ARTIFICIAL INTELLIGENCE BIOINFORMATICS: DEVELOPMENT AND APPLICATION OF TOOLS FOR OMICS AND INTER-OMICS STUDIES

EDITED BY: Angelo Facchiano, Dominik Heider and Davide Chicco

PUBLISHED IN: Frontiers in Genetics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88963-752-2
DOI 10.3389/978-2-88963-752-2

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

ARTIFICIAL INTELLIGENCE BIOINFORMATICS: DEVELOPMENT AND APPLICATION OF TOOLS FOR OMICS AND INTER-OMICS STUDIES

Topic Editors:
Angelo Facchiano, Italian National Research Council, Italy
Dominik Heider, University of Marburg, Germany
Davide Chicco, Krembil Research Institute, Canada

Citation: Facchiano, A., Heider, D., Chicco, D., eds. (2020). Artificial Intelligence Bioinformatics: Development and Application of Tools for Omics and Inter-Omics Studies. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-752-2

Table of Contents

frontiers in Genetics

EDITORIAL
published: 09 April 2020
doi: 10.3389/fgene.2020.00309

Editorial: Artificial Intelligence Bioinformatics: Development and Application of Tools for Omics and Inter-Omics Studies

Davide Chicco [1†‡], *Dominik Heider* [2†‡] *and Angelo Facchiano* [3*†‡]

[1] Krembil Research Institute, University Health Network, Toronto, ON, Canada, [2] Department of Mathematics and Computer Science, Philipps-University of Marburg, Marburg, Germany, [3] Istituto di Scienze dell'Alimentazione, Consiglio Nazionale delle Ricerche (CNR), Avellino, Italy

Keywords: artificial intelligence, bioinformatics, genomics, omics, inter-omics, machine learning, data mining, proteomics

OPEN ACCESS

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

****Correspondence:***
Angelo Facchiano
angelo.facchiano@isa.cnr.it

†ORCID:
Davide Chicco
orcid.org/0000-0001-9655-7142
Dominik Heider
orcid.org/0000-0002-3108-8311
Angelo Facchiano
orcid.org/0000-0002-7077-4912

‡These authors have contributed equally to this work

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

Received: *13 February 2020*
Accepted: *16 March 2020*
Published: *09 April 2020*

Citation:
Chicco D, Heider D and Facchiano A (2020) Editorial: Artificial Intelligence Bioinformatics: Development and Application of Tools for Omics and Inter-Omics Studies. Front. Genet. 11:309. doi: 10.3389/fgene.2020.00309

Editorial on the Research Topic

Artificial Intelligence Bioinformatics: Development and Application of Tools for Omics and Inter-Omics Studies

For half a century, bioinformatics and computational biology have provided tools and data analysis approaches, so the beginning of the omics era represented a novel challenge for researchers, that converged to the area of bioinformatics from the fields of informatics, mathematics, and statistics. In most cases, the solutions offered appeared difficult to use for researchers working in biomedical areas. This occurred in particular when sophisticated approaches from the field of data science and artificial intelligence (AI), were applied to biomedical data (Lisboa et al., 2000).

Machine learning, statistical learning, and soft-computing approaches, such as deep neural networks or genetic algorithms, have also become terms used in the *bio* world, with an incomplete comprehension however, of their potential (Pavel et al., 2016; Lin and Lane, 2017; Zeng and Lumley, 2018). In recent years, omics, multi-omics, and inter-omics experiments have presented a further step toward the investigation in biology, opening the window on personalized medicine, for example for diagnostics (Riemenschneider et al., 2016). The era of big data in medicine is imminent and represents yet a further step forward. Considering this, our Research Topic presents articles on novel developments in the field of artificial intelligence in biology and medicine, and their applications in the analysis of high-throughput data from omics and inter-omics approaches (Facchiano et al.).

1. THE ARTICLE COLLECTION

The Research Topic includes 13 articles:

- 7 Original Research articles (Di Filippo et al.; Kong et al.; Leclercq et al.; Liu et al.; Maj et al.; Simidjievski et al.; Xu et al.)
- 1 Brief Research Report article (Quinn et al.)
- 1 Methods article (Niu et al.)
- 2 Technology and Code articles (Martin and Heider; Wang et al.)
- 1 Review article (M'sch et al.)
- 1 Systematic Review article (Zeng and Bromberg).

The published articles have been evaluated according to each journal editorial policy, by experts of the field. The Research Topic received seven other manuscripts, judged unsuitable for publication and rejected during the review process. The submission deadline was 29th June 2019, therefore any data, experiment, and result presented in the Research Topic articles must be in reference to data, experiments, and results obtained earlier than that date.

1.1. Original Scientific Research and Methods

Simidjievski et al. showed how variational autoencoders (VAEs) can be employed to integrate heterogeneous cancer data. They used these artificial neural networks to integrate multi-omics data such as somatic copy number aberrations (CNA), messenger RNA (mRNA) expressions, and clinical data of patients diagnosed with breast cancer from the METABRIC initiative (Curtis et al., 2012).

Di Filippo et al. developed an R shiny app named HiCeekR that can be used for the analyses of Hi-C data. In contrast to existing tools, HiCeekR represents an easy-to-use graphical user interface to a complete Hi-C data analysis pipeline, including all relevant analysis and visualization steps.

In their article, Niu et al. developed and analyzed a novel pre-training-retraining strategy for deep neural networks and evaluated this strategy based on the prediction of tissue-specific activation of cis-regulatory elements (CREs). This is a very important step as the number of tissue-specific samples is limited. They used all CREs for the pre-training of the net and then used transfer learning to improve tissue-specific predictions.

Maj et al. combined supervised and unsupervised machine learning models on tissue-specific cis-eQTL gene expression data to distinguish mild cognitive impairment and patients with Alzheimer's Disease and to detect potential biological associations.

Kong et al. developed a novel computational model for the prediction of protein-protein interactions (PPIs). The new method, FCTP-WSRC, used a combination of F-vector, composition (C), and transition (T) to numerically encode the protein sequences and subsequently uses principal component analysis (PCA) to extract features. The PCA representation is then used as an input for weighted sparse representation-based classification. FCTP-WSRC has been evaluated on several data sets and shows a superior prediction performance in terms of accuracy and computing time.

Liu et al. used multi-omics data, namely DNA methylation, copy number variation, and gene expression to identify dysfunctional subpathways in cancer and validated their findings with several cancer datasets, for example, liver hepatocellular carcinoma (LIHC), head-neck squamous cell carcinoma (HNSC), cervical squamous cell carcinoma, and endocervical adenocarcinoma.

Xu et al. identified dysregulated competitive endogenous RNA (ceRNA) interactions driven by copy number variation (CNV) in gliomas, and then found their associations with prognosis and histological subtypes by gene set enrichment analysis. Biological functions related to the oncogenesis of malignant gliomas have been detected by the functional analysis of the CNV-driven ceRNA network.

Leclercq et al. proposed BioDiscML, a software program that implements a machine learning method for discovery of biomarkers from multi-omics data. The automatic pipeline built up for mining signatures of diseases by classification, together with the feature selection processes for biomarker discovery, represent the main strengths of this work.

Quinn et al. described an anomaly detector for tissue transcriptomes, aimed to identify cancer without ever seeing a single cancer example. The outlier detection algorithm has been trained on normal samples from a large public data set (Lonsdale et al., 2013) and applied to classify cancer samples from another large public data set (Weinstein et al., 2013).

1.2. Technology Applications

Martin and Heider developed the ContraDRG software, available on a web server, that computationally emulates complex predictions in a reverse-engineering like manner, with intensive calculations using machine learning techniques. ContraDRG can be used to predict partial charges for small molecules based on molecular topology predictions from two commonly used tools, such as PRODRG and ATB. ContraDRG can accurately predict partial charges quickly, and thus can also be applied for screening projects with large amounts of molecules.

Wang et al. used convolutional neural networks to measure conditional relatedness, that is, the degree of the relation of a pair of genes in certain conditions and showed that this approach has a lower false-positive rate compared to traditional co-expression analyses, due to the combination of prior knowledge and co-expression.

1.3. Reviews

In their overview, M'sch et al. reported and described several applications of machine learning methods in immunotherapy, with special attention given to T cell receptor-mediated therapies. They list more than 150 references, which show several data sources and multiple computational intelligence algorithms employed for several goals such as proteasomal cleavage prediction, epitope prediction, and T-cell receptor prediction.

Zeng and Bromberg summarized the recent findings of the functional effects of synonymous mutations in genomes. In particular, they recapped the details and evaluated the performance of nine existing computational methods capable of predicting functional effects for synonymous mutations, also demonstrating the limitations of currently available tools.

2. DISCUSSION

The Research Topic stands out because of its heterogeneity and the diversity of its contents: article authors applied different computational intelligence methods, on different datasets (almost all differing from source and type), to investigate different scientific bioinformatics questions. This diversity confirms the versatility of data mining usage and the huge number of biological subjects that need to be investigated and analyzed.

The Research Topic, in fact, includes original research articles applying statistical learning methods to several dataset types, with gene expression being the most frequent one (Liu et al.; Maj et al.; Quinn et al.; Simidjievski et al.; Wang et al.).

Some authors employed traditional biostatistics techniques, while others took advantage of machine learning methods. In particular, we report the frequent usage of deep learning and artificial neural networks among the applications described in the Research Topic (Leclercq et al.; Maj et al.; Niu et al.; Simidjievski et al.).

The Research Topic articles differ in data and software availability, too. The authors of three articles made their data and software openly public (Maj et al.; Niu et al.; Wang et al.). Two articles have only made their software publicly accessible, but not the data (Leclercq et al.; Simidjievski et al.). The authors of five articles made their datasets available to the scientific community, but not their software (Di Filippo et al.; Kong et al.; Martin and Heider; Quinn et al.; Xu et al.).

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

The Topic Editors thank all the authors and reviewers of the articles submitted to this Frontiers Research Topic.

REFERENCES

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Lin, E., and Lane, H.-Y. (2017). Machine learning and systems genomics approaches for multi-omics data. *Biomark. Res.* 5:2. doi: 10.1186/s40364-017-0082-y

Lisboa, P. J., Ifeachor, E. C., and Szczepaniak, P. S. (2000). *Artificial Neural Networks in Biomedicine*. Berlin: Springer Science & Business Media.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580–585. doi: 10.1038/ng.2653

Pavel, A. B., Sonkin, D., and Reddy, A. (2016). Integrative modeling of multi-omics data to identify cancer drivers and infer patient-specific gene activity. *BMC Syst. Biol.* 10:16. doi: 10.1186/s12918-016-0260-9

Riemenschneider, M., Cashin, K. Y., Budeus, B., Sierra, S., Shirvani-Dastgerdi, E., Bayanolhag, S., et al. (2016). Genotypic prediction of co-receptor tropism of HIV-1 subtypes A and C. *Sci. Rep.* 6:24883. doi: 10.1038/srep24883

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Zeng, I. S. L., and Lumley, T. (2018). Review of statistical learning methods in integrated omics studies (an integrated information science). *Bioinform. Biol. Insights* 12, 1–16. doi: 10.1177/1177932218759292

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 May 2019
doi: 10.3389/fgene.2019.00441

Identification of Cancer Dysfunctional Subpathways by Integrating DNA Methylation, Copy Number Variation, and Gene-Expression Data

Siyao Liu[1]†, Baotong Zheng[1]†, Yuqi Sheng[1]†, Qingfei Kong[2], Ying Jiang[3], Yang Yang[1], Xudong Han[2], Liang Cheng[1], Yunpeng Zhang[1]* and Junwei Han[1]**

[1] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China, [2] College of Basic Medical Science, Harbin Medical University, Harbin, China, [3] College of Basic Medical Science, Heilongjiang University of Chinese Medicine, Harbin, China

A subpathway is defined as the local region of a biological pathway with specific biological functions. With the generation of large-scale sequencing data, there are more opportunities to study the molecular mechanisms of cancer development. It is necessary to investigate the potential impact of DNA methylation, copy number variation (CNV), and gene-expression changes in the molecular states of oncogenic dysfunctional subpathways. We propose a novel method, Identification of Cancer Dysfunctional Subpathways (ICDS), by integrating multi-omics data and pathway topological information to identify dysfunctional subpathways. We first calculated gene-risk scores by integrating the three following types of data: DNA methylation, CNV, and gene expression. Second, we performed a greedy search algorithm to identify the key dysfunctional subpathways within pathways for which the discriminative scores were locally maximal. Finally, a permutation test was used to calculate the statistical significance level for these key dysfunctional subpathways. We validated the effectiveness of ICDS in identifying dysregulated subpathways using datasets from liver hepatocellular carcinoma (LIHC), head-neck squamous cell carcinoma (HNSC), cervical squamous cell carcinoma, and endocervical adenocarcinoma. We further compared ICDS with methods that performed the same subpathway identification algorithm but only considered DNA methylation, CNV, or gene expression (defined as ICDS_M, ICDS_CNV, or ICDS_G, respectively). With these analyses, we confirmed that ICDS better identified cancer-associated subpathways than the three other methods, which only considered one type of data. Our ICDS method has been implemented as a freely available R-based tool (https://cran.r-project.org/web/packages/ICDS).

Keywords: multi-omics data, copy number variation, DNA methylation, subpathway activity, pathway topological information

OPEN ACCESS

Edited by:
Dominik Heider,
University of Marburg, Germany

Reviewed by:
Long Gao,
University of Pennsylvania,
United States
Markus List,
Technische Universität München,
Germany

***Correspondence:**
Liang Cheng
liangcheng@hrbmu.edu.cn
Yunpeng Zhang
zyp19871208@126.com
Junwei Han
hanjunwei1981@163.com

†*These authors have contributed equally to this work*

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

Received: 08 February 2019
Accepted: 29 April 2019
Published: 15 May 2019

Citation:
Liu S, Zheng B, Sheng Y, Kong Q, Jiang Y, Yang Y, Han X, Cheng L, Zhang Y and Han J (2019) Identification of Cancer Dysfunctional Subpathways by Integrating DNA Methylation, Copy Number Variation, and Gene-Expression Data. Front. Genet. 10:441. doi: 10.3389/fgene.2019.00441

INTRODUCTION

Cancer is a complex disease involving multiple biological processes and multiple factors, including genomic, epigenomic, and transcriptomic aberrations associated with cancer formation and development (Forozan et al., 2000; Zhang et al., 2012). Identifying molecular markers of cancer is a major challenge and can effectively clarify diagnosis and treatment. With the development of high-throughput sequencing technology, it is possible to understand the pathogenic mechanisms of cancer at the molecular level (Wang et al., 2014; Liu and Xu, 2015; Zhang et al., 2017). Large-scale cancer genomics projects, such as the Cancer Genome Atlas (TCGA) (Giordano, 2014), provide multi-omics profiles from a large number of patient samples from many cancer types. This may provide a basis for the systematic understanding of the development of cancer. However, both copy number variation (CNV) and DNA methylation changes may affect gene expression, and integration of these data may enhance essential gene characterization in cancer progression (Kim et al., 2010; Xu et al., 2010). Many studies have shown that the use of multi-omics data for integrated analysis helps us to understand the pathogenic mechanisms of cancer. For example, Xu et al. (2010) have shown that the correlation between gene expression and CNV has biological effects on carcinogenesis and cancer progression. Additionally, Zhang et al. (2013) has classified the prognosis of patients with different subtypes of ovarian cancer by integrating four types of molecular data related to gene expression. In view of these works, our goal is to explore the multi-layered genetic and epigenetic regulatory mechanisms of cancer.

Biological pathways are models containing structural information between genes, such as interactions, regulation, modifications, and binding properties. In addition, genes in the same pathway usually coordinately achieve a particular function. With the appearance of some traditional pathway-analysis tools, such as GSEA (Subramanian et al., 2005) and SPIA (Tarca et al., 2009), the pathway-based approach has become the first choice for complex disease analysis to facilitate biological insights. Existing biological-pathway databases provide pathway topological information, such as with the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Wixon and Kell, 2000), which is being updated to suit the needs for practical applications and act a systematic reference knowledge database to understand the metabolism and other cellular processes. Recently, the KEGG pathway database has become one of most widely used resource for biological function annotation (Kanehisa et al., 2017).

Based on pathway topological information, the subpathway concept was proposed in our previous study in which we confirmed that key subpathways – rather than entire pathways – were more suitable for explaining the etiology of diseases (Li et al., 2009, 2013). Subpathways contain fewer components, which enables a more accurate interpretation of the biological function of the disturbance, for the future study of precision medicine. Subpathway-GM (Li et al., 2013) was proposed to identify disease-relevant subpathways by integrating information across genes, metabolites, and pathway structural information within a given pathway; using this, 16 statistically significant subpathways were identified as associated with metastatic prostate cancer. SubpathwayMiner (Li et al., 2009) uses a subgraph-mining method to find subpathways where all of the genes have highly similar functions; this method identified36 dysfunctional subpathways – enriched by differentially expressed genes – as associated with the initiation or progression of lung cancer. Recently, some other methods have been developed to identify subpathways from pathway topology. One example is MIDAS (Lee et al., 2017), which determines condition-specific subpathways and fully utilizes quantitative gene-expression data and network-centrality information across multiple phenotypes. Moreover, the following subpathway-activity measurement tools have been designed to identify activated subpathways between two phenotypes: PATHOME (pathway and transcriptome information) (Nam et al., 2014), TEAK (Topology Enrichment Analysis frameworK) (Judeh et al., 2013), and MinePath (Mining for Phenotype Differential Sub-paths in Molecular Pathways) (Koumakis et al., 2016). Moreover, there is also some other methods proposed network-based analysis to discover *de novo* pathway. For instance, *de novo* pathway enrichment extracted sub-networks enriched in biological entities active by combining experimental data with a large-scale interaction network (Batra et al., 2017). These subpathway-analysis methods mainly identify dysfunctional subpathways only by comparing the expression levels of their involved genes between tumor and normal tissues. In this way, other genetic characterizations of genes, such as CNVs and DNA methylation, are ignored. However, both DNA methylation and CNVs in cancer genomes frequently perturb the expression levels of affected genes and, thus, disrupt pathways controlling normal growth. It is therefore necessary to integrate gene expression information and other genetic information, such as DNA methylation and CNVs, to identify dysfunctional subpathways.

In this study, we propose a novel method, termed Identification of Cancer Dysfunctional Subpathways (ICDS), to identify dysfunctional subpathways by integrating multi-omics data and pathway topological information. In ICDS, the first step is to calculate gene-risk scores to evaluate the contribution of genes to cancer states by considering the following three molecular characterizations: DNA methylation, CNV, and gene expression. In the second step, we convert the KEGG pathway into an undirected-pathway network with genes as nodes and biological relationships as edges, and use a greedy search algorithm to search for candidate dysfunctional subpathways within the pathways for which the discriminative scores are locally maximal. Finally, a perturbation test is used to calculate statistical significance for these dysfunctional subpathways. We applied the ICDS method to liver hepatocellular carcinoma (LIHC), head-neck squamous cell carcinoma (HNSC), and cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) datasets, and compared our results with three analytical methods that only used DNA methylation, CNV, or gene expression to calculate subpathway-activity scores (defined as ICDS_G, ICDS_CNV, ICDS_M, respectively). Through these analyses, we confirmed that ICDS could better identify cancer-associated subpathways compared to the other three methods.

MATERIALS AND METHODS

Datasets

The datasets containing gene expression, CNV, and DNA methylation information were collected from the TCGA website[1]. We downloaded TCGA RNA-seq level-3 data, which were processed and normalized and used the Reads Per Kilobase per Million mapped reads (RPKM) values for the gene-expression levels. Finally, there were 19,754 genes used in 424 LIHC, 546 HNSC, and 309 CESC samples. CNV profiling was estimated using the GISTIC2 method (Mermel et al., 2011), and was annotated to genes using the UCSC cgData HUGO probeMap. For example, the LIHC dataset contains CNVs in 24,776 genes from 373 cancer samples. In this study, we further filtered 364 LIHC samples with matched gene-expression profiles.

We downloaded TCGA level-3 Illumina HumanMethylation450 Bead Array data for DNA methylation. The LIHC DNA methylation level-3 dataset contain β-values for 20,105 genes from 429 samples, which included 50 normal samples and 379 lung-cancer samples. The β-values are calculated by $M/(M+U+100)$ with a range from 0 to 1, in which M is methylated allele frequencies and U is unmethylated allele frequencies. Overall, higher β-values indicate higher methylation. For three datasets, we removed genes with values of zero in more than 80% of the samples. In this paper, we also use the data from HNSC and CESC samples, which were processed using the above procedure. Detailed data information is shown in **Supplementary Table S1**.

The KEGG pathway database contains experimentally verified pathway structural information (e.g., interactions, regulation, modifications, and binding between genes). We collected 294 KEGG biological pathways, and each pathway was converted to an undirected network with genes as nodes and biological relationships as edges on the basis of pathway structural information using the "iSubpathwayMiner" system (Li et al., 2009, 2013).

Calculated Gene Risk Score in Cancer

There are many factors influencing tumorigenesis, such as gene expression, CNV and DNA methylation. For each gene, we calculated its risk score in cancer by considering the following three types of genetic molecular features: gene expression, CNV, and DNA methylation. With the above data, we used the Student's t-test (Hogben, 1964) to calculate the adjusted p-value for differential expression level and differential methylation level of each gene in the tumor and normal samples (denoted by p_{gene} and p_{methy}). According to results of GISTIC2 analysis, the sample was then divided into a copy-number-variated group and an un-variated group for each gene, and then the differential expression level of the gene in the two groups was calculated by Student's t-test (denoted by p_{cnv}). It is difficult to define the quantitative relationship and relative degree of each factor's influence on tumorigenesis, so we assume that gene expression, CNV, and DNA Methylation equally contribute to the cancer development. The gene risk score (RS) was calculated by integrating the above three p-values with Fisher's combined probability test. This method computed a combined statistic S from the adjusted p-values obtained from the three individual datasets as shown in Equation (1). Usually, the statistic S followed a χ^2 distribution with $2k$ degrees of freedom, and we then calculated the null hypothesis p-value of the statistic S. Finally, we converted the p-value to a z-score according to the inverse-normal cumulative-distribution function (CDF), and the z-score was taken as the RS of each gene in cancer.

$$S = -2\log\prod_{m} p_m, \quad m = gene, cnv, methy \tag{1}$$

Calculated Subpathway-Activity Score

Previous studies have confirmed that subpathways can provide more detailed biological information than whole pathways. In this study, we proposed a novel method to combine gene-risk score with pathway topological structure to infer subpathway activities. The RS of genes were obtained by the above method, considering gene expression, CNV and methylation. For a KEGG pathway, we performed a greedy algorithm to search for dysfunctional subpathways within the pathways for which the discriminative scores were locally maximal. Specifically, the search algorithm started from a seed gene i which had a significantly high risk score ($p < 0.001$) and expanded iteratively, after which it selected one of the neighbors of the seed gene to form the current subpathway. For a subpathway k, the activity score (AS_k) was the average of the RS of the member genes in the subpathway, calculated by Equation (2):

$$AS_k = \sum_i \frac{RS_i}{\sqrt{n}} \tag{2}$$

In Equation (2), i is the index of the gene in the subpathway k, while n is the number of genes involved in the subpathway. At each iteration, the algorithm adopted a gene from the neighbors of genes in the current subpathway, which produced maximal increases between AS_{k+1} and AS_k. The search algorithm will stops when no additional gene increases in the score AS_{k+1} over $(1+r)\ AS_k$ or the distance in the current subpathway between any two nodes is greater than 3 in order to keep the search locally. The improvement rate r is chosen to avoid too large subpathway region, resulting in the addition of redundant weak information. The parameter $r = 0.05$ has been demonstrated to be appropriate in the greedy heuristic algorithm applied in the biological network (Chuang et al., 2007). When the Jaccard index between each pair of subpathways in the same pathway was more than 0.6, they were combined, which ensured that the subpathways we found in our method contained more information and reduce redundancy. Furthermore, we only considered subpathways with more than five genes and less than 100 genes, to avoid overly narrow or broad functional subpathways.

Significance Test of the Subpathway

We provided two statistical test methods for each candidate subpathway, of which one was a whole gene-based perturbation, and the other was a local-gene perturbation in a particular

[1] https://tcga-data.nci.nih.gov/tcga/

pathway. Users can choose the test method that they prefer. The first test perturbs the gene labels on the entire gene list in the pathway networks, and recalculates the activity score of the subpathway, denoted as AS_{k_perm1}. This test was used to test the correlation between real subpathways and disease phenotype. In this study, we performed 10,000 perturbations for this test and calculated statistically significant p-value = M/N, in which M is the number of AS_{k_perm1} greater than the real subpathway score AS_k, and N is the number of perturbations. In addition, the second test perturbed the gene names in the pathway to which the subpathway belonged, and recalculated the activity score of the subpathway, denoted as AS_{k_perm2}. This test was used to test the correlation between real subpathways and pathway structure. We also performed 10,000 perturbations and the score of each real AS_k was indexed on the null distribution of all AS_{k_perm2} whose p-values could be evaluated. The p-values were adjusted using the false discovery rate (FDR) method proposed by Benjamini and Hochberg to correct for multiple comparisons (Benjamini et al., 2001). In this study, both FDR at 0.001 was used as the subpathway-significance threshold. We have implemented ICDS as an R-based package that is publicly available on CRAN[2].

RESULTS

Analyses of Hepatocellular Carcinoma Data

A workflow diagram of the ICDS is shown in **Figure 1**. We first applied ICDS to identify dysfunctional subpathways in LIHC. The LIHC dataset was obtained from TCGA, and its detailed information is shown in **Supplementary Table S1**. In the LIHC dataset, we calculated the risk score of 16,207 genes by considering the following three types of genetic molecular features: gene expression, CNV, and DNA methylation. We set the genes with $p < 0.001$ (derived from the combined statistic S) as the seed genes in the pathway network for the subpathway search algorithm (see Materials and Methods). Subpathways were selected which satisfied two permutation tests with FDR1 < 0.001 and FDR2 < 0.001 out of the 10,000 permutations. ICDS identified 19 dysfunctional subpathways associated with LIHC, belonging to 12 entire pathways (**Table 1** and **Supplementary Table S2**), of which up to nine were reported to be associated with tumor occurrence, development, and metastasis.

The most significant subpathway was path 00230_1 in purine metabolism, which contained 61 genes. Some studies have confirmed that the purine-metabolism pathway is highly correlated with the occurrence and metastasis of liver cancer. In multiple cancer cells, a marked imbalance in the enzymic pattern of purine metabolism is linked with transformation or progression, such as in kidney, liver, and colon carcinomas (Weber, 1983). The subregion corresponding to the subpathway included 61 genes (**Supplementary Figure S1A**), such as adenosine monophosphate deaminase 1 (AMPD1) and adenosine kinase (ADK), which are important enzymes involved in purine metabolism. ADK plays a significant role in affecting apoptosis and may become a target for the treatment of cancer (Dzeja et al., 1998). More evidence is mounting regarding the direct relationship between defects in ADK and AMP metabolic signaling (e.g., AMPD) and human diseases (Pavlova and Thompson, 2016), which is a set of collaborative interactions that converts adenosine monophosphate (AMP) to inosine monophosphate (IMP) as part of the process of the purine nucleotide cycle. Compared with normal hepatocytes, the levels of ADK and AMPD1 in LIHC cells were significantly different in expression and methylation (p_{gene} = 6.58e-05 of ADK and p_{gene} = 0.0042 of AMPD1; p_{methy} = 1.05e-05 of ADK and p_{methy} = 9.48e-12 of AMPD1) (**Supplementary Figure S1B**). The abnormality of ADK and AMPD1 changes the metabolic homeostasis of cells and promotes the progression of cancer cells (Pedley and Benkovic, 2017).

To assess the effectiveness of ICDS, we compared our results in LIHC with three other analytical methods in which we calculated the RS of genes by considering only one of the following types of data: gene expression, CNV, or DNA methylation (defined as ICDS-G, ICDS-CNV, or ICDS-M, respectively). Next, we used the same procedure as above to find significant subpathways and used the same parameter settings. Using the methods of ICDS-G and ICDS-M, we obtained three and one significant subpathways, respectively, and the entire pathways they belonged to were all found by the ICDS method (**Table 1**). Using the method ICDS-CNV method, we could not find any significant subpathway. If we consider the genetic differences or expression differences based on a single type of data, we may lose important information. However, ICDS exclusively identified 15 significant subpathways marked with red asterisks in **Figure 2A**, and the KEGG pathways they belong to could not be found based on the three other methods. Some pathways identified by ICDS were the chemokine signaling pathway and focal adhesion, which have been reported to be related to the occurrence and development of hepatocellular carcinoma (Zhao et al., 2011). It has been reported in the literature that the chemokine signaling pathway is involved in the establishment of a tumor-promoting microenvironment and in the development and progression of hepatobiliary cancer (Zlotnik and Yoshie, 2000). Drug targeting of the chemokine pathway is a promising approach for the treatment or even prevention of hepatobiliary cancer. Chemokines play a vital role in tumor progression and metastasis, where the binding of chemokines to their receptors leads to a conformational change, which activates signaling pathways and promotes migration (Zhao et al., 2011). Meanwhile, the subpathway path:04062_1 in the chemokine signaling pathway (**Figure 2B**), exclusively identified by ICDS, included the chemokine family (CC or CXC) and its receptors family (CCR or CXCR). All of these chemokine families exert their biological effects by binding to chemokine receptors that interact with G protein-linked transmembrane receptors (Decaillot et al., 2011). In the subpathway path:04062_1 (**Figure 3A**), the CXC motif chemokine 12 (CXCL12) is a chemokine protein that is differentially expressed between LIHC and normal samples (p_{gene} = 1.53e-35), and both the expression of CCL20 and CCR2 are regulated by differential methylation

[2] https://cran.r-project.org/web/packages/ICDS

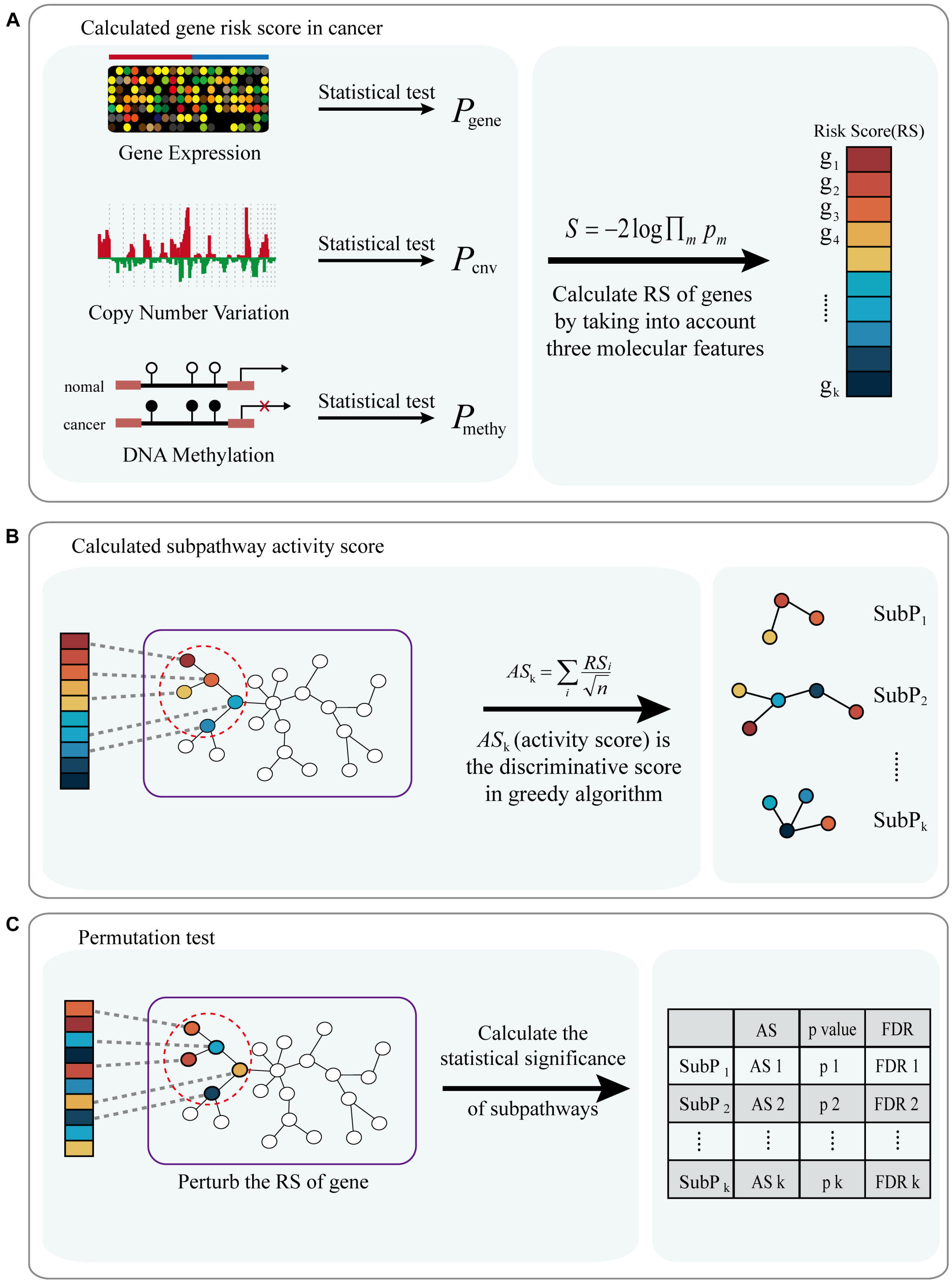


FIGURE 1 | Flow diagram of ICDS methodology. **(A)** Calculated risk score of genes (*RS*) in cancer by considering three types of genetic molecular features: gene expression, CNV and DNA methylation. **(B)** Combine gen-risk score with pathway topological structure to infer the subpathway activity score (AS); subpathways with discriminative activity score in cancer were identified via a greedy search algorithm. **(C)** A permutation test is performed on the risk score of genes, and pathways are prioritized by FDR after permutation tests.

(p_{methy} = 3.07e-18, 2.3e-16). Importantly, the ICDS method not only recognized subregions of differential gene expressions but also detected some genetically or epigenetically diverse regions (e.g., CNVs and methylations). Another subpathway of the chemokine signaling pathway was path:04062_4, which contains 9 genes (**Figure 3B**). We found that four of these genes were mainly influenced by differential expressions and five were mainly influenced by differential methylation. Thus, our method can efficiently find dysfunctional local regions in biological pathways and indicate their perturbation by deriving specific

TABLE 1 | Subpathways identified by ICDS with FDR < 0.001 in the LIHC dataset.

SubpathID	Pathway	Size*	FDR1	FDR2	ICDS-G	ICDS-CNV	ICDS-M
path:00230_1	Purine metabolism	61	<E-11	9.13E-11	√		
path:00240_1	Pyrimidine metabolism	51	<E-11	1.76E-07	√		
path:04380_1	Osteoclast differentiation	13	<E-11	3.29E-06			√
path:00830_1	Retinol metabolism	23	<E-11	3.29E-06			
path:04062_1	Chemokine signaling pathway	24	<E-11	3.46E-06			
path:04510_10	Focal adhesion	8	<E-11	3.46E-06			
path:04152_1	AMPK signaling pathway	24	<E-11	6.34E-06	√		
path:05166_1	HTLV-I infection	16	<E-11	6.34E-06			
path:04062_4	Chemokine signaling pathway	9	<E-11	9.45E-06			
path:00240_3	Pyrimidine metabolism	7	<E-11	1.23E-05			
path:04062_7	Chemokine signaling pathway	10	<E-11	1.31E-05			
path:04110_10	Cell cycle	8	<E-11	2.10E-05			
path:04110_11	Cell cycle	9	<E-11	3.13E-05			
path:04630_4	Jak-STAT signaling pathway	5	<E-11	3.43E-05			
path:00240_2	Pyrimidine metabolism	7	<E-11	3.75E-05			
path:00240_4	Pyrimidine metabolism	8	<E-11	6.61E-05			
path:00230_4	Purine metabolism	10	<E-11	1.10E-04			
path:04110_1	Cell cycle	25	<E-11	1.85E-04			
path:04114_1	Oocyte meiosis	28	<E-11	9.42E-04			

**The number of genes in the subpathway.*

types of molecular aberrations (CNV, differential methylations or differential gene expressions).

Analyses of Head-Neck Squamous Cell Carcinoma Data

The HNSC datasets were obtained from TCGA, and their detailed information is shown in **Supplementary Table S1**. ICDS identified 17 significant dysfunctional subpathways associated with HNSC belonging to 9 entire pathways and the subpathways exclusively identified by the ICDS method are marked with red asterisks in **Figure 4A** (**Table 2**), of which up to eight have been reported to be central to the growth and survival of cancer cells. Subpathways were selected that satisfied two tests with FDR1 < 0.001 and FDR2 < 0.001 (see Materials and Methods).

Path:04919_4 is a significant subpathways (**Figure 4B** and **Supplementary Table S3**) belonging to the thyroid hormone signaling pathway (**Figure 4C**). Many studies have confirmed that the thyroid hormone signaling pathway is a critical component in tumor progression (Kim and Cheng, 2013). Loss of normal function of thyroid-hormone receptors by deletion or mutation can contribute to cancer development, progression and metastasis. Thyroid Hormone Receptor Alpha (THRA) belongs to the nuclear receptor superfamily, is located on different chromosomes, and encodes thyroid hormone (T3) binding thyroid hormone receptor (TR) isoforms, which have been shown to mediate the biological activities of cells (Laudet et al., 1993; Wagner et al., 1995). TRs can function as tumor suppressors, because reduced expression of TRs due to hypermethylation or deletion of TR genes is found in human cancers. The samples had significantly different methylation of THRA (p_{methy} = 4.79e-12) in HNSC, and low expression of THRA is known to activate PIK3R1, which provides instructions for synthesizing a subunit of phosphatidylinositol 3-kinase (PI3K). PI3K signaling is important for many cell activities, including cell growth, division, and migration (Jaiswal et al., 2009). However, we calculated the *RS* of PIK3R1in HNSC, and its contributions with differential methylation were greater than that of differential expression (p_{methy} = 4.78e-12; p_{gene} = 1.46e-06) (**Figure 4B**).

Similarly, we compared the results of HNSC with the three methods above (ICDS-G, ICDS-CNV, and ICDS-M). Using the methods of ICDS-G and ICDS-M, we obtained two significant subpathways and the pathways they belonged to were also found by the ICDS method. However, 13 subpathways identified by ICDS were missing from all of the other methods (ICDS-G, ICDS-CNV, and ICDS-M) (**Table 2**). For example, the subpathway path:00830_3 in retinol metabolism pathway was identified by ICDS but not by ICDS-G, ICDS-CNV, or ICDS-M, and **Supplementary Figures S3**, **S4** show the distribution of the activity score of path:00830_3, combined and individual data source, for the real subpathways and for the randomization cases. The local region of the subpathways was reported to be central to the growth and survival of cancer cells (**Supplementary Figure S2A**). Specifically, vitamin A (retinol) can control mucosal lesions before the occurrence of HNSC and prevent the occurrence of second primary tumors. Therefore, retinol metabolism is essential for the early diagnosis and treatment of HNSC. Retinoic acid (RA) is a critical signaling molecule that regulates gene transcription and the cell cycle (Tzimas and Nau, 2001), and retinal is then metabolized by NAD/NADP-dependent retinal dehydrogenases (RALDH) and by retinal oxidase enzymes to RA (Chen et al., 1995). Additionally, CYP26C1 in the path:00830_3 is involved in the metabolic breakdown of retinoic acid, which could be more effective in the growth inhibition

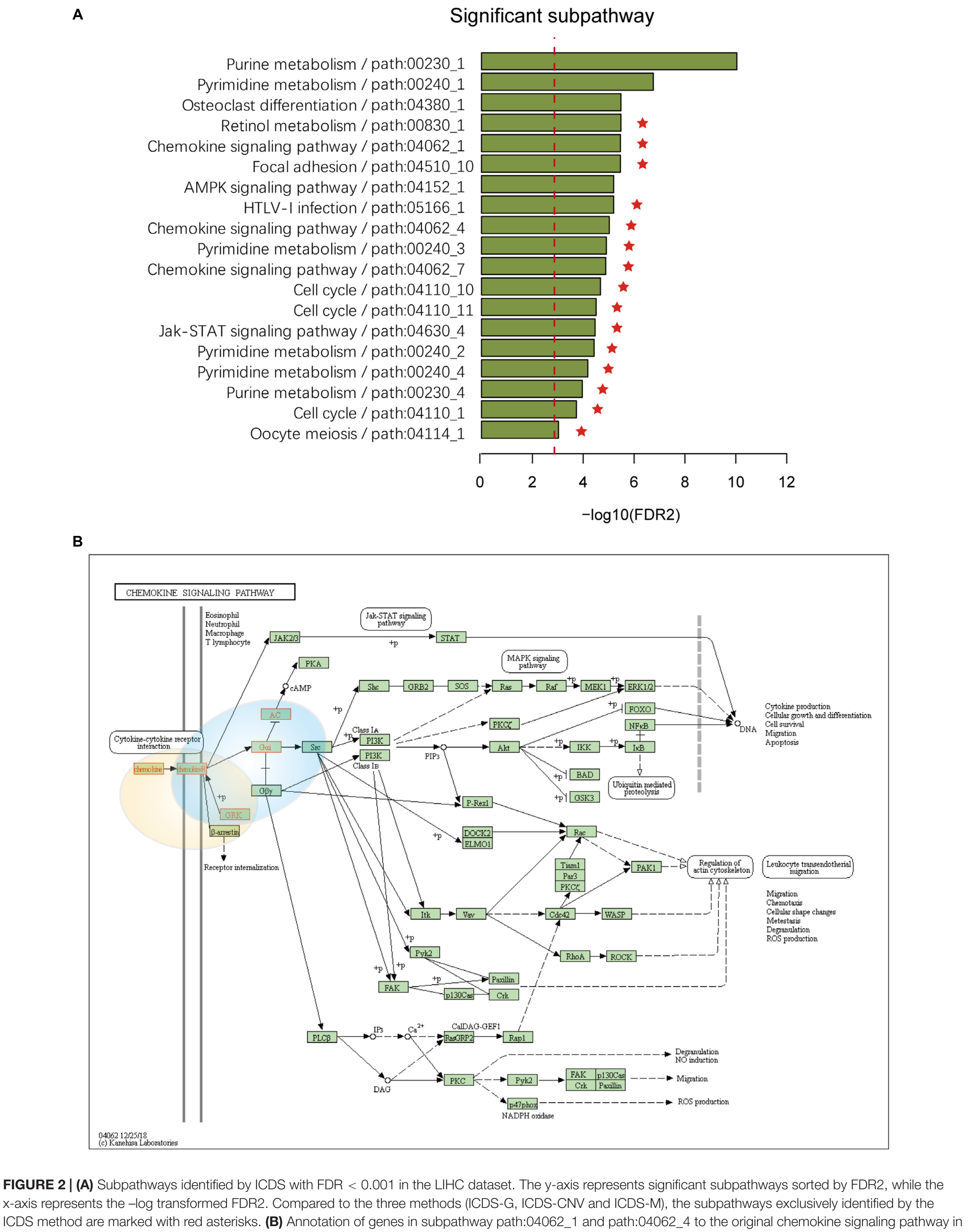


FIGURE 2 | (A) Subpathways identified by ICDS with FDR < 0.001 in the LIHC dataset. The y-axis represents significant subpathways sorted by FDR2, while the x-axis represents the –log transformed FDR2. Compared to the three methods (ICDS-G, ICDS-CNV and ICDS-M), the subpathways exclusively identified by the ICDS method are marked with red asterisks. **(B)** Annotation of genes in subpathway path:04062_1 and path:04062_4 to the original chemokine signaling pathway in KEGG. Genes are marked with red, and the light-yellow circle corresponds to subpathway path:04062_1 and the blue circle to subpathway path:04062_4.

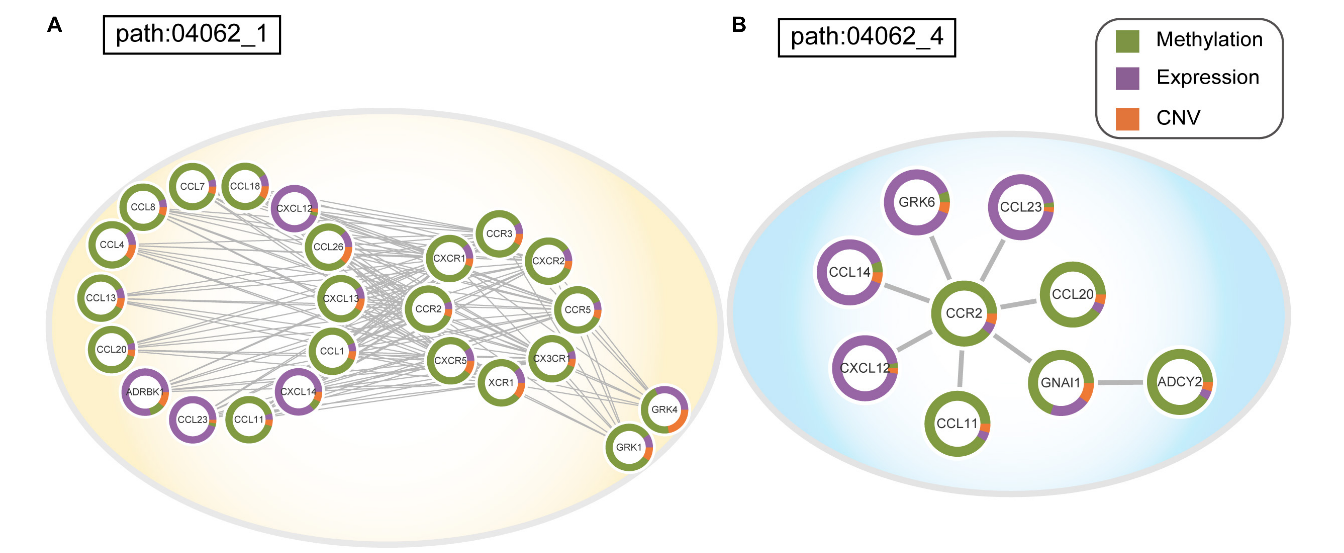


FIGURE 3 | (A) Dysfunctional subpathway (path:04062_1) of chemokine signaling pathway in LIHC. **(B)** Dysfunctional subpathway (path:04062_4) of chemokine signaling pathway in LIHC. The vertex in the subnetwork represents a gene, and green and purple colors in the vertex represent the proportion of the gene's differential expression scores and differential methylation scores between cancer samples and normal samples; orange colors represent the proportion of influence of CNV on gene expression.

of cancer cells (Thatcher and Isoherranen, 2009). Moreover, in the HNSC dataset, some genes mainly showed differences in the degree of methylation compared to normal samples, such as CYP26C1 (p_{methy} = 9.25e-34) and ALDH1A2 (p_{methy} = 1.65e-13). Other components in the same subpathway, path: 00830_3, mainly showed differences in the degree of expression compared to normal samples, such as AOX1 (p_{gene} = 3.11e-18) and ADH4 (p_{gene} = 2.75e-38) (**Supplementary Figure S2B**). Therefore, the ICDS method that we proposed can effectively identify disordered genetic and epigenetic subpathways.

Analyses of Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma Data

We applied ICDS to identify dysfunctional subpathways in CESC (see Materials and Methods). With the threshold of FDR1 < 0.001, we obtained four significant subpathways that had just exceeded the threshold FDR2 (**Supplementary Table S4**), and all of these subpathways were associated with the development and progression of CESC tumors. Meanwhile, using the method of ICDS-G, we obtained three significant subpathways, and the pathways they belonged to were also found by the ICDS method (**Supplementary Tables S4, S5**). Subpathway 04020_1 in the calcium-signaling pathway, identified by ICDS, was simultaneously neglected by the other three methods.

Interestingly, subpathway 04020_1 (**Figure 5A**) in the calcium-signaling pathway is involved many G-protein coupled receptors (GPCRs), such as TACR1, TACR2, and HTR2B, and downstream heterotrimeric guanine nucleotide-binding proteins (G-proteins; GNA14) (**Figure 5B**). In this subpathway, many GPCRs had significant patterns of expression changes in CESC patients, such as TACR1 (p_{gene} = 9.92e-32), TACR2 (p_{gene} = 3.82e-08), and HTR2B (p_{gene} = 2.76e-26). Moreover, with CESC samples, AVPR1A, which is a GPCR in cells, mainly showed differences in methylation and expression compared to normal samples. Many studies have shown that the abnormal expression and activity of GPCRs is associated with the development and progression of cancers (Audigier et al., 2013; Moody et al., 2016). GPCRs play a role as key transducers of signals from the extracellular milieu to the intracellular milieu of cells. It has been confirmed that many GPCRs are highly expressed in specific cancer cells, such as in cervical, breast, and prostate cancer cells (Dey et al., 2010). Similarly, abnormal expression of GPCRs contributes to the development of cancer (Radhika and Dhanasekaran, 2001; O'Hayre et al., 2013). Furthermore, initial signal transduction, such as that of calcium signaling, is achieved primarily by GPCRs activated downstream of heterotrimeric G proteins (Hanlon and Andrew, 2015; Schafer and Blaxall, 2017). Calcium-signaling channels are important for the proliferation, migration, and differentiation of cells, including tumors. CESC is associated with the significant upregulation of calcium-signaling pathways (Perez-Plasencia et al., 2007; Monteith et al., 2012).

Comparison of ICDS With Other Pathway Analysis Methods

In recent years, the pathway-based and subpathway-base approaches have become the first choice for complex disease analysis in order to yield biological insight. To explore whether ICDS could provide new biological insights in identifying important subpathways, we compared ICDS with three widely used pathway-based and subpathway-base approaches including SPIA (Tarca et al., 2009), GSEA (Subramanian et al., 2005), and SubpathwayMiner (Li et al., 2009). These three methods

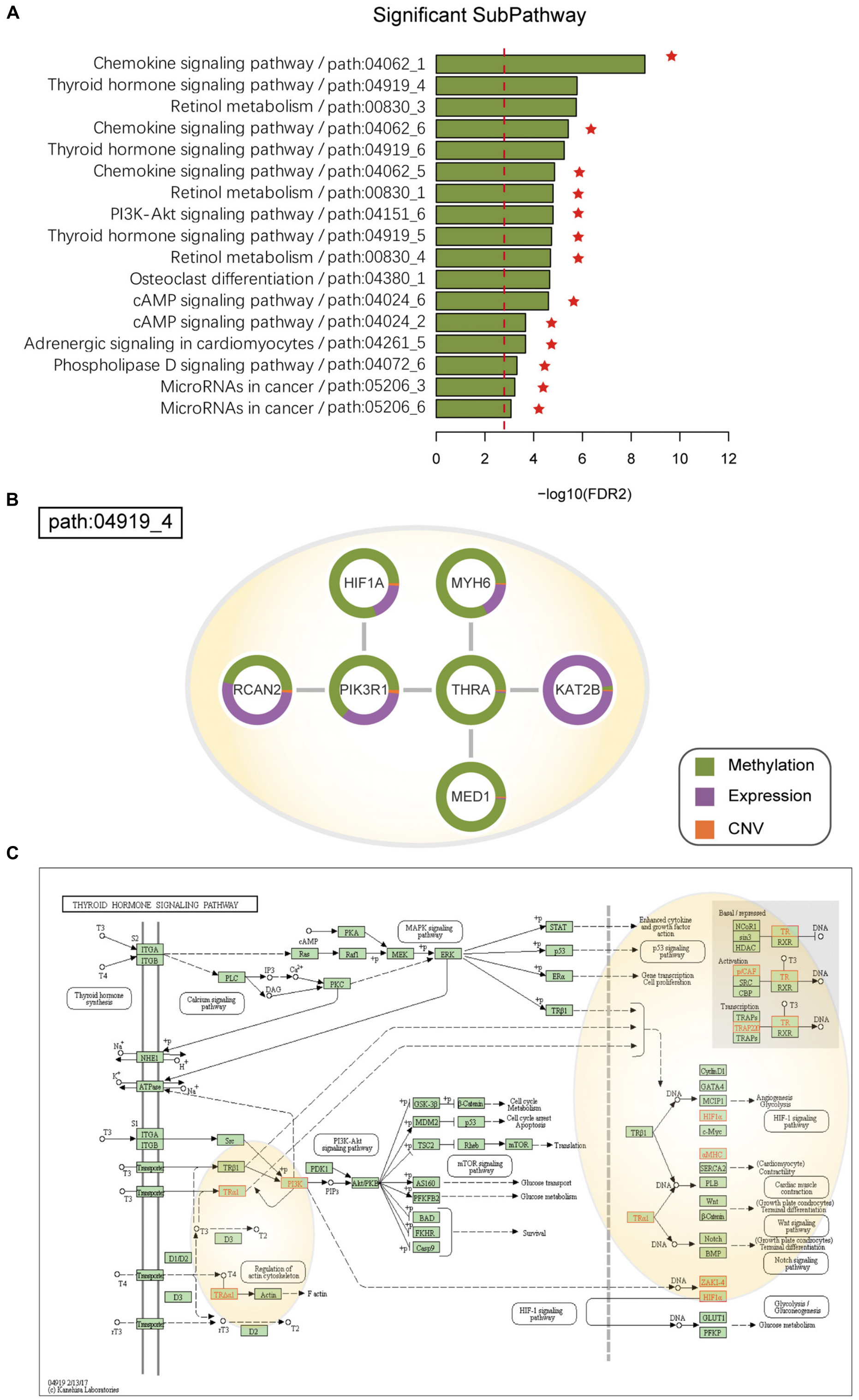


FIGURE 4 | (A) SubPathways identified by ICDS with FDR < 0.001 in the HNSC dataset. The y-axis represents significant subpathways sorted by FDR2, while the x-axis represents the log-transformed FDR2. Compared to the three methods (ICDS-G, ICDS-CNV, and ICDS-M), the subpathways exclusively identified by ICDS method are marked with red asterisks. **(B)** Dysfunctional subpathway (path:04919_4) of thyroid hormone signaling pathway in HNSC. The vertex in the subnetwork represents a gene, and green and purple colors in the vertex represent the proportion of the gene's differential expression scores and differential methylation scores between cancer samples and normal samples; orange colors represent the proportion of influence of CNV on gene expression. **(C)** Annotation of genes in path:04919_4 to the original thyroid hormone signaling pathway in KEGG. Genes are marked with red, and the light-yellow circle corresponds to path:04919_4.

TABLE 2 | Subpathways identified by ICDS with FDR <0.001 in the HNSC dataset.

SubpathID	Pathway	Size*	FDR1	FDR2	ICDS-G	ICDS-CNV	ICDS-M
path:04062_1	Chemokine signaling pathway	41	<E-30	2.73E-09	√		
path:04919_4	Thyroid hormone signaling pathway	7	<E-30	1.67E-06			√
path:00830_3	Retinol metabolism	11	<E-30	1.79E-06			
path:04062_6	Chemokine signaling pathway	10	<E-30	3.82E-06	√		
path:04919_6	Thyroid hormone signaling pathway	5	<E-30	5.59E-06			
path:04062_5	Chemokine signaling pathway	8	<E-30	1.40E-05			
path:00830_1	Retinol metabolism	17	<E-30	1.60E-05			
path:04151_6	PI3K-Akt signaling pathway	10	<E-30	1.60E-05			
path:04919_5	Thyroid hormone signaling pathway	9	<E-30	1.86E-05			
path:00830_4	Retinol metabolism	7	<E-30	2.06E-05			
path:04380_1	Osteoclast differentiation	15	<E-30	2.21E-05			√
path:04024_6	cAMP signaling pathway	9	<E-30	2.48E-05			
path:04024_2	cAMP signaling pathway	11	<E-30	2.17E-04			
path:04261_5	Adrenergic signaling in cardiomyocytes	6	<E-30	2.20E-04			
path:04072_6	Phospholipase D signaling pathway	5	<E-30	4.90E-04			
path:05206_3	MicroRNAs in cancer	5	<E-30	6.0E-04			
path:05206_6	MicroRNAs in cancer	5	<E-30	8.50E-04			

**The number of genes in the subpathway.*

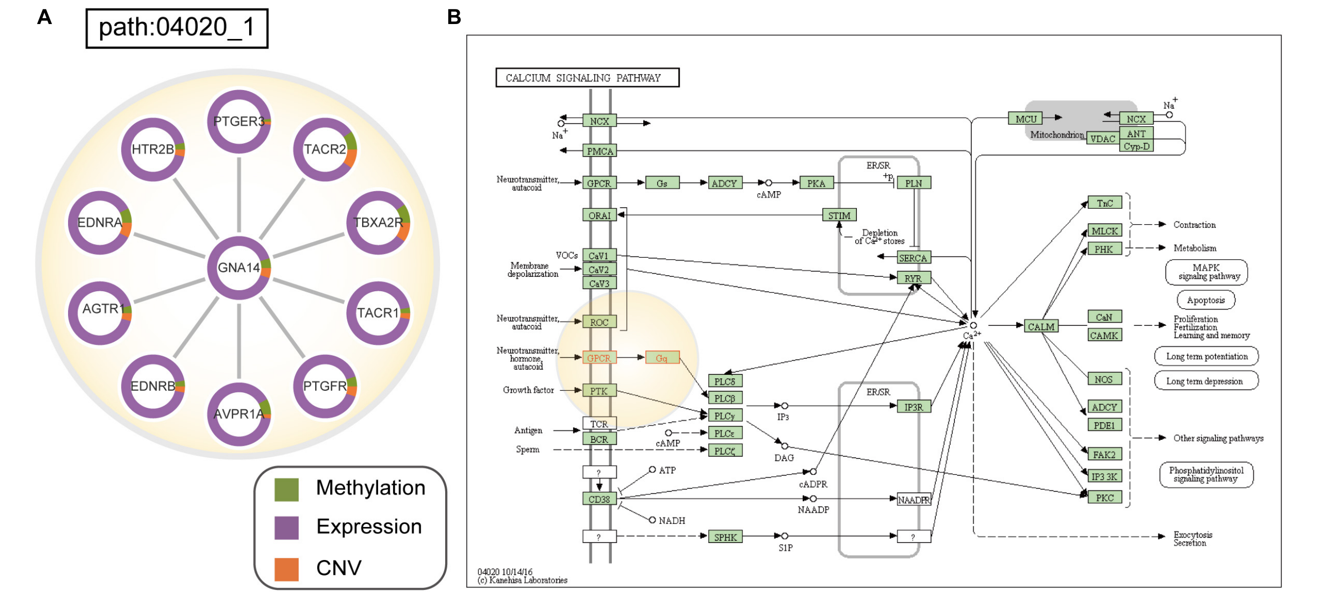


FIGURE 5 | (A) Dysfunctional subpathway (path:04020_1) of calcium signaling pathway in CESC. The vertex in the subnetwork represents a gene, and green and purple colors in the vertex represent the proportion of the gene's differential expression scores and differential methylation scores between cancer samples and normal samples; orange colors represent the proportion of the influence of CNV on gene expression. **(B)** Annotation of genes in path:04020_1 to the original calcium signaling pathway in KEGG. Genes are marked with red, and the light-yellow circle corresponds to path:04020_1.

mainly identify dysregulated pathways or subpathways by using gene expression data, however, the ICDS method identifies the dysregulated subpathways by integrating the three types of data: DNA methylation, CNV, and gene expression. In order to compare the results of the above methods uniformly, we chose to compare the entire pathways identified by them. In HNSC datasets, ICDS identified 17 statistically significant subpathways, which belong to nine entire pathways. SPIA and GSEA found five and eight significant pathways, and SubpathwayMiner did not yield any significant pathways. Through comparing the results of these methods, we found that ICDS identified six statistically significant pathways, which were simultaneously missed by other methods (**Supplementary Table S6**). The significant pathways exclusively identified by ICDS, such as the cAMP signaling pathway, chemokine signaling pathway, Retinol metabolism etc., have been well reported to be associated with the development of HNSC (Tzimas and Nau, 2001; Tanaka et al., 2005). For example, the thyroid hormone signaling pathway and retinol metabolism were reported to be central to the growth and survival of cancer cells. A subpathway of Retinol metabolism identified by ICDS methods (**Supplementary Figure S2A**) is essential for the early diagnosis and treatment of HNSC. These results indicate that the ICDS method may uncover something new dysregulated subpathways.

DISCUSSION

The occurrence and development of diseases, especially cancer, involves a complex biological network (Zou et al., 2016). Genetic variation, epigenetic changes, abnormal gene-expression levels, and many other factors will change in the characteristics of living organisms. With the generation of large-scale sequencing data, more opportunities exist to study the multi-omics molecular mechanisms of cancer development. In systems biology, dysfunctional genes may jointly activate biological pathways. Therefore, the most critical step in exploring complex disease mechanisms is to identify the functional pathways in which these dysregulated genes are located. We proposed the concept of subpathways in our previous work (Li et al., 2009, 2013). The subpathway, defined as a local region of an entire pathway, contains fewer components, which enables a more subtle and accurate interpretation of the biological function of disturbances involved in cancer progression.

In this study, the employed method was based on *a priori* biological pathways (e.g., KEGG), each of which represents a network of interactions between genes, proteins, and chemical molecules. The main purpose of this study was to discover important dysfunctional subregions based on the pathway topological structure. ICDS used Fisher's combined probability test to integrate gene expression, CNVs, and methylation to calculate the *RS* of genes. As the gene expression, CNV and DNA methylation are not completely independent, and thus the independence assumption of Fisher's combined probability test may be violated here. This may be a limitation of our ICDS method. Alternatively, the Brown's method (Poole et al., 2016) can also be used to integrate multiple data source, and it does not suffer from this limitation. A larger *RS* in cancer indicated a greater correlation between the gene and the cancer phenotype. Next, we used a greedy algorithm to search for subpathways in each KEGG pathway network, so that subpathway activities were local maxima. This algorithm have also been used to identified subnetwork markers of breast cancer metastasis in the human protein–protein interaction network previously, and achieved higher accuracy in the classification of metastatic versus non-metastatic tumors (Chuang et al., 2007). To avoid excessive redundancy in the candidate subpathways, we set several parameters, such as seed gene (p-value of combined statistic $S < 0.001$), subpathway size ($5 < \text{size} < 100$), and overlap between subpathways (i.e., Jaccard index between each pair of subpathways in the same pathway < 0.6), which can be set by a user of the ICDS package.

We applied the ICDS method to LIHC, HNSC, and CESC datasets. Based on these analyses, we demonstrated that ICDS can effectively identify dysfunctional subpathways correlated with a cancer phenotype. For the HNSC dataset, the subpathway path:04062_1 was the most significant subpathway and included 41 genes belonging to chemokine signaling pathway. Studies have confirmed that the chemokine signaling pathway is a critical component of tumor progression. These genes did not simultaneously have changes in copy number, methylation, and gene expression. However, these subregions could still be found through our integration algorithm, which is the most prominent advantage of our method. To further validate the ICDS method, we compared it with three other methods that only considered one type of data – gene expression, CNV, or DNA methylation – named as ICDS-G, ICDS-CNV, and ICDS-M, respectively. The results showed that the ICDS method was able to identify new risk subpathways associated with cancer that were simultaneously neglected by the other three methods. Thus, it is essential to integrate multi-omics data to identify additional dysfunctional subpathways in cancer. In the future, we will involve other omics data, such as proteomics, to improve our ICDS method.

To provide users with convenient and simple analytical tools, we have integrated the ICDS, ICDS-G, ICDS-CNV, and ICDS-M methods into an available R-based package on CRAN[3]. If users are considering using the ICDS method, they need to input three datasets of gene expression, copy number, and methylation. The ICDS-package will produce a prioritized list of subpathways. With this method, ICDS is used to find key subpathways related to cancer phenotypes, and it is expected that it can be used to mine for key subnetworks within some prior networks (e.g., the PPI network) based on integrating DNA methylation, CNV, and gene expression data. In addition, ICDS may identify key subpathways as biomarkers to distinguish high and low risk cancer patients. For this purpose, researchers should input the molecular profile of genes with different stage samples, such as patients in different stages of glioma. Therefore, we have developed a free and robust tool to identify dysfunctional subpathways in cancer by integrated multi-omics data.

AUTHOR CONTRIBUTIONS

JH, YZ, and LC conceived and designed the study. SL and BZ developed the software. YY analyzed the data and implemented the methodology. YJ revised the manuscript. YZ provided constructive discussions. JH and LC drafted the manuscript. All the authors read and agreed to the manuscript.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 31401127 and 81804158), the Science and Technology Innovation Talent

[3] https://cran.r-project.org/web/packages/ICDS

Research Foundation of Harbin (Grant No. 2017RAQXJ195), and the National Natural Science Foundation of Heilongjiang Province (Grant No. H2016074).

ACKNOWLEDGMENTS

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript. We also thank the student innovation and entrepreneurship education training center of Harbin Medical University.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00441/full#supplementary-material

REFERENCES

Audigier, Y., Picault, F. X., Chaves-Almagro, C., and Masri, B. (2013). G protein-coupled receptors in cancer: biochemical interactions and drug design. *Prog. Mol. Biol. Transl. Sci.* 115, 143–173. doi: 10.1016/B978-0-12-394587-7.00004-X

Batra, R., Alcaraz, N., Gitzhofer, K., Pauling, J., Ditzel, H. J., Hellmuth, M., et al. (2017). On the performance of de novo pathway enrichment. *NPJ Syst. Biol. Appl.* 3:6. doi: 10.1038/s41540-017-0007-2

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). Controlling the false discovery rate in behavior genetics research. *Behav. Brain Res.* 125, 279–284. doi: 10.1016/s0166-4328(01)00297-2

Chen, H., Namkung, M. J., and Juchau, M. R. (1995). Biotransformation of all-trans-retinol and all-trans-retinal to all-trans-retinoic acid in rat conceptal homogenates. *Biochem. Pharmacol.* 50, 1257–1264. doi: 10.1016/0006-2952(95)02005-w

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., and Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3:140. doi: 10.1038/msb4100180

Decaillot, F. M., Kazmi, M. A., Lin, Y., Ray-Saha, S., Sakmar, T. P., and Sachdev, P. (2011). CXCR7/CXCR4 heterodimer constitutively recruits beta-arrestin to enhance cell migration. *J. Biol. Chem.* 286, 32188–32197. doi: 10.1074/jbc.M111.277038

Dey, S., Hablas, A., Seifeldin, I. A., Ismail, K., Ramadan, M., El-Hamzawy, H., et al. (2010). Urban-rural differences of gynaecological malignancies in Egypt (1999-2002). *BJOG* 117, 348–355. doi: 10.1111/j.1471-0528.2009.02447.x

Dzeja, P. P., Zeleznikar, R. J., and Goldberg, N. D. (1998). Adenylate kinase: kinetic behavior in intact cells indicates it is integral to multiple cellular processes. *Mol. Cell Biochem.* 184, 169–182. doi: 10.1007/978-1-4615-5653-4_13

Forozan, F., Mahlamaki, E. H., Monni, O., Chen, Y., Veldman, R., Jiang, Y., et al. (2000). Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data. *Cancer Res.* 60, 4519–4525.

Giordano, T. J. (2014). The cancer genome atlas research network: a sight to behold. *Endocr. Pathol.* 25, 362–365. doi: 10.1007/s12022-014-9345-4

Hanlon, C. D., and Andrew, D. J. (2015). Outside-in signaling–a brief review of GPCR signaling with a focus on the Drosophila GPCR family. *J. Cell Sci.* 128, 3533–3542. doi: 10.1242/jcs.175158

Hogben, C. A. (1964). A practical and simple equivalent for student's T test of statistical significance. *J. Lab. Clin. Med.* 64, 815–819.

Jaiswal, B. S., Janakiraman, V., Kljavin, N. M., Chaudhuri, S., Stern, H. M., Wang, W., et al. (2009). Somatic mutations in p85alpha promote tumorigenesis through class IA PI3K activation. *Cancer Cell* 16, 463–474. doi: 10.1016/j.ccr.2009.10.016

Judeh, T., Johnson, C., Kumar, A., and Zhu, D. (2013). TEAK: topology enrichment analysis framework for detecting activated biological subpathways. *Nucleic Acids Res.* 41, 1425–1437. doi: 10.1093/nar/gks1299

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kim, H., Huang, W., Jiang, X., Pennicooke, B., Park, P. J., and Johnson, M. D. (2010). Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. *Proc. Natl. Acad. Sci. U.S.A.* 107, 2183–2188. doi: 10.1073/pnas.0909896107

Kim, W. G., and Cheng, S. Y. (2013). Thyroid hormone receptors and cancer. *Biochim. Biophys. Acta* 1830, 3928–3936. doi: 10.1016/j.bbagen.2012.04.002

Koumakis, L., Kanterakis, A., Kartsaki, E., Chatzimina, M., Zervakis, M., Tsiknakis, M., et al. (2016). MinePath: mining for phenotype differential sub-paths in molecular pathways. *PLoS Comput. Biol.* 12:e1005187. doi: 10.1371/journal.pcbi.1005187

Laudet, V., Vanacker, J. M., Adelmant, G., Begue, A., and Stehelin, D. (1993). Characterization of a functional promoter for the human thyroid hormone receptor alpha (c-erbA-1) gene. *Oncogene* 8, 975–982.

Lee, S., Park, Y., and Kim, S. (2017). MIDAS: mining differentially activated subpaths of KEGG pathways from multi-class RNA-seq data. *Methods* 124, 13–24. doi: 10.1016/j.ymeth.2017.05.026

Li, C., Han, J., Yao, Q., Zou, C., Xu, Y., Zhang, C., et al. (2013). Subpathway-GM: identification of metabolic subpathways via joint power of interesting genes and metabolites and their topologies within pathways. *Nucleic Acids Res.* 41:e101. doi: 10.1093/nar/gkt161

Li, C., Li, X., Miao, Y., Wang, Q., Jiang, W., Xu, C., et al. (2009). SubpathwayMiner: a software package for flexible identification of pathways. *Nucleic Acids Res.* 37:e131. doi: 10.1093/nar/gkp667

Liu, Z., and Xu, J. H. (2015). The application of the high throughput sequencing technology in the transposable elements. *Yi Chuan* 37, 885–898. doi: 10.16288/j.yczz.15-140

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12:R41. doi: 10.1186/gb-2011-12-4-r41

Monteith, G. R., Davis, F. M., and Roberts-Thomson, S. J. (2012). Calcium channels and pumps in cancer: changes and consequences. *J. Biol. Chem.* 287, 31666–31673. doi: 10.1074/jbc.R112.343061

Moody, T. W., Nuche-Berenguer, B., Nakamura, T., and Jensen, R. T. (2016). EGFR transactivation by peptide G protein-coupled receptors in cancer. *Curr. Drug Targets* 17, 520–528. doi: 10.2174/1389450116666150107153609

Nam, S., Chang, H. R., Kim, K. T., Kook, M. C., Hong, D., Kwon, C. H., et al. (2014). PATHOME: an algorithm for accurately detecting differentially expressed subpathways. *Oncogene* 33, 4941–4951. doi: 10.1038/onc.2014.80

O'Hayre, M., Vazquez-Prado, J., Kufareva, I., Stawiski, E. W., Handel, T. M., Seshagiri, S., et al. (2013). The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. *Nat. Rev. Cancer* 13, 412–424. doi: 10.1038/nrc3521

Pavlova, N. N., and Thompson, C. B. (2016). The emerging hallmarks of cancer metabolism. *Cell Metab.* 23, 27–47. doi: 10.1016/j.cmet.2015.12.006

Pedley, A. M., and Benkovic, S. J. (2017). A new view into the regulation of purine metabolism: the purinosome. *Trends Biochem. Sci.* 42, 141–154. doi: 10.1016/j.tibs.2016.09.009

Perez-Plasencia, C., Vazquez-Ortiz, G., Lopez-Romero, R., Pina-Sanchez, P., Moreno, J., and Salcedo, M. (2007). Genome wide expression analysis in HPV16 cervical cancer: identification of altered metabolic pathways. *Infect. Agent Cancer* 2:16. doi: 10.1186/1750-9378-2-16

Poole, W., Gibbs, D. L., Shmulevich, I., Bernard, B., and Knijnenburg, T. A. (2016). Combining dependent P-values with an empirical adaptation of Brown's method. *Bioinformatics* 32, i430–i436. doi: 10.1093/bioinformatics/btw438

Radhika, V., and Dhanasekaran, N. (2001). Transforming G proteins. *Oncogene* 20, 1607–1614. doi: 10.1038/sj.onc.1204274

Schafer, A. E., and Blaxall, B. C. (2017). G protein coupled receptor-mediated transactivation of extracellular proteases. *J. Cardiovasc. Pharmacol.* 70, 10–15. doi: 10.1097/FJC.0000000000000475

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Tanaka, T., Bai, Z., Srinoulprasert, Y., Yang, B. G., Hayasaka, H., and Miyasaka, M. (2005). Chemokines in tumor progression and metastasis. *Cancer Sci.* 96, 317–322. doi: 10.1111/j.1349-7006.2005.00059.x

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., et al. (2009). A novel signaling pathway impact analysis. *Bioinformatics* 25, 75–82. doi: 10.1093/bioinformatics/btn577

Thatcher, J. E., and Isoherranen, N. (2009). The role of CYP26 enzymes in retinoic acid clearance. *Expert. Opin. Drug Metab. Toxicol.* 5, 875–886. doi: 10.1517/17425250903032681

Tzimas, G., and Nau, H. (2001). The role of metabolism and toxicokinetics in retinoid teratogenesis. *Curr. Pharm. Des.* 7, 803–831. doi: 10.2174/1381612013397708

Wagner, R. L., Apriletti, J. W., McGrath, M. E., West, B. L., Baxter, J. D., and Fletterick, R. J. (1995). A structural role for hormone in the thyroid hormone receptor. *Nature* 378, 690–697. doi: 10.1038/378690a0

Wang, Q., Wei, L., Guan, X., Wu, Y., Zou, Q., and Ji, Z. (2014). Briefing in family characteristics of microRNAs and their applications in cancer research. *Biochim. Biophys. Acta* 1844, 191–197. doi: 10.1016/j.bbapap.2013.08.002

Weber, G. (1983). Enzymes of purine metabolism in cancer. *Clin. Biochem.* 16, 57–63. doi: 10.1016/s0009-9120(83)94432-6

Wixon, J., and Kell, D. (2000). The kyoto encyclopedia of genes and genomes–KEGG. *Yeast* 17, 48–55. doi: 10.1002/(sici)1097-0061(200004)17:1<48::aid-yea2>3.0.co;2-h

Xu, C., Liu, Y., Wang, P., Fan, W., Rue, T. C., Upton, M. P., et al. (2010). Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. *Mol. Cancer* 9:143. doi: 10.1186/1476-4598-9-143

Zhang, S., Liu, C. C., Li, W., Shen, H., Laird, P. W., and Zhou, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.* 40, 9379–9391. doi: 10.1093/nar/gks725

Zhang, T., Tan, P., Wang, L., Jin, N., Li, Y., Zhang, L., et al. (2017). RNALocate: a resource for RNA subcellular localizations. *Nucleic Acids Res*. 45, D135–D138. doi: 10.1093/nar/gkw728

Zhang, W., Liu, Y., Sun, N., Wang, D., Boyd-Kirkup, J., Dou, X., et al. (2013). Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. *Cell Rep.* 4, 542–553. doi: 10.1016/j.celrep.2013.07.010

Zhao, X., Ning, Q., Sun, X., and Tian, D. (2011). Pokemon reduces Bcl-2 expression through NF-kappa Bp65: a possible mechanism of hepatocellular carcinoma. *Asian Pac. J. Trop. Med.* 4, 492–497. doi: 10.1016/S1995-7645(11)60133-8

Zlotnik, A., and Yoshie, O. (2000). Chemokines: a new classification system and their role in immunity. *Immunity* 12, 121–127. doi: 10.1016/s1074-7613(00)80165-x

Zou, Q., Li, J., Song, L., Zeng, X., and Wang, G. (2016). Similarity computation strategies in the microRNA-disease network: a survey. *Brief. Funct. Genomics* 15, 55–64. doi: 10.1093/bfgp/elv024

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 16 May 2019
doi: 10.3389/fgene.2019.00452

Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data

Mickael Leclercq[1,2]*, *Benjamin Vittrant*[1,2], *Marie Laure Martin-Magniette*[3,4], *Marie Pier Scott Boyer*[1,2], *Olivier Perin*[5], *Alain Bergeron*[1,6], *Yves Fradet*[1,6] and *Arnaud Droit*[1,2]*†

[1] Centre de Recherche du CHU de Québec–Université Laval, Québec City, QC, Canada, [2] Département de Médecine Moléculaire, Université Laval, Québec City, QC, Canada, [3] Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay, Paris Diderot, Sorbonne Paris-Cité, Orsay, France, [4] UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, Paris, France, [5] Digital Sciences Department, L'Oréal Advanced Research, Aulnay-sous-bois, France, [6] Département de Chirurgie, Oncology Axis, Université Laval, Québec City, QC, Canada

OPEN ACCESS

Edited by:
Angelo Facchiano,
Italian National Research Council (CNR), Italy

Reviewed by:
Zhi-Ping Liu,
Shandong University, China
Barbara Di Camillo,
University of Padova, Italy

***Correspondence:**
Mickael Leclercq
mickael.leclercq@gmail.com
Arnaud Droit
arnaud.droit@crchudequebec.ulaval.ca

†Present Address:
Arnaud Droit,
Computational Biology Laboratory,
Centre de Recherche du CHU de Québec – Université Laval,
Québec City, QC, Canada

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

Received: *23 January 2019*
Accepted: *30 April 2019*
Published: *16 May 2019*

Citation:
Leclercq M, Vittrant B, Martin-Magniette ML, Scott Boyer MP, Perin O, Bergeron A, Fradet Y and Droit A (2019) Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. Front. Genet. 10:452. doi: 10.3389/fgene.2019.00452

The identification of biomarker signatures in omics molecular profiling is usually performed to predict outcomes in a precision medicine context, such as patient disease susceptibility, diagnosis, prognosis, and treatment response. To identify these signatures, we have developed a biomarker discovery tool, called BioDiscML. From a collection of samples and their associated characteristics, i.e., the biomarkers (e.g., gene expression, protein levels, clinico-pathological data), BioDiscML exploits various feature selection procedures to produce signatures associated to machine learning models that will predict efficiently a specified outcome. To this purpose, BioDiscML uses a large variety of machine learning algorithms to select the best combination of biomarkers for predicting categorical or continuous outcomes from highly unbalanced datasets. The software has been implemented to automate all machine learning steps, including data pre-processing, feature selection, model selection, and performance evaluation. BioDiscML is delivered as a stand-alone program and is available for download at https://github.com/mickaelleclercq/BioDiscML.

Keywords: machine learning, omics, biomarkers signature, feature selection, precision medicine

INTRODUCTION

The identification of biomarkers that are indicative of a specific biological state is a major research topic in biomedical applications of computational biology (Liu et al., 2014; Beerenwinkel et al., 2016; Zhang et al., 2017). With the emergence of high-throughput molecular profiling technologies and their decreasing costs, traditional medicine is moving to precision medicine to improve disease diagnosis, and to propose tailored interventions to individuals. Research studies involving cohorts of patients aim to discover patterns that establish risk stratification and discriminate patient states, such as diseased vs. controls, disease type, etc. These last years, clinical and biology research turned toward extensive usage of OMICs (i.e., proteomics, transcriptomics, metabolomics, genomics, etc.) technologies, which include microarrays, mass spectrometry, and whole exome/genome and RNA sequencing. Specific patterns associated with a clinical outcome of interest (e.g., disease diagnostic, prognostic), called biomarker signatures, can be derived from these high-dimensional technologies outputs (e.g., gene expression, polymorphisms) (Lin et al., 2017).

These signatures, which are measurable indicators for predicting a biological phenomenon, are usually identified using machine learning (Pasolli et al., 2016) or statistical multivariate analysis approaches (Rohart et al., 2017c).

Biomarker signature identification from disease-derived omics datasets is a challenging task involving many pitfalls. First, the datasets are generally highly unbalanced, where the features (e.g., genes, peptides, metabolites…), also called attributes or variables, largely outnumber the samples. In addition, patients are unequally distributed among measured outcomes. Second, the molecular profiles are often heterogeneous (e.g., subphenotypes in cancer data), of diverse types (e.g., categorical, continuous), and scattered over multiple inputs (Libbrecht and Noble, 2015). To identify sets of predictive biomarker signatures from omics data, a few non-commercial methods have been implemented in R packages (Lê Cao et al., 2009; Taverner et al., 2012; Cun and Fröhlich, 2014; Rohart et al., 2017b). These toolkits have adopted diverse multivariate projection-based methods including principal component analysis (Wold, 1975), independent component analysis (Yao et al., 2012), multi-group partial least squares regression (Eslami et al., 2013), canonical correlation analysis (Hotelling, 1936), K-means clustering (Hartigan and Wong, 1979), and associated visualizations. Recently, other research teams have proposed approaches in machine learning (ML) (Janevski et al., 2009; Cun and Fröhlich, 2013; Lagani et al., 2013; Swan et al., 2013, 2015; Butti et al., 2014; Kong et al., 2014; Kourou et al., 2015), a branch of artificial intelligence that holds a great potential for pattern recognition in complex diseases datasets. ML has already shown its ability to identify key features (markers) and modeling predictive biomarker signature in a variety of fields, including cancer research (Matsumura et al., 2010; Cima et al., 2011; Cui et al., 2011; Roth et al., 2011; Fröhlich and Cun, 2012; Kourou et al., 2015), neurology (Daoqiang and Dinggang, 2012; Deshpande et al., 2013; Fekete et al., 2013), immunology (Sutherland et al., 2011), skin diseases (Johansson et al., 2011), etc. However, all these techniques are complex to use and are out-of-reach for non-programmers and non-ML experts. Furthermore, the software implemented specifically for omics data are still rare and are strictly limited to specific ML algorithms for feature selection (also called "attribute selection") or classification (Butti et al., 2014). Hence, there is an unmet need to develop user-friendly computational approaches for using machine learning in a biomedical context that are dedicated to biologists and clinical researchers. These approaches must be able to identify complex patterns and predict outcomes in various biological or clinical fields (e.g., disease diagnosis, prognosis, therapeutics), thus helping to understand the biology behind a measured outcome.

Considering the complexity of the ML approach, we present in this paper a software called BioDiscML (*Bio*marker *Disc*overy by *M*achine *L*earning), which aims to greatly facilitate the work required for biomarker signature identification from high-dimensional data, such as gene expression, by automating the ML approach. Some non-commercial automatic software already exists to facilitate the choice of learning algorithms and perform hyper-parameter optimization, such as Auto-weka (Thornton et al., 2013), auto-Sklearn (Feurer et al., 2015), autoML (Feurer et al., 2015), and preconfigured pipelines in Orange canvas (Demšar et al., 2013). But they are not explicitly designed to answer biological problems, lack of user-friendly experience for non-ML experts, some focusing only on hyperparameter optimization, and may be complex to parallelize to decrease calculation time. We aim here to fill the gap, providing BioDiscML the capacity to test large number of feature subsets and models in order to obtain the most performant signature to predict a measured outcome. BioDiscML uses an exhaustive search approach, which systematically enumerates a pre-defined set of possible candidates for a solution and test whether each candidate satisfies the problem statement. BioDiscML can also merge files from different sources, search for the most predictive combination of feature subsets and machine learning classifiers, train a model, evaluate predictive performances, parallelize the computation, and search for correlated features.

MATERIALS AND METHODS

BioDiscML is a tool that automates main ML steps by implementing methods for feature and model selection. In this section, we describe the program procedures separated in three main components: preprocessing, feature selection and model selection. We also present all supported models (see **Supplementary Materials**), evaluation metrics, feature search methods, best model selection and correlated features search approaches. Finally, we have summarized the real-life datasets we used to compare BioDiscML against various existing tools.

BioDiscML Software

BioDiscML is a biomarker discovery software that supports classification (categorical class) and regression (numerical class) problems. It is written in JAVA 8 language (Fischer, 2015) and use Weka 3.8 machine learning library (Holmes et al., 1994; Hall et al., 2009; Witten et al., 2016). It automates several machine learning steps aiming to identify predictive models. To this purpose, BioDiscML can routinely perform data preprocessing, features dimension reduction, a combined feature and model selection strategy, identify best models, and search correlated features. All machine learning generated models are evaluated by various cross validation procedures. All steps are configured with editable default parameters. Advanced parameters can also be modified by the user. Some basic information is needed to start the program such as: input dataset(s), class label name, problem type (regression or classification).

BioDiscML pipeline presented in **Figure 1** works as follows: It starts with the preprocessing section. After merging the input datasets when many are submitted, a first sampling step separates the data in a train and a test set (2/3 and 1/3, respectively, by default), this latter will be used after model creation to assess non-overfitting. Then, a feature ranking algorithm sorts the features based on their predictive power with respect to the class. Only the first best 1,000 s features are kept by default. Then, in the feature selection section, for each machine learning algorithm defined in BioDiscML (i.e., the classifiers), and for each optimization evaluation criterion (i.e., a chosen evaluation metric), two types

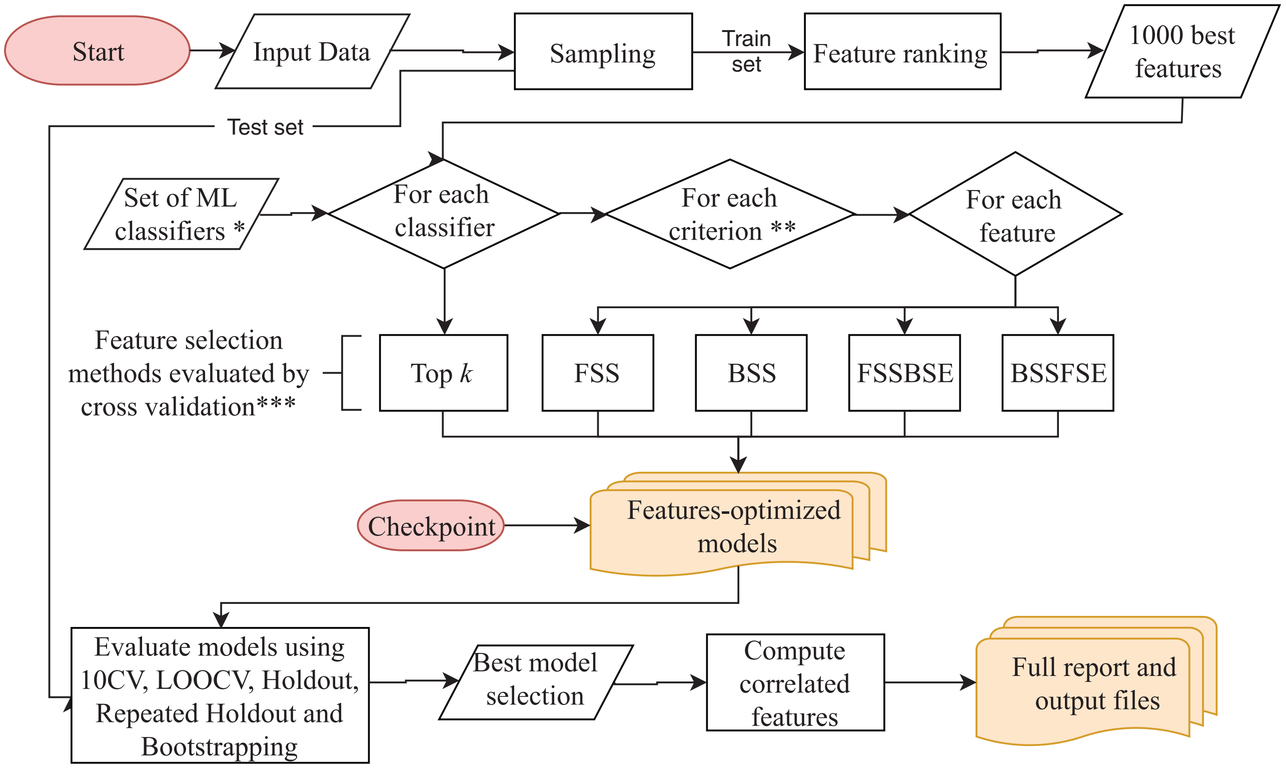

Start
Input Data
Sampling
Train set
Feature ranking
1000 best features
Test set
Set of ML classifiers *
For each classifier
For each criterion **
For each feature
Feature selection methods evaluated by cross validation***
Top k
FSS
BSS
FSSBSE
BSSFSE
Checkpoint
Features-optimized models
Evaluate models using 10CV, LOOCV, Holdout, Repeated Holdout and Bootstrapping
Best model selection
Compute correlated features
Full report and output files

FIGURE 1 | BioDiscML pipeline. Preprocessing and feature selection procedures are fully parallelizable, When all features-optimized models are computed, the model selection starts. The program can be also started from the checkpoint at any moment during the execution. *The Set of ML classifiers is the set of pre-configured commands in classifiers.conf file. All classifiers are listed in the **Supplementary Table S1**. **Criterions are optimized metrics, evaluated by 10-folds cross validation (10 CV), used to assess if a model is improved, such as accuracy, balanced error rate, Matthew's correlation coefficient, area under the curve, sensitivity, specificity, Root Mean Squared Error, etc. (see Evaluation Criterion). ***Feature selection methods include forward stepwise selection (FSS), backward stepwise selection (BSS), forward stepwise selection and backward stepwise elimination (FSSBSE), backward stepwise selection, and Forward stepwise elimination (BSSFSE), and "top *k*" features (see Optimal Feature Subset Search Methods).

of feature search selection are performed: top *k* features and stepwise (see Optimal Feature Subset Search Methods). Top *k* simply select the best *k* elements from the ordered feature set to create a model. In the stepwise approaches, for each element in the ordered set, features are added and/or removed one by one depending on the feature search method. At each iteration, the created model is evaluated by 10-fold cross validation (10 CV) and the combination of selected features is retained if the predictive performance is improved. When all features are tested and the signature is identified, the model is evaluated on other cross-validation/sampling procedures (see Model Evaluation). Once all classifiers are tested, we end with a set of feature-optimized models with their associated performances metrics (see Model Evaluation) and associated features, for each model. In total, about 8,500 models for classification and about 1,800 for regression are tested, but a large part will not be computed because of non-supported data (see **Supplementary Table S1**). Once all models are generated, the program executes the best model(s) selection section. The average performance among some computed metrics (see Model Evaluation) are used to estimate the most efficient model (see Best Model Selection), and correlated features are retrieved from the original dataset (see Correlated Features Search) and compiled in a tabular-separated text file report. Depending computing performances and dataset size, a few hours may be needed for BioDiscML pipeline to finish. Before the end of BioDiscML execution, a user can execute at any time BioDiscML from the checkpoint in parallel to perform the best model selection process, which will retrieve models from the feature-optimized model list generated and updated in real-time.

Data Preprocessing

BioDiscML supports multiple input files (e.g., clinico-pathological information with omics data), as the condition that sample identifiers exist in all files to perform joining. The input datasets are assumed to be clean and consistent, in a flat file format, table-like structure with samples in rows and features in columns (**Figure 2**). Field separator symbols (e.g., tabulation, comma, semicolon) are automatically detected based on the first lines of the file. Feature and instance duplicate names are not allowed. Where multiple datasets are submitted, only one must contain the class label. File contents are composed of instance identifiers (e.g., samples, patients) associated to numerical and/or nominal features (e.g., high/medium/low, effect_A/effect_B, Drug_1/Drug_2). Let be a set of *q* datasets

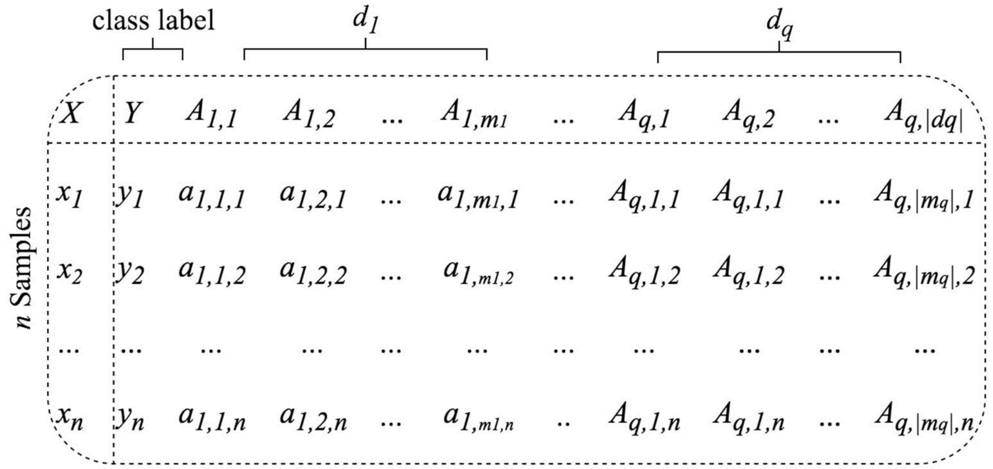

class label
d_1
d_q
n Samples
X Y $A_{1,1}$ $A_{1,2}$ ... $A_{1,m1}$ ... $A_{q,1}$ $A_{q,2}$ ... $A_{q,|dq|}$
x_1 y_1 $a_{1,1,1}$ $a_{1,2,1}$ ... $a_{1,m1,1}$ ... $A_{q,1,1}$ $A_{q,1,1}$ ... $A_{q,|m_q|,1}$
x_2 y_2 $a_{1,1,2}$ $a_{1,2,2}$ ... $a_{1,m1,2}$ ... $A_{q,1,2}$ $A_{q,1,2}$ ... $A_{q,|m_q|,2}$
... ...
x_n y_n $a_{1,1,n}$ $a_{1,2,n}$ ... $a_{1,m1,n}$ .. $A_{q,1,n}$ $A_{q,1,n}$ ... $A_{q,|m_q|,n}$

FIGURE 2 | BioDiscML accepts as input one ({$\boldsymbol{d_1}$} only) or many ({$\boldsymbol{d_1}$, ..,$\boldsymbol{d_q}$}) symbol-separated table-like structured datasets containing samples in row and features in columns.

$\{d_1, d_2, ..., d_q\}$with $q \geq 1$containing $m_q$features. In each dataset the first column is used to create the joining of all datasets and consists of instances unique identifiers. If an identifier does not exist in all datasets, it will be ignored. The class label column Y is required and must be specified by the user. In addition to the class label, the dataset $d_1$contains a set of m_1 features noted $A_1 = \{A_{1,1}, A_{1,2}, ..., A_{1,m_1}\}$where A_{1,m_1},the m_1-th feature of d_1, is a vector denoted $\{a_{1,m_1,1}, a_{1,m_1,2}, ..., a_{1,m_1,n}\}$. Hence the feature vector of the n-th instance of the dataset d_1 is noted $x_{1,n} = \{y_n, a_{1,1,n}, a_{1,2,n}, ..., a_{1,m_1,n}\}$.In case of multiple datasets ($q \geq 2$), the feature vector of the n-th instance of the dataset d_r is noted then $x_{r,n} = \{a_{r,1,n}, a_{r,2,n}, ..., a_{r,m_r,n}\}$, where m_r is its number of features. The resulting set of merged datasets is called D.

Due to experimental errors or partially answered forms by patients, missing data may be present in the dataset. If one wants to conserve the features with missing data, the ML library used by BioDiscML will replace all missing values for nominal and numeric features with the modes (i.e., value that occurs most often) and means from the training data, respectively.

Also, manipulating large files is painful and one would exclude specific features without editing the input files. Thus, we implemented in BioDiscML features exclusion capabilities, where it simply ignores columns entered by the user.

Finally, a stratified sampling, which preserve the initial classes balancing, is applied to generate a test set for further evaluation to assess non-overfitting. It is set by default to create a train set of 2/3 of the input data, from which models will be computed, and 1/3 as a test set. These proportions can be modified by the user, and in case of very low number of instances, sampling can be disabled. A separate test set of the same structure than the train set can also be provided to BioDiscML.

Feature Ranking and Dimension Reduction

Feature ranking (as for feature selection) is essential to identify irrelevant or redundant features, which, once discarded, help to reduce computation time, improve prediction performance, and extract the most informative features (Sasikala et al., 2016). BioDiscML uses Information Gain (Krishnaiah and Kanal, 1982), which evaluates the worth of a feature by measuring the information gain with respect to a class. However, Information Gain is not compatible for regression problems using continuous class. In this case, BioDiscML instead uses ReliefF (Robnik-Sikonja and Kononenko, 1997), an adaptation to the original Relief algorithm (Kira and Rendell, 1992), which is as fast as Information Gain computation. ReliefF evaluates the worth of a feature by repeatedly sampling an instance and considering the value of the given feature for the nearest instance of the same and different class. Both Information Gain and ReliefF are used in conjunction with a ranker search algorithm, which ranks features by their individual evaluations. By default, and to reduce the dimension of the dataset, BioDiscML will only keep informative features (Information Gain >0.01 or |ReliefF| >0.01) or the first 1,000 best features, ordered by their absolute value of their score (ReliefF provides positive and negative correlation scoring with continuous class) (see Algorithm 1).

Feature Subset Selection and Model Search

Selecting a subset of features from a large number of potential variables is a common problem in pattern classification. Some feature subset selection methods involve a criterion to evaluate the capacity of feature subsets to distinguish one class from another, and a search algorithm to explore the potential solution space. At the end of the process, the feature subset generally contains the most important and non-redundant variables. In this context, BioDiscML automates an exhaustive procedure that generates thousands of combinations of ML algorithms and feature subsets defined by various search methods. This technique, which mixes both feature and model search, produces thousands of models associated to an optimal subset of non-redundant features. Many evaluation procedures (e.g., cross validations, resampling, bootstrapping) using train and test sets assess if models do not overfit the train set. All steps are described in Algorithm 2.

Algorithm 1: Dimension reduction by Information Gain and ReliefF

```
Input: train instances of $D$ (merged datasets), classifierType
(classification or regression)
Output: Dataset with ranked best features $S$

for each feature array $A$ do
  if classifierType = classification
    then $meritScore_a$ = Compute information gain
    value of $A$ with respect to classes $Y$
    else $meritScore_a$ = Compute ReliefF value of $a$
    with respect to classes $Y$
  end if
    if $meritScore_a \neq 0$
      then add | $meritScore_a$ |to meritScores
    end if
end for
SortedFeatures = Sort meritScores from largest to smallest values
if $|SortedFeatures| \leq 1000$
  then $S = SortedFeatures$
  else $S = SortedFeatures\{A_1, A_2, ..., A_{1000}\}$
end if
return $S$
```

Available machine learning algorithms

ML classifier algorithms and their hyperparameters (i.e., the options of the learning algorithm) are predefined in BioDiscML with random sets of options, including those provided by default in Weka library. In the current version, about 80 classifiers are available in BioDiscML (**Supplementary Table S1**). Some classifiers exist in various adaptations to support more features or class types. Depending available computing resources, the list of classifiers and hyperparameters can be modified by the user, as well as the spectrum of tested algorithms. In case of non-compatibility between a classifier and the input data or erroneous options, the classifier will be ignored by BioDiscML.

Evaluation criterion

For each classifier, several feature search methods are conducted. Each search method iterates over the features (except "top *k*" features approach) and trains a model at each iteration. To evaluate if a model is improved by adding or removing a feature, an evaluation criterion is measured by 10-fold cross-validation to assess if the prediction performance increases. All metrics are averaged over the folds and by class size, since a classifier usually performs differently over each class. This optimization procedure performed on feature selection either maximize or minimize the criterion, depending if it measures a performance or an error, respectively. Criterions supported by BioDiscML includes accuracy (ACC), balanced error rate (BER), Matthew's correlation coefficient (MCC), area under the curve (AUC), sensitivity, specificity, Root Mean Squared Error (RMSE), Correlation Coefficient (CC), etc. The full criterions list, including their equations, is provided in **Supplementary Table S2**.

Optimal feature subset search methods

For each ML algorithm listed in **Supplementary Table S1**, and for each selected criteria selected in **Supplementary Table S2**, from the ranked features *S* obtained in Algorithm 1, models are trained using several feature search approaches, including: Forward stepwise selection (FSS), Backward stepwise selection (BSS), Forward stepwise selection and Backward stepwise elimination (FSSBSE), Backward stepwise selection and Forward stepwise elimination (BSSFSE), and "top *k*" features. In the stepwise procedures, features having an equal predictive power to the outcome (i.e., distributions similar among classes) and retained in the model may be selected randomly or by order of appearance in the dataset.

Forward stepwise selection (FSS). Also called sequential forward selection (Reunanen, 2003), where features are added one by one to the model. At each added feature, the model is evaluated by 10 CV. If the model is improved, based on a given evaluation criterion, the feature is definitely kept in the model, otherwise it is rejected (Maugis et al., 2011).

Backward stepwise selection (BSS). This approach is similar to the FSS, but instead of starting from the best feature, this algorithm starts the selection from the worst feature. Features are added one by one, if the model is improved (evaluated by 10 CV) the feature is definitely kept in the model, else, it is rejected.

Forward stepwise selection and backward stepwise elimination (FSSBSE). The drawback of FSS and BSS is that once a feature is selected, it cannot be deleted at a later stage. Consequently, redundant features might be selected. To alleviate this problem, we have implemented a FFSBSE algorithm, inspired by previous work (Caruana and Freitag, 1994; Mao, 2004; Zhang, 2011). After each addition of an increasing criterion score feature using FSS, a BSE step removes all previously selected features one by one in reverse order with replacement and test the performance by 10 CV every time. If removing a feature improves the model (evaluated by 10 CV), then the feature is discarded, otherwise it is kept.

Backward stepwise selection and forward stepwise elimination (BSSFSE). Similar to FSSBSE, but instead the algorithm starts from the selection of the worst feature.

"Top k" features This fast method simply trains a model with a subset of *k* best features, with $k = \{1, 5, 10, 15, 20, 25, 30, 40, 50, 75, 100\}$.

Model evaluation

Prediction performance of a model is measured using various evaluation procedures including 10 CV, leave-one-out cross validation (LOOCV), holdout, repeated Holdout, bootstrapping, and 0.632+ bootstrap estimator. For each generated model described in previous sections, and for each evaluation procedure, the following metrics are measured (see **Supplementary Table S2**): ACC, AUC, AUPRC, Sensitivity, Specificity, MCC, BER. In *10 CV evaluation*, the original training set is randomly partitioned into 10 equal sized subsamples. The model is trained on nine subsamples and tested on the remaining one. The CV is repeated 10 times, where each subsample is used exactly once for evaluation. The reported metric scores are their average over all folds. In *LOOCV* each model is trained on all the data except for one instance and a prediction is made for that instance. Average of metric scores are computed over all tested instances. The *holdout* method is the simplest kind of cross validation where the dataset is randomly separated into two sets generated at sampling procedure (see **Figure 1**), called the training set and the testing set. The model is trained using the training set only, then is used to predict the class for the data in the testing set as evaluation. However, this type of evaluation can have a high variance since it depends heavily on which instances end up in the training and test sets. Thus, a *repeated holdout* is also performed 100 times (by default) with random sampling without replacement. *Repeated Holdout* consists of randomly select and hold out a 1/3 of the training sample for testing, build model with only the remaining samples, retrieve its performances, and repeat the process many times. At the end, we report the average all performance metrics. The *bootstrapping* is equivalent, except the random sampling is performed with replacement. Finally, we also provide a 0.632+ bootstrap estimator (Efron, 1983), representing an estimation of the bias of the predictive model, which should tend to 0, hence assessing that the model does not overfit.

In addition to all these metrics, for each feature-optimized generated models, we calculate the average MCC and BER with their associated standard deviation across all evaluations (10 CV, LOOCV, Repeated Holdout, Bootstrap, holdout). For regression, we calculate the average and standard deviation of CC and RMSE.

Best Model Selection

Selecting the best model is not trivial since several good solutions are produced. Moreover, the definition of a "good" model also depends of user needs; for example, one would favor a model with a very low number of features over a model having dozens of feature, even if the latter provides a better overall performance. While BioDiscML proposes an automatic selection of the best model, a manual approach would be appropriate at that step. For this reason, all models are stored in real time in a Microsoft Excel-compatible Comma Separated Value (CSV) file and can be easily ordered by a criterion metric according to the user needs. Identifiers of user-selected models can be then submitted to BioDiscML to generate data files for easy re-use in other programs and full reports (containing the biomarker signature, the model and its hyperparameters, overall performances, and correlated features). Otherwise, by default, BioDiscML best model selection procedure aims to identify the model having a high agreement between the various evaluation methods, hence assessing stability and low overfitting of the model. To this purpose, select the model having the best average MCC with a standard deviation lower than 0.1 (or another adjusted threshold set by the user). The user can change the best model selection strategy at ease in the program configuration file. For example, one would select a trained model on train set having the best

Algorithm 2: Identification of features subsets and feature-optimized models

```
Input: Dataset with ranked best features S, set of ML classifiers
with various hyperparameters, set of criteria, datasets D
Output: Feature-optimized models list L with their identified
features subset

function EVALUATE(model, selectedFeatures, dataset D, list of
models L)
  trainSetEvaluation = Evaluate model using 10CV, LOOCV,
  Bootstrap, Repeated Holdout, 0.632+ estimator on train set
  testSetEvaluation = Extract selectedFeatures from test
  instances of dataset D and perform holdout evaluation
  with model
  performances = trainSetEvaluation, testSetEvaluation
  add model with performances and selectedFeatures to L
  return L
end function

for each classifier in classifiers do
  for each criterion in criteria do
    for each featureSearchMethod in featureSearchMethods{FSS,
    BSS, FSSBSE, BSSFSE)
    do
      if criterion must be maximized
      (see Supplementary Table S2) then
        criterionScore = 0
        rule = "lesser than"
      else
        criterionScore = 1000
        rule = "greater than"
      end if
      if featureSearchMethod = FSS or BSS then
        if featureSearchMethod = BSS then
          S = invert feature rank order of S
        end if
        for each feature A in S do
          Add A to selectedFeatures
          model = Train using classifier with selectedFeatures
          newCriterionScore = perform 10CV evaluation
          if newCriterionScore rule CriterionScore
            then discard a from selectedFeatures
            else keep a in selectedFeatures
            criterionScore = newCriterionScore
          end if
        end for
      else
        if featureSearchMethod = BSSFSE
          then S = invert feature rank order of S
        end if
        for each feature A in S do
          Add A to selectedFeatures
          model = Train using classifier with selectedFeatures
          newCriterionScore = perform 10CV evaluation
```

```
                if newCriterionScore rule CriterionScore
                then discard A from selectedFeatures
                else
                    keep A in selectedFeatures
                    criterionScore = newCriterionScore
                    for each selectedFeature from before last kept
                    feature to the first selected feature in
                    selectedFeatures do
                        remove selectedFeature from selectedFeatures
                        subModel = Train using classifier with
                        selectedFeatures
                        subNewCriterionScore = perform 10CV
                        evaluation
                        if subNewCriterionScore rule
                        NewCriterionScore then
                            discard selectedFeature from selectedFeatures
                            NewCriterionScore = subNewCriterionScore
                        else
                            keep selectedFeature in selectedFeatures
                        end if
                    end for
                end if
            end for
        end if
        L = EVALUATE(model, selectedFeatures, A, L)
    end for
  end for
  # create models without stepwise feature subset selection
  approaches
  selectedFeatures = k first features
  model = Train using classifier with selectedFeatures
  from dataset S
  L = EVALUATE(model, selectedFeatures, A, L)
end for
return L
```

MCC on the test set (TEST_MCC, see readme program file), or on the best bootstrapping using merged training and testing sets (TRAIN_TEST_BS_MCC).Since all generated models have a unique identifier, one would use these identifiers to select the best model based its own criteria.

Ensemble Learning

Since several good models with different features can exist in the results generated by BioDiscML, we also propose a vote classifier able to combine many models together. Different combinations of probability estimates for classification are available, including Average of probabilities, Product of probabilities, Majority voting and Median. As for best model selection, many metrics and correlated features are provided for this ensemble model. We also count the number of occurrences of each features in the combined models. The models to add in the ensemble classifier are dependent of the user choice. They can be selected manually using their unique identifiers, or by setting a metric dependent rule (by default average MCC lower than 0.6) and a maximum number of models to include.

Correlated Features Search

The identified signatures by stepwise search methods will tend to ignore all redundant/correlated features. To use the models as "black box" for pure prediction, this may be optimal, but not for biological interpretation because one would understand why the selected features have a link with the predicted class. To this purpose, from the features in the signature, BioDiscML retrieves all other correlated features from the original dataset using Pearson and Spearman correlations. BioDiscML also identifies all neighbor features discovered during feature ranking procedure by Information Gain and ReliefF methods. Both provide feature ranking scores that are used to detect the features having the same predictive power, i.e., similar behavior among instances. With these techniques, redundant information lost during the feature selection process are recovered, hence helping for further interpretation of the signature.

Gene Set Enrichment Analysis

We performed several Gene Set Enrichment Analysis (GSEA) to characterize the signatures identified by BioDiscML on the test datasets. To this purpose, we used ToppFun tool, from ToppGene suite (Chen et al., 2009), with Bonferroni correction at 0.05 to the probability density function (p-value Method).

Datasets for Benchmarking

Datasets described in **Table 1** have been evaluated to compare the performance of BioDiscML and recent tools. All models and signature information for all tested datasets are presented in **Supplementary Datasets_results.xlsx.**

RESULTS

We compared BioDiscML to various recent approaches dedicated to biomarker discovery and modeling, including MINT (Rohart et al., 2017a), AucPR (Yu and Park, 2014), and RGIFE (Swan et al., 2015) to demonstrate the better predictive performances that BioDiscML offers on various omics datasets. In all cases, BioDiscML outperform these state-of-the-art tools.

BioDiscML vs. Mint

MINT implements a multivariate integrative method able to integrate independent datasets, reduce batch effect, classify instances and identify key discriminant variables. In their study, they performed a feature selection and classification evaluation of a stem cell dataset. According to their published results, they identified a signature of 17 genes which predicted the test and train sets with a BER of 9.4 and 7.1% resp. Using the exact same train set, BioDiscML identified a signature of 19 genes by optimizing the AUC of a Random Forest model with 100 iterations and using the FSSBSE feature search method. The measured BER on the test set was 7%, and on the train set 3.5, 3.6, 6.8, and 7.2% using 10 CV, LOOCV, and repeated holdout and bootstraping resp. To select this model among the 4,710 successfully generated models, we simply retrieved the one

TABLE 1 | Description of the real-world datasets used to evaluate the performance of BioDiscML vs. recent tools.

Name	Description	Features	Instances	References
Stem cells	Fifteen merged transcriptomics microarray sets from multiple platforms. They contain three types of human cells as classes: human Fibroblasts (Fib), embryonic stem cells (ESC), and induced pluripotent stem cells (IPSC)	13,315	Train set: 62 ESC, 105 IPSC, 43 Fib Test set: 33 ESC, 77 IPSC, 22 Fib Total: 210 (train) + 132 (test) = 342 patients	Rohart et al., 2017a
Colon cancer	Transcriptomics microarray available from ColonCA R package in Bioconductor (Gentleman et al., 2006), separated between cancerous from non-cancerous colon tissue	2,000	Sixty-two patients, including 40 tumors and 22 normal cases	Alon et al., 1999
Central nervous system	Microarray gene expression data derived from central nervous system of patients brain tumors to predict embryonal tumor outcome	7,129	Sixty patients, including 39 medulloblastoma survivors, and 21 treatment failures cases	Pomeroy et al., 2002
Diffuse large B-cell lymphoma (DLBCL)	Transcriptomic microarray of pre-treatment biopsies tumor specimens separated in DLBCL and follicular lymphoma	2,647	Seventy-seven patients, including 58 DLBCL and 19 follicular lymphoma	Shipp et al., 2002
Prostate cancer	Microarray expression analysis was used to determine gene expression levels differences between tumor and non-tumor prostate samples	2,135	One hundred two patients, including 52 tumor and 50 normal cases	Singh et al., 2002

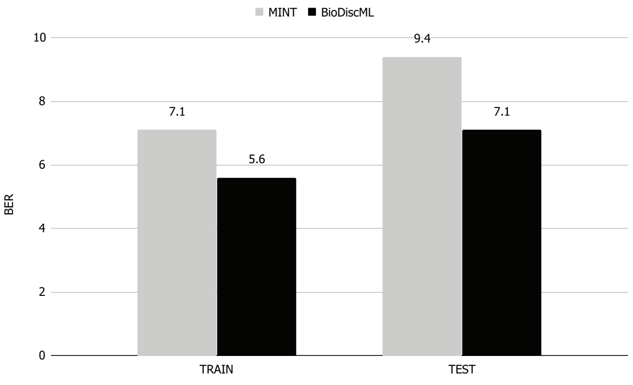


FIGURE 3 | BER comparison of MINT vs. BioDiscML. Train BER value was obtained by LOGOCV performance evaluation and test BER value using holdout validation. Values are in percentage.

having the lowest BER on the holdout method. Thus, on the same test set, the Random Forest model identified by BioDiscML improved the BER from 9.4 to 7%, corresponding to about 25% relative error decrease (see **Figure 3**).

In their paper, MINT authors have provided the signature identified by their method. Although both signatures found by MINT and BioDiscML have no genes in common, most of level 2 biological processes ontologies (see **Supplementary Data**) obtained by these signatures were identical (cellular process, multicellular organismal process, metabolic process, biological regulation, cellular component organization or biogenesis, localization). Specific biological processes were reproduction and immune system in MINT signature, and response to stimulus and developmental process in BioDiscML signature. A long signature of 71 genes can also be obtained using correlated feature search in BioDiscML. Using this long signature, only immune system process was added compared to the short signature, which also exists in the MINT signature. Moreover, this long signature provided perfect predictions on all instances of the test set. We also compared both signatures GSEA (see Methods). MINT signature did not show any significantly enriched ontologies, literature co-citation, co-expression etc. At the opposite, the short signature of BioDiscML found about 20 hits related to stem cells in co-expression databases (GeneSigDB and MSigDB) and co-expression Atlas. Also, about 20 other hits were found in literature co-citation about cognitive diseases (Alzheimer, Parkinson, Schizophrenia). The long signature provided even more hits, in many other categories.

BioDiscML vs. AucPR

In their study, authors of AucPR, an AUC-based approach using penalized regression, have evaluated the performance of their tool against four datasets. While AucPR showed a very good prediction performance on three of four tested datasets, the average AUC on *ColonCA* dataset was about 90% using both best penalization regression approach modes of the tool (Lasso and ElasticNet). Considering AucPR had the lowest performance on this dataset, we tried the performance of BioDiscML on it. In their paper, authors report the boxplots of 100 AUCs obtained by repeated holdout (random separation of 2/3 of the data for training and the remaining for testing) without sampling step. Using the same data and same evaluation method without sampling before training, two models identified by BioDiscML, on the 3,967 successfully generated models, shared the same best average AUC score. We chose the one having the best MCC on repeated holdout, a model based on a Hoeffding Tree (parameters: infogain split, Naive Bayes adaptive leaf prediction strategy, grace period of 200, tie threshold of 0.05) optimized by AUC. This model provided an average AUC of 99.3% (0.632+ rule at 0.047) using 10 genes discovered by FSSBSE. This is an improvement of AUC of about 11%. Both AucPR modes AucL and AucEN selected in comparison 30 and 22 genes resp. The

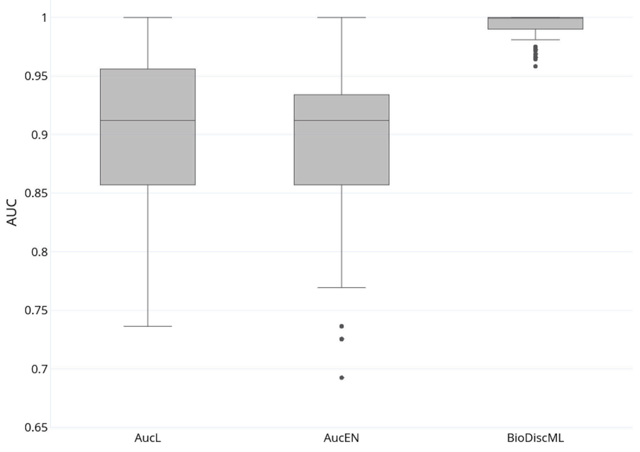


FIGURE 4 | Boxplot of AUCs bootstrapping over 100 iterations of most performant AucPR methods called AucL (AucPR with Lasso) and AucEN (AucPR with ElasticNet), vs. BioDiscML most performant model (Hoeffding Tree).

benchmark comparison of AUCs is reported in **Figure 4**. The model identified by BioDiscML has a much better performance in terms of average AUC and variance over bootstrapping. GSEA was not performed since this dataset didn't provided gene identifiers.

BioDiscML vs. RGIFE

RGIFE is an heuristic method intending to identify reduced panels of biomarkers with highly predictive performance. It first ranks features by their contribution to the generated models, and dynamically removes blocks of features. It also introduces a concept called soft-fail, which considers an iteration successful despite a performance drop within a tolerance level and specific circumstances. We evaluated the performance of BioDiscML on three datasets tested in RGIFE, including Central Nervous System (CNS), DLBCL, and Prostate Cancer datasets. On the 10 tested datasets by RGIFE, the three selected datasets showed accuracies around 60–70% for 10 CV, while BioDiscML identified models and signatures providing prediction performance close to perfection (100% accuracy) with lower number of features. Performances are reported in **Table 2**, where, for each dataset, we identified two models found by BioDiscML. To provide a fair comparison with the RGIFE manuscript we selected models having the best 10 CV accuracy (with best bootstrapping accuracy and lowest number of features in case of models' performance equality), which ended with 100% accurate models. But since this typical measure approach tends to be over-optimistic on the real performance of the models and because overfitting was suspected, we also reported models having the best bootstrapping accuracy. Obtained models show accuracies between 10CV and Bootstrapping more consistent, hence showing models are stable. In any case, 10CV accuracy was always better with BioDiscML results. The two signatures found for CNS dataset presented an overlap of five genes, and a merged list of the signatures show several GSEA significant hits related medulloblastoma and other cancers. For BLBCL dataset, no genes overlapped the two signatures, and we found significant hits related to dehydrogenase activity in the GSEA analysis on the merged list of the signatures, which has a link with follicular lymphoma to diffuse large B-cell lymphoma (Montoto et al., 2007). Finally, the prostate cancer signatures showed no overlap either, but GSEA analysis on the merge lists show several hits related to this cancer.

In terms of computing performances, on a same server containing four Intel(R) Xeon(R) CPU E5-2695 v2 @ 2.40GHz (48 threads), BioDiscML runtime was 28, 387, and 393 min on CNS, DLBCL, and Prostate Cancer datasets resp., and generated 5,751, 6,479, and 6,408 models resp., without exceeding 16 GB memory usage. In comparison, computation time reported by RGIFE in their **Supplementary Data** show ranges about 180–400 min.

DISCUSSION

A Simplified but Customizable Automated ML Tool

BioDiscML tool has been developed to enhance biomarker discovery using an exhaustive ML approach and propose automation of ML steps to perform such task. A large variety of algorithms is available and combinations of strategies are countless if we consider the hyperparameters of all classifiers and feature selection algorithms. This complexity is a barrier to non-expert users attempting to use ML to analyze their data. Thus, we designed BioDiscML to simplify ML steps without penalizing the performance, such as using fast and optimal feature ranking algorithms and feature search methods, limit the number of features after feature ranking, and establish predefined classifiers hyperparameters to reduce computing time. Although editable in BioDiscML configuration file, these intentional limitations provide researchers a program that generate results without intervention within a few hours of calculation on a recent computer.

A Sampling Procedure to Avoid Overtfitted Models

BioDiscML implements a sampling step to assess the non-overfitting and the good performance of identified models and signatures, where it splits the dataset into two stratified (class balancing is preserved) random parts. The program also accepts a second input file as a test dataset, as long as it is in the same format as the train set. In case of very limited instances, it is possible to skip the sampling operation, although not recommended because of the risk to not detect overfitted models. A reasonable number of instances (i.e., samples) should be provided to BioDiscML, else it is expected to obtain models with low performances. For example, we estimate that a highly heterogeneous dataset, such as prostate or breast cancer data, should contain at least half-hundred patients per class, while a dataset based on a study involving cloned living species could be limited to half a dozen individuals per class.

TABLE 2 | Performances of RGIFE vs. BioDiscML measured by accuracy obtained through 10-fold cross validation (10CV_ACC) and bootstrapping (BS_ACC).

	RGIFE			BioDiscML					
Dataset	**10CV_ACC**	**#Features**	**Model**	**10CV_ACC**	**BS_ACC**	**#Features**	**Model**	**Search**	**Criterion**
CNS	77.1	Not reported	KNN	100	80.7	12	A2DE	BSSFSE	AUC
				93.3	98.6	11	HT	FSSBSE	AUC
DLBCL	68	9	RF	100	93	6	A1DE	FSSBSE	MCC
				98.7	98.3	6	NB	FSSBSE	AUC
Prostate cancer	95.2	158	SVM	100	91	12	VFI	BSSFSE	ACC
				99	95.7	10	NB	FSSBSE	AUC

Classifiers evaluated by RGIFE were K-Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machines (SVM). Most performant classifiers identified by BioDiscML were Average two Dependance Estimators (A2DE), Hoeffding Tree (HT), Average 1 Dependance Estimators (A2DE), Voting Features Intervals (VFI), and Naive Bayes (NB). Hyperparameters are described in ***Supplementary Data****. Various criteria were used, including AUC, MCC, and FDR, and two feature search BSSFSE and FSSBSE. The signatures are shown in* ***Supplementary Data****.*

Feature Selection Procedures in BioDiscML Are Fast and Scalable

Omics datasets are generally composed of a thousands of features. To simplify input datasets and save computation time BioDiscML implements a feature ranking and dimension reduction procedure. Many approaches exist (Chandrashekar and Sahin, 2014) and most are applicable to biological problems (Saeys et al., 2007), but we choose to only implement Information Gain (Krishnaiah and Kanal, 1982) for classification, and ReliefF (Robnik-Sikonja and Kononenko, 1997), for regression, since they are fast and highly scalable univariate tests (Saeys et al., 2007). Information Gain shown very good performance on biological data (Li et al., 2004, 2011; Abusamra, 2013), as for ReliefF (Marchiori et al., 2005; He and Yu, 2010; Wang et al., 2016). Besides, their ranking capability provides an easy way to eliminate redundant, non-informative and noisy information, hence our choice to provide only those in BioDiscML.

BioDiscML Uses All Available Classifiers From a Widely Accepted and Efficient ML Library

There is a plethora of ML algorithms specialized in classification (i.e., categorical class) and regression (i.e., continuous class). BioDiscML covers many of them but can also be manually limited to the most known and widely applied in biomedical research for the development of predictive models such as Random Forest, Decision Trees, Rules, Naive Bayes, Artificial Neural Networks, Bayesian Networks and Support Vector Machines. They all resulted in effective and accurate decision-making (Jagga and Gupta, 2015). But the final models created with these classifiers in various studies were all delivered after an exhaustive search work. BioDiscML aims to reduce this search time by providing the models adapted to user datasets. All ML algorithms are provided by an advanced freely available ML library toolkit, called Weka. Besides this library, various ML libraries exist, such as SciKit-Learn (Nelli, 2015) (written in Python) and packages in R (Lesmeister, 2017). BioDiscML implements Weka library for various reasons, including its wide usage in computational biology (Gewehr et al., 2007; Bendl et al., 2014; Bernardi et al., 2015; Arganda-Carreras et al., 2017; Chicco, 2017; Alves et al., 2018), its high citation rate (at August 2018) and its highly versatile object-oriented language JAVA (e.g., easy to parallelize, multi-platform compatibility, GUI integration, generally already installed on clients, etc.), which is much faster (Fourment and Gillings, 2008) and scalable than Python or R. Finally, the user can use Weka GUI (graphical interface) to explore BioDiscML results, generate ROC curves or try other combinations of classifiers by hand. For example, the output files generated by BioDiscML are compatible with Weka and can be loaded in its GUI.

A Combination of Model Search and Feature Search Procedures to Identify Highly Predictive Models

BioDiscML combines the model search and the feature search together to identify biomarker signatures. Using the various search methods (i.e., stepwise and top *k*) and optimized criteria, each model is associated to a signature of features. Forward and backward stepwise search methods return signatures that are optimized on the classifier and the criterion. Note that the backward stepwise search approaches (BSS, BSSFSE) are not the usual “backward elimination” used in the literature (Sutter and Kalivas, 1993) for variables selection since it would be computationally expensive here. Instead, backward selection starts from worst features and will generally return performant models only when most of features have a relatively good univariate information gain or ReliefF score. The signature then reveals a combination of biomarkers which, associated together in a model, provide a highly predictive value of the class.

To assess the overall performance of the models, their robustness and the absence of overfitting, various well-known evaluation methods (Arlot and Celisse, 2010) have been implemented in BioDiscML, because some may not be adapted to all situations. For example, for biomedical studies which generally produce a low number of patients (i.e., instances), bootstrapping is a good alternative to sampling (Chen et al., 2002) (i.e., split in train and test set, involving waste of data). Besides, it is known that *k*-fold cross validation tends to deliver over-optimist performances (Smith et al., 2014). To facilitate the choice of the best models, we provide many performance metrics that can be averaged over all evaluation methods. BioDiscML also provide an ensemble classifier based on a voting system to include many models with different signatures. This method is known to provide better predictive performance than could be obtained from any of the constituent learning algorithms alone (Polikar, 2006).

Signature Interpretation Is Still a Challenge

A biologist will want to interpret and validate *in silico* the signature, since there is an obvious relation between the identified biomarkers in a signature and the predicted class (e.g., outcome). To perform such task, there exist many Gene Set Enrichment Analysis (GSEA) tools, such as ToppGene suite (Chen et al., 2009) or Enrichr (Kuleshov et al., 2016). These GSEA tools will provide a characterization of signature and confirm to the biologist if the signature has a biological meaning with the original study from which the dataset have been generated. Some more extensive literature searches may provide more insights and help linking the signatures' features with the predicted class.

Moreover, in some cases, the biologist, based on its experience and knowledge, may not find the biomarkers he expects in the signatures. This is a consequence of the feature search procedure which produces highly optimized signatures. This optimization tends to ignore all redundant features that could potentially help the biological interpretation of the biomarkers related to the class. To overcome this issue, BioDiscML retrieves all correlated features that could have been excluded during the feature subset selection and model search procedure. It is important to note that adding signature's perfectly correlated features (100% correlated) to the model will maintain its performance. At the opposite, it is expected to have a slight performance drop when adding "almost-correlated" features (95–99% correlation), which can be tested by training and evaluation of the model with the added correlated features.

Some scientific visualization tools would have probably been welcome in BioDiscML, but JAVA visualization libraries are rare. However, to overcome this lack, BioDiscML generates a subset of the input dataset containing only the sample values of the signature' features. This subset in comma-separated values format can be loaded easily in other visualization software such as Microsoft Excel, Orange (Demšar et al., 2013), RapidMiner (Hofmann, 2016), or R (Gardener, 2012) to generate heatmaps or boxplots.

BioDiscML Exhaustive Approach Outperforms Recently Published Tools

We benchmarked BioDiscML against recent tools proposing different approaches to discover biomarker signatures. Benchmarks showed that BioDiscML outperforms these state-of-art methods using same datasets. Because of its exhaustive approach, it was able to identify one or more models with smaller signatures providing much better prediction performances. We also demonstrated in the case of the stem cell dataset that BioDiscML signature contained different genes but similar ontologies than the MINT signature, with a better prediction performance. A GSEA also showed that the BioDiscML signature had much more biological evidence, denoted by the occurrence of stem cells topics in the co-expression databases. The genes in the BioDiscML signatures were also present in neurodegenerative diseases, highlighting the link of these genes with the neuronal system, supported by evidence of efficient stem cell-based therapies for neural repair (Volkman and Offen, 2017). For the other benchmarked datasets which contained gene references, the GSEA analyses also showed supporting evidences assessing the biological relation between the genes found in the signatures and the biological experiment from where they were produced.

It is important to note that short but still very predictive model' signatures can be extended as an "enriched" signature which include the correlated genes. These enriched signatures may increase the accuracy of the signature, but more importantly they can help to better understand the biological meaning of the model. On the MINT dataset, BioDiscML showed a perfect prediction on the test set with the enriched signature and retrieved more ontologies.

Finally, in this paper we benchmarked BioDiscML only on transcriptomics datasets from microarray data provided by the tools we tested. But BioDiscML showed also good performances in other omics datasets tested in other contexts (data not shown).

Performant Models Identified in Minutes

BioDiscML computing performances are highly dependent on the size of the input dataset and the available processors. To generate all models implemented in the software, it requires a few hours of computation. However, it is possible to restrict BioDiscML to a specific list of algorithms, hence reducing the computation time to seconds or minutes. It is also possible to extract the best signatures and models produced since the beginning of BioDiscML execution at any time. We have prioritized the training of the most common and fastest classifiers to propose a large number of computed models shortly after starting BioDiscML. More complex models, such as Multilayer perceptrons, are set in low priority. More running time will simply increase the probability to obtain a better model. The user is informed in the command line output the progression of the program (i.e., the number of models trained and remaining to train). Finally, BioDiscML can be stopped at any moment, especially if the user is not interested to let BioDiscML train complex classifiers.

CONCLUSIONS

This paper introduces BioDiscML, dedicated to identify optimal combination of biomarkers (i.e., features) and machine learning models to predict measured outcomes. It provides a user-friendly and powerful solution to researchers in the medical field looking to identify predictive features, essential to the development of personalized medicine approaches and research of new therapeutic targets. This software has the benefit to exploit a large number of machine learning classifiers within a fully automated process combined with data pre-processing, hence facilitating the work of a non-machine learning experts audience. Expert users have also the possibility to configure advanced options. BioDiscML is a great opportunity to reduce biomarkers search time, by revealing the most adapted classifiers to a given dataset and even proposes new algorithms poorly explored in the literature that could have a great potential to classify biological data. Otherwise, although this program has been tested with omics data and proven its better performances compared to recent computational biology tools created for the same purpose, it is compatible with any other non-biological data. Finally, the ML library used in BioDiscML is highly maintained, hence enabling convenient additions of newly implemented algorithms in future versions.

DATA AND SOFTWARE AVAILABILITY

BioDiscML software project and the datasets analyzed during the current study are available at https://github.com/mickaelleclercq/BioDiscML under GPL-3.0 license. This software written in JAVA is compatible with the main operating systems. Windows, Linux and Mac.

AUTHOR CONTRIBUTIONS

ML designed and implemented BioDiscML software, conducted literature searches, researched data and selected relevant articles. ML also created figures and tables, and wrote, formatted and finalized the article for submission. BV were in charge to test the software and report all bugs. MM-M and MS helped to optimize BioDiscML pipeline and the implemented algorithms. OP helped to improve the manuscript during the reviewing process. AB, YF, and AD supervised and reviewed the design of the study. All authors contributed to writing and reviewing the manuscript.

FUNDING

This project has been funded in part by the Laboratoire d'Uro-Oncologie Expérimentale of the CHU de Québec directed by YF and by the L'Oréal research chair in digital Biology of Université Laval, directed by AD.

ACKNOWLEDGMENTS

We thank Dr Kim-Anh Lê Cao (University of Queensland Diamantina Institute, Translational Research Institute, Brisbane, Australia) who provided us the Stem Cells data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00452/full#supplementary-material

REFERENCES

Abusamra, H. (2013). A comparative study of feature selection and classification methods for gene expression data of glioma. *Procedia Comput. Sci.* 23, 5–14. doi: 10.1016/j.procs.2013.10.003

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6745–6750. doi: 10.1073/pnas.96.12.6745

Alves, P., Liu, S., Wang, D., and Gerstein, M. (2018). Multiple-swarm ensembles: improving the predictive power and robustness of predictive models and its use in computational biology. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 15, 926–933. doi: 10.1109/TCBB.2017.2691329

Arganda-Carreras, I., Kaynig, V., Rueden, C., Eliceiri, K. W., Schindelin, J., Cardona, A., et al. (2017). Trainable weka segmentation: a machine learning tool for microscopy pixel classification. *Bioinformatics* 33, 2424–2426. doi: 10.1093/bioinformatics/btx180

Arlot, S., and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv.* 4, 40–79. doi: 10.1214/09-SS054

Beerenwinkel, N., Greenman, C. D., and Lagergren, J. (2016). Computational cancer biology: an evolutionary perspective. *PLoS Comput. Biol.* 12:e1004717. doi: 10.1371/journal.pcbi.1004717

Bendl, J., Stourac, J., Salanda, O., Pavelka, A., Wieben, E. D., Zendulka, J., et al. (2014). PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. *PLoS Comput. Biol.* 10:e1003440. doi: 10.1371/journal.pcbi.1003440

Bernardi, R. C., Melo, M. C. R., and Schulten, K. (2015). Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta* 1850, 872–877. doi: 10.1016/j.bbagen.2014.10.019

Butti, M. D., Chanfreau, H., Martinez, D., García, D., Lacunza, E., and Abba, M. C. (2014). BioPlat: a software for human cancer biomarker discovery. *Bioinformatics* 30, 1782–1784. doi: 10.1093/bioinformatics/btu111

Caruana, R., and Freitag, D. (1994). "Greedy attribute selection," in *Proceedings of the Eleventh International Conference on Machine Learning* (New Brunswick: Rutgers University), 28–36. doi: 10.1016/B978-1-55860-335-6.50012-X

Chandrashekar, G., and Sahin, F. (2014). A survey on feature selection methods. *Comput. Electr. Eng.* 40, 16–28. doi: 10.1016/j.compeleceng.2013.11.024

Chen, D. R., Kuo, W. J., Chang, R. F., Moon, W. K., and Lee, C. C. (2002). Use of the bootstrap technique with small training sets for computer-aided diagnosis in breast ultrasound. *Ultrasound Med. Biol.* 28, 897–902. doi: 10.1016/S0301-5629(02)00528-8

Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.* 37, W305–W311. doi: 10.1093/nar/gkp427

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Min.* 10:35. doi: 10.1186/s13040-017-0155-3

Cima, I., Schiess, R., Wild, P., Kaelin, M., Schüffler, P., Lange, V., et al. (2011). Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3342–3347. doi: 10.1073/pnas.1013699108

Cui, J., Chen, Y., Chou, W.-C., Sun, L., Chen, L., Suo, J., et al. (2011). An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer. *Nucleic Acids Res.* 39, 1197–1207. doi: 10.1093/nar/gkq960

Cun, Y., and Fröhlich, H. (2013). Network and data integration for biomarker signature discovery via network smoothed T-statistics. *PLoS ONE* 8:e73074. doi: 10.1371/journal.pone.0073074

Cun, Y., and Fröhlich, H. (2014). netClass: an R-package for network based, integrative biomarker signature discovery. *Bioinformatics* 30, 1325–1326. doi: 10.1093/bioinformatics/btu025

Daoqiang, Z., and Dinggang, S. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* 59, 895–907. doi: 10.1016/j.neuroimage.2011.09.069

Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., et al. (2013). Orange: data mining toolbox in python. *J. Mach. Learn. Res.* 14, 2349–2353.

Deshpande, G., Libero, L. E., Sreenivasan, K. R., Deshpande, H. D., and Kana, R. K. (2013). Identification of neural connectivity signatures of autism using machine learning. *Front. Hum. Neurosci.* 7:670. doi: 10.3389/fnhum.2013.00670.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Am. Stat. Assoc.* 78, 316–331. doi: 10.1080/01621459.1983.10477973

Eslami, A., Qannari, E. M., Kohler, A., and Bougeard, S. (2013). "Multi group PLS Regression: Application to Epidemiology," in *Springer Proceedings in Mathematics & Statistics*, (Springer Nature Switzerland AG; Part of Springer Naturel), 243–255. Available online at: https://link.springer.com/bookseries/10533.

Fekete, T., Zach, N., Mujica-Parodi, L. R., and Turner, M. R. (2013). Multiple kernel learning captures a systems-level functional connectivity biomarker signature in amyotrophic lateral sclerosis. *PLoS ONE* 8:e85190. doi: 10.1371/journal.pone.0085190

Feurer, M., Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015). Efficient and robust automated machine learning. *Adv. Neural Inf. Process. Syst.* 28, 2962–2970.

Fischer, R. (2015). "Java 8: It's a Whole New Java," in *Java Closures and Lambda*, (Berkeley, CA: Apress), 1–10. Available online at: https://www.apress.com/us.

Fourment, M., and Gillings, M. R. (2008). A comparison of common programming languages used in bioinformatics. *BMC Bioinformatics* 9:82. doi: 10.1186/1471-2105-9-82

Fröhlich, H., and Cun, Y. (2012). Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics* 13:69. doi: 10.1186/1471-2105-13-69

Gardener, M. (2012). *Beginning R: The Statistical Programming Language.* Indianapolis, IN: John Wiley & Sons.

Gentleman, R., Carey, V., Huber, W., Irizarry, R., and Dudoit, S. (2006). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York, NY: Springer Science & Business Media. doi: 10.1007/0-387-29362-0

Gewehr, J. E., Szugat, M., and Zimmer, R. (2007). BioWeka–extending the Weka framework for bioinformatics. *Bioinformatics* 23, 651–653. doi: 10.1093/bioinformatics/btl671

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newslett.* 11, 10–18. doi: 10.1145/1656274.1656278

Hartigan, J. A., and Wong, M. A. (1979). Algorithm AS 136: A K-means clustering algorithm. *Appl. Stat.* 28, 100–108. doi: 10.2307/2346830

He, Z., and Yu, W. (2010). Stable feature selection for biomarker discovery. *Comput. Biol. Chem.* 34, 215–225. doi: 10.1016/j.compbiolchem.2010.07.002

Hofmann, M. (2016). *RapidMiner Ralf Klinkenberg, RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Available online at: https://www.oreilly.com/library/view/rapidminer/9781482205503/.

Holmes, G., Donkin, A., and Witten, I. H. (1994). *WEKA: A Machine Learning Workbench*. University of Waikato; Department of Computer Science. doi: 10.1109/ANZIIS.1994.396988

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika* 28, 321–377. doi: 10.1093/biomet/28.3-4.321

Jagga, Z., and Gupta, D. (2015). Machine learning for biomarker identification in cancer research – developments toward its clinical application. *Future Med.* 12, 371–387. doi: 10.2217/pme.15.5

Janevski, A., Kamalakaran, S., Banerjee, N., Varadan, V., and Dimitrova, N. (2009). PAPAyA: a platform for breast cancer biomarker signature discovery, evaluation and assessment. *BMC Bioinformatics* 10 (Suppl. 9):S7. doi: 10.1186/1471-2105-10-S9-S7

Johansson, H., Lindstedt, M., Albrekt, A. S., Borrebaeck, C. A. (2011). A genomic biomarker signature can predict skin sensitizers using a cell-based *in vitro* alternative to animal tests. *BMC Genomics* 12:399. doi: 10.1186/1471-2164-12-399

Kira, K., and Rendell, L. A. (1992). "A practical approach to feature selection," in *Proceedings of the Ninth International Workshop on Machine Learning (ML 1992)* (Aberdeen), 249–256. doi: 10.1016/B978-1-55860-247-2.50037-1

Kong, A., Gupta, C., Ferrari, M., Agostini, M., Bedin, C., Bouamrani, A., et al. (2014). Biomarker signature discovery from mass spectrometry data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 766–772. doi: 10.1109/TCBB.2014.2318718

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Krishnaiah, P. R., and Kanal, L. N. (1982). "Handbook of Statistics 2," in *Classification, Pattern Recognition and Reduction of Dimensionality* (North-Holland; Amsterdam; New York, NY: Oxford). Available online at: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.4710270315.

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97. doi: 10.1093/nar/gkw377

Lê Cao, K. A., González, I., and Déjean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics* 25, 2855–2856. doi: 10.1093/bioinformatics/btp515

Lagani, V., Kortas, G., and Tsamardinos, I. (2013). Biomarker signature identification in "omics" data with multi-class outcome. *Comput. Struct. Biotechnol. J.* 6:e201303004. doi: 10.5936/csbj.201303004

Lesmeister, C. (2017). *Mastering Machine Learning with R.* Birmingham: Packt Publishing Ltd.

Li, T., Zhang, C., and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics* 20, 2429–2437. doi: 10.1093/bioinformatics/bth267

Li, X., Peng, S., Zhan, X., Zhang, J., and Xu, Y. (2011). "Comparison of feature selection methods for multiclass cancer classification based on microarray data," in *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)* (Shanghai). doi: 10.1109/bmei.2011.6098612

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Lin, Y., Qian, F., Shen, L., Chen, F., Chen, J., and Shen, B. (2017). Computer-aided biomarker discovery for precision medicine: data resources, models and applications. *Brief. Bioinform.* doi: 10.1093/bib/bbx158

Liu, R., Wang, X., Aihara, K., and Chen, L. (2014). Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers. *Med. Res. Rev.* 34, 455–478. doi: 10.1002/med.21293

Mao, K. Z. (2004). Orthogonal forward selection and backward elimination algorithms for feature subset selection. *IEEE Trans. Syst. Man Cybern. B Cybern.* 34, 629–634. doi: 10.1109/TSMCB.2002.804363

Marchiori, E., Heegaard, N. H. H., West-Nielsen, M., and Jimenez, C. R. (2005). "Feature selection for classification with proteomic data of mixed quality," in *2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology* (La Jolla, CA) doi: 10.1109/cibcb.2005.1594944

Matsumura, K., Opiekun, M., Oka, H., Vachani, A., Albelda, S. M., Yamazaki, K., et al. (2010). Urinary volatile compounds as biomarkers for lung cancer: a proof of principle study using odor signatures in mouse models of lung cancer. *PLoS ONE* 5:e8819. doi: 10.1371/journal.pone.0008819

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2011). Variable selection in model-based discriminant analysis. *J. Multivar. Anal.* 102, 1374–1387. doi: 10.1016/j.jmva.2011.05.004

Montoto, S., Davies, A. J., Matthews, J., Calaminici, M., Norton, A. J., Amess, J., et al. (2007). Risk and clinical implications of transformation of follicular lymphoma to diffuse large B-cell lymphoma. *J. Clin. Oncol.* 25, 2426–2433. doi: 10.1200/JCO.2006.09.3260

Nelli, F. (2015). "Machine learning with scikit-learn," in *Python Data Analytics*, (Berkeley, CA: Apress), 237–264. Available online at: https://www.apress.com/us.

Pasolli, E., Truong, D. T., Malik, F., Waldron, L., and Segata, N. (2016). Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput. Biol.* 12:e1004977. doi: 10.1371/journal.pcbi.1004977

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits Sys. Magazine* 6, 21–45. doi: 10.1109/MCAS.2006.1688199

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415, 436–442. doi: 10.1038/415436a

Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *J. Mach. Learn. Res.* 3, 1371–1382.

Robnik-Sikonja, M., and Kononenko, I. (1997). "An adaptation of Relief for attribute estimation in regression," in *Fourteenth International Conference on Machine Learning* (Nashville, TN), 296–304.

Rohart, F., Eslami, A., Matigian, N., Bougeard, S., and Lê Cao, K.-A. (2017a). MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. *BMC Bioinformatics* 18:128. doi: 10.1186/s12859-017-1553-8

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

Rohart, F., Gautier, B., Singh, A., and Le Cao, K.-A. (2017c). mixOmics: an R package for’omics feature selection and multiple data integration. *PLoS Comput. Biol.* 13:e1005752. doi: 10.1101/108597

Roth, P., Wischhusen, J., Happold, C., Chandran, P. A., Hofer, S., Eisele, G., et al. (2011). A specific miRNA signature in the peripheral blood of glioblastoma patients. *J. Neurochem.* 118, 449–457. doi: 10.1111/j.1471-4159.2011.07307.x

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics* 23, 2507–2517. doi: 10.1093/bioinformatics/btm344

Sasikala, S., Appavu alias Balamurugan, S., and Geetha, S. (2016). Multi filtration feature selection (MFFS) to improve discriminatory ability in clinical data set. *Appli. Comput. Inf.* 12, 117–127. doi: 10.1016/j.aci.2014.03.002

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C. T., et al. (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* 8, 68–74. doi: 10.1038/nm0102-68

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., et al. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1, 203–209. doi: 10.1016/S1535-6108(02)00030-2

Smith, G. C. S., Seaman, S. R., Wood, A. M., Royston, P., and White, I. R. (2014). Correcting for optimistic prediction in small data sets. *Am. J. Epidemiol.* 180, 318–324. doi: 10.1093/aje/kwu140

Sutherland, A., McLean, A., Tang, B., Venter, D., Price, G., Stone, G., et al. (2011). Development and validation of a novel molecular biomarker diagnostic test for the early detection of sepsis. *Crit. Care* 15:R149. doi: 10.1186/cc10274

Sutter, J. M., and Kalivas, J. H. (1993). Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchem. J.* 47, 60–66. doi: 10.1006/mchj.1993.1012

Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S., and Bacardit, J. (2013). Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS* 17, 595–610. doi: 10.1089/omi.2013.0017

Swan, A. L., Stekel, D. J., Hodgman, C., Allaway, D., Alqahtani, M. H., Mobasheri, A., et al. (2015). A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. *BMC Genomics* 16 (Suppl. 1):S2. doi: 10.1186/1471-2164-16-S1-S2

Taverner, T., Karpievitch, Y. V., Polpitiya, A. D., Brown, J. N., Dabney, A. R., Anderson, G. A., et al. (2012). DanteR: an extensible R-based tool for quantitative analysis of -omics data. *Bioinformatics* 28, 2404–2406. doi: 10.1093/bioinformatics/bts449

Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). “Auto-WEKA,” in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD’13* (Chicago, IL). doi: 10.1145/2487575.2487629

Volkman, R., and Offen, D. (2017). Concise review: mesenchymal stem cells in neurodegenerative diseases. *Stem Cells* 35, 1867–1880. doi: 10.1002/stem.2651

Wang, L., Wang, Y., and Chang, Q. (2016). Feature selection methods for big data bioinformatics: a survey from the search perspective. *Methods* 111, 21–31. doi: 10.1016/j.ymeth.2016.08.014

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*. Hamilton, ON: University of Waikato, Morgan Kaufmann.

Wold, H. (1975). “Path models with latent variables: the NIPALS approach,” in *Quantitative Sociology* (Academic Press; Quantitative Sociology), 307–357. Available online at: https://www.sciencedirect.com/book/9780121039509/quantitative-sociology#book-description.

Yao, F., Coquery, J., and Lê Cao, K. A. (2012). Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinformatics* 13:24. doi: 10.1186/1471-2105-13-24

Yu, W., and Park, T. (2014). AucPR: an AUC-based approach using penalized regression for disease prediction with high-dimensional omics data. *BMC Genomics* 15 (Suppl. 10):S1. doi: 10.1186/1471-2164-15-S10-S1

Zhang, P., Cox, A., Cripps, A., and West, N. (2017). “Integrated biomedical data analysis utilizing various types of data for biomarkers identification,” in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Kansas City, MO). doi: 10.1109/BIBM.2017.8217879

Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE Trans. Inf. Theory* 57, 4689–4708. doi: 10.1109/TIT.2011.2146690

Conflict of Interest Statement: OP was employed by company L’Oréal.

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

BRIEF RESEARCH REPORT
published: 02 July 2019
doi: 10.3389/fgene.2019.00599

Cancer as a Tissue Anomaly: Classifying Tumor Transcriptomes Based Only on Healthy Data

Thomas P. Quinn[1,2,3]*, *Thin Nguyen*[1], *Samuel C. Lee*[1] *and Svetha Venkatesh*[1]

[1] *Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, VIC, Australia,* [2] *Centre for Molecular and Medical Research, Deakin University, Geelong, VIC, Australia,* [3] *Bioinformatics Core Research Group, Deakin University, Geelong, VIC, Australia*

Since the turn of the century, researchers have sought to diagnose cancer based on gene expression signatures measured from the blood or biopsy as biomarkers. This task, known as classification, is typically solved using a suite of algorithms that learn a mathematical rule capable of discriminating one group ("cases") from another ("controls"). However, discriminatory methods can only identify cancerous samples that resemble those that the algorithm already saw during training. As such, discriminatory methods may be ill-suited for the classification of cancer: because the possibility space of cancer is definitively large, the existence of a one-of-a-kind gene expression signature is likely. Instead, we propose using an established surveillance method that detects anomalous samples based on their deviation from a learned normal steady-state structure. By transferring this method to transcriptomic data, we can create an anomaly detector for tissue transcriptomes, a "tissue detector," that is capable of identifying cancer without ever seeing a single cancer example. As a proof-of-concept, we train a "tissue detector" on normal GTEx samples that can classify TCGA samples with >90% AUC for 3 out of 6 tissues. Importantly, we find that the classification accuracy is improved simply by adding more healthy samples. We conclude this report by emphasizing the conceptual advantages of anomaly detection and by highlighting future directions for this field of study.

Keywords: machine learning, TCGA, anomaly detection, classification, surveillance

OPEN ACCESS

Edited by:
Angelo Facchiano,
Institute of Food Sciences, National Research Council (CNR-ISA), Italy

Reviewed by:
Shihao Shen,
University of California, Los Angeles, United States
Rosalba Giugno,
University of Verona, Italy

***Correspondence:**
Thomas P. Quinn
contacttomquinn@gmail.com

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

Received: *06 March 2019*
Accepted: *05 June 2019*
Published: *02 July 2019*

Citation:
Quinn TP, Nguyen T, Lee SC and Venkatesh S (2019) Cancer as a Tissue Anomaly: Classifying Tumor Transcriptomes Based Only on Healthy Data. Front. Genet. 10:599.
doi: 10.3389/fgene.2019.00599

1. INTRODUCTION

Cancer is a collection of complex heterogeneous diseases with known genetic and environmental risk factors. Physicians diagnose cancer by carefully weighing evidence collected from patient history, physical examination, laboratory testing, clinical imaging, and biopsy. Computers can aid diagnosis and improve outcomes by mitigating diagnostic errors. Indeed, this objective is actively researched, where studies have shown that computers can reduce the reading errors of mammography (Rangayyan et al., 2007) and commuted tomographic (CT) (Chan et al., 2008) images. Meanwhile, researchers have also sought to use computers to diagnose cancer based on gene expression signatures measured by high-throughput assays like microarray or next-generation sequencing (Alon et al., 1999; Golub et al., 1999). Gene expression signatures are ideal biomarkers because mRNA expression is dynamically altered in response to changes in the cellular environment. However, developing molecular diagnostics requires large data sets which have only recently become available due to reduced assay costs. These data could usher in a new era in clinical diagnostics.

Within the last decade, scientists have produced large transcriptomic data sets containing thousands of clinical samples. Of these, the TCGA stands out as the most comprehensive, having sequenced more than 10,000 unique tissue samples from 33 cancers and healthy tissue controls (Weinstein et al., 2013). Meanwhile, an equally large study, GTEx, has sequenced non-cancerous samples comprising 54 unique human tissue types (Lonsdale et al., 2013). Already, a number of studies have used the TCGA data to build diagnostic classifiers that can determine whether a tissue sample is cancerous or not based only on its gene expression signature (Kourou et al., 2015). This task, known as classification, is typically solved using a suite of algorithms that learn a mathematical rule capable of discriminating one group ("cases") from another ("controls"). This rule is learned from a large portion of the data called the "training set," and then evaluated on withheld data called the "test set." Discriminatory classifiers like artificial neural networks (ANNs), support vector machines (SVMs), and random forests (RFs) have become popular in the biological sciences (Jensen and Bateman, 2011). All of these work well for high-dimensional data, so long as the training set contains enough correctly labeled cases and controls.

Clinicians need to answer questions like, "Is this tissue cancerous or not?" and "Is this cancer malignant or not?" ANNs, SVMs, and RFs can all answer these questions by learning a discriminatory rule from labeled data. However, discriminative methods have two major limitations, both of which apply to cancer classification. The first limitation is theoretical: discriminative methods suffer from the problem of having to see all possible abnormalities in order to make an accurate and generalizable prediction (Sodemann et al., 2012). This is relevant to cancer because there exists countless ways in which a normal cell could become cancerous. As such, the label "cancer" does not encompass a known homogeneous group, but rather a heterogeneous collection of unknown types. It is simply not possible to anticipate the nature or extent of these "unknown unknowns" (Rumsfeld, 2002). The second limitation is practical: even for an ideal homogeneous cancer class, the tumor may occur too rarely for there to exist enough data to learn a meaningful discrimination rule. Discriminatory methods require sufficient sample sizes to learn a rule that tolerates the large variance observed in replicates of transcriptomic data (McIntyre et al., 2011). For these reasons, discriminatory methods are doomed to fail.

On the other hand, we expect that the possibility space for steady-state normal tissue is appreciably smaller than that of the aberrant tumor. By modeling this normal latent structure directly, we could learn a new rule that detects cancerous samples as a departure from normal. This follows the biological intuition that tumors themselves are anomalies of normal cellular physiology. The field of machine learning already has well-established methods that can detect anomalies in high-dimensional data, especially images, for the purpose of surveillance (Budhaditya et al., 2009). By transferring these methods to transcriptomic data, we can create an anomaly detector for tissue transcriptomes, a "tissue detector," that is capable of identifying cancer without ever seeing a single cancer example. In this report, we show that "tissue detectors" are sensible and accurate for the classification of cancer based on gene expression signatures. We do this by training an anomaly detection model on normal GTEx samples, then using it to accurately differentiate normal from cancerous TCGA samples. In presenting these results, we highlight future research directions for the detection of anomalous gene expression signatures.

2. METHODS

2.1. Data Acquisition

We acquired the combined GTEx and TCGA data from Wang et al. (2018), who harmonized them using quantile normalization and svaseq-based batch effect removal (Wang et al., 2018). After downloading the data in fragments per kilobase of transcript per million (FPKM), we chose six tissues that had large sample sizes in both GTEx and TCGA: breast, liver, lung, prostate, stomach, and thyroid. **Table 1** shows the number of healthy and cancer samples for each tissue.

2.2. Model Training

We refer to a predictive model and its threshold as a "tissue detector," of which we trained six (one for each tissue). To train the "tissue detector," we z-score standardized each gene within the GTEx training set, then performed a residual analysis of the GTEx training set. Residual analysis is based on the principle that most data have an underlying structure that can be largely reconstructed using a subset of the principal components, whereby the difference between the reduced representation and the original observations are termed the residues. Residual analysis uses the squared value of the residue as a proven way to measure the degree to which an observation is an outlier. For normally distributed data, the squared value of the residues follows a non-central χ^2 distribution. By comparing the norm of the residue for an unlabeled sample to a procedurally-selected threshold (corresponding to a stipulated false alarm rate), we have a predictive rule that decides whether to reject the null hypothesis and call that sample an anomaly (Jackson and Mudholkar, 1979). Our "tissue detector" method is available from https://github.com/thinng/tissue_detector.

2.3. Model Testing

After training each model on the GTEx data, we evaluated its performance on the respective TCGA data. For each sample in the test set, we calculated an anomaly score based on the distance between that sample and the model reference. We did this by projecting the sample to the principal component space and measuring its residue, where higher residue scores indicate that the sample is more anomalous. If the anomaly score is larger than the anomaly detection threshold, the sample is called abnormal (i.e., an outlier). Otherwise, the sample is called normal (i.e., an inlier). This allows us to differentiate between normal and cancerous TCGA samples without ever seeing a single cancer example. We repeated this procedure for increasingly smaller

TABLE 1 | This table shows the number of samples in each GTEx training set and TCGA test set, alongside the test set performance of that anomaly detector.

	GTEx (N)	TCGA (N)	TCGA (C)	Precision	Recall	Specificity	Accuracy	AUC
Breast	89	110	982	0.975	0.965	0.782	0.947	0.903
Liver	115	48	295	0.986	0.939	0.917	0.936	0.973
Lung	313	59	503	0.987	0.907	0.898	0.906	0.960
Prostate	106	48	426	0.949	0.742	0.646	0.732	0.734
Stomach	192	33	380	0.943	0.966	0.333	0.915	0.547
Thyroid	318	53	441	0.974	0.925	0.792	0.911	0.893

Precision and recall remain high for all classifiers, but specificity suffers for select tissues. This suggests that our "tissue detector," when it fails, has a bias toward viewing all TCGA samples as abnormal. The acronyms N and C refer to number of normal and cancerous samples, respectively.

subsets of the training data, with specificity averaged across ten bootstraps each.

By using the Wang et al. data, we can evaluate the utility of the anomaly detection method with all batch effects already removed. Nevertheless, we chose to use the GTEx data as the "normal" training set so that any residual batch effects between the GTEx and TCGA data would cause the "tissue detector" to call false positives (i.e., to call the healthy TCGA abnormal). For a robust and conservative estimate of performance, we focus our discussion on specificity (which is especially penalized by false positives).

3. RESULTS AND DISCUSSION

3.1. Cancer Is a Tissue Anomaly

For this study, we trained a "tissue detector" on each of the six tissues described in **Table 1**, using only the GTEx samples for training. We then evaluated its performance on withheld TCGA data by calculating an anomaly score for each TCGA sample and comparing it against the anomaly threshold: if the score is greater than the threshold, the sample is considered an anomaly (i.e., cancerous). **Figure 1** shows the (log-)ratio of per-sample anomaly scores relative to the tissue-specific anomaly threshold (y-axis) for each tissue (x-axis), faceted based on whether the sample is cancerous. Especially for breast, liver, lung, and thyroid data, our "tissue detector" not only recognizes most TCGA cancer samples as anomalies, but also recognizes most TCGA healthy samples as normal. On the other hand, anomaly detection is poor for prostate and stomach tissue. **Table 1** shows the precision, recall, and specificity for each "tissue detector." For almost all tissues, recall is better than specificity, meaning false positives are more common than false negatives. **Figure 2** shows the first two principal components of the best performing tissue (breast) with the worst performing tissue (stomach).

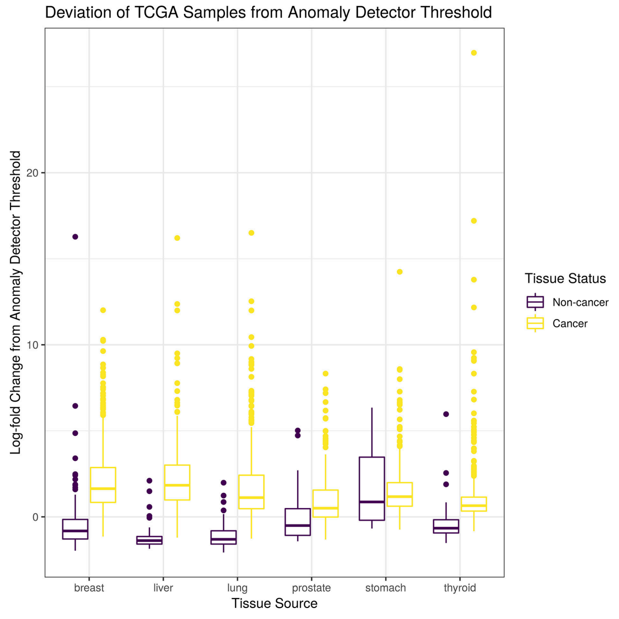


FIGURE 1 | This figure shows the (log-)ratio of per-sample anomaly scores relative to the tissue-specific anomaly threshold (y-axis) for each tissue (x-axis), faceted based on whether the sample is cancerous. The "tissue detector" calls any sample above the x-intercept threshold as an anomaly (i.e., cancerous). The threshold is selected procedurally during model training. This figure shows performance for TCGA test set only; no TCGA samples were included in the training set.

3.2. Detection Improves With More Normal Samples

We hypothesized that increasing the number of normal samples shown to the "tissue detector" during model training would improve its specificity, especially for the poorly performing prostate and stomach detectors. To test this hypothesis, we measured the specificity of each "tissue detector" as trained on increasingly smaller subsets of the GTEx data. **Figure 3** shows the specificity for each "tissue detector' (y-axis) according to the number of samples in the training set (x-axis). A pattern emerges: the inclusion of additional GTEx samples can improve the classification of TCGA samples, up until a point of diminishing returns.

4. CURRENT CHALLENGES

4.1. Translating Concept to Clinic

In this study, we used normal GTEx samples to train a model that could classify TCGA samples. We acknowledge that there is no direct clinical application for this experiment, since it is trivial to differentiate between cancer and non-cancer tissue using simple microscopy. As a proof-of-concept, we chose to use these data

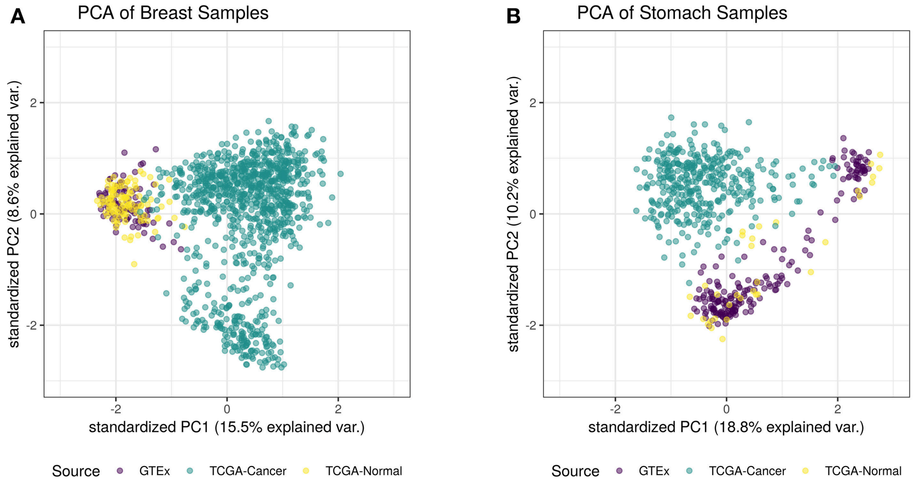


FIGURE 2 | This figure shows the first two principal components of the best performing tissue (breast; **A**) and the worst performing tissue (stomach; **B**), calculated using the log of all tissue data. While the healthy TCGA breast tissue is indistinguishable from normal GTEx tissue, the healthy TCGA stomach falls slightly outside the range of normal GTEx tissue. Although the healthy TCGA stomach tissue is markedly different than the cancer tissue, many of these samples look like anomalies from the perspective of the GTEx "tissue detector".

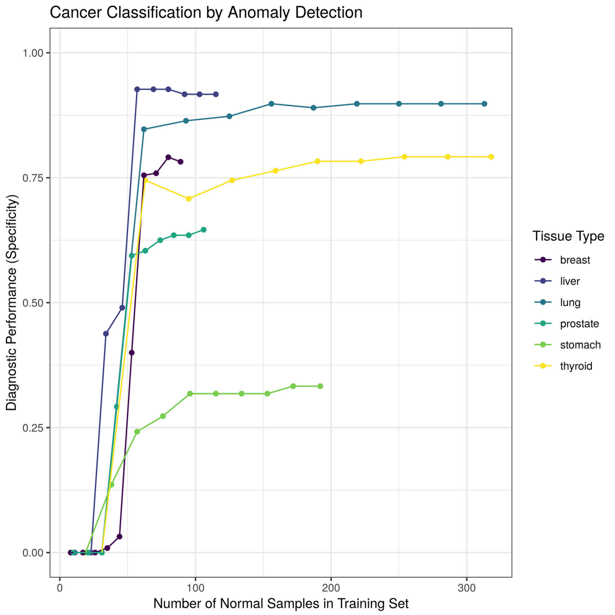


FIGURE 3 | This figure shows the specificity for each "tissue detector" (y-axis) according to the number of samples in the training set (x-axis). Performance is averaged across 10 bootstraps of the GTEx training set. This figure shows performance for TCGA test set only; no TCGA samples were included in the training set.

to demonstrate tissue anomaly detection because the data set is sufficiently large and publicly available. However, anomaly detection could suit many health surveillance applications. By changing the class of samples used in the training set, the meaning of "anomaly" changes. For example, if we include only benign tumors in the training set, then an anomaly detector might identify whether a biopsied tumor is potentially malignant (i.e., not benign). Likewise, using a training set of blood biomarkers for patients with surgically resected tumors might yield an anomaly detector that can identify whether a primary tumor has recurred. Other novel applications might include training a "tissue detector" on healthy lymphatic tissue to screen for lymphatic metastasis or on chemotherapy-sensitive tumor biopsies to screen for emerging drug resistance. Whatever the application, anomaly detection is unique in that it only requires that there exist data for the null state that is under surveillance: it is not necessary that researchers have characterized the full spectrum of the undesired outcome.

4.2. Data Integration

One challenge faced in the detection of anomalous gene expression signatures is the limited amount of data available for training and testing. Even as data sets get larger, anomaly detection will still benefit from the combination of multiple data sets, known as horizontal data integration (Tseng et al., 2012). However, horizontal data integration is complicated because every data set has intra-batch and inter-batch effects caused by systematic or random differences in sample collection. These differences could arise from a variety of biological factors (e.g., biopsy site, age, sex) or technical factors (e.g., RNA extraction protocol, sequencing assay), including latent factors unknown to the investigator (Leek et al., 2012). Although software like ComBat and sva can remove intra-batch biases, inter-batch biases may still remain. Indeed, inter-batch biases could explain why our "tissue detectors," when they fail, tend to view all TCGA samples as abnormal (though the "normal" TCGA samples do all come from sites adjacent to cancerous tissue). Although Wang et al. tried to harmonize the TCGA and GTEx data (Wang et al., 2018), the removal of inter-batch biases is non-trivial and further challenged by the prevailing need to preserve test set independence. Moreover, owing to how next-generation sequencing data measure the relative abundance of gene expression, these data also contain inter-sample biases that sit on top of the intra-batch and inter-batch biases (Soneson and Delorenzi, 2013; Quinn et al., 2018a). It remains an open question of how best to integrate multiple data sets. Non-parametric or compositional PCA-like methods could provide a suitable alternative to anomaly detection that is more robust to inter-batch and inter-sample biases.

4.3. Interpretability

Another challenge faced in the detection of anomalous gene expression signatures is the lack of transparency in the decision-making process. Although the concept of anomaly detection is intuitive, its implementation decomposes high-dimensional data into orthogonal eigenvectors that do not necessarily have any meaning to biologists. When examining these eigenvectors directly, it may be unclear how an anomaly detection model reached its decision. This makes it difficult to formulate new hypotheses to improve the model performance or elucidate the biological system. Future work should aim to improve the interpretability of anomaly detection methods. One approach might involve building a tool that visualizes which eigenvector components contributed maximally to each decision. If some constituent genes are consistently involved in misclassification, this could generate testable hypotheses. Similarly, one could try to characterize the biological importance of the maximally relevant eigenvectors through gene set enrichment analysis (GSEA), as done by Weighted Gene Correlation Network Analysis (Langfelder and Horvath, 2008). This would allow investigators to frame inlier and outlier distributions not only in terms of the constituent genes involved, but also in terms the biological pathways affected. This too could generate testable hypotheses. With these improvements, anomaly detection would become an interpretable and actionable classification strategy for many health surveillance applications.

5. SUMMARY

Technological advances have made it possible to measure the global gene expression signature of any biological sample at little cost. Already, there is a growing body of evidence that gene expression signatures can be used as biomarkers to diagnose cancer (Kourou et al., 2015). In this report, we present a novel application of anomaly detection to classify cancer based on gene expression signatures. By learning the latent structure of normal gene expression from a training set of normal samples, we created a "tissue detector" that can identify cancer without having seen a single cancer example. Our method contrasts with discriminatory methods, widely used in the biological sciences, which can only identify cancerous samples that resemble those that the algorithm already saw during training. In principle, discriminatory methods do not make sense for a disease like cancer where a one-of-a-kind gene expression signature is theoretically possible. Practically speaking, anomaly detection further benefits from normal samples being more readily available and easier to collect than abnormal samples: for any cancer, many more people do not have the cancer than do. Since the inclusion of additional normal samples can improve the specificity of anomaly detection, the curation of large normal data sets could open up the possibility of building diagnostic tests for extremely rare cancers.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: www.dx.doi.org/10.1038/sdata.2018.61.

AUTHOR CONTRIBUTIONS

TQ prepared the figures and drafted the manuscript. TN performed the primary analyses. SL pre-processed the data and supported primary analyses. SV supervised the project. All authors helped design the project and revise the manuscript.

ACKNOWLEDGMENTS

This manuscript has been released as a Pre-Print at Quinn et al. (2018b).

REFERENCES

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et al. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. U.S.A.* 96, 6745–6750.

Budhaditya, S., Pham, D., Lazarescu, M., and Venkatesh, S. (2009). "Effective anomaly detection in sensor networks data streams," in *2009 Ninth IEEE International Conference on Data Mining* (Washington, DC), 722–727.

Chan, H. P., Hadjiiski, L., Zhou, C., and Sahiner, B. (2008). Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography–a review. *Acad. Radiol.* 15, 535–555. doi: 10.1016/j.acra.2008.01.014

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.

Jackson, J. E., and Mudholkar, G. S. (1979). Control procedures for residuals associated with principal component analysis. *Technometrics* 21, 341–349.

Jensen, L. J., and Bateman, A. (2011). The rise and fall of supervised machine learning techniques. *Bioinformatics* 27, 3331–3332. doi: 10.1093/bioinformatics/btr585

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi: 10.1093/bioinformatics/bts034

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45:580. doi: 10.1038/ng.2653

McIntyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J., et al. (2011). RNA-seq: technical variability and sampling. *BMC Genomics* 12:293. doi: 10.1186/1471-2164-12-293

Quinn, T. P., Erb, I., Richardson, M. F., and Crowley, T. M. (2018a). Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* 34, 2870–2878. doi: 10.1093/bioinformatics/bty175

Quinn, T. P., Nguyen, T., Lee, S. C., and Venkatesh, S. (2018b). Cancer as a tissue anomaly: classifying tumor transcriptomes based only on healthy data. *bioRxiv* 426395. doi: 10.1101/426395

Rangayyan, R. M., Ayres, F. J., and Leo Desautels, J. E. (2007). A review of computer-aided diagnosis of breast cancer: toward the detection of subtle signs. *J. Franklin Inst.* 344, 312–348. doi: 10.1016/j.jfranklin.2006.09.003

Rumsfeld, D. (2002). *Department of Defense News Briefing*. Washington, DC: Federal News Source, Inc.

Sodemann, A. A., Ross, M. P., and Borghetti, B. J. (2012). A review of anomaly detection in automated surveillance. *IEEE Trans. Syst. Man Cybernet.* 42, 1257–1272. doi: 10.1109/TSMCC.2012.2215319

Soneson, C., and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14:91. doi: 10.1186/1471-2105-14-91

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* 40, 3785–3799. doi: 10.1093/nar/gkr1265

Wang, Q., Armenia, J., Zhang, C., Penson, A. V., Reznik, E., Zhang, L., et al. (2018). Unifying cancer and normal RNA sequencing data from different sources. *Sci. Data* 5:180061. doi: 10.1038/sdata.2018.61

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

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

Integration of Machine Learning Methods to Dissect Genetically Imputed Transcriptomic Profiles in Alzheimer's Disease

Carlo Maj[1]†, Tiago Azevedo[2]†, Valentina Giansanti[3]†, Oleg Borisov[1], Giovanna Maria Dimitri[2], Simeon Spasov[2], Alzheimer's Disease Neuroimaging Initiative, Pietro Lió[2]* and Ivan Merelli[3]**

[1] Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn, Germany, [2] Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom, [3] National Research Council, Institute for Biomedical Technologies, Milan, Italy

The genetic component of many common traits is associated with the gene expression and several variants act as expression quantitative loci, regulating the gene expression in a tissue specific manner. In this work, we applied tissue-specific cis-eQTL gene expression prediction models on the genotype of 808 samples including controls, subjects with mild cognitive impairment, and patients with Alzheimer's Disease. We then dissected the imputed transcriptomic profiles by means of different unsupervised and supervised machine learning approaches to identify potential biological associations. Our analysis suggests that unsupervised and supervised methods can provide complementary information, which can be integrated for a better characterization of the underlying biological system. In particular, a variational autoencoder representation of the transcriptomic profiles, followed by a support vector machine classification, has been used for tissue-specific gene prioritizations. Interestingly, the achieved gene prioritizations can be efficiently integrated as a feature selection step for improving the accuracy of deep learning classifier networks. The identified gene-tissue information suggests a potential role for inflammatory and regulatory processes in gut-brain axis related tissues. In line with the expected low heritability that can be apportioned to eQTL variants, we were able to achieve only relatively low prediction capability with deep learning classification models. However, our analysis revealed that the classification power strongly depends on the network structure, with recurrent neural networks being the best performing network class. Interestingly, cross-tissue analysis suggests a potentially greater role of models trained in brain tissues also by considering dementia-related endophenotypes. Overall, the present analysis suggests that the combination of supervised and unsupervised machine learning techniques can be used for the evaluation of high dimensional omics data.

Keywords: eQTL, gene expression imputation, GTEx, variational autoencoder, support vector machine, deep learning, recurrent neural networks, Alzheimer's

Edited by:
Dominik Heider,
University of Marburg, Germany

Reviewed by:
Shefali S. Verma,
University of Pennsylvania,United States
Huiluo Cao,
The University of Hong Kong, Hong Kong
Yu Li,
King Abdullah University of Science and Technology, Saudi Arabia,
in collaboration with reviewer HC

Correspondence:
Carlo Maj
cmaj@uni-bonn.de
Pietro Lió
pl219@cam.ac.uk
Ivan Merelli
ivan.merelli@itb.cnr.it

†These authors have contributed equally to this work

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

Received: *18 April 2019*
Accepted: *10 July 2019*
Published: *03 September 2019*

Citation:
Maj C, Azevedo T, Giansanti V, Borisov O, Dimitri GM, Spasov S, Alzheimer's Disease Neuroimaging Initiative, Lió P and Merelli I (2019) Integration of Machine Learning Methods to Dissect Genetically Imputed Transcriptomic Profiles in Alzheimer's Disease. Front. Genet. 10:726. doi: 10.3389/fgene.2019.00726

INTRODUCTION

Nowadays researchers can access omics data at different levels, such as genomics (e.g., dbGaP[1]), transcriptomics (e.g., GEO expression[2]) and also at multi-omics levels (e.g., GTEx[3], Encode[4]). Given the advancement of high-throughput technologies, the increasing availability of omics data can be expected over time. This will allow researchers to better analyze complex systems characterized by many interacting features as the biological systems.

Traditional analytical methods on omics data, such as Genome-wide association study (GWAS) and differential expression analysis, usually rely on univariate approaches with specific statistical modelling (Visscher et al., 2017; McDermaid et al., 2018). These approaches, despite being robust, are limited in detecting potential combinatorial effects in the underlying biological system. Indeed, biological networks can be highly complex with many feedback regulatory loops (Franco and Galloway, 2015). A comprehensive analysis of interaction effects is not feasible with traditional approaches due to the combinatorial explosion of the input factor space (Berger et al., 2013).

On the other hand, machine learning methods have proved to be efficient for the analysis of high dimensional complex systems, although the application of machine learning methods in omics data is still relatively uncommon due to the limited interpretability of the outcome of machine learning frameworks (Li et al., 2016). In this work, we investigate the applicability of different machine learning methods on omics data using, as a case study, matrices of tissue-specific predicted transcriptomic profiles in Alzheimer's disease (AD). AD is a progressive neurodegenerative disorder, representing the predominant form of dementia (Wang et al., 2017), and is characterized by progressive deterioration of memory and cognitive functions that can be tested with different clinical tests (Kirsebom et al., 2017). The pathophysiology of AD involves the formation of the characteristic extracellular amyloid plaques and intracellular neurofibrillary tangles (Kuznetsov and Kuznetsov, 2018).

A lot of research has been done in order to identify the genetics factor contributing to AD. In cases of specific familiar forms of AD, which are recurrent among family members and are characterized by early onset (i.e., age < 65), disease causing mutations in specific genes have been identified, namely amyloid precursor protein (APP), Presenilin 1 PSEN1 and Presenilin 2 PSEN2 (Piaceri et al., 2013). This is not the case of the most common sporadic AD forms, characterized by late onset (age > 65), representing about 95% of AD cases (Bali et al., 2012), for which the “4 allele of Apolipoprotein E (APOE) is the only strong identified genetic risk factor (Dorszewska et al., 2016).

However, the relatively high heritability also of sporadic AD, estimated to be around 60% to 80% (Van Cauwenberghe et al., 2016), combined with the identification of a number of genetic risk loci from GWAS, suggests the presence of a polygenic component in late onset AD (Escott-Price et al., 2015). Indeed, GWAS hits can be associated with different biological pathways, such as cholesterol and lipid metabolism, immune system, inflammatory response, and endosomal vesicle cycling (Lambert et al., 2013). Moreover, several susceptibility loci are localized in gene-dense regions, but it remains unknown which genes of these regions are responsible for the association (Van Cauwenberghe et al., 2016). In fact, identifying the functional role of variants in intergenic regions is not a trivial process, since the related genes might not be the closest to the loci (e.g., chromatin 3D structure can place in proximity relatively distant region in the primary DNA sequence) (Dekker et al., 2013). Moreover, many complex phenotypes have a polygenic architecture, in which many variants have minor effects over a phenotype, and polygenic risk score modeling is capable of finding significant genetic associations for traits with no monogenic causes, but with relatively high heritability (Chatterjee et al., 2016).

Different works show a co-localization between Expression Quantitative Loci (eQTL) and GWAS hits indicating that the biological effect of non-coding variants can be exerted through the regulation of gene expression (Hormozdiari et al., 2016; Wen et al., 2017), that is a polygenic trait in which many variants may be involved. Indeed, different tools model the combined effect of eQTL signals, considering both strong functional SNP effects and additive effects for modest-strength signals (Gamazon et al., 2015; Gusev et al., 2016). Conducting gene association on the basis of the genetic component of gene expression regulation, also called Transcription Wide Association Study (TWAS), proved to be particularly efficient in finding associations with many traits (Gusev et al., 2016).

There are many advantages in testing the genetic component of gene expression rather than evaluating the nominal variant GWAS association: I) the aggregation of multiple eQTL into one gene can boost the association by including additive effect among variants; II) genes are more interpretable biological unit in comparison with variants; III) the statistical power is increased due to the reduction of multiple-comparison tests from hundreds of thousands/million variants (before/after imputation) to the order of thousands of genes (after filtering for gene expression heritability); IV) eQTL are tissue specific and therefore it is possible to perform gene association analysis in the target tissue for the phenotypes and also in secondary tissues for potential peripheral biomarkers (e.g., blood).

Noteworthy, the evaluation of the solely genetic component of gene expression is less comprehensive than the actual gene expression analysis, but has the advantage to focus only on the genetic/heritable component, avoiding environmental confounding effects (Gamazon et al., 2015). Since polygenic effects can be expected also at gene expression level, given the complexity of biochemical systems, performing multi-gene evaluation can provide greater insights concerning potential biological associations (Marigorta et al., 2017). Therefore, machine learning and deep learning methodologies have proved to be efficient at identifying transcriptomic profiles associated with specific phenotypes, considering different input data, such as measured RNA-seq data (Wang et al., 2018), single cell expression (Hu et al., 2016), and also imputed transcriptomic data (Gottlieb et al., 2017).

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

[2] https://www.ncbi.nlm.nih.gov/geo/

[3] https://gtexportal.org/home/index.html

[4] https://www.encodeproject.org/

In this work, we tested multiple machine learning and deep learning approaches to study multi-tissue imputed transcriptomic profiles in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort (Weiner et al., 2013). Noteworthy, the analysis of imputed transcriptomic profiles on ADNI data has been already performed at single gene level identifying, suggesting potential specific gene-tissue associations with amyloid deposition (Hohman et al., 2017). In the following sections we introduce the supervised and unsupervised methods we exploited in this work, the results achieved combining these approaches, and a discussion of the achieved outcomes.

METHODS

Machine Learning Methods in Bioinformatics

Machine Learning (ML) algorithms have proved to be particularly useful for the analysis of complex big biological data (Olson et al., 2017). For instance ML has been applied to detect epistasis within the human genome (McKinney et al., 2006) suggesting that ML can reveal non-linear behavior in biological systems. In the same direction, more recent deep learning approaches have been profitably exploited to analyze genotype/phenotype associations (Min et al., 2017) as well as to extract relevant information from many data modalities, including text, images, and sounds (Li et al., 2019).

Deep learning methods follow a data-driven approach and are therefore well-designed to detect nonlinear-behaviors, which are relatively common in natural systems (Tang et al., 2019). Networks can vary depending on the number of layers and type of nodes and not all of them perform equally well on different data typology. Convolutional Neural Networks (CNN) are generally applied to recognize objects in a pattern, Recurrent Neural Networks (RNN) to analyze temporal data, but it is not mandatory to use any kind of network only for a specific task. For instance, CNNs were successfully used to predict the enhancer-promoter interactions with DNA sequences (Zhuang et al., 2019) and for accurate clustering of sequences (Aoki and Sakakibara, 2018). RNNs were used instead for predicting transcription factor binding sites (Shen et al., 2018) and to dissect the regulation of mRNA to protein-coding translation (Hill et al., 2018).

Noteworthy, also variational autoencoders (VAEs) showed good performance in capturing biologically relevant feature in gene expression data analysis (Way and Greene, 2017a). VAEs are part of a large branch of deep learning architectures, the so called generative models (Goodfellow, 2016). These architectures are based on an encoding-decoding approach and, differently from the standard autoencoders, they assume a stochasticity in the modelling of the data. The original input matrices of features are compressed in a lower dimensional space, the so called encoding phase, and are reconstructed back in a second step, called decoding phase. Both phases are composed by neural networks. VAEs have seen increasing success in many different applications in the last few years, among the unsupervised methodologies recently developed, and they are widely used in different types of data such as time series, images or gene expressions (Goodfellow, 2016; Goodfellow et al., 2016; Way and Greene, 2017b).

Tissue Specific Gene Expression Imputation

Data used for the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). ADNI was launched in 2003 as a public-private partnership led by Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), other biological markers, clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). In the present work, we analyzed the ADNI1-GWAS dataset including gene array genotyping data for 808 samples available on ADNI portal.

Rigorous quality control has been performed. Namely, samples have been checked for sex, missing genotype rates lower than 0.05 and heterozygosity levels $F < 0.2$, while variants with Hardy–Weinberg p-value $< 1e - 10$ have been removed. Then, using the tool by W. Rayner[5] we checked SNPs for strand consistency, allele names, position, Ref/Alt assignments and minor allele frequency (MAF) in comparison to the reference panel. In order to increase the available genetic information, we imputed our data using Sanger Imputation Server[6] exploiting Eagle2 for phasing (Loh et al., 2016) and Positional Burrows–Wheeler Transform (Durbin, 2014), considering Haplotype Reference Consortium version 1.1 (McCarthy et al., 2016) as reference panel. As a postimputation quality control, we removed variants with info quality level < 0.6. Genotype calls with posterior probability < 0.9 were set to missing. Post-QC imputed data was used to estimate gene expression regulation across the different samples.

In order to predict the genetic component of gene expression, we used PrediXcan that evaluates the aggregate effects of cis-regulatory variants (within 1MB upstream or downstream of genes of interest) on gene expression *via* an elastic net regression method (Gamazon et al., 2015). PediXcan needs a reference dataset in which both genome variation and gene expression levels have been measured to build prediction models for gene expression. We exploited already available models trained on GTEX data[7] to impute tissues specific transcriptomic profiles in a total of 42 tissues (we excluded sex specific tissues, e.g., prostate, ovary, etc.). The imputed transcriptomic profiles were subsequently analyzed using different machine learning approaches (**Figure 1**). On the one hand, unsupervised machine learning methods were used to analyze data structure, on the other hand, supervised methods were used to test for the presence of "signal" compared to AD related phenotypes.

[5] http://www.well.ox.ac.uk/wrayner/tools
[6] https://imputation.sanger.ac.uk/
[7] https://gtexportal.org/home

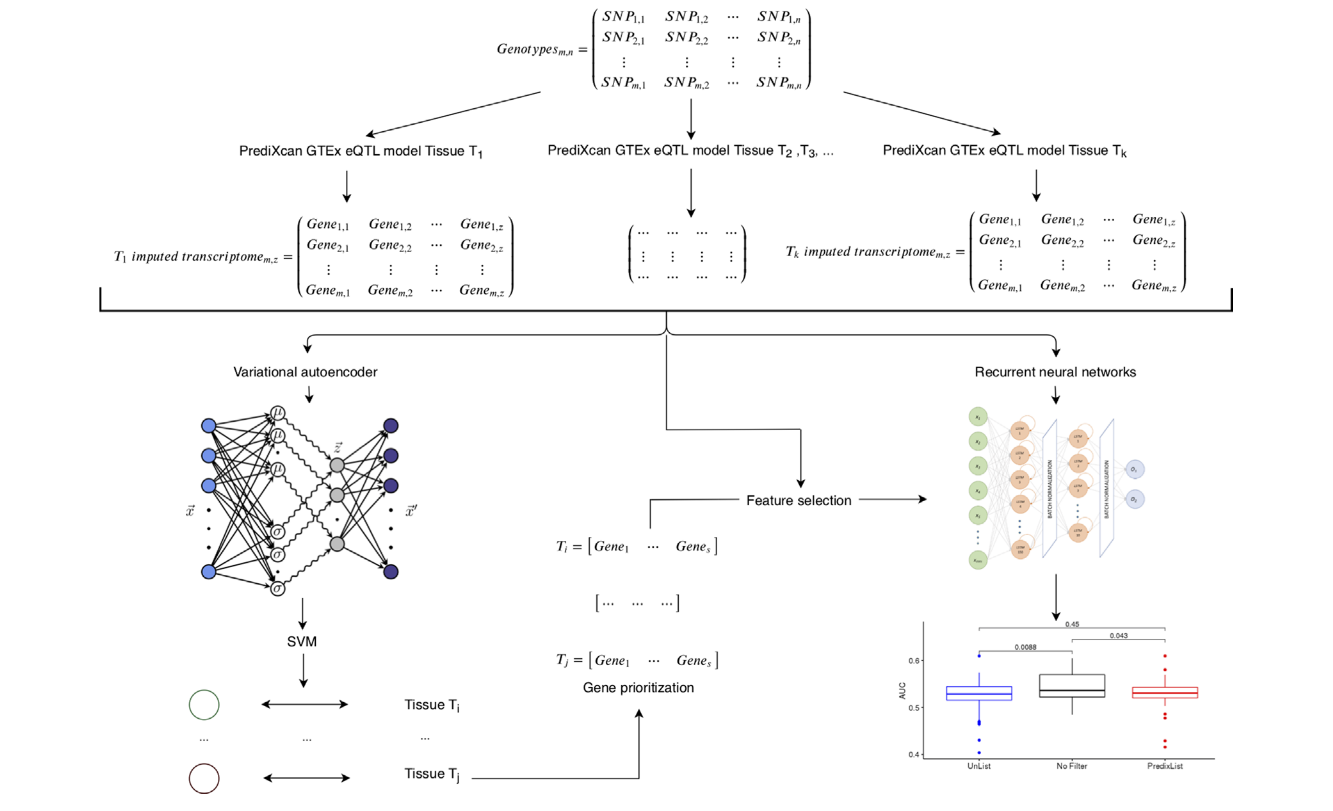

$Genotypes_{m,n} = \begin{pmatrix} SNP_{1,1} & SNP_{1,2} & \cdots & SNP_{1,n} \\ SNP_{2,1} & SNP_{2,2} & \cdots & SNP_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ SNP_{m,1} & SNP_{m,2} & \cdots & SNP_{m,n} \end{pmatrix}$
PrediXcan GTEx eQTL model Tissue T_1
PrediXcan GTEx eQTL model Tissue T_2 ,T_3, ...
PrediXcan GTEx eQTL model Tissue T_k
T_1 imputed transcriptome$_{m,z}$ = $\begin{pmatrix} Gene_{1,1} & Gene_{1,2} & \cdots & Gene_{1,z} \\ Gene_{2,1} & Gene_{2,2} & \cdots & Gene_{2,z} \\ \vdots & \vdots & \vdots & \vdots \\ Gene_{m,1} & Gene_{m,2} & \cdots & Gene_{m,z} \end{pmatrix}$
T_k imputed transcriptome$_{m,z}$ = $\begin{pmatrix} Gene_{1,1} & Gene_{1,2} & \cdots & Gene_{1,z} \\ Gene_{2,1} & Gene_{2,2} & \cdots & Gene_{2,z} \\ \vdots & \vdots & \vdots & \vdots \\ Gene_{m,1} & Gene_{m,2} & \cdots & Gene_{m,z} \end{pmatrix}$
Variational autoencoder
$\vec{x}$
$\vec{z}$
$\vec{x}'$
μ
σ
SVM
Tissue T_i
Tissue T_j
Feature selection
$T_i = [Gene_1 \cdots Gene_s]$
$[\cdots \cdots \cdots]$
$T_j = [Gene_1 \cdots Gene_s]$
Gene prioritization
Recurrent neural networks
BATCH NORMALIZATION
AUC
0.6
0.5
0.4
0.45
0.0088
0.043
UnList
No Filter
PredixList

FIGURE 1 | Framework of integrative analysis of multi-tissues expression profiles. Starting from genotyping data (*m* individuals per *n* variants) we imputed tissues specific transcriptomic profiles (for any tissue T_i, where $i = 1,\ldots, k$) by means of cis-eQTL PrediXcan models trained on GTEx data. Variational autoencoder followed by support vector machine (SVM) latent dimension-tissue match on the imputed gene expression matrices (*m* individuals per *z* genes) is used as a feature selection to identify the most relevant genes per tissue ($T_i = gene_1,\ldots, gene_s$ where *i* is the i_{th} tissue and *s* in the number of prioritized genes) to provide as input of the recurrent neural network classifier.

Gene Prioritization

Gene prioritization was performed considering as input the predicted transcriptomic matrices from ADNI1-GWAS (excluding sex-specific tissues) resulting in a total of 42 tissues with 808 samples each (42 × 808 = 33, 936 samples overall). We performed an independent analysis involving 528 "cases", that included people affected by dementia and/or with cognitive dysfunction (AD and MCI) for a total of 528 × 42 = 22, 176 input data, and 280 controls including healthy subjects for a total of 280 × 42 = 11, 760 input data. Each sample was comprised of 24, 203 genes in total.

To identify relevant genes we used variational autoencoders (VAEs) with a single hidden layer with a dimension of 42 units, hence matching the number of tissues. We adapted the code publicly provided by Way and Greene (2017b) to implement our VAE's architecture. In the encoding phase, the network inputs are the original dataset features representation $\vec{x}$. These are transformed by means of non-linear activation functions in a hidden representation that we denominated $\vec{z}$ and that we assume being characterized by a Gaussian probability density function. In this phase the 2 latent representations of μ and λ of the distribution are learned.

The second part of the architecture that we denoted as the decoder is again built as a neural network. The input this time is the vector $\vec{z}$ i.e. the latent stochastic representation of the input dataset and the output will be the reconstructed representation $\vec{x}'$ of the original input vector $\vec{x}$. A representation of the VAE architecture can be seen in **Figure 1**. The loss function of the VAE consists of two parts: the first part being the reconstruction loss (negative log-likelihood) and the second part being the function expressing the Kullback–Leibler (KL) divergence considering the learned hidden distribution and *a priori* Gaussian distribution (Wetzel, 2017).

The first term of the loss function is considered over the encoder distribution of the hidden representation and it "encourages" the decoding phase to correctly reconstruct the input data (Altosaar, 2019). KL divergence is used to enforce the similarity between the distribution of the latent representation and the normal distribution.

We used separate VAEs to encode the gene expression of the cases and healthy classes. Original data include positive (upregulated genes) and negative values (downregulated genes). In order to compute VAE analysis, input data have been scaled between 0 and 1. Noteworthy, different genes can be present in different tissues while VAE pipeline requires an equal number of gene as input, thus NaN (non-existent/Not a Number) values during VAE input preprocessing were set to 0. The input samples were randomly split in training (80%) and test sets (20%) using a stratified approach to maintain the same proportion of samples per tissue. We used the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0:001 over 75 epochs over the data, rectified linear units during the encoding stage, sigmoid activation during the decoding stage, batch size of 500, and warmup (κ) of 1. Hyperparameters were manually selected using a VAE that was not used further in the analysis, to achieve optimal reconstruction performance without overfitting. The entire autoencoding procedure was repeated 75 times separately for the healthy and AD classes in order to study the repeatability of results.

The main goal of the unsupervised analysis was to identify the up or down-regulation of certain genes in specific tissue types in cases and healthy samples. We used a two-step procedure to achieve this association: we identified the tissue(s) encoded in each latent dimension unit of the VAE models, and then we identified the genes most strongly connected to the given latent dimension unit.

In order to identify the tissue(s) encoded in each latent dimension, we used the activations of the hidden layer in the VAEs as an input feature to 42 binary Support Vector Machine (SVM) classifiers, one for each tissue. We trained each SVM classifier to predict whether the input sample to the VAE belonged to a specific tissue relying on the activation value of a single unit from the embedded latent dimension of the VAE. We repeated this tissue-latent unit association procedure for each tissue and each unit in the hidden VAE layer. We performed a 5-fold stratified cross-validation using a linear SVM ($C = 1$ with class weight balance), thus running a total of 5 × 42 × 42 SVM classifiers for each VAE (a 5-fold cross validation procedure, for 42 binary classifiers, for each one of the 42 hidden layer's unit). We considered a given latent VAE unit to be predictive of a specific tissue type, hence associated with it, if the $F1$ score was greater than 0.8. We found that some hidden units encode more than 1 tissue type.

It is noteworthy to mention that we tried other unsuccessful approaches. Firstly, we tried to use a single VAE with both cases and controls, trying to find subclusters besides the tissues which cluster very well (see **Figure 2**) in the VAE's latent dimension as well as in the original data. We also tried to use a single VAE for each tissue in separate. No obvious structures were found when trying to match the results of t-SNE algorithm with all the available phenotypes, including case/control status. Filtering the input for genes within each tissue that show nominal significance for case/control status using standard simple univariate tests did not improve the results. Filtering genes with $R^2 > 0.15$ of expression prediction using the same threshold as in Hohman et al.′s work (Hohman et al., 2017) did not improve the results as well. In order to understand the features important for classification, we also implemented a saliency map approach. This method is able to detect where the attention of the network (VAEs in our study) is focused (Itti et al., 1998), which can be seen as a sensitivity analysis approach. Saliency maps are generally applied in computer vision but, they can be used in other areas. In our case, the maps were computed on the encoder part of the VAEs and the information extracted is the importance of each gene in the analysis, which is coded as an rgb color code. From this analysis we were not able to identify significative patterns in the input data.

Considering the VAE used in this work, the association of the genes with the latent dimension units can be performed solely relying on the magnitude of the corresponding network weights. Given that each VAE has a single hidden layer, each latent dimension unit is connected directly to every output unit, i.e. reconstructed gene, *via* a linear transformation. Since each reconstructed gene is a summation of the weighted contribution of each latent unit, we could rank the relative importance of the units in the hidden layer relying on the magnitude of the weights. Thus, we selected the 100 most positive and 100 most negative weights for each latent unit encoding a given tissue. This resulted in a set of 100 upregulated and 100 downregulated genes, respectively for each of the trained VAEs. The entire association procedure was performed for the 75 VAEs from healthy and AD samples. We counted the total number of times a given gene was considered up or downregulated by our association procedure and kept it if it appeared more than three times overall. As a result, we produced a list of up or down regulated genes associated with each of the 42 types of tissues. We used this list as an input for pathway enrichment analysis.

In order to perform enrichment analysis, we used Fast Gene Set Enrichment Analysis (FGSEA), a tool developed by Sergushichev et al. (Sergushichev, 2016). The approach implemented by FGSEA deals with quantitative data having inherently directionality like gene expression. The model is based on gene statistic array $S = Si,\ldots Sn$ where N is the number of samples and $Si > 0$ represent over-expression of gene i while $Si < 0$ represent down-expression. The absolute value of Si represents a magnitude of the change. The list of gene sets P of length m usually contains groups of genes that are commonly regulated in certain biological process. To quantify a co-regulation of genes in a gene set p Subramanian et al. (2005) introduced a gene set enrichment score function $sr(p)$ that uses gene rankings (values of S). Given a gene set p the more positive is the value of $sr(p)$ the more enriched the gene set is in positively-regulated genes g with $Sg > 0$, accordingly, negative $sr(p)$ corresponds to enrichment of negatively regulated genes. To deal with multiple-comparison issues an empirical p-value is computed by randomly sampling gene sets of the same size of p.

The lists of downregulated and upregulated genes per tissue (referred as *List-unsupervised*) have been considered also as a feature selection step to build prediction models. We also tested other approaches to identify the most relevant genes as considering: I) nominal significantly associated genes from logistic association test between predicted gene expression levels and phenotype status (referred as *List-PrediXcan*), II) nominal associated genes derived by the combination of single tissue-trait association using generalized Berk–Jones test (referred as *List-UTMOST*) obtained with UTMOST tool (Hu et al., 2019).

Phenotype Prediction Models From Imputed Transcriptomic Matrices

Several supervised analysis techniques were tested in order to understand which one could achieve better results in identifying

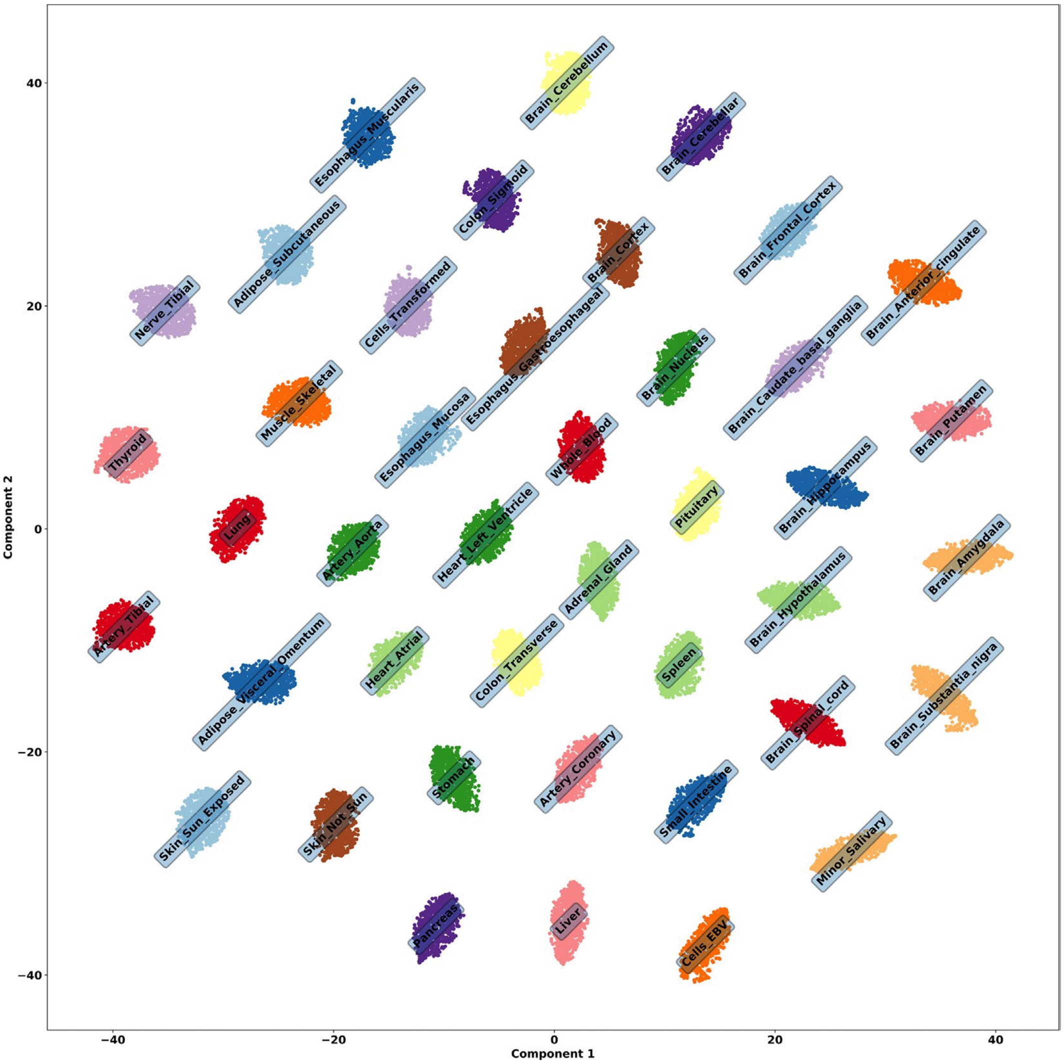


FIGURE 2 | t-SNE embedding of tissue genes, run using the 42 activations on the latent dimension of a VAE to check the embedded structure of all samples. It is obvious that the latent activations are encoding information about each tissue.

cases and controls from the transcriptomic profiles: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Deep Learning networks. The latter are known to achieve better results compared to other machine learning methods, especially when the relationships between the observed features is not supposed to be linear (LeCun et al., 2015).

Since we imputed data according to specific tissues, we searched the model that would perform better among them. For this reason, we randomly selected 6 of the 42 tissues (Adipose Subcutaneous, Artery Aorta, Brain Spinal, Colon Transverse, Thyroid, Whole Blood) and trained the models on 600 of the 808 samples from ADNI1-GWAS, considering that the dataset is slightly unbalanced, as it contains more AD samples (528) than controls (280). SVM, RF and LR were not capable of learning how to classify cases and controls, since they assigned the samples only to the majority class. Concerning Deep Learning, the first accomplishment was understanding the appropriate architecture to elaborate transcriptomic data: we tested two Dense Neural Networks (DNN), two CNNs and an RNN.

The first DNN (DNN-1) consisted of 6 layers with respectively 800, 500, 400, 200, 40 and 2 nodes (called neurons). The second DNN (DNN-2) tested consisted of only three layers with 800, 200 and 2 neurons. The first CNN (CNN-1) had 6 layers: a convolutional layer of 10 filters, a convolutional layer of 5 filters after which a dropout regularization was applied, another convolutional layer of 5 filters, a dense layer of 200 neurons with a dropout, and two dense layers of 100 and 2 neurons in the end. The second CNN (CNN-2) was a pure convolutional network of two convolutional layers of 10 and 5 filters, with a dropout regularization applied between them, and a dense layer with 2 neurons as the output layer. The RNN had 3 layers: two Long Short-Term Memory cells (LSTM) with output dimension of 30 and 20 and a final dense layer of 2 neurons.

Looking at the preliminary training results (**Table 3**) we decided to select and optimize the RNN, manually searching the optimal network's size and then identifying the hyperparameters with the Grid Search algorithm (batch size = 100, epochs = 100). The final architecture consisted of the input and output layers and two hidden LSTM layers of 150 and 10 output dimensions. After every hidden layer a batch normalization was applied to maintain the mean activation close to 0 and the activation standard deviation close to 1. The input layer dimension was equal to the number of genes characterizing the tissue transcriptomic profile, while the output layer was a dense layer of dimension two to make possible the classification of the samples in AD and not-AD.

Considering all the 42 tissues, we had the chance to perform two types of analysis: a tissue-specific analysis and a cross-tissue analysis. In the tissue-specific analysis, we trained models on transcriptomic data specific for each tissue. Therefore, we implemented predictive models that could impute the case/control condition on new transcriptomic data related to the same tissue. The input dimensions of the networks were in the order of thousands, but different for every tissue: the minimum was 2,041 characterizing the Brain Substantia Nigra tissue, and the maximum was reached by the Thyroid tissue with 9,655.

The aim of the cross-tissue analysis was, on the other hand, to observe the similarities between tissues in relationship with the Alzheimer's disease. Models were trained on each single tissue, taking as input the genes shared by all the 42 tissue transcriptomic profiles (24, 203). The column reporting the information for a gene was filled with zeros if it was not possible to impute the transcriptomic profile of that gene in a specific tissue. Comparing the maximum number of genes imputed for the tissues and the total number of genes identified in all the analysis, it was clear that the new arranged matrices of 24, 203 genes for 808 samples were particularly sparse. The models were then used to impute the case/control condition on tissues different to the one used for the training.

Both in single tissue and cross-tissue analyses all the models were trained on 600 samples from ADNI1-GWAS and the tests were performed on the remaining 208 samples. The network architecture was in all cases the one in **Figure 1**, adjusting the input dimension according to the different analysis. A 10-fold cross validation was applied and models compiled with the Adam optimizer and the binary cross-entropy as the optimization score function. The monitored scores were the accuracy, area under the curve (AUC), precision, recall, and *F*1. The saliency map was applied in the first LSTM layer, therefore we could observe if some samples were more informative than other for the classification purpose. Keras[8] and Scikit-learn[9] Python libraries were used, built on top of TensorFlow[10] to implement the networks.

We then worked on features selection to find groups of genes that were likely to improve the model performance regarding the samples partition in case/control, both in the single-tissue and cross-tissues approaches. The identification of such groups in single-tissue analysis can bring to the determination of tissue specific markers, on the other hand in the cross-tissues section we could focus on the set of genes that explained the relationship between tissues. We used three different filter lists: *List-unsupervised*, *List-PrediXcan* and *List-UTMOST* (see **Supplementary Materials Section 3**). Using these lists the input dimensions for all the tissues decreased: the number of unique genes identified by the List-unsupervised was 2,016, 4,984 with List-PrediXcan. List-UTMOST (649 genes) was used only in the cross-tissue analysis as it doesn't provide tissue-specific information.

All the steps described above (except the architecture selection and saliency map) were also performed considering Cognitive Decline over time rather than diagnosis at screening. This dataset consisted of 528 samples (some samples did not have this information), 281 controls and 247 cases. Cognitive Decline has been calculated by considering the difference between the Mini-Mental State Examination (MMSE) score 4 years after recruitment and the MMSE score at recruitment. Then, regardless of the original recruitment diagnosis, we classified the samples into two groups: one group showing no cognitive decline (difference equal or greater than 0) and the other showing a cognitive decline (difference minor than 0). The genes imputed for each tissue were therefore the same in ADNI1-GWAS dataset and Cognitive Decline dataset. To consider the effect of AD related variables, we also performed the same analyses by stratifying by sex and early/late onset for dementia and AD [using 65 years of age as a cutoff (Roberts and Petersen, 2014)] as well as for carrier and noncarrier of APOE ∊4 isoform.

RESULTS

We predicted the genetic component of gene expression across 42 non-sex-specific tissues for all the samples included in ADNI1-GWAS dataset. We exploited tissues specific eQTL models available on precictDB[11] and used PrediXcan tool[12] to derive tissue specific matrices representing individual levels of the genetic component of gene expression. The gene levels obtained by these sample matrices represent transcriptomic profiles based on eQTL across tissues in the analyzed dataset.

In the present work the matrices of imputed expression were analyzed using several machine learning strategies to identify potential tissue specific transcriptomic profiles associated with cognitive decline in Alzheimer's.

Gene-Based Results Per Tissue

We runned t-SNE (Maaten and Hinton, 2008) using the 42 activations on each latent dimension of a VAE to check the embedded structure of all samples, whose result can be seen in **Figure 2**. Although interpretations of Euclidean distances between points in a t-SNE plot is not straightforward (Wattenberg et al., 2016), it is clear from the clusters that information about tissues are being encoded. Indeed, we were able to identify associations between latent dimensions of VAE and tissue.

[8] https://keras.io/

[9] https://scikit-learn.org/stable/

[10] https://www.tensorflow.org/

[11] http://predictdb.org/

[12] https://github.com/hakyimlab/PrediXcan

TABLE 1 | Most upregulated and downregulated genes from the brain nucleus.

	Downregulated		Upregulated	
	AD-MCI	**CTR**	**AD-MCI**	**CTR**
Brain nucleus	ENSG00000230850.3	ABHD14A	ENSAP2	AL356475.1
	GMPR2	ATP2B4	KLF1	F2
	C1QC	BDKRB2	EEF1A1P19	NRIP2
	SUN3	C1QC	RP5-1068B5.3	RP11-704J17.5
	RP11-662J14.1	PXN	RP11-321A17.3	RP11-321A17.3

The evaluation of the weights associated with the latent dimension (see *Methods*) allow us to rank gene importance per tissue considering case/control status. **Table 1** shows the most upregulated and downregulated genes from Brain Nucleus. Check **Supplementary Table S1** for complete information over all 42 tissues.

The saliency map implementation returned not useful information. If taken individually, genes don't have much impact: it is evident also with this result that the AD phenotype is due to a combination of many genes and environmental factors.

In order to investigate the presence of specific gene expression regulation associated with case/control status we considered the lists of tissue-specific up and down regulated genes derived by VAE analysis. Additionally, for each tissue we considered the genes that were differentially regulated in cases but not in controls, that is representing a disease-specific signature. The enrichment analysis have been performed considering Gene ontology[13], KEGG[14] and reactome[15] and pathway databases (Croft et al., 2013; Kanehisa et al., 2016). Complete enrichment analysis results are available as supplementary files (see **Supplementary Materials Section 1**) while significant enrichment tissues specific pathways after FDR correction are shown in **Table 2**.

Interestingly enrichment analysis shows the presence of tissue specific signal in a specific brain tissue (i.e., brain nucleus) concerning pathways involved in gene expression regulation and in immune-related pathways in colon (**Figure S2**). The most significant alterations in brain pathways concern the brain nucleus accumbens (basal ganglia) region. Interestingly, this region has been found to be associated with AD (Nie et al., 2017; Nobili et al., 2017; Li et al., 2018). Instead, the detected downregulation of immune system pathways in cases in comparison to controls could indicate a higher level of inflammation in dementia. This is in line with the association observed between inflammatory bowel diseases and AD (McCaulley and Grush, 2015; Sochocka et al., 2019). Given the pivotal role of APOE (Liu et al., 2013) in AD a specific evaluation was performed to evaluate the effect of APOE related genes.

APOE gene expression is not predicted by gene expression imputation GTEx based models, due to the absence of eQTL explaining a relevant fraction of APOE expression level. However, AD susceptibility due to APOE isoforms (∫2, ∫3 and ∫4), which are well known to confer a different risk for AD depending on the presence of missense coding variants, are associated with APOE gene functionality and can be independent from the genetic component of gene expression regulation. We investigated if other genes directly interacting with APOE, according to string functional database[16], have a significant association in our analysis (see **Supplementary Materials Section S3**).

One of the 11 genes identified, namely *APOC2* (Shao et al., 2018), is among the top differentially regulated genes from variational autoencoder gene prioritization list in brain putamen, an area of the brain associated with AD (Coupé et al., 2019). Interestingly, the same gene is also the only one (among the 11 APOE interacting genes) significantly associated with AD according to a transcription wide association analysis performed according to a GWAS on AD in UK Biobank dataset (Marioni et al., 2018) and public available on TWAS hub[17]. This suggests a potential role for *APOC2* associated with the gene expression regulation and, interestingly, a recent work showed that the methylation profile in such a gene (which in turn affect gene expression) is associated with AD (Shao et al., 2018).

Tissue-Specific and Cross-Tissues Classification

To understand which network performs better on different tissues, we tested five models on six sample tissues. In **Table 3**, accuracy and AUC obtained during their preliminary 10 cross-validation training on 600 of 808 samples are reported: although all methods could perform well at least on one tissue during the training, in that phase only the RNN was capable of reaching an accuracy higher than 90% for all of them. Therefore we decided to optimize the RNN and obtained the network structure described in *Phenotype Prediction Models From Imputed Transcriptomic Matrices*, which was then applied for the single-tissue and cross-tissue analysis on ADNI1-GWAS and Cognitive Decline dataset.

Without the feature selection, we observed a great performance during the training in terms of AUC, accuracy, precision, recall and *F*1 scores (see **Supplementary Materials Section 2**) on both datasets. On test set (composed of 208 samples for tissue for ADNI1-GWAS and 128 for Cognitive Decline) the metrics reached values below expectations, with AUCs near 0:5 especially for ADNI1-GWAS.

[13] http://geneontology.org/

[14] https://www.genome.jp/kegg/pathway.html

[15] https://reactome.org/

[16] https://string-db.org/cgi/network.pl

[17] http://twas-hub.org/

TABLE 2 | Significant tissue-pathways enrichment analysis using Reactome database.

Tissue	Pathway	pval	padj	ES	NES	Genes	
Colon sigmoid	Immune system	3.8E–04	1.2E–02	–5.4E–01	–2.3E+00	CAP1 RASGRP4 RASGRP4 YES1 SIGLEC8 CD47 SELL CALM1	FBXO21 CLEC7A CLEC7A SEC61A1 IL13 HLA-DPB1 KIF11
Brain nucleus	Generic transcription pathway	3.0E–03	1.8E–02	7.2E–01	2.1E+00	ZNF688 ZKSCAN8 ZNF445	RRAGC ZNF697 CASP6
Brain nucleus	RNA polymerase II transcription	3.0E–03	1.8E–02	7.2E–01	2.1E+00	ZNF688 ZKSCAN8 ZNF445	RRAGC ZNF697 CASP6
Brain nucleus	Gene expression (transcription)	3.0E–03	1.8E–02	7.2E–01	2.1E+00	ZNF688 ZKSCAN8 ZNF445	RRAGC ZNF697 CASP6
Colon sigmoid	Adaptive immune system	3.0Ev03	3.3E–02	–6.1E–01	–2.1E+00	FBXO21 SEC61A1 HLA-DPB1 KIF11	YES1 SIGLEC8 SELL CALM1
Colon sigmoid	Innate immune system	2.5E–03	3.3E–02	–6.5E–01	–2.0E+00	CAP1 CLEC7A CD47 CALM1	RASGRP4 YES1 SELL

TABLE 3 | Preliminary networks training performance on six sample tissues: accuracy (Acc) and area under the curve (AUC).

Network	Adipose subcutaneous		Artery aorta		Brain spinal		Colon transverse		Thyroid		Whole blood	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
DNN-1	37.50	0.513	37.33	0.503	64.00	0.538	87.67	0.862	64.50	0.503	39.83	0.516
DNN-2	64.50	0.5	64.50	0.5	90.17	0.892	64.50	0.5	64.50	0.5	64.50	0.5
CNN-1	63.00	0.5	76.92	0.721	77.50	0.901	78.50	0.770	64.08	0.5		0.491
CNN-2	95.83	0.948	64.50	0.5	94.83	0.943	64.50	0.5	96.00	0.95	95.67	0.947
RNN	96.17	0.953	95.67	0.951	94.67	0.942	95.33	0.946	95.33	0.946	94.67	0.939

On ADNI1-GWAS (**Figure 3**), models trained for single-tissue analysis improved their AUCs thanks to the *List-unsupervised* and *List-PrediXcan* feature selection: when the AUCs were below 0:5, the filters application returned a score above the threshold for at least one list. We did not observe a major impact of a list in this phase but the *t*-test confirmed a significant improvement compared to the no filter approach (p-value = 0.001474 for *List-unsupervised* and p-value = $2.693e-06$ for *List-PrediXcan*). Models trained for the cross-tissue analysis instead had a less evident improvement with the lists filter: only the List *unsupervised* returned a slightly significant improvement (p-value = 0.04084). *List-UTMOST* did not give any improvement and, as we could not use it on single-tissue models, we decided not to further analyze it.

Cognitive Decline models performed better than ADNI1-GWAS, both in single-tissue and cross-tissue analysis (**Figure 4**). The lists application on Cognitive Decline models also led to an improvement for tissues with borderline or below the threshold performance (**Figure S5**), reaching AUCs between 0:51 and 0:6. On cross-tissue models we obtained a significant p-value = 0.008766 for List-unsupervised and p-value = 0.04346 for List-PrediXcan.

Comparing the two lists on ADNI1-GWAS, List-unsupervised showed the bigger improvement on cross-tissue models: the *t*-test returned a p-value of 0:009123, but on single-tissue the difference was not significant. Also on Cognitive Decline we observed a slightly major impact of List-unsupervised both for the single-tissue and cross-tissue models. In **Figure 5**, a focus on the improvement achieved with the filter on the Brain tissue is shown in both datasets, in **Figure S4** the evaluation for all tissues is shown.

Figure 6 reports, by columns, the AUC achieved by ADNI1-GWAS cross-tissue models when they were applied on other tissues from the same dataset. The top heatmap describes the relationships between tissue when no filter is applied: we could observe that models trained on Brain

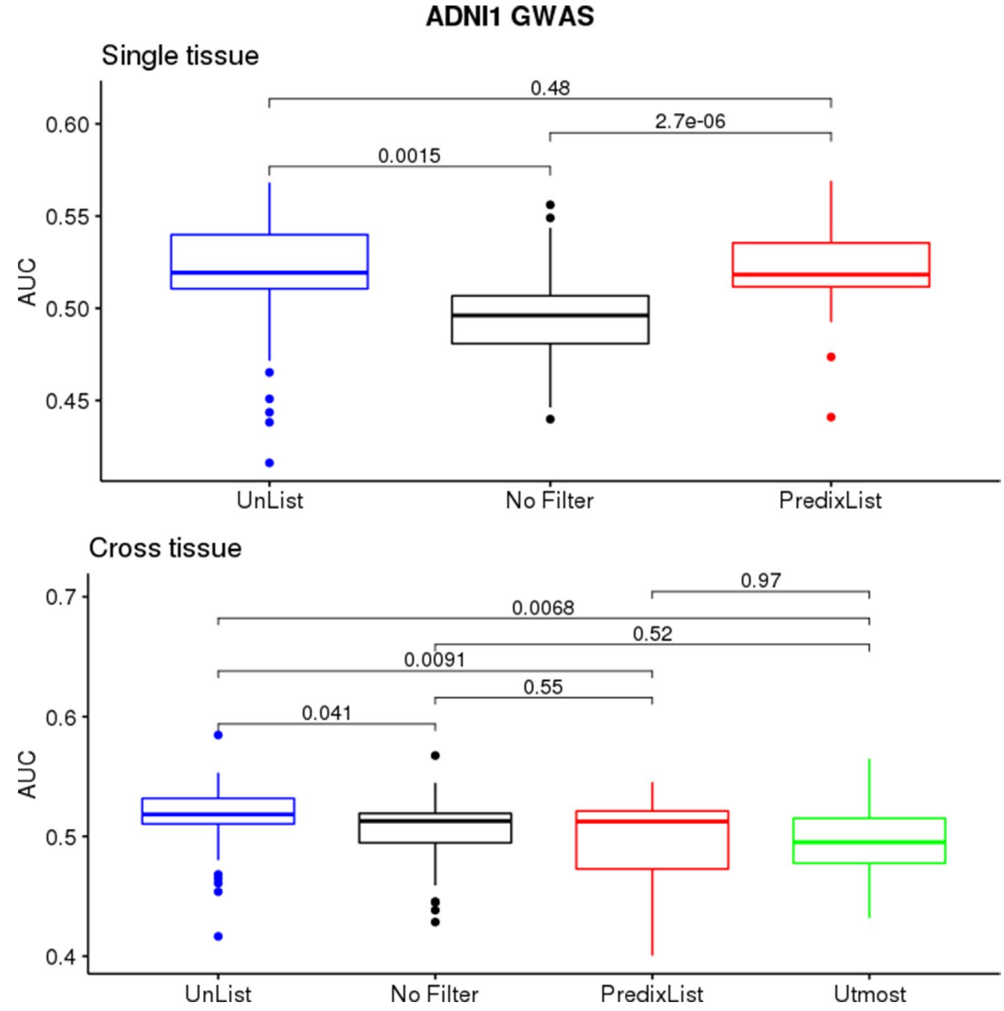


FIGURE 3 | ADNI1-GWAS feature selection evaluation. The single-tissue models (top panel) significantly improved their ability to classify case/control condition thanks to both *List-unsupervised* (blue) and *List-PrediXcan* (red) compared to the no filtering approach (black). On cross-tissue models (bottom panel), where there is also the performance with the *List-UTMOST* (green), the improvement was less evident.

tissues, if they were able to correctly identify the AD subjects on a non-Brain tissue, they could do the same on all the other non-Brain tissues. Instead, models trained on non-Brain tissue could identify AD-MCI/CTRL subjects only on a subset of tissues. We performed the same analysis on ADNI1-GWAS models filtered by List-PrediXcan and List-unsupervised, respectively the middle and bottom heatmaps of **Figure 6**: List-unsupervised removed all the information of cross-tissue relationships, when instead List-PrediXcan mitigate them, pointing out the non-Brain models relationships.

We also tested the stratification for sex, age, APOE effect, and AD condition on ADNI1-GWAS dataset for single-tissue and cross-tissue analysis. It returned no considerable variation in the performance. The saliency map application was also not informative: each sample has the same importance. Lastly, we performed the filter analyses on Cognitive Decline, pointing out the same results (**Figure S6**).

DISCUSSION

In the present work we dissected the tissue specific genetic component of gene expression in association to AD related cognitive decline. Our analysis consisted on the imputation of tissue specific gene expression profiles by using a TWAS-like approach (Mancuso et al., 2017). However, contrary to the standard TWAS analysis, we did not specifically focus on univariate analysis (e.g., gene association based on logistic or linear regression). Instead, we dissected individual transcriptomic levels using different machine learning approaches. We believe that our approach can be of particular interest since is capable of capturing data structure and non-linear behaviour in the system. In fact, it is well known that gene expression levels are not independent, since many genes are actually correlated in terms of regulation (Michalopoulos et al., 2012) and functionality, which means that also epistatic interactions can play a major role in the regulation of biochemical pathways (Sameith et al., 2015).

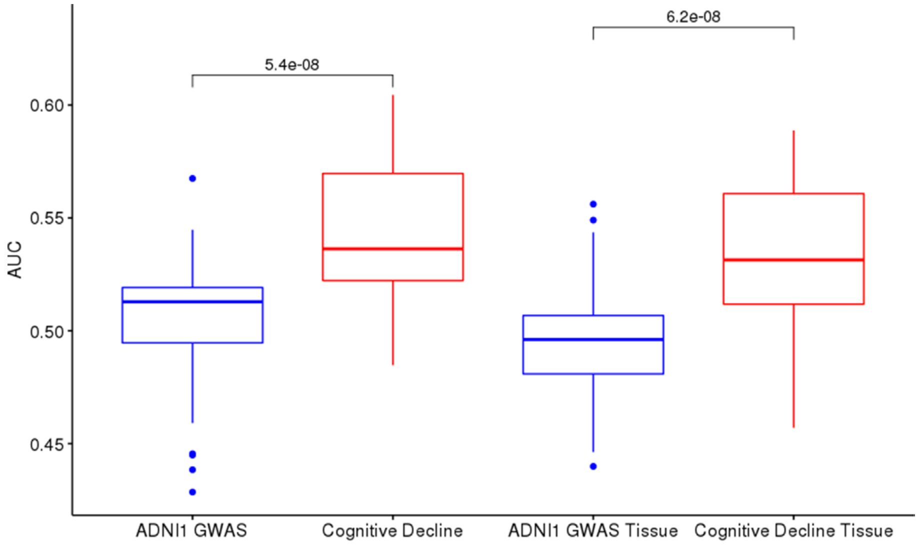


FIGURE 4 | ADNI1-GWAS and Cognitive Decline comparison: Cognitive Decline (red boxes) returns higher AUCs on test sets than ADNI1-GWAS (blue boxes) both in cross-tissue models (left) and in single-tissue models (right).

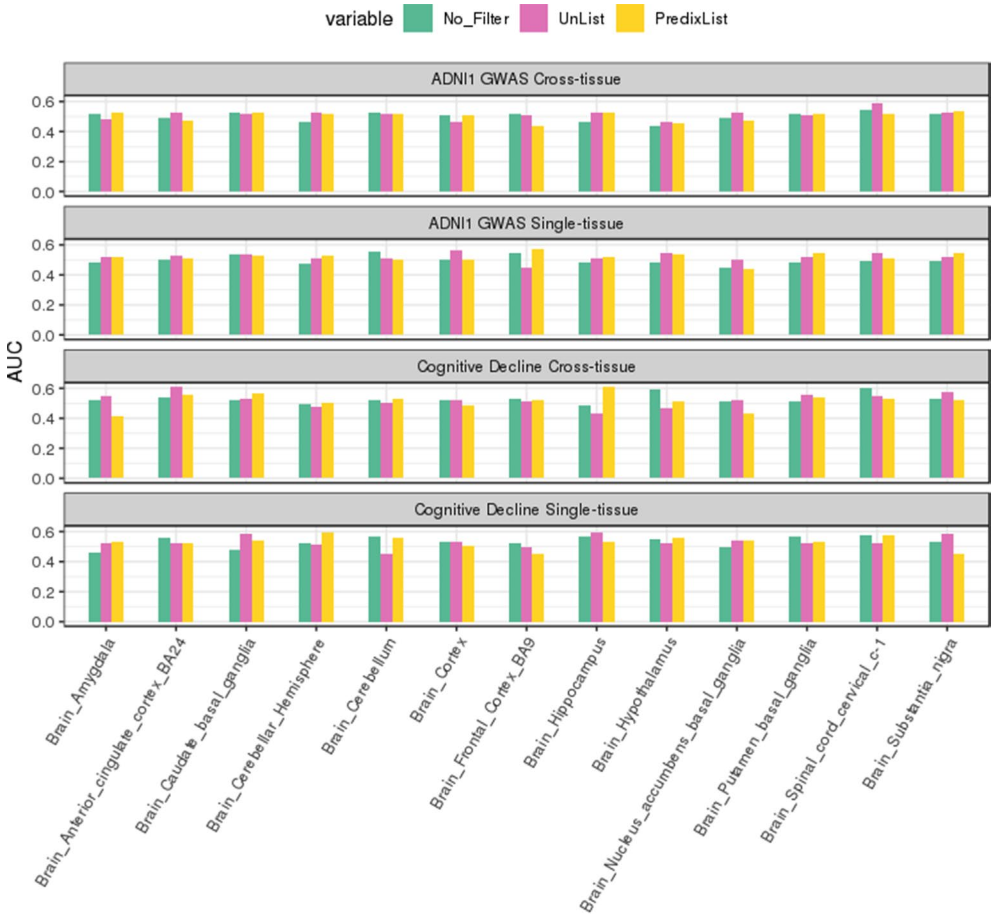


FIGURE 5 | Brain tissues analysis. In green the AUCs on test sets for the no filter application are reported, in red for *List-unsupervised* and in yellow for *List-PrediXcan*. The top two panels report respectively the cross-tissue and single-tissue models performance on ADNI1-GWAS dataset, the third and fourth panels on Cognitive Decline. In both datasets, feature filtering improved the classification in almost all the Brain tissues.

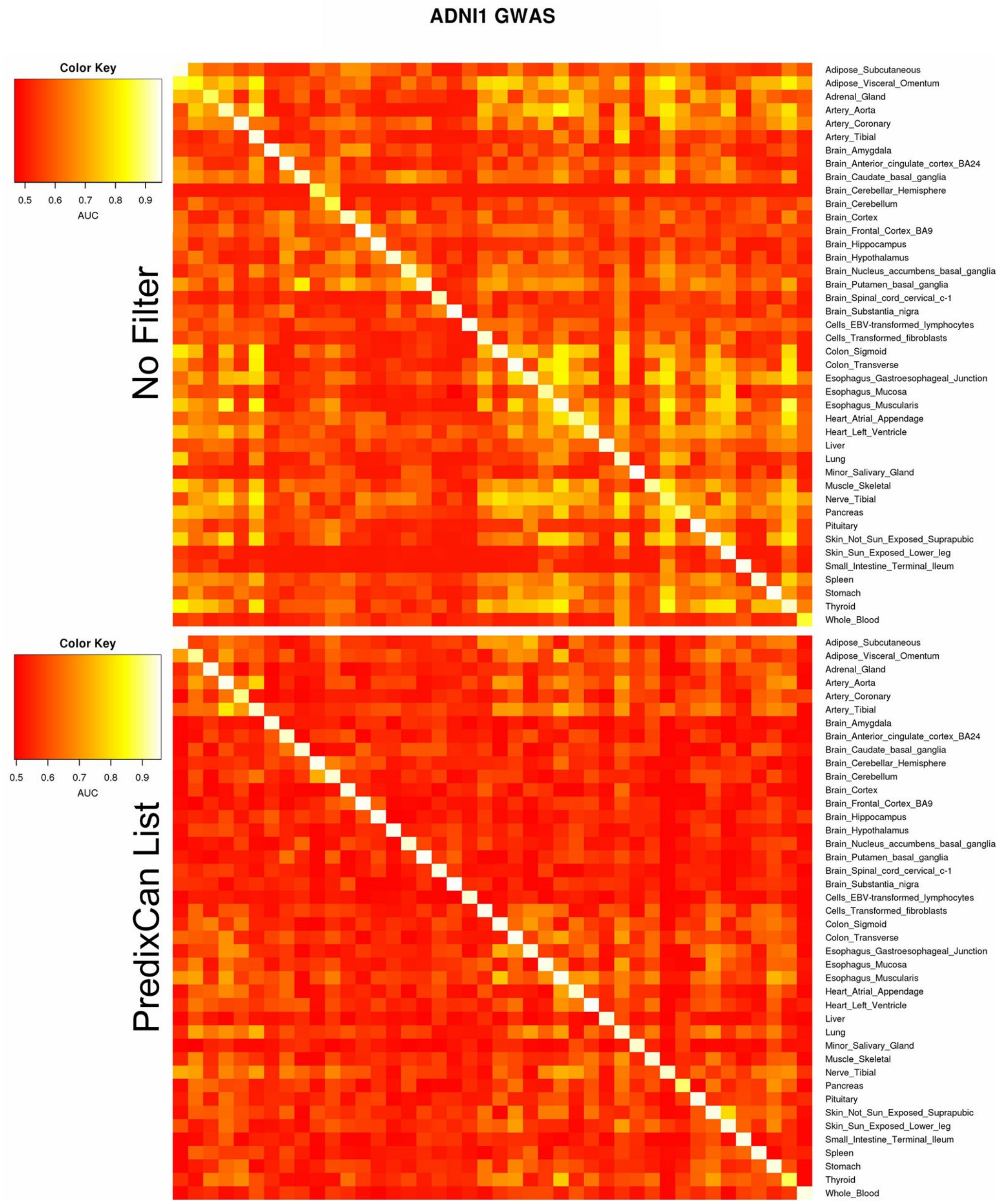


FIGURE 6 | Continued

Interestingly, we observed that a combination of unsupervised and supervised machine learning methods on matrices of predicted expression provided complementary information that can be integrated in order to get new insights in gene expression regulation. On one hand, the VAE combined with enrichment analysis suggests the presence of a specific biochemical pathways alteration in dementia occurring in a specific brain area and in the gut. The identified alteration occur in brain nucleus, a brain region found to be associated with AD by several studies (Cho et al., 2014; Wang et al., 2014; Kuhn et al., 2015; Liu et al., 2015).

This alteration seems to be related to the regulation of gene expression and 436 therefore possibly associated to tissue-specific pathways regulation. Instead, the enriched pathways in gut are related to immune systems and noteworthy, it is well established that immune system dysfunctions can lead to a greater increase of inflammation in AD (Serpente et al., 2014; Heppner et al., 2015; Le Page et al., 2018). These results suggest that our analytical approach can identify relevant biological alterations occurring in AD. Noteworthy, enrichment analysis identified alteration in biological pathway specifically in a brain area and gut, which is in line with the presence of a gut-brain axis dysfunction in AD. Indeed, several researchers pointed out that

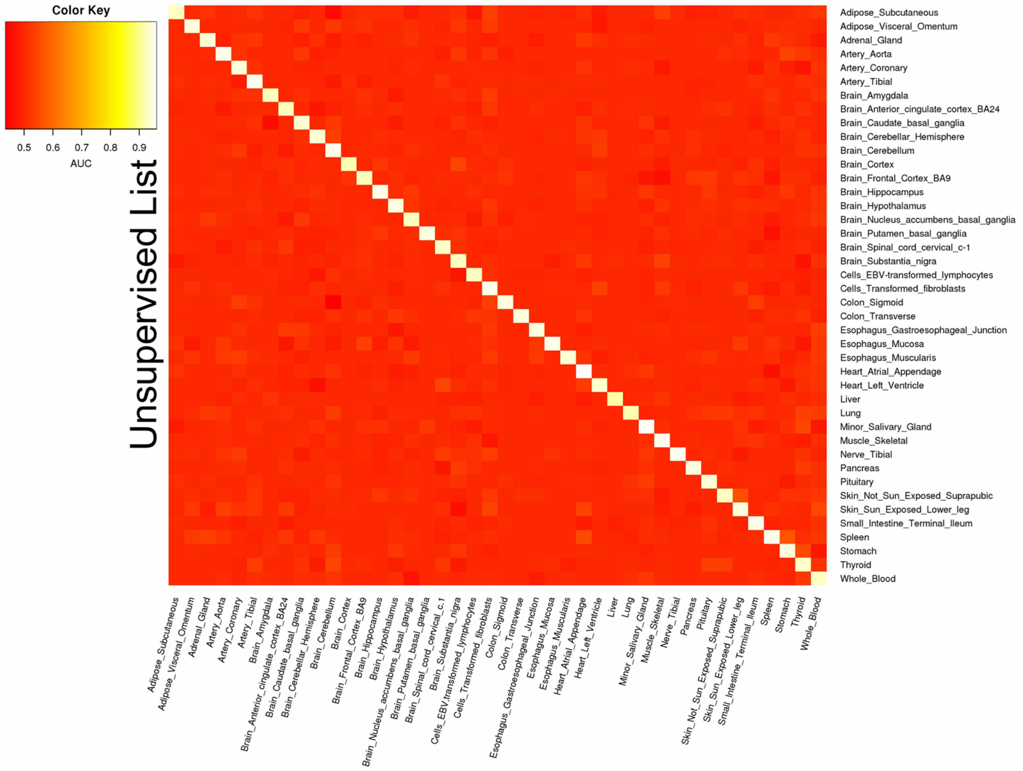


FIGURE 6 | ADNI1-GWAS cross-tissues performance. By column we can observe how much a model trained on a tissue is able to recognize without mistakes (AUC) AD/non-AD subjects from data related to different tissues. On the diagonal for each tissue the AUC obtained for that model during the training is reported. The top panel reports the cross-tissue results without any filter application, the middle panel when using *List-PrediXcan* and the bottom using *List-unsupervised*.

brain-gut axis can be associated with many neurological disorders (Giau et al., 2018; La Rosa et al., 2018).

In the present work, APOE genotype has not been directly included as covariate in prediction models since our aim was to identify other genetic factors that can explain part of the missing heritability on the established polygenic component in AD (Escott-Price et al., 2017; Tosto et al., 2017). However, APOE is expected to be by far the most influencing risk factor for late onset AD. Though estimation of APOE contribution on the heritability component of AD is still not well defined, ranging from 10% to 28% of the overall genetic heritability (Van Cauwenberghe et al., 2016; Stocker et al., 2018). Moreover, in the present work, gene-expression derived genetic signals neglect not-eQTL effects and therefore we have limited analytical power. This justifies the relatively low AUC values in comparison to other prediction models in AD, including the complete genome-wide polygenic signal and using APOE as a covariate (Escott-Price et al., 2017; Tosto et al., 2017). Our aim was indeed to test whether or not there is a genetic signal associated with AD that could be apportioned to tissue specific gene-expression regulation rather than identify a prediction model. It is also known that genetics is just one of the component involved in AD susceptibility and therefore the use of multimodal data (e.g., imaging data, clinical features, metabolomic, and environmental factors) should be taken into account in order to build a reliable classifier in term of translational application (Sapkota et al., 2018). Despite that, our classification models were still capable of finding a signal between cases and controls (overall AUC > 0:5) suggesting that part of the genetic signal in AD related dementia can be associated with tissue-specific gene expression regulation. Moreover, we observed that feature selection can play a major role in the performance of deep learning networks classification.

We are aware that our work presents some limitations. We performed a genetic association with dementia by considering ADNI data evaluating the solely genetic component of gene expression, which neglects other potential genetics effect not related to gene-expression regulation. Our models are also limited by the current version of GTEx data, which has a relatively small size, therefore it is expected that over time new models will optimize eQTL estimation leading to more precise analyses of the genetic component of gene expression. We also focused on non-sex specific tissues, since we wanted to study general potential alterations not involving sex-specific organs, but this could also be a limitation given the different prevalence of AD in females and males (Mazure and Swendsen, 2016).

CONCLUSION

In the present work, we performed an analysis of the predicted genetic component of gene expression in ADNI1-GWAS dataset in association with AD cognitive decline. We dissected the predicted tissue specific gene expression by means of different supervised and unsupervised machine learning approaches. Our results suggest that a framework including unsupervised and supervised methods in data-analysis can provide complementary information and thus leading to better insights into the underlying system.

In particular, variational autoencoder pre-processing of input data proved to be efficient for features selection prior to the implementation of deep learning classification models. However, the limited AUC prediction performance of the developed models suggests that the evaluation of the solely genetic component of gene expression by exploiting up to date available GTEx models is currently under-powered in comparison to genome-wide polygenic risk score modeling.

This is not surprising since we are neglecting the effect of non-eQTL variants. On the other hand, we can disclose tissue specific effects and reveal potential biological mechanisms associated with a given phenotype. In this regard, our analysis showed that brain tissues are more associated with dementia status and that inflammatory processes in brain-gut axis can play a role in AD.

AUTHOR'S NOTE

Data used in preparing this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, many investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: http://adni.loni.usc.edu/. Supplementary data and the code used in this work is available at https://github.com/imerelli/DeepNeuro.

AUTHOR CONTRIBUTIONS

CM, TA and VG equally contributed to the work. They conceived the idea and developed the algorithms. OB, GD, and SS contributed to data analysis. PL and IM supervised the whole study. All authors contributed to final revision of the manuscript.

ACKNOWLEDGMENTS

We would like to thank John Hatton (CNR-ITB) for proofreading the manuscript. ADNI data collection and sharing was funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. TA is funded by the W. D. Armstrong Trust Fund, University of Cambridge, UK. GD is supported by the EPSRC International Doctoral Scholarship. SS is supported 511 by the Engineering and Physical Sciences Research Council [EP/L015889/1].

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00726/full#supplementary-material

REFERENCES

Altosaar, J. (2019). Tutorial - what is a variational autoencoder? Accessed: 2019-03-11.

Aoki, G., and Sakakibara, Y. (2018). Convolutional neural networks for classification of alignments of non-coding RNA sequences. *Bioinformatics* 34, i237–i244. doi: 10.1093/bioinformatics/bty228

Bali, J., Gheinani, A. H., Zurbriggen, S., and Rajendran, L. (2012). Role of genes linked to sporadic Alzheimer's disease risk in the production of β-amyloid peptides. *Proc. Natl. Acad. Sci.* 109, 15307–15311. doi: 10.1073/pnas.1201632109

Berger, B., Peng, J., and Singh, M. (2013). Computational solutions for omics data. *Nat. Rev. Genet.* 14, 333. doi: 10.1038/nrg3433

Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* 17, 392. doi: 10.1038/nrg.2016.27

Cho, H., Kim, J.-H., Kim, C., Ye, B. S., Kim, H. J., Yoon, C. W., et al. (2014). Shape changes of the basal ganglia and thalamus in Alzheimer's disease: a three-year longitudinal study. *J. Alzheimers Dis.* 40, 285–295. doi: 10.3233/JAD-132072

Coupé, P., Manjon, J. V., Lanuza, E., and Catheline, G. (2019). Lifespan changes of the human brain in Alzheimer's disease. *Sci. Rep.* 9, 3998. doi: 10.1038/s41598-019-39809-8

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2013). The reactome pathway knowledgebase. *Nucleic Acids Res.* 42, D472–D477. doi: 10.1093/nar/gkt1102

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14, 390. doi: 10.1038/nrg3454

Dorszewska, J., Prendecki, M., Oczkowska, A., Dezor, M., and Kozubski, W. (2016). Molecular basis of familial and sporadic Alzheimer's disease. *Alzheimers Dis. Res.* 13, 952–963. doi: 10.2174/1567205013666160314150501

Durbin, R. (2014). Efficient haplotype matching and storage using the Positional Burrows–Wheeler Transform (PBWT). *Bioinformatics* 30, 1266–1272. doi: 10.1093/bioinformatics/btu014

Escott-Price, V., Shoai, M., Pither, R., Williams, J., and Hardy, J. (2017). Polygenic score prediction captures nearly all common genetic risk for Alzheimer's disease. *Neurobiol. Aging* 49, 214–2e7. doi: 10.1016/j.neurobiolaging.2016.07.018

Escott-Price, V., Sims, R., Bannister, C., Harold, D., Vronskaya, M., Majounie, E., et al. (2015). Common polygenic variation enhances risk prediction for Alzheimer's disease. *Brain* 138, 3673–3684. doi: 10.1093/brain/awv268

Franco, E., and Galloway, K. E. (2015). "Feedback loops in biological networks," in *Computational Methods in Synthetic Biology* (Springer), 193–214. doi: 10.1007/978-1-4939-1878-2_10

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* 47, 1091. doi: 10.1038/ng.3367

Giau, V., Wu, S., Jamerlan, A., An, S., Kim, S., and Hulme, J. (2018). Gut microbiota and their neuroinflammatory implications in Alzheimer's disease. *Nutrients* 10, 1765. doi: 10.3390/nu10111765

Goodfellow, I. (2016). Nips 2016 tutorial: Generative adversarial networks. arXiv preprintarXiv:1701.00160.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

Gottlieb, A., Daneshjou, R., DeGorter, M., Bourgeois, S., Svensson, P. J., Wadelius, M., et al. (2017). Cohort-specific imputation of gene expression improves prediction of warfarin dose for African Americans. *Genome Med.* 9, 98. doi: 10.1186/s13073-017-0495-0

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* 48, 245. doi: 10.1038/ng.3506

Heppner, F. L., Ransohoff, R. M., and Becher, B. (2015). Immune attack: the role of inflammation in Alzheimer disease. *Nat. Rev. Neurosci.* 16, 358. doi: 10.1038/nrn3880

Hill, S. T., Kuintzle, R., Teegarden, A., Merrill, E., III, Danaee, P., and Hendrix, D. A. (2018). A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Res.* 46, 8105–8113. doi: 10.1093/nar/gky567

Hohman, T. J., Dumitrescu, L., Cox, N. J., Jefferson, A. L., and Alzheimer's Disease Neuroimaging Initiative. (2017). Genetic resilience to amyloid related cognitive decline. *Brain Imaging Behav.* 11, 401–409. doi: 10.1007/s11682-016-9615-5

Hormozdiari, F., Van De Bunt, M., Segre, A. V., Li, X., Joo, J. W. J., Bilow, M., et al. (2016). Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* 99, 1245–1260. doi: 10.1016/j.ajhg.2016.10.003

Hu, Y., Hase, T., Li, H. P., Prabhakar, S., Kitano, H., Ng, S. K., et al. (2016). A machine learning approach for the identification of key markers involved in brain development from single-cell transcriptomic data. *BMC Genomics* 17, 1025. doi: 10.1186/s12864-016-3317-7

Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S. M., et al. (2019). *A statistical framework for cross-tissue transcriptome-wide association analysis*. Tech. rep., Nature Publishing Group. doi: 10.1038/s41588-019-0345-7

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 1254–1259. doi: 10.1109/34.730558

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016). Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361. doi: 10.1093/nar/gkw1092

Kingma, D. P., and Ba, J. (2014). Adam: a method for stochastic optimization.

Kirsebom, B.-E., Espenes, R., Waterloo, K., Hessen, E., Johnsen, S. H., Bråthen, G., et al. (2017). Screening for alzheimer's disease: cognitive impairment in self-referred and memory clinic-referred patients. *J. Alzheimers Dis.* 60, 1621–1631. doi: 10.3233/JAD-170385

Kuhn, J., Hardenacke, K., Lenartz, D., Gruendler, T., Ullsperger, M., Bartsch, C., et al. (2015). Deep brain stimulation of the nucleus basalis of Meynert in Alzheimer's dementia. *Mol. Psychiatry* 20, 353. doi: 10.1038/mp.2014.32

Kuznetsov, I., and Kuznetsov, A. (2018). How the formation of amyloid plaques and neurofibrillary tangles may be related: a mathematical modelling study. *Philos. Trans. R. Soc. Lond. A* 474, 20170777. doi: 10.1098/rspa.2017.0777

La Rosa, F., Clerici, M., Ratto, D., Occhinegro, A., Licito, A., Romeo, M., et al. (2018). The gut-brain axis in Alzheimer's disease and omega-3. a critical overview of clinical trials. *Nutrients* 10, 1267. doi: 10.3390/nu10091267

Lambert, J.-C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., et al. (2013). Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* 45, 1452. doi: 10.1038/ng.2802

Le Page, A., Dupuis, G., Frost, E. H., Larbi, A., Pawelec, G., Witkowski, J. M., et al. (2018). Role of the peripheral innate immune system in the development of Alzheimer's disease. *Exp. Gerontol.* 107, 59–66. doi: 10.1016/j.exger.2017.12.019

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521 (7553), 436. doi: 10.1038/nature14539

Li, Y., Huang, C., Ding, L., Li, Z., Pan, Y., and Gao, X. (2019). Deep learning in bioinformatics: introduction, application, and perspective in big data era. arXiv preprint arXiv:1903.00342. doi: 10.1101/563601

Li, Y., Wu, F.-X., and Ngom, A. (2016). A review on machine learning principles for 596 multi-view biological data integration. *Brief Bioinform.* 19, 325–340. doi: 10.1093/bib/bbw113

Li, Z., Chen, Z., Fan, G., Li, A., Yuan, J., and Xu, T. (2018). Cell-type-specific afferent innervation of the nucleus accumbens core and shell. *Front Neuroanat.* 12. doi: 10.3389/fnana.2018.00084

Liu, A. K. L., Chang, R. C.-C., Pearce, R. K., and Gentleman, S. M. (2015). Nucleus basalis of Meynert revisited: anatomy, history and differential involvement in Alzheimer's and Parkinson's disease. *Acta Neuropathol.* 129, 527–540. doi: 10.1007/s00401-015-1392-5

Liu, C.-C., Kanekiyo, T., Xu, H., and Bu, G. (2013). Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat. Rev. Neurol.* 9, 106. doi: 10.1038/nrneurol.2012.263

Loh, P.-R., Danecek, P., Palamara, P. F., Fuchsberger, C., Reshef, Y. A., Finucane, H. K., et al. (2016). Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* 48, 1443. doi: 10.1038/ng.3679

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

Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* 100, 473–487. doi: 10.1016/j.ajhg.2017.01.031

Marigorta, U. M., Denson, L. A., Hyams, J. S., Mondal, K., Prince, J., Walters, T. D., et al. (2017). Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat. Genet.* 49, 1517. doi: 10.1038/ng.3936

Marioni, R. E., Harris, S. E., Zhang, Q., McRae, A. F., Hagenaars, S. P., Hill, W. D., et al. (2018). GWAS on family history of Alzheimer's disease. *Transl. Psychiatry*. 8, 99. doi: 10.1038/s41398-018-0150-6

Mazure, C. M., and Swendsen, J. (2016). Sex differences in Alzheimer's disease and other dementias. *Lancet Neurol.* 15, 451–452. doi: 10.1016/S1474-4422(16)00067-3

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* 48, 1279. doi: 10.1038/ng.3643

McCaulley, M. E., and Grush, K. A. (2015). Alzheimer's disease: exploring the role of inflammation and implications for treatment. *Int. J. Alzheimers Dis.* 2015, 515248. doi: 10.1155/2015/515248

McDermaid, A., Monier, B., Zhao, J., Liu, B., and Ma, Q. (2018). Interpretation of differential gene expression results of RNA-seq data: review and integration. *Brief Bioinform*. doi: 10.1093/bib/bby067

McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine learning for detecting gene-gene interactions. *Appl. Bioinformatics* 5, 77–88. doi: 10.2165/00822942-200605020-00002

Michalopoulos, I., Pavlopoulos, G. A., Malatras, A., Karelas, A., Kostadima, M.-A., Schneider, R., et al. (2012). Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes. *BMC Res Notes* 5, 265. doi: 10.1186/1756-0500-5-265

Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. *Brief Bioinform.* 18, 851–869. doi: 10.1093/bib/bbw068

Nie, X., Sun, Y., Wan, S., Zhao, H., Liu, R., Li, X., et al. (2017). Subregional structural alterations in hippocampus and nucleus accumbens correlate with the clinical impairment in patients with Alzheimer's disease clinical spectrum: parallel combining volume and vertex-based approach. *Front. Neurol.* 8, 399. doi: 10.3389/fneur.2017.00399

Nobili, A., Latagliata, E. C., Viscomi, M. T., Cavallucci, V., Cutuli, D., Giacovazzo, G., et al. (2017). Dopamine neuronal loss contributes to memory and reward dysfunction in a model of Alzheimer's disease. *Nat. Commun* 8, 14727. doi: 10.1038/ncomms14727

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., and Moore, J. H. (2017). Data-driven advice for applying machine learning to bioinformatics problems. arXiv preprint arXiv:1708.05070. doi: 10.1142/9789813235533_0018

Piaceri, I., Nacmias, B., and Sorbi, S. (2013). Genetics of familial and sporadic Alzheimer's disease. *Front. Biosci. (Elite Ed)* 5, 167–177. doi: 10.2741/E605

Roberts, R. O., and Petersen, R. C. (2014). Predictors of early-onset cognitive impairment. *Brain* 137, 1280–1281. doi: 10.1093/brain/awu089

Sameith, K., Amini, S., Koerkamp, M. J. G., van Leenen, D., Brok, M., Brabers, N., et al. (2015). A high-resolution gene expression atlas of epistasis between gene-specific transcription factors exposes potential mechanisms for genetic interactions. *BMC Biol.* 13, 112. doi: 10.1186/s12915-015-0222-5

Sapkota, S., Huan, T., Tran, T., Zheng, J., Camicioli, R., Li, L., et al. (2018). Alzheimer's biomarkers from multiple modalities selectively discriminate clinical status: relative importance of salivary metabolomics panels, genetic, lifestyle, cognitive, functional health, and demographic risk markers. *Front. Aging Neurosci.* 10, 296. doi: 10.3389/fnagi.2018.00296

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv* 060012. doi: 10.1101/060012

Serpente, M., Bonsi, R., Scarpini, E., and Galimberti, D. (2014). Innate immune system and inflammation in Alzheimer's disease: from pathogenesis to treatment. *Neuroimmunomodulation* 21, 79–87. doi: 10.1159/000356529

Shao, Y., Shaw, M., Todd, K., Khrestian, M., D'Aleo, G., Barnard, P. J., et al. (2018). DNA methylation of TOMM40-APOE-APOC2 in Alzheimer's disease. *J. Hum. Genet.* 63, 459. doi: 10.1038/s10038-017-0393-8

Shen, Z., Bao, W., and Huang, D.-S. (2018). Recurrent neural network for predicting transcription factor binding sites. *Sci. Rep.* 8, 15270. doi: 10.1038/s41598-018-33321-1

Sochocka, M., Donskow-Łysoniewska, K., Diniz, B. S., Kurpas, D., Brzozowska, E., and Leszek, J. (2019). The gut microbiome alterations and inflammation-driven pathogenesis of Alzheimer's disease—a critical review. *Mol. Neurobiol.* 56, 1841–1851. doi: 10.1007/s12035-018-1188-4

Stocker, H., Möllers, T., Perna, L., and Brenner, H. (2018). The genetic risk of Alzheimer's disease beyond APOE ε4: systematic review of Alzheimer's genetic risk scores. *Transl. Psychiatry* 8, 166. doi: 10.1038/s41398-018-0221-8

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Tang, B., Pan, Z., Yin, K., and Khateeb, A. (2019). Recent advances of deep learning in bioinformatics and computational biology. *Front. Genet.* 10. doi: 10.3389/fgene.2019.00214

Tosto, G., Bird, T. D., Tsuang, D., Bennett, D. A., Boeve, B. F., Cruchaga, C., et al. (2017). Polygenic risk scores in familial Alzheimer disease. *Neurology* 88, 1180–1186. doi: 10.1212/WNL.0000000000003734

Van Cauwenberghe, C., Van Broeckhoven, C., and Sleegers, K. (2016). The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Med. Genet.* 18, 421. doi: 10.1038/gim.2015.117

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., et al. (2017). 10 years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* 101, 5–22. doi: 10.1016/j.ajhg.2017.06.005

Wang, D., Li, Y.-Y., Luo, J.-H., and Li, Y.-H. (2014). Age-related iron deposition in the basal ganglia of controls and Alzheimer disease patients quantified using susceptibility weighted imaging. *Arch. Gerontol. Geriatr.* 59, 439–449. doi: 10.1016/j.archger.2014.04.002

Wang, J., Gu, B. J., Masters, C. L., and Wang, Y.-J. (2017). A systemic view of Alzheimer disease—insights from amyloid-β metabolism beyond the brain. *Nat. Rev. Neurol.* 13, 612. doi: 10.1038/nrneurol.2017.111

Wang, L., Xi, Y., Sung, S., and Qiao, H. (2018). RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes. *BMC genomics* 19, 546 683. doi: 10.1186/s12864-018-4932-2

Wattenberg, M., Viégas, F., and Johnson, I. (2016). How to use t-sne effectively. *Distill.* doi: 10.23915/distill.00002

Way, G. P., and Greene, C. S. (2017a). Extracting a biologically relevant latent 685 space from cancer transcriptomes with variational autoencoders. *BioRxiv*, 174474. doi: 10.1101/174474

Way, G. P., and Greene, C. S. (2017b). "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders," in *Biocomputing 2018* (World Scientific). doi: 10.1142/9789813235533_0008

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., et al. (2013). The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimers Dement.* 9, e111–e194. doi: 10.1016/j.jalz.2013.05.1769

Wen, X., Pique-Regi, R., and Luca, F. (2017). Integrating molecular QTL data into genome-wide genetic association analysis: probabilistic assessment of enrichment and colocalization. *PLoS Genet.* 13, e1006646. doi: 10.1371/journal.pgen.1006646

Wetzel, S. J. (2017). Unsupervised learning of phase transitions: From principal component analysis to variational autoencoders. *Phys. Rev. E* 96, 022140. doi: 10.1103/PhysRevE.96.022140

Zhuang, Z., Shen, X., and Pan, W. (2019). A simple convolutional neural network for prediction of enhancer–promoter interactions with DNA sequence data. *Bioinformatics* bty 1050. doi: 10.1093/bioinformatics/bty1050

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

SYSTEMATIC REVIEW
published: 07 October 2019
doi: 10.3389/fgene.2019.00914

Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives

Zishuo Zeng[1,2]* *and Yana Bromberg*[2,3]*

[1] Institute for Quantitative Biomedicine, Rutgers University, Piscataway, NJ, United States, [2] Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ, United States, [3] Department of Genetics, Rutgers University, Human Genetics Institute, Piscataway, NJ, United States

Recent advances in high-throughput experimentation have put the exploration of genome sequences at the forefront of precision medicine. In an effort to interpret the sequencing data, numerous computational methods have been developed for evaluating the effects of genome variants. Interestingly, despite the fact that every person has as many synonymous (sSNV) as non-synonymous single nucleotide variants, our ability to predict their effects is limited. The paucity of experimentally tested sSNV effects appears to be the limiting factor in development of such methods. Here, we summarize the details and evaluate the performance of nine existing computational methods capable of predicting sSNV effects. We used a set of *observed* and artificially *generated* variants to approximate large scale performance expectations of these tools. We note that the distribution of these variants across amino acid and codon types suggests purifying evolutionary selection retaining *generated* variants out of the *observed* set; i.e., we expect the *generated* set to be enriched for deleterious variants. Closer inspection of the relationship between the *observed* variant frequencies and the associated prediction scores identifies predictor-specific scoring thresholds of reliable effect predictions. Notably, across all predictors, the variants scoring above these thresholds were significantly more often *generated* than *observed*. which confirms our assumption that the *generated* set is enriched for deleterious variants. Finally, we find that while the methods differ in their ability to identify severe sSNV effects, no predictor appears capable of definitively recognizing subtle effects of such variants on a large scale.

Keywords: synonymous variants, effect predictors, variant frequency, variant functional effect, machine learning

OPEN ACCESS

Edited by:
Dominik Heider,
University of Marburg, Germany

Reviewed by:
Jan Baumbach,
Technical University of Munich, Germany
Richard Röttger,
University of Southern Denmark, Denmark

****Correspondence:***
Zishuo Zeng
zzeng@bromberglab.org
Yana Bromberg
yana@bromberglab.org

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

Received: *28 June 2019*
Accepted: *29 August 2019*
Published: *07 October 2019*

Citation:
Zeng Z and Bromberg Y (2019) Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. Front. Genet. 10:914. doi: 10.3389/fgene.2019.00914

INTRODUCTION

The vast majority of human genomic variation is accounted for by Single Nucleotide Variants (SNVs) (Bromberg et al., 2013). The roughly 10,000 variants in the coding region of every human genome that have no effect on the resulting product protein sequence are termed synonymous SNVs (sSNVs) (Shen et al., 2013). sSNVs are a product of the degeneracy of genetic code, where amino acids may be encoded by more than one codon. The effects of sSNVs on molecular functionality of the corresponding genes/proteins are often assumed to be minimal. However, earlier studies have argued that sSNVs are as likely to be pathogenic as non-synonymous variants (Chen et al., 2010). sSNVs have been implicated in many diseases, including pulmonary sarcoidosis, attention deficit/hyperactivity disorder, and cancer (Sauna and Kimchi-Sarfaty, 2011; Supek et al., 2014). Synonymous variants can disrupt transcription (Stergachis et al., 2013), splicing (Pagani et al., 2005), co-translational folding (Pechmann and Frydman, 2013), mRNA stability (Presnyak et al., 2015) (**Figure 1**), and cause a plethora of other functionally-relevant changes. In addition, sSNVs can affect transcription and splicing regulatory factors within protein coding regions (Plotkin and Kudla, 2011), thus modulating gene expression (Shabalina et al., 2013; Boël et al., 2016). There is also evidence of evolutionary constraint on both synonymous and non-synonymous variants, which plays a role in shaping codon bias (organism or tissues-specific codon set preference) (Stergachis et al., 2013). An informative experimental approach to evaluating functional effects of sSNVs is saturation genome editing followed by protein function assays (Findlay et al., 2014; Findlay et al., 2018). Unfortunately, there are exceedingly few reports of these experiments in the literature. While there has been a concerted effort in the field to evaluate the effects of non-synonymous single nucleotide variants (nsSNVs) (Mahlich et al., 2017) for the purposes of precision medicine, as well as improving basic understanding of concepts in molecular biology, interpretation of sSNVs is severely lacking. However, considering the significant number of observed synonymous variants, their possible effects, and the dire lack of their systematic experimental interpretations, there is a compelling need for a reliable sSNV effect computational predictor.

In this paper, we review the existing sSNV-effect predictors and apply them to a dataset containing *observed* and artificially *generated* sSNVs. Since there are few experimentally-determined SNVs with deleterious effects, and those that exist have been used as training or testing sets of the predictors, the cornerstone of this study is validating our data set assumption that deleterious sSNVs are enriched in the artificially *generated* set of variants. To support this assumption, in addition to previously published work, e.g., Stergachis et al., 2013, we show that the distributions of observed sSNVs by amino acids and codons are highly

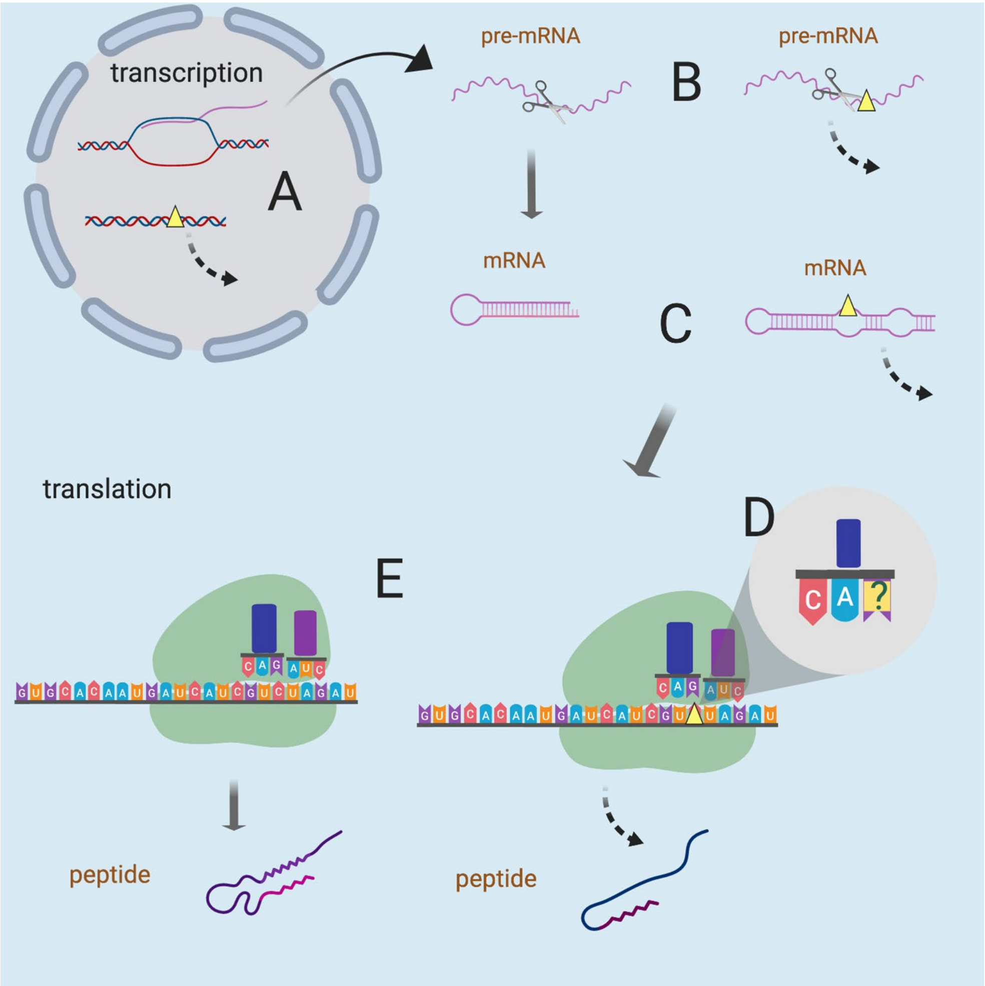


FIGURE 1 | Possible mechanisms of sSNVs impact on biological function. Yellow triangles represent sSNV sites and the dashed lines indicate aberrant processes. sSNVs may affect **(A)** transcription factor binding, **(B)** splicing of pre-mRNA, **(C)** mRNA secondary structure and stability, **(D)** wobble-based tRNA binding, and **(E)** cotranslational folding (and thus the protein structure). Figure was created with BioRender.com.

non-random. We also demonstrate that existing predictor high-scoring variants are enriched among the artificially *generated* sSNVs, additionally validating of our assumption. We finally note that these predictors appear unable to definitely identify subtle effect sSNVs.

METHODOLOGY OF THE PREDICTORS

SNV Predictors Vary by Targeted Variant Type, Training Data, and Descriptive Features

We identified from the literature four sSNV-specific effect predictors: SilVA (Silent Variant Analyzer) (Buske et al., 2013), regSNPs-splicing (Zhang et al., 2017), DDIG-SN (Detecting Disease-causing Genetic SynoNymous variants) (Livingstone et al., 2017), and IDSV (Identification of Deleterious Synonymous Variants) (Shi et al., 2019). Additionally, we considered TraP (Transcript-inferred Pathogenicity) (Gelfman et al., 2017), which addresses both synonymous and intronic variants. Specifically, 1) SilVA was trained on 33 pathogenic and 785 neutral variants from 1000 Genomes Project (1000G) (Birney and Soranzo, 2015), using conservation scores, splicing, DNA, and RNA properties, 2) DDIG-SN and IDSV used positive data from the Human Gene Mutation Database (HGMD) (Cooper et al., 1998; Stenson et al., 2003; Stenson et al., 2009; Stenson et al., 2017) and negative data from 1000G (DDIG-SN) and VariSNP (IDSV) (Schaafsma and Vihinen, 2015) as negative data for training, described using features of translational efficiency and protein properties in addition to those used by SilVA, 3) regSNPs-splicing also used HGMD and 1000G data, but it considers sSNVs only in the context of mRNA splicing and protein function, while 4) TraP was trained on positive data combining SilVA's data with Online Mendelian Inheritance in Man (OMIM) (Hamosh, 2004) variants and negative data from control trios *de novo* variants. TraP uses transcript-affecting features, specific to intronic and synonymous variants.

As opposed to the sSNV-specific tools, more generic predictors, including CADD (Kircher et al., 2014), DANN (Quang et al., 2014), FATHMM-MKL (Shihab et al., 2015), and MutationTaster2 (Schwarz et al., 2014), evaluate synonymous, non-synonymous, regulatory and other kinds of variants. CADD was developed by training a support vector machine (SVM) to differentiate observed *vs.* simulated variants of all variant categories (Kircher et al., 2014). DANN attempts to capture nonlinear signals in (CADD-generated) variant data using a deep neural network (Quang et al., 2014). FATHMM-MKL is a Hidden Markov Model-based method integrating ENCODE (Consortium, 2012) functional annotations of SNVs to evaluate non-coding and synonymous variants (Shihab et al., 2015). MutationTaster2 (Schwarz et al., 2014) uses a naïve Bayes model trained on disease variants vs. variants from 1000G variants to evaluate all SNVs. Notably, these general-purpose predictors are heavily conservation-driven and may lack features to describe the subtle changes induced by sSNVs.

All predictors described here are machine learning-based [using random forests (RFs), SVMs, or deep neural network] and trained to predict pathogenicity, using different data and feature sets (**Table 1**). Supervised machine learning, often used for predicting variant effects, requires selecting a proper training/evaluation set, a number of relevant variant-, gene-, or disease-features, and an appropriate model for identifying feature patterns representative of variant effect/disease-association (Rost et al., 2016).

Available Variant Sets Are Limited in Size and Reliability

Association between genomic variants and diseases can be identified by carefully designed statistical tests, e.g., *via* Genome Wide Association Studies (GWAS) (Visscher et al., 2012). However, unequivocally identifying variants that cause disease are significantly more difficult; this is a particularly hard problem for sSNVs, which carry no corresponding protein sequence changes. Clinical or experimental validation of the causative relationships between genomic variation and disease is either infeasible altogether (as for polygenic disorders) or exceedingly difficult on a large scale due to the necessary resource and time investments. Instead, computational annotation of genomic variant pathogenicity (or simply functional effects) is a cost- and time-efficient substitute, providing a starting point for further experimental validation and discovery.

Most predictors described here (regSNPs-splicing, DDIG-SN, FATHMM-MKL, and MutationTaster2) collect variants identified as causative (positive) from HGMD. The latest version of HGMD (March 2017) comprises over 203,000 variants in over 8,000 genes, manually curated from scientific literature (Stenson et al., 2017). Despite its apparent utility, studies have questioned the reliability of HGMD data. George et al. (2007), for example, point out flaws like inconsistent mutation nomenclature and incomplete incorporation of all applicable data. Yue and Moult (2006) note that some mutations in HGMD are named causes of monogenic disease but are not fully penetrant, while Bell et al. (2011) question disease annotation of recessive variants. In a study of 1,000 exomes, Dorschner et al. (2013) note that only 16 of 585 of HGMD disease-causing variants were actually pathogenic, while in a subsequent study with 6,503 individuals, none of the identified 615 HGMD disease-causing variants were pathogenic (Amendola et al., 2015). Other studies (Xue et al., 2012; Cassa et al., 2013) have shown that many disease-causing variants in HGMD are present in the relatively healthy 1000G individuals (Birney and Soranzo, 2015).

Other sources of positive training/testing data, including OMIM (used by TraP) and ClinVar (used by TraP, regSNPs-splicing, IDSV, CADD, MutationTaster2, and FATHMM-MKL) (Landrum et al., 2014), appear no more reliable. Notably, there is considerable inconsistency between the HGMD and OMIM (George et al., 2007). ClinVar's entries from different sources often conflict between themselves (Landrum and Kattman, 2018), as the reliability of ClinVar's data curation and workflow of medical interpretation has not been proven (Bauer et al., 2018). Substantial discordance between ClinVar and laboratory test results has also been reported (Gradishar et al., 2017).

TABLE 1 | Summary of sSNV-specific predictors.

Ref/Tool name	Training data	Model	Features	Performance
(Buske et al., 2013) SilVA (2013)	33 deleterious from literature, 785 neutral from one 1000 Genomes Project individual	Random forest with 1,001 trees and default number of features	26 in total • conservation • RNA properties • DNA properties • Splicing	**Dataset**: 8 DM from literature and 752 NM from literature and 1000G. **Result**: DM's scores ranked higher than NM's
(Gelfman et al., 2017) TraP (2017)	75 DM from literature and OMIM and 402 *de novo* NM from control trios	Random forest with 1,000 trees, each with	20 in total • Conservation • DNA properties • Splicing	**Dataset**: 66 DM and 4,418 NM from ClinVar. **Result**: AUC = 0.88
(Zhang et al., 2017) regSNPs-splicing (2017)	~655 DM from HGMD and ~655 NM from 1000G	Random forest with 51 trees and 35 features at each node	455 in total • Conservation • RNA properties • protein properties • splicing	**Dataset**: ~325 DM from HGMD and 230 DM from ClinVar, ~325 NM from 1000G and 4,535 NM from ClinVar **Result**: For HGMD vs. 1000G data, AUC = 0.91 for variants in Splice Sites and AUC = 0.82 for all others For ClinVar data, AUC = 0.85 for variants in splice sites and AUC = 0.70 for the all others
(Livingstone et al., 2017) DDIG-SN (2017)	592 DM from HGMD and 10,925 putatively benign from 1000G	Support Vector Machine with radial function kernel	54 in total (including all of the 26 features used in SilVA) • conservation • DNA properties • RNA properties • Protein properties • Splicing	**Dataset**: 279 DM from HGMD and 4,945 NM from 1000G **Result**: AUC = 0.85
(Shi et al., 2019) IDSV (2019)	300 DM from dbDSM and 300 NM from VariSNP	Random forest with 500 trees and 3 features at each split	10 in total • Conservation • DNA properties • Splicing • Translational efficiency	**Dataset**: 153 DM and 5,178 NM from ClinVar **Result**: AUC = 0.87

DM, disease/deleterious mutations; NM, neutral mutations; HGMD, human gene mutation database; 1000G, 1000 genome project; OMIM, online mendelian inheritance in man; AUC, area under the ROC curve (axes in Eqn. 1).

Mutation databases vary drastically (George et al., 2007), not in the least because of experimental interpretation differences; e.g., roughly 17% of the variant effects reported by different laboratories carry contradictory clinical significance (Rehm et al., 2015). Labels of pathogenicity are not fixed, switching from disease to benign and back as evidence accumulates (Shah et al., 2018). As these binary labels also do not provide a quantitative measure of risk (Shah et al., 2018) or penetrance, the term "disease-causing" should be used with caution. One key problem in the field, and a reason for many of the above data limitations, is the absence of a gold standard for identifying disease-causing variants (Dorschner et al., 2013). Moreover, even the "silver-standard" available annotations are far and few between. In fact, while there are many known pathogenic nsSNVs, there are currently much fewer known pathogenic sSNVs available: dbDSM (Wen et al., 2016) (including those from ClinVar, PubMed, NHGRI GWAS catalog (Welter et al., 2013), etc.) contains 1,289 pathogenic sSNVs, and HGMD contains roughly 900 pathogenic sSNVs (Livingstone et al., 2017). Arguably, this number is too small to build a generalizable model for evaluating tens of millions of the possible synonymous variants in human genome. Note that an additional problem is the absence of a true negative set of variants, i.e., those that have been verified to have no effect on protein function or no relationship to some disease (Bromberg et al., 2013).

Use of Allele Frequency to Approximate Variant Effect

SilVA was trained on 33 experimentally defined deleterious and 785 assumed neutral (observed in 1000G) variants. While the former set was very stringently selected, this small number of samples could hardly produce a generalizable model. Other predictors use less well curated data from available databases, but as such run into a problem of reliability. To supplement the lack of experimentally annotated variation, variant population frequency had been suggested as a sign of effect/pathogenicity; i.e., it is generally assumed that disease/effect variants are of low allele frequency (Gibson, 2012), although the precise threshold for "low" is unclear. Predictors (CADD, DANN, FATHMM-MKL, SilVA, regSNP-splicing) often filter out effect variants of higher frequency and/or neutral variants of lower frequency. CADD and DANN training data, for example, contains simulated human variants, appearing after human-chimpanzee divergence, labelled as the effect group (depleted by natural selection) and observed fixed or nearly fixed derived alleles as neutral (Kircher et al., 2014; Quang et al., 2014). Note although simulated variants are likely enriched in deleterious variants, and CADD scores have been shown useful in prioritizing variants in clinical settings (Amendola et al., 2015; Nakagomi et al., 2018; Van Der Velde et al., 2015), it is difficult to directly link the CADD predictions to pathogenicity (Kircher et al., 2014).

Allele frequency, however, is not necessarily correlated with variant effect, particularly when effect being considered is "function change" not "disease." In an earlier study, we found that common [minor allele frequency (MAF) > 5%] non-synonymous variants were more often predicted to have a functional effect than rare (MAF < 1%) ones (Mahlich et al., 2017). Here a high-frequency allele may be beneficial/advantageous and on the way to becoming common, or slightly deleterious and on its way out (Bromberg et al., 2013). Moreover, trivially, allele frequency estimated from the sequenced genomes may be subject to change as the number of samples increases. Thus, 1) low allele frequency is not equivalent to having an effect and 2) although high frequency alleles are unlikely to be disease causing, they may have some impact. Additionally, and perhaps most fundamentally, note that the currently observed SNVs are unlikely a complete set of naturally occurring variants, i.e., many SNVs may be yet unseen.

Features Used Vary From Predictor to Predictor

A variety of features have been considered by predictors as described below. Note that the number of features used in existing predictors ranges from 26 (SilVA) to 1,281 (FATHMM-MKL).

Conservation

Evolutionary conservation, derived from multiple sequence alignments (MSAs) of homologous sequences (Niroula and Vihinen, 2016), is perhaps the most extensively used feature of variant-effect predictors. Commonly used DNA conservation scoring algorithms include GERP (Cooper et al., 2005), phastCons (Siepel et al., 2005), and PhyloP (Pollard et al., 2009) scores. GERP (Genomic Evolutionary Rate Profiling) is a statistical method identifying genomic constrained elements from MSAs. GERP uses a statistical model estimating species divergence times (Hasegawa et al., 1985) and a structural expectation maximization algorithm for phylogenetic inference (Friedman et al., 2002); the later GERP++ is a faster version of the original (Davydov et al., 2010). phastCons fits MSAs to phylogenetic hidden Markov models to identify conserved elements (Siepel et al., 2005). The major difference between phastCons and GERP is that the former models the size and distribution of conserved elements within an MSA, while the latter first individually assesses the conservation at a locus and then searches for clusters of highly conserved loci (Chen et al., 2010). PhyloP combines statistical tests and GERP to detect conservation and acceleration in nucleotide substitution rates (Pollard et al., 2009). All variant effect predictors use at least one of these conservation scoring techniques (**Tables 1**, **2**). DDIG-SN also additionally uses protein conservation as conserved protein positions are often structurally important (Ng, 2003), suggesting possible misfolding due to decreased rate of translation at the relevant codon. Similarly, sSNVs may lead to mistranslation (Kramer and Farabaugh, 2006; Kramer et al., 2010; Komar, 2016) resulting in amino acid substitutions—a particularly problematic occurrence at conserved protein positions.

Conservation is a very important signature of variant effect. For example, for ClinVar's missense dataset the solely-conservation-based component of CADD, GerpS (a derivative of GERP++), as well as PhastCons and PhyloP, attained ROC AUCs (area under the receiver operating characteristic curve) of over 0.82, while CADD's ROC AUC was only slightly higher (0.93) (Kircher et al.,

TABLE 2 | Summary of generalized SNV predictors.

Ref/Tool name	Training data	Model	Features	Performance
(Kircher et al., 2014) CADD (2014)	13,141,299 SNVs, 627,071 insertions, and 926,968 deletions from simulated and observed variant sets	SVM with linear kernel	63 in total • Conservation • Variant consequence • DNA features • Other	**No testing of synonymous variants**
(Quang et al., 2014) DANN (2014)	13,302,220 observed variants; 13,302,220 simulated variants selected from CADD data	Neural network with 3 1,000-node hidden layers	63 features from CADD	**All types of variants, amount of sSNVs not stated** **Dataset**: 162,777 observed and 162,777 simulated variants (including synonymous variants). **Result**: Overall accuracy = 0.66
(Shihab et al., 2015) FATHMM-MKL (2015)	1,073 coding DM from HGMD and 1,073 coding NM from 1000G for 10-feature-group model; 3,000 coding DM from HGMD and 3,000 coding NM from 1000G for 4-feature-group model	Multiple kernel learning	1,281 in total • Conservation • DNA properties • Other	**Coding variants, amount of sSNVs not stated** **Dataset**: 5-fold cross-validation from training data **Result**: AUC = 0.93 and 0.91for 10-feature-group model and 4-feature-group model, respectively
(Schwarz et al., 2014) MutationTaster2 (2014)	122,238 DM from ClinVar and HGMD; 6,807,269 NM from 1000G	Bayesian classifier	~ 7 (not explicitly stated) in total • Conservation • DNA properties • Splicing	**No testing of synonymous variants**

DM, disease/deleterious mutations; NM, neutral mutations; HGMD, human gene mutation database; 1000G, 1000 genome project; AUC, area under the receiver operating characteristic curve.

2014). In FATHMM-MKL's cross-validation on coding variants, its ROC AUCs was = 0.93 while the ROC AUCs for conservation scores alone was = 0.91 (Shihab et al., 2015). Similar results are observed for DDIG-SN (DDIG-SN's ROC AUCs = 0.85, PhyloP's ROC AUCs = 0.76) (Livingstone et al., 2017) and TraP (TraP's ROC AUCs = 0.88, GERP++'s ROC AUCs = 0.87) (Gelfman et al., 2017) datasets. These results suggest that over billions of years of evolution, nature's laboratory has tested and discarded most of the detrimental variants. However, it is important to note that functional tuneability, i.e., development of new or environment-specific versions of functions is an ongoing process, which requires the presence of variants in positions of all levels of conservation, in any given snapshot of a population (Miller et al., 2017; Miller et al., 2019).

DNA Properties

The DNA properties describing the biological effects of sSNVs include but are not limited to localization to transcription factor (TF) binding sites, overall GC content of genes and genomes, and CpG island locations (cytosine followed by guanine in 5' to 3' direction). In more detail: many studies have shown that coding exons can serve as regulatory elements for transcription (Lang et al., 2005; Khan et al., 2012); i.e., roughly 15% of the human genome codons both code for amino acids and specify TF recognition (Stergachis et al., 2013). Thus, synonymous variants in TF-relevant codons can affect TF binding and alter gene transcription rates. Exonic and the flanking intronic region GC architectures can affect DNA methylation and exon recognition (Gelfman et al., 2013). Additionally, CpG sites often host DNA methylation (Bernstein et al., 2007), playing an important role in gene transcription (Gelfman et al., 2013). As mutation rates at CpG dinucleotides are an order higher than at other sites (Nachman and Crowell, 2000), sSNVs can thus alter methylation patterns by disrupting site-specific GC architectures.

All predictors covered in this manuscript, except regSNPs-splicing, incorporate one or more of these DNA properties (**Tables 1**, **2**).

RNA Properties

Codon bias. The preference (frequency of use) of particular codons by specific organisms or tissues is termed codon bias. Codon bias correlates with and informs gene expression levels (Coghlan and Wolfe, 2000; Carbone et al., 2003; Dos Reis et al., 2003; Boël et al., 2016; Komar, 2016), translation rate (Sørensen et al., 1989), as well as protein structure (Zhou et al., 2009) and cotranslational folding (Pechmann and Frydman, 2013; Buhr et al., 2016). There are many different metrics describing codon bias including codon adaptation index (Sharp and Li, 1987), synonymous codon usage order (Angellotti et al., 2007), relative synonymous codon usage (Sharp and Li, 1987), etc. Surprisingly, only SilVA and DDIG-SN have considered codon bias as a factor in their models (**Table 1**).

A related factor governing translation rate is the supply of tRNA during translation. Note that tRNA concentrations are different across organisms and that some organisms lack certain tRNA altogether, supplementing the necessary functionality *via* third position wobble (Novoa et al., 2012). It is hypothesized that codon composition in coding regions coevolved with tRNA abundances to reach the desired translation rates (Plotkin and Kudla, 2011). tRNA adaptation index (tAI) (Reis et al., 2004), used only by IDSV (**Table 1**), is a measure aimed to describe codon bias from the perspective of tRNA supply and demand.

A potentially important feature also missing from all predictors is codon autocorrelation. In codon autocorrelated sequences, same codons follow each other in sequence, i.e., sequence AAABB is more autocorrelated (less anticorrelated) than sequence ABABA, where A and B are two codons of the same amino acid (Cannarozzi et al., 2010). Autocorrelated yeast transcripts are translated faster than anticorrelated ones (Cannarozzi et al., 2010) and many prokaryotes modulate translation through codon correlation (Guo et al., 2012). Thus, using codon correlation may help characterizing sSNV effect.

mRNA structure, stability, and abundance. sSNVs can alter mRNA secondary structure, thus impacting translational efficiency and mRNA decay rate (Hunt et al., 2014), which, in turn, impacts protein production (Komar, 2016) and abundance (Maier et al., 2009). mRNA sequences are more stable than random collections of nucleotides (Seffens, 1999), suggesting that mRNA stability is evolutionarily selected to accommodate sufficient levels of translation before decay. The secondary structure of mRNAs harbors conserved elements (Meyer, 2005) and is tightly interwoven with GC content and codon usage. In fact, an earlier study found that 26% of the expressed genes display differential mRNA stability across individuals (Duan et al., 2013). In these genes, higher GC3 (G or C at the third position of the codon) percentage correlated with higher mRNA stability. This finding is in line with the fact that among the different SNVs, G and C alleles generally result in higher mRNA stability than A and T alleles (Duan et al., 2013). Furthermore, stability is enhanced in mRNA sequences enriched in optimal codons corresponding to tRNAs of higher concentrations (Presnyak et al., 2015).

A number of *in silico* tools have been developed to predict the mRNA structure and stability, including mFold (UNAFold) (Zuker, 2003; Markham and Zuker, 2008), remuRNA (Salari et al., 2012), KineFold (Xayaphoummine et al., 2005), and RNAfold (ViennaRNA package) (Hofacker, 2003). Note, however, that RNA molecules are very thermodynamically flexible and can take on many possible structures. Thus, the predicted RNA structure and its stability depend on the pre-set prediction strategy, which can be aimed to find the minimum free energy structure, the structure closest to other possible structures, or to maximize expected prediction accuracy, which is difficult for RNAs longer than 500 nucleotides (Lorenz et al., 2016). Consequentially, the prediction of RNA structure and stability is inherently uncertain. Among all the sSNV predictors, only SilVA and DDIG-SN use predictive tools to compute the variant-induced changes of energy and structures in pre-mRNA and mature mRNA sequences (**Table 1**).

Note that sSNVs, as well as other variant types (Shah and Gilchrist, 2010), are particularly relevant to functionality of highly expressed genes. Thus, the Genotype-Tissue Expression

(GTEx) project's database containing large-scale human tissue-specific gene expression data (Lonsdale et al., 2013) can be used to establish genes that are likely to manifest sSNV effect. However, none of the predictors described here use expression information to inform their effect predictions.

Splicing Properties

mRNA splicing is a major predictive feature in some predictors, especially regSNPs-splicing and IDSV. It is estimated that up to 15% of disease SNVs cause aberrant splicing (Krawczak et al., 1992). sSNVs can impact exonic splicing enhancers (ESEs) and silencers (ESSs), i.e., short DNA sequence motifs that promote or suppress splicing of pre-mRNA by binding to SR proteins (proteins with long repeats of serine and arginine) (Wang and Burge, 2008). Moreover, sSNVs can change the affinity of pre-mRNA to spliceosomes, leading to false recognition of exon-intron boundaries and producing abnormal mRNAs and dysfunctional proteins (Bali and Bebok, 2015). Taken together, the sSNVs' potential of disrupting splicing is the likely reason for slower evolution at within-ESE sites (Parmley, 2005).

Predictors describe the potential impact of sSNVs on splicing by relying on the identified putative ESE and ESS motifs. Identification of these motifs and the corresponding splicing regulatory proteins has been an ongoing experimental and computational effort (Wang and Burge, 2008; Shepard and Hertel, 2009); identified motifs and regulatory proteins are available *via* public repositories (Desmet et al., 2009; Giulietti et al., 2013; Xing et al., 2016). Tools such as SPANR (Splicing-based Analysis of Variants) (Xiong et al., 2015), have also been developed to predict the splicing effects of SNVs. Splicing is considered by all sSNV-specific predictors, although represented *via* different values.

Protein Properties

One often overlooked aspect in evaluating sSNV effect is the protein structure. Rare codon variants of frequent synonymous codons may slow down the translation rate due to low concentration of tRNAs, slow or stop the elongation of the peptide chain (Zhang et al., 2009), and influence co-translational folding (Kimchi-Sarfaty et al., 2007; Pechmann and Frydman, 2013). Cotranslational folding is closely related to the formation of protein secondary and tertiary structures (Holtkamp et al., 2015); alpha-helix formation can occur in the ribosomal tunnel (Komar, 2009), while tertiary structure formation may take place before the protein completely exits the ribosome (Zhang and Ignatova, 2011). Translationally fast codons are enriched for alpha helices, while beta strands and coil regions prefer translationally slow codons (Thanaraj and Argos, 1996). Optimal codons are enriched in buried and structurally important sites but are negatively correlated with solvent accessible sites (Zhou et al., 2009). Pathogenic sSNVs are generally enriched within the buried sites, intrinsic disorder regions, and alpha-helices, as well as in exons overlapping with known or predicted protein family domains (Zhang et al., 2017). These findings suggest that protein structure should be considered when modelling the effects of sSNVs. However, only regSNPs-splicing and DDIG-SN predictors incorporate protein structural information (**Table 1**).

EVALUATION OF THE PREDICTORS

Collecting the Evaluation Data Set

sSNV effect predictor evaluation is hampered by three major problems: 1) there is no clear definition of neutral and effect variants and 2) available neutral/effect experimental evaluations are limited, and 3) most have been used in predictor development. Here, we created our own data set of variants for evaluation purposes as follows: we collected the *observed* sSNVs [all non-singleton sSNVs that have been observed in either 1000G, ExAC (Lek et al., 2016), or gnomAD (Karczewski et al., 2019)] and the *generated* sSNVs (all possible sSNVs in human genes, excluding *observed* and singleton sSNVs); we thus extracted 1,362,607 *observed* and 24,008,961 *generated* sSNVs. For evaluation purposes, we randomly selected 1,362,607 *generated* variants from our set to create a balanced *observed/generated* variant *Test set* (details in **Supplementary Material).**

There are multiple equally valid reasons for why nearly 95% of all possible sSNVs are not *observed*; the most obvious ones are technical, i.e., insufficient data or sequencing technology bias, and evolutionary, i.e., purifying selection, genetic drift, and genetic hitch-hiking (Smith and Haigh, 1974). As per the latter, we assume that drastically deleterious variants, which would be eliminated on a population scale due to purifying selection, are significantly more frequent in the set of *generated* sSNVs than in *observed* ones. However, the former suggests that we may have simply not (yet) sequenced many of the un-observed *(generated)* variants, which are actually equivalent in potential effect to *observed* ones. Notably, since a large proportion of discovered sSNVs are singletons (Lek et al., 2016), an equivalent proportion of similarly neutral or mild-effect variants can likely be found on the other side of the "sequencing barrier," i.e., they have yet to be sequenced. Moreover, different categories of variants vary in the likelihood of being observed. For example, according to the ExAC project, the discovery of CpG transitions (C- > T, where C is followed by G) is likely close to saturation, while additional transversion and non-CpG transitions are yet to be identified (Lek et al., 2016).

We observe that 1) most of the large effect variants are likely in the *generated* set and either 2a) they make up much of that set, i.e., the *generated* set contains mostly effect variants, or 2b) there are relatively few of them, i.e., the distribution of effect and neutral variants is roughly equivalent across the *generated* and *observed* variants. Note that if (2a) is true, we expect that a precise and sensitive sSNV effect predictor should be able to differentiate the *observed* sSNVs from the *generated* ones, while (2b) would mean that the same predictor would produce similar effect distributions.

Note that our *Test set* data are collected in a somewhat similar, but ultimately very different, way as CADD's (and DANN's) training data. CADD's observed variants are the fixed or nearly fixed alleles at sites where human genes are different from the inferred human-chimpanzee ancestor and thus may encompass our excluded *observed* singletons. CADD's simulated variants follow an estimated *de novo* mutation rate since human-chimpanzee divergence, and thus are a subset of all our variants, including *generated, observed*, and singletons. Importantly, even with down-sampling of *generated* variants to create a balanced set, our *Test set* is much larger (~2.8 million) and more broadly defined than CADD's strictly curated training set (~100,000).

We calculated the enrichment of *observed* sSNVs relative to *generated* sSNVs separately by amino acid (**Figure 2A**) and codon (**Figure 2B**) type. We observe that the distribution of naturally occurring sSNVs is non-random across amino acids and codons. Thus, over a fifth of all tyrosine (Y) and histidine (H) codons in our genome is affected by sSNVs, as compared to roughly 8% of lysine (K) codons. Curiously, the most mutated codons are threonine ACG, serine TCG, and proline CCG (> 43% of each is affected by an sSNV) and alanine GCG (37%). Thus, the CG end-of-codon nucleotide pair seems to indicate least stable codons. On the other hand, the isoleucine ATA codon is almost never mutated (~1%), suggesting that it is preferentially maintained as error free. Moreover, the enrichments of observed sSNVs by amino acids (or codon) are not proportional to the abundance amino acids (or codon) in human transcriptome. The amino acids (e.g., Y, H, N, D) and codons (e.g., ACG, TCG, CCG, GCG, TAC, CAC) with high enrichment of *observed* sSNVs are those of low abundances. This decidedly non-random distribution of variants across codons and amino acids strongly suggests that our *generated* and *observed* variants are likely indeed different from the evolutionary, and thus likely effect, perspective.

Predictors Do Not Distinguish *Observed* and *Generated* sSNVs

To the best of our knowledge, our collection of tools (CADD, DANN, MutationTaster2, FATHMM-MKL, SilVA, TraP, DDIG-SN, regSNP-splicing, and IDSV) make up a complete set of publicly available methods for sSNV analysis. We first evaluated (**Figure S2**) the ability of all predictors (except regSNP-splicing, which was not functional at the time of writing) to differentiate 50,000 *observed* and 50,000 *generated* sSNVs (**Supplementary Materials**). We did not include IDSV for more further analysis because its performance was similar to that of other predictors and it was not available for running it locally or online for the entire set of our variants. Unfortunately, we also had to exclude MutationTaster2, which experienced server problems when running large batches of data.

We used CADD, DANN, FATHMM-MKL, SilVA, TraP, and DDIG-SN to make predictions for our complete variant *Test set*. We calculated the fraction of consensus binary predictions (**Figure 3A**) (FCBP; i.e., the number of predictions agreed upon) for all pairs of predictors and the correlation between scores (**Figure 3B**). As per CADD creators (https://cadd.gs.washington.edu/info), it is hard to threshold its raw scores, while the recommended neutral/deleterious cutoff for phred-scaled scores is 15. For the rest of the predictors, we used 0.5 as the binary threshold (> 0.5 is deleterious). We observed (**Figure 3A**) that the CADD and other sSNV-specific predictors agree with each other because their scores are mostly low (**Figures 3F–H**). Scores from general-purpose predictors do not have high correlation with

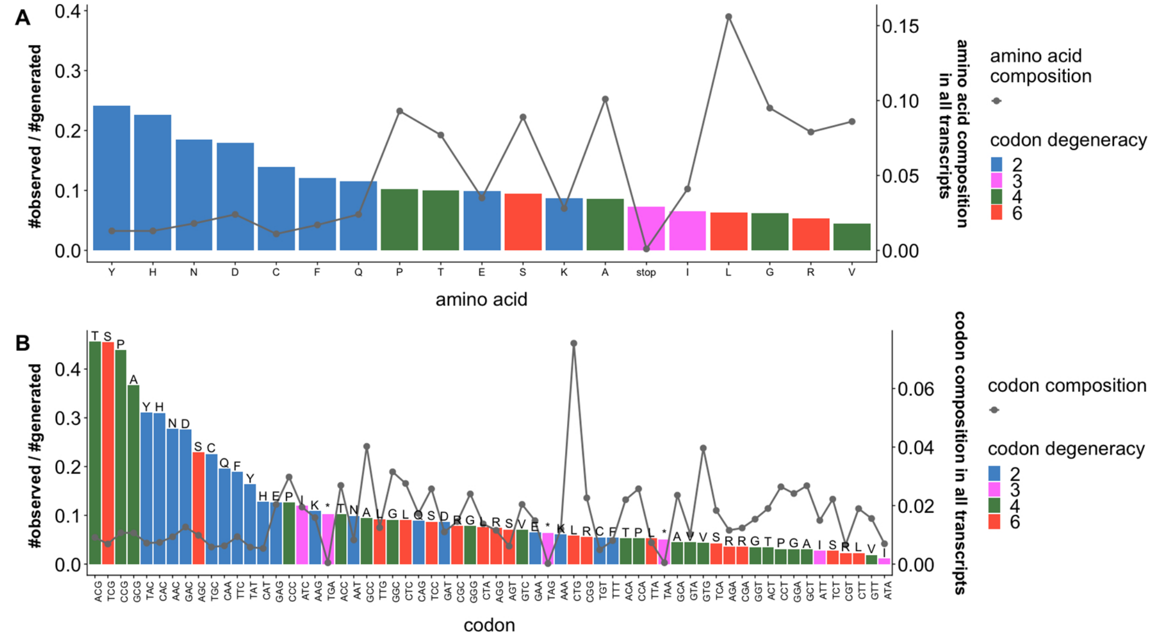


FIGURE 2 | Ratios of *observed* and *generated* sSNVs vary across codons and amino acids. Ratios of *observed* to *generated* sSNVs (barplot, left axis) affecting specific **(A)** amino acids and **(B)** codons in the human transcriptome differ. Lines (right axis) in plots indicate the fractions of **(A)** amino acids and **(B)** codons ("*" is a stop codons). Trivially, 2-codon amino acids are generally enriched for *observed* sSNVs, while higher degeneracy codons are depleted. However, there is a significant difference between the most and least frequent 2-codon amino acid sSNVs. Codons with an NCG pattern (N = any nucleotide) are most often affected by sSNVs. On the other hand, codons with a CGN pattern (also CpG) are relatively rarely affected. Note that amino acid degeneracy is correlated with % composition, although a single codon is often responsible for coding most of each of these amino acids (e.g. Leucine CTG and Valine CTG).

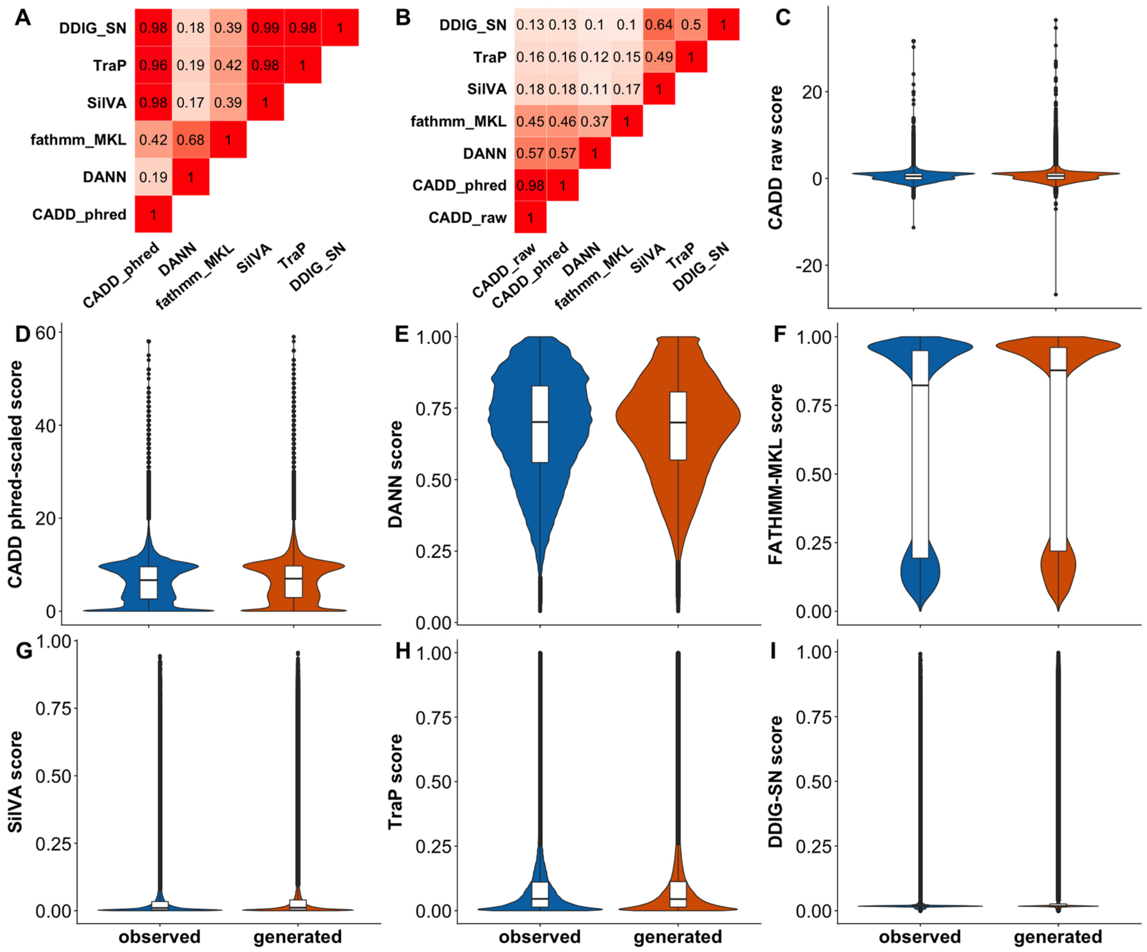


FIGURE 3 | Predictor scores correlate somewhat, but do not differentiate *observed* vs. *generated* sSNVs. Panel **(A)** shows the amount of agreement (i.e., FCBP) for any pair of predictors. High FCBP values indicate that two predictors agree in assigning binary (neutral/deleterious) predictions to variants. Panel **(B)** shows the Pearson correlations among the prediction scores. **(C–I)** Violin/box plots of prediction score distributions across predictors: CADD raw, CADD phred-scaled, DANN, FATHMM-MKL, SilVA, TraP, and DDIG-SN, respectively.

sSNV-specific predictors. Meanwhile, DANN and FATHMM-MKL did not agree with others or between themselves. This lack of agreement across the *Test set* indicates that, in the best case, predictors are orthogonal, correctly identifying a different subset of variants each or, in the worst case, they are mostly unable to recognize effect. Curiously, for each predictor, the distributions of sSNV scores of *observed* and *generated* variants were very similar (**Figure 3**), i.e., predictors disagreed between themselves and with our dataset labels. Note that since the data is large, statistical tests to establish their difference could easily achieve significance and may not be meaningful (Kim and Bang, 2016). Instead, we directly evaluated predictor ability (**Table 3**) to differentiate the two types of variants using ROC AUCs. ROC curve is a plot of true positive rate (TPR) against false positive rate (FPR), which are computed with true positive (TP), false negative (FN), and false positive (FP) (Eqn. 1). No predictor was able to accurately differentiate *generated* and *observed* variants well. To evaluate the variation of different predictors introduced by the sampling of the *generated* set, we also subsampled the *observed* and *generated* sets for 20 times (each with 100,000 samples) and calculated the resulting standard errors of ROC AUCs (**Table 3**).

$$TPR = \frac{TP}{TP+FN},\ FPR = \frac{FP}{FP+TN} \tag{1}$$

TABLE 3 | AUCs of the predictors on sSNVs and nsSNVs.

	***Observed* vs. *generated* sSNVs**		***Observed* vs. *generated* nsSNVs**
	AUC on *Test set*	**Average of AUCs ±SD ***	
CADD raw score	0.518	0.517±0.0012	0.564
CADD phred-scaled score	0.518	0.518±0.0013	0.564
DANN	0.506	0.506±0.0023	0.491
FATHMM-MKL	0.540	0.540±0.0013	0.555
SilVA	0.527	0.527±0.0009	
TraP	0.495	0.496±0.0038	
DDIG-SN	0.535	0.535±0.0012	

**Test set was sampled 20 times (each with 100,000 observed and 100,000 generated variants) to produce averages and standard deviations (SD) of AUCs for sSNVs.*

Predictor Performance Is Only Slightly Better for nsSNVs Than for sSNVs

As mentioned previously, the unexpected inability of predictors (**Figure 3**) to differentiate *observed* and *generated* variants may indicate either the inappropriateness of the data set for the evaluation task or limited predictor abilities. The latter may be related to the specific variant type; i.e., general-purpose predictors, such as CADD and FATHMM-MKL, are good at analyzing non-synonymous variants (Kircher et al., 2014; Shihab et al., 2015), but they may be less sensitive to effects of synonymous variants. To evaluate this possibility, we randomly selected 500,000 each *observed* and *generated* non-synonymous variants from dbNSFP (Liu et al., 2011; Liu et al., 2016) and extracted their associated predictor scores (see **Supplementary Material**). Briefly, an nsSNV was considered *observed* if it was reported by either 1000G, ExAC, or gnomAD; otherwise it was deemed a *generated* nsSNV. While some of the predictors were slightly better at differentiating *generated* from *observed* nsSNVs (**Figure 4**, **Table 3**) than sSNVs, their performance was still not up to the expectations. We also calculated FCBP (**Figure 4A**; cutoffs as above) and score correlation (**Figure 4B**) to find that CADD, DANN, and FATHMM-MKL have a considerably higher agreement on nsSNVs than on sSNVs (**Figure 3A**).

Inferring a Predictor Scoring Threshold From Prediction of Common Variant Effects

The predictor inability to differentiate *observed* and *generated* variants may also be due to the difficulty of defining effect threshold; i.e., variants of low effect are harder to precisely annotate, both computationally and experimentally, and can be equally well classified as effect or neutral. In an effort to increase resolution between the two, predictors often link high allele frequency to absence of effect. In fact, CADD, DANN, FATHMM-MKL, SilVA, and regSNP-splicing effectively label high allele frequency variants as neutral. Taken further, TraP scores were reported (Gelfman et al., 2017) to have negative correlation (−0.51) with bin-average ExAC allele frequencies (Lek et al., 2016). Note that, as mentioned above, this reasoning side-steps evolutionary flow where common (not yet fixed or removed) variants may be advantageous or damaging. To further elaborate on allele frequency relationship with effect predictions, we obtained frequency data from multiple sources (1000G, ExAC, and gnomAD, see **Supplementary Material**) for our *observed* variants. Notably, we saw no correlation, positive or negative, between allele frequency and any predictor score (**Figure 5**). This observation highlights predictor binary classification abilities rather than a continuous spectrum of effect.

For some of the predictors (CADD, SilVA, TraP, DDIG-SN) high scoring variants were overwhelmingly of low frequency. At the same time, many of the low frequency variants were low scoring. Assuming that the predictor scores can be used as reliable indicators of common variant neutrality (low scoring), this result reinforces the idea that low frequency variants are as likely to be pathogenic/effect as neutral/benign. Furthermore, common variant score distributions could help approximate the predictor thresholds of effect. Thus, while variants scoring above a certain threshold can be considered to have an effect, below this threshold binary predictor resolution is questionable.

Predictor thresholds were chosen as the score below which most (99%) of the common variants (allele frequency >0.01) reside (**Figure 5**). Thus, scores above this threshold indicate effect, while scores below the threshold could be effect or neutral.

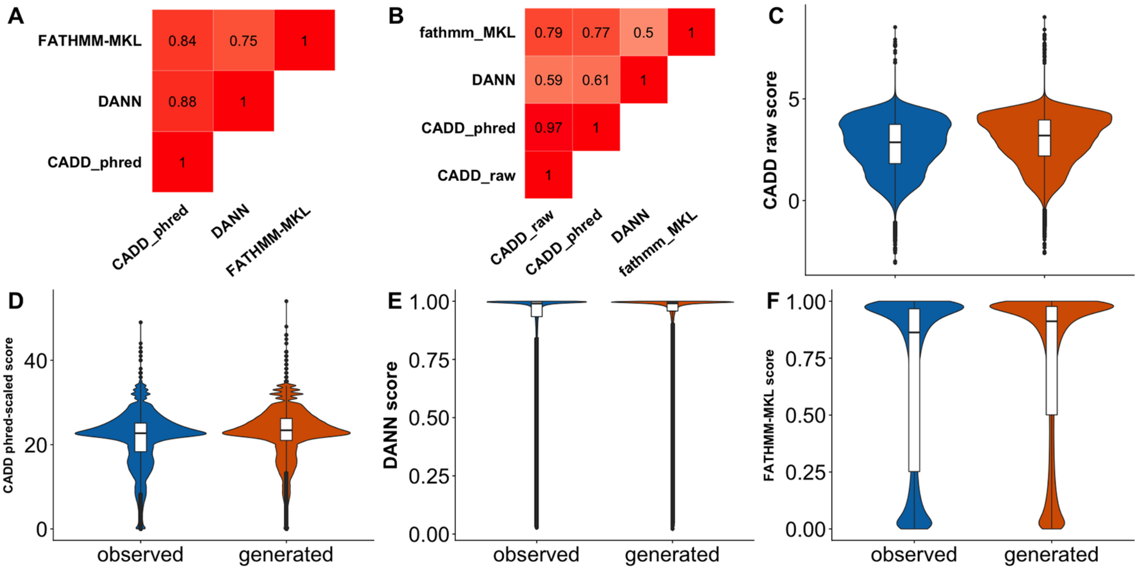


FIGURE 4 | Predictor scores correlate, but do not clearly differentiate *observed* vs. *generated* nsSNVs. Panel **(A)** shows the amount of agreement (i.e., FCBP) for any pair of predictors. High FCBP values indicate that two predictors agree in assigning binary (neutral/deleterious) predictions to variants. Panel **(B)** shows the Pearson correlations among the prediction scores. **(C–F)** Violin/box plots of prediction score distributions across predictors: CADD raw, CADD phred-scaled, DANN, and FATHMM-MKL, respectively.

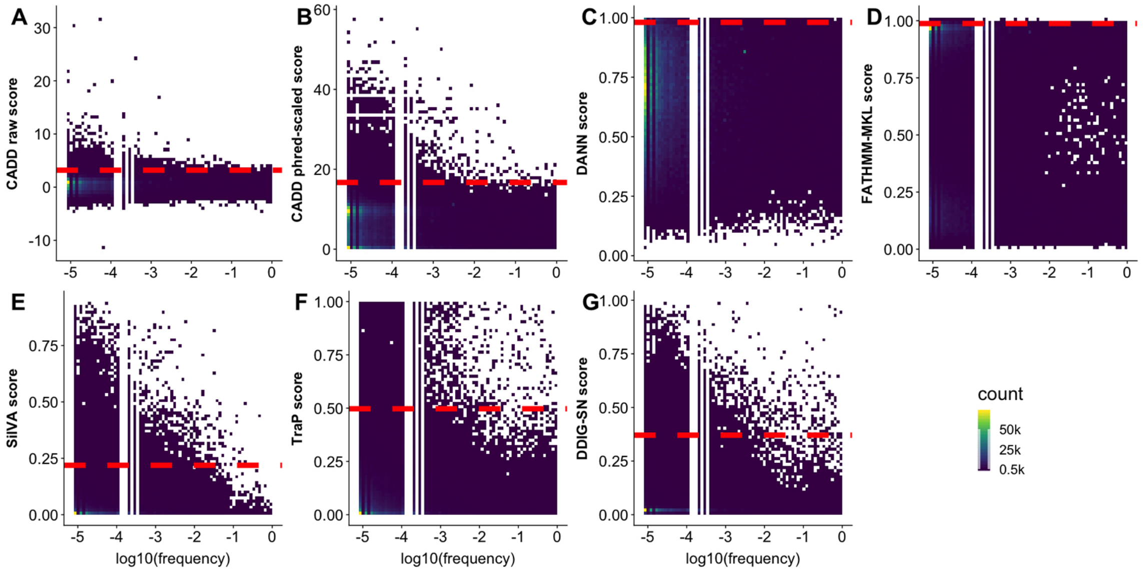


FIGURE 5 | Some predictors assign higher scores to rare variants. In all panels, the scatterplots display the density of *observed* variant prediction scores *vs.* $\log_{10}$(allele frequency). A scoring threshold (red dashed line) for each predictor identifies scores above the threshold as reliable. The threshold is placed at the score that is higher than 99% of common (allele frequency > 0.01) variant scores. **(A-G)** represents the scatterplot for CADD raw, CADD phred-scaled, DANN, FATHMM-MKL, SilVA, TraP, and DDIG-SN, respectively.

We further applied the selected thresholds to both *observed* and *generated* sSNVs (**Table 4**). We define *resolution* (Eqn 2, where "N" stands for number) as a predictor's ability to capture the enrichment of deleterious variants above threshold.

$$resolution = \frac{N_{sSNVs\ above\ the\ threshold}}{N_{observed\ sSNVs}} \times \frac{N_{generated\ sSNVs}}{N_{generated\ sSNVs\ above\ the\ threshold}} \quad (2)$$

The *resolutions* were greater than one for all the predictors, with CADD attaining the highest resolution (> 2). Note that only a small fraction of variants in both sets scored above the threshold, but since the total number of *generated* variants is nearly 18 times higher than the number of *observed* variants, the estimated number of potential identifiably-deleterious sSNVs is only an order of magnitude less than ALL observed sSNVs (~475K vs. ~1.3M). These results suggest that the *generated* set indeed contains many more deleterious variants than the *observed* set and that a new predictor train to recognize these differences may identify deleterious variants more reliably than existing methods.

TABLE 4 | Percentage of sSNVs scoring above threshold and the corresponding predictor resolutions.

	% Above-the-threshold sSNVs in *observed*	% above-the-threshold sSNVs in *generated*	Resolution
CADD raw score	0.871	1.981	2.274
CADD phred-scaled score	0.868	1.979	2.280
DANN	1.594	2.156	1.352
FATHMM-MKL	1.639	2.522	1.538
SilVA	4.902	6.015	1.227
TraP	2.376	2.912	1.226
DDIG-SN	1.764	2.414	1.368

CONCLUSION

Training data is perhaps the most critical component for a machine learning-based variant-effect-predictor. Most sSNV effect predictors we reviewed, retrieved training data from disease mutation databases, such as HGMD and ClinVar. Disease-causing variants can be thought of as severely functionally deleterious, although non-disease variants could also be deleterious or beneficial. Moreover, identifying an sSNV as disease causing, as opposed to associated with disease, is extremely difficult, if not impossible. In fact, studies have revealed flaws of existing disease mutation databases, which may further undermine the reliability of the contained data. Progress in saturation genome mutagenesis may improve data availability in the near future. Currently, however, there is no publicly available, sufficiently large collection of variants with experimentally validated effect annotations that can be used for building a generalizable sSNV effect-predictor.

The lack of gold standard data also prevents proper evaluation of the predictors. Here, we proposed a *Test set* of *observed* and *generated* sSNVs. We assumed that the *generated* set is enriched for deleterious sSNVs due to purifying selection and expected the predictors to differentiate these from the *observed* variants. However, the predictor performance on this data was below our expectations. Note that predictor scores for the variants in our set were poorly correlated and the amount of binary prediction agreement was limited. This observation suggests that predictions may be biased by shared input features, but do not sufficiently well indicate variant effect. We proposed a scoring threshold to separate reliable predictions from the highly uncertain ones for each of the predictor. With the thresholds identified, we further observed that all predictors had significantly more reliably identified sSNVs in the *generated* set than in *observed* set, in line with our earlier expectations of the quality and contents of the

Test set. However, the inability of the predictors to clearly identify effect variants below the severity threshold, suggests that more work is necessary to understand sSNV effects.

We note that our *Test set* is not a gold-standard testing set and is only appropriate for predictor testing only if our underlying biological/data distribution assumptions hold. Thus, we cannot make concrete recommendations of best-practice prediction tools. However, our results clearly indicate that the predictions are highly correlated across sSNV-specific methods, i.e., there is little difference between using SilVA, DDIG-SN, or TraP. On the other hand, outputs of general purpose-predictors (CADD, DANN, and FATHMM-MKL) do not correlate as well. Of these, CADD phred-scaled scores are least likely to classify common variants as having a large effect; i.e., CADD high scores may be deemed reliably non-neutral. Note, however, that this does not mean that CADD low scores indicate variant neutrality – a necessary distinction that evades much of the variant effect literature.

Looking forward to a future sSNV effect-predictor, we note that comparing *observed* vs. *generated,* rather than effect vs. no-effect, variants drastically increases the amount of data useful for training. We also note that this variant collection will need further filtering to address the problem of false positives, i.e., the yet-to-be-*observed generated* variants. Moreover, the transition from *observed* to no-effect and from *generated* to effect annotations will not be trivial. As mentioned earlier, while severe effect variants are likely predominantly confined to the *generated* set, the mild effect variation is probably distributed throughout both *observed* and *generated* collections. Despite these difficulties, the observation that existing predictors identify more higher-scoring effect variants in the *generated* data, suggests that the effect signal can indeed be learnable by models trained to differentiate *observed* vs *generated* variants. Thus, a model using the previously mentioned set of features, possibly in combination with an ensemble of (orthogonal, as evaluated above) existing classifiers, may provide a more reliable description of variant effects.

DATA AVAILABILITY STATEMENT

Public available datasets were analyzed in this study. Human transcripts and genomic coordinate information (GRCh37) can be found at https://grch37.ensembl.org/biomart/martview/e1515959acf51b72adec3001b7e02e59. DANN scores can be found at https://cbcl.ics.uci.edu/public_data/DANN/. TraP scores can be found at http://innovation.columbia.edu/technologies/cu17233_pathogenicity-database-for-identification-of-disease-causing-non-coding-genetic-variations. FATHMM-MKL scores can be found at https://github.com/HAShihab/fathmm-MKL. ANNOVAR annotation tool can be found at http://annovar.openbioinformatics.org/en/latest/. dbNSFP annotation tool can be found at https://sites.google.com/site/jpopgen/dbNSFP/. DDISN-SN server is at http://sparks-lab.org/ddig/. SilVA predictor is at http://compbio.cs.toronto.edu/silva/. MutationTaster2 server is at http://www.mutationtaster.org/StartQueryEngine.html. IDSV can be found at http://bioinfo.ahu.edu.cn:8080/IDSV. Our observed/generated data with predicted scores can be downloaded at https://doi.org/10.5281/zenodo.3471642.

AUTHOR CONTRIBUTIONS

ZZ and YB contributed to the idea conception, analysis design, literature review, and manuscript writing. ZZ conducted data collection, analysis, and visualization.

FUNDING

ZZ and YB Were Supported by the NIH U01 GM115486 Grant.

ACKNOWLEDGMENTS

We would like to thank to thank Dr. Junfeng Xia (Anhui University) for providing IDSV predictions, Dr. Martin Kircher (Berlin Institute of Health) for providing CADD training data, Dr. Jana Marie Schwarz and Dr. Dominik Seelow (both from Charité Berlin) for technical support with MutationTaster2, and Dr. Yunlong Liu (Indiana University) for technical support with DDIG-SN. Particular thanks go to all Bromberg lab members (Dr. Chengsheng Zhu, Dr. Maximillian Miller, Dr. Ariel Aptekmann, Dr. Adrienne Hoarfrost, Dr. Kenneth McGinness, Yannick Mahlich, and Yanran Wang, all Rutgers) for their constructive discussion and advice. We also acknowledge people of the Rutgers Office of Advanced Research Computing (OARC), and particularly Kevin Abbey and Galen Collier, for providing technical support and access to the compute cluster and associated research computing resources necessary for the work reported here. Finally, we would like to thank all researchers that deposit their data into public databases.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00914/full#supplementary-material

REFERENCES

Amendola, L. M., Dorschner, M. O., Robertson, P. D., Salama, J. S., Hart, R., Shirts, B. H., et al. (2015). Actionable exomic incidental findings in 6503 participants: challenges of variant classification. *Genome Res.* 25, 305–315 gr. 183483.114. doi: 10.1101/gr.183483.114

Angellotti, M. C., Bhuiyan, S. B., Chen, G., and Wan, X.-F. (2007). CodonO: codon usage bias analysis within and across genomes. *Nucleic Acids Res.* 35, W132–W136. doi: 10.1093/nar/gkm392

Bali, V., and Bebok, Z. (2015). Decoding mechanisms by which silent codon changes influence protein biogenesis and function. *Int. J. Biochem. Cell Biol.* 64, 58–74. doi: 10.1016/j.biocel.2015.03.011

Bauer, P., Karges, E., Oprea, G., and Rolfs, A. (2018). Unmet needs in human genomic variant interpretation. *Genet. Med.* 20, 376–377. doi: 10.1038/gim.2017.187

Bell, C. J., Dinwiddie, D. L., Miller, N. A., Hateley, S. L., Ganusova, E. E., Mudge, J., et al. (2011). Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Trans. Med.* 3, 65ra4–65ra4. doi: 10.1126/scitranslmed.3001756

Bernstein, B. E., Meissner, A., and Lander, E. S. (2007). The mammalian epigenome. *Cell* 128, 669–681. doi: 10.1016/j.cell.2007.01.033

Birney, E., and Soranzo, N. (2015). Human genomics: the end of the start for population sequencing. *Nature* 526, 52–53. doi: 10.1038/526052a

Boël, G., Letso, R., Neely, H., Price, W. N. Wong, K.-H., Su, M., et al. (2016). Codon influence on protein expression in E. coli correlates with mRNA levels. *Nature* 529, 358. doi: 10.1038/nature16509

Bromberg, Y., Kahn, P. C., and Rost, B., (2013). Neutral and weakly nonneutral sequence variants may define individuality. *Proc. Natl. Acad. Sci.* 110, 14255–14260. doi: 10.1073/pnas.1216613110

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., et al. (2016). Synonymous codons direct cotranslational folding toward different protein conformations. *Mol. Cell* 61, 341–351. doi: 10.1016/j.molcel.2016.01.008

Buske, O. J., Manickaraj, A., Mital, S., Ray, P. N., and Brudno, M. (2013). Identification of deleterious synonymous variants in human genomes. *Bioinformatics* 29, 1843–1850. doi: 10.1093/bioinformatics/btt308

Cannarozzi, G., Schraudolph, N. N., Faty, M., Von Rohr, P., Friberg, M. T., Roth, A. C., et al. (2010). A role for codon order in translation dynamics. *Cell* 141, 355–367. doi: 10.1016/j.cell.2010.02.036

Carbone, A., Zinovyev, A., and Képes, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 19, 2005–2015. doi: 10.1093/bioinformatics/btg272

Cassa, C. A., Tong, M. Y., and Jordan, D. M. (2013). Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum. Mutat.* 34, 1216–1220. doi: 10.1002/humu.22375

Chen, R., Davydov, E. V., Sirota, M., and Butte, A. J. (2010). Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PloS One* 5, e13574. doi: 10.1371/journal.pone.0013574

Coghlan, A., and Wolfe, K. H. (2000). Relationship of codon bias to mRNA concentration and protein length in *Saccharomyces cerevisiae*. *Yeast* 16, 1131– 1145. doi: 10.1002/1097-0061(20000915)16:12<1131::AID-YEA609>3.0.CO;2-F

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

Cooper, D. N., Ball, E. V., and Krawczak, M. (1998). The human gene mutation database. *Nucleic Acids Res.* 26, 285–287. doi: 10.1093/nar/26.1.285

Cooper, G. M., Stone, E. A., Asimenos, G., Green, E. D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.* 15, 901–913. doi: 10.1101/gr.3577405

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* 6, e1001025. doi: 10.1371/journal.pcbi.1001025

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 37, e67–e67. doi: 10.1093/nar/gkp215

Dorschner, M. O., Amendola, L. M., Turner, E. H., Robertson, P. D., Shirts, B. H., Gallego, C. J., et al. (2013). Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am. J. Hum. Genet.* 93, 631–640. doi: 10.1016/j.ajhg.2013.08.006

Dos Reis, M., Wernisch, L., and Savva, R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole Escherichia coli K-12 genome. *Nucleic Acids Res.* 31, 6976–6985. doi: 10.1093/nar/gkg897

Duan, J., Shi, J., Ge, X., Dölken, L., Moy, W., He, D., et al. (2013). Genome-wide survey of interindividual differences of RNA stability in human lymphoblastoid cell lines. *Sci. Rep.* 3, 1318. doi: 10.1038/srep01318

Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. *Nature* 513, 120–123. doi: 10.1038/nature13695

Findlay, G. M., Daza, R. M., Martin, B., Zhang, M. D., Leith, A. P., Gasperini, M., et al. (2018). Accurate classification of BRCA1 variants with saturation genome editing. *Nature* 562, 217–222. doi: 10.1038/s41586-018-0461-z

Friedman, N., Ninio, M., Pe'er, I., and Pupko, T. (2002). A structural EM algorithm for phylogenetic inference. *J. Comput. Biol.* 9, 331–353. doi: 10.1089/10665270252935494

Gelfman, S., Cohen, N., Yearim, A., and AST, G. (2013). DNA-methylation effect on co-transcriptional splicing is dependent on GC-architecture of the exon–intron structure. *Genome Res.* 23, 789–799 gr. 143503.112. doi: 10.1101/gr.143503.112

Gelfman, S., Wang, Q., McSweeney, K. M., Ren, Z., La Carpia, F., Halvorsen, M., et al. (2017). Annotating pathogenic non-coding variants in genic regions. *Nat. Commun.* 8, 236. doi: 10.1038/s41467-017-00141-2

George, R. A., Smith, T. D., Callaghan, S., Hardman, L., Pierides, C., Horaitis, O., et al. (2007). General mutation databases: analysis and review. *J. Med. Genet.* 45, 65–70 doi: 10.1136/jmg.2007.052639

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* 13, 135–145. doi: 10.1038/nrg3118

Giulietti, M., Piva, F., D'Antonio, M., D'Onorio De Meo, P., Paoletti, D., Castrignano, T., et al. (2013). SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res.* 41, D125–D131. doi: 10.1093/nar/gks997

Gradishar, W., Johnson, K., Brown, K., Mundt, E., and Manley, S. (2017). Clinical variant classification: a comparison of public databases and a commercial testing laboratory. *Oncol.* 22, 797–803. doi: 10.1634/theoncologist.2016-0431

Guo, F. B., Ye, Y. N., Zhao, H. L., Lin, D., and Wei, W. (2012). Universal pattern and diverse strengths of successive synonymous codon bias in three domains of life, particularly among prokaryotic genomes. *DNA Res.* 19, 477–485. doi: 10.1093/dnares/dss027

Hamosh, A. (2004). Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33, D514–D517. doi: 10.1093/nar/gki033

Hasegawa, M., Kishino, H., and Yano, T.-A. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22, 160–174. doi: 10.1007/BF02101694

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res.* 31, 3429–3431. doi: 10.1093/nar/gkg599

Holtkamp, W., Kokic, G., Jäger, M., Mittelstaet, J., Komar, A. A., and Rodnina, M. V. (2015). Cotranslational protein folding on the ribosome monitored in real time. *Science* 350, 1104–1107. doi: 10.1126/science.aad0344

Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E., and Kimchi-Sarfaty, C. (2014). Exposing synonymous mutations. *Trends in Genet.* 30, 308–321. doi: 10.1016/j.tig.2014.04.006

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210. doi: 10.1101/531210

Khan, A. H., Lin, A., and Smith, D. J. (2012). Discovery and characterization of human exonic transcriptional regulatory elements. *PloS One* 7, e46098. doi: 10.1371/journal.pone.0046098

Kim, J., and Bang, H. (2016). Three common misuses of P values. *Dent. Hypotheses* 7, 73. doi: 10.4103/2155-8213.190481

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V., et al. (2007). "A" silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* 315, 525–528. doi: 10.1126/science.1135308

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* 46, 310. doi: 10.1038/ng.2892

Komar, A. A. (2009). A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.* 34, 16–24. doi: 10.1016/j.tibs.2008.10.002

Komar, A. A. (2016). The Yin and Yang of codon usage. *Hum. Mol. Genet.* 25, R77–R85. doi: 10.1093/hmg/ddw207

Kramer, E. B., and Farabaugh, P. J. (2006). The frequency of translational misreading errors in E. coli is largely determined by tRNA competition. *RNA*. 13, 87–96. doi: 10.1261/rna.294907

Kramer, E. B., Vallabhaneni, H., Mayer, L. M., and Farabaugh, P. J. (2010). A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. *RNA*. 16, 1797–1808 .doi: 10.1261/rna.2201210

Krawczak, M., Reiss, J., and Cooper, D. N. (1992). The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. *Hum. Genet.* 90, 41–54. doi: 10.1007/BF00210743

Landrum, M. J., and Kattman, B. L. (2018). ClinVar at five years: delivering on the promise. *Hum. Mutat.* 39, 1623–1630. doi: 10.1002/humu.23641

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 42, D980–D985. doi: 10.1093/nar/gkt1113

Lang, G., Gombert, W. M., and Gould, H. J. (2005). A transcriptional regulatory element in the coding sequence of the human Bcl-2 gene. *Immunology* 114, 25–36. doi: 10.1111/j.1365-2567.2004.02073.x

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285–291. doi: 10.1038/nature19057

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum. Mutat.* 32, 894–899. doi: 10.1002/humu.21517

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Hum. Mutat.* 37, 235–241. doi: 10.1002/humu.22932

Livingstone, M., Folkman, L., Yang, Y., Zhang, P., Mort, M., Cooper, D. N., et al. (2017). Investigating DNA-, RNA-, and protein-based features as a means to discriminate pathogenic synonymous variants. *Hum. Mutat.* 38, 1336–1347. doi: 10.1002/humu.23283

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., et al. (2013). The genotype-tissue expression (GTEx) project. *Nat. Genet.* 45, 580. doi: 10.1038/ng.2653

Lorenz, R., Wolfinger, M. T., Tanzer, A., and Hofacker, I. L. (2016). Predicting RNA secondary structures from sequence and probing data. *Methods* 103, 86–98. doi: 10.1016/j.ymeth.2016.04.004

Mahlich, Y., Reeb, J., Hecht, M., Schelling, M., De Beer, T. A. P., Bromberg, Y., et al. (2017). Common sequence variants affect molecular function more than rare variants? *Sci. Rep.* 7, 1608. doi: 10.1038/s41598-017-01054-2

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* 583, 3966–3973. doi: 10.1016/j.febslet.2009.10.036

Markham, N. R., and Zuker, M. (2008). UNAFold. *Bioinformatics* 3-31. doi: 10.1007/978-1-60327-429-6_1

Meyer, I. M. (2005). Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res.* 33, 6338–6348. doi: 10.1093/nar/gki923

Miller, M., Bromberg, Y., and Swint-Kruse, L. (2017). Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Sci. Rep.* 7, 41329. doi: 10.1038/srep41329

Miller, M., Vitale, D., Rost, B., and Bromberg, Y. (2019). fuNTRp: identifying protein positions for variation driven functional tuning. *bioRxiv* 578757. doi: 10.1101/578757

Nachman, M. W., and Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics* 156, 297–304.

Nakagomi, H., Mochizuki, H., Inoue, M., Hirotsu, Y., Amemiya, K., Sakamoto, I., et al. (2018). Combined annotation-dependent depletion score for BRCA1/2 variants in patients with breast and/or ovarian cancer. *Cancer Sci.* 109, 453–461. doi: 10.1111/cas.13464

Ng, P. C. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* 31, 3812–3814. doi: 10.1093/nar/gkg509

Niroula, A., and Vihinen, M. (2016). Variation interpretation predictors: principles, types, performance, and choice. *Hum. Mutat.* 37, 579–597. doi: 10.1002/humu.22987

Novoa, E. M., Pavon-Eternod, M., Pan, T., and De Pouplana, L. R. (2012). A role for tRNA modifications in genome structure and codon usage. *Cell* 149, 202–213. doi: 10.1016/j.cell.2012.01.050

Pagani, F., Raponi, M., and Baralle, F. E. (2005). Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci.* 102, 6368–6372. doi: 10.1073/pnas.0502288102

Parmley, J. L. (2005). Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Molecular biology and evolution* 23, 301–309. doi: 10.1093/molbev/msj035

Pechmann, S., and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Boil.* 20, 237. doi: 10.1038/nsmb.2466

Plotkin, J. B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* 12, 32. doi: 10.1038/nrg2899

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2009). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* 20, 110–121. doi: 10.1101/gr.097857.109

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., et al. (2015). Codon optimality is a major determinant of mRNA stability. *Cell* 160, 1111–1124. doi: 10.1016/j.cell.2015.02.029

Quang, D., Chen, Y., and XIE, X. (2014). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31, 761–763. doi: 10.1093/bioinformatics/btu703

Rehm, H. L., Berg, J. S., Brooks, L. D., Bustamante, C. D., Evans, J. P., Landrum, M. J., et al. (2015). ClinGen—the clinical genome resource. *New Engl. J. Med.* 372, 2235–2242. doi: 10.1056/NEJMsr1406261

Reis, M. D., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32, 5036–5044. doi: 10.1093/nar/gkh834

Rost, B., Radivojac, P., and Bromberg, Y. (2016). Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett.* 590, 2327–2341. doi: 10.1002/1873-3468.12307

Salari, R., Kimchi-Sarfaty, C., Gottesman, M. M., and Przytycka, T. M. (2012). Detecting SNP-induced structural changes in RNA: application to disease studies 241–243. Springer. doi: 10.1007/978-3-642-29627-7_25

Sauna, z. E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* 12, 683. doi: 10.1038/nrg3051

Schaafsma, G. C. P., and Vihinen, M. (2015). VariSNP, a benchmark database for variations from dbSNP. *Hum. Mutat.* 36, 161–166. doi: 10.1002/humu.22727

Schwarz, J. M., Cooper, D. N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* 11, 361. doi: 10.1038/nmeth.2890

Seffens, W. (1999). mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res.* 27, 1578–1584. doi: 10.1093/nar/27.7.1578

Shabalina, S. A., Spiridonov, N. A., and Kashina, A. (2013). Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res.* 41, 2073–2094. doi: 10.1093/nar/gks1205

Shah, N., Hou, Y.-C. C., Yu, H.-C., Sainger, R., Caskey, C. T., Venter, J. C., et al. (2018). Identification of misclassified clinvar variants *via* disease population prevalence. *Am. J. Hum. Genet.* 102, 609–619. doi: 10.1016/j.ajhg.2018.02.019

Shah, P., and Gilchrist, M. A. (2010). Effect of correlated tRNA abundances on translation errors and evolution of codon usage bias. *PLoS Genet.* 6, e1001128. doi: 10.1371/journal.pgen.1001128

Sharp, P. M., and Li, W.-H. (1987). The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15, 1281–1295. doi: 10.1093/nar/15.3.1281

Shen, H., Li, J., Zhang, J., Xu, C., Jiang, Y., Wu, Z., et al. (2013). Comprehensive characterization of human genome variation by high coverage whole-genome sequencing of forty four Caucasians. *PLoS One* 8, e59494. doi: 10.1371/journal.pone.0059494

Shepard, P. J., and Hertel, K. J. (2009). The SR protein family. *Genome Biol.* 10, 242. doi: 10.1186/gb-2009-10-10-242

Shi, F., Yao, Y., Bin, Y., Zheng, C.-H., and Xia, J. (2019). Computational identification of deleterious synonymous variants in human genomes using a feature-based approach. *BMC Med. Genom.* 12, 12. doi: 10.1186/s12920-018-0455-6

Shihab, H. A., Rogers, M. F., Gough, J., Mort, M., Cooper, D. N., Day, I. N., et al. (2015). An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31, 1536–1543. doi: 10.1093/bioinformatics/btv009

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15, 1034–1050. doi: 10.1101/gr.3715005

Smith, J. M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23. doi: 10.1017/S0016672300014634

Sørensen, M. A., Kurland, C., and Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. *J. Mol. Biol.* 207, 365–377. doi: 10.1016/0022-2836(89)90260-X

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., et al. (2003). Human gene mutation database (HGMD®): 2003 update. *Hum. Mutat.* 21, 577–581. doi: 10.1002/humu.10212

Stenson, P. D., Mort, M., Ball, E. V., Evans, K., Hayden, M., Heywood, S., et al. (2017). The human gene mutation database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* 136, 665–677. doi: 10.1007/s00439-017-1779-6

Stenson, P. D., Mort, M., Ball, E. V., Howells, K., Phillips, A. D., Thomas, N. S., et al. (2009). The human gene mutation database: 2008 update. *Genome Med.* 1, 13. doi: 10.1186/gm13

Stergachis, A. B., Haugen, E., Shafer, A., Fu, W., Vernot, B., Reynolds, A., et al. (2013). Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* 342, 1367–1372. doi: 10.1126/science.1243490

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156, 1324–1335. doi: 10.1016/j.cell.2014.01.051

Thanaraj, T., and Argos, P. (1996). Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci.* 5, 1973–1983. doi: 10.1002/pro.5560051003

Van Der Velde, K. J., Kuiper, J., Thompson, B. A., Plazzer, J. P., Van Valkenhoef, G., De Haan, M., et al. (2015). Evaluation of CADD scores in curated mismatch repair gene variants yields a model for clinical validation and prioritization. *Hum. Mutat.* 36, 712–719. doi: 10.1002/humu.22798

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of GWAS discovery. *Am. J. Hum. Genet.* 90(1), 7–24. doi: 10.1016/j.ajhg.2011.11.029

Wang, Z., and Burge, C. B. (2008). Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA* 14, 802–813. doi: 10.1261/rna.876308

Welter, D., Macarthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2013). The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, D1001–D1006. doi: 10.1093/nar/gkt1229

Wen, P., Xiao, P., and Xia, J. (2016). dbDSM: a manually curated database for deleterious synonymous mutations. *Bioinformatics* 32, 1914–1916. doi: 10.1093/bioinformatics/btw086

Xayaphoummine, A., Bucher, T., and Isambert, H. (2005). Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Res.* 33, W605–W610. doi: 10.1093/nar/gki447

Xing, Y., Zhao, X., Yu, T., Liang, D., Li, J., Wei, G., et al. (2016). MiasDB: a database of molecular interactions associated with alternative splicing of human pre-mRNAs. *PloS One* 11, e0155443. doi: 10.1371/journal.pone.0155443

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., et al. (2015). The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347, 1254806–1254806. doi: 10.1126/science.1254806

Xue, Y., Chen, Y., Ayub, Q., Huang, N., Ball, E. V., Mort, M., et al. (2012). Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* 91, 1022–1032. doi: 10.1016/j.ajhg.2012.10.015

Yue, P., and Moult, J. (2006). Identification and analysis of deleterious human SNPs. *J. Mol. Biol.* 356, 1263–1274. doi: 10.1016/j.jmb.2005.12.025

Zhang, G., Hubalewska, M., and Ignatova, Z. (2009). Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Boil.* 16, 274. doi: 10.1038/nsmb.1554

Zhang, G., and Ignatova, Z. (2011). Folding at the birth of the nascent chain: coordinating translation with co-translational folding. *Curr. Opin. Struct. Biol.* 21, 25–31. doi: 10.1016/j.sbi.2010.10.008

Zhang, X., Li, M., Lin, H., Rao, X., Feng, W., Yang, Y., et al. (2017). regSNPs-splicing: a tool for prioritizing synonymous single-nucleotide substitution. *Hum. Genet.* 136, 1279–1289. doi: 10.1007/s00439-017-1783-x

Zhou, T., Weems, M., and Wilke, C. O. (2009). Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol Biol. Evol.* 26, 1571–1580. doi: 10.1093/molbev/msp070

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31, 3406–3415 doi: 10.1093/nar/gkg595

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: 22 October 2019
doi: 10.3389/fgene.2019.01009

Measurement of Conditional Relatedness Between Genes Using Fully Convolutional Neural Network

Yan Wang[1,3], *Shuangquan Zhang*[1], *Lili Yang*[2], *Sen Yang*[1], *Yuan Tian*[3]* *and Qin Ma*[4]

[1] Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun, China, [2] Department of Obstetrics, The First Hospital of Jilin University, Changchun, China, [3] School of Artificial Intelligence, Jilin University, Changchun, China, [4] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, United States

Edited by:
Dominik Heider,
University of Marburg,
Germany

Reviewed by:
Holger Fröhlich,
University of Bonn,
Germany
Leyi Wei,
Tianjin University,
China

Correspondence:
Yuan Tian
yuantian@jlu.edu.cn

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

Received: *18 April 2019*
Accepted: *23 September 2019*
Published: *22 October 2019*

Citation:
Wang Y, Zhang S, Yang L, Yang S, Tian Y and Ma Q (2019) Measurement of Conditional Relatedness Between Genes Using Fully Convolutional Neural Network. Front. Genet. 10:1009. doi: 10.3389/fgene.2019.01009

Measuring conditional relatedness, the degree of relation between a pair of genes in a certain condition, is a basic but difficult task in bioinformatics, as traditional co-expression analysis methods rely on co-expression similarities, well known with high false positive rate. Complement with prior-knowledge similarities is a feasible way to tackle the problem. However, classical combination machine learning algorithms fail in detection and application of the complex mapping relations between similarities and conditional relatedness, so a powerful predictive model will have enormous benefit for measuring this kind of complex mapping relations. To this need, we propose a novel deep learning model of convolutional neural network with a fully connected first layer, named fully convolutional neural network (FCNN), to measure conditional relatedness between genes using both co-expression and prior-knowledge similarities. The results on validation and test datasets show FCNN model yields an average 3.0% and 2.7% higher accuracy values for identifying gene–gene interactions collected from the COXPRESdb, KEGG, and TRRUST databases, and a benchmark dataset of Xiao-Yong *et al.* research, by grid-search 10-fold cross validation, respectively. In order to estimate the FCNN model, we conduct a further verification on the GeneFriends and DIP datasets, and the FCNN model obtains an average of 1.8% and 7.6% higher accuracy, respectively. Then the FCNN model is applied to construct cancer gene networks, and also calls more practical results than other compared models and methods. A website of the FCNN model and relevant datasets can be accessed from https://bmbl.bmi.osumc.edu/FCNN.

Keywords: conditional relatedness between genes, fully convolutional neural network, co-expression similarity, prior-knowledge similarity, gene network

INTRODUCTION

Conditional relatedness between a pair of genes is a degree of the relation between two genes in a certain condition, *e.g.* in cancer tissues or inflammation, implying the probability of these genes jointly involved in a biological process under such cell environment (Wang et al., 2019). It is different from gene–gene interaction meaning a 0/1 (non-interacting/interacting) binary relation between a pair of genes. Measuring such relatedness is a basic tool for understanding the biological and functional relations between genes in a real cell environment (Jelier et al., 2005; Mistry and Pavlidis, 2008). And the measured relatedness is classically used as weights on connections of genes for construction of gene networks in different environments for further biological analysis (Amrine et al., 2015; Li et al., 2018).

Traditionally, expression similarity as well as co-expression is used to measuring conditional relatedness, including Pearson correlation coefficient (PCC) (Eisen et al., 1998), Spearman rank correlation (SRC) (Kumari et al., 2012), mutual information (MI) (Song et al., 2012), partial Pearson correlation (PPC) (Baruch and Albert-László, 2013), and conditional mutual information (CMI) (Kim et al., 2010). PCC can express the linear relationship between a pair of genes, SRC and MI represent the nonlinear relationship, and PPC and CMI indicate the direct linear relationship and the direct nonlinear relationship under the condition of excluding other genes' interferences, respectively. Expression similarities have been successfully applied in measuring conditional relatedness for constructing gene networks, on which Poliakov et al. identify disease-related metabolic pathways (Poliakov et al., 2014). However, when acquiring gene expression data, it often contains some inevitable noise, which causes errors in the calculation of conditional relatedness, well known as high false positive rate.

Another type of similarity, prior-knowledge similarity, is also used to measure gene–gene relatedness, based on the documented biological data and functional annotations in public domain, such as the Gene Ontology (GO) (Consortium, 2004), the KEGG (Kanehisa and Goto, 2000), the Reactome (de Bono et al., 2005), the OrthoDB (Zdobnov et al., 2016), the TRRUST (Han and Puri, 2018), *etc.* It brings high accuracy (ACC) (Diebold and Mariano, 1995), as the prior-knowledge similarity is confirmed by the biological experiment. But the biological experiment is usually conducted in a normal condition, meaning prior-knowledge similarity is hardly used for measuring conditional relatedness.

By the above understanding, integration of expression and prior-knowledge similarities is an effective way to avoid the shortage of using only one category of similarity to measuring conditional relatedness between genes, as a pair of genes with high expression similarity but low prior-knowledge similarity implies their relatedness is most likely a false prediction by co-expression analysis, and the two genes with low expression similarity but high prior-knowledge similarity implies their relatedness is not specific in the condition. The gene pair with both high expression and prior-knowledge similarities should be scored a high rank and recommended by a model. Wang et al. proposed a support vector machine (SVM) model using both expression and prior-knowledge similarities to measure conditional relatedness between a pair of genes, and their computational results showed the proposed model outperforms the existing co-expression analysis methods and other integration models (Wang et al., 2019). The combination of both kinds of similarities has been also succeeded in other related biological issues, *e.g.*, detection of protein–protein interaction (PPI) (Jing and Ng, 2010), measuring functional similarity of gene products (Mistry and Pavlidis, 2008), and identification of disease-causing gene (Mohammadi et al., 2011).

Because of the fast growth of the deep learning technology, deep learning algorithms have outperformed the state-of-the-art traditional machine learning algorithms in many research field of bioinformatics. Babak *et al.* adapted the deep learning convolutional neural network to the task of predicting sequence specificities and showed that they compete favorably with the state of the art (Babak et al., 2015), and their results show that their approach outperforms other state-of-the-art methods. Pan and Shen proposed a deep learning-based framework by using a novel hybrid convolutional neural network and deep belief network to predict the RNA-binding proteins (RBP) interaction sites and motifs on RNAs.They validate their method on 31 large-scale datasets, and their experiments show that the average area under the curve (AUC) (Lobo et al., 2010) can be improved by 8% compared to the best single-source-based predictor (Pan and Shen, 2017). Trebeschi et al. applied the deep learning methods to automatic localization and segmentation of rectal cancers on multiparametric MRI, and their results demonstrate that deep learning can perform accurate localization and segmentation of rectal cancer in multiparametric MRI in the majority of patients (Trebeschi et al., 2017). Gao et al. proposed a new computational approach based on deep neural networks to predict tRNA gene sequences, and their proposed methods outperformed the existing methods under different metrics (Gao et al., 2019).

Motivated by the above mentioned, we develop a novel deep learning model of convolutional neural network (CNN) with a fully connected first layer, named fully convolutional neural network (FCNN), to measure conditional relatedness between genes using both expression and prior-knowledge similarities. The goal of our model is to keep and recommend gene pairs with both high expression and prior-knowledge similarities. The fully connected first layer makes our model extracting more useful information than traditional CNN and the rest CNN structure makes our model easier to train than all fully connected deep learning models. In line of the above two advantages and integrating of co-expression and prior-knowledge similarities, FCNN model calls better results than other models and methods for identifying gene–gene interactions and constructing cancer gene networks. First, the FCNN model acquires an average 3.0% and 2.7% higher ACC values on validation and test samples collected from the COXPRESdb, KEGG, and TRRUST databases and a benchmark dataset of Xiao-Yong et al. research (Xiao-Yong et al., 2010). Then we perform a further verification on the samples from the GeneFriends and DIP databases, and the FCNN model obtains an average of 1.8% and 7.6% higher accuracy, respectively. Finally, the FCNN model is utilized to construct cancer gene networks, which also obtains more practical results, comparing with other models and methods. The source code of FCNN, as well as the datasets and results of this research, are freely available in https://bmbl.bmi.osumc.edu/FCNN.

MATERIALS AND METHODS

Dataset Collection

We take gene pairs with/without expression similarity (co-expression) and prior-knowledge similarity (protein–protein interaction, involvement in a same pathway, and transcriptional regulation) as samples to compose a whole dataset to make our model be trained to predict gene pairs with high expression similarity as well as those with high prior-knowledge similarity at the same time, *i.e.*, to identify gene pairs with both high expression and prior-knowledge similarities. Therefore, the dataset used for training, validation, and test consists of two sub-datasets, so called co-expression sub-dataset and prior-knowledge sub-dataset.

The co-expression sub-dataset is collected from the COXPRESdb database (Release v7.1) (Yasunobu et al., 2015), where co-expressed gene pairs are sorted ascendingly by the mutual rank (MR) (Obayashi and Kinoshita, 2009). The smaller the MR value is, the higher co-expression it has. For each gene, we select the top 30 co-expressed genes to compose 30 co-expressed gene pairs from the Hsa-u.v18-10 and Mmu-u.v18-10 datasets in the COXPRESdb database, respectively. Then we select gene pairs as positive samples that they are commonly co-expressed in Hsa-u.v18-10 and Mmu-u.v18-10 datasets. To relieve the imbalanced problem between positive and negative samples, for each gene, we select middle 60 non-co-expressed genes to compose negative samples, similarly as composing the positives, where negative samples are the non-co-expressed gene pairs with PPC values around 0. There are 32,735 positive samples and 26,782 negative samples in the sub-dataset.

The prior-knowledge sub-dataset is composed of three parts. A) We collect gene-pair samples from the KEGG database (Release Nov 1, 2018) (Kanehisa, 2002) as the first part, where positive samples are gene pairs composited by the genes involved in at least two same pathways, and the negative samples are randomly selected gene pairs composited by the genes never engaged in the same pathway, with the same number of the positives. There are 11,526 positive samples and 11,526 negative samples in the first part. B) Next, for the second part, we use 15,222 gene pairs with PPI from a benchmark dataset of Pan et al. research (Pan et al., 2010) as the positive samples and 21,579 gene pairs without PPI as the negatives. C) In terms of the third part of the sub-dataset, we collect 7,361 gene pairs with the transcriptional regulation records in the TRRUST database (Release v2) (Han et al., 2017) as the positive samples and 7,361 gene pairs by random permutation of the transcription factor and the regulated gene in the positive ones (Nakamura et al., 2004; De et al., 2005; Wang et al., 2019).

Finally, there are a total of 66,844 positive and 67,248 negative samples. Specially, some negative samples were obtained by permutation of the positives and were then selected randomly to ensure the same number of positives for construction of a model with high generalization. And to avoid the bias of random permutation and selection of negative samples, we conduct the above process 100 times, rising to 100 datasets, in each of which a fixed percentage of the samples are used to training, validation, and test, according to the detailed proportion of the sub and sub-sub datasets. Also, the labels for the positive gene pairs are marked as 1s and those for the negatives as 0s. The details of each sub-dataset are showcased in **Table 1**.

TABLE 1 | The structure of FCNN dataset.

Sub dataset	Sub-sub dataset	Type of gene pair	Sample size
Co-expression	Co-expression	Positive	32,735
		Negative	26,782
Prior-knowledge	KEGG	Positive	11,526
		Negative	11,526
	PPI	Positive	15,222
		Negative	21,579
	TRRUST	Positive	7,136
		Negative	7,136
DIP	DIP	Positive	1,396
		Negative	1,396
GeneFriends	GeneFriends	Positive	8,675
		Negative	8,675

For model verification, the gene pairs downloaded from the GeneFriends (Release v3.1) (Sipko et al., 2015) and DIP (Release Feb 13, 2017) (Xenarios et al.) databases are utilized as samples. In the GeneFriends database, we select overall 8,675 co-expressed gene pairs with top 20 PCC values for each gene as the positive samples. Because there is only a small part of genes that are co-expressed in the human genome, we used 8,675 gene pairs obtained by random permutation of the first and second genes in the positive gene pairs as the negative samples. Similarly, 1,396 gene pairs with direct PPI collected from the DIP database are used as the positive samples. Considering gene pairs with real PPIs are rare in the whole human genome, the 1,396 gene pairs by permutation of the two genes in the positive samples are used as the negatives. To avoid the bias of permutation, we conduct the above process 100 times, rising to 100 datasets from the GeneFriends and DIP databases, respectively.

Gene-Pair Features Calculation

To measure conditional relatedness between a pair of genes and avoid the deficiencies of using a single type of feature, we use two kinds of features of gene pairs, including the expression similarities and prior-knowledge similarities.

In the former one, there are seven features, which are the average expression level of each gene of a gene pair, including $Mean_1$ and $Mean_2$, and five co-expression levels, including *PCC*, *SRC*, *PPC*, *MI*, and *CMI*. The expression data for calculation of expression similarities are collected from the GEO datasets (Barrett et al., 2012) based on the Affymetrix Human Genome U133 Plus 2.0 Array platform (released on Nov 2003). Then a pre-processing is executed, including log2 scale and quantile normalization.

The latter one contains five features such as GO similarity (*GOsim*) (Wang et al., 2007), subcellular location similarity (*LCsim*) (Yu et al., 2010), hormonology similarity (*HGsim*) (Chen and Vitkup, 2006), Reactome similarity (*RCsim*) (David et al., 2014), and transcriptional regulation similarity (*FRsim*) (Nagafuchi et al., 1991). The details of these features are defined as follows.

$$GOsim_{i,j} = \max_{g \in G_i, q \in G_j,} \frac{\log(Pms(g,q)^2)}{\log(P(\mathrm{g}) + \log P(q))} \tag{1}$$

where G_i and G_j represent the GO term sets used for annotating gene *i* and *j*, respectively; *p(g)* represents the probability of a gene annotated by an instance of GO term *g*, and *Pms(g,q)* represents the minimum probability of a gene annotated by an instance of a common ancestor GO term of *g* and *q*. The GO terms of genes used here are the biological process GO terms with experimental evidence downloaded from the GO database (Kumari et al., 2012), where a GO tree is built by the relations among GO terms, including "is a", "part of", "has part", and "regulates".

$$LCsim_{i,j} = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \quad (2)$$

where S_i and S_j represent the subcellular sets of two proteins encoded by gene *i* and gene *j*, respectively. The subcellular information of genes is collected from the GO database.

$$HGsim_{i,j} = \frac{L \times K - v_i \times v_j}{\sqrt{(L \times v_i - v_i^2)(L \times v_j - v_j^2)}} \quad (3)$$

where v_i and v_j represent the number of species whose genome contains homologous genes of gene *i* and *j*, respectively; *L* represents the total number of species; and *K* represents the number of species whose genome contains the homologous gene of both gene *i* and *j*.

$$RCsim_{i,j} = 1 - \frac{d_{i,j}}{d_{\max}} \quad (4)$$

where $d_{i,j}$ represents the shortest distance between gene *i* and gene *j* in the graph constructed by gene–gene interactions collected from the Reactome database (Croft et al., 2011), and d_{max} represents the shortest distance of the farthest gene pair.

$$FRsim_{i,j} = \begin{cases} 1, \text{if there is a transcriptional regulation record} \\ 0, \text{otherwise} \end{cases} \quad (5)$$

where $FRsim_{i,j}$ is equal to 1, if there is a transcriptional regulation between gene *i* and *j* recorded in the HTRIdb database (Bovolenta et al., 2012), and is equal to 0, otherwise. Meanwhile, all the databases and relevant data source used to compute these two kinds of gene-pair features are listed in **Supplementary Table S5**.

Model Construction

In the study, we design a model using CNN with a fully connected first layer, named FCNN to measure conditional relatedness of gene pairs shown as **Figure 1**. On one hand, the fully connected first layer of FCNN keeps our model from ignoring important feature combination. On the other hand, the CNN structure makes our model easy to train because of its parameter sharing and sparse connections. In detail, the model contains six layers. The first layer is a fully connected layer with 81 neurons and used for getting as much information as possible. The 12 features $X = [x_1,\ldots,x_{12}]$ as the inputs are fed into this layer to get the activation score a_j of neural *j*:

$$a_j = \sum_{i=1}^{12} \omega_{i,j}^{*} x_i + b_j \quad (6)$$

where ω_{ij} represents the weight between the x_i and neural *j*; and b_j represents the bias. Then we reshape the output $A_1 = [a_1,a_2,\ldots,a_{81}]$ into a 9*9 matrix A_1':

$$A_1' = \begin{bmatrix} a_1 & \cdots & a_9 \\ \vdots & \text{O} & \vdots \\ a_{73} & \cdots & a_{81} \end{bmatrix} \quad (7)$$

which is convenient to operate the convolution. The second layer is a convolutional layer using 20 convolutional kernels of size 2*2 and stride of 1. The output of each neuron of this layer is the

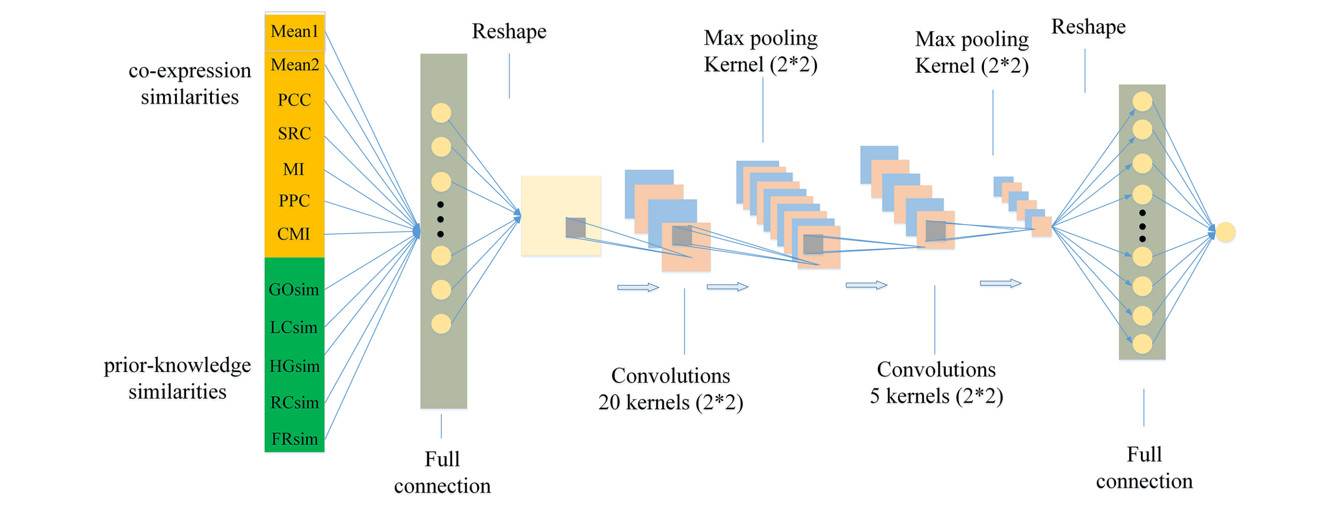


FIGURE 1 | The structure of the FCNN model.

convolution between a kernel matrix and a part of the input. The result A_2 of the second layer is defined as:

$$A_2 = tanh(Conv(A_1')) \tag{8}$$

where $Conv(\cdot)$ represents the convolution operation and $ReLU(\cdot)$ represents the rectified linear unit function. The third layer is a maximum pooling layer with the kernel of size 2*2 and stride of 2, which is used to down sample and reduce the dim of input by selecting the maximum value in each input. The output from the maximum pooling is recorded as A_3:

$$A_3 = Max\_pool(A_2) \tag{9}$$

A dropout operation is used on the third layer to randomly reduce a part of its output to avoid overfitting. The fourth layer is a convolutional layer with five kernels, and its kernel size is 2*2 with stride 1. The fifth layer is a maximum pooling layer with the kernel of size 2*2 and stride of 2. The purpose of using these layers is to further extract the information of the input features and improve the accuracy of the prediction. The results A_4 and A_5 of the fourth and the fifth layers are defined as

$$A_4 = tanh(Conv(A_3)) \tag{10}$$

$$A_5 = Max\_pool(A_4) \tag{11}$$

where *tanh(.)* represents the hyperbolic tangent activation function. The last layer is a fully connected output layer with the predicted conditional relatedness $\hat{y}_k$ of sample *k* defined as

$$\hat{y}_k = Sigmoid(W_f^T \cdot A'_5 + b_f) \tag{12}$$

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{13}$$

where A_5' represents the reshaped vector of A_5; W_f and b_f represent the weight vector and the bias of the final layer, respectively. We apply the Binary Cross Entropy loss (BCEloss) as the loss function of FCNN model defined as

$$BCEloss = -[y_k \log(\hat{y}_k) + (1-y_k)\log(1-\hat{y}_k)] \tag{14}$$

where y_k represents ground true label 1/0 of the positive/negative sample *k*, and *K* represents the total number of all samples. The optimal algorithm is RMSPROP (Zhang et al., 2019).

Based on the CNN structure with a fully connected first layer, our model is trained by grid-search 10-fold cross-validation, and the hyper-parameters with the highest AUC value of the whole cross-validation are employed, including kernel size, stride, *etc.* For the detailed description of the architecture and hyper-parameters, see Optimizing the FCNN Model section.

Experimental Design

Herein, our experiment breaks down four parts, depicted as **Figure 2**, in detail. First, gene-pair samples are collected from three databases and a benchmark dataset to compose the dataset for FCNN construction, which contains co-expression and prior-knowledge sub-datasets. Second, 12 gene-pair features are calculated, including seven expression similarities and five prior-knowledge similarities. Third, FCNN is constructed by grid search in a 10-fold cross-validation process. Finally, FCNN is evaluated by comparing with 12 models and methods in 10-fold cross-validation, test, verification, and construction of gene network.

The 12 compared models and methods consist by seven models, including logistic regression (LR), linear discriminant analysis (LDA), SVM, deep belief network (DBN), fully connected neural network (FNN), CNN, and MFR (Wang et al., 2019), as well as five co-expression analysis methods, including PCC, SRC, MI, PPC, and CMI. In these models and methods, LR, LDA, and SVM are traditional machine learning technologies applied in many fields (Zhang et al., 2006; AndrewCucchiara, 2012; Asafu-Adjei et al., 2013).

Specifically, the SVM model is constructed with the radial basis kernel function. DBN is a classical deep learning generation model, which combines restricted Boltzmann machine (Pang et al., 2018) and neural network structure. Multi-Features Relatedness (MFR) is a SVM-based model with linear kernel function proposed recently, integrating both expression and prior-knowledge similarities to measuring conditional relatedness. And PCC, SRC, MI, PPC, and CMI are traditional methods for measuring conditional relatedness between a pair of genes.

For each model and method, we conduct 10-fold cross-validation using 81% samples in dataset collected from the COXPRESdb, KEGG, and TRRUST databases and a benchmark dataset of Pan et al. research (Pan et al., 2010) for training, 9% samples for validation, and the rest 10% for test. And the results of validation and test are used to compare models and methods in terms of precision. Moreover, samples obtained from the GeneFriends and DIP databases are used for further verification to compare different models or methods in robustness. We also compare the practicability of models and methods in terms of cancer gene network construction. To compare the performance of each model or method, we select the receiver operating characteristic curve (ROC) with its AUC (Lobo et al., 2010) and the ACC value as the criteria.

RESULTS

Optimizing the FCNN Model

We built our parameterized FCNN model using Pytorch (Aorte et al., 2019). The optimal hyper-parameters are obtained from various combinations based on baseline parameters by grid search within 10-fold cross-validation. We test hyper-parameter combinations containing the kernel size, stride, learning rate, activation functions, dropout probability, *etc.*, and get the experimental results of the different hyper-parameters shown as **Table 2**. Specially, the FCNN model is trained by minimizing the BCEloss with RMSprop optimizer (Zhao et al., 2019) in the light of the AUC of validation and test datasets. As shown in **Table 2**, the best hyper-parameters for combination of activation function, the kernel size, stride, the number of neurons in the

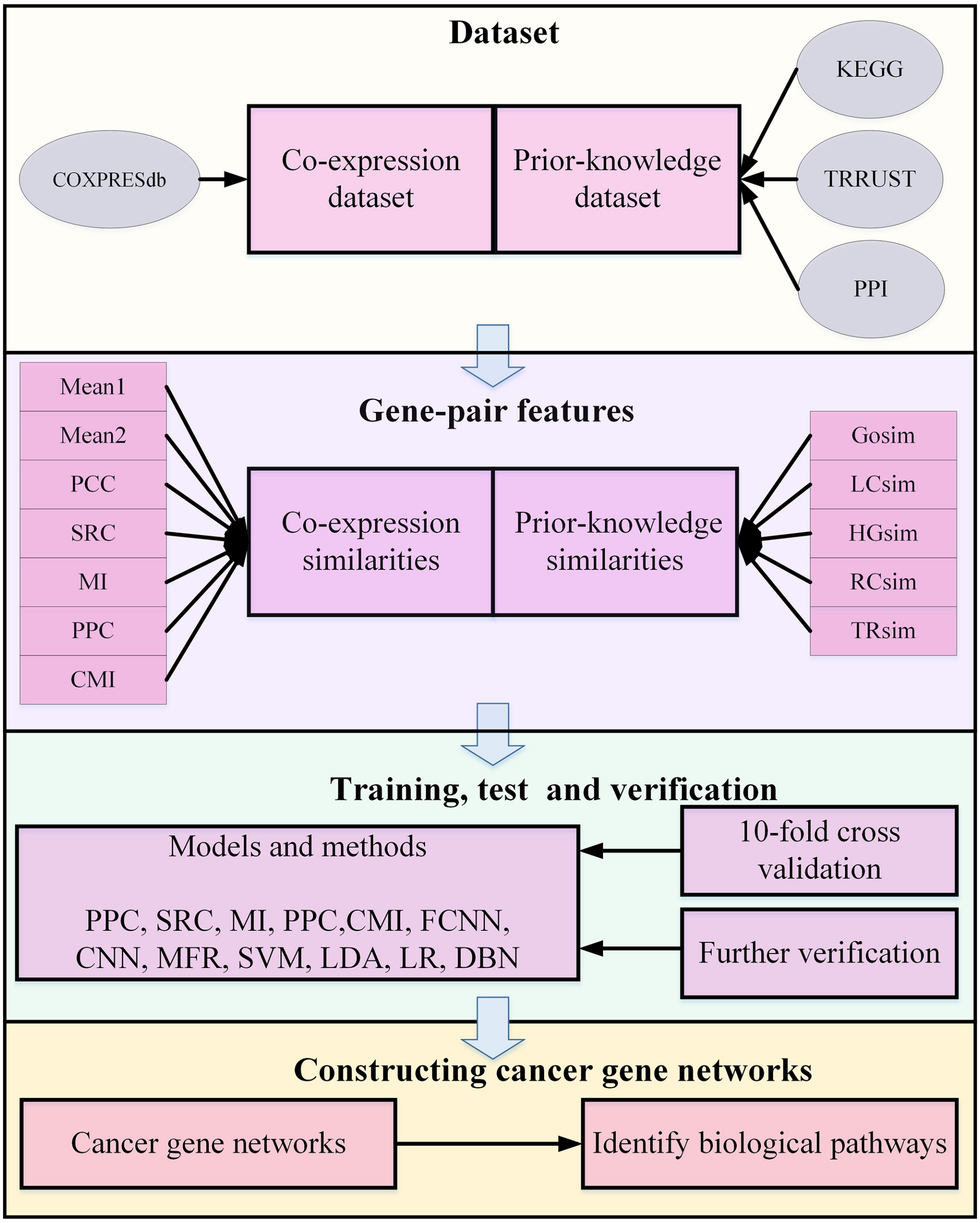


FIGURE 2 | The flowchart of experimental design in biological pathways identification.

first layer, learning rate, the dropout probability, and the batch size is Tanh_Tanh, 2, 1, 81, 0.001, 0.1, and 250, respectively.

Table 2 reflects the experimental results of the combining hyper-parameters. The nine kinds of combination of three activation functions (ReLU, Sigmoid, and Tanh) are evaluated. As a result, combination of Tanh and Tanh is optimal. The kernel size and the stride of the FCNN model are changed to 2 and 3, and 1 and 2, respectively. The kernel of 2 and the stride of 1 are the best suitable for our approach, respectively. The neuron number of the first layer is changed to 5*5, 9*9, and 13*13, and we find 9*9 is optimal. The learning rate is changed to 0.0001, 0.001, and 0.01, and the learning rate of 0.001 shows our approach obtains the best performance in both validation and test AUC. To avoid the overfitting, the dropout probability is applied in our approach, changed to 0.1, 0.2, and 0.3. The dropout probability of 0.1 presents the highest AUC in training and test; meanwhile, the larger the dropout probability, the lower the AUCs on validate and test datasets. And then the batch size for the model is also changed to 200, 250, and 300, which shows that the batch size of 250 gets the best performance. To sum up, the combination of the kernel size of 2, the stride of 1, the neuron number of 81 in the first layer, the learning rate of 0.01, the dropout probability of 0.1, and the batch size of 250 is optimal. And we also list the optimal condition under a single hyper-parameter, based on our experiments.

TABLE 2 | Effects of the varied hyper-parameters through a 10-fold cross-validation in terms of AUC based on the validation and test datasets.

Hyper-parameter	Parameter	Validation	Test
Kernel size	2	**0.8310**	**0.8321**
	3	0.8121	0.8172
Stride	1	**0.8310**	**0.8321**
	2	0.8089	0.8156
Number of neurons	25	0.8191	0.8232
	81	**0.8310**	**0.8321**
	169	0.8189	0.8236
Learning rate	0.01	0.8250	0.8296
	0.001	**0.8310**	**0.8321**
	0.0001	0.7763	0.7802
Dropout probability	0.1	**0.8310**	**0.8321**
	0.2	0.8196	0.8228
	0.3	0.8180	0.8227
Batch size	200	0.8166	0.8231
	250	**0.8310**	**0.8321**
	300	0.8135	0.8209
Activation function	ReLU_ReLU	0.8132	0.8224
	ReLU_Sigmoid	0.8127	0.8210
	ReLU_Tanh	0.8127	0.8242
	Sigmoid_ReLU	0.8224	0.8296
	Sigmoid_ Sigmoid	0.8245	0.8301
	Sigmoid_Tanh	0.8271	0.8308
	Tanh_ReLU	0.8253	0.8297
	Tanh_Sigmoid	0.8245	0.8309
	Tanh_Tanh	**0.8310**	**0.8321**

FCNN model obtains the optimal AUC value, based on the different hyper-parameters combinations.

Comparison With Existing Methods

The best parameters of all models are obtained by grid search within 10-fold cross-validation, and the results of the final models with the best parameters are applied to compare models and methods in terms of precision. As shown in **Figures 3A**, **B**, most machine learning models perform better than the co-expression analysis methods, and our FCNN model has the highest AUC value of 0.831 and ACC of 0.761 than the others. CNN model is better than others except for the FCNN model, with an AUC value of 0.796 and ACC of 0.731, but the DBN model performs the worst among all models and methods. In the light of **Figures 3C–D**, the FCNN model obtains the highest AUC and ACC against all models and methods on the test dataset. And the CNN model yields higher AUC value of 0.799 and ACC value of 0.734, which is better than other models and methods besides the FCNN model.

To test the generalization and robustness of all models and methods on the samples obtained from the GeneFriends and DIP databases, all models applied on this further verification are trained without samples from the GeneFriends and DIP databases. As shown in **Figures 3E–H**, the result on the GeneFriends database reflects the robustness of models and methods in detecting gene–gene interactions from co-expression dataset, and the performance on the DIP database indicates generalization in identifying gene–gene interactions from the prior-knowledge datasets. **Figures 3E–H** shows that FCNN model obtains the third largest AUC value of 0.725 and the highest ACC value of 0.693 among all models and methods on the GeneFriends samples, and AUC and ACC values of FCNN model are better than others on the DIP samples, which are 0.786 and 0.674, respectively.

To clarify the performance of the trained FCNN model on the co-expression, PPI, KEGG, and TRRUST sub-sub datasets, respectively, we applied all models and methods to these four sub-sub datasets and the results shown as **Figure S1**. According to **Figure S1**, our approach achieves the highest AUC of 0.938, 0.578, and 0.532 on the co-expression, PPI, and TRRUST datasets, respectively. For the KEGG dataset, AUC of 0.628 of the FCNN model is a little lower than AUC of 0.63 the CNN model obtained. In light of the above results, it is reasonable that the AUC of FCNN model on the co-expression dataset is higher than on the prior-knowledge dataset, which reflects that our models identify the relationship of genes mainly depending on the co-expression information. And the prior-knowledge information only acts as an auxiliary role in the process of identifying gene relationships. To the best of our knowledge, the co-expression information can powerfully reflect the relatedness of genes in a real cell environment, but possibly contains some error messages. And the prior-knowledge information is invested to relieve these error messages, as the relatedness of gene pairs support by the prior-knowledge messages is global, meaning only a small part of those relatedness is activated on a certain condition. Meanwhile, it also implies our model is not suitable for catching the global relatedness of gene pairs support by the prior knowledge.

Constructing Cancer Gene Networks

Genes act as a vital role in many human diseases, most of which often work with each other and affect human health (Li et al., 2018), and the weighed gene network provides an effective method to study the relationship between genes (Yang et al., 2014). There is a property of gene networks in which the genes involved in related biological processes are connected to each other to compose gene subnetworks with density inside connections and sparse outside connections, *i.e.*, genes in a module should be involved in related biological processes (Matteo et al., 2012). Here, the purpose of measuring conditional relatedness between genes is to detect the probability of these genes jointly involving in a biological process. Therefore, the better conditional relatedness is measured by a model for constructing gene network, the more distinctive such property is. Inspired by the above, we use this property to compare each model or method in the construction of gene networks. The conditional relatedness in this research is utilized to construct cancer gene networks, where nodes indicate genes and weights on edges indicate relatedness. The criterion is the number of metabolic pathways predicted significantly influenced by increased serine metabolism in cancers. We choose reprogramming serine metabolism as it is one of the hallmarks of cancer (Yang and Vousden, 2016). It is reported that serine metabolism is increased in various cancers, especially in bladder cancer (Massari et al., 2016), breast cancer (Locasale et al., 2011; Richard et al., 2011; Kim et al., 2014), colon cancer (Duthie, 2011; Jie et al., 2015; Yoon et al., 2015), and lung cancer (Piskac-Collier et al., 2011; Denicola et al., 2015), and supports several metabolic processes that are crucial for the growth and survival of cancer cells, such as DNA/RNA methylation (Maddocks et al., 2016), glutathione biosynthesis

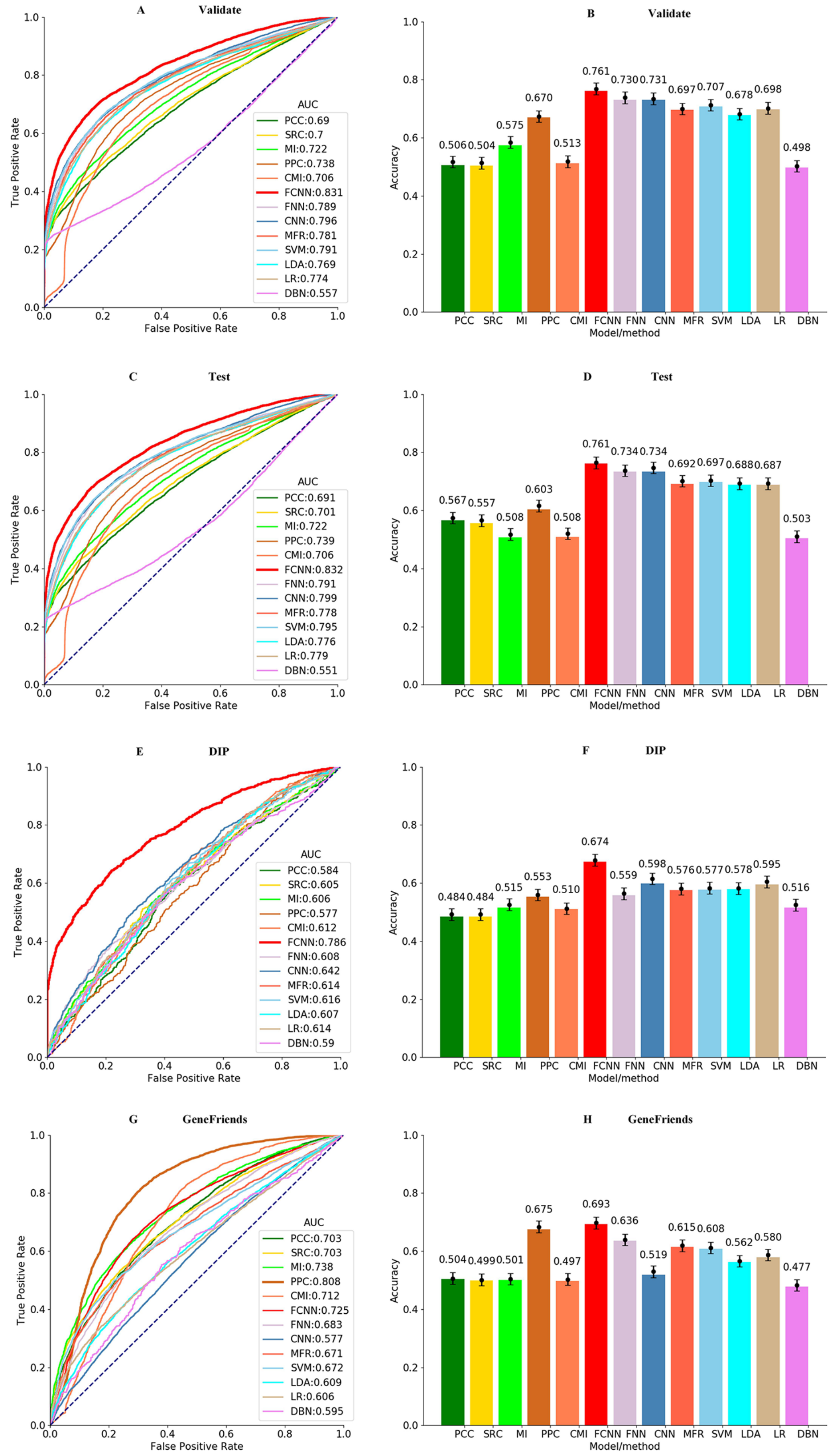


FIGURE 3 | ROCs of all models and methods for identifying gene–gene interactions in the **(A)** validation, **(C)** test, **(E)** DIP, and **(G)** GeneFriends datasets. ACCs of all models and methods for identifying gene–gene interactions in the **(B)** validation, **(D)** test, **(F)** DIP, and **(H)** GeneFriends datasets.

TABLE 3 | The number of samples in cancer and normal tissue.

Caner type	Samples in normal tissue	Samples in cancer
LUAD	515	19
COAD	285	113
BRCA	1095	41
BLCA	408	59

(Amelio et al., 2014), one-carbon metabolism (Yang and Vousden, 2016), *etc.* We conduct enrichment analysis on gene modules identified to be influenced by increased serine metabolism against all the pathways in the KEGG database and obtain significant enriched metabolic pathways (q-value < 0.01) (Storey, 2003). Then we count the number of how many of the significant enriched metabolic pathways are the ones reported to be related to enhanced serine metabolism in cancer tissues. The number shows how well the genes in a module are involved in related biological processes and reflects how well the conditional relatedness is measured by different models for gene network construction.

First, we collect RNA-Seq gene expression data of four cancer types, including bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), and lung adenocarcinoma (LUAD) from the TCGA database (Hampton, 2006), the details of which are shown in **Table 3**. Second, up-regulated genes are identified using Limma t-test (Ritchie et al., 2015), with the fold-change of expression level in cancer *versus* normal tissue > 1.5 and P value < 0.05. Then the relatedness of each pair of up-regulated genes is calculated by FCNN model and 12 other models and methods. Especially, co-expression similarities used as features for each model are calculated using gene expression data in cancers. Third, we construct cancer gene networks, where nodes indicate up-regulated genes, and for each node, we link other nodes with the top 5 relatedness. There are a total of 13*4 gene networks for 13 models and methods in four cancer types. Fourth, we collect 11 enzyme-encoding genes that catalyze biological reactions of serine as the markers for serine metabolism, including *CBS*, *CBSL*, *PTDSS1*, *PTDSS2*, *SDS*, *SDSL*, *SHMT1*, *SHMT2*, *SPTLC1*, *SPTLC2*, *SPTLC3*, and *SRR*. The modules in each network are identified by fast modularity optimization algorithm (Zhang et al., 2009). And the modules with gene markers are defined as modules influenced by increased serine metabolism. We implement gene set enrichment analysis against KEGG pathways on such modules (Christina et al., 2007), by using the hyper-geometric test, with q-value < 0.01. Finally, the metabolic pathways confirmed to be significantly influenced by enhanced serine metabolism in cancer tissues are obtained by intersecting-enriched pathways with the ground truth (see **Supplement Tables 1–4**). As shown in **Figure 4**, we detect 13 significantly influenced pathways in FCNN-based gene network in four cancer types, which is the most among all models and methods.

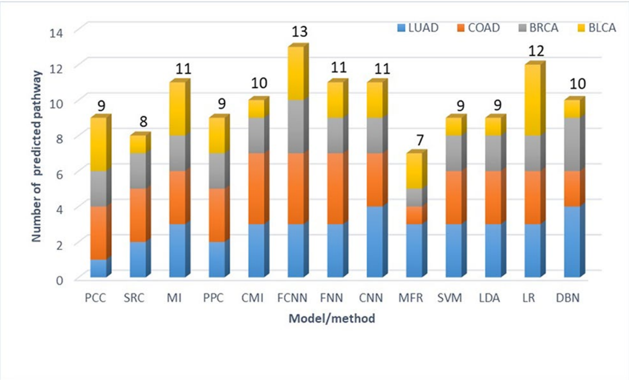


FIGURE 4 | Number of metabolic pathways predicted to be directly influenced by increased serine metabolism in four cancer types.

DISCUSSION

Recent advances in deep learning and bioinformatics stimulate considerable interest in measuring the relatedness of genes, and such pursuit is necessary, which not only speeds up transition from machine learning methods based on measuring correlation to deep learning methods but also can reveal some potential relationship between genes.

Our approach integrates a fully connected layer and the CNN structure for measuring conditional relatedness between genes by integrating co-expression and prior-knowledge similarities. Meanwhile, we demonstrate that this approach is available and effective by experiments on different datasets. To verify our model, we compare the FCNN model with other seven models and five co-expression analysis methods in validation, test, and further verification. The results show that most of machine learning models have higher AUC and ACC values than co-expression analysis methods, implying a combination of both co-expression and prior-knowledge similarities has more obvious advantages in terms of measuring conditional relatedness than using only co-expression similarities. The FCNN model obtains the best performance among machine learning models, which proves deep-learning-based models can more effectively detect the complex map relations between similarities and conditional relatedness than traditional algorithms, such as FNN, MFR, LR, LDA, SVM, and so on. Especially, FCNN model successfully calls a better result than CNN model, which indicates the fully connected first layer persists in our model from ignoring useful combinations of features and the remaining CNN structure with parameter sharing and connection sparsity help our model to be easily trained on the medium-sized dataset. All the above advantages make FCNN model more practical, and as a result, it achieves the best performance in the construction of cancer gene networks. However, PPC and MI obtain higher AUC values on the GeneFriends samples than the FCNN model, mainly because the gene–gene interactions collected from the GeneFriends database are predicted by PCC, making PCC have a natural advantage comparing with other models or methods. And MI has some resemblance with PCC (Yan et al., 2019), which makes it gain the second best result on the GeneFriends dataset.

In line with the performance of the FCNN model, for the next step, we will collect more data, extract more features of gene pairs, and plan to optimize the structure of the model to improve the performance. Meanwhile, we generate some of the negative datasets by random permutation following the way of the references, which may suffer from issue of neglecting tissue specificity; therefore, we will improve this process in our coming researches. Moreover, deep learning is an extremely active research community that is garnering more and more focus from academia, and we expect that deep learning models like this hybrid architecture will be continually explored for the purpose of measuring the relatedness between genes.

CONCLUSION

In conclusion, the FCNN model is a novel deep learning model of CNN with a fully connected first layer, combining co-expression and prior-knowledge similarities to measure conditional relatedness between genes. For benchmarking purposes, we compare the FCNN model to existing models and co-expression analysis methods; our proposed model obtains the best performance of identifying gene–gene interaction invalidation, test, and further verification. Meanwhile, we estimate the performance of all models and methods on the co-expression and prior-knowledge sub-datasets, respectively, which show that the FCNN model is optimal. In terms of constructing gene networks, the FCNN model also outperforms other compared models and methods and achieves more practical results.

DATA AVAILABILITY STATEMENT

The datasets and results of this study, and code of the FCNN model can be freely obtained from https://bmbl.bmi.osumc.edu/FCNN for academic uses and biological analysis.

AUTHOR CONTRIBUTIONS

SZ and YT collected the data and performed the experiments. YW conceived the project. YW and QM designed the study. YT, SZ, LY, and SY wrote the manuscript. All authors read and approved the final manuscript for publication.

FUNDING

This research was funded by the National Natural Science Foundation of China (Nos. 61572227, 61872418) and the Development Project of Jilin Province of China (Nos. 20170203002GX, 20170520063JH, 20180414012GH, 20190201293JC).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01009/full#supplementary-material

REFERENCES

Amelio, I., Markert, E. K., Rufini, A., Antonov, A. V., Sayan, B. S., Tucci, P., et al. (2014). p73 regulates serine biosynthesis in cancer. *Oncogene* 33 (42), 5039–5046. doi: 10.1038/onc.2013.456

Amrine, K. C., Blanco-Ulate, B., and Cantu, D. (2015). Discovery of core biotic stress responsive genes in Arabidopsis by weighted gene co-expression network analysis. *PLoS One* 10 (3), e0118731. doi: 10.1371/journal.pone.0118731

AndrewCucchiara, (2012). Applied logistic regression. *Technometrics* 34 (3), 358–359. doi: 10.2307/1270048

Aorte, F., Dambre, J., Bienstman, P. (2019) Highly parallel simulation and optimization of photonic circuits in time and frequency domain based on the deep-learning framework PyTorch[J]. *Sci. Rep.* 9 (1). doi: 10.1038/s41598-019-42408-2

Asafu-Adjei, J. K., Sampson, A. R., Sweet, R. A., and Lewis, D. A. (2013). Adjusting for matching and covariates in linear discriminant analysis. *Biostatistics* 14 (4), 779–791. doi: 10.1093/biostatistics/kxt017

Babak, A., Andrew, D., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33 (8), 831. doi: 10.1038/nbt.3300

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 (D1), D991–D995. doi: 10.1093/nar/gks1193

Baruch, B., and Albert-László, B. (2013). Network link prediction by global silencing of indirect correlations. *Nat. Biotechnol.* 31 (8), 720–725. doi: 10.1038/nbt.2601

Bovolenta, L. A., Acencio, M. L., and Lemke, N. (2012). HTRIdb: an open-access database for experimentally verified human transcriptional regulation interactions. *BMC Genomics* 13 (1), 405. doi: 10.1186/1471-2164-13-405

Chen, L., and Vitkup, D. (2006). Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol.* 7 (2), R17–R17. doi: 10.1186/gb-2006-7-2-r17

Christina, B., Andreas, K., Jan, K., Benny, K., Nicole, C., Elnakady, Y. A., et al. (2007). GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res.* 35 (Web Server issue), 186–192. doi: 10.1093/nar/gkm323

Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* 32 (suppl_1), D258–D261. doi: 10.1093/nar/gkh036

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., et al. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.* 39 (Database issue), D691. doi: 10.1093/nar/gkq1018

David, C., Antonio Fabregat, M., Robin, H., Marija, M., Joel, W., Guanming, W., et al. (2014). The reactome pathway knowledgebase. *Nucleic Acids Res.* 42 (Database issue), 472–477. doi: 10.1093/nar/gkt1102

de Bono, B., Jassal, B., Birney, E., Schmidt, E., Joshi-Tope, G., Gopinath, G. R., et al. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* 33 (suppl_1), D428–D432. doi: 10.1093/nar/gki072

De, L. U., Jensen, L. J., Fausbøll, A., Jensen, T. S., Bork, P., and Brunak, S. (2005). Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics* 21 (7), 1164–1171. doi: 10.1093/bioinformatics/bti093

Denicola, G. M., Chen, P. H., Mullarky, E., Sudderth, J. A., Hu, Z., Wu, D., et al. (2015). NRF2 regulates serine biosynthesis in non-small cell lung cancer. *Nat. Genet.* 47 (12), 1475. doi: 10.1038/ng.3421

Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *J. Bus. Econ. Stat.* 13 (1), 134–144. doi: 10.1198/073500102753410444

Duthie, S. J. (2011). Folate and cancer: how DNA damage, repair and methylation impact on colon carcinogenesis. *J. Inherit. Metab. Dis.* 34 (1), 101–109. doi: 10.1007/s10545-010-9128-0

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci.* 95 (25), 14863–14868. doi: 10.1073/pnas.95.25.14863

Aorte, F., Dambre, J., Bienstman, P. (2019). Highly parallel simulation and optimization of photonic circuits in time and frequency domain based on the deep-learning framework PyTorch. *Scientific Reports* 9 (1). doi: 10.1038/s41598-019-42408-2

Gao, X., Wei, Z., and Hakonarson, H. (2019). tRNA-DL: a deep learning approach to improve tRNAscan-SE prediction results. *Hum. Heredit.* 83, 163–172. doi: 10.1159/000493215

Hampton, T. (2006). Cancer genome atlas. *JAMA* 296 (16), 1958–1958. doi: 10.1001/jama.296.16.1958-d

Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D., et al. (2017). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* 46 (Database issue), D380–D386. doi: 10.1093/nar/gkx1013

Han, J., and Puri, R. K. (2018). Analysis of the cancer genome atlas (TCGA) database identifies an inverse relationship between interleukin-13 receptor α1 and α2 gene expression and poor prognosis and drug resistance in subjects with glioblastoma multiforme. *J. Neurooncol.* 136 (3), 463–474. doi: 10.1007/s11060-017-2680-9

Jelier, R., Jenster, G., Dorssers, L. C. J., van der Eijk, C. C., van Mulligen, E. M., Mons, B., et al. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics* 21 (9), 2049–2058. doi: 10.1093/bioinformatics/bti268

Jie, C., Shan, W., Dali, H., Wei, D., Chao, X., and Hongliang, G. (2015). MicroRNA-455 inhibits proliferation and invasion of colorectal cancer by targeting RAF proto-oncogene serine/threonine-protein kinase. *Tumour Biol.* 36 (2), 1313–1321. doi: 10.1007/s13277-014-2766-3

Jing, L., and Ng, M. K. (2010). Prior knowledge based mining functional modules from Yeast PPI networks with gene ontology. *BMC Bioinformatics* 11 (11), S3. doi: 10.1186/1471-2105-11-S11-S3

Kanehisa, M. (2002). The KEGG database. *Novartis Found. Symp.* 247 (247), 91–103. doi: 10.1002/0470857897.ch8

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi: 10.1093/nar/28.1.27

Kim, D. C., Wang, X., Yang, C. R., and Gao, J. (2010). Learning biological network using mutual information and conditional independence. *Bmc Bioinformatics* 11 (Suppl 3), S9–S9. doi: 10.1186/1471-2105-11-S3-S9

Kim, S. K., Jung, W. H., and Koo, J. S. (2014). Differential expression of enzymes associated with serine/glycine metabolism in different breast cancer subtypes. *Plos One* 9 (6), e101004. doi: 10.1371/journal.pone.0101004

Kumari, S., Nie, J., Chen, H.-S., Ma, H., Stewart, R., Li, X., et al. (2012). Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PloS One* 7 (11), e50411. doi: 10.1371/journal.pone.0050411

Li, J., Zhou, D., Qiu, W., Shi, Y., Yang, J.-J., Chen, S., et al. (2018). Application of weighted gene co-expression network analysis for data from paired design. *Sci. Rep.* 8 (1), 622. doi: 10.1038/s41598-017-18705-z

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2010). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecol. Biogeogr.* 17 (2), 145–151. doi: 10.1111/j.1466-8238.2007.00358.x

Locasale, J. W., Grassian, A. R., Tamar, M., Lyssiotis, C. A., Mattaini, K. R., Bass, A. J., et al. (2011). Phosphoglycerate dehydrogenase diverts glycolytic flux and contributes to oncogenesis. *Nat. Genet.* 43 (9), 869–874. doi: 10.1038/ng.890

Maddocks, O. D., Labuschagne, C. F., Adams, P. D., and Vousden, K. H. (2016). Serine metabolism supports the methionine cycle and DNA/RNA methylation through *de novo* ATP synthesis in cancer cells. *Mol. Cell* 61 (2), 1–12. doi: 10.1016/j.molcel.2015.12.014

Massari, F., Ciccarese, C., Santoni, M., Iacovelli, R., Mazzucchelli, R., Piva, F., et al. (2016). Metabolic phenotype of bladder cancer. *Cancer Treat. Rev.* 45, 46–57. doi: 10.1016/j.ctrv.2016.03.005

Matteo, D. A., Vera, P., Shruti, S., and Ciccarelli, F. D. (2012). Network of cancer genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Res.* 40 (Database issue), D978–D983. doi: 10.1093/nar/gkr952

Mistry, M., and Pavlidis, P. (2008). Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics* 9 (1), 327. doi: 10.1186/1471-2105-9-327

Mohammadi, A., Saraee, M. H., and Salehi, M. (2011). Identification of disease-causing genes using microarray data mining and Gene Ontology. *BMC Med. Genomics* 4 (1), 12. doi: 10.1186/1755-8794-4-12

Nagafuchi, A., Takeichi, M., and Tsukita, S. (1991). The 102 kd cadherin-associated protein: Similarity to vinculin and posttranscriptional regulation of expression. *Cell* 55 (5), 849–857. doi: 10.1016/0092-8674(91)90392-C

Nakamura, T., Furukawa, Y., Nakagawa, H., Tsunoda, T., Ohigashi, H., Murata, K., et al. (2004). Genome-wide cDNA microarray analysis of gene expression profiles in pancreatic cancers using populations of tumor cells and normal ductal epithelial cells selected for purity by laser microdissection. *Oncogene* 23 (13), 2385–2400. doi: 10.1038/sj.onc.1207392

Obayashi, T., and Kinoshita, K. (2009). Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res.* 16 (5), 249–260. doi: 10.1093/dnares/dsp016

Pan, X., and Shen, H. B. (2017). RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *Bmc Bioinformatics* 18 (1), 136. doi: 10.1186/s12859-017-1561-8

Pan, X. Y., Zhang, Y. N., and Shen, H. B. (2010). Large-scale prediction of human protein–protein interactions from amino acid sequence based on latent topic features. *J. Proteome Res.* 9 (10), 4992–5001. doi: 10.1021/pr100618t

Pang, S., del Coz, J. J., Yu, Z., Luaces, O., and Diez, J., et al. (2018). Deep Learning and Preference Learning for Object Tracking: *A Combined Approach* 47 (3), 859–876. doi: 10.1007/s11063-017-9720-5

Piskac-Collier, A. L., Claudia, M., Lopez, M. S., Andrea, C., Etzel, C. J., Greisinger, A. J., et al. (2011). Variants in folate pathway genes as modulators of genetic instability and lung cancer risk. *Genes Chromosomes Cancer* 50 (1), 1–12. doi: 10.1002/gcc.20826

Poliakov, E., Managadze, D., and Rogozin, I. B. (2014). Generalized Portrait of cancer metabolic pathways inferred from a list of genes overexpressed in cancer. *Genet. Res. Int.* 2014 (4), 646193. doi: 10.1155/2014/646193

Richard, P., Marks, K. M., Shaul, Y. D., Pacold, M. E., Dohoon, K., Kivanç, B., et al. (2011). Functional genomics reveal that the serine synthesis pathway is essential in breast cancer. *Nature* 476 (7360), 346–350. doi: 10.1038/nature10350

Ritchie, M. E., Smyth, G. K., Phipson, B., Wu, D., Hu, Y., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43 (7), e47–e47. doi: 10.1093/nar/gkv007

Sipko, V. D., Thomas, C., and Jo O Pedro, D. M. E. (2015). GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic Acids Res.* 43 (Database issue), 1124–1132. doi: 10.1093/nar/gku1042

Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* 13 (1), 328–328. doi: 10.1186/1471-2105-13-328

Storey, John D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* 31 (6), 2013–2035. doi: 10.1214/aos/1074290335

Trebeschi, S., Griethuysen, J. J. M. V., Lambregts, D. M. J., Lahaye, M. J., Parmar, C., Bakers, F. C. H., et al. (2017). Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci. Rep.* 7 (1), 5301. doi: 10.1038/s41598-017-05728-9

Wang, J. Z., Zhidian, D., Rapeeporn, P., Yu, P. S., and Chin-Fu, C. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics* 23 (10), 1274–1281. doi: 10.1093/bioinformatics/btm087

Wang, Y., Yang, S., Zhao, J., Du, W., Liang, Y., Wang, C., et al. (2019). Using machine learning to measure relatedness between genes: a multi-features model. *Sci. Rep.* 9 (1), 4192. doi: 10.1038/s41598-019-40780-7

Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. DIP: the database of interacting proteins. *Nucleic Acids Res.* 28 (1), 289–291. doi: 10.1093/nar/28.1.289

Xiao-Yong, P., Ya-Nan, Z., and Hong-Bin, S. (2010). Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *J. Proteome Res.* 9 (10), 4992–5001. doi: 10.1021/pr100618t

Yan, W., Sen, Y., Jing, Z., Wei, D., Yanchun, L., Cankun, W., et al. (2019). Using machine learning to measure relatedness between genes: a Multi-Features Model. *Sci. Rep.* 9 (1), 4192. doi: 10.1038/s41598-019-40780-7

Yang, M., and Vousden, K. H. (2016). Serine and one-carbon metabolism in cancer. *Nat. Rev. Cancer* 16 (10), 650. doi: 10.1038/nrc.2016.81

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

Yasunobu, O., Yuichi, A., Takeshi, O., Shu, T., Satoshi, I., Takafumi, N., et al. (2015). COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.* 43 (Database issue), D82. doi: 10.1093/nar/gku1163

Yoon, S., Kim, J. G., Seo, A. N., Park, S. Y., Kim, H. J., Park, J. S., et al. (2015). Clinical Implication of Serine Metabolism-Associated Enzymes in Colon Cancer. *Oncology* 89 (6), 351. doi: 10.1159/000439571

Yu, C. S., Chen, Y. C., and Hwang, J. K. (2010). Prediction of protein subcellular localization. *Proteins* 64 (3), 643–651. doi: 10.1002/prot.21018

Zdobnov, E. M., Tegenfeldt, F., Kuznetsov, D., Waterhouse, R. M., Simao, F. A., Ioannidis, P., et al. (2016). OrthoDB v9. 1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* 45 (D1), D744–D749. doi: 10.1093/nar/gkw1119

Zhang, X., Lu, X., Shi, Q., Xu, X. Q., Leung, H. C. E., Harris, L. N., et al. (2006). Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* 7 (1), 1–13. doi: 10.1186/1471-2105-7-197

Zhang, X. S., Wang, R. S., Wang, Y., Wang, J., Qiu, Y., Wang, L., et al. (2009). Modularity optimization in community detection of complex networks. *Epl* 87 (3), 38002. doi: 10.1209/0295-5075/87/38002

Zhang, W. Y., Gu, H., Liu, C., Hong, S., Xu, W., Yang, J., et al. (2019). Convolutional Neural Network Based Models for Improving Super-Resolution Imaging. *IEEE Access*, 7, 43042–43051. doi: 10.1109/ACCESS.2019.2908501

Zhao, J., Gao, Y., Yang, Z., Li, J., Feng, Y., Qin, Z., et al. (2019). Truck Traffic Speed Prediction under Nonrecurrent Congestion: Based on OptimizedDeep Learning Algorithms and GPS Data. *IEEE Access*, 1 -1. doi: 10.1109/ACCESS.2018.2890414

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: 25 October 2019
doi: 10.3389/fgene.2019.01055

Identification of Dysregulated Competitive Endogenous RNA Networks Driven by Copy Number Variations in Malignant Gliomas

Jinyuan Xu[†]*, Xiaobo Hou*[†]*, Lin Pang*[†]*, Shangqin Sun, Shengyuan He, Yiran Yang, Kun Liu, Linfu Xu, Wenkang Yin, Chaohan Xu** *and Yun Xiao**

College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:
Angelo Facchiano,
Institute of Food Sciences,
National Research Council (CNR-ISA),
Italy

Reviewed by:
Qinghua Jiang,
Harbin Institute of Technology,
China
Joseph Glessner,
Children's Hospital of Philadelphia,
United States

***Correspondence:**
Chaohan Xu
chaohanxu@hrbmu.edu.cn
Yun Xiao
xiaoyun@ems.hrbmu.edu.cn

[†]*These authors have contributed equally to this work and share first authorship*

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

Received: *23 May 2019*
Accepted: *01 October 2019*
Published: *25 October 2019*

Citation:
Xu J, Hou X, Pang L, Sun S, He S, Yang Y, Liu K, Xu L, Yin W, Xu C and Xiao Y (2019) Identification of Dysregulated Competitive Endogenous RNA Networks Driven by Copy Number Variations in Malignant Gliomas. Front. Genet. 10:1055. doi: 10.3389/fgene.2019.01055

Gliomas represent 80% of malignant brain tumors. Because of the high heterogeneity, the oncogenic mechanisms in gliomas are still unclear. In this study, we developed a new approach to identify dysregulated competitive endogenous RNA (ceRNA) interactions driven by copy number variation (CNV) in both lower-grade glioma (LGG) and glioblastoma multiforme (GBM). By analyzing genome and transcriptome data from The Cancer Genome Atlas (TCGA), we first found out the protein coding genes and long non-coding RNAs (lncRNAs) significantly affected by CNVs and further determined CNV-driven dysregulated ceRNA interactions by a customized pipeline. We obtained 13,776 CNV-driven dysregulated ceRNA pairs (including 3,954 mRNAs and 306 lncRNAs) in LGG and 262 pairs (including 221 mRNAs and 11 lncRNAs) in GBM, respectively. Our results showed that most of the ceRNA interactions were weakened by CNVs in both LGG and GBM, and many CNV-driven genes shared the same ceRNAs in the dysregulated ceRNA networks. Functional analysis indicated that the CNV-driven ceRNA network involved in some important mechanisms of tumorigenesis, such as cell cycle, p53 signaling pathway and TGF-beta signaling pathway. Further investigation of the ceRNA pairs in the communities from the dysregulated ceRNA network revealed more detailed biological functions related to the oncogenesis of malignant gliomas. Moreover, by exploring the association of CNV-driven ceRNAs with prognosis and histological subtype, we found that the copy number status of MTAP, KLHL9, and ELAVL2 related to the overall survival in LGG and showed high correlation with histological subtype. In conclusion, this study provided new insight into the molecular mechanisms and clinical biomarkers in gliomas.

Keywords: gliomas, CNV, ceRNA, lncRNA, prognosis

INTRODUCTION

Malignant gliomas are the most common aggressive primary brain tumor (Schwartzbaum et al., 2006; Ostrom et al., 2014). As the most aggressive malignant glioma, glioblastoma multiforme (GBM, WHO grade IV) shows a 5-year survival rate of 5% with the median survival time of 14 months from diagnosis (Parsons et al., 2008). Comparing to GBM, gliomas of WHO (World Health

Organization) grade II and III are less aggressive and have been grouped together by The Cancer Genome Atlas (TCGA) as lower grade gliomas (LGGs). Recently, high-throughput studies have proven that copy number variations (CNVs), which are gains or deletions of genomic segments, are considered important risk factors for human cancers (Xi et al., 2011; Park et al., 2017). CNVs are prominent influential factors for gene expression, which may impact the activities of a variety of oncogenic or tumor suppressive pathways (Liang et al., 2016). Many studies have analyzed the impact of CNVs on gene expression phenotypes. For example, Jornsten et al. combined mRNA regulatory relationships with CNV profiles to construct a CNA-driven network using lasso regression and identified driver copy number alterations (CNAs) and explored their effects on transcription in GBM (Jornsten et al., 2011). Park et al. applied a correlation measure to identify significant relationships between copy number variation regions and mRNAs, and characterized the impact of genotypic variations on phenotype in a genome-wide scale (Park et al., 2012). In fact, DNA CNVs not only influenced the expression of protein-coding genes but also affected the expression levels of long non-coding RNAs and miRNAs (Liang et al., 2016).

Recent studies suggested a new layer of miRNA-mediated regulation that RNAs targeted by the common miRNA could "compete" for the miRNAs and thus indirectly regulate each other (Salmena et al., 2011). Such RNAs are called competing endogenous RNAs (ceRNAs), and their miRNA-mediated interactions are referred to as ceRNA interactions. In addition, examples have been already emerging of non-coding RNAs as ceRNAs, such as lincRNA-p21 (Yoon et al., 2012), lincMD1 (Cesana et al., 2011) and linc-RoR (Wang et al., 2013). Experimental evidence has suggested that the aberration of ceRNA interaction can play important roles in tumorigenesis (Tay et al., 2011). Thus, exploring this novel RNA crosstalk will enhance our insight into gene regulatory networks and contribute to a better understanding of human disease (Tay et al., 2014). The existence and strength of ceRNA interactions may vary significantly in different physiological and cellular conditions (e.g., copy number variation). Most ceRNA studies only considered interactions among ceRNAs and miRNAs while overlooking other important gene regulators, such as transcription factors, DNA methylation, and copy number alteration, which would impede our understanding of ceRNA interactions in cancer (Do and Bozdag, 2018). Therefore, incorporating other types of gene expression regulatory factors, namely copy number alteration, to infer condition-specific dysregulated ceRNA interactions in cancer will be meaningful.

Here, we aimed to discover dysregulated ceRNA interactions driven by CNVs in LGG and GBM. We first got the copy number status of each gene and identified over one hundred protein-coding genes and lncRNAs whose expression levels were significantly affected by CNVs in LGG and GBM. Using a customized program, we identified dysregulated ceRNA interactions driven by CNVs and found some interesting features of the dysregulated ceRNA network. Moreover, by systematically characterizing the functions of the CNV-driven ceRNAs, we found their associations with prognosis and histological subtypes.

MATERIALS AND METHODS

Data Source

The DNA copy number (SNP 6.0), mRNA, and miRNA expression data for the LGG and GBM cohorts were collected from the TCGA data portal (https://tcga-data.nci.nih.gov/tcga), and the lncRNA expression data were derived from TANRIC (Li et al., 2015). We extracted 435 LGG and 152 GBM samples with sample-matching copy number data and gene expression data. For DNA copy number data, we determined five types of discretized copy number calls (−2, −1, 0, 1, 2) for genes in LGG and GBM by GISTIC2.0 (Mermel et al., 2011), and genes with no CNV in more than 10% samples were excluded. The gene expression profiles were normalized by log2(tpm+1) and genes with mean expression lower than 30% of samples or with missing values in more than 10% of samples were filtered.

Identification of CNV-Driven Protein-Coding Genes and lncRNAs

To reduce the influence of noise, we retained high-level amplifications and homozygous deletions discretized by GISTIC2.0 and used the binomial test on the genes that co-existed 2 and −2 status, in which the copy number status with smaller sample size was considered as noise and the copy number status were set to 0 ($P < 0.05$) or deleted ($P \geq 0.05$). Then, for each protein-coding gene or lncRNA, we divided the gene expression data by copy number status and performed the rank-sum test on the two groups. Genes with concordant changes in copy number status and gene expression were considered to be CNV-driven genes ($P < 0.05$, **Supplementary Table 1**).

Identification of Dysregulated ceRNA–ceRNA Interactions Driven by CNV

We developed a computational approach to identify dysregulated ceRNA–ceRNA interactions driven by CNVs (**Supplementary Figure 1**). It consisted of the following steps: (1) Obtaining ceRNA–ceRNA interactions in each copy number state. The interactions of mRNA–miRNA and lncRNA–miRNA were obtained from one confidential online miRNA-target databases: StarBase v2.0 (Li et al., 2014). Using the expression profiles of mRNA, lncRNA, and miRNA in each copy number status (i.e. amplification, deletion, and normal), we calculated Pearson correlation coefficient (PCC) between ceRNA pairs as well as mRNA/lncRNA (ceRNA) and miRNA to measure their expression correlations. The ceRNA pairs with significantly positive correlations (adjusted p-value < 0.05) in which each miRNA-ceRNA interaction showed a significantly negative correlation (adjusted p-value < 0.05) were considered as candidate ceRNA triplets in the status. (2) Calculating difference of ceRNA regulation between copy number status. We assumed that the dysregulation caused by CNV will be reflected in the correlations between ceRNA interactions in each candidate ceRNA triplet. So we compared the correlations of ceRNAs in amplification/deletion samples with normal samples to determine the extent of dysregulation. The extent of dysregulation was defined as:

$$\Delta\text{R} = \left| cor_v(\text{ceRNA1}, \text{ceRNA2}) - cor_n(\text{ceRNA1, ceRNA2}) \right|$$

where cor_v(ceRNA1, ceRNA2) was the PCC estimated from the amplification/deletion samples and cor_n(ceRNA1, ceRNA2) was from normal samples. If a candidate ceRNA triplet existed only in one copy number status, ΔR was also calculated by using the correlation filtered before. (3) Identifying CNV-driven dysregulated ceRNA–ceRNA interactions. To determine whether ΔR was statistically significant, a permutation test was performed. We randomized the labels of copy number status 1000 times and recalculated the changes of correlation coefficients of each ceRNA pair. A P value of 0.05 was used as the cut-off to obtain significantly dysregulated pairs, which were regarded as CNV-driven dysregulated ceRNA–ceRNA interactions. R scripts were available on GitHub (https://github.com/EmeraldG1996/orange-juice/tree/master/ceRNA-interaction).

FUNCTIONAL ENRICHMENT ANALYSIS

For functional enrichment analysis, we first obtained gene expression profiles of LGG/GBM and matched normal samples from TCGA, and calculated the differential expression of genes. Based on the fold change values, we performed gene set enrichment analysis (GSEA) to discover functions kyoto encyclopedia of genes and genomes (KEGG pathways and GO terms) altered in LGG and GBM, respectively. Then, the hypergeometric test was used to further identify what cancer-related functions the ceRNA network (or community) participated in:

$$p = 1 - \sum_{k=0}^{m} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

where N was the number of genes in the gene expression profiles, n was the number of given genes involved in dysregulated ceRNA network or specific community, M was the number of genes that participated in cancer-related KEGG pathway/GO term.

Statistical Analysis of Clinical Data

We downloaded the clinical data of 432 LGG and 124 GBM patients from cBioPortal (http://www.cbioportal.org/). Overall survival curves were constructed by Kaplan–Meier estimation and log-rank tests ($P < 0.05$) were used to identify the significantly survival-related copy number changes. The Cox proportional-hazards regression model was used to investigate the association between the expression of genes and OS. Fisher exact test was performed to detect the clinicopathologic correlates with copy number variations.

RESULTS

Identifying DNA Copy Number Variations in LGG and GBM

To systematically evaluate the copy number variations (CNVs) in LGG and GBM, we performed GISTIC2 on TCGA SNP 6.0 array data to get the copy number status of each gene. After filtering segments with copy ratios less than 0.1, 85 putative CNVs in LGG and 65 in GBM were detected, including a total of 152 and 435 patients, respectively. We divided the identified CNVs into two types, i.e., amplification or deletion, for further analysis (**Table 1**, see *Materials and Methods*). Focal amplifications of pathogenic oncogenes were seen in most of the GBM patients. For example, the amplification of PDGFRA was found in 23 patients, and 71 and 28 patients showed EGFR and CDK4 amplification, respectively. We also found some patients harbored focal deletions of tumor suppressor genes, such as CDKN2A (89) and CDKN2B (84). The amplification of oncogenes across LGG was not as extensive as GBM, but focal deletions of CDKN2A/B were also found in LGG, which were considered as negative cell cycle regulators in gliomas (Simon et al., 1999).

Different Copy Number Status Affected the Expression of Protein-Coding Genes and lncRNAs

To identify protein-coding genes and lncRNAs affected by CNVs, we combined copy number data and expression profiles in LGG and GBM. Based on the rank-sum test, we identified genes whose copy number changes (between different copy number statuses) were concordant with changes in their expression (P value < 0.05, see Materials and Methods, **Supplementary Table 1**). In LGG, the expression of 52 protein-coding genes and 2 lncRNAs were significantly affected by CNVs, including 46 protein-coding

TABLE 1 | Characterization of genomic CNVs detected in LGG and GBM.

Rank	Genomic Location	Size	No. of Genes	Candidate Gene(s)
Amplifications (LGG)				
1	7q32.1	chr7:128,577,665–148,118,090	17	
2	8q24.3	chr8:143,692,404–143,696,833	51	MAF1
Deletions (LGG)				
1	9p21.3	chr9:21,329,669–23,898,052	14	CDKN2A,CDKN2B, MTAP,ELAVL2
Amplifications (GBM)				
1	4q12	chr4:54,009,788–54,740,715	7	PDGFRA,KIT
2	7p11.2	chr7:53,926,675–57,139,864	17	EGFR,VOPP1
3	12q13.3-14.1	chr12:57,520,417–57,957,269	21	CDK4
Deletions (GBM)				
1	9p21.2-21.3	chr9:20,341,664–25,784,562	19	CDKN2A,CDKN2B, MTAP,ELAVL2
2	10q23.31	chr10:87,866,672–87,971,930	1	PTEN

genes and 1 lncRNA showing amplification, and 6 mRNAs and 1 lncRNA associated with deletions. While in GBM, 47 protein-coding genes and lncRNAs were significantly associated with copy number status, including 36 protein-coding genes associated with amplifications, and 9 protein-coding genes and 2 lncRNAs associated with deletions. While our CNV-driven genes were identified between amplification/deletion copy number states and normal state, only several genes were confirmed in previous studies, for example, ELAVL2 in GBM (Bhargava et al., 2017). The genomic localization of these genes showed that the CNVs which significantly affected expression in LGG and GBM could be divided into three and five patterns, respectively (**Figures 1A, B**). In GBM, the CNVs concentrated in 10q23.31 (1), 9p21.2-21.3 (10), 4q12 (5), 12q13.3-14.1 (19), and 7p11.2 (12), consistent with previous reports (Crespo et al., 2012). In these regions, the CNVs of some genes were observed in most patients, including many genes that have been confirmed to be important in the occurrence and development of GBM, such as EGFR, CDKN2A/CDKN2B, and MTAP (Lopez-Gines et al., 2010; Feng et al., 2012; Xu et al., 2017). It has been reported that the deletion of 9p21.3 is related to the occurrence of GBM (Inoue et al., 2004; Alentorn et al., 2015). For LGG, the CNVs concentrated in 9p21.2-21.3 (7), 4q12 (39), 12q13.3-14.1 (8) (**Figure 1A**). Several genes in these regions have been suggested to be important for the prognosis. For example, CDKN2A is an independent predictor of poor survival in diffuse lower-grade gliomas (Aoki et al., 2018).

The expression levels of genes identified as copy number deletion (amplification) were generally decreased (increased) in both LGG and GBM (**Figure 1**), which was consistent with

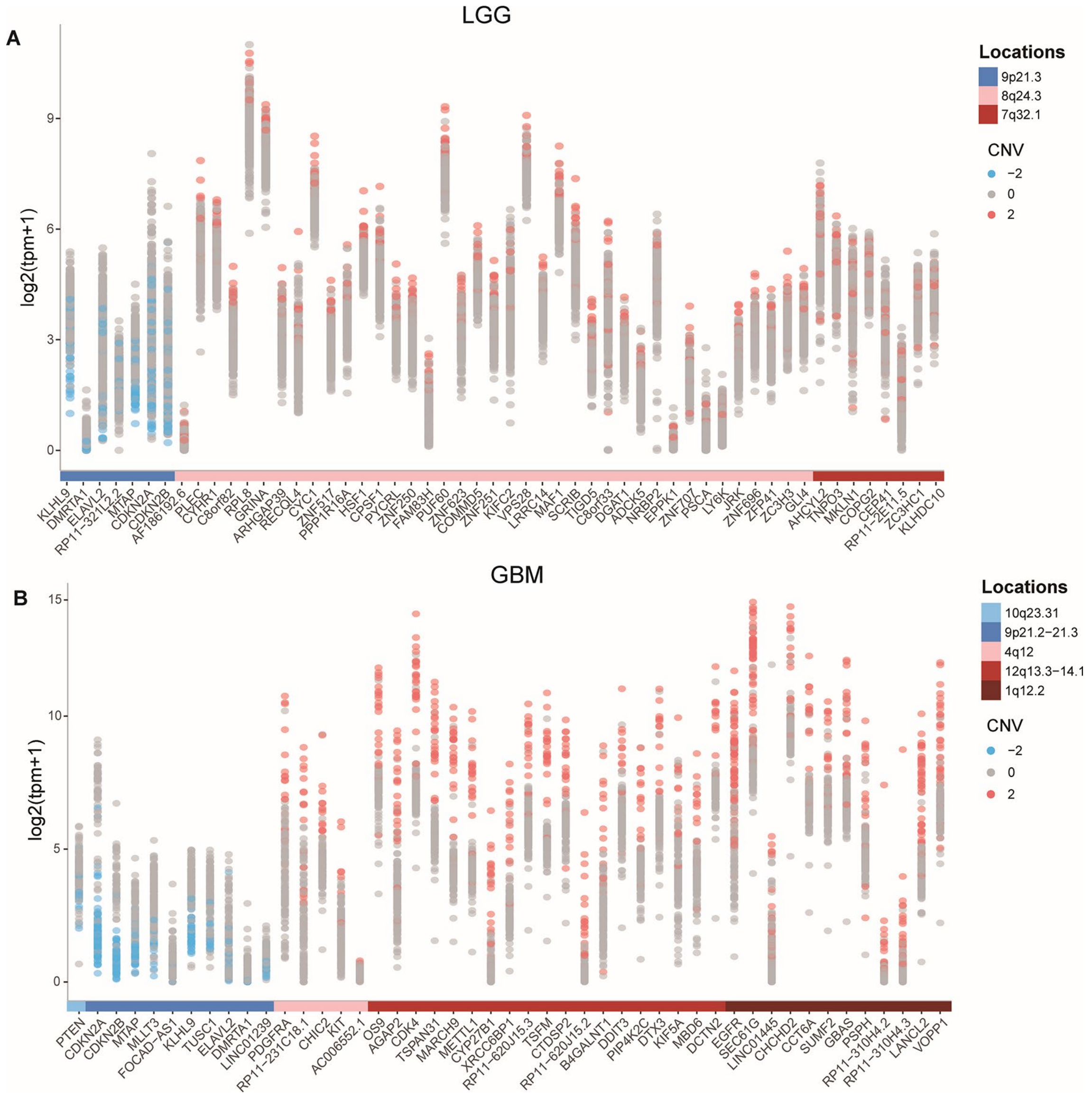


FIGURE 1 | CNVs and the expression of affected protein-coding genes and lncRNAs in GBM and LGG. **(A-B)** Expression of CNV affected genes in GBM **(A)** and in LGG **(B)**.

previous reports (Momtaz et al., 2018). At the same time, we found that the degree of expression changes of different genes within one genomic region was not the same. For example, in GBM, the expression of DMRTA1 and LINC01239, which located in the 9p21.3 region, differed by 10 times when copy number changes.

Identification of the Dysregulated ceRNA Network Driven by CNV

Given the lack of exploration of regulatory factors in existing ceRNA studies, we designed a program to identify dysregulated ceRNA interactions driven by CNV (**Supplementary Figure 1**). The program could be roughly divided into three steps. First, the candidate ceRNA triplets were obtained based on the interactions of mRNA/lncRNA-miRNA in LGG and GBM, respectively. Then, to get ceRNA pairs driven by CNV, we calculated the changes of the correlations of ceRNA pairs in each copy number state (amplification/normal or deletion/normal). If CNV increased the correlation, the ceRNA pair was enhanced by CNV. Conversely, the ceRNA pair was weakened by CNV. Last, we used perturbation test to get significant ceRNA pairs driven by CNV (see Materials and Methods, **Supplementary Table 2**). Through the above three steps, we finally obtained 13776 CNV-driven dysregulated ceRNA pairs in LGG, including 3954 mRNAs and 306 lncRNAs. In GBM, we gained 262 copy number-driven dysregulated ceRNA pairs, including 221 mRNAs and 11 lncRNAs (**Figures 2A**, **B**, **Table 2**).

Next, to gain insights into the dysregulated ceRNA interactions caused by CNV, we visualized the ceRNA network with Cytoscape 3.7.0 (Shannon et al., 2003) (**Figure 2C**). By observing the ceRNA network of GBM, we found most of the ceRNA interactions were

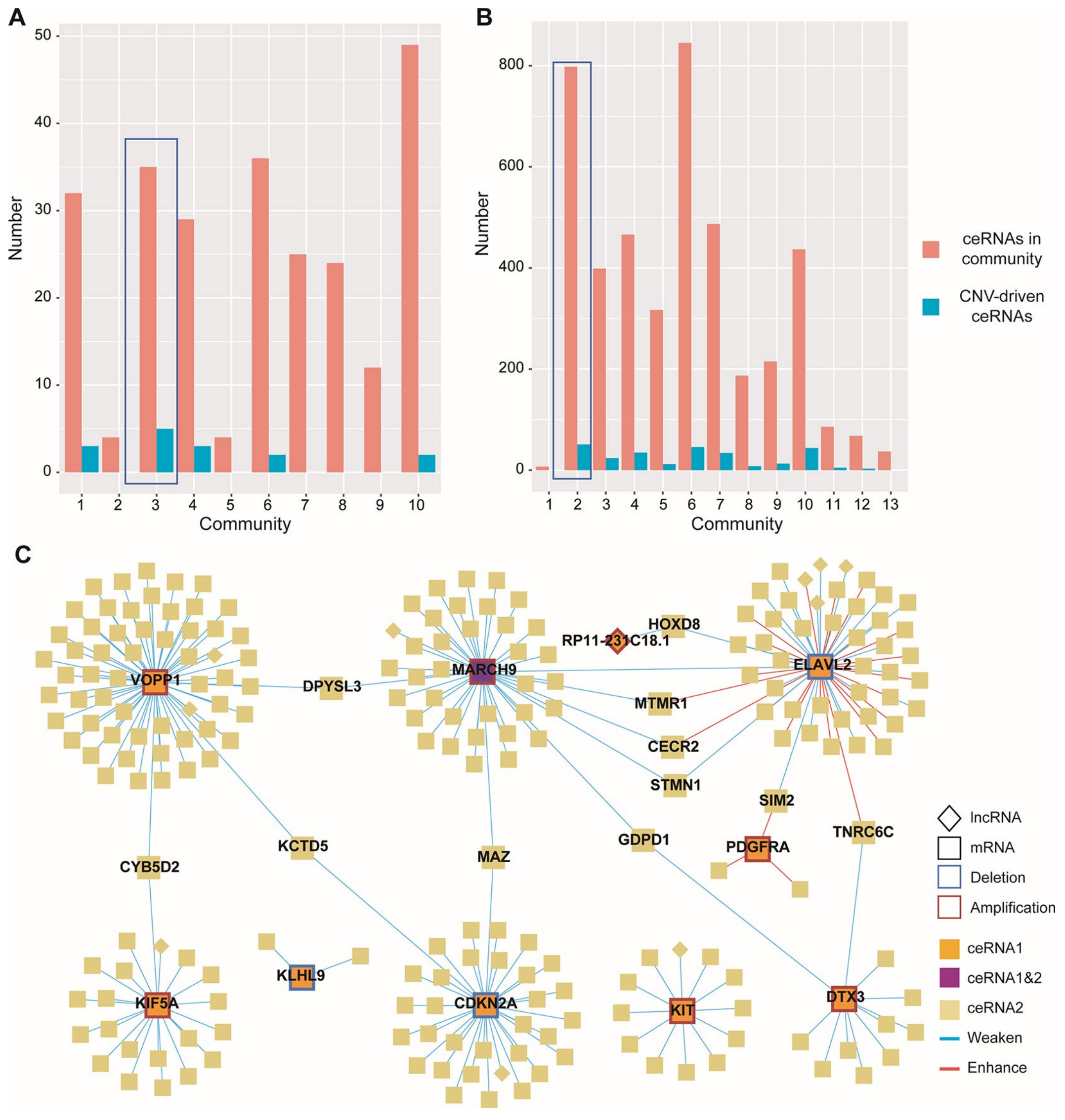


FIGURE 2 | Dysregulated ceRNA pairs driven by CNV. (A–B) Distribution of enhanced and weakened ceRNA pairs in GBM **(A)** and LGG **(B)**. Red for enhanced pairs and blue for weakened pairs. **(C)** Global view of ceRNA network in GBM. The ceRNAs driven by CNV and their ceRNA pairs were colored by orange and light gold, respectively. The CNV-driven ceRNA which was also another CNV-driven ceRNA pair was colored by purple. Square indicated mRNAs and diamond indicated lncRNAs.

TABLE 2 | The information of dysregulated ceRNA pairs driven by CNV in GBM and LGG.

	CNV-driven ceRNAs		ceRNA pairs	mRNAs	lncRNAs	Enhanced pairs	Weakened pairs
LGG	Amplification	34	1488	8031	99	545	943
	Deletion	6	12288	3600	233	549	11739
GBM	Amplification	7	102	147	6	29	73
	Deletion	3	160	83	5	3	157

weakened because of the CNV-driven ceRNAs, and only a few CNV-driven ceRNAs (ELAVL2 and PDGFRA) showed opposite influence (**Figure 2C**). Similar results were also observed in LGG. Interestingly, many CNV-driven genes shared the same ceRNAs in the ceRNA network, and the number of sharing ceRNAs in LGG was larger than GBM. For example, VOPP1 and CDKN2A, which have been proved important in glioma (Xia et al., 2013; Roy et al., 2016), were linked by KCTD5 in GBM (**Figure 2C**). It should be noted that MARCH9, a CNV-driven ceRNA, was also regulated by ELAVL2, and they shared the most ceRNAs (such as MTMR1, STMN1, and CECR2). The interactions between STMN1 and ELAVL2/MARCH9 were weakened by CNV, while in MTMR1 and CECR2 the interactions were weakened by MARCH9 amplification and enhanced by ELAVL2 deletion. In LGG, the ceRNAs shared by MTAP and CDKN2A contained many genes highly associated with gliomas and other cancers, such as IDH1 and CDK4/6 (Cheng et al., 2017). Some studies have shown that co-deletion of CDKN2A and MTAP could be used as markers for glioma stratification, and the deletion of CDKN2A was associated with the expression of CDK4/6 in various tumors (Kaul et al., 2015; Frazao et al., 2018).

Functional Characterization of Dysregulated ceRNAs Driven by CNV

To evaluate the effects of CNV-driven dysregulated ceRNAs, we used a functional analysis pipeline to characterize their aberrant functions in LGG and GBM, respectively (see Materials and Methods). In LGG, the top significant KEGG pathways, such as cell cycle and p53 signaling pathway, have been shown to play a crucial role in tumor occurrence (**Figure 3A**). For example, the activation of tumor suppressor protein p53 was confirmed to

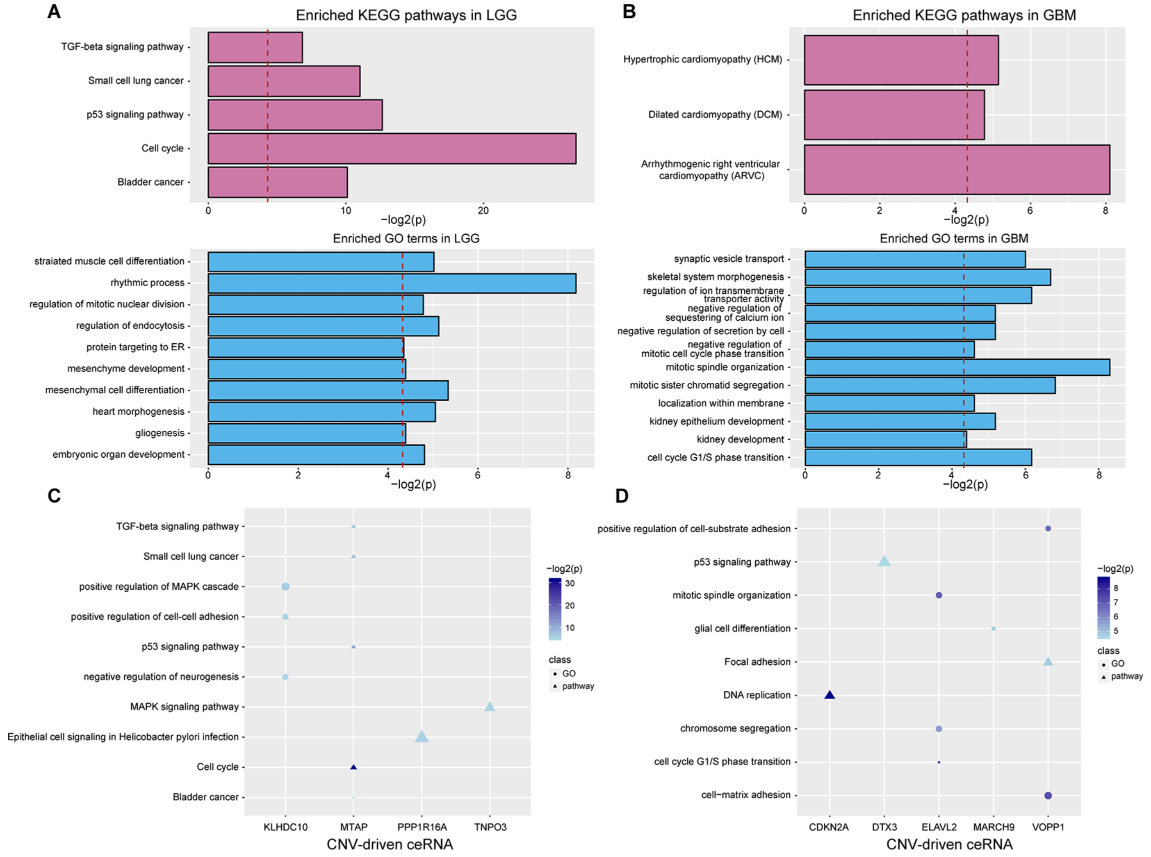


FIGURE 3 | Functional analysis of CNV-driven dysregulated ceRNAs. (A–B) KEGG pathways and GO terms annotated by all dysregulated ceRNAs in LGG **(A)** and GBM **(B)**. The red dotted line indicates that p-value is 0.05. (C-D) KEGG pathways and GO terms annotated by partial CNV-driven ceRNA and its ceRNA pairs in LGG **(C)** and GBM **(D)**. The size of the scatter represents the relative proportion of genes which enriched in the corresponding function.

be regulated by CHK-2 kinase in p53 signaling pathway, which indicated that ceRNA network could reflect the mechanism of tumorigenesis (Harris and Levine, 2005). In GBM, dysregulated ceRNAs were primarily enriched in categories related to cell cycle, e.g. cell cycle G1/S phase transition, and cell division, such as mitotic sister chromatid segregation, negative regulation of mitotic cell cycle phase transition and mitotic spindle organization (**Figure 3B**).

We further investigated the functions of ceRNA pairs driven by each CNV with the same approach. By comparing with functions of all dysregulated ceRNAs, we obtained more detailed tumor-related functions in both LGG and GBM. In LGG, an average of three KEGG pathways and four biological processes were identified ($P < 0.05$, **Figure 3C**). The top enriched results not only contained the pathways enriched by dysregulated ceRNAs but also included pathways that regulated cancer development, such as MAPK signaling, which has been shown to significantly promote the proliferation and migration of glioma cells (Wan and Too, 2010; Zhang et al., 2017). Furthermore, we observed ceRNA pairs enriched in the cell cycle, including CDKN2A (a CNV-driven ceRNA), CDK4 and CDK6. It has been proven that cell cycle was mediated by CDKN2A (Aoki et al., 2018), its dysregulation driven by copy number deletion could inhibit CDK4 and CDK6 and thus blocked traversal from G1 to S-phase (Serrano et al., 1993; Kamb et al., 1994). We also found many cancer-related biological functions in GBM, such as p53 signaling pathway, DNA replication as well as GO terms associated with cell cycle (**Figure 3D**). These results demonstrated that more precise regulatory mechanisms related to glioma could be found by annotating dysregulated ceRNAs.

Exploring Community Structures in the CNV-Driven ceRNA Network

Based on the hypothesis that special topological components in biological networks may provide a new clue to the functional characterization of ceRNAs, we investigated the function of important community structures in the CNV-driven ceRNA network to determine these effects on tumorigenesis (**Figures 2A**, **B**). Here, modules identified from multi-level optimization of modularity were defined as communities (Song et al., 2017).

The largest community in LGG contained 798 nodes, including some glioma-associated genes like IDH1 and CDK4/6 (Cheng et al., 2017), in which most ceRNA pairs were driven by copy number deletion. The functional analysis showed that six GO terms and five KEGG pathways were significantly enriched in this community (p-value < 0.05), such as mesenchyme development, p53 signaling pathway and TGF-beta signaling (**Figure 4**). In this community, BMP-7, as a ceRNA driven by MTAP, has been proved to act as a tumor suppressor that repressed proliferation, self-renewal, and tumor initiation of stem-like glioblastoma cells through suppressing epithelial–mesenchymal transition (EMT) (Zeisberg et al., 2003; Tate et al., 2012). Among all the enriched functions, cell cycle was the most significant (**Figure 4B**), and CDKN2B (Ink4b) drew our attention. As a CNV-driven ceRNA, CDKN2B has been reported to serve as a functional unit in the oncogenesis of malignant gliomas (Shete et al., 2009; Weller et al., 2009), its ceRNA pairs, CDK2 and RBL1, were also annotated in cell cycle and located in the downstream of the pathway (**Figure 4C**). Analogous results were also obtained from the communities of GBM. The largest community of GBM with 34 genes was identified to be relevant to cell cycle-related biological processes (G1/S transition of mitotic cell cycle) and cancer-related pathways (DNA replication) (**Figure 4D**).

CNV-Driven ceRNAs Associated With Prognosis and Histological Subtypes

To further detect the roles of CNV-driven ceRNAs in prognosis, we assessed whether the effects on the clinical outcome of a CNV-driven ceRNA differed by copy number status. We identified some ceRNAs were significantly related to overall survival in LGG (log-rank test p-value < 0.05, **Figure 5**), but regretfully we did not find any significant results in GBM. For LGG, our results showed that the deletion of MTAP, CDKN2A, and CDKN2B had the worse prognosis (with hazard ratios of 1.946, 1.992 and 1.984, respectively). The dysregulated ceRNA network driven by the deletion of CDKN2B was enriched in Epac1/Rap1 pathway, which was proved to be important in glioma cell death (Moon et al., 2012). By using the Cox proportional hazards regression model, we found that the CNV-driven ceRNAs, such as MTAP, KLHL9, and ELAVL2, whose deletion led to worse overall survival also exhibited significant associations between their expression and survival time (**Table 3**, univariate Cox hazard analysis, $P < 0.05$). Seven of them, for example, KLHL9, showed to be independent prognosis factors (**Table 3**, multivariate Cox hazard analysis, $P < 0.05$).

Furthermore, we found that the CNV-driven prognosis factors also showed high correlation with histological subtype (**Table 4**, Fisher exact test, $P < 0.05$). Interestingly, all of the subtype related CNV-driven ceRNAs were located in the deletion region at 9p21. It has been shown that the deletion of 9p21, especially co-deletions of CDKN2A/B and MTAP, could be a marker for different grades of glioma (Frazao et al., 2018). Interestingly, CDKN2B, CDKN2A, MTAP, and KLHL9 also belonged to the largest community in the dysregulated ceRNA network, suggesting their possible role to inhibit the development of glioma together. Besides, we also found a lncRNA, RP11–321l2.2, whose ceRNA pairs were involved in MAPK and PI3K pathways.

DISCUSSION

In this study, we provided a comprehensive catalog of dysregulated ceRNA interactions driven by CNV in both LGG and GBM. We identified the expression of protein-coding genes and lncRNAs affected by CNVs and figured out consistent changes of genes in both cancer subtypes. Based on the CNV-driven genes and ceRNA triplets, dysregulated ceRNA networks driven by copy number amplification/deletion were identified in LGG and GBM. We found that CNV could attenuate the interactions between most ceRNA pairs, and the dysregulated ceRNAs driven by CNV were involved in some critical biological functions in glioma. Furthermore, some CNV-driven ceRNAs showed a

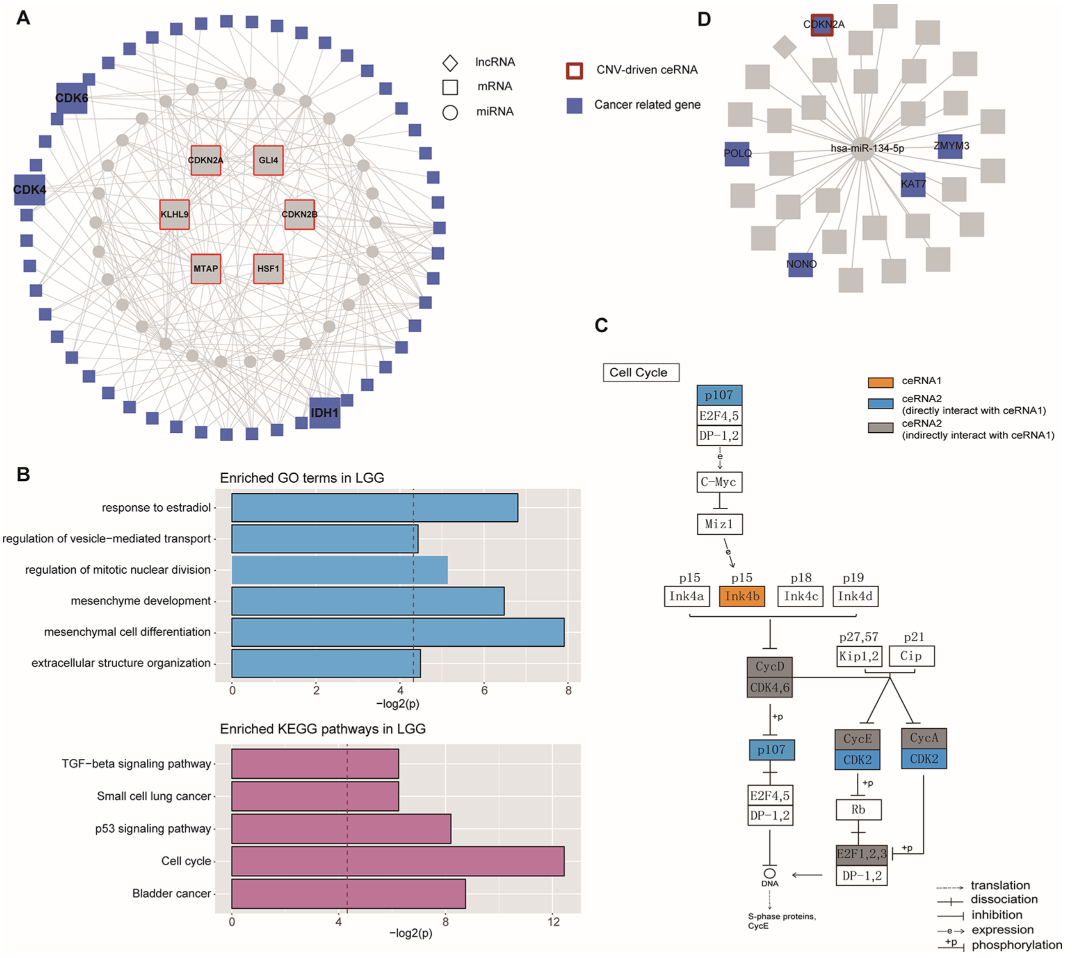


FIGURE 4 | Community analysis in LGG and GBM. **(A)** The architecture of the largest community in LGG. **(B)** Significantly enriched GO terms and KEGG pathways of the largest community in LGG. **(C)** The regulation of ceRNAs involved in cell cycle pathway. **(D)** The architecture of the largest community in GBM.

significant correlation to overall survival, indicating that they may be potential clinical biomarkers of prognosis.

We not only demonstrated that the dysregulated ceRNA network could be influenced by CNV in both LGG and GBM but also obtained some critical biological functions related to the CNV-driven dysregulated ceRNAs. These ceRNAs were significantly enriched in the programs of tumorigenesis, such as cell cycle, p53 signaling pathway. By further functional analysis of each CNV-driven ceRNA sub-network, we identified more detailed tumor-related functions, for example, cell cycle G1/S phase transition. Our study demonstrated a novel finding that the CDKN2B (p15, driven by copy number amplification) could regulate TGF-β signaling pathway in LGG. TGFβR1, which was a ceRNA pair of CDKN2B, is activated by binding with TGF-β (Massague, 1992). Another ceRNA GDNF, a member of TGF-β super-family, has been revealed to strongly induce glioma cell proliferation and migration (Song and Moon, 2006; Ng et al., 2009). These findings could potentially account for the mechanism that TGF-β receptors may be mediated by CDKN2B to influence the glioma occurrence and development. Meanwhile, higher levels of RhoA, another ceRNA member of CDKN2B and a downstream factor in TGF-β/MAPK signaling pathway, can significantly promote glioma cell proliferation and migration (Wan and Too, 2010; Shabtay-Orbach et al., 2015). These results suggest that although the regulation of CDKN2B through TGF superfamily members is not clear, it is worth to determine in the future.

By performing a functional analysis of the largest community in CNV-driven ceRNA network, we could identify key biological functions relevant to LGG pathogenesis. Epithelial–mesenchymal transition (EMT) is known as a facilitator of cellular dissociation and migration, which plays a critical role in cancer metastasis (Cheng et al., 2012; Iwadate, 2016). Our results elucidated a key EMT-related molecule: BMP-7 and discovered a critical ceRNA interaction between MTAP and BMP-7. The ceRNA interactions explain the role of EMT in malignant glioma, which may provide new insight into the mechanism of tumorigenesis. Additionally, the loss of CDKN2B could cause the dysregulation of its relevant community structures, by affecting the expression of its ceRNA partners, including CDK2 and RBL1, and ultimately resulted in

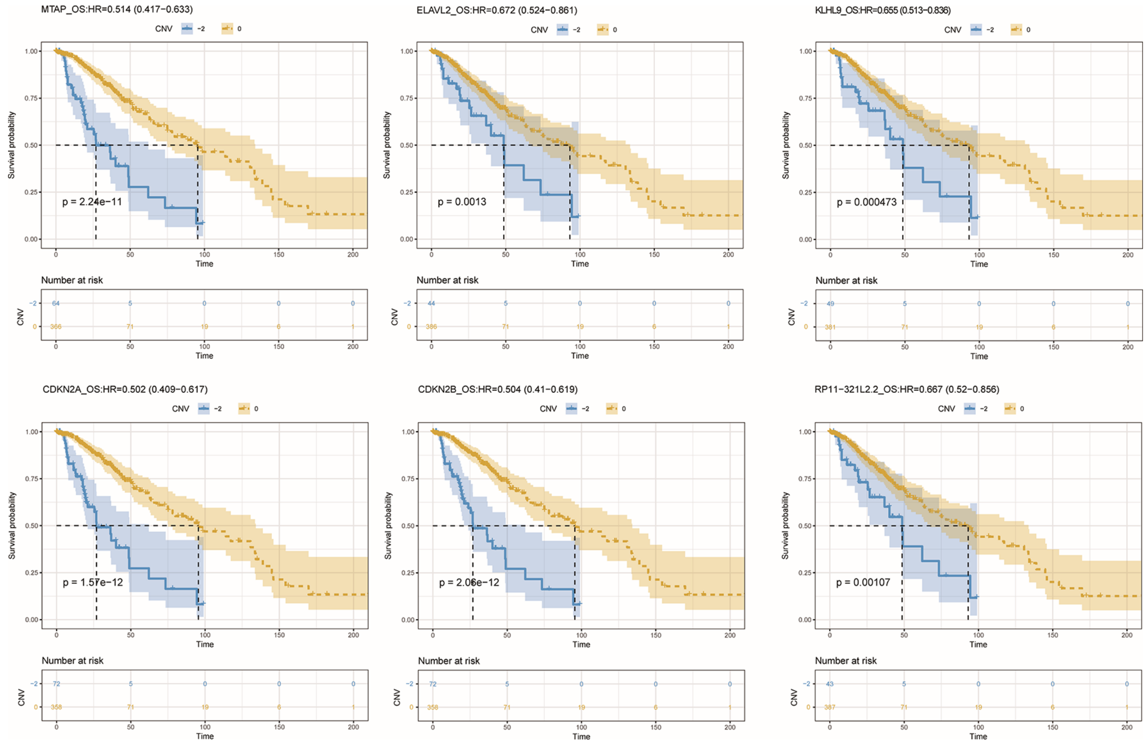


FIGURE 5 | Overall survival among LGG patients (n = 432) stratified by the copy number status of CNV-driven ceRNAs.

TABLE 3 | Univariate and multivariate Cox hazard analyses in LGG.

CNV-driven ceRNA	Univariate analysis		Multivariate analysis	
	HR (95% CI for HR)	P value	HR (95% CI for HR)	P value
MTAP	0.34 (0.2334–0.4953)	<0.0001	1.0602 (0.4489–2.5036)	0.8939
KLHL9	0.4094 (0.3043–0.5509)	<0.0001	0.3836 (0.207–0.7107)	0.0023
ELAVL2	0.6828 (0.5714–0.8159)	<0.0001	1.3713 (0.9268–2.0289)	0.1142
ZNF517	0.6397 (0.4225–0.9687)	0.0348	0.0913 (0.0286–0.2915)	0.0001
SCRIB	1.5802 (1.1698–2.1345)	0.0029	7.2575 (2.4772–21.2626)	0.0003
PUF60	0.542 (0.3389–0.8668)	0.0106	0.0707 (0.0132–0.3791)	0.002
TNPO3	1.476 (1.0102–2.1566)	0.0442	0.2493 (0.0815–0.7628)	0.0149
VPS28	0.4436 (0.2507–0.785)	0.0052	0.1131 (0.0195–0.6546)	0.015
RP11-2E11.5	0.4738 (0.3411–0.6582)	<0.0001	0.5463 (0.3263–0.9146)	0.0215
GRINA	0.4164 (0.2753–0.6299)	<0.0001	0.4701 (0.1931–1.1442)	0.0963
ARHGAP39	1.4211 (1.0099–1.9998)	0.0438	2.0575 (0.8929–4.7411)	0.0903
CEP41	1.5897 (1.1494–2.1987)	0.0051	1.1195 (0.5523–2.2692)	0.7542
CYHR1	1.5474 (1.1329–2.1134)	0.0061	1.5989 (0.6935–3.6862)	0.2708
KLHDC10	1.592 (1.0146–2.4981)	0.0431	0.9731 (0.2574–3.6795)	0.968
LY6K	0.3638 (0.1954–0.6773)	0.0014	0.7473 (0.292–1.9124)	0.5434
PPP1R16A	1.5943 (1.1601–2.191)	0.004	0.6491 (0.2303–1.8295)	0.4136
ZC3HC1	1.6585 (1.0856–2.5338)	0.0193	1.1166 (0.4487–2.779)	0.8126
ZNF707	1.7057 (1.1677–2.4915)	0.0057	1.6517 (0.6609–4.1278)	0.2829
RECQL4	1.6168 (1.2803–2.0417)	0.0001	0.9333 (0.5146–1.6927)	0.8202

cell-cycle dysregulation. These ceRNAs founded by exploring specific community structures could provide new potential therapeutic targets for malignant gliomas.

Our study further revealed the putative influence of CNV-driven ceRNAs in clinicopathologic characteristics. By performing a systematic analysis of the CNV-driven ceRNAs with clinical features, we found that the CNVs of some genes (such as MTAP/CDKN2A/CDKN2B/KLHL9) had significant impacts on histological diagnosis and survival in glioma. Functional analysis of CDKN2B through its influenced ceRNA network further revealed that the dysregulation of specific ceRNA networks driven by CNVs could act as prognostic markers of glioma

TABLE 4 | Fisher exact test of histological subtypes and copy number status of CNV-driven ceRNAs.

ceRNA	CNV	Histological subtypes			p-value
		Astrocytoma	Oligoastrocytoma	Oligodendroglioma	
MTAP	0	117	101	150	5.61E-05
	−2	39	12	13	
ELAVL2	0	128	105	155	0.000373
	−2	28	8	8	
KLHL9	0	125	104	154	0.00018
	−2	31	9	9	
CDKN2A	0	111	101	148	3.50E-06
	−2	45	12	15	
CDKN2B	0	110	101	149	8.65E-07
	−2	46	12	14	
RP11-321L2.2	0	129	105	155	0.000736
	−2	27	8	8	

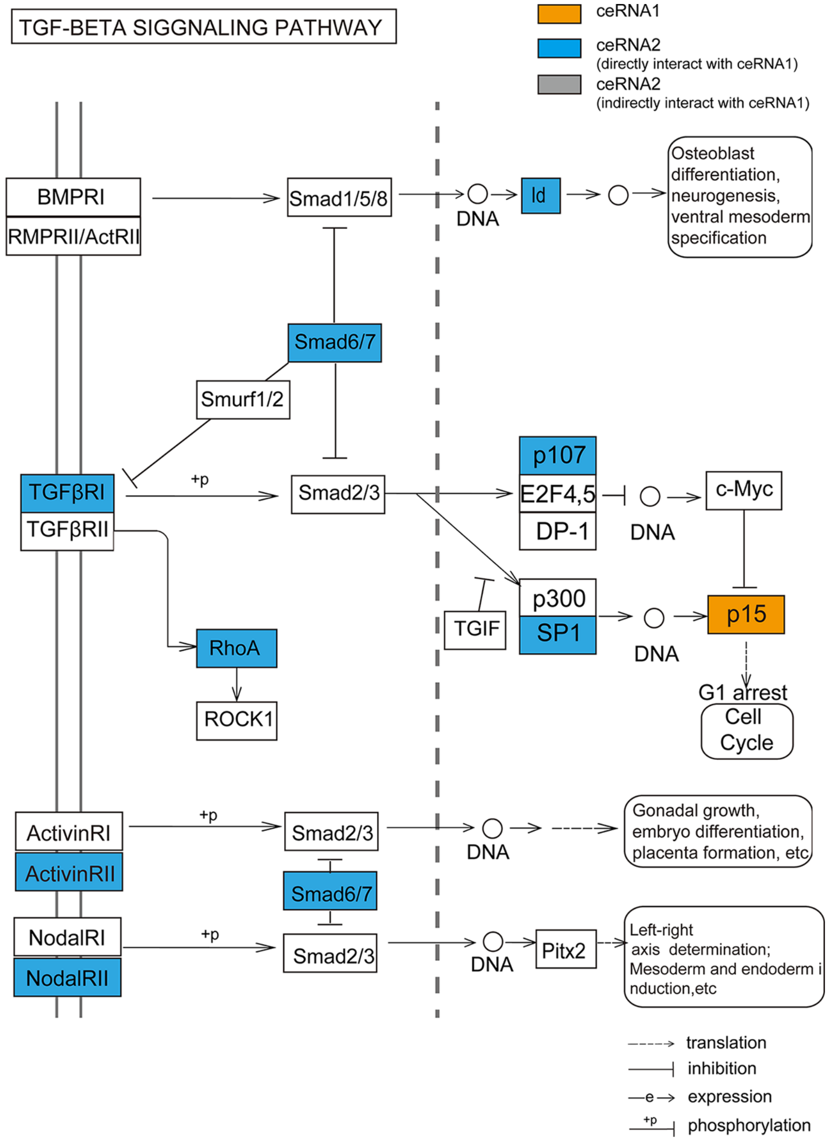

TGF-BETA SIGGNALING PATHWAY
ceRNA1
ceRNA2
(directly interact with ceRNA1)
ceRNA2
(indirectly interact with ceRNA1)
BMPRI
RMPRII/ActRII
Smad1/5/8
DNA
Id
Osteoblast differentiation, neurogenesis, ventral mesoderm specification
Smad6/7
Smurf1/2
TGFβRI
TGFβRII
+p
Smad2/3
p107
E2F4,5
DP-1
DNA
c-Myc
TGIF
p300
SP1
DNA
p15
G1 arrest
Cell Cycle
RhoA
ROCK1
ActivinRI
ActivinRII
+p
Smad2/3
DNA
Gonadal growth, embryo differentiation, placenta formation, etc
Smad6/7
NodalRI
NodalRII
+p
Smad2/3
DNA
Pitx2
Left-right axis determination; Mesoderm and endoderm i nduction,etc
translation
inhibition
expression
phosphorylation

FIGURE 6 | The TGF-beta signaling pathway annotated by ceRNA pairs of CDKN2B (p15). Orange node represents CNV-driven ceRNA. Blue nodes represent the ceRNA members of the CNV-driven ceRNA.

(**Figure 6**). We proposed that the CNV-driven ceRNAs detected to be associated with clinical features may possess clinical functions through regulating other genes by ceRNA networks. The CNV-driven ceRNA network could be used to presume potential prognostic markers of glioma.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://tcga-data.nci.nih.gov/tcga.

AUTHOR CONTRIBUTIONS

YX, CX, and JX conceived and designed this study. XH, LP, SS, and SH collected and analyzed the data. JX, XH, SS, and SH carried out the method and performed the analysis. YY, KL, and LX helped to analyze the results. JX, XH, SS, and YY participated in the discussion of the project. JX, LP and WY revised the manuscript. All authors reviewed, edited, and approved the manuscript.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China [61573122, 31871336]; National Science Foundation of Heilongjiang Province [YQ2019C012]; Heilongjiang Postdoctoral Science Foundation [LBH-Q18099]; Special Funds for the Construction

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01055/full#supplementary-material

SUPPLEMENTARY FIGURE 1 | The computational approach to identify dysregulated ceRNA–ceRNA interactions driven by CNVs.

SUPPLEMENTARY TABLE S1 | List of the CNV-driven genes.

SUPPLEMENTARY TABLE S2 | List of the significant ceRNA pairs driven by CNV.

REFERENCES

Alentorn, A., Dehais, C., Ducray, F., Carpentier, C., Mokhtari, K., Figarella-Branger, D., et al. (2015). Allelic loss of 9p21.3 is a prognostic factor in 1p/19q codeleted anaplastic gliomas. *Neurology* 85, 1325–1331. doi: 10.1212/WNL.0000000000002014

Aoki, K., Nakamura, H., Suzuki, H., Matsuo, K., Kataoka, K., Shimamura, T., et al. (2018). Prognostic relevance of genetic alterations in diffuse lower-grade gliomas. *Neuro-Oncology* 20, 66–77. doi: 10.1093/neuonc/nox132

Bhargava, S., Patil, V., Mahalingam, K., and Somasundaram, K. (2017). Elucidation of the genetic and epigenetic landscape alterations in RNA binding proteins in glioblastoma. *Oncotarget* 8, 16650–16668. doi: 10.18632/oncotarget.14287

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., et al. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 147, 358–369. doi: 10.1016/j.cell.2011.09.028

Cheng, W., Ren, X., Zhang, C., Cai, J., Han, S., and Wu, A. (2017). Gene expression profiling stratifies idh1-mutant glioma with distinct prognoses. *Mol. Neurobiol.* 54, 5996–6005. doi: 10.1007/s12035-016-0150-6

Cheng, W. Y., Kandel, J. J., Yamashiro, D. J., Canoll, P., and Anastassiou, D. (2012). A multi-cancer mesenchymal transition gene expression signature is associated with prolonged time to recurrence in glioblastoma. *PloS One* 7, e34705. doi: 10.1371/journal.pone.0034705

Crespo, I., Tao, H., Nieto, A. B., Rebelo, O., Domingues, P., Vital, A. L., et al. (2012). Amplified and homozygously deleted genes in glioblastoma: impact on gene expression levels. *PloS One* 7, e46088. doi: 10.1371/journal.pone.0046088

Do, D., and Bozdag, S. (2018). Cancerin: a computational pipeline to infer cancer-associated ceRNA interaction networks. *PLoS Comput. Biol.* 14, e1006318. doi: 10.1371/journal.pcbi.1006318

Feng, J., Kim, S. T., Liu, W., Kim, J. W., Zhang, Z., Zhu, Y., et al. (2012). An integrated analysis of germline and somatic, genetic and epigenetic alterations at 9p21.3 in glioblastoma. *Cancer* 118, 232–240. doi: 10.1002/cncr.26250

Frazao, L., do Carmo Martins, M., Nunes, V. M., Pimentel, J., Faria, C., Miguens, J., et al. (2018). BRAF V600E mutation and 9p21: CDKN2A/B and MTAP co-deletions - markers in the clinical stratification of pediatric gliomas. *BMC Cancer* 18, 1259. doi: 10.1186/s12885-018-5120-0

Harris, S. L., and Levine, A. J. (2005). The p53 pathway: positive and negative feedback loops. *Oncogene* 24, 2899–2908. doi: 10.1038/sj.onc.1208615

Inoue, R., Moghaddam, K. A., Ranasinghe, M., Saeki, Y., Chiocca, E. A., and Wade-Martins, R. (2004). Infectious delivery of the 132 kb CDKN2A/CDKN2B genomic DNA region results in correctly spliced gene expression and growth suppression in glioma cells. *Gene Therapy* 11, 1195–1204. doi: 10.1038/sj.gt.3302284

Iwadate, Y. (2016). Epithelial-mesenchymal transition in glioblastoma progression. *Oncol. Letters* 11, 1615–1620. doi: 10.3892/ol.2016.4113

Jornsten, R., Abenius, T., Kling, T., Schmidt, L., Johansson, E., Nordling, T. E., et al. (2011). Network modeling of the transcriptional effects of copy number aberrations in glioblastoma. *Mol. Syst. Biol.* 7, 486. doi: 10.1038/msb.2011.17

Kamb, A., Gruis, N. A., Weaver-Feldhaus, J., Liu, Q., Harshman, K., Tavtigian, S. V., et al. (1994). A cell cycle regulator potentially involved in genesis of many tumor types. *Science* 264, 436–440. doi: 10.1126/science.8153634

Kaul, A., Toonen, J. A., Cimino, P. J., Gianino, S. M., and Gutmann, D. H. (2015). Akt- or MEK-mediated mTOR inhibition suppresses Nf1 optic glioma growth. *Neuro-Oncology* 17, 843–853. doi: 10.1093/neuonc/nou329

Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., et al. (2015). TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.* 75, 3728–3737. doi: 10.1158/0008-5472.CAN-15-0273

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi: 10.1093/nar/gkt1248

Liang, L., Fang, J. Y., and Xu, J. (2016). Gastric cancer and gene copy number variation: emerging cancer drivers for targeted therapy. *Oncogene* 35, 1475–1482. doi: 10.1038/onc.2015.209

Lopez-Gines, C., Gil-Benso, R., Ferrer-Luna, R., Benito, R., Serna, E., Gonzalez-Darder, J., et al. (2010). New pattern of EGFR amplification in glioblastoma and the relationship of gene copy number with gene expression profile. *Mod Pathol. Off. J U. S. Can Acad Pathol. Inc* 23, 856–865. doi: 10.1038/modpathol.2010.62

Massague, J. (1992). Receptors for the TGF-beta family. *Cell* 69, 1067–1070. doi: 10.1016/0092-8674(92)90627-O

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* 12, R41. doi: 10.1186/gb-2011-12-4-r41

Momtaz, R., Ghanem, N. M., El-Makky, N. M., and Ismail, M. A. (2018). Integrated analysis of SNP, CNV and gene expression data in genetic association studies. *Clin. Genet.* 93, 557–566. doi: 10.1111/cge.13092

Moon, E. Y., Lee, G. H., Lee, M. S., Kim, H. M., and Lee, J. W. (2012). Phosphodiesterase inhibitors control A172 human glioblastoma cell death through cAMP-mediated activation of protein kinase A and Epac1/Rap1 pathways. *Life Sci.* 90, 373–380. doi: 10.1016/j.lfs.2011.12.010

Ng, W. H., Wan, G. Q., Peng, Z. N., and Too, H. P. (2009). Glial cell-line derived neurotrophic factor (GDNF) family of ligands confer chemoresistance in a ligand-specific fashion in malignant gliomas. *J. Clin. Neurosci. Off. J. Neurosurgical Soc. Australasia* 16, 427–436. doi: 10.1016/j.jocn.2008.06.002

Ostrom, Q. T., Gittleman, H., Liao, P., Rouse, C., Chen, Y., Dowling, J., et al. (2014). CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2007-2011. *Neuro-Oncology* 16 Suppl 4, iv1–i63. doi: 10.1093/neuonc/nou223

Park, C., Ahn, J., Yoon, Y., and Park, S. (2012). Identification of functional CNV region networks using a CNV-gene mapping algorithm in a genome-wide scale. *Bioinformatics* 28, 2045–2051. doi: 10.1093/bioinformatics/bts318

Park, C., Kim, J. I., Hong, S. N., Jung, H. M., Kim, T. J., Lee, S., et al. (2017). A copy number variation in PKD1L2 is associated with colorectal cancer predisposition in korean population. *Int. J. Cancer* 140, 86–94. doi: 10.1002/ijc.30421

Parsons, D. W., Jones, S., Zhang, X., Lin, J. C., Leary, R. J., Angenendt, P., et al. (2008). An integrated genomic analysis of human glioblastoma multiforme. *Science* 321, 1807–1812. doi: 10.1126/science.1164382

Roy, D. M., Walsh, L. A., Desrichard, A., Huse, J. T., Wu, W., Gao, J., et al. (2016). Integrated genomics for pinpointing survival loci within arm-level somatic copy number alterations. *Cancer Cell* 29, 737–750. doi: 10.1016/j.ccell.2016.03.025

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353–358. doi: 10.1016/j.cell.2011.07.014

Schwartzbaum, J. A., Fisher, J. L., Aldape, K. D., and Wrensch, M. (2006). Epidemiology and molecular pathology of glioma. *Nat. Clin. Pract. Neurol.* 2, 494–503. doi: 10.1038/ncpneuro0289 quiz 491 p following 516.

Serrano, M., Hannon, G. J., and Beach, D. (1993). A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. *Nature* 366, 704–707. doi: 10.1038/366704a0

Shabtay-Orbach, A., Amit, M., Binenbaum, Y., Na'ara, S., and Gil, Z. (2015). Paracrine regulation of glioma cells invasion by astrocytes is mediated by glial-derived neurotrophic factor. *Int. J. Cancer* 137, 1012–1020. doi: 10.1002/ijc.29380

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shete, S., Hosking, F. J., Robertson, L. B., Dobbins, S. E., Sanson, M., Malmer, B., et al. (2009). Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.* 41, 899–904. doi: 10.1038/ng.407

Simon, M., Koster, G., Menon, A. G., and Schramm, J. (1999). Functional evidence for a role of combined CDKN2A (p16-p14(ARF))/CDKN2B (p15) gene inactivation in malignant gliomas. *Acta Neuropathol.* 98, 444–452. doi: 10.1007/s004010051107

Song, C., Zhang, J., Qi, H., Feng, C., Chen, Y., Cao, Y., et al. (2017). The global view of mRNA-related ceRNA cross-talks across cardiovascular diseases. *Sci. Rep.* 7, 10185. doi: 10.1038/s41598-017-10547-z

Song, H., and Moon, A. (2006). Glial cell-derived neurotrophic factor (GDNF) promotes low-grade Hs683 glioma cell migration through JNK, ERK-1/2 and p38 MAPK signaling pathways. *Neurosci. Res.* 56, 29–38. doi: 10.1016/j.neures.2006.04.019

Tate, C. M., Pallini, R., Ricci-Vitiani, L., Dowless, M., Shiyanova, T., D'Alessandris, G. Q., et al. (2012). A BMP7 variant inhibits the tumorigenic potential of glioblastoma stem-like cells. *Cell Death Differ.* 19, 1644–1654. doi: 10.1038/cdd.2012.44

Tay, Y., Kats, L., Salmena, L., Weiss, D., Tan, S. M., Ala, U., et al. (2011). Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* 147, 344–357. doi: 10.1016/j.cell.2011.09.029

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. *Nature* 505, 344–352. doi: 10.1038/nature12986

Wan, G., and Too, H. P. (2010). A specific isoform of glial cell line-derived neurotrophic factor family receptor alpha 1 regulates RhoA expression and glioma cell migration. *J. Neurochem.* 115, 759–770. doi: 10.1111/j.1471-4159.2010.06975.x

Wang, Y., Xu, Z., Jiang, J., Xu, C., Kang, J., Xiao, L., et al. (2013). Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev. Cell* 25, 69–80. doi: 10.1016/j.devcel.2013.03.002

Weller, M., Felsberg, J., Hartmann, C., Berger, H., Steinbach, J. P., Schramm, J., et al. (2009). Molecular predictors of progression-free and overall survival in patients with newly diagnosed glioblastoma: a prospective translational study of the German Glioma Network. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* 27, 5743–5750. doi: 10.1200/JCO.2009.23.0805

Xi, R., Hadjipanayis, A. G., Luquette, L. J., Kim, T. M., Lee, E., Zhang, J., et al. (2011). Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl. Acad. Sci. U.S.A.* 108, E1128–E1136. doi: 10.1073/pnas.1110574108

Xia, H., Yan, Y., Hu, M., Wang, Y., Wang, Y., Dai, Y., et al. (2013). MiR-218 sensitizes glioma cells to apoptosis and inhibits tumorigenicity by regulating ECOP-mediated suppression of NF-kappaB activity. *Neuro-Oncology* 15, 413–422. doi: 10.1093/neuonc/nos296

Xu, H., Zong, H., Ma, C., Ming, X., Shang, M., Li, K., et al. (2017). Epidermal growth factor receptor in glioblastoma. *Oncol. Letters* 14, 512–516. doi: 10.3892/ol.2017.6221

Yoon, J. H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J. L., De, S., et al. (2012). LincRNA-p21 suppresses target mRNA translation. *Mol. Cell* 47, 648–655. doi: 10.1016/j.molcel.2012.06.027

Zeisberg, M., Hanai, J., Sugimoto, H., Mammoto, T., Charytan, D., Strutz, F., et al. (2003). BMP-7 counteracts TGF-beta1-induced epithelial-to-mesenchymal transition and reverses chronic renal injury. *Nat. Med.* 9, 964–968. doi: 10.1038/nm888

Zhang, B. L., Dong, F. L., Guo, T. W., Gu, X. H., Huang, L. Y., and Gao, D. S. (2017). MiRNAs Mediate GDNF-Induced proliferation and migration of glioma cells. *Cell. Physiol. Biochem. Int. J. Exp. Cell. Physiol. Biochem. Pharmacol.* 44, 1923–1938. doi: 10.1159/000485883

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: 30 October 2019
doi: 10.3389/fgene.2019.00990

ContraDRG: Automatic Partial Charge Prediction by Machine Learning

Roman Martin [1,2] *and Dominik Heider* [1]*

[1] *Department of Mathematics and Computer Science, University of Marbug, Marburg, Germany,* [2] *Department of Organic-Analytical Chemistry, TUM Campus Straubing, Straubing, Germany*

In recent years, machine learning techniques have been widely used in biomedical research to predict unseen data based on models trained on experimentally derived data. In the current study, we used machine learning algorithms to emulate computationally complex predictions in a reverse engineering–like manner and developed ContraDRG, a software that can be used to predict partial charges for small molecules based on PRODRG and Automated Topology Builder (ATB) predictions. Both tools generate molecular topology files, including the partial atomic charge, by using different procedures. We show that ContraDRG can accurately predict partial charges in a fraction of the time, because it exploits existing complex models with intensive calculations by using machine learning techniques and thus can also be applied for screening projects with large amounts of molecules. We provide ContraDRG as a web server, which can be used to automatically assign partial charges to incoming user-specified molecules by using our machine learning models. In this study, we compared ContraDRG with PRODRG and ATB in regard of predictivity by statistical methods. ContraDRG allows predicting ATB-derived partial charges with an R^2 value up to 0.980 and for PRODRG up to 1.00. While ATB requires hours or days for the quantum mechanical accurate calculation and refinements, ContraDRG does its approximation within seconds.

Keywords: PRODRG, ATB, machine learning, molecular dynamics simulations, partial charge prediction

OPEN ACCESS

Edited by:
Harinder Singh,
J. Craig Venter Institute,
United States

Reviewed by:
Kumardeep Chaudhary,
Icahn School of Medicine at
Mount Sinai,United States
Sandeep Kumar Dhanda,
La Jolla Institute for Immunology (LJI),
United States

****Correspondence:***
Dominik Heider
dominik.heider@uni-marburg.de

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

Received: *26 June 2019*
Accepted: *18 September 2019*
Published: *30 October 2019*

Citation:
Martin R and Heider D (2019)
ContraDRG: Automatic
Partial Charge Prediction by
Machine Learning.
Front. Genet. 10:990.
doi: 10.3389/fgene.2019.00990

INTRODUCTION

In the last decades, several studies demonstrated how machine learning algorithms were able to create accurate predictions or classifications from experimentally derived data. The applications of machine learning algorithms in biomedical research are diverse (Larrañaga et al., 2006) and range from single-molecule interaction prediction for drug design (Lavecchia, 2015) or omics pattern recognition (Stanke and Morgenstern, 2005), toward the prediction of entire biological systems (D'Alche-Buc and Wehenkel, 2008).

However, in the current study, we used machine learning algorithms to emulate computationally intensive calculations. Precise determination of topology parameters for small molecules, particularly partial charges, is a crucial step for molecular dynamics (MD) simulations and other biochemical and biophysical computations. In particular, MD simulations depend heavily on the accurate parameterization of the molecules; otherwise, the simulations tend to be unreliable and misleading (Lemkul et al., 2010). One main challenge for generating reliable predictions is the ability to create a force field consistent topology for new small molecules since the force fields theory is mostly derived from empirical analysis.

For this purpose, there are different force fields available, based on diverse parameters and underlying theories, such as GROMOS (van Gunsteren et al., 1996; Daura et al., 1998; Scott et al., 1999; Schuler et al., 2001; Oostenbrink et al., 2004), OPLS (Jorgensen and Tirado-Rives, 1988; Jorgensen et al., 1996), CHARMM (Patel and Brooks, 2004; Patel et al., 2004), and AMBER (Cornell et al., 1995; Wang et al., 2004). Parameterization for synthetic small molecules is supported by the general AMBER force field (Wang et al., 2004) and the general CHARMM force field (Patel and Brooks, 2004; Patel et al., 2004), in contrast to GROMOS and OPLS. While detailed information about the GROMOS96 parameter sets is not publicly available, OPLS-AA reveals their entire parameter sets, which includes geometry optimization and quantum chemical calculations (Jorgensen et al., 1996; Kaminski et al., 2001). Thus, users of the GROMOS96 force field rely on empirical parameters and subsequent validations by thermodynamic integration (Oostenbrink et al., 2004).

Over the last years, some freely available tools were developed, refined, and established for automated topology generation. Two commonly used tools are PRODRG (Van Aalten, 1996; Schüttelkopf and Van Aalten, 2004) and the Automated Topology Builder (ATB) (Malde et al., 2011; Koziara et al., 2014; Stroet et al., 2018). Both are frequently used tools that receive user-defined small-molecule files and return parameterized GROMOS-compatible topology files including their partial atomic charges. While PRODRG partial charge determination is based on mapping of building blocks and charge groups onto a database, ATB uses quantum chemical calculations involving electron densities and geometry optimizations (Chandra Singh and Kollman, 1984). However, PRODRG is much faster compared to ATB and produces topologies within seconds, while ATB requires up to multiple days, but generates more precise, more reliable, and more consistent results (Lemkul et al., 2010; Malde et al., 2011). Both tools have been already used for protein–peptide, protein–ligand, protein–lipid, and pharmaceutical drug optimizations (Santos et al., 2017). Although both tools provide free access for automated file parameterization, only ATB supplies a modern application programming interface. Additionally, there are several stand-alone tools, such as Open Babel (O'Boyle et al., 2011) and AutoDock Tools (Morris et al., 2009), which can predict partial charges based on different methods like MMFF94 (Halgren, 1999), based on quantum chemical calculations, or QTPIE (Chen and Martı, 2007), which describes the flow in molecules based on charge transfer variables.

While PRODRG and ATB are proprietary software, they do provide free access for academic purpose. Contrary to that, fully proprietary software like VeraChem's VCharge or Schroedinger's Maestro, which predict, among others, partial charges are also available. VCharge uses a method based on QM-derived electronegativity equalization (Gilson et al., 2003), and Maestro computes the charges according CM1A-BCC (OPLS3e) (Marenich et al., 2012; Roos et al., 2019). Additionally, there is proprietary software such as Amber that requires external tools for partial charges predictions, like the provided and recommended free antechamber (Wang et al., 2006). Antechamber applies usually the AM1-BCC method (Jakalian et al., 2002) for small molecules and can be optimized with provided QM calculations by the RESP method (Bayly et al., 1993).

Engler et al. (2019) showed recently in an innovative approach how to solve two common problems of partial charge determination: (i) the single partial-charge assignment per atom and (ii) the total charge determination. By transferring these problems into a multiple-choice knapsack problem (Dudziński and Walukiewicz, 1987; Kellerer et al., 2004), they were able to predict the partial charges automatically. Moreover, a recent study showed that machine learning prediction based on quantum-chemical calculation can be used to predict partial charges (Bleiziffer et al., 2018).

In the current study, we used small-molecule three-dimensional structures files for prediction of partial charges, based on machine-derived data from the web tools PRODRG and ATB. To this end, we analyzed and compared a set of different machine learning methods and emulated the aforementioned tools. Finally, we compared our predictions with the existing tools. This study demonstrates the usefulness of machine learning models for reverse engineering of costly calculations, which are provided in an easy-to-use online tool.

MATERIALS AND METHODS

Dataset

This study is based on two different datasets, namely, the PRODRG dataset and the ATB dataset. The PRODRG dataset is based on randomly selected molecule structures from the PubChem database (Kim et al., 2018). These molecules were converted into Protein Database Bank format *via* Open Babel (O'Boyle et al., 2011) and subsequently predicted *via* the PRODRG server (v. AA100323.0717). Energy minimization was deactivated, and full charge prediction and chirality enabled. The ATB dataset was collected from the curated molecule and topology files from the ATB (v. 3.0) database (Stroet et al., 2018). We mapped the partial charge predictions from the topology files with the provided all-atom Protein Database Bank files.

We calculated the pairwise Tanimoto similarity coefficient *via* Open Babel (linear seven atoms fragments) for all files to ensure that a diverse set of molecules was used (Kim et al., 2018). The Tanimoto coefficient represents a known indicator for molecular structure similarities (Bajusz et al., 2015). Therefore, we determined the coefficient by comparing every molecule to each other. The resulting coefficients were drawn into a violin plot.

Feature Encoding

In the current study, we focused only on organic elements, namely, carbon, hydrogen, nitrogen, oxygen, phosphorus, sulfur, fluorine, bromine, and iodine (C, H, N, O, P, S, F, Cl, Br, and I). We used 61 different features for the encoding of the molecules, where all atoms are individually analyzed (**Figure 1**). Molecules are internally represented as a cyclic undirected graph, where atoms correspond to vertices, and bonds to edges. These

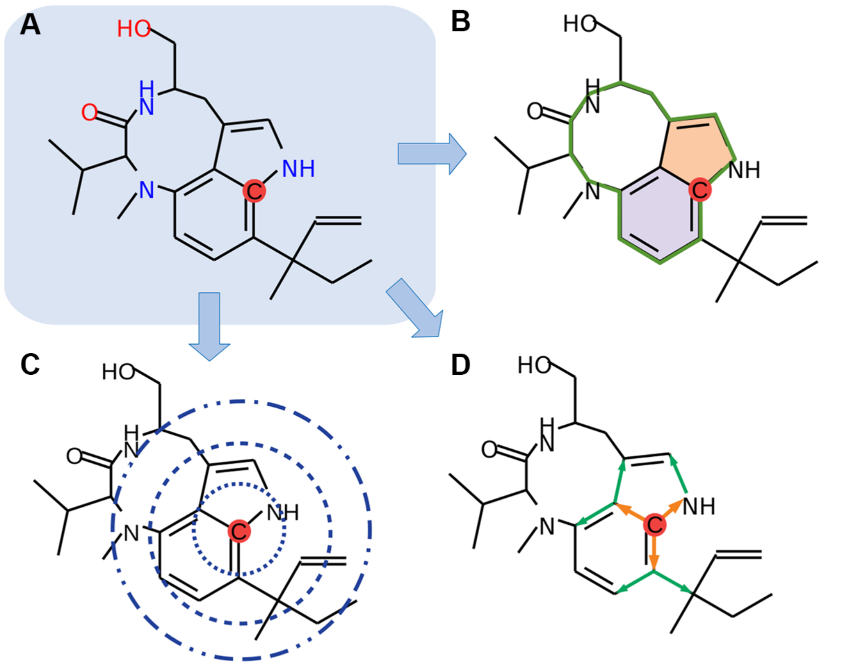


FIGURE 1 | Schematic overview of the feature encoding. **(A)** Each atom will be selected (red dot), and encodings will be generated **(B–D)**. **(B)** Overall circular structures (green line) and nested (colored areas) are detected by a depth-first search. **(C)** Distance searches with three different radii are applied. **(D)** Second-level neighbors path tracing is implemented (orange arrows, first level; green arrows, second level). Chemical structures were drawn with MolView (https://molview.org).

encodings include the hybridization state of carbon atoms, sizes and amounts of nested circles, distances to adjacent atoms, and presence of neighbors through a second-level path tracing. Nested circular structures were identified by a depth-first search derived from the graph theory.

To encode an entire molecule, a list of the positions of the atoms and an adjacency matrix for the bonds are necessary. Protein Database Bank files and SMILE (Weininger, 1988) files can be encoded in such a way easily. However, in contrast to existing approaches, we take explicitly the three-dimensional information into account, thus allowing making prediction also for theoretical molecules.

Machine Learning

We used the R package caret (v. 6.0-81) (Max and Kuhn, 2008) for building the machine learning models. We build models for each element independently. The datasets (one dataset for each element) were split into train and test data with a ratio of 1:4. We trained different models including linear regression, stochastic gradient boosting (Friedman, 2002), random forests (RF) (Breiman, 2001), quantile regression forests (Meinshausen, 2006), weighted k-nearest neighbors (Altman, 1992), and support vector machines (SVMs) (Cortes and Vapnik, 1995) with different kernels. RFs were trained with 500 trees and k-nearest neighbors were built based on $k = 7$ and a Minkowski distance of 2. All other models were trained with default parameters. All models were trained with the partial charge values as labels from PRODRG or ATB, respectively. The models are evaluated based on root median square error (RMSE):

$$\boldsymbol{RMSE} = \sqrt{\frac{\Sigma_{T}^{t=1}(\hat{y}_t - y_t)^2}{T}} \tag{1}$$

Furthermore we used the normalized RMSE:

$$\boldsymbol{NRMSE} = \frac{\sqrt{\frac{\Sigma_{T}^{t=1}(\hat{y}_t - y_t)^2}{T}}}{\sqrt{(\boldsymbol{min}(y) - \boldsymbol{max}(y))^2}} \cdot \mathbf{100} \tag{2}$$

A direct comparison between the different software tools, respectively, the algorithms, is not possible since the applications are using different force fields. However, the aforementioned metrics enable a direct comparison of the machine learning predictions to the original software.

Molecular Dynamics

We tested the ATB-derived random forest models, with 50 randomly chosen molecules from the ATB database with experimental hydration free energy (ΔG^{hyd}). Topologies and coordinate files were obtained by the ATB database. Parameters for the molecule dynamics were taken from the FreeSolv database (Mobley et al., 2009; Mobley, 2013; Mobley and Guthrie, 2014; Duarte Ramos Matos et al., 2017). We used the *gromos54a7_atb.ff* force field according to ATB. Simulations were run under GROMACS (v. 2016.3) with NPT conditions at 298 K and 1 atm. The cutoff for the van der Waals (rvdw) and electrostatic interactions (rcoulomb) was set to

1.2 nm. The simulations were performed with 20 λ-steps and 2 fs per time step, resulting in 12.5-ns simulations per λ-point. GROMACS simulations require removing all nonpolar hydrogens for a united-atoms model. For ContraDRG, original partial charges from ATB were overwritten with ContraDRG predictions. Therefore, we summarized all removed charges into the adjacent remaining atom. Atom-centered partial charge predictions occasionally generate molecules with an excess of net total charges. The excess was eliminated by distributing the excess equally through a molecule. A comparison of the absolute errors between the experimental ΔG^{hyd} free energy and ATB and that between the experimental ΔG^{hyd} free energy with ContraDRG were performed by a Welch *t* test (Welch, 1947). We omitted MD simulations with PRODRG topologies since it has been reported as inaccurate (Lemkul et al., 2010), which could be confirmed in our analyses.

Web Application

The web application ContraDRG is based on an Apache web server (v. 2.4.29) with PHP (v. 7.2.17) and R (v. 3.4.4) as background services. Incoming data will be filtered and converted by Open Babel (v. 2.4.1) into temporary internal PDB files. ContraDRG reads the PDB structures, performs the feature encoding, and applies the trained machine learning models. The final output will be generated by the Open Babel and remapped with partial charge values predicted by ContraDRG determining partial charge values. A two-dimensional graph of the molecule will be displayed after the machine learning prediction. Missing three-dimensional molecules structures, as provided by SMILES formatted molecules, will be computed by Open Babel as well. The partial charge prediction will be performed by the random forest models for each element, which have been shown to outperform the other models.

RESULTS

Overall Approach

The current study aimed to build a reliable and fast prediction model for partial charges. To this end, we used machine learning algorithms to emulate computationally complex predictions in a reverse engineering–like manner and developed ContraDRG, a software that can be used to predict partial charge assignments based on PRODRG and ATB predictions. We collected thousands of randomly selected molecules from PubChem and the ATB database. Finally, we provide the freely accessible web tool ContraDRG, which can be used for partial charge predictions. The resulting predictions provide a reliable approximation of the original tools. However, predictions are carried out in seconds without any user restrictions.

Datasets

We collected 7,000 molecule structures from PubChem with an average size of 19 heavy atoms per molecule (resulting in 132,859 atoms), which were predicted using PRODRG. Seventy percent of the atoms in the PRODRG dataset are carbon, and 13% are oxygen atoms. Moreover, we randomly collected 10,000 molecules from the ATB database with an average size of 25 heavy atoms per molecule. In this ATB dataset, 47% of the atoms are hydrogens, while 35% are carbons. **Figure 2** represents the distribution of all elements in our datasets. Variances in the number of hydrogen atoms between both datasets are due to differences in the underlying model, namely, united-atoms model for PRODRG and all-atoms model for ATB.

To achieve a high variety of different molecules, we analyzed the similarities between every molecule structure to each other by calculating the Tanimoto coefficient in a pairwise manner. The

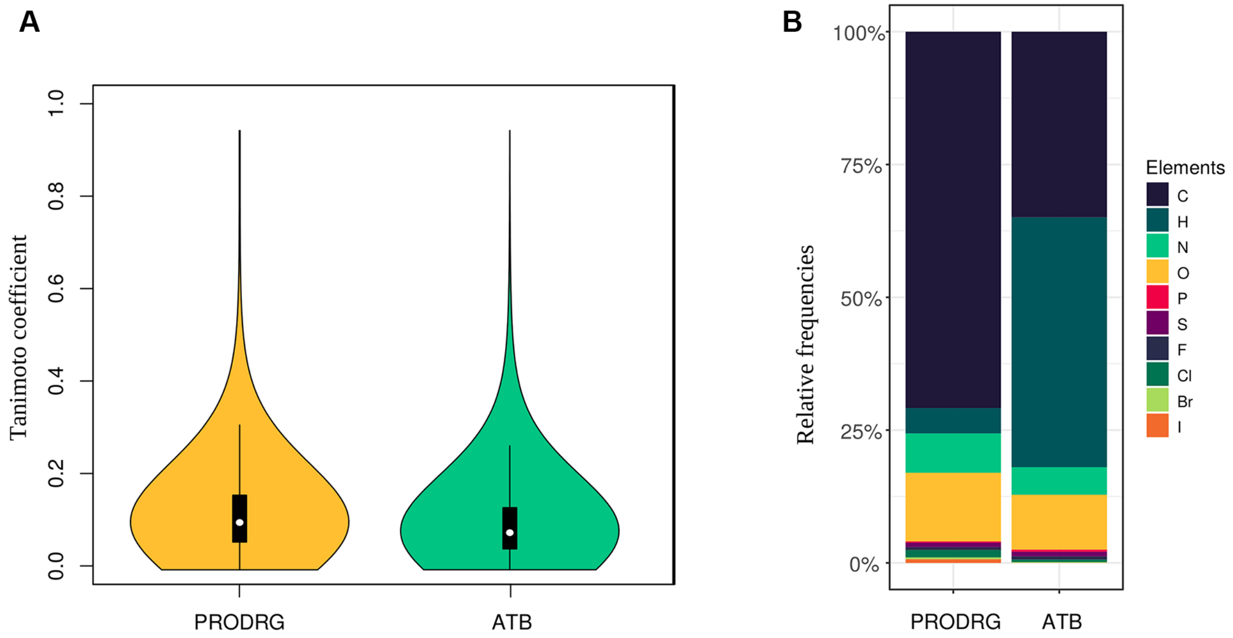


FIGURE 2 | (A) The violin plots show the Tanimoto coefficient for both datasets. The plot width correlates with the relative frequencies of the coefficient. The white dot represents the median, while the black box represents the interquartile range, and the black lines, the 95% confidence intervals. One-sample *t* tests for both sets of Tanimoto coefficients show a $p < 0.001$ for a mean below 0.15. **(B)** The distribution of atom types for each dataset is represented by relative bar plots.

Tanimoto coefficients and their distribution for the PRODRG and the ATB datasets are shown as violin plots in **Figure 2**. The coefficients of all possible pairs of molecules are relatively low, with a median of around 0.11 for the PRODRG and 0.08 for the ATB dataset, indicating a high variance between the incorporated molecules. We used a one-sample *t* test on the Tanimoto coefficients for testing significance against a mean value of 0.15 ($p < 0.001$).

Analysis of the charge distribution through all elements shows a variance in the charge predictions between the different datasets in **Figure 3**. Since the occurrence of molecular constitutions and conformations is limited, the partial charges are not equally distributed over the whole range. Moreover, some atoms tend to act as an electron-pair donor, such as oxygen. Therefore, most oxygen is charged negatively or neutral. Generally, the charge predictions differentiate heavily between the PRODRG and ATB datasets. PRODRGs predictions are more clustered than ATB. This clustering can be observed in the shape of the charge distribution curves by the present peaks of the PRODRG dataset in **Figure 3**. One explanation for the highly clustered charges of PRODRG is the fact that PRODRG maps the molecule to a limited set of building blocks and charge groups, while ATB refines partial charges after an initial determination according to the Merz–Singh–Kollman method (Chandra Singh and Kollman, 1984).

Partial Charge Prediction

We employed several machine learning algorithms for every element on each dataset. Depending on the number of data points, the machine learning algorithm training took several hours up to 10 days on a high-performance cluster, especially for the SVMs and random forest models. Linear regression models turned out to be most inaccurate compared to the random forest models, which mostly outperform all other models in both datasets. For this reason, the ContraDRG web application uses random forest models for the prediction. An exemplary direct side-by-side comparison of ATB-derived ContraDRG prediction with ATB 3.0 is provided in the **Supplementary Material**. For a set of 50 randomly chosen molecules, ATB required an average execution time of 8 h for generating the topology including the partial charges, while ContraDRG required only 9.2 seconds on average for the partial charge prediction per molecule.

Table 1 represents a shortened overview of the best prediction performance. The full-length table is provided in the **Supplementary Material**. The normalized RMSE values allow an easy comparison for each element since they are normalized to the whole range of present partial charge values. Moreover, the predictions for PRODRG-derived data are more accurate than for ATB, which can be observed particularly for underrepresented elements such as iodine in the ATB dataset. The mean R^2 for PRODRG predictions is 0.962 (min. 0.791, max. 1.000) for random forest and 0.685 (0.010–0.985) for SVMs with linear kernel in comparison to the ATB predictions with a mean of R^2 0.908 (0.778–0.982) for random forest and 0.744 (0.520-0.971) for linear SVMs. Overall, the predictions based on the random forest models are more accurate than those based on the other models.

The MD analyses show that the predictions of ContraDRG's ATB-derived random forest models perform as well as ATB in terms of the ΔG^{hyd} free energy calculation. Furthermore, we compared the errors between experimental ΔG^{hyd} values and those derived from ATB with the errors between the experimental data and ATB-derived ContraDRG prediction. No significant differences have been observed by using the Welch *t* test ($p = 0.53$) (Max and Kuhn, 2008). Additional information is provided as **Supplementary Materials**.

DISCUSSION

In summary, we were able to produce partial charge predictions by our fast and unrestricted approach. Depending on the dataset and the frequency of an element in the dataset, reliable predictions are possible. The models for underrepresented elements such as chlorine, bromine, and iodine performed worse compared to those trained on the most abundant elements such as carbon or hydrogen. Surprisingly, linear regression performed better for iodine in the ATB dataset than the corresponding random forest model (see **Supplementary Material**). A possible explanation for that is the fact that iodine atoms are the most underrepresented elements in the ATB dataset, and the random forest models tend to overfit.

Generally, as **Table 1** shows, our predictions for the PRODRG dataset are more accurate than for ATB. There are several possible reasons for that. First, PRODRG is based on a simpler method for assigning partial charges (Altman, 1992). Second, we used molecules from the PubChem database for the PRODRG dataset. The three-dimensional structures of these molecules are all idealized and normalized by PubChem (Bolton et al., 2008). Compared to that, we used curated molecules for the ATB dataset, which mostly originate from the manually curated ChEMBL database (Gaulton et al., 2012; Stroet et al., 2018). Third, ATB performs geometric optimization and remaps the partial charges back to the original structures. Geometry-optimized charges cannot be learned by our model since we do not take geometrical temporary changes into account. Additionally, as shown in **Figure 3**, the partial charges for the ATB data have a higher variance, which makes prediction generally more difficult.

Although our approach is biased to inherit errors from the original tools, the predictions achieve a reliable approximation with low RMSE values. Inconsistent partial charges, which can appear in PRODRG (Lemkul et al., 2010), are unlikely because our models predict the charges along with defined models without determinations of building blocks. Error propagation cannot be avoided; however, by using larger datasets and extended feature sets, the prediction models tend to be more accurate. Our web tool is freely accessible at http://contradrg.heiderlab.de.

CONCLUSION

All existing approaches of partial charges predictions for molecules aim at reconstructing the exact empirical-validated value. Thus, the computations are based on empirical determined data (Mortier et al., 1986; Besler et al., 1990) or on quantum

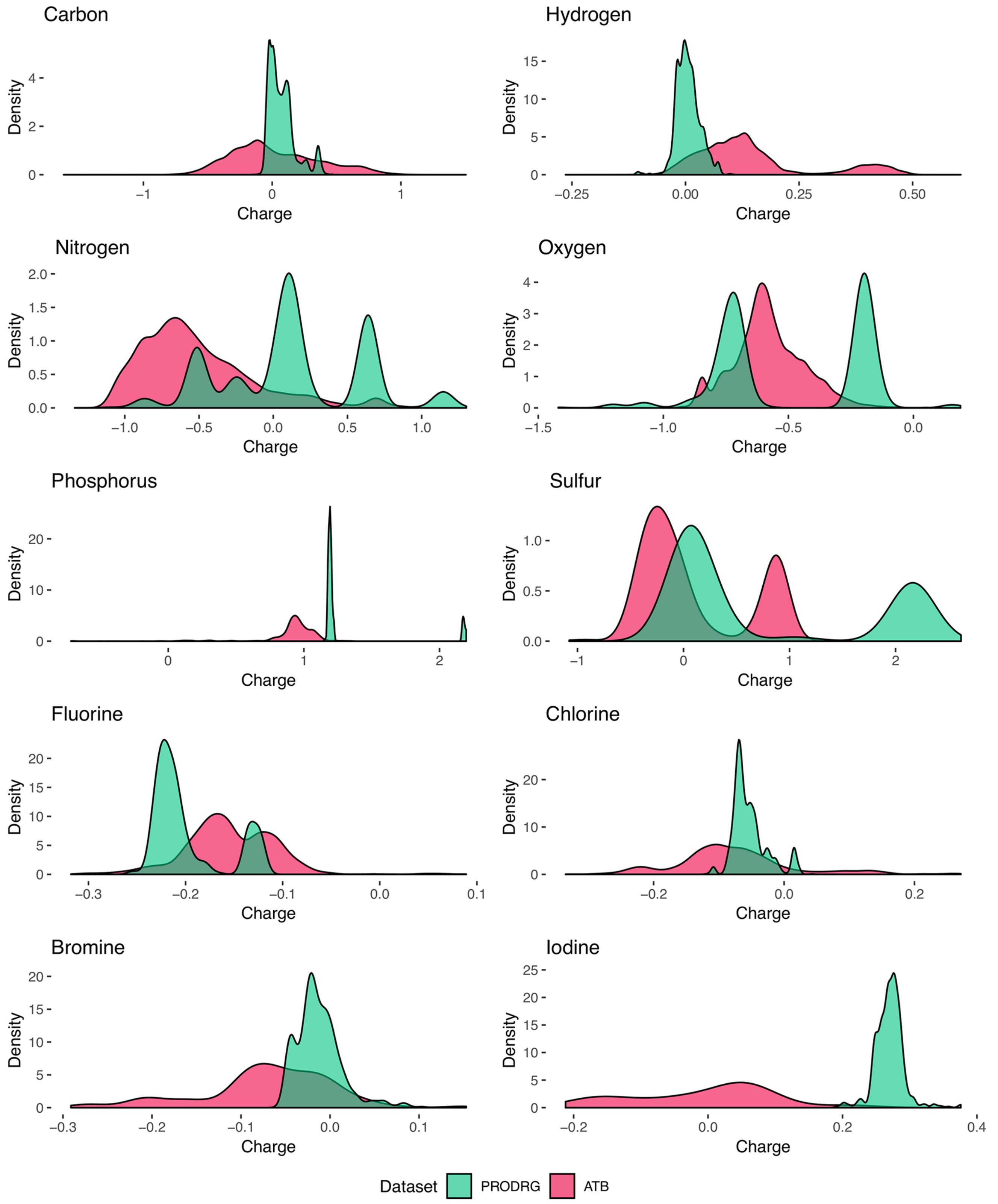


FIGURE 3 | Smoothed kernel density estimates represent the distribution of partial charges (units of *e*) for each molecule in the datasets. Distribution from PRODRGs dataset reveals more clustered peaks (green) than from ATB (red).

TABLE 1 | Performance comparison for partial charge prediction (units of *e*) by random forest and support vector machines with linear kernel of the PRODRG and ATB dataset.

	PRODRG						ATB					
	Random forest			SVM linear			Random forest			SVM linear		
	RMSE	NRMSE	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE	R^2	RMSE	NRMSE	R^2
C	0.011	1.443	0.989	0.054	7.073	0.738	0.069	2.398	0.961	0.152	5.268	0.810
H	0.005	2.878	0.955	0.026	13.924	0.010	0.018	2.313	0.980	0.046	5.794	0.879
N	0.048	1.986	0.990	0.249	10.374	0.730	0.113	5.391	0.919	0.163	7.772	0.834
O	0.051	3.184	0.971	0.153	9.494	0.739	0.047	4.200	0.887	0.071	6.302	0.746
P	0.002	0.152	1.000	0.073	7.157	0.965	0.075	3.712	0.892	0.097	4.803	0.823
S	0.015	0.678	1.000	0.120	5.454	0.985	0.068	3.095	0.982	0.087	3.962	0.971
F	0.003	2.436	0.993	0.007	5.184	0.968	0.017	4.179	0.897	0.037	9.205	0.520
Cl	0.004	2.724	0.980	0.020	15.293	0.415	0.030	5.490	0.895	0.054	9.796	0.705
Br	0.011	8.625	0.791	0.016	12.222	0.589	0.033	8.796	0.778	0.049	13.033	0.531
I	0.004	2.575	0.955	0.010	6.592	0.706	0.036	12.840	0.888	0.062	22.082	0.624
$\bar{x}$)	0.015	2.668	0.962	0.073	9.277	0.685	0.051	5.241	0.908	0.082	8.802	0.744

The root median square error (RMSE) represents the quality of errors while NRMSE shows a normalized RMSE.

mechanical theories (Manz and Sholl, 2010; Manz and Sholl, 2012; Manz and Limas, 2016). However, our approach tries to emulate the algorithm of the predictor without implementing any background knowledge about the underlying theories. Analysis of the input and output data from the web servers with subsequent machine learning approaches are sufficient to easily compute reliable approximations. Our web tool can be used to assign partial charge predictions automatically within seconds. This allows, for example, the correction of precalculated topology files. In the future, we intend to improve our models by using more training data, in particular for those atoms that are underrepresented, and to extend the feature set. Additionally, we intend to generate GROMOS-compatible topology files without geometrical optimization for molecular dynamics simulations.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in http://cdrg.mathematik.uni-marburg.de/data/raw-dataset.zip.

AUTHOR CONTRIBUTIONS

RM performed the data and machine learning analysis. RM drafted the manuscript. DH supervised the project, discussed the results, and revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This study was funded by the European Regional Development Fund, EFRE-Program, European Territorial Cooperation (ETZ) 2014-2020, Interreg V A, Project 41.

ACKNOWLEDGMENTS

We thank the ATB team, particularly Martin Stroet and Alan Mark for providing us access to the ATB database. Calculations were carried out on the MaRC2 high performance cluster of the University of Marburg, which are supported by the State Ministry of Higher Education, Research and the Arts. We especially thank Mr. Sitt of HPC-Hessen for his technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00990/full#supplementary-material

REFERENCES

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *Am. Stat.* 46, 175. doi: 10.2307/2685209

Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminf.* 7, 1–13. doi: 10.1186/s13321-015-0069-3

Bayly, C. I., Cieplak, P., Cornell, W., and Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* 97, 10269–10280. doi: 10.1021/j100142a004

Besler, B. H., Merz, K. M., and Kollman, P. A. (1990). Atomic charges derived from semiempirical methods. *J. Comput. Chem.* 11, 431–439. doi: 10.1002/jcc.540110404

Bleiziffer, P., Schaller, K., and Riniker, S. (2018). Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations. *J. Chem. Inf. Model.* 58, 579–590. doi: 10.1021/acs.jcim.7b00663

Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008). *Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities* Vol. 4. Elsevier B.V, 217–241. Amsterdam, Netherlands. doi: 10.1016/S1574-1400(08)00012-1

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

Chandra Singh, U., and Kollman, Peter A (1984). An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* 5, 129–145. doi: 10.1002/jcc.540050204

Chen, J., and Martı, T. J. (2007). QTPIE: Charge transfer with polarization current equalization. A fluctuating charge model with correct asymptotics. *Chem. Physics Letters* 438, 315–320. doi: 10.1016/j.cplett.2007.02.065

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., et al. (1995). A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* 117, 5179–5197. doi: 10.1021/ja00124a002

Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. *Mach. Learn.* 20, 273–297. doi: 10.1.1.170.5707

D'Alche-Buc, F., and Wehenkel, L. (2008). Machine learning in systems biology. *BMC Proc.* 2 Suppl 4, S1. doi: 10.1186/1753-6561-2-s4-s1

Daura, X., Mark, A. E., and Van Gunsteren, W. F. (1998). Parametrization of aliphatic CHn united atoms of GROMOS96 force field. *J. Comput. Chem.* 19, 535–547. doi: 10.1002/(SICI)1096-987X(19980415)19:5⟨535::AID-JCC6⟩3.0.CO;2-N

Duarte Ramos Matos, G., Kyu, D. Y., Loeffler, H. H., Chodera, J. D., Shirts, M. R., and Mobley, D. L. (2017). Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. *J. Chem. Eng. Data* 62, 1559–1569. doi: 10.1021/acs.jced.7b00104

Dudziński, K., and Walukiewicz, S. (1987). Exact methods for the knapsack problem and its generalizations. *Eur. J. Oper. Res.* 28, 3–21. doi: 10.1016/0377-2217(87)90165-2

Engler, M. S., Caron, B., Veen, L., Geerke, D. P., Mark, A. E., and Klau, G. W. (2019). Automated partial atomic charge assignment for drug-like molecules: a fast knapsack approach. *Algorithms Mol. Biol.* 14, 1. doi: 10.1186/s13015-019-0138-7

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38, 367–378. doi: 10.1016/S0167-9473(01)00065-2

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi: 10.1093/nar/gkr777

Gilson, M. K., Gilson, H. S., and Potter, M. J. (2003). Fast Assignment of Accurate Partial Atomic Charges: An Electronegativity Equalization Method that Accounts for Alternate Resonance Forms. *J. Chem. Inf. Comput. Sci.* 43, 1982–1997. doi: 10.1021/ci034148o

Halgren, T. A. (1999). MMFF VI. MMFF94s option for energy minimization studies. *J. Comput. Chem.* 20, 720–729. doi: 10.1002/(SICI)1096-987X(199905)20:7⟨720::AID-JCC7⟩3.0.CO;2-X

Jakalian, A., Jack, D. B., and Bayly, C. I. (2002). Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* 23, 1623–1641. doi: 10.1002/jcc.10128

Jorgensen, W. L., and Tirado-Rives, J. (1988). The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. *J. Am. Chem. Soc.* 110, 1657–1666. doi: 10.1021/ja00214a001

Jorgensen, W. L., Maxwell, D. S., and Tirado-Rives, J. (1996). Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* 118, 11225–11236. doi: 10.1021/ja9621760

Kaminski, G. A., Friesner, R. A., Tirado-Rives, J., and Jorgensen, W. L. (2001). Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins *via* Comparison with Accurate Quantum Chemical Calculations on Peptides †. *J. Phys. Chem. B* 105, 6474–6487. doi: 10.1021/jp003919d

Kellerer, H., Pferschy, U., and Pisinger, D. (2004). *Knapsack Problems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 548. doi: 10.1007/978-3-540-24777-7

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2018). PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* 47, 1–8. doi: 10.1093/nar/gky1033

Koziara, K. B., Stroet, M., Malde, A. K., and Mark, A. E. (2014). Testing and validation of the Automated Topology Builder (ATB) version 2.0: Prediction of hydration free enthalpies. *J. Comput.-Aided Mol. Design* 28, 221–233. doi: 10.1007/s10822-014-9713-7

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2006). Machine learning in bioinformatics. *Briefings Bioinf.* 7, 86–112. doi: 10.1093/bib/bbk007

Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today* 20, 318–331. doi: 10.1016/j.drudis.2014.10.012

Lemkul, J. A., Allen, W. J., and Bevan, D. R. (2010). Practical Considerations for Building GROMOS-Compatible Small Molecule Topologies. *J. Chem. Inf. Model.* 50, 2221–2235. doi: 10.1021/ci100335w

Malde, A. K., Zuo, L., Breeze, M., Stroet, M., Poger, D., Nair, P. C., et al. (2011). An Automated force field Topology Builder (ATB) and repository: Version 1.0. *J. Chem. Theory Comput.* 7, 4026–4037. doi: 10.1021/ct200196m

Manz, T. A., and Limas, N. G. (2016). Introducing DDEC6 atomic population analysis: Part 1. Charge partitioning theory and methodology. *RSC Advances* 6, 47771–47801. doi: 10.1039/c6ra04656h

Manz, T. A., and Sholl, D. S. (2010). The Electrostatic Potential in Periodic and Nonperiodic Materials. *J. Chem. Theor. Comput.* 6, 2455–2468.

Manz, T. A., and Sholl, D. S. (2012). Improved atoms-in-molecule charge partitioning functional for simultaneously reproducing the electrostatic potential and chemical states in periodic and nonperiodic materials. *J. Chem. Theory Comput.* 8, 2844–2867. doi: 10.1021/ct3002199

Marenich, A. V., Jerome, S. V., Cramer, C. J., and Truhlar, D. G. (2012). Charge model 5: An extension of hirshfeld population analysis for the accurate description of molecular interactions in gaseous and condensed phases. *J. Chem. Theory Comput.* 8, 527–541. doi: 10.1021/ct200866d

Max, K., and Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *J. Stat. Software* 28, 1–26. doi: 10.1053/j.sodo.2009.03.002

Meinshausen, N. (2006). Quantile Regression Forests. *J. Mach. Learn. Res.* 7, 983–999.

Mobley, D. L. (2013). Experimental and Calculated Small Molecule Hydration Free Energies. *UC Irvine Department* 113 (14), 4533–4537.

Mobley, D. L., and Guthrie, J. P. (2014). FreeSolv: A database of experimental and calculated hydration free energies, with input files. *J. Comput.-Aided Mol. Design* 28, 711–720. doi: 10.1007/s10822-014-9747-x

Mobley, D. L., Bayly, C. I., Cooper, M. D., and Dill, K. A. (2009). Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations †. *J. Phys. Chem. B* 113, 4533–4537. doi: 10.1021/jp806838b

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., et al. (2009). Software News and Updates AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* 30, 2785–2791. doi: 10.1002/jcc

Mortier, W. J., Ghosh, S. K., and Shankar, S. (1986). Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *J. Am. Chem. Soc.* 108, 4315–4320. doi: 10.1021/ja00275a013

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. *J. Cheminf.* 3, 33. doi: 10.1186/1758-2946-3-33

Oostenbrink, C., Villa, A., Mark, A. E., and van Gunsteren, W. F. (2004). A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* 25, 1656–1676. doi: 10.1002/jcc.20090

Patel, S., and Brooks, C. L. (2004). CHARMM fluctuating charge force field for proteins: I parameterization and application to bulk organic liquid simulations. *J. Comput. Chem.* 25, 1–15. doi: 10.1002/jcc.10355

Patel, S., Mackerell, A. D., and Brooks, C. L. (2004). CHARMM fluctuating charge force field for proteins: II protein/solvent properties from molecular dynamics simulations using a nonadditive electrostatic model. *J. Comput. Chem.* 25, 1504–1514. doi: 10.1002/jcc.20077

Roos, K., Wu, C., Damm, W., Reboul, M., Stevenson, J. M., Lu, C., et al. (2019). OPLS3e: extending force field coverage for drug-like small molecules. *J. Chem. Theory Comput.* 15, 1863–1874. doi: 10.1021/acs.jctc.8b01026

Santos, P. S., Souza, L. K., Araujo, T. S., Medeiros, J. V. R., Nunes, S. C., Carvalho, R. A., et al. (2017). Methyl*β*-cyclodextrin inclusion complex with *ββ*caryophyllene: Preparation, characterization, and improvement of pharmacological activitie. *ACS Omega* 2, 9080–9094. doi: 10.1021/acsomega.7b01438

Schuler, L. D., Daura, X., and Van Gunsteren, W. F. (2001). An improved GROMOS96 force field for aliphatic hydrocarbons in the condensed phase. *J. Comput. Chem.* 22, 1205–1218. doi: 10.1002/jcc.1078

Schüttelkopf, A. W., and Van Aalten, D. M. F. (2004). PRODRG: A tool for high-throughput crystallography of protein–ligand complexes. *Acta Crystallographica Section D:. Biol. Crystallogr.* 60, 1355–1363. doi: 10.1107/S0907444904011679

Scott, W. R. P., Hünenberger, P. H., Tironi, I. G., Mark, A. E., Billeter, S. R., Fennen, J., et al. (1999). The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* 103, 3596–3607. doi: 10.1021/jp984217f

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi: 10.1093/nar/gki458

Stroet, M., Caron, B., Visscher, K. M., Geerke, D. P., Malde, A. K., Mark, A. E. (2018). Automated Topology Builder Version 3.0: Prediction of Solvation Free Enthalpies in Water and Hexane. *J. Chem. Theory Comput.* 14, 5834–5845. doi: 10.1021/acs.jctc.8b00768

Van Aalten, D. M. (1996). PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *J. Comput.-Aided Mol. Design* 10, 255–262. doi: 10.1007/BF00355047

van Gunsteren, W. F., Billeter, S., Eising, A. A., Hunenberger, P. H., Krüger, P., Mark, A. E., et al. (1996). "*Biomolecular Simulation*," in *The GROMOS96 Manual and User Guide* (Zürich, Switzerland: Vdf Hochschulverlag an der ETH Zürich), 30, 1–1042.

Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A., and Case, D. A. (2004). Development and testing of a general Amber force field. *J. Comput. Chem.* 25, 1157–1174. doi: 10.1002/jcc.20035

Wang, J., Wang, W., Kollman, P. A., and Case, D. A. (2006). Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Model.* 25, 247–260. doi: 10.1016/j.jmgm.2005.12.005

Weininger, D. (1988). SMILES, a chemical language and information system: 1: introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36. doi: 10.1021/ci00057a005

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika* 34, 28–35. doi: 10.1093/biomet/34.1-2.28

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.

ORIGINAL RESEARCH
published: 04 November 2019
doi: 10.3389/fgene.2019.01079

HiCeekR: A Novel Shiny App for Hi-C Data Analysis

*Lucio Di Filippo[1], Dario Righelli[2], Miriam Gagliardi[3,4], Maria Rosaria Matarazzo[4] and Claudia Angelini[2]**

[1] Telethon Institute of Genetics and Medicine (TIGEM), Pozzuoli, Italy, [2] Istituto per le Applicazioni del Calcolo "Mauro Picone," Consiglio Nazionale delle Ricerche, Napoli, Italy, [3] Max Planck Institute for Psychiatry, Munich, Germany, [4] Institute of Genetics and Biophysics "A. Buzzati A. Traverso," Consiglio Nazionale delle Ricerche, Napoli, Italy

The High-throughput Chromosome Conformation Capture (Hi-C) technique combines the power of the Next Generation Sequencing technologies with chromosome conformation capture approach to study the 3D chromatin organization at the genome-wide scale. Although such a technique is quite recent, many tools are already available for pre-processing and analyzing Hi-C data, allowing to identify chromatin loops, topological associating domains and A/B compartments. However, only a few of them provide an exhaustive analysis pipeline or allow to easily integrate and visualize other omic layers. Moreover, most of the available tools are designed for expert users, who have great confidence with command-line applications. In this paper, we present HiCeekR (https://github.com/lucidif/HiCeekR), a novel R Graphical User Interface (GUI) that allows researchers to easily perform a complete Hi-C data analysis. With the aid of the Shiny libraries, it integrates several R/Bioconductor packages for Hi-C data analysis and visualization, guiding the user during the entire process. Here, we describe its architecture and functionalities, then illustrate its capabilities using a publicly available dataset.

Keywords: Hi-C, user-friendly interface, long-range interactions, genome organization, topologically associating domains

Edited by:
Dominik Heider,
University of Marburg,
Germany

Reviewed by:
Pietro Lio',
University of Cambridge,
United Kingdom
Pao-Yang Chen,
Academia Sinica,
Taiwan

****Correspondence:***
Claudia Angelini
claudia.angelini@cnr.it

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

Received: *19 July 2019*
Accepted: *09 October 2019*
Published: *04 November 2019*

Citation:
Di Filippo L, Righelli D, Gagliardi M, Matarazzo MR and Angelini C (2019) HiCeekR: A Novel Shiny App for Hi-C Data Analysis. Front. Genet. 10:1079. doi: 10.3389/fgene.2019.01079

INTRODUCTION

The DNA is organized in a three-dimensional (3D) structure inside the cell nucleus, where chromosomes occupy distinct regions called chromosome territories. Within chromosome territories, the chromatin forms Topological Associated Domains (TADs) characterized by a high frequency of intra-domain loci interactions. Inside the TADs, chromatin loops contain active genes and are physically separated from repressed domains. Investigating the 3D organization of chromatin is important to better understand the higher-order regulation of gene expression and, more in general, the genome functionality.

In the last twenty years, the advent of modern high-throughput technologies has allowed investigating chromatin structure and its hierarchical organization from an individual gene location to the global genome-wide perspective, using either method based on microscopy, such as fluorescent *in situ* hybridization (Solovei et al., 2002), and/or those based on chromosome conformation capture and their evolution. In particular, the original Chromosome Conformation Capture (3C) technique (Dekker, 2002), defined as *One-By-One* approach, enabled to study the 3D chromatin interaction between one region of interest and another single locus that is distant in the linear genome. Over the years, it was improved to expand the number of genomic regions studied in each experiment. Therefore, the Circular Chromosome Conformation Capture (4C) (Zhao et al., 2006) technique was proposed to investigate one locus of interest against all others (i.e. *One-By-All* approach), and later, the Chromosome Conformation Capture Carcon Copy (5C) (Dostie et al., 2006) allowed studying the interactions between multiple sequences (i.e. *Many-By-Many* approach). More recently, by combining proximity-based ligation with massively parallel sequencing, the High-throughput Chromosome Conformation Capture (Hi-C) (Belton et al., 2012; Dekker et al., 2013) allows to simultaneously investigate all genome interactions, therefore providing the *All-By-All* approach. Thanks to Hi-C experiments, it is now possible to study long-range interactions, i.e. physical interactions between chromosomal regions linearly distant that occupy the same spatial location in 3D chromatin conformation, identify chromatin hierarchical structures, and provide high-resolution 3D images of the chromatin architecture and its changes associated to diseases or treatments. However, to comprehensively explore the chromatin structure and its state, the integration of Hi-C results with the global epigenetic landscape is required. Due to the huge amount of data produced during Hi-C experiments, complex work-flows, and sophisticated computational algorithms are necessary to extract information and support the researchers in the interpretation of their computational results. Furthermore, these workflows need to be adapted, in terms of resolution and algorithms, to the specific structures of interest, see Nicoletti et al. (2018); Pal et al. (2019) for general views.

The first step of the data analysis consists of the alignment of the raw reads on a reference genome. However, due to the presence of DNA fragments originated from two distinct genomic loci, that are combined during ligation, the two mates are usually aligned independently and the mapper often requires to incorporate an iterative procedure to better identify the ligation junction. Tools such as HiCUP (Wingett et al., 2015) or the iterative approach described in Imakaev et al. (2012) can be used, instead of classical short-read mappers. The alignment step produces Binary Alignment Map (BAM) files containing the genomic coordinates of each aligned read on the chosen genome. Such files need to be filtered to remove spurious sequences, PCR duplicates, digestion or ligation artifacts, low-quality sequences, and any other sources of technical noise from the sequences of interest.

The analysis is then carried on the retained high-quality sequences. The reference genome is divided into small regions (called bins), that are used to evaluate a square symmetric matrix (known as raw contact matrix) by counting the number of paired-end reads inside each pair of bins. Such a step is often referred to as binning and the contact matrix measures the strength of the interaction between two bins (i.e. the rows and the columns of the contact matrix). The bin width defines the resolution of analysis and, as a consequence, the computational time and the resources required to perform the analysis. The choice of the resolution depends on the organism under investigation, the sequencing depth, the size of the restriction fragment, as well as the available computational resources.

Subsequently, the contact matrix has to be normalized to mitigate bias effects typically present in this type of analysis. Normalization is a crucial step that can have a strong effect on the results (Ay and Noble, 2015). Some normalization algorithms were proposed in Yaffe and Tanay (2011); Hu et al. (2012); Imakaev et al. (2012); Knight and Ruiz (2013). The normalized contact matrices are useful for visualization and are used for further downstream analysis.

The post-processing or downstream analysis defines a wide series of computational procedures aimed at identifying and extracting hierarchical chromatin structures of interest. For example, it is possible to partition the genome in compartments, usually denoted as A and B compartments. Such domains are usually located along the same chromosome and display strong interactions within the same domain and negligible interactions with the other domains. It has been shown that such compartments are connected to active and inactive chromatin states, respectively, and can be related to regions of (gene-dense) euchromatin and regions of (gene-poor) heterochromatin. Compartments are usually identified at a resolution of 100 Kbp or higher. Moreover, by looking at the block-wise structure of the contact matrix, contiguous regions of high self-interactions clearly separated from adjacent regions can be identified. Such regions are usually referred to as tad and the separation boundaries determine their coordinates. tad are usually identified with a resolution of 50 Kbp or higher. Several methods have been proposed for identifying tad boundaries, see Zufferey et al. (2018). With higher-resolution analysis, it is possible to identify specific point-to-point interactions usually referred to as loops. Such interactions can be either *cis*-interactions or *trans*-interactions and appear as spike signals in the contact map. Loops are usually identified with a resolution of 10 Kbp.

Finally, it is also helpful to integrate hic data with other experimental genome-wide datasets [i.e. Chromatin Immunoprecipitation Sequencing (ChIP-Seq) or RNA sequencing (RNA-Seq)] or with other information from an external database to support the researcher in interpreting experimental data, provide evidence of specific regulatory mechanisms and/or insight for novel research hypotheses.

In the last few years, several computational approaches have been proposed to either to perform one or few of the above-mentioned steps or to combine them in more general pipelines. From one hand, the interesting comparative study made in Forcato et al. (2017) provided a clear and detailed description of the advantages and drawbacks of individual methods/algorithms. Indeed, after bench-marking several procedures using different quality indexes, Forcato et al. (2017) showed that several methods reported good performance on some specific steps, although no methods outperformed the others. On the other hand, despite the great effort in the development of tools specifically designed for the analysis of Hi-C, they rarely include all the required functionalities for complete analysis in a single platform. Han and Wei (2017) and Calandrelli et al. (2018) provided a recent list of existing general-purpose tools. In general, most of the available tools are designed for expert users with great confidence about command-line applications. As a consequence, they are not supporting user-friendly data explorations that can lead experimental biologists to easily interpret their results, confirm, or make novel scientific hypotheses. These motivations led us to the development of HiCeekR, a novel computational tool that allows performing most of the above-mentioned steps, through an easy user-friendly graphical interface, combining different algorithms for the analysis of Hi-C data. Moreover, HiCeekR has been designed for guiding the users during the entire analysis process and to provide interactive plots that might help researchers with limited experience in command-line applications, to explore and visualize data and results using a simple *point-and-click* approach.

MATERIALS AND METHODS

In this section, we first describe HiCeekR workflow, then we provide technical details about its implementation and the structure of the Graphical User Interface (GUI). Finally, we illustrate how HiCeekR stores input/output data and results, and describe the internal modular architecture.

HiCeekR Workflow

HiCeekR is a novel Shiny based R package (https://github.com/lucidif/HiCeekR) for Hi-C data analysis. Thanks to its GUI, HiCeekR friendly guides the user during the entire analysis process, allowing him/her to perform a complete data analysis pipeline and to integrate Hi-C data with other omic datasets. Moreover, HiCeekR produces several interactive graphics that allow exploring the results by the usage of the mouse pointer.

As shown in **Figure 1,** HiCeekR analysis starts from already aligned sequence files (in BAM format) obtained from Hi-C experiments, it proceeds through a series of steps from pre-processing and filtering, to the evaluation and normalization of the contact matrices. Once the contact matrices are available, the user can perform the downstream analysis. In particular, HiCeekR allows the identification of genome compartments and tad, the integration of Hi-C data with other omic datasets, such as ChIP-Seq and/or RNA-Seq, the functional analysis, and the visualization of the interaction network. Overall, HiCeekR supports the user in elucidating the functional interplay between chromatin structure and gene regulation by combining and making friendly available a wide bunch of computational and statistical methods.

Through HiCeekR, each step/function can be executed sequentially in a step-by-step analysis (**Figure 1**). After each step, the user can visualize intermediate results, such as summary statistics or graphical representations. However, each step or function can be re-executed by modifying the parameter settings, obtaining consequently updated results. Intermediate and final results (as text files or figures) are stored in pre-organized data structures (see *Data Format and Data Organization*) that can be easily retrieved for future investigations through the HiCeekR GUI.

Pre-Processing

The pre-processing consists of a series of fundamental operations required for the proper execution of HiCeekR. Such operations allow HiCeekR to easily access the information in the subsequent steps and are aimed to reduce the overall execution time. In HiCeekR, the pre-processing is jointly performed

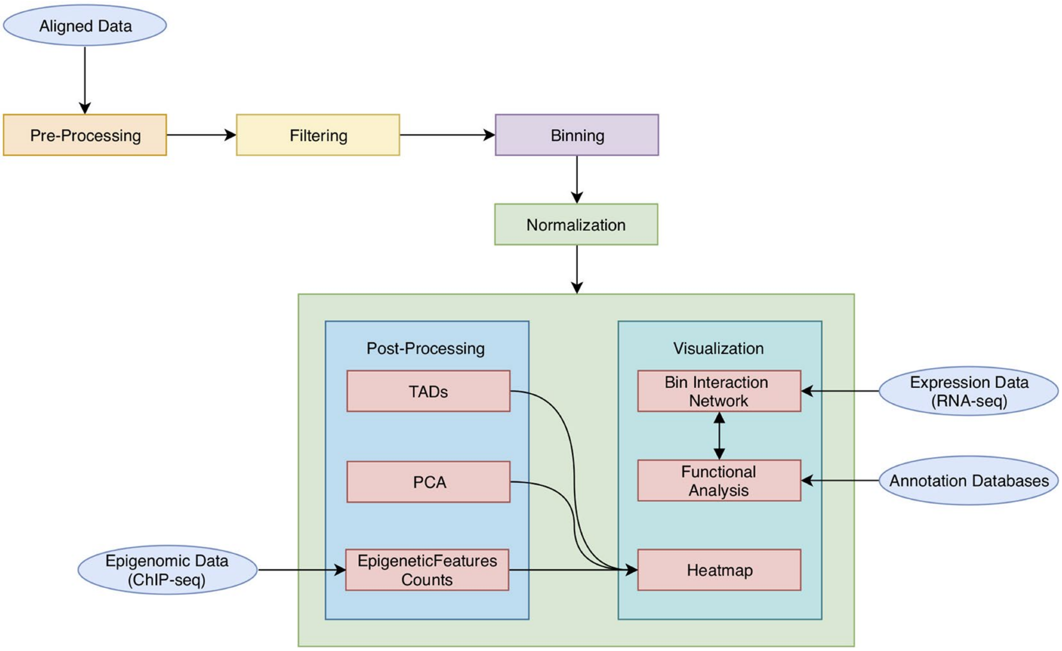


FIGURE 1 | A schematic representation of HiCeekR pipeline. Starting from aligned data, HiCeekR enables to pre-process and filter them to compute (and normalize) the contact matrix. Afterward, it performs several downstream analysis steps in order to detect genome compartments, TADs. Moreover, it also allows the integration of additional epigenetic and transcriptional whole genome datasets, as well as other genome-wide tracks. Finally, it presents the results in interactive graphical forms.

with the creation of a new project (see *Getting Started for the Analysis*), when the user selects the experimental Hi-C files (in BAM format) to work on and the reference genome (in *FASTA* format). At this stage, it is also required to provide the restriction enzyme cutting site and an overhang parameter (in base pairs) that are necessary to split the genome in restriction fragments. The overhang parameter defines the number of base pairs overlapping the restriction enzyme cutting site. Given such information, the restriction fragments are indexed. The coordinates of each detected restriction enzyme cutting site are stored in an index-file (HDF5 file) and associated with one or more mapped read allowing to speed up further computations. The HDF5 file format (https://www.hdfgroup.org/solutions/hdf5/) is chosen for speeding-up heterogeneous data storage and processing, and it is not usually meant to be inspected by a standard user. Note that at this stage, low-quality reads are automatically removed.

At the end of the pre-processing, HiCeekR produces a summary of the statistics for the indexed reads and two diagnostic plots (see **Figures 2A, B**—before filtering) useful to detect artifacts that will be removed during the filtering step. The first plot represents a distribution of the insert lengths over the entire genome, the second shows the distribution of the inward-outward insertion lengths (see *Filtering* for further details).

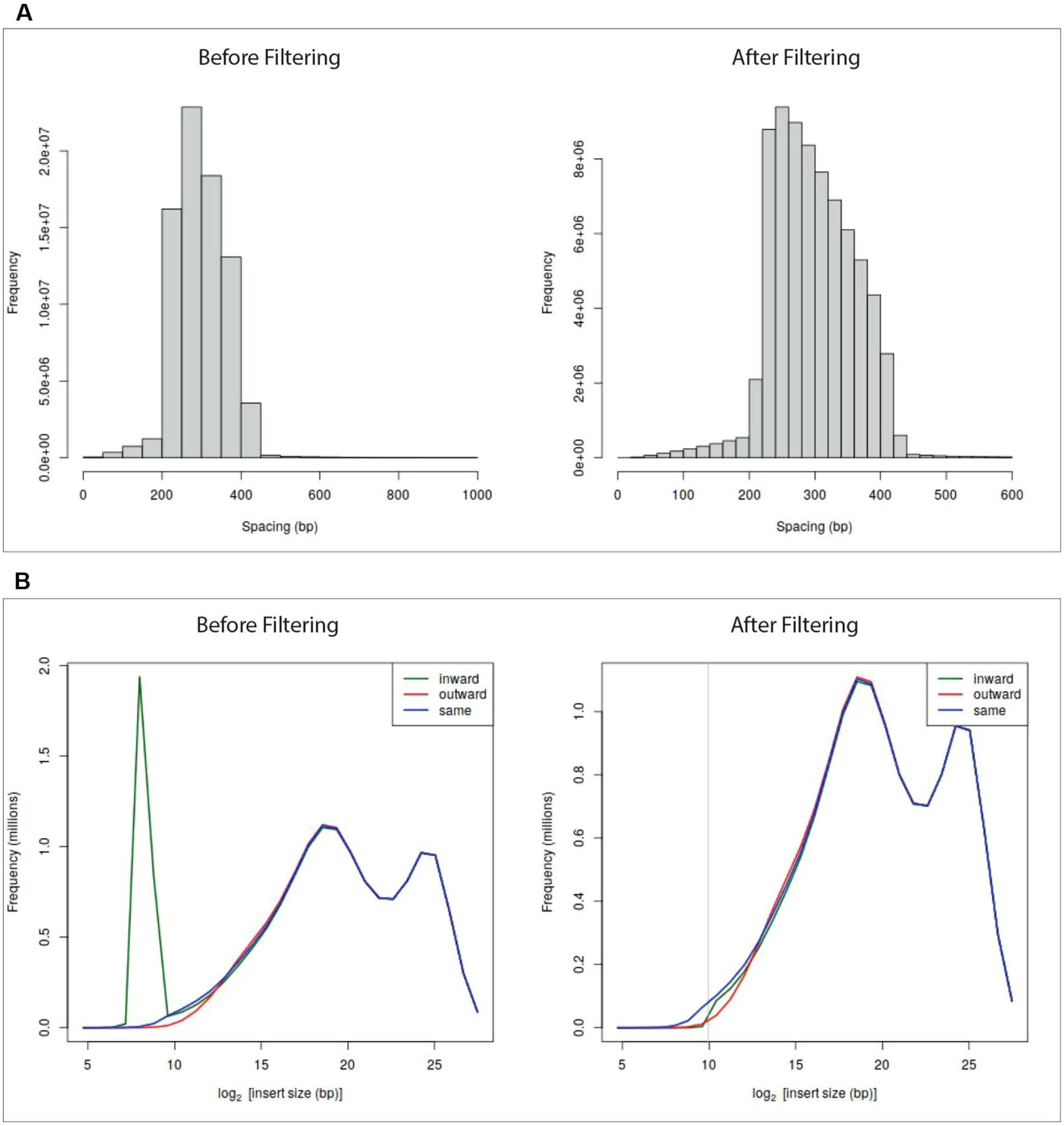


FIGURE 2 | Diagnostic plots and effect of the filtering on sample GSM1608509 (see *A Case Study*). Panel **(A)** shows read length distribution before and after filtering. The plot before filtering indicates that long fragments are present, the corresponding plot after filtering shows that fragments larger than 600 bp were removed. Panel **(B)** shows the read-orientation plot before and after filtering. The plot before filtering suggests possible dangling-end events (green line spike) located at about 2^8 = 256 bp, the corresponding plot after filtering shows that such inward-oriented pair of reads were removed.

Additionally, during the pre-processing, HiCeekR defines the resolution of the entire analysis by the selection of the bin size (default is 6) base pairs), that is used afterward during the binning step.

Filtering

The filtering step is aimed to remove well-recognized artifacts that are produced during library preparation, such as PCR-artifacts, self-circle, and dangling-end fragments (Belton et al., 2012; Ay and Noble, 2015; Lajoie et al., 2015).

In particular, HiCeekR automatically removes PCR duplicates, when previously marked in the BAM files. Marking duplicates can be easily carried out using standard tools.

The identification of self-circle and dangling-end fragments is obtained from the association between read-pairs and restriction fragments that can lead to a two case scenario: the read-pair is associated to different restriction fragments or the same restriction fragment. The former case constitutes the set of valid reads, while the latter occurs when un-ligated dangling-end or circularized self-circle fragments are present into the library preparation. Self-circle (outward strand orientation) and dangling-end (inward strand orientation) fragments can be discriminated each other by looking at the strand orientation of the paired-reads that fall in the same restriction fragment. Since such read-pairs are considered uninformative, they are removed during the filtering step.

HiCeekR removes self-circle and dangling-end fragments by setting a minimum distance for inward pair reads and outward pair reads (*min-inward* and *min-outward* values). It calculates the distance of each associated read from the nearest restriction enzyme site and then estimates the length of the sequencing fragment. Very long fragments, that could be associated with unwanted ligation products, can also be removed by setting a suitable threshold through the *max-frag-length* parameter. By inspecting the diagnostic plot in **Figures 2B**—before filtering), the user can select the *min-inward* and *min-outward* values to remove self-circle and dangling-end products (Lun and Smyth, 2015).

After the filtering process, HiCeekR updates the diagnostic plots (**Figure 2**—after filtering). Results are stored in an HDF5 format.

Binning

The binning step is aimed to perform all those operations required to evaluate the raw contact matrix (Ay and Noble, 2015). To this purpose, the reference genome is divided into n_b bins of approximately non-overlapping and fixed-width w_b (fixed-size bin). Indeed, the exact bin subdivision depends on the locations of the restriction enzyme cutting sites, and few bases of overlap might be allowed between consecutive bins. We recall that the bin size w_b determines the resolution of the analysis (also the resources and the required running time). It is important to select w_b to guarantee good statistical power at an affordable computational cost. Unfortunately, there are no precise guidelines for the selection of w_b, since its choice depends on the sequencing depth and the type of chromatin structure of interest. For these reasons, HiCeekR allows the user to perform the computational analysis at different resolutions, suggesting to first use a low resolution to obtain a general view of the chromatin organization and then repeating and refining the analysis by increasing the resolution while focusing on specific genomic locations of interest (for example, a specific chromosome, or a specific sub-region or two sub-regions located on different chromosomes).

After the bins indexing, HiCeekR assigns the previously filtered-in reads to the genome bins where they better map. Then, it produces the raw contact matrix, a symmetric square matrix $M \in R^{n_b \times n_b}$, by counting the number of reads $M_{i,j}$ that fall within the bins i and j, respectively. To facilitate data exploration, the indexed bins are automatically converted into genomic coordinates. By exploring the raw contact matrix, it is common to observe bins with very large/small values that appear as "outliers" and might due to noise such as low mappability or the presence of many repeated sequences. To reduce this problem, it could be useful to remove "outliers" bins by using a bin-level filtering strategy, as suggested by Lajoie et al. (2015). However, such "outliers" bins can be detected in different ways (Lajoie et al., 2015). The current version of HiCeekR does not implement any bin-level filtering, although we plan to integrate such functionality in future releases.

At the end of binning, HiCeekR stores the bins genomic coordinates as a BED file format and the entire count matrix as a Tab Separated Valuer (TSV) file.

Normalization

The normalization step is aimed to remove technical bias from the raw contact matrix that could lead to false positive/negative findings. The output of such step is a normalized contact matrix, a symmetric square matrix $\hat{M} \in R^{n_b \times n_b}$ of real values, that constitutes one of the main results of the computational data analysis. The current release of HiCeekR implements two different strategies for normalizing the contact matrix: the iterative correction and eigenvector decomposition (ICE) (Imakaev et al., 2012), and the WavSiS (Shavit and Liò, 2014).

ICE is a well-known correction method based on the assumption that the bias in the interaction between two loci can be factorized as the product of the individual biases, affecting each of the two interacting loci (Imakaev et al., 2012). By using such matrix factorization approach, ICE method applies an iterative decomposition algorithm based on the maximum likelihood to convert the raw contact matrix into a normalized one of relative contact probabilities, guaranteeing equal visibility for each region. In particular, the ICE method gives the possibility to Winsorize the matrix to mitigate the effect of the impact of high-abundance bin pairs by using the *Winsor.high* parameter, in combination with the *ignore.low* parameter to not ignore the low abundance bins.

WavSis removes noise by inspecting the variance distribution of the coverage across different physical scales, stabilizing the variance, and applying a wavelet denoising strategy. In particular, the raw contact matrix M (whose entries $M_{i,j}$ are assumed to follow a Poisson distribution) is regarded as a series of decomposed vector coefficients (whose number depends on the number of chromosomes), using the Haar-Fisz transform, which helps in stabilizing the variance. After that, a Gaussian wavelet shrinkage method is used to remove the noise from each set of coefficients and the normalized matrix is reconstructed by inverting the transform. This method is performed independently on each chromosome (selected through the *chromosome of interest* select-box). Additionally, it is possible to remove uncovered regions (detected during this normalization phase) with *NA* values, by using the *remove uncovered* checkbox.

At the end of this process, HiCeekR generates a new tsv file with the normalized count matrix.

Post-Processing

HiCeekR post-processing or downstream analysis supports the user in extracting chromatin structures from the raw or normalized contact matrix and interpreting the results in multiple ways: the detection of A/B-compartments and TADs, the integration with other omic-layers, and the functional interpretation, respectively. These functionalities are available to the user through the modules PCA, directionality index, TopDomTADs, HiCsegTADs, EpigeneticFeatures, and bed2track (in the Post-processing panel), Heatmap, and Network (in the Visualization panel).

HiCeekR detects A/B compartments thanks to the PCA module that performs the principal component analysis (PCA). Large-scale interaction patterns can be identified from the normalized contact matrix by computing the preferential interacting regions (the so-called, compartment A and compartment B). The compartments can be identified by looking at the PCA eigenvector with opposite signs (Lieberman-Aiden et al., 2009; Lajoie et al., 2015). This step requires to select the normalized contact matrix and outputs the PCA eigenvectors (stored as PCA eigenvector matrix) that can be used either to define compartments and for visualization purposes (**Figure 6**). Usually, the first one or two PCA eigenvectors are sufficient to identify the compartments.

Current version of HiCeekR highlights the TADs using three approaches: i) directionality index, ii) TopDom, and iii) HiCseg.

The directionality index module computes the directionality index d_i, as introduced by Dixon et al. (2012). d_i is defined as

$$d_i = \left(\frac{b_i - a_i}{|a_i - b_i|} \right) \left(\frac{(a_i - e_i)^2}{e_i} + \frac{(b_i - e_i)^2}{e_i} \right), \qquad i = 1,....,n_b$$

where a_i and b_i denote the number of mapped reads in the upstream and in the downstream of bin w_i, respectively, and $e_i = \frac{a_i + b_i}{2}$. The directionality index d_i generates a segmentation of the genome, and the TADs are defined as the regions between two sharp changes of directions in such indexes.

The TopDomTADs module implements TopDom algorithm, as proposed in Shin et al. (2015). In particular, it defines a segmentation of the genome based on a three steps procedure: it evaluates the contact frequency signal as the average contact frequency of each bin with its upstream or downstream regions, then selects potential TADs boundaries as the local minima of the contact frequency signal, finally it filters out potential false positive by using Wilcox Rank Sum test under the assumption that the expected contact frequencies of regions within a TADs should be higher than those of a bin in the TADs and a bin outside the TADs, and of those bins outside the TADs. The number of bins to be included in upstream or downstream regions can be controlled by the user with the parameter *Window Size*, which constitute the only tuning parameter of TopDom algorithm.

The HiCsegTADs module implements HiCseg algorithm, as proposed in Lévy-Leduc et al. (2014). In particular, it defines a partition on the contact matrix (either the raw matrix M or the normalized contact matrix $\hat{M}$) with a block structure depending on the unknown TADs boundaries. The parameters of the distributions are estimated by a maximum likelihood approach assuming that the observed contact values, $M_{i,j}$ or $\hat{M}_{i,j}$, within the same TADs share the same distribution parameters. Maximum likelihood estimates are obtained using a dynamic programming algorithm. In this context, Gaussian distributions have to use for modeling normalized contact matrix $\hat{M}$, whereas Poisson or Negative binomial distributions for raw contact matrix M. The user can also choose the maximum number of TADs with the parameter *Kmax* and the structure (i.e. block-diagonal or extended-black diagonal) of the matrix segmentation.

At the end of the TADs processing, HiCeekR automatically generates output files as directionality index track (as a coverage file), and the detected TADs boundaries (in standard BED format). Note that for all modules, the identification of compartments and TADs is performed independently for each chromosome.

As already mentioned, one of the advantages of HiCeekR is given by the possibility to integrate and visualize Hi-C data together with other omic data. To this purpose, in the EpigeneticFeatures module, it is possible to upload one or more aligned BAM files from ChIP-Seq experiments. Then, HiCeekR computes the normalized coverage at the same bin-width resolution chosen for the current Hi-C analysis. Mimicking classical ChIP-Seq coverage, the normalized coverage can be computed either as the number of reads within the bin per million of mapped reads (RPM) or the ratio of the number of reads within the bin in the ChIP-Seq sample over those in the input DNA sample. Additionally, with the bed2track module, it is also possible to process any other genome-wide track in BED format. Such track will be converted by HiCeekR in bin coordinates (i.e. the bin coordinates will be included in the converted track when they intersected the user supplied BED track) to be visualized.

Note that, thanks to the Heatmap module, the user can visualize the normalized contact matrix, the PCA loadings, and/or the directionality index d_i, and/or any bed track (such as those provided as output by TopDomTADs or HiCsegTADs, or converted from user supplied tracks using bed2track), then can add one or more ChIP-Seq coverage tracks to have a more detailed overview of the chromatin state (**Figure 6**).

Finally, in the Network analysis, HiCeekR automatically retrieves the list of genes located within a specific compartment, TADs, or regions of interest. The annotation is obtained overlapping the bins coordinates of the region of interest with the genomic coordinates of the genes (as provided in an annotation file). To this purpose, note that a given bin might be associated with several genes (if the bin overlaps the gene body of more genes), or a given gene might be associated with multiple bins if its gene body is larger than the bin resolution or it overlaps any bin boundary. There are bins not containing genes. The gene-bin association map depends on the annotation and the resolution of the analysis. HiCeekR provides three interactive tables Interaction, Genes, and Enrich. Interaction is a table that contains, for each pair of interacting bins, the corresponding genomic coordinates, the interaction strength, the names of the genes therein contained (if any), and few other information. The gene symbols are hyperlinked to GeneCards (https://www.genecards.org) to facilitate the data interpretation. In the Enrich table, the results of the functional analysis on the identified genes carried out using gProfiler are shown (Raudvere et al., 2019). Identified enriched GO terms, or KEGG and Reactome Pathways are reported together with enriched regulatory motifs/transcription factors (from TRANSFAC), tissue specificity (from human protein Atlas database), Human-specific phenotypes (from Human Phenotype Ontology Database), protein complexes 267 (from CORUM) and results from other interrogated databases. Genes is a table that, among several other information, allows to visualize the gene expression values of the identified genes (only if the user uploaded a gene expression dataset either from RNA-Seq or microarray experiments) that can help in better discriminating chromatin states.

Visualization

It is well known that the visualization of information in a graphical form constitutes one of the most important data exploration tools. However, visualizing Hi-C data can be challenging due to the high-dimensionality of the files and the dimension of the genome. Nowadays, several visualization tools are available, see Yardimci and Noble (2017) for a general review. Nevertheless, HiCeekR provides functions to visualize the obtained results without requiring additional software. Moreover, most of the HiCeekR plots are interactive. In particular, the user can select two main representations: *Heatmap* and *Network* (**Figure 3**).

Using the *Heatmap visualization* the user can explore the raw and the normalized contact matrix using the classic heatmap graphical representation where low and high contact values are depicted using different color intensities. He/she can select a specific chromosome or a pair of chromosomes or, otherwise, a region of interest within each of them. Moreover, it is possible to zoom in/out or to move to another region of interest. Additionally, in the *Heatmap visualization*, the user can add several other genome-wide *tracks* that allow to simultaneously visualize multiple information, such as the loadings of the PCA, the directionality index d_i, any BED format track (i.e. generated by the TADs modules or converted by bed2track module) as well as other omic profiles, such as ChIP-Seq profiles, on the same genome-wide scale, as shown in **Figure 6**.

Using the *Networks visualization* the users can visualize the interactions of a set of bins of interest against all other bins in network form, where the vertices represent the bins and the edges represent the detected interactions. Moreover, the link width is proportional to the strength of the interaction. Additionally, by using user-defined cut-offs, it is to possible to filter-out negligible interactions.

Implementation

HiCeekR is an R-Shiny web GUI which combines several R/Bioconductor packages widely used for Hi-C data analysis and visualization functionalities. In particular, the filtering and the binning steps are implemented using diffHic package (Lun and Smyth, 2015), one of the most used tools for this type of data. Matrix normalization is carried out using ChromeR package (Shavit and Lio', 2014) for the WavSis method and diffHic for the ICE algorithm. The downstream analysis is based on HiTC for the PCA and for the directionality index modules (Servant et al., 2012), TopDom for the TopDomTADs module, HiCseg for the HiCsegTADs module, gProfileR for functional enrichment, and other customized R functions. The graphical output is produced using the ggplot2, plotly, heatmaply, networkD3, and corrplot packages.

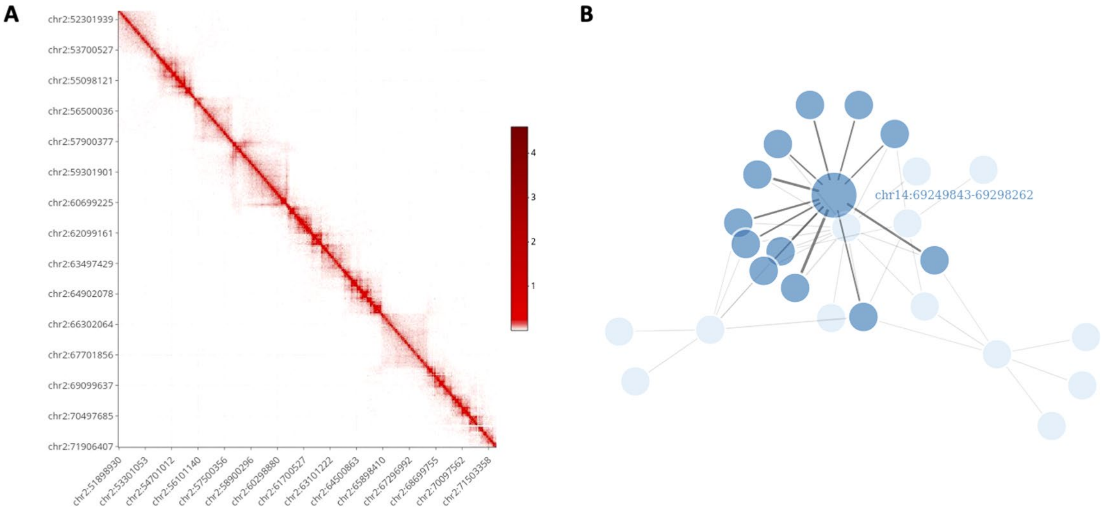


FIGURE 3 | HiCeekR graphical output. **(A)** Heatmap representation on the contact matrix. **(B)** Network representation of selected contacts.

Finally, from the architectural point of view, HiCeekR is open-source, easily expandable with additional functionalities (thanks to the modular structure) and it also allows to integrate third-party functions, as discussed in *Shiny, Modules, and Other Technical Considerations*.

Graphical User Interface

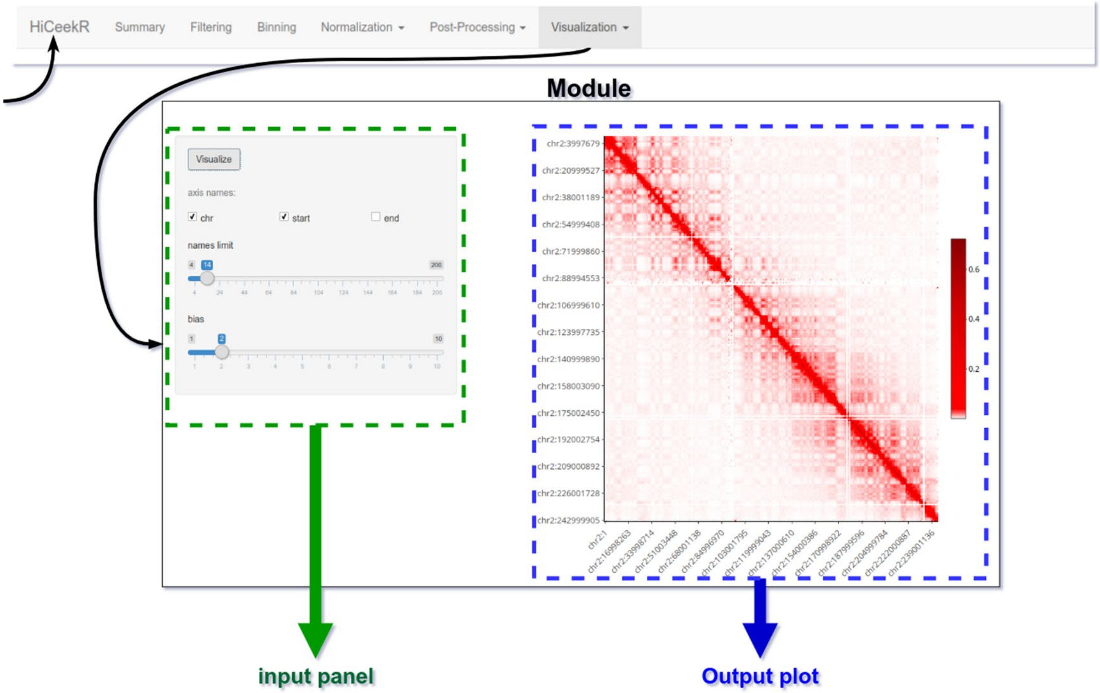
The graphical interface has been designed for guiding the user during the entire analysis process. To this purpose, as shown in **Figure 4**, the upper part of the interface displays the navigation bar illustrating all the main analysis steps in sequential order (i.e. Pre-Processing, Binning, Normalization, Post-Processing, Visualization). Each analysis step panel contains one or more specific functions. By selecting one of them, HiCeekR renders the "Function panel" where input data files, function parameters and/or options (default values are suggested whenever possible) can be set before executing the function (the left side of the interface allows the user to choose all the parameters/options). The results are shown in the "Result panel," that is displayed on the right side of the interface, as plots or tables are automatically saved in a pre-structured way. The graphical representations are interactive and allow exploring the results through point&click and dragging&dropping approach.

Getting Started for the Analysis

At the first HiCeekR execution, the user has to create a configuration file. A dedicated interface will guide him/her by browsing the *working folder*. This step is mandatory for further analyses. Then, each time HiCeekR is executed, the user can either create a new data project or continue/update an already existing project (by selecting the *load* option in the *Welcome* interface). When an experimental dataset is analyzed for the first time, the user will create a new project. HiCeekR will create the data structure, as described in *Data Format and Data Organization* and later results will be stored in a corresponding project name folder. After that, the data analysis can be initiated.

Data Format and Data Organization

HiCeekR allows handling both user experimental data and other information such as the reference genome and annotations. Reference genomes are stored in the *Genomes* folder (in FASTA format), gene annotation in the *Annotation* folder [in Gene Transfer Format (GTF) format]. User experimental data mostly consist in Hi-C sequencing data (i.e. aligned BAM files) obtained from short-read alignment software. However, during the downstream analysis, HiCeekR can use other experimental data such as aligned sequences (i.e. BAM files) obtained from a ChIP-Seq analysis workflow or gene expression values (i.e. TSV file) obtained from RNA-Seq analysis pipeline. We stress that for these additional data the reference genome used during the alignment has to be consistent with the one used for aligning Hi-C data and the gene identifiers have to be consistent with those available in the annotation file. All user experimental data, that refers to the same project, are stored in the *Project data* folder contained in the specific *Project* folder, which has been created by HiCeekR during the pre-processing phase. All user project folders are



axis names:

☑ chr ☑ start □ end

names limit: 14 (4–200)

bias: 2 (1–10)

FIGURE 4 | HiCeekR graphical interface. The upper part of the interface is the navigation bar; on the left side the user can select the parameters of the function, on the right side results will be displayed in form of tables or plots.

saved in the *HiCeekR_projects* main directory. Within each *Project* folder, the results of a specific analysis are organized in the *Analysis* folder, different for each sequence file and resolution (i.e. the width w_b chosen during the binning phase). During each analysis step, HiCeekR stores the results in files in corresponding sub-folders for the specific step. **Figure 5** shows the input/output data organization folder tree.

Shiny, Modules, and Other Technical Considerations

HiCeekR is implemented using R/Shiny library and modular structure. R/Shiny package easily allows developing advanced and practical interfaces in a web-based approach combined with the power of the R statistical instrument. Shiny apps were originally designed for small applications consisting of two main entities: the Shiny User Interface (SUI) that provides all the aesthetic components the user interacts with and the Shiny Server Side (SSS) that performs the required computations. Nevertheless, nowadays it is possible to implement complex applications by combining multiple modules.

A module is conceived as a shiny independent app, with its SSS and SUI. Each HiCeekR interface corresponds to a different module. Overall, the modular structure implemented in HiCeekR allows handling the complexity of the interface and better face the maintainability of the software, not only from a bug-fixing point of view but also when novel functionalities need to be added. Indeed, in this latter case, to add a novel module

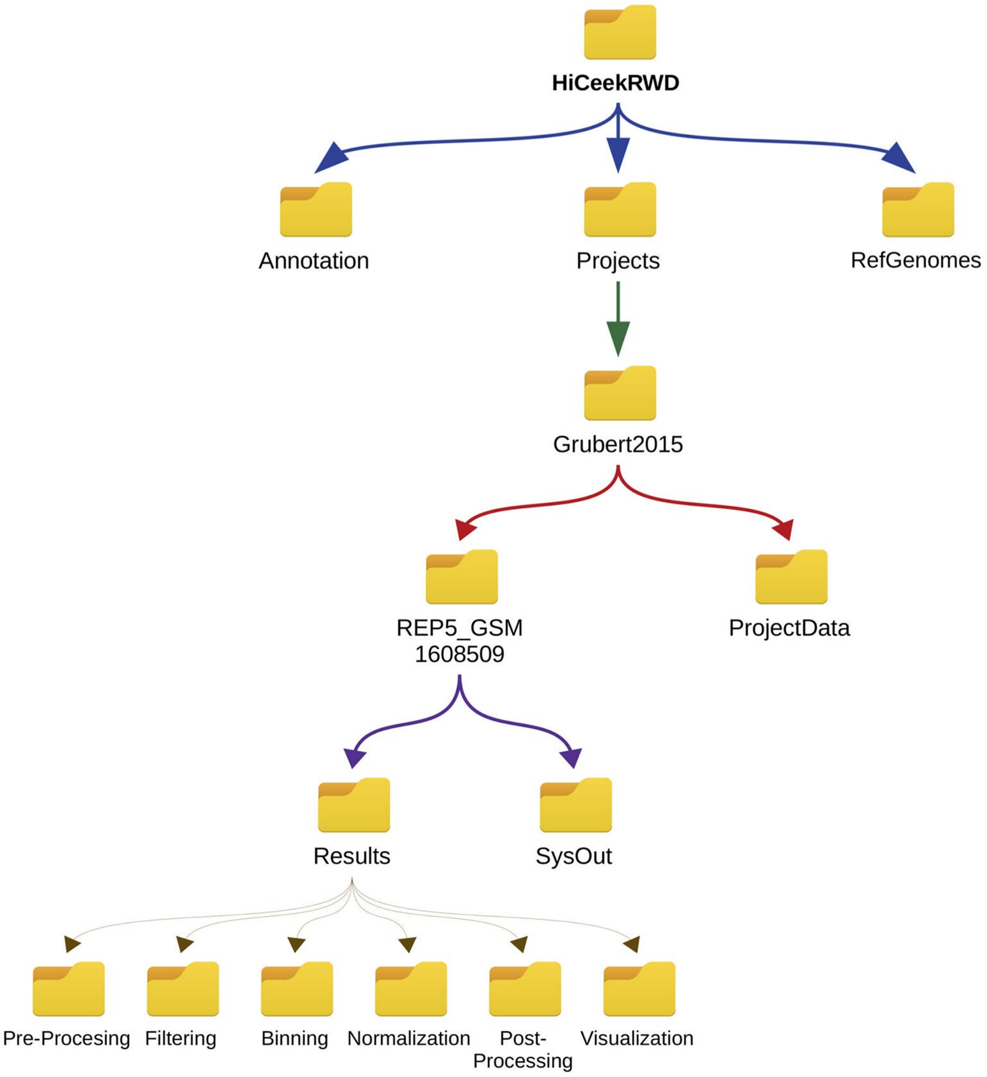


FIGURE 5 | HiCeekR hierarchical data structure. All data are contained in the *HiCeekR* working directory folder and organized in projects. *HiCeekR* working directory folder is created the first time HiCeekR is executed, using the configuration file (see *Getting Started for the Analysis*). Genomes and Annotations can be shared across different projects and are stored in the *Genomes* and *Annotation* folders, respectively. All user projects are saved in the specific *Project_folder* contained in *HiCeekR_projects* main directory. Within the specific *Project_folder* it is possible to create sub-folders related to a specific sample, and/or analysis resolution. Each sub-folder contains *Results* and *SysOut* folders. Folder *Results* contains a sub-folder for each analysis step where intermediate and final results are saved. Folder *SysOut* contains internal logs file and it is not meant for standard users.

it will be necessary only incorporate the novel interface, which implements the required functionalities. Thanks to this choice, HiCeekR results in an easily expansible software.

HiCeekR and Other Available Tools

As mentioned in the *Introduction*, there are relatively few tools that allow performing a comprehensive Hi-C data analysis [see, Calandrelli et al. (2018) and Han and Wei (2017)] for a short list of the most popular tools). Most of them are implemented either in Python, R, Perl, C++, or as a combination of different programming languages. Moreover, they often require several external dependencies to be installed. Out of them, GITAR Calandrelli et al. (2018) and HiCPro Servant et al. (2015) were implemented mostly in Python as command-line. They constitute two useful pipelines designed for expert users (i.e. they allow to perform a specific analysis step or a series of steps). However, they do not have a graphical interface supporting non-expert users. Similarly, HiC-bench Lazaris et al. (2017) provided a well-organized R/Python platform (with a large number of functionalities including those for parameter exploration), but has the same above-mentioned limits for the support of non-expert users. By contrast, HiCdat Schmid et al. (2015) and HiCexplorer Wolff et al. (2018) equipped their tool with a graphical interface. However, the interface of HiCdat is quite naive and limited to the pre-processing step (the higher-order analysis steps have to be performed as command-line). Vice versa, the interface of HiCexplorer is Galaxy based. Hence, it meets the needs of non-expert users as HiCeekR. However, HiCexplorer lacks interactivity in the graphical visualization. Moreover, its local installation is computational demanding. Compared to the above-mentioned alternatives, HiCeekR is completely R based, easy to install and presents a modular graphical interface designed for supporting non-expert users with several functions for interactive visualization of the results.

RESULTS

A Case Study

We illustrate the capability of HiCeekR in analyzing Hi-C data using a dataset from the lymphoblastoid cell line (GM12878) produced from the blood of a female donor, freely available (in FASTQ format) from Gene Expression Omnibus (GEO) (accession number GSE62742). The dataset contains seven biological replicates (including GSM160850 replicates used in the illustrative **Figures 2**, **6**, and **7**), each of them obtained from approximately 25 millions of cells prepared with standard Hi-C library protocol digested with HindIII. The runs were sequenced using Illumina HiSeq 2000 to produce 2×75 paired-end sequences for each library, see Grubert et al. (2015) for details.

Before starting the analysis with HiCeekR, the sequence files were independently aligned to the human reference genome using HiCUP and the hicupmapper script.

In particular, low quality reads (i.e. reads with more than one mismatch in the first 28 bases or the ones with a summed Phred quality score lesser than 70 for all mismatched positions) were removed and only uniquely mapped reads were reported in the BAM files. Duplicated reads were marked using the Picard tools with MarkDuplicates (version 2.18.4). Such BAM files constitute the starting point of the HiCeekR analysis.

We also downloaded a series of ChIP-Seq and RNA-Seq datasets on the same cell line from the ENCODE portal, to illustrate the capability of HiCeekR in integrating other omic data. In particular, we selected already aligned BAM files for the following histone modifications: H3K9Ac, H3K9me3, H4K20me1, H3K27me3, H3K36me3, H3K4me2, H3K4me3, H3K79me2 (ENCSR447YYN series from Bradley Bernstein laboratory at Broad Institute). For simplicity, using the samtools (version 1.9), we merged the three replicates of each modification into a single BAM file, that was sorted and indexed. From RNA-Seq experiment (ENCFF383EXA series from California Institute of Technology or GEO accession number GSE33480) we downloaded the normalized gene expression values and obtained a single two-column tab-delimited file with the gene identifier in the first column and fragments per kilobase of transcript per Million mapped reads (FPKM) in the second one.

All the analyses were performed using as reference genome GRCh37.p13 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.25/) and the gene annotation file obtained from GENCODE gencode.v19.annotation (ENCSR884DHJ).

HiCeekrR Computational Analysis

After creating the new project, we independently analyzed the seven replicates by selecting the corresponding BAM file from the Pre-processing module. For each sample, we selected the reference genome, the cutting enzyme in the *cut site* text-box (*HindIII* site "AAGCTT"), and an *overhang* parameter of 4 bp. Then, we executed the pre-processing and we set 50,000 bp as bin resolution for the rest of the analysis. Therefore, for each BAM file, HiCeekR created a specific folder inside the project folder where the results were saved.

The fragment length and the reads-orientation plots (see **Figure 2**—before filtering) were used to explore the presence of artifacts. We noticed that all the seven replicates show a self-circle spike close to $2^8 = 256$ bp. By using the *Filtering* module, for each BAM file, and setting *min.inward* parameter equal to 1,000 bp, we filtered-out the spike because we are not interested in reads falling in the same restriction fragment. At the same time, since we did not notice dangling-end artifacts, we did not set any *min.outward* threshold to remove it. **Figure 2**—after filtering—illustrates the effect of the applied filtering. Note that, within HiCeekR the figure is interactive, a slide bar allows the user to choose the cut-off directly on the plot.

Afterward, we executed the Binning module using default settings. HiCeekR automatically loaded all required files from the sample under analysis and processed for all the chromosomes. At the end of this step, the detected interactions are shown in the results panel (and saved in the corresponding folder), as bin-to-bin interaction tables.

For this illustrative example, we decided to investigate only chromosomes: 1, 2, 3, 13, 14, 16, since they were previously studied in Martin et al. (2015). Therefore, we selected the corresponding target chromosomes inside "*chromosome of interest*" *selection box* and ticked the *selective bin table* check-box inside the *Export* panel to continue the analysis.

From the Normalization module, we selected the ICE normalization method and set the *Window.high* parameter equal to 0.02 in combination with the *ignore.low = 0* parameter to ignore the low abundance bins. Moreover, to avoid the *NA* values produced by the DiffHiC implementation, we also selected the "Set NA to min" check-box. In such a way HiCeekR sets all the *NAs* to the *min* of the matrix. Afterward, we exported the normalized contact matrix for the chromosomes of our interest. Note that these normalized matrices constitute the starting point of the post-processing analysis.

For brevity, here we illustrate only two cases of usage for the Post-processing: *i)* We first identified compartments and TADs, then we integrated them with ChIP-Seq data, and visualized a region of interest (as in **Figure 6**), *ii)* We converted the normalized contact matrix in a network of interactions for some regions of interests (as in **Figure 7**), then we identified the genes located in each interacting bin and performed gene functional analysis. In this latter case, we also added the gene expression values from RNA-Seq data.

For the first case, we used the PCA module on the normalized contact matrix. Afterward, we used the directionality index module to determine the directional index *di* and TopDomTADs with *Window Size = 20* that provides us a list of TADs boundaries in BED format. Then, we used the EpigeneticFeature module to process ChIP-Seq dataset and compute the normalized coverage at the same genomic resolution of the HiC-Seq analysis (i.e. over bins). Using the *select bin Table file* selector, we chose the BED file corresponding to the chosen bin resolution and the chromosome of interest (here we chose chromosome 2). Then, in the first sub-panel, we selected the first BAM file for the ChIP-Seq data, e.g. the *H3K9Ac* BAM file, through the *BAM file path* selector, and we associated "H3K9Ac" as track label. By checking the *add* checkbox, we added a second track without replacing the previous one. We repeated this operation for H3K9me3, H4K20me1, H3K27me3, H3K36me3, H3K4me2, H3K4me3, H3K79me2. At the end of the process for each sample, HiCeekR generated a vector containing the raw coverage (number of mapped reads) in the bins. Using the second sub-panel, we exported the coverage for all samples as a combined table. To do this, we chose the file name through the *file name* text input and the normalization strategy to use (in the *normalization* checkbox). For this case study, we performed the *RPM* normalization and saved the results using the *export table* button.

Using heatmap module (*layout*), we selected the normalized contact matrix by the *contact matrix* input file widget and we focused the attention on the region 51902204–71950291 of chromosome 2, as illustrative example. From the same panel, we added four additional tracks. In particular, we selected in the first slot the PCA file obtained from the pca module. Since this file contains multiple columns (corresponding to the eigenvectors of the principal components), we selected the eigenvector corresponding to the second principal component (PC2). Note, PC1 or PC2 are usually used to describe compartments, the specific choice depending on the size of the region of interested and the resolution of the analysis. In the second slot, we loaded the directionality index d_i file. After that, we added the bed track of the TADs boundaries as produced by the TopDomTADs. Then, we added the two epigenetic tracks (produced in EpigeneticFeatures module) selecting "H3K9Ac" and "H3K27me3" features columns as an illustrative example. At the end of these uploads, we are able to visualize all the tracks by flagging the *active* checkbox in each slot panel (see **Figure 6**).

In the second case, we used the network module in the Visualization panel and focused the attention on the regions investigated in Martin et al. (2015), listed in **Table 1**. Note that since the regions in **Table 1** are often larger than the bin size chosen for this analysis, each region can correspond to a few bins.

To this purpose, we first selected the normalized contact table (using the *contact table* input file widget), then the gene annotation file (using *Annotation* file input), finally we added the RNA-Seq gene expression data, by selecting the specific file in the *Expression data* file input. By pressing the *set input* button HiCeekR loaded the data and moved into the second tab panel (*show*). Inside this tab panel, we selected the chromosomal coordinates given in **Table 1** (analyzing them individually). For all the interested regions, we set the *normValue* to 0.01 and checked the *global* checkbox (in the left panel). Since the focus of the study was to enlighten long-range interactions, we excluded from the visualization all those regions with a bin distance lower than eight bins, by checking the *intra Chr* checkbox and setting the *min bin distance* text box to 8. Then, HiCeekR visualized the network (see **Figure 7**) and produced three interactive panel-tables (i.e. *Interactions*, *Genes*, and *Functional*), as mentioned in *Pre-Processing*. Within panel-tables *Interactions*, we ranked all the interactions by the interaction strength from the strongest (higher contact matrix value) to the weakest (lower contact matrix value). Therefore, we identified the strongest bin to bin interactions together with the genes therein contained. For the functional analysis, we selected the *hsapiens* database in the *organism* select box.

Analysis Results

Results of the first analysis are summarized in **Figure 6**, where the short p-arm of chromosome 2 (chr2:51,000,000–71,000,000) is displayed in a multi-layer view. The figure includes the normalized contact matrix (on the top) and, in order, the PC2 eigenvector (as a green track), the *di* indices (as a red track), the TADs boundaries as detected by TopDom (as a purple track), and the RPM normalized tracks of the histone marks H3K9Ac, H3K27me3 (as brown and pink tracks), which are associated to transcribed an repressed chromatin, respectively. We highlighted a correlation between the typical rectangular block-shapes in the heatmap and the PC2 loadings allowing detecting the A/B compartments (territories) and categorizing also the TADs thanks to the directionality indexes *di*. Additionally, the histone mark tracks allow us to better characterize the chromatin structure within each pattern. A clear correlation between distinct A/B compartments and the H3K9Ac and H3K27me3 enriched regions is shown at the selected chromosomal region (**Figure 6**).

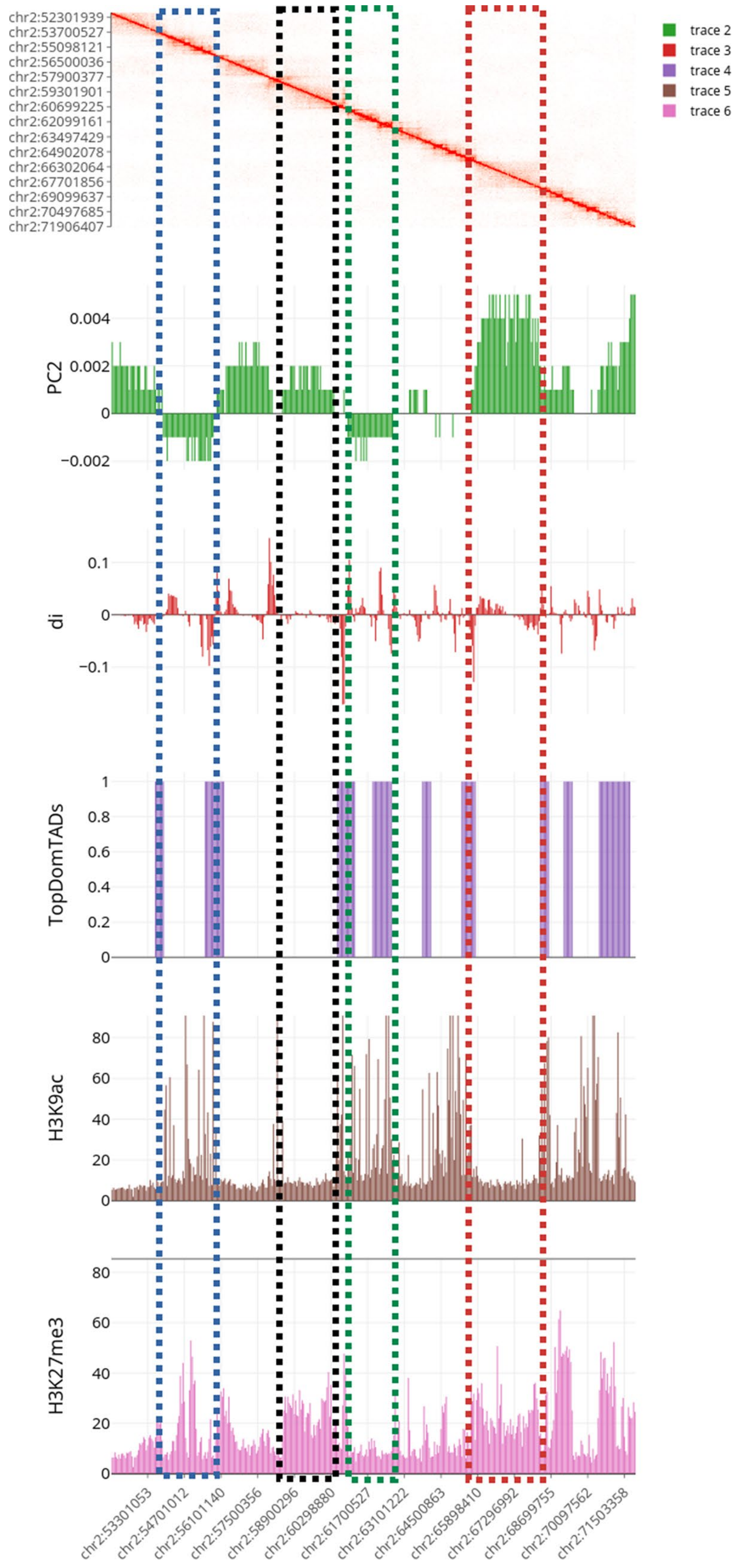


FIGURE 6 | Continued

FIGURE 6 | Multilayer visualization of the region 51902204–71950291 of chromosome 2 (replicate GSM1608509). From the top, the first track shows the normalized contact matrix as a heatmap, where the color intensity is proportional to the strength or the interaction. The green track shows the eigenvector of the second principal components (PC2) that define the putative A/B compartments. The red track displays the directional indexes *di* (that helps in defining TADs). The purple track shows the TADs boundaries as detected by TopDomTADs. The two remaining tracks show the RPM normalized coverage for H3K9Ac (in brown) and H3K4me2 (in pink) histone marks. The H3K9Ac and H3K4me2 enriched regions exhibit a profile similar to PC2 track, indicating that it overall correlates with the 3D organization of the chromosomes in these regions.

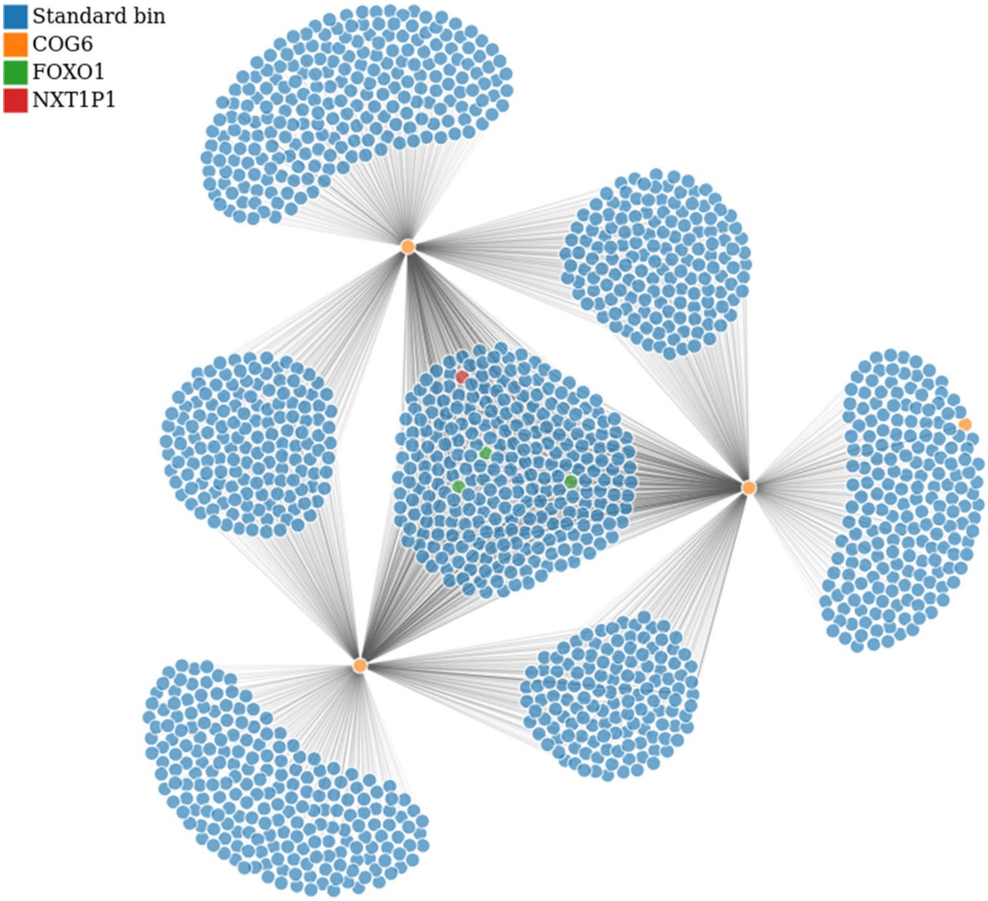


FIGURE 7 | Bin to bin interaction network (evaluated with low stringency). The interaction network was built starting from the region in Tab. 1 containing COG6 gene (bins in orange) and retrieving all interactions within chromosome 13. Additionally, we highlighted the bins containing FOXO1 and NXT1P1, in green and red respectively. This analysis has been performed on GSM160850 replicate.

TABLE 1 | The list of regions identified in Martin et al. (2015) (as chromosome, start, end of the region, and the most relevant genes therein located).

Chr	Start	End	Genes
Chr1	197,473,879	197,744,623	*DENND1B*
Chr3	27,757,440	27,764,206	*EOMES*
Chr13	40,229,764	40,326,765	*COG6*
Chr14	69,262,513	69,454,180	*ZFP36L1*, *ACTN1*
Chr16	11,022,748	11,036,257	*DEXI*

For the second analysis, we report the independent analysis of the regions in **Table 1**. First of all, we noticed that the regions identified in Martin et al. (2015) are often among the strongest interactions (top positions after ranking by strength) identified in our analysis.

In particular, from the panel-table *Interactions*, we easily identified the following gene-bins interactions, where gene-bins means the bins overlapping or containing a given gene. Recall that, based on the chosen resolution and the length of the gene body, each bin might contain few genes, or a given gene might be associated with few bins. We identified that the *EOMES-bins* has multiple strong interactions within chromosome 3, as previously reported Martin et al. (2015). Out of them, the *EOMES-bins* was found to interact with the *AZI2-bins* (such interaction was confirmed for all replicates with strength spanning from 0.020 to 0.025 in the normalized matrices). Additionally, we confirmed the interaction between the *COG6-bins* and the *FOXO1-bins* within chromosome 13, although it is weak (about 0.01 in the normalized matrices). By contrast, we found that the *COG6-bins* presents a strong interaction with the *NXT1P1-bins* (chr13:39697243-39750825) (about 0.017 in the normalized matrices). Such case is illustrated in **Figure 7**, where the *COG6-bins* are depicted in yellow, and the *NXT1P1-bins* and *FOXO1-bins* are depicted in red and green, respectively. Moreover, the *DEXI-bins* on chromosome 16 shows a strong interaction with the *RMI2-bins*, as reported in Martin et al. (2015). Indeed, this interaction was found with strength from 0.0344 to 0.033 in the normalized matrices, being among the strongest interactions that this region shows with distant regions. This region seems also to interact with the *ZC3H7A-bins*, although this interaction is weaker (value close to 0.01) than others. On the other hand, when moving to chromosome 1, the *DENND1B-bins* shows a strong interaction with the *LHX9-bins* (with normalization matrix values spanning from 0.020 to 0.030). Finally, on chromosome 14, we partially confirmed the interaction between the bins containing the *ZFP36L1* and *ACTN1* genes and the *ZFYE26-bins*. This interaction was observed only in a subset of replicates, and, when detected, it shows low strength (normalized value of about 0.01).

From the panel-table *Genes*, we found that, according to the RNA-Seq data, all above mentioned interacting genes are expressed except *LHX9* and show variable expression levels in RNA-Seq: *ZC3H74* gene has the highest RPKM value (186.55), *ZFP36L1*, *ZFYVE26*, and *AZI2* genes show high expression (59.03, 48.67, 37.93 respectively), while *DENDD1B*, *EOMES*, *FOXO1*, *ACTN*, *DEXI*, and *RMI2* genes show a lower level of expression (ranging from 4.21 to 6.87).

Finally, the most interesting results of the functional enrichment analysis performed on the genes interacting with regions in **Table 1** are given in **Table 2**. We can see that *DENND1B* gene, which codifies for a guanine nucleotide exchange factor (GEF) acting as a regulator of T-cell receptor (TCR) internalization in T-cells interacts with LHX9, ATP6V1G3, C1ORF53 genes. They show significant enrichment of binding sites for the transcription factor T-bet, that is a master regulator of the T-helper 1 (Th1) cell development (Kallies and Good-Jacobson, 2017). The zinc-finger *ZFP36L1* gene interacts with *RAD51B* and *ACTN1* genes, which codify for proteins involved in homologous recombination and cell migration, respectively (Lio et al., 2003; Yamaji et al., 2004). Remarkably, the *AZI2* gene, which interacts with the *EOMES* gene, is an important activator of *NF-kB* signaling as also reported in Martin et al. (2015). It shows binding sites for the *FOXJ2* transcription factor, which strictly correlated with *NF-kB* signaling (Lin et al., 2004).

Computational Costs

The analysis of this case-study was executed on an Intel i7-7700HQ processor, with 32Gb RAM system (64bit architecture) on Ubuntu 18.04 LTS, with R version 3.6.1 and Shiny 1.3.2. Other relevant packages are listed in the github page.

The most computationally expensive step is the pre-processing of Hi-C data which requires approximately 20 to 25 min for processing a single BAM file of approximately 150 million of reads. For the binning step, performed on large chromosomes such as human chromosome 1 or 2, with bin size 50,000 bp, the elapsing time is about 3 to 5 min including the output file storage. While for the normalization step the required time is about 30 s. The identification of TADs requires 2 to 5 min per chromosome, depending on the methods and the size of the chosen chromosome. Another time demanding step is the import of indexed ChIP-Seq BAM files that can even take a couple of hours for samples with very high depth such as those obtained after merging different replicates. The computational time is clearly reduced when working with a specific chromosome or at lower bin resolutions or with organisms with smaller genomes.

Software Availability and System Requirements

HiCeekR is freely available as source code package on GitHub (https://github.com/lucidif/HiCeekR), where future releases will be also posted. Moreover, issues and problems can be submitted to the HiCeekR developers through the github issues page to contribute to the development of future releases. The github page also includes a detailed user manual where all HiCeekR modules are described and the data used in the current study that can be used as training example. The current version of HiCeekR was developed and tested on Ubuntu 16/18 and macOS 10.13, using R environment version 3.6.1, and the latest releases of R packages is reported on the github page as Session Info. System requirements strongly depend on the size of the reference genome, sequencing depth and, in particular, on the bin resolution. However, minimal system requirements are Intel i5 4th generation processor and 16Gb RAM.

CONCLUSIONS

Despite the relevance of Hi-C data and the availability of several packages for performing specific steps in their analysis, only a few

TABLE 2 | Results obtained from the functional analysis; the table contains significant terms identified starting from the list of genes contained in the bins strongly interacting with the regions examined by network construction.

Region	term.id	dm	term.name	intersection	*p*-value
DENND1B	TF:M08355	tf	Factor: HOXB2:T-bet	LHX9, ATP6V1G3, C1ORF53	0.0195
EOMES	TF:M08290_1	tf	Factor: FOXJ2:Elf-1	AZI2, ZCWPW2	0.0053
EOMES	TF:M03979_1	tf	Factor: ETV1	AZI2, ZCWPW2	0.0306
EOMES	TF:M07287_1	tf	Factor: FOXO3A	AZI2, ZCWPW2	0.0362
ZFP36L1,ACTN1	CORUM:260	cor	RAD51B-RAD51C complex	RAD51B	0.0497
ZFP36L1,ACTN1	CORUM:4025	cor	Affixin-actinin (alpha) complex	ACTN1	0.0497

comprehensive and user-friendly tools have been developed during the last years (Schmid et al., 2015; Caudai et al., 2018; Wolff et al., 2018). Thanks to its GUI, HiCeekR provides an easy-to-use way to analyze this data type, specifically designed to guide researchers lacking specific training in scientific programming through the different computational steps. Moreover, it also provides multiple approaches for integrating Hi-C data with other omic datasets and a wide series of interactive graphical outputs that can significantly support researches in the interpretation of the huge amount of data produced during Hi-C experiments. The major capabilities of HiCeekR are illustrated by analyzing a publicly available dataset, and integrating ChIP-Seq and RNA-Seq dataset.

Moreover, HiCeekR is implemented in a modular structure. Therefore, other approaches available in literature could be easily encapsulated in further releases. In this regard, an interesting extension is the one proposed by Merelli et al. (2015). In this latter case, by using NuChart tool they build multiple gene-centric graphs starting from Hi-C and transcription data, allowing additional statistical investigations, thanks to the graph-based approach. Such an approach can complement HiCeekR network approach to provide a wider range of methods. It is also clear that post-processing analysis constitutes one of the aspects where artificial intelligence approaches can still greatly contribute to the elucidation of chromatin structure and gene regulation interplay, therefore several other algorithms are expected to be available soon. Hence, we expect that HiCeekR will growth by expanding the number of methods available.

On the other hand, although HiCeekR already implements several methods to facilitate Hi-C data analysis, much work still needs to be done to speed-up the time-demanding computations required for carrying out some specific steps, such as the pre-processing and binning. A possible improvement is the implementation of a parallel version of the algorithms used in HiCeekR or the split-up of the computations on multiple cores/CPUs. In this regards, a good example is given by the NuChart-II R packages, where particular attention is reserved for the implementation of parallel routines for Hi-C data analysis (Merelli et al., 2013; Tordini et al., 2017).

Last but not least, HiCeekR can be improved to better supporting computational reproducible research. Indeed, thanks to its GUI approach, HiCeekR guides the user to perform a complete analysis of Hi-C data, automatically storing input/output data. Despite this is very helpful from the user point of view, it does not provide reproducible research functionalities yet. As mentioned in (Russo et al., 2016b), it is known that the problem of computational reproducibility is very challenging for tools based on GUI, since it becomes hard to precisely trace all the steps/parameters of the analysis workflow when the user can apply a point-and-click approach. However, in the same spirit such that (Russo et al., 2016a) was extending RNASeqGUI (Russo and Angelini, 2014) in the direction of reproducible research, we plan to implement multiple functionalities to automatically produce a comprehensive analysis report incorporating all the executed code and the results (as tables and figures).

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62742. Accession number: GSE62742.

AUTHOR CONTRIBUTIONS

LF designed and implemented HiCeekR, performed analysis of the real cases, and drafted the manuscript. DR contributed to the design and implementation of HiCeekR and wrote the manuscript. MG and MM contributed to the discussion of the real data analysis. CA contributed to the design of HiCeekR. MM and CA guided and supervised all phases of HiCeekR development and wrote the manuscript. All authors read and approved the manuscript.

FUNDING

This work was partially supported by the Italian Flagship project (Epigen) and the Regione Campania Project ADViSE.

ACKNOWLEDGMENTS

This work was partially supported by the Italian Flagship project (Epigen) and the Regione Campania Project ADViSE.

REFERENCES

Ay, F., and Noble, W. S. (2015). Analysis methods for studying the 3D architecture of the genome. *Genome Biol.* 16, 183. doi: 10.1186/s13059-015-0745-7

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58, 268–276. doi: 10.1016/j.ymeth.2012.05.001

Calandrelli, R., Wu, Q., Guan, J., and Zhong, S. (2018). GITAR: an open source tool for analysis and visualization of Hi-C data. *Genomics Proteomics Bioinf.* 16 (5), 365–372. doi: 10.1016/j.gpb.2018.06.006.

Caudai, C., Salerno, E., Zoppe, M., Merelli, I., and Tonazzini, A. (2018). ChromStruct 4: a python code to estimate the chromatin structure from Hi-C Data. *IEEE/ACM Trans. Comput. Biol. Bioinf.* doi: 10.1109/TCBB.2018.2838669

Dekker, J. (2002). Capturing chromosome conformation. *Science* 295, 1306–1311. doi: 10.1126/science.1067799

Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* 14, 390–403. doi: 10.1038/nrg3454

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* 485 (7398), 376–380. doi: 10.1038/nature11082

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., et al. (2006). Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res*. 16 (10), 1299–1309. doi: 10.1101/gr.5571506

Forcato, M., Nicoletti, C., Pal, K., Livi, C. M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. *Nat. Methods* 14, 679–685. doi: 10.1038/nmeth.4325

Grubert, F., Zaugg, J. B., Kasowski, M., Ursu, O., Spacek, D. V., Martin, A. R., et al. (2015). Genetic control of chromatin states in humans involves local and distal chromosomal interactions. *Cell* 162 (5), 1051–1065. doi: 10.1016/j.cell.2015.07.048

Han, Z., and Wei, G. (2017). Computational tools for Hi-C data analysis. *Quant. Biol.* 5 (3), 215–225. doi: 10.1007/s40484-017-0113-6

Hu, M., Deng, K., Selvaraj, S., Qin, Z., Ren, B., and Liu, J. S. (2012). HiCNorm: Removing biases in Hi-C data *via* Poisson regression. *Bioinformatics* 28, 3131–3133. doi: 10.1093/bioinformatics/bts570

Imakaev, M., Fudenberg, G., Mccord, R. P., Naumova, N., Goloborodko, A., Lajoie, B. R., et al. (2012). Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods* 9, 999–1003. doi: 10.1038/nmeth.2148.Iterative

Kallies, A., and Good-Jacobson, K. L. (2017). Transcription factor T-bet orchestrates lineage development and function in the immune system. *Trends Immunol.* 38 (4), 287–297. doi: 10.1016/j.it.2017.02.003

Knight, P. A., and Ruiz, D. (2013). A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* 33 (3), 1029–1047. doi: 10.1093/imanum/drs019

Lajoie, B. R., Dekker, J., and Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* 72, 65–75. doi: 10.1016/j.ymeth.2014.10.031

Lazaris, C., Kelly, S., Ntziachristos, P., Aifantis, I., and Tsirigos, A. (2017). HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking. *BMC Genomics* 18, 22. doi: 10.1186/s12864-016-3387-6

Lévy-Leduc, C., Delattre, M., Mary-Huard, T., and Robin, S. (2014). Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* 30 (17), i386–i392. doi: 10.1093/bioinformatics/btu443

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326 (5950), 289–293. doi: 10.1126/science.1181369

Lin, L., Spoor, M. S., Gerth, A. J., Brody, S. L., and Peng, S. L. (2004). Modulation of Th1 activation and inflammation by the NF-κB repressor Foxj1. *Science* 303 (5660), 1017–1020. doi: 10.1126/science.1093889

Lio, Y. C., Mazin, A. V., Kowalczykowski, S. C., and Chen, D. J. (2003). Complex formation by the human Rad51B and Rad51C DNA repair proteins and their activities *in vitro*. *J. Biol. Chem.* 278 (4), 2469–2478. doi: 10.1074/jbc.M211038200

Lun, A. T., and Smyth, G. K. (2015). diffHic: a bioconductor package to detect differential genomic interactions in Hi-C data. *BMC Bioinf.* 16, 258. doi: 10.1186/s12859-015-0683-0

Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S., et al. (2015). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nat. Commun.* 30 (6), 10069. doi: 10.1038/ncomms10069

Merelli, I., Liò, P., and Milanesi, L. (2013). NuChart: an R package to study gene spatial neighbourhoods with multi-omics annotations. *PloS One* 8 (9), e75146. doi: 10.1371/journal.pone.0075146

Merelli, I., Tordini, F., Drocco, M., Aldinucci, M., Liò, P., and Milanesi, L. (2015). Integrating multi-omic features exploiting chromosome conformation capture data. *Front. Genet.* 6, 1–11. doi: 10.3389/fgene.2015.00040

Nicoletti, C., Forcato, M., and Bicciato, S. (2018). Computational methods for analyzing genome-wide chromosome conformation capture data. *Curr. Opin. Biotechnol.* 54, 98–105. doi: 10.1016/j.copbio.2018.01.023

Pal, K., Forcato, M., and Ferrari, F. (2019). Hi-C analysis: from data generation to integration. *Biophys. Rev.* 11 (1), 67–78. doi: 10.1007/s12551-018-0489-1

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et al. (2019). g: profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* 47 (W1), W191–W198. doi: 10.1093/nar/gkz369

Russo, F., and Angelini, C. (2014). RNASeqGUI: a GUI for analysing RNA-Seq data. *Bioinformatics* 30 (17), 2514–2516. doi: 10.1093/bioinformatics/btu308

Russo, F., Righelli, D., and Angelini, C. (2016a). Advancements in RNASeqGUI towards a reproducible analysis of RNA-Seq experiments. *BioMed. Res. Int.* 2016, 79723510. doi: 10.1155/2016/7972351

Russo, F., Righelli, D., and Angelini, C., (2016b). "Advantages and limits in the adoption of reproducible research and R-tools for the analysis of omic data," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Cham. doi: 10.1007/978-3-319-44332-4_19 .

Schmid, M. W., Grob, S., and Grossniklaus, U. (2015). HiCdat: a fast and easy-to-use Hi-C data analysis tool. *BMC Bioinf.* 16, 277. doi: 10.1186/s12859-015-0678-x

Servant, N., Lajoie, B. R., Nora, E. P., Giorgetti, L., Chen, C. J., Heard, E., et al. (2012). HiTC: exploration of high-throughput 'C' experiments. *Bioinformatics* 28, 2843–2844. doi: 10.1093/bioinformatics/bts521

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., et al (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* 16, 259. doi: 10.1186/s13059-015-0831-x

Shavit, Y., and Lio', P. (2014). Combining a wavelet change point and the Bayes factor for analysing chromosomal interaction data. *Mol. Biosyst.* 10, 1576–1585. doi: 10.1039/C4MB00142G C4MB00142G

Shin, H., Shi, Y., Dai, C., Tjong, H., Gong, K., Alber, F., et al. (2015). TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.* 44 (7), e70. doi: 10.1093/nar/gkv1505

Solovei, I., Cavallo, A., Schermelleh, L., Jaunin, F., Scasselati, C., Cmarko, D., et al. (2002). Spatial preservation of nuclear chromatin architecture during three-dimensional fluorescence *in situ* hybridization (3D-FISH). *Exp. Cell Res.* 276 (1), 10–23. doi: 10.1006/excr.2002.5513

Tordini, F., Drocco, M., Misale, C., Milanesi, L., Liò, P., Merelli, I., et al. (2017). NuChart-II: the road to a fast and scalable tool for Hi-C data analysis. *Int. J. High Perform. Comput. Appl.* 31, 196–211. doi: 10.1177/1094342016668567

Wingett, S. W., Ewels, P., Furlan-Magaril, M., Nagano, T., Schoenfelder, S., Fraser, P., et al. (2015). HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* 4, 1310. doi: 10.12688/f1000research.7334.1

Wolff, J., Bhardwaj, V., Nothjunge, S., Richard, G., Renschler, G., Gilsbach, R., et al. (2018). Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* 46 (W1), W11–W16. doi: 10.1093/nar/gky504

Yaffe, E., and Tanay, A. (2011). Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.* 43, 1059–1065. doi: 10.1038/ng.947

Yamaji, S., Suzuki, A., Kanamori, H., Mishima, W., Yoshimi, R., Takasaki, H., et al. (2004). Affixin interacts with α -actinin and mediates integrin signaling for reorganization of F-actin induced by initial cell-substrate interaction. *J. Cell Biol.* 165 (4), 539–551. doi: 10.1083/jcb.200308141

Yardimci, G. G., and Noble, W. S. (2017). Software tools for visualizing Hi-C data. *Genome Biol.* 18 (1), 26. doi: 10.1186/s13059-017-1161-y

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.* 38 (11), 1341–1347. doi: 10.1038/ng1891

Zufferey, M., Tavernari, D., Oricchio, E., and Ciriello, G. (2018). Comparison of computational methods for the identification of topologically associating domains. *Genome Biol.* 19, 217. doi: 10.1186/s13059-018-1596-9

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 reviewer PL declared a past co-authorship with one of the authors CA to the handling editor.

frontiers
in Genetics

REVIEW
published: 19 November 2019
doi: 10.3389/fgene.2019.01141

Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors

Anja Mösch[1,2], *Silke Raffegerst*[2], *Manon Weis*[2], *Dolores J. Schendel*[2] *and Dmitrij Frishman*[1]*

[1] Department of Bioinformatics, Wissenschaftszentrum Weihenstephan, Technische Universität München, Freising, Germany, [2] Medigene Immunotherapies GmbH, a subsidiary of Medigene AG, Planegg, Germany

In the last years, immunotherapies have shown tremendous success as treatments for multiple types of cancer. However, there are still many obstacles to overcome in order to increase response rates and identify effective therapies for every individual patient. Since there are many possibilities to boost a patient's immune response against a tumor and not all can be covered, this review is focused on T cell receptor-mediated therapies. CD8$^+$ T cells can detect and destroy malignant cells by binding to peptides presented on cell surfaces by MHC (major histocompatibility complex) class I molecules. CD4$^+$ T cells can also mediate powerful immune responses but their peptide recognition by MHC class II molecules is more complex, which is why the attention has been focused on CD8$^+$ T cells. Therapies based on the power of T cells can, on the one hand, enhance T cell recognition by introducing TCRs that preferentially direct T cells to tumor sites (so called TCR-T therapy) or through vaccination to induce T cells *in vivo*. On the other hand, T cell activity can be improved by immune checkpoint inhibition or other means that help create a microenvironment favorable for cytotoxic T cell activity. The manifold ways in which the immune system and cancer interact with each other require not only the use of large omics datasets from gene, to transcript, to protein, and to peptide but also make the application of machine learning methods inevitable. Currently, discovering and selecting suitable TCRs is a very costly and work intensive *in vitro* process. To facilitate this process and to additionally allow for highly personalized therapies that can simultaneously target multiple patient-specific antigens, especially neoepitopes, breakthrough computational methods for predicting antigen presentation and TCR binding are urgently required. Particularly, potential cross-reactivity is a major consideration since off-target toxicity can pose a major threat to patient safety. The current speed at which not only datasets grow and are made available to the public, but also at which new machine learning methods evolve, is assuring that computational approaches will be able to help to solve problems that immunotherapies are still facing.

Keywords: cancer immunotherapy, T cell receptor, neoepitope, neoantigen, cross-reactivity, MHC binding affinity prediction

OPEN ACCESS

Edited by:
Davide Chicco,
Peter Munk Cardiac Centre,
Canada

Reviewed by:
Dhruv Sethi,
Obsidian Therapeutics,
United States
Gustavo Fioravanti Vieira,
Universidade La Salle Canoas,
Brazil

****Correspondence:***
Dmitrij Frishman
d.frishman@wzw.tum.de

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

Received: *26 July 2019*
Accepted: *21 October 2019*
Published: *19 November 2019*

Citation:
Mösch A, Raffegerst S, Weis M, Schendel DJ and Frishman D (2019) Machine Learning for Cancer Immunotherapies Based on Epitope Recognition by T Cell Receptors. Front. Genet. 10:1141. doi: 10.3389/fgene.2019.01141

INTRODUCTION

Immunotherapies have gained more and more importance over the last decades. Checkpoint inhibitors mainly targeting PD1/PDL1 and CTLA4 and personalized cancer vaccines (Gubin et al., 2014; Ott et al., 2017; Sahin et al., 2017) have been and still are heavily investigated in clinical trials. Both depend on patient individual tumor-specific mutations enabling the boost of a cancer-specific T cell-mediated immune response (Snyder et al., 2014; Rizvi et al., 2015; Łuksza et al., 2017). A more direct approach utilizes the adoptive transfer of a patient's autologous T cells, either genetically modified with a transgenic chimeric antigen receptor (CAR) or T cell receptor (TCR). For CAR-T cell as well as TCR-T cell therapy a defined target, the epitope, needs to be identified. CARs, carrying the functional antigen-binding domain of an antibody, recognize three-dimensional peptide structures on the surface of a cell (Sadelain et al., 2013). By contrast, TCRs recognize predominantly linear peptides presented by the major histocompatibility complex (MHC) called human leucocyte antigen (HLA) in humans. For MHC class I presentation and thus CD8$^+$ T cell detection, these peptides come from proteins that are intracellularly processed by either the constitutive proteasome or the IFNγ induced immunoproteasome (Griffin et al., 1998; Neefjes et al., 2011). After cleavage, the peptides are transported to the endoplasmic reticulum (ER) by the transporter associated with antigen processing (TAP) complex, where they are loaded onto MHC class I molecules. The peptide-MHCs (pMHCs) are shuttled to the cell surface where they can potentially be recognized by CD8$^+$ cytotoxic T cells, either naturally carrying or engineered to bear a pMHC-specific TCR (see **Figure 1**). However, there are more than 16,000 different alleles for *HLA-A*, *-B*, and *-C* genes, which bind and present different epitopes (Robinson et al., 2015). Besides MHC class I mediated CD8$^+$ cytotoxic T cell responses, MHC class II bound peptides can induce CD4$^+$ T cell responses that are also reported to play an important role in tumor detection and elimination (Nielsen et al., 2010; Linnemann et al., 2014; Kreiter et al., 2015; Andreatta et al., 2017; Veatch et al., 2018).

A wide spectrum of bioinformatics tools exists for modeling all steps of the MHC class I antigen presentation pathway, including proteasomal cleavage, translocation of the peptides to the ER by TAP, peptide binding to the MHC molecules, and TCR recognition. The overarching goal of these efforts is to enhance our understanding of how T cell epitopes are selected from a virtually unlimited number of short peptides that can be proteolytically generated from the human proteome. The origin of these T cell epitopes can be naturally occurring proteins or peptides derived from somatic mutations. For personalized cancer immunotherapy, these patient- and tumor-specific mutations are usually separately assessed for each patient by exome sequencing, mutation detection and peptide binding prediction (Robbins et al., 2013; Blankenstein et al., 2015; Schumacher and Schreiber, 2015). Predicting these so called neoepitopes or neoantigens is a prevailing challenge for computational methods for immunotherapy and essential for a high-throughput approach to narrow down mutations to be included in vaccines or to be evaluated *in vitro* for T cell recognition, since only very few mutations are truly immunogenic (Yadav et al., 2014; Strønen et al., 2016; Bjerregaard et al., 2017a).

It is also of utmost importance to evaluate potential cross-reactivity of target-candidate epitopes based on various omics data such as proteomics and peptidomics (Haase et al., 2015; Jaravine et al., 2017a; 2017b). However, all existing approaches based on epitope presentation are only a surrogate for T cell recognition, for which no universal and computationally viable approach exists so far, although the first promising results have been published (Jurtz et al., 2018; Ogishi and Yotsuyanagi, 2019). By now, datasets have been generated that allow sequence-based prediction approaches using deep learning (Shugay et al., 2018; Vita et al., 2018).

In this review, we summarize the current state at the development of prediction algorithms and methods for all steps of antigen presentation, evaluate neoepitope prediction approaches, and discuss progress toward sequence-based TCR binding prediction.

PREDICTION OF T CELL EPITOPES

Proteasomal Cleavage Prediction

In order to develop an accurate prediction algorithm for proteosomal cleavages, a thorough mechanistic understanding of

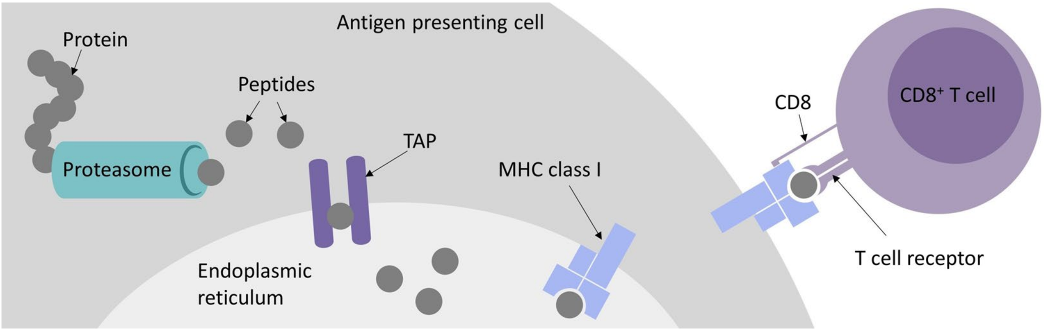


FIGURE 1 | Major histocompatibility complex (MHC) class I antigen presentation pathway for peptides recognized by CD8$^+$ cytotoxic T cells.

the cutting process is required. The PAProC algorithm by Kuttler et al. (Kuttler et al., 2000) relies on a biologically motivated model, which postulates that proteolytic sites are mostly determined by the local sequence context, generally not further away in the sequence than six amino acid residues. The two residues immediately adjacent to the cut make the greatest contribution to the affinity to the active subunits of the proteasome, while the influence of the other surrounding residues is lower. The recognition model is additive in that the total affinity, which ultimately determines the probability of the cut, is considered to be the sum of all individual contributions. Bioinformatics analyses revealed that the amino acids in the six positions preceding the cut and four positions downstream contain sufficient information to reproduce a training dataset of experimentally determined cleavage motifs of 20S proteasomes by a network-based technique. Keşmir et al. (Keşmir et al., 2002) demonstrated that good results in detecting proteasomal cleavage motifs can be achieved by combining experimental data on degradation by the constitutive proteasome with the sequences of peptides bound by the MHC class I molecules, which may be generated either by the constitutive or by the immunoproteasomes. A neural network trained on such a composite dataset, called NetChop, and an updated version NetChop 3.0 (Nielsen et al., 2005), achieved a reasonable accuracy and also yielded useful insights into cleavage-promoting and inhibiting residues as well as into N-terminal extension of peptides after proteasomal cleavage. A recurrent difficulty in predicting proteasomal cleavage is the lack of experimentally verified noncleavage sites. However, such negative data can be artificially generated by considering internal positions of confirmed MHC ligands or randomly generated sites.

TAP Binding Prediction

An early study of Daniel et al. (1998), in which the TAP binding affinity for a large number of peptides of length nine was measured by a peptide binding assay, revealed that positions one to three and nine of the 9-mers make the largest contribution to the selectivity of TAP to peptides. An artificial neural network trained on these data was able to predict the IC_{50} values with high accuracy. The study also found that HLA class I molecules differed significantly with respect to TAP affinities of their ligands. The predictive scope was later extended to peptides of arbitrary length using a stabilized matrix approach and a scoring scheme that only considers the first three N-terminal residues and the last C-terminal residue (Peters et al., 2003). Since it has been established that the selectivity of peptide transport by TAP is entirely determined by the peptide-binding step (Gubler et al., 1998), affinity predictions can be equated with translocation likelihood predictions. A number of further machine learning methods for predicting peptide binding to TAP were trained on 9-mer data, which is the typical length of the peptides that will subsequently bind to the MHC complex (Bhasin, 2004; Zhang et al., 2006; Diez-Rivero et al., 2010; Lam et al., 2010).

Peptide-MHC Binding Prediction

Sequencing of peptides eluted from MHC class I molecules (Falk et al., 1991) as well as mass-spectrometric (MS) (Hunt et al., 1992) and crystallographic (Madden, 1995) evidence revealed common properties of the epitopes, in particular the typical length range of 8–12 residues. Additionally, it showed the existence of MHC allele-specific anchor residues, usually in positions two and nine of the core nonameric segments, as well as auxiliary anchors, where amino acid preferences are less strict (Rammensee et al., 1993).

Starting from the early nineties, efforts were made to collect available information on MHC class I ligands (Brusic et al., 1994; Rammensee et al., 1995,Rammensee et al.,1999) and to predict them using simple motif- and profile-based techniques (Rothbard and Taylor, 1988; Parker et al., 1994; Reche et al., 2002), based on the notion that peptides highly similar in sequence to experimentally characterized ligands will have a higher binding potential than more distantly related peptides and that individual amino acid side chains make independent contributions to the overall binding energy. Machine learning techniques, such as neural networks and hidden Markov models (Bisset and Fierz, 1993; Mamitsuka, 1998; Nielsen et al., 2003) outperform matrix-based methods in predicting peptide binding affinity (Peters et al., 2006; Lin et al., 2008). They are able to deal with peptides of variable length (Lundegaard et al., 2008) and to take into account nonadditive effects, which may arise, e.g., when two amino acids compete for the same site in the peptide-binding groove of the MHC heterodimer. The latest version of the widely used NetMHC algorithm 4.0 (Andreatta and Nielsen, 2016) was trained on many thousands of quantitative affinity measurements for peptides of length 8–11 and the total of 118 MHC class I alleles from human, other primates, and mouse. Neural networks trained on all peptides (allmer networks) significantly outperformed the networks trained on peptides of each individual length separately. The study also suggested specific binding modes for 10- and 11-mers, which are predicted to bulge out of the MHC grove in contrast to 8- and 9-mers, which are strictly linear epitopes. MHCflurry, which relies on affinity measurement and peptide elution MS data, also uses neural networks trained individually for each HLA allele (O'Donnell et al., 2018b). Additionally, it allows users to train networks locally on data of their choice. This can be important especially for cancer immunotherapy applications, since peptide-binding affinity predictions are traditionally focused on viral epitopes.

There is also a growing group of pan-specific methods, including PickPocket (Zhang et al., 2009), NetMHCpan 4.0 (Jurtz et al., 2017), PSSMHCpan (Liu et al., 2017), and ACME (Hu et al., 2019), which take as input both the peptide and the HLA sequence and are able to predict the binding of any peptide to any allele. Most predictions are focused on MHC class I, but there are also methods available for MHC class II, such as NetMHCII 2.3 and NetMHCIIpan 3.2 (Jensen et al., 2018), ProPred (Singh and Raghava, 2001), SMM-align (Nielsen et al., 2007), and NNAlign (Nielsen and Andreatta, 2017), of which the latter also allows to train and use own models, as Garde et al. did for MHC class II prediction using both affinity measurement and MS data (Garde et al., 2019). Many of the aforementioned prediction methods for both MHC class I and II and consensus methods, such as NetMHCcons (Karosiene et al., 2012) and the consensus method by Moutaftsi et al. (Moutaftsi et al., 2006), are integrated into the IEDB epitope analysis resource and can be accessed online (Wang et al., 2010; Fleri et al., 2017; Vita et al., 2018; Dhanda et al., 2019). In addition, combinatory pipelines and frameworks have been published, namely, EpiJen (Doytchinova et al., 2006), NetCTL (Larsen et al., 2007), NetCTLpan (Stranzl et al., 2010), and FRED2 (Schubert et al., 2016), modeling the complete antigen presentation pathway by including proteasomal cleavage and TAP transport predictions.

Epitope presentation, however, is only one step toward T cell recognition. NetMHCstab (Jørgensen et al., 2014) and NetMHCstabpan (Rasmussen et al., 2016) are methods to predict the stability of pMHC complexes, presuming that epitope presentation lasting longer increases the likelihood of T cell recognition and thus immunogenicity. Calis et al. proposed a scoring model to predict true immunogenicity of T cell epitopes (Calis et al., 2013). Despite these efforts, however, true immunogenicity remains far more difficult to predict than mere MHC-binding affinity.

Beyond sequence-based approaches, significant methodological progress has been made in modeling peptide binding to MHC class I molecules on structure level. The diversity of the cognate peptide repertoire and the experimental binding profiles for a particular MHC protein can be accurately captured using both general purpose modeling packages, such as Rosetta (Yanover and Bradley, 2011), and faster specialized methods, such as GradDock (Kyeong et al., 2018), DockTope (Menegatti Rigo et al., 2015), and LYRA (Klausen et al., 2015), of which the latter two are also integrated in the IEDB. Docking experiments are becoming increasingly successful in reproducing crystallographically known peptide-MHC binding geometry (Bordner and Abagyan, 2006; Antunes et al., 2018).

Immunopeptidomics Data

The recent availability of large-scale immunopeptidomics data allowed to explicitly model peptide length distributions and the interdependence between individual sequence positions, leading to more accurate predictions of naturally presented MHC class I ligands (Gfeller et al., 2018). MS profiling provides novel insights into the antigen processing rules, including the discovery of binding motifs, improved description of proteasomal cleavage signatures, cellular localization and sequence features of peptide source proteins, and better understanding of the role of gene expression, protein abundance and degradation (Bassani-Sternberg et al., 2015; Bassani-Sternberg et al., 2017; Abelin et al., 2017). In particular, Abelin et al. (2017) reported that neural networks trained on MS-derived peptides bound to 16 different HLA alleles outperformed affinity-trained predictors.

For immunogenicity, T cell epitope verification by TCRs or TCR-like antibodies would constitute an ideal dataset to train prediction algorithms (Dolan, 2019), but both approaches are highly dependent on specificity and affinity of TCRs and antibodies used and do not reach the high-throughput efficiency of immunopeptidomics. HLA-peptidomics, which is the MS analysis of MHC-eluted peptides, is the most sophisticated method for high-throughput qualitative and quantitative detection of MHC ligands and thereby of potential T cell epitopes (Hunt et al., 1992; Caron et al., 2011; ; Hassan et al., 2014; Álvaro-Benito et al.,2018; Freudenmann et al., 2018).

The isolation of pMHC complexes from cell surfaces (Sugawara et al., 1987;Storkus et al., 1993; Bassani-Sternberg et al., 2015; Marino et al., 2019) or out of serum (Ritz et al., 2016, 2017) is the first critical step for a high-quality MS HLA-peptidome analysis. After elution from pMHC complexes, peptides are purified, separated by high pressure liquid chromatography (HPLC), and directly injected and analyzed in a mass spectrometer followed by computational processing of MS spectra data (see **Figure 2**). Successful peptide detection is determined by various factors, such as HLA enrichment, which is dependent on HLA-antibody quality, efficient elution, and physicochemical characteristics of a peptide defined by its amino acid composition. Relevant peptide properties can be mass, hydrophilicity, and hydrophobicity, its ability to be ionized, as well as cysteine content (Gfeller and Bassani-Sternberg, 2018). Therefore, not all peptides are equally likely to be detected by MS but it is difficult to assess how many peptides are missed. Peptide sequences are often determined by tandem MS: a precursor mass spectrum called MS1 spectrum of the eluted peptides is generated and only peptides with high intensities are isolated for fragmentation and analyzed, resulting in a MS2 or MS/MS spectrum. Observed mass spectra are then compared with theoretical mass spectra in general reference databases. Proteogenomic computational pipelines using customized reference datasets also allow the identification of peptides originating from noncanonical and allegedly noncoding reading frames (Laumont and Perreault, 2017; Laumont et al., 2018), unconventional, genomic coding-sequences (Erhard et al.,

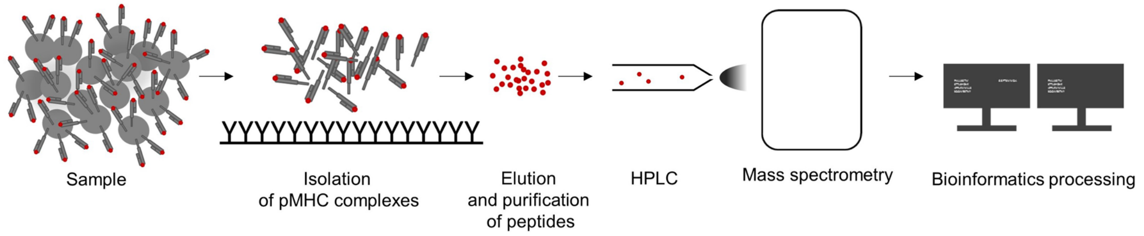

Sample
Isolation of pMHC complexes
Elution and purification of peptides
HPLC
Mass spectrometry
Bioinformatics processing

FIGURE 2 | Workflow to analyze of major histocompatibility complex (MHC)-eluted peptides by mass-spectrometric (MS). A sample is lysed, pMHC complexes are captured and peptides are purified by immunoaffinity purification using MHC-specific immobilized antibodies. Eluted peptides are separated by high pressure liquid chromatography (HPLC), analyzed by MS, and the resulting data are computationally processed.

2018) as well as neoepitopes from somatic alterations (Yadav et al., 2014; Carreno et al., 2015) or intron retentions (Smart et al., 2018). In addition, the generation of customized spectral library databases of high confidence peptides can be used for data-independent acquisition approaches (Ritz et al., 2017), resulting in increased reproducibility and sensitivity.

Peptides are often assigned to the HLA molecule from which they were originally eluted by predicting the binding affinity (Freudenmann et al., 2018; Bilich et al., 2019). For common HLA alleles, usually a sufficient number of peptides are identified as binders, resulting in datasets large enough to train prediction algorithms. However, for less frequent HLA alleles, the pool of identified and correctly assigned peptides is more limited, which leads to variability in performance of prediction techniques depending on the rarity of each HLA allele (O'Donnell et al., 2018b). If MS datasets annotated by binding affinity predictions are used to train machine learning algorithms, a self-amplifying bias is introduced. MS profiling of mono-allelic cells (Giam et al., 2015; Abelin et al., 2017) as well as deconvolution approaches (Bassani-Sternberg and Gfeller, 2016) can circumvent this problem and improve the quality of available training data and prediction performance.

IMMUNOTHERAPY-SPECIFIC APPLICATIONS OF EPITOPE PREDICTION

Neoepitope Identification

Cancer-specific mutations have been demonstrated to be viable targets for tumor-infiltrating lymphocytes (TILs) enabled by checkpoint inhibitors that block CTLA4 or PD1/PDL1 or by vaccine-induced immune responses (van Rooij et al., 2013; Carreno et al., 2015; Cohen et al., 2015; Gros et al., 2016; McGranahan et al., 2016; Ott et al., 2017; Zacharakis et al., 2018; Hilf et al., 2019). These mutations alter amino acid sequences of proteins and are recognized as so called neoepitopes or neoantigens, with both terms used ambiguously and oftentimes synonymously in the literature. Here, we use the term neoepitopes for epitopes predicted to be presented by a certain MHC and the term neoantigens for confirmed immunogenic mutations. By definition, neoantigens are tumor-specific, which makes them ideal immunotherapy targets, but they are also to a large degree patient-specific. Despite many efforts, only very few shared neoantigens such as KRAS$^{G12D/V}$ or BRAF^{V600E}, could be identified, making an off-the-shelf therapy approach hardly feasible (Warren and Holt, 2010; Angelova et al., 2015; Tran et al., 2015; Thorsson et al., 2018). Furthermore, a high individual tumor mutation burden and the ambition to provide personalized medicine for more patients do not allow for testing the immunogenicity of every mutation *in vitro*. Therefore, the current standard procedure for individual patients relies on exome sequencing followed by mutation calling and machine learning-based neoepitope prediction, which represents the main application of pMHC-binding prediction algorithms in the field of cancer immunotherapy. Here, we reviewed more than 70 publications using binding prediction algorithms to identify neoepitopes of which 49, that provided quantifiable data, are shown in **Table 1**. Not all studies stated all steps of their neoepitope selection process, including which algorithm parameters were used, how many neoepitopes were found when applying a threshold or how many and what types of mutation were used for predicting neoepitopes, which makes quantitative evaluation and reproducibility difficult. This is aggravated by the large variance in ratio of predicted neoepitopes per mutation, which is caused by thresholds of varying strictness, the number of features used for filtering, and the approach to counting neoepitopes or neoantigens, i.e., if a mutation was counted only once even if presented by more than one HLA allele or contained in multiple epitopes predicted to be immunogenic. Furthermore, some studies could only experimentally validate a subset of predicted neoepitopes and experimental validation was determined by biological assays of varying sensitivity from MHC-ligand confirmation to ELISPOT assays using patient-specific TILs.

Not surprisingly, most publications investigated cancer types known for high mutation loads, such as non-small cell lung carcinoma and melanoma, but glioblastoma and chronic lymphocytic leukemia were also shown to harbor neoantigens identified by neoepitope prediction (Rajasagi et al., 2014; Hilf et al., 2019; Keskin et al., 2019). Regarding mutation types, the focus clearly lies on single nucleotide variants (SNVs) considering their abundance in tumors above all other types of mutation, their comparatively easy detection by mutation calling software and easier computational generation of mutated and wild-type peptide sequences (Bailey et al., 2018; Ellrott et al., 2018). However, larger indels, frameshifts, and other more complex mutation types can be the source of more neoepitopes that are also less similar to self and thus highly interesting immunotherapeutic targets. More recent studies from Kahles et al., Koster et al., and Schischlik et al. investigated these types of mutation, benefitting from improvements on sequencing and mutation calling techniques (Kahles et al., 2018; Koster and Plasterk, 2019; Schischlik et al., 2019). Nevertheless, identification of cancer-specific mutation remains a critical step in every neoepitope identification pipeline and the number of mutations obtained varies greatly dependent on the software and thresholds employed (Tran et al., 2015; Karasaki et al., 2017).

The focus of most publications lies on MHC class I presented neoepitopes that can be detected by CD8$^+$ T cells. MHC class I prediction algorithms are more commonly used but there is clear evidence that MHC class II mediated CD4$^+$ T cell responses play a major role in neoantigen immune responses and thus should also be considered for neoepitope detection. (Linnemann et al., 2014; Kreiter et al., 2015; Tran et al., 2015; Hugo et al., 2016; Ott et al., 2017; Reuben et al., 2017; Sahin et al., 2017; Sonntag et al., 2018; Vrecko et al., 2018).

All studies, except Koster et al., who investigated 10-mers only, looked at peptides with a length of 8–10 or 8–11 amino acids or just at 9-mers alone, which are the majority of peptides presented by MHC class I (Trolle et al., 2016). Most studies also relied on matching HLA types for the samples used, often determined by one of the following HLA typing algorithms: ATHLATES, HLAminer, OptiType, PHLAT, POLYSOLVER, and seq2HLA (Boegel et al., 2012; Warren et al., 2012; Liu et al., 2013; Szolek et al., 2014; Shukla et al., 2015; Bai et al., 2018). In contrast,

TABLE 1 | Publications describing the application of machine learning approaches to neoepitope prediction.

Publication	Indication	Sample type and number	number of HLAs used	Estimated ratio of predicted neoepitopes from mutations	Estimated ratio of experimentally confirmed neoantigens	Number of features	Algorithms
(Segal et al., 2008)	BRCA/CRC	11 patients	1	0.17	N/A	1	NetMHC, SYFPEITHI, BIMAS, RANKPEP
(Castle et al., 2012)	MEL	1 murine cell line	N/S	0.05	0.32^{T}	2	NetMHC
(Khalili et al., 2012)	various	312 genes (COSMIC)	57	1.40	N/A	2	NetMHC 3.2
(Robbins et al., 2013)	MEL	3 patients	2	0.18	0.03 ^{T}	3	NetMHCpan 2.4
(van Rooij et al., 2013)	MEL	1 patient	4	0.42	<0.01 ^{T}	3	NetChop, NetMHC 3.2
(Boegel et al., 2014)	various	167 cancer cell lines	6	0.44	N/A	1	IEDB 2.9
(Duan et al., 2014)	SARC	2 murine tumors	3	0.75	0.56 ^{T}	2	NetMHC 3.0
(Snyder et al., 2014)	MEL	64 patients	6	0.42	<0.01 ^{T}	3	NetMHC 3.4, RANKPEP, IEDB immunogenicity, CTLPred
(Yadav et al., 2014)	CRC/PRAD	2 murine cell lines	2	0.03	0.02 ^{T}	3	NetMHC 3.4
(Angelova et al., 2015)	CRC	552 TCGA patients	6	0.41	N/A	2	NetMHCpan
(Carreno et al., 2015)	MEL	7 samples/3 patients	1	0.04	0.43 ^{B}	3	NetMHC 3.4
(Cohen et al., 2015)	MEL	8 patients	2	0.02	0.02 ^{T}	2	IEDB
(Rizvi et al., 2015)	NSCLC	34 patients	6	0.62	<0.01 ^{T}	2	NetMHC 3.4
(Rooney et al., 2015)	various	4250 TCGA patients	6	0.14	N/A	2	NetMHCpan 2.4
(Tran et al., 2015)	GIC	10 patients	12	0.03	0.21 ^{T}	2	NetMHCpan 2.8, NetMHCIIpan 3.0
(Van Allen et al., 2015)	MEL	110 patients	6	1.56	N/A	2	NetMHCpan 2.4
(van Gool et al., 2015)	UCEC	245 TCGA patients	1	0.06	N/A	3	NetMHCpan 2.8
(Bassani-Sternberg and Gfeller, 2016)	MEL	1 patient	6	1.43	<0.01 ^{B}	1	NetMHCpan 2.8
(Goh et al., 2016)	MCC	49 patients	4	0.09	N/A	1	NetMHC 3.4
(Gros et al., 2016)	MEL	3 patients	6	0.03	0.55 ^{T}	2	IEDB
(Hugo et al., 2016)	MEL	38 patients	12	0.06	N/A	3	NetMHCpan 2.8, NetMHCIIpan 3.0
(Kalaora et al., 2016)	MEL	1 patient	6	5.30	<0.01 ^{B}	1	NetMHCpan 2.8
(Karasaki et al., 2016)	NSCLC	15 patients	6	0.62	N/A	1	NetMHCpan 2.8
(Löffler et al., 2016)	CHOL	1 patient	6	3.68	0 ^{B}	2	NetMHC 3.4, NetMHCpan 2.8, SYFPEITHI
(Strønen et al., 2016)	MEL	3 patients	1	0.05	0.19 ^{T}	4	NetChop, NetMHC 3.2, NetMHCpan 2.0
(Anagnostou et al., 2017)	NSCLC	10 patients	6	0.76	<0.01 ^{T}	4	SYFPEITHI, NetMHCpan, NetCTLpan
(Chang et al., 2017)	PED	540 patients	6	0.42	N/A	2	NetMHCcons 1.1
(Karasaki et al., 2017)	NSCLC	4 patients	6	0.20	N/A	2	NetMHCpan 2.8
(Kato et al., 2017)	BRCA	5 patients	6	0.47	N/A	2	NetMHC 3.4, NetMHCpan 2.8
(Miller et al., 2017)	MM	664 patients	6	0.16	N/A	3	NetMHC 4.0
(Ott et al., 2017)	MEL	6 patients	6	0.01	0.60 ^{T}	3	NetMHCpan 2.4
(Sahin et al., 2017)	MEL	13 patients	10	0.02	0.60 ^{T}	2	IEDB 2.5 (MHC class I & II)
(Zhang et al., 2017)	BRCA	3 patients	6	0.01	0.16 ^{T}	3	NetMHC 3.2
(Kalaora et al., 2018)	MEL	15 patients/cell lines	6	9.57	0.15 ^{T}	2	NetMHCpan 3.0
(Kinkead et al., 2018)	PAAD	1 murine cell line	2	0.27	0.16 ^{T}	2	NetMHC 3.2/3.4, NetMHCpan 2.8
(Martin et al., 2018)	OV	1 patient	6	1.57	0,09 ^{T}	2	NetMHCpan 2.4
(O'Donnell et al., 2018a)	OV	92 patients	6	0.02	N/A	2	NetMHCpan 2.8
(Sonntag et al., 2018)	PDAC	1 patient	10	2.00	0.75 ^{T}	3	NetMHC, NetMHCIIpan 3.1, SYFPEITHI

(*Continued*)

TABLE 1 | Continued

Publication	Indication	Sample type and number	number of HLAs used	Estimated ratio of predicted neoepitopes from mutations	Estimated ratio of experimentally confirmed neoantigens	Number of features	Algorithms
(Thorsson et al., 2018)	various	8546 TCGA patients	6	0.74	N/A	2	NetMHCpan 3.0, pVAC-Seq 4.0.8
(Vrecko et al., 2018)	HCC	1 patient	3	0.05	0.15 ^{T}	2	SYFPEITHI, IEDB (MHC class II)
(Wu et al., 2018)	various	7748 TCGA samples	100	1.18	N/A	1	NetMHCpan 4.0
(Bulik-Sullivan et al., 2019)	NSCLC	7 patients	6	0.10	0.08 ^{T}	>4	EDGE
(Hilf et al., 2019)	GBM	10 patients	1	0.03	0.85 ^{T}	3	IEDB 2.5
(Keskin et al., 2019)	GBM	8 patients	6	0.20	0.07 ^{T}	3	NetMHCpan 2.4
(Koster and Plasterk, 2019)	various	10186 TCGA patients	1	0.02	N/A	2	NetMHC 4.0
(Liu et al., 2019)	OV	20 patients	12	0.15	0.24 ^{T}	3	NetMHCpan 3.0, NetMHCIIpan 3.1
(Löffler et al., 2019)	HCC	16 patients	6	1.79	0 ^{B}	2	NetMHC 4.0, NetMHCpan 3.0, SYFPEITHI
(Rosenthal et al., 2019)	NSCLC	164 samples/64 patients	6	0.86	N/A	2	NetMHC 4.0, NetMHCpan 2.8
(Schischlik et al., 2019)	PNMN	113 patients	6	2.53	0.66 ^{B}	2	NetMHCpan

N/S means not specified. Cancer type abbreviations: adenocarcinoma (AC), breast cancer (BRCA), cholangiocarcinoma (CHOL), colorectal cancer (CRC), glioblastoma (GBM), gastrointestinal cancer (GIC), hepatocellular carcinoma (HCC), merkel cell carcinoma (MCC), melanoma (MEL), multiple myeloma (MM), non-small cell lung cancer (NSCLC), ovarian cancer (OV), pancreatic ductal adenocarcinoma (PDAC), pediatric cancers (PED), Ph-negative myeloproliferative neoplasms (PNMN), prostate adenocarcinoma (PRAD), sarcoma (SARC) and uterine corpus endometrial cancer (UCEC). ^{T} indicates experimentally confirmed T cell responses (e.g., IFNγ ELISPOT), ^{B} indicates experimentally confirmed major histocompatibility complex (MHC) binding (e.g., mass spectrometric [MS] of eluted peptides), and N/A indicates that no experimental validation was done. Features are mutated peptide binding prediction, wild-type peptide binding prediction, gene expression, sequence-based features like sequence similarity scores, and immunogenicity predictions. If available, version information of algorithms is included.

Wu et al. made predictions based on the 100 most frequent HLA alleles in their dataset and Wood et al. based on the general 145 most frequent alleles (Wood et al., 2018; Wu et al., 2018). Whether or not such approaches yield substantial information gain is a debatable issue since most immunogenic mutations are highly individual and restricted by a patient's individual HLA type (Marty et al., 2017; McGranahan et al., 2017; Rosenthal et al., 2019). HLA-A*02:01 has been extensively studied since it is the most common allele in Caucasian populations and therefore was exclusively used by Segal et al. for their analysis (Segal et al., 2008). Since predictions for A*02:01 still belong to the best performing group and can be more easily validated compared to other alleles due to established *in vitro* protocols and reagents, Carreno et al., Spranger et al., Strønen et al., van Gool et al., and Hilf et al. also only used A*02:01 for their predictions and the studies that carried out experimental validation accomplished high confirmation of predicted neoepitopes (Carreno et al., 2015; van Gool et al., 2015; Spranger et al., 2016; Strønen et al., 2016; Hilf et al., 2019). Similarly, Koster et al. only used A*02:01 for an unfiltered TCGA dataset although they did not perform experimental validation. Similar to Wood et al., they did not use HLA typing information for TCGA samples, which has been generated but can only be obtained by applying for access to restricted data (Shukla et al., 2015; Charoentong et al., 2017; Marty et al., 2017).

For most studies, algorithms from the NetMHC family were chosen as they are widely known and represent the state-of-the-art prediction methods for binding of a peptide to a given MHC molecule. Van Allen et al. showed that out of 17 validated neoantigens, 14 passed the 500 nM standard threshold, indicating high sensitivity (van Buuren et al., 2014). However, only a handful of the predicted binders will also be recognized by T cells, which requires additional filtering or prediction improvement (Anonymous, 2017). Indeed, using more filtering criteria leads to fewer predicted neoepitopes per mutation, as seen in **Figure 3A**, although the false negative rate remains unknown. Only a few publications rely on predicting the binding affinity of mutated peptides alone and most use at least one additional threshold criterion, of which gene expression as a premise for antigen recognition is the most common. As RNA-Seq data was not available for Anagnostou et al., Le et al. and Reuben et al., they used TCGA expression data as a proxy to further filter the mutations to test for immunogenicity. Binding of the wild-type peptide was also considered by some studies, but not always used for filtering. Duan et al. proposed a "differential agretopicity index" (DAI), which is the difference between the predicted mutated and wild-type binding affinity, to use as a filtering criterion for neoepitope prediction. Although it yielded promising results based on their mouse data, it seemed less reliable in further investigations by Bjerregaard et al. and Koşaloğlu-Yalçın et al. using human data (Duan et al., 2014; Bjerregaard et al., 2017b; Koşaloğlu-Yalçın et al., 2018). In another study by Ghorani et al., DAI was more predictive for

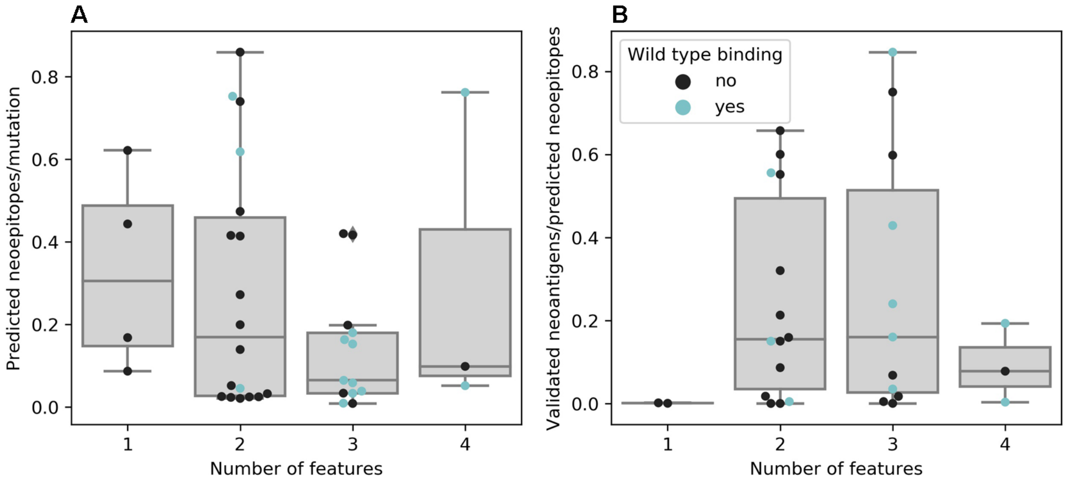


FIGURE 3 | (A) Neoepitopes per mutation grouped by the number of features used for neoepitope selection. Data based on publications that offered comparable data, e.g., not obviously counting a neoepitope predicted to be presented by multiple major histocompatibility complexes (MHCs) multiple times (n = 38). **(B)** Ratio of confirmed to predicted neoepitopes grouped by the number of features used for neoepitope selection. Data based on publications that experimentally validated all predicted neoepitopes (n = 30)

immune infiltration in melanoma and lung cancer compared to neoantigen or mutation load, suggesting that while some neoepitope responses might be enhanced by a reduced cross-reactivity potential, there are also many validated neoantigens whose wild-type counterparts are predicted to bind comparably strong (Ghorani et al., 2018; Koşaloğlu-Yalçın et al., 2018).

There is evidence that taking more than one feature into account promises greater success for experimentally validating predicted neoepitopes (see **Figure 3B**). However, the results of experimental validation are dependent on the sensitivity of the technique used and the reactivity of neoantigen-specific TILs can additionally be hampered by other factors, such as tumor immune suppression or T cell exhaustion (Anonymous, 2017; Bulik-Sullivan et al., 2019).

Some studies chose a quantitative approach, mostly linking neoepitope load and survival (Brown et al., 2014; Rizvi et al., 2015; Miller et al., 2017; Ghorani et al., 2018). It has to be mentioned that neoepitope load and mutational burden are usually highly correlated (Pearson r = 0.89 based on 38 publications with less than 1 neoepitope per mutation from **Table 1**) and although it can be assumed that an increased survival is linked to the immunogenicity of mutations, quantifying predicted neoepitopes does not necessarily transport more information than mutation burden alone (Nathanson et al., 2017). There are, however, also studies that correlated survival with neoepitopes but not mutational burden or found contradictory results depending on patient cohorts (Snyder et al., 2014; Ghorani et al., 2018).

Among well-described approaches for neoepitope identification based on affinity binding prediction algorithms, there are also pipelines available that automate all analytic steps and rank potential neoepitopes based on peptide affinity prediction and other features (see **Table 2**). They differ greatly as to their properties and outputs, thus offering choices depending on research questions and dataset sizes. Their availability demonstrates how important neoepitope prediction has become as an application for binding affinity prediction algorithms.

TABLE 2 | Neoepitope prediction pipelines based on mutation data input. Additional features are cancer driver status of the mutated gene used by MuPeXI; differential agretopicity index (DAI), sequence-based immunogenicity score, and more used by Neopepsee; DAI, cleavage, and stability prediction used by pVACtools.

	MuPeXI	**CloudNeo**	**Neopepsee**	**pVACTools**
Algorithms	NetMHCpan	NetMHCpan	NetCTLpan, IEDB Bayes classifier	8 MHC class I predictors 4 MHC class II predictors
Input	VCF gene expression TSV	VCF BAM	VCF RNA-Seq FASTQ	VCF BAM (RNA and DNA)
HLA typing	user input	integrated	user input or integrated	user input or integrated
Mutation types	SNVs indels frameshifts	SNVs	SNVs	SNVs indels fusions (additional input)
Wild type peptide	yes	yes	yes	yes
Gene expression	yes (optional)	no	yes	yes
Additional features	yes	no	yes	yes
Availability	local, webserver	cloud	local	local
Reference	(Bjerregaard et al., 2017a)	(Bais et al., 2017)	(Kim et al., 2018)	(Hundal et al., 2019)

Since a variety of different neoepitope identification approaches exist and it is not clear which features are predictive for immunogenicity, Koşaloğlu-Yalçın et al. and Kim et al. integrated and compared features additional to the standard MHC binding affinity by either comparing areas under the curve of receiver operating characteristics or evaluating feature importance derived from trained classifiers (Kim et al., 2018; Koşaloğlu-Yalçın et al., 2018). Both studies found that binding affinity prediction performs best or is the most informative feature. This is not surprising for viral epitopes constituting a major part of data on which most prediction algorithms are trained nor for neoantigens from literature mainly selected by predicted binding affinity, which introduces a bias toward this feature. It still remains unclear how many potential neoantigens are not detected because their binding affinity is predicted to lie beyond thresholds. An approach avoiding this bias has been proposed by Bulik-Sullivan et al. (Bulik-Sullivan et al., 2019). Like the most recent generation of neural network binding prediction algorithms, they developed a deep learning neural network trained on MS data, but apart from improved peptide sequence modeling, they also included features unrelated to the pMHC interaction, namely, quantified gene expression, flanking sequence, and protein family. Although their model is currently limited to HLA alleles of the training data, the approach demonstrated an increased performance of neoepitope discovery over peptide binding prediction and can also be expanded to MHC class II presented antigens.

Cross-Reactivity Assessment

A major challenge for immunotherapies introducing TCRs into patient recipient T cells is the choice of safe target antigens. If an engineered TCR-T cell cross-reacts with self-antigens in healthy tissue, the side-effects can be devastating. Possible TCR toxicity scenarios can be generally divided into on-target and off-target toxicities. On-target toxicities include all aspects of a specific target antigen or epitope expression that lead to an unintentional TCR-mediated destruction of healthy tissues. An example of on-target toxicity is melanocyte destruction, hearing loss, and retina infiltration mediated by MART1-targeting TCR-T cells relating to the same epitope in all cases (Johnson et al., 2009).

Off-target toxicities, in contrast, can appear by unexpected recognition of alternative epitopes that contain amino acid exchanges (mismatches) compared to the known epitope sequence. In rare cases, these mismatched peptides are presented identically on corresponding MHC molecules and are recognized equally well by deployed TCRs.

Targeting epitope sequences of proteins originating from highly homologous family members can cause unforeseen tissue damage as exemplified by the study performed by Morgan et al. (Morgan et al., 2013). Using autologous anti-MAGEA3 TCR-T cells, adoptive transfer led to severe neurotoxicity in several patients. The MAGEA3-specific TCR used in this clinical trial also recognized a MAGEA12, which was retrospectively found to be expressed in the brain. In the Linette et al. study, clinicians adoptively transferred MAGEA3-TCR-modified lymphocytes that also recognized an alternative epitope derived from the protein titin, causing fatal heart failure in two patients (Linette et al., 2013). Each of these examples underline the importance and need of comprehensive preclinical target and TCR analysis to prevent potential adverse events at later stages of clinical development.

With Expitope, we presented the first web server for assessing epitope sharing when evaluating new potential target candidates (Haase et al., 2015). Based on predictions for proteasomal cleavage, TAP transport, and MHC class I binding affinity, Expitope lists peptides with a given number of mismatches including the original target peptide. For these peptides, which are linked to genes by transcripts, the expression values in various healthy tissues, representing all vital human organs, are extracted from RNA-Seq data. However, transcript abundance only indirectly indicates protein expression. Meanwhile, proteome-wide human protein abundance data has become available and now facilitates a more direct approach for the prediction of potential cross-reactivity. The development of a new version 2.0 of Expitope, which computes all possible, naturally occurring epitopes of a peptide sequence and the corresponding cross-reactivity indices using both protein and transcript abundance levels weighted by a proposed hierarchy of importance of various human tissues, should help addressing this issue (Jaravine et al., 2017a). Cross-reactivity potential can also be assessed by calculating structural similarities between pMHC complexes obtained by molecular docking (Antunes et al., 2010) and by clustering pMHC complexes based on their electrostatic properties and the accessible surface area (Mendes et al., 2015). A comprehensive review by Baker et al. (2012) is covering these aspects in great detail.

TCR BINDING PREDICTION

The final piece of the epitope recognition puzzle is the interaction of the pMHC complex with the TCR, which represents a very difficult problem for modeling studies and sequence-based predictions. One reason for that is the complex and noncontiguous nature of the interaction interface, with the CDR1 and CDR2 regions of the TCR α and β chains making contacts with the MHC class I molecule and the CDR3 regions directly interacting with the bound peptide (see **Figure 4**). Another major hurdle in predicting TCR recognition is the scarcity of experimentally confirmed TCR complementarity determining regions and the sequences of their respective binding partners on the pMHC complex. For example, one of the first feasibility studies of CDR3 sequence patterns was only based on two immunogenic HIV peptides (De Neuter et al., 2018). An additional complication is posed by the fact that repertoire sequencing combined with immune assays determines antigen-specific clonotypes, but does not yield negative controls, i.e., validated pairs of CDRs and pMHC complexes that do not bind each other.

CDR3β chains appear to always be in contact with the antigen bound to the MHC class I molecule, whereas the direct contact of CDR3α chains to the peptide is not always required (Glanville et al., 2017). The involvement of short linear stretches of CDR3β sequence in peptide-TCR interactions creates the opportunity to cluster TCRs in groups of common specificity

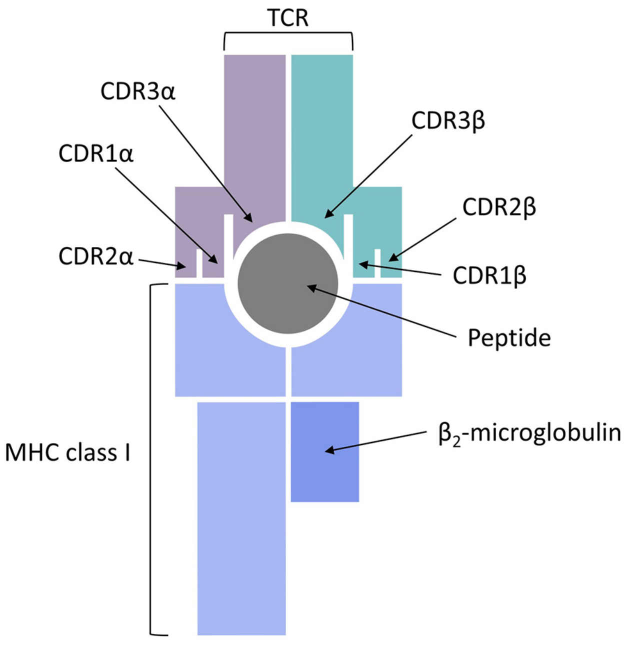


FIGURE 4 | T cell receptor (TCR) binding to a peptide presented by major histocompatibility complex (MHC) class I.

(Dash et al., 2017; Glanville et al., 2017) and also serves as the basis for developing specialized algorithms for sequence-based prediction of pMHC/TCR binding. Two recent publications addressed this problem from two completely different perspectives. Jurtz et al. presented a proof of concept study, in which they predicted TCR interactions with their cognate HLA-A*02:01-presented peptide targets (Jurtz et al., 2018). A machine learning approach, called NetTCR, was trained on 8,920 TCRβ CDR3 sequences and 91 cognate peptide targets obtained from IEDB and from the immune assay data published by Klinger et al. (2015). A dataset of negative interactions was assembled by randomly matching TCR and peptide pairs. The NetTCR project in its current form is limited to a small number of peptides and it does not consider CDR1/CDR2 interactions with the MHC molecules or CDR3α sequences, but it is an important step forward because it demonstrates that TCR recognition of pMHCs is specific enough to be captured by sequence-level prediction tools.

Ogishi and Yotsuyanagi exploited the existence of immunodominant epitopes, which are targeted by the adaptive immune system in different individuals and would therefore be expected to exhibit some prominent features that make them especially prone to be recognized by T cells (Ogishi and Yotsuyanagi, 2019). The idea behind their repertoire-wide TCR-epitope contact potential profiling is that intermolecular contacts between relevant portions of the epitope and the TCR CDR3β region that closely resemble the contact structure of the interactions involving immunodominant peptides would be more likely to be immunogenic. To quantitatively assess the interaction affinity, they used physicochemical properties of amino acids and an energetic potential, calculated as the sum of all pairwise contact potentials for individual amino acids. The latter were obtained from several previously published amino acid contact potential scales, available from the AAINDEX database (Kawashima et al., 2007). These features were converted to immunogenicity scores using machine learning. It should be noted that the knowledge-based potentials, derived from crystal structures of proteins and protein complexes, reflect either intramolecular interactions driving protein folding and stability or contacts at protein interfaces and may only be a coarse approximation of peptide-TCR interactions. Yet, Ogishi and Yotsuyanagi demonstrated that the most informative contact-based and property-based features strongly correlate with experimentally measured TCR-peptide affinities.

Both approaches by Jurtz et al. and Ogishi and Yotsuyanagi are solely based on CDR3β chains and do not incorporate CDR3α sequence information. This is due to the fact that most datasets and databases such as IEDB and VDJdb did, until recently, consist mainly of CDR3β sequences (**Figure 5**)

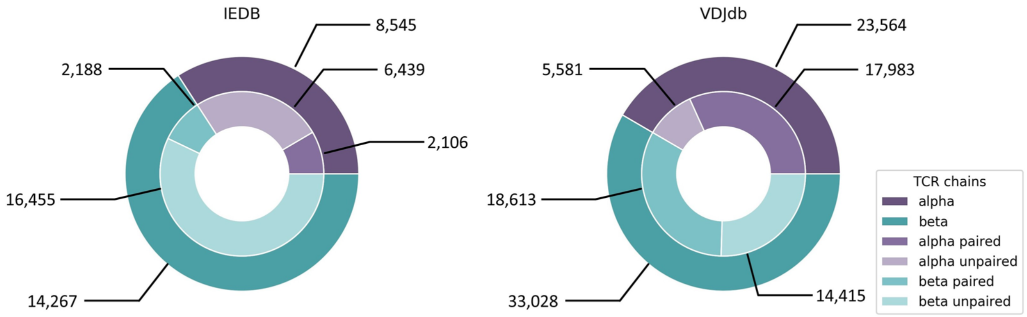


FIGURE 5 | IEDB and VDJdb contents of CDR3α and CDR3β sequences of human origin. IEDB contains 386 unique epitopes linked to CDR3α sequences and 426 unique epitopes linked to CDR3β sequences. For VDJdb there are 93 and 177 unique epitopes, respectively. IEDB data was downloaded from https://www.iedb.org on September 30th, 2019 with the following query parameters: Current Filters: No B cell assays, No major histocompatibility complex (MHC) ligand assays, Restriction Type: Class I, Host: Homo sapiens (human). VDJdb data was taken from https://vdjdb.cdr3.net/overview (last updated on August 7th, 2019).

derived from bulk sequencing (Shugay et al., 2018; Vita et al., 2018), since identifying functional TCR pairing in repertoire data is technically challenging (Holec et al., 2018). Single cell sequencing eliminates this problem and a large dataset has just been added to VDJdb, which is, however, dominated by only few epitopes and HLA alleles. Another problem regarding TCR-epitope data is the lack of true negative datasets and the inclusion of cross-reactivity information, since many TCRs are able to recognize more than one epitope, which has been elaborated in section "Cross-reactivity assessment." For this reason, pMHC/TCR binding prediction would also add valuable information to the detection of potential cross-reactivity for clinical candidate TCRs.

Further light on the details of pMHC/TCR interactions can be shed by molecular dynamics simulations. This entails understanding the role of hydrogen bonds, hydrophobic contacts, and interactions with the solvent in determining the specificity and cross-reactivity of each individual complex and proposing specific models of TCR engagement with the CDR1, CDR2, and CDR3 regions (Cuendet et al., 2011). Moreover, molecular modeling can help to compare the surface morphology between the complexed wild-type and mutated peptides and their relationship with immunogenicity (Park et al., 2013) and can also help to predict affinity-enhancing TCR mutations (Malecek et al., 2014). In cases where three-dimensional structures are not yet available, accurate models of pMHC/TCR complexes can be obtained by homology modeling (Zoete et al., 2013; Lanzarotti et al., 2019). Finally, a number of both rigid and flexible pMHC/TCR docking protocols have been proposed, which, in many cases, are able to produce accurate complex models starting from unbound structures (Pierce and Weng, 2013).

CONCLUSION AND OUTLOOK

Machine learning has become an indispensable tool for immunotherapeutic applications over the last decades. The established core method is peptide binding affinity prediction and thus target identification for TCR-T therapy or personalized neoantigen vaccination. The constant evolution of available training data as well as machine learning techniques, building on growing computational power, has improved the quality of binding affinity predictions. Focus has been on CD8$^+$ cytotoxic T cells, but the substantial role of CD4$^+$ T cells is increasingly gaining attention and efforts are made to also improve predictions for MHC class II presented epitopes, which poses a more challenging task compared to MHC class I binding due to the larger variety in peptide length and the open binding groove (Brown et al., 1993).

Additional challenges which can be tackled by machine learning remain. Immunogenicity is still an elusive aim for prediction tools, especially when it comes to personalized therapies relying on neoepitope identification. This is owed to the fact that patient immune systems and tumors undergo a process of mutual influence and therefore are highly individual and heterogeneous. The identification of features derived from the immune system that affect T cell recognition of individual epitopes within a tumor could be the key toward more reliable personalized immunotherapy predictions, thereby opening the process to a broader number of patients. Although neoantigens are currently in the focus of cancer immunotherapy, the detection of shared tumor antigens beyond coding DNA regions remains necessary since not all tumors harbor enough immunogenic mutations and the creation of potent TCRs for individual patients is currently impossible. Another challenge, which can be tackled with the help of ongoing data acquisition, is TCR binding prediction. Being able to reliably predict which TCR will recognize which epitope is extremely valuable not only for target epitope identification for TCR-T therapies, but also especially for TCR safety assessment, since it can speed up the process of selecting TCRs for the clinic by reducing *in vitro* screening of TCR candidates.

As the TCR-T adoptive immunotherapy community grows and data on the impact of sequence variations in both TCR alpha and beta chains on peptide fine specificity, sensitivity of peptide-MHC recognition and TCR cross-reactivity for partially mismatched epitopes emerge, artificial intelligence in the form of machine learning will be critical to advance understanding of pMHC/TCR interactions for many types of antigen and many different HLA allotypes. In particular, these issues will become additionally relevant as this form of immunotherapy is developed for patient populations worldwide. High-throughput TCR discovery platforms, yielding TCR sequence information from natural repertoires of T cells or through TCR mutational analyses, coupled with functional assessment of peptide variants as a means to assess cross-reactivity, offer many opportunities to continually improve understanding of pMHC/TCR interactions that will not only advance the cause of basic science but also help to meet medical needs for patients with cancer, infectious diseases or autoimmunity, where it is envisioned that TCR-Ts have the potential to provide improved therapies worldwide.

In particular, the push to couple TCR sequence data with neoantigen recognition for single patients through analysis of individual tumor samples in order to develop more potent cancer vaccines or TCR-T immunotherapies has already fostered strong collaborations and commercial endeavors to advance the interplay of machine learning and TCR recognition. While it currently seems daunting to imagine how the enormous and fast flow of information now emerging from many sources can be accessed and assembled to rapidly support the broader needs for personalized patient-individualized TCR-based immunotherapies, this review summarizes the challenges as well as the substantial progress that has already been achieved in defining some of the most relevant parameters in the complex cell biology of antigen processing and presentation and pMHC interactions with TCRs that lead to successful immune recognition. Important gaps have also been defined, alerting the community to the types of control data that may already exist in many laboratories, or could be collected, that would help in the refinement of prediction tools to achieve better results in the future. Increased interest and collaborative efforts of machine learning and HLA and TCR specialists will certainly foster further developments to support the rapidly expanding field of T cell-based immunotherapy of high medical relevance.

With the support of bioinformatic tools and improved prediction algorithms, immunotherapy holds the potential to become more precise, more personalized, and more effective than current cancer treatments—and potentially with fewer side effects.

AUTHOR CONTRIBUTIONS

AM, SR, MW, DS, and DF all contributed to the writing and all approved the content of this review article.

REFERENCES

Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., et al. (2017). Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* 46, 315–326. doi: 10.1016/j.immuni.2017.02.007

Álvaro-Benito, M., Morrison, E., Abualrous, E. T., Kuropka, B., and Freund, C. (2018). Quantification of HLA-DM-dependent major histocompatibility complex of class II immunopeptidomes by the peptide landscape antigenic epitope alignment utility. *Front. Immunol.* 9, 872. doi: 10.3389/fimmu.2018.00872

Anagnostou, V., Smith, K. N., Forde, P. M., Niknafs, N., Bhattacharya, R., White, J., et al. (2017). Evolution of neoantigen landscape during immune checkpoint blockade in non-small cell lung cancer. *Cancer Discovery* 7, 264–276. doi: 10.1158/2159-8290.CD-16-0828

Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 32, 511–517. doi: 10.1093/bioinformatics/btv639

Andreatta, M., Jurtz, V. I., Kaever, T., Sette, A., Peters, B., and Nielsen, M. (2017). Machine learning reveals a non-canonical mode of peptide binding to MHC class II molecules. *Immunology* 152, 255–264. doi: 10.1111/imm.12763

Angelova, M., Charoentong, P., Hackl, H., Fischer, M. L., Snajder, R., Krogsdam, A. M., et al. (2015). Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* 16, 64. doi: 10.1186/s13059-015-0620-6

Anonymous (2017) The problem with neoantigen prediction. *Nat. Biotechnol.* 35, 97–97. doi: 10.1038/nbt.3800

Antunes, D. A., Devaurs, D., Moll, M., Lizée, G., and Kavraki, L. E. (2018). General Prediction of peptide-mhc binding modes using incremental docking: a proof of concept. *Sci. Rep.* 8, 4327. doi: 10.1038/s41598-018-22173-4

Antunes, D. A., Vieira, G. F., Rigo, M. M., Cibulski, S. P., Sinigaglia, M., and Chies, J. A. B. (2010). Structural allele-specific patterns adopted by epitopes in the MHC-I cleft and reconstruction of mhc:peptide complexes to cross-reactivity assessment. *PloS One* 5, e10353. doi: 10.1371/journal.pone.0010353

Bai, Y., Wang, D., and Fury, W., (2018). "PHLAT: Inference of high-resolution HLA types from RNA and whole exome sequencing," in *in HLA Typing*. Ed. Boegel, S. ((New York, NY: Springer New York), 193–201. doi: 10.1007/978-1-4939-8546-3_13

Bailey, M. H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., et al. (2018). Comprehensive characterization of cancer driver genes and mutations. *Cell* 173, 371–385.e18. doi: 10.1016/j.cell.2018.02.060

Bais, P., Namburi, S., Gatti, D. M., Zhang, X., and Chuang, J. H. (2017). CloudNeo: a cloud pipeline for identifying patient-specific tumor neoantigens. *Bioinformatics* 33, 3110–3112. doi: 10.1093/bioinformatics/btx375

Baker, B. M., Scott, D. R., Blevins, S. J., and Hawse, W. F. (2012). Structural and dynamic control of T-cell receptor specificity, cross-reactivity, and binding mechanism. *Immunol Rev.* 250, 10–31. doi: 10.1111/j.1600-065X.2012.01165.x

Bassani-Sternberg, M., and Gfeller, D. (2016). Unsupervised HLA peptidome deconvolution improves ligand prediction accuracy and predicts cooperative effects in peptide–HLA interactions. *J. Immunol.* 197, 2492–2499. doi: 10.4049/jimmunol.1600808

Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P. O., et al. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PloS Comput. Biol.* 13, e1005725. doi: 10.1371/journal.pcbi.1005725

Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* 14, 658–673. doi: 10.1074/mcp.M114.042812

Bhasin, M. (2004). Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci.* 13, 596–607. doi: 10.1110/ps.03373104

Bilich, T., Nelde, A., Bichmann, L., Roerden, M., Salih, H. R., Kowalewski, D. J., et al. (2019). The HLA ligandome landscape of chronic myeloid leukemia delineates novel T-cell epitopes for immunotherapy. *Blood* 133, 550–565. doi: 10.1182/blood-2018-07-866830

Bisset, L. R., and Fierz, W. (1993). Using a neural network to identify potential HLA-DR1 binding sites within proteins. *J. Mol. Recognit.* 6, 41–48. doi: 10.1002/jmr.300060105

Bjerregaard, A.-M., Nielsen, M., Hadrup, S. R., Szallasi, Z., and Eklund, A. C. (2017a). MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother*. 66 (9), 1123–1130 doi: 10.1007/s00262-017-2001-3

Bjerregaard, A.-M., Nielsen, M., Jurtz, V., Barra, C. M., Hadrup, S. R., Szallasi, Z., et al. (2017b). An analysis of natural t cell responses to predicted tumor neoepitopes. *Front. Immunol.* 8, 1566. doi: 10.3389/fimmu.2017.01566

Blankenstein, T., Leisegang, M., Uckert, W., and Schreiber, H. (2015). Targeting cancer-specific mutations by T cell receptor gene therapy. *Curr. Opin. Immunol.* 33, 112–119. doi: 10.1016/j.coi.2015.02.005

Boegel, S., Löwer, M., Bukur, T., Sahin, U., and Castle, J. C. (2014). A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *OncoImmunology* 3, e954893. doi: 10.4161/21624011.2014.954893

Boegel, S., Löwer, M., Schäfer, M., Bukur, T., de Graaf, J., Boisguérin, V., et al. (2012). HLA typing from RNA-Seq sequence reads. *Genome Med.* 4, 102. doi: 10.1186/gm403

Bordner, A. J., and Abagyan, R. (2006). Ab initio prediction of peptide-MHC binding geometry for diverse class I MHC allotypes. *Proteins: Struct Funct. Bioinf.* 63, 512–526. doi: 10.1002/prot.20831

Brown, J. H., Jardetzky, T. S., Gorga, J. C., Stern, L. J., Urban, R. G., Strominger, J. L., et al. (1993). Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1. *Nature* 364, 33–39. doi: 10.1038/364033a0

Brown, S. D., Warren, R. L., Gibb, E. A., Martin, S. D., Spinelli, J. J., Nelson, B. H., et al. (2014). Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Res.* 24, 743–750. doi: 10.1101/gr.165985.113

Brusic, V., Rudy, G., and Harrison, L. C. (1994). MHCPEP: a database of MHC-binding peptides. *Nucleic Acids Res.* 22, 3663–3665. doi: 10.1093/nar/22.17.3663

Bulik-Sullivan, B., Busby, J., Palmer, C. D., Davis, M. J., Murphy, T., Clark, A., et al. (2019). Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* 37, 55–63. doi: 10.1038/nbt.4313

Calis, J. J. A., Maybeno, M., Greenbaum, J. A., Weiskopf, D., De Silva, A. D., Sette, A., et al. (2013). Properties of MHC class I presented peptides that enhance immunogenicity. *PloS Comput. Biol.* 9, e1003266. doi: 10.1371/journal.pcbi.1003266

Caron, E., Vincent, K., Fortier, M.-H., Laverdure, J.-P., Bramoulle, A., Hardy, M.-P., et al. (2011). The MHC I immunopeptidome conveys to the cell surface an integrative view of cellular regulation. *Mol. Syst. Biol.* 7, 533–533. doi: 10.1038/msb.2011.68

Carreno, B. M., Magrini, V., Becker-Hapak, M., Kaabinejadian, S., Hundal, J., Petti, A. A., et al. (2015). A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* 348, 803–808. doi: 10.1126/science.aaa3828

Castle, J. C., Kreiter, S., Diekmann, J., Lower, M., van de Roemer, N., de Graaf, J., et al. (2012). Exploiting the mutanome for tumor vaccination. *Cancer Res.* 72, 1081–1091. doi: 10.1158/0008-5472.CAN-11-3722

Chang, T.-C., Carter, R. A., Li, Y., Li, Y., Wang, H., Edmonson, M. N., et al. (2017). The neoepitope landscape in pediatric cancers. *Genome Med.* 9, 78. doi: 10.1186/s13073-017-0468-3

Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., et al. (2017). Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* 18, 248–262. doi: 10.1016/j.celrep.2016.12.019

Cohen, C. J., Gartner, J. J., Horovitz-Fried, M., Shamalov, K., Trebska-McGowan, K., Bliskovsky, V. V., et al. (2015). Isolation of neoantigen-specific T cells from tumor and peripheral lymphocytes. *J. Clin. Invest.* 125, 3981–3991. doi: 10.1172/JCI82416

Cuendet, M. A., Zoete, V., and Michielin, O. (2011). How T cell receptors interact with peptide-MHCs: A multiple steered molecular dynamics study. *Proteins: Struct Funct. Bioinf.* 79, 3007–3024. doi: 10.1002/prot.23104

Daniel, S., Brusic, V., Caillat-Zucman, S., Petrovsky, N., Harrison, L., Riganelli, D., et al. (1998). Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. *J. Immunol.* 161, 617–624.

Dash, P., Fiore-Gartland, A. J., Hertz, T., Wang, G. C., Sharma, S., Souquette, A., et al. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* 547, 89–93. doi: 10.1038/nature22383

De Neuter, N., Bittremieux, W., Beirnaert, C., Cuypers, B., Mrzic, A., Moris, P., et al. (2018). On the feasibility of mining CD8$^+$ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics*. 70 (3), 159–168 doi: 10.1007/s00251-017-1023-5

Dhanda, S. K., Mahajan, S., Paul, S., Yan, Z., Kim, H., Jespersen, M. C., et al. (2019). IEDB-AR: immune epitope database—analysis resource in 2019. *Nucleic Acids Res.* 47, W502–W506. doi: 10.1093/nar/gkz452

Diez-Rivero, C. M., Chenlo, B., Zuluaga, P., and Reche, P. A. (2010). Quantitative modeling of peptide binding to TAP using support vector machine. *Proteins: Struct Funct. Bioinf.* 78, 63–72. doi: 10.1002/prot.22535

Dolan, B. P. (2019). "Quantitating MHC Class I ligand production and presentation using TCR-like antibodies," in *in Antigen Processing*. Ed. van Endert, P. ((New York, NY: Springer New York), 149–157. doi: 10.1007/978-1-4939-9450-2_12

Doytchinova, I. A., Guan, P., and Flower, D. R. (2006). EpiJen: a server for multistep T cell epitope prediction. *BMC Bioinf.* 7, 131. doi: 10.1186/1471-2105-7-131

Duan, F., Duitama, J., Al Seesi, S., Ayres, C. M., Corcelli, S. A., Pawashe, A. P., et al. (2014). Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* 211, 2231–2248. doi: 10.1084/jem.20141308

Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., et al. (2018). Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* 6, 271–281.e7. doi: 10.1016/j.cels.2018.03.002

Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D. J., Weekes, M. P., et al. (2018). Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods* 15, 363–366. doi: 10.1038/nmeth.4631

Falk, K., Rötzschke, O., Stevanovié, S., Jung, G., and Rammensee, H.-G. (1991). Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. *Nature* 351, 290–296. doi: 10.1038/351290a0

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

Freudenmann, L. K., Marcu, A., and Stevanović, S. (2018). Mapping the tumour human leukocyte antigen (HLA) ligandome by mass spectrometry. *Immunology* 154, 331–345. doi: 10.1111/imm.12936

Garde, C., Ramarathinam, S. H., Jappe, E. C., Nielsen, M., Kringelum, J. V., Trolle, T., et al. (2019). Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics*. 71 (7), 445–454. doi: 10.1007/s00251-019-01122-z

Gfeller, D., and Bassani-Sternberg, M. (2018). Predicting antigen presentation—what could we learn from a million peptides? *Front. Immunol.* 9, 1716. doi: 10.3389/fimmu.2018.01716

Gfeller, D., Guillaume, P., Michaux, J., Pak, H.-S., Daniel, R. T., Racle, J., et al. (2018). The length distribution and multiple specificity of naturally presented HLA-I ligands. *J. Immunol.* 201, 3705–3716. doi: 10.4049/jimmunol.1800914

Ghorani, E., Rosenthal, R., McGranahan, N., Reading, J. L., Lynch, M., Peggs, K. S., et al. (2018). Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. *Ann. Oncol.* 29, 271–279. doi: 10.1093/annonc/mdx687

Giam, K., Ayala-Perez, R., Illing, P. T., Schittenhelm, R. B., Croft, N. P., Purcell, A. W., et al. (2015). A comprehensive analysis of peptides presented by HLA-A1: A comprehensive analysis of peptides. *Tissue Antigens* 85, 492–496. doi: 10.1111/tan.12565

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., et al. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature* 547, 94–98. doi: 10.1038/nature22976

Goh, G., Walradt, T., Markarov, V., Blom, A., Riaz, N., Doumani, R., et al. (2016). Mutational landscape of MCPyV-positive and MCPyV-negative Merkel cell carcinomas with implications for immunotherapy. *Oncotarget* 7, 3403–3415. doi: 10.18632/oncotarget.6494

Griffin, T. A., Nandi, D., Cruz, M., Fehling, H. J., Kaer, L. V., Monaco, J. J., et al. (1998). Immunoproteasome assembly: cooperative incorporation of interferon γ (IFN-γ)-inducible subunits. *J. Exp. Med.* 187, 97–104. doi: 10.1084/jem.187.1.97

Gros, A., Parkhurst, M. R., Tran, E., Pasetto, A., Robbins, P. F., Ilyas, S., et al. (2016). Prospective identification of neoantigen-specific lymphocytes in the peripheral blood of melanoma patients. *Nat. Med.* 22, 433–438. doi: 10.1038/nm.4051

Gubin, M. M., Zhang, X., Schuster, H., Caron, E., Ward, J. P., Noguchi, T., et al. (2014). Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens. *Nature* 515, 577–581. doi: 10.1038/nature13988

Gubler, B., Daniel, S., Armandola, E. A., Hammer, J., Caillat-Zucman, S., and van Endert, P. M. (1998). Substrate selection by transporters associated with antigen processing occurs during peptide binding to TAP. *Mol. Immunol.* 35, 427–433. doi: 10.1016/s0161-5890(98)00059-5

Haase, K., Raffegerst, S., Schendel, D. J., and Frishman, D. (2015). Expitope: a web server for epitope expression. *Bioinformatics* 31, 1854–1856. doi: 10.1093/bioinformatics/btv068

Hassan, C., Kester, M. G. D., Oudgenoeg, G., de Ru, A. H., Janssen, G. M. C., Drijfhout, J. W., et al. (2014). Accurate quantitation of MHC-bound peptides by application of isotopically labeled peptide MHC complexes. *J. Proteomics* 109, 240–244. doi: 10.1016/j.jprot.2014.07.009

Hilf, N., Kuttruff-Coqui, S., Frenzel, K., Bukur, V., Stevanović, S., Gouttefangeas, C., et al. (2019). Actively personalized vaccination trial for newly diagnosed glioblastoma. *Nature* 565, 240–245. doi: 10.1038/s41586-018-0810-y

Holec, P. V., Berleant, J., Bathe, M., and Birnbaum, M. E. (2018). A Bayesian framework for high-throughput T cell receptor pairing. *Bioinformatics*. 35 (8), 1318–1325. doi: 10.1093/bioinformatics/bty801

Hu, Y., Wang, Z., Hu, H., Wan, F., Chen, L., Yuanpeng, X., et al. (2019). ACME: Pan-specific peptide-MHC class I binding prediction through attention-based deep neural networks. *Bioinformatics*. doi: 10.1093/bioinformatics/btz427

Hugo, W., Zaretsky, J. M., Sun, L., Song, C., Moreno, B. H., Hu-Lieskovan, S., et al. (2016). Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell* 165, 35–44. doi: 10.1016/j.cell.2016.02.065

Hundal, J., Kiwala, S., McMichael, J., Miller, C. A., Wollam, A. T., Xia, H., et al. (2019). pVACtools: a computational toolkit to identify and visualize cancer neoantigens. *bioRxiv*. doi: 10.1101/501817

Hunt, D. F., Henderson, R. A., Shabanowitz, J., Sakaguchi, K., Michel, H., Sevilir, N., et al. (1992). Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. *Science* 255, 1261–1263. doi: 10.1126/science.1546328

Jaravine, V., Mösch, A., Raffegerst, S., Schendel, D. J., and Frishman, D. (2017a). Expitope 2.0: a tool to assess immunotherapeutic antigens for their potential cross-reactivity against naturally expressed proteins in human tissues. *BMC Cancer* 17, 892. doi: 10.1186/s12885-017-3854-8

Jaravine, V., Raffegerst, S., Schendel, D. J., and Frishman, D. (2017b). Assessment of cancer and virus antigens for cross-reactivity in human tissues. *Bioinformatics* 33, 104–111. doi: 10.1093/bioinformatics/btw567

Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., et al. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 154, 394–406. doi: 10.1111/imm.12889

Johnson, L. A., Morgan, R. A., Dudley, M. E., Cassard, L., Yang, J. C., Hughes, M. S., et al. (2009). Gene therapy with human and mouse T-cell receptors mediates cancer regression and targets normal tissues expressing cognate antigen. *Blood* 114, 535–546. doi: 10.1182/blood-2009-03-211714

Jørgensen, K. W., Rasmussen, M., Buus, S., and Nielsen, M. (2014). NetMHCstab - predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* 141, 18–26. doi: 10.1111/imm.12160

Jurtz, V. I., Jessen, L. E., Bentzen, A. K., Jespersen, M. C., Mahajan, S., Vita, R., et al. (2018). NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. doi:10.1101/433706

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: improved peptide–MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* 199, 3360–3368. doi: 10.4049/jimmunol.1700893

Kahles, A., Lehmann, K.-V., Toussaint, N. C., Hüser, M., Stark, S. G., Sachsenberg, T., et al. (2018). Comprehensive analysis of alternative splicing across tumors from 8,705 patients. *Cancer Cell* 34, 211–224.e6. doi: 10.1016/j.ccell.2018.07.001

Kalaora, S., Barnea, E., Merhavi-Shoham, E., Qutob, N., Teer, J. K., Shimony, N., et al. (2016). Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* 7, 5110–5117. doi: 10.18632/oncotarget.6960

Kalaora, S., Wolf, Y., Feferman, T., Barnea, E., Greenstein, E., Reshef, D., et al. (2018). Combined analysis of antigen presentation and t-cell recognition reveals restricted immune responses in melanoma. *Cancer Discovery* 8, 1366–1375. doi: 10.1158/2159-8290.CD-17-1418

Karasaki, T., Nagayama, K., Kawashima, M., Hiyama, N., Murayama, T., Kuwano, H., et al. (2016). Identification of individual cancer-specific somatic mutations for neoantigen-based immunotherapy of lung cancer. *J. Thoracic Oncol.* 11, 324–333. doi: 10.1016/j.jtho.2015.11.006

Karasaki, T., Nagayama, K., Kuwano, H., Nitadori, J., Sato, M., Anraku, M., et al. (2017). Prediction and prioritization of neoantigens: integration of RNA sequencing data with whole-exome sequencing. *Cancer Sci.* 108, 170–177. doi: 10.1111/cas.13131

Karosiene, E., Lundegaard, C., Lund, O., and Nielsen, M. (2012). NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 64, 177–186. doi: 10.1007/s00251-011-0579-8

Kato, T., Park, J.-H., Kiyotani, K., Ikeda, Y., Miyoshi, Y., and Nakamura, Y. (2017). Integrated analysis of somatic mutations and immune microenvironment of multiple regions in breast cancers. *Oncotarget* 8, 62029–62038. doi: 10.18632/oncotarget.18790

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2007). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36, D202–D205. doi: 10.1093/nar/gkm998

Keskin, D. B., Anandappa, A. J., Sun, J., Tirosh, I., Mathewson, N. D., Li, S., et al. (2019). Neoantigen vaccine generates intratumoral T cell responses in phase Ib glioblastoma trial. *Nature* 565, 234–239. doi: 10.1038/s41586-018-0792-9

Keşmir, C., Nussbaum, A. K., Schild, H., Detours, V., and Brunak, S. (2002). Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.* 15, 287–296. doi: 10.1093/protein/15.4.287

Khalili, J. S., Hanson, R. W., and Szallasi, Z. (2012). In silico prediction of tumor antigens derived from functional missense mutations of the cancer gene census. *OncoImmunology* 1, 1281–1289. doi: 10.4161/onci.21511

Kim, S., Kim, H. S., Kim, E., Lee, M. G., Shin, E., Paik, S., et al. (2018). Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann. Oncol.* 29 (4), 1030–1036. doi: 10.1093/annonc/mdy022

Kinkead, H. L., Hopkins, A., Lutz, E., Wu, A. A., Yarchoan, M., Cruz, K., et al. (2018). Combining STING-based neoantigen-targeted vaccine with checkpoint modulators enhances antitumor immunity in murine pancreatic cancer. *JCI Insight* 3, e122857. doi: 10.1172/jci.insight.122857

Klausen, M. S., Anderson, M. V., Jespersen, M. C., Nielsen, M., and Marcatili, P. (2015). LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Res.* 43, W349–W355. doi: 10.1093/nar/gkv535

Klinger, M., Pepin, F., Wilkins, J., Asbury, T., Wittkop, T., Zheng, J., et al. (2015). Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PloS One* 10, e0141561. doi: 10.1371/journal.pone.0141561

Koşaloğlu-Yalçın, Z., Lanka, M., Frentzen, A., Logandha Ramamoorthy Premlal, A., Sidney, J., Vaughan, K., et al. (2018). Predicting T cell recognition of MHC class I restricted neoepitopes. *OncoImmunology* 7, e1492508. doi: 10.1080/2162402X.2018.1492508

Koster, J., and Plasterk, R. H. A. (2019). A library of Neo Open Reading Frame peptides (NOPs) as a sustainable resource of common neoantigens in up to 50% of cancer patients. *Sci. Rep.* 9, 6577. doi: 10.1038/s41598-019-42729-2

Kreiter, S., Vormehr, M., van de Roemer, N., Diken, M., Löwer, M., Diekmann, J., et al. (2015). Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* 520, 692–696. doi: 10.1038/nature14426

Kuttler, C., Nussbaum, A. K., Dick, T. P., Rammensee, H.-G., Schild, H., and Hadeler, K.-P. (2000). An algorithm for the prediction of proteasomal cleavages. *J. Mol. Biol.* 298, 417–429. doi: 10.1006/jmbi.2000.3683

Kyeong, H.-H., Choi, Y., and Kim, H.-S. (2018). GradDock: rapid simulation and tailored ranking functions for peptide-MHC Class I docking. *Bioinformatics* 34, 469–476. doi: 10.1093/bioinformatics/btx589

Lam, T., Mamitsuka, H., Ren, E., and Tong, J. (2010). TAP Hunter: a SVM-based system for predicting TAP ligands using local description of amino acid sequence. *Immunome Res.* 6, S6. doi: 10.1186/1745-7580-6-S1-S6

Lanzarotti, E., Marcatili, P., and Nielsen, M. (2019). T-cell receptor cognate target prediction based on paired α and β chain sequence and structural CDR loop similarities. *Front. Immunol.* 10, 2080. doi: 10.3389/fimmu.2019.02080

Larsen, M. V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O., and Nielsen, M. (2007). Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinf.* 8, 424. doi: 10.1186/1471-2105-8-424

Laumont, C. M., and Perreault, C. (2017). Exploiting non-canonical translation to identify new targets for T cell-based cancer immunotherapy. *Cell. Mol. Life Sci*. 75 (4), 607–621. doi: 10.1007/s00018-017-2628-4

Laumont, C. M., Vincent, K., Hesnard, L., Audemard, É., Bonneil, É., Laverdure, J.-P., et al. (2018). Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Trans. Med.* 10, eaau5516. doi: 10.1126/scitranslmed.aau5516

Lin, H., Ray, S., Tongchusak, S., Reinherz, E. L., and Brusic, V. (2008). Evaluation of MHC class I peptide binding prediction servers: Applications for vaccine research. *BMC Immunol.* 9, 8. doi: 10.1186/1471-2172-9-8

Linette, G. P., Stadtmauer, E. A., Maus, M. V., Rapoport, A. P., Levine, B. L., Emery, L., et al. (2013). Cardiovascular toxicity and titin cross-reactivity of affinity-enhanced T cells in myeloma and melanoma. *Blood* 122, 863–871. doi: 10.1182/blood-2013-03-490565

Linnemann, C., van Buuren, M. M., Bies, L., Verdegaal, E. M. E., Schotte, R., Calis, J. J. A., et al. (2014). High-throughput epitope discovery reveals frequent recognition of neo-antigens by CD4$^+$ T cells in human melanoma. *Nat. Med.* 21, 81–85. doi: 10.1038/nm.3773

Liu, C., Yang, X., Duffy, B., Mohanakumar, T., Mitra, R. D., Zody, M. C., et al. (2013). ATHLATES: accurate typing of human leukocyte antigen through exome sequencing. *Nucleic Acids Res.* 41, e142–e142. doi: 10.1093/nar/gkt481

Liu, G., Li, D., Li, Z., Qiu, S., Li, W., Chao, C., et al. (2017). PSSMHCpan: a novel PSSM-based software for predicting class I peptide-HLA binding affinity. *GigaScience* 6, 1–11. doi: 10.1093/gigascience/gix017

Liu, S., Matsuzaki, J., Wei, L., Tsuji, T., Battaglia, S., Hu, Q., et al. (2019). Efficient identification of neoantigen-specific T-cell responses in advanced human ovarian cancer. *J. Immuno Ther Cancer* 7, 156. doi: 10.1186/s40425-019-0629-6

Löffler, M. W., Chandran, P. A., Laske, K., Schroeder, C., Bonzheim, I., Walzer, M., et al. (2016). Personalized peptide vaccine-induced immune response associated with long-term survival of a metastatic cholangiocarcinoma patient. *J. Hepatol* 65, 849–855. doi: 10.1016/j.jhep.2016.06.027

Löffler, M. W., HEPVAC Consortium Mohr, C., Bichmann, L., Freudenmann, L. K., Walzer, M., et al. (2019). Multi-omics discovery of exome-derived neoantigens in hepatocellular carcinoma. *Genome Med.* 11, 28. doi: 10.1186/s13073-019-0636-8

Łuksza, M., Riaz, N., Makarov, V., Balachandran, V. P., Hellmann, M. D., Solovyov, A., et al. (2017). A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature*. 551 (7681), 517–520. doi: 10.1038/nature24473

Lundegaard, C., Lund, O., and Nielsen, M. (2008). Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics* 24, 1397–1398. doi: 10.1093/bioinformatics/btn128

Madden, D. R. (1995). The three-dimensional structure of peptide-MHC complexes. *Annu. Rev. Immunol.* 13, 587–622. doi: 10.1146/annurev.iy.13.040195.003103

Malecek, K., Grigoryan, A., Zhong, S., Gu, W. J., Johnson, L. A., Rosenberg, S. A., et al. (2014). Specific Increase in Potency via Structure-Based Design of a TCR. *J. Immunol.* 193, 2587–2599. doi: 10.4049/jimmunol.1302344

Mamitsuka, H. (1998). Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. *Proteins* 33, 460–474. doi: 10.1002/(sici)1097-0134(19981201)33:4<460::aid-prot2>3.0.co;2-m

Marino, F., Chong, C., Michaux, J., and Bassani-Sternberg, M., (2019). "High-throughput, fast, and sensitive immunopeptidomics sample processing for mass spectrometry," in *in Immune Checkpoint Blockade*. Ed. Pico de Coaña, Y. (New York, NY: Springer New York), 67–79. doi: 10.1007/978-1-4939-8979-9_5

Martin, S. D., Wick, D. A., Nielsen, J. S., Little, N., Holt, R. A., and Nelson, B. H. (2018). A library-based screening method identifies neoantigen-reactive T cells in peripheral blood prior to relapse of ovarian cancer. *OncoImmunology* 7, e1371895. doi: 10.1080/2162402X.2017.1371895

Marty, R., Kaabinejadian, S., Rossell, D., Slifker, M. J., van de Haar, J., Engin, H. B., et al. (2017). MHC-I genotype restricts the oncogenic mutational landscape. *Cell* 171, 1272–1283.e15. doi: 10.1016/j.cell.2017.09.050

McGranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., et al. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* 351, 1463–1469. doi: 10.1126/science.aaf1490

McGranahan, N., Rosenthal, R., Hiley, C. T., Rowan, A. J., Watkins, T. B. K., Wilson, G. A., et al. (2017). Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* 171, 1259–1271.e11. doi: 10.1016/j.cell.2017.10.001

Mendes, M. F. A., Antunes, D. A., Rigo, M. M., Sinigaglia, M., and Vieira, G. F. (2015). Improved structural method for T-cell cross-reactivity prediction. *Mol. Immunol.* 67, 303–310. doi: 10.1016/j.molimm.2015.06.017

Menegatti Rigo, M., Amaral Antunes, D., Vaz de Freitas, M., Fabiano de Almeida Mendes, M., Meira, L., Sinigaglia, M., et al. (2015). DockTope: a web-based tool for automated pMHC-I modelling. *Sci. Rep.* 5, 18413. doi: 10.1038/srep18413

Miller, A., Asmann, Y., Cattaneo, L., Braggio, E., Keats, J., Auclair, D., et al. (2017). High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. *Blood Cancer J.* 7, e612. doi: 10.1038/bcj.2017.94

Morgan, R. A., Chinnasamy, N., Abate-Daga, D., Gros, A., Robbins, P. F., Zheng, Z., et al. (2013). Cancer regression and neurological toxicity following anti-MAGE-A3 TCR gene therapy. *J. Immuno ther* 36, 133–151. doi: 10.1097/CJI.0b013e3182829903

Moutaftsi, M., Peters, B., Pasquetto, V., Tscharke, D. C., Sidney, J., Bui, H.-H., et al. (2006). A consensus epitope prediction approach identifies the breadth of murine T_{CD8+}-cell responses to vaccinia virus. *Nat. Biotechnol.* 24, 817–819. doi: 10.1038/nbt1215

Nathanson, T., Ahuja, A., Rubinsteyn, A., Aksoy, B. A., Hellmann, M. D., Miao, D., et al. (2017). Somatic mutations and neoepitope homology in melanomas treated with CTLA-4 blockade. *Cancer Immunol. Res.* 5, 84–91. doi: 10.1158/2326-6066.CIR-16-0019

Neefjes, J., Jongsma, M. L. M., Paul, P., and Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol*. 11 (12), 823–36. doi: 10.1038/nri3084

Nielsen, M., and Andreatta, M. (2017). NNAlign: a platform to construct and evaluate artificial neural network models of receptor–ligand interactions. *Nucleic Acids Res.* 45, W344–W349. doi: 10.1093/nar/gkx276

Nielsen, M., Justesen, S., Lund, O., Lundegaard, C., and Buus, S. (2010). NetMHCIIpan-2.0 - Improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. *Immunome Res.* 6, 9. doi: 10.1186/1745-7580-6-9

Nielsen, M., Lundegaard, C., and Lund, O. (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinf.* 8, 238 . doi: 10.1186/1471-2105-8-238

Nielsen, M., Lundegaard, C., Lund, O., and Keşmir, C. (2005). The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 57, 33–41. doi: 10.1007/s00251-005-0781-7

Nielsen, M., Lundegaard, C., Worning, P., Lauemøller, S. L., Lamberth, K., Buus, S., et al. (2003). Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci.* 12, 1007–1017. doi: 10.1110/ps.0239403

O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018b). MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* 7 (1), 129–132.e4. doi: 10.1016/j.cels.2018.05.014

O'Donnell, T., Christie, E. L., Ahuja, A., Buros, J., Aksoy, B. A., Bowtell, D. D. L., et al. (2018a). Chemotherapy weakly contributes to predicted neoantigen expression in ovarian cancer. *BMC Cancer* 18, 87. doi: 10.1186/s12885-017-3825-0

Ogishi, M., and Yotsuyanagi, H. (2019). Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space. *Front. Immunol.* 10, 827. doi: 10.3389/fimmu.2019.00827

Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 547, 217–221. doi: 10.1038/nature22991

Park, M.-S., Park, S. Y., Miller, K. R., Collins, E. J., and Lee, H. Y. (2013). Accurate structure prediction of peptide–MHC complexes for identifying highly immunogenic antigens. *Mol. Immunol.* 56, 81–90. doi: 10.1016/j.molimm.2013.04.011

Parker, K. C., Bednarek, M. A., and Coligan, J. E. (1994). Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* 152, 163–175.

Peters, B., Bui, H.-H., Frankild, S., Nielson, M., Lundegaard, C., Kostem, E., et al. (2006). A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PloS Comput. Biol.* 2, e65. doi: 10.1371/journal.pcbi.0020065

Peters, B., Bulik, S., Tampe, R., van Endert, P. M., and Holzhutter, H.-G. (2003). Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol.* 171, 1741–1749. doi: 10.4049/jimmunol.171.4.1741

Pierce, B. G., and Weng, Z. (2013). A flexible docking approach for prediction of T cell receptor-peptide-MHC complexes. *Protein Sci.* 22, 35–46. doi: 10.1002/pro.2181

Rajasagi, M., Shukla, S. A., Fritsch, E. F., Keskin, D. B., DeLuca, D., Carmona, E., et al. (2014). Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* 124, 453–462. doi: 10.1182/blood-2014-04-567933

Rammensee, H. G., Falk, K., and Rötzschke, O. (1993). Peptides naturally presented by MHC class I molecules. *Annu. Rev. Immunol.* 11, 213–244. doi: 10.1146/annurev.iy.11.040193.001241

Rammensee, H. G., Friede, T., and Stevanoviíc, S. (1995). MHC ligands and peptide motifs: first listing. *Immunogenetics* 41, 178–228. doi: 10.1007/bf00172063

Rammensee, H., Bachmann, J., Emmerich, N. P., Bachor, O. A., and Stevanović, S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50, 213–219. doi: 10.1007/s002510050595

Rasmussen, M., Fenoy, E., Harndahl, M., Kristensen, A. B., Nielsen, I. K., Nielsen, M., et al. (2016). Pan-specific prediction of peptide-MHC class I complex stability, a correlate of T cell immunogenicity. *J. Immunol.* 197, 1517–1524. doi: 10.4049/jimmunol.1600582

Reche, P. A., Glutting, J.-P., and Reinherz, E. L. (2002). Prediction of MHC class I binding peptides using profile motifs. *Hum. Immunol.* 63, 701–709. doi: 10.1016/s0198-8859(02)00432-9

Reuben, A., Gittelman, R., Gao, J., Zhang, J., Yusko, E. C., Wu, C.-J., et al. (2017). TCR Repertoire intratumor heterogeneity in localized lung adenocarcinomas: an association with predicted neoantigen heterogeneity and postsurgical recurrence. *Cancer Discovery* 7, 1088–1097. doi: 10.1158/2159-8290.CD-17-0256

Ritz, D., Gloger, A., Neri, D., and Fugmann, T. (2017). Purification of soluble HLA class I complexes from human serum or plasma deliver high quality immuno peptidomes required for biomarker discovery. *PROTEOMICS* 17, 1600364. doi: 10.1002/pmic.201600364

Ritz, D., Gloger, A., Weide, B., Garbe, C., Neri, D., and Fugmann, T. (2016). High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. *PROTEOMICS* 16, 1570–1580. doi: 10.1002/pmic.201500445

Rizvi, N. A., Hellmann, M. D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J. J., et al. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348, 124–128. doi: 10.1126/science.aaa1348

Robbins, P. F., Lu, Y.-C., El-Gamil, M., Li, Y. F., Gross, C., Gartner, J., et al. (2013). Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat. Med.* 19, 747–752. doi: 10.1038/nm.3161

Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. E. (2015). The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.* 43, D423–D431. doi: 10.1093/nar/gku1161

Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G., and Hacohen, N. (2015). Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* 160, 48–61. doi: 10.1016/j.cell.2014.12.033

Rosenthal, R., Cadieux, E. L., Salgado, R., Moore, D. A., Lund, T., Tanić, M., et al. (2019). Neoantigen-directed immune escape in lung cancer evolution. *Nature* 567, 479–485. doi: 10.1038/s41586-019-1032-7

Rothbard, J. B., and Taylor, W. R. (1988). A sequence pattern common to T cell epitopes. *EMBO J.* 7, 93–100. doi: 10.1002/j.1460-2075.1988.tb02787.x

Sadelain, M., Brentjens, R., and Rivière, I. (2013). The basic principles of chimeric antigen receptor design. *Cancer Discovery* 3, 388–398. doi: 10.1158/2159-8290.CD-12-0548

Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.-P., Simon, P., Löwer, M., et al. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* 547, 222–226. doi: 10.1038/nature23003

Schischlik, F., Jäger, R., Rosebrock, F., Hug, E., Schuster, M. K., Holly, R., et al. (2019). Mutational landscape of the transcriptome offers putative targets for immunotherapy of myeloproliferative neoplasms. *Blood*, blood.2019000519. doi:10.1182/blood.2019000519

Schubert, B., Walzer, M., Brachvogel, H.-P., Szolek, A., Mohr, C., and Kohlbacher, O. (2016). FRED 2: an immunoinformatics framework for Python. *Bioinformatics* 32, 2044–2046. doi: 10.1093/bioinformatics/btw113

Schumacher, T. N., and Schreiber, R. D. (2015). Neoantigens in cancer immunotherapy. *Science* 348, 69–74. doi: 10.1126/science.aaa4971

Segal, N. H., Parsons, D. W., Peggs, K. S., Velculescu, V., Kinzler, K. W., Vogelstein, B., et al. (2008). Epitope landscape in breast and colorectal cancer. *Cancer Res.* 68, 889–892. doi: 10.1158/0008-5472.CAN-07-3095

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., et al. (2018). VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res.* 46, D419–D427. doi: 10.1093/nar/gkx760

Shukla, S. A., Rooney, M. S., Rajasagi, M., Tiao, G., Dixon, P. M., Lawrence, M. S., et al. (2015). Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* 33, 1152–1158. doi: 10.1038/nbt.3344

Singh, H., and Raghava, G. P. S. (2001). ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 17, 1236–1237. doi: 10.1093/bioinformatics/17.12.1236

Smart, A. C., Margolis, C. A., Pimentel, H., He, M. X., Miao, D., Adeegbe, D., et al. (2018). Intron retention is a source of neoepitopes in cancer. *Nat. Biotechnol.* 36 (11), 1056–1058. doi: 10.1038/nbt.4239

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J. M., Desrichard, A., et al. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. *New Engl. J. Med.* 371, 2189–2199. doi: 10.1056/NEJMoa1406498

Sonntag, K., Hashimoto, H., Eyrich, M., Menzel, M., Schubach, M., Döcker, D., et al. (2018). Immune monitoring and TCR sequencing of CD4 T cells in a long term responsive patient with metastasized pancreatic ductal carcinoma treated with individualized, neoepitope-derived multipeptide vaccines: a case report. *J. Trans. Med.* 16, 23. doi: 10.1186/s12967-018-1382-1

Spranger, S., Luke, J. J., Bao, R., Zha, Y., Hernandez, K. M., Li, Y., et al. (2016). Density of immunogenic antigens does not explain the presence or absence of the T-cell–inflamed tumor microenvironment in melanoma. *Proc. Natl. Acad. Sci.* 113, E7759–E7768. doi: 10.1073/pnas.1609376113

Storkus, W. J., Zeh, H. J., Salter, R. D., and Lotze, M. T. (1993). Identification of T-cell epitopes: rapid isolation of class I-presented peptides from viable cells by mild acid elution. *J. Immuno ther Emphasis Tumor Immunol.* 14, 94–103.

Stranzl, T., Larsen, M. V., Lundegaard, C., and Nielsen, M. (2010). NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics* 62, 357–368. doi: 10.1007/s00251-010-0441-4

Strønen, E., Toebes, M., Kelderman, S., van Buuren, M. M., Yang, W., van Rooij, N., et al. (2016). Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science* 352, 1337–1341. doi: 10.1126/science.aaf2288

Sugawara, S., Abo, T., and Kumagai, K. (1987). A simple method to eliminate the antigenicity of surface class I MHC molecules from the membrane of viable cells by acid treatment at pH 3. *J. Immunol. Methods* 100, 83–90. doi: 10.1016/0022-1759(87)90175-x

Szolek, A., Schubert, B., Mohr, C., Sturm, M., Feldhahn, M., and Kohlbacher, O. (2014). OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* 30, 3310–3316. doi: 10.1093/bioinformatics/btu548

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T.-H., et al. (2018). The immune landscape of cancer. *Immunity* 48, 812–830.e14. doi: 10.1016/j.immuni.2018.03.023

Tran, E., Ahmadzadeh, M., Lu, Y.-C., Gros, A., Turcotte, S., Robbins, P. F., et al. (2015). Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* 350, 1387–1390. doi: 10.1126/science.aad1253

Trolle, T., McMurtrey, C. P., Sidney, J., Bardet, W., Osborn, S. C., Kaever, T., et al. (2016). The length distribution of class I-restricted t cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *J. Immunol.* 196, 1480–1487. doi: 10.4049/jimmunol.1501721

Van Allen, E. M., Miao, D., Schilling, B., Shukla, S. A., Blank, C., Zimmer, L., et al. (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 350, 207–211. doi: 10.1126/science.aad0095

van Buuren, M. M., Calis, J. J., and Schumacher, T. N. (2014). High sensitivity of cancer exome-based CD8 T cell neo-antigen identification. *OncoImmunology* 3, e28836. doi: 10.4161/onci.28836

van Gool, I. C., Eggink, F. A., Freeman-Mills, L., Stelloo, E., Marchi, E., de Bruyn, M., et al. (2015). POLE proofreading mutations elicit an antitumor immune response in endometrial cancer. *Clin. Cancer Res.* 21, 3347–3355. doi: 10.1158/1078-0432.CCR-15-0057

van Rooij, N., van Buuren, M. M., Philips, D., Velds, A., Toebes, M., Heemskerk, B., et al. (2013). Tumor exome analysis reveals neoantigen-specific T-Cell reactivity in an ipilimumab-responsive melanoma. *J. Clin. Oncol.* 31, e439–e442. doi: 10.1200/JCO.2012.47.7521

Veatch, J. R., Lee, S. M., Fitzgibbon, M., Chow, I.-T., Jesernig, B., Schmitt, T., et al. (2018). Tumor-infiltrating BRAF^{V600E}-specific CD4^{+} T cells correlated with complete clinical response in melanoma. *J. Clin. Invest.* 128, 1563–1568. doi: 10.1172/JCI98689

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2018). The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* 47, D339–D343. doi: 10.1093/nar/gky1006

Vrecko, S., Guenat, D., Mercier-Letondal, P., Faucheu, H., Dosset, M., Royer, B., et al. (2018). Personalized identification of tumor-associated immunogenic neoepitopes in hepatocellular carcinoma in complete remission after sorafenib treatment. *Oncotarget* 9, 35394–35407. doi: 10.18632/oncotarget.26247

Wang, P., Sidney, J., Kim, Y., Sette, A., Lund, O., Nielsen, M., et al. (2010). Peptide binding predictions for HLA DR, DP and DQ molecules. *BMC Bioinf.* 11, 568. doi: 10.1186/1471-2105-11-568

Warren, R. L., and Holt, R. A. (2010). A census of predicted mutational epitopes suitable for immunologic cancer control. *Hum. Immunol.* 71, 245–254. doi: 10.1016/j.humimm.2009.12.007

Warren, R. L., Choe, G., Freeman, D. J., Castellarin, M., Munro, S., Moore, R., et al. (2012). Derivation of HLA types from shotgun sequence datasets. *Genome Med.* 4, 95. doi: 10.1186/gm396

Wood, M. A., Paralkar, M., Paralkar, M. P., Nguyen, A., Struck, A. J., Ellrott, K., et al. (2018). Population-level distribution and putative immunogenicity of cancer neoepitopes. *BMC Cancer* 18, 414. doi: 10.1186/s12885-018-4325-6

Wu, J., Zhao, W., Zhou, B., Su, Z., Gu, X., Zhou, Z., et al. (2018). TSNAdb: a database for tumor-specific neoantigens from immunogenomics data analysis. *Genomics Proteomics Bioinf.* 16, 276–282. doi: 10.1016/j.gpb.2018.06.003

Yadav, M., Jhunjhunwala, S., Phung, Q. T., Lupardus, P., Tanguay, J., Bumbaca, S., et al. (2014). Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* 515, 572–576. doi: 10.1038/nature14001

Yanover, C., and Bradley, P. (2011). Large-scale characterization of peptide-MHC binding landscapes with structural simulations. *Proc. Natl. Acad. Sci. U.S.A.* 108, 6981–6986. doi: 10.1073/pnas.1018165108

Zacharakis, N., Chinnasamy, H., Black, M., Xu, H., Lu, Y.-C., Zheng, Z., et al. (2018). Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer. *Nat. Med.* 24, 724–730. doi: 10.1038/s41591-018-0040-8

Zhang, G. L., Petrovsky, N., Kwoh, C. K., August, J. T., and Brusic, V. (2006). PRED(TAP): a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res.* 2, 3. doi: 10.1186/1745-7580-2-3

Zhang, H., Lund, O., and Nielsen, M. (2009). The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* 25, 1293–1299. doi: 10.1093/bioinformatics/btp137

Zhang, X., Kim, S., Hundal, J., Herndon, J. M., Li, S., Petti, A. A., et al. (2017). Breast cancer neoantigens can induce CD8$^+$ T-cell responses and antitumor immunity. *Cancer Immunol. Res.* 5, 516–523. doi: 10.1158/2326-6066.CIR-16-0264

Zoete, V., Irving, M., Ferber, M., Cuendet, M. A., and Michielin, O. (2013). Structure-based, rational design of T cell receptors. *Front. Immunol.* 4, 268. doi: 10.3389/fimmu.2013.00268

Conflict of Interest: AM, SR, and MW are employees and DS is a Managing Director of Medigene Immunotherapies GmbH, a subsidiary of Medigene AG, Planegg, Germany.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 11 December 2019
doi: 10.3389/fgene.2019.01205

Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice

Nikola Simidjievski[1]†, Cristian Bodnar[1]†, Ifrah Tariq[1,2]†, Paul Scherer[1], Helena Andres Terre[1], Zohreh Shams[1], Mateja Jamnik[1] and Pietro Liò[1]*

[1] Department of Computer Science and Technology, University of Cambridge, Cambridge, United Kingdom, [2] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, United States

OPEN ACCESS

Edited by:
Davide Chicco,
Peter Munk Cardiac Centre,
Canada

Reviewed by:
Emanuel Weitschek,
Università Telematica Internazionale
Uninettuno,
Italy
Samir B. Amin,
Jackson Laboratory for Genomic
Medicine, United States

***Correspondence:**
Nikola Simidjievski
nikola.simidjievski@cl.cam.ac.uk

†These authors have contributed equally to this work

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

Received: 29 July 2019
Accepted: 31 October 2019
Published: 11 December 2019

Citation:
Simidjievski N, Bodnar C, Tariq I, Scherer P, Andres Terre H, Shams Z, Jamnik M and Liò P (2019) Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice. Front. Genet. 10:1205. doi: 10.3389/fgene.2019.01205

International initiatives such as the Molecular Taxonomy of Breast Cancer International Consortium are collecting multiple data sets at different genome-scales with the aim to identify novel cancer bio-markers and predict patient survival. To analyze such data, several machine learning, bioinformatics, and statistical methods have been applied, among them neural networks such as autoencoders. Although these models provide a good statistical learning framework to analyze multi-omic and/or clinical data, there is a distinct lack of work on how to integrate diverse patient data and identify the optimal design best suited to the available data.In this paper, we investigate several autoencoder architectures that integrate a variety of cancer patient data types (e.g., multi-omics and clinical data). We perform extensive analyses of these approaches and provide a clear methodological and computational framework for designing systems that enable clinicians to investigate cancer traits and translate the results into clinical applications. We demonstrate how these networks can be designed, built, and, in particular, applied to tasks of integrative analyses of heterogeneous breast cancer data. The results show that these approaches yield relevant data representations that, in turn, lead to accurate and stable diagnosis.

Keywords: machine learning, cancer–breast cancer, variational autoencoder, deep learning, integrative data analyses, artificial intelligence, bioinformactics, multi-omic analysis

INTRODUCTION

The rapid technological developments in cancer research yield large amounts of complex heterogeneous data on different scales—from molecular to clinical and radiological data. The limited number of samples that can be collected are usually noisy, incompletely annotated, sparse, and high-dimensional (many variables). As much as these high-throughput data acquisition approaches challenge the data-to-discovery process, they drive the development of new sophisticated computational methods for data analysis and interpretation. In particular, the synergy of cancer research and machine learning has led to groundbreaking discoveries in diagnosis, prognosis, and treatment planning for cancer patients (Vial et al., 2018; Levine et al., 2019). Typically, such machine learning methods are developed to address particular complexities inherent in individual data types, separately. While relevant, this approach is sub-optimal since it fails to exploit the inter-dependencies between the different data silos, and is thus often not extendable to analyzing and modeliing more complex biological phenomena (Gomez-Cabrero et al., 2014; Hériché et al., 2019).

To capitalize on the inter-dependencies and relations across heterogeneous types of data about each patient (Yuan et al., 2011; Miotto et al., 2016), integrating multiple types and sources of data is essential. The data-integration paradigm focuses on a fundamental concept—that a complex biological process is a combination of many simpler processes and its function is greater than the sum of its parts. Hence, integrating and simultaneously analyzing different data types offers better understanding of the mechanisms of a biological process and its intrinsic structure. Many studies have addressed and highlighted the importance of data integration at different scales (Gomez-Cabrero et al., 2014; Huang et al., 2017; Karczewski and Snyder, 2018; López de Maturana et al., 2019; Žitnik et al., 2019). In the context of analyzing cancer data, it has been shown that such integrative approaches yield improved performance for accurate diagnosis, survival analysis, and treatment planning (Shen et al., 2009; Kristensen et al., 2014; Thomas et al., 2014; Gevaert et al., 2016; Vial et al., 2018). In particular, Wang et al. (2014) show that, for the case of five different cancer profiles, integrating mRNA expression, DNA methylation, and miRNA data leads to more accurate survival profiles than each of the individual types of data alone. These findings are in line with the ones of (Amin et al., 2014), where the authors point out that gene expression profiles alone are sub-optimal for predicting complete response in patients with multiple myeloma.

In this paper we design and systematically analyze several deep-learning approaches for data integration based on Variational Autoencoders (VAEs) (Kingma and Welling, 2014). VAEs provide an *unsupervised* methodology for generating meaningful (disentangled) latent representations of integrated data. Such approaches can be utilized in two ways. First, the generated latent representations of integrated data can be exploited for analysis by any machine learning technique. Second, our architectures can be deployed on other heterogeneous data sets. We illustrate the functionality and benefit of the designed approaches by applying them to cancer data—this paves the way to improve survival analysis and bio-marker discovery.

There are several existing machine learning approaches that integrate diverse data. These can be classified into three different categories based on how the data is being utilized (Pavlidis et al., 2002; Gevaert et al., 2006): (i) output (or late) integration, (ii) partial (or intermediate) integration, and (iii) full (or early) integration. Output integration relates to methods that model different data separately, the output of which is subsequently combined (Gevaert et al., 2006; Yang et al., 2010; Qi, 2012). Partial integration refers to specifically designed and developed methods that produce a joint model learned from multiple data simultaneously (Gevaert et al., 2006; Wang et al., 2014; Žitnik and Zupan, 2015). Finally, full-integration approaches focus on combining different data before applying a learning algorithm, either by simply aggregating them or learning a common latent representation (Shen et al., 2009; Bengio et al., 2013). Our work presented here falls into this third category, namely full (or early) integration.

Recently, many deep learning approaches have been proposed for analyzing cancer data (Levine et al., 2019). Typically, they rely on extracting valuable features using deep convolutional neural networks for analyzing and classifying tasks of radiological data (Ardila et al., 2019; Esteva et al., 2019). However, these methods often relate to supervised learning, and require many labeled observations in order to perform well. In contrast, unsupervised approaches learn representations by identifying patterns in the data and extracting meaningful knowledge while overcoming data complexities. Particular variants of deep learning networks, referred to as autoencoders, have demonstrated good performance for unsupervised representation learning (Bengio et al., 2013).

Autoencoders learn a compressed representation (embedding/code) of the input data by reconstructing it on the output of the network. The hope is that such a compressed representation captures the structure of the data (i.e., intrinsic relationships between the data variables) and therefore allows for more accurate downstream analyses (Belkin and Niyogi, 2003). Autoencoders have been deployed on a variety of tasks across different data types such as dimensionality reduction, data denoising, compression, and data generation. In the context of cancer data integration, several studies highlighted their utility in combining data on different scales for identifying prognostic cancer traits such as liver (Chaudhary et al., 2018), breast (Tan et al., 2015) and neuroblastoma cancer (Zhang et al., 2018) sub-types. The focus of these studies is to apply autoencoders to specific problems of cancer-data integration.

In contrast, in this paper we investigate approaches that build upon probabilistic autoencoders which implement Variational Bayesian inference for unsupervised learning of latent data representations. Instead of only learning a compressed representation of the input data, VAEs learn the parameters of the underlying distribution of the input data. VAEs can be utilized as methods for full/early integration of data: this allows for learning representations from heterogeneous data on different scales from different sources. In this paper we mainly focus on the data integration aspect, so we utilize VAEs together with other sophisticated machine learning methods for modeling and analyzing breast cancer data. We perform a systematic evaluation (we evaluate 1296 different network configurations) of different aspects of data integration based on VAEs. We investigate and evaluate four different integrative VAE architectures and their components. We analyze and demonstrate their functionality by integrating multi-omics and clinical data for different breast-cancer analysis tasks on data from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort. In summary, the contribution of this paper is two-fold: (i) novel architectures for integrating data; and (ii) methodologies for choosing architectures that best suit the data in hand.

MATERIALS AND METHODS

Many machine learning methodologies have been applied to cancer medicine to improve and personalize diagnosis, survival analysis, and treatment of cancer patients. These include linear and non-linear, as well as supervised and unsupervised techniques like regression, principal component analysis (PCA), support vector machines (SVMs), deep neural networks, and autoencoders (Kourou et al., 2015).

Some are more suitable for integrating diverse types of data than others. In our work we use VAEs and combine them into a number of different architectures for a deep analysis and comparison with respect to specific data features and tasks at hand. VAEs are particularly suitable in this setting since they are generative, non-linear, unsupervised, and amenable to integrating diverse data.

We deploy our architectures on the case of integrating multi-omic and clinical cancer data. There are a number of candidate initiatives for big data collection of cancer data such as The Cancer Genome Atlas (TCGA) and METABRIC. In our work we use the METABRIC data set because it is one of the largest among genetic data sets, it is reasonably well annotated, and it is well analyzed. We particularly focus on the integration of gene expressions, copy number alterations, and clinical data.

In this section we describe theoretical aspects of VAEs and the specialized architectures that we use to integrate data. Next, we describe the data and the suite of experiments used to evaluate the methodological and computational frameworks for investigating cancer traits in clinical applications.

Variational Autoencoders

Generally, an autoencoder consists of two networks, an *encoder* and a *decoder*, which broadly perform the following tasks:

- **Encoder:** Maps the high dimensional input data into a latent variable embedding which has lower dimensions than the input.
- **Decoder:** Attempts to reconstruct the input data from the embedding.

The model contains a decoder function $f(\cdot)$ parameterized by θ and an encoder function $g(\cdot)$ parameterized by ϕ. The lower dimensional embedding learned for an input x in the bottleneck layer is $h = g_\phi(x)$ and the reconstructed input is $x' = f_\theta(g_\phi(x))$.

The parameters $\langle\theta,\phi\rangle$ are learned together to output a reconstructed data sample that is ideally the same as the original input $x \approx f_\theta(g_\phi(x))$. There are various metrics used to quantify the error between the input and output such as cross entropy (CE) or simpler metrics such as mean squared error:

$$L_{AE}(\theta,\phi) = \frac{1}{n}\sum_{i=0}^{n}(x_i - f_0(g_\phi(x_i)))^2.$$

The main challenge when designing an autoencoder is its sensitivity to the input data. While an autoencoder should learn a representation that embeds the key data traits as accurately as possible, it should also be able to encode traits which generalize beyond the original training set and capture similar characteristics in other data sets.

Thus, several variants have been proposed since autoencoders were first introduced. These variants mainly aim to address shortcomings such as improved generalization, disentanglement, and modification to sequence input models. Some significant examples include the Denoising Autoencoder (DAE) (Vincent et al., 2008), Sparse Autoencoder (Coates et al., 2011; Makhzani and Frey, 2014), and more recently the VAE (Kingma and Welling, 2014).

The VAE (**Figure 1**) uses stochastic inference to approximate the latent variables z as probability distributions. These distributions represent and capture relevant features from the input. VAEs are scalable to large data sets, and can deal with intractable posterior distributions by fitting an approximate inference or recognition model, using a reparameterized variational lower bound estimator. They have been broadly tested and used for data compression or dimensionality reduction. Their adaptability and potential to handle non-linear behavior has made them particularly well suited to work with complex data.

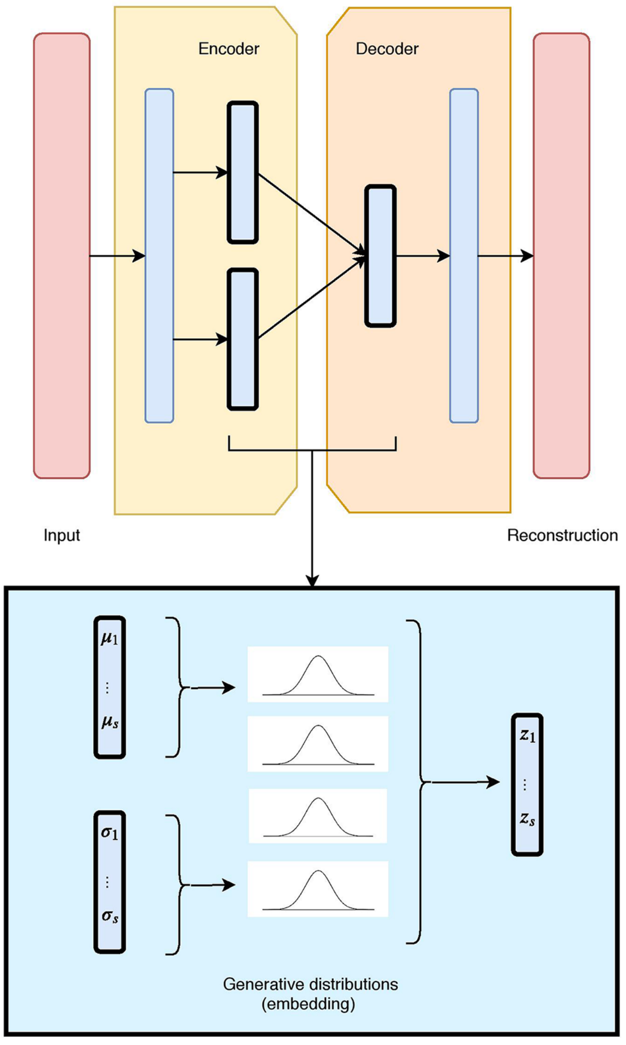


FIGURE 1 | The unimodal Variational Autoencoder (VAE) architecture and latent embedding: the red layers correspond to the input and reconstructed data, given and generated by the model. The hidden layers are in blue, with the embedding framed in black. Each latent component is made of two nodes (mean and standard deviation), which define a Gaussian distribution. The combination of all Gaussian constitutes the VAE generative embedding.

A VAE builds upon a probabilistic framework where the high dimensional data x is drawn from a random variable with distribution $p_{data}(x)$. It assumes that the natural data x also lies in a lower dimensional space, that can be characterized by an unobserved continuous random variable z. In the Bayesian approach, the prior $p_\theta(z)$ and conditional (or likelihood) $p_\theta(x|z)$

typically come from a family of parametric distributions, with Probability Density Functions differentiable almost everywhere with respect to both θ and z. While the true parameters θ and the values of the latent variables z are unknown, the VAE approximates the often intractable true posterior $p_\theta(x|z)$ by using a recognition model $q_\theta(z|x)$ and the learned parameters ϕ represented by the weights of a neural network.

More specifically, a VAE builds an inference or a recognition model $q_\theta(z|x)$, where given a data-point x it produces a distribution over the latent values z from where it could have been drawn. This is also called a probabilistic encoder. A probabilistic decoder will then, given a certain value of z, produce a distribution over the possible corresponding values of x, therefore constructing the likelihood $p_\theta(x|z)$. Note that the decoder is also a generative model, since the likelihood $p_\theta(x|z)$ can be used to map from the latent to the original space and learn to reconstruct the inputs as well as generate new ones.

Typically, VAE model assumes latent variables to be the centred isotropic multivariate Gaussian $p_\phi(z) = N(z;0, I)$, and $p_\theta(x|z)$ a multivariate Gaussian (for numerical values) or Bernoulli (for categorical values) with parameters approximated by using a fully connected neural network. Since the true posterior $p_\theta(z|x)$ is intractable, we assume it takes the form of a Gaussian with an approximately diagonal covariance. This allows the variational inference alternative to approximate the true posterior, as it converts the inference problem into an optimization one. In particular, instead of solving intractable integrals, this relates to maximizing a likelihood. In such cases, the variational approximate posterior will also need to be a multivariate Gaussian with diagonal covariate structure:

$$q_\phi\left(z\,|\,x^{(i)}\right) = N\left(z;\mu^{(i)},\sigma^{(i)}I\right)$$

where the mean $\mu^{(i)}$ and standard deviation $\sigma^{(i)}$ are outputs of the encoder.

Since $p_\theta(z)$ and $q_\phi(z|x^{(i)})$ are Gaussian, the discrepancy between them can be directly computed and differentiated. The resulting likelihood for this model on data-point $x^{(i)}$ is:

$$l_i(\theta,\phi) = -E_{q_\phi(z|x^{(i)})}[\log p_\theta(x\,|\,z)] + \mathrm{KL}(q_\phi(z\,|\,x^{(\mathrm{i})})\,||\,p_\theta(z)),$$

where the first term corresponds to the reconstruction loss, which encourages the decoder to learn to reconstruct the data from the embedding space. The second term is regularization, and measures the divergence between the encoding distributions $q(z|x)$ and $p(z)$, and penalizes the entanglement between components in the latent space. It is typically estimated by the Kullback–Leibler (KL) divergence, a measure of discrepancy between two probability distributions, which in this case is applied between the prior and the representation.

While in this paper we focus on a standard Gaussian prior due to its simplicity, there are several, more sophisticated, alternatives for the choice of a prior. In particular, Dilokthanakul et al. (2016) propose a mixture of Gaussians in order to achieve more flexible priors, and Tomczak and Welling (2018) realize this by estimating the prior as a mixture of approximate posteriors. Nalisnick and Smyth (2017) employ a Dirichlet process as a non-parametric prior through stick-breaking process, which generalizes over the generative process and allows for better representations. Johnson et al. (2016) utilize graphical models as a prior to train a VAE model. These alternative approaches to the choice of a prior require more sophisticated model training techniques in the learning phase. On the other hand, there are also approaches that instead of the prior, they focus on more flexible posteriors, therefore leading to better (and disentangled) representations. These include normalizing flows (Rezende and Mohamed, 2015), auto-regressive flows (Chen et al., 2017), and inverse auto-regressive flows (Kingma et al., 2016).

In a similar context, research has shown that the entanglement factor can play a crucial role in the quality of the representations. In response, Higgins et al. (2017) control the influence of the disentanglement factor using a parameter β. Moreover, some approaches have experimented with different regularization terms, such as the InfoVAE (Zhao et al., 2017), where Maximum Mean Discrepancy (MMD) is employed as an alternative to KL divergence. MMD (Gretton et al., 2007) is based on the concept that two distributions are identical if, and only if. all their moments are identical. Therefore, by employing MMD *via* the kernel embedding trick, the divergence can be defined as the discrepancy between the moments of two distributions $p(z)$ and $q(z)$ as:

$$\begin{aligned} MMD(q(z)\,||\,p(z)) = {} & E_{p(z),p(z')}[k(z,z')] \\ & + \mathrm{E}_{q(z),q(z')}[k(z,z')] - 2E_{q(z),p(z')}[\mathrm{k}(z,z')] \end{aligned}$$

where $k(z,z')$ denotes any universal kernel (Zhao et al., 2019). In this paper, we employ a Gaussian kernel $k(z,z') = e^{-\frac{\|z-z'\|^2}{2\sigma^2}}$ when considering MMD regularization in the objective function.

Variational Autoencoders for Data Integration

We designed and evaluated four different architectures for data integration: we present them here each with two diverse data sources (depicted in **Figures 2**, **3**, **4**, and **5** as red and green boxes on the left).

The first architecture, **Variational Autoencoder with Concatenated Inputs (CNC-VAE)** in **Figure 2**, is a simple approach to integration, where the encoder is directly trained from different data sets, aligned, and concatenated at input. While such architecture is a straightforward and not a novel way to data integration, we employ it both, as a benchmark and a proof-of-principle for learning a homogeneous representation from heterogeneous data sources.

Besides the concatenated input, the rest of the CNC-VAE network utilizes a standard VAE architecture. As depicted in **Figure 2**, the input data is first scaled, aligned, and concatenated before being fed to the network. CNC-VAE has one objective function that reconstructs the combined data rather than a

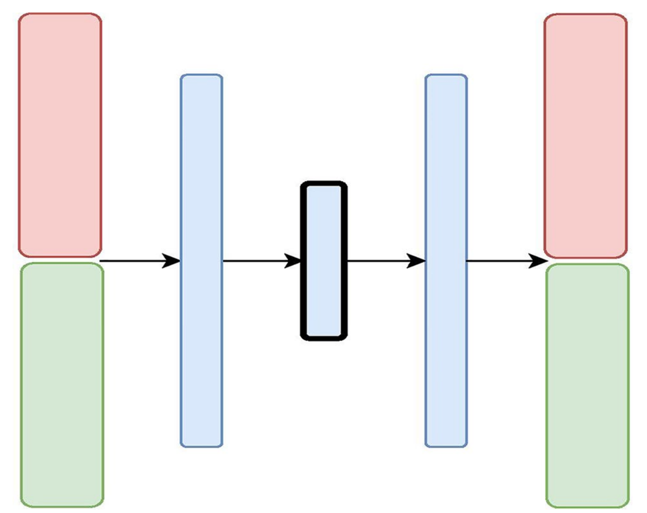


FIGURE 2 | The Variational Autoencoder with Concatenated Inputs (CNC-VAE) Architecture: the red and green layers on the left correspond to two inputs from different data sources. The blue layers are shared, with the embedding being framed in black.

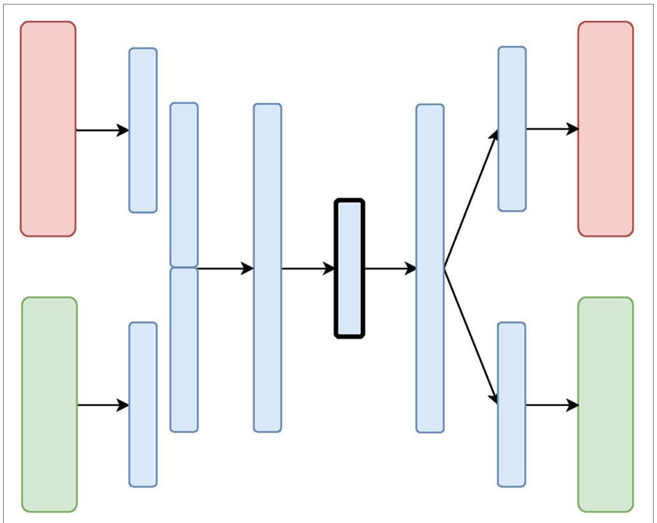


FIGURE 3 | The X-shaped Variational Autoencoder (X-VAE) Architecture: the red and green layers on the left correspond to two inputs from different data sources. The blue layers are shared, with the embedding being framed in black.

separate objective function for each input data source. Therefore, CNC-VAE aims at reducing redundancies and extracting meaningful structure across all input sources, regardless of the scales or modalities of the data. While the CNC-VAE architecture may be simplistic, the complexity lies in highly domain-specific preprocessing of the data. Indeed, in some real-world settings, utilizing a single objective function of combined heterogenious inputs may not be optimal or even feasible.

Unlike CNC-VAE, the next three architectures aim at more sophisticated means to data integration. In particular, all of them consider data integration in the hidden layers. The

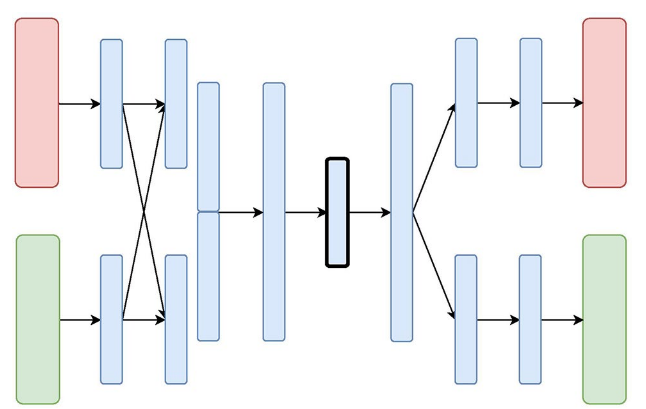


FIGURE 4 | The Mixed-Modal Variational Autoencoder (MM-VAE) Architecture: the red and green layers on the left correspond to two inputs from different data sources. The blue layers are shared, with the embedding being framed in black.

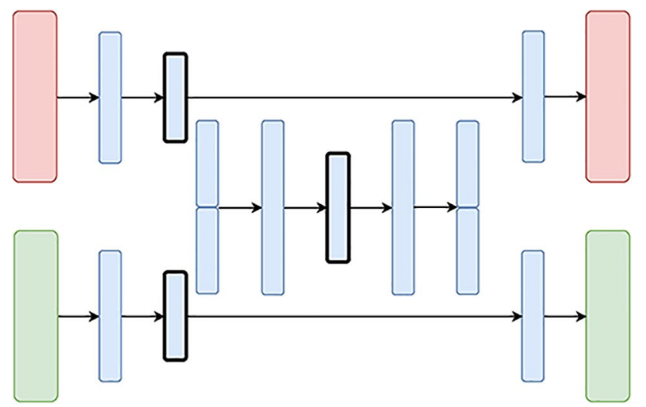


FIGURE 5 | The Hierarchical Variational Autoencoder (H-VAE) Architecture: the red and green layers on the left correspond to two inputs from different data sources. The blue layers are shared, with the embedding being framed in black.

X-shaped Variational Autoencoder (X-VAE) merges high-level representations of several heterogeneous data sources into a single latent representation by learning to reconstruct the input data from the common homogeneous representation. The architecture is depicted in **Figure 3** and consists of individual branches (one for each data source: red and green) that are combined into one before the bottleneck layer. In the decoding phase, the merged branch splits again into several branches that produce individual reconstructions of the inputs. X-VAE takes into account different data modalities by combining different loss functions for each data source in the objective function. This allows for learning better and more meaningful representations.

While, in principle, X-VAE is able to take into account many possible interactions between multiple data sources, its performance is sensitive to the properties of the data being integrated. In particular, X-VAE is prone to poor performance when employed to integrate unbalanced data sets with low number of observations. As a consequence, the objective function might also be unbalanced, focusing on some sources more if the distribution of the input data varies substantially across the data sources. A similar limitation can also result from a poor choice of loss function for each of the data sources.

The **Mixed-Modal Variational Autoencoder (MM-VAE)** attempts to address some of the limitations of X-VAE, by employing a more gradual integration in the hidden layers of the encoder. More specifically, it builds upon the concept of transfer learning, where learned concepts from one domain are re-purposed and shared for learning tasks in others domains. **Figure 4** presents the architecture of MM-VAE. Similarly to X-VAE, it also consists of branches that individually reconstruct the input data sources. Here, however, the important difference is that the branches share information with each other in the encoding phase. In particular, higher-level learned concepts of each branch are shared between all the branches, and used deeper in the network. This allows for information from the different sources to be combined more gradually before being compressed into a single homogeneous embedding.

The objective function combines different reconstruction loss functions that correspond to the data types at input. Similarly to X-VAE, MM-VAE's performance is limited when small and unbalanced data sets are being considered. While the additional integration layers may help to stabilize the objective function, poor choice of reconstruction loss terms may still impede the performance in general.

The **Hierarchical Variational Autoencoder (H-VAE)** builds upon traditional meta-learning approaches for combining multiple individual models. H-VAE, depicted in **Figure 5**, is comprised of several low-level VAEs that relate to each data source separately, and the result is assembled together in a high-level VAE. More specifically, each of the low-level VAEs is employed to learn a representation of an individual data source. These individual representations are then merged together and fed to a high-level VAE that produces the integrated data representation. We use the same architecture for each low-level VAE, but in principle, these could be independently designed and further refined for a specific data-source and task at hand.

H-VAE is designed to improve on some of the shortcomings of X-VAE and MM-VAE, since it simplifies the individual network branches. In particular, the input to the high-level autoencoder is composed of representations learned from several individual low-level autoencoders. These low-level autoencoders already implement distribution regularization terms in each of them separately, thus the input to the high-level autoencoder already consists of approximated multivariate standard normal distributions characterizing the general traits of the individual input modalities. Moreover, since each data source is handled in a modular fashion, H-VAEs are capable of handling data sets which make best use of specialized low-level autoencoders. However, constructing an H-VAE adds a substantial computational overhead compared to the other three architectures as it involves a two-stage learning process where low-level VAEs must be trained first, and then the final high-level representation can be learned on the outputs of the low-level encoders.

Data

To demonstrate how the proposed VAE architectures can be utilized in the integration of heterogeneous cancer data types, we conducted our study utilizing multi-omics data found on somatic copy number aberrations (CNA), mRNA expression data, as well as on the clinical data of breast cancer patient samples from the METABRIC cohort (Curtis et al., 2012).

Providing effective treatment takes such heterogeneity of data into account, and our VAE architectures enable us to do just that. Finding driver events which help stratify breast cancers into different subgroups has been of great focus within the research community lately, particularly the identification of genomic profiles that stratify patients.

In the context of genomic and transcriptomic studies, the acquired somatic mutations and the inherited genomic variation contribute jointly to tumorigenesis, disease onset, and progression (Curtis et al., 2012; Tan et al., 2015; Pereira et al., 2016). For example, despite somatic CNAs being the dominant feature found in sporadic breast cancer cases, the elucidation of driver events in tumorigenesis is hampered by the large variety of random non-pathogenic passenger alterations and copy number variants (Leary et al., 2008; Bignell et al., 2010).

This has led to the argument that integrative approaches for the available information are necessary to make richer assessments of disease sub-categorization (Curtis et al., 2012). A pioneering work that advocates this perspective in breast cancer research is the METABRIC initiative. The METABRIC project is a Canada–UK initiative that aims to group breast cancers based on multiple genomic, transcriptomic, and image data types recorded over 2000+ patient samples. This data set represents one of the largest global studies of breast cancer tissues performed to date. Similarly to (Curtis et al., 2012) we focus on integrating CNA and mRNA expression data, but in addition integrate clinical data too. We use integrative VAEs to showcase how such architectures can be designed, built, and used for cancer studies of this kind.

Experimental Setup

What follows is an outline of our experimental evaluation used to verify that the studied approaches produce valid representations and can be employed for data integration. The aim of this evaluation is threefold. First, for each of the architectures, we seek the optimal configuration in terms of choosing an appropriate objective function and parameters of the network. Second, we aim to evaluate and choose the most appropriate architectures for our data-integration tasks. In particular, we perform a comparative quantitative analysis of the representations obtained from each of the architectures based on different data sets at input. Finally, we discuss the findings in terms of their application to cancer data integration and provide a qualitative (visual) analysis of the obtained representations.

In particular, we tackle several classification tasks by integrating three data types from the METABRIC data—CNA, mRNA expression, and clinical data. We evaluate the predictive performance of the integrative approaches by combining clinical and mRNA data, CNA and mRNA data as well as clinical and CNA data, separately. The METABRIC data consists of 1,980 breast-cancer patients assigned to different groups according to:

- two immunohistochemistry (IHC) sub-types (ER+ and ER−),
- six intrinsic gene-expression sub-types (PAM50) (Prat et al., 2010), and
- 10 IntegrativeCluster (IntClust) sub-types (Curtis et al., 2012).

These patients are also assigned to two groups based on whether or not the cancer metastasised to another organ after the initial treatment (i.e., Distance Relapse). The three cancer sub-types and the distance relapse variable (described with gene expression profiles, CNA profiles, and clinical variables for each patient), are used as target variables in the classification tasks performed in the study.

To control our study, we followed Curtis et al. (2012) and used a pre-selected set of the input CNA and mRNA features. In particular, we used the most significant *cis*-acting genes that are significantly associated with CNAs determined by a gene-centric ANOVA test. We selected the genes with the most significant Bonferroni adjusted p-value from the Illumina database containing 30,566 probes. After missing-data removal, the input data sets consisted of 1000 features of normalized gene expression numerical data, scaled to [0,1], and 1000 features of copy number categorical data. The clinical data included various categorical and numerical features such as: age of the patient at diagnosis, breast tumor laterality, the Nottingham Prognostic Index, inferred menopausal state, number of positive lymph nodes, size and grade of the tumor, as well as chemo-, hormone-, and radio-therapy regimes. Numerical features were discretized and subsequently one-hot encoded. This was combined with the categorical features, yielding 350 clinical features. Finally, all three data sets were sampled into five-fold cross-validation splits for each classification tasks separately, stratified according to the class distribution of the four target variables, respectively. Note that these splits remained the same for all experiments in the study.

While our four architectures differ in some key aspects related to how and where (on which level) they integrate data, for experimental purposes of this study, the depth of the architectures remained moderate, and constant across all experiments. In particular, in all designs except for MM-VAE, the encoder and decoder were symmetric and consisted of compression/decompression dense layers placed before and after data merging. MM-VAE implemented an additional data-merging layer in the encoder network. Therefore, all of the architectures had a moderate depth between two and four hidden layers. The optimal output size of these layers was evaluated for different values of 128,256 and 512. Moreover, all layers used batch normalization (Ioffe and Szegedy, 2015) with Exponential Linear Unit (Clevert et al., 2016) activations (except for the bottleneck and the output layers). All of the architectures also employed a hidden dropout component with a rate of 0.2. Note that the final layers of the CNA and clinical branches employed sigmoid activation function. The models were trained for 150 epochs using an Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.001 (with exponential decay rates of first- and second-moment estimates β_1 $_=$ 0.9 and β_2 $_=$ 0.999) and a batch size of 64. Furthermore, we also investigated the performance of representations with different sizes. For each of the architectures and their configurations, we learned and evaluated representations with sizes 16, 32, and 64.

In the experiments we also considered choosing an optimal objective function that would improve the disentanglement between the embedded components. The objective functions consider both the reconstruction loss and a regularization term. For the former, given that we integrated heterogeneous data, we incorporated Binary Cross Entropy loss for the categorical and Mean Squared Error loss for the continuous data. Note that, while the CNA data is categorical and so multivariate categorical distribution would be suitable, an approach such as one-hot encoding would substantially increase the data dimensionality. Therefore, we employed label smoothing (Salimans et al., 2016), where the form of $p_\theta(x_{cna}|z)$ is a multivariate Bernoulli distribution, with values of x_{cna} scaled to [0,1]. For the regularization terms, we evaluated different options which include weighted KL divergence and weighted MMD. We tested different values of weight β, $\beta \in \{1,10,15,25,50,100\}$, for each of the two regularization terms.

To make optimal design decisions, we evaluated the quality of the representations obtained from our four integrative architectures on three integrative tasks, each of these with 108 different network configurations with respect to the hyper-parameters outlined above. In particular, we evaluated the performance of a given configuration by training a predictive model on the produced representations and measuring its predictive performance on a binary classification task of IHC cancer sub-types (ER+ and ER−). For all network configurations, we trained and evaluated a Gaussian naive Bayes classifier, since it does not require tuning of additional hyper-parameters for the downstream task. We performed a five-fold cross-validation and report the average accuracy.

Once we identified the appropriate configuration for each of the architectures, we evaluated the quality of the learned representation in terms of predictive performance on the remaining three classification tasks. In particular, we evaluated the performance of three different methods trained on different representation. These included Gaussian naive Bayes classifier, SVMs (with RBF kernel C = 1.5 and gamma set to $1/N_f$, where N_f denotes the number of features) and Random Forest (with 50 trees and 1/2 of the features considered at every split). For all three classification tasks we also performed a five-fold cross-validation and report the average accuracy. We also compared these results with the performance of predictive models trained on: (i) the raw (un-compressed) data, as well as (ii) data transformed using PCA (a linear method for data transformation).

The integrative VAE architectures are implemented using the Keras deep learning library (Chollet et al., 2015) with Tensorflow backend. The code for training and evaluating the performance of the VAE networks is available on this repository.[1]

Finally, we visually inspected the learned representations of the whole data set obtained from each of the architectures, and compared them to the uncompressed data. For this task we employed the t-distributed stochastic neighboring embedding (tSNE) (van der Maaten and Hinton, 2008) algorithm.

[1] https://github.com/CancerAI-CL/IntegrativeVAEs

RESULTS

We present and discuss the results of the empirical evaluation. First, we report on the analyses for identifying the suitable design choices within the integrative approaches. Next, we present the results of the analyses of predictive performance of three different predictive methods applied to representations obtained from our VAE architectures with the optimal configuration. Finally, we present a visual analysis of the learned representations obtained from the evaluated architectures.

Design of Integrative VAEs

For each integrative task, we investigated 108 different configurations for each architecture. These highlighted the effect of the size of the learned embedding, the optimal size of each of the dense layers, the most appropriate regularization in the objective function, and how much this regularization should influence the overall loss. We evaluated these configurations for all four architectures on three integrative tasks, by comparing the average train and test performance of classifying IHC sub-typed patients. The results, in general, indicate that properties of these configurations for each architecture are consistent across the three integrative tasks. Therefore, for brevity, here we only present the results when combining clinical and mRNA data. The rest of the results, namely for combining CNA and mRNA, and CNA and clinical data are given in the **Supplementary Material**.

Figure 6 presents the downstream performance of predictive models, trained on the representations produced by the integrative VAEs on clinical and mRNA data. In particular, **Figures 6A–D** compare the performance from representations obtained from CNC-VAE, X-VAE, MM-VAEm and H-VAE, respectively. In general, the configurations regularized with MMD yield better representations that lead to substantially more accurate predictions than the configurations regularized with KL. In terms of the weight of the regularization term, the configurations are robust in general, with moderately large weights (β = [25,50]) leading to slightly better results.

In term of the size of the dense layers, all architectures except H-VAE exhibit stable behavior, with moderate sizes of (*size* = [128,256]) leading to slightly better representations than the ones with dense layer size of 512 in the case of X-VAE and MM-VAE. In the case of H-VAE, the quality of the representations is more affected by the size of the layer where smaller sizes lead to better performance than larger ones.

Considering the size of the latent space, the networks that produce higher-dimensional encodings lead to better predictive performance. This is particularly the case for X-VAE and MM-VAE architectures, while the other two are mostly unaffected. Note however, that the influence of the size of the representations on the overall performance is also related to the integrative task. More specifically, for this particular classification task, higher-dimensional representations when integrating clinical and mRNA data yield better and more stable performance overall. In contrast, when integrating clinical/CNA or CNA/mRNA data lower-dimensional representations are better.

In summary, based on these results, we made the following design decisions for configuring the integrative VAE architectures for the rest of the experimental analyses. First, the networks were trained using the MMD regularization with β = 50, since in all cases using MMD exhibited better performance than the networks trained using KL divergence with various levels of β. Next, we set the size of the dense layers to 256. Finally, since large sizes of the latent space yielded better performance, we set it to be 64.

Quality of the Learned Representations

In this set of experiments, we focused on testing our central hypothesis that the integrative VAE architectures are able to produce representations that yield stable and improved predictive performance. We evaluated their performance in three classification tasks: predicting IC10, PAM50 sub-types, and Distance Relapse.

We used three standard predictive methods: Naive Bayes, SVM, and Random Forest. These were deployed: (i) on representations learned (compressed) from data integrated through our four VAE architectures; (ii) on embedded combined data using PCA with 64 components; (iii) on combined raw (un-compressed) data; and (iv) on each of the data sources separately in order to evaluate the integrative effect. Apart from this last case, the data sources for integration were CNA/mRNA, clinical/mRNA, and clinical/CNA data, as before.

Table 1 summarises the results of this analysis. In general, all of the VAE integrative architectures outperform the baselines on all three predictive tasks when integrating CNA/mRNA, clinical/mRNA data, and clinical/CNA. Overall, all architectures produce better representations when integrating clinical and mRNA data. This result is consistent across all three tasks, where the learned representations coupled with SVMs yield the best predictive performance. This finding is also supported by the benchmark approaches, where combining clinical and mRNA data yields better results than CNA/mRNA and clinical/CNA. Note that, for the task of predicting Distance Relapse, integrating clinical/CNA exhibits, in general, slightly worse but comparable performance to the one produced for clinical/mRNA. These results suggest that for our particular classification tasks, some data types are more beneficial to integrate than others.

We note that while VAEs lead to more accurate predictions, this performance improvement is not significant when compared to PCA. We conjecture that this might be an artifact of many linear relations present in the data, which are captured by the PCA. In contrast, the integrative VAEs are also able to model the non-linearities in the data, which gives them a performance advantage.

Comparing the performance of the four VAE architectures, H-VAE and X-VAE mostly yield more accurate predictions, however, the difference is not significant. Overall, for these three tasks, H-VAE produces more stable and better quality predictions when applied for integrating clinical and mRNA data, given the design decisions outlined previously. While for simplicity we made the same design choices for all architectures, the performance of these models can be further improved, with

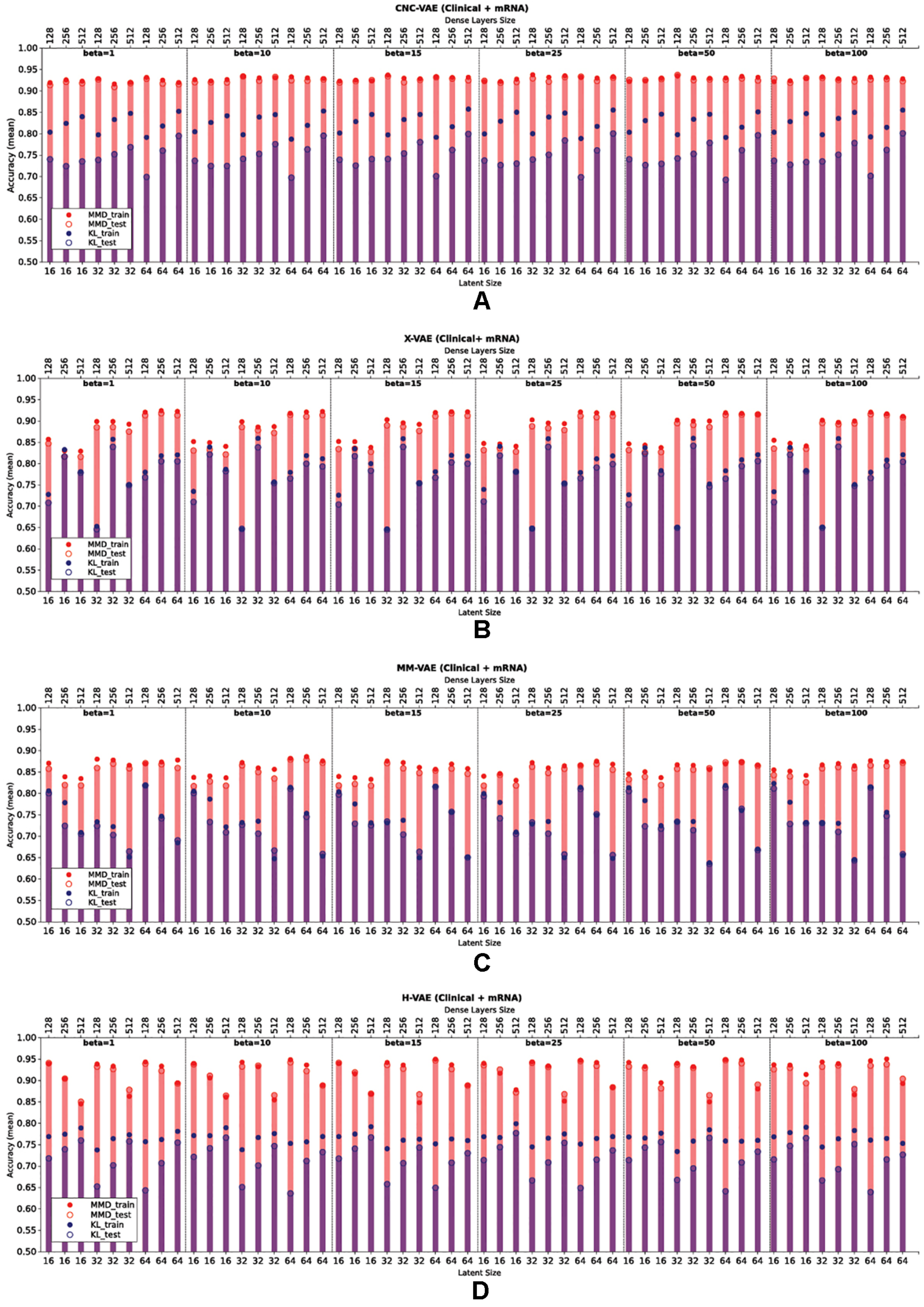


FIGURE 6 | Comparison of the downstream performance on the IHC classification tasks of a predictive model trained on the representations produced by integrating clinical and mRNA data using **(A)** CNC-VAE, **(B)** X-VAE, **(C)** MM-VAE, and **(D)** H-VAE. Full circles denote the training accuracy, while empty circles and bars denote the test accuracy averaged over five-fold cross-validation. Red and blue colors denote the configurations when Maximum Mean Discrepancy (MMD) and Kullback–Leibler (KL) are employed, respectively. Bottom x-axis depicts the size of the latent dimension, while the top x-axis the size of the dense layers of each configuration.

TABLE 1 | Comparison of the downstream predictive performance (on three classification tasks) of the three predictive models trained on raw and PCA-transformed data as well as representations produced by the four integrative Variational Autoencoders (VAEs) by integrating copy number aberration (CNA)/mRNA, clinical/mRNA, and clinical/CNA data.

		DR			PAM50			IC10		
		NB	**SVM**	**RF**	**NB**	**SVM**	**RF**	**NB**	**SVM**	**RF**
CNC-VAE	CNA + mRNA	0.648	0.687	0.684	0.731	0.789	0.749	0.742	0.823	0.784
	Clin. + mRNA	0.732	0.750	0.711	0.784	**0.827**	0.750	*0.829*	0.834	0.781
	Clin. + CNA	0.682	0.751	0.711	0.563	0.624	0.503	0.612	0.657	0.485
X-VAE	CNA + mRNA	0.639	0.687	0.685	0.715	0.788	0.751	0.747	0.835	0.785
	Clin. + mRNA	0.751	**0.774**	0.735	0.787	0.816	0.758	0.821	**0.858**	0.781
	Clin. + CNA	0.695	0.772	0.724	0.576	0.628	0.517	0.627	0.679	0.487
MM-VAE	CNA + mRNA	0.659	0.693	0.688	0.739	0.774	0.759	0.774	0.841	*0.799*
	Clin. + mRNA	0.744	0.756	0.731	*0.803*	0.800	0.760	0.824	0.838	0.781
	Clin. + CNA	0.746	0.770	0.732	0.587	0.605	0.508	0.604	0.621	0.477
H-VAE	CNA + mRNA	0.656	0.687	0.683	0.724	0.792	0.744	0.746	0.816	0.792
	Clin. + mRNA	0.748	**0.774**	0.746	0.790	**0.827**	*0.768*	0.794	0.839	0.776
	Clin. + CNA	0.728	0.761	0.732	0.525	0.579	0.469	0.477	0.594	0.393
PCA	CNA + mRNA	0.628	0.694	0.682	0.595	0.696	0.632	0.639	0.766	0.675
	Clin. + mRNA	0.729	0.754	0.724	0.708	0.771	0.693	0.761	0.828	0.702
	Clin. + CNA	0.673	0.745	0.733	0.562	0.621	0.560	0.601	0.669	0.606
Raw data	CNA + mRNA	0.618	0.696	0.677	0.528	0.581	0.730	0.723	0.664	0.763
	Clin. + mRNA	0.754	0.696	0.748	0.492	0.596	0.739	0.344	0.530	0.780
	Clin. + CNA	0.757	0.696	*0.763*	0.407	0.539	0.617	0.517	0.615	0.646
Raw data	Only CNA	0.609	0.696	0.647	0.430	0.523	0.568	0.621	0.604	0.624
	Only mRNA	0.612	0.696	0.687	0.646	0.604	0.730	0.769	0.633	0.774
	Only clinical	*0.757*	0.708	0.747	0.265	0.363	0.437	0.110	0.181	0.259

Italic typeface denotes the best performance obtained by a particular method for a particular classification task. Bold typeface denotes the best-performing method for the particular classification task.

CNC-VAE, Variational Autoencoder with Concatenated Inputs; X-VAE, X-shaped Variational Autoencoder; MM-VAE, Mixed-Modal Variational Autoencoder; H-VAE, Hierarchical Variational Autoencoder.

careful calibration of both the architecture components as well as the hyper-parameters of the classifier considered.

Qualitative Analyses

In the last set of experiments, we visually inspected the learned representations of the whole data set, obtained from the H-VAE by integrating clinical/mRNA data. Using tSNE diagrams, shown in **Figure 7**, we compared the level of disentanglement of the embedded data with both, raw (uncompressed) data as well as PCA-transformed data. The tSNE projections clearly show that H-VAE is able to produce more sparse and disentangled representations in comparison to raw or PCA transformed data. Note that the t-SNE projections of the raw and PCA-transformed data also indicate data separability. This may explain the competitive performance produced by the benchmark classifiers in the previous section, as well as the advantage of integrating clinical and mRNA data.

DISCUSSION

In this study we investigated and evaluated aspects of VAE architectures important for integrative data analyses. We designed and implemented four integrative VAE architectures, and demonstrated their utility in integrating multi-omics and clinical breast-cancer data. We systematically experimented (we evaluated 1296 different network configurations) with how the data should be integrated as well as what appropriate architecture parameters produce high-quality, low-dimensional representations. In the case of integrating breast-cancer data we found that the choice of an appropriate regularization when training the autoencoders is imperative. Our results show that the integrative VAEs yield better (and more disentangled) representations when MMD is employed, which also corresponds to findings from other studies (Zhao et al., 2017; Chen et al., 2018). Moreover, we found that giving a moderately large weight to this regularization term further improves the quality of the learned representations. The results show that the quality of the representations is mostly invariant to the size of the hidden layers and the embedding dimension, suggesting that the investigated integrative architectures are robust. Note however, that such parameters are task-specific, and therefore it is recommended that they are tuned according to the dimensionality of the input data as well as the depth of the network.

In the context of performance, all four integrative VAE architectures are generally able to produce better representations of the data when compared to a linear transformation approach. This suggests that the integrative VAEs are able to accurately model the non-linearities present in the integrated data, while still being able to reduce the data-dimensionality, leading to good representations. When comparing the different architectures, the results showed that overall the H-VAE and X-VAE exhibit the best performance, followed by the simple CNC-VAE and MM-VAE. This indicates that, while all of the architectures are able to accurately model the data, H-VAE exhibits more stable behavior. Moreover, given that H-VAE is a hierarchical model,

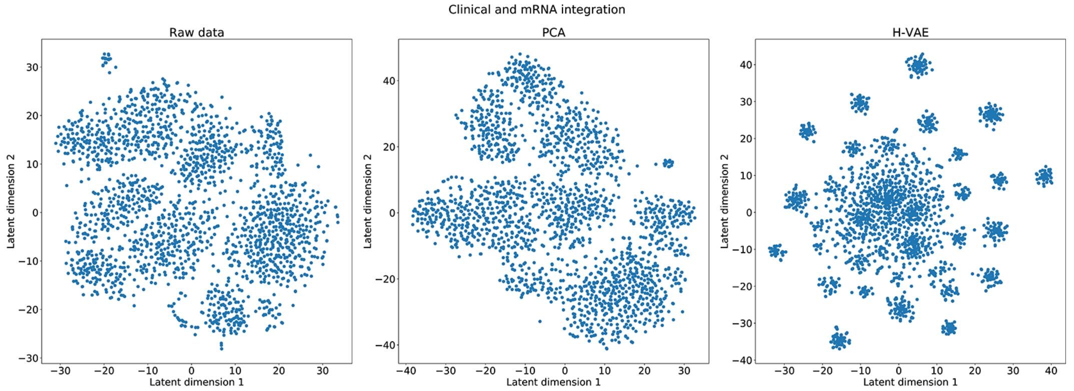


FIGURE 7 | Qualitative comparison of the learned representations with H-VAE, raw data, and PCA-transformed data when integrating clinical and mRNA data.

all of the learned representations (including the intermediate ones from the low-level autoencoders) can be further utilized for more delicate, interpretable analyses. Note however, when employing H-VAE, there is a trade-off between the quality of the learned representations and the time required for learning them. Therefore, when time or resources are limited, employing X-VAE or even the simple CNC-VAE will yield favourable results.

In terms of integrative analyses of breast-cancer data, the results indicate that, for the particular classification tasks considered in our study, some data types are more amenable to integrating than others. More specifically, utilizing the VAEs for integrating clinical and mRNA data coupled with the right classification method led to better downstream predictive performance than the alternative integration of CNA and mRNA data. This highlights an important aspect of this study: for premium results in such integrative data analyses, one should not only focus on the choice and tuning of appropriate predictive methods, but also on the type of data at input. In other words, rather than considering separate components of the analysis, one should focus on the whole end-to-end integrative process.

Autoencoders have been used for learning representations and analyzing transcriptomic cancer data before. In particular, our work relates to Way and Greene (2018), since it employs VAEs for constructing latent representations and analyzing transcriptomic cancer data. The authors show that VAEs can be utilized for knowledge extraction from gene expression pan-cancer TCGA data (TCGA et al., 2013), thus reducing the dimensionality of the single, homogeneous data source while still being able to identify patterns related to different cancer types. Our work is also related to Tan et al. (2015), where the authors deploy DAE for integrating and analyzing gene-expression data from TCGA (TCGA et al., 2013) and METABRIC (Curtis et al., 2012). Tan et al. (2015) also employ DAE for learning latent features from multiple data sets. The latent features are used to identify genes relevant to two different breast cancer sub-types.

In contrast to Curtis et al. (2012) and Tan et al. (2015), we designed novel VAE architectures for integrating heterogeneous data, hence enabling learning patterns that relate to the intrinsic relationships between different data types. While DAEs aim at learning an embedded representation of the input, the VAEs focus on learning the underlying distribution of the input data. Therefore, besides data integration, the methods proposed in this paper can be also employed for data generation.

More generally, our work relates to other approaches based on autoencoders for data integration on various tasks of cancer diagnosis and survival analysis. These include using DAEs for integrating various types of electronic health records (Miotto et al., 2016) as well as custom designed autoencoders for analyses of liver (Chaudhary et al., 2018), bladder (Poirion et al., 2018), and neuroblastoma (Zhang et al., 2018) cancer types.

In a broader context, our work is related to the long tradition of data integration approaches for addressing various challenges in cancer analyses. In particular, Curtis et al. (2012) present an approach for clustering breast-cancer patients based on integrated data from the METABRIC cohort. The approach uses the Integrative Clustering method (Shen et al., 2009) which produces clusters from a multi-omic joint latent embedding. These clusters are then utilized for identifying mutation-driver genes (Pereira et al., 2016) and survival analyses (Rueda et al., 2019). In this context, the work presented in this paper can be readily applied to similar tasks. In particular, the integrative VAEs can be used to learn common representations of the heterogeneous data at input, which can then be used for constructing clusters that address the aforementioned analysis tasks. In contrast to the Integrative Clustering method, the integrative VAEs can handle high-dimensional data sources, which provide better integration and therefore may further improve the overall performance.

In a similar context, the Similarity Network Fusionmethod by Wang et al. (2014) successfully addresses intermediate heterogeneous data integration for identifying cancer sub-types for various kinds of cancers including glioblastoma, breast, kidney, and lung carcinoma. Similarity Network Fusion first constructs graphs from the individual data sources, which are in turn combined into a single, integrative, graph using nonlinear similarity approach. Such graphs can be also used in conjunction with the integrative VAEs. More specifically, by using such graphs will impose a structure of the integrative data, which in turn may lead to far better (and disentangled) representations. Next, Gevaert et al. (2006) present a data integration approach with Bayesian networks for predicting breast cancer prognosis. The authors report that employing Bayesian networks for intermediate integration yields better performance for the particular predictive task. Since our proposed VAE approaches address full data integration, they can also be readily used together with the aforementioned integrative approaches.

We identified several additional directions for future work. First, the experiments reported in this study are limited to integrating heterogeneous multi-omics data from two sources. While in principle the autoencoder designs allow for integrating heterogeneous data from many more sources simultaneously, we intend to empirically evaluate the generality of our approaches and extend them to other types of data such as imaging data. Next, considering the specific architecture decisions made in this paper, we plan to further refine the designed architecture and fine-tune the learning hyper-parameters in ordered to improve the quality of the learned representations. This includes experimenting with deeper architectures as well as implementing methods that allow for more sophisticated priors as well as methods that focus on more flexible posteriors (Rezende and Mohamed, 2015; Kingma et al., 2016). Finally, we intend to ensemble the various proposed architectures which should yield more stable and robust findings, and take a step further towards producing more meaningful and interpretable findings.

While VAEs are capable of generating useful representations for vast amounts of complex heterogeneous data, in terms of interpretability, the biological relevance of the learned representations has to be verified if they are to be used in clinical decision support systems. Previous work (Tan et al., 2015) has attempted to interpret latent features, wherein features which were most influential in deciding clinical phenomena such ER/IHC status were extracted and identified. However, the actual interpretations of these features have received comparatively little attention. In order to interpret extracted VAE features and bring explanation to the learned representations, biological and biomedical ontologies such as gene ontology (GO[2]) have proven very useful (Titus et al., 2018; Way and Greene, 2018). An immediate continuation of the work presented in this paper is performing enrichment analysis on genes most related to each VAEs' learned embedding to investigate the joint effects of various gene sets within specific biological pathways. Tools such as ShinyGo[3] allow KEGG Pathway Mapping[4], where the relationships between genes and human disease including various types of cancer can be identified. Using this approach to interpretability can potentially offer a qualitative metric to evaluate and compare different VAE architectures based on the biological relevance of the features extracted from learned representations to breast cancer and other cancer types in general.

CONCLUSION

In conclusion, in this study we demonstrate the utility of VAEs for full data integration. The design and the analyses of different integrative VAE architectures and configurations, and in particular their application to the tasks of integrative modeliing and analyzing heterogeneous breast cancer data, are the main contributions of this paper.

The studied approaches have several distinguishing properties. First, they are able to produce representations that capture the structure (i.e., intrinsic relationships between the data variables) of the data and therefore allow for more accurate downstream analyses. Second, they are able to reduce the dimensionality of the input data without loss of quality or performance. Therefore, in the process of compressing the input data, they can reduce noise implicitly present in the data. Third, they are modular and easily extendable to handle integration of a multitude of heterogeneous data sets. Next, while the integrative VAEs can be used as a data pre-proccessing approach for learning representations, they can also be utilized in a more generative setting for producing surrogate data, which can be used for more in-depth analysis. Finally, we show that VAEs can be successfully applied to learn representations in complex integrative tasks, such as integrative analyses of breast cancer data, that ultimately lead to more accurate and stable diagnoses.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher. The code used in this study is available at https://github.com/CancerAI-CL/IntegrativeVAEs.

AUTHOR CONTRIBUTIONS

MJ and PL initiated the study. All authors designed the study. CB and IT designed the methods. NS, CB, and IT implemented the methods. NS and PS performed the analysis. All authors analyzed the results. NS, MJ, PS, HT, and ZS wrote the manuscript. All authors reviewed and refined the manuscript.

ACKNOWLEDGMENTS

This work was supported by The Mark Foundation Institute for Integrated Cancer Medicine (MFICM). MFICM is hosted at the University of Cambridge, with funding from The Mark Foundation for Cancer Research (NY, U. S. A.) and the Cancer Research UK Cambridge Centre [C9685/A25177] (UK). We thank Dr. Jean Abraham and Dr. Oscar Rueda for the helpful feedback and discussions on the work presented in this paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01205/full#supplementary-material

[2] http://geneontology.org
[3] http://bioinformatics.sdstate.edu/go/
[4] https://www.genome.jp/kegg/pathway.html#mapping

REFERENCES

Amin, S. B., Yip, W.-K., Minvielle, S., Broyl, A., Li, Y., Hanlon, B., et al. (2014). Gene expression profile alone is inadequate in predicting complete response in multiple myeloma. *Leukemia* 28, 2229–2234. doi: 10.1038/leu.2014.140

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., et al. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* 25, 954–961. doi: 10.1038/s41591-019-0447-x

Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15, 1373–1396. doi: 10.1162/089976603321780317

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Bignell, G. R., Greenman, C. D., Davies, H., Butler, A. P., Edkins, S., Andrews, J. M., et al. (2010). Signatures of mutation and selection in the cancer genome. *Nature* 463, 893. doi: 10.1038/nature08768

Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer research: an Off. J. Am. Assoc. Cancer Res.* 24, 1248–1259. doi: 10.1158/1078-0432.CCR-17-0853

Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., et al. (2017). "Variational lossy autoencoder," in *Proceedings of 5th International Conference on Learning Representations*, ICLR 2017. (Toulon, France: OpenReview.net Conference Track Proceedings).

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders, in *Advances in Neural Information Processing Systems 31*. Eds. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Red Hook, NY, U. S. A.: Curran Associates, Inc.), 2610–2620.

Chollet, F., et al. (2015). *Keras*, . Tech. rep. Available at: https://keras.io/getting-started/faq/#how-should-i-cite-keras

Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (ELUs). In *Proceedings of the 4th International Conference on Learning Representations*, *ICLR 2016*, San Juan, Puerto Rico, Eds. Y. Bengio and Y. LeCun (Conference Track Proceedings).

Coates, A., Ng, A., and Lee, H. (2011). "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Eds. G. Gordon, D. Dunson, and M. Dudík (Fort Lauderdale, FL, USA: PMLR), 215–223. vol. 15 of *Proceedings of Machine Learning Research*.

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Dilokthanakul, N., Mediano, P. A. M., Garnelo, M., Lee, M. C. H., Salimbeni, H., Arulkumaran, K., et al. (2016). Deep unsupervised clustering with gaussian mixture variational autoencoders. *CoRR*.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., et al. (2019). A guide to deep learning in healthcare. *Nat. Med.* 25, 24–29. doi: 10.1038/s41591-018-0316-z

Gevaert, O., Smet, F. D., Timmerman, D., Moreau, Y., and Moor, B. D. (2006). Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics* 22, e184–e190. doi: 10.1093/bioinformatics/btl230

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 Suppl 2, I1–I1. doi: 10.1186/1752-0509-8-S2-I1

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). "A kernel method for the two-sample-problem," in *Advances in Neural Information Processing Systems 19*. Eds. B. Schölkopf, J. C. Platt, and T. Hoffman (Cambridge, MA, U. S. A.: MIT Press), 513–520.

Hériché, J.-K., Alexander, S., and Ellenberg, J. (2019). Integrating imaging and omics: Computational methods and challenges. *Annu. Rev. Biomed. Data Sci.* 2, null. doi: 10.1146/annurev-biodatasci-080917-013328

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., et al. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework, in *Proceedings of 5th International Conference on Learning Representations*, *ICLR 2017*. (Toulon, France: OpenReview.net Conference Track Proceedings).

Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. In Genet.* 8, 84–84. doi: 10.3389/fgene.2017.00084

Ioffe, S., and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift, in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*. Eds. F. R. Bach and D. M. Blei (Lille, France: JMLR.org Workshop and Conference Proceedings), 37.

Johnson, M. J., Duvenaud, D., Wiltschko, A. B., Datta, S. R., and Adams, R. P. (2016). Structured vaes: Composing probabilistic graphical models and variational autoencoders, in *Advances in Neural Information Processing Systems 29*. Eds. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett (NY, U. S. A.: Curran Associates, Inc., Red Hook), 2946–2954.

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 Review Article.

Kingma, D. P., and Ba, J. (2015). Adam: A method for stochastic optimization, in *Proceedings of the 3rd International Conference on Learning Representations*, *ICLR 2015*. Eds. Y. Bengio, and Y. LeCun (San Diego, CA, U. S. A.: Conference Track Proceedings).

Kingma, D. P., and Welling, M. (2014). "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations*, *ICLR 2014*. Eds. Y. Bengio, and Y. LeCun (Banff, AB, Canada: Conference Track Proceedings).

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., and Welling, M. (2016). Improving variational autoencoders with inverse autoregressive flow, in *Advances in Neural Information Processing Systems 29*, Eds. Lee, D. D. and Sugiyama, M. and Luxburg, U. V. and Guyon, I. and Garnett, R. (NY, U. S. A.: Curran Associates, Inc., Red Hook), 4743–4751.

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Kristensen, V. N., Lingjærde, O. C., Russnes, H. G., Vollan, H. K. M., Frigessi, A., and Børresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. *Nat. Rev. Cancer* 14, 299. doi: 10.1038/nrc3721

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

Leary, R. J., Lin, J. C., Cummins, J., Boca, S., Wood, L. D., Parsons, D. W., et al. (2008). Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16224–16229. doi: 10.1073/pnas.0808041105

Levine, A. B., Schlosser, C., Grewal, J., Coope, R., Jones, S. J. M., and Yip, S. (2019). Rise of the machines: Advances in deep learning for cancer diagnosis. *Trends In Cancer* 5, 157–169. doi: 10.1016/j.trecan.2019.02.002

Makhzani, A., and Frey, B. J. (2014). k-sparse autoencoders. In *Proceedings of the 2nd International Conference on Learning Representations*, *ICLR 2014*. Eds. Y. Bengio and Y. LeCun (Banff, AB, Canada: Conference Track Proceedings)

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* 6, 26094. doi: 10.1038/srep26094

Nalisnick, E., and Smyth, P. (2017). Stick-breaking variational autoencoders, in *Proceedings of 5th International Conference on Learning Representations*, *ICLR 2017*. (Toulon, France: OpenReview.net Conference Track Proceedings).

Pavlidis, P., Weston, J., Cai, J., and Noble, W. S. (2002). Learning gene functional classifications from multiple data types. *J. Comput. Biol.* 9, 401–411. doi: 10.1089/10665270252935539

Pereira, B., Chin, S.-F., Rueda, O. M., Vollan, H.-K. M., Provenzano, E., Bardwell, H. A., et al. (2016). The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* 7, 11479. doi: 10.1038/ncomms11479

Poirion, O. B., Chaudhary, K., and Garmire, L. X. (2018). Deep learning data integration for better risk stratification models of bladder cancer. *AMIA Jt Summits Trans. Sci. Proc.* 2017, 197–206.

Prat, A., Parker, J. S., Karginova, O., Fan, C., Livasy, C., Herschkowitz, J. I., et al. (2010). Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res.* 12, R68. doi: 10.1186/bcr2635

Qi, Y. (2012). *Random Forest for Bioinformatics* (Boston, MA: Springer US), 307–323. doi: 10.1007/978-1-4419-9326-7_11

Rezende, D., and Mohamed, S. (2015). "Variational inference with normalizing flows," in *Proceedings of the 32nd ICML*. Eds. F. Bach, and D. Blei (Lille, France: PMLR). vol. 37 of *Proceedings of Machine Learning Research*, 1530–1538.

Rueda, O. M., Sammut, S.-J., Seoane, J. A., Chin, S.-F., Caswell-Jin, J. L., Callari, M., et al. (2019). Dynamics of breast-cancer relapse reveal late-recurring er-positive genomic subgroups. *Nature* 567, 399–404. doi: 10.1038/s41586-019-1007-8

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans, in *Advances in Neural Information Processing Systems 29*. Eds. D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon and R. Garnett (NY, U. S. A.: Curran Associates, Inc., Red Hook), Inc., 2234–2242.

Shen, R., Olshen, A. B., and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25, 2906–2912. doi: 10.1093/bioinformatics/btp543

Tan, J., Ung, M., Cheng, C., and Greene, C. (2015). Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac. Symp. Biocomput.* 20, 132–143. doi: 10.1142/9789814644730_0014

TCGA, N., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi: 10.1038/ng.2764

Thomas, M., De Brabanter, K., Suykens, J. A. K., and De Moor, B. (2014). Predicting breast cancer using an expression values weighted clinical classifier. *BMC Bioinf.* 15, 411–411. doi: 10.1186/s12859-014-0411-1

Titus, A. J., Wilkins, O. M., Bobak, C. A., and Christensen, B. C. (2018). An unsupervised deep learning framework with variational autoencoders for genome-wide dna methylation analysis and biologic feature extraction applied to breast cancer. *bioRxiv*. doi: 10.1101/433763

Tomczak, J. M., and Welling, M. (2018). Vae with a vampprior, in *Proceedings of the International Conference on Artificial Intelligence and Statistics, AISTATS 2018*. Eds. A. J. Storkey and F. Perez-Cruz (Lanzarote, Canary Islands, Spain: Proceedings of Machine Learning Research, PMLR) vol. 84.

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

Vial, A., Stirling, D., Field, M., Ros, M., Ritz, C., Carolan, M., et al. (2018). The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review. *Trans. Cancer Res.* 7 (3), 803–816. doi: 10.21037/tcr.2018.05.02

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders, in *Proceedings of the 25th ICML (ACM), ICML '08*, (New York, NY, U. S. A.: ACM International Conference Proceeding Series, ACM).1096–1103. doi: 10.1145/1390156.1390294

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 EP–. doi: 10.1038/nmeth.2810

Way, G. P., and Greene, C. (2018). Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. *Pac. Symp. Biocomput.* 23, 80–91. doi: 10.1142/9789813235533_0008

Yang, P., Yang, Y. H., B. Zhou, B., and Y.Zomaya, A. (2010). A review of ensemble methods in bioinformatics. *Curr. Bioinf.* 5, 296–308. doi: 10.2174/157489310794072508

Yuan, Y., Savage, R. S., and Markowetz, F. (2011). Patient-specific data fusion defines prognostic cancer subtypes. *PloS Comput. Biol.* 7, 1–12. doi: 10.1371/journal.pcbi.1002227

Zhang, L., Lv, C., Jin, Y., Cheng, G., Fu, Y., Yuan, D., et al. (2018). Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. *Front. In Genet.* 9, 477–477. doi: 10.3389/fgene.2018.00477

Zhao, S., Song, J., and Ermon, S. (2019). InfoVAE: Balancing Learning and Inference in Variational Autoencoders. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019*, Honolulu, Hawaii, U. S. A. Palo Alto, CA. USA; AAAI Press, 5885–5892. doi: 10.1609/aaai.v33i01.33015885

Žitnik, M., and Zupan, B. (2015). Data fusion by matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 41–53. doi: 10.1109/TPAMI.2014.2343973

Žitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. *Inf. Fusion* 50, 71– 91. doi: 10.1016/j.inffus.2018.09.012

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

METHODS
published: 08 January 2020
doi: 10.3389/fgene.2019.01305

A Pretraining-Retraining Strategy of Deep Learning Improves Cell-Specific Enhancer Predictions

Xiaohui Niu[†], *Kun Yang*[†], *Ge Zhang*, *Zhiquan Yang* and *Xuehai Hu*[*]

College of Informatics, Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan, China

Deciphering the code of cis-regulatory element (CRE) is one of the core issues of today's biology. Enhancers are distal CREs and play significant roles in gene transcriptional regulation. Although identifications of enhancer locations across the whole genome [discriminative enhancer predictions (DEP)] is necessary, it is more important to predict in which specific cell or tissue types, they will be activated and functional [tissue-specific enhancer predictions (TSEP)]. Although existing deep learning models achieved great successes in DEP, they cannot be directly employed in TSEP because a specific cell or tissue type only has a limited number of available enhancer samples for training. Here, we first adopted a reported deep learning architecture and then developed a novel training strategy named "pretraining-retraining strategy" (PRS) for TSEP by decomposing the whole training process into two successive stages: a pretraining stage is designed to train with the whole enhancer data for performing DEP, and a retraining strategy is then designed to train with tissue-specific enhancer samples based on the trained pretraining model for making TSEP. As a result, PRS is found to be valid for DEP with an AUC of 0.922 and a GM (geometric mean) of 0.696, when testing on a larger-scale FANTOM5 enhancer dataset *via* a five-fold cross-validation. Interestingly, based on the trained pretraining model, a new finding is that only additional twenty epochs are needed to complete the retraining process on testing 23 specific tissues or cell lines. For TSEP tasks, PRS achieved a mean GM of 0.806 which is significantly higher than 0.528 of gkm-SVM, an existing mainstream method for CRE predictions. Notably, PRS is further proven superior to other two state-of-the-art methods: DEEP and BiRen. In summary, PRS has employed useful ideas from the domain of transfer learning and is a reliable method for TSEPs.

Keywords: deep learning, pretraining, retraining, tissue-specific enhancers, prediction

Edited by:
Dominik Heider,
University of Marburg, Germany

Reviewed by:
Giuliano Armano,
University of Cagliari, Italy
Nagarajan Raju,
Vanderbilt University Medical Center, United States

***Correspondence:**
Xuehai Hu
huxuehai@mail.hzau.edu.cn

[†]*These authors have contributed equally to this work*

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

Received: *06 June 2019*
Accepted: *26 November 2019*
Published: *08 January 2020*

Citation:
Niu X, Yang K, Zhang G, Yang Z and Hu X (2020) A Pretraining-Retraining Strategy of Deep Learning Improves Cell-Specific Enhancer Predictions. Front. Genet. 10:1305. doi: 10.3389/fgene.2019.01305

INTRODUCTION

One of the core issues of today's biology is to decipher the code of cis-regulatory element (CRE) (Yáñez-Cuna et al., 2013). Enhancers are important distal CREs and play significant roles in gene transcriptional regulation (Bulger and Groudine, 2011). The regulation of gene expression by enhancers acts as a binding platform for recruiting transcriptional factors and cofactors to activate transcriptions of target genes (Shlyueva et al., 2014; Li et al., 2016).

Accurate identification of enhancer locations across the whole human genome is extremely important and is currently of great interest based on two facts: (1) ENCODE project indirectly identified >500,000 putative enhancers (Hoffman et al., 2012; Ernst and Kellis, 2012) and their total length might reach 12% of the human genome (Fishilevich et al., 2017), suggesting the enhancer element is a nonnegligible component of the human genome, and (2) genome-wide association studies (GWAS) in the past decade locked over 55% of the disease-associated SNPs in the non-coding DNA (Maurano et al., 2012). Some of them were reported to be exactly located within the enhancer regions, implying strong relationships between human diseases and the enhancer element. For example, a cancer-associated SNP of rs6983267 identified by human GWAS of intestinal tumors was reported to be contained within a Myc enhancer regulatory element (Sur et al., 2012). However, because of two hallmarks of enhancers, it is a challenging problem to distinguish them from other CREs: regulating manners of long-distance and bidirectionality. Typically, distal enhancers are located more than 10kb away from the target genes they regulate (Bulger and Groudine, 2011), and on the other hand, an enhancer can bidirectionally function both at the upstream and downstream of the target gene, which doubles the searching difficulty (Li et al., 2016).

In the past two decades, researchers have developed several distinct experimental strategies from different viewpoints for inferring the locations of active enhancers, such as transgenic mouse assay (Visel et al., 2007), using chromatin features from ENCODE data (Heintzman et al., 2009; Ernst and Kellis, 2012; Hoffman et al., 2012), massively parallel report assay (MPRA) employing barcode-containing transcripts (Melnikov et al., 2012; Kwasnieski et al., 2014; Shen et al., 2015), STARR-seq using self-transcribing transcripts (Arnold et al., 2013), and cap analysis of gene expression (CAGE), utilizing enhancer RNA (eRNA) (Andersson et al., 2014).

An alternative way for identifying enhancers is by computational methods, which try to learn intrinsic features from credible enhancer sequence samples and then build reliable prediction models for making evaluation and discovery. This mechanistic approach is feasible because DNA sequence is both sufficient and necessary for enhancer activity: (1) an enhancer sequence can still drive gene expressions when being removed from its endogenous context to upstream of a reporter gene (Kvon et al., 2012), suggesting its sufficiency; (2) a disruption of core motif within an enhancer sequence would substantially reduce enhancer activity (Kwasnieski et al., 2014), implying its necessity. As a matter of fact, a series of studies have already addressed this issue in the past decade (Lee et al., 2011; Kleftogiannis et al., 2014; Liu et al., 2016; Beer, 2017; Yang et al., 2017). A pioneer finding is that *k*-mer features of length 6 are predictive sequence features for discriminative enhancer prediction (DEP) when using ChIP-seq data of P300 (Lee et al., 2011). An advanced version of *k*-mer tool named gkm-SVM, which is one of the most popular method for regulatory sequence predictions (Ghandi et al., 2014), was recently employed for DEP (Beer, 2017). iEnhancer-2L proposed to use pseudo *k*-tuple nucleotide composition features for identifying enhancers and their strengths (Liu et al., 2016). Notably, BiRen (Yang et al., 2017) recently introduced more advanced tools including convolutional neural network (CNN) and bidirectional recurrent neural network (BRNN) for DEP. The above methods were all developed for DEP and they would give no answers about tissue-specific enhancer prediction (TSEP). At this point, DEEP (Kleftogiannis et al., 2014) integrated three resources of enhancer data, ENCODE, FANTOM5, and VISTA, and developed an ensemble model for DEP as well as for TSEP.

Although deep learning methods including BiRen were adopted for DEP, they have some problems that should be addressed for the task of TSEP. In the past 5 years, deep learning tools were successfully applied in some areas of biology from genomics and imaging to electronic medical records (Webb, 2018). Particularly, CNN has become a dominating method in various prediction problems, including predicting transcriptional factor binding sites (TFBS) (Alipanahi et al., 2015; Quang and Xie, 2016; Zeng et al., 2016) and predicting chromatin effects of DNA variants (Zhou and Troyanskaya, 2015; Kelley et al., 2016; Liu et al., 2018; Min et al., 2017). However, these successful experiences might not be directly transferred to TSEP by the following dilemma: on the one hand, a given enhancer for one specific tissue might not be activated in another tissue, so it is impossible to make multiple TSEPs only with one deep learning model; on the other hand, if we divide the whole enhancer dataset into multiple tissue-specific enhancer datasets and then build multiple prediction models, the sample size of each tissue is only several hundred or a few thousands, which is far less than the number of parameters (often hundreds of thousands) needed to be trained, suggesting that the built models might take high risks of falling into overfitting.

Here, we proposed a novel deep learning training strategy named pretraining-retraining strategy (PRS), which is especially appropriate for the task of TSEP. To address the problem of multiple TSEPs, we decomposed the training process into two successive stages: a pretraining stage and a retraining stage. The pretraining stage is designed for learning an appropriate network structure with optimal model hyperparameters of one model by using the whole enhancer data. Subsequently, a retraining stage is adopted only with a given tissue-specific enhancer dataset based on the trained pretraining model, suggesting a novel training pattern of one pretraining model together with multiple retraining models. To address the problem of overfitting, PRS allows all the hyperparameters to learn reasonable values when the pretraining stage is completed. And those reasonable values are good initial values of the retraining process, which enable the retraining model to converge very fast even with limited number of tissue-specific enhancer samples. PRS was tested on FANTOM5 enhancer data and was proven to be a powerful model for TESP.

MATERIALS AND METHODS

Datasets Preparation

In this work, the FANTOM5 enhancer data was used for performing prediction tasks. FANTOM consortium released a large-scale enhancer dataset that contains 65,423 enhancer activities (measured by TPM (tag per million mapped reads) of their expressions of eRNA) in 1,829 distinct tissues or cell lines in human (Andersson et al., 2014), which was recorded as a matrix $E_{65423\times 1829}$ with 65,423 rows and 1,829 columns (http://fantom.gsc.riken.jp/5/datafiles/latest/extra/Enhancers/).

In the pretraining stage, we used the following strategy for constructing a large-scale enhancer dataset: at first, we took a cut-off criterion of $TPM_{min} \geq 0.08$ (presents the minimal nonzero value of TPM across all tissues and cell lines of a given enhancer) to select most active enhancers, leaving only 5386 enhancers passing this criterion. Secondly, we excluded enhancers shorter than 100bp and fixed the enhancer sequence length at 1000bp with 4667 enhancers. Finally, we employed a redundancy reduction procedure CD-HIT (Huang et al., 2010) with a cutoff threshold of 0.8 and 4653 enhancers were remaining as the final positive samples. The length distribution of all 4,653 enhancer positive samples can be found in **Supplementary Figure 1**. We randomly selected 46,530 DNA sequences with length of 1,000 bp as negative samples from non-enhancer intergenic regions (obtained from the GRCh37 reference genome by excluding exon, intron and known enhancers) to meet a consensus of recent studies (Kleftogiannis et al., 2014; Liu et al., 2016; Yang et al., 2017).

In the retraining stage, 23 representative tissues or cell lines were chosen for showing cell-specific enhancer prediction performances. We also took a cut-off criterion of TPM 0.8.0.8 is the 75% quantile of the whole TPM distribution, implying that the condition of larger than 0.8 guarantees activity of enhancer) to select most active enhancers for each tissue or cell line. Ten times of the amount of each positive sample were selected as the corresponding negative samples.

Learning Subsequence Features With CNN

CNN is a modern combination of convolutional operator and classic neural network by introduction of some advanced techniques including rectified linear unit (ReLU), pooling and dropout. Convolutional operator is very powerful for detecting significant local features that are further denoised by ReLU and pooling. When performing prediction with neural network, CNN was proven efficient and successful in various image recognition tasks including handwriting recognition, face recognition (LeCun et al., 2015). Here, we adopted a similar framework with DeepBind (Alipanahi et al., 2015) to perform CNN model, which in turn includes three layers: a convolution layer (Conv), a activation layer (ReLU), a pooling layer (Pool), where the outputs of the final layer are regarded as selected features of the inputs (**Figure 1**).

Learning Dependencies With Bidirectional GRU

Recurrent neural network (RNN) is one kind of the advanced ANN model that has a "memory" which could capture the previous information, which is appropriate to analyze the sequential data (Schuster and Paliwal, 1997). Over the years, more advanced architectures of RNNs were developed to overcome shortcomings of the classic RNN model. Among them, bidirectional RNN (BRNN) is designed for those situations where output at time step is not only associated with the previous states, but also with future information. Because of the forward and inverse strand in enhancer sequences with bidirectional regulation function, BRNN model was proven to be very efficient to deal with regulatory sequence prediction problems (Quang and Xie, 2016).

However, BRNN still suffers a vanishing gradient problem that makes it hard to capture the long-term dependencies in the sequential data. For solving this problem, a gated recurrent units (GRU) was proposed by Bahdanau et al. (2014) by introducing some new concepts including update gate, reset gate and candidate "memory" layer. In this study, the bi-directional gated recurrent unit (Bi-GRU) was designed to connect with the last layer of CNN (the dropout layer) and six matrices WU will be learned by data (**Figure 1**).

Model Design and PRS

Previous studies on TFBS predictions reported that the convergent filter matrices of the CNN layer are exactly consistent with TF binding motif (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015; Kelley et al., 2016; Quang and Xie, 2016), suggesting CNN is efficient for learning local subsequence features. More importantly, a recent study (Quang and Xie, 2016; Yang et al., 2017) had used RNN layer to effectively address the dependence of the adjacent features in a sequence. Here, we adopted a similar deep learning model of BiRen (Yang et al., 2017) that added an RNN layer following the CNN layer (**Figure 1**). We expect to firstly learn local subsequence features (TF motifs) of an enhancer sequence with CNN, and then to learn how to combine these motifs (dependence of motifs) to form an enhancer sequence with RNN.

To solve the problem of TSEP, we proposed a novel PRS. Our idea is that we firstly use the whole FANTOM5 enhancer data (containing all tissues and cell lines) to determine an optimal network structure and all the model parameters, based on which we construct and record the pretraining model. Theoretically, such a pretraining model is only valid for discriminating enhancer from non-enhancers. For a given tissue, we will then take a retraining strategy by redoing training process with its tissue-specific enhancer data based on the pretraining model.

Pretraining With the Whole FANTOM5 Enhancer Data

We performed a pretraining process with the whole FANTOM5 enhancer data of Enhancer4653, which contains 4653 enhancer sequences and 46530 non-enhancer sequences. Firstly, we divided the whole dataset into three portions: 10/12 as training set E_train for training model), 1/12 as validation set E_va (for determining an optimal epoch) and 1/12 as testing set E_test (for evaluating model). To begin with a CNN structure, the initial values of model hyperparameters including filter number M, filter length m and pooling size p were set to be 64, 5 and 3 respectively. Subsequently, the output of CNN is turned as the input of RNN. Finally, a neural network with 32 neurons (a weight matrix of WM) was designed to be followed with the

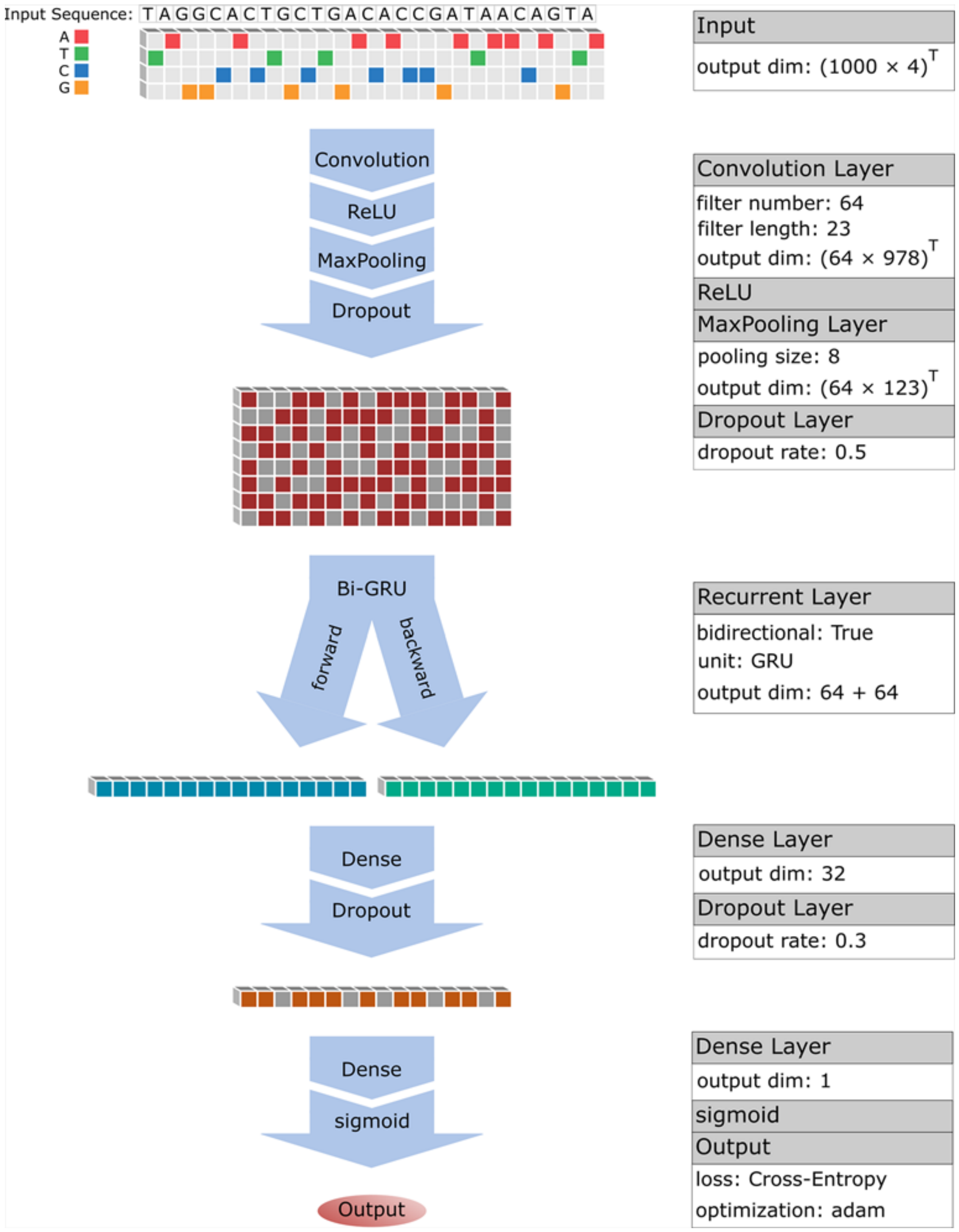


FIGURE 1 | Flow chart of hybrid deep learning architecture.

RNN layer and the output of the neural network NN will further be processed by a sigmoid function for mapping the predicted values into interval [0,1] (**Figure 1**):

$$\hat{y} = sigmoid(NN) = \frac{1}{1 + e^{-NN}}$$

which is considered as the final predicted value of each sample. This is the end of forward computation.

Here we took a rational strategy for preventing overfitting, which aims to find an optimal epoch minimizing objective $_{va}$ as:

$$\text{objective}_{\text{va}} = \text{crossentropy}_{\text{va}} + \lambda_1 \|M\|_1 + \lambda_2 \|WU\|_1 + \lambda_3 \|WM\|_1,$$

$$\text{crossentropy}_{\text{va}} = -\frac{1}{n} \sum_{y_i \in E\_va} [y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_t)],$$

where those $y_i \in E\_va$ belong to the validation set E_va and they never appeared in the training process. The strategy of minimizing objective $_{va}$ not objective $_{train}$, will effectively prevent overfitting and finally obtain the pretraining model (we call it the FANTOM model) with all the model parameters and hyperparameters determined. We finally evaluated effectiveness of the FANTOM model with predicting accuracy on all elements belonging to the testing set E_test.

Retraining With Specific Tissue (Cell Lines) Enhancer Data

Once we have the FANTOM model, we next implement a retraining strategy to predict tissue-specific enhancer based on it. A hypothesis of the retraining strategy is that a specific tissue enhancer dataset has similar pattern with the whole FANTOM5 enhancer dataset, which implies that the predicting model of tissue-specific enhancer might share the same network structure and all the model hyperparameters of FANTOM model. The only differences between them are the updated values of those parameters including filter matrices M and weight matrix WM.

Being different from regular training process that starts with random initial parameters, our novel retraining strategy will start with the convergent values of parameters obtained in the FANTOM model. The retraining strategy has some advantages when comparing with regular training: (1) it will rapidly reach optimal prediction accuracy with only dozens of epochs, implying it is time-saving; (2) the optimal prediction accuracy will be significantly better than that of a direct training (not begin with the pretraining model).

Evaluation of the Prediction Performance

Here, we used five indices for evaluating the prediction performance of models: sensitivity (Sens or recall), specificity (Spec), precision, accuracy (ACC), geometric mean (GM) value and Matthew's correlation coefficient (MCC):

$$\begin{cases} Sens = recall = \dfrac{TP}{TP + FN}, \\ Spec = \dfrac{TN}{TN + FP}, \\ precision = \dfrac{TP}{TP + FP}, \\ ACC = \dfrac{TP + TN}{TP + FP + TN + FN}, \\ GM = \sqrt{precision \cdot recall} \\ MCC = \dfrac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \end{cases}$$

To test the balance between true positive and false positive rates, another evaluating index is the Area Under the ROC Curve (AUC). Because of the imbalance between the positive and negative dataset, we applied GM as an important index to assess the performance.

RESULTS

Predicting Housekeeping Enhancers With the FANTOM Model

We first determined optimal values of three model hyperparameters including filter number M, filter length m, and pooling size p within the CNN layer with the training data E_train the validation set E_va and the testing set E_test When considering the optimal filter number, some previous works reported their choices. DeepBind (Alipanahi et al., 2015) used 16 filters for learning TF motifs; DeepSEA (Zhou and Troyanskaya, 2015) adopted three layers of CNN and took 320, 480, and 960 filters for learning chromatin features respectively; Basset (Kelley et al., 2016) employed three layers of CNN of 300, 200, 200 filters for chromatin accessibility prediction. Based on these existing experiences, we executed a parameter optimization strategy using grid search on the combinations of filter number (32, 64, 128, 256) and filter length [all odd numbers in (5, 25)] (**Figure 2**). Although researchers often used ACC or AUC value for evaluating prediction model (Liu et al., 2016; Beer, 2017; Yang et al., 2017), we here employed GM for evaluation because assessment with GM is more appropriate for extremely imbalance dataset (Kleftogiannis et al., 2014) (1:10 in this study). As a result, a maximal GM value of 0.821 was achieved at the combination of filter number of 64 and filter length of 23. Although, another high GM value of 0.815 was also achieved at the combination of filter number of 64 and filter length of 21, we finally determined the optimal filter number as 64.

After fixing filter number of 64, we then took a further grid search on the combinations of filter length with all odd numbers in [5,25] and pooling size of 3, 5, 8, 11, 14 and 17. We here employed GM value (**Figure 2**) together with AUC value (**Figure 2**) for a comprehensive evaluation. As a result, a maximal GM value of 0.815 was achieved at the combination of filter length of 15 and pooling size of 3 and the combination of filter length of 23 and pooling size of 8 achieved the second rank with GM value of 0.796. We noted that GM values exhibit a decreasing trend when pooling size is increasing (the column means of 3, 5, 8 and 11 are 0.750, 0.738, 0.732 and 0.731 respectively). In addition of the fact that larger pooling size would lose more information, we discarded the situations when pooling size is larger than 8 and only considered the situations with pooling size of 3, 5 and 8. We next focus on another evaluation indicator, AUC, for further searching. Interestingly, AUC values perpetuate an opposite trend when pooling size is increasing: the column means of 3, 5 and 8 are 0.912, 0.931 and 0.942 respectively, indicating that we should choose pooling size with 8. Although the maximal AUC value of 0.954 was achieved at filter length of 11 when fixing pooling size with 8. A comprehensive evaluation both using GM value and AUC value finally confirmed that the optimal filter length is 23 and the optimal pooling size is 8 because GM value of filter length of 11 was only 0.707 (significantly lower than 0.796 of filter length of 23).

In summary, we successively determined three important model hyperparameters as follows: filter number of 64, filter length of 23 and pooling size of 8. After confirming them, the FANTOM model was reevaluated *via* a 5-fold-cross-validation for a more objective assessment (**Table 1**). In the large-scale imbalanced enhancer dataset, the FANTOM model achieved a great AUC value of 0.922 (**Supplementary Figure 3**), an acceptable MCC value of 0.527, and an acceptable AUPRC value of 0.619 (**Supplementary Figure 2**) for this imbalanced dataset. In a word, the FANTOM model is a reliable prediction model on dataset of Enhancer4653, which consists of 4653

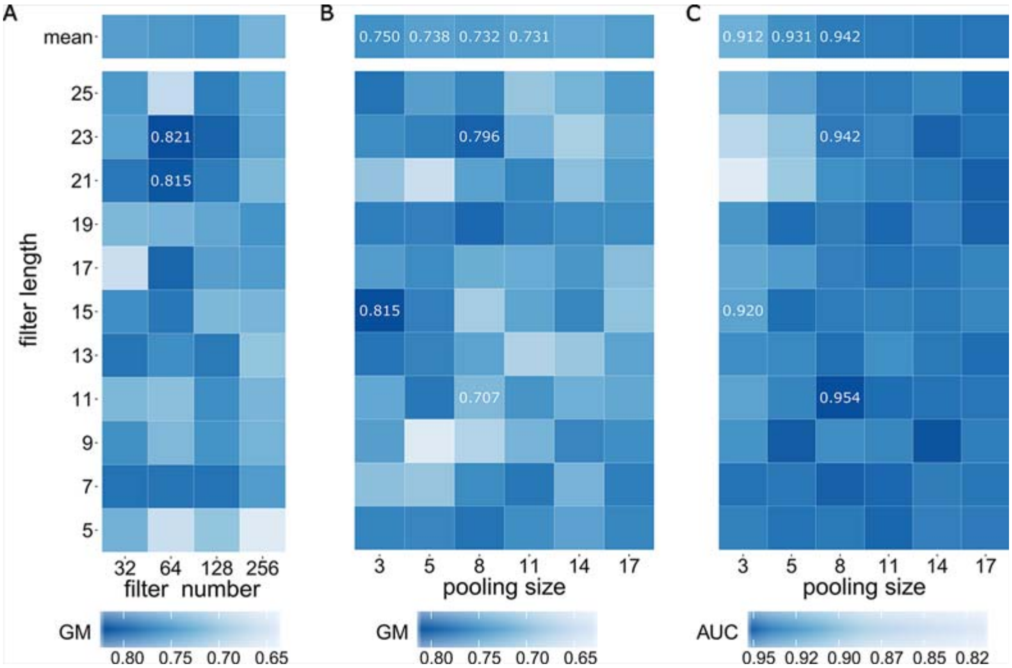


FIGURE 2 | Determining optimal model hyperparameters of filter number, filter length, and pooling size. **(A)** GM values of grid search on the combinations of filter number and filter length. **(B)** GM values of grid search on the combinations of filter length and pooling size. **(C)** AUC values of grid search on the combinations of filter length and pooling size.

TABLE 1 | Prediction performances of pretraining stage with large-scale FANTOM5 enhancer data via a five-fold-cross-validation.

Enhancer dataset	Sample size	ACC	AUC	SEN	SPE	MCC	GM
FANTOM5 enhancer data	4653 + 46530	0.929	0.922	0.499	0.972	0.527	0.696

housekeeping enhancers (Zabidi et al., 2015) and 46530 non-enhancers, implying it has potential to be a reliable model for housekeeping enhancer prediction.

Predicting Tissue-Specific Enhancers With a Retraining Strategy

Next we proposed to predict tissue-specific enhancers with a retraining strategy, which aims to build an updated model based on the pretraining model when adding a given tissue-specific enhancer dataset. Similar as before, a training epoch containing a cycle of forward computation and backpropagation was adopted to perform updating.

Next two specific problems which arise are: how many epochs is at least required and how many epochs is optimal? To answer these, based on the FANTOM model, we designed four groups of retraining with four distinct numbers of epochs: 10 epochs named FANTOM-ep10, 20 epochs named FANTOM-ep20, 50 epochs named FANTOM-ep50 and 100 epochs named FANTOM-ep100. Meanwhile, we performed another four groups of *ab initio* training (not based on the FANTOM model): 10 epochs named None-ep10, 20 epochs named None-ep20, 50 epochs named None-ep50, and 100 epochs named None-ep100. Training on 23 selected groups of tissue-specific enhancer datasets (Materials and methods), a total of eight boxplots representing their GM values is given in **Figure 3**, from which we found two interesting facts: (1) GM values of four pretraining-retraining models (starting with FANTOM-) are far greater than those of *ab initio* training models (starting with None-), suggesting the importance and necessity of PRS; (2) among four pretraining-retraining models, GM values of FANTOM-ep20 are relatively higher, though no significant difference was found between FANTOM-ep20 and FANTOM-ep10 (one-sided t-test, p-value = 0.31). However, significant difference was found between FANTOM-ep20 and FANTOM-ep50 (one-sided t-test, p-value = 0.036), suggesting FANTOM-ep50 (and FANTOM-ep100) model might fall into a problem of overfitting. In a word, retraining with 10 epochs is at least required and retraining with 20 epochs might be a good choice. It is not necessary to retrain with epochs larger than 50, which is not only time-consuming but also is easy to fall into overfitting.

After determining the optimal retraining epochs as 20, let us show the superiority of FANTOM-ep20 model by precisely comparing it to None-ep100 model (the best model within None models). From **Figure 3**, it is obvious that all the points

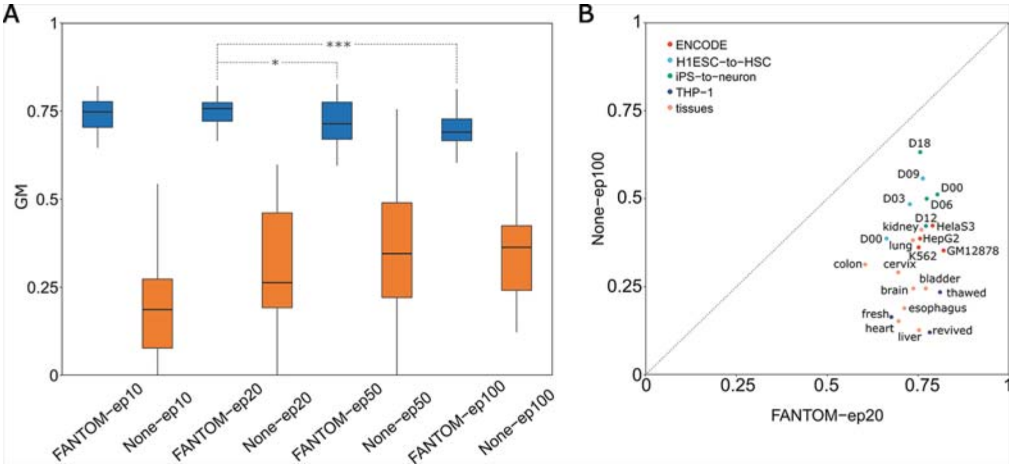


FIGURE 3 | Determining optimal pretraining-retraining model and comparison with classic model with no pretraining stage. **(A)** Comparison analysis determines FANTOM-ep20 model to be the optimal pretraining-retraining model. **(B)** Comparison of GM values between FANTOM-ep20 models and None-ep100 models on 23 different tissues or cell lines.

located below the line y = x, suggesting that FANTOM-ep20 model is superior to None-ep100 model at each tissue. Furthermore, 23 FANTOM-ep20 models take their GM values between 0.606 and 0.822 (with a mean of 0.746), whereas 23 GM values of None-ep100 models distribute from 0.122 to 0.634 with a mean of 0.345. A statistical t-test showed that the former is extremely greater than the latter (p-value = 1.44e-12), suggesting the difference between these two is huge. Without a pretraining stage, TSEPs using deep learning model are bad due to very low Sens values. It is widely accepted that positive sample predictions are hard when training on an extremely imbalanced dataset. The mean of 23 Sens values of None-ep100 models has a very low mean of 0.141, suggesting only 14% of positive samples were accurately predicted. By contrast, when taking PRS, 23 Sens values of FANTOM-ep20 models has a mean of 0.580, implying FANTOM-ep20 model accurately identified about 60% of positive samples. In summary, the prediction on tissue-specific enhancer will be unreliable if a pretraining stage was absent, whereas it will be much better and more acceptable by adding a pretraining stage.

We investigated the resource consumption of prediction of enhancer samples by running our script on a test computer with Ubuntu 18.04 on processors of Intel(R) Core(TM) i7-7700 CPU @ 3.60GHz, GPU of GeForce GTX 1080 Ti and 24 GB RAM. When running on 4616 testing sequences with a length of 1000 bp, a total of 1.28s was needed for such predictions, implying that the average computation time of each DNA sequence was about 2.77 × 10^{-4} second.

Comparisons With Other Existing Methods

To further show the superiority of our method, comprehensive comparisons with three state-of-the-art methods, gkm-SVM (Ghandi et al., 2014; Ghandi et al., 2016; Beer, 2017), DEEP (Kleftogiannis et al., 2014), and BiRen (Yang et al., 2017), were performed. There are two distinct strategies for such a comparison: one is to run other tools on our dataset; the other is to run our method on existing dataset that other method used.

We first adopted the former comparison strategy for gkm-SVM. Gkm-SVM is one of the most popular methods for regulatory sequence prediction (Ghandi et al., 2014) and has gradually become a dominating method in this area (Ghandi et al., 2016). We downloaded its R package from the website https://cran.r-project.org/web/packages/gkmSVM/index.html and then run it on our 23 tissue-specific enhancer datasets with its default parameters of L=10, K=6. A direct comparison with our best model of FANTOM-ep20 can be found in **Figure 4**, which shows the point-to-point comparisons of GM values on 23 tissues or cell lines. It is obvious that all the blue points representing those GM values (a mean of 0.806) achieved by FANTOM-ep20 models are above the orange points (a mean of 0.528) by gkm-SVM, suggesting our FANTOM-ep20 model is superior to gkm-SVM on GM values. This is further confirmed by the box-plots of these two and a t-test between them with a p-value of 1.725e-15 in **Figure 4**, though AUC values of gkm-SVM (a mean of 0.969) are slightly greater than those of our FANTOM-ep20 model (a mean of 0.957).

We next applied the later comparison strategy for DEEP and BiRen. DEEP (Kleftogiannis et al., 2014) trained many individual models for 36 different tissues from FANTOM enhancer data but it only provided the detailed prediction results on three specific tissues: heart, liver, and brain, which were chosen for comparisons. Using the latest version of FANTOM5 enhancer data, we set the cutoff thresholds with $TPM > 1$;$TPM > 4$;$TMP > 1$ to select three groups of tissue-specific enhancers whose numbers are closest to those numbers provided by DEEP (**Table 2**). To be consistent with

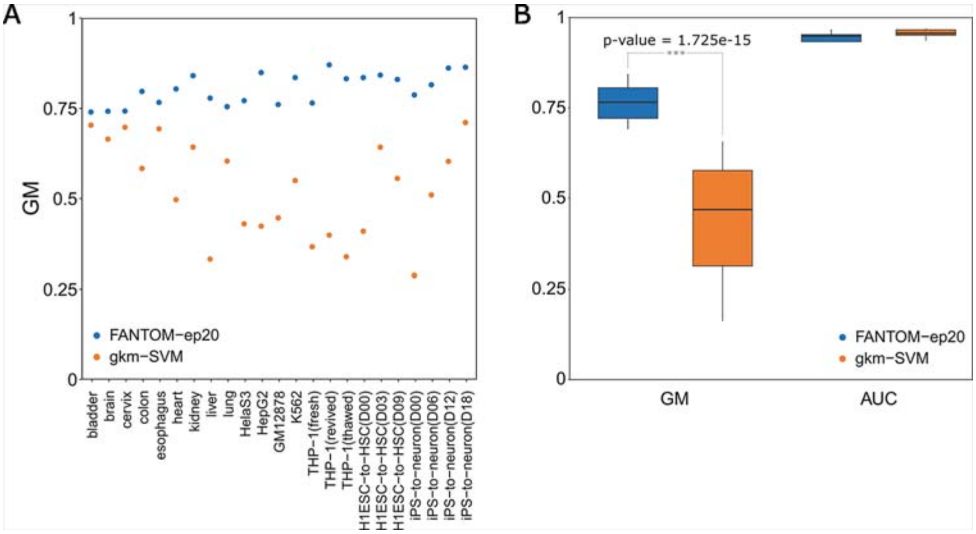


FIGURE 4 | Comparisons between our FANTOM-ep20 model and gkm-SVM tool on 23 different tissues or cell lines. **(A)** One-to-one direct comparison of GM value on each tissue or cell line. **(B)** Distribution comparisons of GM values and AUC values with box plots.

TABLE 2 | Comprehensive comparisons of FANTOM-ep20 model with DEEP and BiRen.

Comparison targets	Data source	Sample size	Method	ACC	AUC	Sens	Spec	MCC	GM
DEEP	Heart	295 + 2950	DEEP	0.822	NA	0.802	0.824	NA	0.812
		239 + 2390	FANTOM-ep20[a]	0.946	0.963	0.664	0.976	0.669	0.805
	Liver	84 + 840	DEEP	0.745	NA	0.740	0.755	NA	0.741
		75 + 750	FANTOM-ep20	0.982	0.990	0.905	0.989	0.891	0.946
	Brain	639 + 6390	DEEP	0.853	NA	0.832	0.855	NA	0.843
		619 + 6190	FANTOM-ep20	0.906	0.915	0.630	0.933	0.501	0.766
BiRen	VISTA	1747 + 17470	BiRen	NA	0.957	NA	NA	NA	NA
	VISTA	1848 + 18480	FANTOM-ep20	0.946	0.958	0.650	0.975	0.655	0.796

[a]*our best pretraining-retraining model by pretraining with large scale FANTOM enhancer data and retraining with 20 epochs; 'NA' represents 'not provided by original publications'.*

DEEP, the negative samples were chosen from random intergenic regions with 10 times number of positive samples of each tissue. After performing the optimal testing strategy (40% for training and 60% for testing) of DEEP, ACC values of FANTOM-ep20 models of heart, liver, and brain were 0.946, 0.982, and 0.906, respectively, which are greater than 0.822, 0.745, and 0.853 of DEEP (**Table 2**), suggesting our model has higher prediction accuracy compared with DEEP. In their article, DEEP claimed that great superiority of their model is prediction balance on imbalance dataset, which is measured by GM value. While comparing GM values, our FANTOM-ep20 models of heart, liver and brain achieved 0.805, 0.946 and 0.766, which are comparable with 0.812, 0.741and 0.843 of DEEP respectively (**Table 2**).

For comparison with BiRen, we applied our FANTOM-ep20 model on VISTA enhancer data that BiRen used. We visited the updated version of VISTA enhancer browser https://enhancer.lbl.gov/ and downloaded 959 positive human enhancer sequences and 889 negative ones, summing 1,848 human enhancer sequences. To be consistent with BiRen, a non-enhancer dataset containing 10 times the number of random genomic fragments (18,480 non-enhancer sequences) were selected from the whole genome (the GRCh37 reference genome) by excluding exon, intron and known enhancers. As a result, our FANTOM-ep20 model achieved an average AUC value of 0.958, which is slightly larger than 0.957 of BiRen by evaluating *via* a five-fold cross validation test. Moreover, additional evaluation indices including ACC, GM, Sens, and Spec of our FANTOM-ep20 model are also provided in **Table 2**, from which we found that a GM value of 0.796 was achieved, suggesting our FANTOM-ep20 model remains robust prediction performance on VISTA enhancer data.

DISCUSSION

Enhancers are important CREs and play significant roles in gene transcriptional regulation. Majority of enhancers have strong cell or tissue specificity, which highlights the importance of TSEP. In this paper, we developed a novel training strategy of deep learning named with PRS, which was proven to be a reliable prediction model for TSEP. Finally, we conclude that PRS brings some new contributions or findings into the area of TSEP:

New contribution to training strategy: a specific cell or tissue type has only hundreds or a few thousands of specific enhancer samples, which might make existing deep learning methods to fall into overfitting problem. PRS employs a large scale FANTOM enhancers data to construct a pretraining model with optimal model hyperparameters, and then uses each small sample dataset of tissue-specific enhancers to retrain, based on the trained pretraining model. Testing results on 23 different cell or tissue types demonstrate that PRS is superior to classic training strategy without pretraining, which enable us to conclude that PRS is a reliable method for TSEP.

New findings on optimal retraining epochs: we found that 20 additional epochs are optimal when retraining a new source of tissue-specific enhancer samples based on the trained pretraining model. Either too few or too many additional epochs are not the good choices, because too few epochs like FANTOM-ep10 has not fully learned features of the new source data, whereas too many epochs like FANTOM-ep50 might has a big problem of overfitting.

New contribution to transfer learning: when comparing the best model of PRS named with FANTOM-ep20 with existing tool names with BiRen, we noted an interesting fact: FANTOM-ep20 achieved a greater AUC value with a different enhancer data source of VISTA enhancer data in the retraining stage. VISTA enhancer data was generated with a totally different biological assay and has distinct distribution or source domain with FANTOM enhancer data. Our FANTOM-ep20 model took pretraining with FANTOM enhancer data and then performed retaining with VISTA enhancer data. This shows that our PRS model has good performance of transfer learning, which implies that PRS might provide helpful ideas for transfer learning studies.

Although notable successes were achieved in the current study, some drawbacks or limitations still need further investigations in the future works. For example, this method is not appropriate for enhancers with sequences shorter than 100bp and greater than 1000bp. In addition, there are totally three main sources of enhancer data: FANTOM, Vista, and ENCODE. In the current study, we only trained on FANTOM enhancer data and tested on Vista enhancer data. The comprehensive combinations of training and testing between three sources are the future directions of DEP and TSEP.

DATA AVAILABILITY STATEMENT

We developed our scripts and pipeline with the “Keras” deep learning framework in Python. We deposited our data, codes, and trained models at the following github website: https://github.com/yangg-kun/enhancer_retraining.

AUTHOR CONTRIBUTIONS

XH and XN designed the research. KY, GZ and ZQ performed the research and analyzed the data. XH and XN wrote the manuscript. All authors revised the manuscript.

FUNDING

XN was partially supported by the Fundamental Research Funds for the Central Universities HZAU (Grant No. 2662017JC048). XH was partially supported by the National Natural Science Foundation of China (NSFC) (Grant No. 11671003).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.01305/full#supplementary-material

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831. doi: 10.1038/nbt.3300

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., et al. (2014). An atlas of active enhancers across human cell types and tissues. *Nature* 507, 455–461. doi: 10.1038/nature12787

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M, Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* 339, 1074–1077. doi: 10.1126/science.1232542

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by Jointly Learning to align and translate. - arXiv preprint arXiv:1409.0473. *Comput. Sci.*

Beer, M. A. (2017). Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.* 38, 1251–1258. doi: 10.1002/humu.23185

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell* 144, 327–339. doi: 10.1016/j.cell.2011.01.024

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* 9, 215. doi: 10.1038/nmeth.1906

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., et al. (2017). GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* 2017. doi: 10.1093/database/bax028

Ghandi, M., Lee, D., Mohammad-Noori, M., and Beer, M. A. (2014). Enhanced regulatory sequence prediction using gapped k-mer features. *PloS Comput. Biol.* 10, e1003711. doi: 10.1371/journal.pcbi.1003711

Ghandi, M., MohammadNoori, M., Ghareghani, N., Lee, D., Garraway, L., and Beer, M. A. (2016). gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* 32, 2205–2207. doi: 10.1093/bioinformatics/btw203

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459, 108–112. doi: 10.1038/nature07829

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* 9, 473–476. doi: 10.1038/nmeth.1937

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26, 680–682. doi: 10.1093/bioinformatics/btq003

Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26, 990. doi: 10.1101/gr.200535.115

Kleftogiannis, D., Kalnis, P., and Bajic, V. B. (2014). DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res.* 43, e6–e6. doi: 10.1093/nar/gku1058

Kvon, E. Z., Stampfel, G., Yáñez-Cuna, J. O., Dickson, B. J., and Stark, A. (2012). HOT regions function as patterned developmentals enhancers and have a distinct cis-regulatory signature. *Genes Dev.* 26, 908–913. doi: 10.1101/gad.188052.112

Kwasnieski, J. C., Fiore, C., Chaudhari, H. G., and Cohen, B. A. (2014). High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* 24, 1595–1602. doi: 10.1101/gr.173518.114

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436. doi: 10.1038/nature14539

Lee, D., Karchin, R., and Beer, M. A. (2011). Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* 21, 2167–2180. doi: 10.1101/gr. 121905.111

Li, W., Notani, D., and Rosenfeld, M. G. (2016). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.* 17, 207. doi: 10.1038/nrg.2016.4

Liu, B., Fang, L., Long, R., Lan, X., and Chou, K. C. (2016). iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* 32, 362–369. doi: 10.1093/bioinformatics/btv604

Liu, Q., Xia, F., Yin, Q., and Jiang, R. (2018). Chromatin accessibility prediction *via a* hybrid deep convolutional neural network. *Bioinformatics*. 34 (5), 732–738. doi: 10.1093/bioinformatics/btx679

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337 (6099), 1190–1195. doi: 10.1126/science.1222794

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., et al. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* 30, 271–277. doi: 10.1038/nbt.2137

Min, X., Zeng, W., Chen, N., Chen, T., and Jiang, R. (2017). Chromatin accessibility prediction *via* convolutional long short-term memory networks with k-mer embedding. *Bioinformatics* 33, i92–i101. doi: 10.1093/bioinformatics/btx234

Quang, D., and Xie, X. (2016). DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* 44, e107–e107. doi: 10.1093/nar/gkw226

Schuster, M., and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process* 45, 2673–2681. doi: 10.1109/78.650093

Shen, S. Q., Myers, C. A., Hughes, A. E., Byrne, L. C., Flannery, J. G., and Corbo, J. C. (2015). Massively parallel cis-regulatory analysis in the mammalian central nervous system. *Genome Res.* 26, 238–255. doi: 10.1101/gr.193789.115

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* 15, 272.

Sur, I. K., Hallikas, O., Vähärautio, A., Yan, J., Turunen, M., Enge, M., et al. (2012). Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* 338, 1360–1363. doi: 10.1126/science.1228606

Visel, A., Minovitsky, S., Dubchak, I., and Penacchio, L. A. (2007). VISTA enhancer browser—a database of tissue-specific human enhancers. *Nucleic Acids Res.* 35, D88–D92. doi: 10. 1093/nar/gkl822

Webb, S. (2018). Deep learning for biology. *Nature* 554, 555–557. doi: 10.1038/d41586-018-02174-z

Yáñez-Cuna, J. O., Kvon, E. Z., and Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. *Trends Genet.* 29, 11–22. doi: 10.1016/j.tig.2012.09.007

Yang, B., Liu, F., Ren, C., Ouyang, Z., Xie, Z., Bo, X., et al. (2017). BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics* 33 (13), 1930–1936. doi: 10.1093/bioinformatics/btx105

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., et al. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* 518, 556–559. doi: 10.1038/nature13994

Zeng, H., Edwards, M. D., Liu, G., and Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* 32, i121. doi: 10.1093/bioinformatics/btw255

Zhou, J., and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12, 931. doi: 10.1038/nmeth.3547

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: 04 February 2020
doi: 10.3389/fgene.2020.00018

FCTP-WSRC: Protein–Protein Interactions Prediction *via* Weighted Sparse Representation Based Classification

Meng Kong[1], Yusen Zhang[1], Da Xu[1], Wei Chen[1] and Matthias Dehmer[2,3,4]*

[1] School of Mathematics and Statistics, Shandong University at Weihai, Weihai, China, [2] University of Applied Sciences Upper Austria, School of Management, Steyr, Austria, [3] College of Artificial Intellegience, Nankai University, Tianjin, China, [4] Department of Biomedical Computer Science and Mechantronics, UMIT Hall, Tyrol, Austria

OPEN ACCESS

Edited by:
Dominik Heider,
University of Marburg,
Germany

Reviewed by:
Sitanshu Sekhar Sahu,
Birla Institute of Technology,
India
Markus List,
Technical University of Munich,
Germany

****Correspondence:***
Yusen Zhang
zhangys@sdu.edu.cn

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

Received: *09 July 2019*
Accepted: *07 January 2020*
Published: *04 February 2020*

Citation:
Kong M, Zhang Y, Xu D, Chen W and Dehmer M (2020) FCTP-WSRC: Protein–Protein Interactions Prediction via Weighted Sparse Representation Based Classification. Front. Genet. 11:18. doi: 10.3389/fgene.2020.00018

The task of predicting protein–protein interactions (PPIs) has been essential in the context of understanding biological processes. This paper proposes a novel computational model namely FCTP-WSRC to predict PPIs effectively. Initially, combinations of the F-vector, composition (C) and transition (T) are used to map each protein sequence onto numeric feature vectors. Afterwards, an effective feature extraction method PCA (principal component analysis) is employed to reconstruct the most discriminative feature subspaces, which is subsequently used as input in weighted sparse representation based classification (WSRC) for prediction. The FCTP-WSRC model achieves accuracies of 96.67%, 99.82%, and 98.09% for *H. pylori*, *Human* and *Yeast* datasets respectively. Furthermore, the FCTP-WSRC model performs well when predicting three significant PPIs networks: the single-core network (CD9), the multiple-core network (Ras-Raf-Mek-Erk-Elk-Srf pathway), and the cross-connection network (Wnt-related Network). Consequently, the promising results show that the proposed method can be a powerful tool for PPIs prediction with excellent performance and less time.

Keywords: protein–protein interactions, principal component analysis, sparse representation, prediction, crossover network

INTRODUCTION

Investigating protein–protein interactions (PPIs) relate to examine the correlation between proteins involved in various aspects of life processes such as signal transduction, gene expression regulation, energy metabolism, and cell cycle regulation. The traditional way of studying individual proteins has failed to meet the requirements of the post-genome era because the performance of proteins is diverse and dynamic when performing physiological functions. Therefore, proteins should be studied at the global, network, and dynamic levels. Only by studying the sum of all proteins can we support the understanding of life's behavioral processes, disease prevention, and development of new drugs (Long et al., 2019). In recent years, some researchers predict PPIs by biological methods such as yeast two-hybrid screening (Ito et al., 2001; Pazos and Valencia, 2002) and affinity purification (Gavin et al., 2002). However, the results obtained by wet-lab experiments usually contain a large amount of false positive and false negative data, and these methods are time consuming and costly. These limitations motivate the development of effective machine learning methods to predict large-scale PPIs.

Up to now, D.S. Huang et al. predicts PPIs utilizing different information sources such as tertiary structure of proteins, phylogenetic profiles, and protein domains (De-Shuang and Chun-Hou, 2006; De-Shuang and Ji-Xiang, 2008). However, these computational methods require prior knowledge of the target protein (An et al., 2016). In recent years, protein sequence-based methods (Yu et al., 2017) are becoming the most widely applied technique for predicting PPIs due to the availability of protein sequence data. Liu et al. (2012) designs a sequence analysis method to represent protein sequences based on hypergeometric series using the q-Wiener index (Xu et al., 2017). X. Li et al. employs a global encoding approach (GE) to describe global information of amino sequence (Li et al., 2009).

Since the effectiveness of machine learning algorithms has been continuously verified in recent years, the use of machine learning methods for predicting PPIs has become a new research area. Yanzhi et al. proposes a support vector machine (SVM) prediction method based on auto covariance (AC) (Wold et al., 1993; Yanzhi et al., 2008) Davies et al. designs a model based on k-nearest neighbor (KNN) with local descriptor (LD) (Juan et al., 2007; Davies et al., 2008; Tong and Tammi, 2008; Lei et al., 2010). Juwen et al. using SVM with conjoint triad method predicting PPIs (Juwen et al., 2007). In addition, algorithms that use machine learning include: random forest (RF) with multi scale continuous and discontinuous local descriptor (MCD) (You et al., 2014), deep neural networks (DNNs) with pseudo amino acid physicochemical property descriptors(APAAC) (Kuo-Chen, 2005; Du et al., 2017) and so forth. These methods to perform PPIs prediction use solely amino acid sequence data. In addition, different representation methods can extract distinct characteristic information of protein sequences, and it is known that the feature information extracted by these representation methods can be complementary. Thus, for PPIs prediction, we advocate combining multiple descriptors, which can capture more information than a single descriptor (Deng et al., 2015). EnsDNN is a multi-descriptor combining method based on deep neural network (Xenarios et al., 2002). These descriptors such as auto-covariance descriptor (AC), local descriptor (LD) and multi-scale continuous and discontinuous local descriptor (MCD). It achieved a high accuracy of 95.25% on the *Saccharomyces cerevisiae* dataset. Despite this, there is still room to improve the accuracy and efficiency.

Previous works have pointed out that using feature selection or feature extraction before conduction the classification tasks can improve the classification accuracy (Zhang et al., 2012). The software EFS (Ensemble Feature Selection) makes use of multiple feature selection methods and combines their normalized outputs to a quantitative ensemble importance. Currently, eight different feature selection methods have been integrated in EFS, which can be used separately or combined in an ensemble (Neumann et al., 2017). What's more, several evolutionary based methods are proposed for dimensionality reduction (Chuang et al., 2016). A multi-objective differential evolution method (called MODEMDR) was proposed to merge the various contingency table measures based on MDR to detect significant gene-gene interactions (Yang et al., 2017). In this paper, principal component analysis (PCA) is utilized to do the feature extraction which projects the original feature space into a new space. The effectiveness of the proposed FCTP-WSRC is examined in terms of classification accuracy on the PPI dataset.

The main contribution of this paper is to develop a new computational tool called FCTP-WSRC to predict PPIs efficiently. More precisely: (1) Combinations of the F-vector, composition (C) and transition (T) are used to map each protein sequence on numeric feature vectors. (2) An effective feature extraction method PCA (principal component analysis) is employed to reconstruct the most discriminative feature subspaces, which is subsequently used as input in weighted sparse representation based classification (WSRC) for prediction. We obtain a unique 60-dimensional feature vector of each protein pair. (3) The FCTP-WSRC model can predict newly discovered protein-protein interactions with unknown biological functions using only protein sequence information.

METHODOLOGY

Reduced Sequence and F-Vector

In this paper, a computational model based on multivariate mutual information is designed to represent the protein sequence and obtain the feature vector. The model describes the protein sequence as a fixed length feature vector containing key information, which can be used as an effective input for machine learning algorithm. Therefore, the design of the F vector, the composition and transition (CT) descriptors is combined to map each protein sequence to a digital feature vector. F-vector of protein sequence is constructed in the following manner.

First, we generate reduced amino acid sequences according to their physicochemical properties such as hydrophobicity and polarity. When studying Shannon entropy of residue properties, instead of treating the amino acids as distinct symbols in the entropy calculation, six groups have proposed partitioning the amino acids into stereo chemically defined sets, and then computing the entropy of the column with respect to these sets. According to Capra JA et al. (Capra and Singh, 2007), we classify residues into six different classes. The six classes of amino acids are: aliphatic (AVLIMC), aromatic (FWYH), polar (STNQ), positive (KR), negative (DE), and special (reflecting their special conformational properties) (GP) (Mirny and Shakhnovich, 1999), as depicted in **Table 1**.

TABLE 1 | Amino acid classification.

Descriptor	Property	Classification
A1	Aliphatic amino acid	A,V,L,I,M,C
A2	Aromatic amino acid	F,W,Y,H
A3	Polar amino acid	S,T,N,Q
A4	Positive amino acid	K,R
A5	Negative amino acid	D,E
A6	Special conformations	G,P

The plane rectangular coordinate system has four quadrants. Dividing 20 amino acids into four groups can use the formula (1) to map the protein sequence to the unit circle. However, 20 amino acids are divided into six classes. Thus, we recombine six types of amino acids. Three classes of amino acids are selected from the six classes of amino acids as one group and the remaining three classes are unchanged. In this way, we can get four groups of amino acids, and there are a total of 20 combination patterns. It is found through experiments that the 20 patterns will cause too many features and affect the operation efficiency. Selecting the top 10 combination patterns got good results.

Then, we use a binary space (V, F) to describe amino acid sequences. Here, V is the feature space of the sequence information, and each amino acid combined pattern v_i represents a sort of quad type; F is the feature vector corresponding to V. The size of V should be 10; thus, I = 1,2, …, 10. We describe ten amino acid combined patterns by the letters B, J, O and U in **Table 2**. The detailed definition and description for (V, F) are illustrated by the Equations (1)-(4). Clearly, each protein has a corresponding F vector.

$$S_q(v_i) \to \begin{cases} \left(\cos\left(\frac{\pi}{2}\frac{B_j}{B_n+1}\right), \sin\left(\frac{\pi}{2}\frac{B_j}{B_n+1}\right)\right) & if\ S_q = B \\ \left(\cos\left(\frac{\pi}{2}+\frac{\pi}{2}\frac{J_j}{J_n+1}\right), \sin\left(\frac{\pi}{2}+\frac{\pi}{2}\frac{J_j}{J_n+1}\right)\right) & if\ S_q = J \\ \left(\cos\left(\pi+\frac{\pi}{2}\frac{O_j}{O_n+1}\right), \sin\left(\pi+\frac{\pi}{2}\frac{O_j}{O_n+1}\right)\right) & if\ S_q = O \\ \left(\cos\left(\frac{3\pi}{2}+\frac{\pi}{2}\frac{U_j}{U_n+1}\right), \sin\left(\frac{3\pi}{2}+\frac{\pi}{2}\frac{U_j}{U_n+1}\right)\right) & if\ S_q = U \end{cases} \quad (1)$$

We suppose each reduced sequence $S=S_1S_2S_3\cdots S_n$, $S_q\in\{B,J,O,U\}$, and q = 1, 2,…, n. B_n is the number of B in the sequence S by using the pattern v_i. B_j is the number of B in the first j characters when $S_j = B$. According to Equation (1), we introduce Equation (2):

$$S(v_i) \to \begin{cases} M_x = \frac{1}{n}\sum_{q=1}^{n} x_q \\ M_y = \frac{1}{n}\sum_{q=1}^{n} y_q \\ V_x = \sqrt{\frac{1}{n-1}\sum_{q=1}^{n}(x_q - M_x)^2} \\ V_y = \sqrt{\frac{1}{n-1}\sum_{q=1}^{n}(y_q - M_y)^2} \end{cases} \quad (2)$$

Here x_q and y_q (q = 1,2,⋯, n) are derived from Equation (1).

For example, sequence $METKDGIRWA$ can be expressed as $BOBJOUBJBB$ based on v_1, so it is mapped to the unit circle as shown in **Figure 1**. The reduced sequence corresponds to a one-to-one curve in the unit circle. So, the invariant of the curve can be used as the characteristic value of the sequence. Finally, the F-vector can be expressed by:

$$F = (F(v_i), F(v_2), \cdots, F(v_{10})) \quad (3)$$

The vector $F(v_i)$ is as follows:

$$F(v_i) = (M_x, M_y, V_x, V_y), i = 1, 2, \cdots, 10 \quad (4)$$

TABLE 2 | Ten amino acid combined patterns described by the letters B, J, O, and U.

	B	J	O	U
v_1	$\{A_1, A_2, A_3\}$	A_4	A_5	A_6
v_2	$\{A_1, A_2, A_4\}$	A_3	A_5	A_6
v_3	$\{A_1, A_2, A_5\}$	A_3	A_4	A_6
v_4	$\{A_1, A_2, A_6\}$	A_3	A_4	A_5
v_5	$\{A_1, A_3, A_4\}$	A_2	A_5	A_6
v_6	$\{A_1, A_3, A_5\}$	A_2	A_4	A_6
v_7	$\{A_1, A_3, A_6\}$	A_2	A_4	A_5
v_8	$\{A_1, A_4, A_5\}$	A_2	A_3	A_6
v_9	$\{A_1, A_4, A_6\}$	A_2	A_3	A_5
v_{10}	$\{A_1, A_5, A_6\}$	A_2	A_3	A_4

Thus, a 40-dimensional vector is obtained to characterize each amino acid sequence.

The Composition and Transition of Protein Sequence (CT)

In this section, we put forward a new description approach using binary coding sequences. First of all, the amino acid sequence is mapped to a sparse matrix. Then the composition (C) and transition (T) of characteristic sequence are extracted from the obtained sparse matrix. The protein sequence is scanned from left to right by the step of one amino acid at a time. Suppose a protein sequence with n amino acid residues is given: $S=S_1S_2S_3\cdots S_n$;D = $\{A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y,V\}$. Now we derive the matrix A of this sequence:

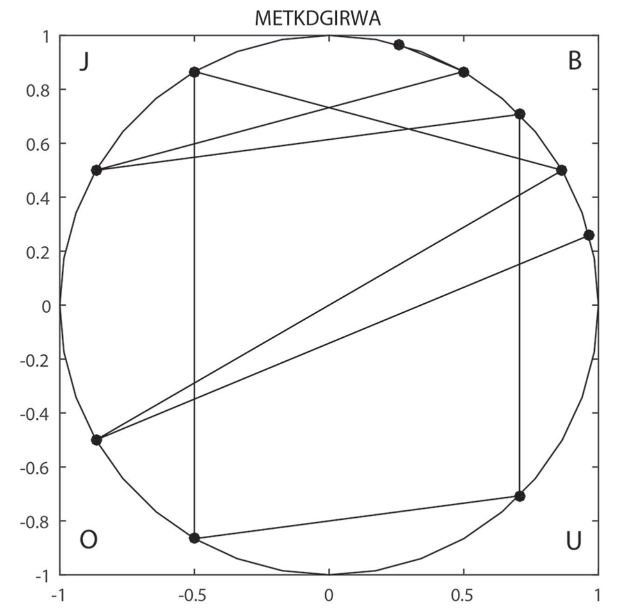


FIGURE 1 | 2-D Unit circle mapping representation of "METKDGIRWA" under pattern.

$$A = \left(\begin{array}{ccccccc} & S_1 & S_2 & S_3 & \cdots & S_{n-1} & S_n \\ A & a_{11} & a_{12} & a_{13} & \cdots & a_{1,n-1} & a_{1,n} \\ R & a_{21} & a_{22} & a_{23} & \cdots & a_{2,n-1} & a_{2,n} \\ N & a_{31} & a_{32} & a_{33} & \cdots & a_{3,n-1} & a_{3,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ Y & a_{19,1} & a_{19,2} & a_{19,3} & \cdots & a_{19,n-1} & a_{19,n} \\ V & a_{20,1} & a_{20,2} & a_{20,3} & \cdots & a_{20,n-1} & a_{20,n} \end{array}\right)_{20*n}$$

$$a_{i,j} = \begin{cases} 1, \ if \ D(i) = S(j) \\ 0, \ others \end{cases} \tag{5}$$

where $D(i)$ is the i-th kind of amino acid in the arranged letter sequence D.

For each row vector of matrix A with length n, we divide the sequence into L sub-vectors. For each characteristic sub-vector, the composition (C) consists of four parts: frequency of "0", frequency of "1", frequency of "11" and frequency of "111", respectively. The descriptor (T) is the frequency of "0" followed by "1" or "1" followed by "0". An example regarding the composition (C) of the sub-vector with respect to amino acid A is shown in the **Figure 2**. The subsequence "AATWTFAAACATAPDAADAG" with respect to amino acid A is replaced by "11000011101010011010". We see that there exists ten "1", ten "0", four "11", and one "111". The composition for these four parts is 10×100%/(10 + 10) = 50%, 10×100%/(10 + 10) = 50%, 4×100%/19 = 21.05%, and 1 × 100%/18 = 5.56%. The transition for "1-0" and "0-1" is (6 + 5)×100%/19 = 57.89%. Thus, a protein sequence is transformed into a 4×20×5 = 400 dimensional vector with L = 4 and 20 amino acids.

Reconstructing Feature Vectors

So far, we combine the descriptor F-vector (40 dimension) and descriptor CT (400 dimension) for a protein sequence into a 440-dimensional vector. However, if this vector is used as input of the classifier directly, the efficiency is likely to be low. Therefore, in this section we discuss how to reconstruct new feature vectors using principal component analysis (PCA). Principal component analysis (PCA) is a widely used dimensional compression technique. The main idea of PCA is to sequentially find a set of mutually orthogonal coordinate axes from the original space, which is closely related to the data itself. When 30 dimensional features are selected, the contribution rate of features can reach more than 90%. It can not only ensure the accuracy, but also improve the calculation efficiency. Therefore, we use PCA to reduce 440 dimension vector to 30 dimension. We connect the feature vectors of two proteins (V_A and V_B) to describe their interaction information (V_{AB}):

$$\{V_{AB}\} = \{V_A\} \oplus \{V_B\} \tag{6}$$

Thus, a pair of proteins can be expressed by a 60 dimensional vector.

Weighted Sparse Representation Based Classification (WSRC)

In recent years, inspired by the theory of compressed sensing, Wright et al. (2009) proposed a sparse representation based classification (SRC). The algorithm has been proven useful and reliable for many applications. Later, Fan et al. (2015) proposed a weighted sparse representation based classification (WSRC), which introduced sample weights into training samples and enhanced the robustness of classification. Usually the representation result of WSRC is sparser than that of SRC, so better recognition results can be obtained. Here we give a brief introduction towards WSRC.

Suppose that training samples are classified into C classes. Let $X = [X_1, X_2,\ldots, X_c] \in R^{d \times n}$, where $X_i \in R^{d \times n\ i}$ is the n_i training sample of class i. Given a test sample $y \in R^d$: $y = X\alpha$, where $\alpha = [\alpha_1, \alpha_2,\ldots, \alpha_c]$, α_i is the representation coefficient vector associated with the i-th class. WSRC keeps data relativity while

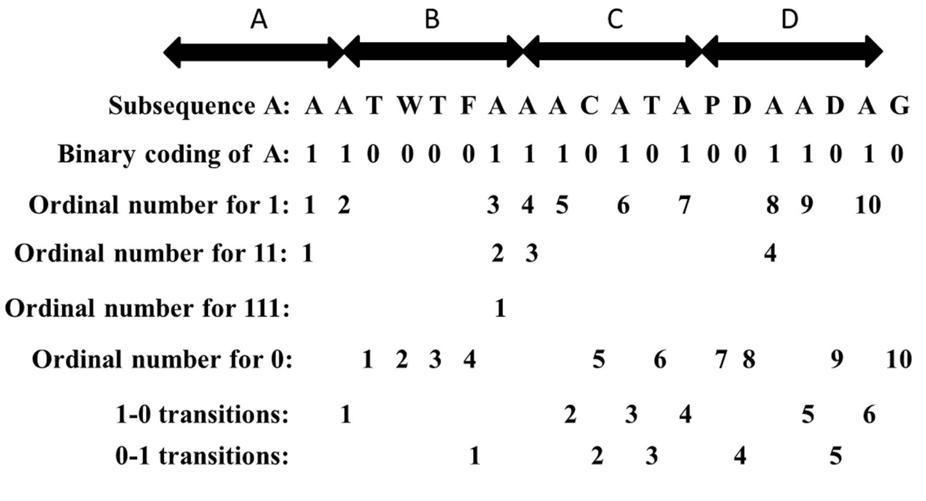

A
B
C
D
Subsequence A: A A T W T F A A A C A T A P D A A D A G
Binary coding of A: 1 1 0 0 0 0 1 1 1 0 1 0 1 0 0 1 1 0 1 0
Ordinal number for 1: 1 2 3 4 5 6 7 8 9 10
Ordinal number for 11: 1 2 3 4
Ordinal number for 111: 1
Ordinal number for 0: 1 2 3 4 5 6 7 8 9 10
1-0 transitions: 1 2 3 4 5 6
0-1 transitions: 1 2 3 4 5

FIGURE 2 | The composition and transition of subsequence "AATWTFAAACATAPDAADAG" with respect to amino acid A.

sparse representation makes coding localized and allows more neighboring samples to express the samples to be tested. The training samples nearer to the test samples should be given smaller weights to make their corresponding coefficients larger. The objective function is:

$$(Weighted\ \ l^1) : \min\ ||\mathrm{W}\alpha||_1 \tag{7}$$

subject to

$$y = X\alpha \tag{8}$$

Dealing with occlusion, the Equations (7) and (8) should be extended to the stable *l*\s\do5(1)–*minimization* problem:

$$\hat{\alpha} = \arg \quad min \quad ||\alpha||_1 \tag{9}$$

subject to

$$\| y - X\alpha \| \leq \quad \epsilon\,. \tag{10}$$

ε > 0 is the tolerance of reconstruction error. After obtaining the sparsest solution $\hat{\alpha}$, we assign a test sample *y* to the class *i* by the following rule:

$$min_i r_i(y) = \quad \| y - X\hat{\alpha}^i \|, i = 1, 2, \ldots, c\,. \tag{11}$$

and specifically,

$$diag(W) = [d(y, x_1^1), \ldots, d(y, x_{n_c}^c)]\,. \tag{12}$$

W is a diagonal matrix used to adjust the weight of training samples to express the test samples and n_c is the sample number of training set in class *c*. WSRC calculates the Gaussian similarities between the test sample and the entire training samples, which are used as the weight of each training sample. The Gaussian similarity between two samples, a1 and a2, could be defined as follows:

$$d(a_1, a_2) = \exp\left(-\frac{\| a_1 - a_2 \|^2}{2\sigma^2}\right) \tag{13}$$

where σ means the Gaussian kernel width. In this paper, we take the parameters ϵ = 0.005, σ = 1.5. The WSRC algorithm can be described as follows:

ALGORITHM 1 | Weighted sparse representation based classification (WSRC).

```
INPUT:
  The matrix of training samples $X \in R^{d\times n}$ and a test sample $y \in R^d$.
OUTPUT:
  The prediction label of $y$ as $\textit{identify}(y) = \arg\min_i r_i(y)$.
1: Normalize each column of $X$ to have the unit $l_2$ norm.
2: Calculate the Gaussian similarity between $y$ and each sample in $X$ and obtain the weight matrix $W$.
3: Solve the stable $l_1$—minimization problem described in Equation (7).
4: Calculate residual error: $\min_i r_i(y) = \| y - X\hat{\alpha}^i \|, \quad i = 1, 2, \ldots, c$.
5: return $y$;
```

DATASET

In this paper, *H. pylori*, *Yeast*, and *Human* PPIs datasets are downloaded from the DIP database (Xenarios et al., 2002). Cd-hit (Li et al., 2001) is a tool for protein sequence clustering that clusters sequences based on their similarity. This article uses the cd-hit tool to remove redundant sequences such that the protein interaction dataset has less than 40% homology and builds a non-redundant dataset (Shawn et al., 2005). Thus, the *H. pylori* dataset contains 1,428 pairs of interacting proteins, the *Yeast* dataset contains 5,594 pairs of interacting proteins, and the *Human* dataset contains 3,899 pairs of interacting proteins. The choice of negative samples is crucial. This paper constructs a non-interacting dataset (negative sample) based on the protein interaction dataset (positive sample) that has been obtained (Yanzhi et al., 2008; You et al., 2015). Sequences in non-interacting protein pairs are randomly selected from a positive samples, but several conditions need to be met: (1) Non-interacting sequence pairs cannot appear in the interaction dataset. (2) The number of protein pairs in a non-interacting dataset should be balanced with the interacting dataset. (3) The contribution of each protein sequence in the non-interacting dataset should be as consistent as possible. Through this strategy, 1458 negative samples of *H. pylori*, 5,594 negative samples of *Yeast*, and 4,262 negative samples of *Human* are obtained. Thus, the *H. pylori* dataset has a total of 2,916 pairs of protein sequences, the *Yeast* dataset has a total of 11,188 pairs of protein sequences, and the *Human* dataset has a total of 8,161 pairs of protein sequences. Furthermore, in order to construct a PPIs network model, three significant PPIs network datasets are performed: the single-core network (CD9), the multiple-core network (Ras-Raf-Mek-Erk-Elk-Srf pathway), and the cross-connection network (Wnt-related Network).

EVALUATION OF THE PREDICTION PERFORMANCE

Here, we employ five fold cross validation to evaluate the performance of the FCTP-WSRC model. The entire dataset is divided into five groups randomly, four of which are used as the training samples and the remaining one as the test samples. The average performance on five sets is used as the performance of our method. Several evaluation indicators are used to evaluate the performance of the development methods of this article. Brief descriptions of these metrics are as follows: (1) sensitivity (Sn) is the percentage of correctly identified interacting protein pairs; (2) specificity (Sp) is the percentage of correctly identified non-interacting protein pairs; (3) accuracy (Acc) is the percentage of correctly identified protein pairs; (4) matthew's correlation coefficient (Mcc) is a stricter evaluation standard considering both under and over predictions. Some concepts and terms to explain this parameters are defined as follows (You et al., 2013):

$$\begin{cases} Sn = \frac{TP}{TP+FN} \\ Sp = \frac{TN}{TN+FP} \\ Acc = \frac{TP+TN}{TP+FP+TN+FN} \\ Mcc = \frac{(TP)(TN)-(FP)(FN)}{\sqrt{[TP+FP][TP+FN][TN+FP][TN+FN]}} \end{cases} \tag{14}$$

where TP is the number of true positive; FN is the number of false negative; TN is the number of true negative; and FP is the number of false positive. In addition, the ROC curve and the area under an ROC curve (AUC) (Huang et al., 2016a) are employed to evaluate the performance of the FCTP-WSRC approach.

DISCUSSION

Prediction Ability

For the sake of testing the stability and reliability of the results, we employ a fivefold cross validation for three typical dataset. For the practicality and effectiveness of our proposed method, we conduct ten times five fold cross validations and use the average results as the final experimental results. We obtain the final results of Acc, Sn, Sp, and Mcc of 96.67%, 95.42%, 97.85%, and 93.56% on the *H. pylori* dataset. Moreover, we obtain excellent performance of average accuracy, sensitivity, specificity, and Mcc of 99.82%, 99.88%, 99.77%, 99.63% on the *Human* dataset and 98.09%, 99.45%, 96.82%, 96.25% on the *Yeast* dataset, respectively. What's more, I have compared the feature selection PCA with the current state-of-the-art feature selection methods EFS on the *Human* dataset. The Acc, Sn, Sp and Mcc of EFS are 0.9499, 0.9601, 0.9448, and 0.9045, respectively, which are lower than our method PCA+WSRC. The comparison of the effects of different feature numbers based on PCA is shown in **Figure 3**.

The Prediction Performance Comparison of FCTP-WSRC With FCTP-SVM

To further verify the effectiveness of the FCTP-WSRC approach, we compare the predictions with the frequently used classifier support vector machine (SVM). The kernel functions commonly used in support vector machines are: linear kernel, polynomial kernel and radial basis kernel function. Linear kernel is mainly used in the case of linear separability. The dataset in this paper has a low feature dimension and is linear inseparability. Compared with the polynomial kernel function, the radial basis kernel function needs to determine fewer parameters, and the more parameters the more complicated the model. Through experiments, we use the LIBSVM (Chang and Lin, 2011) implementation of SVM with the radial basis kernel function:

$$k\quad(x,y) = exp(\frac{\| x- \quad y\|^2 \,\|}{2\sigma^2}) \tag{15}$$

The prediction results of the SVM and WSRC methods on the H. pylori, Human and Yeast datasets are shown in **Table 3**, and the bar chart is displayed in **Figure 5A**. From these results, we can see that the WSRC classifier is significantly better than the SVM classifier. In addition, the ROC (receive operator characteristic) curve illustrating the performance of different classification methods. The curve presents the sensitivity (the true positive rate) against the specificity (the false positive rate). The ROC curves of FCTP-WSRC on the H.

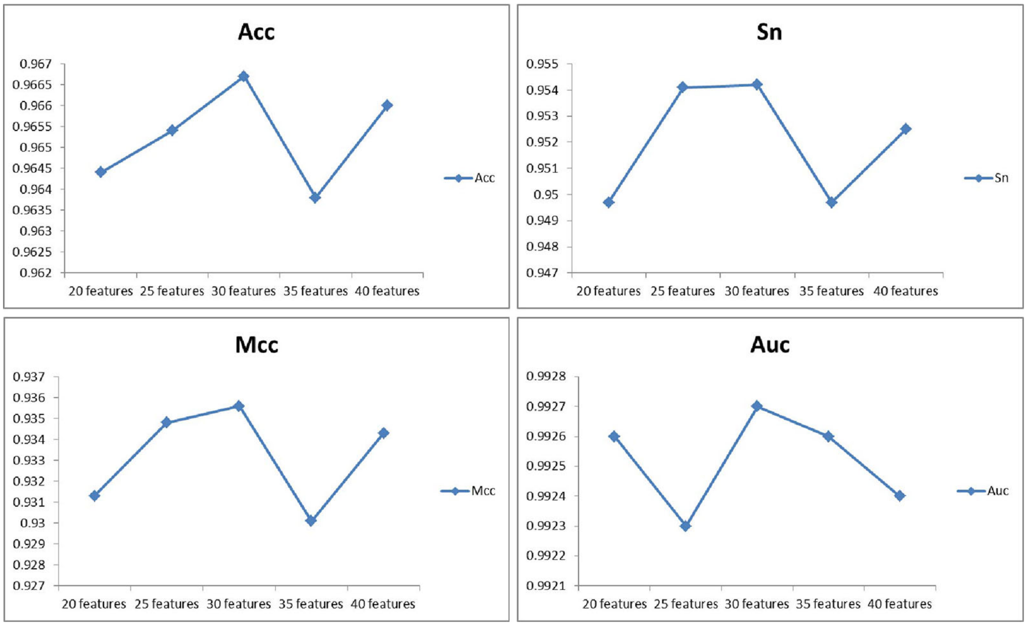


FIGURE 3 | The comparison of the effects of different feature numbers based on principal component analysis (PCA).

TABLE 3 | The prediction performance comparison of FCTP-WSRC with FCTP-SVM.

Dataset	Classification model	Acc	Sn	Sp	Mcc	AUC
H. pylori dataset	SVM	0.9215	0.9191	0.9235	0.8552	0.9718
	WSRC	**0.9667**	**0.9542**	**0.9785**	**0.9356**	**0.9927**
Human dataset	SVM	0.9914	0.9911	0.9925	0.9830	0.9992
	WSRC	**0.9982**	**0.9988**	**0.9977**	**0.9963**	**1**
Yeast dataset	SVM	0.9482	0.9560	0.9411	0.9019	0.9846
	WSRC	**0.9809**	**0.9945**	**0.9682**	**0.9625**	**0.9986**

Bolded texts are used to emphasize the results of the method designed in this article.

pylori, Human and Yeast datasets are shown in **Figure 4A** and those of FCTP-SVM are shown in **Figure 4B**. Good performance is reflected in curves with stronger bending towards the upper-left corner of the ROC graph, that is, high sensitivity is achieved with a low false positive rate. For all models, the areas under an ROC curves (AUC) are > 97.18%. It can be seen from **Figure 4** that the ROC curves of the WSRC classifier are significantly better than those of the SVM classifier. This clearly prove that the WSRC classifier of the proposed method is an accurate and robust classifier for predicting PPIs. The increased classification performance of the WSRC classifier compared with the SVM classifier can be explained by two reasons: (1) the obvious advantage of WSRC is that it does not need to select and compute kernel functions. (2) Protein sequence data expressed by FCTP method is very sparse, so it is suitable for PPIs prediction by sparse representation classifier.

Comparison With Other Methods

Tables 4–6 compare the prediction performance by the proposed method (FCTP-WSRC) and some outstanding works on the *H. pylori*, *Yeast* and *Human* dataset. **Table 4** describes the average accuracies of other seven methods including HKNN (Nanni, 2005), Signature products (Shawn et al., 2005), Ensemble of HKNN (Nanni and Lumini, 2006), PCA+ELM (You et al., 2013), WSRC+GE (Nanni and Lumini, 2006), HOG +SVD+RF (Ding et al., 2016), and RVM+BiGP (An et al., 2016). **Table 5** describes the average accuracies of other seven methods including LDA+RF (Xiao-Yong et al., 2010), LDA+RoF (Xiao-Yong et al., 2010), AC+RF (Xiao-Yong et al., 2010), AC+RoF [41], WSRC+GE (Huang et al., 2016a), and HOG+SVD+RF (Ding et al., 2016). **Table 6** describes the average accuracies of other seven methods including AutoCC (Yanzhi et al., 2008), SVM+LD (Guo et al., 2015), RF+PR+LPQ (Wong et al., 2015), PCA+ELM (You et al., 2013), WSRC+PSM (Huang et al., 2016b), HOG+SVD+RF (Ding et al., 2016), and RVM+BiGP (An et al., 2016). These results using distinct methods on three datasets are intuitively shown by **Figure 5B**. All the results prove that our method improves predictions by using fixed-length feature vectors.

Network Prediction

An effective application of a good PPIs prediction method should have a good ability to predict PPI networks. Up to now, many machine learning approaches have been applied to predict PPIs networks. Despite this, there is still room to improve the accuracy and stability. Therefore, we have extended the prediction method of PPI networks consisting of PPI pairs: the single-core network (CD9), the multiple-core network (Ras-Raf-Mek-Erk-Elk-Srf pathway), and the cross-connection network (Wnt-related Network). The prediction results and the networks are shown in **Figures 6–8**. The black line is predicted correctly,

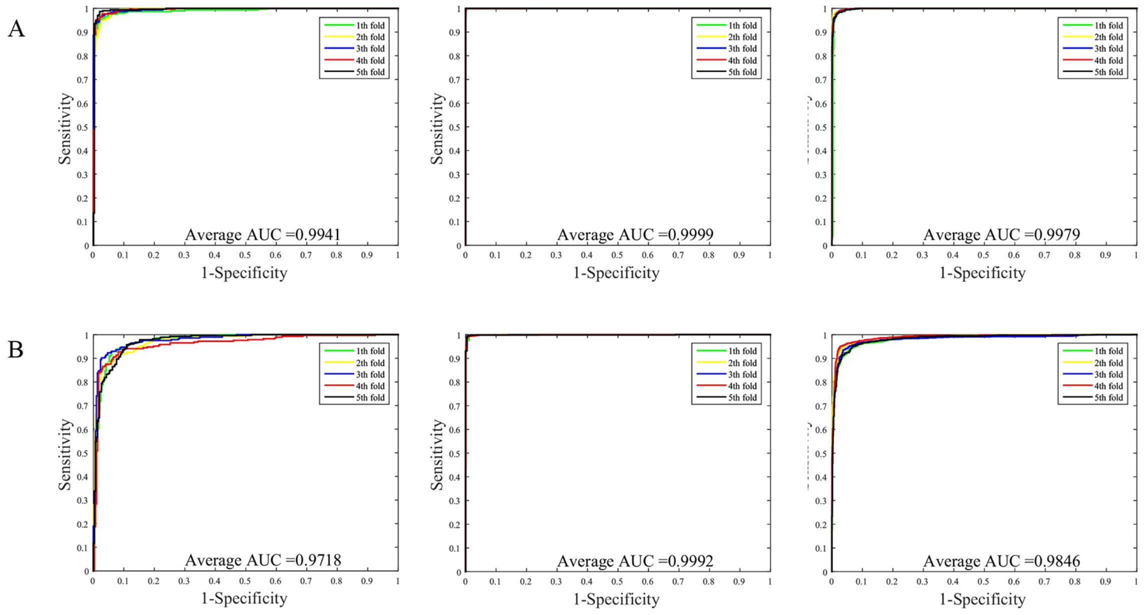


FIGURE 4 | **(A)** ROC curve of FCTP-WSRC on the H. pylori, Human and Yeast datasets. **(B)** ROC curve of FCTP-SVM on the H. pylori, Human and Yeast datasets.

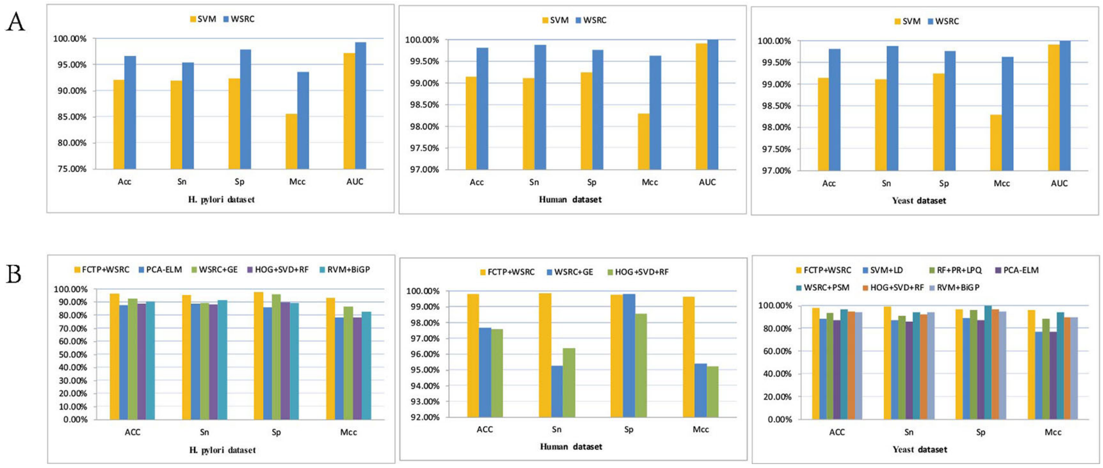


FIGURE 5 | (A) Results using FCTP encoding on the H. pylori, Human and Yeast datasets with different classifiers. **(B)** Results using different methods on three datasets.

the red line is predicted error, and the yellow node is the core protein.

CD9 is a four-pass transmembrane protein superfamily composed of multiple homologous membrane proteins, which is widely distributed in different tissues of human body and participates in the regulation of sperm-egg binding. It plays an important role in cell membrane biology in connection with cell support, adhesion, movement, proliferation, fusion and metastasis of tumor cells. This paper uses the CD9 single-core network dataset, where a protein interacts radially with other proteins (Yang et al., 2006). The result indicates that all 16 PPIs could be identified by our method. The accuracy of this method is 18.75% higher than that of Shen's work (Juwen et al., 2007).

The Ras-Raf-Mek-Erk-Elk-Srf pathway is a widely activated mitogen-activated protein kinase signaling pathway that is complex, highly conserved and widely found in eukaryotic cells. It can transmit extracellular signals into the nucleus, causing changes in the expression profile of specific proteins in the cells, which in turn affects cell fate, and is closely related to the development of tumors (Davis, 2010). Ras, Raf, Mek, Erk, Elk, and Srf act as core proteins that determine signal transduction. Our method has a prediction accuracy of 95.96%, which is better than 85.19% of Shen's work (Juwen et al., 2007).

The Wnt signaling pathway is a group of multiple downstream channel signaling pathways that are excited by the binding of the ligand protein Wnt and membrane protein receptors. In biology, most PPIs network is the cross-connection network. While Wnt-related pathways are essential for signal transduction, the use of scientific computing methods to predict Wnt-related network has important practical significance (Stelzl et al., 2005). The accuracy of Shen's work is 96.04% in the network, our method is 100% which is best.

TABLE 4 | Comparing the prediction performance by the proposed method (FCTP-WSRC) and some state-of-art works on the *H. pylori* dataset.

Model	ACC	Sn	Sp	Mcc
Our method	**0.9667**	**0.9542**	**0.9785**	**0.9356**
HKNN	0.8400	0.8600	0.8400	N/A
Signature products	0.8340	0.7990	0.8570	N/A
Ensemble of HKNN	0.8660	0.8670	0.8500	N/A
PCA+ELM	0.8750	0.8895	0.8615	0.7813
WSRC+GE	0.9283	0.8932	0.9613	0.8643
HOG+SVD+RF	0.8906	0.8815	0.8979	0.7815
RVM+BiGP	0.9057	0.9188	0.8955	0.8291

Here, N/A means not available. Bolded texts are used to emphasize the results of the method designed in this article.

TABLE 5 | Comparing the prediction performance by the proposed method (FCTP-WSRC) and some state-of-art works on the *Human* dataset.

Model	ACC	Sn	Sp	Mcc
Our method	**0.9982**	**0.9988**	**0.9977**	**0.9963**
LDA+RF	0.9640	0.9420	N/A	0.9280
LDA+RoF	0.9570	0.9760	N/A	0.9180
AC+RF	0.9550	0.9400	N/A	0.9140
AC+RoF	0.9510	0.9330	N/A	0.9100
WSRC+GE	0.9766	0.9528	0.9981	0.9541
HOG+SVD+RF	0.9760	0.9637	0.9859	0.9521

N / A means that the result of this indicator is not queried.

TABLE 6 | Comparing the prediction performance by the proposed method (FCTP-WSRC) and some state-of-art works on the *Yeast* dataset.

Model	ACC	Sn	Sp	Mcc
Our method	**0.9809**	**0.9945**	**0.9682**	**0.9625**
AutoCC	0.8933	0.8993	0.8887	N/A
SVM+LD	0.8856	0.8737	0.8950	0.7715
RF+PR+LPQ	0.9392	0.9110	0.9645	0.8856
PCA+ELM	0.8700	0.8615	0.8759	0.7736
WSRC+PSM	0.9709	0.9433	1	0.9433
HOG+SVD+RF	0.9483	0.9240	0.9710	0.8977
RVM+BiGP	0.9457	0.9427	0.9486	0.8974

N / A means that the result of this indicator is not queried.

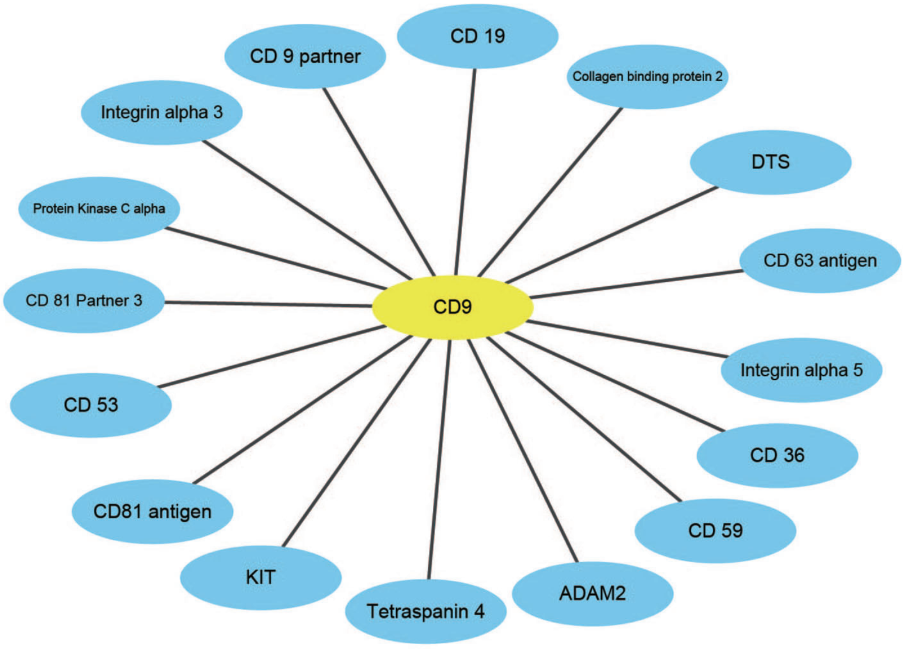


FIGURE 6 | The prediction results of single-core network of CD9.

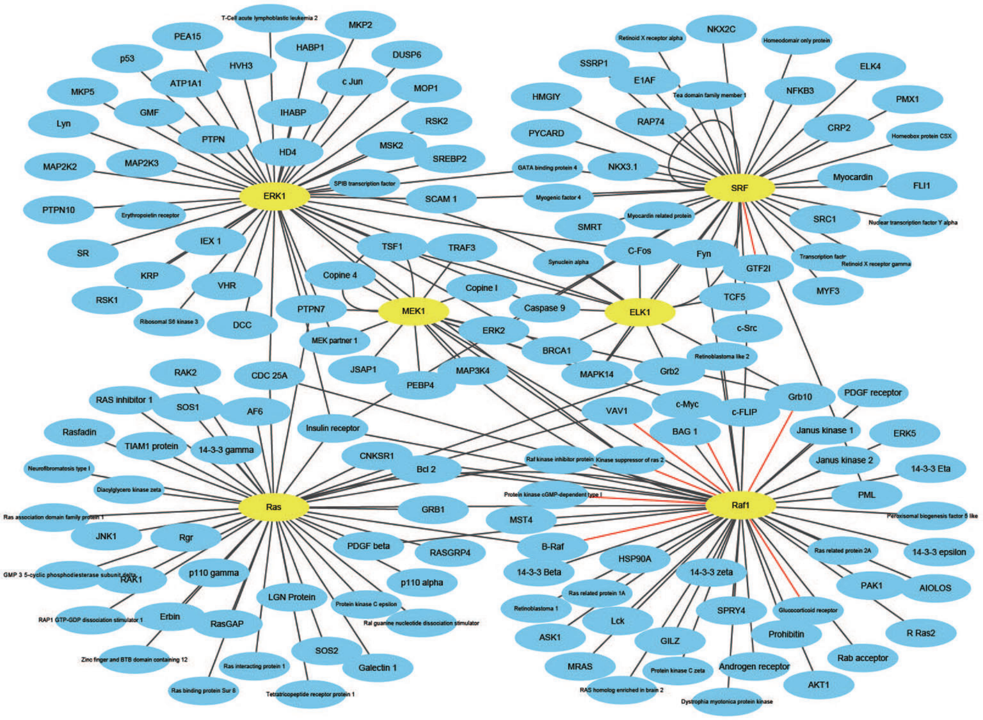


FIGURE 7 | The prediction results of multi-core network of Ras-Raf-Mek-Erk-Elk-Srf pathway.

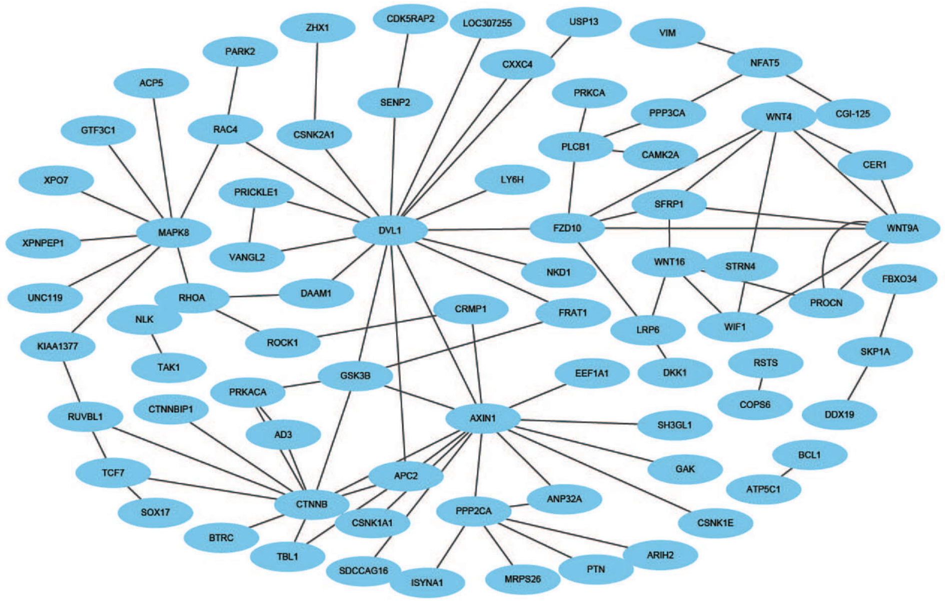


FIGURE 8 | The prediction results of cross-connection network of Wnt-related pathway.

TABLE 7 | Protein-protein interaction information obtained by a web tool PIE.

Protein-protein interaction	PMID	PPI score
CD9-CD19	9804823	0.7703
CD9-CD9 partner	16690612	0.9999
CD9-Integrin alpha 3	7790364	0.9999
CD9-Protein Kinase C alpha	11325968	0.7479
CD9-CD81 Partner 3	16690612	0.9999
CD9-CD53	23500527	0.818
CD9-CD81 antigen	16690612	0.9999
CD9-KIT	12036870	0.7073
CD9-Tetraspanin 4	27993971	0.9502
CD9-ADAM2	10518536	0.557
CD9-CD59	15625824	-0.0798
CD9-CD36	17684062	0.6525
CD9-Integrin alpha 5	10811835	0.8497
CD9-CD63 antigen	19640571	0.7556
CD9-DTS	8367482	0.1173
CD9-Collagen binding protein 2	9931299	0.5501

Evaluating the Performance of FCTP-WSRC by PIE Software

PIE (Protein Interaction information Extraction) the search is a web service to extract PPI-relevant articles from MEDLINE (Sun et al., 2012), which can be used *via* a web application at http://www.ncbi.nlm.nih.gov/IRET/PIE/. It implement a competition-winning approach utilizing word and syntactic analyses by machine learning techniques. For easy user access, PIE the search provides a PubMed-like search environment, but the output is the list of articles prioritized by PPI confidence scores. PPI score is a relative value between 1.0 (highly likely) and -1.0 (highly unlikely) among retrieved articles. From **Table 7**, we can see that only CD9-CD59 is negative 0.0798, which is very close to zero obtained by the web tool PIE. That is to see, PPI-relevant articles extracted by the PIE cannot predict the relationship between CD9 and CD59. This also shows that our method can be used to predict potential PPI.

Conclusion

The problem of predicting PPIs has been tackled extensively. Given the fact that computational tools for predicting PPIs have been used over years, only a few of them are able to predict easily, quickly, and accurately. Above all, we have explored a novel computational tool called FCTP-WSRC to predict PPIs efficiently. We characterize a fixed-length feature vector of protein sequence using descriptors F-vector, composition (C), and transition (T).

Our numerical results demonstrate that the WSRC classifier model is feasible to perform PPIs detection. We see that FCTP-WSRC perform significantly well when it comes to distinguish positive samples and negative samples of protein pairs. That is to say, these results support the notion that our FCTP-WSRC model is a highly effective proteomics research support tool. In the future, we will extend our approach to more significant PPI networks with unknown biological functions.

Code is programmed by MATLAB, which can be downloaded from https://github.com/wowkiekong/PPI-prediction. User-friendly and publicly accessible web-servers represent the future direction for developing practically more useful computational tools and enhancing their impact (Chou, 2017). Our future efforts will be to establish a web-server for the prediction method reported in this paper.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

MK, YZ, and DX contributed conception and design of the study. YZ and WC performed the data processing. MK and DX constructed the protein–protein interactions prediction model. MK wrote the first draft of the manuscript. YZ, WC, DX, and MD wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

ACKNOWLEDGMENTS

We gratefully acknowledge the anonymous reviewers who read our paper and gave some constructive comments. This work is supported by the National Natural Science Foundation of China (Nos. 61877064).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00018/full#supplementary-material

REFERENCES

An, J. Y., Meng, F. R., You, Z. H., Chen, X., Yan, G. Y., and Hu, J. P. (2016). Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model. *Protein Sci. A Publ. Protein Soc.* 25, 1825–1833. doi: 10.1002/pro.2991

Capra, J. A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics* 23, 1875–1882. doi: 10.1093/bioinformatics/btm270

Chang, C. C., and Lin, C. J. (2011). Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST)* 2, 1–27. doi: 10.1145/1961189.1961199

Chou, K. C. (2017). An unprecedented revolution in medicinal chemistry driven by the progress of biological science. *Curr. Top. In Med. Chem.* 17, 2337–2358. doi: 10.2174/1568026617666170414145508

Chuang, L. Y., Moi, S. H., Lin, Y. D., and Yang, C. H. (2016). A comparative analysis of chaotic particle swarm optimizations for detecting single nucleotide polymorphism barcodes. *Artif. Intell. In Med.* 73, 23–33. doi: 10.1016/j.artmed.2016.09.002

Davies, M. N., Secker, AA Andrew, Andrew, F., Clark, E., Timmis, J., and Flower, D. R. (2008). Optimizing amino acid groupings for gpcr classification. *Bioinformatics* 24, 1980–1986. doi: 10.1093/bioinformatics/btn382

Davis, R. J. (2010). Transcriptional regulation by map kinases. *Mol. Reprod. Dev.* 42, 459–467. doi: 10.1002/mrd.1080420414

Deng, S. P., Zhu, L., and Huang, D. S. (2015). Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks. *BMC Genomics* 16, S4. doi: 10.1186/1471-2164-16-S3-S4

De-Shuang, H., and Chun-Hou, Z. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* 22, 1855–1862. doi: 10.1093/bioinformatics/btl190

De-Shuang, H., and Ji-Xiang, D. (2008). A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. *IEEE Trans. Neural Networks* 19, 2099–2115. doi: 10.1109/TNN.2008.2004370

Ding, Y., Tang, J., and Guo, F. (2016). Identification of protein-protein interactions *via a* novel matrix-based sequence representation model with amino acid contact information. *Int. J. Mol. Sci.* 17, 1623. doi: 10.3390/ijms17101623

Du, X., Sun, S., Hu, C., Yao, Y., Yan, Y., and Zhang, Y. (2017). Deepppi: Boosting prediction of protein-protein interactions with deep neural networks. *J. Chem. Inf. Model.* 57, 1499–1510. doi: 10.1021/acs.jcim.7b00028. acs.jcim.7b00028.

Fan, Z., Ming, N., Qi, Z., and Liu, E. (2015). Weighted sparse representation for face recognition. *Neurocomputing* 151, 304–309. doi: 10.1016/j.neucom.2014.09.035

Gavin, A., Bösche, M., Krause, R., Grandi, P., Marzioch, M., and Andreas, B. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147. doi: 10.1038/415141a

Guo, F., Ding, Y., Li, Z., and Tang, J. (2015). Identification of protein-protein interactions by detecting correlated mutation at the interface. *J. Chem. Inf. Model.* 55, 2042–2049. doi: 10.1021/acs.jcim.5b00320

Huang, Y. A., You, Z. H., Chen, X., Chan, K., and Luo, X. (2016a). Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. *BMC Bioinf.* 17, 184. doi: 10.1186/s12859-016-1035-4

Huang, Y. A., You, Z. H., Hu, P., Li, S., Luo, X., and Wong, L. (2016b). Construction of reliable protein-protein interaction networks using weighted sparse representation based classifier with pseudo substitution matrix representation features. *Neurocomputing* 218, 131–138. doi: 10.1016/j.neucom.2016.08.063

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A* 98, 4569–4574. doi: 10.1073/pnas.061034498

Juan, C., Yi, H. L., Hu, L., Choong Yong, U., Qun, T. Z., Juan, Z. C., et al. (2007). Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol. Immunol.* 44, 514–520. doi: 10.1016/j.molimm.2006.02.010

Juwen, S., Jian, Z., Xiaomin, L., Weiliang, Z., Kunqian, Y., Kaixian, C., et al. (2007). Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U. S. A* 104, 4337–4341. doi: 10.1073/pnas.0607879104

Kim, S., Kwon, D., Shin, S.Y., and Wilbur, W.J. (2012). . Pie the search: searching pubmed literature for protein interaction information. *Bioinformatics* 28, 597–598. doi: 10.1093/bioinformatics/btr702

Kuo-Chen, C. (2005). Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* 21, 10–19. doi: 10.1093/bioinformatics/bth466

Lei, Y., Jun-Feng, X., and Jie, G. (2010). Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein Pept. Lett.* 17, 1085–1090. doi: 10.2174/092986610791760306

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17, 282–283. doi: 10.1093/bioinformatics/17.3.282

Li, X., Liao, B., Shu, Y., Zeng, Q., and Luo, J. (2009). Protein functional class prediction using global encoding of amino acid sequence. *J. Theor. Biol.* 261, 290–293. doi: 10.1016/j.jtbi.2009.07.017

Liu, J., Gutman, I., Mu, Z., and Zhang, Y. (2012). Q-analog of wiener index. *Appl. Math. Comput.* 218, 9528–9535. doi: 10.1016/j.amc.2012.03.048

Long, Z., Yu, G., Xia, D., and Wang, J. (2019). Protein-protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* 324, 10–19. doi: 10.1016/j.neucom.2018.02.097

Mirny, L. A., and Shakhnovich, E. I. (1999). Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. *J. Mol. Biol.* 291, 177–196. doi: 10.1006/jmbi.1999.2911

Nanni, L., and Lumini, A. (2006). An ensemble of k-local hyperplanes for predicting protein-protein interactions. *Bioinformstics* 22, 1207–1210. doi: 10.1093/bioinformatics/btl055

Nanni, L. (2005). Hyperplanes for predicting protein-protein interactions. *Neurocomputing* 69, 257–263. doi: 10.1016/j.neucom.2005.05.007

Neumann, U., Genze, N., and Heider, D. (2017). Efs: an ensemble feature selection tool implemented as r-package and web-application. *Biodata Min.* 10, 21. doi: 10.1186/s13040-017-0142-8

Pazos, F., and Valencia, A. (2002). In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins Struct. Funct. Bioinf.* 47, 219–227. doi: 10.1002/prot.10074

Shawn, M., Diana, R., and Jean-Loup, F. (2005). Predicting protein-protein interactions using signature products. *Bioinformatics* 21, 218–226. doi: 10.1093/bioinformatics/bth483

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., and Goehler, H. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968. doi: 10.1016/j.cell.2005.08.029

Tong, J. C., and Tammi, M. T. (2008). Prediction of protein allergenicity using local description of amino acid sequence. *Front. In Biosci. A J. Virtual Library* 13, 6072. doi: 10.2741/3138

Wold, S., Jonsson, J., Sjörström, M., Sandberg, M., and Rännar, S. (1993). Dna and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Analytica Chim. Acta* 277, 239–253. doi: 10.1016/0003-2670(93)80437-P

Wong, L., You, Z. H., Li, S., Huang, Y. A., and Liu, G. (2015). Detection of protein-protein interactions from amino acid sequences using a rotation forest model with a novel pr-lpq descriptor. *Lecture Notes In Comput. Sci.* 9227, 713–720. doi: 10.1007/978-3-319-22053-6\s\do5(7)5

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 210–227. doi: 10.1109/TPAMI.2008.79

Xenarios, I., Salwínski, L., Duan, X. J., Higney, P., Kim, S. M., and Eisenberg, D. (2002). Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303. doi: 10.1093/nar/30.1.303

Xiao-Yong, P., Ya-Nan, Z., and Hong-Bin, S. (2010). Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *J. Proteome Res.* 9, 4992–5001. doi: 10.1021/pr100618t

Xu, C., Ge, L., Zhang, Y., Dehmer, M., and Gutman, I. (2017). Prediction of therapeutic peptides by incorporating q-wiener index into chou's general pseaac. *J. Biomed. Inf.* 75, 63–69. doi: 10.1016/j.jbi.2017.09.011

Yang, X. H., Kovalenko, O. V., Kolesnikova, T. V., Andzelm, M. M., Rubinstein, E., Strominger, J. L., et al. (2006). Contrasting effects of ewi proteins, integrins, and protein palmitoylation on cell surface cd9 organization. *J. Biol. Chem.* 281, 12976–12985. doi: 10.1074/jbc.M510617200

Yang, C. H., Chuang, L. Y., and Lin, Y. D. (2017). Multiobjective differential evolution-based multifactor dimensionality reduction for detecting gene-gene interactions. *Sci. Rep.* 7, 12869. doi: 10.1038/s41598-017-12773-x

Yanzhi, G., Lezheng, Y., Zhining, W., and Menglong, L. (2008). Using support vector machine combined with auto covariance to predict protein-protein interactions from protein sequences. *Nucleic Acids Res.* 36, 3025–3030. doi: 10.1093/nar/gkn159

You, Z. H., Lei, Y. K., Zhu, L., Xia, J., and Wang, B. (2013). Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinf.* 14, S10. doi: 10.1186/1471-2105-14-S8-S10

You, Z. H., Zhu, L., Zheng, C. H., Yu, H. J., Deng, S. P., and Ji, Z. (2014). Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinf.* 15, S9. doi: 10.1186/1471-2105-15-s15-s9

You, Z. H., Li, J., Gao, X., He, Z., Zhu, L., Lei, Y. K., et al. (2015). Detecting protein-protein interactions with a novel matrix-based protein sequence representation and support vector machines. *BioMed. Res. Int.* 2015, 1–9. doi: 10.1155/2015/867516

Yu, L., Zhang, Y., Gutman, I., Shi, Y., and Dehmer, M. (2017). Erratum: Protein sequence comparison based on physicochemical properties and the position-feature energy matrix. *Sci. Rep.* 7, 46237. doi: 10.1038/srep46237

Zhang, S., Ye, F., and Yuan, X. (2012). Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences *via* pssm. *J. Biomol. Struct. Dyn.* 29, 1138–1146. doi: 10.1080/07391102.2011.672627

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership