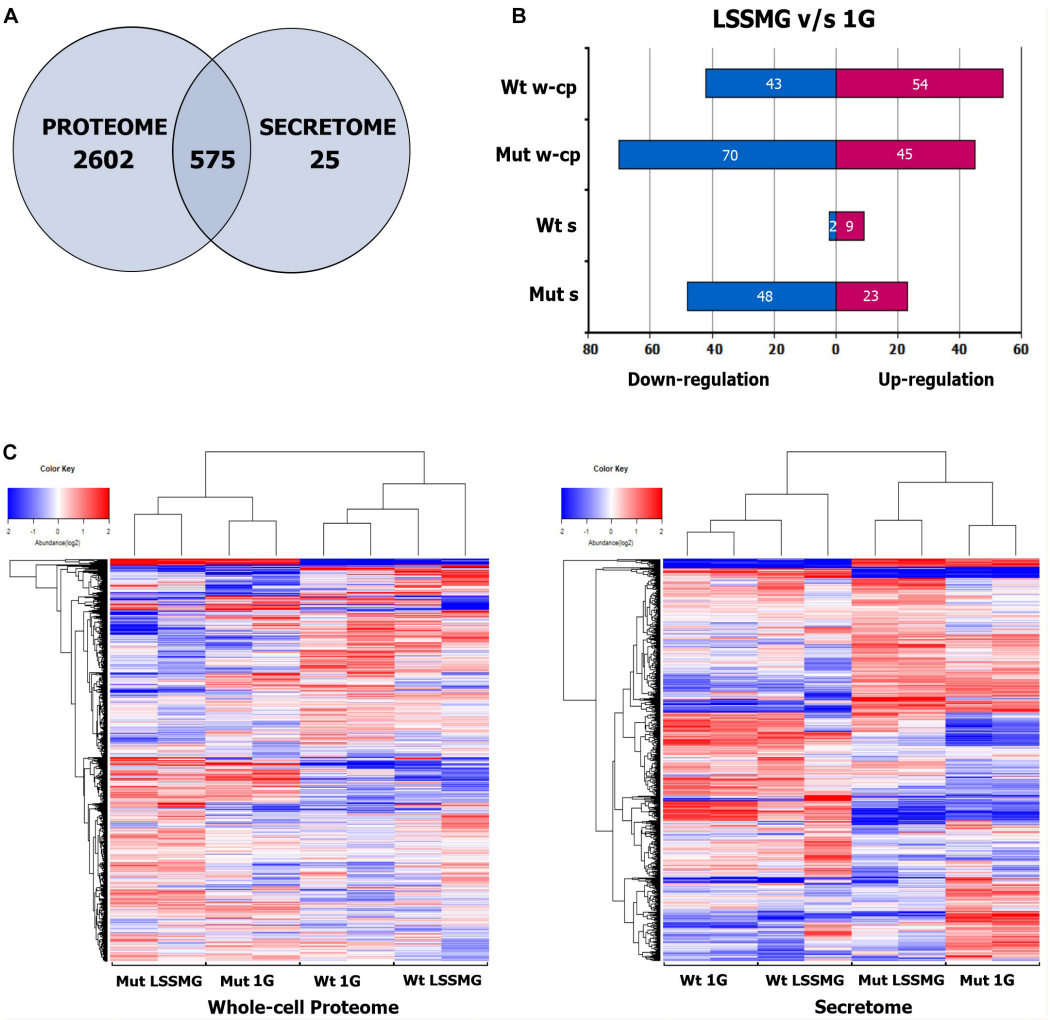

FIGURE 4 | Proteins identified and detected as regulated in the whole-cell proteome and secretome of *Knufia chersonesos* Wt and Mut following exposure to LSSMG. **(A)** Overlap of proteins between whole-cell proteome and secretome, identified in two separate TMT LC/MS experiments. **(B)** Number of differentially abundant proteins for each strain under LSSMG compared to 1G (control) (fold change of $\geq \pm 1.5$, $p \leq 0.05$). **(C)** Heat map hierarchical clustering of co-varying proteins quantified in each strain. The color key indicates the values that represented the protein abundance ratio (log2) and the number of proteins counted in each value. A total of 6 runs – 3 technical replicates and 2 biological replicates – were performed for each sample. Wt, wild type; Mut, mutant; w-cp, whole-cell proteome; s, secretome. Heatmaps were generated using R.

detected as the smallest groups. A distribution analysis for cellular component GO terms indicated a prevalence of intracellular, membrane, and cell wall proteins (i.e., CC GO: envelope) also in the culture supernatant—i.e., the ratio of total proteins with predicted extracellular to intracellular localization was approximately 25%:75% in the wild type and 6%:94% in the mutant—where the detection of cytoplasmic proteins is possibly due to mechanical stress and cell death (Miura and Ueda, 2018). Nonetheless, the presence of these proteins in the secreted fraction could also depend on active secretion of proteins with no predicted signal peptide, i.e., non-classical protein secretion, which has been recognized in various organisms including fungi, bacteria and plants based on the large number of leaderless proteins detected in extracellular compartments, including extracellular vesicles (EVs) (Regente et al., 2012; Vallejo et al., 2012; Sun and Su, 2019). The 25 proteins uniquely detected in the secretome encompassed a number of carboxylic ester hydrolases with predicted extracellular localization alongside cytoplasmic, cell wall– and plasma membrane-associated proteins (**Supplementary Table 3**).

In order to detect significant rearrangements in the protein pool upon exposure to LSSMG, a quantitative analysis was carried out by comparing experimental groups. The clustering of the samples, biological and technical replicates included, was confirmed by Hierarchical Cluster (HC) analysis (**Figure 4C**). However, one of the LSSMG-exposed secretomes from *K. chersonesos* Wt appeared to deviate from the biological duplicate of the same condition. The number of differentially abundant proteins for each strain is summarized in **Figure 4** (fold change of $\geq \pm 1.5$, $p \leq 0.05$), where (B) shows the count

of regulated proteins following direct comparison of treatment and control whole-cell proteome and secretome samples. The mutant proteome contained the largest number of modulated proteins (115) compared to the mutant secretome (71), the wild type proteome (97), and the wild type secretome (11).

Effects of LSSMG on *Knufia chersonesos* Whole-Cell Proteome

The proteomic quantitative analysis of LSSMG-exposed *K. chersonesos* Wt and Mut revealed 54-up and 43-down and 45-up and 70-down regulated proteins, respectively, when compared to the 1G-exposed counterparts (fold change of $\geq \pm 1.5$, $p \leq 0.05$) (**Supplementary Tables 4**, **5**). Distribution of over-represented BP GO terms among differentially expressed proteins is displayed in **Figure 5A**. Most significantly upregulated biological processes included "lipid and carbohydrate metabolism" in both strains (24 to 28% of all upregulated proteins). However, 15% of downregulated proteins were also in the carbohydrate metabolism category. By contrast, proteins involved in lipid metabolic processes were not found to be decreased. Further, "transcription" (12%) and "transport" (27%) were additional highly represented categories of upregulated proteins in the wild type, whereas "transport" (28%) and "amino acid metabolic process" (16%) were well represented categories of upregulated proteins in the mutant.

Proteins involved in lipid metabolism included CF317_002955-T1/C7ZGD6_NECH7—predicted to be phospholipase A2—an acetyl-CoA desaturase (CF317_004579-T1/A0A072PJ62_9EURO) and the uncharacterized protein CF317_000532-T1/A0A0D2C8U2_9EURO upregulated in both strains, along with an inositol-3-phosphate synthase (CF317_006213-T1/A0A0D2GI50_9EURO), which was instead solely observed in the mutant. These proteins participate in cellular pathways for the biosynthesis of unsaturated fatty acids (KEGG pathway 01040) and the metabolism of phospho- and glycerophospholipids (KEGG 00564), inositol phosphate (KEGG 00562), arachidonic and linoleic acid, among others (**Table 1**; Passoth, 2017). Pyruvate metabolism (KEGG 00620), glycolysis/gluconeogenesis (KEGG 00010) and starch and sucrose metabolism, were possibly also involved in the response to LSSMG, as revealed by the KEGG pathway analysis based on the increased abundance of proteins like alcohol dehydrogenase 1 (CF317_005767-T1/A0A0N0NIW7_9EURO) and alpha/beta-glucosidase agdC (CF317_006579-T1/A0A178BWA1_9EURO) in the wild type. Glutamine synthetase and protein FYV10, the first involved in glyoxylate and dicarboxylate metabolism (KEGG 00630) and in nitrogen metabolism (Zhang et al., 2017) and the latter mediating the degradation of enzymes of the gluconeogenesis pathway (Braun et al., 2011), were instead detected as upregulated in the mutant. Decreased levels of a number of hydrolytic enzymes destined for secretion were observed in the LSSMG-exposed proteomes for the carboxylic ester hydrolases CF317_0002308-T1/A0A1J9RJA8_9PEZI, CF317_007618-T1/A0A6A6HNE7_9PEZI, CF317_007621-T1/A0A0D2AG04_9PEZI (in Wt) and CF317_002086-T1/A0A0D2BHT9_9EURO (in Mut). The same was observed for the cell wall–degradation enzymes endo- and extracellular glucanases CF317_0009683-T1/W9ZBZ7_9EURO and CF317_009779-T1/H6BQE2_EXODN (in Wt) and the 3,2-*trans*-enoyl-CoA isomerase CF317_000932-T1/A0A1C1D1N7_9EURO (in Mut), the latter of which is involved in the metabolism of unsaturated fatty acids in beta oxidation (Janssen et al., 1994).

A differential abundance of proteins involved in transport was also observed (**Table 1**). Transmembrane ammonium (CF317_007500-T1/A0A072PWW7_9EURO) and iron transporters (i.e., HemS domain-containing protein CF317_008807-T1/L7HNA7_MAGOY) and the choline transport protein (CF317_009191-T1/A0A2P8A4S9_9PEZI) were present in the LSSMG-exposed whole-cell proteome of both strains, but they were more increased in the mutant (over 3-, 2. 4-, and 1.5-folds, respectively) than in the wild type (slightly over 1.5-folds). Additionally, two AA permease 2 domain-containing proteins (CF317_005369-T1/A0A0D2FEL3_9EURO and CF317_002132-T1/A0A438N399_EXOME) responsible for transmembrane amino acid transport were also upregulated in the mutant, whereas proteins with a predicted role in vesicle-mediated transport—i.e., VPS37 C-terminal domain-containing protein (CF317_007988-T1/A0A0D2CPQ3_9EURO) and WD_REPEATS_REGION domain-containing protein (CF317_001709-T1/A0A438MTR2_EXOME)—were exclusively detected in the wild type counterpart. Downregulated proteins included POT family proton-dependent oligopeptide transporter (CF317_001542-T1/A0A072P870_9EURO) and major facilitator superfamily transporters (MFS) (CF317_001911-T1/W9Z1V7_9EURO), involved in transportation of substrate molecules, including sugars, drugs, metabolites, amino acids, vitamins, and both organic and inorganic ions, or small peptides (Pao and Paulsen, 1998). Together with ABC transporters, MFS transporters are often detected among the cell wall components undergoing changes in response to the shift from normal gravity to microgravity (Sathishkumar et al., 2016).

Low-shear simulated microgravity additionally triggered upregulation of some proteins involved in cellular amino acid metabolic processes and downregulation of others (**Table 1**). Glutamate dehydrogenase was threefold and fourfold upregulated in *K. chersonesos* Wt and Mut, respectively. Together with the glutamine synthase, whose levels were also found to be increased in the mutant, the enzyme is reportedly involved in the primary nitrogen metabolism and in the biosynthesis and metabolism of several amino acids (i.e., arginine biosynthesis, glutamine, alanine and aspartate metabolism) (Meti et al., 2011). Protein D-3-phosphoglycerate dehydrogenase (CF317_004816-T1/A0A0D2CHE2_9EURO), participating in cysteine, methionine, glycine, serine, and threonine metabolism, was also more enriched in LSSMG-exposed mutant samples. However, a higher number of enzymes involved in cellular amino acid metabolic processes was found to be decreased under LSSMG. Nine were downregulated exclusively in the mutant, while protein CF317_006402-T1 was observed also in the wild type. According to the KEGG analysis, some of these proteins are involved with more

FIGURE 5 | Biological processes GOs categories of differentially expressed proteins in *Knufia chersonesos* under LSSMG. **(A)** Whole-cell proteome **(B)** Secretome. Proteins with changed abundance (FC $\geq \pm$ 1.5, $p \leq 0.05$) were annotated with terms representing various biological processes using OmicsBox v. 1.4.11 (https://www.biobam.com). GO terms were thereafter summarized using sequence distribution/GO multilevel pie charts (Filtered by sequence count, Cutoff = 2).

than one pathway: 4 proteins are implicated in arginine and proline metabolism (CF317_000519-T1, CF317_007970-T1, CF317_007070-T1, CF317_008030-T1), 3 in tyrosine metabolism (CF317_007412-T1, CF317_006402-T1, CF317_008924-T1), and 2 in histidine metabolism (CF317_007412-T1, CF317_006712-T1), arginine biosynthesis (CF317_002693-T1, CF317_000519-T1) and phenylalanine biosynthesis and metabolism (CF317_007412-T1, CF317_006402-T1). The putative indoleamine 2,3-dioxygenase, involved instead in tryptophan catabolic process to the NAD precursor kynurenine, represented the most decreased protein in both strains, with an over 2-fold downregulation.

Some proteins falling into the GO category "DNA-templated transcription" were enriched exclusively in the LSSMG-exposed wild type proteome. They were mostly phosphatases—known to also regulate pathways important for stress (Serra-Cardona et al., 2015)—having abundances nearly 2.5-fold higher than in the control samples (i.e., Alkaline

TABLE 1 | Most abundant biological process GO categories regulated under LSSMG in *Knufia chersonesos* Wt and Mut whole-cell proteome. Differentially expressed proteins included in each category are also shown.

***Protein* accession No.[a]**	**Putative Protein function**	**Protein relative abundance***	***p*-value**
Carbohydrate and lipid metabolic process			
Wild type			
CF317_003648-T1	GP-PDE domain-containing protein	1.1133	4.79E-02
CF317_002955-T1	Phospholipase A2	1.1032	3.26E-01
CF317_005767-T1	Alcohol dehydrogenase 1	0.8542	9.37E-02
CF317_006579-T1	Alpha/beta-glucosidase agdC	0.8177	6.11E-03
CF317_009031-T1	Carboxylic ester hydrolase	0.8024	4.22E-02
CF317_000532-T1	Uncharacterized protein	0.775	3.78E-01
CF317_007434-T1	Serine/threonine-protein kinase ppk6	0.6797	1.44E-01
CF317_004579-T1	Acyl-CoA desaturase	0.6334	1.95E-05
CF317_0002308-T1	Carboxylic ester hydrolase	−0.6053	1.47E-02
CF317_0009683-T1	Endo-1,3(4)-beta-glucanase	−0.6918	2.10E-02
CF317_007618-T1	Carboxylic ester hydrolase	−0.7365	4.22E-02
CF317_009779-T1	Extracellular cell wall glucanase Crf1/allergen Asp F9	−0.8811	2.70E-01
CF317_009003-T1	Serine/threonine-protein kinase TOR	−0.9558	9.75E-02
CF317_007621-T1	Cutinase	−1.0366	7.69E-04
Mutant			
CF317_009031-T1	Carboxylic ester hydrolase	1.1163	4.22E-02
CF317_007501-T1	Glutamine synthetase	0.9516	2.84E-01
CF317_006213-T1	Inositol-3-phosphate synthase	0.8851	2.34E-01
CF317_004579-T1	Acyl-CoA desaturase	0.8258	1.95E-05
CF317_000532-T1	Uncharacterized protein	0.7363	3.78E-01
CF317_007133-T1	Protein FYV10	0.6298	1.26E-02
CF317_002086-T1	AB hydrolase-1 domain-containing protein	−0.6016	1.98E-01
CF317_009017-T1	NodB homology domain-containing protein	−0.6135	5.38E-02
CF317_005813-T1	4HBT domain-containing protein	−0.6213	7.10E-02
CF317_001523-T1	Glucose-6-phosphate 1-epimerase	−0.6226	6.42E-02
CF317_000932-T1	3,2-*trans*-enoyl-CoA isomerase	−0.6255	1.20E-02
CF317_009683-T1	Endo-1,3(4)-beta-glucanase	−1.0502	2.10E-02
Transport			
Wild type			
CF317_004246-T1	Phosphate transporter	1.3345	4.85E-01
CF317_007988-T1	VPS37 C-terminal domain-containing protein	0.8635	5.08E-01
CF317_0005675-T1	Mitochondrial thiamine pyrophosphate carrier 1	0.7750	8.05E-03
CF317_001330-T1	Phosphate transporter	0.7548	1.35E-01
CF317_001709-T1	WD_REPEATS_REGION domain-containing protein	0.7277	5.37E-01
CF317_003773-T1	Zinc-regulated transporter 1	0.6633	1.42E-02
CF317_009191-T1	Choline transport protein	0.6590	9.34E-03
CF317_007500-T1	Ammonium transporter	0.6548	2.80E-01
CF317_008807-T1	HemS domain-containing protein	0.5860	2.75E-01
CF317_0006744-T1	Ribosomal protein L37e	−0.6649	1.96E-01
CF317_0001542-T1	POT family proton-dependent oligopeptide transporter	−0.6830	8.37E-03
CF317_0009683-T1	Endo-1,3(4)-beta-glucanase	−0.6918	2.10E-02
CF317_001911-T1	MFS transporter, SP family, major inositol transporter	−0.8427	1.84E-02
Mutant			
CF317_007500-T1	Ammonium transporter	1.6152	2.80E-01
CF317_009191-T1	Choline transport protein	1.2684	9.34E-03
CF317_005369-T1	AA_permease domain-containing protein	0.9504	4.30E-01
CF317_003222-T1	CNT family concentrative nucleoside transporter	0.7381	4.62E-01
CF317_008807-T1	HemS domain-containing protein	0.7308	2.75E-01
CF317_002497-T1	SEC7 domain-containing protein	0.7191	7.38E-02
CF317_009070-T1	Stress response protein NST1	0.6604	1.86E-01

(Continued)

TABLE 1 | Continued

Protein accession No.[a]	Putative Protein function	Protein relative abundance*	*p*-value
CF317_002132-T1	AA_permease domain-containing protein	0.6505	4.55E-02
CF317_001911-T1	MFS transporter, SP family, major inositol transporter	−0.7301	1.84E-02
CF317_001542-T1	POT family proton-dependent oligopeptide transporter	−0.8932	8.37E-03
CF317_009683-T1	Endo-1,3(4)-beta-glucanase	−1.0502	2.10E-02
Transcription, RNA and cellular amino acid metabolic processes			
Wild type			
CF317_004755-T1	Glutamate dehydrogenase	1.7627	3.54E-01
CF317_004246-T1	Phosphate transporter	1.3345	4.85E-01
CF317_008227-T1	Alkaline phosphatase	1.2107	4.76E-01
CF317_003892-T1	Alkaline phosphatase	1.0475	4.25E-01
CF317_005207-T1	Ribonuclease T1	0.9813	3.38E-01
CF317_005767-T1	Alcohol dehydrogenase 1	0.8542	9.37E-02
CF317_004551-T1	Carboxypeptidase	0.6315	5.66E-01
CF317_009661-T1	Fungal_*trans* domain-containing protein	−0.5936	4.26E-01
CF317_009392-T1	Fungal_*trans* domain-containing protein	−0.6382	1.93E-01
CF317_0006744-T1	Ribosomal protein L37e	−0.6649	1.96E-01
CF317_000112-T1	60S ribosomal protein L34-B	−0.7009	9.84E-02
CF317_000184-T1	Large subunit ribosomal protein L24e	−0.7218	2.77E-01
CF317_003175-T1	Poly(A) polymerase	−0.7450	9.79E-03
CF317_006402-T1	4-hydroxyphenylpyruvate dioxygenase	−0.7730	4.57E-04
CF317_009779-T1	Extracellular cell wall glucanase Crf1/allergen Asp F9	−0.8811	2.70E-01
CF317_009003-T1	Serine/threonine-protein kinase TOR	−0.9558	9.75E-02
CF317_002178-T1	Indoleamine 2,3-dioxygenase	−1.1838	5.88E-04
Mutant			
CF317_004755-T1	Glutamate dehydrogenase	1.9810	3.54E-01
CF317_007501-T1	Glutamine synthetase	0.9516	2.84E-01
CF317_001165-T1	Malic enzyme	0.6220	4.41E-01
CF317_004816-T1	D-3-phosphoglycerate dehydrogenase	0.6044	5.51E-03
CF317_002693-T1	Ornithine transcarbamylase	−0.5899	2.51E-02
CF317_000519-T1	Arginase	−0.6439	1.12E-02
CF317_006712-T1	Histidinol dehydrogenase	−0.6459	2.39E-02
CF317_008030-T1	D-amino-acid oxidase domain-containing protein	−0.6799	2.92E-02
CF317_008924-T1	Homogentisate 1,2-dioxygenase	−0.6970	5.01E-01
CF317_007970-T1	Multifunctional fusion protein	−0.7153	4.07E-01
CF317_000649-T1	Delta(3,5)-Delta(2,4)-dienoyl-CoA isomerase, mitochondrial	−0.7416	1.03E-02
CF317_007412-T1	Histidinol-phosphate aminotransferase	−0.7514	5.09E-02
CF317_003892-T1	Alkaline phosphatase	−0.8825	4.25E-01
CF317_007070-T1	Multifunctional fusion protein	−0.9090	6.03E-02
CF317_003175-T1	Poly(A) polymerase	−0.9882	9.79E-03
CF317_006402-T1	4-hydroxyphenylpyruvate dioxygenase	−1.1274	4.57E-04
CF317_002178-T1	Indoleamine 2,3-dioxygenase	−1.4328	5.88E-04
Response to stress and chemical			
Wild type			
CF317_004755-T1	Glutamate dehydrogenase	1.7627	3.54E-01
CF317_006388-T1	Nitric oxide dioxygenase	0.9459	3.16E-01
CF317_005767-T1	Alcohol dehydrogenase 1	0.8542	9.37E-02
CF317_003242-T1	Catalase	−0.6330	7.85E-02
CF317_009003-T1	Serine/threonine-protein kinase TOR	−0.9558	9.75E-02
CF317_007018-T1	3-phytase	−0.9947	4.67E-04
CF317_000687-T1	Catalase	−1.1410	6.34E-02
Mutant			
CF317_004755-T1	Glutamate dehydrogenase	1.9810	3.54E-01
CF317_006388-T1	Nitric oxide dioxygenase	0.7290	3.16E-01

(Continued)

TABLE 1 | Continued

Protein accession No.[a]	Putative Protein function	Protein relative abundance*	*p*-value
CF317_004909-T1	Adenylosuccinate synthetase	−0.6917	8.11E-02
CF317_007970-T1	Multifunctional fusion protein	−0.7153	4.07E-01
CF317_008612-T1	Aldedh domain-containing protein	−0.7312	2.05E-01
CF317_007412-T1	Histidinol-phosphate aminotransferase	−0.7514	5.09E-02
CF317_008426-T1	Glutathione reductase	−0.7680	8.19E-02
CF317_007018-T1	3-phytase	−0.8087	4.67E-04
CF317_000687-T1	Catalase	−0.8987	6.34E-02
CF317_008773-T1	Catalase-peroxidase	−0.9459	3.42E-02
Proteolysis			
Wild type			
CF317_008840-T1	Carboxypeptidase	1.0950	6.46E-01
CF317_004157-T1	Peptidase_M14 domain-containing protein	0.8180	2.52E-01
CF317_008998-T1	Putative fumarylacetoacetate hydrolase	0.7082	7.54E-02
CF317_000103-T1	Zinc carboxypeptidase	0.6512	1.88E-01
CF317_004551-T1	Carboxypeptidase	0.6315	5.66E-01
Cytoskeleton and organelle organization			
Wild type			
CF317_009683-T1	Endo-1,3(4)-beta-glucanase	−0.6918	2.10E-02
CF317_009003-T1	Serine/threonine-protein kinase TOR	−0.9558	9.75E-02
Mutant			
CF317_008503-T1	Thiamine thiazole synthase	0.8255	2.30E-01
CF317_007856-T1	WD_REPEATS_REGION domain-containing protein	0.7141	1.88E-02

**Log2 fold change of SMG-exposed compared to unexposed proteome ($p \leq 0.05$).*
[a]*Protein accession number in the K. chersonesos database of ab initio translated proteins.*

phosphatase CF317_008227-T1/A0A0D2BGD8_9EURO and CF317_003892-T1/A0A0D2BGD8_9EURO) (**Table 1**). Within the same category, downregulation was observed for a putative serine/threonine-protein kinase TOR and two *trans* domain-containing proteins involved in DNA binding and transcription. Decreased levels of structural protein constituents of the large ribosomal subunit—i.e., ribosomal protein L37e, L34-B and L24e—were instead detected upon LSSMG exposure exclusively in the wild type, alongside the upregulation of the ribonuclease T1 (CF317_005207-T1/W9XBV2_9EURO) involved in RNA degradation. Also significantly decreased was poly(A)polymerase (CF317_003175-T1/A0A1C1CN69_9EURO), an important component in mRNA synthesis, responsible for the addition of the 3′ polyadenine tail to a newly synthesized pre-messenger RNA.

In the group of proteins involved in response to stress and chemical, most of the modulated proteins were decreased by LSSMG exposure (**Table 1**). Downregulated stress response proteins included catalases (i.e., CF317_000687-T1/R8BNJ4_TOGMI, levels 2.7- and 1.9-fold lower than in the unexposed wild type and mutant proteome, respectively; CF317_003242-T1/R4XE87_TAPDE and CF317_008773-T1/U1G9G0_ENDPU), notably involved in the response to oxidative stress, and the 3-phytase CF317_007018-T1/A0A0D2IHR9_9EURO—a phosphatase enzyme with a predicted role in counteracting phosphate deficiency and osmotic stress (Belgaroui et al., 2018)—in both strains. Further, the glutathione-recycling enzyme glutathione reductase (CF317_008426-T1/A0A0D1ZJU8_9EURO) (Couto et al., 2016), the histidinol-phosphate aminotransferase (CF317_007412-T1/H6BQ73_EXODN) and the adenylosuccinate synthetase (ASS; F317_004909-T1/W9W243_9EURO), showed lower levels in the mutant under LSSMG conditions. The latter, best known as the first enzyme in the *de novo* synthesis of AMP from inosine-5′-monophosphate (IMP), is responsive to salt stress, as reported in plants (Zhao et al., 2013). Among the proteins with increased abundance in both fungal strains, was glutamate dehydrogenase (CF317_004755-T1), involved in amino acid metabolism and for which a stress-dependent regulation has been demonstrated (e.g., abiotic stress generated ROS) (Skopelitis et al., 2006). Further, the putative cytosolic nitric oxide dioxygenase (CF317_006388-T1/S7Z8A1_PENO1)—reported to play a role in nitric oxide detoxification mechanisms by the conversion of this signaling molecule to nitrate (Cánovas et al., 2016)—was present in the wild type and mutant LSSMG-exposed proteomes at levels around 2-fold higher than that of the unexposed ones. Along with the above-mentioned proteins, a number of enzymes, specifically peptidases involved in proteolysis, were found to be upregulated exclusively in *K. chersonesos* wild type (**Table 1**). The putative carboxypeptidase CF317_008840-T1/W9XS38_9EURO and CF317_004157/W2S3B3_9EURO; and CF317_000103/A0A6A6DDQ8_9PEZI and CF317_004551/A0A0D2GVY9_9EURO, known to catalyze reactions that are important to various physiological processes, such as the cell cycle, cell growth and differentiation, apoptosis, and stress response (Neto et al., 2018), were 2-fold and

1.5-fold upregulated, respectively. Lastly, proteins involved in cytoskeleton and organelle organization showed opposite regulation in the two *K. chersonesos* strains: The first was decreased in the wild type, whereas the latter was upregulated in the mutant (**Table 1**).

Effects of LSSMG on *Knufia chersonesos* Secretome

The proteomic characterization of *K. chersonesos* secretome upon exposure to LSSMG for 7 days revealed solely 9-up and 2-down regulated proteins in the wild type and 23-up and 48-down regulated proteins in the mutant, when compared to the 1G-exposed counterparts (fold-change of $\geq \pm$ 1.5, $p \leq 0.05$) (**Figure 4B**). Among these proteins, the ratio of proteins with predicted extracellular to intracellular localization was approximately 20%:80% in both wild type and mutant (**Supplementary Tables 6**, **7**). Out of 14 proteins with predicted signal peptides in the mutant secretome, 12 were regulated following LSSMG-exposure. In the wild type secretome, 2 out of 61 proteins with predicted extracellular localization were regulated in response to microgravity.

The distribution of differentially expressed proteins among BP GO terms is presented in **Figure 5B**. In the wild type, the regulated proteins were prevalently involved in transport and metabolic processes, whereas in the mutant strain, regulated proteins were mostly involved with biological processes such as carbohydrate and lipid metabolism (15 proteins), response to chemical and stress (13 proteins), translation, and ribosome biogenesis and transcription (15 proteins). Interestingly, the majority of proteins involved in the above-mentioned GO BP categories exhibited downregulation in LSSMG-exposed secretomes of the mutant and exhibited upregulation in the wild type (**Table 2**). One protein exhibiting opposite regulation in the two analyzed strains, was the putative scytalone dehydratase CF317_002654-T1, homolog of *Phialophora attinorum* A0A0N1P280_9EURO. This enzyme, which is reportedly involved in the biosynthesis of dihydroxynaphthalene (DNH) melanin from endogenous substrate (Eisenman and Casadevall, 2012), was 1.5-fold increased in the wild type and 1.6-fold decreased in the melanin-deficient mutant. In the wild type, upregulated proteins also included the adenylosuccinate lyase CF317_004119-T1 and the M20_dimer domain-containing protein (Carboxypeptidase S) CF317_006825-T1, both implicated in amino acid (alanine, aspartate and glutamate) metabolism (KEGG pathway pae00250), and the plasma membrane ATPase CF317_002664-T1, partaking in oxidative phosphorylation (**Table 2**). Also in the wild type, the casein kinase I 1 CF317_004827-T1 and the histone H2A CF317_001633-T1, both involved in DNA repair (Skoneczna et al., 2018), were found in the LSSMG-exposed secretome at levels 1.7- and 1.6-fold higher than in the unexposed one. The two downregulated proteins were protein CF317_008131-T1, homolog of the arabinan endo-1,5-alpha-L-arabinosidase from *Aspergillus wentii DTO 134E9*, and protein CF317_002949-T1, homolog of the feruloyl esterase A0A177BY00_9PLEO from *Paraphaeosphaeria sporulosa,* both with a predicted extracellular localization and reportedly involved in polysaccharide degradation (Tesei et al., 2020).

In the *K. chersonesos* mutant, LSSMG triggered the upregulation of just a low number of proteins. The malate dehydrogenase CF317_003266-T1/A0A1C1CNC5_9EURO, 1,3-beta-D-glucan-UDP glucosyltransferase CF317_000614-T1/A0A438MTW5_EXOME and alpha, alpha-trehalase CF317_004588-T1/H6C927_EXODN, involved in the metabolism of starch, sucrose, and sphingolipids, were 2. 4-, 1.9-, and 1.8-fold upregulated, respectively. Conversely, proteins with decreased abundance encompassed, among others, enzymes catalyzing different steps throughout the process of glycolysis/gluconeogenesis (KEGG pathway hsa00010)—such as the putative phosphotransferase CF317_008448/W2RU25_9EURO, the phosphoglycerate kinase CF317_004751/W2RJE3_9EURO and the 2-phosphoglycerate dehydratase CF317_003513/H6BNI5_EXODN—fructose and mannose (hsa00051), pyruvate (hsa00620) and glycerolipid metabolism (hsa00561), and the pentose phosphate pathway (hsa00030) (i.e., phosphotransferase, epimerase domain-containing protein CF317_008979 and 6-phosphogluconate dehydrogenase, decarboxylating CF317_005481). Further, the CF317_007096/A0A0U1M0E3_TALIS 6-phosphofructo-2-kinase, acting as an activator of the glycolysis/gluconeogenesis pathway (Raben and Wolfgang, 2019), was the most decreased protein in the GO category of carbohydrate and lipid metabolism, with a 3-fold downregulation.

Increased levels of proteins with a role in the response to stress and chemicals (**Table 2**) were observed in the exposed secretomes for the plasma membrane protein CF317_000051-T1/A0A178DA36_9EURO with reported phosphatase activity and the guanine nucleotide-binding protein subunit beta CF317_007665-T1/A0A0D2CV56_9EURO with a role in signal transduction, among others. However, as already observed for the whole-cell proteome, the majority of proteins involved in stress response showed decreased abundance upon exposure to LSSMG. The DNA-repair protein histone H2A (CF317_001633-T1), upregulated in the wild type secretome, was at least twofold downregulated in the mutant, together with the nucleoside diphosphate kinase CF317_007147-T1/A0A178C9G0_9EURO. The latter, ubiquitous enzyme involved in nucleotides biosynthetic process (Janin and Deville-Bonne, 2002), has been shown to participate in the metabolism of selective drugs like Isoniazid and Fluorouracil (KEGG map00983) in *Homo sapiens*. The same was observed for the S-formylglutathione hydrolase, whose activity is mainly linked to methane metabolism and the recycling of glutathione in cell detoxification pathways (Haslam et al., 2002; Yurimoto et al., 2003). Similarly, the thioredoxin domain-containing protein CF317_000980-T1 and the 6-phosphogluconate dehydrogenase, decarboxylating CF317_005481-T1/A0A0N0NLG1_9EURO, both involved in cell redox homeostasis and glutathione metabolism (Li et al., 2019), showed an around twofold downregulation.

Also in the categories of transcription, translation, and ribosome biogenesis, downregulated proteins outnumbered the upregulated ones (**Table 2**). The most increased protein (almost twofold) was the putative D-amino-acid

TABLE 2 | Most abundant biological process GO categories regulated under LSSMG in *Knufia chersonesos* Wt and Mut whole-cell secretome. Differentially expressed proteins included in each category are shown.

***Protein* accession No.[a]**	**Putative Protein function**	**Protein relative abundance***	***p*-value**
Carbohydrate and lipid metabolic processes			
Wild type			
CF317_004827-T1	Casein kinase I 1	0.656	4.63E-02
CF317_002949-T1	Feruloyl esterase	−0.859	1.16E-03
CF317_008131-T1	Arabinan endo-1,5-alpha-L-arabinosidase	−0.998	4.32E-02
Mutant			
CF317_003266-T1	Malate dehydrogenase	1.244	3.48E-02
CF317_000614-T1	1,3-beta-D-glucan-UDP glucosyltransferase	0.893	2.26E-02
CF317_004588-T1	Alpha, alpha-trehalase	0.813	2.42E-02
CF317_003037-T1	Acyl-CoA-dependent ceramide synthase	0.797	1.16E-03
CF317_000059-T1	PSDC domain-containing protein	−0.634	7.14E-03
CF317_004317-T1	3′(2′),5′-bisphosphate nucleotidase	−0.650	4.52E-02
CF317_008979-T1	Epimerase domain-containing protein	−0.727	2.36E-02
CF317_003714-T1	Mannitol-1-phosphate 5-dehydrogenase	−0.739	9.36E-05
CF317_008448-T1	Phosphotransferase	−0.790	4.34E-04
CF317_003513-T1	2-phosphoglycerate dehydratase	−0.870	1.53E-03
CF317_006165-T1	Esterase/lipase	−0.917	9.78E-03
CF317_003202-T1	Concanavalin A-like lectin/glucanase	−0.918	4.79E-02
CF317_005481-T1	6-phosphogluconate dehydrogenase, decarboxylating	−0.960	7.57E-04
CF317_004751-T1	Phosphoglycerate kinase	−1.143	8.27E-04
CF317_007096-T1	6-phosphofructo-2-kinase	−1.494	5.59E-03
Transport			
Wild type			
CF317_001911-T1	MFS transporter, SP family, major inositol transporter	0.747	2.92E-02
CF317_002664-T1	Plasma membrane ATPase	0.702	3.25E-02
CF317_004827-T1	Casein kinase I 1	0.656	4.63E-02
Mutant			
CF317_000614-T1	1,3-beta-D-glucan-UDP glucosyltransferase	0.893	2.26E-02
CF317_001401-T1	t-SNARE coiled-coil homology domain-containing protein	0.843	4.08E-02
CF317_004472-T1	Endoplasmic reticulum transmembrane protein	0.742	4.09E-04
CF317_006347-T1	Inorganic phosphate transport protein PHO88	0.588	4.10E-02
CF317_002531-T1	Putative inorganic phosphate transporter C8E4.01c	−0.591	1.01E-03
CF317_007808-T1	NTF2 domain-containing protein	−0.756	2.95E-02
CF317_008448-T1	Phosphotransferase	−0.790	4.34E-04
Transcription, RNA and cellular amino acid metabolic processes			
Wild type			
CF317_001633-T1	Histone H2A	0.752	9.78E-03
CF317_002935-T1	Aromatic amino acid aminotransferase	0.633	1.18E-02
Mutant			
CF317_008030-T1	D-amino-acid oxidase domain-containing protein	0.783	1.10E-02
CF317_003734-T1	Ribosomal_L23eN domain-containing protein	0.771	4.09E-02
CF317_007665-T1	Guanine nucleotide-binding protein subunit beta	0.775	2.98E-03
CF317_005931-T1	Aspartate–tRNA ligase	0.742	2.84E-02
CF317_004317-T1	3′(2′),5′-bisphosphate nucleotidase	−0.650	4.52E-02
CF317_003885-T1	Putative RNA-binding protein	−0.668	4.83E-03
CF317_001861-T1	40S ribosomal protein S20	−0.709	4.62E-03
CF317_008924-T1	Homogentisate 1,2-dioxygenase	−0.730	4.76E-02
CF317_005516-T1	60S ribosomal protein L7	−0.734	1.03E-03
CF317_008448-T1	Phosphotransferase	−0.790	4.34E-04
CF317_004501-T1	40S ribosomal protein S24	−0.796	1.53E-02
CF317_001633-T1	Histone H2A	−1.084	9.78E-03
CF317_007960-T1	60S acidic ribosomal protein P1	−1.277	9.18E-03

(Continued)

TABLE 2 | Continued

Protein accession No.[a]	Putative Protein function	Protein relative abundance*	*p*-value
CF317_008933-T1	Elongation factor EF-1 beta subunit	−1.706	2.09E-02
CF317_000379-T1	Elongation factor EF-1 gamma subunit	−1.838	2.56E-03
Response to stress and chemical			
Wild type			
CF317_001633-T1	Histone H2A	0.752	9.78E-03
CF317_004827-T1	Casein kinase I 1	0.656	4.63E-02
Mutant			
CF317_000051-T1	Plasma membrane phosphatase required for sodium stress response	0.906	1.73E-02
CF317_004588-T1	Alpha, alpha-trehalase	0.813	2.42E-02
CF317_007665-T1	Guanine nucleotide-binding protein subunit beta	0.775	2.98E-03
CF317_006969-T1	HRXXH domain-containing protein	0.696	4.88E-02
CF317_004317-T1	3′(2′),5′-bisphosphate nucleotidase	−0.650	4.52E-02
CF317_003570-T1	S-formylglutathione hydrolase	−0.720	1.69E-02
CF317_008979-T1	Epimerase domain-containing protein	−0.727	2.36E-02
CF317_003714-T1	Mannitol-1-phosphate 5-dehydrogenase	−0.739	9.36E-05
CF317_003513-T1	2-phosphoglycerate dehydratase	−0.870	1.53E-03
CF317_005481-T1	6-phosphogluconate dehydrogenase, decarboxylating	−0.960	7.57E-04
CF317_000980-T1	Thioredoxin domain-containing protein	−0.983	2.08E-02
CF317_001633-T1	Histone H2A	−1.084	9.78E-03
CF317_007147-T1	Nucleoside diphosphate kinase	−1.745	8.70E-04
Biological processes and secondary metabolic process			
Wild type			
CF317_001633-T1	Histone H2A	0.752	9.78E-03
CF317_004119-T1	Adenylosuccinate lyase	0.703	1.58E-02
CF317_006825-T1	M20_dimer domain-containing protein	0.588	1.22E-02
CF317_002654-T1	Scytalone dehydratase	0.587	3.70E-02
CF317_002949-T1	Feruloyl esterase	−0.859	1.16E-03
Mutant			
CF317_000614-T1	1,3-beta-D-glucan-UDP glucosyltransferase	0.893	2.26E-02
CF317_006969-T1	HRXXH domain-containing protein	0.696	4.88E-02
CF317_003885-T1	Putative RNA-binding protein	−0.668	4.83E-03
CF317_002654-T1	Scytalone dehydratase	−0.679	3.70E-02
CF317_006992-T1	Serine/threonine-protein kinase	−0.682	2.90E-02
CF317_008448-T1	Phosphotransferase	−0.790	4.34E-04
CF317_003513-T1	2-phosphoglycerate dehydratase	−0.870	1.53E-03
CF317_000980-T1	Thioredoxin domain-containing protein	−0.983	2.08E-02
CF317_006154-T1	Pyruvate decarboxylase	−1.138	6.87E-03
CF317_003229-T1	Putative versicolorin reductase	−1.176	2.61E-03
CF317_007960-T1	60S acidic ribosomal protein P1	−1.277	9.18E-03
CF317_006677-T1	HIT domain-containing protein	−1.770	8.90E-03

**Log2 fold change of SMG-exposed compared to unexposed proteome ($p \leq 0.05$).*
[a]*Protein accession number in the K. chersonesos database of ab initio translated proteins.*

oxidase (DAO) domain-containing protein CF317_008030-T1/A0A0D1YYX4_9EURO, also involved in penicillin and cephalosporin biosynthesis (KEGG 00311). Elongation factor EF-1 beta and gamma subunit (CF317_008933-T1 and CF317_000379-T1), homologs of *Exophiala dermatitidis* CBS 525.76 H6BVG8_EXODN and H6BY84_EXODN proteins and involved in purine metabolism (KEGG 00230), were detected in the mutant LSSMG-exposed secretome at levels at least 3.5-fold lower than that of the unexposed ones. Further, four ribosomal proteins—i.e., 40S ribosomal protein S20 and S24 (CF317_001861-T1 and CF317_004501-T1); 60S ribosomal protein P1 and L7 (CF317_007960-T1 and CF317_005516-T1)—showed decreased abundance.

Additionally downregulated in response to LSSMG were the uncharacterized HIT domain-containing protein CF317_006677-T1 (over 3-fold regulated) and the putative versicolorin reductase CF317_003229-T1 (over 2-fold regulated), the latter known to mediate aflatoxins biosynthetic processes (Nakamura et al., 2011). Modulation was also observed in proteins involved in transport such as the above-mentioned

1,3-beta-D-glucan-UDP glucosyltransferase CF317_000614-T1 with a role on starch and sucrose metabolism, the endoplasmic reticulum transmembrane protein CF317_004472-T1, and the inorganic phosphate transport protein PHO88 CF317_006347-T1, which all appeared to be upregulated. Conversely, the transmembrane transporter "putative inorganic phosphate transporter C8E4.01c" CF317_002531-T1, the uncharacterized NTF2 domain-containing protein CF317_007808-T1, and the glycolytic enzyme phosphotransferase CF317_008448-T1 were detected as downregulated.

Knufia chersonesos Wt and Mut Exhibit Opposite Regulation of Several Differentially Expressed Proteins

The comparative analysis of all examined samples to a reference sample—a pool of all established experimental conditions—revealed qualitative and quantitative differences between *K. chersonesos* Wt and Mut at the proteome and secretome level. This was especially evident when examining the top 10 differentially regulated proteins i.e., the 10 most up- and downregulated proteins at each experimental condition. A number of these proteins were found to undergo opposite modulation in the two strains, mainly in the whole-cell proteome. Distribution of the top differentially regulated whole-cell proteins is summarized in **Supplementary Table 8**: out of 41 different proteins, 6 were found to be top regulated in wild type and mutant under both normal gravity and microgravity condition, 10 only in the Wt and 10 exclusively in the Mut. Around 25% represented ribosomal proteins; 10% were proteins whose identity or function could not be elucidated after homology search; and the remaining proteins were involved in RNA binding, transcription, translation, transport, and carbohydrate metabolism. Out of the 49 top differentially regulated proteins from the secreted fraction, 2 were found in both strains at all experimental conditions, 17 only in the Wt and 17 only in the Mut (**Supplementary Table 9**). Similar to what was observed for the proteome, 26% of the detected proteins were ribosomal, followed by enzymes involved in carbohydrate metabolism (18%)—secreted esterases and lipases included—and in transport (10%).

Altogether, the eight top differentially regulated proteins common to wild type and mutant in both normal gravity and microgravity—i.e., the large subunit (LSU) ribosomal proteins L28 (CF317_000935-T1), L32 (CF317_002712-T1), L35 (CF317_001879-T1) and L36 (CF317_007709-T1), the small subunit (SSU) ribosomal protein S30 (CF317_005321-T1), the U1 small nuclear ribonucleoprotein C (CF317_001622-T1), the murein transglycosylase (CF317_006383-T1) and the uncharacterized protein CF317_007965-T1—displayed an LSSMG-dependent regulation that also appeared to be specific to the Wt strain (**Supplementary Figure 3**). As shown in **Supplementary Figure 3**, all proteins that are upregulated in the Wt strain appear to be downregulated in the Mut strain and vice versa.

A protein-protein interaction analysis was performed to verify the occurrence of these 8 proteins in common pathways, using STRING. A 6-node network was obtained for the proteins which matched homologs (sequence homology) in the STRING database (**Supplementary Figure 4**) and an interaction was confirmed for ribosomal proteins S30 (homolog of *E. dermatitidis* XP_009156842.1) and L35 (homolog of XP_009155966.1). As a component of the SSU, S30 has a role in mRNAs binding and selection of cognate aminoacyl-transferase RNA (tRNA), whereas L35 is required for polymerization of the amino acids delivered by tRNAs into a polypeptide chain (Ben-Shem et al., 2011). The involvement of protein L35 in cell growth regulation and apoptosis has additionally been reported (Bommer and Stahl, 2005). The proteins' functional link (combined score: 0.999) was supported by evidence such as the co-expression of orthologs (score 0.963) and their interaction in other organisms (experimental/biochemical data, score: 0.904). Further interactions could not be detected possibly because functional characterization of these proteins in *E. dermatitidis* is yet to be thoroughly achieved.

Knufia chersonesos Wt and Mut Comparative Analysis Under Normal Gravity (1G)

A comparative analysis of wild type and mutant was performed to evaluate different responses at the proteome and secretome level under control condition, i.e., normal gravity. The analysis of the whole-cell proteome yielded a total of 100 proteins with increased and 40 proteins with decreased abundance in *K. chersonesos* Mut, compared with the wild type (fold change of $\geq \pm 1.5$, $p \leq 0.05$).

As shown in **Supplementary Figure 5A**, "protein translation," "organelle organization," and "regulation of biological processes" were detected as the prevalent processes in the mutant, encompassing more than 50% of all upregulated proteins. This includes several ribosomal proteins and ribonucleoproteins. Ribosomal proteins L32 (CF317_002712), L36 (CF317_007709), S30 (CF317_005321) and L37 (CF317_004524-T1) and the small nuclear ribonucleoprotein C (CF317_001622-T1) were among the most upregulated proteins, being present in the mutant at level 8.7 to 3.6 (3.1 to 1.85 log2 fold change) higher than in the wild type (**Supplementary Table 10**). Increased levels were also observed for proteins involved in RNA metabolic processes and in transcription such as the U3 small nucleolar RNA-associated protein 11 (A0A0D2DFT3_9EURO), the U6 snRNA-associated Sm-like protein LSm4 (CF317_002219-T1) and the pre-mRNA-splicing factor isy1 (CF317_006097-T1),—partaking in nucleolar processing of pre-18S ribosomal RNA and in pre-mRNA splicing (Zhou et al., 2014)—the transcription factors C2H2 and RfeG (CF317_009591-T1 and CF317_000265-T1) and the 2′-phosphotransferase (CF317_002731-T1). Protein modulation also affected enzymes involved in the oxidative phosphorylation, whose levels resulted to be increased in *K. chersonesos* Mut, i.e., the cytochrome c oxidase assembly factor 6 (CF317_000645-T1 and CF317_000273-T1), cytochrome b-c1 complex subunit 7 (CF317_008279-T1), NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit CF317_000197-T1. Regulation of proteins involved in chemical and stress response was also observed. The histone H2A (CF317_001633-T1)—reported to have a role on DNA repair (Moore et al., 2007)—and the Nuclear transcription

factor Y alpha (CF317_006017-T1)—with multiple roles in development and stress response (Zhao et al., 2017)—were found at levels at least 1.6-fold higher than in the melanized counterpart. Conversely, downregulated processes at normal gravity condition prevalently encompassed carbohydrate, lipid and general cellular metabolism (**Supplementary Figure 5A**). Decreased levels were observed for glycoside hydrolases and esterolytic enzymes including the alpha-galactosidase (CF317_004369-T1) and cutinases (CF317_007621-T1)—levels 9.7- and 7-fold lower than in the wild type—for the glucan 1,3-beta glucosidase (CF317_002004-T1; 1.69-folds decreased) and for the murein transglycosylase (CF317_004960-T1, CF317_003460-T1, CF317_006383-T1; 1.83-, 2- and 4-fold decreased), enzymes responsible for polymers and microbial degradation by cleavage at the peptidoglycan or the cell wall polysaccharides level (Lincoln et al., 1997). Additionally, proteins involved in lipid metabolic process—i.e., acetyl-CoA C-acetyltransferase (CF317_005988-T1), isocitrate lyase (CF317_004212-T1) and lipase (CF317_005885-T1)—and in protein catabolism—i.e., proteasome subunit alpha type (CF317_008861-T1), proteasome endopeptidase complex (CF317_004788-T1) and peptidase S53 domain-containing protein (CF317_005049-T1)—showed decreased levels between 1.8- and 1.5-fold.

At the secretome level, the quantitative analysis of *K. chersonesos* Mut v/s Wt under normal gravity conditions, resulted in the detection of 104 proteins with increased and 87 proteins with decreased abundance in the mutant secretome. Of these proteins, 24% have predicted extracellular localization based on the presence of a signal peptide (**Supplementary Table 11**). Distribution of over-represented biological process GO terms among differentially expressed proteins is displayed in **Supplementary Figure 5B**. Most of proteins upregulated in the mutant were involved with carbohydrate metabolism and regulation of biological process (24% of all upregulated proteins), transport (9%), response to stress (8%) and transcription (8%). A number of hydrolases required for the breakdown of β-glucan chains and other cell-wall components— i.e., carboxylic ester hydrolyses (CF317_004653-T1, CF317_008040-T1, CF317_002308-T1, CF317_005799-T1), feruloyl esterase (CF317_002949-T1), alongside glucan 1,3-beta-glucosidase (CF317_002004-T1), murein transglycosylase (CF317_006383-T1) and cutinase (CF317_007621-T1) which were downregulated in the mutant proteome—were present in the secretome at levels comprised between 1.5- and 4-folds higher than in the wild type (**Supplementary Table 11**). The same for mannan endo-1,6-alpha-mannosidase (CF317_001276-T1) and chitin deacetylase (CF317_003258-T1), which play multiple roles in the function of the fungal cell wall (Spreghini et al., 2003; Mouyna et al., 2020). In the category transport, the multicopper oxidase (CF317_009669-T1; 1.9-fold), the chloride channel protein (CF317_004136-T; 1.7-folds) and the nitrate/nitrite transporter (CF317_004846-T1; 1.6-fold) represented the most upregulated proteins. Increased levels of stress-response proteins were observed for the Histone H2A protein (CF317_001633-T1; 1.7-fold)—found to be overexpressed also in the mutant proteome—alongside the redox protein thioredoxin (CF317_006109-T1; 1.7-fold), catalase A (CF317_009216-T1; 1.7-fold) and the housekeeping enzyme nucleoside diphosphate kinase (CF317_007147-T1; 2.3-fold), whose involvement in the signaling pathway of oxidative stresses has been reported (Desorption et al., 2000). Normal gravity condition was also characterized by increased levels of peptidyl-prolyl *cis-trans* isomerase (CF317_005659-T1), EF-1 gamma subunit (CF317_000379-T1), and alkaline phosphatase (CF317_008227-T1; CF317_003892-T1), all involved in transcription. Conversely, decreased levels were observed in the secretome almost exclusively for proteins involved in translation and ribosome biogenesis (**Supplementary Figure 5B**). Out of 87 downregulated proteins, more than 50 encompassed ribosomal proteins. Of these, the most downregulated ones—the 60S ribosomal proteins L32 (CF317_006070-T1), L35 (CF317_001879-T1) and L33-A (CF317_005595-T1)—were present at levels below 5-fold than their counterparts in the wild type secretome. Other proteins were prevalently involved in stress/chemical response, e.g., plasma membrane phosphatase (CF317_000051-T1), CipC-like antibiotic response protein (CF317_003991-T1), small heat shock protein (SHSP) domain-containing protein (CF317_001628-T1). These findings pose interesting questions regarding the presence of intracellular proteins in *K. chersonesos* secretome: while it may be indicative of cell lysis during cultivation, it may also suggest that ribosomal and other non-classical secretory proteins are found in the culture supernatant due to EVs-mediated secretion (Sun and Su, 2019). The co-isolation of these and classical secretory proteins during sample preparation can occur if no prior separation of the vesicle containing fraction from the secretome, is performed (Vallejo et al., 2012).

DISCUSSION

To date, investigations into the responses of microorganisms to space and Mars-like conditions have been performed with few black fungi species; some of them involved stress simulation in ground-based facilities (Pacelli et al., 2017), while others carried out the exposure of fungal strains inside or outside the International Space Station (Onofri et al., 2012, 2019; Pacelli et al., 2016).

With only one exception (Zakharova et al., 2014)—the comparative study of 2D protein patterns under Mars-like conditions—the majority of astrobiological work carried out with black fungi has mainly focused on the analysis of ultrastructural alterations and DNA integrity. Hence, this study represents the first qualitative and quantitative proteomic characterization of a black fungus response to simulated space conditions, i.e., microgravity, which has significance to exobiology and implications to planetary protection policy. In-depth understanding of proteome-related alterations in cell physiology is crucial to gaining new insight into the evolution of extremophiles and the actual limits for life. Further, the comparative analysis of a melanin-deficient mutant and the wild type is important to evaluate the role of melanization in stress survival.

Under microgravity conditions, a development toward a more clump-like growth was observed in both *K. chersonesos* Wt and

Mut and may be due to the low shear and low turbulence suspension culture environment created by LSSMG. However, morphological differences were not detected via FE-SEM in cells grown under LSSMG compared to those grown in 1G. This is not surprising, as black fungi reportedly resort to strategies to minimize efforts at both the morphological and physiological level when exposed to stress conditions (Sterflinger, 2006). Microcolonial growth and the switch among different growth forms—i.e., budding and hyphae formation—ensure that these organisms have a lifestyle versatility to cope with a variety of stress in their natural habitats (Sterflinger et al., 1999; Gostinčar et al., 2011). Unaltered morphology upon exposure to LSSMG was previously described in filamentous fungi in both suspension (e.g., *Aspergillus niger* and *Penicillium chrysogenum*) (Sathishkumar et al., 2014) and agar cultures (*A. niger*, *Candida albicans*) (Yamazaki et al., 2012), albeit LSSMG-induced phenotypic changes have been to date substantiated by a higher number of studies on fungal species e.g., *Pleurotus* sp., *Candida* sp., *Cladosporium* sp., *Ulocladium* sp., *Basipetospora* sp., etc (Miyazaki et al., 2010; Searles et al., 2011; Gomoiu et al., 2013). Interestingly, the sole morphological response to microgravity detected in the present study was the early switch to hyphae formation observed in *K. chersonesos* Wt (at day 5 instead of day 7, as in normal gravity; **Figure 2**). This is consistent with previous reports of increased filamentous growth under LSSMG in the opportunistic fungal pathogen *C. albicans* (Altenburg et al., 2008). The observation of early hyphal development in non-pathogenic species in response to LSSMG is rather suggestive of biofilm formation as a strategy for enhanced resistance to stress and for the forage for nutrients (Searles et al., 2011). Both LSSMG-cultured *K. chersonesos* Mut and their 1G controls showed extensive cell self-aggregation, i.e., cell clumping and occasional filamentation that are characteristics of black fungi (Nai et al., 2013). Significant variations were also not found in the total cell number, and cell size was not affected by microgravity in both strains (**Supplementary Tables 1**, **2**), unlike what has been documented by studies on a variety of bacteria (Huang et al., 2018) and yeasts (Crabbé et al., 2013).

The results of proteomics analysis revealed that exposure to LSSMG altered the proteome and secretome of both *K. chersonesos* Wt and Mut when compared to the 1G counterparts, having an impact on different pathways. Interestingly, the mutant response mainly involved protein downregulation, which might suggest a general slowing of the metabolic rate (**Figure 4B**). In contrast, more subtle rearrangements in the protein repertoire were observed in the wild type, especially in the secreted fraction, which possibly reflect a fine-tuning of the regulation of protein expression. Regardless, both strains showed increased abundance of proteins involved in carbohydrate metabolism, especially at the whole-cell proteome level (**Figure 5A**). Glycolysis/gluconeogenesis and pyruvate metabolism were found to be promoted in the wild type, whereas the glyoxylate shunt, ancillary cycle to TCA cycle and essential for growth on two-carbon compounds (Lorenz and Fink, 2001), was upregulated in the mutant. Similar alterations in carbohydrate metabolism were previously observed in melanotic filamentous fungi exposed to simulated Mars conditions (SMC) or to ISS-conditions (Romsdahl et al., 2018, 2019; Blachowicz et al., 2019a). Here, increased abundance was also observed for several starvation-induced glycoside hydrolases with roles in nutrient acquisition from biopolymers and in the recycling of cell wall components to support cell maintenance. Starvation response was thereby suggested as crucial adaptation to space conditions, especially oligotrophy. Conversely, one characteristic of both *K. chersonesos* Wt and Mut was decreased levels of glycoside hydrolases and cell wall–degradation enzymes such as the endo-1,3(4)-beta-glucanase (CF317_0009683-T1) and the extracellular cell wall glucanase Crf1 (CF317_009779-T1). The same was observed for carbohydrate-active enzymes like cutinases and for lipases at both the proteome and the secretome level, which suggests that, under space conditions, black fungi opt for strategies different than those adopted by filamentous fungi.

GO analysis and pathways prediction further revealed upregulation of proteins involved in the biosynthesis of unsaturated fatty acids (USFA) and in the metabolism of phospho- and glycerophospholipids. This may be explained by modifications in membrane lipid composition aimed at maintaining membranes stability against stress (Xia et al., 2019). More specifically, an LSSMG-dependent increase in membrane fluidity attributable to a higher USFA/SFA ratio was previously demonstrated by studies in plain lipid membranes, various microorganisms e.g., *Escherichia coli* and plants upon exposure to microgravity (Sieber et al., 2014; Kim and Rhee, 2016; Kordyum and Chapman, 2017). Potentially, this could affect the function of membrane-integrated proteins, thereby leading to altered transporter activity. Whether the activity of uptake transporters is altered in microgravity conditions is currently unknown (Eyal and Derendorf, 2019). However, similarly to what was observed in other fungi e.g., *C. albicans* (Crabbé et al., 2013) and bacteria e.g., *Bacillus subtilis* (Morrison et al., 2019) in space, *K. chersonesos* Wt and Mut ion-channels and integral membrane proteins were found to be upregulated. The increased abundance of specific transporters generally maximizes the uptake of nutrients (e.g., phosphate and nitrogen). As such, it possibly represents an adaptive response to temporary starvation caused either by zones of nutrient depletion developing around the colony or by the partial loss of contact of the cells with the culture medium, which may occur under microgravity conditions (Mazars et al., 2014; Senatore et al., 2018). Other types of transporters—i.e., ammonium transporters and permeases for transmembrane amino acid transport—could provide a link between nitrogen assimilation and proteins synthesis (Martzivanou et al., 2006). Indeed, upregulation of proteins involved in amino acid biosynthesis and metabolic processes was recorded under LSSMG conditions, but a higher number of these proteins was found to be decreased, especially in the mutant proteome.

Decreases in the levels of structural ribosomal proteins involved in cytoplasmic translation (i.e., L23a, L37a, and L34) and of the poly(A)polymerase, as well as the increase in RNA degrading ribonuclease T1, were additionally detected in microgravity. Together, the regulation of these proteins

may suggest that protein translation via ribosomal translational machinery is reduced upon exposure to LSSMG, a phenomenon that has been indicated as a widespread response to microgravity, space, and Mars-like conditions not only in fungi (Sheehan et al., 2007; Willaert, 2013; Feger et al., 2016; Kamal et al., 2018, 2019; Blachowicz et al., 2019a). Remarkably, in our study, a decrease in translation and ribosome biogenesis was only observed in the wild type whole-cell proteome and, to a minor extent, in the mutant secretome (**Figure 5B**). However, ribosomes are very dynamic organelles and additional tests will be needed to confirm that the 2-fold regulation of the above-mentioned proteins is actually indicative of reduced protein synthesis under microgravity.

In a similar fashion, proteins taking part in functional organization of cell organelles and cytoskeleton—for which a wide spectrum of microgravity-dependent changes has been reported (Zhang et al., 2015)—showed opposite regulation in the wild type and the mutant. The same was observed in the secretome for the DNA-repair Histone H2A and the scytalone dehydratase, the latter involved in the biosynthesis of DNH melanin from endogenous substrate (Eisenman and Casadevall, 2012), which were upregulated in the wild type and downregulated in the mutant. This is suggestive of increased measures for cell protection and is in line with the fact that melanin pigmentation and enhanced melanin synthesis is most often a feature of fungi living on space stations (Dadachova and Casadevall, 2008; De Middeleer et al., 2019). Further, versicolorin reductase (CF317_003229-T1), a protein mediating aflatoxins biosynthetic processes, was also decreased only in the mutant. Although production of aflatoxins in *K. chersonesos* has hitherto not been reported, this finding can be indicative of reduced production of allergenic or toxic metabolites (e.g., polyketides). Also in the mutant, the upregulation of hydrolytic enzymes such as the alpha, alpha-trehalase suggests a recourse to store carbohydrates as a carbon source. By contrast, production of compounds that increase the osmotolerancy (e.g., trehalose) represents a quite common protective measure implemented by microorganisms under microgravity and other types of stress (Willaert, 2013).

Opposite protein regulation was especially evident when examining the top 10 proteins differentially regulated in both strains at all experimental conditions, which included several ribosomal proteins (**Supplementary Figure 3**). These discrepancies suggest that strategies to cope with suboptimal conditions of growth may be strain-specific and that diverse rearrangements of proteome repertoire are possibly related to the presence or absence of melanin in the cell wall. Indeed, differences between wild type and mutant proteomic profiles were even observed at normal gravity condition. Proteins involved in translation, transcription and RNA metabolic process were significantly more upregulated in the mutant proteome than in the wild type (up to 9-fold, as in the case of ribosomal proteins L32, L36, S30 and L37), whereas the mutant secretome showed increased abundance of a number of hydrolases required for the breakdown of biopolymers and cell wall components (**Supplementary Figure 5**).

A number of proteins differentially expressed under simulated microgravity were involved in stress response, including glutamate dehydrogenase (GDH; CF317_004755-T1) and nitric oxide dioxygenase (NOD; CF317_006388-T1), over-represented in the proteome of both fungi, and chemical stress component proteins with proteolytic and phosphatase activity. While its role as a stress-responsive enzyme was speculated for GDH (i.e., ROS) (Skopelitis et al., 2006), NOD participates in mechanisms of detoxification of nitric oxide, a gas with multiple roles in cellular metabolism ranging from defense to signaling. The observation of peptidases, which were enriched in *K. chersonesos* Wt proteome under exposure to LSSMG, may instead suggest the removal of damaged proteins to enhance cellular fitness and maximize survival (Bonham-Carter et al., 2013; Zhang et al., 2015). In this respect, it should be noted that a role of altered gravity in the breakdown of protein structures has been previously reported (Trotter et al., 2015). However, a major part of detected stress response proteins involved in cell redox homeostasis, glutathione metabolism and recycling and osmotic stress defense—i.e., catalases, glutathione reductases, 3-phytases and s-formylglutathione hydrolases—appeared to be decreased under exposure to LSSMG. As previously suggested in plants (Zhang et al., 2015), downregulation of general stress response proteins under microgravity might indicate an impaired activation of the defense response components. However, decreased levels of common stress proteins and lack of a heat shock response (HSP) represent a key component of black fungi response to a variety of suboptimal conditions of growth and have been attributed to an energy-saving strategy that relies on a fine-tuning regulation of protein abundance (Tesei et al., 2012, 2015; Zakharova et al., 2013, 2014; Blasi et al., 2017).

One further interesting aspect was the high number of proteins traditionally recognized as cytoplasmic in *K. chersonesos* secretome, especially in the mutant (i.e., 6% of total secreted proteins with predicted signal peptide v/s 25% in the wild type; 20% of regulated secreted proteins had signal peptides in both strains). Such a discrepancy between wild type and mutant was also reported in a study of *K. chersonesos* secretome aimed at the screening for novel polyesterases (Tesei et al., 2020), hence it seems to suggest that a low number of extracellular proteins may be a peculiarity of the mutant secretome. The presence of cytoplasmic proteins in the culture supernatant may indicate that the mutant is more prone to cell lysis than the wild type (Miura and Ueda, 2018). However, given that proteins with no signal peptide (leaderless) can also be found in the secreted fraction as a result of non-classical protein secretion, e.g., via EVs-mediated pathways (Vallejo et al., 2012; Sun and Su, 2019), it may as well indicate co-isolation of vesicle proteins and classical secretory proteins during sample preparation procedures. These proteins (e.g., ribosomal proteins, proteins involved in carbohydrate metabolism, response to stress, signaling, cell division and transport, etc.), for which vesicular transport might be the only route of extracellular delivery, are generally predicted as non-secretory (Bleackley et al., 2019).

Together, the aspects of protein modulation observed in *K. chersonesos* Wt and Mut suggest that the basic

energy metabolism was upregulated in the Wt strain. Here, rearrangements of the protein repertoire resembled the classic response to microgravity, but with no evidence of significant activation of stress components or starvation response. The mutant mostly engaged in protein downregulation, without affecting their cell growth and survivability. This study therefore indicates the ability of black fungi to cope with microgravity conditions and suggests that cell wall melanization may ultimately influence the metabolic response to microgravity. Point mutation will be needed to confirm whether mutagenesis played a role in protein downregulation in *K. chersonesos* Mut. To further our understanding about the impact of microgravity on ribosome biogenesis and protein translation, future work shall involve proteomics workflows such as radio-labeled protein expression assays.

DATA AVAILABILITY STATEMENT

WGS data for *K. chersonesos* MA5789 are available in NCBI GenBank (GCA_002319055.1), under BioSample accession number SAMN07326825 and BioProject accession number PRJNA393270. The proteomics datasets generated during the current study are accessible through the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via PRIDE with the data set identifier PXD022898.

AUTHOR CONTRIBUTIONS

DT designed the study, drafted the manuscript, carried out the LSSMG exposure experiments, and performed data analysis and interpretation. AC and MK conducted protein sample processing, LC/MS analyses, and proteome data processing. JS annotated the genome of *K. chersonesos* MA5789 for proteome analysis. GBMM conducted FE-SEM analysis. KV designed the study and critically reviewed the manuscript. DT, KS, and KV contributed to funding acquisition. All authors contributed to the review, editing, and approval of the final version of the manuscript.

FUNDING

This research was funded by the Austrian Science Fund (FWF, Der Wissenschaftsfonds) project T872 Firnberg-Programm awarded to DT. A 2012 NASA Space Biology grant no. 19-12829-26 under Task Order NNN13D111T was awarded to KV which also supported GBMM. The Mass Spectrometry and Proteomics Core Facility at the City of Hope was supported in part by the National Cancer Institute of the National Institutes of Health under award no. P30CA033572. AC and MK were supported by Jet Propulsion Laboratory (JPL) subcontract no. 1611422.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the Austrian Science Fund (FWF, Der Wissenschaftsfonds) for sponsoring DT's postdoctoral position and for the funding. Part of the research described in this publication was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA. The authors thank Emily Klonicki, Jet Propulsion Laboratory (JPL) summer student, for contribution to the optimization of fungal cultivation under simulated microgravity and the members of the Planetary Protection group at Jet Propulsion Laboratory (JPL) for their technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.638708/full#supplementary-material

REFERENCES

Aguilera, A., and González-Toril, E. (2019). *Fungi in Extreme Environments: Ecological Role and Biotechnological Significance*. Cham: Springer, 21–38. doi: 10.1007/978-3-030-19030-9

Altenburg, S. D., Nielsen-Preiss, S. M., and Hyman, L. E. (2008). Increased filamentous growth of *Candida albicans* in simulated microgravity. *Genomics Proteomics Bioinforma* 6, 42–50. doi: 10.1016/S1672-0229(08)60019-4

Ametrano, C. G., Grewe, F., Crous, P. W., Goodwin, S. B., Liang, C., Selbmann, L., et al. (2019). Genome-scale data resolve ancestral rock- inhabiting lifestyle in Dothideomycetes (*Ascomycota*). *IMA Fungus* 10:19.

Belgaroui, N., Lacombe, B., Rouached, H., and Hanin, M. (2018). Phytase overexpression in *Arabidopsis* improves plant growth under osmotic stress and in combination with phosphate deficiency article. *Sci. Rep.* 8:1137. doi: 10.1038/s41598-018-19493-w

Ben-Shem, A., de Loubresse, N. G., Melnikov, S., Jenner, L., Yusupova, G., and Yusupov, M. (2011). The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* 334, 1524–1529.

Blachowicz, A., Chiang, A. J., Elsaesser, A., Kalkum, M., Ehrenfreund, P., Stajich, J. E., et al. (2019a). Proteomic and metabolomic characteristics of extremophilic fungi under simulated Mars conditions. *Front. Microbiol.* 10:1013. doi: 10.3389/fmicb.2019.01013

Blachowicz, A., Chiang, A. J., Romsdahl, J., Kalkum, M., Wang, C. C. C., and Venkateswaran, K. (2019b). Proteomic characterization of *Aspergillus fumigatus* isolated from air and surfaces of the International Space Station. *Fungal Genet. Biol.* 124, 39–46. doi: 10.1016/j.fgb.2019.01.001

Blasi, B., Tafer, H., Kustor, C., Poyntner, C., Lopandic, K., and Sterflinger, K. (2017). Genomic and transcriptomic analysis of the toluene degrading black yeast Cladophialophora immunda. *Sci. Rep.* 7:11436. doi: 10.1038/s41598-017-11807-8

Bleackley, M. R., Dawson, C. S., and Anderson, M. A. (2019). Fungal extracellular vesicles with a focus on proteomic analysis. *Proteomics* 19, 1–14. doi: 10.1002/pmic.201800232

Bommer, U. A., and Stahl, J. (2005). *Ribosomal Proteins in Eukaryotes*. New York, NY: John Wiley & Sons, Ltd. doi: 10.1038/npg.els.0003867

Bonham-Carter, O., Pedersen, J., Najjar, L., and Bastola, D. (2013). "Modeling the effects of microgravity on oxidation in mitochondria: a protein damage assessment across a diverse set of life forms," in *Proceedings of the IEEE 13th International Conference on Data Mining Workshops ICDMW 2013*, (Dallas, TX), 250–257. doi: 10.1109/ICDMW.2013.149

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). "UniProtKB/Swiss-Prot BT," in *Plant Bioinformatics: Methods and Protocols*, ed. D. Edwards (Totowa, NJ: Humana Press), 89–112. doi: 10.1007/978-1-59745-535-0_4

Braun, B., Pfirrmann, T., Menssen, R., Hofmann, K., Scheel, H., and Wolf, D. H. (2011). Gid9, a second RING finger protein contributes to the ubiquitin ligase activity of the Gid complex required for catabolite degradation. *FEBS Lett.* 585, 3856–3861. doi: 10.1016/j.febslet.2011.10.038

Cánovas, D., Marcos, J. F., Marcos, A. T., and Strauss, J. (2016). Nitric oxide in fungi: is there NO light at the end of the tunnel? *Curr. Genet.* 62, 513–518. doi: 10.1007/s00294-016-0574-6

Checinska, A., Probst, A. J., Vaishampayan, P., White, J. R., Kumar, D., Stepanov, V. G., et al. (2015). Microbiomes of the dust particles collected from the International Space Station and Spacecraft Assembly Facilities. *Microbiome* 3:50. doi: 10.1186/s40168-015-0116-3

Checinska-Sielaff, A., Urbaniak, C., Babu, G., Mohan, M., Stepanov, V. G., Tran, Q., et al. (2019). Characterization of the total and viable bacterial and fungal communities associated with the International Space Station surfaces. *Microbiome* 7:50.

Choudhry, P. (2016). High-Throughput method for automated colony and cell counting by digital image analysis based on edge detection. *PLoS One* 11:e0148469. doi: 10.1371/journal.pone.0148469

Couto, N., Wood, J., and Barber, J. (2016). Free radical biology and medicine the role of glutathione reductase and related enzymes on cellular redox homoeostasis network. *Free Radic. Biol. Med.* 95, 27–42. doi: 10.1016/j.freeradbiomed.2016.02.028

Crabbé, A., Nielsen-Preiss, S. M., Woolley, C. M., Barrila, J., Buchanan, K., McCracken, J., et al. (2013). Spaceflight enhances cell aggregation and random budding in *Candida albicans*. *PLoS One* 8:e80677. doi: 10.1371/journal.pone.0080677

Dadachova, E., and Casadevall, A. (2008). Ionizing radiation: how fungi cope, adapt, and exploit with the help of melanin. *Curr. Opin. Microbiol.* 11, 525–531. doi: 10.1016/j.mib.2008.09.013

De Middeleer, G., Leys, N., Sas, B., and De Saeger, S. (2019). Fungi and mycotoxins in space—a review. *Astrobiology* 19, 915–926. doi: 10.1089/ast.2018.1854

Desorption, M. L., Mass, I. T., Song, E. J., Kim, Y. S., Chung, J. Y., Kim, E., et al. (2000). Oxidative modification of nucleoside diphosphate kinase and its identification by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Biochemistry* 39, 10090–10097.

Eisenman, H. C., and Casadevall, A. (2012). Synthesis and assembly of fungal melanin. *Appl. Microbiol. Biotechnol.* 93, 931–940. doi: 10.1007/s00253-011-3777-2

El-gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Potter, S. C., Qureshi, M., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res* 47, 427–432. doi: 10.1093/nar/gky995

Eyal, S., and Derendorf, H. (2019). Medications in space: in search of a pharmacologist's guide to the galaxy. *Pharm. Res.* 36:148. doi: 10.1007/s11095-019-2679-3

Feger, B. J., Thompson, J. W., Dubois, L. G., Kommaddi, R. P., Foster, M. W., Mishra, R., et al. (2016). Microgravity induces proteomics changes involved in endoplasmic reticulum stress and mitochondrial protection. *Sci. Rep.* 6:34091. doi: 10.1038/srep34091

Gomoiu, I., Chatzitheodoridis, E., Vadrucci, S., and Walther, I. (2013). The effect of spaceflight on growth of ulocladium chartarum colonies on the International Space Station. *PLoS One* 8:e62130. doi: 10.1371/journal.pone.0062130

Gonçalves, V. N., Cantrell, C. L., Wedge, D. E., Ferreira, M. C., Soares, M. A., Jacob, M. R., et al. (2016). Fungi associated with rocks of the Atacama Desert: taxonomy, distribution, diversity, ecology and bioprospection for bioactive compounds. *Environ. Microbiol.* 18, 232–245. doi: 10.1111/1462-2920.13005

Gorbushina, A. A. (2007). Life on the rocks. *Environ. Microbiol.* 9, 1613–1631. doi: 10.1111/j.1462-2920.2007.01301.x

Gorbushina, A. A., Kotlova, E. R., and Sherstneva, O. A. (2008). Cellular responses of microcolonial rock fungi to long-term desiccation and subsequent rehydration. *Stud. Mycol.* 61, 91–97. doi: 10.3114/sim.2008.61.09

Gostinčar, C., Grube, M., and Gunde-Cimerman, N. (2011). Evolution of fungal pathogens in domestic environments? *Fungal Biol.* 115, 1008–1018. doi: 10.1016/j.funbio.2011.03.004

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi: 10.1093/nar/gkn176

Gunde-Cimerman, N., Sonjak, S., Zalar, P., Frisvad, J. C., Diderichsen, B., and Plemenitaš, A. (2003). Extremophilic fungi in arctic ice: a relationship between adaptation to low temperature and water activity. *Phys. Chem. Earth Parts A/B/C* 28, 1273–1278. doi: 10.1016/J.PCE.2003.08.056

Gunde-Cimerman, N., Zalar, P., de Hoog, G. S., and Plemenitaš, A. (2000). Hypersaline waters in saltern – natural ecological niches for halophilic black yeasts. *FEMS Microbiol. Ecol.* 32, 235–240.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

Haslam, R., Rust, S., Pallett, K., Cole, D., and Coleman, J. (2002). Cloning and characterisation of S-formylglutathione hydrolase from *Arabidopsis thaliana*: a pathway for formaldehyde detoxification. *Plant Physiol. Biochem.* 40, 281–288. doi: 10.1016/S0981-9428(02)01378-5

Herranz, R., Anken, R., Boonstra, J., Braun, M., Christianen, P. C. M., de Geest, M., et al. (2013). Ground-based facilities for simulation of microgravity: organism-specific recommendations for their use, and recommended terminology. *Astrobiology* 13, 1–17. doi: 10.1089/ast.2012.0876

Huang, B., Li, D. G., Huang, Y., and Liu, C. T. (2018). Effects of spaceflight and simulated microgravity on microbial growth and secondary metabolism. *Mil. Med. Res.* 5, 1–14. doi: 10.1186/s40779-018-0162-9

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., et al. (2019). EggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47, D309–D314. doi: 10.1093/nar/gky1085

Janin, J., and Deville-Bonne, D. (2002). "7 – Nucleoside-diphosphate kinase: structural and kinetic analysis of reaction pathway and phosphohistidine intermediate," in *Enzyme Kinetics and Mechanism*, ed. D. L. Purich (Cambridge, MA: Academic Press), 118–134. doi: 10.1016/S0076-6879(02)54009-X

Janssen, U., Fink, T., Lichter, P., and Stoffel, W. (1994). Human mitochondrial 3,2-trans-enoyl-CoA isomerase (DCI): gene structure and localization to chromosome 16p13.3. *Genomics* 23, 223–228. doi: 10.1006/geno.1994.1480

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., et al. (2009). STRING 8 — a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* 37, 412–416. doi: 10.1093/nar/gkn760

Kamal, K. Y., Herranz, R., Van Loon, J. J. W. A., and Medina, F. J. (2018). Simulated microgravity, Mars gravity, and 2g hypergravity affect cell cycle regulation, ribosome biogenesis, and epigenetics in *Arabidopsis* cell cultures. *Sci. Rep.* 8:6424. doi: 10.1038/s41598-018-24942-7

Kamal, K. Y., van Loon, J. J. W. A., Medina, F. J., and Herranz, R. (2019). Differential transcriptional profile through cell cycle progression in *Arabidopsis* cultures under simulated microgravity. *Genomics* 111, 1956–1965. doi: 10.1016/j.ygeno.2019.01.007

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Kim, H. W., and Rhee, M. S. (2016). Influence of low-shear modeled microgravity on heat resistance, membrane fatty acid composition, and heat stress-related gene. *Appl. Environ. Microbiol.* 82, 2893–2901. doi: 10.1128/AEM.00050-16. Editor

Knox, B. P., Blachowicz, A., Palmer, J. M., Romsdahl, J., Huttenlocher, A., Wang, C. C. C., et al. (2016). Characterization of *Aspergillus fumigatus* isolates from air and surfaces of the International Space Station. *Clin. Sci. Epidemiol.* 1, 1–15.

Kordyum, E. L., and Chapman, D. K. (2017). Plants and microgravity: patterns of microgravity effects at the cellular and molecular levels. *Cytol. Genet.* 51, 108–116. doi: 10.3103/S0095452717020049

Li, H., Ericsson, M., Rabasha, B., Budnik, B., Chan, S. H., Freinkman, E., et al. (2019). 6-phosphogluconate dehydrogenase links cytosolic carbohydrate metabolism to protein secretion via modulation of glutathione levels. *Cell Chem. Biol.* 26, 1306–1314.e5. doi: 10.1016/j.chembiol.2019.05.006

Lincoln, S. P., Fermor, T. R., and Wood, D. A. (1997). Production and detection of muramidase and acetylglucosaminidase from *Agaricus bisporus*. *Lett. Appl. Microbiol.* 25, 24–29. doi: 10.1046/j.1472-765X.1997.00163.x

Lorenz, M. C., and Fink, G. R. (2001). The glyoxylate cycle is required for fungal virulence. *Nature* 412, 83–86. doi: 10.1038/35083594

Martzivanou, M., Babbick, M., Cogoli-Greuter, M., and Hampp, R. (2006). Microgravity-related changes in gene expression after short-term exposure of *Arabidopsis thaliana* cell cultures. *Protoplasma* 229, 155–162. doi: 10.1007/s00709-006-0203-1

Mazars, C., Briére, C., Grat, S., Pichereaux, C., Rossignol, M., Pereda-Loth, V., et al. (2014). Microgravity induces changes in microsome-associated proteins of *Arabidopsis* seedlings grown on board the international space station. *PLoS One* 9:e91814. doi: 10.1371/journal.pone.0091814

Meti, R. S., Ambarish, S., and Khajure, P. V. (2011). Enzymes of ammonia assimilation in fungi: an overview. *Sci. Technol.* 2, 28–38.

Miura, N., and Ueda, M. (2018). Evaluation of unconventional protein secretion by *Saccharomyces cerevisiae* and other Fungi. *Cells* 7:128. doi: 10.3390/cells7090128

Miyazaki, Y., Sunagawa, M., Higashibata, A., Ishioka, N., Babasaki, K., and Yamazaki, T. (2010). Differentially expressed genes under simulated microgravity in fruiting bodies of the fungus *Pleurotus ostreatus*. *FEMS Microbiol. Lett.* 307, 72–79. doi: 10.1111/j.1574-6968.2010.01966.x

Moore, J. D., Yazgan, O., Ataian, Y., and Krebs, J. E. (2007). Diverse roles for histone H2A modifications in DNA damage response pathways in yeast. *Genetics* 25, 15–25. doi: 10.1534/genetics.106.063792

Morrison, M. D., Fajardo-Cavazos, P., and Nicholson, W. L. (2019). Comparison of *Bacillus subtilis* transcriptome profiles from two separate missions to the International Space Station. *NPJ Microgravity* 5:1. doi: 10.1038/s41526-018-0061-0

Mouyna, I., Dellière, S., Beauvais, A., Gravelat, F., Carrion, S. D. J., Pearlman, E., et al. (2020). What are the functions of chitin deacetylases in *Aspergillus fumigatus*? *Front. Cell. Infect. Microbiol.* 10:28. doi: 10.3389/fcimb.2020.00028

Nai, C., Wong, H. Y., Pannenbecker, A., Broughton, W. J., Benoit, I., de Vries, R. P., et al. (2013). Nutritional physiology of a rock-inhabiting, model microcolonial fungus from an ancestral lineage of the Chaetothyriales (*Ascomycetes*). *Fungal Genet. Biol.* 56, 54–66. doi: 10.1016/j.fgb.2013.04.001

Nakamura, H., Narihiro, T., Tsuruoka, N., Mochimaru, H., Matsumoto, R., Tanabe, Y., et al. (2011). Evaluation of the aflatoxin biosynthetic genes for identification of the *Aspergillus* section Flavi. *Microbes Environ.* 26, 367–369. doi: 10.1264/jsme2.ME11201

Neto, Y. A. A. H., Garzon, N. G. D. R., Pedezzi, R., and Cabral, H. (2018). Specificity of peptidases secreted by filamentous fungi. *Bioengineered* 9, 30–37. doi: 10.1080/21655979.2017.1373531

Onofri, S., Barreca, D., Selbmann, L., Isola, D., Rabbow, E., Horneck, G., et al. (2008). Resistance of Antarctic black fungi and cryptoendolithic communities to simulated space and Martian conditions. *Stud. Mycol.* 61, 99–109. doi: 10.3114/sim.2008.61.10

Onofri, S., de la Torre, R., de Vera, J.-P., Ott, S., Zucconi, L., Selbmann, L., et al. (2012). Survival of rock-colonizing organisms after 1.5 years in outer space. *Astrobiology* 12, 508–516. doi: 10.1089/ast.2011.0736

Onofri, S., Selbmann, L., de Hoog, G. S., Grube, M., Barreca, D., Ruisi, S., et al. (2007). Evolution and adaptation of fungi at boundaries of life. *Adv. Space Res.* 40, 1657–1664. doi: 10.1016/j.asr.2007.06.004

Onofri, S., Selbmann, L., Pacelli, C., Zucconi, L., Rabbow, E., and De Vera, J. P. (2019). Survival, DNA, and ultrastructural integrity of a cryptoendolithic Antarctic fungus in Mars and Lunar Rock analogs exposed outside the International Space Station. *Astrobiology* 19, 170–182. doi: 10.1089/ast.2017.1728

Pacelli, C., Bryan, R. A., Onofri, S., Selbmann, L., Zucconi, L., Shuryak, I., et al. (2018). Survival and redox activity of Friedmanniomyces endolithicus, an Antarctic endemic black meristematic fungus, after gamma rays exposure. *Fungal Biol.* 122, 1222–1227. doi: 10.1016/j.funbio.2018.10.002

Pacelli, C., Selbmann, L., Zucconi, L., De Vera, J.-P., Rabbow, E., Horneck, G., et al. (2016). BIOMEX experiment: ultrastructural alterations, molecular damage and survival of the fungus cryomyces antarcticus after the experiment verification tests. *Orig. Life Evol. Biosph.* 47, 187–202. doi: 10.1007/s11084-016-9485-2

Pacelli, C., Selbmann, L., Zucconi, L., Raguse, M., Moeller, R., Shuryak, I., et al. (2017). Survival, DNA integrity, and ultrastructural damage in antarctic cryptoendolithic eukaryotic microorganisms exposed to ionizing radiation. *Astrobiology* 17, 126–135. doi: 10.1089/ast.2015.1456

Pao, S. S., and Paulsen, I. T. S. M. (1998). Major facilitator superfamily. *Microbiol. Mol. Biol. Rev.* 62, 1–34.

Passoth, V. (2017). "Lipids of yeasts and filamentous fungi and their importance for biotechnology BT," in *Biotechnology of Yeasts and Filamentous Fungi*, ed. A. A. Sibirny (Cham: Springer International Publishing), 149–204. doi: 10.1007/978-3-319-58829-2_6

Rabbow, E., Rettberg, P., Barczyk, S., Bohmeier, M., Parpart, A., Panitz, C., et al. (2012). EXPOSE-E: an ESA astrobiology mission 1.5 years in space. *Astrobiology* 12, 374–386. doi: 10.1089/ast.2011.0760

Rabbow, E., Rettberg, P., Parpart, A., Panitz, C., Schulte, W., Molter, F., et al. (2017). EXPOSE-R2: the astrobiological ESA mission on board of the international space station. *Front. Microbiol.* 8:1533. doi: 10.3389/fmicb.2017.01533

Raben, D. M., and Wolfgang, M. J. (2019). *Phosphofructokinase-2/Fructose Bisphosphatase-2☆*. Amsterdam: Elsevier. doi: 10.1016/B978-0-12-801238-3.11340-6

Regente, M., Pinedo, M., and Elizalde, M. (2012). Apoplastic exosome-like vesicles: a new way of protein secretion in plants? *Plant Signal Behav.* 7, 544–546.

Romsdahl, J., Blachowicz, A., Chiang, A. J., Chiang, Y. M., Masonjones, S., Yaegashi, J., et al. (2019). International Space Station conditions alter genomics, proteomics, and metabolomics in *Aspergillus nidulans*. *Appl. Microbiol. Biotechnol.* 103, 1363–1377. doi: 10.1007/s00253-018-9525-0

Romsdahl, J., Blachowicz, A., Chiang, A. J., Singh, N., Stajich, J. E., Kalkum, M., et al. (2018). Characterization of *Aspergillus niger* isolated from the International Space Station. *mSystems* 3, 1–13. doi: 10.1128/msystems.00112-18

Rosenzweig, J. A., Abogunde, O., Thomas, K., Lawal, A., Nguyen, Y. U., Sodipe, A., et al. (2010). Spaceflight and modeled microgravity effects on microbial growth and virulence. *Appl. Microbiol. Biotechnol.* 85, 885–891. doi: 10.1007/s00253-009-2237-8

Sathishkumar, Y., Krishnaraj, C., Rajagopal, K., Sen, D., and Lee, Y. S. (2016). High throughput de novo RNA sequencing elucidates novel responses in *Penicillium chrysogenum* under microgravity. *Bioprocess Biosyst. Eng.* 39, 223–231. doi: 10.1007/s00449-015-1506-4

Sathishkumar, Y., Velmurugan, N., Lee, H. M., Rajagopal, K., Im, C. K., and Lee, Y. S. (2014). Effect of low shear modeled microgravity on phenotypic and central chitin metabolism in the filamentous fungi *Aspergillus niger* and *Penicillium chrysogenum*. *Antonie van Leeuwenhoek* 106, 197–209. doi: 10.1007/s10482-014-0181-9

Savojardo, C., Martelli, P. L., Fariselli, P., Profiti, G., and Casadio, R. (2018). BUSCA: an integrative web server to predict subcellular localization of proteins. *Nucleic Acids Res.* 46, W459–W466. doi: 10.1093/nar/gky320

Scalzi, G., Selbmann, L., Zucconi, L., Rabbow, E., Horneck, G., Albertano, P., et al. (2012). LIFE experiment: isolation of cryptoendolithic organisms from Antarctic colonized sandstone exposed to space and simulated Mars conditions on the International Space Station. *Orig. Life Evol. Biosph.* 42, 253–262. doi: 10.1007/s11084-012-9282-5

Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012). NIH image to imageJ: 25 years of image analysis. *Nat. Methods* 9, 671–675. doi: 10.1038/nmeth.2089

Searles, S. C., Woolley, C. M., Petersen, R. A., Hyman, L. E., and Nielsen-Preiss, S. M. (2011). Modeled microgravity increases filamentation, biofilm formation, phenotypic switching, and antimicrobial resistance in *Candida albicans*. *Astrobiology* 11, 825–836. doi: 10.1089/ast.2011.0664

Selbmann, L., De Hoog, G. S., Mazzaglia, A., Friedmann, E. I., and Onofri, S. (2005). Fungi at the edge of life: cryptoendolithic black fungi from Antarctic desert. *Stud. Mycol.* 51, 1–32.

Selbmann, L., de Hoog, G. S., Zucconi, L., Isola, D., Ruisi, S., Gerrits van den Ende, A. H. G., et al. (2008). Drought meets acid: three new genera in a dothidealean clade of extremotolerant fungi. *Stud. Mycol.* 61, 1–20. doi: 10.3114/sim.2008.61.01

Selbmann, L., Zucconi, L., Isola, D., and Onofri, S. (2014). Rock black fungi: excellence in the extremes, from the Antarctic to space. *Curr. Genet.* 61, 335–345. doi: 10.1007/s00294-014-0457-7

Senatore, G., Mastroleo, F., Leys, N., and Mauriello, G. (2018). Effect of microgravity & space radiation on microbes. *Future Microbiol.* 13, 831–847. doi: 10.2217/fmb-2017-0251

Serra-Cardona, A., Canadell, D., and Ariño, J. (2015). Coordinate responses to alkaline pH stress in budding yeast. *Microb. Cell* 2, 182–196. doi: 10.15698/mic2015.06.205

Sheehan, K. B., McInnerney, K., Purevdorj-Gage, B., Altenburg, S. D., and Hyman, L. E. (2007). Yeast genomic expression patterns in response to low-shear modeled microgravity. *BMC Genomics* 8:3. doi: 10.1186/1471-2164-8-3

Shtarkman, Y. M., Koçer, Z. A., Edgar, R., Veerapaneni, R. S., D'Elia, T., Morris, P. F., et al. (2013). Subglacial Lake Vostok (Antarctica) accretion ice contains a

diverse set of sequences from aquatic, marine and sediment-inhabiting bacteria and eukarya. *PLoS One* 8:e67221. doi: 10.1371/journal.pone.0067221

Sieber, M., Hanke, W., and Kohn, F. P. M. (2014). Modification of membrane fluidity by gravity. *Open J. Biophys.* 04, 105–111. doi: 10.4236/ojbiphy.2014.44012

Skoneczna, A., Krol, K., and Skoneczny, M. (2018). "How do yeast and other fungi recognize and respond to genome perturbations?," in *Stress Response Mechanisms in Fungi*, ed. M. Skoneczny (Cham: Springer), 87–130. doi: 10.1007/978-3-030-00683-9_3

Skopelitis, D. S., Paranychianakis, N. V., Paschalidis, K. A., Pliakonis, E. D., Delis, I. D., Yakoumakis, D. I., et al. (2006). Abiotic stress generates ROS that signal expression of anionic glutamate dehydrogenases to form glutamate for proline synthesis in tobacco and grapevine. *Plant Cell* 18, 2767–2781. doi: 10.1105/tpc.105.038323

Spreghini, E., Davis, D. A., Subaran, R., Kim, M., and Mitchell, A. P. (2003). Roles of *Candida albicans* Dfg5p and Dcw1p cell surface proteins in growth and hypha formation. *Eukaryot. Cell* 2, 746–755. doi: 10.1128/EC.2.4.746

Sterflinger, K. (2006). "Black yeasts and meristematic fungi: ecology, diversity and identification," in *Biodiversity and Ecophysiology of Yeasts. The Yeast Handbook*, eds G. Péter and C. Rosa (Heidelberg: Springer). doi: 10.1007/3-540-30985-3_20

Sterflinger, K., de Hoog, G. S., and Haase, G. (1999). Phylogeny and ecology of meristematic ascomycetes. *Stud. Mycol.* 43, 5–22.

Sterflinger, K., Tesei, D., and Zakharova, K. (2012). Fungi in hot and cold deserts with particular reference to microcolonial fungi. *Fungal Ecol.* 5, 453–462. doi: 10.1016/j.funeco.2011.12.007

Sun, X., and Su, X. (2019). Harnessing the knowledge of protein secretion for enhanced protein production in filamentous fungi. *World J. Microbiol. Biotechnol.* 35, 1–10. doi: 10.1007/s11274-019-2630-0

Taylor, P. W. (2015). Impact of space flight on bacterial virulence and antibiotic susceptibility. *Infect. Drug Resist.* 8, 249–262. doi: 10.2147/IDR.S67275

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* 18, 1979–1990. doi: 10.1101/gr.081612.108

Tesei, D., Marzban, G., Marchetti-Deschmann, M., Tafer, H., Arcalis, E., and Sterflinger, K. (2015). Proteome of tolerance fine-tuning in the human pathogen black yeast *Exophiala dermatitidis*. *J. Proteomics* 128, 39–57. doi: 10.1016/j.jprot.2015.07.007

Tesei, D., Marzban, G., Zakharova, K., Isola, D., Selbmann, L., and Sterflinger, K. (2012). Alteration of protein patterns in black rock inhabiting fungi as a response to different temperatures. *Fungal Biol.* 116, 932–940. doi: 10.1016/j.funbio.2012.06.004

Tesei, D., Quartinello, F., Guebitz, G. M., Ribitsch, D., Nöbauer, K., Razzazi-Fazeli, E., et al. (2020). Shotgun proteomics reveals putative polyesterases in the secretome of the rock-inhabiting fungus Knufia chersonesos. *Sci. Rep.* 10:9770. doi: 10.1038/s41598-020-66256-7

Tesei, D., Tafer, H., Poyntner, C., Piñar, G., Lopandic, K., and Sterflinger, K. (2017). Draft genome sequences of the black rock fungus Knufia petricola and its spontaneous nonmelanized mutant. *Genome Announc.* 5, 1–2. doi: 10.1128/genomeA.01242-17

Trotter, B., Otte, K. A., Schoppmann, K., Hemmersbach, R., Fröhlich, T., Arnold, G. J., et al. (2015). The influence of simulated microgravity on the proteome of *Daphnia magna*. *NPJ Microgravity* 1:15016. doi: 10.1038/npjmgrav.2015.16

Urbaniak, C., van Dam, P., Zaborin, A., Zaborina, O., Gilbert, J. A., Torok, T., et al. (2019). Genomic characterization and virulence potential of two *Fusarium oxysporum* isolates cultured from the International Space Station. *mSystems* 4:e00345-18. doi: 10.1128/msystems.00345-18 e00345-18

Vallejo, M. C., Nakayasu, E. S., Matsuo, A. L., Sobreira, T. J. P., Longo, L. V. G., Ganiko, L., et al. (2012). Vesicle and vesicle-free extracellular proteome of *Paracoccidioides brasiliensis*: comparative analysis with other pathogenic fungi. *J. Proteome Res.* 11, 1676–1685.

Voigt, O., Knabe, N., Nitsche, S., Erdmann, E. A., Schumacher, J., and Gorbushina, A. A. (2020). An advanced genetic toolkit for exploring the biology of the rock – inhabiting black fungus Knufia petricola. *Sci. Rep.* 10:22021. doi: 10.1038/s41598-020-79120-5

Willaert, R. G. (2013). The growth behavior of the model eukaryotic yeast *Saccharomyces cerevisiae* in microgravity. *Curr. Biotechnol.* 2, 226–234.

Xia, Z., Zhou, X., Li, J., Li, L., Ma, Y., Wu, Y., et al. (2019). Multiple-omics techniques reveal the role of glycerophospholipid metabolic pathway in the response of *Saccharomyces cerevisiae* against hypoxic stress. *Front. Microbiol.* 10:1398. doi: 10.3389/fmicb.2019.01398

Yamaguchi, N., Roberts, M., Castro, S., Oubre, C., Makimura, K., Leys, N., et al. (2014). Microbial monitoring of crewed habitats in space-current status and future perspectives. *Microbes Environ.* 29, 250–260. doi: 10.1264/jsme2.ME14031

Yamazaki, T., Yoshimoto, M., Nishiyama, Y., Okubo, Y., and Makimura, K. (2012). Phenotypic characterization of *Aspergillus niger* and *Candida albicans* grown under simulated microgravity using a three-dimensional clinostat. *Microbiol. Immunol.* 56, 441–446. doi: 10.1111/j.1348-0421.2012.00471.x

Yurimoto, H., Lee, B., Yano, T., Sakai, Y., and Kato, N. (2003). Physiological role of S-formylglutathione hydrolase in C1 metabolism of the methylotrophic yeast *Candida boidinii*. *Microbiology* 149, 1971–1979. doi: 10.1099/mic.0.26320-0

Zakharova, K., Marzban, G., de Vera, J.-P., Lorek, A., and Sterflinger, K. (2014). Protein patterns of black fungi under simulated Mars-like conditions. *Sci. Rep.* 4:5114. doi: 10.1038/srep05114

Zakharova, K., Tesei, D., Marzban, G., Dijksterhuis, J., Wyatt, T., and Sterflinger, K. (2013). Microcolonial fungi on rocks: a life in constant drought? *Mycopathologia* 175, 537–547.

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simao, 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

Zhang, Y., Wang, L., Xie, J., and Zheng, H. (2015). Differential protein expression profiling of *Arabidopsis thaliana* callus under microgravity on board the Chinese SZ-8 spacecraft. *Planta* 241, 475–488. doi: 10.1007/s00425-014-2196-x

Zhang, Z., Mao, C., Shi, Z., and Kou, X. (2017). The amino acid metabolic and carbohydrate metabolic pathway play important roles during salt-stress response in tomato. *Front. Plant Sci.* 8:1231. doi: 10.3389/fpls.2017.01231

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

Zhao, Q., Zhang, H., Wang, T., Chen, S., and Dai, S. (2013). Proteomics-based investigation of salt-responsive mechanisms in plant roots. *J. Proteomics* 82, 230–253.

Zhou, L., Hang, J., Zhou, Y., Wan, R., Lu, G., Yin, P., et al. (2014). Crystal structures of the Lsm complex bound to the 3′ end sequence of U6 small nuclear RNA. *Nature* 506, 116–120. doi: 10.1038/nature12803

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: 26 March 2021
doi: 10.3389/fgene.2021.633073

In silico Design for Systems-Based Metabolic Engineering for the Bioconversion of Valuable Compounds From Industrial By-Products

Albert Enrique Tafur Rangel[1,2], Wendy Ríos[1], Daisy Mejía[1], Carmen Ojeda[1], Ross Carlson[3], Jorge Mario Gómez Ramírez[1] and Andrés Fernando González Barrios[1]**

[1] Grupo de Diseño de Productos y Procesos, Department of Chemical and Food Engineering, Universidad de los Andes, Bogotá, Colombia, [2] Grupo de Investigación CINBIOS, Department of Microbiology, Universidad Popular del Cesar, Valledupar, Colombia, [3] Center for Biofilm Engineering, Montana State University, Bozeman, MT, United States

OPEN ACCESS

Edited by:
Gorji Marzban,
University of Natural Resources and Life Sciences Vienna, Austria

Reviewed by:
Bashir Sajo Mienda,
University of Tübingen, Germany
Long Liu,
Jiangnan University, China

***Correspondence:**
Albert Enrique Tafur Rangel
ae.tafur@uniandes.edu.co
Andrés Fernando González Barrios
andgonza@uniandes.edu.co

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

Received: 24 November 2020
Accepted: 23 February 2021
Published: 26 March 2021

Citation:
Tafur Rangel AE, Ríos W, Mejía D, Ojeda C, Carlson R, Gómez Ramírez JM and González Barrios AF (2021) In silico Design for Systems-Based Metabolic Engineering for the Bioconversion of Valuable Compounds From Industrial By-Products. Front. Genet. 12:633073. doi: 10.3389/fgene.2021.633073

Selecting appropriate metabolic engineering targets to build efficient cell factories maximizing the bioconversion of industrial by-products to valuable compounds taking into account time restrictions is a significant challenge in industrial biotechnology. Microbial metabolism engineering following a rational design has been widely studied. However, it is a cost-, time-, and laborious-intensive process because of the cell network complexity; thus, it is important to use tools that allow predicting gene deletions. An *in silico* experiment was performed to model and understand the metabolic engineering effects on the cell factory considering a second complexity level by transcriptomics data integration. In this study, a systems-based metabolic engineering target prediction was used to increase glycerol bioconversion to succinic acid based on *Escherichia coli*. Transcriptomics analysis suggests insights on how to increase cell glycerol utilization to further design efficient cell factories. Three *E. coli* models were used: a core model, a second model based on the integration of transcriptomics data obtained from growth in an optimized culture media, and a third one obtained after integration of transcriptomics data from adaptive laboratory evolution (ALE) experiments. A total of 2,402 strains were obtained with fumarase and pyruvate dehydrogenase being frequently predicted for all the models, suggesting these reactions as essential to increase succinic acid production. Finally, based on using flux balance analysis (FBA) results for all the mutants predicted, a machine learning method was developed to predict new mutants as well as to propose optimal metabolic engineering targets and mutants based on the measurement of the importance of each knockout's (feature's) contribution. Glycerol has become an interesting carbon source for industrial processes due to biodiesel business growth since it has shown promising results in terms of biomass/substrate yields. The combination of transcriptome, systems metabolic modeling, and machine

learning analyses revealed the versatility of computational models to predict key metabolic engineering targets in a less cost-, time-, and laborious-intensive process. These data provide a platform to improve the prediction of metabolic engineering targets to design efficient cell factories. Our results may also work as a guide and platform for the selection/engineering of microorganisms for the production of interesting chemical compounds.

Keywords: systems metabolic engineering, transcriptomics, machine learning, adaptive laboratory evolution, metabolic modeling resources/frameworks

INTRODUCTION

Shifting from petrochemical sources to renewable, abundant, and inexpensive feedstocks to obtain valuable chemicals has become a promising goal for the chemical industry (Vlysidis et al., 2011). The biodiesel industry has increased in the last years by using renewable raw materials, but it generates large amounts of glycerol, which has become a burden. The bioconversion of glycerol is a potential route to increasing the use of bio-based succinic acid, a critical building block chemical with an attractive market. The availability of three pathways for succinic acid production (**Figure 1**; Chen et al., 2013a), the adaptability to different environments, and the accessibility of metabolic engineering and omics tools make *Escherichia coli* an attractive cell factory. However, some challenges, such as low growth rate and yield, the use of a rich medium, the generation of by-products, and various anaerobic requirements, need to be overcome for bio-based succinic acid production, considering cost-effective issues, as compared with the petroleum-based approach.

The main goal of using microbial cell factories is to design cheap and high-yield biotechnology-based conversion processes. A significant problem to be solved is how to enhance cell growth while using its capabilities to obtain a high-yield target chemical product. A classical approach for that is adaptive laboratory evolution (ALE), which is based on the selection of microorganisms with superior production capability after random mutagenesis screening. Another approach to strain improvement is metabolic engineering, which uses genetic manipulation to optimize the production of desired compounds. Metabolic engineering selects targets that increase productivity based on the rationality of trial-and-error development cycles and an understanding of the routes playing a significant role in the synthesis. Strain design with this method has been extensively applied to use and/or produce interesting compounds (Kern A. et al., 2007; Chen et al., 2013a,b; Förster and Gescher, 2014; Woo and Park, 2014), including bio-based organic acids by substrate transport enhancement, gene overexpression, and deletion (Shams Yazdani and Gonzalez, 2008; Zhang B. et al., 2012; Buschke et al., 2013; Förster and Gescher, 2014; Yin et al., 2015; Zhu and Jackson, 2015). However, making the strain industrially competitive requires much time, effort, and high cost (Rangel et al., 2020).

When DNA was discovered in the last century, a new approach called metabolic network modeling for the study of cellular metabolism was developed (O'Brien et al., 2015). It allows to determine how several pathways in a cell can interact, as well as to elucidate basic microbial processes (Haggart et al., 2011). The first genome-scale metabolic network was described in 1999, and in 2002, the use of metabolic modeling to analyze recombinant pathways was reported (Carlson et al., 2002). Several models have been developed ever since with significant accuracy and useful predictions (Portela et al., 2013) that can be used to guide experimental studies (Pharkya and Maranas, 2006; O'Brien et al., 2015). COnstraints-Based Reconstruction and Analysis (COBRA) methods make it possible to predict, given a cellular objective function, attractive targets to increase or maximize biochemical yields, and to determine perturbations after genetic manipulations of the cell (Kim, 2012; Ruckerbauer et al., 2015). OptKnock, OptStrain, OptForce, and OptReg are some COBRA methods developed to predict metabolic engineering targets for cell optimization by using gene–protein reaction (GPR) relationship (Burgard et al., 2003; Pharkya et al., 2003; Pharkya et al., 2004; Pharkya and Maranas, 2006; Ranganathan et al., 2010).

OptKnock applies a flux balance analysis (FBA) approach for simulating genome-scale metabolic models (GEMs). It assumes that each organism's metabolic network has been tuned through evolution for some objective function, be it a maximal growth rate or energy efficiency (e.g., minimal ATP utilization). While this assumption may be valid for wild-type (WT) organisms that have evolved over many hundreds or thousands of generations, it may be less appropriate for engineered mutants (KO) because they have been engineered in a controlled environment and unexposed to the same evolutionary forces. Hypothesizing that mutant organisms are unable to immediately adapt their metabolic network to achieve the WT objective function, computational tools such as minimization of metabolic adjustment (MOMA) were developed (Segre et al., 2002). This approach is mathematically formalized as a quadratic programming (QP) problem, finding a suboptimal flux profile that is a minimal Euclidean distance from the WT (WT-FBA) and the genetically perturbed (KO-FBA) organisms. FBA combined with MOMA evaluation after OptKnock prediction could provide a more accurate prediction of the immediate metabolic response to KO than FBA does on its own. However, a large list of knockout combinations could be obtained when computational tools are used, and select which test in a lab can be laborious.

Several approaches to optimize cell factories have been developed, but conventional and computational approaches

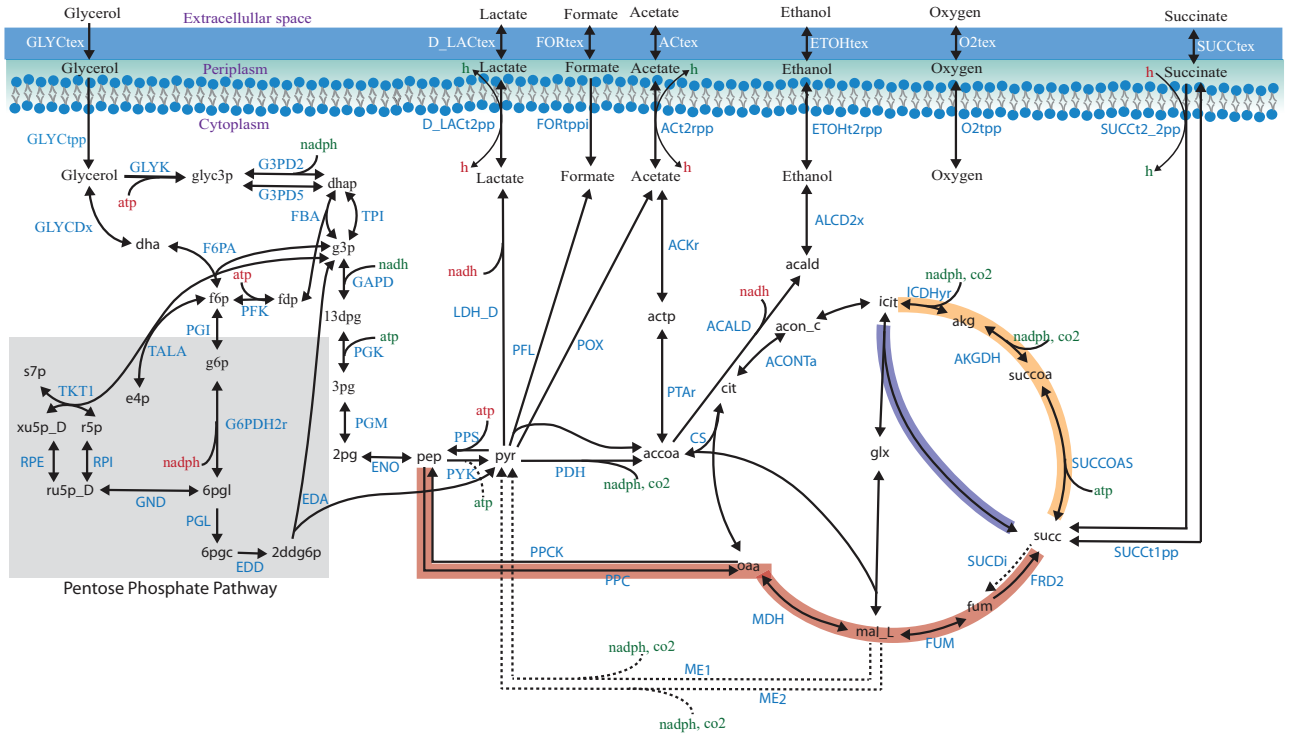

FIGURE 1 | Succinic acid pathways from glycerol in *Escherichia coli*. The three pathways for succinic acid production are indicated by the thick red (the PEP–pyruvate–oxaloacetate node—the reductive TCA branch), yellow (the oxidative TCA branch), and blue (the glyoxylate shunt) arrows. Relevant biochemical reactions are represented based on the ID BIGG database names (King et al., 2016).

have not always been successful due to unexpected changes in the cell where an intracellular complex interconnected network of genes, proteins, and reactions exists. Systems metabolic engineering has emerged as an approach that integrates metabolic engineering and combined metabolic and "omics" network models. This approach could be beneficial for genome-scale modeling because it reduces the solution space and generates accurate predictions (Nordlander et al., 2008; Feist et al., 2010; Blazier and Papin, 2012; Machado and Herrgård, 2014; Rangel et al., 2020). Mainly considering that under certain environmental conditions, there are a limited number of reactions that are active according to transcriptional responses and other regulation phenomena to provide beneficial improvements for the cell bioconversion process (Fong and Marciniak, 2003; Fong et al., 2005; Lee and Palsson, 2010; Conrad et al., 2011; Wang et al., 2011; Zhang J. et al., 2012; Bao et al., 2014).

In this study, systems metabolic engineering for overproduction of succinic acid from glycerol in *E. coli* ATCC 8739 was used through integration of transcriptomics data to metabolic models and classification tree analysis using the random forest to classify gene targets predicted by OptKnock. Our strategy took advantage of transcriptomics data obtained from an evolved *E. coli* in glycerol and an optimized culture media. These data were subsequently integrated into a metabolic network model to predict targets using OptKnock. Predicted combinations were then evaluated using FBA, flux variability analysis, and MOMA to determine the effects of gene reaction knockout in the cell. Finally, predicted target reactions were evaluated using random forest to determine the importance of each target using succinic acid production, growth rate, and Euclidian distance between the WT strain and each mutant as response variables.

MATERIALS AND METHODS

Strains and Culture Conditions

E. coli ATCC 8739 was used in this study. It was obtained commercially from the American Type Culture Collection (ATCC). A glycerol-based medium containing the following components (per liter) was used as the reference culture condition: 5 g of yeast extract, 2.5 g of NaCl, 5 ml of trace metal solution [0.55 g/L $CaCl_2$, 0.10 g/L $MnCl_2$ $4H_2O$, 0.17 g/L $ZnCl_2$, 0.043 g/L $CuCl_2$ $2H_2O$, 0.06 g/L $CoCl_2$ $6H_2O$, 0.06 g/L Na_2MoO_4 $2H_2O$, 0.06 g/L $Fe(NH_4)_2(SO4)_2$ $6H_2O$, 0.20 g/L $FeCl_3$ $6H_2O$], 5 ml of $MgSO_4$ (1 M), and 30 g of glycerol. A 50 ml culture was carried out in a 250 ml baffled-conical Erlenmeyer flask and cultivated aerobically at 37°C and 200 rpm.

Two conditions were evaluated in this study: an adapted *E. coli* on high glycerol concentrations (30, 40, 50, 60 g/L) and an optimized culture condition. For the first condition, four *E. coli* cultures were continuously subcultured each for 72 h in Luria–Bertani (LB) medium (5 g/L of NaCl, 10 g/L of tryptone, and

5 g/L of yeast extract, supplemented with 30, 40, 50, or 60 g/L). After every three subcultured rounds (216 h), the concentration of tryptone was decreased from 1 until reaching 0 g/L. Then, 10 subcultured rounds each for 72 h were carried out. During the complete experiment, a 50 ml culture was carried out in a 250 ml non-baffled-conical Erlenmeyer flask and cultivated aerobically at 37°C and 200 rpm. For each subcultured round, an OD ∼0.33 600 nm was considered as inoculum starting point. At the end of each tryptone decreasing, 1 ml of culture was kept at –80°C and used for further evaluation of growth and glycerol uptake. For the optimized culture condition, the glycerol-based medium was supplemented with 1 g of NH_4Cl, 6 g of Na_2HPO_4, and 3 g of K_2HPO_4 at the same conditions as the reference culture.

Differential Expression Analysis

RNA-Seq was carried out in triplicate for all conditions. For the adapted strain, the culture conditions for RNA-Seq were the same as those for the optimized culture medium condition. To harvest cells for total RNA purification, the culture sample was first treated with RNAprotect Bacteria Reagent (Cat No./ID: 76506), and enzymatic lysis and proteinase K digestion of the bacteria were carried out following the manufacturer's protocol. Then, the Qiagen RNeasy Mini kit (Cat No./ID: 74104), following the manufacturer's protocol, was used to obtain the total RNA for further analysis. Each sample was treated with DNase following the protocol in order to remove the DNA. The samples were sent to commercial RNA-Seq services for further sample processing and sequencing (Genewiz, South Plainfield, NJ).

Clean, raw data was obtained by removing the reads containing adapters using Trimmomatic. The sequence RefSeq: NC_CP010468 was employed for mapping. RNA reads were mapped using the software bowtie2, and featureCounts was employed to read counts. SARTools (Statistical Analysis of RNA-Seq data Tools) (Varet et al., 2016) was used for statistical RNA-Seq analysis. Differentially expressed genes (DEGs) were identified using the DESeq2 R Package. The functional classification of the DEGs was performed using Gene Ontology (GO) analysis by Blast2GO (Götz et al., 2008). The data discussed in this publication have been deposited in NCBI's Gene Expression Omnibus (Edgar, 2002) and are accessible through GEO Series accession number GSE140847.

Genome-Scale Metabolic Network Reconstruction

In order to obtain metabolic engineering targets to overproduce succinic acid from glycerol, two *E. coli* models were used: *EColiCore2* (ECC2) (data under peer review) and iTA1338 for *E. coli* ATCC 8739 (**Supplementary File 1**). Gene associations for both models were modified to ECOLC_RS number based on the sequence RefSeq: NC_CP010468 to facilitate the integration of transcriptomics data. Extensive manual curation was conducted, including (i) adding/eliminating transport reactions and extracellular metabolites and (ii) filling pathway gaps. GapFind and GapFill, two optimization problems that search for root metabolite problems that are not connected in the network and that solve them, were used to fill gaps in iTA1338, including biomass reaction BIOMASS_Ec_iML1515_WT_75p37M (**Supplementary File 1**). All optimization problems were solved using the COBRA Toolbox v.3.0 (Heirendt et al., 2019).

Transcriptomics Integration and Metabolic Engineering Target Prediction

The gene inactivity moderated by metabolism and expression (GIMME) (Becker and Palsson, 2008) method was used to integrate transcriptome data with the *E. coli* metabolic model. This method then minimized the usage of low-expression reactions while keeping the objective (e.g., biomass) above a certain value. Expressed genes were considered according to their expression level with log2 fold change (FC) $\geq |1|$. Next, according to the GPR rules and the defined gene expression states, a specific activity state for each reaction was identified. Finally, a specific context model was obtained from the transcriptomic data. Metabolic engineering targets were obtained using OptKnock. However, MOMA was used to understand the probability of those mutants predicted to be adapted and to reach the optimal state (predicted succinic and growth flux) considering the Euclidean distance. It because OptKnock predicts an optimal state, but after genetic manipulation cell are not in this state. The maximum uptake rate of glycerol was set to 13.3 mmol/g DW h^{-1}. The OptKnock, GIMME, and MOMA methods were conducted using COBRA Toolbox v.3.0 (Heirendt et al., 2019) in MATLAB 2017b and Gurobi 8.0.1.

Machine Learning to Determine Potential Metabolic Engineering Targets

Random forest models are supervised machine learning approaches, which have the advantage of giving a summary of the importance of each variable. This approach is based on a randomized variable selection process. An estimation of variable importance is provided by *IncNodePurity*, which measures the decrease in tree node purity that results from all splits of a given variable over all trees (Li et al., 2015). For interpretation purposes, this measure can be used to rank variables by the strength of their relation to the response variable (Li et al., 2015). A matrix of binary values was built from m mutant predicted and n reactions in the set of possible reactions to be knocked out. In this matrix, one represents the presence of one specific reaction to be deleted in the mutant and zero the absence in the combination of reactions to be deleted in the mutant. The matrix was partitioned into training and test sets; the training set was used to build a random forest model to predict succinic acid production, growth rate, or the growth rate Euclidean distance between the mutant and WT strains as response variables. For the training set, succinic acid production, growth rate variable response was initially predicted using FBA, and the growth rate Euclidean distance between the mutant and WT strains was predicted using MOMA. Next, the model performance was assessed using the testing set. Finally, we used the random forest to determine the importance of each target reaction over the three evaluated response variables.

RESULTS

Glycerol Consumption of *E. coli* After Adaptive Laboratory Evolution

Luria–Bertani is one of the most common cultures used industrially for the growth of *E. coli*. In order to increase glycerol consumption by *E. coli* on LB media, an ALE experiment was carried out. Results obtained in this study, before the ALE experiments, suggest that even when high cell density cultures are reached, a low consumption of glycerol is observed. For all the four conditions (supplementation of 30, 40, 50, or 60 g/L of glycerol), a growth curve was carried out, showing that a maximum of 7 g/L consumption of glycerol could be achieved naturally by *E. coli*. Nevertheless, after the ALE experiments, an increase of 3 g/L in the glycerol consumption was observed for the strain growing in a supplementation of 30 g/L of glycerol. Despite this data showing an increase of around 30% in glycerol consumption, it is far below that obtained in the optimized culture, which reaches a consumption of 30 g/L of glycerol (data under peer review).

Transcriptional Response of *E. coli* for Aerobic Glycerol Consumption

A cell is considered a complex system where a large number of processes are carried out. These processes then involve an interaction between genes, transcripts, proteins, metabolites, and reactions, among others (Lee et al., 2012; Furusawa et al., 2013; Rangel et al., 2020). Metabolic models are reconstructed by using genome information; however, it is well known that metabolism is given by environmental conditions by passing through a cell regulation process. This causes some genes to be turned on and off under certain conditions. To determine which reactions are active to obtain high accurate models, two transcriptomic profiles were obtained from an ALE experiment and an optimized culture medium.

DEGs were determined using the DESeq2 statistical package after filtering out low count reads with an average value of <100. Significant DEGs were defined as those whose abundance had at least a log2 fold change [(log2 FC) > | 2|] between the reference condition (glycerol-based medium) and a chosen experimental condition (optimized culture medium and evolved strain) at a false discovery rate (FDR)-corrected $P < 0.05$. Relevant genes with log2 FC > | 1| for glycerol metabolism or under the same regulon were taken into account. **Figure 2** shows the distribution of DEGs using a log2 FC $\geq$ |2| for one strain growing in the optimized culture medium and one evolved strain growing in the same optimized medium. This analysis determined that 478 genes were differentially expressed, with 222 genes downregulated and 256 upregulated for the optimized medium, and 431 DEGs for the evolved strain, of which 223 genes were downregulated and 208 genes were upregulated. When comparing DEGs in the optimized medium and those in the evolved strain, 59 downregulated genes were found to be unique in the evolved strain and 58 unique genes for the optimized medium. In this context, 47 and 95 upregulated genes were found to be unique in the evolved strain and the optimized medium, respectively (**Figure 2**).

DEGs were classified into the following three groups using GO analysis: biological processes, molecular functions, and cellular components. The shared downregulated genes predominantly included those involved in the metabolic process (cellular, organic substances, nitrogen compounds, and primary metabolic processes), chemicals, stress and stimulus responses, and heterocyclic compound systems. Between downregulated genes, we found *phoB* and *phoR*, which are involved in phosphorous uptake and metabolism since, under excess phosphorous, PhoR inactivates *phoB* (Makino et al., 1989). **Figure 3** shows the level 2 GO terms for unique down- and upregulated genes in both conditions using Blast2GO (Götz et al., 2008). The 117 unique downregulated genes at log2 FC $\geq$ |2| and an adjusted $P \leq 0.05$ were classified into 15 functional groups. Two GO terms, signaling and locomotion, were only present for the evolved strain, and one GO term, multiorganism processes, was only present for the optimized culture condition in downregulated genes (**Figure 3A**).

GO analysis revealed that shared upregulated DEGs (**Figure 2**) are involved mostly in the metabolic process (51%), including GO terms such as cellular, organic substances, primary, and nitrogen compound processes; 11% of the upregulated genes were associated with biosynthetic processes and the establishment of localization. The main GO terms for molecular functions were those involved in a binding activity (66%), counting ions, heterocyclic compounds, organic cyclic compounds, small molecules, and protein binding, followed by transferase activity (10%) and transmembrane transporter activity (9%). About 42% of the DEGs categorized in cellular functions were implicated in membrane GO terms, with 17% in the cell periphery and 16% in the cytoplasm.

Glycerol metabolism in *E. coli* is mediated by *glp* operons. In consequence, transcriptomic analysis shows shared upregulation of *glpBCFKQTX* genes. The changes in bacterial gene expression in response to glycerol utilization are summarized in **Table 1**. During glycerol utilization, GlpF permease facilitates glycerol entry into *E. coli* for further transformation into glycerol-3-phosphate (Gly-3-P) by GlpK under aerobic conditions. Comparing *glpK* expression with the values obtained for other genes in the glp regulon showed that *glpK* was one of the most highly expressed genes. However, a difference of ∼1 log2 FC between the evolved strain and the optimized culture condition was exhibited in the *glpFKX* operon (**Table 1**). As a consequence of the regulatory network, an increase in the expression of *glpX* was detected (2.76 log2 FC), which is part of the *glpFKX* operon and works as an alternative fructose-1,6-bisphosphatase involved in gluconeogenesis by catalyzing the hydrolysis of fructose-1,6-bisphosphate to fructose 6-phosphate (Booth, 2014). Overexpression of *glpX* has been shown to increase hydrogen production (Kim et al., 2011). Additionally, transcriptomic analysis showed upregulation of both flavin oxidases *glpD* and *glpABC*.

The electron-transport chains of *E. coli* are composed of many different dehydrogenases and terminal reductases. Glycerol metabolism in *E. coli* uses oxygen as the main electron acceptor, but it could also employ fumarate under anaerobic conditions by encoding a fumarate reductase complex under anaerobic

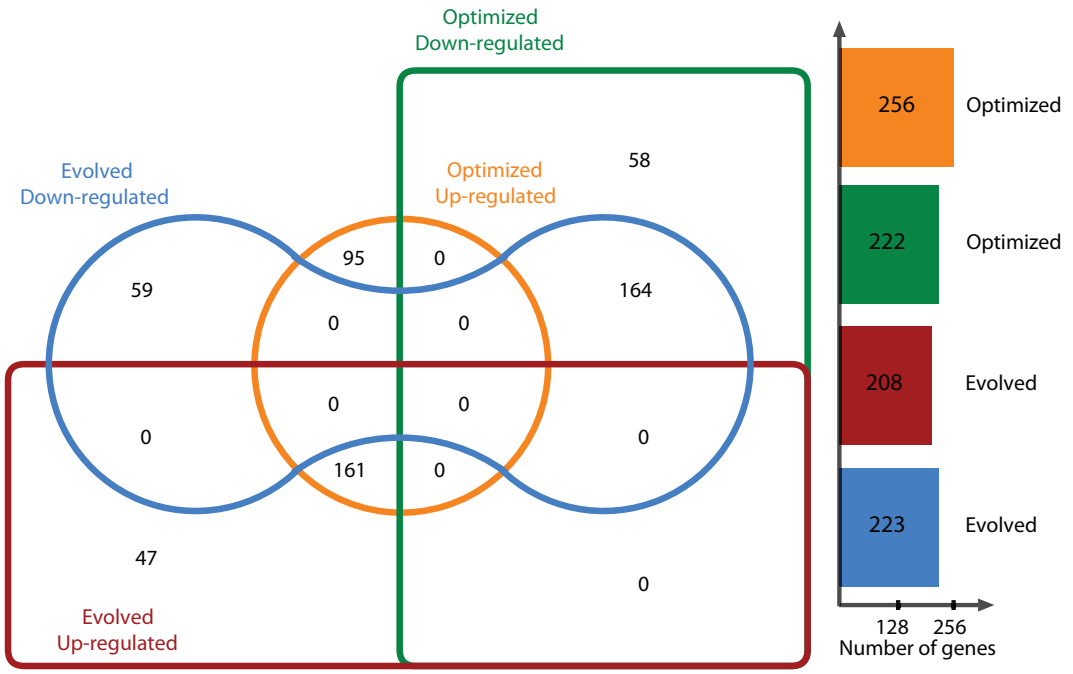

FIGURE 2 | *E. coli* ATCC 8739 differentially expressed genes in the optimized medium vs. glycerol-based medium and the evolved strain vs. glycerol-based medium.

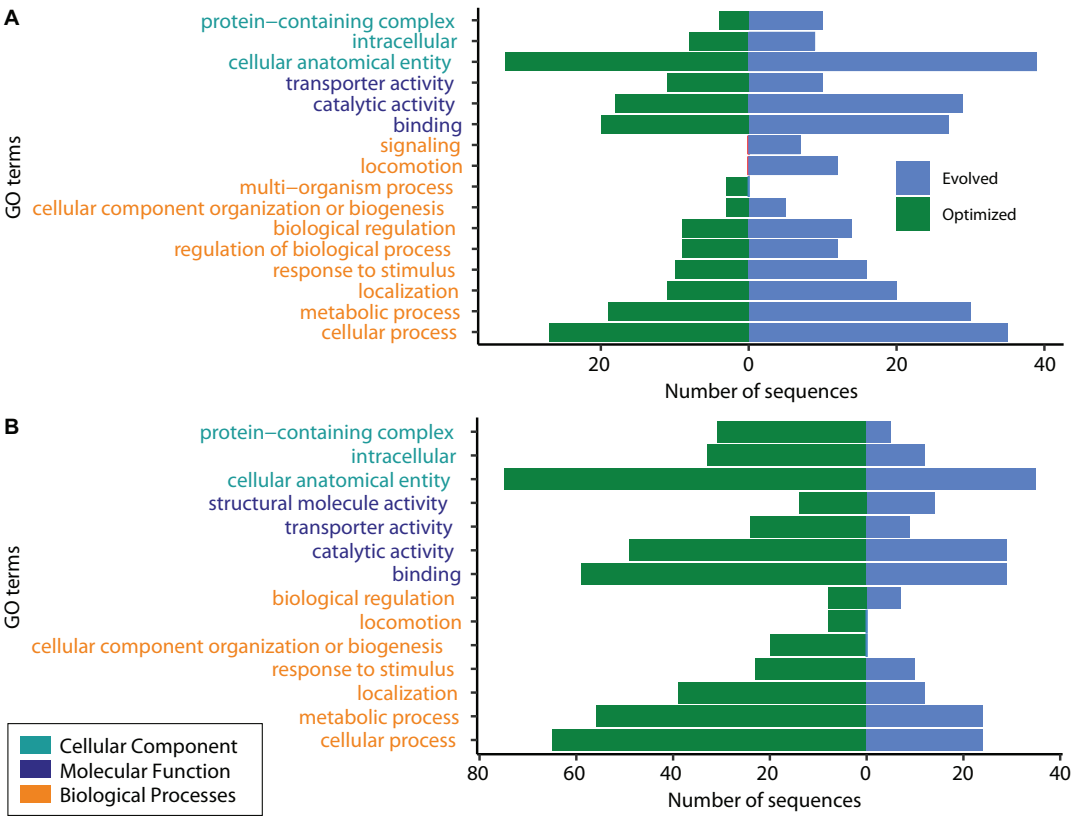

FIGURE 3 | Gene Ontology (GO) analysis of unique differentially expressed genes. **(A)** Downregulated genes. **(B)** Upregulated genes.

TABLE 1 | Differential expression of genes involved in glycerol metabolism.

RefSeq tag (ECOLC_RS)	Gene name	Old locus tag	Product	Log2 FC Exp 3	Log2 FC evolved
01540	*glpD*	EcolC_0288	Aerobic glycerol-3-phosphate dehydrogenase	1.69	1.21
07540	*glpC*	EcolC_1408	Anaerobic glycerol-3-phosphate dehydrogenase subunit C	2.61	3.59
07545	*glpB*	EcolC_1409	Anaerobic glycerol-3-phosphate dehydrogenase subunit B	2.87	3.74
07550	*glpA*	EcolC_1410	Sn-glycerol-3-phosphate dehydrogenase subunit A	1.28	2.00
07555	*glpT*	EcolC_1411	Glycerol-3-phosphate transporter	5.41	4.36
07560	*glpQ*	EcolC_1412	Glycerophosphoryl diester phosphodiesterase	5.25	5.52
22045	*glpF*	EcolC_4091	Aquaporin	4.17	3.03
22050	*glpK*	EcolC_4092	Glycerol kinase	5.36	4.37
22055	*glpX*	EcolC_4093	Fructose-1,6-bisphosphatase	2.76	2.05
10840	*fumA*	EcolC_2018	Fumarate hydratase	1.46	−0.60
20740	*frdA*	EcolC_3856	Fumarate reductase flavoprotein subunit	0.68	1.71
20745	*frdB*	EcolC_3857	Fumarate reductase iron-sulfur subunit	0.89	1.87
20750	*frdC*	EcolC_3858	Fumarate reductase subunit C	0.74	1.72
20755	*frdD*	EcolC_3859	Fumarate reductase subunit D	0.16	1.09

conditions (Jones and Gunsalus, 1987; Cecchini et al., 2002). **Table 1** shows log2 FC for *fumA* and *frdABCD* genes in *E. coli*. The *fumA* gene was encoded for abundant fumarase, predominantly expressed in the optimized culture medium (1.55 log2 FC), but not for the evolved strain (−0.53 log2 FC). FumA has been reported to be predominantly expressed under aerobic conditions (Chen et al., 2012). Under aerobic conditions, the catalysis of succinate to fumarate interconversion is mediated by the succinate dehydrogenase complex encoded by *sdhABCD* (Cecchini et al., 2002). However, in this study, *sdhABCD* genes were not found to be differentially expressed in any of the culture conditions. Interestingly, among the upregulated genes in the adapted strain, a difference of ∼1 log2 FC in the expression of the fumarate reductase genes (*frdABCD*), which is used in anaerobic growth, was observed over the optimized culture condition.

The maltose operon of *E. coli* consists of genes that encode proteins involved in the uptake and metabolism of maltose and maltodextrins. These genes have been found to be highly associated with upregulation under glycerol utilization as a carbon source, and changes in the level of *glpK* transcription had a significant effect on *malT* transcription (Chagneau et al., 2001). In this study, *malEFKMTPQ* genes were shown to be upregulated in both conditions. For *malT*, the log2 FC was more highly expressed in the optimized culture condition than in the evolved strain. The same behavior was observed for *glpK*. Thus, a high expression of this regulon in this study could be presumably linked to the high expression of the *glpK* gene since they showed similar log2 FC.

As a result of glycerol metabolism, acetate is mainly generated. In our analysis, the phosphate acetyltransferase encoded by *pta*, which catalyzes the reversible conversion between acetyl-CoA and acetylphosphate (Lin et al., 2005; Blankschien et al., 2010), was found to be upregulated (∼2.30 log2 FC). Also, the *atpABCDEFGH* genes have a role in the generation of ATP from ADP and phosphate. These genes were observed to be upregulated, with similar log2 FC, except for *atpA*, which had a difference of around 1 log2 FC in the optimized culture medium with respect to the evolved strain.

Predicting Potential Metabolic Engineering Targets for Succinic Acid Overproduction

Genome-scale metabolic models (GEMs) are defined as a complete set of reactions involved in cell metabolism, given by genome annotation, regardless of whether the annotated metabolic genes are expressed in a given environment. This assumption could be correct in genome-scale models because core models represent the central metabolism, but the full potential of GEMs remains unexploited mainly (Ataman et al., 2017). To avoid this situation and to evaluate the effects of using a core or a large model to predict metabolic engineering targets, three models were used: a core model (ECC2) and two models obtained after the integration of transcriptomics data that can help to elucidate the actual state of the metabolic network *in vivo* for further metabolic engineering.

Metabolic Model Reconstruction and Transcriptomics Integration

For the integration process, a reconstruction of the metabolic model for *E. coli* ATCC 8739 was carried out based on the iEcolC 1368 (Monk et al., 2013), iEC1349_Crooks (Monk et al., 2016), and iML1515 models (Monk et al., 2017). Extensive manual curation was conducted to fill pathway gaps. Transport and exchange reactions were added or eliminated, enabling nutrient uptake and by-product secretions. Finally, the resulting model was designated iTA1338, and it involved 2,032 metabolites, 2,804 reactions, and 1,338 genes (**Supplementary File 1**). After that, using GIMME, context-specific metabolic networks were constructed departing from the iTA1338 model for two types of strains: (1) WT *E. coli* ATCC 8739 growing in an optimized culture medium (iTA818) (**Supplementary File 1**) and (2) *E. coli*

ATCC 8739 strains evolved to grow on glycerol (iTA821) (**Supplementary File 1**). Manual curation was carried for the iTA821 model based on GapFind and GapFill results.

Figure 4 illustrates the number of reactions obtained for each model after transcriptomic integration. The same growth rate was observed after integration; however, flux distribution in 24 reactions was exhibited (**Figure 4B**). The reactions only present in iTA821 are mainly associated with the inner membrane transport (14). Other unique reactions in iTA821 were mapped to be linked to the citric acid cycle, cofactor and prosthetic group biosynthesis, glutamate metabolism, inorganic ion transport and metabolism, the nucleotide salvage pathway, oxidative phosphorylation, and pyruvate metabolism, among others. Unusual reactions of iTA818 were mainly associated with transport, including the transport outer membrane porin (218), transport inner membrane (50), and transport outer membrane (15), followed by cell envelope biosynthesis (37), the nucleotide salvage pathway (24), glycerophospholipid metabolism (14), alternate carbon metabolism (12), and cofactor and prosthetic group biosynthesis (7), among others.

In silico Systems Metabolic Engineering Targets Prediction

To predict *E. coli* strains that overproduce succinic acid from glycerol, OptKnock was used (Burgard et al., 2003). Before predicting the reaction target to overproduce succinic acid, both metabolic networks were preprocessed. The goal of preprocessing was to obtain a smaller set of selected reactions that could serve as valid targets for gene knockouts. First, all reactions displaying maximum and minimum fluxes equal to zero were removed from the set of potential reactions to be knocked out. Next, all reactions that had been experimentally found to be essential for growth were removed from consideration (Joyce et al., 2006). Also, the reactions that were found to be computationally essential were not considered, as well as non-gene-associated reactions.

Ten OptKnock rounds of mutant prediction were carried out. In each round, the set of reactions was set up to 1, 2, 3, . . . 10, and 100 mutants were requested per round. ECC2, iTA818, and iTA821 models were used to predict mutants of succinic acid overproducers; 811, 806, and 785 possible mutants were obtained from the ECC2, iTA818, and iTA821 models, respectively (**Supplementary File 2**). **Figure 5** describes the frequency of the reactions predicted in all the possible mutants. It can be seen that 30 reactions were above the average frequency. Reactions acetate kinase (ACKr), fructose 6-phosphate aldolase (F6PA), fumarase (FUM), pyruvate dehydrogenase (PDH), pyruvate formate lyase (PFL), phosphotransacetylase (PTAr), succinate dehydrogenase (SUCDi), triosephosphate isomerase (TPI), glycerol-3-phosphate dehydrogenase-NADP (G3PD2), and glycerol dehydrogenase (GLYCDx) were frequently predicted for the all models. It is important to mention that G6PDH2r, LDH_D, PGL, and POX were not predicted to be part of models iTA818 and iTA821 after integration.

Interestingly, in the complete set of reactions predicted, PDH was the most frequent target reaction, followed by FUM in all the models (**Figure 5**), and minimal variations in the knockout frequency were observed for these reactions. **Figure 5A** shows the plot of the first two principal components of the principal components analysis (PCA), representing the variability of 89% of the data. This analysis shows how PDH and FUM knockouts

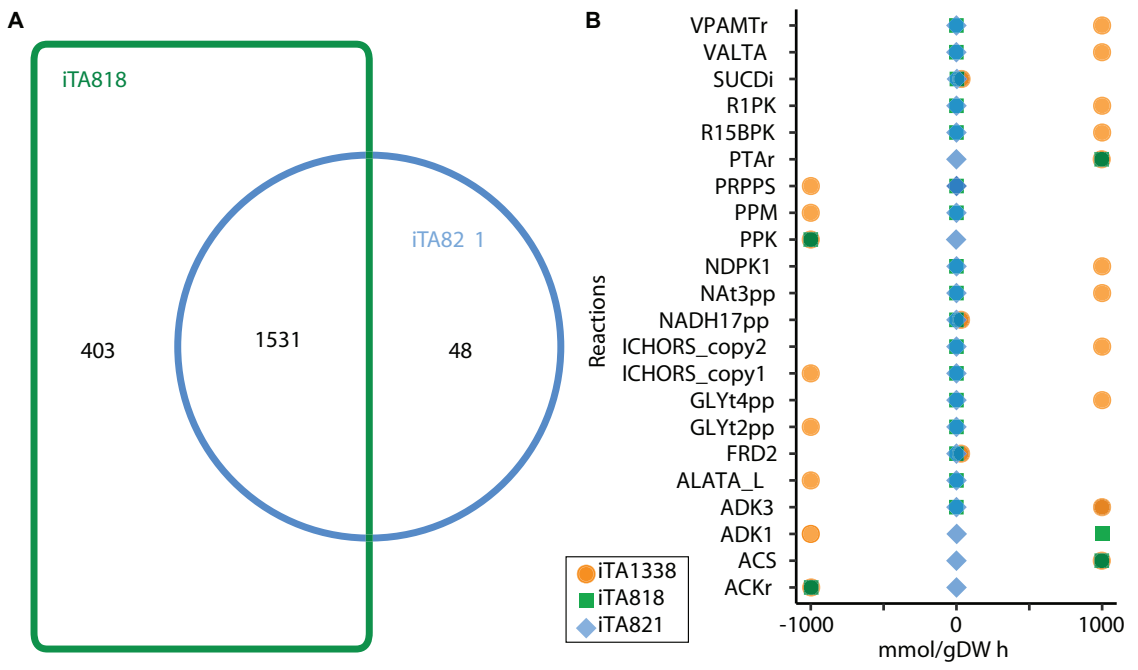

FIGURE 4 | Models comparison after transcriptomic integration using gene inactivity moderated by metabolism and expression (GIMME) under aerobic conditions. **(A)** Venn diagram or reactions included in the model. **(B)** Flux Balance Analysis under aerobic conditions.

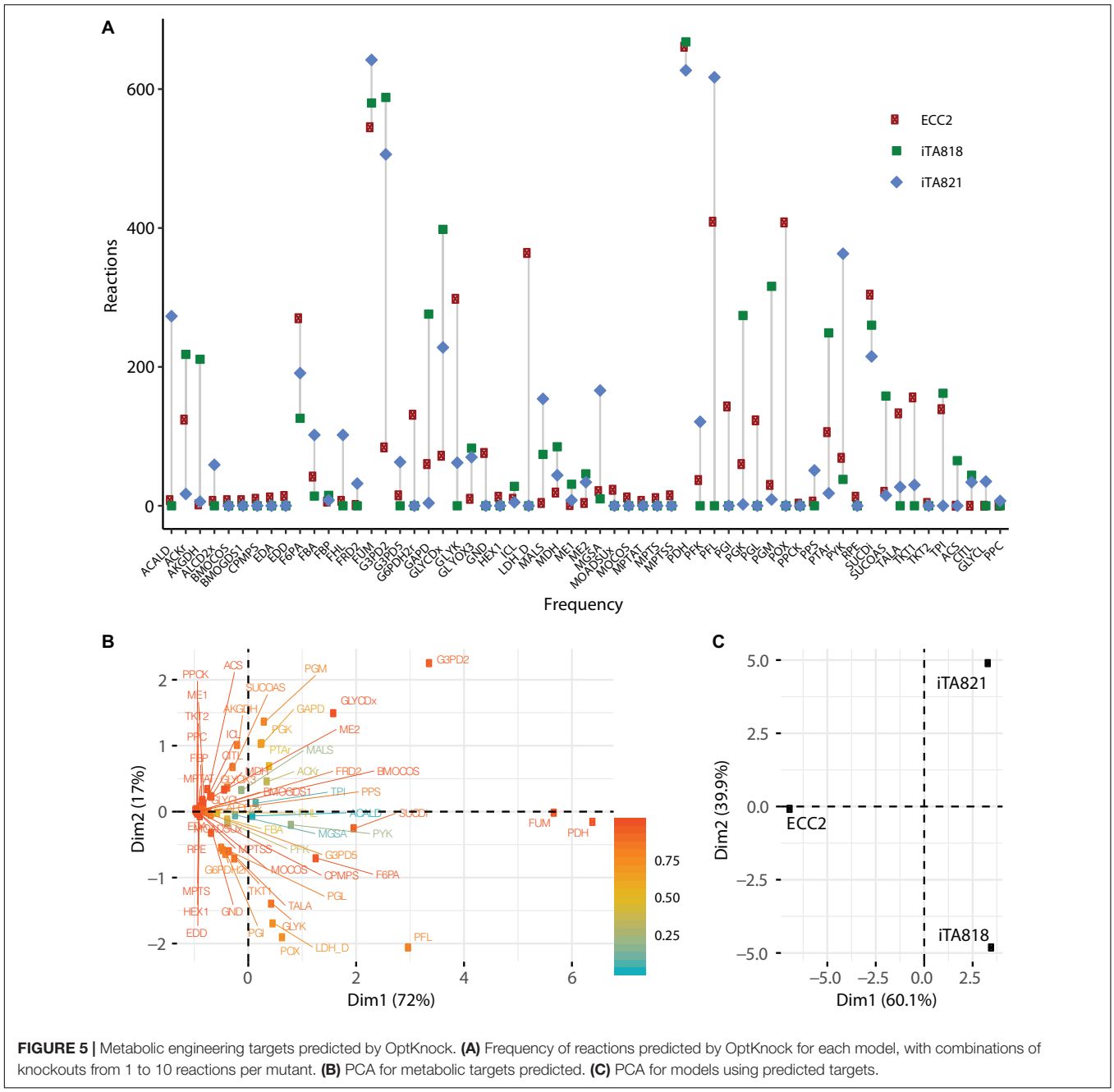

FIGURE 5 | Metabolic engineering targets predicted by OptKnock. **(A)** Frequency of reactions predicted by OptKnock for each model, with combinations of knockouts from 1 to 10 reactions per mutant. **(B)** PCA for metabolic targets predicted. **(C)** PCA for models using predicted targets.

are closely related to succinic acid overproduction from glycerol. Regions of high variability are clustered along with the first principal component, presenting a value of zero for the first principal component. This indicates that the factors that make up the first principal component are critical for high titers. The contributions of different models to the first two principal components of the PCA are shown in **Figure 5B**, and they are indicative of the relative influence on the variability in knockout predictions given by transcriptomic integration.

A cluster analysis between the reaction frequency for each *k* deletion showed that elimination of acetate, formate, and lactate by-products mediated by POX, PFL, and LDH_D is highly related to PDH and FUM deletion (**Figure 6**). This phenomenon, probably due to PDH deletion, results in reduced conversion of pyruvate to acetyl-CoA, which is the main substrate in ACKr and PTAr reaction to generate acetate (**Figure 1**), a competitive by-product on succinate production (Blankschien et al., 2010). Then, if PDH deletion is not carried out, ACKr and PTAr knockouts would become essential to increasing succinate production, as well as minimizing costs in the separation process (Kurzrock and Weuster-Botz, 2010; López-Garzón and Straathof, 2014).

Since metabolic manipulation of cells results in a stressful process, the negative impact of deletions on the maximum growth rate can be observed. To determine the effects of reaction

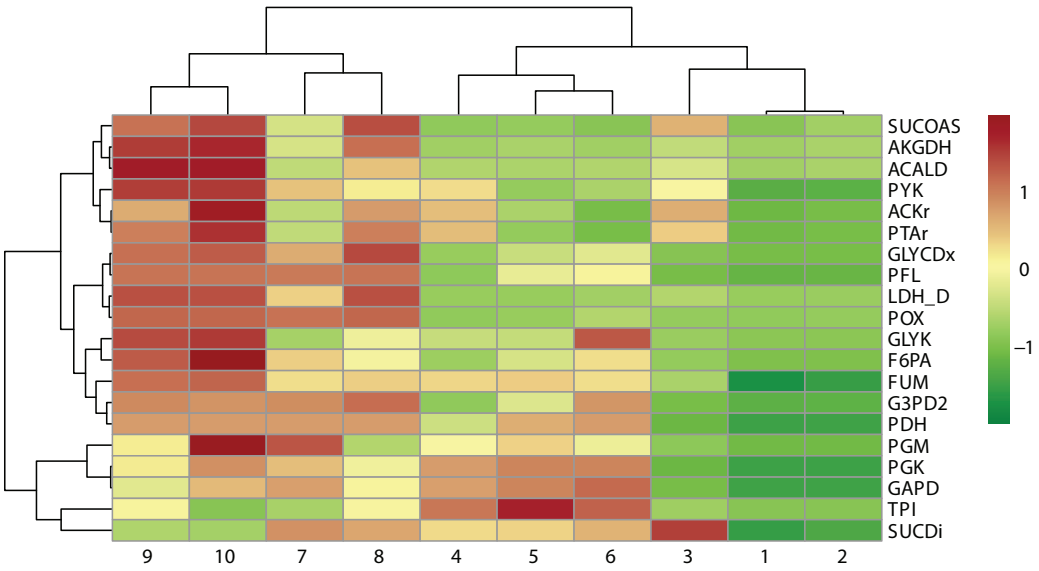

FIGURE 6 | Pearson correlation of reaction frequency by knockout numbers (*k*), predicted using OptKnock.

knockouts over the cell, FBA was carried out and Euclidean distance was calculated for each mutant predicted. **Figure 7** illustrates the relationship between the number of knockouts, succinic acid production, and growth rate using FBA and the Euclidian distance between the WT and mutant strains using MOMA. It can be seen that the number of reactions knocked is highly related to high succinic acid production rates due to the elimination of competitive by-products, such as acetate, formate, and lactate, requiring at least three to four deletions. The highest succinate production (∼8.5 mmol/g DW h^{-1}) was observed in mutants predicted in ECC2 when 9 or 10 reactions were deleted. However, this implies a substantial reduction in the growth rate to ∼4% compared with the WT strain. Thus, selecting these mutants is unrealistic for the industrial production of succinic acid. The same behavior in the reduction of the growth rate was observed for those mutants that required more than six deletions in mutants predicted in iTA818 and iTA821. In contrast, a considerable reduction in the growth rate (28% of the WT growth rate) as well as an increasing succinic acid production rate (around 30% more than those with 9–10 knockouts) for those mutants with six knockouts was observed. In addition, it was observed that there is no direct correlation, in the same magnitude for all the mutants, between the Euclidean distance and the numbers of knockouts in each mutant. However, **Figure 7** shows amplifications in the Euclidian distance between the WT and the mutants when succinate production and knockout numbers increase and growth rate decreases.

Identification of Critical Metabolic Targets and Potential Mutants

OptKnock results are a large list of knockout combinations where maximum product synthesis occurs at a maximum growth rate reachable (Burgard et al., 2003). However, it has been observed that the optimal solution of the target given by OptKnock is not necessarily growth-coupled, and some mutants predicted do not increase the product target. Consequently, selecting a mutant to be tested in the lab could be really difficult and probably result in a laborious process. Assuming that each mutant product growth "coupled" predicted will result in a successful biological production, these mutants can ensure high productivity over time and initially solve this situation (Shabestary and Hudson, 2016). To identify growth-coupled production solutions, a COBRA Toolbox function was used to verify the minimum and maximum production rates given a set of reactions to be knocked out. As a result, the same minimum and maximum flux for the desired product should be obtained when the maximum growth rate is achieved. One thousand seven hundred ninety-nine (1,799) mutants were predicted to be growth-coupled, 539 to be growth-coupled non-unique (maximum flux - minimum flux > 0.1), and 64 mutants were categorized as not growth-coupled (maximum flux < 0.1). For the mutants categorized as growth-coupled non-unique, an FBA was carried out to predict the succinic acid production rate (**Figure 7**), where 279 mutants were predicted to have a difference between the maximum production rate predicted by the function and FBA < 2, resulting in 2,078 *in silico* mutants that overproduce succinic acid.

In order to filter and select potential mutants to be tested in the lab, a random forest model to predict the importance of each reaction knockout was developed based on the OptKnock predictions. Each possible combination of reactions using binary values that increase the succinic acid production was associated with the flux of the extracellular succinic acid and biomass reaction obtained by FBA and the Euclidian distance obtained by MOMA. The dataset was divided into two groups: 70% for training and 30% for the test. Following feature selection and cross-validation, a robust model that associated any combination of 58 reaction variables to a predicted growth rate and succinic

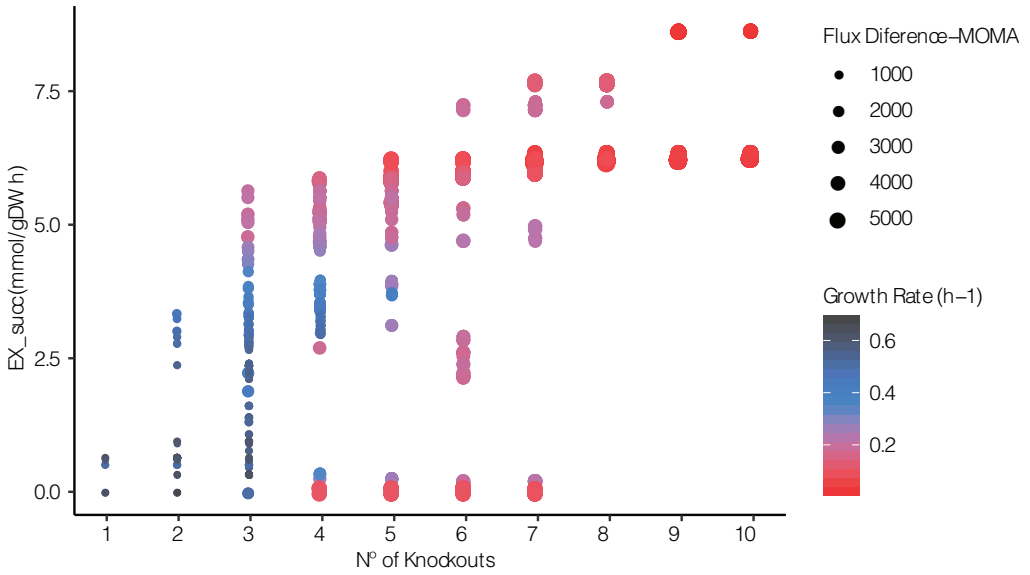

FIGURE 7 | Relationship between the number of knockouts, succinic acid production, growth rate, and flux difference.

acid production ratio was obtained. A measure of the importance of the contribution of each feature to the random forest model is shown in **Figure 8** indicated by *IncNodePurity*. This model exhibited a mean square error (MSE) value of 0.293 when using the reaction flux of EX_succ_e flux obtained by FBA as a variable response. For growth rate (biomass reaction) as a response variable, the MSE value was 0.0002. Finally, when the Euclidian distance for each mutant was used as the response variable, the MSE value was 9,175.158, indicating that the Euclidian distance is not a good response variable to predict cell behavior when using the random forest model. Moreover, this result allows the use of machine learning models to predict the largest number of mutants than those obtained by OptKnock in terms of growth rate and succinic acid production since OptKnock is more time-consuming.

Figure 8A shows that PFL, LDH_D, GLYCDx, G3PD2, PDH, and POX are the most important reactions to increase the amount of succinic acid. These reactions are mainly associated with the GldA–DhaKLM fermentative route and the Gly-3-P route (**Figure 1**) in glycerol utilization (Blankschien et al., 2010), as well as acetyl-CoA generation given by the PDH knockout. In around 24% of the mutants predicted, a combination of GLYCDx and G3PD2 reactions was found to increase succinic acid production. However, POX and LDH_D reactions were not present in iTA818 and iTA821 models, and PDH, G3PD2, and PFL were also found to be the most important reactions, predicted to have an effect on growth rate (**Figure 8B**).

The pyruvate dehydrogenase complex is a critical connection point between glycolysis and the TCA cycle, both of which function during aerobic respiration through catalyzing the conversion of pyruvate to acetyl coenzyme A (acetyl-CoA) (Schutte et al., 2015). PDH deactivation results in PFL carrying the flux from pyruvate to acetyl-CoA (Khodayari et al., 2015). Simple reaction knockouts show that PDH deletion results in a growth rate reduction of ∼5%. Additionally, five reactions (FUM, GAPD, PGK, PGM, and TPI) were predicted to have the most significant reduction (8–10%) in growth rate during glycerol utilization. Of those reactions, only FUM has a significant positive effect over succinate production when this deletion was carried out alone. However, in mutants in which both FUM and PDH were predicted (59.45%), TPI appeared in around 12.60% (**Figure 5**). Then, the deletion of genes associated with TPI in addition to FUM and PDH reactions could negatively affect growth rate.

DISCUSSION

Glycerol metabolism in *E. coli* has been described in the literature (Murarka et al., 2008; Booth, 2014). However, cell changes are carried out as a response to stressful situations. In this study, two conditions were tested for transcription response in *E. coli* to further integrate to metabolic network modeling. Gene expression-wide analyses reveal how cells have the ability to avoid glycerol toxicity, increasing consumption. The most striking response to glycerol consumption and the possible mechanism to optimize succinic acid production from glycerol were revealed by the combination of the transcriptome, metabolic modeling, and machine learning analyses.

After glycerol incorporation in the cell mediated by GlpF, glycerol can be metabolized through two pathways. The first is mediated by the glycerol kinase GlpK through phosphorylation of glycerol to Gly-3-P, followed by GlpD activity under aerobic conditions, leading to dihydroxyacetone phosphate (DHAP) (**Figure 1**). The alternative pathway consists of an oxidation step by glycerol dehydrogenase (GldA) to yield dihydroxyacetone (DHA), followed by phosphorylation by DHA kinase (DhaK) to yield DHAP as well. In this study, overexpression of *glpK* was

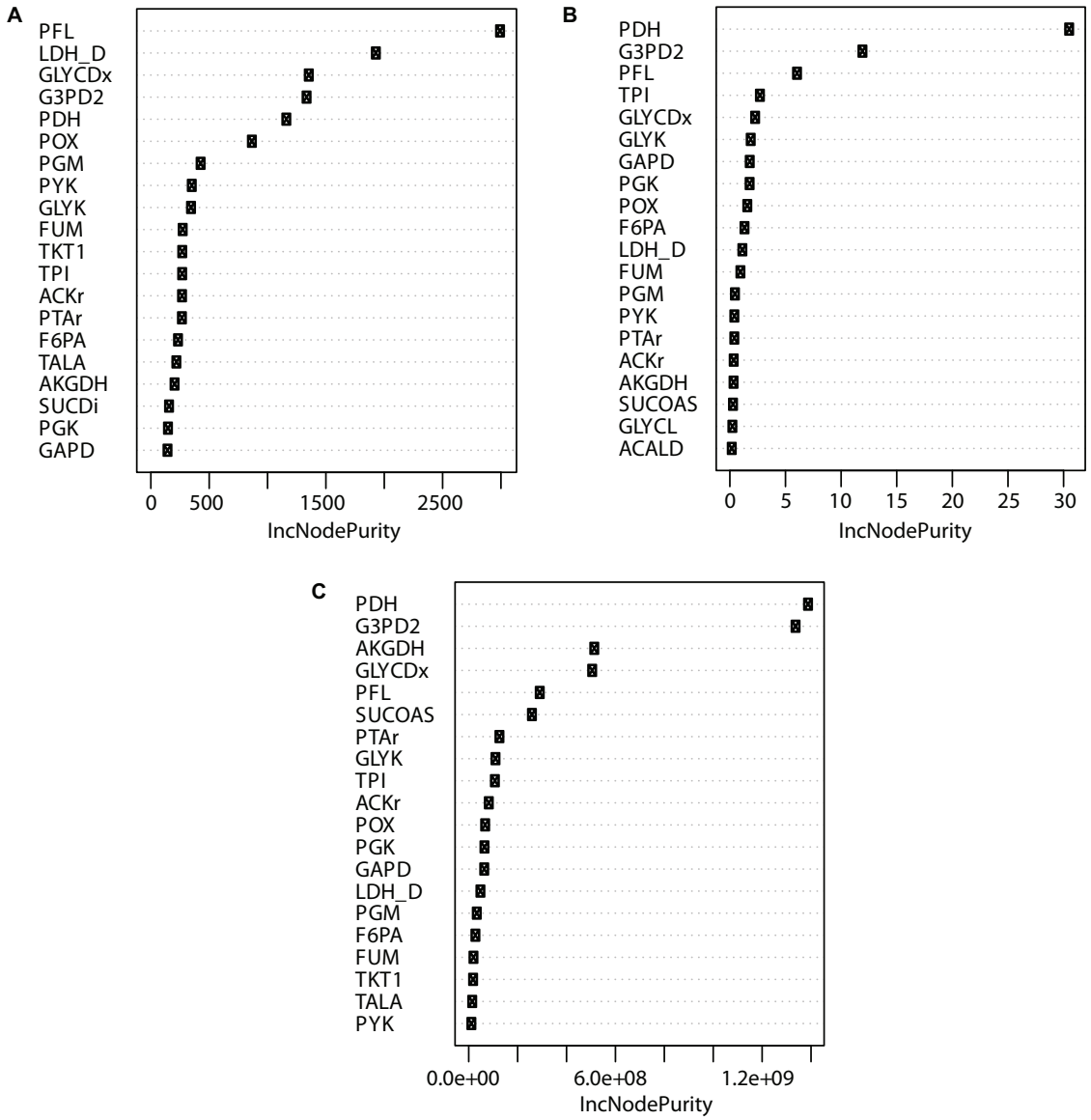

FIGURE 8 | Top 20 node purities obtained using random forest for reactions predicted by OptKnock. **(A)** Reaction knockout importance on succinic acid production. **(B)** Reaction knockout importance in growth rate reduction. **(C)** Reaction knockout importance in Euclidian distance.

observed in both conditions, with a difference of around 20%. This result is not surprising since the GlpK-mediated reaction is a rate-limiting step in glycerol utilization (Herring et al., 2006). However, it has been observed that under the optimized culture conditions, the glycerol utilization rate is higher than that in the evolved conditions, suggesting that other mechanisms should exit in the cell to enhance glycerol utilization. Gly-3-P is the first intermediate between the glycerol pathway and the TCA cycle, as well as between the biosynthesis and catabolism of lipids; however, accumulation of Gly-3-P can become toxic. Thus, it is carefully regulated (Booth, 2014). The export of Gly-3-P could be mediated by *phoE* and *ompF* membrane porins; however, downregulation of *phoE* (−8.67 and −9.04 log2 FC for the optimized culture and the evolved strain, respectively) and upregulation of *ompF* (log2 FC 2.43) in the optimized culture suggest that it could play an essential role in *E. coli* ATCC 8739 glycerol metabolism at high uptake rates avoiding toxicity.

The marked upregulation of *glpQ* (5.351 and 5.597 log2 FC for the optimized culture and evolved strain, respectively), which catalyzes the hydrolysis of glycerol-phosphodiesters to alcohol plus Gly-3-P together with *ompF*, could explain the partially higher transcript abundance of *glpT* since the externally generated (or supplied) Gly-3-P activates GlpT (Wong and Kwan, 1992; Lemieux et al., 2005). This protein exchanges Gly-3-P for phosphate, avoiding the toxicity of both Gly-3-P and the inorganic phosphates (Booth, 2014). As a result and considering that phosphate is necessary to increase glycerol utilization, autoregulation of the PhoB/PhoR two-component

regulatory system needs to be down-expressed. Downregulation of PhoB/PhoR was observed in this study, which could explain the achievement of optimal density (Gao and Stock, 2013), as well as contribute to the regulation of glycerol phosphate metabolism (Baek and Lee, 2007).

The transcriptional analysis also identified the differential expression of both flavin oxidases *glpD* and *glpABC*. Once Gly-3-P is in the cytoplasm, it is oxidized to dihydroxyacetone phosphate by one of two flavin-dependent oxidases encoded by *glpD* or *glpABC* genes under aerobic or anaerobic conditions, respectively (Blankschien et al., 2010; Booth, 2014). In the presence of oxygen or nitrates, GlpD transfers electrons to the respective terminal oxidized. In contrast, under anaerobic conditions, the GlpABC system transfers the electrons to fumarate or nitrates (Unden and Bongaerts, 1997). GlpD upregulation was expected since culture conditions were under aerobic conditions, but a higher expression of the *glpABC* system was surprising. Overexpression of *glpABC* under aerobic conditions could be elucidated because of the activation of fumarate reductase enzymes (**Table 1**) in the evolved strain as a result of high cell densities during the ALE process. However, in glycerol fermentation studies, the Δ*frdA* mutant has been shown to be beneficial for glycerol fermentation because it prevents the negative impact of hydrogen by maintaining suitable redox conditions (Murarka et al., 2008). Moreover, its activity could be supported by *sdhABCD* since they are structurally and functionally homologous (Guest, 1981). Therefore, we hypothesized that *frdABCD* upregulation could be the reason why enhancement in glycerol utilization was not observed in the evolved strain, even when an optimized culture medium was employed.

Insights on the molecular adaptive responses of *E. coli* to glycerol consumption revealed by the transcriptional datasets identified a marked *hdeAB* upregulation only in the evolved strain. This is attractive since HdeAB are periplasmic proteins that play a role in optimal protection at low pH (Masuda and Church, 2003; Kern R. et al., 2007). Therefore, differences in *hdeAB* upregulation in the evolved strain and the optimized culture medium probably occur because acetate is the main product in glycerol utilization, and under ALE conditions, pH was not controlled. Moreover, the addition of a phosphate buffer system using the salts Na_2HPO_4 and KH_2PO_4 provides the culture medium used directly for the optimized condition with a buffering capacity.

It was observed that the main and preferable route for glycerol consumption is the pathway mediated by GlpK since this gene was highly overexpressed in high glycerol consumption cultures. Moreover, *glpK* deletion has also been observed to be essential for glycerol utilization as the sole carbon source (Velur Selvamani et al., 2014). Then, the deletion of this gene could result in a non-effective bioconversion process. As a result, this gene should not be taken into account for engineered *E. coli* strains using glycerol as the carbon source even when the GLYK reaction was repeatedly predicted to be knocked by OptKnock in ECC2 and iTA821 since two pathways for glycerol utilization in *E. coli* exist.

Based on OptKnock and random forest model predictions, four critical control points, glycolysis, pyruvate metabolism, the pentose phosphate pathway, and the TCA cycle, are associated with the overproduction of succinic acid. FUM and SUCDi appear to be the most significant keys in the TCA cycle for succinate overexpression. The results of this study suggest that they are mutually exclusive. Parallelly, the knockout of by-products such as acetate, formate, and lactate by deleting POX, ACKr, PTAr, PFL, and LDH_D was highly predicted to be knocked out. Those results are interesting since one of the bottlenecks for industrial production of bio-based products is the elimination of by-products, which could facilitate the recovery and purification process. These results and those obtained in the transcriptional responses suggest that deletion of the *pta* needs to be, almost as mandatory, carried out since acetate production becomes a competitive pathway in glycerol metabolism for succinic acid production (Zhang et al., 2010).

The pyruvate dehydrogenase complex is a critical connection point between glycolysis and the TCA cycle, both of which function during aerobic respiration through catalyzing the conversion of pyruvate to acetyl coenzyme A (acetyl-CoA) (Schutte et al., 2015). PDH deactivation results in PFL carrying the flux from pyruvate to acetyl-CoA (Khodayari et al., 2015). Simple reaction knockouts show that PDH deletion results in a growth rate reduction of ∼5%. Additionally, five reactions (FUM, GAPD, PGK, PGM, and TPI) were predicted to have the most significant reduction (8–10%) in growth rate during glycerol utilization. Of those reactions, only FUM has a significant positive effect over succinate production when this deletion was carried out alone. These results indicate that those mutants predicted by OptKnock, where FUM and PDH are predicted, need to betested in the lab because it has been observed that a low growth rate could negatively affect the profitability of industrial bio-based production products (Chen et al., 2013a; Tafur Rangel et al., 2018). However, in mutants in which both FUM and PDH were predicted (59.45%), TPI appeared in around 12.60% (**Figure 5**). Then, the deletion of genes associated with TPI in addition to FUM and PDH reactions could negatively affect the growth rate. This is because in the absence of TpiA, DHAP is converted to methylglyoxal, which, even at submillimolar concentrations, is a toxic compound (Booth, 2014). DHAP is the result of the alternative pathway on glycerol metabolization consisting of an oxidation step by glycerol dehydrogenase (GldA). DHAP must be transformed into the general glycolytic pathway through isomerization by triosephosphate isomerase (TpiA) as glyceraldehyde-3-phosphate (GA3P). Therefore, deletion of *tpiA* could result in growth inhibition and cell death in the presence of glycerol as the only carbon source (Velur Selvamani et al., 2014). However, since FBA is not able to capture regulation, this situation could not be predicted by OptKnock.

Finally, computational models suggest that deletions of just six to seven reaction knockouts are beneficial for industrial production since the growth rate does not decrease extremely. It is important to consider that a similar succinate production could be achieved if six to eight reactions are knocked out for all models. An assumption using optimization methods to predict cell capabilities is that the cell could quickly adjust the metabolism to maximize growth under certain conditions. This affirmation could be true for WT strains because FBA predicts an

optimal condition. However, in metabolically engineered strains, the cell attempts to compensate for the genetic changes carried out by the fewest changes in gene regulation until it achieves an optimal state that could be predicted using FBA (Senger et al., 2015). Then, FBA in engineered strains predicts a long-term evolved state. Thus, an alternative to evaluate unevolved mutants is the MOMA method (Segre et al., 2002). MOMA solves this problem by finding the solution that is most similar to the WT state (maximization of WT growth rate). **Figure 7** shows a jump in the Euclidian distance between the WT and mutant strains when succinate production increases. This result could imply that after genetic manipulation, microbial cell factories require to be evolutionarily engineered. ALE studies have shown to provide the cell with the ability to grow under selection pressure to go up from a suboptimal state to optimal growth rate predicted using *in silico* models (Ibarra et al., 2002). Moreover, since OptKnock seeks to maximize the flux of a target chemical while maximizing the growth rate, our predictions could be beneficial for further ALE experiments because microbial cell factories have naturally evolved to maximize the growth rate. Thus, the succinic acid production rate would increase as biomass formation increases (Shabestary and Hudson, 2016) by using ALE rounds after metabolically engineering cells (Graf et al., 2019).

CONCLUSION

By adopting tools from various disciplines, computational methods for systems metabolic engineering have been developed to understand cell behavior and how level systems (RNA, proteins, and metabolites, among others) can interact inside the cell for industrial purposes. In the same way, *E. coli* has been extensively studied to become a cell factory for the production of useful bio-based chemicals and materials through its native capabilities. However, there are some challenges that still need to be overcome.

This study proposes that computational tools can accelerate the optimization of cell factories by identifying metabolic engineering targets (genes/reactions) and not just by predicting mutants that may be biologically unviable. Therefore, systems metabolic engineering reduces time in rational strain design and guides in the selection of metabolic engineering targets based on cell behavior under experimental conditions. Simultaneously, departing from traditional computational tools, new methods such as machine learning could be proposed as an interesting alternative for the reduction of computational demand. However, these techniques are dependent on the level of completeness and accuracy of the metabolic model considered, which could be improved by using omics data.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. The RNA datasets presented in this study can be found in the NCBI's Gene Expression Omnibus database and are accessible through GEO Series accession number GSE140847. Further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

AT: conceptualization, methodology, software, validation, investigation, visualization, formal analysis, and writing—original draft preparation. WR, DM, and CO: investigation. RC: methodology, software, and writing—review and editing. JG: supervision and writing—review and editing. AG: conceptualization, supervision, and writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was supported by the Gobernación del Cesar Program of Science, Technology, and Innovation for Higher Education through Ph.D. scholarships from the Colombian Ministry of Science, Technology, and Innovation.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.633073/full#supplementary-material

REFERENCES

Ataman, M., Hernandez Gardiol, D. F., Fengos, G., and Hatzimanikatis, V. (2017). redGEM: Systematic reduction and analysis of genome-scale metabolic reconstructions for development of consistent core metabolic models. Senger RS, editor. *PLoS Comput. Biol.* 13:e1005444. doi: 10.1371/journal.pcbi.1005444

Baek, J. H., and Lee, S. Y. (2007). Transcriptome analysis of phosphate starvation response in *Escherichia coli*. *J. Microbiol. Biotechnol.* 17, 244–252.

Bao, H., Liu, R., Liang, L., Jiang, Y., Jiang, M., Ma, J., et al. (2014). Succinic acid production from hemicellulose hydrolysate by an *Escherichia coli* mutant obtained by atmospheric and room temperature plasma and adaptive evolution. *Enzyme Microb. Technol.* 66, 10–15. doi: 10.1016/j.enzmictec.2014.04.017

Becker, S. A., and Palsson, B. O. (2008). Context-specific metabolic networks are consistent with experiments. *PLoS Comput. Biol.* 4:e1000082. doi: 10.1371/journal.pcbi.1000082

Blankschien, M. D., Clomburg, J. M., and Gonzalez, R. (2010). Metabolic engineering of *Escherichia coli* for the production of succinate from glycerol. *Metab. Eng.* 12, 409–419. doi: 10.1016/j.ymben.2010.06.002

Blazier, A. S., and Papin, J. A. (2012). Integration of expression data in genome-scale metabolic network reconstructions. *Front. Physiol.* 3:299. doi: 10.3389/fphys.2012.00299

Booth, I. R. (2014). Glycerol and methylglyoxal metabolism. *EcoSal. Plus* 1, 1–8.

Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003). OptKnock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84, 647–657.

Buschke, N., Schäfer, R., Becker, J., and Wittmann, C. (2013). Metabolic engineering of industrial platform microorganisms for biorefinery applications - Optimization of substrate spectrum and process robustness by rational and evolutive strategies. *Bioresour. Technol.* 135, 544–554. doi: 10.1016/j.biortech.2012.11.047

Carlson, R., Fell, D., and Srienc, F. (2002). Metabolic pathway analysis of a recombinant yeast for rational strain development. *Biotechnol. Bioeng.* 79, 121–134. doi: 10.1002/bit.10305

Cecchini, G., Schröder, I., Gunsalus, R. P., and Maklashina, E. (2002). Succinate dehydrogenase and fumarate reductase from *Escherichia coli*. *Biochim. Biophys. Acta - Bioenerg.* 1553, 140–157. doi: 10.1016/s0005-2728(01)00238-9

Chagneau, C., Heyde, M., Alonso, S., Portalier, R., and Laloi, P. (2001). External-pH-dependent expression of the maltose regulon and ompF gene in *Escherichia coli* is affected by the level of glycerol kinase, encoded by glpK. *J. Bacteriol.* 183, 5675–5683. doi: 10.1128/jb.183.19.5675-5683.2001

Chen, X., Xu, G., Xu, N., Zou, W., Zhu, P., Liu, L., et al. (2013b). Metabolic engineering of Torulopsis glabrata for malate production. *Metab. Eng.* 19, 10–16. doi: 10.1016/j.ymben.2013.05.002

Chen, X., Zhou, L., Tian, K., Kumar, A., Singh, S., Prior, B. A., et al. (2013a). Metabolic engineering of *Escherichia coli*: a sustainable industrial platform for bio-based chemical production. *Biotechnol. Adv.* 31, 1200–1223. doi: 10.1016/j.biotechadv.2013.02.009

Chen, Y. P., Lin, H. H., Yang, C. D., Huang, S. H., and Tseng, C. P. (2012). Regulatory role of cAMP receptor protein over *Escherichia coli* fumarase genes. *J. Microbiol.* 50, 426–433. doi: 10.1007/s12275-012-1542-6

Conrad, T. M., Lewis, N. E., and Palsson, B. O. (2011). Microbial laboratory evolution in the era of genome-scale science. *Mol. Syst. Biol.* 7:509. doi: 10.1038/msb.2011.42

Edgar, R. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210.

Feist, A. M., Zielinski, D. C., Orth, J. D., Schellenberger, J., Herrgard, M. J., and Palsson, B. O. (2010). Model-driven evaluation of the production potential for growth-coupled products of *Escherichia coli*. *Metab. Eng.* 12, 173–186. doi: 10.1016/j.ymben.2009.10.003

Fong, S. S., Joyce, A. R., and Palsson, B. Ø (2005). Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth. *Genome Res.* 15, 1365–1372. doi: 10.1101/gr.3832305

Fong, S. S., and Marciniak, J. Y. (2003). Description and interpretation of adaptive evolution of *Escherichia coli* K-12 MG1655 by using a genome-scale in silico metabolic model. *Society* 185, 6400–6408. doi: 10.1128/jb.185.21.6400-6408.2003

Förster, A. H., and Gescher, J. (2014). Metabolic engineering of *Escherichia coli* for production of mixed-acid fermentation end products. *Front. Bioeng. Biotechnol.* 2:16. doi: 10.3389/fbioe.2014.00016

Furusawa, C., Horinouchi, T., Hirasawa, T., and Shimizu, H. (2013). Systems metabolic engineering: the creation of microbial cell factories by rational metabolic design and evolution. *Adv. Biochem. Eng. Biotechnol.* 131, 1–23. doi: 10.1007/10_2012_137

Gao, R., and Stock, A. M. (2013). Evolutionary tuning of protein expression levels of a positively autoregulated two-component system. *PLoS Genet.* 9:e1003927. doi: 10.1371/journal.pgen.1003927

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi: 10.1093/nar/gkn176

Graf, M., Haas, T., Müller, F., Buchmann, A., Harm-Bekbenbetova, J., Freund, A., et al. (2019). Continuous adaptive evolution of a fast-growing corynebacterium glutamicum strain independent of protocatechuate. *Front. Microbiol.* 10:1648. doi: 10.3389/fmicb.2019.01648

Guest, J. R. (1981). Partial replacement of succinate dehydrogenase function by phage- and plasmid-specified fumarate reductase in *Escherichia coli*. *J. Gen. Microbiol.* 122, 171–179. doi: 10.1099/00221287-122-2-171

Haggart, C. R., Bartell, J. A., Saucerman, J. J., and Papin, J. A. (2011). Whole-genome metabolic network reconstruction and constraint-based modeling. *Methods Enzymol.* 500, 411–433. doi: 10.1016/b978-0-12-385118-5.00021-9

Heirendt, L., Arreckx, S., Pfau, T., Mendoza, S. N., Richelle, A., Heinken, A., et al. (2019). Creation and analysis of biochemical constraint-based models using the COBRA Toolbox v.3.0. *Nat. Protoc.* 14, 639–702.

Herring, C. D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M. K., Joyce, A. R., et al. (2006). Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat. Genet.* 38, 1406–1412. doi: 10.1038/ng1906

Ibarra, R. U., Edwards, J. S., and Palsson, B. O. (2002). *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 420, 186–189. doi: 10.1038/nature01149

Jones, H. M., and Gunsalus, R. P. (1987). Regulation of *Escherichia coli* fumarate reductase (frdABCD) operon expression by respiratory electron acceptors and the fnr gene product. *J. Bacteriol.* 169, 3340–3349. doi: 10.1128/jb.169.7.3340-3349.1987

Joyce, A. R., Reed, J. L., White, A., Edwards, R., Osterman, A., Baba, T., et al. (2006). Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *J. Bacteriol.* 188, 8259–8271. doi: 10.1128/jb.00740-06

Kern, A., Tilley, E., Hunter, I. S., Legiša, M., and Glieder, A. (2007). Engineering primary metabolic pathways of industrial micro-organisms. *J. Biotechnol.* 129, 6–29. doi: 10.1016/j.jbiotec.2006.11.021

Kern, R., Malki, A., Abdallah, J., Tagourti, J., and Richarme, G. (2007). *Escherichia coli* HdeB is an acid stress chaperone. *J. Bacteriol.* 189, 603–610. doi: 10.1128/jb.01522-06

Khodayari, A., Chowdhury, A., and Maranas, C. D. (2015). Succinate overproduction: a case study of computational strain design using a comprehensive *Escherichia coli* kinetic model. *Front. Bioeng. Biotechnol.* 2:76. doi: 10.3389/fbioe.2014.00076

Kim, J. (2012). *Development and Applications of Integrated Metabolic and Transcriptional Regulatory Network Models*. Madison, WI: University of Wisconsin–Madison.

Kim, Y. M., Cho, H. S., Jung, G. Y., and Park, J. M. (2011). Engineering the pentose phosphate pathway to improve hydrogen yield in recombinant *Escherichia coli*. *Biotechnol. Bioeng.* 108, 2941–2946. doi: 10.1002/bit.23259

King, Z. A., Lu, J., Dräger, A., Miller, P., Federowicz, S., Lerman, J. A., et al. (2016). BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* 44(D1), D515–D522. doi: 10.1093/nar/gkv1049

Kurzrock, T., and Weuster-Botz, D. (2010). Recovery of succinic acid from fermentation broth. *Biotechnol. Lett.* 32, 331–339. doi: 10.1007/s10529-009-0163-6

Lee, D. H., and Palsson, B. O. (2010). Adaptive evolution of *escherichia coli* K-12 MG1655 during growth on a nonnative carbon source, L-l,2-propanediol. *Appl. Environ. Microbiol.* 76, 4158–4168. doi: 10.1128/aem.00373-10

Lee, J. W., Na, D., Park, J. M., Lee, J., Choi, S., and Lee, S. Y. (2012). Systems metabolic engineering of microorganisms for natural and non-natural chemicals. *Nat. Chem. Biol.* 8, 536–546. doi: 10.1038/nchembio.970

Lemieux, M. J., Huang, Y., and Wang, D. N. (2005). Crystal structure and mechanism of GlpT, the glycerol-3-phosphate transporter from E. coli. *J. Electron. Microsc. (Tokyo).* 54(Suppl. 1), 43–46.

Li, J., Poursat, M. A., Drubay, D., Motz, A., Saci, Z., Morillon, A., et al. (2015). A dual model for prioritizing cancer mutations in the non-coding genome based on germline and somatic events. *PLoS Comput Biol.* 11:e1004583. doi: 10.1371/journal.pcbi.1004583

Lin, H., Bennett, G. N., and San, K.-Y. (2005). Metabolic engineering of aerobic succinate production systems in *Escherichia coli* to improve process productivity and achieve the maximum theoretical succinate yield. *Metab. Eng.* 7, 116–127. doi: 10.1016/j.ymben.2004.10.003

López-Garzón, C. S., and Straathof, A. J. J. (2014). Recovery of carboxylic acids produced by fermentation. *Biotechnol. Adv.* 32, 873–904. doi: 10.1016/j.biotechadv.2014.04.002

Machado, D., and Herrgård, M. (2014). Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput. Biol.* 10:e1003580. doi: 10.1371/journal.pcbi.1003580

Makino, K., Shinagawa, H., Amemura, M., Kawamoto, T., Yamada, M., and Nakata, A. (1989). Signal transduction in the phosphate regulon of *Escherichia coli* involves phosphotransfer between PhoR and PhoB proteins. *J. Mol. Biol.* 210, 551–559. doi: 10.1016/0022-2836(89)90131-9

Masuda, N., and Church, G. M. (2003). Regulatory network of acid resistance genes in *Escherichia coli*. *Mol. Microbiol.* 48, 699–712. doi: 10.1046/j.1365-2958.2003.03477.x

Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., et al. (2013). Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl. Acad. Sci. U.S.A.* 110, 20338–20343. doi: 10.1073/pnas.1307797110

Monk, J. M., Koza, A., Campodonico, M. A., Machado, D., Seoane, J. M., Palsson, B. O., et al. (2016). Multi-omics quantification of species variation of *Escherichia coli* links molecular features with strain phenotypes. *Cell Syst.* 3, 238–251.e12.

Monk, J. M., Lloyd, C. J., Brunk, E., Mih, N., Sastry, A., King, Z., et al. (2017). iML1515, a knowledgebase that computes *Escherichia coli* traits. *Nat. Biotechnol.* 35, 904–908. doi: 10.1038/nbt.3956

Murarka, A., Dharmadi, Y., Yazdani, S. S., and Gonzalez, R. (2008). Fermentative utilization of glycerol by *Escherichia coli* and its implications for the production of fuels and chemicals. *Appl. Environ. Microbiol.* 74, 1124–1135. doi: 10.1128/aem.02192-07

Nordlander, B., Krantz, M., and Hohmann, S. (2008). Hog1-mediated metabolic adjustments following hyperosmotic shock in the yeast. *Current* 20, 51–79. doi: 10.1007/4735_2007_0256

O'Brien, E. J., Monk, J. M., and Palsson, B. O. (2015). Using genome-scale models to predict biological capabilities. *Cell* 161, 971–987. doi: 10.1016/j.cell.2015.05.019

Pharkya, P., Burgard, A. P., and Maranas, C. D. (2003). Exploring the overproduction of amino acids using the bilevel optimization framework OptKnock. *Biotechnol. Bioeng.* 84, 887–899. doi: 10.1002/bit.10857

Pharkya, P., Burgard, A. P., and Maranas, C. D. (2004). OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* 14, 2367–2376. doi: 10.1101/gr.2872004

Pharkya, P., and Maranas, C. D. (2006). An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. *Metab. Eng.* 8, 1–13. doi: 10.1016/j.ymben.2005.08.003

Portela, C., Villas-Bôas, S., Rocha, I., and Ferreira, E. C. (2013). Genome scale metabolic network reconstruction of the pathogen Enterococcus faecalis. *IFAC Proc. Volumes* 46, 131–136. doi: 10.3182/20131216-3-in-2044.00067

Ranganathan, S., Suthers, P. F., and Maranas, C. D. (2010). OptForce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Comput. Biol.* 6:e1000744. doi: 10.1371/journal.pcbi.1000744

Rangel, A. E. T., Gómez Ramírez, J. M., and González Barrios, A. F. (2020). From industrial by-products to value-added compounds: the design of efficient microbial cell factories by coupling systems metabolic engineering and bioprocesses. *Biofuels Bioprod. Biorefin.* 14, 1228–1238. doi: 10.1002/bbb.2127

Ruckerbauer, D. E., Jungreuthmayer, C., and Zanghellini, J. (2015). Predicting genetic engineering targets with Elementary Flux Mode Analysis: a review of four current methods. *N. Biotechnol.* 32, 534–546. doi: 10.1016/j.nbt.2015.03.017

Schutte, K. M., Fisher, D. J., Burdick, M. D., Mehrad, B., Mathers, A. J., Mann, B. J., et al. (2015). *Escherichia coli* pyruvate dehydrogenase complex is an important component of CXCL10-mediated antimicrobial activity. *Infect. Immun.* 84, 320–328. doi: 10.1128/iai.00552-15

Segre, D., Vitkup, D., and Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15112–15117. doi: 10.1073/pnas.232349399

Senger, R., Yen, J., Tanniche, I., Fisher, A., Gillaspy, G., and Bevan, D. (2015). Designing metabolic engineering strategies with genome-scale metabolic flux modeling. *Adv. Genomics Genet.* 5:93. doi: 10.2147/agg.s58494

Shabestary, K., and Hudson, E. P. (2016). Computational metabolic engineering strategies for growth-coupled biofuel production by Synechocystis. *Metab. Eng. Commun.* 3, 216–226. doi: 10.1016/j.meteno.2016.07.003

Shams Yazdani, S., and Gonzalez, R. (2008). Engineering *Escherichia coli* for the efficient conversion of glycerol to ethanol and co-products. *Metab. Eng.* 10, 340–351. doi: 10.1016/j.ymben.2008.08.005

Tafur Rangel, A. E., Camelo Valera, L. C., Gómez Ramírez, J. M., and González Barrios, A. F. (2018). Effects of metabolic engineering on downstream processing operational cost and energy consumption: the case of *Escherichia coli*'s glycerol conversion to succinic acid. *J. Chem. Technol. Biotechnol.* 93, 2011–2020. doi: 10.1002/jctb.5432

Unden, G., and Bongaerts, J. (1997). Alternative respiratory pathways of *Escherichia coli*: energetics and transcriptional regulation in response to electron acceptors. *Biochim. Biophys. Acta - Bioenerg.* 1320, 217–234. doi: 10.1016/s0005-2728(97)00034-0

Varet, H., Brillet-Guéguen, L., Coppée, J. Y., and Dillies, M. A. (2016). SARTools: a DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data. Mills K, editor. *PLoS One* 11:e0157022. doi: 10.1371/journal.pone.0157022

Velur Selvamani, R. S., Telaar, M., Friehs, K., and Flaschel, E. (2014). Antibiotic-free segregational plasmid stabilization in *Escherichia coli* owing to the knockout of triosephosphate isomerase (tpiA). *Microb. Cell Fact.* 13:58. doi: 10.1186/1475-2859-13-58

Vlysidis, A., Binns, M., Webb, C., and Theodoropoulos, C. (2011). A techno-economic analysis of biodiesel biorefineries: assessment of integrated designs for the co-production of fuels and chemicals. *Energy* 36, 4671–4683. doi: 10.1016/j.energy.2011.04.046

Wang, Y., Manow, R., Finan, C., Wang, J., Garza, E., and Zhou, S. (2011). Adaptive evolution of nontransgenic *Escherichia coli* KC01 for improved ethanol tolerance and homoethanol fermentation from xylose. *J. Ind. Microbiol. Biotechnol.* 38, 1371–1377. doi: 10.1007/s10295-010-0920-5

Wong, K. K., and Kwan, H. S. (1992). Transcription of glpT of *Escherichia coli* K12 is regulated by anaerobiosis and fnr. *FEMS Microbiol. Lett.* 94, 15–18.

Woo, H. M., and Park, J. B. (2014). Recent progress in development of synthetic biology platforms and metabolic engineering of Corynebacterium glutamicum. *J. Biotechnol.* 180, 43–51. doi: 10.1016/j.jbiotec.2014.03.003

Yin, X., Li, J., Shin, H. D., Du, G., Liu, L., and Chen, J. (2015). Metabolic engineering in the biotechnological production of organic acids in the tricarboxylic acid cycle of microorganisms: advances and prospects. *Biotechnol. Adv.* 33, 830–841. doi: 10.1016/j.biotechadv.2015.04.006

Zhang, B., Skory, C. D., and Yang, S. T. (2012). Metabolic engineering of Rhizopus oryzae: effects of overexpressing pyc and pepc genes on fumaric acid biosynthesis from glucose. *Metab. Eng.* 14, 512–520. doi: 10.1016/j.ymben.2012.07.001

Zhang, J., Wu, C., Du, G., and Chen, J. (2012). Enhanced acid tolerance in Lactobacillus casei by adaptive evolution and compared stress response during acid stress. *Biotechnol. Bioprocess Eng.* 17, 283–289.

Zhang, X., Shanmugam, K. T., and Ingram, L. O. (2010). Fermentation of glycerol to succinate by metabolically engineered strains of *escherichia coli*. *Appl. Environ. Microbiol.* 76, 2397–2401.

Zhu, Q., and Jackson, E. N. (2015). Metabolic engineering of Yarrowia lipolytica for industrial applications. *Curr. Opin. Biotechnol.* 36, 65–72. doi: 10.1016/j.copbio.2015.08.010

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

REVIEW
published: 06 April 2021
doi: 10.3389/fgene.2021.637141

Understanding Omics Driven Plant Improvement and *de novo* Crop Domestication: Some Examples

Rakesh Kumar[1]†, Vinay Sharma[2†], Srinivas Suresh[1], Devade Pandurang Ramrao[1], Akash Veershetty[1], Sharan Kumar[1], Kagolla Priscilla[1], BhagyaShree Hangargi[1], Rahul Narasanna[1], Manish Kumar Pandey[2], Gajanana Ramachandra Naik[1], Sherinmol Thomas[3] and Anirudh Kumar[4]**

[1] Department of Life Science, Central University of Karnataka, Kalaburagi, India, [2] International Crops Research Institute for the Semi-Arid Tropics, Hyderabad, India, [3] Department of Biosciences & Bioengineering, Indian Institute of Technology Bombay, Mumbai, India, [4] Department of Botany, Indira Gandhi National Tribal University, Amarkantak, India

OPEN ACCESS

Edited by:
Fatemeh Maghuly,
University of Natural Resources and Life Sciences, Vienna, Austria

Reviewed by:
Uday Chand Jha,
Indian Institute of Pulses Research (ICAR), India
Atsushi Fukushima,
RIKEN, Japan

***Correspondence:**
Rakesh Kumar
rakeshkumar@cuk.ac.in
Anirudh Kumar
anirudh.kumar@igntu.ac.in

† These authors have contributed equally to this work

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

Received: 02 December 2020
Accepted: 02 March 2021
Published: 06 April 2021

Citation:
Kumar R, Sharma V, Suresh S, Ramrao DP, Veershetty A, Kumar S, Priscilla K, Hangargi B, Narasanna R, Pandey MK, Naik GR, Thomas S and Kumar A (2021) Understanding Omics Driven Plant Improvement and de novo Crop Domestication: Some Examples. Front. Genet. 12:637141. doi: 10.3389/fgene.2021.637141

In the current era, one of biggest challenges is to shorten the breeding cycle for rapid generation of a new crop variety having high yield capacity, disease resistance, high nutrient content, etc. Advances in the "-omics" technology have revolutionized the discovery of genes and bio-molecules with remarkable precision, resulting in significant development of plant-focused metabolic databases and resources. Metabolomics has been widely used in several model plants and crop species to examine metabolic drift and changes in metabolic composition during various developmental stages and in response to stimuli. Over the last few decades, these efforts have resulted in a significantly improved understanding of the metabolic pathways of plants through identification of several unknown intermediates. This has assisted in developing several new metabolically engineered important crops with desirable agronomic traits, and has facilitated the *de novo* domestication of new crops for sustainable agriculture and food security. In this review, we discuss how "omics" technologies, particularly metabolomics, has enhanced our understanding of important traits and allowed speedy domestication of novel crop plants.

Keywords: omics, metabolomics, *de novo* domestication, crop improvement, domesticated-genes

INTRODUCTION

The process of crop domestication began approximately 12,000 years ago, and was an important milestone during human civilization and led the foundation of modern agriculture. In the 21st century, most of the cultivated crops are domesticated from their wild ancestral progenitors (Meyer et al., 2012). During the domestication process plants were selected based on visible traits guided by needs of the time which led to the selection of only few desired alleles and dilution of the genetic variation present within the crop (**Figure 1**). For example, in cereals like wheat and rice, traits such as increase in the number of seeds per plant, uniform seed maturation, and non-shattering of seeds were preferred over the size of kernels during early domestication (Si et al., 2016). However, the selection of such traits varies greatly from plant to plant or between crops. For instance, in fleshy fruits or berries such as tomato, eggplant and apple, the size of the fruits and berries were preferred over overall yield (Zhu et al., 2018). Likewise, in tuber producing plants such as potato

the tuber size is one of the preferred traits (Fernie and Yan, 2019). Surprisingly, cultivated plant species represent only 250 fully domesticated species among 2500 species, which have undergone the process of domestication, and represent 160 plant families (Smýkal et al., 2018). This proportion is even starker considering the total plant diversity available for the cultivation or those, which are already being cultivated in different places (semi-cultivated species). For example, around 400,000 semi-cultivated plant species are currently known which can be utilized for designing future crops (Smýkal et al., 2018; Fernie and Yan, 2019).

The process of domestication of a species is a very slow and steady process. In fact, the modern cultivars available were generated following a long series of events: (a) Neolithic Revolution, (b) Columbian Exchange, (c) Industrial Revolution, (d) Green Revolution, and (e) Genomic Revolutions (Smýkal et al., 2018). Presently, to feed an ever-growing global population in the face of climate change is challenge for agriculture especially due to reduction of the arable lands due consistent conversion of lands into semi-arid areas and nutrient deficient land along with increase in soil pH or salinity. Therefore, a more rapid method of developing elite climate smart cultivars with desired traits is required. This could be achieved through integrated OMICS approaches, which include genomics, transcriptomics, proteomics, metabolomics and phenomics integrated with bioinformatics analyses (Kumar et al., 2017, 2018; Sharma et al., 2021). Plant OMICs based research have played very important role in deciphering metabolic pathways and their molecular regulators, which govern key traits and several plant developmental processes (Kumar et al., 2017; Razzaq et al., 2019). In the past decade there has been a significant progress in the field of both sequencing (van Dijk et al., 2018; Kumar et al., 2020; Schmidt et al., 2020) and analytical methods for the detection of molecules (Ren et al., 2018; Gilmore et al., 2019; Macklin et al., 2020), which has not only improved detection throughput but also the accuracy and the sensitivity (Kumar et al., 2017; Chiang et al., 2018; Qi et al., 2019).

In the past, for the selection of traits breeding programs involved phenotypic selection of plants (which are guided by metabolic pathways) (Kiszonas and Morris, 2018). For instance, during the Green Revolution (from 1960 to 1980), development of semi-dwarf high yielding varieties of rice and wheat involved phenotypic selections of improved lines which actually involved selection of gibberellic acid pathway genes including the *GA20 oxidase* and *DELLA* protein encoding genes (Silverstone and Sun, 2000). In fact, most of the traits, which were targeted for the Green Revolution, are controlled by one or more metabolic pathways. Therefore, precise editing of these metabolic pathways can help in the development of varieties in a very short time (Rodríguez-Leal et al., 2017; Zhang Y. et al., 2018; Fernie and Yan, 2019). Previously, most of the reviews on plant omics have focused on the instrumentation involved and results obtained by different researchers (Kumar et al., 2017; Mangul et al., 2019; Misra et al., 2019; Tang and Aristilde, 2020). In this review, we represent how this omics knowledge can be utilized for development of improved cultivars by targeting metabolic pathways and also emphasize the use of this information for *de novo* domestication of wild ancestral species for sustainable food security.

ROLE OF OMICS DATA IN UNDERSTANDING PLANT TRAITS

Genomics plays an important role in the identification of quantitative trait loci (QTLs) and genes controlling important traits, particularly in domesticated crops (Fernie and Yan, 2019). Moving forward, great insights have been gleaned from genome sequencing and re-sequencing programs examining wild ancestral species of domesticated crops (Unamba et al., 2015). In plant genomics, Next Generation Sequencing (NGS) has played a very important role and provided opportunities in the field of functional genomics due to the availability of reference genomes for several model and crop plant species. These resources combined with high quality re-sequencing offers huge potential for discovery of causal genes and mechanisms associated with the key agronomic traits (Thudi et al., 2016; Chen et al., 2019; Varshney et al., 2019). Re-sequencing also enriched the availability of SNPs data and can be utilized for genomics-based studies such as GWAS (genome wide association study) and QTL-seq (Kumar et al., 2020), both of which are useful tools for trait mapping (Rivas et al., 2011; Zhu et al., 2011; Zhang et al., 2021). With the advent of these technologies combined with advances in metabolomics such as integration of GWAS with metabolomics efficient means for dissecting underlying molecular mechanisms involved in the growth and development are available (**Table 1**; Fang and Luo, 2019).

Sequencing and QTL-seq Based Trait Discovery

Presently, QTL-seq is one of the most successful approach developed for the gene discovery and trait dissection (Kumar et al., 2020; Pandey et al., 2020). This approach offers preliminary idea for positional cloning for linked genetic factors or genes, and it has excellent success in marker-assisted selection for crop improvement programs (Xu F. et al., 2015). With the advancements in NGS technologies new approaches like quantitative trait locus sequencing (QTL-seq) has been utilized for exploring rapid QTL or gene identification (Takagi et al., 2013). In QTL-seq approach, the extreme highest and lowest genotypes are selected from the mapping population for target traits, followed by mixing an equal amount of DNA from each bulk to build up two extreme bulk (High bulk and low bulk). Then, each bulk is sequenced and assembled and gene annotation is performed. This approach combined with Bulked segregant analysis, accompanied by whole genome re-sequencing technologies, is more effective and capable than the previous QTL detection methods (Takagi et al., 2013). Utilizing QTL-seq approach several QTLs and genes for different rice phenotypes (Takagi et al., 2013; Daware et al., 2016; Ogiso-Tanaka et al., 2017; Yang et al., 2017; Kadambari et al., 2018; Qin et al., 2018; Bommisetty et al., 2020; Nubankoh et al., 2020), soybean (Song et al., 2017; Zhang X. et al., 2018), chickpea (Singh et al., 2016; Deokar et al., 2019), tomato (Illa-Berenguer et al., 2015),

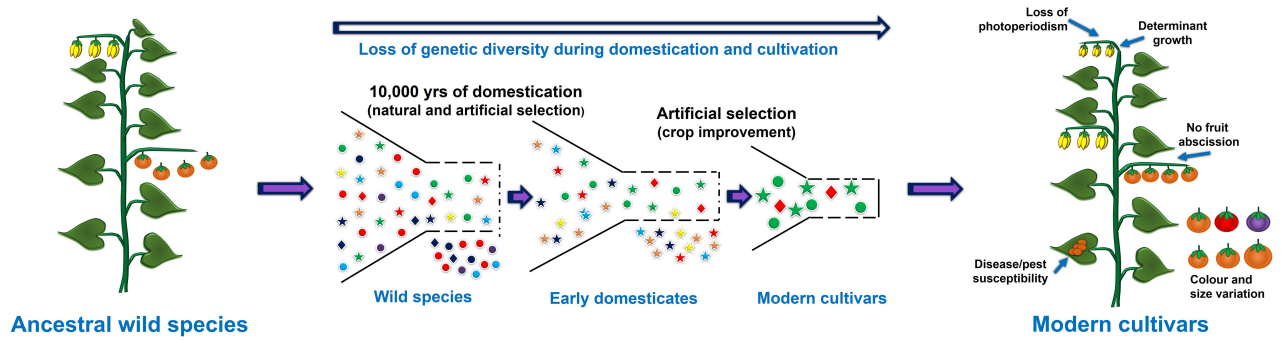

FIGURE 1 | Representation of domestication process and the loss of useful genetic variation due to selective breeding and selection of few alleles.

groundnut (Kumar et al., 2020; Luo et al., 2019; Zhao et al., 2020), have been effectively identified. This approach has also been deployed across in several crops due to its inherent ability to understand both qualitative and quantitative traits (**Table 2**). For instance, Kumar et al. (2020) identified the role of two genes *RING-H2 finger protein* and *zeaxanthin epoxidase* in fresh seed dormancy in groundnut; both genes are known to control abscisic acid (ABA) accumulation. Very recently, Ramos et al. (2020) identified three QTLs (genomic regions) *viz* QtlPC-C04, QtlPC-C11 and QtlPC-C14 (linked to genes *R1R2R3*) associated with resistance to *Phytophthora capsici* Leonian which causes crown rot in squash (*Cucurbita moschata*). The most significant benefit of whole genome sequencing is that it allows the identification of causative mutations in target chromosomal regions. Additionally, this method identifies molecular markers which are located inside the harboring chromosomal region that can also be used to narrow down the genomic region which will help in mining the trait linked genes.

RNA-seq Based Trait Discovery

Advances in RNA sequencing (RNA-seq) technologies and approaches have made significant impact toward trait discovery, and have enabled plant developmental studies characterizing expression patterns of all the functional genes as well as regulatory RNAs (Nayak et al., 2019). RNA-seq is a more robust approach for precise measurement of transcripts and has been widely used for transcript profiling in several plant species (Cloonan et al., 2008; Wang et al., 2009). This data is vital for improving genome annotations, and offers precise gene position information for functional characterization and genome editing. The RNA-seq approach has been deployed for molecular characterization of several important agronomic traits such as seed size (Wan et al., 2017), seed coat color (Wan et al., 2018), seed coat cracking (Wan et al., 2016), seed and bud dormancy (Qi et al., 2015; Zhu et al., 2015; Khalil-Ur-Rehman et al., 2017), fatty acid biosynthesis and oil quality (Nayak et al., 2019), nutritional quality traits (Reddy and Ulaganathan, 2015), etc., which can offer precise gene information for developing designer crops for future. Also, RNA-seq have been used to understand the molecular mechanisms associated with salt tolerance in rice (Zhou et al., 2016; Lei et al., 2020); Chinese rye grass (Sun et al., 2013), and maize (Liang and Schnable, 2016). In wheat, RNA-seq study reported the drought responsive molecular pathways along with key candidate genes and molecular markers in the root tissue (Iquebal et al., 2019). RNA-seq has also been shown to be highly useful in combination with other -omics methods for gene discovery and pathway investigations.

Proteomics Enabled Genetic Trait Understanding

Knowledge of proteomics is being used to map the translated genes and loci controlling the expression of respective genes. It helps in identification of proteins responsible for bringing intricate phenotypic variations. High throughput proteomics has gone beyond the identification of individual proteins, to quantitative profiling, post translational modification studies, signaling, protein–protein interaction and many more areas. Photosynthesis plays major role in biomass production and yield. Change in protein profile studies was performed in chlorophyll deficient *Brassica napus* leaves which established the relationship between chlorophyll biosynthesis and photosynthesis (Chu et al., 2015). Xylem sap proteomics has revealed several insights related to cell wall formation (Zhang M. et al., 2014), leaf senescence (Wang et al., 2012) biotic and abiotic stress response (Alvarez et al., 2008; González et al., 2012), cell to cell communication (Agrawal et al., 2010), and root–shoot communication (Krishnan et al., 2011). The enhanced level of redox proteins, stress and defense related proteins, calcium ion regulation proteins, signaling G-protein and RNA metabolism related proteins were induced in phloem sap study. Recently, proteomics study revealed that low light stress obstructs carbon fixation and *OsGAPB* overexpression augment tolerance to low light stress conceivably by increasing CO_2 assimilation and chlorophyll accumulation in rice (Liu et al., 2020). Interestingly, simultaneous upregulation of both biotic and abiotic stress responsive protein has been observed during bacterial blight infection in rice, which indicate the activation of common pathway (Kumar et al., 2015). Whereas in case of rice-*M. oryzae* interaction PBZ1, OsPR-10, SalT, Glu1, Glu2, and TLP proteins were up-regulated (Kim et al., 2004). iTRAQ proteomics study of *Oryza officinalis* provided evidences that proteins related to biosynthesis of secondary metabolites and carbon metabolism were mostly enriched after

TABLE 1 | List of selected studies involved mQTL and mGWAS approach.

Plant	Population/ accessions	Approach	Tissue	Study	Significant outcome	References
Apple (*Malus domestica*)	Prima × Fiesta	LC-MS	Fruit	mQTL	Identified 669 mQTLs, includes a major mQTL hotspot on LG16 contains gene *leucoanthocyanidin reductase* belong to the phenylpropanoid pathway.	Khan et al., 2012
Arabidopsis thaliana	Col-0 × C24 (RIL), ILs	GC-MS	Leaf	mQTL	Identified 385 mQTL for 136 metabolites	Lisec et al., 2009
	A. thaliana accessions	LC-MS	Leaf	mGWAS	Identification of 123 mQTL and 70 candidate genes	Wu et al., 2018
	314 natural accessions	GC-MS	Leaf	mGWAS	Identify two candidate genes (AT5G53120 and AT4G39660) involved in the β-alanine metabolic pathway	Wu et al., 2016
	Bay × Sha (RIL)	GC-MS	Leaf	mQTL	Identified 11 mQTL clusters linked to the plant central metabolism.	Rowe et al., 2008
	RILs and ILs	GC-MS	Seedling	mQTL	Identified 153 QTLs for augmented additive (Z1) and 83 QTL for dominance effects (Z2) in RIL	Lisec et al., 2009
	96 accessions	HPLC-DAD	Leaf	mGWAS	Identified two major QTLs controlling glucosinolate variation; and *AOP* and *MAM as* candidate genes	Chan et al., 2010
	313-ecotype association panel	LC-MS	Seed	mGWAS	Identified two significant associated genomic regions (One region is linked with serine-related trait and second region is linked with four histidine-related traits)	Angelovici et al., 2017
	Col-0 × C24	GC-MS	Seed	mQTL	Identified 786 mQTLs and candidate genes including *bZIP10* as regulator of seed metabolism	Knoch et al., 2017
Barley (*Hordeum vulgare*)	Diverse set of barley accessions	LC-MS	Flag leaf	mGWAS	Reported three mQTLs for metabolites (γ-tocopherol, glutathione, and succinate content) involved in antioxidative defense	Templer et al., 2017
	Maresi **x** CamB (RIL)	LC-MS	Leaf	mQTL	Identified 138 mQTLs for 98 traits. Annotation of mQTL identified genomic region with stress response related genes	Piasecka et al., 2017
	Qingke and barley accessions including wild	LC-MS	Leaf and Seed	mGWAS	Identified 90 significant mGWAS loci for variation of phenylpropanoid content	Zeng et al., 2020
Blueberry (*Cyanococcus*)	886 blueberry genotypes	GC-MS	Fruits	mGWAS	Identified 519 significant SNPs linked to 11 volatile organic compounds	Ferrão et al., 2020
Maize (*Zea mays* L.)	By804 × B73 (RIL)	GC-MS	Seedling, Leaf, Kernel	mQTL	Detected 297 QTL and candidate genes to the amino acid biosynthetic and catabolic pathways, tricarboxylic acid cycle and carbohydrate metabolism	Wen et al., 2015
	Inbred lines	GC-MS	Leaf	mGWAS	Identified 26 distinct metabolites strong associations with leaf complex trait such as dry mass, lignin composition etc.	Riedelsheimer et al., 2012
	Inbred lines	HPLC	Grain	mGWAS	Identified *ZmVTE4* haplotype and three new gene targets for increasing antioxidant and vitamin E levels. Also identified two additional genes, *ZmHGGT1* and one prephenate dehydratase parolog that modestly contribute to tocotrienol variation	Lipka et al., 2013
	Inbred lines	UP-LC	Kernel	mGWAS	Identified 74 loci functionally associated with kernel oil content and fatty acid composition; Also identified genes involved in oil biosynthesis (*DGAT1-2*, *FATB* and *FAD2*), members of the oil metabolic pathway (*FAD2*, *LCACS*, *ACP*, and *COPII*) and one transcription factor (*WRI1a*)	Li et al., 2013
	Inbred lines	HPLC	Kernel	mGWAS	Nine carotenoid compounds measured in grain samples, the most abundant was zeaxanthin; Identified 58 candidate genes involved in biosynthesis and retention of carotenoids in maize.	Owens et al., 2014
	Inbred lines and RIL population	LC-MS	Mature Kernel	mGWAS	Identified 1,459 significant locus–trait associations across three environments through metabolite-based genome-wide association mapping, identified potential causal variants for five candidate genes involved in metabolic traits	Wen et al., 2014

(Continued)

TABLE 1 | Continued

Plant	Population/ accessions	Approach	Tissue	Study	Significant outcome	References
	Inbred diversity panel	LC-MS	Kernel	mGWAS	Identified 19 modules which shows significant associations with genetic control of biochemical networks within the kernel.	Shen et al., 2013
	513 diverse inbred lines association panel	GC-MS	Seedling, Leaf, Kernel	mGWAS	Identified 153 significant loci linked to primary metabolism	Wen et al., 2018
Potato (*Solanum tuberosum*)	Diversity panel	LC-MS	Tuber	mGWAS	Identified 472 features in which significant SNPs have been associated to glycoalkaloids (α-chaconine, β-chaconine, and α-solamarine) reported on chromosomes 2, 7, and 8	Levina et al., 2020
	C (*S. phureja* × S. tuberosum) × E (*S. vernei* × *S. tuberosum*)	GC-MS	Tuber	mQTL	Identified 87 mQTLs associated to primary metabolism	Carreno-Quintero et al., 2012
Rapeseed (*Brassica napus*)	EXPRESS × SWU07 (DH)	NIRS	Seed	mQTL	Identified four QTLs for Glucosinolates content between	He et al., 2018
	Tapidor × Ningyou7 (DH)	HPLC	Leaf and Seed	mQTL	105 mQTLs related to glucosinolate biosynthesis in rapeseed seed and leaves have been observed	Feng et al., 2012
Rice (*Oryza sativa*)	ZS97 × MH63 (RIL)	LC-MS	Flag leaf, germinating Seed	mQTL	Identified 1,884 mQTLs in flag leaf and 937 mQTLs in germinating seed samples	Gong et al., 2013
	Sasanishiki × Habatak (BIL)	GC-MS, LC-MS, CE-MS	Seed	mQTL	Identified 802 mQTLs for 759 metabolic traits; including mQTL hotspot on chromosome 3 regulating amino acids content	Matsuda et al., 2012
	Landraces accessions and subpopulations rice subspecies *indica* and *japonica*	LC-MS	Five-leaf stage	mGWAS	Identified 36 candidate genes controlling metabolites level and nutritional composition	Chen et al., 2014
	Landraces accessions	LC-MS	Leaf/ seedling	mGWAS	Identified 323 associations, demonstrating that phytochemicals produced in rice cultivars are diverse	Matsuda et al., 2015
	Landraces and elite varieties of indica and japonica	LC-MS	Grains	mGWAS	More than 30 candidate genes were identified, associated with metabolic and/or morphological traits.	Chen et al., 2016
	156 Landrace	LC-MS	Leaf/root and other tissue parts of rice	mGWAS	Identified two *spermidine hydroxyl-cinnamoyltransferases* (Os12g27220 and Os12g27254) that might underlie the natural variation levels of spermidine conjugates in rice	Dong and Wang, 2015
	ZS97 × MH63 (RIL)	LC-MS	Leaf and Seed	mQTL	Provided over 2,800 highly resolved metabolic quantitative trait loci for 900 metabolites; associated 24 candidate genes to various metabolic quantitative trait loci by data mining, including ones regulating important morphological traits and bio-logical processes	Gong et al., 2013
	Three CSSL populations (N/Z, M/Z, and A/Z)	LC-MS	Flag leaf	mQTL	Identified 1,587 mQTL, of which 684 in (A/Z), 479 in (M/Z), and 722 in(N/Z) have been detected among three CSSL population	Chen et al., 2018
	Lemont × Teqing (RIL)	GC-MS	Leaf	mQTL	Identified two mQTL hotspots which have opposing effects on carbon and nitrogen rich metabolites, and regulate carbon and nitrogen partitioning.	Li et al., 2016
Strawberry (*Fragaria* × *ananassa*)	*F.* x *ananassa* 232 × 1392 (F1)	LC-MS	Fruit	mQTL	Reported 309 mQTLs for 77 polar secondary metabolites.	Pott et al., 2020
	232 × 1392 (F1)	GC-MS	Fruit	mQTL	Reported 133 unique mQTLs for 44 traits with PVE% range from 9.6% to 46.1%. RNA seq analysis identified candidate gene *Mannose-6-phosphate isomerase* responsible for natural variation in L-ascorbic acid in fruit	Vallarino et al., 2019
Tomato (*Solanum lycopersicum*)	Introgression lines	LC-MS	Fruit	mQTL	Detected 216 canalization metabolite quantitative trait loci (cmQTLs) for secondary metabolites and 93 cmQTLfor primary metabolites.	Alseekh et al., 2017

(Continued)

TABLE 1 | Continued

Plant	Population/ accessions	Approach	Tissue	Study	Significant outcome	References
	Introgression lines	UPLC	Fruit	mQTL	Identified 679 mQTLs for primary metabolites and secondary metabolites	Alseekh et al., 2015
	Introgression lines	GC-MS	Seed	mQTL	Identified 46 mQTLs in IL population and propose post transcriptional regulation	Toubiana et al., 2012
	Tomato accessions including wild	GC-MS	Fruit	mGWAS	Identified a total 44 loci associated with 19 traits, including sucrose, ascorbate, malate and citrate levels.	Sauvage et al., 2014
	Tomato accessions including wild	GC-MS	Fruit	mGWAS	Identified 388 suggestive association loci (including 126 significant loci) for 92 metabolic traits including nutrition and flavor-related loci by genome-wide association study	Ye et al., 2019
	IL12-3 × M82	LC-MS	Fruit and leaf	mQTL	Reported 1528 mQTLs in fruit and 428 mQTL in leaf; Major mQTL involved in the regulation of diacylglycerols and triacylglycerols have been detected on chromosome 12	Garbowicz et al., 2018
	76 ILs + recurrent parent M82	LC-MS	Seed	mQTL	Identified 338 mQTL for flavonoids, steroidal glycoalkaloids, and specialized metabolites content	Alseekh et al., 2020
	IL4-4 × M82	GC-MS, HPLC, LC-MS	Fruit	mQTL	Identified 72 mQTL, where major mQTLs linked to twenty genes which have a broad effect on several metabolic pathways.	Liu et al., 2016
	ILs	GC-MS	Fruit	mQTL	Reported 889 fruit traits related mQTLs and 326 yield-related traits mQTLs	Schauer et al., 2006
	IL and heterozygous ILH	GC-MS	Fruit	mQTL	Identified 332 putative mQTL associated with metabolite accumulation (174 mQTLs is dominantly inherited, 61 mQTLs is additively inherited and 80 mQTLs is recessively inherited and negligible number of mQTL showing the feature of over dominant inheritance)	Schauer et al., 2008
	S. lycopersicum (M82) × *S. pennellii* Ils	GC-MS, LC-MS, HPLC-PDA, NMR	Fruit	mQTL	Detected mQTL for carotenoids and tocopherols	Perez-Fons et al., 2014
Wheat (*Triticum aestivum*)	KN9204 × J411 (RIL)	LC-MS	Kernel	mQTL	Identified 1005 mQTLs and 24 genes candidate gene related to grain related traits	Shi et al., 2020
	Excalibur × Kukri (DH)	LC-MS	Flag leaf	mQTL	Identified mQTLs for 238 metabolites across 159 intervals on genetic map; two mQTLs on chromosome 7A coordinating the genetic control of various metabolites	Hill et al., 2015
	Winter elite lines (135)	GC-MS, LC-MS	Flag leaf	mGWAS	Identified significant associations 17 SNPs with six metabolic traits, namely oxalic acid, ornithine, L-arginine, pentose alcohol III, L-tyrosine, and a sugar oligomer (oligo II)	Matros et al., 2017
	Natural accessions	LC-MS	Mature seeds	mGWAS	A total of 1098 mGWAS associations were detected with large effects, within which 26 candidate genes for flavonoid decoration pathway	Chen et al., 2020
	Doubled haploid lines	GC-MS	Flag leaf	mQTL	Identified 112 mQTLs for 95 metabolites, of which 43 are known compounds	Hill et al., 2013

planthopper infestation (Zhang et al., 2019c). Several proteomics and transcriptomics study conducted on seed dormancy study revealed the important role of antioxidant mechanism, protein thiol and redox regulation along with hormonal signaling in rice, wheat and barley (Hu et al., 2015). Mass spectrometry (MS) based proteomics study demonstrated the cultivar specific induction of proteins in salt stress condition such as glutathione-based detoxification of ROS was highly induced in tolerant variety whereas proteins involved in iron uptakes were more expressed in salt sensitive variety in barley (Witzel et al., 2009). Similarly, the role of *OsCYP2* in moderating the antioxidant enzymes was established in transgenic rice overexpressing cyclophilin during salt stress (Ruan et al., 2011). Seed proteomics of various developmental stages during different stresses have revealed the process involved in seed dormancy, seed germination, and seed development (Finnie et al., 2011). Proteomics related to environmental changes and abiotic stress focused on water supply responsive proteins in wheat against drought, high temperature, low temperature, frost, salt and heavy metals have been carried out (Yang et al., 2011; Han et al., 2013; Kosová et al., 2013;

TABLE 2 | List of important QTL-seq studies in crop plants.

Crop	Population	Target Trait	QTL/Gene mapped	References
Oryza sativa	IR 64 × Sonasal	Grain Weight	Two genes LOC_Os05g15880 (glycosyl hydrolase) and LOC_Os05g18604 (serine carboxypeptidase)	Daware et al., 2016
	Nipponbare × BIL-55	Late heading under short-day conditions	Zinc finger B-box domain containing protein (Os04t0540200-01), WD40-repeat-domain–containing proteins (Os04t0555500-01, Os04t0555600-01, Os04t0564700-01), flowering locus D (Os04t0560300-01), CCAAT-binding-domain–containing protein (Os06t0498450-00)	Ogiso-Tanaka et al., 2017
	H12-29 × FH212	Grain Length and Weight	*qTGW5.3* (1.13 Mb)	Yaobin et al., 2018
	LND384 × INRC10192	Plant height	*asd1* (67.51 Kb)	Kadambari et al., 2018
	M9962 × Sinlek	Spikelet fertility	*qSF1, qSF2, and qSF3 (LOC_Os01g59420, LOC_Os02g31910, LOC_Os02g32080, LOC_Os03g50730)*	Nubankoh et al., 2020
	BPT5204 × MTU3626	Grain weight	*qGW8* (LOC_Os08g01490 (Cytochrome P450), and LOC_Os08g01680 (WD domain, G-beta repeat domain containing protein)	Bommisetty et al., 2020
Triticum aestivum	GY448 × GY115	Awnless trait	*Qal.nwipb-5AL* (25-bp indel in *B1* gene promoter region)	Wang et al., 2021
Zea mays	CMS-C lines × A619	Fertility Restoration	*qRf8-1* (17.93-Mb)	Zheng et al., 2020
Brassica napus	Huyou19 × Purler	Branch angle	Branch angle 1 (BnaA0639380D, a homolog of AtYUCCA6)	Wang et al., 2016
	Cabriolet × Darmor	Vernalization	FLOWERING LOCUS C (*BnaFLC.A02*) and FLOWERING LOCUS T (*BnaFT.A02*)	Tudor et al., 2020
Brassica rapa	Zicaitai × Caixin	Purple Trait	*BrMYBL2.1* gene	Zhang X. et al., 2020
Glycine max	Zhonghuang × Jiyu 102	Seed cotyledon color	qCC1 (30.7-kb) and qCC2 (67.7-kb)	Song et al., 2017
	CSSL3228 × NN1138–2	Plant height	Glyma.13 g249400 candidate gene	Zhang X. et al., 2018
Arachis hypogaea	ZH8 × ZH9	Testa color	*AhTc1*, encoding a R2R3-MYB transcription factor	Zhao et al., 2020
	ICGV 00350 × ICGV 97045	Fresh seed dormancy	RING-H2 finger protein and zeaxanthin epoxidase	Kumar et al., 2020
	Yuanza 9102 × Xuzhou 68-4	Shelling percentage	Nine candidate genes in 10 SNPs	Luo et al., 2019
Cicer arietinum	ICC 4958 × ICC 1882	100-seed weight	Two genes *Ca_0436* and *Ca_04607*	Singh et al., 2016
	ICCV 96029 × CDC Frontier and ICCV 96029 × Amit	Ascochyta blight	Six candidate genes on chromosomes Ca2 and Ca4	Deokar et al., 2019
Solanum lycopersicum	Three populations (12S139, 12S143 and 12S75)	Fruit weight and locule number	Three fruit weight (*fw1.1*, *fw3.3*, *fw11.2*) and one locule number (*lcn6.1*) QTLs	Illa-Berenguer et al., 2015
Cucumis melo	MR-1 × M1-32	Stigma Color	GS8.1 (268 kb) MELO3C003149, MELO3C003158, and MELO3C003165	Qiao et al., 2021
Cucumis sativus	PM-R × PM-S	Powdery mildew resistance	*Two QTLs pm5.2* and *pm6.1* (*CsGy5G015660*)	Zhang et al., 2021

Alvarez et al., 2014; Capriotti et al., 2014; Kang et al., 2015). These studies offered novel insights and better understanding of crop physiology and metabolism during various kinds of stresses.

Metabolomics Based Trait Understanding

Holistic metabolomics based to study trails in plants were started late, particularly this approach was started through the introduction of untargeted metabolome detection (Alonso et al., 2015). Several studies have been reported where a particular metabolic pathways have been mapped (Sharma et al., 2021). For instance, the substantial changes in the various phytohormones was investigated in poplar leaf (Novák et al., 2008), rice various aerial organs (Kojima et al., 2009), rosemary leaves et al. (Müller and Munné-Bosch, 2011), Arabidopsis developing seeds (Kanno et al., 2010), strawberry fruits (Gu et al., 2019), potato tuber (Peivastegan et al., 2019), wheat developing seeds (Matsuura et al., 2019), watermelon fruit (Kojima et al., 2021), etc. The targeted approach has been also applied to explore the carotenoid pathway (Fernandez-Orozco et al., 2013; Kim et al., 2016; Mibei et al., 2017; Yoo et al., 2017; Price et al., 2018; Di Lena et al., 2019), flavonoid pathways (Karimi et al., 2011; Dong X. et al., 2014; Torres et al., 2019), amino acids (Torres et al., 2019; Praveen et al., 2020), and fatty acids (Talebi et al., 2013; Vidigal et al., 2018). Such profiling studies has helped in improving several important traits in plants by targeting specific pathways. Almost 10 years back Liu et al. (2011) targeted fatty acids biosynthesis pathways for enhancing biofuel production. Very recently and *fatty acid desaturase 2* was targeted in several crops such as canola (Okuzaki et al., 2018), peanut (Yuan et al., 2019), rice (Abe et al., 2018), false flax (Morineau et al., 2017), and Soybean (Wu et al., 2020), for enhanced production of oleic acid (C18:1), respectively.

Several un-targeted metabolomics approach has been utilized to target multiple class of metabolites (amines, sugars, organic acids, etc.) from a sample extracted from various tissues of the model and crop plants such as Arabidopsis, apple, groundnut, kiwi fruit, alpine bird's-foot-trefoil, strawberry, grapes, mango,

maize, medicago, orange, pear, sunflower, soybean, tomato, rice, white lupin, etc. (Sharma et al., 2021). Now, the targeted and un-targeted metabolomics approach have been coupled with genomics data for carrying out metabolomics-based quantitative trait locus (mQTL) and metabolic genome-wide association studies (mGWAS) studies (Wen et al., 2015; Chen et al., 2016); which simultaneously identifies the genomic region, causal genes and key metabolites and associated metabolic pathways that govern particular trait in plants. Recently, Li K. et al. (2019) identified 65 primary metabolites *viz* 22 amino acids, 21 organic acids, 12 sugars, four amines and six miscellaneous metabolites in the leaf of teosinte (an ancestor of maize) and identifies advantageous genes present in the wild relative associated with grain yield and shape trait in maize. In tomato, for one of the important trait accumulation of secondary metabolite in fruit was analyzed, and reported several subset of mQTLs- including mQTLs associated with acyl-sugar, hydroxycinnamates, naringenin chalcone, and a range of glycoalkaloids (Alseekh et al., 2015). Likewise, there are several studies which identified key genomic regions, candidate genes and mQTLs related to important traits through mQTL and mGWAS based studies including some domesticated traits, this was extensively reviewed by Sharma et al. (2021).

Previously, a combined transcriptome, proteome and metabolomics approach was used to investigate the ripening process with a final aim of extending tomato fruit shelf life (Osorio et al., 2011). This study showed a strong relationship between metabolites and their associated transcripts controlling ripening such as sugars, organic acids, and cell wall metabolism pathways. Similar studies have been done for banana which led to identification of genes including *ERF1B*, *fructose-1,6-bisphosphatase* and *polygalacturonase* as key regulators of pulp ripening (Li T. et al., 2019). Recently, a combined transcriptome and metabolome study was deployed to study the molecular aspects of resistance and the interaction between *Trichoderma harzianum* strain T22 with tomato during defense responses against aphids (Coppola et al., 2019). This study demonstrated the importance of plant transcription factor families such as ZIP, MYB, NAC, AP2-ERF, and WRKY in biotic stress resistance. These examples show the potential of the -omics studies, working in tandem to characterize complex molecular interactions. These data have been used for the development of several gene expression/proteome/metabolome atlases to facilitate omics-driven crop improvement (**Table 3**).

MOLECULAR REGULATIONS OF DOMESTICATION RELATED TRAITS: SELECTED EXAMPLES

Over the past two decades the molecular regulation and the associated metabolic pathways of several agronomic traits has been revealed because of intensive research and the deployment of omics tools (**Table 4**). For the major domesticated traits their associated genes pathways have been linked with metabolic networks; however, more focused research is required to understand their specific role in particular metabolic pathways. Here, we review progress in omics-based investigations of several important domestications related traits.

Transcriptional Control for Loss of Seed Shattering Trait in Cereal

From an evolutionary viewpoint, natural selection allows wild plant species to have specific functions to disperse seeds and fruits. *Although from the agronomic viewpoint, natural seed dispersal is an undesirable trait in crops as it leads to significant seed loss in harvest. Consequently, natural seed dispersal was strongly chosen against by ancient humans to ensure productive cultivation during the domestication period (*Purugganan and Fuller, 2009; Lenser and Theißen, 2013). The non-shattering traits were considered as the landmark of domestication in seed crops, as it makes the domesticated species mostly rely on human activity for propagation and enables the fixation of other domestication traits (Purugganan and Fuller, 2009). Seed crops have established their reduction of seed shattering ability independently and it is a convergent morphological adaptation to artificial selection (Purugganan and Fuller, 2009; Olsen and Wendel, 2013).

In cereal, seed shattering or fruit dehiscence is enacted through an abscission layer in the lemma-pedicel joint. Various transcription factors (TFs) coding genes were found in rice (*Oryza sativa*), which are involved in decreasing seed shattering. *Shattering4* (*Sh4*) encodes the TF with Myb3 homology and is important for the formation of a functional abscission layer in the pedicle (Li et al., 2006). *A single change of amino acid in DNA binding domain of Sh4 is intimately linked to the reduced seed shattering in domesticated rice. Also, the expression of the domesticated allele has been substantially reduced compared to the wild allele* (Li et al., 2006). *Thus, the combination of coding and regulatory alteration of* Sh4 *seems to affect the formation of the abscission layer, and consequently tries to weaken the shattering phenotype* (Li et al., 2006). *qSH1* is a major QTL on chromosome 1 involved in seed shattering in rice. The main gene, *qSH1*, codes a homeobox transcription factor-like BEL1 which is homologous to AtRPL (Konishi et al., 2006). A single nucleotide polymorphism (SNP) in the 5′-regulatory region effectively nullifies *qSH1* expression in the preliminary abscission layer in the early development stage and contributes to non-shattering traits of rice (Konishi et al., 2006). Interestingly, the regulatory SNP in the homologs of RPL promoter are also amenable for distinct structures of seed dispersal based on natural selection of Brassica species with diminished replum development (Arnaud et al., 2011). These studies show a notable convergent mechanism where the same regulatory SNP could describe developmental variations in seed dispersal structures, which are important for both domestication and natural selection in distant species (Arnaud et al., 2011; Gasser and Simon, 2011). *SH5* is another homeobox type *BEL1* gene with a high *qSH1* homology. SH5 has been expressed in the abscission layer (Yoon et al., 2014). Knockout of *SH5* inhibits abscission layer formation and prevents seed shattering. Over-expression of *SH5* leads to higher seed shattering, a consequence of decreased pedicel lignin levels (Yoon et al., 2014). The regulatory pathway of abscission layer formation has recently been expanded to include *Shattering abortion 1*

TABLE 3 | List of gene-expression, proteome and metabolome atlas developed in plant.

Plant name	Scientific name	Tissue/cell type	Gene/Proteins/ Metabolites	Citations	DOI
Gene expression atlas			*Genes*		
Chickpea	*Cicer arietinum*	27	15,947	Kudapa et al., 2018	10.1111/pce.13210
Peanut	*Arachis hypogaea*	19	NA	Sinha et al., 2020	10.1111/pbi.13374
Soybean	*Glycine max*	14	66210	Libault et al., 2010 Severin et al., 2010	10.1111/j.1365-313X.2010.04222.x 10.1186/1471-2229-10-160
Wheat	*Triticum aestivum*	32	94,114	International Wheat Genome Sequencing Consortium (IWGSC)	10.1126/science.aar7191
Rice	*Oryza sativa*	40	∼30,000	Jiao et al., 2009	10.1038/ng.282
Maize	*Zea may*	11	22,151	Sekhon et al., 2013	10.1371/journal.pone.0061005
Bryophyte	*Physcomitrella patens*	10	∼32500	Ortiz-Ramírez et al., 2016	10.1016/j.molp.2015.12.002
Proteome atlas			*Proteins*		
Arabidopsis	*Arabidopsis thaliana*	9	13,029	Baerenfaller et al., 2008	10.1126/science.1157956
Rice	*Oryza sativa*	3	2,528	Koller et al., 2002	10.1073/pnas.172183199
Wheat	*Triticum aestivum*	24	46,016	Duncan et al., 2017	10.1111/tpj.13402
Metabolome atlas					
Arabidopsis	*Arabidopsis thaliana*			Wu et al., 2018	10.1016/j.molp.2017.08.012

(*SHAT1*), an AP2 transcription factor encoding gene (Zhou et al., 2012). *SHAT1 is needed for seed shattering by specifying abscission layer.* Sh4 *positively regulates the SHAT 1 expression in the abscission layer.* qSH1 *expression is lost in abscission layer in both the shat1 and sh4 mutant background, indicating qSH1 acts downstream of the shat1 and sh4 in the abscission layer establishment* (Zhou et al., 2012). *Intriguingly,* qSH 1 *is also needed in the abscission layer for expression of* SH1 *and* Sh4. *Thus the* qSH 1 *possibly takes part in a positive feedback loop of SH1 and Sh4 by establishing the* SHAT1 *and* Sh4 *expression in the abscission layer* (Zhou et al., 2012). *While* SH5 *and* SHAT1 *play a role in differentiating the abscission layer, the question remains whether both are artificially selected domestication genes. Like rice, decrease of seed shattering in domesticated sorghum is a result of loss of abscission in the joint that connects the seed hull with the pedicel. In sorghum, seed shattering is regulated by a single gene,* Shattering1 *(*Sh1*), which encodes a transcription factor YABBY. The non-shattering trait can be accounted for by any one of the three different loss-of-function mutations selected independently during sorghum domestication process* (Lin et al., 2012). *The notable mutations in* Sh1 *orthologs in rice and maize may be related to the shattering decrease in these crops (*Lin et al., 2012). Whether *Sh1* has been rewired into an SH5-directed seed shattering network in rice remains to be investigated in the future. In a wild relative of sorghum (*Sorghum propinquum*), seed shattering is conferred by the *SpWRKY* gene. It is believed that *SpWRKY* controls cell wall biosynthesis genes negatively in the abscission layer. Even so, *SpWRKY* was not crafted by artificial selection to contribute to the non-shattering characteristic for domesticated sorghum (Tang et al., 2013). These above studies together have raised a fascinating potential that the convergent domestication of non-shattering crops may have achieved the same underlying genetic goals by parallel selection (Lenser and Theißen, 2013).

In domesticated wheat (*Triticum aestivum*) free-threshing trait (loss of spike shattering tendency) is conferred by important *Q* gene (Simons et al., 2006). Q-*gene encodes the AP2-family transcription factor. The domesticated Q allele is abundantly transcribed than the wild q allele. Besides, both alleles differ in single amino acid, which significantly improves the homo-dimerization ability of the cultivated allele* (Simons et al., 2006). *Similar to Sh4, the development of the free-threshing character in cultivated wheat might also have been due to the combination of the coding and regulatory changes in the cultivated gene. The difference of expression between Q and q seems more significant as it can clarify the free threshing character in cultivated wheat* (Simons et al., 2006; Zhang et al., 2011). Even though mutation which gives rise to Q has a significant effect on the process of wheat domestication, as it helps farmers to harvest the grain more effectively, the exact cellular cause contributing to free-threshing character is still unclear. Similar research has been progressed in non-cereals crop such as overexpression *AtFUL* to make the pods shattering resistance in *Brassica juncea* (Østergaard et al., 2006).

Cross-Talk Between Phytohormones and Related Genes Regulating Seed Shattering and Dehiscence Zones (DZ)

Hormonal homeostasis and interactions have been found recently as direct downstream effects of the core genetic network. As an example *indehiscent (IND)* expression is involved in the formation of local auxin minimum at the margin of the valve by regulating the auxin efflux in the separation layer cells (Sorefan et al., 2009). Further findings reveal that another b-HLH class SPATULA (SPT) transcription factor, required for carpel fusion early in the female reproductive organ development, may interact physically with IND (Girin et al., 2011). Auxins and cytokinins play an antagonistic role in plant growth and development (Bishopp et al., 2011). This scenario also indicates that the cytokinin signaling pathway is active at the valve margins and such a signaling pathway is interrupted in the shp1/2 and *ind* mutant. Consequently, local application of cytokinins in the fruit development can help to restore valve

TABLE 4 | List of genes domesticated in the past and associated metabolic pathways.

Crops	Traits	Domesticated Genes	Involvement in the metabolic pathways	References
Rice	Plant architecture	*sd1*	Encodes gibberellin 20-oxidase (Gibberellin pathway gene)	Spielmeyer et al., 2002
	Seed shattering	*sh4*	Abscisic acid response elements (ABREs) have been identified which is involved in ABA hormone signal pathways	Yan et al., 2015
		qSH1	APETALA2-like transcription factor SUPERNUMERARY BRACT (SNP) positively regulates the expression of two rice genes, *qSH1* and *SH5* (SNB-involved regulating pathway)	Jiang et al., 2019
	Awn	*LABA1 / An-2*	*An-2* encodes a cytokinin synthesis enzyme that modulates awn length	Gu et al., 2015; Hua et al., 2015
		qAWNL2	N.A	Amarasinghe et al., 2020
	Seed and hull color	*Rc and Rd*	Involved in proanthocyanidin synthesis via the flavonoid pathway	Sweeney et al., 2006; Furukawa et al., 2007
	Seed dormancy	*Sdr4*	Zinc finger protein, *OsVP1* activates *Sdr4* expression to control rice seed dormancy via the ABA signaling pathway	Sugimoto et al., 2010; Chen et al., 2020
	Grain size	*qSW5/GW5*	GW5/ *qSW5* involved in brassinosteroid signaling pathway to regulate grain width and weight (Novel nuclear protein)	Shomura et al., 2008; Weng et al., 2008; Liu et al., 2017
		Gn1a	Encodes cytokinin oxidase	Ashikari et al., 2005
Maize	Plant architecture	*tb1 (teosinte branched1)*	Two maize mutants, *teosinte branched1* (*tb1*) and *grassy tillers1* (*gt1*), have been shown affected in the regulation of auxin biosynthesis pathway	Doebley et al., 1997; Whipple et al., 2011
		br2	Gene modulates the transport of auxin	Zhang et al., 2019b
	Inflorescence architecture	*ra1 (ramosa1), Tga1*	R*A1* involved in the *ramosa* pathway (Transcription factor)	Sigmon and Vollbrecht, 2010
	Grain filling	*ZmSWEET4c*	Hexose transporter, SWEET4c is important for the Glc to the starch biosynthesis in the endosperm during embryogenesis	Sosso et al., 2015
Wheat	Vernalization	*Vrn2 (ZCCT1 and ZCCT1)*	Likely to coordinate with GA, ABA, cytokinin, and JA signaling pathway	Yan et al., 2004; Deng et al., 2015
		Vrn1	Central gene in vernalization pathway similar to *APETALA* of *Arabidopsis*. Linked with GA, ABA, Cytokinin, and JA signaling pathway	Yan et al., 2003; Deng et al., 2015
	Free threshing	*Q* and *homeologs*	Involved in secondary cell wall synthesis pathways and regulation of chemical composition of glumes	Zhang Z. et al., 2020
	Plant architecture	*Rht-1*	Repressor of gibberellic acid pathway	Thomas, 2017
Sorghum	Plant architecture	*dw3*	Gene modulates the transport of auxin	Multani et al., 2003
	Grain pigmentation	*Tannin1 (Tan 1)*	*Tan1* gene, encoding a WD40 protein, that regulate the tannin biosynthesis	Wu et al., 2012
Barley	Inflorescence architecture	*Vrs2*	*Vrs2* expression influences the expression of genes that regulate biosynthesis and metabolism of auxin and cytokinin (Transcription factor, HD-ZIP)	Komatsuda et al., 2007; Youssef et al., 2017
	Naked (free-threshing) grains	*Nud*	ERF family transcription factor gene regulating a lipid biosynthesis pathway (Transcription factor)	Taketa et al., 2008
Soybean	Determinate growth habit	*Dt2*	Plant height of semi-determinate plants is associated with GA signaling	Zhang et al., 2019a
Tomato	Fruit size	*fw2.2*	Similar to human RAS, *SlKLUH* is the causal gene for the *fw3.2* QTL and encodes a CYP450 of the 78A class	Frary et al., 2000
		SUN	Regulating auxin biosynthetic and responsive pathway	Xiao et al., 2008; Wang et al., 2019
Mustard	Flowering Time	*BrFLC1*	Interacts with the vernalization pathway (MADS-box transcription factor) and coordinate with gibberellic acid pathway	Yuan et al., 2009

margin formation and further enhance dehiscence in shp1/2 and ind mutants, suggesting that cytokinins play a crucial role in valve margin differentiation (Marsch-Martínez et al., 2012). Recent studies reveal gibberellins (GAs) are also involved in the establishment of separation layer cell identity, in addition to auxins and cytokinins (Arnaud et al., 2010). As per the "relief of restraint" model, GA-mediated degradation of DELLA protein is important for GA signaling and also necessary to trigger expression of downstream genes (Harberd, 2003; Sun and Gubler, 2004). *GA3ox1*, which facilitates the final step in bioactive GAs synthesis, is shown as the direct target of IND. ALC interacts physically with DELLA repressors and local GAs production destabilizes the DELLA protein and relieves ALC to play its role in SL cell specification (Arnaud et al., 2010). In summary,

these findings show that many phytohormones participate in the DZ specification and indicate that precise balance between biosynthesis and response is important. Notwithstanding the studies where the function of hormones in the development of DZ have been elucidated, very few studies about how such hormonal signals are coordinated in DZ have been carried out. One of the key challenges is to unravel the complete context of the molecular mechanisms and interactions of plant hormones underlying DZ-specification.

There are many ways for minimizing crop losses due to crop shattering ranging from conventional parental selection with minimum shattering to the screening of mutants and gene editing methods. By advancing the next-generation sequencing and the marker traits associations, many genes involved in pod dehiscence were found, and a series of mutations underlying shattering resistance in several crops and their wild relatives have been identified (Fuller and Allaby, 2009; Dong and Wang, 2015). Attempts have been made to improve shattering resistance in Brassica, which include interfering in the dehiscence process by manipulating the molecular and hormonal control pathways (Fuller and Allaby, 2009; Altpeter et al., 2016) and developing transgenic lines with pod-shattering resistance (Liljegren et al., 2000, 2004). In future, studies should focus, alongside gene-editing methods, on fine-tuning of the degree of shatter-resistance with RNA interference or the use of mutated forms of genes related to shattering in various crops.

Key Genes Targeted for Dwarfing of Cereal to Enhance the Productivity

The plant architecture is genetically controlled by a set of genes which subsequent affect yield and productivity of crop plant species. Often, mutation or knockdown of a single gene could also lead to significant change in the overall plant growth and development, subsequently plant architecture (Spielmeyer et al., 2002). In 1960s, the agricultural transformation that increased the production of rice and wheat was via the introduction of cultivars with a genetic predisposition to a short stature due to restricted elongation of stem (Silverstone and Sun, 2000). This phenotype enabled a significant partitioning of photosynthate produced from photosynthesis to sink organs like grains (Sun and Frelich, 2011).

Currently introduction of dwarfing genes is the most important aspect deployed in modern cereal breeding. The stems of tall wheat and rice crops are not strong enough to sustain heavy grains of the high yielding cultivars, which result in significant yield losses. In addition, the proportion of assimilates partitioned in grain increases yields. Genes associated with the semi-dwarf growth of the wheat and rice cultivars have been studied. In wheat, *Reduced height* (*Rht*) gene has been identified which is shown to interfere with GA signaling transduction pathway (Peng et al., 1999). Subsequently, three research groups investigated *semi dwarf1* (*SD1*) gene from rice and found that the same hormone impair the biosynthesis (Monna et al., 2002; Sasaki et al., 2002; Spielmeyer et al., 2002). Thus, gibberellin hormone appears to be central to plant stature control.

Wheat *Rht* Gene and Gibberellin Signaling

The Green Revolution's wheat dwarfing genes originated in Japan (Gale et al., 1985). The *Norin 10* dwarfing genes are now available worldwide in 70% of current commercial wheat cultivars. Norin10 contains two dwarfing genes that are semi-dominant homologous alleles on Chromosomes B and D. These alleles are labeled as *Rht-B1b* (formerly *Rht1*) and *Rht-D1b* (*Rht2*) to reflect their chromosome position (Boerner et al., 1996). The Rht alleles cause a reduced response to the plant hormone GA class (Gale et al., 1985). These plant hormones are diterpenoid carboxylic acids, that are involved in several processes of development in higher plants, including stem elongation (Hooley, 1994). The Rht gene is an ortholog of *Arabidopsis GA-Insensitive (GAI)* and maize *dwarf 8* genes, for which mutations result in GA-insensitive dwarfs (Peng et al., 1999). *Rht-1a/d8/GAI* (wild type protein) is a subgroup of the GRAS family of proteins that are thought to act as transcriptional regulators (Pysh et al., 1999). Peng et al. (1999) reported base substitutions in the *Rht-B1b* and *Rht-D1b* alleles that insert stop codons within the DELLA region. They mentioned that translational re-initiation at one of several methionines which follow the stop codon could lead to the formation of truncated Rht protein without the DELLA domain, which functions as a constituent (GA insensitive) growth repressor. The D8 (Peng et al., 1999) and GAI mutations (Peng et al., 1997) also lead to partial or complete deletion from one or both of the conserved domains. The Rht-1a/d8/GAI proteins thus function as negative GA signaling regulators and suppress GA function, provided N-terminal domains are present (Harberd et al., 1998; Dill et al., 2001). To support this concept, ectopic expression of GAI (Peng et al., 1999) in rice caused dwarfism and loss of function mutations in Rht-like genes in some cases produces an over-growth phenotype (Ikeda et al., 2001; Chandler et al., 2002). Besides *d8*, *Rht-1a* orthologs were reported in rice (known as *OsGAI* or *SLR1*) (Ogawa et al., 2000; Ikeda et al., 2001) and barley (SLN1) (Chandler et al., 2002). While cereals have a single case of Rht-1a/d8/GAI type proteins, Arabidopsis contains a gene family encoding RGA proteins and three RGA-like proteins (RGL1, -2, -3) in addition to GAI. The Arabidopsis homologues seem to overlap in their function in various GA-regulated developmental processes (Olszewski et al., 2002). It is unknown how a single protein in cereals crops is functionally equivalent to five proteins in Arabidopsis; such variation may indicate major functional redundancy in Arabidopsis or fundamental differences in GA signaling pathways between Arabidopsis and Gramineae members. Recently, some progress was made in understanding the function of Rht-like proteins and their GA repression. RGA (Dill et al., 2001), SLR1 (Itoh et al., 2002), and SLN1 (Gubler et al., 2002) are found in the nucleus and thus rapidly degraded with GA presence, the DELLA domain needed for this process. Rht's upstream signal transduction pathway is still unknown, but GA-induced degradation is believed to involve ubiquitin-mediated proteolysis (Chandler et al., 2002).

Rice *sd1* Gene and Gibberellin Biosynthesis

Unlike *Rht*, the *sd1* mutation of rice is recessive and normal height can be restored in mutants using GA application showing

that they have been impaired in GA production (Ashikari et al., 2002). Three research groups independently isolated the *sd1* gene and showed it encodes GA 20-oxidase (GA20ox), an enzyme involved in biosynthesis of GA (Monna et al., 2002; Sasaki et al., 2002; Spielmeyer et al., 2002). Two of these research groups have used positional cloning to detect a GA20ox open reading frame close to the sd1 locus on the long chromosome arm (Monna et al., 2002; Spielmeyer et al., 2002). *They also reported mutations in corresponding genes from semi-dwarf varieties. The third group, which had inferred the gene's identity by the effect of GA content mutations, used PCR to amplify DNA fragments, corresponding to two GA20ox genes, one of which mapped to the sd1 loci (*Sasaki et al., 2002; Ashikari et al., 2002). Semi-dwarf rice cultivars with Dee-geo-woo-gen sd1 allele contain a 383-bp deletion in the *GA20ox* gene (known as *OsGA20ox2*), which incorporates stop codon that is likely to result in a highly truncated, inactive enzyme. Gibberellin 20-oxidases are 2-oxoglutarate-dependent dioxygenases catalyzing carbon-20 depletion in the penultimate stage in biosynthesis of GA (Hedden and Phillips, 2000). These oxidases are encoded by small gene families, members of which have partial functional redundancy due to overlapping (but different) expression profiles or because of movement of the intermediates synthesized by enzymes between tissues. Therefore, loss-of-function GA20ox mutants are relatively less GA-deficient and are semi-dwarfs, unlike significant GA-deficient plants, which are extremely dwarfed and sometimes sterile. Two *GA20ox* genes were defined in rice: *OsGA20ox1* (Toyomasu et al., 1997) and *OsGA20ox2*. Remarkably, selection for semi-dwarfism in rice has consistently yielded mutations in *OsGA20ox2* instead of *OsGA20ox1* or another GA-biosynthesis gene (for example, *GA 3-oxidase* is also encoded by a multi-gene family). Mutations in other genes might have a severe developmental impact or have negative impact on yield, and thus have been not selected in breeding programs. Genetic and functional analyses of *SLR1/RHT* and *SD1* genes in rice and wheat have enormously improved the understanding of GA biosynthesis and signals, resulting in a strong methodology for manipulating the plant height of major crops. Both cases illustrate the central role played by GAs in controlling developmental processes. Therefore, GA signaling pathways (biosynthesis and signal transduction) are key aspects for manipulation in pursuit of further crop yield improvements. The yields of existing cereal crops seem to be approaching their limit, and new interventions are required if population is not to outstrip the food supply. Targeted genetic engineering/modification using newly emerged genomics, genome-editing technologies may be part of the next Green Revolution.

Achieving Submergence Tolerance

The incidences of uncertain rain and flood have been increased due to continued climate change. Today, more than 30 percent of the rice-planting land is vulnerable to flooding resulting in crop loss. In 1960s, the development of semi-dwarf variety was one of greatest achievement which significantly addressed the issue of global hunger threat caused due to human population explosion. The suppression of GAs production in the stem reportedly made high yielding semi-dwarf rice varieties susceptible to one of the most important abiotic stress "water logging." These developed semi-dwarf rice varieties lacked submergence tolerance. The lower nodes of these varieties unable to produce enough gibberellins to trigger elongation of the internode.

Genomics Based Discovery of Genomic Regions Associated With Submergence Tolerance

Submergence stress causes several adverse impacts on a plant such as low light intensity, hypoxia, nutrient effusion, physical injury, susceptibility to pathogen and pests attacks (Angaji et al., 2010). Several QTL mapping studies reported number of QTLs controlling submergence tolerance (Xu and Mackill, 1996; Nandi et al., 1997; Toojinda et al., 2003). A major QTL (Sub1) for submergence tolerance has been identified on chromosome 9 with LOD 36 and 69% of phenotypic variance explained (PVE) (Xu and Mackill, 1996). Sequencing of Sub1 genomic region identified three genes which encodes a ERFs (Sub1A, Sub1B, and Sub1C) in which Sub1A has been reported as a key component of submergence tolerance (Xu et al., 2006). Further cloning and characterization of Sub1 QTL helping in the detection of responsible genes and also help to discover tightly linked gene-based markers for molecular breeding program (Siangliw et al., 2003; Toojinda et al., 2005; Neeraja et al., 2007). Furthermore, in other studies major QTLs namely qAG9-2 on L.G. 9 and qAG7-1 on L.G. 7 were reported (Angaji et al., 2010; Septiningsih et al., 2013). Later on, qAG9-2 QTL has been fine mapped and found a candidate gene OsTPP7 which encodes a trehalose-6-phosphate phosphatase which is responsible to regulate anaerobic generation (Kretzschmar et al., 2015). Both Sub1 and qAG9-2 major QTLs are widely used in rice breeding programs to improve submergence tolerance at germination and vegetative stages. Utilizing genomics resources several breeding efforts are also made in developing submergence tolerance varieties to sustain rice production. Various landraces and traditional genotypes namely, Kurkaruppan, FR13A, Thavalu, Goda Heenati, etc., were reported to be a suitable source of alleles which is associated with submergence tolerance (Miro and Ismail, 2013).

Precise Characterization of Genes Governing Submergence Tolerance

In recent years significant progressed have been made toward understanding the physiological, biochemical and genetic basis of submergence tolerance, to identify the causal gene(s) that are crucial for submergence tolerance (Oladosu et al., 2020). Recently, Kuroha et al. (2018) identified the gene *SD1* (*SEMIDWARF*) responsible for submergence-induced elongation of internode by producing gibberellins mainly GA4. Another study identified genes *SNORKEL 1 (SK1)* and *SK2* which encodes for ERFs, appeared to trigger submergence tolerance via ethylene signaling (Hattori et al., 2009). Both gene products further facilitate the internode elongation through GAs. Previous study identified a submergence tolerance gene *SUB1A* (an *Ethylene-response-factor-like* gene) on chromosome 9 which encodes ERFs (Xu et al., 2006; Fukao et al., 2006). During flash floods, *SUB1A* inhibits plant elongation at the seedling stage.

Flash floods usually last for a few weeks. Cultivars carrying *SUB1A* tolerance gene show stunted growth and can survive in submerged conditions for a few weeks. Both *SNORKEL 1* and *SNORKEL 2* (*SK1/2*) genes and *SUB1A* encode ERFs which are associated with GAs, but they act in opposite ways in controlling plant development in response to submergence. Further more research is required to uncover the various pathways associated with *SK1*; *SK2* and *SUB1A*. Furthermore, recently two genes have been identified *ACCELERATOR OF INTERNODE ELONGATION 1* (*ACE1*) and *DECELERATOR OF INTERNODE ELONGATION 1* (*DEC1*) which are responsible to control stem elongation (Nagai et al., 2020). *ACE1* gene encoding an unknown function protein which is associated with internodes elongation via GAs, whereas, *DEC1* gene encoding a zinc – finger TF, which suppresses internodes elongation. Both the genes influence gibberellin-activated cell division in stem nodes. The expression of *ACE1* gene during submergence conditions in rice triggers elongation of internodes within a cell-division zone of the plant. This results in an increased number of elongated internodes and increased plant height. Further gene *ACE1C9285* is controlled by SUB1C, a gibberellin-activated TF which is upregulated in response to submergence (Fukao and Bailey-Serres, 2008). *SUB1C* expression level seemingly low in cultivars that contain the *SUB1A-1* regulator gene, a homolog to *SUB1C*. In short rice cultivars expressing gene *SUB1A-1*, GAs responsiveness altered, subsequently use carbon pool for leaves elongation, and restrict overall plant development and enter to transient quiescent stage during flooding, an adaptation to overcome deep floods (Fukao et al., 2006; Xu et al., 2006). In semi-dwarf cultivars, internodes elongation only takes place in the upper internodes during growth stage. Nagai et al. (2020) reported a gene *ACE1-LIKE1*, which triggers upper internodes growth in deep-water. Presently, these omics study based information on the genetic basis of submergence tolerance is the base of rapid improvement of plant architecture to design a high yielding crop tolerant submergence.

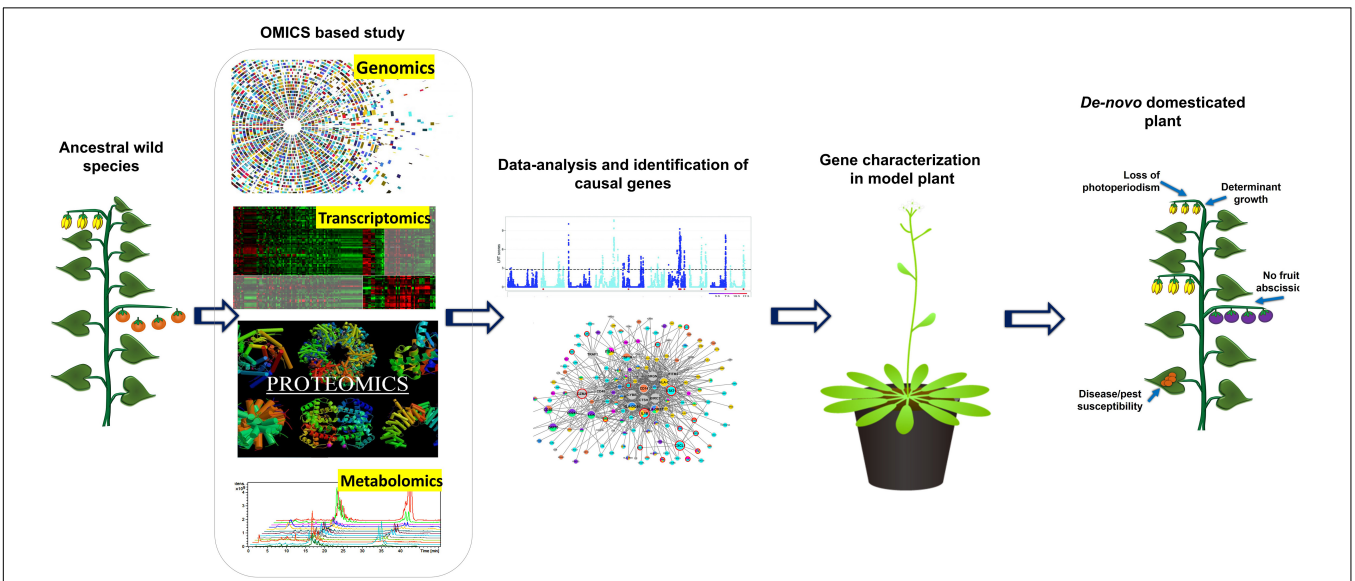

FIGURE 2 | Schematic diagram representing the role of OMICS based research in gene characterization and development of designer crops using *de novo* domesticated crops approach.

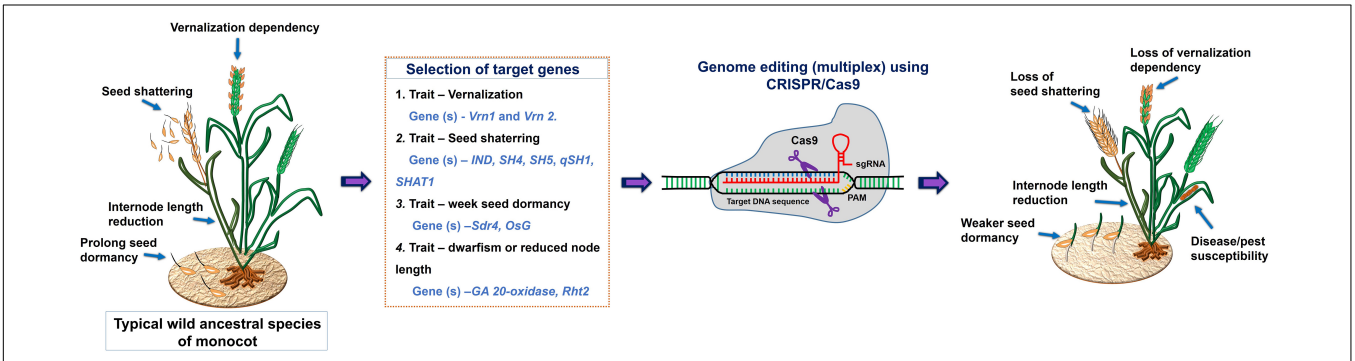

FIGURE 3 | A schematic representation of a draft model for the selection of target genes for CRISPR/Cas9 mediated domestication of wild ancestral species of monocot.

TRANSLATION OF OMICs DRIVEN DATA FOR RE-DOMESTICATION AND *DE NOVO* DOMESTICATION: UTILIZATION OF GENOME/GENE EDITING TOOL

Gene-editing technologies have become choice of a researcher to domesticate neglected crops and wild relatives in a short period (Fernie and Yan, 2019). Traditionally, plant domestication and the development of productive cultivars required decades of breeding, which is also the key reason why so many breeding programs over the last 100 years focused on further improvement of a relatively small number of crops. Recent identification of several major domestication genes and scientific breakthroughs in integrating various genomic changes in plants concurrently with CRISPR/Cas9 editing has allowed re-domestication of existing crop plants and *de-novo* domestication wild species to be domesticated within a single generation (**Figure 2**) (Schindele et al., 2020). *De-novo* domestication has contributed to agro-biodiversity and diet quality, with possible future environmental and nutritional benefits (Singh et al., 2019). In the history of crop domestication amid higher yield selection and breeding, international germplasm exchange; multiple local resistance and resilience genes of wild species have been lost or have never been completely incorporated into breeding lines (Fernie and Yan, 2019). In other words, wild relatives of domesticated plants have significantly higher variable gene pool than that of domesticated ones (Hickey et al., 2019). As we start to uncover more about the framework of crop genomes and the loci of quality traits, there are chances of incorporating valuable characters into existing crop species and ways to quickly re-domesticate new crops. This step can be effectively achieved using breakthrough CRISPR-Cas9 gene-editing technologies, in particular, to introduce beneficial alleles without linkage drag (Li et al., 2018), to produce novel quantitative variations (Rodríguez-Leal et al., 2017), direct deletion of deleterious alleles (Johnsson et al., 2019), and/or higher recombination rates (Mieulet et al., 2018). Recently, gene editing has been shown to enhance plant architecture, flower development, and fruit size in *Physalis pruinosa* (Lemmon et al., 2018). Gene editing is a promising method to generate diversity and to compensate for the genetic hitchhiking effects in germplasm. For reference, associated selection of traits such as fruit weight and disease resistance altered the tomato metabolome, providing an opportunity for precise breeding to alter nutritional and flavor traits (Zhu et al., 2018). These hitchhiking effects and others, such as those found in rice and maize, represent promising goals for genetic modification to fettle linkage drag (Palaisa et al., 2004). For instance, African rice landrace Kabre possess superior resistance to pests and tolerance to drought; however, during domestication the plant architecture compromised affecting their overall yield potential. To address this Lacchini et al. (2020) targeted multiples genes which control plant architecture (*HTD1*) and control seed size and/or yield (*GS3, GW2,* and *GN1A*) by generating knockouts through multiplex CRISPR/Cas9. In knockouts, mutation in *HTD1* gene caused reduced plant high to diminish lodging and improved tillering, whereas mutations in *GS3, GW2,* and *GN1A* resulted increased panicle and length along with improved seed girth. Earlier, Hu et al. (2019) demonstrated generation of semi-dwarf rice lines by targeting gene *SD1* and *Photosensitivity5* (*SE5)* in elite landraces Kasalath. In this post genomics, the technique CRISPR/Cas has received overwhelming response and till dates several knockouts of rice elite varieties are available with improved traits by targeting specific genes which were characterized due to viability of several omics approached era. Some of the examples for the targeted traits and gene targets in rice are *LAZY1* for tiller-spreading, *Gn1a, GS3,* and *DEP1* for improved grain number, size and dense erect panicles, *SBEIIb* for High amylose content, *OsERF922* for enhanced blast resistance, *OsSEC3A* for resistance against blast causing pathogen *Magnaporthe oryzae, OsSWEET13 for* bacterial blight resistance, *ALS* and *EPSPS* for herbicide resistance, *OsPDS, OsMPK2, OsMPK5, OsBADH2, OsAOX1a, OsAOX1b, OsAOX1c,* and *OsBEL* for tolerance against various abiotic stress, OsHAK-1 for low cesium accumulation, and *OsPRX2* for potassium deficiency tolerance (Shan et al., 2013; Xie and Yang, 2013; Shan et al., 2014; Xu et al., 2014; Zhang H. et al., 2014; Zhou et al., 2014; Woo et al., 2015; Meng et al., 2017; Nieves-Cordones et al., 2017; Mao et al., 2018; Ma et al., 2018). Likewise, in wheat *EDR1*, *TaMLOA1, TaMLOB1,* and *TaMLOD1* targeted for resistance to powdery mildew, and *GW2* and *TaGW2* targeted for increased grain size, weight and protein content (Shan et al., 2014; Wang et al., 2014; Gil-Humanes et al., 2017; Kim et al., 2018; Wang et al., 2018). In orphan crops cassava and flax herbicide resistance has been introduced by targeting a gene *EPSPS* (Sauer et al., 2016; Hummel et al., 2018); whereas *ALS* was targeted in soybean (Cai et al., 2015). Similarly, many traits have been introduced or improved by targeting various genes in some economically important crops plants such as maize, tomato, potato, grapes, orange, cucumber, tea, etc. (Adhikari and Poudel, 2020; Bhatta and Malla, 2020).

TABLE 5 | List of genes targeted in wild ancestral species of tomato and strawberry to demonstrate *de novo* domestication.

Wild relative	Target Gene	Traits modification	References
Solanum pimpinellifolium	*CLV3, WUS, SP, SP5G, and GGP1*	Plant height and response to phtotoperiodism, flower numbers, and fruit size and shape, and ascorbic acid content	Zsögön et al., 2018
	OVATE, MULT, FAS, SP, and CycB	Plant architecture and habitat, flower numbers, and fruit size and shape, and lycopene content	Li et al., 2018
Fragaria vesca	*FveTAR1* and Fve*YUC10*	Auxin biosynthetic and signaling genes affecting plant growth and reproductive organ development	Feng et al., 2019
	FveTAA1 and *FveARF 8*	Auxin biosynthetic and signaling genes affecting plant growth and reproductive organ development	Zhou et al., 2018

The wild ancestral species of crop plants such as *Solanum pimpinellifolium* for tomato; *Solanum demissum* and *S. stoloniferum* of potato; *Fragaria vesca* of strawberry; *Teosinte* and *Tripsacum* of maize; *Triticum dicoccoides,* and *T. turgidum L.* ssp. *Durum* of wheat; *Oryza rufipogon* and *O. longistaminata* of rice; *Manihot glaziovii* and *M. neosana* and *Glycine soja* of soybean have been used for introgression key agronomic important traits into cultivars though breeding program (Zsögön et al., 2017). Moreover, most of the domesticated related traits and associated genes well characterized and has been linked with the metabolic pathway(s), and/or hormone biosynthesis and signaling (**Table 4**); therefore, integrated omics approach

TABLE 6 | A model representing state of art for selecting the genes which can be edited to domesticate crop wild ancestral species through CRISPR/Cas9 approach.

Crop Name	Target Gene	Function	References
Zea Mays	*Tb1*	TCP-gene family TF which is involved in suppression of side branching changes the source/sink relationships; yields increase.	Doebley et al., 1997; Studer et al., 2011
	tga1	SBP-box TF have a key role in alteration of the encased kernel to naked kernel	Wang et al., 2015
	CCT	CCT domain-containing protein gene involved in decrease of photoperiod sensitivity	Yang et al., 2013; Huang et al., 2018
Glycine max	*DT1*	CETS is a family of regulatory genes which are involved in transforming indeterminate growth to determinate, resulting in developing a compact crop.	Tian et al., 2010; Cai et al., 2018
	GA20ox	Key enzyme involved in Gibberellin biosynthesis and identified as its association with seed weight	Lu et al., 2016
	SHAT1-5	Plant specific NAC gene family TF involved in the biosynthesis of secondary cell wall which facilitating fiber cell cap thickening result in a decreasing the rate of pod shattering	Dong Y. et al., 2014
Solanum lycopersicum	*ARF19*	Auxin response factor 19 TF reported being a negative regulator of fruit set	De Jong et al., 2009
	BRC1a	*BRANCHED1a* gene encoding a TCP family TF which involved in the regulation of lateral shoot outgrowth	Martín-Trillo et al., 2011
	CHI	Chalcone Isomerase is associated with flavonoid biosynthesis	Willits et al., 2005
	S	*Compound inflorescence (s)* encodes a homeobox TF which controls the number of flower/fruits per inflorescence architecture	Lippman et al., 2008
	CKX	Cytokinin oxidase enzyme associated gene is involved in the inactivation of bioactive cytokinin	Ashikari et al., 2005
	FAS	*CLAVATA3* encoded the *Fasciated* gene which is associated with controlling locules number and size in fruit	Xu C. et al., 2015
	GLK2	Golden2-like TF belongs to GARP family which play a key role in the regulation of chloroplast development in fruits	Powell et al., 2012
	J1	*JOINTLESS* belongs to MADS-box gene family controlling the development of the abscission zone in pedicels	Mao et al., 2000
	Cyc-B	Lycopene β-cyclase involved in the catalyzes the conversion of lycopene into β-carotene	Ronen et al., 2000
	NOR	*Non-ripening* gene associated with the initiation of the normal fruit ripening	Seymour et al., 2013
	O	*OVATE* is a regulatory gene involved in the regulation of fruit shape	Liu et al., 2002
	PRO	*PROCERA* gene involved in suppression of gibberellin signaling	Jasinski et al., 2008
	RIN	*RIPENING INHIBITOR* gene belongs MADS-box family; key role in controlling biosynthesis of ripening -related ethylene	Seymour et al., 2013
	SP	*SELF-PRUNING* gene is a developmental regulator associated with indeterminate and sympodial growth habit in tomato	Pnueli et al., 1998
	SFT	*SINGLE FLOWER TRUSS* gene involved in regulation of flowering	Lifschitz et al., 2006
	CLV3	*CLAVATA3* key meristematic gene, regulating locule numbers in fruit	Rodríguez-Leal et al., 2017
	PSY1	Phytoene synthase 1 gene involved in the biosynthesis of carotenoid resulting in yellow flesh fruit	Hayut et al., 2017
	ANT1	*Anthocyanin mutant 1* gene encodes a *Myb* TF which involve in increasing anthocyanin content	Čermák et al., 2015
	GAD2, GAD3	Key genes encoding an enzyme glutamate decarboxylase for biosynthesis of γ-aminobutyric acid (GABA) in fruit	Nonaka et al., 2017
	ALMT9	*Al-ACTIVATED MALATE TRANSPORTER9* gene involved in decreasing the malate content accumulation in fruit	Ye et al., 2017
	MBP21	*MBP21* is a MADS-box protein controlling formation of abscission zone in pedicel	Roldan et al., 2017
	BOP1, BOP2, BOP3	*BLADE ON PETIOLE* gene reported being associated with early flowering with simplified inflorescences	Xu et al., 2016
	SP5G	*SELF-PRUNING 5G* gene is a flowering repressor linked involved in the development of day-length-sensitive tomato plant	Soyk et al., 2017
Cucumis sativus	*WIP1*	*WIP1* is a C2H2 zinc finger TF gene involved in development of gynoecious plant	Hu et al., 2017
Actinidia chinensis	*CEN*	*CENTRORADIALIS* like gene associated with the development of compact plant with early terminal flowering and fruit development	Varkonyi-Gasic et al., 2019

which also involved metabolomics study has provided insights into the molecular basis of trait domestication. One can target these domesticated genes in wild ancestral plants for their speedy domestication. Now through CRISPR-Cas9 method these wild relative can be directly used for re-domestication or *de-novo* domestication (**Figure 3** and **Tables 5**, **6**). One of the important case study of *de novo* domestication in tomato has been done by Zsögön et al. (2018) by targeting important domestication related genes through CRISPR-Cas9 in tomato wild ancestral species *S. pimpinellifolium*. Zsögön et al. (2018) targeted *SELFPRUNING* (*SP*, control general plant growth habit), *OVATE* (*O*, regulate fruit shape); *FASCIATED* (*FAS*), *FRUIT WEIGHT* 2.2 and *CLAVATA3* (*CLV3*) (control fruit size and locule numbers), *MULTIFLORA* (*MULT*, regulate fruit number), and *LYCOPENE BETA CYCLASE* (*CycB*). The engineered *S. pimpinellifolium* lines and achieved remarkable change in the plant overall phenotype with important traits essential for the commercial purpose such as increased lycopene content, enhanced fruit shape and determinant growth of plant; moreover, this was achieved in just single generation. Another study involved editing of multiples genes *SP, SP5G* (control day-length insensitivity), *CLV3,* WUSCHEL (*WUS*) and *GDP-L-galactose phosphorylase 1* (*GGP1,* control biosynthesis of ascorbic acid) in *S. pimpinellifolium* (Li et al., 2018). This study clearly showed how selective editing of domesticated related genes can completely alter the plant architecture and improves the nutritional quality of fruits and makes convert wild relative into domesticated crop with retained biotic and abiotic stress tolerance properties (Li et al., 2018). Very recently, in the wild strawberry (*Fragaria vesca*) few attempts has been made to demonstrate the procedure of the re-domestication or *de novo* domestication (Zhou et al., 2018; Feng et al., 2019). These attempts involved editing of genes *tryptophan aminotransferase of Arabidopsis 1* (*TAA1,* converts tryptophan to indole-3-pyruvic acid), *Auxin response factor* 8 (*ARF8, repressor of auxin signaling*) and *YUCCA10* (*YUC10,* family of flavin-containing monooxygenases convert IPyA to IAA), key auxin biosynthetic and signaling pathways genes. Rice has five allotetraploids (BBCC, CCDD, HHJJ, HHKK, and KKLL) wild species which are also valuable genetic resources for improving of elite rice varieties. Among them the CCDD (species from South America genome) possess much stronger biotic and abiotic resistance and larger biomass compared to the cultivated diploid rice. Recently Yu et al. (2021) demonstrated *de novo* domestication of wild allotetraploid rice PPR1 (*O. alta*; CCDD type genome) by improving six agronomically important traits *viz* nutrition use efficiency, abiotic stress tolerance, grain yield and quality, heading date, biotic stress resistance and sterility by genome editing targeting multiple genes including *OaSD1-CC, OaSD1-DD, OaAn-1-CC,* and *OaAn-1-DD* by CRISPR/Cas9 method. This suggests that CRISPR/Cas is a promising approach tool for the domestication of crops (Crews and Cattani, 2018), and is highly important for characters of defined selective sweeps in related species. These achievements were possible due to precise prediction of causal genes and metabolic pathways achieved by interpretation of data generated through genomics, transcriptomics, metabolomics, etc.

CONCLUSION

Omics have helped plant biologists to dissect important developmental clues and gene characterization. Presently, multidimensional omics approach where the biological sample can be analyzed for transcriptomics, proteomics and metabolomics in parallel, etc; offers plant biologists a complete understanding of plant metabolism by revisiting the metabolic pathways or identification of newer pathways. In the past 20 years, plant biologists have gathered significant amount of data relevant to genomes, transcriptome, proteome, and metabolome. Recent attempts are on development of gene-expression and proteome atlas. Altogether, this would strengthen the knowledge of the metabolic pathways, which have played crucial role during domestication of crop as well as trait improvement. Now, this knowledge has been translated to develop designer crops with desired traits by editing metabolic pathways of wild ancestral species (rich resource of genetic variations) called as *de novo*-crop domestication. Domestication of wild or semi domesticated crop (tolerant to stress responses) would be feasible by multi step process were few important traits need to be improved first using genome editing; later the homologous lines can be selected for next level of trait modification. Such approach would be able to deliver a commercial line in 5 to 10 years. The CRISPR/Cas technique need to be explored in full extent by targeting several traits such as bio-fortification of nutrition's; because the current growing population also demand nutritional security. To achieve this, analysis of resequencing data available for the several crops is important; including GWAS which can identify high quality SNPs and haplotypes associated with target trait. Therefore, we expected in next 20 years' omics technology driven *de-novo* crop domestication will play very important role in the field of plant biotechnology.

AUTHOR CONTRIBUTIONS

RK received the invitation and conceived the plan for the manuscript. RK and VS wrote the manuscript. AK, SS, DR, SK, KP, BH, AV, RK, MP, ST, and GN improved the section and developed the table and figures. ST helped in developing the revised version. All the authors have read the manuscript before submission.

REFERENCES

Abe, K., Araki, E., Suzuki, Y., Toki, S., and Saika, H. (2018). Production of high oleic/low linoleic rice by genome editing. *Plant Physiol. Biochem.* 131, 58–62. doi: 10.1016/j.plaphy.2018.04.033

Adhikari, P., and Poudel, M. (2020). CRISPR-Cas9 in agriculture: Approaches, applications, future perspectives, and associated challenges. *Malays. J. Halal Res.* 3, 6–16. doi: 10.2478/mjhr-2020-0002

Agrawal, G. K., Jwa, N. S., Lebrun, M. H., Job, D., and Rakwal, R. (2010). Plant secretome: unlocking secrets of the secreted proteins. *Proteomics* 10, 799–827.

Alonso, A., Marsal, S., and Julià, A. (2015). Analytical methods in untargeted metabolomics: state of the art in 2015. *Front. Bioeng. Biotechnol.* 3:23. doi: 10.3389/fbioe.2015.00023

Alseekh, S., Ofner, I., Liu, Z., Osorio, S., Vallarino, J., Last, R. L., et al. (2020). Quantitative trait loci analysis of seed-specialized metabolites reveals seed-specific flavonols and differential regulation of glycoalkaloid content in tomato. *Plant J.* 103, 2007–2024. doi: 10.1111/tpj.14879

Alseekh, S., Tohge, T., Wendenberg, R., Scossa, F., Omranian, N., Li, J., et al. (2015). Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. *Plant Cell* 27, 485–512. doi: 10.1105/tpc.114.132266

Alseekh, S., Tong, H., Scossa, F., Brotman, Y., Vigroux, F., Tohge, T., et al. (2017). Canalization of tomato fruit metabolism. *Plant Cell* 29, 2753–2765. doi: 10.1105/tpc.17.00367

Altpeter, F., Springer, N. M., Bartley, L. E., Blechl, A. E., Brutnell, T. P., Citovsky, V., et al. (2016). Advancing crop transformation in the era of genome editing. *Plant Cell* 28, 1510–1520.

Alvarez, S., Marsh, E. L., Schroeder, S. G., and Schachtman, D. P. (2008). Metabolomic and proteomic changes in the xylem sap of maize under drought. *Plant Cell Environ.* 31, 325–340. doi: 10.1111/j.1365-3040.2007.01770.x

Alvarez, S., Roy Choudhury, S., and Pandey, S. (2014). Comparative quantitative proteomics analysis of the ABA response of roots of drought-sensitive and drought-tolerant wheat varieties identifies proteomic signatures of drought adaptability. *J. Proteome Res.* 13, 1688–1701. doi: 10.1021/pr401165b

Amarasinghe, Y. P. J., Kuwata, R., Nishimura, A., Phan, P. D. T., Ishikawa, R., and Ishii, T. (2020). Evaluation of domestication loci associated with awnlessness in cultivated rice. Oryza sativa. *Rice* 13:26

Angaji, S. A., Septiningsih, E. M., Mackill, D. J., and Ismail, A. M. (2010). QTLs associated with tolerance of flooding during germination in rice (*Oryza sativa* L.). *Euphytica* 172, 159–168. doi: 10.1007/s10681-009-0014-5

Angelovici, R., Batushansky, A., Deason, N., Gonzalez-Jorge, S., Gore, M. A., Fait, A., et al. (2017). Network-guided GWAS improves identification of genes affecting free amino acids. *Plant Physiol.* 173, 872–886. doi: 10.1104/pp.16.01287

Arnaud, N., Girin, T., Sorefan, K., Fuentes, S., Wood, T. A., Lawrenson, T., et al. (2010). Gibberellins control fruit patterning in Arabidopsis thaliana. *Genes Dev.* 24, 2127–2132. doi: 10.1101/gad.593410

Arnaud, N., Lawrenson, T., Østergaard, L., and Sablowski, R. (2011). The same regulatory point mutation changed seed-dispersal structures in evolution and domestication. *Curr. Biol.* 21, 1215–1219. doi: 10.1016/j.cub.2011.06.008

Ashikari, M., Sakakibara, H., Lin, S., Yamamoto, T., Takashi, T., Nishimura, A., et al. (2005). Cytokinin oxidase regulates rice grain production. *Science* 309, 741–745. doi: 10.1126/science.1113373

Ashikari, M., Sasaki, A., Ueguchi-Tanaka, M., Itoh, H., Nishimura, A., Datta, S., et al. (2002). Loss-of-function of a rice gibberellin biosynthetic gene, GA20 oxidase (GA20ox-2), led to the rice 'green revolution'. *Breed. Sci.* 52, 143–150. doi: 10.1270/jsbbs.52.143

Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., et al. (2008). Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* 320, 938–941. doi: 10.1126/science.1157956

Bhatta, B. P., and Malla, S. (2020). Improving horticultural crops via CRISPR/Cas9: current successes and prospects. *Plants* 9:1360. doi: 10.3390/plants9101360

Bishopp, A., Benková, E., and Helariutta, Y. (2011). Sending mixed messages: auxin-cytokinin crosstalk in roots. *Curr. Opin. Plant Biol.* 14, 10–16. doi: 10.1016/j.pbi.2010.08.014

Boerner, A., Plaschke, J., Korzun, V., and Worland, A. J. (1996). The relationships between the dwarfing genes of wheat and rye. *Euphytica* 89, 69–75. doi: 10.1007/bf00015721

Bommisetty, R., Chakravartty, N., Bodanapu, R., Naik, J. B., Panda, S. K., Lekkala, S. P., et al. (2020). Discovery of genomic regions and candidate genes for grain weight employing next generation sequencing based QTL-seq approach in rice (*Oryza sativa* L.). *Mol. Biol. Rep.* 47, 8615–8627. doi: 10.1007/s11033-020-05904-7

Cai, Y., Chen, L., Liu, X., Guo, C., Sun, S., Wu, C., et al. (2018). CRISPR/Cas9-mediated targeted mutagenesis of GmFT2a delays flowering time in soya bean. *Plant Biotech. J.* 16, 176–185. doi: 10.1111/pbi.12758

Cai, Y., Chen, L., Liu, X., Sun, S., Wu, C., Jiang, B., et al. (2015). CRISPR/Cas9-mediated genome editing in soybean hairy roots. *PLoS One* 10:e0136064. doi: 10.1371/journal.pone.0136064

Capriotti, A. L., Borrelli, G. M., Colapicchioni, V., Papa, R., Piovesana, S., Samperi, R., et al. (2014). Proteomic study of a tolerant genotype of durum wheat under salt-stress conditions. *Anal. Bioanal. Chem.* 406, 1423–1435. doi: 10.1007/s00216-013-7549-y

Carreno-Quintero, N., Acharjee, A., Maliepaard, C., Bachem, C. W., Mumm, R., Bouwmeester, H., et al. (2012). Untargeted metabolic quantitative trait loci analyses reveal a relationship between primary metabolism and potato tuber quality. *Plant Physiol.* 158, 1306–1318. doi: 10.1104/pp.111.188441

Čermák, T., Baltes, N. J., Čegan, R., Zhang, Y., and Voytas, D. F. (2015). High-frequency, precise modification of the tomato genome. *Genome Biol.* 16:232.

Chan, E. K., Rowe, H. C., Hansen, B. G., and Kliebenstein, D. J. (2010). The complex genetic architecture of the metabolome. *PLoS Genet.* 6:e1001198. doi: 10.1371/journal.pgen.1001198

Chandler, P. M., Marion-Poll, A., Ellis, M., and Gubler, F. (2002). Mutants at the slender1 locus of barley cv Himalaya. molecular and physiological characterization. *Plant Physiol.* 129, 181–190. doi: 10.1104/pp.010917

Chen, J., Wang, J., Chen, W., Sun, W., Peng, M., Yuan, Z., et al. (2018). Metabolome analysis of multi-connected biparental chromosome segment substitution line populations. *Plant Physiol.* 178, 612–625. doi: 10.1104/pp.18.00490

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., et al. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* 46:714. doi: 10.1038/ng.3007

Chen, W., Wang, W., Lyu, Y., Wu, Y., Huang, P., Hu, S., et al. (2020). OsVP1 activates Sdr4 expression to control rice seed dormancy via the ABA signaling pathway. *Crop J.* 9, 68–78. doi: 10.1016/j.cj.2020.06.005

Chen, W., Wang, W., Peng, M., Gong, L., Gao, Y., Wan, J., et al. (2016). Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nat. Commun.* 7:12767.

Chen, X., Lu, Q., Liu, H., Zhang, J., Hong, Y., Lan, H., et al. (2019). Sequencing of cultivated peanut, Arachis hypogaea, yields insights into genome evolution and oil improvement. *Mol. Plant* 12, 920–934. doi: 10.1016/j.molp.2019.03.005

Chiang, S., Zhang, W., and Ouyang, Z. (2018). Paper spray ionization mass spectrometry: recent advances and clinical applications. *Expert Rev. Proteomics* 15, 781–789. doi: 10.1080/14789450.2018.1525295

Chu, P., Yan, G. X., Yang, Q., Zhai, L. N., Zhang, C., Zhang, F. Q., et al. (2015). iTRAQ-based quantitative proteomics analysis of Brassica napus leaves reveals pathways associated with chlorophyll deficiency. *J. Proteomics* 113, 110–126.

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., et al. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods* 5, 613–619. doi: 10.1038/nmeth.1223

Coppola, M., Diretto, G., Digilio, M. C., Woo, S. L., Giuliano, G., Molisso, D., et al. (2019). Transcriptome and metabolome reprogramming in tomato plants by *Trichoderma harzianum* strain T22 primes and enhances defence responses against aphids. *Front. Physiol* 10:745. doi: 10.3389/fphys.2019.00745

Crews, T. E., and Cattani, D. J. (2018). Strategies, advances, and challenges in breeding perennial grain crops. *Sustainability* 10:2192. doi: 10.3390/su10072192

Daware, A., Das, S., Srivastava, R., Badoni, S., Singh, A. K., Agarwal, P., et al. (2016). An efficient strategy combining SSR markers-and advanced QTL-seq-driven QTL mapping unravels candidate genes regulating grain weight in rice. *Front. Plant Sci.* 7:1535. doi: 10.3389/fpls.2016.01535

De Jong, M., Wolters-Arts, M., Feron, R., Mariani, C., and Vriezen, W. H. (2009). The Solanum lycopersicum auxin response factor 7 (SlARF7) regulates auxin signaling during tomato fruit set and development. *Plant J.* 57, 160–170. doi: 10.1111/j.1365-313x.2008.03671.x

Deng, W., Casao, M. C., Wang, P., Sato, K., Hayes, P. M., Finnegan, E. J., et al. (2015). Direct links between the vernalization response and other key traits of cereal crops. *Nat. Commun.* 6:5882

Deokar, A., Sagi, M., Daba, K., and Tar'an, B. (2019). QTL sequencing strategy to map genomic regions associated with resistance to ascochyta blight in chickpea. *Plant Biotechnol. J.* 17, 275–288. doi: 10.1111/pbi.12964

Di Lena, G., Casini, I., Lucarini, M., and Lombardi-Boccia, G. (2019). Carotenoid profiling of five microalgae species from large-scale production. *Food Res. Int.* 120, 810–818. doi: 10.1016/j.foodres.2018.11.043

Dill, A., Jung, H. S., and Sun, T. P. (2001). The DELLA motif is essential for gibberellin-induced degradation of RGA. *Proc. Natl. Acad. Sci. U.S.A.* 98, 14162–14167. doi: 10.1073/pnas.251534098

Doebley, J., Stec, A., and Hubbard, L. (1997). The evolution of apical dominance in maize. *Nature* 386, 485–488. doi: 10.1038/386485a0

Dong, X., Chen, W., Wang, W., Zhang, H., Liu, X., and Luo, J. (2014). Comprehensive profiling and natural variation of flavonoids in rice. *J. Integr. Plant Biol.* 56, 876–886. doi: 10.1111/jipb.12204

Dong, Y., and Wang, Y. Z. (2015). Seed shattering: from models to crops. *Front. Plant Sci.* 6:476. doi: 10.3389/fpls.2015.00476

Dong, Y., Yang, X., Liu, J., Wang, B. H., Liu, B. L., and Wang, Y. Z. (2014). Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. *Nat. Commun.* 5:3352.

Duncan, O., Trösch, J., Fenske, R., Taylor, N. L., and Millar, A. H. (2017). Resource: mapping the *Triticum aestivum* proteome. *Plant J.* 89, 601–616.

Fang, C., and Luo, J. (2019). Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. *Plant J.* 97, 91–100. doi: 10.1111/tpj.14097

Feng, J., Dai, C., Luo, H., Han, Y., Liu, Z., and Kang, C. (2019). Reporter gene expression reveals precise auxin synthesis sites during fruit and root development in wild strawberry. *J. Exp. Bot.* 70, 563–574. doi: 10.1093/jxb/ery384

Feng, J., Long, Y., Shi, L., Shi, J., Barker, G., and Meng, J. (2012). Characterization of metabolite quantitative trait loci and metabolic networks that control glucosinolate concentration in the seeds and leaves of *Brassica napus*. *New Phytol.* 193, 96–108. doi: 10.1111/j.1469-8137.2011.03890.x

Fernandez-Orozco, R., Gallardo-Guerrero, L., and Hornero-Méndez, D. (2013). Carotenoid profiling in tubers of different potato (*Solanum* sp) cultivars: accumulation of carotenoids mediated by xanthophyll esterification. *Food Chem.* 141, 2864–2872. doi: 10.1016/j.foodchem.2013.05.016

Fernie, A. R., and Yan, J. (2019). De novo domestication: an alternative route toward new crops for the future. *Mol. Plant* 12, 615–631. doi: 10.1016/j.molp.2019.03.016

Ferrão, L. F. V., Johnson, T. S., Benevenuto, J., Edger, P. P., Colquhoun, T. A., and Munoz, P. R. (2020). Genome-wide association of volatiles reveals candidate loci for blueberry flavor. *New Phytol.* 226, 1725–1737. doi: 10.1111/nph.16459

Finnie, C., Sultan, A., and Grasser, K. D. (2011). From protein catalogues towards targeted proteomics approaches in cereal grains. *Phytochemistry* 72, 1145–1153. doi: 10.1016/j.phytochem.2010.11.014

Frary, A., Nesbitt, T. C., Grandillo, S., Knaap, E., Cong, B., Liu, J., et al. (2000). fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. *Science* 289, 85–88. doi: 10.1126/science.289.5476.85

Fukao, T., and Bailey-Serres, J. (2008). Submergence tolerance conferred by Sub1A is mediated by SLR1 and SLRL1 restriction of gibberellin responses in rice. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16814–16819. doi: 10.1073/pnas.0807821105

Fukao, T., Xu, K., Ronald, P. C., and Bailey-Serres, J. (2006). A variable cluster of ethylene response factor–like genes regulates metabolic and developmental acclimation responses to submergence in rice. *Plant Cell* 18, 2021–2034. doi: 10.1105/tpc.106.043000

Fuller, D. Q., and Allaby, R. (2009). Seed dispersal and crop domestication: Shattering, germination and seasonality in evolution under cultivation. *Annu. Plant Rev.* 38, 238–295. doi: 10.1002/9781444314557.ch7

Furukawa, T., Maekawa, M., Oki, T., Suda, I., Iida, S., Shimada, H., et al. (2007). The Rc and Rd genes are involved in proanthocyanidin synthesis in rice pericarp. *Plant J.* 49, 91–102. doi: 10.1111/j.1365-313x.2006.02958.x

Gale, M. D., Youssefian, S., and Russell, G. E. (1985). Dwarfing genes in wheat. *Progr. Plant Breed.* 1, 1–35. doi: 10.1016/b978-0-407-00780-2.50005-9

Garbowicz, K., Liu, Z., Alseekh, S., Tieman, D., Taylor, M., Kuhalskaya, A., et al. (2018). Quantitative trait loci analysis identifies a prominent gene involved in the production of fatty acid-derived flavor volatiles in tomato. *Mol. Plant* 11, 1147–1165. doi: 10.1016/j.molp.2018.06.003

Gasser, C. S., and Simon, M. K. (2011). Seed dispersal: same gene, different organs. *Curr. Biol.* 21, R546–R548.

Gil-Humanes, J., Wang, Y., Liang, Z., Shan, Q., Ozuna, C. V., Sánchez-León, S., et al. (2017). High-efficiency gene targeting in hexaploid wheat using DNA replicons and CRISPR/Cas9. *Plant J.* 89, 1251–1262. doi: 10.1111/tpj.13446

Gilmore, I. S., Heiles, S., and Pieterse, C. L. (2019). Metabolic imaging at the single-cell scale: recent advances in mass spectrometry imaging. *Annu. Rev. Anal. Chem.* 12, 201–224. doi: 10.1146/annurev-anchem-061318-115516

Girin, T., Paicu, T., Stephenson, P., Fuentes, S., Körner, E., O'Brien, M., et al. (2011). INDEHISCENT and SPATULA interact to specify carpel and valve margin tissue and thus promote seed dispersal in *Arabidopsis*. *Plant Cell* 23, 3641–3653. doi: 10.1105/tpc.111.090944

Gong, L., Chen, W., Gao, Y., Liu, X., Zhang, H., Xu, C., et al. (2013). Genetic analysis of the metabolome exemplified using a rice population. *Proc. Natl. Acad. Sci. U.S.A.* 110, 20320–20325. doi: 10.1073/pnas.1319681110

González, J. F., Degrassi, G., Devescovi, G., De Vleesschauwer, D., Höfte, M., Myers, M. P., et al. (2012). A proteomic study of *Xanthomonas oryzae* pv *oryzae* in rice xylem sap. *J. Proteomics* 75, 5911–5919. doi: 10.1016/j.jprot.2012.07.019

Gu, B., Zhou, T., Luo, J., Liu, H., Wang, Y., Shangguan, Y., et al. (2015). An-2 encodes a cytokinin synthesis enzyme that regulates awn length and grain production in rice. *Mol. Plant* 8, 1635–1650. doi: 10.1016/j.molp.2015.08.001

Gu, T., Jia, S., Huang, X., Wang, L., Fu, W., Huo, G., et al. (2019). Transcriptome and hormone analyses provide insights into hormonal regulation in strawberry ripening. *Planta* 250, 145–162. doi: 10.1007/s00425-019-03155-w

Gubler, F., Chandler, P. M., White, R. G., Llewellyn, D. J., and Jacobsen, J. V. (2002). Gibberellin signaling in barley aleurone cells. control of SLN1 and GAMYB expression. *Plant Physiol.* 129, 191–200. doi: 10.1104/pp.010918

Han, Q., Kang, G., and Guo, T. (2013). Proteomic analysis of spring freeze-stress responsive proteins in leaves of bread wheat (*Triticum aestivum* L.). *Plant Physiol. Biochem.* 63, 236–244. doi: 10.1016/j.plaphy.2012.12.002

Harberd, N. P. (2003). Relieving DELLA restraint. *Science* 299, 1853–1854. doi: 10.1126/science.1083217

Harberd, N. P., King, K. E., Carol, P., Cowling, R. J., Peng, J., and Richards, D. E. (1998). Gibberellin: inhibitor of an inhibitor of…? *Bioessays* 20, 1001–1008. doi: 10.1002/(sici)1521-1878(199812)20:12<1001::aid-bies6>3.0.co;2-o

Hattori, Y., Nagai, K., Furukawa, S., Song, X. J., Kawano, R., and Sakakibara, H. (2009). The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. *Nature* 460, 1026–1030. doi: 10.1038/nature08258

Hayut, S. F., Bessudo, C. M., and Levy, A. A. (2017). Targeted recombination between homologous chromosomes for precise breeding in tomato. *Nat. Commun.* 8:15605.

He, Y., Fu, Y., Hu, D., Wei, D., and Qian, W. (2018). QTL mapping of seed glucosinolate content responsible for environment in *Brassica napus*. *Front. Plant Sci.* 9:891. doi: 10.3389/fpls.2018.00891

Hedden, P., and Phillips, A. L. (2000). Gibberellin metabolism: new insights revealed by the genes. *Trends Plant Sci.* 5, 523–530. doi: 10.1016/s1360-1385(00)01790-8

Hickey, L. T., Hafeez, A. N., Robinson, H., Jackson, S. A., Leal-Bertioli, S. C., Tester, M., et al. (2019). Breeding crops to feed 10 billion. *Nat. Biotechnol.* 37, 744–754. doi: 10.1038/s41587-019-0152-9

Hill, C. B., Taylor, J. D., Edwards, J., Mather, D., Bacic, A., Langridge, P., et al. (2013). Whole-genome mapping of agronomic and metabolic traits to identify novel quantitative trait loci in bread wheat grown in a water-limited environment. *Plant Physiol.* 162, 1266–1281. doi: 10.1104/pp.113.217851

Hill, C. B., Taylor, J. D., Edwards, J., Mather, D., Langridge, P., Bacic, A., et al. (2015). Detection of QTL for metabolic and agronomic traits in wheat with adjustments for variation at genetic loci that affect plant phenology. *Plant Sci.* 233, 143–154. doi: 10.1016/j.plantsci.2015.01.008

Hooley, R. (1994). Gibberellins: perception, transduction and responses. *Plant Mol. Biol.* 26, 1529–1555. doi: 10.1007/bf00016489

Hu, B., Li, D., Liu, X., Qi, J., Gao, D., Zhao, S., et al. (2017). Engineering non-transgenic gynoecious cucumber using an improved transformation protocol and optimized CRISPR/Cas9 system. *Mol. Plant* 10, 1575–1578. doi: 10.1016/j.molp.2017.09.005

Hu, J., Rampitsch, C., and Bykova, N. V. (2015). Advances in plant proteomics toward improvement of crop productivity and stress resistance. *Front. Plant Sci.* 6:209. doi: 10.3389/fpls.2015.00209

Hu, X., Cui, Y., Dong, G., Feng, A., Wang, D., Zhao, C., et al. (2019). Using CRISPR-Cas9 to generate semi-dwarf rice lines in elite landraces. *Sci. Rep.* 9:19096.

Hua, L., Wang, D. R., Tan, L., Fu, Y., Liu, F., Xiao, L., et al. (2015). LABA1, a domestication gene associated with long, barbed awns in wild rice. *Plant Cell* 27, 1875–1888. doi: 10.1105/tpc.15.00260

Huang, C., Sun, H., Xu, D., Chen, Q., Liang, Y., Wang, X., et al. (2018). ZmCCT9 enhances maize adaptation to higher latitudes. *Proc. Natl. Acad. Sci. U.S.A.* 115, E334–E341.

Hummel, A. W., Chauhan, R. D., Cermak, T., Mutka, A. M., Vijayaraghavan, A., Boyher, A., et al. (2018). Allele exchange at the EPSPS locus confers glyphosate tolerance in cassava. *Plant Biotechnol. J.* 16, 1275–1282. doi: 10.1111/pbi.12868

Ikeda, A., Ueguchi-Tanaka, M., Sonoda, Y., Kitano, H., Koshioka, M., Futsuhara, Y., et al. (2001). slender rice, a constitutive gibberellin response mutant, is caused by a null mutation of the SLR1 gene, an ortholog of the height-regulating gene GAI/RGA/RHT/D8. *Plant Cell* 13, 999–1010. doi: 10.2307/3871359

Illa-Berenguer, E., Van Houten, J., Huang, Z., and van der Knaap, E. (2015). Rapid and reliable identification of tomato fruit weight and locule number loci by QTL-seq. *Theor. Appl. Genet.* 128, 1329–1342. doi: 10.1007/s00122-015-2509-x

Iquebal, M. A., Sharma, P., Jasrotia, R. S., Jaiswal, S., Kaur, A., Saroha, M., et al. (2019). RNAseq analysis reveals drought-responsive molecular pathways with candidate genes and putative molecular markers in root tissue of wheat. *Sci. Rep.* 9:13917.

Itoh, H., Ueguchi-Tanaka, M., Sato, Y., Ashikari, M., and Matsuoka, M. (2002). The gibberellin signaling pathway is regulated by the appearance and disappearance of slender rice1 in nuclei. *Plant Cell* 14, 57–70. doi: 10.1105/tpc.010319

Jasinski, S., Tattersall, A., Piazza, P., Hay, A., Martinez-Garcia, J. F., Schmitz, G., et al. (2008). Procera encodes a DELLA protein that mediates control of dissected leaf form in tomato. *Plant J.* 56, 603–612. doi: 10.1111/j.1365-313x.2008.03628.x

Jiang, L., Ma, X., Zhao, S., Tang, Y., Liu, F., Gu, P., et al. (2019). The APETALA2-like transcription factor SUPERNUMERARY BRACT controls rice seed shattering and seed size. *Plant Cell* 31, 17–36. doi: 10.1105/tpc.18.00304

Jiao, Y., Tausta, S. L., Gandotra, N., Sun, N., Liu, T., Clay, N. K., et al. (2009). A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nat. Genet.* 41, 258–263. doi: 10.1038/ng.282

Johnsson, M., Gaynor, R. C., Jenko, J., Gorjanc, G., de Koning, D. J., and Hickey, J. M. (2019). Removal of alleles by genome editing (RAGE) against deleterious load. *Genet. Sel.* 51, 1–18.

Kadambari, G., Vemireddy, L. R., Srividhya, A., Nagireddy, R., Jena, S. S., Gandikota, M., et al. (2018). QTL-Seq-based genetic analysis identifies a major genomic region governing dwarfness in rice (*Oryza sativa* L.). *Plant Cell Rep.* 37, 677–687. doi: 10.1007/s00299-018-2260-2

Kang, G., Li, G., Wang, L., Wei, L., Yang, Y., Wang, P., et al. (2015). Hg-responsive proteins identified in wheat seedlings using iTRAQ analysis and the role of ABA in Hg stress. *J. Proteome Res.* 14, 249–267. doi: 10.1021/pr5006873

Kanno, Y., Jikumaru, Y., Hanada, A., Nambara, E., Abrams, S. R., Kamiya, Y., et al. (2010). Comprehensive hormone profiling in developing *Arabidopsis* seeds: examination of the site of ABA biosynthesis, ABA transport and hormone interactions. *Plant Cell Physiol.* 51, 1988–2001. doi: 10.1093/pcp/pcq158

Karimi, E., Jaafar, H. Z., and Ahmad, S. (2011). Phenolics and flavonoids profiling and antioxidant activity of three varieties of Malaysian indigenous medicinal herb *Labisia pumila* Benth. *J. Med. Plant Res.* 5, 1200–1206.

Khalil-Ur-Rehman, M., Sun, L., Li, C. X., Faheem, M., Wang, W., and Tao, J. M. (2017). Comparative RNA-seq based transcriptomic analysis of bud dormancy in grape. *BMC Plant Biol.* 17:18. doi: 10.1186/s12870-016-0960-8

Khan, S. A., Chibon, P. Y., de Vos, R. C., Schipper, B. A., Walraven, E., and Beekwilder, J. (2012). Genetic analysis of metabolites in apple fruits indicates an mQTL hotspot for phenolic compounds on linkage group 16. *J. Exp. Bot.* 63, 2895–2908. doi: 10.1093/jxb/err464

Kim, D., Alptekin, B., and Budak, H. (2018). CRISPR/Cas9 genome editing in wheat. *Funct. Integr.* 18, 31–41. doi: 10.1007/s10142-017-0572-x

Kim, J. S., An, C. G., Park, J. S., Lim, Y. P., and Kim, S. (2016). Carotenoid profiling from 27 types of paprika (*Capsicum annuum L.*) with different colors, shapes, and cultivation methods. *Food Chem.* 201, 64–71. doi: 10.1016/j.foodchem.2016.01.041

Kim, S. T., Kim, S. G., Hwang, D. H., Kang, S. Y., Kim, H. J., Lee, B. H., et al. (2004). Proteomic analysis of pathogen-responsive proteins from rice leaves induced by rice blast fungus, *Magnaporthe grisea*. *Proteomics* 4, 3569–3578. doi: 10.1002/pmic.200400999

Kiszonas, A. M., and Morris, C. F. (2018). Wheat breeding for quality: a historical review. *Cereal Chem.* 95, 17–34.

Knoch, D., Riewe, D., Meyer, R. C., Boudichevskaia, A., Schmidt, R., and Altmann, T. (2017). Genetic dissection of metabolite variation in *Arabidopsis* seeds: evidence for mQTL hotspots and a master regulatory locus of seed metabolism. *J. Exp. Bot.* 68, 1655–1667. doi: 10.1093/jxb/erx049

Kojima, K., Andou, D., and Ito, M., (2021). Plant hormone changes in growing small watermelon fruit. *Hort. J.* UTD-209. doi: 10.2503/hortj.UTD-209

Kojima, M., Kamada-Nobusada, T., Komatsu, H., Takei, K., Kuroha, T., Mizutani, M., et al. (2009). Highly sensitive and high-throughput analysis of plant hormones using MS-probe modification and liquid chromatography–tandem mass spectrometry: an application for hormone profiling in *Oryza sativa*. *Plant Cell Physiol.* 50, 1201–1214. doi: 10.1093/pcp/pcp057

Koller, A., Washburn, M. P., Lange, B. M., Andon, N. L., Deciu, C., Haynes, P. A., et al. (2002). Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. U.S.A.* 99, 11969–11974. doi: 10.1073/pnas.172183199

Komatsuda, T., Pourkheirandish, M., He, C., Azhaguvel, P., Kanamori, H., Perovic, D. et al. (2007). Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1424–1429. doi: 10.1073/pnas.0608580104

Konishi, S., Izawa, T., Lin, S. Y., Ebana, K., Fukuta, Y., Sasaki, T., et al. (2006). An SNP caused loss of seed shattering during rice domestication. *Science* 312, 1392–1396. doi: 10.1126/science.1126410

Kosová, K., Vítámvás, P., Planchon, S., Renaut, J., Vanková, R., and Prášil, I. T. (2013). Proteome analysis of cold response in spring and winter wheat (*Triticum aestivum*) crowns reveals similarities in stress adaptation and differences in regulatory processes between the growth habits. *J. Proteome Res.* 12, 4830–4845. doi: 10.1021/pr400600g

Kretzschmar, T., Pelayo, M. A. F., Trijatmiko, K. R., Gabunada, L. F. M., Alam, R., Jimenez, R. R. et al. (2015). A trehalose-6-phosphate phosphatase enhances anaerobic germination tolerance in rice. *Nat. Plants* 1, 1–5.

Krishnan, H. B., Natarajan, S. S., Bennett, J. O., and Sicher, R. C. (2011). Protein and metabolite composition of xylem sap from field-grown soybeans (Glycine max). *Planta* 233, 921–931. doi: 10.1007/s00425-011-1352-9

Kudapa, H., Garg, V., Chitikineni, A., and Varshney, R. K. (2018). The RNA-seq-based high resolution gene expression atlas of chickpea (*Cicer arietinum L.*) reveals dynamic spatio-temporal changes associated with growth and development. *Plant Cell Environ.* 41, 2209–2225.

Kumar, A., Bimolata, W., Kannan, M., Kirti, P. B., Qureshi, I. A., and Ghazi, I. A. (2015). Comparative proteomics reveals differential induction of both biotic and abiotic stress response associated proteins in rice during *Xanthomonas oryzae* pv. oryzae infection. *Funct. Integr. Genomic* 15, 425–437. doi: 10.1007/s10142-014-0431-y

Kumar, R., Bohra, A., Pandey, A. K., Pandey, M. K., and Kumar, A. (2017). Metabolomics for plant improvement: status and prospects. *Front. Plant Sci.* 8:1302. doi: 10.3389/fpls.2017.01302

Kumar, R., Janila, P., Vishwakarma, M. K., Khan, A. W., Manohar, S. S., Gangurde, S. S. et al. (2020). Whole-genome resequencingbased QTL-seq identified candidate genes and molecular markers for fresh seed dormancy in groundnut. *Plant Biotechnol. J.* 18, 992–1003. doi: 10.1111/pbi.13266

Kumar, R., Tamboli, V., Sharma, R., and Sreelakshmi, Y. (2018). NAC-NOR mutations in tomato Penjar accessions attenuate multiple metabolic processes and prolong the fruit shelf life. *Food Chem.* 259, 234–244. doi: 10.1016/j.foodchem.2018.03.135

Kuroha, T., Nagai, K., Gamuyao, R., Wang, D. R., Furuta, T., Nakamori, M., et al. (2018). Ethylene-gibberellin signaling underlies adaptation of rice to periodic flooding. *Science* 361, 181–186. doi: 10.1126/science.aat1577

Lacchini, E., Kiegle, E., Castellani, M., Adam, H., Jouannic, S., Gregis, V., et al. (2020). CRISPR-mediated accelerated domestication of African rice landraces. *PLoS One* 15:e0229782. doi: 10.1371/journal.pone.0229782

Lei, L., Zheng, H., Bi, Y., Yang, L., Liu, H., Wang, J., et al. (2020). Identification of a Major QTL and candidate gene analysis of salt tolerance at the bud burst stage in rice (*Oryza sativa L.*) using QTL-seq and RNA-seq. *Rice* 13:55.

Lemmon, Z. H., Reem, N. T., Dalrymple, J., Soyk, S., Swartwood, K. E., Rodriguez-Leal, D., et al. (2018). Rapid improvement of domestication traits in an orphan crop by genome editing. *Nat. Plants* 4, 766–770. doi: 10.1038/s41477-018-0259-x

Lenser, T., and Theißen, G. (2013). Molecular mechanisms involved in convergent crop domestication. *Trends Plant Sci.* 18, 704–714. doi: 10.1016/j.tplants.2013.08.007

Levina, A. V., Hoekenga, O., Gordin, M., Broeckling, C., and De Jong, W. S. (2020). Genetic analysis of potato tuber metabolite composition: genome-wide

association studies applied to a non-targeted metabolome. *Crop Sci.* 61, 591-603. doi: 10.1002/csc2.20398

Li, B., Zhang, Y., Mohammadi, S. A., Huai, D., Zhou, Y., and Kliebenstein, D. J. (2016). An integrative genetic study of rice metabolism, growth and stochastic variation reveals potential C/N partitioning loci. *Sci. Rep.* 6:30143.

Li, C., Zhou, A., and Sang, T. (2006). Rice domestication by reducing shattering. *Science* 311, 1936–1939. doi: 10.1126/science.1123604

Li, H., Peng, Z., Yang, X., Wang, W., Fu, J., Wang, J., et al. (2013). Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat. Genet.* 45, 43–50. doi: 10.1038/ng.2484

Li, K., Wen, W., Alseekh, S., Yang, X., Guo, H., and Li, W. (2019). Large-scale metabolite quantitative trait locus analysis provides new insights for high-quality maize improvement. *Plant J.* 99, 216–230.

Li, T., Yang, X., Yu, Y., Si, X., Zhai, X., Zhang, H., et al. (2018). Domestication of wild tomato is accelerated by genome editing. *Nat. Biotechnol.* 36, 1160–1163. doi: 10.1038/nbt.4273

Li, T., Yun, Z., Wu, Q., Qu, H., Duan, X., and Jiang, Y. (2019). Combination of transcriptomic, proteomic, and metabolomic analysis reveals the ripening mechanism of banana pulp. *Biomolecules* 9:523. doi: 10.3390/biom9100523

Liang, Z., and Schnable, J. C. (2016). RNA-seq based analysis of population structure within the maize inbred B73. *PLoS One* 11:e0157942. doi: 10.1371/journal.pone.0157942

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., and Franklin, L. D. (2010). An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *Plant J.* 63, 86–99.

Lifschitz, E., Eviatar, T., Rozman, A., Shalit, A., Goldshmidt, A., Amsellem, Z. et al. (2006). The tomato FT ortholog triggers systemic signals that regulate growth and flowering and substitute for diverse environmental stimuli. *Proc. Natl. Acad. Sci. U.S.A.* 103, 6398–6403. doi: 10.1073/pnas.0601620103

Liljegren, S. J., Ditta, G. S., Eshed, Y., Savidge, B., Bowman, J. L., and Yanofsky, M. F. (2000). SHATTERPROOF MADS-box genes control seed dispersal in *Arabidopsis*. *Nature* 404, 766–770. doi: 10.1038/35008089

Liljegren, S. J., Roeder, A. H., Kempin, S. A., Gremski, K., Østergaard, L., Guimil, S., et al. (2004). Control of fruit patterning in *Arabidopsis* by INDEHISCENT. *Cell J.* 116, 843–853. doi: 10.1016/s0092-8674(04)00217-x

Lin, Z., Li, X., Shannon, L. M., Yeh, C. T., Wang, M. L., Bai, G., et al. (2012). Parallel domestication of the shattering1 genes in cereals. *Nat. Genet.* 44, 720–724. doi: 10.1038/ng.2281

Lipka, A. E., Gore, M. A., Magallanes-Lundback, M., Mesberg, A., Lin, H., Tiede, T., et al. (2013). Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. *G3* 3, 1287–1299. doi: 10.1534/g3.113.006148

Lippman, Z. B., Cohen, O., Alvarez, J. P., Abu-Abied, M., Pekker, I., Paran, I. et al. (2008). The making of a compound inflorescence in tomato and related nightshades. *PLoS Biol.* 6:288. doi: 10.1371/journal.pbio.0060288

Lisec, J., Steinfath, M., Meyer, R. C., Selbig, J., Melchinger, A. E., Willmitzer, L., et al. (2009). Identification of heterotic metabolite QTL in *Arabidopsis thaliana* RIL and IL populations. *Plant J.* 59, 777–788. doi: 10.1111/j.1365-313x.2009.03910.x

Liu, J., Chen, J., Zheng, X., Wu, F., Lin, Q., Heng, Y., et al. (2017). GW5 acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. *Nat. Plants* 3:17043.

Liu, J., Van Eck, J., Cong, B., and Tanksley, S. D. (2002). A new class of regulatory genes underlying the cause of pear-shaped tomato fruit. *Proc. Natl. Acad. Sci. U.S.A.* 99, 13302–13306. doi: 10.1073/pnas.162485999

Liu, X., Sheng, J., and Curtiss, R. III (2011). Fatty acid production in genetically modified cyanobacteria. *Proc. Natl. Acad. Sci. U.S.A.* 108, 6899–6904. doi: 10.1073/pnas.1103014108

Liu, Y., Pan, T., Tang, Y., Zhuang, Y., Liu, Z., Li, P., et al. (2020). Proteomic Analysis of rice subjected to low light stress and overexpression of OsGAPB increases the stress tolerance. *Rice* 13:30.

Liu, Z., Alseekh, S., Brotman, Y., Zheng, Y., Fei, Z., Tieman, D. M., et al. (2016). Identification of a *Solanum pennellii* chromosome 4 fruit flavor and nutritional quality-associated metabolite QTL. *Front. Plant Sci.* 7:1671. doi: 10.3389/fpls.2016.01671

Lu, X., Li, Q. T., Xiong, Q., Li, W., Bi, Y. D., Lai, Y. C., et al. (2016). The transcriptomic signature of developing soybean seeds reveals the genetic basis of seed trait adaptation during domestication. *Plant J.* 86, 530–544. doi: 10.1111/tpj.13181

Luo, H., Pandey, M. K., Khan, A. W., Guo, J., Wu, B., Cai, Y., et al. (2019). Discovery of genomic regions and candidate genes controlling shelling percentage using QTL-seq approach in cultivated peanut (*Arachis hypogaea L.*). *Plant Biotechnol. J.* 17, 1248–1260. doi: 10.1111/pbi.13050

Ma, J., Chen, J., Wang, M., Ren, Y., Wang, S., Lei, C., et al. (2018). Disruption of OsSEC3A increases the content of salicylic acid and induces plant defense responses in rice. *J. Exp. Bot.* 69, 1051–1064. doi: 10.1093/jxb/erx458

Macklin, A., Khan, S., and Kislinger, T. (2020). Recent advances in mass spectrometry based clinical proteomics: applications to cancer research. *Clin. Proteomics* 17:17.

Mangul, S., Martin, L. S., Hill, B. L., Lam, A. K. M., Distler, M. G., Zelikovsky, A., et al. (2019). Systematic benchmarking of omics computational tools. *Nat. Commun.* 10:1393.

Mao, L., Begum, D., Chuang, H. W., Budiman, M. A., Szymkowiak, E. J., Irish, E. E., et al. (2000). JOINTLESS is a MADS-box gene controlling tomato flower abscission zone development. *Nature* 406, 910–913. doi: 10.1038/35022611

Mao, X., Zheng, Y., Xiao, K., Wei, Y., Zhu, Y., Cai, Q., et al. (2018). OsPRX2 contributes to stomatal closure and improves potassium deficiency tolerance in rice. *Biochem. Biophys. Res. Commun.* 495, 461–467. doi: 10.1016/j.bbrc.2017.11.045

Marsch-Martínez, N., Ramos-Cruz, D., Irepan Reyes-Olalde, J., Lozano-Sotomayor, P., Zúñiga-Mayo, V. M., and de Folter, S. (2012). The role of cytokinin during *Arabidopsis* gynoecia and fruit morphogenesis and patterning. *Plant J.* 72, 222–234. doi: 10.1111/j.1365-313x.2012.05062.x

Martín-Trillo, M., Grandío, E. G., Serra, F., Marcel, F., Rodríguez-Buey, M. L., Schmitz, G., et al. (2011). Role of tomato BRANCHED1-like genes in the control of shoot branching. *Plant J.* 67, 701–714. doi: 10.1111/j.1365-313x.2011.04629.x

Matros, A., Liu, G., Hartmann, A., Jiang, Y., Zhao, Y., Wang, H., et al. (2017). Genome–metabolite associations revealed low heritability, high genetic complexity, and causal relations for leaf metabolites in winter wheat (*Triticum aestivum*). *J. Exp. Bot.* 68, 415–428.

Matsuda, F., Nakabayashi, R., Yang, Z., Okazaki, Y., Yonemaru, J., Ebana, K., et al. (2015). Metabolome-genome-wide association study (mGWAS) dissects genetic architecture for generating natural variation in rice secondary metabolism. *Plant J.* 81, 13–23. doi: 10.1111/tpj.12681

Matsuda, F., Okazaki, Y., Oikawa, A., Kusano, M., Nakabayashi, R., Kikuchi, J., et al. (2012). Dissection of genotype–phenotype associations in rice grains using metabolome quantitative trait loci analysis. *Plant J.* 70, 624–636. doi: 10.1111/j.1365-313x.2012.04903.x

Matsuura, T., Mori, I. C., Himi, E., and Hirayama, T. (2019). Plant hormone profiling in developing seeds of common wheat (*Triticum aestivum* L.). *Breed. Sci.* 69, 601–610. doi: 10.1270/jsbbs.19030

Meng, Y., Hou, Y., Wang, H., Ji, R., Liu, B., Wen, J., et al. (2017). Targeted mutagenesis by CRISPR/Cas9 system in the model legume medicago truncatula. *Plant Cell Rep.* 36, 371–374. doi: 10.1007/s00299-016-2069-9

Meyer, R. S., DuVal, A. E., and Jensen, H. R. (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *New Phytol.* 196, 29–48. doi: 10.1111/j.1469-8137.2012.04253.x

Mibei, E. K., Ambuko, J., Giovannoni, J. J., Onyango, A. N., and Owino, W. O. (2017). Carotenoid profiling of the leaves of selected African eggplant accessions subjected to drought stress. *Food Sci. Nutr.* 5, 113–122. doi: 10.1002/fsn3.370

Mieulet, D., Aubert, G., Bres, C., Klein, A., Droc, G., Vieille, E., et al. (2018). Unleashing meiotic crossovers in crops. *Nat. Plants* 4, 1010–1016. doi: 10.1038/s41477-018-0311-x

Miro, B., and Ismail, A. M. (2013). Tolerance of anaerobic conditions caused by flooding during germination and early growth in rice (*Oryza sativa* L.). *Front. Plant Sci.* 4:269. doi: 10.3389/fpls.2013.00269

Misra, B. B., Langefeld, C., Olivier, M., and Cox, L. A. (2019). Integrated omics: tools, advances and future approaches. *J. Mol. Endocrinol.* 62, R21–R45.

Monna, L., Kitazawa, N., Yoshino, R., Suzuki, J., Masuda, H., Maehara, Y., et al. (2002). Positional cloning of rice semidwarfing gene, sd-1: rice "green revolution gene" encodes a mutant enzyme involved in gibberellin synthesis. *DNA Res.* 9, 11–17. doi: 10.1093/dnares/9.1.11

Morineau, C., Bellec, Y., Tellier, F., Gissot, L., Kelemen, Z., Nogué, F., et al. (2017). Selective gene dosage by CRISPR-Cas9 genome editing in hexaploid *Camelina sativa*. *Plant Biotechnol. J.* 15, 729–739. doi: 10.1111/pbi.12671

Müller, M., and Munné-Bosch, S. (2011). Rapid and sensitive hormonal profiling of complex plant samples by liquid chromatography coupled to electrospray ionization tandem mass spectrometry. *Plant Methods* 7, 1–11.

Multani, D. S., Briggs, S. P., Chamberlin, M. A., Blakeslee, J. J., Murphy, A. S., and Johal, G. S. (2003). Loss of an MDR transporter in compact stalks of maize br2 and sorghum dw3 mutants. *Science* 302, 81–84. doi: 10.1126/science.1086072

Nagai, K., Mori, Y., Ishikawa, S., Furuta, T., Gamuyao, R., Niimi, Y., et al. (2020). Antagonistic regulation of the gibberellic acid response during stem growth in rice. *Nature* 584, 109–114. doi: 10.1038/s41586-020-2501-8

Nandi, S., Subudhi, P. K., Senadhira, D., Manigbas, N. L., Sen-Mandi, S., and Huang, N. (1997). Mapping QTLs for submergence tolerance in rice by AFLP analysis and selective genotyping. *Mol. Gen. Genet.* 255, 1–8. doi: 10.1007/s004380050468

Nayak, S. N., Hebbal, V., Soni, P., Kumar, R., Pandey, A. K., Wan, L., et al. (2019). *Groundnut Kernel Transcriptome. In Comprehensive Foodomics*. Amsterdam: Elsevier.

Neeraja, C. N., Maghirang-Rodriguez, R., Pamplona, A., Heuer, S., Collard, B. C., Septiningsih, E. M., et al. (2007). A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. *Theor. Appl. Genet.* 115, 767–776. doi: 10.1007/s00122-007-0607-0

Nieves-Cordones, M., Mohamed, S., Tanoi, K., Kobayashi, N. I., Takagi, K., Vernet, A., et al. (2017). Production of low-Cs+ rice plants by inactivation of the K+ transporter Os HAK 1 with the CRISPR-Cas system. *Plant J.* 92, 43–56. doi: 10.1111/tpj.13632

Nonaka, S., Arai, C., Takayama, M., Matsukura, C., and Ezura, H. (2017). Efficient increase of γ-aminobutyric acid (GABA) content in tomato fruits by targeted mutagenesis. *Sci. Rep.* 7:7057.

Novák, O., Hauserová, E., Amakorová, P., Doležal, K., and Strnad, M. (2008). Cytokinin profiling in plant tissues using ultra-performance liquid chromatography–electrospray tandem mass spectrometry. *Phytochemistry* 69, 2214–2224. doi: 10.1016/j.phytochem.2008.04.022

Nubankoh, P., Wanchana, S., Saensuk, C., Ruanjaichon, V., Cheabu, S., Vanavichit, A., et al. (2020). QTL-seq reveals genomic regions associated with spikelet fertility in response to a high temperature in rice (*Oryza sativa* L.). *Plant Cell Rep.* 39, 149–162. doi: 10.1007/s00299-019-02477-z

Ogawa, M., Kusano, T., Katsumi, M., and Sano, H. (2000). Rice gibberellin-insensitive gene homolog, OsGAI, encodes a nuclear-localized protein capable of gene activation at transcriptional level. *Gene* 245, 21–29. doi: 10.1016/s0378-1119(00)00018-4

Ogiso-Tanaka, E., Tanaka, T., Tanaka, K., Nonoue, Y., Sasaki, T., Fushimi, E., et al. (2017). Detection of novel QTLs qDTH4. 5 and qDTH6. 3, which confer late heading under short-day conditions, by SSR marker-based and QTL-seq analysis. *Breed. Sci.* 67, 101–109. doi: 10.1270/jsbbs.16096

Okuzaki, A., Ogawa, T., Koizuka, C., Kaneko, K., Inaba, M., Imamura, J., et al. (2018). CRISPR/Cas9-mediated genome editing of the fatty acid desaturase 2 gene in Brassica napus. *Plant Physiol. Biochem.* 131, 63–69. doi: 10.1016/j.plaphy.2018.04.025

Oladosu, Y., Rafii, M. Y., Arolu, F., Chukwu, S. C., Muhammad, I., Kareem, I., et al. (2020). Submergence tolerance in rice: review of mechanism, breeding and, future prospects. *Sustainability* 12:1632. doi: 10.3390/su12041632

Olsen, K. M., and Wendel, J. F. (2013). A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* 64, 47–70. doi: 10.1146/annurev-arplant-050312-120048

Olszewski, N., Sun, T. P., and Gubler, F. (2002). Gibberellin signalling: biosynthesis, catabolism, and response pathways. *Plant Cell* 14(Suppl. 1), S61–S80.

Ortiz-Ramírez, C., Hernandez-Coronado, M., Thamm, A., Catarino, B., Wang, M., and Dolan, L. (2016). A transcriptome atlas of *Physcomitrella* patens provides insights into the evolution and development of land plants. *Mol. Plant* 9, 205–220. doi: 10.1016/j.molp.2015.12.002

Osorio, S., Alba, R., Damasceno, C. M., Lopez-Casado, G., Lohse, M., Zanor, M. I., et al. (2011). Systems biology of tomato fruit development: combined transcript, protein, and metabolite analysis of tomato transcription factor (nor, rin) and ethylene receptor (Nr) mutants reveals novel regulatory interactions. *Plant Physiol.* 157, 405–425. doi: 10.1104/pp.111.175463

Østergaard, L., Kempin, S. A., Bies, D., Klee, H. J., and Yanofsky, M. F. (2006). Pod shatter-resistant Brassica fruit produced by ectopic expression of the FRUITFULL gene. *Plant Biotechnol. J.* 4, 45–51. doi: 10.1111/j.1467-7652.2005.00156.x

Owens, B. F., Lipka, A. E., Magallanes-Lundback, M., Tiede, T., Diepenbrock, C. H., Kandianis, C. B., et al. (2014). A foundation for provitamin A biofortification of maize: genome-wide association and genomic prediction models of carotenoid levels. *Genetics* 198, 1699–1716. doi: 10.1534/genetics.114.169979

Palaisa, K., Morgante, M., Tingey, S., and Rafalski, A. (2004). Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9885–9890. doi: 10.1073/pnas.0307839101

Pandey, M. K., Pandey, A. K., Kumar, R., Nwosu, C. V., Guo, B., Wright, G. C., et al. (2020). Translational genomics for achieving higher genetic gains in groundnut. *Theore. Appl. Genet.* 133, 1679–1702. doi: 10.1007/s00122-020-03592-2

Peivastegan, B., Hadizadeh, I., Nykyri, J., Nielsen, K. L., Somervuo, P., Sipari, N., et al. (2019). Effect of wet storage conditions on potato tuber transcriptome, phytohormones and growth. *BMC Plant Biol.* 19:262. doi: 10.1186/s12870-019-1875-y

Peng, J., Carol, P., Richards, D. E., King, K. E., Cowling, R. J., Murphy, G. P., et al. (1997). The Arabidopsis GAI gene defines a signalling pathway that negatively regulates gibberellin responses. *Genes Dev.* 11, 3194–3205. doi: 10.1101/gad.11.23.3194

Peng, J., Richards, D. E., Hartley, N. M., Murphy, G. P., Devos, K. M., Flintham, J. E., et al. (1999). 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* 400, 256–261. doi: 10.1038/22307

Perez-Fons, L., Wells, T., Corol, D. I., Ward, J. L., Gerrish, C., et al. (2014). A genome-wide metabolomic resource for tomato fruit from *Solanum pennellii*. *Sci. Rep.* 4:3859.

Piasecka, A., Sawikowska, A., Kuczyńska, A., Ogrodowicz, P., Mikołajczak, K., Krystkowiak, K., et al. (2017). Drought-related secondary metabolites of barley (*Hordeum vulgare* L.) leaves and their metabolomic quantitative trait loci. *Plant J.* 89, 898–913. doi: 10.1111/tpj.13430

Pnueli, L., Carmel-Goren, L., Hareven, D., Gutfinger, T., Alvarez, J., Ganal, M., et al. (1998). The SELF-PRUNING gene of tomato regulates vegetative to reproductive switching of sympodial meristems and is the ortholog of CEN and TFL1. *Development* 125, 1979–1989.

Pott, D. M., Vallarino, J. G., Cruz-Rus, E., Willmitzer, L., Sánchez-Sevilla, J. F., Amaya, I., et al. (2020). Genetic analysis of phenylpropanoids and antioxidant capacity in strawberry fruit reveals mQTL hotspots and candidate genes. *Sci. Rep.* 10:20197.

Powell, A. L., Nguyen, C. V., Hill, T., Cheng, K. L., Figueroa-Balderas, R., Aktas, H., et al. (2012). Uniform ripening encodes a Golden 2-like transcription factor regulating tomato fruit chloroplast development. *Science* 336, 1711–1715. doi: 10.1126/science.1222218

Praveen, A., Pandey, A., and Gupta, M. (2020). Protective role of nitric oxide on nitrogen-thiol metabolism and amino acids profiling during arsenic exposure in *Oryza sativa* L. *Ecotoxicology* 29, 825–836. doi: 10.1007/s10646-020-02250-z

Price, E. J., Bhattacharjee, R., Lopez-Montes, A., and Fraser, P. D. (2018). Carotenoid profiling of yams: clarity, comparisons and diversity. *Food Chem.* 259, 130–138. doi: 10.1016/j.foodchem.2018.03.066

Purugganan, M. D., and Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature* 457, 843–848. doi: 10.1038/nature07895

Pysh, L. D., Wysocka-Diller, J. W., Camilleri, C., Bouchez, D., and Benfey, P. N. (1999). The GRAS gene family in *Arabidopsis*: sequence characterization and basic expression analysis of the SCARECROW-LIKE genes. *Plant J.* 18, 111–119. doi: 10.1046/j.1365-313x.1999.00431.x

Qi, C., Jiang, H., Xiong, J., Yuan, B., and Feng, Y. (2019). On-line trapping/capillary hydrophilic-interaction liquid chromatography/mass spectrometry for sensitive determination of RNA modifications from human blood. *Chin. Chem. Lett.* 30, 553–557. doi: 10.1016/j.cclet.2018.11.029

Qi, J., Sun, P., Liao, D., Sun, T., Zhu, J., and Li, X. (2015). Transcriptomic analysis of American ginseng seeds during the dormancy release process by RNA-Seq. *PLoS One* 10:e0118558. doi: 10.1371/journal.pone.0118558

Qiao, A., Fang, X., Liu, S., Liu, H., Gao, P., and Luan, F. (2021). QTL-seq identifies major quantitative trait loci of stigma color in melon. *Hortic. Plant J.* doi: 10.1016/j.hpj.2021.01.004

Qin, Y., Cheng, P., Cheng, Y., Feng, Y., Huang, D., Huang, T., et al. (2018). QTL-Seq identified a major QTL for grain length and weight in rice using near isogenic F2 population. *Rice Sci.* 25, 121–131. doi: 10.1016/j.rsci.2018.04.001

Ramos, A., Fu, Y., Michael, V., and Meru, G. (2020). QTL-seq for identification of loci associated with resistance to *Phytophthora* crown rot in squash. *Sci. Rep.* 10:5326.

Razzaq, A., Sadia, B., Raza, A., Khalid Hameed, M., and Saleem, F. (2019). Metabolomics: a way forward for crop improvement. *Metabolites* 9:303. doi: 10.3390/metabo9120303

Reddy, M. M., and Ulaganathan, K. (2015). RNA-Seq analysis of urea nutrition responsive transcriptome of *Oryza sativa* elite indica cultivar RP Bio 226. *Genom. Data* 6, 112–113. doi: 10.1016/j.gdata.2015.08.025

Ren, J. L., Zhang, A. H., Kong, L., and Wang, X. J. (2018). Advances in mass spectrometry-based metabolomics for investigation of metabolites. *RSC Adv.* 8, 22335–22350. doi: 10.1039/c8ra01574k

Riedelsheimer, C., Lisec, J., Czedik-Eysenberg, A., Sulpice, R., Flis, A., Grieder, C., et al. (2012). Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. *Proc. Natl. Acad. Sci. U.S.A.* 109, 8872–8877. doi: 10.1073/pnas.1120813109

Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K., et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* 43, 1066–1073. doi: 10.1038/ng.952

Rodríguez-Leal, D., Lemmon, Z. H., Man, J., Bartlett, M. E., and Lippman, Z. B. (2017). Engineering quantitative trait variation for crop improvement by genome editing. *Cell* 171, 470–480. doi: 10.1016/j.cell.2017.08.030

Roldan, M. V. G., Périlleux, C., Morin, H., Huerga-Fernandez, S., Latrasse, D., Benhamed, M., et al. (2017). Natural and induced loss of function mutations in SlMBP21 MADS-box gene led to jointless-2 phenotype in tomato. *Sci. Rep.* 7:4402.

Ronen, G., Carmel-Goren, L., Zamir, D., and Hirschberg, J. (2000). An alternative pathway to β-carotene formation in plant chromoplasts discovered by map-based cloning of Beta and old-gold color mutations in tomato. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11102–11107. doi: 10.1073/pnas.190177497

Rowe, H. C., Hansen, B. G., Halkier, B. A., and Kliebenstein, D. J. (2008). Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. *Plant Cell* 20, 1199–1216. doi: 10.1105/tpc.108.058131

Ruan, S. L., Ma, H. S., Wang, S. H., Fu, Y. P., Xin, Y., Liu, W. Z., et al. (2011). Proteomic identification of OsCYP2 a rice cyclophilin that confers salt tolerance in rice (*Oryza sativa* L.) seedlings when overexpressed. *BMC Plant Biol.* 11:34. doi: 10.1186/1471-2229-11-34

Sasaki, A., Ashikari, M., Ueguchi-Tanaka, M., Itoh, H., Nishimura, A., Swapan, D., et al. (2002). A mutant gibberellin-synthesis gene in rice. *Nature* 416, 701–702. doi: 10.1038/416701a

Sauer, N. J., Narváez-Vásquez, J., Mozoruk, J., Miller, R. B., Warburg, Z. J., Woodward, M. J., et al. (2016). Oligonucleotide-mediated genome editing provides precision and function to engineered nucleases and antibiotics in plants. *Plant Physiol.* 170, 1917–1928. doi: 10.1104/pp.15.01696

Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., et al. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. *Plant Physiol.* 165, 1120–1132. doi: 10.1104/pp.114.241521

Schauer, N., Semel, Y., Balbo, I., Steinfath, M., Repsilber, D., Selbig, J., et al. (2008). Mode of inheritance of primary metabolic traits in tomato. *Plant Cell* 20, 509–523. doi: 10.1105/tpc.107.056523

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., et al. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nat. Biotechnol.* 24, 447–454. doi: 10.1038/nbt1192

Schindele, A., Dorn, A., and Puchta, H. (2020). CRISPR/Cas brings plant biology and breeding into the fast lane. *Curr. Opin. Biotechnol.* 61, 7–14. doi: 10.1016/j.copbio.2019.08.006

Schmidt, J., Blessing, F., Fimpler, L., and Wenzel, F. (2020). Nanopore sequencing in a clinical routine laboratory: challenges and opportunities. *Clin. Lab.* 66:1097.

Sekhon, R. S., Briskine, R., Hirsch, C. N., Myers, C. L., Springer, N. M., Buell, C. R., et al. (2013). Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays. *PLoS One* 8:e61005. doi: 10.1371/journal.pone.0061005

Septiningsih, E. M., Ignacio, J. C. I., Sendon, P. M., Sanchez, D. L., Ismail, A. M., and Mackill, D. J. (2013). QTL mapping and confirmation for tolerance of anaerobic conditions during germination derived from the rice landrace Ma-Zhan Red. *Theor. Appl. Genet.* 126, 1357–1366. doi: 10.1007/s00122-013-2057-1

Severin, A. J., Woody, J. L., Bolon, Y. T., Joseph, B., Diers, B. W., Farmer, A. D., et al. (2010). RNA-seq atlas of glycine max: a guide to the soybean transcriptome. *BMC Plant Biol.* 10:160. doi: 10.1186/1471-2229-10-160

Seymour, G. B., Chapman, N. H., Chew, B. L., and Rose, J. K. (2013). Regulation of ripening and opportunities for control in tomato and other fruits. *Plant Biotech. J.* 11, 269–278. doi: 10.1111/j.1467-7652.2012.00738.x

Shan, Q., Wang, Y., Li, J., and Gao, C. (2014). Genome editing in rice and wheat using the CRISPR/Cas system. *Nat. Protoc.* 9, 2395–2410. doi: 10.1038/nprot.2014.157

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., et al. (2013). Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat. Biotechnol.* 31, 686–688. doi: 10.1038/nbt.2650

Sharma, V., Gupta, P., Priscilla, K., SharanKumar, A., Hangargi, B., Veershetty, A., et al. (2021). metabolomics intervention towards better understanding of plant traits. *Cells* 10:346. doi: 10.3390/cells10020346

Shen, M., Broeckling, C. D., Chu, E. Y., Ziegler, G., Baxter, I. R., Prenni, J. E., et al. (2013). Leveraging non-targeted metabolite profiling via statistical genomics. *PLoS One* 8:e57667. doi: 10.1371/journal.pone.0057667

Shi, T., Zhu, A., Jia, J., Hu, X., Chen, J., Liu, W., et al. (2020). Metabolomics analysis and metabolite-agronomic trait associations using kernels of wheat (*Triticum aestivum*) recombinant inbred lines. *Plant J.* 103, 279–292. doi: 10.1111/tpj.14727

Shomura, A., Izawa, T., Ebana, K., Ebitani, T., Kanegae, H., Konishi, S., et al. (2008). Deletion in a gene associated with grain size increased yields during rice domestication. *Nat. Genet.* 40, 1023–1028. doi: 10.1038/ng.169

Si, L., Chen, J., Huang, X., Gong, H., Luo, J., Hou, Q., et al. (2016). OsSPL13 controls grain size in cultivated rice. *Nat. Genet.* 48, 447–456. doi: 10.1038/ng.3518

Siangliw, M., Toojinda, T., Tragoonrung, S., and Vanavichit, A. (2003). Thai jasmine rice carrying QTLch9 (Sub QTL) is submergence tolerant. *Ann. Bot.* 91, 255–261. doi: 10.1093/aob/mcf123

Sigmon, B., and Vollbrecht, E. (2010). Evidence of selection at the ramosa1 locus during maize domestication. *Mol. Ecol.* 19, 1296–1311. doi: 10.1111/j.1365-294x.2010.04562.x

Silverstone, A. L., and Sun, T. (2000). Gibberellins and the green revolution. *Trends Plant Sci.* 5, 1–2. doi: 10.1016/s1360-1385(99)01516-2

Simons, K. J., Fellers, J. P., Trick, H. N., Zhang, Z., Tai, Y. S., Gill, B. S., et al. (2006). Molecular characterization of the major wheat domestication gene Q. *Genetics* 172, 547–555. doi: 10.1534/genetics.105.044727

Singh, A., Dubey, P. K., Chaurasia, R., Dubey, R. K., Pandey, K. K., Singh, G. S., et al. (2019). Domesticating the undomesticated for global food and nutritional security: four steps. *Agron J.* 9:491. doi: 10.3390/agronomy9090491

Singh, V. K., Khan, A. W., Jaganathan, D., Thudi, M., Roorkiwal, M., Takagi, H., et al. (2016). QTL-seq for rapid identification of candidate genes for 100-seed weight and root/total plant dry weight ratio under rainfed conditions in chickpea. *Plant Biotechnol. J.* 14, 2110–2119. doi: 10.1111/pbi.12567

Sinha, P., Bajaj, P., Pazhamala, L., Nayak, S., Pandey, M. K., Chitikineni, A., et al. (2020). The *Arachis hypogaea* gene expression atlas (AhGEA) for accelerating translational research in cultivated groundnut. *Plant Biotechnol. J.*

Smýkal, P., Nelson, M. N., Berger, J. D., and Von Wettberg, E. J. (2018). The impact of genetic changes during crop domestication. *Agronomy* 8:119. doi: 10.3390/agronomy8070119

Song, J., Li, Z., Liu, Z., Guo, Y., and Qiu, L. J. (2017). Next-generation sequencing from bulked-segregant analysis accelerates the simultaneous identification of two qualitative genes in soybean. *Front. Plant Sci.* 8:919. doi: 10.3389/fpls.2017.00919

Sorefan, K., Girin, T., Liljegren, S. J., Ljung, K., Robles, P., Galván-Ampudia, C. S., et al. (2009). A regulated auxin minimum is required for seed dispersal in *Arabidopsis*. *Nature* 459, 583–586. doi: 10.1038/nature07875

Sosso, D., Luo, D., Li, Q. B., Sasse, J., Yang, J., Gendrot, G., et al. (2015). Seed filling in domesticated maize and rice depends on SWEET-mediated hexose transport. *Nat. Genet.* 47, 1489–1493. doi: 10.1038/ng.3422

Soyk, S., Müller, N. A., Park, S. J., Schmalenbach, I., Jiang, K., Hayama, R., et al. (2017). Variation in the flowering gene SELF PRUNING 5G promotes day-neutrality and early yield in tomato. *Nat. Genet.* 49, 162–168. doi: 10.1038/ng.3733

Spielmeyer, W., Ellis, M. H., and Chandler, P. M. (2002). Semidwarf (sd-1),"green revolution" rice, contains a defective gibberellin 20-oxidase gene. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9043–9048. doi: 10.1073/pnas.132266399

Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* 43, 1160–1163. doi: 10.1038/ng.942

Sugimoto, K., Takeuchi, Y., Ebana, K., Miyao, A., Hirochika, H., Hara, N., et al. (2010). Molecular cloning of Sdr4, a regulator involved in seed dormancy and domestication of rice. *Proc. Natl. Acad. Sci. U.S.A.* 107, 5792–5797. doi: 10.1073/pnas.0911965107

Sun, S., and Frelich, L. E. (2011). Flowering phenology and height growth pattern are associated with maximum plant height, relative growth rate and stem tissue mass density in herbaceous grassland species. *J. Ecol.* 99, 991–1000. doi: 10.1111/j.1365-2745.2011.01830.x

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

Sun, Y., Wang, F., Wang, N., Dong, Y., Liu, Q., Zhao, L., et al. (2013). Transcriptome exploration in *Leymus chinensis* under saline-alkaline treatment using 454 pyrosequencing. *PLoS One* 8:e53632. doi: 10.1371/journal.pone.0053632

Sweeney, M. T., Thomson, M. J., Pfeil, B. E., and McCouch, S. (2006). Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. *Plant Cell* 18, 283–294. doi: 10.1105/tpc.105.038430

Takagi, H., Uemura, A., Yaegashi, H., Tamiru, M., Abe, A., Mitsuoka, C., et al. (2013). MutMap-Gap: whole-genome resequencing of mutant F 2 progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene Pii. *New Phytol.* 200, 276–283. doi: 10.1111/nph.12369

Taketa, S., Amano, S., Tsujino, Y., Sato, T., Saisho, D., Kakeda, K., et al. (2008). Barley grain with adhering hulls is controlled by an ERF family transcription factor gene regulating a lipid biosynthesis pathway. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4062–4067. doi: 10.1073/pnas.0711034105

Talebi, A. F., Mohtashami, S. K., Tabatabaei, M., Tohidfar, M., Bagheri, A., Zeinalabedini, M., et al. (2013). Fatty acids profiling: a selective criterion for screening microalgae strains for biodiesel production. *Algal Res.* 2, 258–267. doi: 10.1016/j.algal.2013.04.003

Tang, H., Cuevas, H. E., Das, S., Sezen, U. U., Zhou, C., Guo, H., et al. (2013). Seed shattering in a wild *Sorghum* is conferred by a locus unrelated to domestication. *Proc. Natl. Acad. Sci.* 110, 15824–15829. doi: 10.1073/pnas.1305213110

Tang, Y. J., and Aristilde, L. (2020). Editorial overview: analytical biotechnology in the era of high-performance omics, synthetic biology, and machine learning. *Curr. Opin. Biotechnol.* 64, iii–vi. doi: 10.1016/j.copbio.2020.07.009

Templer, S. E., Ammon, A., Pscheidt, D., Ciobotea, O., Schuy, C., McCollum, C., et al. (2017). Metabolite profiling of barley flag leaves under drought and combined heat and drought stress reveals metabolic QTLs for metabolites associated with antioxidant defense. *J. Exp. Bot.* 68, 1697–1713. doi: 10.1093/jxb/erx038

Thomas, S. G. (2017). Novel Rht-1 dwarfing genes: tools for wheat breeding and dissecting the function of DELLA proteins. *J. Exp. Bot.* 68, 354–358. doi: 10.1093/jxb/erw509

Thudi, M., Chitikineni, A., Liu, X., He, W., Roorkiwal, M., Yang, W., et al. (2016). Recent breeding programs enhanced genetic diversity in both desi and kabuli varieties of chickpea (*Cicer arietinum* L.). *Sci. Rep.* 6:38636.

Tian, Z., Wang, X., Lee, R., Li, Y., Specht, J. E., Nelson, R. L., et al. (2010). Artificial selection for determinate growth habit in soybean. *Proc. Natl. Acad. Sci. U.S.A.* 107, 8563–8568. doi: 10.1073/pnas.1000088107

Toojinda, T., Siangliw, M., Tragoonrung, S., and Vanavichit, A. (2003). Molecular genetics of submergence tolerance in rice: QTL analysis of key traits. *Ann. Bot.* 91, 243–253. doi: 10.1093/aob/mcf072

Toojinda, T., Tragoonrung, S., Vanavichit, A., Siangliw, J. L., Pa-In, N., et al. (2005). Molecular breeding for rainfed lowland rice in the Mekong region. *Plant Prod. Sci.* 8, 330–333. doi: 10.1626/pps.8.330

Torres, N., Hilbert, G., Antolín, M. C., and Goicoechea, N. (2019). Aminoacids and flavonoids profiling in Tempranillo berries can be modulated by the arbuscular mycorrhizal fungi. *Plants* 8:400. doi: 10.3390/plants8100400

Toubiana, D., Semel, Y., Tohge, T., Beleggia, R., Cattivelli, L., Rosental, L., et al. (2012). Metabolic profiling of a mapping population exposes new insights in the regulation of seed metabolism and seed, fruit, and plant relations. *PLoS Genet.* 8:e1002612. doi: 10.1371/journal.pgen.1002612

Toyomasu, T., Kawaide, H., Sekimoto, H., Von Numers, C., Phillips, A. L., Hedden, P., et al. (1997). Cloning and characterization of a cDNA encoding gibberellin 20-oxidase from rice (*Oryza sativa*) seedlings. *Physiol. Plant* 99, 111–118. doi: 10.1034/j.1399-3054.1997.990116.x

Tudor, E. H., Jones, D. M., He, Z., Bancroft, I., Trick, M., Wells, R., et al. (2020). QTL-seq identifies BnaFT. A02 and BnaFLC. A02 as candidates for variation in vernalization requirement and response in winter oilseed rape (*Brassica napus*). *Plant Biotechnol. J.* 18, 2466–2481. doi: 10.1111/pbi.13421

Unamba, C. I., Nag, A., and Sharma, R. K. (2015). Next generation sequencing technologies: the doorway to the unexplored genomics of non-model plants. *Front. Plant Sci.* 6:1074. doi: 10.3389/fpls.2015.01074

Vallarino, J. G., Pott, D. M., Cruz-Rus, E., Miranda, L., Medina-Minguez, J. J., Valpuesta, V., et al. (2019). Identification of quantitative trait loci and candidate genes for primary metabolite content in strawberry fruit. *Hortic. Res.* 6:4.

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends Genet.* 34, 666–681. doi: 10.1016/j.tig.2018.05.008

Varkonyi-Gasic, E., Wang, T., Voogd, C., Jeon, S., Drummond, R. S., Gleave, A. P., et al. (2019). Mutagenesis of kiwifruit CENTRORADIALIS-like genes transforms a climbing woody perennial with long juvenility and axillary flowering into a compact plant with rapid terminal flowering. *Plant Biotech. J.* 17, 869–880. doi: 10.1111/pbi.13021

Varshney, R. K., Thudi, M., Roorkiwal, M., He, W., Upadhyaya, H. D., Yang, W., et al. (2019). Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits. *Nat. Genet.* 51, 857–864. doi: 10.1038/s41588-019-0401-3

Vidigal, P., Duarte, B., Cavaco, A. R., Caçador, I., Figueiredo, A., Matos, A. R., et al. (2018). Preliminary diversity assessment of an undervalued tropical bean (*Lablab purpureus* (L.) Sweet) through fatty acid profiling. *Plant Physiol. Biochem.* 132, 508–514. doi: 10.1016/j.plaphy.2018.10.001

Wan, L., Li, B., Lei, Y., Yan, L., Huai, D., Kang, Y., et al. (2018). Transcriptomic profiling reveals pigment regulation during peanut testa development. *Plant Physiol. Biochem.* 125, 116–125. doi: 10.1016/j.plaphy.2018.01.029

Wan, L., Li, B., Lei, Y., Yan, L., Ren, X., Chen, Y., et al. (2017). Mutant transcriptome sequencing provides insights into pod development in peanut (*Arachis hypogaea* L.). *Front. Plant Sci.* 8:1900. doi: 10.3389/fpls.2017.01900

Wan, L., Li, B., Pandey, M. K., Wu, Y., Lei, Y., Yan, L., et al. (2016). Transcriptome analysis of a new peanut seed coat mutant for the physiological regulatory mechanism involved in seed coat cracking and pigmentation. *Front. Plant Sci.* 7:1491. doi: 10.3389/fpls.2016.01491

Wang, D., Cao, D., Zong, Y., Li, Y., Wang, J., Li, Z., et al. (2021). Bulked QTL-Seq identified a major QTL for the awnless trait in spring wheat cultivars in Qinghai. *China. Biotechnol. Biotechnol. Equip.* 35, 124–130. doi: 10.1080/13102818.2020.1857661

Wang, H., Cheng, H., Wang, W., Liu, J., Hao, M., Mei, D., et al. (2016). Identification of BnaYUCCA6 as a candidate gene for branch angle in *Brassica napus* by QTL-seq. *Sci. Rep.* 6:38493.

Wang, H., Studer, A. J., Zhao, Q., Meeley, R., and Doebley, J. F. (2015). Evidence that the origin of naked kernels during maize domestication was caused by a single amino acid substitution in tga1. *Genetics* 200, 965–974. doi: 10.1534/genetics.115.175752

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. *Science* 343, 80–84. doi: 10.1126/science.1246981

Wang, W., Pan, Q., He, F., Akhunova, A., Chao, S., Trick, H., et al. (2018). Transgenerational CRISPR-Cas9 activity facilitates multiplex gene editing in allopolyploid wheat. *CRISPR J.* 1, 65–74. doi: 10.1089/crispr.2017.0010

Wang, Y., Clevenger, J. P., Illa-Berenguer, E., Meulia, T., van der Knaap, E., and Sun, L. (2019). A comparison of sun, ovate, fs8. 1 and auxin application on

tomato fruit shape and gene expression. *Plant Cell Physiol.* 60, 1067–1081. doi: 10.1093/pcp/pcz024

Wang, Y., Li, B., Du, M., Eneji, A. E., Wang, B., Duan, L., et al. (2012). Mechanism of phytohormone involvement in feedback regulation of cotton leaf senescence induced by potassium deficiency. *J. Exp. Bot.* 63, 5887–5901. doi: 10.1093/jxb/ers238

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Wen, W., Jin, M., Li, K., Liu, H., Xiao, Y., Zhao, M., et al. (2018). An integrated multi-layered analysis of the metabolic networks of different tissues uncovers key genetic components of primary metabolism in maize. *Plant J.* 93, 1116–1128. doi: 10.1111/tpj.13835

Wen, W., Li, D., Li, X., Gao, Y., Li, W., Li, H., et al. (2014). Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat. Commun.* 5:3438.

Wen, W., Li, K., Alseekh, S., Omranian, N., Zhao, L., Zhou, Y., et al. (2015).). Genetic determinants of the network of primary metabolism and their relationships to plant performance in a maize recombinant inbred line population. *Plant Cell* 27, 1839–1856. doi: 10.1105/tpc.15.00208

Weng, J., Gu, S., Wan, X., Gao, H., Guo, T., Su, N., et al. (2008). Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. *Cell Res.* 18, 1199–1209. doi: 10.1038/cr.2008.307

Whipple, C. J., Kebrom, T. H., Weber, A. L., Yang, F., Hall, D., Meeley, R., et al. (2011). grassy tillers1 promotes apical dominance in maize and responds to shade signals in the grasses. *Proc. Natl. Acad. Sci. U.S.A.* 108, E506–E512.

Willits, M. G., Kramer, C. M., Prata, R. T., De Luca, V., Potter, B. G., et al. (2005). Utilization of the genetic resources of wild species to create a non-transgenic high flavonoid tomato. *J. Agric. Food Chem.* 53, 1231–1236. doi: 10.1021/jf049355i

Witzel, K., Weidner, A., Surabhi, G. K., Börner, A., and Mock, H. P. (2009). Salt stress-induced alterations in the root proteome of barley genotypes with contrasting response towards salinity. *J. Exp. Bot.* 60, 3545–3557. doi: 10.1093/jxb/erp198

Woo, J. W., Kim, J., Kwon, S. I., Corvalán, C., Cho, S. W., Kim, H., et al. (2015). DNA-free genome editing in plants with preassembled CRISPR-Cas9 ribonucleoproteins. *Nat. Biotechnol.* 33, 1162–1164. doi: 10.1038/nbt.3389

Wu, N., Lu, Q., Wang, P., Zhang, Q., Zhang, J., Qu, J., et al. (2020). Construction and analysis of GmFAD2-1A and GmFAD2-2A soybean fatty acid desaturase mutants based on CRISPR/Cas9 technology. *Int. J. Mol. Sci.* 21:1104. doi: 10.3390/ijms21031104

Wu, S., Alseekh, S., Cuadros-Inostroza, Á, Fusari, C. M., Mutwil, M., Kooke, R., et al. (2016). Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in *Arabidopsis thaliana*. *PLoS Genet.* 12:e1006363. doi: 10.1371/journal.pgen.1006363

Wu, S., Tohge, T., Cuadros-Inostroza, Á, Tong, H., Tenenboim, H., Kooke, R., et al. (2018). Mapping the *Arabidopsis* metabolic landscape by untargeted metabolomics at different environmental conditions. *Mol. Plant* 11, 118–134. doi: 10.1016/j.molp.2017.08.012

Wu, Y., Li, X., Xiang, W., Zhu, C., Lin, Z., Wu, Y., et al. (2012). Presence of tannins in sorghum grains is conditioned by different natural alleles of Tannin1. *Proc. Natl. Acad. Sci. U.S.A.* 109, 10281–10286. doi: 10.1073/pnas.1201700109

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., and Van Der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* 319, 1527–1530. doi: 10.1126/science.1153040

Xie, K., and Yang, Y. (2013). RNA-guided genome editing in plants using a CRISPR–Cas system. *Mol. Plant.* 6, 1975–1983. doi: 10.1093/mp/sst119

Xu, C., Liberatore, K. L., MacAlister, C. A., Huang, Z., Chu, Y. H., Jiang, K., et al. (2015). A cascade of arabinosyltransferases controls shoot meristem size in tomato. *Nat. Genet.* 47, 784–792. doi: 10.1038/ng.3309

Xu, C., Park, S. J., Van Eck, J., and Lippman, Z. B. (2016). Control of inflorescence architecture in tomato by BTB/POZ transcriptional regulators. *Genes Dev.* 30, 2048–2061. doi: 10.1101/gad.288415.116

Xu, F., Sun, X., Chen, Y., Huang, Y., Tong, C., and Bao, J. (2015). Rapid identification of major QTLs associated with rice grain weight and their utilization. *PloS One* 10:e0122206. doi: 10.1371/journal.pone.0122206

Xu, K., and Mackill, D. J. (1996). A major locus for submergence tolerance mapped on rice chromosome 9. *Mol. Breeding* 2, 219–224. doi: 10.1007/bf00564199

Xu, K., Xu, X., Fukao, T., Canlas, P., Maghirang-Rodriguez, R., Heuer, S., et al. (2006). Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* 442, 705–708. doi: 10.1038/nature04920

Xu, R., Li, H., Qin, R., Wang, L., Li, L., Wei, P., et al. (2014). Gene targeting using the *Agrobacterium tumefaciens*-mediated CRISPR-Cas system in rice. *Rice* 7:5.

Yan, H., Ma, L., Wang, Z., Lin, Z., Su, J., and Lu, B. R. (2015). Multiple tissue-specific expression of rice seed-shattering gene SH4 regulated by its promoter pSH4. *Rice* 8:12.

Yan, L., Loukoianov, A., Blechl, A., Tranquilli, G., Ramakrishna, W., et al. (2004). The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. *Science* 303, 1640–1644. doi: 10.1126/science.1094305

Yan, L., Loukoianov, A., Tranquilli, G., Helguera, M., Fahima, T., and Dubcovsky, J. (2003). Positional cloning of the wheat vernalization gene VRN1. *Proc. Natl. Acad. Sci. U.S.A.* 100, 6263–6268. doi: 10.1073/pnas.0937399100

Yang, F., Jørgensen, A. D., Li, H., Søndergaard, I., Finnie, C., Svensson, B., et al. (2011). Implications of high-temperature events and water deficits on protein profiles in wheat (Triticum aestivum L. cv. Vinjett) grain. *Proteomics* 11, 1684–1695. doi: 10.1002/pmic.201000654

Yang, Q., Li, Z., Li, W., Ku, L., Wang, C., Ye, J., et al. (2013). CACTA-like transposable element in ZmCCT attenuated photoperiod sensitivity and accelerated the post-domestication spread of maize. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16969–16974. doi: 10.1073/pnas.1310949110

Yang, X., Xia, X., Zhang, Z., Nong, B., Zeng, Y., Xiong, F., et al. (2017). QTL mapping by whole genome re-sequencing and analysis of candidate genes for nitrogen use efficiency in rice. *Front. Plant Sci.* 8:1634. doi: 10.3389/fpls.2017.01634

Yaobin, Q., Peng, C., Yichen, C., Yue, F., Derun, H., Tingxu, H., et al. (2018). QTL-Seq identified a major QTL for grain length and weight in rice using near isogenic F2 population. *Rice Sci.* 25, 121–131. doi: 10.1016/j.rsci.2018.04.001

Ye, J., Li, W., Ai, G., Li, C., Liu, G., Chen, W., et al. (2019). Genome-wide association analysis identifies a natural variation in basic helix-loop-helix transcription factor regulating ascorbate biosynthesis via D-mannose/L-galactose pathway in tomato. *PLoS Genet.* 15:e1008149. doi: 10.1371/journal.pgen.1008149

Ye, J., Wang, X., Hu, T., Zhang, F., Wang, B., Li, C., et al. (2017). An InDel in the promoter of Al-ACTIVATED MALATE TRANSPORTER9 selected during tomato domestication determines fruit malate contents and aluminum tolerance. *Plant Cell* 29, 2249–2268. doi: 10.1105/tpc.17.00211

Yoo, H. J., Park, W. J., Lee, G. M., Oh, C. S., Yeam, I., Won, D. C., et al. (2017). Inferring the genetic determinants of fruit colors in tomato by carotenoid profiling. *Molecules* 22:764. doi: 10.3390/molecules22050764

Yoon, J., Cho, L. H., Kim, S. L., Choi, H., Koh, H. J., and An, G. (2014). The BEL 1-type homeobox gene SH 5 induces seed shattering by enhancing abscission-zone development and inhibiting lignin biosynthesis. *Plant J.* 79, 717–728. doi: 10.1111/tpj.12581

Youssef, H. M., Eggert, K., Koppolu, R., Alqudah, A. M., Poursarebani, N., Fazeli, A., et al. (2017). VRS2 regulates hormone-mediated inflorescence patterning in barley. *Nat. Genet.* 49, 157–161. doi: 10.1038/ng.3717

Yu, H., Lin, T., Meng, X., Du, H., Zhang, J., Liu, G., et al. (2021). A route to de novo domestication of wild allotetraploid rice. *Cell* 184, 1156–1170.

Yuan, M., Zhu, J., Gong, L., He, L., Lee, C., Han, S., et al. (2019). Mutagenesis of FAD2 genes in peanut with CRISPR/Cas9 based gene editing. *BMC Biotechnol.* 19:24. doi: 10.1186/s12896-019-0516-8

Yuan, Y. X., Wu, J., Sun, R. F., Zhang, X. W., Xu, D. H., Bonnema, G., et al. (2009). A naturally occurring splicing site mutation in the *Brassica rapa* FLC1 gene is associated with variation in flowering time. *J. Exp. Bot.* 60, 1299–1308. doi: 10.1093/jxb/erp010

Zeng, X., Yuan, H., Dong, X., Peng, M., Jing, X., Xu, Q., et al. (2020). Genome-wide dissection of co-selected UV-B responsive pathways in the UV-B adaptation of qingke. *Mol. Plant* 13, 112–127. doi: 10.1016/j.molp.2019.10.009

Zhang, C., Badri Anarjan, M., Win, K. T., Begum, S., and Lee, S. (2021). QTL-seq analysis of powdery mildew resistance inmildew resistance in a Korean cucumber inbred line. *Theor. Appl. Genet.* 134, 435–451. doi: 10.1007/s00122-020-03705-x

Zhang, D., Wang, X., Li, S., Wang, C., Gosney, M. J., Mickelbart, M. V., et al. (2019a). A post-domestication mutation, Dt2, triggers systemic modification of divergent and convergent pathways modulating multiple agronomic traits in soybean. *Mol. Plant* 12, 1366–1382. doi: 10.1016/j.molp.2019.05.010

Zhang, H., Zhang, J., Wei, P., Zhang, B., Gou, F., Feng, Z., et al. (2014). The CRISPR/C as9 system produces specific and homozygous targeted gene editing in rice in one generation. *Plant Biotechnol. J.* 12, 797–807. doi: 10.1111/pbi.12200

Zhang, M., Lv, D., Ge, P., Bian, Y., Chen, G., Zhu, G., et al. (2014). Phosphoproteome analysis reveals new drought response and defense mechanisms of seedling leaves in bread wheat (*Triticum aestivum* L.). *J. Proteomics* 109, 290–308. doi: 10.1016/j.jprot.2014.07.010

Zhang, X., Hou, X., Liu, Y., Zheng, L., Yi, Q., Zhang, H., et al. (2019b). Maize brachytic2 (br2) suppresses the elongation of lower internodes for excessive auxin accumulation in the intercalary meristem region. *BMC Plant Biol.* 19:589. doi: 10.1186/s12870-019-2200-5

Zhang, X., Wang, W., Guo, N., Zhang, Y., Bu, Y., Zhao, J., et al. (2018). Combining QTL-seq and linkage mapping to fine map a wild soybean allele characteristic of greater plant height. *BMC Genom.* 19:226. doi: 10.1186/s12864-018-4582-4

Zhang, X., Yin, F., Xiao, S., Jiang, C., Yu, T., Chen, L., et al. (2019c). Proteomic analysis of the rice (*Oryza officinalis*) provides clues on molecular tagging of proteins for brown planthopper resistance. *BMC Plant Biol.* 19:30. doi: 10.1186/s12870-018-1622-9

Zhang, X., Zhang, K., Wu, J., Guo, N., Liang, J., Wang, X., et al. (2020). QTL-Seq and sequence assembly rapidly mapped the gene BrMYBL2. 1 for the purple trait in *Brassica rapa*. *Sci. Rep.* 10:2328.

Zhang, Y., Massel, K., Godwin, I. D., and Gao, C. (2018). Applications and potential of genome editing in crop improvement. *Genome Biol.* 19:210.

Zhang, Z., Belcram, H., Gornicki, P., Charles, M., Just, J., Huneau, C., et al. (2011). Duplication and partitioning in evolution and function of homoeologous Q loci governing domestication characters in polyploid wheat. *Proc. Natl. Acad. Sci. U.S.A.* 108, 18737–18742. doi: 10.1073/pnas.1110552108

Zhang, Z., Li, A., Song, G., Geng, S., Gill, B. S., Faris, J. D., et al. (2020). Comprehensive analysis of Q gene near-isogenic lines reveals key molecular pathways for wheat domestication and improvement. *Plant J.* 102, 299–310. doi: 10.1111/tpj.14624

Zhao, Y., Ma, J., Li, M., Deng, L., Li, G., Xia, H., et al. (2020). Whole-genome resequencing-based QTL-seq identified AhTc1 gene encoding a R2R3-MYB transcription factor controlling peanut purple testa colour. *Plant Biotechnol. J.* 18, 96–105. doi: 10.1111/pbi.13175

Zheng, M., Yang, T., Liu, X., Lü, G., Zhang, P., Jiang, B., et al. (2020). qRf8-1, a Novel QTL for the fertility restoration of maize CMS-C identified by QTL-seq. *G3* 10, 2457–2464. doi: 10.1534/g3.120.401192

Zhou, H., Liu, B., Weeks, D. P., Spalding, M. H., and Yang, B. (2014). Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. *Nucleic Acids Res.* 42, 10903–10914. doi: 10.1093/nar/gku806

Zhou, J., Wang, G., and Liu, Z. (2018). Efficient genome editing of wild strawberry genes, vector development and validation. *Plant Biotech. J.* 16, 1868–1877. doi: 10.1111/pbi.12922

Zhou, Y., Lu, D., Li, C., Luo, J., Zhu, B. F., Zhu, J., et al. (2012). Genetic control of seed shattering in rice by the APETALA2 transcription factor shattering abortion1. *Plant Cell* 24, 1034–1048. doi: 10.1105/tpc.111.094383

Zhou, Y., Yang, P., Cui, F., Zhang, F., Luo, X., and Xie, J. (2016). Transcriptome analysis of salt stress responsiveness in the seedlings of Dongxiang wild rice (*Oryza rufipogon* Griff.). *PLoS One* 11:e0146242. doi: 10.1371/journal.pone.0146242

Zhu, C., Li, X., and Yu, J. (2011). Integrating rare-variant testing, function prediction, and gene network in composite resequencing-based genome-wide association studies (CR-GWAS). *G3* 1, 233–243. doi: 10.1534/g3.111.000364

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., et al. (2018). Rewiring of the fruit metabolome in tomato breeding. *Cell* 172, 249–261. doi: 10.1016/j.cell.2017.12.019

Zhu, Y., Li, Y., Xin, D., Chen, W., Shao, X., Wang, Y., et al. (2015). RNA-Seq-based transcriptome analysis of dormant flower buds of Chinese cherry (*Prunus pseudocerasus*). *Gene* 555, 362–376. doi: 10.1016/j.gene.2014.11.032

Zsögön, A., Čermák, T., Naves, E. R., Notini, M. M., Edel, K. H., Weinl, S., et al. (2018). De novo domestication of wild tomato using genome editing. *Nat. Biotechnol.* 36, 1211–1216. doi: 10.1038/nbt.4272

Zsögön, A., Cermak, T., Voytas, D., and Peres, L. E. P. (2017). Genome editing as a tool to achieve the crop ideotype and de novo domestication of wild relatives: case study in tomato. *Plant Sci.* 256, 120–130. doi: 10.1016/j.plantsci.2016.12.012

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: 15 April 2021
doi: 10.3389/fgene.2021.635971

Three-in-One Simultaneous Extraction of Proteins, Metabolites and Lipids for Multi-Omics

Jianing Kang[1,2,3†], *Lisa David*[2,3†], *Yangyang Li*[2,4], *Jing Cang*[1] *and Sixue Chen*[2,3,5,6*]

[1] College of Life Science, Northeast Agricultural University, Harbin, China, [2] Department of Biology, University of Florida, Gainesville, FL, United States, [3] University of Florida Genetics Institute, Gainesville, FL, United States, [4] College of Horticulture, Shenyang Agricultural University, Shenyang, China, [5] Plant Molecular and Cellular Biology Program, University of Florida, Gainesville, FL, United States, [6] Proteomics and Mass Spectrometry, Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL, United States

Elucidation of complex molecular networks requires integrative analysis of molecular features and changes at different levels of information flow and regulation. Accordingly, high throughput functional genomics tools such as transcriptomics, proteomics, metabolomics, and lipidomics have emerged to provide system-wide investigations. Unfortunately, analysis of different types of biomolecules requires specific sample extraction procedures in combination with specific analytical instrumentation. The most efficient extraction protocols often only cover a restricted type of biomolecules due to their different physicochemical properties. Therefore, several sets/aliquots of samples are needed for extracting different molecules. Here we adapted a biphasic fractionation method to extract proteins, metabolites, and lipids from the same sample (3-in-1) for liquid chromatography-tandem mass spectrometry (LC-MS/MS) multi-omics. To demonstrate utility of the improved method, we used bacteria-primed *Arabidopsis* leaves to generate multi-omics datasets from the same sample. In total, we were able to analyze 1849 proteins, 1967 metabolites, and 424 lipid species in single samples. The molecules cover a wide range of biological and molecular processes, and allow quantitative analyses of different molecules and pathways. Our results have shown the clear advantages of the multi-omics method, including sample conservation, high reproducibility, and tight correlation between different types of biomolecules.

Keywords: multi-omics, 3-in-1 method, proteomics, metabolomics, lipidomics, *Arabidopsis*, disease

OPEN ACCESS

Edited by:
Gorji Marzban, University of Natural Resources and Life Sciences, Vienna, Austria

Reviewed by:
Candice Ulmer, Centers for Disease Control and Prevention (CDC), United States
Julie Ann Reisz, University of Colorado, United States

***Correspondence:**
Sixue Chen
schen@ufl.edu

† These authors have contributed equally to this work

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

Received: *30 November 2020*
Accepted: *18 February 2021*
Published: *15 April 2021*

Citation:
Kang J, David L, Li Y, Cang J and Chen S (2021) Three-in-One Simultaneous Extraction of Proteins, Metabolites and Lipids for Multi-Omics. Front. Genet. 12:635971. doi: 10.3389/fgene.2021.635971

INTRODUCTION

Systems biology, the comprehensive study of biological components and their interactions within a cell or a tissue, is indispensable toward understanding complex cellular functions and processes. Multi-omic measurements and integration of the resulting information can transform our understanding of complex biological systems (Dai and Chen, 2012; He et al., 2012; Mostafa et al., 2016; Meng et al., 2019). Multiple layers of information (DNA, RNA, protein, metabolite, and lipid) can provide important insights into cellular molecular networks. In recent years, rapid progress has been made in genomics and transcriptomics. Nevertheless, proteomics, metabolomics, and lipidomics have emerged as cornerstones in the field of systems biology because the essential

information at protein, metabolite, and lipid levels cannot be predicted or deduced from genomics and transcriptomics (He et al., 2012; Geng et al., 2016, 2017; Mostafa et al., 2016; Meng et al., 2019).

To conduct multi-omics, aliquots of the same sample are required for different extraction procedures optimized for different biomolecules. In addition to increased effort inherent to different parallel sample handling, the required sample amounts for multiple extractions are often not available. Meanwhile, the multi-components extracted from parallel sets of replicates can decrease consistency and comparability when performing multi-omics integration. Consequently, a versatile extraction method providing robust and reliable recovery of the molecular components from a single sample is desirable. Such a method decreases sample handling time and thus increases throughput. Importantly, it conserves critical samples and improves data accuracy and comparability because different molecules are all derived from the same sample. Common methods employed for fractionated extractions are based on a two-phase lipid extraction method developed in 1957 (Folch et al., 1957). It uses chloroform/methanol/water partitioning of polar and hydrophobic metabolites and was designed to increase the purity of lipids. Here we modified and optimized this method to obtain high quality proteins, metabolites, and lipids from a single sample, as a 3-in-1 method (**Figures 1**, **2**). This method can be easily applied to many types of materials. It should be noted that when applying to other sample types, the amount of samples may vary based on the types of samples and their water content, etc. Regardless of the source material, proteins, lipids, and metabolites have the same physicochemical properties, therefore, this method has broad application potential.

To test the utility of our 3-in-1 extraction method, we used leaves of *Arabidopsis thaliana* (WS ecotype) that had been primed by a pathogenic *Pseudomonas syringae pv. tomato* DC3000 (*Pst* DC3000). Systemic Acquired Resistance (SAR), a salicylic acid (SA)-dependent immune response, improves immunity of systemic tissues after prior localized exposure to a pathogen. *Arabidopsis* knockout mutants defective in SAR response differ in disease resistance when compared to wild type plants. Here we used *Arabidopsis* wild type and a knockout mutant of a lipid transfer protein DIR1 (defective in induced resistance 1) to examine SAR in whole leaves. We have successfully used the 3-in-1 method and annotated 424 lipids, which cover most of the lipid classes. In addition, we have identified 1967 metabolites using a LC-MSn method, and obtained 1849 protein identifications. These results demonstrate the superior 3-in-1 method can greatly facilitate multi-omics studies in systems biology.

MATERIALS AND METHODS

Plant Growth and Bacterial Injection

Arabidopsis thaliana wild type (WS ecotype) and *dir1* mutant (in WS background) were grown in 8 h light/16 h dark with a light intensity of 140 μmol/m^2 s. One mature rosette leaf of the 5-week-old plants was injected with either *Pst* DC3000 in 10 mM $MgCl_2$ (OD600 = 0.02) for treated plants or 10 mM $MgCl_2$ for mock plants using a needleless syringe. Fully expanded distal rosette leaves that were not injected were collected at 48 h after infiltration and directly frozen in liquid nitrogen and stored in −80°C for 3-in-1 extraction. Three biological replicates of treated and three biological replicates of mock leaves were used.

Multi-Omics Sample Preparation

Three hundred milligrams fresh weight leaves were quickly immersed in glass tubes with 3 ml pre-heated 75°C methanol (MeOH) and 0.01% butylated hydroxytoluene (BHT) and incubated for 15 min. Internal standards were added to each sample as follows: for proteins: 60 fmol bovine serum albumin (BSA) tryptic peptides per 1 μg sample protein; for metabolites: 10 μL 0.1 nmol/μL lidocaine and camphorsulfonic acid; and for lipids: 10 μL 0.2 μg/μL deuterium labeled 15:0-18:1(d7) phosphatidylethanolamine (PE) and 15:0-18:1(d7) diacylglycerol (DG).

For extraction of proteins, metabolites and lipids, 6 ml of chloroform and 2 ml of water (3:1, vol/vol) were added to each tube and 500 μl of MeOH was added to replenish the methanol that evaporated from boiling (Folch et al., 1957). Samples were vortexed at 4°C for 1 h. The liquid was transferred from the extracts to glass centrifuge tubes for further phase separation. To improve collection of all 3 components, 2 ml of chloroform/methanol (2:1 vol/vol) with 0.01% BHT was added to the leaves in the glass tubes and agitated for another 30 min at 4°C. This liquid was combined with the previous into the glass centrifuge tubes. This last extraction procedure was repeated twice on all the samples until the leaves appeared white. After extraction, leaves were dried at 105°C overnight and weighed for dry weights.

For phase separation, extracts were centrifuged at 10,000 rpm for 10 min at 4°C. First the upper (metabolites in MeOH) phase was collected and transferred to plastic 2 ml centrifuge tubes, then the bottom (lipids in chloroform) phase was removed and transferred to glass tubes, leaving the middle (protein) layer for protein collection. The lipid extract was evaporated by the Nitrogen gas and dried sample tube filled with nitrogen gas was placed at −80°C. The lipid extract was dissolved in 1 ml isopropanol (IPA) for LC-MS analysis. Metabolites were lyophilized to dryness, then the tubes were filled with argon and placed at −80°C. Metabolites were solubilized in 100 μl of 0.1% formic acid (FA) for LC MS/MS analysis.

Protein Precipitation and Trypsin Digestion

Proteins were precipitated by addition of 80% acetone in the glass centrifuge tubes in −20°C. After 16 h, the samples were centrifuged at 10,000 rpm for 10 min at 4°C. After removing acetone, proteins were resuspended in 100 μl of 50 mM ABC, reduced using 10 mM dithiothreitol (DTT) for 1 h at 22°C, and then alkylated with 55 mM iodoacetamide (IAM) for 1 h in darkness. The samples were digested with trypsin (1:100 w/w) for 16 h. All the samples were acidified by addition of 0.1% FA to stop the digestion and stored at −80°C.

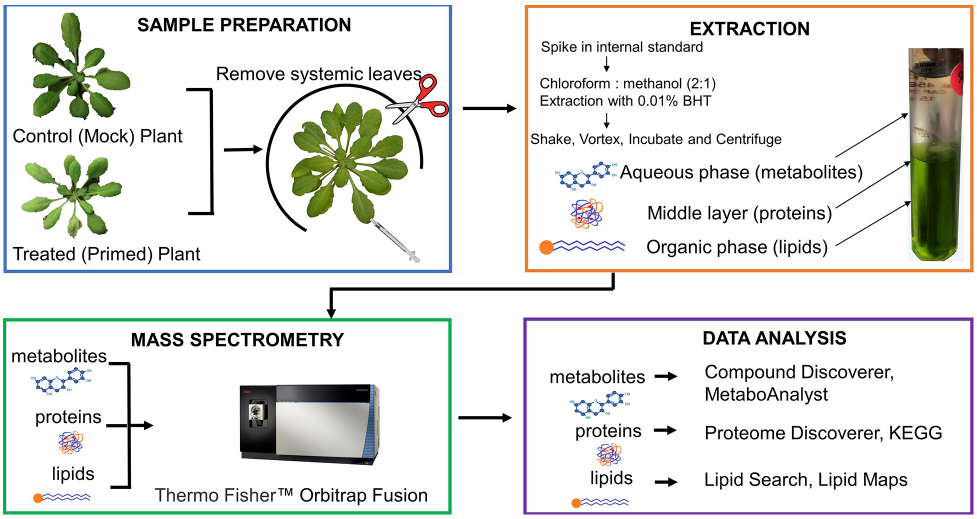

FIGURE 1 | Diagram of 3-in-1 sample preparation method for profiling proteins, metabolites, and lipids from control and primed *Arabidopsis* leaves. The biphasic fractionation separates three types of biomolecules simultaneously, which are analyzed on the same mass spectrometry platform. The data are also analyzed using the same vendor's software. A more detailed workflow of the extraction is shown in **Figure 2**.

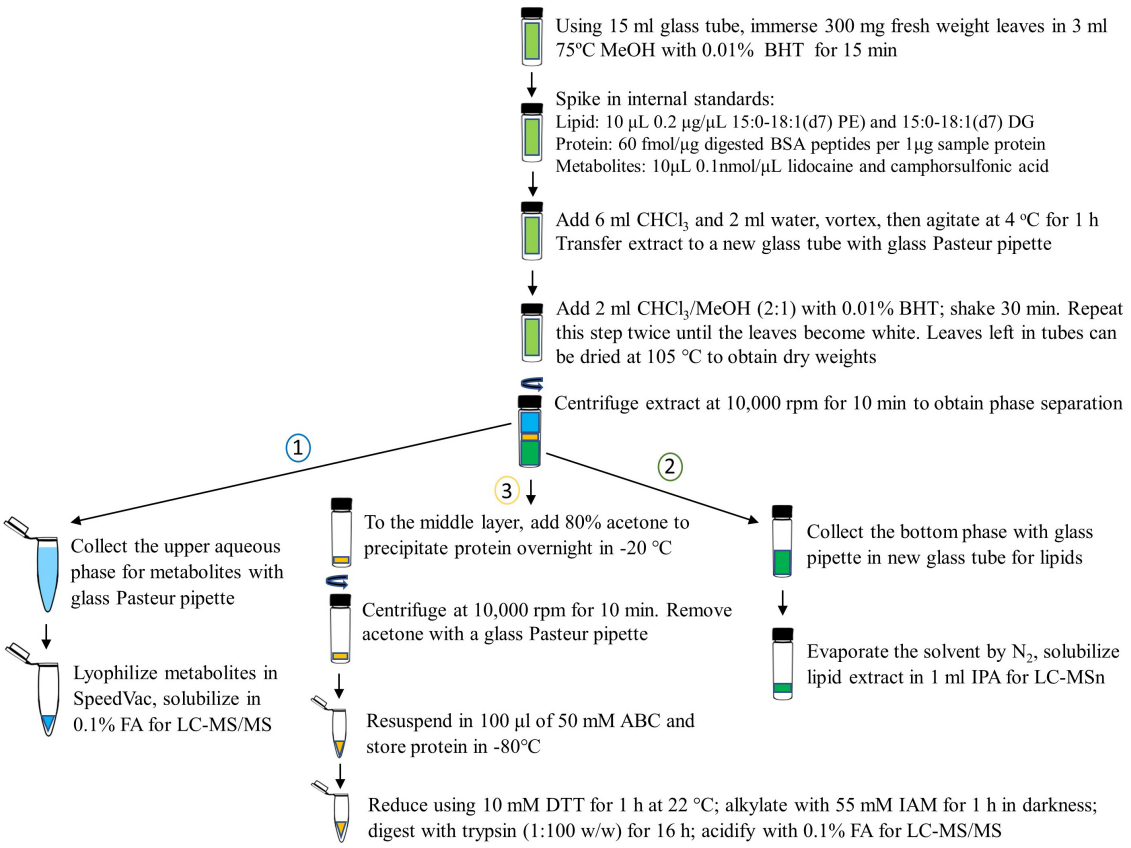

FIGURE 2 | Detailed workflow of 3-in-1 sample extraction of proteins, metabolites, and lipids from control and primed *Arabidopsis* leaves.
A chloroform/methanol/water extraction is used to separate the three fractions and each layer is carefully isolated using supplies of glass materials to avoid plastic contaminants in samples. The order of fractionated is important and labeled. Butylated hydroxytoluene (BHT) is added at the start of the extraction to avoid oxidation of lipids during the procedure. MeOH, methanol; PE, phosphatidylethanolamine; DG, diacylglycerol; BSA, bovine serum albumin; $CHCl_3$, chloroform; FA, formic acid; LC-MS/MS, liquid chromatography tandem mass spectrometry; IPA, isopropanol; ABC, ammonium bicarbonate; DTT, dithiothreitol; IAM, iodoacetamide.

Liquid Chromatography Mass Spectrometry (LC-MS) and Omics Data Analysis

Untargeted metabolomic, lipidomic, and proteomic methods were run on an Orbitrap Fusion Tribrid mass spectrometer (Thermo Fisher Scientific, Bremen, Germany). A Vanquish^{TM} UHPLC was used for lipids and metabolites, and an Easy-nLC was used for peptides. An Accucore C18 (100 mm × 2.1 mm, 2 μm) column and an Acclaim C30 (2.1 mm × 250 mm, 3 μm) were used for metabolites and lipids, respectively. The column chamber temperature was 55°C, and the pump flow rate was 0.45 ml/min. For metabolomics, solvent A (0.1% FA) and solvent B (0.1% FA and 99.9% acetonitrile) were used. The LC gradient is set to 0 min: 1% of solvent B, 5 min: 1% of B, 6 min: 40% of B, 7.5 min: 98% of B, 8.5 min: 98% of B, 9 min: 0.1% of B, 10 min stop run. To enhance identification, an AcquireX MSn data acquisition strategy was used which employs replicate injections for exhaustive sample interrogation and increases the number of compounds in the sample with distinguishable fragmentation spectra for identification (David et al., 2021). Pooled samples were created using equal volumes of all the samples for quality control, and were run after each sample set. Electrospray ionization spray voltage for positive ions was 3500 and for negative ions was 2500. Sheath gas was set to 50, auxiliary gas was set at 1 and sweep gas was set to 1. The ion transfer tube temperature was set at 325°C and the vaporizer temperature was set at 350°C. Full MS1 used the Orbitrap mass analyzer with a resolution of 120,000, scan range (m/z) of 55–550, maximum injection time (MIT) of 50, automatic gain control (AGC) target of 2e5, 1 microscan, and RF lens set to 50. For lipidomics, solution A consisted of 0.1% FA, 10 mM ammonium formate, and 60% acetonitrile. Solution B consisted of 0.1% FA, 10 mM ammonium formate, and 90:10 acetonitrile: isopropyl alcohol. The LC gradient is set to 0 min: 32% of solvent B (i.e., 68% of solvent A), 1.5 min: 45% of B, 5 min: 52% of B, 8 min: 58% of B, 11 min: 66% of B, 14 min: 70% of B, 18 min: 75% of B, 21 min: 97% of B, 26 min: 32% of B, 32 min stop run. Full MS1 used the Orbitrap ion trap mass analyzer with a resolution of 70,000, 1 microscan, AGC target set to 1e6, and a scan range from 200 to 2000 m/z for positive and negative polarity. The dd-MS^2 scan used 1 microscan, resolution of 35,000, AGC target 5e5, MIT of 46 ms, and loop count of 3.

The column used for peptides was the Acclaim PepMap^{TM} 100 pre-column (75 μm × 2 cm, nanoViper C18, 3 μm, 100 A) combined with an Acclaim PepMap^{TM} RSLC (75 μm × 25 cm, nanoViper C18, 2 μm, 100 A) analytical column. The LC runs a linear gradient of solvent B (0.1% FA, 99.9% Acetonitrile) from 1 to 30% for 90 min at 250 nL/min. The solvent A was 0.1% FA. The MS was operated in data-dependent acquisition mode with a cycle time of 3 s. Eluted peptides were detected in the Orbitrap MS at a 120,000 resolution with a scan range of 350–1800 m/z. Most abundant ions bearing 2–7 charges were selected for MS/MS analysis. AGC for the full MS scan was set as 2e5 with MIT as 50 ms, and AGC Target of 1e4 and MIT of 35 ms were set for the MS/MS scan. The normalized collision energy was 35, and ions were detected with an Ion Trap detector. A dynamic exclusion time of 30 s was applied to prevent repeated sequencing of the most abundant peptides.

Proteome Discoverer^{TM} 2.4, Compound Discover^{TM} 3.0, and Lipid Search 4.1^{TM} software (Thermo Fisher Scientific, Bremen, Germany) were used for proteomics, metabolomics and lipidomics data analyses, respectively (**Figure 1**). Software scoring parameters used for metabolite, lipid, and protein identifications are briefly described here with references provided to previous publications with more details (Geng et al., 2016, 2017; Breitkopf et al., 2017). Briefly, for proteomic data analysis, MS/MS spectra were searched against *Arabidopsis* TAIR10 database with 10 ppm mass tolerance for MS1 and 0.02 Da tolerance for MS2, two missed cleavage sites, fixed modification of cysteine carbamidomethylation (+57.021), and dynamic modifications of methionine oxidation (+15.996). Peptide confidence level was set at 1% false discovery rate with at least two unique peptides. Relative protein abundance in treated and mock samples was measured using label-free quantification in the Proteome Discoverer^{TM} 2.4. For metabolomics data, metabolite identification included predicting compositions, searching mzCloud spectra database, and assigning compound annotations by searching ChemSpider, Pathway mapping to KEGG and Metabolika pathways was used for functional analysis. The metabolites were scored by applying mzLogic and the best score was kept. Peak areas were normalized by the positive and negative mode internal standards (lidocaine and camphorsulfonic acid, respectively) (Geng et al., 2017). For lipidomics data, raw files from three replicates of mock and treated were uploaded to Lipid Search 4.1^{TM} for annotation of lipids found in all the samples. A mass list was generated for uploading to Compound Discover^{TM} 3.0 Software. This mass list was used for metabolite identification along with predicted compositions, mzCloud database matching, and compound annotations. Lipid Search scoring algorithms considering lipid fragmentation ions related to headgroup, fatty acids and backbone, as well as precursor and product ion accuracy of 5 ppm were used. Peak areas were normalized by median-based normalization.

Statistical analyses were done by normalizing peak areas by internal standards spiked in the samples. The average areas of three biological replicates of each group were compared as a ratio and two criteria were used to determine significantly altered components: (1) *p*-value from an unpaired student's *t*-test less than 0.05, and (2) increase or decrease of 2-fold (*dir1* primed/wild type primed) (**Supplementary Table 1**). All protein MS raw data and search results have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the data set PXD023094. All the metabolomics and lipidomics MS raw data and search results have been deposited to the MetaboLights repository with the data set identifier MTBLS2303.

RESULTS AND DISCUSSION

The multi-omics sample preparation workflow that we have developed has allowed us to increase the number of lipids,

proteins, and metabolites identified from a single sample (**Figure 1**). Previous extraction methods applied to *Arabidopsis* leaves identified 1987 proteins (Nakayasu et al., 2016), 2638 proteins (Salem et al., 2016), 150 metabolites and 200 lipid species (Salem et al., 2016, 2017; **Table 1**). Our method was able to identify 1849 confident proteins with 2 or more unique peptides at FDR of less than 1%. Our method greatly increases the number of polar metabolites to 1967, and non-polar lipids to 424 lipid species (**Table 1**). This represents a more than 10-fold increase in the identified metabolites and more than twice the number of identified lipid species, when compared to previous *Arabidopsis* papers (Salem et al., 2016, 2017). The number of identified proteins in this work appears to be lower than reported in a previous paper (Salem et al., 2016), but we used stringent criteria for high confidence. Otherwise, we could have identified 2778 proteins (**Table 1**). We also compared our method to other three-part extraction methods developed for mammalian cell lines (Coman et al., 2016; Nakayasu et al., 2016). Again, our method stands out considering the large numbers of identified polar and non-polar metabolites. While the overall number of identified proteins in our samples is lower than those reported in Coman et al. (2016) and Nakayasu et al. (2016) (**Table 1**), we are fully aware that such a comparison may not be sensible because of species and protein database differences. For instance, the mouse proteome is larger with 55,152 entries in UniProt, while the TAIR10 database contains 35,386 entries (Zhang et al., 2019). Nevertheless, we can reasonably expect that our 3-in-1 method will lead to valuable results when applied to mammalian cells. Among all the 3-in-1 methods in **Table 1**, our method is most similar to Nakayasu et al. (2016), which used human epithelial Calu-3 cells. For *Arabidopsis*, they only reported identification of 1987 proteins using an-house software. Since it is not clear about their FDR and unique peptide criteria, it may be reasonable to assume that our protein data of 2778 proteins (with 1% FDR) and 1894 proteins (after applying additional two unique peptide filter) are comparable, if not better. Importantly, we identified nearly 20 times more metabolites and more than twice the lipid species (**Table 1** and **Supplementary Table 2**). Here are some technical improvements in our method: (1) we added a reductant at the first step to preserve lipids and extracted for longer time; (2) we did three chloroform/methanol extraction steps until the leaves looked white in color, while Nakayasu et al. (2016) only extracted one time; (3) we lyophilized the fractions of metabolites and lipids (under nitrogen gas) before reconstitution and LC-MSn, while they collected the lipid and metabolite layers directly into autosampler vials; (4) we used a new AcquireX LC-MSn data acquisition strategy (David et al., 2021), which enhanced the coverage of metabolome and lipidome; (5) their metabolomics was done using GC-MS, which is known to cover a small number of central metabolites (Gowda and Djukovic, 2014; Geng et al., 2017); and (6) this work may have also benefited from the use of Compound Discoverer software with access to a large MzCloud database of MS2 spectra.

Here we also compared our 3-in-1 extraction method to previously published methods targeted to a single component, including proteins (Zybailov et al., 2009; Niehl et al., 2013), metabolites (Fiehn et al., 2000; Wu et al., 2018), and lipids (Higashi et al., 2015; Kehelpannala et al., 2020). We found that our method allows for similar numbers of identified proteins, and increased numbers of metabolites and lipids when compared to these single component extraction methods (**Table 1**). Please note, that protein work mentioned in **Table 1** include gel-based sample prefractionation step to improve coverage, but in our study we obtained similar numbers of proteins without this fractionation step. We can attribute the improved extraction and identification of metabolites, lipids, and proteins to three factors: 1. advanced instrumentation by using the Orbitrap tribrid mass spectrometer; 2. deep sampling and fragmentation of analytes using the AcquireX technology resulting in improved level 2 identification by MS2 and MS3; and 3. preservation of each layer by using Nitrogen gas for evaporation of chloroform in lipid samples and addition of reductant to avoid lipid and metabolite oxidation, as well as avoiding disruption of middle layer to preserve for protein precipitation in acetone and the use of only glass materials to avoid plastic contaminations during extraction (**Table 1** and **Figure 2**). This procedure also requires careful removal of each component layer so as not to disrupt and disperse the middle layer that contains the proteins. This was achieved by avoiding agitation of the glass tube after removing from the centrifuge and by carefully sliding the glass pipette along the side of the tube to draw off the metabolite and lipid layers sequentially, leaving the protein layer intact (**Figure 2**).

Increased identification of proteins, metabolites and lipids is essential for understanding the interconnected molecular networks that mediate cellular responses. **Figure 3A** shows that different molecules (proteins, metabolites, and lipids) from a specific pathway can be examined together to gauge potential regulations and activities of the pathway. This is important because protein abundance data do not reflect the activity of the protein, but when combined with the information for metabolites and lipids, the activities of enzymes leading to synthesis of the metabolites can be deduced. **Figure 3B** shows that the identified proteins from *Arabidopsis* leaves cover a wide range of molecular pathways (129 out of 541 KEGG pathways), in addition to pathways covered by the identified metabolites and lipids, highlighting the complementary nature of different "omics." In **Figure 3C**, principal component analysis shows unsupervised clustering of wild type samples and mutant samples separately, and also that mock versus treated samples grouping together for proteins, metabolites and lipids. The results clearly indicate high reproducibility of the 3-in-1 method and its application to capturing biological differences related to *Arabidopsis* systemic acquired resistance.

When comparing to previous methods (**Table 1**), our new method clearly stands out in the high coverage of metabolome and lipidome. For example, both low abundant (methionine, tryptophan, and tyrosine) and high abundant amino acids (arginine and glutamic acid) in plants (Kumar et al., 2017) were identified. In addition, metabolites with a variety of chemical properties were covered, including polar (e.g., glutamine and tyrosine), non-polar (e.g., methionine), aromatic amino acids (e.g., tryptophan and tyrosine), cofactors (e.g.,

TABLE 1 | Comparison of the three-in-one method in this study with previously published targeted and three-in-one methods.

	Proteins		Metabolites		Lipids		Simultaneous extraction of proteins, metabolites, and lipids				
References	**Zybailov et al., 2009**	**Niehl et al., 2013**	**Fiehn et al., 2000**	**Wu et al., 2018**	**Higashi et al., 2015**	**Kehelpannala et al., 2020**	**Coman et al., 2016[a]**	**Nakayasu et al., 2016[a]**	**Salem et al., 2016**	**Salem et al., 2017**	**This work**
Materials	*Arabidopsis* ecotype Col-0	*Arabidopsis* ecotype Col-0	Arabidopsis ecotypes Col-2 and C24	*Arabidopsis* ecotype Col-0	*Arabidopsis* ecotypes Col-0 and Nossen	*Arabidopsis* ecotype Col-0	Mouse bone marrow cells	*Arabidopsis* Human epithelial Calu-3 cells	*Arabidopsis* ecotype Col-0	*Arabidopsis* ecotype Col-0	*Arabidopsis* ecotype WS
Extraction	Tris buffer with 5% SDS	Trizol and acetone precipitation	Chloroform: methanol: H_2O	Methyl-tert-butyl-ether: methanol: H_2O	Chloroform: methanol: H_2O	Chloroform: methanol: H_2O	Methyl-tert-butyl- ether: methanol: H_2O	Chloroform: methanol: H_2O	Methyl-tert-butyl- ether: methanol: H_2O	Methyl-tert-butyl- ether: methanol: H_2O	Chloroform: methanol: H_2O
Fractional on	Gel electrophoresis into 12 fractions	Gel electrophoresis into 8 fractions	Lipophilic and polar phases	Aqueous phase	Lipophilic phase	Lipophilic phase	SIMPLEx containing 3 phases	MPLEx containing 3 phases	Polar and non-polar liquid and liquid fractional	Polar and non-polar liquid and liquid fractional on	Triphasic fractionation
Chromato-graphy	Ultimate LC with 90 min gradient	Picotip with 50 min LC gradient	Gas chromatography 8000	Waters Acquity LC with 44 min gradient	Shimadzu LC with 40 min gradient	Agilent 1290 LC with 30 min gradient	Ultimate 3000 LC with 45 min gradient	Nano-/Cap-LC with 90 min gradient, Agilent GC-MS	Ultimate LC with 110 min gradient	Ultimate LC with 110 min gradient	Ultimate LC with 90 min gradient
Mass spectrometer	LTQ Orbitrap MS/MS	LTQ Qrbitrap MS/MS	Voyager mass spectrometer	Exactive Qrbitrap MS	Ion trap-Time-of-Flight (TOF) MS	Quadrupole-TOF MS/MS	LTQ Orbitrap Velos and QTRAP 6500 MS/MS	LTQ-Orbitrap Velos MS/MS	Q-Exactive Orbitrap MS/MS	Q-Exactive Orbitrap MS/MS	Orbitrap Fusion Tribrid MSn and AquireX
Software	Mascot 2.2	Mascot 2.3	MassLab FindTarget and Pirouette	REFINER MS 10.0	Profiling Solution and in-house Perl script	MS-DIAL	Progenesis 4.1	VIPER (in-house)	Mascot 2.5	Mascot 2.5	Proteome Discoverer 2.4
							MultiQuant 3.0	Metabolite Detector	Target Search	Target Search	Compound Discover 3.0
							Chipsoft 8.3.1	LIQUID (in-house)	Progenesis QI2.2	Progenesis QI 2.2	Lipid Search 4.1
Level[b]	2	2	1	2	1 and 2	2	1 and 2	2	2	2	1 and 2
Identification	2800 proteins	1474 proteins	326 metabolites	123 metabolites	66 lipids	208 lipids	3327 proteins	1987/2670 proteins[c]	2638 proteins	Not available	2778/1849 proteins[d]
							75 metabolites	51 metabolites	150 metabolites	50 metabolites	1967 metabolites
							360 lipids	236/171 lipids[e]	200 lipids	200 lipids	424 lipids

Our three-part extraction method is compared to other single component extraction methods and to other three-component extraction methods. Extraction technique, sample preservation, instrumentation method, and data analysis software all play a role in the improved profiling of proteins, metabolites, and lipids. a. this paper used mammalian cells. All other samples are Arabidopsis leaves. b. Level 1, Authentic standards (identification); Level 2, MS/MS data matching to library/database. c. 1987 proteins from Arabidopsis (of unknown ecotype), 2670 proteins identified from Human epithelial Calu-3 cells. d. 2778 proteins identified only with 1% FDR, and 1894 proteins after applying a two unique peptide filter. e. while it is written in the paper text that there were 236 lipids, only a total of 171 lipids could be found in the ***Supplementary Tables****.*

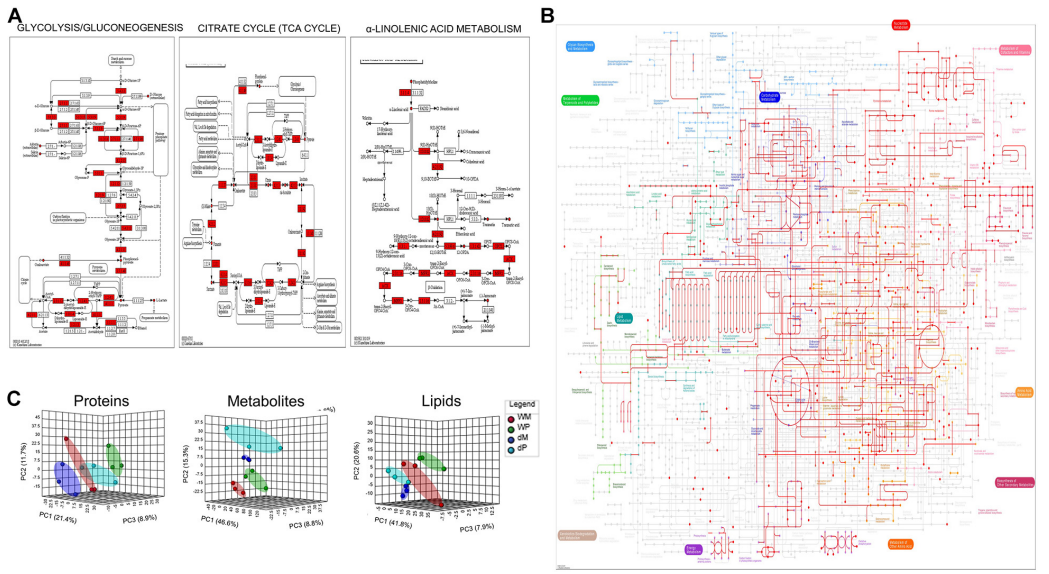

FIGURE 3 | Evaluation of KEGG pathway coverage, data quality, and performance improvement with the 3-in-1 extraction method. **(A)** Enhanced coverage of specific molecular pathways by the identified proteins, metabolites, and lipids. In The red-colored boxes represent identified proteins and the red-colored circles are lipids and metabolites. **(B)** Mapping of the quantified proteins, metabolites, and lipids onto the KEGG metabolic pathways. **(C)** Principal component analysis (PCA) of relative levels of proteins, metabolites, and lipids obtained from three biological replicates under the four experimental conditions (WM, wild type mock; WP, wild type primed; dM, *dir1* mock; dP, *dir1* primed).

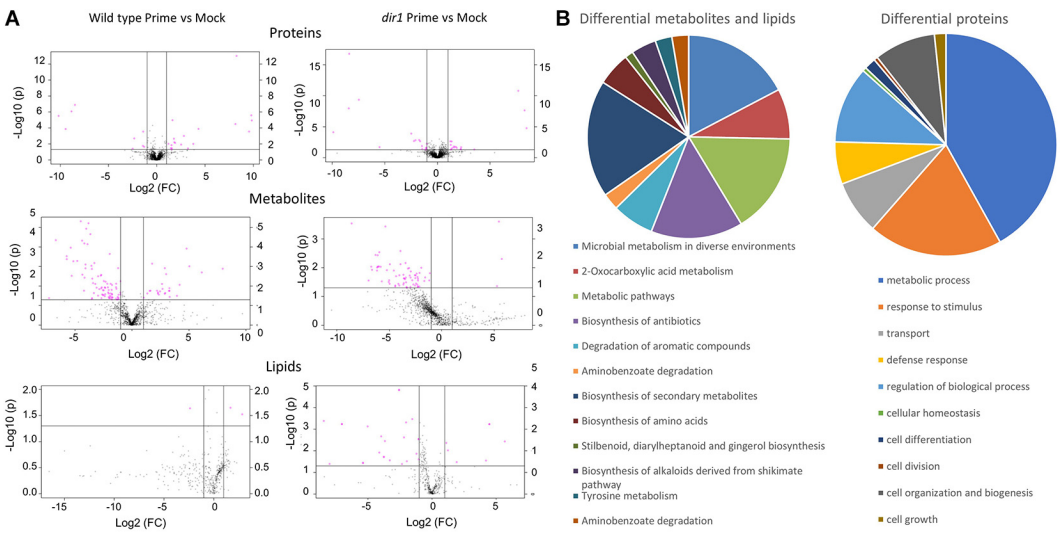

FIGURE 4 | Significant changes of proteins, metabolites, and lipids in *Arabidopsis* leaves of wild type and *dir1* mutant primed by *Pst DC 3000* treatment. **(A)** Volcano plots displaying differential changes of proteins, metabolites, and lipids in wild type and *dir1* mutant. Pink dots indicate differential molecules. **(B)** Biological functions of the differential metabolites/lipids and proteins in wild type versus *dir1* primed leaves.

NAD^+, ATP), and plant hormones (e.g., SA and jasmonic acid). Moreover, lipids also spanned a range of lipid classes and different concentrations in the cells. They include major lipid classes, such as glycerolipids: monoradylglycerolipids (MG), diradylglycerolipids (DG), and triradylglycerolipids (TG); glycerophospholipids: glycerophosphoserines (PS), glycerophosphoinositols (PI), glycerophosphoglycerols (PG), glycerophosphoethanolamines (PE), glycerophosphocholines (PC), lyso-glycerophosphoethanolamines (LPE); sphingolipids: ceramides (Cer); and galactolipids: monogalactosyldiacylglycerol (MGDG), digalactosyldiacylglycerol (DGDG), and digalactosylmonoacylglycerol (DGMG). Interestingly, the relative abundances of the lipid classes correlate well with those detected in previous publications (**Supplementary Table 2**), in spite the ecotype differences between this study and the other studies (**Table 1**).

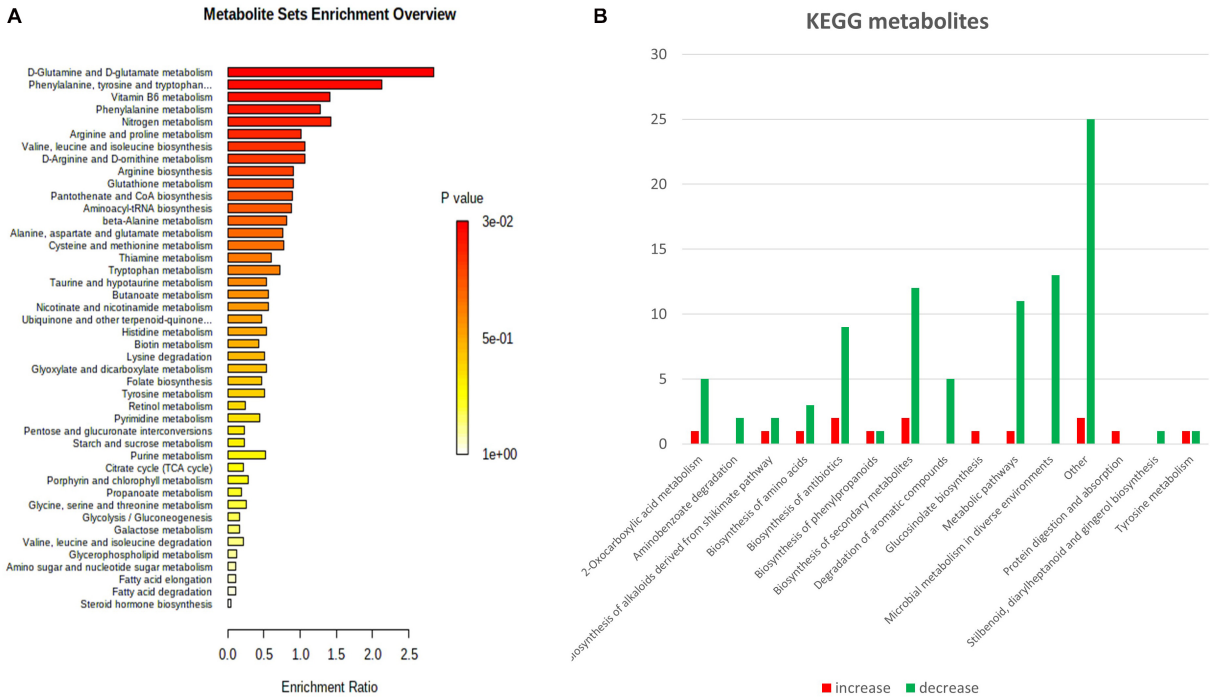

FIGURE 5 | Enriched pathways and differential metabolites and lipids in *Arabidopsis* leaves of wild type and *dir1* mutant primed by *Pst* DC 3000 treatment. **(A)** MetaboAnalyst analysis of pathway enrichment for metabolites and lipids. **(B)** KEGG pathways of metabolites/lipids that are differentially abundant in wild type versus *dir1* primed leaves.

A successful multi-omics study should not only allow for large-scale discovery of biomolecules at different abundances, but also uncover meaningful biological processes and significance. Here we employ the 3-in-1 method in a proof-of-concept study to measure changes of proteins, metabolites and lipids from each sample during SAR. A volcano plot of the protein, metabolite, and lipid data from wild type SAR (primed/control) versus SAR in the *dirl* mutant showed many differential molecular changes with significant *p*-values of less than 0.05 (**Figure 4A**). The method also showed decent reproducibility even with biological replicate samples. Of the 113 differentially abundant proteins between *dir1* primed/WS primed, 112 had coefficient of variation (CV) less than 20%. Of the 135 differential metabolites and 15 lipids, they were 91 and all 15 less than 20%, respectively. Differential metabolites and lipids were grouped and mapped to KEGG pathways and differential proteins were separately mapped to KEGG pathways (**Figure 4B**). Interestingly, the largest group of differential proteins mapped to metabolic process and metabolic pathways was the second most abundant biological process for the differential metabolites (**Figure 4B**). Protein differences in the *dir1* primed versus wild type primed plants indicate that the altered *dir1* defense responses may account for its susceptibility when compared to wild type plants. Proteins in response to stimulus and defense response pathways were the second and sixth most abundant groups, respectively (**Figure 4B**). When examining metabolites and lipids that were different between the *dir1* and wild type primed leaves, we found the largest groups related to biosynthesis of secondary metabolites, biosynthesis of antibiotics, and biosynthesis of amino acids as the first, fourth, and seventh most abundant groups, respectively. Secondary metabolites and amino acids play well-known roles in plant defense responses (Zeier, 2013; Rojas et al., 2014; Kadotani et al., 2016; Erb and Kliebenstein, 2020). Additionally, biosynthesis of antibiotics can be correlated to defense response against the biological pathogen *Pst* during priming.

To further investigate the roles of the differential metabolites and lipids, we performed a pathway enrichment analysis (**Figure 5A**), revealing enrichment of multiple amino acid metabolic pathways including: glutamine and glutamate, phenylalanine, tyrosine, tryptophan, arginine, proline, valine, leucine, isoleucine, and lysine metabolism (**Figure 5A**). They were largely decreased in the susceptible *dir1* mutant in the category of amino acid biosynthesis (**Figure 5B**). Interestingly, the protein level changes corroborate the metabolomics data (**Figure 5A**), indicating translational regulation of amino acid metabolism. The *dir1* mutant also had lower abundance of other defense related metabolites, e.g., antibiotics and secondary metabolites (**Figure 5B**). These results can help explain the susceptibility of the *dir1* mutant and the critical role of DIR1 in plant defense response. In contrast to the *dir1* mutant, the wild type plants increased the levels of these defense-related metabolites.

Since amino acid metabolism was dramatically affected in the *dir1* mutant during SAR, here we focused on mapping proteins and metabolites onto the KEGG pathways for amino acid biosynthesis (**Figure 6B**). Six proteins and two metabolites

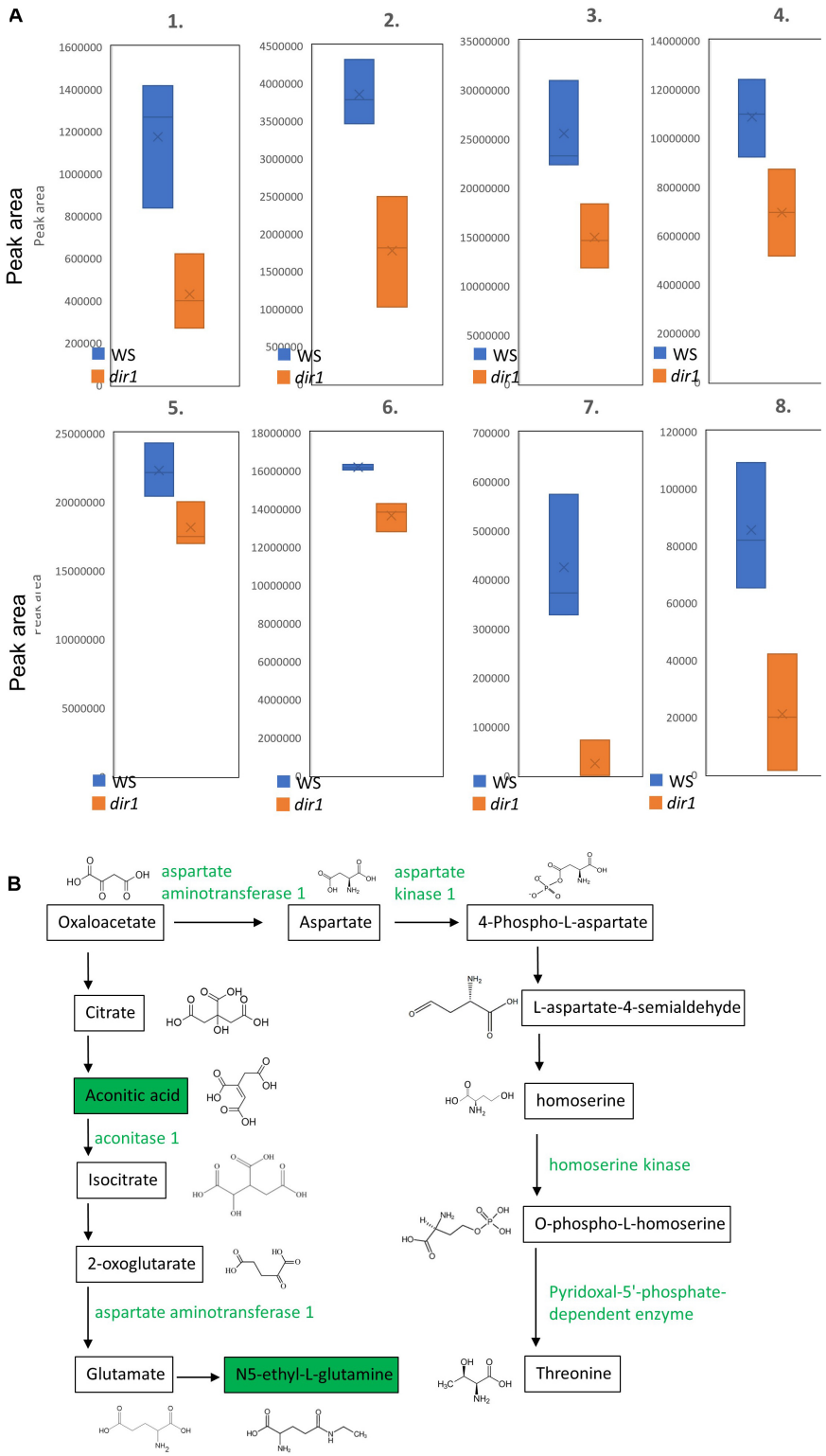

FIGURE 6 | Amino acid biosynthesis pathways with differential metabolites and proteins in leaves of wild type and *dir1* mutant primed by *Pst* DC 3000 treatment. **(A)** Box plots showing lower abundance of six proteins and two metabolites in the glutamate and threonine biosynthetic pathways. 1. aspartate kinase 1; 2. homoserine kinase; 3. pyridoxal-5′-phosphate-dependent enzyme; 4. argininosuccinate synthase; 5. aconitase 1; 6. aspartate aminotransferase 1; 7. aconitic acid; 8. N5-ethyl-L-glutamine. **(B)** KEGG pathways of metabolites/proteins related to glutamate and threonine that are differentially abundant in wild type versus *dir1* primed leaves. Green color indicates decreased abundance.

were mapped to amino acid biosynthesis, and they were related to glutamate and glutamine, the top enriched pathway for the metabolite analysis (**Figure 5A**). All the six proteins and two metabolites were decreased in the primed *dir1* mutant when compared to primed wild type plants (**Figure 6A**). As amino acid biosynthesis is closely related to plant disease resistance (Zeier, 2013; Rojas et al., 2014; Erb and Kliebenstein, 2020), DIR1 may play a role in regulating amino acid during SAR priming. Amino acid metabolism is inhibited during the SAR response of the *dir1* mutant (**Figure 5B**). A previous metabolomic study revealed that the levels of several amino acids were significantly increased in *Arabidopsis* leaves inoculated with SAR-inducing *P.syringae*, including aromatic amino acids, branched-chain amino acids, Thr and Lys, whereas Asp was decreased (Zeier, 2013). Here we found a decrease in threonine biosynthesis in the *dir1* mutant (**Figure 6B**). Additionally, Kadotani et al. (2016) found that exogenous application of glutamate to rice leaves was sufficient to induce systemic resistance against rice blast. These results are consistent with our finding that compromised amino acid metabolism may contribute to the disease susceptibility of the *dir1* mutant. The potential connection between DIR1 and amino acid metabolism is a new discovery, which needs to be further characterized in future studies.

CONCLUSION

Multi-omics has advanced our understanding of the complex molecular mechanisms underlying genetic diseases, host-pathogen interactions, and metabolic disorders important to human health and crop production. The 3-in-1 sample preparation method greatly facilitates application of proteomics, metabolomics and lipidomics technologies to tackling fundamental biological and systems biology questions. Here we demonstrated the utility and robustness of the improved method using *Arabidopsis* leaves from wild type and *dir1* mutant challenged with *Pseudomonas* pathogen that causes crop diseases. In total, we were able to profile 1849 proteins, 1967 metabolites and 424 lipids from single samples, and integrate them into pathways and networks. The high coverage of molecules has not been achieved before. In addition, integration of the data has generated interesting questions and testable hypotheses. For example, how DIR1 regulates amino acid metabolism is intriguing. Apparently, the extraction of proteins, metabolites and lipids simultaneously from the same sample (3-in-1) has the following advantages: (1) inexpensive and easy to perform as this method does not require any special reagents or kits; (2) reducing technical variations related to sample preparation of different molecules; (3) conservation of sample amount (e.g., in case of single-cell types and clinical biopsies); (4) enhancing multi-omics by high coverage, reproducibility and tight correlation between different biomolecules; (5) broadly applicable to any other cells or tissue types. Therefore, this newly improved method has great value to multi-omics and systems biology toward understanding cellular molecular networks (through hypothesis generation and hypothesis testing) important for biological functions, traits and phenotypes.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: All protein MS raw data and search results have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the data set identifier PXD023094. All metabolite and lipid MS raw data and search results have been deposited to the MetaboLights data repository with the data set identifier MTBLS2303.

AUTHOR CONTRIBUTIONS

SC and JC conceived the idea and designed and supervised the experiments. JK and LD did the sample preparation, acquired all the proteomics, metabolomics and lipidomics data, and analyzed the data. YL assisted with data analysis and interpretation. JK wrote the first draft of the manuscript. LD improved the writing. JK, YL, and SC revised the manuscript. SC finalized the manuscript for publication. All authors participated in data interpretation, manuscript preparation, and read and approved the final version of the manuscript.

FUNDING

This material is based upon work supported by the National Science Foundation under Grant No. 1920420. This work was also supported by United States Department of Agriculture Grant No. 2020-67013-32700/project accession no. 1024092 from the USDA National Institute of Food and Agriculture.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Craig Dufresne and John Orlando at the Thermo Fisher Scientific for technical assistance. The authors would also like to thank Dr. Tongjun Gu from Bioinformatics Core of Interdisciplinary Center for Biotechnology Research, University of Florida for advice in data analysis. Credit also goes to Ms. Ame Ishitani for the drawings used in **Figure 2** using ibis Paint X drawing application. This manuscript has been much improved owing to the invaluable advice from the two reviewers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.635971/full#supplementary-material

Supplementary Table 1 | Proteins, metabolites, and lipids identified in SAR primed wild type (WS ecotype) and *dir1* knockout mutant leaves.

Supplementary Table 2 | Comparison of the identified lipid classes and relative abundances with other published papers.

REFERENCES

Breitkopf, S. B., Ricoult, S. J. H., Yuan, M., Xu, Y., Peake, D. A., Manning, B. D., et al. (2017). A relative quantitative positive/negative ion switching method for untargeted lipidomics via high resolution LC-MS/MS from any biological source. *Metabolomics* 13:30. doi:10.1007/s11306-016-1157-8

Coman, C., Solari, F. A., Hentschel, A., Sickmann, A., Zahedi, R. P., and Ahrends, R. (2016). Simultaneous metabolite, protein, lipid extraction (SIMPLEX): a combinatorial multimolecular omics approach for systems biology. *Mol. Cell. Proteomics* 15, 1453–1466. doi: 10.1074/mcp.M115.053702

Dai, S., and Chen, S. (2012). "Information processing at the proteomics level," in *Springer Handbook of Bio- and Neuroinformatics (HBBNI)*, Chap. 4, ed. Nikola Kasabov (Berlin: Springer), 57–72.

David, L., Kang, J., and Chen, S. (2021). Untargeted metabolomics of *Arabidopsis* stomatal immunity. *Methods Mol. Biol.* 2200, 413–424. doi: 10.1007/978-1-0716-0880-7_20

Erb, M., and Kliebenstein, D. J. (2020). Plant secondary metabolites as defenses, regulators, and primary metabolites: the blurred functional trichotomy. *Plant Physiol.* 184, 39–52. doi: 10.1104/pp.20.00433

Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. (2000). Metabolite profiling for plant functional genomics. *Nat. Biotech.* 18, 1157–1161. doi: 10.1038/81137

Folch, J., Lees, M., and Sloane Stanley, G. H. (1957). A simple method for the isolation and purification of total lipides from animal tissues. *J. Biol. Chem.* 226, 497–509.

Geng, S., Misra, B. B., de Armas, E., Huhman, D. V., Alborn, H. T., Sumner, L. W., et al. (2016). Jasmonate-mediated stomatal closure under elevated CO2 revealed by time-resolved metabolomics. *Plant J.* 88, 947–962. doi: 10.1111/tpj.13296

Geng, S., Yu, B., Zhu, N., Dufresne, C., and Chen, S. (2017). Metabolomics and proteomics of *Brassica napus* guard cells in response to low CO2. *Front. Mol. Biosci.* 4:51. doi: 10.3389/fmolb.2017.00051

Gowda, G. A. N., and Djukovic, D. (2014). Overview of mass spectrometry-based metabolomics: opportunities and challenges. *Met. Mol. Biol.* 1198, 3–12. doi: 10.1007/978-1-4939-1258-2_1

He, Y., Dai, S., Dufresne, C. P., Zhu, N., Pang, Q., and Chen, S. (2012). Integrated proteomics and metabolomics of *Arabidopsis* acclimation to gene-dosage dependent perturbation of isopropylmalate dehydrogenases. *PLoS One* 8:e57118. doi: 10.1371/journal.pone.0057118

Higashi, Y., Okazaki, Y., Myouga, F., Shinozaki, K., and Saito, K. (2015). Landscape of the lipidome and transcriptome under heat stress in *Arabidopsis* thaliana. *Sci. Rep.* 5:10533. doi: 10.1038/srep10533

Kadotani, N., Akagi, A., Takatsuji, H., Miwa, T., and Igarashi, D. (2016). Exogenous proteinogenic amino acids induce systemic resistance in rice. *BMC Plant Biol.* 3:60. doi: 10.1186/s12870-016-0748-x

Kehelpannala, C., Rupasinghe, T. W., Hennessy, T., Bradley, D., Ebert, B., and Roessner, U. (2020). A comprehensive comparison of four methods for extracting lipids from *Arabidopsis* tissues. *Plant Methods* 16:55. doi: 10.1186/s13007-020-00697-z

Kumar, V., Sharma, A., Kaur, R., Thukral, A. K., Bhardwaj, R., and Ahmad, P. (2017). Differential distribution of amino acids in plants. *Amino Acids* 49, 821–869. doi: 10.1007/s00726-017-2401-x

Meng, L., Zhang, T., Geng, S., Scott, P. B., Li, H., and Chen, S. (2019). Jasmonate ZIM domain 7 regulated proteomic and metabolomic changes in *Arabidopsis* drought tolerance. *J. Proteomics* 196, 81–91.

Mostafa, I., Zhu, N., Yoo, M. J., Balmant, K. M., Misra, B. B., Dufresne, C., et al. (2016). New nodes and edges in the glucosinolate molecular network revealed by proteomics and metabolomics of *Arabidopsis* myb28/29 and cyp79B2/B3. *J. Proteomics* 138, 1–19. doi: 10.1016/j.jprot.2016.02.012

Nakayasu, E. S., Nicora, C. D., Sims, A. C., Burnum-Johnson, K. E., Kim, Y. M., Kyle, J. E., et al. (2016). MPLEx: a robust and universal protocol for single-sample integrative proteomic, metabolomic, and lipidomic analyses. *mSystems* 1:e43-16. doi: 10.1128/mSystems.00043-16

Niehl, A., Zhang, Z. J., Kuiper, M., Peck, S. C., and Heinlein, M. (2013). Label-free quantitative proteomic analysis of systemic responses to local wounding and virus infection in *Arabidopsis* thaliana. *J. Proteome Res.* 12, 2491–2503. doi: 10.1021/pr3010698

Rojas, C. M., Senthil-Kumar, M., Tzin, V., and Mysore, K. S. (2014). Regulation of primary plant metabolism during plant-pathogen interactions and its contribution to plant defense. *Front. Plant Sci.* 5:17. doi: 10.3389/fpls.2014.00017

Salem, M., Bernach, M., Bajdzienko, K., and Giavalisco, P. (2017). A simple fractionated extraction method for the comprehensive analysis of metabolites, lipids, and proteins from a single sample. *J. Vis. Exp.* 124:55802. doi: 10.3791/55802

Salem, M. A., Juppner, J., Bajdzienko, K., and Giavalisco, P. (2016). Protocol: a fast, comprehensive and reproducible one-step extraction method for the rapid preparation of polar and semi-polar metabolites, lipids, proteins, starch and cell wall polymers from a single sample. *Plant Methods* 12:45. doi: 10.1186/s13007-016-0146-2

Wu, S., Tohge, T., Cuadros-Inostroza, Á, Tong, H., Tenenboim, H., Kooke, R., et al. (2018). Mapping the *Arabidopsis* metabolic landscape by untargeted metabolomics at different environmental conditions. *Mol. Plant* 11, 118–134. doi: 10.1016/j.molp.2017.08.012

Zeier, J. (2013). New insights into the regulation of plant immunity by amino acid metabolic pathways. *Plant Cell Environ.* 36, 2085–2103. doi: 10.1111/pce.12122

Zhang, H., Liu, P., Guo, T., Zhao, H., Bensaddek, D., Aebersold, R., et al. (2019). *Arabidopsis* proteome and the mass spectral assay library. *Sci. Data* 6:278. doi: 10.1038/s41597-019-0294-0

Zybailov, B., Friso, G., Kim, J., Rudella, A., Rodríguez, V. R., Asakura, Y., et al. (2009). Large scale comparative proteomics of a chloroplast Clp protease mutant reveals folding stress, altered protein homeostasis, and feedback regulation of metabolism. *Mol. Cell Proteomics* 8, 1789–1810. doi: 10.1074/mcp.M900104-MCP200

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: 19 May 2021
doi: 10.3389/fgene.2021.675260

Aluminum or Low pH – Which Is the Bigger Enemy of Barley? Transcriptome Analysis of Barley Root Meristem Under Al and Low pH Stress

Miriam Szurman-Zubrzycka[1], Karolina Chwiałkowska[2], Magdalena Niemira[3], Mirosław Kwaśniewski[2], Małgorzata Nawrot[1], Monika Gajecka[1], Paul B. Larsen[4] and Iwona Szarejko[1]**

[1] Institute of Biology, Biotechnology and Environmental Protection, Faculty of Natural Sciences, University of Silesia in Katowice, Katowice, Poland, [2] Centre for Bioinformatics and Data Analysis, Medical University of Bialystok, Bialystok, Poland, [3] Clinical Research Centre, Medical University of Bialystok, Bialystok, Poland, [4] Department of Biochemistry, University of California, Riverside, Riverside, CA, United States

OPEN ACCESS

Edited by:
Joanna Jankowicz-Cieslak, International Atomic Energy Agency, Austria

Reviewed by:
Zhong-Hua Chen, Western Sydney University, Australia
Loriana Demecsová, Center of Plant Science and Biodiversity, Institute of Botany, Slovak Academy of Sciences, Slovakia

***Correspondence:**
Miriam Szurman-Zubrzycka
miriam.szurman@us.edu.pl
Iwona Szarejko
iwona.szarejko@us.edu.pl

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

Received: 02 March 2021
Accepted: 13 April 2021
Published: 19 May 2021

Citation:
Szurman-Zubrzycka M, Chwiałkowska K, Niemira M, Kwaśniewski M, Nawrot M, Gajecka M, Larsen PB and Szarejko I (2021) Aluminum or Low pH – Which Is the Bigger Enemy of Barley? Transcriptome Analysis of Barley Root Meristem Under Al and Low pH Stress. Front. Genet. 12:675260. doi: 10.3389/fgene.2021.675260

Aluminum (Al) toxicity is considered to be the most harmful abiotic stress in acidic soils that today comprise more than 50% of the world's arable lands. Barley belongs to a group of crops that are most sensitive to Al in low pH soils. We present the RNA-seq analysis of root meristems of barley seedlings grown in hydroponics at optimal pH (6.0), low pH (4.0), and low pH with Al (10 μM of bioavailable Al^{3+} ions). Two independent experiments were conducted: with short-term (24 h) and long-term (7 days) Al treatment. In the short-term experiment, more genes were differentially expressed (DEGs) between root meristems grown at pH = 6.0 and pH = 4.0, than between those grown at pH = 4.0 with and without Al treatment. The genes upregulated by low pH were associated mainly with response to oxidative stress, cell wall organization, and iron ion binding. Among genes upregulated by Al, overrepresented were those related to response to stress condition and calcium ion binding. In the long-term experiment, the number of DEGs between hydroponics at pH = 4.0 and 6.0 were lower than in the short-term experiment, which suggests that plants partially adapted to the low pH. Interestingly, 7 days Al treatment caused massive changes in the transcriptome profile. Over 4,000 genes were upregulated and almost 2,000 genes were downregulated by long-term Al stress. These DEGs were related to stress response, cell wall development and metal ion transport. Based on our results we can assume that both, Al^{3+} ions and low pH are harmful to barley plants. Additionally, we phenotyped the root system of barley seedlings grown in the same hydroponic conditions for 7 days at pH = 6.0, pH = 4.0, and pH = 4.0 with Al. The results correspond to transcriptomic data and show that low pH itself is a stress factor that

causes a significant reduction of root growth and the addition of aluminum further increases this reduction. It should be noted that in acidic arable lands, plants are exposed simultaneously to both of these stresses. The presented transcriptome analysis may help to find potential targets for breeding barley plants that are more tolerant to such conditions.

Keywords: barley, RNA-Seq, transcriptome, low pH, aluminum (Al), stress, root meristem

INTRODUCTION

One of the biggest problems of modern agronomy and a constraint for world agriculture is the progressive acidification of arable lands, caused by industrial pollution and overuse of ammonia- and amide-containing fertilizers. It is estimated that up to 50% of arable lands worldwide are acidic, with a pH below 5.5 (Von Uexküll and Mutert, 1995; Singh et al., 2017; Barros et al., 2020). The majority of crops growing in acidic soils show significant yield losses - up to 80%, depending on the species (Sade et al., 2016). The primary factor responsible for reduced yield in acidic soils is aluminum (Al), the third most abundant element (after oxygen and silicon) and the most common metal in the Earth's crust. In alkaline and near-neutral soils, Al is bound in various minerals or occurs in forms that are mostly harmless to plants. However, in acidic soils, Al is released from clay minerals in the form of $[Al(H_2O)_6]^{3+}$, for simplicity often referred to as Al^{3+} ions, that are bioavailable for plants and highly phytotoxic (Bhalerao and Prabhu, 2013; Sade et al., 2016; Rahman et al., 2018).

The first symptom of Al toxicity in acidic soils is reduction of root growth, resulting from inhibition of both elongation and division rates of root cells. As a consequence, the plant suffers from reduced water and nutrient uptake, which leads to plant growth retardation and, finally, yield reduction. It has been shown that Al^{3+} ions are highly reactive and there are many potential Al binding sites in plant cells. Al^{3+} ions interact with the cell wall, cell membrane, and symplastic components; therefore they interfere with a broad spectrum of physical and cellular processes (Kochian et al., 2005, 2015). The first structure in roots that Al^{3+} ions interact with is the apoplast. Aluminum ions directly cross-link the negatively charged carboxyl groups of pectins in the cell wall, which leads to its stiffening and inhibition of cell elongation (Kopittke et al., 2015). A significant part of absorbed Al (30–90%) is accumulated in the apoplast (Silva, 2012; Gupta et al., 2013). Al^{3+} ions interact also with the negatively charged surface of the plasmalemma and displace other ions like Ca^{2+} from phospholipid head groups, which destabilizes the cell membrane and alters its fluidity. It also leads to depolarization of the plasmalemma, which affects cellular ion homeostasis. Additionally, the replacement of Ca^{2+} by Al^{3+} in the plasma membrane increases Ca^{2+} content in the apoplast and therefore stimulates callose deposition. Accumulation of callose inhibits intercellular transport through plasmodesmata (Kochian et al., 2005).

A fraction of Al^{3+} that enters the cytosol may interact with cytoskeletal elements and disturb its dynamics directly or indirectly through modification of e.g., Ca^{2+} signaling cascade. The disturbances in spatial orientation of the cytoskeleton may affect cell expansion and lead to morphological changes and distortion of roots (Sade et al., 2016). Moreover, there is extensive evidence that Al^{3+} ions enter the nucleus, cause DNA damage (Silva et al., 2000; Min et al., 2009; Jaskowiak et al., 2018), and activate the DDR (DNA Damage Response) pathway, which additionally leads to inhibition of cell divisions (Rounds and Larsen, 2008; Nezames et al., 2012; Szurman-Zubrzycka et al., 2019). Furthermore, exposure to Al induces oxidative stress. It promotes the overproduction of Reactive Oxygen Species (ROS) and alters the activity of enzymes responsible for maintaining ROS homeostasis in cells, such as superoxide dismutase and ascorbate peroxidase (Yamamoto et al., 2003; Guo et al., 2004; Jones et al., 2006). The Al-induced overproduction of ROS leads to the peroxidation of lipids and proteins and further DNA damage (Achary and Panda, 2009).

In general, plants evolved two main strategies to cope with Al ions: (1) Al exclusion mechanisms and (2) Al tolerance mechanisms. The first one is based on the production of organic acids (OAs) and their exudation outside the cell. The OAs, such as citric and malic acids, chelate Al in the rhizosphere which prevents its entrance to the root cells. The second strategy deals with Al that entered the cell. The internal OAs and other organic compounds form Al-complexes that are detoxified in vacuoles or reallocated to the upper, less Al-sensitive parts of the plant (reviewed in Kochian et al., 2015; Riaz et al., 2018).

Taken together Al induces a broad spectrum of changes and responses in plant cells. Al stress is considered as the main growth-limiting factor in acidic soils and the second, after drought, most serious abiotic stress to crop production worldwide (Kochian et al., 2015). Barley (*Hordeum vulgare* L.), which is the 4th most important cereal crop, is known to be one of the most sensitive to Al cereal species (Ishikawa et al., 2000; Wang et al., 2006), but its response to Al has not been studied at the whole transcriptome level. Besides, our preliminary studies have shown that barley is very sensitive not only to phytotoxic Al^{3+} ions in acidic conditions, but also to the low pH of growth medium alone. The low pH causes so called H^+ or proton toxicity. In naturally occurring acidic arable lands, plants are exposed simultaneously to both of these stressors (low pH and Al), as Al becomes soluble at pH below 5.5. However, growing plants in the hydroponic solution makes it possible to examine at the gene expression level the plant response to the stress triggered by low pH without Al, and to reveal changes caused by Al toxicity itself.

Here we show, for the first time, the global transcriptome profile of barley root tips grown in hydroponics at the optimal pH (6.0), low pH (4.0), and low pH with Al (10 μM of bioavailable Al^{3+} ions) in two independent, short-term and long-term, experiments.

MATERIALS AND METHODS

Plant Material

The spring barley (*Hordeum vulgare* L.) cultivar 'Sebastian' bred by the Danish company Sejet Plantbreeding was used as plant material in the presented study. This cultivar is a parent variety of barley TILLING population (*Hor*TILLUS) that was developed at the Department of Genetics, University of Silesia in Katowice (Szurman-Zubrzycka et al., 2018) and is extensively used in functional genomics studies.

Examination of Root Parameters of Barley Seedlings Grown at Low pH and Treated With Aluminum

Hydroponic Experiment

The low pH and aluminum treatments were performed in a hydroponic environment as described previously (Szurman-Zubrzycka et al., 2019). Briefly, seeds of barley cv. 'Sebastian' were surface-sterilized in 5% sodium hypochlorite and incubated in the dark at 4°C for stratification. Then the seeds were put on Petri dishes filled with moist filter paper and placed in a growth chamber at 25°C in the dark. After 48 h, the germinated seeds were transferred to 4.5 L hydroponic containers with Magnavaca solution (Magnavaca et al., 1987) at pH = 6.0, pH = 4.0, or pH = 4.0 with 10 μM of bioavailable Al^{3+} ions. The concentration of 10 μM of bioavailable Al^{3+} ions was calculated with GEOCHEM-EZ software (Shaff et al., 2010) and it corresponds to 50 μM of nominal $AlCl_3$ added to the Magnavaca medium at pH = 4.0. The maximum of 12 seedlings were placed in one container that was considered as one replicate and each experimental combination was set up as three replicates. The seedlings were grown in hydroponics for 7 days (7 d) under controlled conditions: 20°C/18°C (day/night), 16/8 h photoperiod, 250 μM m^{-2} s^{-1} light intensity.

Root System Scanning and Analysis

After 7 days, the seedlings were removed from containers and their roots were preserved in 50% ethanol and scanned in water in waterproof trays. For scanning, the EPSON PERFECTION V700 PHOTO scanner with a dual-lens system was used accompanied by WinRHIZO software (Regent Instruments). The root parameters were calculated, based on the obtained scans, with the use of WinRHIZO and SmartRoot[1] software. Statistical analyses were performed using ANOVA ($P < 0.05$) followed by Tukey's Honest Significant Difference test (Tukey HSD test, $P < 0.05$).

[1]https://smartroot.github.io/SmartRoot-Installation/

Analysis of Root Meristem Transcriptome of Barley Seedlings Grown at Low pH and Treated With Aluminum

Two independent experiments, short- and long-term, were performed for transcriptome analysis.

Short-Term Experiment

The seeds of barley cv. 'Sebastian' were germinated as described in section "Hydroponic Experiment." Germinated seeds were then transferred to 4.5 L hydroponic containers with Magnavaca medium at pH = 6.0 (three containers) and pH = 4.0 without aluminum (six containers). A maximum of 12 seedlings were placed in one container and this was considered as one replicate. After 48 h of seedlings growth, the root meristems (of approximately 1–2 mm length) were collected from three containers with solution at pH = 6.0 and three containers with solution at pH = 4.0, as control samples without Al. Subsequently, the aluminum (10 μM of bioavailable Al^{3+} ions) was added to the remaining three containers with Magnavaca solution at pH = 4.0. After 24 h of Al treatment the root meristems were collected, as Al-treated samples (**Figure 1A**). The collected root meristems were stored in RNAlater at 4°C for several days for further RNA isolation.

Long-Term Experiment

The seeds of barley cv. 'Sebastian' were germinated as described in section "Hydroponic Experiment." Similarly as in the short-term experiment, germinated seeds were transferred to 4.5 L hydroponic containers with Magnavaca solution adjusted to pH = 6.0 (three containers) and pH = 4.0 without aluminum (six containers). After 48 h, 10 μM of bioavailable Al^{3+} ions were added to three containers at pH = 4.0. After further 7 days of seedlings growth, the root meristematic tissue was collected in RNAlater (Invitrogen), as pH = 6.0, pH = 4.0 and Al-treated samples (**Figure 1B**).

RNA Isolation, Preparation of RNA-seq Libraries and Sequencing

For RNAseq analysis, mRNA was isolated from root tips with the use of the Dynabeads mRNA DIRECT Micro Kit (Thermo Fisher Scientific). Root meristems from at least eight plants from one hydroponic container were considered as one repetition (with an average of five root meristems per plant). The RNA-seq libraries were prepared using the TruSeq Stranded mRNA kit (Illumina) according to manufacturer's instructions. The quality of the prepared RNA-seq libraries was assessed using the TapeStation device (Agilent) and the High Sensitivity DNA ScreenTape kit (Agilent). The concentration of fragments in the libraries was measured with a Qubit fluorimeter (Thermo Fisher Scientific).

For cluster generation, the barcoded libraries were pooled with equimolar concentrations. The libraries from the short-term experiment were sequenced in the paired end (PE) mode 2 × 76 bp, six barcoded samples per lane in the Illumina HiSeq 4000 sequencer at the Genomics and Epigenomics

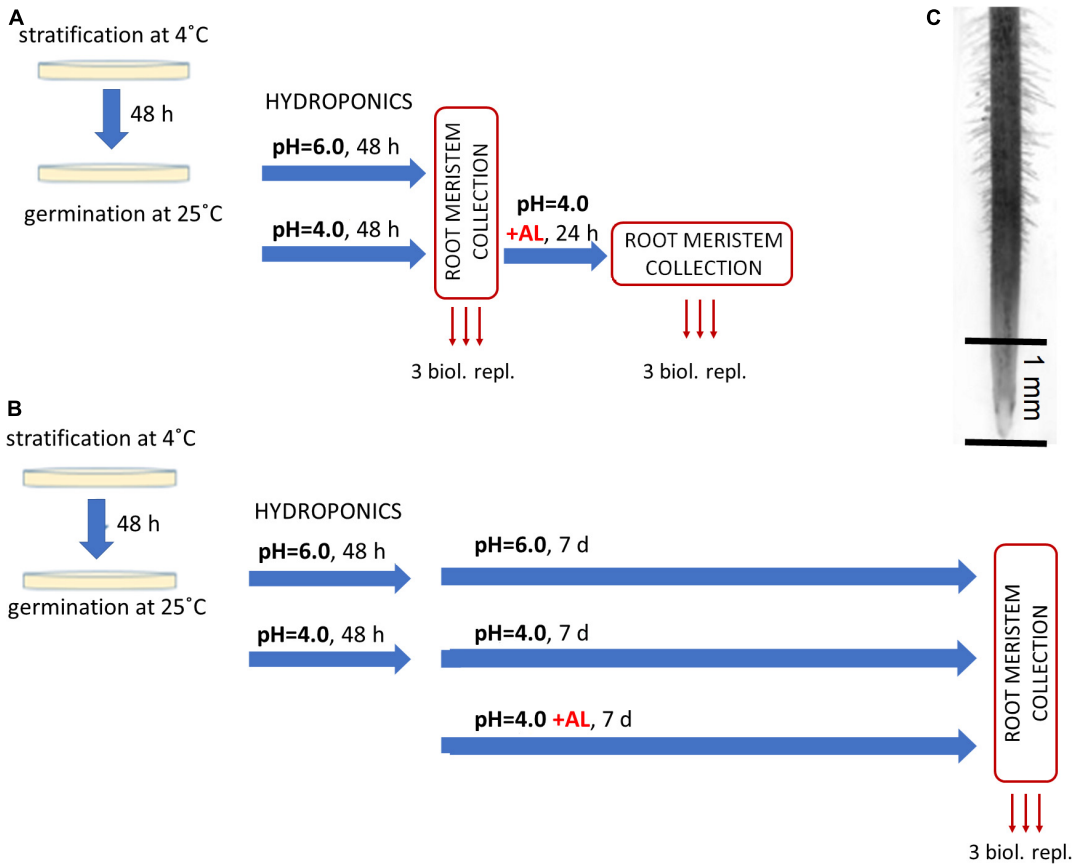

FIGURE 1 | Experimental setup for RNA-Seq analysis. After 48 h of germination on Petri dishes barley seedlings were transferred to hydroponics with Magnavaca medium adjusted to pH = 6.0 or pH = 4.0. After another 48 h $AlCl_3$ was added to the medium (10 μM of bioavailable Al^{3+} ions). **(A)** Short-term Al treatment. Root meristems were taken for RNA isolation just before Al addition (as control at pH = 6.0 and pH = 4.0) and after 24 h of Al treatment. **(B)** Long-term Al treatment. Root meristems were taken for RNA isolation from seedlings 7 days after Al addition and from seedlings grown at pH = 6.0 and pH = 4.0 without Al at the same time point. **(C)** The picture of barley root with an indication of its part (root meristem, ∼1 mm of root tip) that was taken for RNA isolation.

Laboratory, Clinical Research Centre of the Medical University of Bialystok (Poland). The libraries from the long-term experiment were sequenced in the paired end (PE) mode 2 × 150 by the Novogene company (Illumina platform). On average, 59.3 (±14.6) mln reads were obtained per each sample (single biological replicate).

Bioinformatic Analysis of RNA-seq Data

BCL files were demultiplexed and converted to fastq files using bcl2fastq (Illumina, San Diego, CA, United States) with an adapters removal step. The quality of the obtained sequencing data was assessed before the analysis and after each of its stages, using the FastQC (The Babraham Institute, Cambridge, United Kingdom) and MultiQC (Ewels et al., 2016) tools. Due to different read lengths in both experiment batches, reads were initially trimmed to the length of 75 bp with BBduk (DOE Joint Genomes Institute, Walnut Creek, CA, United States). Then, quality trimming and filtering was preformed using Sickle tool[2] under PHRED of 15, N bases removal and minimal length of 20 bp for one mate in the PE mode. The remaining ribosomal RNA reads were then removed using the SortMeRNA software (Kopylova et al., 2012). Filtered non-rRNA reads were mapped to the second version of the reference genome sequence assembly of barley cv. 'Morex' (Morex V2; Leibniz Institute of Plant Genetics and Crop Plant Research – IPK; Monat et al., 2019) with the splice-aware aligner STAR (Dobin et al., 2013) using two pass mode without non-canonical motifs. Mapping parameters were adjusted to the Morex V2 genome annotation from gff3 file provided by IPK, with regard to mates gaps and intron lengths. Only uniquely mapping reads were allowed with maximum 0.05 mismatch rate over read length. The quality of mapping was assessed with QualiMap (Okonechnikov et al., 2016) as well as SAMStat (Lassmann et al., 2011). We applied the high confidence (HC) set of gene annotations in the Morex V2 assembly and counted reads mapping to genes annotated in the gff3 using GeneCounts from the quantMode in STAR mapper. The analysis of differences in gene expression levels between samples was performed with the DESeq2 tool (Love et al., 2014). Raw read count matrices were used as

[2]https://github.com/najoshi/sickle

an input and genes without any expression detected were removed. Libraries size factors were estimated using median ratio method and further used in all size normalization steps. Then DESeq function was called on the whole dataset and covered the following steps: sequencing depth normalization between the samples, gene – wise dispersion estimation across all samples, and fitting a negative binomial generalized linear model (GLM) under Wald statistics to each gene. Using a formula with condition factors we applied contrasts for each desired comparison to the results with usage of Cook's cut-off and independent filtering. Statistical analyzes were performed based on the results obtained from three biological replicates. Differentially Expressed Genes (DEGs) were identified under $\alpha = 0.05$ after *P*-value correction for multiple comparisons using the Benjamini and Hochberg False Discovery Rate procedure (FDR) and $\log_2$FoldChange ($\log_2$FC) ≥ 1 or ≤ -1. Exploratory analysis of RNA-Seq data including clustering analysis and Principal Component Analysis (PCA) were carried out with the use of R environment tools. For data inspection and visualization, counts were subjected to regularized logarithm transformation (rlog) to get log2-scaled data that is approximately homoscedastic and normalized with respect to library size. PCA was performed with prcomp function and results were visualized as bi-plots using 'ggplot2' library. Hierarchical clustering of samples was performed based on distance expressed as an inverse of Pearsons's correlation coefficient and applying Ward D2 linkage algorithm. Normalized and rlog transformed expression values were scaled and centered to be relatively represented as z-scores. Heatmaps were visualized with 'heatmap.2' function from 'gplots' R library. For k-means clustering we have identified an optimal number of samples clusters with Silhouette (Rousseeuw, 1987), Elbow method (Halkidi et al., 2001), and Hubert statistics (Dalton et al., 2009), and applied a cluster number shown by minimum two of three used models. K-means clustering was conducted using 'k-means' function from 'clusters' R package with 1000 initial resampling and 20 iterations. Each gene scores were calculated as correlations with the cluster cores. Expression profiles were visualized with 'ggplot2' library. To identify overrepresented biological processes, gene annotation and Gene Ontology (GO) enrichment analysis were carried out using 'clusterProfiler' R package and hypergeometric test under $\alpha = 0.05$ after *P*-value correction for multiple comparisons using FDR. A set of all genes detected under investigated conditions in all of the samples was used as a background for over-representation analyses. Gene Ontology terms were recovered from the gff3 file deposited in the IPK database with Morex V2 reference genome assembly. Over-representation results were visualized on dot-plots using internal plotting function from 'clusterProfiler.'

RT-qPCR Analysis of Gene Expression

The RNAqueous Kit (Thermo Fisher Scientific) was used for RNA isolation from root meristems for RT-qPCR analysis. Root meristems were isolated in the same way and at the same time points of experiments as for RNA-seq analysis. Isolated samples were evaluated using ND-1000 spectrophotometer (Thermo Fisher Scientific). Five hundred ng of total RNA was taken for RQ1 DNase (Promega) treatment and reverse transcribed using a RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher Scientific) with Oligo(dT) primers in a 20 μL reaction mix. The RT-qPCR reaction was prepared in a 10 μL volume using a LightCycler® 480 SYBR Green I Master (Roche) in two technical replicates. A volume of 2.5 μL of obtained cDNA diluted beforehand fivefold was added to the reaction mix. The primers used in the analysis were designed with Primer3 (Untergasser et al., 2012) and are listed in **Supplementary Material 1**. The RT-qPCR analysis was performed using a LightCycler 480 (Roche) under the following reaction conditions: initial denaturation 5 min at 95°C, followed by 10 s at 95°C, 20 s at a temperature specific for the primers, 10 s at 72°C, repeated in 40 cycles. Denaturation for the melt curve analysis was conducted for 5 s at 95°C, followed by 1 min at 65°C and heating up to 98°C (0.1°C/s for the fluorescence measurement). The qPCR efficiency and the Ct values were determined using LinRegPCR (Ruijter et al., 2009) and used for calculation of relative expression level. Two genes, *H2A* (*Histone H2A*) and *EF1* (*Translation Elongation Factor 1-a*) used as internal controls were selected based on the stability of their expression using NormFinder (Andersen et al., 2006) and BestKeeper (Pfaffl et al., 2004). The relative expression level was calculated using the $\Delta\Delta$Ct method (Livak and Schmittgen, 2001) and calibrated to root meristems sampled from pH = 6.0 or pH = 4.0. The *t*-Student test was applied to determine the significant differences (at $P < 0.05$) between the compared samples.

RESULTS

Changes in the Barley Root System in Response to Low pH and Al Stress

To evaluate the influence of low pH and aluminum on barley root growth, we have examined in detail the root system of seedlings grown for 7 days in hydroponic conditions at pH = 6.0, pH = 4.0, and pH = 4.0 with 10 μM of bioavailable Al^{3+} ions. It has been clearly shown that low pH alone causes a significant reduction of root growth, and the addition of aluminum further inhibits its development (**Figure 2**).

Neither the low pH nor the aluminum caused any change in the number of seminal roots (**Figure 3A**). However, the length of the seminal roots was significantly affected by both stressors. The longest root of plants grown at pH = 4.0 were half shorter than those grown at pH = 6.0, and the longest root of plants grown in a medium with 10 μM of bioavailable Al were half shorter than those grown at pH = 4.0 (**Figure 3B**). Similarly, the total length of all seminal roots was reduced almost by 50% by low pH and further reduced by Al (**Figure 3C**). The development and growth of lateral roots of barley seedlings were affected even more. The number of lateral roots produced by the plant decreased from 385 to 152 due to the low pH (60% reduction), and to 52 due to Al exposure (65% reduction in relation to pH = 4.0) (**Figure 3D**). These roots were also drastically shortened. The summary length of all lateral roots was reduced by half by low pH and further reduced by 95% under aluminum stress compared to low pH conditions (**Figure 3E**).

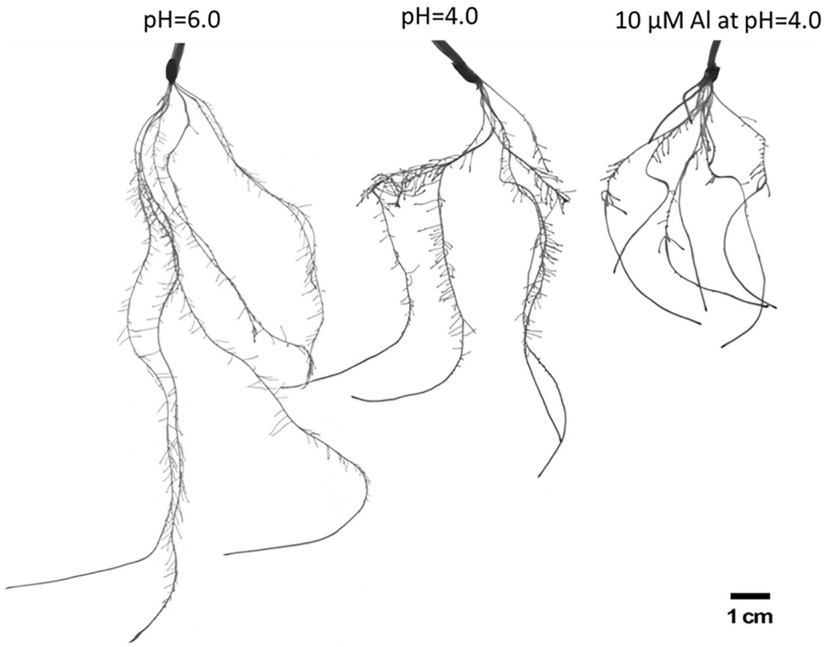

FIGURE 2 | The comparison of root systems of barley cv. 'Sebastian' after 7 days growth in hydroponics in Magnavaca medium at pH = 6.0, pH = 4.0, and with 10 µM of bioavailable Al^{3+} ions at pH = 4.0.

As a consequence, the total length of the whole root system was reduced to 53% by low pH itself and to 17% by Al stress at low pH, compared to optimal conditions of pH = 6.0 (**Figure 3F**). Interestingly, the diameter of the roots was also altered. Both factors, low pH and Al, caused a slight increase of root diameter (**Figure 3G**).

RNAseq Data Processing Statistics

Two independent experiments, short- and long-term, were performed for transcriptome analysis with RNAseq. Nine samples were collected in the short-term experiment: three samples from root meristems grown at pH = 6.0, three samples from root meristems grown at pH = 4.0 and three samples from root meristems treated with Al for 24 h at pH = 4.0. Similarly, another nine samples were collected in the long-term experiment: from root meristems of plants grown at pH = 6.0 (three samples), pH = 4.0 (three samples) and plants treated for 7 days with Al (three samples). In total, RNA-Seq libraries were constructed from 18 samples and subjected to sequencing in the paired-end mode (PE).

In the short-term experiment, soft trimming, filtering and exclusion of reads originating from rRNAs (main source of discarded reads) yielded a final mean per sample value of 19.2 (±3.8) mln paired end (PE) reads. On average 95.6% (±0.3%) of them were uniquely mapped to the reference genome, which indicates a high mapping rate (**Table 1**). In the long-term experiment, an average of 13.8 (±3.3) million clean PE reads was obtained, and a high rate of 88.4% (±4.3%) of them uniquely mapped to the barley genome (**Table 1**). PCA of obtained RNA-Seq data showed the significant differentiation of samples grown at pH = 6.0, pH = 4.0, and treated with Al in both experiments (**Supplementary Material 2**). Biological replicates from the same time-point clustered together, and PC1 explaining most of the variability (70.2% in the short-term experiment and 88.1% in the long-term experiment) corresponded to the applied treatment. The differences in gene expression were analyzed with DESeq2 tool and DEGs were identified under $\alpha = 0.05$ after *P*-value FDR correction. We further analyzed genes with $\log_2$FoldChange ($\log_2$FC) ≥ 1 or ≤ -1 as DEGs. In the short-term experiment, 1899 genes were differentially expressed in root meristems grown at low pH (4.0) when compared to those grown at pH = 6.0 and 986 genes were differentially expressed after exposure to Al for 24 h. In the long-term experiment, 870 genes were differentially expressed by low pH and 5873 by Al treatment for 7 days. The statistical significance of the results and magnitude of changes are shown on Volcano plots (**Figure 4**). To confirm obtained RNA-Seq results, four differentially expressed genes (DEGs) were checked using RT-qPCR method. The results confirmed the direction of change of expression as detected by RNA-seq (**Supplementary Material 3**).

Global Transcriptome Analysis of Barley Root Meristems in the Short-Term Experiment

Surprisingly, in the short-term experiment, more genes were differentially expressed ($\log_2$FC ≥ 1 or ≤ -1) in root meristems of barley plants grown at pH = 4.0 in relation to pH = 6.0, than in plants treated for 24 h with Al compared to plants grown at pH = 4.0 without Al (**Figure 5**). In total, the expression of 1899

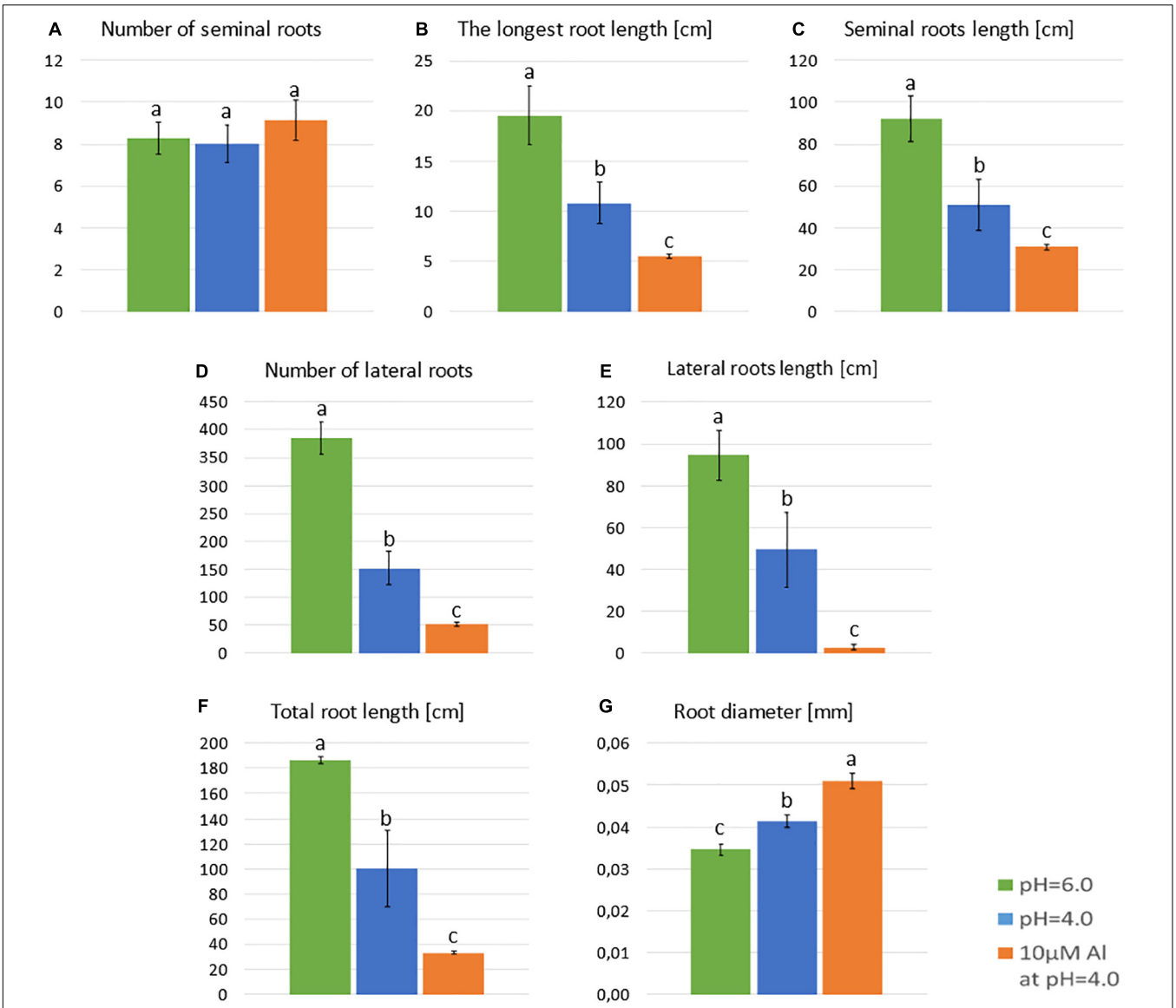

FIGURE 3 | Main parameters of 'Sebastian' roots after 7 days growth in hydroponics in Magnavaca medium at pH = 6.0, pH = 4.0, and with 10 μM of bioavailable Al^{3+} ions at pH = 4.0. **(A)** The number of seminal roots per plant. **(B)** The length of the longest seminal root. **(C)** Total length of seminal roots. **(D)** The number of lateral roots per plant. **(E)** The length of all lateral roots. **(F)** The summary length of the whole root system (seminals + laterals). **(G)** The average diameter of roots. Statistical analyses were performed using ANOVA ($P < 0.05$) followed by Tukey's honestly significant difference test (Tukey HSD test, $P < 0.05$). Significant differences are indicated by different letters.

genes was altered after 48 h of growth at low pH. Among them, 1361 were upregulated and 538 were downregulated. Treatment with 10 μM Al at pH = 4.0 for 24 h led to a change of expression of 986 genes. Majority of these genes (883) were upregulated, whereas 103 were downregulated. These numbers suggest that growing of barley seedlings in a short-term hydroponics at the low pH (4.0) has a great impact on the transcriptome profile of root meristems, even greater than short term (24 h) Al exposure at pH = 4.0 when compared to low pH conditions without Al. Correspondingly, the length of seminal roots in the short-term experiment was more affected by low pH itself than by addition of aluminum for 24 h (**Supplementary Material 4**).

Genes With Expression Altered by Low pH in the Short-Term Experiment

After 48 h of seedling growth in hydroponics at pH = 4.0, 72% of DEGs were upregulated and 38% were downregulated compared to seedlings grown at pH = 6.0 (**Figures 5**, **6A**). The GOs term enrichment allowed identification of overrepresented groups of up- and downregulated genes (**Figure 7A**). Among upregulated ones, a cluster of genes related to maintaining REDOX homeostasis stood out the most. The peroxidase HORVU.MOREX.r2.2HG0129730 had the highest fold change in the gene expression level (log_2FoldChange = 7.88, **Supplementary Material 5**). Almost sixty other genes encoding

TABLE 1 | The statistics of data filtering and mapping steps for 18 analyzed PE RNA-Seq samples.

Short-term experiment				Long-term experiment			
Sample	**Filtered reads**	**Mapped reads**	**Mapping rate [%]**	**Sample**	**Filtered reads**	**Mapped reads**	**Mapping rate [%]**
pH6_1	22747533	21674211	95,28159	pH6_1	18337028	17344706	94,58843
pH6_2	25895857	24867094	96,02731	pH6_2	11004027	10292225	93,53144
pH6_3	16383269	15661893	95,59687	pH6_3	11440974	9777453	85,45997
pH4_1	18405817	17654616	95,91868	pH4_1	15675008	14481444	92,38556
pH4_2	13897267	13300449	95,7055	pH4_2	19387421	16840937	86,86528
pH4_3	20728621	19856210	95,79127	pH4_3	9900989	8133049	82,1438
Al_1	21361770	20383233	95,41921	Al_1	13081680	11328414	86,59755
Al_2	18438423	17588979	95,39308	Al_2	13855686	11808184	85,22266
Al_3	15398531	14667448	95,25225	Al_3	11499869	10276899	89,36536
Average	**19250788**	**18406015**	**95,59842**	**Average**	**13798078**	**12253701**	**88,46223**

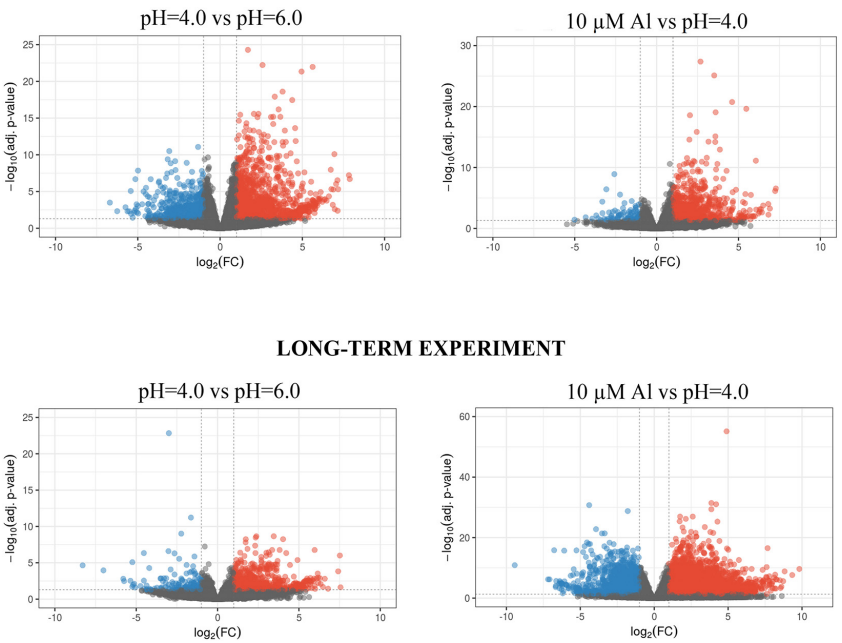

FIGURE 4 | Volcano plots of differential expression analysis of barley root meristems grown at pH = 4.0 compared to pH = 6.0 and treated with 10 µM of bioavailable Al^{3+} compared to pH = 4.0 in the short-term and long-term experiments. Each dot represents one gene. Dashed lines determine cut-off of | log2FC| ≥ 1 and FDR adjusted *P*-value ≤ 0.05. Blue dots correspond to gene expression down-regulation, whereas red dots – up-regulation.

proteins of peroxidase activity were highly upregulated by low pH. The great number of genes with the function assigned in cellular oxidant detoxification, hydrogen peroxide catabolic process or response to oxidative stress also showed increased expression. Among them are genes encoding, e.g., cytochrome P450 proteins that are monooxygenases involved in the formation of ROS (32 genes upregulated, with the highest $\log_2$FC = 5.68); laccases that are multicopper oxidases (12 genes upregulated, with the highest $\log_2$FC = 7.17); glutathione S-transferases (GSTs) that are detoxifying enzymes helping to protect cells from oxidative damage (10 genes upregulated, with the highest $\log_2$FC = 5.15); or aldehyde oxidases (3 genes upregulated, with the highest $\log_2$FC = 7.17). On the other hand, some other genes encoding proteins maintaining ROS levels were downregulated. The expression of 10 genes for cytochrome P450 (out of 16 downregulated) was highly decreased, with $\log_2$FoldChange > –3.0, three genes encoding GSTs and one gene encoding laccase were also downregulated, which further indicates that acidification of the environment contributes to ROS balance disruption. These data clearly show that lowering the pH from 6.0 to 4.0 induces oxidative stress in barley root meristems.

Another cluster of significantly overrepresented groups of genes upregulated at pH = 4.0 was related to cell wall development. The GOs term enrichment indicated that there are groups of genes involved, for instance, in

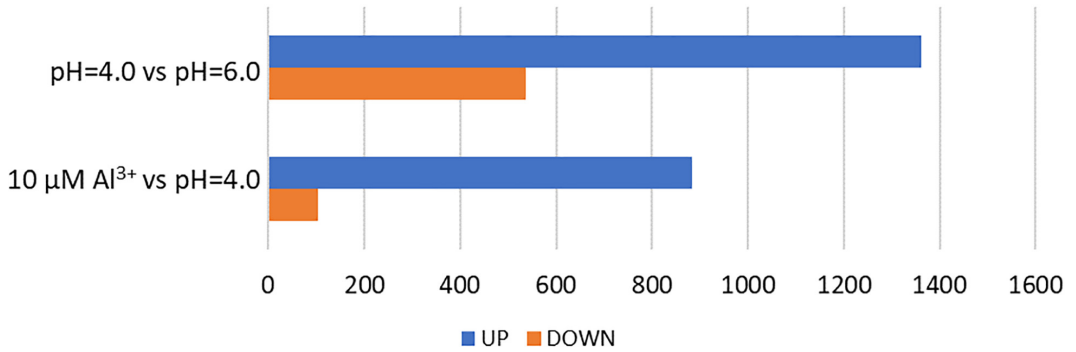

FIGURE 5 | The number of differentially expressed genes (DEGs) in the transcriptome of barley root meristems in the short-term experiment. DEGs were determined for plants grown at pH = 4.0 compared to pH = 6.0; and after 24 h treatment with 10 µM Al at pH = 4.0 compared to pH = 4.0 without Al^{3+} ions (DEGs were identified under $\alpha = 0.05$ with FDR adjusted *P*-value and $\log_2 FC \geq 1$ or ≤ -1).

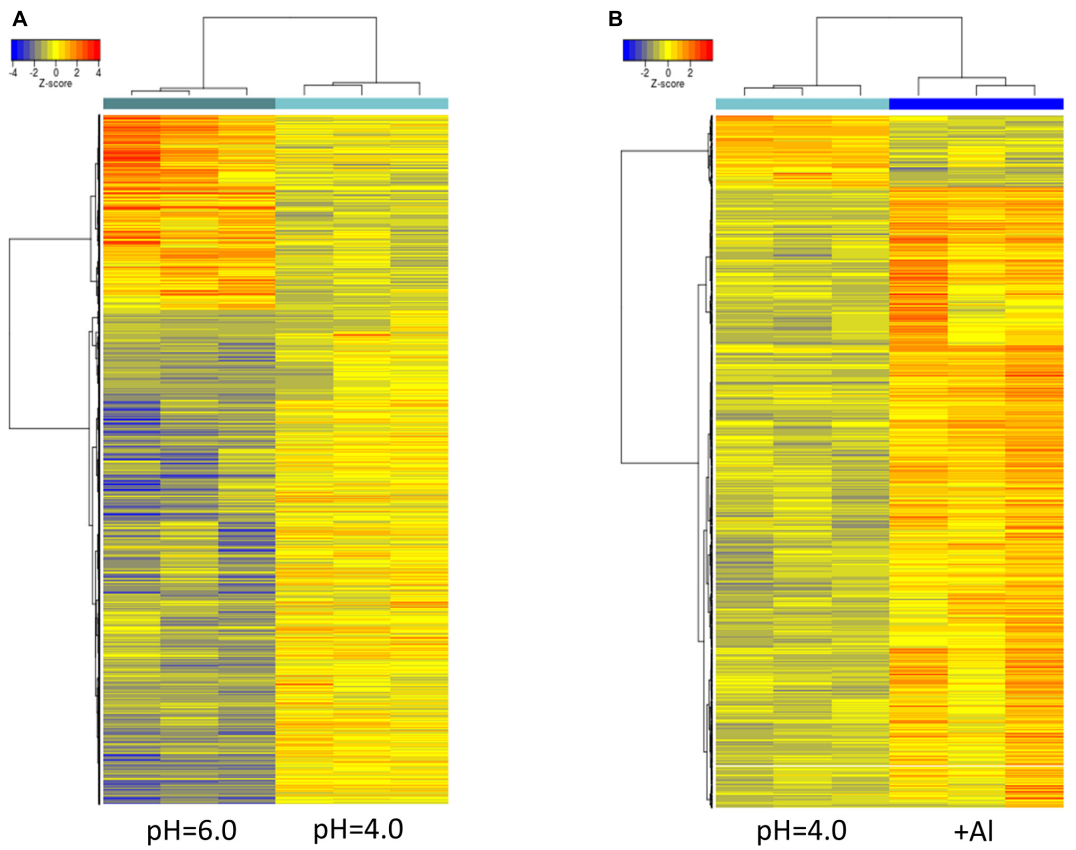

FIGURE 6 | Hierarchical clustering of samples using distance based on Pearson's correlation coefficient with Ward D2 linkage algorithm for a set of genes differentially expressed in barley root meristems in the short-term experiment. Gene counts were regularized log (rlog) transformed with library size-wise normalization, scaled and centered to be represented as *z*-scores in the log2 scale. The heat map represents the relative expression levels of DEGs. **(A)** The comparison between pH = 4.0 and pH = 6.0. **(B)** The comparison between Al-treatment at pH = 4.0 and pH = 4.0 without Al.

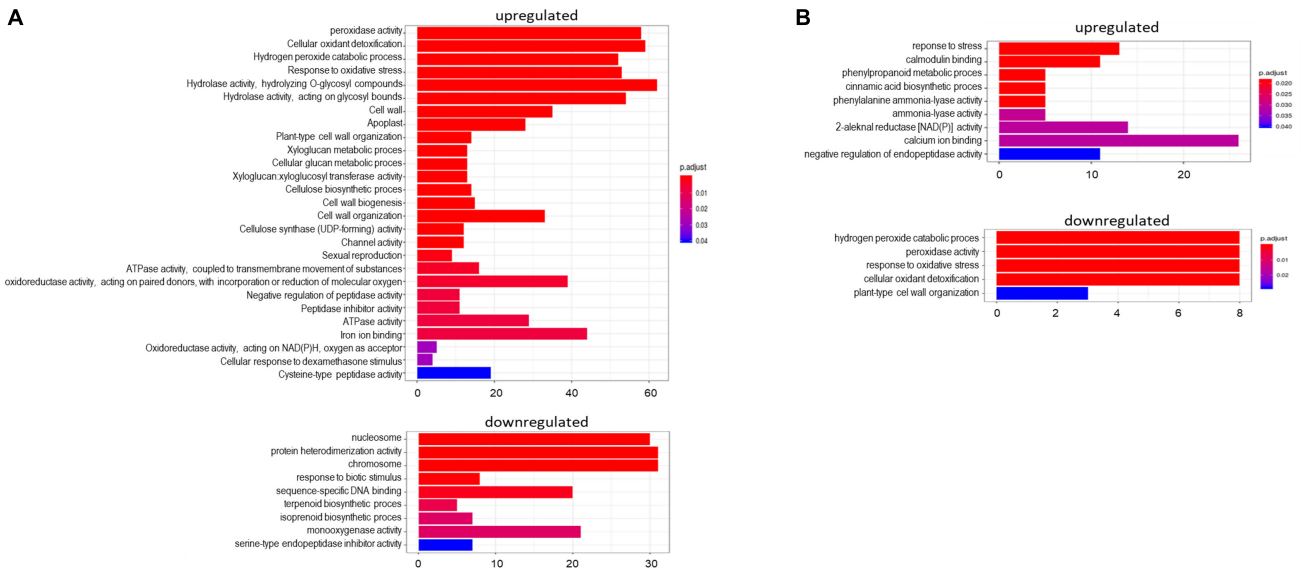

FIGURE 7 | GO term enrichment of genes upregulated and downregulated in the short-term experiment by low pH **(A)** and Al **(B)**. The numbers on the *X* axis indicates the number of genes; p.adjust means FDR adjusted *p*-value.

cell wall biogenesis, cellular glucan metabolic process, xyloglucan metabolic process, xyloglucan:xyloglucosyl transferase activity, cellulose biosynthetic process, or cellulose synthase (UDP-forming) activity. Within this group, the highest change in the gene expression showed xyloglucan endotransglucosylase/hydrolase (XTH) HORVU.MOREX.r2.4HG0348650 ($\log_2$FoldChange = 7.83). Fifteen genes encoding different further XTHs were identified as upregulated by low pH. These enzymes are known to cut and rejoin hemicellulose chains in the cell wall. More than 20 genes encoding expansins that are engaged in modifying the elasticity of the cell wall, and over a dozen genes encoding cellulose synthases showed increased expression pattern at low pH. Such a huge amount of DEGs related to cell wall organization evidently indicates that maintaining optimal pH is crucial for the proper development of this structure in barley roots.

Moreover, low pH influenced the signaling pathways in barley root meristems by alteration of expression of genes encoding protein kinases and transcription factors (TFs). TFs with changed expression belong mainly to WRKY (4 upregulated, 17 downregulated), MYB (7 upregulated, 6 downregulated), bHLH (7 upregulated), and NAC (5 upregulated, 2 downregulated) TFs families. Interestingly, in response to low pH, a group of genes related to chromatin organization was significantly downregulated, as for example genes encoding basic histones (H2A, H2B, and H4) or enzymes that posttranslationally modify histones, like histone-lysine *N*-methyltransferases.

The full lists of genes upregulated and downregulated by low pH (4.0) in the short-term experiment, with $\log_2$FoldChange ≥ 1 or ≤ -1, are provided as **Supplementary Materials 5, 6**.

Genes With Expression Altered by Al Treatment in the Short-Term Experiment

After 24 h of growth in hydroponics with 10 μM of bioavailable Al^{3+} ions at pH = 4.0, almost 90% of DEGs were upregulated and only 10% were downregulated compared to pH = 4.0 without Al (**Figures 5**, **6B**). Interestingly, the number of genes with expression affected by Al in the presented short-term experiment was twice lower than the number of genes with expression altered by low pH alone.

The GOs term enrichment identified the overrepresented groups of up- and downregulated genes after 24 h Al treatment (**Figure 7B**). Among them those related to the stress response were overrepresented in both, up- and downregulated groups. Out of DEGs encoding enzymes of peroxidase activity, 13 were upregulated and 7 were downregulated by 24 h of Al treatment. Four genes for glutathione S-transferases, the detoxifying enzymes, were highly upregulated. Thirteen and seven genes for cytochrome P450 were up- and downregulated, respectively. As was indicated earlier, low pH alone is already a stress factor to barley roots and this data suggest that 24 h of Al treatment additionally increases the stress.

The next overrepresented groups of upregulated genes were related to calcium homeostasis. It is well known that Al disturbs homeostasis of Ca^{2+} ions. Here, genes related to Ca^{2+} ion binding and calmodulin binding were upregulated. Calmodulin (calcium-modulated protein), activated by Ca^{2+}, modifies downstream proteins such as kinases and phosphatases in the calcium signal transduction pathway.

The expression of many transcription factors was also altered (mainly upregulated) by 24 h Al treatment. The most abundant were WRKY (12 upregulated), NAC (10 upregulated), and

MYB (7 upregulated) TFs. Therefore they are assumed to play important roles in regulating the expression of downstream genes involved in Al response.

The full lists of genes upregulated and downregulated by Al in the short-term experiment, with $\log_2$FoldChange ≥ 1 or ≤ -1, are provided as **Supplementary Materials 7, 8**.

Global Transcriptome Analysis of Root Meristems in the Long-Term Experiment

In the long-term experiment, the material for RNA isolation was collected 7 days after Al addition to the hydroponic solution. The seedlings were grown under conditions of optimal pH (6.0), low pH (4.0) without Al, and low pH (4.0) with 10 μM of bioavailable Al. Contrary to the results obtained for the short-term experiment, in the long-term experiment more genes were differentially expressed in root meristems of barley plants exposed to Al^{3+} ions than in plants stressed with low pH alone (**Figure 8**). In total, the expression of 870 genes was altered by low pH. Among them, 720 were upregulated and 150 were downregulated. Seven day treatment with 10 μM Al at pH = 4.0 led to a change of expression of a huge number of 5873 genes, of which 4116 were upregulated, whereas 1757 were downregulated. These numbers indicate that barley plants seem to adapt to low pH, at least at the transcriptome level, while the prolonged exposure to Al causes massive changes of transcriptome profile.

Genes With Expression Altered by Low pH in the Long-Term Experiment

At the 7 days time point of hydroponics at pH = 4.0, 82% of DEGs were upregulated and 18% were downregulated in relation to pH = 6.0 (**Figures 8**, **9A**). The GOs term enrichment allowed identification of overrepresented groups of up- and downregulated genes in plants exposed to low pH (**Figure 10A**). The results show that among upregulated genes were those related to transporter activity, such as e.g., HORVU.MOREX.r2.3HG0242890 gene that encodes a copper transporter whose expression was highly induced by low pH ($\log_2$FC = 7.2). Few other genes related to copper ion maintenance were also upregulated in these conditions ($CuSO_4$ is one of the components of Magnavaca solution). Another overrepresented group of upregulated genes was related to transferase activity and inhibitory regulation of peptidase activity.

The expression of genes encoding various transcription factors was also changed by low pH in the long-term experiment, however, their number was not as high as in the short-term experiment. They encoded TFs derived from the same TF families with most abundant those belonging to NAC family – eight upregulated TFs. The lower number of TFs with altered expression translated to the lower, than in the short-term experiment, number of total DEGs after exposure to pH = 4.0.

It is worth highlighting that in the short-term experiment many genes related to oxidative stress response were highly upregulated by low pH, whereas in the long-term experiment, these groups of genes were not overrepresented. For example, in total there were only 11 genes encoding enzymes of peroxidase activity upregulated after long-term exposure to pH = 4.0, in comparison to over 60 peroxidase genes upregulated by low pH in the short-term experiment. This also applies to genes encoding cytochrome P450, with over 30 of them upregulated after 48 h of hydroponics at pH = 4.0, whereas after long exposure to low pH this number dropped to 6. These findings suggest that at low pH plants are exposed to a huge oxidative stress and they need time to adapt to it.

The full lists of genes upregulated and downregulated by low pH in the long-term experiment, with $\log_2$FoldChange ≥ 1 or ≤ -1, are provided as **Supplementary Materials 9, 10**.

Genes With Expression Altered by Al in the Long-Term Experiment

The extremely high number of genes had altered expression after 7 days of growth in hydroponics with 10 μM of bioavailable Al^{3+} ions at pH = 4.0. The majority (70%) of DEGs were upregulated and 30% were downregulated in relation to pH = 4.0 without Al (**Figures 8**, **9B**). In the long-term experiment, the expression of significantly more genes was affected by Al than by low pH itself.

The GOs term enrichment allowed identification of overrepresented groups of genes up- and downregulated after long-term treatment with Al (**Figure 10B**). Based on

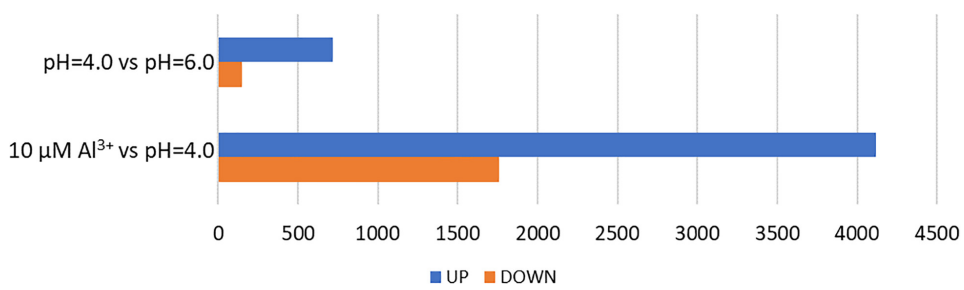

FIGURE 8 | The number of differentially expressed genes (DEGs) in the transcriptome of barley root meristems in the long-term experiment. DEGs were determined for plants grown at pH = 4.0 compared to pH = 6.0; and after treatment with 10 μM Al at pH = 4.0 compared to pH = 4.0 without Al^{3+} ions (DEGs were identified under $\alpha = 0.05$ with FDR adjusted *P*-value and $\log_2$FC ≥ 1 or ≤ -1).

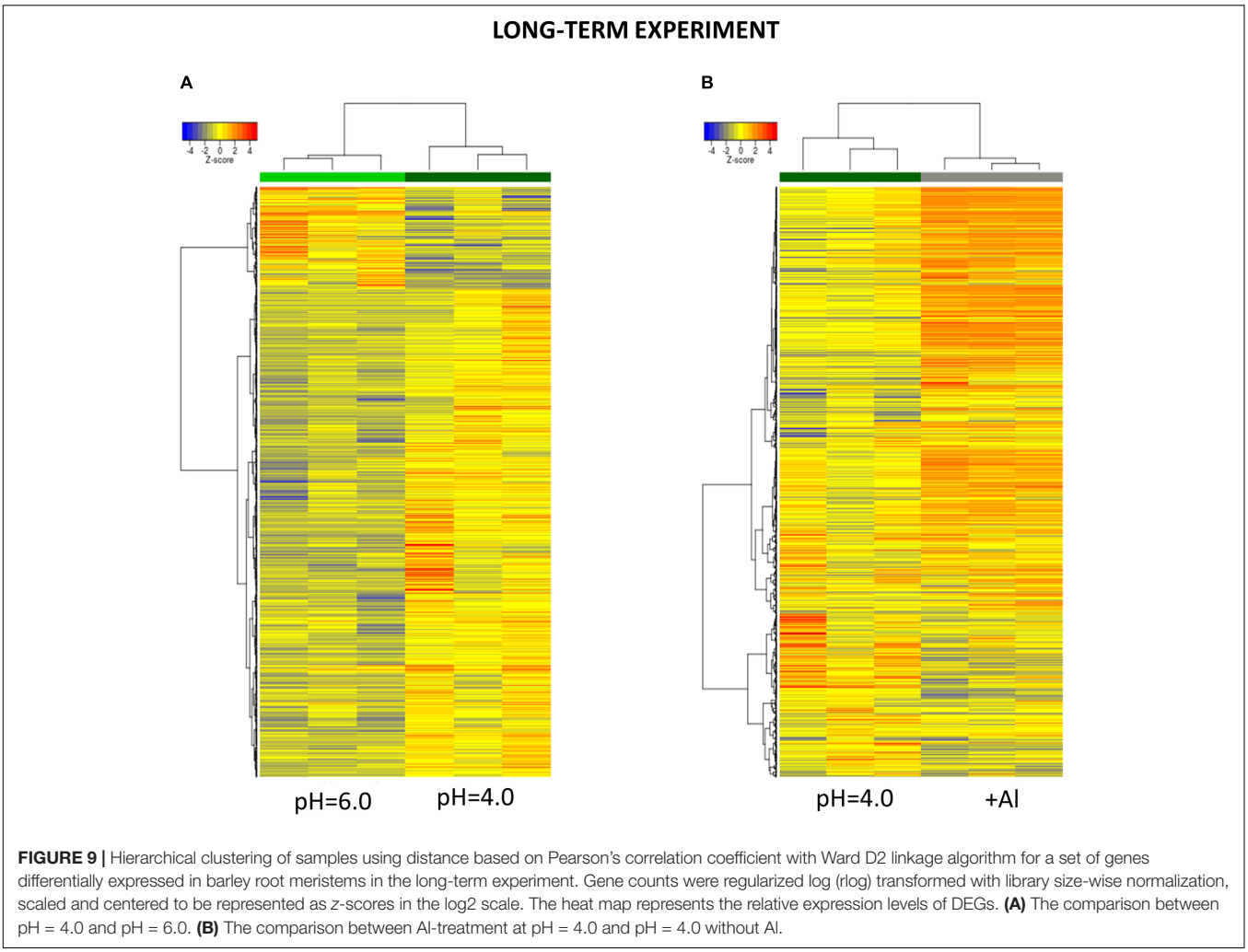

FIGURE 9 | Hierarchical clustering of samples using distance based on Pearson's correlation coefficient with Ward D2 linkage algorithm for a set of genes differentially expressed in barley root meristems in the long-term experiment. Gene counts were regularized log (rlog) transformed with library size-wise normalization, scaled and centered to be represented as *z*-scores in the log2 scale. The heat map represents the relative expression levels of DEGs. **(A)** The comparison between pH = 4.0 and pH = 6.0. **(B)** The comparison between Al-treatment at pH = 4.0 and pH = 4.0 without Al.

GO term enrichment, Al seemingly causes strong oxidative stress to barley roots. Hundreds of genes involved in oxidation processes were up- and downregulated. Among them were genes with ontologies defined as e.g., monooxygenase activity, dioxygenase activity, oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, peroxidase activity, cellular oxidant detoxification, or response to oxidative stress. Out of genes encoding peroxidases, 37 were upregulated, whereas 39 were downregulated. 72 genes encoding different proteins belonging to cytochrome P450 were upregulated and 16 were downregulated. The most upregulated gene HORVU.MOREX.r2.1HG0002460 ($\log_2$FC = 9.8) encodes a glutathione S-transferase and the most downregulated gene HORVU.MOREX.r2.2HG0099410 ($\log_2$FC = –9.4) encodes a peroxidase. Taken together, it shows that oxidative balance was disturbed in barley root meristem cells after prolonged aluminum treatment in hydroponics.

One of the overrepresented groups of genes that were upregulated in the long-term experiment was related to metal ion transport (GO:0030001). Within this GO term there are genes that may be potentially involved in the transport of any metal ion with an electric charge (therefore potentially also Al^{3+} ions) within a cell or between the cells. Three metal transporters from NRAMP (Natural resistance-associated macrophage protein) family were identified. Other metal transporters identified as differentially expressed after Al treatment were potassium, zinc, copper, magnesium, or calcium transporters. The elevated expression of genes involved in calmodulin binding further confirmed the disturbance in calcium homeostasis. Additionally, the two largest groups of upregulated transporters were heavy metal transport/detoxification superfamily (>30 upregulated, the highest $\log_2$FC = 5.09) and ABC transporter family proteins (>30 upregulated, the highest $\log_2$FC = 7.41). They might be potentially involved in Al ion transport. The downregulation of 12 genes encoding other proteins belonging to the heavy metal transport/detoxification superfamily further indicates that metal homeostasis was disturbed by exposure of roots to Al.

Upon prolonged Al treatment, genes encoding malate and citrate synthases (HORVU.MOREX.r2.2HG0146360 and HORVU.MOREX.r2.7HG0610760) were highly upregulated (with $\log_2$FC = 4,2 and $\log_2$FC = 5.5, respectively), which suggests that barley produces organic acids (OAs) in

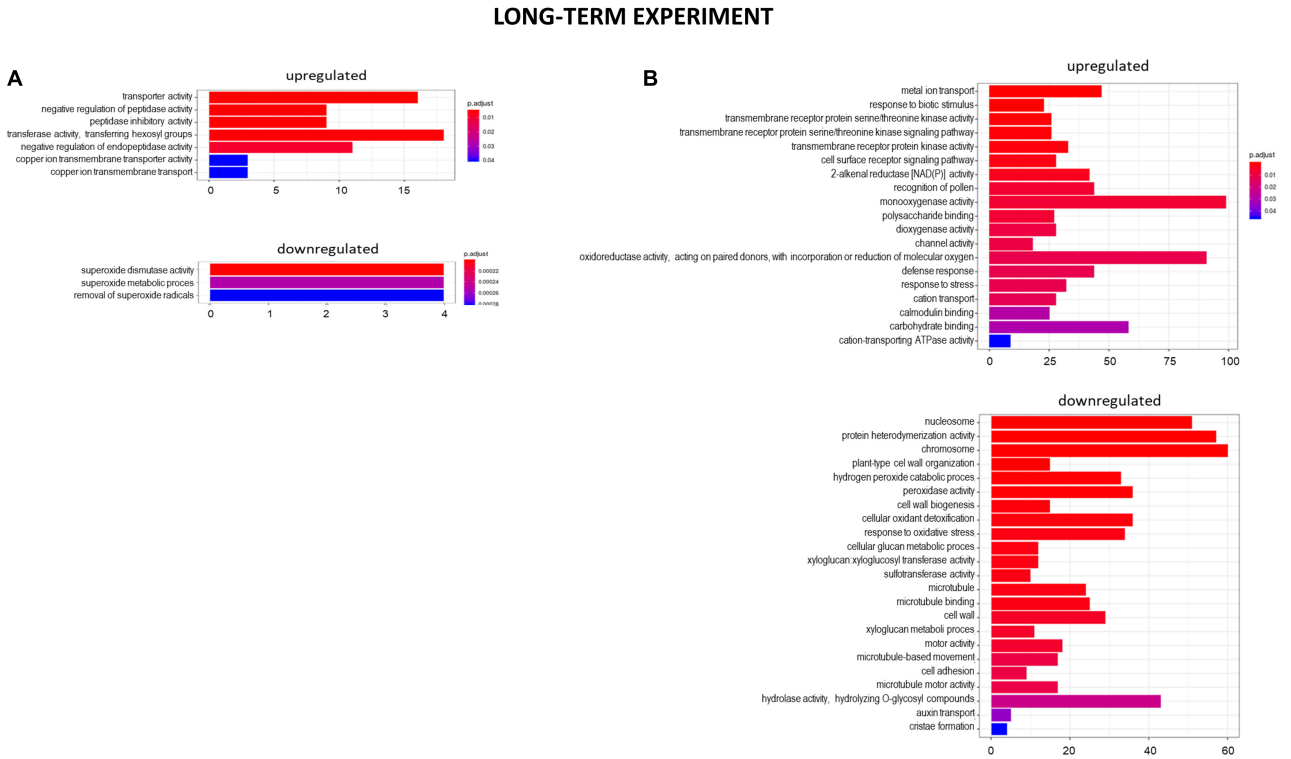

FIGURE 10 | GO term enrichment of genes upregulated and downregulated in the long-term experiment by low pH **(A)** and Al **(B)**. The numbers on the *X* axis indicates the number of genes; p.adjust means FDR adjusted *p*-value.

response to Al, probably to chelate Al ions in the process of detoxification. However, the gene encoding aluminum activated citrate transporter, which is a membrane protein involved in the exudation of citrate outside the root cells, was not upregulated, and aluminum activated malate transporter was even downregulated ($\log_2$FC = −1.83).

Among genes downregulated by Al were those related to chromosome organization, e.g., genes encoding basic histones (H2A, H2B, H3, and H4) and enzymes that modify histones, e.g., histone deacetylase 2 or histone *N*-methyltransferases. The other overrepresented group of downregulated genes was related to the cell wall development, which is consistent with the assumption that Al inhibits cell wall growth. What is more, many genes that were downregulated upon Al treatment for 7 days were involved in microtubule binding, movement, and activity. Al binds to the cytoskeleton and disrupts spatial orientation of the cytoskeleton, which disturbs cell expansion and is consistent with our observation that genes involved in microtubule organization are also Al-responsive. Taken together, the downregulation of the mentioned genes clearly indicates that the growth of cells in root meristem is disturbed and slowed down.

Genes encoding kinases were highly overrepresented within upregulated genes, which indicates activation of signaling pathways. The expression of many various transcription factors was also strongly altered by 7 days Al treatment which further resulted in the extremely high number of DEGs. Significantly more TFs were upregulated than downregulated and among them were those belonging to e.g., MYB (41 upregulated, 8 downregulated), WRKY (35 upregulated, 2 downregulated), NAC (26 upregulated, 1 downregulated), and bZIP (12 upregulated, 3 downregulated) TF families.

The full lists of genes upregulated and downregulated by Al treatment in the long-term experiment, with $\log_2$FoldChange ≥ 1 or ≤ -1, are provided as **Supplementary Materials 11, 12**.

Common Genes With Expression Altered by Low pH and Al Treatment

The comparison of DEGs between transcriptomes of low pH- and Al-treated barley root meristems showed that only a small group of DEGs was shared and the expression of much more genes was changed specifically by each treatment (**Figure 11**). It indicates the activation of distinct molecular mechanisms in response to these stresses. Moreover, the comparative analysis of GO terms enrichment further indicated that low pH and Al stress altered the expression of different groups of genes with diverse molecular functions, both in the short- and long-term experiments (**Figures 12**, **13**).

However, in the short-term experiment, there were 153 DEGs common for low pH and Al treatment (146 upregulated and 7 downregulated). Among them were some genes related to oxidative stress response (e.g., encoding peroxidases or alpha-dioxygenase 2) and cell wall development (e.g., encoding pectate lyase, pectinesterase, xyloglucan endotransglucosylase/hydrolase

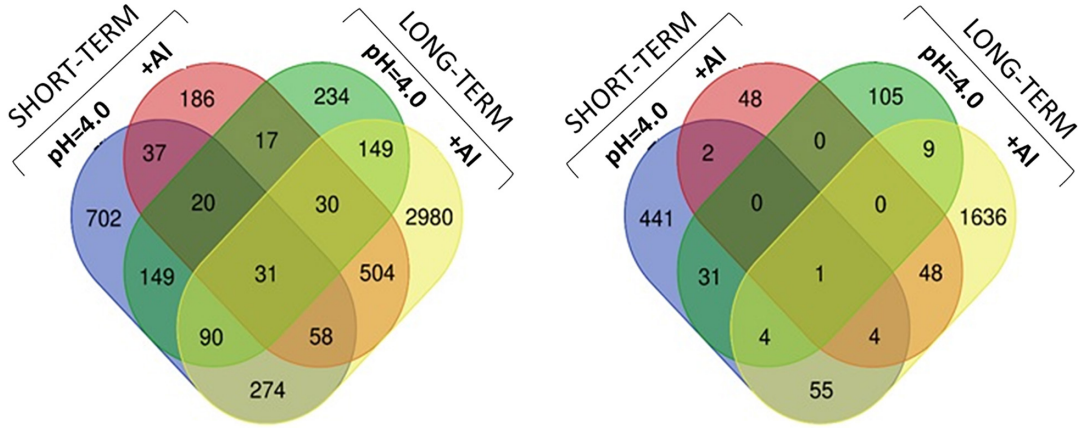

FIGURE 11 | Venn diagrams showing common and specific DEGs upregulated and downregulated between all experimental combinations.

and expansin). Additionally, several transcription factors were also upregulated by both analyzed stresses (low pH and Al), which indicates that some common mechanisms of response might be activated. The lists of common genes with expression altered by low pH and Al in the short-term experiment is provided together with their annotations as **Supplementary Material 13**. Additionally, to illustrate the prevalent expression patterns of DEGs in the short-term experiment, we performed the analysis of gene expression profiles using k-means clustering. Four clusters of genes with prevailing expression patterns have been identified (**Figure 14A**). DEGs common for both analyzed factors, with expression upregulated by low pH and further upregulated by Al are grouped within the Cluster 3. The GO term enrichment of these group of genes also showed that they are mainly related to oxidative stress and cell wall development (**Supplementary Material 14**). The GO term enrichment of DEGs from remaining clusters is also provided in **Supplementary Material 14**.

In the long-term experiment, there were 314 DEGs common for low pH and Al treatment (300 upregulated and 14 downregulated). Similarly as in the short-term experiment, among common genes with expression altered after long low pH and Al treatments there were DEGs related to oxidative stress (e.g., encoding peroxidases or cytochrome P450 family proteins) and cell wall development (e.g., encoding xyloglucan endotransglucosylases/hydrolases or aldehyde dehydrogenase). The list of common genes with expression altered by low pH and Al in the long-term experiment is provided together with their annotations as **Supplementary Material 15**. This data indicate the existence of common responses to low pH and Al, nevertheless, the majority of DEGs showed expression changes specifically in response to one of these factor, with the highest number of genes with expression affected after long Al exposure. This is also illustrated by k-means clustering which showed four clusters of genes with prevalent expression patterns in the long-term experiment (**Figure 14B**). The presented heatmap shows that in most clusters (except for a small Cluster 2) aluminum altered the gene expression to the greatest extent. DEGs common for low pH and Al with expression upregulated by both analyzed factors independently are grouped within the Cluster 3. The overrepresented GO terms in Cluster 3 were e.g., monooxygenase activity, iron ion binding, oxidoreductase activity, transmembrane transporter activity, calcium ion binding and metal ion transport. The GO term enrichment of DEGs from all clusters is provided as **Supplementary Material 16**.

DISCUSSION

Barley (*Hordeum vulgare* L.) is the most Al-sensitive species among small grain cereals, but still there are differences in Al tolerance among barley cultivars, which are mostly correlated with the ability of the genotype to secrete citrate (Zhao et al., 2003; Furukawa et al., 2007). Cv. 'Sebastian' used in our study is relatively tolerant to Al when compared to other barley cultivars (Vega et al., 2019). Nonetheless, even the micromolar concentration of bioavailable Al^{3+} ions (10 μM Al^{3+}) applied in hydroponic solution at pH = 4.0 for 7 days, extremely reduced (by 83%) the total length of 'Sebastian' roots, compared to root length of untreated plants grown at optimal pH = 6.0. However, without a doubt, the reduction of root growth was not caused by Al^{3+} ions only, but also by the low pH and proton/H^+ toxicity, as growing plants at pH = 4.0 without addition of Al reduced the total root length of 'Sebastian' seedlings almost by half. It has also been previously reported that barley is very sensitive to H^+ toxicity (Zhao et al., 2003; Guo et al., 2004). Similarly, higher H^+ activity, significantly decreased the root length of rice seedlings grown at pH = 3.5 and pH = 4.5 compared to pH = 5.5 (Zhang et al., 2015). However, even though it was reported that proton rhizotoxicity can be more harmful than Al rhizotoxicity in natural

FIGURE 12 | The comparison of GO terms over-representation analysis of differentially expressed genes in the short-term experiment using hypergeometric test under $\alpha = 0.05$ with FDR adjusted *P*-value.

acid soils (Kinraide, 2008), the effect of the low pH itself has been understudied in the aspect of Al toxicity. It has to be stressed that in natural conditions, in acidic arable lands, Al toxicity and H^+ toxicity coexist and together negatively affect barley performance and yield. To discriminate Al effect from a low pH effect in our RNA-Seq analysis, we compared transcriptomes of Al treated root meristems to those grown without Al at pH = 4.0. Additionally, we also compared the transcriptomes of barley plants grown at pH = 4.0 to those grown at pH = 6.0 (both without Al).

Our results show that the low pH caused global changes in barley transcriptome profile when seedlings were grown in hydroponics for 48 h (short-term experiment), whereas after a prolonged time of growth under low pH (further 7 days), the number of DEGs significantly decreased, suggesting that partial adaptation of plants to this stress occurred. Interestingly, the opposite effect was seen in regards to aluminum toxicity. After 24 h of Al treatment, many genes were up- and downregulated in root meristems, however, their number increased extremely after 7 days of Al treatment, which suggests that remodeling of the transcriptome following Al stress is a long-lasting and dynamic process. These results are in agreement with the microarray analysis of *Arabidopsis thaliana* transcriptome profiles in response to Al stress, where more transcripts were Al-responsive after 48 h than 6 h treatment (Kumari et al., 2008).

Low pH and Al as Oxidative Stressors

Different abiotic stresses, such as drought, cold, salt, and heat, can disrupt the balance of ROS content and lead to their accumulation in the cell, which results in oxidative stress (reviewed in You and Chan, 2015). It has long been known that aluminum also induces oxidative stress in plants. The

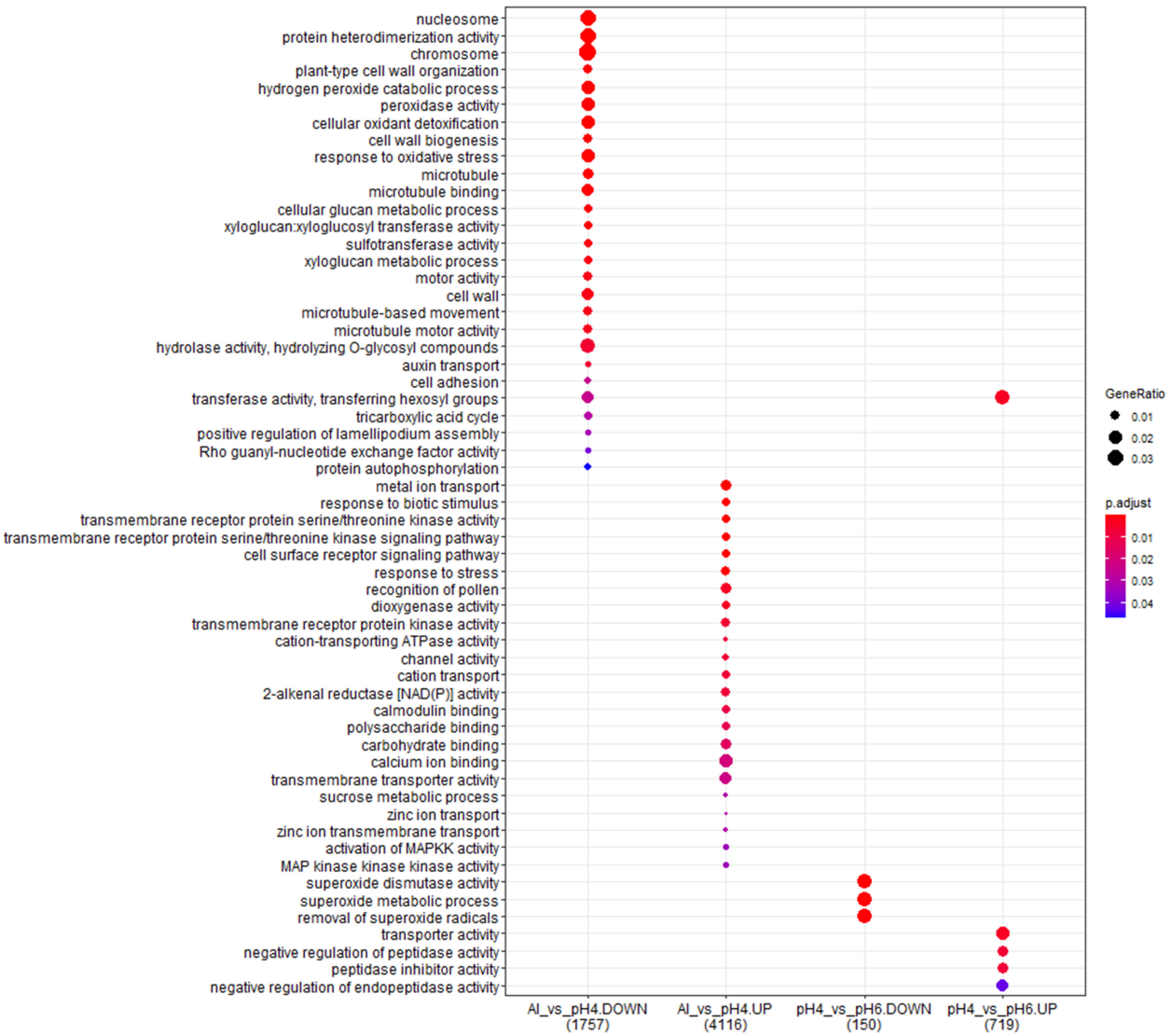

FIGURE 13 | The comparison of GO terms over-representation analysis of differentially expressed genes in the long-term experiment using hypergeometric test under $\alpha = 0.05$ with FDR adjusted *P*-value.

first genes related to oxidative stress, which were identified as being upregulated by Al in *Arabidopsis thaliana*, encoded peroxidase, glutathione-S-transferase, and protein homologous to the reticuline:oxygen oxidoreductase enzyme (Richards et al., 1998). Al was found to influence reactive oxygen intermediates, lipid peroxidation, protein oxidation, and activities of antioxidant enzymes in many different plant species, including *Allium cepa* (Achary et al., 2008), *Triticum aestivum* (Xu et al., 2012; Sun et al., 2017), and *Zea mays* (Boscolo et al., 2003; Giannakoula et al., 2010). In the presented study, we show that both analyzed factors, low pH and Al, led to the alteration of oxidative stress genes expression in barley roots.

In our 'low pH only' study, the number of DEGs related to oxidative stress response was very high after 48 h of growth at pH = 4.0, but in the long-term experiment, the number of DEGs significantly decreased, suggesting that barley plants adapt to oxidative stress caused by low pH (H^+ toxicity) over time. The study performed in rice (*Oryza sativa*) has shown that growing plants for 2 weeks at pH lowered to 3.5 led to the serious lipid peroxidation and increases of H_2O_2 and MDA (malondialdehyde) content in rice roots. At the transcriptomic level, it led to downregulation of copper/zinc superoxide dismutases (*Cu/Zn SOD1*, *Cu/Zn SOD2*) and catalases (*CATA* and *CATB*), and upregulation of ascorbate peroxidase 1 (*APX1*). Correspondingly, the activity of these enzymes was also altered. It was assumed that higher activity of APX can contribute to adaptation of rice to low pH (Zhang et al., 2015). In our study, the expression of genes encoding SODs was not altered after 48 h of growth at low pH, but the extension of hydroponic culture to 7 days caused a significant decrease in expression level of four

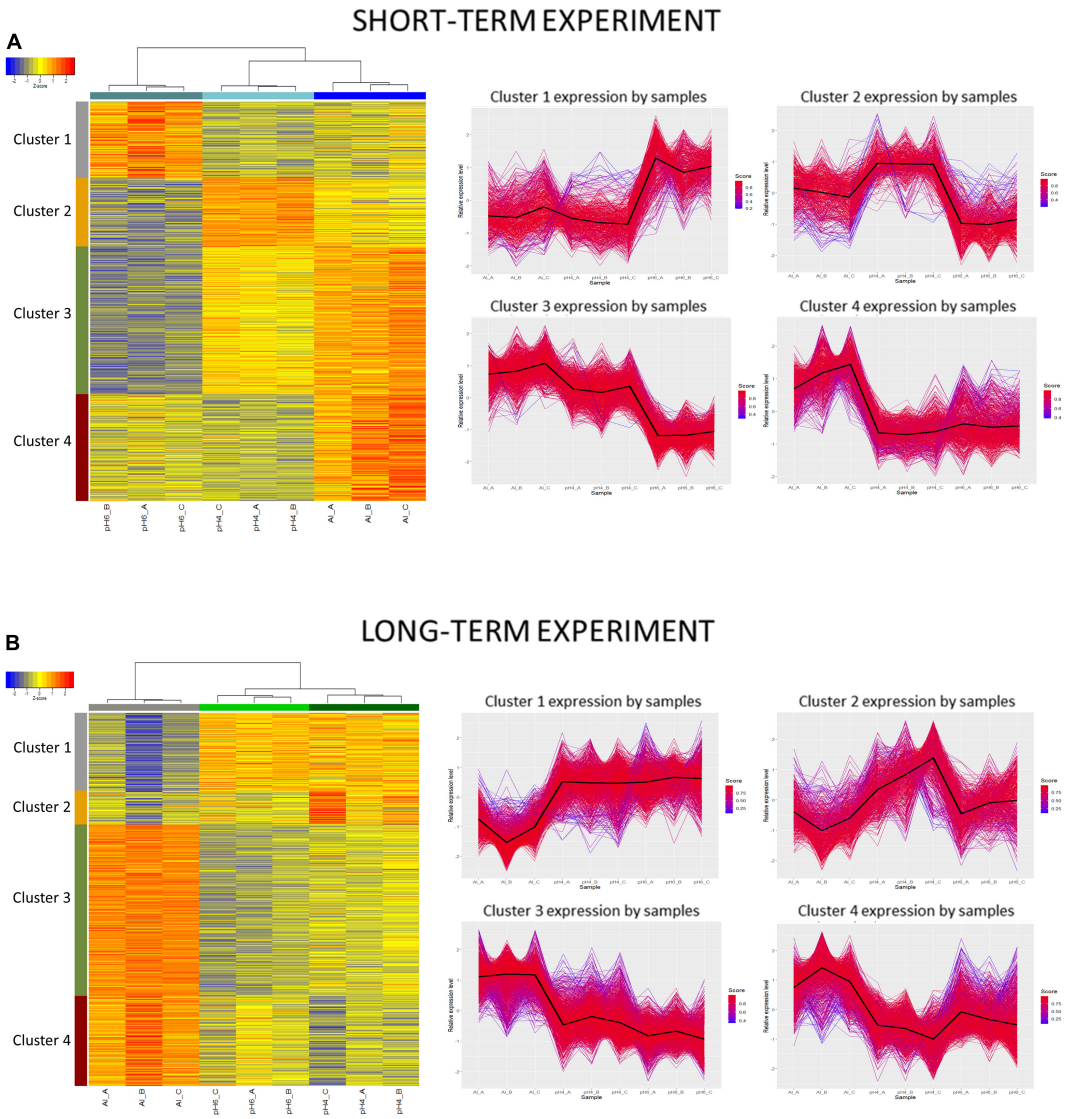

FIGURE 14 | Clustering analysis using hierarchical clustering of samples and k-means clustering of genes. Clustering was performed using a selected set of genes differentially expressed in barley root meristems grown at pH = 4.0 compared to pH = 6.0 and treated with 10 μM of bioavailable Al^{3+} compared to pH = 4.0. Gene counts were regularized log (rlog) transformed with library size-wise normalization, scaled and centered to be represented as *z*-scores in the log2 scale. Hierarchical clustering of samples was performed using distance based on Pearson's correlation coefficient with Ward D2 linkage algorithm. Heat-map represent their expressional patterns. Color bars on the left are corresponding to each consequent cluster identified: from 1 on the top to the 4 at the bottom. For each gene cluster detected with k-means clustering, a plot of relative gene expression profile is shown on the right with black lines indicating each cluster centroids. Scores represent correlations of each gene with the cluster core. **(A)** Short-term experiment; **(B)** long-term experiment.

of Cu/Zn SODs, similarly to rice. However, the genes encoding ascorbate peroxidase or CATA were not found among DEGs and the barley ortholog of *CATB* was even highly upregulated ($log_2FC = 4.56$), which may be related to the higher sensitivity of barley to low pH compared to rice.

In the case of Al treatment, a high number of genes related to oxidative stress were differentially expressed in barley roots, especially in the long-term experiment, where hundreds of these genes were highly up- and downregulated. Among them there were genes encoding peroxidases (PODs), superoxide dismutases (SODs), cytochrome P450 monooxygenases, glutathione S-transferases (GSTs), thioredoxins (TRX), and others. Studies performed on other species, e.g., Arabidopsis, cucumber (*Cucumis sativus*), rice, wheat (*Triticum aestivum*) and citrus (*Citrus sinensis* and *Citrus grandis*) also showed that Al induces strong oxidative stress and upregulates the activity of antioxidative mechanisms (Kumari et al., 2008; Pereira et al., 2010; Ma et al., 2012; Guo et al., 2017; Liu et al., 2018; Awasthi et al., 2019). In general, when different genotypes were compared after Al treatment, the Al-tolerant lines were characterized by a higher activity of the antioxidative system than the Al-sensitive ones. However, RNA-seq analysis in maize showed that the total

number of genes related to oxidative stress upregulated by Al treatment was higher in the Al-sensitive than the Al-tolerant genotype, suggesting that upregulation of these genes was merely a consequence of Al toxicity, not the activation of Al tolerance mechanisms (Maron et al., 2008). Such a huge number of DEGs from this category in our experiment emphasizes the very high level of Al-sensitivity of barley compared to other species, even though 'Sebastian' cultivar belongs to the Al-tolerant group among barley cultivars.

Peroxidases are known to be enzymatic antioxidants, hence massive upregulation in their expression means that the plant is under oxidative stress. In the presented study, 61 genes encoding peroxidases were upregulated after 48 h growth at low pH, but after further 7 days of low pH hydroponics, this number dropped to 11 DEGs. The same period (7 days) of Al treatment caused alteration of expression of many more genes encoding different peroxidases, which were both up- (37 POD genes) and downregulated (39 POD genes). It should be noted that PODs have more diverse functions, e.g., they are involved in cross-linking of the cell wall constituents (Bakalovic et al., 2006). The oxidative cross-linking of the cell wall components managed by some classes of PODs may increase cell wall stiffening and decrease its extensibility which is associated with inhibition of root growth by Al (Ma et al., 2012). In the presented experiments, the 7-day treatment of barley seedlings with Al^{3+} ions caused a significant reduction of root growth accompanied by the increase of root diameter. Moreover, the activity of some PODs that leads to H_2O_2 formation may be a potential mechanism of Al tolerance, because production of H_2O_2 may be used to restructure the cell wall and block Al entry by decreasing cell wall porosity (Maron et al., 2008). What is more, Tamás et al. (2005) presented that the production of H_2O_2 mediated by PODs in response to Al led to cell death of barley root border cells and hence protected root tips by chelating Al in the dead cells. The cell wall-bound PODs were also found to be involved in lignin biosynthesis, which is known to be one of the symptoms of Al stress (Li et al., 2003). The contrasting pattern of increased and decreased expression of different peroxidase genes in response to Al treatment has been observed also in other transcriptomic studies (Kumari et al., 2008; Maron et al., 2008; Li et al., 2017). It shows the complex and diverse roles of peroxidases in Al stress response.

Another example of enzymes involved in oxidative stress response is thioredoxins (TRX) - thiol-oxidoreductases that function in maintaining redox homeostasis (Meyer et al., 2012). They are induced by a variety of oxidative stimuli and their overexpression protects the cell from cytotoxicity caused by oxidative stress (Nishinaka et al., 2001). Lately, the *AtTRX1* (*Thioredoxin H-type 1*) gene was identified in GWAS studies in Arabidopsis and confirmed by reverse-genetics and co-expression gene network analysis as associated with Al-tolerance (Nakano et al., 2020). However, to the best of our knowledge, to date there was no report about TRX involvement in Al tolerance or response in monocots. Interestingly, in our study the expression of *THR* genes was altered mainly by Al (two genes upregulated in the short-term and 11 up- and 3 down-regulated in the long-term experiment), whereas low pH affected the expression of only two of *THR* genes in both, the short- and the long-term treatment. The Arabidopsis knock-out mutant in the *TRX1* gene was hypersensitive to Al, but not to proton (low pH) toxicity (Nakano et al., 2020). These data together suggest that thioredoxins are involved in the protection of cells from Al-induced oxidative stress rather than from the proton-induced one.

Cell Wall Related Genes Regulated by Low pH and Al

The cell wall is suggested to be a primary target of Al toxicity and the majority of Al absorbed by the root tissue is localized in the apoplast. Aluminum binds to the negatively charged carboxylic groups of pectins and changes the cell wall properties, which may cause inhibition of the root cell elongation and growth (Kochian, 1995; Zheng and Yang, 2005; Silva, 2012). Our RNA-Seq data show that the expression of several genes encoding enzymes that directly modify pectins (with GO:0042545 – cell wall modification) was affected, mainly by the prolonged Al treatment. However low pH itself also changed the expression profile of some genes from this group, although to a lesser extent, which suggests that Al stress has a larger impact on modifying pectins in barley root meristematic cells than low pH alone. It is in line with our previous study showing that aluminum changes the pectin cell wall composition in barley root cells (Jaskowiak et al., 2019). Barley plants exposed to a long Al exposure showed the changes in content and localization of the pectic epitopes involved in maintenance of cell wall flexibility, stiffening of the wall and firmness of the cells. In the presented study, among DEGs related to pectin modification, those encoding pectinesterases and pectin lyases were the most abundant, especially after long-term Al treatment. Pectinesterases (also known as pectin methylesterases) belong to a large family of isozymes that catalyze the de-esterification of pectins. In our study, the expression of several genes encoding pectinesterases was upregulated by both, low pH and Al. It is consistent with our previous studies where we show, by analyzing LM19 and LM20 antibodies, that unesterified homogalacturonans (HGs) were more abundant in the Al-treated roots compared to the not treated ones (Jaskowiak et al., 2019). Similar results were obtained previously for maize, which additionally supports the hypothesis that the difference in Al tolerance among maize genotypes may depend on the level of methyl-esterification of pectins (Eticha et al., 2005).

It has been reported that Al stress enhances the incorporation of lignin into the cell wall in roots of many plant species, including wheat and rice (Hossain et al., 2005; Sasaki et al., 2006; Wang and Kao, 2007). The deposition of lignin provides the rigidity and mechanical resistance of the plant cell wall by creating a barrier that limits the radial movement of metals and pathogens (Gavnholt and Larsen, 2002). Phenylalanine ammonia-lyase (PAL) is an enzyme involved in the biosynthesis of lignin. In our study, genes encoding PALs were upregulated specifically after Al treatment (five and seven genes upregulated in the short- and long-term experiment, respectively). As indicated earlier, some DEGs encoding peroxidases with

expression altered in the presented RNA-seq data may also be related to lignin biosynthesis. Another group of enzymes that may be involved in lignin deposition are laccases, because of their localization in lignifying cell walls and their potential to oxidize lignin precursors (Gavnholt and Larsen, 2002). In our study, the genes encoding laccases were upregulated by both, low pH and Al, with the highest response after 7 days Al treatment. They were also found to be up-regulated by Al in maize (Maron et al., 2008). These data indicate that lignin deposition plays a role in plant response to Al toxicity as a potential cause of root growth inhibition. It can possibly play a role in Al tolerance by blocking the entrance of Al inside the root tissue.

Transcription Factors Modulated by Low pH and Al

Various TFs were overrepresented among genes with expression changed by low pH and Al. They belong mainly to WRKY, MYB, and NAC families and they all may play complementary roles in regulating the expression patterns of low pH- and Al-responsive genes. The differences in TFs expression profiles between Al and low pH treatments indicate that various responses may be activated upon these two stresses.

It is well documented that one transcription factor is of special importance in both, low pH and Al tolerance in Arabidopsis – the C_2H_2 zinc-finger protein STOP1 (Sensitive To Proton rhizotoxicity 1). STOP1 regulates multiple genes protecting Arabidopsis from H^+ and Al toxicities and *stop1* mutants (T-DNA insertional, as well as missense) are H^+- and Al-hypersensitive. Their hypersensitivity is related to downregulation of *AtALMT1* (*Aluminum-Activated Malate Transporter1*), *AtALS3* (*Aluminum-Sensitive 3*) that encodes an ABC transporter possibly involved in redistribution of Al, and other genes involved in ion homeostasis and metabolic pathways regulating pH (Sawaki et al., 2009). OsART1 (Aluminum Resistance Transcription factor 1), the ortholog of AtSTOP1, activates multiple genes involved in Al tolerance in rice, including those implicated in external and internal Al detoxification, e.g., *STAR1* (*Sensitive to Al rhizotoxicity 1*) encoding ABC transporter. However, unlike STOP1 in Arabidopsis, OsART1 is involved specifically in Al response only, not in response to the stress caused by low pH (Yamaji et al., 2009). More intriguingly, the VuSTOP1 (ortholog found in rice bean, *Vigna umbellata*) is involved mainly in response to H^+ toxicity (Fan et al., 2015). Expression of *AtSTOP1* and *OsART1* turned out to be constitutive and not affected by proton or Al stress, hence these TFs are thought to be regulated posttranslationally. It was recently confirmed that in Arabidopsis AtSTOP1 function is regulated by SUMOylation (Fang et al., 2020). On the contrary, the expression of *VuSTOP1* is induced by both, H^+ and Al^{3+} (reviewed in Fan et al., 2016). What is more, in wheat three homoeologous *TaSTOP1* genes display differential expression patterns: *TaSTOP1-A* is induced by Al^{3+}, *TaSTOP1-B* is constitutively expressed and *TaSTOP1-C* is induced by H^+ (Garcia-Oliveira et al., 2013). We used *TaSTOP1* sequence to search for potential barley orthologs and we found one barley *STOP1* ortholog: HORVU.MOREX.r2.3HG0249360. It encodes a zinc finger protein with DNA-binding transcription factor activity. The expression pf *HvSTOP1* was not affected in the presented study neither by low pH nor by Al, however its GOs indicated that it is involved in both, low pH and Al response (GO:0010044 – response to aluminum ion, GO:0010447 – response to acidic pH). Thus, it may be assumed that barley *HvSTOP1* gene is regulated posttranslationally, similarly to the STOP1 in Arabidopsis.

Transporters Specific for Al Response

Other interesting groups among DEGs encode different types of transporters that were differentially expressed especially in long-term experiments. Three metal transporters, specifically upregulated only by Al, encode NRAMP proteins (Natural Resistance-Associated Macrophage Protein). One of them, HORVU.MOREX.r2.7HG0610240 (log_2FC = 1.35) is homologous to *ZmNrat1* (*nramp Aluminum Transporter 1*) that is known to be a membrane transporter of aluminum in maize (Guimaraes et al., 2014; Matonyei et al., 2020). Similar to the barley gene, *ZmNrat1* was also upregulated by Al treatment. It is suggested that NRAT1 membrane proteins are involved in the Al response mechanism by being responsible for the transport of Al from outside to inside the cell, which reduces Al concentration in the apoplast.

In our study genes encoding potassium, zinc, or copper transporters were found to be differentially expressed by both applied stresses. However, magnesium transporters were activated only by Al. The examples are genes: HORVU.MOREX.r2.3HG0249560 encoding magnesium transporter MRS2-like protein, which is an ortholog of *OsMGT1* (*Magnesium Transporter 1*), upregulated after both short and long Al treatment and HORVU.MOREX.r2.2HG0180770 encoding another putative magnesium transporter whose expression increased significantly only after long Al treatment. Similarly, which was demonstrated in rice, *OsMGT1* expression was rapidly upregulated by Al, but not by low pH and was found to be regulated by OsART1 (Chen et al., 2012). This transporter is responsible for Mg uptake in the roots and increasing internal Mg^{2+} concentration was demonstrated to be crucial for conferring Al tolerance (reviewed in Rengel et al., 2015).

ABC (ATP-Binding Cassette) transporters are a large family of ubiquitous transmembrane proteins responsible for the active transport of various ligands across membranes (reviewed in Linton, 2007). Some representatives of this group are confirmed to be involved in detoxifying Al. A great number of genes encoding ABC transporters were upregulated in our study by Al. Also many were upregulated by low pH, but only in the short-term treatment experiment. Two genes encoding ABC transporters that were found among Al specific DEGs were homologous to *OsALS1* (*Aluminum Sensitive1,* Os03g0755100), namely HORVU.MOREX.r2.5HG0424840 and HORVU.MOREX.r2.5HG0424850 (lof_2FC = 1.4 and 3.15, respectively). *OsALS1* encodes a tonoplast-localized ABC transporter and is regulated by OsART1. Its expression in rice was also specifically induced by Al, not by low pH, as it is responsible for sequestration of Al into the vacuole (Huang et al., 2012).

However, the expression of Arabidopsis ortholog, *AtALS1*, was not Al inducible (Larsen et al., 2007). Because of the increase of the expression of two barley *ALS1* orthologs in response to Al (similarly as in rice), it may be assumed that both of them are potentially involved in internal Al detoxification in barley.

OsSTAR1/STAR2 complex is another example of an ABC transporter that is responsible for Al detoxification in rice. *OsSTAR1* encodes an ATP-binding protein that forms a complex with a transmembrane protein OsSTAR2. The STAR1/STAR2 complex is responsible for the transport of UDP (uridine diphosphate)-glucose that can modify cell walls and therefore mask Al-binding sites. Both *OsSTAR1* and *OsSTAR2* are upregulated upon Al stress in rice (Huang et al., 2009). In Arabidopsis knock-out of *AtSTAR1* resulted in increased sensitivity to Al, however its expression was constitutive in roots and shoots and was not induced by Al (Huang et al., 2010). However, the expression of *AtALS3*, that is homologous to *OsSTAR2*, increased in roots following Al exposure (Larsen et al., 2005). The homologs of *OsSTAR1* and *OsSTAR2* were identified in the barley genome but only a homolog of *OsSTAR1*, HORVU.MOREX.r2.4HG0339800, was upregulated in our RNA-seq experiment after a prolonged Al treatment ($\log_2$FC = 1.35). These data suggest that the pathway of Al tolerance based on Al detoxification is not as efficient in barley as in rice, which is consistent with rice being a more highly Al-tolerant cereal.

OA Related Genes

So far, the best-documented mechanisms that help higher plants to cope with Al toxicity rely on organic acid (OA) exudation, which can chelate and thus neutralize Al^{3+} ions. OAs can act either in the rhizosphere when they are released outside the root tissue (Al exclusion mechanism), or inside the cell where they take part in Al detoxification (Al tolerance mechanism). Different plant species may secrete different OAs from roots, mainly citrate, malate, and/or oxalate anions (reviewed in Yang et al., 2013, 2019). In general, species or varieties that are tolerant to Al can secrete high levels of OAs when exposed to Al stress (e.g., Li et al., 2000; Yang et al., 2000; You et al., 2005; Dong et al., 2008). In barley, the most Al-sensitive among small grain cereals, the differential Al tolerance observed among different cultivars is correlated mainly with citrate secretion (Zhao et al., 2003). Correspondingly, in our study, the prolonged Al treatment caused a very high upregulation of a gene encoding citrate synthase (HORVU.MOREX.r2.7HG0610760 with $\log_2$FC = 5.5), but also an increased expression of a gene encoding malate synthase (HORVU.MOREX.r2.2HG0146360 with $\log_2$FC = 4.2). Interestingly, low pH alone also caused upregulation of citrate synthase (with $\log_2$FC = 6.0), but only in the short term experiment.

Organic acids produced by plants are exuded outside the root to the rhizosphere through membrane transporters. The first OA transporter, a malate transporter ALMT1 (Aluminum-activated Malate Transporter 1) was discovered in wheat. The *TaALMT1* gene is constitutively highly expressed in the Al-tolerant wheat cultivars and its expression is not upregulated by Al (Sasaki et al., 2004). In a tea plant that is highly tolerant to Al, four genes encoding ALMT homologs were found and contrary to wheat, all of them were upregulated by Al (Li et al., 2017). In our study the *HvALMT1* gene was not upregulated and was even slightly downregulated by Al. It is in line with the fact that in response to Al, barley plants release only citrate but not malate to the rhizosphere (Zhao et al., 2003). The increase in the expression level of malate synthase after Al treatment may indicate that malate is involved in internal detoxification of Al. Nevertheless, the overexpression of the *TaALMT1* gene increased the malate secretion and Al tolerance in transgenic barley (Delhaize et al., 2004).

The transmembrane transporters releasing citrate anions outside the cells were first identified in barley and sorghum (*Sorghum bicolor*) and named, respectively: HvAACT1 (Aluminum Activated Citrate Transporter 1) (Furukawa et al., 2007) and SbMATE1 (Magalhaes et al., 2007). Afterward they were identified in many other plant species, including wheat, maize, rye, rice, and rice bean (Ryan et al., 2009; Maron et al., 2010; Yokosho et al., 2010, 2011; Yang et al., 2011). These citrate transporters belong to the MATE (Multidrug And Toxic Compound Extrusion) family that is one of the largest plant transporter families. In the majority of plant species, genes encoding these transporters are upregulated by Al. Surprisingly, *HvAACT1* was not found to be upregulated by Al stress in barley (Furukawa et al., 2007). The barley cultivars that are relatively tolerant to Al were characterized by constitutive high expression of *HvAACT1*. In our study, we also did not find *HvAACT1* among DEGs in any experimental combination, which indicates that its expression is not altered by low pH or Al.

CONCLUSION

Here we show for the first time the global transcriptome analysis of root meristematic cells of barley *Hordeum vulgare* L. grown at low pH and treated with Al. We provide a full list of differentially expressed genes that may be useful for studying mechanisms of H^+ and Al toxicity in this important crop species. The obtained results provide new insights into the very complex mechanisms underlying H^+ and Al tolerance in barley, suggesting that there are several common, but many more specific genetic pathways launched in response to these stresses. The fact that many various mechanisms are activated indicates that the pyramiding of genes for H^+- and Al-tolerance to obtain higher tolerance in barley is possible. Based on our results we can definitely say that both factors, low pH and Al, are the enemies of barley. However, aluminum causes more changes at transcriptome level when plants are exposed for this stress for a long time. It should be noted that plants grown on acidic soils are simultaneously exposed to low pH and Al throughout their life.

DATA AVAILABILITY STATEMENT

RNA-Sequencing data reported in this article has been deposited in the Gene Expression Omnibus under the accession no. GSE167271.

AUTHOR CONTRIBUTIONS

IS, MS-Z, MK, and PL conceived the study. IS, MS-Z, and MK designed the experiments. MS-Z and MNa performed the low pH and Al experiments. MNa analyzed the root parameters. MS-Z, KC, and MG isolated RNA for RNA-seq and qPCR analysis. MG performed the qPCR analysis. MNi prepared the libraries for RNA-seq and performed sequencing in the short-term experiment. KC and MK performed the bioinformatic analysis. MS-Z, IS, and PL interpreted the results. MS-Z wrote the manuscript. IS and PL revised and edited the manuscript. All authors have approved the manuscript.

FUNDING

This work was supported by the National Centre for Research and Development, Poland (grant ERA-CAPS-II/2/2015) and by the National Science Centre, Poland (grant Beethoven Life1 2018/31/F/NZ2/03952).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.675260/full#supplementary-material

REFERENCES

Achary, V., Jena, S., Panda, K., and Panda, B. (2008). Aluminium induced oxidative stress and DNA damage in root cells of *Allim cepa* L. *Ecotoxicol. Environ. Saf.* 70, 300–310. doi: 10.1016/j.ecoenv.2007.10.022

Achary, V., and Panda, B. (2009). Aluminium-induced DNA damage and adaptive response to genotoxic stress in plant cells are mediated through reactive oxygen intermediates. *Mutagenesis* 25, 201–209. doi: 10.1093/mutage/gep063

Andersen, C. L., Jensen, J. L., and Ørntoft, T. F. (2006). Characterizing vascular parameters in hypoxic regions: a combined magnetic resonance and optical imaging study of a human prostate cancer model. *Cancer Res.* 66, 9929–9936. doi: 10.1158/0008-5472.CAN-06-0886

Awasthi, J., Saha, B., Panigrahi, J., Yanase, E., Koyama, H., and Panda, S. (2019). Redox balance, metabolic fingerprint and physiological characterization in contrasting North East Indian rice for Aluminum stress tolerance. *Sci. Rep.* 9:8681. doi: 10.1038/s41598-019-45158-3

Bakalovic, N., Passardi, F., Ioannidis, V., Cosio, C., Penel, C., Falquet, L., et al. (2006). PeroxiBase: a class III plant peroxidase database. *Phytochemistry* 67, 534–539. doi: 10.1016/j.phytochem.2005.12.020

Barros, A., Chandnani, R., de Sousa, S., Maciel, L., Tokizawa, M., Guimaraes, C., et al. (2020). Genetic factors conditioning tolerance to multiple stresses for crops cultivated on acidic tropical soils. *Front. Plant Sci.* 11:565339. doi: 10.3389/fpls.2020.565339

Bhalerao, S., and Prabhu, D. (2013). Aluminium toxicity in plants – a review. *J. Appl. Chem.* 2, 447–474.

Boscolo, P., Menossi, M., and Jorge, R. (2003). Aluminum-induced oxidative stress in maize. *Phytochemistry* 62, 181–189. doi: 10.1016/S0031-9422(02)00491-0

Chen, Z. C., Yamaji, N., and Motoyam, R. (2012). Up-regulation of a magnesium transporter gene OsMGT1 is required for conferring aluminum tolerance in rice. *Plant Physiol.* 159, 16224–11633. doi: 10.1104/pp.112.199778

Dalton, L., Ballarin, V., and Brun, M. (2009). Clustering Algorithms: on learning, validation, performance, and applications to genomics. *Curr. Genomic* 10, 430–445. doi: 10.2174/138920209789177601

Delhaize, E., Ryan, P., Hebb, D., Yamamoto, Y., Sasaki, T., and Matsumoto, H. (2004). Engineering high-level aluminum tolerance in barley with the ALMT1 gene. *PNAS* 101, 15249–15254. doi: 10.1073/pnas.0406258101

Dobin, A., Davis, C., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Dong, X., Shen, R., Chen, R., Zhu, Z., and Ma, J. (2008). Secretion of malate and citrate from roots is related to high Al-resistance in Lespedeza bicolor. *Plant Soil* 306, 139–147. doi: 10.1007/s11104-008-9564-x

Eticha, D., Stass, A., and Horst, W. (2005). Cell-wall pectin and its degree of methylation in the maize root-apes: significance for genotypic differences in aluminum resistance. *Plant Cell Environ.* 28, 1410–1420. doi: 10.1111/j.1365-3040.2005.01375.x

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32, 3047–3048. doi: 10.1093/bioinformatics/btw354

Fan, W., Lou, H., Gong, Y., Liu, M., Cao, M., Liu, Y., et al. (2015). Characetrization of an inducible C2H2-type zinc finger transcription factor VuSTOP1 in rice bean (*Vigna umbellata*) reveals differential regulation between low pH and aluminum tolerance mechanisms. *New Phytol.* 208, 456–468. doi: 10.1111/nph.13456

Fan, W., Lou, H., Yang, J., and Zheng, S. (2016). The roles of STOP1-like transcription factors in aluminum and proton tolerance. *Plant Signal. Behav.* 11:e1131371. doi: 10.1080/15592324.2015.1131371

Fang, Q., Zhang, J., Zhang, Y., Fan, N., van den Burg, H., and Huang, C. (2020). Regulation of aluminium resistance in *Arabidopsis* involves the SUMOylation of the zinc finger transcription factor STOP1. *Plant Cell* 32, 3921–3938. doi: 10.1105/tpc.20.00687

Furukawa, J., Yamaji, N., Wang, H., Mitani, N., Murata, Y., Sato, K., et al. (2007). An aluminum-activated citrate transporter in barley. *Plant Cell Physiol.* 48, 1081–1091. doi: 10.1093/pcp/pcm091

Garcia-Oliveira, A., Benito, C., Prieto, P., de Andrade Menezes, R., Rodrigues-Pousada, C., Guedes-Pinto, H., et al. (2013). Molecular characterization of TaSTOP1 homoeologues and their response to aluminium and proton (H(=)) toxicity in bread wheat (*Triticum aestivum* L.). *BMC Plant Biol.* 13:134. doi: 10.1186/1471-2229-13-134

Gavnholt, B., and Larsen, K. (2002). Molecular biology of plant laccases in relation to lignin formation. *Physiol. Plant.* 116, 273–280. doi: 10.1034/j.1399-3054.2002.1160301.x

Giannakoula, A., Moustakas, M., Syros, T., and Yupsanis, T. (2010). Aluminum stress induces up-regulation of an efficient antioxidant system in the Al-tolerant maize line but not in the Al-sensitive line. *Environ. Exp. Bot.* 67, 487–494. doi: 10.1016/j.envexpbot.2009.07.010

Guimaraes, C., Simoes, C., Pastina, M., Maron, L., Magalhaes, J., Vasconcellos, R., et al. (2014). Genetic dissection of Al tolerance QTLs in the maize genome by high density SNP scan. *BMC Genomics* 15:153. doi: 10.1186/1471-2164-15-153

Guo, P., Qi, Y., Yang, L., Lai, N., Ye, X., Yang, Y., et al. (2017). Root adaptive responses to aluminum-treatment revealed by RNA-Seq in two citrus species with different aluminum-tolerance. *Front. Plant. Sci.* 8:330. doi: 10.3389/fpls.2017.00330

Guo, T., Zhang, G., Zhou, M., Wu, F., and Chen, J. (2004). Effects of aluminum and cadmium toxicity on growth and antioxidant enzyme activities of two barley genotypes with different Al resistance. *Plant Soil* 258, 241–248. doi: 10.1023/B:PLSO.0000016554.87519.d6

Gupta, N., Gaurav, S., and Kumar, A. (2013). Molecular basis of aluminium toxicity in plants: a review. *Am. J. Plant Sci.* 4, 21–37. doi: 10.4236/ajps.2013.4.12A3004

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *J. Intellig. Inform. Syst.* 17, 107–145. doi: 10.1023/A:1012801612483

Hossain, M., Hossain, A., Kihara, T., Koyama, H., and Hara, T. (2005). Aluminum-induced lipid peroxidation and lignin deposition are associated with an increase in H2O2 generation in wheat seedlings. *Soil Sci. Plant Nutr.* 51, 223–230. doi: 10.1111/j.1747-0765.2005.tb00026.x

Huang, C., Yamaji, N., Chen, Z., and Ma, F. (2012). A tonoplast-localized half-size ABC transporter is required for internal detoxification of aluminum in rice. *Plant J.* 69, 857–867. doi: 10.1111/j.1365-313X.2011.04837.x

Huang, C., Yamaji, N., and Ma, J. (2010). Knockout of a bacterial-type ATP-binding cassette transporter gene, AtSTAR1, results in increased aluminum sensitivity in Arabidopsis. *Plant Physiol.* 153, 1669–1677. doi: 10.1104/pp.110.155028

Huang, C., Yamaji, N., Mitani, N., Yano, M., Nagamura, Y., and Ma, J. (2009). A bacterial-type ABC transporter is involved in aluminium tolerance in rice. *Plant Cell* 21, 655–667. doi: 10.1105/tpc.108.064543

Ishikawa, S., Wagatsuma, T., Sasaki, R., and Ofei-Manu, P. (2000). Comparison of the amount of citric and malic acids in Al media of seven plant species and two cultivars each in five plant species. *Soil Sci. Plant Nutr.* 46, 751–758. doi: 10.1080/00380768.2000.10409141

Jaskowiak, J., Kwasniewska, J., Milewska-Hendel, A., Kurczynska, E., Szurman-Zubrzycka, M., and Szarejko, I. (2019). Aluminum alters the histology and pectin cel wall composition of barley roots. *Int. J. Mol. Sci.* 20:3039. doi: 10.3390/ijms20123039

Jaskowiak, J., Tkaczyk, O., Slota, M., Kwasniewska, J., and Szarejko, I. (2018). Analysis of aluminium toxicity in Hordeum vulgare roots with an emphasis on DNA integrity and cell cycle. *PLoS One* 13:e0193156. doi: 10.1371/journal.pone.0193156

Jones, D., Blancflor, E., Kochian, L., and Gilroy, S. (2006). Spatial coordination of aluminium uptake, production of reactive oxygen species, callose production and wall rigidification in maize roots. *Plant Cell Environ.* 29, 1309–1318. doi: 10.1111/j.1365-3040.2006.01509.x

Kinraide, T. (2008). Toxicity factors in acidic forest sils: attempts to evaluate separately the toxic effectis of excessive Al3+ and H+ and insufficient Ca2+ and Mg2+ upon root elongation. *Eur. J. Soil Sci.* 54, 323–333. doi: 10.1046/j.1365-2389.2003.00538.x

Kochian, L. (1995). Cellular mechanisms of aluminum toxicity and resistance in plants. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 46, 237–260. doi: 10.1146/annurev.pp.46.060195.001321

Kochian, L., Piñeros, M., and Hoekenga, O. (2005). The physiology, genetics and molecular biology of plant aluminum resistance and toxicity. *Planta and Soil* 274, 175–195. doi: 10.1007/s11104-004-1158-7

Kochian, L., Piñeros, M., Liu, J., and Magalhaes, J. (2015). Plant adaptation to acid soils: the molecular basis for crop aluminum resistance. *Annu. Rev. Plant Biol.* 66, 571–598. doi: 10.1146/annurev-arplant-043014-114822

Kopittke, P., Moore, K., Lombi, E., Gianoncelli, A., Ferguson, B., Blamey, F., et al. (2015). Identification of the primary lesion of toxic aluminium in plant roots. *Plant Physiol.* 167, 1402–1411. doi: 10.1104/pp.114.253229

Kopylova, E., Noé, L., and Touzet, H. (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28, 3211–3217. doi: 10.1093/bioinformatics/bts611

Kumari, M., Taylor, G., and Deyholos, M. (2008). Transcriptomic responses to aluminum stress in roots of *Arabidopsis thaliana*. *Mol. Genet. Genomics* 279, 339–357. doi: 10.1007/s00438-007-0316-z

Larsen, P., Cancel, J., Rounds, M., and Ochoa, V. (2007). Arabidopsis ALS1 encodes a root tip and stele localized half type ABC transporter required for root growth in aluminum toxic environment. *Planta* 225, 1447–1458. doi: 10.1007/s00425-006-0452-4

Larsen, P., Geisler, M., Jones, C., Williams, K., and Cancel, J. (2005). ALS3 encodes a phloem-localized ABC transporter-like protein that is required for aluminum tolerance in *Arabidopsis*. *Plant J.* 41, 353–363. doi: 10.1111/j.1365-313X.2004.02306.x

Lassmann, T., Hayashizaki, Y., and Daub, C. (2011). SAMStat: monitoring biases in next generation sequencing data. *Bioinformatics* 27, 130–131. doi: 10.1093/bioinformatics/btq614

Li, X., Ma, J., and Matsumoto, H. (2000). Pattern of aluminium-induced secretion of organic acids differs between rye and wheat. *Plant Physiol.* 123, 1537–1544. doi: 10.1104/pp.123.4.1537

Li, Y., Huang, J., Song, X., Zhang, Z., Jiang, Y., Zhu, Y., et al. (2017). An RNA-Seq transcriptome analysis revealing novel insights into aluminum tolerance and accumulation in tea plant. *Planta* 246, 91–103. doi: 10.1007/s00425-017-2688-6

Li, Y., Kajita, S., Kawai, S., Katayama, Y., and Morohoshi, N. (2003). Downregulation of an anionic peroxidase in transgenic Aspen and its effect on lignin characteristics. *J. Plant Res.* 116, 175–182. doi: 10.1007/s10265-003-0087-5

Linton, K. (2007). Structure and function of ABC transporters. *Physiology* 22, 122–130. doi: 10.1152/physiol.00046.2006

Liu, W., Xu, F., Lv, T., Zhou, W., Chen, Y., Jin, C., et al. (2018). Spatial responses of antioxidative system to aluminum stress in roots of wheat (*Triticum aestivum* L.) plants. *Sci. Total Environ.* 15, 462–469. doi: 10.1016/j.scitotenv.2018.01.021

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. *Methods* 25, 402–408. doi: 10.1006/meth.2001.1262

Love, M., 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

Ma, B., Gao, L., Zhang, H., Cui, J., and Shen, Z. (2012). Aluminum-induced oxidative stress and changes in antioxidant defenses in the roots of rice varieties differing in Al tolerance. *Plant Cell Rep.* 31, 687–696. doi: 10.1007/s00299-011-1187-7

Magalhaes, J., Liu, J., Guimarães, C., Lana, U., Alves, V., Wang, Y., et al. (2007). A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. *Nat. Genet.* 39, 1156–1161. doi: 10.1038/ng2074

Magnavaca, R., Gardner, C., and Clark, R. (1987). "Evaluation of inbred maize lines for aluminum tolerance in nutrient solution," in *Genetic Aspects of Plant Mineral Nutrition*, eds H. Gabelman and B. Longhman (Dordrecht: Martinus Nijhoff Publishers), 255–265. doi: 10.1007/978-94-009-3581-5_23

Maron, L., Kirst, M., Mao, C., Milner, M., Menossi, M., and Kochian, L. (2008). Transcriptional profiling of aluminum toxicity and tolerance responses in maize roots. *New Phytol.* 179, 116–128. doi: 10.1111/j.1469-8137.2008.02440.x

Maron, L., Piñeros, M., Guimarães, C., Magalhaes, J., Pleima, J., Mao, C., et al. (2010). Two functionally distinct members of the MATE (multi-drug and toxic compound extrusion) family of transporters potentially underlie two major aluminum tolerance QTLs in maize. *Plant J.* 61, 728–740. doi: 10.1111/j.1365-313X.2009.04103.x

Matonyei, T., Barros, B., Guimaraes, R., Ouma, E., Cheprot, R., Apolinário, L., et al. (2020). Aluminum tolerance mechanisms in Kenyan maize germplasm are independent from the citrate transporter ZmMATE1. *Sci. Rep.* 10:7320. doi: 10.1038/s41598-020-64107-z

Meyer, Y., Belin, C., Delorme-Hinous, V., Reichheld, J., and Riondet, C. (2012). Thioredoxin and glutaredoxin systems in plants: molecular mechanisms, crosstalks, and functional significance. *Antioxid. Redox Signal.* 17, 1124–1160. doi: 10.1089/ars.2011.4327

Min, Y., Huilan, Y., Honhai, L., and Lihua, W. (2009). Aluminum induces chromosomes aberrations, micronuclei and cell cycle dysfunction in root cells of *Vicia faba*. *Environ. Tox.* 25, 124–129. doi: 10.1002/tox/20482

Monat, C., Padmarasu, S., Lux, T., Wicker, T., Gundlach, H., Himmelbach, A., et al. (2019). TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. *Genome Biol.* 20:284. doi: 10.1186/s13059-019-1899-5

Nakano, Y., Kusunoki, K., Hoekenga, O., Tanaka, K., Iuchi, S., Sakata, Y., et al. (2020). Genome-wide association study and genomic prediction eluciadte the distinct genetic architecture of aluminum and proton tolerance in *Arabidopsis thaliana*. *Front. Plant Sci.* 11:405. doi: 10.3389/fpls.2020.00405

Nezames, C., Sjorgen, C., Barajas, J., and Larsen, P. (2012). The Arabidopsis cell cycle checkpoint regulators TANMEI/ALT2 and ATR mediate the active process of aluminum-dependent root growth inhibition. *The Plant Cell* 24, 608–621. doi: 10.1105/tpc.112.095596

Nishinaka, Y., Masutani, H., Nakamura, H., and Yodoi, J. (2001). Regulatory roles of thioredoxin in oxidative stress-induced cellular responses. *Redox Repoty.* 6, 289–295. doi: 10.1179/135100001101536427

Okonechnikov, K., Conesa, A., and García-Alcalde, F. (2016). QualiMap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32, 292–294. doi: 10.1093/bioinformatics/btv566

Pereira, L., Mazzanti, C., Gonçalves, J., Cargnelutti, D., Tabaldi, L., Becker, A., et al. (2010). Aluminum-induced oxidative stress in cucumber. *Plant Physiol. Biochem.* 48, 683–689. doi: 10.1016/j.plaphy.2010.04.008

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

Rahman, M., Lee, S.-H., Ji, H., Kabir, A., Jones, C., and Lee, K.-W. (2018). Importance of mineral nutrition for mitigating aluminum toxicity in plants on acidic soils: current status and opportunities. *Int. J. Mol. Sci.* 19:3073. doi: 10.3390/ijms19103073

Rengel, Z., Bose, J., Chen, Q., and Tripathi, B. (2015). Magnesium alleviates plant toxicity of aluminium and heavy metals. *Crop Pasture Sci.* 66, 1298–1307. doi: 10.1071/CP15284

Riaz, M., Yan, L., Wu, X., Hussain, S., Aziz, O., and Jiang, C. (2018). Mechanisms of organic acids and boron induced tolerance of aluminum toxicity: a review. *Ecotoxicol. Environ. Saf.* 165, 25–35. doi: 10.1016/j.ecoenv.2018.08.087

Richards, K., Schott, E., Sharma, Y., Davis, K., and Gardner, R. (1998). Aluminum induces oxidative stress genes in *Arabidopsis thaliana*. *Plant Physiol.* 116, 409–418. doi: 10.1104/pp.116.1.409

Rounds, M., and Larsen, P. (2008). Aluminum-dependent root-growth inhibition in *Arabidopsis* results from AtATR-regulated cell-cycle arrest. *Curr. Biol.* 18, 1495–1500. doi: 10.1016/j.cub.2008.08.050

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20, 53–65. doi: 10.1016/0377-0427(87)90125-7

Ruijter, J. M., Ramakers, C., Hoogaars, W. M. H., Karlen, Y., Bakker, O., van den Hoff, M. J. B., et al. (2009). Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. *Nucleic Acids Res.* 37:e45. doi: 10.1093/nar/gkp045

Ryan, P., Raman, H., Gupta, S., Horst, W., and Delhaize, E. (2009). A second mechanism for aluminum resistance in wheat relies on the constitutive efflux of citrate from roots. *Plant Physiol.* 149, 340–351. doi: 10.1104/pp.108.129155

Sade, H., Meriga, B., Surapu, V., Gadi, J., Sunita, M., Suravajhala, P., et al. (2016). Toxicity and tolerance of aluminum in plants: tailoring plants to suit to acid soils. *Biometals* 29, 187–210. doi: 10.1007/s10534-016-9910-z

Sasaki, M., Yamamoto, Y., and Matsumato, H. (2006). Lignin deposition induced by aluminum in wheat (*Triticum aestivum*) roots. *Physiol. Palnt.* 96, 193–198. doi: 10.1111/j.1399-3054.1996.tb00201.x

Sasaki, T., Yamamoto, Y., Ezaki, B., Katsuhara, M., and Ahn, S. (2004). A wheat gene encoding an aluminum-activated malate transporter. *Plant J.* 37, 645–653. doi: 10.1111/j.1365-313X.2003.01991.x

Sawaki, Y., Iuchi, S., Kobayashi, Y., Kobayashi, Y., Ikka, T., Sakurai, N., et al. (2009). STOP1 regulates multiple genes that protect Arabidopsis from proton and aluminum toxicities. *Plant Physiol.* 150, 281–294. doi: 10.1104/pp.108.134700

Shaff, J., Schultz, B., Craft, E., Clark, R., and Kochian, L. (2010). GEOCHEM-EZ: a chemical speciation program with greater power and flexibility. *Plant Soil* 330, 207–214. doi: 10.1007/s11104-009-0193-9

Silva, I., Smyth, T., Moxley, D., Carter, T., Allen, N., and Rufty, T. (2000). Aluminum accumulation at nuclei of cells in the root tip. Fluorescence detection using lumogallion and confocal laser scanning microscopy. *Plant Physiol.* 123, 543–552. doi: 10.1104/pp/123.2.543

Silva, S. (2012). Aluminium toxicity targets in plants. *J. Bot.* 2012:219462. doi: 10.1155/2012/219462

Singh, S., Tripathi, D., Singh, S., Sharma, S., Dubey, N., Chauhan, D., et al. (2017). Toxicity of aluminium on various levels of plant cells and organism: a review. *Environ. Exp. Bot.* 137, 177–193. doi: 10.1016/j.envexpbot.2017.01.005

Sun, C., Liu, L., Zhou, W., Lu, L., Jin, C., and Kin, X. (2017). Aluminum induces distinct chamges in the metabolism od reactive oxygen and nitrogen species in the roots of two wheat genotypes with different aluminum resistance. *J. Agric. Food Chem.* 65, 9419–9427. doi: 10.1021/acs.jafc.7b03386

Szurman-Zubrzycka, M., Nawrot, M., Jelonek, J., Dziekanowski, M., Kwasniewska, J., and Szarejko, I. (2019). ATR, a DNA damage signaling kinase, is involved in aluminum response in barley. *Fron. Plant Sci.* 10:1299. doi: 10.3389/fpls.2019.01299

Szurman-Zubrzycka, M., Zbieszczyk, J., Marzec, M., Jelonek, J., Chmielewska, B., Kurowska, M., et al. (2018). HorTILLUS – a rich and renewable source of induced mutations for forward/reverse genetics and pre-breeding programs in barley (*Hordeum vulgare* L.). *Front. Plant Sci.* 9:216. doi: 10.3389/fpls.2018.00216

Tamás, L., Budíková, S., Huttová, J., Mistrík, I., Simonovicová, M., and Siroká, B. (2005). Aluminum-induced cell death of barley-root border cells is correlated with peroxidase- and oxalate oxidase-mediated hydrogen peroxide production. *Plant Cell Rep.* 24, 189–194. doi: 10.1007/s00299-005-0939-7

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Res.* 40, 1–12. doi: 10.1093/nar/gks596

Vega, I., Nikolic, M., Pontigo, S., Godoy, K., de La Luz Mora, M., and Cartes, P. (2019). Silicon improves the production of high antioxidant or structural phenolic compounds in barley cultivars under aluminum stress. *Agronomy* 9:388. doi: 10.3390/agronomy9070388

Von Uexküll, H., and Mutert, E. (1995). Global extent, development and economic impact of acidic soils. *Plant Soil* 171, 1–15. doi: 10.1007/bf00009558

Wang, J., and Kao, C. (2007). Protective effect of ascorbic acid and glutathione on AlCl3-inhibited growth of rice roots. *Biol. Plant.* 51, 493–500. doi: 10.1007/s10535-007-0104-y

Wang, J., Raman, H., Zhang, G., Mendham, N., and Zhou, M. (2006). Aluminum tolerance in barley (*Hordeum vulgare* L.): physiological mechanisms, genetics and screening methods. *J. Zhejiang Univ. Sci. B* 7, 769–787. doi: 10.1631/jzus.2006.B0769

Xu, F., Li, G., Jin, C., Liu, W., Zhang, S., Zhang, Y., et al. (2012). Aluminum-induced changes in reactive oxygen species accumulation, lipid peroxidation and antioxidant capacity in wheat root tips. *Biol. Plant* 51, 89–96. doi: 10.1007/s10535-012-0021-6

Yamaji, N., Huang, C., and Nagao, S. (2009). A zinc finger transcription factor ART1 regulated multiple genes implicated in aluminum tolerance in rice. *Plant Cell* 21, 3339–3349. doi: 10.1105/tpc.109.070771

Yamamoto, Y., Kobayashi, Y., Davi, S., Rikiishi, S., and Matsumoto, H. (2003). Oxidative stress triggered by aluminum in plant roots. *Plant Soil* 255, 239–243. doi: 10.1023/A:1026127803156

Yang, J., Fan, W., and Zheng, S. (2019). Mechanisms and regulation of aluminum-induced secretion of organic acid anions from plant roots. *J. Zhejinag Univ. Sci. B* 20, 513–527. doi: 10.1631/jzus.B1900188

Yang, L., Qi, Y., Jiang, H., and Chen, L. (2013). Roles of organic acid anion secretion in aluminium tolerance of higher plants. *BioMed. Res. Int.* 2013:173682. doi: 10.1155/2013/173682

Yang, X., Yang, J., Zhou, Y., Piñeros, M., Kochian, L., Li, G., et al. (2011). A de novo synthesis citrate transporter, *Vigna umbellata* multidrug and toxic compound extrusion, implicates in Al-activated citrate efflux in rice bean (*Vigna umbellata*) root apex. *Plant Cell Environ.* 34, 2138–2148. doi: 10.1111/j.1365-3040-2011.02410.x

Yang, Z., Sivaguru, M., Horst, W., and Matsumoto, H. (2000). Aluminium tolerance is achieved by exudation of citric acid from roots of soybean (*Glycine max*). *Physiol. Plant.* 110, 72–77. doi: 10.1034/j.1399-3054.2000.110110.x

Yokosho, K., Yamaji, N., and Ma, J. (2010). Isolation and characterisation of two MATE genes in rye. *Funct. Plant Biol.* 37, 296–303. doi: 10.1071/FP09265

Yokosho, K., Yamaji, N., and Ma, J. (2011). An Al-inducible MATE gene is involved in externanl detoxification of Al in rice. *Plant J.* 68, 1061–1069. doi: 10.1111/j.1365-313x.2011.04757.x

You, J., and Chan, Z. (2015). ROS regulation during abiotic stress responses in crop plants. *Front. Plant Sci.* 6:1092. doi: 10.3389/flps.2015.01092

You, J., He, Y., Yang, J., and Zheng, J. (2005). A comparison of aluminum resistance among Polygonum species originating on strongly acidic and neutral soils. *Plant Soil* 276, 143–151. doi: 10.1007/s11104-005-3786-y

Zhang, Y., Zhu, D., Zhang, Y., Chen, H., Xiang, J., and Lin, X. (2015). Low pH-induced changes of antioxidant enzyme and ATPase activities in the roots of rice (*Oryza sativa* L.) seedling. *PLoS One* 10:e0116971. doi: 10.1371/journal.pone.0116971

Zhao, Z., Ma, J., Sato, K., and Takeda, K. (2003). Differential Al resistance and citrate secretion in barley (*Hordeum vulgare* L.). *Planta* 217, 794–800. doi: 10.1007/s00425-003-1043-2

Zheng, S., and Yang, J. (2005). Target sited of aluminum phytotoxicity. *Biol. Plant.* 49, 321–331. doi: 10.1007/s10535-005-0001-1

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.

The handling editor declared a past co-authorship with the authors MS-Z and IS.

frontiers in Genetics

ORIGINAL RESEARCH
published: 19 May 2021
doi: 10.3389/fgene.2021.623419

Identification of Rice Blast Loss-of-Function Mutant Alleles in the Wheat Genome as a New Strategy for Wheat Blast Resistance Breeding

Huijun Guo[1†], Qidi Du[2†], Yongdun Xie[1], Hongchun Xiong[1], Linshu Zhao[1], Jiayu Gu[1], Shirong Zhao[1], Xiyun Song[2], Tofazzal Islam[3] and Luxiang Liu[1]*

[1] National Engineering Laboratory for Crop Molecular Breeding, National Center of Space Mutagenesis for Crop Improvement, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China, [2] College of Life Sciences, Qingdao Agricultural University, Qingdao, China, [3] Institute of Biotechnology and Genetic Engineering (IBGE), Bangabandhu Sheikh Mujibur Rahman Agricultural University, Gazipur, Bangladesh

OPEN ACCESS

Edited by:
Gorji Marzban,
University of Natural Resources and Life Sciences, Vienna, Austria

Reviewed by:
Momina Hussain,
National Institute for Biotechnology and Genetic Engineering, Pakistan
Bradley Till,
University of California, Davis, United States
Alejandra Landau,
Instituto de Genética "Ewald A. Favret", Instituto Nacional de Tecnología Agropecuaria, Argentina

***Correspondence:**
Luxiang Liu
liuluxiang@caas.cn

[†] These authors have contributed equally to this work

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

Received: 30 October 2020
Accepted: 13 April 2021
Published: 19 May 2021

Citation:
Guo H, Du Q, Xie Y, Xiong H, Zhao L, Gu J, Zhao S, Song X, Islam T and Liu L (2021) Identification of Rice Blast Loss-of-Function Mutant Alleles in the Wheat Genome as a New Strategy for Wheat Blast Resistance Breeding. Front. Genet. 12:623419. doi: 10.3389/fgene.2021.623419

Blast is caused by the host-specific lineages of the fungus *Magnaporthe oryzae* and is the most important destructive disease in major crop plants, including rice and wheat. The first wheat blast outbreak that occurred in Bangladesh in 2016 and the recent epidemic in Zambia were caused by the *M. oryzae Triticum* (*MoT*) pathotype, a fungal lineage belonging to *M. oryzae*. Although a few reported wheat cultivars show modest resistance to *MoT*, the patterns of genetic variation and diversity of this pathotype make it crucial to identify additional lines of resistant wheat germplasm. Nearly 40 rice blast resistant and susceptible genes have so far been cloned. Here, we used BLAST analysis to locate two rice blast susceptible genes in the wheat reference genome, *bsr-d1* and *bsr-k1*, and identified six identical homologous genes located on subgenomes A, B, and D. We uncovered a total of 171 single nucleotide polymorphisms (SNPs) in an ethyl methanesulfonate (EMS)-induced population, with mutation densities ranging from 1/1107.1 to 1/230.7 kb through Targeting Induced Local Lesions IN Genomes (TILLING) by sequencing. These included 81 SNPs located in exonic and promoter regions, and 13 coding alleles that are predicted to have severe effects on protein function, including two pre-mature mutants that might affect wheat blast resistance. The loss-of-function alleles identified in this study provide insights into new wheat blast resistant lines, which represent a valuable breeding resource.

Keywords: wheat, rice blast, wheat blast, TILLING, mutant allele, deleterious effect

INTRODUCTION

Wheat blast is now a serious threat to food and nutritional security in three different continents, namely South America, Asia, and Africa (Islam et al., 2020). The first ever reported wheat blast epidemic occurred in Brazil in 1985 (Igarashi et al., 1986) and have taken place in the other South American countries in following decades, and subsequently spread to the neighboring wheat growing areas in Argentina, Bolivia, and Paraguay. In February 2016, a major outbreak affected 16% of the wheat planting area in Bangladesh, leading to an almost complete crop failure across 15000

hectares (Islam et al., 2016). Finally, during the 2017–2018 growing season, a widespread epidemic significantly affected most cultivars in both experimental and farming fields in Zambia (Tembo et al., 2020). It has been demonstrated that the pathogen *Magnaporthe oryzae pathotype Triticum* (*MoT*) was responsible for the outbreaks in both Bangladesh and Zambia, and that this lineage is closely related to those responsible for the wheat blast outbreak that occurred in South America (Islam et al., 2016; Tembo et al., 2020). Ceresini et al. (2018) assumed that wheat blast disease was introduced in Bangladesh through wheat grain trading from Brazil. In fact, previous research has shown that *M. oryzae* jumped from a native grass host to wheat during the 1980s in Brazil, after which a mutation in one of the isolates causing increased pathogenicity and the functional loss of resistance genes led to widespread *MoT* in wheat cultivars (Inoue et al., 2017).

Due to the relatively recent emergence of *Triticum*, there are only a few known resistant (R) genes available against this destructive pathogen in natural wheat varieties or germplasm (Islam et al., 2020). Beyond the well-characterized 2NS/2AS translocation genotypes that were acquired from *Aegilops ventricosa* (Cruz et al., 2016b; Cruppe et al., 2019; Juliana et al., 2020), the genes Rmg8 and RmgGR119, from the Albanian accession GR119, seemingly confer high blast resistance at both the heading stage and under high temperature conditions (Anh et al., 2015; Wang et al., 2018). While these genes are crucial to the current efforts to breed blast resistant wheat varieties, it has been shown that Rmg8 can be suppressed by *MoT*'s effector gene PWT4 (Inoue et al., 2020), and that resistance of 2NS translocation was eroded by new *MoT* virulence groups (Cruz et al., 2016a), which means other resistant mechanisms might become obsolete with the evolution of *MoT* in the near future. Hence, it is urgent to develop durable blast resistant wheat varieties and, especially, to identify novel non-2NS R genes in order to effectively control the threat posed by *MoT*. One possibility is through mutation induction, a mechanism that has been shown to be effective in creating novel alleles (Campbell et al., 2012; Lu et al., 2018) and germplasms (Xiong et al., 2017; Guo et al., 2019), and that can also be used to generate new *MoT* resistant varieties.

Over one hundred rice blast R and susceptible (S) genes and QTLs have so far been discovered or cloned, including *Ptr*, *Pi-ta*, *Pi-b*, and *Pi-21* (Srivastava et al., 2017; Zhao et al., 2018). In the case of the S gene *Pi-21* (Os04g0401000), the simultaneous deletions of 18- and 48-bp confer non-specific and durable resistance to rice blast. However, the gene is tightly linked with a locus associated with poor eating quality, which makes its use less than ideal to improve disease resistance (Fukuoka et al., 2009). Another example is BSR-K1, a protein that contains five tetratricopeptide repeats (TPRs) and binds to the mRNA of defense-related genes. The genotypes that encode for *Bsr-k1* (*Os10g0548200*) are susceptible to rice blast, while those encoding the *bsr-k1* allele, a pre-mature termination mutation, show broad resistance against both blast and bacterial blight (Zhou et al., 2018). Finally, *bsr-d1* (*Os03g32230*) is a loss of function allele that confers broad spectrum rice blast resistance in natural rice varieties. The gene encodes a putative C2H2-like transcription factor in the nucleus and is regulated by a MYB family transcription factor. Importantly, in this case, no unfavorable genes are known to be closely linked (Li W. et al., 2017).

Loss-of-function mutations therefore represent one of the ways to obtain fungal disease resistance in both natural populations and breeding scenarios. One example is the well-known *Fhb1* (*His*), a gene which encodes a histidine-rich calcium-binding protein and that originated in the lower reaches of the Yangtze Valley of China. The gene contains a 752-bp deletion within its 5′ end that confers resistance against Fusarium head blight (Li et al., 2019) and has been utilized worldwide as one of the best genetic resources in wheat breeding (Hao et al., 2020). Another example is the mildew resistant locus o (*mlo*) where resistance-conferring missense and knockout mutations against powdery mildew were induced in the conserved region of the gene by ethyl methanesulfonate (EMS) mutagen treatment and gene editing approaches (Wang et al., 2014; Acevedo-Garcia et al., 2017). Notwithstanding, *Tamlo* alleles were more susceptible to *MoT* (Gruner et al., 2020).

It has been demonstrated that chemical and physical mutagens are able to induce nucleotide changes, including substitutions, insertion, or deletions (Ahloowalia and Maluszynski, 2001; Du et al., 2017; Krasileva et al., 2017; Ichida et al., 2019), that represent loss-of-function mutations resulting in favorable, fungal-resistant phenotypes (Acevedo-Garcia et al., 2017; Hussain et al., 2018). Targeting Induced Local Lesions IN Genomes (TILLING) is a reverse genetic approach to identify mutant allele (McCallum et al., 2000), and it has been used to discover mutant alleles in wheat, rice, barley and many other species. The target traits, such as wheat starch quality (Slade et al., 2005, 2012; Hazard et al., 2012), rice phytic acid and starch (Kim and Tai, 2014; Kim et al., 2018), have been improved through the approach. There are several different methods have been developed to TILL mutant alleles, such as gel electrophoresis based on enzyme digestion (Till et al., 2006), high resolution melting (Dong et al., 2009; Acanda et al., 2014), and the higher throughput TILLING by sequencing (Tsai et al., 2011).

Here, we tried to establish a new strategy aimed at identifying *MoT* resistance in wheat based on knowledge associated with rice blast resistance. Specifically, we took advantage of the close evolutionary relationship between *MoT* and *M. oryzae (MoO)*, BLASTed rice blast S orthologs in the wheat reference genome, and analyzed their functional domains. We then used EMS mutagen treatment and TILLING by sequencing in order to identify mutant single nucleotide polymorphisms (SNPs) in the M2 population that severely impact gene function and that might have the potential to enhance blast resistance in wheat. Our approach provides a new strategy to enhance the genetic diversity of wheat blast resistant germplasm.

MATERIALS AND METHODS

Plant Materials

Wheat (*Triticum aestivum* L.) cultivar Jing411 and its EMS-induced M2 mutated population (Guo et al., 2017)

were used to identify mutant alleles in target genes. Five biological replicates of wild type (WT) were used as reference.

A total of 2,300 M2 individuals were used for mutation screening. M1 plants were strictly self-crossed by bagging, and a single seed was harvested from each plant to develop the M2 population, leaves of each M2 individual plants were sampled to extract DNA. All samples were normalized to the same concentration (50 ng/μl) and placed in 96-well plates. A two dimensional pooling scheme was used following protocol of Till et al. (2006) with modification, the 12 samples in each column were pooled into one sample, and the eight samples in each row were pooled into another (**Supplementary Figure 1**), a total of 571 pooled samples were obtained. All pooling samples were then used for TILLING by sequencing.

The M3 mutants which were predicted to have severe impacts were used to validate variations, each mutant line was planted 20–40 seeds according to their total seed amount. The seeds were planted in the experimental field of Institute of Crop Sciences, Chinese Academy of Agricultural Sciences. Seedling leaf of each individual was sampled to extract DNA for mutation confirmation.

Sequence Blast and Analysis

We used the DNA sequences of rice blast S genes *Os10g0548200*, *Os03g32230*, and *Os04g0401000*, from the Rice Annotation Project Database[1], as templates to BLAST in the wheat reference genome Version 2.0[2]. The wheat orthologs found across the three sub-genomes were then analyzed in NCBI's database[3] in order to access their conserved functional domains, which were used as the target sequences of mutation detection by TILLING. Specific primers (**Supplementary Table 1**) were designed using the software GenoPlexs Primer Designer (Molbreeding Company, China), and used to amplify pooling samples.

TILLING by Sequencing

The PCR reaction and library construction was prepared using the GenoPlexs Multiplex-PCR Library Prep Kit (Molbreeding Company, China), each step was performed according to the kit manual. The PCR reaction included 50 ng DNA, 1× T PCR Master Mix with improved high-fidelity pfu thermostable DNA polymerase and the primer mix. Amplification conditions included denaturation at 95°C for 5 min, followed by 32 cycles of 95°C for 30 s, 60°C for 30 s, and 72°C for 5 min on an ABI 9700 thermal cycler. The PCR products were then fragmented with an ultrasonic cleaner (Xinzhi Biotechnology, Ningbo, China, Scientz08-III) and, the fragment size and concentration were detected by agarose gel electrophoresis. After normalization, the products were purified with AMPure XP (Beckman Coulter, A63880).

The purified products were further used to add adaptor and barcode. Firstly the ends were repaired with Repair Enzyme by incubating 20 min on an ABI 9700 thermal cycler, and the A base was added to 3′ ends at the same time; then the adapters were added, which was incubated at 22°C for 60 min on an ABI 9700 thermal cycler. Then, a second purification round was followed before adding barcode. Finally, the barcode was added in conditions of denaturation at 98°C for 2 min, followed by 12 cycles of 98°C for 30 s, annealing for 30 s, and 72°C for 40 s, final extension at 72°C for 4 min. The sequence of barcode was AGTCGGAGGCCAAGCGGTCTTAGGAAGACAANNNNNN NNNNCAACTCCTTGGCTCACA, and the bottom adapter was TTGTCTTCCTAAGGAACGACATGGCTACGATCCGACT.

After a third purification round and fragment size detection, the library was sequenced by MGISEQ2000 (MGI Tech Co., Ltd., China).

Mutation Detection

We filtered the raw reads to fetch clean reads using the software fastp V0.20.0 with parameters -n 10 -q 20 -u 40 (Chen et al., 2018). The Clean reads (BioProject ID PRJCA004347, deposited at National Genomics Data Center)[4] were mapped to amplicon sequences of Chinese Spring (IWGSC RefSeq V1.0) using BWA-mem with default parameters[5] (Li and Durbin, 2009). Sorting were performed with Picard (Version 2.1.1)[6]. GATK's (version v3.5-0-g36282e4) module UnifiedGenotyper was used to call SNPs with parameters: -dcov 1000000 -minIndelFrac 0.15 -glm BOTH -l INFO; and module VariantFiltration was used to filter variants with parameters: -filterExpression "MQ0 ≥ 4 & ((MQ0/(1.0 * DP)) > 0.1)," -filterName "HARD_TO_VALIDATE," -filterExpression "DP < 5 || QD < 2," -filterName "LOW_READ_SUPPORT." Variants were discovered from the VCF file (**Supplementary File 1**) using Perl scripts (**Supplementary File 2**). SNPs with <5× sequencing depth were treated as missing data. The variations between WT and Chinese Spring were filtered out.

The called SNPs were further corrected with frequency. All of the heterozygous sites in WT were considered to be false positive, and they were firstly filtered out before correction with ratio of alter alleles depth to read depth ≤0.20 or ≥0.80, which was higher than those of mutant call. Then SNPs were corrected in mutant pooling samples with the following threshold, when the ratio of alter alleles depth to read depth ≤0.05, the SNPs were considered to be homozygous and identity with reference sites, with the ratio ≥0.95 were considered to be homozygous mutation sites, and with the remainder being considered as heterozygous mutation sites. The mutant SNPs identified in both the row-pooling-sample and the line-pooling-sample were considered to represent true mutations, while those that were only detected in either the row-pooling-sample or the line-pooling-sample were considered to be false positives (**Supplementary Figure 1** and **Supplementary File 3**). Those of SNPs identified in the antisense strands were substituted by complementary bases in the sense strands, and listed in tables.

The mutation density of each gene was calculated by dividing the total number of SNPs by the total sequenced

[1]https://rapdb.dna.affrc.go.jp

[2]https://wheat-urgi.versailles.inra.fr/

[3]https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

[4]https://bigd.big.ac.cn/gsub/

[5]http://bio-bwa.sourceforge.net/bwa.shtml

[6]http://broadinstitute.github.io/picard/

length (sequenced length of the gene multiplied by number of sampled individuals).

Prediction of Mutation Effects

The SNPs classified as true positives were then classified into promoter, exon and intron regions according to their respective location on the genes, and the effects on protein translation of those lying in the coding region analyzed. The impacts of missense mutations were predicted using the online software PROVEAN (Protein Variation Effect Analyzer)[7].

Structure Prediction of Mutant Proteins

The secondary protein structure was predicted using the website http://www.prabi.fr/. The three-dimensional (3D) structures of non-sense and missense mutations were predicted using the SWISS-MODEL server and 3D models generated from multiple threading alignments of amino acid sequences (Bordoli et al., 2009; Biasini et al., 2014). The protein structures were edited and visualized using the software Deepview/Swiss PDB Viewer V.4.1.0.

Validation of Mutant Lines by Sanger Sequencing

Specific primers for each SNP were designed manually according to the specificity of 3′ end. The PCR reaction included 1× Taq Plus Master Mix II (Vazyme Biotech Co., Ltd.), 10 μm primer mix and 100 ng/μl genomic DNA. Amplification conditions included denaturation at 95°C for 3 min, followed by 35 cycles of 95°C for 15 s, annealing for 20 s, and 72°C for 1 min. The PCR products were then detected by 1% agarose gel electrophoresis, those with single band were further sequenced to detect the specificity of primers. Finally, individual samples of each mutant were amplified by the specific primers with two biological repeats, and sequenced by Sanger sequencing to validate SNP variation.

RESULTS

Identification of Homologous Rice Blast S Genes in Wheat

Through BLAST in the wheat reference genome, orthologs of rice blast S gene *Bsr-k1* (*Os10g0548200*) were identified in the first homologous group 1A (TraesCS1A02G207700), 1B (TraesCS1B02G221400), and 1D (TraesCS1D02G211000) (**Supplementary Figure 2A**), while *Bsr-d1* (*Os03g32230*) orthologs were present on the seventh homologous group 7A (TraesCS7A02G160700), 7B (TraesCS7B02G065700), and 7D (TraesCS7D02G161800) (**Supplementary Figure 2B**). However, no *Pi-21* (Os04g0401000) orthologs were identified in the wheat reference genome (**Supplementary Table 2**).

The three *Bsr-k1* wheat orthologs consist of 20 exons and 19 introns (**Supplementary Figure 2A**), and include five sets of conserved functional TPR domains that were observed in wheat homologous genes (**Figure 1**). Their respective observed protein sequence identity was higher than 97%, and more than 80% when compared to BSR-K1.

We have also observed that the sequence identity of BSR-D1 with its wheat orthologs was only 62.3–64.1%. However, the C2H2-type zinc finger domains of *Bsr-d1* were highly conserved in three wheat orthologs (**Figure 2**), whereby its function might be maximally preserved in wheat.

Mutation Density and Substitution Types of Target Fragments

The density of mutations in the six target fragments ranged from 1/1107.1 to 1/230.7 kb (**Table 1**), with an average of 1/309.5 kb. The lowest mutation density was found in the gene TraesCS7A01G160700, where only three mutants were detected.

More than 90% of base substitutions detected were transitions, and the remainder were transversions. All transversions occurred in intronic regions and corresponded to mutations from C, T, or A into G, A, or C. The only exception was found in the 5′UTR region and included a C > G transversion that resulted in a start-codon gain in line E1354. No deletions or insertions were detected in the population.

The Effects of SNPs in *Bsr-k1* Wheat Orthologs

The PCR amplicons were verified by agarose gel electrophoresis (**Supplementary Figure 3** and not shown) and, after fragmentation, purification and adding adaptor, the products were sequenced by next-generation sequencing, and the sequences of WT were submitted to National Center for Biotechnology Information (NCBI) database (**Table 1**). In total, we identified 146 mutated SNPs in the three *Bsr-k1* wheat orthologs in total in the M2 population. The mutations were distributed across the promoter, exonic and intronic regions (**Figure 3**). The mutations overlapping the coding region (CDS) were classified into silent, missense and non-sense mutation types due to their respective effects on amino acid translation. A total of 10 mutants were predicted to have severe effects on gene function.

Moreover, a total of 55 SNPs in the gene TraesCS1A02G207700 were identified, including 15 and 40 SNPs located in exons and introns, respectively (**Table 2**, **Figure 3**, and **Supplementary Table 3**). Among the 11 SNPs found in the CDS region, five resulted in missense mutations and two (line E758 and E325) were predicted to severely impact gene function, while the other six represented silent mutations. We also found a start-codon-gain mutant in the 5′UTR region, which resulted in a 132-base advance of the starting codon without any downstream frameshift.

A total of 45 mutated SNPs distributed across promoter, exonic, and intronic regions were identified in the gene TraesCS1B02G221400 (**Table 2**, **Supplementary Table 3**, and **Figure 3**). Among these, we found 10 missense mutations, 4 of which were predicted to have a deleterious impact. Furthermore, there were four mutations in the promoter region that may also lead to variations in gene function.

[7]http://provean.jcvi.org/index.php

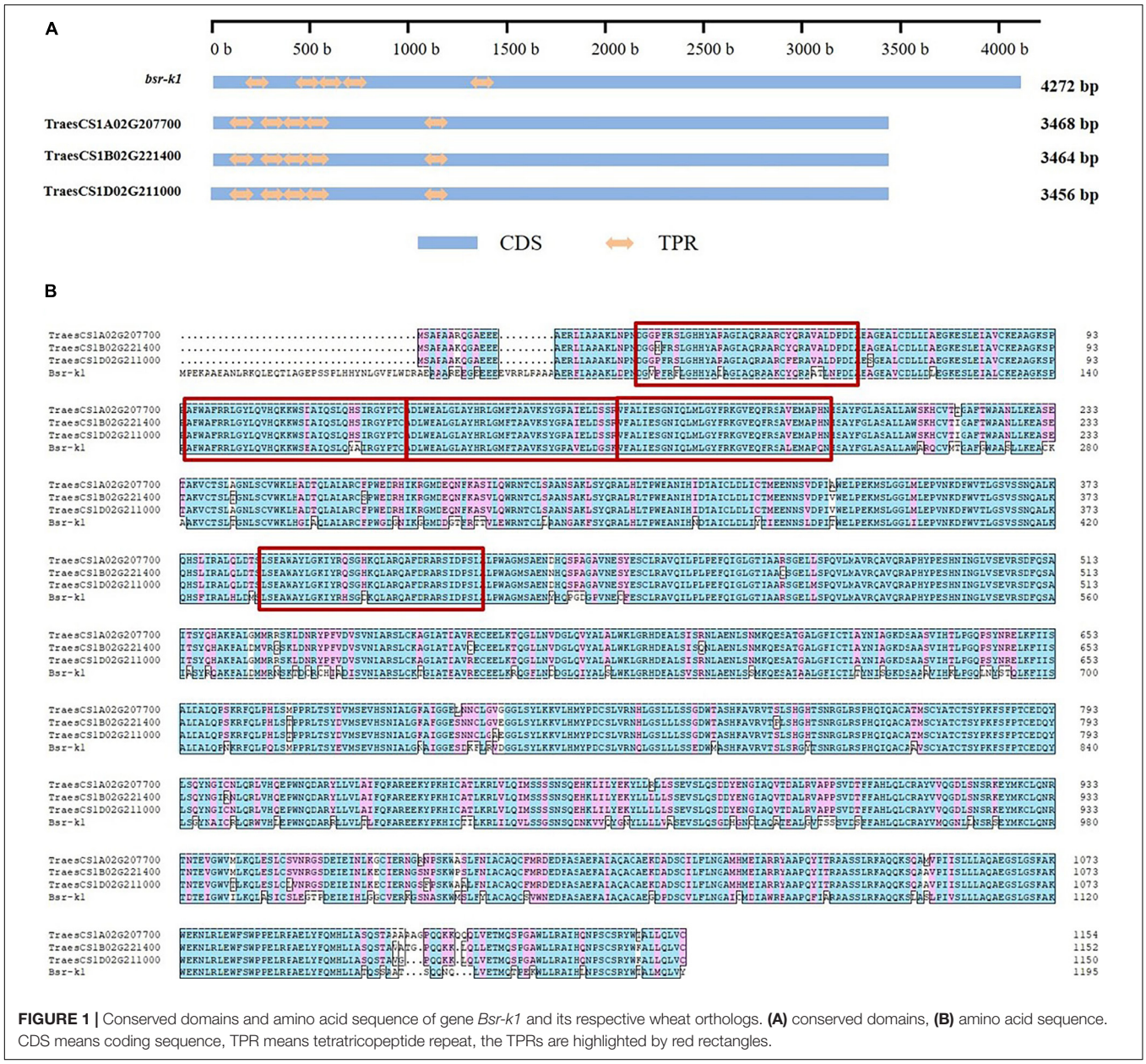

FIGURE 1 | Conserved domains and amino acid sequence of gene *Bsr-k1* and its respective wheat orthologs. **(A)** conserved domains, **(B)** amino acid sequence. CDS means coding sequence, TPR means tetratricopeptide repeat, the TPRs are highlighted by red rectangles.

Finally, we identified 20 SNPs in the exonic regions of the gene TraesCS1D02G211000 (**Table 2**, **Supplementary Table 3**, and **Figure 3**), of which two were non-sense and 10 missense mutations. Importantly, the two stop-gained mutants (E60 and E724) as well as C604T (E91) and C2740T (E315) might lead to severe impacts on function.

The Effects of SNPs in *Bsr-d1* Wheat Orthologs

A total of 25 SNPs were found in the three wheat orthologs of *Bsr-d1* in the M2 population, including one start-codon loss, 10 missense mutations, and several others located in the UTR and promoter regions (**Table 3**). PROVEAN analysis predicted that the loss of the start-codon (G135A) is neutral due to the existence of an alternative start codon within 12 base pairs without any downstream frameshift. This analysis also predicted that the C488T mutation in TraesCS7A02G160700 and the C304T and G497A mutations in TraesCS7D02G161800 have a severe effect on protein function.

Verification of SNPs With Severe Impacts in M3

A total of 13 SNPs with severe impacts were discovered in the six target genes, 12 of them and five of those located in UTR and promoter region were further validated, except mutant line E044-3 because of insufficient seeds. A total of 12 sets of specific primers were used after electrophoresis and sequencing evaluation (**Supplementary Table 4** and **Supplementary Figure 4**). 100%

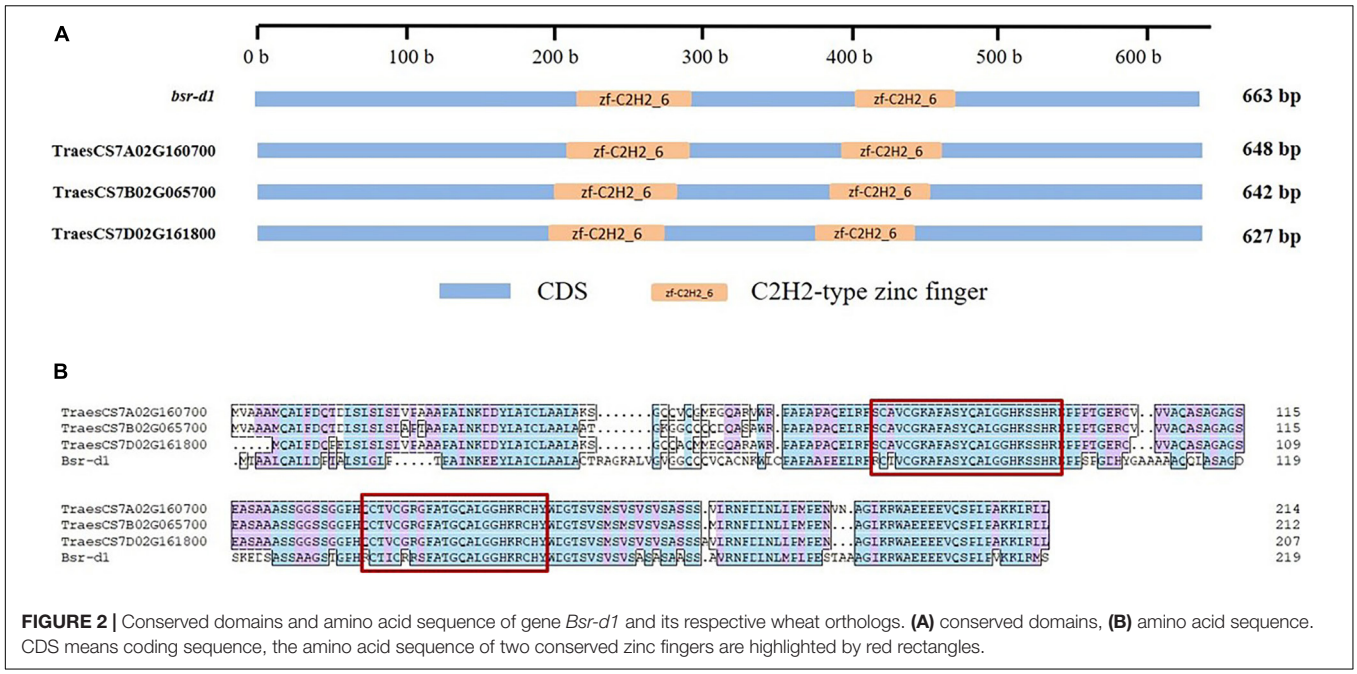

FIGURE 2 | Conserved domains and amino acid sequence of gene *Bsr-d1* and its respective wheat orthologs. **(A)** conserved domains, **(B)** amino acid sequence. CDS means coding sequence, the amino acid sequence of two conserved zinc fingers are highlighted by red rectangles.

TABLE 1 | Mutation densities of *Bsr-k1* and *Bsr-d1* wheat orthologs in the M2 population after EMS treatment.

Gene	NCBI accession number	Gene size (kb)	Sequenced fragment size (kb)	Mutation number	Mutation density
TraesCS1A02G207700	MW388661	10.948	6.950	55	1/290.6 kb
TraesCS1B02G221400	MW388662	8.444	4.514	45	1/230.7 kb
TraesCS1D02G211000	MW388663	8.260	4.884	46	1/244.2 kb
TraesCS7A02G160700	MW388664	1.163	1.444	3	1/1107.1 kb
TraesCS7B02G065700	MW388665	0.904	1.469	10	1/337.9 kb
TraesCS7D02G161800	MW388666	0.905	1.427	12	1/273.5 kb

of the SNPs were confirmed by Sanger sequencing, and all of the SNPs in M3 were consistent with those from pooled M2 population (**Table 4**).

Secondary and 3D Structure Variation of Target Proteins

Using online software, the predicted three-dimensional protein models of BSR-K1 wheat orthologs showed homology to the *Saccharomyces cerevisiae* Ski2-3-8 complex with multiple alpha helices (**Figure 4**, **Supplementary Figure 5**, and **Supplementary Table 5**). The amino acid change Ala407Thr in line E325, located in the fifth TPR region, which was a change from hydrophobic residue to hydrophilic residue and, resulted in the formation of a random coil instead of an alpha helix (**Supplementary Figure 5A** and **Supplementary Table 5**). On the contrary, the mutation Ser162Phe in line E786, hydrophilic residue to hydrophobic Phe residue, resulted in a reduced random coil that enabled more residues to form an alpha helix and less to participate in an extended strand (**Supplementary Figure 5B** and **Supplementary Table 5**). The truncation mutation found in line E724 led to the loss of the fifth TPR region and remaining residues (**Figure 4C** and **Supplementary Table 5**), which might significantly affect protein function.

DISCUSSION

Our Mutated Population Resulted in the Discovery of Multiple Mutations in One Line and the Same SNP in Multiple Lines

Mutational types and their frequency are often correlated with the mutagens and the species where they occur. Based on high-throughput data from exome capture and whole-genome sequencing, transitions generally represent over 90% of EMS treatment induced mutations (Henry et al., 2014; Krasileva et al., 2017), compared to just ∼40–50% using heavy ion beams and fast neutrons (Li G.T. et al., 2017; Ichida et al., 2019). In our study, the proportion of transition mutations observed were 91 and 98% in the genome and exon/promoter regions, respectively, which is consistent with previously reported EMS-induction results (Henry et al., 2014; Krasileva et al., 2017). While an individual mutant line can carry thousands of mutated alleles (Krasileva et al., 2017; Li G.T. et al., 2017; Hussain et al., 2018),

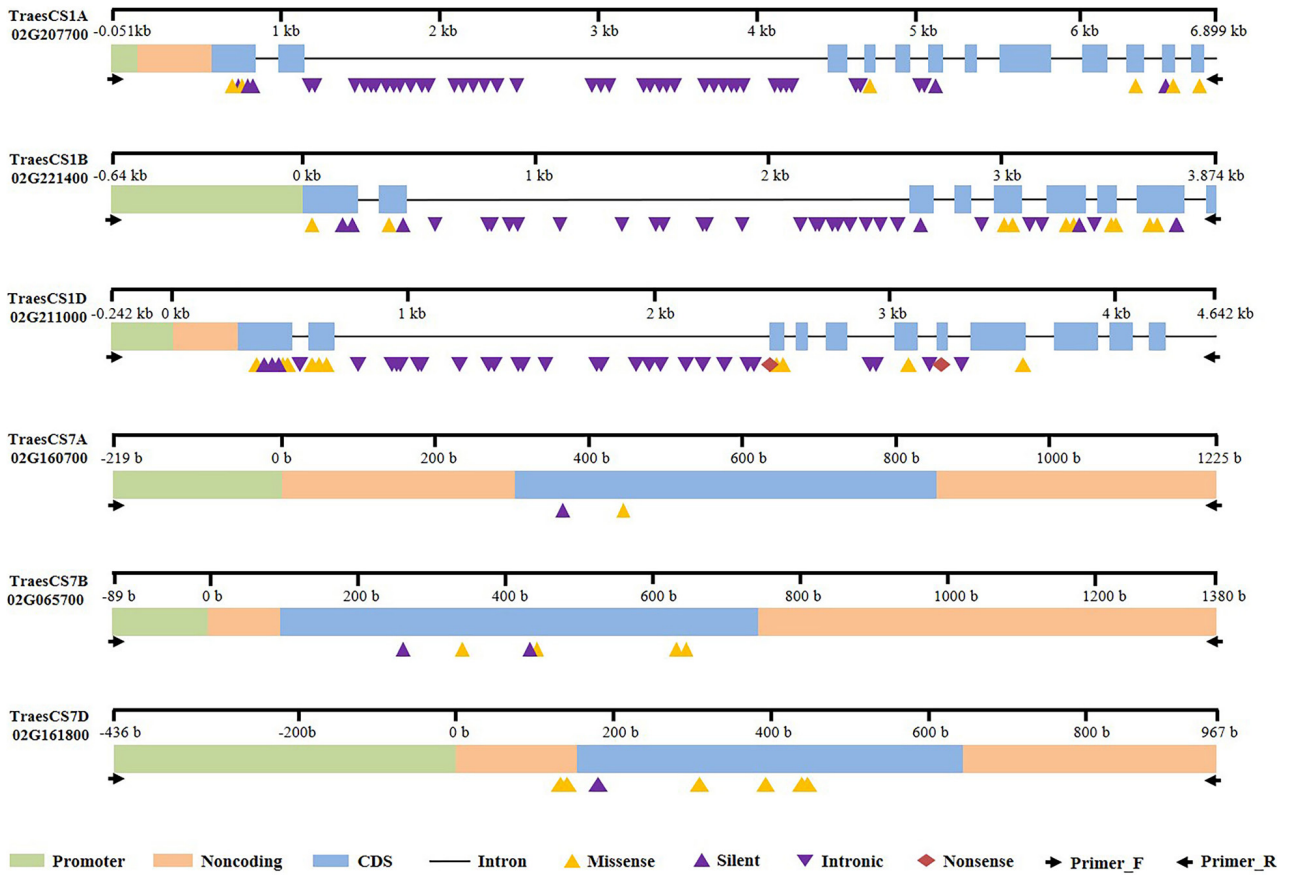

FIGURE 3 | SNP distribution in the six target fragments of *Bsr-k1* and *Bsr-d1* wheat orthologs.

we found that 12 out of 175 lines carried more than one mutated SNPs (specifically two mutations in each line). Most of them in our experiment were either located in intronic regions, represented silent mutations or had neutral effects. Only lines E48 and A53 carried mutations in the promoter region of TraesCS1B02G221400 that might affect gene function, no loss-of-function double mutant line of the target genes was directly created in the current M2 population. These results confirmed that multiple mutations existed in one individual line.

In addition, a previous study focusing on mutated tetraploid and hexaploid wheat populations, identified 1.4 out of 10 million SNPs (i.e., around 14%) in more than one individual line through exome capture (Krasileva et al., 2017). In this study, we uncovered 8 SNPs (5%) in 2–5 lines, most of which are located in intronic regions, probably due to the lesser constraint affecting intron evolution. These results demonstrated that the same SNP can be found in multi-individuals of the same mutated population even if through EMS treatment. As mentioned above, the EMS mutagen induces transition-type mutations such as G > A and C > T, transversions are thought to represent non-EMS mutations and instead result from genetic heterogeneity or sequencing errors associated with lower coverage (King et al., 2015; Krasileva et al., 2017), whereas it has been reported that transversions in different species induced by chemical mutagens including EMS were presented with lower percentage (Spencer-Lopes et al., 2018). In our study, five out of eight mutations were non-EMS type. Since we have excluded positions with less than 5× depth, it is unlikely that these mutations result from sequencing error. In addition, seeds used for EMS treatment and WT were derived from the same branch, so the probability of genetic heterogeneity is very low. Taken together, these transversions probably derived from EMS treatment.

Using Rice Blast Susceptible Genes Opens a New Window to Promote Wheat Blast Resistance Breeding Through Mutation Induction

The promoter region controls the transcription of genes through the binding of specific transcription factors. Accordingly, variations in the genomic sequence of both transcription factors and promoter might alter gene function. *WRKY76* is a transcription factor that binds to W-box elements and its overexpression results in decreased resistance to rice blast (Yokotani et al., 2013). At the same time, a SNP in the promoter region (-618) and consequent *bsr-d1* knockout leads to an increased binding affinity with the transcription factor MYBS1, which, in turn, enhances blast resistance (Li W. et al., 2017). The

TABLE 2 | SNPs identified in *Bsr-k1* wheat orthologs and their predicted impact on protein function.

Line	Region	Allele[a]	Mutation Type	Variation in Amino Acid[b]	PROVEAN Score	Prediction
TraesCS1A02G207700						
A32	5′UTR	C121T				
E1354	5′UTR	C159G	start codon gained			
A408	5′UTR	C196T				
A410	5′UTR	G254A				
E333	CDS1	C301T	Missense	P4S	0.788	Neutral
E038-9	CDS1	C342T	Silent	L17=		
E049-1	CDS1	G346A	Missense	A19T	−0.433	Neutral
E054-10	CDS1	C399T	Silent	H36=		
E439	CDS1	C483T	Silent	A65=		
E758	CDS4	C5070T	Missense	A149V	−3.825	Deleterious
E203	CDS6	G5413A	Silent	Q188=		
E325	CDS10	G6403A	Missense	A407T	−3.28	Deleterious
E1258	CDS11	C6629T	Silent	C448=		
E049-1	CDS11	G6641A	Silent	V452=		
E630	CDS12	G6807A	Missense	G472D	1.357	Neutral
TraesCS1B02G221400						
A53	promoter	G-429A				
E48	promoter	C-212T				
E046-3	promoter	C-192T				
E041-12	promoter	C-84T				
E1180	CDS1	C11T	Missense	P4I	0.108	Neutral
E833	CDS1	G87A	Silent	G29=		
E1015	CDS1	C162T	Silent	A54=		
E038-16	CDS2	G353A	Missense	G76E	−4.464	Deleterious
E607	CDS2	C438T	Silent	Y104=		
E889	CDS3	G2567A	Silent	Q117=		
E1272	CDS5	G2885A	Missense	S161R	−1.465	Neutral
E786	CDS5	C2888T	Missense	S162F	−3.81	Deleterious
E536	CDS6	G3113A	Missense	E192K	−0.464	Neutral
E1294	CDS6	G3195A	Missense	G219E	−3.792	Deleterious
E1171	CDS6	G3196A	Silent	G219=		
E1418	CDS7	G3305A	Missense	E230K	−1.289	Neutral
E410	CDS7	G3308A	Missense	A231T	−2.888	Deleterious
E148	CDS8	G3506A	Missense	D267N	−0.167	Neutral
E601	CDS8	G3522A	Missense	R272K	0.184	Neutral
E118	CDS8	C3625T	Silent	R306=		
E496	CDS8	C3625T	Silent	R306=		
TraesCS1D02G211000						
E653	5′UTR	C5T				
E1344	5′UTR	C39T				
E028-15 (II)	5′UTR	C83T				
E044-9	5′UTR	C109T				
E136	5′UTR	G203A				
A316	CDS1	C242T	Missense	A3V	−0.649	Neutral
E042-1	CDS1	C246T	Silent	P4=		
E1151	CDS1	C333T	Silent	S33=		
E316	CDS1	C342T	Silent	H36=		
E1184	CDS1	C356T	Missense	A41V	0.926	Neutral
E972	CDS1	G380A	Missense	R49K	0.625	Neutral
E1180	CDS2	G547A	Missense	A74T	−0.034	Neutral
E91	CDS2	C604T	Missense	P93S	−3.459	Deleterious
E958	CDS2	G626A	Missense	R100Q	−1.297	Neutral
E60	CDS3	C2686T	Non-sense	Q109stop	−6.915	Deleterious
E203	CDS3	G2701A	Missense	D114N	1.075	Neutral
E315	CDS3	C2740T	Missense	P127S	−5.116	Deleterious
E054-9	CDS6	G3275A	Missense	E192K	−0.597	Neutral
E724	CDS7	G3525A	Non-sense	W249stop	−16.106	Deleterious
E539	CDS8	C3849T	Missense	T327I	−1.544	Neutral

(Continued)

TABLE 2 | Continued

[a]*: Start from the initiation site of the gene.*
[b]*: "=" means Synonymous change.*

TABLE 3 | SNPs identified in *Bsr-d1* wheat orthologs and their predicted impact on protein function.

Line	Region	Allele[a]	Mutation Type	Variation in Amino Acid[b]	PROVEAN score	Prediction
TraesCS7A02G160700						
E038-14	5′UTR	G217A				
E035-7	CDS1	C412T	Silent	S16=		
E038-6	CDS1	C488T	Missense	L42F	−4	Deleterious
TraesCS7B02G065700						
E049-4	5′UTR	C38T				
A305	CDS1	C220T	Silent	D32=		
E051-2	CDS1	G314A	Missense	A64T	−1.002	Neutral
A146	CDS1	G466A	Silent	G114=		
E1300	CDS1	G470A	Missense	E116K	−0.675	Neutral
A17	CDS1	G620A	Missense	V166M	−0.445	Neutral
E035-13	CDS1	C633T	Missense	A170V	−1.283	Neutral
E053-12	3′UTR	C859T				
E023-10	3′UTR	C1202T				
E024-11	3′UTR	G1235A				
TraesCS7D02G161800						
A42	promoter	G-370A				
A196	promoter	C-361T				
A259	promoter	C-258T				
A417	promoter	G-236A				
E024-3	promoter	G-143A				
E054-8	CDS1	G135A	start codon lost	M1I	−0.584	Neutral
E040-14	CDS1	C197T	Missense	P22L	−1.62	Neutral
A277	CDS1	C246T	Silent	A38=		
E044-3	CDS1	C304T	Missense	P58S	−2.879	Deleterious
A34	CDS1	C421T	Missense	R97W	−0.816	Neutral
E149	CDS1	G491A	Missense	S119D	−1.235	Neutral
E044-2	CDS1	G497A	Missense	G122D	−3.815	Deleterious

[a]*: Start from the initiation site of the gene.*
[b]*: "=" means Synonymous change.*

bsr-d1 wheat orthologs reported here maintained the C2H2-type zinc finger functional domain, and we report mutations in the promoter and coding regions of the gene that have the potential to enhance *MoT* resistance.

TABLE 4 | Validation of SNPs in M3 generation.

Gene	Mutant	Allele	Total Number of tested individuals	Mutants	Non-mutants
TraesCS1A02G207700	E758	C5070T	35	34	1
	E325	G6403A	30	30	0
TraesCS1B02G221400	E038-16	G353A	18	15	3
	E786	C2888T	27	26	1
	E1294	G3195A	27	27	0
	E410	G3308A	32	32	0
TraesCS1D02G211000	E91	C604T	30	26	4
	E60	C2686T	16	16	0
	E315	C2740T	34	33	1
	E724	G3525A	17	14	3
TraesCS7A02G160700	E038-14	G217A	37	36	1
	E038-6	C488T	30	30	0
TraesCS7D02G161800	A42	G-370A	24	23	1
	A196	C-361T	24	8	16
	A259	C-258T	24	17	7
	E024-3	G-143A	33	32	1
	E044-3	C304T	31	18	13

A majority of disease resistant genes encode for conservative proteins containing a nucleotide binding site with leucine rich repeats. In contrast, while TPR mediate an alternative immune response mechanism in plants, the loss-of-function BSR-K1 TPR protein is unable to bind to the mRNA of the *OsPAL* gene family, resulting in blast resistance in rice (Zhou et al., 2018). We found that the BSR-K1 TPR protein is highly conserved in wheat with over 80% sequence identity. Moreover, we identified five tandem repeats, multiple truncation and missense mutations with deleterious effects in the three sub-genomes that lead to the destruction of the TPR domain in a similar fashion to what is observed in rice. The susceptible powdery mildew gene *Mlo* found in barley is conserved across plant species (Kusch et al., 2016), and its loss-of-function mutation in wheat and other species leads to enhanced powdery mildew disease resistance (Acevedo-Garcia et al., 2017). The mutants identified in this study might also provide enhanced immunity and resistance to wheat blast. Although the resistance level needs to be validated under infected-field conditions, these alleles have not been previously reported in the literature, and might represent a valuable new resource for wheat blast (or even other fungi) disease resistance breeding.

As an hexaploid species, a mutation on one of the three sub-genomes may or may not lead to phenotypic variation in wheat. Hence, it is necessary to pyramide the three homologs before evaluating resistance, and it would be particularly beneficial to pyramide the deleterious mutations reported in the five genes mentioned above in order to evaluate their interactions against *MoT* and other fungal diseases.

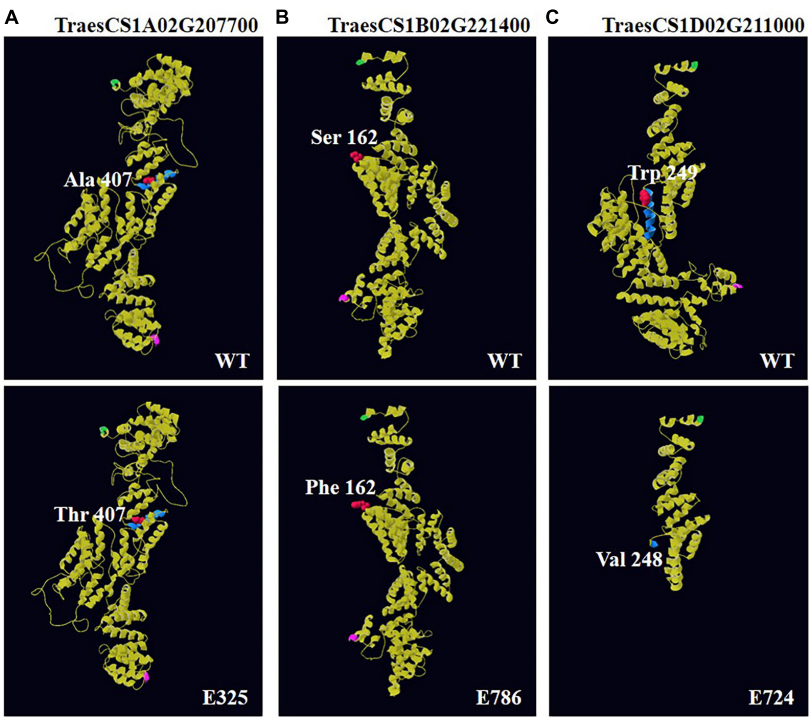

FIGURE 4 | Three-dimensional (3D) models of BSR-K1 wheat orthologs and mutants. The models were constructed using template 4buj.2.B, a *S. cerevisiae* Ski2-3-8 complex. **(A)** 3D structure of TraesCS1A02G207700 and its mutant E325; **(B)** 3D structure of TraesCS1B02G221400 and its mutant E786; **(C)** 3D structure of TraesCS1D02G211000 and its mutant E724. The N-terminal is highlighted in green, the C-terminal in pink, and the residue immediately before and after each mutation is shown in red, its secondary structure in blue.

CONCLUSION

We obtained six wheat orthologs of two rice blast susceptible genes through homologous gene comparison and identified loss-of-function mutations in these genes in a M2 population. We discovered that 13 mutant alleles have deleterious effects and might enhance wheat blast resistance. Our research provides a new strategy and novel gene resources to tackle disease resistant wheat breeding.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

HG and LL designed the experiments. HG and QD analyzed the data, prepared all tables and figures, and wrote the manuscript with input from all co-authors. HG, QD, YX, HX, LZ, JG, SZ, XS, and TI performed the experiments. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by IAEA CRP project (23087), the National Key Research and Development Project of China (2016YFD0102101), and China Agriculture Research System (Grant No. CARS-03).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.623419/full#supplementary-material

Supplementary Figure 1 | Diagram showing the sample pooling procedure on a 96-well-plate. The 12 samples corresponding to the gray line C were pooled into a single line-pooling-sample, while the eight samples in the blue row seven were pooled into a separate row-pooling-sample. The green sample detected in both line- and row-pooling-samples was considered a positive mutant, whereas the pink sample only detected in the row-pooling-samples was considered a negative mutant.

Supplementary Figure 2 | Structure of orthologous genes of **(A)** *Bsr-k1* (located on chromosomes 1A, 1B, and 1C) and **(B)** *Bsr-d1* (located on chromosomes 7A, 7B, and 7D).

Supplementary Figure 3 | Verification of the amplicons of gene TraesCS1A02G207700 by agarose gel electrophoresis. Forty-eight samples (E1–H12) from one of the 96-well pooling plates. M, DNA marker with fragment size of 15,000, 8,000, 6,000, 4,000, 3,000, 2,000, 1,000, and 500 bp.

Supplementary Figure 4 | An example of amplification using specific primers in M3 mutant lines of E325, E786, and E60 by agarose gel electrophoresis. E325-1, E325-2 mean the different M3 individuals of mutant E325, and so on down the gel.

Supplementary Figure 5 | Secondary structure of BSR-K1 wheat orthologs and mutants. **(A)** WT and mutant E325 of TraesCS1A02G207700; **(B)** WT and mutant E786 of TraesCS1B02G221400; **(C)** WT and mutant E724 of TraesCS1D02G211000. The mutant residue is highlighted by black triangle, and the variation secondary structure is highlighted in red rectangle.

REFERENCES

Acanda, Y., Martinez, O., Prado, M. J., Gonzalez, M. V., and Rey, M. (2014). EMS mutagenesis and qPCR-HRM prescreening for point mutations in an embryogenic cell suspension of grapevine. *Plant Cell Rep.* 33, 471–481. doi: 10.1007/s00299-013-1547-6

Acevedo-Garcia, J., Spencer, D., Thieron, H., Reinstadler, A., Hammond-Kosack, K., Phillips, A. L., et al. (2017). mlo-based powdery mildew resistance in hexaploid bread wheat generated by a non-transgenic TILLING approach. *Plant Biotechnol. J.* 15, 367–378. doi: 10.1111/pbi.12631

Ahloowalia, B. S., and Maluszynski, M. (2001). Induced mutations – A new paradigm in plant breeding. *Euphytica* 118, 167–173.

Anh, V. L., Anh, N. T., Tagle, A. G., Vy, T. T., Inoue, Y., Takumi, S., et al. (2015). Rmg8, a New Gene for Resistance to Triticum Isolates of Pyricularia oryzae in Hexaploid Wheat. *Phytopathology* 105, 1568–1572. doi: 10.1094/phyto-02-15-0034-r

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., et al. (2014). SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* 42, W252–W258.

Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., and Schwede, T. (2009). Protein structure homology modeling using SWISS-MODEL workspace. *Nat. Prot.* 4, 1–13. doi: 10.1038/nprot.2008.197

Campbell, J., Zhang, H., Giroux, M. J., Feiz, L., Jin, Y., Wang, M., et al. (2012). A mutagenesis-derived broad-spectrum disease resistance locus in wheat. *Theor. Appl. Genet.* 125, 391–404. doi: 10.1007/s00122-012-1841-7

Ceresini, P. C., Castroagudin, V. L., Rodrigues, F. A., Rios, J. A., Eduardo Aucique-Perez, C., Moreira, S. I., et al. (2018). Wheat Blast: Past, Present, and Future. *Annu. Rev. Phytopathol.* 56, 427–456.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890.

Cruppe, G., Cruz, C. D., Peterson, G., Pedley, K., Asif, M., Fritz, A., et al. (2019). Novel Sources of Wheat Head Blast Resistance in Modern Breeding Lines and Wheat Wild Relatives. *Plant Dis.* 104, 35–43. doi: 10.1094/pdis-05-19-0985-re

Cruz, C. D., Magarey, R. D., Christie, D. N., Fowler, G. A., Fernandes, J. M., Bockus, W. W., et al. (2016a). Climate Suitability for Magnaporthe oryzae Triticum Pathotype in the United States. *Plant Dis.* 100, 1979–1987. doi: 10.1094/pdis-09-15-1006-re

Cruz, C. D., Peterson, G. L., Bockus, W. W., Kankanala, P., Dubcovsky, J., Jordan, K. W., et al. (2016b). The 2NS Translocation from Aegilops ventricosa Confers Resistance to the Triticum Pathotype of Magnaporthe oryza. *Crop Sci.* 56, 990–1000. doi: 10.2135/cropsci2015.07.0410

Dong, C., Vincent, K., and Sharp, P. (2009). Simultaneous mutation detection of three homoeologous genes in wheat by High Resolution Melting analysis and Mutation Surveyor. *BMC Plant Biol.* 9:143. doi: 10.1186/1471-2229-9-143

Du, Y., Luo, S., Li, X., Yang, J., Cui, T., Li, W., et al. (2017). Identification of Substitutions and Small Insertion-Deletions Induced by Carbon-Ion Beam Irradiation in Arabidopsis thaliana. *Front. Plant Sci.* 8:1851. doi: 10.3389/fpls.2017.01851

Fukuoka, S., Saka, N., Koga, H., Ono, K., Shimizu, T., Ebana, K., et al. (2009). Loss of Function of a Proline-Containing Protein Confers Durable Disease Resistance in Rice. *Science* 325, 998–1001. doi: 10.1126/science.1175550

Gruner, K., Esser, T., Acevedo-Garcia, J., Freh, M., Habig, M., Strugala, R., et al. (2020). Evidence for Allele-Specific Levels of Enhanced Susceptibility of WheatmloMutants to the Hemibiotrophic Fungal PathogenMagnaporthe oryzaepv.Triticum. *Genes* 11:517. doi: 10.3390/genes11050517

Guo, H., Liu, Y., Li, X., Yan, Z., Xie, Y., Xiong, H., et al. (2017). Novel mutant alleles of the starch synthesis gene TaSSIVb-D result in the reduction of starch granule number per chloroplast in wheat. *BMC Genom.* 18:358. doi: 10.1186/s12864-017-3724-4

Guo, H., Xiong, H., Xie, Y., Zhao, L., Gu, J., Zhao, S., et al. (2019). Functional mutation allele mining of plant architecture and yield-related agronomic traits and characterization of their effects in wheat. *BMC Genet.* 20:102. doi: 10.1186/s12863-019-0804-2

Hao, Y., Rasheed, A., Zhu, Z., Wulff, B. B. H., and He, Z. (2020). Harnessing Wheat Fhb1 for Fusarium Resistance. *Trends Plant Sci.* 25, 1–3. doi: 10.1016/j.tplants.2019.10.006

Hazard, B., Zhang, X., Colasuonno, P., Uauy, C., Beckles, D. M., and Dubcovsky, J. (2012). Induced Mutations in the Starch Branching Enzyme II (SBEII) Genes Increase Amylose and Resistant Starch Content in Durum Wheat. *Crop Sci.* 52, 1754–1766.

Henry, I. M., Nagalakshmi, U., Lieberman, M. C., Ngo, K. J., Krasileva, K. V., Vasquez-Gross, H., et al. (2014). Efficient Genome-Wide Detection and Cataloging of EMS-Induced Mutations Using Exome Capture and Next-Generation Sequencing. *Plant Cell* 26, 1382–1397. doi: 10.1105/tpc.113.121590

Hussain, M., Iqbal, M. A., Till, B. J., and Rahman, M. U. (2018). Identification of induced mutations in hexaploid wheat genome using exome capture assay. *PLoS One* 13:e0201918. doi: 10.1371/journal.pone.0201918

Ichida, H., Morita, R., Shirakawa, Y., Hayashi, Y., and Abe, T. (2019). Targeted exome sequencing of unselected heavy-ion beam-irradiated populations reveals less-biased mutation characteristics in the rice genome. *Plant J.* 98, 301–314. doi: 10.1111/tpj.14213

Igarashi, S., Utimada, C., Igarashi, L., Kazuma, A., and Lopes, R. (1986). *Pyricularia* in wheat. I. Occurrence of *Pyricularia* sp. in Parana State. *Fitopatol. Bras.* 11, 351–352.

Inoue, Y., Trinh, V. P. T., Tani, D., and Tosa, Y. (2020). Suppression of wheat blast resistance by an effector of Pyricularia oryzae is counteracted by a host specificity resistance gene in wheat. *New Phytol.* 229, 488–500. doi: 10.1111/nph.16894

Inoue, Y., Vy, T. T. P., Yoshida, K., Asano, H., Mitsuoka, C., Asuke, S., et al. (2017). Evolution of the Wheat Blast Fungus Through Functional Losses in a Host Specificity Determinant. *Science* 357, 80–83. doi: 10.1126/science.aam9654

Islam, M. T., Croll, D., Gladieux, P., Soanes, D. M., Persoons, A., Bhattacharjee, P., et al. (2016). Emergence of wheat blast in Bangladesh was caused by a South American lineage of Magnaporthe oryzae. *BMC Biol.* 14:84. doi: 10.1186/s12915-016-0309-7

Islam, M. T., Gupta, D. R., Hossain, A., Roy, K. K., He, X., Kabir, M. R., et al. (2020). Wheat blast: a new threat to food security. *Phytopathol. Res.* 2:28.

Juliana, P., He, X., Kabir, M. R., Roy, K. K., Anwar, M. B., Marza, F., et al. (2020). Genome-wide association mapping for wheat blast resistance in CIMMYT's international screening nurseries evaluated in Bolivia and Bangladesh. *Sci. Rep.* 10:15972.

Kim, H., Yoon, M.-R., Chun, A., and Tai, T. H. (2018). Identification of novel mutations in the rice starch branching enzyme I gene via TILLING by sequencing. *Euphytica* 214:94.

Kim, S.-I., and Tai, T. H. (2014). Identification of novel rice low phytic acid mutations via TILLING by sequencing. *Mole. Breed.* 34, 1717–1729. doi: 10.1007/s11032-014-0127-y

King, R., Bird, N., Ramirez-Gonzalez, R., Coghill, J. A., Patil, A., Hassani-Pak, K., et al. (2015). Mutation Scanning in Wheat by Exon Capture and Next-Generation Sequencing. *PLoS One* 10:e0137549. doi: 10.1371/journal.pone.0137549

Krasileva, K. V., Vasquez-Gross, H. A., Howell, T., Bailey, P., Paraiso, F., Clissold, L., et al. (2017). Uncovering hidden variation in polyploid wheat. *Proc. Natl. Acad. Sci. U. S. A.* 114, E913–E921.

Kusch, S., Pesch, L., and Panstruga, R. (2016). Comprehensive Phylogenetic Analysis Sheds Light on the Diversity and Origin of the MLO Family of Integral Membrane Proteins. *Genome Biol. Evolut.* 8, 878–895. doi: 10.1093/gbe/evw036

Li, G., Zhou, J., Jia, H., Gao, Z., Fan, M., Luo, Y., et al. (2019). Mutation of a histidine-rich calcium-binding-protein gene in wheat confers resistance to Fusarium head blight. *Nat Genet.* 51, 1106–1112. doi: 10.1038/s41588-019-0426-7

Li, G. T., Jain, R., Chern, M., Pham, N. T., Martin, J. A., Wei, T., et al. (2017). The Sequences of 1504 Mutants in the Model Rice Variety Kitaake Facilitate Rapid Functional Genomic Studies. *Plant Cell* 29, 1218–1231. doi: 10.1105/tpc.17.00154

Li, W., Zhu, Z., Chern, M., Yin, J., Yang, C., Ran, L., et al. (2017). A Natural Allele of a Transcription Factor in Rice Confers Broad-Spectrum Blast Resistance. *Cell* 170, 114–126. doi: 10.1016/j.cell.2017.06.008

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

Lu, H. P., Luo, T., Fu, H. W., Wang, L., Tan, Y. Y., Huang, J. Z., et al. (2018). Resistance of rice to insect pests mediated by suppression of serotonin biosynthesis. *Nat. Plants* 4, 338–344. doi: 10.1038/s41477-018-0152-7

McCallum, C. M., Comai, L., Greene, E. A., and Henikoff, S. (2000). Targeting Induced Local Lesions IN Genomes (TILLING) for Plant Functional Genomics. *Plant Physiol.* 123, 439–442. doi: 10.1104/pp.123.2.439

Slade, A. J., Fuerstenberg, S. I., Loeffler, D., Steine, M. N., and Facciotti, D. (2005). A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. *Nat. Biotechnol.* 23, 75–81. doi: 10.1038/nbt1043

Slade, A. J., Mcguire, C., Loeffler, D., Mullenberg, J., Skinner, W., Fazio, G., et al. (2012). Development of high amylose wheat through TILLING. *BMC Plant Biol.* 12:69. doi: 10.1186/1471-2229-12-69

Spencer-Lopes, M. M., Forster, B. P., and Jankuloski, L. (eds) (2018). *Manual on Mutation Breeding - Third edition*. Rome: Food and Agriculture Organization of the United Nations, 58.

Srivastava, D., Shamim, M., Kumar, M., Mishra, A., Pandey, P., Kumar, D., et al. (2017). Current Status of Conventional and Molecular Interventions for Blast Resistance in Rice. *Rice Sci.* 24, 299–321. doi: 10.1016/j.rsci.2017.08.001

Tembo, B., Mulenga, R. M., Sichilima, S., M'siska, K. K., Mwale, M., Chikoti, P. C., et al. (2020). Detection and characterization of fungus (Magnaporthe oryzae pathotype Triticum) causing wheat blast disease on rain-fed grown wheat (Triticum aestivum L.) in Zambia. *PLoS One* 15:e0238724. doi: 10.1371/journal.pone.0238724

Till, B. J., Zerr, T., Comai, L., and Henikoff, S. (2006). A protocol for TILLING and Ecotilling in plants and animals. *Nat. Prot.* 1, 2465–2477. doi: 10.1038/nprot.2006.329

Tsai, H., Howell, T., Nitcher, R., Missirian, V., Watson, B., Ngo, K. J., et al. (2011). Discovery of Rare Mutations in Populations: TILLING by Sequencing. *Plant Physiol.* 156, 1257–1268. doi: 10.1104/pp.110.169748

Wang, S. Z., Asuke, S., Vy, T. T. P., Inoue, Y., Chuma, I., Win, J., et al. (2018). A New Resistance Gene in Combination with Rmg8 Confers Strong Resistance Against Triticum Isolates of Pyricularia oryzae in a Common Wheat Landrace. *Phytopathology* 108, 1299–1306. doi: 10.1094/phyto-12-17-0400-r

Wang, Y. P., Cheng, X., Shan, Q. W., Zhang, Y., Liu, J. X., Gao, C. X., et al. (2014). Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. *Nat. Biotechnol.* 32, 947–951. doi: 10.1038/nbt.2969

Xiong, H., Guo, H., Xie, Y., Zhao, L., Gu, J., Zhao, S., et al. (2017). RNAseq analysis reveals pathways and candidate genes associated with salinity tolerance in a spaceflight-induced wheat mutant. *Scient. Rep.* 7:2731.

Yokotani, N., Sato, Y., Tanabe, S., Chujo, T., Shimizu, T., Okada, K., et al. (2013). WRKY76 is a rice transcriptional repressor playing opposite roles in blast disease resistance and cold stress tolerance. *J. Exp. Bot.* 64, 5085–5097. doi: 10.1093/jxb/ert298

Zhao, H., Wang, X., Jia, Y., Minkenberg, B., Wheatley, M., Fan, J., et al. (2018). The rice blast resistance gene Ptr encodes an atypical protein required for broad-spectrum disease resistance. *Nat. Commun.* 9:2039.

Zhou, X., Liao, H., Chern, M., Yin, J., Chen, Y., Wang, J., et al. (2018). Loss of function of a rice TPR-domain RNA-binding protein confers broad-spectrum disease resistance. *Proc. Natl. Acad. Sci. U. S. A.* 115, 3174–3179. doi: 10.1073/pnas.1705927115

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

REVIEW
published: 22 June 2021
doi: 10.3389/fgene.2021.675686

Mechanisms of Genome Maintenance in Plants: Playing It Safe With Breaks and Bumps

Aamir Raina[1,2], Parmeshwar K. Sahu[3], Rafiul Amin Laskar[4], Nitika Rajora[5], Richa Sao[3], Samiullah Khan[1] and Rais A. Ganai[6]**

[1] Mutation Breeding Laboratory, Department of Botany, Aligarh Muslim University, Aligarh, India, [2] Botany Section, Women's College, Aligarh Muslim University, Aligarh, India, [3] Department of Genetics and Plant Breeding, Indira Gandhi Agriculture University, Raipur, India, [4] Department of Botany, Bahona College, Jorhat, India, [5] National Agri-Food Biotechnology Institute, Mohali, India, [6] Watson-Crick Centre for Molecular Medicine, Islamic University of Science and Technology, Awantipora, India

OPEN ACCESS

Edited by:
Fatemeh Maghuly,
University of Natural Resources and Life Sciences, Vienna, Austria

Reviewed by:
Sravankumar Thula,
Central European Institute of Technology (CEITEC), Czechia
Vijee Mohan,
University of North Texas, United States
Cécile Raynaud,
UMR 9213 Institut des Sciences des Plantes de Paris Saclay (IPS2), France

***Correspondence:**
Aamir Raina
aamir854@gmail.com
Rais A. Ganai
rais.ganai@islamicuniversity.edu.in

Specialty section:
This article was submitted to Systems Biology Archive, a section of the journal Frontiers in Genetics

Received: 03 March 2021
Accepted: 04 May 2021
Published: 22 June 2021

Citation:
Raina A, Sahu PK, Laskar RA, Rajora N, Sao R, Khan S and Ganai RA (2021) Mechanisms of Genome Maintenance in Plants: Playing It Safe With Breaks and Bumps. Front. Genet. 12:675686. doi: 10.3389/fgene.2021.675686

Maintenance of genomic integrity is critical for the perpetuation of all forms of life including humans. Living organisms are constantly exposed to stress from internal metabolic processes and external environmental sources causing damage to the DNA, thereby promoting genomic instability. To counter the deleterious effects of genomic instability, organisms have evolved general and specific DNA damage repair (DDR) pathways that act either independently or mutually to repair the DNA damage. The mechanisms by which various DNA repair pathways are activated have been fairly investigated in model organisms including bacteria, fungi, and mammals; however, very little is known regarding how plants sense and repair DNA damage. Plants being sessile are innately exposed to a wide range of DNA-damaging agents both from biotic and abiotic sources such as ultraviolet rays or metabolic by-products. To escape their harmful effects, plants also harbor highly conserved DDR pathways that share several components with the DDR machinery of other organisms. Maintenance of genomic integrity is key for plant survival due to lack of reserve germline as the derivation of the new plant occurs from the meristem. Untowardly, the accumulation of mutations in the meristem will result in a wide range of genetic abnormalities in new plants affecting plant growth development and crop yield. In this review, we will discuss various DNA repair pathways in plants and describe how the deficiency of each repair pathway affects plant growth and development.

Keywords: DNA damage, DNA repair pathways, mutations, genome integrity, DNA replication

INTRODUCTION

DNA replication is a fundamental process required for all organisms to divide and grow. It encompasses the precise duplication of DNA into two identical copies for the preservation of genetic information (Burgers and Kunkel, 2017). DNA is constantly subjected to numerous diverse kinds of insults that alter its sequence and its chemical nature, affecting the conservation of this information (Carusillo and Mussolino, 2020). The primary source of this alteration is the occasional incorporation of errors during the duplication of DNA by enzymes called DNA polymerases (Ganai and Johansson, 2016). These sporadically incorporated incorrect nucleotides

in the newly synthesized DNA occasionally escape the proofreading by the exonuclease site of the DNA polymerases, thereby generating errors (Joyce, 1997). These errors during the process of cell division can have severe consequences on the fitness and viability of an offspring. Remarkably, the errors introduced by DNA polymerase are limited because of the high selectivity by the snugly fit active site of these enzymes and the accompanying ability to excise the incorrect nucleotides (Hogg et al., 2014). In addition to the replication-mediated errors, DNA is constantly exposed to endogenous and exogenous DNA-damaging agents affecting the biochemical and physical properties of the DNA (Aguilera and García-Muse, 2013; **Table 1**). The mutations arising from these errors can have a catastrophic effect leading to the initiation of genetic and age-related diseases such as cancer and aging. Interestingly, some errors that escape these repair processes can at times act as a source of genetic diversity and pave way for the selection of a better and fitter organism (Karthika et al., 2020).

In mammals, the mechanism of DNA damage response and repair has been well studied because of its role in the initiation of cancers and its applications in cancer therapeutics (Tian et al., 2015). In plants, the DNA damage response is understudied but over the last decade has attracted enormous attention largely because of its consequences on the growth and development of plants (Manova and Gruszka, 2015). Plants exposed to excess DNA damage displayed a significant reduction in productivity and crop yield. It appears that the core components of the DNA damage response pathway are similarly organized in plants. Orthologous genes exist for master DNA damage response genes such as ataxia telangiectasia mutated (ATM) (Kurzbauer et al., 2021), ATM and Rad3 related (ATR), and meiotic recombination 11 (MRE11)–radiation-sensitive 50 (RAD50)–Nijmegen breakage syndrome 1 (MRE11-RAD50-NBS1) (MRN) complex (Cools and De-Veylder, 2009). The deletion of ATM and

TABLE 1 | List of major DNA-damaging agents associated with different DNA repair pathways and their sources.

Repair pathway	DNA damages	Source
Direct reversal repair	6-4PP (dinucleoside monophosphate 6-4 photoproduct)	UV radiation
	CPD (cyclobutane pyrimidine nucleoside phosphate dimer)	UV radiation
	O^6-alkylG (O^6-alkyl-2′-deoxyguanosine-5′-monophosphate)	Alkylating agents
	Pyrimidine dimer (dipyrimidine nucleoside phosphate dimer)	UV radiation
	Thymidine dimer (dithymidine nucleoside phosphate dimer)	UV radiation
	1,*N*6-ethenoA (1,*N*6-etheno-2′-deoxyadenosine-5′-monophosphate)	Vinyl chloride metabolites Chloroethylene oxide Chloroacetaldehyde
	3,*N*4-ethenoC (3,*N*4-etheno-2′-deoxycytidine-5′-monophosphate)	Vinyl chloride metabolites Chloroethylene oxide Chloroacetaldehyde
	1,*N*2-ethenoG (1,*N*2-etheno-2′-deoxyguanosine-5′-monophosphate)	Vinyl chloride metabolites Chloroethylene oxide Chloroacetaldehyde β-Carotene oxidation products
	1 mA (1-methyl-2′-deoxyadenosine-5′-monophosphate)	Alkylating agents
	1 mG (1-methyl-2′-deoxyguanosine-5′-monophosphate)	Alkylating agents
	3 mC (3-methyl-2′-deoxycytidine-5′-monophosphate)	Alkylating agents
	3 mT (3-methyl-2′-deoxythymidine-5′-monophosphate)	Alkylating agents
Mismatch repair	Base mismatch	Polymerase mistake Spontaneous deamination Homologous recombination
	Small deletion loop	Polymerase mistake
	Large deletion loop	Polymerase mistake
	Large insertion loop	Polymerase mistake
	Small insertion loop	Polymerase mistake
Base excision repair	Base mismatch (base mismatch)	Polymerase mistake Spontaneous deamination Homologous recombination
	Single-strand break (single-stranded DNA break)	UV radiation Enzymatic cleavage Ionizing radiation
	Nick (nick)	Enzymatic cleavage
	AP site (apurinic site)	Spontaneous Unstable adducts Base excision repair
	dU (2′-deoxyuridine-5′-monophosphate)	Base excision repair Spontaneous deamination

(Continued)

Abbreviation:DDR, DNA damage repair; ATM, ataxia telangiectasia mutated; ATR, ataxia telangiectasia mutated and Rad3 related; MRE11, meiotic recombination 11; RAD50, radiation sensitive 50; NBS1, Nijmegen breakage syndrome 1; MRN, Mre11-Rad50-Nbs1; ROS, reactive oxygen species; BER, base excision repair; NER, nucleotide excision repair; MMR, mismatch repair; HRR, homologous recombination repair; NHEJ, non-homologous end-joining; ICL, interstrand cross-links; DRR, direct reversal repair; ssDNA, single-stranded DNA; dsDNA, double-stranded DNA; CPD, cyclobutane pyrimidine dimers; MTHFpolyGlu, *N*5, *N*10 methenyl-tetrahydrofolylpolyglutamate; FADH, flavin adenine dinucleotide; 6-4 PP, 6-4 Photoproducts; MGMT, O^6-methylguanine-DNA methyltransferase; 1 meA, 1-methyladenine; MMS, methyl methanesulfonate; XRCC1, X-ray repair cross-complementing protein 1; Pol δ/ε, DNA polymerase δ/ε; PCNA, proliferating cell nuclear antigen; FEN1, flap endonuclease 1; EXO1, exonuclease 1; RPA, replication protein A; 8-oxoG, 7,8-dihydro-8-oxoguanine; AP, apurinic/apyrimidinic; Pol β, DNA polymerase B; AtLIG1, *Arabidopsis* DNA ligase 1; PARP, poly(ADP-ribose) polymerase; GGR, global genomic repair; TCR, transcription-coupled repair; XPC, xeroderma pigmentosum group C; AtCEN2, *Arabidopsis thaliana* CENTRIN2; DSB, double-strand break; SSB, single-strand break; DSBR, double-strand break repair; dHJ, double Holliday junction; c-NHEJ, classical/canonical NHEJ; b-NHEJ, backup-NHEJ pathway; Alt-NHEJ, alternative NHEJ; ncRNA, non-coding RNA; aRNA, aberrant transcripts; qiRNA, quelling-induced RNA; diRNA, DSB-induced small RNA; siRNA, small interfering RNA; DDB2, DNA damage-binding protein 2; AGO1, argonaute 1; DCL4, Dicer-like-4; I-SceI, intron-encoded endonuclease from *Saccharomyces cerevisiae;* ZFNs, zinc-finger nucleases; TALENs, transcription activator–like effector nucleases; CRISPR-Cas9, clustered regularly interspaced short palindromic repeats/CRISPR-associated protein 9; tracrRNA, transactivating crRNA; sgRNA, single-guide RNA.

TABLE 1 | Continued

Repair pathway	DNA damages	Source
	Thymidine glycol (5,6-dihydroxy-5,6-dihydrothymidine-5′-monophosphate)	UV radiation Reactive oxygen species
	3 mA (3-methyl-2′-deoxyadenosine-5′-monophosphate)	Alkylating agents
	3 mG (3-methyl-2′-deoxyguanosine-5′-monophosphate)	Alkylating agents
	7 mA (7-methyl-2′-deoxyadenosine-5′-monophosphate)	Alkylating agents
	8-oxoG (8-oxo-2′-deoxyguanosine-5′-monophosphate)	Reactive oxygen species
	FapyA (4,6-diamino-5-formamidopyrimidine-2′-deoxynucleoside-5′-monophosphate)	Reactive oxygen species Ionizing radiation
	FapyG (2,6-diamino-4-hydroxy-5-formamidopyrimidine-2′-deoxynucleoside-5′-monophosphate)	Reactive oxygen species Ionizing radiation
	7 mG (7-methyl-2′-deoxyguanosine-5′-monophosphate)	Alkylating agents
Nucleotide excision repair	6-4PP (dinucleoside monophosphate 6-4 photoproduct)	UV radiation
	CPD (cyclobutane pyrimidine nucleoside phosphate dimer)	UV radiation
	Bulky adduct	Large polycyclic hydrocarbon
Homologous recombination repair	DNA gaps; DNA double-stranded breaks (Dsbs); DNA interstrand crosslinks	Ionizing radiation, chemical agents, ultraviolet light
Non-homolog end-joining	Partially single-stranded DNA; double-stranded breaks	Enzymatic digestion

ATR in *Arabidopsis thaliana* presented no discernible phenotype *per se*. However, these plants are sensitive to DNA damaging agents such as aphidicolin, radiations, and alkylating agents. Furthermore, similar to mammals the activation of ATR and ATM is dependent on the MRN complex because the mutants of *rad50* and *mre11* are unable to activate ATR and ATM. Moreover, *rad50* and *mre11* mutants are sterile, indicating the inability of these plants to repair DNA damages affecting their ability to reproduce by either accumulation of mutations in meristem or by an unknown essential function in meiosis during gamete formation (Amiard et al., 2010). Furthermore, *ku80* mutants exhibited increased homologous recombination when exposed to increased stress conditions (Yao et al., 2013). Likewise, increased expression of DNA Pol lambda was observed in plants treated with excess hydrogen peroxide and sodium chloride (Roy et al., 2013). Taken together, these observations indicate that DNA damage response pathways are critical for the growth and development of plants by preventing the accumulation of mutations.

Plants are constantly exposed to adverse environmental settings such as heavy metals, drought, ultraviolet (UV) light, heat, lack of nutrients, and changing temperatures. Because of the sessile and autotrophic nature of the plant life cycle, they are unable to evade and escape these stressful conditions. For instance, the autotropic trait necessitates them to harness the sunlight for the production of food at the expense of exposure to UV light, resulting in the formation of toxic cyclobutane dimers in DNA (Dany et al., 2001). The photosynthetic and metabolic processes result in significant production of metabolic byproducts including reactive oxygen species (ROS) (Tuteja et al., 2009; Li et al., 2019). Production of ROS triggers single- and double-stranded breaks (SSBs and DSBs) in the DNA either directly through destruction of bases or modifications of bases. In some crop plants, oxidative stress imbalances ROS production and consequently promotes developmental defects and growth reduction (Rybaczek et al., 2021). This results in a significant decrease in plant productivity and crop quality. However, to prevent the toxic effects of ROS, plants normally keep a balance between the generation of free radicals and their eradication through the antioxidant system formed by superoxide dismutase, catalase, and ascorbate peroxidase (Li et al., 2019; Wang et al., 2019). These enzymes are vital for limiting the cellular accumulation of ROS. For instance, the mutants of *apx1* and *cat1* exhibit increased DNA damage demonstrating that ROS production has direct effects on the stability of plant DNA (Vanderauwera et al., 2011; Hu et al., 2016). Taken together, these observations underline the importance of DNA repair pathways for the prevention and accumulation of mutations on exposure to adverse environmental conditions. In exceptional cases, the mutations accumulate at an enormous rate upon many cell divisions and generations, separating one generation from the next affecting the plant viability. For instance, 6-year-old *Crepis tectorum* seeds showed reduced germination and a wide range of developmental abnormalities in the seedlings and mature plants (Navashin and Shkvarnikov, 1933). The phenotypic effects were exacerbated when seeds were stored at elevated temperatures. The mutant phenotypes from the plant phenocopies X-ray treated cells indicating accumulation of DNA damages in these seeds (Navashin and Shkvarnikov, 1933; Bray and West, 2005). Besides, the exposure of cereals and *Arabidopsis* to severe DNA damage results in DNA duplication without the ensuing cell division producing polyploid cells. The production of polyploid cells signifies permanent differentiation of cells (Galbraith et al., 1991). However, the same phenomenon of re-replication and severe DNA damage in meristems promotes cell death to avoid

the transfer of these DNA damages to the next generation. Therefore, it appears that maintenance of genetic integrity is key to the survival of plants and for the transfer of accurate genetic information to subsequent generations. Surprisingly, despite the elevated exposure to DNA-damaging agents, it appears that the frequency of the mutation rate in plants is very low. Thus, plants must actively engage numerous genes in different DNA repair pathways to protect DNA from endogenous and exogenous stress (**Table 2**). In this review, we will summarize these complex mechanisms by which plants repair their DNA from severe exposure to biotic and abiotic stress.

DNA REPAIR PATHWAYS

The integrity of DNA is under constant assault from endogenous and exogenous DNA-damaging factors including radiations, chemical mutagens, or spontaneously arising mutations. However, it appears that regardless of these assaults on DNA, the rate of mutation is exceptionally low because of the efficacy with which these alterations are fixed. To date, several pathways are known for repairing DNA damages; however, a few general assumptions can be made about these DNA repair mechanisms. First, most DNA repair pathways require a template strand for copying information into the damaged strand. The second general feature of DNA repair is the redundancy in repairing these damages, implying that a particular DNA error can be repaired by more than one repair pathway. The redundancy increases the likelihood of DNA repair and partly guaranteeing that practically almost all errors are corrected. At least five major DNA repair pathways viz. base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), homologous recombination repair (HRR), and non-homologous end-joining (NHEJ) repair are active throughout different stages of the cell cycle, allowing the cells to repair the DNA damage (Chatterjee and Walker, 2017). Direct chemical reversal and interstrand crosslink (ICL) repair pathways may also be exploited to clear unique lesions. These repair mechanisms are important for the genetic stability of cells. In this section, we will discuss general DNA repair mechanisms by which plants repair diverse kinds of DNA insults.

Direct Reversal Repair

Direct reversal repair (DRR) removes certain DNA and RNA modifications, without excision, resynthesis, or ligation (Ahmad et al., 2015). It is an error-free repair pathway that retains the original genetic information because it does not involve the breaking of the phosphodiester backbone. To date, three major DRR mechanisms have been identified: (i) photoreactivation repair, (ii) direct DNA repair by alkyltransferase, and (iii) direct DNA repair by AlkB family dioxygenases (Yi and He, 2013).

Photoreactivation Repair

The exposure of organisms to sunlight in the blue or UV-A spectrum results in the formation of cyclobutane pyrimidine dimers (CPD) such as thymidine–thymidine dimers. However, a process known as photoreactivation significantly decreases the

TABLE 2 | List of key genes that play vital roles in different DNA repair pathways.

Repair pathway	Genes	References
Direct reversal repair	ALKBH2 alkB, alkylation repair homolog 2 (*Escherichia coli*)	Duncan et al., 2002; Yang et al., 2008; Lenz et al., 2020; Toh et al., 2020
	ALKBH3 alkB, alkylation repair homolog 3 (*E. coli*)	Duncan et al., 2002; Yang et al., 2008; Fedeles et al., 2015; Lenz et al., 2020
	MGMT *O*-6-methylguanine-DNA methyltransferase	Tano et al., 1990; Mitra and Kaina, 1993; Ibrahim Al-Obaide et al., 2021
	PHR	Husain and Sancar, 1987; Li and Sancar, 1990; Sancar, 2016
	ADA	Jeggo, 1979; Shevell and Walker, 1991; Mielecki and Grzesiuk, 2014
Mismatch repair	EXO1 Exonuclease 1	Wilson et al., 1998; Lee et al., 2002; Sertic et al., 2020
	MLH3 mutL homolog 3 (*E. coli*)	Lipkin et al., 2000; Hawken et al., 2010; Hayward et al., 2020
	PMS1 PMS1 postmeiotic segregation increased 1	Hong et al., 2010; Li et al., 2020
	POLD1 Polymerase (DNA directed), delta 1, catalytic subunit 125 kDa	Dresler et al., 1988; Tsurimoto et al., 2005; Rytkonen et al., 2006; Nichols-Vinueza et al., 2021
	POLE Polymerase (DNA directed), epsilon	Rytkonen et al., 2006; Ewing et al., 2007; León-Castillo et al., 2020
Base excision repair	APEX1 APEX nuclease (multifunctional DNA repair enzyme) 1	Demple et al., 1991; Beernink et al., 2001; Coughlin, 2019;
	APEX2 APEX nuclease (apurinic/apyrimidinic endonuclease) 2	Rual et al., 2005; Burkovics et al., 2006; Briggs et al., 2010; Mengwasser et al., 2019
	FEN1 Flap structure-specific endonuclease 1	Murray et al., 1994; Zheng et al., 2008; Lu et al., 2020
	HUS1 HUS1 checkpoint homolog (*Schizosaccharomyces pombe*)	Volkmer and Karnitz, 1999; Liu C. Y. et al., 2010; Zhou et al., 2019
	MBD4 Methyl-CpG binding domain protein 4	Hendrich and Bird, 1998; Screaton et al., 2003; Sannai et al., 2019
	MPG *N*-methylpurine-DNA glycosylase	Miao et al., 2000; Ewing et al., 2007; Ryu et al., 2020
	NEIL1 Nei endonuclease VIII–like 1 (*E. coli*)	Das et al., 2007; Sengupta et al., 2018; Saini et al., 2020
	OGG1 8-Oxoguanine DNA glycosylase	Radicella et al., 1997; Lindahl and Wood, 1999; Ewing et al., 2007; Miglani et al., 2021
	PARP1 Poly(ADP-ribose) polymerase 1	Dantzer et al., 1998; Kanno et al., 2007; Wong et al., 2009; Lavrik, 2020

(Continued)

TABLE 2 | Continued

Repair pathway	Genes	References
	PNKP Polynucleotide kinase 3 and phosphatase	Jilani et al., 1999; Karimi-Busheri et al., 1999; Kalasova et al., 2020
	RAD1 RAD1 homolog (*S. pombe*)	Parker et al., 1998; Zou and Elledge, 2003; Huangteerakul et al., 2021
Nucleotide excision repair	DDB1 Damage-specific DNA-binding protein 1	Keeney et al., 1993; Marini et al., 2006; Kim et al., 2016
	ERCC6 Excision repair cross-complementing rodent repair deficiency, complementation group 6	Selby and Sancar, 1997; Thorslund et al., 2005; Fousteri et al., 2006; Faridounnia et al., 2018; Apelt et al., 2020
	ERCC8 Excision repair cross-complementing rodent repair deficiency, complementation group 8	Henning et al., 1995; Selby and Sancar, 1997; Groisman et al., 2003; Fousteri et al., 2006; Lu et al., 2018; Moslehi et al., 2020
	MFD Mutation frequency decline	Selby et al., 1991; Oller et al., 1992; Martin et al., 2019; Leyva-Sánchez et al., 2020
Homologous recombination repair	EME1 Essential meiotic endonuclease 1 homolog 1 (*S. pombe*)	Briggs et al., 2010; Liu Y. et al., 2010; Wang et al., 2016
	FANCA Fanconi anemia, complementation group A	Kupfer et al., 1997; Bailey et al., 2010; Román-Rodríguez et al., 2019
	MRE11 Meiotic recombination 11 homolog A (*Saccharomyces cerevisiae*)	Paull and Gellert, 1998; Gatei et al., 2000; Lee and Paull, 2005; Chansel-Da Cruz et al., 2020
	RAD50	Bhaskara et al., 2007; Ghosal and Muniyappa, 2007; Chansel-Da Cruz et al., 2020; Völkening et al., 2020
Non-homologous end-joining	DCLRE1C DNA cross-link repair 1C (PSO2 homolog, *S. cerevisiae*)	Ma et al., 2002; Briggs et al., 2010; Liu Y. et al., 2010; Richter et al., 2019
	NHEJ1 Non-homologous end-joining factor 1	Ahnesorg et al., 2006; Buck et al., 2006; Esmaeilzadeh et al., 2019
	XRCC6 X-ray repair complementing defective repair in Chinese hamster cells 6	Cooper et al., 2000; Kim et al., 2008; Balinska et al., 2019
	YKU80	Ruan et al., 2005; Sabourin et al., 2007; Carballar et al., 2020

biological consequences of these UV radiations by repairing these damages. A class of enzymes called photolyases specifically binds to these CPDs and directly reverses this damage in an error-free manner. Instead of removing the DNA-damaged region, photoreactivation reverses DNA damage to its original form in an error-free manner. In early life forms, it is believed to be the first evolved DNA repair mechanism and is still preserved in diverse species such as bacteria, yeast, plants, and animals (Lucas-Lledó and Lynch, 2009). In *Escherichia coli* energy derived from blue spectrum, light is absorbed by chromophores [*N*5, *N*10 methenyl-tetrahydro folylpolyglutamate and flavin adenine dinucleotide (MTHFpolyGlu and $FADH^-$)] followed by sequential electron transfer from FADH to pyrimidine dimer. Finally, electronic rearrangement generates an unstable dimer radical that hydrolyses to yield the monomeric pyrimidines (**Figure 1**). Plants that are specialized in selectively reversing 6-4 photoproducts (6-4 PPs) or CPD, two distinct forms of photolyase enzymes such as 6-4 photolyase and class II photolyase, have been identified. These photolyases repair the lesions by binding at their respective DNA-damaged site in a light-independent manner and obtaining energy from the blue or near UV-A spectrum (Brettel and Byrdin, 2010). The photolyase genes are considered to be useful in modern agriculture to enhance the UV resistance and production of improved cultivars.

Direct DNA Repair by Alkyltransferases

Alkylating agents react with the DNA and add alkyl groups preferably at *O*- and *N*- positions of nitrogenous bases. To combat the mutagenic effects of alkylating agents, organisms employ direct repair in which alkylated bases are screened followed by direct transfer of alkyl group from the nitrogenous base to the cysteine of an enzyme called O^6-methylguanine-DNA methyltransferase (MGMT or AGT). MGMT binds in the minor groove of DNA, scans the DNA, repairs the alkylated bases, and therefore provides a quick repair for such DNA lesion. The MGMT protein, whose bacterial analog is called Ogt, specifically reverses guanine base methylation by removing methyl groups from the guanine (Pegg, 1990; Esteller et al., 2000; Ahmad et al., 2015). As each MGMT molecule can be used only once, the procedure is costly; the reaction is stoichiometric rather than catalytic (Ibrahim Al-Obaide et al., 2021). MGMTs are ubiquitous in both bacteria and higher organisms except fission yeast and plants (Pegg, 2011). The adaptive response in bacteria is a generic response to methylating agents that confers a degree of tolerance to alkylating agents by upregulating alkylation repair enzymes after prolonged exposure. The methylation of the bases cytosine and adenine by ALKBH2 and ALKBH3 is the DNA damage that cells can repair (Yang et al., 2008; Fedeles et al., 2015; Lenz et al., 2020; **Figures 2**, **3A**). To date, no homologs for MGMT have been reported in plants; however, plants have evolved a mechanism for the removal of alkylated bases, and recent research implicates BER as a substitute for MGMT activity (Manova and Gruszka, 2015).

Direct DNA Repair by the AlkB Family Dioxygenases

AlkB family dioxygenases scan the genome and have the ability to alkylation lesions by flipping the alkylated or damaged base in both single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA). In the event of oxidative dealkylation, the AlkB family dioxygenases require iron as cofactor and 2-oxoglutarate as cosubstrate for activation of dioxygen molecule for various oxidative reactions. The activated dioxygen molecule then oxidizes and removes the alkyl group from N1 adenine (1-methyladenine) or N3 cytosine (3-methylcytosine), to yield

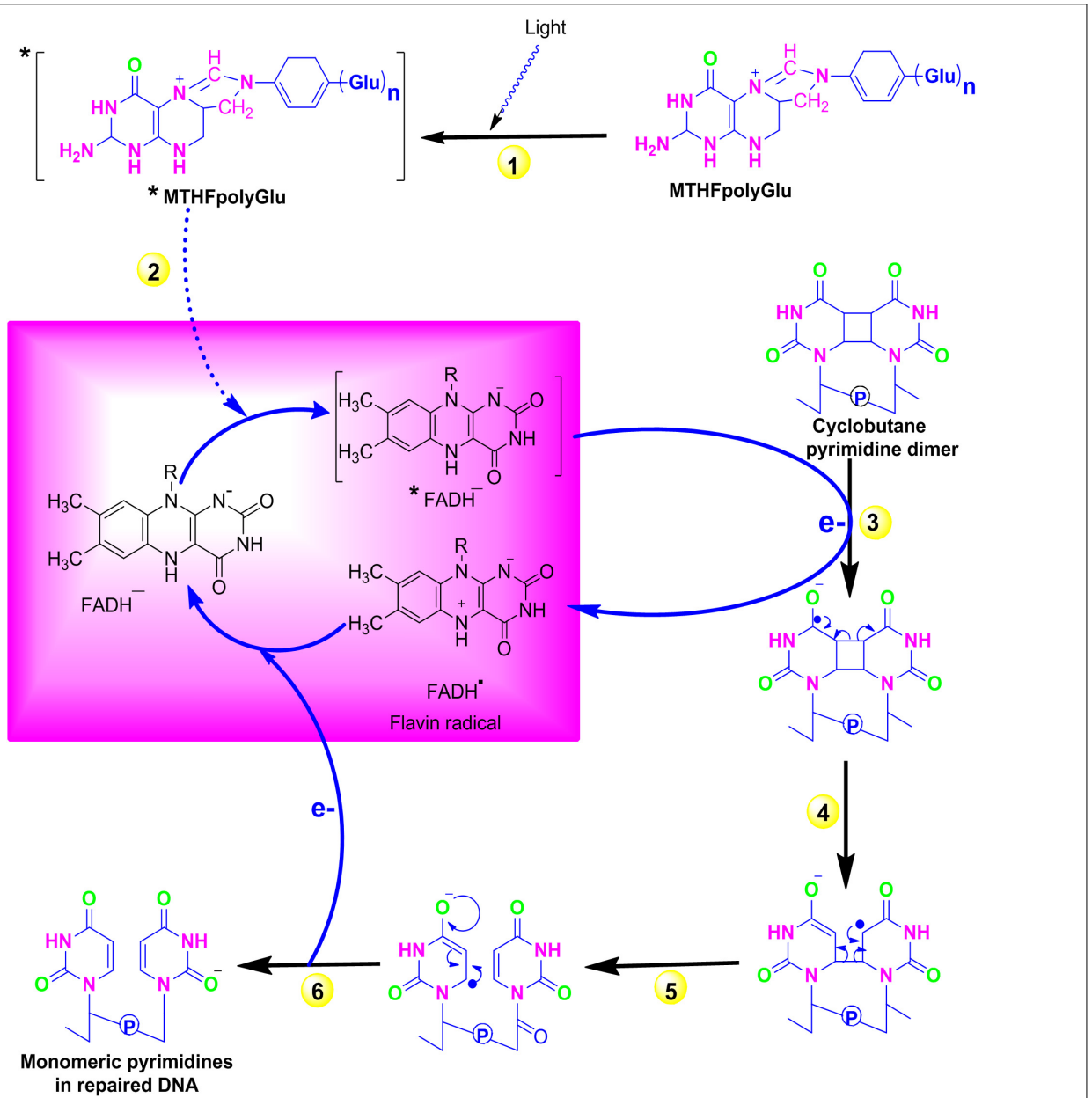

FIGURE 1 | Repair of pyrimidine dimers with photolyase. (1) A blue-light photon is absorbed by the first chromophore MTHFpolyGlu, which functions as a photoantenna. (2) The electron from the excited MTFH* is then transferred to second chromophore $FADH^-$. (3) The excited electron from $*FADH^-$ is then transferred to pyrimidine dimer and converts it into pyrimidine monomers. (4–5) Electronic rearrangement restores the monomeric pyrimidines, and (6) the electron is transferred back to the flavin radical to regenerate $FADH^-$. Source: Figure adapted and modified from Bray and West (2005). [MTFH (*N*5, *N*10 methenyl-tetrahydro folate); $FADH^-$ (flavin adenine dinucleotide)].

an unmodified base (**Figures 2**, **3B,C**). The *E. coli* AlkB protein (EcAlkB) repairs the 1-methyladenine (1-meA) and 3-methylcytosine. ALKBH2 and ALKBH3 are the mammalian homologs of *E. coli* AlkB with ALKBH2 as the main repair enzyme for 1-meA (Yi and He, 2013). Plants have also evolved an adaptive mechanism that is similar to other eukaryotes to repair alkylated nitrogenous bases. Meza et al. (2012) have reported several AlkB homologs such as AT2G22260, which revealed sequence similarity to both ALKBH2 and ALKBH3 in *A. thaliana*. The *Arabidopsis* ALKBH2 protein also displayed *in vitro* repair activities on hydroxylated methyl and ethyl groups covalently linked to DNA. Furthermore, seedlings raised from alkbh2 knockout plants developed abnormally when grown in the presence of methyl methanesulfonate (MMS).

Mismatch Repair

DNA replication–mediated errors that escape fraying by the exonuclease activity of the DNA polymerase are corrected

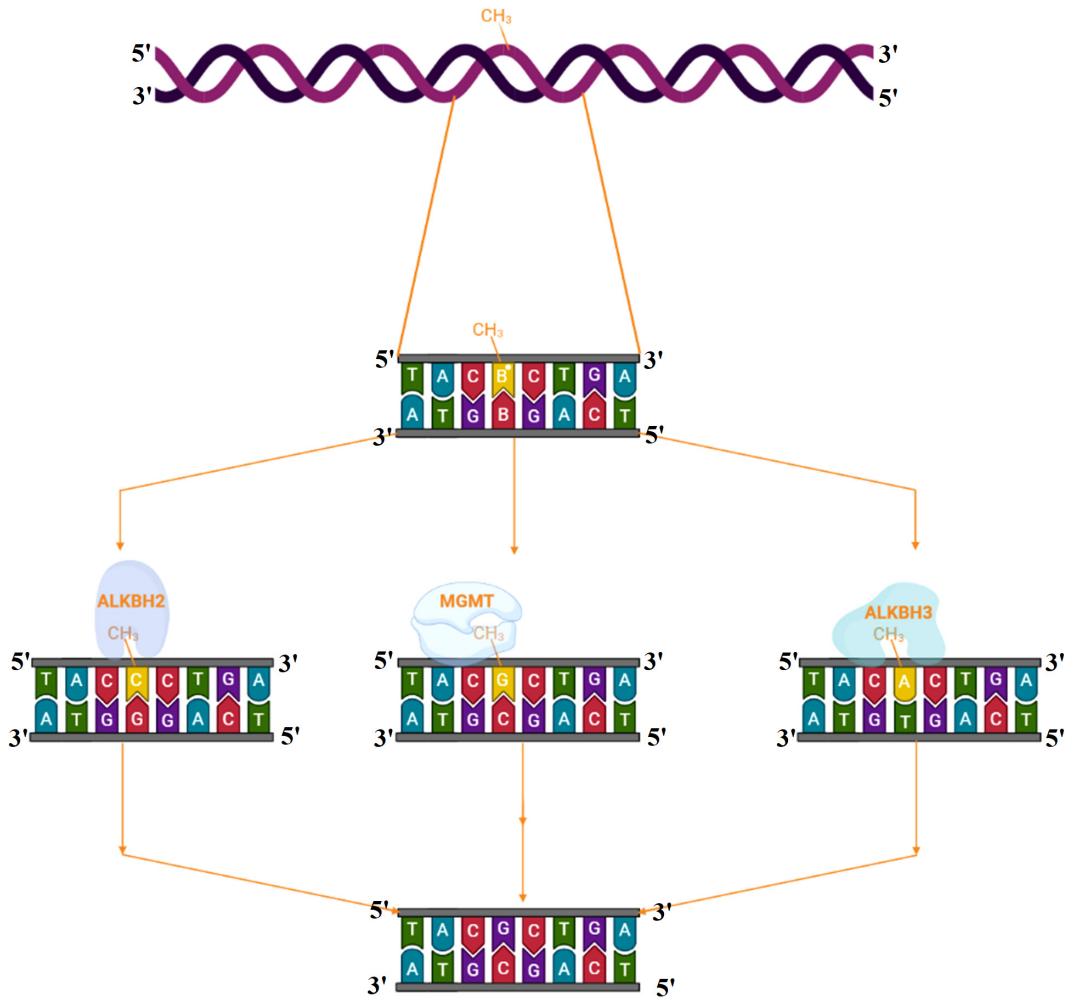

FIGURE 2 | Direct reversal of N alkylated DNA bases by alkyltransferase and dioxygenase.

via an MMR system. In the MMR system, specific enzymes excise the newly incorporated incorrect nucleotide and replace it with the correct nucleotide. The key biological function of MMR system is to correct errors introduced during DNA replication. Besides, MMR is actively involved in the repair of mispaired intermediate bases, insertion–deletion, loops, elimination of unnecessary heteroduplexes, psoralen-induced ICLs, and oxidative DNA damage (Manova and Gruszka, 2015). Overall, MMR enables the cell to preserve genome integrity by increasing the DNA replication fidelity, decreases the frequency of mutations, and regulates the dynamics of short repetitive sequences, homologous recombination, and normal meiosis (Spampinato et al., 2009). MMR is strongly conserved in all living species as an important protection mechanism for preserving genomic integrity, although certain differences within the kingdoms appear to exist. In prokaryotes, MMR is majorly carried about by the concerted action of three main enzymes mutator (Mut)S, MutL, and MutH that direct the recognition and removal of the mismatch. MutS recognizes a G-T mismatch followed by a cut near the mismatch by MutH. The region containing mismatch is removed by exonuclease I, and a new DNA segment is synthesized by DNA polymerase III to fill the gap (**Figure 4**). In eukaryotes, MMR machinery mainly consists of MutSα/β comprising of (MutS homologs) *MSH2*, *MSH3*, *MSH5*, and *MSH6*, and MutL homolog comprising of *MLH1*, *PMS1 (MLH2)*, *MLH3*, and *PMS2 (MLH4)*. Plants have an additional MSH gene called *MSH7* (Culligan and Hays, 2000). The general mechanism by which MMR functions in eukaryotes begins by the recognition of the mismatch by MutSα/β followed by the incision of the nick by MutLα. This allows for the recruitment of exonuclease 1 (EXO1), replication protein A (RPA), and Pol δ for the replacement of specific DNA segments through strand displacement synthesis. The role of MMR factors during postreplicative and recombination MMR is well known in plants. MSH2 deficiency in *Arabidopsis* prevents homologous but enhances homologous recombination and microsatellite instability in germline cells (Leonard et al., 2003; Li et al., 2006), whereas MSH7 regulates meiotic recombination, and its downregulation impairs meiotic recombination and fertility in cereals (Lloyd et al., 2006; Lario et al., 2015).

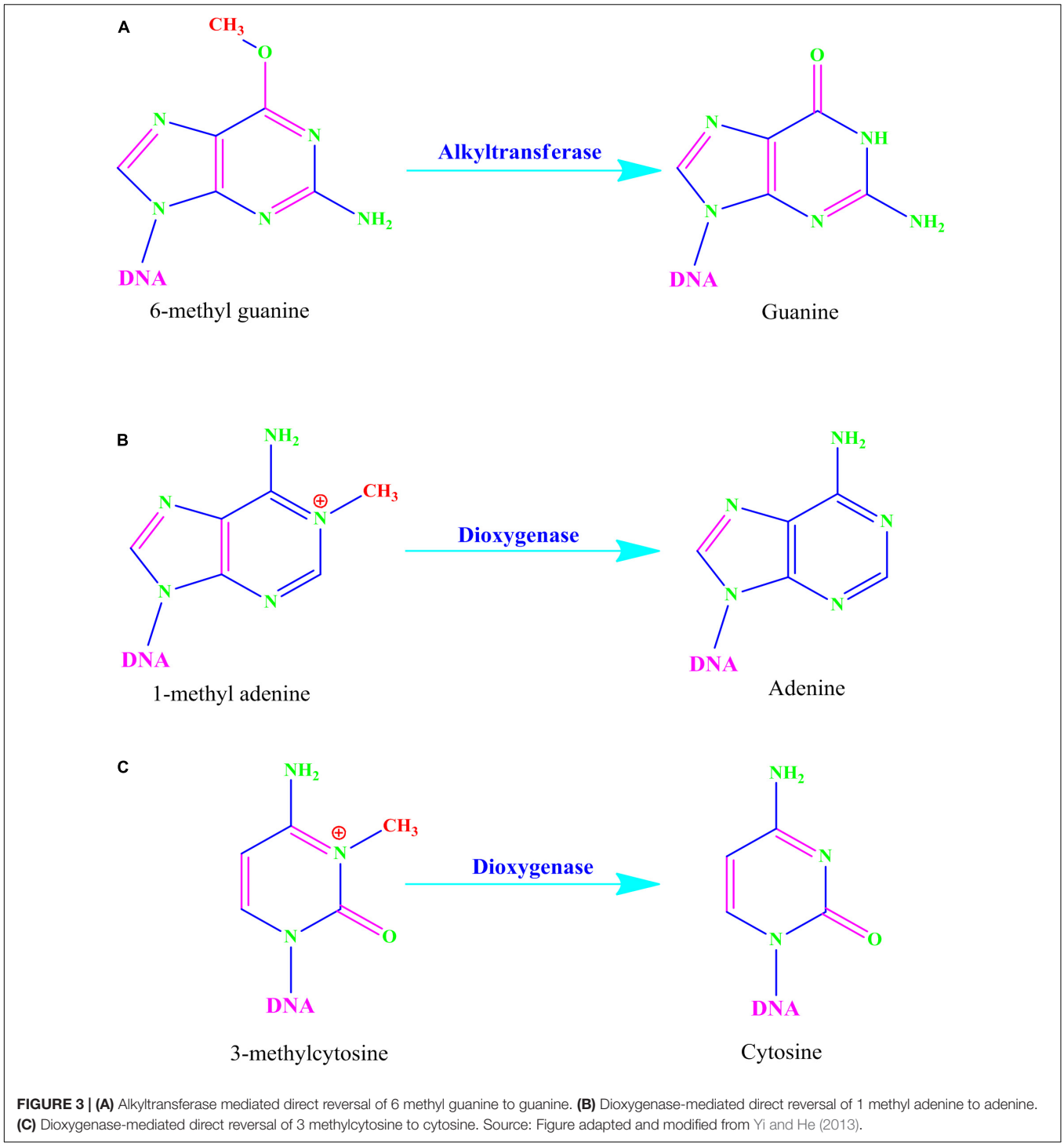

FIGURE 3 | (A) Alkyltransferase mediated direct reversal of 6 methyl guanine to guanine. **(B)** Dioxygenase-mediated direct reversal of 1 methyl adenine to adenine. **(C)** Dioxygenase-mediated direct reversal of 3 methylcytosine to cytosine. Source: Figure adapted and modified from Yi and He (2013).

Excision Repair

Unlike photoreactivation, other DNA repair pathways do not undo the DNA damage directly but instead substitute the damaged DNA with an appropriate nucleotide. Excision repair involves the removal of the damaged nucleotide by dual incision of the DNA strand containing the lesion (Waterworth et al., 2019). The incision is made on both sides of the lesion, followed by repair using the intact strand as a template. A common four-step pathway is used by these repair mechanisms that include (1) the initial detection of the DNA damage, (2) excision of the damaged nucleotide by the incision of a nick and subsequent removal of the damaged nucleotide(s), (3) filling of the gap by DNA polymerase using the exposed 3-OH as primer, and (4) finally sealing of the nick by DNA ligase. The mechanisms

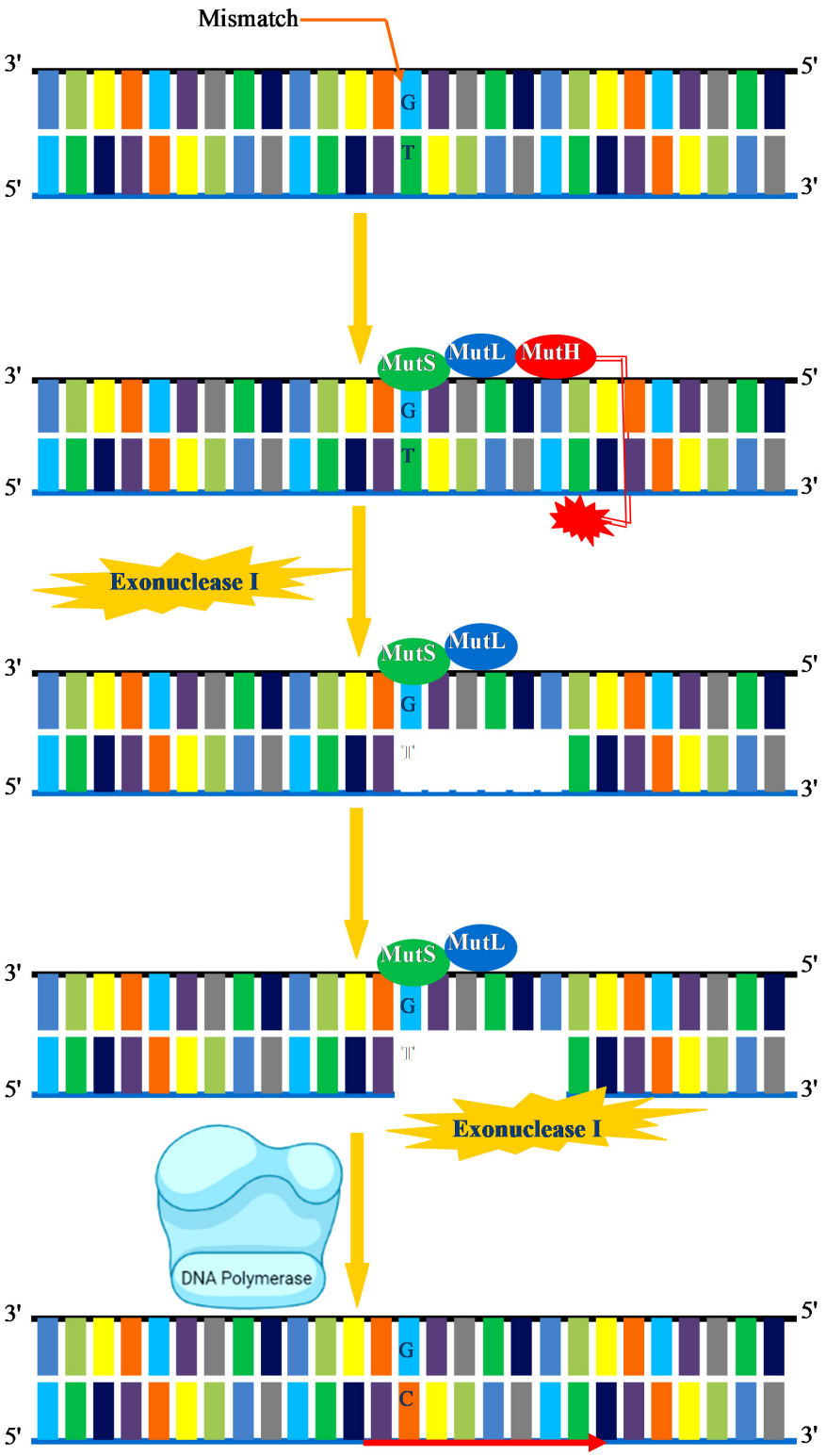

FIGURE 4 | Mismatch repair. A G-T mismatch is recognized by MutS in association with MutL. MutH cleaves in the vicinity of mismatch. Exonuclease I initiates removal of DNA segment containing the incorrect base DNA. Exonuclease I completes the removal of damaged DNA. DNA polymerase III then synthesizes the new DNA and fills the gap.

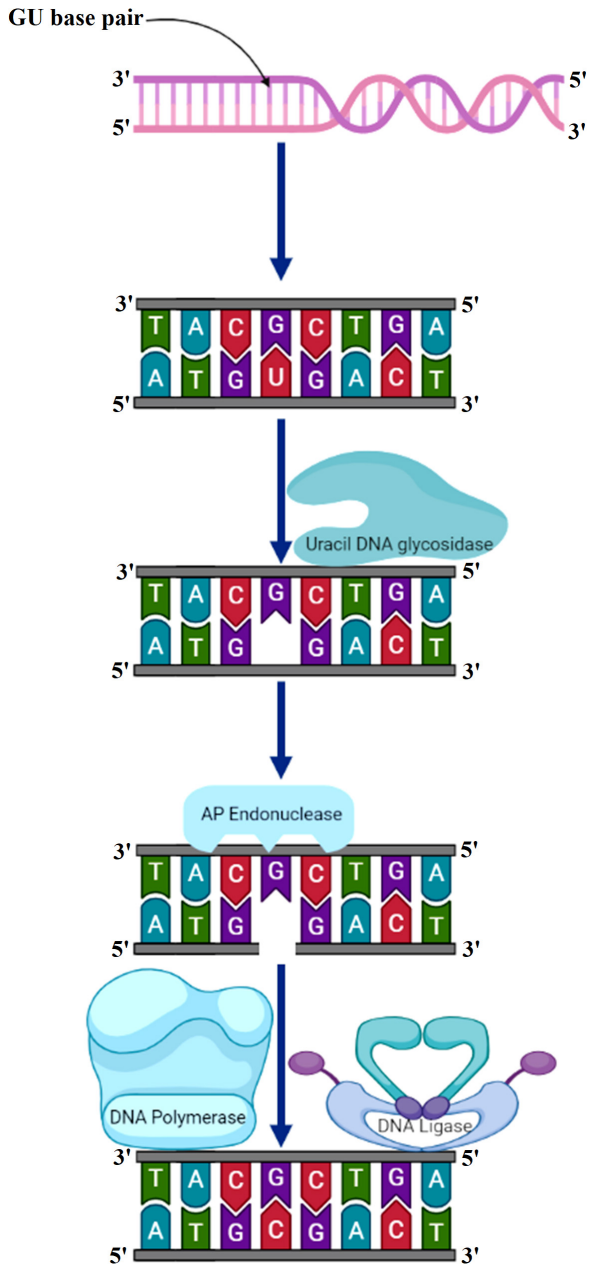

FIGURE 5 | Uracil bases in DNA, formed by the deamination of cytosine, are excised and replaced by cytosine by the combined action of uracil DNA glycosylase, AP endonuclease, DNA polymerase, and DNA ligase. AP, apurinic/apyrimidinic.

that determine the distance of the nick from the damage and the subsequent removal of the incorrect nucleotide permit the classification of this type of repair into two types, BER and NER.

Base Excision Repair

The primary function of BER is to clear the genome of minute non-helix-distorting base lesions (Wallace, 2014). Bulky helix-distorting lesions are repaired by the associated NER pathway. BER acts on a variety of lesions including apurinic sites [apurinic/apyrimidinic (AP sites)], damaged and modified bases (Manova and Gruszka, 2015; Waterworth et al., 2019). Mechanistically in base-excision repair, the DNA glycosylase enzymes recognize and remove the modified/damaged bases from the DNA (Sakumi and Sekiguchi, 1990). This is followed by the removal of the nucleotide and a replacement of the polynucleotide strand. So far, in plants, several lesion-specific DNA glycosylases have been described; for instance, uracil glycosylase recognizes and removes uracil formed due to spontaneous deamination of cytosine (**Figure 5**). In *Arabidopsis*, whole-cell extract DNA containing uracil is repaired by the BER pathway in combination with uracil-DNA glycosylases. In particular, *in vitro* reconstitution of DNA repair reactions carried out with isolated cell extracts from *Arabidopsis* or other plants has been extremely helpful in identifying several structural and functional aspects of BER. Hypoxanthine, 3-methyladenine, 7-methylguanine, and other modified bases are recognized by other glycosylases. The first cloned plant DNA repair gene, *Arabidopsis* 3-methyladenine-DNA glycosylase, has been shown to eliminate MMS-induced DNA lesions (Santerre and Britt, 1994). Other bifunctional glycosylases, such as 8-oxoG DNA glycosylase/AP lyase, cut the DNA backbone on the 3′ side of the AP site followed by repair of 7,8-dihydro-8-oxoguanine (8-oxoG), a guanine oxidation product in *Arabidopsis* (Bray and West, 2005). More specifically, the lesion-specific DNA glycosylase hydrolyses the *N*-glycosidic bond linking the modified/damaged base to the 1′-carbon atom of deoxyribose sugar, without altering the DNA sugar-phosphate backbone. This results in the creation of an abasic site, which is then recognized by an AP endonuclease or AP lyase, which cuts the DNA backbone by cleaving the phosphodiester bond at the AP site (**Figure 6**). Subsequently, depending on the nature of the lesion and the enzyme involved, the repair response can either continue by "short" or "long" patch mechanisms. In mammalian cells, BER's "short" mode exploits DNA polymerase β (Pol β), XRCC1 (X-ray repair cross-complementing protein 1), and LIG3α to repair a single-nucleotide gap The BER "long patch" removes 10 nucleotides surrounding the lesion and relies on the involvement of the DNA polymerase δ/ε-proliferating cell nuclear antigen-flap endonuclease 1 (δ/ε-PCNA-FEN1) complex. In plants, short-patch BER is an important DNA repair mechanism for uracil elimination in mitochondrial DNA (Boesch et al., 2009). The short-patch repair is less conserved because of the lack of plant homologs of DNA Pol β or DNA ligase III. Notably, considering the absence of Pol β and ligase III homologs in plants, all BER modes can occur after the initial incision stages, and the repair reactions are completed by the *Arabidopsis* DNA ligase 1 (AtLIG1) ligation (Cordoba-Cañero et al., 2009; Cordoba-Cañero et al., 2011). However, DNA polymerase λ in rice showed *in vitro* deoxyribose phosphate (dRP) lyase activity and sequence similarity with human Pol λ and therefore may be a substitute for Pol β (Uchiyama et al., 2004). Furthermore, XRCC1-like protein isolated from *Arabidopsis* is devoid of domains that mediate in the interaction of XRCC1 with Pol β, and LIG3α in mammals, however, possesses a conserved BRCT domain that mediates interaction with poly(ADP-ribose) polymerase (PARP) (Taylor et al., 1998; Uchiyama et al., 2004). There are at least

two PARP activities in plants that may play role in BER and recombinational repair pathways (Amor et al., 1998; Babiychuk et al., 2001). It is pertinent to mention that SSBs in DNA during BER are inevitable intermediates and can act as substrates for nucleotide excision and recombination repair (Memisoglu and Samson, 2000). Several findings indicate that BER plays a critical role in repairing seed storage-induced oxidative DNA lesions in germinating embryos (Macovei et al., 2011; Cordoba-Cañero et al., 2014). Further understanding of these processes will help enhance the means of protecting seeds and discover new ways of preserving their capacity to germinate.

Nucleotide Excision Repair

Nucleotide excision repair is used to repair bulky types of DNA damage, such as steric changes in DNA duplex structure or base dimers, in which an oligonucleotide of 30 bases is excised followed by DNA polymerase mediated resynthesis in the single-stranded region (Kusakabe et al., 2019; Ferri et al., 2020). This pathway can also recognize polymerase-blocking lesions using stalled RNA polymerase, which is then fed into the NER pathway (Waterworth et al., 2019). The minute details underlying mechanisms of NER have been explored by comprehensive studies in both simple and complex organisms. Mostly NER genes and associated repair proteins share a similar pattern of organization in both crop and model plants. In general, NER plays a critical role in corrections of structural alterations in regular DNA double-helix, and hence, it is conserved in both prokaryotic and eukaryotic organisms. For instance, UV-induced photo products such as pyrimidine dimers and 6-4 PPs that produce significant conformational changes in DNA are key substrates of NER. The serious human disorders caused by inborn genetic defects in NER proteins, such as xeroderma pigmentosum and Cockayne syndrome, demonstrate the significance of this repair process (Lehmann et al., 2018; Krokidis et al., 2020). NER eliminates these adducts by making an incision on both sides of the adduct followed by the removal of this incised stretch of DNA through a helicase (Marteijn et al., 2014). The gap is eventually filled by DNA Pol δ with the help of RPA, PCNA, and FEN1 (**Figure 7**). However, in plants, the homolog of human DNA Pol δ is not yet clear, and further research is required to demonstrate the enzyme that fills the gaps created by the removal of ssDNA on each side of the lesion. NER varies in two ways from BER: first, the diversity of DNA damage products recognized by the NER is strikingly large, and second, the repair complex initiates repair by creating nicks on the affected strand. These nicks occur at both 5′ and 3′ ends of the lesion at a particular distance, which is then excised as an oligonucleotide by the action of a helicase. Recent work suggests that DNA/RNA helicases can mitigate the negative effects of multiple abiotic stress factors (Gill et al., 2015). In eukaryotes, OsXPB2, a member of the strongly conserved helicase superfamily 2, is involved in DNA metabolism, such as transcription and repair (Umate et al., 2011). With differing efficiencies, the excision repair complex cleaves almost every DNA structure abnormally from very thin, non-distorting lesions (such as O^6-methylguanine or abasic sites) to very bulky adducts (thymine-psoralen adducts or pyrimidine dimers). For every potential lesion, it is not feasible for a cell to create a particular repair enzyme; therefore, this pathway has evolved to deal with diverse kinds of damages. The efficacy of NER varies, depending on the nature of the DNA lesion and its genomic location. There are two separate NER subpathways: (a) global genomic repair (GGR) that repairs alterations in chromatin structure and DNA-associated proteins, (b) transcription-coupled repair (TCR) that eliminates transcription-locking lesions from the heavily expressed genes (Hanawalt, 2002). The two NER modes share the same repair proteins, however, differ primarily in sensing DNA damages. In higher eukaryotes, TCR recognizes stalled RNA Pol II complex on the transcribed strand after encountering DNA damage, and hence only this DNA strand is fixed quickly, whereas GGR recognizes damages on the coding strand that persist for longer durations (Tornaletti, 2005). GGR is dependent on xeroderma pigmentosum group C (XPC)/hHR23B complex stabilized by hCEN2 that mediates recognition of DNA damages (Thoma and Vasquez, 2003). Whereas TCR is independent of XPC is initiated on encountering stalled RNA polymerase II (Mu and Sancar, 1997). *Arabidopsis* deficient in AtCEN2 revealed reduced repair of UV-C–caused DNA damage *in vitro* (Molinier et al., 2004). As part of the *Arabidopsis* homolog of the human XPC protein (AtRAD4) recognition complex, the *A. thaliana* CENTRIN2 (AtCEN2) gene was implicated in the early stages of GGR, thereby modulating both NER and HRR. A relation between NER and HRR has also been shown to be an alternate mechanism for CPD repair in plants (Molinier et al., 2004; Liang et al., 2006). Hence, it can be concluded that several NER genes are related to factors involved in homologous recombination and photo repair in plants, and such a complex interplay of different DNA repair pathways could improve the plasticity and adaptability of the plant genome to a wide range of ecologies (Manova and Gruszka, 2015). For plants, the selective activity of excision repair mechanisms at the level of actively transcribed genes tends to be very important, and it may be useful to investigate the role of gene-specific repair in augmenting UV tolerance in crop species.

The key discrepancies in the mismatch, base excision, and nucleotide–excision repair mechanisms are in the identification and mode of excision of damaged nucleotide. In BER and MMR, a single nick is created in the sugar-phosphate backbone on one side of the damage, whereas in NER, nicks are made on both sides of the DNA damage. Furthermore, in BER, DNA polymerase displaces the old nucleotides when it extends the exposed 3' end of the nick; in MMR, the old nucleotides are degraded, and in NER, nucleotides are displaced by helicase enzymes. DNA polymerase and ligase are used by all three pathways to fill in the gap created by the excision and for sealing the nick, respectively.

Homologous Recombination Repair and Non-Homologous End-Joining

The DNA repair mechanisms mentioned previously occasionally fail to completely repair the lesions, resulting in SSBs or DSBs. Additionally, these breaks can also be induced by the exposure of cells to exogenous agents such as ionizing radiation. DSBs are the most damaging of all the lesions, and a few unrepaired DSBs can lead to chromosomal fragmentation and even cell death (Dudáš and Chovanec, 2004; Sonoda and Hochegger, 2006).

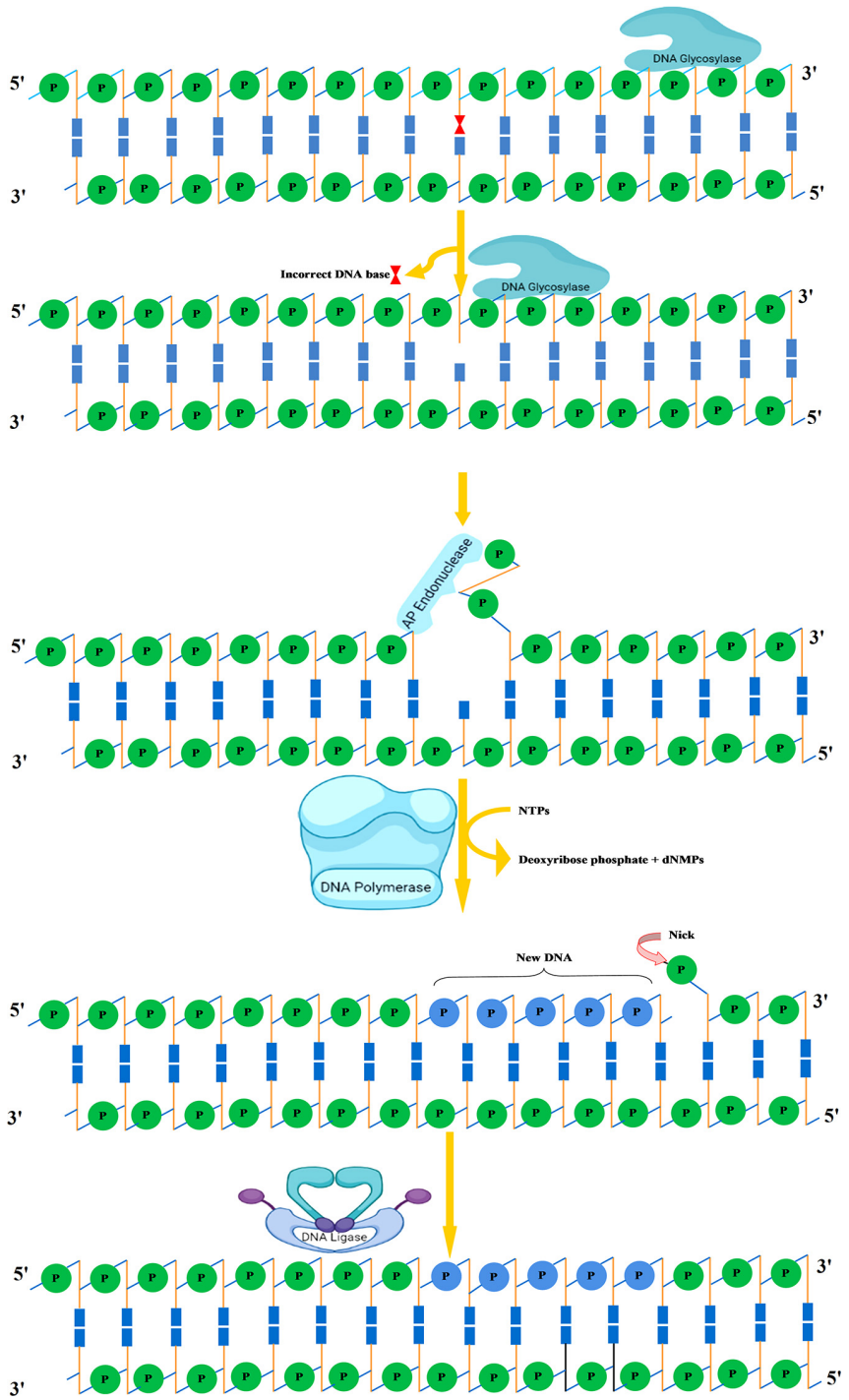

FIGURE 6 | Base excision repair. Recognition followed by removal of damaged DNA base by DNA glycosylase resulting in the formation of AP site. An AP endonuclease nicks the phosphodiester backbone near the AP site. DNA polymerase I replaces the damaged portion with a new DNA. Finally, DNA ligase seals the nick. AP, apurinic/apyrimidinic.

DSBs usually occur spontaneously within a cell, particularly during DNA replication and when the cell is under oxidative stress (Waterworth et al., 2019). These breaks in S-phase can obstruct the progression of the moving replication fork, resulting in a replication fork blockade (Hochegger et al., 2004). To circumvent the toxic effects of DSBs, organisms have evolved two pathways viz. homologous recombination and NHEJ for the repair of DNA breaks.

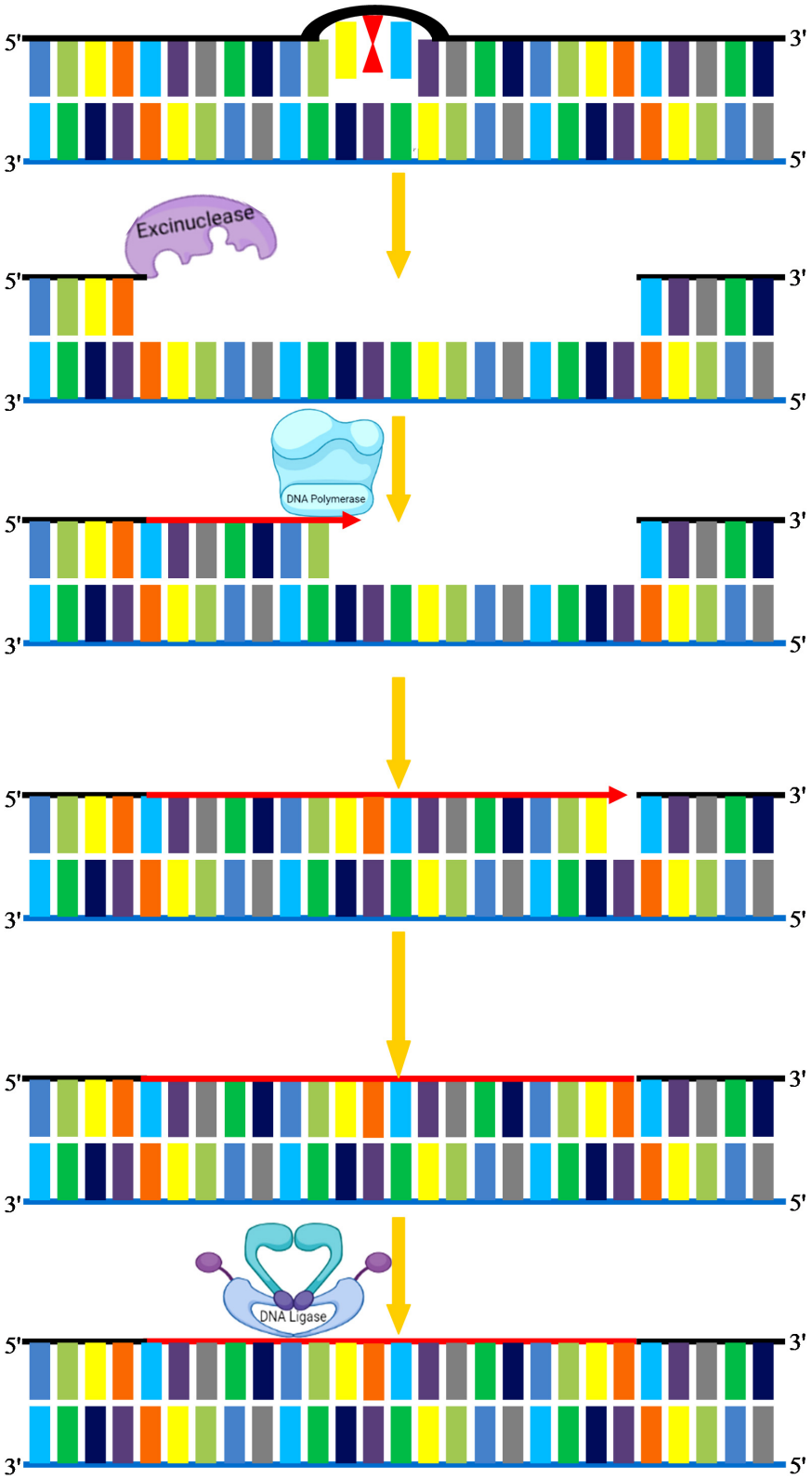

FIGURE 7 | Nucleotide excision repair. NER eliminates these adducts by making an incision on both sides of adduct by excinucleases followed by the removal of this incised stretch of DNA. DNA polymerase I replaces the damaged portion with a new DNA. The gap is eventually filled by DNA ligase.

Homologous recombination repair

A homologous recombination is a form of genetic recombination in which nucleotide sequences are swapped between two DNA molecules that are either related or identical. Cells normally use it to repair toxic double-strand breaks that occur on both strands of the DNA. During meiosis, the mechanism by which eukaryotes including animals and many plants make sperm and egg cells, homologous recombination creates new variations of DNA sequences. These new DNA combinations create genetic diversity in offspring, which allows populations to respond to changing conditions over time. The HRR pathway is a "flawless" DNA repair mechanism that repairs DSBs by using information encoded by homologous sequence. HRR is enabled by DSBs that occur inside replicated DNA (replication-independent DSBs) or at broken replication forks (replication-dependent DSBs). Production of the ends of the DNA double-strand break, homologous DNA pairing, and strand exchange, repair DNA synthesis, and resolution of the heteroduplex molecules are all part of HRR. To initiate the repair of the DSBs by homologous recombination, the DNA breaks must first be recognized, and an appropriate signal must be sent to the repair machinery for checkpoint activation. The repair initiates with the recruitment of the MRN complex at the site of DSBs (Charbonnel et al., 2010). MRN complex facilitates the recruitment of key regulators of DSB repair, protein kinases belonging to the phosphatidyl-inositol 3-kinase (PI3-kinase) family, ATM, and ATR (**Figure 8**). The MRN complex starts processing the DNA ends by the exonucleolytic degradation of the 3′ end followed by the activation of ATM/ATR that, in turn, phosphorylate the Sae2/CtIP and hundreds of other target protein involved in DSB repair and checkpoint activation (Charbonnel et al., 2010). The recruitment of these proteins is essentially required to generate the free 3′ ends and stabilization of DSBs. These 3′ overhangs produced by the excision of the 5′ end by MRN complex are coated with RPA to prevent its exonuclease-mediated degradation (Schmidt et al., 2019). This is followed by the binding o breast cancer 1/2 (BRCA1/2), which subsequently recruits RAD51 at the site of DSBs. RAD51 displaces the bound RPA and facilitates strand invasion into the homologous template (Mannuss et al., 2012). Next, the 3′ overhang coated with RAD51 locates the homologous sequence and invades the dsDNA by displacing the second strand of the template generating the "D-Loop" (displacement loop) (Dudáš and Chovanec, 2004; Mannuss et al., 2012; Ganai et al., 2016). After the formation of "D-Loop," breaks can be either repaired by the synthesis-dependent strand-annealing (SDSA) model or double-strand break repair (DSBR) in which double Holliday junction (dHJ) intermediates are formed. The dHJ intermediates are resolved by resolvases that cut the crossed or non-crossed strands, resulting in the crossover or non-crossover products (Dudáš and Chovanec, 2004). The SDSA method uses the donor strand to fill the gap by using its sequence information, thus realigning the invasive strand to the original break site (Schmidt et al., 2019). In contrast to DSBR, the repaired end products always consist of non-crossovers. HRR uses the undamaged sister chromatid to restore the missing genetic information due to DSBs (Dudáš and Chovanec, 2004). As homologous sister chromatids are needed to repair the damage, it is therefore believed that HRR is only active during the S and G2 phases of the cell cycle (Davis and Chen, 2013). Interestingly, Gallego et al. (2005) identified six RAD51 paralogs in *Arabidopsis*, of which the expression of three is upregulated upon treatment with γ-irradiation. Thus, indicating that RAD51 paralogs play a central role in repairing γ-rays induced DSB through the HRR pathway.

Non-homologous end-joining

Non-homologous end-joining accounts for the most common form of DSB repair mechanism in plants (Puchta, 2005; Pannunzio et al., 2018). It involves the direct joining of two broken DNA ends. In comparison to homologous recombination, which involves a homologous sequence to guide repair, NHEJ directly ligates two ends without the need for a homologous sequence. NHEJ can be subdivided into two classes depending on the pathway used to repair the damage. The first one is KU-dependent classical/canonical NHEJ (c-NHEJ) repair, which encompasses direct ligation of the broken ends generally yielding error-free repair; however, occasionally small (usually less than a few nucleotides) insertions or deletions occur. In c-NHEJ, the KU heterodimer consisting of two subunits with 70- and 80-kDa molecular weight; i.e., KU70 (XRCC6) and KU80 (XRCC5) bind to the DSB to initiate the repair (Shen et al., 2017). As NHEJ involves rejoining the broken ends, the binding of KU not only prevents the damage of the free DNA ends but also assists in aligning the ends closer to each other (Mannuss et al., 2012). Subsequently, KU recruits other key proteins such as ligase IV, protein kinases C to repair the free DNA ends (Mannuss et al., 2012; Shen et al., 2017). In *Arabidopsis*, it was observed that AtKU70 and AtKU80 mRNAs increased threefold after induction of the DSBs (Mannuss et al., 2012). Thus, indicating that KU plays a crucial role in repairing DSB through the NHEJ pathway (**Figure 9**). Another NHEJ repair pathway works without the requirement of KU, and this pathway is referred to as backup-NHEJ pathway (b-NHEJ) or alternative NHEJ (Alt-NHEJ) or microhomology-mediated NHEJ because it acts in the absence of c-NHEJ. Very little is known about the mechanism of the b-NHEJ pathway, which involves multiple components such as polymerase (ADP-PARP1), but the function of PARP1 in c-NHEJ is not clear as it appears to be involved in a KU-dependent manner too (Shen et al., 2017). This pathway uses microhomologous sequences during the alignment of broken ends before ligating them together, thus resulting in deletions flanking either side of the original break. There are two conflicting reports regarding the repair of DSBs in plants. A study conducted in *A. thaliana* revealed that the predominant repair mechanism for DSBs is mediated by Alt-NHEJ exploiting DNA polymerase θ (PolQ) (Van Kregten et al., 2016). However, a second study reported that there are dissimilar mechanisms for the repair of DSBs in somatic and germ cells (Faure, 2021; Nishizawa-Yokoi et al., 2021). In the case of *A. thaliana* germ cells, the repair is completely dependent on Pol Q by Alt-NHEJ. However, the same authors in *A. thaliana* and rice somatic cells suggest the lack of an absolute requirement of Pol Q for the repair of DSBs revealing HRR is perhaps the predominant mechanism. Overall, these studies point toward the existence of a

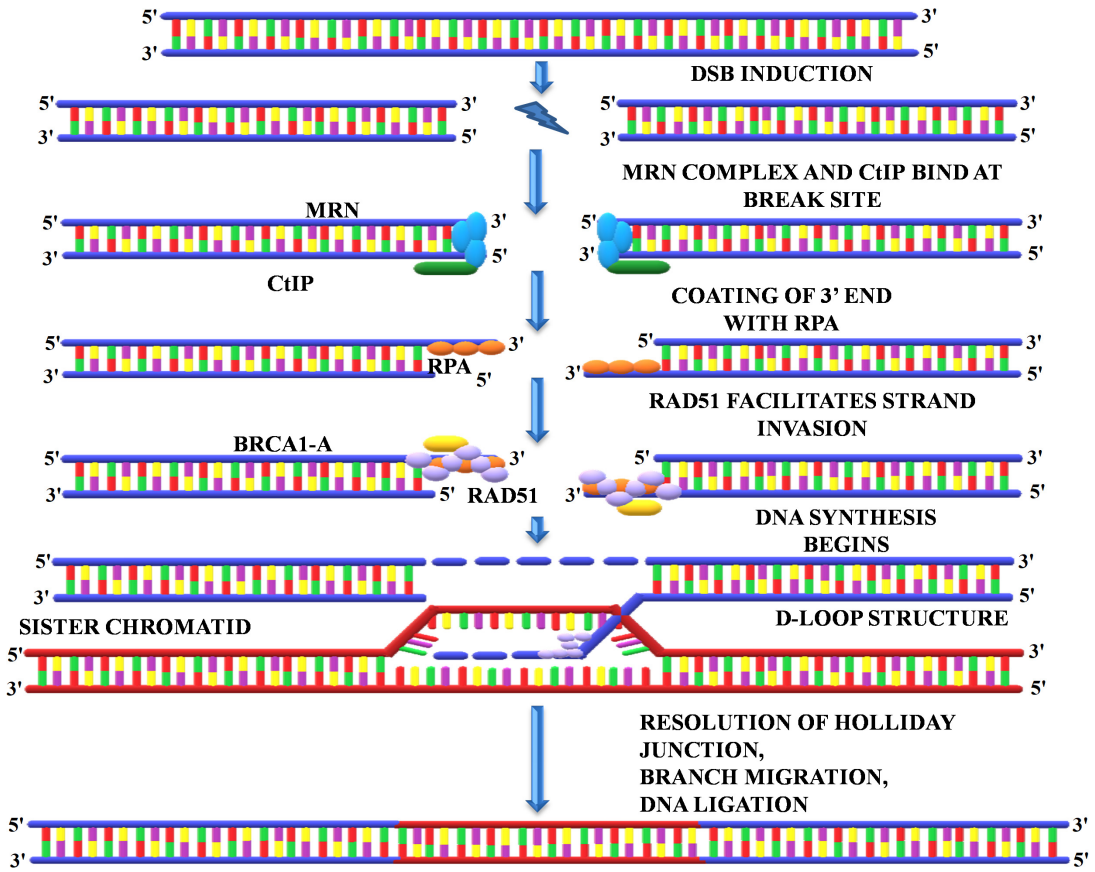

FIGURE 8 | Homologous recombination repair. The HRR initiates with the recruitment of MRN, and CtIP complex at the repair site activates the kinases such as ATM and ATR. The MRN complex degrades the 3' end followed by coating with replication protein A and binding of BRCA1/2, which subsequently recruits RAD51 and initiates DNA synthesis. RAD51 displaces the bound RPA and facilitates strand invasion into the homologous template that generates the D-Loop and Holliday junction, which are eventually resolved by resolvase. MRN, Mre11-Rad50-Nbs1; CtIP, carboxy-terminal interacting protein; ATM, ataxia telangiectasia mutated; ATR, ataxia telangiectasia mutated and Rad3 related; BRCA1/2, breast cancer1/2; RAD51, radiation sensitive 51.

different mechanism for the repair of DSBs in plants. However, in mosses, the repair of DSBs predominantly occurs through HRR (Mara et al., 2019). *Pol q* deletion mutants do not show any developmental or genetic instability phenotype in mosses. Furthermore, these mutants showed the same sensitivity as wild type to DNA-damaging agents such as MMS, cisplatin UV rays except for bleomycin for which it was less sensitive than the wild type. These Pol Q mutants displayed enhanced HRR compared to wild type, indicating Pol Q acts as an inhibitor of the HR repair pathway. Taken together, these studies suggest that in mosses repair of DSBs predominantly occurs through HRR than Alt-NHEJ.

ROLE OF SMALL RNAs IN DNA DAMAGE RESPONSE

So far, we have explored the roles of various proteins in DNA damage response and maintenance of genome integrity. The vital roles of RNA in regulating DNA repair have started to emerge and reflect their importance in maintaining genome integrity by signaling DNA repair cascade via a mechanism not understood yet. However, recent evidence suggests a conserved and crucial role of RNA molecules, RNA processing enzymes, and other factors in DNA repair. It appears that most of the genome gets transcribed, but many of these transcripts do not code for proteins. These transcripts are called non-coding RNAs (ncRNAs), and some of these ncRNAs remain associated with chromatin in a sequence-specific manner to control many cellular pathways such as gene expression (di Fagagna, 2014). Recent studies about ncRNA reveal its additional role in refining the DDR. The structural integrity of DNA depends on small ncRNAs acting at the site of DNA damage. These small RNAs are recruited to the site of DNA damage and help transduce the signal for the recruitment of proteins at the site of DNA damage for accurate DNA repair.

The chemically induced replication stress led to an interaction between non-canonical small RNAs and DDR that led to subsequent production of small RNAs from actively transcribed ribosomal loci in *Neurospora crassa*, and this event was assisted by an ortholog of argonaute protein and RdRPs. These small RNAs were produced from the degradation of longer RNA species. The

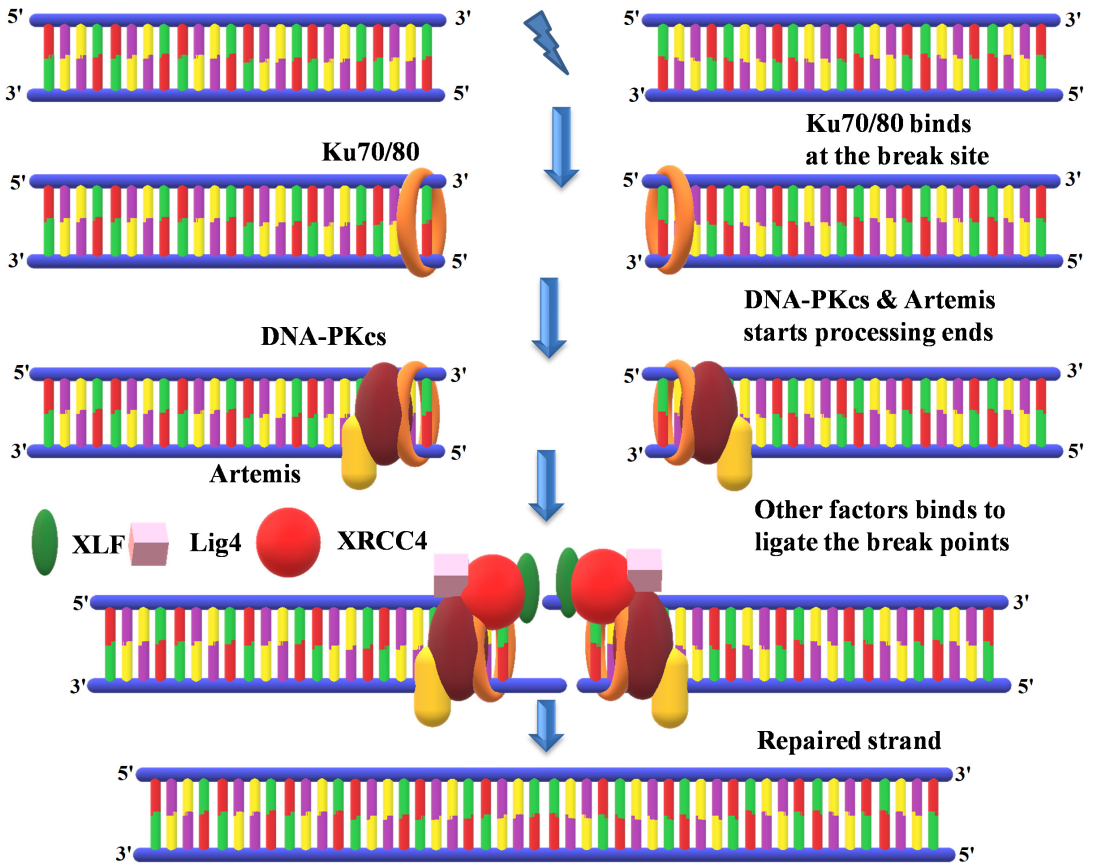

FIGURE 9 | Non-homologous end-joining recombinational repair. KU70 and KU80 bind at the repair site followed by the processing of DNA ends by DNA-PKcs and Artemis. This is followed by the synthesis of new DNA in association with key proteins such as XLF, ligase 4, and XRCC4 at free DNA ends. KU70 and KU80, heterodimer protein with 70- and 80-kDa molecular weight; DNA-PKcs, DNA-dependent protein kinase, catalytic subunit; XRCC4, X-ray repair cross-complementing protein; XLF, XRCC4-like factor.

aberrant transcripts ("aRNA") transcribed as a result of DNA damage are unresponsive to RNA polymerase inhibitors and are amplified by RdRPs and then processed into small RNA known as quelling-induced RNA (qiRNA). These qiRNAs then facilitate the degradation of aRNA, similar to the small interfering RNA (siRNA) amplification cycle (Schalk et al., 2017).

Wei et al. (2012) reported the production of diRNAs (DSB-induced small RNAs) in an *Arabidopsis* transgenic line. DSB repair through SSA (single-strand annealing) mechanism restores *b*-glucuronidase expression, which provides a visible and quantitative readout of DSB repair events (Wielgoss et al., 2013). The biogenesis of diRNAs requires the PI3-kinase ATR, RNA Pol IV, and Dicer-like proteins. Also, any kind of changes or directed mutagenesis in these proteins has resulted in a significant reduction in DSB repair efficiency, which confirmed the role of small RNA in DNA repair efficiency. As discussed in the above sections, UV radiations induce the formation of CPDs and 6-4 PP, which damage DNA structure and disturb cell/genome integrity by distorting regular DNA double-helical structure. However, plants have evolved a mechanism to escape and mitigate UV-induced irreversible DNA damage at their growing points. For instance, in UV-irradiated *A. thaliana*, the DNA damage-binding protein 2 (DDB2) and argonaute 1 (AGO1) form a chromatin-bound complex together with 21-nucleotide-long siRNAs, which perhaps assist in recognizing damage sites in an RNA/DNA complementary strand-specific manner. Synthesis of siRNA, which is associated with the PPs, involves the unusual concerted action of RNA polymerase IV, RNA-dependent RNA polymerase-2, and Dicer-like-4 (DCL4). Moreover, the association/dissociation of the DDB2-AGO1 complex with chromatin is under the control of siRNA abundance and DNA damage signaling, thus providing a view on the interplay between small RNAs and DNA repair recognition factors at damaged sites (Schalk et al., 2017).

SCOPE OF DNA REPAIR MECHANISMS IN CROP IMPROVEMENT

Biotic and abiotic stresses frequently affect various developmental stages of crop plants and reduce their economical yield. Additionally, these stressful conditions also influence the efficiency of DNA repair pathways resulting in increased mutation frequency and genetic variability. Higher genetic

variability in any species may evolve new phenotypes that can significantly enhance the adaptability to a range of ecologies (Wielgoss et al., 2013). DNA repair pathways have played an important role in induced mutagenesis as mutagens induce a wide range of DNA damages, which can have disastrous consequences on the integrity of the genome. However, some of these erroneous mutations can have beneficial consequences as well and are chosen by natural selection. These mutations have played an immense role in crop improvement programs by increasing genetic variability and developing new mutant varieties with improved traits within a short period, which can be further explored by the plant breeders (Oladosu et al., 2016). To date, it has made an immense contribution in the improvement of yield, maturity durations, and biotic and abiotic stress resistance and still utilized by plant breeders across the globe for crop improvement (Oladosu et al., 2016). Moreover, the improved mutant varieties play a vital role in crop biodiversity and offer useful breeding material for further crop improvement (Chaudhary et al., 2019; Raina et al., 2020).

Significant advancement in food production and quality has been recorded over the last six decades with the help of available genetic variation and diversity in crop plants. Although looking at the rising human population and reduced cultivable lands, further improvement in food production and nutritional quality is required in the near future. Expanding the knowledge of DNA repair processes in plants will possibly pave the way for interesting biotechnological applications aimed at improving stress tolerance in crops. Several researchers have reported the role of various enzymes and genes in DNA repair and subsequent productivity of plants.

Alterations in the expression pattern of genes have been reported to promote several beneficial activities in *Arabidopsis*. Kaiser et al. (2009) reported that the overexpression of photolyase enzyme may increase total biomass production under elevated UV-B radiation. Vanderauwera et al. (2007) demonstrated that reduced PARP levels in transgenic *Arabidopsis* led to enhanced tolerance to a wide range of abiotic stresses. Kimura and Sakaguchi (2006) reported the UV tolerance in *Arabidopsis* and rice by overexpression of the gene encoding the CPD photolyase enzyme. Similarly, the activity of helicases is usually up-regulated during stress conditions in plants. Vashisht and Tuteja (2006) demonstrated the overexpression of helicase enzyme in high salinity stress.

The disruption of MMR activities in plants through RNAi, CRISPR/Cas9, zinc-finger nucleases (ZFN), transcription activator–like effector nucleases (TALEN), or any other genetic engineering tools may perhaps create huge genetic variation and diversity as required for crop improvement. This phenomenon may generate novel plant types with desirable traits. The depletion of the nuclear-encoded DNA MMR protein MSH1 causes desirable and heritable changes in plant development. Several researchers reported that disruption of MSH1 genes in *Arabidopsis*, rice, potato, tomato, soybean, sorghum, and tobacco may drastically change their phenotypes and produce a wide range of novel plant types (Santamaria et al., 2014; Virdi et al., 2015; Rakosy et al., 2019; Jiang et al., 2020). A different spectrum of mutations gradually accumulates in MMR-deficient genotypes and increases generation after generation (Chao et al., 2008). However, stabilization of these mutations is quite complicated and still a big challenge to plant biologists. Stabilization can be achieved by bringing back active MMR proteins in the genetically reprogrammed plants or by crossing the mutant with their immediate parent. Moreover, the active MMR gene may stabilize the indels or mutations that occurred in the previous generation and produce genetically reorganized plants (Virdi et al., 2015; Yang et al., 2015).

SITE-DIRECTED MUTAGENESIS

Genome editing has emerged as one of the finest innovations in the field of plant biotechnology. The method encompasses the induction of site-specific DSBs by nucleases in the genome followed by exploitation of the repair of these breaks that lead to a generation of desired mutations. Smih et al. (1995) are the pioneers who induced DSBs in mammalian cells to study DNA repair by expressing I-SceI (intron-encoded endonuclease from *Saccharomyces cerevisiae*). Subsequently, various endonucleases such as meganucleases, ZFNs, transcription activator–like effector nucleases (TALENs), and clustered regularly interspaced short palindromic repeats/CRISPR-associated protein (CRISPR-Cas9) were used to induce site-specific DSBs in the genome (**Figure 10**).

Zinc-finger nucleases contain a DNA-binding domain through which they bind to DNA by recognizing three base pairs at the target site (Kim et al., 1996; Kumawat et al., 2019). To induce DSB, FokI, a type IIS restriction endonuclease isolated from *Flavobacterium okeanokoites*, must act as a dimer; therefore, each FokI monomer is attached to two DNA-binding ZFNs that recognize different DNA sequences (Townsend et al., 2005; Kumawat et al., 2019). When the monomers are closer to each other, FokI is activated and creates a DSB. Bibikova et al. (2001) used ZFNs to induce site-specific DSB in *Xenopus* oocytes that stimulate the HR repair pathway. Interestingly they also showed that targeted mutagenesis could be achieved by NHEJ as a result of ZFN-induced DSB in *Drosophila* (Bibikova et al., 2002). Later, Lloyd et al. (2005) utilized this technique to induce mutations at specific sites in *Arabidopsis*. Targeted mutations conferring herbicide resistance were achieved by altering the sequences of the endogenous acetohydroxyacid synthase (*SuRA* and *SuRB*) genes in the tobacco plant (Wright et al., 2009).

Transcription activator–like effector nucleases is another class of nuclease used for site-directed mutagenesis that focuses on a single nucleotide as opposed to three for ZFNs (Kim et al., 1996; Boch et al., 2009). The structural feature of the TALEN protein is unique in many ways, making it compatible with the design editing tool because it includes the nuclear location signal, N-terminal translocation signal, the acid activation domain, and the central repeat domain that binds DNA (Li et al., 2011a). Li et al. (2011b) designed hybrid TALEN to induce DSB in tobacco leaves.

CRISPR/Cas is the most promising and efficient genome-editing technique than the nucleases discussed above. CRISPR/Cas was discovered in bacteria or archaea as a type II prokaryotic adaptive immune system, which provides bacteria immunity against invading phages (Jinek et al., 2012;

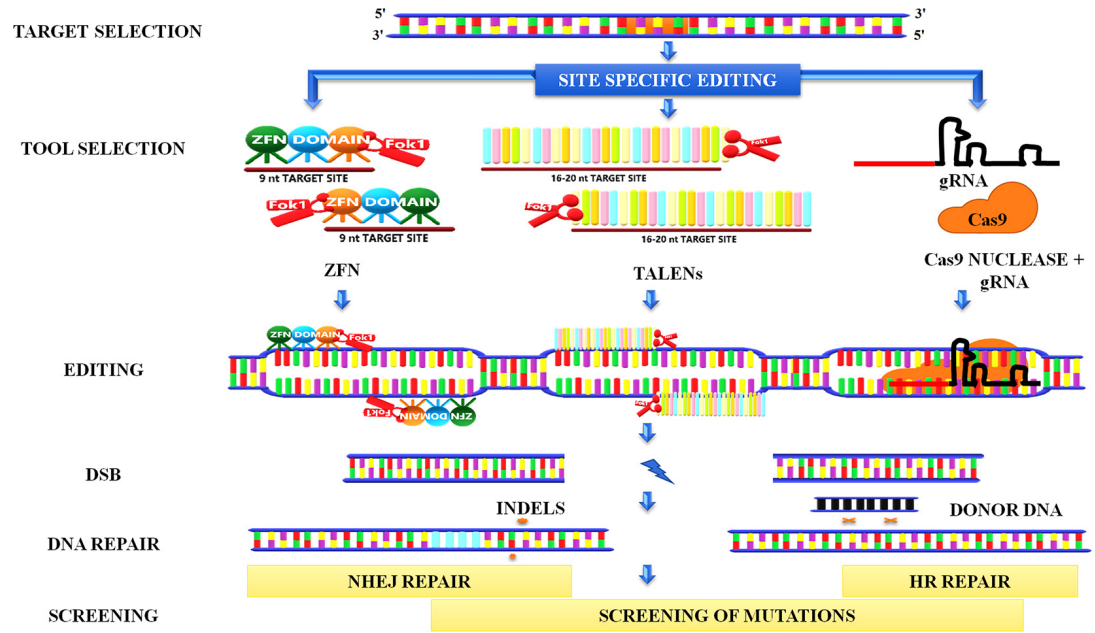

FIGURE 10 | Genome editing. Site-directed genome editing involves the induction of site-specific double-stranded breaks in the genome followed by the recruitment of endonucleases such as ZFNs, TALENs, CRISPR-Cas9. ZFN recognizes nine nucleotide sites on binding and creates a break when two FokI monomers are in proximity to each other. TALENS also works in a similar manner. In the CRISPR system, a gRNA binds to the target site in the genomic region and forms a complex with Cas 9 nuclease to create a break. These breaks can further be repaired by NHEJ, which inserts indels in the sequence, or by HRR pathways in which homolog donor sequence could be used to modify the target sequence. ZFNs, Zinc-finger nucleases; TALENs, transcription activator–like effector nucleases CRISPR-Cas9, clustered regularly interspaced short palindromic repeats/CRISPR-associated protein 9; gRNA, guide RNA; FokI, a type IIS restriction endonuclease isolated from *Flavobacterium okeanokoites*; NHEJ, non-homologous end-joining.

Kumawat et al., 2019). The mechanism of immunizing bacteria against viral attack starts with the incorporation of protospacer, which are small fragments of a foreign sequence in the host chromosome at the proximal end of the CRISPR array (Jinek et al., 2012). The protospacer consists of identical repeats, the transcription product of these repeats results in the generation of precursor CRISPR RNA (pre-crRNA). Later, enzymatic cleavage leads to the formation of crRNA, which has the ability to complementarily base pair with the protospacer sequence of the invasive viral target (Jinek et al., 2012; Kumawat et al., 2019). After recognition of target and complementary base pairing, Cas9 nuclease digests the target sequence and directs the silencing of viral sequences. In bacteria, there are three types of CRISPR/Cas systems known to date, viz. types I, II, and III. Type II system is most commonly used in genome editing. In the type II system, transactivating crRNA (tracrRNA), which is complementary to the pre-crRNA, in the presence of Cas9 tracrRNA helps in the maturation by processing with the ds-RNA–specific ribonuclease RNase III (Jinek et al., 2012; Kumawat et al., 2019). For efficient genome editing, single-guide RNAs (sgRNAs) are synthesized by combining the tracrRNA and crRNAs in which 5′ sequence of sgRNA binds to the target sequence and 3′ sequence binds to the Cas9 nuclease (Kumawat et al., 2019). The targeted mutagenesis by CRISPR/Cas9 is achieved by generating the sgRNAs complementary to the desired site, which allows binding of Cas9 to the desired site. The Cas9 enzyme subsequently cleaves the DNA at the desired site, resulting in the DSB, which is repaired by the HRR or NHEJ pathway leading to small indels. To confirm the role of KU in the NHEJ pathway in plants, Shen et al. (2017) utilized the CRISPR/Cas9 system to induce DSB in two genes, i.e., *Arabidopsis cruciferin 3* (CRU3) and protoporphyrinogen oxidase and observed larger deletions in mutants lacking KU.

FUTURE DIRECTIONS AND CONCLUDING REMARKS

The plant DNA damage response is evolving as a key process influencing plant growth and development in response to adverse environmental cues. The DNA damage response directly influences genome stability by preventing the accumulation of mutations within the organism. The literature discussed in this review reflects the dearth of data regarding the process of genome stability in plants compared to bacteria, yeast, and human. Given the climate change and the stress it imposes on plant growth and productivity, future research in this area will provide important insights into how plants maintain genome stability under stressful conditions. Characterizing various novel interactions between DNA repair proteins in response to stress will open new avenues for crop improvement. Furthermore, with the advent of CRISPR-Cas9 screens, it will be exciting to identify novel genes involved in DNA repair in plants not otherwise possible by classical genetics. Another promising line of research is to understand the link between DNA repair and

chromatin dynamics. DNA repair proteins and processes require access to the DNA damage, which requires extensive chromatin remodeling and epigenetic modifications at the site of the DNA damage. It will be fascinating to uncover such modifications and further determine if such chromatin states are stable and heritable during stressful conditions. These heritable states will allow plants to acclimatize to such adverse environmental conditions. Future work would thus require understanding the mechanism of the initiation of these epigenetic states and designing assay systems that will allow us to study the heritable nature of these epigenetic states.

AUTHOR CONTRIBUTIONS

All authors wrote the initial draft with the following contribution—RG: introduction, future directions, and concluding remarks; AR and RL: DNA repair pathways; NR: recombinational repair; RS: role of small RNAs in DNA damage response; and PS: site-directed mutagenesis and scope of DNA repair mechanisms in crop improvement. After the initial draft was framed, AR and RG rewrote the review. SK throughout the process of writing contributed to the overall assessment of the manuscript.

REFERENCES

Aguilera, A., and García-Muse, T. (2013). Causes of genome instability. *Annu. Rev. Genet.* 47, 1–32. doi: 10.1146/annurev-genet-111212-133232

Ahmad, A., Nay, S. L. and O'Connor, T. R. (2015). "Direct reversal repair in mammalian cells," *Advances in DNA Repair* ed. C. Chen (IntechOpen), 95–128.

Ahnesorg, P., Smith, P., and Jackson, S. P. (2006). XLF interacts with the XRCC4-DNA ligase IV complex to promote DNA nonhomologous end-joining. *Cell.* 124, 301. doi: 10.1016/j.cell.2005.12.031

Amiard, S., Charbonnel, C., Allain, E., Depeiges, A., White, C. I., and Gallego, M. E. (2010). Distinct roles of the ATR kinase and the Mre11-Rad50-Nbs1 complex in the maintenance of chromosomal stability in Arabidopsis. *Plant Cell* 22, 3020–3033. doi: 10.1105/tpc.110.078527

Amor, Y., Babiychuk, E., Inzé, D., and Levine, A. (1998). The involvement of poly (ADP-ribose) polymerase in the oxidative stress responses in plants. *Febs Letters* 440, 1–7. doi: 10.1016/s0014-5793(98)01408-2

Apelt, K., White, S. M., Kim, H. S., Yeo, J. E., Kragten, A., Wondergem, A. P., et al. (2020). ERCC1 mutations impede DNA damage repair and cause liver and kidney dysfunction in patients. *J. Exp. Med.* 218, e20200622. doi: 10.1084/jem.20200622

Babiychuk, E., Van Montagu, M., and Kushnir, S. (2001). N-terminal domains of plant poly (ADP-ribose) polymerases define their association with mitotic chromosomes. *The Plant Journal* 28, 245–255. doi: 10.1046/j.1365-313x.2001.01143.x

Bailey, S. D., Xie, C., Do, R., Montpetit, A., Diaz, R., Mohan, V., et al. (2010). Variation at the NFATC2 locus increases the risk of thiazolidinedione-induced edema in the Diabetes REduction Assessment with ramipril and rosiglitazone Medication (DREAM) study. *Diabetes care.* 33, 2250. doi: 10.2337/dc10-0452

Balinska, K., Wilk, D., Filipek, B., Mik, M., Zelga, P., Skubel, P., et al. (2019). Association of XRCC6 C1310G and LIG4 T9I polymorphisms of NHEJ DNA repair pathway with risk of colorectal cancer in the Polish population. *Pol. Prz. Chir.* 91, 15. doi: 10.5604/01.3001.0013.1030

Beernink, P. T., Segelke, B. W., Hadi, M. Z., Erzberger, J. P., Wilson Iii, D. M., and Rupp, B. (2001). Two divalent metal ions in the active site of a new crystal form of human apurinic/apyrimidinic endonuclease, Ape1: implications for the catalytic mechanism. *J. Mol. Biol.* 307, 1023. doi: 10.1006/jmbi.2001.4529

Bhaskara, V., Dupré, A., Lengsfeld, B., Hopkins, B. B., Chan, A., Lee, J. H., et al. (2007). Rad50 adenylate kinase activity regulates DNA tethering by Mre11/Rad50 complexes. *Mol. Cell.* 25, 647. doi: 10.1016/j.molcel.2007.01.028

Bibikova, M., Carroll, D., Segal, D. J., Trautman, J. K., Smith, J., Kim, Y. G., et al. (2001). Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. *Mol. cell bio* 21, 289–297. doi: 10.1128/mcb.21.1.289-297.2001

Bibikova, M., Golic, M., Golic, K. G., and Carroll, D. (2002). Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases. *Genetics* 161, 1169–1175. doi: 10.1093/genetics/161.3.1169

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., et al. (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science.* 326, 1509–1512. doi: 10.1126/science.1178811

Boesch, P., Ibrahim, N., Paulus, F., Cosset, A., Tarasenko, V., and Dietrich, A. (2009). Plant mitochondria possess a short-patch base excision DNA repair pathway. *Nucleic Acids Res.* 37, 5690–5700. doi: 10.1093/nar/gkp606

Bray, C. M., and West, C. E. (2005). DNA repair mechanisms in plants: crucial sensors and effectors for the maintenance of genome integrity. *New Phytologist* 168, 511–528. doi: 10.1111/j.1469-8137.2005.01548.x

Brettel, K., and Byrdin, M. (2010). Reaction mechanisms of DNA photolyase. *Curr.Opin.Struct.biol* 20, 693. doi: 10.1016/j.sbi.2010.07.003

Briggs, F. B., Goldstein, B. A., McCauley, J. L., Zuvich, R. L., De Jager, P. L., Rioux, J. D., et al. (2010). Variation within DNA repair pathway genes and risk of multiple sclerosis. *Am. J. epidemiol.* 172, 217. doi: 10.1093/aje/kwq086

Buck, D., Malivert, L., de Chasseval, R., Barraud, A., Fondanèche, M. C., and Sanal, O. (2006). Cernunnos, a novel nonhomologous end-joining factor, is mutated in human immune deficiency with microcephaly. *Cell.* 124, 287. doi: 10.1016/j.cell.2005.12.030

Burgers, P. M., and Kunkel, T. A. (2017). Eukaryotic DNA replication fork. *Annu. Rev. Biochem.* 86, 417–438. doi: 10.1146/annurev-biochem-061516-044709

Burkovics, P., Szukacsov, V., Unk, I., and Haracska, L. (2006). Human Ape2 protein has a 3'-5' exonuclease activity that acts preferentially on mismatched base pairs. *Nucleic Acids Res.* 34, 2508. doi: 10.1093/nar/gkl259

Carballar, R., Martínez-Láinez, J. M., Samper, B., Bru, S., Bállega, E., Mirallas, O., et al. (2020). CDK-mediated Yku80 Phosphorylation Regulates the Balance Between Non-homologous End Joining (NHEJ) and Homologous Directed Recombination (HDR). *J. Mol. Biol.* 432, 166715. doi: 10.1016/j.jmb.2020.11.014

Carusillo, A., and Mussolino, C. (2020). DNA Damage: From Threat to Treatment. *Cells.* 9, 1665. doi: 10.3390/cells9071665

Chansel-Da Cruz, M., Hohl, M., Ceppi, I., Kermasson, L., Maggiorella, L., Modesti, M., et al. (2020). A Disease-Causing Single Amino Acid Deletion in the Coiled-Coil Domain of RAD50 Impairs MRE11 Complex Functions in Yeast and Humans. *Cell Reports.* 33, 108559. doi: 10.1016/j.celrep.2020.108559

Chao, E. C., Velasquez, J. L., Witherspoon, M. S. L., Rozek, L. S., Peel, D., Ng, P., et al. (2008). Accurate classification ofMLH1/MSH2missense variants with multivariate analysis of protein polymorphisms-mismatch repair (MAPP-MMR). *Hum. Mutat.* 29, 852–860. doi: 10.1002/humu.20735

Charbonnel, C., Gallego, M. E., and White, C. I. (2010). Xrcc1-dependent and Ku-dependent DNA double-strand break repair kinetics in Arabidopsis plants. *Plant J.* 64, 280. doi: 10.1111/j.1365-313X.2010.04331.x

Chatterjee, N. and Walker, G. C. (2017). Mechanisms of DNA damage, repair, and mutagenesis. *Environ. Mol. Mutagen.* 58 235–263. doi: 10.1002/em.22087

Chaudhary, J., Deshmukh, R., and Sonah, H. (2019). Mutagenesis approaches and their role in crop improvement. *Plants.* 8, 467. doi: 10.3390/plants8110467

Cools, T., and De-Veylder, L. (2009). DNA stress checkpoint control and plant development. *Curr.Opin. Plant Biol.* 12, 23–28. doi: 10.1016/j.pbi.2008.09.012

Cooper, M. P., Machwe, A., Orren, D. K., Brosh, R. M., Ramsden, D., and Bohr, V. A. (2000). Ku complex interacts with and stimulates the Werner protein. *Genes Dev.* 14, 907–912.

Cordoba-Cañero, D., Morales-Ruiz, T., Roldan-Arjona, T., and Ariza, R. R. (2009). Single-nucleotide and long-patch base excision repair of DNA damage in plants. *Plant J.* 60, 176. doi: 10.1111/j.1365-313X.2009.03994.x

Cordoba-Cañero, D., Roldan-Arjona, T., and Ariza, R. R. (2011). Arabidopsis ARP endonuclease functions in a branched base excision DNA repair pathway completed by LIG1. *Plant J.* 68, 493. doi: 10.1111/j.1365-313X.2011.04720.x

Cordoba-Cañero, D., Roldan-Arjona, T., and Ariza, R. R. (2014). Arabidopsis ZDP DNA 3′-phosphatase and ARP endonuclease function in 8-oxoG repair

initiated by FPG and OGG1 DNA glycosylases. *Plant J.* 79, 824. doi: 10.1111/tpj.12588

Coughlin, S. S. (2019). "Epidemiology of breast cancer in women," in *Breast cancer metastasis and drug resistance. Advances in Experimental Medicine and Biology*, Vol. 1152, ed. A. Ahmad (Cham: Springer), doi: 10.1007/978-3-030-20301-6_2

Culligan, K. M., and Hays, J. B. (2000). Arabidopsis thaliana MutS-homolog proteins—atMSH2, atMSH3, atMSH6, and a novel atMSH7 protein—form three distinct heterodimers with different specificities for mismatched DNA. *Plant Cell.* 12, 991–1002. doi: 10.2307/3871224

Dantzer, F., de Murcia, G., Ménissier-de Murcia, J., Nasheuer, H. P., and Vonesch, J. L. (1998). Functional association of poly (ADP-ribose) polymerase with DNA polymerase α-primase complex: a link between DNA strand break detection and DNA replication. *Nucleic Acids res.* 26, 1891. doi: 10.1093/nar/26.8.1891

Dany, A. L., Douki, T., Triantaphylides, C., and Cadet, J. (2001). Repair of the main UV-induced thymine dimeric lesions within Arabidopsis thaliana DNA: evidence for the major involvement of photoreactivation pathways. *J. Photochem. Photobiol. B. Biol* 65, 127–135. doi: 10.1016/s1011-1344(01)00254-8

Das, A., Boldogh, I., Lee, J. W., Harrigan, J. A., Hegde, M. L., Piotrowski, J., et al. (2007). The human Werner syndrome protein stimulates repair of oxidative DNA base damage by the DNA glycosylase NEIL1. *J. Biol. Chem.* 282, 26591. doi: 10.1074/jbc.M703343200

Davis, A., and Chen, D. (2013). DNA double strand break repair via non-homologous end-joining.Transal. *Cancer Res.* 2, 130–143.

Demple, B., Herman, T., and Chen, D. S. (1991). Cloning and expression of APE, the cDNA encoding the major human apurinic endonuclease: definition of a family of DNA repair enzymes. *Proc. Nat.l Acad. Sci.* 88, 11450–11454. doi: 10.1073/pnas.88.24.11450

di Fagagna, F. D. A. (2014). A direct role for small non-coding RNAs in DNA damage response. *Trends Cell Biol.* 24, 171–178. doi: 10.1016/j.tcb.2013.09.008

Dresler, S. L., Gowans, B. J., Robinson-Hill, R. M., and Hunting, D. J. (1988). Involvement of DNA polymerasedelta in DNA repair synthesis in human fibroblasts at late times after ultraviolet irradiation. *Biochemistry.* 27, 6379. doi: 10.1021/bi00417a028

Dudáš, A., and Chovanec, M. (2004). DNA double-strand break repair by homologous recombination. *Mutat. Res.-Rev. Mutat. Res.* 566, 131. doi: 10.1016/j.mrrev.2003.07.001

Duncan, T., Trewick, S. C., Koivisto, P., Bates, P. A., Lindahl, T., and Sedgwick, B. (2002). Reversal of DNA alkylation damage by two human dioxygenases.Proc. *Natl. Acad. Sci.* 99, 16660. doi: 10.1073/pnas.262589799

Esmaeilzadeh, H., Bordbar, M. R., Hojaji, Z., Habibzadeh, P., Afshinfar, D., Miryounesi, M., et al. (2019). An immunocompetent patient with a nonsense mutation in NHEJ1 gene. *BMC med. genet* 20:1. doi: 10.1186/s12881-019-0784-0

Ewing, R. M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., et al. (2007). Large-scale mapping of human protein–protein interactions by mass spectrometry. *Mol. Syst. Biol.* 3, 89. doi: 10.1038/msb4100134

Esteller, M., Garcia-Foncillas, J., Andion, E., Goodman, S. N., Hidalgo, O. F., Vanaclocha, V., et al. (2000). Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* 343, 1350–1354. doi: 10.1056/NEJM200011093431901

Faridounnia, M., Folkers, G. E., and Boelens, R. (2018). Function and interactions of ERCC1-XPF in DNA damage response. *Molecules.* 23, 3205. doi: 10.3390/molecules23123205

Faure, D. (2021). Is there a unique integration mechanism of Agrobacterium T-DNA into a plant genome? *New Phytologist* 229, 2386–2388. doi: 10.1111/nph.17184

Fedeles, B. I., Singh, V., Delaney, J. C., Li, D., and Essigmann, J. M. (2015). The AlkB family of Fe (II)/α-ketoglutarate-dependent dioxygenases: repairing nucleic acid alkylation damage and beyond. *J. Biol. Chem.* 290, 20734. doi: 10.1074/jbc.R115.656462

Ferri, D., Orioli, D., and Botta, E. (2020). Heterogeneity and overlaps in nucleotide excision repair disorders. *Clin.Genet.* 97, 12. doi: 10.1111/cge.13545

Fousteri, M., Vermeulen, W., van Zeeland, A. A., and Mullenders, L. H. (2006). Cockayne syndrome A and B proteins differentially regulate recruitment of chromatin remodeling and repair factors to stalled RNA polymerase II in vivo. *Mol. Cell.* 23, 471. doi: 10.1016/j.molcel.2006.06.029

Galbraith, D. W., Harkins, K. R., and Knapp, S. (1991). Systemic endopolyploidy in Arabidopsis thaliana. *Plant Physio* 96, 985–989. doi: 10.1104/pp.96.3.985

Gallego, C., Estévez, A. M., Fárez, E., Ruiz-Pérez, L. M., and González-Pacanowska, D. (2005). Overexpression of AP endonuclease protects Leishmania major cells against methotrexate induced DNA fragmentation and hydrogen peroxide. *Mol. Biochem. parasitol.* 141, 191. doi: 10.1016/j.molbiopara.2005.03.002

Ganai, R. A., and Johansson, E. (2016). DNA replication—a matter of fidelity. *Mol. Cell.* 62, 745. doi: 10.1016/j.molcel.2016.05.003

Ganai, R. A., Zhang, X. P., Heyer, W. D., and Johansson, E. (2016). Strand displacement synthesis by yeast DNA polymerase ε. *Nucleic Acids Res.* 44, 8229–8240. doi: 10.1093/nar/gkw556

Gatei, M., Young, D., Cerosaletti, K. M., Desai-Mehta, A., Spring, K., Kozlov, S., et al. (2000). ATM-dependent phosphorylation of nibrin in response to radiation exposure. *Nat. Genet.* 25, 115. doi: 10.1038/75508

Ghosal, G., and Muniyappa, K. (2007). The characterization of Saccharomyces cerevisiae Mre11/Rad50/Xrs2 complex reveals that Rad50 negatively regulates Mre11 endonucleolytic but not the exonucleolytic activity. *J. Mol. Biol.* 372, 864. doi: 10.1016/j.jmb.2007.07.013

Groisman, R., Polanowska, J., Kuraoka, I., Sawada, J., Saijo, M., Drapkin, R., et al. (2003). The ubiquitin ligase activity in the DDB2 and CSA complexes is differentially regulated by the COP9 signalosome in response to DNA damage. *Cell.* 113, 357. doi: 10.1016/s0092-8674(03)00316-7

Hanawalt, P. (2002). Sub-pathways of nucleotide excision repair and their regulation. *Oncogene* 21, 8949. doi: 10.1038/sj.onc.1206096

Hawken, S. J., Greenwood, C. M., Hudson, T. J., Kustra, R., McLaughlin, J., and Yang, Q. (2010). The utility and predictive value of combinations of low penetrance genes for screening and risk prediction of colorectal cancer. *Hum. Genet.* 128, 89. doi: 10.1007/s00439-010-0828-1

Hayward, B. E., Steinbach, P. J., and Usdin, K. (2020). A point mutation in the nuclease domain of MLH3 eliminates repeat expansions in a mouse stem cell model of the Fragile X-related disorders. *Nucleic Acids res.* 48, 7856. doi: 10.1093/nar/gkaa573

Hendrich, B., and Bird, A. (1998). Identification and characterization of a family of mammalian methyl CpG-binding proteins. *Genet. Res.* 72, 59–72. doi: 10.1017/s0016672398533307

Henning, K. A., Li, L., Iyer, N., McDaniel, L. D., Reagan, M. S., Legerski, R., et al. (1995). The Cockayne syndrome group A gene encodes a WD repeat protein that interacts with CSB protein and a subunit of RNA polymerase II TFIIH. *Cell.* 82, 555–564. doi: 10.1016/0092-8674(95)90028-4

Hochegger, H., Sonoda, E., and Takeda, S. (2004). Post-replication repair in DT40 cells: translesion polymerases versus recombinases. *Bioessays.* 26, 151–158. doi: 10.1002/bies.10403

Hogg, M., Osterman, P., Bylund, G. O., Ganai, R. A., Lundstrom, E. B., and Sauer-Eriksson, A. E. (2014). Structural basis for processive DNA synthesis by yeast DNA polymerase ε. *Nat. Struct. Mol. Biol.* 21, 49. doi: 10.1038/nsmb.2712

Hong, K. W., Jin, H. S., Lim, J. E., Kim, S., Go, M. J., and Oh, B. (2010). Recapitulation of two genome wide association studies on blood pressure and essential hypertension in the Korean population.J. *Hum.Genet.* 55, 336. doi: 10.1038/jhg.2010.31

Hu, Z., Cools, T., and De Veylder, L. (2016). Mechanisms used by plants to cope with DNA damage. *Annu. Rev. Plant biol.* 67, 439–462. doi: 10.1146/annurev-arplant-043015-111902

Huangteerakul, C., Aung, H. M., Thosapornvichai, T., Duangkaew, M., Jensen, A. N., Sukrong, S., et al. (2021). Chemical-Genetic Interactions of Bacopa monnieri Constituents in Cells Deficient for the DNA Repair Endonuclease RAD1 Appear Linked to Vacuolar Disruption. *Molecules* 26, 1207. doi: 10.3390/molecules26051207

Husain, I., and Sancar, A. (1987). Photoreactivation in phr mutants of *Escherichia coli* K-12. *J Bacteriol* 169, 2367–2372. doi: 10.1128/jb.169.6.2367-2372.1987

Ibrahim Al-Obaide, M. A., Arutla, V., Bacolod, M. D., Wang, W., Zhang, R., and Srivenugopal, K. S. (2021). Genomic Space of MGMT in Human Glioma Revisited: Novel Motifs, Regulatory RNAs, NRF1, 2, and CTCF Involvement in Gene Expression. *Int. J. Mol. Sci.* 22, 2492. ndoi: 10.3390/ijms22052492

Jeggo, P. (1979). Isolation and characterization of *Escherichia coli* K-12 mutants unable to induce the adaptive response to simple alkylating agents.J. *Bacteriol* 139, 783. doi: 10.1128/JB.139.3.783-791.1979

Jiang, M., Wu, X., Song, Y., Shen, H., and Cui, H. (2020). Effects of OsMSH6 mutations on microsatellite stability and homeologous recombination in rice. *Front.Plant Sci.* 11:220.

Jilani, A., Ramotar, D., Slack, C., Ong, C., Yang, X. M., and Scherer, S. W. (1999). Molecular cloning of the human gene, PNKP, encoding a polynucleotide kinase 3′-phosphatase and evidence for its role in repair of DNA strand breaks caused by oxidative damage. *J. Biol.Chem.* 274, 24176. doi: 10.1074/jbc.274.34.24176

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science.* 337, 816–821. doi: 10.1126/science.1225829

Joyce, C. M. (1997). Choosing the right sugar: how polymerases select a nucleotide substrate. *Proc. Natl. Acad. Sci. U.S.A.* 94, 1619–1622. doi: 10.1073/pnas.94.5.1619

Kaiser, G., Kleiner, O., Beisswenger, C., and Batschauer, A. (2009). Increased DNA repair in Arabidopsis plants overexpressing CPD photolyase. *Planta.* 230, 505. doi: 10.1007/s00425-009-0962-y

Kalasova, I., Hailstone, R., Bublitz, J., Bogantes, J., Hofmann, W., Leal, A., et al. (2020). Pathological mutations in PNKP trigger defects in DNA single-strand break repair but not DNA double-strand break repair. *Nucleic acids res.* 48, 6672. doi: 10.1093/nar/gkaa489

Kanno, S. I., Kuzuoka, H., Sasao, S., Hong, Z., Lan, L., Nakajima, S., et al. (2007). A novel human AP endonuclease with conserved zinc-finger-like motifs involved in DNA strand break responses. *EMBO J.* 26, 2094. doi: 10.1038/sj.emboj.7601663

Karimi-Busheri, F., Daly, G., Robins, P., Canas, B., Pappin, D. J., Sgouros, J., et al. (1999). Molecular characterization of a human DNA kinase. *J. Biol. Chem.* 274, 24187. doi: 10.1074/jbc.274.34.24187

Karthika, V., Babitha, K. C., Kiranmai, K., Shankar, A. G., Vemanna, R. S., and Udayakumar, M. (2020). Involvement of DNA mismatch repair systems to create genetic diversity in plants for speed breeding programs. *Plant Physiology Reports* 25, 185–199. doi: 10.1007/s40502-020-00521-9

Keeney, S., Chang, G. J., and Linn, S. (1993). Characterization of a human DNA damage binding protein implicated in xeroderma pigmentosum E. *J. Biol. Chem.* 268, 21293–21300. doi: 10.1016/s0021-9258(19)36923-6

Kim, E., Li, K., Lieu, C., Tong, S., Kawai, S., Fukutomi, T., et al. (2008). Expression of apolipoprotein C-IV is regulated by Ku antigen/peroxisome proliferator-activated receptor γ complex and correlates with liver steatosis. *J. Hepatol.* 49, 787. doi: 10.1016/j.jhep.2008.06.029

Kim, W., Lee, S., Son, Y., Ko, C., and Ryu, W. S. (2016). DDB1 stimulates viral transcription of hepatitis B virus via HBx-independent mechanisms. *J. Virol.* 90, 9644. doi: 10.1128/JVI.00977-16

Kim, Y. G., Cha, J., and Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. *Proc. Nat. Acad. Sci. U.S.A.* 93, 1156. doi: 10.1073/pnas.93.3.1156

Kimura, S., and Sakaguchi, K. (2006). DNA repair in plants. *Chem. Rev.* 106, 753. doi: 10.1021/cr040482n

Krokidis, M. G., D'Errico, M., Pascucci, B., Parlanti, E., Masi, A., Ferreri, C., et al. (2020). Oxygen-Dependent Accumulation of Purine DNA Lesions in Cockayne Syndrome Cells. *Cells* 9, 1671. doi: 10.3390/cells9071671

Kumawat, S., Rana, N., Bansal, R., Vishwakarma, G., Mehetre, S., Das, B. K., et al. (2019). Fast Neutron Mutagenesis in Plants: Advances, Applicability and Challenges. *Plants.* 8, 164. doi: 10.3390/plants8060164

Kupfer, G. M., Näf, D., Suliman, A., Pulsipher, M., and D'Andrea, A. D. (1997). The Fanconi anaemia proteins, FAA and FAC interact to form a nuclear complex. *Nat. genet.* 17, 487. doi: 10.1038/ng1297-487

Kurzbauer, M. T., Janisiw, M. P., Paulin, L. F., Prusén Mota, I., Tomanov, K., Krsicka, O., et al. (2021). ATM controls meiotic DNA double-strand break formation and recombination and affects synaptonemal complex organization in plants. *Plant Cell* koab045. ***Q,

Kusakabe, M., Onishi, Y., Tada, H., Kurihara, F., Kusao, K., Furukawa, M., et al. (2019). Mechanism and regulation of DNA damage recognition in nucleotide excision repair. *Genes. Environ.* 41, 1–6. doi: 10.1186/s41021-019-0119-6

Lario, L. D., Botta, P., Casati, P., and Spampinato, C. P. (2015). Role of AtMSH7 in UV-B-induced DNA damage recognition and recombination. *J. Exp. Bot.* 66, 3019. doi: 10.1093/jxb/eru464

Lavrik, O. I. (2020). PARPs' impact on base excision DNA repair. *DNA repair (Amst).* 93, 102911. doi: 10.1016/j.dnarep.2020.102911

Lee, B. I., Nguyen, L. H., Barsky, D., Fernandes, M., and Wilson, D. M. III (2002). Molecular interactions of human Exo1 with DNA. *Nucleic acids res.* 30, 942. doi: 10.1093/nar/30.4.942

Lee, J. H., and Paull, T. T. (2005). ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. *Science.* 308, 551–554. doi: 10.1126/science.1108297

Lehmann, J., Seebode, C., Martens, M. C., and Emmert, S. (2018). Xeroderma pigmentosum–facts and perspectives. *Anticancer Res.* 38, 1159. doi: 10.21873/anticanres.12335

Lenz, S. A., Li, D., and Wetmore, S. D. (2020). Insights into the Direct Oxidative Repair of Etheno Lesions: MD and QM/MM Study on the Substrate Scope of ALKBH2 and AlkB. *DNA repair (Amst).* 96, 102944. doi: 10.1016/j.dnarep.2020.102944

Leonard, J. M., Bollmann, S. R., and Hays, J. B. (2003). Reduction of Stability of Arabidopsis Genomic and Transgenic DNA-Repeat Sequences (Microsatellites) by Inactivation of AtMSH2 Mismatch-Repair Function. *Plant Physiol.* 133, 328. doi: 10.1104/pp.103.023952

León-Castillo, A., Britton, H., McConechy, M. K., McAlpine, J. N., Nout, R., Kommoss, S., et al. (2020). Interpretation of somatic POLE mutations in endometrial carcinoma. *J.Pathol* 250, 323. doi: 10.1002/path.5372

Leyva-Sánchez, H. C., Villegas-Negrete, N., Abundiz-Yañez, K., Yasbin, R. E., Robleto, E. A., and Pedraza-Reyes, M. (2020). Role of Mfd and GreA in Bacillus subtilis base excision repair-dependent stationary-phase mutagenesis. *J. bacteriol* 202, 9. doi: 10.1128/JB.00807-19

Li, L., Jean, M., and Belzile, F. (2006). The impact of sequence divergence and DNA mismatch repair on homeologous recombination in Arabidopsis. *Plant J.* 45, 908. doi: 10.1111/j.1365-313X.2006.02657.x

Li, T., Huang, S., Jiang, W. Z., Wright, D., Spalding, M. H., Weeks, D. P., et al. (2011a). TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. *Nucleic Acids Res.* 39, 359. doi: 10.1093/nar/gkq704

Li, T., Huang, S., Zhao, X., Wright, D. A., Carpenter, S., Spalding, M. H., et al. (2011b). Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes. *Nucleic acids res.* 39, 6315. doi: 10.1093/nar/gkr188

Li, X., Wu, Y., Suo, P., Liu, G., Li, L., Zhang, X., et al. (2020). Identification of a novel germline frameshift mutation p. D300fs of PMS1 in a patient with hepatocellular carcinoma: A case report and literature review. *Medicine.* 99, e19076. doi: 10.1097/MD.0000000000019076

Li, Y. F., and Sancar, A. (1990). Active site of *Escherichia coli* DNA photolyase: mutations at Trp277 alter the selectivity of the enzyme without affecting the quantum yield of photorepair. *Biochemistry* 29, 5698–5706. doi: 10.1021/bi00476a009

Li, Z. Q., Li, J. T., Bing, J., and Zhang, G. F. (2019). The role analysis of APX gene family in the growth and developmental processes and in response to abiotic stresses in Arabidopsis thaliana.Chinese. *Yi Chuan* 41, 534. doi: 10.16288/j.yczz.19-026

Liang, L., Flury, S., Kalck, V., Hohn, B., and Molinier, J. (2006). CENTRIN2 interacts with the Arabidopsis homolog of the human XPC protein (AtRAD4) and contributes to efficient synthesis-dependent repair of bulky DNA lesions. *Plant mol. biol.* 61, 345. doi: 10.1007/s11103-006-0016-9

Lindahl, T., and Wood, R. D. (1999). Quality control by DNA repair. *Science.* 286, 1897. doi: 10.1126/science.286.5446.1897

Lipkin, S. M., Wang, V., Jacoby, R., Banerjee-Basu, S., Baxevanis, A. D., Lynch, H. T., et al. (2000). MLH3: a DNA mismatch repair gene associated with mammalian microsatellite instability. *Nat. Genet.* 24, 27. doi: 10.1038/71643

Liu, C. Y., Wu, M. C., Chen, F., Ter-Minassian, M., Asomaning, K., Zhai, R., et al. (2010). A Large-scale genetic association study of esophageal adenocarcinoma risk. *Carcinogenesis.* 31, 1259. doi: 10.1093/carcin/bgq092

Liu, Y., Shete, S., Wang, L. E., El-Zein, R., Etzel, C. J., Liang, F. W., et al. (2010). Gamma-radiation sensitivity and polymorphisms in RAD51L1 modulate glioma risk. *Carcinogenesis.* 31, 1762. doi: 10.1093/carcin/bgq141

Lloyd, A., Nafees, B., Narewska, J., Dewilde, S., and Watkins, J. (2006). Health state utilities for metastatic breast cancer. *Br. J. Cancer.* 95, 683. doi: 10.1038/sj.bjc.6603326

Lloyd, A., Plaisier, C. L., Carroll, D., and Drews, G. N. (2005). Targeted mutagenesis using zinc-finger nucleases in Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.* 102, 2232–2237.

Lu, X., Chen, F., Liu, X., Yuan, D., Zi, Y., He, X., et al. (2018). Detection and clinical significance of DNA repair gene ERCC8 tag SNPs in gastric cancer. *Turk. J. Gastroenterol.* 29, 392. doi: 10.5152/tjg.2018.17662

Lu, X., Liu, R., Wang, M., Kumar, A. K., Pan, F., He, L., et al. (2020). MicroRNA-140 impedes DNA repair by targeting FEN1 and enhances chemotherapeutic response in breast cancer. *Oncogene.* 39, 234. doi: 10.1038/s41388-019-0986-0

Lucas-Lledó, J. I., and Lynch, M. (2009). Evolution of mutation rates: phylogenomic analysis of the photolyase/cryptochrome family. *Mol. Biol. Evol* 26, 1143. doi: 10.1093/molbev/msp029

Ma, Y., Pannicke, U., Schwarz, K., and Lieber, M. R. (2002). Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V (D).J. recombination. *Cell.* 108, 781. doi: 10.1016/s0092-8674(02)00671-2

Macovei, A., Balestrazzi, A., Confalonieri, M., Faè, M., and Carbonera, D. (2011). New insights on the barrel medic MtOGG1 and MtFPG functions in relation to oxidative stress response in planta and during seed imbibition. *Plant Physiol.Biochem.* 49, 1040. doi: 10.1016/j.plaphy.2011.05.007

Mannuss, A., Trapp, O., and Puchta, H. (2012). Gene regulation in response to DNA damage. *Biochim. Biophys. Acta.* 1819, 154. doi: 10.1016/j.bbagrm.2011.08.003

Manova, V., and Gruszka, D. (2015). DNA damage and repair in plants – from models to crops. *Front. Plant Sci.* 6:885. doi: 10.3389/fpls.2015.00885

Mara, K., Charlot, F., Guyon-Debast, A., Schaefer, D. G., Collonnier, C., Grelon, M., et al. (2019). POLQ plays a key role in the repair of CRISPR/Cas9-induced double-stranded breaks in the moss Physcomitrella patens. *New Phytologist* 222, 1380–1391. doi: 10.1111/nph.15680

Marini, F., Nardo, T., Giannattasio, M., Minuzzo, M., Stefanini, M., Plevani, P., et al. (2006). DNA nucleotide excision repair-dependent signaling to checkpoint activation. *Proc. Natl. Acad. Sci. U.S.A. doi* 103, 17325–17330. doi: 10.1073/pnas.0605446103

Marteijn, J. A., Lans, H., Vermeulen, W., and Hoeijmakers, J. H. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat. Rev. Mol. Cell. Biol.* 15, 465. doi: 10.1038/nrm3822

Martin, H. A., Porter, K. E., Vallin, C., Ermi, T., Contreras, N., Pedraza-Reyes, M., et al. (2019). Mfd protects against oxidative stress in Bacillus subtilis independently of its canonical function in DNA repair. *BMC Microbiol.* 19:26. doi: 10.1186/s12866-019-1394-x

Memisoglu, A., and Samson, L. (2000). Base excision repair in yeast and mammals. *Mutat Res. Fund. Mol* 451, 39. doi: 10.1016/s0027-5107(00)00039-7

Mengwasser, K. E., Adeyemi, R. O., Leng, Y., Choi, M. Y., Clairmont, C., and D'Andrea, A. D. (2019). Genetic Screens Reveal FEN1 and APEX2 as BRCA2 Synthetic Lethal Targets. *Mol. Cell.* 73, e6. doi: 10.1016/j.molcel.2018.12.008

Meza, T. J., Moen, M. N., Vågbø, C. B., Krokan, H. E., Klungland, A., Grini, P. E., et al. (2012). The DNA dioxygenase ALKBH2 protects Arabidopsis thaliana against methylation damage. *Nucleic Acids Res.* 40, 6620–6631. doi: 10.1093/nar/gks327

Miao, F., Bouziane, M., Dammann, R., Masutani, C., Hanaoka, F., Pfeifer, G., et al. (2000). 3-Methyladenine-DNA glycosylase (MPG protein) interacts with human RAD23 proteins. *J. Biol. Chem.* 275, 28433. doi: 10.1074/jbc.M001064200

Mielecki, D., and Grzesiuk, E. (2014). Ada response - a strategy for repair of alkylated DNA in bacteria. *FEMS Microbiol.Lett.* 355, 1. doi: 10.1111/1574-6968.12462

Miglani, K., Kumar, S., Yadav, A., Aggarwal, N., and Gupta, R. (2021). OGG1 DNA Repair Gene Polymorphism as a Biomarker of Oxidative and Genotoxic DNA Damage. *Iran Biomed J* 25, 47. doi: 10.29252/ibj.25.1.47

Mitra, S., and Kaina, B. (1993). Regulation of repair of alkylation damage in mammalian genomes.Prog. *Nucleic Acid Res. Mo.l Biol.* 44, 109. doi: 10.1016/s0079-6603(08)60218-4

Molinier, J., Ramos, C., Fritsch, O., and Hohn, B. (2004). CENTRIN2 modulates homologous recombination and nucleotide excision repair in Arabidopsis. *The Plant Cell* 16, 1633–1643. doi: 10.1105/tpc.021378

Moslehi, R., Tsao, H. S., Zeinomar, N., Stagnar, C., Fitzpatrick, S., and Dzutsev, A. (2020). Integrative genomic analysis implicates ERCC6 and its interaction with ERCC8 in susceptibility to breast cancer. *Sci Rep.* 10, 21276. doi: 10.1038/s41598-020-77037-7

Mu, D., and Sancar, A. (1997). Model for XPC-independent transcription-coupled repair of pyrimidine dimers in humans. *Journal of Biological Chemistry* 272, 7570–7573. doi: 10.1074/jbc.272.12.7570

Murray, J. M., Tavassoli, M., Al-Harithy, R., Sheldrick, K. S., Lehmann, A. R., Carr, A., et al. (1994). Structural and functional conservation of the human homolog of the Schizosaccharomyces pombe rad2 gene, which is required for chromosome segregation and recovery from DNA damage. *Mol. Cell Biol* 14, 4878. doi: 10.1128/mcb.14.7.4878

Navashin, M., and Shkvarnikov, P. (1933). Process of mutation in resting seeds accelerated by increased temperature. *Nature.* 132, 482–483. doi: 10.1038/132482c0

Nichols-Vinueza, D. X., Delmonte, O. M., Bundy, V., Bosticardo, M., Zimmermann, M. T., Dsouza, N. R., et al. (2021). POLD1 Deficiency Reveals a Role for POLD1 in DNA Repair and T and B Cell Development. *J. Clin.Immunol.* 41, 270. doi: 10.1007/s10875-020-00903-6

Nishizawa-Yokoi, A., Saika, H., Hara, N., Lee, L. Y., Toki, S., and Gelvin, S. B. (2021). Agrobacterium T-DNA integration in somatic cells does not require the activity of DNA polymerase θ. *New Phytologist* 229, 2859–2872. doi: 10.1111/nph.17032

Oladosu, Y., Rafii, M. Y., Abdullah, N., Hussin, G., Ramli, A., Rahim, H. A., et al. (2016). Principle and application of plant mutagenesis in crop improvement: A Review. *Biotechnol. Biotechnol. Equip.* 30, 1. doi: 10.1080/13102818.2015.1087333

Oller, A. R., Fijalkowska, I. J., Dunn, R. L., and Schaaper, R. M. (1992). Transcription-repair coupling determines the strandedness of ultraviolet mutagenesis in *Escherichia coli*. *Proc. Natl. Acad. Sci. U.S.A.* 89, 11036. doi: 10.1073/pnas.89.22.11036

Pannunzio, N. R., Watanabe, G., and Lieber, M. R. (2018). Nonhomologous DNA end-joining for repair of DNA double-strand breaks. *J. Biol. Chem.* 293, 10512–10523. doi: 10.1074/jbc.TM117.000374

Parker, A. E., Van de Weyer, I., Laus, M. C., Oostveen, I., Yon, J., Verhasselt, P., et al. (1998). A human homologue of the Schizosaccharomyces pombe rad1+ checkpoint gene encodes an exonuclease. *J. Biol. Chem* 273, 18332. doi: 10.1074/jbc.273.29.18332

Paull, T. T., and Gellert, M. (1998). The 3' to 5' exonuclease activity of Mre 11 facilitates repair of DNA double-strand breaks. *Mol. Cell.* 1, 969. doi: 10.1016/s1097-2765(00)80097-0

Pegg, A. E. (1990). Properties of mammalian O6-alkylguanine-DNA transferases. *Mutat. Res.* 233, 165–175. doi: 10.1016/0027-5107(90)90160-6

Pegg, A. E. (2011). Multifaceted roles of alkyltransferase and related proteins in DNA repair, DNA damage, resistance to chemotherapy, and research tools. *Chem. Res. Toxicol.* 24, 618–639. doi: 10.1021/tx200031q

Puchta H. (2005). The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *J. Exp. Bot.* 56, 1–14. doi: 10.1093/jxb/eri025

Radicella, J. P., Dherin, C., Desmaze, C., Fox, M. S., and Boiteux, S. (1997). Cloning and characterization of hOGG1, a human homolog of the OGG1 gene of Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U S A.* 94, 8010–8015. doi: 10.1073/pnas.94.15.8010

Raina, A., Laskar, R. A., Tantray, Y. R., Khursheed, S., Wani, M. R., Khan, S., et al. (2020). Characterization of induced high yielding cowpea mutant lines using physiological, biochemical and molecular markers. *Scientific reports.* 10, 1–22.

Rakosy, T. E., Lorincz, B. E., Molnár, I., Thieme, R., Hartung, F., Sprink, T., et al. (2019). New phenotypes of potato co-induced by mismatch repair deficiency and somatic hybridization. *Front. Plant Sci.* 10:3.

Richter, C., Marquardt, S., Li, F., Spitschak, A., Murr, N., Edelhäuser, B. A. H., et al. (2019). Rewiring E2F1 with classical NHEJ via APLF suppression promotes bladder cancer invasiveness. *J. Exp. Clin. Cancer Res.* 38, 292. doi: 10.1186/s13046-019-1286-9

Román-Rodríguez, F. J., Ugalde, L., Álvarez, L., Díez, B., Ramírez, M. J., Risueño, C., et al. (2019). NHEJ-Mediated Repair of CRISPR-Cas9-Induced DNA Breaks Efficiently Corrects Mutations in HSPCs from Patients with Fanconi Anemia. *Cell Stem Cell* 25, 607. doi: 10.1016/j.stem.2019.08.016

Roy, S., Choudhury, S. R., Sengupta, D. N., and Das, K. P. (2013). Involvement of AtPolλ in the repair of high salt- and DNA cross-linking agent-induced double strand breaks in Arabidopsis. *Plant Physiol.* 162, 195–210.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature.* 437, 1173. doi: 10.1038/nature04209

Ruan, C., Workman, J. L., and Simpson, R. T. (2005). The DNA repair protein yKu80 regulates the function of recombination enhancer during yeast mating type switching. *Mol. Cell Biol.* 25, 8476. doi: 10.1128/MCB.25.19.8476-8485

Rybaczek, D., Musiałek, M. W., Vrána, J., Petrovská, B., Pikus, E. G., and Doležel, J. (2021). Kinetics of DNA Repair in Vicia faba Meristem Regeneration Following Replication Stress. *Cells* 10, 88. doi: 10.3390/cells10010088

Rytkonen, A. K., Vaara, M., Nethanel, T., Kaufmann, G., Sormunen, R., Laara, E., et al. (2006). Distinctive activities of DNA polymerases during human DNA replication. *FEBS J.* 273, 2984. doi: 10.1111/j.1742-4658.2006.05310

Ryu, C. S., Bae, J., Kim, I. J., Kim, J., Oh, S. H., Kim, O. J., et al. (2020). MPG and NPRL3 Polymorphisms are Associated with Ischemic Stroke Susceptibility and Post-Stroke Mortality. *Diagnostics* 10, 947. doi: 10.3390/diagnostics10110947

Sabourin, M., Tuzon, C. T., and Zakian, V. A. (2007). Telomerase and Tel1p preferentially associate with short telomeres in S. *Cerevisiae. Mol. Cell.* 27, 550. doi: 10.1016/j.molcel.2007.07.016

Saini, D., Sudheer, K. R., Kumar, P. R. V., Soren, D. C., Jain, V., Koya, P. K. M., et al. (2020). Evaluation of the influence of chronic low-dose radiation on DNA repair gene polymorphisms [XRCC1, XRCC3, PRKDC (XRCC7), LIG1, NEIL1] in individuals from normal and high level natural radiation areas of Kerala Coast. *Int. J. Radiat. Biol.* 96, 734. doi: 10.1080/09553002.2020.1739771

Sakumi, K., and Sekiguchi, M. (1990). Structures and functions of DNA glycosylases. *Mutat. Res.* 236, 161–172. doi: 10.1016/0921-8777(90)90003-n

Sancar, A. (2016). Mechanisms of DNA Repair by Photolyase and Excision Nuclease (Nobel Lecture). *Angew. Chem. Int. Ed. Engl.* 55, 85. doi: 10.1002/anie.201601524

Sannai, M., Doneddu, V., Giri, V., Seeholzer, S., Nicolas, E., and Yip, S. C. (2019). Modification of the base excision repair enzyme MBD4 by the small ubiquitin-like molecule SUMO1. *DNA Repair.* 82, 102687. doi: 10.1016/j.dnarep.2019.102687

Santamaria, R., Shao, M. R., Wang, G., Nino-Liu, D. O., Kundariya, H., Wamboldt, Y., et al. (2014). MSH1-induced non-genetic variation provides a source of phenotypic diversity in Sorghum bicolor. *PLoS ONE.* 9:e108407. doi: 10.1371/journal.pone.0108407

Santerre, A., and Britt, A. B. (1994). Cloning of a 3-methyladenine-DNA glycosylase from Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U.S.A.* 91, 2240–2244. doi: 10.1073/pnas.91.6.2240

Gill, S. S., Anjum, N. A., Gill, R., Jha, M., and Tuteja, N. (2015). DNA damage and repair in plants under ultraviolet and ionizing radiations. *Sci. World J.* 2015:250158. doi: 10.1155/2015/250158

Schalk, C., Cognat, V., Graindorge, S., Vincent, T., Voinnet, O., and Molinier, J. (2017). Small RNA-mediated repair of UV-induced DNA lesions by the DNA damage-binding protein 2 and ARGONAUTE 1. *Proc. Natl. Acad. Sci. U.S.A.* 114, e2965. doi: 10.1073/pnas.1618834114

Schmidt, C., Pacher, M., and Puchta, H. (2019). "DNA break repair in plants and its application for genome engineering," in *Transgenic Plants. Methods in Molecular Biology*, Vol. 1864, eds S. Kumar, P. Barone, and M. Smith (New York, NY: Humana Press), 237–266. doi: 10.1007/978-1-4939-8778-8_17

Screaton, R. A., Kiessling, S., Sansom, O. J., Millar, C. B., Maddison, K., Bird, A., et al. (2003). Fas-associated death domain protein interacts with methyl-CpG binding domain protein 4: a potential link between genome surveillance and apoptosis. *Proc. Natl. Acad. Sci. U S A.* 100, 5211. doi: 10.1073/pnas.0431215100

Selby, C. P., and Sancar, A. (1997). Human transcription-repair coupling factor CSB/ERCC6 is a DNA-stimulated ATPase but is not a helicase and does not disrupt the ternary transcription complex of stalled RNA polymerase II. *J. Biol. Chem.* 272, 1885. doi: 10.1074/jbc.272.3.1885

Selby, C. P., Witkin, E. M., and Sancar, A. (1991). *Escherichia coli* mfd mutant deficient in "mutation frequency decline" lacks strand-specific repair: in vitro complementation with purified coupling factor. *Proc. Natl. Acad. Sci. U S A.* 88, 11574. doi: 10.1073/pnas.88.24.11574

Sengupta, S., Yang, C., Hegde, M. L., Hegde, P. M., Mitra, J., Pandey, A., et al. (2018). Acetylation of oxidized base repair-initiating NEIL1 DNA glycosylase required for chromatin-bound repair complex formation in the human genome increases cellular resistance to oxidative stress. *DNA Repair.* 66, 1. doi: 10.1016/j.dnarep.2018.04.001

Sertic, S., Quadri, R., Lazzaro, F., and Muzi-Falconi, M. (2020). EXO1: A tightly regulated nuclease. *DNA Repair.* 93, 102929. doi: 10.1016/j.dnarep.2020.102929

Shen, H., Strunks, G. D., Klemann, B. J., Hooykaas, P. J., and de Pater, S. (2017). CRISPR/Cas9-induced double-strand break repair in Arabidopsis nonhomologous end-joining mutants. *G3.* 7, 193–202. doi: 10.1534/g3.116.035204

Shevell, D. E., and Walker, G. C. (1991). A region of the Ada DNA-repair protein required for the activation of ada transcription is not necessary for activation of alkA. *Proc. Natl. Acad. Sci. U S A.* 88, 9001. doi: 10.1073/pnas.88.20.9001

Smih, F., Rouet, P., Romanienko, P. J., and Jasin, M. (1995). Double-strand breaks at the target locus stimulate gene targeting in embryonic stem cells. *Nucleic Acid Res.* 3, 5012–5019. doi: 10.1093/nar/23.24.5012

Sonoda, E., and Hochegger, H. (2006). Differential usage of non-homologous end-joining and homologous recombination in double strand break repair. *DNA Repair.* 5, 1021. doi: 10.1016/j.dnarep.2006.05.022

Spampinato, C. P., Gomez, R. L., Galles, C., and Lario, L. D. (2009). From bacteria to plants: a compendium of mismatch repair assays. *Mut. Res* 682, 110–128. doi: 10.1016/j.mrrev.2009.07.001

Tano, K., Shiota, S., Collier, J., Foote, R. S., and Mitra, S. (1990). Isolation and structural characterization of a cDNA clone encoding the human DNA repair protein for O6-alkylguanine. *Proc. Natl. Acad. Sci. U S A.* 87, 686–690. doi: 10.1073/pnas.87.2.686

Taylor, R. M., Hamer, M. J., Rosamond, J., and Bray, C. M. (1998). Molecular cloning and functional analysis of the Arabidopsis thaliana DNA ligase I homologue. *The Plant Journal* 14, 75–81. doi: 10.1046/j.1365-313x.1998.00094.x

Thoma, B. S., and Vasquez, K. M. (2003). Critical DNA damage recognition functions of XPC-hHR23B and XPA-RPA in nucleotide excision repair. *Molecular Carcinogenesis* 38, 1–13. doi: 10.1002/mc.10143

Thorslund, T., Kobbe, C., Harrigan, J. A., Indig, F. E., Christiansen, M., Stevnsner, T., et al. (2005). Cooperation of the Cockayne syndrome group B protein and poly(ADP-ribose) polymerase 1 in the response to oxidative stress. *Mol. Cell Biol.* 25, 7625. doi: 10.1128/MCB.25.17.7625-7636.2005

Tian, H., Gao, Z., Li, H., Zhang, B., Wang, G., Zhang, Q., et al. (2015). DNA damage response–a double-edged sword in cancer prevention and cancer therapy. *Cancer Lett.* 358, 8–16. doi: 10.1016/j.canlet.2014.12.038

Toh, J. D. W., Crossley, S. W. M., Bruemmer, K. J., Ge, E. J., He, D., Iovan, D. A., et al. (2020). Distinct RNA N-demethylation pathways catalyzed by nonheme iron ALKBH5 and FTO enzymes enable regulation of formaldehyde release rates. *Proc. Natl. Acad. Sci. U S A.* 117, 25284. doi: 10.1073/pnas.2007349117

Tornaletti, S. (2005). Data from: Transcription arrest at DNA damage sites. *Mutat. Res.* 577, 131–145. doi: 10.1016/j.mrfmmm.2005.03.014

Townsend, B. J., Poole, A., Blake, C. J., and Llewellyn, D. J. (2005). Antisense suppression of a (+)-δ-cadinene synthase gene in cotton prevents the induction of this defense response gene during bacterial blight infection but not its constitutive expression. *Plant physiol* 138, 516–528. doi: 10.1104/pp.104.056010

Tsurimoto, T., Shinozaki, A., Yano, M., Seki, M., and Enomoto, T. (2005). Human Werner helicase interacting protein 1 (WRNIP1) functions as a novel modulator for DNA polymerase δ. *Genes to Cells.* 10, 13. doi: 10.1111/j.1365-2443.2004.00812

Tuteja, N., Ahmad, P., Panda, B. B., and Tuteja, R. (2009). Genotoxic stress in plants: shedding light on DNA damage, repair and DNA repair helicases. *Mut. Res.* 681, 134–149. doi: 10.1016/j.mrrev.2008.06.004

Uchiyama, Y., Kimura, S., Yamamoto, T., Ishibashi, T., and Sakaguchi, K. (2004). Plant DNA polymerase λ, a DNA repair enzyme that functions in plant meristematic and meiotic tissues. *European Journal of Biochemistry* 271, 2799–2807. doi: 10.1111/j.1432-1033.2004.04214.x

Umate, P., Tuteja, N., and Tuteja, R. (2011). Genome-wide comprehensive analysis of human helicases. *Commun. Integr. Biol.* 4, 118–137. doi: 10.4161/cib.4.1.13844

Van Kregten, M., de Pater, S., Romeijn, R., van Schendel, R., Hooykaas, P. J., and Tijsterman, M. (2016). T-DNA integration in plants results from polymerase-θ-mediated DNA repair. *Nature plants* 2, 1–6. doi: 10.1016/j.dnarep.2012.10.004

Vanderauwera, S., De Block, M., Van de Steene, N., van de Cotte, B., Metzlaff, M., and Van Breusegem, F. (2007). Silencing of poly (ADP-ribose) polymerase in

plants alters abiotic stress signal transduction. *Proc. Natl. Acad. Sci. U S A.* 104, 15150–15155. doi: 10.1073/pnas.0706668104

Vanderauwera, S., Suzuki, N., Miller, G., van de Cotte, B., Morsa, S., Ravanat, J. L., et al. (2011). Extranuclear protection of chromosomal DNA from oxidative stress. *Proc. Natl. Acade. Sci. U.S.A.* 108, 1711–1716. doi: 10.1073/pnas.1018359108

Vashisht, A. A., and Tuteja, N. (2006). Stress responsive DEAD-box helicases, a new pathway to engineer plant stress tolerance. *J. Photochem. Photobiol. B.* 84, 150–160. doi: 10.1016/j.jphotobiol.2006.02.010

Virdi, K. S., Laurie, J. D., Xu, Y. Z., Yu, J., Shao, M. R., Sanchez, R., et al. (2015). Arabidopsis MSH1 mutation alters the epigenome and produces heritable changes in plant growth. *Nature communications.* 6, 1–9.

Völkening, L., Vatselia, A., Asgedom, G., Bastians, H., Lavin, M., Schindler, D., et al. (2020). RAD50 regulates mitotic progression independent of DNA repair functions. *FASEB J.* 34, 2812. doi: 10.1096/fj.201902318

Volkmer, E., and Karnitz, L. M. (1999). Human homologs of Schizosaccharomyces pombe rad1, hus1, and rad9 form a DNA damage-responsive protein complex. *J. Biol. Chem.* 274, 567. doi: 10.1074/jbc.274.2.567

Wallace, S. S. (2014). Base excision repair: a critical player in many games. *DNA Repair.* 19, 14. doi: 10.1016/j.dnarep.2014.03.030

Wang, W., Cheng, Y., Chen, D., Liu, D., Hu, M., Dong, J., et al. (2019). The Catalase Gene Family in Cotton: Genome-Wide Characterization and Bioinformatics Analysis. *Cells.* 8, 86. doi: 10.3390/cells8020086

Wang, X., Wang, H., Guo, B., Zhang, Y., Gong, Y., Zhang, C., et al. (2016). Gen1 and Eme1 Play Redundant Roles in DNA Repair and Meiotic Recombination in Mice. *DNA Cell Biol.* 35, 585. doi: 10.1089/dna.2015.3022

Waterworth, W. M., Bray, C. M., and West, C. E. (2019). Seeds and the art of genome maintenance. *Front. Plant Sci.* 10:706. doi: 10.3389/fpls.2019.00706

Wei, W., Ba, Z., Gao, M., Wu, Y., Ma, Y., Amiard, S., et al. (2012). A role for small RNAs in DNA double-strand break repair. *Cell.* 149, 101–112.

Wielgoss, S., Barrick, J. E., Tenaillon, O., Wiser, M. J., Dittmar, J., Cruveiller, S., et al. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proc. Natl. Acad. Sci. USA.* 110, 222–227. doi: 10.1073/pnas.1219574110

Wilson, D. M., Carney, J. P., Coleman, M. A., Adamson, A. W., Christensen, M., and Lamerdin, J. E. (1998). Hex1: a new human Rad2 nuclease family member with homology to yeast exonuclease 1. *Nucleic Acids Res.* 26, 3762. doi: 10.1093/nar/26.16.3762

Wong, R. H., Chang, I., Hudak, C. S., Hyun, S., Kwan, H. Y., and Sul, H. S. (2009). A role of DNA-PK for the metabolic gene regulation in response to insulin. *Cell* 136, 1056. doi: 10.1016/j.cell.2008.12.040

Wright, T. R., Lira, J. M., Merlo, D. J., and Hopkins, N. (2009). *Novel Herbicide Resistance Genes. U.S. Patent No. 2009/0093366.* ***q.

Yang, C. G., Yi, C., Duguid, E. M., Sullivan, C. T., Jian, X., He, C., et al. (2008). Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. *Nature.* 452, 961. doi: 10.1038/nature06889

Yang, X., Kundariya, H., Xu, Y. Z., Sandhu, A., Yu, J., Hutton, S. F., et al. (2015). MutS HOMOLOG1-derived epigenetic breeding potential in tomato. *Plant Physiol.* 168, 222–232. doi: 10.1104/pp.15.00075

Yao, Y., Andriy, B., Viktor, T., Andrey, G., and Igor, K. (2013). Genome stability of Arabidopsis atm, ku80 and rad51b mutants: somatic and transgenerational responses to stress. *Plant Cell Physiol.* 54, 982–989. doi: 10.1093/pcp/pct051

Yi, C., and He, C. (2013). DNA repair by reversal of DNA damage. *Cold Spring Harb. Perspect. Biol.* 5:a012575. doi: 10.1101/cshperspect.a012575

Zheng, L., Zhou, M., Guo, Z., Lu, H., Qian, L., Dai, H., et al. (2008). Human DNA2 is a mitochondrial nuclease/helicase for efficient processing of DNA replication and repair intermediates. *Mol. Cell.* 32, 325. doi: 10.1016/j.molcel.2008.09.024

Zhou, Z. Q., Zhao, J. J., Chen, C. L., Liu, Y., Zeng, J. X., and Wu, Z. R. (2019). HUS1 checkpoint clamp component (HUS1) is a potential tumor suppressor in primary hepatocellular carcinoma. *Mol. Carcinog.* 58, 76. doi: 10.1002/mc.22908

Zou, L., and Elledge, S. J. (2003). Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. *Science.* 300, 1542. doi: 10.1126/science.1083430

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: frontiersin.org/about/contact

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