



COMPUTATIONAL LEARNING MODELS AND METHODS DRIVEN BY OMICS FOR BIOLOGY FOR “THE FIFTH CHINA COMPUTER SOCIETY BIOINFORMATICS CONFERENCE”

EDITED BY: Wang Guohua, Jun Wan, Fa Zhang, Chunhou Zheng and
Liang Cheng

PUBLISHED IN: Frontiers in Genetics



frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence.

The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714

ISBN 978-2-88974-603-3

DOI 10.3389/978-2-88974-603-3

About Frontiers

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

Frontiers Journal Series

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

Dedication to Quality

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

What are Frontiers Research Topics?

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

COMPUTATIONAL LEARNING MODELS AND METHODS DRIVEN BY OMICS FOR BIOLOGY FOR “THE FIFTH CHINA COMPUTER SOCIETY BIOINFORMATICS CONFERENCE”

Topic Editors:

Wang Guohua, Harbin Institute of Technology, China

Jun Wan, Indiana University, Purdue University Indianapolis, United States

Fa Zhang, Chinese Academy of Sciences (CAS), China

Chunhou Zheng, Anhui University, China

Liang Cheng, Harbin Medical University, China

Citation: Guohua, W., Wan, J., Zhang, F., Zheng, C., Cheng, L., eds. (2022).

Computational Learning Models and Methods Driven by Omics for Biology for “The Fifth China Computer Society Bioinformatics Conference”.

Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-603-3

Table of Contents

05	<i>Sc-GPE: A Graph Partitioning-Based Cluster Ensemble Method for Single-Cell</i>	Xiaoshu Zhu, Jian Li, Hong-Dong Li, Miao Xie and Jianxin Wang
14	<i>An Efficient and Easy-to-Use Network-Based Integrative Method of Multi-Omics Data for Cancer Genes Discovery</i>	Ting Wei, Botao Fa, Chengwen Luo, Luke Johnston, Yue Zhang and Zhangsheng Yu
29	<i>Identification of Potential Prognostic Competing Triplets in High-Grade Serous Ovarian Cancer</i>	Jian Zhao, Xiaofeng Song, Tianyi Xu, Qichang Yang, Jingjing Liu, Bin Jiang and Jing Wu
40	<i>Research on Components Assembly Platform of Biological Sequences Alignment Algorithm</i>	Haihe Shi, Gang Wu, Xuchu Zhang, Jun Wang, Haipeng Shi and Shenghua Xu
48	<i>Efficient Multiple Sequences Alignment Algorithm Generation via Components Assembly Under PAR Framework</i>	Haipeng Shi, Haihe Shi and Shenghua Xu
55	<i>Joint L_p-Norm and $L_{2,1}$-Norm Constrained Graph Laplacian PCA for Robust Tumor Sample Clustering and Gene Network Module Discovery</i>	Xiang-Zhen Kong, Yu Song, Jin-Xing Liu, Chun-Hou Zheng, Sha-Sha Yuan, Juan Wang and Ling-Yun Dai
69	<i>TMP- SSurface2: A Novel Deep Learning-Based Surface Accessibility Predictor for Transmembrane Protein Sequence</i>	Zhe Liu, Yingli Gong, Yuanzhao Guo, Xiao Zhang, Chang Lu, Li Zhang and Han Wang
79	<i>Advances in the Identification of Circular RNAs and Research Into circRNAs in Human Diseases</i>	Shihu Jiao, Song Wu, Shan Huang, Mingyang Liu and Bo Gao
87	<i>iEnhancer-EBLSTM: Identifying Enhancers and Strengths by Ensembles of Bidirectional Long Short-Term Memory</i>	Kun Niu, Ximei Luo, Shumei Zhang, Zhixia Teng, Tianjiao Zhang and Yuming Zhao
97	<i>A Novel Biomarker Identification Approach for Gastric Cancer Using Gene Expression and DNA Methylation Dataset</i>	Ge Zhang, Zijong Xue, Chaokun Yan, Jianlin Wang and Huimin Luo
108	<i>Application of Multilayer Network Models in Bioinformatics</i>	Yuanyuan Lv, Shan Huang, Tianjiao Zhang and Bo Gao
115	<i>Enhancement and Imputation of Peak Signal Enables Accurate Cell-Type Classification in scATAC-seq</i>	Zhe Cui, Ya Cui, Yan Gao, Tao Jiang, Tianyi Zang and Yadong Wang
125	<i>PMDFI: Predicting miRNA–Disease Associations Based on High-Order Feature Interaction</i>	Mingyan Tang, Chenzhe Liu, Dayun Liu, Junyi Liu, Jiaqi Liu and Lei Deng

138 *SIns: A Novel Insertion Detection Approach Based on Soft-Clipped Reads*

Chaokun Yan, Junyi He, Junwei Luo, Jianlin Wang, Ge Zhang and
Huimin Luo

146 *RetroScan: An Easy-to-Use Pipeline for Retrocopy Annotation and
Visualization*

Zhaoyuan Wei, Jiahe Sun, Qinhui Li, Ting Yao, Haiyue Zeng and Yi Wang



Sc-GPE: A Graph Partitioning-Based Cluster Ensemble Method for Single-Cell

Xiaoshu Zhu^{1,2}, Jian Li¹, Hong-Dong Li², Miao Xie¹ and Jianxin Wang^{2*}

¹ School of Computer Science and Engineering, Yulin Normal University, Yulin, China, ² Hunan Provincial Key Laboratory on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China

OPEN ACCESS

Edited by:

Chunhou Zheng,
Anhui University, China

Reviewed by:

Xiujuan Lei,
Shaanxi Normal University, China
Jin-Xing Liu,
Qufu Normal University, China
Yannan Bin,
Anhui University, China

*Correspondence:

Jianxin Wang
jxwang@mail.csu.edu.cn

Specialty section:

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

Received: 10 September 2020

Accepted: 23 November 2020

Published: 15 December 2020

Citation:

Zhu X, Li J, Li H-D, Xie M and Wang J
(2020) Sc-GPE: A Graph
Partitioning-Based Cluster Ensemble
Method for Single-Cell.
Front. Genet. 11:604790.
doi: 10.3389/fgene.2020.604790

Clustering is an efficient way to analyze single-cell RNA sequencing data. It is commonly used to identify cell types, which can help in understanding cell differentiation processes. However, different clustering results can be obtained from different single-cell clustering methods, sometimes including conflicting conclusions, and biologists will often fail to get the right clustering results and interpret the biological significance. The cluster ensemble strategy can be an effective solution for the problem. As the graph partitioning-based clustering methods are good at clustering single-cell, we developed Sc-GPE, a novel cluster ensemble method combining five single-cell graph partitioning-based clustering methods. The five methods are SNN-cliq, PhenoGraph, SC3, SSNN-Louvain, and MPGS-Louvain. In Sc-GPE, a consensus matrix is constructed based on the five clustering solutions by calculating the probability that the cell pairs are divided into the same cluster. It solved the problem in the hypergraph-based ensemble approach, including the different cluster labels that were assigned in the individual clustering method, and it was difficult to find the corresponding cluster labels across all methods. Then, to distinguish the different importance of each method in a clustering ensemble, a weighted consensus matrix was constructed by designing an importance score strategy. Finally, hierarchical clustering was performed on the weighted consensus matrix to cluster cells. To evaluate the performance, we compared Sc-GPE with the individual clustering methods and the state-of-the-art SAME-clustering on 12 single-cell RNA-seq datasets. The results show that Sc-GPE obtained the best average performance, and achieved the highest NMI and ARI value in five datasets.

Keywords: single-cell clustering, cluster ensemble, consensus matrix, importance score, graph partitioning

INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) data measures the gene expression level in individual cells instead of the average gene expression level in bulk RNA-seq cells (Stuart and Satija, 2019). So, it has advantages in accurately identifying the transcriptomic signatures for cell types (Grün et al., 2015). Along with the rapid development of scRNA-seq technologies, the cost of sequencing is reduced, and larger datasets are generated, carrying a higher error rate (Vitak et al., 2017). The development brought some computational challenges (Kiselev et al., 2019; Zhu et al., 2019a), for example, (1) high noise. The drop-out rate from reverse transcription failure and sequencing depth would reach 80% (Soneson and Robinson, 2018; Andrews and Hemberg, 2019); (2) high dimension. The dimension

usually exceeds 10,000, making it difficult to measure the similarity of cell pairs; (3) larger sample size. The sample size increases from dozens to hundreds of thousands, which raises the time and complexity involved in identifying cell types (Grun, 2020).

Clustering is an efficient way of analyzing scRNA-seq data to identify novel cell types, and some single-cell clustering methods are proposed (Xu et al., 2019; Yip et al., 2019). However, it can be observed that the clustering results from various clustering methods are different in the number of clusters and cell assignments. Meanwhile, no method performs best on all scRNA-seq datasets. The reason is that the existing methods focus on a different step in identifying cell types, including data denoising (Wang et al., 2018), dimensionality reduction (Wang and Gu, 2018; Becht et al., 2019), similarity measurement (Kim et al., 2019) and clustering (Qi et al., 2019; Zhu et al., 2019b). Notably, the similarity measurement plays an important role in identifying cell types. Some graph partitioning-based clustering methods achieved better performance for the accurate similarity measurement. For example, SNN-cliq (Xu and Su, 2015) constructed a weighted shared nearest neighbor (SNN) graph; and clustered cells by partitioning the cliques on the graph. PhenoGraph (Levine et al., 2015) performed another weighted strategy to generate an SNN graph; and partitioned the graph using the Louvain community detection method. SSNN-Louvain (Zhu et al., 2020) integrated the structural information to construct a structural SNN graph; and clustered cells by modifying the Louvain community detection method. The cells are sorted as per their importance in the initialization step of Louvain community detection method. MPGS-Louvain (Zhu et al., 2019c) constructed a novel global and path-based similarity graph, and also partitioned it using a modified Louvain community detection method. Therefore, it is a challenge to enhance the accuracy of clustering by combining more efficient clustering information in multiple views.

An increasing number of research shows that the cluster ensemble method is a good idea, which integrates the information of each clustering method in a different view (Kuncheva and Vetrov, 2006; Vega-Pons and Ruiz-Shulcloper, 2011; Liu et al., 2019). ISSCE (Yu et al., 2016) designed a clustering ensemble strategy to cluster high dimensional data, including three steps: firstly, the incremental approach was implemented to select clustering members; secondly, the random subspace division was applied to handle high dimensional data; finally, the constraint propagation method was used to integrate prior knowledge. Recently, some cluster ensemble methods for scRNA-seq data have been proposed. SC3 (Kiselev et al., 2017) ensembled several clustering results from *k*-means algorithm into a consensus matrix; and clustered cells using hierarchical clustering (HC). SAFE-clustering (Yang et al., 2019) implemented a hypergraph-based strategy to ensemble CIDR, Seurat, tSNE, and SC3 to construct a consensus matrix. *k*-means was used to cluster cells. They also proposed the SAME-clustering (Huh et al., 2020) methods by using a consensus matrix-based strategy to ensemble the same four clustering methods and combining the Expectation-Maximization algorithm to cluster cells. We find that these cluster ensemble methods are based

on hypergraph-based or voting-based integrated learning and do not consider the different importance of the individual clustering method.

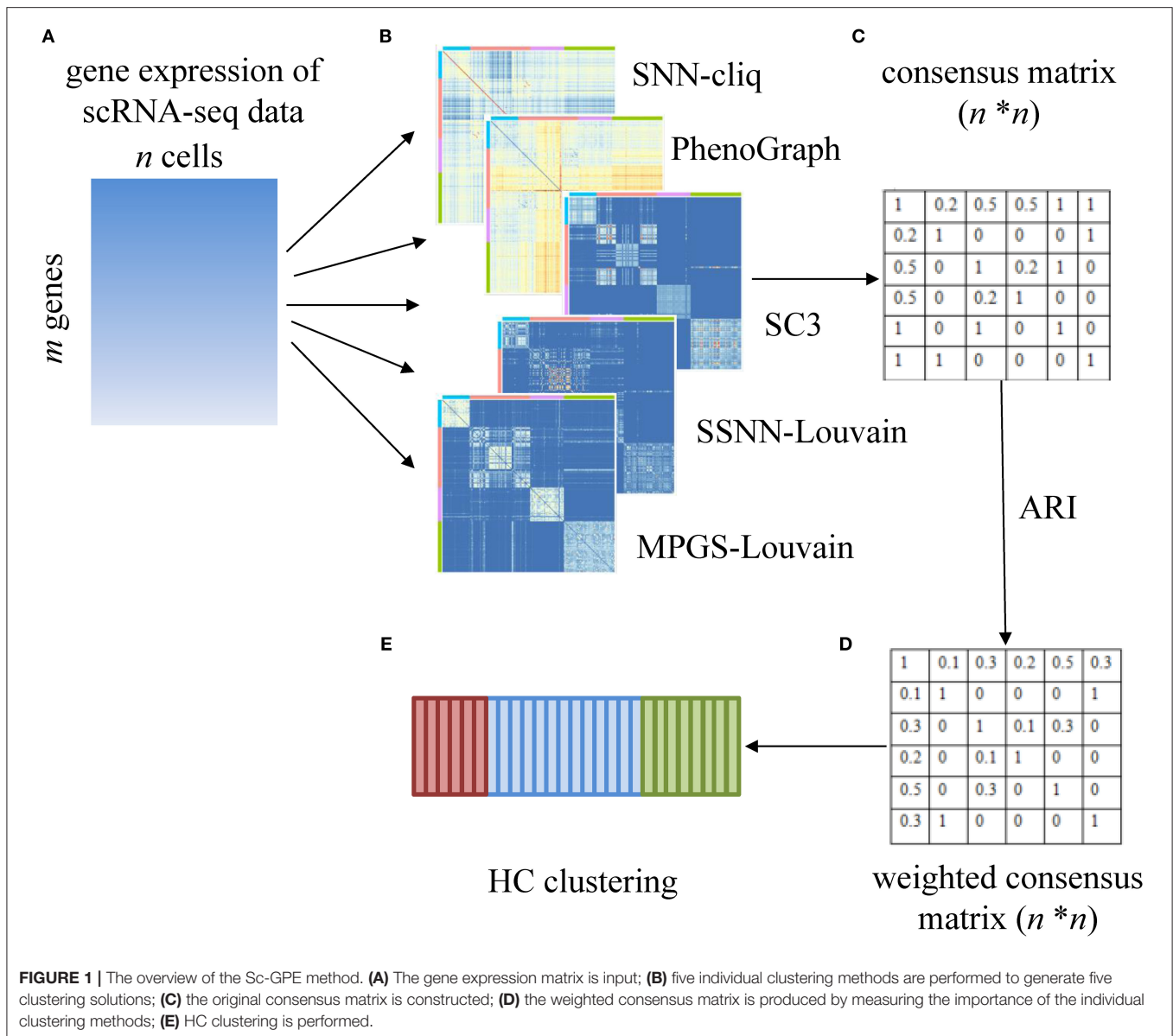
According to the principle that the minority is subordinate to the majority, we assume that the more consistent the cluster labels predicted by different clustering methods are, the more accurate they will be. That is, the individual clustering method with a higher similarity to others would be more important in the cluster ensemble strategy. Based on this assumption, we propose a novel graph partitioning-based ensemble method for single-cell clustering (Sc-GPE), integrating SNN-cliq, PhenoGraph, SSNN-Louvain, MPGS-Louvain, and SC3 by a weighted voting-based method. To measure the importance of the individual clustering method, we design a scoring strategy based on the adjusted rand index (ARI) (Hubert and Arabie, 1985). Then we construct a weighted consensus matrix, the weight is a score of the importance of each method. Finally, HC is performed to cluster cells. To prove the performance, Sc-GPE is compared to the five original clustering methods and the state-of-the-art cluster ensemble method "SAME-clustering." The results demonstrate that Sc-GPE outperforms other methods.

MATERIALS AND METHODS

According to the analysis above, we can find that integrating multiple clustering results would merge more information in different views. Moreover, different clustering methods play different roles in integration. Inspired by these ideas, we propose the Sc-GPE method by ensembling five graph partitioning-based clustering methods which are SNN-cliq, PhenoGraph, SSNN-Louvain, MPGS-Louvain, and SC3. The main reasons for choosing the five clustering methods are as follows: firstly, the first four clustering methods are graph partitioning-based methods, and the last one is the consensus matrix-based method. Their good performance provides the basis to improve the accuracy of the cluster ensemble. Secondly, in the five clustering methods, different strategies of similarity graph construction and graph partitioning have been implemented, respectively. They would enhance the generalization ability of clustering. Sc-GPE has three following advantages: (1) it does not need to deal with the problem of different cluster labels from different cluster methods, so it is suitable for unsupervised clustering lacking the true cluster labels; (2) It is easy to implement since no special parameters need to be adjusted; (3) The weighted strategy is comprehensible and effective.

Sc-GPE

In Sc-GPE, a gene expression matrix with *m* rows (genes) and *n* columns (cells) is the input of the five clustering methods. The five clustering results sets are achieved and ensembled into a consensus matrix with *n* rows (cells) and *n* columns (cells). Then, based on the consensus matrix, a weighted consensus matrix is constructed by measuring the importance of the individual clustering method. That is, the voting strategy in the original consensus matrix is replaced as a weighted voting strategy, and the weight is determined according to the similarity of the



clustering result pairs. The overview of Sc-GPE method is shown in **Figure 1**.

Cells are defined as set $C = \{c_1, \dots, c_n\}$, where n is the number of cells. Let k be the number of individual clustering methods, the clustering results set is defined as $R = \{R^1, \dots, R^k\}$. So, in the k clustering methods, the i -th cell c_i is assigned to k predicted cluster labels, denoted as $R(c_i) = \{R^1(c_i), \dots, R^k(c_i)\}$. The detail of Sc-GPE is described as follows.

Firstly, the original consensus matrix is constructed. The consensus matrix $I_{x,y}$ is calculated based on Equations (1) and (2). In Equations (1) and (2), when the cell c_x and cell c_y are assigned into the same cluster in the l -th method, the value of $\delta(R^l(c_x), R^l(c_y))$ is equal to 1, otherwise is 0. The element of the consensus matrix presents the probability of cell pairs divided into the same cluster by each method. For example, when k is 5, the element of the consensus matrix $I_{x,y}$ equals the

sum of $\delta(R^l(c_x), R^l(c_y))$ in the five methods multiplying by the same weight $1/5$. Because this represents the probability of the occurrence of cell pairs in the same cluster, this strategy does not need to solve the problem that each cell achieves different cluster labels from the individual clustering methods.

$$I_{x,y} = \frac{1}{k} \sum_{l=1}^k \delta(R^l(c_x), R^l(c_y)) \quad (1)$$

$$\delta(X, Y) = \begin{cases} 0, & \text{if } X \neq Y \\ 1, & \text{if } X = Y, \end{cases} \quad (2)$$

where c_x and c_y are cell pairs in cells set C . k is the number of individual clustering methods. R^l is the clustering results in the l -th method.

Next, based on the assumption that the more consistent cluster labels predicted by all the clustering methods are more accurate, we design an importance score of the individual clustering methods. As ARI is a popular index for measuring the consensus of two clustering solutions, we use ARI to measure the importance of the individual clustering method. The importance score is defined as Equations (3) and (4). In Equations (3) and (4), ω_l denotes the importance of the l -th clustering method in all k methods. r_l represents the similarity between the l -th clustering method and other methods, which is calculated by averaging the ARI between predicted clusters in the l -th clustering method and the ones in each of the other methods.

$$\omega_l = \frac{r_l}{\sum_{j=1}^k r_j} \quad (3)$$

$$r_l = \frac{1}{k-1} \sum_{j=1, j \neq l}^k ARI(R^l, R^j), \quad (4)$$

where ω_l is the importance score of the l -th clustering method. r_l is the average of ARI between predicted clusters from the l -th method and other methods, and k is the number of individual clustering methods.

Then, the weighted consensus matrix is constructed by introducing the importance score of the individual clustering method to the original consensus matrix. The weighted consensus matrix $I_{x,y}'$ is defined as Equation (5). In Equation (5), the weighted consensus matrix $I_{x,y}'$ multiplies the importance score ω_l of the individual clustering methods, instead of the constant $1/k$ in the original consensus matrix.

$$I_{x,y}' = \sum_{l=1}^k \omega_l \times \delta(R^l(c_x), R^l(c_y)), \quad (5)$$

Finally, the HC method is performed to cluster cells on the weighted consensus matrix.

Evaluation Indices

We use two popular indices to evaluate the performance of clustering methods, including Normalized Mutual Information (NMI) (Estévez et al., 2009) and Adjusted Rand Index (ARI) (Hubert and Arabie, 1985). The two criteria are statistic-based indicators, showing the consensus of the predicted labels and the true ones in different views. NMI demonstrates the difference by calculating Mutual Information and Entropy between the two clustering solutions, with the range of values from 0 to 1. ARI presents the probability that a data pair will appear in the same cluster in the true clusters and the predicted clusters, with the range of values from -1 to 1 . The higher the NMI or ARI value obtained, the better performance the method has.

$$NMI(P, Q) = 2 \frac{I(P; Q)}{H(P) + H(Q)}, \quad (6)$$

where $I(P; Q)$ is the mutual information between P and Q . $H(P)$ and $H(Q)$ is the entropy of P and Q , respectively.

$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}{\frac{1}{2} \left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2} \right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2} \right] / \binom{n}{2}}, \quad (7)$$

where n is the number of cells. In the contingency table resulting from the overlap between true clusters and predicted ones, n_{ij} is the element in the i -th row and the j -th column, a_i is the summation of the elements in the i -th row, and b_j is the summation of the elements in the j -th column.

Datasets

We collected 12 published scRNA-seq datasets. Generally, they serve as gold standard datasets with true labels. They are available from Gene Expression Omnibus (GEO) and European Bioinformatics Institute (EMBL-EBI), respectively. These datasets have been normalized to various units, such as Transcripts Per Million reads (TPM), Fragments Per Kilobase of

TABLE 1 | The detail of scRNA-seq datasets.

Accessed ID	Datasets	Data unit	#Cells	#Genes	#Cell types	References
GSE38495	Ramskold	RPKM	33	21042	7	Ramsköld et al., 2012
GSE57249	Biase	FPKM	49	25384	3	Biase et al., 2014
GSE36552	Yan	RPKM	90	20214	6	Yan et al., 2013
E-MTAB-3321	Goolam	RPM	124	40315	5	Goolam et al., 2016
GSE70657	Grover	RPKM	135	15158	2	Grover et al., 2016
GSE70605	Liu	RPKM	145	18855	25	Liu et al., 2016
GSE51372	Ting	RPM	187	21583	7	Ting et al., 2014
GSE85908	Yeo	TPM	214	27473	4	Song et al., 2017
E-MTAB-2805	Pollen	TPM	249	6982	11	Pollen et al., 2014
GSE45719	Deng	RPKM	259	22147	10	Deng et al., 2014
GSE52529	Trapnell	FPKM	372	35988	4	Trapnell et al., 2014
GSE67835	Darmanis	CPM	466	22085	9	Darmanis et al., 2015

transcript per Million fragments mapped (FPKM), and Reads Per Kilobase per Million mapped reads (RPKM), etc. The details of the datasets are presented in **Table 1**.

EXPERIMENTS AND RESULTS

Implementation of the Five Clustering Methods

For optimal performance, we performed the five clustering methods with the default parameters in the references. The details of the parameters are described as follows.

For SNN-clq, the nearest neighbor parameter k is set to 3; the connectivity parameter of quasi-cliques r is set to 0.7; the threshold of the overlap of quasi-cliques m is set to 0.5.

For PhenoGraph, the surface marker expression data is normalized based on dividing by the maximum values. To construct the SNN graph, the nearest neighbor parameter k is set to 50.

For SC3, the log-transformed normalized $\log_2(x+1)$ is performed.

For SSNN-Louvain and MPGS-Louvain, SIMLR is performed with the default parameters in the initial similarity measurement step. The width parameter of the Gaussian kernel function σ is set to 1.0, 1.25, 1.5, 1.75, and 2. The nearest neighbor parameter k is set to 10, 12, 14... 30. (σ, k) pair resulting in 55 Gaussian kernels. In SSNN-Louvain, to construct the structural SNN graph, the nearest neighbor parameter k is set to $0.1n$ (n is the number of nodes). In MPGS-Louvain,

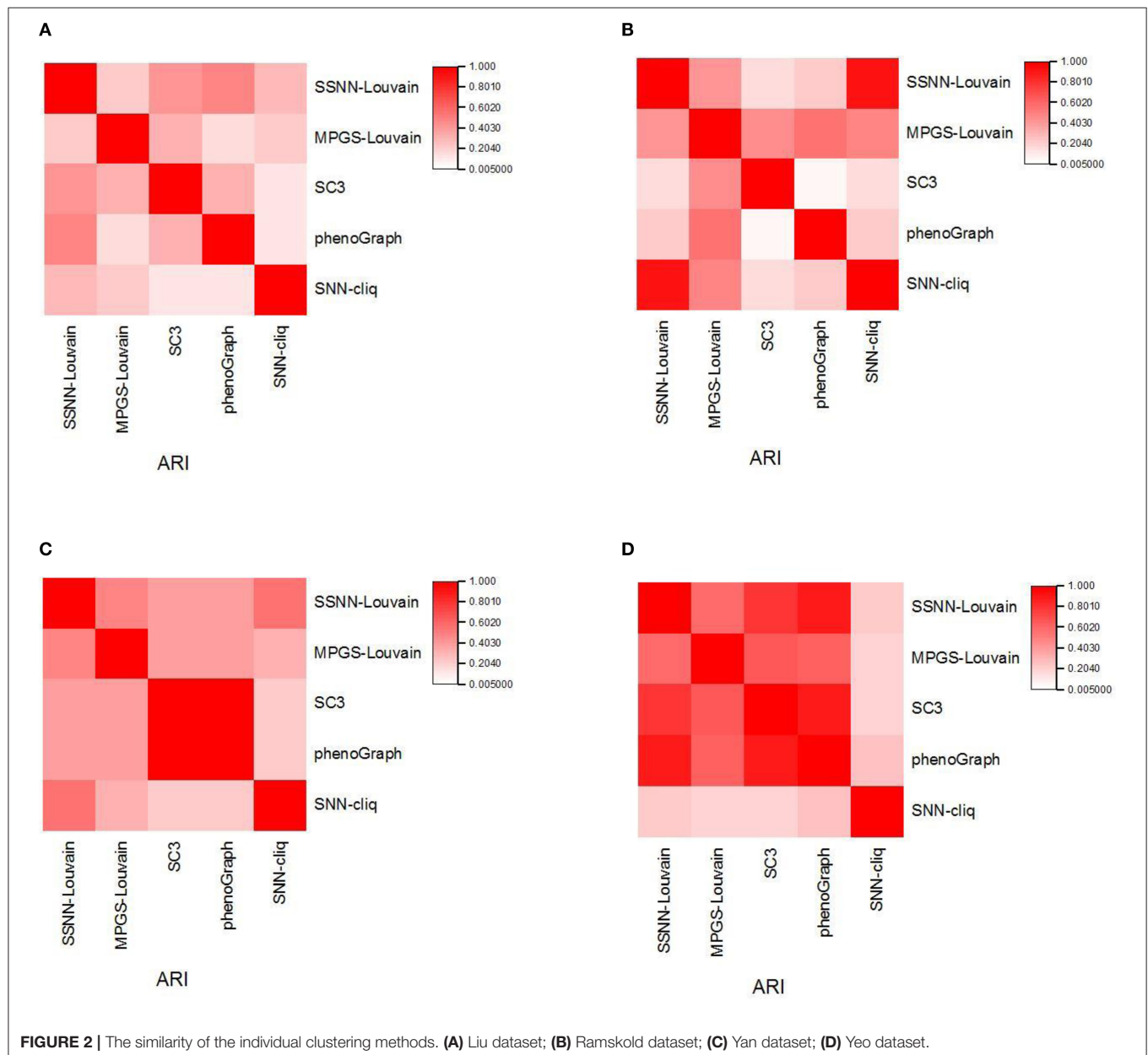


FIGURE 2 | The similarity of the individual clustering methods. (A) Liu dataset; (B) Ramskold dataset; (C) Yan dataset; (D) Yeo dataset.

the path length l is set to 2 for high performance and low time complexity.

Furthermore, in SNN-clq, PhenoGraph, SSNN-Louvain, and MPGS-Louvain, the number of categories can be automatically estimated by using quasi-clique partition or Louvain community detection, without a priori true categories.

Similarity Measurement of the Individual Clustering Methods

To analyze the difference of predicted results between the individual clustering methods, we calculate the ARI between the different clustering results and provide the consensus matrix heatmap. We select four scRNA-seq datasets: Ramskold, Yan, Yeo, and Liu, in which the Ramskold dataset is easy to partition

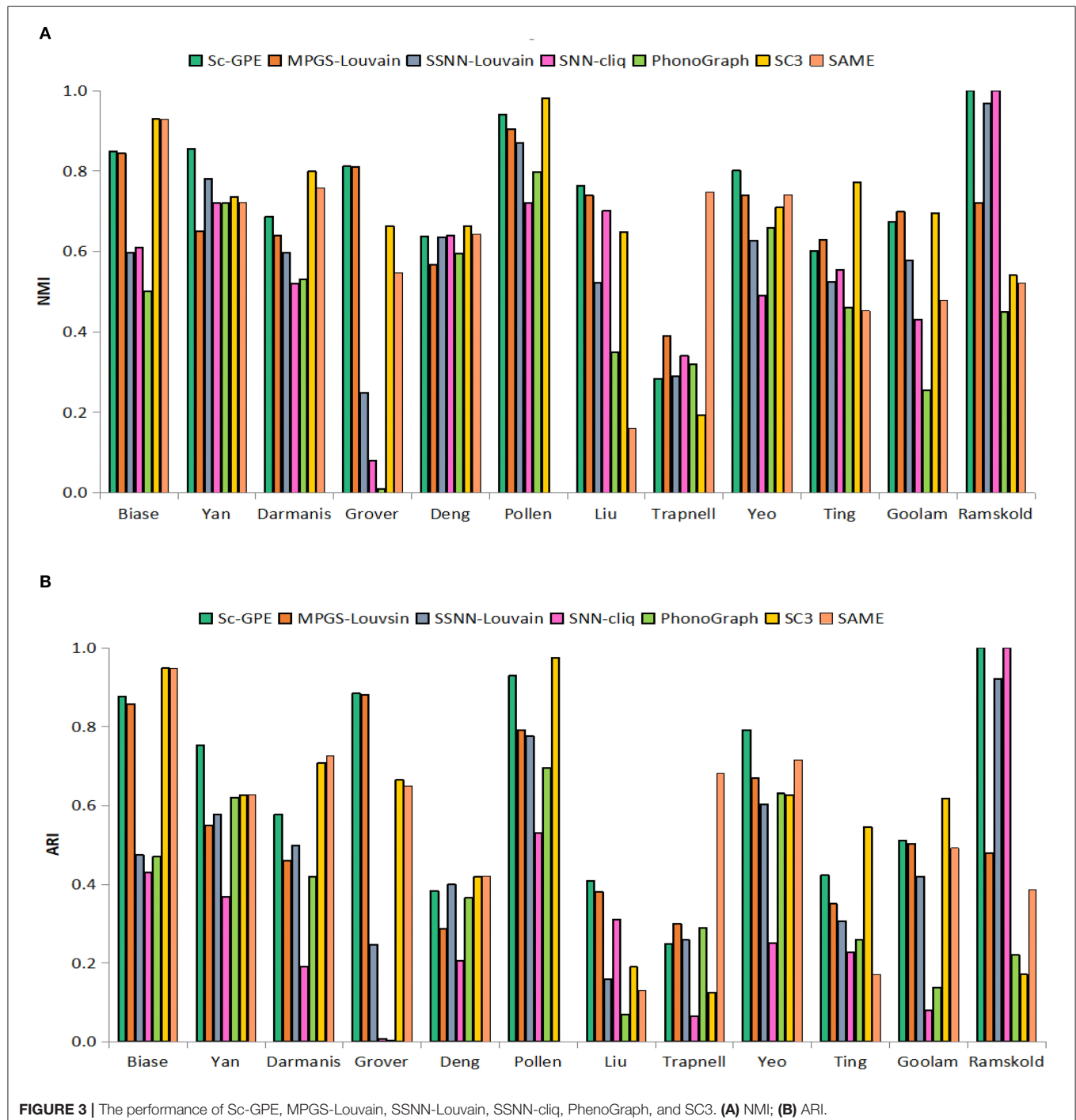


FIGURE 3 | The performance of Sc-GPE, MPGS-Louvain, SSNN-Louvain, SNN-clq, PhenoGraph, and SC3. **(A)** NMI; **(B)** ARI.

while the Liu dataset is hard to cluster. The first three datasets have a smaller number of true categories from four to seven, and the latter dataset has the true categories 25. The heatmaps are shown in **Figure 2**.

From **Figure 2**, it is observed that some faint similarity exists among the solutions of the individual clustering methods, which is consistent with the results from Yang et al. (2019). In different datasets, the similarities between the results of the individual clustering methods vary. For example, SSNN-Louvain shows relatively high similarity with SC3 and PhenoGraph on the Liu

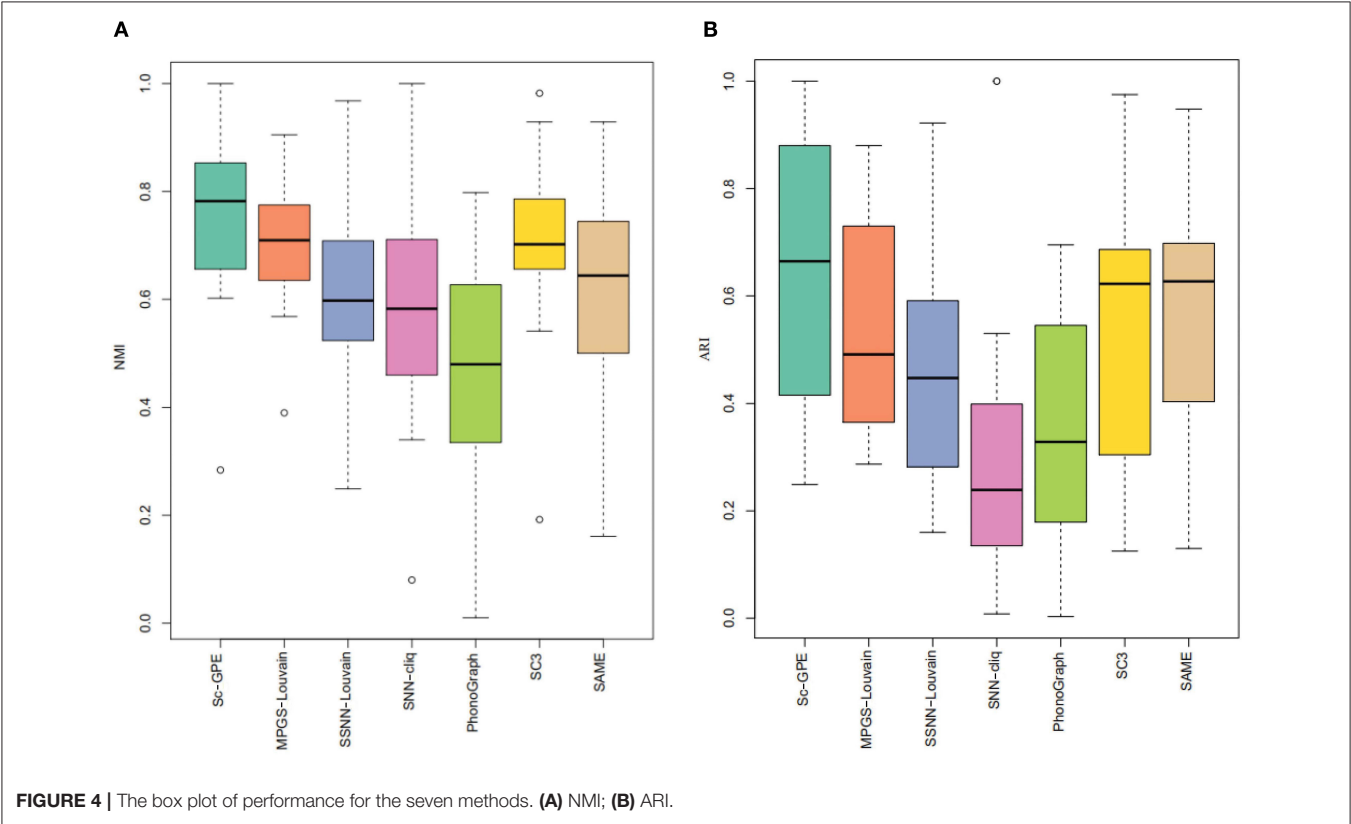
dataset. MPGS-Louvain shows a higher similarity than other clustering methods to the Ramskold dataset. SC3 is observed in the high similar to PhenoGraph on the Yan dataset. SNN-cliq shows a low similarity with other methods on the Yeo dataset. The difference between SC3 and PhenoGraph varies greatly in different datasets. The similarity between SC3 and PhenoGraph is close to one on the Yan and Yeo datasets, but the opposite results are achieved on the Liu and Ramskold datasets.

Furthermore, we can observe big differences between SNN-cliq and SC3, PhenoGraph on the four datasets. Therefore, we can

TABLE 2 | The comparison of the number of clusters from seven methods.

Datasets	Sc-GPE	MPGS-Louvain	SSNN-Louvain	SNN-cliq	PhenoGraph	SC3	SAME-clustering
Ramskold	7	3	8	7	2	2	2
Biase	3	3	4	6	2	3	3
Yan	6	6	8	18	3	3	3
Goolam	5	5	6	25	4	2	3
Grover	2	2	3	12	3	3	2
Liu	25	15	7	26	3	6	4
Ting	7	8	7	21	5	11	4
Yeo	4	5	3	28	3	5	3
Pollen	11	11	7	9	7	11	NA*
Deng	10	10	7	43	6	6	5
Trapnell	4	5	6	56	6	10	4
Darmanis	9	8	5	38	6	12	5

*SAME-Clustering method achieves NA on the Pollen dataset for that the clustering member Seurat in SAME-Clustering failed to run on this dataset.



find that different clustering methods would capture information about scRNA-seq data from different perspectives.

Comparisons With the Individual Clustering Methods and SAME-Clustering

To test the performance of our proposed Sc-GPE method, we compare it with both the five clustering methods and the state-of-the-art clustering ensemble algorithm SAME-clustering on 12 scRNA-seq datasets in terms of NMI and ARI. The results are shown in **Figure 3**. SAME-Clustering achieves the NA value of NMI and ARI on the Pollen dataset, because the clustering member Seurat in SAME-Clustering failed to run on this dataset.

From the experimental results, Sc-GPE achieves the highest average of NMI and ARI in all methods. Sc-GPE outperforms the six methods on five scRNA-seq datasets: Yan, Grover, Liu, Yeo, and Ramskold, while SC3 achieves the best performance on five scRNA-seq datasets: Biase, Deng, Pollen, Ting, and Goolam. The averages of NMI and ARI obtained by Sc-GPE are 6.92 and 17.79% higher than those of SC3, respectively. SAME-Clustering works best on three datasets: Biase, Darmanis, and Trapnell. The averages of NMI and ARI obtained by Sc-GPE are 21.84 and 20.19% higher than those of SAME-clustering, respectively. A large difference in clustering performance can be observed on the Grover, Liu, and Goolam datasets. The results show that Sc-GPE performs well and outperforms other methods.

Moreover, we compare the number of clusters in the seven methods, shown in **Table 2**. It can be observed that the number of predicted clusters has an obvious influence on the clustering solutions. For example, the clustering number of SNN-cliq and PhenoGraph is quite different from that of other methods, which is in consensus with their relatively poor performance on most datasets. SNN-cliq achieves the clustering numbers commonly more than the true categories except for the pollen dataset, PhenoGraph is just the opposite.

To further demonstrate the performance of Sc-GPE, we provide a box plot of the seven methods for 12 datasets, measured by NMI and ARI, shown in **Figure 4**. The box plot clearly shows that Sc-GPE outperforms the other six methods. The worse ARI value of 0.249 in Sc-GPE is from the Trapnell dataset, where some cells are misallocated resulting from two poor clustering solutions. SNN-cliq achieves the worst results in terms of ARI, and PhenoGraph performs worst on the NMI.

CONCLUSIONS

Currently, various single-cell clustering algorithms have been proposed with the advantage of accurately representing cell heterogeneity. However, there is a problem that the predicted cluster results from different clustering methods are quite different, which would limit the generalization capabilities. Combining the information from different cluster results would be a good resolution to improve the performance of clustering.

REFERENCES

Andrews, T. S., and Hemberg, M. (2019). M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* 35, 2865–2867. doi: 10.1093/bioinformatics/bty1044

So, we propose a novel cluster ensemble method Sc-GPE, which integrating five clustering methods: SNN-cliq, PhenoGraph, SSNN-Louvain, MPGS-Louvain, and SC3.

In Sc-GPE, a consensus matrix-based ensemble model is performed. It is a good statistics approach that can solve the problem of the different cluster labels generated in the individual clustering methods making it difficult to determine the correspondence cluster labels across all methods, which usually exists in the hypergraph-based cluster ensemble method. Furthermore, a weighted strategy is designed to measure the importance of individual clustering methods according to the similarity with other methods. A weighted consensus matrix is constructed based on the weighted strategy, which can distinguish the role of the individual clustering methods.

Sc-GPE provides close-to-the-best clustering solutions by combing the clustering methods that perform various similarity measurements and graph partitioning algorithms. The experimental results from twelve scRNA-seq datasets show that Sc-GPE outperforms the five individual clustering methods and state-of-the-art SAME-clustering method. However, the relatively small number of individual clustering methods may provide insufficient information and limit the performance of the Sc-GPE, and how to choose more optimal individual clustering methods should be researched in future work.

DATA AVAILABILITY STATEMENT

The datasets analyzed in this work are available in the following repositories: GEO: <https://xenabrowser.net/datapages/>; EMBL-EBI: <https://www.ebi.ac.uk/> and details of the datasets can be found in **Table 1**.

AUTHOR CONTRIBUTIONS

XZ and JW: conceptualization and design. XZ and H-DL: writing. H-DL and MX: data acquisition. XZ and JL: methodology. All authors: contributed to the article and approved the submitted version.

FUNDING

This research was supported by the National Natural Science Foundation of China (Nos: 61762087, 61702555, 61662028, and 61772557), Hunan Provincial Science and Technology Program (No. 2018WK4001), 111 Project (No. B18059), and Natural Science Foundation of Guangxi Province (No. 2018JJA170175).

ACKNOWLEDGMENTS

This paper is recommended by the 5th CCF Bioinformatics Conference.

Becht, E., McInnes, L., Healy, J., Dutertre, C., Kwok, I. W. H., Ng, L. G., et al. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* 37, 38–44. doi: 10.1038/nbt.4314

Biase, F. H., Cao, X., and Zhong, S. (2014). Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell

- RNA sequencing. *Genome Res.* 24, 1787–1796. doi: 10.1101/gr.177725.114
- Darmanis, S., Sloan, S. A., Zhang, Y., Enge, M., Caneda, C., Shuer, L. M., et al. (2015). A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl. Acad. Sci. U.S.A.* 112, 7285–7290. doi: 10.1073/pnas.1507125112
- Deng, Q., Ramsköld, D., Reinis, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* 343, 193–196. doi: 10.1126/science.1245316
- Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. (2009). Normalized mutual information feature selection. *IEEE Trans. Neural Netw.* 20, 189–201. doi: 10.1109/TNN.2008.2005601
- Goolam, M., Scialdone, A., Graham, S. J., Macaulay, I. C., Jedrusik, A., Hupalowska, A., et al. (2016). Heterogeneity in Oct4 and Sox2 targets biases cell fate in 4-cell mouse embryos. *Cell* 165, 61–74. doi: 10.1016/j.cell.2016.01.047
- Grover, A., Sanjua, Pla, A., Thongjuea, S., Carrelha, J., Giustacchini, A., Gambardella, A., et al. (2016). Single-cell RNA sequencing reveals molecular and functional platelet bias of aged haematopoietic stem cells. *Nat. Commun.* 7:11075. doi: 10.1038/ncomms11075
- Grun, D. (2020). Revealing dynamics of gene expression variability in cell state space. *Nat. Methods* 17, 45–49. doi: 10.1038/s41592-019-0632-3
- Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., et al. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature* 525, 251–255. doi: 10.1038/nature14966
- Hubert, L., and Arabie, P. (1985). Comparing partitions. *J. Classif.* 2, 193–218. doi: 10.1007/BF01908075
- Huh, R., Yang, Y., Jiang, Y., Shen, Y., and Li, Y. (2020). SAME-clustering: Single-cell Aggregated clustering via Mixture Model Ensemble. *Nucleic Acids Res.* 48, 86–95. doi: 10.1093/nar/gkz959
- Kim, T., Chen, I. R., Lin, Y., Wang, A. Y., Yang, J. Y. H., and Yang, P. (2019). Impact of similarity metrics on single-cell RNA-seq data clustering. *Brief. Bioinform.* 20, 2316–2326. doi: 10.1093/bib/bby076
- Kiselev, V. Y., Andrews, T. S., and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* 20, 273–282. doi: 10.1038/s41576-018-0088-9
- Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T. S., Yiu, A., Chandra, T., et al. (2017). SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* 14, 483–486. doi: 10.1038/nmeth.4236
- Kuncheva, L. I., and Vetrov, D. P. (2006). Evaluation of stability of k-means cluster ensembles with respect to random initialization. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 1798–1808. doi: 10.1109/TPAMI.2006.226
- Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D., et al. (2015). Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* 162, 184–197. doi: 10.1016/j.cell.2015.05.047
- Liu, W., Liu, X., Wang, C., Gao, Y., Gao, R., Kou, X., et al. (2016). Identification of key factors conquering developmental arrest of somatic cell cloned embryos by combining embryo biopsy and single-cell sequencing. *Cell Discov.* 2, 1–15. doi: 10.1038/celldisc.2016.10
- Liu, Z., Liu, F., Hong, C., Gao, M., Chen, Y., Liu, S., et al. (2019). “Detection of cell types from single-cell RNA-seq data using similarity via kernel preserving learning embedding,” in *2019 IEEE International Conference on Bioinformatics and Biomedicine* (San Diego, CA: IEEE). doi: 10.1109/BIBM47256.2019.8983395
- Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., et al. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* 32, 1053–1058. doi: 10.1038/nbt.2967
- Qi, R., Ma, A., Ma, Q., and Zou, Q. (2019). Clustering and classification methods for single-cell RNA-sequencing data. *Brief. Bioinform.* 21, 1196–1208. doi: 10.1093/bib/bbz062
- Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., et al. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* 30, 777–782. doi: 10.1038/nbt.2282
- Soneson, C., and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* 15, 255–261. doi: 10.1038/nmeth.4612
- Song, Y., Botvinnik, O. B., Lovci, M. T., Kakaradov, B., Liu, P., Xu, J. L., et al. (2017). Single-cell alternative splicing analysis with expedition reveals splicing dynamics during neuron differentiation. *Mol. Cell* 67, 148–161. doi: 10.1016/j.molcel.2017.06.003
- Stuart, T., and Satija, R. (2019). Integrative single-cell analysis. *Nat. Rev. Genet.* 20, 257–272. doi: 10.1038/s41576-019-0093-7
- Ting, D. T., Wittner, B. S., Ligorio, M., Jordan, N. V., Shah, A. M., Miyamoto, D. T., et al. (2014). Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* 8, 1905–1918. doi: 10.1016/j.celrep.2014.08.029
- Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., et al. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* 32, 381–386. doi: 10.1038/nbt.2859
- Vega-Pons, S., and Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. *Int. J. Pattern Recogn. Artif. Intell.* 25, 337–372. doi: 10.1142/S0218001411008683
- Vitak, S. A., Torkenczy, K. A., Rosenkrantz, J. L., Fields, A. J., Christiansen, L., Wong, M. H., et al. (2017). Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat. Methods* 14, 302–308. doi: 10.1038/nmeth.4154
- Wang, B., Pourshafeie, A., Zitnik, M., Zhu, J., Bustamante, C., Batzoglou, S., et al. (2018). Network enhancement as a general method to denoise weighted biological networks. *Nat. Commun.* 9:3108. doi: 10.1038/s41467-018-05469-x
- Wang, D., and Gu, J. (2018). VASC: dimension reduction and visualization of single-cell RNA-seq data by deep variational autoencoder. *Genom. Proteom. Bioinform.* 16, 320–331. doi: 10.1016/j.gpb.2018.08.003
- Xu, C., and Su, Z. (2015). Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* 31, 1974–1980. doi: 10.1093/bioinformatics/btv088
- Xu, Y., Li, H.-D., Pan, Y., Luo, F., and Wang, J. (2019). “BioRank: a similarity assessment method for single cell clustering,” in *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (Madrid), 157–162. doi: 10.1109/TCBB.2019.2931582
- Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., et al. (2013). Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* 20, 1131–9. doi: 10.1038/nsmb.2660
- Yang, Y., Huh, R., Culpepper, H. W., Lin, Y., Love, M. I., and Li, Y. (2019). SAFE-clustering: single-cell aggregated (from ensemble) clustering for single-cell RNA-seq data. *Bioinformatics* 35, 1269–1277. doi: 10.1093/bioinformatics/bty793
- Yip, S. H., Sham, P. C., and Wang, J. (2019). Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief. Bioinform.* 20, 1583–1589. doi: 10.1093/bib/bby011
- Yu, Z., Luo, P., You, J., Wong, H.-S., Leung, H., Wu, S., et al. (2016). Incremental semi-supervised clustering ensemble for high dimensional data clustering. *IEEE Trans. Knowl. Data Eng.* 28, 701–714. doi: 10.1109/TKDE.2015.2499200
- Zhu, X., Guo, L., Xu, Y., Li, H., Liao, X., Wu, F., et al. (2019c). A global similarity learning for clustering of single-cell RNA-seq data. *2019 IEEE International Conference on Bioinformatics and Biomedicine* (San Diego, CA: IEEE). doi: 10.1109/BIBM47256.2019.8983200
- Zhu, X., Li, H.-D., Guo, L., Wu, F.-X., and Wang, J. (2019a). Analysis of single-cell RNA-seq data by clustering approaches. *Curr. Bioinf.* 14, 314–322. doi: 10.2174/1574893614666181120095038
- Zhu, X., Li, H.-D., Xu, Y., Guo, L., Wu, F.-X., Duan, G., and Wang, J. (2019b). A hybrid clustering algorithm for identifying cell types from single-cell RNA-Seq data. *Genes* 10:98. doi: 10.3390/genes10020098
- Zhu, X., Zhang, J., Xu, Y., Wang, J., Peng, X., and Li, H. (2020). Single-cell clustering based on shared nearest neighbor and graph partitioning. *Interdiscip. Sci. Computat. Life Sci.* 12, 117–130. doi: 10.1007/s12539-019-00357-4

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.

Copyright © 2020 Zhu, Li, Li, Xie and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



An Efficient and Easy-to-Use Network-Based Integrative Method of Multi-Omics Data for Cancer Genes Discovery

Ting Wei^{1,2}, Botao Fa^{1,2}, Chengwen Luo^{1,2}, Luke Johnston², Yue Zhang^{1,2} and Zhangsheng Yu^{1,2*}

¹ Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai, China, ² SJTU-Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University, Shanghai, China

OPEN ACCESS

Edited by:

Chunhou Zheng,
Anhui University, China

Reviewed by:

Junfeng Xia,
Anhui University, China
Zhu-Hong You,
Xinjiang Technical Institute of Physics
and Chemistry, Chinese Academy
of Sciences (CAS), China
Jin-Xing Liu,
Qufu Normal University, China

*Correspondence:

Zhangsheng Yu
yuzhangsheng@sjtu.edu.cn

Specialty section:

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

Received: 01 October 2020

Accepted: 25 November 2020

Published: 08 January 2021

Citation:

Wei T, Fa B, Luo C, Johnston L,
Zhang Y and Yu Z (2021) An Efficient
and Easy-to-Use Network-Based
Integrative Method of Multi-Omics
Data for Cancer Genes Discovery.
Front. Genet. 11:613033.
doi: 10.3389/fgene.2020.613033

Identifying personalized driver genes is essential for discovering critical biomarkers and developing effective personalized therapies of cancers. However, few methods consider weights for different types of mutations and efficiently distinguish driver genes over a larger number of passenger genes. We propose MinNetRank (Minimum used for Network-based Ranking), a new method for prioritizing cancer genes that sets weights for different types of mutations, considers the incoming and outgoing degree of interaction network simultaneously, and uses minimum strategy to integrate multi-omics data. MinNetRank prioritizes cancer genes among multi-omics data for each sample. The sample-specific rankings of genes are then integrated into a population-level ranking. When evaluating the accuracy and robustness of prioritizing driver genes, our method almost always significantly outperforms other methods in terms of precision, F1 score, and partial area under the curve (AUC) on six cancer datasets. Importantly, MinNetRank is efficient in discovering novel driver genes. SP1 is selected as a candidate driver gene only by our method (ranked top three), and SP1 RNA and protein differential expression between tumor and normal samples are statistically significant in liver hepatocellular carcinoma. The top seven genes stratify patients into two subtypes exhibiting statistically significant survival differences in five cancer types. These top seven genes are associated with overall survival, as illustrated by previous researchers. MinNetRank can be very useful for identifying cancer driver genes, and these biologically relevant marker genes are associated with clinical outcome. The R package of MinNetRank is available at <https://github.com/weitinging/MinNetRank>.

Keywords: multi-omics, network-based methods, cancer gene prediction, driver genes, tumor stratification

INTRODUCTION

Rapid technological advances in high-throughput sequencing have driven the development of omics field. Omics data types include genomics, transcriptomics, proteomics, epigenomics, and metabolomics (Hasin et al., 2017). However, a single type of “omics” only provides limited insights into the biological mechanisms of diseases. Additionally, the different omics data events

are somewhat interdependent. An integrative study of multi-omics data contributes to a holistic understanding of the molecular function (Sun and Hu, 2016). An essential question in cancer genomics is distinguishing driver genes, which are causally implicated in oncogenesis, from biologically neutral passenger genes that are immaterial to neoplasia (Greenman et al., 2007). Passenger mutations can become driver mutations (and vice versa) under changing environmental conditions and selection pressures, increasing the complexity of intratumor heterogeneity (Yap et al., 2012). Accumulating evidence suggests that identifying personalized driver genes is essential for the development of effective personalized therapies and realizing the goals of precision medicine (Dagogo-Jack and Shaw, 2018). A critical but challenging step is to incorporate different omics data in a meaningful and efficient way to discover cancer driver genes and elucidate potential causative changes of cancer (Huang et al., 2017). The main approaches for distinguishing driver genes from passenger genes can be divided into frequency-based methods and network-based approaches.

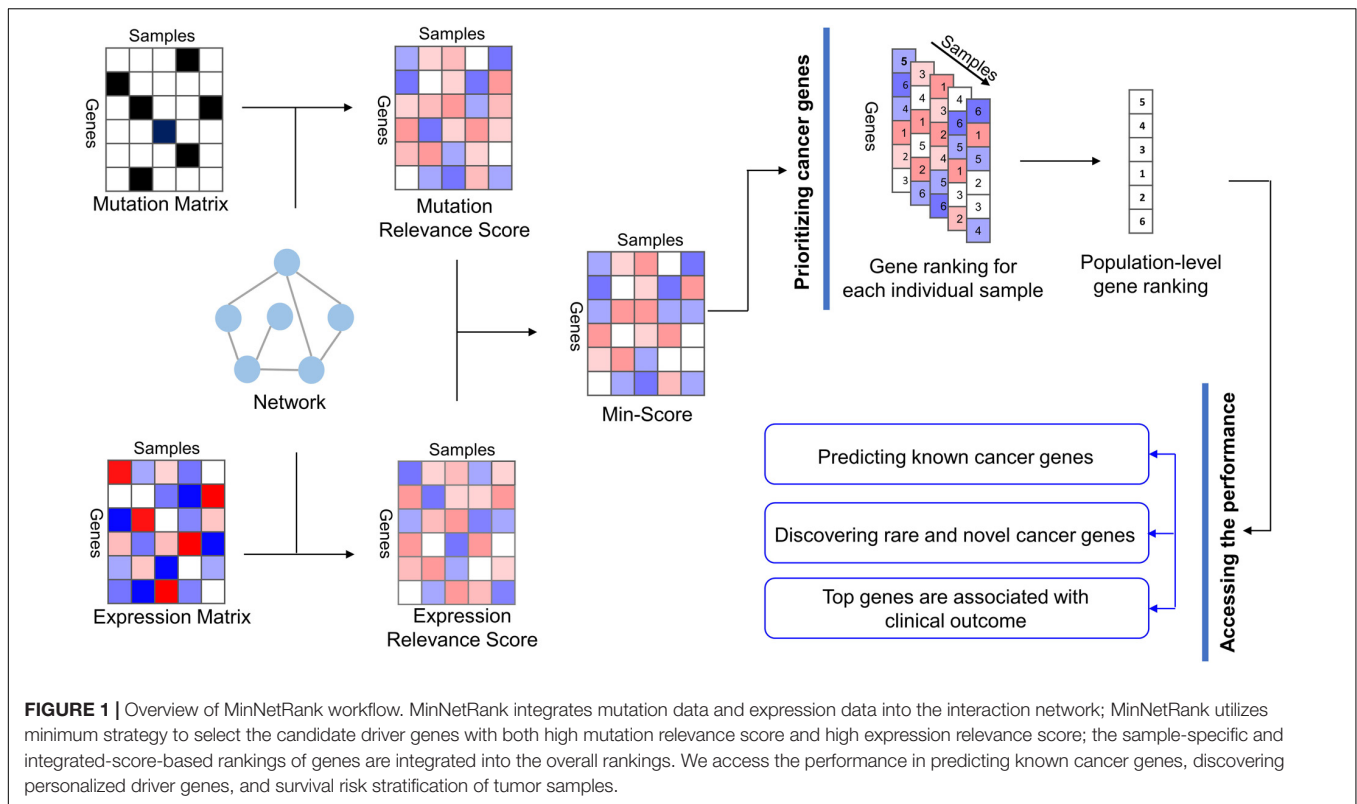
Frequency-based methods estimate the background mutation rate (BMR) representing the rate of random passenger mutations and identify driver genes that harbor significantly more somatic mutations than BMR (Ding et al., 2008; Pon and Marra, 2015). However, accurately estimating BMR is difficult because of the variability among cancer types, among samples of the same cancer type, and between genomes (Pon and Marra, 2015). Subsequent frequency-based methods, such as MuSiC and MutSigCV, have been developed to correct for one or more of these factors (Dees et al., 2012; Lawrence et al., 2013). Somatic mutations are characterized by a small number of frequently mutated genes and many infrequently mutated genes. Moreover, more than 99.9% of the somatic mutations in tumors are passengers (Vogelstein et al., 2013). It is challenging to identify infrequent or rare driver genes by methods based only on mutation frequency.

Network-based approaches have emerged as promising and powerful methods to detect low-frequency and high-frequency mutated driver genes due to their ability to model gene interactions. For network-based approaches, nodes representing genes and edges are links between two genes if there is an interaction between them (Huang et al., 2017). Network-based methods have been successfully applied to many biomedical fields, such as the discovery of mutation subnetwork (Vandin et al., 2011), prediction of drug–target interaction, and cancer gene prioritization (Bashashati et al., 2012; Chen et al., 2012; Yu et al., 2013). HotNet2 uses a network diffusion model to simultaneously assess the frequency of somatic mutation and the local topology of the interaction network and detects significantly mutated subnetworks (Leiserson et al., 2015). Mutations for Functional Impact on Network Neighbors (MUFFIN) is a method for prioritizing cancer genes accounting for mutation frequency of genes and their direct neighbors in functional network (Cho et al., 2016). Both HotNet2 and MUFFIN use mutation data only without integrating other omics data. DawnRank is a single patient approach to rank potential driver genes based on their impact on downstream differential expression genes in the interaction network (Hou and Ma,

2014). NetICS predicts mediator genes affected by proximal upstream-located aberrant genes and proximal downstream-located differentially expressed genes (Dimitrakopoulos et al., 2018). Both DawnRank and NetICS consider only incoming degree or outgoing degree of interaction network for single omics. For example, DawnRank only considers incoming degree for expression data. It is desirable to use incoming and outgoing degree simultaneously. Driver_IRW (Driver genes discovery with Improved Random Walk method) assigns different transition probabilities for different genes of the interaction network (Wei et al., 2020). DeepDriver predicts cancer driver genes based on mutation-based features and gene similarity networks using deep convolutional neural networks (Luo et al., 2019). None of these methods consider the different weights for the different types of mutations; however, the weighting method is essential for sample-specific study. Furthermore, none of these methods investigate the relationship between the top rankings of genes and overall survival. Therefore, we develop a more meaningful and efficient method that considers different weight coefficients for the various types of mutations, simultaneously considers the incoming and outgoing degree of interaction network for single omics, and uses minimum strategy to integrate multi-omics data.

We present a new method called MinNetRank that uses minimum strategy among multi-omics data to prioritize cancer genes (**Figure 1**). The main steps of MinNetRank include (1) single-omics data analysis: calculating mutation relevance scores and expression relevance scores of genes for each sample using network diffusion based on incoming and outgoing degree. We further consider different weight coefficients for the different types of mutations and propose Weighted_MinNetRank. (2) The integration of multi-omics data: calculating the minimum value of mutation relevance score and expression relevance score as an integrated score for each gene in each sample. A higher minimum value reflects a higher mutation relevance score and expression relevance score simultaneously; (3) prioritizing driver genes: aggregating the sample-specific and integrated-score-based rankings of genes into a robust population-level gene ranking.

We apply Weighted_MinNetRank and MinNetRank to analyze five The Cancer Genome Atlas (TCGA) datasets (hepatocellular carcinoma, stomach adenocarcinoma, bladder urothelial carcinoma, lung adenocarcinoma, and skin cutaneous melanoma) and one International Cancer Genome Consortium (ICGC) dataset (hepatocellular carcinoma). We select the top 50 genes of population-level ranking as candidate driver genes. We systematically examine the performance of Weighted_MinNetRank and MinNetRank from three aspects. Firstly, Weighted_MinNetRank and MinNetRank outperform other methods [Mean, Maximum, DawnRank, NetICS, and a commonly used frequency-based method (Freq)] in terms of precision, F1 score, and partial area under the curve (AUC) value of selecting cancer driver genes. Secondly, Weighted_MinNetRank and MinNetRank detect rare and novel candidate driver genes (e.g., SP1 in hepatocellular carcinoma). Finally, the top seven genes can be used as prognostic biomarkers for risk stratification. The survival difference between two



subtypes (low-risk and high-risk groups) is statistically significant in all six datasets.

RESULTS

We propose a new method (MinNetRank) that uses minimum strategy among multi-omics data to prioritize cancer genes. For comparison, we also add the performance of mean (Mean) and maximum (Maximum) to integrate the mutation data and expression data. All mutations have the same weight for MinNetRank. We further consider different weight coefficients for the different types of mutations (Weighted_MinNetRank). In this study, Weighted_MinNetRank and MinNetRank are compared with other five methods [Mean, Maximum, DawnRank (Hou and Ma, 2014), NetICS (Dimitrakopoulos et al., 2018), and Freq] on five types of cancer (liver hepatocellular carcinoma, stomach adenocarcinoma, lung adenocarcinoma, bladder urothelial carcinoma, and skin cutaneous melanoma). Freq is a simple and common method based only on mutation frequency, which compares the mutation frequency of genes in tumor patient (Dimitrakopoulos et al., 2018; Guo et al., 2018). Weighted_MinNetRank and MinNetRank are an efficient and easy-to-use network-based method for cancer genes discovery by integrating multi-omics data, as shown in the subsequent results.

Overview of MinNetRank

The schematic in **Figure 1** illustrates the three-step procedure of our new method MinNetRank. MinNetRank requires three input

files: gene mutations, gene expression for tumor and normal samples, and the interaction network.

Step 1: calculating mutation relevance score and expression relevance score using RWR (Random Walker with Restart) algorithm. The $n \times m$ matrix S^M is the gene mutation status for each sample, where n is the number of genes, and m is the number of samples. $S_{ik}^M = 1$ if gene i is mutated in sample k and $S_{ik}^M = 0$ otherwise. We further consider different weight coefficients for the different types of mutations and supplement a new method (Weighted_MinNetRank). We normalize each column of S^M by $S^M / \text{colSum}(S^M)$. We define the $n \times m$ mutation relevance score matrix W^M as multiplication between diffused matrix D and S^M :

$$W^M = DS^M. \quad (1)$$

The D_{ij} reflects the connectivity between gene i and gene j , and S_{ik}^M reflects the mutation status of gene i in sample k . The product W_{ik}^M is gene i 's mutation relevance score in sample k , defined as the proximity of gene i to mutation genes.

Similarly, the $n \times m$ matrix S^E is RNA differential expression score (Absolute value of Log2 Fold-Change, $ALFC$) for each sample. We define the expression relevance score matrix W^E as,

$$W^E = DS^E. \quad (2)$$

Step 2: minimum value of mutation relevance score and expression relevance score. To integrate multi-omics data (gene mutation and expression data), the mutation relevance score and

expression relevance score are combined to produce a gene min-score for each sample. The min-score is the minimum value of W_{ik}^M and W_{ik}^E :

$$W = \text{pmin}(W^M, W^E). \quad (3)$$

pmin is R function and returns the minimum of the corresponding elements of the two input vectors. W_{ik} is the minimum value of W_{ik}^M and W_{ik}^E ($i \in 1 \cdots n$, $k \in 1 \cdots m$), where n is the number of genes, and m is the number of samples. The high score of W_{ik} means that gene i is proximal to many mutation genes and differentially expressed genes for each k . The minimum value is a meaningful and efficient way to integrate multi-omics data for the following two reasons:

Firstly, the minimum strategy reduces extreme values that may be potential outliers in highly skewed distributions. The probability distribution of W_{*k}^M (the mutation relevance scores for genes in sample k) and W_{*k}^E (the expression relevance scores for genes in sample k) is a positively skewed distribution. This means that some genes have extremely high scores. These high scores may be due to the technical noise of high-throughput sequencing and the incomplete interaction network. For example, as shown in **Figure 2**, sample TCGA-BC-A10X has three mutated genes in TCGA-LIHC, and only one gene (*OR2C3*) of these is in the interaction network. The *OR2C3* mutation relevance score in TCGA-BC-A10X is evidently high ($W_{ik}^M = 0.48$, $i = \text{OR2C3}$ and $k = \text{TCGA-BC-A10X}$) and is ranked 1st. Meanwhile, the *OR2C3* expression relevance score in TCGA-BC-A10X is 3.24-06 and is ranked 8, 221st. Henceforth, the high mutation relevance score needs to be cautiously processed. Lastly, the min-score of *OR2C3* mutation relevance score and expression relevance score is ranked 1, 943rd. *OR2C3* is an olfactory receptor protein and probably is not a potential driver gene (Malnic et al., 2004; Riessland et al., 2017).

Secondly, the minimum (“double high”) strategy is necessary to prioritize cancer genes having a higher biological relevance. If one gene has a relatively high mutation relevance score but low expression relevance score (such as *OR2C3* in TCGA-BC-A10X), this gene may not be a potential driver gene since differential gene expression is the downstream events of DNA mutation (Sager, 1997). In the other case, the *SI* expression relevance score in TCGA-DD-AAE2 is ranked 8th ($W_{ik}^E = 0.0012$, $i = \text{SI}$, and $k = \text{TCGA-DD-AAE2}$), and the mutation relevance score is ranked last. Only *MGAM* interacts with *SI* in the interaction network, and TCGA-DD-AAE2 has no *SI* or *MGAM* mutation. We hope the candidate driver genes have a high mutation relevance score and high expression relevance score.

MinNetRank used a minimum strategy to integrate multi-omics data (mutation data and expression data). We further investigated which data have the greatest effect on the minimum score. We calculated the proportion of mutation relevance score and expression relevance score in minimum scores for the top 50 candidate cancer genes. The proportion of mutation relevance score was 0.657 in all six datasets, and expression relevance score was 0.347. Mutation relevance score affected the minimum score more.

Step 3: integrating sample-specific rankings of genes into a population-level ranking. We transform the min-scores into

rankings, since min-scores indicate the relative importance of each sample's genes. To integrate the sample-specific rankings of genes into a robust population-level ranking, we calculate the sum of per-sample ranking. Each step of MinNetRank is based on single sample analysis, such as using the per-sample network diffusion, calculating the minimum value of mutation relevance score and expression relevance score for each gene in each sample, and transforming min-scores into rankings for each sample. We calculate the sum of per-sample ranking as the population-level ranking.

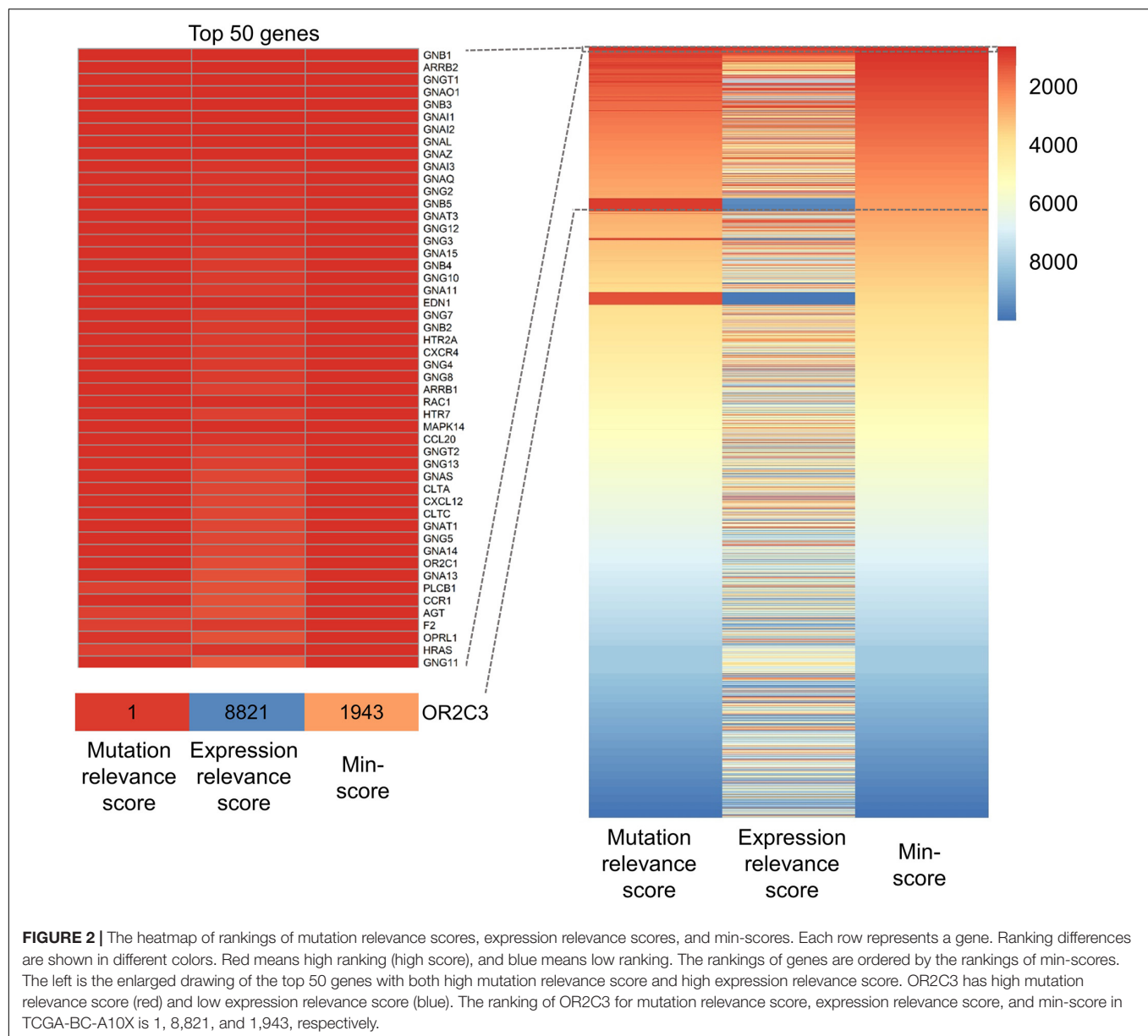
To perform a systematic comparison of seven methods (Weighted_MinNetRank, MinNetRank, Mean, Maximum, DawnRank, NetICS, and Freq), the 576 genes annotated in cancer gene census (CGC) are used as the gold standard cancer driver gene set, and the genes not in CGC are the negative set. The evaluation metrics (precision, F1 score, and partial AUC value) are based on the top 50 genes of six different datasets (five TCGA datasets and one ICGC dataset). The five TCGA datasets are regarding hepatocellular carcinoma (TCGA-LIHC), stomach adenocarcinoma (TCGA-STAD), bladder urothelial carcinoma (TCGA-BLCA), lung adenocarcinoma (TCGA-LUAD), and skin cutaneous melanoma (TCGA-SKCM), respectively. The one ICGC dataset includes hepatocellular carcinoma data from LIRI-JP (Liver Cancer-RIKEN, JP) project (LIRI-LIHC) (Fa et al., 2019). Skin cutaneous melanoma, lung adenocarcinoma, bladder urothelial carcinoma, and stomach adenocarcinoma have a high mutation burden (Martincorena and Campbell, 2015), and LIHC has two different datasets. Both are common cancer types and pose increasing public concerns. The detailed descriptions of six datasets are provided in **Table 1**. The somatic mutations include non-synonymous simple nucleotide variation (SNV) and insertions and deletions (InDels) in coding regions.

MinNetRank Accurately Predicted Cancer Gene

In general, considering the weights for the different types of mutations (Weighted_MinNetRank) had a better performance than other six methods (MinNetRank, Mean, Maximum, NetICS, DawnRank, and Freq) in all six cancer datasets (TCGA-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, TCGA-SKCM, and LIRI-LIHC). Weighting for the different types of mutations was essential for a personalized analysis. As shown in **Figure 3** (for datasets TCGA-LIHC and LIRI-LIHC), **Supplementary Figure 1** (for datasets TCGA-STAD and TCGA-BLCA), and **Supplementary Figure 2** (for datasets TCGA-LUAD and TCGA-SKCM), Weighted_MinNetRank and MinNetRank achieved a higher precision, F1 score, and AUC in all six datasets, namely, Weighted_MinNetRank and MinNetRank could rank the known gold standard cancer driver genes higher. The AUC of Freq was not calculated as the mutation frequency for some genes were the same.

MinNetRank Robustly Predicted Cancer Gene

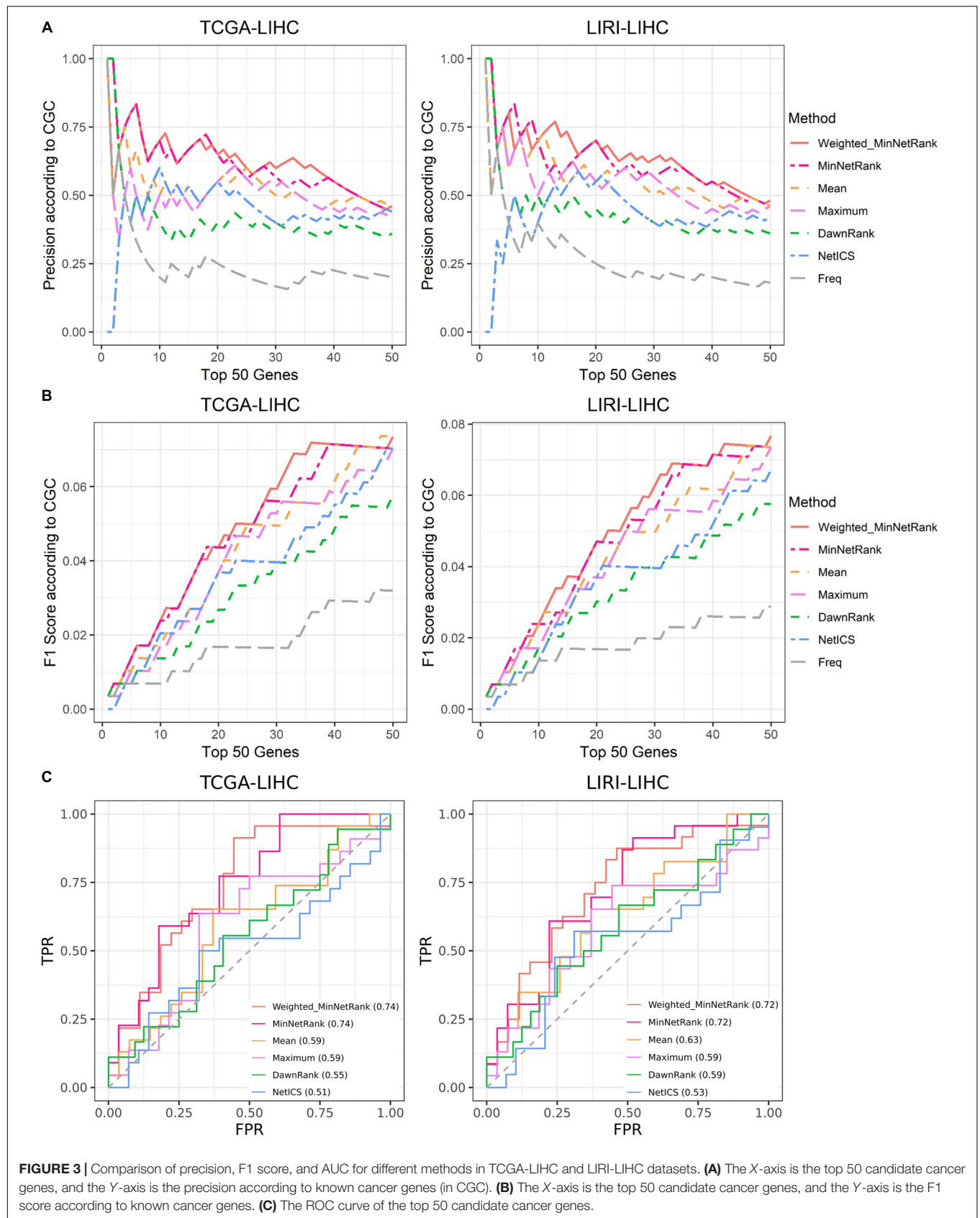
The Weighted_MinNetRank and MinNetRank also had the advantage of obtaining robust and stable results using the subset



of samples with different sample sizes. We calculated the mean and standard deviation (SD) of the precision values P (mean precision of the top 50 genes), F1 scores, and partial AUC values after 10 runs. The precision value was proportional to the area under the precision curve (**Figure 3A**). All six methods used the same subset of samples, and the subset of samples was randomly selected from all samples by R. Using the same subset of samples, we compared the results of six methods. The mean of the precision, F1 score, and partial AUC for Weighted_MinNetRank and MinNetRank was higher than other methods, and the SD was smaller [**Figure 4** (for datasets TCGA-LIHC and LIRI-LIHC), **Supplementary Figure 3** (for datasets TCGA-STAD and TCGA-BLCA), and **Supplementary Figure 4** (for datasets TCGA-LUAD and TCGA-SKCM)]. The performance in all six datasets and different sample sizes showed the robustness of our method.

Furthermore, Weighted_MinNetRank and MinNetRank still performed well, even with a smaller number of samples.

In order to evaluate the contribution of each part of Weighted_MinNetRank and MinNetRank (calculating the relevance score using both incoming and outgoing degree of the interaction network for single omics, using minimum strategy to integrate multi-omics data, and the different weighted methods), we calculated the precision, F1 score, and partial AUC value of the top 50 candidate cancer genes. We also added network metrics (degree centrality, betweenness centrality, and the mean of degree and betweenness centrality). We needed to calculate the baselines of the network only once, and the results were the same for all datasets. As shown in **Table 2**, Weighted_MinNetRank had a better performance than all other methods in terms of precision, F1 score,



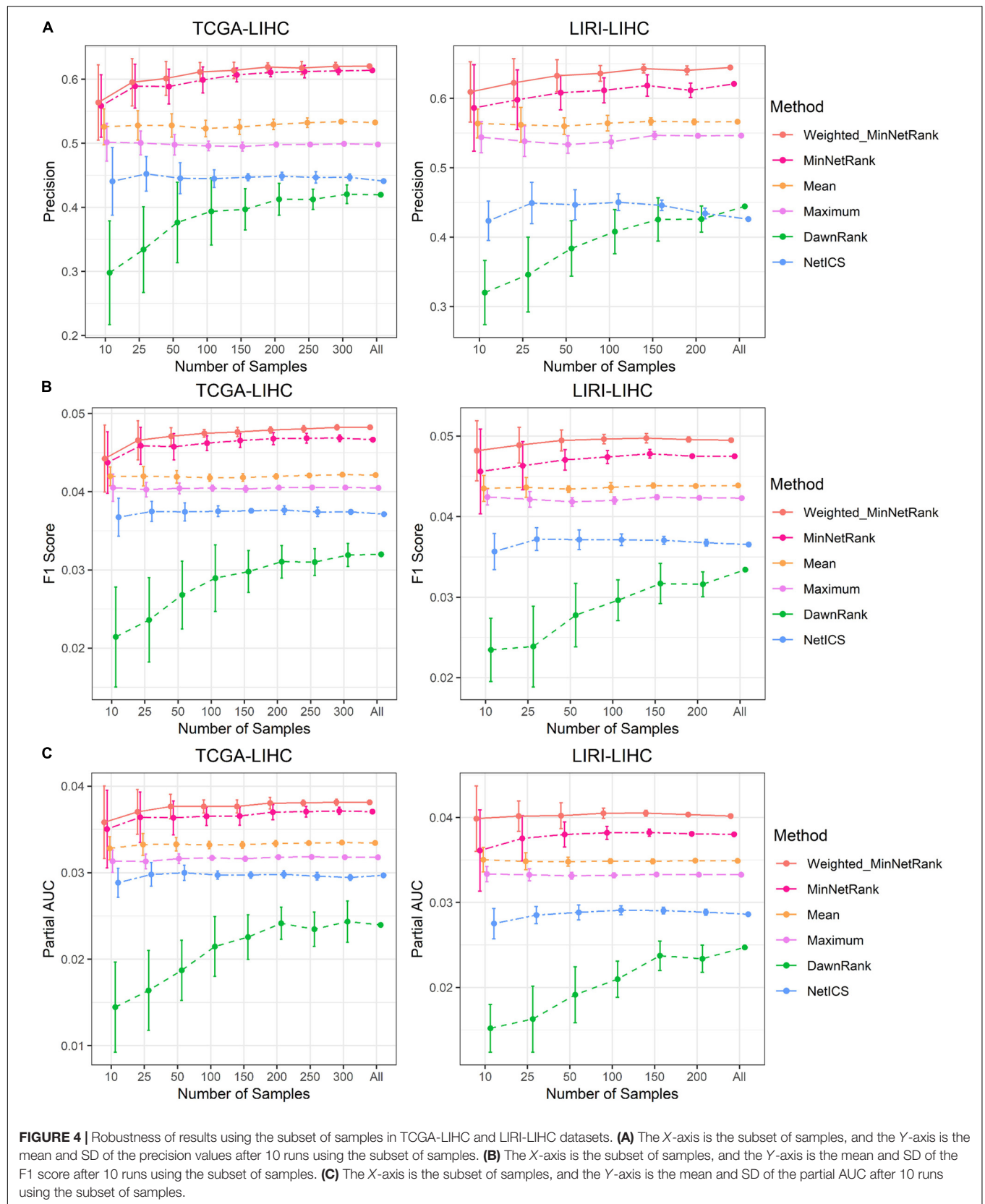


TABLE 1 | Six datasets used in MinNetRank.

Datasets	Data type	Samples	Website
TCGA-LIHC	Mutation	363	https://portal.gdc.cancer.gov/projects/TCGA-LIHC
	RNA expression (tumor)	371	
	RNA expression (normal)	50	
LIRI-LIHC	Mutation	258	https://dcc.icgc.org/projects/LIRI-JP
	RNA expression (tumor)	230	
	RNA expression (normal)	197	
TCGA-STAD	Mutation	437	https://portal.gdc.cancer.gov/projects/TCGA-STAD
	RNA expression (tumor)	375	
	RNA expression (normal)	32	
TCGA-BLCA	Mutation	412	https://portal.gdc.cancer.gov/projects/TCGA-BLCA
	RNA expression (tumor)	408	
	RNA expression (normal)	19	
TCGA-LUAD	Mutation	565	https://portal.gdc.cancer.gov/projects/TCGA-LUAD
	RNA expression (tumor)	513	
	RNA expression (normal)	59	
TCGA-SKCM	Mutation	467	https://portal.gdc.cancer.gov/projects/TCGA-SKCM
	RNA expression (tumor)	468	
	RNA expression (normal)	1	

and partial AUC in all six datasets. For weighted methods, Weighted_MinNetRank_PrCID had better performance than PrDSM weighted methods (Weighted_MinNetRank_PrDSM and Weighted_MinNetRank_Filter_PrDSM) in all datasets. There was no significant difference between Weighted_MinNetRank_PrCID and Weighted_MinNetRank. There were some possible reasons for this phenomenon. Firstly, there were many synonymous mutations in all datasets (32,381 synonymous mutations on average); however, the percentage of deleterious synonymous mutations was relatively small (9.76% in the study of PrDSM) (Cheng et al., 2019). Many benign synonymous mutations increased noise. We may need to pre-process the scores of synonymous mutations (Weighted_MinNetRank_Filter_PrDSM performed better than Weighted_MinNetRank_PrDSM). Secondly, the number of missense mutations was the largest, and the number of frameshift mutations was small, so Weighted_MinNetRank weighting for missense mutations had almost the same performance as Weighted_MinNetRank_PrCID weighting for missense mutations and frameshift mutations. LIRI-LIHC dataset did not provide the position information of frameshift mutations in cDNA, so Weighted_MinNetRank_PrCID was not available for LIRI-LIHC dataset.

MinNetRank Discovered Rare and Novel Driver Genes

In addition to obtaining the accurate and robust results, one of the main advantages of MinNetRank was to discover rare and personalized cancer genes. Personalized driver genes could contribute to the development of personalized medicine.

A gene was considered as a rare gene if the gene was mutated in a small number of samples (<5%). For the top 50 candidate driver genes of MinNetRank, the numbers of rare genes in TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA,

TCGA-LUAD, and TCGA-SKCM were 48 (96%), 48 (96%), 42 (84%), 44 (88%), 48 (96%), and 42 (84%), respectively. Among rare genes, 28 genes (58.33%), 27 genes (56.25%), 27 genes (64.28%), 27 genes (61.36%), 27 genes (56.25%), and 27 genes (64.28%) have not been classified as known cancer gene in TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively. We further investigated the rare genes in CGC (gold standard cancer driver gene set), and there were 98.00, 97.95, 85.05, 90.79, 91.73, and 82.11% rare genes in TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively. The proportion of rare genes in CGC was high, and the proportion of rare genes for all CGC known cancer genes was approximately the same as the proportion of rare genes for the top 50 candidate driver genes.

MinNetRank also identified novel cancer driver genes that have not been classified as drivers by other methods. Taking an example for *SP1*, *SP1* was considered as a cancer gene only by MinNetRank and was ranked 3rd, 3rd, 3rd, 2nd, 3rd, and 1st in TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively (**Supplementary Table 1**). The mutation frequency of *SP1* was 8.26×10^{-3} , 1.60×10^{-2} , 2.43×10^{-2} , 8.85×10^{-3} , and 1.07×10^{-2} (ranked 2903rd, 6393rd, 1599th, 7892nd, and 10330th in terms of the mutation frequency) in TCGA-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively. *SP1* was a zinc finger transcription factor and was reported to be associated with cell differentiation, proliferation, and apoptosis (Beishline and Azizkhan-Clifford, 2015; Safe et al., 2018). Using pathway enrichment analysis, we found that *SP1* was involved in multiple pathways enriched by known cancer genes, such as the transforming growth factor (TGF)-beta signaling pathway and choline metabolism in cancer and breast cancer.

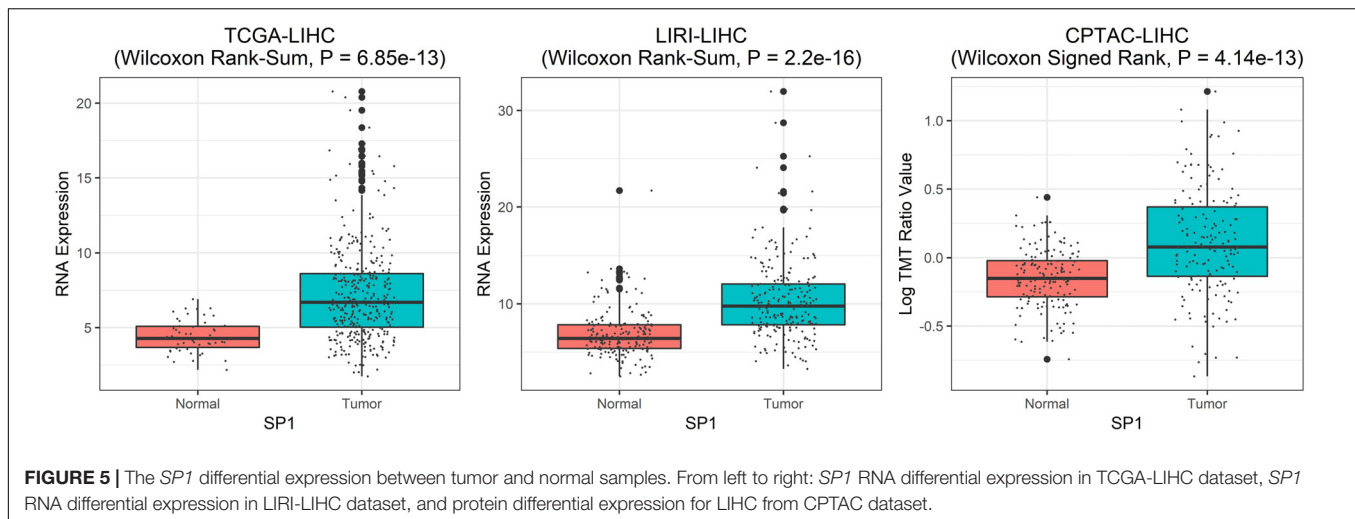
As shown in **Figure 5** (for datasets TCGA-LIHC and LIRI-LIHC), **Supplementary Figure 5** (for datasets TCGA-STAD

TABLE 2 | The performance of each part of MinNetRank according to the precision, F1 score, and partial AUC value.

Metrics	Methods	TCGA-LIHC	LIRI-LIHC	TCGA-STAD	TCGA-BLCA	TCGA-LUAD	TCGA-SKCM
Precision	Weighted_MinNetRank	0.620	0.645	0.602	0.623	0.583	0.533
	Weighted_MinNetRank_PrDSM	0.615	0.633	0.591	0.613	0.573	0.523
	Weighted_MinNetRank_Filter_PrDSM	0.621	0.629	0.599	0.621	0.575	0.528
	Weighted_MinNetRank_PrCID	0.628	–	0.594	0.630	0.580	0.533
	MinNetRank	0.614	0.621	0.585	0.608	0.576	0.515
	MinNetRank (mutation)	0.569	0.576	0.514	0.563	0.445	0.390
	MinNetRank (expression)	0.574	0.580	0.479	0.517	0.512	0.549
	DawnRank	0.420	0.444	0.473	0.586	0.405	0.404
	NetICS	0.441	0.426	0.437	0.453	0.393	0.161
	Mean	0.532	0.566	0.461	0.520	0.414	0.411
	Maximum	0.498	0.546	0.452	0.483	0.405	0.420
	Freq	0.255	0.277	0.249	0.511	0.194	0.149
	Degree centrality	0.189	0.189	0.189	0.189	0.189	0.189
	Betweenness centrality	0.521	0.521	0.521	0.521	0.521	0.521
	Mean of degree and betweenness	0.493	0.493	0.493	0.493	0.493	0.493
F1 score	Weighted_MinNetRank	0.048	0.049	0.046	0.048	0.044	0.042
	Weighted_MinNetRank_PrDSM	0.047	0.049	0.045	0.047	0.044	0.041
	Weighted_MinNetRank_Filter_PrDSM	0.048	0.048	0.046	0.047	0.044	0.041
	Weighted_MinNetRank_PrCID	0.048	–	0.045	0.047	0.044	0.042
	MinNetRank	0.047	0.047	0.045	0.046	0.043	0.041
	MinNetRank (mutation)	0.043	0.044	0.042	0.044	0.039	0.036
	MinNetRank (expression)	0.045	0.046	0.039	0.040	0.040	0.043
	DawnRank	0.032	0.033	0.039	0.043	0.029	0.027
	NetICS	0.037	0.037	0.037	0.037	0.035	0.016
	Mean	0.042	0.044	0.038	0.041	0.037	0.037
	Maximum	0.040	0.042	0.037	0.039	0.037	0.039
	Freq	0.018	0.018	0.017	0.038	0.012	0.011
	Degree centrality	0.013	0.013	0.013	0.013	0.013	0.013
	Betweenness centrality	0.044	0.044	0.044	0.044	0.044	0.044
	Mean of degree and betweenness	0.042	0.042	0.042	0.042	0.042	0.042
Partial AUC	Weighted_MinNetRank	0.038	0.040	0.035	0.038	0.034	0.033
	Weighted_MinNetRank_PrDSM	0.037	0.039	0.034	0.037	0.034	0.032
	Weighted_MinNetRank_Filter_PrDSM	0.038	0.039	0.035	0.038	0.034	0.032
	Weighted_MinNetRank_PrCID	0.038	–	0.034	0.038	0.034	0.033
	MinNetRank	0.037	0.038	0.034	0.037	0.034	0.032
	MinNetRank (mutation)	0.033	0.036	0.032	0.035	0.031	0.029
	MinNetRank (expression)	0.034	0.035	0.031	0.031	0.031	0.034
	DawnRank	0.024	0.025	0.032	0.036	0.022	0.021
	NetICS	0.030	0.029	0.030	0.029	0.029	0.011
	Mean	0.033	0.035	0.031	0.034	0.029	0.031
	Maximum	0.032	0.033	0.029	0.031	0.028	0.030
	Freq	0.011	0.011	0.010	0.026	0.007	0.006
	Degree centrality	0.007	0.007	0.007	0.007	0.007	0.007
	Betweenness centrality	0.035	0.035	0.035	0.035	0.035	0.035
	Mean of degree and betweenness	0.033	0.033	0.033	0.033	0.033	0.033

and TCGA-BLCA), and **Supplementary Figure 6** (for datasets TCGA-LUAD and TCGA-SKCM), *SP1* RNA expression of tumor samples was statistically higher than normal samples in TCGA-LIHC (Wilcoxon Rank-Sum, $P = 6.85e-13$), LIRI-LIHC (Wilcoxon Rank-Sum, $P = 2.2e-16$), and TCGA-STAD (Wilcoxon Rank-Sum, $P = 5.89e-10$). The differential expression was not significant in TCGA-BLCA (Wilcoxon Rank-Sum,

$P = 0.17$), TCGA-LUAD (Wilcoxon Rank-Sum, $P = 0.95$), and TCGA-SKCM (Wilcoxon Rank-Sum, $P = 0.21$). We further validated *SP1* expression on the protein level, and the differential protein expression between tumor and normal samples was significant in LIHC (Wilcoxon Signed Rank test, $P = 4.14e-13$). Only LIHC had protein expression data from CPTAC (The National Cancer Institute's Clinical Proteomic Tumor Analysis



Consortium) dataset. These results suggested that *SP1* can be the biomarker of hepatocellular carcinoma.

Top Genes of MinNetRank Were Associated With Clinical Outcome

For each dataset, we selected seven genes with top ranking and high SD as biomarkers for tumor stratification (mentioned in the section “Materials and Methods”). We performed unsupervised K-means clustering using obtained biomarkers to assign each patient into either high-risk or low-risk groups. The Kaplan–Meier survival curves of the two groups are well separated, and the log-rank P-values of the survival difference between two groups are $9.21\text{e-}04$, $1.23\text{e-}05$, $2.42\text{e-}03$, $3.75\text{e-}03$, $9.21\text{e-}04$, and $4.19\text{e-}02$ for TCGA-LIHC, LIRI-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM, respectively [Figure 6 (for datasets TCGA-LIHC and LIRI-LIHC), Supplementary Figure 7 (for datasets TCGA-STAD and TCGA-BLCA), and Supplementary Figure 8 (for datasets TCGA-LUAD and TCGA-SKCM)].

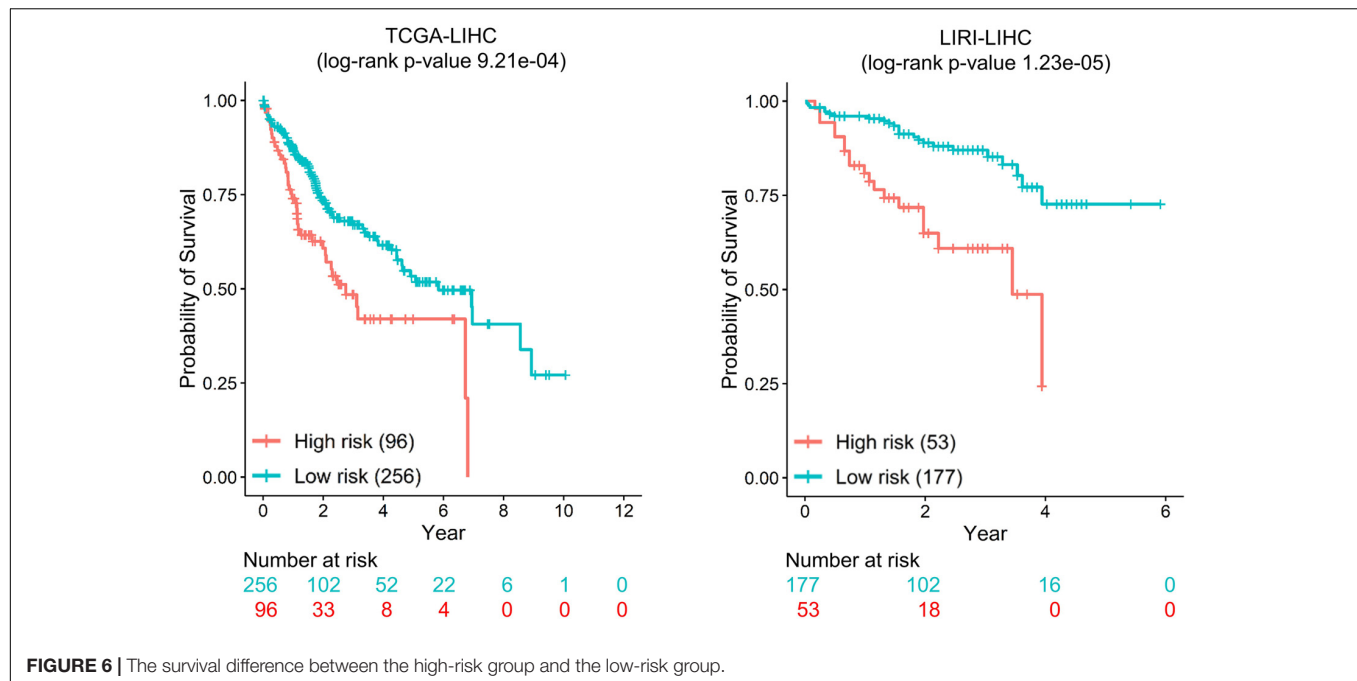
In the two liver cancer datasets (TCGA-LIHC and LIRI-LIHC), there were six shared genes (*CTNNB1*, *JUN*, *PIK3R1*, *RAC1*, *SRC*, and *TP53*). All these genes used for tumor stratification are biologically relevant. *CTNNB1* regulated cell growth and adhesion and was predictive for recurrence in aggressive fibromatosis (van Broekhoven et al., 2015). *JUN* (AP-1 Transcription Factor Subunit) participated in regulating a diverse array of cellular processes, including proliferation, apoptosis, differentiation, and survival (Trop-Steinberg and Azar, 2017). *PIK3R1* was a prognostic biomarker for breast cancer (Cizkova et al., 2013). *RAC1* regulated a wide range of cellular events, including the control of cell growth and the activation of protein kinases (Lou et al., 2018). *SRC* was prognostic relevant to colon cancer and rectal cancer (Martínez-Pérez et al., 2017). *TP53* was one of the most frequent alterations and potential prognostic markers in human cancers (Olivier et al., 2010). *GRB2* was the special biomarker for TCGA-LIHC, and *MAPK14* was for LIRI-LIHC. *GRB2* was evaluated as a prognostic marker for lung adenocarcinoma (Toki et al., 2016). *MAPK14* was a

member of the MAP kinase family. MAPK pathway regulated cell proliferation, differentiation, and development (Fang and Richardson, 2005). The seven biomarkers are the same in TCGA-STAD and TCGA-BLCA (*CTNNB1*, *GRB2*, *JUN*, *RAC1*, *SP1*, *SRC*, and *TP53*). These seven genes were reported to be related to prognosis (Hang et al., 2016). For TCGA-LUAD and TCGA-SKCM, there were six shared genes (*CTNNB1*, *JUN*, *RAC1*, *SRC*, *TP53*, and *GRB2*). *GNB1* was the special biomarker for TCGA-LUAD, and *FYN* was for TCGA-SKCM. *FYN* was tyrosine kinases and was an essential molecule in cancer pathogenesis and drug resistance (Elias and Ditzel, 2015). In summary, the top seven genes were associated with clinical outcome and were biologically relevant in all six datasets. These results suggested that MinNetRank could also be a promising method for tumor stratification.

NetICS and DawnRank did not investigate the prognostic value of top genes in cancer. To evaluate the performance of predicting the clinical outcome for different methods, we used the same criterion to choose the top seven genes for each method in six datasets. Compared with NetICS and DawnRank, only Weighted_MinNetRank and MinNetRank obtained a statistically significant survival risk difference between the high-risk and low-risk groups in all six datasets (Supplementary Table 2).

DISCUSSION

Extensive genetic heterogeneity exists between tumors of different tissues and between individuals with the same tumor type (Burrell et al., 2013). The personalized mutation profile is the key to advance personalized disease diagnosis and therapy in the clinic (Sheng et al., 2015; Olivier et al., 2019). However, few methods could efficiently prioritize driver genes over many passenger genes in a single patient. The critical challenge facing today is to predict rare and even personalized driver genes with higher accuracy. We develop MinNetRank, an efficient and easy-to-use method that integrates the mutation data, expression data and interaction network to prioritize each sample's driver genes.



Weighted_MinNetRank further considers the different weights for the different types of mutations.

Weighted_MinNetRank and MinNetRank achieve a higher precision, F1 score, and partial AUC value of prioritizing cancer genes in five TCGA datasets (TCGA-LIHC, TCGA-STAD, TCGA-BLCA, TCGA-LUAD, and TCGA-SKCM). We also utilize an additional liver cancer cohort (LIRI-LIHC) to validate the result of TCGA-LIHC. Better performance in all datasets demonstrates the proposed approach's robustness (Figure 3 and Table 2). We use top candidate driver genes for pathway enrichment analysis and find some signaling pathways previously studied in cancer, such as the Ras signaling pathway and ErbB signaling pathway. Furthermore, we first investigate the relationship between the top seven genes and clinical outcome and find the statistically significant survival difference between the low-risk and high-risk groups in all six datasets only for Weighted_MinNetRank and MinNetRank. The top seven genes are biologically relevant and could be used as biomarkers for survival risk stratification. Accurate outcome prediction is important for personalized cancer therapies in clinical practice, for instance, a low-risk patient can be advised to select a less radical therapy.

We demonstrate that MinNetRank can discover rare and novel cancer genes. Personalized driver genes could contribute to developing personalized diagnosis and therapy. *SP1* is considered a candidate driver gene only by MinNetRank and is ranked top three in all six datasets. The RNA expression of *SP1* is significantly higher in LIHC tumor samples (TCGA-LIHC and LIRI-LIHC datasets) and STAD tumor samples (TCGA-STAD dataset). The differential expression is further validated on the protein level in LIHC. *SP1* is the biomarker for tumor stratification in TCGA-STAD and TCGA-BLCA, and *SP1* RNA expression is associated with survival outcome in TCGA-STAD

dataset (Cox proportional hazards model, $P = 0.02$). These results are in accordance with the reports in literatures (Shi and Zhang, 2019). Targeting *SP1* is highly promising strategy in cancer chemotherapy (Vizcaino et al., 2015).

Using both the incoming and outgoing degree of interaction network, the minimum strategy and weighting for the different types of mutations all contribute to the accuracy and robustness of prioritizing driver genes. Known cancer genes have a higher incoming and outgoing degree, and simultaneously considering incoming and outgoing degree is rational. MinNetRank adopts a minimum strategy to prioritize cancer genes with a high mutation relevance score and high expression relevance score. These enable our method to select more relevant genes and avoid the potential outliers, which are common in high-throughput sequencing technologies due to the positively skewed distributions of mutation and expression relevance scores. Weighting for different types of mutations is essential for sample-specific study and finding personalized driver genes.

There are some limitations to MinNetRank and similar methods. Firstly, MinNetRank largely depends on the interaction network. Although many interaction sources exist, such as experiment, co-expression, and text mining, the interaction network is still incomplete. If the mutation gene or differentially expressed gene is not in the interaction network, this gene would not be used for network diffusion and not be as a candidate cancer gene. Secondly, MinNetRank uses paired tumor and normal samples to calculate *ALFC*; however, TCGA datasets have a limited number of normal samples with expression data. Thirdly, MinNetRank only integrates mutation data and expression data into the interaction network. Besides mutation data, other events, such as miRNA differential expression, epigenetic changes, copy number variation, and structure variation, could also contribute to cancer progression. Differential expression data, including

RNA expression data and protein expression data, could be combined. We may need to improve MinNetRank from two aspects in the future. On one hand, we could integrate the gene co-expression network with the interaction networks (Hou et al., 2019; Wei et al., 2020). We also need to incorporate additional types of omics data (genomics, transcriptomics, proteomics, epigenomics, and images). On the other hand, Weighted_MinNetRank only considers mutations in coding region. We may need to incorporate non-coding mutations. We also need to give weight coefficients for all mutations through multiple techniques.

Integrating different types of omics data is often used to better elucidate the molecular function. However, sound study designs and solid analytical strategies are needed to advance human disease research further. For example, the mean precision of the top 50 cancer genes is 0.61 (MinNetRank) and 0.56 (NetICS) in TCGA-LIHC and 0.61 (MinNetRank) and 0.54 (NetICS) in TCGA-BLCA. The top 50 candidate cancer genes of NetICS used here are from the published paper (Dimitrakopoulos et al., 2018). In this article, NetICS integrates different types of data that include somatic mutation, copy number variation, methylation, miRNA expression, gene expression, and protein expression. Although MinNetRank only focuses on integrating the mutation data and expression data, the mean precision of MinNetRank is still higher than that of NetICS.

CONCLUSION

This article developed a new method (denoted as MinNetRank) by setting weights for different types of mutations and using the minimum strategy to integrate multi-omics for cancer genes discovery. Minimum strategy reduced the influence of extreme scores in highly skewed distributions and was the “double high” strategy to prioritize cancer genes, having a relatively high mutation score and expression score. Different weight coefficients for the different types of mutations contributed to the better performance. We demonstrated our method’s accuracy and robustness in prioritizing driver genes on five TCGA datasets and one ICGC dataset. Besides, MinNetRank has the advantage of discovering rare and personalized cancer genes. The top seven candidate driver genes stratified patients into two subtypes (high-risk and low-risk groups) exhibiting significant survival differences and could be used as prognostic biomarkers for survival. Of course, our method has room for improvement. Gene co-expression network and more types of omics data should be incorporated, and different weight coefficients should be considered.

MATERIALS AND METHODS

Dataset

The genes annotated in the CGC can be used to benchmark known cancer genes (Tate et al., 2019). This gold standard known cancer gene set includes 576 genes (July 2019)¹. Many cancer

studies use CGC genes as the benchmark for the evaluation (Bashashati et al., 2012; Hou and Ma, 2014; Bertrand et al., 2015; Wei et al., 2017; Guo et al., 2018).

Interaction Network

We used the interaction network that has been widely used in the related paper (Hou and Ma, 2014; Guo et al., 2018). The interaction network integrated a variety of resources, including the network used in MEMO as well as the up-to-date information from Reactome (Croft et al., 2011; Ciriello et al., 2012), the NCI-Nature Pathway Interaction Database (Schaefer et al., 2009), and KEGG (Kanehisa et al., 2016). The resulting interaction network consisted of 11,648 genes and 211,794 edges. The average degree centrality of interaction network was 34.20, and the average betweenness centrality was 1.58E-04.

MinNetRank

MinNetRank uses an interaction network that could discover cancer driver genes more efficiently (Leiserson et al., 2015). One of the main reasons for this is the high connectivity (high incoming degree and outgoing degree) of known cancer genes in the interaction network. For example, the mean and median of incoming degree for known cancer genes (in CGC) are 36.06 and 17, which are much higher than those of the genes that are not classified as known cancer genes (17.41 and 3, respectively). Also, the mean and median outgoing degree of known cancer genes are 30.37 and 12, which are much higher than those of the genes that are not in CGC (17.66 and 4, respectively). To a certain extent, this is expected since genes with high connectivity could exert a more significant influence on the biological system (Winter et al., 2012). RWR algorithm models how closely related the two genes are and measures both the direct and indirect neighbors of each gene in the interaction network, making it more sensitive for prioritizing cancer driver genes (Dimitrakopoulos et al., 2018). Unlike NetICS and DawnRank, we consider both incoming and outgoing degree of interaction network for single omics.

Diffused Matrix

Let A be the $n \times n$ adjacency matrix of an interaction network where n represents the number of nodes (the number of genes in the interaction network). A is a 0–1 matrix and $a_{ij} = 1$ if there is a directed edge from node j to node i . A' is the transpose of matrix A and $a_{ji} = 1$ if there is a directed edge from node i to node j . We denote $deg_j^{out} = \sum_{i=1}^N a_{ij}$ as the outgoing degree of node j or the number of outgoing edges. While $deg_j^{in} = \sum_{i=1}^N a_{ji}$ is the incoming degree of node j . MinNetRank considers both the incoming degree and outgoing degree, so we define the normalized adjacency matrix A^{norm} as,

$$A^{norm} = \begin{pmatrix} \frac{a_{11}+a_{11}}{deg_1^{out}+deg_1^{in}} & \cdots & \frac{a_{1n}+a_{n1}}{deg_n^{out}+deg_n^{in}} \\ \vdots & \ddots & \vdots \\ \frac{a_{n1}+a_{1n}}{deg_1^{out}+deg_1^{in}} & \cdots & \frac{a_{nn}+a_{nn}}{deg_n^{out}+deg_n^{in}} \end{pmatrix}. \quad (4)$$

We define the diffused matrix D as,

$$D = \beta [I - (1 - \beta)A^{norm}]^{-1} \quad (5)$$

¹<https://cancer.sanger.ac.uk/census>

The value of D_{ij} lies between 0 and 1 and reflects the connectivity between nodes j and i . Higher score means that two genes are more closely related. The restart probability of β ($0 \leq \beta \leq 1$) determines the degree of diffusion, namely, how far the random walker can move in the network. When $\beta=1$, there is no diffusion, namely, we do not use the information of the interaction network. When $\beta=0$, gene mutation score or differential expression score (see below) diffuses to the whole network. β depends on the interaction network and is independent of any mutation data or expression data. We chose β to balance diffusion and retainment (Leiserson et al., 2015), and β is 0.48 in this study. The diffused matrix D needs to be computed only once for a given interaction network.

ALFC

For each patient k , we calculate the Absolute value of Log2 Fold-Change ($ALFC$) of gene i for the paired tumor and normal samples as a differential expression score. The fold change, or relative difference, is widely used to measure differential gene expression (Love et al., 2014). The absolute value of fold change is taken in order to capture both upregulation and downregulation.

$$ALFC_{ik} = \begin{cases} \left| \log_2 \frac{\text{gene } i \text{ expression of tumor sample in patient } k}{\text{gene } i \text{ expression of normal sample in patient } k} \right| & \text{paired tumor and normal samples} \\ \left| \log_2 \frac{\text{gene } i \text{ expression of tumor sample in patient } k}{\text{the mean of gene } i \text{ expression of all normal samples}} \right| & \text{unpaired} \end{cases} \quad (6)$$

Weighted_MinNetRank

Weighted_MinNetRank uses SIFT scores (between 0 and 1) as the weight coefficients for missense mutations and gives the same weight with 1 to other mutations (stop-gain, stop-loss, frameshift, and non-frameshift) (Ng and Henikoff, 2001). Although synonymous mutations do not alter amino acids, some deleterious synonymous mutations play important roles in cancer (Wen et al., 2016). We further incorporate synonymous mutations and use PrDSM scores as the weights for synonymous mutations (Weighted_MinNetRank_PrDSM). We also use PrDSM scores greater than 0.38 as the weights (Weighted_MinNetRank_Filter_PrDSM). If a PrDSM score is greater than 0.308, the corresponding synonymous mutation is considered as deleterious (Cheng et al., 2019). Besides, we use PredCID scores as the weights for frameshift mutations (Weighted_MinNetRank_PrCID) (Yue et al., 2020).

Assessing the Performance in Predicting Known Cancer Genes

In order to assess the performance in predicting known cancer genes, our method (Weighted_MinNetRank and MinNetRank) was compared with NetICS (Dimitrakopoulos et al., 2018), DawnRank (Hou and Ma, 2014), and Freq. The top 50 genes of the population-level ranking were identified as candidate driver genes and compared with the positive genes in CGC. We used the precision, F1 score, and partial AUC value to evaluate the performance. The precision was defined as expression (7) and can be viewed as the measure of exactness. The recall was the percentage of total known cancer genes correctly predicted by MinNetRank. F1 score combined recall and precision using

the harmonic mean. There were many more negative genes than positive genes (positives/negatives = 0.052) and even fewer positive genes when we considered cancer type-specific known cancer genes (positives/negatives \approx 0.0029). It was more informative to use partial AUC, which considered the number of true positives scored higher than the n th highest scoring negatives, measured for all values from 1 to n (Dimitrakopoulos et al., 2018). Precision, F1 score, and partial AUC were based on the top 50 genes.

$$\text{precision} = \frac{(\text{CGC genes}) \cap (\text{Top } N \text{ predicted driver genes})}{\text{Top } N \text{ predicted driver genes}} \quad (7)$$

$$\text{recall} = \frac{(\text{CGC genes}) \cap (\text{Top } N \text{ predicted driver genes})}{\text{CGC genes}} \quad (8)$$

$$F1 \text{ Score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (9)$$

$$AUC_n = \frac{1}{nT} \sum_{i=1}^n T_i, \quad (10)$$

where T was the total number of known cancer genes in CGC, and T_i was the number of positives scored higher than the i th highest scoring negatives.

Assessing the Robustness Using the Subset of Samples

In order to further compare these methods, we calculated the precision, F1 score, and partial AUC using the subset of samples with different sample sizes. We experimented with sample sizes of $n = 10, 25, 50 \times 1, 50 \times 2, \dots, 50 \times \lceil N/50 \rceil$, and N was the total sample size of multi-omics data. For each sample size, we performed 10 random samples. We defined the precision value $P = \text{mean}(p_i)$, where p_i was the precision of top i candidate cancer gene, $i = 1, 2, \dots, 50$. The mean and SD of precision value, F1 score, and partial AUC value for 10 runs were used to measure the robustness.

Tumor Stratification

Some papers used gene mutation data and expression data to identify genes that were indicators for survival. Using these biomarkers, patients can be stratified into subtypes (Haider et al., 2014). We further investigated the relationship between the top genes of population-level ranking and patients' survival time. Genes whose expression with a low variation between tumors provided very limited information for tumor stratification (Winter et al., 2012). According to the genes' rankings, we selected the top seven genes with a greater SD of expression than five as biomarkers for each dataset (Winter et al., 2012). Using these seven biomarkers, K-means clustering (unsupervised learning algorithm) assigned each patient to one of the two clusters (high-risk and low-risk groups). The log-rank test was then used to compare the survival differences of the two groups (R survival package).

DATA AVAILABILITY STATEMENT

The mutation data, expression data, and clinical data of the TCGA dataset are available in the TCGA Data Portal (<https://portal.gdc.cancer.gov/projects/>). Those from LIRI-JP are available in ICGC Data Portal (<https://dcc.icgc.org/projects/LIRI-JP>). The LIHC protein expression data are from CPTAC Data Portal (<https://proteomics.cancer.gov/data-portal>). The detail descriptions of these data are provided in **Table 1**. The example data used to demonstrate MinNetRank are available at <https://github.com/weitingting/MinNetRank>.

AUTHOR CONTRIBUTIONS

ZY and TW: conceptualization, methodology, and validation. TW: software, formal analysis, investigation, and writing—original draft preparation. BF and TW: data curation. ZY, CL, LJ, and TW: writing—review and editing. ZY: supervision and project administration. ZY and YZ: funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China 11671256 and Shanghai Jiao Tong University STAR Grant.

REFERENCES

- Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., Rosner, J., et al. (2012). DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.* 13:R124. doi: 10.1186/gb-2012-13-12-r124
- Beishline, K., and Azizkhan-Clifford, J. (2015). Sp1 and the 'hallmarks of cancer'. *FEBS J.* 282, 224–258. doi: 10.1111/febs.13148
- Bertrand, D., Chng, K. R., Sherbaf, F. G., Kiesel, A., Chia, B. K., Sia, Y. Y., et al. (2015). Patient-specific driver gene prediction and risk assessment through integrated network analysis of cancer omics profiles. *Nucleic Acids Res.* 43:e44. doi: 10.1093/nar/gku1393
- Burrell, R. A., McGranahan, N., Bartek, J., and Swanton, C. (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* 501, 338–345. doi: 10.1038/nature12625
- Chen, X., Liu, M. X., and Yan, G. Y. (2012). Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.* 8, 1970–1978. doi: 10.1039/c2mb00002d
- Cheng, N., Li, M., Zhao, L., Zhang, B., Yang, Y., Zheng, C.-H., et al. (2019). Comparison and integration of computational methods for deleterious synonymous mutation prediction. *Brief. Bioinform.* 21, 970–981. doi: 10.1093/bib/bbz047
- Cho, A., Shim, J. E., Kim, E., Supek, F., Lehner, B., and Lee, I. (2016). MUFFINN: cancer gene discovery via network analysis of somatic mutation data. *Genome Biol.* 17:129. doi: 10.1186/s13059-016-0989-x
- Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* 22, 398–406. doi: 10.1101/gr.125567.111
- Cizkova, M., Vacher, S., Meseure, D., Trassard, M., Susini, A., Mlcuchova, D., et al. (2013). PIK3R1 underexpression is an independent prognostic marker in breast cancer. *BMC Cancer* 13:545. doi: 10.1186/1471-2407-13-545

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.613033/full#supplementary-material>

Supplementary Figure 1 | Comparison of precision and F1 score for different methods in TCGA-STAD and TCGA-BLCA datasets.

Supplementary Figure 2 | Comparison of precision and F1 score for different methods in TCGA-LUAD and TCGA-SKCM datasets.

Supplementary Figure 3 | Robustness of results using the subset of samples in TCGA-STAD and TCGA-BLCA datasets.

Supplementary Figure 4 | Robustness of results using the subset of samples in TCGA-LUAD and TCGA-SKCM datasets.

Supplementary Figure 5 | The SP1 differential expression between tumor and normal in TCGA-STAD and TCGA-BLCA.

Supplementary Figure 6 | The SP1 differential expression between tumor and normal in TCGA-LUAD and TCGA-SKCM.

Supplementary Figure 7 | The survival difference between high-risk group and low-risk group in TCGA-STAD and TCGA-BLCA.

Supplementary Figure 8 | The survival difference between high-risk group and low-risk group in TCGA-LUAD and TCGA-SKCM.

Supplementary Table 1 | Top 50 driver genes of six methods.

Supplementary Table 2 | The log-rank *P*-value of tumor stratification for each method in six datasets.

- 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, D691–D697. doi: 10.1093/nar/gkq1018
- Dagogo-Jack, I., and Shaw, A. T. (2018). Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* 15, 81–94. doi: 10.1038/nrclinonc.2017.166
- Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 22, 1589–1598. doi: 10.1101/gr.134635.111
- Dimitrakopoulos, C., Hindupur, S. K., Häfliger, L., Behr, J., Montazeri, H., Hall, M. N., et al. (2018). Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* 34, 2441–2448. doi: 10.1093/bioinformatics/bty148
- Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455, 1069–1075. doi: 10.1038/nature07423
- Elias, D., and Ditzel, H. J. (2015). Fyn is an important molecule in cancer pathogenesis and drug resistance. *Pharmacol. Res.* 100, 250–254. doi: 10.1016/j.phrs.2015.08.010
- Fa, B., Luo, C., Tang, Z., Yan, Y., Zhang, Y., and Yu, Z. (2019). Pathway-based biomarker identification with crosstalk analysis for robust prognosis prediction in hepatocellular carcinoma. *EBioMedicine* 44, 250–260. doi: 10.1016/j.ebiom.2019.05.010
- Fang, J. Y., and Richardson, B. C. (2005). The MAPK signalling pathways and colorectal cancer. *Lancet Oncol.* 6, 322–327. doi: 10.1016/S1470-2045(05)70168-6
- Greenman, C., Stephens, P., Smith, R., Dalgleish, G. L., Hunter, C., Bignell, G., et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153–158. doi: 10.1038/nature05610
- Guo, W. F., Zhang, S. W., Liu, L. L., Liu, F., Shi, Q. Q., Zhang, L., et al. (2018). Discovering personalized driver mutation profiles of single samples in

- cancer by network control strategy. *Bioinformatics* 34, 1893–1903. doi: 10.1093/bioinformatics/bty006
- Haider, S., Wang, J., Nagano, A., Desai, A., Arumugam, P., Dumartin, L., et al. (2014). A multi-gene signature predicts outcome in patients with pancreatic ductal adenocarcinoma. *Genome Med.* 6:105. doi: 10.1186/s13073-014-0105-3
- Hang, J., Hu, H., Huang, J., Han, T., Zhuo, M., Zhou, Y., et al. (2016). Sp1 and COX2 expression is positively correlated with a poor prognosis in pancreatic ductal adenocarcinoma. *Oncotarget* 7, 28207–28217. doi: 10.18632/oncotarget.8593
- Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biol.* 18:83. doi: 10.1186/s13059-017-1215-1
- Hou, J. P., and Ma, J. (2014). DawnRank: discovering personalized driver genes in cancer. *Genome Med.* 6:56. doi: 10.1186/s13073-014-0056-8
- Hou, M.-X., Gao, Y.-L., Liu, J.-X., Shang, J., Zhu, R., and Yuan, S.-S. (2019). A new method for mining information of co-expression network based on multi-cancers integrated data. *BMC Med. Genomics* 12(Suppl. 7):155. doi: 10.1186/s12920-019-0608-2
- Huang, S., Chaudhary, K., and Garmire, L. X. (2017). More is better: recent progress in multi-omics data integration methods. *Front. Genet.* 8:84. doi: 10.3389/fgene.2017.00084
- Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* 44, D457–D462. doi: 10.1093/nar/gkv1070
- Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499, 214–218. doi: 10.1038/nature12213
- Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., et al. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* 47, 106–114. doi: 10.1038/ng.3168
- Lou, S., Wang, P., Yang, J., Ma, J., Liu, C., and Zhou, M. (2018). Prognostic and clinicopathological value of Rac1 in cancer survival: evidence from a meta-analysis. *J. Cancer* 9, 2571–2579. doi: 10.7150/jca.24824
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8
- Luo, P., Ding, Y., Lei, X., and Wu, F. X. (2019). deepDriver: predicting cancer driver genes based on somatic mutations using deep convolutional neural networks. *Front. Genet.* 10:13. doi: 10.3389/fgene.2019.00013
- Malnic, B., Godfrey, P. A., and Buck, L. B. (2004). The human olfactory receptor gene family. *Proc. Natl. Acad. Sci. U.S.A.* 101, 2584–2589. doi: 10.1073/pnas.0307882100
- Martincorena, I., and Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science* 349, 1483–1489. doi: 10.1126/science.aab4082
- Martínez-Pérez, J., Lopez-Calderero, I., Saez, C., Benavent, M., Limon, M. L., Gonzalez-Exposito, R., et al. (2017). Prognostic relevance of Src activation in stage II-III colon cancer. *Hum. Pathol.* 67, 119–125. doi: 10.1016/j.humpath.2017.05.025
- Ng, P. C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. *Genome Res.* 11, 863–874. doi: 10.1101/gr.176601
- Olivier, M., Asmis, R., Hawkins, G. A., Howard, T. D., and Cox, L. A. (2019). The need for multi-omics biomarker signatures in precision medicine. *Int. J. Mol. Sci.* 20:4781. doi: 10.3390/ijms20194781
- Olivier, M., Hollstein, M., and Hainaut, P. (2010). TP53 mutations in human cancers: origins, consequences, and clinical use. *Cold Spring Harb. Perspect. Biol.* 2:a001008. doi: 10.1101/cshperspect.a001008
- Pon, J. R., and Marra, M. A. (2015). Driver and passenger mutations in cancer. *Annu. Rev. Pathol.* 10, 25–50. doi: 10.1146/annurev-pathol-012414-040312
- Riessland, M., Kaczmarek, A., Schneider, S., Swoboda, K. J., Lohr, H., Bradler, C., et al. (2017). Neurocalcin delta suppression protects against spinal muscular atrophy in humans and across species by restoring impaired endocytosis. *Am. J. Hum. Genet.* 100, 297–315. doi: 10.1016/j.ajhg.2017.01.005
- Safe, S., Abbruzzese, J., Abdelrahim, M., and Hedrick, E. (2018). Specificity protein transcription factors and cancer: opportunities for drug development. *Cancer Prev. Res. (Phila)* 11, 371–382. doi: 10.1158/1940-6207.CAPR-17-0407
- Sager, R. (1997). Expression genetics in cancer: shifting the focus from DNA to RNA. *Proc. Natl. Acad. Sci. U.S.A.* 94, 952–955. doi: 10.1073/pnas.94.3.952
- Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et al. (2009). PID: the pathway interaction database. *Nucleic Acids Res.* 37, D674–D679. doi: 10.1093/nar/gkn653
- Sheng, J., Li, F., and Wong, S. T. (2015). Optimal drug prediction from personal genomics profiles. *IEEE J. Biomed. Health Inform.* 19, 1264–1270. doi: 10.1109/JBHI.2015.2412522
- Shi, S., and Zhang, Z. G. (2019). Role of Sp1 expression in gastric cancer: a meta-analysis and bioinformatics analysis. *Oncol. Lett.* 18, 4126–4135. doi: 10.3892/ol.2019.10775
- Sun, Y. V., and Hu, Y. J. (2016). Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. *Adv. Genet.* 93, 147–190. doi: 10.1016/bs.adgen.2015.11.004
- Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47, D941–D947. doi: 10.1093/nar/gky1015
- Toki, M. I., Carvajal-Hausdorf, D. E., Altan, M., McLaughlin, J., Henick, B., Schalper, K. A., et al. (2016). EGFR-GRB2 protein colocalization is a prognostic factor unrelated to overall EGFR expression or EGFR mutation in lung adenocarcinoma. *J. Thorac. Oncol.* 11, 1901–1911. doi: 10.1016/j.jtho.2016.06.025
- Trop-Steinberg, S., and Azar, Y. (2017). AP-1 expression and its clinical relevance in immune disorders and cancer. *Am. J. Med. Sci.* 353, 474–483. doi: 10.1016/j.amjms.2017.01.019
- van Broekhoven, D. L., Verhoef, C., Grünhagen, D. J., van Gorp, J. M., den Bakker, M. A., Hinrichs, J. W., et al. (2015). Prognostic value of CTNNB1 gene mutation in primary sporadic aggressive fibromatosis. *Ann. Surg. Oncol.* 22, 1464–1470. doi: 10.1245/s10434-014-4156-x
- Vandin, F., Upfal, E., and Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.* 18, 507–522. doi: 10.1089/cmb.2010.0265
- Vizcaino, C., Mansilla, S., and Portugal, J. (2015). Sp1 transcription factor: a long-standing target in cancer chemotherapy. *Pharmacol. Ther.* 152, 111–124. doi: 10.1016/j.pharmthera.2015.05.008
- Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A. Jr., and Kinzler, K. W. (2013). Cancer genome landscapes. *Science* 339, 1546–1558. doi: 10.1126/science.1235122
- Wei, P. J., Wu, F. X., Xia, J., Su, Y., Wang, J., and Zheng, C. H. (2020). Prioritizing cancer genes based on an improved random walk method. *Front. Genet.* 11:377. doi: 10.3389/fgene.2020.00377
- Wei, P.-J., Zhang, D., Li, H.-T., Xia, J., and Zheng, C.-H. (2017). DriverFinder: a gene length-based network method to identify cancer driver genes. *Complexity* 2017:4826206. doi: 10.1155/2017/4826206
- 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
- Winter, C., Kristiansen, G., Kersting, S., Roy, J., Aust, D., Knosel, T., et al. (2012). Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput. Biol.* 8:e1002511. doi: 10.1371/journal.pcbi.1002511
- Yap, T. A., Gerlinger, M., Futreal, P. A., Pusztai, L., and Swanton, C. (2012). Intratumor heterogeneity: seeing the wood for the trees. *Sci. Transl. Med.* 4:127s110. doi: 10.1126/scitranslmed.3003854
- Yu, D., Kim, M., Xiao, G., and Hwang, T. H. (2013). Review of biological network data and its applications. *Genomics Inform.* 11, 200–210. doi: 10.5808/GI.2013.11.4.200
- Yue, Z., Chu, X., and Xia, J. (2020). PredCID: prediction of driver frameshift indels in human cancer. *Brief. Bioinform.* bbaa119. doi: 10.1093/bib/bbaa119

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.

Copyright © 2021 Wei, Fa, Luo, Johnston, Zhang and Yu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Identification of Potential Prognostic Competing Triplets in High-Grade Serous Ovarian Cancer

Jian Zhao¹, Xiaofeng Song^{1*}, Tianyi Xu¹, Qichang Yang¹, Jingjing Liu¹, Bin Jiang² and Jing Wu^{3*}

¹ Department of Biomedical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, ² College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing, China, ³ School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China

OPEN ACCESS

Edited by:

Fa Zhang,
Chinese Academy of Sciences
(CAS), China

Reviewed by:

Qi Zhao,
University of Science and Technology
Liaoning, China
Hao Lin,
University of Electronic Science and
Technology of China, China

*Correspondence:

Xiaofeng Song
xfsong@nuaa.edu.cn
Jing Wu
wujing@njmu.edu.cn

Specialty section:

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

Received: 18 September 2020

Accepted: 19 November 2020

Published: 13 January 2021

Citation:

Zhao J, Song X, Xu T, Yang Q, Liu J,
Jiang B and Wu J (2021) Identification
of Potential Prognostic Competing
Triplets in High-Grade Serous Ovarian
Cancer. *Front. Genet.* 11:607722.
doi: 10.3389/fgene.2020.607722

Increasing lncRNA-associated competing triplets were found to play important roles in cancers. With the accumulation of high-throughput sequencing data in public databases, the size of available tumor samples is becoming larger and larger, which introduces new challenges to identify competing triplets. Here, we developed a novel method, called LncMiM, to detect the lncRNA-miRNA-mRNA competing triplets in ovarian cancer with tumor samples from the TCGA database. In LncMiM, non-linear correlation analysis is used to cover the problem of weak correlations between miRNA-target pairs, which is mainly due to the difference in the magnitude of the expression level. In addition, besides the miRNA, the impact of lncRNA and mRNA on the interactions in triplets is also considered to improve the identification sensitivity of LncMiM without reducing its accuracy. By using LncMiM, a total of 847 lncRNA-associated competing triplets were found. All the competing triplets form a miRNA-lncRNA pair centered regulatory network, in which ZFAS1, SNHG29, GAS5, AC112491.1, and AC099850.4 are the top five lncRNAs with most connections. The results of biological process and KEGG pathway enrichment analysis indicates that the competing triplets are mainly associated with cell division, cell proliferation, cell cycle, oocyte meiosis, oxidative phosphorylation, ribosome, and p53 signaling pathway. Through survival analysis, 107 potential prognostic biomarkers are found in the competing triplets, including FGD5-AS1, HCP5, HMGN4, TACC3, and so on. LncMiM is available at <https://github.com/xiaofengsong/LncMiM>.

Keywords: lncRNA, ceRNA, competing triplet, LncMiM, ovarian cancer

INTRODUCTION

Non-coding RNAs (ncRNAs) were once considered as junk RNAs; however, evidence has increasingly shown that ncRNAs can perform diverse functions (Slack and Chinnaiyan, 2019; Yao et al., 2019; Chen et al., 2020; Nair et al., 2020). Among ncRNAs, the most intensively studied subclass are microRNAs (miRNAs, usually 19–24 nucleotides long), which can regulate gene expression posttranscriptionally by destabilizing target mRNAs via the RNA-induced silencing complex (RISC) (Bartel, 2009; Gebert and MacRae, 2019). The miRNA-based regulation has been reported to be involved in many pathologies including cancer (Peng and Croce, 2016; Dhawan et al., 2018; Huang et al., 2019). By contrast, the other class of abundant ncRNAs, lncRNAs (>200 nucleotides long), are still less understood, although much larger numbers of lncRNAs have

been identified using high-throughput sequencing techniques in recent years (Fang et al., 2018; Frankish et al., 2019; Volders et al., 2019). Nevertheless, the existing well-characterized lncRNAs have demonstrated their important roles in various critical biological processes, such as chromatin remodeling, genomic splicing, cell proliferation, and cell differentiation (Fatica and Bozzoni, 2014; Han and Chang, 2015; Romero-Barrios et al., 2018; Rossi et al., 2019; Yao et al., 2019). In addition, dysregulation of lncRNAs is implicated in various human diseases (Schmitt and Chang, 2016; Bao et al., 2019; Gao et al., 2019).

Recent studies prove that lncRNAs participate in the posttranscriptional regulation by acting as competing endogenous RNAs (ceRNAs) (Song et al., 2017; He et al., 2019). The lncRNAs that share miRNA response elements (MREs) with mRNAs can compete for miRNA binding, thereby alleviating the inhibitory effect of miRNAs on their mRNA targets. To date, considerable lncRNA-associated competing triplets (lncRNA–miRNA–mRNA) have been reported to be involved in cancer progression (Du et al., 2016; Cong et al., 2019; Wang et al., 2019). For example, the lncRNA *MEG3* functions as a ceRNA of oncogenic miR-181 to regulate gastric cancer progression (Peng et al., 2015, 3). The lncRNA *UCA1* upregulates the expression of *ERBB4* through competitively “sponging” miR-193a–3p and functions as an oncogene in non-small cell lung cancer (NSCLC) (Nie et al., 2016, 1). The *XIST*/miR-92b/*Smad7* triplet is found to play an important role in the progression of hepatocellular carcinoma (Zhuang et al., 2016). Hence, lncRNA associated competing triplets attract more and more attention in cancer research.

At present, several computational methods have been proposed for identifying competing triplets (Le et al., 2017; Hornakova et al., 2018). In general, people usually use linear correlations between gene–gene and/or gene–miRNA pairs to identify ceRNA triplets, since it requires a small sample size and fewer computations (Wang et al., 2015). However, the linear correlation-based methods do not measure the impact of the miRNA on the gene–gene interaction within triplets, resulting in reduced credibility of competing triplet identification results. In order to overcome this problem, several methods based on partial correlation (PC) or conditional mutual information (CMI) have been developed. Among them, two typical methods are often employed: sensitivity correlation and HERMES (Sumazin et al., 2011; Paci et al., 2014). Sensitivity correlation calculates the difference between linear correlation and partial correlation for ceRNA pairs, while HERMES calculates the difference in mutual information for each gene–gene interaction between high and low miRNA expression levels. Despite the constant increase in available methods (Wen et al., 2020), identification of competing triplets through utilizing RNA-seq and miRNA-seq data remains a challenging issue.

With the widespread application of high-throughput sequencing technology, a great deal of data has been accumulated in public databases (Lonsdale et al., 2013; Weinstein et al., 2013). The increasing data lead to more competing triplets identified by the existing methods (Wang et al., 2019); however, they also introduce some new problems needed to be solved. First, the

bigger the data size, the fewer the number of linear correlated miRNA–gene pairs we could find. It seems that the relationship between the expression patterns of miRNA and its target gene is not a linear correlation as assumed by the existing methods. Second, it is noted that competing gene–gene interactions may be regulated by several miRNAs, and thus, the increased data size would make it harder to evaluate the impact of the miRNA on gene–gene interactions by using PC and CMI. In addition, besides the impact of the miRNA on gene pairs, the influence of the gene on the relationship between miRNA and other target genes should be also considered.

Here, for large data sets, we present a powerful method, named LncMiM, to identify lncRNA-associated competing triplets with a new framework addressing the above issues. From the large scale of gene and miRNA expression profiles derived from the TCGA database, 847 competing triplets were identified by using LncMiM, while only a few triplets were identified as competing ones by linear correlation-based methods. The enrichment analysis shows that they are mainly involved in cell proliferation process, cell division process, cell cycle, and ribosome pathways. Among them, 18 competing triplets were found to be associated with prognosis in high-grade serous ovarian cancer. Our method will help in identifying more lncRNA-associated competing triplets in cancer and may contribute to reveal the potential post-transcriptional regulatory mechanism of lncRNAs.

MATERIALS AND METHODS

Data Collection and Pre-processing

As shown in **Figure 1**, paired RNA-seq and miRNA-seq data of ovarian cancer (379 samples from 373 patients) are downloaded from the Cancer Genome Atlas (TCGA) (Weinstein et al., 2013). The RNA-seq data type is “Gene Expression Quantification,” and its workflow type is “HTSeq-FPKM.” The miRNA-seq data type is “Isoform Expression Quantification,” and its workflow type is “BCGSC miRNA Profiling.” The RPM (reads per million mapped reads) value was used to evaluate the expression level of miRNAs. For different samples from the same patient, we merged them by calculating the mean FPKM or RPM value for each lncRNA, mRNA, and miRNA. Finally, we got 376 samples with both the RNA-seq data and miRNA-seq data.

The annotation files of the protein-coding transcripts and the long non-coding transcripts were downloaded from the GENCODE (version 33) database (Harrow et al., 2012). With the transcript annotation, we extracted the mRNA expression data and the lncRNA expression data from the RNA-seq data, and the mRNAs without 3′ UTR annotation were abandoned. Human miRNA sequences and annotation were downloaded from the miRBase (release 22.1) database (Kozomara et al., 2019), and the seed and mature sequences of miRNAs in the miRNA-seq data were both extracted. In order to reduce the computation burden and avoid false-positive identification, we filtered out all the lower expressed RNA (mRNA, lncRNA, and miRNA) based on an artificial criterion. The remaining expressed RNAs need to be satisfied with the following conditions: (a) RNA's expressed value should be >0 in more than 75% of the 376 samples; (b) RNA's expressed value should be >5 in more than 25% of the samples;

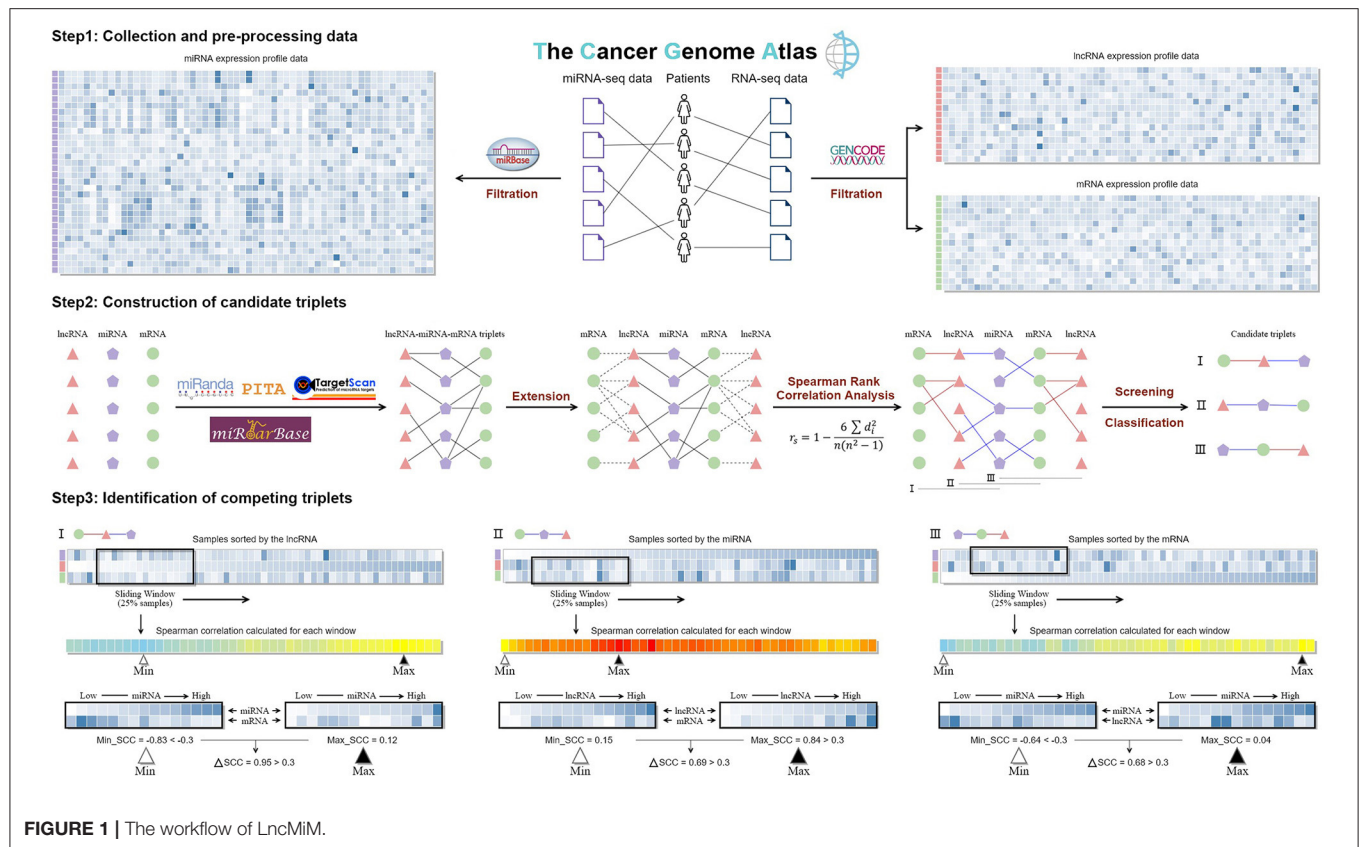


FIGURE 1 | The workflow of LncMiM.

and (c) the expression variation across samples ($\log_2\text{IQR}$) should be >0.58 . As a result, the expression data of 8,076 mRNAs, 225 lncRNAs, and 387 miRNAs were used for further analysis.

Construction of Candidate Triplets

TargetScan, PITA, and miRanda are three commonly used methods to predict miRNA–target interactions (Figure 1). Due to their distinct miRNA–target predicting strategies, these methods are exclusive to any single one alone (Chiu et al., 2015). Thus, TargetScan (version 7.2) (Agarwal et al., 2015), PITA (version 6) (Kertesz et al., 2007), and miRanda (v3.3a) (Miranda et al., 2006) were all applied to predict miRNA–target genes. The parameters of TargetScan and PITA were set to the default values, while the score threshold of miRanda was set to 120 to get a larger miRNA–target gene pool. In addition, the experimentally validated miRNA–target interactions derived from the miRTarBase database (release 8.0) were also added into the miRNA–target gene dataset (Huang et al., 2020).

The lncRNA–miRNA–mRNA triplets were constructed based on the interaction relationship of miRNA–lncRNA and miRNA–mRNA; then the lncRNA and mRNA in each triplet were extracted as lncRNA–mRNA pairs. The Spearman's rank correlation coefficient (SCC) was calculated for the miRNA–lncRNA, miRNA–mRNA, and lncRNA–mRNA pairs to evaluate the regulatory relationships between miRNA, mRNA, and lncRNA in each triplet. Through a rigid screening, only 0.1% pairs were remained as functional interactions, and the cutoff

values for the miRNA–lncRNA, miRNA–mRNA, and lncRNA–mRNA pairs are -0.305 , -0.311 , and 0.520 , respectively. Based on the types of remaining interactions, candidate triplets are grouped into three classes: I, “lncRNA-centered” triplets with miRNA–lncRNA and lncRNA–mRNA interactions; II, “miRNA-centered” triplets with lncRNA–miRNA and miRNA–mRNA interactions; and III, “mRNA-centered” triplets with miRNA–mRNA and mRNA–lncRNA interactions.

Workflow of LncMiM for Identifying Competing Triplets

For identifying competing triplets from the three types of candidate triplets, specific workflows were respectively built to evaluate the centered miRNA, lncRNA, and mRNA on the relationship between the other RNAs (Figure 1). In each workflow, samples were firstly sorted in an ascending order based on the expression of the centered RNA in the candidate triplet. The SCC of the other RNAs was calculated on the samples within the sliding window, whose size is set to 94 (25% of the total samples) and step is set to 1. And then, the maximum and minimum SCCs were calculated. Based on the type of candidate triplets, different filtering criteria were set to identify competing triplets. For the “lncRNA-centered” and “mRNA-centered” triplets, their minimum SCC should be <-0.311 and -0.305 , respectively. For the “miRNA-centered” triplets, their maximum SCC should be more than 0.520 . In addition, the difference between the maximum and minimum

SCC should be >0.300 . Finally, all the candidate triplets meeting their corresponding filtering criteria were identified as competing triplets.

In addition, to assess the statistical significance of the correlation coefficient difference (ΔCor), a series of null hypotheses were tested by measuring the ΔCor distribution over random conditions. That is, for each candidate triplet, two non-overlapping sample subsets were randomly chosen from the complete dataset, rather than based on the expression of miRNA, and then the correlation coefficient and ΔCor were calculated for these two random sample subsets. This process was repeated 100 times. The p -value is defined as the fraction of ΔCor in random condition that was larger than that on the specified conditions mentioned above; p -values were Bonferroni-corrected for the total number of candidate triplets. The triplets with adjusted p -values <0.01 are statistically significant.

Functional and Survival Analysis of the Competing Triplet

With the competing triplets, the integrative regulatory network was built and visualized by Cytoscape (Shannon et al., 2003). The size of the node and the width of the line are determined by the number of competing triplets containing them. The circular layout was produced by using the yFiles layout Algorithms. DAVID 6.8 (<https://david.ncifcrf.gov>) was used to perform the enrichment analysis of biological processes and KEGG pathways (Huang et al., 2009). For the enriched biological process terms, their adjusted p -values should be <0.05 .

The clinical profiles of 373 patients with high-grade serous ovarian cancer were downloaded from the TCGA database. The patients' ID, age at initial pathologic diagnosis, vital status, days to death, days to last follow-up, neoplasm histologic grade, and clinical stage were extracted from the clinical profiles. Based on data integrity, 369 patients' clinical data were screened out for the following survival analysis. The days to death together with the days to last follow-up make up the overall survival time of patients. Both the single variate and multivariate survival analyses used the Cox proportional hazards (PH) regression. In addition, to investigate the impact of specific genes on the survival time, patients were classified into different groups through four ways based on their expression levels. The survival analysis and visualization were performed by using the "survminer" R package.

RESULTS

Investigation of the Expression Relationship Between miRNA and Target Gene

In general, miRNAs are assumed to be linearly correlated with their target genes. Thus, the Pearson correlation coefficient (PCC) was initially used to identify negatively correlated miRNA-mRNA and miRNA-lncRNA pairs. With the threshold of -0.30 , from the 74,086 miRNA-lncRNA pairs and 2,608,237 miRNA-mRNA pairs (Figures 2A,E), only 3 miRNA-lncRNA pairs and 443 miRNA-mRNA pairs were found to be negatively

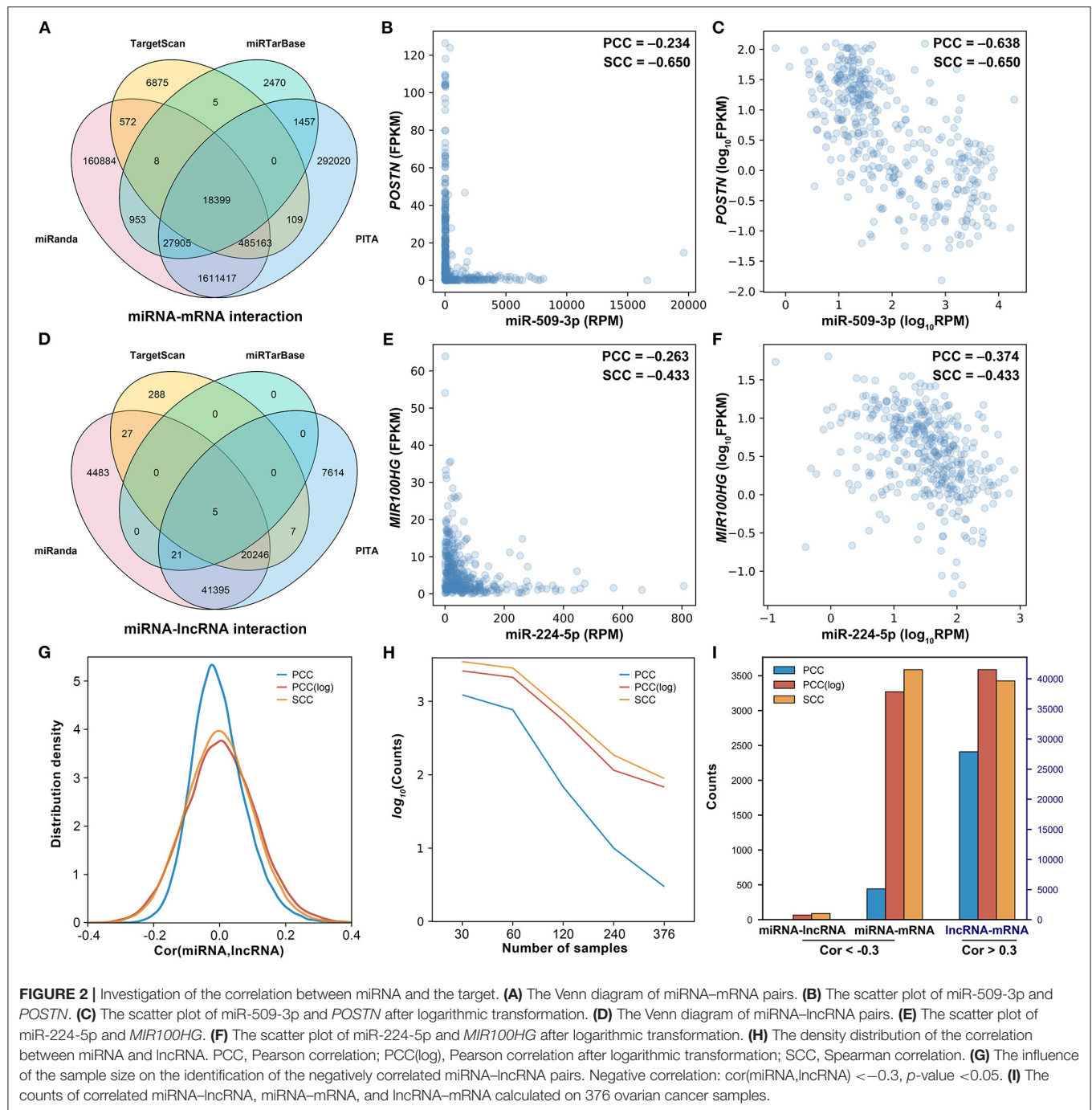
correlated, which are far less than expected. As shown in Figure 2B, there is a negative regulatory relationship between miR-509-3p and *POSTN*, but the PCC is only -0.234 . Similarly, miR-224-5p is also shown to be negatively correlated with *MIR100HG*; their PCC is -0.263 (Figure 2E). If the expression values were normalized by a logarithmic transformation, however, the PCCs of miR-509-3p-*POSTN* and miR-224-5p-*MIR100HG* change to -0.638 and -0.374 , respectively (Figures 2C,F). As shown in Figures 2G,I, after the logarithmic transformation, more negatively correlated miRNA-target gene pairs were detected. In addition, with the increase in the sample size, the number of negatively correlated miRNA-lncRNA pairs ($\text{PCC} < -0.3$, $P\text{-value} < 0.05$) significantly decreases (Figure 2H). These results implied that PCC is not appropriate for the evaluation of the regulatory relationship between miRNA and target gene, especially for large sample data.

Here, we assumed that the relationship between miRNA and the target is not linear. As shown in Figures 2B,C,E,F, as compared with the PCC, the SCC is more accurate for assessing the relationship between miRNA and the target. In addition, the SCC is less affected by the sample size (Figure 2H) and can detect more negatively correlated miRNA-target gene pairs (Figure 2I). Thus, the SCC was used to screen negatively correlated miRNA-target pairs. From the 74,086 miRNA-lncRNA pairs and 2,608,237 miRNA-mRNA pairs, only 0.1% were respectively screened out as the negatively correlated miRNA-target pairs. A total of 72 negatively correlated miRNA-lncRNA pairs and 2,608 negatively correlated miRNA-mRNA pairs were selected, respectively, with the thresholds -0.311 and -0.305 . Besides the miRNA-target pairs, with threshold 0.52, 1,806 positively correlated mRNA-lncRNA pairs were screened out from 1,816,605 candidate mRNA-lncRNA pairs.

Investigation of the Impact on Pairwise Interaction by the Other One in Triplets

With the strictly selected negatively and positively correlated interactions, 256 competing triplets can be found by using the traditional strategy. If a miRNA is negatively correlated to two positively correlated target genes, then they form a competing triplet. As this traditional strategy ignores the mediating effect of miRNA on the positive relationship between target genes, several competing triplets may be fake ones. For example, miR-185-3p is negatively correlated to the two positively correlated target genes (Figures 3A-C); however, the positive correlation between SNHG29 and RPLP0 is not related to miR-185-3p (Figure 3D). According to the ceRNA hypothesis, SNHG29-miR-185-3p-RPLP0 is a fake competing triplet. Thus, the impact of miRNA on the interaction between ceRNA pairs should be considered.

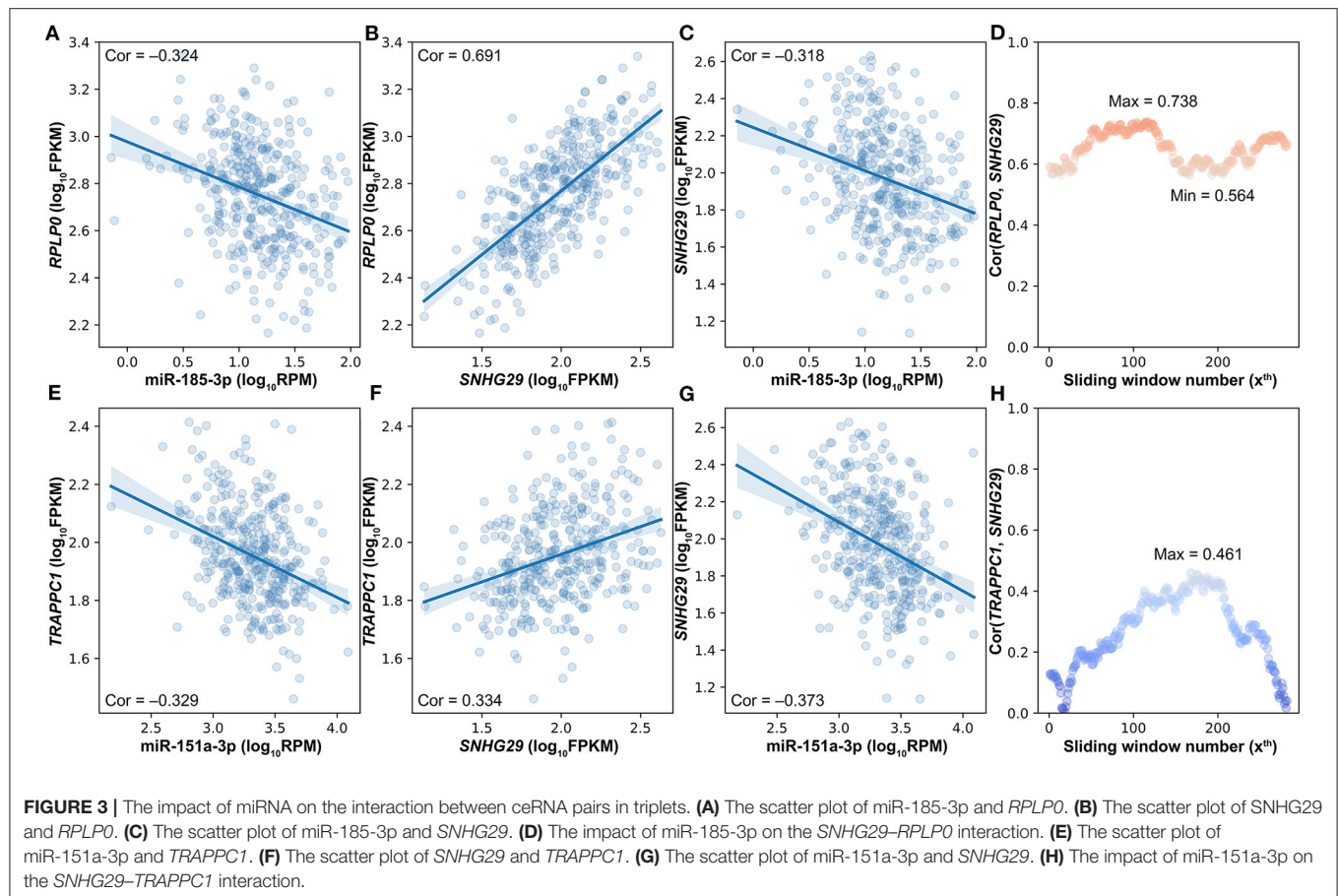
To determine whether the interaction between target genes is derived from their relationship with miRNA, a commonly used method is to compare the correlation coefficients of target gene pairs under conditions of high and low miRNA expression levels. Accordingly, the differences of lncRNA-mRNA pairs' SCCs on the first and last quarter of samples sorted by miRNA expression were calculated, and 15 of the 256 competing triplets were identified to be true. A hidden hypothesis of this method is that



the strength of the interaction between lncRNA and mRNA is linearly correlated with the miRNA expression level. However, according to the ceRNA hypothesis, both extremely high and extremely low miRNA expressions would impair the interaction between ceRNA pairs and even make them unrelated with each other. For example, miR-151a-3p is negatively correlated to the two positively correlated target genes (Figures 3E-G). The SCC between *TRAPPC1* and *SNHG29* is not linearly correlated with the expression level of miR-151a-3p (Figure 3H). The SCC achieves the maximum value at about the median

miRNA expression level. Therefore, in LncMiM, all the miRNA expression levels, rather than only the highest and lowest ones, are considered when evaluating the impact of miRNA on the interaction between target gene pairs.

Besides the impact of miRNA on the lncRNA-mRNA interaction, lncRNA and mRNA can also affect the miRNA-target interactions. As shown in Figure 4, miR-151a-3p is negatively correlated to the two positively correlated target genes (*RPS6* and *SNHG29*). The SCC between *RPS6* and *SNHG29* is significantly changed with the rise of miR-151a-3p expression levels



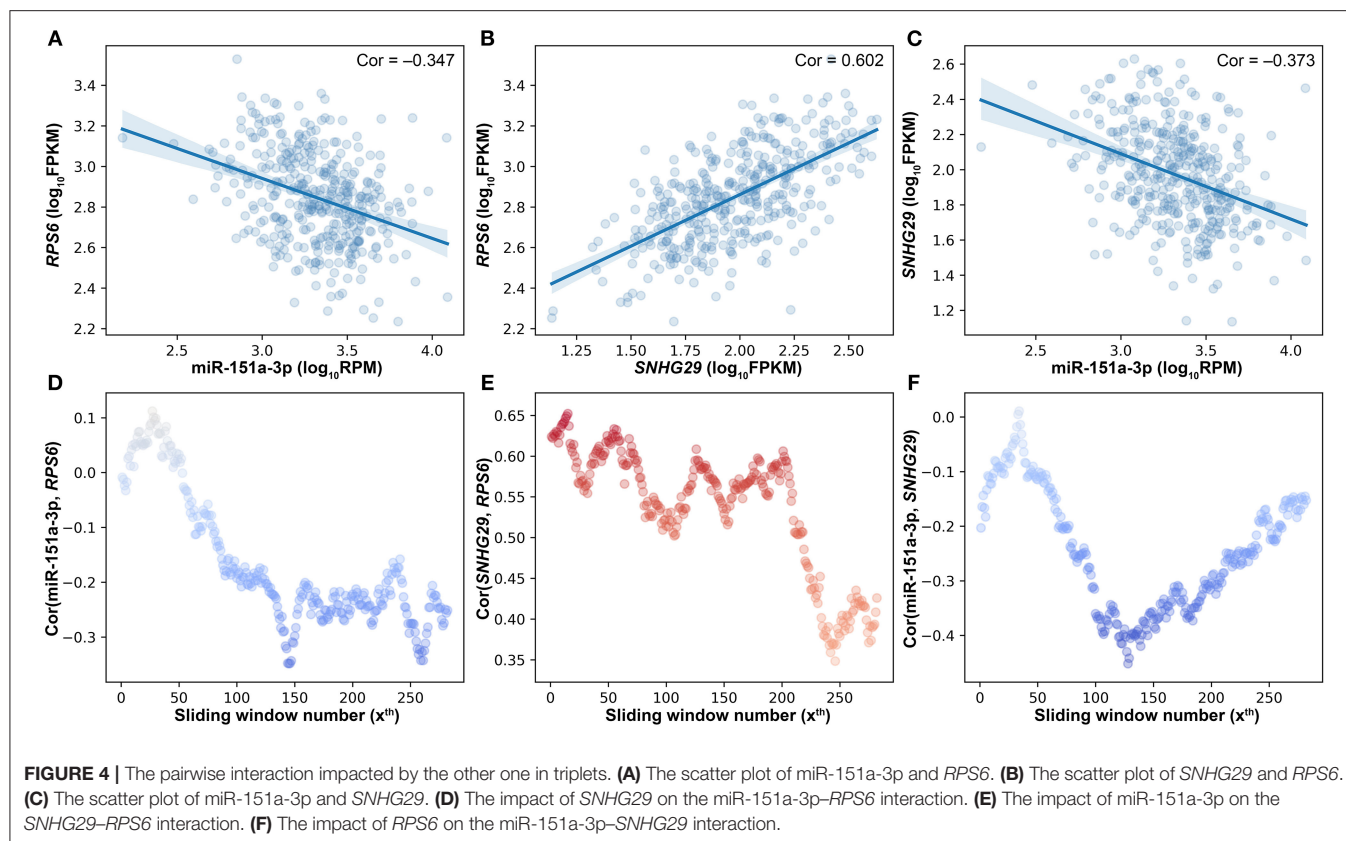
(Figure 4E). Moreover, the correlation relationship between *RPS6* and miR-151a-3p is impacted by the *SNHG29* (Figure 4D), and the interaction between *SNHG29* and miR-151a-3p is influenced by the *RPS6* (Figure 4F). As the pairwise interactions are impacted by the other one in the triplets, it is not enough to assess the real relationship between each pair only based on their own expression profiles, especially when the sample size is very large. The triplet with two correlated pairs may also be a competing triplet; thus, three types of candidate triplets were analyzed in LncMiM. Using the selected miRNA–target and lncRNA–mRNA pairs, 2060 “miRNA-centered” triplets, 1944 “lncRNA-centered” triplets, and 1537 “mRNA-centered” triplets were assembled. By using LncMiM, 231 “miRNA-centered” triplets, 339 “lncRNA-centered” triplets, and 439 “mRNA-centered” triplets were identified as competing triplets (Supplementary Table 1). In total, 847 competing triplets were found, including 38 miRNAs, 36 lncRNAs, and 236 mRNAs.

Functional Analysis of the lncRNA-Associated Competing Triplets in Ovarian Cancer

In the competing triplets, a considerable number of lncRNAs, miRNAs, and mRNAs have been reported to be associated with

ovarian cancer. By searching related literature and databases, about 30% lncRNAs have been verified to play roles in the regulation of proliferation, invasion, and migration of ovarian cancer cells, including *ZFAS1*, *SNHG1*, *GAS5*, *EMX20S*, *GIHCG*, *TP53TG1*, *EPB41L4A-AS1*, *SNHG8*, *SNHG6*, and *HCP5* (Zhan et al., 2018; Gao et al., 2019; Wu et al., 2019; Miao et al., 2020; Wang et al., 2020). In addition, some lncRNAs (e.g., *SNHG29*, *FGD5-AS1*, *TRIM52-AS1*, *EPB41L4A-AS1*, *RNASEH1-AS1*, *SNHG7*, *SPINT1-AS1*, *MAPKAPK5-AS1*, and *PITPNA-AS1*) are reported to be involved in other types of cancers (Wang et al., 2018; Gao et al., 2019, 2; Han et al., 2019; Zhou et al., 2020). Through retrieving the miRCancer database (version june2020) (Xie et al., 2013), 60.5% miRNAs in the competing triplets have been proved to be associated with ovarian cancer. In the mRNAs, 29 ovarian cancer oncogenes were found by searching the OCGene database (Liu et al., 2015). These results indicate that the lncRNA-associated competing triplets play important roles in the progression of ovarian cancer.

To analyze the regulatory relationship between lncRNA, miRNA, and mRNA in ovarian cancer, a comprehensive network was established through combining the 847 lncRNA-associated competing triplets (Figure 5A). In the network, 310 nodes are connected by 1,182 edges, including 132 miRNA–lncRNA edges, 539 miRNA–mRNA edges, and 511



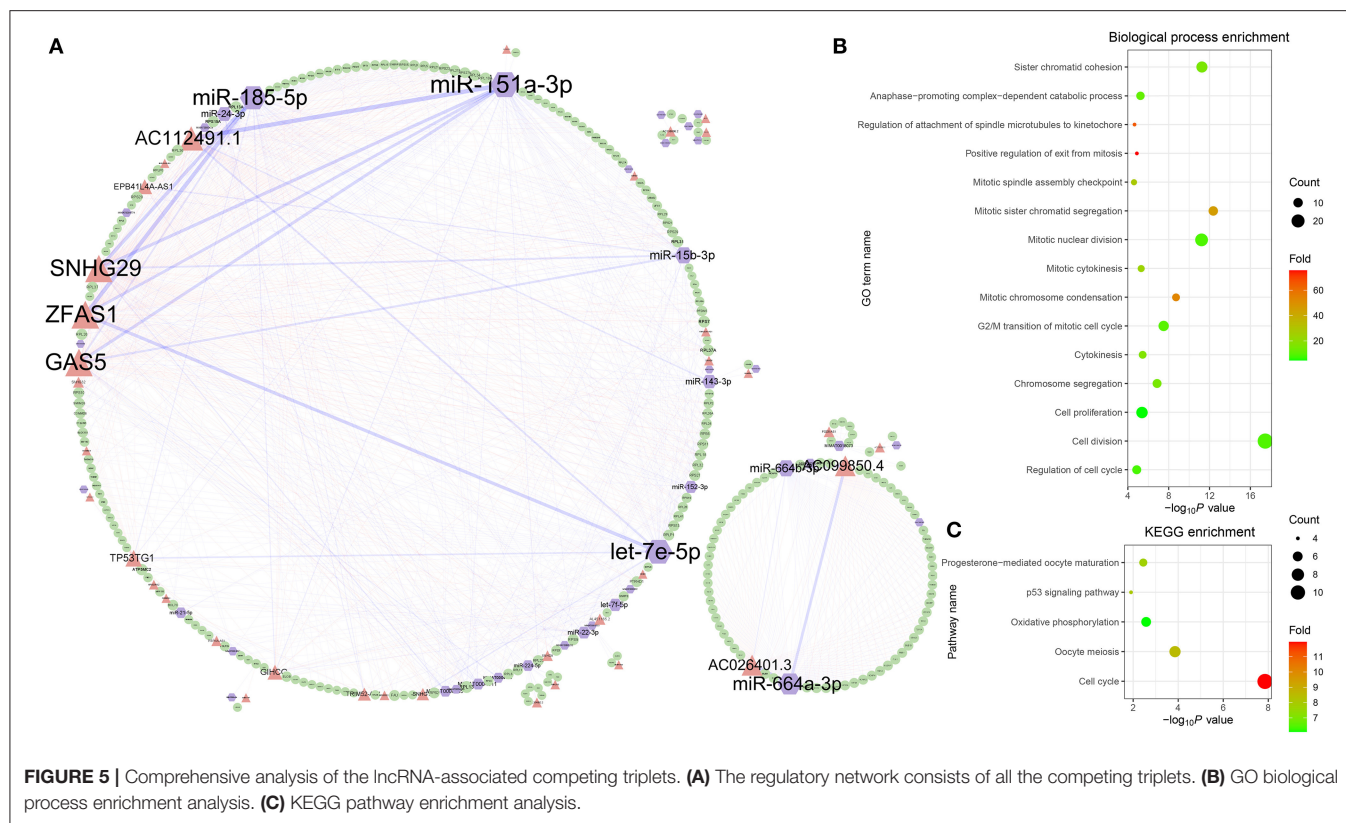
lncRNA–mRNA edges. Among them, the top 10 nodes with most connections are miR-151a-3p, *ZFAS1*, *SNHG29*, miR-185-5p, *GAS5*, *AC112491.1*, let-7e-5p, miR-664a-3p, *AC099850.4*, and miR-15b-3p. The top 10 edges connected with most nodes are miR-151a-3p–*AC112491.1*, miR-185-5p–*ZFAS1*, miR-185-5p–*SNHG29*, miR-151a-3p–*GAS5*, let-7e-5p–*ZFAS1*, miR-664a-3p–*AC026401.3*, miR-151a-3p–*SNHG29*, miR-151a-3p–*ZFAS1*, miR-664b-3p–*AC099850.4*, and miR-15b-3p–*GAS5*. Based on the connections, the nodes are divided into two groups. The small group is mainly regulated by the miR-664a-3p and *AC026401.3* pair, while the ribosome protein-related mRNAs are all located in the large group.

Among the mRNAs, there are 39 *RPL* and 27 *RPS* genes, which indicates that the triplets are involved in the ribosome biogenesis. Except the *RPs*, the GO biological process enrichment analysis of the other genes shows that the competing triplets are also involved in cell division, cell proliferation, regulation of cell cycle, anaphase-promoting complex-dependent catabolic process, cytokinesis, chromosome segregation, and other nine processes related with cell mitosis (Figure 5B). In addition, the competing triplets are found to be mainly enriched in five KEGG pathways, including cell cycle, oocyte meiosis, oxidative phosphorylation, p53 signaling pathway, and progesterone-mediated oocyte maturation (Figure 5C). All the results suggest that the lncRNA-associated competing triplets mediate ovarian cancer progression through regulating ribosome biogenesis, cell cycle, cell division, and cell proliferation, and they may be

associated with survival in patients with high-grade serous ovarian cancer.

Identification of Potential Prognostic Competing Triplets

The Cox PH analysis was used to identify survival time associated miRNAs, mRNAs, and lncRNAs in the competing triplets. The result of univariate Cox PH analysis indicates that the lncRNA *FGD5-AS1* ($p = 0.0008$) is a potential prognostic biomarker for all patients with ovarian cancer. For patients in grade G2, *C12orf45*, *NDUFB8*, *POLR2J*, *SNRPE*, and *SNRPF* are found to be associated with survival time ($p < 0.001$). By multivariate analysis with patient age at diagnosis, more potential prognostic biomarkers are found, including *FGD5-AS1*, *GABPA*, *MRPS27*, *NR1D2*, and *NR2C2*. For patients in grade G2, only *SNRPF* is related to the survival time with the diagnosis age. For patients in grade G3, *FGD5-AS1*, *LETMD1*, *MAPKAPK5-AS1*, *MRPS27*, and *SDHC* are screened out as prognostic biomarkers with the diagnosis age. *FGD5-AS1* and *MRPS27* are found to be associated with the survival time of patients in stage IIIC, while *B9D1*, *RNASEH1-AS1*, *SPINT1-AS1*, and *ZWINT* are associated with the survival time of patients in stage IV. The association between the survival time and the triplet was also evaluated by using multivariate Cox PH analysis. With the threshold $p < 0.001$, miR-224-5p/*AL354892.2/ZBTB12* is found to be survival associated competing triplets. Considering the age of the patient at the initial pathologic diagnosis, 18 competing triplets are found to



be associated with the overall survival time of patients in ovarian cancer, including miR-224-5p/*AL354892.2/ZBTB12*, miR-3653-3p/*FGD5-AS1/NR1D2*, miR-224-5p/*AC112491.1/NDUFB8*, and so on (Supplementary Table 2).

In addition, the Kaplan–Meier survival analysis was also performed to evaluate the potential prognostic power of miRNAs, lncRNAs, and mRNAs in the competing triplets. Considering the large data size, for each gene, the tumor samples were divided into two or three groups according to their expression levels by four ways (Figure 6A). By different grouping modes, a total of 107 RNAs are found to be associated with survival time, including 13 miRNAs, 10 lncRNAs, and 84 mRNAs (Supplementary Table 3). As show in Figure 6B, each grouping mode has its unique results. Especially, the grouping mode b has the least common results with the other modes, which indicates that there is a more complicated relationship between the patient survival time and the gene expression value. For each grouping mode, the most significant genes are *HMGH4*, *TACC3*, *RNF111*, and *VGLL4* (Figures 6B,C,E,F). The survival associated genes are involved in 368 competing triplets, which are found to be enriched in cell division, cell proliferation, ribosome, and cell cycle.

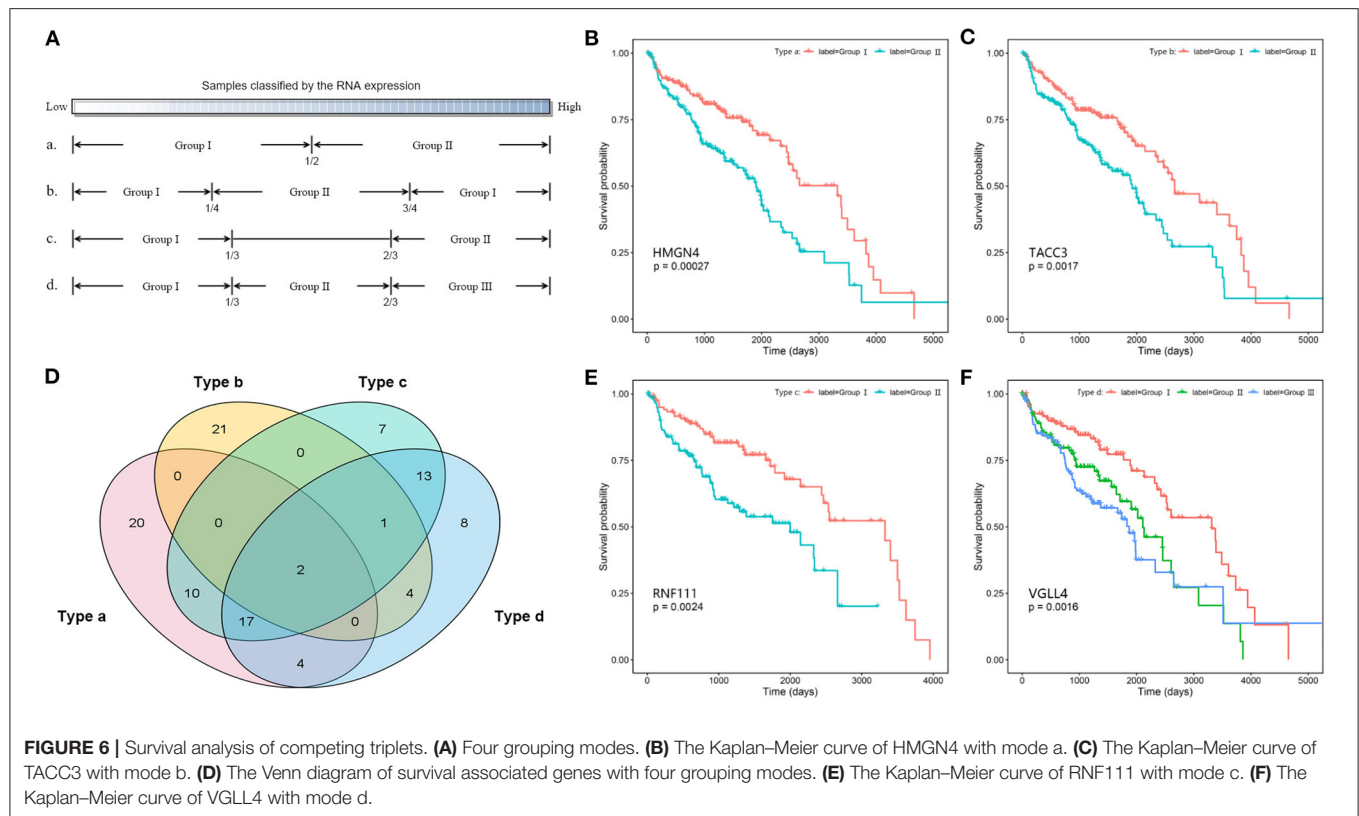
DISCUSSION

In this study, TargetScan, PITA, miRanda, and miRTarBase were used together to predict miRNA–target pairs, and a total of 2,608,237 miRNA–mRNA and 74,086 miRNA–lncRNA

interactions were found (Supplementary Table 4). As shown in Figures 2A,D, each tool exclusively predicted a fraction of miRNA–target interactions. Although a vast number of miRNA–target interactions were predicted by TargetScan, PITA, and miRanda, there are still several experimentally validated miRNA–target interactions predicted by none of these tools. miRNA–mRNA pairs together with miRNA–lncRNA pairs could construct a huge number of triplets ($\sim 1.69\text{E}+9$). Considering the computation and time cost, miRNA–target pairs were firstly filtered by correlation relationships.

Through the miRNA–target relationship, 1,816,605 indirect interactions between mRNA and lncRNA were established. Based on the linear relationship calculated by the PCC, 3 negatively correlated miRNA–lncRNA pairs ($PCC = -0.3$), 443 negatively correlated miRNA–mRNA pairs ($PCC = -0.3$), and 27,897 positively correlated lncRNA–mRNA pairs ($PCC > 0.3$) were screened out. With the linearly correlated pairs, 64 competing triplets were established. The impact of miRNA on the linear relationship between lncRNA and mRNA was only found in seven competing triplets. In contrast, based on the non-linear relationship assessed by the SCC, 89 negatively correlated miRNA–lncRNA pairs ($SCC = -0.3$), 3,586 negatively correlated miRNA–mRNA pairs ($SCC < -0.3$), and 33,267 positively correlated lncRNA–mRNA pairs ($SCC > 0.3$) were screened out. Comparing with the PCC, more negatively correlated miRNA–target pairs are found by the SCC.

In most of the scatter plots of the negatively correlated miRNA–target pairs, the points are mainly located in the



bottom left corner, which looks like a triangle other than a line (**Figure 2E**). By comparison, after normalizing expression values by a logarithmic transformation, the points become more dispersed and scatter around a line. This result indicates that the linear correlation between miRNA and the target is impacted by the large span of the expression values, which is brought by the large sample size. In addition, the different orders of magnitude of the expression value between miRNA and the target gene are also an impact factor. The expression value of miRNA is calculated by RPM (max value: 8.23E5), while the expression values of mRNA and lncRNA are calculated by FPKM (max value of mRNA: 2.15E4, max value of lncRNA: 1.85E3). Therefore, it is better to assess the relationship between miRNA and the target by the non-linear correlation, especially on the large scale of data.

The bigger the size of the patient data, the more complex the relationships between ceRNAs we can observe. According to the ceRNA hypothesis, the strength of the competing relationship between ceRNAs is not constant but depends on the amount of miRNA (**Figures 3H, 4E**). Similarly, the strength of the interaction between miRNA and ceRNA is also impacted by the amount of the other ceRNA (**Figures 4D,F**). In 231 competing triplets, miRNAs are negatively correlated to the mRNAs and lncRNAs. Although the positive correlation between mRNA and lncRNA is not significant on the whole samples, their correlation is changed with the expression level of miRNA, and a significant positive correlation can be observed on a specific subset of samples. In 778 competing triplets, the negative correlation between miRNA and ceRNA is not significant on the whole

samples, but there is a significant negative correlation on a specific subset of samples, and the correlation is influenced by the other ceRNA. Thus, besides the impact of miRNA on the interaction between ceRNAs, the impact of ceRNA on the correlation between miRNA and other ceRNAs should also be considered.

However, there is still no method considering the impact of both the miRNA and the ceRNAs when identifying competing triplets. The method, sensitivity partial Pearson correlation (SPPC), only estimates the impact (sensitivity) of miRNA on the interactions between ceRNAs (Paci et al., 2014). However, when using SPPC on “miRNA-centered” candidate triplets, no competing triplets were identified. JAMI is a conditional mutual information-based method, which can only estimate the impact of ceRNA on the interaction between miRNA and other ceRNAs (Hornakova et al., 2018). With JAMI, 87 competing triplets were filtered out from 1,507 “mRNA-centered” candidate triplets, and 385 competing triplets were identified from 1,944 “lncRNA-centered” candidate triplets. The JAMI results only show that the centered ceRNA has a significant influence on the relationship between miRNA and the other RNA in a candidate triplet, but it is still unknown if the other RNA is a ceRNA that should be negatively correlated with miRNA. In addition, the SNHG29/miR-151a-3p/RPS6 competing triplet is not identified by JAMI (**Figure 4**).

Considering the drawbacks of the existing tool, we developed a novel method named LncMiM to identify lncRNA-associated competing triplets in ovarian cancer. Besides the impact

of miRNA on the interaction between ceRNA pairs, the impact of ceRNA on the interaction between miRNA and the other ceRNA is also used to identify competing triplets. As compared with other tools, LncMiM shows better performance (**Supplementary Table 5**). By using LncMiM, 231 competing triplets were identified from 2,060 “miRNA-centered” candidate triplets, 339 competing triplets were identified from 1,944 “lncRNA-centered” triplets, and 439 competing triplets were identified from 1,507 “mRNA-centered” triplets. In final, a total of 847 lncRNA-associated competing triplets were found. The functional enrichment analysis shows that the competing triplets are mainly involved in cell division, cell proliferation, and regulation of cell cycle. The KEGG pathway analysis shows that they are associated with ribosome, cell cycle, oocyte meiosis, oxidative phosphorylation, p53 signaling pathway, and progesterone-mediated oocyte maturation. Among them, 18 competing triplets are found to be significantly correlated with the overall survival in ovarian cancer.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Materials**, further inquiries can be directed to the corresponding authors.

REFERENCES

- Agarwal, V., Bell, G. W., Nam, J.-W., and Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife* 4:e05005. doi: 10.7554/eLife.05005.028
- Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* 47, D1034–D1037. doi: 10.1093/nar/gky905
- Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136, 215–233. doi: 10.1016/j.cell.2009.01.002
- Chen, J., Brunner, A.-D., Cogan, J. Z., Nuñez, J. K., Fields, A. P., Adamson, B., et al. (2020). Pervasive functional translation of noncanonical human open reading frames. *Science* 367, 1140–1146. doi: 10.1126/science.aay0262
- Chiu, H.-S., Llobet-Navas, D., Yang, X., Chung, W.-J., Ambesi-Impimbatto, A., Iyer, A., et al. (2015). Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. *Genome Res.* 25, 257–267. doi: 10.1101/gr.178194.114
- Cong, Z., Diao, Y., Xu, Y., Li, X., Jiang, Z., Shao, C., et al. (2019). Long non-coding RNA linc00665 promotes lung adenocarcinoma progression and functions as ceRNA to regulate AKR1B10-ERK signaling by sponging miR-98. *Cell Death Dis.* 10, 1–15. doi: 10.1038/s41419-019-1361-3
- Dhawan, A., Scott, J. G., Harris, A. L., and Buffa, F. M. (2018). Pan-cancer characterisation of microRNA across cancer hallmarks reveals microRNA-mediated downregulation of tumour suppressors. *Nat. Commun.* 9:5228. doi: 10.1038/s41467-018-07657-1
- Du, Z., Sun, T., Hacisuleyman, E., Fei, T., Wang, X., Brown, M., et al. (2016). Integrative analyses reveal a long noncoding RNA-mediated sponge regulatory network in prostate cancer. *Nat. Commun.* 7:10982. doi: 10.1038/ncomms10982
- Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., et al. (2018). NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* 46, D308–D314. doi: 10.1093/nar/gkx1107
- Fatica, A., and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nat. Rev. Genet.* 15, 7–21. doi: 10.1038/nrg3606
- Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., et al. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* 47, D766–D773. doi: 10.1093/nar/gky955
- Gao, Y., Wang, P., Wang, Y., Ma, X., Zhi, H., Zhou, D., et al. (2019). Lnc2Cancer v2.0: updated database of experimentally supported long non-coding RNAs in human cancers. *Nucleic Acids Res.* 47, D1028–D1033. doi: 10.1093/nar/gky1096
- Gebert, L. F. R., and MacRae, I. J. (2019). Regulation of microRNA function in animals. *Nat. Rev. Mol. Cell Biol.* 20, 21–37. doi: 10.1038/s41580-018-0045-7
- Han, L., Li, Z., Jiang, Y., Jiang, Z., and Tang, L. (2019). SNHG29 regulates miR-223-3p/CTNND1 axis to promote glioblastoma progression via Wnt/β-catenin signaling pathway. *Cancer Cell Int.* 19:345. doi: 10.1186/s12935-019-1057-x
- Han, P., and Chang, C.-P. (2015). Long non-coding RNA and chromatin remodeling. *RNA Biol.* 12:1094. doi: 10.1080/15476286.2015.1063770
- Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., et al. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22, 1760–1774. doi: 10.1101/gr.135350.111
- He, R.-Z., Luo, D.-X., and Mo, Y.-Y. (2019). Emerging roles of lncRNAs in the post-transcriptional regulation in cancer. *Genes Dis.* 6, 6–15. doi: 10.1016/j.gendis.2019.01.003
- Hornakova, A., List, M., Vreeken, J., and Schulz, M. H. (2018). JAMI: fast computation of conditional mutual information for ceRNA network analysis. *Bioinformatics* 34, 3050–3051. doi: 10.1093/bioinformatics/bty221
- Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi: 10.1093/nar/gkn923
- Huang, H.-Y., Lin, Y.-C.-D., Li, J., Huang, K.-Y., Shrestha, S., Hong, H.-C., et al. (2020). miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. *Nucleic Acids Res.* 48, D148–D154. doi: 10.1093/nar/gkz896
- Huang, Z., Shi, J., Gao, Y., Cui, C., Zhang, S., Li, J., et al. (2019). HMDD v3.0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res.* 47, D1013–D1017. doi: 10.1093/nar/gky1010
- Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat. Genet.* 39, 1278–1284. doi: 10.1038/ng2135

AUTHOR CONTRIBUTIONS

JZ developed the LncMiM method, identified competing triplets, and wrote the manuscript. TX and QY collected and calculated the RNA expression profiles of ovarian cancer. JL and BJ identified the potential prognostic triplets. JW and XS conceived the study, supervised the work, manuscript writing, and editing. All authors read and approved the final manuscript.

FUNDING

This study was supported by grants from China Postdoctoral Science Foundation (No. 2019M661817), the National Natural Science Foundation of China (No. 61973155, No. 61901225, No. 62003165), and Fundamental Research Funds for the Central Universities (No. NP2018109). This paper is recommended by the 5th Computational Bioinformatics Conference.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.607722/full#supplementary-material>

- Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. *Nucleic Acids Res.* 47, D155–D162. doi: 10.1093/nar/gky1141
- Le, T. D., Zhang, J., Liu, L., and Li, J. (2017). Computational methods for identifying miRNA sponge interactions. *Brief. Bioinformatics* 18, 577–590. doi: 10.1093/bib/bbw042
- Liu, Y., Xia, J., Sun, J., and Zhao, M. (2015). OCGene: a database of experimentally verified ovarian cancer-related genes with precomputed regulation information. *Cell Death Dis.* 6:e2036. doi: 10.1038/cddis.2015.380
- 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
- Miao, W., Lu, T., Liu, X., Yin, W., and Zhang, H. (2020). LncRNA SNHG8 induces ovarian carcinoma cells cellular process and stemness through Wnt/ β -catenin pathway. *Cancer Biomark.* 28, 459–471. doi: 10.3233/CBM-190640
- Miranda, K. C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., et al. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 126, 1203–1217. doi: 10.1016/j.cell.2006.07.031
- Nair, L., Chung, H., and Basu, U. (2020). Regulation of long non-coding RNAs and genome dynamics by the RNA surveillance machinery. *Nat. Rev. Mol. Cell Biol.* 21, 123–136. doi: 10.1038/s41580-019-0209-0
- Nie, W., Ge, H., Yang, X., Sun, X., Huang, H., Tao, X., et al. (2016). LncRNA-UCA1 exerts oncogenic functions in non-small cell lung cancer by targeting miR-193a-3p. *Cancer Lett.* 371, 99–106. doi: 10.1016/j.canlet.2015.11.024
- Paci, P., Colombo, T., and Farina, L. (2014). Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. *BMC Syst. Biol.* 8:83. doi: 10.1186/1752-0509-8-83
- Peng, W., Si, S., Zhang, Q., Li, C., Zhao, F., Wang, F., et al. (2015). Long non-coding RNA MEG3 functions as a competing endogenous RNA to regulate gastric cancer progression. *J. Exp. Clin. Cancer Res.* 34:79. doi: 10.1186/s13046-015-0197-7
- Peng, Y., and Croce, C. M. (2016). The role of MicroRNAs in human cancer. *Signal Transduc. Target. Therapy* 1, 1–9. doi: 10.1038/sigtrans.2015.4
- Romero-Barrios, N., Legascue, M. F., Benhamed, M., Ariel, F., and Crespi, M. (2018). Splicing regulation by long noncoding RNAs. *Nucleic Acids Res.* 46, 2169–2184. doi: 10.1093/nar/gky095
- Rossi, M., Bucci, G., Rizzotto, D., Bordo, D., Marzi, M. J., Puppo, M., et al. (2019). LncRNA EPR controls epithelial proliferation by coordinating Cdkn1a transcription and mRNA decay response to TGF- β . *Nat. Commun.* 10:1969. doi: 10.1038/s41467-019-09754-1
- Schmitt, A. M., and Chang, H. Y. (2016). Long noncoding RNAs in cancer pathways. *Cancer Cell* 29, 452–463. doi: 10.1016/j.ccell.2016.03.010
- 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
- Slack, F. J., and Chinnaiyan, A. M. (2019). The role of non-coding RNAs in oncology. *Cell* 179, 1033–1055. doi: 10.1016/j.cell.2019.10.017
- Song, Y., Sun, J., Zhao, J., Yang, Y., Shi, J., Wu, Z., et al. (2017). Non-coding RNAs participate in the regulatory network of CLDN4 via ceRNA mediated miRNA evasion. *Nat. Commun.* 8:289. doi: 10.1038/s41467-017-00304-1
- Sumazin, P., Yang, X., Chiu, H.-S., Chung, W.-J., Iyer, A., Llobet-Navas, D., et al. (2011). An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 147, 370–381. doi: 10.1016/j.cell.2011.09.041
- Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdaghe, P., et al. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res.* 47, D135–D139. doi: 10.1093/nar/gky1031
- Wang, L., He, M., Fu, L., and Jin, Y. (2020). Role of lncRNA HCP5/microRNA-525-5p/PRC1 crosstalk in the malignant behaviors of ovarian cancer cells. *Exp. Cell Res.* 394:112129. doi: 10.1016/j.yexcr.2020.112129
- Wang, P., Li, X., Gao, Y., Guo, Q., Wang, Y., Fang, Y., et al. (2019). LncACTdb 2.0: an updated database of experimentally supported ceRNA interactions curated from low- and high-throughput experiments. *Nucleic Acids Res.* 47, D121–D127. doi: 10.1093/nar/gky1144
- Wang, P., Ning, S., Zhang, Y., Li, R., Ye, J., Zhao, Z., et al. (2015). Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer. *Nucleic Acids Res.* 43, 3478–3489. doi: 10.1093/nar/gkv233
- Wang, Y., Yang, F., and Zhuang, Y. (2018). Identification of a progression-associated long non-coding RNA signature for predicting the prognosis of lung squamous cell carcinoma. *Exp. Ther. Med.* 15, 1185–1192. doi: 10.3892/etm.2017.5571
- 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
- Wen, X., Gao, L., and Hu, Y. (2020). LACEModule: identification of competing endogenous rna modules by integrating dynamic correlation. *Front. Genet.* 11:235. doi: 10.3389/fgene.2020.00235
- Wu, Y., Deng, Y., Guo, Q., Zhu, J., Cao, L., Guo, X., et al. (2019). Long non-coding RNA SNHG6 promotes cell proliferation and migration through sponging miR-4465 in ovarian clear cell carcinoma. *J. Cell. Mol. Med.* 23, 5025–5036. doi: 10.1111/jcmm.14359
- Xie, B., Ding, Q., Han, H., and Wu, D. (2013). miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* 29, 638–644. doi: 10.1093/bioinformatics/btt014
- Yao, R.-W., Wang, Y., and Chen, L.-L. (2019). Cellular functions of long noncoding RNAs. *Nat. Cell Biol.* 21, 542–551. doi: 10.1038/s41556-019-0311-8
- Zhan, L., Li, J., and Wei, B. (2018). Long non-coding RNAs in ovarian cancer. *J. Exp. Clin. Cancer Res.* 37:120. doi: 10.1186/s13046-018-0793-4
- Zhou, C., Chen, Z., Peng, C., Chen, C., and Li, H. (2020). Long noncoding RNA TRIM52-AS1 sponges miR-514a-5p to facilitate hepatocellular carcinoma progression through increasing MRPS18A. *Cancer Biother. Radiopharm.* doi: 10.1089/cbr.2019.3271. [Epub ahead of print].
- Zhuang, L. K., Yang, Y. T., Ma, X., Han, B., Wang, Z. S., Zhao, Q. Y., et al. (2016). MicroRNA-92b promotes hepatocellular carcinoma progression by targeting Smad7 and is mediated by long non-coding RNA XIST. *Cell Death Dis.* 7:e2203. doi: 10.1038/cddis.2016.100

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.

Copyright © 2021 Zhao, Song, Xu, Yang, Liu, Jiang and Wu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Research on Components Assembly Platform of Biological Sequences Alignment Algorithm

Haihe Shi^{1*}, Gang Wu¹, Xuchu Zhang¹, Jun Wang¹, Haipeng Shi^{2,3} and Shenghua Xu²

¹ School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China, ² School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China, ³ School of Software, Jiangxi Normal University, Nanchang, China

OPEN ACCESS

Edited by:

Fa Zhang,
Chinese Academy of Sciences
(CAS), China

Reviewed by:

Yushan Qiu,
Shenzhen University, China
Taolue Chen,
Birkbeck, University of London,
United Kingdom
Yanjie Wei,
Chinese Academy of Sciences
(CAS), China

*Correspondence:

Haihe Shi
haiheshi@jxnu.edu.cn

Specialty section:

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

Received: 18 November 2020

Accepted: 21 December 2020

Published: 21 January 2021

Citation:

Shi H, Wu G, Zhang X, Wang J, Shi H
and Xu S (2021) Research on
Components Assembly Platform of
Biological Sequences Alignment
Algorithm. *Front. Genet.* 11:630923.
doi: 10.3389/fgene.2020.630923

After years of development, the complexity of the biological sequence alignment algorithm is gradually increasing, and the lack of high abstract level domain research leads to the complexity of its algorithm development and improvement. By applying the idea of software components to the design and development of algorithms, the development efficiency and reliability of biological sequence alignment algorithms can be effectively improved. The component assembly platform applies related assembly technology, which simplifies the operation difficulty of component assembly and facilitates the maintenance and optimization of the algorithm. At the same time, a friendly visual interface is used to intuitively complete the assembly of algorithm components, and an executable sequence alignment algorithm program is obtained, which can directly carry out alignment computing.

Keywords: biological sequence alignment algorithm, component, component model, component assembly platform, B/S architecture

INTRODUCTION

Bioinformatics is an interdisciplinary subject involving life sciences, mathematics, and computer science. Its main research work lies in the acquisition, processing, and storage of biological information, and further includes distribution, analysis, and interpretation. Its research methods are to use various technologies and tools of computer, biology and mathematics to mine and understand the biological significance contained in the massive data (Wang et al., 2015; Liu, 2018). After years of development, bioinformatics has shaped big data of biological information. As a basic method of mining biological sequence information, sequence alignment algorithms have received extensive attention from researchers in recent years.

Sequence alignment algorithms can be divided into pairwise sequence alignment algorithms and multiple-sequence alignment algorithms (Zhan et al., 2019, 2020). The most classic solution of the pairwise sequence alignment algorithm is the dynamic programming algorithm, and the multiple-sequence alignment algorithm is due to its NP completeness (Wang and Jiang, 1994), the current research is dedicated to finding the best approximate solution, but there is a lack of research on the level of algorithm domain. In recent years, the complexity and development difficulty of the newly proposed sequence alignment algorithm program have been increasing, and the efficiency of algorithm development and maintenance cannot be guaranteed. The idea of Component-Based Software Development (CBSD) (Yin, 2017) is viewed as an effective means to solve the “software crisis.” It is also one of the current development trends of software development. Its greatest

advantage is that it can reuse the existing development results and improve software development efficiency. Algorithm is the core of software, which embodies the wisdom of software developers. The development efficiency and running efficiency of the algorithm have a crucial impact on the final quality of software. Therefore, the development idea of CBSD can be applied to algorithm development to further improve the development activities of algorithm programs.

Don Batory proposed an algorithmic component development method, connecting the feature model, grammar, and proposition formula to achieve the purpose of defining arbitrary constraints and using satisfiability solvers to debug feature models. In addition, a logical truth maintenance system is introduced to propagate the constraint characteristics of feature selection. Finally, based on these theoretical foundations, a product line development tool set that supports feature modularization and its combination is developed, and the combination development of graph algorithm is described (Batory, 2005).

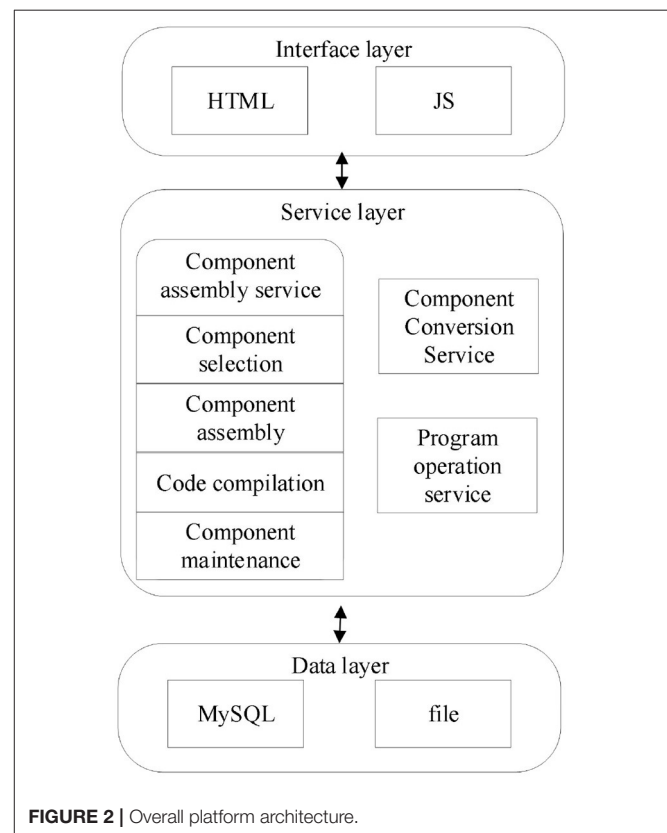
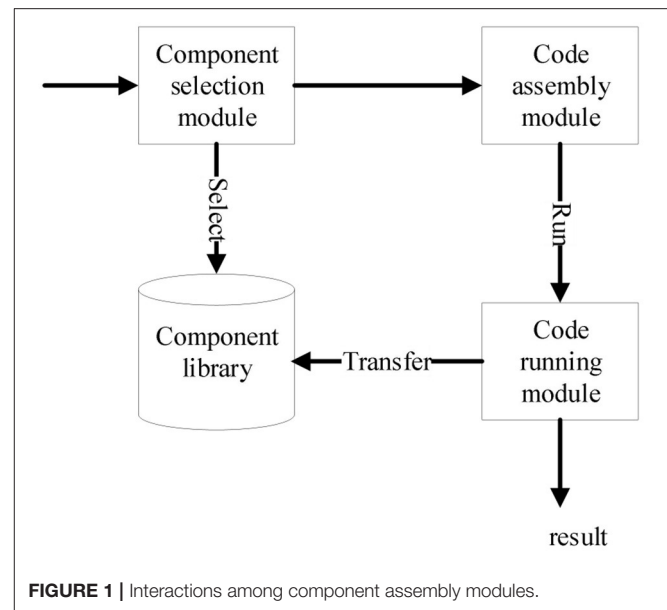
Through in-depth study, we found that the first step of component-based algorithm development is to complete the domain analysis of algorithm family in a certain domain, and obtain a domain feature model that can guide component design and implementation. Next step is the structural design and interaction design of the components according to the requirements shown in the feature model. Finally it is to implement models using a suitable development language and provide corresponding component assembly services. Under the guidance of generative programming (Czarnecki and Eisenecker, 2000), FODM (Zhang and Mei, 2003) domain modeling method and PAR (Xue, 1993, 1997, 1998, 2016; Wang and Xue, 2009; Xue et al., 2018), domain modeling activities, component design activities and component implementation activities for common sequence alignment algorithms are almost done by our research team. Based on the existing results, the paper presents the assembly platform of sequence alignment algorithm components. The platform mainly provides the assembly services for the developed algorithm components, which greatly improves the automation of the algorithm component assembly, and further reduces the complexity of the algorithm development.

PLATFORM CONSTRUCTION

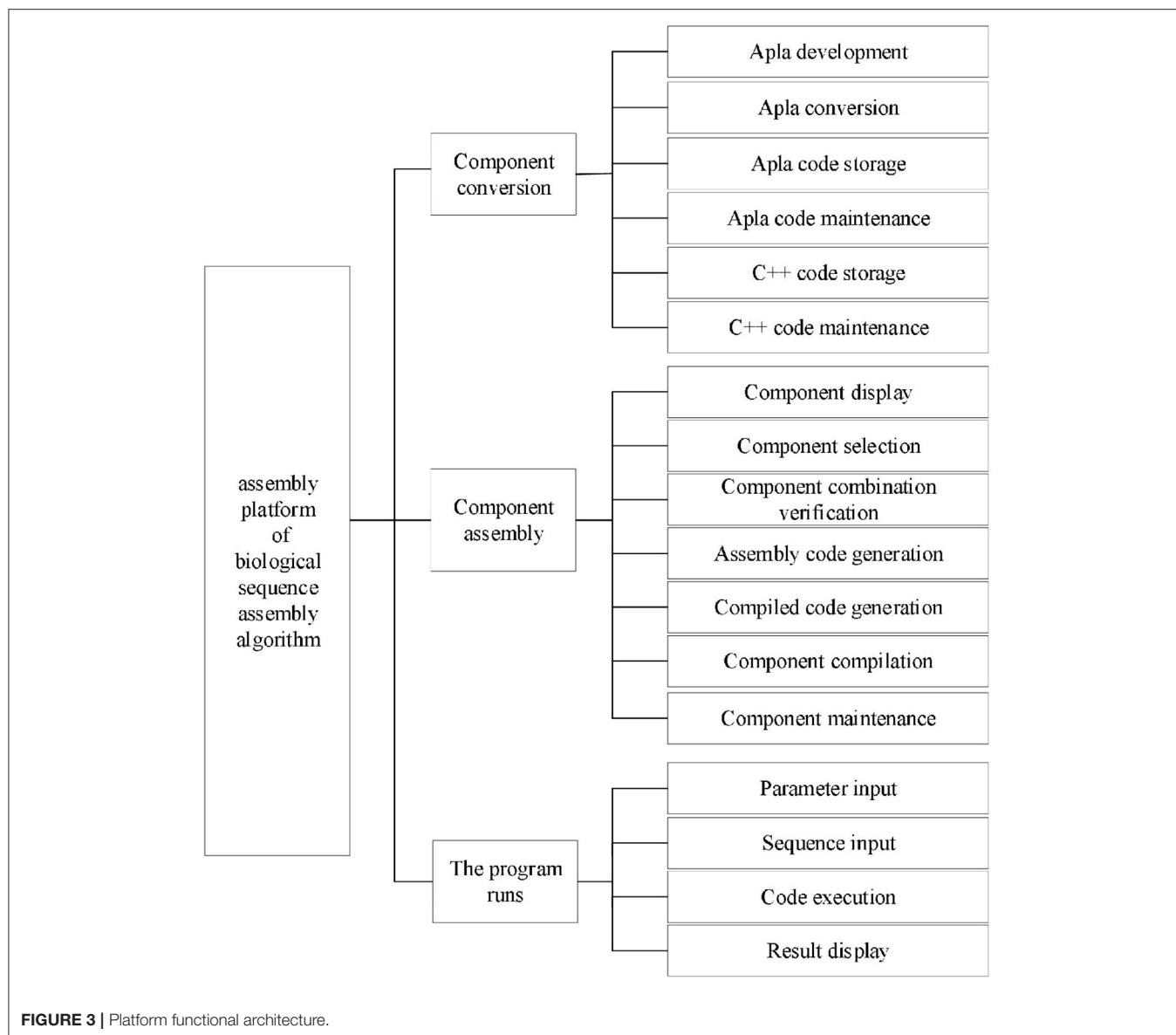
Preliminaries

Software Reuse

With the development of computer technology, its influence in human society is gradually improved. While the complexity and security of software are becoming increasingly prominent. Researchers are difficult to grasp the efficiency, cost, quality and future maintenance of software development. As early as 1968, the North Atlantic Treaty Organization (NATO) has put forward the definition of software crisis. And then the research of software engineering (Wang et al., 2018) also develops rapidly. Software reuse (Zhang and Mei, 2003, 2014; Zhang et al., 2005; Barros-Justo et al., 2019; Feng et al., 2019) is considered to be a feasible technology to improve the level of software industrial production and effectively solve the software crisis.



The idea of software reuse is to reuse the existing software in accordance with the specifications in the development process. When developing other systems in the same field, it is not necessary to develop from scratch, but on the basis of reusable resources to carry out efficient reuse development. In this process,



abstraction is the basic element (Zhu, 2017), and efficient reuse cannot lack high-abstraction modeling of related reuse fields. The scope of reusable resources covers various forms of products, including software design documents, domain models, software patterns, code components, software architecture, software implementation documents, application generators and so on.

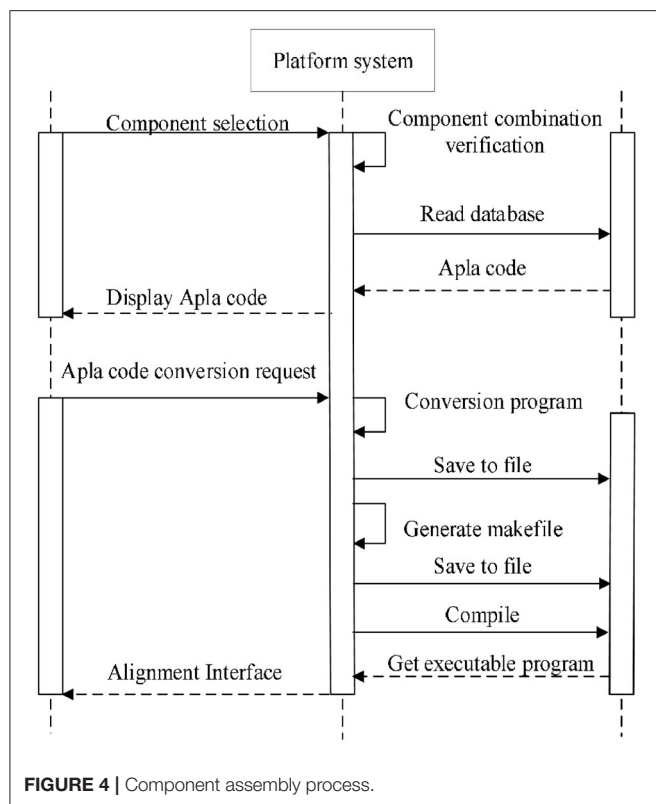
Component Technology

Component assembly technology (Zhang, 2018; Wu, 2019) is the core part of realizing CBSD. After completing a series of component design and development work, the final goal of CBSD is to assemble the components. From the current research (Xu et al., 2006; Chen et al., 2012; Zhen et al., 2014), the technology has achieved some research results.

The component assembly forms mainly include black box assembly, white box assembly and gray box assembly. The main

difference is whether the components need to be modified before assembly. Black box assembly is the most suitable assembly method for component encapsulation, but it also reduces the adaptability of components. White box assembly emphasizes the adaptability of components. The assembly is flexible and can achieve greater composability. However, due to too many implementation details exposed, the ease of use of components will be reduced, and improper modifications will occur, so that the final assembly cannot achieve the expected. Gray box assembly is the most widely used assembly method currently. It combines black box assembly and white box assembly and can be adapted to a variety of application scenarios.

The difference between sequence alignment algorithm component and software component is that the former often has higher coupling degree, and the algorithm component often needs to be modified to adapt to the relevant application

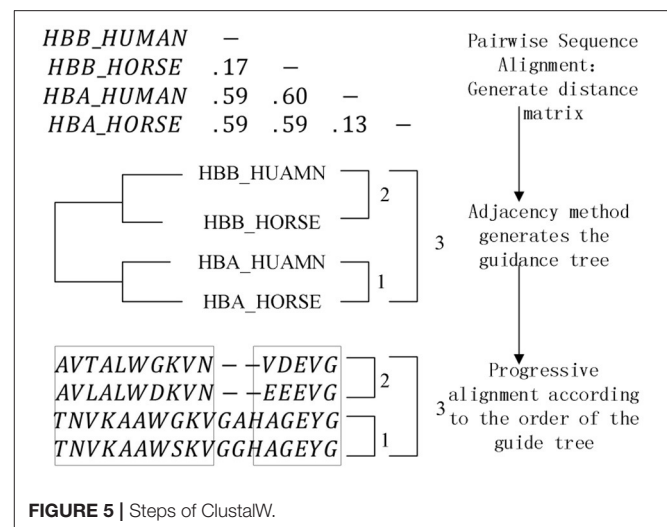


scenarios. However, the idea of sequence alignment algorithm is complex. If the assembler does not have a good understanding of the algorithm, new errors will be made when modifying it. Therefore, before assembling algorithm components, it is necessary to conduct a detailed domain analysis, formally describe the algorithm component, and form a structure framework to guide the algorithm component assembly. Finally, the gray box assembly of sequence alignment algorithm components is completed under the guidance of domain model, formal specification and algorithm framework.

Platform Design Requirement Analysis

The goal of CBSD is that the software system can be automatically generated from a series of software components according to the system requirements supported by the generator. The purpose of developing the component assembly platform for sequence alignment algorithm is also to reduce the manual assembly workload as much as possible and improve the automation level of the whole component system.

The platform mainly includes component transformation, component assembly and code running. By means of C++ program generation system of PAR, the component transformation module can transform Apla components into C++ components, see details in Xue et al. (2018). The component assembly function and code running function are composed of four modules, i.e., component library, component



selection, code assembly, and code running. The interactive relationship among the modules is shown in Figure 1.

Component library module includes two parts, one is the source code of algorithm components that have been transformed and stored in the files, and the other is the Apla component assembly code that needs to be manually developed or modified in the database. In addition, the component library module also plays a management role, such as adding, deleting, modifying, and checking components, supporting further component expansion and modification in the future.

Component selection module reads the components in the component library and displays them on the platform interface according to their required and optional features and the type characteristics of their affiliation. After selecting the components, the validity of component composition is checked. According to the multi-choice one or multi-choice relationship of feature dimension and the dependency relationship between components, the component composition is constrained accordingly to prevent illegal combination from the subsequent process.

After completing the component selection, code assembly module obtains the required Apla assembly code from the database, and the user can make appropriate modifications to correctly call the selected component in the component library. After the assembly code is developed, the Apla code is converted into the corresponding C++ code through the transformation system of PAR platform. Finally, the assembly and compilation of executable codes is completed in the sever of platform to generate executable algorithmic programs.

After the user inputs the alignment data, the code running module executes the corresponding executable alignment algorithm program. It enables the user to directly perform sequence alignment operations on the platform, and displays the final algorithm running results on the page.

System Design

The platform is developed using B/S architecture following J2EE specifications, the Java language is used, and the Spring

MULTIPLE SEQUENCE ALIGNMENT

MANDATORY

☒ seq_check

PAIRWISE MODE

☒ DP
☐ fast

ALIGN MODE

☒ progressive
☐ iterative

PROGRESSIVE_MODE

☒ phylogenetic tree
☐ extend_lib
☐ center_star

TREE_ALGORITHM

☒ NJ
☐ UPGMA

MULTI_RESULT_OP

☒ tree_op
☒ align_op

generate

FIGURE 6 | Component selection interface.

Boot (Wang et al., 2016) and MyBatis (Rong, 2015) as well as Thymeleaf framework, which are currently popular in web development, are adopted, and the stability of their architecture has been tested by practical applications.

The overall architecture of the assembly platform is shown in **Figure 2**, including the data layer, service layer, and interface layer. The data layer mainly uses MySQL and files to store the Apla program and component source code required by the platform. The service layer mainly includes component assembly service, component transformation service, and program run service. The application layer mainly uses HTML to display the platform, uses JS to implement the relevant interaction logic.

According to the requirement analysis, the functional architecture of the platform is shown in **Figure 3**. Component transformation module includes the functions of Apla development, Apla transformation, Apla code storage and maintenance, and C++ code storage and maintenance. As the core of the platform, component assembly module consists of the functions of component display, component selection, component combination verification, assembly code generation, compiling code generation, component compilation and component maintenance. Program running module is composed of the functions of sequence input, parameter input, code execution and result display.

Detailed Platform Design

Through system function requirement analysis, overall architecture design, and functional architecture design, the sequence alignment algorithm component assembly platform is outlined. Next is to give a detailed design of the platform system based on the operating sequence of each module. The platform mainly includes component conversion process, component selection process, component assembly process and algorithm execution process. This section mainly describes the process of component assembly, as shown in **Figure 4**.

Analysis of Key Platform Algorithms

The pairwise sequence alignment algorithm and heuristic multiple sequence alignment algorithm based on dynamic programming have been implemented in the platform, and the most critical one is the progressive multiple sequence alignment algorithm based on phylogenetic tree. The most classic ClustalW (Thompson et al., 1994) algorithm in the algorithm thought is implemented in 1994 by Thompson and Higgins. Its operation steps are described as follows, and the algorithm diagram is shown in **Figure 5**.

- (1) Pairwise sequence alignment. The sequence group is aligned between two pairs, and the distance matrix is established by the pairwise sequence alignment score to indicate the distance between the sequences.
- (2) Generate a phylogenetic tree. Using the information in the distance matrix, a phylogenetic tree is established through the corresponding clustering algorithm to guide the subsequent multiple sequence alignment operations.
- (3) Progressive alignment. The previous operation has generated a guide tree, and the last step is to gradually complete the alignment of all sequences in the form of keeping gaps, starting from the close evolutionary relationship according to the alignment sequence reflected by the guide tree.

The components involved in the algorithm are sequence validity check component, pairwise sequence alignment component, distance matrix component, phylogenetic tree component, progressive alignment component, and alignment result output component. Since the distance matrix component and the pairwise sequence alignment component are highly coupled, the pairwise sequence alignment component is designed as a generic operation parameter of the distance matrix, and the corresponding distance matrix can be generated by instantiating different pairwise sequence alignment algorithms. The phylogenetic tree component also includes a clustering algorithm selection sub-component, which is also designed as a

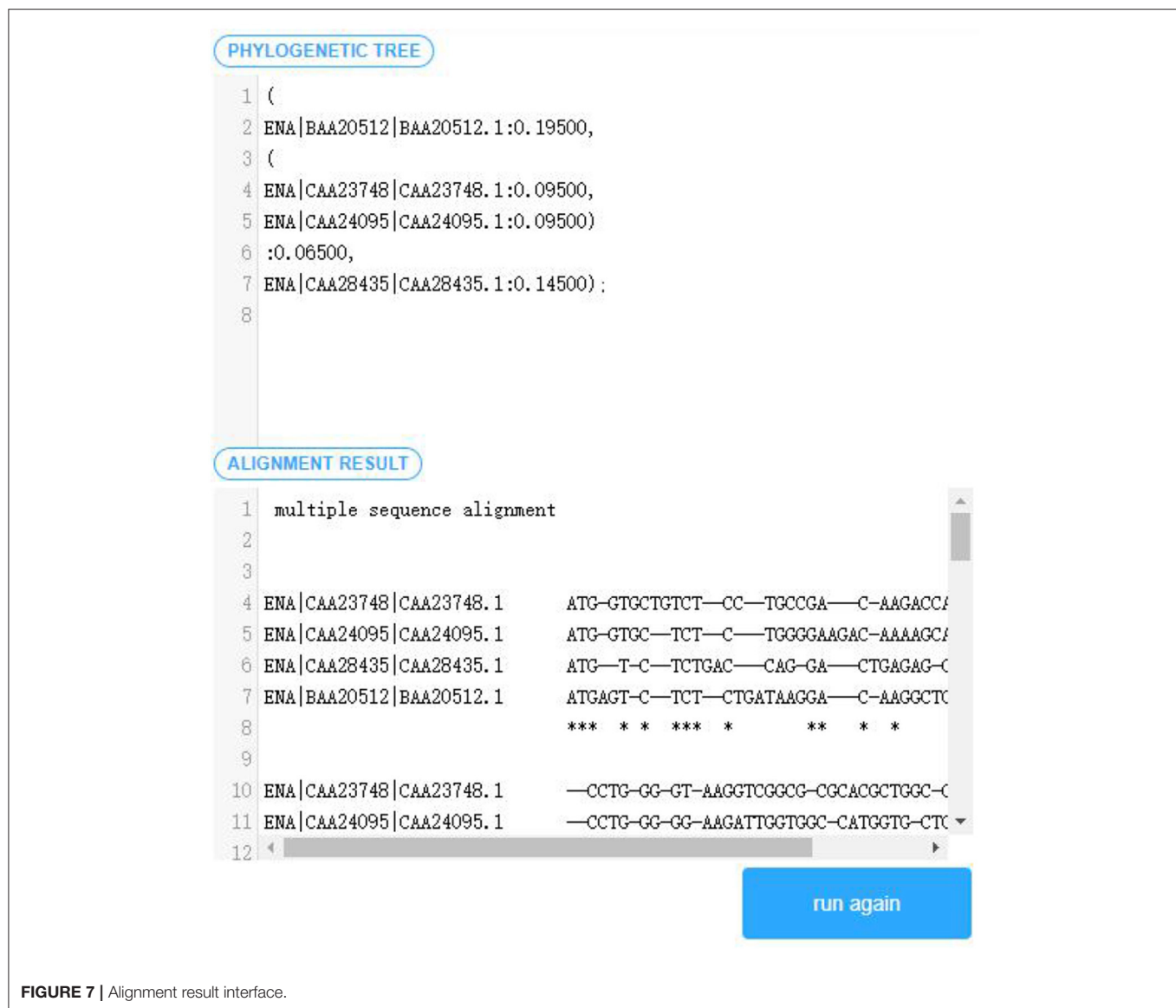


FIGURE 7 | Alignment result interface.

generic operation parameter. The commonly used instantiation algorithms are the NJ algorithm (Saitou and Nei, 1987) and the UPGMA algorithm (Zhang et al., 2018). The objective function (Carrillo and Lipman, 1988; Notredame et al., 1998) is also designed as a generic operation parameter while performing a progressive alignment, here we aims to expand the scope of algorithm components assembly.

ASSEMBLY EXAMPLE

We will carry out an example of the assembly for the progressive alignment algorithm based on phylogenetic tree to demonstrate how the modules of the platform work together and how they interact with each other.

- (1) Transform Apla components except those for assembling. The transformation system of PAR platform is used to convert the developed Apla components into C++ components and store them to the platform's local files.
- (2) Visually select several existing components satisfying the composable constraints according to the established domain feature model and component interaction model. The components are grouped by the required or optional attribute. In order to prevent the selection of illegal combinations from the subsequent assembly, the distinction between multi-choice-one or multi-choice-multi is carried out in the optional components group. The component selection interface is shown in Figure 6.
- (3) Based on the interaction relationship between the components, read the corresponding Apla assembly code in the database and display it on the page following the selection of component combination. The user can check and modify the component assembly code, and then submit an Apla conversion request and store the converted C++ assembly code as the local file.
- (4) After all the Apla component conversion and assembly code conversion, the *makefile* script file for compilation is

generated automatically, and is executed to compile and link the C++ components. The parameter input interface of sequence the **Presentation 1** for details.

- (5) After the user inputs the sequence data, and the replacement matrix as well as the penalty model required by the multiple sequence alignment, the algorithm program generated will be executed, and the alignment output displayed in the user interface. As shown in **Figure 7**.

SUMMARY

With the development of CBSD, component-based development technology has been verified in many practical applications. It can exactly improve the development efficiency and maintenance of software. In this paper, the component development technology is applied to the development of biological sequence alignment algorithms. Under the guidance of domain modeling, generative programming and PAR, the formal transformation of sequence alignment algorithm components is carried out, and a B/S-based visual assembly platform for the gray box assembly of algorithm components is constructed. On top of our previous study results, the components required by the sequence alignment algorithm are classified and displayed, and the corresponding combination constraints are designed and implemented. After the legal component combination is selected, the assembly code can be modified and compiled to form an executable algorithm program. In addition, the algorithm can run directly on the platform to facilitate users to conveniently conduct sequence alignment studies.

Next, we will release out codes in GitHub. Future work also includes the improvement of the biological sequence alignment algorithm component assembly platform from the following aspects.

- (1) The algorithm components of this platform will be further expanded to enlarge the scope of algorithms generated from component assembly.
- (2) The combination constraints in the platform have not been explicitly implemented. We will restrict the combination constraints of algorithm components to XML files, and shape the corresponding combination constraint documents to make it easier for users to assemble.

REFERENCES

- Barros-Justo, J. L., Benitti, F. B. V., and Matalonga, S. (2019). Trends in software reuse research: a tertiary study. *Comput. Standards Interf.* 66:103352. doi: 10.1016/j.csi.2019.04.011
- Batory, D. (2005). "Feature models, grammars, and propositional formulas," in *Proceedings of the Software Product Lines: 9th International Conference*, eds H. Obbink, and K. Pohl (Berlin, Heidelberg: Springer), 7–20. doi: 10.1007/11554844_3
- Carrillo, H., and Lipman, D. (1988). The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.* 48, 1073–1082. doi: 10.1137/0148063
- Chen, X., Wu, Y. J., Peng, X., and Zhao, W. Y. (2012). A collaborative approach for web application development by using component composition. *J. Front. Comput. Sci. Technol.* 7, 114–125. doi: 10.3778/j.issn.1673-9418.1209002

- (3) With the richer component library, the algorithm component library needs to have an efficient component search function. Recent years, the recommendation algorithm based on artificial intelligence has developed rapidly. The feasibility of introducing this technology into the platform to improve the ease and automation level of algorithm assembly platform will be carefully studied.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

HS instructed the whole research work and revised the paper. GW, XZ, and JW did the codes work and the experiments. HS did the experiments. SX proofread the full text. All authors read and approved the final manuscript and were agreed to be accountable for all aspects of the work.

FUNDING

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61662035, 62062039, and 71561010, the Natural Science Foundation of Jiangxi Province under Grand No. 20202BAB202024, and the Post-graduate Innovation Project of Jiangxi Province under Grand No. YC2016-B061.

ACKNOWLEDGMENTS

We thank the reviewers of CBC2020 for their helpful comments and recommendation for publication in the Journal of Frontiers in Genetics.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.630923/full#supplementary-material>

- Czarnecki, K., and Eisenecker, U. (2000). *Generative Programming: Methods, Tools, and Applications*. New Jersey: Addison-Wesley.
- Feng, H. W., Du, P. Y., and Liu, Y. (2019). Software reuse technology and its application in software development. *Electronic Technol. Softw. Eng.* 6:51.
- Liu, Z. (2018). Application of computer technology in bioinformatics. *Sci. Technol. Innov.* 10, 67–68.
- Notredame, C., Holm, L., and Higgins, D. G. (1998). COFFEE: an objective function for multiple sequence alignments. *Bioinformatics* 14, 407–422. doi: 10.1093/bioinformatics/14.5.407
- Rong, Y. D. (2015). Application research of mybatis persistence layer framework. *Inf. Security Technol.* 6, 86–88. doi: 10.3969/j.issn.1674-9456.2015.12.031
- Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence

- weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680. doi: 10.1093/nar/22.22.4673
- Wang, C. J., and Xue, J. Y. (2009). “Formal derivation of a generic algorithmic program for solving a class of extremum problems,” in *2009 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking and Parallel/Distributed Computing* (Daegu), 100–105. doi: 10.1109/SNPD.2009.46
- Wang, G., Liu, Y., Zhu, D., Klau, G. W., and Feng, W. (2015). Bioinformatics methods and biological interpretation for next-generation sequencing data. *Biomed Res. Int.* 2015:690873. doi: 10.1155/2015/690873
- Wang, L., and Jiang, T. (1994). On the complexity of multiple sequence alignment. *J. Comput. Biol.* 1, 337–348. doi: 10.1089/cmb.1994.1.337
- Wang, Q., Ping, J., and Ban, Y. (2018). Application of software engineering technology in system software development process. *China N. Telecommun.* 21, 96–97. doi: 10.3969/j.issn.1673-4866.2019.09.080
- Wang, Y. H., Zhang, J. S., and Deng, A. M. (2016). Spring boot research and application. *Inf. Commun.* 10, 91–94. doi: 10.3969/j.issn.1673-1131.2016.10.045
- Wu, M. (2019). *Instantiated Component Assembly Platform*. Jilin university.
- Xu, F., Liu, Y., and Huang, H. (2006). Research of component composition technology based on software architecture connector. *Comput. Appl.* 26, 836–839.
- Xue, J. (1993). Two new strategies for developing loop in variants and their applications. *J. Comput. Sci. Technol.* 8, 147–154. doi: 10.1007/BF02939477
- Xue, J. (1997). A unified approach for developing efficient algorithmic programs. *J. Comput. Sci. Technol.* 12, 314–329. doi: 10.1007/BF02943151
- Xue, J. (1998). Formal derivation of graph algorithmic programs using partition-and-recur. *J. Comput. Sci. Technol.* 13, 553–561. doi: 10.1007/BF02946498
- Xue, J. (2016). “Genericity in PAR Platform,” in *International Workshop on Structured Object-Oriented Formal Language and Method* (Cham: Springer), 3–14. doi: 10.1007/978-3-319-31220-0_1
- Xue, J., Zheng, Y., Hu, Q., You, Z., Xie, W., Cheng, Z. (2018). “PAR: a practicable formal method and its supporting platform,” in *Formal Methods and Software Engineering*, eds J. Sun, and M. Sun (Gold Coast, QLD: ICFEM) 70–86. doi: 10.1007/978-3-030-02450-5_5
- Yin, H. G. (2017). *Component-Based Software Development*. Jiangsu University of Science and Technology.
- Zhan, Q., Fu, Y., Jiang, Q. H., Liu, B., Peng, J., and Wang, Y. D. (2020). SpliVert: a protein multiple sequence alignment refinement method based on splitting-splicing vertically. *Protein Pept. Lett.* 27, 295–302. doi: 10.2174/0929866526666190806143959
- Zhan, Q., Wang, N., Jin, S., Tan, R., Jiang, Q. H., and Wang, Y. D. (2019). ProbPFP: a multiple sequence alignment algorithm combining hidden Markov model optimized by particle swarm optimization with partition function. *BMC Bioinformatics*. 20(Suppl. 18):573. doi: 10.1186/s12859-019-3132-7
- Zhang, F. W. (2018). *Research and Application of Component-Based Software Reuse Technology*. University of north.
- Zhang, R., Wang, Y., Zhu, X., Yin, J., Han, C., and Yang, Y. (2018). Study on optimization initial center K-Means Algorithm based on UPGMA. *Comput. Technol. Dev.* 28, 50–58. doi: 10.19720/j.cnki.issn.1005-9369.2005.05.027
- Zhang, S., Li, H., and Gu, Z. G. (2005). A powerful tool for functional genomics research—Comparative genomics. *J. Northeast Agric. Univ.* 36, 664–668.
- Zhang, W., and Mei, H. (2003). A feature-oriented domain model and its modeling process. *Softw. J.* 14, 1345–1356. doi: 10.13328/j.cnki.jos.2003.08.001
- Zhang, W., and Mei, H. (2014). Feature-oriented software reuse technology-state of the art (in Chinese). *Chin. Sci. Bull.* 59, 21–42. doi: 10.1360/972013-341
- Zhen, D. W., Shen, L. W., Peng, X., and Zhao, W. Y. (2014). Component composition technology and tool based on AJAX for web application. *Comput. Sci.* 41, 152–156. doi: 10.11896/j.issn.1002-137X.2014.11.030
- Zhu, H. Q. (2017). Discussion on computer software reuse technology. *The Digital World* 12:425.

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.

Copyright © 2021 Shi, Wu, Zhang, Wang, Shi and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Efficient Multiple Sequences Alignment Algorithm Generation via Components Assembly Under PAR Framework

Haipeng Shi^{1,2}, Haihe Shi^{3*} and Shenghua Xu¹

¹School of Information Management, Jiangxi University of Finance and Economics, Nanchang, China, ²School of Software, Jiangxi Normal University, Nanchang, China, ³School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China

OPEN ACCESS

Edited by:

Wang Guohua,
Harbin Institute of Technology, China

Reviewed by:

Pu-Feng Du,
Tianjin University, China
Yushan Qiu,
Shenzhen University, China

*Correspondence:

Haihe Shi
haiheshi@jxnu.edu.cn

Specialty section:

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

Received: 11 November 2020

Accepted: 29 December 2020

Published: 04 February 2021

Citation:

Shi H, Shi H and Xu S (2021) Efficient Multiple Sequences Alignment Algorithm Generation via Components Assembly Under PAR Framework. *Front. Genet.* 11:628175. doi: 10.3389/fgene.2020.628175

As a key algorithm in bioinformatics, sequence alignment algorithm is widely used in sequence similarity analysis and genome sequence database search. Existing research focuses mainly on the specific steps of the algorithm or is for specific problems, lack of high-level abstract domain algorithm framework. Multiple sequence alignment algorithms are more complex, redundant, and difficult to understand, and it is not easy for users to select the appropriate algorithm; some computing errors may occur. Based on our constructed pairwise sequence alignment algorithm component library and the convenient software platform PAR, a few expansion domain components are developed for multiple sequence alignment application domain, and specific multiple sequence alignment algorithm can be designed, and its corresponding program, i.e., C++/Java/Python program, can be generated efficiently and thus enables the improvement of the development efficiency of complex algorithms, as well as accuracy of sequence alignment calculation. A star alignment algorithm is designed and generated to demonstrate the development process.

Keywords: multiple sequence alignment algorithm, domain component, algorithm generation, convenient software development platform, bioinformatics

INTRODUCTION

Alignment is a common and important approach in biology study. In the research of bioinformatics (Wang et al., 2015), biological sequence alignment is one of the important processes of similarity analysis between unknown and known molecular sequences, the basis of biological sequence analysis and database search, and used in the sequence assembly. It is the key link to apply high-performance computing to biology.

Sequence alignment is a technique for identifying regions of sequence similarity by arranging genome sequences to obtain the function, structure, or evolutionary relationship between the sequences to be aligned. With the implementation of the Human Genome Project, the development of sequencing technology has produced a large amount of raw sequence data about biological molecules. For example, Illumina HiSeqX Ten can generate approximately 3 billion 2×150 bp paired-end sequencing data within 3 days (Illumina, 2016). Challenged with such a wealth of

genome sequence data, to efficiently process and analyze these data, to compare similar regions and conserved sites between the two sequences, to seek sequence homology structures, and to reveal biological heredity, variation, and evolution, etc., have become the main motivations for the research of sequence alignment algorithms.

At present, most of the research on alignment algorithms focus on specific problems (Isa et al., 2014; Cattaneo et al., 2015; Chattopadhyay et al., 2015; Huo et al., 2016) or specific algorithm optimization (Farrar, 2007; Houtgast et al., 2017; Junid et al., 2017) in the field of sequence similarity analysis, but less on the whole problem domain, so it is difficult to get an algorithm component library with a higher level of abstraction and suitable for the whole field of sequence similarity analysis. To some extent, this leads to the redundancy of the sequence alignment algorithm and the errors that may be caused by the artificial selection algorithm. It also makes it difficult for people to effectively understand the algorithm structure and ensure the correct use of the algorithm, which reduces the accuracy of the sequence similarity analysis. Because of the specificity and low-level abstraction of existing algorithms, researchers need to spend a lot of time to learn and use such algorithms, and it is also difficult to locate and solve the errors generated by the algorithms; thus, maintainability and reusability of the algorithms are reduced, and the burden of sequence similarity analysis is increased.

Sequence alignment algorithms can be divided into pairwise alignment algorithms and multiple sequence alignment algorithms (Zhan et al., 2019, 2020). Among them, the most classic solution to the pairwise sequence alignment algorithm is dynamic programming. We studied the field of dynamic programming-based pairwise sequence alignment algorithm (DPPSAA) in the early stage and established a domain component library (Shi and Zhou, 2019), which has been successfully applied to the problem of pairwise sequence alignment algorithm. However, the multisequence alignment algorithm is rather complex. Because of its non-deterministic polynomial (NP)-complete (Wang and Jiang, 1994), current researches are all devoted to finding the optimal approximate solution. With the increase of the complexity and difficulty of the multisequence alignment algorithm, the reliability and efficiency of the algorithm are difficult to be guaranteed.

Based on the previous work, this article adopts the formal method PAR (Xue, 1997, 2016; Shi and Xue, 2009, 2012; Xue et al., 2018) to describe, construct, transform, and refine the components, models, and frameworks related to the multisequence alignment algorithm and expand PAR platform to support the generation of effective multiple sequence alignment algorithm *via* component assembly. The multilevel different models in the algorithm development process are unified under the PAR framework to effectively ensure the reliability of the resulting algorithm and improve the efficiency of algorithm development.

Through in-depth analysis of the field of multiple sequence alignment algorithms, based on the component library of the DPPSAA domain, some algorithm components have been improved and added, and a component library of multiple

sequence alignment algorithms on top of the component library of the DPPSAA domain was established. Finally, an example, the successful assembly of the star alignment algorithm and the automatic generation of the C++ program, is shown.

ALGORITHM GENERATION UNDER THE PAR FRAMEWORK

Related Work

On the basis of the component library in the DPPSAA domain, this article has carried out the research on the algorithm design and program generation of multiple sequence alignment algorithms under the PAR framework.

PAR

The PAR framework includes two parts: software formal method and convenient software development platform. The PAR method is composed of a generic algorithm design language Radl, a generic abstract programming language Apla, systematic methodology for algorithms and programming. It combines two high-efficiency techniques, i.e., partition and recursion used in special problems, covering a variety of known algorithm design techniques such as dynamic programming, greedy, divide and conquer, and so on. It can be used as a unified method of algorithm generation to avoid the difficulty of making choices among the existing algorithm design methods. The PAR platform is composed of Apla to C++/C#/Java/Python program generation systems and realizes the automatic generation of algorithmic programs such as sequential programs, parallel/concurrent programs, and database applications.

Practice has proven that the productivity of complex algorithm program and database application software can be greatly improved by using the language, method, series algorithm, and program automatic generation tool provided by PAR. Many military departments, such as the National General Equipment Department, Beijing Military Region, and armored academy, have taken the lead in applying these achievements to the construction of China's important military projects and have achieved remarkable military and economic benefits. The PAR framework has been appraised by the expert group of the Ministry of Science and Technology of China as "having the international advanced level, among which the theoretical framework of the correctness of the complex algorithm program has the international leading level."

DPPSAA Domain Model and Component Library

In Shi and Zhou (2019), we analyzed the characteristics of DPPSAA, extracted the common and variable features and the constraints and dependencies between them, established the DPPSAA domain model and its algorithm component interaction model, and further implemented the models using the abstract programming language Apla to form a highly abstract DPPSAA component library, in order to automatically or semiautomatically assemble components to generate sequence alignment algorithms for specific fields, thereby reducing the error rate and time

cost of manual selection algorithms for sequence similarity analysis, improving the efficiency of algorithm execution, and even assembling a more efficient new sequence alignment algorithm based on dynamic programming.

The experimental results show that the DPPSAA algorithm component library has a certain degree of practicability and has good expected results. It can be seen from the domain realization process that the domain feature model is a formal description at a higher level of abstraction, which not only makes the specific composition characteristics and dependencies of the algorithm clearly displayed, but also is very helpful for understanding the overall architecture of the algorithm. Moreover, the establishment of the feature interaction model makes it easier to specify the specific configuration knowledge required by the algorithm in the domain implementation process and then automatically assemble the components in the DPPSAA algorithm component library to design the desired algorithm, without paying too much attention to the details of algorithm implementation.

Algorithm Generation Process

Based on a large amount of practical work carried out in the early stage, combined with related methodologies such as PAR and domain engineering, the development of multisequence alignment algorithms can be divided into two parts: reuse-oriented development and application reuse development.

For reuse-oriented development, it can be divided into the following steps:

1. Analyze the algorithm family in the field of multiple sequence alignment, and establish the domain model.
2. Formally describe the component function specifications.
3. Use the PAR method to design abstract Apla algorithm components, use the PAR platform to obtain highly reliable executable language-level components, and expand the PAR platform component library in a self-expanding manner.

The process of designing a specific problem-solving algorithm and generating a program is a development process of application reuse:

1. Analyze and (formally) characterize the specific problem to be solved.
2. Determine the algorithm components required for assembly.
3. The Apla abstract language is used to describe the assembly process, and the executable program corresponding to the specific algorithm is automatically generated through the PAR platform.

The introduction of the PAR framework reduces the operational difficulty of algorithm component assembly and improves the automation of algorithm component assembly.

STAR ALIGNMENT ALGORITHM

Algorithm Idea

The star alignment algorithm (Zou et al., 2009, 2015) is a heuristic fast approximation algorithm for typical multisequence alignment.

It compares all sequences in pairs and selects the sequence with the highest alignment score with other sequences as the central sequence. Then, continue to compare with other sequences to obtain the final alignment result. When adding subsequent sequences to the alignment process, follow the “leave blank once, leave blank everywhere” rule, which cannot guarantee the ultimate result of the alignment.

For example, for the sequence $s1 = \text{CGCT}$, $s2 = \text{GCGT}$, $s3 = \text{CCTG}$, the pairwise alignment results of the sequences $s1$, $s2$, and $s3$ are shown in **Table 1**.

The star alignment algorithm adds the alignment scores of each sequence to other sequences and selects the sequence with the largest score as the central sequence. Therefore, in this case, $s1$ is selected as the center sequence, and the best alignment result and the final merge result with $S2$ and $S3$ are shown in **Figure 1**.

Algorithm Component and Apla Implementation

Using feature modeling knowledge and performing process analysis on star alignment algorithms, we will know that multisequence alignment is mainly used as the core service of star alignment algorithms in the star alignment process. The multiple sequence alignment service is mainly based on the pairwise sequence alignment, by selecting the optimal pairwise sequence alignment result as the central sequence, and then continuously adding the suboptimal sequence to the alignment until the final multisequence alignment result is obtained. After analyzing the execution process of the star alignment algorithm, the multisequence alignment operation service mainly consists of the following features (the component name of the corresponding feature in parentheses): sequence legality check (*msa_check*), distance matrix (*dist_Matrix*), pairwise alignment manipulation (*align_manipulation*), center sequence selection (*msa_center*), remember alignment space

TABLE 1 | Distance matrix of $s1$, $s2$, and $s3$.

	s1	s2	s3	Score
s1		−1	−1	−2
s2	−1		−2	−3
s3	−1	−2		−3

The best alignment result:

s1	[-]	C	G	C	T		s1	C	G	C	T	[-]
s2	G	C	G	[-]	T		s3	C	[-]	C	G	T

The final alignment result:

s1	[-]	C	G	C	T	[-]
s2	G	C	G	[-]	T	[-]
s3	[-]	C	[-]	C	G	T

FIGURE 1 | Result of star sequence alignment.

(rmb_space), multisequence alignment result output (msa_op_result), and so on. Among them, sequence legality check, pairwise alignment manipulation, distance matrix, and center sequence selection are mandatory features in the star alignment algorithm, and the multisequence alignment result output feature mainly depends on the remember alignment space feature; that is, when the assembly algorithm contains a multisequence alignment result output component, it will include and implement the remember alignment space component by default.

Taking DPPSAA as the basis of sequence alignment, generic programming language Apla is used to abstractly represent the star alignment algorithm, which can realize star alignment algorithm by standardized assembly. Here, we expand on the basis of the component library in the DPPSAA domain, so that the component library in this domain can be used to assemble and implement the star alignment algorithm. We perform Apla representation of the extended component as follows:

1. Sequence legality check

msa_check is an extension based on the check component in the DPPSAA field that can be used to detect multiple sequences. The Apla process statement is:

```
procedure msa_check(String str[]);
```

where str[] represents the base string array for multiple sequence alignment.

2. Distance matrix

dist_Matrix means that all pairwise alignment scores participating in multisequence alignments are returned as distance matrix elements, and the component uses pairwise sequence alignment operations as its generic parameters. The prototype of the Apla function is as follows:

```
function dist_Matrix (procalign_manipulation(sometype
elemMatrix; ADT dp_mode(eM:elemMatrix); op_mode (func
score_op():integer; proc traceback (proc print_align(); proc
print_extrude() =NULL)); result:boolean; eM: elemMatrix;
s:String; t:String)):integer[ ][ ].
```

3. Center sequence selection

The msa_center component is an important part of the components library of multiple sequence alignment algorithm. This component can be used to select the best alignment in all pairwise alignments; take the best alignment sequence in the alignment as the center sequence, and then iteratively add the remaining sequences to obtain the best multiple sequence alignment results. The function prototype is as follows:

```
function msa_center(dist[][]: integer):integer;
```

The array dist represents the array returned by the distance matrix, and the component returns the index value of the center sequence.

4. Remember alignment space

In the star alignment algorithm, the algorithm follows the rule of “leave blank once, leave blank everywhere” when adding subsequent sequences to the alignment process. Therefore, the role of the rmb_space component is to remember the space inserted during each sequence alignment. The function prototype is as follows:

```
function rmb_space(): integer[][];
```

5. Multisequence alignment result output

This component inserts the space index value obtained in (4) into the sequence to output the final multisequence alignment result. This component can be implemented with the following Apla process:

```
procedure msa_op_result(space[][]:integer);
```

Star Alignment Algorithm Generation

Using the Apla-C++ conversion system, the aforementioned component library is converted into the corresponding C++ component through the combination of automatic conversion and manual conversion, which can be used to generate the star alignment algorithm program and conduct test analysis to obtain experimental results. This section shows only the three main components: dist_Matrix component, msa_center component, and rmb_space component.

As the star alignment algorithm requires the pairwise sequence alignment manipulation, and the alignment score result value is used as the element of the distance matrix, the dist_matrix component needs to use the sequence alignment manipulation in DPPSAA as its generic parameter to obtain the score value of the pairwise alignment of all sequences. In the process of converting the Apla program to the C++ program, it is first necessary to assemble the components in DPPSAA to form a pairwise sequence alignment algorithm and design the pairwise sequence alignment algorithm as an independent function as the function pointer parameter of the distance matrix component, which reduces the dependency between the pairwise sequence alignment algorithm and the distance matrix. Here, we set the pairwise sequence alignment algorithm to NW algorithm and return the pairwise sequence alignment scores. The C++ code is as follows:

```
class MsaNW{//NW algorithm assembly
public:
    int Msa_NW(Score_matrix_manip& matrix,const std::string&
s,const std::string& t){
        matrix.apply_memory();
        matrix.Memory_Score_of_Matrix(&Init_Score_matrix::Init_
Score_matrix1, matrix.get_Matrix(), matrix.getPenaltyMatrix(),
matrix.get_length_s(), matrix.get_length_t());
        dp_mode dp_NW;
        dp_NW.align_and_score(matrix,&set_and_remember::set_
and_remember1);
        return matrix.the_Last_element_score();
    }
}
```

The C++ program obtained by transforming the dist_matrix component is as follows:

```
class dist_Matrix{
    int** dist; //distance matrix
    int* row_sum;//sum of row
    int seqs_num;//number of sequences
public:
    void Dist_Matrix(int(MsaNW::*Msa_NW)(Score_matrix_
mani&, const std::string&, const std::string&),std::string* seqs,
Score_matrix_manip** matrix)//final score {...}
    void sum_row(){...}
}
```


Among them, the class `dist_Matrix` contains three attributes; `dist` represents the distance matrix, for example, the element `dist[0][1] = 1`, which represents the pairwise sequence alignment score value of the first sequence, and the second sequence is 1; `row_sum` represents the sum of the scores of each sequence after pairwise alignment with other sequences, that is, the row sum of `dist`; `seqs_num` represents the number of sequences participating in the alignment. In the method `Dist_Matrix`, `seqs` represents a string pointer to all sequences participating in the alignment, `matrix` represents a two-dimensional matrix composed of score matrix objects obtained after pairwise alignment of all sequences, and the method `sum_row()` is used for calculation `row_sum` value.

At the same time, `msa_center` component is transformed into a class `msa_center`. The attribute `center_index` of this class records the index of the center sequence. The method `Msa_center` is used to calculate the `center_index`, and the distance matrix object is used as its parameter. The C++ representation of this component is as follows:

```
class msa_center{
private:
int center_index; //record center sequence index
public:
int Msa_center(dist_Matrix distM){...}
}
```

`rmb_space` component is also converted to the class `rmb_space` in C++, where the attribute `Msa_space_loc` represents the gap position inserted when the center sequence is aligned with other sequences, and the attribute `msa_ret_str` means the sequence alignment result after inserting gaps in all sequences according to the “leave blank once, leave blank this time” rule. The C++ representation is as follows:

```
class rmb_space{
int** Msa_space_loc;//the position of the space when each
sequence is aligned with the center sequence
std::string* msa_ret_str;//MAS alignment result
```

public:

```
void Msa_add_space(MsaCenterSeq mcs, Dist_Matrix distM,
Msa_Sequence* seqs, Score_matrix_manip** matrix){..}
}
```

Through the above conversion, we can obtain the complete component library to assemble and generate the star alignment algorithm. The process of assembling and generating the star alignment algorithm is listed below, where `Star` represents the parameter matrix of the method `Dist_Matrix` used to construct the distance matrix in the star alignment algorithm, that is, the score matrix operation object in the NW algorithm.

```
int main{
std::string s[3]={"CGCT", "CCTG", "GCGT"};
int seq_num = sizeof(s)/sizeof(s[0]);
Msa_check().check_dna(s, seq_num);
Star star(s, seq_num);
dist_Matrix distM(seq_num);
distM.Dist_Matrix(&MsaNW:Msa_NW,s, star.get_matrix());
distM.sum_row();
msa_center mc;
mc.Msa_center_seq(distM);
RmbSpace rs(seq_num, star.get_Seqs()->max_length());
rs.Msa_add_space(mc, distM, star.get_seqs(), star.
get_matrix());
Msa_print_align().msa_print_align(rs.get_ret_str(),
seq_num);
}
```

Experiment Analysis

We downloaded four *Escherichia coli* DNA data with a length of approximately 200 characters from NCBI's Genbank gene database website for experimental testing. The basic configuration of the machine is 3.40 GHz, Intel Core i7 processor, 8 GB RAM, and Windows 7 operating system. The result of the experiment is shown in **Figure 2**. The comparison takes 11.318 s.

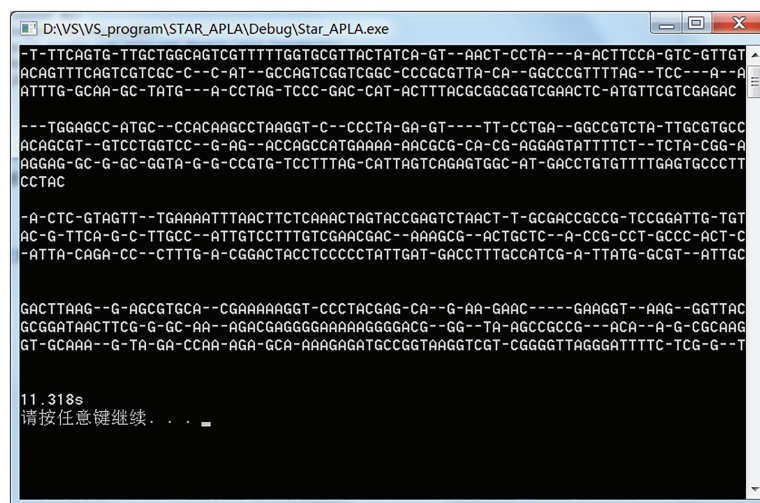


FIGURE 2 | Snapshot of the alignment result.

The running kr alignment algorithm generated by the assembly can perform multisequence alignment better and has obtained results similar to the original star alignment algorithm, which verifies the practicability of the star alignment algorithm generated by the assembly.

CONCLUSION

Sequence alignment algorithms are widely used. Because of the complexity of multiple sequence alignment problems and the diversity of algorithm design strategies, it is difficult to guarantee the development efficiency and reliability of multiple sequence alignment algorithm programs.

This article takes the problem of multiple sequence alignment as a special field and works on the algorithm development and program generation under PAR framework. Through the analysis of problem characteristics, the generality of the domain algorithm family is extracted, the features are described, and abstract algorithm components are designed. Based on the research of the pairwise sequence alignment algorithm family, the method and platform under the PAR framework are used to assemble the specific multisequence alignment algorithms and generate programs automatically. As a case study, assembly of the star alignment algorithm is given to demonstrate the generation process of the specific algorithm program, which further proves the practicability of the component library in the related field and the reliability and efficiency of the algorithm generation under the PAR framework.

REFERENCES

- Cattaneo, G., Petrillo, U. F., Giancarlo, R., and Roscigno, G. (2015). "Alignment-free sequence comparison over hadoop for computational biology" in *Proceedings of the 44th International Conference on Parallel Processing Workshops*. September 1-4, 2015; Piscataway, NJ: IEEE, 184-192.
- Chatopadhyay, A. K., Nasiev, D., and Flower, D. R. (2015). A statistical physics perspective on alignment-independent protein sequence comparison. *Bioinformatics* 31, 2469-2474. doi: 10.1093/bioinformatics/btv167
- Farrar, M. (2007). Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* 23, 156-161. doi: 10.1093/bioinformatics/btl582
- Houtgast, E., Sima, V. M., and Al-Ars, Z. (2017). "High performance streaming Smith-Waterman implementation with implicit synchronization on Intel FPGA using OpenCL" in *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Bioengineering*. October 23-25, 2017; Piscataway, NJ: IEEE, 492-496.
- Huo, H., Sun, Z., Li, S., Vitter, J. S., Wang, X., Yu, Q., et al. (2016). "CS2A: A compressed suffix array-based method for short read alignment" *Proceedings of the 2016 IEEE Data Compression Conference*. March 30-April 1, 2016; Piscataway, NJ: IEEE, 271-278.
- Illumina (2016). HiSeqX instrument performance parameters [EB/OL]. Available at: <https://www.illumina.com/systems/sequencing-platforms/hiseq-x/specifications.html>
- Isa, M. N., Murad, S. A. Z., Ismail, R. C., Ahmad, M. I., Jambek, A. B., and Kamil, M. M. (2014). "An efficient processing element architecture for pairwise sequence alignment" in *Proceedings of the 2nd International Conference on Electronic Design*. August 19-21, 2014; Piscataway, NJ: IEEE, 461-464.
- Junid, S. A. M. A., Idros, M. F. M., Razak, A. H. A., Osman, F. N., and Tahir, N. M. (2017). "Parallel processing cell score design of linear

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: <https://www.ncbi.nlm.nih.gov/genbank/>.

AUTHOR CONTRIBUTIONS

HpS did the codes work and the experiments. HhS instructed the whole research work and revised the paper. SX proofread the full text. All authors read and approved the final manuscript and are agree to be accountable for all aspects of the work.

FUNDING

This work was supported by the National Natural Science Foundation of China under grant nos. 61662035, 62062039, and 71561010, the Natural Science Foundation of Jiangxi Province under grant no. 20202BAB202024, and the Postgraduate Innovation Project of Jiangxi Province under grant no. YC2016-B061.

ACKNOWLEDGMENTS

We thank the reviewers of CBC2020 for their helpful comments and recommendation for publication in the Journal of Frontiers in Genetics.

- gap penalty smith-waterman algorithm" in *Proceedings of the 13th IEEE International Colloquium on Signal Processing & ITS Applications*. March 10-12, 2017; PBatu Ferringhi: IEEE, 299-302.
- Shi, H., and Xue, J. (2009). PAR-based formal development of algorithms. *Chin. J. Comput.* 32, 982-991. doi: 10.3724/SP.J.1016.2009.00982
- Shi, H., and Xue, J. (2012). Research on automated sorting algorithms generation based on PAR. *J. Softw.* 23, 2248-2260. doi: 10.3724/SP.J.1001.2012.04164
- Shi, H., and Zhou, W. (2019). Design and implementation of pairwise sequence alignment algorithm components based on dynamic programming. *J. Comput. Res. Dev.* 56, 1907-1917. doi: 10.7544/issn1000-1239.2019.20180835
- Wang, L., and Jiang, T. (1994). On the complexity of multiple sequence alignment. *J. Comput. Biol.* 1, 337-348. doi: 10.1089/cmb.1994.1.337
- Wang, G., Liu, Y., Zhu, D., Klau, G. W., and Feng, W. (2015). Bioinformatics methods and biological interpretation for next-generation sequencing data. *Biomed. Res. Int.* 2015, 1-2. doi: 10.1155/2015/690873
- Xue, J. (1997). A unified approach for developing efficient algorithmic programs. *J. Comput. Sci. Technol.* 12, 314-329. doi: 10.1007/BF02943151
- Xue, J. (2016). "PAR: a model driven engineering platform for generating algorithms and software" in *Symposium Programming: Logics, Models, Algorithms and Concurrency to recognize Jayadev Misra's Accomplishments*. April 29-30, 2016; University of Texas.
- Xue, J., Zheng, Y., Hu, Q., You, Z., Xie, W., and Cheng, Z. (2018). "PAR: a practicable formal method and its supporting platform" in *Proceedings of the 20th International Conference on Formal Engineering Methods (ICFEM 2018)*. eds. J. Sun and M. Sun. November 12-16, 2018; LNCS 11232, 70-86.
- Zhan, Q., Fu, Y., Jiang, Q. H., Liu, B., Peng, J., and Wang, Y. D. (2020). SpliVert: a protein multiple sequence alignment refinement method based on splitting-splicing vertically. *Protein Pept. Lett.* 27, 295-302. doi: 10.2174/0929866526666190806143959

- Zhan, Q., Wang, N., Jin, S., Tan, R., Jiang, Q. H., and Wang, Y. D. (2019). ProbPPF: a multiple sequence alignment algorithm combining hidden Markov model optimized by particle swarm optimization with partition function. *BMC Bioinformatics* 20(Suppl. 18):573. doi: 10.1186/s12859-019-3132-7
- Zou, Q., Guo, M., and Wang, X. (2009). An algorithm for DNA multiple sequence alignment based on center star method and keyword tree. *Acta Electron. Sin.* 37, 1746–1750.
- Zou, Q., Hu, Q., Guo, M., and Wang, G. (2015). HAlign: fast multiple similar DNA/RNA sequence alignment based on the centre star strategy. *Bioinformatics* 31, 2475–2481. doi: 10.1093/bioinformatics/btv177

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.

Copyright © 2021 Shi, Shi and Xu. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Joint Lp-Norm and $L_{2,1}$ -Norm Constrained Graph Laplacian PCA for Robust Tumor Sample Clustering and Gene Network Module Discovery

Xiang-Zhen Kong, Yu Song, Jin-Xing Liu*, Chun-Hou Zheng*, Sha-Sha Yuan, Juan Wang and Ling-Yun Dai

School of Computer Science, Qufu Normal University, Rizhao, China

OPEN ACCESS

Edited by:

Xianwen Ren,
Peking University, China

Reviewed by:

Xiaojuan Shao,
National Research Council Canada
(NRC-CNRC), Canada
Yongcui Wang,
Northwest Institute of Plateau Biology
(CAS), China

*Correspondence:

Jin-Xing Liu
sdcavell@126.com
Chun-Hou Zheng
zhengch99@126.com

Specialty section:

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

Received: 26 October 2020

Accepted: 29 January 2021

Published: 23 February 2021

Citation:

Kong X-Z, Song Y, Liu J-X,
Zheng C-H, Yuan S-S, Wang J and
Dai L-Y (2021) Joint Lp-Norm
and $L_{2,1}$ -Norm Constrained Graph
Laplacian PCA for Robust Tumor
Sample Clustering and Gene Network
Module Discovery.
Front. Genet. 12:621317.
doi: 10.3389/fgene.2021.621317

The dimensionality reduction method accompanied by different norm constraints plays an important role in mining useful information from large-scale gene expression data. In this article, a novel method named Lp-norm and $L_{2,1}$ -norm constrained graph Laplacian principal component analysis (PL21GPCA) based on traditional principal component analysis (PCA) is proposed for robust tumor sample clustering and gene network module discovery. Three aspects are highlighted in the PL21GPCA method. First, to degrade the high sensitivity to outliers and noise, the non-convex proximal Lp-norm ($0 < p < 1$) constraint is applied on the loss function. Second, to enhance the sparsity of gene expression in cancer samples, the $L_{2,1}$ -norm constraint is used on one of the regularization terms. Third, to retain the geometric structure of the data, we introduce the graph Laplacian regularization item to the PL21GPCA optimization model. Extensive experiments on five gene expression datasets, including one benchmark dataset, two single-cancer datasets from The Cancer Genome Atlas (TCGA), and two integrated datasets of multiple cancers from TCGA, are performed to validate the effectiveness of our method. The experimental results demonstrate that the PL21GPCA method performs better than many other methods in terms of tumor sample clustering. Additionally, this method is used to discover the gene network modules for the purpose of finding key genes that may be associated with some cancers.

Keywords: Lp-norm, graph regularization, sparse constraint, principal component analysis, tumor clustering, gene network modules, $L_{2,1}$ -norm

INTRODUCTION

High-throughput sequencing technologies, including genome-wide measurements, have enabled large-scale gene expression profiles to accumulate faster (Goodwin et al., 2016). It is of great significance to obtain useful information from these data. Reliable and precise identification of cancer types and obtaining key pathogenic genes are very important for cancer diagnosis and treatment (Koboldt et al., 2012). Generally, gene expression data have a typical characteristic of “high dimension, low sample” size (West, 2003), which is a challenge for most traditional statistical methods. Too many variables and some uncorrelated noise variables in the gene expression data may all have a negative effect on the tumor clustering performance regardless of whether

supervised or unsupervised clustering methods are used. Despite these problems, many researchers have demonstrated the effectiveness of tumor-type identification and feature selection by leveraging many machine learning algorithms (Hochreiter et al., 2010; Lee et al., 2010; Liu J. X. et al., 2015; Bunte et al., 2016; Kong et al., 2017; Wang et al., 2017; Chen et al., 2019). Among them, algorithms based on principal component analysis (PCA) (Collins, 2002; Jolliffe, 2002) have been widely used to process gene expression data successfully (Liu et al., 2013; Liu J. X. et al., 2015; Wang et al., 2017; Feng et al., 2019) for dimension reduction and denoising. However, PCA-based algorithms, including sparse principal component analysis (SPCA) (Zou et al., 2006; Shen and Huang, 2008; Journee et al., 2010; Liu et al., 2016; Feng et al., 2019) and robust principal component analysis (RPCA) (Candès et al., 2009; Liu et al., 2013; Liu J. X. et al., 2015; Wang et al., 2017), mainly deal with data that lie in a linear data manifold (Jiang et al., 2013). Many methods that can handle data lying in a non-linear manifold have been proposed, such as locality preserving projections (LPP) (He et al., 2005), locally linear embedding (LLE) (Roweis and Saul, 2000), local tangent space alignment (Zhang and Zha, 2002), Laplacian eigenmap (LE) (Belkin and Niyogi, 2002, 2003; Spielman, 2007) and latent variable model (LELVM) (Keyhanian and Nasersharif, 2015). The purpose of these non-linear dimensionality reduction techniques is to find a representation of points (samples) in a low-dimensional space, in which all points (samples) still maintain the similarity in the original high-dimensional space.

In recent years, optimization models that combine linear and non-linear dimensionality reduction methods, especially graph Laplacian embedding, have demonstrated their effectiveness. Liu et al. (2017) constructed a graph Laplacian matrix for semisupervised feature extraction. Cai et al. (2011) proposed a method named graph regularized non-negative matrix factorization (GNMF), which combined graph structure and non-negative matrix factorization for an improved compact representation of the original data. Jiang et al. (2013) developed graph-Laplacian PCA (gLPCA), which sought a low-dimensional representation of image data with significant improvement in clustering and image reconstruction by incorporating graph structures and PCA. Feng et al. (2017) employed pgLPCA based on graph Laplacian regularization and Lp-norm for feature selection and tumor clustering. Wang et al. (2019a) used Laplacian regularized low-rank representation (LLRR), which considers the intrinsic geometric structure of gene expression data to cluster the tumor samples. In addition, many methods benefit from norm constraints. For example, Journee et al. (2010) employed the L_0 -norm constraint based on PCA to stress the sparse expression of genes in samples. The L_1 -norm (Tibshirani, 1996) was introduced as the regularization function in sparse singular value decomposition (SSVD) (Lee et al., 2010; Kong et al., 2017) and the mix-norm optimization model proposed by Wang et al. (2019b). Feng et al. (2016) employed the $L_{1/2}$ -norm constraint in their model to select characteristic genes. However, there remain some facets to be improved: for example, the robustness of the algorithm should be enhanced further, and the sparse representation of the data should be highlighted. For these purposes, the Lp-norm

(Chartrand, 2012; Nie et al., 2013; Feng et al., 2017; Kong et al., 2017) constraint was used in the optimization model to degrade the sensitivity of outliers of the data. The $L_{2,1}$ -norm (Xiang et al., 2012; Yang et al., 2012) constraint was used by Liu et al. (2017) and Wang et al. (2019b) to generate the row sparsity.

Motivated by the literature mentioned above, especially (Tibshirani, 1996; Chartrand, 2012; Xiang et al., 2012; Nie et al., 2013; Feng et al., 2017; Kong et al., 2017), we propose a new method named PL21GPCA incorporating traditional PCA, graph Laplacian embedding and different norm constraints on the loss function and the regularization function for robust tumor sample clustering and gene network module discovery. Five gene expression datasets, including one benchmark dataset, two single-cancer datasets from The Cancer Genome Atlas (TCGA), and two integrated datasets of multiple cancers from TCGA, are used to evaluate the effectiveness of our method. The experimental results demonstrate that the PL21GPCA method outperforms many existing methods in terms of tumor sample clustering. Additionally, this method is employed to discover gene network modules to identify the key genes with close relationships to some cancers.

We organize the rest of this paper as follows. Section “Related Works” provides the related works containing the non-convex proximal Lp-norm, $L_{2,1}$ -norm and graph regularized PCA. The optimization model of PL21GPCA is presented, and the solution procedure is detailed in section “Methodology.” Section “Experiments and Discussion” presents the parameter selections, experimental results and some discussions. The tumor sample clustering and gene network analysis are also described in this section. In Section “Conclusion and Suggestions,” we present the conclusion for this article and propose some suggestions for future research.

RELATED WORKS

Definitions of the Proximal Lp-Norm and $L_{2,1}$ -Norm

Let $X \in \mathbb{R}^{p \times n}$ be a data matrix, the proximal Lp-norm of X is defined as follows:

$$\|X\|_p = \left(\sum_i^p \sum_j^n |x_{ij}|^p \right)^{\frac{1}{p}} \quad (0 < p < 1) \quad (1)$$

The Lp-norm with $0 < p < 1$ is a function with three typical characteristics: globally non-differentiable, non-convex, and non-smooth (Chartrand, 2012; Zhang et al., 2015). Many researchers have made suggestions to deal with Lp-norm ($0 < p < 1$) minimization (Chartrand, 2012; Guo et al., 2013; Qin et al., 2013). Since Lp-norm minimization can result in a sparser solution than the L_1 -norm and perform better in terms of robustness to outliers than the L_2 -norm in a sense, we use it to constrain the loss function of the PL21GPCA optimization model. The generalized shrinkage operation proposed by Chartrand (2012) is adopted to solve the function effectively in our method.

The $L_{2,1}$ -norm of matrix X is as follows:

$$\|X\|_{2,1} = \sum_{i=1}^p \sqrt{\sum_{j=1}^n x_{ij}^2} = \sum_{i=1}^p \|x_i\|_2 \quad (2)$$

where x_i (corresponding to feature i) is the i th row of matrix X . Yang et al. (2012) provided an intuitive explanation of the $L_{2,1}$ -norm in the literature. To solve the $L_{2,1}$ -norm, we can compute the L_2 -norm of each row of X first, record it as a vector $b(X) = (\|x_1\|_2, \|x_2\|_2, \dots, \|x_p\|_2)$, and then compute the L_1 -norm of vector $b(X)$. The components of vector b indicate the importance of each feature. The $L_{2,1}$ -norm favors obtaining a small number of non-zero rows in matrix X , and then feature selection will be achieved.

PCA and Graph Laplacian Embedding

Principal Component Analysis (PCA)

Let $X = (x_1, \dots, x_n) \in R^{p \times n}$ ($p \gg n$) be a matrix whose rows represent genes and columns represent samples. PCA is usually used to find the optimal principal directions $V^T = (v_1, \dots, v_n) \in R^{k \times n}$ ($V^T V = I$) that define the low-dimensional (k -dim) subspace. And the projected data points in the low subspace V can be denoted as the elements of the matrix $U_{p \times k} = (u_1, \dots, u_k) \in R^{p \times k}$. The traditional PCA finds U and V with the squared Frobenius norm:

$$\arg \min_{U, V} \|X - UV^T\|_F^2 \quad s.t. \quad V^T V = I \quad (3)$$

In our optimization model, the proximal L_p -norm $\|g\|_p$ ($0 < p < 1$) (Chartrand, 2012; Nie et al., 2013; Feng et al., 2017) is used instead of the traditional quadratic loss function $\|g\|_F$ to reduce the influence of outliers and noise. PCA naturally relates closely to the classic clustering means known as K-means (Ding and He, 2004). The optimal principal components contained in matrix V provide the solution of the K-means clustering method. It inspired us to combine PCA with Laplacian embedding, whose principal purpose is also clustering.

Graph Laplacian Embedding

Principal component analysis can find an approximate set of basis vectors in the case where data usually lie in a linear manifold (Jiang et al., 2013). In consideration of the local invariance of the intrinsic geometric structure of the data distribution, graph Laplacian embedding is a popular method among recent studies in non-linear manifold learning theory (Belkin and Niyogi, 2002, 2003; Spielman, 2007). The assumption of local invariance is that if two points (samples) are close in the intrinsic geometry of the original data distribution, the representations of these two points (samples) in the new coordinate are also close to each other. The local geometric structure can be modeled through a nearest neighbor graph on a scatter of data points. Given the data matrix $X = (x_1, \dots, x_n) \in R^{p \times n}$, $x_i (i = 1, \dots, n)$ can be regarded as one data point (one vertex in the graph). For each data point x_i , we find its k' nearest neighbors and put edges between x_i and its neighbors. Then, a graph with n vertices can be constructed, on which the weight matrix

$W \in R^{n \times n}$ is defined. w_{ij} is the weight between vertices x_i and x_j , it is used to measure the closeness of two points x_i and x_j , and it is a symmetric similarity matrix. There are three popular choices defining the weight matrix on the graph: heat kernel weighting, 0-1 weighting, and dot-product weighting. If nodes i and j are connected, using heat kernel weighting, $w_{ij} = e^{-\frac{\|x_i - x_j\|^2}{\sigma}}$, $w_{ij} = 1$ using 0-1 weighting and $w_{ij} = w_i^T w_j$ using dot-product weighting. The different similarity measures are suitable for different situations. Detailed information about the different weighting schemes can be found in the literature (Cai et al., 2011).

Let $Z^T = (z_1, z_2, \dots, z_n) \in R^{k \times n}$ represent the n data points in the k -dim embedding coordinates $V^T = (v_1, \dots, v_n) \in R^{k \times n}$ ($V^T V = I$), i.e., the representation of x_i in the new low-dimensional basis is $z_i = [v_{i1}, \dots, v_{ik}]$. The “dissimilarity” of the two data points in the low basis can be measured by the Euclidean distance or the divergence distance. The Euclidean distance is adopted in our method. Define the “dissimilarity” of the two points in the low basis as $d(z_i, z_j) = \|z_i - z_j\|^2$, combined with the weight matrix W , and the smoothness of the low-dimensional representation can be measured by minimizing:

$$\begin{aligned} S &= \frac{1}{2} \sum_{i,j=1}^n \|z_i - z_j\|^2 w_{ij} \\ &= \sum_{i=1}^n z_i^T z_i D_{ii} - \sum_{i,j=1}^n z_i^T z_j w_{ij} \\ &= \text{Tr}(V^T D V) - \text{Tr}(V^T W V) = \text{Tr}(V^T L V) \end{aligned} \quad (4)$$

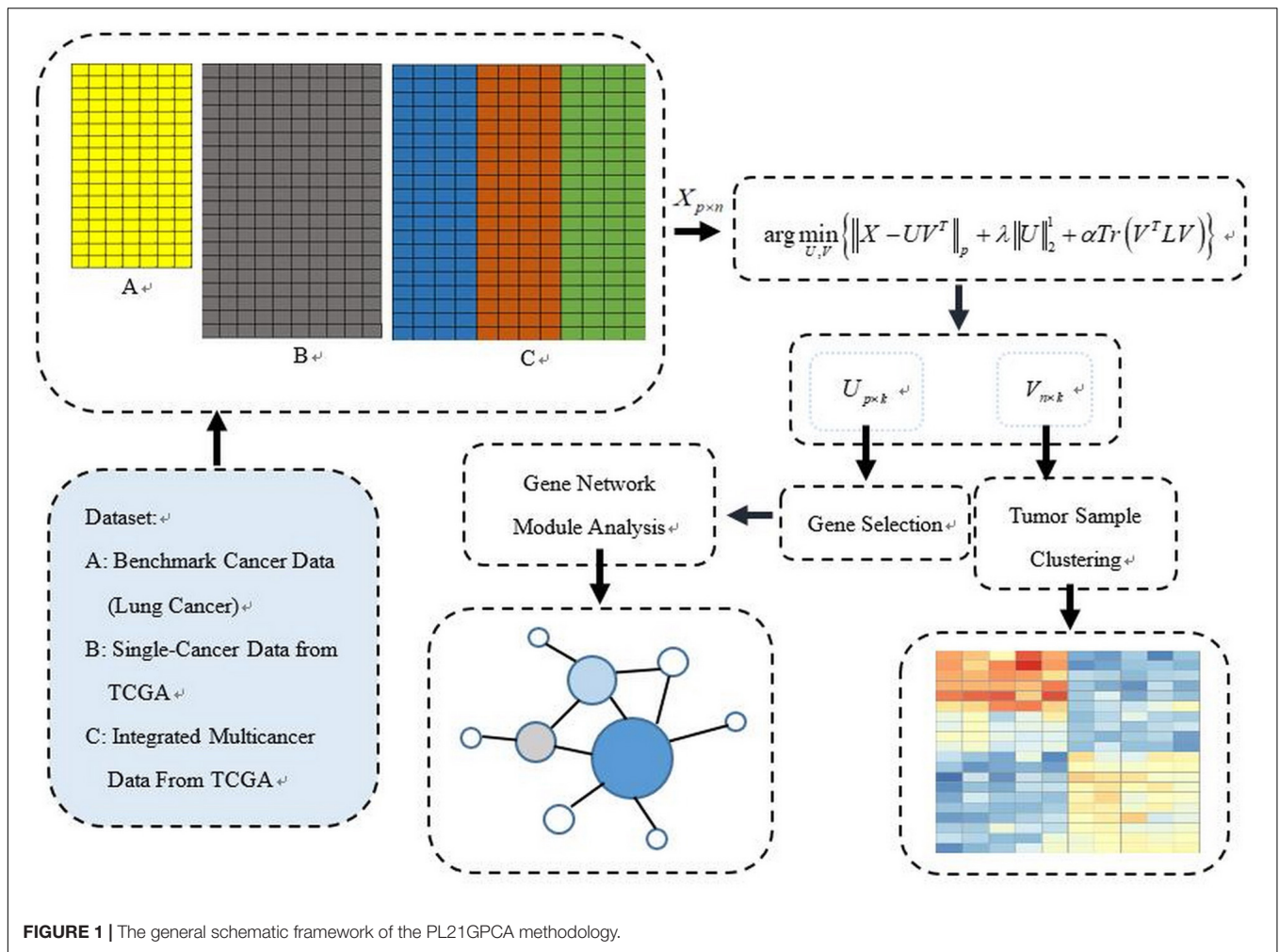
where $\text{Tr}(\bullet)$ is the trace of a matrix, $D = \text{diag}(d_1, \dots, d_n)$ is a diagonal matrix, and $d_i = \sum_{j=1}^n w_{ij}$. We call the $L = D - W$ the Laplacian matrix (Spielman, 2007).

METHODOLOGY

The PL21GPCA procedure is presented in this section. Figure 1 illustrates our general research framework. In brief, our work includes two steps. The first is obtaining the optimal projected matrix $U_{p \times k}$ and the principal directions matrix $V_{k \times n}$ via PL21GPCA. The second is to evaluate the validity of PL21GPCA. In this step, based on the principal directions matrix $V_{k \times n}$ obtained by PL21GPCA, the classic clustering method K-means is employed for tumor sample clustering. According to the projected matrix $U_{p \times k}$, the differentially expressed genes are selected to carry out gene network analysis to find the key genes with close relationships to some cancers.

To summarize, three aspects are highlighted in our method:

- (1) To reduce the influence of outliers and noise, the non-convex proximal L_p -norm $\|g\|_p$ ($0 < p < 1$) is used on the loss function, which could improve the robustness of the optimization model effectively compared with the other constraints.



- (2) To enhance the sparsity of gene expression in cancer samples, the $L_{2,1}$ -norm is used on the projected matrix $U_{p \times k}$.
- (3) To retain the intrinsic geometric structure of the data points (samples), the graph regularization item is recommended in the optimization model.

Assume the input matrix $X = (x_1, \dots, x_n) \in \mathbb{R}^{p \times n}$ ($p \gg n$), which denotes p genes' expression in n samples. Our goal is to find the optimal low-dimensional (k -dim) subspace denoted as $V^T = (v_1, \dots, v_n) \in \mathbb{R}^{k \times n}$ ($V^T V = I$) and the projected matrix $U_{p \times k} = (u_1, \dots, u_k) \in \mathbb{R}^{p \times k}$ in the low subspace. The traditional PCA finds U and V with the squared Frobenius norm in the solution. In our optimization model, the proximal L_p -norm $\|g\|_p$ ($0 < p < 1$) (Chartrand, 2012; Nie et al., 2013; Feng et al., 2017) replaces the traditional quadratic loss function $\|g\|_F$ to reduce the influence of outliers and noise. The $L_{2,1}$ -norm is used on one of the regularization terms to enhance the sparse gene expression in cancer samples. The graph Laplacian regularization item emphasizing the local invariance of the intrinsic geometric structure is recommended in the optimization model.

The objective function of this method is designed as follows:

$$\arg \min_{U, V} \{ \|X - UV^T\|_p + \lambda \|U\|_2^1 + \alpha \text{Tr}(V^T L V) \} \quad (5)$$

$$\text{s.t. } V^T V = I, 0 < p < 1, \lambda > 0, \alpha > 0$$

Clearly, the objective function is somewhat intractable because it is non-convex and non-smooth. We adopt the augmented Lagrangian multiplier (ALM) (Hestenes, 1969; Bertsekas, 1982; Spielman, 2007; Lin et al., 2010) to address this optimization problem. Researchers have proven that the ALM algorithm possesses Q-linear convergence properties under some conditions (Bertsekas, 1982).

When using the ALM method to obtain the optimal solution of (5), we replace $X - UV^T$ with E . Eq. (5) can be equivalently written as:

$$\arg \min_{E, U, V} \{ \|E\|_p + \lambda \|U\|_2^1 + \alpha \text{Tr}(V^T L V) \} \quad (6)$$

$$\text{s.t. } E - X + UV^T = 0, V^T V = I$$

According to the ALM method, eq. (6) is equivalent to minimizing:

$$L_{\mu,Y}(E, U, V) = \|E\|_p + \frac{\mu}{2} \|E - X + UV^T + \frac{Y}{\mu}\|_F^2 + \lambda \|U\|_2^2 + \alpha \text{Tr}(V^T LV) \quad (7)$$

where γ is the Lagrangian multiplier, and μ is the step size of the update rule. In (7), there are three variables to be solved. The alternating direction method (ADM) (Gabay and Mercier, 1976) is adopted to tackle this thorny problem because the equation with only one variable is easily solved when the others are fixed. By this means, (7) naturally results in three subproblems.

Problem 1: When U and V are fixed, (7) is written as follows:

$$L_{\mu,Y}(E, U, V) = \|E\|_p + \frac{\mu}{2} \|E - X + UV^T + \frac{Y}{\mu}\|_F^2 \quad (8)$$

where $0 < p < 1$. Eq. (8) can be solved by the proximal shrink operator denoted as follows:

$$\text{shrink}_p(t, \delta) := \max\{0, |t| - \delta|t|^{p-1}\} \frac{t}{|t|} \quad (9)$$

Let $t = X - UV^T - \frac{Y}{\mu}$, $\delta = \frac{1}{\mu}$. Then, according to the shrinkage operation (soft thresholding) proposed by Chartrand (2012), E is updated as:

$$E^{r+1} = \text{shrink}_p\left\{X - U^r (V^r)^T - \frac{Y^r}{\mu^r}, \frac{1}{\mu^r}\right\} \quad (10)$$

Problem 2: When E and V are fixed, (7) is simplified as follows:

$$L_{\mu,Y}(E, U, V) = \frac{\mu}{2} \|E - X + UV^T + \frac{Y}{\mu}\|_F^2 + \lambda \|U\|_2^2 \quad (11)$$

To simplify (11), let $H = X - E - \frac{Y}{\mu}$. Then, (11) is written as:

$$L_{\mu,Y}(E, U, V) = \frac{\mu}{2} \|UV^T - H\|_F^2 + \lambda \|U\|_2^2 \quad (12)$$

The partial derivatives of L with respect to U are:

$$\frac{\partial L}{\partial U} = \mu(UV^T - H)V + 2\lambda QU \quad (13)$$

where $Q \in \mathbb{R}^{p \times p}$ is a diagonal matrix with $q_{i,i} = \frac{1}{\|U_{(i,:)}\|_2}$ ($i = 1, \dots, p$) (Xiang et al., 2012). Letting (13) be equal to 0, the following update rule for U is then obtained:

$$U^{r+1} = \left(I + \frac{2\lambda}{\mu^r} Q^r\right)^{-1} H^r V^r \quad (14)$$

To simplify (14), let $A^r = \left(I + \frac{2\lambda}{\mu^r} Q^r\right)^{-1}$, and then (14) is written as:

$$U^{r+1} = A^r H^r V^r \quad (15)$$

Problem 3: When E and V are fixed, (7) is simplified as follows:

$$L_{\mu,Y}(E, U, V) = \frac{\mu}{2} \|E - X + UV^T + \frac{Y}{\mu}\|_F^2 + \alpha \text{Tr}(V^T LV) \quad (16)$$

With respect to the settings $H = X - E - \frac{Y}{\mu}$, (16) can be written equivalently as:

$$\begin{aligned} L_{\mu}(E, U, V) &= \frac{\mu}{2} \|UV^T - H\|_F^2 + \alpha \text{Tr}(V^T LV) \\ &= \frac{\mu}{2} \text{Tr}((UV^T - H)(UV^T - H)^T) + \alpha \text{Tr}(V^T LV) \end{aligned} \quad (17)$$

Based on (17), V is found by minimizing:

$$V = \arg \min_V \text{Tr}(V^T (\frac{\alpha}{\mu} L - H^T A H) V) \quad (18)$$

Therefore, V^{r+1} can be obtained as follows:

$$V^{r+1} = (v_1, \dots, v_k) \quad (19)$$

where (v_1, \dots, v_k) are the k eigenvectors corresponding to the smallest k eigenvalues of the matrix $\frac{\alpha}{\mu} L - H^T A H$. Thus, based on the ALM, ADM and the shrinkage operation, the solution to solve the optimization model described in (5) is shown in **Algorithm 1**. In the optimization model, there are six parameters $k, p, \lambda, \alpha, \rho, \mu$ to be pre-determined, among them. As the parameters used to control the step size in the update rule of AML, we set $\mu = 10^{-2}$ and $\rho = 1.2$ for all gene expression datasets experiments (Feng et al., 2016). The parameter k is determined referring to the number of prior categories of each dataset. For the three essential parameters p, λ, α , to be determined in (5), we choose them corresponding to different situations for the best clustering performance through extensive experiments. Different parameters are chosen for different datasets. Detailed parameter selections and discussions are described in section “Experiments and Discussion.”

EXPERIMENTS AND DISCUSSION

Gene Expression Datasets

Five gene expression datasets, which include one benchmark dataset, two single-cancer datasets from TCGA, and two integrated multicancer datasets from TCGA, are used to evaluate

ALGORITHM 1 | The solution to optimized (5).

Input:

Gene expression data matrix: $X_{p \times n}$,

Parameters: $k, p, \lambda, \alpha, \rho, \mu$

Output:

$U_{p \times k}, V_{n \times k}$

Initialize:

E, Y, U, V

Do

Update U by (14)

Update V by (19)

Update E by (10)

Update μ by $\mu = \rho\mu$

Update Y by $Y^{r+1} = Y^r + \mu^r (E^r - X + U^r (V^r)^T)$

Update μ by $\mu^{r+1} = \rho\mu^r$

Until convergence

the performance of PL21GPCA. The verified experiments consist of two aspects: “tumor sample clustering” and “gene network module discovery.” Based on the optimal low-dimensional (k -dim) subspace denoted as $V^T = (v_1, \dots, v_n) \in \mathbb{R}^{k \times n}$ ($V^T V = I$), the classical clustering method K-means is then used for tumor clustering. For comparison, extensive experiments are also performed using existing dimensionality reduction methods, including SPCA (Journée et al., 2010), RPCA (Candès et al., 2009), gLPCA (Jiang et al., 2013), pgLPCA (Feng et al., 2017) and GNMF (Cai et al., 2011). Among the compared methods, some are based on PCA, and some introduce the graph Laplacian regularization item. Based on the optimal projected matrix $U_{p \times k}$, the differentially expressed genes are selected for gene network analysis to find key genes with close relationships to some cancers.

The details of the five data sets are as follows. The benchmark gene expression dataset is lung cancer data (Bhattacharjee et al., 2001) that have often been employed by researchers to evaluate their algorithms (Lee et al., 2010; Kong et al., 2017), consisting of 12,625 genes of 56 samples. There are four types of lung cancer in the 56 samples: pulmonary carcinoid (20), colon metastases (13), small cell lung carcinoma samples (6) and normal lung samples (17). The two single-cancer datasets and the two integrated multicancer datasets are all from The Cancer Genome Atlas (TCGA) which is known as the largest tumor specimens database. The genomic data provided by TCGA include DNA methylation, microRNA expression, gene expression, protein expression, and DNA copy number, etc. We downloaded gene expression datasets (at level 3) of five different cancers from TCGA: colorectal cancer (CRC), cholangiocarcinoma (CHOL), squamous cell carcinoma of head and neck (HNSC), pancreatic cancer (PAAD), and esophageal cancer (ESCA). Each dataset consists of 20,502 genes expressed in different numbers of samples. In our experiments, CRC and CHOL are used as single-cancer datasets to evaluate the performance of the PL21GPCA method. There are 281 samples for CRC and 45 for CHOL. Each of these two datasets contains two types of cancer samples, “negative” and “positive.” “Negative” or “NT” represents normal samples. “Positive” or “TP” represents diseased samples. There are 262 “TP” samples in the CRC data and 36 in the CHOL data, and the rest are “NT” samples. Two integrated datasets are used to further verify the performance of the PL21GPCA method. Each integrated dataset consists of 3 types of cancers. One of the integrated datasets, H_C_P, contains 836 “TP” samples, among which the sample numbers of the three cancers are 398 (HNSC), 262 (CRC), and 176 (PAAD). The other integrated dataset, E_C_C, contains 481 “TP” samples, in which the sample numbers of the three cancers are 183 (ESCA), 36 (CHOL), and 262 (CRC). The statistics of these datasets are summarized in Table 1.

Tumor Sample Clustering

Evaluation Metric

Based on the optimal principal directions $V^T = (v_1, \dots, v_n) \in \mathbb{R}^{k \times n}$ ($V^T V = I$), the K-means algorithm is then employed for tumor sample clustering. The accuracy (ACC) and the normalized mutual information (NMI) are the two most

commonly used metrics to evaluate the clustering results (Cai et al., 2005). For the i th sample, we use p_i to denote the prior label and r_i to denote the obtained clustering label. The metric ACC is defined as follows:

$$ACC = \frac{\sum_{i=1}^n \theta(p_i, \text{map}(r_i))}{n}, \quad (20)$$

where n denotes the total number of samples in every dataset. The function $\theta(x, y)$ equals 1 if $x = y$ and 0 otherwise. The function $\text{map}(r_i)$ maps each obtained cluster label r_i to the equivalent prior label. Let C be the prior set of clusters and C' be the obtained set from our algorithm. Define their mutual information metric $MI(C, C')$ as:

$$MI(C, C') = \sum p(c_i, c'_j) \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdots p(c'_j)} \quad (21)$$

where $p(c_i)$ and $p(c'_j)$ are the probabilities that a sample arbitrarily selected from the dataset belongs to clusters c_i and c'_j , respectively, and $p(c_i, c'_j)$ is the joint probability. In the experiments, the metric NMI is defined as follows:

$$NMI(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))} \quad (22)$$

where $H(C)$ and $H(C')$ are the entropies of C and C' , respectively. Therefore, the metric $NMI(C, C')$ ranges from 0 to 1. $NMI = 1$ if the two sets of clusters are identical, and if the two sets are independent, $NMI = 0$.

However, a problem that needs to be resolved is that the K-means algorithm may or may not converge to the same solution in each run with random initial conditions. Therefore, the evaluated metrics ACC and NMI obtained by only once-running of k-means is not enough to explain the result. To solve this problem, for the given cluster number k , K-means was run 50 times on each dataset, and the average performance was computed. As a reference, we also recorded the maximum values of ACC and NMI of the 50 runs. Thus, four metrics, ACC_max, ACC_mean, NMI_max and NMI_mean, are used to evaluate our experiments. Generally, the larger the mean value is, the better is the clustering performance, and the better are the stability and robustness of the clustering. This also indicates that the corresponding dimension reduction method has good robustness and sparse effect.

TABLE 1 | Statistical information on the experimental data.

Dataset		# of genes (p)	# of samples (n)	# of classes (k)
Benchmark Data	Lung Cancer	12625	56	4
Single-Cancer Data from TCGA	CRC	20502	281	2
	CHOL	20502	45	2
Integrated Cancer Data from TCGA	H_C_P	20502	836	3
	E_C_C	20502	481	3

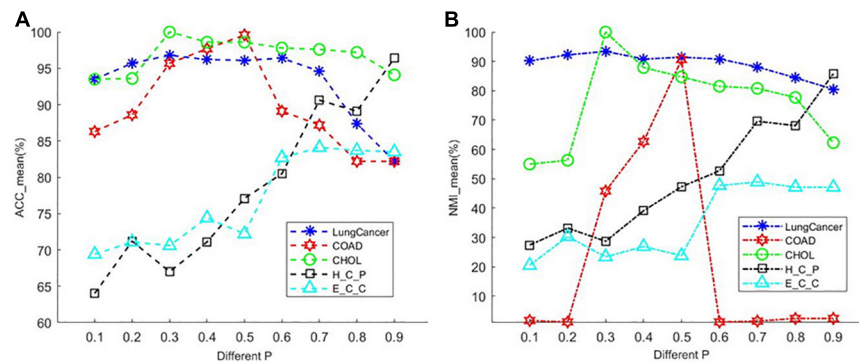


FIGURE 2 | The average performance taking the essential parameter at nine different values from 0.1 to 0.9. **(A)** The mean value of ACC for different cancer datasets. **(B)** The mean value of NMI for different cancer datasets.

Parameter Selection

The PL21GPCA model has three essential parameters, p , λ , and α , which need to be determined in (5). The range of each parameter is $0 < p < 1$, $\lambda > 0$, $\alpha > 0$. When determining the optimal value of one parameter, the other two parameters are fixed. We focus on the influence of the value of p on the performance. PL21GPCA achieves consistently good performance when the two regularization parameters λ and α are varied from 10 to 1,000 on all three datasets. **Figure 2** shows how the average performance varies when taking the essential parameter p at nine different values from 0.1 to 0.9. For every dataset, extensive experiments are carried out to seek the appropriate parameters to achieve the best performance for tumor sample clustering. Thus, different parameters are chosen for different datasets (see **Table 2**).

There is another parameter that is not appear in the objective function of PL21GPCA. However, it is also an important parameter affecting the performance of our method. It is parameter k' , the number of nearest neighbors of every point when constructing the graph in the step of graph Laplacian embedding. Setting this parameter too small may cause overfitting, and too large may increase the error. By extensive experiments, we find that the appropriate value for this parameter is near the square root of the sample number for different datasets.

Clustering Results

Tables 3–5 show the clustering results on the lung cancer data, single-cancer data from TCGA (CRC and CHOL datasets), and integrated cancer data (H_C_P and E_C_C datasets), comparing the PL21GPCA-based method with the competitors. For each

dataset with a given cluster number k , the K-means algorithm was run 50 times to randomize the experiments. The maximum and the mean value metrics are all presented in the tables. The performance of the PL21GPCA-based method is highlighted in bold in the tables. Regardless of the datasets, the PL21GPCA-based method always results in the best performance on the mean value metrics ACC_mean and NMI_mean. As mentioned above, the mean value is more meaningful than the maximum value, which is for reference only. By leveraging the power of three

TABLE 3 | Clustering performance on lung cancer.

Methods	ACC (%)		NMI (%)	
	ACC_Max	ACC_mean	NMI_Max	NMI_mean
SPCA	100	84.39	100	83.07
RPCA	100	86.25	100	84.77
GNMF	85.71	79.71	75.57	69.62
gLPCA	89.29	78.5	80.82	69.86
pgLPCA	100	82	100	80.05
PL21GPCA	100	96.82	100	93.44

TABLE 4 | Clustering performance on CRC and CHOL.

Data	Method	ACC (%)		NMI (%)	
		ACC_Max	ACC_mean	NMI_Max	NMI_mean
CRC	SPCA	92.17	87.57	35.3	22.57
	RPCA	98.22	67.95	69.82	24.33
	GNMF	88.61	60.5	30.79	18.93
	gLPCA	90.75	87.01	22.7	15
	pgLPCA	94.31	78.65	42.67	20.1
	PL21GPCA	99.64	99.64	90.55	90.55
CHOL	SPCA	100	93.38	100	60.65
	RPCA	100	100	100	100
	GNMF	100	100	100	100
	gLPCA	100	78.04	100	54.87
	pgLPCA	100	81.87	100	59.83
	PL21GPCA	100	100	100	100

TABLE 2 | Values of the three parameters p , λ , and α for different datasets.

Dataset	Lung Cancer	CRC	CHOL	H_C_P	E_C_C
Parameter selections	$p = 0.3$	$p = 0.5$	$p = 0.3$	$p = 0.9$	$p = 0.7$
	$\lambda = 10$	$\lambda = 100$	$\lambda = 10$	$\lambda = 100$	$\lambda = 10$
	$\alpha = 100$	$\alpha = 100$	$\alpha = 100$	$\alpha = 100$	$\alpha = 100$

TABLE 5 | Clustering performance on H_C_P and E_C_C.

Data	Method	ACC (%)		NMI (%)	
		ACC_Max	ACC_mean	NMI_Max	NMI_mean
H_C_P	SPCA	55.26	51.82	17.85	14.98
	RPCA	<i>91.87</i>	<i>77.3</i>	<i>71.43</i>	<i>68</i>
	GNMF	57.3	54.02	29.59	22.59
	gLPCA	55.62	52.96	29.43	16.89
	pgLPCA	86.96	70.26	58.42	45.4
	PL21GPCA	96.41	96.41	85.77	85.75
E_C_C	SPCA	71.52	67.9	19.28	15.06
	RPCA	<i>81.08</i>	<i>76.17</i>	<i>55.72</i>	<i>32.47</i>
	GNMF	68.4	62.05	19.03	9.29
	gLPCA	70.69	69.58	23.14	19.7
	pgLPCA	79.63	72.72	41.33	31.35
	PL21GPCA	85.65	84.09	60.31	47.15

measures, including taking the proximal L_p -norm $\|g\|_p$ ($0 < p < 1$) on the loss function, employing the $L_{2,1}$ -norm regularization item to insure feature selection, and introducing the Laplacian regularization item to emphasize the geometrical structure of the data, the PL21GPCA-based method can always get a better clustering performance.

For the different types of data used in the experiments, a number of meaningful points need to be emphasized further.

The benchmark data

For the lung cancer dataset, **Table 3** shows that the PL21GPCA-based method achieves the same performance as SPCA, RPCA and pgLPCA considering the maximum value metrics (the ACC_max and the NIM_max are also 100%) but is obviously superior to the other methods in terms of the mean value metric (ACC_mean reaches 96.82% and the NIM_mean reaches 93.44%).

Single-cancer data from TCGA

Table 4 shows the clustering performance of the two single-cancer datasets from TCGA. For the CRC dataset, our method presents very superior performance compared with other methods, with the ACC_mean reaching 99.64% as well as the ACC_max. The good average performance shows the robustness of the PL21GPCA method. In addition, the two NMI metrics (all reaching 90.55%) also go far beyond the performance of other methods. For the CHOL dataset, all the methods achieve the same results (100%) when considering the maximum value metrics. Our method achieves the same performance (100%) as GNMF and RPCA in terms of the mean value metrics. A surmise is reported that there may be distinct discriminations for the two kinds of samples in the original CHOL data (Kong et al., 2017).

Integrated multicancer data from TCGA

Table 5 reports the estimation results on the two integrated datasets. It shows that the PL21GPCA method performs much better than the competitors. As highlighted in bold in **Table 5**, for H_C_P data, the ACC_max and the ACC_mean all reach 96.41%, and the NMI_max and the NMI_mean are also superior to the corresponding values for other methods. For

E_C_C data, our method is still outstanding; taking the ACC metric as an example, the ACC_max reaches 85.65%, and the ACC_mean reaches 84.09%. Based on the excellent performance on these two integrated datasets, should we speculate that the PL21GPCA method is more suitable for learning the compact representation of higher-dimensional and more complex data than its competitors, which needs further verification.

Finally, as we can see from **Tables 3–5**, among the compared methods, the RPCA method performs second to our method and better than the other competitors, such as SPCA, GNMF, gLPCA, and pgLPCA. The performance of RPCA is in italics in the tables. If the intrinsic geometric structure is introduced to RPCA, will the performance be improved further? This question is also worth further verification.

Embedding Evaluation

To further show the performance of the novel dimensionality reduction method compared others, a visualized data distribution of the low-dimensional embedding corresponding to the first two components of the PCA-based method are demonstrated. Besides the proposed method PL21GPCA, the results of three other methods including SPCA, gLPCA, pgLPCA are compared because these methods are also the direct extensions of PCA. **Figure 3** presents the sample clustering results in a two-dimensional space. We choose two representative datasets CRC data and H-C-P data to show the results. **Figures 3A–D** are the results of the compared methods SPCA, gLPCA, pgLPCA and PL21GPCA, respectively, on the CRC dataset. **Figures 3E–H** are the compared results of the four methods on the H-C-P dataset. No matter for the CRC data which contains two types of cancer samples, or for the H-C-P data which contains three types of cancer samples, SPCA and gLPCA make the samples from different categories being mixed together, and the pgLPCA can only separate the samples into categories roughly, so they have unideal clustering results. However, PL21GPCA make the embeddings of samples in clearer distribution. Therefore, the clustering results is better than the compared methods. The visualized results verified the robustness and the flexibility of the proposed model.

Experiments on Simulated data

Experiments on simulation data are also carried out to evaluate the effectiveness of PL21GPCA. The simulation data used in the experiment is a matrix $X_{3000 \times 80}$ generated by *rand* function in Matlab. In order to simulate the representation of features in different types of samples, based on the generated matrix $X_{3000 \times 80}$, some changes have also been made. Firstly, we add 1 to the values of columns 1 to 20 in rows $i^*30 - 29$ ($i = 1, \dots, 100$) of matrix $X_{3000 \times 80}$, add 2 to the values of columns 21 to 40 in rows $i^*30 - 19$ ($i = 1, \dots, 100$), add 3 to the values of columns 41 to 60 in rows $i^*30 - 9$ ($i = 1, \dots, 100$), add 4 to the values of columns 61 to 80 in rows $i^*30 - 5$ ($i = 1, \dots, 100$), add 2 to the values of columns 21 to 40 in rows $i^*30 - 25$ ($i = 1, \dots, 100$), add 1 to the values of columns 1 to 20 in rows $i^*30 - 15$ ($i = 1, \dots, 100$), which means that the 80 samples in the simulation data contain four categories. Secondly, we use the function *imnoise* in matlab to add different sizes of Gaussian white noise

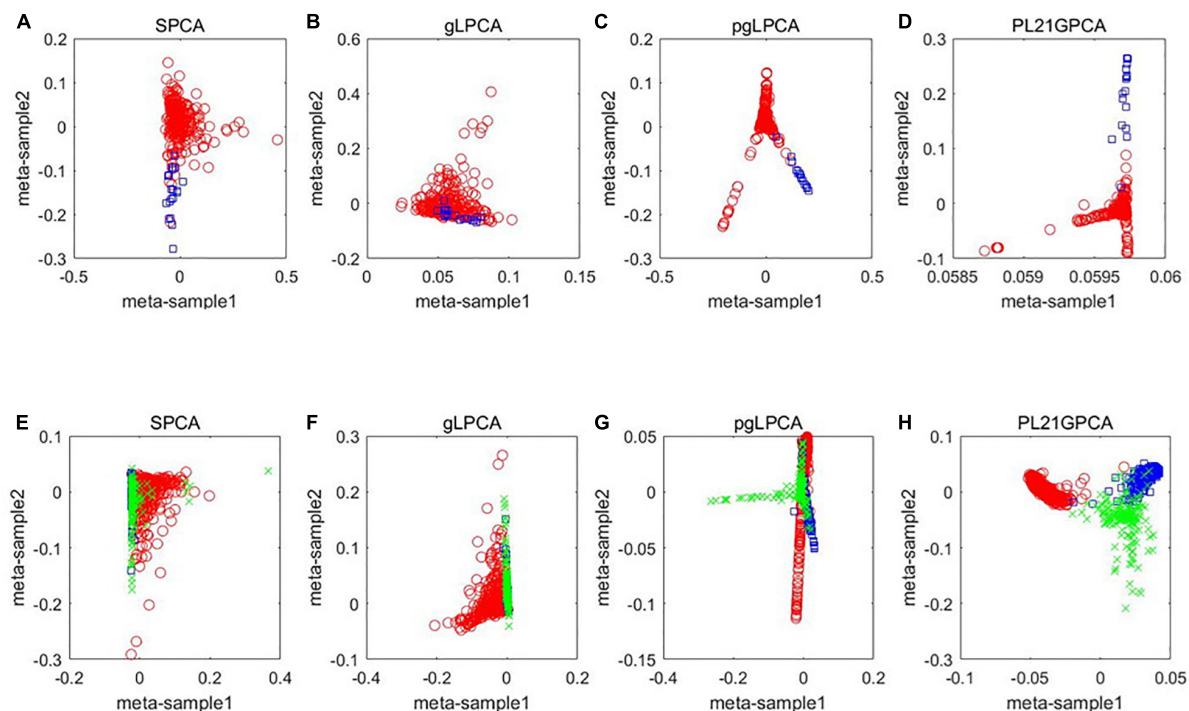


FIGURE 3 | A visualized comparison of low-dimensional embeddings by SPCA, gLPCA, pgLPCA, and PL21GPCA on COAD and H-C-P datasets. **(A–D)** Are the results of the compared methods SPCA, gLPCA, pgLPCA, and PL21GPCA respectively on the CRC dataset. **(E–H)** Are the compared results of the four methods on the H-C-P dataset.

to X . The mean value of the added Gaussian white noise is 0 and the variance σ^2 is chosen in the range of $[0.4 \sim 1.2]$. Next, we use the proposed method PL21GPCA and the compared methods to reduce the dimension and denoise the simulated data, and then use the K-means method to cluster the denoised data, the evaluation metric ACC_mean mentioned above is used to test the effectiveness of the method. the K-means algorithm is run 50 times to randomize the experiments.

Table 6 shows the experiments results on simulated data. It can be seen evidently that the performances of all methods change with the increase of noise. The best performance of different methods when adding different noises are marked with black bold. Although the performance of pl21GPCA is second only to RPCA when the noise is low ($\sigma^2 = 0.4$), with the increase of Gaussian white noise, the effect of our proposed method is mostly ahead of other methods especially when $\sigma^2 = 0.6, 0.8, 1.2$, which shows that the new method has better denoising ability and robustness.

TABLE 6 | Clustering performance on simulated data with different Gaussian white noise.

Simulated data	SPCA	RPCA	GNNF	gLPCA	gpLPCA	PL21GPCA
$\sigma^2=0.4$	96.6	99.75	95.35	87.37	89.47	99.45
$\sigma^2=0.6$	94.35	91.33	94.35	84.68	86.45	97.45
$\sigma^2=0.8$	85.87	91.1	93.85	83.2	85.57	94.35
$\sigma^2=1.0$	80.12	90.85	93.4	86.48	85.67	93.33
$\sigma^2=1.2$	70.25	76.43	73.58	85.83	82.12	87.15

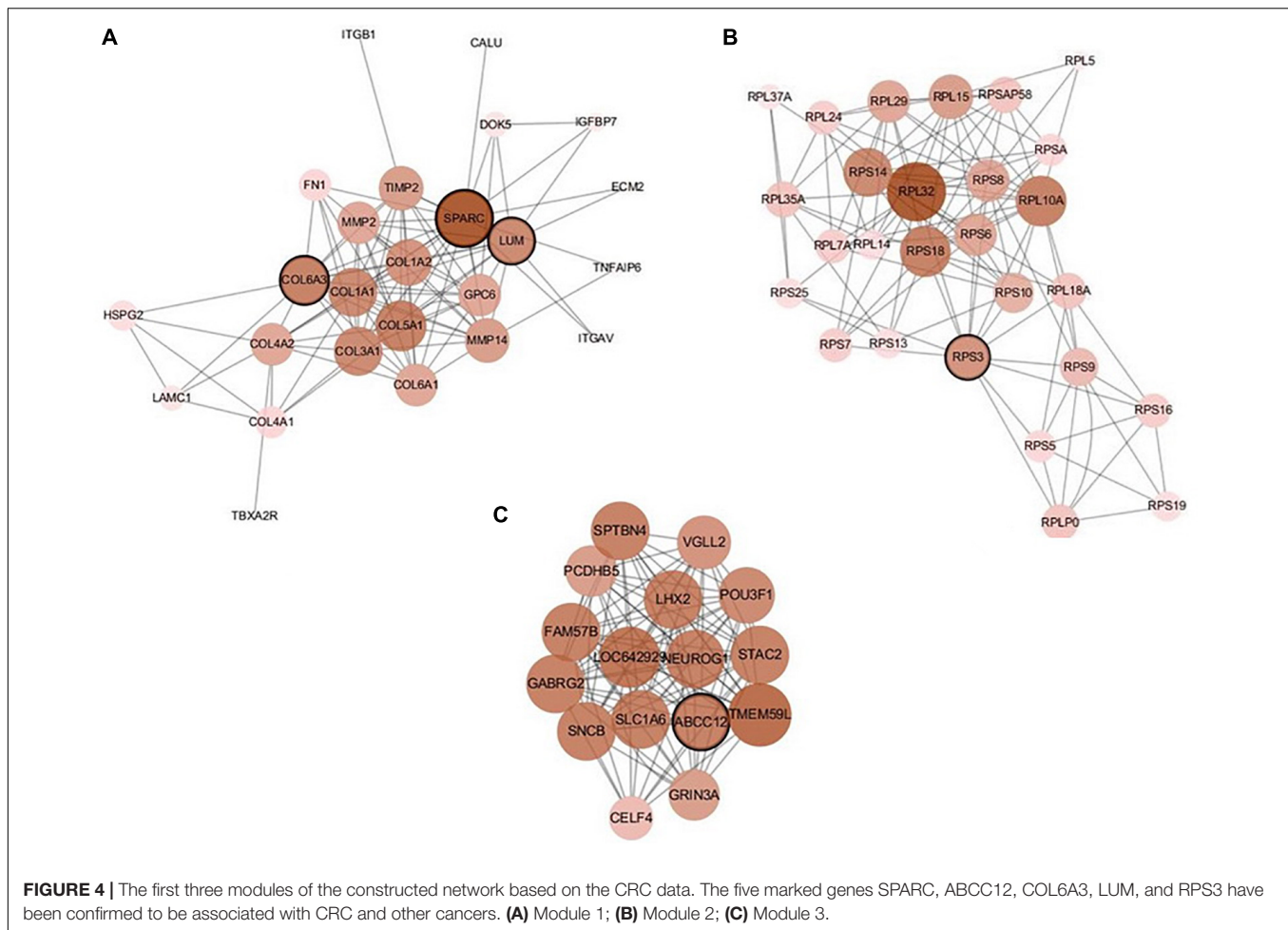
Gene Network Module Discovery

Due to the outstanding performance of our method on the CRC dataset and the integrated H_C_P dataset, the construction and analysis of the gene network are based on these two datasets. The strategy of gene network module discovery involves two steps. First, the genes for constructing the co-expression gene networks are selected. Second, based on the filtered genes, co-expression networks are established, and then the key genes that may be closely related to some cancers are analyzed.

Gene Selection

In this step, there are two problems to be solved: one is how to select genes, and the other is how many to select. It is known that among thousands of genes, only a handful of them regulate a specific biological process (Delbert et al., 2005; Liu et al., 2013). These minority of genes are called differentially expressed genes (Liu J. et al., 2015). In this article, the differentially expressed genes are selected to carry out gene network analysis according to the projected matrix $U_{p \times k}$. Now, we mark the optimal projected matrix $U_{p \times k}$ as \tilde{U} ; therefore, these differentially expressed genes can be identified according to \tilde{U} (Liu J. et al., 2015; Feng et al., 2016). We denote \tilde{U} as follows:

$$\tilde{U} = \begin{bmatrix} \tilde{u}_{11} & \tilde{u}_{12} & \cdots & \tilde{u}_{1k} \\ \tilde{u}_{21} & \tilde{u}_{22} & \cdots & \tilde{u}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \tilde{u}_{p1} & \tilde{u}_{p2} & \cdots & \tilde{u}_{pk} \end{bmatrix} \quad (23)$$



The upregulated genes are reflected by the positive value in the matrix \tilde{U} , and the downregulated genes are reflected by the negative value (Liu et al., 2013). Therefore, the absolute value of the items in \tilde{U} is used to identify the differentially expressed genes. The items of each row in \tilde{U} are summed, and then the evaluating vector denoted as \hat{U} is obtained:

$$\hat{U} = \left[\sum_{j=1}^k \tilde{u}_{1j} \quad \sum_{j=1}^k \tilde{u}_{2j} \quad \cdots \quad \sum_{j=1}^k \tilde{u}_{pj} \right]^T \quad (24)$$

The larger item in \hat{U} indicates the more strongly differentially expressed gene. Therefore, we sort the elements in \hat{U} in descending order and take the top l ($l \ll p$) elements. In many studies, it has been unclear how many genes should be selected for gene network analysis. Since only a small number of genes can regulate a specific biological process, these genes may play a decisive role in the clustering results of tumor samples. In this paper, the number of genes used for constructing the gene network is determined according to the clustering performance based on the selected genes. Through experimentally investigating the clustering performance with the number of selected genes varied from 500 to 2000, it is found that

the clustering results corresponding to **1600** genes are best for the CRC data and **700** for the H_C_P data.

Construction of Gene Networks

Suppose l differentially expressed genes are used to construct the gene network. Let matrix $R_{l \times n}$ denote the l gene expression in n samples. We use the Pearson correlation coefficient (PCC) (Hou et al., 2019) to measure the correlation of any two genes in $R_{l \times n}$. The values in the PCC matrix vary in the range of $[0, 1]$. The larger the PCC value is, the higher the correlation is. Based on matrix $R_{l \times n}$, an adjacency matrix $A_{l \times l}$ can be calculated. According to the adjacency matrix, an intuitive visualized graph of the gene interaction network composed of several modules is obtained.

Analysis of Gene Network Modules

There are 39 modules, including 218 nodes and 504 edges, in the constructed network based on the CRC data. We analyzed the top 10 nodes (genes) with higher degrees in the first three modules that retained more relevant interactions. The degree of the node (gene) shows its role in the network modules. The larger the degree of the node (gene) is, the more important the node (gene) is, and such nodes (genes) may retain the

tight connectivity of the network. **Figure 4** shows the main part of the first three gene network modules in which a small number of nodes whose degree is very low have been removed. The roles of the top ten genes in the first three modules are illustrated in **Figure 4**. The degree value of a node in **Figure 4** is represented by its size and color. The larger the node is, the darker its color is, which corresponds to a larger degree of the node. Referring to GeneCard with its website <http://www.genecards.org/>, we list the annotations of the top ten genes in **Table 7**. Five of the top ten genes have been validated as associated with multiple cancers: SPARC, ABCC12, COL6A3, LUM, and RPS3. The corresponding nodes of these genes are marked with a black outline in **Figure 4** and are also shown in bold in **Table 7**. In the literature (Liu Q. Z. et al., 2015), the gene SPARC has been recommended as a predictor of colorectal cancer. The gene ABCC12 is a human ATP binding cassette (ABC) transporter and is a multidrug resistance protein (MRP9). However, MRP9 has been recognized as an important target for the immunotherapy of breast cancer (Bera et al., 2002). Studies have shown that colorectal cancer can be predicted by the gene COL6A3 because it is overexpressed in samples of colorectal cancer. Therefore, COL6A3 is considered a potential diagnostic and prognostic marker gene for colorectal cancer (Qiao et al., 2015). As one of the members of the leucine-rich proteoglycan family, the gene Lumican (LUM) is overexpressed in many kinds of cancers, including colorectal, neuroendocrine, cervical, carcinoid, breast, and pancreatic cancer. LUM also causes the growth and invasion of pancreatic cancer (Ishiwata et al., 2007). The ribosomal protein gene S3 (RPS3) is also overexpressed in colorectal cancer. Researchers found an increase

in ribosome synthesis in patients with colorectal cancer (Pogue-Geile et al., 1991). Although the other five genes RPL32, TMEM59L, LOC642929, LHX2, and TLCD3B have not been identified in clinical studies indicating their effect on cancers, they may be considered candidate oncogenes because of their high ranking in our constructed gene network modules. By constructing co-expression gene network modules based on the CRC dataset, we found some disease-causing genes for colorectal cancer and other related cancers. It shows that constructing gene network modules via the genes filtered based on PL21GPCA can help us discover the key oncogenes.

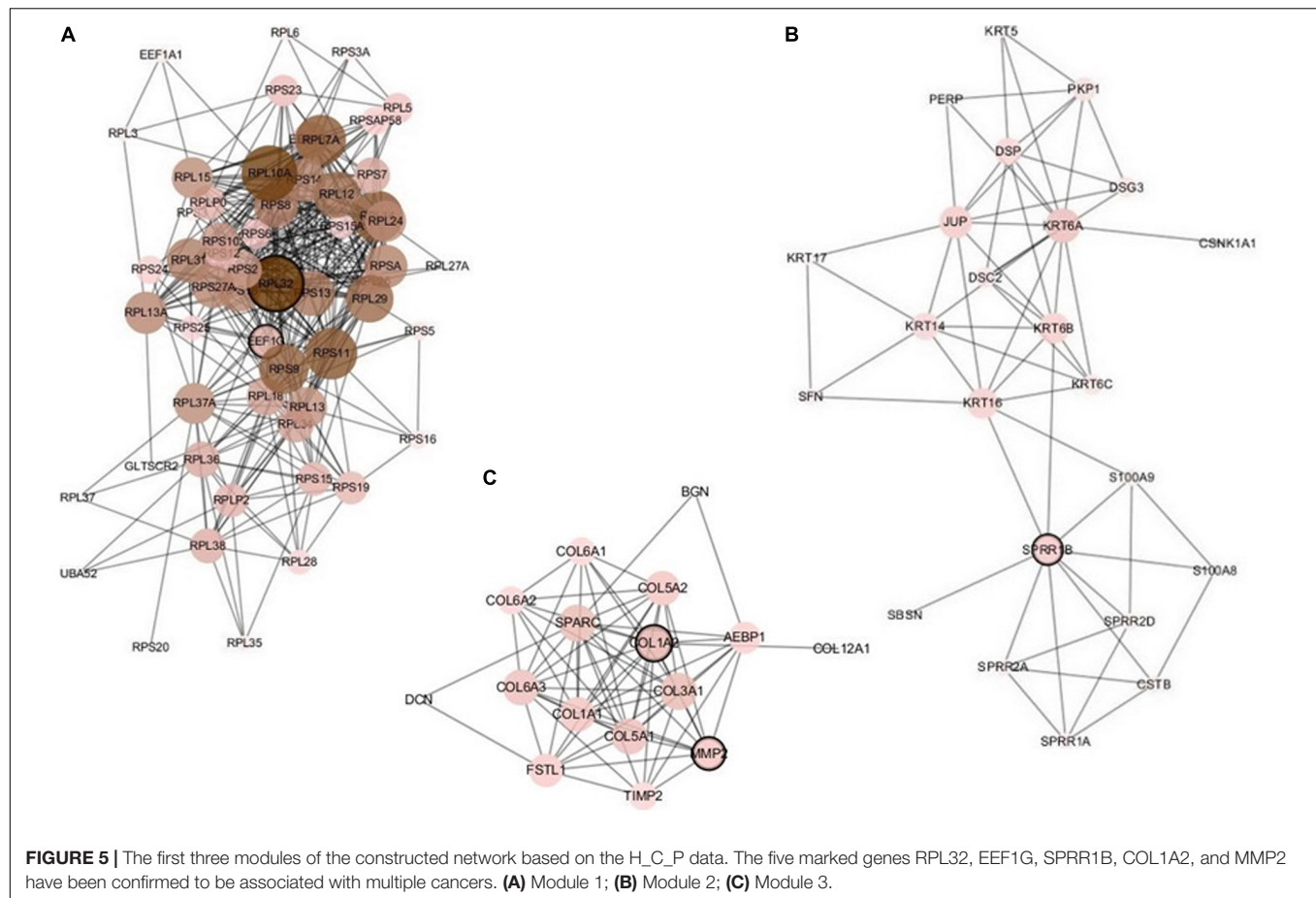
The constructed network based on the integrated data H_C_P includes 157 nodes and 644 edges. We analyzed the five important nodes (genes) with higher degrees in the first three modules that retained more relevant interactions. **Figure 5** illustrates the main part of the first three gene network modules in which the nodes of very low degree have also been removed. Referring to GeneCards, their annotations are listed in **Table 8**. The five genes RPL32, EEF1G, SPRR1B, COL1A2, and MMP2 have been recognized to be related to multiple cancers. The corresponding nodes of these genes are marked with a black outline in **Figure 5**. Wan et al. (2004) conducted large-scale experiments on human liver cancer cells. Research has shown that RPL32 is one of the potential genes that affect human cell growth and cancer formation and provides an important tool for diagnostic markers and drug targets (Wan et al., 2004). EEF1G has been thought to be a characteristic gene for colorectal cancer; it is highly expressed in most colorectal cancers and could be considered a marker gene for colorectal cancer detection (Matassa et al., 2013). In addition, the expression level of EEF1G in pancreatic tumor cells was higher than that in normal cells (Lew et al., 1992). SPRR1B is overexpressed in human oral squamous cells. It has been experimentally proven that SPRR1B overexpression in cells will signal MAP kinases but inhibit MAP kinase signals, so SPRR1B can affect cell growth and maintenance (Michifuri et al., 2013). Kiyoshi Misawa and other researchers mainly studied the expression of COL1A2 in head and neck squamous cell carcinoma (HNSC) and found that hypermethylation of CpG may cause inactivation of the gene COL1A2. Therefore, the COL1A2 gene may affect the formation and development of HNSC and could become a major biomarker (Misawa et al., 2011). As a member of the matrix metalloproteinase (MMP) gene family, MMP2 is relevant to the generation of malignant tumors, including colorectal cancer, lung cancer, and breast cancer (Yu et al., 2002; Arajo et al., 2015; Ren et al., 2015). Analysis through the gene network constructed based on integrated multicancer data is helpful for mining the interrelationships between different cancers and genes. It may provide an important reference for the diagnosis and treatment of various diseases.

CONCLUSION AND SUGGESTIONS

In this article, we propose a new dimensionality reduction method named PL21GPCA based on PCA for robust tumor sample clustering and gene network module discovery. Based

TABLE 7 | Annotations of the top ten genes in the first three network modules based on CRC data.

Gene	Summary
RPL32	A protein coding gene. Diseases associated with RPL32 include frontal convexity meningioma and retinitis pigmentosa 49
SPARC	Diseases associated with SPARC include osteogenesis imperfecta, type xvii and osteogenesis imperfecta, type iv
TMEM59L	TMEM59L (Transmembrane Protein 59 Like) is a protein coding gene. An important paralog of this gene is TMEM59
LOC642929	LOC642929 (General Transcription Factor II, I Pseudogene) is a pseudogene
ABCC12	Diseases associated with ABCC12 include familial cold autoinflammatory syndrome 1 and episodic kinesigenic dyskinesia 1. An important paralog of this gene is ABCC11
COL6A3	A protein coding gene. An important paralog of this gene is COL6A6
LUM	Among its related pathways are defective ST3GAL3, which causes MCT12 and EIEE15, and keratin sulfate/keratin metabolism
LHX2	LHX2 (LIM Homeobox 2) is a protein coding gene. Diseases associated with LHX2 include schizencephaly and retinitis pigmentosa
TLCD3B	TLCD3B (TLC Domain Containing 3B) is a protein coding gene. An important paralog of this gene is TLCD3A
RPS3	Diseases associated with RPS3 include eumycotic mycetoma and Waardenburg syndrome, type 3



on the traditional PCA, the non-convex proximal L_p -norm $\|g\|_p$ ($0 < p < 1$) is applied on the loss function to decrease the sensitivity to outliers and noise. The $L_{2,1}$ -norm is used on the projected matrix to enhance the sparse gene expression in cancer samples. The graph regularization item is introduced to the optimization model to retain the geometric structure of the data. Five gene expression datasets, including one benchmark dataset, two higher-dimensional single-cancer datasets from TCGA, and two integrated multicancer datasets from TCGA, are used to evaluate the performance of our method. The compared experiments demonstrate that the PL21GPCA method outperforms many existing methods in terms of tumor sample clustering. Moreover, this method is employed to discover gene network modules to find the key genes with close relationships to cancers. The results of our study may be a useful reference for clinical diagnosis.

There are some suggestions for future research. First, in the optimization model of PL21GPCA, the constraint used on the loss function is the non-convex proximal L_p -norm $\|g\|_p$ ($0 < p < 1$), since L_p -norm minimization can result in a sparser solution than the L_1 -norm and perform better in terms of robustness to outliers than the L_2 -norm. However, in addition to the generalized shrinkage operation proposed by Chartrand (2012), there are some other suggestions to address

the L_p -norm ($0 < p < 1$) minimization (Guo et al., 2013; Qin et al., 2013) problems. Therefore, we will continue to explore other solutions to the optimization model with the L_p -norm $\|g\|_p$ ($0 < p < 1$). Second, we will evaluate the performance of PL21GPCA as a compact representation method combined with other methods, including supervised and unsupervised clustering methods such as spectral clustering, support vector machine (SVM) or their improved versions. Third, as mentioned above, the PL21GPCA method gets especially outstanding performance

TABLE 8 | Annotations of the most important five genes in the first three network modules based on H_C_P data.

Gene	Summary
RPL32	A protein coding gene. Diseases associated with RPL32 include frontal convexity meningioma and retinitis pigmentosa 49
EEF1G	Diseases associated with EEF1G include gastrointestinal carcinoma. Among its related pathways are viral mRNA translation and gene expression
SPRR1B	A protein coding gene. An important paralog of this gene is SPRR1A
COL1A2	Among its related pathways are ERK signaling and IL4-mediated signaling events
MMP2	Among its related pathways are direct p53 effectors and development endothelin-1/EDNRA signaling

for processing the integrated data, so we will use the PL21GPCA method to process many other integrated data to verify its performance further.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: The lung cancer data <http://www.unc.edu/~haipeng>. The TCGA data <http://www.tcg.org/>.

AUTHOR CONTRIBUTIONS

X-ZK conceived and designed the experiments. YS and X-ZK performed the experiments and contributed to the writing of the manuscript. S-SY, JW, and L-YD analyzed the data.

REFERENCES

- Arajo, R. F., Lira, G. A., Vilaa, J. A., Guedes, H. G., Leito, M. C. A., Lucena, H. F., et al. (2015). Prognostic and diagnostic implications of MMP-2, MMP-9, and VEGF-a expressions in colorectal cancer. *Pathol. Res. Practice* 211, 71–77. doi: 10.1016/j.prp.2014.09.007
- Belkin, M., and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. *Adv. Neural Inf. Process. Syst.* 14, 585–591.
- Belkin, M., and Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* 15, 1373–1396. doi: 10.1162/089976603321780317
- Bera, T. K., Iavarone, C., Kumar, V., Lee, S., Lee, B., and Pastan, I. (2002). MRP9, an unusual truncated member of the ABC transporter superfamily, is highly expressed in breast cancer. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6997–7002. doi: 10.1073/pnas.102187299
- Bertsekas, D. P. (1982). *Constrained Optimization and Lagrange Multiplier Methods*. Cambridge, MA: Academic Press.
- Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc. Natl. Acad. Sci. U.S.A.* 98, 13790–13795. doi: 10.1073/pnas.191502998
- Bunte, K., Leppaahto, E., Saarinen, I., and Kaski, S. (2016). Sparse group factor analysis for biclustering of multiple data sources. *Bioinformatics* 32, 2457–2463. doi: 10.1093/bioinformatics/btw207
- Cai, D., He, X. F., and Han, J. W. (2005). Document clustering using locality preserving indexing. *IEEE Trans. Knowl. Data Eng.* 17, 1624–1637. doi: 10.1109/tkde.2005.198
- Cai, D., He, X. F., Han, J. W., and Huang, T. S. (2011). Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 1548–1560. doi: 10.1109/tpami.2010.231
- Candès, E. J., Li, X., Ma, Y., and Wright, J. (2009). Robust principal component analysis? *J. ACM* 58, 1–37.
- Chartrand, R. (2012). Nonconvex splitting for regularized low-rank + sparse decomposition. *IEEE Trans. Signal Process.* 60, 5810–5819. doi: 10.1109/tsp.2012.2208955
- Chen, X. J., Huang, J. Z., Wu, Q. Y., and Yang, M. (2019). Subspace weighting co-clustering of gene expression data. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 16, 352–364. doi: 10.1109/tcbb.2017.2705686
- Collins, M. (2002). “A generalization of principal component analysis to the exponential family,” in *Proceedings of the 14th International Conference on Advances in Neural Information Processing Systems*, Cambridge, MA.
- Delbert, D., Morris, Q. D., and Frey, B. J. (2005). Multi-way clustering of microarray data using probabilistic sparse matrix factorization. *Bioinformatics* 21(Suppl. 1), i144–i151.
- Ding, C., and He, X. F. (2004). “K-Means Clustering Via Principal Component Analysis,” in *Proceedings of the 21st International Conference on Machine Learning (ICML)*, New York, NY 1, 29.
- Feng, C., Gao, Y. L., Liu, J. X., Zheng, C. H., and Yu, J. (2017). PCA based on graph laplacian regularization and P-norm for gene selection and clustering. *IEEE Trans. Nanobiosci.* 16, 257–265. doi: 10.1109/tnb.2017.2690365
- Feng, C. M., Liu, J. X., Gao, Y. L., Wang, J., Wang, D. Q., and Du, Y. (2016). “A graph-laplacian pca based on L1/2-norm constraint for characteristic gene selection,” in *Proceedings of the 2016th IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM 2016)*, Shenzhen, 1258–1263.
- Feng, C. M., Xu, Y., Liu, J. X., Gao, Y. L., and Zheng, C. H. (2019). Supervised discriminative sparse PCA for corn-characteristic gene selection and tumor classification on multiview biological data. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 2926–2937. doi: 10.1109/tnnls.2019.2893190
- Gabay, D., and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.* 2, 17–40. doi: 10.1016/0898-1221(76)90003-1
- Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* 17, 333–351. doi: 10.1038/nrg.2016.49
- Guo, S., Wang, Z., and Ruan, Q. (2013). Enhancing sparsity via l_p ($0 < p < 1$) minimization for robust face recognition. *Neurocomputing* 99, 592–602. doi: 10.1016/j.neucom.2012.05.028
- He, X., Yan, S., Hu, Y., Niyogi, P., and Zhang, H. J. (2005). Face recognition using laplacian faces. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 328–340.
- Hestenes, M. R. (1969). Multiplier and gradient methods. *J. Optim. Theory Appl.* 4, 303–320. doi: 10.1007/bf00927673
- Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., et al. (2010). FABIA: factor analysis for bicluster acquisition. *Bioinformatics* 26, 1520–1527. doi: 10.1093/bioinformatics/btq227
- Hou, M. X., Gao, Y. L., Liu, J. X., Dai, L. Y., Kong, X. Z., and Shang, J. H. (2019). Network analysis based on low-rank method for mining information on integrated data of multi-cancers. *Comput. Biol. Chem.* 78, 468–473. doi: 10.1016/j.compbiolchem.2018.11.027
- Ishiwata, T., Cho, K., Kawahara, K., Yamamoto, T., Fujiwara, Y., Uchida, E., et al. (2007). Role of lumican in cancer cells and adjacent stromal tissues in human pancreatic cancer. *Oncol. Rep.* 18, 537–543.
- Jiang, B., Ding, C., Luo, B., and Tang, J. (2013). “Graph-laplacian PCA: closed-form solution and robustness,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Portland.
- Jolliffe, I. (2002). *Principal Component Analysis*. New York, NY: Springer.
- Journee, M., Nesterov, Y., Richtarik, P., and Sepulchre, R. (2010). Generalized Power method for sparse principal component analysis. *J. Mach. Learn. Res.* 11, 517–553.

J-XL and C-HZ contributed to reagents, materials, and analysis tools. All authors contributed to the article and approved the submitted version.

FUNDING

This work was funded by grants from the National Natural Science Foundation of China (No. 61702299) and jointly in part by the National Natural Science Foundation of China, Nos. 61872220, 61701279, and 61902215.

ACKNOWLEDGMENTS

Thanks a lot for my co-tutor Yong Xu who is now a professor in Harbin Institute of Technology, Shenzhen, China.

- Keyhanian, S., and NaserSharif, B. (2015). "Laplacian eigenmaps latent variable model modification for pattern recognition," in *Proceedings of the 23rd Iranian Conference on Electrical Engineering (ICEE)*, Tehran.
- Koboldt, D. C., Fulton, R. S., McLellan, M. D., Schmidt, H., Kalicki-Verizer, J., McMichael, J. F., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70. doi: 10.1038/nature11412
- Kong, X. Z., Liu, J. X., Zheng, C. H., Hou, M. X., and Wang, J. (2017). Robust and efficient biomolecular clustering of tumor based on ℓ_p -norm singular value decomposition. *IEEE Trans. Nanobiosci.* 16, 341–348. doi: 10.1109/tnb.2017.2705983
- Lee, M., Shen, H., Huang, J. Z., and Marron, J. S. (2010). Biclustering via sparse singular value decomposition. *Biometrics* 66, 1087–1095. doi: 10.1111/j.1541-0420.2010.01392.x
- Lew, Y., Jones, D. V., Mars, W. M., Evans, D., Byrd, D., and Frazier, M. L. (1992). Expression of elongation factor-1 gamma-related sequence in human pancreatic cancer. *Pancreas* 7, 144–152. doi: 10.1097/00006676-199203000-00003
- Lin, Z. C., Chen, M. M., and Ma, Y. (2010). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv [Preprint]* 9. doi: 10.1016/j.jsb.2012.10.010
- Liu, J., Liu, J. X., Gao, Y. L., Kong, X. Z., Wang, X. S., and Wang, D. (2015). A P-Norm robust feature extraction method for identifying differentially expressed genes. *PLoS One* 10:e0133124. doi: 10.1371/journal.pone.0133124
- Liu, J. X., Wang, D., Gao, Y. L., Zheng, C. H., Shang, J. L., Liu, F., et al. (2017). A joint-L2,1-norm-constraint-based semi-supervised feature extraction for RNA-Seq data analysis. *Neurocomputing* 228, 263–269. doi: 10.1016/j.neucom.2016.09.083
- Liu, J. X., Wang, Y. T., Zheng, C. H., Sha, W., Mi, J. X., and Xu, Y. (2013). Robust PCA based method for discovering differentially expressed genes. *BMC Bioinformatics* 14 Suppl. 8:S3.
- Liu, J. X., Xu, Y., Gao, Y. L., Zheng, C. H., Wang, D., and Zhu, Q. (2016). A class-information-based sparse component analysis method to identify differentially expressed genes on RNA-seq data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 13, 392–398. doi: 10.1109/tcbb.2015.2440265
- Liu, J. X., Xu, Y., Zheng, C. H., Kong, H., and Lai, Z. H. (2015). RPCA-based tumor classification using gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 12, 964–970. doi: 10.1109/tcbb.2014.2383375
- Liu, Q. Z., Gao, X. H., Chang, W. J., Wang, H. T., Wang, H., Cao, G. W., et al. (2015). Secreted protein acidic and rich in cysteine expression in human colorectal cancer predicts postoperative prognosis. *Eur. Rev. Med. Pharmacol. Sci.* 19, 1803–1811.
- Matassa, D. S., Amoroso, M. R., Agliarulo, I., Maddalena, F., Sisinni, L., Paladino, S., et al. (2013). Translational control in the stress adaptive response of cancer cells: a novel role for the heat shock protein TRAP1. *Cell Death Dis.* 4:e851. doi: 10.1038/cddis.2013.379
- Michifuri, Y., Hirohashi, Y., Torigoe, T., Miyazaki, A., Fujino, J., Tamura, Y., et al. (2013). Small proline-rich protein-1B is overexpressed in human oral squamous cell cancer stem-like cells and is related to their growth through activation of MAP kinase signal. *Biochem. Biophys. Res. Commun.* 439, 96–102. doi: 10.1016/j.bbrc.2013.08.021
- Misawa, K., Kanazawa, T., Misawa, Y., Imai, A., and Mineta, H. (2011). Hypermethylation of collagen $\alpha 2$ (I) gene (COL1A2) is an independent predictor of survival in head and neck cancer. *Cancer Biomark.* 10, 135–144. doi: 10.3233/cbm-2012-0242
- Nie, F. P., Wang, H., Huang, H., and Ding, C. (2013). Joint Schatten ℓ_p -norm and ℓ_q -norm robust matrix completion for missing value recovery. *Knowl. Inf. Syst.* 42, 525–544. doi: 10.1007/s10115-013-0713-z
- Pogue-Geile, K., Geiser, J. R., Shu, M., Miller, C., and Pipas, J. M. (1991). Ribosomal protein genes are overexpressed in colorectal cancer: Isolation of a cDNA clone encoding the human S3 ribosomal protein. *Mol. Cell. Biol.* 11, 3842–3849. doi: 10.1128/mcb.11.8.3842
- Qiao, J., Fang, C. Y., Chen, S. X., Wang, X. Q., and Liu, F. (2015). Stroma derived COL6A3 is a potential prognosis marker of colorectal carcinoma revealed by quantitative proteomics. *Oncotarget* 6, 29929–29946. doi: 10.18632/oncotarget.4966
- Qin, L., Lin, Z., She, Y., and Chao, Z. (2013). A comparison of typical ℓ_p minimization algorithms. *Neurocomputing* 119, 413–424. doi: 10.1016/j.neucom.2013.03.017
- Ren, F., Tang, R., Xin, Z., Mihiranganee, M. W., Luo, D., Dang, Y., et al. (2015). Overexpression of MMP family members functions as prognostic biomarker for breast cancer patients: a systematic review and meta-analysis. *Plos One* 10:e0135544. doi: 10.1371/journal.pone.0135544
- Roweis, S. T., and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 2323–2326. doi: 10.1126/science.290.5500.2323
- Shen, H. P., and Huang, J. H. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multiv. Anal.* 99, 1015–1034. doi: 10.1016/j.jmva.2007.06.007
- Spelman, D. A. (2007). "Spectral graph theory and its applications. foundations of computer science, 2007," in *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science FOCS '07*, Providence, RI.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc.* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x
- Wan, D., Gong, Y., Qin, W., Zhang, P., Li, J., Wei, L., et al. (2004). Large-scale cDNA transfection screening for genes related to cancer development and progression. *Proc. Natl. Acad. Sci. U.S.A.* 101, 15724–15729. doi: 10.1073/pnas.0404089101
- Wang, J., Liu, J. X., Kong, X. Z., Yuan, S. S., and Dai, L. Y. (2019a). Laplacian regularized low-rank representation for cancer samples clustering. *Comput. Biol. Chem.* 78, 504–509. doi: 10.1016/j.compbiolchem.2018.11.003
- Wang, J., Liu, J. X., Zheng, C. H., Wang, Y. X., Kong, X. Z., and Wen, C. G. (2019b). A mixed-norm laplacian regularized low-rank representation method for tumor samples clustering. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 172–182. doi: 10.1109/tcbb.2017.2769647
- Wang, Y. X., Gao, Y. L., Liu, J. X., Kong, X. Z., and Li, H. J. (2017). Robust principal component analysis regularized by truncated nuclear norm for identifying differentially expressed genes. *IEEE Trans. Nanobiosci.* 16, 447–454. doi: 10.1109/tnb.2017.2723439
- West, M. (2003). *Bayesian Factor Regression Models in the "Large p, Small n" Paradigm*, Vol. 7. Oxford: Oxford University Press, 723–732.
- Xiang, S. M., Nie, F. P., Meng, G. F., Pan, C. H., and Zhang, C. S. (2012). Discriminative least squares regression for multiclass classification and feature selection. *IEEE Trans. Neural Netw. Learn. Syst.* 23, 1738–1754. doi: 10.1109/tnnls.2012.2212721
- Yang, S. Z., Hou, C. P., Nie, F. P., and Wu, Y. (2012). Unsupervised maximum margin feature selection via $L_{2,1}$ -norm minimization. *Neural Comput. Appl.* 21, 1791–1799. doi: 10.1007/s00521-012-0827-3
- Yu, C., Pan, K., Xing, D., Liang, G., and Lin, D. (2002). Correlation between a single nucleotide polymorphism in the matrix metalloproteinase-2 promoter and risk of lung cancer. *Cancer Res.* 62, 6430–6433.
- Zhang, Z., Xu, Y., Yang, J., Li, X., and Zhang, D. (2015). A survey of sparse representation: algorithms and applications. *IEEE Access* 3, 490–530. doi: 10.1109/access.2015.2430359
- Zhang, Z., and Zha, H. (2002). Principal manifolds and nonlinear dimension reduction via local tangent space alignment. *J. Shang. Univ.* 8, 406–424. doi: 10.1007/s11741-004-0051-1
- Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Stat.* 15, 265–286.

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.

Copyright © 2021 Kong, Song, Liu, Zheng, Yuan, Wang and Dai. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



TMP-SSurface2: A Novel Deep Learning-Based Surface Accessibility Predictor for Transmembrane Protein Sequence

Zhe Liu^{1,2,3}, Yingli Gong⁴, Yuanzhao Guo², Xiao Zhang⁵, Chang Lu², Li Zhang^{1*} and Han Wang²

¹ School of Computer Science and Engineering, Changchun University of Technology, Changchun, China, ² School of Information Science and Technology, Institute of Computational Biology, Northeast Normal University, Changchun, China, ³ Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China, ⁴ College of Intelligence and Computing, Tianjin University, Tianjin, China, ⁵ College of Computing and Software Engineering, Kennesaw State University, Kennesaw, GA, United States

OPEN ACCESS

Edited by:

Wang Guohua,
Harbin Institute of Technology, China

Reviewed by:

Hongjie Wu,
Suzhou University of Science
and Technology, China
Xiujuan Lei,
Shaanxi Normal University, China

*Correspondence:

Li Zhang
lizhang@ccut.edu.cn

Specialty section:

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

Received: 20 January 2021

Accepted: 22 February 2021

Published: 15 March 2021

Citation:

Liu Z, Gong Y, Guo Y, Zhang X,
Lu C, Zhang L and Wang H (2021)
TMP-SSurface2: A Novel Deep
Learning-Based Surface Accessibility
Predictor for Transmembrane Protein
Sequence. *Front. Genet.* 12:656140.
doi: 10.3389/fgene.2021.656140

Transmembrane protein (TMP) is an important type of membrane protein that is involved in various biological membranes related biological processes. As major drug targets, TMPs' surfaces are highly concerned to form the structural biases of their material-bindings for drugs or other biological molecules. However, the quantity of determinate TMP structures is still far less than the requirements, while artificial intelligence technologies provide a promising approach to accurately identify the TMP surfaces, merely depending on their sequences without any feature-engineering. For this purpose, we present an updated TMP surface residue predictor TMP-SSurface2 which achieved an even higher prediction accuracy compared to our previous version. The method uses an attention-enhanced Bidirectional Long Short Term Memory (BiLSTM) network, benefiting from its efficient learning capability, some useful latent information is abstracted from protein sequences, thus improving the Pearson correlation coefficients (CC) value performance of the old version from 0.58 to 0.66 on an independent test dataset. The results demonstrate that TMP-SSurface2 is efficient in predicting the surface of transmembrane proteins, representing new progress in transmembrane protein structure modeling based on primary sequences. TMP-SSurface2 is freely accessible at <https://github.com/NENUBioCompute/TMP-SSurface-2.0>.

Keywords: transmembrane protein, deep learning, relative accessible surface area, attention mechanism, long short term memory

INTRODUCTION

Transmembrane Proteins (TMPs) are the gatekeepers to the cells and control the flow of molecules and information across the membrane (Goddard et al., 2015). The function of MPs is crucial for a wide range of physiological processes like signal transduction, electron transfer, and neurotransmitter transport (Roy, 2015). They span the entire biological membrane with segments exposed on both the outside and inside of aqueous spaces and have a profound effect on the pharmacokinetics of various drugs (Padmanabhan, 2014), cell mechanics regulation

(Stillwell, 2016), molecule transport (Oguro and Imaoka, 2019; Puder et al., 2019) and so on. Also, the evidence is pointing toward TMPs associating with a wide range of diseases, including dyslipidemia, autism, epilepsy (Rafi et al., 2019; Tanabe et al., 2019; Weihong et al., 2019), and multiple cancers (Moon et al., 2019; Yan et al., 2019). Moreover, based on the current therapeutics market, it is evaluated that more than one-third of future drug targets would be TMPs (Studer et al., 2014) and the surface of TMPs is always identified as an interaction interface according to statistical reports (Lu et al., 2019b).

The quantitative approach for measuring the exposure of residues is to calculate the relatively accessible surface area (rASA) of the residues (Tarafder et al., 2018). rASA reflects the exposure of a single residue to the solvent, making it a directive reference of protein structures. Predicting rASA of TMPs is a rewarding task to biological problems like function annotation, structural modeling, and drug discovery (Zhang et al., 2019). In this case, accurate sequence-based computational rASA predictors need to be developed urgently to provide more support for structure prediction.

Many rASA predictors had been reported performing well on soluble proteins but the structural differences between the two protein types are significant, especially when interacting with the phospholipid bilayer. There are a few methods released to predict rASA of TMP residues based on their primary sequences. Beuming and Weinstein (2004) firstly proposed a knowledge-based method to predict the binary state (buried or exposed) of residues in terms of a preassigned cutoff in the transmembrane region of α -TMPs, it is the first rASA predictor of TMPs. After that, a series of methods using machine learning including SVC, SVR, and SVM emerged, which can be automatically divided into two categories according to their functionality: binary classifier and rASA real value predictor. All of these machine learning-based methods were designed for α -TMPs, some methods were just effective with the transmembrane region of the proteins restrictedly, such as TMX (Liwicki et al., 2007; Wang et al., 2011), TMExpoSVC (Lai et al., 2013), and TMExpoSVR (Lai et al., 2013), only MPRAP (Illergård et al., 2010) and MemBrane-Rasa (Xiao and Shen, 2015; Yin et al., 2018) were able to predict rASA of the entire sequence. Our previous work (Lu et al., 2019a) combined Inception blocks with CapsNet, proving that deep learning takes many advantages for the prediction but there is still room for accuracy improvement.

The predictors mentioned above including our previous version all applied common methods like SVM and feed-forward neural networks. However, these non-sequential models do not naturally handle sequential data and have trouble capturing long-term dependencies of a certain sequence (Sønderby and Winther, 2014), thus being a bottleneck in rASA prediction tasks, calling for more suitable models. In recent years, various Long Short Term Memory (LSTM) models have already employed to learn temporal information of protein secondary structure, confirming the amazing ability of LSTM in handling protein sequences through experimental verification (Sønderby and Winther, 2014; Sønderby et al., 2015; Heffernan et al., 2017). When it comes to sequence level issues, LSTM is definitely a better choice. Furthermore, previous tools did not have measures

for reinforcing effective features, resulting in lower inefficiency of model learning. Additionally, various input restrictions and long waiting times also made the predictors less friendly to users.

In this study, we proposed an attention-enhanced bidirectional LSTM network named TMP-SSurface2 to predict rASA of TMPs at the residue level, which was implemented on top of the CNN-based Z-coordinate predictor TM-ZC (Lu et al., 2020). TMP-SSurface2 was trained and tested against the non-redundant benchmark dataset we created with primary sequences as input, improving the Pearson correlation coefficients (CC) value performance of the old version from 0.584 to 0.659, and reduced the mean absolute error (MAE) from 0.144 to 0.140. Apart from state-of-the-art prediction accuracy, TMP-SSurface2 also achieved the highest output efficiency compared to existing methods with no length restriction of input. The source codes of TMP-SSurface2 and the corresponding materials can be freely accessed at <https://github.com/NENUBioCompute/TMP-SSurface-2.0>.

MATERIALS AND METHODS

Benchmark Dataset

A total of 4,007 TMPs were downloaded from PDBTM (version: 2019-01-04). We removed the proteins which contained unknown residues such as "X" or whose length was less than 30 residues since too short a sequence may not form a representative structure. To avoid the redundancy of data and reduce the influence of homology bias, CD-HIT (Li and Godzik, 2006) was utilized to eliminate the duplicate structures with a 30% sequence identity cut-off resulting in 704 protein chains (618 α protein chains and 86 β protein chains) left. These proteins were randomly divided into a training set of 604 proteins, a validation set of 50 proteins, and a test set of 50 proteins, respectively. In this work, five-fold cross-validation experiments were performed and the results were compared against other predictors.

The residue solvent accessibility surface area (ASA) is defined as the surface accessibility of a certain residue when exposed to water or lipid. Several tools are capable of calculating ASA, such as Naccess (Lee and Richards, 1971), PSAIA (Mihel et al., 2008), MSMS (Sanner et al., 1996), and Dictionary of Protein Secondary Structure (DSSP) (Kabsch and Sander, 1983).

The ASA of residues was calculating by DSSP, using a probe with a radius of 1.4 Å. A residue's ASA is divided by the corresponding standard maximum accessible surface area (MaxASA), which is the ASA of extended tri-peptides (Gly-X-Gly) (Tien et al., 2013), to generate rASA values. rASA can be calculated by the following formula:

$$rASA = \frac{ASA}{MaxASA} \quad (1)$$

Features and Encoding

To make the prediction more accurate, it is vital to provide useful features to deep learning-based methods. In our experiments, we carefully select two encoding features to represent the protein fragment: one-hot code and PSSM.

Prediction of transmembrane protein residues' rASA is a classical regression problem, which can be formulated as follows: for a given primary sequence of a TMP, a sliding window of k residues was used to predict the real value of central residue's rASA. For instance, if k is 19, then each protein is subsequently sliced into fragments of 19 amino acids.

For each residue in protein sequences, one-hot code is a 20-dimension vector (see **Figure 1**), using a 19 dimensional "0" vector with a "1" corresponding to the amino acid at the index of a certain protein sequence. In this way, each protein fragment can be mapped into an exclusive and undisturbed coding within its relative position information (He et al., 2018). It is proved that a one-hot code is extremely easy to generate while effective for protein function prediction associated problems (Ding and Li, 2015).

A position-specific scoring matrix (PSSM) reflects the evolutionary profile of the protein sequence based on a search against a certain database. Highly conserved regions during evolution are always functional regions according to the researches (Jeong et al., 2010; Zeng et al., 2019), so PSSM has been widely used in many bioinformatics problems and achieves commendable results. In our study, PSI-BLAST (Altschul et al., 1997) was utilized to generate PSSM searching against the uniref50 (version: 2019-01-16) database with 3 iterations and a 0.01 E -value cutoff. For a given protein sequence, the PSSM feature is a 20-dimension matrix with each column representing a profile and each row representing a residue.

As shown in **Figure 2**, each amino acid in the protein sequence is represented as a vector of 41 numbers, including 20 from one-hot code (represented as binary numbers), 20 from PSSM, and 1 Noseq label (representing a gap) (Fang et al., 2018) in the last column to improve the prediction performance of the residues located on both ends of protein while using a sliding window. In order to facilitate the window sliding operation, the first and last parts of the sequence are, respectively, padded with 1 and 0 s, which length is half of the sliding windows size. For each protein with L residues, we can get L matrices.

Model Design

In this section, a novel compound deep learning network is presented. **Figure 3A** shows the proposed pipeline. The input features for TMP-SSurface2 are the one-hot code and the PSSM matrix. The CNN whose structure and parameters are all same as TM-ZC is used to generate the Z-coordinate of TMP residues. Z-coordinate, which is an important constituent in the field of MP structure prediction, is often implemented to stand for a residue's relative position concerning the membrane (Yin et al., 2018). After that, the final feature map containing a one-hot code, PSSM, and Z-coordinate will be put into a bidirectional LSTM (BiLSTM) network for training and testing.

To further optimize the model, we also attached an attention mechanism (Baron-Cohen, 1995) layer to the top of BiLSTM, which is motivated by how we pay visual attention to different regions of an image or correlate words in one sentence, to help LSTM focus on a certain region that relatively deserves more attention. The detailed structure of the mentioned LSTM network is shown in **Figure 3B**.

Formula (2) to formula (9) describe the forward recursions for a single LSTM layer, where \odot equals to the elementwise multiplication, x_t means input from the previous layer, i_t , f_t , o_t represent "input gate," "forget gate" and "output gate," respectively. h_{t-rec} stands for the output forwarded to the next time slice, and h_t is passed upwards in a multilayer LSTM (Sønderby and Winther, 2014). Attention neural networks have recently demonstrated popularity in a wide range of tasks ranging from natural language processing to computer vision (Chorowski et al., 2014; Rocktäschel et al., 2015; Sharma et al., 2015). Inspired by these projects, we attached an attention mechanism to LSTM for feature capturing. As shown in formula (10), the combination of attention mechanism enables the model to re-assign the weight (W_{att}) of the feature vector (V), indicating that the next output vector (V') should focus more on which part of the input sequence, and then generate the next output according to the focus region.

$$i_t = \sigma(x_t W_{xi} + h_{t-1} W_{hi} + b_i) \quad (2)$$

$$f_t = \sigma(x_t W_{xf} + h_{t-1} W_{hf} + b_f) \quad (3)$$

$$o_t = \sigma(x_t W_{xo} + h_{t-1} W_{ho} + b_o) \quad (4)$$

$$g_t = \tanh(x_t W_{xg} + h_{t-1} W_{hg} + b_g) \quad (5)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (6)$$

$$h_t = o_t \odot \tanh(c_t) \quad (7)$$

$$h_{t-rec} = h_t + \text{feedforwardnet}(h_t) \quad (8)$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)} \quad (9)$$

$$V' = W_{att} \odot V \quad (10)$$

Our model was implemented, trained, and tested using Keras and Tensorflow. Main hyperparameters (sliding window size, training dropout rate, number of LSTM units, and layers of LSTM) were explored. The early stopping and save-best strategy were applied when the validation loss did not reduce in 10 epochs during training time, the process would stop and save the best model parameters. We used Adam optimizer to dynamically transform the learning rate while the model was training. All the experiments were performed using an Nvidia 1080Ti GPU.

Performance Evaluation

To quantitatively evaluate the predictions of TMP-SSurface2, Pearson correlation coefficients (CC) and mean absolute error (MAE) were used in this study. CC undertook the task of measuring the linear correlation between real values and

Sequence Length

A	R	D	C	Q	E	H	I	G	N	L	K	M	F	P	S	T	W	Y	V
0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
⋮																			
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIGURE 1 | One-hot code of protein residues.

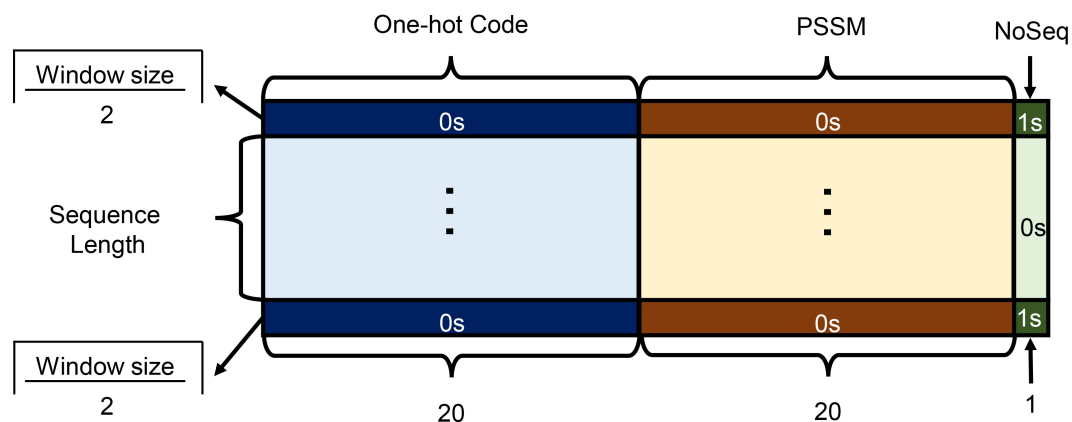


FIGURE 2 | Encoding features as the model input.

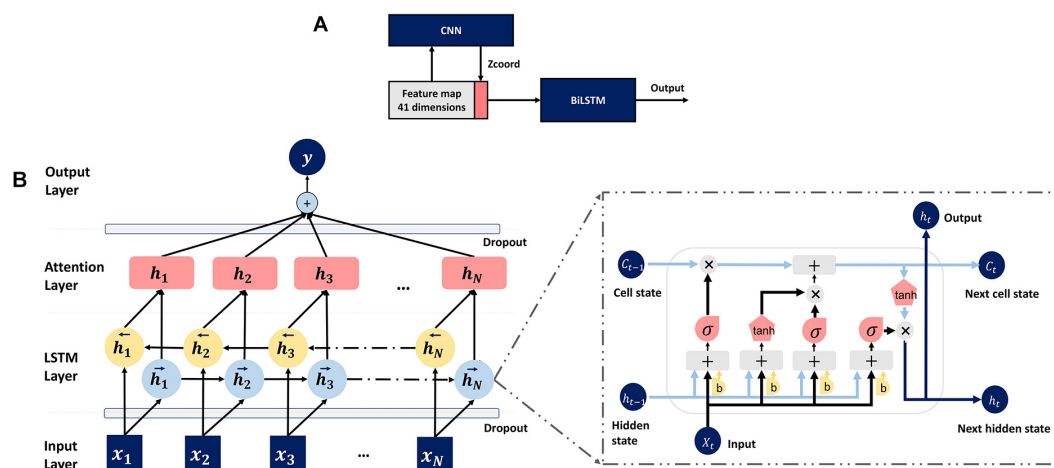


FIGURE 3 | (A) Pipeline of the deep learning model. (B) The attention-enhanced bidirectional LSTM network.

predicting values. CC ranges from -1 to 1 , where -1 indicates an abstract negative correlation, 1 positive correlation, and 0 absolutely no correlation. Formula (11) shows the definition of CC, where L represents the number of residues, x_i and y_i define the observed and predicted rASA value severally, \bar{x} and \bar{y} equal to the corresponding mean value, respectively.

$$CC = \frac{\sum_{i=1}^L (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^L (x_i - \bar{x})^2\right] \left[\sum_{i=1}^L (y_i - \bar{y})^2\right]}} \quad (11)$$

Mean absolute error measures the closeness of prediction values to real values. As shown in formula (12), MAE is defined as the average difference between predicted and observed rASA values of all residues.

$$MAE = \frac{1}{L} \sum_{i=1}^L |y_i - x_i| \quad (12)$$

RESULTS

Feature Analysis

As we all know, it is the features, instead of model structures, determine the upper-performance limit of deep learning. To investigate the different features' contribution to the predictor TMP-SSurface2, we tested both independent features used in the predictor and their various combinations on our valid dataset.

Table 1 illustrates that all of the three independent features (Z-coordinate, one-hot, and PSSM) contain useful information for predicting rASA by themselves, among which PSSM achieves the best overall results (CC = 0.631 and MAE = 0.144). It is suggested that PSSM is an important feature in rASA prediction mainly because of the inclusion of evolutionary knowledge. When combining these different features, as was indicated by a former study, the CC values are almost linearly related to the MAE values (Yuan et al., 2006), the maximum CC values always accompany the minimum MAE. Experimental investigation shows that every single feature made a contribution to the prediction and achieved the most considerable performance (CC = 0.659 and MAE = 0.140) when they were combined.

Hyperparameter Tuning and Model Performance

Tables 2–5 summarizes the exploration of the attention-enhanced bidirectional LSTM network with various

hyperparameters on the validation dataset. The object of doing these experiments was to find out a better configuration of our method. The tested hyperparameters were carefully selected and only the major factors which would greatly influence the model were explored on the validation dataset.

A sliding window approach is utilized to append useful neighborhood information to improve prediction accuracy. **Table 2** shows how the length of the sliding window affects the performance of our network. Since the contexts fed into the proposed deep learning model relies on the length of the sliding window, the prediction accuracy would be directly influenced by its value. In general, when the window size becoming larger, it will cost more time for training, but the prediction performance may not be better as the window length increases. Historically, if a

TABLE 2 | Effect of sliding window length on CC performance.

Window Length	CC	MAE
13	0.642	0.141
15	0.641	0.143
17	0.645	0.143
19	0.648	0.140
21	0.646	0.141
23	0.640	0.142

**Bold fonts represent the best experimental results.*

TABLE 3 | Effect of dropout rate on CC performance.

Dropout rate	Train CC	Test CC	Test MAE
No	0.851	0.632	0.143
0.2	0.806	0.640	0.143
0.3	0.782	0.648	0.140
0.4	0.762	0.641	0.141
0.5	0.725	0.638	0.143

**Bold fonts represent the best experimental results.*

TABLE 4 | Effect of LSTM units' number on CC performance.

Num of units	CC	MAE	Num of Parameters
500	0.639	0.142	2,191,381
600	0.641	0.142	3,109,591
700	0.648	0.140	4,187,781
800	0.643	0.143	5,425,981
900	0.646	0.140	6,824,181

**Bold fonts represent the best experimental results.*

TABLE 5 | Effect of the number of LSTM layers on CC performance.

LSTM Layers	CC	MAE	Num of parameters
1	0.648	0.140	4,187,781
2	0.659	0.140	15,953,381
3	0.642	0.141	27,718,981
4	0.646	0.141	39,484,581

**Bold fonts represent the best experimental results.*

TABLE 1 | Prediction performance based on individual input features and their various combinations.

Feature	CC	MAE
Z-coordinate	0.310	0.191
one-hot	0.417	0.180
PSSM	0.631	0.144
one-hot+PSSM	0.641	0.142
one-hot+PSSM+ Z-coordinate	0.659	0.140

**Bold fonts represent the best experimental results.*

sliding window was utilized by sequence-based protein structure predicting tasks, the peak of performance often occurred when its length was between about 13 and 23 residues (Fang et al., 2018; Lu et al., 2019a). We searched the window length from 13 to 23 by a step of two residues, finding the best result when the number is 19 and it was chosen as the final window length in this section.

Table 3 shows how the dropout rate affects the model performance when the window size is 19. Deep learning neural networks are much easier to overfit a training dataset with few examples, dropout regularization will help reducing overfitting and improve the generalization of deep neural

TABLE 6 | Comparison of TMP-SSurface2 with the previous predictors on the independent dataset.

Predictor	CC	MAE	Failure	Time Cost (min)
MPRAP	0.397	0.176	9	6.5
MemBrane-Rasa	0.545	0.153	7	23.7
TMP-SSurface	0.584	0.144	0	4.7
TMP-SSurface2	0.659	0.140	0	4.3

**Bold fonts represent the best experimental results.*

TABLE 7 | Performance of TMP-SSurface2 on different types of TMPs.

TMP Types	Protein number	CC	MAE
α -helical TMPs	45	0.674	0.138
β -barrel TMPs	5	0.562	0.151
all-TMPs	50	0.659	0.140

networks (Dahl et al., 2013). The dropout rates in the range of 0.2–0.4 are all acceptable according to the training and testing prediction performance. Finally, we chose 0.3 as our dropout rate, and the concatenation network in our study is regularized using a 30% dropout.

In the LSTM network, the number of LSTM units is also an important parameter, which determines the output dimension of different layers just like ordinary neural networks. When the number of LSTM units in one layer changes, the scale of parameters and prediction accuracy of the model will immediately be affected. To find the best choice of LSTM units, we tried different values at the same time. The results are shown in **Table 4**, we chose 700 as the number of LSTM units in a simple layer.

As it can be seen in **Table 5**, when the LSTM network has two bidirectional layers (i.e., four simple layers, two forward and two backward), the model performs best on the validation set. However, the prediction accuracy of the model may not grow as the number of LSTM layers increases. It is suspected that a large number of model parameters will lead to the

TABLE 8 | Contribution of attention mechanism.

Model	CC	MAE
No attention	0.637	0.150
Attention with LSTM	0.659	0.140
Attention with Dropout	0.645	0.141

**Bold fonts represent the best experimental results.*

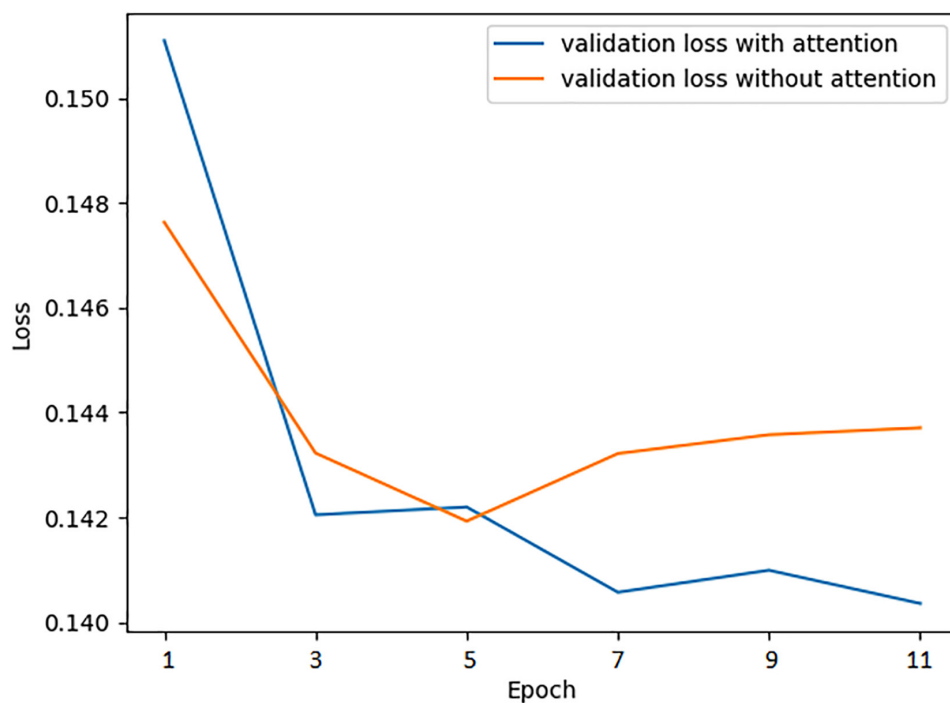
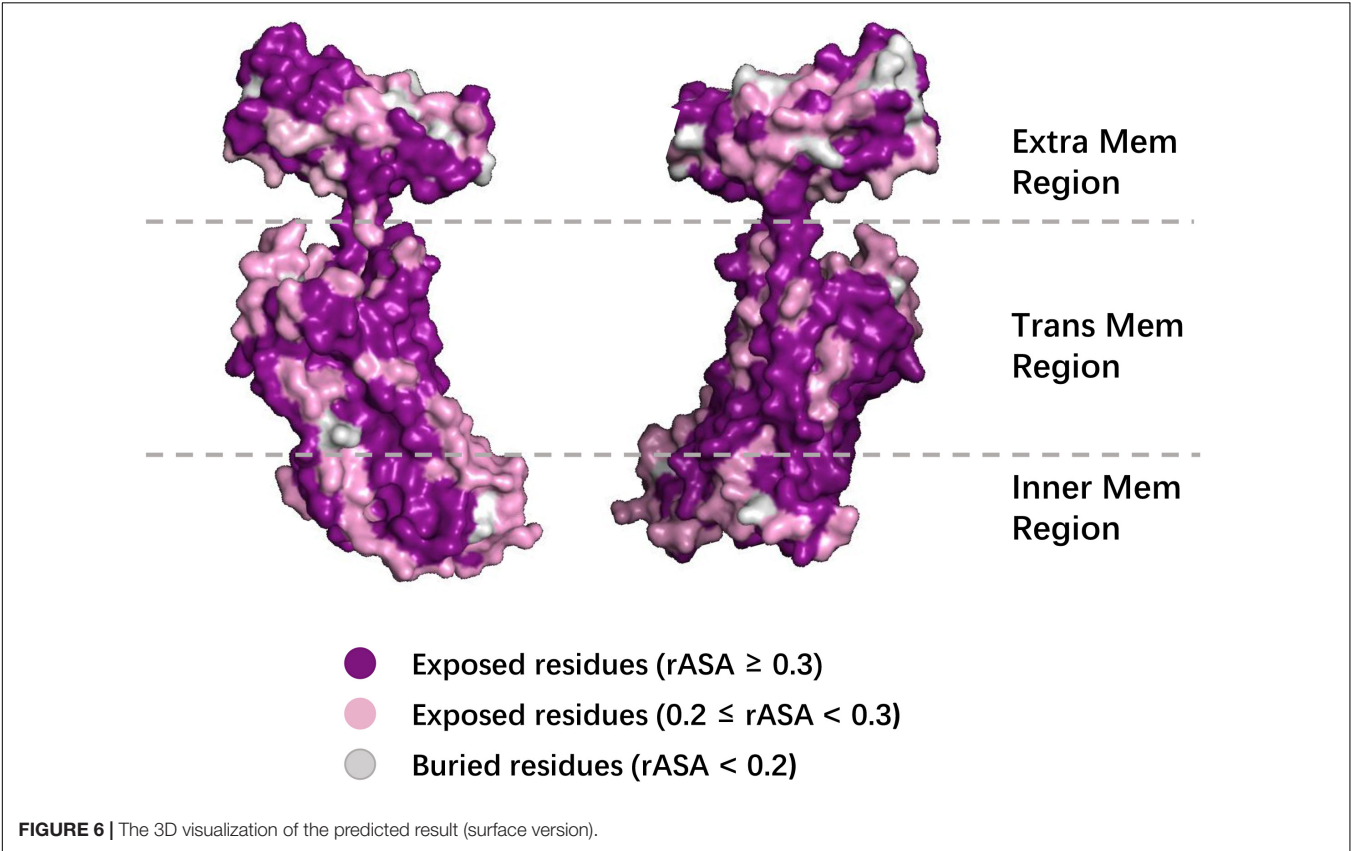
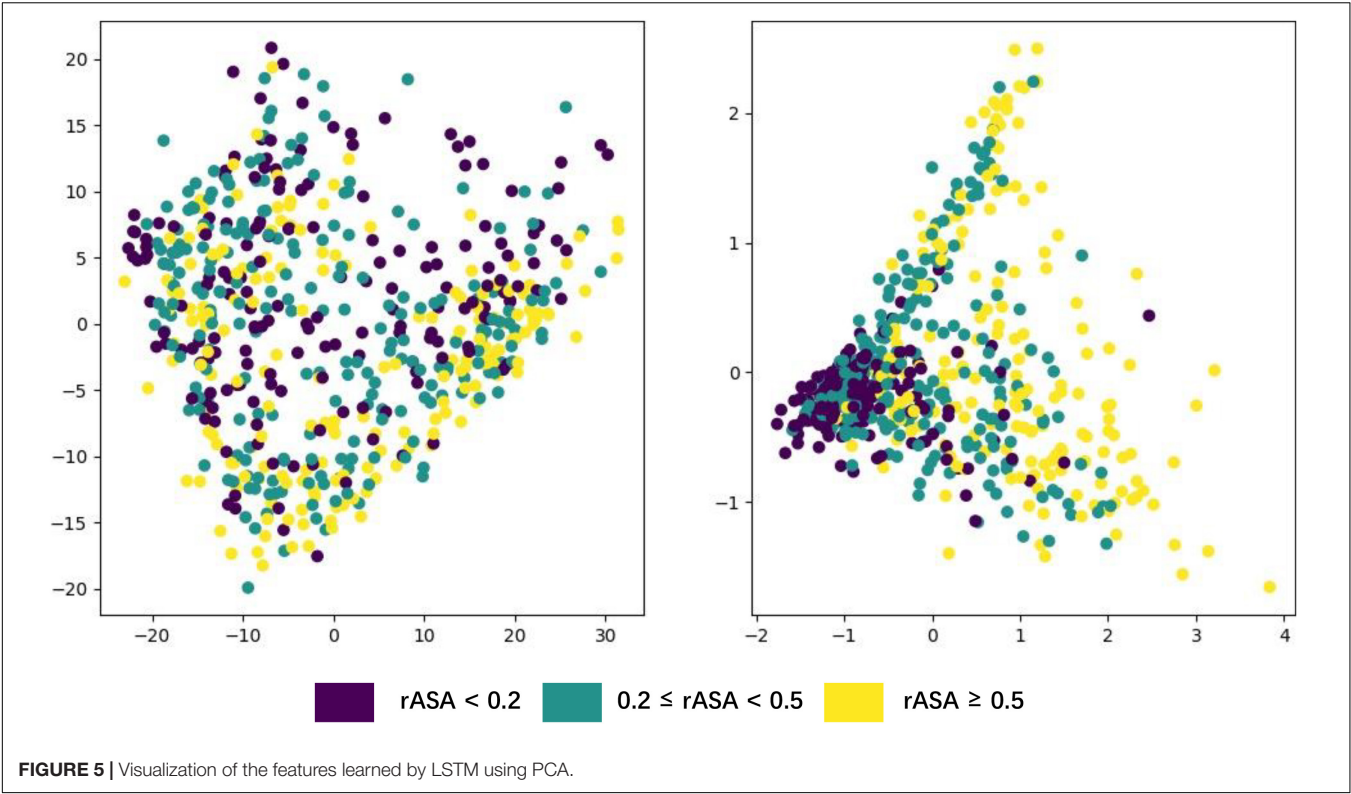


FIGURE 4 | Validation loss curve of the training process with and without attention mechanism.



overfitting of LSTM on the training set, thus reducing the generalization ability of it.

Comparison With Previous Predictors

In this section, we list the existing methods that can be used to predict the rASA of TMP in the full chain and compare TMP-SSurface2 with them. **Table 6** shows the performance improvement of the proposed TMP-SSurface2 after implementing the new model relative to the old version and the other tools. During testing MPRAP and MemBrane-Rasa on the independent dataset, we figured out that not every sequence fed into these predictors can get a corresponding output since some third-party tools might cause the failure. Just like TMP-SSurface, the new version is reliable in getting prediction results because of the simple coding scheme. Furthermore, TMP-SSurface2 significantly outperformed the previous predictors and has the quickest predicting speed. The details of the comparison are shown in **Table 6**.

TMP Type Test

Statistical results show that most of the existing methods only focused on α -helical TMPs while ignored β -barrel TMPs, which made it inconvenient for the users who cannot distinguish the protein type. As described previously, the data set we used contains both α -helical and β -barrel TMPs, making our predictor more suitable for all types of TMP. **Table 7** illustrates that when TMP-SSurface2 meets either of these two different TMPs, the prediction performance on the independent testing dataset was both considerable and reliable.

Contribution of Attention Mechanism

The attention mechanism promotes the model to extract features more effectively, speeding up the prediction accuracy to the peak, even improving the performance at the same time. To verify the positive effect of the attention mechanism, we monitoring the mean absolute error loss curve of the validation dataset with or without the attention layer, respectively, using the preselected best hyperparameters while training. As is shown in **Figure 4**, when the network is attention-enhanced,

the convergence speed and accuracy of the training set were significantly improved.

Moreover, we also combined attention mechanisms with various network layers to verify whether or how much the attention mechanism would improve the prediction performance. Firstly, we removed the attention layer and tested the trained model on the test set. Meanwhile, we attached the attention mechanism to the bidirectional LSTM layer and the Dropout layer, respectively, to conduct experiments, the results are shown in **Table 8**. It can be seen that the combination of attention mechanism and bidirectional LSTM layer reached the best performance, which is related to the fact that the LSTM layer had learned the most abundant features. In essence, the attention mechanism is to enhance the feature extraction process, so it will achieve the best effect when combined with the network layer that is the most effective for feature extraction.

Visualization of the Features Learnt by LSTM

Deep neural networks can learn high-level abstract features from original inputs, to verify whether the extracted features are generalizable, we utilized PCA (Wold, 1987) to visualize the input features and each LSTM unit's output in one bidirectional layer with test data. **Figure 5** shows the PCA scatter diagram of the test data before and after fed into LSTM, respectively. The input data had 42 features (i.e., 42 dimensions), PCA reduced its dimensionality and visualized it, but there was no clear cluster. The bidirectional LSTM layer we used contained 1,400 dimensions (twice of units in a simple LSTM layer) and the trend toward clustering had occurred, which demonstrates that LSTM had effectively captured useful and powerful features needed in this work.

Generally, buried residues are under stronger evolutionary constraints than exposed ones irrespectively of the environment (Kauko et al., 2008). The diagram shows that the residues whose rASA was lower than 0.2 narrowed down to a small area through PCA, which means these residues' rASA values stayed closely

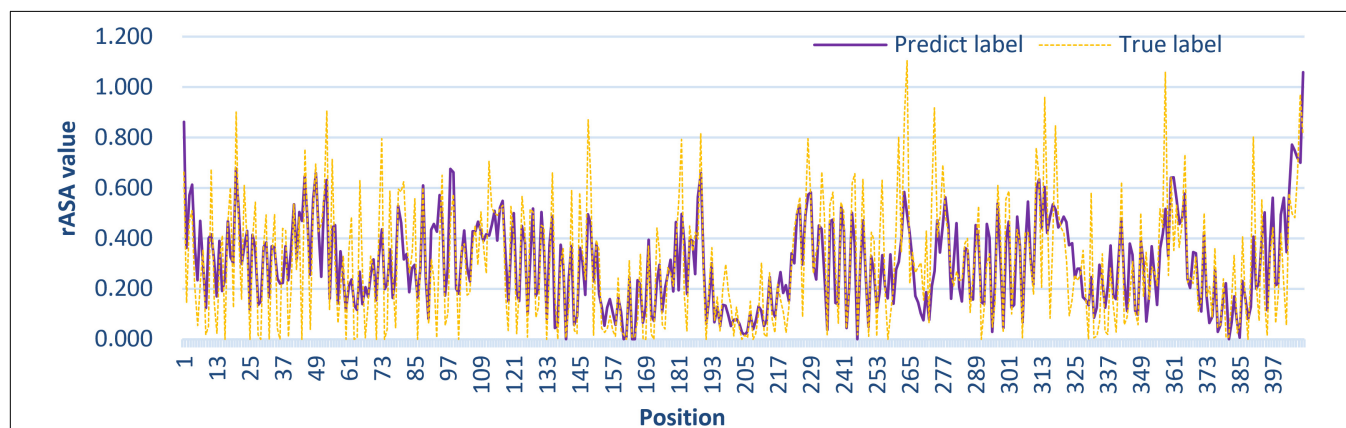


FIGURE 7 | The comparison between the TMP-SSurface2-predicted rASA values and real rASA values.

aligned with the features derived from their sequence, just proved the previous statement.

Case Studies

To further demonstrate the effectiveness of TMP-SSurface2, we take 4n6h_A as an example of case studies. 4n6h_A is an *Escherichia coli* α -TMP (subgroup: G protein-coupled receptor) containing 408 residues as the receptor of multiple ligands like sodium ion, heme, and so on (Fenalti et al., 2014). **Figure 6** shows the 3D visualization of the predicted result (surface version) and **Figure 7** illustrates the comparison between the TMP-SSurface2-predicted rASA values and real rASA values. As were shown in figures, the overall trend of rASA has been appropriately captured, but TMP-SSurface2 seems conservative in predicting some fully exposed or buried residues' rASA. It is suspected that TMP-SSurface2 may confuse these residues with the ones located on water-soluble regions, resulting in low prediction performance of them.

CONCLUSION

In this study, we proposed an updated TMP-SSurface predictor, which aimed to predict transmembrane protein residues' rASA from primary sequences. Apart from classical feed-forward neural networks, we developed an attention-enhanced bidirectional LSTM network on top of the CNN-based Z-coordinate predictor to process sequential data and improved the CC value performance of the old version from 0.58 to 0.66 on the independent test dataset. The improvement of LSTM directly indicates that the order of residues in a sequence would exactly influence the protein structure and LSTM has a more powerful ability to process sequential data than CapsNet. The Z-coordinate feature was explored and applied in TMP-SSurface2 and proved to be useful, which means the z-coordinate has a lifting effect on rASA prediction, indicating that structural features can support each other. We also appended various

important experiments like feature visualization and case study to visualize the effectiveness of the model. TMP-SSurface2 had no constraints with input since it could handle all types of TMPs at any length. The predicted rASA would make contributions to TMPs' structure analysis, TMP-ligand binding prediction, TMP function identification and so on.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZL, YGo, and XZ conceived the idea of this research, collected the data, implemented the predictor, and wrote the manuscript. YGu and CL tuned the model and tested the predictor. LZ and HW supervised the research and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Science and Technology Research Project of the Education Department of Jilin Province (No. JJKH20191309KJ), Jilin Scientific and Technological Development Program (No. 20180414006GH), and Fundamental Research Funds for the Central Universities (Nos. 2412019FZ052 and 2412019FZ048).

ACKNOWLEDGMENTS

This article is recommended by the 5th CCF Bioinformatics Conference.

REFERENCES

- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25, 3389–3402. doi: 10.1093/nar/25.17.3389
- Baron-Cohen, S. (1995). "The eye direction detector (EDD) and the shared attention mechanism (SAM): Two cases for evolutionary psychology," in *proceeding at the Portions of this paper were presented at the Society for Research in Child Development Conference, New Orleans, Mar 1993; the British Psychological Society, Welsh Branch, "Faces" Conference, U Wales Coll of Cardiff, Sep 1993; and the British Society for the Philosophy of Science "Roots of Joint Reference" Conference, U Bristol, Nov 1993*, (Mahwah: Lawrence Erlbaum Associates, Inc).
- Beuming, T., and Weinstein, H. (2004). A knowledge-based scale for the analysis and prediction of buried and exposed faces of transmembrane domain proteins. *Bioinformatics* 20, 1822–1835. doi: 10.1093/bioinformatics/bth143
- Chorowski, J., Bahdanau, D., Cho, K., and Bengio, Y. (2014). End-to-end continuous speech recognition using attention-based recurrent nn: first results. *arXiv [Preprint]* arXiv: 1412.1602,
- Dahl, G. E., Sainath, T. N., and Hinton, G. E. (2013). "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *Proceeding of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, (Vancouver, BC: IEEE), 8609–8613.
- Ding, H., and Li, D. (2015). Identification of mitochondrial proteins of malaria parasite using analysis of variance. *Amino acids* 47, 329–333. doi: 10.1007/s00726-014-1862-4
- Fang, C., Shang, Y., and Xu, D. (2018). Improving protein gamma-turn prediction using inception capsule networks. *Sci. Rep.* 8:15741.
- Fenalti, G., Giguere, P. M., Katritch, V., Huang, X.-P., Thompson, A. A., Cherezov, V., et al. (2014). Molecular control of δ -opioid receptor signalling. *Nature* 506, 191–196.
- Goddard, A. D., Dijkman, P. M., Adamson, R. J., dos Reis, R. I., and Watts, A. (2015). Reconstitution of membrane proteins: a GPCR as an example. *Methods Enzymol.* 556, 405–424.
- He, F., Wang, R., Li, J., Bao, L., Xu, D., and Zhao, X. (2018). Large-scale prediction of protein ubiquitination sites using a multimodal deep architecture. *BMC Syst. Biol.* 12:109. doi: 10.1186/s12918-018-0628-0
- Heffernan, R., Yang, Y., Paliwal, K., and Zhou, Y. (2017). Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone

- angles, contact numbers and solvent accessibility. *Bioinformatics* 33, 2842–2849. doi: 10.1093/bioinformatics/btx218
- Illergård, K., Callegari, S., and Elofsson, A. (2010). MPRAP: an accessibility predictor for α -helical transmembrane proteins that performs well inside and outside the membrane. *BMC Bioinformatics* 11:333. doi: 10.1186/1471-2105-11-333
- Jeong, J. C., Lin, X., and Chen, X.-W. (2010). On position-specific scoring matrix for protein function prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 8, 308–315. doi: 10.1109/tcbb.2010.93
- Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22, 2577–2637. doi: 10.1002/bip.360221211
- Kauko, A., Illergård, K., and Elofsson, A. (2008). Coils in the membrane core are conserved and functionally important. *J. Mol. Biol.* 380, 170–180. doi: 10.1016/j.jmb.2008.04.052
- Lai, J.-S., Cheng, C.-W., Lo, A., Sung, T.-Y., and Hsu, W.-L. (2013). Lipid exposure prediction enhances the inference of rotational angles of transmembrane helices. *BMC Bioinformatics* 14:304. doi: 10.1186/1471-2105-14-304
- Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* 55, 379–400. doi: 10.1016/0022-2836(71)90324-x
- Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158
- Liwicki, M., Graves, A., Fernández, S., Bunke, H., and Schmidhuber, J. (2007). “A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks,” in *Proceedings of the 9th International Conference on Document Analysis and Recognition, ICDAR 2007, Parana*. *
- Lu, C., Gong, Y., Liu, Z., Guo, Y., Ma, Z., and Wang, H. (2020). TM-ZC: a deep learning-based predictor for the Z-coordinate of residues in α -helical transmembrane proteins. *IEEE Access* 8, 40129–40137. doi: 10.1109/ACCESS.2020.2976797
- Lu, C., Liu, Z., Kan, B., Gong, Y., Ma, Z., and Wang, H. (2019a). TMP-SSurface: a deep learning-based predictor for surface accessibility of transmembrane protein residues. *Crystals* 9:640. doi: 10.3390/cryst9120640
- Lu, C., Liu, Z., Zhang, E., He, F., Ma, Z., and Wang, H. (2019b). MPLs-pred: predicting membrane protein-ligand binding sites using hybrid sequence-based features and ligand-specific models. *Int. J. Mol. Sci.* 20:3120. doi: 10.3390/ijms20133120
- Mihel, J., Šikić, M., Tomić, S., Jeren, B., and Vlahović, K. (2008). PSAIA—protein structure and interaction analyzer. *BMC Struct. Biol.* 8:21. doi: 10.1186/1472-6807-8-21
- Moon, Y. H., Lim, W., and Jeong, B. C. (2019). Transmembrane protein 64 modulates prostate tumor progression by regulating Wnt3a secretion. *Oncol. Lett.* 18, 283–290.
- Oguro, A., and Imaoka, S. (2019). Thioredoxin-related transmembrane protein 2 (TMX2) regulates the ran protein gradient and importin- β -dependent nuclear cargo transport. *Sci. Rep.* 9:15296.
- Padmanabhan, S. (2014). *Handbook of Pharmacogenomics and Stratified Medicine*. London: Academic Press.
- Puder, S., Fischer, T., and Mierke, C. T. (2019). The transmembrane protein fibrocytin/polyductin regulates cell mechanics and cell motility. *Phys. Biol.* 16:066006. doi: 10.1088/1478-3975/ab39fa
- Rafi, S. K., Fernández-Jaén, A., Álvarez, S., Nadeau, O. W., and Butler, M. G. (2019). High functioning autism with missense mutations in synaptotagmin-like protein 4 (sytl4) and transmembrane protein 187 (tmem187) genes: sytl4-protein modeling, protein-protein interaction, expression profiling and microRNA studies. *Int. J. Mol. Sci.* 20:3358. doi: 10.3390/ijms20133358
- Rocktäschel, T., Grefenstette, E., Hermann, K. M., Koëisk, T., and Blunsom, P. (2015). Reasoning about entailment with neural attention. *arXiv [preprint]* arXiv: 1509.06664.
- Roy, A. (2015). Membrane preparation and solubilization. *Methods Enzymol.* 557, 45–56. doi: 10.1016/bs.mie.2014.11.044
- Sanner, M. F., Olson, A. J., and Spehner, J. C. (1996). Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* 38, 305–320. doi: 10.1002/(sici)1097-0282(199603)38:3<305::aid-bip4>3.0.co;2-y
- Sharma, S., Kiro, R., and Salakhutdinov, R. (2015). Action recognition using visual attention. *arXiv [preprint]* arXiv:1511.04119.
- Sønderby, S. K., Sønderby, C. K., Nielsen, H., and Winther, O. (2015). “Convolutional LSTM networks for subcellular localization of proteins,” in *Proceeding of the International Conference on Algorithms for Computational Biology*, (Springer), 68–80. doi: 10.1007/978-3-319-21233-3_6
- Sønderby, S. K., and Winther, O. (2014). Protein secondary structure prediction with long short term memory networks. *arXiv [preprint]* arXiv:1412.7828.
- Stillwell, W. (2016). *An Introduction to Biological Membranes: Composition, Structure and Function*. Elsevier. Available online at: [https://books.google.com/books?hl=en&lr=&id=Q_WpCwAAQBAJ&oi=fnd&pg=PP1&dq=Stillwell,+W.+\(2016\).+An+Introduction+to+Biological+Membranes:&ots=NCr6lWYhDS&sig=VHh16aKevDFW2U96K7XEPIWG_u4#v=onepage&q=Stillwell%2C%20W.%20\(2016\).%20An%20Introduction%20to%20Biological%20Membranes%3A&f=false](https://books.google.com/books?hl=en&lr=&id=Q_WpCwAAQBAJ&oi=fnd&pg=PP1&dq=Stillwell,+W.+(2016).+An+Introduction+to+Biological+Membranes:&ots=NCr6lWYhDS&sig=VHh16aKevDFW2U96K7XEPIWG_u4#v=onepage&q=Stillwell%2C%20W.%20(2016).%20An%20Introduction%20to%20Biological%20Membranes%3A&f=false)
- Studer, G., Biasini, M., and Schwede, T. (2014). Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). *Bioinformatics* 30, i505–i511.
- Tanabe, Y., Taira, T., Shimotake, A., Inoue, T., Awaya, T., Kato, T., et al. (2019). An adult female with proline-rich transmembrane protein 2 related paroxysmal disorders manifesting paroxysmal kinesigenic choreoathetosis and epileptic seizures. *Rinsho shinkeigaku* 59, 144–148. doi: 10.5692/clinicalneuro.1001228
- Tarafder, S., Ahmed, M. T., Iqbal, S., Hoque, M. T., and Rahman, M. S. (2018). RBSURFPred: modeling protein accessible surface area in real and binary space using regularized and optimized regression. *J. Theoretical Biol.* 441, 44–57. doi: 10.1016/j.jtbi.2017.12.029
- Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., and Wilke, C. O. (2013). Maximum allowed solvent accessibilities of residues in proteins. *PloS one* 8:e80635. doi: 10.1371/journal.pone.0080635
- Wang, C., Li, S., Xi, L., Liu, H., and Yao, X. (2011). Accurate prediction of the burial status of transmembrane residues of α -helix membrane protein by incorporating the structural and physicochemical features. *Amino acids* 40, 991–1002. doi: 10.1007/s00726-010-0727-8
- Weihong, C., Bin, C., and Jianfeng, Y. (2019). Transmembrane protein 126B protects against high fat diet (HFD)-induced renal injury by suppressing dyslipidemia via inhibition of ROS. *Biochem. Biophys. Res. Commun.* 509, 40–47. doi: 10.1016/j.bbrc.2018.12.003
- Wold, H. (1987). Response to DA freedman. *J. Educ. Stat.* 12, 202–205. doi: 10.3102/10769986012002202
- Xiao, F., and Shen, H.-B. (2015). Prediction enhancement of residue real-value relative accessible surface area in transmembrane helical proteins by solving the output preference problem of machine learning-based predictors. *J. Chem. Inf. Mod.* 55, 2464–2474. doi: 10.1021/acs.jcim.5b00246
- Yan, J., Jiang, Y., Lu, J., Wu, J., and Zhang, M. (2019). Inhibiting of proliferation, migration, and invasion in lung cancer induced by silencing interferon-induced transmembrane protein 1 (IFITM1). *BioMed Res. Int.* 2019:9085435.
- Yin, X., Yang, J., Xiao, F., Yang, Y., and Shen, H.-B. (2018). MemBrain: an easy-to-use online webserver for transmembrane protein structure prediction. *Nanomicro Lett.* 10:2.
- Yuan, Z., Zhang, F., Davis, M. J., Bodén, M., and Teasdale, R. D. (2006). Predicting the solvent accessibility of transmembrane residues from protein sequence. *J. Proteome Res.* 5, 1063–1070. doi: 10.1021/pr050397b
- Zeng, B., Hoenigschmid, P., and Frishman, D. (2019). Residue co-evolution helps predict interaction sites in α -helical membrane proteins. *J. Struct. Biol.* 206, 156–169. doi: 10.1016/j.jsb.2019.02.009
- Zhang, J., Zhang, Y., and Ma, Z. (2019). In-silico prediction of human secretory proteins in plasma based on discrete firefly optimization and application to cancer biomarkers identification. *Front. Genet.* 10:542. doi: 10.3389/fgene.2019.00542

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.

Copyright © 2021 Liu, Gong, Guo, Zhang, Lu, Zhang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Advances in the Identification of Circular RNAs and Research Into circRNAs in Human Diseases

Shihu Jiao^{1,2†}, Song Wu^{3†}, Shan Huang⁴, Mingyang Liu^{5*} and Bo Gao^{6*}

¹ Hainan Key Laboratory for Computational Science and Application, Hainan Normal University, Haikou, China, ² Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou, China, ³ Director of Preventive Treatment of Disease Centre, Qinhuangdao Hospital of Traditional Chinese Medicine, Qinhuangdao, China, ⁴ Department of Neurology, The Second Affiliated Hospital, Harbin Medical University, Harbin, China, ⁵ Department of Internal Medicine-Oncology, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China, ⁶ Department of Radiology, The Second Affiliated Hospital, Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:

Fa Zhang,
Institute of Computing Technology,
Chinese Academy of Sciences (CAS),
China

Reviewed by:

Xiangxiang Zeng,
Hunan University, China
Yushan Qiu,
Shenzhen University, China

*Correspondence:

Mingyang Liu
redrumff@163.com
Bo Gao
1678729588@qq.com

[†]These authors have contributed
equally to this work

Specialty section:

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

Received: 07 February 2021

Accepted: 01 March 2021

Published: 19 March 2021

Citation:

Jiao S, Wu S, Huang S, Liu M and
Gao B (2021) Advances in
the Identification of Circular RNAs
and Research Into circRNAs
in Human Diseases.
Front. Genet. 12:665233.
doi: 10.3389/fgene.2021.665233

Circular RNAs (circRNAs) are a class of endogenous non-coding RNAs (ncRNAs) with a closed-loop structure that are mainly produced by variable processing of precursor mRNAs (pre-mRNAs). They are widely present in all eukaryotes and are very stable. Currently, circRNA studies have become a hotspot in RNA research. It has been reported that circRNAs constitute a significant proportion of transcript expression, and some are significantly more abundantly expressed than other transcripts. CircRNAs have regulatory roles in gene expression and critical biological functions in the development of organisms, such as acting as microRNA sponges or as endogenous RNAs and biomarkers. As such, they may have useful functions in the diagnosis and treatment of diseases. CircRNAs have been found to play an important role in the development of several diseases, including atherosclerosis, neurological disorders, diabetes, and cancer. In this paper, we review the status of circRNA research, describe circRNA-related databases and the identification of circRNAs, discuss the role of circRNAs in human diseases such as colon cancer, atherosclerosis, and gastric cancer, and identify remaining research questions related to circRNAs.

Keywords: circRNAs, database, machine learning, circRNAs identification, diseases

INTRODUCTION

Circular RNAs (circRNAs) are endogenous non-coding RNAs (ncRNAs) that have gained increasing attention in recent years. circRNAs are formed by exon or intron cyclization that ligates the 5' terminal cap and 3' terminal poly(A) tail to form a circular structure. They are mainly located in the cytoplasm or stored in exosomes, are unaffected by RNA exonucleases, are more stably expressed and less susceptible to degradation, and have been shown to exist in a wide variety of eukaryotic organisms (Li Y. et al., 2015; Pradeep et al., 2020). The widespread existence of circRNAs suggests that they have certain biological functions as lncRNAs and microRNAs (miRNAs) play (Jiang et al., 2009, 2014, 2015; Wang et al., 2014; Cheng L. et al., 2019; Liang et al., 2019; Wei and Liu, 2020; Yang et al., 2020). In recent years, studies have shown a diversity of formation mechanisms

and biological functions of circRNAs. circRNAs are formed by various mechanisms; for example, spliceosomes (intracellular protein–RNA complexes) catalyze splicing as follows (Salgia et al., 2003): first, the spliceosome recognizes introns, which are flanked by the splice donor (or 5' splice site) and the splice acceptor (or 3' splice site) with specific sequences at the 5' and 3' ends; then, the 2' hydroxyl group of the downstream sequence attacks the splice donor, resulting in a circular intron lariat structure; finally, the 3' hydroxyl group of the upstream exon splice donor attacks the splice acceptor, the upstream and downstream exons are sequentially spliced to form a linear structure, and the intron lariat structure is usually degraded rapidly by debranching enzyme. Variable splicing is the process by which a precursor mRNA (pre-mRNA) can be transcribed from different RNA splicing methods; that is, different combinations of splice sites, to produce mutually exclusive mRNA splice isoforms, which in turn are translated to produce different protein products (Pan et al., 2008). This is the main function of RNA cyclization. Cyclization of circRNAs can be divided into intron and exon cyclization (Sanger et al., 1976), and the current mainstream cyclization mechanisms are categorized as follows: (1) exon skipping, (2) direct back-splicing of intron, (3) circRNA formation by RNA-binding proteins (RBPs; Chen, 2016; Zhang et al., 2018), and (4) circular intron RNA cyclization (Stoddard, 2014); the detailed mechanisms are shown in **Figure 1**. The diversity of circRNAs, and thus their diverse biological functions, is a direct result of these multiple formation mechanisms. For example, circRNAs can act as miRNA sponges (Hansen et al., 2013; Memczak et al., 2013; Zhao et al., 2020a), be translated into proteins (Yang et al., 2017), bind functional proteins (Li Z. et al., 2015), regulate RNA splicing (Conn et al., 2017), and regulate transcription (Chao et al., 1998; Memczak et al., 2013). Therefore, the identification of circRNAs contributes to our understanding of the formation and biological functions of circRNAs.

In 1976, Kolakofsky (1976) observed, for the first time, defective interfering RNAs in parainfluenza virus particles using electron microscopy. Sanger et al. (1976) discovered that plant-infecting viroids are a class of single-stranded, circular RNA molecules that have characteristics such as high thermal stability and a natural circular structure by self-complementary. In 1979, similar circular transcripts were found in HeLa cells and yeast mitochondria by electron microscopy (Hsu and Coca-Prados, 1979). In 1981, a ribosomal RNA (rRNA) gene was discovered in *Tetrahymena* that contained an intron sequence that formed a circular RNA after splicing. In 1988, the intron of 23S rRNA in archaea was found to be spliced at a specific site to form a stable circular RNA and to function as a transposon. In 1991, researchers identified several circular transcripts formed by different splicing patterns in the human oncogene DCC (Nigro et al., 1991), and these circular RNAs were then found in human *ETS1* gene, mouse *Sry* (sex-determining region Y) gene, rat cytochrome P450 *2C24* gene and human P450 *2C18* gene.

Despite their early discovery, research on circRNAs has been slow in recent decades. Although circRNAs were discovered decades ago, they could not be detected by molecular techniques that relied on poly(A) enrichment because they did not have free 3' and 5' ends. Instead, cyclizable exons were spliced

by reverse splicing, which was different from regular linear splicing. Moreover, the mapping algorithm of early transcriptome analysis could not directly map the sequenced fragments to the genome, leading to the idea that circRNAs were byproducts of missplicing. With the development of high-throughput sequencing and bioinformatics technologies, it was first proposed in 2012 that circRNAs are circular transcripts generated by reverse splicing of mRNA precursors, which are found to exist in large quantities in different types of human cells. In 2013, it was found that circRNAs can act as a sponge for miRNAs (Hansen et al., 2013; Memczak et al., 2013), which regulate the growth and development of organisms. Since then, circRNAs have rapidly become a research hotspot. To identify circRNAs, in addition to high-throughput techniques (RNA-seq), common analytical and computational methods are used, such as CIRC (Gao et al., 2015), segemehl (Hoffmann et al., 2014), Mapslice (Wang et al., 2010), and CircSeq (Guo et al., 2014). In recent years, researchers have developed machine learning methods to identify circRNAs based on the above methods (Yin et al., 2021). Feature selection is an important part of these machine learning models. Feature selection, aiming to select a subset of features by eliminating redundant and noise features, is an important preprocessing step in bioinformatics. Recently, Su et al. (2018) proposed a binomial distribution based method to perform feature selection in computational genomics. The effectiveness of their method has been proved by predicting lncRNA subcellular localizations (Su et al., 2018). Since both nucleotide and amino acid composition obey binomial distribution, this method is suggested to be used for genomic and proteomic analysis. We provide here an overview of the research progress of circRNAs, including the development of circRNA databases, identification of circRNAs, and the role of circRNAs in human diseases such as colon cancer, atherosclerosis, and gastric cancer.

circRNA-RELATED DATABASES

In recent years, as circRNA research has progressed, an increasing number of circRNAs have been discovered in different species, and circRNA-related databases have been created. Some of the main circRNA databases published so far are listed below.

- (1) circBase collects and merges public circRNA datasets and provides evidence of the genomic catalog of their expression, as well as scripts to identify circRNAs in sequencing data¹ (Glazar et al., 2014).
- (2) Circ2Trait is a comprehensive database that includes potential associations of circRNAs with diseases and traits by studying the interaction network of circRNAs with miRNAs and calculating their internal SNPs and Argonaute (Ago) interaction sites² (Ghosal et al., 2013).
- (3) deepBase contains about 150,000 circRNA genes from organisms, including human, mouse, *Drosophila*, and nematode. This database also constructs the

¹<http://www.circbase.org/>

²<http://gyanxet-beta.com/circdb/>

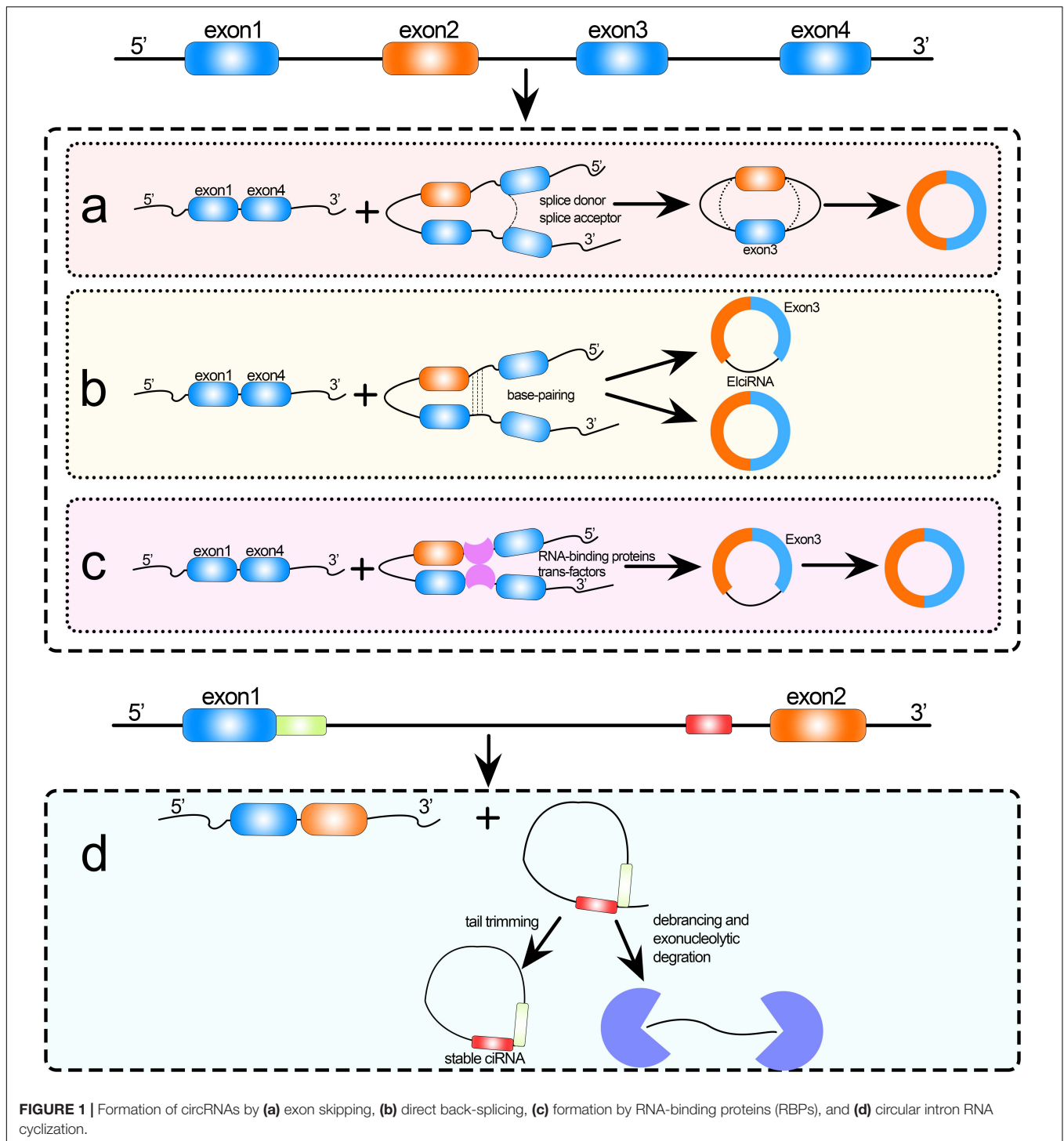


FIGURE 1 | Formation of circRNAs by (a) exon skipping, (b) direct back-splicing, (c) formation by RNA-binding proteins (RBPs), and (d) circular intron RNA cyclization.

most comprehensive expression map of circRNAs³ (Yang et al., 2010).

- (4) CirNet mainly includes RNA-seq data of more than 400 samples from 26 tissues collected from the sequence read archive database. This database not only includes basic information on circRNAs but also provides expression

profile data of circRNAs in different tissues and the competing endogenous (ce)RNA regulatory network of circRNAs-miRNA-gene⁴ (Liu et al., 2016).

- (5) starBase v2.0 integrates published circRNA data and constructs interaction networks of miRNAs with circRNAs and circRNAs with RBPs. In addition, the database looks

³<http://deepbase.sysu.edu>

⁴<http://syslab5.nchu.edu.tw/CircNet>

for potential miRNA–ncRNA, miRNA–mRNA, ncRNA–RNA, RBP–ncRNA, and RBP–mRNA interactions through high-throughput data. starBase also predicts the function of ncRNAs from miRNA-mediated (ceRNA) regulatory networks (miRNAs, lncRNAs, and pseudogenes) and protein-coding genes using the online tools miRFunction and ceRNAFunction⁵ (Li et al., 2014).

TOOLS FOR RECOGNITION OF circRNAs

Because of the low expression level of circRNAs and limitations of previous computational methods, these RNA molecules were only found in small numbers in individual genes and therefore initially thought to be products of missplicing, byproducts of RNA splicing, incidental in animals, or precursors of linear RNAs. In recent years, with improved experimental and computational methods for circRNAs and the use of next-generation high-throughput sequencing technologies (Wang et al., 2009; Zeng et al., 2017, 2019), a large number of stable circRNAs have now been found in a variety of cells, and 85% of circRNAs can be mapped to known genes, of which 84% overlap with coding exons (Memczak et al., 2013). Because of the special structure of circRNAs—they lack a 5′ terminal cap and a 3′ terminal poly(A) tail and have a closed-loop structure with covalent bonds—and their maturation mechanism, early sequencing methods could not easily detect such molecules. Improvements in sequencing analysis techniques and computational methods have made detection more efficient (Malysiak-Mrozek et al., 2019; Mrozek, 2020). Therefore, studies on the identification of circRNAs are reviewed from two aspects: (1) identification based on sequencing data and (2) identification based on sequence features and machine learning methods.

Identification of circRNAs Based on Sequencing

Many algorithms exist for circRNA identification, including CIRC (Gao et al., 2015), segemehl (Hoffmann et al., 2014), Msplice (Wang et al., 2010), CircSeq (Guo et al., 2014), and find_circ (Memczak et al., 2013). Using these algorithms, researchers have identified a large number of circRNAs in human, mouse, nematode, archaea, and other organisms (Yang et al., 2011; Jeck and Sharpless, 2014). We describe here several of these commonly used sequencing-based tools for identification of circRNAs.

CIRI (Stoddard, 2014) was developed by Gao et al. (2015) to comprehensively identify circRNAs, and it is based on the novel chiastic clipping signal algorithm. CIRI can accurately detect circRNAs from transcriptomic data without bias through multiple filtering strategies. This tool is mainly used to identify and annotate circRNAs from RNA-seq data. Unlike other methods for annotating circRNAs, CIRI eliminates false positives by using a new algorithm based on paired cross-clip signal detection in the BWA-MEM sequence alignment/map and combining it with systematic filtering.

⁵<http://starbase.sysu.edu.cn/>

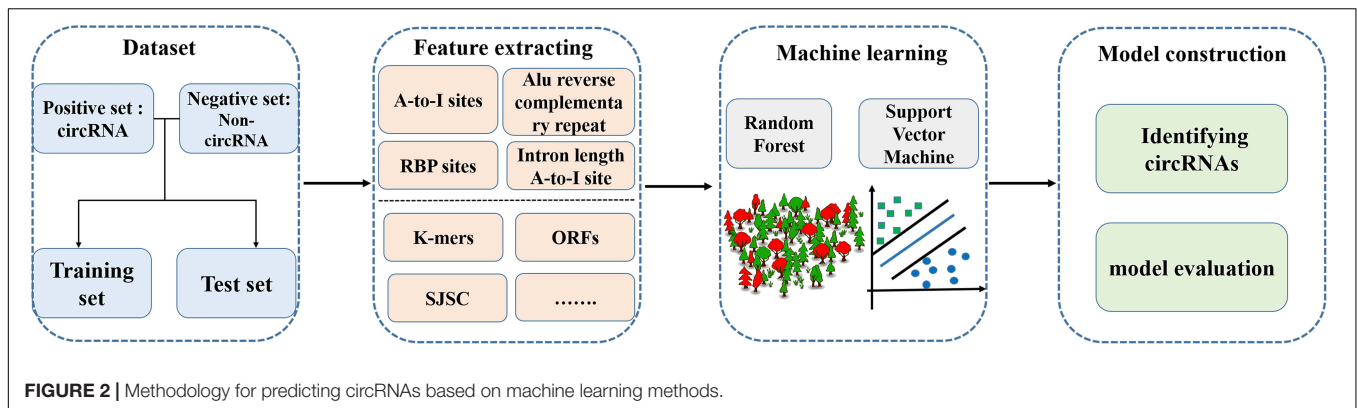
CIRCexplorer, a tool for identifying circRNAs developed by Zhang et al. (2014), was the first to elucidate the regulatory mechanism of complementary sequences on production of exon-derived circRNAs. This tool revealed that regulation of variable cyclization was mediated by competitive pairing of complementary sequences, providing a new theoretical perspective on the complexity and diversity of gene expression at the transcriptional and posttranscriptional levels. Nearly 10,000 circRNAs were identified in human embryonic stem cell line H9 using a special nuclease to enrich circRNAs in combination with computational analysis software, demonstrating exon cyclization mediated by the complementary sequence of intron RNA. Competitive pairing of complementary sequences between different regions can selectively generate either linear RNAs or circRNAs.

CircSeq, a tool developed by Guo et al. (2014) to identify and characterize mammalian circRNAs, is a computational pipeline to identify and quantify the relative abundance of circRNAs from RNA-seq databases. Compared with other identification tools, CircSeq does not require available gene annotation to identify circRNAs. The application of the identification tool to non-polyA-selected RNA sequencing data in the ENCODE project proved its ability to classify and globally characterize more than 7000 human circRNAs.

The above sequencing methods all identify back-splicing sites from high-throughput sequencing data to detect circRNAs. In comparing some of the above identification tools, Hansen et al. (2016) and Sekar et al. (2019) found that only a small percentage of circRNAs could be predicted simultaneously by these tools, indicating significant differences and species variability. Therefore, the above tools developed around high-throughput sequencing technology have poor identification performance and low consistency. Moreover, these tools generally have high false-positive rates and low sensitivity (Hansen et al., 2016). To address these shortcomings, researchers have developed tools to identify circRNAs on the basis of sequence features and machine learning.

Identification of circRNAs Based on Sequence Features and Machine Learning

Identifying circRNAs using sequence features that distinguish circRNAs from linear RNAs (especially mRNAs that encode proteins) is an urgent problem to be solved in bioinformatics. In recent years, the combination of sequence features and machine learning has been successfully used to solve biological problems such as the prediction of gene regulatory sites and splice sites (Wang et al., 2008; Xiong et al., 2015), and protein function (Cao et al., 2017; Gbenro et al., 2020; Hippe, 2020; Zhai et al., 2020), etc (Mrozek et al., 2007, 2009; Wei et al., 2017b,c, 2018; Jin et al., 2019; Stephenson et al., 2019; Su et al., 2019a,b; Liu B. et al., 2020; Liu Y. et al., 2020; Smith et al., 2020; Zhao et al., 2020b,c). Some tools have been developed to identify circRNAs using sequence features and machine learning methods. The basic framework of using machine learning methods to predict circRNAs is shown in **Figure 2**.



One study selected 100 RNA circularization-related sequence features, including length, adenosine-to-inosine (A-to-I) density, and Alu sequences of introns upstream and downstream of the splice site, and established a machine learning model to identify circRNAs in the human genome. The classification abilities of two machine learning methods, random forest (RF; Cheng et al., 2019b; Liu et al., 2019) and support vector machine (SVM; Jiang et al., 2013; Wei et al., 2014, 2017a, 2019; Zhao et al., 2015; Cheng, 2019; Hong et al., 2020; Li and Liu, 2020; Shao and Liu, 2020), were also compared. The results showed that the selected sequence features could effectively identify RNA circularization and that different sequence features contribute differently to the classification and prediction ability of the model. The RF method showed better classification than the SVM method.

In 2021, Yin et al. (2021) constructed a tool, named PCirc, to identify circRNAs using multiple sequence features and RF classification. This tool specifically targets the identification of circRNAs in plants, mainly from RNA sequence data. The tool encodes the sequence information of rice circRNAs by using three feature-encoding methods: k-mers, open reading frames, and splicing junction sequence coding (SJSC). The accuracy of the encoded information is greater than 80% when using the RF method for identification. The identification model can be used not only for the identification of rice circRNAs, but also for the recognition of circRNAs in plants such as *Arabidopsis thaliana*.

circRNAs AND HUMAN DISEASES

In terms of disease diagnosis, studies have found that the exosomes released by cancer cells contain abundant circRNAs, suggesting that circRNAs might be used as biological markers for clinical diagnosis. The key when using circRNAs for disease prediction is to identify the interaction site between the circRNA and miRNA or RBP, and then indirectly determine the association between the circRNA and disease by analyzing the relationship between the miRNA or RBP and disease (Jiang et al., 2010; Cheng et al., 2018; Liu, 2020; Zeng et al., 2020; Zuo et al., 2020).

In 2015, Li Y. et al. (2015) reported that exosomes are enriched with circRNAs, so it is possible that diseases such as colon cancer could be diagnosed by detecting circRNAs in serum. Aberrant expression of circRNAs in colorectal cancer and pancreatic ductal

adenocarcinoma has been used as a diagnostic or predictive biomarker. By studying their expression profile, it was found that circRNAs may be associated with the molecular pathogenesis of cutaneous basal cell carcinoma (Sand et al., 2016).

The first validated circRNA, cANRIL, is closely related to a single nucleotide polymorphism (SNP) that is thought to alter the splicing of cANRIL, leading to expression of the *INK4A/ARF* loci, resulting in an increased incidence of atherosclerosis (Burd et al., 2010). Hypoxia is one of the key factors contributing to the development of atherosclerosis, and is therefore also regulated by circRNA (Boeckel et al., 2015).

Xu et al. (2015) showed that mice of a transgenic line overexpressing the *miR-7* gene in β -cells developed diabetes mellitus. The same study showed that overexpression of the circRNA ciRS-7 inhibited miR-7 function and thus improved insulin secretion. Potential target genes of *miR-7* have been identified by bioinformatics analysis and include *Myrip* (a gene regulating insulin secretory granules) and *Pax6* (a gene enhancing insulin transcription).

A study by Li P. et al. (2015) identified the circRNA hsa-circ002059 as being associated with gastric cancer. In that study, expression of this circRNA was downregulated in gastric tissues of patients compared with healthy controls. In addition, hsa-circ002059 was found at significantly lower levels in plasma of patients with gastric cancer than in healthy controls.

In bladder cancer, circRNAs have been identified using high-throughput microarray technology. Using this approach, Zhong et al. (2016) found two downregulated circRNAs (circFAM169A and circTRIM24) and 4 upregulated circRNAs (circTCF25, circZFR, circPTK2, and circBC048201) in bladder cancer tissue compared with adjacent non-tumor tissues. In addition, in the cancer tissues, circTCF25 could increase expression of the *CDK6* gene by modulating miR-103a-3p and miR-107. This is closely related to the development of cancer.

Qin et al. (2016) identified hsa-cir0001649 in hepatocellular carcinoma (HCC) and found that its expression was significantly decreased compared with that in adjacent normal liver tissue. In contrast, Shang et al. (2016) found that another circRNA, hsa-cir0005075, was significantly downregulated in HCC compared with adjacent normal tissue.

Exosomes are highly enriched with circRNAs. Exosomes are extracellular vesicles, 40 to 160 nm in diameter, that function

as important intercellular signaling pathways (Li Y. et al., 2015; Kalluri and LeBleu, 2020). The exosome database exoRBase included 92 sequenced samples of serum exosomes, including samples from healthy volunteers and patients with coronary heart disease and colon cancer. The exosome samples contained 58,330 circRNAs and 18,333 mRNAs (Li et al., 2018). Zhang et al. (2019) demonstrated that circNRIP1, when secreted via exosome, can be taken up by gastric cancer cells and promote their proliferation, migration, and invasion. Therefore, exosomes can be regarded as *in vivo* carriers of circRNAs that can amplify their biological functions.

CHALLENGES AND PROSPECTS

Compared with long non-coding RNAs and miRNAs, research on circRNAs is still in its infancy and many questions remain to be answered, primarily in four areas:

- (1) Transport and degradation: because circRNAs can resist RNase digestion and are stable in cells, the process of their degradation is unclear.
- (2) Formation: it is unknown whether circRNAs are produced during or after transcription.
- (3) Expression, translation, and function of circRNAs: circRNAs have stable structures and are highly conserved, underpinning their ability to play important roles in different organisms. Their unconfirmed roles, including acting as miRNA sponges, regulating gene expression, and targeting RBPs, require comprehensive and extensive elucidation.
- (4) Research methodology: the experimental methodologies and bioinformatics used to identify circRNAs are challenging. For example, in experimental methods, general RNA-seq procedures such as reverse transcription may cause technical mis-ligation and generate a large number of artificial circRNAs. These pseudo circRNAs can

account for 34–55% of the sequencing quantity, seriously affecting the accuracy of the data. As for methods that use machine learning and sequence features, only a few identification tools exist and their accuracy needs to be improved. These tools are not stable across different species. Therefore, in the future, stable identification models and deep learning methods are needed to establish identification tools for circRNAs and improve the robustness of the models.

Accurate identification will help determine additional biological functions of circRNAs. The unique features of circRNAs such as ceRNA may provide new ideas for drug discovery and development. The tissue specificity and stability of circRNAs make them potentially useful biomarkers. In the near future, it is likely that circRNAs will play important roles in the prevention, diagnosis, and treatment of various diseases.

AUTHOR CONTRIBUTIONS

ML and BG: conceptualization, writing—review and editing, and supervision. SJ, SH, and SW: investigation and writing—original draft preparation. All authors have read and agreed to the published version of the manuscript.

FUNDING

The work was supported by National Natural Science Foundation of China (No. 62002087).

ACKNOWLEDGMENTS

We thank Louise Adam, ELS(D), from Liwen Bianji, Edanz Editing China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

REFERENCES

- Boeckel, J. N., Jae, N., Heumüller, A. W., Chen, W., Boon, R. A., Stellos, K., et al. (2015). Identification and characterization of hypoxia-regulated endothelial circular RNA. *Circ. Res.* 117, 884–890.
- Burd, C. E., Jeck, W. R., Liu, Y., Sanoff, H. K., Wang, Z., and Sharpless, N. E. (2010). Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet.* 6:e1001233. doi: 10.1371/journal.pgen.1001233
- Cao, R., Freitas, C., Chan, L., Sun, M., Jiang, H., and Chen, Z. (2017). ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network. *Molecules* 22:1732. doi: 10.3390/molecules22101732
- Chao, C. W., Chan, D. C., Kuo, A., and Leder, P. (1998). The mouse formin (Fmn) gene: abundant circular RNA transcripts and gene-targeted deletion analysis. *Mol. Med.* 4, 614–628. doi: 10.1007/bf03401761
- Chen, L. L. (2016). The biogenesis and emerging roles of circular RNAs. *Nat. Rev. Mol. Cell Biol.* 17, 205–211. doi: 10.1038/nrm.2015.32
- Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/156652321904191022113307
- Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002
- Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucl. Acids Res.* 47, D140–D144.
- Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019b). Computational methods for identifying similar diseases. *Mol. Ther. Nucl. Acids.* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019
- Conn, V. M., Hugouvieux, V., Nayak, A., Conos, S. A., Capovilla, G., Cildir, G., et al. (2017). A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. *Nat. Plants* 3:17053.
- Gao, Y., Wang, J., and Zhao, F. (2015). CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biol.* 16:4.
- Ghenro, S., Hippe, K., and Cao, R. (2020). “HMMeta: Protein function prediction using hidden markov models,” in *Proceedings of the BCB '20: 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (New York, NY: Association for Computing Machinery).
- Ghosal, S., Das, S., Sen, R., Basak, P., and Chakrabarti, J. (2013). Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front. Genet.* 4:283. doi: 10.3389/fgene.2013.00283

- Glazar, P., Papavasiliou, P., and Rajewsky, N. (2014). circBase: a database for circular RNAs. *RNA* 20, 1666–1670. doi: 10.1261/rna.043687.113
- Guo, J. U., Agarwal, V., Guo, H., and Bartel, D. P. (2014). Expanded identification and characterization of mammalian circular RNAs. *Genome Biol.* 15:409.
- Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K., et al. (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495, 384–388. doi: 10.1038/nature11993
- Hansen, T. B., Veno, M. T., Damgaard, C. K., and Kjems, J. (2016). Comparison of circular RNA prediction tools. *Nucl. Acids Res.* 44:e58. doi: 10.1093/nar/gkv1458
- Hippe, K. (2020). “Sola gbenro; renzhi cao in *prolango2: protein function prediction with ensemble of encoder-decoder networks*,” in *Proceedings of the BCB '20: 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (New York, NY: Association for Computing Machinery).
- Hoffmann, S., Otto, C., Dose, G., Tanzer, A., Langenberger, D., Christ, S., et al. (2014). A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biol.* 15:R34.
- Hong, Z., Zeng, X., Wei, L., and Liu, X. (2020). Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 36, 1037–1043.
- Hsu, M. T., and Coca-Prados, M. (1979). Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature* 280, 339–340. doi: 10.1038/280339a0
- Jeck, W. R., and Sharpless, N. E. (2014). Detecting and characterizing circular RNAs. *Nat. Biotechnol.* 32, 453–461. doi: 10.1038/nbt.2890
- Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* 4(Suppl. 1):S2. doi: 10.1186/1752-0509-4-S1-S2
- Jiang, Q., Ma, R., Wang, J., Wu, X., Jin, S., Peng, J., et al. (2015). lncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. *BMC Genomics* 16(Suppl. 3):S2. doi: 10.1186/1471-2164-16-S3-S2
- Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* 8, 282–293. doi: 10.1504/ijdm.2013.056078
- Jiang, Q., Wang, J., Wang, Y., Ma, R., Wu, X., and Li, Y. (2014). TF2lncRNA: identifying common transcription factors for a list of lncRNA genes from ChIP-Seq data. *Biomed Res. Int.* 2014:317642.
- Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., et al. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucl. Acids Res.* 37, D98–D104.
- Jin, Q., Meng, Z., Tuan, D. P., Chen, Q., Wei, L., and Su, R. (2019). DUNet: a deformable. *Knowl. Based Syst.* 178, 149–162. doi: 10.1016/j.knsys.2019.04.025
- Kalluri, R., and LeBleu, V. S. (2020). The biology, function, and biomedical applications of exosomes. *Science* 367:eau6977. doi: 10.1126/science.aau6977
- Kolakofsky, D. (1976). Isolation and characterization of Sendai virus DI-RNAs. *Cell* 8, 547–555. doi: 10.1016/0092-8674(76)90223-3
- Li, C. C., and Liu, B. (2020). MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. *Brief. Bioinform.* 21, 2133–2141. doi: 10.1093/bib/bbz133
- 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. *Nucl. Acids Res.* 42, D92–D97.
- Li, P., Chen, S., Chen, H., Mo, X., Li, T., Shao, Y., et al. (2015). Using circular RNA as a novel type of biomarker in the screening of gastric cancer. *Clin. Chim. Acta* 444, 132–136. doi: 10.1016/j.cca.2015.02.018
- Li, S., Li, Y., Chen, B., Zhao, J., Yu, S., Tang, Y., et al. (2018). exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucl. Acids Res.* 46, D106–D112.
- Li, Y., Zheng, Q., Bao, C., Li, S., Guo, W., Zhao, J., et al. (2015). Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Res.* 25, 981–984. doi: 10.1038/cr.2015.82
- Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., et al. (2015). Exon-intron circular RNAs regulate transcription in the nucleus. *Nat. Struct. Mol. Biol.* 22, 256–264. doi: 10.1038/nsmb.2959
- Liang, C., Changlu, Q., He, Z., Tongze, F., and Xue, Z. (2019). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucl. Acids Res.* 48:7603.
- Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucl. Acids Res.* 47:e127. doi: 10.1093/nar/gkz740
- Liu, B., Zhu, Y., and Yan, K. (2020). Fold-LTR-TCP: protein fold recognition based on triadic closure principle. *Brief. Bioinform.* 21, 2185–2193. doi: 10.1093/bib/bbz139
- Liu, Y. C., Li, J. R., Sun, C. H., Andrews, E., Chao, R. F., Lin, F. M., et al. (2016). CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucl. Acids Res.* 44, D209–D215.
- Liu, Y., Huang, Y., Wang, G., and Wang, Y. (2020). A deep learning approach for filtering structural variants in short read sequencing data. *Brief Bioinform.* doi: 10.1093/bib/bbaa370
- Liu, Z. P. (2020). Predicting lncRNA-protein interactions by machine learning methods: a review. *Curr. Bioinform.* 15, 831–840. doi: 10.2174/157489361566200224095925
- Malysiak-Mrozek, B., Baron, T., and Mrozek, D. (2019). Spark-IDPP: high-throughput and scalable prediction of intrinsically disordered protein regions with Spark clusters on the cloud. *Cluster Comput. J. Net. Softw. Tools Appl.* 22, 487–508. doi: 10.1007/s10586-018-2857-9
- Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., et al. (2013). F le noble., N rajewsky, circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495, 333–338. doi: 10.1038/nature11928
- Mrozek, D. (2020). A review of cloud computing technologies for comprehensive microRNA analyses. *Comput. Biol. Chem.* 88:107365. doi: 10.1016/j.compbiolchem.2020.107365
- Mrozek, D., Malysiak, B., and Kozielski, S. (2007). “An optimal alignment of proteins energy characteristics with crisp and fuzzy similarity awards,” in *Proceedings of the 2007 IEEE International Conference on Fuzzy Systems*, Vol. 1-4 (London: IEEE), 1513–1518.
- Mrozek, D., Malysiak-Mrozek, B., and Kozielski, S. (2009). *Alignment of Protein Structure Energy Patterns Represented as Sequences of Fuzzy Numbers*. Cincinnati, OH: IEEE, 35–40.
- Nigro, J. M., Cho, K. R., Fearon, E. R., Kern, S. E., Ruppert, J. M., Oliner, J. D., et al. (1991). Scrambled exons. *Cell* 64, 607–613. doi: 10.1016/0092-8674(91)90244-s
- Pan, Q., Shai, O., Lee, L. J., Frey, J., and Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* 40, 1413–1415. doi: 10.1038/ng.259
- Pradeep, C., Nandan, D., Das, A. A., and Velayutham, D. (2020). Comparative transcriptome profiling of disruptive technology, single-molecule direct RNA sequencing. *Curr. Bioinf.* 15, 165–172. doi: 10.2174/1574893614666191017154427
- Qin, M., Liu, G., Huo, X., Tao, X., Sun, X., Ge, Z., et al. (2016). Hsa_circ_0001649: a circular RNA and potential novel biomarker for hepatocellular carcinoma. *Cancer Biomark.* 16, 161–169.
- Salgia, S. R., Singh, S. K., Gurha, P., and Gupta, R. (2003). Two reactions of *Haloferax volcanii* RNA splicing enzymes: joining of exons and circularization of introns. *RNA* 9, 319–330. doi: 10.1261/rna.2118203
- Sand, M., Bechara, F. G., Sand, D., Gambichler, T., Hahn, S. A., Bromba, M., et al. (2016). Circular RNA expression in basal cell carcinoma. *Epigenomics* 8, 619–632. doi: 10.2217/epi-2015-0019
- Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J., and Kleinschmidt, A. K. (1976). Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *PNAS* 73, 3852–3856. doi: 10.1073/pnas.73.11.3852
- Sekar, S., Geiger, P., Cuyugan, L., Boyle, A., Serrano, G., Beach, T. G., et al. (2019). Identification of circular RNAs using RNA sequencing. *J. Vis. Exp.* 14:e59981. doi: 10.3791/59981
- Shang, X., Li, G., Liu, H., Li, T., Liu, J., Zhao, Q., et al. (2016). Comprehensive circular RNA profiling reveals that hsa_circ_0005075, a new circular RNA biomarker, is involved in hepatocellular carcinoma development. *Medicine* 95:e3811.

- Shao, J., and Liu, B. (2020). ProtFold-DFG: protein fold recognition by combining directed fusion graph and pagerank algorithm. *Brief. Bioinform.* doi: 10.1093/bib/bbaa192
- Smith, J., Conover, M., Stephenson, N., Eickholt, J., Si, D., Sun, M., et al. (2020). TopQA: a topological representation for single-model protein quality assessment with machine learning. *J. Int. J. Comput. Biol. Drug Des.* 13:144. doi: 10.1504/ijcbdd.2020.10026784
- Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., et al. (2019). Survey of machine learning techniques in drug discovery. *Curr. Drug Metab.* 20, 185–193. doi: 10.2174/1389200219666180820112457
- Stoddard, B. L. (2014). Homing endonucleases from mobile group I introns: discovery to genome engineering. *Mobile DNA* 5:7. doi: 10.1186/1759-8753-5-7
- Su, R., Liu, X., Wei, L., and Zou, Q. (2019a). Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response. *Methods (San Diego, Calif.)* 166, 91–102. doi: 10.1016/j.ymeth.2019.02.009
- Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2019b). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 16, 1231–1239. doi: 10.1109/tcbb.2018.2858756
- Su, Z. D., Huang, Y., Zhang, Z. Y., Zhao, Y. W., Wang, D., Chen, W., et al. (2018). iLoc-lncRNA: predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC. *Bioinformatics* 34, 4196–4204.
- Wang, G., Wang, Y., Feng, W., Wang, X., Yang, J. Y., Zhao, Y., et al. (2008). Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genom.* 9 (Suppl. 2):S22. doi: 10.1186/1471-2164-9-S2-S22
- Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., et al. (2010). MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucl. Acids Res.* 38:e178. doi: 10.1093/nar/gkq622
- Wang, P. L., Bao, Y., Yee, M. C., Barrett, S. P., Hogan, G. J., Olsen, M. N., et al. (2014). Circular RNA is expressed across the eukaryotic tree of life. *PLoS One* 9:e90859. doi: 10.1371/journal.pone.0090859
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary. *Nat. Rev. Genet.* 10, 57–63.
- Wei, H., and Liu, B. (2020). iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Brief. Bioinform.* 21, 1356–1367. doi: 10.1093/bib/bbzw057
- Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018). Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* 117, 212–217.
- Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 192–201. doi: 10.1109/tcbb.2013.146
- Wei, L., Tang, J., and Zou, Q. (2017a). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inf. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026
- Wei, L., Wan, S., Guo, J., and Wong, K. K. L. (2017c). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005
- Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019). Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 16, 1264–1273. doi: 10.1109/tcbb.2017.2670558
- Wei, L., Xing, P., Zeng, J., Chen, J. X., Su, R., and Guo, F. (2017b). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001
- Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., et al. (2015). RNA splicing: the human splicing code reveals new insights into the genetic determinants of disease. *Science* 347, 1254806.
- Xu, H., Guo, S., Li, W., and Yu, P. (2015). The circular RNA Cdr1as, via miR-7 and its targets, regulates insulin transcription and secretion in islet cells. *Sci. Rep.* 5:12453.
- Yang, J. H., Shao, P., Zhou, H., Chen, Y. Q., and Qu, L. H. (2010). deepBase: a database for deeply annotating and mining deep sequencing data. *Nucl. Acids Res.* 38, D123–D130.
- Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G., and Chen, L. L. (2011). Genomewide characterization of non-polyadenylated RNAs. *Genome Biol.* 12:R16.
- Yang, Q., Wu, J., Zhao, J., Xu, T., Han, P., and Song, X. (2020). The expression profiles of lncRNAs and their regulatory network during smek1/2 knockout mouse neural stem cells differentiation. *Curr. Bioinform.* 15, 77–88. doi: 10.2174/1574893614666190308160507
- Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., et al. (2017). Extensive translation of circular RNAs driven by N-6-methyladenosine. *Cell Res.* 27, 626–641. doi: 10.1038/cr.2017.31
- Yin, S., Tian, X., Zhang, J., Sun, P., and Li, G. (2021). PCirc: random forest-based plant circRNA identification software. *BMC Bioinf.* 22:10. doi: 10.1186/s12859-020-03944-1
- Zeng, X. X., Lin, W., Guo, M. Z., and Zou, Q. (2019). Details in the evaluation of circular RNA detection tools: reply to Chen and Chuang. *PLoS Comput. Biol.* 15:5. doi: 10.1371/journal.pcbi.1006916
- Zeng, X., Lin, W., Guo, M., and Zou, Q. (2017). A comprehensive overview and evaluation of circular RNA detection tools. *PLoS Comput. Biol.* 13:e1005420. doi: 10.1371/journal.pcbi.1005420
- Zeng, X., Zhong, Y., Lin, W., and Zou, Q. (2020). Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief. Bioinform.* 21, 1425–1436. doi: 10.1093/bib/bbz080
- Zhai, Y., Chen, Y., Teng, Z., and Zhao, Y. (2020). Identifying antioxidant proteins by using amino acid composition and protein-protein interactions. *Front. Cell Dev. Biol.* 8:591487. doi: 10.3389/fcell.2020.591487
- Zhang, X., Wang, S., Wang, H., Cao, J., Huang, X., Chen, Z., et al. (2019). Circular RNA circNRP1 acts as a microRNA-149-5p sponge to promote gastric cancer progression via the AKT1/mTOR pathway. *Mol. Cancer* 18:20.
- Zhang, X.-Q., Wang, H.-B., Zhang, Y., Lu, X., Chen, L.-L., and Yang, L. (2014). Complementary sequence-mediated exon circularization. *Cell* 159, 134–147. doi: 10.1016/j.cell.2014.09.001
- Zhang, Z., Yang, T., and Xiao, J. (2018). Circular RNAs: promising biomarkers for human diseases. *Ebiomedicine* 34, 267–274. doi: 10.1016/j.ebiom.2018.07.036
- Zhao, T., Hu, Y., and Cheng, L. (2020a). Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief. Bioinform.* doi: 10.1093/bib/bbaa212
- Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020b). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428
- Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020c). ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. *BMC Bioinformatics* 21:43. doi: 10.1186/s12859-020-3388-y
- Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA promoter identification in arabidopsis using multiple histone markers. *Biomed Res. Int.* 2015:861402.
- Zhong, Z., Lv, M., and Chen, J. (2016). Screening differential circular RNA expression profiles reveals the regulatory role of circTCF25-miR-103a-3p/miR-107-CDK6 pathway in bladder carcinoma. *Sci. Rep.* 6:30919.
- Zuo, Y., Zou, Q., Li, J., Jiang, M., and Liu, X. (2020). 2lpiRNAPred: a two-layered integrated algorithm for identifying piRNAs and their functions based on LFE-GM feature selection. *RNA biology* 17, 892–902. doi: 10.1080/15476286.2020.1734382

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.

Copyright © 2021 Jiao, Wu, Huang, Liu and Gao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



iEnhancer-EBLSTM: Identifying Enhancers and Strengths by Ensembles of Bidirectional Long Short-Term Memory

Kun Niu^{1†}, Ximei Luo^{2†}, Shumei Zhang¹, Zhixia Teng¹, Tianjiao Zhang^{1*} and Yuming Zhao^{1*}

¹ College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, ² School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

OPEN ACCESS

Edited by:

Fa Zhang,
Institute of Computing Technology,
Chinese Academy of Sciences (CAS),
China

Reviewed by:

Juan Wang,
Inner Mongolia University, China
Xuefeng Cui,
Shandong University, China

*Correspondence:

Tianjiao Zhang
Ztj.hit@gmail.com
Yuming Zhao
zym@nefu.edu.cn

[†] These authors have contributed
equally to this work

Specialty section:

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

Received: 08 February 2021

Accepted: 01 March 2021

Published: 23 March 2021

Citation:

Niu K, Luo X, Zhang S, Teng Z,
Zhang T and Zhao Y (2021)
iEnhancer-EBLSTM: Identifying
Enhancers and Strengths by
Ensembles of Bidirectional Long
Short-Term Memory.
Front. Genet. 12:665498.
doi: 10.3389/fgene.2021.665498

Enhancers are regulatory DNA sequences that could be bound by specific proteins named transcription factors (TFs). The interactions between enhancers and TFs regulate specific genes by increasing the target gene expression. Therefore, enhancer identification and classification have been a critical issue in the enhancer field. Unfortunately, so far there has been a lack of suitable methods to identify enhancers. Previous research has mainly focused on the features of the enhancer's function and interactions, which ignores the sequence information. As we know, the recurrent neural network (RNN) and long short-term memory (LSTM) models are currently the most common methods for processing time series data. LSTM is more suitable than RNN to address the DNA sequence. In this paper, we take the advantages of LSTM to build a method named iEnhancer-EBLSTM to identify enhancers. iEnhancer-ensembles of bidirectional LSTM (EBLSTM) consists of two steps. In the first step, we extract subsequences by sliding a 3-mer window along the DNA sequence as features. Second, EBLSTM model is used to identify enhancers from the candidate input sequences. We use the dataset from the study of Quang H et al. as the benchmarks. The experimental results from the datasets demonstrate the efficiency of our proposed model.

Keywords: enhancer, identification, classification, recurrent neural network, long short-term memory

INTRODUCTION

Enhancers, as cis-acting DNA sequences, are small pieces of DNA that are surrounded by specific proteins that often boost the expression of specific genes, and the specific proteins are always transcription factors (TFs) (Sen and Baltimore, 1986; Krivega and Dean, 2012; Pennacchio et al., 2013; Liu B. et al., 2016, 2018; Nguyen et al., 2019). In fact, enhancers play a highly important role *in vivo*. As we know, enhancers can increase the gene expression by interacting with TFs. By activating the transcription of genes, one way that enhancers influence target gene transcription is by bringing enhancers close to target genes by forming chromatin loops, and the other way is through self-transcription. Either way will bring about increasing of gene expression (Krivega and Dean, 2012). Moreover, it is well known that enhancers can influence human health and many human diseases. Recently, researchers have shown that under evolutionary

constraints, approximately 85% of human DNA corresponds to non-protein-coding sequences with a significant portion constituting cis-regulatory elements. It is therefore not surprising that genetic variations within these regulatory sequences may lead to phenotypic variations and serve as the etiological basis of human disease (Shen and Zou, 2020). This indicates that enhancers might contribute to evolution.

As the amount of histone modifications and other biological data available on public databases and the development of bioinformatics, gene expression and gene control have become increasingly well known (Kleinjan and Lettice, 2008; Liu G. et al., 2016, 2018; Liu et al., 2017; Wang et al., 2020), and study about enhancers is a hot spot currently, especially how to identify enhancers and their strength (Zou et al., 2016; Zacher et al., 2017; Zhang T. et al., 2020). However, there remain many challenges to identify enhancers. For example, enhancers locate in the non-coding regions that occupy 98% of the human genome. This feature leads to a large search space and increases the difficulty. It is also a formidable challenge that enhancers are located 20 kb away from the target genes, or even in another chromosome, unlike promoters are located somewhere around the transcription start sites of genes. These features make identifying the enhancers more difficult (Pennacchio et al., 2013). As a result, in recent years, a large number of researchers have turned their attention to this topic. In 2017, Zacher et al. proposed a hidden Markov model named Genomic State ANotation (GenoSTAN), which is a new unsupervised genome segmentation algorithm that overcomes many limitations, such as unrealistic data distribution assumptions. Although the experience has shown that chromatin state annotation is more effective in predicting enhancers than the transcription-based definition, sensitivity (SN) remains poor (Wang et al., 2020). There are also other algorithms that can be used for enhancer identification and classification. Liu et al. built a predictor that has two layers named “iEnhancer-2L,” which is the first predictor that can identify enhancers with the strength information. The authors used pseudo k-tuple nucleotide composition (PseKNC) to encode the DNA sequences and then made full use of support vector machine (SVM) to build a classifier (Liu B. et al., 2016). In 2018, a new predictor called “iEnhancer-EL” was proposed by Bin Liu et al. iEnhancer-EL is formed through k-mer, subsequence profile, or PseKNC and SVM. Then it obtains the key classifiers and final predictor for layers 1 and 2 (Liu B. et al., 2018; Nguyen et al., 2019). This bioinformatics tool is equivalent to an advanced version of iEnhancer-2L and therefore has better performance than Enhancer-2L. Last year, Quang H. et al. proposed a new model called iEnhancer-ECNN that uses both one-hot encoding and k-mer to encode the sequence and ensembles of convolutional neural networks as the predictor. In our view, it has great improvements in many metrics.

In this study, we build a prediction network named iEnhancer-ensembles of bidirectional long short-term memory (EBLSTM) to identify enhancers and predict their strengths at the same time. We use 3-mer to encode the input DNA sequences. Then we predict enhancers by EBLSTM. Although we only use DNA sequence information, the experimental results prove the effectiveness of our method.

MATERIALS AND METHODS

Benchmark Dataset

The dataset used in our study is collected from previous studies by Liu B. et al. (2016), Liu B. et al. (2018), and Nguyen et al. (2019) and consists of the chromatin states of nine cell lines, including H1ES, K562, GM12878, HepG2, HUVEC, HSMM, NHLF, NHEK, and HMEC (Liu B. et al., 2016). The dataset is divided into two parts; one part is used to train the model. We called this dataset as the development set. The other part is an independent test dataset. As shown in **Figure 1A**, the development set consists of 1484 enhancer samples and 1484 negative samples and it is also the layer 1 dataset for enhancer identification. Moreover, 1484 enhancer samples can be divided into 742 strong enhancer samples and 742 weak enhancer samples, and it is the layer 2 dataset for enhancer classification. As shown in **Figure 1B**, the independent test set contains 200 enhancer samples (100 strong and 100 weak) and 200 negatives. At the same time, the dataset can be presented as follows:

$$Dataset = Dataset_+ \cup Dataset_- \quad (1)$$

$$Dataset_+ = Dataset_{strong} \cup Dataset_{weak} \quad (2)$$

where the *Dataset* is all the data that we used, *Dataset₊* means the positive dataset, which is the enhancers in our study, and *Dataset₋* means the negative dataset, which is the non-enhancer dataset in our study. Therefore, these two formulas mean the *Dataset* consists of *Dataset₊* and *Dataset₋*, and *Dataset₊* consists of *Dataset_{strong}* and *Dataset_{weak}*.

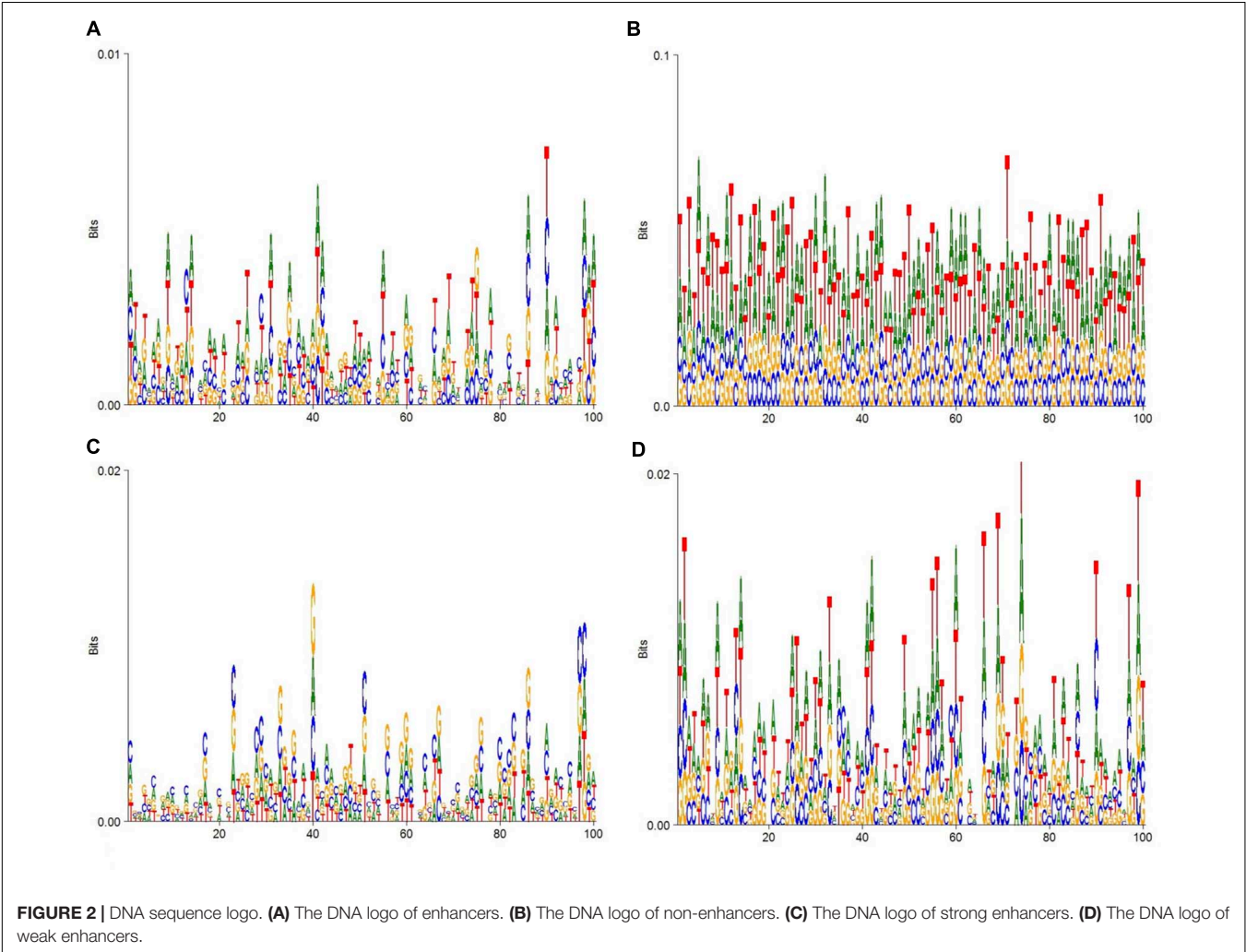
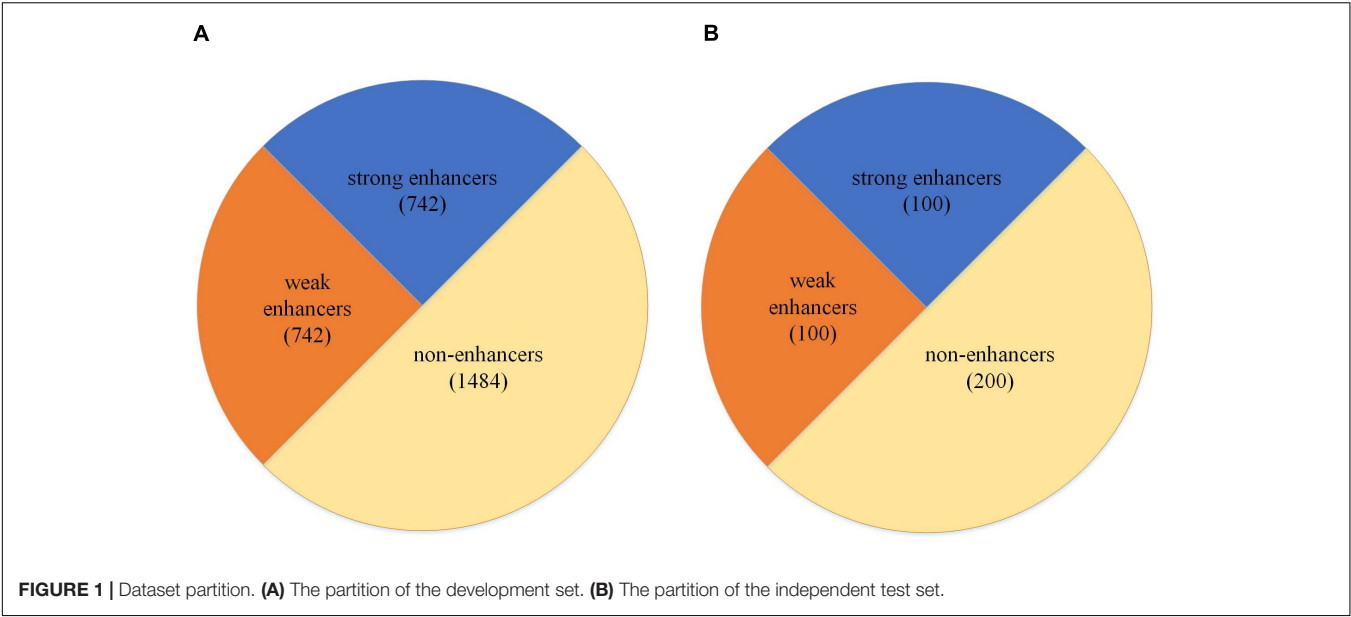
To display the datasets of this experiment more intuitively, DNA consensus sequences of enhancers (**Figure 2A**), non-enhancers (**Figure 2B**), strong enhancers (**Figure 2C**), and weak enhancers (**Figure 2D**) are calculated. As **Figure 2** shows, the specific distributions of A, T, C, and G on these four datasets are different. This means that differences in DNA sequence can be used to distinguish these four types of sequences.

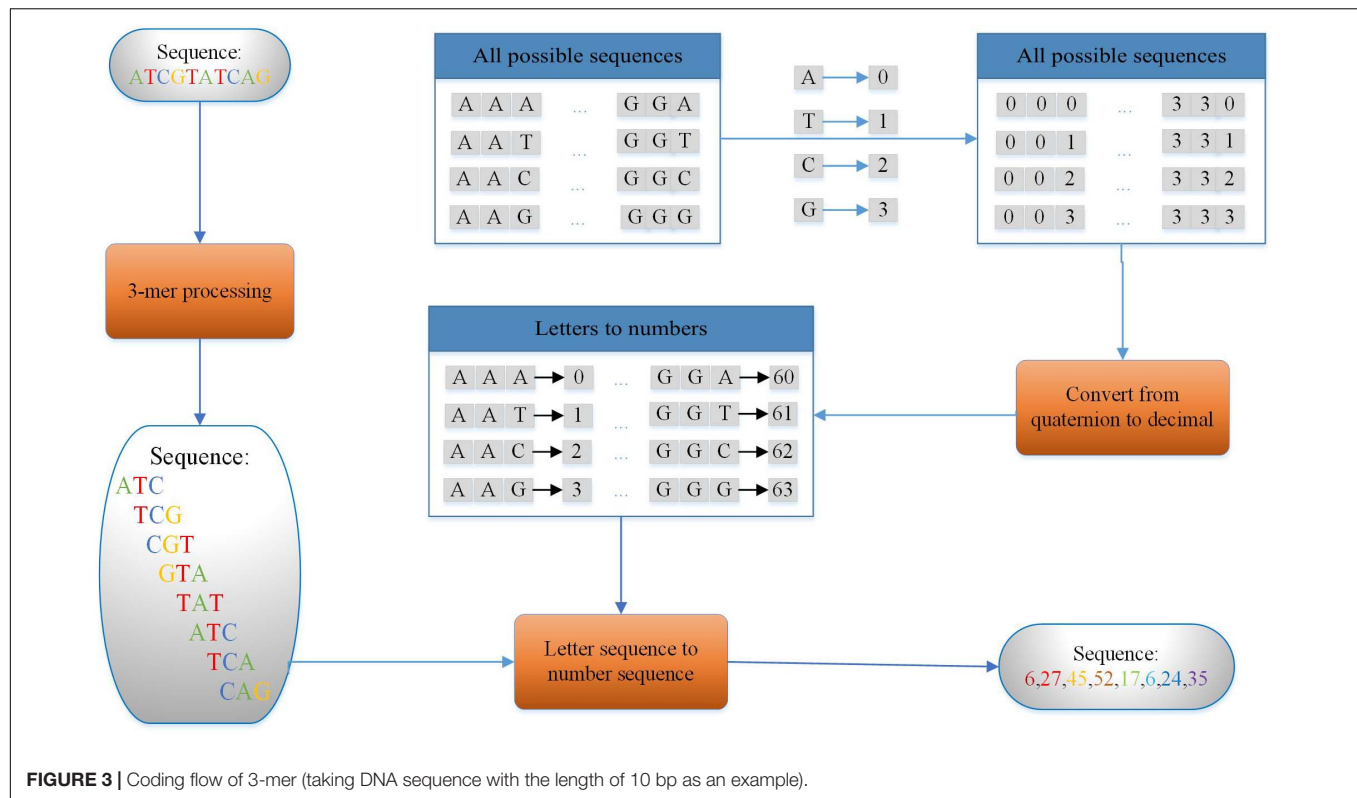
Every enhancer sample has the same length of 200 bp. In the process of building the model, the development set will be divided into five parts, no matter whether in layer 1 or in layer 2, and each part will be the validation in turn and other four parts will be the training set.

Sequence Encoding Scheme

In this study, we use the principle of k-mer (Liu et al., 2019; Zou et al., 2019; Yang et al., 2020; Zhang Z. Y. et al., 2020), which means dividing the nucleic acid sequence into many shorter subsequences with length of k to encode the 200-bp enhancer sequence. As we know, enhancers are a type of DNA sequence and are composed of two kinds of purines (including adenine and guanine) and two kinds of pyrimidines (including cytosine and thymine). Thus, we can encode the obtained sequence of a length of 200 using k-mer ($k = 3$) as a sequence with a length of 198 by the encoding method shown in **Figure 3**. For example, the DNA sequence D is shown as follows:

$$D = \{ATCGTATCAG\} \quad (3)$$





3-Mers are extracted by sliding a 3-mer window along the original DNA sequence with one step as features. The example sequence could be cut into eight such short sequences in S_1 .

$$S_1 = \{ATC, TCG, CGT, GTA, TAT, ATC, TCA, CAG\} \quad (4)$$

Then, eight numbers are used to represent eight short sequences with a strategy that makes each different 3-bp subsequence corresponds to a different number as shown in **Figure 3**. The DNA sequence can be transformed as a number sequence as follows:

$$S_2 = \{6, 27, 45, 52, 17, 6, 24, 35\} \quad (5)$$

Finally, a number sequence of length 8 can be extracted from a 10-bp DNA sequence. Thus, a sequence of 200 bp in the experiment is encoded in this way and a sequence of 198 digits is produced. Using the sequence ATC in S_1 as an example, ATC is regarded as a quaternary three-digit number, A as 0, T as 1, C as 2, and G as 3. Then convert the number in base 3 to base 10. So 64 different 3-mers can be represented by 0–63.

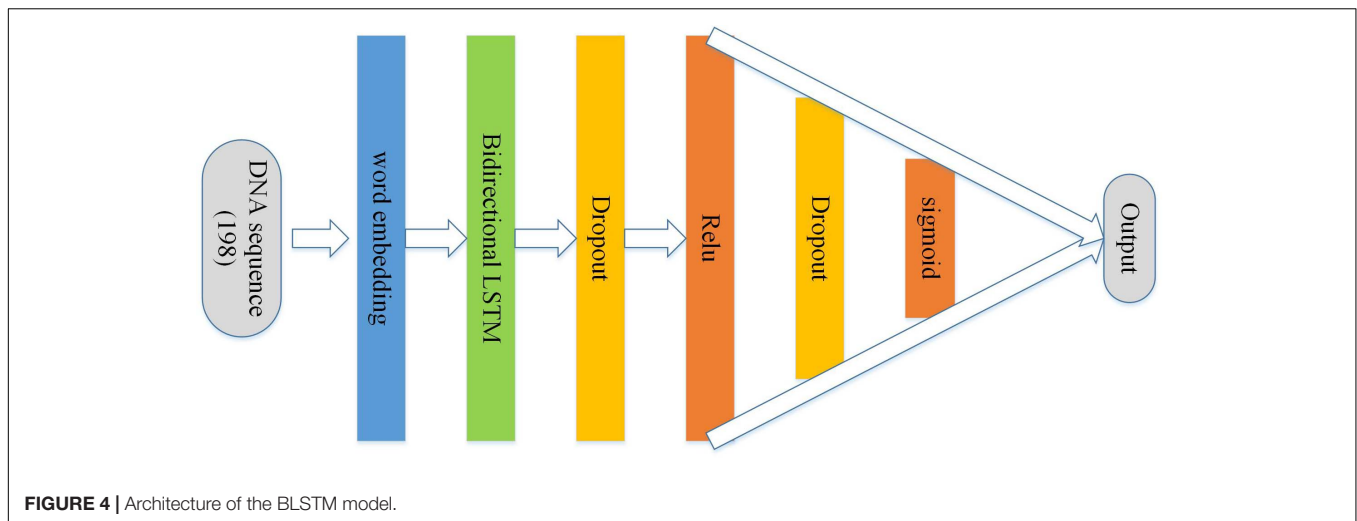
BLSTM Architecture

As **Figure 4** shows, a sequence of numbers with the sequence encoding scheme with the length 198 followed by the body of the structure is used as input to BLSTM. It is mainly composed of an embedding layer, a bidirectional LSTM, a dropout layer, the rectified linear unit (relu), a dropout layer, and sigmoid activation functions. In the architecture, the main purpose of embedding term training is to incorporate into the model to form an end-to-end structure, and the vector trained by the

embedding layer can better adapt to the corresponding tasks (Kleijnjan and Lettice, 2008; Liu G. et al., 2016, 2018; Liu et al., 2017; Zhang T. et al., 2020). The recurrent neural network (RNN) is a network of nerves that processes sequential data. Compared with the ordinary neural network, it can process the sequence variation data (Zou et al., 2016; Zacher et al., 2017). Long short-term memory (LSTM) is a special RNN, and it is mainly used to solve the problem of gradient explosion and disappearance. In short, LSTM performs better than normal RNN if the sequence is long (Liu et al., 2019; Zou et al., 2019; Yang et al., 2020; Zhang Z. Y. et al., 2020). Bidirectional LSTM is equivalent to the LSTM upgraded version, which means that time sequence data are used to input history and future data simultaneously. In contrast to time sequence, two cyclic neural networks are connected to the same output, and the output layer can obtain historical and future information at the same time (Bian et al., 2014; Goldberg and Levy, 2014; Juntao and Zou, unpublished; Tang et al., 2014). The function of dropout layer is preventing model overfitting. In addition, after relu and sigmoid layers (Gers et al., 1999; Graves and Schmidhuber, 2005; Sundermeyer et al., 2012; Zaremba et al., 2014; Huang et al., 2015; Xingjian et al., 2015; Li and Liu, 2020; Sherstinsky, 2020), a probability of whether the sequence is an enhancer or not can be calculated.

Ensemble Model

There are two algorithms in ensemble learning: boosting and bagging (Li et al., 2020; Lv Z. B. et al., 2020; Sultana et al., 2020; Zhu et al., 2020). In our experiment, the data from each experiment are relatively independent and the bagging algorithm



is more suitable. First, the basis learner models are trained independently by using subsamples. Finally, the strong learner model is made by different ensemble methods. The testing result shows that bagging is better than boosting. The entire workflow of bagging is in perfect agreement with our experimental procedure. After that, through several experiments, compared with the voting and median methods, the average method (**Figure 5**) can improve most of the metrics in our experiment in the process of selecting the ensemble method.

In our experiment, the dataset is divided into five parts according to fivefold cross-validation and each part is used as the validation set (Cheng et al., 2019; Dao et al., 2020a; Tang et al., 2020; Zhang D. et al., 2020; Zhao et al., 2020), respectively, and the remaining four parts are used as the training set for the experiment. Five different sets of parameters and models are obtained in these five experiments, and then five sets of models are used to test and obtain the prediction results. The final prediction probability value of the five prediction results is obtained by the average method, and then the prediction results is obtained by comparing with the threshold value of 0.5.

Measurement

To get the performance of the model, some evaluation metrics are used, such as accuracy (ACC), SN, specificity (SP), Matthews's correlation coefficient (MCC), and area under the ROC curve (AUC) (Jiang et al., 2013; Cheng, 2019; Liang et al., 2019; Dao et al., 2020b; Lv H. et al., 2020; Shao and Liu, 2020; Shao et al., 2020; Su et al., 2020; Lv et al., 2021; Zhang et al., 2021). In the formulas of these metrics, TP, TN, FP, and FN mean true positive, true negative, false positive, and false negative, respectively. As we know, ACC is a description of systematic errors, a measure of statistical bias, and it always evaluates a model objectively when the dataset is balanced. SN and SP can support the model more accurately when the data are not balanced. The ROC curve is based on a confounding factors matrix, and the abscissa and the ordinate of the ROC curve are the false positive rate (FPR) and true positive rate (TPR), respectively, and AUC is the area under the curve. When comparing the different classification models,

the ROC curve of each model can be drawn to obtain the value of the AUC, which can be used as an important indicator to evaluate the quality of a model (Gers et al., 1999; Graves and Schmidhuber, 2005; Sundermeyer et al., 2012; Wei et al., 2014, 2017a,b, 2019; Zaremba et al., 2014; Jin et al., 2019; Su et al., 2019; Ao et al., 2020a,b; Li and Liu, 2020; Sherstinsky, 2020; Yu et al., 2020a,b,c). The higher the AUC value is, the better the model is. The MCC is used as a measure of the quality of binary classifications and it is always used in the field of bioinformatics and machine learning. The reason why it is seen as a balanced measure is that MCC can take into account TP, TN, FP, and FN and we can get more ACC results by this way. MCC is a value between +1 and -1. +1 means a perfect prediction, 0 represents that the method does not work, and -1 indicates that the prediction was the exact opposite. These evaluation metrics are calculated from the count of TP, TN, FP, and FN.

$$ACC = \frac{TN + TP}{TP + FN + TN + FP} \quad (6)$$

$$SN = \frac{TP}{TP + FN} \quad (7)$$

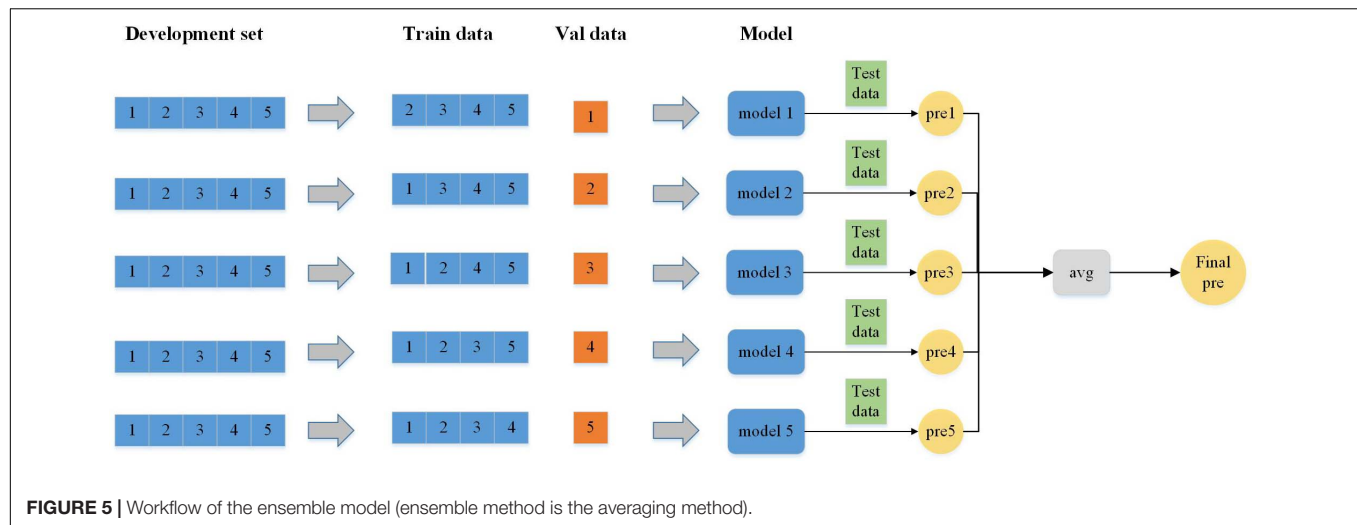
$$SP = \frac{TN}{TN + FP} \quad (8)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (9)$$

RESULTS

Two-Layer Classification Framework

To finish the work in an orderly way, a two-layer classification framework is proposed, which is composed in two steps: identifying enhancer and classifying strong enhancer from weak enhancers. In fact, layers 1 and 2 have the same encoding scheme and network structure. The only difference between the two layers is the input dataset. In layer 1, all data are used as the training set, enhancer set, and non-enhancer set, as part of all data and considered the positive set and negative set, respectively.



In layer 2, only the enhancers are used in the experiment. The strong enhancer and weak enhancer are used, respectively, as the positive set and negative set.

Layer 1: Enhancer Identification

As we know, enhancer identification is extremely important in the field of enhancers. Now it is a hot topic in bioinformatics. In this study, the process of identification can be regarded as preparation for next step. To illustrate it, before judging whether a DNA sequence is a strong enhancer or a weak enhancer, the first thing is to judge if the sequence is an enhancer or not. If it is an enhancer, then the model predicts if it is strong or weak. Through this process, it becomes easier to understand its characteristics. Compared with layer 2 (enhancer classification), layer 1 will have higher ACC. For the reason, there are more differences between enhancer and non-enhancer than strong enhancer and weak enhancer. The more the difference, the easier it is to distinguish. In the process of the experiment, all of the datasets (enhancer + non-enhancer) are divided into five parts. Data division strategy is shown in **Table 1**.

Layer 2: Enhancer Classification

The differences between strong enhancers and weak enhancers are small. Hence, for layer 2, enhancer classification is more difficult than layer 1. Enhancer's biological function and distinguishing the enhancer's strength are an important component in understanding its physical and chemical

properties. For layer 2, more effort is paid in to study it. In this layer, the enhancer dataset (strong + weak) is split into five parts as layer 1, but the amount of enhancer data is smaller (**Table 2**). Compared with layer 1, the layer 2 data are characterized by smaller differences and smaller quantities.

Comparison of Different Encoding Schemes

In the second part of our study, we compared the encoding methods that we introduced the sequence and encoding scheme. The encoding method adopted in this article is to encode the letters in the sequence into the numbers by 3-mer. Meanwhile, several other coding methods have also been tested, such as 2-mer, one-hot, and encoding by correspondence between letters and numbers.

k-Mer is obtained by sliding on the DNA sequence with a step size of 1 bp. In our experiment, take 3-mer ($k = 3$) as an example. When k is 3, 198 3-mers can be extracted from DNA sequence of length 200. Each 3-mer consists of the three letters taken as a whole, so it is possible to encode the original letter sequence into a sequence of numbers of length 198 based on the encoding method shown in **Figure 3**. In addition, the purpose of k-mer is to enhance the relationship between adjacent letters so that the model can learn better. The same is true for 2-mer, except that we end up with a sequence of 199 digits. Another method is to encode the

TABLE 1 | The specific division of the dataset into five parts for identifying enhancers and non-enhancers.

Part	Enhancers	Non-enhancers
1	296	296
2	296	296
3	296	296
4	296	296
5	300	300
Total	1484	1484

TABLE 2 | The specific division of the dataset into five parts for classifying strong enhancers and weak enhancers.

Part	Strong	Weak
1	148	148
2	148	148
3	148	148
4	148	148
5	150	150
Total	742	742

letters directly in the sequence into the corresponding numbers according to the one-to-one correspondence between letters and numbers (A->0, T->1, C->2, G->3). One-hot coding, in fact, means that there are N state registers used to encode N states. Each state has an independent register bit, and only one of these register bits is valid. In other words, there can only be one state. This method ignores the relationship between adjacent sequences.

As shown in **Table 3**, one-hot encoding scheme showed poor effect in every metric. Adjacent sequences are separated in this process and coding these sequences by one-hot into the EBLSTM may not be a good idea. The other three methods have a similar effect by careful observation, and SN of letters to numbers and 3-mer is equal. But in other metrics, 3-mer is undoubtedly the best one. Similarly, as shown in **Table 4**, in the process of enhancer classification, the difference among different encoding schemes will be more obvious. It can be seen that 3-mer performs better than the others for each item; thus, we think 3-mer is a more suitable encoding method for this experiment.

TABLE 3 | Result of comparison of using different encoding schemes in layer 1 (enhancers identification) under 10 trials.

Encoding scheme	ACC	AUC	SN	SP	MCC
Letters to numbers	0.753	0.824	0.755	0.750	0.500
One-hot	0.565	0.611	0.494	0.642	0.132
2-Mer	0.758	0.827	0.735	0.762	0.505
3-Mer	0.772	0.835	0.755	0.795	0.534

TABLE 4 | Result of comparison of using different encoding schemes in layer 2 (enhancers classification) under 10 trials.

Encoding scheme	ACC	AUC	SN	SP	MCC
Letters to numbers	0.640	0.650	0.784	0.512	0.302
One-hot	0.526	0.522	0.438	0.412	0.116
2-Mer	0.645	0.662	0.786	0.498	0.304
3-Mer	0.658	0.688	0.812	0.536	0.324

TABLE 5 | Result of comparison of using different architectures in layer 1 (enhancers identification) under 10 trials.

Architecture type	ACC	AUC	SN	SP	MCC
Simple RNN	0.721	0.791	0.732	0.760	0.488
Bidirectional RNN	0.745	0.801	0.767	0.751	0.492
Simple LSTM	0.742	0.812	0.802	0.746	0.512
Bidirectional LSTM	0.772	0.835	0.755	0.795	0.534

TABLE 6 | Result of comparison of using different architectures in layer 2 (enhancers classification) under 10 trials.

Architecture type	ACC	AUC	SN	SP	MCC
Simple RNN	0.617	0.634	0.801	0.591	0.249
Bidirectional RNN	0.628	0.617	0.792	0.612	0.276
Simple LSTM	0.634	0.626	0.770	0.578	0.302
Bidirectional LSTM	0.658	0.688	0.812	0.536	0.324

Comparison of Different Architectures

In this experiment, we tried eight different internal structures, including simple RNN, bidirectional RNN, simple LSTM, and bidirectional LSTM, and then, on the basis of the four networks doubled, respectively, which means that another four structures are two layers of RNNs, bidirectional RNNs, simple LSTMs, and bidirectional LSTMs. After this step, a model that has the best performance would be chosen that with higher metrics than other seven models. Then the dropout layer is added to produce the final architecture.

Tables 5, 6 show the different architecture results in layers 1 and 2, respectively. The results are shown from the results in **Table 5**. Except for SN, the bidirectional LSTM has better effect based on the four other evaluation metrics. The reasons may be that bidirectional LSTM is more complex than the other three architectures and more features can be captured by it. In fact, we also do the other four experiments, as mentioned in the previous paragraph. But increasing the number of layers in this

TABLE 7 | Result of comparison of using different ensemble models in layer 1 (enhancers identification) under 10 trials.

Ensemble method	ACC	AUC	SN	SP	MCC
Median	0.728	0.788	0.774	0.726	0.498
Voting	0.765	0.762	0.792	0.738	0.517
Averaging	0.772	0.835	0.755	0.795	0.534

TABLE 8 | Result of comparison of using different ensemble models in layer 2 (enhancers classification) under 10 trials.

Ensemble method	ACC	AUC	SN	SP	MCC
Median	0.622	0.664	0.740	0.572	0.310
Voting	0.638	0.644	0.794	0.562	0.311
Averaging	0.658	0.688	0.812	0.536	0.324

TABLE 9 | Result of comparison with existing state-of-the-art methods in layer 1 (enhancers identification).

Method	ACC	AUC	SN	SP	MCC	Source
iEnhancer-2L	0.730	0.806	0.710	0.750	0.460	Liu B. et al., 2016
EnhancerPred	0.740	0.801	0.735	0.745	0.480	Jia and He, 2016
iEnhancer-EL	0.748	0.817	0.710	0.785	0.496	Liu B. et al., 2016; Liu G. et al., 2018
iEnhancer-ECNN	0.769	0.832	0.785	0.752	0.537	Nguyen et al., 2019
iEnhancer-EBLSTM	0.772	0.835	0.755	0.795	0.534	This study

TABLE 10 | Result of comparison with existing state-of-the-art methods in layer 2 (enhancers classification).

Method	ACC	AUC	SN	SP	MCC	Source
iEnhancer-2L	0.605	0.668	0.470	0.740	0.218	Liu G. et al., 2016
EnhancerPred	0.550	0.579	0.450	0.650	0.102	Jia and He, 2016
iEnhancer-EL	0.610	0.680	0.540	0.680	0.222	Liu G. et al., 2018
iEnhancer-ECNN	0.678	0.748	0.791	0.564	0.368	Nguyen et al., 2019
iEnhancer-EBLSTM	0.658	0.688	0.812	0.536	0.324	This study

architecture also raises the processing time longer. The efficiency will be reduced. Therefore, the results of these four experiments were added to the table. A similar situation occurs in **Table 6**, where bidirectional LSTM is also the better choice in many metrics, except SP. Together, these results provide important insights into the idea that bidirectional LSTM is the best fit for the experiment.

Comparison of Different Ensemble Models

As mentioned in Section “Ensemble Model,” during the experiment, we tested three ensemble strategies. Each method has advantages and disadvantages. To explore which kind of strategy is more suitable for enhancers DNA sequences characteristics identification, median, voting, and averaging are tested. Set of indicators across the different methods are assessed. In **Table 7**, the voting and averaging methods are significantly better than the median method, and their performance of the two is very similar, but AUC and MCC in the averaging method are higher than those in the voting method, which shows that the predictive effect and stability of the average method are more advantageous than those of the voting method. In addition, in **Table 8**, the averaging method is still the best of these three ensemble methods. Combining these two tables to draw a conclusion, the indicators for the averaging method are better than the other two methods. The averaging method is the best one, and finally in our model, this method is applied to achieve ensemble learning.

Comparison With Existing State-of-the-Art Methods

There are several excellent methods for the prediction of enhancers, and the well-known methods are iEnhancer-2L, EnhancerPred, iEnhancer-EL, and iEnhancer-ECNN. **Tables 9, 10** show the results of the comparison with existing state-of-the-art methods in layers 1 and 2.

As **Table 9** shows, compared with the previous three experimental methods, all the results of the metrics are significantly improved, especially in AUC and MCC. Moreover, compared with iEnhancer-ECNN in 2019, in this study, the results for ACC, AUC, and SP are slightly higher, but the results for SN and MCC are slightly lower. As seen in **Table 10**, iEnhancer-EBLSTM remains a reliable method that has better performance than iEnhancer-2L, iEnhancer-EL, and EnhancerPred, especially for SN and MCC; this method has been greatly improved. From the experimental results, we can see that both iEnhancer-EBLSTM and iEnhancer-ECNN are significantly better than the previous methods. I think the reason lies in the fact that the deep learning model itself has certain advantages, which can capture features more accurately and learn more efficiently. The model obtained can have more accurate

parameters, so as to obtain higher results. However, compared with iEnhancer-ECNN, the data for AUC in our experiment are lower than the result of them, but the data for SN are higher. Overall, these results indicate that iEnhancer-EBLSTM performs best in enhancer identification and classification.

DISCUSSION

In this paper, we proposed the prediction model called iEnhancer-EBLSTM to identify enhancers and their strengths. In addition, this model uses the principle of 3-mer to encode the DNA sequence and EBLSTM to get the predictive result. The biggest advantage of this method is that it only uses DNA sequence information and does not rely on other features such as chromosome status, gene expression data, and histone modification. This greatly facilitates researchers to use it. iEnhancer-EBLSTM could be used not only for identifying enhancers but also for distinguishing strong enhancers from weak enhancers. In the first layer, the predictor can identify whether the DNA sequence is enhancer or not, and the ACC is 0.772. In the second layer, the predictor can classify that the enhancer is strong or weak, and the ACC is 0.658. A lot of work still needs to be done on the second layer. There is little difference between strong and weak enhancers. More and more information of DNA sequences, physical and chemical needs to be mined. The characteristic information can be recorded more completely, and the various models can be built based on this information in more detail.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

YZ conceived and designed the project. KN and XL conducted the experiments and analyzed the data. KN and YZ wrote the manuscript. TZ and YZ revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61971119 and 61901103) and the Natural Science Foundation of Heilongjiang Province (Grant No. LH2019F002).

REFERENCES

- Ao, C., Yu, L., and Zou, Q. (2020a). Prediction of bio-sequence modifications and the associations with diseases. *Briefin. Funct. Genom.* 20:1201.
- Ao, C., Zhou, W., Gao, L., Dong, B., and Yu, L. (2020b). Prediction of antioxidant proteins using hybrid feature representation method and random forest. *Genomics* 89, 256–178.
- Bian, J., Gao, B., and Liu, T.-Y. (2014). “Knowledge-powered deep learning for word embedding, in Joint European conference on machine learning and knowledge discovery in databases,” in *Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2014. Lecture Notes in Computer Science*,

- Vol. 8724, eds T. Calders, F. Esposito, E. Hüllermeier, and R. Meo (Berlin: Springer).
- Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/156652321904191022113307
- Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational Methods for identifying similar diseases. *Mol. Ther. Nucleic Acids Res.* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019
- Dao, F. Y., Lv, H., Yang, Y. H., Zulfikar, H., Gao, H., and Lin, H. (2020a). Computational identification of N6-methyladenosine sites in multiple tissues of mammals. *Comput. Struct. Biotechnol. J.* 18, 1084–1091. doi: 10.1016/j.csbj.2020.04.015
- Dao, F. Y., Lv, H., Zulfikar, H., Yang, H., Su, W., Gao, H., et al. (2020b). A computational platform to identify origins of replication sites in eukaryotes. *Brief. Bioinform.* 17:bbaa017.
- Gers, F. A., Schmidhuber, J., and Cummins, F. (1999). *Learning to Forget: Continual Prediction with LSTM*.
- Goldberg, Y., and Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv [Preprint]*. arXiv:1402.3722.
- Graves, A., and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neur. Netw.* 18, 602–610. doi: 10.1016/j.neunet.2005.06.042
- Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv [Preprint]*. arXiv:1508.01991.
- Jia, C., and He, W. (2016). EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features. *Sci. Rep.* 6, 1–7.
- Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* 8, 282–293. doi: 10.1504/ijdm.2013.056078
- Jin, Q., Mengad, Z., Phamb, T. D., Chena, Q., Weic, L., and Sua, R. (2019). DUNet: A deformable network for retinal vessel segmentation. *Knowl. Based Syst.* 178, 149–162.
- Kleinjan, D. A., and Lettice, L. A. (2008). Long-range gene control and genetic disease. *Adv. Genet.* 61, 339–388. doi: 10.1016/s0065-2660(07)00013-2
- Krivega, I., and Dean, A. (2012). Enhancer and promoter interactions—long distance calls. *Curr. Opin. Genet. Dev.* 22, 79–85. doi: 10.1016/j.gde.2011.11.001
- Li, C.-C., and Liu, B. (2020). MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. *Brief. Bioinform.* 21, 2133–2141. doi: 10.1093/bib/bbz133
- Li, J., Wei, L., Guo, F., and Zou, Q. (2020). EP3: an ensemble predictor that accurately identifies type III secreted effectors. *Brief. Bioinform.* doi: 10.1093/bib/bbaa008
- Liang, C., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2019). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48:7603.
- 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, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* 47:e127. doi: 10.1093/nar/gkz740
- Liu, B., Li, K., Huang, D. S., Chou, K. C., and Enhancer, E. L. (2018). Identifying enhancers and their strength with ensemble learning approach. *Bioinformatics* 34, 3835–3842. doi: 10.1093/bioinformatics/bty458
- Liu, G., Hu, Y., Jin, S., and Jiang, Q. (2017). Genetic variant rs763361 regulates multiple sclerosis CD226 gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 114, E906–E907.
- Liu, G., Hu, Y., Jin, S., Zhang, F., Jiang, Q., and Hao, J. (2016). Cis-eQTLs regulate reduced LST1 gene and NCR3 gene expression and contribute to increased autoimmune disease risk. *Proc. Natl. Acad. Sci. U.S.A.* 113, E6321–E6322.
- Liu, G., Zhang, Y., Wang, L., Xu, J., Chen, X., Bao, Y., et al. (2018). Alzheimer's disease rs11767557 variant regulates EPHA1 gene expression specifically in human whole blood. *J. Alzheimers Dis.* 61, 1077–1088. doi: 10.3233/jad-170468
- Lv, H., Dao, F. Y., Guan, Z.-X., Yang, H., Y-W, Li, and Lin, H. (2020). Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. *Brief. Bioinform.* 2:bbaa255.
- Ly, H., Dao, F. Y., Zulfikar, H., Su, W., Ding, H., Liu, L., et al. (2021). A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. *Brief. Bioinform.* 21:bbab031. doi: 10.1093/bib/bbab031
- Ly, Z. B., Wang, D. H., Ding, H., Zhong, B. N., and Xu, L. (2020). *Escherichia Coli* DNA N-4-methylcytosine site prediction accuracy improved by light gradient boosting machine feature selection technology. *IEEE Access* 8, 14851–14859. doi: 10.1109/access.2020.2966576
- Nguyen, Q. H., Nguyen-Vo, T.-H., Le, N. Q. K., Do, T. T. T., Rahardja, S., and Nguyen, B. P. (2019). iEnhancer-ECNN: identifying enhancers and their strength using ensembles of convolutional neural networks. *BMC Genom.* 20:951.
- Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: five essential questions. *Nat. Rev. Genet.* 14, 288–295. doi: 10.1038/nrg3458
- Sen, R., and Baltimore, D. (1986). Multiple nuclear factors interact with the immunoglobulin enhancer sequences. *Cell* 46, 705–716. doi: 10.1016/0092-8674(86)90346-6
- Shao, J., and Liu, B. (2020). ProtFold-DFG: protein fold recognition by combining Directed Fusion Graph and PageRank algorithm. *Brief. Bioinform.* 2:bbaa192. doi: 10.1093/bib/bbaa192
- Shao, J., Yan, K., and Liu, B. (2020). FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network. *Brief. Bioinform.* 2:bbaa144. doi: 10.1093/bib/bbaa144
- Shen, Z., and Zou, Q. (2020). Basic polar and hydrophobic properties are the main characteristics that affect the binding of transcription factors to methylation sites. *Bioinformatics* 36, 4263–4268. doi: 10.1093/bioinformatics/btaa492
- Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Phys. D Nonlin. Phenom.* 404:132306. doi: 10.1016/j.physd.2019.132306
- Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2019). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE Acm. Transact. Comput. Biol. Bioinform.* 16, 1231–1239. doi: 10.1109/tcbb.2018.2858756
- Su, W., Wang, F., Tan, J. X., Dao, F. Y., Yang, H., and Ding, H. (2020). The prediction of human DNase I hypersensitive sites based on DNA sequence information. *Chemometr. Intel. Labor. Syst.* 209:104223. doi: 10.1016/j.chemolab.2020.104223
- Sultana, N., Sharma, N., Sharma, K. P., and Verma, S. A. (2020). Sequential ensemble model for communicable disease forecasting. *Curr. Bioinform.* 15, 309–317. doi: 10.2174/1574893614666191202153824
- Sundermeyer, M., Schlüter, R., and Ney, H. (2012). “LSTM neural networks for language modeling” in *Proceedings of the Thirteenth Annual Conference of the International Speech Communication Association*.
- Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B., et al. (2014). “Learning sentiment-specific word embedding for twitter sentiment classification,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1, Baltimore, MD.
- Tang, Y.-J., Pang, Y.-H., and Liu, B. (2020). IDP-Seq2Seq: identification of intrinsically disordered regions based on sequence to sequence learning. *Bioinformatics* 23:btba667. doi: 10.1093/bioinformatics/bta a667
- Wang, H., Liu, Y., Guan, H., and Fan, G.-L. (2020). The regulation of target genes by co-occupancy of transcription factors, c-Myc and Mxi1 with max in the mouse cell line. *Curr. Bioinform.* 15, 581–588. doi: 10.2174/1574893614666191106103633
- Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set. *IEEE ACM Transact. Comput. Biol. Bioinform.* 11, 192–201. doi: 10.1109/tcbb.2013.146
- Wei, L., Ranbc, S., Bingd, W., Xiuting, L., Quana, Z., and Gaof, X. (2019). Integration of deep feature representations and handcrafted features to improve the prediction of N-6-methyladenosine sites. *Neurocomputing* 324, 3–9. doi: 10.1016/j.neucom.2018.04.082
- Wei, L., Tang, J., and Zou, Q. (2017a). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026
- Wei, L., Xing, P., Zeng, J., Xiu, J., Ran, C., and Guo, S. F. (2017b). Improved prediction of protein-protein interactions using novel negative samples,

- features, and an ensemble classifier. *Artif. Intel. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001
- Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W. -k., and Woo, W. -c. (2015). *Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. Advances in Neural Information Processing Systems.*
- Yang, H., Yang, W., Dao, F. Y., Lv, H., Ding, H., Chen, W., et al. (2020). A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*. *Brief Bioinform.* 21, 1568–1580. doi: 10.1093/bib/bbz123
- Yu, L., Xu, F., and Gao, L. (2020a). Predict New therapeutic drugs for hepatocellular carcinoma based on gene mutation and expression. *Front. Bioeng. Biotechnol.* 8:8.
- Yu, L., Zhoua, D., Gaoa, L., and Zhab, Y. (2020b). Prediction of drug response in multilayer networks based on fusion of multiomics data. *Methods (San Diego CA)* 5:256.
- Yu, L., Zou, Y., Wang, Q., Zheng, S., and Gao, L. (2020c). Exploring drug treatment patterns based on the action of drug and multilayer network model. *Int. J. Mol. Sci.* 21:5014. doi: 10.3390/ijms21145014
- Zacher, B., Michel, M., Schwalb, B., Cramer, P., Tresch, A., Gagneur, J., et al. (2017). Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. *PLoS One* 12:e0169249. doi: 10.1371/journal.pone.0169249
- Zaremba, W., Sutskever, I., and Vinyals, O. (2014). Recurrent neural network regularization. *arXiv [Preprint]*. arXiv:1409.2329.
- Zhang, D., Chen, H. D., Zulfiqar, H., Yuan, S. S., Huang, Q. L., Zhang, Z. Y., et al. (2021). iBLP: a XGBoost-based predictor for identifying bioluminescent proteins. *Comput. Mathemat. Methods Med.* 2021: 6664362.
- Zhang, D., Xu, Z. C., Su, W., Yang, Y. H., Yang, H., and Lin, H. (2020). iCarPS: a computational tool for identifying protein carbonylation sites by novel encoded features. *Bioinformatics* 7:btaa702.
- Zhang, T., Wang, R., Jiang, Q., and Wang, Y. (2020). An information gain-based method for evaluating the classification power of features towards identifying enhancers. *Curr. Bioinform.* 15, 574–580. doi: 10.2174/1574893614666191120141032
- Zhang, Z. Y., Yang, Y.-H., Ding, H., Wang, D., Chen, W., and Lin, H. (2020). Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief Bioinform.* 22, 526–535. doi: 10.1093/bib/bbz177
- Zhao, T., Hu, Y., Peng, J., Cheng, L., and Martelli, P. L. (2020). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428
- Zhu, H., Du, X., and Yao, Y. (2020). ConvsPPIS: identifying protein-protein interaction sites by an ensemble convolutional neural network with feature graph. *Curr. Bioinform.* 15, 368–378. doi: 10.2174/1574893614666191105155713
- Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* 10:114.
- Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

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.

Copyright © 2021 Niu, Luo, Zhang, Teng, Zhang and Zhao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



A Novel Biomarker Identification Approach for Gastric Cancer Using Gene Expression and DNA Methylation Dataset

Ge Zhang, Zijing Xue, Chaokun Yan*, Jianlin Wang* and Huimin Luo

School of Computer and Information Engineering, Henan University, Kaifeng, China

OPEN ACCESS

Edited by:

Wang Guohua,
Harbin Institute of Technology, China

Reviewed by:

Wei Lan,
Guangxi University, China
Xiulong Liu,
Tianjin University, China

*Correspondence:

Chaokun Yan
ckyan@henu.edu.cn
Jianlin Wang
jlwang@henu.edu.cn

Specialty section:

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

Received: 21 December 2020

Accepted: 16 February 2021

Published: 25 March 2021

Citation:

Zhang G, Xue Z, Yan C, Wang J and
Luo H (2021) A Novel Biomarker
Identification Approach for Gastric
Cancer Using Gene Expression and
DNA Methylation Dataset.
Front. Genet. 12:644378.
doi: 10.3389/fgene.2021.644378

As one type of complex disease, gastric cancer has high mortality rate, and there are few effective treatments for patients in advanced stage. With the development of biological technology, a large amount of multiple-omics data of gastric cancer are generated, which enables computational method to discover potential biomarkers of gastric cancer. That will be very important to detect gastric cancer at earlier stages and thus assist in providing timely treatment. However, most of biological data have the characteristics of high dimension and low sample size. It is hard to process directly without feature selection. Besides, only using some omic data, such as gene expression data, provides limited evidence to investigate gastric cancer associated biomarkers. In this research, gene expression data and DNA methylation data are integrated to analyze gastric cancer, and a feature selection approach is proposed to identify the possible biomarkers of gastric cancer. After the original data are pre-processed, the mutual information (MI) is applied to select some top genes. Then, fold change (FC) and *T*-test are adopted to identify differentially expressed genes (DEG). In particular, false discover rate (FDR) is introduced to revise *p*-value to further screen genes. For chosen genes, a deep neural network (DNN) model is utilized as the classifier to measure the quality of classification. The experimental results show that the approach can achieve superior performance in terms of accuracy and other metrics. Biological analysis for chosen genes further validates the effectiveness of the approach.

Keywords: gastric cancer, omics data, biomarkers, feature selection, deep neural network, machine learning

1. INTRODUCTION

Gastric cancer is one of the most common malignant tumors of the digestive system (Nogueira et al., 2017). The pathogenesis is mainly relevant to helicobacter pylori infection, diet, environment, and genetic factors. It remains one of the most deadly cancers worldwide, especially among older males (Siegel et al., 2020). Generally speaking, early detection of cancer is crucial for increasing the chances for successful treatment and prolonging the patient's life. The 5-year survival rate of early-stage gastric cancer can reach more than 95% (Song et al., 2017). However, the early stage of gastric cancer is hard to monitor because of rare symptoms and some potential patients' cancer may be advanced when they are first diagnosed. Therefore, early targeting and treatment are very important in clinical practice of gastric cancer (Wang et al., 2020). In recent years, with the

development of sequencing technology, the genome data of cancer patients can be obtained easily. These genomic data have been used to study the association between genetic changes and diseases and contribute to diagnosis and prognosis. However, these data always have the characteristics of high dimensions and low sample size (HDLSS) (Han et al., 2019). It is hard to process these data directly (Yan et al., 2018). Therefore, feature selection technology is usually adopted to assist in analyzing the possible cancer-causing genes, also called biomarkers, from massive cancer data. The biomarkers can facilitate us to understand the pathogenesis of diseases at a detailed molecular level and play an auxiliary role in clinical diagnosis.

Till now, many researchers have applied the feature selection methods to the field of gene expression data analysis (Ding and Peng, 2005; Lu et al., 2017; Zhao et al., 2020). However, it is incomprehensive to analyze cancer only using gene expression data. The rapid accumulation of omics data can provide disparate, partially independent, and complementary information about the entire genome (Zhang et al., 2016). The multi-omic data can lay an important foundation for mining informative biomarkers for cancer (Ruffalo et al., 2015). Among these omics data, DNA methylation is an important epigenetic event that affects gene expression during the development in various diseases such as cancer (Bird, 1986; Wang et al., 2018). In general, DNA methylation status is more reliable than gene expression (Paziewska et al., 2014). The combination of DNA methylation data and gene expression data is more beneficial to explain the pathogenesis of gastric cancer. Therefore, these two kinds of data are utilized to identify the biomarkers of gastric cancer in our study.

In this paper, we propose a novel gastric cancer biomarker identification approach, referred to GCBMI, to discover the possible biomarkers of gastric cancer. First, the gene expression data and DNA methylation data of gastric cancer are collected and processed. Then, fold change, statistical test, and mutual information are utilized to identify the differentially expressed genes of gastric cancer and the selected genes can serve as guidelines to reduce the dimension of omics data. At last, the DNN model is adopted as the classifier to measure the quality of classification. Experimental results indicate that GCBMI can obtain more favorable performance than other state-of-art methods.

The main contributions of this study are summarized as follows:

- For gastric cancer, a novel feature selection approach is proposed to identify the potential biomarkers. Here, DNA methylation data is integrated with the gene expression data effectively to obtain a comprehensive analysis to discover the relationship between gastric cancer and potential biomarkers.
- Besides *T*-test and FC, mutual information is introduced as a preliminary screening method to filter out redundant genes and FDR is adopted to revise *p*-value to further screen genes.
- The experimental results suggest that our approach can achieve improvement in different evaluation indicators than other state-of-art methods. In addition to evaluating accuracy, GO analysis, heatmap, and literature review are executed.

The above biological validation is able to demonstrate that the genes selected by our approach are associated with gastric cancer.

The remainder of this paper is organized as follows: In section 2, we review related works of feature selection methods. The proposed approach is introduced in section 3. section 4 introduces the experimental design. Experimental results and biological analysis are described in section 5. Finally, we summarize the paper and make a vision for the future in section 6.

2. RELATED WORK

With the development of sequencing technology, massive amounts of cancer genome data have been accumulated at an accelerated speed. A number of feature selection methods have been extensively applied to cancer data. Traditional feature selection methods can be divided into two categories: filter methods and wrapper methods. Among them, the filter method has the advantage of low time consumption. So far, some filter methods had been well-applied to gene expression data.

Principal Component Analysis (PCA) is an effective dimensionality reduction method (Wold et al., 1987). Ding et al. combined feature extraction with feature selection in gene expression data (Ding et al., 2009). The relief was utilized to feature selection, and PCA was used to extract features. Then, they used the support vector machines (SVM) for classification. Experimental results illustrated that their method is effective to reduce the classification error rate in eight cancer datasets. But such methods cannot guarantee that the features still remain the corresponding biological significance. For example, the dimensionality reduction of features by PCA is equivalent to mapping the new features on the original features, and the features obtained after PCA are different from the original genes (Shen and Huang, 2008). Thus, it is often difficult to interpret the results.

Hsu et al. used extremely randomized trees (ET) to calculate the weight of the features (Hsu and Si, 2018). Feature selection was achieved by selecting features with high weight. Then, the linear SVM was combined to achieve about 95% accuracy on TCGA datasets. Lee et al. developed a novel filter method to identify the biomarkers of lung cancer and confirmed seven possible biomarkers (Lee et al., 2011).

In addition to filter methods, the wrapper methods utilize classification accuracy as a measurement standard for evaluation and find the optimal feature subset by iteration of meta-heuristic algorithms (Rodrigues et al., 2014). A lot of meta-heuristic algorithms had been well-applied to wrapper methods for feature selection of cancer such as bat algorithm (BA), recursive memetic algorithm (RMA), binary krill herd algorithm (MBKH), and so on (Dashtban et al., 2018; Ghosh et al., 2019; Zhang et al., 2020).

Dashtban et al. proposed MOBBA-LS which utilized fisher criterion and BA (Dashtban et al., 2018). They tested their method on three microarray cancer datasets. The accuracy achieved 100, 97, and 100% on leukemia, prostate, and SRBCT datasets, respectively. Ghosh et al. developed a recursive memetic

algorithm (RMA) model for feature selection (Ghosh et al., 2019), and Zhang et al. proposed a pre-screening method of feature ranking, IG-MBKH, which is based on information gain (IG) and an improved binary krill herd (MBKH) (Zhang et al., 2020). The above methods can obtain favorable classification accuracy on microarray data of cancer.

Multiple-omics data can enable to provide a more comprehensive analysis of the entire genome. Among them, DNA methylation is one of the important epigenetic regulatory mechanisms (Luo et al., 2020). Especially, it is considered as a molecular factor that controls and regulates gene expression levels near the CpG sites. Its status is closely associated with diverse diseases and is generally more stable than gene expression (Ding et al., 2019). Therefore, the function of DNA methylation data was widely recognized. Increasing feature selection methods, which are based on gene expression data and DNA methylation data, were proposed.

For Alzheimer's disease, Park et al. proposed a biomarker prediction model, which integrated multi-omic data (Park et al., 2020). They used the Limma package to select possible biomarkers. Experimental results showed that their method can achieve better accuracy than using single data, and some chosen genes were reported in AlzGene database.

Mallik et al. proposed a method to identify biomarkers of cancer based on omics data (Mallik et al., 2017). The maximal relevance and minimal redundancy (mRMR) and parameter test like *T*-test were used to select the genes. The results suggested that their method had stable performance on different classifiers and classification accuracy can achieve about 95 and 90% in gene expression data and DNA methylation data, respectively.

Wang et al. proposed a feature selection method based on gene expression data and DNA methylation data of the six types of cancer (Wang et al., 2020). Their method can be divided into three steps. First, the correlation between gene expression profile and methylation profile of each gene was calculated to screen genes initially. Then, the genes were further filtered by *T*-test and FDR value. Finally, the genes selected in first two steps are filtered by Elastic Net. Finally, support vector machine was utilized as the classifier. The accuracy can be as high as 98% for the training set and 97% for the independent test set.

3. THE PROPOSED APPROACH

In this section, the proposed approach GCBMI is introduced. The overall workflow of GCBMI is shown in **Figure 1**. GCBMI consists of three stages: data pre-processing, selection of DEG and data combination, and using deep neural network as the classifier.

3.1. Data Pre-processing

In this section, we regularize the gene expression data, and then merge the individual gene expression data files. In addition, on the basis of annotation file of the gene chip, the column (feature) name of each sample is converted to the gene name, and the label column is added. In the annotation file of the gene chip, the gene name corresponding to each probe is stored. If a gene corresponds to multiple probes, we take the median of expression value as new expression value of the gene. After

that, the genes with null values are further removed. In order to eliminate the influence of outliers, the dataset is standardized by z-score according to the following formula (Zhang et al., 2014). Finally, the datasets are divided into training set and test set in our experiment.

$$x' = \frac{x - \bar{x}}{\sigma} \quad (1)$$

where x and x' represent a column of data before and after standardization. \bar{x} and σ represent the mean and standard deviation of a column of data in training set.

Likewise, DNA methylation data are also processed accordingly to eliminate the influence of outliers.

3.2. Selection of Differentially Expressed Genes and Data Combination

In this section, how to identify DEG in our approach is introduced. For gene expression data, the characteristics of high dimension and low sample size make it hard to construct a prediction model directly and may lead to the over-fitting (Ma and Zhang, 2019). For this issue, an appropriate method is required to reduce the size of feature space and the risk of over-fitting.

In GCBMI, the DEG and the differentially methylated positions (DMP) are utilized to train the model. The overall process contains three steps as follows.

First, MI (Liu H. et al., 2009) is applied to select TopN genes for gene expression data and DNA methylation data, respectively. It is a classic filter method of feature selection, which has been successfully applied to many feature selection problems (Peng and Fan, 2017). In order to avoid redundancy, the MI is adopted to filter out irrelevant genes. N is set to 3,000 through the subsequent experiments.

Second, FC and *T*-test are adopted to do identify DEG and DMP. What is more, the FDR is applied to revise the p -value. Taking DEG as an example, FC value for each selected genes in the first step is calculated. Since the data obey the normally distributed by Z-score standardization. Parametric statistics like *T*-test can work well on this kind of data. Then, Levene-test (Ankarali et al., 2009) is applied to verify whether the samples with variance homogeneity or not. If they have variance homogeneity, performing the standard *T*-test (Gauvreau and Pagano, 1993) to calculate p -value. Otherwise, the Welch's *T*-test (Algina et al., 1994) is executed to calculate the p -value. After that, the FC value and significant p -value for each gene are obtained. Finally, FDR is utilized to revise p -value to further screen candidate genes. A suitable threshold for FC value, p -value, and FDR are set to filter genes. And then we can obtain DEG. Similarly, DMP can be obtained. As shown in **Figure 1**, in gene expression data, the $|FC| > 2.1$ and $p\text{-value} < 0.05$. The $|FC| > 1.8$ and $p < 0.05$ in DNA methylation data. The FDR threshold value of both experimental datasets is set as 0.01. A hypothesis is made that if the gene is differentially expressed and occur hypermethylated and hypomethylated in different samples. This gene may have a potential relationship with gastric cancer. So the overlapping genes in DEG and DMP are the possible biomarkers of gastric cancer.

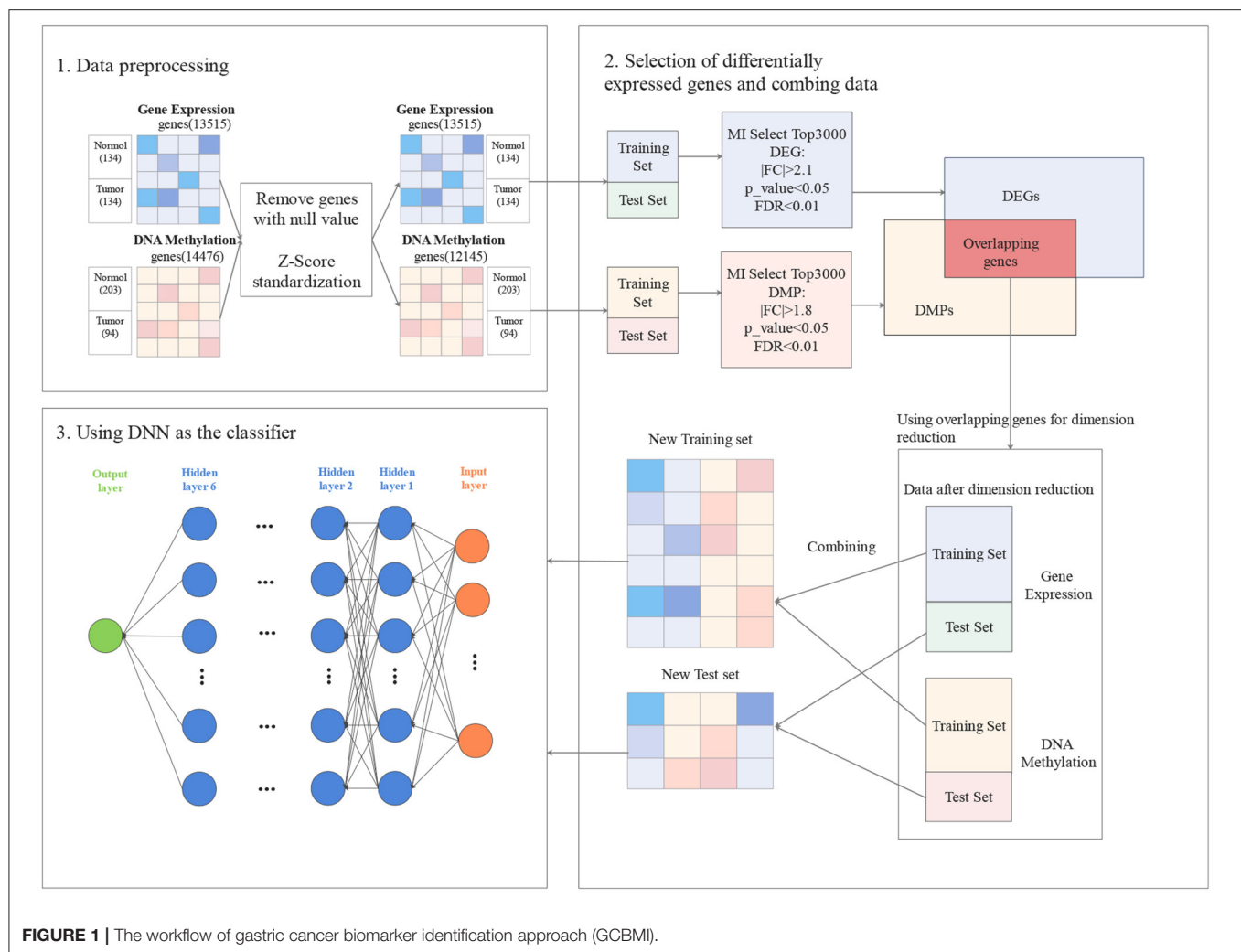


FIGURE 1 | The workflow of gastric cancer biomarker identification approach (GCBMI).

Finally, in order to extend training samples, all possible pairs of gene expression data and DNA methylation data for tumor and normal samples are utilized to merge into a new dataset. As shown in **Figure 2**, Cartesian product (Emelyanov and Ponomaryov, 2017) is performed on the gene expression data and DNA methylation data. The gene expression data and methylation data that labeled as tumor are combined into new tumor samples, and which labeled as normal are combined into new normal samples. In this way, the gene expression matrix and DNA methylation matrix are combined into a new expression matrix. This matrix has a large sample size. For example, in one of the cross-validation, the training set of gene expression data has 214 samples, which contains 112 tumor samples and 102 normal samples. DNA methylation data have 237 samples, which contains 160 tumor samples and 77 normal samples. After the combination, we will obtain 17,920 tumor samples and 7,854 normal samples. Taking them as new tumor samples and normal samples, so the new training set contains 25,774 samples, including 17,920 tumor samples and 7,854 normal samples.

3.3. Using Deep Neural Network as the Classifier

DNN model has excellent classification performance compared with traditional classifiers in previous studies, such as (Chen et al., 2020; Singh and Yamada, 2020). Here, the DNN also adopted as the classifier and the parameters of the DNN are determined through experiments.

In this section, the structure of the network is introduced. Our DNN model consists of three parts: input layer, hidden layer, and output layer. The input layer consists of two parts, corresponding to gene expression data and DNA methylation data, respectively. Then we add six hidden layers that applied ReLU as the activation function. Each layer contains 100 nodes and a additional bias nodes. The dropout is added for each hidden layer to avoid overfitting, which refers to drop some neurons randomly according to a certain probability during the learning iteration. It is equivalent to train a sparser network than the original network. Each of iterations is training a different network model to prevent overfitting. Finally, since our data only have two categories, the output layer with one node is sufficient. Sigmoid

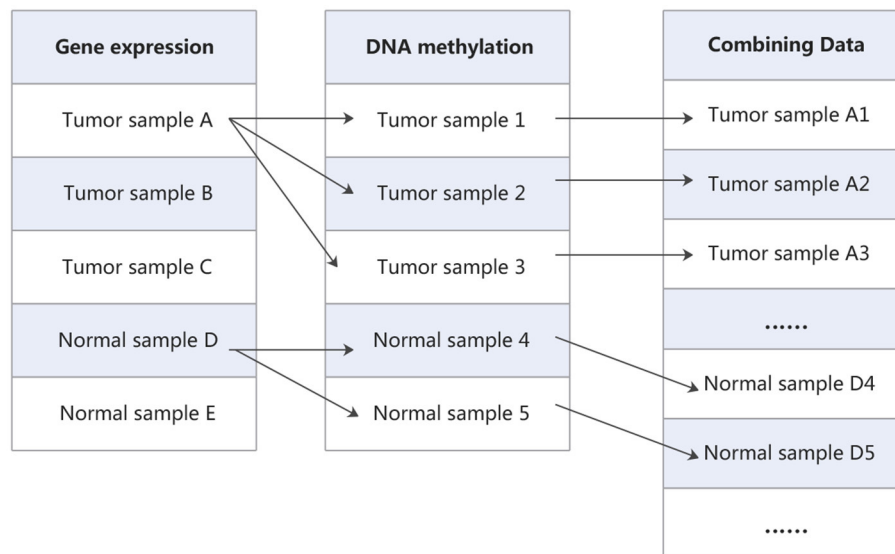


FIGURE 2 | The process of combining data.

function is adopted as the activation function of the output layer to make the output value between 0 and 1.

In the DNN model, the loss function is binary cross entropy and cost function is the reduced average value of cross entropy. Adam algorithm is applied to optimize the parameters of the network model. The formula of the loss function and cost function are as follows:

$$L(\hat{y}, y) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}) \quad (2)$$

$$J(w, b) = \frac{1}{m} \sum_{i=1}^m (-y^i \log(\hat{y}^i) - (1 - y^i) \log(1 - \hat{y}^i)) \quad (3)$$

where y and \hat{y} represent the true value and the predicted value of a sample. \hat{y} is the result of sigmoid regression. m is the total number of samples and i represents the index of the sample. w and b represent weights and biases, respectively.

4. EXPERIMENTAL SETTING

The experiments can be divided into two parts. First, we compare GCBMI with other state-of-art methods. The ET (Hsu and Si, 2018), Elastic Net (Wang et al., 2020), IG-MBKX (Zhang et al., 2020), and MOBAA-LS (Dashtban et al., 2018) are selected as the baselines. A detailed description of the comparison methods is as follows:

- ET was proposed by Hsu et al. They used ET to calculate the weight of the features and select features with high weight. SVM was combined to evaluate the feature subsets. This method achieved about 95% accuracy on TCGA datasets.
- Elastic Net was a novel method that integrates the Pearson correlation coefficient, T -test, and FDR. The data are based on gene expression data and DNA methylation data. In six types

of omics-data, the accuracy can up to about 98% by combining with SVM.

- IG-MBKX was presented and applied to feature selection for high-dimensional datasets. This method combined IG and krill herd algorithm and they used K-Nearest Neighbor (KNN) classifier to evaluate the classification accuracy. The accuracy of classification on nine different cancer datasets was more than 90%.
- MOBAA-LS is based on fisher criterion and BA. The accuracy achieved 100, 97, and 100% on leukemia, prostate, and SRBCT datasets, respectively.

Second, we investigate the prediction performance of DNN in biomarker identification for gastric cancer and how our method using different classifiers can affect the classification accuracy. We undertake experiments to compare our method using DNN classifier compared with using the traditional classifiers, such as KNN (Tahir et al., 2007), SVM (Vieira et al., 2013), and Naive Bayesian (NB) (Bielza and Larrañaga, 2014).

4.1. Dataset

We select the GEO database, which is an authoritative database of cancer applied in many previous studies (Zouridis et al., 2012; Wang et al., 2013) as the benchmark database. And the gene expression data GSE29272 (Li et al., 2014) and DNA methylation data GSE30601 (Lei et al., 2013; Kurashige et al., 2016) of gastric cancer are downloaded to construct our experiment dataset. As shown in **Table 1**, there are 268 samples of gene expression data including 134 tumor samples, 134 normal samples, and 13,515 features. And DNA methylation data contains 203 tumor samples, 94 normal samples, and 14,476 features.

4.2. Parameter Setting

The experiments are conducted on Intel Dual Core CPU, 8 GB RAM, Windows 7 operating system. The procedure

TABLE 1 | Benchmark dataset.

Dataset	Gene expression	DNA methylation
GEO ID	GSE29272	GSE30601
Normal samples	134	203
Tumor samples	134	94
Features	13515	14476

TABLE 2 | Parameter setting.

Methods	Parameter setting
GCBMI	MI: $n = 3,000$; Gene expression: $ FC > 2, p < 0.05, FDR < 0.01$; DNA methylation: $ FC > 1.8, p < 0.05, FDR < 0.01$
ET	Default parameters
IG-MBKH	$N = 20$; Iterations = 400; TopM = 80; Nmax = 4; Vf = 0.02; Dmax = 0.005
Elastic Net	$p < 0.05, FDR < 0.01$, ElasticNetCV (cv = 10)
MOBBA-LS	opN = 500, Population = 20, iteration = 300, alpha = 0.9, sigma = 0.7, injRate = 0.01, extRate = 0.01

is implemented under the programming environment Python version 3.6. The feature selection algorithms, statistical detection methods, and classifiers are provided by the Scikit-learn package and scipy package and the DNN is built by Keras package. Related parameters are given as follows: DNN is set as described in the Section 3.3; SVM: degree = 3, gamma = auto, kernel = “rbf,” cache_size = 200; KNN: $K = 5$. The parameters of methods are set according to the original literature (Dashtban et al., 2018; Hsu and Si, 2018; Wang et al., 2020; Zhang et al., 2020). The specific settings are shown in **Table 2**.

According to Park et al. (2020), all experiments use five-fold cross validation. The dataset is divided into five parts, and one part is taken as the test set in order and the rest parts are taken as the training set in each cross validation. After the Cartesian product is executed, there are average 8,053 normal samples, 17,400 tumor samples as training set, and 496 normal samples, 1,079 tumor samples as test set. The accuracy, precision, recall, F1-score and Area Under Curve (AUC) are utilized to evaluate the classification results of the model (Tanzi et al., 2020). These evaluation indicators are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

$$Prediction = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F1 - Score = \frac{2 \cdot Prediction \cdot Recall}{Prediction + Recall} \quad (7)$$

The positive samples are tumor samples and the negative samples are normal samples. True positive (TP) indicates the number

TABLE 3 | Performance comparison on different metrics (the accuracy, precision, recall, F1-score, and AUC value are average).

Methods	Accuracy	Precision	Recall	F1-score	AUC
GCBMI + DNN	0.9870	0.9971	0.9836	0.9903	0.9891
ET + SVM	0.9259	0.8571	1.0	0.9230	0.9333
Elastic Net + SVM	0.8922	0.9003	0.9433	0.9210	0.8598
IG-MBKH + KNN	0.9518	0.9730	0.9166	0.9437	0.9483
MOBBA-LS + SVM	0.94	0.9477	0.9327	0.9401	0.9412

The bold values represent the highest value of each metrics.

of tumor samples that have been correctly classified, false positive (FP) indicates the number of normal samples which are misclassified as tumor samples, true negative (TN) indicates the number of correctly classified normal samples, and false negative (FN) indicates the number of tumor samples, which are misclassified as normal samples.

5. RESULTS AND DISCUSSION

5.1. Comparison of Other State-of-Art Methods

In this section, GCBMI is compared with other state-of-art methods, and the experimental results are shown in **Table 3**. The accuracy of GCBMI achieved is 98.7%. The Elastic net also applies omics data, but the accuracy of GCBMI is 9% higher than the Elastic net. The performance of two wrapper methods IG-MBKH and MOBBA-LS are similar in our experiment. In terms of accuracy, these two methods are about 5% lower than our approach. The accuracy of extremely randomized trees achieved is 93%. What is more, in terms of precision and recall, GCBMI also has the highest precision and the second highest recall. This indicates FP and FN appear less frequently and the classification performance of GCBMI is superior to other state-of-art methods.

F1-score and AUC value are often applied to evaluate the stability and robustness of models. The two indicators of GCBMI can achieve about 99%. It is 5–7% higher than other state-of-art methods. In order to display the advantages of our method more intuitively, the histogram of experimental results is plotted in **Figure 3**.

Overall, GCBMI can get better performance on different evaluation indicators than other feature selection methods, which indicates that the genes identified by GCBMI have more sufficient capacity to classify gastric cancer. The high F1-score and AUC value also illustrate that our model has better stability. The experimental results suggest that combined omics data are meaningful, and it may reveal some causal relationships between different biological layers.

5.2. The Impact of Classifiers on Performance

In this section, the impact of different classifiers is evaluated on our feature selection method. **Table 4** displays the experimental results, which indicates that DNN model compared with the other three classifiers has better performance in different evaluation indicators. The performance of KNN is similar to

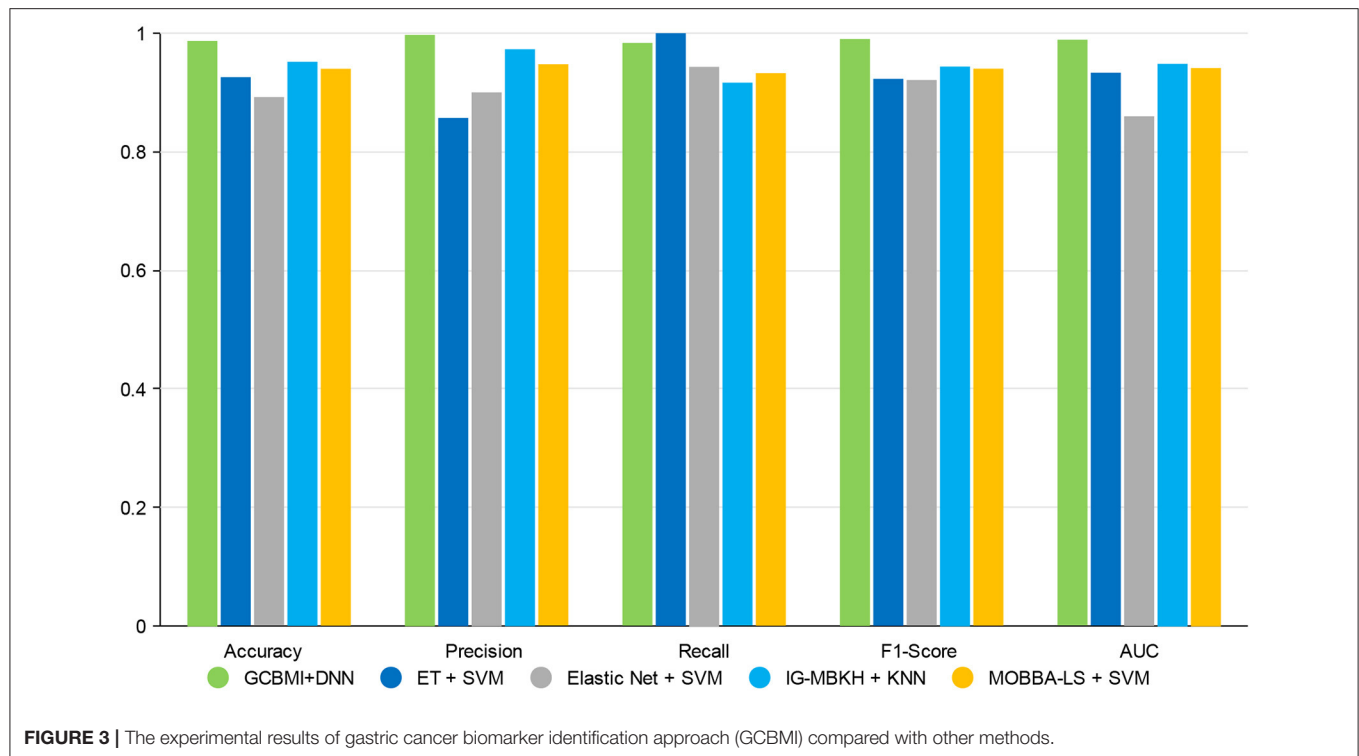


TABLE 4 | Results with different classifiers (the accuracy, precision, recall, F1-score, and AUC value are average).

Classifiers	Accuracy	Precision	Recall	F1-score	AUC
DNN	0.9870	0.9971	0.9836	0.9903	0.9891
KNN	0.9776	0.9934	0.9729	0.9830	0.9795
SVM	0.9819	0.9878	0.9826	0.9862	0.9803
NB	0.9651	0.9698	0.9777	0.9737	0.9557

The bold values represent the highest value of each metrics.

SVM and NB is worst but still reaches 96%. The performance of our method is stable in different classifiers. GCBMI integrates gene expression data and DNA methylation data and expands the number of samples. In this way, the DNN model can be trained better and achieves superior results than other classifiers.

On the whole, when compared with the KNN, SVM, and NB, our deep neural network model has better performance in different metrics, which indicates the validity of our feature selection approach. All the experimental results indicate that DNN model is a more appropriate classifier to feature selection in our approach. **Figure 4** shows the histogram of the average accuracy, F1 score, and AUC value of GCBMI with different classifiers, respectively. The classification advantage of DNN model has been shown in it, which has demonstrated the effectiveness of GCBMI.

5.3. Gene Analysis

In our experiment, the overlapped genes are recorded, which are shown in **Table 5**. In each fold of cross-validation, about 20 genes are selected. These genes are the intersections of DEG and DMP. Among them, eight genes appear in each intersection

and they are thought to be biomarkers of gastric cancer. In this section, the selected genes are further analyzed to understand the biological relevance.

Through literature retrieving, we can find the coding protein of PGC is a digestive enzyme produced by the stomach and it is the main component of the gastric mucosa. Polymorphism of this gene is associated with gastric cancer susceptibility. Serum levels of this enzyme are used as the biomarker for certain stomach diseases, including *Helicobacter pylori* associated gastritis (Sun et al., 2009). Moreover, Liu et al. discovered PGC was positively expressed in normal gastric mucosa (100%), and the expression rate was 6.45% in gastric cancer (Liu D. et al., 2009). The results suggested that PGC has important application value in the diagnosis of gastric cancer.

For gene PSCA, relevant research demonstrated that proteins encoded by PSCA play an important role in cell proliferation. In addition to being highly expressed in the prostate, it is also expressed in differentiating gastric epithelial cells. This gene includes a polymorphism that results in an upstream start codon in some individuals; this polymorphism is thought to be associated with a risk for gastric cancers (Bahrenberg et al., 2000; Sakamoto et al., 2008).

Except for PGC and PSCA, gene PDGFD as a member of PDGF family (Huang et al., 2014), its signaling pathway has been considered as a new target for the treatment of gastric cancer (Wang et al., 2009). Besides, gene KCNE2 is expressed mainly in the cytoplasm of parietal cells. Kuwahara et al. discovered that the loss of KCNE2 expression could cause gastric adenocancer (Kuwahara et al., 2013).

For these eight genes identified, in order to observe their expression level, gene expression heatmap is constructed. As

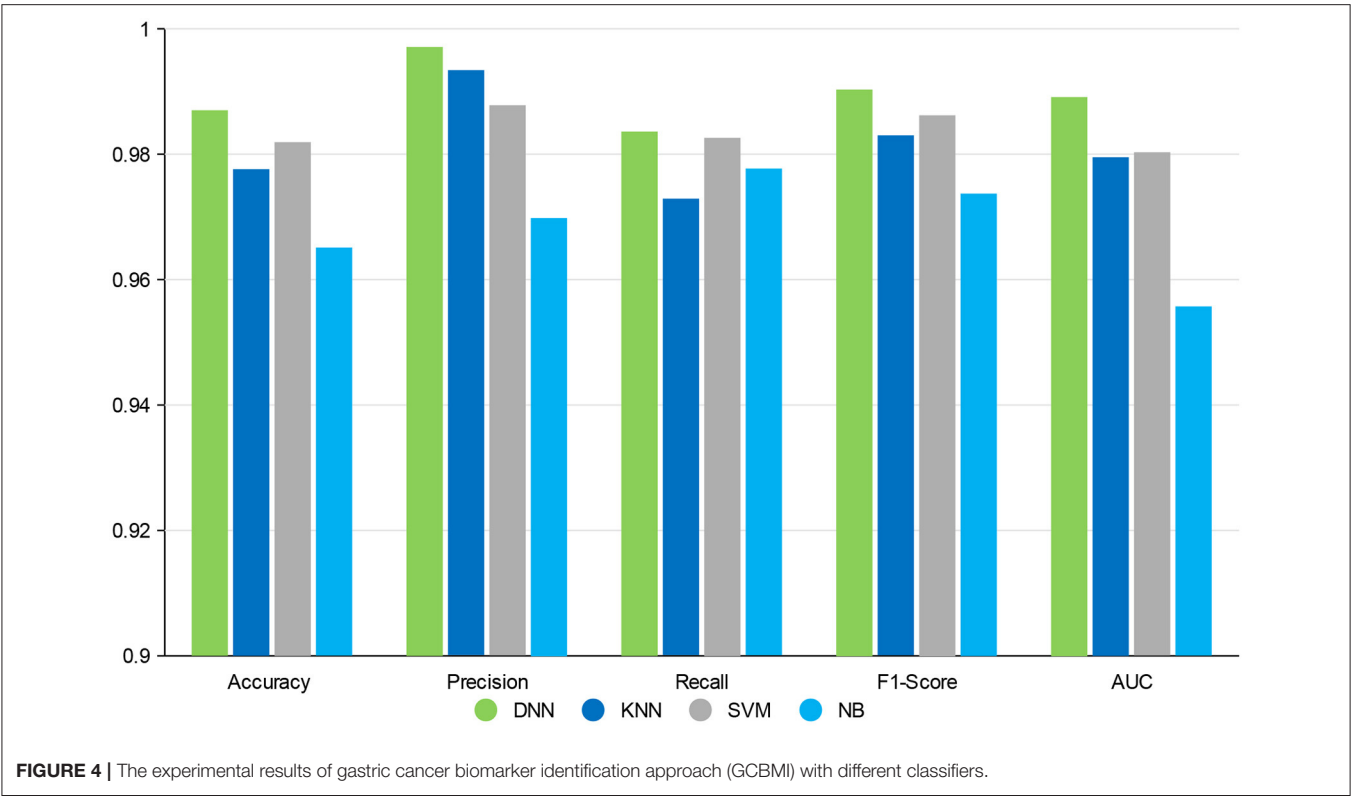


TABLE 5 | Selected genes from integrating gene expression and DNA methylation dataset.

K-fold	Number of overlapping genes	Selected genes
K = 1	17	FAHD2A,PGC,FIGF,PPAP2B,FOXA1,IFITM2,HOXC10, GPRC5C,CLEC3B,FBN1,LIF,C5,PSCA,PDGFD,KCNE2,RORC,C3
K = 2	19	PGC,FIGF,NID2,PPAP2B,IFITM2,RAB31,RORC,GPRC5C,FSCN1,TEAD4,CLEC3B,RAB17,IGFALS,C5,PSCA,PDGFD,KCNE2,COL4A1,C3
K = 3	17	FAHD2A,PGC,PPAP2B,FOXA1,IFITM2,IGFALS,GPRC5C, TEAD4,DNM1,ORM1,PTPRN2,FBN1,PSCA,PDGFD,KCNE2,RORC,C3
K = 4	24	PGC,FIGF,PDGFRB,PSMA7,TEAD4,C5,RORC,ADA, IFITM1,FAHD2A,PPAP2B,IGFALS,SLC1A2,GPRC5C,CLEC3B,CAPN9,KCNE2,PSCA,IFITM2,FSCN1,RPRM, PDGFD,SERPINA4,FBN1
K = 5	17	IFITM1,PGC,FIGF,PPAP2B,KCNE2,IFITM2,HOXC10, GPRC5C,CAPN9,FBN1,HRAS,C5,PSCA,PDGFD,SERPINA4,RORC,C3
Overlapped genes in 5-CV	8	PGC,RORC,GPRC5C,PDGFD,KCNE2,PSCA,IFITM2, PPAP2B

shown in **Figure 5**, the expression levels of these eight genes in all samples are demonstrated. The first half of the heatmap are normal samples, and others are tumor samples. Basically, the result shows that these genes have different expression in normal and tumor samples. Some of these genes differed significantly between the two classes and may have some relationship with gastric cancer.

What is more, the enrichment analysis is conducted by DAVID database for selected genes. As shown in **Table 6**, biological significance of the genes are reported through Gene Ontology (GO). “GO:0008284 positive regulation of cell proliferation,” “GO:0046597 negative regulation of viral entry into host cell,” “GO:0030335 positive regulation of cell migration” are common biological activities in human cancer (Dyrskjot

et al., 2009). Among them, there have some items about platelet, some studies have suggested that gastric cancer may lead to changes in platelet count and morphology (Matowicka-Karna et al., 2013). In addition, some studies also have been pointed out that interferon (Ferrantini et al., 2007) and other related factors may have relationship with the occurrence of cancer.

6. CONCLUSION

In this work, we propose a novel feature selection approach, GCBMI, which uses gene expression and DNA methylation data for identifying the biomarkers of gastric cancer. GCBMI consists of three main parts, namely data pre-processing, selection of differentially expressed genes and data combination, and

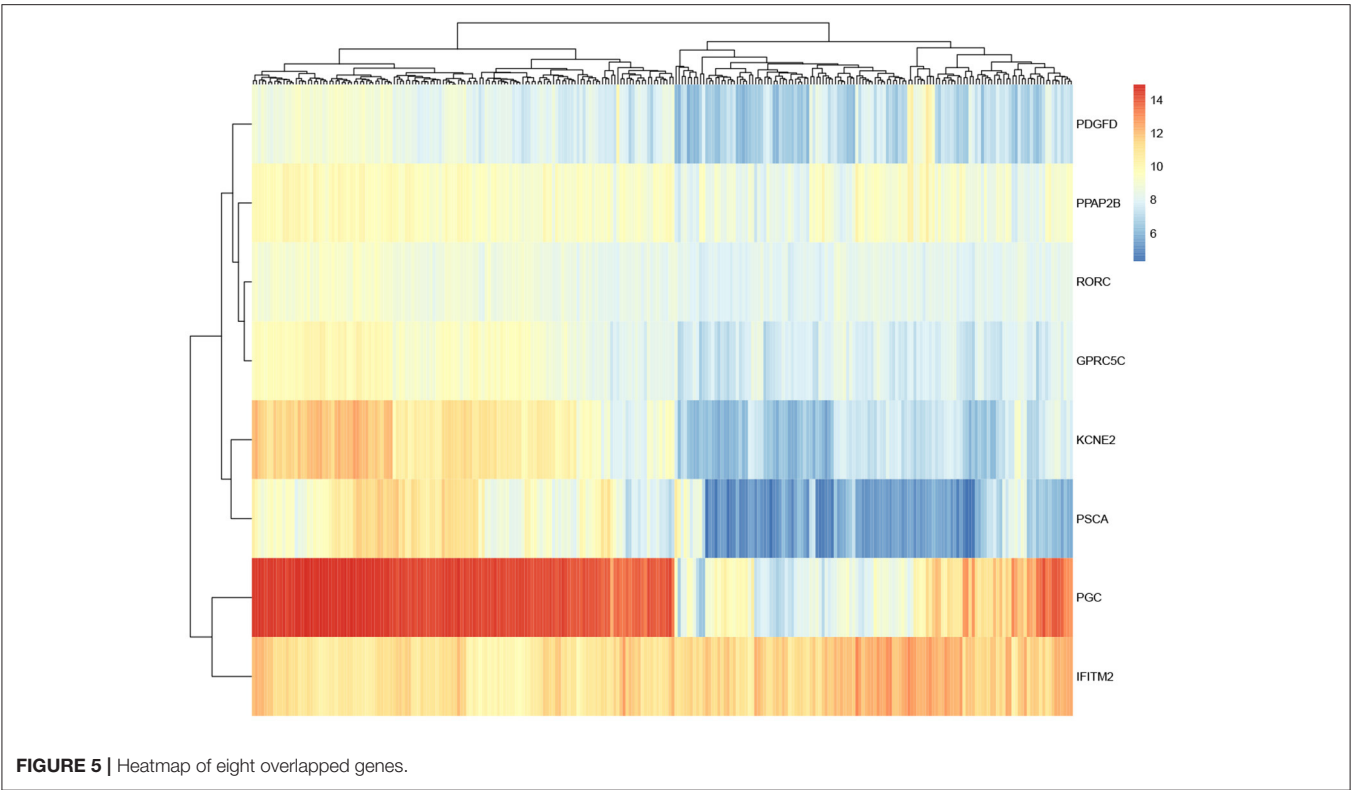


TABLE 6 | GO analysis of selected genes.

Category	Term	p-value	Gene
GOTERM_BP_DIRECT	GO:0071560 cellular response to transforming growth factor beta stimulus	0.003912643	CLEC3B,FBN1, PDGFD
GOTERM_BP_DIRECT	GO:0043406 positive regulation of MAP kinase activity	0.005625548	HRAS,PDGFRB, PDGFD
GOTERM_BP_DIRECT	GO:0008284 positive regulation of cell proliferation	0.01138237	LIF,HOXC10,HRAS, PDGFRB,PDGFD
GOTERM_BP_DIRECT	GO:0002576 platelet degranulation	0.016395992	ORM1,CLEC3B, SERPINA4
GOTERM_BP_DIRECT	GO:0035456 response to interferon-beta	0.017024892	IFITM1,IFITM2
GOTERM_BP_DIRECT	GO:0035455 response to interferon-alpha	0.018899122	IFITM1,IFITM2
GOTERM_MF_DIRECT	GO:0048407 platelet-derived growth factor binding	0.020021643	COL4A1,PDGFRB
GOTERM_MF_DIRECT	GO:0005102 receptor binding	0.026443684	LIF,C3,C5,PDGFRB
GOTERM_MF_DIRECT	GO:0005161 platelet-derived growth factor receptor binding	0.02720561	PDGFRB,PDGFD
GOTERM_BP_DIRECT	GO:0036120 cellular response to platelet-derived growth factor stimulus	0.033768846	PDGFRB,PDGFD
GOTERM_BP_DIRECT	GO:0046597 negative regulation of viral entry into host cell	0.033768846	IFITM1,IFITM2
GOTERM_BP_DIRECT	GO:0030335 positive regulation of cell migration	0.047784333	HRAS,PDGFRB, PDGFD
GOTERM_BP_DIRECT	GO:0048008 platelet-derived growth factor receptor signaling pathway	0.053858697	PDGFRB, PDGFD

deep neural network as the classifier. Differential expression analysis, statistical test, and MI are integrated to obtain comprehensive view to implement the biomarkers identification after data pre-processing. MI is introduced to filter out irrelevant gene, and FC and *T*-test are utilized to select differentially expressed genes. In particular, FDR is applied to revise the *p*-value to further screen genes. After that, Cartesian product is performed to expand samples. Moreover, GCBMI adopts DNN as the classifier to evaluate the classification ability of selected genes. Experimental results on GEO dataset indicate that the proposed approach outperforms other state-of-the-art

feature methods. The results of biological relevant verification indicate the status of the selected gene as the biomarkers of gastric cancer.

What is more, the performance of combined with omics data tends to be more superior than using a single omics data alone. In the future, some other omics data will be combined such as copy number variation (CNV) data to identify cancer biomarkers, and our methods will be applied to other fields as well (Liu et al., 2020). Besides, some measures will also be taken to improve our method so that its classification performance can be improved further.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

CY and ZX conceived and designed the approach. ZX performed the experiments. HL analyzed the data. GZ and ZX wrote the manuscript. CY and JW supervised the whole study process and revised the manuscript. All authors have read and approved the final version of manuscript.

REFERENCES

- Algina, J., Oshima, T., and Lin, W.-Y. (1994). Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. *J. Educ. Stat.* 19, 275–291. doi: 10.3102/10769986019003275
- Ankarali, H., Yazici, A. C., and Ankarali, S. (2009). A bootstrap confidence interval for skewness and kurtosis and properties of t-test in small samples from normal distribution. *Med. J. Trakya Univ.* 26, 297–305. doi: 10.1620/tjem.219.337
- Bahrenberg, G., Brauers, A., Joost, H.-G., and Jakse, G. (2000). Reduced expression of psc, a member of the ly-6 family of cell surface antigens, in bladder, esophagus, and stomach tumors. *Biochem. Biophys. Res. Commun.* 275, 783–788. doi: 10.1006/bbrc.2000.3393
- Bielza, C., and Larrañaga, P. (2014). Discrete bayesian network classifiers: a survey. *ACM Comput. Surv.* 47, 1–43. doi: 10.1145/2576868
- Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature* 321, 209–213. doi: 10.1038/321209a0
- Chen, Z., Pang, M., Zhao, Z., Li, S., Miao, R., Zhang, Y., et al. (2020). Feature selection may improve deep neural networks for the bioinformatics problems. *Bioinformatics* 36, 1542–1552. doi: 10.1093/bioinformatics/btz763
- Dashbani, M., Balafar, M., and Suravajhala, P. (2018). Gene selection for tumor classification using a novel bio-inspired multi-objective approach. *Genomics* 110, 10–17. doi: 10.1016/j.ygeno.2017.07.010
- Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.* 3, 185–205. doi: 10.1142/S0219720005001004
- Ding, W., Bu, H., Zheng, S., and Qian, F. (2009). "Tumor classification by using PCA with relief wrapper" in *2009 2nd IEEE International Conference on Computer Science and Information Technology* (Beijing: IEEE), 514–517. doi: 10.1109/ICCSIT.2009.5234895
- Ding, W., Chen, G., and Shi, T. (2019). Integrative analysis identifies potential DNA methylation biomarkers for pan-cancer diagnosis and prognosis. *Epigenetics* 14, 67–80. doi: 10.1080/15592294.2019.1568178
- Dyrskjot, L., Ostfeld, M. S., Bramsen, J. B., Silaharoglu, A. N., Lamy, P., Ramanathan, R., et al. (2009). Genomic profiling of microRNAs in bladder cancer: miR-129 is associated with poor outcome and promotes cell death *in vitro*. *Cancer Res.* 69, 4851–4860. doi: 10.1158/0008-5472.CAN-08-4043
- Emelyanov, P., and Ponomaryov, D. (2017). "Cartesian decomposition in data analysis," in *2017 Siberian Symposium on Data Science and Engineering (SSDSE)* (Novosibirsk: IEEE), 55–60. doi: 10.1109/SSDSE.2017.8071964
- Ferrantini, M., Capone, I., and Belardelli, F. (2007). Interferon- α and cancer: mechanisms of action and new perspectives of clinical use. *Biochimie* 89, 884–893. doi: 10.1016/j.biochi.2007.04.006
- Gauvreau, K., and Pagano, M. (1993). Student's t test. *Nutrition* 9:386.
- Ghosh, M., Begum, S., Sarkar, R., Chakraborty, D., and Maulik, U. (2019). Recursive memetic algorithm for gene selection in microarray data. *Expert Syst. Appl.* 116, 172–185. doi: 10.1016/j.eswa.2018.06.057

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 61802113, 61802114, and 61972134), Science and Technology Development Plan Project of Henan Province (Nos. 202102210173 and 212102210091), China Postdoctoral Science Foundation (No. 2020M672212), and Henan Province Postdoctoral Research Project Founding.

ACKNOWLEDGMENTS

This paper is recommended by the 5th Computational Bioinformatics Conference.

- Han, F., Tang, D., Cheng, Z., Jiang, J., and Li, Q.-W. (2019). A hybrid gene selection method based on gene scoring strategy and improved particle swarm optimization. *BMC Bioinformatics* 20:289. doi: 10.1186/s12859-019-2773-x
- Hsu, Y.-H., and Si, D. (2018). "Cancer type prediction and classification based on RNA-sequencing data," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Honolulu, HI: IEEE), 5374–5377. doi: 10.1109/EMBC.2018.8513521
- Huang, F., Wang, M., Yang, T., Cai, J., Zhang, Q., Sun, Z., et al. (2014). Gastric cancer-derived msc-secreted pdgf-dd promotes gastric cancer progression. *J. Cancer Res. Clin. Oncol.* 140, 1835–1848. doi: 10.1007/s00432-014-1723-2
- Kurashige, J., Hasegawa, T., Niida, A., Sugimachi, K., Deng, N., Mima, K., et al. (2016). Integrated molecular profiling of human gastric cancer identifies ddr2 as a potential regulator of peritoneal dissemination. *Sci. Rep.* 6:22371. doi: 10.1038/srep22371
- Kuwahara, N., Kitazawa, R., Fujiishi, K., Nagai, Y., Haraguchi, R., and Kitazawa, S. (2013). Gastric adenocarcinoma arising in gastritis cystica profunda presenting with selective loss of *kcne2* expression. *World J. Gastroenterol.* 19:1314. doi: 10.3748/wjg.v19.i8.1314
- Lee, I.-H., Lushington, G. H., and Visvanathan, M. (2011). A filter-based feature selection approach for identifying potential biomarkers for lung cancer. *J. Clin. Bioinform.* 1:11. doi: 10.1186/2043-9113-1-11
- Lei, Z., Tan, I. B., Das, K., Deng, N., Zouridis, H., Pattison, S., et al. (2013). Identification of molecular subtypes of gastric cancer with different responses to pi3-kinase inhibitors and 5-fluorouracil. *Gastroenterology* 145, 554–565. doi: 10.1053/j.gastro.2013.05.010
- Li, W.-Q., Hu, N., Burton, V. H., Yang, H. H., Su, H., Conway, C. M., et al. (2014). PLCE1 mRNA and protein expression and survival of patients with esophageal squamous cell carcinoma and gastric adenocarcinoma. *Cancer Epidemiol. Prevent. Biomark.* 23, 1579–1588. doi: 10.1158/1055-9965.EPI-13-1329
- Liu, D., Wu, J., and Wu, H.-X. (2009). Expression of MG7 and PGC in gastric cancer and precancerous lesion and its significance. *China Cancer*, 1.
- Liu, H., Sun, J., Liu, L., and Zhang, H. (2009). Feature selection with dynamic mutual information. *Pattern Recogn.* 42, 1330–1339. doi: 10.1016/j.patcog.2008.10.028
- Liu, X., Chen, S., Liu, J., Qu, W., Xiao, F., Liu, A. X., et al. (2020). Fast and accurate detection of unknown tags for RFID systems – hash collisions are desirable. *IEEE/ACM Trans. Network.* 28, 126–139. doi: 10.1109/TNET.2019.2957239
- Lu, H., Chen, J., Yan, K., Jin, Q., Xue, Y., and Gao, Z. (2017). A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing* 256, 56–62. doi: 10.1016/j.neucom.2016.07.080
- Luo, X., Wang, F., Wang, G., and Zhao, Y. (2020). Identification of methylation states of DNA regions for Illumina methylation BeadChip. *BMC Genomics* 21:672. doi: 10.1186/s12864-019-6019-0
- Ma, T., and Zhang, A. (2019). "Affinitynet: semi-supervised few-shot learning for disease type prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI), 1069–1076. doi: 10.1609/aaai.v33i01.33011069
- Mallik, S., Bhadra, T., and Maulik, U. (2017). Identifying epigenetic biomarkers using maximal relevance and minimal redundancy based

- feature selection for multi-omics data. *IEEE Trans. Nanobiosci.* 16, 3–10. doi: 10.1109/TNB.2017.2650217
- Matowicka-Karna, J., Kamocki, Z., Polńska, B., Osada, J., and Kemona, H. (2013). Platelets and inflammatory markers in patients with gastric cancer. *Clin. Dev. Immunol.* 2013:6. doi: 10.1155/2013/401623
- Nogueira, C., Mota, M., Gradiz, R., Cipriano, M. A., Caramelo, F., Cruz, H., et al. (2017). Prevalence and characteristics of epstein-barr virus-associated gastric carcinomas in portugal. *Infect. Agents Cancer* 12:41. doi: 10.1186/s13027-017-0151-8
- Park, C., Ha, J., and Park, S. (2020). Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset. *Expert Syst. Appl.* 140:112873. doi: 10.1016/j.eswa.2019.112873
- Paziewska, A., Dabrowska, M., Goryca, K., Antoniewicz, A., Dobruch, J., Mikula, M., et al. (2014). DNA methylation status is more reliable than gene expression at detecting cancer in prostate biopsy. *Brit. J. Cancer* 111, 781–789. doi: 10.1038/bjc.2014.337
- Peng, H., and Fan, Y. (2017). Feature selection by optimizing a lower bound of conditional mutual information. *Informat. Sci.* 418, 652–667. doi: 10.1016/j.ins.2017.08.036
- Rodrigues, D., Pereira, L. A., Nakamura, R. Y., Costa, K. A., Yang, X.-S., Souza, A. N., et al. (2014). A wrapper approach for feature selection based on bat algorithm and optimum-path forest. *Expert Syst. Appl.* 41, 2250–2258. doi: 10.1016/j.eswa.2013.09.023
- Ruffalo, M., Koyutürk, M., and Sharan, R. (2015). Network-based integration of disparate omic data to identify “silent players” in cancer. *PLoS Comput. Biol.* 11:e1004595. doi: 10.1371/journal.pcbi.1004595
- Sakamoto, H., Yoshimura, K., Saeki, N., Katai, H., Shimoda, T., Matsuno, Y., et al. (2008). Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nat. Genet.* 40, 730–740. doi: 10.1038/ng.152
- Shen, H., and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *J. Multivar. Anal.* 99, 1015–1034. doi: 10.1016/j.jmva.2007.06.007
- Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *Ca A Cancer J. Clin.* 70, 7–30. doi: 10.3322/caac.21590
- Singh, D., and Yamada, M. (2020). FSNet: feature selection network on high-dimensional biological data. *arXiv [preprint] arXiv:2001.08322*.
- Song, Z., Wu, Y., Yang, J., Yang, D., and Fang, X. (2017). Progress in the treatment of advanced gastric cancer. *Tumor Biol.* 39:1010428317714626. doi: 10.1177/1010428317714626
- Sun, L.-P., Gong, Y.-H., Dong, N.-N., Wang, L., and Yuan, Y. (2009). Correlation of pepsinogen c (PGC) gene insertion/deletion polymorphism to PGC protein expression in gastric mucosa and serum. *Chin. J. Cancer* 28, 487–492.
- Tahir, M. A., Bouridane, A., and Kurugollu, F. (2007). Simultaneous feature selection and feature weighting using hybrid tabu search/k-nearest neighbor classifier. *Pattern Recogn. Lett.* 28, 438–446. doi: 10.1016/j.patrec.2006.08.016
- Tanzi, L., Vezzetti, E., Moreno, R., Aprato, A., Audisio, A., and Massé, A. (2020). Hierarchical fracture classification of proximal femur x-ray images using a multistage deep learning approach. *Eur. J. Radiol.* 133:109373. doi: 10.1016/j.ejrad.2020.109373
- Vieira, S. M., Mendonça, L. F., Farinha, G. J., and Sousa, J. M. (2013). Modified binary pso for feature selection using svm applied to mortality prediction of septic patients. *Appl. Soft Comput.* 13, 3494–3504. doi: 10.1016/j.asoc.2013.03.021
- Wang, G., Hu, N., Yang, H. H., Wang, L., Su, H., Wang, C., et al. (2013). Comparison of global gene expression of gastric cardia and noncardia cancers from a high-risk population in china. *PLoS ONE* 8:e63826. doi: 10.1371/journal.pone.0063826
- Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2018). MeDReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151. doi: 10.1093/nar/gkx1096
- Wang, X., Shang, W., Li, X., and Chang, Y. (2020). Methylation signature genes identification of cancers occurrence and pattern recognition. *Comput. Biol. Chem.* 85:107198. doi: 10.1016/j.compbiolchem.2019.107198
- Wang, Z., Kong, D., Li, Y., and Sarkar, F. H. (2009). PDGF-D signaling: a novel target in cancer therapy. *Curr. Drug Targets* 10, 38–41. doi: 10.2174/138945009787122914
- Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometr. Intell. Lab. Syst.* 2, 37–52. doi: 10.1016/0169-7439(87)80084-9
- Yan, C., Ma, J., Luo, H., and Wang, J. (2018). A hybrid algorithm based on binary chemical reaction optimization and tabu search for feature selection of high-dimensional biomedical data. *Tsinghua Sci. Technol.* 23, 733–743. doi: 10.26599/TST.2018.9010101
- Zhang, C., Cai, H., Huang, J., and Song, Y. (2016). nbCNV: a multi-constrained optimization model for discovering copy number variants in single-cell sequencing data. *BMC Bioinformatics* 17:384. doi: 10.1186/s12859-016-1239-7
- Zhang, G., Hou, J., Wang, J., Yan, C., and Luo, J. (2020). Feature selection for microarray data classification using hybrid information gain and a modified binary krill herd algorithm. *Interdiscipl. Sci. Comput. Life Sci.* 12, 288–301. doi: 10.1007/s12539-020-00372-w
- Zhang, Z., Cheng, Y., and Liu, N. C. (2014). Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of web of science subject categories. *Scientometrics* 101, 1679–1693. doi: 10.1007/s11192-014-1294-7
- Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. *BMC Bioinformatics* 21:43. doi: 10.1186/s12859-020-3388-y
- Zouridis, H., Deng, N., Ivanova, T., Zhu, Y., Wong, B., Huang, D., et al. (2012). Methylation subtypes and large-scale epigenetic alterations in gastric cancer. *Sci. Transl. Med.* 4:156ra140. doi: 10.1126/scitranslmed.3004504

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.

Copyright © 2021 Zhang, Xue, Yan, Wang and Luo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Application of Multilayer Network Models in Bioinformatics

Yuanyuan Lv^{1,2}, Shan Huang³, Tianjiao Zhang^{4*} and Bo Gao^{5*}

¹ Hainan Key Laboratory for Computational Science and Application, Hainan Normal University, Haikou, China, ² Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Quzhou, China, ³ Department of Neurology, The Second Affiliated Hospital of Harbin Medical University, Harbin, China, ⁴ College of Information and Computer Engineering, Northeast Forestry University, Harbin, China, ⁵ Department of Radiology, The Second Affiliated Hospital, Harbin Medical University, Harbin, China

OPEN ACCESS

Edited by:

Chunhou Zheng,
Anhui University, China

Reviewed by:

Jin-Xing Liu,
Qufu Normal University, China
Chunyu Wang,
Harbin Institute of Technology, China

*Correspondence:

Tianjiao Zhang
ztj.hit@gmail.com
Bo Gao
1678729588@qq.com

Specialty section:

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

Received: 06 February 2021

Accepted: 26 February 2021

Published: 31 March 2021

Citation:

Lv Y, Huang S, Zhang T and
Gao B (2021) Application of Multilayer
Network Models in Bioinformatics.
Front. Genet. 12:664860.
doi: 10.3389/fgene.2021.664860

Multilayer networks provide an efficient tool for studying complex systems, and with current, dramatic development of bioinformatics tools and accumulation of data, researchers have applied network concepts to all aspects of research problems in the field of biology. Addressing the combination of multilayer networks and bioinformatics, through summarizing the applications of multilayer network models in bioinformatics, this review classifies applications and presents a summary of the latest results. Among them, we classify the applications of multilayer networks according to the object of study. Furthermore, because of the systemic nature of biology, we classify the subjects into several hierarchical categories, such as cells, tissues, organs, and groups, according to the hierarchical nature of biological composition. On the basis of the complexity of biological systems, we selected brain research for a detailed explanation. We describe the application of multilayer networks and chronological networks in brain research to demonstrate the primary ideas associated with the application of multilayer networks in biological studies. Finally, we mention a quality assessment method focusing on multilayer and single-layer networks as an evaluation method emphasizing network studies.

Keywords: multilayer networks, bioinformatics, brain network structure, biological systems, chronological networks

INTRODUCTION

In recent years, the formulation of multilayer networks has provided new methods for the study of multilevel network systems. Many biological systems comprise interconnected units that can be effectively modeled as networks, which are mathematical structures describing connections between points (Jing et al., 2019; Liu B. et al., 2020; Shao et al., 2020). Complex network systems provide powerful research tools and methods for studying biological fields (Kumari and Verma, 2020; Liu X. et al., 2020; Shao and Liu, 2020), from interactions between genes and proteins (Zhang et al., 2019; Li Z. et al., 2020; Zhai et al., 2020), to the study of tissue and organ functions (Yang et al., 2020), and even human brain study (Zhang J. et al., 2020). The complexity and evolutionary nature of biological systems enable the extensive application of multilayer networks and associated methods. Additionally, ecosystems and evolutionary systems evolve and change over time, and the corresponding network structures for these systems change correspondingly. Furthermore, the reasons for all these changes, particularly topological changes in the course of

network structure change, and the importance of network feedback in network structure analyses are all topics worthy of exploration.

A network representation is a simplified description of a more complex, multifaceted system. A social system can include different types of interactions of different biological significance (e.g., cooperation or competition), while standard network approaches usually ignore these interactions or achieve integration through analyzing networks with different edge types separately. In bioinformatics studies using network structures, the progress of each biological system relies on the amount of data and/or new discoveries about unknown biological areas. For example, in the study of transcription-translation relationships between genes and proteins, genes and proteins are represented by nodes, and the correspondence between genes and proteins is represented by links in the network. Therefore, it is necessary to first understand the characteristics of each individual gene and protein, and the methods used to identify these relationships (Lin et al., 2019; Zhang D. et al., 2021; Zhang Z.Y. et al., 2021; Zulfiqar et al., 2021). Only then can the most relevant genes and corresponding proteins for a disease or symptom be identified through clustering or linkage analyses of the network, which further enables the investigation of target therapies for symptoms of disease (Zhu et al., 2018; Iliopoulos et al., 2020). These applications all rely on the data set and on the biological correspondence of genes and proteins.

The definition of a multilayer network varies slightly from one application to another. All edges and nodes in a single network are homogeneous, but in the real world, there is heterogeneity in both the objects represented by the nodes and the connections represented by the edges. Multilayer networks add additional tagging capabilities to traditional networks. That is, tagging terms are added to the traditional network, which can be understood as a composite of simple (single-layer) networks with different tags for complex networks. This is a relatively easy way to understand the definition of complex networks on different systems. Currently, according to different applications and subjects, multilayer networks can be divided into the following types:

- (1) Multiplex networks: Networks in which the nodes on different layers are connected by inter-layer edges.
 - (a) In multi-relational networks, each layer represents a different type of interaction, i.e., different relationships are the distinguishing dimension for building a multilayer network, and the relationships are the tagged labels.
 - (b) In a temporal network, each layer encodes the same type of interaction at different time points or time windows. That is, time series (time windows) are the tagged labels between layers in a multilayer network.
- (2) Interconnected networks: Nodes in different layers do not necessarily represent the same entities and inter-layer edges between different types of nodes may exist.
 - (a) The network of networks consists of subsystems, which are themselves networks. They are interconnected by interlayer edges between subsystem nodes.

- (b) In a contextual network, each layer is interpreted to represent a different type of node. For example, interactions between males in one layer, interactions between females in another layer, and interactions between the sexes in a third layer. These interactions are represented by inter-layer edges.
- (c) Spatial networks (also known as geometric networks) can be connected by ecological networks of the subjects moving between various locations.

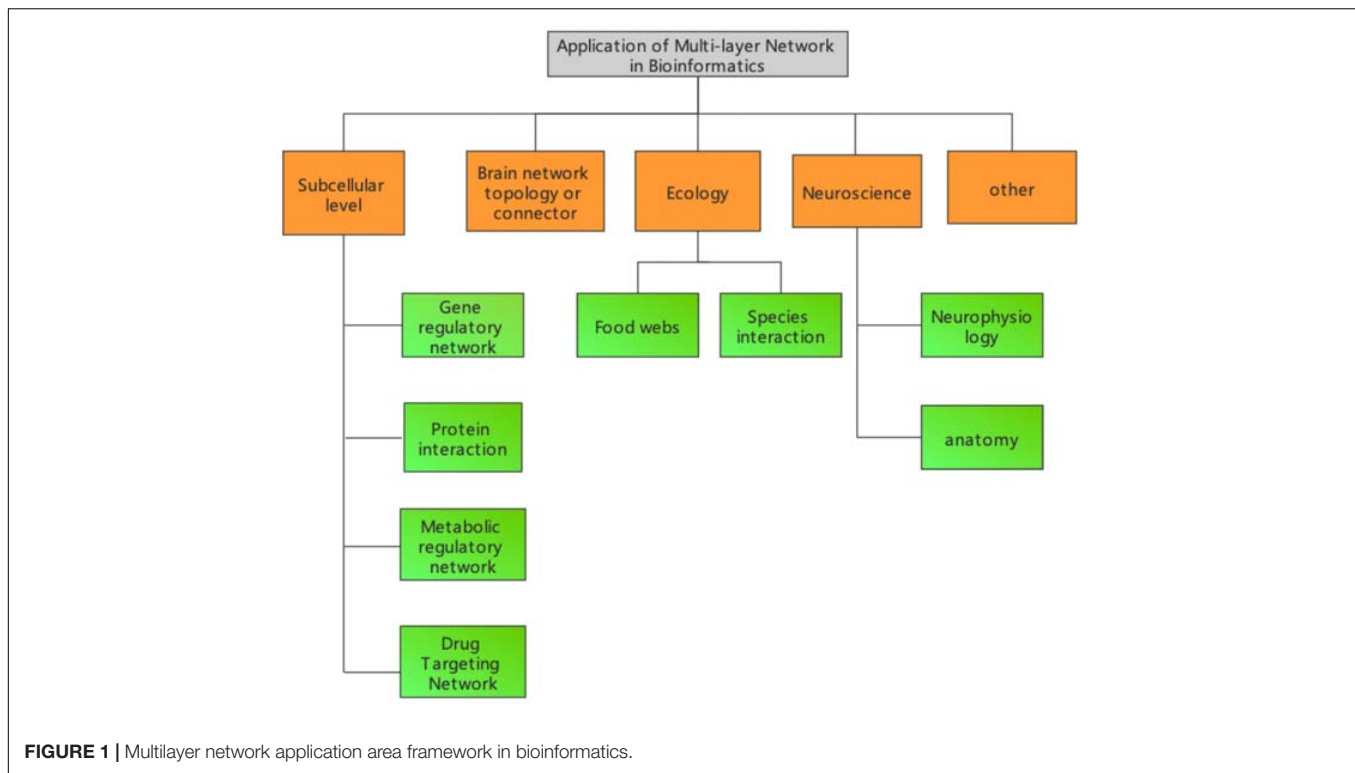
Multilayer networks are currently used in various fields including physics, chemistry, biology, technology, finance, and social systems because of their inherent structural and functional characteristics. In this review, we briefly introduce the development of multilayer network concepts, techniques, and applications in bioinformatics by reviewing multilayer network applications in bioinformatics, and we summarize the outlook and development of multilayer networks in bioinformatics by analyzing current research.

MULTILAYER NETWORK APPLICATIONS

The definition and methodology of multilayer network models in bioinformatics depends on the specific research problem. Organisms can be classified into different systems under different levels, and that system usually changes dynamically with time. Therefore, usually the representation of bioinformatics related networks varies depending on the specific biological system. In this review, we classify research topics into different categories according to the different levels of biological systems. As shown in **Figure 1**, multilayer networks in bioinformatics can be classified into five major categories.

As the understanding of DNA structure and function has gradually improved (Liu M.L. et al., 2020), understanding the relationships between genes and proteins, genes and disease, and disease and drugs has greatly evolved. For example, the correspondence between genes and disease has been mined through network structures, where the method utilizes a joint learning approach using the functional and connectivity patterns of proteins to predict disease-gene relationships using human interactome networks. In contrast to other data structures, interactomes are characterized by a high degree of incompleteness and lack of explicit negative knowledge, which makes prediction particularly challenging. To maximize potential information in the network, a second-order random walk procedure named random walker (RW²) is applied in these studies. The random walker is able to learn rich representations of disease gene (or gene product) characteristics. This method has successfully compared with the best-known disease gene prediction systems and other state-of-the-art graph-based methods.

A large number of candidate disease-causing genes can be sequenced and checked for variation to help determine relationships between disease and genes (Zhang Z.M. et al., 2020). Many different computational methods have been developed to address this challenge. The observation that genes associated with



similar diseases have a higher probability of interaction, many of these methods rely on the analysis of the topological properties of biological networks. However, the incomplete and noisy nature of biological networks is an important challenge. Two-step framework for disease gene ranking: (1) Construct a reliable functional connectivity network using sequence information and machine learning techniques. And (2) rank disease-gene relationships on the basis of that constructed functional connectivity network. Unlike other functional connectivity network-based frameworks that use functional connectivity networks based on the integration of various low-quality biological data, protein sequences can be used as comprehensive data to construct a reliable initial network. Additionally, the physicochemical properties of amino acids can be used to supplement hypotheses of protein function. In conclusion, our assessment of these methods indicates high efficiency and effectiveness for constructing functional linkage networks for disease genes (Wang et al., 2008; Jiang et al., 2010, 2013; Cheng et al., 2018; Zeng et al., 2018).

Gene function can also be determined by collecting biological data. For example, the *Drosophila* ovary epidermal cells (ECs) externally control the maintenance and progeny differentiation of germ line stem cells (GSC). In this study, the role of 173 EC genes that control GSC maintenance and progeny differentiation were identified using a *Drosophila in vivo* systemic RNAi approach (Zeng et al., 2016; Zou et al., 2016; Wang et al., 2019). Among the identified genes, 10 and 163 genes were required by ECs for GSC maintenance and progeny differentiation, respectively. The genes required for progeny differentiation were classified into different functional categories,

including transcription, mRNA splicing, protein degradation, signal transduction, and cytoskeleton regulation (Cao et al., 2019). In addition, GSC progeny differentiation defects caused by defective ECs were often associated with BMP signaling elevation, indicating that preventing BMP signaling is a general functional feature of the differentiation niche. Finally, EC exon junction complex (EJC) components were identified as required for EC maintenance and the prevention of BMP signaling, and thus the promotion of GSC progeny differentiation. Therefore, this study identifies the major regulators of the differentiation niche and provides important insights into the external control of stem cell progeny differentiation.

Corresponding network structures for different biological data and specific subjects can also be designed to analyse specific biological systems (Zeng et al., 2016; Jiang et al., 2017; Liu et al., 2017). Currently, at the subcellular level, these networks mainly include gene regulatory networks (Wang et al., 2010; Ding et al., 2011; Jiang et al., 2014; Cheng et al., 2019; Konda et al., 2019; Liu L. et al., 2019; Mortezaeefar et al., 2019; Hong et al., 2020), protein functional networks (Guo et al., 2011, 2013, 2014; Sikandar et al., 2019; Tao et al., 2020; Liu et al., 2021), metabolic regulatory networks (Jin et al., 2020), and drug targeting networks (Wei et al., 2014; Ding et al., 2017, 2019, 2020a,b; Jin Q. et al., 2019; Jin S. et al., 2019; Srivastava et al., 2019; Zhao et al., 2019; Zeng et al., 2020).

The study of human brain functional and structural mechanisms using brain networks is also a hot field. Currently, research has mainly studied brain function by acquiring the brain waves of subjects, and the functional partitions of the brain have been predominantly obtained by functional experiments or

magneto encephalography. This portion of our review will be introduced in detail in the next section.

Modern network theory is increasingly applied to neuroscience to understand neurophysiology and anatomy at different scales and under experimentally attainable physiological and pathophysiological conditions. The first attempt was made at the micro anatomical level of individual neurons. Watts and Strogatz analyzed the anatomical connections of the nervous system of *Caenorhabditis elegans* where neurons represent the nodes and synaptic or gap connections of a neural network. Their study revealed a highly clustered and efficient network, thus representing the first evidence of a real neural system with a small-world network. Later graph-theoretical approaches focused on morphological representations or dynamic correlations of the electrical stimulation activity of neuronal networks.

Network-based analyses have also been useful to address several questions in ecology and issues in conservation. The first study was carried out in a contextual network of so-called species interactions. Food webs are one of the fundamental issues in ecological studies, and despite the rather high variability detected in network structure, food webs present a complex topology similar to other types of real networks and host-parasite networks. One of the main advantages of these approaches is the direct assessment of the robustness and sensitivity of a given ecosystem to species loss or other perturbations. Another network type widely used in ecology is connected landscape mapping, where nodes typically represent patches on the landscape. The resulting spatial networks describe the linkages between processes and patterns that characterize the landscape, thus providing an effective way to assess important issues such as the effects of species dispersal or habitat loss.

Understanding the interactions between different species in a community and responses to environmental change is a central goal of ecology. However, defining the network structure of microbial communities is very challenging because of associated extremely diverse and unexplored states. Although recent developments in metagenomic technologies, such as high-throughput sequencing and functional gene arrays, have provided revolutionary tools for analyzing microbial community structure, it is still difficult to study network interactions in microbial communities based on high-throughput meta genomic data. A mathematical and bioinformatics framework for constructing molecular ecological networks (MENs) based on Random Matrix Theory (RMT) has been proposed. The remarkable feature of this approach compared with other network construction methods is that the network is automatically defined and robust to noise, thus providing a good solution to several common problems associated with high throughput.

APPLICATIONS OF MULTILAYER NETWORKS IN BRAIN RESEARCH

The brain is the control center of most animal activities, and it has been the goal of many researchers to unravel the mystery of the brain and simulate the human brain with external devices such as

computers. Before that, the structure and mechanism of the brain needs to be clarified, and it is costly to study the human brain because of its complexity. The human brain is a complex system organized by the structural and functional relationships among its components (Liu et al., 2018; Song et al., 2018; Liu G. et al., 2019). Recent experimental advances have led to unprecedented amounts of data that describe the structure and function of the brain, and it is now possible to model the brain as a network by measuring pairwise interactions between its various units. This modeling can be performed across multiple scales, where network nodes represent units of the brain, including proteins, neurons, brain regions, or other measurement units. Recording techniques such as functional magnetic resonance imaging (fMRI), magneto encephalography (MEG), and electroencephalography (EEG) are capable of capturing brain dynamics across time and across multiple frequency bands.

Recent neuroscience research has also exploited the versatility of multilayer frameworks to model complex relationships in neural data. For example, given fMRI and diffusion tensor imaging (DTI) for a single subject, a multilayer network can be constructed, with one layer representing the fMRI network and another layer representing the DTI network. Using the fMRI data, a functional network can be constructed in which the nodes represent brain regions and the edges represent the coherence between regional activities. On the basis of DTI data, a structural network can be constructed by dividing the brain into regions and then measuring the strength of physical connections between these regions. Finally, considering each network as a layer in a multilayer network, the edges of a brain region in the fMRI layer can be added to the DTI layer to form a multilayer network.

The brain is an inherently dynamic system, and the performance of cognition requires dynamic reconfiguration of a highly evolved network of brain regions, which interact in complex and transient communications. However, an accurate description of these reconfiguration processes during human cognitive function remains elusive (Liu and Jiang, 2016). Therefore, many studies have used temporal networks to investigate the dynamic cluster structure of brain networks and reveal the underlying human brain dynamic changes during learning. Temporal networks that contain temporal information have the advantage of retaining the full information of the data without aggregating connections into individual networks.

When we complete different cognitive vision tasks, we subdivide the regional time series into time windows of variable size, and determine the impact of the time window size on the observed dynamics. Specifically, we applied a multilayer community detection algorithm to identify temporal communities, and we computed network flexibility to quantify the changes in these communities over time. Within our frequency band of interest, large and small time windows were associated with the brain network flexibility within a narrow range, while medium time windows were associated with broad network flexibility. Using medium time windows of 75–100 s, we identified brain regions with low flexibility (considered core regions and observed in visual and attentional areas) and brain regions with high flexibility (considered peripheral regions and observed in subcortical and temporal lobe regions) by

comparison with appropriate control dynamic network models. In general, this work demonstrates the effect of time window size on the network dynamics observed during task execution, providing practical considerations when selecting time windows in dynamic network analysis. More generally, this work reveals organizing principles for functional brain connections that are inaccessible to static network approaches.

The hypothesis that human executive functions arise from the dynamic interactions of multiple networks has been tested in previous research (Ding et al., 2019). To corroborate this research, we investigated a key executive function (FCD), namely arbitrary visuomotor mapping. MEG and intracranial EEG were recorded using high gamma activity brain connectivity analysis. We then generated visuomotor mapping using the dynamic interactions of three partially overlapping cortico-cortical and cortico-subcortical functional connectivity (FC) networks. First, visual and parietal regions were coordinated with sensorimotor and premotor regions. Second, dorsal fronto parietal circuits dominated by sensorimotor and associative frontal striatal networks were incorporated. Finally, bilateral sensory-motor areas were coordinated with the cortico-cortical hemisphere between the left fronto parietal network and the visual areas. Our study argued that these networks reflect the processing of visual information, the emergence of visuomotor plans, and the processing of somatosensory responses or action outcomes. Thus, our study demonstrates that visuomotor integration exists in the dynamic reconfiguration of multiple cortico-cortical and cortico-subcortical FC networks. More generally, the approach demonstrates that optokinetic-related FC is unstable and shows task performance-related switching dynamics and regional flexibility on a time scale. In addition, our optokinetic-related FC has sparse connectivity with a density of 10%. Taken together, these findings shed light on the relationship between dynamic network reconfiguration and short-time executive function and provide a candidate start point for the better understanding of cognitive structure.

A vast number of multilayer network applications exists in bioinformatics, but the application of multilayer networks in any subfield of bioinformatics still relies on the acquisition and accumulation of bioinformatics data, and brain research is no exception. Therefore, interdisciplinary collaboration is a very efficient and necessary option. Brain structure and functionality are gradually understood, driven by brain data acquisition. According to these studies, the dynamic modeling of brain function by combining temporal dimensions is an effective means of study. Perhaps as research progresses, new data dimensions will be added (Wang et al., 2018, 2020; Wei et al., 2018a; Ding et al., 2019; Liu B. et al., 2019; Su et al., 2019b; Dao et al., 2020; Li J. et al., 2020; Lv et al., 2020).

CONCLUSION AND PERSPECTIVES

Multilayer (complex) networks have been an effective tool for studying complex problems in recent years and are currently being used in a variety of fields. As systems biology develops, multilayer networks are applicable to many aspects and research

areas within the field. Because of dataset availability, these networks are currently more often applied to genetics and brain research. However, as research progresses, it should become easier to unravel structural and functional fogs in biology on one hand, and on the other hand, research in this area will prove beneficial to the understanding of biological principles in general to better serve all people. In view of current research status, our review has presented the following ideas and prospects:

- (1) The development of biology is promoted by the joint development of various fields, and the application of multilayer networks in bioinformatics depends on the accumulation of biological data and the development of computer-related theories. Therefore, as an interdisciplinary subject, it needs the collaborative work of interdisciplinary experts.
- (2) Because of the complexity and dynamic change of biological systems, time-series multilayer networks with the addition of temporal information will have more and more applications in the simulation of dynamic processes in the study of genes, disease, drug discovery, and brain research.
- (3) Exploring the communication mode between tissue cells in the form of multi-layer network is to study the interaction (functionality) between structures on the basis of the network represented by the structure.

In addition to the structural and functional aspects of multilayer network research, methods to efficiently evaluate and assess the results of multilayer networks remains an important issue. The evaluation of the algorithmic complexity of multilayer networks has been proposed to assess if and when the multilayer representation of a system is qualitatively superior to classical single-layer aggregation network approaches (Wei et al., 2017a,b,c, 2018b, 2019; Su et al., 2019a, 2020a,b; Wang D. et al., 2021; Wang H. et al., 2021; Zhao et al., 2017).

AUTHOR CONTRIBUTIONS

YL contributed to conception and design of the study and wrote the first draft of the manuscript. SH organized the database. BG performed the statistical analysis. TZ, YL, and BG wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 62002087).

ACKNOWLEDGMENTS

We thank Steven M. Thompson, from LiwenBianji, Edanz Editing China (www.liwenbianji.cn/ac), for editing the English text of a draft of this manuscript.

REFERENCES

- Cao, S., Wang, Y., and Tang, Z. (2019). Adaptive elman model of gene regulation network based on time series data. *Curr. Bioinform.* 14, 551–561. doi: 10.2174/1574893614666190126145431
- Cheng, L., Hu, Y., Sun, J., Zhou, M., and Jiang, Q. (2018). DincRNA: a comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* 34, 1953–1956. doi: 10.1093/bioinformatics/bty002
- Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144.
- Dao, F. Y., Lv, H., Zhang, D., Zhang, Z. M., Liu, L., and Lin, H. (2020). DeepYY1: a deep learning approach to identify YY1-mediated chromatin loops. *Brief. Bioinform.* doi: 10.1093/bib/bbaa356 Online ahead of print.
- Ding, H., Luo, L. F., and Lin, H. (2011). Entropy production rate changes in lysogeny/lysis switch regulation of *Bacteriophage Lambda*. *Commun. Theor. Phys.* 55, 371–375. doi: 10.1088/0253-6102/55/2/31
- Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inform. Sci.* 418, 546–560. doi: 10.1016/j.ins.2017.08.045
- Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028
- Ding, Y., Tang, J., and Guo, F. (2020a). Identification of drug-target interactions via dual laplacian regularized least squares with multiple kernel fusion. *Knowl. Based Syst.* 204:106254. doi: 10.1016/j.knsys.2020.106254
- Ding, Y., Tang, J., and Guo, F. (2020b). Identification of drug-target interactions via fuzzy bipartite local model. *Neural Comput. Appl.* 23, 10303–10319. doi: 10.1007/s00521-019-04569-z
- Guo, F., Li, S. C., and Wang, L. (2011). Protein-protein binding sites prediction by 3D structural similarities. *J. Chem. Inform. Modeling* 51, 3287–3294. doi: 10.1021/ci200206n
- Guo, F., Li, S. C., Du, P., and Wang, L. (2014). Probabilistic models for capturing more physicochemical properties on protein-protein interface. *J. Chem. Inform. Modeling* 54, 1798–1809. doi: 10.1021/ci5002372
- Guo, F., Li, S. C., Ma, W., and Wang, L. (2013). Detecting protein conformational changes in interactions via scaling known structures. *J. Comput. Biol.* 20, 765–779. doi: 10.1089/cmb.2013.0069
- Hong, Z., Zeng, X., Wei, L., and Liu, X. (2020). Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics* 36, 1037–1043.
- Iliopoulos, A. C., Beis, G., Apostolou, P., and Papasotiriou, I. (2020). Complex networks, gene expression and cancer complexity: a brief review of methodology and applications. *Curr. Bioinform.* 15, 629–655. doi: 10.2174/1574893614666191017093504
- Jiang, Q., Hao, Y., Wang, G., Juan, L., Zhang, T., Teng, M., et al. (2010). Prioritization of disease microRNAs through a human phenome-microRNAome network. *BMC Syst. Biol.* 4(Suppl. 1):S2.
- Jiang, Q., Jin, S., Jiang, Y., Liao, M., Feng, R., Zhang, L., et al. (2017). Alzheimer's disease variants with the genome-wide significance are significantly enriched in immune pathways and active in immune cells. *Mol. Neurobiol.* 54, 594–600. doi: 10.1007/s12035-015-9670-8
- Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* 8, 282–293. doi: 10.1504/ijdm.2013.056078
- Jiang, Q., Wang, J., Wang, Y., Ma, R., Wu, X., and Li, Y. (2014). TF2LncRNA: identifying common transcription factors for a list of lncRNA genes from ChIP-Seq data. *Biomed. Res. Int.* 2014:317642.
- Jin, Q., Meng, Z., Tuan, D. P., Chen, Q., Wei, L., and Su, R. (2019). DUNet: a deformable network for retinal vessel segmentation. *Knowl. Based Syst.* 178, 149–162. doi: 10.1016/j.knsys.2019.04.025
- Jin, S., Zeng, X., Fang, J., Lin, J., Chan, S. Y., and Erzurum, S. C. (2019). applications, a network-based approach to uncover microRNA-mediated disease comorbidities and potential pathobiological implications. *NPJ Syst. Biol. Appl.* 5, 1–11.
- Jin, S., Zeng, X., Xia, F., Huang, W., and Liu, X. (2020). Application of deep learning methods in biological networks. *Brief. Bioinform.* doi: 10.1093/bib/bbaa043 Online ahead of print.
- Jing, F., Zhang, S.-W., and Zhang, S. (2019). Brief survey of biological network alignment and a variant with incorporation of functional annotations. *Curr. Bioinform.* 14, 4–10. doi: 10.2174/1574893612666171020103747
- Konda, A. K., Sabale, P. R., Soren, K. R., Subramaniam, S. P., Singh, P., Rathod, S., et al. (2019). Systems biology approaches reveal a multi-stress responsive WRKY transcription factor and stress associated gene co-expression networks in chickpea. *Curr. Bioinform.* 14, 591–601. doi: 10.2174/1574893614666190204152500
- Kumari, N., and Verma, A. (2020). Analysis of oncogene protein structure using small world network concept. *Curr. Bioinform.* 15, 732–740. doi: 10.2174/1574893614666191113143840
- Li, J., Pu, Y., Tang, J., Zou, Q., and Guo, F. (2020). DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides. *IEEE J. Biomed. Health Inform.* 24, 3012–3019. doi: 10.1109/jbhi.2020.2977091
- Li, Z., Zhang, T., Lei, H., Wei, L., Liu, Y., and Shi, Y. (2020). Research on gastric cancer's drug-resistant gene regulatory network model. *Curr. Bioinform.* 15, 225–234. doi: 10.2174/1574893614666190722102557
- Lin, H., Liang, Z. Y., Tang, H., and Chen, W. (2019). Identifying Sigma70 promoters with novel pseudo nucleotide composition. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 1316–1321. doi: 10.1109/tcbb.2017.2666141
- Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* 47:e127. doi: 10.1093/nar/gkz740
- Liu, B., Zhu, Y., and Yan, K. (2020). Fold-LTR-TCP: protein fold recognition based on triadic closure principle. *Brief. Bioinform.* 21, 2185–2193. doi: 10.1093/bib/bbz139
- Liu, G., and Jiang, Q. (2016). Alzheimer's disease CD33 rs3865444 variant does not contribute to cognitive performance. *Proc. Natl. Acad. Sci. U.S.A.* 113, E1589–E1590.
- Liu, G., Hu, Y., Han, Z., Jin, S., and Jiang, Q. (2019). Genetic variant rs17185536 regulates SIM1 gene expression in human brain hypothalamus. *Proc. Natl. Acad. Sci. U.S.A.* 116, 3347–3348. doi: 10.1073/pnas.1821550116
- Liu, G., Jin, S., Hu, Y., and Jiang, Q. (2018). Disease status affects the association between rs4813620 and the expression of Alzheimer's disease susceptibility gene TRIB3. *Proc. Natl. Acad. Sci. U.S.A.* 115, E10519–E10520.
- Liu, G., Zhang, F., Jiang, Y., Hu, Y., Gong, Z., Liu, S., et al. (2017). Integrating genome-wide association studies and gene expression data highlights dysregulated multiple sclerosis risk pathways. *Mult. Scler.* 23, 205–212. doi: 10.1177/1352458516649038
- Liu, L., Li, Q. Z., Jin, W., Lv, H., and Lin, H. (2019). Revealing gene function and transcription relationship by reconstructing gene-level chromatin interaction. *Comput. Struct. Biotechnol. J.* 17, 195–205. doi: 10.1016/j.csbj.2019.01.011
- Liu, L., Zhang, L. R., Dao, F. Y., Yang, Y. C., and Lin, H. (2021). A computational framework for identifying the transcription factors involved in enhancer-promoter loop formation. *Mol. Ther. Nucleic Acids* 23, 347–354. doi: 10.1016/j.omtn.2020.11.011
- Liu, M. L., Su, W., Wang, J. S., Yang, Y. H., Yang, H., and Lin, H. (2020). Predicting preference of transcription factors for Methylated DNA using sequence information. *Mol. Ther. Nucleic Acids* 22, 1043–1050. doi: 10.1016/j.omtn.2020.07.035
- Liu, X., Hong, Z., Liu, J., Lin, Y., Alfonso, R.-P., Zou, Q., et al. (2020). Computational methods for identifying the critical nodes in biological networks. *Brief. Bioinform.* 21, 486–497. doi: 10.1093/bib/bbz011
- Lv, H., Dao, F.-Y., Guan, Z.-X., Yang, H., Li, Y.-W., and Lin, H. (2020). Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. *Brief. Bioinform.* bbaa255. doi: 10.1093/bib/bbaa255
- Mortezaeefar, M., Fotovat, R., Shekari, F., and Sasani, S. (2019). Comprehensive understanding of the interaction among stress hormones signalling pathways by gene co-expression network. *Curr. Bioinform.* 14, 602–613. doi: 10.2174/1574893614666190226160742
- Shao, J., and Liu, B. (2020). ProtFold-DFG: protein fold recognition by combining Directed Fusion Graph and PageRank algorithm. *Brief. Bioinform.* doi: 10.1093/bib/bbaa192 Online ahead of print.

- Shao, J., Yan, K., and Liu, B. (2020). FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network. *Brief. Bioinform.* doi: 10.1093/bib/bbaa144 Online ahead of print.
- Sikandar, A., Anwar, W., and Sikandar, M. (2019). Combining sequence entropy and subgraph topology for complex prediction in protein protein interaction (PPI) network. *Curr. Bioinform.* 14, 516–523. doi: 10.2174/1574893614666190103100026
- Song, T., Rodríguez-Patón, A., Zheng, P., and Zeng, X. (2018). Spiking Neural P systems with colored spikes. *IEEE Trans. Cogn. Dev. Syst.* 10, 1106–1115. doi: 10.1109/tcds.2017.2785332
- Srivastava, N., Mishra, B. N., and Srivastava, P. (2019). In-silico identification of drug lead molecule against pesticide exposed-neurodevelopmental disorders through network-based computational model approach. *Curr. Bioinform.* 14, 460–467. doi: 10.2174/1574893613666181112130346
- Su, R., Liu, X., and Wei, L. (2020a). MinE-RFE: determine the optimal subset from RFE by minimizing the subset-accuracy-defined energy. *Brief. Bioinform.* 21, 687–698. doi: 10.1093/bib/bbz021
- Su, R., Liu, X., Wei, L., and Zou, Q. (2019a). Deep-resp-forest: a deep forest model to predict anti-cancer drug response. *Methods* 166, 91–102. doi: 10.1016/j.ymeth.2019.02.009
- Su, R., Liu, X., Xiao, G., and Wei, L. (2020b). Meta-GDBP: a high-level stacked regression model to improve anticancer drug response prediction. *Brief. Bioinform.* 21, 996–1005. doi: 10.1093/bib/bbz022
- Su, R., Wu, H., Xu, B., Liu, X., and Wei, L. (2019b). Developing a multi-dose computational model for drug-induced hepatotoxicity prediction based on toxicogenomics data. *IEEE ACM Trans. Comput. Biol. Bioinform.* 16, 1231–1239. doi: 10.1109/tcbb.2018.2858756
- Tao, Z., Li, Y., Teng, Z., and Zhao, Y. (2020). A method for identifying vesicle transport proteins based on LibSVM and MRMD. *Comput. Math. Methods Med.* 2020:8926750.
- Wang, D., Zhang, Z., Jiang, Y., Mao, Z., Wang, D., Lin, H., et al. (2021). DM3Loc: multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. *Nucleic Acids Res.* doi: 10.1093/nar/gkab016 Online ahead of print.
- Wang, G., Wang, Y., Feng, W., Wang, X., Yang, X. Y., Zhao, Y., Wang, Y., Liu, Y., et al. (2008). Transcription factor and microRNA regulation in androgen-dependent and -independent prostate cancer cells. *BMC Genomics* 9(Suppl. 2):S22.
- Wang, G., Luo, X., Wang, J., Wan, J., Xia, S., Zhu, H., et al. (2018). MedReaders: a database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* 46, D146–D151.
- Wang, G., Wang, Y., Teng, M., Zhang, D., Li, L., and Liu, Y. (2010). Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon gamma-stimulated HeLa cells. *PLoS One* 5:e11794. doi: 10.1371/journal.pone.0011794
- Wang, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt Independence criterion. *Neurocomputing* 383, 257–269. doi: 10.1016/j.neucom.2019.11.103
- Wang, H., Tang, J., Ding, Y., and Guo, F. (2021). Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment. *Brief. Bioinform.* doi: 10.1093/bib/bbaa409 Online ahead of print.
- Wang, L., Xuan, Z., Zhou, S., Kuang, L., and Pei, T. (2019). A Novel Model for predicting LncRNA-disease associations based on the LncRNA-MIRNA-disease interactive network. *Curr. Bioinform.* 14, 269–278. doi: 10.2174/1574893613666180703105258
- Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018a). Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* 117, 212–217.
- Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and promising identification of human micrnas by incorporating a high-quality negative set. *IEEE ACM Trans. Comput. Biol. Bioinform.* 1, 192–201. doi: 10.1109/tcbb.2013.146
- Wei, L., Tang, J., and Zou, Q. (2017a). Local-DPP: an improved DNA-binding protein prediction method by exploring local evolutionary information. *Inform. Sci.* 384, 135–144. doi: 10.1016/j.ins.2016.06.026
- Wei, L., Wan, S., Guo, J., and Wong, K. K. L. (2017b). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005
- Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019). Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE ACM Trans. Comput. Biol. Bioinform.* 16, 1264–1273. doi: 10.1109/tcbb.2017.2670558
- Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017c). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001
- Wei, L., Zhou, C., Chen, H., Song, J., and Su, R. (2018b). ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics* 34, 4007–4016.
- Yang, Q., Wu, J., Zhao, J., Xu, T., Han, P., and Song, X. (2020). The expression profiles of lncRNAs and their regulatory network during Smek1/2 knockout mouse neural stem cells differentiation. *Curr. Bioinform.* 15, 77–88. doi: 10.2174/1574893614666190308160507
- Zeng, X. X., Liu, L., Lu, L. Y., and Zou, Q. (2018). Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 34, 2425–2432. doi: 10.1093/bioinformatics/bty112
- Zeng, X., Song, X., Ma, T., Pan, X., Zhou, Y., and Hou, Y. (2020). Repurpose open data to discover therapeutics for COVID-19 using deep learning. *J. Proteome Res.* 19, 4624–4636. doi: 10.1021/acs.jproteome.0c00316
- Zeng, X., Zhang, X., and Zou, Q. (2016). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinform.* 17, 193–203. doi: 10.1093/bib/bbv033
- Zhai, Y., Chen, Y., Teng, Z., and Zhao, Y. (2020). Identifying antioxidant proteins by using amino acid composition and protein-protein interactions. *Front. Cell Dev. Biol.* 8:591487.
- Zhang, D., Chen, H.-D., Zulfikar, H., Yuan, S.-S., Huang, Q.-L., Zhang, Z.-Y., et al. (2021). iBLP: an XGBoost-based predictor for identifying bioluminescent proteins. *Comput. Math. Methods Med.* 2021:6664362.
- Zhang, J., Feng, J., and Wu, F.-X. (2020). Finding community of brain networks based on neighbor index and DPSO with dynamic crossover. *Curr. Bioinform.* 15, 287–299. doi: 10.2174/1574893614666191017100657
- Zhang, W., Li, W., Zhang, J., and Wang, N. (2019). Data integration of hybrid microarray and single cell expression data to enhance gene network inference. *Curr. Bioinform.* 14, 255–268. doi: 10.2174/1574893614666190104142228
- Zhang, Z. M., Tan, J. X., Wang, F. Y., Dao, F. Y., Zhang, Z. Y., and Lin, H. (2020). Early diagnosis of hepatocellular carcinoma using machine learning method. *Front. Bioeng. Biotechnol.* 8:254.
- Zhang, Z. Y., Yang, Y. H., Ding, H., Wang, D., Chen, W., and Lin, H. (2021). Design powerful predictor for mRNA subcellular location prediction in *Homo sapiens*. *Brief. Bioinform.* 22, 526–535. doi: 10.1093/bib/bbz177
- Zhao, X., Chen, L., Guo, Z.-H., and Liu, T. (2019). Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinform.* 14, 709–720. doi: 10.2174/1574893614666190220114644
- Zhao, Y., Wang, F., Chen, S., Wan, J., and Wang, G. (2017). Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network. *Biomed Res. Int.* 2017:7049406.
- Zhu, L., Su, F., Xu, Y., and Zou, Q. (2018). Network-based method for mining novel HPV infection related genes using random walk with restart algorithm. *BBA Mol. Basis Dis.* 1864, 2376–2383. doi: 10.1016/j.bbdis.2017.11.021
- 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.
- Zulfikar, H., Masoud, M. S., Yang, H., Han, S.-G., Wu, C.-Y., and Lin, H. (2021). Screening of prospective plant compounds as HIR and CLIR inhibitors and its antiallergic efficacy through molecular docking approach. *Comput. Math. Methods Med.* 2021:6683407.

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.

Copyright © 2021 Lv, Huang, Zhang and Gao. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Enhancement and Imputation of Peak Signal Enables Accurate Cell-Type Classification in scATAC-seq

Zhe Cui^{1†}, Ya Cui^{2†}, Yan Gao¹, Tao Jiang¹, Tianyi Zang^{1*} and Yadong Wang^{1*}

¹ Centre for Bioinformatics, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China,

² College of Life Science, University of Chinese Academy of Sciences, Beijing, China

OPEN ACCESS

Edited by:

Fa Zhang,
Chinese Academy of Sciences (CAS),
China

Reviewed by:

Bingqiang Liu,
Shandong University, China
Quan Zou,
University of Electronic Science
and Technology of China, China

*Correspondence:

Tianyi Zang
tianyi.zang@hit.edu.cn
Yadong Wang
ydwang@hit.edu.cn

[†]These authors have contributed
equally to this work

Specialty section:

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

Received: 25 January 2021

Accepted: 22 February 2021

Published: 06 April 2021

Citation:

Cui Z, Cui Y, Gao Y, Jiang T,
Zang T and Wang Y (2021)
Enhancement and Imputation of Peak
Signal Enables Accurate Cell-Type
Classification in scATAC-seq.
Front. Genet. 12:658352.
doi: 10.3389/fgene.2021.658352

Single-cell Assay Transposase Accessible Chromatin sequencing (scATAC-seq) has been widely used in profiling genome-wide chromatin accessibility in thousands of individual cells. However, compared with single-cell RNA-seq, the peaks of scATAC-seq are much sparser due to the lower copy numbers (diploid in humans) and the inherent missing signals, which makes it more challenging to classify cell type based on specific expressed gene or other canonical markers. Here, we present svmATAC, a support vector machine (SVM)-based method for accurately identifying cell types in scATAC-seq datasets by enhancing peak signal strength and imputing signals through patterns of co-accessibility. We applied svmATAC to several scATAC-seq data from human immune cells, human hematopoietic system cells, and peripheral blood mononuclear cells. The benchmark results showed that svmATAC is free of literature-based markers and robust across datasets in different libraries and platforms. The source code of svmATAC is available at <https://github.com/mrcuizhe/svmATAC> under the MIT license.

Keywords: scATAC-seq, classification, machine learning, support vector machine, cell-type annotation

INTRODUCTION

With the technological progress in Single-cell Assay Transposase Accessible Chromatin sequencing (scATAC-seq) (Buenrostro et al., 2013), which has overcome the previous limitations and is able to generate thousands of single cells chromatin accessibility data at lower cost (Chen et al., 2019), a certain number of scATAC-seq datasets have been sequenced with different techniques in diverse libraries. For example, the Chromium Single Cell ATAC technology from 10X genomics (10X Genomic, 2020) can profile hundreds to tens of thousands of nuclei in one chip and finish the process from sample to sequencing-ready library in 1 day. For single-cell RNA-sequencing (scRNA-seq) and scATAC-seq data, the processing steps typically start with unsupervised clustering cells from coordinate-based peak matrix and then identify cell types from clustered groups. Thus, many methods requiring a training dataset labeled with corresponding cell populations for classifier training have been developed to get rid of the requirement of prior knowledge in scRNA-seq (Kiselev et al., 2018; Lieberman et al., 2018; Lopez et al., 2018; Boufea et al., 2019; Johnson et al., 2019; Ma and Pellegrini, 2019; Tan and Cahan, 2019). Support vector machine (SVM) performs

the best among machine learning methods for classifying cell types in scRNA-seq (Abdelaal et al., 2019), and a lot of SVM-based tools have been proved effective and efficient (Pedregosa et al., 2011; Alquicira-Hernandez et al., 2019). However, the low copy number of DNA molecule in a cell results in only 1–10% of the accessible peaks in scATAC-seq being detectable, while the percentage for expressed genes detected in scRNA-seq is about 10–45% (Liu et al., 2019; Mereu et al., 2020). When clustering in scATAC-seq, such severe signal loss in a massive sparse space makes it more challenging to annotate cluster groups through gene-related canonical markers, which is practical and well-received in scRNA-seq. This missing of signal makes the SVM with linear kernel hard to work (Stewart et al., 2018) because this method starts with dimensionality reduction and feature selection, which is largely dependent on the accuracy and integrity of the dataset. Even so, SVM still outperformed other popular machine learning methods on cell-type classification of scATAC-seq (Cui et al., 2020), though the classification results of these methods (including SVM) are all performing at a low level. Since the signal missing will affect the quality of feature selection and then affects the construction of the classification model, the data recovery and signal strength enhancement are essential for SVM-based methods in scATAC-seq datasets (Yan et al., 2020).

Statistical methods such as imputing dropouts and correcting excess zero-counts have already been applied to scATAC-seq datasets, and this type of enhancing and recovering of missing signals has shown great power for downstream analysis. SCALE (Xiong et al., 2019) constructs a probabilistic Gaussian Mixture Model to characterize data, followed by denoising and imputing missing values in clustered subgroups. scOpen (Li et al., 2019) recovers the dropout signal in a particular cell using positive-unlabeled learning. However, these methods basically are using the statistic-based model, which may require an extra prior knowledge or time-consuming globally statistics. Since the repertoire of accessible regulatory elements in cell lines or tissues is unique, this type of data imputation is then considered as a kind of molecular signature for identifying. For example, Cicero (Pliner et al., 2018) is able to predict cis-regulatory DNA interactions through scATAC-seq from a single experiment.

Here, we present svmATAC, an automatic cell classification SVM-based method for scATAC-seq data. svmATAC enhances the data from cluster/group data first, followed by imputing the signal linkage according to the co-accessibility scores from Cicero. The enhanced and imputed data will then be input to SVM (linear kernel) classifier for model training and cell-type prediction (Figure 1). We applied svmATAC to several typical scATAC-seq datasets containing different cell types, including human immune cell (hereafter Corces2016) (Corces et al., 2016), human hematopoietic system cell (hereafter Buenrostro et al., 2018), and peripheral blood mononuclear cell (hereafter 10 × PBMCs) (Genomic, 2020), to evaluate its classification ability. With fivefold cross-validation, svmATAC showed a great advance on prediction accuracy and surpassed 7.13–21.34% compared to SVM (linear kernel). In inter-dataset experiments, svmATAC also maintained great predictive power to accurately and quickly identify cell types based on a pre-trained model. We

believe that svmATAC has great potential to handle complex cell-type identification problems in practical and realistic scenarios.

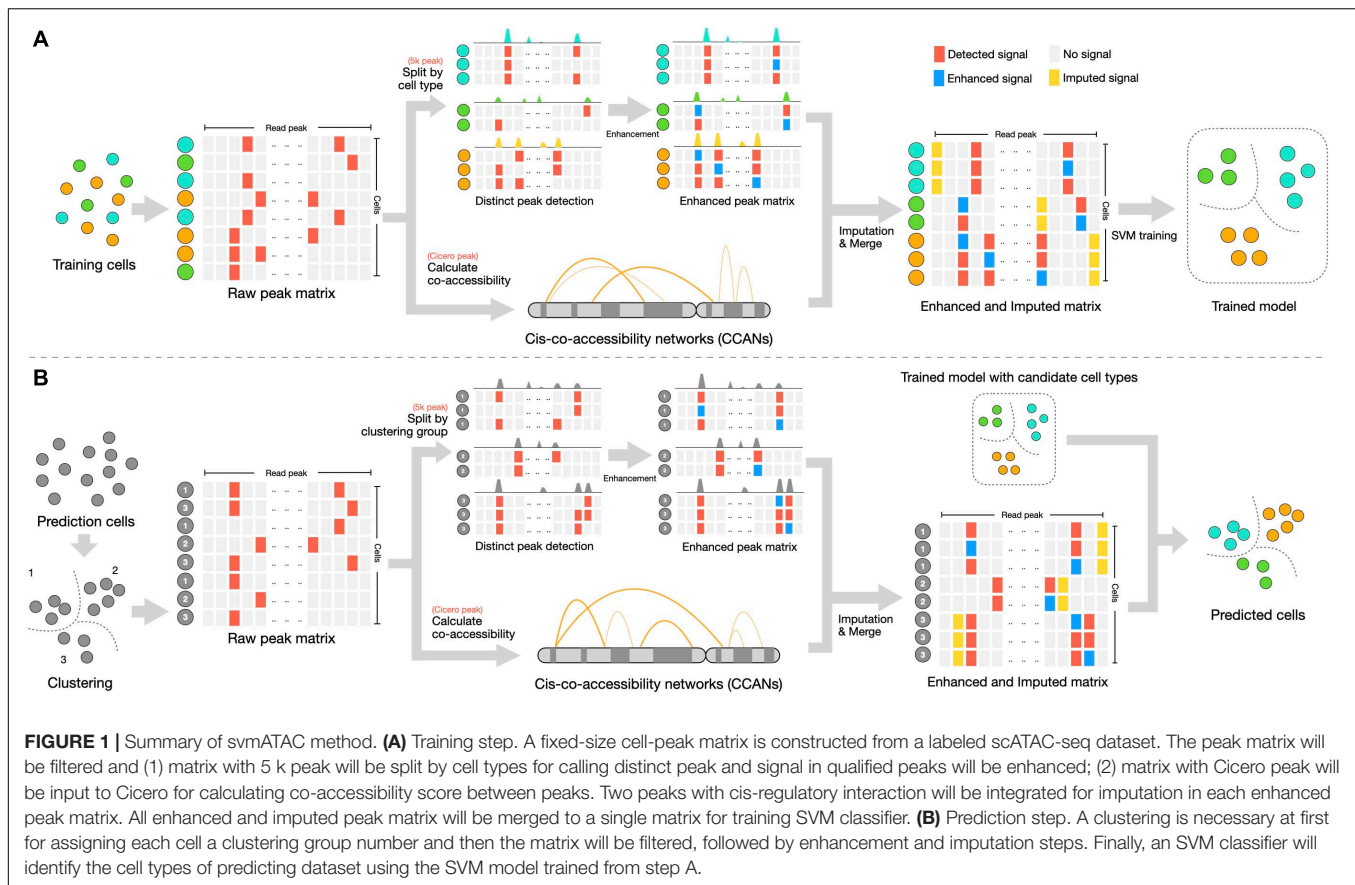
RESULTS

svmATAC as a General Framework for Classification of scATAC-seq

svmATAC applies two pivotal functions, i.e., group-based read signal enhancement and cis-regulatory relationship-based imputation to cell-peak matrix, followed by training model and predicting cell types using SVM classifier (Figure 1). With this specific design, the peak signals of scATAC-seq are strengthened and related by extra biological connections, which improves the feature selection in lower dimensional space. svmATAC consists of three main steps: (1) It applies a specific design enhancement method to establish cell-peak matrix. The peak value 0 will be set to 1 when the peak (column) signal rate is larger than prior knowledge cutoff in a cell-type/cluster group. This step is able to correctly classify some of the cell types (Supplementary Tables 1–10), compared to directly using raw dataset, but it is still not good enough. (2) An imputation method, i.e., Cicero, is applied to construct the cis-regulatory relationship between peaks and to compute the co-accessibility scores. Two peaks of a cell-type/cluster group will be integrated for imputation when its co-accessibility score ≥ 0.25 (Pliner et al., 2018). That is, the value 1 will be assigned for both peaks if any one peak is distinct. (3) The cell-peak matrix processed by the two pivotal functions will be used as input for an SVM classifier to perform model training. With the trained SVM model, svmATAC can achieve the final prediction of cell types in unlabeled dataset. In order to give a comprehensive evaluation on the performance of svmATAC, we, respectively, designed an intra-dataset experiment and an inter-dataset experiment as below.

Benchmark Results on Intra-Dataset Experiments

We evaluated the performance of svmATAC in an intra-dataset experiment by applying a fivefold cross-validation across each dataset after cell filtering. We randomly divided all the cells into fivefold with equal proportions of each cell population in each fold. The first and smallest dataset we used is from the human immune cells (hereafter Corces2016). This dataset consists of 576 immune cells from four isolated cell populations including leukemic blasts (Blast), lymphoid-primed multipotent progenitors (LMPP), leukemia stem cells (LSC), and monocytes. The gold standard labels we used here are from the original paper and predicted by enhancer cytometry. Compared to the SVM (linear kernel), we found an improvement on the predicted results when using svmATAC. The percentage of correctly predicted cells in all populations are all increased by at most 19.79% (from 75 to 94.79%) in monocyte (Figure 2A); the F1 scores are also improved in all population with monocyte increased the most from 0.85 to 0.97 (Figure 3A). The details for confusion matrix and F1 score list for Corces2016 are presented in Supplementary Tables 1, 2.



The second dataset we used is from the human hematopoietic system, which consists of 2,034 labeled hematopoietic cells from 10 cell populations including hematopoietic stem cells (HSC), multipotent progenitors (MPP), lymphoid-primed multipotent progenitors (LMPP), common myeloid progenitors (CMP), granulocyte-macrophage progenitors (GMP), GMP-like cells, megakaryocyte-erythroid progenitors (MEP), common lymphoid progenitors (CLP), monocytes (mono), and plasmacytoid dendritic cells (pDC). In order to test the ability of identifying the cells from different batches, we divided the LMPPs into two groups: LMPP-O: generated and first published in Corces2016; LMPP: newly generated and first published in Buenrostro2018. We used the FACS-sorting labels as the gold standard for this dataset. All cells in this dataset are correctly classified using svmATAC. Similar to the results on Corces2016, the percentage of correctly predicted cells in all population are increased by at most 86% in MPP (**Figure 2B**), and the F1 scores are also improved in all populations, with MPP increased the most from 0.25 to 1 (**Figure 3B**), compared to SVM (linear kernel). The details for confusion matrix and F1 score list for Buenrostro2018 are presented in **Supplementary Tables 3, 4**.

The last two datasets we used are from the peripheral blood mononuclear cells. These two datasets were generated from the same healthy donor but prepared in different libraries. In total, there are 3,917 cells profiled in 10× PBMCs v1 dataset and

4,585 cells were profiled in 10× PBMCs Next Gem dataset but both datasets are unlabeled. Based on recent studies (Bravo González-Blas et al., 2019; Pliner et al., 2019), we expected eight populations in each dataset, so we clustered cells into eight groups and use these cluster IDs as the gold label for training and testing (**Supplementary Figures 1, 2**). However, though cells with the same cluster ID may be predicted together into one group, we cannot check whether these predicted cell-types are true positives when only cluster ID is available. Thus, we assigned cell types using Seurat v3 (Stuart et al., 2019) based on a labeled scRNA-seq dataset from the same sample and then selected the high-confidence labels as gold standard for scATAC-seq datasets. We totally labeled 2,927 cells for the 10× PBMCs v1 dataset and 3,670 cells for the 10× PBMCs Next Gem dataset. For the Seurat labeled 10× PBMCs v1 dataset, the percentage of correctly predicted cells in each population increases to 100%, while $CD8^+$ T, DC, and $FCGR3A^+$ Mono are barely correctly identified at first using SVM (linear kernel) (**Figure 2C**); the F1 scores also improved in all populations, notably from 0 to 1 in $CD8^+$, 0.09–1 in DC, and 0–1 in $FCGR3A^+$ Mono (**Figure 3C**), compared to SVM (linear kernel). For the Seurat labeled 10× PBMCs Next Gem dataset, the percentage of correctly predicted cells in all population increases by at most 91% in $CD8^+$ T (**Figure 2D**); the F1 scores also improved and $CD8^+$ T increased the most from 0.16 to 1 (**Figure 3D**). The details for confusion matrix and F1 score list for the Seurat labeled 10× PBMCs v1 dataset and the

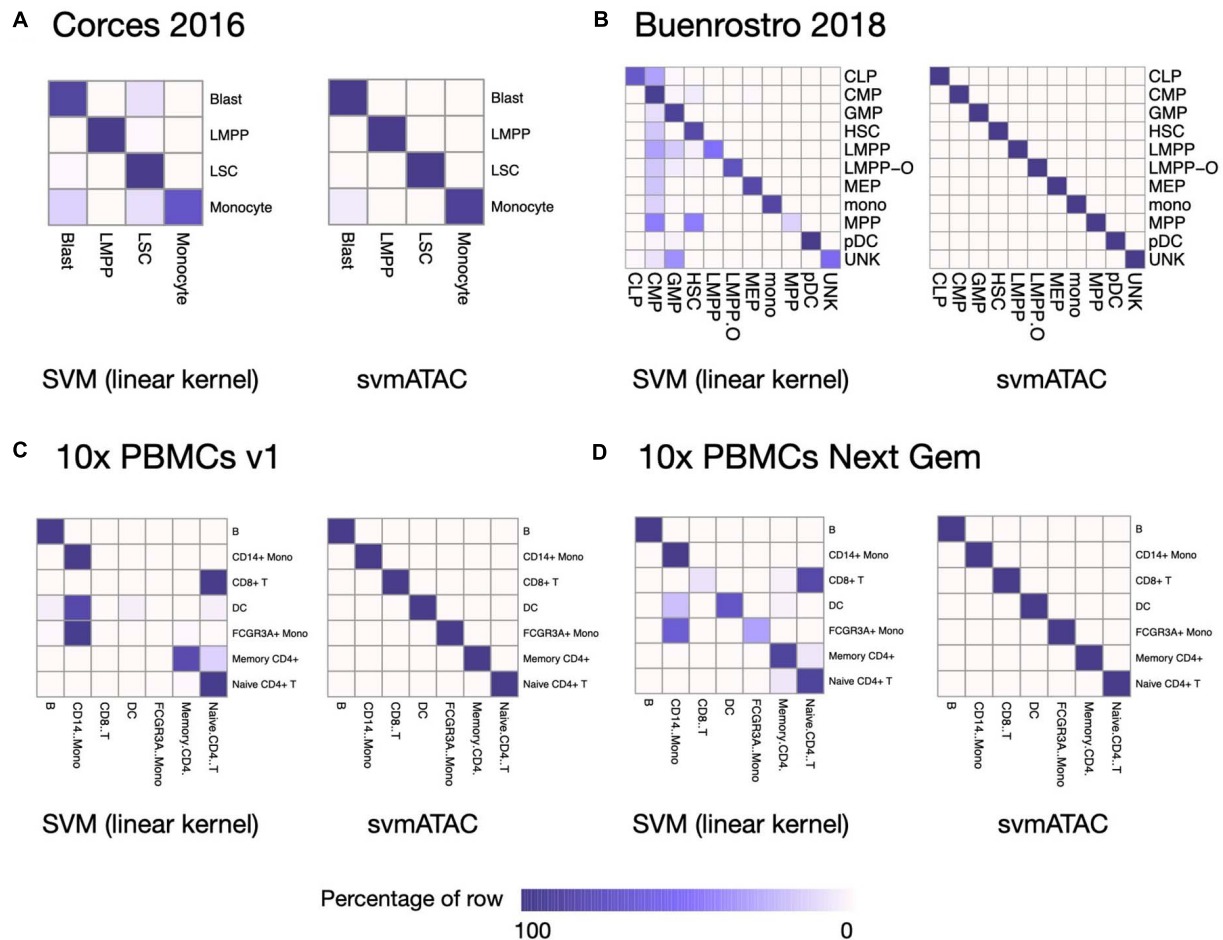


FIGURE 2 | Heatmap comparing the SVM (linear kernel) and svmATAC predicted cells versus true label of intra-dataset experiment. **(A)** The experiment on Corces2016. In monocyte, the percentage of correctly predicted cells by svmATAC is increased the most, by 19.79% (from 75 to 94.79%), while the percentage of LSC is increased the least, by only 0.52% (from 98.96 to 99.48%), compared to SVM (linear kernel). **(B)** The experiment on Buenrostro2018. All cells are correctly classified by svmATAC, and the percentage of correctly predicted cells in all population increase by at most 86% in MPP, compared to the SVM (linear kernel). **(C)** The experiment on 10× PBMCs v1. All cells are correctly classified by svmATAC. The cells of CD8⁺ T and FCGR3A⁺ Mono, which are totally incorrectly classified by the SVM (linear kernel), are all correctly classified by svmATAC. **(D)** The experiment on 10× PBMCs Next Gem. All cells are correctly classified by svmATAC. The cells of CD8⁺ T, DC, and FCGR3A⁺ Mono, most of which are incorrectly classified by the SVM (linear kernel), are all correctly classified by svmATAC. Colors represent the percentages of cells of a specific reported type labeled as each type by svmATAC.

Seurat labeled 10 × PBMCs Next Gem dataset are presented in **Supplementary Tables 5–8**.

Benchmark Results on Inter-Dataset Experiments

In order to evaluate the ability of svmATAC to control or even overcome the deviation between different datasets such as batch effect, tissue type, and other technical factors, we designed the inter-dataset experiment, in which two datasets are generated from the same tissue, but prepared in different libraries and sequenced from different platforms.

We used Seurat labeled 10× PBMCs v1 to train a model first and then classify the labels of 10× PBMCs Next Gem based on this model. We compared the predicted labels with

Seurat labels to evaluate the performance of svmATAC, and we found that although the model of the v1 dataset was trained on sparser molecular data from a different method and instrument, svmATAC is robust, performing well across datasets, and capable of overcoming batch effect and technical bias.

svmATAC accurately classified 99.95% (3,668 out of 3,670) cells in the 10× PBMCs Next Gem dataset (**Supplementary Tables 9, 10**), compared to 47.96% using SVM (linear kernel) (**Figure 4A**). We also notice that all cells in the 10× PBMCs Next Gem dataset are correctly classified by svmATAC, even though the cells of CD8⁺ T and FCGR3A⁺ Mono are barely correctly classified when using SVM (linear kernel). Therefore, the F1 scores for all populations in svmATAC are all improved and CD8⁺ T and FCGR3A⁺ Mono increase the most by 0.996 (from 0 to 0.996) and 0.96 (from 0.04 to 1), respectively (**Figure 4B**).

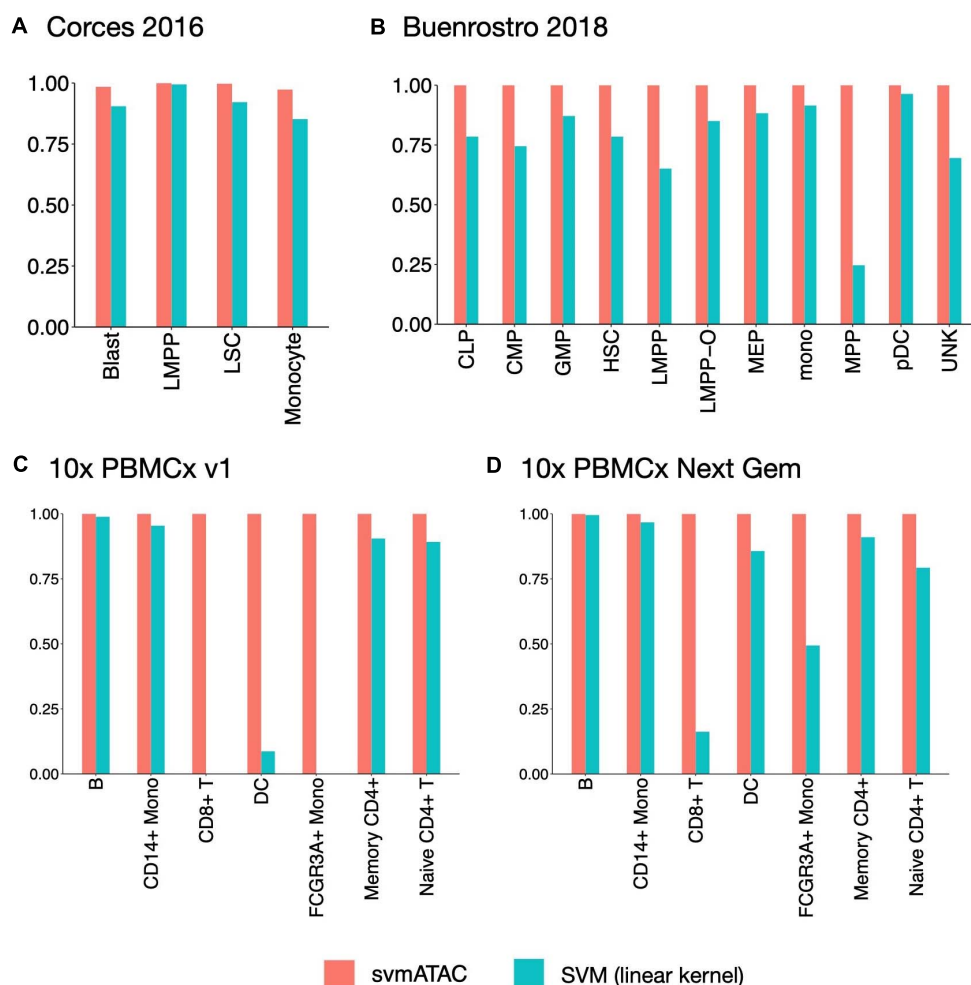


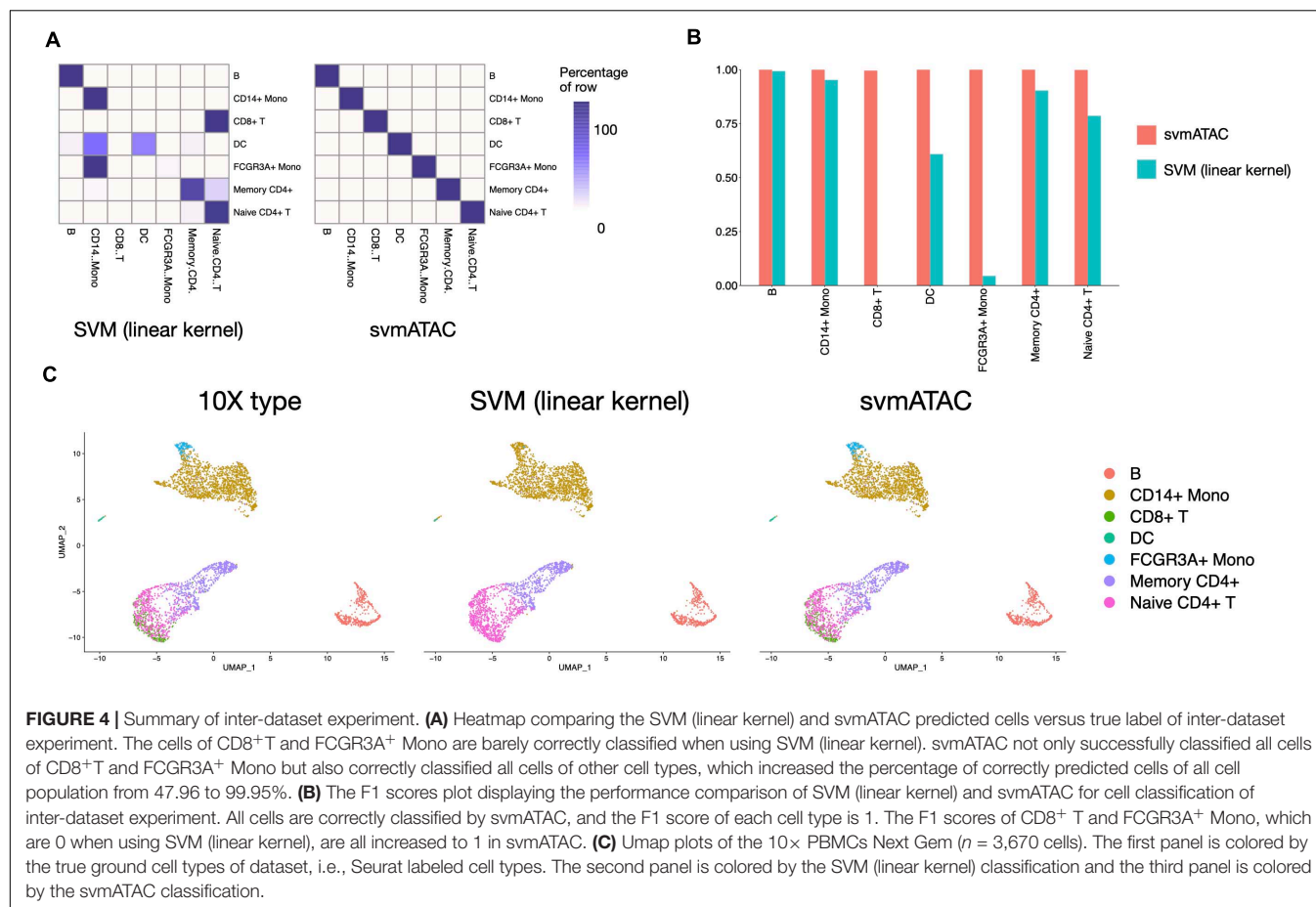
FIGURE 3 | The F1 scores plot showing the performance comparison of SVM(linear kernel) and svmATAC for cell classification of intra-dataset experiment. **(A)** The experiment on Corces2016. svmATAC performed best on LMPP and its F1 score is 1 and the F1 score of monocyte increased the most, by 0.12 (from 0.85 to 0.97), compared to SVM (linear kernel). **(B)** The experiment on Buenrostro2018. The F1 scores of all cell types are 1 for svmATAC, which means that all cells are correctly classified and the F1 scores of all populations are increased by at most 0.75 (from 0.25 to 1) in MPP, compared to SVM (linear kernel). **(C)** The experiment on Seurat labeled 10x PBMCs v1. All cells are correctly classified by svmATAC and the F1 score of each cell type is 1. The F1 scores of CD8⁺ T and FCGR3A⁺ Mono, which are 0 when using SVM (linear kernel), are all increased to 1 for svmATAC. **(D)** The experiment on Seurat labeled 10x PBMCs Next Gem. All cells are correctly classified by svmATAC and the F1 score of each cell type is 1. The F1 scores of CD8⁺ T and FCGR3A⁺ Mono increased most by 0.84 (from 0.16 to 1) and 0.51 (from 0.49 to 1) when using SVM (linear kernel), compared to SVM (linear kernel). The red panel represents the results for svmATAC, and the blue panel represents the results for SVM (linear kernel) on unenhanced and imputed data.

We next investigated qualitatively the obtained classification results, using the respective feature matrices to project the cells onto a 2-D space using UMAP (McInnes et al., 2018) and colored them based on the obtained classification results or the gold standard labels. We found a high distribution consistency between true labels and svmATAC classified labels (**Figure 4C**), while SVM (linear kernel) misclassified most of the cells into two similar cell groups. Because of the close spatial distribution in lower-dimensional feature space, SVM (linear kernel) misclassified almost all cells of FCGR3A⁺ Mono and CD8⁺ T to CD14⁺ Mono and Naive CD4⁺ T, respectively. svmATAC not only successfully classified the almost all cells of these

two cell types but also correctly classified all cells of other cell types.

DISCUSSION

Single-cell ATAC sequencing is a new technology in the area of the chromatin accessibility profile of individual cells and gives a new perspective of the identification and characterization of cell types (Cusanovich et al., 2015). Here, we introduced svmATAC, a specially designed method for scATAC-seq data to classify single cells based on readout enhancement, imputation, and a SVM model. The benchmark results show that svmATAC is able to



accurately classify cells in both intra- and inter-datasets. The outstanding achievements of svmATAC are mainly due to its two pivotal modules: (1) the peak signal enhancement can overcome the disadvantage of read loss by sequencing technology; (2) the biological cis-regulatory relationship-based imputation can establish connections between significant regions.

However, there are still a few shortcomings for svmATAC that cannot be ignored. (1) In the current version of svmATAC, the accuracy and sensitivity of cell-type classification are highly relying on the manually selected cutoff for enhancement and imputation, which does exist a gap for applying svmATAC to more complex scATAC-seq datasets. We will develop an automatic cutoff adjustment for svmATAC in the future. (2) We also notice that a certain number of noisy read signals are added by mistake to the enhancement and imputation processes and decreases the performance especially in the inter-dataset experiment. This is another point for future work about how to avoid adding useless signal in enhancement and imputation steps. (3) Although svmATAC shows its potential on overcoming the batch effect on inter-dataset experiments using 10× datasets, we still expect more datasets coming from the same tissue or sample but generated through different sequencing pipelines.

Moreover, svmATAC also supports the user-defined classification model from all kinds of machine learning algorithms, which has great potentials in the adaptability

in various scATAC-seq datasets. Therefore, svmATAC is a promising approach and benefits cutting-edge genomic studies.

MATERIALS AND METHODS

Construction of Cell-Peak Matrix

Several region definitions for cell-peak matrix have been broadly used (Chen et al., 2019), including peaks on bulk data or aggregate single-cell data, pseudo-bulk data, regions around insertion sites, and fixed-size bins. The regions from bulk or aggregate scATAC-seq data are based on peak calling, and this process only keeps those areas covered by at least one read. The pseudo-bulk clades created by hierarchical clustering is different in the way of calling peaks, but the peaks are still generated from sequencing data. These regions around insertion sites do not rely on calling peaks from sequencing reads; however, this kind of peak region still only covers a part of the whole genome reference. These types of regions selection may be suitable for the developer's application scenarios, such as clustering the cells into groups but cannot fulfill the requirement of svmATAC. This is because one of the most common scenarios for svmATAC is to predict the cell types for a dataset using a pre-trained classifier, which requires the two datasets used in training and predicting to

share the same peak regions to ensure the compatibility of selected features.

We generated two types of cell-peak matrix containing different peak regions. One peak region is applying fixed-size peak regions (hereafter 5 k peak) for the training and predicting of the classifier process, in which we detected the read signal every 5,000 bp and therefore split the whole genome reference(hg19) into more than 600 k pieces. Note that some other tools may filter out the peaks with no read signal detected for saving memory and computing time, and we kept all the peaks here to make sure all regions for training data and predicting data are the same for compatibility of data structure. The other peak region is designed for Cicero; it is because we found that Cicero cannot process matrix with large regions spanning too large, such as 5 k here. We obtained a much smaller peak region (hereafter Cicero peak) from published data or bulk ATAC-seq data for matrix construction. This data matrix is only used for computing the co-accessibility score in the imputation process.

For the Coces2106 dataset, we first downloaded it from the NCBI database (GSE74310) and aligned it to hg19 using BWA-MEM (version 0.7.17-r1188) (Li, 2013) and enabled Picard (Broad_Institute, 2019) and Samtools (version 1.9) (Li et al., 2009) to remove the duplicated reads. Only duplication remove is applied to 10× PBMCs v1 and 10× PBMCs Next Gem dataset because these datasets are obtained in bam format from https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_pbmc_5k_v1 and https://support.10xgenomics.com/single-cell-atac/datasets/1.2.0/atac_pbmc_5k_nextgem, respectively. These two 10× PBMCs datasets are downloaded with only cluster group ID available, but no true cell label was provided; we assigned labels to each cell by Seurat v3 as it can convincingly assign labels for scATAC-seq data when its scRNA-seq and labels are available. The peak and count file of Buenrostro2018 is available at GSE96769, and we obtained the aligned data from https://github.com/pinellolab/scATAC-benchmarking/tree/master/Real_Data/Buenrostro_2018.

Based on aligned and duplication-removed data and the cell labels provided in the datasets, we then estimated read coverage for each peak to build a cell-peak binary count matrix, in which each value 1 or value 0 represents whether a read signal was detected from the cell in this bin (1) or not (0). There is no limit to the number of cell types or the number of cells. Peaks that overlap ENCODE-defined blacklist regions are all set to zero. Cell populations with a size smaller than 10 were filtered. Note that for both kinds of peak region (5 k or Cicero), we did not filter out columns when all values are 0, which could be a kind of feature of classifier training.

Each cell matrix is represented in a compressed, sparse, column-oriented numeric matrix (dgCMatrx class in R). All these matrices are stored in RDS files and publicly available at <https://github.com/mrcuizhe/svmATAC>.

Signal Strength Enhancement

The massive loss of read signal in scATAC-seq leads to incorrect zero counts of the cell-peak matrix, which may influence the training and prediction of the SVM classifier. Recovering the loss signal in data is a popular and workable way to strengthen

the classification ability of machine learning classifiers, and this method has already been broadly accepted and developed in scRNA-seq data analysis, whose loss rate is a quarter less than that of scATAC-seq.

The enhancement process in svmATAC is trying to recover the inherent loss signal caused by sequencing techniques or experimental bias and then enhance the peak signal strength of each group. The enhancement procedure is a group-based step, in which data must be first divided into several groups based on its cell labels or clustering group numbers.

We first separated the cell-peak matrix by cell types into an $n \times m$ matrix by cell types, i.e., a data matrix with n cells and m peaks. Then, we enhanced the read signal by recovering the missing signal using the following formula:

- When the fraction of non-zero cells of the i_{th} peak is larger than the cutoff for enhancement (i.e., $c_{enh} \leq \frac{\sum_{j=1}^n C_{i,j}}{n} \leq 1$), we will treat all counts in the i_{th} peak as follows:

$$S_i = [C_{i,1}, \dots, C_{i,j}, \dots, C_{i,n}], C_{i,j} = 1 \quad (1)$$

where S_i represents the read count for the i_{th} column (peak) in cell-peak matrix, $C_{i,j}$ represents the read count of the j_{th} cell in the i_{th} column in matrix and $i \in [1, m]$, $j \in [1, n]$. c_{enh} represents the cutoff for enhancement, and we recommend 0.1 here based on the read loss rate of scATAC-seq (Mereu et al., 2020; Liu et al., 2019) and experiment results (Supplementary Tables 1–10), which also shows that the enhancement step is efficient and necessary on scATAC-seq data for cell-type classification.

- When the percentage of non-zero cells of a peak is less than the cutoff for enhancement (i.e., $0 \leq \frac{\sum_{j=1}^n C_{i,j}}{n} \leq c_{enh}$), we will not change S_i and keep it intact.

Signal Imputation

Apart from the enhancement of read signal, another way frequently applied in scATAC-seq data analysis is imputing read signal based on iconic biomarkers or biologic relationships, which may benefit the selection of features for each cell type. The imputation in svmATAC is also group-based and includes two steps:

- Compute the co-accessibility score for every two peaks. Co-accessibility scores represent the patterns and linkages of co-accessible pairs of DNA elements, such as distal elements and promoters. We use Cicero (v3.11, with default parameters) here to compute the co-accessibility scores for every two peaks. The co-accessibility score of each two peaks ranges from 0 to 1, indicating the strength of Cicero co-accessibility links. Scores closer to 1 indicate that two elements (peaks) are more co-accessible and vice versa.
- Imputing read signal based on cis-regulatory relationship into each group from co-accessibility score. Two peaks from enhanced data matrix will be considered as significantly connected if its co-accessibility score is higher than a threshold value. We first separated the enhanced cell-peak matrix by cell types into an $n \times m$ matrix, i.e., a

data matrix with n cells and m peaks; then, all Cicero-linked peaks will be integrated for imputation using the following formula:

When the Cicero co-accessibility score for the linkage between the i_{th} peak and k_{th} peak is higher than the cutoff for imputation and there is no zero cell for the k_{th} peak (i.e., $L_{ik} \geq c_{int}$ and $\sum_{j=1}^n C_{k,j} = n$), we will treat all counts in the i_{th} peak as follows:

$$S_i = [C_{i,1}, \dots, C_{i,j}, \dots, C_{i,n}] \quad C_{i,j} = 1 \quad (2)$$

where S_i represents the i_{th} column (peak) in cell-peak matrix, $C_{i,j}$ represents the read count of the j_{th} cell in the i_{th} peak in matrix and $i, k \in [1, m], j \in [1, n]$. c_{int} represents the cutoff for co-accessibility score, and we recommend 0.25 here based on the prior knowledge from the Cicero paper and experiment results (**Supplementary Tables 1–10**), which also shows that the enhancement step is efficient and necessary on scATAC-seq data for cell-type classification. L_{ik} represents the Cicero co-accessibility score for the linkage between the i_{th} peak and the k_{th} peak. When either the Cicero co-accessibility score for the linkage between the i_p peak and k_p peak is lower than the cutoff for imputation or there is more than one non-zero cell for the k_{th} peak (i.e., $L_{ij} < c_{int}$ or $\sum_{j=1}^n C_{k,j} \neq n$), we will not change the count value in the i_{th} peak and keep S_i intact.

Note that the matrix for computing Cicero co-accessibility score is based on Cicero peaks, which is different from the 5 k peak used for enhanced data matrix. Only the first (leftmost) 5 k peak will be considered for imputation if a peak from Cicero peaks is overlapped with multiple 5 k peaks. All imputed matrixes should be merged back into one matrix by columns(peaks) for downstream training and predicting.

Classifier Training and Predicting

The enhanced and imputed cell-peak matrix will be used as input for SVM to train a classifier, and the trained classifier will then be used to predict cell types in an unlabeled dataset. We totally designed two types of experiments including intra-dataset and inter-dataset for evaluating the performance and adjusting the parameters in svmATAC.

In intra-dataset experiments, we performed a fivefold cross-validation on four datasets, including Corces2016, Buenrostro2018, 10× PBMCs v1, and 10× PBMCs Next Gem, to evaluate the classification ability of svmATAC. The folds were divided in a stratified manner to keep equal proportions of each cell population in each fold. The training and testing folds were same for all methods.

To evaluate the performance of svmATAC in more realistic scenarios (batch effect, technical factors, etc.), we designed an inter-dataset experiment, in which we trained a classifier based on 10× PBMCs v1 dataset and used this classifier to predict the cells of 10× PBMCs Next Gem dataset. Note that for the predicting dataset, since there are no known labels before classification and our process of enhancement and imputation are both group-based, a clustering is recommended to assign the cells a group number for following enhancement and imputation.

Performance Evaluation Metrics

In this paper, we evaluated and compared the performance of SVM (linear kernel) and svmATAC using the following two metrics:

For all datasets, we compared the F1 scores across different cell types and evaluated the performance of each method using mean F1 scores.

F1 score is defined as:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

where Precision is defined as:

$$\text{Precision} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalsePositives}} \quad (4)$$

Similarly, Recall (or the ratio of TPs to total calls in the truth set) is defined as:

$$\text{Recall} = \frac{\text{TruePositives}}{\text{TruePositives} + \text{FalseNegatives}} \quad (5)$$

We represented the percentage of cells of a specific reported type labeled as each type in a heatmap, which flatly and intuitively showed the confusion matrix and the percentage of correctly/incorrectly classified cells.

The percentage of cells of a specific reported type labeled as each type is defined as:

$$\text{Percentage}_{\text{cell_type}_i, \text{cell_type}_j} = \frac{N_{\text{cell_type}_i, \text{cell_type}_j}}{N_{\text{cell_type}_i}} \quad (6)$$

where $\text{Percentage}_{\text{cell_type}_i, \text{cell_type}_j}$ represents the percentage of those cell_type_i cells labeled as cell_type_j , $N_{\text{cell_type}_i, \text{cell_type}_j}$ represents the number of those cell_type_i cells labeled as cell_type_j , and $N_{\text{cell_type}_i}$ represents the total number of cell_type_i .

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

ZC and YC conceived the project. ZC, TZ, and YW supervised the project. ZC developed svmATAC. YG and TJ provided valuable suggestions for experiments design. ZC, TZ, and YW wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Key Research and Development Program of China (Grant Nos. 2017YFC1201201, 2018YFC0910504, and 2017YFC0907503), the National Natural Science Foundation of China (Grant No. 32000467), the China

Postdoctoral Science Foundation (Grant No. 2020M681086), and the Heilongjiang Postdoctoral Foundation (Grant No. LBH-Z20014).

ACKNOWLEDGMENTS

We acknowledge members of the Wang lab for constructive discussions and support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.658352/full#supplementary-material>

Supplementary Figure 1 | Heatmap and F1-score comparing the SVM (linear kernel) and svmATAC predicted cells cluster versus original cluster in intra-dataset experiment. **(A)** Heatmap displaying the confusion matrix of predicted cell cluster ID versus original cluster ID in 10× PBMCs v1 with cluster ID dataset. **(B)** Heatmap displaying the confusion matrix of predicted cell cluster ID versus original cluster ID in 10× PBMCs Next Gem with cluster ID dataset. **(C)** Bar plot displaying the f1 scores of 10× PBMCs v1 with cluster ID. **(D)** Bar plot displaying the f1 score of 10× PBMCs Next Gem with cluster ID. Colors of **(A,B)** represent the percentages of cells of a specific reported type labeled as each type by svmATAC. In **(C,D)**, the red panel represent the results for svmATAC, and blue panel represents the results for general SVM on unenhanced and unimputed data.

Supplementary Figure 2 | Heatmap comparing the SVM (linear kernel) and svmATAC predicted cells cluster versus original cluster ID in inter-dataset experiment. **(A)** 10× PBMCs v1 with cluster ID dataset. **(B)** 10× PBMCs Next

Gem with cluster ID dataset. Colors represent the percentages of cells of a specific reported type labeled as each type by svmATAC.

Supplementary Table 1 | F1 scores of intra-dataset experiment using Corces2016 dataset with different enhancement and imputation cutoffs.

Supplementary Table 2 | The confusion matrix across different enhancement and imputation cutoffs.

Supplementary Table 3 | F1 scores of intra-dataset experiment using Buenrostro2018 dataset with different enhancement and imputation cutoffs.

Supplementary Table 4 | The confusion matrix across different enhancement and imputation cutoffs for Buenrostro2018 dataset.

Supplementary Table 5 | F1 scores of intra-dataset experiment using 10× PBMCs v1 Seurat Labeled dataset with different enhancement and imputation cutoffs.

Supplementary Table 6 | The confusion matrix across different enhancement and imputation cutoffs for 10× PBMCs v1 Seurat Labeled dataset.

Supplementary Table 7 | F1 scores of intra-dataset experiment using 10× PBMCs Next Gem Seurat Labeled dataset with different enhancement and imputation cutoffs.

Supplementary Table 8 | The confusion matrix across different enhancement and imputation cutoffs for 10× PBMCs Next Gem Seurat Labeled dataset.

Supplementary Table 9 | F1 scores of inter-dataset experiment that training with 10× PBMCs v1 Seurat Labeled dataset and predicting in 10× PBMCs Next Gem Seurat Labeled dataset with different enhancement and imputation cutoffs.

Supplementary Table 10 | The confusion matrix across different enhancement and imputation cutoffs for inter-dataset experiment that training with 10× PBMCs v1 Seurat Labeled dataset and predicting in 10× PBMCs Next Gem Seurat Labeled dataset.

REFERENCES

- Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H. L., Reinders, M. J. T., et al. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* 20:194.
- Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., and Powell, J. E. (2019). scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol.* 20:264. doi: 10.1186/s13059-019-1862-5
- Boufela, K., Seth, S., and Batada, N. N. (2019). scID: identification of transcriptionally equivalent cell populations across single cell RNA-seq data using discriminant analysis. *bioRxiv* [preprint] doi: 10.1101/470203
- Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., et al. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* 16, 397–400. doi: 10.1038/s41592-019-0367-1
- Broad_Institute (2019). *Picard toolkit*. Broad Institute, GitHub repository. Available online at: <http://broadinstitute.github.io/picard/> (accessed September 29, 2020).
- Buenrostro, J. D., Corces, M. R., Lareau, C. A., Wu, B., Schep, A. N., and Aryee, M. J. (2018). Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation. *Cell* 173, 1535–1548.e16. doi: 10.1016/j.cell.2018.03.074
- Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10, 1213–1218. doi: 10.1038/nmeth.2688
- Chen, H. D., Lareau, C. A., Andreani, T., Vinyard, M. E., Garcia, S. P., Clement, K., et al. (2019). Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* 20:241.
- Corces, M. R., Buenrostro, J. D., Wu, B., Greenside, P. G., Chan, S. M., and Koenig, J. L. (2016). Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* 48, 1193–1203. doi: 10.1038/ng.3646
- Cui, Z., Juan, L., Jiang, T., Liu, B., Zang, T., and Wang, Y. (2020). “Assessment of machine learning methods for classification in single cell ATAC-seq,” in *Proceeding of the 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Seoul: IEEE).
- Cusanovich, D. A., Daza, R., Adey, A., Pliner, H. A., Christiansen, L., Gunderson, K. L., et al. (2015). Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* 348, 910–914.
- Genomic (2020). *Genomic, 10X*. Available online at: <https://support.10xgenomics.com/single-cell-atac/> (accessed April 3, 2020).
- Johnson, T. S., Wang, T., Huang, Z., Yu, C. Y., Wu, Y., Han, Y., et al. (2019). LAMBDA: label ambiguous domain adaptation dataset integration reduces batch effects and improves subtype detection. *Bioinformatics* 35, 4696–4706. doi: 10.1093/bioinformatics/btz295
- Kiselev, V. Y., Yiu, A., and Hemberg, M. (2018). scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* 15, 359–362. doi: 10.1038/nmeth.4644
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* [preprint] arXiv 1303.3997.
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Li, Z., Kuppe, C., Cheng, M., Menzel, S., Zenke, M., Kramann, R., et al. (2019). scOpen: chromatin-accessibility estimation of single-cell ATAC data. *bioRxiv* [preprint] doi: 10.1101/865931
- Lieberman, Y., Rokach, L., and Shay, T. (2018). CaSTLe-classification of single cells by transfer learning: harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. *PLoS One* 13:e0205499. doi: 10.1371/journal.pone.0205499

- Liu, F. L., Zhang, Y. Y., Zhang, L., Li, Z. Y., Fang, Q., Gao, R. R., et al. (2019). Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biol.* 20: 242.
- Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat. Methods* 15, 1053–1058. doi: 10.1038/s41592-018-0229-2
- Ma, F., and Pellegrini, M. (2019). Automated identification of cell types in single cell RNA sequencing. *bioRxiv* [preprint] doi: 10.1101/532093
- McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv* [preprint] arXiv 1802.03426.
- Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D. J., Alvarez-Varela, A., et al. (2020). Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* 38, 747–755. doi: 10.1038/s41587-020-0469-4
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *JMLR* 12, 2825–2830.
- Pliner, H. A., Packer, J. S., McFaline-Figueroa, J. L., Cusanovich, D. A., Daza, R. M., Aghamirzaie, D., et al. (2018). Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* 71, 858–871.e8. doi: 10.1016/j.molcel.2018.06.044
- Pliner, H. A., Shendure, J., and Trapnell, C. (2019). Supervised classification enables rapid annotation of cell atlases. *Nat. Methods* 16, 983–986. doi: 10.1038/s41592-019-0535-3
- Stewart, T. G., Zeng, D., and Wu, M. C. (2018). Constructing support vector machines with missing data. *WIREs Comput. Stat.* 10:e1430. doi: 10.1002/wics.1430
- Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., et al. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888–1902.e21. doi: 10.1016/j.cell.2019.05.031
- Tan, Y., and Cahan, P. (2019). Single cell net: a computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Syst.* 9, 207–213.e2. doi: 10.1016/j.cels.2019.06.004
- Xiong, L., Xu, K., Tian, K., Shao, Y., Tang, L., Gao, G., et al. (2019). SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* 10:4576. doi: 10.1038/s41467-019-12630-7
- Yan, F., Powell, D. R., Curtis, D. J., and Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* 21, 22–22. doi: 10.1186/s13059-020-1929-3

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.

Copyright © 2021 Cui, Cui, Gao, Jiang, Zang and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



PMDFI: Predicting miRNA–Disease Associations Based on High-Order Feature Interaction

Mingyan Tang[†], Chenzhe Liu[†], Dayun Liu, Junyi Liu, Jiaqi Liu* and Lei Deng*

School of Computer Science and Engineering, Central South University, Changsha, China

OPEN ACCESS

Edited by:

Wang Guohua,
Harbin Institute of Technology, China

Reviewed by:

Yanglan Gan,
Donghua University, China
Hao Lin,
University of Electronic Science and
Technology of China, China

*Correspondence:

Lei Deng
leideng@csu.edu.cn
Jiaqi Liu
liujiaqi@csu.edu.cn

[†]These authors have contributed
equally to this work and share first
authorship

Specialty section:

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

Received: 20 January 2021

Accepted: 18 February 2021

Published: 09 April 2021

Citation:

Tang M, Liu C, Liu D, Liu J, Liu J and
Deng L (2021) PMDFI: Predicting
miRNA–Disease Associations Based
on High-Order Feature Interaction.
Front. Genet. 12:656107.
doi: 10.3389/fgene.2021.656107

MicroRNAs (miRNAs) are non-coding RNA molecules that make a significant contribution to diverse biological processes, and their mutations and dysregulations are closely related to the occurrence, development, and treatment of human diseases. Therefore, identification of potential miRNA–disease associations contributes to elucidating the pathogenesis of tumorigenesis and seeking the effective treatment method for diseases. Due to the expensive cost of traditional biological experiments of determining associations between miRNAs and diseases, increasing numbers of effective computational models are being used to compensate for this limitation. In this study, we propose a novel computational method, named PMDFI, which is an ensemble learning method to predict potential miRNA–disease associations based on high-order feature interactions. We initially use a stacked autoencoder to extract meaningful high-order features from the original similarity matrix, and then perform feature interactive learning, and finally utilize an integrated model composed of multiple random forests and logistic regression to make comprehensive predictions. The experimental results illustrate that PMDFI achieves excellent performance in predicting potential miRNA–disease associations, with the average area under the ROC curve scores of 0.9404 and 0.9415 in 5-fold and 10-fold cross-validation, respectively.

Keywords: miRNA–disease associations, high-order features, feature interactions, random forest, logistic regression

1. INTRODUCTION

MiRNAs are short non-coding RNAs with length about 19–25 nucleotides (Ambros, 2001, 2004; Bartel, 2004). Since the first miRNA (lin-4) was discovered by Victor Ambros in 1993 (Lee et al., 1993), miRNA has been the most widely studied class of non-coding RNAs now (Esteller, 2011). Besides, it has been confirmed that miRNAs commonly exist in plants, animals, viruses, and human beings, and have an essential effect on cell growth, differentiation, and apoptosis because of its post-transcriptionally gene regulation by affecting the translation of mRNAs (Wienholds and Plasterk, 2005; Das et al., 2014; Zhao et al., 2017). The important influence of miRNAs on biological processes is manifested in most intronic miRNAs sharing promoter regions with host genes (Zhao et al., 2015). Therefore, it is natural for scientists to link miRNAs with human diseases and use them as biomarkers in the treatment of diseases. For example, miR-164a is highly expressed in pediatric acute lymphoblastic leukemia and pediatric acute myeloid leukemia (Zhang et al., 2009; Li et al., 2010). Studies demonstrated that miR-21 plays a crucial role in a plethora of biological diseases

including cancer, cardiovascular diseases, and inflammation (Kumarswamy et al., 2011). Guay and Regazzi (2015) and Horsham et al. (2015) observed that the deregulation of miR-7 expression can potentially affect the adaptive capacity of β cells, contributing to the development of diabetes. The model-based computational approach proposed by Wang et al. (2008) identified five transcription factors and 7 miRNAs to be potentially responsible for the level of androgen dependency. Although miRNAs are proved to have close relationship with human disorders, the traditional biological methods to detect the underlying association between miRNAs and diseases are laboratory based, costly, and time consuming. Therefore, it is urgent and essential to apply computational methods to solve this issue. Nowadays, many computational methods are proposed to predict the novel association between miRNAs and diseases, and they are mainly divided into two categories: one is based on the assumption that the functional similarity of miRNAs tends to relate to similar diseases, and the other is based on machine learning.

According to the hypothesis that the functionally related miRNAs have a positive relationship with corresponding diseases, Chen and Zhang (2013) presented three methods based on the microRNA similarity, phenotype similarity, and network consistency similarity obtained by both of the two above similarity values, which are named as MBSI, PBSI, and NetCBI, respectively. Among these methods, NetCBI is better than the others with area under the ROC curve (AUC) of 0.8066, which still needs to be improved. Li et al. (2017) provided DeepWalk method that utilizes similarities within a known miRNA-disease association bipartite network to predict the unidentified miRNA-disease association when biological information, such as miRNA functional similarity and disease semantic similarity is unavailable. Although this method could reach the highest AUC of 0.937, it is incapable to predict associations of new miRNA or diseases that do not exist in the known network. Shen et al. (2017) integrated miRNA functional similarity, disease semantic similarity, and known miRNA-disease association, and then employed collaborative matrix factorization to predict the unknown miRNA-disease association (CMFMDA). CMFMDA could predict undiscovered miRNAs and diseases without known associations, but it may bias to miRNAs with more verified associated diseases. Chen et al. (2016) developed WBSMDA to reveal the novel miRNA-disease associations by integrating confirmed miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile (GIP) kernel similarity of diseases and miRNAs, and obtained an average AUC of 0.8031. Then, they further raised the AUC to 0.9035 with an original method called HAMDA (Chen et al., 2017), which employs the hybrid graph-based recommendation algorithm to uncover the unrecognized associations between miRNAs and diseases.

As for methods based on machine learning, Peng et al. (2019) proposed a learning-based model named MDA-CNN. The method generates a three-layer network, including miRNA similarity network, disease similarity network, and protein-protein interaction network, to extract features and integrates an autoencoder and a convolutional network to select features

and predict miRNA-disease association, respectively. Although the highest AUC the MDA-CNN achieved is 0.8897, the method performs well at the miRNA-phenotype association prediction. Zheng et al. (2019) presented a model based on machine learning named MLMDA, which utilizes miRNA sequence information extracted by k-mer sparse matrix, combining with other similarities of diseases and miRNAs. Besides, the MLMDA adopts a deep autoencoder to glean more latent features and uses the random forest (RF) to predict novel miRNA-disease associations. Chen et al. (2019) developed a method called EDTMDA, which applies principal component analysis (PCA) to reducing the dimension of features and utilizes ensemble learning to gain ultimate scores between miRNAs and diseases. EDTMDA's AUC could reach 0.9309 in LOOCV, but the dependence on the known associations between miRNAs and diseases may lead to a preference for miRNAs that have more associated diseases. Jiang et al. (2013) proposed an SVM-based method to identify disease-related microRNAs, which can distinguish positive microRNA-disease associations from negative microRNA-disease associations. In 10-fold cross-validation procedure, this method achieved the AUC of up to 0.8884. Zhang et al. (2019) proposed an unsupervised deep learning method implemented by variational autoencoder. The method combines miRNA similarity and disease similarity with identified associations to get two spliced matrices as the input of variational autoencoder, and then obtains the association scores of miRNA and disease. The model is not affected by the dearth of negative samples, but is hard to interpret.

In conclusion, the aforementioned computational methods could predict the underlying miRNA-disease associations effectively, but each one still has its own limits. In this paper, we propose a novel method called PMDFI, which is an ensemble approach for miRNA-disease associations prediction based on feature interaction learning. Our model can be divided into four parts: data set collection and processing, high-level feature extraction, feature interaction, and an integrated learning model. In detail, we gather miRNA-disease associations from HMDD v2.0, and calculate miRNA functional similarity, disease semantic similarity, GIP kernel similarity for miRNA, and disease. Then, after using the stacked autoencoder to extract the high-order features, we send them to the feature interactive layer to gain cross features. Finally, we design an ensemble model combining multiple RFs and logistic regression to predict potential miRNA-disease associations. In the experimental results, PMDFI has achieved excellent performance in predicting potential miRNA-disease associations, with AUC of 0.9404 and 0.9415 under 5-fold and 10-fold cross-validation, respectively.

2. MATERIALS AND METHODS

2.1. Datasets for MDA Prediction

The experimentally supported miRNA-disease associations come from HMDD v2.0, which is derived from Li et al.'s work (Li et al., 2014). HMDD v2.0, a manual collected database, is used to annotate in details the miRNA-disease associations from genetics, epigenetics, circulating miRNAs, and miRNA-target interactions. We gather 5430 miRNA-disease association pairs encompassing 495 miRNAs and 383 diseases from the HMDD

v2.0. In order to represent the associations between miRNA $m(i)$ and disease $d(j)$, we construct an adjacency matrix $A_{495 \times 383}$, where element $A(i, j) = 1$ indicates that miRNA has a definite association with disease, and element $A(i, j) = 0$ indicates that the association between miRNA and disease is uncertain. Matrix A is a sparse matrix with 5,430 of “1,” i.e., 5,430 miRNA–disease association pairs, and we take these pairs as positive samples. As for the negative samples, according to Zhou et al. (2020), all “0”s (miRNA–disease pairs with no definite association) in the matrix A are divided into 23 clusters with k-means clustering, and the same amount of samples are randomly selected from each cluster to form 5,418 negative samples. It is worth noting that, in order to ensure the validity of comparative experiments, the positive and negative samples in our datasets are the same as Zhou et al.’s work.

2.2. MiRNA and Disease Information Profiles

2.2.1. MiRNA Functional Similarity

The miRNA functional similarity is useful to predict the functions of unknown miRNAs and study the interactions between miRNAs, because miRNAs with similar functions tend to trigger pathologically similar diseases. The miRNA functional similarity matrix can be represented as follows:

$$FS = [m_1, m_2, \dots, m_{nm}]^T, m_i \in \mathbb{R}^{km} \quad (1)$$

where nm is the number of miRNAs and km is the size of the vector that represents an miRNA.

Here, we download miRNA function similarity between miRNA pairs directly from <http://www.cuilab.cn/files/images/cuilab/misim.zip>, which calculated by Wang et al.’s work based on advanced MISIM method (Wang et al., 2010). The miRNA functional similarity matrix FS is a matrix with 495 rows and 495 columns, and element $FS(m_i, m_j)$ represents the functional similarity between $miRNA(i)$ and $miRNA(j)$.

2.2.2. Disease Semantic Similarity

If an miRNA has been proved to be linked to a certain disease, it is possible that the miRNA is also related to other diseases with similar phenotypes. Therefore, the semantic similarity of the disease is effective in large-scale research on the association between disease and miRNA. The disease semantic similarity is described as directed acyclic graph (DAG), and

$$DAG(d) = \{d, T(d), E(d)\} \quad (2)$$

where d is the disease itself, $T(d)$ is a set of nodes consisting of disease D and all its ancestor nodes, and $E(d)$ corresponds to the edge set of the direct link from the parent node to the child node.

We collect disease semantic similarity from MeSH database (<http://www.ncbi.nlm.nih.gov/>), which has been widely adopted to study miRNA–disease associations (Zou et al., 2016). And each disease in DAG can be calculated as follows:

$$\begin{cases} D1_D(d) = 1 & \text{if } d = D \\ D1_D(d) = \max \{0.5 \times D1_D(d') \mid d' \in \text{child of } d\} & \text{if } d \neq D \end{cases} \quad (3)$$

and

$$DV(D) = \sum_{d \in T(d)} D_D(d) \quad (4)$$

Then the semantic similarity score between $diseases(i)$ and $diseases(j)$ is defined as follows:

$$SS(d(i), d(j)) = \frac{\sum_{t \in T(d(i)) \cap T(d(j))} (D_{d(i)}(t) + D_{d(j)}(t))}{DV(d(i)) + DV(d(j))}. \quad (5)$$

2.2.3. GIP Kernel Similarly for miRNA and Disease

GIP kernel similarity originates from the topological structure of the known interaction network, which is beneficial for predicting the miRNA–disease associations (Wang et al., 2010). We adopt a binary vector $IP(d)$, a row in the adjacency matrix, to express the interaction profile of disease d with each miRNA, and the disease GIP kernel similarity between disease $d(i)$ and $d(j)$ can be calculated as follows:

$$GS_d(d_i, d_j) = \exp \left(-\gamma_d \|IP(d_i) - IP(d_j)\|^2 \right) \quad (6)$$

and

$$\gamma_d = \lambda'_d / \left(\frac{1}{n} \sum_{i=1}^n \|IP(d_i)\|^2 \right) \quad (7)$$

where n is the number of human diseases and equals to 383, γ_d is an adjustable parameter of the kernel bandwidth, and $\lambda'_d = 1$ according to van Laarhoven et al.’s work (van Laarhoven et al., 2011). Similarly, we can use a binary vector $IP(m)$ to express the interaction profile of miRNA m with each disease, and the GIP kernel similarity between miRNA $m(i)$ and $m(j)$ can be calculated as follows:

$$GS_m(m_i, m_j) = \exp \left(-\gamma_m \|IP(m_i) - IP(m_j)\|^2 \right) \quad (8)$$

and

$$\gamma_m = \lambda'_m / \left(\frac{1}{m} \sum_{i=1}^m \|IP(m_i)\|^2 \right) \quad (9)$$

where m is the number of miRNAs and equals to 495, for the same reason, λ'_m is set to 1.

2.3. PMDFI Framework

In this study, we construct a model named PMDFI to predict potential miRNA–disease associations. The flowchart of PMDFI is shown in **Figure 1**. In the data set collection and processing stage, we gather 495 miRNAs and 383 diseases from the HMDD v2.0 database to form an adjacency matrix $A_{495 \times 383}$, including 5430 miRNA–disease pairs with definite associations. Then, we acquire miRNA functional similarity (FS), disease semantic similarity (SS), and GIP kernel similarity for miRNA (GS_m) and disease (GS_d). For each miRNA–disease pair, we extract four one-dimensional features, which include a 1×495 miRNA functional similarity feature, a 1×383 diseases semantic similarity feature, and a 1×495 and 1×383 GIP kernel similarity for

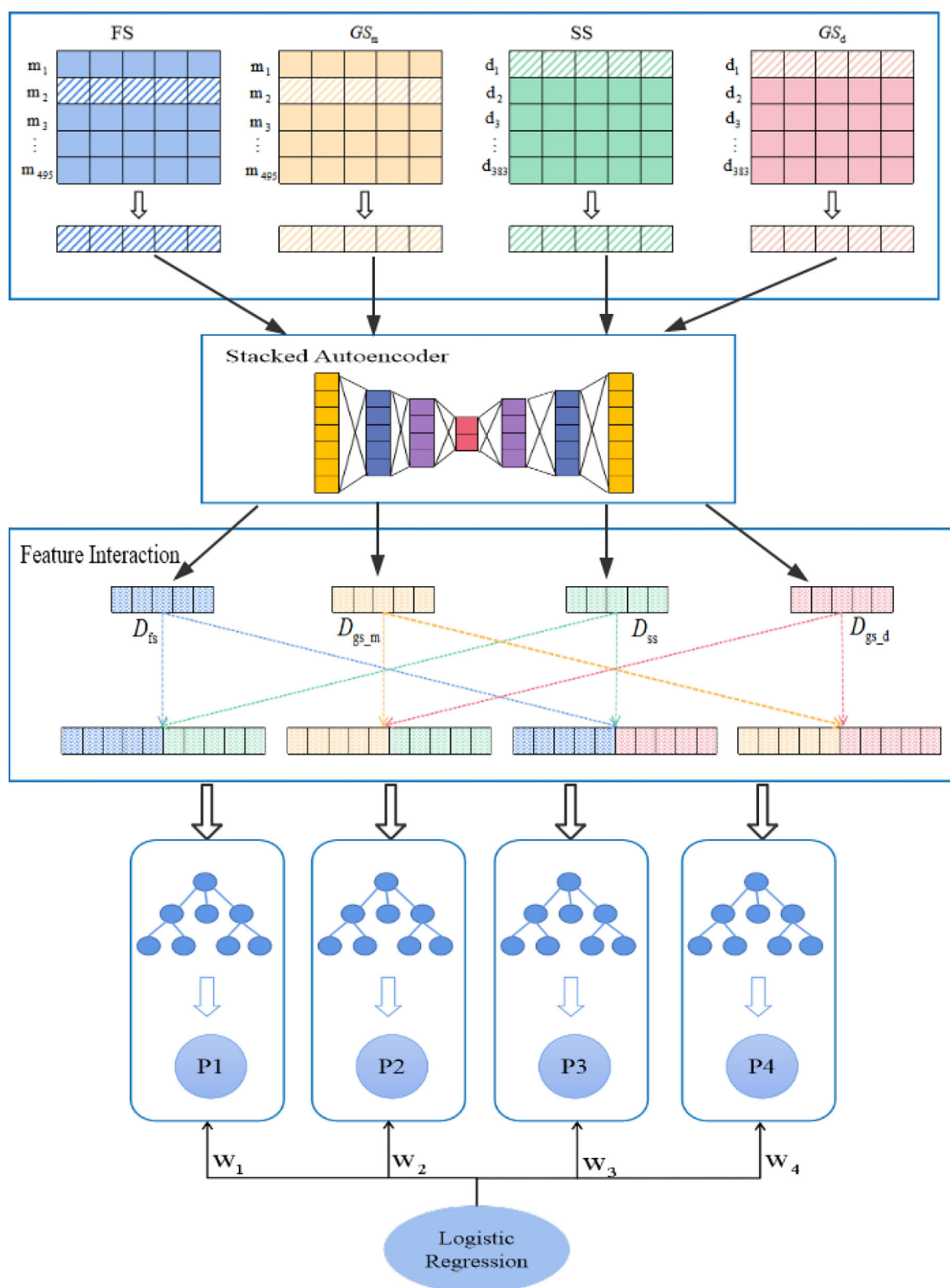


FIGURE 1 | Flowchart of PMDFI model to predict potential microRNAs (miRNAs)–diseases associations. The model can be divided into four parts: data set collection and processing, high-order feature extraction, feature interaction, and an integrated learning model. First, we gather miRNA–disease associations from HMDD v2.0, and form the similarity matrix between miRNA and disease; second, we adopt a stacked autoencoders to extract high-order features; then, we use the interaction features layer to learn the interaction between different features. Finally, we combine multiple random forest (RF) with logistic regression to predict potential miRNA–disease associations.

miRNAs and disease. Then these features are input in parallel into the stacked autoencoder to extract high-order features, instead of directly concatenating and averaging them. In this way, our method has the ability to learn the internal deep connections in the feature matrix, which have been previously ignored due to the lack of miRNA functional similarity or diseases semantic similarity. In the feature interaction layer, the high-order features derived from stacked autoencoder are sent to perform feature interaction learning, which aims at obtaining four cross features containing the internal potential relationship of miRNA (disease) and the interaction information among those features. Finally, the obtained cross features are independently input into the four RF models for training, and a set of four prediction scores is calculated for each sample input. During each iteration, we constantly adjust the weight of each RF model, and adopt a logistic regression to make a final comprehensive prediction.

2.3.1. Stacked Autoencoder to Extract High-Order Features

These four similarities matrix information (FS, SS, GS_m, and GS_d) have inevitable restriction that they are unable to present the inner deep connections among different miRNAs (diseases) due to low-order feature representations. To tackle this obstacle, inspired by Song et al.'s work (Song et al., 2019), we use a stacked autoencoder to extract meaningful high-order features for miRNA and disease from the established similarity network. The autoencoder is an artificial neural network that can learn the efficient representation of input data through unsupervised learning (Vincent et al., 2008; Shu et al., 2018). As a powerful feature detector, the autoencoder encodes the original input feature and reduces the dimensionality to find implicit associations between the input feature, and extracts expressive high-order features. As shown in **Figure 2**, the stacked encoder consists of two parts: an encoder (also known as the recognition network) and a decoder (also known as the generation network). The encoder converts the input feature into an internal representation, and the decoder converts the internal indicates conversion to output.

In order to learn high-order features, we build a stacked autoencoder that includes three hidden layers with 256, 128, and 64 units. The stacked autoencoder means that the feature vectors in the previous autoencoder are used as the input of the next autoencoder, and the whole training process is greedy in a layered manner. In our model, the feature information of FS = {fs₁, fs₂, ..., fs₄₉₅}, SS = {ss₁, ss₂, ..., ss₄₉₅}, GS_d = {d₁, d₂, ..., d₃₈₃} and GS_m = {m₁, m₂, ..., m₄₉₅} is input into stacked autoencoder H1, H2, H3, and H4, respectively, and divided into four parallel groups for high-order feature extraction by minimizing the discrepancy between the input features and the reconstruction ones.

Initially, we set N_L and N_{G_i} as the number of units in the input layer and the i th hidden layer, and use one feature vector $x \in R^{N_L \times 1}$ to represent those input feature vectors. Subsequently, during the encoding process, the autoencoder transforms x into a latent representation $g^{(i)}$ through a composite mapping of linear transformation and non-linear activation function f , as shown in

the following equation:

$$g^{(i)} = f(W_1^{(i)}x + b_1^{(i)}) \quad (10)$$

where i is i th hidden layer, $g^{(i)} \in R^{N_{G_i}}$ is the latent feature, $W_1^{(i)} \in R^{N_{G_i} \times N_L}$ is the encoding weight matrix, $b_1^{(i)} \in R^{N_{G_i}}$ is the bias vector, and $f(\cdot)$ is the sigmoid function.

Here, we adopt three hidden layers, i.e., $i = 3$. Then there is the process of decoding, which learns features inverse mapping. The latent representation $y^{(i)}$ is mapped to a feature vector as follows:

$$y^{(i)} = f(W_2^{(i)}g^{(i)} + b_2^{(i)}) \quad (11)$$

similarly, $g^{(i)}$ is the latent data, $W_2^{(i)} \in R^{N_L \times N_{G_i}}$ is the decoding weight matrix, $b_2^{(i)}$ is the bias vector.

Given a training feature vector $x(k)$, which can be shown as: $x(k) = \{f_s(k), ss(k), d(k), m(k)\}$ (Denoted as $\chi = \{FS, SS, GS_d, GS_m\}$), we can learn the underlying features by minimizing the reconstruction error of the cost function:

$$H_N(X, Y, \theta) = \frac{1}{2} \sum_{k=1}^m \|x(k) - y(k)\|_2^2 + \lambda \|\theta\|_2^2 \quad (12)$$

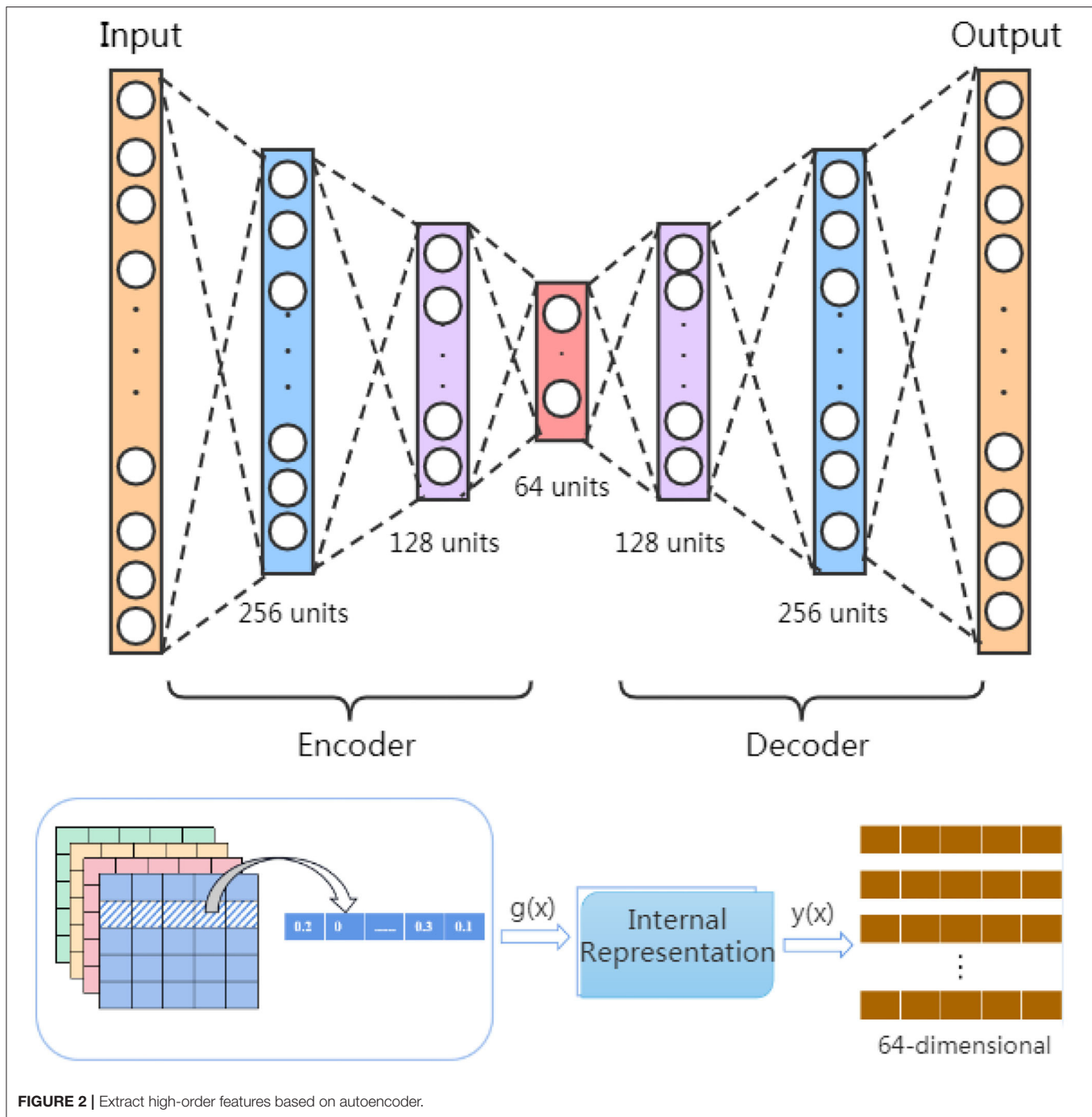
where $N = 1, 2, 3, 4$, and Y represents all the reconstructed feature vectors, $y(k)$ is the k th reconstructed feature vector, $x(k)$ is the k th training feature vector, m is the number of training feature vectors, λ is the weight decay parameter, $\theta = \{W, b\}$, W is the weight, and b is the biases of the autoencoder.

2.3.2. Feature Interaction

In the previous section, we have obtained four different types of high-order features (D_{fs} , D_{ss} , D_{gs_m} , and D_{gs_d}) derived from miRNA functional similarity, disease semantic similarity, and GIP kernel similarity for miRNA and disease. However, these four features are unilateral feature representations, which only express the degree of closeness among different miRNAs (diseases) and extract their meaningful latent connections. An effective prediction accuracy not only depends on valuable high-order features, but also on the feature interactive information. Therefore, we obtain cross features by combining different high-order features and use them to learn feature interaction information.

In our model, a feature interaction layer is adopted to gain the interaction information between different high-order features. Considering the miRNA-disease associations, we combine the two features of miRNA with the two features of disease, respectively, and gain a total of four cross features. In order to predict the association between a specific miRNA and a certain disease, D_{fs} and D_{ss} are simultaneously mapped to the same space to obtain cross features, which can be expressed as:

$$D_1 = \begin{bmatrix} D_{fs} \\ D_{ss} \end{bmatrix}^T \quad (13)$$



Similarly, the other three cross features are shown as follows:

$$D_2 = \begin{bmatrix} D_{gs_m} \\ D_{ss} \end{bmatrix}^T \quad (14)$$

$$D_3 = \begin{bmatrix} D_{fs} \\ D_{gs_d} \end{bmatrix}^T \quad (15)$$

$$D_4 = \begin{bmatrix} D_{gs_m} \\ D_{gs_d} \end{bmatrix}^T \quad (16)$$

As a result, the high-order features of miRNA and disease are mapped to different spaces for feature interaction, and four unilateral high-order features are converted into four cross features with deep interactivity.

2.3.3. Ensemble Model Based on Multiple RF and Logistic Regression

An RF consists of an set of classification trees, and each tree divides the feature space into different regions based on the division of each node in the tree. During the training process,

the randomness allows the trees to give independent estimates, which collectively contribute to achieve accurate and robust results. Here, we use four RFs and each RF is consisted of 300 independent trees. The core idea of our model is to input four interactive cross features into respective RF in parallel for self-learning and model building, and then merge the four RFs with logistic regression to make comprehensive predictions.

Our dataset includes 5,430 positive samples labeled as “1,” and 5,418 negative samples labeled as “0.” The input sample x_k of each four cross features covers diversified feature information and the four cross features could be represented as $f_k = \{D_1^{(k)}, D_2^{(k)}, D_3^{(k)}, D_4^{(k)}\}$, ($D_N \in \mathbb{R}^{1 \times 64}$, ($N = 1, 2, 3, 4$)). And we use $\theta_R = \{[x_1; f_1], [x_2; f_2], \dots, [x_m; f_m]\}$ to denote all training miRNA–disease pairs, where m is the number of all training sample pairs. In order to train a robust model, all samples are randomly input into the random forest for pre-training. For a sample x_k , the interactive cross features f_k are input into the corresponding RF, and a set of prediction score can be obtained and expressed as, $p^{(k)} = \{p_1^{(k)}, p_2^{(k)}, p_3^{(k)}, p_4^{(k)}\}$. $p_N^{(k)}$ is a probability score between 0 and 1, which represents the degree of association between a miRNA and a disease. Subsequently, we use logistic regression to do the final classification task for each miRNA–disease pair, instead of simply averaging the probability score of the four RF regression models. We consider the score $P^{(k)}$ of each sample pair x_k as a new feature $x'^{(k)} = \{x_1'^{(k)}, x_2'^{(k)}, x_3'^{(k)}, x_4'^{(k)}\}$ and assign it a weight $W^{(k)} = \{w_1^{(k)}, w_2^{(k)}, w_3^{(k)}, w_4^{(k)}\}$, and constantly update the weights during each iteration. After logistic regression training, the comprehensive prediction performance can be expressed as: $Y = w^T x' + b$, where b is a constant. Finally, We conduct 5-fold cross-validation and 10-fold cross-validation on all samples to test the performance of our method.

3. RESULTS AND DISCUSSION

3.1. Evaluation Criteria

To assess the performance of PMDFI, we adopt 5-fold cross-validation (5-CV) and 10-fold cross-validation (10-CV) as well as several widely used measures, including recall, precision, F1-score, AUC, and area under the PR curve (AUPR). And these measures are calculated as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (17)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (18)$$

$$F1 - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (19)$$

where TP , FP , TN , and FN represent the true positive, false positive, true negative, and false negative, respectively.

3.2. Prediction of miRNA–Disease Association Based on PMDFI

We use 5-fold and 10-fold cross-validation to evaluate the performance of PMDFI in predicting miRNA–disease

associations. In 5-CV (10-CV), all sample pairs are randomly divided into five (10) equal groups, and four (nine) groups of them are regarded as training samples, and the remaining one group is used as test samples. **Table 1** lists the results of 5-CV and 10-CV obtained by PMDFI, and indicates that under 5-CV (10-CV), the AUC, AUPR, Precision, Recall, and F1-score of PMDFI are 0.9404 (0.9415), 0.9373 (0.9385), 0.8663 (0.8669), 0.8812 (0.8832), and 0.8736 (0.8748), respectively. The average AUC of our model exceeds 0.94 in either the 5-fold cross-test or the 10-fold cross-test. Therefore, the results fully demonstrate that PMDFI has a good performance in predicting the latent associations between miRNAs and diseases.

3.3. Comparison With Existing State-of-the-Art Methods

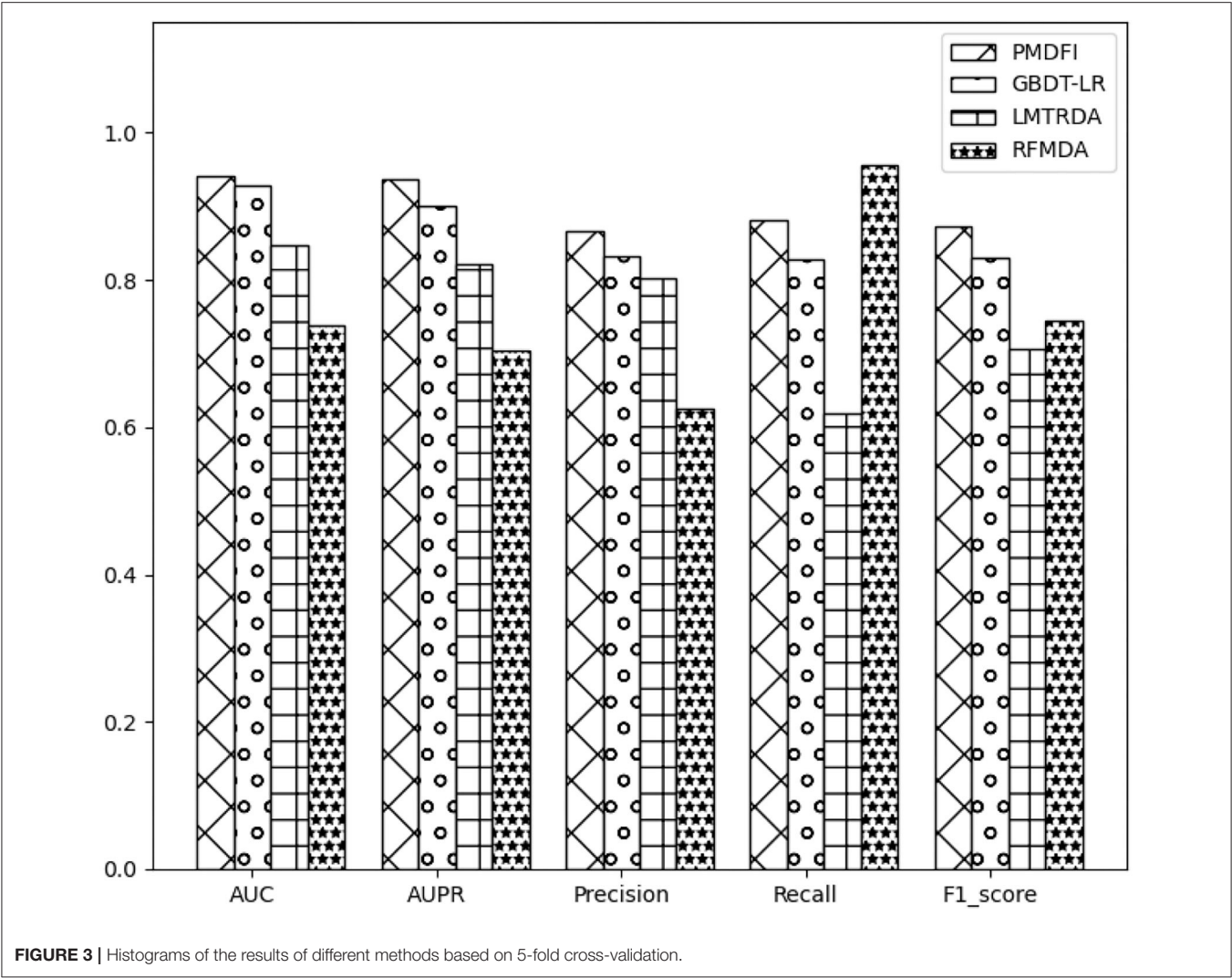
In order to systematically evaluate the performance of PMDFI, we compare our method with other state-of-the-art computational models, such as GBDT-LR (Zhou et al., 2020), LMTRDA (Wang et al., 2019), and RFMDA (Chen et al., 2018). GBDT-LR is a original model that combines gradient boosting decision tree with logistic regression to prioritize miRNA candidates for diseases. LMTRDA is a logistic model tree used to predict miRNA–disease associations by fusing multi-source information. RFMDA is a computational model of random forest for miRNA–disease associations prediction based on machine learning. The comparison between PMDFI and these models is carried out based on 5-CV and illustrated specifically in **Table 2**. From the table, PMDFI, GBDT-LR, LMTRDA, and RFMDA models achieve AUC of 0.9404, 0.9274, 0.8479, and 0.7388, respectively, and PMDFI presents the best performance. PMDFI outperforms GBDT-LR by 1.3%, LMTRDA by 9.25%, and RFMDA by 20.16% in terms of AUC. **Figure 3** further describes the comparison of our method with other methods in 5-CV with the format of histograms, and the leftmost one represents our method. In conclusion, except that the recall is 0.0736 lower than RFMDA, PMDFI makes a significant improvement in the field of prediction for potential miRNA–disease associations.

TABLE 1 | The results of 5-fold and 10-fold cross-validation obtained by PMDFI.

C. val.	AUC	AUPR	Precision	Recall	F1-score
5-CV	0.9404	0.9373	0.8663	0.8812	0.8736
10-CV	0.9415	0.9385	0.8669	0.8832	0.8748

TABLE 2 | The comparison of different methods based on 5-fold cross-validation.

Method	AUC	AUPR	Precision	Recall	F1-score
PMDFI	0.9404	0.9373	0.8663	0.8812	0.8736
GBDT-LR	0.9274	0.9014	0.8315	0.8273	0.8302
LMTRDA	0.8479	0.8217	0.8013	0.6190	0.7076
RFMDA	0.7388	0.7034	0.6253	0.9548	0.7453



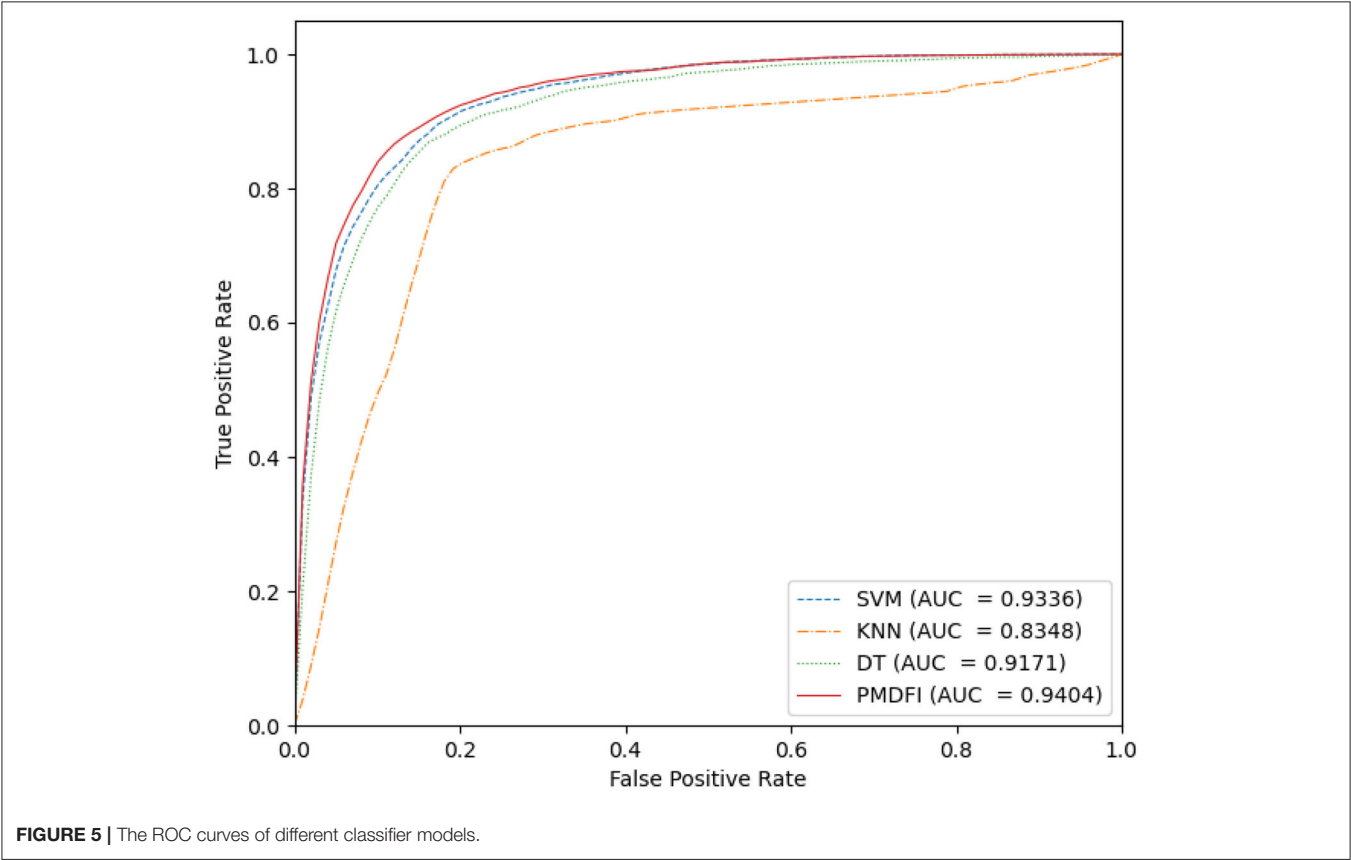
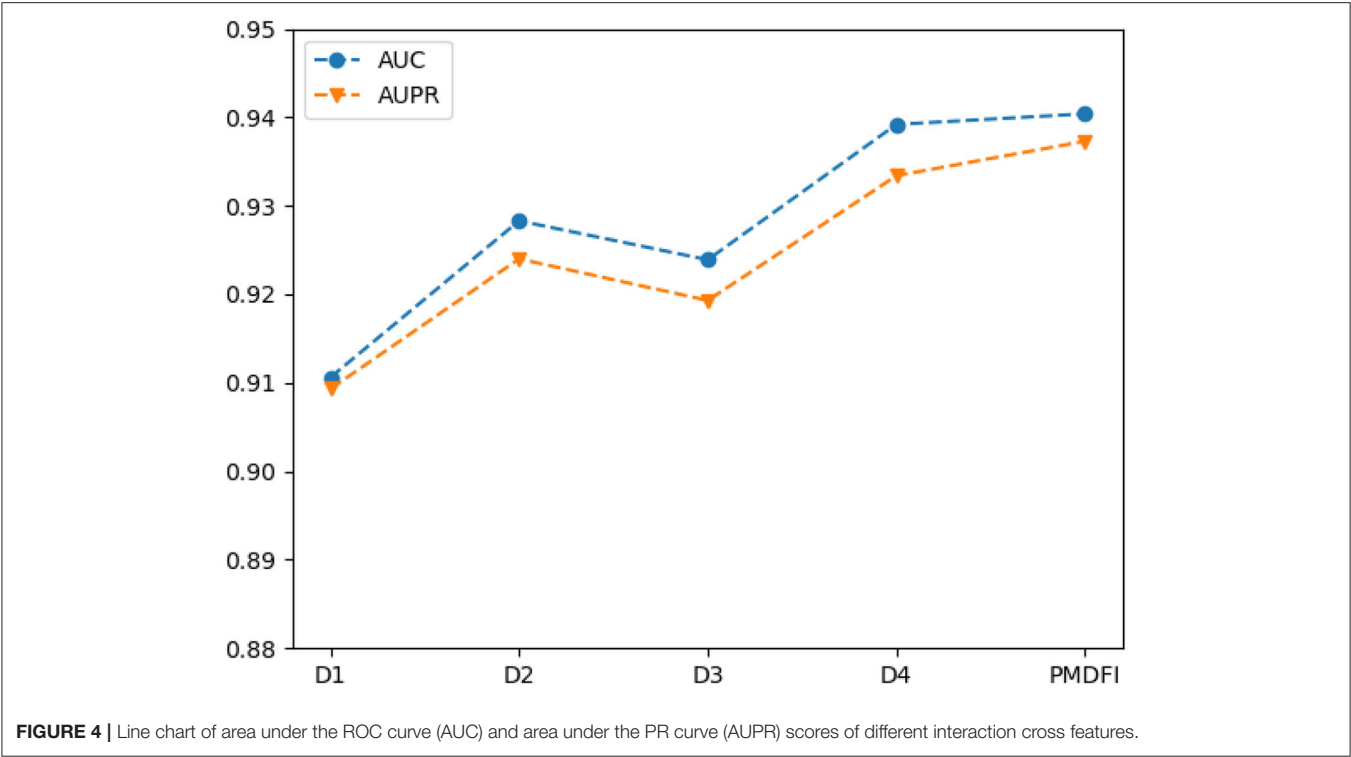
3.4. Comparison With Different Interactive Cross Features

In order to further illustrate the contribution of distinct interactive cross features to the potential miRNA–disease associations prediction, we separately input cross features $D1 (D_{fs} \oplus D_{ss})$, $D2 (D_{fs} \oplus D_{gs-d})$, $D3 (D_{gs-m} \oplus D_{ss})$, and $D4 (D_{gs-m} \oplus D_{gs-d})$ into the RF model for training, without integrating the overall performance of the four cross features. **Table 3** displays the performance of each interactive cross features on miRNA–disease potential association prediction. In the table, the AUC and AUPR score of the four interactive cross features fluctuate in the range of 0.9249 ± 0.0143 and 0.9213 ± 0.0121 , respectively. And the cross feature $D1$ has the worst performance with an AUC of 0.9106, which is 2.98% lower than the optimal score. Besides, the $D4$ cross feature has the best performance compared to other three, and its AUC, AUPR, Precision, Recall, and F1-score are 0.9392, 0.9334, 0.8630, 0.8834, and 0.8730, respectively. Although $D4$ is the best performer

TABLE 3 | Comparison of the performance of four interactive cross features.

Method	AUC	AUPR	Precision	Recall	F1-score
D1 ($D_{fs} \oplus D_{ss}$)	0.9106	0.9093	0.8289	0.8388	0.8338
D2 ($D_{fs} \oplus D_{gs-d}$)	0.9283	0.9240	0.8513	0.8692	0.8601
D3 ($D_{gs-m} \oplus D_{ss}$)	0.9239	0.9193	0.8381	0.8642	0.8509
D4 ($D_{gs-m} \oplus D_{gs-d}$)	0.9392	0.9334	0.8630	0.8834	0.8730
PMDFI	0.9404	0.9373	0.8663	0.8812	0.8736

among the four cross features, the performance of it is still slightly worse than that of the integration of the whole four features. For a clearer comparison, we also draw a line graph of the four interactive cross features and their combinations in terms of AUC and AUPR values. **Figure 4** gives a clue that the performance of integrating the four interactive cross features is the best, and its AUC and AUPR values are both at the highest point.



3.5. Comparison With Different Classifier Models

In our method, we use an ensemble learning model composed of multiple RFs to predict the potential miRNA–disease associations. To confirm the excellence of the RF-based ensemble learning model, we compare it with several common classifier models, such as SVM, k-nearest neighbor (KNN), and decision tree (DT), using a common data set and feature set. **Figure 5** is the ROC curve of these four classifier models, where the AUC of SVM, KNN, DT, and PMDFI are 0.9336, 0.8348, 0.9171, and 0.9404, respectively. From the picture, the performance of

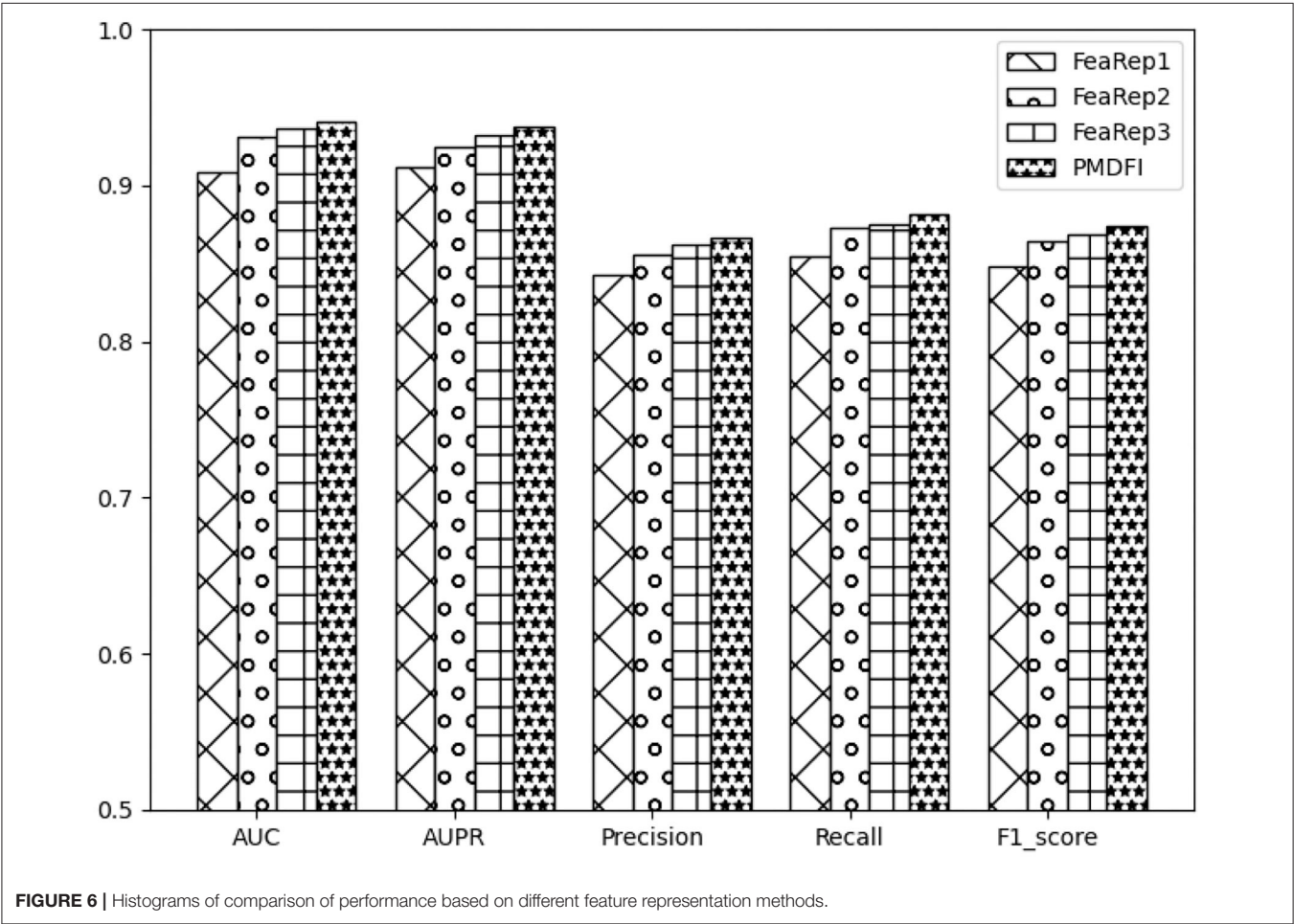
SVM is slightly worse than PMDFI; the AUC of DT is 2.33% lower than PMDFI; the performance of KNN is the worst among them, and its AUC is 10.56% lower than PMDFI. In summary, our method, RF-based PMDFI, has a curve above all the other three ones, which stands for the best performance in predicting miRNA–disease associations.

3.6. Analysis of High-Order Feature Extraction and Feature Interaction

Unlike other models that directly use miRNA and disease similarity feature information, our method PMDFI utilizes high-order feature extraction and feature interaction to represent features. In order to verify the validity of the proposed feature representation approach, we compare it with other three methods. The first one is DBNMDA (Chen et al., 2020), which directly extracts the features of all miRNA–disease pairs to pre-train the Restricted Boltzmann Machine (RBM). The second one is DBMDA (Zheng et al., 2020), which utilizes the autoencoder to resize the miRNA (disease) similarity features and then fuses the features during the feature set construction stage. The third one is GBDT-LR (Zhou et al., 2020), which uses gradient boosting decision tree (GBDT) to extract distinguishing features and

TABLE 4 | The specific outcomes based on different feature representation methods.

Method	AUC	AUPR	Precision	Recall	F1-score
FeaRep1	0.9083	0.9119	0.8430	0.8543	0.8486
FeaRep2	0.9307	0.9252	0.8554	0.8731	0.8641
FeaRep3	0.9367	0.9327	0.8619	0.8746	0.8682
PMDFI	0.9404	0.9373	0.8663	0.8812	0.8736



feature combinations. We name the feature representation in each of the aforementioned three methods as FeaRep1 (based on DBNMDA), FeaRep2 (based on DBMDA), and FeaRep3 (based on GBDT-LR). **Table 4** reveals in details the outcome of distinct feature representation methods. The AUC of the feature representation method used in the PMDFI are 3.21, 0.97, and 0.37% higher than FeaRep1, FeaRep2, and FeaRep3, respectively. And we plot more straightforward histograms to illustrate the results of the comparison, as shown in **Figure 6**. From the figure, the feature representation method used by PMDFI, the rightmost one, is superior to the other three methods in all evaluation dimensions. To summarize, the experiment further demonstrates that high-order feature extraction and feature interaction have profound contributions to predicting the potential relevance of miRNA–disease.

3.7. Case Studies

To analyze the prediction performance of PMDFI in practical situations, we conduct several common disease case studies with PMDFI, including breast cancer, melanoma, and lymphoma. We initially train all known miRNA–disease associations in the HMDD v.2.0 with PMDFI, and then list top-10 predicted miRNAs for validation using two other databases, namely dbDEMC 2.0 (Yang et al., 2017) and miRCancer (Xie et al., 2013). The dbDEMC 2.0 is a database designed to store and display differentially expressed miRNAs in detected human cancers, which contains 2,224 differentially expressed miRNAs in 36 cancer types. And the miRCancer is a microRNA–cancer association database, which currently records 878 relationships between 236 miRNAs and 79 human cancers.

According to recent studies, we choose three prevalent diseases as our case studies and the results are listed in **Table 5**. The first one is breast cancer, as the most common cancer affecting women, which accounts for 23% of all cancers and 14% of cancer deaths (Jemal et al., 2011; Anastasiadi et al., 2017). The studies have shown that loss of the tumor suppressor miRNA or overexpression of the oncogenic miRNA may lead to the occurrence or metastasis of breast cancer (Serpico et al., 2014). Therefore, finding the relationship between miRNAs and breast cancer offers a direction for the diagnosis and treatment of breast cancer. From **Table 5**, we can see that nine out of the 10 predicted breast cancer related miRNAs appear in dbDEMC 2.0 or miRCancer. The second disease is Melanoma, which is the most serious type of skin cancer. It is caused by the cancerous transformation of skin cells when prolonged exposing under the ultraviolet light (Rastrelli et al., 2014). Pencheva et al. (2012) have identified a set of miRNAs that are deregulated in independent metastatic lines derived from multiple patients with melanoma, which manifests the importance to research the association between miRNAs and melanoma. The data from the middle line of **Table 5** illustrate that the PMDFI model has accurately predict all the top 10 melanoma-related miRNAs. The last disorder is malignant lymphoma, which is a large group of tumors with considerable heterogeneity. Although it occurs in the lymph nodes, due to the distribution characteristics of the lymphatic system, lymphoma is a systemic disease that can invade almost any tissue and organ in the body (Dean et al., 2005;

TABLE 5 | The candidate miRNAs associated with breast cancer, melanoma, and lymphoma.

Diseases	miRNA	Evidence
Breast cancer	hsa-mir-150	dbDEMC 2.0;miRCancer
	hsa-mir-15b	dbDEMC 2.0
	hsa-mir-130a	dbDEMC 2.0;miRCancer
	hsa-mir-196b	dbDEMC 2.0
	hsa-mir-98	dbDEMC 2.0;miRCancer
	hsa-mir-106a	dbDEMC 2.0;miRCancer
	hsa-mir-142	miRCancer
	hsa-mir-378a	Unconfirmed
	hsa-mir-30e	miRCancer
Melanoma	hsa-mir-372	dbDEMC 2.0;miRCancer
	hsa-mir-150	miRCancer
	hsa-mir-373	miRCancer
	hsa-mir-127	dbDEMC 2.0
	hsa-mir-181b	dbDEMC 2.0
	hsa-mir-10b	dbDEMC 2.0;miRCancer
	hsa-mir-224	dbDEMC 2.0;miRCancer
	hsa-mir-101	dbDEMC 2.0;miRCancer
	hsa-mir-223	dbDEMC 2.0
Lymphoma	hsa-mir-27a	dbDEMC 2.0;miRCancer
	hsa-mir-30c	dbDEMC 2.0
	hsa-mir-34a	dbDEMC 2.0;miRCancer
	hsa-mir-34c	Unconfirmed
	hsa-mir-9	dbDEMC 2.0;miRCancer
	hsa-mir-29a	dbDEMC 2.0;miRCancer
	hsa-mir-222	dbDEMC 2.0
	hsa-mir-7a	dbDEMC 2.0
	hsa-mir-29b	dbDEMC 2.0;miRCancer
	hsa-mir-181b	dbDEMC 2.0
	hsa-mir-145	dbDEMC 2.0;miRCancer
	hsa-mir-221	dbDEMC 2.0

Paydas et al., 2016). Zheng et al. (2018) list several examples to describe miRNAs' role in the development of B-cell lymphoma, both as oncogenes and tumor suppressor genes, and nine out of the 10 predicted lymphoma-associated miRNAs are verified in dbDEMC 2.0 or miRCancer.

4. CONCLUSION

Given the significance that the miRNA–diseases associations make to the diagnosis of diseases and superiority that computer have compared to biological experiments, emerging computational models pop up in the miRNA–disease associations prediction realm. In this paper, we propose a novel computational model called PMDFI, which is an ensemble learning method to predict the miRNA–disease associations based on feature interactive learning. Our method not only integrates the four RF models of separated cross features, but also incorporates logistic regression to provide comprehensive predictions by assigning adjustable weights. Moreover, we

apply stacked autoencoders to extracting meaningful high-order features from miRNA functional similarity, disease semantic similarity, and GIP kernel similarity of miRNA and disease. And we also construct a feature interaction layer to promote the interactions between distinct features. As a result, PMDFI reaches the average AUC of 0.9404 and 0.9415 under 5-fold and 10-fold cross-validation and successfully predicted miRNA–disease associations within three case studies.

However, there is room for improvement in the future. First, with the rapid development of sequencing technology, all types of data have exploded, and we will integrate those multi-source data to dramatically improve the robustness of the model. Second, in future researches, we would devote ourselves to discovering more original features of miRNAs and diseases to boost the performance and explore some brand-new feature calculation methods. Third, concerning the negative samples, we randomly select them from unlabeled samples, which may include unreliable false samples. To offset these negative effect on the eventual prediction, we would introduce the measurement of reliable negative samples in the future.

REFERENCES

- Ambros, V. (2001). microRNAs: tiny regulators with great potential. *Cell* 107, 823–826. doi: 10.1016/S0092-8674(01)00616-X
- Ambros, V. (2004). The functions of animal microRNAs. *Nature* 431, 350–355. doi: 10.1038/nature02871
- Anastasiadi, Z., Lianos, G. D., Ignatiadou, E., Harissis, H. V., and Mitsis, M. (2017). Breast cancer in young women: an overview. *Updat. Surg.* 69, 313–317. doi: 10.1007/s13304-017-0424-1
- Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297. doi: 10.1016/S0092-8674(04)00045-5
- Chen, H., and Zhang, Z. (2013). Similarity-based methods for potential human microRNA-disease association prediction. *BMC Med. Genomics* 6:12. doi: 10.1186/1755-8794-6-12
- Chen, X., Li, T. H., Zhao, Y., Wang, C. C., and Zhu, C. C. (2020). Deep-belief network for predicting potential miRNA-disease associations. *Brief. Bioinformatics* 16:bbaa186. doi: 10.1093/bib/bbaa186
- Chen, X., Niu, Y. W., Wang, G. H., and Yan, G. Y. (2017). Hamda: hybrid approach for miRNA-disease association prediction. *J. Biomed. Inform.* 76, 50–58. doi: 10.1016/j.jbi.2017.10.014
- Chen, X., Wang, C. C., Yin, J., and You, Z. H. (2018). Novel human miRNA-disease association inference based on random forest. *Mol. Ther. Nucleic Acids* 13, 568–579. doi: 10.1016/j.omtn.2018.10.005
- Chen, X., Yan, C. C., Zhang, X., You, Z. H., Deng, L., Liu, Y., et al. (2016). WBSMDA: within and between score for miRNA-disease association prediction. *Sci. Rep.* 6:21106. doi: 10.1038/srep21106
- Chen, X., Zhu, C. C., and Yin, J. (2019). Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput. Biol.* 15:e1007209. doi: 10.1371/journal.pcbi.1007209
- Das, J., Podder, S., and Ghosh, T. C. (2014). Insights into the miRNA regulations in human disease genes. *BMC Genomics* 15:1010. doi: 10.1186/1471-2164-15-1010
- Dean, R. M., Fowler, D. H., Wilson, W. H., Odom, J., Steinberg, S. M., Chow, C., et al. (2005). Efficacy of reduced-intensity allogeneic stem cell transplantation in chemotherapy-refractory non-hodgkin lymphoma. *Biol. Blood Marrow Transplant.* 11, 593–599. doi: 10.1016/j.bbmt.2005.04.005
- Esteller, M. (2011). Non-coding RNAs in human disease. *Nat. Rev. Genet.* 12, 861–874. doi: 10.1038/nrg3074
- Guay, C., and Regazzi, R. (2015). MicroRNAs and the functional β cell mass: for better or worse. *Diabet. Metab.* 41, 369–377. doi: 10.1016/j.diabet.2015.03.006

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

AUTHOR CONTRIBUTIONS

LD, MT, and JiL conceived the prediction method. MT and JuL wrote the paper. MT, CL, and DL developed the computer programs. CL and DL analyzed the results and revised the paper. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by National Natural Science Foundation of China (Grant Nos. 61972422 and 61672541) and the Fundamental Research Funds for the Central Universities of Central South University (Grant No. 2020zzts590).

- Horsham, J. L., Ganda, C., Kalinowski, F. C., Brown, R. A., Epis, M. R., and Leedman, P. J. (2015). MicroRNA-7: a miRNA with expanding roles in development and disease. *Int. J. Biochem. Cell Biol.* 69, 215–224. doi: 10.1016/j.biocel.2015.11.001
- Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., and Forman, D. (2011). Global cancer statistics. *CA Cancer J. Clin.* 61, 69–90. doi: 10.3322/caac.20107
- Jiang, Q., Wang, G., Jin, S., Li, Y., and Wang, Y. (2013). Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Mining Bioinform.* 8, 282–293. doi: 10.1504/IJDMB.2013.056078
- Kumarswamy, R., Volkmann, I., and Thum, T. (2011). Regulation and function of miRNA-21 in health and disease. *RNA Biol.* 8, 706–713. doi: 10.4161/rna.8.5.16154
- Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75, 843–854. doi: 10.1016/0092-8674(93)90529-Y
- Li, G., Luo, J., Xiao, Q., Liang, C., Ding, P., and Cao, B. (2017). Predicting microRNA-disease associations using network topological similarity based on deepwalk. *IEEE Access* 5, 24032–24039. doi: 10.1109/ACCESS.2017.2766758
- Li, L., Chen, X. P., and Li, Y. J. (2010). MicroRNA-146a and human disease. *Scand. J. Immunol.* 71, 227–231. doi: 10.1111/j.1365-3083.2010.02383.x
- Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., et al. (2014). HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* 42, D1070–D1074. doi: 10.1093/nar/gkt1023
- Paydas, S., Acikalin, A., Ergin, M., Celik, H., Yavuz, B., and Tanriverdi, K. (2016). Micro-RNA (miRNA) profile in hodgkin lymphoma: association between clinical and pathological variables. *Med. Oncol.* 33:34. doi: 10.1007/s12032-016-0749-5
- Pencheva, N., Tran, H., Buss, C., Huh, D., Drobnjak, M., Busam, K., et al. (2012). Convergent multi-miRNA targeting of ApoE drives LRP1/LRP8-dependent melanoma metastasis and angiogenesis. *Cell* 151, 1068–1082. doi: 10.1016/j.cell.2012.10.028
- Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019). A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* 35, 4364–4371. doi: 10.1093/bioinformatics/btz254
- Rastrelli, M., Tropea, S., Rossi, C. R., and Alaiab, M. (2014). Melanoma: epidemiology, risk factors, pathogenesis, diagnosis and classification. *In Vivo* 28, 1005–1011. doi: 10.11648/j.jctr.20160401.11
- Serpico, D., Molino, L., and Di Cosimo, S. (2014). microRNAs in breast cancer development and treatment. *Cancer Treat. Rev.* 40, 595–604. doi: 10.1016/j.ctrv.2013.11.002

- Shen, Z., Zhang, Y. H., Han, K., Nandi, A. K., Honig, B., and Huang, D. S. (2017). miRNA-disease association prediction with collaborative matrix factorization. *Complexity* 2017:2498957. doi: 10.1155/2017/2498957
- Shu, Z., Xin, S., Xu, X., Liu, L., and Kavan, L. (2018). Detecting 3d points of interest using multiple features and stacked auto-encoder. *IEEE Trans. Vis. Comput. Graph.* 25, 2583–2596. doi: 10.1109/TVCG.2018.2848628
- Song, W., Shi, C., Xiao, Z., Duan, Z., Xu, Y., Zhang, M., et al. (2019). “AutoInt: automatic feature interaction learning via self-attentive neural networks,” in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing), 1161–1170. doi: 10.1145/3357384.3357925
- van Laarhoven, T., Nabuurs, S. B., and Marchiori, E. (2011). Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 27, 3036–3043. doi: 10.1093/bioinformatics/btr500
- Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P. A. (2008). “Extracting and composing robust features with denoising autoencoders,” in *Proceedings of the 25th International Conference on Machine Learning* (Helsinki), 1096–1103. doi: 10.1145/1390156.1390294
- Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26, 1644–1650. doi: 10.1093/bioinformatics/btq241
- Wang, G., Wang, Y., Feng, W., Wang, X., Yang, J. Y., Zhao, Y., et al. (2008). Transcription factor and microRNA regulation in androgen-dependent and-independent prostate cancer cells. *BMC Genomics* 9:S22. doi: 10.1186/1471-2164-9-S2-S22
- Wang, L., You, Z. H., Chen, X., Li, Y. M., Dong, Y. N., Li, L. P., et al. (2019). LMTRDA: Using logistic model tree to predict miRNA-disease associations by fusing multi-source information of sequences and similarities. *PLoS Comput. Biol.* 15:e1006865. doi: 10.1371/journal.pcbi.1006865
- Wienholds, E., and Plasterk, R. H. (2005). MicroRNA function in animal development. *FEBS Lett.* 579, 5911–5922. doi: 10.1016/j.febslet.2005.07.070
- Xie, B., Ding, Q., Han, H., and Wu, D. (2013). miRCancer: a microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* 29, 638–644. doi: 10.1093/bioinformatics/btt014
- Yang, Z., Wu, L., Wang, A., Tang, W., Zhao, Y., Zhao, H., et al. (2017). dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res.* 45, D812–D818. doi: 10.1093/nar/gkw1079
- Zhang, H., Luo, X. Q., Zhang, P., Huang, L. B., Zheng, Y. S., Wu, J., et al. (2009). MicroRNA patterns associated with clinical prognostic parameters and cns relapse prediction in pediatric acute leukemia. *PLoS ONE* 4:e7826. doi: 10.1371/journal.pone.0007826
- Zhang, L., Chen, X., and Yin, J. (2019). Prediction of potential miRNA-disease associations through a novel unsupervised deep learning framework with variational autoencoder. *Cells* 8:1040. doi: 10.3390/cells8091040
- Zhao, Y., Wang, F., Chen, S., Wan, J., and Wang, G. (2017). Methods of microRNA promoter prediction and transcription factor mediated regulatory network. *BioMed Res. Int.* 2017:7049406. doi: 10.1155/2017/7049406
- Zhao, Y., Wang, F., and Juan, L. (2015). MicroRNA promoter identification in arabidopsis using multiple histone markers. *BioMed Res. Int.* 2015:861402. doi: 10.1155/2015/861402
- Zheng, B., Xi, Z., Liu, R., Yin, W., Sui, Z., Ren, B., et al. (2018). The function of microRNAs in B-cell development, lymphoma, and their potential in clinical practice. *Front. Immunol.* 9:936. doi: 10.3389/fimmu.2018.00936
- Zheng, K., You, Z. H., Wang, L., Zhou, Y., Li, L. P., and Li, Z. W. (2019). MLMDA: a machine learning approach to predict and validate microRNA-disease associations by integrating of heterogenous information sources. *J. Transl. Med.* 17:260. doi: 10.1186/s12967-019-2009-x
- Zheng, K., You, Z. H., Wang, L., Zhou, Y., Li, L. P., and Li, Z. W. (2020). DBMDA: A unified embedding for sequence-based miRNA similarity measure with applications to predict and validate miRNA-disease associations. *Mol. Ther. Nucleic Acids* 19, 602–611. doi: 10.1016/j.omtn.2019.12.010
- Zhou, S., Wang, S., Wu, Q., Azim, R., and Li, W. (2020). Predicting potential miRNA-disease associations by combining gradient boosting decision tree with logistic regression. *Comput. Biol. Chem.* 85:107200. doi: 10.1016/j.compbiolchem.2020.107200
- 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/bfpg/elv024

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.

Copyright © 2021 Tang, Liu, Liu, Liu and Deng. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



Slms: A Novel Insertion Detection Approach Based on Soft-Clipped Reads

Chaokun Yan¹, Junyi He¹, Junwei Luo^{2*}, Jianlin Wang¹, Ge Zhang¹ and Huimin Luo^{1*}

¹ School of Computer and Information Engineering, Henan University, Kaifeng, China, ² College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

OPEN ACCESS

Edited by:

Wang Guohua,
Harbin Institute of Technology, China

Reviewed by:

Hailin Chen,
East China Jiaotong University, China
Minzhu Xie,
Hunan Normal University, China

*Correspondence:

Junwei Luo
luojunwei@hpu.edu.cn
Huimin Luo
luohuimin@henu.edu.cn

Specialty section:

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

Received: 09 February 2021

Accepted: 06 April 2021

Published: 30 April 2021

Citation:

Yan C, He J, Luo J, Wang J,
Zhang G and Luo H (2021) Slms:
A Novel Insertion Detection Approach
Based on Soft-Clipped Reads.
Front. Genet. 12:665812.
doi: 10.3389/fgene.2021.665812

As a common type of structural variation, an insertion refers to the addition of a DNA sequence into an individual genome and is usually associated with some inherited diseases. In recent years, many methods have been proposed for detecting insertions. However, the accurate calling of insertions is also a challenging task. In this study, we propose a novel insertion detection approach based on soft-clipped reads, which is called Slms. First, based on the alignments between paired reads and the reference genome, Slms extracts breakpoints from soft-clipped reads and determines insertion locations. The insert size information about paired reads is then further clustered to determine the genotype, and Slms subsequently adopts Minia to assemble the insertion sequences. Experimental results show that Slms can achieve better performance than other methods in terms of the F-score value for simulated and true datasets.

Keywords: structural variation, alignment, short read, the next generation sequencing technology, soft-clipped read

INTRODUCTION

Although single-nucleotide polymorphisms (SNPs) represent the most frequent genomic variation, it is generally acknowledged that human genomes show more differences as a consequence of structural variations (SVs) (Gusnanto et al., 2012). SVs generally refer to genome sequence changes greater than 50 bp and can be further categorized as insertions, deletions, duplications, inversions, and translocations, among others, as well as combinations of these categories (Feuk et al., 2006; Alkan et al., 2011; Baker, 2012). Some studies have shown that phenotypic changes and some diseases are caused by SVs, e.g., autism, Parkinson's disease, and schizophrenia (Suzuki et al., 2011). Therefore, the accurate detection of SVs is of great significance for gene expression analysis and related disease research (MacConaill and Garraway, 2010). However, until a few years ago, there were no efficient methods for the detection of SVs with high precision. The development of next-generation sequencing (NGS) technology has allowed researchers to obtain a large amount of sequence data, which has improved research on SV detection (The 1000 Genomes Project Consortium, 2010; Zhang et al., 2010; Guan and Sung, 2016; Kosugi et al., 2019).

As one type of SV, an insertion refers to the addition of a DNA sequence to the genome. This sequence might be novel or could exist in the original genome, which would be equivalent to translocation or duplication. In general, insertions can be divided into two types: (i) novel insertions refer to the insertion of a sequence that cannot be found or mapped to the reference genome, and (ii) mobile element insertions or duplications constitute insertions in which the sequence comes from the original sequence. The sequence of this second type of insertion can be obtained

through a comparison with the reference genome. Based on the identification of discordant patterns in sequence data, some SV detection methods can currently be utilized to detect insertions. In general, these methods can be categorized into the following four classes: (i) paired-end mapping (PEM-based methods, such as BreakDancer (Chen et al., 2009), PEMer (Korbel et al., 2009) and GASV (Sindi et al., 2009)), which is based on the physical position and distance information of paired-end or mate-pair reads (Lee et al., 2009; Hormozdiari et al., 2010); (ii) split read (SR)-based methods, which search for split alignments of unmapped or clipped reads, and an example is CREST, which uses clipped reads to identify structural variations through multiple alignments and assembly (Wang et al., 2011); (iii) depth of coverage (DoC)-based methods such as SegSeq (Chiang et al., 2009), EWT (Yoon et al., 2009) and CNVnator (Abyzov et al., 2011)), which provide a macroscopic view of whether there is a high coverage area on the genome; and (iv) *de novo* assembly, which uses related reads to recover insertion sequences. The latter methods, such as ANISE and BASIL (Holtgrewe et al., 2015), SvABA (Wala et al., 2018), EPGA (Luo et al., 2015b) and EPGA2 (Luo et al., 2015a), require a coverage depth that is not less than 40X and have a high cost. However, these methods usually focus on abnormal information, such as variations in the insertion size and soft-clipped information, and thus cannot yield accurate detection results for insertions with variable sizes.

Some hybrid methods have been proposed for the detection of insertions with variable sizes in recent years. For example, Pindel, as a classical method, is mainly designed for deletions and small insertions and uses PEM and SR signatures to locate the breakpoints (Ye et al., 2009). However, for large insertions over 50 bp, Pindel does not perform well and yields many false positive results. MindTheGap uses a k-mer-based method to detect the insertion site and recovers insertion sequences through an assembly of k-mers (Rizk et al., 2014). This method enables the detection of small and large insertions, but the methods finds it difficult to locate a breakpoint when other polymorphisms occur near the insertion site, which leads to a high number of false negative results. As an insertion detection approach based on breakpoints, BreakSeek applies a Bayesian model for the PEM and SR signatures to find the accurate position of an insertion (Zhao and Zhao, 2015). The BreakSeek method can obtain accurate breakpoint results and genotypes without assembly, but the coverage depth of the dataset has some impact on the performance. In addition, although some insertion detection methods, such as PopIns (Kehr et al., 2016) and Pamir (Kavak et al., 2017), perform well, they may require a large number of data points.

In this paper, we propose an insertion detection approach called SIns, which is based on soft-clipped reads and achieves high insertion detection accuracy. SIns adopts PEM to identify and correct the breakpoints from a previous analysis of soft-clipped reads and clusters the insert size to determine the genotype. For sequence assembly, SIns directly extracts all abnormal reads and uses Minia to recover the insertion sequences. We conducted experiments using simulated data and real datasets, and the results show that SIns exhibits high accuracy in breakpoint detection and genotype determination.

The rest of this paper is organized as follows: in Section 2, we introduce the proposed method in detail. The experimental results are shown in Section 3, and we summarize and discuss the findings in Section 4.

METHODS

In this study, we propose a novel insertion detection approach named SIns for the detection of insertions based on soft-clipped reads. In general, SIns performs the following three steps: (i) breakpoint detection, determining the location of insertions based on comprehensive information; (ii) genotyping, identifying the genotype of the insertion based on clustering results; and (iii) assembly of insertion sequences. The overall pipeline of SIns is shown (Figure 1).

Breakpoint Detection

Breakpoint detection is an important step in SIns. In this study, the breakpoints can be obtained through the following steps.

Step 1 Selection of Soft-Clipped Reads

For each soft-clipped read, SIns first obtains its clipped part, S_c , and then extracts a sequence S_r from the reference genome, which corresponds to S_r . Note that the length of S_r equals that of S_c .

Based on the Smith-Waterman algorithm, a score matrix between S_c and S_r can then be constructed to reflect their detailed matching degree. Moreover, SIns can obtain the maximum score from the matrix, which refers to the length of the longest successive sequence. To identify and screen out real soft-clipped reads, a threshold parameter c is then set to select those reads whose S_c and S_r exhibit higher similarity. This parameter c can be computed using the following equation:

$$c = \begin{cases} 1, & \max \text{ score} < \text{cliplength} * m \\ 0, & \max \text{ score} \geq \text{cliplength} * m \end{cases} \quad (1)$$

where m represents the mappability ($m \in [0,1]$). If c equals 1, SIns selects it for the following steps; otherwise, SIns abandons it. A larger m indicates greater similarity between S_c and S_r . The default value for the parameter m is 0.5.

Step 2 Determination of Candidate Breakpoints

In our study, the soft-clipped reads were further divided into four types, namely, LL, LR, RL, and RR, which are shown in Figure 2. Taking “LL” as an example, the first L means that the left mate read is soft-clipped, and the second “L” specifies that this read is clipped on its head, whereas “RR” indicates that the right mate read is soft-clipped on its tail.

A true insertion might be related to the four types of soft-clipped reads. These soft-clipped reads can provide similar breakpoint information. In general, an insertion breakpoint is regarded strongly as true if the four types of soft-clipped reads mentioned above exist. However, it is difficult to find all types of soft-clipped reads for a true insertion, particularly if the DoC is low. In this paper, SIns defines four types of breakpoints, which are represented as {LL, LR}, {LL, RL}, {RL, LR}, and {RL, RR}. For a breakpoint, SIns collects all related soft-clipped reads that are

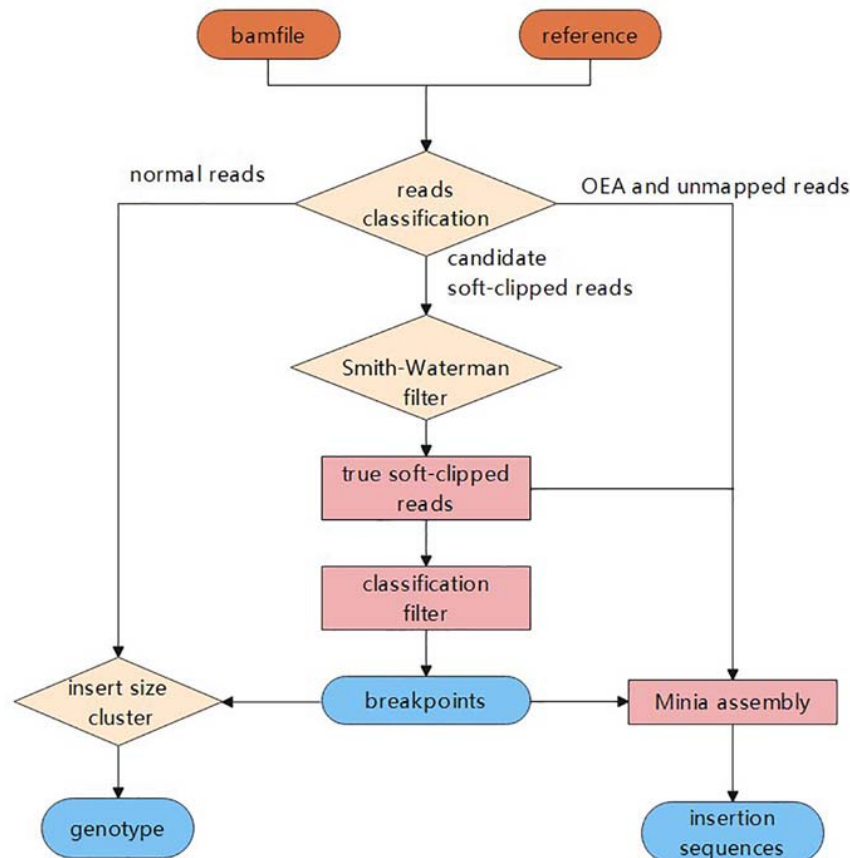


FIGURE 1 | The process of SIns.

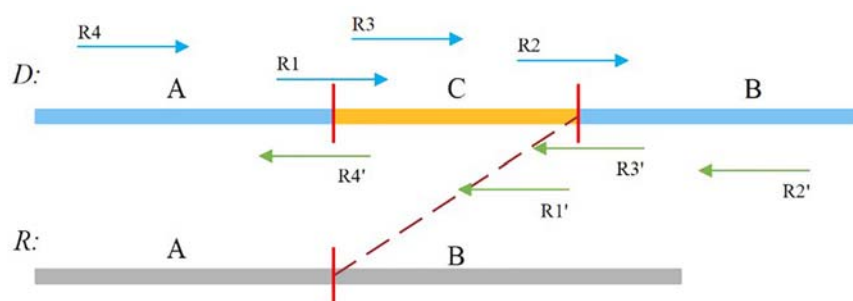


FIGURE 2 | Sequence A and B is normal, and sequence C is insertion sequence. R1, R2, R4', and R3' are soft-clipped reads. R1 belongs to the LL type, R2 belongs to the LR type, R4' belongs to the RL type, and R3' belongs to the RR type.

kept to PSD and determines their types, and SIns then uses the following equation to determine whether a breakpoint is true:

$$J = (LL \vee RL) \wedge (LR \vee RR) \quad (2)$$

where $LL \wedge LR$ indicates that the PSD of a breakpoint contains LL and LR, and $LL \vee RL$ indicates that it contains LL or RL. Subsequently, SIns obtains a list of breakpoints using the above-described method. However, the method yields

some false positive breakpoints, which can be due to a high GC content, sequencing error or SNPs. Therefore, even though their proportion is small, these breakpoints should be checked and filtered.

Step 3 Filtering of the Breakpoints

Through the above-described steps, SIns can obtain candidate breakpoints, which might include some false breakpoints. SIns then uses a filter method based on the insertion size to further

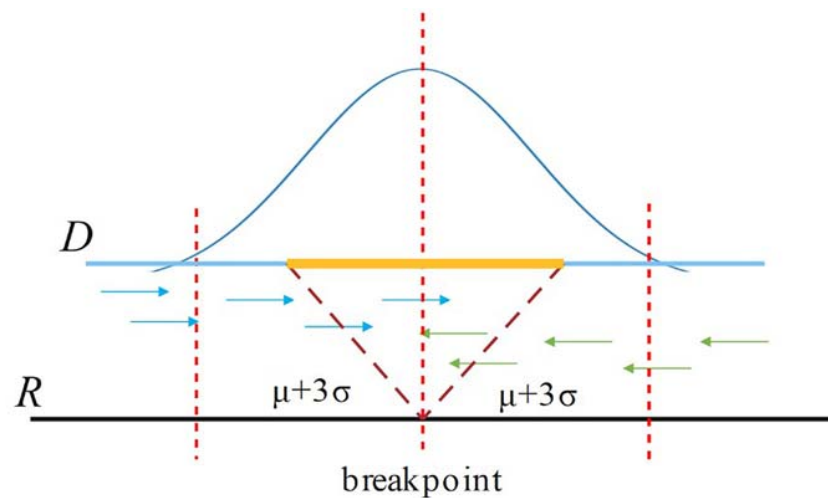


FIGURE 3 | For a breakpoint, SIns only consider reads aligned in the region $[p - (\mu + 3\sigma), p + (\mu + 3\sigma)]$, where p is the position of the breakpoint.

improve the precision of these breakpoints. An insertion usually causes a series of abnormal reads with an anomalous insert size distribution.

For a candidate breakpoint, SIns first finds the paired reads that span this breakpoint and OEA reads (one-end-anchored reads). Note that these reads should be aligned in the region $[p - (\mu + 3\sigma), p + (\mu + 3\sigma)]$, where p is the position of the breakpoint, μ is the insert size of the read library, and σ is the standard deviation of μ as shown in **Figure 3**. If the sum of paired reads and OEA reads is larger than $\text{Cov}/2$, SIns treats this breakpoint as true, otherwise, the method considers the breakpoint to be false. Cov is the coverage of the read library.

Genotyping

Genotyping is a necessary step of SIns. In a polyploid, the genotype is divided into heterozygous and homozygous genotypes. Taking diploid as an example, a heterozygous variation is only included in one chromosome and not the other one contains. In contrast, homozygosity indicates that the same variation is found in both chromosomes.

Genotyping can provide great convenience for subsequent studies, and many approaches, particularly assembly-based methods, are available for genotyping; however, all the assembly-based methods usually require considerable time and memory. Here, SIns adopts a cluster-based method, which can save as much time as possible.

If an insertion occurs, it will inevitably cause a change in the insert size for paired reads around the breakpoint, such as OEA reads, and a decrease in the normal insert size. For a heterozygous insertion, the insert size is difficult to determine because the paired reads might originate from two different chromosomes. Some paired reads contain insertions, whereas others do not. We defined $P(P_l, P_r, \text{ and } i)$ for a paired read spanning the breakpoint, where P_l is the aligning position of the left mate read, P_r is the aligning position of the right mate read and i is the insert size value around this paired read. After obtaining P

for all paired reads spanning the breakpoint, SIns applies the DBSCAN for clustering. In DBSCAN, the parameter $\text{eps} = 50$, $\text{min_samples} = 2$ in default, and these parameters can be adjusted. And, SIns determines a breakpoint as heterozygote if there is one cluster in the clustering result, otherwise, the breakpoint is deemed as homozygous. Two types of insert size distributions are shown in **Figure 4**.

Assembly Insertion Sequences

In the assembly stage, SIns extracts OEA, soft-clipped and unmapped reads for a breakpoint to recover all possible insertion sequences. After applying the Minia (Boeva et al., 2012) algorithm to these abnormal reads, SIns generate a series of sequences with overlap, which contain insertion sequences. SIns then maps these sequences to the reference genome and obtains the insertion sequence results. For example, if the CIGAR value of a candidate sequence is 132M186I130M, the algorithm finds the length of this insertion, i.e., 186 bp, and determines that the sequence content is 133–318 bases.

EXPERIMENTS AND ANALYSIS

To verify the performance of SIns, we used SURVIVOR (Jeffares et al., 2017) and ART (Huang et al., 2012) to simulate a large number of insertions on human chromosome 22 ranging in size from 50 to 1,500 bp and in coverage from 5X to 50X. The recent popular detection methods MindTheGap and BreakSeek were compared with the proposed SIns method. In addition, the real human dataset NA12878 was selected to test the performance of SIns.

Experimental Settings

Simulation Datasets and Parameter Setting

The simulation dataset was based on human chromosome 22, and the error rate of the dataset was set to 0.1%. SURVIVOR was used

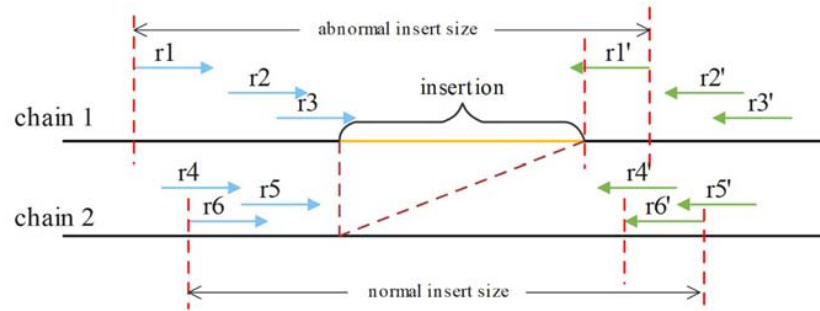


FIGURE 4 | The paired reads (r_1, r_1'), (r_2, r_2'), and (r_3, r_3') are obtained on the first chain, which contains an insertion. The other paired reads were obtained on the normal chain as shown. These insert sizes can be clustered into two clusters.

to simulate the structural variation. Here, we selected insertions for the simulation, and other types of structural variations were set to 0. ART was used to simulate different read sets from the simulated chromosome 22 containing insertions. We first generated some simulations of chromosome 22 containing insertions of different sizes, namely, 50–300 bp, 301–600 bp, 601–1,000 bp, and 1,001–1,500 bp, and ART was then used to simulate read sets with different coverages, i.e., 5X, 10X, 20X, 30X, 40X, and 50X. The read length was uniformly set to 150 bp, the inset size was 500 bp, and the standard deviation was 50. Using the above parameters, we can understand the detection ability of SIns under various conditions.

Evaluation Metrics

If the difference between the detected breakpoint and the simulated breakpoint does not exceed 10 bp, we consider it a positive result, which is represented by TP; otherwise, the result is represented by FP. True breakpoints that were not detected are indicated by FN. To clearly show the detection performance of various methods, we used the metrics precision (Pr), recall (Rc) and F-score as follows:

$$P_r = \frac{TP}{TP + FP} \times 100\% \quad (3)$$

$$R_c = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

The F-score was defined as the harmonic average of precision and recall:

$$F_{score} = \frac{2P_r \times R_c}{P_r + R_c} \times 100\% \quad (5)$$

Simulation Dataset

Results on Homozygous Dataset

We compared SIns with MindTheGap and BreakSeek, selected chromosome 22 as the reference and simulated a chromosome containing 1,051 insertions of 50–300 bp, a chromosome containing 597 insertions of 301–600 bp, a chromosome containing 597 insertions of 601–1,000 bp and a chromosome containing 790 insertions of 1,001–1,500 bp. Based on different coverages, we simulated six read sets for

each simulated chromosome. The experimental results are shown in **Table 1**.

As shown in **Table 1**, the performances of SIns and BreakSeek in detecting insertions of 50–300 bp were better. Although the precision of BreakSeek was generally higher than that of SIns, its F-score was only better than that of SIns when the coverages of the read set were 40X and 50X. We also found that SIns has a higher recall, which means that SIns can detect more true insertions. SIns exhibited higher precision and recall regardless of the coverage and the length of insertions. In addition, none of the methods worked well with low DoCs. However, for the case with a low coverage ($\text{DoC} \leq 10\text{X}$), SIns showed better performance than the other methods.

Results on Heterozygous Dataset

To verify the performance of SIns in detecting heterozygous insertions, we simulated read sets of chromosome 22. Simulations of chromosome 22 containing insertions of 50–300 bp were used to produce these read sets, and other simulations of chromosome 22 containing an insertion of 301–600 bp were also used to generate other read sets. We then combine the read sets from the normal chromosome 22 and the simulations of chromosome 22. Note that the read sets were simulated with different coverages: 10X, 20X, 40X, 60X, and 80X. The experimental results are shown in **Table 2**.

As illustrated in **Table 2**, the detection results obtained with MindTheGap were less effective than those obtained with homozygous detection because MindTheGap has more sequences to choose from when selecting k-mers, which will yield some conflicting issues. The performance of BreakSeek on these two datasets was not as good as the results obtained with homozygotes, and a reason for this finding might be that normal reads extracted from the reference genome, which contained many contradictory PEM and SR information, were added. When BreakSeek iteratively analyses the PEM signature, there is too much contradictory information that can be used, and thus, the result cannot show the most authentic SV information. In contrast, when SIns extracts breakpoint information at the initial stage, the method relies more on SR information and thus experiences less interference from contradictory information. At the subsequent filtering stage, due to the addition of normal

TABLE 1 | Comparison of three tools for four ranges.

Doc	Tool	50-300			301-600			601-1,000			1,001-1,500		
		Pr	Rc	F-score	Pr	Rc	F-score	Pr	Rc	F-score	Pr	Rc	F-score
5X	Slms	99.784	87.726	93.367	100	64.992	78.782	100	61.977	76.525	100	63.924	77.992
	BreakSeek	99.791	45.48	62.484	100	14.405	25.183	98.592	11.725	20.958	100	11.899	21.267
	MindTheGap	11.949	26.546	16.48	2.317	27.471	4.274	3.104	26.801	5.563	4.551	29.494	7.885
10X	Slms	99.412	96.48	97.924	99.815	90.62	94.996	100	89.615	94.523	100	90.127	94.807
	BreakSeek	99.892	87.631	93.36	100	61.809	76.398	99.701	55.946	71.674	99.774	55.823	71.591
	MindTheGap	30.356	64.986	41.381	20.918	65.662	31.728	21.315	67.337	32.38	25.962	67.468	37.496
20X	Slms	99.037	97.812	98.42	99.65	95.477	97.519	100	93.802	96.802	99.868	95.57	97.671
	BreakSeek	99.603	95.433	97.473	99.27	91.122	95.022	99.259	89.782	94.283	99.447	91.013	95.043
	MindTheGap	85.845	80.209	82.932	75.955	79.899	77.878	73.242	80.235	76.579	79.597	80	79.798
30X	Slms	98.848	98.002	98.423	99.308	96.147	97.702	100	94.807	97.334	99.867	95.316	97.539
	BreakSeek	99.509	96.384	97.922	99.298	94.807	97.001	99.284	92.965	96.021	99.459	93.165	96.209
	MindTheGap	86.829	81.541	84.102	77.564	81.072	79.279	75.425	81.742	78.457	80.73	81.139	80.934
40X	Slms	98.102	98.382	98.242	100	96.482	98.21	99.825	95.477	97.603	99.868	95.949	97.87
	BreakSeek	99.708	97.431	98.556	99.123	94.64	96.829	99.295	94.305	96.735	99.597	93.797	96.61
	MindTheGap	86.917	81.541	84.143	77.404	80.905	79.115	75.889	82.245	78.939	80.832	81.139	80.985
50X	Slms	98.57	98.382	98.476	98.969	96.482	97.71	100	95.477	97.686	99.869	96.203	98.001
	BreakSeek	99.708	97.431	98.556	98.614	95.31	96.934	99.118	94.137	96.564	99.338	94.937	97.087
	MindTheGap	87.018	81.637	84.242	77.28	80.905	79.051	75.153	82.077	78.463	80.881	81.392	81.136

The bold values represent the highest value of each data set in different depth.

TABLE 2 | Result of 50–300 and 301–600 bp heterozygous insertions.

50-300	Tool	50-300			301-600		
		Pr	Rc	F-score	Pr	Rc	F-score
10X	Slms	100	92.959	96.351	100	89.782	94.616
	BreakSeek	100	33.111	49.75	100	21.441	35.31
	MindTheGap	11.275	21.789	14.86	5.211	22.111	8.435
20X	Slms	99.903	97.907	98.895	100	96.985	98.469
	BreakSeek	99.707	64.7	78.477	100	48.576	65.389
	MindTheGap	88.596	57.659	69.856	79.669	56.449	66.078
40X	Slms	99.807	98.573	99.186	100	97.99	98.985
	BreakSeek	98.847	65.271	78.625	98.805	41.541	58.491
	MindTheGap	98.609	67.46	80.113	97.387	68.677	80.55
60X	Slms	99.425	98.763	99.093	100	97.99	98.985
	BreakSeek	98.389	63.939	77.509	98.214	46.064	62.714
	MindTheGap	99.349	72.598	83.892	98.42	73.032	83.846
80X	Slms	99.616	98.858	99.236	100	97.99	98.985
	BreakSeek	98.503	62.607	76.556	98.264	47.404	63.955
	MindTheGap	98.84	72.978	83.963	98.42	73.032	83.846

The bold values represent the highest value of each data set in different depth.

reads, the filtering conditions were more rigorous and precise, which explains why the precision of SIns increased, whereas the recall value decreased.

Experiments Based on Real Dataset

NA12878 is the gold standard dataset commonly used in genomics. Experiments with NA12878 (ERR194147 50X¹) samples were conducted using the SIns, MindTheGap and

BreakSeek methods. We extracted the reads with a probability of 0.1 because the coverage was too high. The generally recognized VCF file of this sample contains 50,016 insertion reports larger than 50 bp. The corresponding vcf file can be downloaded from NCBI. We only selected the detected results in the file records as true values. The test results are shown in Table 3.

We have filtered out the SNPs and Indels of this data set. The above results show that SIns has good performance on

TABLE 3 | Results obtained with NA12878.

	SIns	MindTheGap	BreakSeek
chr1	123	98	90
chr2	180	136	74
chr3	107	57	38
chr4	105	87	37
chr5	94	68	44
chr6	117	84	43
chr7	134	91	44
chr8	73	72	43
chr9	77	69	48
chr10	101	62	42
chr11	88	46	41
chr12	99	65	46
chr13	66	27	36
chr14	51	29	28
chr15	42	44	29
chr16	88	63	69
chr17	67	46	29
chr18	72	42	27
chr19	67	46	23
chr20	38	50	25
chr21	57	16	24
chr22	28	27	21

¹ <http://www.ebi.ac.uk/ena>

TABLE 4 | Homozygote results obtained with four ranges.

Doc	50–300			301–600			601–1,000			1,001–1,500		
	Mind TheGap	Break Seek	SIns	Mind TheGap	Break Seek	SIns	Mind TheGap	Break Seek	SIns	Mind TheGap	Break Seek	SIns
5X	176s	1868s	20s	174s	1842s	35s	178s	2130s	29s	177s	2127s	31s
10X	217s	1868s	40s	216s	2250s	68s	227s	2156s	65s	212s	2089s	61s
20X	243s	2177s	77s	242s	2178s	119s	242s	2054s	142s	235s	4349s	123s
30X	264s	2249s	116s	264s	2109s	180s	257s	3723s	191s	203s	5281s	184s
40X	284s	2415s	154s	286s	2589s	250s	292s	4948s	240s	204s	2736s	245s
50X	304s	2577s	193s	310s	2943s	343s	310s	3207s	319s	211s	2539s	307s

TABLE 5 | Heterozygous results obtained with four ranges.

Doc	50–300			301–600		
	Mind TheGap	Break Seek	SIns	Mind TheGap	Break Seek	SIns
10X	140s	1997s	38s	139s	2020s	19s
20X	152s	2041s	76s	154s	1990s	47s
40X	171s	2224s	150s	180s	2495s	84s
60X	190s	2779s	227s	193s	2869s	122s
80X	212s	2703s	305s	215s	3294s	204s
100X	227s	3634s	425s	254s	3719s	259s

most chromosomes compared with MindTheGap and BreakSeek. Although the detection number of insertions on chromosome 15 and 20 are lower than that of MindTheGap, we can find the result on the rest of chromosomes are better than other two methods. And the average of F-score on all 22 chromosomes is 5.46% for SIns. MindTheGap is 2.42%, and BreakSeek is 2.85%. The average of F-score shows the same conclusion.

Running Time Comparison

Here we list the time comparison results of homozygote and heterozygous experiments.

Although clustering is useful in the SIns process, it does not require as many iterations as in BreakSeek, MindTheGap and other methods; thus, SIns exhibits a relatively obvious advantage in terms of running time. As shown in **Tables 4, 5**, all the methods were run in the same machine and a single thread by default. As a result, SIns exhibited better performance than the other two methods in most cases. The main time-consuming step of SIns is the third step: the reads used for assembly are extracted from the original read collection, which is the most work-intensive step. If the assembly is not considered and the method aims to just detect breakpoints and judge genotypes, SIns can complete the task within a short time.

DISCUSSION

In this article, we propose an insertion detection method named SIns based on the comprehensive processing of soft-clipped read information. SIns can provide more precise detection of breakpoints and can perform relatively accurate genotyping. In addition, SIns uses the Minia algorithm for assembly of the insertion sequence, and the successfully assembled sequence is then filtered and tailored according to the

breakpoint information. After these steps, the complete insertion sequence is provided.

Most of the existing methods show effectiveness in detecting small insertions but show poor performance in cases of low coverage. These methods usually are difficult to detect all types of SVs of all sizes. SIns focuses on the detection of insertions of different sizes. We tested the detection performance of SIns using various simulated datasets and compared it with MindTheGap and BreakSeek. In most cases, the performance of SIns was better than those of the other two methods. Comparing with the other two methods, SIns performs well both on low and high coverage data sets and different size insertions. The experimental results using a real dataset show that SIns exhibits good detection capability.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <http://www.ebi.ac.uk/ena>.

AUTHOR CONTRIBUTIONS

CY and JL conceived and designed the approach. JH performed the experiments. JW and GZ analyzed the data. JH and JL wrote the manuscript. JL and HL supervised the whole study process and revised the manuscript. All authors have read and approved the final version of manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (Nos. 61972134, 61802113, and 61802114). Science and Technology Development Plan Project of Henan Province, (Nos. 202102210173 and 212102210091). China Postdoctoral Science Foundation (No. 2020M672212). Henan Province Postdoctoral Research Project Funding.

ACKNOWLEDGMENTS

This paper is recommended by the 5th Computational Bioinformatics Conference.

REFERENCES

- Abyzov, A., Urban, A. E., Snyder, M., and Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res.* 21, 974–984. doi: 10.1101/gr.114876.110
- Alkan, C., Coe, B. P., and Eichler, E. E. (2011). Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376. doi: 10.1038/nrg2958
- Baker, M. (2012). Structural variation: the genome's hidden architecture. *Nat. Methods* 9, 133–137. doi: 10.1038/nmeth.1858
- Boeva, V., Popova, T., Bleakley, K., Chiche, P., Cappel, J., Schleiermacher, G., et al. (2012). Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 28, 423–425. doi: 10.1093/bioinformatics/btr670
- Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods* 6, 677–681. doi: 10.1038/nmeth.1363
- Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L., et al. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods* 6, 99–103. doi: 10.1038/nmeth.1276
- Feuk, L., Carson, A. R., and Scherer, S. W. (2006). Structural variation in the human genome. *Nat. Rev. Genet.* 7, 85–97.
- Guan, P., and Sung, W.-K. (2016). Structural variation detection using next-generation sequencing data: a comparative technical review. *Methods* 102, 36–49. doi: 10.1016/j.ymeth.2016.01.020
- Gusnanto, A., Wood, H. M., Pawitan, Y., Rabbitts, P., and Berri, S. (2012). Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics* 28, 40–47. doi: 10.1093/bioinformatics/btr593
- Holtgrewe, M., Kuchenbecker, L., and Reinert, K. (2015). Methods for the detection and assembly of novel sequence in high-throughput sequencing data. *Bioinformatics* 31, 1904–1912. doi: 10.1093/bioinformatics/btv051
- Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., et al. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. *Bioinformatics* 26, i350–i357.
- Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics* 28, 593–594. doi: 10.1093/bioinformatics/btr708
- Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., et al. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* 8:14061.
- Kavak, P., Lin, Y.-Y., Numanagiae, I., Asghari, H., Güngör, T., Alkan, C., et al. (2017). Discovery and genotyping of novel sequence insertions in many sequenced individuals. *Bioinformatics* 33, i161–i169.
- Kehr, B., Melsted, P., and Halldórsson, B. V. (2016). PopIns: population-scale detection of novel sequence insertions. *Bioinformatics* 32, 961–967. doi: 10.1093/bioinformatics/btv273
- Korbel, J. O., Abyzov, A., Mu, X. J., Carriero, N., Cayting, P., Zhang, Z., et al. (2009). PEmr: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.* 10:R23.
- Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M., and Kamatani, Y. (2019). Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.* 20:117.
- Lee, S., Hormozdiari, F., Alkan, C., and Brudno, M. (2009). MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods* 6, 473–474. doi: 10.1038/nmeth.f.256
- Luo, J., Wang, J., Li, W., Zhang, Z., Wu, F.-X., Li, M., et al. (2015a). EPGA2: memory-efficient de novo assembler. *Bioinformatics* 31, 3988–3990.
- Luo, J., Wang, J., Zhang, Z., Wu, F.-X., Li, M., and Pan, Y. (2015b). EPGA: de novo assembly using the distributions of reads and insert size. *Bioinformatics* 31, 825–833. doi: 10.1093/bioinformatics/btu762
- MacConaill, L. E., and Garraway, L. A. (2010). Clinical implications of the cancer genome. *J. Clin. Oncol.* 28:5219. doi: 10.1200/jco.2009.27.4944
- Rizk, G., Gouin, A., Chikhi, R., and Lemaitre, C. (2014). MindTheGap: integrated detection and assembly of short and long insertions. *Bioinformatics* 30, 3451–3457. doi: 10.1093/bioinformatics/btu545
- Sindi, S., Helman, E., Bashir, A., and Raphael, B. J. (2009). A geometric approach for classification and comparison of structural variants. *Bioinformatics* 25, i222–i230.
- Suzuki, S., Yasuda, T., Shiraishi, Y., Miyano, S., and Nagasaki, M. (2011). ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics* 12(Suppl 14):S7. doi: 10.1186/1471-2105-12-S14-S7
- The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* 467:1061. doi: 10.1038/nature09534
- Wala, J. A., Bandopadhyay, P., Greenwald, N. F., O'Rourke, R., Sharpe, T., Stewart, C., et al. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* 28, 581–591. doi: 10.1101/gr.221028.117
- Wang, J., Mullighan, C. G., Easton, J., Roberts, S., Heatley, S. L., Ma, J., et al. (2011). CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat. Methods* 8, 652–654. doi: 10.1038/nmeth.1628
- Ye, K., Schulz, M. H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25, 2865–2871. doi: 10.1093/bioinformatics/btp394
- Yoon, S., Xuan, Z., Makarov, V., Ye, K., and Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* 19, 1586–1592. doi: 10.1101/gr.092981.109
- Zhang, Q., Ding, L., Larson, D. E., Koboldt, D. C., McLellan, M. D., Chen, K., et al. (2010). CMDS: a population-based method for identifying recurrent DNA copy number aberrations in cancer from high-resolution data. *Bioinformatics* 26, 464–469. doi: 10.1093/bioinformatics/btp708
- Zhao, H., and Zhao, F. (2015). BreakSeek: a breakpoint-based algorithm for full spectral range INDEL detection. *Nucleic Acids Res.* 43, 6701–6713. doi: 10.1093/nar/gkv605

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.

Copyright © 2021 Yan, He, Luo, Wang, Zhang and Luo. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.



RetroScan: An Easy-to-Use Pipeline for Retrocopy Annotation and Visualization

Zhaoyuan Wei^{1,2}, Jiahe Sun², Qinhui Li¹, Ting Yao¹, Haiyue Zeng² and Yi Wang^{1,2*}

¹ State Key Laboratory of Silkworm Genome Biology, Biological Science Research Center, Southwest University, Chongqing, China, ² Biological Science Research Center, Southwest University, Chongqing, China

OPEN ACCESS

Edited by:

Chunhou Zheng,
Anhui University, China

Reviewed by:

Margaret Woodhouse,
Agricultural Research Service,
United States Department
of Agriculture, United States
Izabela Makalowska,
Adam Mickiewicz University, Poland
Jin-Xing Liu,
Qufu Normal University, China

*Correspondence:

Yi Wang
yiwang28@swu.edu.cn

Specialty section:

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

Received: 02 June 2021

Accepted: 26 July 2021

Published: 16 August 2021

Citation:

Wei Z, Sun J, Li Q, Yao T, Zeng H
and Wang Y (2021) RetroScan: An
Easy-to-Use Pipeline for Retrocopy
Annotation and Visualization.
Front. Genet. 12:719204.
doi: 10.3389/fgene.2021.719204

Retrocopies, which are considered “junk genes,” are occasionally formed via the insertion of reverse-transcribed mRNAs at new positions in the genome. However, an increasing number of recent studies have shown that some retrocopies exhibit new biological functions and may contribute to genome evolution. Hence, the identification of retrocopies has become very meaningful for studying gene duplication and new gene generation. Current pipelines identify retrocopies through complex operations using alignment programs and filter scripts in a step-by-step manner. Therefore, there is an urgent need for a simple and convenient retrocopy annotation tool. Here, we report the development of RetroScan, a publicly available and easy-to-use tool for scanning, annotating and displaying retrocopies, consisting of two components: an analysis pipeline and a visual interface. The pipeline integrates a series of bioinformatics software programs and scripts for identifying retrocopies in just one line of command. Compared with previous methods, RetroScan increases accuracy and reduces false-positive results. We also provide a Shiny app for visualization. It displays information on retrocopies and their parental genes that can be used for the study of retrocopy structure and evolution. RetroScan is available at <https://github.com/Vicky123wzy/RetroScan>.

Keywords: retrocopy, pipeline, evolution, visualization, genome

INTRODUCTION

Gene duplications, which are generated by DNA- or RNA-mediated mechanisms (Innan and Kondrashov, 2010; Sakai et al., 2011), are a major source of the origination of new genes (Long et al., 2003) and play pivotal roles in genome evolution, new biological process origination and functional diversification (Flagel and Wendel, 2009). Retrocopies are a special type of RNA-mediated duplication (Brosius, 1991) in which the reverse transcripts of mRNAs derived from parental genes are occasionally reinserted at an ectopic location in the genome (Long et al., 2003). Retrocopies are new sequence fragments formed by retrotransposition events. Most retrocopies are non-functional due to their insertion at inappropriate sites or a lack of parental gene features such as introns or regulatory elements and are believed to be retropseudogenes (Lynch and Conery, 2000; Navarro and Galante, 2013). Another group of retrocopies may inherit the complete open reading frames (ORFs) of the parental genes or recruit regulatory elements such as promoters, enhancers and coding sequences from flanking regions to generate a functional retrogene (Pan and Zhang, 2009). Furthermore, the fusion of a retrocopy with coding sequences near the

insertion site generates a chimeric gene (Betran et al., 2002; Wang et al., 2002). Recent studies have systematically identified a substantial number of retrocopies in the genomes of fruit flies (Bai et al., 2007), *Caenorhabditis elegans* (Schridder et al., 2011), humans (Ohshima et al., 2003; Zhang et al., 2003; Vinckenbosch et al., 2006), zebrafish (Fu et al., 2010), and other mammals (Pan and Zhang, 2009). Some studies have searched for retrocopies in plant genomes, mainly in *Arabidopsis thaliana* (Zhang et al., 2005), rice (Sakai et al., 2011), poplar (Zhu et al., 2009), and green algae (Jąkowski et al., 2016). Moreover, some functions of retrocopies have been verified through experiments; for example, Jingwei functions in the metabolism of recruitment pheromones and juvenile hormones in fruit flies (Long and Langley, 1993; Zhang et al., 2010), and CYP98A8 and CYP98A9 are involved in pollen development in *Arabidopsis thaliana* (Matsuno et al., 2009). Retrocopies not only contribute to the diversity of genome sequences but can also cause rapid and significant changes in the genome by altering genome structures. Therefore, they are an important driving force for the origination of new genes (Carelli et al., 2016) and provide evidence of evolutionary innovations (Navarro and Galante, 2015). With the rapid development of next-generation sequencing technology, many studies have assembled chromosome-level genomes of new species, and a tool for annotating retrocopies at the genome-wide level would help us to fully understand their positions in the genome and the process of their production. Such a tool would be highly significant for studying genome evolution and subsequently analyzing the function of retrocopies (Kaessmann et al., 2009).

Since retrocopies have often lost introns and but are otherwise highly similar to their parental genes, the identification of retrocopies in the whole genome is generally based on the use of protein sequences as templates for sequence alignment. Current retrocopy identification pipelines are based mainly on the TBLASTN, BLAT, and paralog methods (Casola and Betrán, 2017). Most studies of retrocopies are based on the TBLASTN method, which aligns the annotated protein-coding sequences to whole-genome sequences. Candidate hits are determined by alignment with parental genes to determine the numbers of lost introns, point mutations and frameshift mutations using FASTA (Pearson and Lipman, 1988) and GENEWISE (Birney et al., 2004). This method has been used to find retrocopies in humans (Vinckenbosch et al., 2006), *Caenorhabditis elegans* (Abdelsamad and Pecinka, 2014), *Arabidopsis thaliana* (Zhang et al., 2005), rice (Sakai et al., 2011), poplar (Zhu et al., 2009), and green algae (Jąkowski et al., 2016). However, the speed of the TBLASTN method is relatively slow, and scanning a large genome often takes several days or even a few weeks. But Kabza et al. (2014) were the first use LAST to identify retrocopies instead of TBLASTN, which greatly improved the speed of alignment. The use of BLAT to align genomic sequences with cDNA sequences instead of proteins is also a good option. The BLAT method directly estimates the number of missing introns according to the alignment results without additional programs. However, compared with the TBLASTN method, the BLAT method shows lower accuracy, and some positive retrocopies will be ignored. This is not conducive to further evolutionary analysis because the BLAT method cannot get the proteins mutations

information between parental genes and retrocopies. Navarro and Galante (2015) used the BLAT method to scan for retrocopies in seven primate genomes, and the PlantRGDB database provides annotations for the retrocopies of 49 plant genomes (Wang, 2017). Moreover, a new method developed by Abdelsamad and Pecinka (2014) divides the annotated genes into two types, intron-free genes and intron-containing genes, and then aligns them using paralogs to identify retrocopies. Compared to the previous two methods, this approach can find more retrocopies in intron-free genes but also produces more false-positive results. It is impossible to find retropseudogenes via the paralog method because it focuses only on annotated genes rather than genome sequences. All of the above methods for identifying retrocopies present some disadvantages. Therefore, there is an urgent need to develop a comprehensive and uncomplicated tool for identifying, annotating and analyzing retrocopies in the genome which could facilitate in-depth research on retrocopies.

In the development of an easy-to-use retrocopy identification pipeline, the following requirements must be met. First, the increasing number of genome sequences generated by high-throughput sequencing technology have brought retrocopy research a new era, so the new pipeline must be suitable for various species, including large-scale genomes. Second, it must be convenient for users to configure and run, requiring few extra operations. Third, it should effectively reduce false-positive results. Finally, all results should be clearly displayed in the form of clear figures. To meet all of these design needs, we developed a convenient and accurate tool, RetroScan,¹ which is based on the method of aligning protein sequences with genome sequences to recognize retrocopies by integrating multiple software programs and scripts. Next, RetroScan was used to explore the expression, age distribution and functions of the retrocopies. Finally, we constructed a reliable graphical interface to display the results, thus helping researchers to easily obtain information on retrocopies and achieve a deep understanding them.

MATERIALS AND METHODS

RetroScan is an easy-to-use tool for retrocopy identification that integrates a series of bioinformatics tools [LAST (Kielbasa et al., 2011), BEDtools (Quinlan and Hall, 2010), ClustalW2 (Larkin et al., 2007), KaKs_Calculator (Wang et al., 2010), HISAT2 (Kim et al., 2015), StringTie (Pertea et al., 2015), SAMtools (Li et al., 2009), and Shiny] and scripts. It scans retrocopies based on alignments between protein-coding genes and whole-genome sequences. This tool can also analyze heterosense substitution and synonymous substitution, compare gene structure between parental genes and retrocopies, and calculate corresponding expression values. Moreover, RetroScan has a user-friendly visualization interface that provides overall statistical information, a retrocopy structure diagram, the non-synonymous/synonymous substitution (Ka/Ks) ratio distribution and the fragments

¹<https://github.com/Vicky123zyw/RetroScan>

per kilobase per million (FPKM) heatmap using the Shiny package in R.

Retrocopy Identification

RetroScan mainly relies on the identification of genomic intronless alignments from mature transcripts (mRNAs) for the reason that retrocopies are processed copies of multiexon proteins. It requires at least two input files (**Figure 1**): a genome sequence file (FASTA format) and a corresponding annotation file (GFF format), from which it can provide detailed information on retrocopies and parental genes in the genome. If users wish to obtain the expression values of retrocopies, they need to submit additional RNA-Seq data.

According to genome sequences and GFF file (**Figure 1**), RetroScan first employs the peptide sequences used as queries in similarity searches against complete genome sequences using LAST to identify candidate hits. To avoid duplicate results, the longest transcripts of each gene for alignment are retained for the next step. Multiexon proteins are selected for subsequent analysis because the parental genes must lose at least two introns. According to the alignment results from the previous step, users can set the sequence identity, coverage and alignment length parameters to consider the specific conditions of the species. Multiple alignment hits to the same genomic locus are clustered using BEDTools. When the distance between the hits is less than a certain length, indicating that they are unlikely to be separated by introns, adjacent homology hits are merged using BEDTools. The gap default is 40 bp in RetroScan, but if users want to change this threshold, they should take into consideration that the length of most introns ought to be larger than the threshold.

Next, the merged sequences are aligned back to multiexon proteins using LAST, and the best hits are retained as putative parental genes. Finally, the number of lost introns is estimated to obtain reliable results according to the alignment output. We calculate the position of the introns on the protein sequences according to the annotation file. RetroScan only retains parental genes (excluding the first and last 10 amino acids) that span at least two introns and single-exon retrocopies. We discard any cases involving possible DNA-based duplications by aligning retrocopy sequences back to genome sequences to minimize the number of false-positive results. If a retrocopy shows multiple highly similar sequences in the genome, it will be deleted.

In addition, retrocopies with either premature stop codons or frameshift mutations are defined as retropseudogenes; otherwise, they are defined as intact retrocopies. If one intact retrocopy can recruit novel regulatory elements or new protein-coding exons and evolve into a functional retrogene, it can be defined as a chimeric retrogene. RetroScan is more convenient and easier to use, which integrates multiple softwares and there is no need for the user to call the softwares at each step. Compared with the traditional processes, LAST alignment is faster. We also align the results of retrocopy back to the genome to avoid rertocopy caused by DNA duplication, which effectively reduces false positives.

Ka/Ks Analysis

The age distribution of the retrocopies (**Figure 1**) is determined by calculating Ka, Ks and the Ka/Ks ratio between each retrocopy

and its parental gene. The coding sequence (CDS) information of the retrocopies and their parental genes based on the annotation file are extracted for Ka/Ks calculation. Then, RetroScan performs multiple alignments between the corresponding protein sequences using ClustalW2. Finally, the Ka, Ks, and Ka/Ks values are calculated using KaKs_calculator_2.0.

Retrocopy Expression Analysis

Although the sequences of the parental genes and retrocopies are similar, some retrocopies are not expressed, which implies that they have no function. Some retrocopies exhibit expression patterns similar to those of their parental genes and may have similar functions, and some retrocopies exhibit much higher expression values than their parental genes, which means that they may replace the parental gene function. Therefore, analyzing the expression of retrocopies in different tissues and organs is helpful for exploring their functions. As retrocopies show high similarity with their parental genes, the expression values of them might be biased by the lack of RNA-seq reads mapping uniquely to either copy. There are two factors that could possibly cause this. First, it is well known that retrocopies have very low expression and are usually limited to one or a few tissues (Carelli et al., 2016). Secondly, sequences that matched equally well to a given retrogene progenitor were excluded what additionally reduced the number of positive results (Rosikiewicz et al., 2017). To estimate the expression values of retrogenes (**Figure 1**), RetroScan uses HISAT2, SAMtools, and StringTie to analyze the RNA-Seq data based on retrocopy and parental gene position information, which has the advantages of high accuracy and fast speed. After the reads are mapped to the corresponding annotated sequences using HISAT2, RetroScan converts SAM files into BAM files and sorts them using SAMtools. Finally, StringTie calculates FPKM values, which are helpful for analyzing differential expression. All programs are run with the default settings.

Visualization

We developed a visual interface that can clearly display retrocopy structure, the ka/ks distribution, expression levels, sequence alignments and statistical figures. We use R to analyze the RetroScan results, while the web pages are mainly built with Shiny and a series of R packages such as ggplot2, UpSetR, ggmsa, VennDiagram, dplyr, DT, shinydashboard, Biostrings, muscle, pheatmap, stringr, shinyjs, RColorBrewer, ape, etc. The interface layout is divided into four parts: summary, retrocopy, KaKs and expression. Users can upload the RetroScan result files generated by RetroScan through the START button on the homepage.

The “Summary” page shows the RetroScan results and related statistical information which are mainly displayed in the form of tables, histograms, pie charts, line graphs, Venn diagrams, heat maps, and so on. There is a table containing all of the information for retrocopies and their parental genes, including the retrocopy ID, chromosome, start site and end site of the retrocopy; the parental gene ID, identity, coverage, and description; and the host gene ID (**Figure 2A**). The other parts of the page show seven statistical figures illustrating the chromosome distribution of the parental genes corresponding

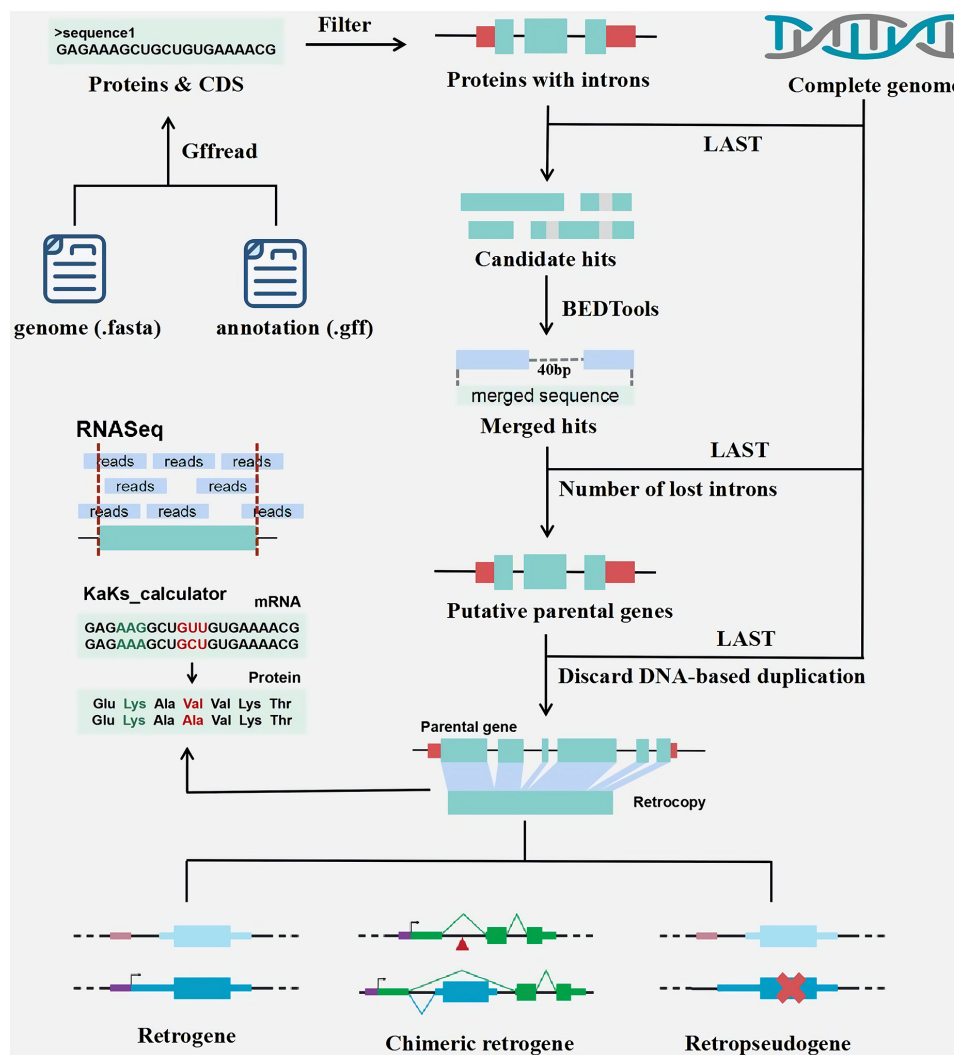


FIGURE 1 | The pipeline of retrocopy annotation.

to the retrocopies on each chromosome (**Figure 2B**), the distribution of the number of retrocopies of by each parental gene, the retrocopy length distribution, the percentage of identity (**Figure 2C**), the percentage of coverage and the percentage of retrospseudogenes, intact retrocopies and chimeric retrocopies. The static UpSet plot (Conway et al., 2017; **Figure 2D**) visualizes the intersections of datasets showing an identity $\geq 90\%$, ≥ 3 lost introns, host genes, a $Ka/Ks \leq 0.1$, and coverage $\geq 90\%$ in a matrix layout and introduces aggregates based on groupings and queries. The upper bar graph corresponds to the lower dot matrix graph including the intersections of related datasets.

The “Retrocopy” page includes a search box where users can enter any retrocopy ID. The search result integrates the detailed information, sequence structure, alignment and expression of a certain retrocopy. The structure figure (**Figure 2E**) shows the structural differences in the gene sequences among the parental genes, retrocopies and host genes so that users can clearly understand the formation of retrocopies from parental genes.

The sequence section contains the sequences of the retrocopy gene and protein sequences (**Figure 2F**). The alignment section shows the sequence alignment between the retrocopy and the parental gene to allow users to identify the differences in bases (**Figure 2G**). The expression patterns in different developmental stages and tissues could be used as a basis for judging whether a retrocopy has a biological function and whether there is functional correlation between the retrocopy and its parental gene. The page displays the expression values in a line chart in which two lines represent the expression of the retrocopy and the parental gene (**Figure 2H**).

A Ka/Ks table and four statistical figures are provided to investigate the origin and evolution of retrocopies on the “ Ka/Ks ” page. Users can view the table of Ka , Ks , and Ka/Ks values and set reasonable thresholds for filtering retrocopies. The age distribution is shown with a Ks histogram and is estimated by comparing the protein sequences of the parental genes and retrocopies (**Figure 2I**). Another Ks histogram shows the



FIGURE 2 | Visualization of the retrocopy results. **(A)** The table contains all information on retrocopies and parental genes. **(B)** The chromosome distribution of retrocopies. **(C)** The percentage of identity. **(D)** The UpSet plot visualizes the intersections of datasets showing an identity $\geq 90\%$, ≥ 3 lost introns, host genes, $Ka/Ks \leq 0.1$ and coverage $\geq 90\%$. **(E)** The structure figure shows the differences in the gene sequences between the parental genes and retrocopies. **(F)** The retrocopy sequence and the parental gene mRNA and protein sequences. **(G)** Sequence alignment between the retrocopy and parental gene. **(H)** The expression values of retrocopy and parental genes. **(I)** Ks distribution histograms. **(J)** Histogram showing the mean FPKM values of retrocopies (blue bar) and parental genes (brown bar) in all tissues. **(K)** Heatmap showing the expression of all retrocopies.

Ks distribution in three categories: retropseudogenes, intact retrocopies and chimeric retrocopies.

The expression page provides information on estimated retrocopy expression. The table shows the accurate FPKM values of the retrocopies and their parental genes. The histogram shows the mean FPKM values for each tissue (Figure 2J). Moreover, the heatmap shows the expression of all retrocopies (Figure 2K). The heatmap clearly shows the tissues in which retrocopies are highly expressed or not expressed, so that user can explore the function of retrocopies and whether their expression shows an organizational preference.

Users can filter the data based on any table column on each page and can directly search for keywords in the search box above the tables. All image colors and text sizes can be adjusted according to users' needs. All information tables and figures can be downloaded by clicking the download tabs.

RESULTS

Test

RetroScan is suitable for species with available scaffold-level or chromosome-level genome assemblies and detailed annotation information. If users upload the relevant RNA-Seq data, they can further explore the expression values of retrocopies. A well-developed retrocopy annotation tool requires tests to examine its accuracy and improve its applicability. Here, we selected six eukaryotic species for verification, including two vertebrates [*Homo sapiens* (Falconer et al., 2012), *Danio rerio* (Howe et al., 2013)], two plants [*Arabidopsis thaliana* (Theologis et al., 2000), *Oryza sativa* (Sasaki and International Rice Genome Sequencing Project, 2005)] and two insects [*Drosophila melanogaster* (Adams et al., 2000), *Aedes aegypti* (Nene et al., 2007)]. The data were all downloaded from NCBI (Supplementary Table 1). In addition, we also tested species genomes from databases such as JGI (Phytozome), Ensembl and FlyBase (Supplementary Table 2). In our tests, RetroScan performed well and was suitable for genomic data of various databases. The running time and results of RetroScan are listed in Table 1. We ran RetroScan by entering the genome sequence files and corresponding annotation files. For evaluation, the programs were run on a dedicated Linux machine with Ubuntu18.04 running no other job, using the GNU time command to obtain real time. The machine had 16 GB of physical RAM and a six core Intel i7 CPU. We set all parameters to the default settings (thread = 1, identity ≥ 50%, coverage_rate ≥ 50%, coverage_len ≥ 50 aa, intron_loss_num ≥ 2, gap_len ≥ 40 bp, parent_loss_intron_len ≥ 60 bp, retro_one_exon_len ≤ 30, kaksmethod = NG). The size of the genomes ranged from 121 M to 3.3 G, and the number of retrocopy results reached 7048. The size of the genome, the number of annotated proteins and the proportion of repeated sequences have the greatest impact on the running time.

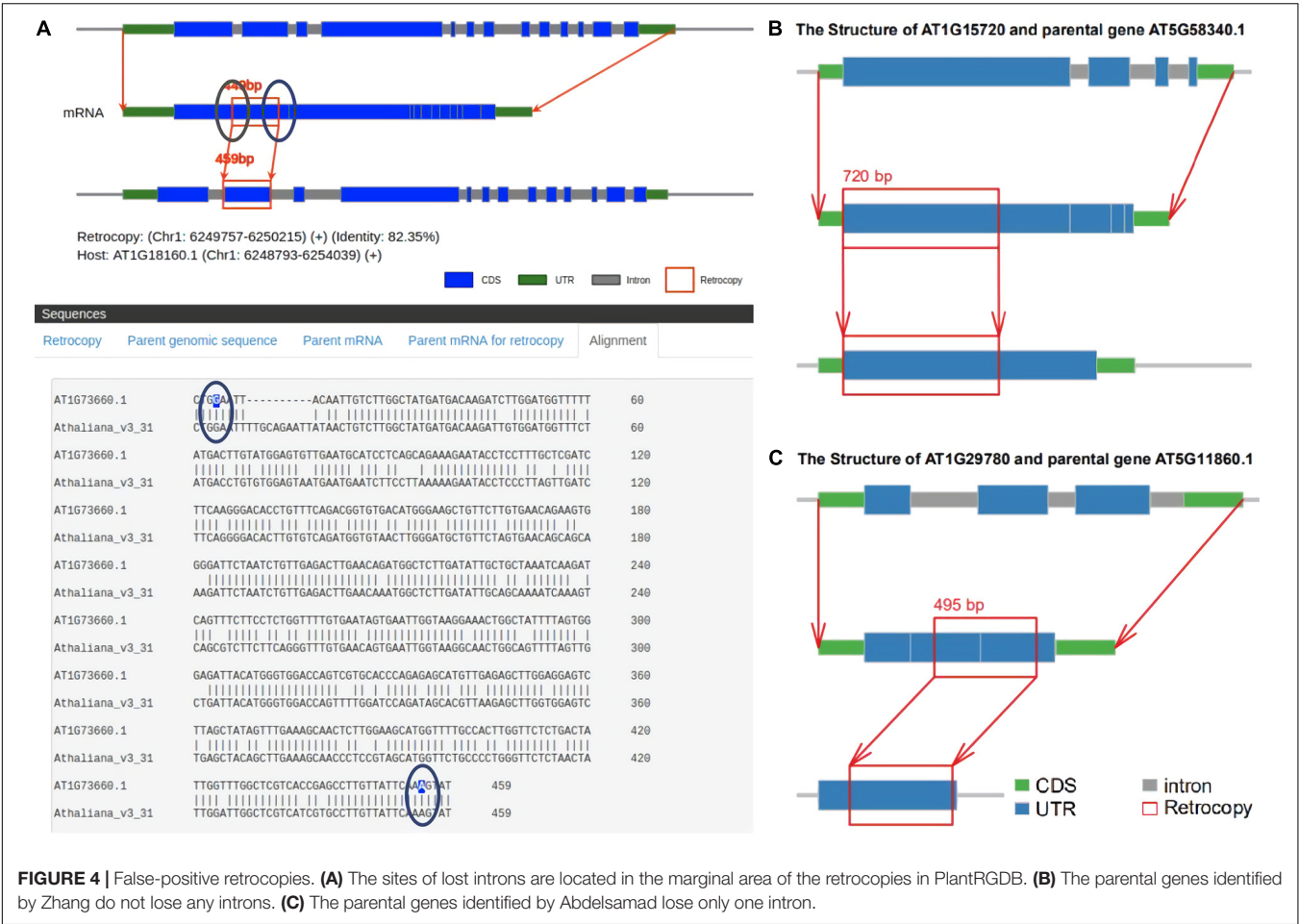
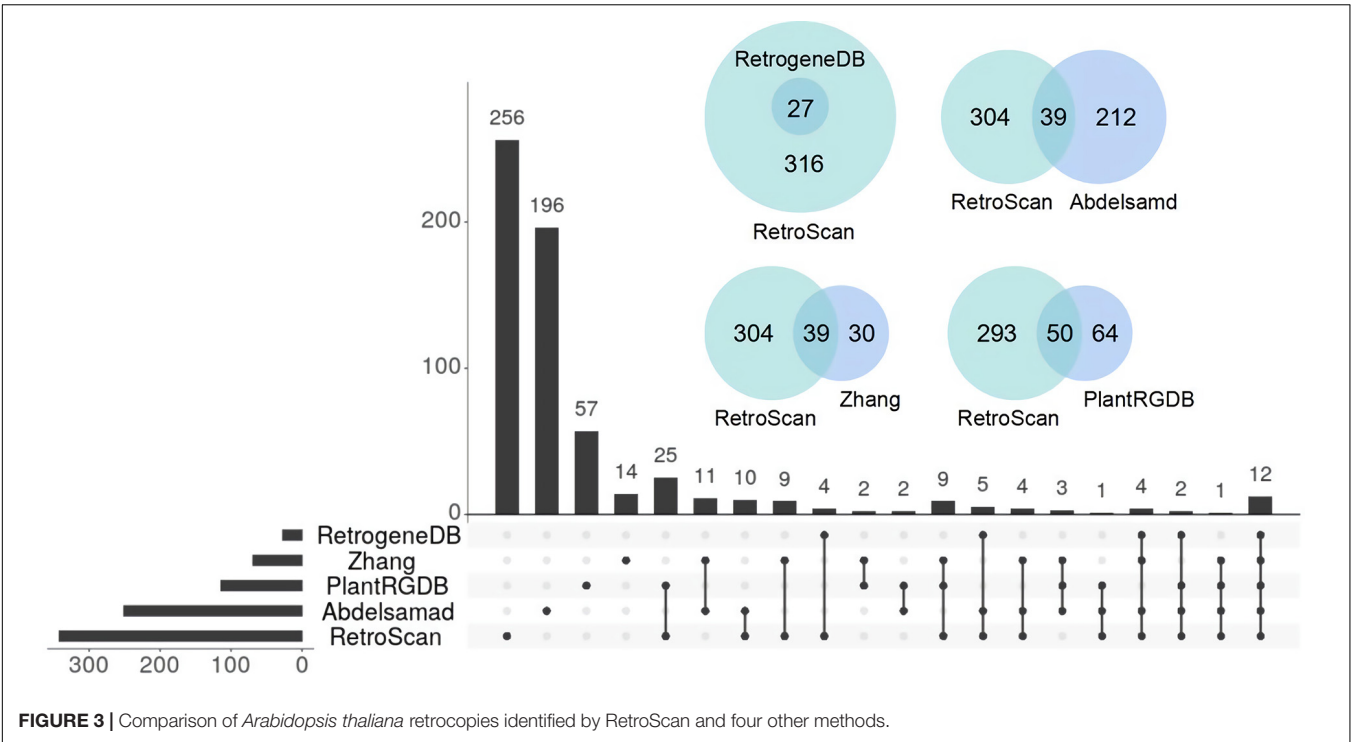
Comparison With Previous Studies

There is a lack of a uniform definition of retrocopy identity. The criteria for judging retrocopies are based mainly on the core definition that the sequences of retrocopies and their

TABLE 1 | RetroScan results for retrocopies in *Homo sapiens*, *Danio rerio*, *Arabidopsis thaliana*, *Oryza sativa*, *Drosophila melanogaster*, and *Aedes aegypti*.

Species	Genome size	Protein number	Time	Retrocopy number
<i>Homo sapiens</i>	3.3 G	1,302,060	768 min	7,048
<i>Danio rerio</i>	1.7 G	659,618	256 min	449
<i>Drosophila melanogaster</i>	145.7 M	336,015	11 min	221
<i>Aedes aegypti</i>	1.3 G	330,718	61 min	410
<i>Arabidopsis thaliana</i>	121.2 M	259,756	13 min	343
<i>Oryza sativa</i>	387.6 M	189,861	27 min	661

parental genes are highly similar but the parental genes lose multiple introns. Current retrocopy identification pipelines are based on the TBLASTN, BLAT, and paralog methods, and we selected representative studies in these pipelines to compare with RetroScan: RetrogeneDB (Rosikiewicz et al., 2017) for TBLASTN, PlantRGDB (Wang, 2017) for BLAT and the study of Abdelsamad and Pecinka (2014) and Zhang et al. (2005) for paralog (Supplementary Table 3). The results between these methods vary greatly, so we used *Arabidopsis thaliana* as an example to explain the reasons for these differences. RetroScan includes 343 retrocopies, RetrogeneDB includes 27, PlantRGDB includes 114 (duplicates have been removed), Zhang includes 69 and Abdelsamad includes 251. To compare other results with those of RetroScan, we considered any two retrocopies that overlapped at the same genomic position in which the overlap region was more than 50% of their sequence length to be the same retrocopy. An UpSet plot was generated to represent the intersections between five datasets (Figure 3). The total number of retrocopies in all studies was 627. Among the RetroScan retrocopies, 87 were shared with retrocopies from other pipelines, and 256 were novel (Figure 3). The 256 novel retrocopies consisted partly of retropseudogenes, which were mainly distributed in non-coding regions. Other novel retrocopies were newly discovered retrocopies that were ignored by the other four pipelines. We observed that all of the RetrogeneDB retrocopies overlapped with the RetroScan results because that study applied a similar pipeline to directly align protein-coding sequences with genome sequences using LAST. However, RetrogeneDB involved more stringent criteria (e.g., regarding alignment length, identity and coverage), and few retrocopies could be found in non-coding regions. RetroScan and PlantRGDB showed only 50 overlapping results, as PlantRGDB used the BLAT tool to identify retrocopies in plants. The BLAT method is not as accurate as BLASTN and will result in the loss of some positive results. The parental genes identified by the BLAT method do indeed lose multiple introns, but the sites of lost introns are located in the marginal area of the retrocopies, which are easily excluded in RetroScan (Figure 4A). Abdelsamad and Zhang developed a new method for identifying retrocopies. The method mainly compares intron-free genes and intron-genes with paralogs to find retrocopies. The paralog method can find more retrocopies in intron-free genes than the previous two methods but also produces more false-positive results. Therefore, only 39 overlapping results were observed with the results of this method. Moreover, it cannot find



retropseudogenes because it only uses annotated genes rather than genomic sequences. A portion of the retrocopies identified by the paralog method were found in the ortholog clusters shared with other species, such as rice. Another possibility is that parental genes with multiple exons do not lose any introns (**Figure 4B**) or lose only one intron (**Figure 4C**) in the region corresponding to the retrocopies. RetroScan can solve most of the above problems. Because two alignments are performed, mapping proteins to genome sequences and confirming lost introns, RetroScan guarantees that the results are accurate and reliable.

DISCUSSION

Retrocopies are fragments of genomic sequences which are highly similar to protein coding genes. They were considered as non-functional pseudogenes at some time in the past. Approaches established to identify pseudogenes include PseudoPipe (Zheng and Gerstein, 2006), HAVANA method (Searle et al., 2004), PseudoFinder (Chen et al., 2011), RetroFinder (Zheng et al., 2007), GIS-PET method (Ng et al., 2005), and consensus method (Zheng et al., 2007). These methods were developed by different teams, which mainly use alignment tools such as Blast, Blastz, and Blat to align DNA, protein, cDNA, and mRNA sequences and then accord to homology, intron-exon structure, existence of stop codons or frameshifts and so on to judge whether it is a pseudogene. However, not all retrocopies are pseudogenes, which are formed by retrotransposition and partly play some regulatory or other important roles in genome. Therefore, based on the identification of pseudogenes, researchers have developed new identification methods specifically for retrocopies by exhaustively aligning of genomic sequences against all possible parental genes. But different prediction methods often result in different numbers or sets of retrocopies because each researcher uses different criteria for identification.

Here, we draw up the criteria for judging retrocopies by RetroScan, which is a promising software developed to scan, annotate and display retrocopies. Regarding the coverage, similarity, the number of lost introns and other parameters between the parental genes and retrocopies, users can set according to the species situation. Compared to previous approaches, our new computational analysis tool shows increased accuracy and speed and is more convenient to use, especially when processing species with large-scale genomes. RetroScan is faster than the BLAT method and produces fewer false positives, similar to the paralog method. We used six species data to compare the results of RetroScan and three classic pipelines. Compared the sequence structure of retrocopies with parental genes, we found that RetroScan had the lowest false positives. At the same time, we ensure that the final results have nothing to do with DNA duplication by comparing the results back to the genome and deleting retrocopies with a large number of duplicates. It involves only one step and requires at least two input files (genome sequence file and annotation file). If RNA-Seq data are provided, it can further calculate the expression values of retrocopies. We used multiple sets of model species genomes

for testing, and the results proved that RetroScan is effective for the identification of retrocopies. In addition, our study is the first to provide a user-friendly visual interface that displays results, including information on retrocopies, Ka/Ks values, retrocopy structure and expression. Our approach shows great potential for retrocopy identification and will make an important contribution to evolutionary research, providing a powerful tool for promoting research on the duplication of genes and the origination of new genes and new functions.

Unlike RetroScan that identifies retrocopies of a single species, there are studies that focus on the genetic variations between groups. Schrider et al. (2013) describe a computational approach leveraging next-generation sequence data to detect gene copy-number variants caused by retrotransposition (retroCNVs), and find that these variants account for a substantial number of gene copy-number differences between individuals, and that gene retrotransposition may often result in both deleterious and beneficial mutations. Miller et al. (2021) exploit sideRETRO, a pipeline dedicated to detecting retroCNVs in whole-genome sequencing data and revealing their insertion sites, zygosity and genomic context and classifying them as somatic or polymorphic events. These tools focus on identifying the CNVs of retrocopy in the population, while RetroScan contributes greatly to research on retrocopies in individual organisms, which is of great significance for establishing a foundation for the future analysis of retroCNVs between subgroups.

In summary, RetroScan is a comprehensive, efficient and one-step retrocopy identification tool developed for users. We firmly believe that RetroScan will be useful for further comparative and evolutionary studies.

DATA AVAILABILITY STATEMENT

RetroScan is available at <https://github.com/Vicky123wzy/RetroScan> and can be installed directly by Conda. Users can also download the source code from GitHub, install related software and manually configure RetroScan. The data used in this study were retrieved from NCBI (<https://www.ncbi.nlm.nih.gov/>). Further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ZW developed the tool and drafted the manuscript. JS packaged, uploaded, and tested the tool. QL and TY participated in data testing. HZ revised the manuscript. YW designed and supervised the study and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundation of China (31871330), the Scientific Research Starting Foundation of Southwest University (SWU118103), and the Fundamental Research Funds for the Central Universities (XDJK2019TJ003).

ACKNOWLEDGMENTS

We thank Guoqing Zhang and Anqiang Jia for their help with the collection of the genome data. We also thank Hailong Guo and Fang Lu for their help with the use of the RetroScan applications.

REFERENCES

- Abdelsamad, A., and Pecinka, A. (2014). Pollen-specific activation of Arabidopsis retrogenes is associated with global transcriptional reprogramming. *Plant Cell* 26, 3299–3313. doi: 10.1105/tpc.114.126011
- Adams, M. D., Celniker, S. E., Holt, R. A., and Evans, C. A. (2000). The genome sequence of *Drosophila melanogaster*. *Science* 287:2185. doi: 10.1126/science.287.5461.2185
- Bai, Y., Casola, C., Feschotte, C., and Betran, E. (2007). Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. *Genome Biol.* 8:R11. doi: 10.1186/gb-2007-8-1-r11
- Betran, E., Thornton, K., and Long, M. (2002). Retroposed new genes out of the X in *Drosophila*. *Genome Res.* 12, 1854–1859. doi: 10.1101/gr.6049
- Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504
- Brosius, J. (1991). Retroposons—seeds of evolution. *Science* 251:753. doi: 10.1126/science.1990437
- Carelli, F. N., Hayakawa, T., Go, Y., Imai, H., Warnefors, M., and Kaessmann, H. (2016). The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* 26, 301–314. doi: 10.1101/gr.198473.115
- Casola, C., and Betrán, E. (2017). The genomic impact of gene retrocopies: what have we learned from comparative genomics, population genomics, and transcriptomic analyses? *Genome Biol. Evol.* 9, 1351–1373. doi: 10.1093/gbe/evx081
- Chen, S.-M., Ma, K.-Y., and Zeng, J. (2011). Pseudogene: lessons from PCR bias, identification and resurrection. *Mol. Biol. Rep.* 38, 3709–3715. doi: 10.1007/s11033-010-0485-4
- Conway, J. R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 2938–2940. doi: 10.1093/bioinformatics/btx364
- Falconer, E., Hills, M., Naumann, U., Poon, S. S. S., Chavez, E. A., Sanders, A. D., et al. (2012). DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat. Methods* 9, 1107–1112. doi: 10.1038/nmeth.2206
- Flagel, L. E., and Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytol.* 183, 557–564. doi: 10.1111/j.1469-8137.2009.02923.x
- Fu, B., Chen, M., Zou, M., Long, M., and He, S. (2010). The rapid generation of chimerical genes expanding protein diversity in zebrafish. *BMC Genomics* 11:657. doi: 10.1186/1471-2164-11-657
- Howe, K., Clark, M. D., Torroja, C. F., and Torrance, J. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496, 498–503. doi: 10.1038/nature12111
- Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11, 97–108. doi: 10.1038/nrg2689
- Jakalski, M., Takeshita, K., Deblieck, M., Koyanagi, K. O., Makalowska, I., Watanabe, H., et al. (2016). Comparative genomic analysis of retrogene repertoire in two green algae *Volvox carteri* and *Chlamydomonas reinhardtii*. *Biol. Direct* 11, 35–35. doi: 10.1186/s13062-016-0138-1
- Kabza, M., Ciombarowska, J., and Makalowska, I. (2014). RetrogeneDB—a database of animal retrogenes. *Mol. Biol. Evol.* 31, 1646–1648. doi: 10.1093/molbev/msu139
- Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. *Nat. Rev. Genet.* 10, 19–31. doi: 10.1038/nrg2487
- Kielbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493. doi: 10.1101/gr.113985.110
- Kim, D., Langmead, B., and Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi: 10.1038/nmeth.3317
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948. doi: 10.1093/bioinformatics/btm404
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352
- Long, M., Betran, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* 4, 865–875. doi: 10.1038/nrg1204
- Long, M., and Langley, C. H. (1993). Natural selection and the origin of jingwei, a chimeric processed functional gene in *Drosophila*. *Science* 260:91. doi: 10.1126/science.7682012
- Lynch, M., and Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290:1151. doi: 10.1126/science.290.5494.1151
- Matsuno, M., Compagnon, V., Schoch, G. A., Schmitt, M., Debayle, D., Bassard, J.-E., et al. (2009). Evolution of a novel phenolic pathway for pollen development. *Science* 325:1688. doi: 10.1126/science.1174095
- Miller, T. L. A., Orpinelli Rego, F., Buzzo, J. L. L., and Galante, P. A. F. (2021). sideRETRO: a pipeline for identifying somatic and polymorphic insertions of processed pseudogenes or retrocopies. *Bioinformatics* 37, 419–421. doi: 10.1093/bioinformatics/btaa689
- Navarro, F. C., and Galante, P. A. (2013). RCPedia: a database of retrocopied genes. *Bioinformatics* 29, 1235–1237. doi: 10.1093/bioinformatics/btt104
- Navarro, F. C. P., and Galante, P. A. F. (2015). A genome-wide landscape of retrocopies in primate genomes. *Genome Biol. Evol.* 7, 2265–2275. doi: 10.1093/gbe/evv142
- Nene, V., Wortman, J. R., Lawson, D., and Haas, B. (2007). Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science (New York, N.Y.)* 316, 1718–1723. doi: 10.1126/science.1138878
- Ng, P., Wei, C.-L., Sung, W.-K., Chiu, K. P., Lipovich, L., Ang, C. C., et al. (2005). Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* 2, 105–111. doi: 10.1038/nmeth733
- Ohshima, K., Hattori, M., Yada, T., Gojobori, T., Sakaki, Y., and Okada, N. (2003). Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol.* 4, R74–R74. doi: 10.1186/gb-2003-4-11-r74
- Pan, D., and Zhang, L. (2009). Burst of young retrogenes and independent retrogene formation in mammals. *PLoS One* 4:e5040. doi: 10.1371/journal.pone.0005040
- Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 85, 2444–2448. doi: 10.1073/pnas.85.8.2444
- Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi: 10.1038/nbt.3122
- Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842. doi: 10.1093/bioinformatics/btq033
- Rosikiewicz, W., Kabza, M., Kosinski, J. G., Ciombarowska-Basheer, J., Kubiak, M. R., and Makalowska, I. (2017). RetrogeneDB—a database of plant and animal retrocopies. *Database (Oxford)* 2017:bax038. doi: 10.1093/database/bax038
- Sakai, H., Mizuno, H., Kawahara, Y., Wakimoto, H., Ikawa, H., Kawahigashi, H., et al. (2011). Retrogenes in rice (*Oryza sativa* L. ssp. *japonica*) exhibit correlated expression with their source genes. *Genome Biol. Evol.* 3, 1357–1368. doi: 10.1093/gbe/evr111

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2021.719204/full#supplementary-material>

- Sasaki, T., and International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. *Nature* 436, 793–800. doi: 10.1038/nature03895
- Schrider, D. R., Navarro, F. C., Galante, P. A., Parmigiani, R. B., Camargo, A. A., Hahn, M. W., et al. (2013). Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet.* 9:e1003242. doi: 10.1371/journal.pgen.1003242
- Schrider, D. R., Stevens, K., Cardeno, C. M., Langley, C. H., and Hahn, M. W. (2011). Genome-wide analysis of retrogene polymorphisms in *Drosophila melanogaster*. *Genome Res.* 21, 2087–2095. doi: 10.1101/gr.116434.110
- Searle, S. M., Gilbert, J., Iyer, V., and Clamp, M. (2004). The otter annotation system. *Genome Res.* 14, 963–970. doi: 10.1101/gr.1864804
- Theologis, A., Ecker, J. R., Palm, C. J., and Federspiel, N. A. (2000). Sequence and analysis of chromosome 1 of the plant *Arabidopsis thaliana*. *Nature* 408, 816–820. doi: 10.1038/35048500
- Vinckenbosch, N., Dupanloup, I., and Kaessmann, H. (2006). Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3220–3225. doi: 10.1073/pnas.0511307103
- Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinform.* 8, 77–80. doi: 10.1016/s1672-0229(10)60008-3
- Wang, W., Brunet, F. G., Nevo, E., and Long, M. (2002). Origin of sphinx, a young chimeric RNA gene in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U.S.A.* 99, 4448–4453. doi: 10.1073/pnas.072066399
- Wang, Y. (2017). PlantRGDB: a database of plant retrocopied genes. *Plant Cell Physiol.* 58:e2. doi: 10.1093/pcp/pcw210
- Zhang, J., Yang, H., Long, M., Li, L., and Dean, A. M. (2010). Evolution of enzymatic activities of testis-specific short-chain dehydrogenase/reductase in *Drosophila*. *J. Mol. Evol.* 71, 241–249. doi: 10.1007/s00239-010-9384-5
- Zhang, Y., Wu, Y., Liu, Y., and Han, B. (2005). Computational identification of 69 retroposons in *Arabidopsis*. *Plant Physiol.* 138, 935–948. doi: 10.1104/pp.105.060244
- Zhang, Z., Harrison, P. M., Liu, Y., and Gerstein, M. (2003). Millions of years of evolution preserved: a comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* 13, 2541–2558. doi: 10.1101/gr.1429003
- Zheng, D., Frankish, A., Baertsch, R., Kapranov, P., Reymond, A., Choo, S. W., et al. (2007). Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res.* 17, 839–851. doi: 10.1101/gr.5586307
- Zheng, D., and Gerstein, M. B. (2006). A computational approach for identifying pseudogenes in the ENCODE regions. *Genome Biol.* 7(Suppl. 1), S13.11–S13.10. doi: 10.1186/gb-2006-7-s1-s13
- Zhu, Z., Zhang, Y., and Long, M. (2009). Extensive structural renovation of retrogenes in the evolution of the *Populus* genome. *Plant Physiol.* 151, 1943–1951. doi: 10.1104/pp.109.142984
- Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
- Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.
- Copyright © 2021 Wei, Sun, Li, Yao, Zeng and Wang. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers



OPEN ACCESS

Articles are free to read
for greatest visibility
and readership



FAST PUBLICATION

Around 90 days
from submission
to decision



HIGH QUALITY PEER-REVIEW

Rigorous, collaborative,
and constructive
peer-review



TRANSPARENT PEER-REVIEW

Editors and reviewers
acknowledged by name
on published articles

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org

Contact us: frontiersin.org/about/contact



REPRODUCIBILITY OF RESEARCH

Support open data
and methods to enhance
research reproducibility



DIGITAL PUBLISHING

Articles designed
for optimal readership
across devices



FOLLOW US

@frontiersin



IMPACT METRICS

Advanced article metrics
track visibility across
digital media



EXTENSIVE PROMOTION

Marketing
and promotion
of impactful research



LOOP RESEARCH NETWORK

Our network
increases your
article's readership