DEEP LEARNING FOR BIOLOGICAL NETWORK ANALYSIS

EDITED BY: Jianye Hao, Zhongyu Wei, Jiajie Peng and Yulan He
PUBLISHED IN: Frontiers in Genetics and
Frontiers in Bioengineering and Biotechnology

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88974-294-3
DOI 10.3389/978-2-88974-294-3

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

DEEP LEARNING FOR BIOLOGICAL NETWORK ANALYSIS

Topic Editors:
Jianye Hao, Tianjin University, China
Zhongyu Wei, Fudan University, China
Jiajie Peng, Northwestern Polytechnical University, China
Yulan He, University of Warwick, United Kingdom

Citation: Hao, J., Wei, Z., Peng, J., He, Y., eds. (2022). Deep Learning for Biological Network Analysis. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88974-294-3

Table of Contents

frontiers in Genetics

ORIGINAL RESEARCH
published: 28 November 2019
doi: 10.3389/fgene.2019.01191

Peptide-Major Histocompatibility Complex Class I Binding Prediction Based on Deep Learning With Novel Feature

Tianyi Zhao[1], *Liang Cheng*[2], *Tianyi Zang*[1]* *and Yang Hu*[1]*

[1] *Department of Computer Science and Technology, School of Life Science and Technology, Harbin Institute of Technology, Harbin, China,* [2] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*

Peptide-based vaccine development needs accurate prediction of the binding affinity between major histocompatibility complex I (MHC I) proteins and their peptide ligands. Nowadays more and more machine learning methods have been developed to predict binding affinity and some of them have become the popular tools. However most of them are designed by the shallow neural networks. Bengio said that deep neural networks can learn better fits with less data than shallow neural networks. In our case, some of the alleles only have dozens of peptide data. In addition, we transform each peptide into a characteristic matrix and input it into the model. As we know when dealing with the problem that the input is a matrix, convolutional neural network (CNN) can find the most critical features by itself. Obviously, compared with the traditional neural network model, CNN is more suitable for predicting binding affinity. Different from the previous studies which are based on blocks substitution matrix (BLOSUM), we used novel feature to do the prediction. Since we consider that the order of the sequence, hydropathy index, polarity and the length of the peptide could affect the binding affinity and the properties of these amino acids are key factors for their binding to MHC, we extracted these information from each peptide. In order to make full use of the data we have obtained, we have integrated different lengths of peptides into 15mer based on the binding mode of peptide to MHC I. In order to demonstrate that our method is reliable to predict peptide-MHC binding, we compared our method with several popular methods. The experiments show the superiority of our method.

Keywords: peptide-major histocompatibility complex class I binding prediction, deep learning, convolutional neural network, epitope prediction, human leukocyte antigen

OPEN ACCESS

Edited by:
Jianye Hao,
Tianjin University, China

Reviewed by:
Deepika Mathur,
Columbia University,
United States
Yuansong Zhao,
Boston University, United States
Hao Lin,
University of Electronic Science and
Technology of China, China

****Correspondence:***
Tianyi Zang
tianyi.zang@hit.edu.cn
Yang Hu
huyang@hit.edu.cn

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

Received: *15 August 2019*
Accepted: *28 October 2019*
Published: *28 November 2019*

Citation:
Zhao T, Cheng L, Zang T and
Hu Y (2019) Peptide-Major
Histocompatibility Complex Class I
Binding Prediction Based on Deep
Learning With Novel Feature.
Front. Genet. 10:1191.
doi: 10.3389/fgene.2019.01191

INTRODUCTION

Many scholars try to find personalized treatment for melanoma and other cancers through major histocompatibility complex (MHC) (Kreiter et al., 2015; Bentzen et al., 2016; Johnson et al., 2016). Two successful phase I clinical trials proved that cancer vaccines are not a dream. These studies showed that 66.7 and 61.5% of resected melanoma patients have been cured during the period of 20–32 months and 12–23 months separately following vaccination (Ott et al., 2017; Sahin et al., 2017). These works were published in Nature, which have attracted more attention to personalized neoantigen vaccines (Chu et al., 2018).

Since neoantigens are ideal targets for immunotherapy, understanding the binding affinity between specific peptides and MHC alleles is an essential step in designing vaccines (Rolland et al., 2011; Cheng et al., 2017). The large number of peptide chains makes the research time-consuming and laborious. With the improvement of sequencing technology and bioinformatics, the binding affinity between predicted peptides and MHC alleles has become more flexible and economical (Jensen et al., 2018).

MHC is a gene family found in most vertebrate genomes and is closely related to the immune system. The MHC of humans is also known as human leukocyte antigen (HLA). There are two types of MHC; the first type of MHC processes internal decomposition of the protein (such as the virus), the second type of MHC is only located on antigen-presenting cells (APC), such as macrophages. For example, if there is bacterial invasion in the tissue, and the macrophage is swallowed, the bacterial fragments are prompted by MHC to the helper T cells to initiate an immune response. The regulated DNA is located on chromosome 6 (6p21.31) (Cheng et al., 2018; Cheng et al., 2019) and includes a series of tightly linked loci that are closely related to human immune system function (Neefjes et al., 2011). Some of these genes encode cell surface antigens, which are the "characteristics" that are not confusing for each person's cells. They are the basis for the immune system to distinguish itself from foreign bodies. The HLA complex is located in the 21.31 region (6p21.31) on the short arm of chromosome 6, and is composed of 3.6 million base pairs. It is the region with the highest gene density and the most polymorphic region in human chromosomes. Known as "chemical fingerprints in humans".

Recently, many researchers have focused on the field of predicting the binding affinity between peptide and MHC alleles. Some of them focused on the MHC-I and some of them focused on the MHC-II. There are also lots of tools and algorithms which are developed for this work. We classified these methods into three categories: Machine learning, neural network and deep learning.

Machine learning methods extracted features and constructed models to predict peptide-MHC interactions. Giguere S. et al. (2013) used kernel ridge regression to predict peptide-protein binding affinity. Uslan V and Seker H. (Uslan and Seker, 2016) used support vector regression (SVR) based on fuzzy model to do this work. Pavel P. Kuksa et al. (Kuksa et al., 2015) proposed a high-order semi-RBM to pretrain feed-forward high-order neural network (HONN). After that, high-order nuclear SVM was used to predict peptide-MHC binding. Although these methods can capture nonlinear interactions between different peptides, they fail to model the direct strong high-order interactions between features.

Recently, neural network (Hao et al., 2016; Hao et al., 2017) and deep learning (Peng et al., 2019a; Peng et al., 2019b) are the most common used methods in this field. Kasper W. Jorgensen (Jørgensen et al., 2014) developed a novel tool-NetMHCstab to predict stability of peptide-MHC complexes. They used Artificial neural network (ANN) to identify the stability of 10 different HLA class I molecules. Recently more studies tried to integrate peptides of different lengths into a machine-learning frame. These methods such as MHCflurry (O'Donnell et al., 2018) and NetMHCpan (Trolle et al., 2015) can involve more training data into their model and become popular tools for this task (Jurtz et al., 2017). NetMHC trained models for each MHC allele and this model is based on allele-specific approach (Andreatta and Nielsen, 2015). Whereas NetMHCIIpan (Jensen et al., 2018) is based on the pan-allele approach. Actually, they both used basic ANN with the immune epitope database (IEDB) (Vita et al., 2018; Salimi et al., 2019). NNAlign (Alvarez et al., 2018) which is a method based on neural network has been a common method to build models. Barra et al. (2018), Garde et al. (2019) all developed their own methods based on NNAlign. With the development of Mass Spectrometry (MS), the precision of identifying MHC ligands has been improved. Some researchers have proved that using MS data to do the training the model could be more robust. In the most recently released NetMHCpan 4.0 (Jurtz et al., 2017), they added MS data into their training set and improved their prediction accuracy.

Deep learning methods have shown their powerful ability of prediction and classification in recent years and have attracted more and more scholars' attention (Peng et al., 2019c). Zeng and Gifford (2019) purposed a deep residual network-based computational approach that quantifies uncertainty in peptide-MHC affinity prediction. Sidhom et al. (2018) present Allele-Integrated MHC (AI-MHC), a deep learning architecture for human Class I and Class II MHC binding prediction. More researchers' work (Bulik-Sullivan et al., 2019; Phloyphisut et al., 2019; Tran et al., 2019) have proved that deep learning methods have better performance than shallow neural networks.

The other important step to predict peptide-MHC binding affinity is extraction of feature. In the previous studies, most of the studies focused on the 9-mer peptides because most presented MHC class-I ligands are 9 mer (Bassani-Sternberg et al., 2015). However, for some alleles, they prefer other lengths of peptides. For example, Mamu-A2*05 preferentially binds 8-mer peptides (de Groot et al., 2017) and HLA-B*44:03 (Rist et al., 2013) prefers 10 and 11 mer peptides. Recently more and more researchers found methods to make all peptides into the same length so they can train their models with more data. Massimo Andreatta and Nielsen et al. (2015) added or deleted the primary sequence to ensure all the peptides are 9 mer. As a result, they involved the length of the deletion/insertion and the length and the composition of the peptide flanking regions in the feature. Youngmahn Han and Dongsup Kim (Han and Kim, 2017) considered each peptide as an image and each data in the feature is a pixel.

Although most previous studies have achieved high accuracy of prediction, there should be a novel method to use chemical properties of peptides to predict the binding affinity. In this paper, we used sequence comparison based on BLOSUM62 coding and to chemical properties of peptides extract feature and used convolutional neural network (CNN) to build models.

METHODS

Feature Extraction

For the MHC-I complex, the alpha chain has three domains, wherein the grooves formed by the $\alpha 1$ and $\alpha 2$ regions can bind to an antigen peptide and the α3 region is a CD8 binding region. The β chain has only one domain of β2, forming a microglobulin structure. As shown in **Figure 1**, the binding core of nine amino acids plays a major role in the binding of the MHC-I molecule to the affinity peptide. At the same time, the peptide flanking residues (PFR) on both sides also plays a certain role in the binding. In the binding core, positions one, four, six, seven, nine are called "anchors" and play a more important role in binding than other locations. Based on this theory, we proposed a novel method that can convert the 8–14mer peptide to 15mer. Since one, four, six, nine are much more important than the other locations, we try to ensure that the two sequences of one to four and six to nine are not inserted into the new 'amino acid' (X). As we can see in **Figure 2**, we take 9–12mer peptide as an example. X is an artificial amino acid which is only related to itself and not related to the other 20 amino acids.

After converting all peptides to 15mer, all the peptides should be encoded by BLOSUM62 matrix (Styczynski et al., 2008). X is encoded as a vector of zeros but the score between X and itself is one. Then the feature of each peptide is a matrix 15*21.

The chemical properties of peptides have been reported to strongly affect the binding affinity. When the body is infected, inflammatory factors such as IFN-γ can change the β subunit composition of the proteasome 20S, making the proteasome more likely to cleave hydrophobic and alkalinous amino acids (so that the peptide is more easily bound to MHC-I). As said by Udaka et al. (1995) there is a general preference for hydrophobic amino acids. They also divided MHC-I into eight positions and found that the dominance of amino acids with hydrophobic side chains is unequivocal for some positions. Conversely, neutral or positively charged hydrophilic side chains are preferred in some other positions. In addition, Some positions allow hydrophobic as well as hydrophilic amino acids and appear to be less constrained than other positions.

Therefore, we proposed a novel way to extract the feature of peptides. We extracted four kinds of features: Sequence, Hydropathy index, Polarity, Length.

For the first feature: Sequence, we sorted the 21 kinds of amino acids by the BLOSUM62. 'A'', 'R', 'N', 'D', 'C', 'Q', 'Ev, 'G', 'H', 'I', 'L', 'K', 'M', 'F', 'P', 'S', 'T', 'W', 'Y', 'V', 'X' are represented by the numbers 1 to 21 respectively.

For the second feature: Hydropathy index, we used Eisenberg consensus scale (ECS) (Eisenberg, 1984) to value each amino acid's hydropathy index. X's hydropathy index is zero. **Table 1** shows the score of every amino acid.

For the third feature: Polarity, we divided 21 amino acids into five classes. According to the polarity of R group or the trend of interaction with water at physiological pH (approaching pH 7.0), they can be divided into non-polarity, polarity without charge, positive charge (alkalinity) and negative charge (acidity) (Wolfenden et al., 2015). X's class is zero. **Table 2** shows the classification of every amino acid.

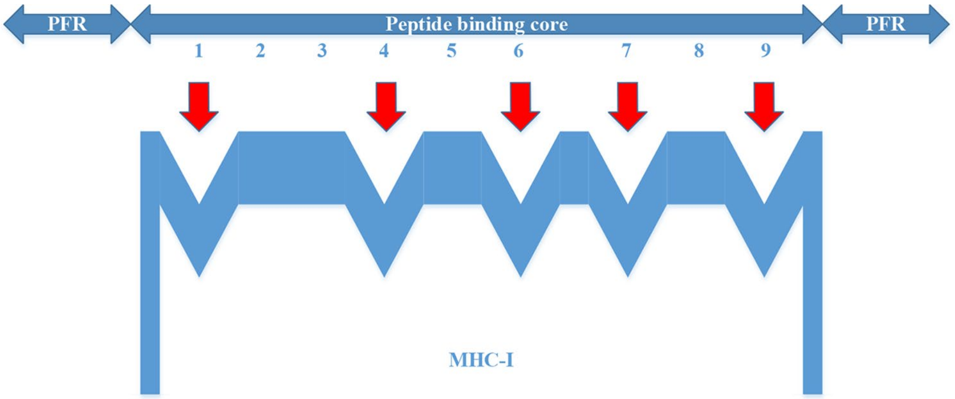


FIGURE 1 | Binding of major histocompatibility complex -I molecules to affinity peptides.

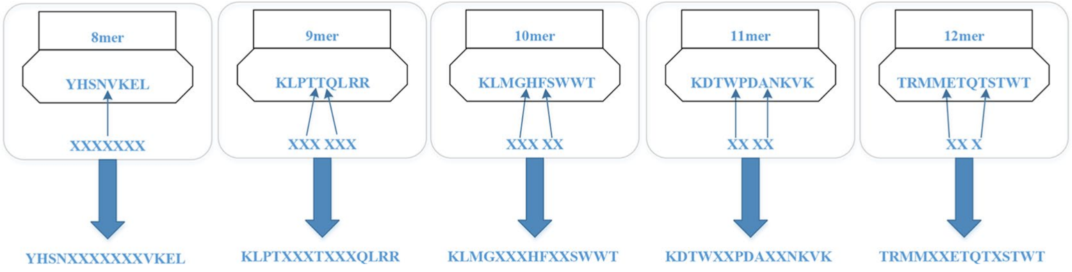


FIGURE 2 | Encoding peptides of different lengths.

TABLE 1 | Hydropathy Index of 21 amino acids.

Amino Acids	Hydropathy Index	Amino acids	Hydropathy Index
R	−2.5	K	−1.5
D	−0.9	Q	−0.85
N	−0.78	E	−0.74
H	0.40	S	−0.18
T	−0.05	P	0.12
Y	0.26	C	0.29
G	0.48	A	0.62
M	0.64	W	0.81
L	1.1	V	1.1
F	1.2	I	1.4
X	0		

TABLE 2 | Five Classes of amino acids based on polarity.

Class	Label	Amino acids
NONE	0	X
Polarity without charge	1	A, G, I, L, F, P, V
Non-polarity	2	N, C, Q, S, T, W, Y, M
Negative charge (acidity)	3	D, E
Positive charge (alkalinity)	4	R, H, K

For the fourth feature: Length, we use the length of peptide as a feature.

The detailed flow is as in the following **Figure 3**.

As shown in **Figure 3**, each peptide would be encoded as a 4*15 matrix. N is the number of training set.

Building Model by Convolutional Neural Network

Each peptide could be put into the CNN as a “picture” whose size is N*15. So we should set the structure of CNN firstly.

Figure 4 shows the structure of CNN. It contains two convolution layers. Each convolution layers have 20 filters. We used rectified linear unit (‘ReLu‘) as the activation function in the activation layer. ‘Max’ method is used in the Pool layer.

We built four models for different lengths of the peptides. We grouped the peptides by their length (L). The four groups are L < = 8, L = 9, L = 10 and L = > 11.

RESULTS

Data Description

We downloaded three different datasets. The detailed information is shown in **Table 3**.

We totally obtained 525,672 peptides and the data include their allele, peptide, measurement value, measurement inequality, measurement type, measurement source, and original allele.

We only selected those alleles whose number of peptides are larger than 20. Then 522,268 peptides are left. These peptides belongs to 193 kinds of alleles. As shown in **Figure 5**, one allele has more than 60,000 peptide data and some alleles’ data are much smaller.

Among these 522,268 peptides, there are 338,978 positive peptides. As we know, different alleles have different preferences for length of peptides. As shown in **Figure 6**, we found that most of the alleles prefer the length nine.

Therefore, it is much reasonable to put length of peptide into the feature matrix.

Evaluation of the Convolutional Neural Network & Based on New Feature

We used both binding affinity (BA) data and eluted ligand (EL) data. After integrating the two data sets together, in order to prevent the uneven distribution of the negative and positive peptides, we sorted the data in disorder. Then, we did fivecross validation.

HLA type alleles are the data we care about most. There are 43 HLA-A alleles and 82 HLA-B alleles in our dataset. In the Youngmahn Han and Dongsup Kim’s paper (Han and Kim, 2017), they used Deep CNN to compare with NetMHCPan, SMM(47), ANN, and PickPocket (Zhang et al., 2009). We used their statistical data and evaluated our CNN which is based on the novel feature. We call our method CNN-NF.

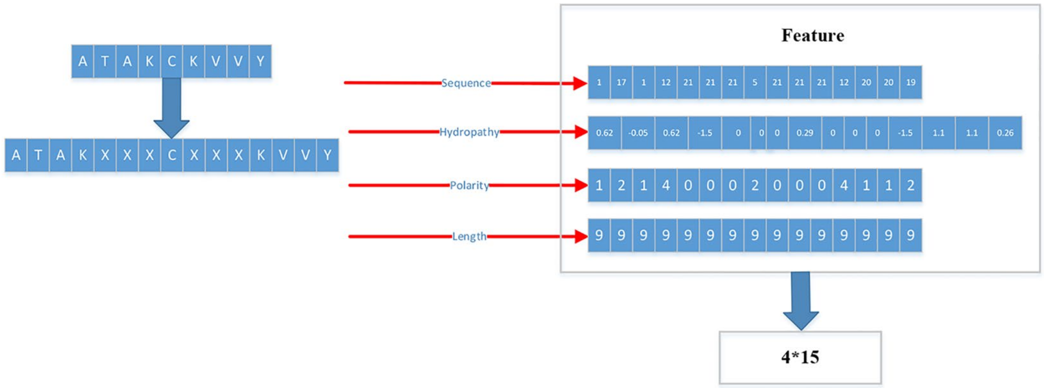


FIGURE 3 | Detailed flow of generating training set and testing set.

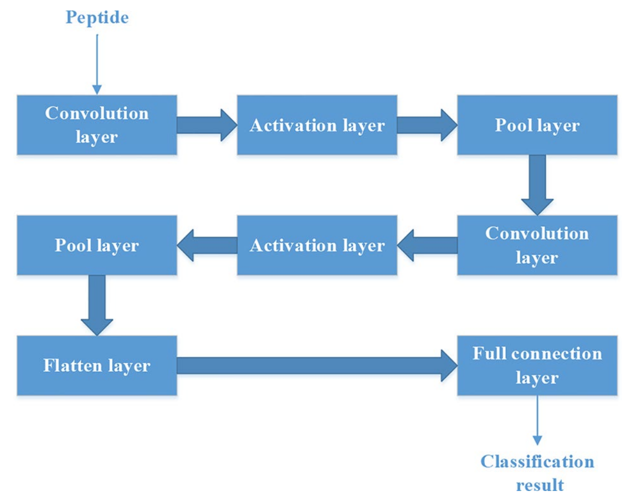


FIGURE 4 | The structure of convolutional neural network.

TABLE 3 | Detailed information of data.

Name	Source
IEDB affinity data	Vita et al. (2018)
BD2013	Kim et al. (2014)
MS data	Abelin et al. (2017)

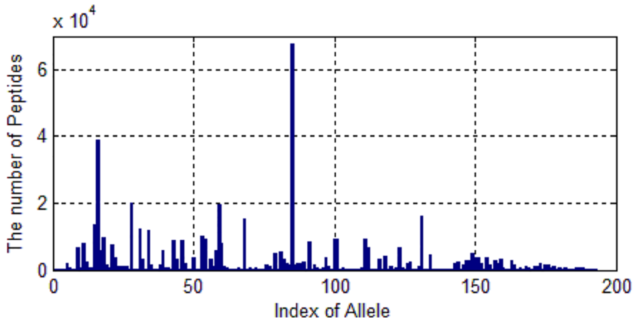


FIGURE 5 | The distribution of the number of peptides of 193 alleles.

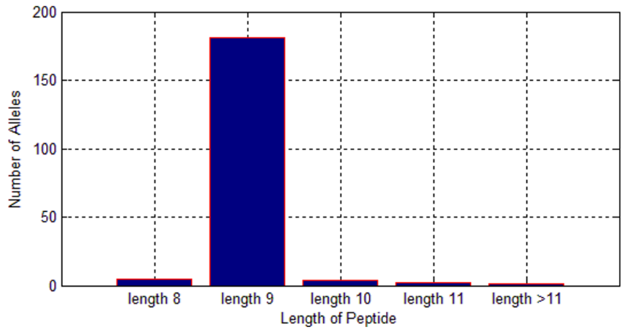


FIGURE 6 | Length preference of 193 alleles.

F1 score is used to evaluate models. It can be calculated as:

$$F1 = \frac{2TP}{2TP + FN + FP} \tag{1}$$

Here, true positive (TP) denotes positive samples whose predictions are positive. false negative (FN) denotes positive samples whose predictions are negative. false positive (FP) denotes negative samples whose predictions are positive.

As we can see in **Table 4**, **Tables 4A, B** summarize the prediction results for HLA-A and HLA-B alleles, respectively. The mean values of the F1 Score of the CNN-NF were 0.643 and 0.692. The values are slightly higher than those of other methods. In addition to that, the standard deviation of the two experiments are lower than those of other methods' either. It means that CNN-NF is more stable.

Since we totally obtain 193 alleles, we calculated 193 F1 scores. As shown in **Figure 7**, there are 19% alleles whose F1 score are more than 0.9. In addition, there are 34% alleles whose F1 score are lower than 0.5. We can know that different alleles have different accuracy and even polarization.

We also are interested in the area under curve (AUC) of the 193 allele experiments. We draw **Figure 8** for each allele's performance of AUC and another figure for the distribution of AUC in 193 experiments.

As we can see in **Figure 9**, although there are some alleles whose accuracy are lower than 0.5, most of the alleles have an accuracy more than 0.7. The low accuracy of some alleles may be due to the small amount of data. It may also be caused by the extreme imbalance of data.

Peptide-Length Preference Of Major Histocompatibility Complex Molecules

Although we have known that most of the alleles mostly prefer the nine length peptide, different alleles have different preferences in 8,9,10,11,12,13,14,15mer peptides. We should verify the ability of our method to capture peptide long preferences for different MHC molecules. Therefore, we randomly generated 10,000 peptides for each MHC molecules. These 10,000 peptides' length range from 8 to 15. The number of peptides of each length is the same so each length has 1,250 peptides. Then we put these artificial peptides into the models and the models would tell us the probability of being positive. We selected the top 2% probabilities and calculated the distribution of different lengths.

As shown in **Figures 10–12**, we randomly selected an allele for each HLA-A, B, and C coding site to verify the ability of our method to capture peptide long preferences for different MHC molecules.

CNN-NF prefer to identify the 9mer peptide as the binding peptide. Besides, if the number of the specific length peptide is small, CNN-NF can hardly give a high score. We can consider this phenomenon as a way that CNN guarantee the training accuracy.

CONCLUSIONS

In this paper, we purposed a novel method for peptide-MHC-I binding prediction. Since deep learning is developing fast, we consider that it has more advantages than shallow neural

TABLE 4 | Prediction results for human leukocyte antigen-1 (HLA-I) alleles(A).

(A) Summary of prediction results for HLA-A alleles (F1 Score)

	CNN-NF	DCNN	NetMHCPan	SMM	ANN	PickPocket
Mean	0.643	0.638	0.608	0.601	0.579	0.561
Median	0.603	0.696	0.667	0.667	0.667	0.625
Standard Deviation	0.166	0.23	0.267	0.250	0.286	0.318

(B) Summary of prediction results for HLA-B alleles (F1 Score)

	CNN-NF	DCNN	NetMHCPan	SMM	ANN	PickPocket
Mean	0.692	0.593	0.606	0.578	0.606	0.560
Median	0.621	0.667	0.625	0.615	0.643	0.593
Standard Deviation	0.228	0.286	0.286	0.302	0.290	0.277

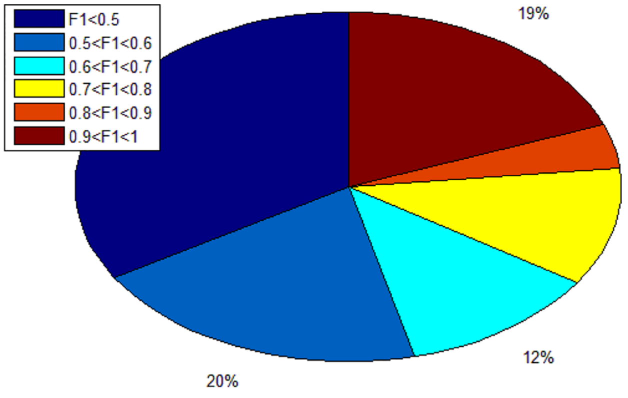


FIGURE 7 | The distribution ratio of F1 score.

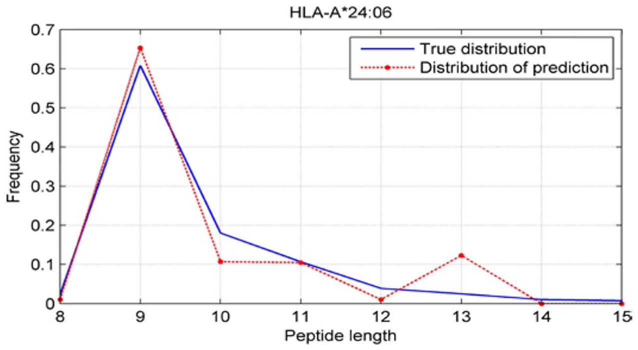


FIGURE 10 | Predicted length preference of HLA-A*24:06.

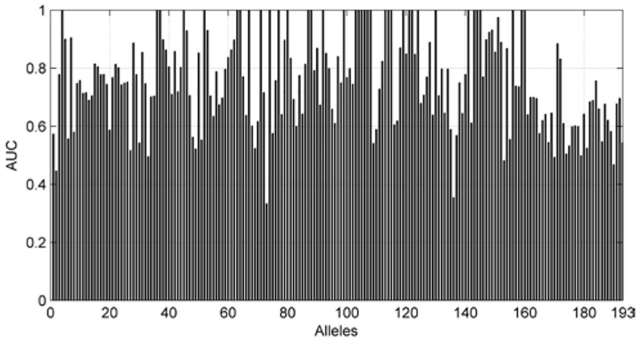


FIGURE 8 | AUC of each allele.

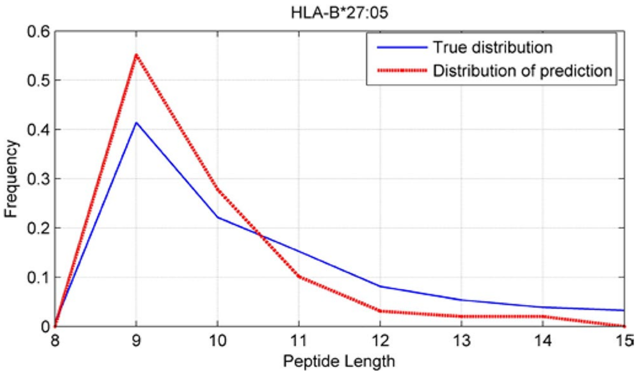


FIGURE 11 | Predicted length preference of HLA-B*27:05.

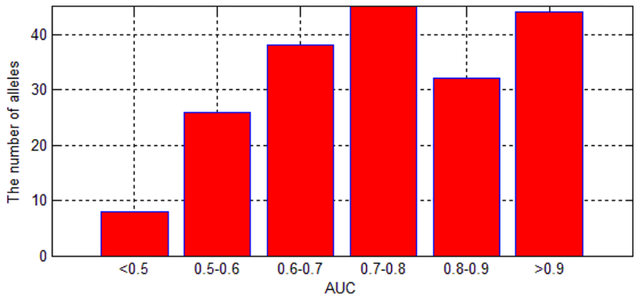


FIGURE 9 | The distribution of AUC in 193 experiments.

networks. The other more important reason to introduce CNN to this field is that the most commonly used format of feature for each peptide is a matrix. Therefore most researchers usually first convert the feature matrix into a line or a column. However, CNN could find out the real feature of each peptide by the initial feature matrix. In brief, CNN is more suitable for predicting peptide-MHC-I binding affinity.

Another novel thought of our paper is the way of extracting feature. The most common way to extract feature is based on BLOSUM nowadays. Although BLOSUM is a typical way to do sequence alignment, the order of the sequence and the characteristic of the acid amino would undoubtedly affect the binding of peptides to genes. Therefore, we extracted four kinds

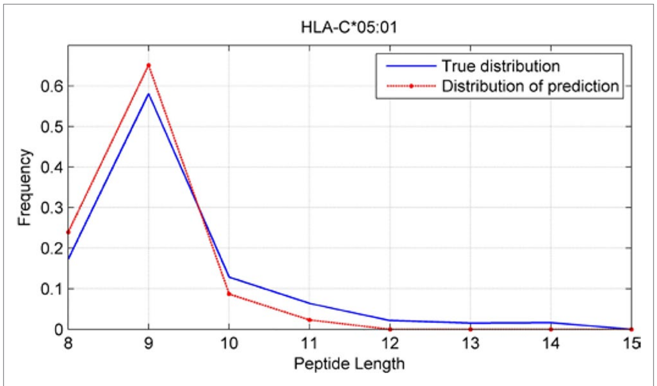


FIGURE 12 | Predicted length preference of HLA-C*05:01.

of feature for each peptide. They are the order of the sequence, hydropathy index, polarity, and length.

Our work flow can be concluded in three steps. Firstly, we convert every length of peptide into 15mer based on the binding mode of peptide to MHC I. Then, we extracted feature of each peptide based on the order of the sequence, hydropathy index, polarity, and length. For each peptide, the feature of it should be a matrix with 4 * 15 dimension. Finally, we built a frame of CNN and put these features and their corresponding label into it.

We put three data sets together and obtain 525,672 peptides. We built model for each alleles so we totally built 193 models. To verify the accuracy of our model, we did five cross validation. We compared our method with DCNN, NetMHCPan4.0, SMM, ANN and PickPocket. In most cases, the accuracy of CNN-NF is higher than that of other methods. In addition, we also use our model to test the preference of different alleles to length. The length preference obtained by prediction is very close to the true preference.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://tools.iedb.org/mhci/download/. Code and data are available at https://github.com/zty2009/MHC-I/tree/master.

AUTHOR CONTRIBUTIONS

TZh wrote this paper and did experiments. LC provided important ideas. This whole work is guided by TZa and YH. TZa and YH also provided all the materials and environment to complete this work.

FUNDING

TZ and YH are the corresponding authors. TZh and LC are co-first authors. This work was supported by the National Natural Science Foundation of China (Nos: 61571152 and 61502125), the National High-tech R&D Program of China (863 Program) [Nos: 2014AA021505, 2015AA020101, 2015AA020108], the National Science and Technology Major Project [Nos: 2013ZX03005012 and 2016YFC1202302], Heilongjiang Postdoctoral Fund (Grant No. LBH-Z15179), and China Postdoctoral Science Foundation (Grant No. 2016M590291).

REFERENCES

Abelin, J. G., Keskin, D. B., Sarkizova, S., Hartigan, C. R., Zhang, W., Sidney, J., et al. (2017). Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 46, 315–326. doi: 10.1016/j.immuni.2017.02.007

Alvarez, B., Barra, C., Nielsen, M., and Andreatta, M. (2018). Computational tools for the identification and interpretation of sequence motifs in immunopeptidomes. *Proteomics* 18, 1700252. doi: 10.1002/pmic.201700252

Andreatta, M., and Nielsen, M. (2015). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 32, 511–517. doi: 10.1093/bioinformatics/btv639

Barra, C., Alvarez, B., Paul, S., Sette, A., Peters, B., Andreatta, M., et al. (2018). Footprints of antigen processing boost MHC class II natural ligand predictions. *Genome Med.* 10, 84. doi: 10.1186/s13073-018-0594-6

Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L. J., and Mann, M. (2015). Mass spectrometry of HLA-I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics*, M114. 042812. doi: 10.1074/mcp.M114.042812

Bentzen, A. K., Marquard, A. M., Lyngaa, R., Saini, S. K., Ramskov, S., Donia, M., et al. (2016). Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat. Biotechnol.* 34, 1037. doi: 10.1038/nbt.3662

Bulik-Sullivan, B., Busby, J., Palmer, C. D., Davis, M. J., Murphy, T., Clark, A., et al. (2019). Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* 37, 55. doi: 10.1038/nbt.4313

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2017). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Briefings In Bioinf.* 20, 203–209. doi: 10.1093/bib/bbx103

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., 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.* doi: 10.1093/nar/gkz843

Chu, Y., Liu, Q., Wei, J., and Liu, B. (2018). Personalized cancer neoantigen vaccines come of age. *Theranostics* 8, 4238. doi: 10.7150/thno.24387

de Groot, N. G., Heijmans, C. M., de Ru, A. H., Janssen, G. M., Drijfhout, J. W., Otting, N., et al. (2017). A Specialist Macaque MHC Class I Molecule with HLA-B* 27-like Peptide-Binding Characteristics. *J. Immunol.* 199(10), 3679–3690. doi: 10.4049/jimmunol.1700502

Eisenberg, D. (1984). Three-dimensional structure of membrane and surface proteins. *Annu. Rev. Biochem.* 53, 595–623. doi: 10.1146/annurev.biochem.53.1.595

Garde, C., Ramarathinam, S. H., Jappe, E. C., Nielsen, M., Kringelum, J. V., Trolle, T., et al. (2019). Improved peptide-MHC class II interaction prediction through integration of eluted ligand and peptide affinity data. *Immunogenetics*, 71(7), 1–10. doi: 10.1007/s00251-019-01122-z

Giguere, S., Marchand, M., Laviolette, F., Drouin, A., and Corbeil, J. (2013). Learning a peptide-protein binding affinity predictor with kernel ridge regression. *BMC Bioinf.* 14, 82. doi: 10.1186/1471-2105-14-82

Han, Y., and Kim, D. (2017). Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC Bioinf.* 18, 585. doi: 10.1186/s12859-017-1997-x

Hao, J., Kang, E., Sun, J., Wang, Z., Meng, Z., Li, X., et al. (2016). An adaptive Markov strategy for defending smart grid false data injection from malicious attackers. *IEEE Trans. Smart Grid* 9, 2398–2408. doi: 10.1109/TSG.2016.2610582

Hao, J., Huang, D., Cai, Y., and Leung, H.-f. (2017). The dynamics of reinforcement social learning in networked cooperative multiagent systems. *Eng. Appl. Artif. Intell.* 58, 111–122. doi: 10.1016/j.engappai.2016.11.008

Jørgensen, K. W., Rasmussen, M., Buus, S., and Nielsen, M. (2014). Net MHC stab-predicting stability of peptide-MHC-I complexes; impacts for cytotoxic T lymphocyte epitope discovery. *Immunology* 141, 18–26. doi: 10.1111/imm.12160

Jensen, K. K., Andreatta, M., Marcatili, P., Buus, S., Greenbaum, J. A., Yan, Z., et al. (2018). Improved methods for predicting peptide binding affinity to MHC class II molecules. *Immunology* 154(3): 394–406. doi: 10.1111/imm.12889

Johnson, D. B., Estrada, M. V., Salgado, R., Sanchez, V., Doxie, D. B., Opalenik, S. R., et al. (2016). Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy. *Nat. Commun.* 7, 10582. doi: 10.1038/ncomms10582

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J. Immunol.* 199(9): 3360–3368. ji1700893. doi: 10.1101/149518

Kim, Y., Sidney, J., Buus, S., Sette, A., Nielsen, M., and Peters, B. (2014). Dataset size and composition impact the reliability of performance benchmarks for peptide-MHC binding predictions. *BMC Bioinf.* 15, 241. doi: 10.1186/1471-2105-15-241

Kreiter, S., Vormehr, M., Van de Roemer, N., Diken, M., Löwer, M., Diekmann, J., et al. (2015). Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* 520, 692. doi: 10.1038/nature14426

Kuksa, P. P., Min, M. R., Dugar, R., and Gerstein, M. (2015). High-order neural networks and kernel methods for peptide-MHC binding prediction. *Bioinformatics* 31, 3600–3607. doi: 10.1093/bioinformatics/btv371

Neefjes, J., Jongsma, M. L., Paul, P., and Bakke, O. (2011). Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* 11, 823. doi: 10.1038/nri3084

Nielsen, M., Lundegaard, C., and Lund, O. (2007). Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinf.* 8, 238. doi: 10.1186/1471-2105-8-238

O'Donnell, T. J., Rubinsteyn, A., Bonsack, M., Riemer, A. B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* 7, 129–132. e4. doi: 10.1016/j.cels.2018.05.014

Ott, P. A., Hu, Z., Keskin, D. B., Shukla, S. A., Sun, J., Bozym, D. J., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 547, 217. doi: 10.1038/nature22991

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019a). A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics*. 35(21), 4364–4371 doi: 10.1101/276048

Peng, J., Wang, X., and Shang, X. (2019b). Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data. *BMC Bioinf.* 20, 284. doi: 10.1186/s12859-019-2769-6

Peng, J., Guan, J., and Shang, X. (2019c). Predicting Parkinson's disease genes based on node2vec and autoencoder. *Front. In Genet.* 10, 226. doi: 10.3389/fgene.2019.00226

Phloyphisut, P., Pornputtapong, N., Sriswasdi, S., and Chuangsuwanich, E. (2019). MHCSeqNet: a deep neural network model for universal MHC binding prediction. *BMC Bioinf.* 20, 270. doi: 10.1186/s12859-019-2892-4

Rist, M. J., Theodossis, A., Croft, N. P., Neller, M. A., Welland, A., Chen, Z., et al. (2013). HLA peptide length preferences control CD8+ T cell responses. *J. Immunol.* 191(2): 561–571. doi: 10.4049/jimmunol.1300292

Rolland, M., Tovanabutra, S., Frahm, N., Gilbert, P. B., Sanders-Buell, E., Heath, L., et al. (2011). Genetic impact of vaccination on breakthrough HIV-1 sequences from the STEP trial. *Nat. Med.* 17, 366. doi: 10.1038/nm.2316

Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.-P., Simon, P., Löwer, M., et al. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* 547, 222. doi: 10.1038/nature23003

Salimi, N., Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., et al. (2019). The Immune Epitope Database enables and accelerates research. *J. Immunol.* 202, 131.20–131.20

Sidhom, J.-W., Pardoll, D., and Baras, A. (2018). AI-MHC: an allele-integrated deep learning framework for improving Class I & Class II HLA-binding predictions. *bioRxiv*, 318881. doi: 10.1101/318881

Styczynski, M. P., Jensen, K. L., Rigoutsos, I., and Stephanopoulos, G. (2008). BLOSUM62 miscalculations improve search performance. *Nat. Biotechnol.* 26, 274. doi: 10.1038/nbt0308-274

Tran, N. H., Qiao, R., Xin, L., Chen, X., Liu, C., Zhang, X., et al. (2019). Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* 16, 63–66. doi: 10.1038/s41592-018-0260-3

Trolle, T., Metushi, I. G., Greenbaum, J. A., Kim, Y., Sidney, J., Lund, O., et al. (2015). Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics* 31, 2174–2181. doi: 10.1093/bioinformatics/btv123

Udaka, K., Wiesmüller, K.-H., Kienle, S., Jung, G., and Walden, P. (1995). Tolerance to amino acid variations in peptides binding to the major histocompatibility complex class I protein H-2Kb. *J. Biol. Chem.* 270, 24130–24134. doi: 10.1074/jbc.270.41.24130

Uslan, V., and Seker, H. (2016). Quantitative prediction of peptide binding affinity by using hybrid fuzzy support vector regression. *Appl. Soft Comput.* 43, 210–221. doi: 10.1016/j.asoc.2016.01.024

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2018). The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.* 47, D339–D343. doi: 10.1093/nar/gky1006

Wolfenden, R., Lewis, C. A., Yuan, Y., and Carter, C. W. (2015). Temperature dependence of amino acid hydrophobicities. *Proc. Natl. Acad. Sci.* 112, 7484–7488. doi: 10.1073/pnas.1507565112

Zeng, H., and Gifford, D. K. (2019). Quantification of uncertainty in peptide-MHC binding prediction improves high-affinity peptide Selection for therapeutic design. *Cell Syst.* 9(2), 159–166.e3. doi: 10.1016/j.cels.2019.05.004

Zhang, H., Lund, O., and Nielsen, M. (2009). The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* 25, 1293–1299. doi: 10.1093/bioinformatics/btp137

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 09 April 2020
doi: 10.3389/fbioe.2020.00298

Deep Reinforcement Learning for Data Association in Cell Tracking

Junjie Wang[1], Xiaohong Su[1], Lingling Zhao[1]* and Jun Zhang[2]*

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China, [2] Department of Rehabilitation, Heilongjiang Province Land Reclamation Headquarters General Hospital, Harbin, China

Accurate target detection and association are vital for the development of reliable target tracking, especially for cell tracking based on microscopy images due to the similarity of cells. We propose a deep reinforcement learning method to associate the detected targets between frames. According to the dynamic model of each target, the cost matrix is produced by conjointly considering various features of targets and then used as the input of a neural network. The proposed neural network is trained using reinforcement learning to predict a distribution over the association solution. Furthermore, we design a residual convolutional neural network that results in more efficient learning. We validate our method on two applications: the multiple target tracking simulation and the ISBI cell tracking. The results demonstrate that our approach based on reinforcement learning techniques could effectively track targets following different motion patterns and show competitive results.

Keywords: cell tracking, linear assignment problem, deep learning, deep reinforcement learning, data association, residual CNN

OPEN ACCESS

Edited by:
Jiajie Peng,
Northwestern Polytechnical University, China

Reviewed by:
Xishuang Dong,
Prairie View A&M University, United States
Naifeng Wen,
Dalian Nationalities University, China
Zhong Li,
Institute of Disaster Prevention, China

****Correspondence:***
Lingling Zhao
zhaoll@hit.edu.cn
Jun Zhang
zhangjun13902003@163.com

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

Received: *15 January 2020*
Accepted: *20 March 2020*
Published: *09 April 2020*

Citation:
Wang J, Su X, Zhao L and Zhang J (2020) Deep Reinforcement Learning for Data Association in Cell Tracking. Front. Bioeng. Biotechnol. 8:298. doi: 10.3389/fbioe.2020.00298

1. INTRODUCTION

Tracking individual cells in a group is the fundamental of many biomedical analysis tasks, including understanding how genotypes are related to phenotypes, tracking the early development of organs and meristems, and potentially tracking the development of cancerous tumors (Cheng et al., 2019, 2020; Han et al., 2019; Hu et al., 2019). It is often necessary to identify individual cells and follow them over time to gain biological insights from time-lapse microscopy recordings of cell behavior. Microscopic target tracking can provide technical support for the analysis of other features in biological and medical research (Cheng, 2019; Zhao et al., 2020). Therefore, it is of great significance to find an automatic and reliable way to track multiple cells.

There are many procedures and methods for tracking objects at the microscopic level. Tracking-by-detection methods are widely used in multi-target tracking, in which detection and association are two primary issues. Extensive research efforts have focused on detection, especially in cell-tracking applications. For target association between frames, the naïve nearest-neighbor method is commonly adopted but provides unsatisfactory association accuracy. Target association is a combinatorial optimization problem, which is widely studied in computer science and mathematics and many such problems are NP-hard. In general, the linear assignment problem is to find the optimal assignment that maximizes or minimizes the sum of the costs in a cost matrix. Classic algorithms for the linear assignment include the Hungarian method (Kuhn, 1955), auction algorithms (Bertsekas, 1992), and certain variant algorithms.

Recently, some data-driven methods have been proposed to solve combinatorial problems. Vinyals et al. (2015) first proposed a pointer network (PN) to solve combinatorial problems such as the traveling salesman problem and convex hulls. Inspired by the PN, in Milan et al. (2017), a recurrent neural network is used to find the marginal probability based on the cost matrix. A deep Hungarian network based on the recurrent neural network has also been proposed for multi-object tracking (Xu et al., 2019).

Substantial progress in artificial intelligence has been made in supervised learning, where systems are trained on vast amounts of labeled data (Peng et al., 2019a,b,c, 2020). However, supervised learning predominantly works in domains with an abundance of human-labeled data. In many challenging domains, supervised learning fails due to a lack of available data. Reinforcement learning (RL) seeks to create intelligent agents that adapt to an environment by analyzing their own experiences. Bello et al. (2016) and Khalil et al. (2017) suggested using the RL method to train the network without the ground-truth labels. Because it is difficult to obtain optimal solutions for certain NP-hard combinatorial problems. RL is a branch of machine learning that focuses on obtaining an optimal policy to solve specific problems. Following the work of Bello et al. (2016), some researchers have proposed different deep reinforcement learning (DRL)-based methods for solving combinatorial problems that have yielded good performance (Emami and Ranka, 2018; Nazari et al., 2018; Fu et al., 2019).

This work is motivated by several recent proposed DRL methods for NP-hard problems. We propose a DRL approach to automatically search for assignment solutions for a given cost matrix. Specifically, we first modeled the association of cells between frame as an linear assignment problem and formulated the assignment problem with the one-to-one constraint as a DRL problem. Then, with the objective of minimizing the sum of the assignment costs, we used DRL to obtain the optimal assignment solution. To convert the cost matrix into a finite action space, we employ the residual learning and convolutional neural network (CNN) to extract features from a set of training samples and use the pointing mechanism (Bello et al., 2016) to satisfy the one-to-one constraints of the linear assignment problem. Then, the CNN is trained with the REINFORCE algorithm (Williams, 1992) to search for assignment solutions, and the sum of the cost matrix of the selected solution is used as a reward to adjust the parameters of the neural networks.

Our contributions are the following:(1) A simple framework for cell detection and association based on the idea of (2) We introduce a formulation that translates the decision making in the linear assignment problem algorithm into an RL problem. (3) We propose a novel neural network architecture that end-to-end maps the inputs to the decision outputs.

The organization of this paper is as follows. Related work is introduced in section 2. The framework of the proposed method and training details are presented in section 3. In section 4, some experiments are conducted to evaluate the performance of our proposed method. The conclusion is given in section 5.

2. RELATED WORK

2.1. Cell Tracking

A large variety of cell tracking methods have been described in the existing literature. These cell tracking methods can be broadly grouped into two categories: (i) tracking by model evolution and (ii) tracking by detection.

In tracking by model evolution methods, cell segmentation and tracking are solved simultaneously in each frame of a cell video. Typically, these methods are driven by data in some feature space and make a regularity assumption on the smoothness of the curve. In this framework, cells are represented by parametric or implicit active contour models. Parametric models utilize the explicit representations of cell boundaries such as Gaussian Mixture Models (GMM) (Amat et al., 2014), active meshes (Dufour et al., 2010), or active contours (Zimmer et al., 2002). Implicit models often use the level set to represent the cell contours (Dzyubachyk et al., 2010). These cell tracking methods have some shortcomings. For example, the parametric method depends on the chosen parameterization, and the implicit method is computationally expensive.

Existing cell tracking methods generally adopt the tracking by detection strategy. The tracking by detection method typically consists of two stages: the cell detection stage and cell association stage. In the first stage, the cells are detected by image segmentation methods. Subsequently, in the second stage, detected cells are associated with neighboring frames in real-time or all frames offline. Cell detection can be achieved by classic image segmentation algorithms based on intensity features, gradient features, or texture features (Chenouard et al., 2013; Xing and Yang, 2016). Recently, several deep learning approaches have shown significant success in cell segmentation tasks (Ronneberger et al., 2015; Falk et al., 2019; Gupta et al., 2019).

3. METHODS

In this section we present a tracking by detection approach to construct the cell trajectories from a time-series microscopy image sequence. The framework consists of two modules: cell detection and cell association. The U-Net segmentation method is employed to detect all the cells in each frame, and then we adopt the traditional single hypothesis tracking method with Kalman filter and frame-by-frame data association to produce the cell trajectories.

3.1. Initial Cell Segmentation

Ronneberger et al. (2015) proposed a new neural network for cell segmentation, namely U-Net, which has achieved state-of-the-art results on a wide array of biomedical image segmentation tasks (Ronneberger et al., 2015; Falk et al., 2019). Since then, most attempts to improve the performance of cell tracking methods have been based on the U-Net architecture (Li et al., 2018). In our approach, cell segmentation is performed using the U-Net implementation of Ronneberger et al. (2015).

3.2. Cell Time-Series Model

In this work, we assume that each cell can be modeled as a discrete-time Markov process:

$$x_t = Ax_{t-1} + Q_t \tag{1}$$

where A is the transition matrix and Q_t is the process noise matrix, which follows a Gaussian distribution. Once the detected cells are retrieved, the detection results Z_t can be viewed as the measurements, where each measurement $z_t^i \in Z_t$ is defined as

$$z_t^i = Hx_t^i + R_t \tag{2}$$

A Kalman filter can be adopted to use those cell detection results to predict the state of cells, which can then be used to formulate the cell association between frames as a linear assignment problem.

3.3. Deep Reinforcement Learning Based Cell Association

To solve the target association problem by DRL, we present our solution architecture in three parts: (1) Problem Formulation. We formulate the procedure for selecting an assignment solution as an RL problem to associate target states and measurements. (2) Neural network architecture. An end-to-end architecture that maps from the state space to the action space is designed. (3) Training algorithm. We present the RL algorithm used for the policy search.

3.3.1. Problem Formulation

3.3.1.1. The formulation of linear assignment problem

Assume that the cell trajectories can be denoted as a set $\Omega_{t-1} = \{\omega_{t-1}^1, \omega_{t-1}^2, ..., \omega_{t-1}^{M_{t-1}}\}$ at time $t-1$. Each element of Ω_{t-1} corresponds to a cell trajectory. To find their associated new measurements at time step t, each trajectory would be predicted by a Kalman filter and then find the possible association between predicted cell states and new measurements. Let the set $B = \{\hat{x}_{t|t-1}^1, \hat{x}_{t|t-1}^2 \cdots, \hat{x}_{t|t-1}^{M_{t-1}})\}$ represent the predicted states for all the existing cells at time $t-1$. Then the association mapping from set B to the measurement set $Z_t = \{z_t^1, z_t^2, ..., z_t^{N_t}\}$ can be treated as an assignment problem.

The values of the cost matrix D are calculated through the location distance between the elements of set B and the measurements as shown in **Figure 1**. Unlike the conventional association cost matrix, we construct a new cost matrix that considers the association event. To be specific, matrix D is defined as

$$\mathcal{D} = \begin{pmatrix} \Lambda & \Upsilon \\ \Gamma & \Lambda^T \end{pmatrix} \tag{3}$$

where D is a $(M_{t-1} + N_t) \times (M_{t-1} + N_t)$ square matrix, with the row and column indices representing the M_{t-1} prediction from trajectories and N_t measurements. The matrix D consists of four sub-matrices $\Lambda(M_{t-1} \times N_t)$, $\Upsilon(M_{t-1} \times M_{t-1})$, and $\Gamma(N_t \times N_t)$ implies that the corresponding target's state is judged as "Tracked", "Lost," and "New," respectively. In the sub-matrices $\Upsilon(M_{t-1} \times M_{t-1})$ and $\Gamma(N_t \times N_t)$, we define the value of the diagonal element as a distance threshold and other elements to be ∞. Here, when a predicted state is highly self-associated, we consider it to be lost. An estimated state that highly associates itself is considered as a new target. The elements of the sub-matrix $\Lambda(M_{t-1} \times N_t)$ are the distances between the prediction state and measurements.

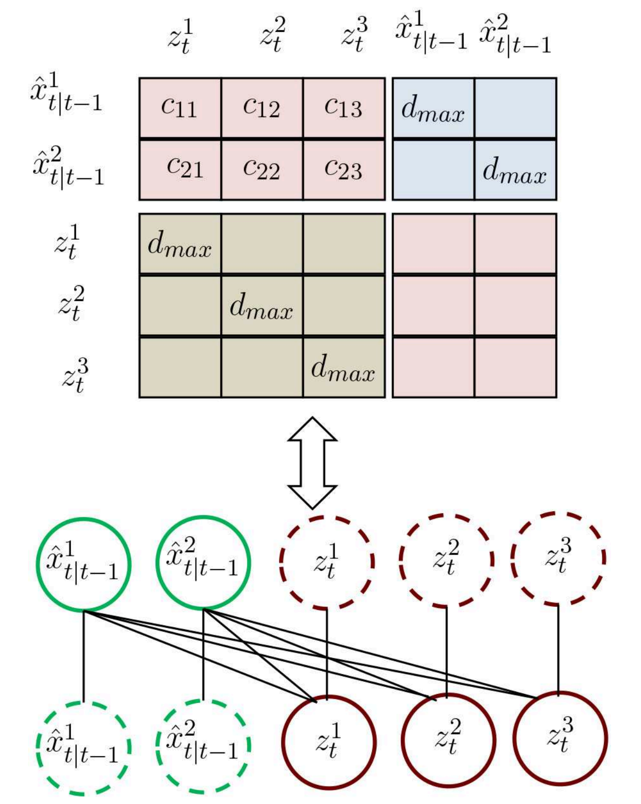


FIGURE 1 | Illustration of the proposed association matrix. $\{\hat{x}_{t|t-1}^i\}_{i=1}^2$ is the prediction by the Kalman filter. $\{\hat{z}_t^i\}_{i=1}^3$ are the measurements.

3.3.1.2. RL formulating for the linear assignment problem

The standard RL formulation starts with an MDP: at time step $t \geq 0$, an agent is in a state $s_t \in \mathcal{S}$, takes an action $a_t \in \mathcal{A}$, receives an instant reward $r_t \in \mathbb{R}$ and transitions to the next state $s_{t+1} \sim p(\cdot|s_t, a_t)$. A policy $\pi : \mathcal{S} \mapsto \mathcal{P}(\mathcal{A})$ gives a mapping from any state to a distribution over actions $\pi(\cdot|s_t)$. The objective of RL is to search for a policy that maximizes the expected cumulative rewards over a horizon T, i.e., $\max_\pi \; J(\pi) : = \mathbb{E}[\sum_{t=0}^{T-1} r_t \gamma^t; \pi]$, where $\gamma \in (0, 1]$ is a discount factor and the expectation is w.r.t. randomness in the policy π as well as the environment [e.g., the transition dynamics $p(\cdot|s_t, a_t)$]. In practice, we consider parameterized policies π_θ and aim to find $\theta^* = \arg\max J(\pi_\theta)$.

To formulate the procedure of selecting assignment solution algorithms into an MDP, we specify below the state space $\mathcal{S}$, action space $\mathcal{A}$, reward function r_t and transition dynamics $s_{t+1} \sim p(\cdot|s_t, a_t)$.

State Space $\mathcal{S}$. The set of states ($\mathcal{S}$) is defined as all costs of the predicted cell assigned to the detected cell. In this sense, the set $\mathcal{S}$ varies according to the number of tasks in the instance.

Action Space $\mathcal{A}$. The agent can choose to either assign a predicted cell to a detected cell or not. Thus, we define the action space as

$\mathcal{A} = \{0, 1\}$, where 1 represents the predicted cell assigned to a detected cell and 0 represents otherwise.

Reward r_t. For most RL applications, designing a reward function is always a critical part, especially when the agent needs to precisely perform actions in a complicated task. A good reward function will make the agent learn more efficiently and achieve better results. By contrast, an agent with a poor reward function may suffer slow convergence or even produce undesirable results. The objective of the linear assignment problem is to minimize the total cost of the assignment solution. To achieve this objective, we design the reward function as the sum of the assignment cost after producing an assignment solution. Given a cost matrix $C = \{c_{ij}\}, i = 1, ..., N, j = 1, ...N$ and a selected assignment solution $X = \{x_{ij}\}, i = 1, ..., N, j = 1, ...N$, the reward r_t can be defined as $r_t = \sum_{i=1}^{N} \sum_{j=1}^{N} c_{ij}x_{ij}$.

Transition s_t. In our work, the state transition is deterministic after an action has been chosen because it can directly assign the corresponding task to the person.

3.3.2. Architecture Details

The input of our residual CNN (ResCNN) is a cost matrix C that can be treated as the sum of a probability distribution for matching X and a noise V as

$$C = X + V \tag{4}$$

The sequence-to-sequence models the linear assignment problem (Milan et al., 2017; Emami and Ranka, 2018) with the aim of learning a mapping function $\mathcal{F}(C) = X$ to directly predict the probability distribution. For ResCNN, we adopt the residual learning framework to train a residual mapping $\mathcal{R}(C) \approx V$, and then we have $X = C - \mathcal{R}(C)$. **Figure 2** illustrates the architecture of the proposed ResCNN for learning $\mathcal{R}(C)$. In the following, we explain the architecture of ResCNN.

Our proposed neural network is similar to the image denoising network introduced in Zhang et al. (2017). The input of the neural network is a cost matrix that can be regarded as a single-channel image. With the cost matrix C as input, the following ResCNN consists of a series of different types of fundamental blocks. The first block consists of a convolution layer (Conv) and a rectified linear unit (ReLU) (Krizhevsky et al., 2012) layer, where the convolution layer utilizes 8 filters of size $3 \times 3 \times 1$ to generate 8 feature maps. Then, the 8 feature maps are fed into three Conv+BN+ReLU-type blocks. For these three blocks, 8 filters of size $3 \times 3 \times 64$ are used, and batch normalization (BN) (Ioffe and Szegedy, 2015) is added between convolution and ReLU. Then, the noise V is computed by the last convolution layer, and the probability distribution of the assignment matrix X is subtracted from its input (cost matrix). Finally, the probability distribution is clipped by the tanh activation function so that the intensities of the output lie in the range [–1,1].

In summary, the main feature of our ResCNN is the adoption of residual learning to learn $\mathcal{R}(\mathbf{C})$ rather than the probability distribution directly. In addition, borrowing the idea of Zhang et al. (2017), batch normalization is incorporated into the ResCNN to speed up the training procedure and improve the performance.

In the following, we will give some important details about our network design and training.

3.3.2.1. Integration of Residual Learning and Batch Normalization

Batch normalization is a standard technique that is widely used in image classification CNN models. Training a deep neural network model is often difficult not only because of the gradient vanishing/exploding problem but also because the distribution of data changes between layers, which is called the "internal covariate shift" phenomenon. Batch normalization is a technique that can relieve this phenomenon by introducing several simple operations to the input data. The goal of the normalization step for batch normalization is to transform the layer input t before non-linearity as follows:

$$t' = \frac{t - \mathrm{E}[t]}{\sqrt{\mathrm{Var}[t]}} \tag{5}$$

where $\mathrm{E}[t]$ and $\mathrm{Var}[t]$ are the expectation and variance computed over all training data. It is usually impractical to exactly calculate $\mathrm{E}[t]$ and $\mathrm{Var}[t]$ with stochastic optimization. Batch normalization instead approximates $\mathrm{E}[t]$ and $\mathrm{Var}[t]$ via the mini-batch statistics during training. It would be beneficial if the mini-batch statistics agree well with the full training data statistics.

Batch normalization and residual learning are two important algorithms for designing a neural network architecture. Residual learning and batch normalization can benefit from each other (Zhang et al., 2017). In this paper, we adopt this strategy by integrating these two technologies. Specifically, such an integration not only can significantly increase the training speed but also tends to improve the performance.

3.3.2.2. Zero Padding to Avoid Boundary Artifacts

In the linear assignment problem, the input and output need to be consistent. However, due to the characteristics of convolution, the neural network is prone to producing boundary artifacts without proper handling. There are two common ways to solve this problem: symmetrical padding and zero padding. In our work, we select zero padding to maintain a consistent matrix size.

3.3.2.3. Pointing Mechanism to Satisfy the Constraints

Unlike ordinary visual tasks, for the linear assignment problem, one major characteristic is that one detected cell can only be assigned to one predicted cell. The neural network output should satisfy one-to-one constraints. Let $X = C - V$ denote the outputs of the neural networks. To avoid collisions whereby one task may be assigned to multiple cells simultaneously, we use a mask to set the probability of detected cell that have already been assigned to a predicted cell to $-\infty$, as shown in Equation (6)

$$u_{ij} = \begin{cases} Y_{ij} & \text{if } j \neq \pi_{i'} \quad \forall i' < i \\ -\infty & \text{otherwise.} \end{cases} \tag{6}$$

where u_{ij} is the probability that predicted cell i at time $t - 1$ is assigned to detected cell j at time t. $\pi_{i'}$ is the solution for cell i'.

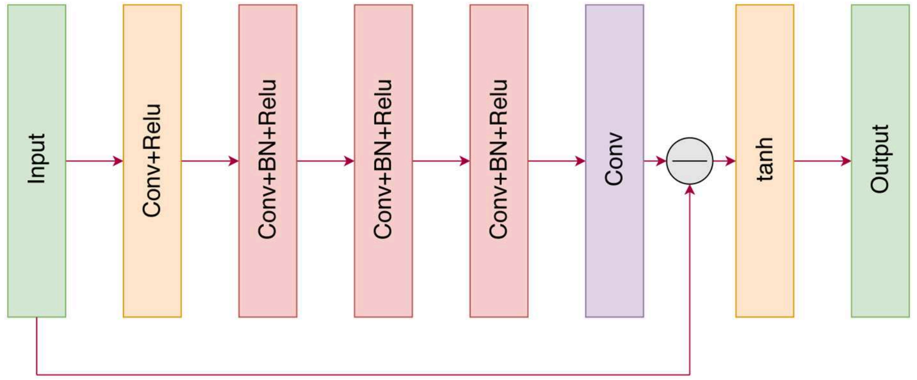


FIGURE 2 | The architecture of the proposed ResCNN network.

Next, a normalized softmax operation is applied to u to compute the final output probability matrix.

3.3.3. Training With Policy Gradients

In this paper, we utilize the RL to train the neural network. The input of the network can be denoted as $C = c_{ij}$. The output of the network is the assignment solution π. In this work, we use the sum of the selected costs $AC(\mathcal{S}|C)$ as the reward. More specifically, the parameters of the neural network can be denoted as θ, and the goal of training is the expected reward, which is given by an input cost matrix C defined as follows:

$$J(\theta|C) = E_{\pi\sim p(\pi|C;\theta)}AC(\pi|C) \tag{7}$$

In our work, $p(\pi|C;\theta)$ is the stochastic policy of a neural network with parameters θ. We learn θ using the Adam optimizer based on the REINFORCE algorithm (Williams, 1992). REINFORCE can make weight adjustments in a direction that lies along the gradient of expected reinforcement. Based on REINFORCE, in each step of training, if the reward, baseline value and probability distribution of prediction are obtained, then the parameters of the neural network, θ, are incremented by an amount

$$\nabla_\theta J(\theta|C) = E_{\pi\sim p_{\theta(\cdot|C)}}\left[(AC(\pi|C) - b(C))\nabla_\theta \log p_\theta(\pi|C)\right] \tag{8}$$

where $b(C)$ denotes the baseline value of the assignment cost and is used to reduce the variance of the gradients. If we randomly obtain M *i.i.d.* samples, then the above gradients can be approximated by

$$\nabla_\theta J(\theta|C) \approx \frac{1}{M}\sum_{i=1}^{M}\left[\left(AC\left(\pi_i|C_i\right) - b\left(C_i\right)\right)\nabla_\theta \log p_\theta\left(\pi_i|C_i\right)\right] \tag{9}$$

For a cost matrix, the baseline value $b(C_i)$ is initialized by calculating the sum of the cost of the assignment solution that is generated by the neural network. In each step, the baseline value is updated as follows:

$$b'\left(C_i\right) = b\left(C_i\right) + \alpha\left(AC\left(\pi_i|C_i\right) - b\left(C_i\right)\right) \tag{10}$$

Algorithm 1 gives the pseudo-code of the training procedure of the neural network.

Algorithm 1: Training Procedure

```
Training set $\{C_i\}_{i=1}^{M}$, number of training steps $T$, batch size $B$.
Initialize the neural net params $\theta$.
Initialize baseline value.
for $t = 1$ to $T$ do
    Select a batch of samples $C_i$ for $i \in \{1, \cdots, B\}$.
    Sample solution $\pi_i$ based on $p_\theta(\cdot|C_i)$ for $i \in \{1, \cdots, B\}$.
    Let $g_\theta = \frac{1}{B}\sum_{i=1}^{B}[(AC(\pi_i|C_i) - b(C_i))\nabla_\theta log p_\theta(\pi_i|C_i)]$.
    Update $\theta = ADAM(\theta, g_\theta)$.
    Update baseline $b(C_i) = b(C_i) + \alpha(AC(\pi_i|C_i) - b(C_i))$ for $i \in \{1, \cdots, B\}$.
end for
return neural net parameters $\theta$.
```

4. EXPERIMENTS

To evaluate the performance of our proposed method, we consider two applications of the linear assignment problem: maximum weight matching (MWM) and data association for multi-target tracking. We first compare our method with the state-of-the-art DRL method for maximum weight matching. Then, we test our method on a multi-target tracking scenario. Finally, we evaluate our proposed method on three cell microscopy datasets, Fluo-N2DH-GOWT1, PhC-C2DH-U373, and Fluo-N2DH-SIM+ from the ISBI 2015 Cell Tracking Challenge (Maška et al., 2014). Each datasets contains 2 training sequences and 2 challenge sequences. Since it's hard to get the ground truth of segmentation and trajectories in the challenge datasets, we performed tracking experiments on testing datasets.

In all experiments, we used 500, 000 training samples for the data association. To produce the training samples, we randomly sample $M + N$ points in the euclidean space to simulate the data association between two frames. We use the same hyper-parameters to train our model. The initial learning rate for the

TABLE 1 | Median optimality ratios on the MWM test set.

	***N* = 15**	***N* = 20**	***N* = 25**
AC+Matching	0.935	0.897	0.725
SPG+Matching	0.904	0.895	0.889
Ours	0.977	0.968	0.965

TABLE 2 | Average OSPA-T distance and IDSW for different methods over 100 random runs.

Method	$R = 0.01\mathbf{I}_2$		$R = 0.05\mathbf{I}_2$		$R = 0.1\mathbf{I}_2$	
	OSPA-T	**IDSW**	**OSPA-T**	**IDSW**	**OSPA-T**	**IDSW**
JPDA	0.19(0.05)	0.90(0.88)	0.34(0.11)	0.70(0.82)	0.41(0.12)	0.40(0.70)
JPDA$_{10}$	0.23(0.10)	0.70(0.67)	0.37(0.11)	0.90(0.88)	0.43(0.09)	1.10(0.99)
JPDA-HA	0.28(0.06)	0.60(0.84)	0.37(0.10)	0.70(0.95)	0.46(0.14)	1.30(0.82)
JPDA-RL	0.28(0.06)	0.60(0.84)	0.36(0.08)	0.60(0.70)	0.45(0.13)	1.10(0.99)
LSTM	0.11(0.01)	1.07(0.84)	0.21(0.01)	1.00(0.74)	0.37(0.11)	0.60(0.89)

The standard deviations are given in parentheses.

Adam optimizer is 10^{-3} and decays every 5,000 steps by a factor of 0.96.

4.1. Maximum Weight Matching

Define a weighted bipartite graph $G = (V = \{V_1, V_2\}, E)$, where V is the vertex set containing two disjoint vertex sets V_1 and V_2, with $|V_1| = N$ and $|V_2| = N$, and E is the set of all edges between every node $v_1 \in V_1$ and $v_2 \in V_2$. Let $w_{ij}, i \leq i \leq N, 1 \leq j \leq N$ denote the associated weight for the edges in the graph. Then, a matching in a graph G is a subset of E such that no two edges share a common vertex. A maximum weight matching is a matching such that the sum of the weights of the edges in the matching is maximal (Emami and Ranka, 2018). In our simulation, each vertex of the graph is represented by a point (x_i, y_i), and W_{ij} is the Euclidean distance between vertex i and j. We select the optimality ratio as $\frac{\text{predicted matching weight}}{\text{optimal matching weight}} \in [0, 1]$ to measure the performance of our proposed method. The optimal matching weight is computed by the Hungarian algorithm, and the predicted matching weight is obtained by our method.

We trained our method on MWM with $N = \{15, 20, 25\}$. The results are compared with SPG+Matching and AC+Matching (Emami and Ranka, 2018), two DRL method solvers for the MWM problem. The results in **Table 1** are the median optimality ratios on the test set. As a baseline, the performances of SPG+Matching and AC+Matching also are presented in **Table 1**. We observe drastic drops in median optimality ration for the AC+Matching methods with an increasing number of nodes. By contrast, the performances of SPG+Matching and our method show less drastic drops. The results clearly show that our model is competitive with AC+Matching and SPG+Matching methods.

4.2. Simulated Multiple Target Tracking

One major application of linear assignment is data association for multi-target tracking. Therefore, we set up a simulated multi-target tracking scenario to evaluate the performance of the proposed method similar to Milan et al. (2017). Five targets cross each other at a certain time. The track state x is represented by a vector $\left[\, x \;\, y \;\, \dot{x} \;\, \dot{y} \,\right]$, which contains the position (x, y) and velocity $(\dot{x}, \dot{y})$ information. **Figure 3A** shows the ground truth of the five targets. The measurements provide noisy positions for the targets, i.e., $z_t = Hx_t + v_t$, where $H = \left[\, 1 \;\, 0 \,\right] \otimes \mathbf{I}_{2\times2}$ and $v_t \sim \mathcal{N}\{0, R\}$. **Figure 3B** gives the measurements for $R = 0.05\mathbf{I}_{2\times2}$.

We replace the data association part of JPDA with our method and call it JPDA-RL. The input matrix $C \in \mathbb{R}^{N\times N}$ is the Mahalanobis distance between the estimated target states and the measurements. We compare JPDA-RL with the traditional joint probabilistic data association (JPDA) filter (Fortmann et al., 1980), an approximation of the JPDA filter with the 10 best association hypotheses (Hamid Rezatofighi et al., 2015), an approximation of the JPDA filter with the Hungarian algorithm used to solve the association probabilities and the supervised LSTM used to solve the association problem in Milan et al. (2017).

Figure 3 shows the tracking results from the traditional JPDA filter and our proposed method with the JPDA filter of a single run. The traditional JPDA filter cannot handle the coalescence phenomenon. Our method can correctly distinguish the targets after they have crossed each other.

We employ two metrics to evaluate the tracking results: the Optimal Sub-pattern Assignment metric for track (OSPA-T) and Number of Identity Switch (IDSW). The OSPA-T distance (Ristic et al., 2011) is a metric used to evaluate differences between the real tracks $T_t = \left\{X_t^1, \ldots, X_t^m\right\}$ and the estimated tracks $\widehat{T}_t = \left\{\hat{X}_t^1, \ldots, \hat{X}_t^n\right\}$ by computing the quantity

$$\begin{aligned} d_p^{(c)}\left(T_t, \widehat{T}_t\right) = &\left(\frac{1}{n}\left(\min_{\pi\in\Pi_n}\sum_{i=1}^{m} d^{(c)}\left(X_t^i, \widehat{X}_t^{\pi(i)}\right)^p \right.\right. \\ &\left.\left. + c^p(n-m)\right)\right)^{1/p} \quad \text{if } m \leq n \\ d_p^{(c)}\left(T_t, \widehat{T}_t\right) = &d_p^{(c)}\left(\widehat{T}_t, T_t\right) \qquad \text{elsewhere} \end{aligned} \tag{11}$$

where $d(\cdot,\cdot)$ is the L2-norm, Π_n is the permutations in $\{1, \ldots, n\}$ and $d^{(c)}\left(X_t^i, \hat{X}_t^j\right)$ is the distance between X_t^i and $\hat{X}_t^j$ such that

$$d^{(c)}\left(X_t^i, \widehat{X}_t^j\right) = \min\left(c, d\left(X_t^i, \widehat{X}_t^j\right)\right) \tag{12}$$

To compute the OSPA-T distance for the estimated tracks and true tracks, two parameters, the cardinality penalty c and outlier sensitivity p, need to be set. In our simulations, we set $c = 1$ and $p = 1$.

In **Table 2**, we present a comprehensive comparison of the average OSPA-T distance and IDSW for different algorithms for different measurement noise levels. Interestingly, the IDSW of our method is lower for other algorithms at low measurement noise levels.

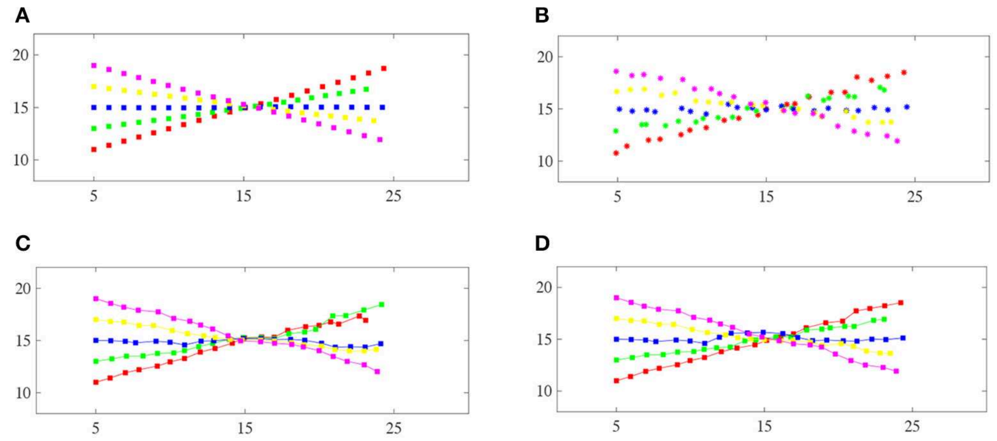


FIGURE 3 | Comparison of the track maintenance performance of different algorithms: **(A)** Ground-truth trajectories of the five targets, **(B)** the measurements of the five targets, **(C)** the JPDA filter, **(D)** our proposed method. Each color corresponds to a particular target. Note that our method correctly resolves this crossing case, whereas the JPDA filter switches the two trajectories after the targets cross.

TABLE 3 | TRA, SEG and OPT performance for our method, CPN, KTH (Magnusson and Jaldén, 2012), BLOB (Akram et al., 2016), U-Net (Ronneberger et al., 2015), U-Net-S (Gupta et al., 2019), and GC-ME (Bensch and Ronneberger, 2015).

		TRA	SEG	OPT
Fluo-N2DH-GOWT1-01	CPN	0.9864	0.8506	0.9185
	BLOB	0.9733	0.7415	0.8574
	KTH	0.9462	0.6849	0.8155
	Ours	**0.9875**	**0.8585**	**0.9230**
Fluo-N2DH-GOWT1-02	CPN	**0.9719**	0.8725	0.9222
	BLOB	0.9628	0.9046	0.9337
	KTH	0.9452	0.8942	0.9197
	Ours	0.9575	**0.9181**	**0.9378**
PhC-C2DH-U373-01	CPN	0.9594	0.7336	0.8456
	U-Net	0.9869	**0.9375**	**0.9622**
	GC-ME	0.9779	0.8748	0.9264
	Ours	**0.9919**	0.8527	0.9223
PhC-C2DH-U373-02	CPN	0.9346	0.7376	0.8361
	U-Net	**0.9547**	**0.8303**	0.8925
	GC-ME	0.9040	0.7567	0.8304
	Ours	0.9318	0.7735	0.8527
Fluo-N2DH-SIM+-01	U-Net-S	**0.9862**	**0.8866**	**0.9364**
	Ours	0.9841	0.8854	0.9348
Fluo-N2DH-SIM+-02	U-Net-S	0.9597	0.7381	0.8489
	Ours	**0.9618**	**0.7616**	**0.8617**

The best TRA and SEG values for each sequence are highlighted.

4.3. Cell Tracking

The segmentation task by U-Net and data association by DRL are conducted on AMD Ryzen 9 3900X 12 core processors with a GeForce GTX 2060 graphics card. For comparison, segmentation (SEG), tracking (TRA) accuracy measures and overall performance (OP) are adopted to evaluate the tracking performance. For TRA, Acyclic Oriented Graph Matching (AOGM) is used to count the changes needed to transform the cell tracking family tree into the ground-truth graph. OP is defined as the mean of TRA and SEG.

The results of this work are compared against the best performing available methods for each dataset. For the *Fluo-N2DH-GOWT1-01* dataset, we compare our method with the two tracking-by-detection [*CPN* (Akram et al., 2017) *KTH* (Magnusson and Jaldén, 2012)] and one joint cell detection and tracking [*BLOB* (Akram et al., 2016)] methods as the baselines. For the *PhC-C2DH-U373* dataset, we use the best performing *U-Net* (Ronneberger et al., 2015) and a graph cuts and model evolution-based tracking method (*GC-ME*) (Bensch and Ronneberger, 2015) as the baselines. For the *Fluo-N2DH-SIM+* dataset, we use a Siamese matching-based tracker based on the U-Net segmentation results (*U-Net-S*) (Gupta et al., 2019) as the baseline.

Table 3 lists the TRA, SEG and OPT scores for all methods over three datasets. It can be observed that our method yields the best TRA, SEG and OPT over the *Fluo-N2DH-GOWT1-01* sequence. However, our method has a lower TRA score over the *Fluo-N2DH-GOWT1-02* sequence. One reason for the lower TRA score of our method is that the *Fluo-N2DH-GOWT1-02* sequence has multiple cell events, including mitosis, apoptosis and cell fusion. Our method does not consider the complex process of cell differentiation.

For the *PhC-C2DH-U373* sequences, the *U-Net* tracking method uses the cell segmentation model trained from two sequences. Therefore, the SEG score of *U-Net* is the best among all algorithms over the *PhC-C2DH-U373* sequences. However, even with that advantage, our method still obtains a higher TRA score on the *PhC-C2DH-U373-01* sequence. *U-Net* produces very accurate cell segmentation masks on *PhC-C2DH-U373* sequences, but for the data association step, it often fails to associate correctly. The reason is that *U-Net* utilizes the greedy search method to link the cell segmentation between frames.

For *Fluo-N2DH-SIM+* sequences, our method has similar performance with *U-Net-S*. Once the cells have been detected, our method for cell tracking is able to achieve high overall accuracy in linking the cells between frames.

5. CONCLUSION

In this paper, we presented a solution to the problem of data association in cell tracking using the deep reinforcement learning. We formulated the data association problem into a linear assignment problem and then proposed a deep reinforcement learning framework which utilizes a residual CNN neural network. In simulation results, we compare the proposed method with other state-of-the-art approaches on various cell tracking datasets, and the results show that the proposed method achieves better comprehensive performance. Thus, our method likely has applications in the field of biomedical engineering. There are also some limitations of our tracking method that leave room for improvement. In future research, we plan to improve the data association method to deal with one-to-many and many-to-one association problems.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

LZ and JW substantially contributed to the conception and design of the study. XS analyzed and interpreted the data. LZ, JW, and JZ drafted the article.

FUNDING

This work was supported by the Tou-Yan Innovation Team Program of Heilongjiang Province (2019-15), the National Natural Science Foundation of China [61871160], and the Heilongjiang Province Postdoctoral Fund [LBH-TZ20].

ACKNOWLEDGMENTS

We thank all of the reviewers for their valuable comments and suggestions.

REFERENCES

Akram, S. U., Kannala, J., Eklund, L., and Heikkilä, J. (2016). "Joint cell segmentation and tracking using cell proposals," in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)* (Prague: IEEE), 920–924.

Akram, S. U., Kannala, J., Eklund, L., and Heikkilä, J. (2017). Cell tracking via proposal generation and selection. *arXiv [preprint] arXiv*:1705.03386.

Amat, F., Lemon, W., Mossing, D. P., McDole, K., Wan, Y., Branson, K., et al. (2014). Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nat. Methods* 11, 951–958. doi: 10.1038/nmeth.3036

Bello, I., Pham, H., Le, Q. V., Norouzi, M., and Bengio, S. (2016). Neural combinatorial optimization with reinforcement learning. *arXiv [preprint] arXiv*:1611.09940.

Bensch, R., and Ronneberger, O. (2015). "Cell segmentation and tracking in phase contrast images using graph cut with asymmetric boundary costs," in *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)* (New York, NY: IEEE), 1220–1223.

Bertsekas, D. P. (1992). Auction algorithms for network flow problems: a tutorial introduction. *Comput. Optim. Appl.* 1, 7–66. doi: 10.1007/BF00247653

Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/156652321904191022113307

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutmdisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* 48, D554–D560. doi: 10.1093/nar/gkz843

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* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

Chenouard, N., Bloch, I., and Olivo-Marin, J. (2013). Multiple hypothesis tracking for cluttered biological image sequences. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 2736–3750. doi: 10.1109/TPAMI.2013.97

Dufour, A., Thibeaux, R., Labruyere, E., Guillen, N., and Olivo-Marin, J.-C. (2010). 3-d active meshes: fast discrete deformable models for cell tracking in 3-d time-lapse microscopy. *IEEE Trans. Image Process.* 20, 1925–1937. doi: 10.1109/TIP.2010.2099125

Dzyubachyk, O., Van Cappellen, W. A., Essers, J., Niessen, W. J., and Meijering, E. (2010). Advanced level-set-based cell tracking in time-lapse fluorescence microscopy. *IEEE Trans. Med. Imaging* 29, 852–867. doi: 10.1109/TMI.2009.2038693

Emami, P., and Ranka, S. (2018). Learning permutations with sinkhorn policy gradient. *arXiv [preprint] arXiv*:1805.07010.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., et al. (2019). U-Net: deep learning for cell counting, detection, and morphometry. *Nat. Methods* 16, 67–70. doi: 10.1038/s41592-018-0261-2

Fortmann, T. E., Bar-Shalom, Y., and Scheffe, M. (1980). "Multi-target tracking using joint probabilistic data association," in *1980 19th IEEE Conference on Decision and Control including the Symposium on Adaptive Processes* (Albuquerque, NM: IEEE), 807–812.

Fu, H., Tang, H., Hao, J., Lei, Z., Chen, Y., and Fan, C. (2019). Deep multi-agent reinforcement learning with discrete-continuous hybrid action spaces. *arXiv [preprint] arXiv*:1903.04959.

Gupta, D. K., de Bruijn, N., Panteli, A., and Gavves, E. (2019). Tracking-assisted segmentation of biological cells. *arXiv [preprint] arXiv*:1910.08735.

Hamid Rezatofighi, S., Milan, A., Zhang, Z., Shi, Q., Dick, A., and Reid, I. (2015). "Joint probabilistic data association revisited," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago), 3047–3055.

Han, J., Han, X., Kong, Q., and Cheng, L. (2019). pssubpathway: a software package for flexible identification of phenotype-specific subpathways in cancer progression. *Bioinformatics*. doi: 10.1093/bioinformatics/btz894. [Epub ahead of print].

Hu, Y., Zhao, T., Zang, T., Zhang, Y., and Cheng, L. (2019). Identification of Alzheimer's disease-related genes based on data integration method. *Front. Genet.* 9:703. doi: 10.3389/fgene.2018.00703

Ioffe, S., and Szegedy, C. (2015). "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning* (Lille), 448–456.

Khalil, E., Dai, H., Zhang, Y., Dilkina, B., and Song, L. (2017). "Learning combinatorial optimization algorithms over graphs," in *Advances in Neural Information Processing Systems* (Long Beach, CA), 6348–6358.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, (Lake Tahoe, CA), 1097–1105.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Res. Logist. Q.* 2, 83–97. doi: 10.1002/nav.3800020109

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.-W., and Heng, P.-A. (2018). H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Trans. Med. Imaging* 37, 2663–2674. doi: 10.1109/TMI.2018.2845918

Magnusson, K. E. G., and Jaldén, J. (2012). "A batch algorithm using iterative application of the viterbi algorithm to track cells and construct cell lineages," in *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)* (Barcelona), 382–385.

Maška, M., Ulman, V., Svoboda, D., Matula, P., Matula, P., Ederra, C., et al. (2014). A benchmark for comparison of cell tracking algorithms. *Bioinformatics* 30, 1609–1617. doi: 10.1093/bioinformatics/btu080

Milan, A., Rezatofighi, S. H., Garg, R., Dick, A., and Reid, I. (2017). "Data-driven approximations to np-hard problems," in *Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, CA).

Nazari, M., Oroojlooy, A., Snyder, L., and Takác, M. (2018). "Reinforcement learning for solving the vehicle routing problem," in *Advances in Neural Information Processing Systems* (Montreal, QC), 9839–9849.

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019a). A learning-based framework for mirna-disease association identification using neural networks. *Bioinformatics* 35, 4364–4371. doi: 10.1101/276048

Peng, J., Lu, J., Hoh, D., Dina, A. S., Shang, X., Kramer, D. M., et al. (2020). Identifying emerging phenomenon in long temporal phenotyping experiments. *Bioinformatics* 36, 568–577. doi: 10.1093/bioinformatics/btz559

Peng, J., Wang, X., and Shang, X. (2019b). Combining gene ontology with deep neural networks to enhance the clustering of single cell rna-seq data. *BMC Bioinform.* 20:284. doi: 10.1186/s12859-019-2769-6

Peng, J., Zhu, L., Wang, Y., and Chen, J. (2019c). Mining relationships among multiple entities in biological networks. *IEEE ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/TCBB.2019.2904965. [Epub ahead of print].

Ristic, B., Vo, B.-N., Clark, D., and Vo, B.-T. (2011). A metric for performance evaluation of multi-target tracking algorithms. *IEEE Trans. Signal Process.* 59, 3452–3457. doi: 10.1109/TSP.2011.2140111

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Munich), 234–241.

Vinyals, O., Fortunato, M., and Jaitly, N. (2015). "Pointer networks," in *Advances in Neural Information Processing Systems* (Munich), 2692–2700.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learn.* 8, 229–256. doi: 10.1007/BF00992696

Xing, F., and Yang, L. (2016). Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE Rev. Biomed. Eng.* 9, 234–263. doi: 10.1109/RBME.2016.2515127

Xu, Y., Ban, Y., Alameda-Pineda, X., and Horaud, R. (2019). Deepmot: a differentiable framework for training multiple object trackers. *arXiv [preprint] arXiv*:1906.06618.

Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a gaussian denoiser: residual learning of deep cnn for image denoising. *IEEE Trans. Image Process.* 26, 3142–3155. doi: 10.1109/TIP.2017.2662206

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 Bioinform.* 21:43. doi: 10.1186/s12859-020-3388-y

Zimmer, C., Labruyere, E., Meas-Yedid, V., Guillén, N., and Olivo-Marin, J.-C. (2002). Segmentation and tracking of migrating cells in videomicroscopy with parametric active contours: a tool for cell-based drug testing. *IEEE Trans. Med. imaging* 21, 1212–1221. doi: 10.1109/TMI.2002.806292

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 15 April 2020
doi: 10.3389/fbioe.2020.00324

Characterization and Classification of Electrophysiological Signals Represented as Visibility Graphs Using the Maxclique Graph

Erika Elizabeth Rodriguez-Torres[1], *Ulises Paredes-Hernandez*[1], *Enrique Vazquez-Mendoza*[2], *Margarita Tetlalmatzi-Montiel*[1], *Consuelo Morgado-Valle*[3], *Luis Beltran-Parrazal*[3] *and Rafael Villarroel-Flores*[1]*

[1] *Área Académica de Matemáticas y Física, Universidad Autónoma del Estado de Hidalgo, Pachuca, Mexico,* [2] *Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Mexico City, Mexico,* [3] *Centro de Investigaciones Cerebrales, Dirección General de Investigaciones, Universidad Veracruzana, Xalapa, Mexico*

OPEN ACCESS

Edited by:
Jiajie Peng,
Northwestern Polytechnical University, China

Reviewed by:
Liang Zhao,
Hubei University of Medicine, China
Fuhai Li,
Washington University in St. Louis, United States

****Correspondence:***
Rafael Villarroel-Flores
rafaelv@uaeh.edu.mx

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

Received: *17 January 2020*
Accepted: *24 March 2020*
Published: *15 April 2020*

Citation:
Rodriguez-Torres EE, Paredes-Hernandez U, Vazquez-Mendoza E, Tetlalmatzi-Montiel M, Morgado-Valle C, Beltran-Parrazal L and Villarroel-Flores R (2020) Characterization and Classification of Electrophysiological Signals Represented as Visibility Graphs Using the Maxclique Graph. Front. Bioeng. Biotechnol. 8:324. doi: 10.3389/fbioe.2020.00324

Detection, characterization and classification of patterns within time series from electrophysiological signals have been a challenge for neuroscientists due to their complexity and variability. Here, we aimed to use graph theory to characterize and classify waveforms within biological signals using maxcliques as a feature for a deep learning method. We implemented a compact and easy to visualize algorithm and interface in Python. This software uses time series as input. We applied the maxclique graph operator in order to obtain further graph parameters. We extracted features of the time series by processing all graph parameters through K-means, one of the simplest unsupervised machine learning algorithms. As proof of principle, we analyzed integrated electrical activity of XII nerve to identify waveforms. Our results show that the use of maxcliques allows identification of two distinct types of waveforms that match expert classification. We propose that our method can be a useful tool to characterize and classify other electrophysiological signals in a short time and objectively. Reducing the classification time improves efficiency for further analysis in order to compare between treatments or conditions, e.g., pharmacological trials, injuries, or neurodegenerative diseases.

Keywords: visibility graphs, graph theory, maxcliques, electrophysiological signals, deep learning, pre-Bötzinger complex, XII nerve, sigh

1. INTRODUCTION

To understand brain functioning neuroscientists use electrophysiological techniques (e.g., macropatch and patch-clamp recordings) to assess activity of neurons. Whereas, sharp-electrode and patch-clamp techniques are used to record the activity of a single neuron, extracellular field recordings and macropatch techniques allow recording the activity of many neurons within a population. Macropatch suction electrodes are widely used to record motor nerve activity. The inspiratory phase of the respiratory rhythm is generated in the pre-Bötzinger complex (pre-BötC), a neuronal network in the ventrolateral medulla. In an *in vitro* preparation containing the pre-BötC, inspiratory-related motor output can be recorded from the XII nerve. Nerve activity is integrated and used to classify and characterize the inspiratory-related burst. Frequently, researchers made this manually; however, this is a time-consuming and very subjective task. Spike sorting, traditionally, is made measuring properties of the waveform (e.g., peak latency, spike half-width, amplitude), determining which of these properties or features are relevant (e.g., principal component analysis) and performing cluster analysis (Rey et al., 2015). In the literature, one can find several algorithms employed for the spike sorting, following different steps and approaches. For instance, some of the techniques are based on wavelets, or combinations of wavelets and different approaches of principal components, for a review one can see Rey et al. (2015) and Lefebvrea et al. (2016). A more recent approach is based on the shape, phase, and distribution features of each spike and a clustering algorithm based on k-means (Caro-Martín et al., 2018). Along with the spike sorting algorithms, methods that validate them are necessary (Einevoll et al., 2012). However, in the field of respiratory rhythm there are not automated methods for the identification of sighs. In plethysmographic recordings, sighs are identified visually by the expert. In electrophysiological recordings from reduced preparations *in vitro*, the criteria for defining a sigh are determined by the researcher and therefore vary between research groups. Some groups consider the amplitude as a relevant parameter (Lieske et al., 2000; Lieske and Ramirez, 2006a,b; Ruangkittisakul et al., 2008); others the presence of biphasic burst (Kam et al., 2013; Li P. et al., 2016). Here, based on our analysis we propose to use graph theory to characterize and classify waveforms within biological signals using maxcliques as a feature for a deep learning method.

A network or graph is one of the most intuitive, explicit and clear representation of a complex system. Such graphs consist of *nodes* and *links* representing the participating elements and the interactions among them. Therefore, graphs characterize the structure of complex systems and how its elements interact. That is, they can reflect the dynamics or functions of the complex system if states and transitions are represented by nodes and links, respectively (Gershenson and Niazi, 2013). If one can understand the relationship between structure and function, then the characterization and classification of complex systems can be studied further.

In this work, each time series is associated with a simple graph called *visibility graph*, as defined in Lacasa et al. (2008) and studied in Lacasa and Flanagan (2015). As a remark, this graphs inherits either the periodicity or the randomness of the original time series. Even more, fractal time series are transformed into scale-free graphs (Lacasa et al., 2008). We aimed to use Graph Theory to characterize and classify visibility graphs using maxcliques as a feature for a deep learning method. Here, we analyzed *in vitro* recordings of XII nerve inspiratory activity to classify sighs and non-sighs waveforms. The visibility graph of a non-sighs shows a simpler structure than sighs. The interest of the authors in sighs is its relevance in preventing lung collapses.

Recently, visibility graphs have been employed to analyze the resulting time series from physiological data as in Hou et al. (2016), Jiang et al. (2013), and Shao (2010), in the analysis of complex networks for cardiorespiratory interactions (Long, 2015) or a modified visibility graph for the suicidal tendency (Bhaduri et al., 2016). However, in these works, the concept of maxcliques from graph theory was not implemented in the characterization and classification of waveforms.

We present a graphical interface, written in Python, that helped in the process of constructing the visibility graph from the time series and determining several parameters of the resulting graph. Python is an open source interpreted programming language. The simplicity of Python syntax makes its code readable and understandable, facilitating its learning. There are several Python libraries, such as pandas, numpy, SciPy, and others that allow the user to process and analyze data easily and quickly. Although a Python package with the implementation of the algorithm described in Lacasa et al. (2008) can be found in García-Herrera (2015), the one showed in this work is more convenient and, as a consequence, easier to visualize. The interface employs two Python libraries: *NetworkX* (https://networkx.github.io/), where several algorithms of Graph Theory have been already implemented, and *matplotlib* (https://matplotlib.org/) for the graphs.

We claim that several aspects of a time series can be deduced from certain parameters of the associated visibility graphs. In this work in particular, that was the case with the maximum degree, the clique number, and the number of cliques. They allowed to tell sighs from non-sighs in the time series obtained in the waveforms from *in vitro* recordings of XII nerve inspiratory activity.

2. MATERIALS AND METHODS

2.1. Graph Theory

As a mathematical concept, a *graph* G is composed by a set of points denoted with $V(G)$, and a set denoted by $E(G)$ whose elements are unordered pairs of elements of $V(G)$. The elements of $V(G)$ are called *vertices* or *nodes*, and the elements of $E(G)$ are called *edges* or *links*. The number of vertices in a graph G is called the *order* of the graph G and is denoted by $|G|$. If the nodes v_1, v_2 are such that $\{v_1, v_2\} \in E(G)$, we say that the vertices v_1, v_2 are *adjacent*, and we denote that by $v_1 \sim v_2$. Given a vertex v, the number of vertices adjacent to v is called the *degree* of v. As a starting point for the concepts from graph theory, we recommend Harary (1969) and McKee and McMorris (1999).

Lacasa et al. (2008) associated for the first time a graph to a given time series by a procedure they called the *visibility algorithm*, which we now describe. Given a time series with data pairs $\{(t_a, y_a)\}$, they obtain the *visibility graph* of the time series as the graph where the vertex set is the set of all data pairs, and define that the pairs $(t_a, y_a), (t_b, y_b)$ are adjacent whenever we have:

$$y_c < y_b + (y_a - y_b)\frac{t_b - t_c}{t_b - t_a}, \qquad (1)$$

for all data pairs (t_c, y_c) with $t_a < t_c < t_b$. The geometric visualization of this condition is shown in **Figures 1A,B**.

Given a graph G, a *maxclique* C is a subset of its nodes such that every two nodes in C are adjacent, and there is no vertex in G not in C that is adjacent to all the vertices of C. We follow McKee and McMorris (1999) in the use of the term "*maxclique*," in order to avoid the ambiguity found in the literature on the meaning of the word "clique."

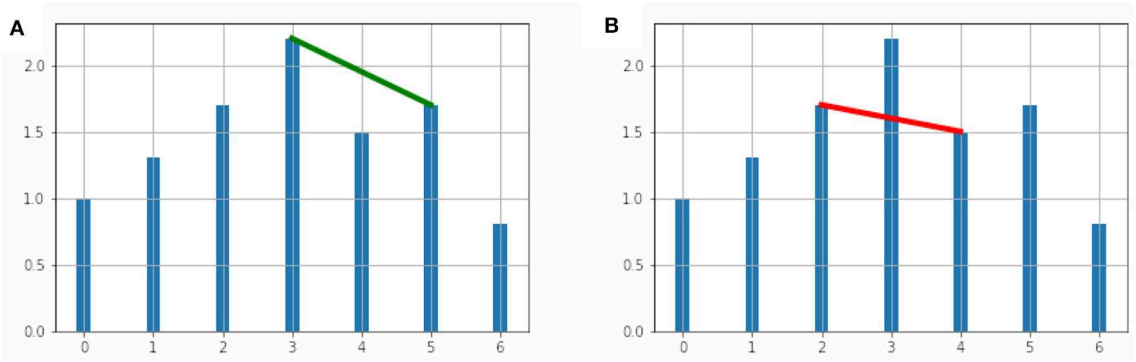


FIGURE 1 | (A) The vertices (3, 2.2) and (5, 1.7) are adjacent in the visibility graph. **(B)** The vertices (2, 1.7) and (4, 1.5) are not adjacent in the visibility graph.

The *maxclique graph* is the graph that has as vertices the maxcliques of G, and where two maxcliques C_1, C_2 are adjacent whenever there is at least a vertex of G that belongs to both C_1 and C_2. As a reference for maxclique graphs we mention Szwarcfiter (2003). The maxclique graph of G will be denoted as $K(G)$. It follows then that a graph G has $|K(G)|$ maxcliques.

We now define further parameters of a graph G that will be considered in this work:

- **Maximum degree:** This is denoted by $\Delta(G)$, and is the maximum among all degrees of vertices of G.
- **Clique number:** This is the number of elements of the largest maxclique of G. It is denoted by $\omega(G)$.

As an example of the concepts described here, consider the time series given by

$$[(0,1),(1,1.3),(2,1.7),(3,2.2),(4,1.5),(5,1.7),(6,0.8)] \qquad (2)$$

In **Figure 1A**, we show two vertices adjacent in the visibility graph and in **Figure 1B** we show two non-adjacent vertices.

The visibility graph G of this time series is shown in **Figure 2A**. The vertex with maximum degree is the vertex 3, and its degree is 5, and so $\Delta(G) = 5$. The graph G has three maxcliques, so that $|K(G)| = 3$. The three maxcliques are: $\{0,1,2,3\}$, $\{3,4,5\}$ and $\{5,6\}$, with 2, 3, and 4 vertices each. Since the greatest maxclique of G has four elements, we obtain that $\omega(G) = 4$. Finally, note that the second clique intersects each of the other two, and the first and the third do not intersect. So the graph $K(G)$ has three vertices, as it is shown in **Figure 2B**.

2.2. Interface to NetworkX in Python

The graph algorithms described in section 2.1 were implemented in a Python interface using the PyQT5 library. The supported files are of one or two columns (*.txt* or *.csv* format). One can select the percentage of sampling frequency (recommended for large signals), visibility graph style and an option to create the maxclique graph (**Figure 3**).

With the signals loaded and setting the parameters, the visibility graph G is created. The visibility graph G, the maxclique graph $K(G)$ (in format *.png*) and the parameters that are calculated in each algorithm (in format *.txt*) are saved to the signals folder. The interface has also a tool to segment or auto segment signals (**Figure 3**). The button *Start segmentation* enables a bar to select a region in the signal loaded. For auto segment signals the user must introduce an upper threshold, lower threshold, segment width and distance between spikes. In **Figure 3**, we show a schematic representation of the process to classify electrophysiological signals using maxclique graph parameters.

2.3. Experiment

The pre-BötC (pre-Bötzinger complex) is a heterogeneous network of interneurons. In rats this contains a population of ∼1,000 neurons. In synaptic interactions between pre-BötC neurons each neuron produces inspiratory rhythmic activity in the form of synchronous depolarization of 10–20 mV with a duration of 0.3–0.8 s and with waveforms called inspiratory bursts. In addition to its role in the generation of the respiratory rhythm, pre-BötC is essential for the formation of the respiratory pattern. The protocol for obtaining respiratory rhythm records consists in sectioning the brain stem of neonatal rats under the microscope until the ambiguous nucleus and the inferior olive appear (**Figure 4**).

We describe the electrophysiology in brief. Coronal sections were cut (500–600μm) and the rhythmic activity was recorded from the roots of the XII nerve (XIIn). Then the signal of the XIIn motor neurons excited by pre-BötC neurons is transmitted, obtaining the rhythmic activity of the XIIn (**Figure 4**). Once baseline activity was established, drug application was performed in the slice bath. In each experiment, two time series were obtained, the first corresponding to control respiratory activity (**Figure 5A**) and the second when the pre-BötC slice was exposed to bombesin (**Figure 5B**). In **Figure 4** we can observe two components: normal respiratory rhythm (non-sigh) and long inspirations known as sighs. Sighs are biphasic inspiratory bursts. However, sighs can fulfill important regulatory functions. More specifically, a sigh acts as a general restorative of the respiratory system (Patroniti et al., 2002). In general, the pre-BötC generates a normal inspiratory burst every 7–8 s (non-sigh) and every 30–40 s generates a disturbance called a sigh. For more information on how the experiment was done see Munoz-Ortiz et al. (2016).

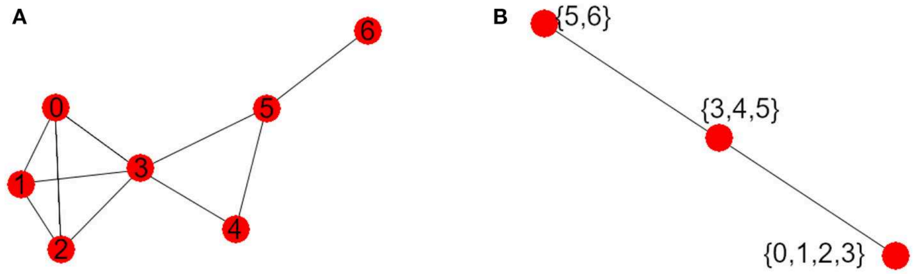
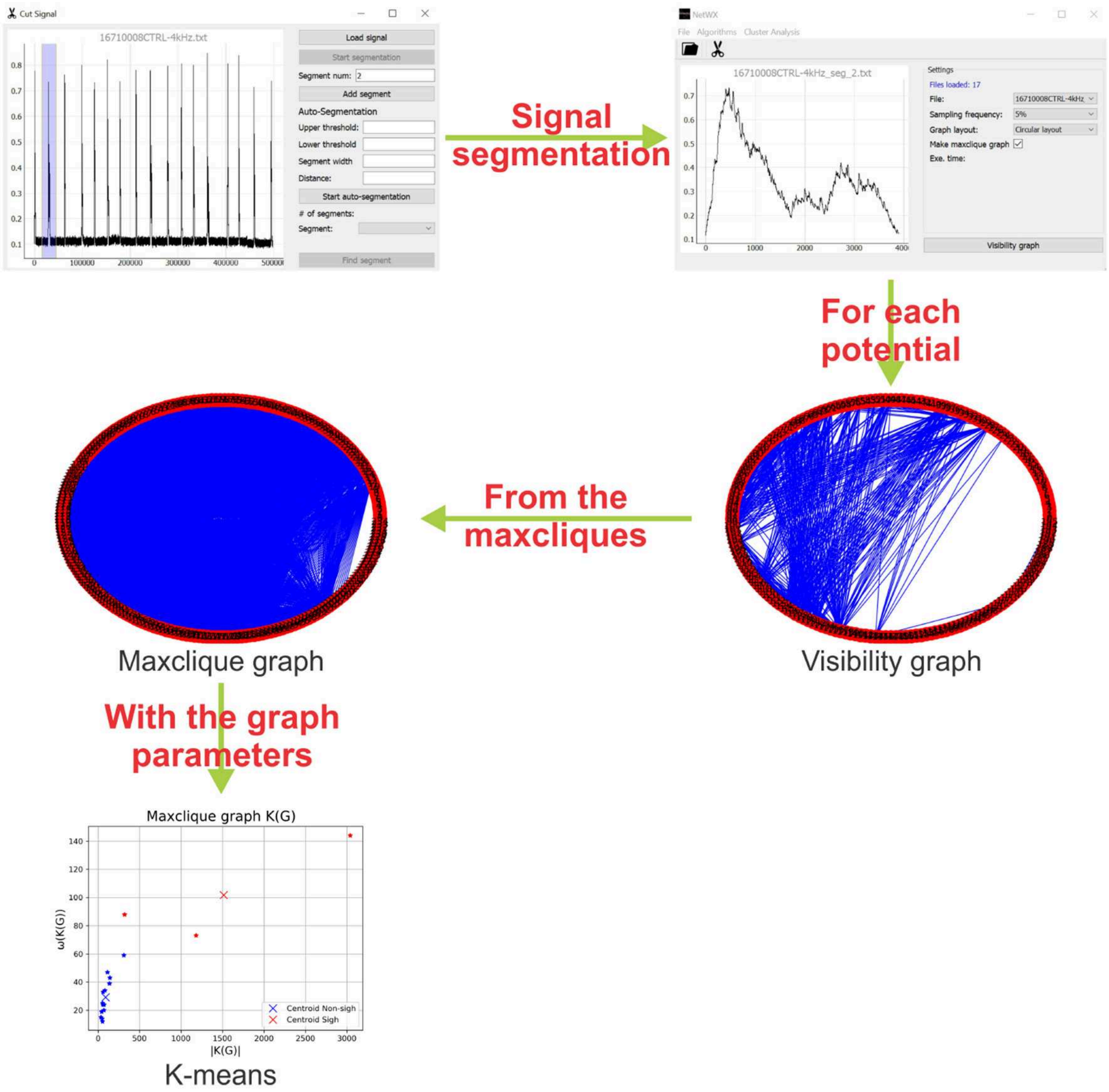


FIGURE 2 | (A) Visibility graph *G*. **(B)** Maxclique graph *K*(*G*).



Segment num: 2

Auto-Segmentation

Upper threshold:

Lower threshold

Segment width

Distance:

# of segments:

Segment:

Settings

Files loaded: 17

File: 16710008CTRL-4kHz

Sampling frequency: 5%

Graph layout: Circular layout

Make maxclique graph ☑

Exe. time:

FIGURE 3 | Schematic representation of the methodology. First, the interface identify and segment each potential of the electrophysiology recording. Then, a visibility graph is created for each potential, for large signals a reduction of the sampling frequency is recommended. After that, from the maxcliques determined of the visibility graph, the maxclique graph is created and its parameters are estimated. Finally, a K-means clustering is performed on the maxclique graph parameters. In this work, the result is a classification of the potentials as sighs or non-sighs.

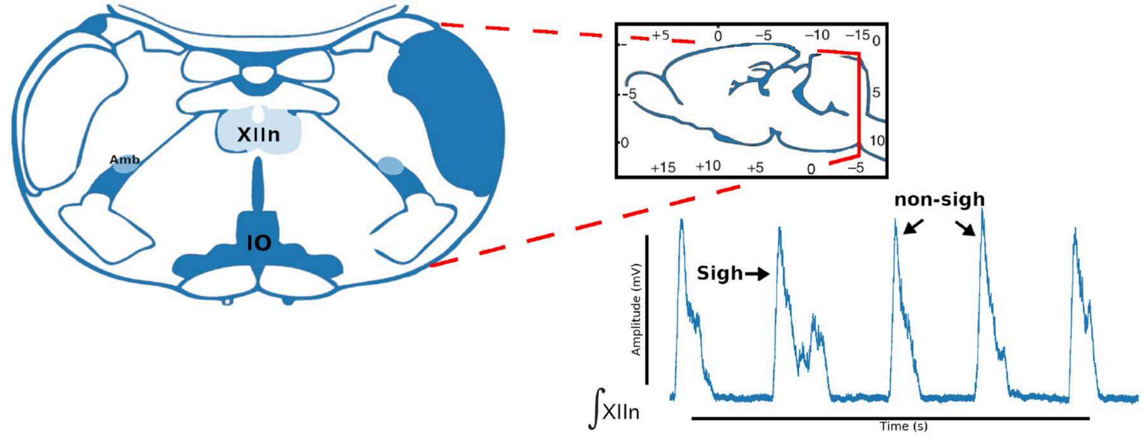


FIGURE 4 | Coronal brainstem section that presents the anatomical marks to locate the pre-Bötzinger complex. Representative integrated activity of the XII nerve showing characteristic waveform of sigh and non-sigh.

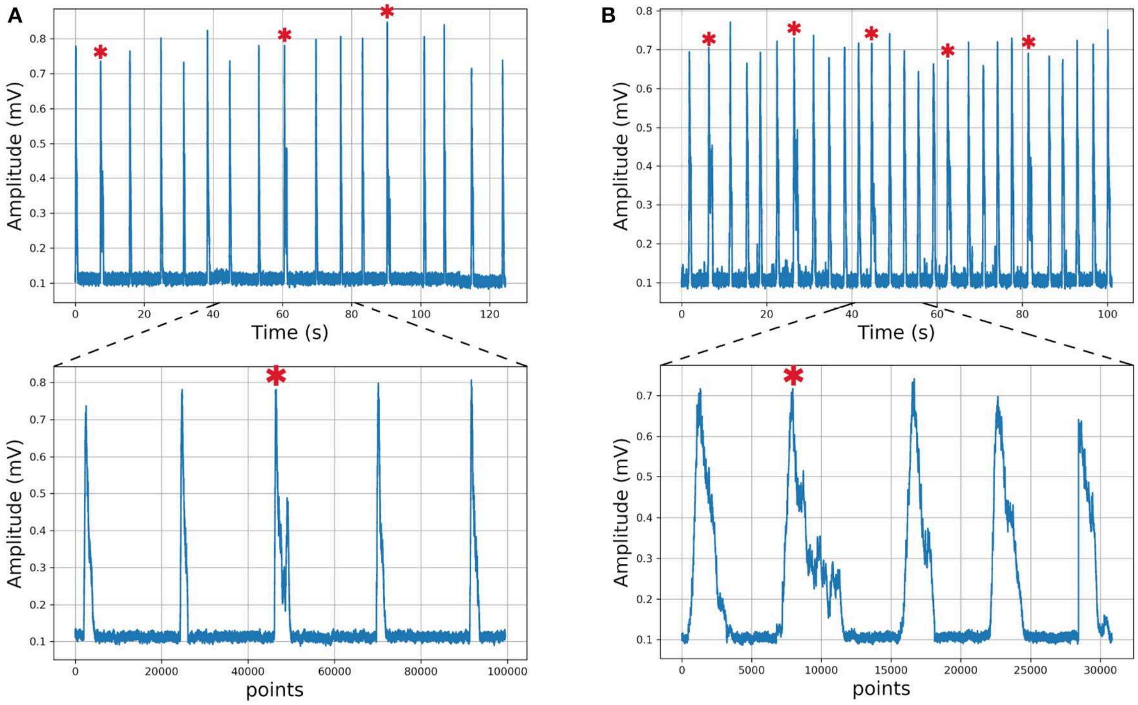


FIGURE 5 | Respiratory rhythm. **(A)** Control record and **(B)** Bombesin record. In asterisk (*) are shown sighs.

2.4. Statistical Analysis

Given that data did not follow a normal distribution (Shapiro-Wilk test), the Box-Cox transformation was used, as implemented in the R package fpp. With that, $\lambda = -0.475$ was determined as the value that maximized the log-likelihood function and yield the best transformation to normality. Some parameters of the visibility graph G associated to the time series ($\Delta(G)$, $\omega(G)$, and $|G|$) and of the maxclique graph $K(G)$ ($\Delta(K(G))$, $\omega(K(G))$, and $|K(G)|$) were compared between sigh and non-sigh using a two-way ANOVA, followed by a Bonferroni's multiple comparisons test. To evaluate the performance of classification based on visibility or maxclique graph parameters, we compared the number of sighs and non-sighs identified by the three classifiers performing a chi-squared test and a pairwise comparison with Bonferroni's correction. Then, we compared both classifications vs. the classifications based on an expert determining the number of successes and failures of each classification. Then, we performed a McNemar's test. Two-way ANOVA was performed in GraphPad Prism (v. 6.00, GraphPad Software, Ca, USA). Box-Cox transformation,

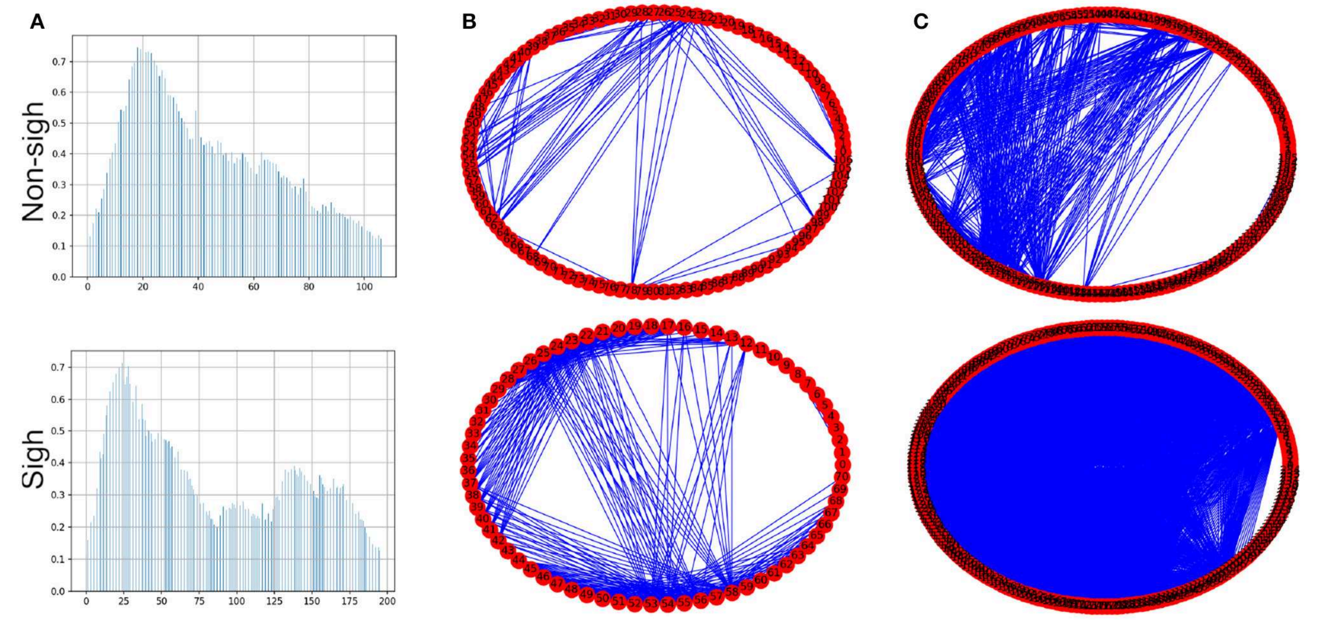


FIGURE 6 | (A) Inspiratory burst recordings with its sampling frequency reduced to 5% of non-sigh and sigh time series from respiratory rhythm *in vitro* recordings. **(B)** Circle visibility graphs constructed from time series shown in **(A)**, for non-sigh G_1 and sigh G_2, respectively. The non-sigh circle visibility graph may appear to show fewer connections than the sigh one. **(C)** Maxclique graphs for non-sigh and sigh. In this case, it is apparent that the number of connections (that is, edges) is much larger for sigh ($K(G_2)$) than non-sigh ($K(G_1)$).

chi-squared and McNemar's tests were performed in R (v. 3.6.1—"Action of the Toes"). Significant differences were considered at $P \leq 0.05$. Data is showed as mean ± S.E.M.

3. RESULTS AND DISCUSSION

3.1. Results

As an example of usefulness, we employed *in vitro* recordings from XII nerve respiratory rhythm activity of rats in order to obtained time series describing burst amplitude. In this time series, we can differentiate between sigh and non-sigh waveforms, which were recorded in control and bombesin conditions. First of all, we wanted to determine if the classification between sigh and non-sighs was correct, independently of the experimental condition. To achieve the latter, we used a short time series composed of 17 potentials of control recording (**Figure 5A**), which were previously classified by an expert in 14 non-sighs and 3 sighs. Likewise, we used a bombesin recording composed of 27 inspiratory bursts (**Figure 5B**), 22 non-sighs and 5 sighs.

To create the visibility graphs the sampling frequency of each inspiratory burst was reduced to 5%, in both sighs and non-sighs waveforms (**Figure 6A**). The visibility graph of the non-sigh and sigh will be denoted by G_1 and G_2 (**Figure 6B**), respectively. Now, for each visibility graph, G_1 and G_2, we constructed their maxclique graphs, denoted as $K(G_1)$ and $K(G_2)$ (**Figure 6C**), respectively. From both graphs, we calculated their maximum degree $\Delta(G_1), \Delta(G_2)$, clique number $\omega(G_1), \omega(G_2)$, and number of cliques $|K(G_1)|, |K(G_2)|$.

Classification of waveforms was performed using K-means clustering analysis with the three graph parameters [clique number: $\omega(G)$, number of maxcliques: $|K(G)|$, and maximum degree: $\Delta(G)$] of each graph [visibility, G and maxclique, $K(G)$], comparing in pairs. Of these parameters, we observed that clique number and number of maxcliques classify better both waveforms, independently of experimental condition.

K-means clustering analysis with visibility graph parameters resulted in 13 non-sighs and 4 sighs in the control recording, and 21 non-sighs and 6 sighs in the bombesin recording (**Figure 7A**). In contrast, K-means clustering analysis with maxclique graph parameters resulted in 14 non-sighs and 3 sighs in the control recording, and 21 non-sighs and 6 sighs in the bombesin recording (**Figure 7B**). In **Figures 7C,D**, we show the inspiratory bursts as classified by the maxclique graph parameters, in both control and bombesin condition, which shows that this classification is accurate. Altogether, these results show that the clique number and the number of max cliques of the maxclique graph have a better classifying waveforms performance.

In the previous description, we used an expert delimited and classified waveforms. However, we created an automatic segmentation and performed the same analysis to evaluate if the classification remained consistent. In this case, we used a time series composed of 39 and 99 inspiratory bursts, recorded in control and bombesin conditions, respectively (**Figure 8**).

The automatic segmentation identified every single burst. Classification based on visibility graph parameters resulted in 10 sighs and 29 non-sighs, in the control recording and 42

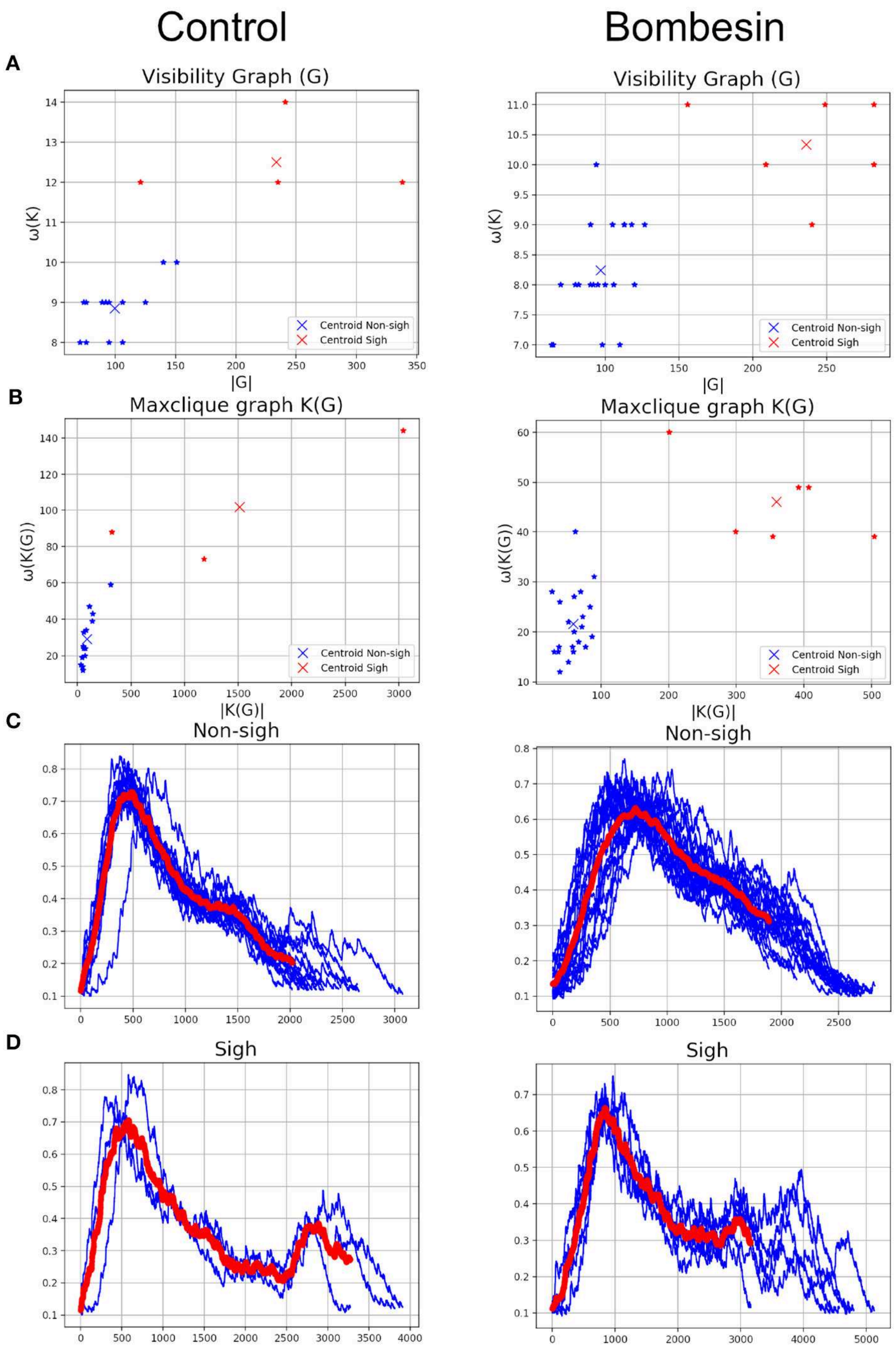


FIGURE 7 | Classification according with clique's parameters. Control on first column and bombesin recording on second column. **(A)** Visibility graph and **(B)** Maxclique graph K-means cluster analysis, **(C)** non-sighs, and **(D)** sighs inspiratory bursts. In red are shown the means of all inspiratory bursts classified with Maxclique graph.

sighs and 57 non-sighs, in the bombesin recording. On the other hand, classification based on maxclique graph parameters resulted in 5 sighs and 34 non-sighs, in the control recording and 6 sighs and 93 non-sighs, in the bombesin recording. The inspiratory bursts as classified by the maxclique graph parameters, in both control and bombesin condition, are shown in **Figures 9A,B**, respectively. This suggests that automatic segmentation properly identifies potentials, regardless of the waveform and experimental condition.

On previous results, we observed that maxclique parameters seem to classify more accurately between both waveforms. Thus, in order to determine if this is robust enough, we performed the

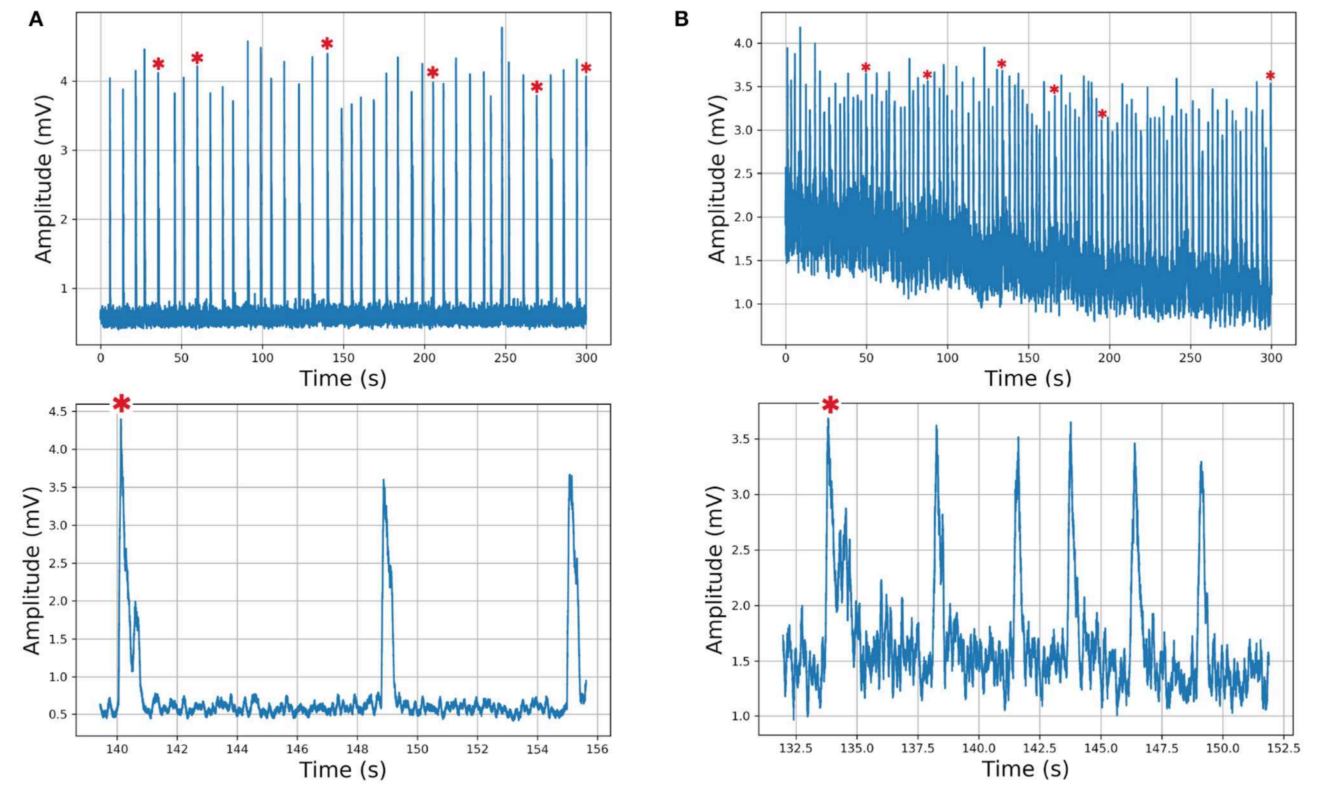


FIGURE 8 | Respiratory rhythm recording and corresponding visibility graphs for **(A)** control with 39 and **(B)** bombesin with 99 inspiration burst. In asterisk (*) are shown sighs.

analysis with a larger time series, composed of 182 potentials. After K-means classification based on visibility or maxclique parameters we compared between putative sigh (20 potentials) and non-sigh (162 potentials) waveforms. Our analysis showed that both visibility and maxclique graph parameters show statistical difference between sigh (S) and non-sigh (NS) (graph parameters, $F_{5,1080} = 579.2$, $P < 0.0001$; waveform, $F_{1,1080} = 508.4$, $P < 0.0001$; graph parameter*waveform, $F_{5,1080} = 14.66$, $P < 0.0001$). Bonferroni's *post-hoc* test showed that G max degree (S, 44.10 ± 1.60 vs. NS, 26.62 ± 0.48; $P < 0.0001$; **Figure 10**), G clique num (S, 11.00 ± 0.27 vs. NS, 8.77 ± 0.10; $P < 0.0001$; **Figure 10**), G number of max cliques (S, 260.60 ± 14.01 vs. NS, 92.10 ± 2.21; $P < 0.0001$; **Figure 10**), $K(G)$ max degree (S, 137.40 ± 10.26 vs. NS, 43.10 ± 1.42; $P < 0.0001$; **Figure 10**), $K(G)$ clique num (S, 71.55 ± 6.69 vs. NS, 24.25 ± 0.85; $P < 0.0001$; **Figure 10**), and $K(G)$ number of max cliques (S, 779.40 ± 142.80 vs. NS, 66.32 ± 3.94; $P < 0.0001$; **Figure 10**) differed between sighs and non-sighs. This suggest that the groups generated by the K-means are authentic groups.

However, the above does not imply that these groups represent real sighs and non-sighs. First, we compared the number of sighs and non-sighs classified with both parameters and by an expert, which resulted to be different ($\chi^2 = 40.84$; $df = 2$; $P < 0.0001$). Our pairwise comparison analysis showed that classification based on visibility graph parameters (S, 61; NS, 162) is statistically different from that performed by the expert (S, 20; NS, 162; $\chi^2 = 25.41$; $df = 1$; $P < 0.0001$). In contrast, the classification based on maxclique graph parameter (S, 20; NS, 162) did not differ from the classification performed by the expert (S, 20; NS, 162; $\chi^2 = 0$; $df = 1$; $P = 1$).

Although our previous results showed that maxclique parameters identify the same number of sigh and non-sigh as the expert, we determined the number of successes and failures to assess the accuracy of classification. Our results showed that the classification based on maxclique graph parameters had six failures (three sighs and three non-sighs) and 176 success, whereas classification based on visibility graph parameters had 41 failures (all non-sighs) and 141 successes. McNemar's test showed that maxclique graph parameters were better to correctly identify and classify sigh and non-sigh waveforms (McNemar's $\chi^2 = 82.747$, $df = 1$, $P < 0.0001$). Altogether, these results indicate that the classification based on maxclique graph parameters is robust to classify accurately between sighs and non-sighs. Also, this suggests that these parameters should be used to classify other waveforms.

3.2. Discussion

In this paper, we have presented a classification and characterization of electrophysiological signals using graph parameters applied to visibility graphs and to the result of a graph operator called the maxclique graph, which is denoted by $K(G)$. The parameter $\omega(G)$, and the enumeration of the maximal

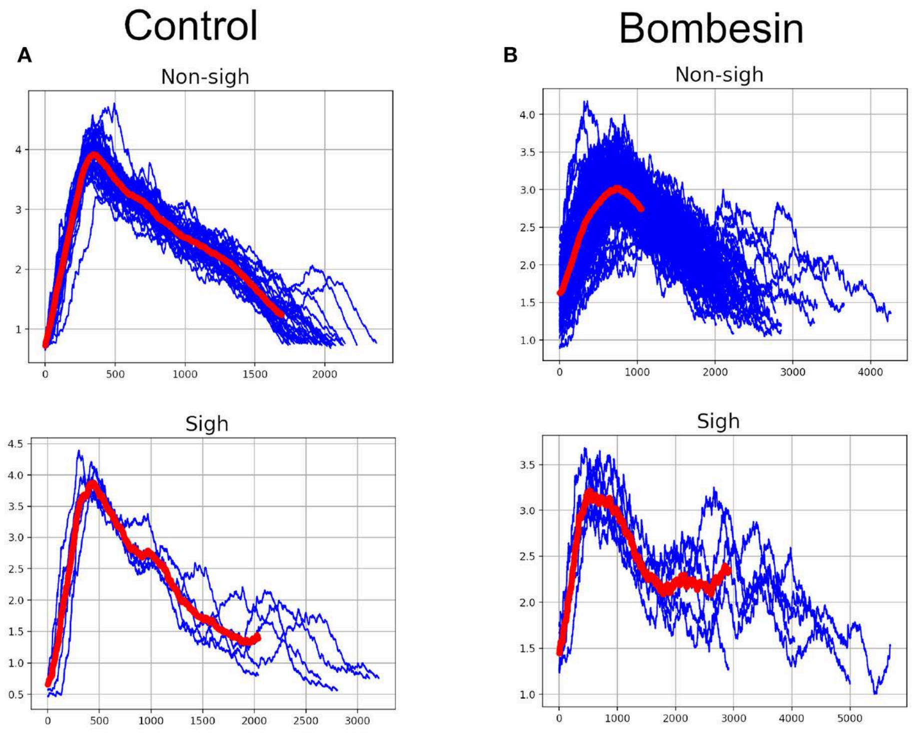


FIGURE 9 | Inspiratory burst classification according with clique's parameters in Maxclique graph. **(A)** Control with 39 inspiratory burst and **(B)** bombesin with 99 inspiratory burst. In red are shown the means of all inspiratory bursts classified.

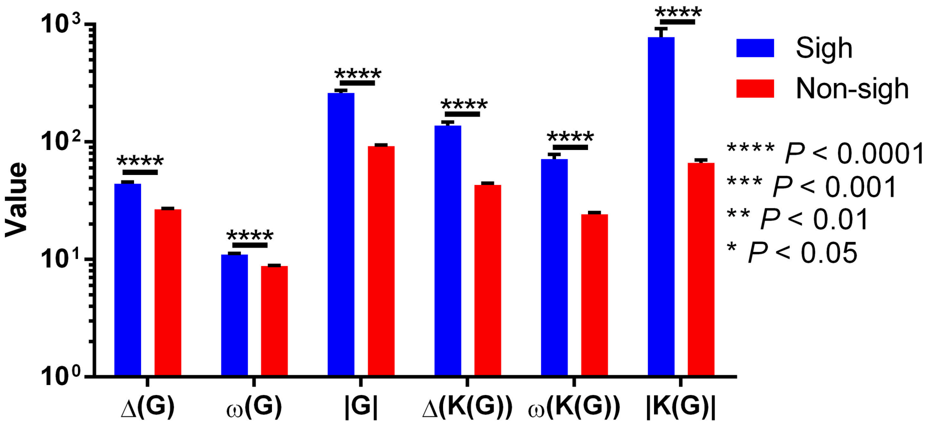


FIGURE 10 | Visibility ($\Delta(G)$, $\omega(G)$, and $|G|$) and maxclique ($\Delta(K(G))$, $\omega(K(G))$, and $|K(G)|$) graph parameters of sigh and non-sigh waveforms. Data is showed as mean ± S.E.M. Significant differences between non-sigh and sigh were determined using a two-way ANOVA, followed by Bonferroni's multiple comparisons. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$. Sighs, $n = 20$; non-sighs, $n = 162$.

cliques have already been considered in bioinformatics, for example in proteins and genes (see Tomita et al., 2011).

The maxclique graph operator has already been applied to Loop Quantum Gravity (for example see Requardt, 2000). To the best of our knowledge, this is the first time that the maxclique graph operator has been used in electrophysiological signals characterization. We have verified the usefulness of this operator for the task of identifying sighs and non-sighs waveforms, using *in vitro* recordings of XII nerve respiratory rhythm, and implementing in Python an interface using the algorithms described in this work. We think that it is apparent that this software can also be applied to characterize other electrophysiological recordings. The advantage of using cliques is the following:

- As shown in **Figures 7A,B**, the maxclique graph $K(G)$ allows us to differentiate sighs and non-sighs better than the visibility graph alone.

These results suggest that maxclique graph ($K(G)$), and particularly its parameters of number of cliques ($|K(G)|$), and clique number ($\omega(G)$) have a better performance characterizing and classifying these electrophysiological signals than a visual inspection of the time series. This is because if the time series has many small fluctuations (like sighs), then the visibility graph will have many small cliques, therefore, the graph parameter $|K(G)|$ will be relatively big and the parameter $\omega(G)$ will be relatively small. On the other hand, if in the time series there are few fluctuations and a value of the data much larger than the others, then there will be a big clique in the visibility graph, resulting in a small value of $|K(G)|$ and a larger value of $\omega(G)$ (like non-sighs). Sighs, and other breathing patterns are embedded within eupneic (normal breathing) signals. Unbiased detection of patterns is a challenge for electrophysiologist. The use of visibility graphs and maxclique analysis provides a tool for sorting waveforms probing a larger number of parameters, instead of commonly used peak amplitude, burst durations or the presence of biphasic shape.

Our statistical analysis showed that visibility and maxclique parameters differ between sigh and non-sigh. Nevertheless, we need further studies to correlate these parameters with their biological meaning to determine what these differences could mean in physiology. Allowing us to implement these graph parameters to compare between different conditions and treatments.

3.3. Conclusion

Applying graph theory to electrophysiological recordings we were able to characterize and classify sighs and non-sighs. The visibility graphs and maximum degree allowed to characterize and classify between sighs and non-sighs. Even though the visibility graphs were not effective, the maxclique graphs and parameters of clique algorithm generated a characterization more effective with more successes. Altogether, these results suggest that maxclique graphs and its parameters are more suitable to characterize and classify electrophysiological signals. Likewise, the graphical interface developed allows applying this methodology to other electrophysiological signals.

DATA AVAILABILITY STATEMENT

The data employed to support the findings of this study have been deposited in the *github* repository mentioned before, i.e., https://github.com/Ulipaeh/vgraph.

ETHICS STATEMENT

The animal study was reviewed and approved by Norma Oficial Mexicana (NOM)-062-ZOO-1999 NIH Guidelines for the Euthanasia of Rodent Fetuses and Neonates.

AUTHOR CONTRIBUTIONS

ER-T, UP-H, EV-M, MT-M, CM-V, LB-P, and RV-F conceived and designed the study, and contributed in typing the manuscript. UP-H produced the graphical interface. EV-M contributed the statistical analysis and writing. CM-V and LB-P contributed the electrophysiological experiments. RV-F applied the graph theory concepts and properties.

REFERENCES

Bhaduri, S., Chakraborty, A., and Ghosh, D. (2016). Speech emotion quantification with chaos-based modified visibility possible precursor of suicidal tendency. *J. Neurol. Neurosci.* 7:100. doi: 10.21767/2171-6625.1000100

Caro-Martín, C. R., Delgado-García, J. M., Gruart, A., and Sánchez-Campusano, R. (2018). Spike sorting based on shape, phase, and distribution features, and K-TOPS clustering with validity and error indices. *Sci. Rep.* 8:17796. doi: 10.1038/s41598-018-35491-4

Einevoll, G. T., Franke, F., Hagen, E., Pouzat, C., and Harris, K. D. (2012). Towards reliable spike-train recordings from thousands of neurons with multielectrodes. *Curr. Opin. Neurobiol.* 22, 11–17. doi: 10.1016/j.conb.2011.10.001

García-Herrera, R. (2015). *Visibility_Graph 0.4*. Available online at: https://pypi.org/project/visibility_graph/

Gershenson, C. and Niazi, M. A. (2013). Multidisciplinary applications of complex networks modeling, simulation, visualization, and analysis. *Complex Adapt. Syst. Model.* 1:17. doi: 10.1186/2194-3206-1-17

Harary, F. (1969). *Graph Theory*. Reading, MA; Menlo Park, CA; London: Addison-Wesley Publishing Co.

Hou, F. Z., Li, F. W., Wang, J., and Yan, F. R., (2016). Visibility graph analysis of very short-term heart rate variability during sleep. *Phys. A* 458, 140–145. doi: 10.1016/j.physa.2016.03.086

Jiang, S., Bian, C., Ning, X., and Qianli, D. Y. M. (2013). Visibility graph analysis on heartbeat dynamics of meditation training. *Appl. Phys. Lett.* 102:253702. doi: 10.1063/1.4812645

Kam, K., Worrell, J. W., Janczewski, W. A., Cui, Y., and Feldman, J. L. (2013). Distinct inspiratory rhythm and pattern generating mechanisms in the pre-Bötzinger complex. *J. Neurosci.* 33, 9235–9245. doi: 10.1523/JNEUROSCI.4143-12.2013

Lacasa, L. and Flanagan, R. (2015). Time reversibility from visibility graphs of non-stationary processes. *Phys. Rev.* E92:022817. doi: 10.1103/PhysRevE.92.022817

Lacasa, L., Luque, B., Ballesteros, F., Luque, J., and Nuño, J. C. (2008). From time series to complex networks: the visibility graph. *Proc. Natl. Acad. Sci. U.S.A.* 105, 4972–4975. doi: 10.1073/pnas.0709247105

Lefebvrea, B., Ygera, P., and Marre, O. (2016). Recent progress in multi-electrode spike sorting methods. *J. Physiol. Paris* 110, 327–335. doi: 10.1016/j.jphysparis.2017.02.005

Li, P., Janczewski, W. A., Yackle, K., Kam, K., Pagliardini, S., Krasnow, M. A., et al. (2016). The peptidergic control circuit for sighing. *Nature* 530, 293–297. doi: 10.1038/nature16964

Lieske, S. P. and Ramirez, J.-M. (2006a). Pattern-specific synaptic mechanisms in a multifunctional network. I. Effects of alterations in synapse strength. *J. Neurophysiol.* 95, 1323–1333. doi: 10.1152/jn.00505.2004

Lieske, S. P. and Ramirez, J.-M. (2006b). Pattern-specific synaptic mechanisms in a multifunctional network. II. Intrinsic modulation by metabotropic glutamate receptors. *J. Neurophysiol.* 95, 1334–1344. doi: 10.1152/jn.00506.2004

Lieske, S. P., Thoby-Brisson, M., Telgkamp, P., and Ramirez, J. M. (2000). Reconfiguration of the neural network controlling multiple breathing patterns: eupnea, sighs and gasps. *Nat. Neurosci.* 3, 600–607. doi: 10.1038/75776

Long, X. (2015). *On the analysis and classification of sleep stages from cardiorespiratory activity* (Ph.D. thesis), Department of Electrical Engineering, Proefschrift.

McKee, T. A. and McMorris, F. R. (1999). *Topics in Intersection Graph Theory. SIAM Monographs on Discrete Mathematics and Applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Munoz-Ortiz, J., Munoz-Ortiz, E., López-Meraz, M. L., Beltran-Parraza, L., and Morgado-Valle, C. (2016). Complejo pre-Bötzinger: generación y modulación del ritmo respiratorio. *Neurología* 34, 461–468. doi: 10.1016/j.nrl.2016.05.011

Patroniti, N., Foti, G., Cortinovis, B., Maggioni, E., Bigatello, L. M., Cereda, M., et al. (2002). Sigh improves gas exchange and lung volume in patients with acute respiratory distress syndrome undergoing pressure support ventilation. *Anesthesiology* 96, 788–794. doi: 10.1097/00000542-200204000-00004

Requardt, M. (2000). (Quantum) spacetime as a statistical geometry of lumps in random networks. *Classic. Quant. Gravity* 17, 2029–2057. doi: 10.1088/0264-9381/17/10/301

Rey, H. G., Pedreira, C., and Quiroga, R. Q. (2015). Past, present and future of spike sorting techniques. *Brain Res. Bull.* 119, 106–117. doi: 10.1016/j.brainresbull.2015.04.007

Ruangkittisakul, A., Schwarzacher, S. W., Secchia, L., Yonglie Ma, N. B., Poon, B. Y., Funk, G. D., et al. (2008). Generation of Eupnea and Sighs by a spatiochemically organized inspiratory network. *J. Neurosci.* 28, 2447–2458. doi: 10.1523/JNEUROSCI.1926-07.2008

Shao, Z.-G. (2010). Network analysis of human heartbeat dynamics. *Appl. Phys. Lett.* 96:073703. doi: 10.1063/1.3308505

Szwarcfiter, J. L. (2003). "A survey on clique graphs," in *Recent Advances in Algorithms and Combinatorics, Volume 11 of CMS Books Math./Ouvrages Math. SMC*, eds B. A. Reed and C. Linhares-Sales (New York, NY: Springer), 109–136.

Tomita, E., Akutsu, T., and Matsunaga, T. (2011). "Efficient algorithms for finding maximum and maximal cliques: Effective tools for bioinformatics," in *Biomedical Engineering, Trends in Electronics, Communications and Software*, ed A. N. Laskovski (Rijeka: InTech), 625–640.

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

ORIGINAL RESEARCH
published: 21 April 2020
doi: 10.3389/fgene.2020.00328

Predicting the Disease Genes of Multiple Sclerosis Based on Network Representation Learning

Haijie Liu[1,2,3]*, Jiaojiao Guan[4], He Li[5], Zhijie Bao[6], Qingmei Wang[3], Xun Luo[7,8] and Hansheng Xue[4]*

[1] Department of Neurology, Xuanwu Hospital, Capital Medical University, Beijing, China, [2] Department of Physical Medicine and Rehabilitation, Tianjin Medical University General Hospital, Tianjin, China, [3] Stroke Biological Recovery Laboratory, Department of Physical Medicine and Rehabilitation, Spaulding Rehabilitation Hospital, The Teaching Affiliate of Harvard Medical School Charlestown, Boston, MA, United States, [4] School of Computer Science, Northwestern Polytechnical University, Xi'an, China, [5] Department of Automation, College of Information Science and Engineering, Tianjin Tianshi College, Tianjin, China, [6] School of Textile Science and Engineering, Tiangong University, Tianjin, China, [7] Kerry Rehabilitation Medicine Research Institute, Shenzhen, China, [8] Shenzhen Dapeng New District Nan'ao People's Hospital, Shenzhen, China

OPEN ACCESS

Edited by:
Zhongyu Wei,
Fudan University, China

Reviewed by:
Fan Yang,
Harvard Medical School,
United States
Xiaoke Ma,
Xidian University, China

***Correspondence:**
Haijie Liu
haijieliu@163.com
Hansheng Xue
xhs1892@gmail.com

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

Received: 23 January 2020
Accepted: 19 March 2020
Published: 21 April 2020

Citation:
Liu H, Guan J, Li H, Bao Z, Wang Q, Luo X and Xue H (2020) Predicting the Disease Genes of Multiple Sclerosis Based on Network Representation Learning. Front. Genet. 11:328.
doi: 10.3389/fgene.2020.00328

Multiple sclerosis (MS) is an autoimmune disease for which it is difficult to find exact disease-related genes. Effectively identifying disease-related genes would contribute to improving the treatment and diagnosis of multiple sclerosis. Current methods for identifying disease-related genes mainly focus on the hypothesis of guilt-by-association and pay little attention to the global topological information of the whole protein-protein-interaction (PPI) network. Besides, network representation learning (NRL) has attracted a huge amount of attention in the area of network analysis because of its promising performance in node representation and many downstream tasks. In this paper, we try to introduce NRL into the task of disease-related gene prediction and propose a novel framework for identifying the disease-related genes multiple sclerosis. The proposed framework contains three main steps: capturing the topological structure of the PPI network using NRL-based methods, encoding learned features into low-dimensional space using a stacked autoencoder, and training a support vector machine (SVM) classifier to predict disease-related genes. Compared with three state-of-the-art algorithms, our proposed framework shows superior performance on the task of predicting disease-related genes of multiple sclerosis.

Keywords: multiple sclerosis, network embedding, disease gene prediction, PPI network, deep learning

1. INTRODUCTION

Multiple sclerosis (MS) is an autoimmune disease that disrupts the myelin and axons, which leads to inflammatory disorder of the brain and spinal cord (Compston and Coles, 2002), and it is difficult to find exact pathogens and disease-related genes. In recent studies, some of the disease-related genes of multiple sclerosis have been collected and made available, such as in the DisGeNet database (Pinero et al., 2017). However, there are still many unknown MS disease-related genes that need to be discovered. Identifying such genes will effectively contribute to discovering the inner molecular mechanisms of MS as a disease and will help researchers to learn more about MS. Thus, it is essential and of importance to develop a novel algorithm to identify the disease-related genes of MS rapidly and effectively.

Predicting disease-related genes has attracted a huge amount of attention in recent years, and many computational methods have been proposed because of the natural advantages of such methods in terms of time and money saved (Peng et al., 2017, 2019a, 2020a; Ma et al., 2018a; Hu et al., 2019; Xue et al., 2019b). Furthermore, computational methods are effective and precise enough to guide wet experiments (Liu et al., 2019a,b; Peng et al., 2019c). Thus, it is necessary to explore the area of predicting disease-related genes using computational methods. Most of the existing methods for predicting disease-related genes are based on the assumption of the guilt-by-association hypothesis (Peng et al., 2019a). Specifically, genes associated with the same or similar diseases usually have a higher probability of sharing the same topological structure or similar neighbors as others in the gene interaction networks. Thus, based on this guilt-by-association hypothesis, the core of predicting disease-related genes is calculating the distance or similarity between candidate genes and disease-related genes effectively and correctly.

Many approaches have been proposed to measure distance or similarity between gene nodes. The simplest method is direct neighborhood counting (Oti et al., 2006), which mainly counts the number of disease-related genes among their neighborhoods. If the neighbors of gene *g* are associated with multiple sclerosis disease, gene *g* is likely to be a disease-related gene. However, this method overlooks disease-related genes that do not connect with g in the protein-protein-interaction (PPI) network. To solve this problem, several methods are proposed to utilize the shortest path length model to measure the distance between genes (Krauthammer et al., 2004). However, these methods have not achieved satisfying performance, because both the directing neighborhood counting and shortest path length methods only consider the local topological structure of the PPI network instead of the global information of the network topology. Many papers suggest that global topological information would be able to improve the performance of gene node presentation and downstream tasks (Ma et al., 2018b, 2019; Peng et al., 2019b, 2020b; Xue et al., 2019a). Thus, some papers have tried to capture global topological information through random walk with restart (Li and Patra, 2010; Ma et al., 2017; Peng et al., 2018). Borrowing ideas from random walk with restart, we aim, in the current study, to introduce network representation learning (NRL) methods, which represent genes in the network as low-dimensional features, into the task of predicting the disease-related genes of MS.

In this paper, we implement an existing NRL method, termed NRL-based algorithms, for the task of predicting MS disease-related genes and transform non-linear feature vectors into low-dimensional space with a stacked autoencoder. The contributions of this paper can be listed as follows:

- NRL-based algorithms learn global non-linear topological information of the protein-protein-interaction network based on node2vec, DeepWalk, and LINE.
- The deep learning model of a stacked autoencoder is implemented in our proposed framework to extract low-dimensional feature vectors.
- NRL-based algorithms show superior performance in the task of predicting the disease-related genes of MS.

2. METHODS

In this paper, we introduce NRL algorithms, termed NRL-based algorithms, for the task of predicting the disease-related genes of MS. The framework used contains three main parts: NRL-based algorithms, a Stacked AutoEncoder (Bengio et al., 2006), and a Support Vector Machine (SVM) (Chang and Lin, 2011). Here, we use three classical NRL algorithms to transform the PPI network into high-dimensional feature space, namely node2vec (Grover and Leskovec, 2016), DeepWalk (Perozzi et al., 2014), and LINE (Tang et al., 2015). After obtaining the PPI network embedding features, we run a stacked autoencoder model to extract useful feature vectors into low-dimensional space. Finally, a SVM classifier is implemented to predict the disease-related genes of MS. The whole workflow of the model is shown in **Figure 1**.

2.1. NRL-Based Protein-Protein Interaction Network Embedding

In our method, we use three classical NRL algorithms (node2vec, DeepWalk, and LINE) to capture the global features of the PPI network and represent genes as non-linear feature vectors. The details of the three algorithms are introduced in the next part.

DeepWalk (Perozzi et al., 2014) is the first-proposed NRL algorithm. It tries to represent nodes as novel latent feature vectors. It first learns topological information from the network using a random walk algorithm. Then, it can be treated as a natural language process problem. The learned sequence information is inputted into the Skip-Gram model. The aim of the DeepWalk model is to maximize the probability of neighbors of the node n_i in the walk sequence. The objective function can be shown as:

$$max_{\varphi} Pr(\{n_{i-w}, ..., n_{i+w}\} \setminus n_i|\varphi(n_i)) = \prod_{j=i-w,j\neq i}^{i+w} Pr(n_j|\varphi(n_i)) \quad (1)$$

where w is the size of the window and $\varphi(n_i)$ and $\{n_{i-w}, ..., n_{i+w}\}$ are the current feature representation and neighborhood nodes of n_i, respectively. Finally, the DeepWalk algorithm uses hierarchical softmax to generate the low-dimensional representation vectors. The overall overflow can be seen in **Figure 2A**. node2vec (Grover and Leskovec, 2016) is an extended version of the DeepWalk algorithm. In the process of learning the network topology, node2vec integrates two neighborhood sampling strategies, Breadth-First Search (BFS) and Depth First Search (DFS). These two strategies for capturing topological information are shown in **Figure 2B**. The node2vec algorithm proposes a novel random walk strategy with two parameters, p and q. The random walk procedure of node2vec can be seen in **Figure 2C**. Parameter p mainly controls the probability of revisiting a node in the process of random walk, and q controls the possibility of capturing "local" or "global" nodes. In particular, if $p = 1.0$ and $q = 1.0$, then the node2vec algorithm can be seen similarly as the DeepWalk method.

LINE (Tang et al., 2015) is designed for large-scale NRL, mainly capturing the first-order and second-order topological

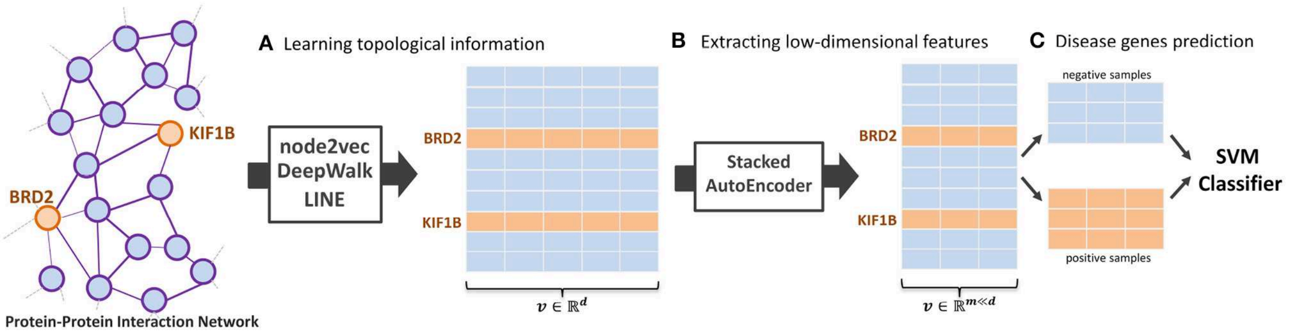

Protein-Protein Interaction Network
KIF1B
BRD2
A Learning topological information
node2vec
DeepWalk
LINE
BRD2
KIF1B
$v \in \mathbb{R}^d$
B Extracting low-dimensional features
Stacked
AutoEncoder
BRD2
KIF1B
$v \in \mathbb{R}^{m \ll d}$
C Disease genes prediction
negative samples
positive samples
SVM
Classifier

FIGURE 1 | The workflow of the proposed NRL-based framework. The framework contains three main parts: **(A)** learning the topological structure of the protein-protein-interaction network, **(B)** transforming network embedding features into low-dimensional space, and **(C)** training the support vector machine classifier to predict disease-related genes.

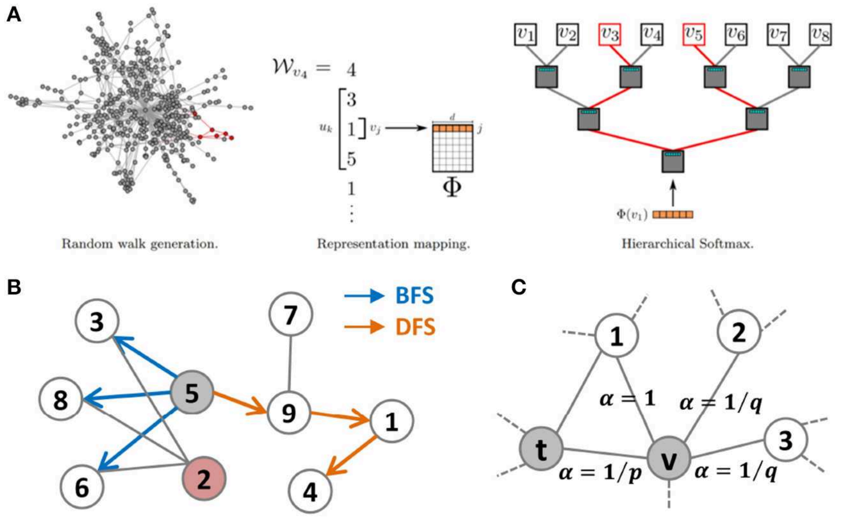


FIGURE 2 | (A) Overview of DeepWalk. It consists of three main parts: random walk generation, representation learning, and hierarchical softmax. This figure was extracted from the original paper. **(B)** Two types of search strategies from node 5, BFS and DFS. **(C)** The random walk procedure in node2vec.

information. The idea of second-order information in LINE can be learned from **Figure 2B**. In this figure, nodes 5 and 2 have the same neighborhood, 3, 8, and 6. Although nodes 2 and 5 are not linked directly, we think that they are similar to each other. The first-order and second-order topological information between two nodes n_i and n_j can be measured as:

$$P_1(n_i, n_j) = \frac{1}{1 + exp(-u_i^T u_j)} \qquad P_2(n_j|n_i) = \frac{exp(\bar{u}_j^T \bar{u}_i)}{\sum_k exp(\bar{u}_k^T \bar{u}_i)} \tag{2}$$

where u_i describes the representation of node n_i. By optimizing the KL-divergence of these first-order and second-order distributions, we can obtain the final representations of gene nodes.

2.2. Extracting Low-Dimensional Feature Vectors

In our NRL-based MS disease-related gene prediction model, we use a stacked autoencoder model to transform high-dimensional non-linear features learned by NRL-based algorithms into low-dimensional feature space. Commonly, many models use Principal Component Analysis (PCA) (Abdi and Williams, 2010) or Independent Component Analysis (ICA) (Hyvärinen and Oja, 2000) to reduce the dimensionality of the feature matrix. However, these methods cannot capture non-linear feature vectors effectively. Also, these linear dimensionality reduction methods would distort the original data structure and cannot keep original features in the low-dimensional feature space. A stacked autoencoder (SAE) model can address these shortcomings.

An autoencoder is an unsupervised model that is widely used in feature extraction and dimensionality reduction. An autoencoder contains two main parts, an encoder and a decoder, and its aim is to minimize the reconstruction error between input and output. The encoded features of the hidden layer are the final low-dimensional output that is used in the downstream tasks. Assuming that the $i-th$ input node vector is x_i, the reconstructed node vector can be described as $\hat{x}_i = g(W' \cdot f(W \cdot x_i + b) + b')$,

TABLE 1 | The experimental results of NRL-based methods and other baselines.

	Abc	F1	AUROC	AUPRC
ED	0.6032 (0.0165)	0.5933 (0.0204)	0.6439 (0.0163)	0.6356 (0.0216)
SPL	0.6136 (0.0296)	0.6033 (0.0198)	0.6703 (0.0205)	0.6531 (0.0208)
RWR	0.5312 (0.0113)	0.5203 (0.0305)	0.5431 (0.0195)	0.5321 (0.0233)
LINE-SAE-SVM	0.5527 (0.0102)	0.5403 (0.0218)	0.5838 (0.0106)	0.5716 (0.0198)
node2vec-SAE-SVM	**0.7011 (0.0212)**	**0.6944 (0.0138)**	**0.7647 (0.0186)**	0.7472 (0.0283)
DeepWalk-SAE-SVM	0.6941 (0.0288)	0.6914 (0.0315)	0.7554 (0.0204)	**0.7478 (0.0243)**

The bold values indicate the best performance.

where f and g are activation functions, and $\Theta = \{W, b, W', b'\}$ are the parameters to be learned. Then, the loss function of a three-layer autoencoder can be represented as follows:

$$\arg\min_{\theta \in \Theta} \sum_{i=1}^{n} \parallel \hat{x}_i - x_i \parallel_2^2 \tag{3}$$

The stacked autoencoder has been widely used in many areas to extract feature vectors and reduce the dimensionality (Peng et al., 2019b). Thus, we also add a stacked autoencoder model in our framework to improve the performance of predicting MS disease-related genes.

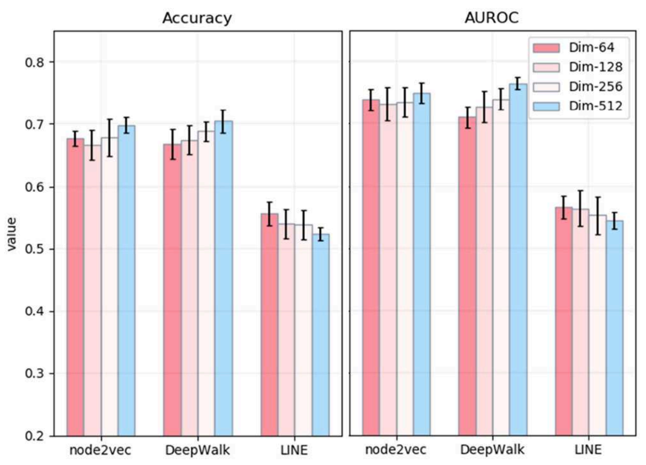


FIGURE 3 | Accuracy and AUPRC values of three network representation learning algorithms with four different numbers of dimensions. The x-axis represents three different methods. The y-axis represents the values of Accuracy **(left)** and AUROC **(right)**.

2.3. Predicting Disease-Related Genes Based on an SVM Classifier

After obtaining low-dimensional gene feature vectors, we train the SVM algorithm to predict the disease-related genes of MS. This prediction task can be treated as a label classification problem. SVM is applied widely on many classification tasks because of its stability, simplicity, and effectiveness. Here, we also select SVM as the classifier for our model. The disease-related genes of MS are chosen as positive samples, and then we randomly select several unrelated genes as negative samples from the PPI network. The number of negative samples is the same as that of positive samples.

In order to evaluate the performance of the SVM classifier in the task of MS disease-related gene prediction, we randomly select 80% of the dataset as a training dataset and 20% as the test dataset. We choose the standard RBF kernel for the SVM classifier and use the grid search method to select the optimal hyper-parameters.

3. RESULTS

3.1. Datasets and Baselines

In the experimental part, we mainly use two datasets: the protein-protein interaction network (PPI) and the disease-related genes of MS. The PPI network contains 13,460 nodes and 141,296 edges, which is the same as in the paper (Menche et al., 2015). Candidate genes associated with MS disease were downloaded from the DisGeNet database (https://www.disgenet.org/browser/0/1/1/C0026769) (Pinero et al., 2017). After preprocessing, we can obtain 924 genes that relate to MS disease. In order to evaluate the performance of our proposed method, we compare NRL-based methods with three classical methods, including Random Walk with Restart (RWR) (Li and Patra, 2010), Shortest Path Length (SPL) (Krauthammer et al., 2004) and Euclidean distance (ED) (Díaz-Uriarte and de Andrés, 2006). Random walk with restart is a classical path learning method, which is widely used in biological network analysis to capture the topological structure of the network. Shortest path length and Euclidean distance are both typical path-based disease-related gene prediction methods. We, in this paper, compare NRL-based methods with these path-based methods to validate the superiority of NRL on the task of disease-related gene prediction.

On the task of disease-related gene prediction, we adopt accuracy, F1, area under the ROC curve (AUROC), and area under the PR curve (AUPRC) as the evaluation criterion. All of the experiments adopt five-fold cross-validation. After several experimental validations, the optimal number of dimensions of the PPI network embedding and the final dimensionality of features after running stacked autoencoder are 512 and 64, respectively.

3.2. Performance in Predicting Disease-Related Red Genes of MS

In order to validate the performance of NRL-based algorithms on the task of predicting the disease-related genes of MS, we

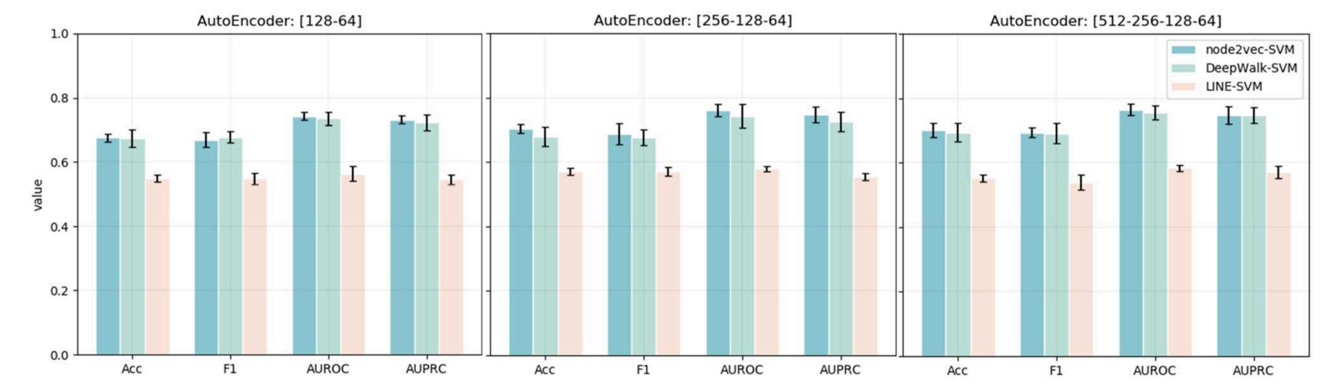


FIGURE 4 | Accuracy, F1, AUPRC, and AUPRC values of three network representation learning algorithms with four different numbers of dimensions and different autoencoder structures. The x-axis represents four different evaluation metrics. The y-axis represents the value of the evaluation metric.

TABLE 2 | The experimental results of NRL-based methods with different classifiers.

		Acc	F1	AUROC	AUPRC
Logistic Regression	LINE	0.5272(0.0131)	0.5172(0.0125)	0.5596(0.0138)	0.5391(0.0248)
	node2vec	0.6483(0.0163)	0.6483(0.0163)	0.6899(0.0236)	0.6409(0.0208)
	DeepWalk	0.5793(0.0250)	0.5793(0.0150)	0.6658(0.0216)	0.6153(0.0200)
Random Forest	LINE	0.6176(0.0188)	0.6276(0.0188)	0.6208(0.0216)	0.6057(0.0263)
	node2vec	0.7172(0.0117)	0.7012(0.0217)	0.7400(0.0126)	0.7191(0.0203)
	DeepWalk	0.6959(0.0215)	0.6759(0.0163)	0.7336(0.0185)	0.7008(0.0202)

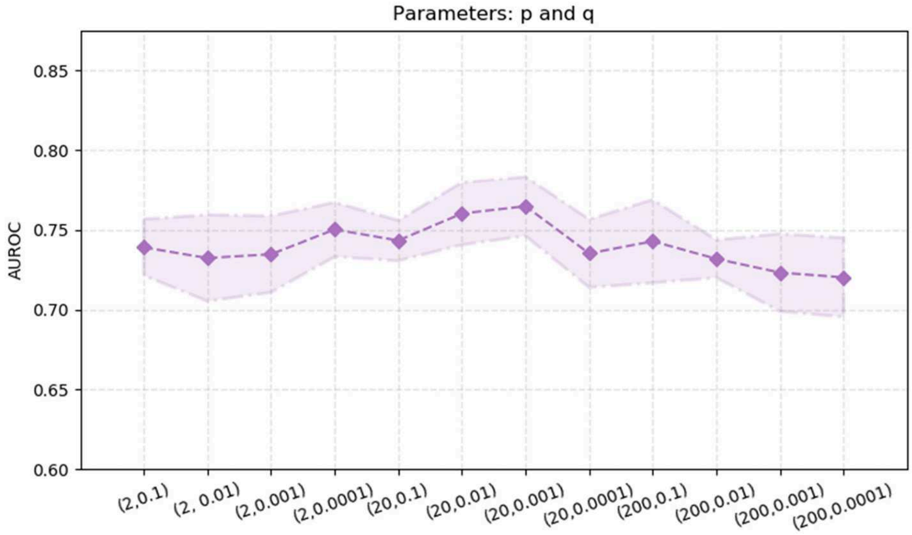


FIGURE 5 | AUROC with different parameter combinations of *p* and *q* in the node2vec algorithm. The x-axis represents different parameter combinations. The y-axis represents the value of AUROC.

compare our model with three classical methods: random walk with restart, shortest path length, and Euclidean distance. The experimental results of the NRL-based methods and baselines are shown in **Table 1**. The node2vec-based and DeepWalk-based methods are obviously superior to the other algorithms. For node2vec, the values of accuracy and AUROC reach 0.7011 and 0.7647, respectively, much higher than the three classical methods. The performance of DeepWalk is similar to that of node2vec, and the AUPRC value of DeepWalk is the highest among the six algorithms. However, the performance of LINE is not as good as the other two NRL-based methods. LINE mainly considers the first-order and second-order information of the network topology in the process of embedding. The PPI network is very sparse and many isolated nodes exist, which may lead to the poor performance of LINE. Overall, the NRL-based methods contribute to improving the performance of MS disease-related gene prediction.

3.3. Effects of Different Parameters on Disease-Related Gene Prediction

The whole process of the NRL-based methods consists of three main parts: capturing the topological information of the PPI network, extracting low-dimensional features, and predicting disease-related genes based on the SVM classifier. Among different parameters, the most influential is the number of dimensions of embedding. Thus, we mainly explore the effects of the number of embedding dimensions on the task of disease-related gene prediction. In detail, we run three NRL algorithms with four different numbers of dimensions, namely 64, 128, 256, and 512. The experimental results are shown in **Figure 3**. In general, the values of accuracy and AUROC are stable, and the number of embedding dimensions has less impact on the experimental results in predicting the disease-related genes of MS. For node2vec, the values of accuracy and AUROC are around 0.67 and 0.73, respectively, in the case of the four different dimensionalities.

Except for the dimensionality of network embedding, we also consider the effects of the stacked autoencoder. Here, we also embed the PPI network with four different numbers of dimensions. We, then, implement the stacked autoencoder to transform high-dimensional features into low-dimensional space. The final number of dimensions through the stacked autoencoder is 64. The experimental results are shown in **Figure 4**. Comparing the experimental results with the model without an autoencoder, we can clearly see the effects of the autoencoder on extracting low-dimensional features. Besides, with the increase in the number of autoencoder layers, the model shows better performance in the task of predicting MS disease-related genes. Thus, we adopt five layers [512-256-128-64] as our model's stacked autoencoder structure. In the third part, an SVM classifier is used in our model to predict disease-related genes. This step is flexible: we can train other classifiers to finish prediction tasks. Here, we also train Logistic Regression and Random Forest classifiers to predict the disease-related genes of MS. The detailed experimental results are shown in **Table 2**.

node2vec performs better than the other two algorithms, DeepWalk and LINE. Thus, we also explore the effects of the two parameters in the node2vec algorithm, p and q. We randomly select parameters $p \in \{2.0, 20.0, 200\}$ and $q \in \{0.1, 0.01, 0.001, 0.0001\}$. The experimental results are shown in **Figure 5**. The AUROC values are fluctuating within a certain range $[0.72, 0.77]$. When $p = 20$ and $q = 0.01$, the AUROC value of the node2vec algorithm achieve its maximum (0.7647).

4. CONCLUSION

Identifying the disease-related genes of MS effectively is essential for the treatment and diagnosis of MS. In this paper, we introduce NRL methods into the task of identifying disease-related genes and propose a novel NRL-based framework to predict the disease-related genes of MS. The NRL-based algorithms consist of three main components: capturing the global topological structure of the PPI, encoding non-linear representation vectors into low-dimensional feature space using a stacked autoencoder, and training a SVM classifier to predict disease-related genes. We compare our proposed method with three classical algorithms. The experimental results show the superior performance of the NRL-based algorithms. Moreover, the proposed NRL-based algorithms are scalable and robust enough to be applied to many other tasks of disease-related gene prediction.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.disgenet.org/browser/0/1/1/C0026769.

AUTHOR CONTRIBUTIONS

HLiu formulated the study concept and designed the study. HX, JG, and HLi performed research and implemented the algorithm. HX and HLi wrote the paper. QW, ZB, and XL designed the experiments and wrote the paper. All authors read and approved the final manuscript.

FUNDING

This work was supported by the National Natural Science Foundation of China (81701189 to HLiu).

ACKNOWLEDGMENTS

We are thankful to the reviewers for their helpful comments.

REFERENCES

Abdi, H., and Williams, L. J. (2010). Principal component analysis. *Wiley Interdiscipl. Rev. Comput. Stat*. 2, 433–459. doi: 10.1002/wics.101

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2006). "Greedy layer-wise training of deep networks," in *Proceedings of the 19th International Conference on Neural Information Processing Systems, NIPS'06* (Cambridge, MA: MIT Press), 153–160.

Chang, C.-C., and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2:27. doi: 10.1145/1961189.1961199

Compston, A., and Coles, A. (2002). Multiple sclerosis. *Deutsche Medizinische Wochenschrift* 359, 1221–1231. doi: 10.1016/S0140-6736(02)08220-X

Diaz-Uriarte, R., and de Andres, S. A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7:3. doi: 10.1186/1471-2105-7-3

Grover, A., and Leskovec, J. (2016). "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference* *on Knowledge Discovery and Data Mining* (San Francisco, CA: ACM), 855–864. doi: 10.1145/2939672.2939754

Hu, Y., Zhao, T., Zang, T., Zhang, Y., and Cheng, L. (2019). Identification of alzheimer's disease-related genes based on data integration method. *Front. Genet*. 9:703. doi: 10.3389/fgene.2018.00703

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural Netw*. 13, 411–430. doi: 10.1016/s0893-6080(00)00026-5

Krauthammer, M., Kaufmann, C. A., Gilliam, T. C., and Rzhetsky, A. (2004). Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. *Proc. Natl. Acad. Sci. U.S.A*. 101, 15148–15153. doi: 10.1073/pnas.0404315101

Li, Y., and Patra, J. C. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics* 26, 1219–1224. doi: 10.1093/bioinformatics/btq108

Liu, J., Liu, Q., Zhang, L., Su, S., and Liu, Y. (2019a). Enabling massive XML-based biological data management in hbase. *IEEE/ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/TCBB.2019.2915811. [Epub ahead of print].

Liu, J., Qu, Z., Yang, M., Sun, J., Su, S., and Zhang, L. (2019b). Jointly integrating VCF-based variants and owl-based biomedical ontologies in MONGODB. *IEEE/ACM Trans. Comput. Biol. Bioinform*. doi: 10.1109/TCBB.2019.2951137. [Epub ahead of print].

Ma, X., Dong, D., and Wang, Q. (2019). Community detection in multi-layer networks using joint nonnegative matrix factorization. *IEEE Trans. Knowl. Data Eng*. 31, 273–286. doi: 10.1109/TKDE.2018.2832205

Ma, X., Sun, P., and Qin, G. (2017). Identifying condition-specific modules by clustering multiple networks. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 15, 1636–1648. doi: 10.1109/TCBB.2017.2761339

Ma, X., Sun, P. G., and Zhang, Z. Y. (2018a). An integrative framework for protein interaction network and methylation data to discover epigenetic modules. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 16, 1855–1866. doi: 10.1109/TCBB.2018.2831666

Ma, X., Tang, W., Wang, P., Guo, X., and Gao, L. (2018b). Extracting stage-specific and dynamic modules through analyzing multiple networks associated with cancer progression. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 15, 647–658. doi: 10.1109/TCBB.2016.2625791

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Oti, M., Snel, B., Huynen, M. A., and Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *J. Med. Genet*. 43, 691–698. doi: 10.1136/jmg.2006.041376

Peng, J., Bai, K., Shang, X., Wang, G., Xue, H., Jin, S., et al. (2017). Predicting disease-related genes using integrated biomedical networks. *BMC Genomics* 18(Suppl. 1):1043. doi: 10.1186/s12864-016-3263-4

Peng, J., Guan, J., and Shang, X. (2019a). Predicting Parkinson's disease genes based on node2vec and autoencoder. *Front. Genet*. 10:226. doi: 10.3389/fgene.2019.00226

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019b). A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics*. 35, 4364–4371. doi: 10.1101/276048

Peng, J., Lu, J., Hoh, D., Dina A. S., Shang, X., Kramer, D. M., et al. (2020a). Identifying emerging phenomenon in long temporal phenotyping experiments. *Bioinformatics* 36, 568–577. doi: 10.1093/bioinformatics/btz559

Peng, J., Wang, X., and Shang, X. (2019c). Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data. *BMC Bioinformatics* 20:284. doi: 10.1186/s12859-019-2769-6

Peng, J., Xue, H., Wei, Z., Tuncali, I., Hao, J., and Shang, X. (2020b). Integrating multi-network topology for gene function prediction using deep neural networks. *Brief. Bioinform*. bbaa036. doi: 10.1093/bib/bbaa036

Peng, J., Zhang, X., Hui, W., Lu, J., Li, Q., Liu, S., et al. (2018). Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. *BMC Syst. Biol*. 12:18. doi: 10.1186/s12918-018-0539-0

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: online learning of social representations," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY), 701–710. doi: 10.1145/2623330.2623732

Pinero, J., Bravo, l., Queralt-Rosinach, N., Gutierrez-Sacristan, A., Deu-Pons, J., Centeno, E., Garcia-Garcia, J., et al. (2017). Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*. 45, D833–D839. doi: 10.1093/nar/gkw943

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). *Line: Large-Scale Information Network Embedding*. doi: 10.1145/2736277.2741093

Xue, H., Peng, J., Li, J., and Shang, X. (2019a). "Integrating multi-network topology via deep semi-supervised node embedding," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19* (Beijing: ACM), 2117–2120. doi: 10.1145/3357384.3358164

Xue, H., Peng, J., and Shang, X. (2019b). Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO. *BMC Syst. Biol*. 13:34. doi: 10.1186/s12918-019-0697-8

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 23 April 2020
doi: 10.3389/fbioe.2020.00349

Detecting Cancer Survival Related Gene Markers Based on Rectified Factor Network

Lingtao Su[1,2], Guixia Liu[2], Juexin Wang[1], Jianjiong Gao[3] and Dong Xu[1]*

[1] Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, United States, [2] Department of Computer Science and Technology, Jilin University, Changchun, China, [3] Memorial Sloan Kettering Cancer Center, New York, NY, United States

Detecting gene sets that serve as biomarkers for differentiating patient survival groups may help diagnose diseases robustly and develop multi-gene targeted therapies. However, due to the exponential growth of search space imposed by gene combinations, the performance of existing methods is still far from satisfactory. In this study, we developed a new method called BISG (BIclustering based Survival-related Gene sets detection) based on a rectified factor network (RFN) model, which allows efficiently biclustering gene subsets. By correlating genes in each significant bicluster with patient survival outcomes using a log-rank test and multi-sampling strategy, multiple survival-related gene sets can be detected. We applied BISG on three different cancer types, and the resulting gene sets were tested as biomarkers for survival analyses. Secondly, we systematically analyzed 12 different cancer datasets. Our analysis shows that the genes in all the survival-related gene sets are mainly from five gene families: microRNA protein coding host genes, zinc fingers C2H2-type, solute carriers, CD (cluster of differentiation) molecules, and ankyrin repeat domain containing genes. Moreover, we found that they are mainly enriched in heme metabolism, apoptosis, hypoxia and inflammatory response-related pathways. We compared BISG with two other methods, GSAS and IPSOV. Results show that BISG can better differentiate patient survival groups in different datasets. The identified biomarkers suggested by our study provide useful hypotheses for further investigation. BISG is publicly available with open source at https://github.com/LingtaoSu/BISG.

Keywords: rectified factor network, biclustering, survival analysis, biomarker, variational inference

OPEN ACCESS

Edited by:
Zhongyu Wei,
Fudan University, China

Reviewed by:
Zhi-Ping Liu,
Shandong University, China
Qin Ma,
The Ohio State University,
United States

****Correspondence:***
Dong Xu
xudong@missouri.edu

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

Received: *31 January 2020*
Accepted: *30 March 2020*
Published: *23 April 2020*

Citation:
Su L, Liu G, Wang J, Gao J and Xu D (2020) Detecting Cancer Survival Related Gene Markers Based on Rectified Factor Network. Front. Bioeng. Biotechnol. 8:349. doi: 10.3389/fbioe.2020.00349

INTRODUCTION

Identifying biomarker genes for survival risk prediction allows earlier detection of mortality risk and design of individualized therapy (Wang and Liu, 2018). Due to the exponential growth of search space imposed by the combination explosion of genes, most proposed survival prediction models mainly focus on a single gene. However, the genes perform their functions as groups rather than individually. Identifying robust gene sets that can consistently predict a patient's survival outcome has become a main challenge in the field.

In gene expression experiments, functionally related genes often exhibit a similar pattern in only a subset of samples or under specific experimental conditions (Padilha and Campello, 2017). This problem can be solved by biclustering, which can be used to detect latent row and column groups of different response patterns (Zhang et al., 2017; Saelens et al., 2018). By combining patient survival information, whether the resulting subset of genes are related to patient survival can be tested. Sparse coding has demonstrated its advantage in biclustering gene expression data (Hochreiter et al., 2010). Using sparse representations, the biclustering model tends to have a smaller number of row and column groups since a large amount of variation is already explained by these observed covariates (Blei et al., 2017). In fact, sparse coding has been well-developed in deep learning obtained by rectified linear units (ReLU) (Xu et al., 2016) and dropout (Srivastava et al., 2014). Recently, the rectified factor network (RFN) model (Clevert et al., 2015) was introduced, which aims at finding a sparse, non-negative representation of the input, and extracting the covariance structure of the data. The RFN model uses the posterior regularization method (Ganchev et al., 2010), which separates model characteristics from data dependent characteristics and restricts the posterior means to be non-negative. As computing posterior is very time consuming, variational inference is utilized in RFN model, which approximates probability densities through optimization. Furthermore, by utilizing the projected Newton and projected gradient update strategies during optimization, RFN can efficiently carry out biclustering with high accuracy.

In this study, we adapted RFN for biclustering analysis of integrated mutation and gene expression datasets from the same sets of samples, and developed a new method called BISG (BIclustering based **S**urvival-related **G**ene sets detection). As in Hochreiter et al. (2010), a bicluster is defined as a pair of a row (gene) set and a column (sample) set for which the rows are similar to each other on the selected columns and vice versa. The motivation for developing BISG is to predict such biclusters using gene expression data and associate these biclusters with diseases and disease subtypes. BISG is a rectified factor analysis model, which extracts the covariance structure of the input data and enforces the posterior has to be non-negative and normalized. Non-negative constraints lead to sparse and non-linear codes, while normalization constraints scale the signal part of each hidden unit. For computing the posterior, a family of variational distribution Q of allowed posterior distributions is introduced. In this way, we transform the biclustering problem into an optimization problem, which is optimized by a generalized alternating minimization algorithm (Gunawardana and Byrne, 2005). To speed up computation in the generalized expectation maximization algorithm, we perform a gradient step in both E-step and M-step with fast GPU implementations. We correlate genes in each significant bicluster with patient survival outcomes using a log-rank test and multi-sampling strategy, and only keep the gene sets that can differentiate sample groups by their significantly different survival curves in training and validation datasets. The identified biomarkers suggested by our study can be used as hypotheses for further investigation in improving cancer patient survival.

MATERIALS AND METHODS

Methods Overview

The overall design of BISG is shown in **Figure 1**. BISG mainly comprises of four parts: (1) data preprocessing, (2) bicluster detection, (3) survival analysis, and (4) result analysis. BISG takes RNAseq data, single nucleotide polymorphisms (SNP) data and sample survival data as input. In the data preprocessing, only genes having at least one SNP mutation and samples with survival information are kept. The expression data are normalized to a range between 0 and 1. Each time 90% of the samples are iteratively used as a training set to detect significant biclusters, and the remaining 10% are then used as a validation set. For bicluster detection, a multi-sampling strategy is applied. Each time we randomly select expression data of 100 different samples from the training set to detect significant biclusters using the RFN model, bicluster extraction, quality control and significance test methods. Biclusters passing all these tests are then used for survival analysis. Based on the genes in each bicluster, BISG separates samples (patients) in the training set into two groups G1 (with over 80% bicluster genes significantly up-regulated) and G2 (with all bicluster genes express normally). The survival curves of the two groups are statistically tested by a log-rank test. A multi-sampling strategy is also used in this test, i.e., each time we randomly select the same number of samples from G2 as in G1 (or from G1 as in G2, depending on which one has more samples). If a bicluster gene set can differentiate sample groups by their significantly different survival curves in 80% samplings in the training set, we then validate whether the bicluster genes can separate patients in the validation set into two different survival groups. We random sample 1,000, 5,000, and 10,000 times respectively, and after all iterations only commonly occurred significant bicluster gene sets that can well separate patients in the validation set into different survival groups are selected as biomarkers. In the result analysis, we conduct an independent test of biomarkers with new datasets from GEO (Gene Expression Omnibus) database, and do KEGG and hallmark gene sets enrichment analysis, and also identify common gene families of all the biomarker genes.

Data Preprocessing

Table 1 summarizes the data of the 12 cancer types used in training and validation of BISG. We downloaded their RNAseq median Z-score datasets, SNP mutation datasets and clinical datasets from the cBioPortal database (Cerami et al., 2012; Gao et al., 2013). Based on the median Z-score value we normalized each gene expression values to a range between 0 and 1 (0 means no change, 1 means highly up-regulated).

After the biomarkers were predicted, we utilized three microarray datasets GSE16011 (Gravendeel et al., 2009), GSE3494 (Palazon et al., 2017), and GSE11969 (Takeuchi et al., 2006), as well as their corresponding sample survival information from the GEO as independent test datasets to confirm these biomarkers detected in gliomas, breast cancer and lung adenocarcinoma, respectively. Two datasets, GSE1456 (Pawitan et al., 2005), which was used by GSAS (Varn et al., 2015) but not BISG, and GSE32062 (Yoshihara et al., 2012),

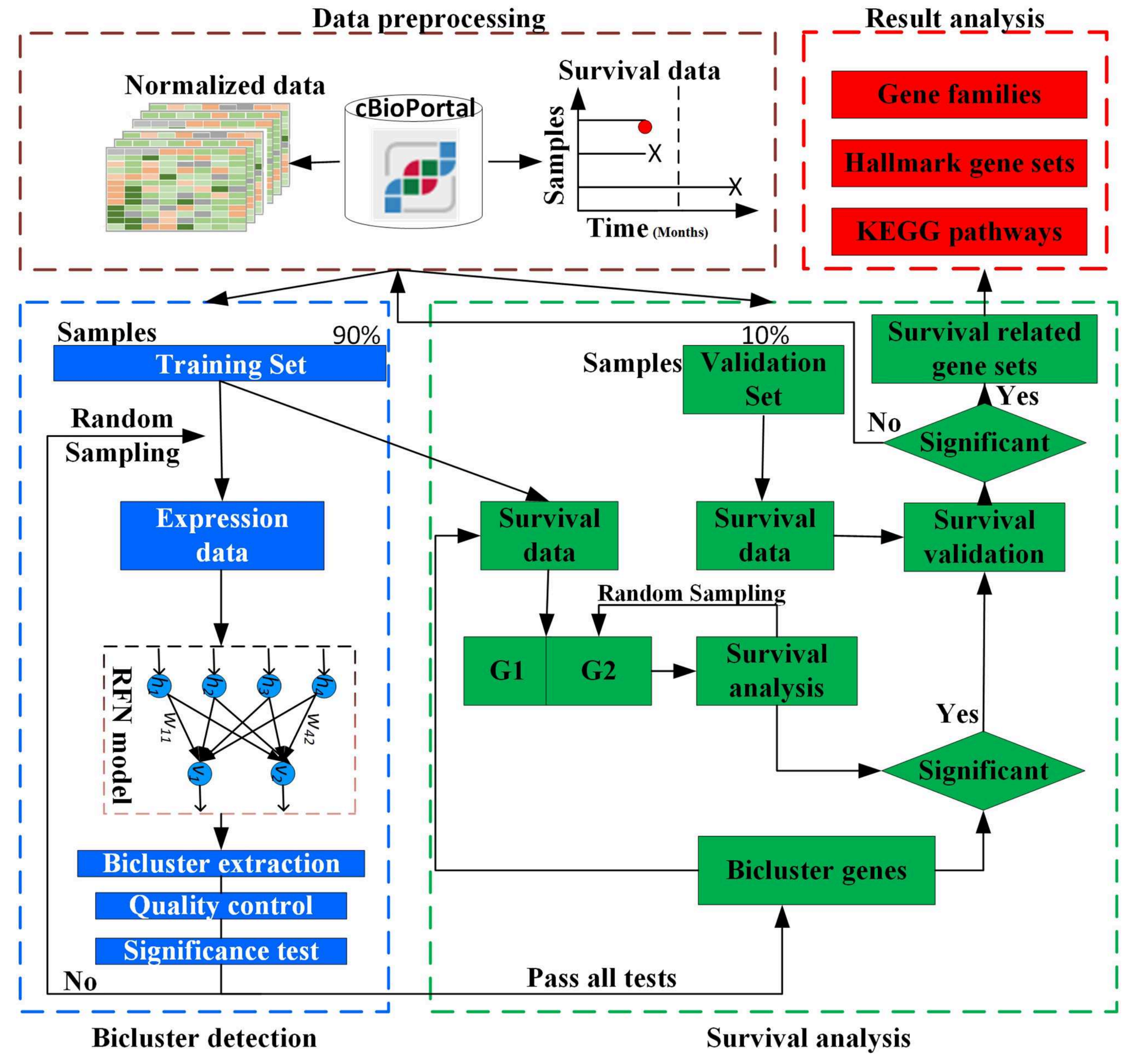


FIGURE 1 | Overview of BISG.

which was used by IPSOV (Shen et al., 2019) but not BISG, were used to compared the classification performance of gene sets detected by BISG, GSAS, and IPSOV. Another dataset GSE3494 (new data for BISG and GSAS) was used to test whether the core gene set detected by GSAS and the top-ranked gene set identified by BISG with breast cancer datasets from cBioPortal database can differentiate samples in GSE3494 into different survival groups. These datasets were normalized the same as in the cBioPortal database, and the datasets were shown in **Table 2**.

Bicluster Detection

Given a normalized gene expression matrix, $V = (X, Y)$, with a set of rows $X = \{x_1, \ldots, x_N\}$, a set of columns $Y = \left\{y_1, \ldots y_M\right\}$, and the element $v_{ij} \in V$ represents the expression value of gene i in sample j. A bicluster $B = (I, J)$ is a $n \times m$ submatrix of V, where $I = (i_1, \ldots i_n) \subset X$ is a subset of genes and $J = (j_1, \ldots j_m) \subset Y$ is a subset of samples. The biclustering aims to identify a set of biclusters $B = \{B_1, \ldots B_s\}$ such that each bicluster $B_k = (I_k, J_k)$ satisfies specific homogeneity criteria. The RFN model is a single or stacked factor analysis model as in Equation (1), which extracts

TABLE 1 | Cancer data used for training and validating biomarkers.

ID	Cancer type	Gene number	SNP number	Sample number
1	Brain lower grade glioma	2,511	3,141	282
2	Colorectal adenocarcinoma	10,680	23,982	222
3	Glioblastoma	4,148	5,974	130
4	Head and neck squamous cell carcinoma	11,767	27,742	500
5	Kidney renal clear cell carcinoma	6,572	9,923	435
6	Lung adenocarcinoma	8,180	16,625	221
7	Ovarian serous cystadenocarcinoma	3,641	4,573	183
8	Pancreatic adenocarcinoma	6,101	9,415	150
9	Papillary thyroid carcinoma	1,320	1,437	313
10	Prostate adenocarcinoma	7,673	12,658	496
11	Thyroid carcinoma	1,656	1,835	395
12	Breast Invasive Carcinoma	7,079	11,089	448

TABLE 2 | Independent test datasets used for confirming predicted biomarkers and for comparison.

ID	Cancer name	Gene number	Sample number
GSE3494	Breast cancer	4,883	236
GSE11969	Lung Adenocarcinoma	5,273	149
GSE16011	Gliomas	2,061	264
GSE1456	Breast cancer	14,204	159
GSE32062	Ovarian cancer	19,592	260

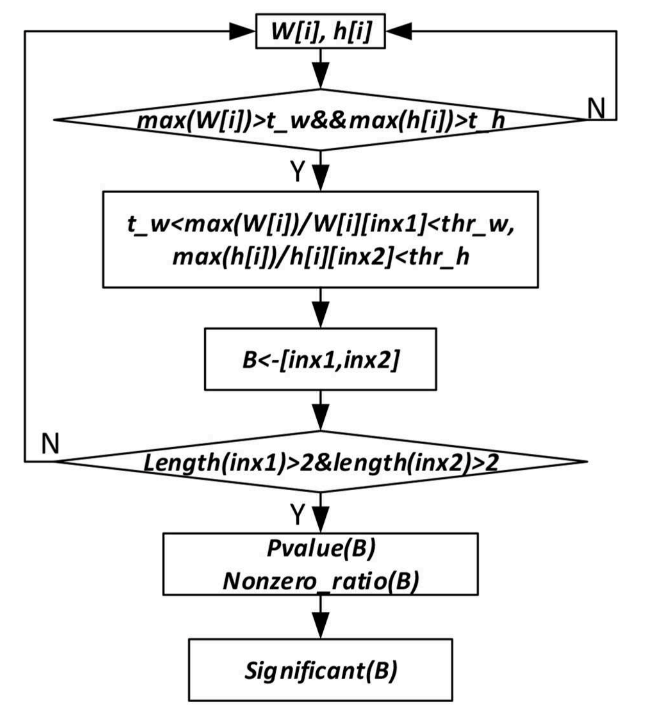

W[i], h[i]
max(W[i])>t_w&&max(h[i])>t_h
N
Y
t_w<max(W[i])/W[i][inx1]<thr_w,
max(h[i])/h[i][inx2]<thr_h
B<-[inx1,inx2]
N
Length(inx1)>2&length(inx2)>2
Y
Pvalue(B)
Nonzero_ratio(B)
Significant(B)

FIGURE 2 | Significant bicluster extraction process. *W*[*i*] and *h*[*i*] are the gene and sample membership vectors. *max* (*W*[*i*]) and *max*(*h*[*i*]) are maximum values of *W*[*i*] and *h*[*i*], respectively. *t_w*, *t_h*, *thr_w,* and *thr_h* are threshold values used to filter bicluster membership genes and samples. *B* represents bicluster. *P-value* (*B*) is *p*-value of a bicluster *B*. *Nonzero_ratio* (*B*) is used for bicluster quality control, which is calculated as the ratio of non-zero elements in a bicluster.

the covariance structure of the data.

$$V = Wh + \varepsilon \tag{1}$$

where $V = \{V_1, \ldots V_N)$ is the input data (visible units), $h \sim N(0, I)$ is the hidden unit (where N is a normal distribution), W is the weight matrix, $\varepsilon \sim \mathbb{N}(0, \Upsilon)$ is the noise error vector, and Υ is the noise covariance matrix. The parameters of the model are W and Υ. If h is given, then only the noise ε is a random variable and we have $V|h \sim N(Wh, \Upsilon)$.

Let E denote the expectation of the data including the prior distribution of the factors and the noise distribution. We can get $E(VV^T) = WW^T + \Upsilon$. The marginal distribution for V is $V \sim N(0, WW^T + \Upsilon)$. The log-likelihood of the input data is given in Equation (2).

$$\begin{aligned}\log \prod_{i=1}^{n} p(V_i) = -\tfrac{nm}{2}\log(2\pi) - \tfrac{n}{2}\log|WW^T + \Upsilon| \\ -\tfrac{1}{2}\sum_{i=1}^{n} V_i^T (WW^T + \Upsilon)^{-1} V_i\end{aligned} \tag{2}$$

For the mean-centered input vector V, the posterior $p(h_i|V_i)$ is Gaussian with the mean vector $(u_p)_i$ and covariance matrix K_{pp} as in Equation (3):

$$\begin{aligned}(u_p)_i &= (I + W^T\Upsilon^{-1}W)^{-1} W^T \Upsilon^{-1} V_i, K_{pp} \\ &= (I + W^T\Upsilon^{-1}W)^{-1}\end{aligned} \tag{3}$$

To maximize the likelihood, we introduce a variational distribution Q, and the objective function $\mathbb{F}$ of our model is shown in Equation (4):

$$\begin{aligned}\mathbb{F} &= \frac{1}{n}\sum_{i=1}^{n} \log p(V_i) - \frac{1}{n}\sum_{i=1}^{n} D_{KL}(Q(h_i|V_i)||p(h_i|V_i)) \\ &= \frac{1}{n}\sum_{i=1}^{n} \int Q(h_i|V_i)\log p(V_i|h_i)dh_i - \frac{1}{n}\sum_{i=1}^{n} D_{KL}(Q(h_i|V_i)||p(h_i))\end{aligned} \tag{4}$$

where Q is a variational distribution for the approximate of the posterior $p(h_i|V_i)$. We constrain Q to the family of rectified and normalized Gaussian distributions. $D_{KL} > 0$ is the KL distance. $\mathbb{F}$ is the objective of the EM algorithm. The E-step maximizes $\mathbb{F}$ with respect to Q; therefore, the E-step minimizes $D_{KL}(Q(h_i|V_i)||p(h_i|V_i))$. The M-step maximizes $\mathbb{F}$ respect to the parameters (W, Υ); therefore, the M-step maximizes $\int Q(h_i|V_i)\log p(V_i|h_i)dh_i$. Considering the quadratic problem of the posterior regularization method, to speed up the computation using fast GPU implementations, we perform a gradient step in both E- and M-steps. In the E-step, we use the projected Newton method as in Equation (5).

$$\min_{\mu_i} \frac{1}{n}\sum_{i=1}^{n}\left(\mu_i-(\mu_p)_i\right)^T\left(\mu_i-(\mu_p)_i\right),\ s.t.\ \mu_i \geq 0,$$
$$\frac{1}{n}\sum_{i=1}^{n}\mu_{ij}^2 = 1 \tag{5}$$

In Equation (5), with $\frac{1}{n}\sum_{i=1}^{n}\mu_{ij}^2 = 1, \mu_i \geq 0$ we constrain the variational distributions to the family of normal distributions with non-negative mean components, and can avoid the explaining away problem as shown in Clevert et al. (2015).

In M-step, we decrease the expected reconstruction error, as in Equation (6).

$$\varepsilon = \frac{1}{2}\left(m\log(2\pi)\right)+\log|\Upsilon|+Tr\left(\Upsilon^{-1}C\right)-2Tr\left(\Upsilon^{-1}W^T\right)$$
$$+\ Tr\left(W^T\Upsilon^{-1}WZ\right) \tag{6}$$

Where $P = \frac{1}{n}\sum_{i=1}^{n}V_i\mu_i^T$, $Z = \frac{1}{n}\sum_{i=1}^{n}V_i\mu_i^T + K_{pp}$ and $C = \frac{1}{n}\sum_{i=1}^{n}V_iV_i^T$. In combination, we get the updates for E-step: $E_Q(h_i) = \mu_i, E_Q(h_ih_i^T) = \mu_i\mu_i^T + K_{pp}$ and M-step: $W^{new} = PZ^{-1}, \Upsilon^{new} = C - PW^T - WP^T + WZW^T$.

To get the sparse, non-negative and non-linear of the input representations, and also to model the covariance structure of the input, we choose the maximum likelihood factor analysis as the model and apply the posterior regularization method (Ganchev et al., 2010). To enforce sparse codes, a Laplace prior on the weight matrix and dropout strategy are used. To further enforce sparseness of the sample and gene membership vectors, we propose a new bicluster extraction strategy as shown in **Figure 2**. For each gene and sample membership vectors, firstly, we get their maximum values, and then for each non-zero element, we get the ratio between the maximum value and the element. If the ratio fulfills the threshold value and at least two genes and two samples are included, then the bicluster is filtered for quality and significance test. For each bicluster passing the quality measure, a *p*-value (Equation 7) is calculated and the Bonferroni correction is used to control the overall type I error.

$$\Pr(B(m,n,q)\geq k)\geq \Pr(B\left(m,n,q\right)\geq mnq\left(1+\frac{k}{mnq}-1\right)) \tag{7}$$

According to Koyuturk et al. (2004), if there is no association in a data matrix, each element can be assumed to an outcome of an independent Bernoulli trial with success probability q. Given a normalized gene expression matrix V with M rows, N columns and K none zero elements, we look for a subset of rows and columns such that a bicluster induced by these rows and columns is dense enough to be considered statistically significant. Assume that $Pr(V(i,j) \neq 0) = q$, where q can be estimated by the density of the matrix, i.e., $q = K/MN$. For an arbitrary bicluster, with m rows and n columns, we assume that the number of non-zero elements is k. Then $k$follows a binormal distribution. The *p*-value of statistical significance test for an $m \times n$ bicluster is given in Equation (7). By using Chernoff's bound (Theodosopoulos, 2007), we get:

$$\Pr(k \geq mnp\,(1+\delta)\,) \leq e^{-mnp\delta^2/3} \tag{8}$$

where $\delta > 0$. Assume that the probability of observing k non-zero elements in the bicluster is less than P^*, then by Equation (8), the bicluster is significant if $k \geq mnp(1+\delta)$, and $\delta \geq \sqrt{3(-\ln P^*)/mnp}$. In summary, according to Koyuturk et al. (2004) the bicluster is statistically significant if:

$$C(m,n,k) = k - mnp - \sqrt{3(-\ln P^*)/mnp} \geq 0 \tag{9}$$

For each bicluster identified, the Bonferroni correction is used to control the overall type I error. The level of significance is set at $\frac{0.05}{b}$, where b is the number of biclusters identified. Besides, we use the none zero ratio in a bicluster to do quality control of the biclustering results. As defined above, the higher the k value, the better the quality of the identified bicluster.

Survival Analysis

We use Kaplan-Meier plots (Goel et al., 2010) to visualize survival curves and with a log-rank test (Singh and Mukhopadhyay, 2011) to compare the survival curves of patients with and without changed expression of the bicluster gene sets. The survival probability, also known as the survivor function S(t), is the probability that an individual survives from the time origin (e.g., diagnosis of cancer) to a specified future time t. The survival probability at time $\mathrm{t_i}$, $\mathrm{S(t_i)}$ is calculated as below:

$$S(t_i) = S(t_{i-1})(1 - d_i/n_i) \tag{10}$$

where $S(t_{i-1})$ is the probability of being alive at t_{i-1}. n_i is the number of patients alive just before t_i. d_i is the number of events at t_i. $t_0 = 0$ and $S(0) = 1$.

Considering genes in each significant bicluster, both samples in the training set and validation set can be divided into two groups G1 (with over 80% bicluster genes significantly changed) and G2 (with bicluster genes express normally). To test the survival difference of samples in G1 and G2, a multi-sampling strategy is utilized, each time the same number of samples are selected. The survival curves of the two selected sample groups can be compared statistically by testing the null hypothesis i.e., there is no difference regarding survival among two groups. This null hypothesis is statistically tested by a log-rank test. In the log-rank test, we calculate the expected number of events in each group, i.e., E1 and E2, while O1 and O2 are the total number of observed events in each group, respectively. The test statistic is:

$$Log-rank\ test = (O_1-E_1)^2/E_1 + (O_2-E_2)^2/E_2 \tag{11}$$

The test statistic and the significance can be drawn by comparing the calculated value with the critical value (using the chi-square table). To guarantee that the bicluster genes are more likely survival-related, for each significant bicluster, considering samples in the training set, we repeat the log-rank test 100 times. If the genes in the bicluster can separate patient groups in more than 80% sampling times, then we use the validation datasets to test whether they can also separate them into two different survival groups. Only bicluster gene sets passing all these significance tests are filtered out as the final biomarkers. We also confirm some biomarkers with independent datasets from the GEO database. In this study, the log-rank test and survival analysis are conducted based on functions in the *lifelines* python package.

RESULTS

Biomarker Gene Sets in Brain Lower Grade Glioma, Lung Adenocarcinoma, and Breast Invasive Carcinoma

We applied BISG on the datasets of brain lower grade glioma, lung adenocarcinoma and breast invasive carcinoma from the cBioPortal database (**Table 1**). Under the default and the same parameter setting as in Su et al. (2019), we identified 24, 7, and 6 significant cancer survival-related biomarker gene sets for lower grade glioma, lung adenocarcinoma and breast invasive carcinoma, respectively (as shown in **Figure 3**, and **Supplementary Figures S1, S2** and **Supplementary Table S4**). The identified gene sets include 109, 82, and 58 genes, respectively. Multiple cancer survival-related genes were found in these genes, including CDH17 (Qiu et al., 2019), PTPRJ (D'Agostino et al., 2018), SLC16A14 (Elsnerova et al., 2017), TMTC2 (He et al., 2018), and NOTCH4 (Wang et al., 2018). Moreover, the results of gene set enrichment analysis and pathway analysis showed that most of the genes have known involvement in cancers. The survival curves of patients with (over 80% bicluster genes significantly upregulated) and without (others) top-ranked four most significant biclusters for each of the three cancer types are shown in **Supplementary Figure S3**, where the bicluster gene sets identified by our methods can well separate patients into two different survival groups.

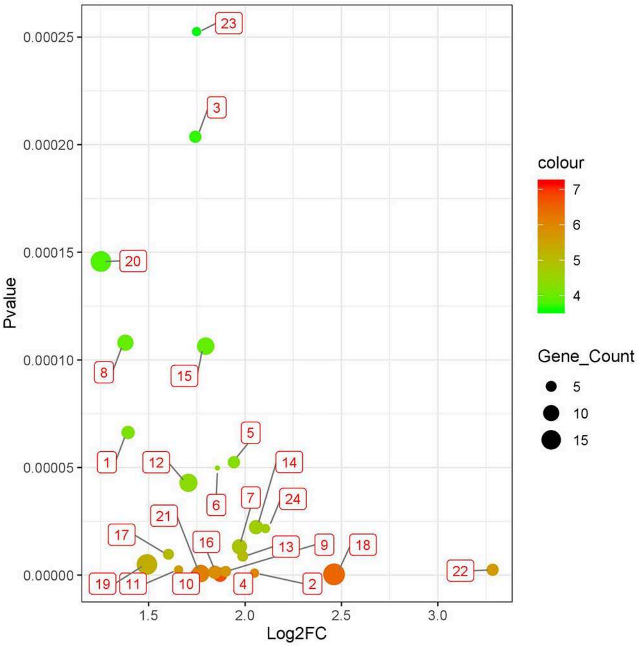


FIGURE 3 | Twenty four significant survival-related gene sets detected in brain lower grade glioma with datasets from the cBioPortal database (**Table 1**). The corresponding genes of each gene set are shown in **Supplementary Tables S1, S4**.

System Analysis Survival-Related Biomarker Gene Sets in 12 Different Cancer Types

We systematically detected significant survival-related biomarker genes sets in 12 different cancer types with datasets in **Table 1**. The number of significant biomarker gene sets and their corresponding gene IDs for each cancer are shown in **Supplementary Table S2**. To find their relationships and functions of these significant biomarker gene sets, firstly, we conducted a function enrichment analysis with the GSEA hallmark gene sets from MSigDB (Liberzon et al., 2015). As shown in **Figure 4**, the function enrichment is mostly in heme metabolism, apoptosis, hypoxia, and inflammatory response. These are consistent with current findings. For example, according to Kalainayakan et al. (2019), cyclopamine tartrate suppresses tumor growth in the lung by inhibiting heme metabolism and OXPHOS (oxidative phosphorylation). A hallmark of cancer is the ability of malignant cells to evade apoptosis (Hanahan and Weinberg, 2011). Avoiding apoptosis is integral to tumor development and resistance to therapy. According to Muz et al. (2015), hypoxia stimulates a complex cell signaling network in cancer cells, including the HIF, PI3K, MAPK, and NFγB pathways. According to Nishijima et al. (2019), inflammatory markers are predictive of poorer survival, independent of traditional prognostic factors in older adults with cancer.

We also analyzed the enriched KEGG pathways of all the bicluster gene sets. As shown in **Supplementary Figure S4**, focal adhesion, neuroactive ligand receptor interaction, endocytosis and pathways in cancer are the most commonly enriched pathways by these gene sets. Finally, we systematically analyzed gene family information of all the biomarker gene sets of each cancer type. Results were shown in **Supplementary Table S3**. According to our analysis, genes in all the survival-related gene sets mainly from five gene families: microRNA protein-coding host genes, zinc fingers C2H2-type, solute carriers, CD molecules and ankyrin repeat domain-containing genes. Many of these genes are known survival-related (detailed information and the corresponding literature are shown in **Supplemental Material**). Furthermore, we found that many cancer survival-related genes identified so far are also from these gene families. For example, LEMD1 and EPHB2 are microRNA protein coding host genes, and SLC2A3 from solute carriers (Martinez-Romero et al., 2018). Other two survival-related genes RAD21 and CKS2 are microRNA protein coding host genes (van't Veer et al., 2002). In addition, CDH1 is from CD molecule (Gao et al., 2019). Of the 68 cancer survival-related gene sets in Varn

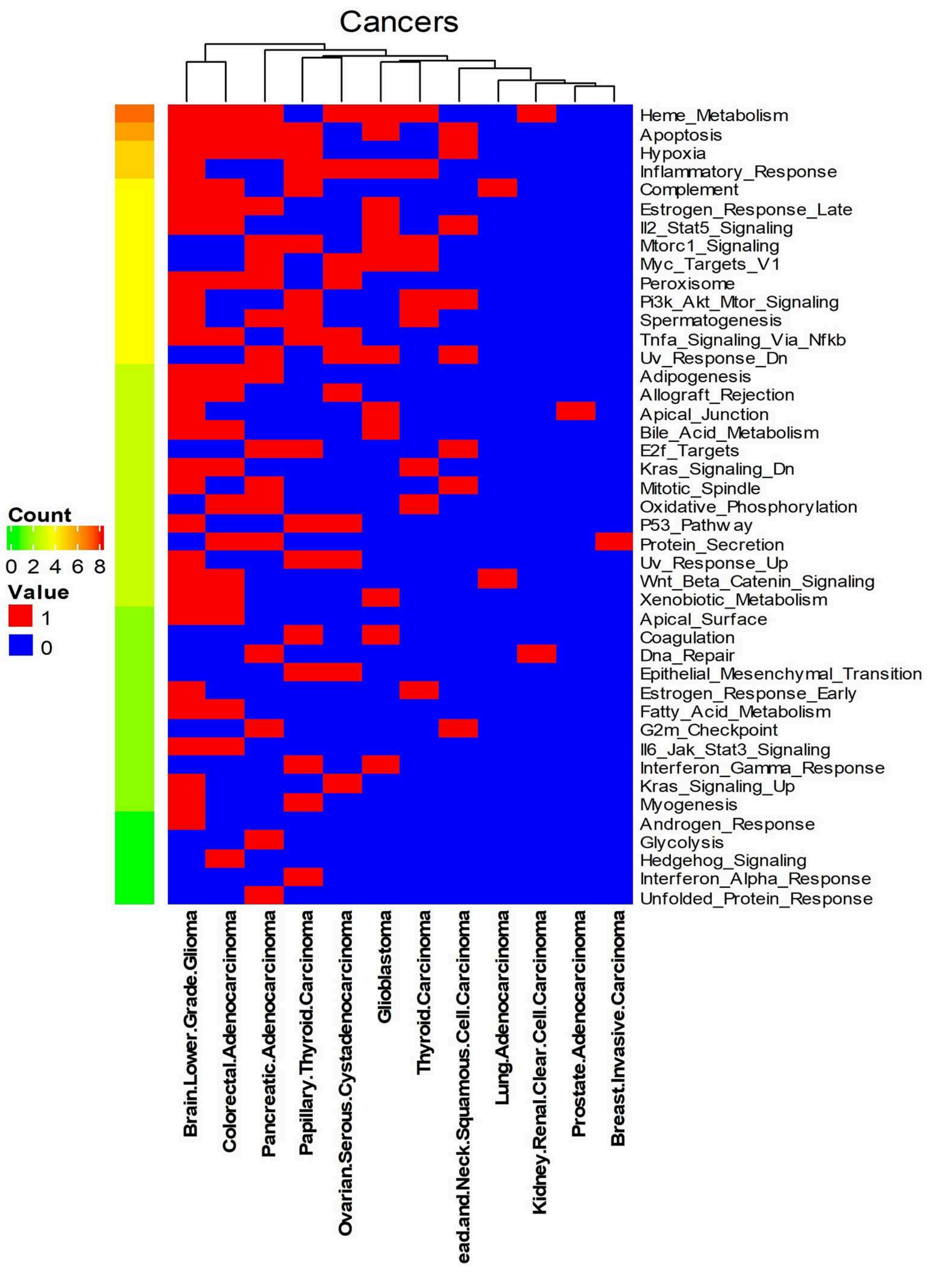


FIGURE 4 | Enriched GSEA hallmark gene sets of all the biomarker gene sets of all the 12 cancer types. Names on the right Y-axis are the hallmark gene sets. Names on the bottom X-axis are the names of the 12 cancer types. Count means the number of cancers whose significant gene sets enriched in corresponding hallmark gene sets. Values in this figure are 0 or 1. Zero means the biomarker gene sets of the corresponding cancer are not enriched in the hallmark gene sets.

et al. (2015), HMMR from CD molecules, MCM7 and CKS2 are microRNA protein coding host genes. Of the 129 ovarian cancer survival-related genes in Shen et al. (2019), 17 are from CD molecules gene family, 7 from microRNA protein-coding host genes, 1 from ankyrin repeat domain-containing gene family.

Results Independent Tests

To test whether biomarker gene sets detected by BISG with datasets from cBioPortal database can differentiate patients into different survival groups with new independent datasets, we collected three microarray datasets GSE16011, GSE3494, and GSE11969, as well as their corresponding sample survival

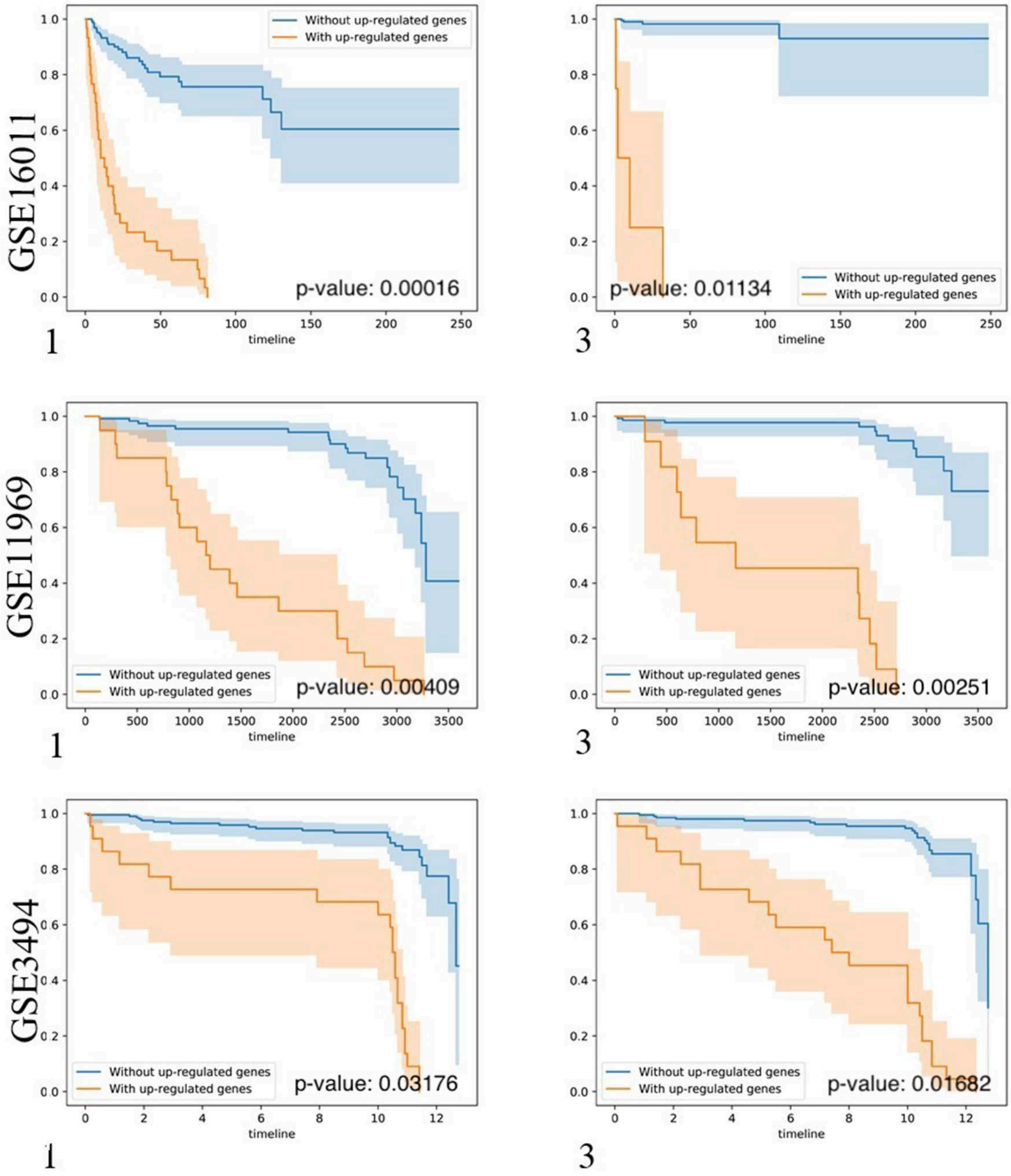


FIGURE 5 | Kaplan-Meier plots of the survival analysis of the samples from brain lower grade glioma (GSE16011), lung adenocarcinoma (GSE11969), and breast invasive carcinoma (GSE3494) patients. 1, 3 means the first and the third top-ranked biomarker gene sets detected by BISG with corresponding cBioPortal datasets. The patients were separated into two groups according to the expression profiles of biomarker genes in the selected biomarker gene set. These genes provided the best split between patients of high and low risk based on their expression levels. In the case of genes in biomarker gene sets (labeled in brown) the over-expression is correlated with poor survival (only up-regulated genes were considered); and in the case of patients without biomarker genes (labeled in blue) the over-expression is correlated with good survival. In all cases the adjusted *p*-values of the analyses are highly significant, indicating that the two populations represented by the two curves have a very clear difference in their overall survival.

information (**Table 2**) from GEO as independent test datasets to confirm the biomarkers detected in gliomas, breast cancer and lung adenocarcinoma, respectively. For comparison, we selected the top-ranked first and third biomarker gene sets (as shown in **Figure 3**, and **Supplementary Figures S2, S3**) for each of the three cancer types. For any selected biomarker gene set, patients can be separated into two groups, one group with biomarker genes significantly changed, and the other with bicluster genes

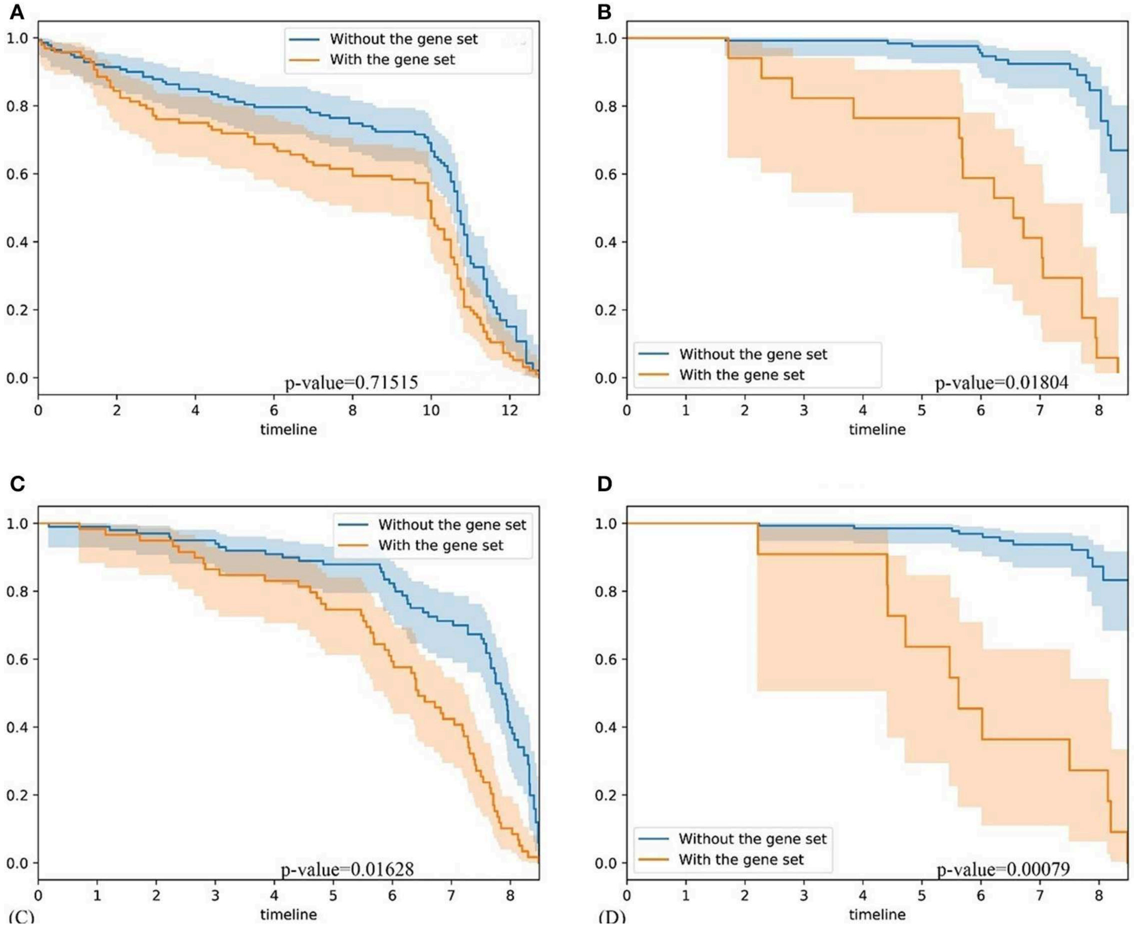


FIGURE 6 | Comparison of gene set based patient survival group classification. "With gene set" means patients with over 80% expression of genes in the gene set significantly changed. "Without the gene set" means patients with the expression of genes in gene set are normal. **(A)** The survival curve of core gene set identified by the GSAS algorithm applied on the GSE1456 dataset. **(B)** The survival curve of the top-ranked gene set identified by our method applied on the GSE1456 dataset. **(C)** The survival curve of core gene set identified by the GSAS algorithm applied on the GSE3494 dataset. **(D)** The survival curve of the top-ranked gene set identified by our method applied on the GSE3494 dataset.

express normally. For survival analysis, we randomly selected the same number of patients from the two groups and test whether their survival curves are significantly different. As shown in **Figure 5**, the biomarker genes can well separate patients into different survival groups.

Comparison With GSAS and IPSOV

To further validate our method, firstly, we compared our methods with GSAS. GSAS quantitatively assesses a gene set's activity score with the BASE algorithm (Cheng et al., 2007), along with patient time-to-event data, to perform survival analyses to identify the gene sets that are significantly correlated with patient survival. Different from our method, they got gene sets directly from MSigDB. By applying on seven independent datasets, one core gene set with 68 genes were filtered out as most related to breast cancer survival. For comparison, we test whether the core gene set detected by GSAS and the top-ranked gene set identified by BISG with breast cancer datasets from cBioPortal database can different samples in GSE1456 (used by GSAS but not BISG) and GSE3494 (new to both two methods) into different survival groups. We run each method many times, and each time we randomly selected the same number of genes from their respective gene sets. The best performing results of each method are shown in **Figure 6**, where the gene set identified by BISG can better separate patients into different survival groups. In **Figures 6A,C**, patients with and without the biomarker genes based on GSAS have similar survival rates, while as shown in (B) and (D), the patients with biomarker genes identified by BISG have different survival rates from the rest. In this comparison, all the datasets are new and independent data that were not used in training BISG. Results indicate that the gene sets identified by BISG can better separate patients into different survival groups.

Furthermore, we also compared BISG with IPSOV. We tested whether the ovarian cancer survival-related gene sets detected by IPSOV (with data from GSE32062) and the top-ranked gene set identified by BISG with ovarian cancer datasets from the cBioPortal database can differentiate samples in GSE32062 (used by GSAS but not BISG) into different survival groups. Detailed results are shown in **Supplementary Figure S5**. Results showed that the biomarker gene set identified by BISG can better separate patients into different survival groups. Again, all the samples for comparison with GSAS were not used by BISG for the selection of biomarker gene sets, which means the biomarker genes identified by BISG are more likely cancer survival related genes.

Based on the fast GPU implementation of the RFN model, BISG can do biclustering analysis of large input datasets in a fast and accurate way, which enables BISG using a multi-sampling strategy to iteratively detect survival-related biomarker gene sets. In contrast to the standard clustering, the samples of a bicluster are only similar to each other on a subset of genes. As a result, genes in each significant bicluster can better differentiate samples into different survival groups. Compared with GSAS and IPSOV, the biomarker gene sets of our method are directly detected from biclustering analysis of the expression datasets, which can well capture the dynamic change of gene sets, and can reflect the real relationships of these genes.

CONCLUSION

In this paper, we proposed BISG for identifying cancer survival-related biomarker gene sets. BISG can efficiently conduct biclustering for high-dimensional gene expression matrix, and along with patient time-to-event data perform survival analyses. To speed up computation, BISG performs a generalized alternating minimization algorithm with GPU implementations. In this way, BISG can efficiently construct very sparse, non-linear, high-dimensional representations of the input via their posterior means. To identify robust biomarker gene sets, multiple iterations and a random sampling strategy were utilized, and each time only bicluster genes that can significantly differentiate patient survival groups were kept. To detect patterns in survival-related gene sets, we systematically analyzed 12 different cancer types, and identified their enriched pathways and their gene families. The results indicated that the identified gene families and genes are biologically meaningful and consistent with the existing scientific findings. With several independent test datasets, identified biomarkers were confirmed. We also compared BISG with two related methods, and BISG outperformed them. The predicted biomarker gene sets can be further investigated for improving cancer patient survival. BISG is now based on a simple factor analysis model, which can be further extended into multi-layers with a deep learning network structure.

Our method has the potential to be extended for single-cell RNA-seq analysis, which has been widely applied in studying cell heterogeneity such as cells of different cancer types or subtypes. A pertinent question in such analyses is to identify cell subpopulations. Our methods can conduct biclustering effectively and efficiently especially for big expression matrices. Ongoing consortium efforts have generated extensive atlases of single-cell datasets covering diverse biological contexts with thousands of samples (Xie et al., 2019), and our methods may be suitable for analyzing them. We will explore applications of our method on single-cell RNA-seq analyses as our future work.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: GSE3439, GSE11969, GSE16011, GSE1456, and GSE32062, https://www.cbioportal.org/.

AUTHOR CONTRIBUTIONS

LS, DX, and GL contributed conception and design of the study. LS, JW, and JG downloaded and organized datasets. LS performed the statistical and result analysis. LS wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work was supported by the National Nature Science Foundation of China to GL (Nos. 61373051 and 61772226), US National Institutes of Health (R35-GM126985) to DX, and US National Cancer Institute Cancer Center Core Grant to JG (P30-CA008748).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00349/full#supplementary-material

REFERENCES

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* 112, 859–877. doi: 10.1080/01621459.2017.1285773

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., and Aksoy, B. A. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 960–960. doi: 10.1158/2159-8290.Cd-12-0326

Cheng, C., Yan, X., Sun, F., and Li, L. M. (2007). Inferring activity changes of transcription factors by binding association with sorted expression profiles. *BMC Bioinformatics* 8:452. doi: 10.1186/1471-2105-8-452

Clevert, D. A., Mayr, A., Unterthiner, T., and Hochreiter, S. (2015). Rectified factor networks. *Adv. Neural. Inf. Process. Syst.* 28:2028. doi: 10.5555/2969442.2969447

D'Agostino, S., Lanzillotta, D., Varano, M., Botta, C., Baldrini, A., Bilotta, A., et al. (2018). The receptor protein tyrosine phosphatase PTPRJ negatively modulates the CD98hc oncoprotein in lung cancer cells. *Oncotarget* 9, 23334–23348. doi: 10.18632/oncotarget.25101

Elsnerova, K., Bartakova, A., Tihlarik, J., Bouda, J., Rob, L., Skapa, P., et al. (2017). Gene expression profiling reveals novel candidate markers of ovarian carcinoma intraperitoneal metastasis. *J. Cancer* 8, 3598–3606. doi: 10.7150/jca.20766

Ganchev, K., Graca, J., Gillenwater, J., and Taskar, B. (2010). Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.* 11, 2001–2049. doi: 10.5555/1756006.1859918

Gao, C. D., Zhuang, J., Zhou, C., Li, H. Y., Liu, C., Liu, L. J., et al. (2019). SNP mutation-related genes in breast cancer for monitoring and prognosis of patients: a study based on the TCGA database. *Cancer Med.* 8, 2303–2312. doi: 10.1002/cam4.2065

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

Goel, M. K., Khanna, P., and Kishore, J. (2010). Understanding survival analysis: kaplan-meier estimate. *Int. J. Ayurveda Res.* 1, 274–278. doi: 10.4103/0974-7788.76794

Gravendeel, L. A. M., Kouwenhoven, M. C. M., Gevaert, O., de Rooi, J. J., Stubbs, A. P., Duijm, J. E., et al. (2009). Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Research.* 69, 9065–9072. doi: 10.1158/0008-5472.Can-09-2307

Gunawardana, A., and Byrne, W. (2005). Convergence theorems for generalized alternating minimization procedures. *J. Mach. Learn. Res.* 6, 2049–2073. doi: 10.5555/1046920.1194913

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144, 646–674. doi: 10.1016/j.cell.2011.02.013

He, R. Q., Zhou, X. G., Yi, Q. Y., Deng, C. W., Gao, J. M., Chen, G., et al. (2018). Prognostic signature of alternative splicing events in bladder urothelial carcinoma based on spliceseq data from 317 cases. *Cell Physiol. Biochem.* 48, 1355–1368. doi: 10.1159/000492094

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

Kalainayakan, S. P., Ghosh, P., Dey, S., Fitzgerald, K. E., Sohoni, S., Konduri, P. C., et al. (2019). Cyclopamine tartrate, a modulator of hedgehog signaling and mitochondrial respiration, effectively arrests lung tumor growth and progression. *Sci. Rep.* 9:1405. doi: 10.1038/s41598-018-38345-1

Koyuturk, M., Szpankowski, W., and Grama, A. (2004). "Biclustering gene-feature matrices for statistically significant dense patterns," in *2004 IEEE Computational Systems Bioinformatics Conference Proceedings* (Stanford, CA), 480–484.

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Martinez-Romero, J., Bueno-Fortes, S., Martin-Merino, M., de Molina, A. R., and de Las Rivas, J. (2018). Survival marker genes of colorectal cancer derived from consistent transcriptomic profiling. *BMC Genomics* 19:857. doi: 10.1186/s12864-018-5193-9

Muz, B., de la Puente, P., Azab, F., and Azab, A. K. (2015). The role of hypoxia in cancer progression, angiogenesis, metastasis, and resistance to therapy. *Hypoxia* 3, 83–92. doi: 10.2147/HP.S93413

Nishijima, T. F., Deal, A. M., Lund, J. L., Nyrop, K. A., Muss, H. B., and Sanoff, H. K. (2019). Inflammatory markers and overall survival in older adults with cancer. *J. Geriatr. Oncol.* 10, 279–284. doi: 10.1016/j.jgo.2018.08.004

Padilha, V. A., and Campello, R. J. G. B. (2017). A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics* 18:55. ARTN 55 doi: 10.1186/s12859-017-1487-1

Palazon, A., Tyrakis, P. A., Macias, D., Velica, P., Rundqvist, H., Fitzpatrick, S., et al. (2017). An HIF-1alpha/VEGF-A axis in cytotoxic T cells regulates tumor progression. *Cancer Cell* 32, 669–683. doi: 10.1016/j.ccell.2017.10.003

Pawitan, Y., Bjohle, J., Amler, L., Borg, A. L., Egyhazi, S., Hall, P., et al. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.* 7, R953–R964. doi: 10.1186/bcr1325

Qiu, H. B., Zhang, L. Y., Ren, C., Zeng, Z. L., Wu, W. J., Luo, H. Y., et al. (2019). Targeting CDH17 suppresses tumor progression in gastric cancer by downregulating Wnt/beta-catenin signaling. *PloS ONE* 14:e56959. doi: 10.1371/journal.pone.0056959

Saelens, W., Cannoodt, R., and Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. *Nat. Commun.* 9:1090. doi: 10.1038/s41467-018-03424-4

Shen, S. P., Wang, G. R., Zhang, R. Y., Zhao, Y., Yu, H., Yongyue, W., et al. (2019). Development and validation of an immune gene-set based prognostic signature in ovarian cancer. *EBioMedicine* 40, 318–326. doi: 10.1016/j.ebiom.2018.12.054

Singh, R., and Mukhopadhyay, K. (2011). Survival analysis in clinical trials: basics and must know areas. *Perspect. Clin. Res.* 2, 145–148. doi: 10.4103/2229-3485.86872

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Su, L., Liu, G., Wang, J., and Xu, D. (2019). A rectified factor network based biclustering method for detecting cancer-related coding genes and miRNAs, and their interactions. *Methods* 166, 22–30. doi: 10.1016/j.ymeth.2019.05.010

Takeuchi, T., Tomida, S., Yatabe, Y., Kosaka, T., Osada, H., Yanagisawa, K., et al. (2006). Expression profile-defined classification of lung adenocarcinoma shows close relationship with underlying major genetic changes and clinicopathologic behaviors. *Int. J. Clin. Oncol.* 24, 1679–1688. doi: 10.1200/Jco.2005.03.8224

Theodosopoulos, T. (2007). A reversion of the chernoff bound. *Stat. Probabil. Lett.* 77, 558–565. doi: 10.1016/j.spl.2006.09.003

van't Veer, L. J., Dai, H. Y., van de Vijver, M. J., He, Y. D. D., Hart, A. A. M., Mao, M., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536. doi: 10.1038/415530a

Varn, F. S., Ung, M. H., Lou, S. K., and Cheng, C. (2015). Integrative analysis of survival-associated gene sets in breast cancer. *BMC Med. Genomics* 8:11. ARTN 11 doi: 10.1186/s12920-015-0086-0

Wang, J. W., Wei, X. L., Dou, X. W., Huang, W. H., Du, C. W., and Zhang, G. J. (2018). The association between Notch4 expression, and clinicopathological characteristics and clinical outcomes in patients with breast cancer. *Oncol. Lett.* 15, 8749–8755. doi: 10.3892/ol.2018.8442

Wang, W., and Liu, W. (2018). Integration of gene interaction information into a reweighted random survival forest approach for accurate survival prediction and survival biomarker discovery. *Sci. Rep.* 8:13202. doi: 10.1038/s41598-018-31497-0

Xie, J., Ma, A., Zhang, Y., Liu, B., Cao, S., Wang, C., et al. (2019). QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. *Bioinformatics* 36, 1143–1149. doi: 10.1093/bioinformatics/btz692

Xu, L., Choy, C. S., and Li, Y. W. (2016). "Deep sparse rectifier neural networks for speech denoising," in *2016 IEEE International Workshop on Acoustic Signal Enhancement* (Xi'an: Iwaenc).

Yoshihara, K., Tsunoda, T., Shigemizu, D., Fujiwara, H., Hatae, M., Fujiwara, H., et al. (2012). High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway. *Clin. Cancer Res.* 18, 1374–1385. doi: 10.1158/1078-0432.Ccr-11-2725

Zhang, Y., Xie, J., Yang, J. Y., Fennell, A., Zhang, C., and Ma, Q. (2017). QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics* 33, 450–452. doi: 10.1093/bioinformatics/btw635

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

ORIGINAL RESEARCH
published: 05 May 2020
doi: 10.3389/fbioe.2020.00390

Protein Interaction Network Reconstruction Through Ensemble Deep Learning With Attention Mechanism

Feifei Li[1], *Fei Zhu*[1,2]*, *Xinghong Ling*[1] *and Quan Liu*[1]

[1] *School of Computer Science and Technology, Soochow University, Suzhou, China,* [2] *Provincial Key Laboratory for Computer Information Processing Technology, Soochow University, Suzhou, China*

OPEN ACCESS

Edited by:
Jiajie Peng,
Northwestern Polytechnical University, China

Reviewed by:
Pu-Feng Du,
Tianjin University, China
Vijaykumar Muley,
National Autonomous University of Mexico, Mexico

***Correspondence:**
Fei Zhu
zhufei@suda.edu.cn

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

Received: *21 January 2020*
Accepted: *07 April 2020*
Published: *05 May 2020*

Citation:
Li F, Zhu F, Ling X and Liu Q (2020) Protein Interaction Network Reconstruction Through Ensemble Deep Learning With Attention Mechanism. Front. Bioeng. Biotechnol. 8:390. doi: 10.3389/fbioe.2020.00390

Protein interactions play an essential role in studying living systems and life phenomena. A considerable amount of literature has been published on analyzing and predicting protein interactions, such as support vector machine method, homology-based method and similarity-based method, each has its pros and cons. Most existing methods for predicting protein interactions require prior domain knowledge, making it difficult to effectively extract protein features. Single method is dissatisfactory in predicting protein interactions, declaring the need for a comprehensive method that combines the advantages of various methods. On this basis, a deep ensemble learning method called EnAmDNN (Ensemble Deep Neural Networks with Attention Mechanism) is proposed to predict protein interactions which is an appropriate candidate for comprehensive learning, combining multiple models, and considering the advantages of various methods. Particularly, it encode protein sequences by the local descriptor, auto covariance, conjoint triad, pseudo amino acid composition and combine the vector representation of each protein in the protein interaction network. Then it takes advantage of the multi-layer convolutional neural networks to automatically extract protein features and construct an attention mechanism to analyze deep-seated relationships between proteins. We set up four different structures of deep learning models. In the ensemble learning model, second layer data sets are generated with five-fold cross validation from basic learners, then predict the protein interaction network by combining 16 models. Results on five independent PPI data sets demonstrate that EnAmDNN achieves superior prediction performance than other comparing methods.

Keywords: protein-protein interaction network, protein-protein interaction, ensemble learning, deep learning, attention mechanism, multi-layer convolutional neural network

INTRODUCTION

Protein interactions and interaction networks take part in vital activities of each living cell, including signal transduction, immune response, metabolism of energy substance, cell cycle control, etc. (Keskin et al., 2016). The exact identification of protein interactions is therefore important not only to understanding the functions of proteins but also to structure-based drug design and treatment of diseases (Li et al., 2009).

Majority of existing methods for predicting PPI are based on Gene Ontology and annotations, phylogenetic profile, gene fusion, the interacting proteins co-evolution pattern and the similarity of proteins in sequence, structure, domain and subcellular localization (Boxem et al., 2008; Zhang et al., 2012; Planas-Iglesias et al., 2013; Sun et al., 2017). However, as their accuracy and reliability depend heavily on collected prior knowledge, they are hardly applied widely. Several methods based on amino acid sequence computation have been explored to predict PPI, such as support vector machine with traditional auto-correlation, k-nearest neighbor (kNN) with local description (LD) (Yang et al., 2010), support vector machine (SVM) with conventional auto covariance (AC) (Guo et al., 2008) or local description (LD) (Zhou et al., 2011), deep neural network with amphiphilic Pseudo amino acid composition (PseAAC) descriptor (Du et al., 2017b) and so on. The above methods provide different techniques of protein sequences such as AC, LD, MCD, PseAAC, with each technique extracting different feature information of protein interactions (Zhang et al., 2019a). AC and CT considered the physical properties of amino acids and their dipole and side-chain volumes respectively. Then LD uses triples to describe composition, transition and distribution of sequence, while PseAAC further studies order information of sequences. We propose to combine different descriptors to achieve PPI prediction to obtain more information from protein interactions.

Ensemble learning is a machine learning method, which uses a series of learners and uses some rules to integrate the learning results so as to obtain better performance than a single learner. And ensemble learning has broad application prospects in many fields such as protein phosphorylation site prediction, genome function prediction and cancer prediction in bioinformatics (Gomes et al., 2017; Krawczyk et al., 2017). The previous works also demonstrate the effectiveness of classifier ensemble and provide some guidelines to generate an ensemble classification model (Martin et al., 2005; Han and Huang, 2006; Huang and Zheng, 2006; Huang and Du, 2008). Wang used a boosting technique to generate multiple classifiers iteratively to solve the problem of imbalance between positive and negative data when predicting the phosphorylation sites (Wang et al., 2017). Wang took a random forest and voting method as a basic classifier integration strategy separately to predict PPI sites (Wang et al., 2019). You et al. (2019) chose the basic classifiers with optimal performance, leaving the classifiers with small differences and using the max-wins voting (MWV) strategy to predict DNA binding proteins. Zhang et al. (2019a) trained 27 models by combining AC, MCD, LD with 9 DNN models of different configurations, and integrated these models through Double-layers BP Neural Network.

Furthermore, when exploring protein interactions and interaction networks, it is nonnegligible to quantify the interaction/non-interaction relationship between two proteins. One solution is to directly concatenate the features of the two proteins to form a feature vector (Zhang et al., 2019b), which lacks the information characteristics of the interaction/non-interaction between two proteins; another solution is to extract two features with two different networks and combine the features to form a new feature vector as the input of the model (Du et al., 2017b; Hashemifar et al., 2018), which is incapable of learning inherent relation of the proteins. Recently in natural language processing domain, researches have shown that attention mechanisms can effectively emphasize the relatively important parts of the input sentences and help boost the performance of relation extraction (Chen et al., 2017; Du et al., 2017a). In bioinformatics, attention mechanism is also adopted in chemical-protein interaction (CPI) (Zhang et al., 2019b), kinase-specific phosphorylation site prediction (Wang et al., 2017) and so on. In Xuan et al. (2019) model, exploiting the attention mechanism module to learn features or extract the relationship between IncRNA and disease provides more information. Wang et al. (2017) designed a two-dimensional independent attention mechanism for predicting phosphorylation sites, which enabled the model, called MusiteDeep, to automatically search important positions of the output sequences to estimate the contribution of each element in the sequences and feature dimensions. However, the above researches concern only single attention mechanism in the deep neural network model, which can be replaced by the multi-head attention mechanism that can exert attention multiple times and divide attention information into multiple heads. Liu et al. (2018) integrated attention pooling

TABLE 1 | Division of amino acid into seven groups based on the dipoles and volumes of the side chains.

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7
A, G, V	C	F, I, L, P	M, S, T, Y	H, N, Q, W	K, R	D, E

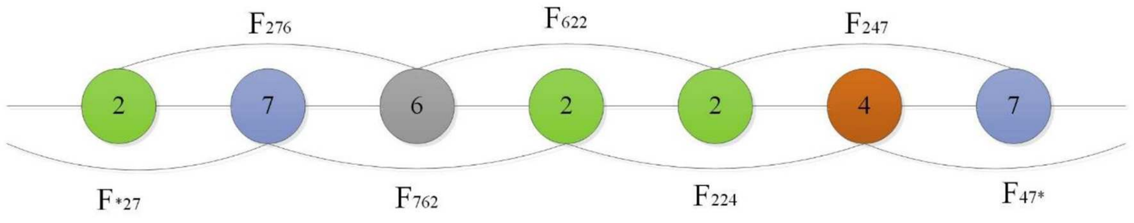


FIGURE 1 | For a grouped sequence “2762247,” the numerical code string of consecutive amino acids are “276,” “762,” “622,” “224,” “247,” and “*27,” “47*” according to Shen, and “*” is considered to be the first or second amino acid of an amino acid in a continuous amino acid. So its triad types are F276, F762, F622, F224, F427.

into the gated recurrent unit (GRU) model to extract CPIs. Verga et al. (2018) combined the Multi-head attention with convolution neural networks to construct a transformer model to extract the document-level biomedical relations. Thus, a multi-head attention mechanism will make it easier to capture the relevant important information for deep neural networks in PPI extraction.

Motivated by attention mechanisms and ensemble learning, we propose an algorithm called EnAmDNN, which at first extracted the biophysical-chemical information of protein sequences through AC, CT, LD, and PseAAC and association with the interactive description of each protein in protein interaction network; then it automatically extracted the protein features by multi-layer convolutional neural network, adopted attention mechanism to analyze deep-seated relationship of proteins and then forms the feature vectors. In EnAmDNN, 16 kinds of DNN models are trained through 4 characteristic bases which are the inputs of 4 DNNs with different layers and different neurons. In the integration module, the outputs of 16 DNNs are taken as the inputs of deep neural networks finally, and the five-fold cross validation is adopted to comprehensively predict protein interactions and interaction networks. Our contributions can be summarized as follows: (1) the new network structure can automatically extract highly abstract representations and detect the sequence specificity of proteins; (2) the attention mechanism is adopted to analyze internal links between the two proteins and the network description of each protein, instead of directly concatenating the two proteins, to improve the prediction accuracy; (3) ensemble learning considers the advantages of different descriptors and different DNNs to achieve comprehensive learning.

PRELIMINARIES

Deep Neural Network

It turns out that deep neural network (DNN) plays an important role in bioinformatics (Alipanahi et al., 2015; Zhou and Troyanskaya, 2015; Liu et al., 2016), i.e., predicting inner-organization and trans-organization RNA splicing patterns (Leung et al., 2014). DeepMind applied DNN to the detection of sequence specificity of the DNA-RNA binding protein, which is superior to other methods (Alipanahi et al., 2015); DeepSEA applied DNN to learn the code of regulatory sequences from chromatin map sequences in order to discern priorities of other functional varieties (Zhou and Troyanskaya, 2015); other examples include genome informatics extraction, detection of protein structure and medicine discovery. In short, compared with other sequence-based methods, DNN has the following advantages: (1) it can automatically learn certain protein sequences; (2) it can reduce the influence of noise on the raw data and extract the hidden high dimension representation (Bengio et al., 2013). However, the performance of DNN is closely related to the network configuration and may vary greatly for different configurations.

Protein Representation Technique

Different representation techniques of protein features may have a strong impact on the performance of PPI prediction, making it a challenge to effectively express the protein features and describe the connections of two proteins. We choose four representative protein techniques instead of one to avoid the limitation brought by a single technique.

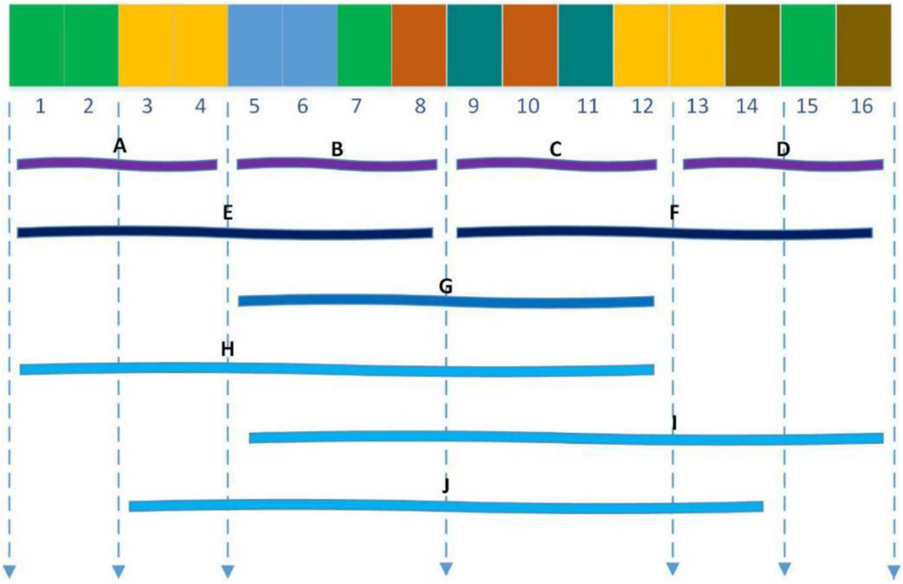


FIGURE 2 | Ten regions **(A–J)** of the entire protein sequence. The regions **(A–D)** are generated by dividing the whole sequence into four equal regions, and regions **(E,F)** are generated by dividing the whole sequence into two equal regions. The region **(G)** stand for the central 50% of the entire sequence. And the regions **(H–J)** stand for the first 75%, the final 75% and the central 75% of the entire sequence respectively.

Auto Covariance Technique

Two proteins interact with each other through electrostatic, hydrophobic, steric and hydrogen bond, which can be reflected by the seven physicochemical properties of amino acids, including hydrophobicity (H_1), hydrophilicity (H_2), volumes of side chains of amino acids (VSC), polarity (P_1), polarizability (P_2), solvent-accessible surface area and net charge index of side chains. The above properties are exploited by the auto-covariance method to transform amino acid sequence into uniform matrices which reveal the special connection of two residues under a certain distance and are widely applied in protein-encoding. For example, a protein sequence of length L is calculated as follows (Guo et al., 2008):

$$AC(lag, j) = \frac{1}{L - lag} \sum_{i=1}^{L-lag} (X_{i,j} - \frac{1}{L} \sum_{i=1}^{L} X_{i,j}) \times (X_{(i+lag),j} - \frac{1}{L} \sum_{i=1}^{L} X_{i,j}) \quad (1)$$

Xij represents the *j*-th physical property of the *i*-th amino acid in the protein sequence; *lag* represents the distance between residues; then proteins of various lengths are encoded as vectors of equal length $lg * $ p, where lg is the maximum *lag* (lag = 1, 2, ..., lg), p is the number of physical properties. In this study, p was 7, reflecting the characteristics of the seven amino acids. As with Guo, we set the log to 30 (Guo et al., 2008). Therefore, each protein sequence is represented as a 210-dimensional vector.

Conjoint Triad Technique

Shen et al. (2007) introduced a conjoint triad technique to represent sequence information of each protein, in which any three contiguous amino acids are regarded as a unit and the characteristics of one amino acid and its vicinal amino acids are fully considered. First, the conjoint triad divides 20 standard amino acids into 7 groups according to their dipole and side-chain volumes, then the triads can be distinguished according to the type of amino acid. According to Shen's settings, there are 343 (7 × 7 × 7) triad types (Shen et al., 2007), as shown in **Figure 1**.

Finally, the PPI information of protein sequences are projected into the homogeneous vector space according to the

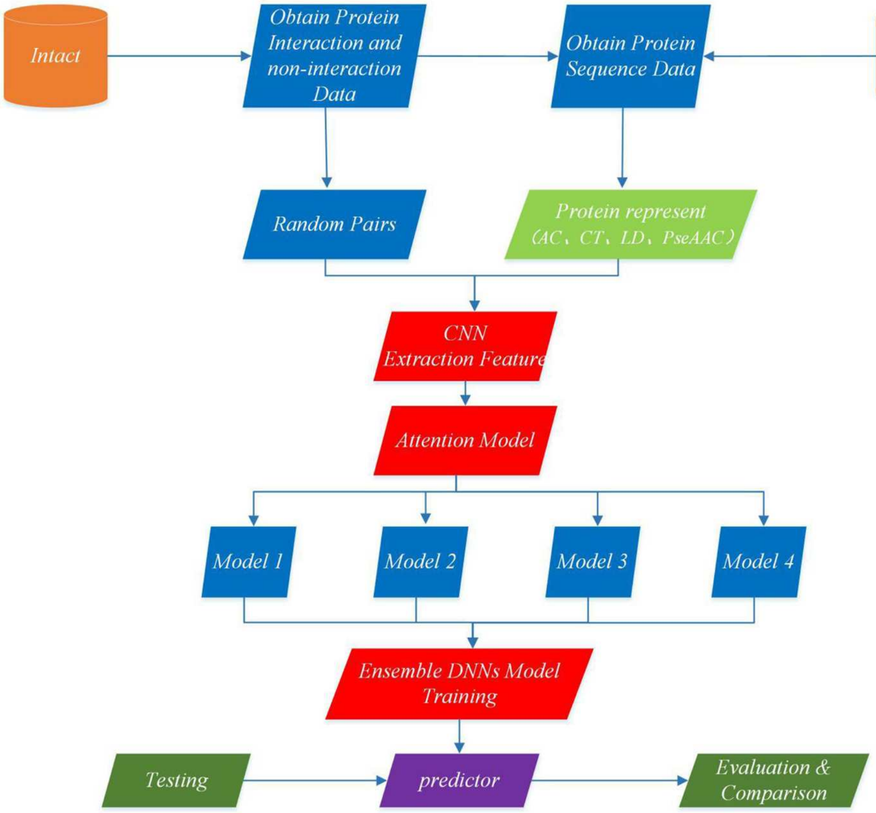

Intact
Obtain Protein Interaction and non-interaction Data
Obtain Protein Sequence Data
Uniprot
Random Pairs
Protein represent (AC、CT、LD、PseAAC)
CNN Extraction Featur
Attention Model
Model 1
Model 2
Model 3
Model 4
Ensemble DNNs Model Training
Testing
predictor
Evaluation & Comparison

FIGURE 3 | Flowchart of EnAmDNN for predicting protein-protein interactions. First, the interaction pairs and non-interaction pairs of related proteins are obtained from IntAct, and all protein sequence data of UniProt are obtained; the appropriate proportion of interaction pairs and non-interaction pairs are selected, and each group of protein pairs (including interaction pairs and non-interaction pairs) is vectorization by AC, CT, LD, and PseAAC techniques; put vector protein into convolution neural network for feature extraction of each protein; extracted features are transferred to attention mechanism module for deep analysis of interaction between each group of protein pairs; then the analyzed features are input into deep neural network of different models for training; finally, the final prediction results are obtained by integrating the prediction results of different models.

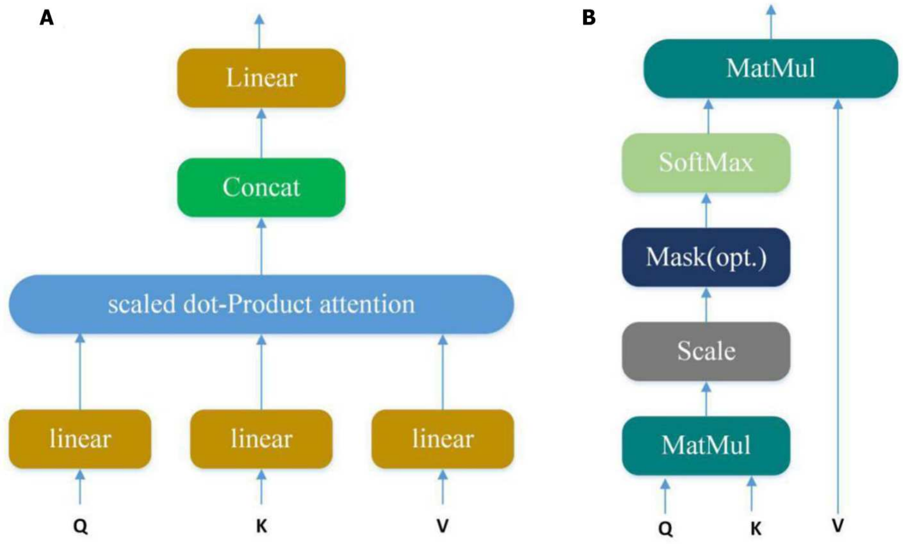


FIGURE 4 | (A) Multi-Head Attention consists of several attention layers. (Vaswani et al., 2017) First, query, key and value go through a linear transformation, and then enter them into scaled dot-Product attention to generate many heads; then concatenate these heads to keep relevant information in different representation subspaces. **(B)** Scaled dot-product attention (Vaswani et al., 2017). Obtain weights by similarity calculation between query and each key, and the weights are normalized by softmax function; then attention is obtained by the weight and the corresponding value.

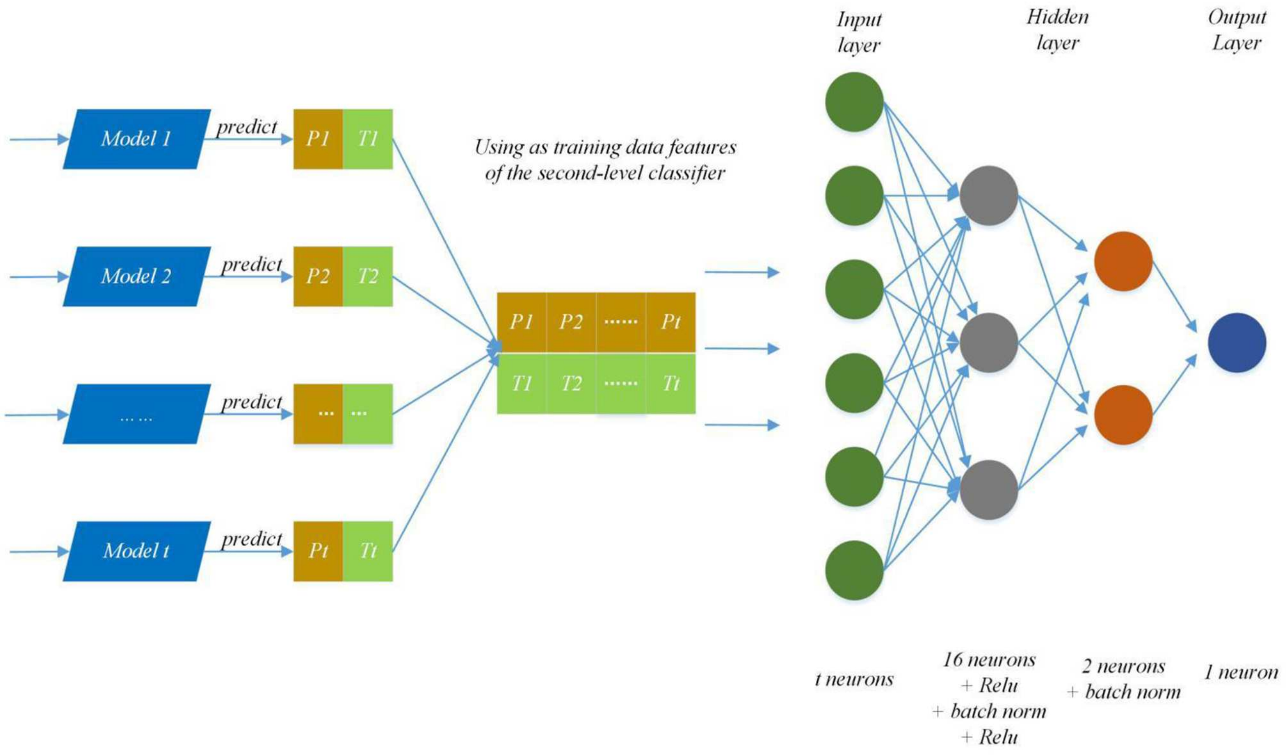


FIGURE 5 | Ensemble strategy composed of deep neural networks. The first layer results (P_n, T_n) ($0 < n < t+1$) are predicted by T primary learners, where P_n and T_n stand for training data and the prediction result; then use recombined (P_n, T_n) as training data features of the second-level classifier and put it into deep neural networks to predict protein interaction networks.

frequency of each triad type, where each protein is represented by a 343-dimensional vector.

Local Descriptor Technique

The Local descriptor technique (Zhou et al., 2011) also divided 20 standard amino acids into 7 groups as shown in **Table 1** and divided the entire protein sequence into 10 regions as shown in **Figure 2**. For each sub-sequence, three descriptors, composition (C), transition (T) and distribution (D), are applied to describe its trait where C represents the proportion of each amino acid group; T represents the frequency with which amino acids in one group are followed by amino acids of another group; D measures the proportion of chain length where the top 25, 50, 75, and 100% of the amino acids of a particular group are located. For the local descriptor method, each region produces 63 values, where C represents 7, T represents 7, and D represents 35, and then each protein is encoded as a 630-dimensional vector.

For example, according to **Table 1**, the sequence "ACLACLCCLAALLCCCLALALAAALL" is converted into the amino acid group "12312322311332321311313111333133" so that the sub-sequence contains 9 "1", 7 "2," and 10 "3."For feature C, 9/(9 + 7 + 10) = 0.3461, 7/(9 + 7 + 10) = 0.2693, 10/(9 + 7 + 10) = 0.3846; for feature T, there are 2 cases that "1" is converted to "2" or "2" is converted to "1," then 2/25 = 0.08; similarly, transitions between "3" and "1" as well as "2" and "3" are 3/25 = 0.12, 6/25 = 0.24, respectively; for feature D, there are nine "1"s, then the D descriptor for 1 is 1/26 = 0.0384, [0.25*9 + 0.5]/26 = 0.0769, [0.5*9 + 0.5]/26=0.1923, [0.75*9 + 0.5]/26 = 0.2692, [1*9 + 0.5]/26 = 0.3462.

Pseudo Amino Acid Composition (PseAAC) Technique

Tian et al. (2019) used a sequence encoding technique based on pseudo amino, that is, PseAAC. Given a protein sequence P with L amino acid residues:

$$S_1S_2S_3S_4.......S_L$$

where S_i represents the ith residue of the protein P, $1 \leq i \leq L$.

According to the PseAAC technique, the protein P can be formulated as

$$P = [x_1, x_2, ..., x_{20}, x_{21}, ..., x_{20+\lambda}]^T, \ (\lambda < L) \tag{2}$$

where the $20 + \lambda$ components are given by

$$x_k = \begin{cases} \dfrac{f_k}{\sum\limits_{i=1}^{20} f_i + \omega \sum\limits_{j=1}^{\lambda} \theta_j}, (1 \leq k \leq 20) \\ \dfrac{\omega \theta_{k-20}}{\sum\limits_{i=1}^{20} f_i + \omega \sum\limits_{j=1}^{\lambda} \theta_j}, (21 \leq k \leq 20 + \lambda) \end{cases} \tag{3}$$

TABLE 2 | AD with various methods.

Methods	Accuracy	Recall	Precision	F1	AUC
EnsAmDNN	**0.9467**	**0.9329**	**0.9541**	**0.9433**	**0.9463**
SVM_AC	0.7886	0.6369	0.8956	0.7435	0.7831
kNN_LD	0.7549	0.7507	0.7602	0.7542	0.7556
SVM_LD	0.804	0.6735	0.9174	0.7754	0.805
NDDs_APAAC	0.898	0.9093	0.8907	0.8993	0.8977
EnsDNN	0.9372	0.9255	0.9531	0.9388	0.938

The highest score of each evaluation criteria is emphasized in bold.

TABLE 3 | PD with various methods.

Methods	Accuracy	Recall	Precision	F1	AUC
EnsAmDNN	**0.895**	**0.8568**	0.9275	**0.8903**	**0.8951**
SVM_AC	0.797	0.62	0.9373	0.7444	0.7904
kNN_LD	0.8057	0.8042	0.7906	0.7958	0.8064
SVM_LD	0.8315	0.7337	0.9084	0.8108	0.831
NDDs_APAAC	0.8773	0.8247	0.9058	0.8627	0.8906
EnsDNN	0.8917	0.8433	**0.9311**	0.8846	0.8915

The highest score of each evaluation criteria is emphasized in bold.

TABLE 4 | Cancer with various methods.

Methods	Accuracy	Recall	Precision	F1	AUC
EnsAmDNN	**0.8502**	**0.8062**	**0.8863**	**0.8436**	**0.8508**
SVM_AC	0.6524	0.4848	0.7347	0.5811	0.6545
kNN_LD	0.6475	0.6761	0.6202	0.6458	0.6471
SVM_LD	0.6673	0.5591	0.715	0.6263	0.6681
NDDs_APAAC	0.7551	0.7224	0.7764	0.7474	0.7555
EnsDNN	0.8008	0.7549	0.8362	0.7925	0.802

The highest score of each evaluation criteria is emphasized in bold.

TABLE 5 | Cancer with various methods.

Methods	Accuracy	Recall	Precision	F1	AUC
EnsAmDNN	**0.907**	0.8523	**0.96**	**0.9018**	**0.9088**
SVM_AC	0.7356	0.6316	0.806	0.7037	0.7371
kNN_LD	0.7671	0.7246	0.7934	0.7565	0.7675
SVM_LD	0.7819	0.7545	0.7962	0.774	0.7816
NDDs_APAAC	0.8454	0.8326	0.837	0.8339	0.8446
EnsDNN	0.9039	**0.8747**	0.9252	0.899	0.9034

The highest score of each evaluation criteria is emphasized in bold.

TABLE 6 | Diabetes with various methods.

Methods	Accuracy	Recall	Precision	F1	AUC
EnsAmDNN	**0.8333**	0.871	0.7941	**0.8308**	**0.8355**
SVM_AC	0.7891	0.6667	0.7941	0.7128	0.7819
kNN_LD	0.7568	0.8571	0.75	0.8	0.7411
SVM_LD	0.7813	0.8269	0.7361	0.7742	0.7885
NDDs_APAAC	0.803	**0.9118**	0.7561	0.8267	0.7996
EnsDNN	0.8030	0.7576	**0.8333**	0.7937	0.8030

The highest score of each evaluation criteria is emphasized in bold.

In Equation (3), $f_k(k = 1, 2, \ldots, 20)$ are the normalized occurrence frequencies of 20 amino acids in protein P; ω is the weighting factor set to 0.05 in general work; and $\theta_j(j = 1, 2, \ldots, \lambda)$ denotes the order relationship between two residues that are j residues apart, which is shown as follows:

$$\theta_j = \frac{1}{L-j}\sum_{i=1}^{L-j} J_{i,i+j}(j < L) \tag{4}$$

$$J_{i,i+j} = \frac{1}{3}\sum_{p=1}^{3}\left[H_p(A_{i+j}) - H_p(A_i)\right] \tag{5}$$

where J_{ij} denotes the order relationship function between amino acid A_i and A_j, $H_p(A_i)$ denotes the pth property of A_i. $H_1(A_i)$, $H_2(A_i)$ and $H_3(A_i)$ are the hydrophobicity value, hydrophilicity value and side-chain mass for the amino acid, respectively. This coding method contains more sequence characteristics because it considers not only the physicochemical properties of the protein but also the order information of sequences.

MATERIALS AND METHODS

Data Sets

We collect the dataset information of Parkinson's disease (PD), Alzheimer's disease (AD), cancer, cardiac and diabetes, whose interactive information is from IntAct database (Kerrien et al., 2007) and sequence information from Uniprot (Bairoch et al., 2004). We are concerned about positive-negative selection in our work. For the positive set, proteins and protein pairs that contain less than 50 amino acids and 40% of sequence identity are removed to eliminate the variance caused by minor bias proteins to the model. The negative set was obtained by pairing proteins whose subcellular localization is different (Guo et al., 2008) or GO Cellular Component (CC) and Biological Process (BP) ontology with experimental evidence codes (Muley and Ranjan, 2012). The subcellular location information on the proteins is extracted from Uniprot. According to this information, a protein can be divided into several types of localization cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, Golgi apparatus, peroxisome and vacuole. The way to construct negative set must meet the following requirements: (1) the non-interacting pairs cannot appear in the positive data set; (2) the contribution of proteins in the negative set should be as

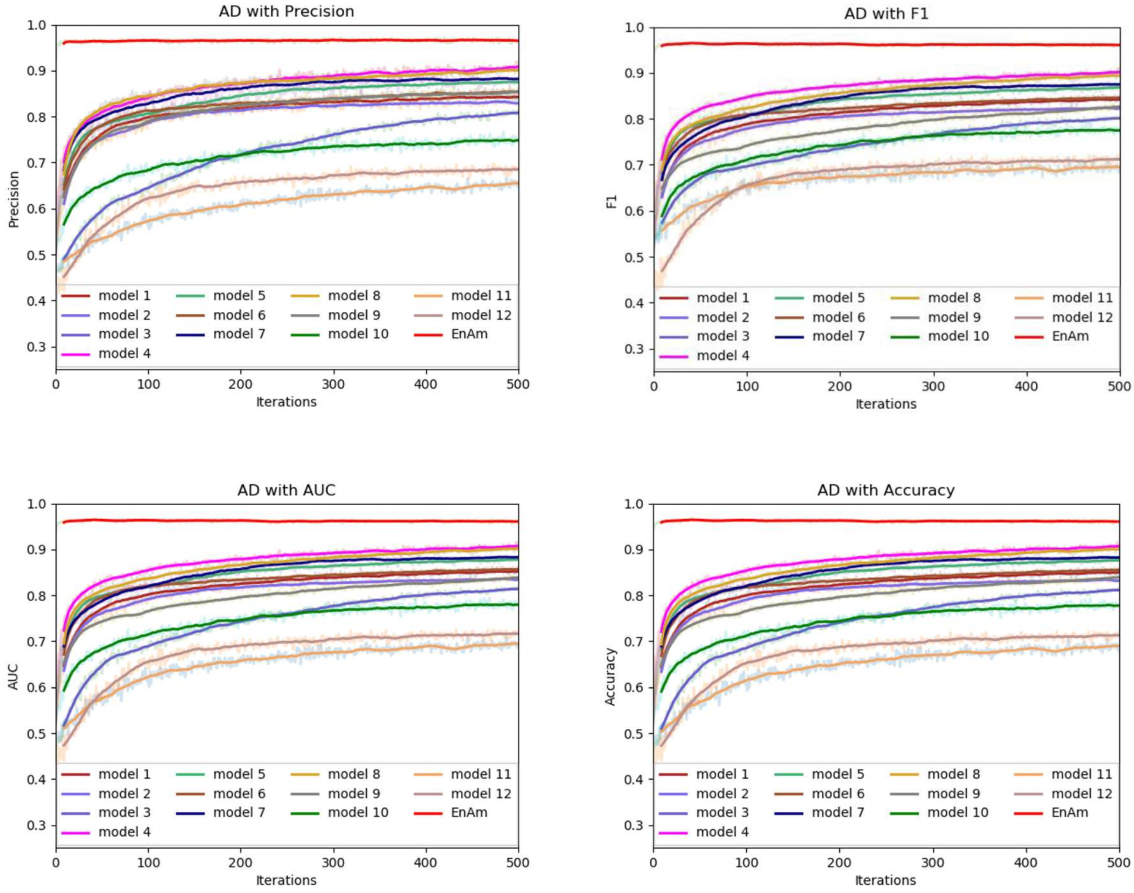


FIGURE 6 | Comparison of evaluation indexes of each basic model and ensemble model with AD data set.

harmonious as possible. In our work, the ratio between positive and negative set is 1:1, where the negative sets are randomly chosen from non-interactive pairs.

Finally, we have five independent PPI datasets: Parkinson's disease (PD) (4,127 interacting pairs and 4,127 non-interacting pairs), Alzheimer disease (AD) (4,096 interacting pairs and 4,096 non-interacting pairs), Cancer (6,352 interacting pairs and 6,352 non-interacting pairs), Cardiac (2,663 interacting pairs and 2,663 non-interacting pairs) and Diabetes (163 interacting pairs and 163 non-interacting pairs).

Evaluation Criteria

The following metrics are taken into account to perform evaluation: Overall Prediction Accuracy, Recall, Precision, F_1 score values, and Area under the ROC Curve (AUC) (Zhang et al., 2019a). The first four metrics are defined as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{F}_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{9}$$

where TP (true positive) is the number of true interacting pairs found in the positive data set, TN (true negative) is the number of true non-interacting pairs with correct prediction, FP (false positive) is the number of the predicted interacting pairs not found in the positive data set, and FN (false negative) is the number of the true interacting pairs with false prediction.

Ensemble Deep Neural Networks

This section describes EnAmDNN model that predicts PPI based on protein sequences, which consists of the input module, the convolution module, the attention mechanism module, the DNN module and the integration module. Each protein sequence is encoded, by the input module, through the protein representation techniques, as a vector, whose feature is extracted by the convolution module. Then, the internal link in the protein pair is detected through the attention module, and then the analyzed protein pair is provided to 16 dependent learners. After the training is completed, these learners will be integrated

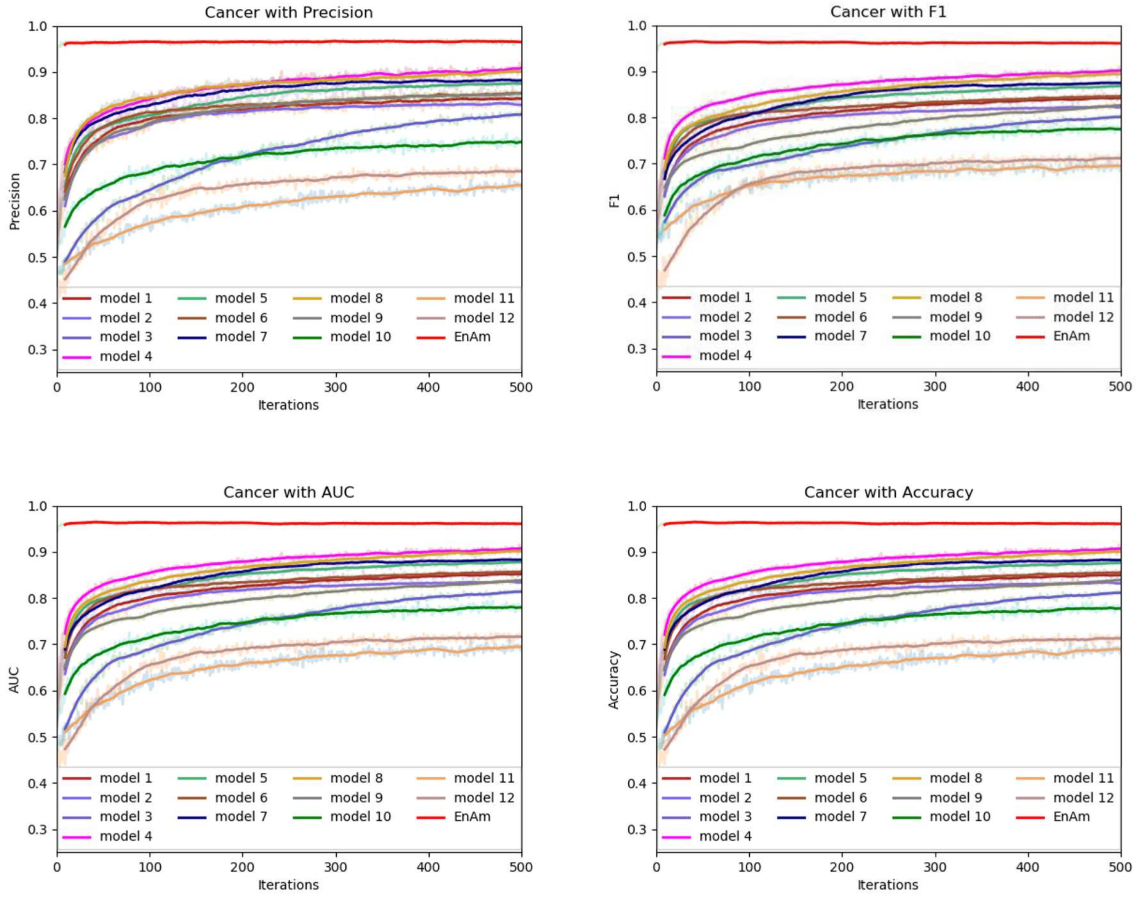


FIGURE 7 | Comparison of evaluation indexes of each basic model and ensemble model with Cancer data set.

through a two-layer hidden layer neural network. The working process of an EnAmDNN is shown as **Figure 3**.

Deep Convolutional Module

The convolution module is a batch of normalized layers, a stack of convolutional layers and activation layers, which can automatically extract features of vectorized protein sequences. In our model, the output of the convolution module is calculated by an expression that starts with a convolution layer and ends with a convolution layer:

$$In_{t+1} = RELU\left(Batch_{\beta,\delta}\left(Conv_{\lambda}\left(In_t\right)\right)\right) \quad (10)$$

$$Out = Conv_{\gamma}\left(In_n\right) \quad (11)$$

Repeat In n times and enter the result In_n into Equation 11. Where Out is the output vector, In is the input vector and β, γ, δ, λ are the parameters of batch normalization and convolution layers.

The convolution layer searches the sequences according to their input orders and outputs the corresponding features; the batch normalization layer takes in the feature vectors and normalizes their mean values and the variances; ReLU layer takes in the normalized vectors and introduces non-linearity to achieve complex function approximation. Then the above processes repeat n times to obtain the feature vector.

Attention Module

Convolution layer can automatically learn potential features from input sequences, but only a small part of these potential features are very important in PPI. In our model, we use the multi-attention mechanism to adjust the weight of the input sequences to further emphasize the relatively crucial features. Applying the attention multiple times may learn more important features than single attention and allowing the model to learn relevant information in different representation subspaces (Vaswani et al., 2017). It can be understood that attention selectively selects a small amount of important information that is beneficial to PPI from a large amount of information and focuses on important information, ignoring most of the insignificant information. We choose the mechanism of multiple attention rather than directly connecting the two protein eigenvectors to increase the exploration of protein pairs and further use the indirect relationship between residues to obtain more accurate information. The Multi-head attention calculates the output based on the query and a set of key-value pairs, where Q, K, V denote query, key, and value respectively. The specific structure is shown in **Figure 4**:

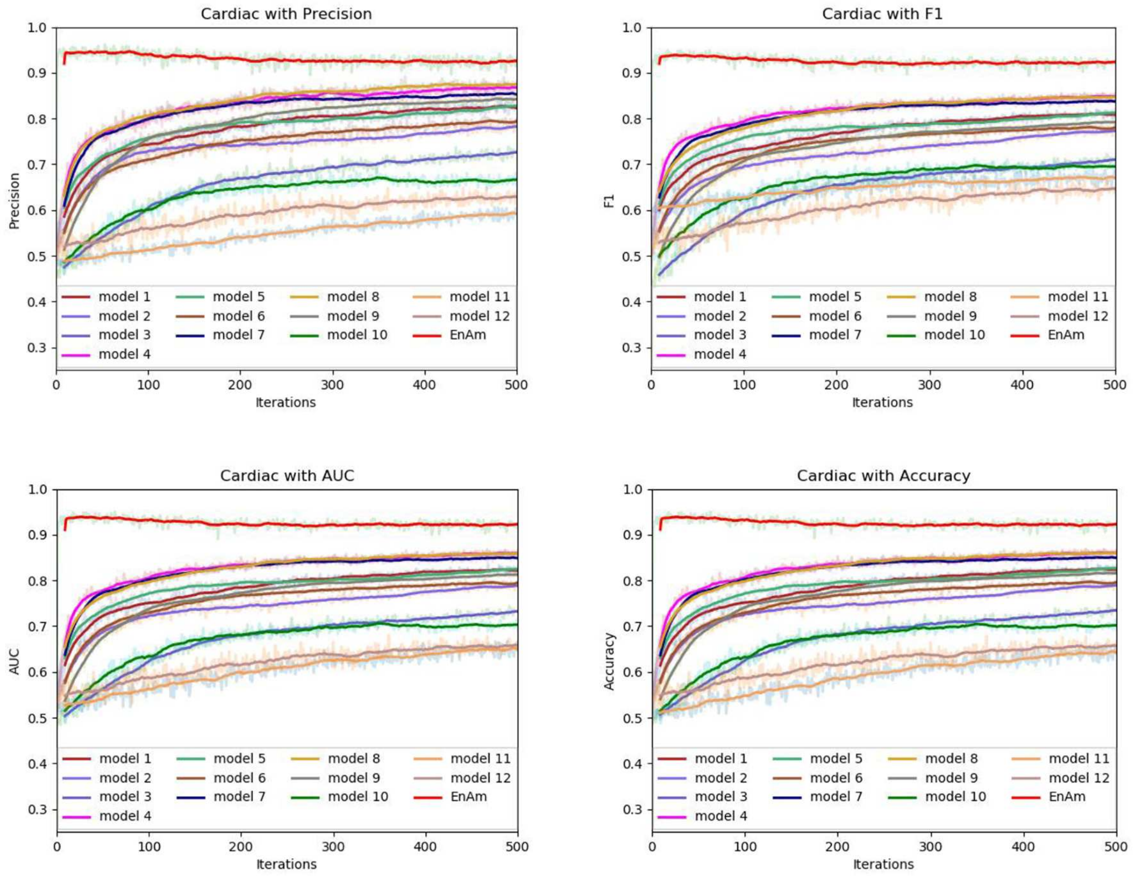


FIGURE 8 | Comparison of evaluation indexes of each basic model and ensemble model with Cardiac data set.

Query, key and value go through a linear transformation first, and then enter them into scaled dot-Product attention. At this time, the attention calculation formula is as follows:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK'}{\sqrt{d_k}})V \tag{12}$$

Where $\sqrt{d}$ is scaling factor. The core of the Multi-head attention is employing the above attention multiple times, and one time attention means one head. Suppose the Multi-head attention needs to be done h times to generate h heads, the $\mathrm{Att}_{\mathrm{head}}$ can be calculated as follows:

$$Att_{headi}(Q, K, V) = Attention(QW_i^Q, KW_i^K VW_i^V), 1 < i \leq h \tag{13}$$

where W_i^Q, W_i^K, W_i^V are parameter matrices. Finally, these heads are concatenated and once again linearly transformed by

$$MultiHead(Q, K, V) = Linear(Concat(Att_{head1}, ..., Att_{head\mathrm{h}})W^{\mu}) \tag{14}$$

In order to keep the invariance of features, we introduce average pooling and maximum pooling to reduce the errors caused by model parameters and retain information of global and local features.

$$\begin{aligned} &newMultiHead(Q, K, V) \\ &= Concat(AvgPool(MatMul(MultiHead(Q, K, V), Q)), \\ &MaxPool(MatMul(MultiHead(Q, K, V), Q))) \end{aligned} \tag{15}$$

where *AvgPool* is the function of average pooling and *MaxPool* is the function of maximum pooling.

For a protein pair ($\mathrm{P}_1, \mathrm{P}_2$), it is expressed as S_1, S_2 respectively after convolution layer. We use the merge layer to fuse the protein pairs that are redistributed by the attention mechanism. The calculation formula of the merge layer is as follows:

$$\begin{aligned} S_1{}' &= newMultiHead(S_1, S_2, S_1), \\ S_2{}' &= newMultiHead(S_2, S_1, S_2) \end{aligned} \tag{16}$$

$$Merge(S_1{}', S_2{}') = Concat(\frac{S_1{}' \cdot S_2{}'}{|S_1{}'| \times |S_2{}'|}, S_1{}'AS_2{}', S_1{}', S_2{}') \tag{17}$$

where A is weight.

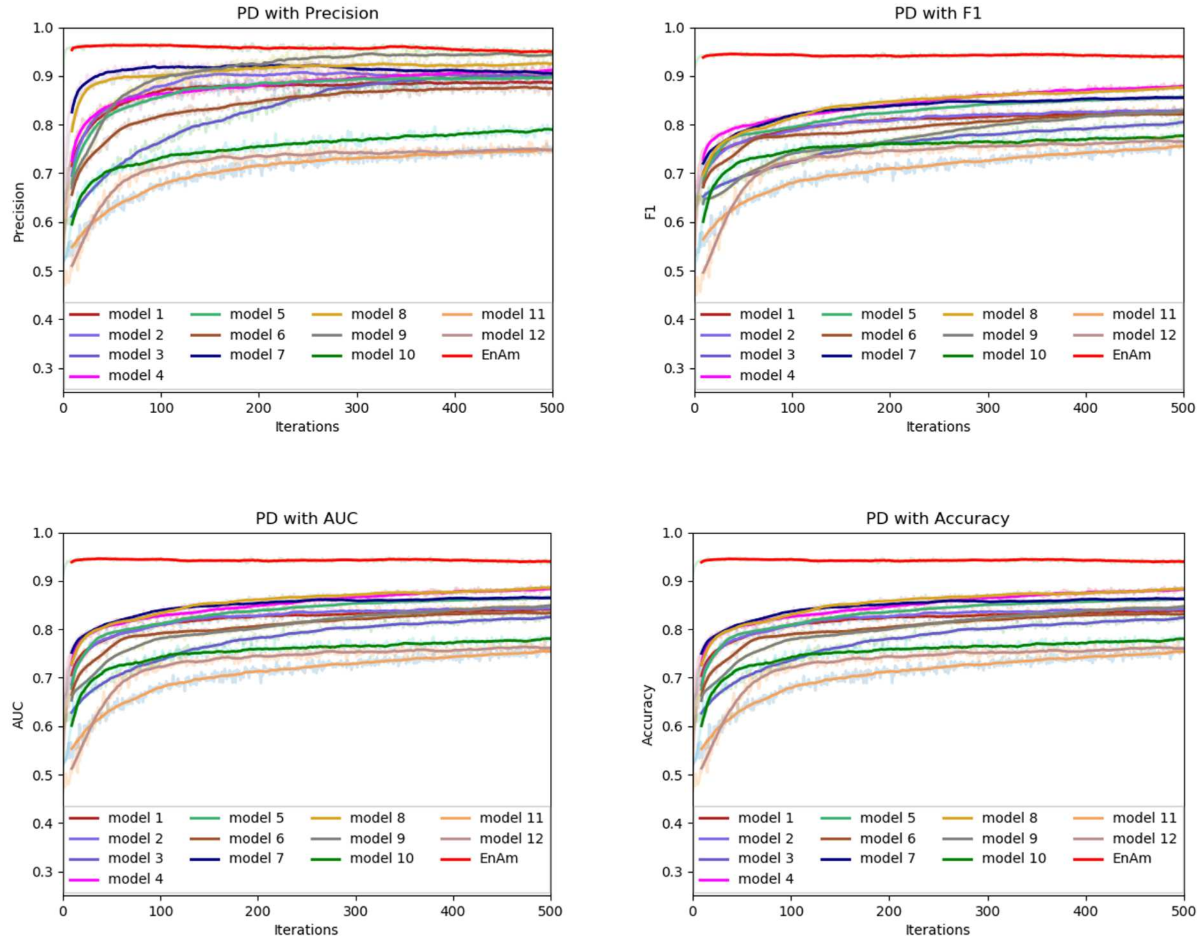


FIGURE 9 | Comparison of evaluation indexes of each basic model and ensemble model with PD data set.

The basic model of our work is mainly constructed by 3.3.1 the deep convolution module, 3.3.2 the attention mechanism module and the deep neural network. The specific basic learning algorithm is shown in algorithm 1.

Algorithm 1: Basic learning algorithm based on multi-head attention with pooling

Input:interaction data $D = \{x_i, y_i\}_{i=1}^{n}$, encoded protein sequence pair$x_i = (s_{i1}, s_{i2})$
Return: basic learning algorithm parameters W, b

```
for $i = 1$ to $n$:
    $S_1, S_2 \leftarrow x_i(s_{i1}, s_{i2})$ represent encoded protein $p_1$ and protein $p_2$
    for $iter = 1$ to $t$:
        $S_{1,iter+1} = RELU\left(Batch_{\beta,\delta}\left(Conv_{\lambda}\left(S_{1,iter}\right)\right)\right)$
        $S_{2,iter+1} = RELU\left(Batch_{\beta,\delta}\left(Conv_{\lambda}\left(S_{2,iter}\right)\right)\right)$
    end for
    $S_1 \leftarrow Conv_{\gamma}\left(S_{1,t}\right)$
    $S_2 \leftarrow Conv_{\gamma}\left(S_{2,t}\right)$
    calculate the importance value of single attention by
        $Attenhead_{S_1}^{i} = \text{softmax}\left(\frac{S_1 S_2{}'}{\sqrt{d_k}}\right) S_1,$
        $Attenhead_{S_2}^{i} = \text{softmax}\left(\frac{S_2 S_1{}'}{\sqrt{d_k}}\right) S_2$
    connect multiple attentions by $S_1^{*} = MultiHead(S_1, S_2, S_1), S_2^{*} = MultiHead(S_2, S_1, S_2)$
    Calculate new feature importance value of multi-head attention with pooling by
        $S_1^{**} = newMultiHead(S_1^*, S_2^*, S_1^*),$
        $S_2^{**} = newMultiHead(S_2^*, S_1^*, S_2^*)$
    merge the sequence feature pairs as the input of the network by $Concat(\frac{S_1^{**} \cdot S_2^{**}}{|S_1^{**}| \times |S_2^{**}|}, S_1^{**} A S_2^{**}, S_1^{**}, S_2^{**})$
    $output = \text{softmax}\left(W \cdot merge + b\right)$
end for
Return $W, b$
```

Ensemble Strategy

Ensembles of independent deep neural networks can improve the performance of a single network (Bairoch et al., 2004). In the Otto group product classification challenge, the first one won the championship by stacking 30 models. The model achieved remarkable results, and we also adopted the stacking method of ensemble learning in our work. Secondly, to predict the effect better, the trainers of each primary model keep stability and diversity as much as possible.

We modify the internal structure of the algorithm and learn from different feature representations, which are two strategies to maintain diversity and achieved improvement, so we also take the same measures (Zhang et al., 2019a). In practice, we choose four feature representations to quantify the characteristics of each protein and set different parameters of DNNS according to the characteristics of each representation. Then we use the stacking method to combines with five-fold cross validation, the primary learners are trained from the initial data set, and a new data set is generated by the primary learners for training the secondary learner. It means the output of each primary trainer is input as an example to the secondary trainer for fusion output and PPI prediction. Here, the secondary trainer is composed of deep neural networks. Its structure is shown in **Figure 5** and ensemble strategy is described in algorithm 2.

TABLE 7 | Comparison of EnAm-Con and EnAm-Sep.

Data sets	EnsAC	EnsLD	EnsCT	EnsPseAAC	EnsCon	EnAMDNN
AD	0.8502	0.9029	0.8949	0.7759	0.9338	**0.9456**
PD	0.8378	0.8788	0.8820	0.8059	**0.9242**	0.8951
Cancer	0.7638	0.7911	0.7889	0.6928	0.8468	**0.8508**
Cardiac	0.8201	0.8789	0.8888	0.7455	0.9020	**0.9088**
Diabetes	0.7516	0.7576	0.7917	0.6921	0.8459	**0.8787**

The highest score of each evaluation criteria is emphasized in bold.

RESULT

Comparing the Prediction Performance With Other Methods

All the experiment were carried out on a computer with CentOs, Cuda version 10.1.243, CuDnn version 7.0 and software environment python3.7+keras2.0+torch1.3.

In order to evaluate the performance of EnAmDNN, we compared it with the approaches proposed by Guo et al. (2008), Zhou et al. (2011), Du et al. (2017b), Yang et al. (2010), Zhang et al. (2019a) emphasized the highest score of each evaluation criteria in bold and present the results in **Tables 2–6**, which separately utilize AC, LD, CT, APAAC, PseAAC to encode amino acid sequence, and predicted PPI with SVM, k-nearest neighbor (kNN) or DNNs, all of which share the same training sets and the same testing sets. It can be seen from **Table 2** that EnAmDNN generally outperforms these predictors, where EnAmDNN achieved optimal prediction performance in all the datasets, especially in AD, with an accuracy of 94.67%, and a recall rate of 93.29%. The accuracy is 95.41%, F1 is 94.33%, and AUC is 94.63%. This is because, in EnAmDNN, feature representations in protein sequences are coordinated, and new features are obtained through different classifiers. Compared with the recent EnsDNN model, in five independent data sets, the AUC index DnAmDNN has increased by 0.89, 0.41, 0.61, 0.6, 3.90%, and the accuracy of PPI prediction are relatively high. The EnAmDNN model takes advantage of the multi- head attention mechanism, that is, extracts the internal links of the PPI, thereby improving the performance of the model.

To further demonstrate the effect of ensemble strategy, five-fold cross-validation is employed to improve the reliability of the results. **Figure 6** shows the performance of each basic learner, where it can be observed, taking AD dataset as an example,

Algorithm 2: EAM algorithm

Input: interaction data; basic learning algorithm $\psi_1, \psi_2, \ldots, \psi_M$; L layers; learning rate η; Max iterations and iteration threshold ε; class k$\in$ {0, 1}

Return: ensemble strategy parameters W, b

```
initialize weight matrix $W$ and Bias $b$, $iter$ as 1
extract protein sequence from uniprot
vectorize protein sequence by AC, LD, CT, PseAAC
divide training data $TrD = \{D_i\}_{i=1}^{K} = \{x_i, y_i\}_{i=1}^{n}$ and testing data $TeD$
for $m = 1$ to $M$ do
    for $k = 1$ to $K$ do
        train $h_m = \phi_m(\bar{D}_k)$ with $\bar{D}_k$
        predict $D_k$ and require predicted value $P_{mk} = h_m(D_k)$
        predict $T$ and and require predicted value $T_k = h_m(T)$
    end for
    splicing predicted value of training data through $m$th basic model by $P_m = (P_{m1}, P_{m2}, ..., P_{mK})^T$
    splicing predicted value of testing data through $m$th basic model by $T_m = (T_1 + T_2 + ... + T_K)/K$
end for
construct new training data by
and construct new testing data by $newT = (T_1, T_2, ..., T_M)$
Repeat:
    $iter$+1
    for i = 1 to $n$:
        set $a^{(1)} = x_i^*$
        for $l$=2 to $L$: $a^l = \sigma(W^l a^{l-1} + b^l)$
        calculate probability $p_k$ divided into class k
        calculate $loss^L = -\sum_{k=0}^{1} y_k^L \cdot \log(p_k)$
        for $l$=2 to $L$:
    end for
    for $l$=2 to $L$:
        $\nabla W^l = \mu \sum_{i=1}^{n} loss^l (a^{l-1})^T, \nabla b^l = \mu \sum_{i=1}^{n} loss^l$
        $W^l \leftarrow W^l - \nabla W^l, b^l \leftarrow b^l - \nabla b^l$
    end for
Until $\nabla W, \nabla b < \varepsilon$
Return $W$, $b$
```

that each basic learner, associated with five-fold cross-validation method, shows fairish prediction performance, which is reflected on all the evaluation criteria Accuracy, Recall, Precision, F1, AUC. The result indicates that our model extracts and trains the features produced by basic learners and that the shortcoming of each basic learner is overcome to a certain degree. It's also confirmed with PD, Cardiac and Cancer in **Figures 7–9**.

Performance of PPI Prediction

To further study the effectiveness of the ensemble strategy, we designed two different network architectures: (a) concatenating four feature representations (AC, LD, CT, PseAAC) as the input to the first layer classifiers (namely EnAm-Con) and (b) separately taking one feature representation as the input to the first layer classifiers (namely EnAm-Sep including EnsAC, EnsLD, EnsCT, EnsPseAAC). EnAm-Con first concatenates four feature representations and then integrated 12 trained DNNs in the same way as EnAmDNN. For EnAm-Sep, we separately trained 12-model DNNs based AC, LD, CT, and PseAAC, and integrated these DNNs in the same way as EnAmDNN. The performance of EnAm-Con and EnAm-Sep which emphasized in bold are also listed in **Table 7** where it can be observed that the LD method performs better than AC and PseAAC method. The LD method of AUC value obtained from the first four data sets are 6.2, 4.89, 3.57, 7.17, 0.8 and are 16.37, 9.05, 14.19, 17.89, 9.46% higher than AC and PseAAC methods separately, which is because LD can encode more interaction information. It can be seen from **Table 6** that EnAm-Con performs better than EnAm-Sep, proving that concatenating different feature representations as new feature vector can improve the accuracy of the ensemble strategy. It can be concluded that these four representations are complementary to each other and our ensemble strategy is effective and feasible.

The number of basic learners greatly influences the overall prediction performance, where the efficiency of the model continues to grow as the number of learners increases, to a point that the performance tends to be stable. To evaluate the influence of the EnAmDNN, we assign different numbers of DNNs to protein represent technique, such as 1, 3, 5, 7, 9. The result is presented in **Figure 10**, where it can be observed that the AUC of the EnAmDNN tends to be stable when the number reaches 16. The efficiency of the EnAmDNN may also be affected by the performance of each basic learner, for which we prepare 16 basic learners, iterate them for 600 times, and combine them through deep neural network. The result is shown as follows. It can be seen that the prediction performance improves as the iteration continues and the model tends to remain stable at the point of 200.

Table 8 reports the process running time of EnAmDNN based on fold cross-validation with 16 basic learners and iterate each basic learner for 600 times.

Meanwhile, to further investigate the contribution of using an ensemble predictor with fold cross-validation, we integrated the simplified EnAmDNN, which don't use fold cross-validation. To reduce the impact of data dependency in the experiment, we constructed data sets on Cardiac based on Muley and Ranjan (2012) to observe the performance of proposed model. From **Table 9**, we can see that EnAmDNN achieves competent prediction performance with an average accuracy of 85.66%, precision of 89.47%, F_1 of 85.16%, and AUC of 85.76. It has better performance than simplified EnAmDNN across evaluation metrics. The prediction results show that EnAmDNN with fold cross-validation is capable of predicting PPIs.

CONCLUSIONS

In this paper, we propose an ensemble deep learning framework (EnAmDNN) with an attention mechanism that aims to predict protein interaction networks. EnAmDNN firstly extracts the

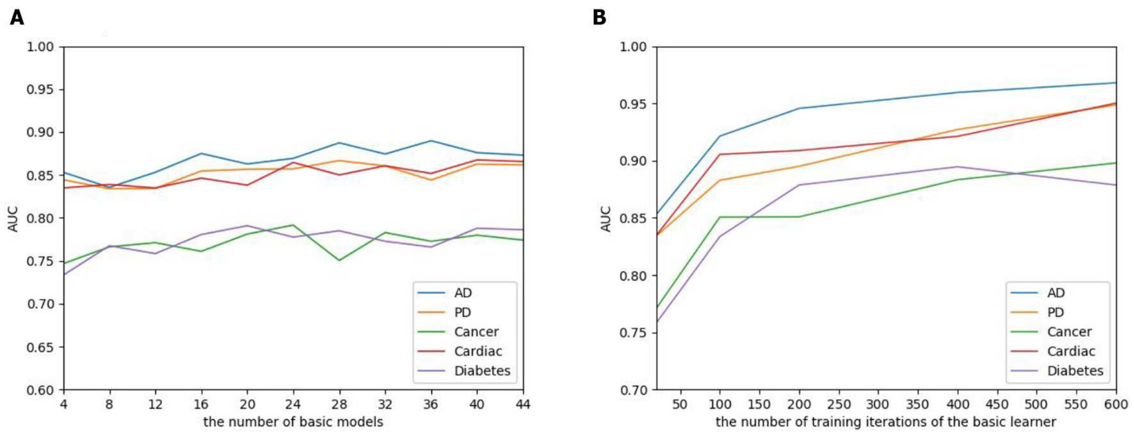


FIGURE 10 | (A) The influence of the number of basic learners on the EnAmDNN. It can be observed that the AUC of the EnAmDNN tends to be stable when the number reaches 16. **(B)** The influence of the number of training iterations of the basic learners on the EnAmDNN. It can be seen that the model tends to remain stable at the point of 200.

TABLE 8 | Running time of EnAmDNN based on fold cross-validation.

Date sets	AD	PD	Cancer	Cardiac	Diabetes
Time (s)	1,273,589.80	835,292.29	2,618,297.11	1,145,029.32	101,784.60

TABLE 9 | Comparison of EnAmDNN and simplified EnAmDNN.

Models	Accuracy	Recall	Precision	F1	AUC
EnAmDNN	0.8749	0.8138	0.9337	0.8688	0.8765
Simplified EnAmDNN	0.8566	0.8142	0.8947	0.8516	0.8576

feature information of protein sequences through AC, LD, CT, and PseAAC, and projects the information into various feature spaces to segment information of AC, LD, CT, PseAAC amino acid from different perspective; then the multi-head attention mechanism is adopted to capture the internal connections of interactions; each technique is assigned 4 independent DNNs with different configurations, resulting in 16 basic learners, and finally combined by deep neural network. To further evaluate the prediction performance of EnAmDNN, we apply it to 5 independent data sets, where improvements of various degrees can be observed for indicators AUC, ACC, Recall, Precision, F1, from which it can be concluded that EnAmDNN learns better than previous approaches from different DNNs and representations.

DATA AVAILABILITY STATEMENT

Alzheimer disease data was downloaded from the IntAct database (https://www.ebi.ac.uk/intact/query/annot:%22dataset:alzheimers%22?conversationContext=6) under search term annot: "dataset:alzheimers." Cardiac data was downloaded from the IntAct database (https://www.ebi.ac.uk/intact/query/annot:%22dataset:cardiac%22?conversationContext=7) under search term annot: 'dataset:cardiac." Diabetes data was downloaded from the IntAct database (https://www.ebi.ac.uk/intact/query/annot:%22dataset:diabetes%22?conversationContext=8) under search term annot: "dataset:diabetes." Parkinson's disease data was downloaded from the IntAct database (https://www.ebi.ac.uk/intact/query/annot:%22dataset:parkinsons%22?conversationContext=9) under search term annot: "dataset:parkinsons." Cancer disease data was downloaded from the IntAct database (https://www.ebi.ac.uk/intact/query/annot:%22dataset:cancer%22?conversationContext=b) under search term annot: "dataset:cancer." Protein sequence information were downloaded from the Uniprot database (https://www.uniprot.org/downloads).

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments, performed the experiments, and contributed reagents, materials, analysis tools: FL. Algorithm design and analysis and analyzed the data: FZ and FL. Wrote the paper: FL, FZ, XL, and QL.

FUNDING

This work was supported by National Natural Science Foundation of China (61303108); The Natural Science Foundation of Jiangsu Higher Education Institutions of China (17KJA520004); Suzhou Key Industries Technological Innovation-Prospective Applied Research Project (SYG201804); Program of the Provincial Key Laboratory for Computer Information Processing Technology (Soochow University) (KJS1524); A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

REFERENCES

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33, 831–838. doi: 10.1038/nbt.3300

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). The universal protein resource (UniProt). *Nucleic Acids Res.* 33, 154–159. doi: 10.1093/nar/gki070

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Boxem, M., Maliga, Z., Klitgord, N., Li, N., Lemmens, I., Mana, M., et al. (2008). A protein domain-based interactome network for C. Elegans early embryogenesis. *Cell* 134, 534–545. doi: 10.1016/j.cell.2008.07.009

Chen, Q., Zhu, X., Ling, H.-Z., Wei, S., Jiang, H., and Inkpen, D. (2017). Recurrent neural network-based sentence encoder with gated attention for natural language inference. *arXiv* arXiv:1708.01353. Available online at: https://arxiv.org/abs/1705.00106

Du, X., Shao, J., and Cardie, C. (2017a). Learning to ask: neural question generation for reading comprehension. *arXiv: Comput. Lang.* arXiv:1705.00106v7. Available online at: https://arxiv.org/abs/1708.01353

Du, X., Sun, S., Hu, C., Yao, Y., Yan, Y., Zhang, Y. (2017b). DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. *J. Chem. Inf. Model* 57, 1499–1510. doi: 10.1021/acs.jcim.7b00028

Gomes, H. M., Barddal, J. P., Enembreck, F., and Bifet, A. (2017). A survey on ensemble learning for data stream classification. *ACM Comput. Surv.* 50:23. doi: 10.1145/3054925

Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res.* 36, 3025–3030. doi: 10.1093/nar/gkn159

Han, F., and Huang, D. S. (2006). Improved extreme learning machine for function approximation by encoding a priori information. *Neurocomputing* 69, 2369–2373. doi: 10.1016/j.neucom.2006.02.013

Hashemifar, S., Neyshabur, B., Khan, A. A., and Xu, J. (2018). Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* 34, 802–810. doi: 10.1093/bioinformatics/bty573

Huang, D. S., and Du, J. X. (2008). A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. *IEEE Trans. Neural Netw.* 19, 2099–2115. doi: 10.1109/TNN.2008.2004370

Huang, D. S., and Zheng, C. H. (2006). Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* 22, 1855–1862. doi: 10.1093/bioinformatics/btl190

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., et al. (2007). IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.* 35, 561–565. doi: 10.1093/nar/gkl958

Keskin, O., Tuncbag, N., and Gursoy, A. (2016). Predicting protein–protein interactions from the molecular to the proteome level. *Chem. Rev.* 116, 4884–4909. doi: 10.1021/acs.chemrev.5b00683

Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., and Wozniak, M. (2017). Ensemble learning for data stream analysis. *Inform Fus.* 37, 132–156. doi: 10.1016/j.inffus.2017.02.004

Leung, M. K., Xiong, H. Y., Lee, L. J., and Frey, B. J. (2014). Deep learning of the tissue-regulated splicing code. *Bioinformatics* 30, 121–129. doi: 10.1093/bioinformatics/btu277

Li, J., Zhu, X., and Chen, J. Y. (2009). Building disease-specific drug-protein connectivity maps from molecular interaction networks and pubmed abstracts. *PLoS Comput. Biol.* 5:e1000450. doi: 10.1371/journal.pcbi.1000450

Liu, F., Ren, C., Li, H., Zhou, P., Bo, X., and Shu, W. (2016). *De novo* identification of replication-timing domains in the human genome by deep learning. *Bioinformatics* 32, 641–649. doi: 10.1093/bioinformatics/btv643

Liu, S., Shen, F., Komandur Elayavilli, R., Wang, Y., Rastegar-Mojarad, M., Chaudhary, V., et al. (2018). Extracting chemical–protein relations using attention-based neural networks. *Database* 2018:bay102. doi: 10.1093/database/bay102

Martin, S., Roe, D., and Faulon, J. L. (2005). Predicting protein–protein interactions using signature products. *Bioinformatics* 21, 218–226. doi: 10.1093/bioinformatics/bth483

Muley, V. Y., and Ranjan, A. (2012). Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. *PLoS ONE* 7:e42057. doi: 10.1371/journal.pone.0042057

Planas-Iglesias, J., Bonet, J., García-García, J., Marín-López, M. A., Feliu, E., and Oliva, B. (2013). Understanding protein–protein interactions using local structural features. *J. Mol. Biol.* 425, 1210–1224. doi: 10.1016/j.jmb.2013.01.014

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., et al. (2007). Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U. S. A.* 104, 4337–4341. doi: 10.1073/pnas.0607879104

Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinform.* 18, 277. doi: 10.1186/s12859-017-1700-2

Tian, B., Wu, X., Chen, C., Qiu, W., Ma, Q., and Yu, B. (2019). Predicting protein-protein interactions by fusing various chou's pseudo components and using wavelet denoising approach. *J. Theor. Biol.* 462, 329–346. doi: 10.1016/j.jtbi.2018.11.011

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). *Attention is all You Need. Neural Information Processing Systems (NIPS 2017)*. Long Beach, CA: MIT Press.

Verga, P., Strubell, E., and Mccallum, A. (2018). Simultaneously self-attending to all mentions for full-abstract biological relation extraction. *N Am. Chapter Assoc. Comput. Linguist.* 1, 872–884. doi: 10.18653/v1/N18-1080

Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., et al. (2017). MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. *Bioinformatics* 33, 3909–3916. doi: 10.1093/bioinformatics/btx496

Wang, X., Yu, B., Ma, A., Chen, C., Liu, B., and Ma, Q. (2019). Protein–protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique. *Bioinformatics* 35, 2395–2402. doi: 10.1093/bioinformatics/bty995

Xuan, P., Cao, Y., Zhang, T., Kong, R., and Zhang, Z. (2019). Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes. *Front. Genet.* 10:416. doi: 10.3389/fgene.2019.00416

Yang, L., Xia, J. F., and Gui, J. (2010). Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein Pept. Lett.* 17, 1085–1090. doi: 10.2174/092986610791760306

You, W., Yang, Z., Guo, G., Wan, X., and Ji, G. (2019). Prediction of DNA-binding proteins by interaction fusion feature representation and selective ensemble. *Knowledge Based Syst.* 163, 598–610. doi: 10.1016/j.knosys.2018.09.023

Zhang, L., Yu, G., Xia, D., and Wang, J. (2019a). Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* 324, 10–19. doi: 10.1016/j.neucom.2018.02.097

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., et al. (2012). Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature* 490, 556–560. doi: 10.1038/nature11503

Zhang, Y., Lin, H., Yang, Z., Wang, J., and Sun, Y. (2019b). Chemical–protein interaction extraction via contextualized word representations and multihead attention. *Database* 2019:baz054. doi: 10.1093/database/baz054

Zhou, J., and Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12, 931–934. doi: 10.1038/nmeth.3547

Zhou, Y. Z., Gao, Y., and Zheng, Y. Y. (2011). Prediction of protein-protein interactions using local description of amino acid sequence. *Comm. Comput. Inf.* 202, 254–262. doi: 10.1007/978-3-642-22456-0_37

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

frontiers
in Bioengineering and Biotechnology

ORIGINAL RESEARCH
published: 06 May 2020
doi: 10.3389/fbioe.2020.00418

Disease Module Identification Based on Representation Learning of Complex Networks Integrated From GWAS, eQTL Summaries, and Human Interactome

Tao Wang, Qidi Peng, Bo Liu, Yongzhuang Liu* and Yadong Wang**

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

OPEN ACCESS

Edited by:
Zhongyu Wei,
Fudan University, China

Reviewed by:
Junwei Han,
Harbin Medical University, China
Zhen Tian,
Zhengzhou University, China

***Correspondence:**
Bo Liu
bo.liu@hit.edu.cn
Yongzhuang Liu
yongzhuang.liu@hit.edu.cn
Yadong Wang
ydwang@hit.edu.cn

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

Received: *06 March 2020*
Accepted: *14 April 2020*
Published: *06 May 2020*

Citation:
Wang T, Peng Q, Liu B, Liu Y and Wang Y (2020) Disease Module Identification Based on Representation Learning of Complex Networks Integrated From GWAS, eQTL Summaries, and Human Interactome. Front. Bioeng. Biotechnol. 8:418. doi: 10.3389/fbioe.2020.00418

The study of disease-relevant gene modules is one of the main methods to discover disease pathway and potential drug targets. Recent studies have found that most disease proteins tend to form many separate connected components and scatter across the protein-protein interaction network. However, most of the research on discovering disease modules are biased toward well-studied seed genes, which tend to extend seed genes into a single connected subnetwork. In this paper, we propose N2V-HC, an algorithm framework aiming to unbiasedly discover the scattered disease modules based on deep representation learning of integrated multi-layer biological networks. Our method first predicts disease associated genes based on summary data of Genome-wide Association Studies (GWAS) and expression Quantitative Trait Loci (eQTL) studies, and generates an integrated network on the basis of human interactome. The features of nodes in the network are then extracted by deep representation learning. Hierarchical clustering with dynamic tree cut methods are applied to discover the modules that are enriched with disease associated genes. The evaluation on real networks and simulated networks show that N2V-HC performs better than existing methods in network module discovery. Case studies on Parkinson's disease and Alzheimer's disease, show that N2V-HC can be used to discover biological meaningful modules related to the pathways underlying complex diseases.

Keywords: disease module identification, GWAS, eQTL, node2vec, hierarchical clustering

1. INTRODUCTION

The genome-wide association studies (GWAS) have successfully identified vast of variants associated with complex diseases (Visscher et al., 2017). However, the gene targets responsible for GWAS signals largely remain elusive, which hinders the way of illuminating molecular mechanisms of complex diseases and developing novel drug targets (Gallagher and Chen-Plotkin, 2018; Cheng et al., 2019b). The challenge of transforming GWAS signals into clinical useful gene targets is mainly due to the fact that most susceptibility variants locate in non-coding regions and thus do not alter the protein sequence directly. Emerging evidence has shown that regulation of gene expression is important mechanism associated with disease susceptibility variants (Westra et al., 2013; GTEx Consortium, 2017; Watanabe et al., 2017). Thus, to understand the molecular mechanism underlying GWAS signals, there is an urgent need to investigate the genes regulated by disease-associated variants and gene modules which could be disturbed by these potential disease genes.

The development of genome-wide assay of genetic variants and gene expressions, makes it possible to systematically associate genetic variations with quantitive levels of gene expression in a population, which is known as expression quantitative trait loci (eQTL) analysis (GTEx Consortium, 2017). Advances in eQTL studies enable rapid identification of potential casual genes (i.e., eQTL regulated genes, egenes) genome-widely in relevant tissues of complex diseases (Fairfax et al., 2012; Cheng et al., 2018b; Dong et al., 2018; Wang et al., 2019a,b). The public available eQTL and other molecular signatures have become useful resources to nominate candidate casual genes of complex diseases (GTEx Consortium, 2017; Cheng et al., 2019a, 2020). However, the detailed understanding of the molecular mechanisms through which these egenes jointly affect disease phenotypes remains largely unclear, and their discovery is a challenging computational task (Cheng et al., 2019b; Peng et al., 2020a). Instead of analyzing binary relationships between single SNP and single gene, network-based analyses provide valuable insights into the higher-order structure of gene communities or pathways that those potential disease genes may work together in the etiology of complex diseases (Fagny et al., 2017; Cheng et al., 2019b; Peng et al., 2020b; Wang et al., 2020). And advances in deep learning and graph representation learning technologies improve the accuracy of identifying disease related biomarkers (Peng et al., 2019a,b). In this paper, our purpose is to derive disease related modules from an integrated network with multi-layer information including human interactome (mainly protein-protein interactions, PPI), and summaries of GWAS and eQTL studies. To aid this purpose, we present a novel algorithm named N2V-HC, which could learn deep representing features of nodes in the integrated molecular network, and unbiasedly detect gene communities enriched with potential disease genes (i.e., egenes in the context).

The identification of disease modules is driven by the primary observation that disease-related proteins tend to interact closely in biological network (Agrawal et al., 2018). In recent years, many studies have applied network-based methodologies to predict disease modules (Califano et al., 2012; Mäkinen et al., 2014; Ghiassian et al., 2015; Sharma et al., 2015; Calabrese et al., 2017). However, there are several challenges in current disease modules detection methods: (1) most methods rely on seed genes to expand the connected module. They adapt "seed-extend" strategy, starting from the well-studied disease genes and expanding the module by adding directly connected neighborhood. However, some complex diseases have no or only a few known disease genes, such as neurodegenerative disorders (e.g., Parkinson's disease, Alzheimer's disease etc.). This makes the process biased toward well-studied disease genes, and the discovery ability is largely limited by selected seed genes. (2) Recent studies have shown that most disease pathways do not correspond to single well-connected component in PPI network. Instead, disease proteins tend to form many separate connected components and scatter across the network (Agrawal et al., 2018). However, current methods tend to extend the seed genes into a large connected component or sub-network which might be less sufficient for discovering global disease modules. (3) The principle of node similarity measurement in current methods is mainly based on homophily, while ignoring the structural equivalence. Under the homophily hypothesis, nodes in the same module have higher similarity while under the structural equivalence hypothesis, nodes that have similar structural roles in network also have higher similarity. Studies have shown that the structural equivalence is also an important feature in measuring node similarity (Perozzi et al., 2014; Grover and Leskovec, 2016), which should also be considered.

To solve these challenges, our proposed method, N2V-HC, first predicts the disease genes based on genetic associations from summaries of GWAS and eQTL studies and integrates eQTL SNP (eSNP), eQTL regulated gene (egene) with human interactome network. Second, we use node2vec (Grover and Leskovec, 2016), an advanced network embedding method, to learn node features through a biased random walk process. The embedding process considers both the homophily and structural equivalence of nodes in the network. Third, nodes are clustered based on their embedding features using an iterative hierarchical clustering strategy. Modules are determined by a dynamic tree-cut strategy, and candidate disease modules are prioritized by evaluating whether the module is enriched for predicted disease genes. To evaluate the clustering performance of N2V-HC, we compared it with several state-of-the-art graph clustering methods including Markov clustering (MCL) (Enright et al., 2002), affinity propagation (AP) (Frey and Dueck, 2007), spectral clustering (Shi and Malik, 2000), mCODE (Bader and Hogue, 2003), GLay (Su et al., 2010), and hierarchical clustering on several real-world networks with ground truth labels, and also on multiple simulated networks. The experimental results showed that our method has better clustering performance than compared methods. We also performed case studies on Parkinson's disease (PD) and Alzheimer's disease (AD), and found biological meaningful modules associated with PD and AD, which might help to explain the pathology of diseases.

2. METHODS

2.1. Overview

In order to pinpoint key disease related modules, we propose a novel algorithm named N2V-HC, which could learn global connectivity features for nodes in an integrated molecular network, and automatically detect gene communities enriched with potential disease genes. The N2V-HC algorithm mainly consists of three steps as shown in **Figure 1**. Step 1: construction of integrated complex network. The integrated network is constructed based on known experimental molecular interaction networks, such as PPI network, and additional edges are added based on disease relevant signals from GWAS and the eQTL links between GWAS signals to network genes (section 2.2). Step 2: representation learning in network. N2V-HC learns features or embeddings for each node in the network by using node2vec (section 2.3). Step 3: identification of disease modules.

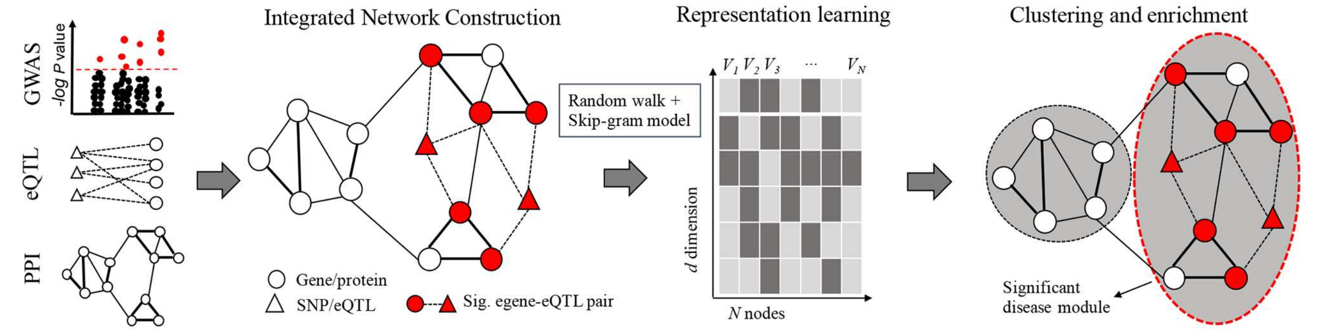


FIGURE 1 | Framework of N2V-HC algorithm. The left-most panel shows input data sources of the integrated network: summary statistics of GWAS and eQTL studies, and PPI network or other types of networks. The edge width represents weight on edge. Representation learning step extracts global connectivity features for *N* nodes of the integrated network by using a biased random walk technology and the Skip-gram model. Each feature is a numeric vector of *d* dimension. Unsupervised hierarchical clustering method and dynamic tree-cut method are applied in an iterative module convergence process. The circle with red dash line represents the disease module which is significantly enriched with egenes.

Unsupervised hierarchical clustering method and dynamic tree-cut method are applied to partition network nodes into modules, and an iterative module convergence strategy is used. The disease module is finally prioritized by enrichment performance (section 2.4). Other methods are also detailed here (sections 2.5–2.7).

2.2. Construction of Integrated Complex Network

We project the eQTLs significantly associated with specific disease onto a gene interaction network, i.e., a PPI network in this work, and generate an integrated biological complex network, where disease modules are discovered. To make the network construction procedures more clear, we use susceptibility variants of Parkinson's disease (PD) and Alzheimer's disease (AD) as cases to illustrate the whole process.

2.2.1. GWAS Data Preparation

First, we extract GWAS index SNPs of PD and AD from the most recent and largest GWAS papers conducted by Nalls et al. (2019) and Jansen et al. (2019). Second, we calculate proxy SNPs in linkage disequilibrium (LD) with index SNPs by setting LD $R^2 \geq 0.6$ using EUR population of 1000G genome reference panel (Genomes Project Consortium, 2015). Proxy SNPs are derived separately for PD and AD using SNiPA platform (https://snipa.helmholtz-muenchen.de/snipa3/?task=proxy_search), and other parameters are set in default.

2.2.2. eQTL Data Preparation

As eQTL and gene expression are tissue-specific and PD and AD are also relevant to brain tissue, we first download eQTL summaries of brain frontal cortex from GTEx portal (https://gtexportal.org/). Then, we extract associations involving those GWAS-derived SNPs (index SNPs and their proxies). FDR is calculated based on the nominal *P*-values of the extracted eQTL associations. We use $FDR \leq 0.05$ as cutoff to determine significant eQTL-egene associations.

2.2.3. Human Interactome Preparation

First, we use the molecular physical interaction network complied by Menche et al. (2015), consisting of 141,296 physical interactions and 13,460 proteins. The edges of the network are experimentally documented in human cells, including protein-protein and regulatory interactions, metabolic pathway, and kinase-substrate interactions. Since some genes are not active in human brain, we filtered out 2,736 genes with low expression levels in frontal cortex based on the gene expression profiles in GTEx portal.

2.2.4. Network Integration

We first projected the significant eQTL-egene pairs onto the human interactome. Since the input proxy SNPs can be tagged by index SNPs, we used the corresponding index SNPs to replace the proxy SNPs in the merged network.

2.3. Representation Learning of Network Structure

Node2vec (Grover and Leskovec, 2016) is applied to learn the global features or representations of nodes in the network. Node2vec is a network embedding method based on random walk, which has been successfully applied in bioinformatics applications (Grover and Leskovec, 2016; Cheng et al., 2018a). It learns node representations following two principles: nodes in the same community have similar embeddings (i.e., homophily); and nodes sharing similar structure roles have similar embeddings (i.e., structural equivalency).

Node2vec extends the Skip-gram model to networks. Given a graph $G = (V, E)$, it learns the representation $\vec{z}_u = f(u)$ of node u by optimizing the objective function given by Equation 1, where $N_S(u)$ represents network neighborhood of node u generated by a sampling strategy S, and $f : V \to R^{n\times d}$, where d is the dimension of the embedding space (i.e., the feature dimension of nodes). By making assumptions of conditional independence and symmetry of feature space, the objective function is further transformed into

Equation (2).

$$\max_{f} \sum_{u \in V} \log P(N_S(u)|f(u)) \tag{1}$$

$$\max_{f} \sum_{u \in V} \{-\log [\sum_{v \in V} \exp(f(u) \cdot f(v))] + \sum_{n_i \in N_S(u)} f(n_i) \cdot f(u)\} \tag{2}$$

In order to obtain the node neighborhood $N_S(u)$, node2vec uses a biased random walk method, which can perform flexible trade-offs between DFS and BFS. It calculates the node neighborhood by simulating a random walk of length l. Suppose the current position is node v, the previous position is node t, and the next step is to walk to node x. To determine the next node x, the transition probability is designed as shown in Equation (3), where $\alpha_{pq}(t, x)$ is given by Equation (4) and $d_{tx} = \{0, 1, 2\}$ represents the shortest path distance from node t to node x, and the p and q parameters constrain the direction of random walk (that is, a large p indicates closer to DFS, while a large q indicates closer to BFS). Let c_i represents the walker in step i, then the probability of visiting node x is given by Equation (5). Among them, Z represents a normalized constant, that is, $Z = \sum_{(v,x) \in E} \pi_{vx}$.

$$\pi_{vx} = \alpha_{pq}(t, x) \cdot w_{vx} \tag{3}$$

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p}, d_{tx} = 0; \\ 1, d_{tx} = 1; \\ \frac{1}{q}, d_{tx} = 2. \end{cases} \tag{4}$$

$$P(c_i = x | c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z}, (v, x) \in E; \\ 0, othersize. \end{cases} \tag{5}$$

2.4. Identification of Disease Modules

2.4.1. Hierarchical Clustering and Dynamic Dendrogram-Cut

After learning the global connectivity features for each node in the network, we perform bottom-up hierarchical clustering to distinct modules. The hierarchical clustering initially treats each node as a cluster, and then iteratively merges the two clusters that have best similarity until the last one. Typically, N2V-HC uses Euclidean distance and average linkage method by default to construct the dendrogram. Then we apply Dynamic Hybrid tree-cut method on the dendrogram to obtain a flexible number of clusters.

The Dynamic Hybrid tree-cut method adopts bottom-up merging strategy (Langfeldera et al., 2008). Let N be the total number of nodes in a cluster, and N_0 be the minimum number of nodes in a cluster. The cluster core is defined as the lowest N_c nodes in the cluster, where $N_c = \min\{int(\frac{N_0}{2} + \sqrt{N - \frac{N_0}{2}}), N\}$. The core scatter $\bar{d}$ is the average dissimilarity of the node pairs in the cluster core. The cluster gap g is the difference between $\bar{d}$ and the height of the cluster. The first step of the "Dynamic Hybrid" method is to merge the nodes/branches in the dendrogram bottom to up to get initial clusters. These clusters should satisfy the following four conditions: (1) $N > N_0$; (2) the height of the cluster is less than the maximum tree height h_{max}; (3) the cluster's core scatter $\bar{d} < d_{max}$; (4) The cluster gap $g > g_{min}$. $(N_0, h_{max}, d_{max}, g_{min})$ can be specified by the user. This will leave out some single nodes or tiny clusters (cluster that meet the above conditions except $N > N_0$), which are called outliers. The second step is to merge these outlier into the clusters generated in the first step. For these outliers, the outlier-cluster dissimilarity is calculated one by one, and is classified into the cluster most similar to it (Langfeldera et al., 2008).

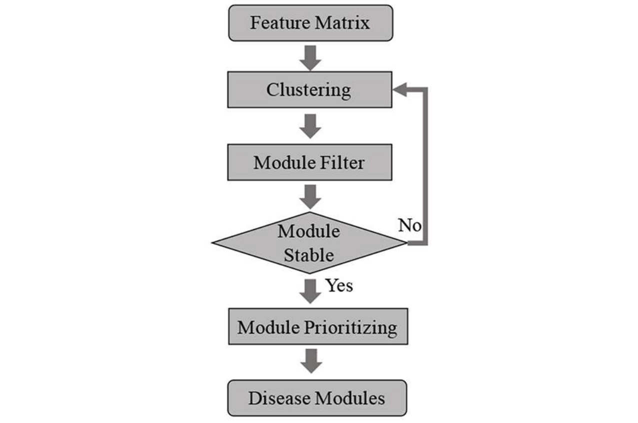


FIGURE 2 | Steps of disease module identification.

2.4.2. Iterative Module Selection Process

After global clustering, the initial clusters are generated, some of which may be enriched with disease associated egenes, while other may not consist of any egenes. To boil down the number of candidate modules, we filter out modules that do not consist of any disease relevant egenes. The genes in left modules are then extracted as a subnetwork, and we repeat the clustering and dynamic dendrogram-cut processes. These steps will be iteratively performed until the modules are stable, which means current clustering results stay same with last clustering results. After the process is convergent, all left modules consist of disease relevant egenes, which are the candidate disease modules. The iterative module selection process is shown in **Figure 2**.

2.4.3. Prioritizing Disease Modules by Enrichment Analysis

We then test whether egenes are enriched in the candidate disease modules. The enrichment analysis is performed by Fisher's exact test. All genes shown in the network with size n are used as background genes, and are assigned to four cells of a two by two contingency table, according to if a gene is in a module or not, and if it is a egene or not. For example, given a module M, suppose a is the number of genes that are in module M and are egenes; b represents the number of genes that are egenes but not in M; c is number of genes in module M, but are not egenes; d represents number of genes that are not egenes and not in module M, the fisher's exact test P-value is given by the Equation 6:

$$P = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \tag{6}$$

2.5. Module Mapping

To evaluate the performance of module detection on ground truth datasets or simulated datasets, it is essential to first match the modules discovered by methods under evaluation with the ground truth modules. We model this module mapping problem by a classical task assignment algorithm. The task assignment problem is a fundamental combinatorial optimization problem. Suppose there are N agents and N tasks, each agent will be assigned to perform a task, and there will be a cost generated for each agent-task assignment, the object is to find the best task assignment strategy to minimize the cost. In context of the module mapping problem, our purpose is to find the best bijection between predicted module set and ground truth module set, which maximize the size of module intersections. In formula, let the intersection matrix as $\{S_{i,j}\}_{N*N}$, where $s_{i,j} = 1$ represents the number of overlapping nodes between module i and module j, and the binary matrix as $\{M_{i,j}\}_{N*N}$, where $m_{i,j} = 1$ if and only if module i is matched with module j, otherwise $m_{i,j} = 0$. To guarantee one-to-one correspondence, two conditions are needed: $\sum_{i=1}^{N} m_{i,j} = 1$ and $\sum_{j=1}^{N} m_{i,j} = 1$. The objective is to optimize the binary matching matrix $\{M_{i,j}\}_{N*N}$ which maximizes $\sum_{i=1}^{N}\sum_{j=1}^{N} s_{i,j} * m_{i,j}$.

In addition, there is a common case that the number of predicted modules is not equal to the module number in ground truth. And this is an unbalanced task assignment problem. As a solution, we manually add empty modules to the short module sequence, to make sure the two module sequences have same length. Then, the problem is transformed to balanced task assignment problem, as described above.

2.6. Micro F1 Score

In binary classification problem, the F1 score is commonly used performance indicator, as shown in Equation (7), where $precision = \frac{TP}{TP + FP}$, and $recall = \frac{TP}{TP + FN}$.

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{7}$$

The module detection is similar to multi-label classification problem. To compare the module detection performance of different methods on ground truth datasets, we use micro F1 score as the indicator. The micro F1 score is a variant of F1 score, as shown in Equation (8), where $precision_{Micro}$ is defined in Equation (9) and $recall_{Micro}$ is defined in Equation (10). Suppose there are N predicted modules, the TP_i, FP_i, FN_i in Equations (9) and (10) represent the number of true positive nodes, false positive nodes and false negative nodes in module i, respectively.

$$F1_{Micro} = \frac{2 * recall_{Micro} * precision_{Micro}}{recall_{Micro} + precision_{Micro}} \tag{8}$$

$$precision_{Micro} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i + FP_i)} \tag{9}$$

$$recall_{Micro} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N}(TP_i + FN_i)} \tag{10}$$

2.7. Gene Set Enrichment Analysis

Gene enrichment analysis is performed by overlapping genes in a module with Gene Ontology (GO) gene sets using GSEA with the C2 and C5 collection of the MSigDB. Genes shown in candidate disease modules are mapped onto MSigDB and are evaluated by fisher's exact test. The top 50 significantly enriched terms are used.

3. RESULTS AND DISCUSSION

The accuracy of disease module detection in N2V-HC largely depends on the unsupervised clustering process. In this section, we first compared N2V-HC with several classical graph clustering methods, including Affinity propagation, GLay, MCL, Spectral clustering, mCODE, and Hierarchical clustering on various types of testing networks with labels of ground truth modules. Next, we applied N2V-HC to Parkinson's disease and Alzheimer's disease with PPI network, the latest GWAS summaries and brain eQTL summaries. We found (1) our method significantly performs better than compared methods; (2) most of the identified disease modules correspond to core disease-relevant pathways, which often comprise therapeutic targets.

3.1. Clustering Performance on Real-World Networks

To evaluate the clustering performance of N2V-HC, we compared it with several state-of-the-art graph clustering methods, including Markov clustering (MCL) (Enright et al., 2002), affinity propagation (AP) (Frey and Dueck, 2007), spectral clustering (Shi and Malik, 2000), mCODE (Bader and Hogue, 2003), GLay (Su et al., 2010), and hierarchical clustering. Six real-world networks with various sizes, densities, types (weighted/unweighted, directed/undirected) and ground truth cluster labels were used as testing datasets, including: Zachary's karate club network (Zachary, 1977), UKfaculty social network (Nepusz et al., 2008), Dolphin Social Network (Lusseau et al., 2003), College football game network (Girvan and Newman, 2002), US Political Books network (Krebs, 2004), and Cora citation network (Fakhraei et al., 2015). The six real-world networks are summarized in **Table 1**.

To evaluate their performance, micro F1 score was chosen as the indicator of performance (see section 2). We first map the predicted modules with ground truth modules by maximizing the overlap size of all modules (see section 2). Then, true positive (TP), false positive (FP), true negative (TN) and false negative (FN) number of nodes in each predicted module were calculated and leveraged into the micro F1 score (see section 2). To be

TABLE 1 | Summary of real-world network datasets.

Dataset	#Nodes	#Edges	Density	#Clusters	Graph type	References
Karate	34	78	1.4E-1	2	w, ud	Zachary, 1977
Dolphins	62	159	8.4E-2	2	uw, ud	Lusseau et al., 2003
UKfaculty	81	817	2.5E-1	4	w, ud	Nepusz et al., 2008
Polbooks	105	441	8.1E-2	3	uw, ud	Krebs, 2004
Football	115	613	9.4E-2	12	uw, ud	Girvan and Newman, 2002
Cora	2,708	5,429	1.4E-3	7	uw, ud	Fakhraei et al., 2015

w, weighted graph; uw, unweighted graph; ud, undirected graph.

TABLE 2 | Clustering performance on real-world networks.

Datasets	AP	GLay	MCL	SC	HC	mCODE	N2V-HC (MMS, DS, NPC)
Karate	0.844	0.847	0.529	0.588	0.588	0.623	**0.941** (10, 2, 2)
Dolphins	0.935	0.804	0.677	0.613	0.565	0.533	**0.984** (10, 2, 2)
UKfaculty	0.494	0.889	0.951	0.370	0.333	0.397	**0.963** (10, 2, 3)
Polbooks	0.609	0.816	0.838	0.400	0.438	0.451	**0.848** (10, 2, 4)
Football	0.113	0.583	**0.930**	0.235	0.235	0.435	0.922 (5, 2, 11)
Cora	0.356	0.512	0.294	0.298	0.287	0.295	**0.661** (100, 0, 6)

AP, affinity propagation; MCL, Markov cluster; SC, spectral clustering; HC, hierarchical clustering; MMS, minModuleSize; DS, DeepSplit; NPC, number of predicted clusters. Parameter setting: MCL inflation factor setting: Karate 2.0, Dolphins 2.0, UKfaculty 2.5, Polbooks 2.1, Football 2.0, Cora 1.8. Parameters in AP, GLay, SC, HC, and mCODE were in default except that cluster number was set to the ground truth if available. Bold Values indicate the best micro F1 scores.

noted, we fine-tuned the corresponding parameters of N2V-HC and compared methods to make the number of predicted modules close to the true module numbers. The experiment results were summarized in **Table 2**. As we can see, our method performs significantly better than most compared methods in the six real-world networks.

As a case, we illustrated the clustering effect of N2V-HC on Dolphins social network as shown in **Figure 3**. The original Dolphins social network is shown on the left panel, with red and blue colors representing two ground truth modules. The right panel shows the hierarchical dendrogram constructed by N2V-HC, where the leaf nodes represent the original dolphin members in the network, and the two predicted modules are also colored in red and blue. Only one node, with label "40," is wrongly classified into opposite module, which is colored in yellow. However, we can see from the original network, the node "40" actually appears at the border of both modules, and could be arbitrarily classified.

3.2. Clustering Performance on Simulated Networks

We then evaluated the performance of N2V-HC on simulated networks in various scales. We used the network simulation tool LFR-benchmark (Lancichinetti et al., 2008), to generate small-to-large scale networks, with weighted and directed edges. The character of simulated networks can be adjusted by function LFR(N, k, $maxk$, muw, $t1$, $t2$), where N controls the number of network nodes, k controls the average degree of the node, $maxk$ controls the maximum degree of the node, muw controls the mixing parameter for the weight, $t1$ controls minus exponent for the degree sequence, and $t2$ controls minus exponent for the community size distribution. We set $muw = 0.5$, $t1 = 2$, $t2 = 1$ in their default values. By setting different combination of parameters N, k, and $maxk$, we generated five networks in different scales (shown in **Table 3**). Then we run N2V-HC and compared methods on these five networks, the resulting micro F1 score is shown in **Table 4**. We can see that N2V-HC still performs much better than compared methods in different schema. With the network getting larger and more complex, the performance of compared methods tend to dramatically decline, while our method has better stability, indicating the robustness of N2V-HC. Combining the above experiments, we can conclude that N2V-HC can accurately extract the intrinsic network modules, which enables the ability to predict disease-relevant modules.

3.3. Case Studies on Parkinson's Disease and Alzheimer's Disease

Alzheimer's disease and Parkinson's disease are the top two neurodegenerative disorders, whose etiological mechanisms are still unclear. To predict the disease-relevant modules, we first constructed the networks integrated from GWAS, eQTL data, and human interactome by following steps (see section 2): (1) 90 and 32 independent GWAS index SNPs were obtained from the latest largest-scale to date GWAS of PD (Nalls et al., 2019) and AD (Jansen et al., 2019), respectively. (2) 7,194 and 1,270 proxy SNPs were derived separately based on 1000G EUR population for PD and AD. (3) eQTL associations were extracted for those GWAS-derived SNPs (index SNPs and their proxies) from summaries of GTEx brain frontal

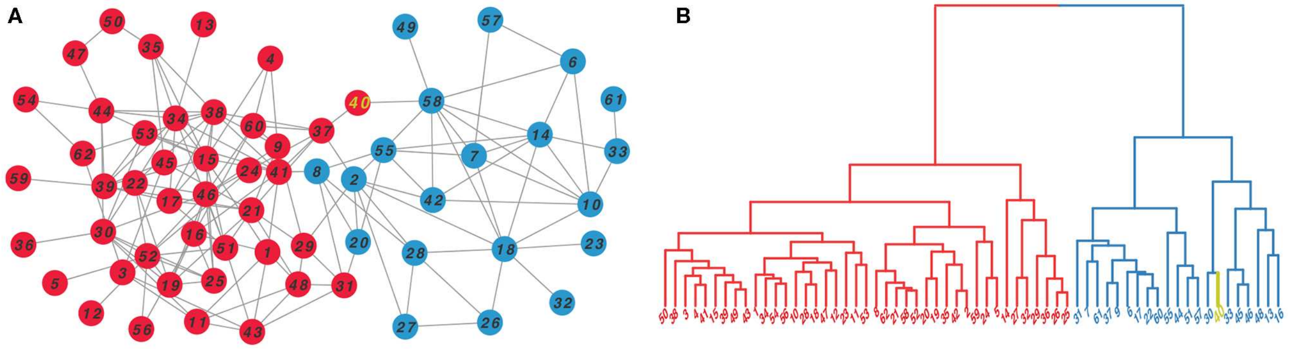


FIGURE 3 | The clustering effect of N2V-HC on Dolphins social network (Lusseau et al., 2003). **(A)** The topology of original network, with colors represents the ground truth communities. **(B)** The hierarchical clustering dendrogram constructed by N2V-HC, where each leaf node represents a member in original network. Two predicted modules are colored in red and blue. Node "40," which is misclassified, is labeled in yellow.

TABLE 3 | Summary of LFR simulated networks.

LFR($N, k, maxk$)	Nodes	Edges	Density	Clusters
LFR (100, 10, 30)	100	1,047	0.212	7
LFR (500, 10, 50)	500	5,269	0.042	36
LFR (1000, 20, 100)	1,000	19,115	0.038	39
LFR (2000, 30, 200)	2,000	60,946	0.030	34

TABLE 4 | Clustering performance on LFR-benchmark datasets.

Datasets	AP	GLay	MCL	SC	HC	mCODE	N2V-HC(MMS, DS, NPC)
LFR (100, 10, 30)	0.304	0.131	0.350	0.28	0.26	0.35	**0.615** (6, 2, 8)
LFR (500, 10, 50)	0.090	0.127	0.120	0.128	0.14	0.138	**0.496** (4, 3, 38)
LFR (1,000, 20, 100)	0.097	0.075	**0.692**	0.103	0.109	0.145	0.620 (6, 3, 40)
LFR (2,000, 30, 200)	0.092	0.033	0.651	0.080	0.082	0.135	**0.682** (5, 2, 34)

AP, affinity propagation; MCL, Markov cluster; SC, spectral clustering; HC, hierarchical clustering; MMS, minModuleSize; DS, DeepSplit; NPC, number of predicted clusters. Parameter setting: MCL inflation factor was set in default (2.5) for all networks. Parameters in AP, GLay, SC, HC, and mCODE were in default except that cluster number was set to the ground truth if available. Bold Values indicate the best micro F1 scores.

cortex (version V7). After filtering by threshold $FDR \leq 0.05$, 41,538 significant associations, representing 248 egenes and 4,821 eSNPs were extracted for PD; and 370 significant associations, representing 19 egenes and 150 eSNPs were extracted for AD. (4) We downloaded the molecular physical interaction network complied by Menche et al. (2015), which consists of 110,913 physical interactions and 10,724 proteins after removing genes with low expression levels in frontal cortex. (5) Finally, we projected the significant eQTL-egene pairs onto the human interactome. Since the input proxy SNPs can be tagged by index SNPs, we used the corresponding index SNPs to replace the proxy SNPs in the merged network. The outcome integrated network for PD consists of 10,912 nodes, including 10,852 genes and 60 independent PD susceptibility SNPs, and 111,038 edges. The outcome integrated network for AD consists of 10,736 nodes, including 10,727 genes and 9 independent AD susceptibility SNPs, and 110,803 edges. Then we performed N2V-HC on these two integrated networks, by setting the dimension of representing features as 128, and the Dynamic Hybrid tree-cut parameter as $minModuleSize = 20$ and $deepSplit = 2$.

For integrated network of PD, the module detection process converged after four iterations, resulting in 51 candidate disease modules containing at least one egene (**Table S1**). Fisher's exact test was conducted for each module to test whether egenes were over-expressed in the module. And FDR was calculated to evaluate the enrichment significance. After filtering by $FDR \leq 0.05$, 15 modules were predicted as the PD disease modules, which on average covered 80 genes. We next investigated the module function by performing gene set enrichment analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005). Specifically, we computed the overlaps between module genes and gene sets in C2 (curated gene sets) and C5 (GO gene sets) categories of MSigDB (Liberzon et al., 2015). Among the 15 predicted PD modules, 12 (80%) modules have been annotated with

TABLE 5 | Gene set enrichment analysis of PD modules.

ID	# Gene	# PD egene	*P*-value	FDR	GSEA inferred module function	PD-relevant evidence
PD36	39	20	2.94E-23	1.50E-21	GPCR ligand binding	Martin et al., 2005
PD41	33	17	5.62E-20	9.55E-19	Retinoic acid biosynthesis	Jacobs et al., 2007; Esteves et al., 2015,
PD42	32	13	7.47E-14	9.52E-13	GPI-anchor biosynthesis, ER/Golgi trafficking, Membrane lipid biosynthesis	Wang et al., 2014, Abeliovich and Gitler, 2016
PD12	126	19	5.45E-11	5.56E-10	Endocytosis, Immune response	Mosley et al., 2012; Abeliovich and Gitler, 2016
PD20	80	13	2.57E-08	2.18E-07	Immune response, Integrin cell surface	Wu and Reddy, 2012
PD37	38	9	1.28E-07	9.35E-07	Potassium channels, Glycogen metabolism	Chen et al., 2018
PD44	30	7	3.75E-06	2.12E-05	Hemoglobin complex	Freed and Chakrabarti, 2016
PD10	135	13	1.18E-05	6.00E-05	Oxidoreductase activity	Parker et al., 2008
PD34	42	7	3.94E-05	1.82E-04	Glycosaminoglycans biosynthesis	Lehri-Boufala et al., 2015
PD45	29	5	4.43E-04	1.74E-03	Immune response, Natural killer cell mediated immunity	Mihara et al., 2008
PD35	42	5	2.49E-03	9.08E-03	Lysosome, Sphingolipic metabolism	Dehay et al., 2013, Lin et al., 2019
PD46	29	4	3.96E-03	1.34E-02	WNT signaling pathway, Dopaminergic neuron differentiation	Arenas, 2014

# Gene, number of genes in a module; # PD egene, number of egene regulated by PD susceptibility variants in a module.

TABLE 6 | Gene set enrichment analysis of AD modules.

ID	# Gene	# AD egene	*P*-value	FDR	GSEA inferred module function	AD-relevant evidence
AD1	88	6	6.36E-09	5.09E-08	Immune response	Wang et al., 2018
AD2	42	3	5.16E-05	2.07E-04	WNT signaling pathway,Dopaminergic neuron differentiation	dos Santos and Smidt, 2011
AD3	177	4	2.28E-04	6.08E-04	Immune response,JAK/STAT signaling pathway	Nicolas et al., 2013
AD4	52	2	3.73E-03	7.47E-03	ER/Golgi trafficking,Glycosaminoglycans metabolism	Placido et al., 2014

functions relevant to known PD pathways (**Table 5**, **Table S1**). For example, the cellular pathways including oxidative stress, immune response, endosomal-lysosomal dysfunction, intra-cellular trafficking stress etc., have been widely reported associated with PD pathology in literatures (Parker et al., 2008; Mosley et al., 2012; Dehay et al., 2013; Abeliovich and Gitler, 2016).

Similarly, we also obtained eight candidate modules associated with AD, among which four modules had $FDR \leq 0.05$ based on Fisher's exact test (**Table 6**, **Table S2**). These molecular pathways include immune response, WNT signaling pathway, JAK/SAT signaling pathway and intra-cellular trafficking, which also have been reported associated with AD pathology in literatures (dos Santos and Smidt, 2011; Nicolas et al., 2013; Placido et al., 2014; Wang et al., 2018). Interestingly, the predicted AD modules and PD modules have similar functions, for example, AD1, AD3, PD12, PD20, and PD45 are all associated with immune response; AD2 and PD46 are associated with WNT signaling pathway and dopaminergic neuron differentiation; AD4 and PD42 are associated with intracellular trafficking. Three module pairs have high similarity including (AD1, PD20), (AD2, PD46), and (AD4, PD34), whose intersection size and Jaccard index are (67, 0.68), (21, 0.44), and (18, 0.26), respectively. There is no similarity (Jaccard index = 0) or very low similarity (Jaccard index < 0.05) between other AD-PD module pairs. These evidence indicate that AD and PD might share remarkably similar dysregulated pathways; and multiple modules may work together in the same disease pathway (e.g., immune response), where shared modules might be involved between AD and PD pathology.

In order to investigate the relationship between the predicted disease modules, our method is able to built the dendrogram of all candidate modules based on the module eigen feature, defined as the eigen vector of node features in a module corresponding with the first principle component. For example, the module dendrogram of Parkinson's disease was shown in **Figure 4**. We found several module blocks (modules with high similarity covered by shaded rectangle as shown in **Figure 4**) are annotated with similar functions. For example, PD10, PD15, and PD48 are related to oxidative stress; PD3, PD20, PD31, and PD45 are related to immune response; PD4, PD14, PD19, and PD43 are related to intracellular trafficking; PD33 and PD34 are related to glycosaminoglycans biosynthesis. Especially, PD24 and PD49 are both annotated as Parkinson's disease pathway (GSEA FDR

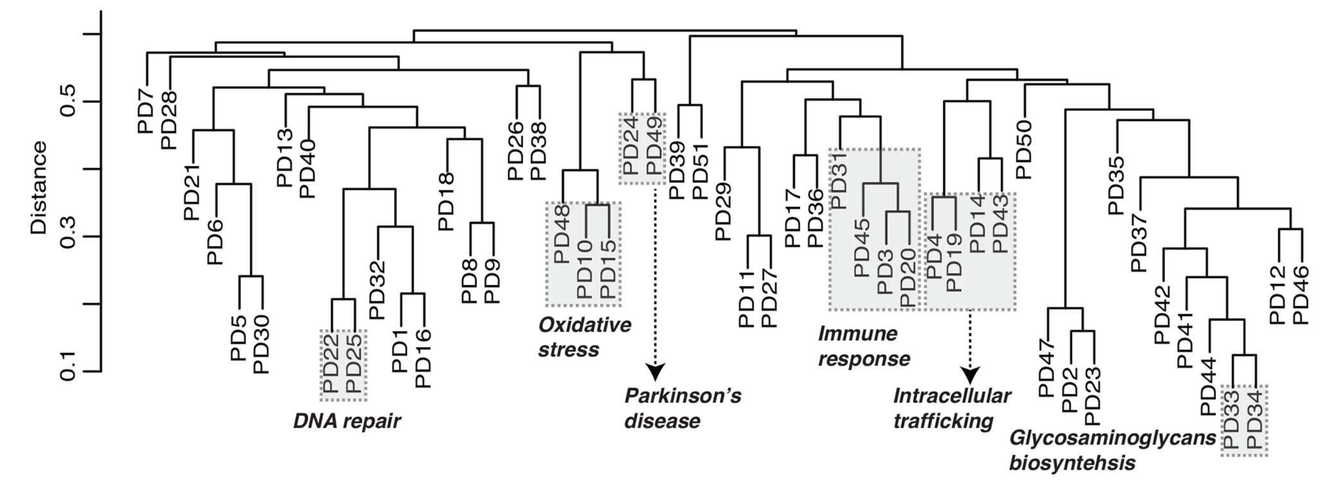


FIGURE 4 | Module dendrogram of Parkinson's disease. Dendrogram of modules is built based on module eigen feature, i.e., the eigen vector corresponding with the first principle component of node features in a module. Distance is measured as one minus Pearson's correlation coefficient. Modules covered by the shaded rectangle share similar functions as illustrated.

$= 1.2 \ast 10^{128}$ and $7 \ast 10^{15}$) and mitochondrial process (GSEA FDR $= 5.7 \ast 10^{141}$ and $1.5 \ast 10^{20}$) by GSEA. The module dendrogram provide guidance to merge multiple modules into a super module, and can also be used to infer module functions.

As a secondary finding, we found some of the provisionally insignificant candidate modules were also associated with functions relevant to AD and PD pathology. For example, two modules were directly annotated as Parkinson's disease pathway (PD24, GSEA FDR $= 1.2 \ast 10^{128}$) and Alzheimer's disease pathway (AD6, GSEA FDR $= 2 \ast 10^{8}$). We also found modules associated with autophagy (PD13), apoptosis (PD1), post-synapse (PD11), SNARE binding (PD19), and mitochondria (PD15, PD48, PD49, PD9), which are believed to have played a role in PD etiology (Dehay et al., 2013; Abeliovich and Gitler, 2016).

Furthermore, our method generates disease modules without bias toward the seed genes. The traditional methods adapt "seed-extend" strategy, starting from the disease seed genes and expanding the module by adding neighborhood. For example, the DIAMOnD algorithm (Ghiassian et al., 2015) first defines the disease module as the subnetwork only consisting of the well-studied disease genes (seed genes). Next, for each iteration, one gene (named DIAMOnD gene) with highest connectivity score with the module will be added to grow the module, until all genes in the network are added. The first added *N* DIAMOnD genes (*N* is arbitrarily defined by user) together with the seed genes will form the final disease module. Thus, the module generated under "seed-extend" strategy is biased toward seed genes. However, in our N2V-HC method, the seed genes are masked during the hierarchical clustering procedure. In other words, our module generation process is not based on seed genes. Instead, we use seed genes as posterior knowledge to prioritize modules based on enrichment significance.

4. CONCLUSIONS

Disease module identification is often a crucial step to discover disease pathway and potential drug targets. In this article, we present a new algorithm framework, named N2V-HC, to predict disease modules based on deep feature learning of biological complex networks. Our method includes three steps: First, integrating a network from GWAS, eQTL summaries, and human interactome; Second, learning the node representing features in the integrated network; Third, detecting modules based on hierarchical clustering, and evaluating whether some of modules may be candidates for specific disease by determining their enrichment with egenes that are regulated by disease susceptibility variants. Experiments on network datasets with ground true labels suggest our method has better performance in module detection than compared methods. In addition, we apply N2V-HC on Parkinson's disease and Alzheimer's disease, and find significant modules associated with PD and AD. In general, our method can be used to incorporate with other types of networks beside PPI. We believe it will be a powerful tool for researchers to understand the molecular mechanisms of complex diseases in the post-GWAS era.

DATA AVAILABILITY STATEMENT

GTEx eQTL datasets can be downloaded at the GTEx portal (https://gtexportal.org/). The implementation of N2V-HC can be freely downloaded at Github (https://github.com/QidiPeng/N2V-HC).

AUTHOR CONTRIBUTIONS

TW designed the study, analyzed the data, and wrote the paper. QP implemented the algorithm framework,

co-analyzed the data, and co-wrote the paper. BL, YL, and YW supervised the research, provided funding support, and revised the paper.

FUNDING

This work has been supported by the National Key Research and Development Program of China (Nos. 2017YFC0907503 and 2017YFC1201201).

ACKNOWLEDGMENTS

We are thankful to the reviewers for their helpful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbioe.2020.00418/full#supplementary-material

REFERENCES

Abeliovich, A., and Gitler, A. D. (2016). Defects in trafficking bridge Parkinson's disease pathology and genetics. *Nature* 539, 207–216.doi: 10.1038/nature20414

Agrawal, M., Zitnik, M., and Leskovec, J. (2018). "Large-scale analysis of disease pathways in the human interactome," in *PSB* (Hawaii: World Scientific), 111–122. doi: 10.1142/9789813235533_0011

Arenas, E. (2014). Wnt signaling in midbrain dopaminergic neuron development and regenerative medicine for Parkinson's disease. *J. Mol. Cell Biol*. 6, 42–53. doi: 10.1093/jmcb/mju001

Bader, G. D., and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* 4:2. doi: 10.1186/1471-2105-4-2

Calabrese, G. M., Mesner, L. D., Stains, J. P., Tommasini, S. M., Horowitz, M. C., Rosen, C. J., et al. (2017). Integrating GWAS and co-expression network data identifies bone mineral density genes SPTBN1 and MARK3 and an osteoblast functional module. *Cell Syst*. 4, 46–59. doi: 10.1016/j.cels.2016.10.014

Califano, A., Butte, A. J., Friend, S., Ideker, T., and Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nat. Genet*. 44, 841–847. doi: 10.1038/ng.2355

Chen, X., Xue, B., Wang, J., Liu, H., Shi, L., and Xie, J. (2018). Potassium channels: a potential therapeutic target for Parkinson's disease. *Neurosci. Bull*. 34, 341–348. doi: 10.1007/s12264-017-0177-3

Cheng, L., Jiang, Y., Ju, H., Sun, J., Peng, J., Zhou, M., and Hu, Y. (2018a). Infacront: calculating cross-ontology term similarities using information flow by a random walk. *BMC Genomics* 19:919. doi: 10.1186/s12864-017-4338-6

Cheng, L., Qi, C., Zhuang, H., Fu, T., and Zhang, X. (2020). gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res*. 48, D554-D560. doi: 10.1093/nar/gkz843

Cheng, L., Yang, H., Zhao, H., Pei, X., Shi, H., Sun, J., et al. (2019a). MetSigDis: a manually curated resource for the metabolic signatures of diseases. *Brief. Bioinformatics* 20, 203–209. doi: 10.1093/bib/bbx103

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019b). Computational methods for identifying similar diseases. *Mol. Ther. Nucleic Acids*. 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

Cheng, L., Zhuang, H., Yang, S., Jiang, H., Wang, S., and Zhang, J. (2018b). Exposing the causal effect of c-reactive protein on the risk of type 2 diabetes mellitus: a Mendelian randomization study. *Front. Genet*. 9:657. doi: 10.3389/fgene.2018.00657

Dehay, B., Martinez-Vicente, M., Caldwell, G. A., Caldwell, K. A., Yue, Z., Cookson, M. R., et al. (2013). Lysosomal impairment in Parkinson's disease. *Mov. Disord*. 28, 725–732. doi: 10.1002/mds.25462

Dong, X., Liao, Z., Gritsch, D., Hadzhiev, Y., Bai, Y., Locascio, J. J., et al. (2018). Enhancers active in dopamine neurons are a primary link between genetic variation and neuropsychiatric disease. *Nat. Neurosci*. 21, 1482–1492. doi: 10.1038/s41593-018-0223-0

dos Santos, M. T. A., and Smidt, M. P. (2011). En1 and Wnt signaling in midbrain dopaminergic neuronal development. *Neural Dev*. 6:23. doi: 10.1186/1749-8104-6-23

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Esteves, M., Cristóv ao, A. C., Saraiva, T., Rocha, S. M., Baltazar, G., Ferreira, L., et al. (2015). Retinoic acid-loaded polymeric nanoparticles induce neuroprotection in a mouse model for Parkinson's disease. *Front. Aging Neurosci*. 7:20. doi: 10.3389/fnagi.2015.00020

Fagny, M., Paulson, J. N., Kuijjer, M. L., Sonawane, A. R., Chen, C.-Y., Lopes-Ramos, C. M., et al. (2017). Exploring regulation in tissues with eqtl networks. *Proc. Natl. Acad. Sci. U.S.A*. 114, E7841-E7850. doi: 10.1073/pnas.1707375114

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., et al. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet*. 44, 502–510. doi: 10.1038/ng.2205

Fakhraei, S., Foulds, J., Shashanka, M., and Getoor, L. (2015). "Collective spammer detection in evolving multi-relational social networks," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW: ACM), 1769–1778. doi: 10.1145/2783258.2788606

Freed, J., and Chakrabarti, L. (2016). Defining a role for hemoglobin in Parkinson's disease. *NPJ Parkinson's Dis*. 2, 1–4. doi: 10.1038/npjparkd.2016.21

Frey, B. J., and Dueck, D. (2007). Clustering by passing messages between data points. *Science* 315, 972–976. doi: 10.1126/science.1136800

Gallagher, M. D., and Chen-Plotkin, A. S. (2018). The post-GWAS era: from association to function. *Am. J. Hum. Genet*. 102, 717–730. doi: 10.1016/j.ajhg.2018.04.002

Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Ghiassian, S. D., Menche, J., and Barabási, A.-L. (2015). A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome. *PLoS Comput. Biol*. 11:e1004120. doi: 10.1371/journal.pcbi.1004120

Girvan, M., and Newman, M. E. (2002). Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U.S.A*. 99, 7821–7826. doi: 10.1073/pnas.122653799

Grover, A., and Leskovec, J. (2016). "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA: ACM), 855–864. doi: 10.1145/2939672.2939754

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature 24277

Jacobs, F. M., Smits, S. M., Noorlander, C. W., von Oerthel, L., van der Linden, A. J., Burbach, J. P. H., et al. (2007). Retinoic acid counteracts developmental defects in the *Substantia nigra* caused by Pitx3 deficiency. *Development* 134, 2673–2684. doi: 10.1242/dev.02865

Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., et al. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet*. 51, 404–413. doi: 10.1038/s41588-018-0311-9

Krebs, V. (2004). *Books About Us Politics*. Unpublished. Available online at: http://www.orgnet.com

Lancichinetti, A., Fortunato, S., and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Phys. Rev. E* 78:046110. doi: 10.1103/PhysRevE.78.046110

Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 24, 719–720. doi: 10.1093/bioinformatics/btm563

Lehri-Boufala, S., Ouidja, M.-O., Barbier-Chasseféire, V., Hénault, E., Raisman-Vozari, R., Garrigue-Antar, L., et al. (2015). New roles of glycosaminoglycans in α-synuclein aggregation in a cellular model of Parkinson disease. *PLoS ONE* 10:e116641. doi: 10.1371/journal.pone.0116641

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Syst*. 1, 417–425. doi: 10.1016/j.cels.2015.12.004

Lin, G., Wang, L., Marcogliese, P. C., and Bellen, H. J. (2019). Sphingolipids in the pathogenesis of Parkinson's disease and Parkinsonism. *Trends Endocrinol. Metab*. 30, 106–117. doi: 10.1016/j.tem.2018.11.003

Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behav. Ecol. Sociobiol*. 54, 396–405. doi: 10.1007/s00265-003-0651-y

Mäkinen, V.-P., Civelek, M., Meng, Q., Zhang, B., Zhu, J., Levian, C., et al. (2014). Integrative genomics reveals novel molecular pathways and gene networks for coronary artery disease. *PLoS Genet*. 10:e1004502. doi: 10.1371/journal.pgen.1004502

Martin, B., De Maturana, R. L., Brenneman, R., Walent, T., Mattson, M. P., and Maudsley, S. (2005). Class II G protein-coupled receptors and their ligands in neuronal function and protection. *Neuromol. Med*. 7, 3–36. doi: 10.1385/NMM:7:1-2:003

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J., et al. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science* 347:1257601. doi: 10.1126/science.1257601

Mihara, T., Nakashima, M., Kuroiwa, A., Akitake, Y., Ono, K., Hosokawa, M., et al. (2008). Natural killer cells of Parkinson's disease patients are set up for activation: a possible role for innate immunity in the pathogenesis of this disease. *Parkinsonism Relat. Disord*. 14, 46–51. doi: 10.1016/j.parkreldis.2007.05.013

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., et al. (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet*. 34, 267–273. doi: 10.1038/ng1180

Mosley, R. L., Hutter-Saunders, J. A., Stone, D. K., and Gendelman, H. E. (2012). Inflammation and adaptive immunity in Parkinson's disease. *Cold Spring Harb. Perspect. Med*. 2:a009381. doi: 10.1101/cshperspect.a009381

Nalls, M. A., Blauwendraat, C., Vallerga, C. L., Heilbron, K., Bandres-Ciga, S., Chang, D., et al. (2019). Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol*. 18, 1091–1102. doi: 10.1016/S1474-4422(19)30320-5

Nepusz, T., Petróczi, A., Négyessy, L., and Bazsó, F. (2008). Fuzzy communities and the concept of bridgeness in complex networks. *Phys. Rev. E* 77:016107. doi: 10.1103/PhysRevE.77.016107

Nicolas, C. S., Amici, M., Bortolotto, Z. A., Doherty, A., Csaba, Z., Fafouri, A., et al. (2013). The role of JAK-STAT signaling within the CNS. *JAK-STAT* 2:e22925. doi: 10.4161/jkst.22925

Parker, W. D. Jr, Parks, J. K., and Swerdlow, R. H. (2008). Complex I deficiency in Parkinson's disease frontal cortex. *Brain Res*. 1189, 215–218. doi: 10.1016/j.brainres.2007.10.061

Peng, J., Hui, W., Li, Q., Chen, B., Hao, J., Jiang, Q., et al. (2019a). A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* 35, 4364–4371. doi: 10.1093/bioinformatics/btz254

Peng, J., Lu, J., Hoh, D., Dina, A. S., Shang, X., Kramer, D. M., et al. (2020a). Identifying emerging phenomenon in long temporal phenotyping experiments. *Bioinformatics* 36, 568–577. doi: 10.1093/bioinformatics/btz559

Peng, J., Wang, X., and Shang, X. (2019b). Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-seq data. *BMC Bioinformatics* 20:284. doi: 10.1186/s12859-019-2769-6

Peng, J., Xue, H., Wei, Z., Tuncali, I., Hao, J., and Shang, X. (2020b). Integrating multi-network topology for gene function prediction using deep neural networks. *Brief. Bioinformatics* bbaa036. doi: 10.1093/bib/bbaa036

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 701–710. doi: 10.1145/2623330.2623732

Placido, A., Pereira, C., Duarte, A., Candeias, E., Correia, S., Santos, R., et al. (2014). The role of endoplasmic reticulum in amyloid precursor protein processing and trafficking: implications for Alzheimer's disease. *Biochim. Biophys. Acta* 1842, 1444–1453. doi: 10.1016/j.bbadis.2014.05.003

Sharma, A., Menche, J., Huang, C. C., Ort, T., Zhou, X., Kitsak, M., et al. (2015). A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet*. 24, 3005–3020. doi: 10.1093/hmg/ddv001

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell*. 22, 888–905. doi: 10.1109/34.868688

Su, G., Kuchinsky, A., Morris, J. H., States, D. J., and Meng, F. (2010). Glay: community structure analysis of biological networks. *Bioinformatics* 26, 3135–3137. doi: 10.1093/bioinformatics/btq596

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

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

Wang, M.-M., Miao, D., Cao, X.-P., Tan, L., and Tan, L. (2018). Innate immune activation in Alzheimer's disease. *Ann. Transl. Med*. 6:177. doi: 10.21037/atm.2018.04.20

Wang, S., Zhang, S., Liou, L.-C., Ren, Q., Zhang, Z., Caldwell, G. A., et al. (2014). Phosphatidylethanolamine deficiency disrupts α-synuclein homeostasis in yeast and worm models of Parkinson disease. *Proc. Natl. Acad. Sci. U.S.A*. 111, E3976–E3985. doi: 10.1073/pnas.1411694111

Wang, T., Peng, J., Peng, Q., Wang, Y., and Chen, J. (2020). FSM: Fast and scalable network motif discovery for exploring higher-order network organizations. *Methods* 173, 83–93. doi: 10.1016/j.ymeth.2019.07.008

Wang, T., Peng, Q., Liu, B., Liu, X., Liu, Y., Peng, J., and Wang, Y. (2019a). eQTLMAPT: fast and accurate eQTL mediation analysis with efficient permutation testing approaches. *Front. Genet*. 10:1309. doi: 10.3389/fgene.2019.01309

Wang, T., Ruan, J., Yin, Q., Dong, X., and Wang, Y. (2019b). "An automated quality control pipeline for eQTL analysis with RNA-seq data," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (San Diego, CA: IEEE), 1780–1786. doi: 10.1109/BIBM47256.2019.8983006

Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun*. 8:1826. doi: 10.1038/s41467-017-01261-5

Westra, H.-J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet*. 45, 1238–1243. doi: 10.1038/ng.2756

Wu, X., and Reddy, D. S. (2012). Integrins as receptor targets for neurological disorders. *Pharmacol. Therap*. 134, 68–81. doi: 10.1016/j.pharmthera.2011.12.008

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *J. Anthropol. Res*. 33, 452–473. doi: 10.1086/jar.33.4.3629752

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 September 2020
doi: 10.3389/fgene.2020.539292

Weighted Gene Co-expression Network Analysis of Key Biomarkers Associated With Bronchopulmonary Dysplasia

Yao Cai[1]†, Fei Ma[1]†, LiuHong Qu[2,3]†, Binqing Liu[1], Hui Xiong[1], Yanmei Ma[4], Sitao Li[1]*‡ and Hu Hao[1]*‡

[1] Department of Neonatology, The Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China, [2] Department of Neonatology, The Maternal and Child Health Care Hospital of Huadu, Guangzhou, China, [3] Huadu Affiliated Hospital of Guangdong Medical University, Guangzhou, China, [4] Laboratory of Inborn Metabolism Errors, The Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou, China

Bronchopulmonary dysplasia (BPD) is a complex disorder resulting from interactions between genes and the environment. The accurate molecular etiology of BPD remains largely unclear. This study aimed to identify key BPD-associated genes and pathways functionally enriched using weighted gene co-expression network analysis (WGCNA). We analyzed microarray data of 62 pre-term patients with BPD and 38 pre-term patients without BPD from Gene Expression Omnibus (GEO). WGCNA was used to construct a gene expression network, and genes were classified into definite modules. In addition, the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analyses of BPD-related hub genes were performed. Firstly, we constructed a weighted gene co-expression network, and genes were divided into 10 modules. Among the modules, the yellow module was related to BPD progression and severity and included the following hub genes: *MMP25*, *MMP9*, *SIRPA*, *CKAP4*, *SLCO4C1*, and *SLC2A3*; and the red module included some co-expression molecules that displayed a continuous decline in expression with BPD progression and included the following hub genes: *LEF1*, *ITK*, *CD6*, *RASGRP1*, *IL7R*, *SKAP1*, *CD3E*, and *ICOS*. GO and KEGG analyses showed that high expression of inflammatory response-related genes and low expression of T cell receptor activation-related genes are significantly correlated with BPD progression. The present WGCNA-based study thus provides an overall perspective of BPD and lays the foundation for identifying potential pathways and hub genes that contribute to the development of BPD.

Keywords: bronchopulmonary dysplasia, weighted gene co-expression network analysis, hub gene, biological process, biomarkers

OPEN ACCESS

Edited by:
Zhongyu Wei,
Fudan University, China

Reviewed by:
Huiyun Gu,
Wuhan University, China
Xiaoguang Zhou,
Nanjing Children's Hospital, China

****Correspondence:***
Sitao Li
sitaoli918@126.com
Hu Hao
freehaotiger@126.com

†These authors have contributed equally to this work

‡These authors have contributed equally to this work and share senior authorship

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

Received: *29 February 2020*
Accepted: *18 August 2020*
Published: *09 September 2020*

Citation:
Cai Y, Ma F, Qu L, Liu B, Xiong H, Ma Y, Li S and Hao H (2020) Weighted Gene Co-expression Network Analysis of Key Biomarkers Associated With Bronchopulmonary Dysplasia. Front. Genet. 11:539292. doi: 10.3389/fgene.2020.539292

Abbreviations: AUC, area under the curve; BPD, bronchopulmonary dysplasia; GEO, Gene Expression Omnibus; GS, gene significance; MEs, module eigengenes; MM, module group members; PPI, protein–protein interaction; ROC, receiver operating characteristic; TOM, topological overlap matrix; WGCNA, weighted gene co-expression network analysis.

INTRODUCTION

Bronchopulmonary dysplasia (BPD) is a chronic lung disease in pre-term infants that is characterized by arrested lung development due to early lung injury (Jobe, 1999; Speer, 2006a). Since its first description in Northway et al. (1967), with the survival of increasing number of premature babies having very low birth weights, the incidence of BPD has remained high. However, few specific treatments are available for reducing the burden of the disease (Lemons et al., 2001; Farstad et al., 2011). Survivors of BPD have an increased risk of pulmonary hypertension, growth retardation, neurodevelopmental delay, and other long-term sequelae that have a major impact on families and health care system (Bhandari and McGrath, 2013). However, the mechanism of BPD formation is complex and includes many processes, such as inflammation, oxidative stress, abnormal angiogenesis, and damaged lung repair. Some of these processes remain to be elucidated (Coalson, 2003; Chess et al., 2006; Collins et al., 2017; Yang et al., 2017).

Weighted gene co-expression network analysis (WGCNA) is an algorithm that defines modules of genes with similar expression patterns in complex diseases (Langfelder and Horvath, 2008). WGCNA can effectively integrate gene expression and clinical trait data to appraise functional pathways and candidate molecular biomarkers (Presson et al., 2008). WGCNA facilitates a global interpretation of gene expression data through the construction of gene networks based on the similarity of expression profiles among samples (Oldham et al., 2008). WGCNA has been used for the study of gene-network signatures, co-expression modules, and hub genes involved in human respiratory syncytial virus infection (Vieira et al., 2019), autoimmune diseases (Medina and Lubovac-Pilav, 2016; Ma et al., 2017), and various cancers (Giulietti et al., 2016; Tian et al., 2019). A hub gene is a key gene that plays a vital role in regulatory pathways; the regulation of other genes is often affected by this gene (Luscombe et al., 2004). Thus, WGCNA can be used to elucidate gene-network signatures and hub genes associated with BPD to better understand the pathogenesis of this disease. To our knowledge, however, WGCNA, as a system biology approach, has not been applied to the analysis of BPD-derived data thus far.

In the present study, we used WGCNA to explore the gene-network signatures of peripheral blood from pre-term infants with and without BPD. The pathogenesis of BPD was explored using pathway enrichment analysis to investigate the biological pathways and key hub genes that were associated with BPD. Finally, enrichment analysis was used to determine the potential functions of these hub genes and to identify key genes potentially involved in the pathogenesis of BPD.

MATERIALS AND METHODS

Data Collection and Preprocessing

A flow chart illustrating the data preparation, processing, and analysis is displayed in **Figure 1**. We used 'bronchopulmonary dysplasia' as the key word to search the Gene Expression Omnibus (GEO) database and to select datasets containing samples from different pathological stages and normal controls. Finally, the dataset GSE32472[1], was found to meet our requirements and was therefore downloaded. To identify the molecular networks and hub genes related to the pathological progress of BPD, WGCNA was conducted. GSE32472 provided microarray profiles of blood samples of newborns with BPD, including microarray assessment of gene expression at approximately the 5th, 14th, and 28th days of life. To ensure the stability of the selection, 100 blood samples at about the 28th day were selected, when a more definite diagnosis of BPD had been made. These samples were obtained from 38 controls, and 38 mild, 10 moderate, and 14 severe BPD cases. The expression data were normalized using quantile normalization function in limma package of R software (Ritchie et al., 2015). The genes with the highest variance in expression values (top 25%) were selected for co-expression network construction. Cluster analysis using the Pearson's correlation matrices and the average linkage method were conducted to detect whether outlier samples existed for the purpose of ensuring the reliability of the network construction. A brief design of the study is shown in **Figure 1A**.

Construction of a Co-expression Network

A co-expression network was constructed using the WGCNA algorithm package in R (Presson et al., 2008). First, the Pearson's correlation matrices were constructed for all pair-wise genes. Next, a weighted adjacency matrix was constructed by using the power function amn = $|$ Cmn$|^{\beta}$ (Cmn = Pearson correlation between gene m and gene n; a_{mn} = adjacency between gene m and gene n) (Zhang and Horvath, 2005). The parameter β served as a soft threshold parameter to expand strong correlations and penalize weak correlations between genes. To ensure a scale-free topology of the network, β was selected when the scale independence value was equal to 0.9. The adjacency was transformed into a topological overlap matrix (TOM) to measure the network connectivity of a gene, which is defined as the sum of its adjacency with all other genes. Hierarchical clustering was performed according to TOM-based dissimilarity to distribute genes with similar expression patterns into modules with a minimum cluster size of 50 (Ravasz et al., 2002). Highly similar modules were merged with a cut-off of 0.25.

Identification of Modules Significantly Associated With BPD Severity

Module eigengenes (MEs) were considered the major component in the principal component analysis for each gene module, and the expression patterns of all genes could be summarized into a single characteristic expression profile within a given module. To identify modules significantly associated with BPD severity, the correlation between MEs and BPD stage was evaluated by the Pearson's correlation test with $p < 0.05$ as the cut-off. The modules most significantly related to BPD severity were considered as key modules and subjected to further analysis.

[1] http://www.ncbi.nlm.nih.gov/geo/

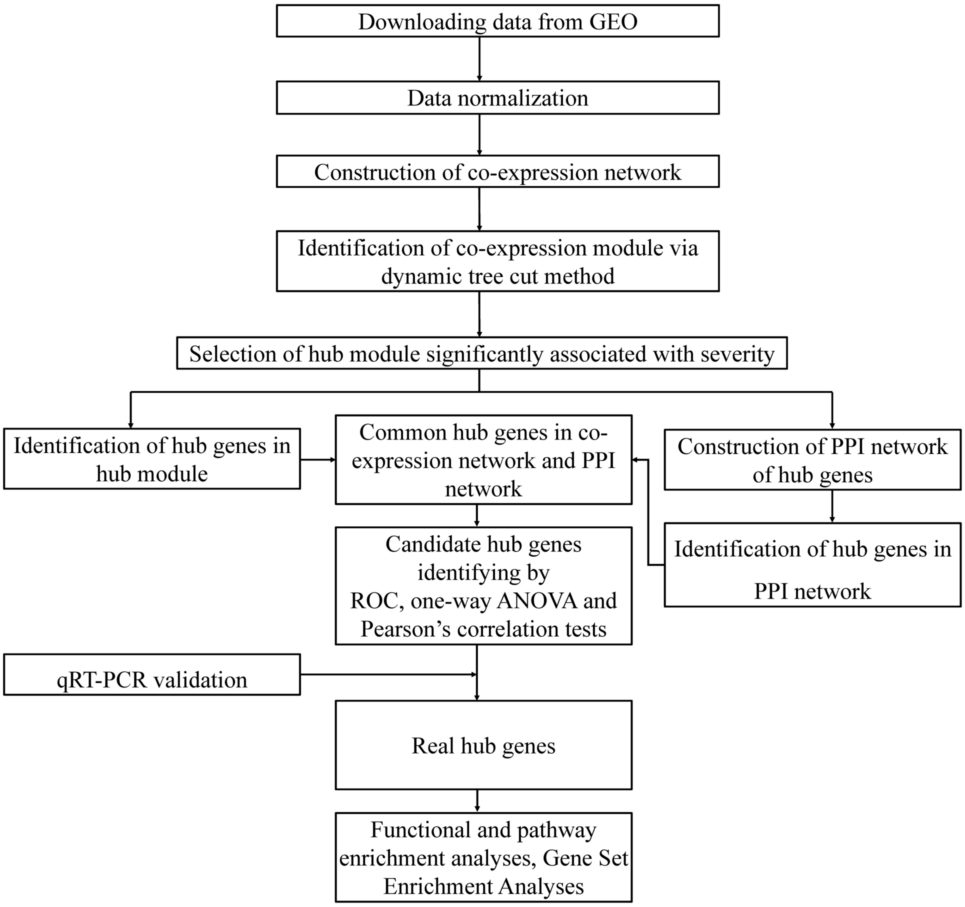

Downloading data from GEO
Data normalization
Construction of co-expression network
Identification of co-expression module via dynamic tree cut method
Selection of hub module significantly associated with severity
Identification of hub genes in hub module
Common hub genes in co-expression network and PPI network
Construction of PPI network of hub genes
Identification of hub genes in PPI network
Candidate hub genes identifying by ROC, one-way ANOVA and Pearson's correlation tests
qRT-PCR validation
Real hub genes
Functional and pathway enrichment analyses, Gene Set Enrichment Analyses

FIGURE 1 | Outline of the study design.

Identification of Candidate Hub Genes

A module hub gene is a highly connected in-module gene that has the highest module member (MM) score of its corresponding module (Horvath and Dong, 2008). The MM score for every gene was calculated by the WGCNA function KME, which correlates the expression profile of a gene with the ME of a module to quantify the relationship between a gene and a given module. The absolute value of gene significance (GS) represents the Pearson's correlation between a given gene and clinical features. We removed hub genes based on the cut-off criteria ($|\text{MM}| \geq 0.85$, $|\text{GS}| \geq 0.45$). Further, all genes in key modules were uploaded to STRING[2] to acquire information about the interaction between genes. Protein–protein interaction (PPI) networks were constructed with the species limited to '*Homo sapiens*' and a confidence > 0.9. In the PPI network, genes with a degree ≥ 10 were defined as hub nodes. Hub genes common in both co-expression network and PPI network were selected for candidate hub genes identification.

Hub genes common in both co-expression network and PPI network were analyzed by ROC curve, and area under the curve (AUC) was calculated to distinguish the control group from the BPD group. In addition, one-way ANOVA and Pearson's correlation tests were conducted to explore the relevance of the hub genes common in both co-expression network and PPI network in terms of disease severity. Candidate hub genes were identified using the following criteria: (1) a significant *P* value in the one-way ANOVA test and the Pearson's correlation and (2) an AUC > 0.8.

qRT-PCR Validation and Real Hub Genes Identification

To validate the candidate hub genes obtained by WGCNA, pre-term infants with or without BPD blood samples were collected from the Department of Neonatology of the Sixth Affiliated Hospital of Sun Yat-sen University. This research was approved by the ethics review board of Sixth Affiliated Hospital of Sun Yat-sen University (2019ZSLYEC-80), and written informed consent was provided by the participants' legal guardians. From each sample, 100 ng of cDNA was obtained for RT-PCR amplification reaction, and the expression of an endogenous control (housekeeping gene:

[2] http://string-db.org

GAPDH) was used for the determination of the relative expression levels of the hub genes. Primer sequences for related hub genes are listed in **Supplementary Table S1**. Real hub genes were identified if the results of RT-PCR have significant difference.

Functional and Pathway Enrichment Analyses

To gain further insights into the functions of hub genes in the module most related to BPD, we performed biological process analysis and KEGG pathway enrichment analysis with 'c2.cp.kegg.v7.1.symbols' as background[3].

Gene Set Enrichment Analysis (GSEA) for Hub Biological Pathways Confirmation

Mapping to KEGG (Kyoto Encyclopedia of Genes and Genomes) database[4], GSEA[5] (Subramanian et al., 2007) was performed between control and BPD groups to confirm the expression pattern of hub biological pathways.

Statistical Analysis

Non-parametric tests or *t*-tests based on data distribution characteristics were used to analyze the statistical significance of the difference in hub gene expression levels between the two groups. Analyses were conducted in GraphPad Prism 8.0.2. $P < 0.05$ was considered statistically significant.

RESULTS

Weighted Co-expression Network Construction

Figure 1 shows the flow chart of data preparation, processing, analysis, and validation for this study. The data were normalized using the limma package of R software (**Figure 2A** and R code in **Supplementary Table S2**). The co-expression analysis included 100 samples with clinical information, sample information, and expression matrix. Input files are provided in **Supplementary Tables S3, S4**. Sample clustering was performed based on Pearson's correlation matrices and the average linkage method. No outliers were detected (**Figure 2B**). The genes showing the highest expression variance (top 25%) were selected for subsequent WGCNA using the WGCNA package in R software. Genes with similar expression patterns were then grouped by average linkage hierarchical clustering. In our study, $\beta = 23$ (scale-free $R^2 > 0.901$) was selected as the soft threshold to ensure a scale-free network (**Figures 3A,B**). Next, we constructed a systematic clustering tree using the WGCNA package. In **Figure 3C**, each short vertical line represents a gene, and each color represents one module composed of genes with similar expression patterns. The genes shown in gray were the genes that could not be merged. A total of 10 modules were identified (**Figure 3C**).

Identification of Key Modules Associated With BPD Severity

We tested the relevance of each module for BPD clinical information, focusing on different BPD stages. As displayed in **Figure 3D**, the yellow module (P = 5e-08, R^2 = 0.51) was most significantly and positively correlated with BPD severity, whereas the red module (P = 3e-11, R^2 = −0.60) showed the opposite result. The correlation between the yellow module and BPD severity gradually increased and finally became positive. The red module showed the opposite pattern. Based on the above findings, the red and yellow modules were identified as key modules correlated with BPD severity and were thus, further analyzed.

PPI Network Construction With Corresponding Module Genes

PPI networks of the red and yellow modules were constructed with a cutoff confidence > 0.9 (**Figures 4A,B**). A total of 31 genes in the red module and 41 genes in the yellow module were identified with a degree $\geq$ 10 as hub genes in each PPI network. Based on $|$ MM$|$ $\geq$ 0.85 and $|$ GS$|$ $\geq$ 0.45, a total of 76 genes in the red module and 77 genes in the yellow module were selected as hub genes in each co-expression network (**Figures 4C,D**). A total of 21 genes in the red module and 13 genes in the yellow module were identified in both the PPI and co-expression networks (**Figures 4E,F**). All GS, MM, and intramodule connectivity values of each identified module are listed in **Supplementary Tables S5–S7.**

Identification of Real Hub Genes

All the 21 candidate hub genes in the red module showed significance in the one-way ANOVA. A total of nine genes had an AUC $\geq$ 0.80, and 19 showed a significant correlation with disease severity in the Pearson's correlation analysis. Eventually, nine genes in the red module with an AUC $\geq$ 0.80 and significant P values in the Pearson's correlation as well as one-way ANOVA were regarded as candidate hub genes (**Figure 4G**). Similarly, of the 13 candidate hub genes in the yellow module, 12 showed significance in the one-way ANOVA, 11 genes had an AUC $\geq$ 0.80, and 10 showed a significant correlation with disease severity in the Pearson's correlation analysis. Nine genes in the yellow module had an AUC $\geq$ 0.80 and significant P values in the Pearson's correlation as well as in the one-way ANOVA, and were thus, selected as candidate hub genes (**Figure 4H**). Detailed information about the red and yellow modules in relation to the Pearson's correlation, ROC, and one-way ANOVA has been provided in **Supplementary Tables S8–S11.** The severity plot for the candidate hub genes is shown in **Figure 5A**. The expression levels of candidate hub genes in the yellow module increased with disease severity, and the expression

[3] https://www.gsea-msigdb.org/gsea/downloads.jsp

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

[5] http://software.broadinstitute.org/gsea/index.jsp

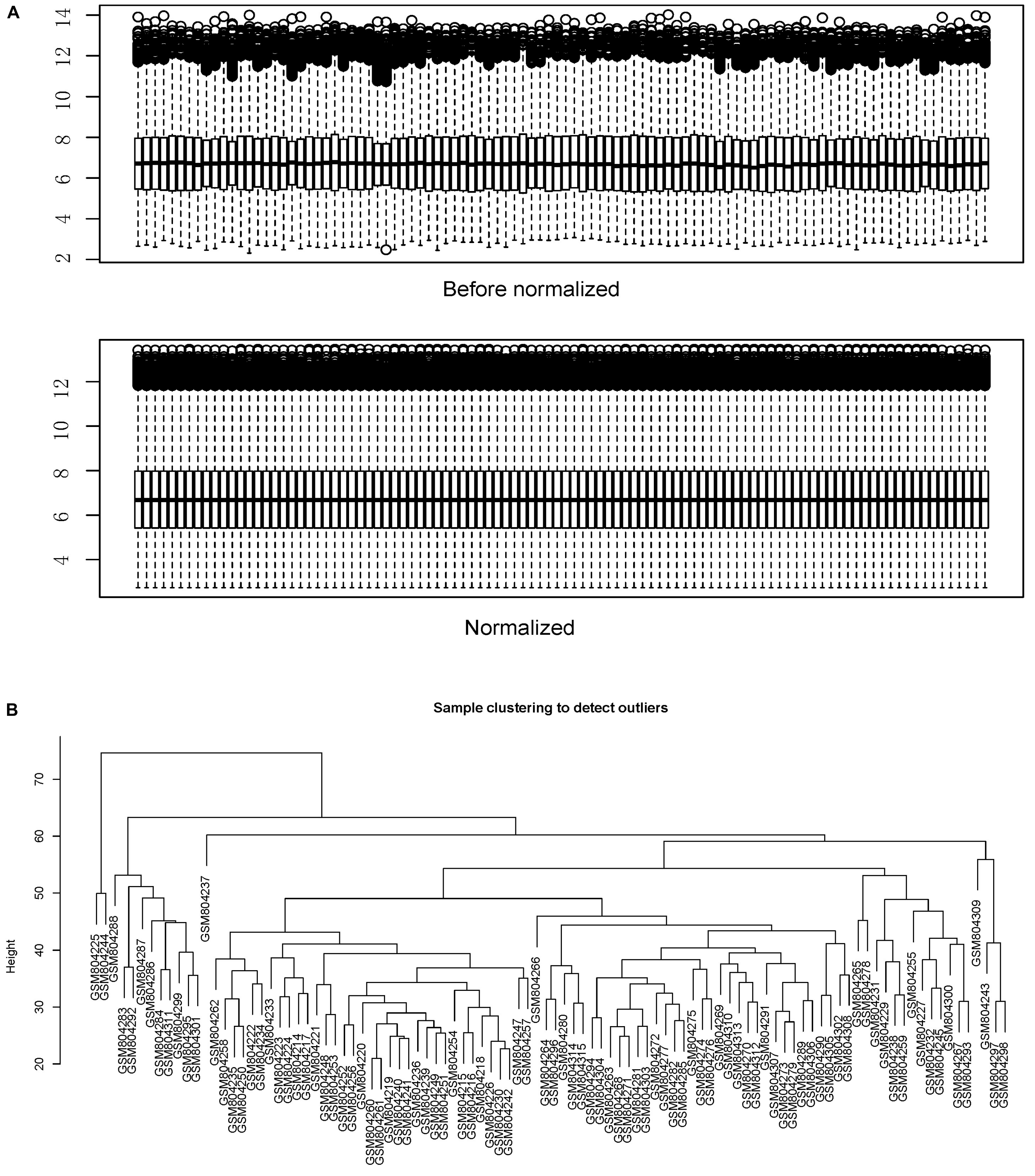


FIGURE 2 | Data normalization and sample clustering dendrogram. **(A)** Data were normalized using the limma package of the R software. **(B)** Sample clustering was performed using the Pearson's correlation matrices and the average linkage method.

levels of candidate hub genes in this module were significantly increased in different BPD severity conditions compared with those of normal controls. In contrast, candidate hub genes in the red module showed decreasing expression levels with greater disease severity, and markedly decreased expression levels in different BPD severity conditions (**Figure 5B**). To further

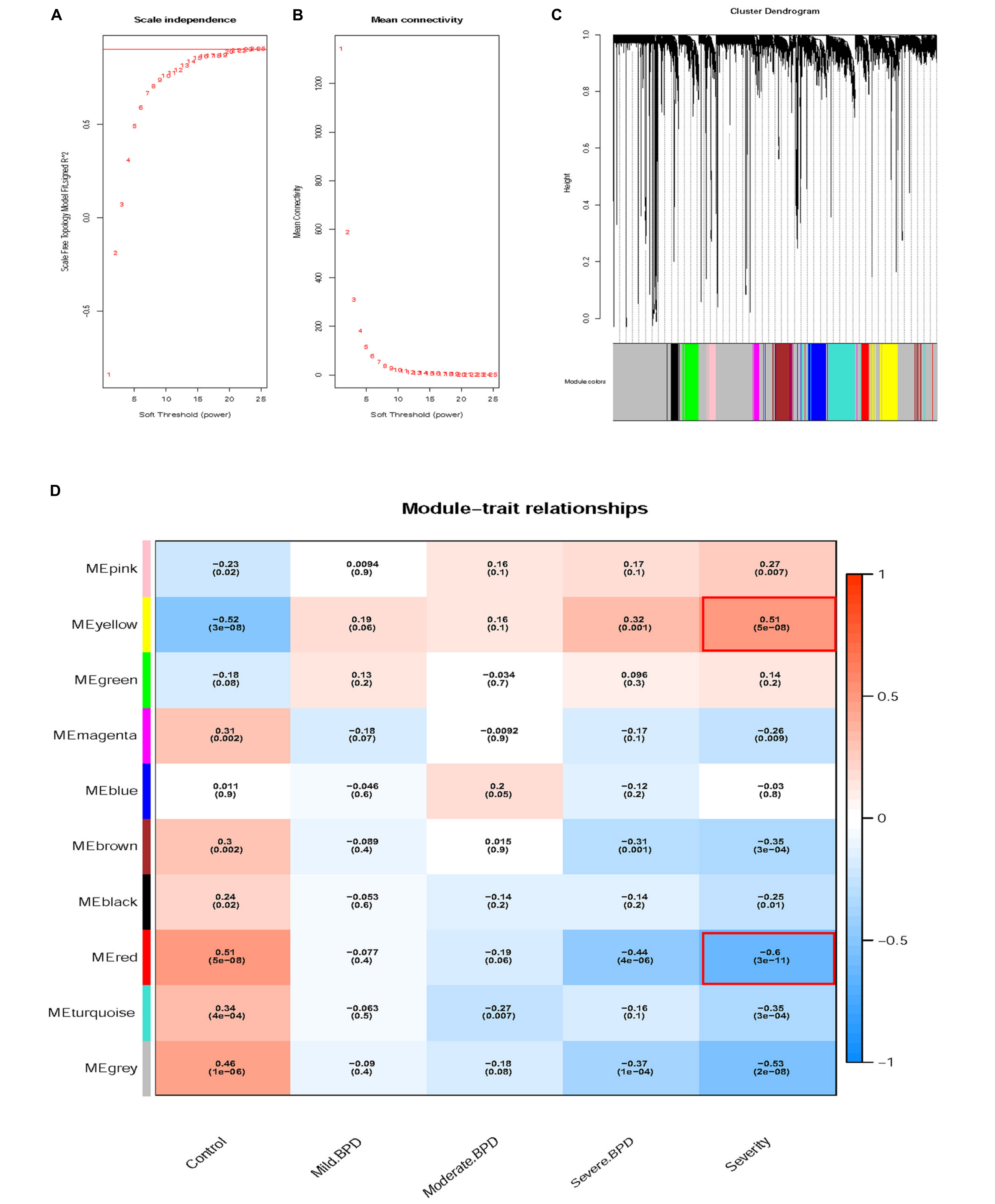


FIGURE 3 | Determination of soft-thresholding power and grouping of genes with similar expression into modules using weighted gene co-expression network analysis (WGCNA). **(A)** Analysis of the scale-free fit index for soft-thresholding powers (β). **(B)** Analysis of the mean connectivity for soft-thresholding powers. **(C)** Dendrogram of clustered genes. **(D)** Identification of modules associated with clinical information.

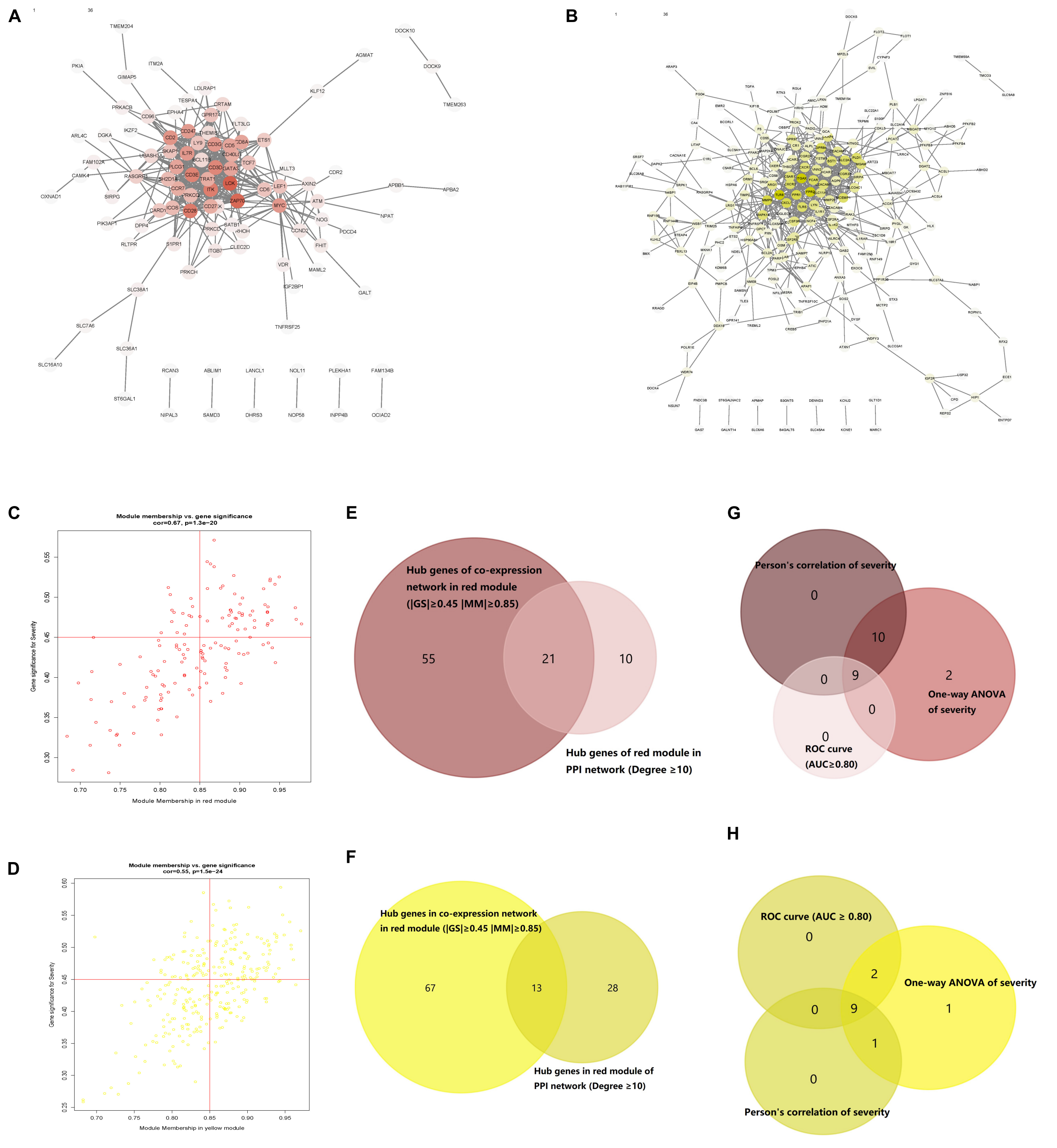


FIGURE 4 | Protein–protein interaction (PPI) networks of genes corresponding to the two key modules. **(A)** PPI network of the nodes in the red module. **(B)** PPI network of the nodes in the yellow module. **(C)** Scatter plot of module eigengenes (MEs) in the red module. **(D)** Scatter plot of MEs in the yellow module. **(E)** Common red hub genes in the co-expression and PPI networks. **(F)** Common yellow hub genes in the co-expression and PPI networks. **(G)** Common genes in the red module shared characteristics with an area under the curve (AUC) ≥ 0.80 and had significant *P* values in the Pearson's correlation and one-way ANOVA tests. **(H)** Common genes in the yellow module shared characteristics with an AUC ≥ 0.80 and had significant *P* values in the Pearson's correlation and one-way ANOVA tests.

clarify the clinical significance and identify real hub genes, we collected the BPD patient's blood for qRT-PCR validation *in vitro*. The results showed that most of these genes had statistically significant differences and were considered as real hub genes, except for MAPK14, CEACAM3, CSF2RB, and CD3G (**Figure 6**).

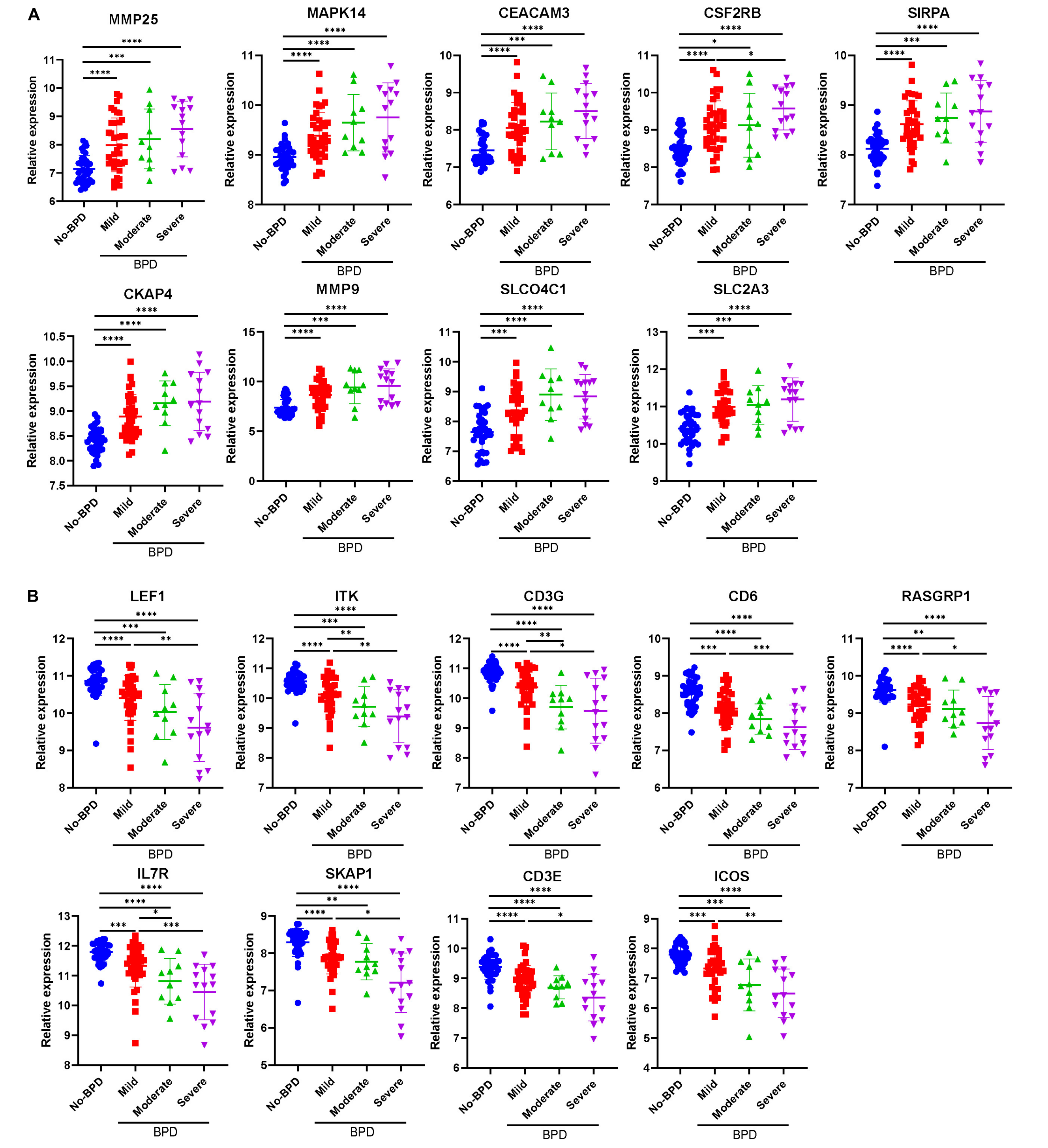


FIGURE 5 | Severity plot of the real hub genes. **(A)** Severity plot of the identified hub genes in the yellow module. **(B)** Severity plot of the identified hub genes in the red module. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, and ****$P < 0.0001$.

Functional and Pathway Enrichment Analysis

To learn more about the function of the identified hub genes, they were subjected to perform the biological process and KEGG pathway enrichment analyses. Real hub genes in the red module, which exhibited a negative correlation with disease severity, were significantly enriched in 20 BPs: T cell activation, positive regulation of leukocyte cell-cell adhesion, positive regulation

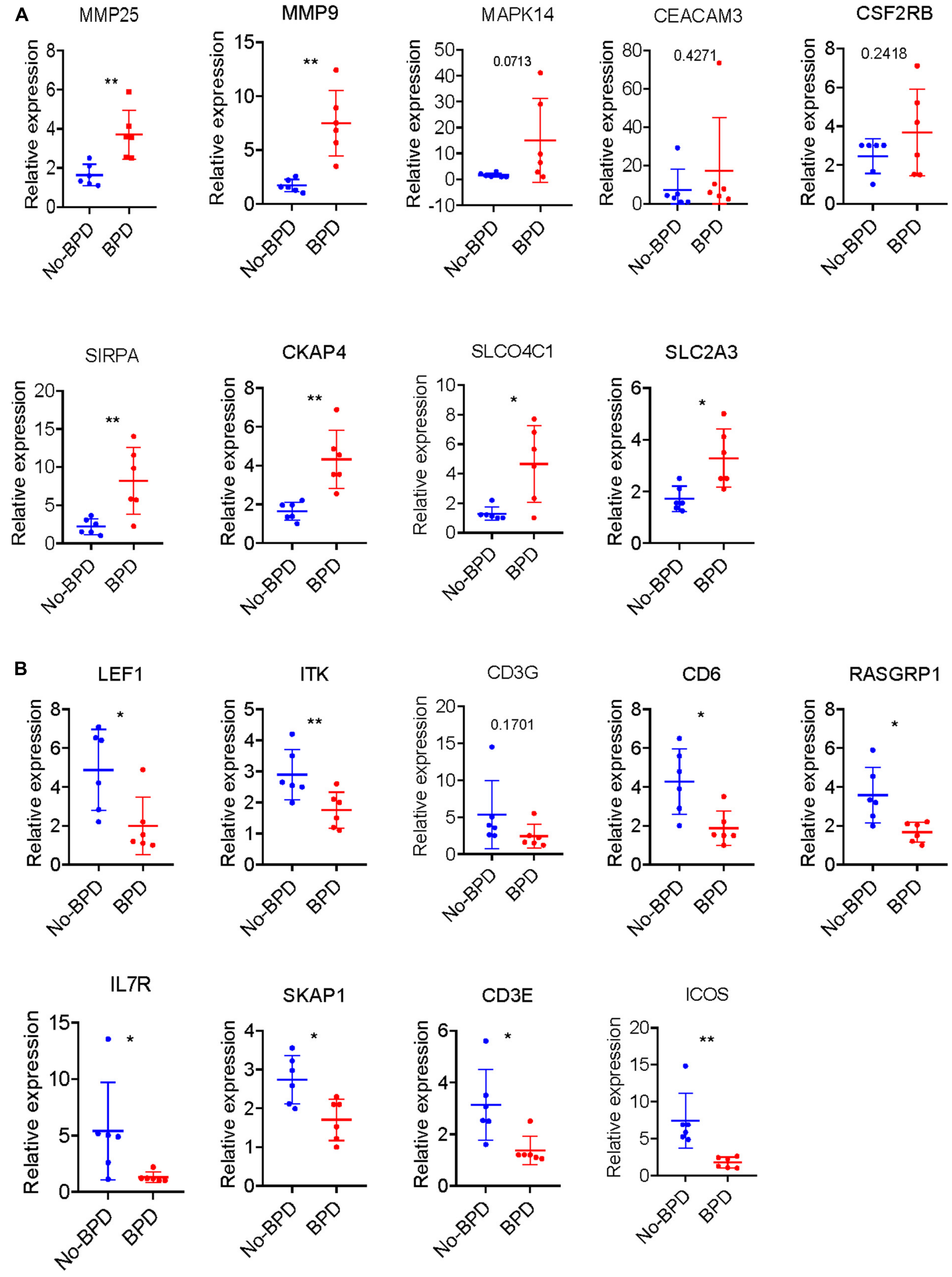


FIGURE 6 | Validation of hub genes by qRT-PCR. **(A)** Severity plot of the identified hub genes in the yellow module analyses by qRT-PCR. **(B)** Severity plot of the identified hub genes in the red module analyses by qRT-PCR. **P* < 0.05 and ***P* < 0.01.

of cell-cell adhesion, regulation of leukocyte cell-cell adhesion, leukocyte cell-cell adhesion, positive regulation of cell adhesion, regulation of cell-cell adhesion, positive regulation of T cell activation, regulation of cell-cell adhesion, positive regulation of T cell activation, positive regulation of lymphocyte activation, positive regulation of leukocyte activation, regulation of T cell activation, positive regulation of cell activation, regulation of lymphocyte activation, regulation of leukocyte activation, T cell differentiation, lymphocyte differentiation, interleukin-4 production, T cell differentiation in thymus, positive regulation of T cell differentiation in thymus, and regulation of cell-cell adhesion mediated by integrin (**Figure 7A**). The real hub genes were also enriched in three KEGG pathways: T cell receptor signaling pathway, primary immunodeficiency, and hematopoietic cell lineage (**Figure 7B**). The real hub genes in the yellow module, which showed a positive correlation with disease severity, were enriched in 4 BPs, neutrophil degranulation, neutrophil activation involved in immune response, neutrophil mediated immunity, and neutrophil activation (**Figure 7C**). KEGG pathway enrichment analysis showed that real hub genes in the yellow module were enriched in the bladder cancer pathway, and leukocyte transendothelial migration (**Figure 7D**).

Gene Set Enrichment Analysis for Hub Biological Pathways Confirmation

As bladder cancer pathway is not related to BPD, it was not considered in GSEA confirmation. According to the results of the GSEA, the terms 'T cell receptor signaling pathway' and 'primary immunodeficiency' were significantly enriched in the

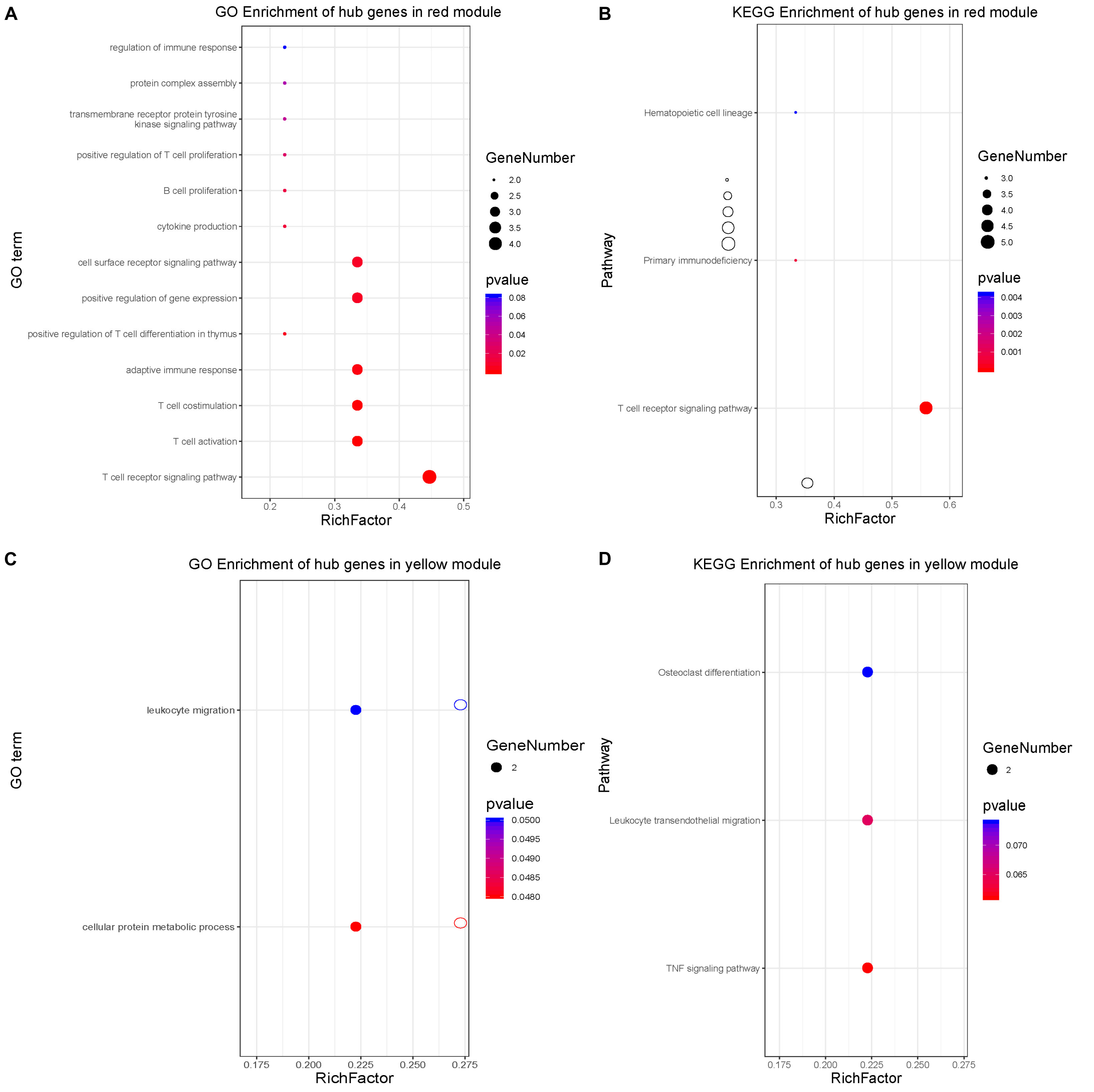


FIGURE 7 | Biological process and KEGG pathway enrichment analyses of hub genes. **(A)** Biological process analysis for the hub genes in the red module. **(B)** KEGG pathway enrichment for the hub genes in the red module. **(C)** Biological process enrichment for the hub genes in the yellow module. **(D)** KEGG pathway enrichment for the hub genes in the yellow module.

control group while the term 'hematopoietic cell lineag' was not (**Figures 8A,B** and **Supplementary Table S12**). On the contrary, the term 'leukocyte transendothelial migration' was significantly enriched in the BPD group (**Figure 8C** and **Supplementary Table S13**). These results successfully confirmed the expression pattern of hub biological pathways.

DISCUSSION

To our knowledge, our study reports the first application of WGCNA to construct a BPD-related gene-network. We found two key gene modules and several hub genes that were associated with BPD progression. This research provides new insights into the molecular etiology of BPD, as well as potential therapeutic targets for this disease. Ten co-expression modules were obtained through WGCNA. The yellow module was associated with progression and severity of BPD and the red module included co-expressed genes that displayed a continuous decline in expression with BPD progression.

Among the 10 modules, the yellow module was especially involved in BPD pathogenesis. Some genes showed greater positive association with the progression of BPD including *MMP25*, *MMP9*, *SIRPA*, *CKAP4*, *SLCO4C1*, and *SLC2A3*. The red module contained genes showing greater negative association with the progression of BPD including *LEF1*, *ITK*, *CD6*, *RASGRP1*, *IL7R*, *SKAP1*, *CD3E*, and *ICOS*. These genes can be considered as hub genes and also play important roles in other co-expression modules.

Functional enrichment analysis is widely used to classify biological entities into functionally related groups (Rue-Albrecht et al., 2016). In the present study, we used the GO and KEGG analyses to elucidate the biological functions of hub genes in the yellow module, that were significantly up-regulated with the increase of BPD severity. The genes in the yellow module were mainly enriched in the response to cellular protein metabolic processes, leukocyte migration, and TNF signaling pathway. The inflammatory response plays critical roles in the development of BPD (Shahzad et al., 2016). Consistent with previous reports (Ma et al., 2017), we found a significant increase in levels of *MMP9* and *MMP25* in infants with BPD compared with those in infants without BPD. This consistency not only further demonstrates the reliability of our results, but also provides additional confirmation of the pivotal role of MMP proteins in BPD progression. Disease-related gene expression analysis revealed signaling pathways involved in BPD progression, including protein kinase A, MAPK, and neuromodulin/epidermal growth factor receptor signal. In a newborn Sprague-Dawley rat BPD model, activation of the MAPK and PI3K/AKT signaling pathways in lung tissues was monitored during prolonged exposure of newborn rats to hyperoxia (Liu et al., 2018). This previous study suggested that MAPK14 could be used as a biological marker to monitor disease progression.

The most notable down-regulated pathway in BPD progression is the T cell receptor signaling pathway. Our data showed that the expression of T cell receptor molecules, including *CD3E*, *CD6*, and *ICOS*, decreased significantly during BPD progression. These molecules had not been confirmed in previous studies. T cell response depends on the type of ligand that binds to the receptor, the duration of cooperation, and the presence of co-receptors or co-inhibitors (Cheng et al., 2011; James et al., 2011). In our study, transcription factors and related pathways, such as *CD3E*, *CD6*, and *ICOS*, were under-expressed in children with BPD, suggesting that reduced T receptor expression may lead to decreased receptor density at the cell surface, which in turn may be a risk factor for bacterial translocation and further infection. These results are consistent with the fact that pulmonary infection is a risk factor for BPD (Speer, 2006b).

Enrichment analysis revealed the signaling pathways that may be related to the pathogenesis of the disease. The results can be considered in two ways. One is by placing our findings in the context of the existing knowledge, and the other is by studying genes known to be potentially involved in the pathogenic mechanism of BPD. The overexpression of pathways involved in inflammatory cytokine production and leukocyte migration in the present study confirms the generally accepted

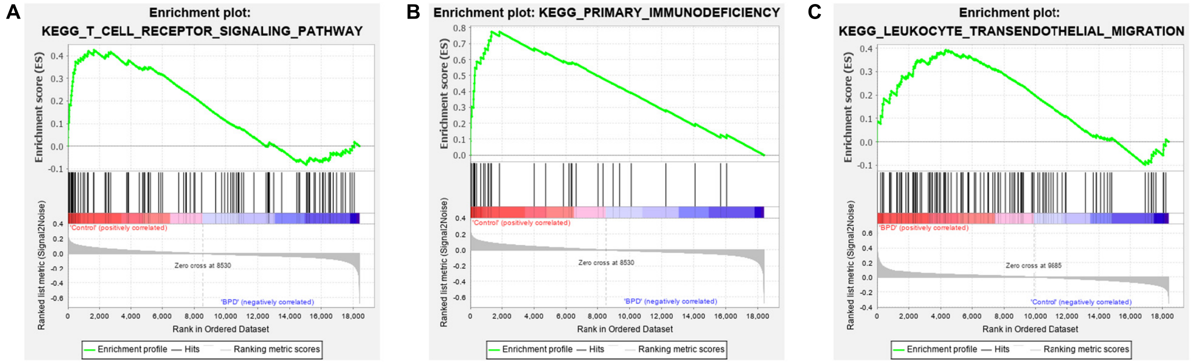


FIGURE 8 | Gene set enrichment analysis of real hub genes. **(A)** Enrichment result of 'T cell receptor signaling pathway' between the control group and the bronchopulmonary dysplasia (BPD) group. **(B)** Enrichment result of the term 'primary immunodeficiency' between the control group and the BPD group. **(C)** Enrichment result of 'leukocyte transendothelial migration' pathway between the control group and the BPD group.

contribution of inflammatory responses to the etiology of BPD. By contrast, we found a low expression of genes related to other immune response pathways, including the T cell receptor pathway. Pietrzyk et al. (2013) reported that overexpression of pathways involving cytokines and their receptors confirms the widely accepted role of inflammatory responses in the etiology of BPD, and that T cell response pathways are closely related to infant maturity (Pietrzyk et al., 2013). Therefore, based on the above-mentioned research studies, our research has revealed more specific regulatory molecules to provide new targets for the prediction of BPD and for targeted interventions.

In summary, this study applied WGCNA to a large dataset to explore BPD-related co-expression gene networks. Our results revealed the roles of key co-expression module genes, hub genes, and functional biological pathways were associated with the down-regulation of the T cell receptor signaling pathway, the enrichment of the TNF signaling pathway and leukocyte migration in BPD pathogenesis, thus providing new insights into the development of BPD. However, the exact molecular mechanisms connecting hub genes and functional pathways of BPD need further exploration.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material.**

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Review Board of Sixth Affiliated Hospital of Sun Yat-sen University (2019ZSLYEC-80), and written informed consent was provided by the participants' legal guardians. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SL and HH had full access to all the study data and took responsibility for data integrity and the accuracy of data analysis results. YC, FM, LQ, BL, HX, and YM contributed to the conception, study design, and data collection. All authors participated in the drafting of the article or its critical revision for important intellectual content and approved the publication of the final version of the manuscript.

FUNDING

This study was funded by the National Natural Science Foundation of China (No. 81801206) and the Science and Technology Foundation of Guangzhou, China (Nos. 201704020230, 201704020086, and 201604020154).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.539292/full#supplementary-material

REFERENCES

Bhandari, A., and McGrath, M. S. (2013). Long-term pulmonary outcomes of patients with bronchopulmonary dysplasia. *Semin. Perinatol.* 37, 132–137. doi: 10.1053/j.2013.01.010

Cheng, J., Montecalvo, A., and Kane, L. P. (2011). Regulation of NF-κB induction by TCR/CD28. *Immunol. Res.* 50, 113–117. doi: 10.1007/s12026-011-8216-z

Chess, P. R., D'Angio, C. T., Pryhuber, G. S., and Maniscalco, W. M. (2006). Pathogenesis of bronchopulmonary dysplasia. *Semin. Perinatol.* 30, 171–178. doi: 10.1053/j.semperi.2006.05.003

Coalson, J. J. (2003). Pathology of new bronchopulmonary dysplasia. *Semin. Neonatol.* 8, 73–81. doi: 10.1016/s1084-2756(02)00193-8

Collins, J. J. P., Tibboel, D., de Kleer, I. M., Reiss, I. K. M., and Rottier, R. J. (2017). The future of bronchopulmonary dysplasia: emerging pathophysiological concepts and potential new avenues of treatment. *Front. Med.* 4:61. doi: 10.3389/fmed.2017.00061

Farstad, T., Bratlid, D., Medbø, S., and Markestad, T. (2011). Bronchopulmonary dysplasia- prevalence, severity and predictive factors in a national cohort of extremely premature infants. *Acta Paediatr.* 100, 53–58. doi: 10.1111/j.1651-2227.2010.01959.x

Giulietti, M., Occhipinti, G., Principato, G., and Piva, F. (2016). Weighted gene co-expression network analysis reveals key genes involved in pancreatic ductal adenocarcinoma development. *Cell. Oncol.* 9, 379–388. doi: 10.1007/s13402-016-0283-7

Horvath, S., and Dong, J. (2008). Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.* 4:e1000117. doi: 10.1371/journal.pcbi.1000117

James, J. R., McColl, J., Oliveira, M. I., Dunne, P. D., Huang, E., Jansson, A., et al. (2011). The T cell receptor triggering apparatus is composed of monovalent or monomeric proteins. *J. Biol. Chem.* 286, 31993–32001. doi: 10.1074/jbc.M111.219212

Jobe, A. J. (1999). The new BPD: an arrest of lung development. *Pediatr. Res.* 46, 641–643. doi: 10.1203/00006450-199912000-00007

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

Lemons, J. A., Bauer, C. R., Oh, W., Korones, S. B., Papile, L. A., Stoll, B. J., et al. (2001). Very low birth weight outcomes of the National Institute of Child health and human development neonatal research network, January 1995 through December 1996. NICHD neonatal research network. *Pediatrics* 107:E1. doi: 10.1542/peds.107.1.e1

Liu, D., Liu, Y., Dou, L., Sun, M., Jiang, H., and Yi, M. (2018). Spatial and temporal expression of SP-B and TGF-β1 in hyperoxia-induced neonatal rat lung injury. *Int. J. Clin. Exp. Pathol.* 11, 232–239.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431, 308–312. doi: 10.1038/nature02782

Ma, C., Lv, Q., Teng, S., Yu, Y., Niu, K., and Yi, C. (2017). Identifying key genes in rheumatoid arthritis by weighted gene co-expression network analysis. *Int. J. Rheum. Dis.* 20, 971–979. doi: 10.1111/1756-185X.13063

Medina, R. I., and Lubovac-Pilav, Z. (2016). Gene co-expression network analysis for identifying modules and functionally enriched pathways in type 1 diabetes. *PLoS One* 11:e0156006. doi: 10.1371/journal.pone.0156006

Northway, W. H. Jr., Rosan, R. C., and Porter, D. Y. (1967). Pulmonary disease following respiratory therapy of hyaline-membrane disease. Bronchopulmonary dysplasia. *N. Engl. J. Med.* 276, 357–368. doi: 10.1056/NEJM196702162760701

Oldham, M. C., Konopka, G., Iwamoto, K., Langfelder, P., Kato, T., Horvath, S., et al. (2008). Functional organization of the transcriptome in human brain. *Nat. Neurosci.* 11, 1271–1282. doi: 10.1038/nn.2207

Pietrzyk, J. J., Kwinta, P., Wollen, E. J., Bik-Multanowski, M., Madetko-Talowska, A., Günther, C. C., et al. (2013). Gene expression profiling in preterm infants: new aspects of bronchopulmonary dysplasia development. *PLoS One* 8:e78585. doi: 10.1371/journal.pone.0078585

Presson, A. P., Sobel, E. M., Papp, J. C., Suarez, C. J., Whistler, T., Rajeevan, M. S., et al. (2008). Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC Syst. Biol.* 2:95. doi: 10.1186/1752-0509-2-95

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* 297, 1551–1555. doi: 10.1126/science.1073374

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Rue-Albrecht, K., McGettigan, P. A., Hernández, B., Nalpas, N. C., Magee, D. A., Parnell, A. C., et al. (2016). GOexpress: an R/Bioconductor package for the identification and visualisation of robust gene ontology signatures through supervised learning of gene expression data. *BMC Bioinformatics* 17:126. doi: 10.1186/s12859-016-0971-3

Shahzad, T., Radajewski, S., Chao, C. M., Bellusci, S., and Ehrhardt, H. (2016). Pathogenesis of bronchopulmonary dysplasia: when inflammation meets organ development. *Mol. Cell. Pediatr.* 3:23. doi: 10.1186/s40348-016-0051-9

Speer, C. P. (2006a). Inflammation and bronchopulmonary dysplasia: a continuing story. *Semin. Fetal Neonatal Med.* 11, 354–362. doi: 10.1016/j.siny.2006.03.004

Speer, C. P. (2006b). Pulmonary inflammation and bronchopulmonary dysplasia. *J. Perinatol.* 26, S57–S62. doi: 10.1038/sj.jp.7211476

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J. P. (2007). GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics* 23, 3251–3253.. doi: 10.1093/bioinformatics/btm369

Tian, A., Pu, K., Li, B., Li, M., Liu, X. G., Gao, L. P., et al. (2019). Weighted gene coexpression network analysis reveals hub genes involved in cholangiocarcinoma progression and prognosis. *Hepatol. Res.* 49, 1195–1206. doi: 10.1111/hepr.13386

Vieira, S. E., Bando, S. Y., de Paulis, M., Oliveira, D. B. L., Thomazelli, L. M., Durigon, L. M., et al. (2019). Distinct transcriptional modules in the peripheral blood mononuclear cells response to human respiratory syncytial virus or to human rhinovirus in hospitalized infants with bronchiolitis. *PLoS One* 14:e0213501. doi: 10.1371/journal.pone.0213501

Yang, M., Chen, B. L., Huang, J. B., Meng, Y. N., Duan, X. J., Chen, L., et al. (2017). Angiogenesis-related genes may be a more important factor than matrix metalloproteinases in bronchopulmonary dysplasia development. *Oncotarget* 8, 18670–18679. doi: 10.18632/oncotarget.14722

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:Article17. doi: 10.2202/1544-6115.1128

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

frontiers in Genetics

ORIGINAL RESEARCH
published: 09 September 2020
doi: 10.3389/fgene.2020.01034

Identification and Characterization of lncRNAs Related to the Muscle Growth and Development of Japanese Flounder (*Paralichthys olivaceus*)

Shuxian Wu[1,2], Jingru Zhang[1,2], Binghua Liu[1,2], Yajuan Huang[1,2], Siping Li[1,2], Haishen Wen[1,2], Meizhao Zhang[1,2], Jifang Li[1,2], Yun Li[1,2] and Feng He[1,2]*

[1] The Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao, China, [2] Fisheries College, Ocean University of China, Qingdao, China

OPEN ACCESS

Edited by:
Jiajie Peng,
Northwestern Polytechnical University, China

Reviewed by:
Hao Zhang,
Jilin University, China
Xungang Tan,
Chinese Academy of Sciences, China
Quanqi Zhang,
Ocean University of China, China

****Correspondence:***
Feng He
hefengouc@ouc.edu.cn

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

Received: *22 August 2019*
Accepted: *11 August 2020*
Published: *09 September 2020*

Citation:
Wu S, Zhang J, Liu B, Huang Y, Li S, Wen H, Zhang M, Li J, Li Y and He F (2020) Identification and Characterization of lncRNAs Related to the Muscle Growth and Development of Japanese Flounder (Paralichthys olivaceus). Front. Genet. 11:1034. doi: 10.3389/fgene.2020.01034

Long noncoding RNAs (lncRNAs) play an important role in many life activities, but the expression pattern and function of lncRNAs in Japanese flounder skeletal muscle are unclear. In this study, 751 lncRNAs were selected from skeletal muscle in different development stages of the Japanese flounder [stage A: larval 7 days post hatching (dph); stage B: juvenile about 90 dph; stage C (female) and stage D (male): adult about 24 months] using coding potential analysis methods. In total, 232, 211, 194, 28, 29, and 14 differentially expressed lncRNAs and 9549, 8673, 9181, 1821, 1080, and 557 differentially expressed mRNAs were identified in comparisons of A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D, respectively. We identified the *cis*- and *trans*-regulatory target genes of differentially expressed lncRNAs, and lncRNA–gene interaction networks were constructed using the Cytoscape program. In total, there were 200, 200, 200, 93, 47, and 11 *cis*-regulation relationships, and 29, 19, 24, 38, 8, and 47 *trans*-regulation relationships in the comparisons between A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D, respectively. These results indicate that lncRNA may participate in the development of Japanese flounder skeletal muscle through *cis*- or *trans*-acting mechanisms, thus providing a scientific basis for further study of the biological function of lncRNA in Japanese flounder skeletal muscle. Based on these relationships, functional annotation of the related lncRNAs was performed by gene ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis. Differentially expressed genes associated with muscle development were enriched in multiple pairs of comparisons (e.g., differentially expressed genes LOC109640370, LOC109634180, LOC109643555, rusc1, and LOC109626999 were enriched in the actin-binding term, and differentially expressed genes LOC109640370, was, LOC109644970, LOC109643555, and itga9 were enriched in the regulation of the actin cytoskeleton pathway in the KEGG pathway analysis in the comparison of stages C and D). We predicted lncRNA–mRNA, miRNA–mRNA, and lncRNA–miRNA regulatory relationships and constructed interactive networks using Cytoscape software. Co-expression networks show that most lncRNAs interact with one or two predicted miRNAs involved in muscle growth and development. These results provide a basis for further study of the function of lncRNAs on skeletal muscle in different developmental stages of Japanese flounder.

Keywords: Japanese flounder, skeletal muscle, lncRNA, integrated analysis, transcriptome

INTRODUCTION

LncRNAs are defined as transcripts that are more than 200 nucleotides in length and are not translated into proteins (Perkel, 2013). This length limitation distinguishes long ncRNAs from small noncoding RNAs, such as microRNAs (miRNAs), small interfering RNAs (siRNAs), Piwi interacting RNAs (piRNAs), small nucleolar RNAs (snoRNAs), and other short RNAs (Lina et al., 2013). Note that only one fifth of transcriptions are associated with protein-coding genes in the human genome (Philipp et al., 2007). Large-scale cDNA library sequencing and transcriptome sequencing indicate that tens of thousands of intergenic sites are transcribed to noncoding RNAs in mammals. Approximately 78% of lncRNAs are tissue specific, and only ∼19% of mRNAs are tissue specific (Cabili et al., 2011). At present, more and more lncRNAs have been found in mammals, such as humans (Pieter-Jan et al., 2013; Iyer et al., 2015), mice (Phillip et al., 2013; Lv et al., 2015), and sheep (Bakhtiarizadeh et al., 2016) as well as in plants, such as rice (Zhang et al., 2014). It has been reported that lncRNA plays an important role in the regulation of gene transcription (Goodrich and Kugel, 2006), post-transcriptional regulation (Ming-De et al., 2005), epigenetic regulation (Mercer and Mattick, 2013; Morlando et al., 2014), and aging and disease (Lukiw et al., 1992). Although there is growing evidence that most of them may be functional (Mercer et al., 2009), only a relatively small proportion has been shown to be biologically relevant.

Skeletal muscle is a striated muscle tissue composed of muscle cells with contractile capacity. It is well known that the fetal stage is the main stage of skeletal muscle development, and there is no net increase in the number of muscle fibers after birth (Nishina et al., 2003). Muscle development is a complex process that requires interactions between multiple factors (Buckingham, 2006). At present, studies on skeletal muscle growth and development generally focus on the expression and function of related coding genes (Thomas and Mathias, 2011; Eng et al., 2013). The skeletal muscle fiber phenotype is regulated by various independent signaling pathways, including the mitogen-activated protein kinase (MAPK) pathway (Keren et al., 2006), calcineurin (Naya et al., 2000), calcium/calmodulin-dependent protein kinase IV (Takayuki et al., 2004), and the peroxisome proliferator γ coactivator 1 (PGC-1) (Handschin et al., 2003). Studies show that some miRNAs are also involved in the development of skeletal muscle (Tang et al., 2015; Jebessa et al., 2018). Several recent studies show that lncRNAs also play a crucial role in skeletal muscle development (Zhan et al., 2016; Liang et al., 2017; Zhou et al., 2018). In addition, lncRNAs can interact as a competitive endogenous RNA (ceRNA) with miRNAs involved in the regulation of target gene expression, thereby regulating muscle development (Cesana et al., 2011).

Japanese flounder is a valuable marine fish and an important economic fish for marine aquaculture in Asia. Therefore, it is important to reveal the molecular mechanisms of Japanese flounder skeletal muscle formation and development. Studies show that some coding genes play important roles during the development of Japanese flounder skeletal muscle (Huang et al., 2018; Wu et al., 2018). In addition, some studies focus on the effects of noncoding RNA on skeletal muscle development of Japanese flounder. Some micRNAs (mir-1, mir-133, mir-206) play an important role in muscle development during Japanese flounder metamorphosis (Fu et al., 2011, 2012). However, information on lncRNAs related to skeletal muscle development in the Japanese flounder is still limited.

In this study, we used the Illumina HiSeq 2500 platform to identify lncRNAs and mRNAs involved in skeletal muscle development in Japanese flounder. Our study provides useful information for further study of the function of lncRNA during skeletal muscle development in a fish species, and these results will help study skeletal muscle development from the perspective of noncoding RNAs.

MATERIALS AND METHODS

Ethics Statement

The study was approved by the respective Animal Research and Ethics Committees of Ocean University of China. The field studies did not involve endangered or protected species. The fish were all euthanized by tricaine methanesulfonate (MS-222) prior to experimentation.

Experimental Animal Collection

Japanese flounder were collected from the Donggang District Institute of marine treasures in Rizhao, Shandong province. The fish were transported to the Ocean University of China and temporarily reared in a 500-L white bucket for 24 h. The Japanese flounder were collected at various stages: larval 7 days post hatching (dph) (stage A), juvenile ∼90 dph (stage B), female adult ∼24 months (stage C), and male adult ∼24 months. In our experiment and data analysis, 3 fish were used in all stages except for stage A (here fish were very small in size, so ∼50 individuals were combined and considered to be one sample). All fish were euthanized with MS-222, and tissue samples were collected. In stage A, we used a microscope to cut off redundant tissue and only retain muscle tissue. Samples were immediately frozen in liquid nitrogen and then stored at −80°C until further processing.

Illumina Deep Sequencing and Sequence Analysis

Total RNA for RNA sequencing (RNA-seq) was extracted using TRIzol Reagent (Invitrogen, Carlsbad, CA, United States) according to the manufacturer's protocol. The concentration of RNA was quantified by the nucleic acid analyzer Biodropsis BD-1000 (OSTC, China) and the integrity by agarose gel electrophoresis examination. Ribosomal RNA (rRNA) was removed from the total RNA using the Epicenter Ribo-Zero^{TM} rRNA Removal Kit (Epicenter, Madison, WI, United States) following the manufacturer's instructions. The constructed cDNA library was quality tested on an Agilent Bioanalyzer 2100 system, and then high-throughput sequencing was performed on the Illumina HiSeqTM 2500 platform. The paired-end sequencing raw reads were cleared by removing reads containing adapters, including ploy-N reads and low-quality reads to obtain clean reads. At the same time, the Phred score (Q20), Q30, and GC contents of the clean data were calculated. All the downstream analyses were based on the high-quality clean data. The clean reads were mapped to the Japanese flounder reference genome[1] using the Tophat2 software.

Reconstructing transcripts for clean readings was based on probabilistic models using Cufflinks 2.0.2 software. Based on the characteristics of lncRNA, we used a rigorous three-step screening method to obtain candidate lncRNAs (**Figure 1A**). First, Cuffcompare software was used to screen out transcripts that were perfectly matched or similar to other ncRNAs, mRNAs, etc., while clarifying the location type of the remaining transcripts. We then retained transcripts annotated as "i" (intergenic lncRNA), "u" (intronic lncRNA), "x" (anti-sense lncRNA), and "o" (sense-overlapping lncRNA) by screening for candidate lncRNA transcripts. Second, single-exon transcripts and transcripts < 200 bp long were removed. Finally, we used four analytical tools, including CPC (encoding potential calculator) (Lei et al., 2007), CNCI (coding-non-coding-index) (Liang et al., 2013), Pfam Scan (Finn et al., 2014), and PLEK (Li et al., 2014) to predict the coding potential of the transcripts. CPC score $\leq$ 0, CNCI score $\leq$ 0, Pfam: *E*-value $\leq$ 0.001, and coding_potential_score $\leq$ 0 were conditions for screening lncRNA. The transcript expression levels were calculated using the fragments per kb per million (FPKM) reads method, which is the number of fragments per kilobase length from a gene per million fragments. The transcript differential expression was calculated according to the negative binomial distribution test in the DESeq (Anders and Huber, 2012) software[2]. Transcripts with $p < 0.05$ and | (fold change) | $\geq$ 2 were designated as differentially expressed. The sequencing data obtained from RNA-seq were released to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under the accession numbers SRR12102079, SRR12102078, SRR12102077, SRR12102076, SRR12102075, SRR12102074, SRR12102073, SRR12102072, SRR12102071, SRR12102070, SRR12102069, and SRR12102068.

Cis- and *Trans*-Analyses and Enrichment Analysis

We searched for all of the coding genes 100 kb upstream and downstream of differentially expressed lncRNA that had significant co-expression (Pearson correlation calculation) with the lncRNA. These genes that are genomically adjacent and coexpressed in the expression pattern are likely to be the *cis*-target genes of the lncRNA. Based on the results of differential co-expression, lncRNA and mRNA not on the same chromosome were selected as candidate targets. The RNA interaction software RIsearch-2.0 was used to predict the binding of candidate lncRNA and mRNA at the nucleic acid level. The number of bases in which two nucleic acid molecules directly interact with each other is not less than 10 and the free energy of base binding is not more than −50 were used as screening conditions, and they determined the potential that the lncRNA was *trans*-acting. Differentially expressed lncRNAs and their corresponding differentially expressed *cis*- and *trans*-target genes were used to construct lncRNA–gene interaction networks using the Cytoscape program. Predicting the main functions of lncRNA was done by functional enrichment analysis of lncRNA target mRNA genes. We performed gene ontology (GO) enrichment analysis (Young et al., 2010). The number of differential transcripts included in each GO entry was counted, and the significance of differential transcript enrichment in each GO entry was calculated using the hypergeometric distribution test method. The Kyoto Encyclopedia of Genes and Genomes (KEGG) (Minoru et al., 2008) is the main public database of Pathway and is used to perform Pathway analysis on differential transcripts (combined with KEGG annotation results). The significance of differential transcript enrichment in each Pathway entry is calculated by a hypergeometric distribution test.

Quantitative Real-Time-PCR (qRT-PCR) Analysis

Four differentially expressed lncRNAs target genes that affect muscle development were selected for qRT-PCR verification, including calcium voltage-gated channel subunit alpha1 D (cacna1d) (Krasnyi and Ozernyuk, 2011), actin-related protein 3B-like (Tseng et al., 2002), kin of IRRE like 2 (kirrel2) (Durcan et al., 2014), myosin-7B-like (Girgenrath et al., 2005), and their corresponding lncRNA regulatory factors (TCONS_00034769, TCONS_00041871, TCONS_00089031, and TCONS_00038291). The qRT-PCR primers for these lncRNAs and genes are shown in **Table 1**. Quantitative real-time PCR was conducted with the Roche LightCycler480 (Germany) and SYBR Premix Ex Taq^{TM} (TliRNaseH Plus) Kit (Takara, Japan) to determinate the relative expressions of each gene. Primers are listed in **Table 1**, and the Japanese flounder 18 s (GenBank accession no. EF126037.1) was used as the endogenous reference gene. The amplified system was formed from 10μL SYBR®Premix Ex Taq (TliRNaseH Plus), 0.4μL ROX reference dye, 0.4μL PCR forward primer, 0.4μL PCR reverse primer, 2μL cDNA template, and 20 μL of RNase-free water. The reaction was completed according to the following procedure: 95°C for 30 s, 40 cycles of 95°C for 5 s, and Tm for 30 s. All samples were run in triplicate. Then, we calculated

[1] https://www.ncbi.nlm.nih.gov/genome/?term=Japanese+flounder

[2] http://bioconductor.org/packages/release/bioc/html/DESeq.html

TABLE 1 | Primers used for real time PCR.

LncRNA and gene	Primer sequence (5′-3′)	Product length (bp)	Annealing temperature (°C)
TCONS_ 00041871	F: CTCCTGAACCCTTTTCTCCT R:GCTCAGTCTGACTTTAGTGCC	153	60
TCONS_ 00034769	F: ACTGCTCTGGCCTGAGGATG R: CGGCTCTATTGTGGGGAACC	198	65
TCONS_ 00089031	F: CTCACTGTGGGTTTTCAAGC R: TTTGAGCCAGAACAGAGGGT	173	65
TCONS_ 00038291	F:GACGCAGAGGAAAGAAGCAC R: GGAGCAACTTCCTCAGACCT	176	65
Actin-related protein 3B-like	F:TGAGTGGAGGACGGATAAAG R: TCGGACCAATCTCATCGTAG	159	60
Cacna1d	F: ACGCTACTCTGTTTGCTCTG R:AACTTCCCCACTGTTACCTC	178	60
Kirrel2	F: CGTGGTGCTCAGTAATGGTA R: CGTCTGCTGTGATGATAGGT	155	60
Myosin-7B-like	F:AGATTGAGGGGATAGAGTGG R: CCCAGATGGTTGTCATAGAG	169	60
18S	F:ATTGACGGAAGGGCACCAC R:ATGCACCACCACCCACAGA	134	65

the relative expression by the method of comparative threshold $(2-\Delta\Delta Ct)$ (Livak and Schmittgen, 2001).

LncRNA–miRNA–mRNA Network Construction

Cytoscape is an open source software for biological network integration, visualization, and analysis, loading molecular and genetic interaction data sets in many standard formats in the fields of molecular and systems biology, genomics, and proteomics (Saito et al., 2012). In this study, lncrRNA–mRNA, miRNA–mRNA, and lncRNA–miRNA relationship pairs with regulatory relationships were predicted, and Cytoscape software was used to construct the interaction network diagram among them. The nodes in the network diagram are the mRNAs related to muscle development in the GO enrichment and KEGG pathways; the lncRNAs that have a regulatory relationship with mRNAs; the miRNAs that have a regulatory relationship with mRNAs. The sequencing data obtained from miRNA-seq were released to the NCBI SRA database under the accession numbers SRR11968806, SRR11968805, SRR11968804, SRR11968803, SRR11968802, SRR11968801, SRR11968800, SRR11968799, SRR11968798, SRR11968797, SRR11968796, and SRR11968795.

Statistical Analysis

The data were presented as means ± SEM. The statistical differences were analyzed by one-way ANOVA and Duncan's multiple range tests in SPSS 19.0 software. $P < 0.05$ was considered to be statistically significant.

RESULTS

Overview of RNA-Sequencing

High-throughput RNA-seq was performed on the Illumina Hiseq 2500 platform. Each library produced more than 90 million raw reads. After filtering low-quality reads, clean reads still accounted for more than 93% of the raw reads. More than 67.32% of the clean reads perfectly mapped to the reference genome of the Japanese flounder. The uniquely mapped reads ranged from 62.16 to 75.55% of the clean reads (**Table 2**).

Identification of lncRNAs in Japanese Flounder Skeletal Muscle

According to the characteristics of lncRNAs, RNA-seq produced 751 lncRNA transcripts after strict screening and filtering of RNAs that did not meet the requirements (**Figure 1B**). The length of lncRNAs ranged from 201 to 9381 bp; the length of lncRNAs between 201 and 1000 bp was 53.3%, 1000–2000 bp was 30.5%, and the average length of lncRNAs was 1243 bp (**Figure 1C**). The lncRNAs with 2 exons were 68% and with 3 exons were 18% (**Figure 1D**). The number of predicted lncRNAs types was 436 for intergenic lncRNA (u), 146 for intronic lncRNA (i), 85 for anti-sense lncRNA (x), and 84 for sense-overlapping lncRNA (o) (**Figure 1E**).

Differential Expression Analysis of lncRNAs and mRNAs

In order to display the information on differentially expressed lncRNAs and mRNAs more intuitively, the differential expression of lncRNAs and mRNAs in the same differential comparison group was shown by using Circos software (**Figure 2** and **Supplementary Figure S1**). In the A versus B comparison, 232 differentially expressed lncRNAs were detected, 67 of which were upregulated and 165 were downregulated. A total of 9549 differentially expressed mRNAs were detected, 3041 of which were upregulated and 6508 were downregulated (**Figure 2** and **Supplementary Table S1**). In the A versus C comparison, 211 differentially expressed lncRNAs were detected, 60 of which were upregulated and 151 were downregulated. A total of 8673 differentially expressed mRNAs were detected, 2665 of which were upregulated and 6008 were downregulated (**Supplementary Figure S1** and **Supplementary Table S2**). In the A versus D comparison, 194 differentially expressed lncRNAs were detected, 63 of which were upregulated and 131 were downregulated, and 9181 differentially expressed mRNAs were detected, 2854 of which were upregulated and 6327 were downregulated (**Supplementary Figure S1** and **Supplementary Table S3**). In the B versus C comparison, 28 differentially expressed lncRNAs were detected, 13 of which were upregulated and 15 were downregulated, and 1821 differentially expressed mRNAs were detected, 978 of which were upregulated and 843 were downregulated (**Supplementary Figure S1** and **Supplementary Table S4**). In the B versus D comparison, 29 differentially expressed lncRNAs were detected, 12 of which were upregulated and 17 were downregulated, and 1080 differentially expressed mRNAs were detected, 532 of which were upregulated and 548 were downregulated (**Supplementary Figure S1** and **Supplementary Table S5**). In the C versus D comparison, 14 differentially expressed lncRNAs were detected, 7 of which were upregulated and 7 were downregulated, and 557 differentially expressed mRNAs were detected, 245 of which were upregulated

TABLE 2 | Summary of draft reads of 12 libraries by RNA-sequencing.

Sample	Raw reads	Clean reads	Total mapped	Multiple mapped	Uniquely mapped
A_1_1	95,070,396	90,140,748 (94.81%)	68,921,334 (76.46%)	2,510,198 (2.78%)	66,411,136 (73.67%)
A_1_2	95,386,604	90,562,700 (94.94%)	71,317,689 (78.75%)	2,896,374 (3.20%)	68,421,315 (75.55%)
A_1_3	98,029,082	92,699,826 (94.56%)	72,020,838 (77.69%)	2,679,887 (2.89%)	69,340,951 (74.80%)
B_2_1	96,622,226	91,173,512 (94.36%)	67,805,686 (74.37%)	6,443,523 (7.07%)	61,362,163 (67.30%)
B_2_2	96,239,174	91,922,114 (95.51%)	66,710,930 (72.57%)	5,541,344 (6.03%)	61,169,586 (66.55%)
B_2_3	96,343,382	90,217,708 (93.64%)	60,737,386 (67.32%)	4,662,151 (5.17%)	56,075,235 (62.16%)
C_3_1	95,721,552	91,153,382 (95.23%)	71,011,936 (77.90%)	10,055,963 (11.03%)	60,955,973 (66.87%)
C_3_2	95,830,852	91,285,636 (95.26%)	72,297,477 (79.20%)	1,278,4703 (14.01%)	59,512,774 (65.19%)
C_3_3	97,999,946	93,132,198 (95.03%)	73,007,033 (78.39%)	12,142,118 (13.04%)	60,864,915 (65.35%)
D_4_1	98,391,502	92,247,994 (93.76%)	65,722,331 (71.25%)	8,110,885 (8.79%)	57,611,446 (62.45%)
D_4_2	98,380,634	93,523,450 (95.06%)	73,108,525 (78.17%)	10,076,310 (10.77%)	63,032,215 (67.40%)
D_4_3	96,840,342	92,390,402 (95.40%)	73,089,193 (79.11%)	9,858,949 (10.67%)	63,230,244 (68.44%)

and 312 were downregulated (**Supplementary Figure S1** and **Supplementary Table S6**).

lncRNA–Gene Interaction Network Construction

To address how lncRNA interacts with its target gene (mRNA) to regulate Japanese flounder muscle development and identify key molecular players in the process, we predicted *cis*- and *trans*-targets of differentially expressed lncRNAs and constructed the possible regulatory networks for these interactions. Previous studies have demonstrated that lncRNAs regulate the expression of adjacent protein-coding genes via a *cis*-acting mechanism (Han et al., 2012; Qian et al., 2012). In the present study, we screened for all the coding genes in the 100k that were upstream and downstream of the differentially expressed lncRNAs and significantly coexpressed with the lncRNAs (Pearson correlation calculation, **Supplementary Tables S14–S19**). These genes that are genomically adjacent and coexpressed in expression patterns are predicted to be *cis*-target genes of lncRNAs. In addition, lncRNAs regulate the expression of genes located on other chromosomes through a *trans*-acting mechanism (Han et al., 2012). Based on the results of differential co-expression, the lncRNAs and mRNAs that are not on the same chromosome were selected as candidate targets. The RNA interaction software RIsearch-2.0 was used to predict the binding of candidates of lncRNA and mRNA at the nucleic acid level. The lncRNAs and mRNAs that may have direct regulation were screened, and these genes were predicted to be *trans*-target genes. For the comparison of A and B, the lncRNA-gene interaction network contained 304 network nodes, 83 lncRNAs, 221 protein-coding genes, 200 pairs of *cis*-regulation relations, and 29 pairs of *trans*-regulation relations (**Figure 3** and **Supplementary Table S7**). For the comparison of A and C, the lncRNA–gene interaction network contained 285 network nodes, 70 lncRNAs, 215 protein encode genes, 200 pairs of *cis*-regulation relations, and 19 pairs of *trans*-regulation relations (**Supplementary Figure S2** and **Supplementary Table S7**). For the comparison of A and D, the lncRNA–gene interaction network consists of 278 network nodes, 69 lncRNAs, 209 protein-coding genes, 200 pairs of *cis*-regulation relations, and 24 pairs of *trans*-regulation relations (**Supplementary Figure S2** and **Supplementary Table S7**). For the comparison of B and C, the lncRNA–gene interaction network contained 157 network nodes, 26 lncRNAs, 131 protein-coding genes, 93 pairs of *cis*-regulation relations, and 38 pairs of *trans*-regulation relations (**Supplementary Figure S2** and **Supplementary Table S7**). For the comparison of B and D, the lncRNA–gene interaction network contained 108 network nodes, 22 lncRNAs, 86 protein encode genes, 47 pairs of *cis*-regulation relations, and 8 pairs of *trans*-regulation relations (**Supplementary Figure S2** and **Supplementary Table S7**). For the comparison of C and D, the lncRNA–gene interaction network contained 97 network nodes, 12 lncRNAs, 85 protein-coding genes, 11 pairs of *cis*-regulation relations, and 47 pairs of *trans*-regulation relations (**Supplementary Figure S2** and **Supplementary Table S7**). We then analyzed the expression correlation between the network lncRNA and its corresponding target gene. In the network constructed from the differentially expressed lncRNAs and target genes identified from the A versus B comparison, 212 lncRNA–gene linkages were positively correlated, and the other 18 linkages were negatively correlated (**Figure 3** and **Supplementary Table S7**). For the A versus C comparison, 210 lncRNA–gene connections were positively correlated, and 9 connections were negatively correlated (**Supplementary Figure S2** and **Supplementary Table S7**). For the comparison of A and D, 213 lncRNA–gene connections were positively correlated, and 11 connections were negatively correlated (**Supplementary Figure S2** and **Supplementary Table S7**). For the B versus C comparison, 73 lncRNA–gene connections were positively correlated, and 58 connections were negatively correlated (**Supplementary Figure S2** and **Supplementary Table S7**). For the B versus D comparison, 46 lncRNA–gene connections were positively correlated, and 9 connections were negatively correlated (**Supplementary Figure S2** and **Supplementary Table S7**). For the C versus D comparison, 24 lncRNA-gene connections were positively correlated, and 34 connections were negatively correlated (**Supplementary Figure S2** and **Supplementary Table S7**). The directional analysis shows that the positive correlation number between lncRNA–gene pairs

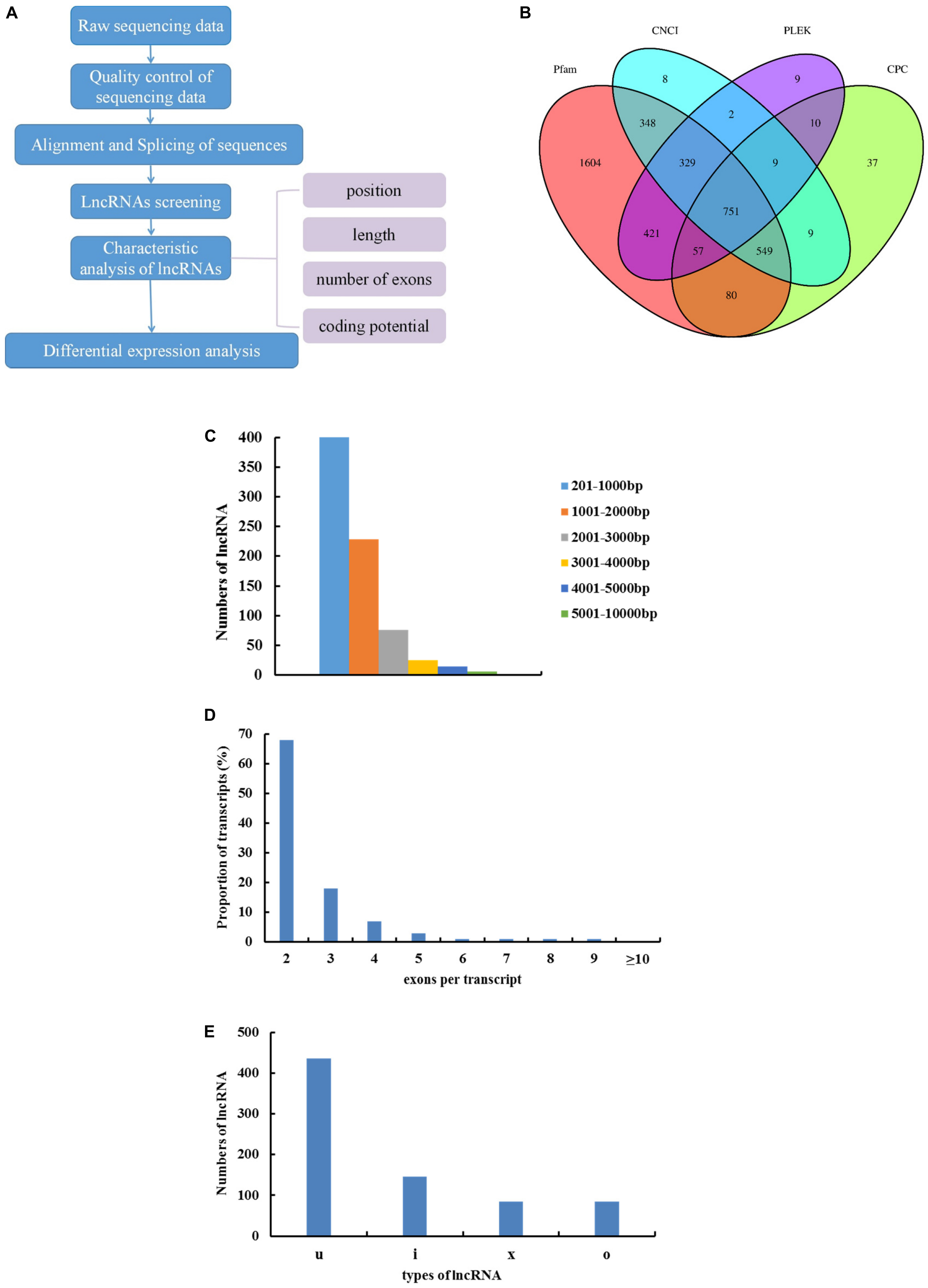


FIGURE 1 | The features of Japanese flounder lncRNAs. **(A)** Identification and verification of lncRNA in skeletal muscle of Japanese Flounder. **(B)** Venn diagram of Candidate lncRNA coding ability prediction result. **(C)** Exon length distribution of Japanese flounder lncRNAs. **(D)** Exon numbers per transcript of Japanese flounder lncRNAs. **(E)** The lncRNAs types of Japanese flounder skeletal muscle, “I” (intergenic lncRNA), “u” (intronic lncRNA), “x” (anti-sense lncRNA), “o” (sense-overlapping lncRNA).

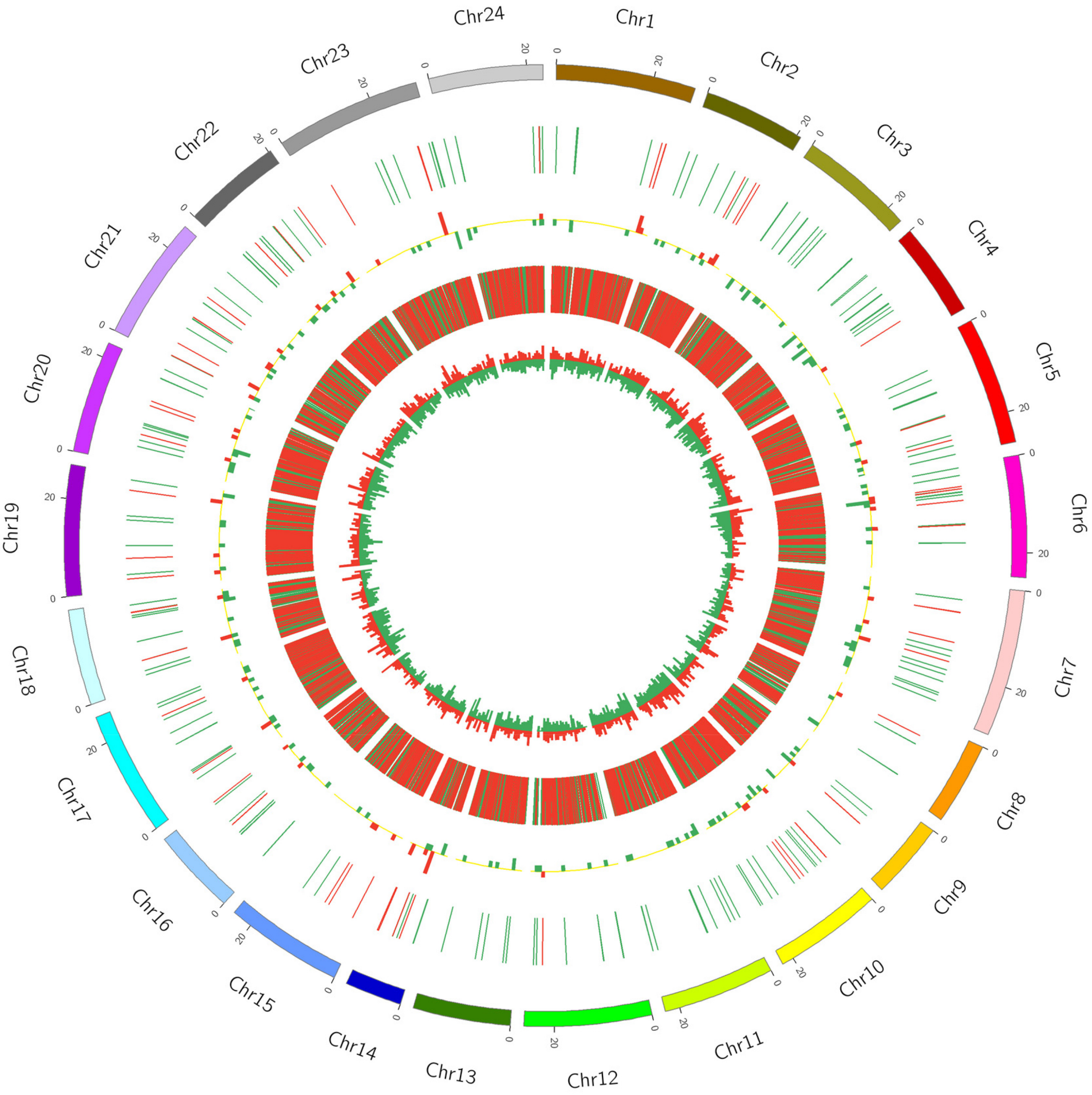


FIGURE 2 | Circos diagram of differential expression of lncRNA and mRNA in A vs. B. In the figure, the outermost circle is the autosomal distribution of the Japanese flounder; the second circle is the lncRNA of differential expression on the chromosome, the red line indicates up-regulation, the green line indicates down-regulation; the third circle is the histogram of differentially expressed lncRNAs at different positions. Red indicates up-regulation, green indicates down-regulation, and the higher the column, indicates the more differentially expressed gene numbers. The fourth circle is the distribution of differentially expressed mRNAs on the chromosome, and the color distribution is the same as lncRNA; the innermost circle is the column with differentially expressed mRNAs at different positions, color distribution is the same as lncRNA.

was higher than the negative correlation number except for comparison networks of C and D.

GO and KEGG Pathway Analysis

We enriched the biological processes and pathways in all comparisons. In the A versus B comparison, 3805 terms were enriched, and 3720 terms were enriched in the A versus C comparison, 3627 terms in A versus D, 1779 terms in B versus C, 1780 terms in B versus D, and 741 terms were enriched in the C versus D comparison ($p < 0.05$). We selected the top 30 terms in the GO enrichment analysis for each comparison (screening GO entries with corresponding transcript numbers greater than 2, sorting from large to small according to the corresponding −log10P-value for each entry

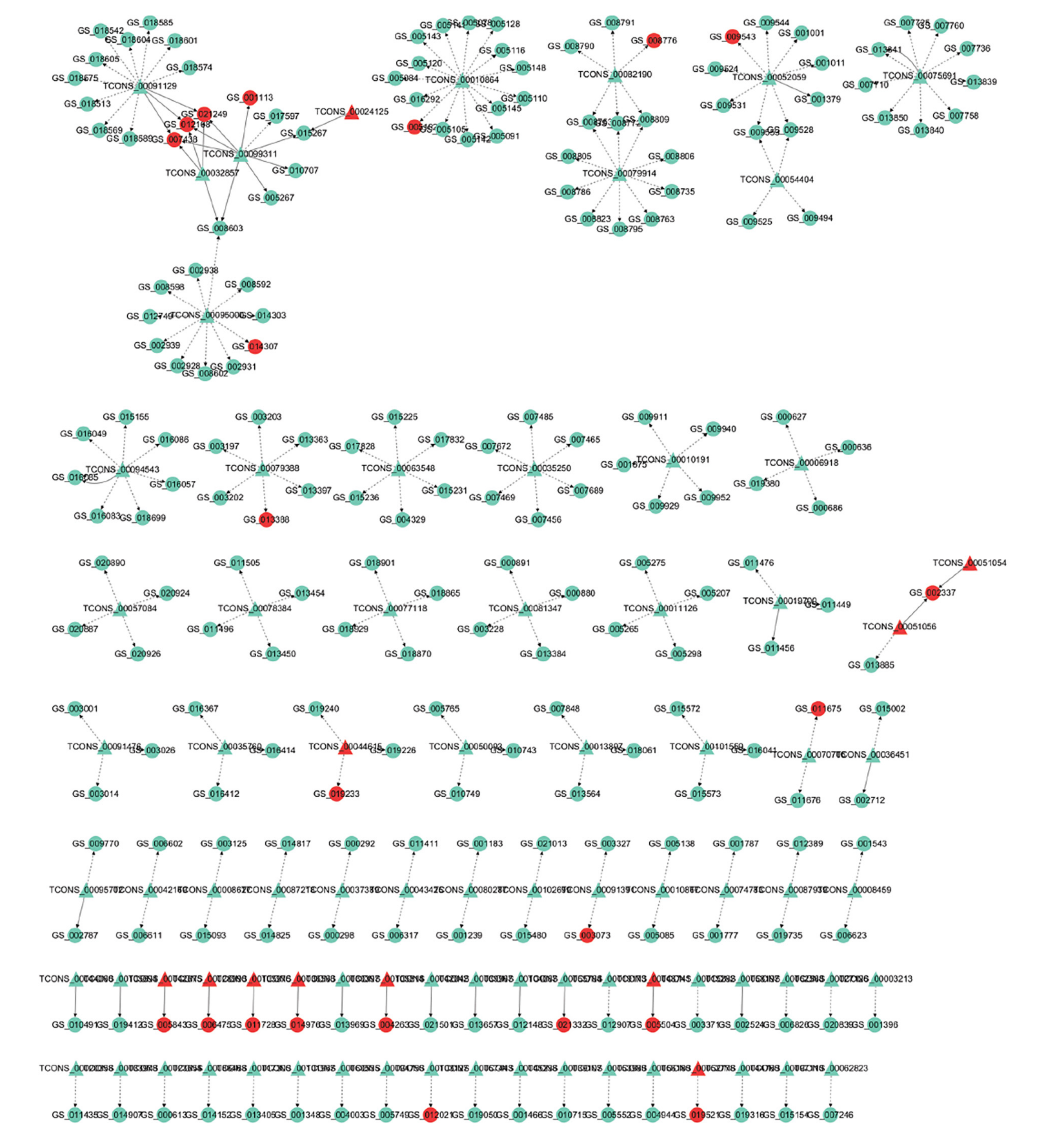


FIGURE 3 | LncRNA – gene interaction network diagram in A vs. B. Red indicates up-regulation, green indicates down-regulation, triangles represent lncRNA, and circles indicate mRNA. The dashed line indicates the interaction between the differentially expressed lncRNA and its corresponding cis target gene, while the solid line indicates the interaction between the differentially expressed lncRNA and its corresponding trans target gene.

and then selecting 10 terms in each of the three categories) analysis (**Supplementary Tables S8–S13**). Many GO terms that were enriched in more than one comparison were related to myosin filament, epidermal cell differentiation, and calcium- and calmodulin-responsive adenylate cyclase activity. In all of these GO terms, there were enriched muscle-related terms, such as muscle myosin complex, myosin filament, and actin binding. In all of these GO terms, muscle-related terms, such as muscle myosin complex, myosin filament, and actin binding, were enriched. **Figure 4A** is a GO enrichment map of the C versus D comparison in which actin binding is significantly enriched in the top 30. For A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D comparisons, 79, 85, 72, 31, 31, and 31 enrichment pathways were detected, respectively ($p < 0.05$). We selected the top 20 terms in the KEGG pathway analysis for each comparison (screening the pathway entries with

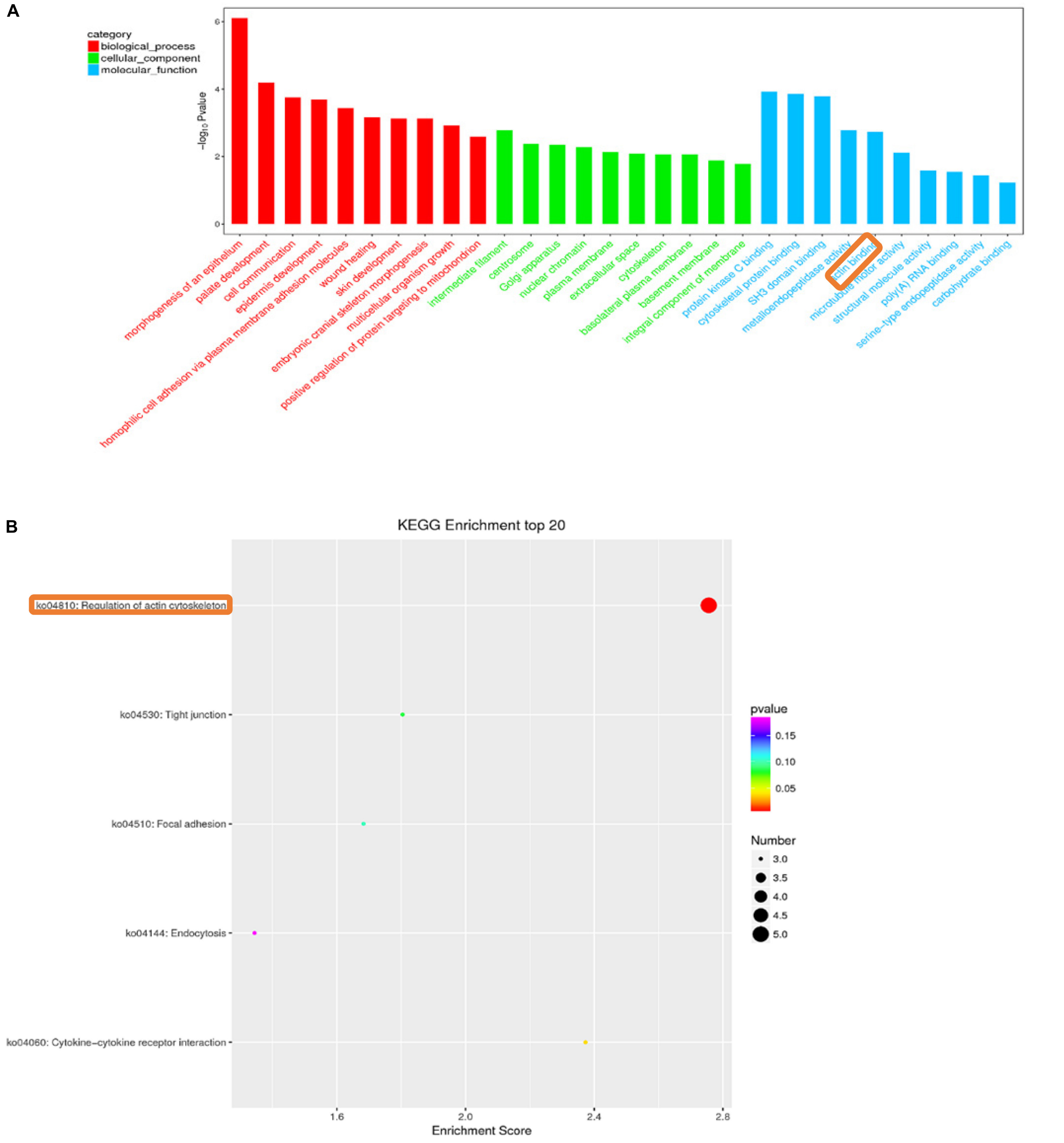


FIGURE 4 | Histogram of gene ontology (GO) classification **(A)**. GO analysis of C vs. D differentially expressed lncRNAs target genes. The horizontal axis indicates the GO entry name and the vertical axis indicates −log10Pvalue. Red bars: biological process; Green bars: cellular component; Blue bars: molecular function. KEGG pathway enrichment of C vs. D differentially expressed lncRNAs target genes **(B)**. The horizontal axis is the enrichment score, the vertical axis indicates the name of the pathway.

transcript numbers greater than 2 and sorting from large to small according to the corresponding −log10P-value for each entry) (**Supplementary Tables S8–S13**). The regulation of the actin cytoskeleton pathway was significantly enriched in the C versus D comparison (**Figure 4B**). These results suggest that some lncRNAs may be involved in the growth and development of skeletal muscle.

Verification of Gene Expression Profiles Using qRT-PCR

To confirm the accuracy and reproducibility of differentially expressed lncRNAs and differentially expressed gene expression levels obtained from RNA-seq, we selected four differentially expressed lncRNA target genes that affect muscle development and performed qRT-PCR verification. The expression of cacna1d was downregulated in stage C compared with that in stage A. The expression of TCONS_00034769 was downregulated in stage C compared with that in stage A. The expression of actin-related protein 3B-like was upregulated in stage B compared with that in stage A. The expression of TCONS_00041871 was downregulated in stage B compared with that in stage A. The expression of kirrel2 was downregulated in stage B compared with that in stage A. The expression of TCONS_00089031 was downregulated in stage B compared with that in stage A. The expression of myosin-7B-like was downregulated in stage D compared with that in stage A. The expression of TCONS_00038291 was upregulated in stage D compared with that in stage A. All four lncRNAs and their target genes showed similar expression patterns compared to RNA-seq data, indicating the reliability of our RNA-seq data (**Figure 5**).

Bioinformatics Analysis of lncRNA–miRNA–mRNA Networks

Recent studies show that lncRNAs can act as a competitive endogenous RNA affecting post-transcriptional regulation by interfering with the miRNA pathway. To further elucidate the role of lncRNAs in the growth and development of skeletal muscle in Japanese flounder, we used TargetScan and miRanda to predict miRNAs that have a regulatory relationship with the given muscle development–associated mRNAs and the differentially expressed lncRNAs and established the basic lncRNA linkage (**Figure 6**). The lncRNA–miRNA–mRNA interaction network (**Figure 7**, **Supplementary Figure S2,S3** and **Supplementary Table S7**) was constructed using Cytoscape software. In the A versus B comparison, there were 18 miRNAs, 13 mRNAs, and 101 lncRNAs, which had at least one predicted target miRNA. There were 18 miRNAs, 13 mRNAs, and 110 lncRNAs in the A versus C comparison. There were 18 miRNAs, 13 mRNAs, and 95 lncRNAs in the A versus D comparison. There were 19 miRNAs, 13 mRNAs, and 11 lncRNAs in the B versus C comparison. There were 19 miRNAs, 13 mRNAs, and 12 lncRNAs in the B versus D comparison. There were 19 miRNAs, 13 mRNAs, and 4 lncRNAs in the C versus D comparison. Some of these lncRNAs (including TCONS_00003213, TCONS_00006684, and TCONS_00023918) were found to interact with at least three target miRNAs.

DISCUSSION

More and more lncRNAs have been discovered in different tissues and cells by high-throughput sequencing technology, some of which have been proven to play important roles in the growth or disease development of some mammals (Geng et al., 2013; White et al., 2014; Gao et al., 2017) and other model organisms (Pauli et al., 2012; Grote et al., 2013). In this study, only 67.32% of the B_2_3 sample in the RNA-seq results was mapped to the Japanese flounder genome, and the mapping rate was a bit low. The main reason for the low reference genome-mapping rate is that the genome assembly of the reference species is not ideal. NCBI has two reference genomes of Japanese flounder, and the assembly quality is not high. ContigN50 is 30K and 36K, respectively. In

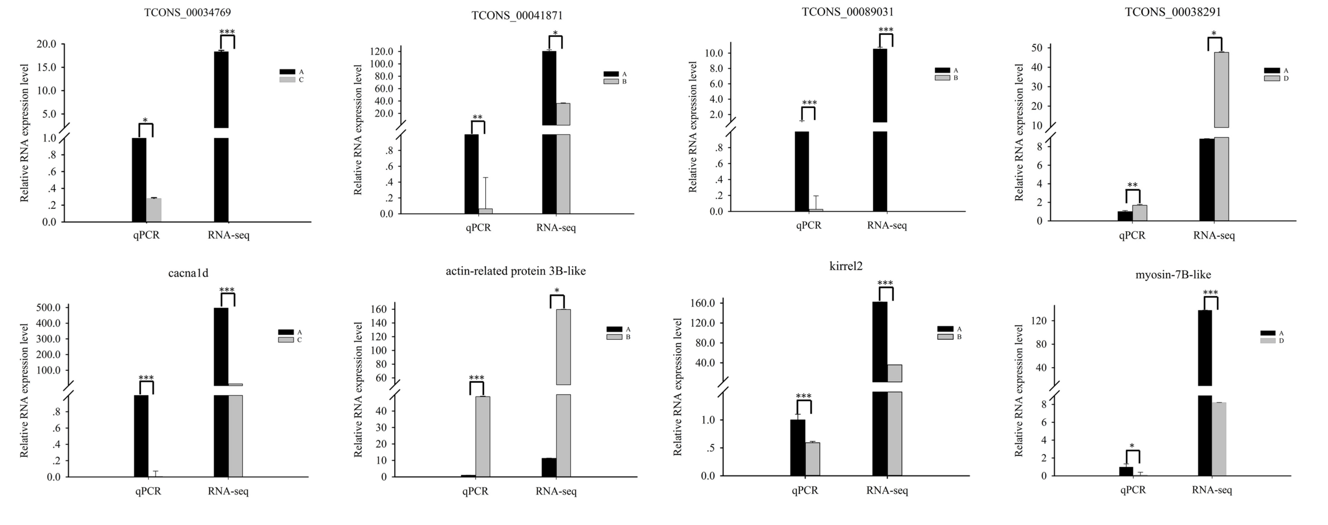


FIGURE 5 | Quantitative real-time PCR validation of differentially expressed lncRNAs and their corresponding target genes. Values are shown as mean with SEM. Significant differences between stages are indicated by *. *, **, and *** indicate $p < 0.01$, $p < 0.001$, and $p < 0.0001$, respectively.

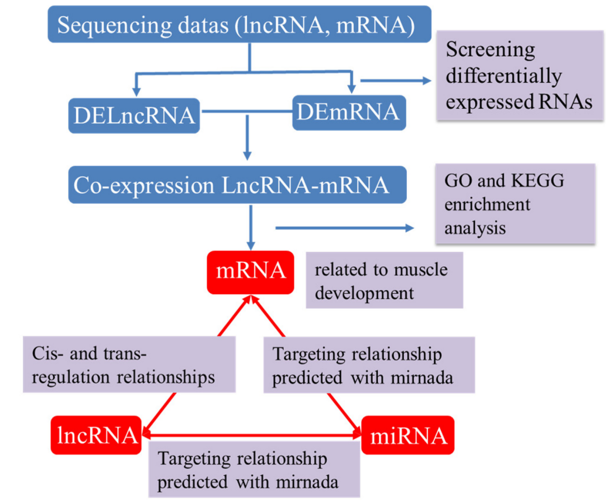

Sequencing datas (lncRNA, mRNA)
Screening differentially expressed RNAs
DELncRNA
DEmRNA
Co-expression LncRNA-mRNA
GO and KEGG enrichment analysis
mRNA
related to muscle development
Cis- and trans-regulation relationships
Targeting relationship predicted with mirnada
lncRNA
miRNA
Targeting relationship predicted with mirnada

FIGURE 6 | The process of building lncRNA – miRNA-mRNA interactive network. DElncRNA: differentially expressed lncRNA; DEmRNA: differentially expressed mRNA.

general, the longer the contigN50, the better the assembly result. Previous reports have shown that the mapping rates were only 74 and 85% (Zhang et al., 2016; Sun et al., 2020), which are relatively low. There are still few analyses of Japanese flounder with the reference genome, most of which are Denovo strategies, which also reflects that the Japanese flounder genome is not very good.

The identification and characterization of lncRNAs in Japanese flounder, especially in skeletal muscle development, are very limited compared to those of lncRNAs in mammals. In this study, we identified 751 lncRNAs in four stages (A, B, C, D) of Japanese flounder skeletal muscle development through high-throughput sequencing. We also identified differentially expressed mRNAs and lncRNAs among the A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D comparisons, respectively, 9549, 8673, 9181, 1821, 1080, and 557 differentially expressed mRNAs and 232, 211, 194, 28, 29, and 14 differentially expressed lncRNAs. These may have specific biological functions in skeletal muscle development of Japanese flounder. In recent years, the biological effects of some lncRNAs in muscle have been reported. For example, it is reported that lncRNA AK 017368 can promote the proliferation and inhibit the differentiation of skeletal muscle myoblasts (Liang et al., 2017). Linc-MD1 is the first identified lncRNA specifically involved in muscle differentiation (Cesana et al., 2011). In addition, the newly identified lncRNA MAR promotes muscle differentiation and regeneration and may also be a novel therapeutic target for the treatment of aging or muscle atrophy (Zhang et al., 2018). Therefore, differentially expressed lncRNAs identified here may also affect the development of skeletal muscle in Japanese flounder.

It is well known that lncRNAs can function by targeting protein-encoding genes. In this study, we hypothesized their potential biological function by predicting the *cis*- and *trans*-regulated target genes of lncRNAs. Recent studies have also shown that lncRNAs are involved in *cis*-regulatory activity in muscle development. Studies have shown that lncRNA-Six1, located 432 bp upstream of the protein-encoding gene Six homeobox 1 (Six1), promotes cell proliferation and participates in muscle growth through *cis*-acting regulation of genes (Cai et al., 2017). Previous studies show that the lncRNA Dum located upstream of the developmental pluripotency-associated 2 (Dppa2) gene is involved in myogenic differentiation and muscle regeneration (Wang et al., 2015). Here, we screened all the coding genes 100k upstream and downstream of the differentially expressed lncRNAs as *cis*-targets. In the comparisons of A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D, 200, 200, 200, 93, 47, and 11 pairs of lncRNAs and genes were found to have *cis*-regulation relations, respectively. The identified *cis*-regulatory genes have important significance for predicting the function of lncRNAs. Thus, the regulatory role of lncRNAs on muscle development needs to be investigated further. Many studies have shown that lncRNA participates in muscle development through *trans*-acting. The synergistic activity of the two DNA enhancer elements CE and DRR located in the ∼24 kb upstream region of MYOD1 regulates the level of MyoD expression in the myogenic lineage, whereas DDRRNA promotes the expression of myogenin by *trans*-acting (Mousavi et al., 2013). MUNC, also known as DRR (eRNA), is located 5 kb upstream of the transcription start site of MyoD and promotes the function of MyoD during skeletal muscle development (Mueller et al., 2015). Based on the results of differential co-expression, the lncRNAs and mRNAs that are not on the same chromosome were selected as candidate targets. In the comparisons of A versus B, A versus C, A versus D, B versus C, B versus D, and C versus D, 29, 19, 24, 38, 8, and 47 pairs of LncRNAs and genes were found to have *trans*-regulation relations. These results suggest that lncRNA may participate in the development of Japanese flounder skeletal muscle through *cis*- or *trans*-acting mechanisms.

To identify the potential function of lncRNAs, in this study, we performed GO and KEGG enrichment analysis on the predicted *cis*- and *trans*-target genes of differentially expressed lncRNAs to understand the function of these differentially expressed lncRNA target genes and further predict the biological function of lncRNAs. In these enrichment analysis results, we find that the *cis*-acting regulatory gene LOC109635692 is enriched for the term "myosin VII complex, actin filament binding." Furthermore, LOC109635692 is also found to be the *cis*-target gene predicted by lncRNA TCONS_00056200 and TCONS_00056180. The gene shows a downward trend in the comparison of A and B. During fish growth, increasing muscle fiber number or muscle size is the major muscle growth process. This gene is related to one of the factors that affect muscle development. A negative correlation is observed between LOC109635692 and TCONS_00056200, and a positive correlation is observed between LOC109635692 and TCONS_00056180, indicating that different lncRNAs may regulate the same target gene in different ways to regulate muscle development. Some coding genes regulated by lncRNA *trans*-action are also enriched in terms and pathways related to muscle development. For example, LOC109626851 identified in the comparison of B and C is enriched in the myosin filament

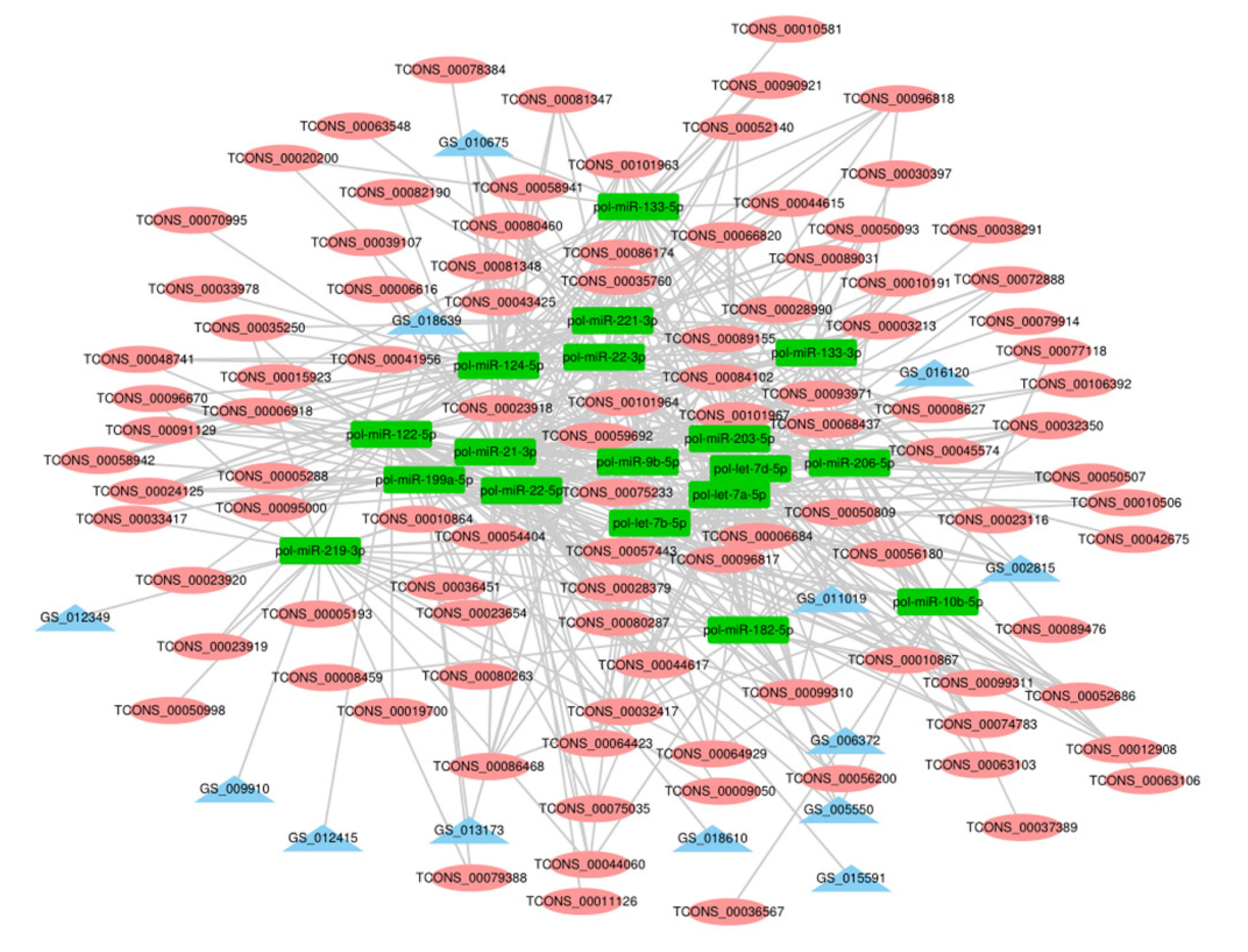


FIGURE 7 | The lncRNA – miRNA-mRNA interaction network in A vs. B. In the network, blue triangles represent mRNAs, red ovals represent lncRNAs, and green rectangles represent miRNAs.

term, LOC109648097 identified in the comparison of B and D is also enriched in the term of myosin filament, and the genes identified in the comparison of C and D, such as LOC109626999 and LOC109643555, are enriched in terms, such as actin binding. These genes are all *cis*- or *trans*-regulated with lncRNA. These results indicate that lncRNA may participate in the development of skeletal muscle in Japanese flounder.

Previous studies show that cacna1d is an important regulator of muscle development (Krasnyi and Ozernyuk, 2011, Park et al., 2018). In the lncRNA–gene network comparing A and C, the cacna1d gene -s the predicted *cis*-target of TCONS_00034769 (**Figure 3B**). In addition, it is enriched in the term of "skeletal muscle fiber development" in GO enrichment analysis. Moreover, the expression levels of cacna1d and TCONS_00034769 in stage A are higher than those in stage C. These results indicate that TCONS_00034769 targets cacna1d through a *cis*-regulatory mechanism to regulate skeletal muscle development in Japanese flounder. In the lncRNA–gene network comparing A and B, the actin-related protein 3B-like gene is the predicted *cis*-target of TCONS_00041871 (**Figure 3A**). In addition, it is enriched in the terms of "regulation of myosin II filament organization, positive regulation of actin filament polymerization, and actin binding" in GO enrichment analysis. In the skeletal muscle of Japanese flounder, the expression level of actin-related protein 3B-like in stage B is higher than that in stage A, indicating that this gene may be related to one of the factors that affect muscle development. The predicted regulatory lncRNA, TCONS_00041871, can control the expression of actin-related protein 3B-like through the *cis*-regulatory mechanism and is expressed at a higher level in skeletal muscle in stage A than in stage B. This negative correlation between LncRNA and target genes indicates that lncRNA regulates the muscle development of Japanese flounder by inhibiting the expression of the gene. Kin of IRRE like 2 (kirrel2) is predicted to be a *cis*-acting target of TCONS_00089031 (**Figure 3A**), and it is enriched in the terms of "myosin binding and positive regulation of actin filament polymerization" in GO enrichment analysis. In the skeletal muscle of Japanese flounder, the expression level of kirrel2 in stage A is higher than that in stage B. The predicted regulatory lncRNA, TCONS_00089031 can regulate the expression of kirrel2 through the *cis*-regulatory mechanism and is expressed at a higher level in skeletal muscle in stage A than in stage B. This positive correlation between lncRNA and target genes indicates that lncRNA regulates the muscle development of *Paralichthys olivaceus* by promoting the expression of the gene. In the lncRNA–gene network comparing A and D, the myosin-7B-like gene is the predicted *cis*-target of TCONS_00038291 (**Figure 3C**). In addition, it is enriched in the term of "myosin filament" in GO enrichment analysis. These results indicate that these lncRNAs play a role in the development of Japanese flounder skeletal muscle by regulating the related genes.

Both lncRNA and miRNA have their own regulatory networks. Their regulatory networks are not independent but are intertwined and interdependent, and they also have regulatory relationships and form complex regulatory networks with mRNA. In the study of pancreatic cancer, the large-scale effects of interrelated miRNAs are revealed by establishing an lncRNA–miRNA–mRNA regulatory network and constructing a model for predicting the disease mechanisms of miRNAs (Ye et al., 2014). In this study, we also predicted the biological function of lncRNAs by establishing lncRNA–miRNA–mRNA networks. Co-expression networks show that most lncRNAs interact with one or two predicted miRNAs that are involved in muscle growth and development. Some of these lncRNAs (including TCONS_00093971, TCONS_00096817, TCONS_00032744, etc.) have established interactions with at least three target miRNAs. Although these lncRNAs require further experimental validation, this information may help us explore the potential regulatory mechanisms of lncRNAs during Japanese flounder skeletal muscle growth and development.

DATA AVAILABILITY STATEMENT

The sequencing data obtained from the RNA-seq were released to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database under the accession numbers SRR12102079, SRR12102078, SRR12102077, SRR12102076, SRR12102075, SRR12102074, SRR12102073, SRR12102072, SRR12102071, SRR12102070, SRR12102069, and SRR12102068.

ETHICS STATEMENT

The animal study was reviewed and approved by the Respective Animal Research and Ethics Committees of Ocean University of China.

AUTHOR CONTRIBUTIONS

SW finished the experiment, analyzed the data, designed the tables and figures, and drafted the manuscript. BL, JZ, YH, SL, MZ, JL, YL, and HW revised the manuscript. FH conceived the study and revised the manuscript. All authors contributed to manuscript revision and approved the submitted version.

FUNDING

This research was supported by the National Nature Science Foundation of China (31672642) and the Natural Science Foundation of Shandong Province (ZR2014CM018). It is appreciated that the comments from editors and reviewers have greatly improved our manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.01034/full#supplementary-material

REFERENCES

Anders, S., and Huber, W. (2012). *Differential Expression of RNA-Seq data at the Gene Level–the DESeq Package*. Heidelberg: EMBL.

Bakhtiarizadeh, M. R., Hosseinpour, B., Arefnezhad, B., Shamabadi, N., and Salami, S. A. (2016). In silico prediction of long intergenic non-coding RNAs in sheep. *Genome* 59, 263–275. doi: 10.1139/gen-2015-0141

Buckingham, M. (2006). Myogenic progenitor cells and skeletal myogenesis in vertebrates. *Curr. Opin. Gen. Dev.* 16, 525–532. doi: 10.1016/j.gde.2006.08.008

Cabili, M. N., Cole, T., Loyal, G., Magdalena, K., Barbara, T. V., Aviv, R., et al. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927. doi: 10.1101/gad.17446611

Cai, B., Li, Z., Ma, M., Wang, Z., Han, P., Abdalla, B. A., et al. (2017). LncRNA-Six1 encodes a micropeptide to activate Six1 in Cis and is involved in cell proliferation and muscle growth. *Front. Physiol.* 8:230. doi: 10.3389/fphys.2017.00230

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., et al. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 147, 358–369. doi: 10.1016/j.cell.2011.09.028

Durcan, P. J., Conradie, J. D., Devyver, M. V., and Myburgh, K. H. (2014). Identification of novel Kirrel3 gene splice variants in adult human skeletal muscle. *BMC Physiol.* 14:11. doi: 10.1186/s12899-014-0011-3

Eng, D., Ma, H. Y., Gross, M. K., and Kioussi, C. (2013). gene networks during skeletal myogenesis. *ISRN Dev. Biol.* 2013, 1–8. doi: 10.1155/2013/348704

Finn, R. D., Alex, B., Jody, C., Penelope, C., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, 222–230.

Fu, Y., Shi, Z., Wu, M., Zhang, J., Jia, L., and Chen, X. (2011). Identification and differential expression of MicroRNAs during metamorphosis of the japanese flounder (*Paralichthys olivaceus*). *PLoS One* 6:e22957. doi: 10.1371/journal.pone.0022957

Fu, Y. S., Shi, Z. Y., Wang, G. Y., Li, W. J., Zhang, J. L., and Jia, L. (2012). Expression and regulation of miR-1, -133a, -206a, and MRFs by thyroid hormone during larval development in Paralichthys olivaceus. *Comp. Biochem. Physiol. Part B Biochem. Mol. Biol.* 161, 226–232. doi: 10.1016/j.cbpb.2011.11.009

Gao, X., Ye, J., Yang, C., Zhang, K., Li, X., Luo, L., et al. (2017). Screening and evaluating of long noncoding RNAs in the puberty of goats. *BMC Genomics* 18:164. doi: 10.1186/s12864-017-3578-9

Geng, C., Ziyun, W., Dongqing, W., Chengxiang, Q., Mingxi, L., Xing, C., et al. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* 41, D983–D986.

Girgenrath, S., Song, K., and Whittemore, L. A. (2005). Loss of myostatin expression alters fiber-type distribution and expression of myosin heavy chain isoforms in slow- and fast-type skeletal muscle. *Muscle Nerve* 31, 34–40. doi: 10.1002/mus.20175

Goodrich, J., and Kugel, J. (2006). Non-coding-RNA regulators of RNA polymerase II transcription. *Nat. Rev. Mol. Cell Biol.* 7, 612–616. doi: 10.1038/nrm1946

Grote, P., Wittler, L., Hendrix, D., Koch, F., Währisch, S., Beisaw, A., et al. (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev. Cell* 24, 206–214. doi: 10.1016/j.devcel.2012.12.012

Han, L., Zhang, K., Shi, Z., Zhang, J., Zhu, J., Zhu, S., et al. (2012). LncRNA profile of glioblastoma reveals the potential role of lncRNAs in contributing to glioblastoma pathogenesis. *Int. J. Oncol.* 40, 2004–2012.

Handschin, C., Rhee, J., Lin, J., Tarr, P. T., and Spiegelman, B. M. (2003). An autoregulatory loop controls peroxisome proliferator-activated receptor γ coactivator 1α expression in muscle. *Proc. Natl. Acad. Sci. U.S.A.* 100, 7111–7116. doi: 10.1073/pnas.1232352100

Huang, Y., Wen, H., Zhang, M., Hu, N., Si, Y., Li, S., et al. (2018). DNA methylation status of MyoD and IGF-I gene correlated with its muscle growth during different development stages of Japanese flounder (*Paralichthys olivaceus*). *Comp. Biochem. Physiol. Part B Biochem. Mol. Biol.* 219-220, 33–43.

Iyer, M. K., Niknafs, Y. S., Rohit, M., Udit, S., Anirban, S., Yasuyuki, H., et al. (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* 47, 199–208.

Jebessa, E., Ouyang, H., Abdalla, B. A., Li, Z., Abdullahi, A. Y., Liu, Q., et al. (2018). Characterization of miRNA and their target gene during chicken embryo skeletal muscle development. *Oncotarget.* 9, 17309–17324. doi: 10.18632/oncotarget.22457

Keren, A., Tamir, Y., and Bengal, E. (2006). The p38 MAPK signaling pathway: a major regulator of skeletal muscle development. *Mol. Cell. Endocrinol.* 252, 224–230. doi: 10.1016/j.mce.2006.03.017

Krasnyi, A. M., and Ozernyuk, N. D. (2011). The expression of genes encoding the voltage-dependent L-type Ca 2+ channels in proliferating and differentiating C2C12 myoblasts of mice. *Biol. Bull.* 38, 292–296. doi: 10.1134/s1062359011030071

Lei, K., Yong, Z., Zhi-Qiang, Y., Xiao-Qiao, L., Shu-Qi, Z., Liping, W., et al. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* 35, W345–W349.

Li, A., Zhang, J., and Zhou, Z. (2014). PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k- mer scheme. *BMC Bioinformatics* 15:311. doi: 10.1186/1471-2105-15-311

Liang, S., Haitao, L., Dechao, B., Guoguang, Z., Kuntao, Y., Changhai, Z., et al. (2013). Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41, e166–e166.

Liang, T., Zhou, B., Shi, L., Wang, H., Chu, Q., Xu, F., et al. (2017). lncRNA AK017368 promotes proliferation and suppresses differentiation of myoblasts in skeletal muscle development by attenuating the function of miR-30c. *FASEB J.* 32, 377–389. doi: 10.1096/fj.201700560rr

Lina, M., Bajic, V. B., and Zhang, Z. (2013). On the classification of long non-coding RNAs. *RNA Biol.* 10, 924–933.

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

Lukiw, W. J., Handley, P., Wong, L., and Mclachlan, D. R. C. (1992). BC200 RNA in normal human neocortex, non-Alzheimer dementia (n.d.), and senile dementia of the Alzheimer type (AD). *Neurochem. Res.* 17, 591–597. doi: 10.1007/bf00968788

Lv, J., Liu, H., Yu, S., Liu, H., Cui, W., Gao, Y., et al. (2015). Identification of 4438 novel lincRNAs involved in mouse pre-implantation embryonic development. *Mol. Gen. Genom.* 290, 685–697. doi: 10.1007/s00438-014-0952-z

Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009). Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* 10, 155–159.

Mercer, T. R., and Mattick, J. S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. *Nat. Struct. Mol. Biol.* 20, 300–307. doi: 10.1038/nsmb.2480

Ming-De, Y., Chih-Chen, H., Gi-Ming, L., Ann-Lii, C., Ya-Wen, L., and Shuang-En, C. (2005). Identification and characterization of a novel gene Saf transcribed from the opposite strand of Fas. *Hum. Mol. Genet.* 14, 1465. doi: 10.1093/hmg/ddi156

Minoru, K., Michihiro, A., Susumu, G., Masahiro, H., Mika, H., Masumi, I., et al. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* 36, D480-D484.

Morlando, M., Ballarino, M., Fatica, A., and Bozzoni, I. (2014). The role of long noncoding RNAs in the epigenetic control of gene expression. *Chemmedchem* 9, 505–510. doi: 10.1002/cmdc.201300569

Mousavi, K., Zare, H., Dell'Orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., et al. (2013). eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol. Cell.* 51, 606–617. doi: 10.1016/j.molcel.2013.07.022

Mueller, A. C., Cichewicz, M. A., Dey, B. K., Layer, R., Reon, B. J., Gagan, J. R., et al. (2015). MUNC, a long noncoding RNA that facilitates the function of MyoD in skeletal myogenesis. *Mol. Cell. Biol.* 35, 498–513. doi: 10.1128/mcb.01079-14

Naya, F. J., Mercer, B., Shelton, J., Richardson, J. A., Williams, R. S, and Olson, E. N. (2000). Stimulation of slow skeletal muscle fiber gene expression by calcineurin in vivo. *J. Biol. Chem.* 275, 4545–4548. doi: 10.1074/jbc.275.7.4545

Nishina, H., Green, L. R., Mcgarrigle, H. H., Noakes, D. E., Poston, L., and Hanson, M. A. (2003). Effect of nutritional restriction in early pregnancy on isolated femoral artery function in mid-gestation fetal sheep. *J Physiol.* 553(Pt 2), 637–647. doi: 10.1113/jphysiol.2003.045278

Park, J. W., Lee, J. H., Kim, S. W., Han, J. S., and Park, T. S. (2018). Muscle differentiation induced up-regulation of calcium-related gene expression in quail myoblasts. *Asian Australas. J. Anim. Sci.* 31, 1507–1515. doi: 10.5713/ajas.18.0302

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., et al. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* 22, 577–591. doi: 10.1101/gr.133009.111

Perkel, J. M. (2013). Visiting "noncodarnia". *Biotechniques* 54, 303–304.

Philipp, K., Jill, C., Sujit, D., Nix, D. A., Radharani, D., Willingham, A. T., et al. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* 316, 1484–1488. doi: 10.1126/science.1138341

Phillip, G., Lars, W., David, H., Frederic, K., Sandra, W. H., Arica, B., et al. (2013). The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. *Dev. Cell* 24, 206–214. doi: 10.1016/j.devcel.2012.12.012

Pieter-Jan, V., Kenny, H., Xiaowei, W., BjoRn, M., Lennart, M., Kris, G., et al. (2013). LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.* 41, 246–251.

Qian, B., Zhengtao, H., Feng, C., Ruiming, Z., Yi, D., Xue, S., et al. (2012). Transcriptome analysis of long non-coding RNAs of the nucleus accumbens in cocaine-conditioned mice. *J. Neurochem.* 123, 790–799. doi: 10.1111/jnc.12006

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L., Lotia, S., et al. (2012). A travel guide to cytoscape plugins. *Nat. Methods* 9, 1069–1076. doi: 10.1038/nmeth.2212

Sun, B., Li, X., Ning, X., and Sun, L. (2020). Transcriptome analysis of paralichthys olivaceus erythrocytes reveals profound immune responses induced by edwardsiella tarda infection. *Int. J. Mol. Sci.* 21:3094. doi: 10.3390/ijms21093094

Takayuki, A., Ribar, T. J., Williams, R.S., and Zhen, Y. (2004). Skeletal muscle adaptation in response to voluntary running in Ca2+/calmodulin-dependent protein kinase IV-deficient mice. *Am. J. Physiol. Cell Physiol.* 287, C1311–C1319.

Tang, Z., Yang, Y., Wang, Z., Zhao, S., Mu, Y., and Li, K. (2015). Integrated analysis of miRNA and mRNA paired expression profiling of prenatal skeletal muscle development in three genotype pigs. *Sci. Rep.* 5:15544.

Thomas, B., and Mathias, G. (2011). Transcriptional mechanisms regulating skeletal muscle differentiation, growth and homeostasis. *Nat. Rev. Mol. Cell Biol.* 12, 349–361. doi: 10.1038/nrm3118

Tseng, B. S., Zhao, P., Pattison, J. S., Gordon, S. E., Granchelli, J. A., Madsen, R. W., et al. (2002). Regenerated mdx mouse skeletal muscle shows differential mRNA expression. *J. Appl. Physiol.* 93, 537–545. doi: 10.1152/japplphysiol.00202.2002

Wang, L., Zhao, Y., Bao, X., Zhu, X., Kwok, Y. K.-Y., Sun, K., et al. (2015). LncRNA Dum interacts with Dnmts to regulate Dppa2 expression during myogenic differentiation and muscle regeneration. *Cell Res.* 25, 335–350. doi: 10.1038/cr.2015.21

White, N. M., Cabanski, C. R., Silva-Fisher, J. M., Dang, H. X., Govindan, R., and Maher, C. A. (2014). Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. *Genome Biol.* 15, 429.

Wu, S., Huang, Y., Li, S., Wen, H., Zhang, M., Li, J., et al. (2018). DNA methylation levels and expression patterns of Smyd1a and Smyd1b genes during Metamorphosis of the Japanese Flounder (*Paralichthys olivaceus*). *Comp. Biochem. Physiol. B. Biochem. Mol. Biol.* 223, 16–22. doi: 10.1016/j.cbpb.2018.05.002

Ye, S., Yang, L., Zhao, X., Song, W., Wang, W., and Zheng, S. (2014). Bioinformatics method to predict two regulation mechanism: TF–miRNA–mRNA and lncRNA–miRNA–mRNA in pancreatic cancer. *Cell Biochem. Biophys.* 70, 1849–1858. doi: 10.1007/s12013-014-0142-y

Young, M. D., Wakefield, M. J., Smyth, G. K., and Alicia, O. (2010). Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* 11, R14–R14.

Zhan, S., Dong, Y., Zhao, W., Guo, J., Zhong, T., Wang, L., et al. (2016). Genome-wide identification and characterization of long non-coding RNAs in developmental skeletal muscle of fetal goat. *BMC Genomics* 17:666. doi: 10.1186/s12864-016-3009-3

Zhang, W., Liu, Y., Yu, H., Du, X., Zhang, Q., Wang, X., et al. (2016). Transcriptome analysis of the gonads of olive flounder (*Paralichthys olivaceus*). *Fish Physiol. Biochem.* 42, 1581–1594. doi: 10.1007/s10695-016-0242-2

Zhang, Y. C., Liao, J. Y., Li, Z. Y., Yu, Y., Zhang, J. P., Li, Q. F., et al. (2014). Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. *Genome Biol.* 15:512.

Zhang, Z. K., Li, J., Guan, D., Liang, C., Zhuo, Z., Liu, J., et al. (2018). A newly identified lncRNA MAR1 acts as a miR-487b sponge to promote skeletal muscle differentiation and regeneration. *J. Cachexia Sarcopenia Muscle* 9, 613–626. doi: 10.1002/jcsm.12281

Zhou, R., Wang, Y., Long, K., Jiang, A., and Jin, L. (2018). Regulatory mechanism for lncRNAs in skeletal muscle development and progress on its research in domestic animals. *Yi Chuan* 40, 292–304.

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

The reviewer QZ declared a shared affiliation, with no collaboration, with several of the authors, FH, SW, JZ, BL, YH, SL, HW, MZ, JL, YL, to the handling editor at the time of review.

frontiers in Genetics

ORIGINAL RESEARCH
published: 30 September 2020
doi: 10.3389/fgene.2020.00869

Learning Cell-Type-Specific Gene Regulation Mechanisms by Multi-Attention Based Deep Learning With Regulatory Latent Space

Minji Kang [1†], Sangseon Lee [1†], Dohoon Lee [2] and Sun Kim [1,2,3]*

[1] Bioinformatics Institute, Seoul National University, Seoul, South Korea, [2] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea, [3] Department of Computer Science and Engineering, Institute of Engineering Research, Seoul National University, Seoul, South Korea

Epigenetic gene regulation is a major control mechanism of gene expression. Most existing methods for modeling control mechanisms of gene expression use only a single epigenetic marker and very few methods are successful in modeling complex mechanisms of gene regulations using multiple epigenetic markers on transcriptional regulation. In this paper, we propose a multi-attention based deep learning model that integrates multiple markers to characterize complex gene regulation mechanisms. In experiments with 18 cell line multi-omics data, our proposed model predicted the gene expression level more accurately than the state-of-the-art model. Moreover, the model successfully revealed cell-type-specific gene expression control mechanisms. Finally, the model was used to identify genes enriched for specific cell types in terms of their functions and epigenetic regulation.

Keywords: gene regulation mechanism, gene regulatory network, multi-omics, deep learning, cell-type-specific

OPEN ACCESS

Edited by:
Jiajie Peng,
Northwestern Polytechnical University,
China

Reviewed by:
Eunhee Choi,
Harvard Medical School,
United States
Tianyi Zhao,
Harbin Institute of Technology, China

****Correspondence:***
Sun Kim
sunkim.bioinfo@snu.ac.kr

† These authors share first authorship

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

Received: *07 February 2020*
Accepted: *16 July 2020*
Published: *30 September 2020*

Citation:
Kang M, Lee S, Lee D and Kim S (2020) Learning Cell-Type-Specific Gene Regulation Mechanisms by Multi-Attention Based Deep Learning With Regulatory Latent Space. Front. Genet. 11:869. doi: 10.3389/fgene.2020.00869

1. INTRODUCTION

Epigenetic gene regulation is a major control mechanism of gene expression. Histone modifications one of the most versatile modes of chromatin regulation among diverse epigenetic regulatory mechanisms are defined as covalent modifications of a set of specific amino acids at N-terminal tails of histone proteins. Combinations of the type of amino acids and their modifications constitute "histone codes" that are distributed across the genome and are known to regulate overall chromatin states. On the other hand, DNA methylation occurs directly at the cytosine bases of DNA and regulates gene expression in part by altering the binding affinity of most of the transcription factors. Besides the individual effect of each epigenetic modification, the complexity of epigenetic gene regulation mostly arises from the crosstalk between the different types of epigenetic modifications. For example, positive interplay between histone marks (1) H2BK120u1 and H3K4me3, and (2) H3K4me3 and H3/H4 acetylation (Zhang et al., 2015) is an example of the complex epigenetic regulation. Furthermore, some histone modifications are known to be associated with DNA methylation (Cedar and Bergman, 2009). *De novo* DNA methyltransferases, DNMT3A and DNMT3B, are known to physically interact with specific histone marks, H3K36me3 and H3K4me0, through their internal PWWP and ADD domain, respectively. Methyl-CpG-binding domain (MBD) proteins have been reported to "read" methylated CpG, and recruit chromatin-modifying complexes such as SWI/SNF components (Fatemi and Wade, 2006). Subtle epigenetic interactions between different types of histone modifications and DNA methylation can therefore be regarded as a major determinant of the general chromatin structure of cells that govern the accessibility of transcription factors to the chromatin.

Given the essential role of epigenetic alterations in regulating gene expression, a number of studies on modeling the regulatory effects of these epigenetic markers have been performed. However, existing modeling methods utilize only a single epigenetic marker. Some studies have investigated the role of histone marks in the context of gene regulation. DeepChrome (Singh et al., 2016) used a Convolutional Neural Network based model to model gene regulation. It was the first deep learning approach to predict the gene expression level, and it captured local characteristics of histone marks. Another study, AttentiveChrome (Singh et al., 2017), proposed a hierarchy of multiple Long Short-Term Memory modules with an attention mechanism to predict gene expression levels. AttentiveChrome predicted gene expression more accurately than DeepChrome, and it showed which histone marks or which gene loci were used, using an attention mechanism. Both studies used individual deep learning approaches to understand gene regulation but utilized histone marks only. There have also been studies to identify relationships between genome-wide DNA methylation and gene expression. Wagner et al. (2014) investigated the relationships between DNA methylation and the gene expression profile of primary fibroblast samples from 62 individuals. More recently, Zhong et al. (2019) predicted gene expression using DNA methylation in human populations, using linear regression-based methods. Recent studies investigated the relationships between mutation and gene expression. Zeng et al. (2017) used a linear regression-based model to predict gene expression with cis-SNPs. Xie et al. (2017) examined the effectiveness of a deep auto-encoder to predict the gene expression profile measured in yeast with SNP. These prior studies on gene expression prediction revealed relationships between gene expression and a single epigenetic marker of histone marks, DNA methylation, or SNP. However, these studies were not designed to model complex transcriptional control mechanisms involving the interplay of various epigenetic regulatory modules.

We therefore introduce an *explainable* deep learning model with a multi-attention network for epigenetic regulation mechanisms. Our model integrates multiple markers such as histone marks, DNA methylation, and transcription factors and explains the complex interactions between the molecular regulators. The attention network modules of our model allow human experts to understand the gene regulation mechanisms. Moreover, the model characterizes cell-type-specific gene regulation mechanisms for 18 cell lines, based on the weights of the Multi-Attention network. In summary, the proposed model provides a better understanding of cell-type-specific gene regulation.

2. MATERIALS AND METHODS

We propose a two-step ensemble deep learning model for gene expression prediction and the architecture is illustrated in **Figure 1**. At the first layer of the model, separate neural networks vectorize epigenetic and transcriptional markers with different strategies, and then at the second layer, output vectors from the first layer are integrated by a Multi-Attention network. To predict gene expression, we used the same outputs previously used in DeepChrome (Singh et al., 2016) and AttentiveChrome (Singh et al., 2017). All genes are divided into highly expressed genes (HEG) and lowly expressed genes (LEG) according to their expression levels, which formulates the problem as a binary classification task.

To begin, separate models embed histone marks, DNA methylation, and transcription factors into a regulatory latent space. First, histone marks are embedded into the latent space by a Convolutional Neural Network (CNN) followed by a Bi-directional Long Short-Term Memory (LSTM) network with attention. Second, DNA methylation is vectorized by a Dynamic Bi-directional LSTM with attention. Lastly, a Self-Attention Network (SAN) embeds the transcription factors. After embedding features in three vectors, a Multi-Attention network combines these vectors to predict whether a gene would be highly expressed or lowly expressed. While the end-to-end model predicts the gene expression level as a whole, the Multi-Attention network determines which types of epigenetic markers are most influential for controlling gene expression and how epigenetic features interact with each other in each cell type.

We used datasets from the Roadmap Epigenomics Projects (Kundaje et al., 2015) to predict the gene expression level of 18 cell lines, for which data measuring levels of histone marks, DNA methylation, and transcription factors are available (**Table 1**, **Supplementary Figure 1**). The epigenetic and transcriptional markers near the transcription start site (TSS) mainly involve in gene expression. We therefore focused on the gene region of 4,000 base-pair (bp) around the TSS for histone markers or DNA methylation and 200 bp around the TSS for transcription factors. To implement the model, we used Pytorch, an open-source machine learning library based on Python. Implementation of our model can be found at Github (https://github.com/pptnz/deeply-learning-regulatory-latent-space).

In the following sections, deep learning models for each of the epigenetic and transcriptional markers are explained.

2.1. Embedding Histone Marks

We used seven core histone marks: H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K27ac, and H3K9ac. Among 31 histone marks in the Roadmap Epigenomics Projects, the seven core histone marks had been profiled and investigated the most. Each of the seven histone marks were profiled for more than 62 cell lines, whereas other histone marks were profiled for less than 24 cell lines (**Supplementary Figure 2**). To investigate cell-type-specific gene regulation mechanisms, we used the seven histone marks with abundant cell line data.

To vectorize the histone marks, we used CNN, followed by Bi-directional LSTM with an attention mechanism. CNN is a deep learning architecture proposed for extracting local features of various sizes in two-dimensional images (Min et al., 2016). In this model, CNN captures local patterns of the seven histone marks. RNN is a deep learning architecture with a cyclic structure, which has caught the limelight in natural language processing

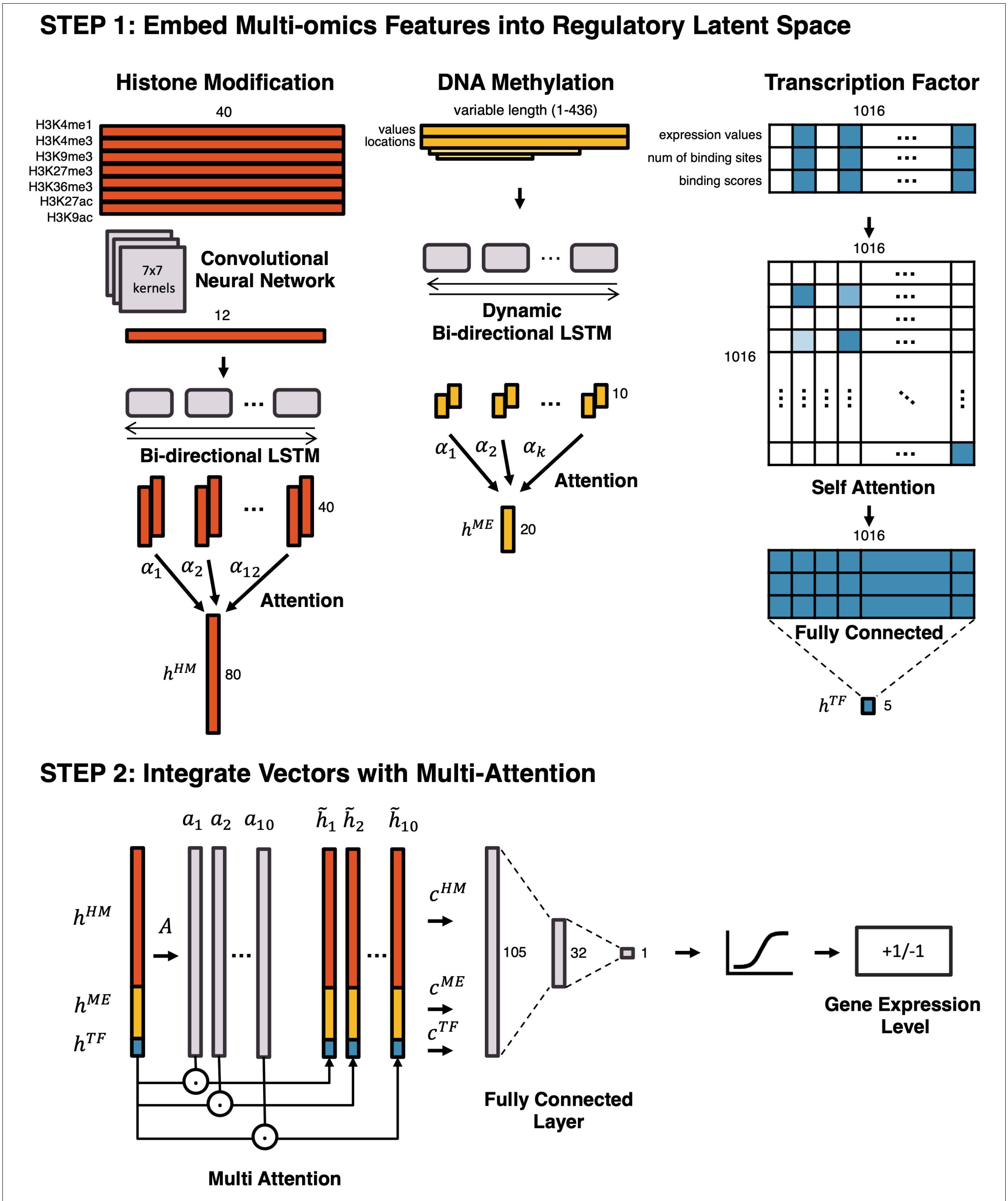


FIGURE 1 | An overview of the proposed model. To predict gene expression level and to model the regulation mechanism, a Multi-Attention based deep learning model with regulatory latent space is designed. It consists of two steps: (1) embedding multi-omics features into a regulatory latent space, and (2) integrating latent vectors with a Multi-Attention network. In the first step, different deep learning architectures are utilized to reflect the characteristics of each omics feature. By omics-specific layers, multi-omics features are transformed into latent vectors in the regulatory latent space. In the second step, the latent vectors are integrated by a Multi-Attention network. The attention weights of multi-omics features represent their effects on the gene regulation.

TABLE 1 | Cell line data used in this study.

Cell line	Group	Standardized epigenome name
E003	ESC	H1 Cells
E004	ES-deriv	H1 BMP4 derived mesendoderm cultured cells
E005	ES-deriv	H1 BMP4 derived trophoblast cultured cells
E006	ES-deriv	H1 derived mesenchymal stem cells
E007	ES-deriv	H1 derived neuronal progenitor cultured cells
E011	ES-deriv	hESC derived CD184+ endoderm cultured cells
E016	ESC	HUES64 cells
E038	Blood & T-cell	Primary T helper naive cells from peripheral blood
E047	Blood & T-cell	Primary T CD8+ naive cells from peripheral blood
E066	Tissue & Primary cell	Liver
E087	Tissue & Primary cell	Pancreatic islets
E114	Cancer cell line	A549 EtOH 0.02pct lung carcinoma cell line
E116	Cancer cell line	GM12878 lymphoblastoid cells
E117	Cancer cell line	HeLa-S3 cervical carcinoma cell line
E118	Cancer cell line	HepG2 hepatocellular carcinoma cell line
E119	Tissue & Primary cell	HMEC mammary epithelial primary cells
E120	Tissue & Primary cell	HSMM skeletal muscle myoblasts cells
E123	Cancer cell line	K562 leukemia cells

fields (Min et al., 2016). LSTM is one of the architectures of the Recurrent Neural Network, proposed for considering long-term dependencies (Gers et al., 2000). Unlike other RNN architectures, LSTM has a forget gate, which allows the model to forget irrelevant parts of a sequence and deal with a long sequence. In our model, LSTM captures sequential patterns. The attention mechanism reveals important gene loci.

The histone marks in a gene region of 4,000 bp around TSS are divided into 40 bins with a bin size of 100 bp. On each bin, log read counts are calculated for each histone mark, respectively. The preprocessed histone mark matrix of size 7 × 40 is fed into a CNN that consists of a convolutional layer, a batch normalization layer, and a 1D max-pooling layer. In the convolutional layer, 100 kernels of size 7 × 7 are used, so that a vector of size 1 × 34 is produced. In the max-pooling layer, a kernel with size 3 and stride 3 is used with left and right padding. Afterward, the output vector of the CNN is fed into the Bi-directional Long Short-Term Memory (LSTM) with attention, producing a h^{HM} of size 80.

2.2. Embedding DNA Methylation

We used methylation values at all CpG sites within up/down-stream of 2,000 bp from TSS. DNA methylation is vectorized by a Dynamic Bi-directional LSTM with attention. The number of CpG sites vary for different genes. Thus, the "Dynamic" LSTM deals with the variable number of CpGs, and the "Bi-directional" LSTM considers both directions of the DNA strands. Dynamic Bi-directional LSTM produces the output vector h^{ME} of a fixed size 20.

2.3. Embedding Transcription Factors

We first selected candidate binding transcription factors (TFs) for each gene, based on prior knowledge of human transcription factors in Lambert et al. (2018), and the motif detection tool, HOMER (Heinz et al., 2010). We utilized TFs that have their binding sites within the region of 200 bp around the TSS. Based on this configuration, an input matrix for TFs is processed as a matrix of size 3 x 1016. Three rows of an input matrix represent TF expression values, the number of binding sites, and the binding scores of TFs by HOMER. One-thousand-and-sixteen columns of the matrix represent human transcription factors. Except for the candidate binding transcription factors, all columns are masked to zero.

Since the data of transcription factors are discrete rather than sequential, CNN or LSTM cannot be employed. Thus, a Self-Attention Network (SAN) is used to embed the input matrix in vector a h^{TF} of size 5. As a result of SAN, the attention weight matrix is produced, providing vital information about relationships and interactions between transcription factors.

2.4. Integrating Latent Vectors

To integrate latent vectors, we used the Multi-Attention Block from the Multi-Attention Recurrent Network (MARN) (Zadeh et al., 2018). MARN was proposed for the comprehension of human communication with multi-modal data (language modality, vision modality, and acoustic modality). As it was designed to deal with data with different characteristics, MARN is suitable for dealing with three latent vectors from different multi-omics data.

First, all three latent vectors h^{HM}, h^{ME}, and h^{TF} are concatenated. The concatenated vector h is fed into a fully connected layer A. Multiple attention weights $a_1, a_2, ..., a_k$ are then produced, where k is the number of attentions. The k attention weights are multiplied to the concatenated vector, the vectors $\tilde{h}_1, \tilde{h}_2, ..., \tilde{h}_k$ are produced by element-wise multiplication of the concatenation and the k attention weights as $\tilde{h}_i = h \otimes a_i$.

Finally, a fully connected layer produces the predicted labels, which represents whether a gene is highly expressed (HEG, +1) or lowly expressed (LEG, -1).

3. RESULTS

To evaluate our proposed model, we split 18,070 genes into four-folds for each cell line. The first and second folds were used as a test and validation set, respectively, and the remaining two folds were used as a training set. Every result is averaged from a 4-fold cross-validation.

3.1. Performance Evaluation of Models With Histone Modification Only

We set a baseline with the state-of-the-art method for gene expression level prediction, AttentiveChrome (Singh et al., 2017). As AttentiveChrome was designed for histone marks only, instead of multiple epigenetic features, we trained both our model and AttentiveChrome using only seven histone marks for a fair comparison. We evaluated them with two metrics. (1) First, we performed a classification task on whether a gene is highly or lowly expressed in that cell line. (2) Second, gene expression value prediction was performed in terms of rank concordance

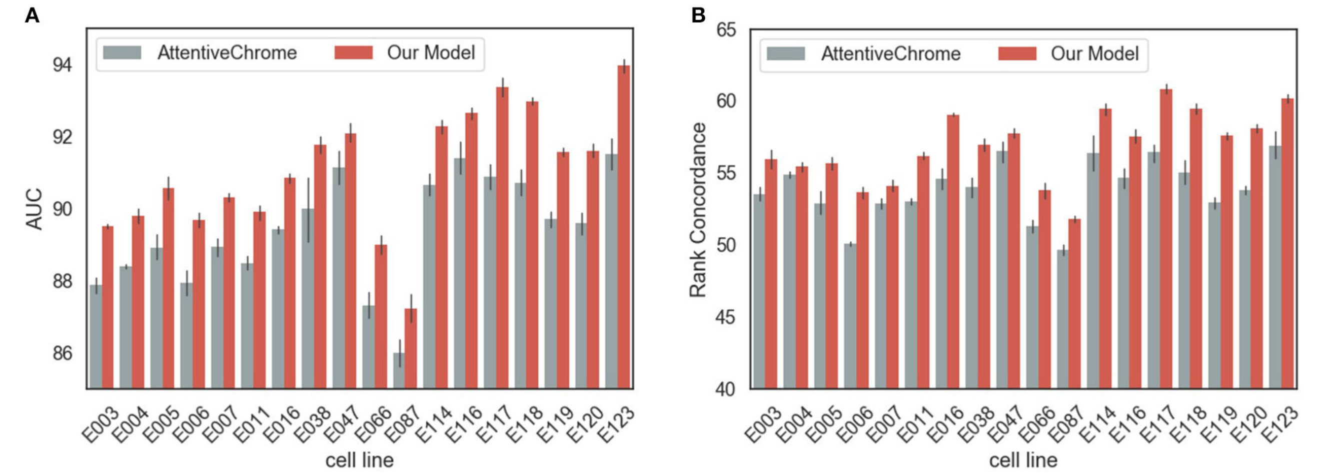


FIGURE 2 | Performance evaluation of models in predicting the gene expression levels. The performance of our model surpassed that of a baseline model, AttentiveChrome, in both criteria of **(A)** AUC and **(B)** Rank Concordance for two models for every cell line.

between the gene expression values and the final output values of the model.

In terms of AUC, rank concordance, and AUPR, our model outperformed AttentiveChrome for every cell line (**Figure 2**, **Supplementary Figure 3**). On average, the proposed model achieved 91.05% of AUC, while AttentiveChrome achieved 89.72%. Moreover, the model demonstrated its robustness by showing higher rank concordances between the gene expression value and the final output of the model. Our model showed 56.83% of rank concordance on average, while AttentiveChrome showed only 53.85%. We conjecture that the performance difference is due to the difference in model architectures. AttentiveChrome first uses an individual LSTM structure for each histone mark, and then integrates histone marks with an additional LSTM. On account of the individual LSTM, the interactions of numerous types of histone marks are likely to be neglected. Consequently, AttentiveChrome was not successful in capturing the local characteristics of seven histone marks. On the other hand, our model used both CNN and LSTM to capture local and sequential features of histone marks in a single model. Our model is therefore suitable for modeling not only the roles of histone marks but also interactions among them.

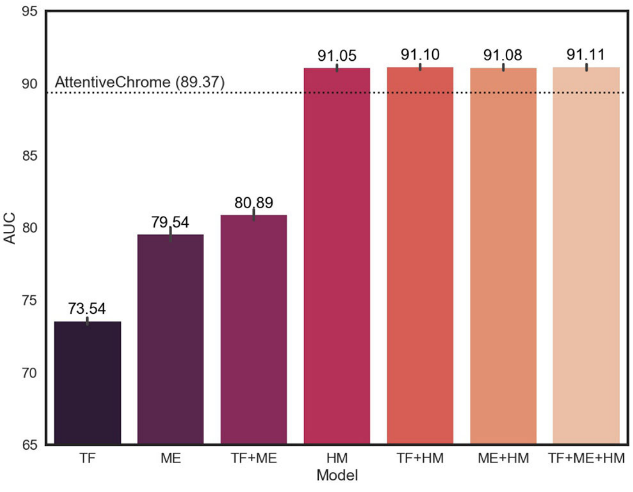


FIGURE 3 | Average AUC of 18 cell lines for different subsets of multi-omics features. TF, ME, and HM stand for transcription factors, DNA methylation, and histone marks, respectively. The model showed improvement in AUC, adding multiple epigenetic markers.

3.2. Performance Evaluation of Models With Multi-Omics Markers

Since our model is designed to utilize multi-omics biomarkers, we measured performance in terms of the average AUC and AUPR of our models that were trained on all possible combinations of multi-omics features (**Figure 3**, **Supplementary Figure 4**). The average AUC of the model improved when adding and integrating multi-omics features. In particular, the model with histone marks (HM, TF+HM, ME+HM, and TF+ME+HM) showed remarkable levels of AUC, exceeding the AUC of AttentiveChrome. This result is attributed to the fact that genes can be expressed if chromatins are opened, and thus histone marks are a major determinant of chromatin regulation.

3.3. Modeling Gene Regulation Mechanisms Using Multi-Omics Markers

Multi-omics markers are required to model gene regulation mechanisms. We focused on HeLa cell since its accuracy has been improved significantly by adding multiple markers (**Supplementary Figure 5**). In the HeLa cell, genes exist that cannot be predicted correctly using histone modification marks alone. **Figure 4** shows the average ChIP-seq reads of histone modification marks of genes with the same labels and predictions of the HM model. The model prediction is consistent: genes

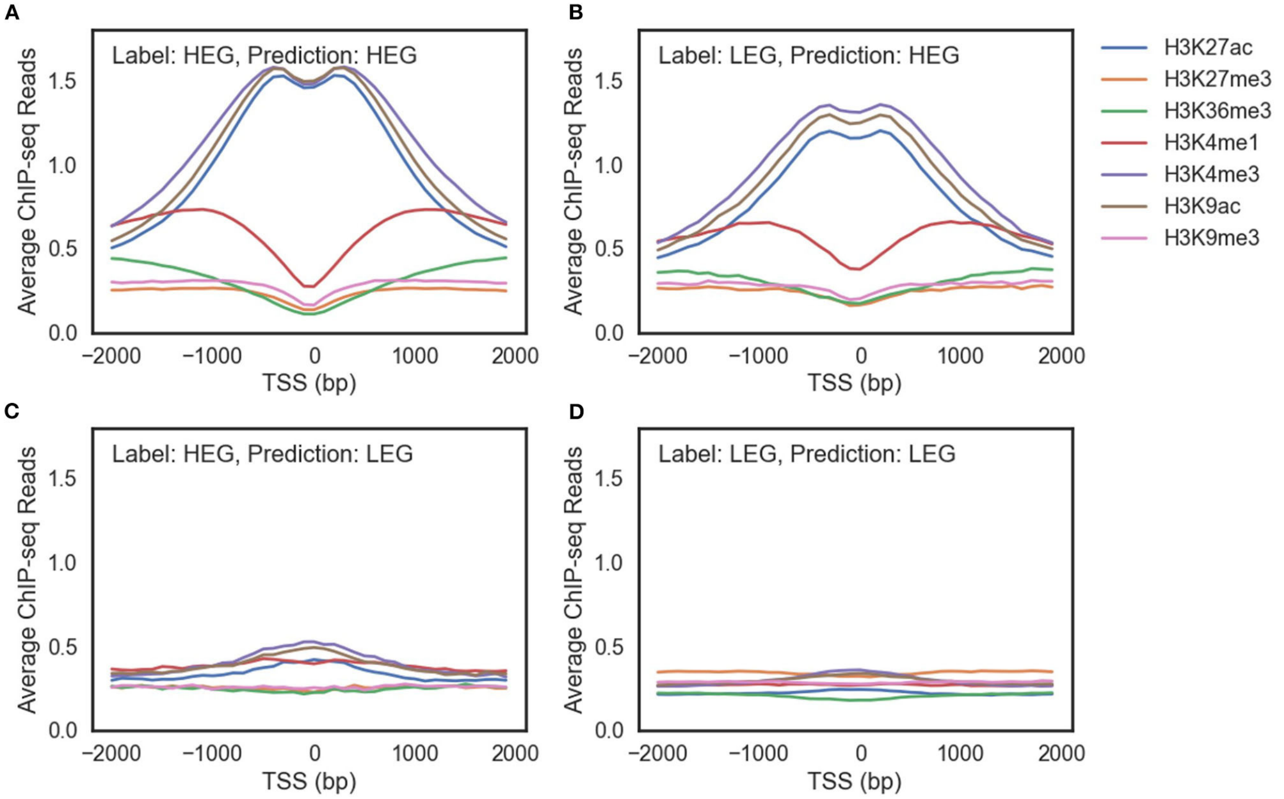


FIGURE 4 | The average ChIP-seq reads of seven histone modification marks of **(A)** highly expressed genes (HEG) predicted as HEG. **(B)** lowly expressed genes (LEG) predicted as HEG. **(C)** HEG predicted as LEG. **(D)** LEG predicted as LEG.

with the same prediction curves have common characteristics regardless of their true labels.

There were three differences between genes predicted as highly expressed genes (HEG) (**Figures 4A,B**) and genes predicted as lowly expressed genes (LEG) (**Figures 4C,D**). First, ChIP-seq reads of the genes predicted as HEG had a larger scale than the genes predicted as LEG. Second, the genes predicted as HEG had a noticeable peak around TSS for each histone mark associated with the activation of genes (H3K27ac, H3K4me3, and H3K9ac). Third, the genes predicted as HEG had a higher value of H3K4me1 around TSS, which is related to enhancers.

There are LEG even with an open chromatin state (**Figure 4B**) and HEG with weak signals of activation histone marks (**Figure 4C**). A model that only uses histone marks is limited in both predicting the gene expression level and characterizing the gene regulation mechanisms. In other words, multiple epigenetic markers such as DNA methylation and transcription factors are required to understand the complex gene regulation mechanisms.

RNF212, one of the enriched genes in the HeLa cell, epitomizes a gene that can be fully understood only by the multi-omics model, especially the TF+ME+HM model. The gene is a highly expressed gene but the model with histone marks alone failed to predict the gene expression level. This is due to the weak activation of histone marks. The intensities of histone marks associated with the activation of genes (H3K27ac, H3K4me3, and H3K9ac) were much smaller than those of other HEGs (i.e., *SLF1*) (**Figure 5A**). However, the gene was predicted correctly by the model with three epigenetic markers (TF+ME+HM). This is because the model could learn the regulation mechanisms of DNA methylation and transcription factors. **Figure 5B** illustrates the DNA methylation levels of *RNF212* and the attention weights of Dynamic LSTM. Surprisingly, the attention weight near TSS was high, and the region was unmethylated. The unmethylated promoter region enabled transcription factors to bind to the gene. **Figure 5C** shows all the possible binding transcription factors of the promoter region. These transcription factors were up-regulated especially in the HeLa cell compared to the other 17 cell lines (**Figure 5D**). Therefore, we can infer that *RNF212* could be highly expressed, despite weak signals of activation marks, thanks to the help of the "highly expressed" transcription factors, which bound to the unmethylated promoter region. As shown in the example, our multi-omics model reflected the multiple gene regulation mechanisms of histone marks, DNA methylation, and transcription factors.

3.4. Characterizing Cell-Type-Specific Gene Regulation Mechanisms

Figure 6 demonstrates the weights of the Multi-Attention Block in 18 cell lines. Every weight is normalized by the average weight

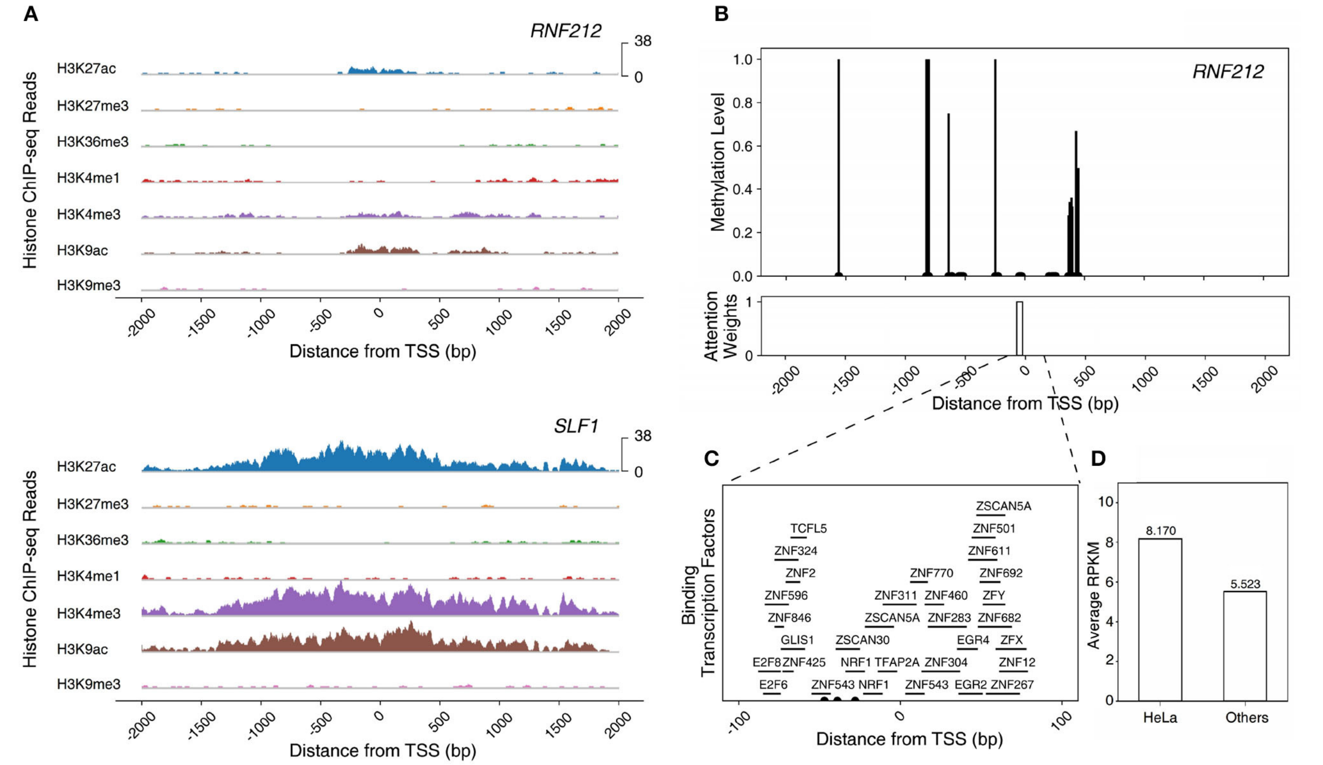


FIGURE 5 | Modeling the epigenetic regulation mechanism of *RNF212* in HeLa cell line. **(A)** The ChIP-seq reads of histone modification marks of *RNF212* and *SLF1*. **(B)** The DNA methylation levels and attention weights of *RNF212*. **(C)** The candidate binding transcription factors of the promoter region of *RNF212*. **(D)** The average expression value of the binding transcription factors of *RNF212* in HeLa cell line and other cell lines.

of 18 cell-lines in order to compare the importance of markers in each cell line. The attention weight of each feature shows how the model attends to the feature to predict gene expression level. The attention weight of each epigenetic marker therefore represents the importance of the markers in gene regulation. We compared the attention weights to reveal the important regulatory mechanisms in 18 cell lines and 5 cell types: ESC, ES-deriv, Blood & T-cell, Tissue & Primary Cell, and Cancer Cell Line.

There was no big difference in the weights of histone marks between the 18 cell lines. This is because histone modification plays a key role in the activation of genes, irrespective of cell type. Only after the chromatin structure of the gene is opened, can the gene be expressed. The higher AUC of the HM model (91.05) compare to that of the TF model (73.54) or the ME model (79.54), supports the importance of histone marks.

In contrast, weights of DNA methylation or transcription factors vary among cell types. In other words, DNA methylation and transcription factors determine the cell-type-specific gene regulatory mechanism. In general, cancer cell lines showed high attention weights of DNA methylation. The result is intuitive because DNA methylation is important in the development of cancer (Wajed et al., 2001; Kulis and Esteller, 2010). The abnormal patterns of methylation can inhibit gene expression and increase the probability of mutation (Wajed et al., 2001; Kulis and Esteller, 2010). It is commonly known that the hypermethylation of CpG islands inactivates tumor suppressor genes. Moreover, global hypomethylation significantly contributes to genome instability and aberrant gene expression.

In addition, embryonic stem cells showed the high attention weights of transcription factors. This reflects the crucial role of transcription factors in determining the fate of stem cells between self-renewal and differentiation. Transcriptional circuitry involving transcription factors like *OCT4, SOX2*, and *NANOG* is well-known to be a core regulatory mechanism of stem cells to maintain their stemness (Pan et al., 2002; Li, 2010). Furthermore, the significance of transcriptional regulation in embryonic stem cells has been highlighted since the prominent discovery, showing that ectopic overexpression of four essential transcription factors (*OCT4, SOX2, KLF4, MYC*), which are often referred to as "Yamanaka factors," are sufficient to induce the pluripotency of somatic cells.

Furthermore, we evaluated the cell-type-specificity and compatibility of our model, by training on one cell line and testing on other cell lines. For each cell line, the greatest AUC of the model was achieved when the model was trained on the cell line, demonstrating the cell-type-specificity. Moreover, it is notable that cell lines in the same group showed similar AUC patterns (**Figure 7**). By performing hierarchical clustering with

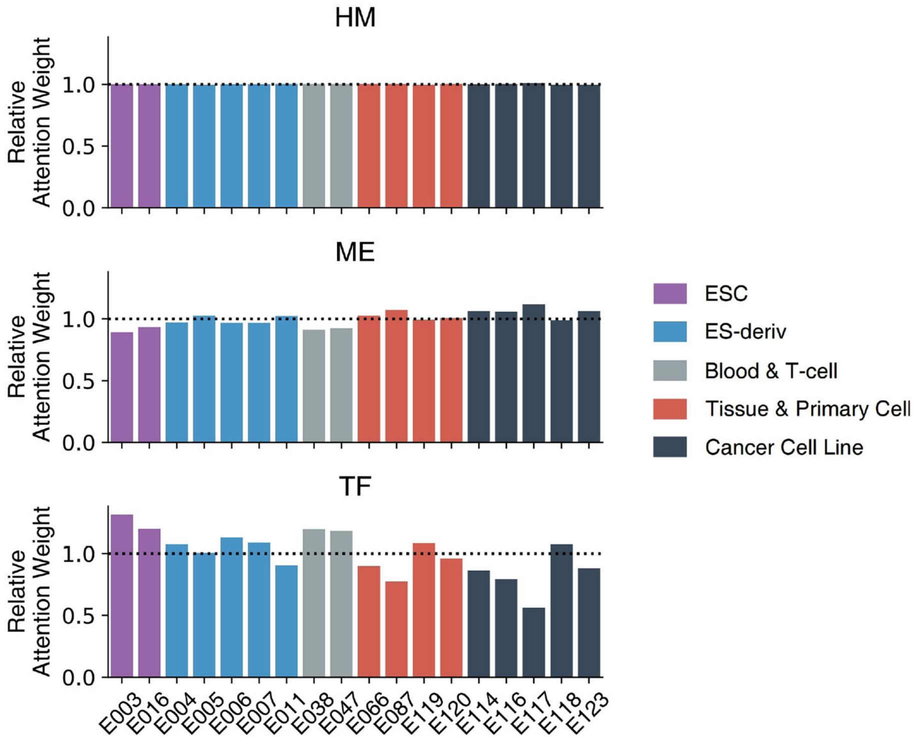


FIGURE 6 | The weights of the Multi-Attention Block in 18 cell lines. Based on relative attention weights, the importance of histone marks was not significantly different among cell lines. Moreover, methylation and transcription factors showed quite different weights among each cell line. In case of methylation, cancer cell lines were more focused on methylation compared to other cell lines. In case of transcription factor, ESC type cell lines had higher attention weights than those of others.

the Euclidean distance, cell lines in the same group were clustered together. The result highlights the transferability between the models in the same group, i.e., *transfer learning*. In other words, each cell line can be explained well by the model of the other cell lines if they are in the same group. For instance, the blood and T-cell group, E038 and E047, showed the best AUC for each other's model. This is probably because the cell lines in the same group tend to have similar gene regulation mechanisms.

3.5. Identification of Enriched Genes: Case Studies on HeLa and K562

Performances of the multi-omics model on the HeLa cell line and K562 cell line were quite improved compared to AttentiveChrome (**Supplementary Figure 5**). In addition, the multi-omics model better captured cell line enriched genes that were obtained from the Human Protein Atlas (http://www.proteinatlas.org; Uhlen et al., 2017). In the case of the HeLa cell line, the multi-omics model predicted 12 genes correctly among 20 enriched genes, while 9-10 genes were predicted correctly by the HM, TF+HM, and ME+HM models (**Supplementary Table 1**). On the other hand, in the case of the K562 cell line, 38 out of 62 genes were predicted correctly with the multi-omics model. Similar to the HeLa cell case, other models showed poor performances (34-37 genes, **Supplementary Table 2**). The number of correctly predicted HEG by each model is summarized in **Supplementary Figures 6, 7** for HeLa and K562, respectively.

We further investigated functions and epigenetic regulation mechanisms of cell-type enriched genes on the HeLa and K562 cell lines (**Figure 8**). *RNF212* was one of the HeLa cell enriched genes and was predicted correctly with the multi-omics only model. *RNF212* creates a cellular memory of DNA damage by tagging the lingering breaks (Qiao et al., 2018) and is known as a prognostic marker in cervical cancer in The Human Protein Atlas http://www.proteinatlas.org. The TF+HM and ME+HM model failed to predict the expression level of *RNF212*, while the TF+ME+HM model predicted it as an expressed gene. This result therefore implies that the expression of *RNF212* may be modulated by DNA methylation and transcription factors. It is also shown in the weights of the Multi-Attention Block in **Figure 8**.

PPARGC1A was also predicted correctly by the multi-omics model of the HeLa cell line. *PPARGC1A* belongs to the PCG-1 family that is associated with the regulation of mitochondrial biogenesis, promoting cell growth, proliferation, and evasion of the apoptosis signal (Lin et al., 2005; Jones et al., 2012). In particular, *PPARGC1A* modulates telomere function and the DNA damage mechanism in diabetes and cardiovascular disease (Lai et al., 2008; Xiong et al., 2015). Interestingly, the TF+ME+HM and ME+HM model, but not the TF+HM model, correctly predicted expression of the gene. In

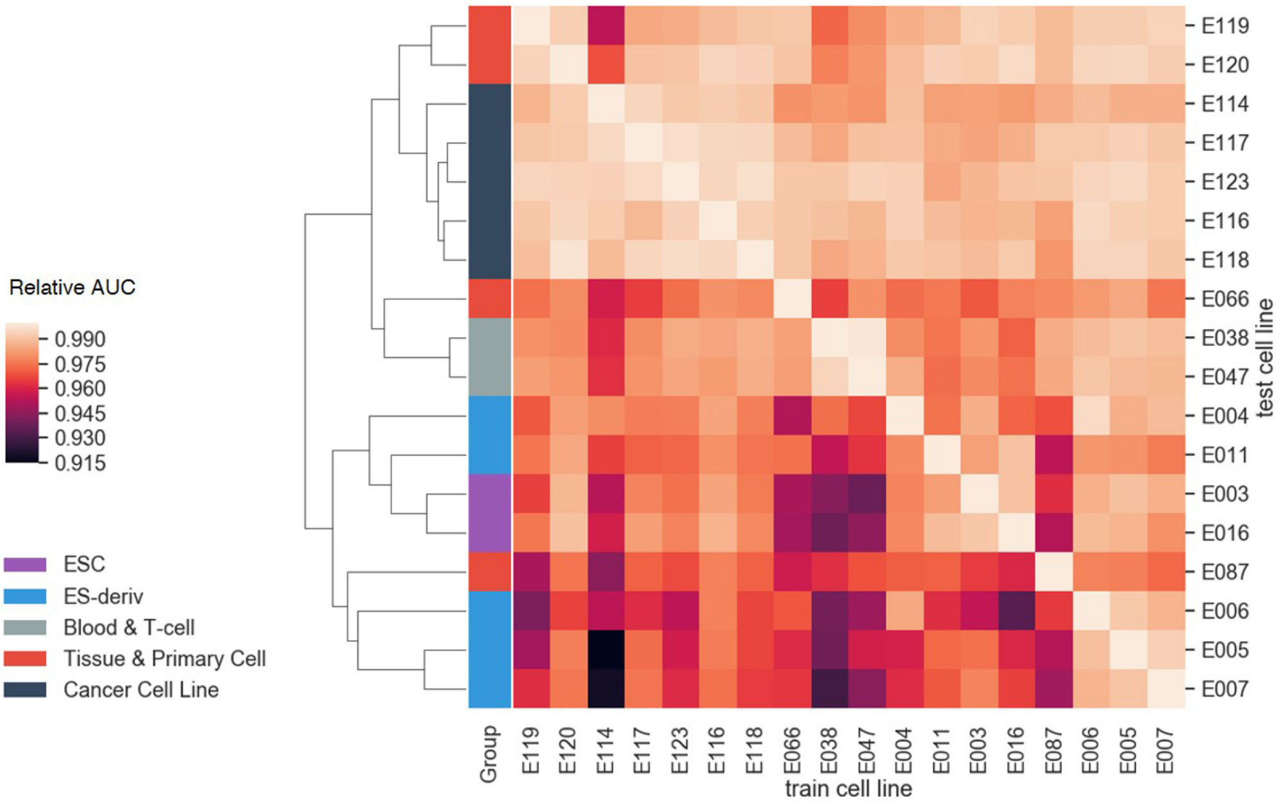


FIGURE 7 | The compatibility test results between cell lines. The AUC values of each test cell line were normalized with the AUC value of the model trained on that cell line. As a result of hierarchical clustering, cell lines in the same group showed similar AUC patterns. This result highlights the transferability between the models in the same group.

Figure 8, methylation was relatively more highlighted than other omics data. According to this observation, we speculated that methylation is one of the main regulators of *PPARGC1A*. This epigenetic regulation was already reported in other tissues such as brown adipose tissue or skeletal muscle tissue (Gillberg et al., 2013, 2014; Gill and La Merrill, 2017).

In the case of the K562 cell line, the *TRIM6* gene was captured by the multi-omics model. It belongs to the Tripartite motif (TRIM) family which is related to the cancer stem cell self-renewal process. *TRIM6*, more specifically, directly interacts with the *MYC* gene to modulate stem cell differentiation (Jaworska et al., 2019). Attention weights of TF were relatively higher than the weights of ME. Besides the TF+ME+HM model, the TF+HM only model predicted the activation of genes correctly. Therefore, it is thought that HM and TF co-regulate the expression of *TRIM6*.

Lastly, *CDKN1A*, also known as p21, is a kind of tumor suppressor gene. *CDKN1A* plays a crucial role in regulating cell cycles to prevent cancer progression. In **Figure 8**, histone and methylation were relatively highlighted. Based on this observation, we thought that methylation might be a key factor in epigenetic regulation of *CDKN1A*. In addition to the TF+ME+HM model, the ME+HM model also predicted correctly. From previous studies, expression of *DNMT1* and *CDKN1A* showed a negative regulation mechanism on chronic myelogenous leukemia (Kaufman-Szymczyk et al., 2019). It was also reported that *DNMT3B* knock-down induced up-regulation of a number of tumor suppressor genes including *CDKN1A* (Poole et al., 2017).

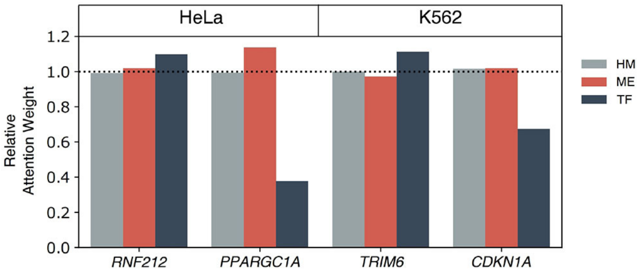


FIGURE 8 | The epigenetic mechanisms of highly expressed genes in the HeLa and K562 cell lines. Among correctly predicted highly expressed genes by the TF+ME+HM model, four genes are selected for a further case study of epigenetic gene regulation mechanism: *RNF212* and *PPARGC1A* for HeLa, and *TRIM6* and *CDKN1A* for the K562 cell line. To elucidate the importance of each omics on the gene regulation, the relative attention weights of the Multi-Attention Block were used.

Based on the case study of enriched genes of the HeLa and K562 cell line, we could investigate the epigenetic regulatory mechanisms of gene expressions by the weights of the Multi-Attention Blocks. Although it was possible to infer the involvement of histone marks, DNA methylation, and transcription factor for each gene, and to analyze their importance, there are other epigenetic and transcriptional factors that regulate gene expression. MicroRNA (miRNA) is one of the famous epigenetic factors that was not included in the model. MiRNAs are actually genes that are controlled by epigenetic mechanisms and TFs. For example, EWS is known to regulate Drosha, which controls biogenesis of miRNA (Kim et al., 2014). miRNA can then affect the transcription and translation of genes. To study the effects of miRNAs, we collected the genes that were correctly predicted by the TF+ME+HM model, not by the HM model in the HeLa cell line and 275 genes were selected as candidate genes. Using a biomedical literature search platform, BEST (Lee et al., 2016), 14 genes were related to miRNA in the context of the HeLa cell line, cervical cancer, or ovarian cancer (**Supplementary Table 3**). For example, the expression of LPAR2 was repressed by miR-377, and oncogenic processes such as cell proliferation or migration are known to be repressed by that inhibition mechanism (Zhang et al., 2020). As another example, ITGB1 was targeted by miR-183. It is known that miR-183 may play a role in tumor suppressors, such as the inhibition of cell invasion or the decrease of migration capacities of HeLa cells (Li et al., 2010). Incorporation of miRNA in our deep learning model can certainly be helpful in understanding complex gene regulation mechanisms. We plan to investigate how roles of miRNA can be seamlessly integrated into our deep learning model.

4. CONCLUSION

In summary, the proposed model learned cell-type-specific gene regulation mechanisms through Multi-Attention based deep learning strategies. To the best of our knowledge, the model is the first of its kind to use multiple epigenetic and transcriptional markers for predicting gene expressions. Our model achieved higher prediction accuracy than the state-of-the-art model. Additionally, the proposed method provided useful insight into cell-type-specific gene regulation mechanisms. Specifically, the weights of the Multi-Attention Block revealed the relative importance of each marker in the specific cell line. Lastly, we identified the mechanism of enriched genes in HeLa and K562 cell lines.

Our model investigated the roles of three markers: histone marks, DNA methylation, and transcription factors. However, the gene regulatory network may also involve additional epigenetic and transcriptional markers such as microRNA, competing endogenous RNA, or long non-coding RNA. Thus, future studies on other epigenetic markers need to be conducted.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://egg2.wustl.edu/roadmap/data/byDataType/rna/expression/, https://egg2.wustl.edu/roadmap/data/byDataType/dnamethylation/, and https://egg2.wustl.edu/roadmap/data/byFileType/alignments/consolidated/.

AUTHOR CONTRIBUTIONS

SK conceived the experiment. MK and SL conducted the experiment and drafted the manuscript. MK, SL, and DL processed data and analyzed results. All authors read and approved the final manuscript.

FUNDING

This research is supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (No.NRF-2017M3C4A7065887), the Collaborative Genome Program for Fostering New Post-Genome Industry of the NRF funded by the Ministry of Science and ICT (No.NRF2014M3C9A3063541), the Bio & Medical Technology Development Program of the NRF funded by the Ministry of Science & ICT (NRF-2019M3E5D307337511), the Bio & Medical Technology Development Program of the NRF funded by the Ministry of Science & ICT (NRF-2019M3E5D4065965), and a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI15C3224).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.00869/full#supplementary-material

REFERENCES

Cedar, H., and Bergman, Y. (2009). Linking dna methylation and histone modification: patterns and paradigms. *Nat. Rev. Genet.* 10, 295–304. doi: 10.1038/nrg2540

Fatemi, M., and Wade, P. A. (2006). Mbd family proteins: reading the epigenetic code. *J. Cell Sci.* 119, 3033–3037. doi: 10.1242/jcs.03099

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: continual prediction with lstm. *Neural Comput.* 12, 2451–2471. doi: 10.1162/089976600300015015

Gill, J., and La Merrill, M. A. (2017). An emerging role for epigenetic regulation of pgc-1α expression in environmentally stimulated brown adipose thermogenesis. *Environ. Epigenet.* 3:dvx009. doi: 10.1093/eep/dvx009

Gillberg, L., Jacobsen, S., Ribel-Madsen, R., Gjesing, A. P., Boesgaard, T. W., Ling, C.,et al. (2013). Does DNA methylation of ppargc1a influence insulin action in first degree relatives of patients with type 2 diabetes? *PLoS ONE* 8:e58384. doi: 10.1371/journal.pone.0058384

Gillberg, L., Jacobsen, S. C., Rönn, T., Brøns, C., and Vaag, A. (2014). Ppargc1a dna methylation in subcutaneous adipose tissue in low birth weight subjects impact of 5 days of high-fat overfeeding. *Metabolism* 63, 263–271. doi: 10.1016/j.metabol.2013.10.003

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004

Jaworska, A., Wlodarczyk, N., Mackiewicz, A., and Czerwinska, P. (2019). The role of trim family proteins in the regulation of cancer stem cell self-renewal. *Stem Cells* 38, 165–173. doi: 10.1002/stem.3109

Jones, A. W., Yao, Z., Vicencio, J. M., Karkucinska-Wieckowska, A., and Szabadkai, G. (2012). Pgc-1 family coactivators and cell fate: roles in cancer, neurodegeneration, cardiovascular disease and retrograde mitochondria–nucleus signalling. *Mitochondrion* 12, 86–99. doi: 10.1016/j.mito.2011.09.009

Kaufman-Szymczyk, A., Majda, K., Szuławska-Mroczek, A., Fabianowska-Majewska, K., and Lubecka, K. (2019). Clofarabine-phytochemical combination exposures in cml cells inhibit dna methylation machinery, upregulate tumor suppressor genes and promote caspase-dependent apoptosis. *Mol. Med. Rep.* 20, 3597–3608. doi: 10.3892/mmr.2019.10619

Kim, K. Y., Hwang, Y. J., Jung, M.-K., Choe, J., Kim, Y., Kim, S., et al. (2014). A multifunctional protein ews regulates the expression of drosha and micrornas. *Cell Death Diff.* 21, 136–145. doi: 10.1038/cdd.2013.144

Kulis, M., and Esteller, M. (2010). DNA methylation and cancer. *Adv. Genet.* 70, 27–56. doi: 10.1016/B978-0-12-380866-0.60002-2

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248

Lai, C.-Q., Tucker, K. L., Parnell, L. D., Adiconis, X., García-Bailo, B., Griffith, J., et al. (2008). Ppargc1a variation associated with dna damage, diabetes, and cardiovascular diseases: the boston puerto rican health study. *Diabetes* 57, 809–816. doi: 10.2337/db07-1238

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., et al. (2018). The human transcription factors. *Cell* 172, 650–665. doi: 10.1016/j.cell.2018.01.029

Lee, S., Kim, D., Lee, K., Choi, J., Kim, S., Jeon, M., et al. (2016). Best: next-generation biomedical entity search tool for knowledge discovery from biomedical literature. *PLoS ONE* 11:e0164680. doi: 10.1371/journal.pone.0164680

Li, G., Luna, C., Qiu, J., Epstein, D. L., and Gonzalez, P. (2010). Targeting of integrin $\beta 1$ and kinesin 2α by microrna 183. *J. Biol. Chem.* 285, 5461–5471. doi: 10.1074/jbc.M109.037127

Li, Y.-Q. (2010). Master stem cell transcription factors and signaling regulation. *Cell. Reprogramming* 12, 3–13. doi: 10.1089/cell.2009.0033

Lin, J., Handschin, C., and Spiegelman, B. M. (2005). Metabolic control through the pgc-1 family of transcription coactivators. *Cell Metab.* 1, 361–370. doi: 10.1016/j.cmet.2005.05.004

Min, S., Lee, B., and Yoon, S. (2016). Deep learning in bioinformatics. *Brief. Bioinform.* 18, 851–869. doi: 10.1093/bib/bbw068

Pan, G. J., Chang, Z. Y., Schöler, H. R., and Duanqing, P. (2002). Stem cell pluripotency and transcription factor oct4. *Cell Res.* 12, 321–329. doi: 10.1038/sj.cr.7290134

Poole, C. J., Zheng, W., Lodh, A., Yevtodiyenko, A., Liefwalker, D., Li, H., et al. (2017). Dnmt3b overexpression contributes to aberrant dna methylation and myc-driven tumor maintenance in t-all and burkitt's lymphoma. *Oncotarget* 8:76898. doi: 10.18632/oncotarget.20176

Qiao, H., Rao, H. P., Yun, Y., Sandhu, S., Fong, J. H., Sapre, M., et al. (2018). Impeding dna break repair enables oocyte quality control. *Mol. Cell* 72, 211–221. doi: 10.1016/j.molcel.2018.08.031

Singh, R., Lanchantin, J., Robins, G., and Qi, Y. (2016). Deepchrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 32, i639–i648. doi: 10.1093/bioinformatics/btw427

Singh, R., Lanchantin, J., Sekhon, A., and Qi, Y. (2017). "Attend and predict: Understanding gene regulation by selective attention on chromatin," in *Advances in Neural Information Processing Systems*, 6785–6795.

Uhlen, M., Zhang, C., Lee, S., Sjöstedt, E., Fagerberg, L., Bidkhori, G., et al. (2017). A pathology atlas of the human cancer transcriptome. *Science* 357:eaan2507. doi: 10.1126/science.aan2507

Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014). The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.* 15:R37. doi: 10.1186/gb-2014-15-2-r37

Wajed, S. A., Laird, P. W., and DeMeester, T. R. (2001). Dna methylation: an alternative pathway to cancer. *Ann. Surg.* 234:10. doi: 10.1097/00000658-200107000-00003

Xie, R., Wen, J., Quitadamo, A., Cheng, J., and Shi, X. (2017). A deep auto-encoder model for gene expression prediction. *BMC Genomics* 18:845. doi: 10.1186/s12864-017-4226-0

Xiong, S., Patrushev, N., Forouzandeh, F., Hilenski, L., and Alexander, R. W. (2015). Pgc-1α modulates telomere function and dna damage in protecting against aging-related chronic diseases. *Cell Rep.* 12, 1391–1399. doi: 10.1016/j.celrep.2015.07.047

Zadeh, A., Liang, P. P., Poria, S., Vij, P., Cambria, E., and Morency, L.-P. (2018). "Multi-attention recurrent network for human communication comprehension," in *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zeng, P., Zhou, X., and Huang, S. (2017). Prediction of gene expression with cis-snps using mixed models and regularization methods. *BMC Genomics* 18:368. doi: 10.1186/s12864-017-3759-6

Zhang, T., Cooper, S., and Brockdorff, N. (2015). The interplay of histone modifications–writers that read. *EMBO Rep.* 16, 1467–1481. doi: 10.15252/embr.201540945

Zhang, Y., Liu, Y., Guo, X., Hu, Z., and Shi, H. (2020). Interfering human papillomavirus e6/e7 oncogenes in cervical cancer cells inhibits the angiogenesis of vascular endothelial cells via increasing mir-377 in cervical cancer cell-derived microvesicles. *Oncotargets Ther.* 13:4145. doi: 10.2147/OTT.S239979

Zhong, H., Kim, S., Zhi, D., and Cui, X. (2019). Predicting gene expression using DNA methylation in three human populations. *PeerJ* 7:e6757. doi: 10.7717/peerj.6757

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.

METHODS
published: 27 October 2020
doi: 10.3389/fgene.2020.538492

MGMIN: A Normalization Method for Correcting Probe Design Bias in Illumina Infinium HumanMethylation450 BeadChips

*Zhenxing Wang, Yongzhuang Liu and Yadong Wang**

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

The Illumina Infinium HumanMethylation450 Beadchips have been widely utilized in epigenome-wide association studies (EWAS). However, the existing two types of probes (type I and type II), with the distribution of measurements of probes and dynamic range different, may bias downstream analyses. Here, we propose a method, MGMIN (*M*-values Gaussian-MIxture Normalization), to correct the probe designs based on *M*-values of DNA methylation. Our strategy includes fitting Gaussian mixture distributions to type I and type II probes separately, a transformation of *M*-values into quantiles and finally a dilation transformation based on *M*-values of DNA methylation to maintain the continuity of the data. Our method is validated on several public datasets on reducing probe design bias, reducing the technical variation and improving the ability to find biologically differential methylation signals. The results show that MGMIN achieves competitive performances compared to BMIQ which is a well-known normalization method on β-values of DNA methylation.

Keywords: DNA methylation, design bias, normalization, M-value, Gaussian mixture model, Illumina Infinium 450K

OPEN ACCESS

Edited by:
Zhongyu Wei,
Fudan University, China

Reviewed by:
Teng Zhixia,
Northeast Forestry University, China
Zhen Tian,
Zhengzhou University, China

****Correspondence:***
Yadong Wang
ydwang@hit.edu.cn

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

Received: *27 February 2020*
Accepted: *21 September 2020*
Published: *27 October 2020*

Citation:
Wang Z, Liu Y and Wang Y (2020) MGMIN: A Normalization Method for Correcting Probe Design Bias in Illumina Infinium HumanMethylation450 BeadChips. Front. Genet. 11:538492. doi: 10.3389/fgene.2020.538492

1. INTRODUCTION

DNA methylation, as a well-known epigenetic marker, plays an essential role in biological processes and complex genetic diseases like cancer and diabetes (Irizarry et al., 2009; Paul et al., 2016). The Illumina Infinium HumanMethylation450 (450K) BeadChip (Bibikova et al., 2011) provides measurements of the level of methylation at over 480K CpG sites and has been widely used in epigenome-wide association studies (EWAS) and large-scale projects, such as The Cancer Genome Atlas (TCGA). The probes in the Infinium 450K BeadChip come in two different designs, type I (n = 135,501) and type II (n = 350,076), in order to increase the genomic coverage of the assay. However, the methylation values (β-values or M-values) derived from the two types of designs exhibit different distributions. Particularly, the type I probes possess a larger range of measurement than the type II probes (Dedeurwaerder et al., 2011). The differences between the two types of probe designs may impact the downstream analyses.

Several approaches have been published to correct the probe design bias. A peak-based correction (PBC) method normalizes type II probes to render them comparable with type I probes (Dedeurwaerder et al., 2011). In fact, PBC gets poor performance when the density distribution of methylation values does not show well-defined peaks. SQN (Touleimat and Tost, 2012) and SWAN (Maksimovic et al., 2012) select subset of probes with similar biological category to adjust the probe design bias. Beta MIxture Quantile dilation (BMIQ) is a model-based normalization approach to correct β-values of type II probes according to the beta distribution of β-values of type I probes, which appears to outperform PBC, SQN, and SWAN (Teschendorff et al., 2012).

In this work, we propose a method to correct the probe design bias based on the Gaussian Mixture Model (GMM) of the M-values of DNA methylation, which is called M-value Gaussian-MIxture Normalization (MGMIN). The method includes three steps: (i) fit Gaussian-mixture distributions to type I and type II probes separately, (ii) utilize a transformation of M-values into quantiles, (iii) perform a dilation transformation based on M-values to maintain the continuity of the data. We evaluate MGMIN using several independent datasets in terms of reducing the replicate technical variance and correcting the type II bias. By comparison with BMIQ, the results show that MGMIN improves the overall performance of normalization.

TABLE 1 | Comparison of MGMIN and BMIQ on detecting the differentially methylated probes (DMPs) associated with HPV status was performed by counting the number of DMPs (Dataset 2), the number of validated differentially methylated probes (nTPs) (Dataset 3: GSE38266 and Dataset 4: GSE95036) and corresponding estimates for the positive predictive value (PPV = nTP/nDMPs).

Metric	Raw	BMIQ	MGMIN
nDMP	51 (51[a])	239 (252[a])	220
nTP (GSE38266)	16 (13[a])	55 (51[a])	37
PPV (GSE38266)	0.31 (0.25[a])	0.23 (0.20[a])	0.17
nTP (GSE95036)	3	13	27
PPV (GSE95036)	0.06	0.05	0.12

[a] *Values reported in* Teschendorff et al. (2012).

2. MATERIALS AND METHODS

2.1. Measure DNA Methylation With *M*-value

The β-value of DNA methylation for each probe is defined by the ratio of the methylated intensity (M) and the overall intensity (sum of methylated intensity and unmethylated intensity: M + U):

$$\beta - value = \frac{M}{M + U + \alpha}$$

where α is a constant offset (by default, α = 100) to regularize the β-value when the overall intensity is low. The β-value falls between 0 and 1 which follows a Beta distribution naturally. A β-value of 0 indicates the CpG site of the measured sample is fully unmethylated and a value of 1 indicates that the CpG site is completely methylated.

The M-value is calculated by the log2 ratio of the methylated intensity (M) vs. the unmethylated intensity (U):

$$M - value = \log_2(\frac{M + \alpha}{U + \alpha})$$

where α here is also an offset (by default, α = 1) to counteract the big changes caused by small intensity estimation errors. An M-value close to zero indicates that the measured CpG site is about hemimethylated. A positive M-value suggests that more copies of the measured CpG site are methylated than unmethylated and a negative M-value means more copies of the CpG site are unmethylated. The M-value has been widely used in two-color expression microarray analysis (Du et al., 2010).

Due to more than 95% CpG sites have intensities more than 1,000 in Illumina methylation data, the α in β-value and M-value has an insignificant effect on observed results. So the relationship between β-value and M-value is shown as (with α ignored):

$$\beta = \frac{2^M}{2^M + 1}; M = \log_2(\frac{\beta}{1 - \beta})$$

According to the conclusions in Du et al. (2010), the M-value is more statistically valid in an analysis by modeling the distribution of M-values because of it's *homoscedastic*. So we choose to adjust the M-values of type II probes into the distribution property of type I probes to correct the probe design bias.

2.2. MGMIN: *M*-value Gaussian-MIxture Normalization

Gaussian Mixture Model (GMM) has been widely applied as a clustering method in analyzing gene-expression microarray data (Yeung et al., 2001; Pan et al., 2002) and used to detect differential gene expression (McLachlan et al., 2006). In this paper, we apply GMM to distinguish different methylation states of CpG sites for further correction. The M-values of a single 450K microarray can be viewed as a finite Gaussian mixture model of several methylation states (hypomethylated-U, hemimethylated-H, hypermethylated-F). The probability density function of the M-value for a single CpG site (M_i) is defined as:

$$p(M_i; \theta) = \sum_{k=1}^{K} \pi_k N(M_i|\mu_k, \sigma_k^2) \tag{1}$$

where $p(M_i, \theta)$ represents the model density for M_i with unknown parameter vector θ, K is the number of different methylation states (components), $N(M_i|\mu_k, \sigma_k^2)$ is the probability density function of the kth Gaussian component, and π_k is the mixing proportions which satisfy the constraint that $\sum_{k=1}^{K} \pi_k = 1$ and $0 \leq \pi_k \leq 1$. The parameter vector θ consists of the mixing proportions π_k, the mean value μ_k and the standard deviation σ_k, which can be estimated by the EM algorithm.

Next, we describe MGMIN in detail. First, M-values of type I and type II probes are modeled by GMM separately. Let μ_T^S and σ_T^S denote the mean value and standard deviation where $S \in (U, H, F)$ and $T \in (I, II)$. K_I and K_{II} are the numbers of components for type I and type II probes, which are both set as 3 by default.

Second, each probe is assigned to hypomethylated (U_T), hemimethylated (H_T), or hypermethylated (F_T) states by using the maximum probability criterion. Let U_T^L (U_T^R) denote the U_T probes with M-values smaller (larger) than μ_T^U, and let F_T^L (F_T^R) represent the F_T probes with M-values smaller (larger) than μ_T^F where $T \in (I, II)$. Then, we calculate the probabilities of

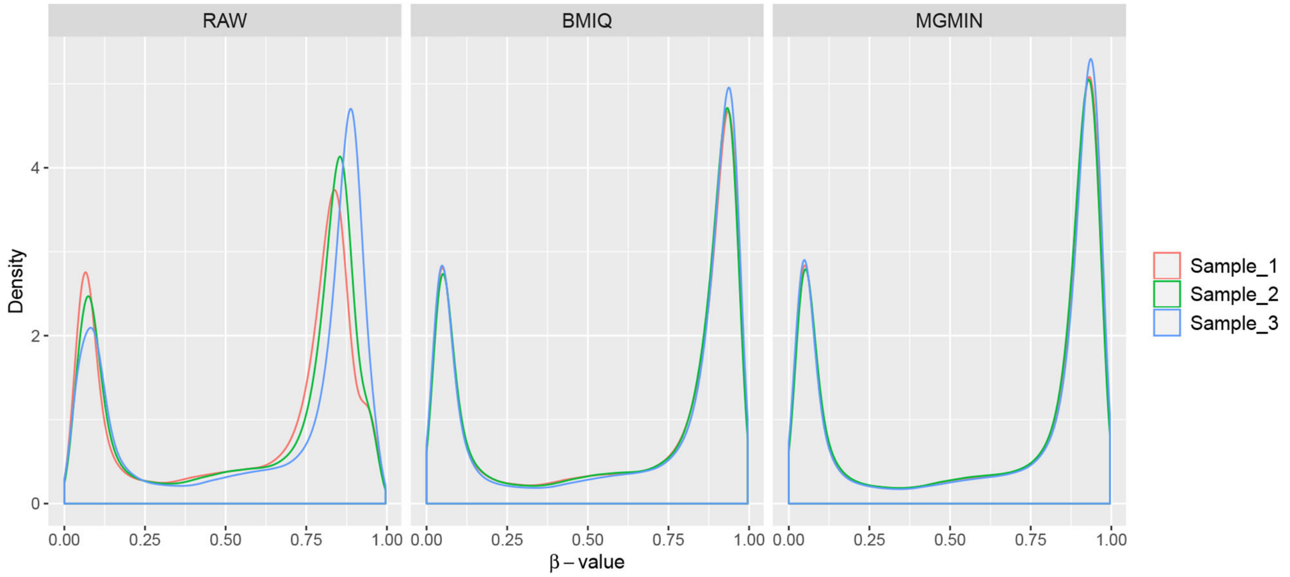


FIGURE 1 | The density curves of β-values for the three replicates in Dataset 1. The left panel is for the case of raw data with no normalization, middle panel for BMIQ and right panel for MGMIN.

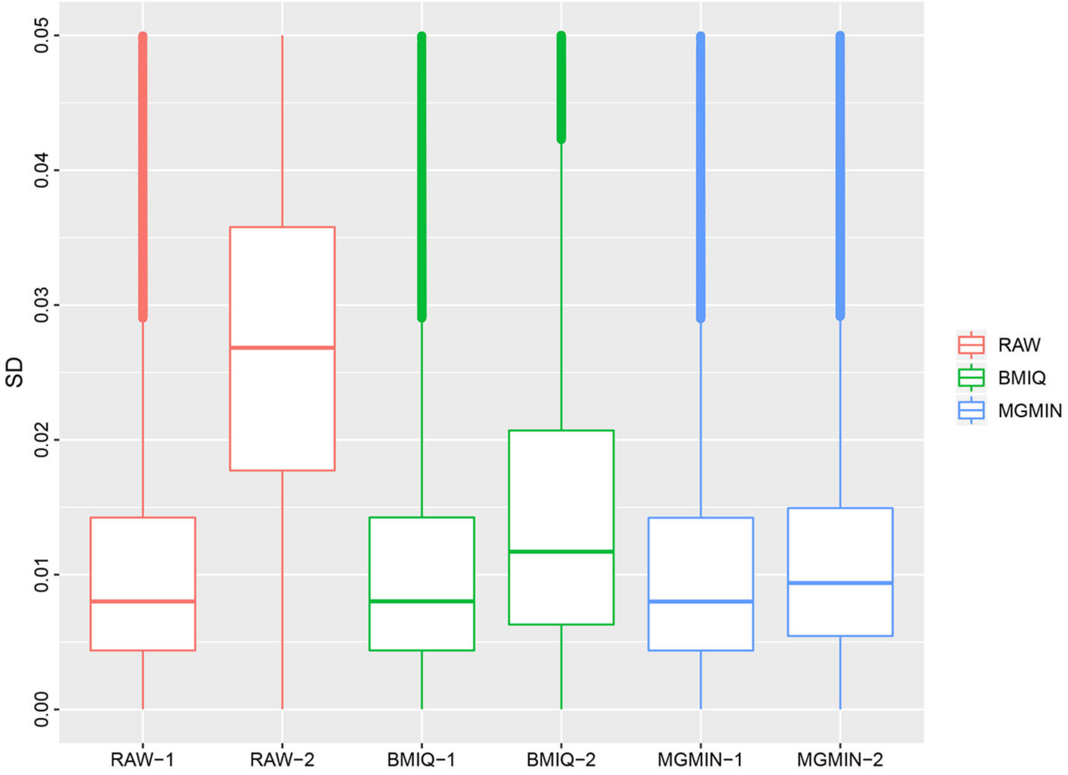


FIGURE 2 | Boxplots of the standard deviations of β-values for the three replicates in Dataset 1, for raw β-values (RAW), normalized β-values by BMIQ (BMIQ), and normalized β-values by MGMIN (MGMIN). RAW-1 represents the type I of raw values and RAW-2 represents the type II of raw values, and so on.

U_{II}^{L} probes, i.e.,

$$p = P(M_{U_{II}^{L}} | \mu_{II}^{U}, (\sigma_{II}^{U})^2) \qquad (2)$$

where P represents the cumulative distribution function of the Gaussian component. These probabilities are transformed back to quantiles (M-value) by using the parameters μ_{I}^{U} and σ_{I}^{U} of

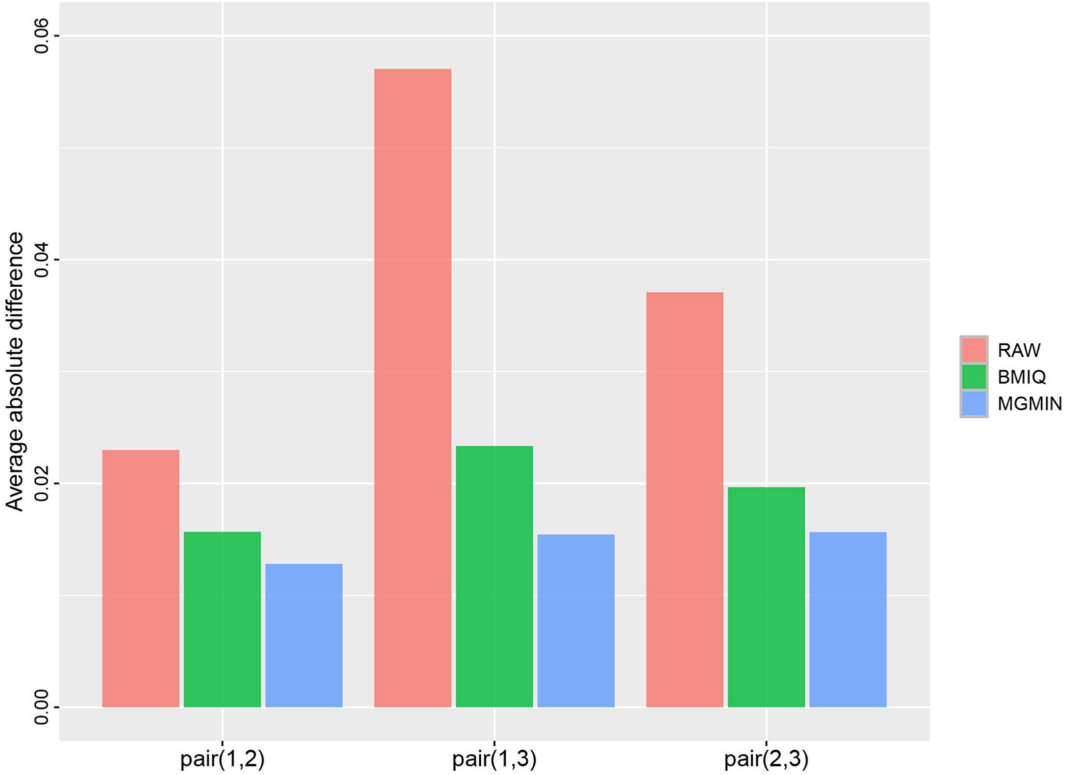


FIGURE 3 | Barplots of the average absolute difference in β-values of type II probes between two samples in each of the three pairs of the three replicates in Dataset 1.

type I probes, i.e.,

$$q = P^{-1}(p|\mu_I^U, (\sigma_I^U)^2) \tag{3}$$

where P^{-1} returns the value of the inverse cumulative density function given the probability p and q is the normalized M-values for U_{II}^L. The similar operation is performed on F_{II}^R probes.

Then, we merge the U_{II}^R, H_{II}, and F_{II}^L probes into one set G on which a conformal (shift + dilation) transformation is performed. Some parameters are identified as $minG = \min\{M_G\}$, $maxG = \max\{M_G\}$ and $\Delta_G^M = maxG - minG$. Similarly, the minimum value of F_{II}^R and the maximum value of U_{II}^L are also identified, i.e., $minF = \min\{F_{II}^R\}$ and $maxU = \max\{U_{II}^L\}$. Two distance values can be calculated as

$$\Delta_{UG} = minG - maxU$$

$$\Delta_{GF} = minF - maxG$$

The new normalized maximum and minimum values of G-probes are expected to satisfy the constraint that

$$maxG' = \min\{F_{II}^{R\prime}\} - \Delta_{GF}$$

$$minG' = \max\{U_{II}^{L\prime}\} + \Delta_{UG}$$

where $F_{II}^{R\prime}$ and $U_{II}^{L\prime}$ are new normalized values for F_{II}^R and U_{II}^L, respectively. So the new normalized range value of set G is $\Delta_G^{M\prime} = maxG' - minG'$. The normalized M-values of set G, $M_{G_{II}}{'}$, is calculated by

$$M_{G_{II}}{'} = minG' + d_f(M_{G_{II}} - minG) \tag{4}$$

where $d_f = \Delta_G^{M\prime}/\Delta_G^M$ is the dilation factor. So, the normalized M-values for type II probes consist of q for U_{II}^L, $M_{G_{II}}{'}$, and q for F_{II}^R.

$$M_{II}{'} = (q_{U_{II}^L}, M_{G_{II}}{'}, q_{F_{II}^R})$$

Lastly, the normalized M-values are transformed to β-values.

There are some important points to notice: (i) the initial values for μ and σ in EM algorithm are set as (−4,0,4) and (1,1,1) and small perturbations to the initial μ and σ will not affect the final model because MGMIN captures the natural property of the M-value of DNA methylation, (ii) K_I will be changed to 4 automatically when $\mu_I^F - \sigma_I^F$ is smaller than $\mu_{II}^F - \sigma_{II}^F$ in order to ensure that μ_I^F can always be larger than μ_{II}^F and avoid the presence of an unexpected peak in transformed M-values of hypermethylated type II probes, (iii) if $K_I = 4$, the F_I will be the set of probes belonging to the component with the largest μ, while the U_I contains the probes belonging to the component with the smallest μ and the other two components are assigned

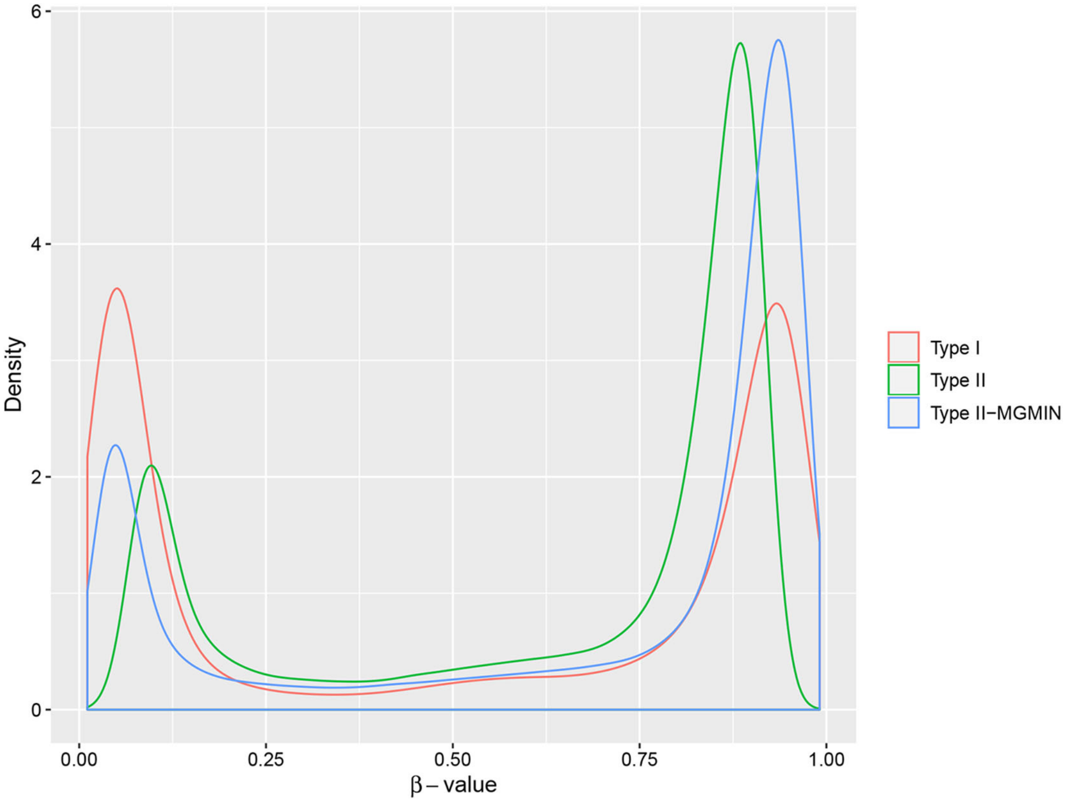


FIGURE 4 | The density curves of β-values for type I probes, type II probes and normalized type II probes (type II-MGMIN) for sample GSM815138 from GEO29290.

to H_I, (iv) no thresholds need to be set by default or estimated by manual to distinguish the three different states of DNA methylation.

2.3. Datasets

We selected several public 450K datasets as following:

Dataset 1: GSE29290 downloaded from GEO considered in Dedeurwaerder et al. (2011). We used the three replicates (GSM15136, GSM15137 and GSM15138) from the HCT116WT cell-line and matched bisulfite pyrosequencing (BPS) date for nine type II probes of sample GSM815138 (r3) (Table 1 in Dedeurwaerder et al., 2011) to evaluate the performance of different methods.

Dataset 2: GSE38268 downloaded from GEO considered in Lechner et al. (2013) which consists of 6 fresh frozen HNC samples. We selected 5 samples as same as (Teschendorff et al., 2012), of which 2 were HPV+ and 3 HPV− (GSM937820 to GSM937824).

Dataset 3: GSE38266 downloaded from GEO considered in Lechner et al. (2013) which contains 21 FFPE HPV+ HNSCC samples and 21 FFPE HPV− HNSCC samples. Note that the entire quality of the dataset GSE38266 is not high.

Dataset 4: GSE95036 downloaded from GEO considered in Degli Esposti et al. (2017) which contains 6 HPV+ HNC samples and 5 HPV− HNC samples.

3. RESULTS

3.1. MGMIN Needs No Default Initial Values of Parameters

Similar to the mixture model of BMIQ, MGMIN applies Gaussian mixture models for *M*-values instead of beta-mixture models for β-values. MGMIN also uses quantile information to correct the *M*-values of the type II probes into a distribution which is comparable with that of type I probes. MGMIN complies the inherent Gaussian mixture distributions for *M*-values of type I and type II probes to avoid setting any parameters manually, which is different from the default breakpoints in BMIQ. Thus, MGMIN needs less manual intervention than BMIQ. However, MGMIN is slightly inferior to BMIQ on some dataset (**Table 1**) due to the entire low quality of the dataset. Note that the PPV of BMIQ on Dataset 3 is lower than that of no normalization (RAW).

3.2. MGMIN Reduces Technical Variation

MGMIN is applied to Dataset 1 to identify the ability to improve reproducibility. The standard deviation (SD) for each probe across the three replicates was computed using no normalization (RAW), BMIQ, and MGMIN separately. As can be seen in **Figure 1**, both MGMIN and BMIQ almost made the density curves for the three replicates coincide with each other and reduced the technical variation significantly

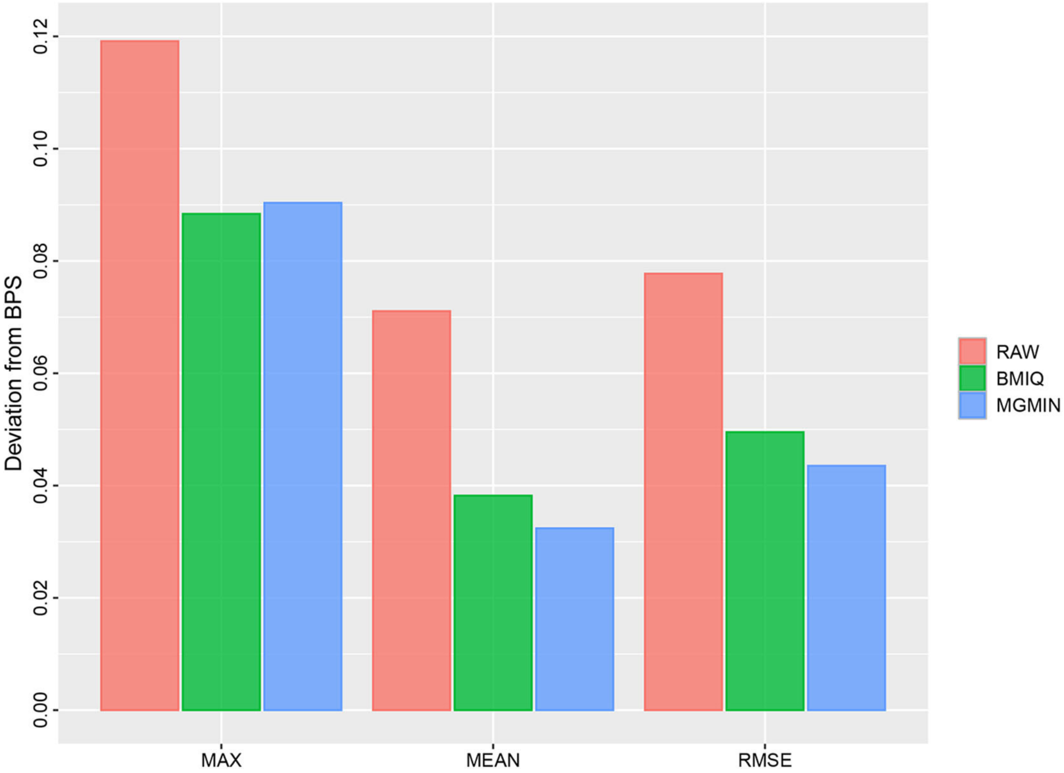


FIGURE 5 | Barplots for the maximum (MAX), mean (MEAN) and root mean square error (RMSE) of the absolute deviation from the matched BPS values of nine type II probes for GSM815138 (r3) in Dataset 1 considered in Dedeurwaerder et al. (2011) using no normalization (RAW), BMIQ, and MGMIN, respectively.

compared to no normalization. Compared to BMIQ, the standard deviation for type II probes adjusted by MGMIN is smaller (**Figure 2**). MGMIN also provided significant reduction of average absolute difference in β-values of type II probes between two samples in each of the three pairs of the three replicates (**Figure 3**).

3.3. MGMIN Reduces Probe Design Bias

MGMIN reduces the probe design bias via correcting the M-values of the type II probes such that the distribution curves for the M-values of the type I and type II probes show similar dynamic ranges and peaks (**Figure 4**). In Dedeurwaerder et al. (2011), the β-values for nine probes of type II by bisulfite pyrosequencing technique for sample GSM815138 (r3) were provided, which can be used as a gold-standard to evaluate the performance of different correction methods. Hence, we compared the normalized results of the nine type II probes in 450K arrays by MGMIN and BMIQ. As shown in **Figure 5**, although MGMIN performed slightly worse than BMIQ at the maximum value of the absolute deviation from BPS data, MGMIN significantly reduced the type II bias than BMIQ and raw data in terms of mean and root mean square error (RMSE) of the absolute deviation from the matched BPS values.

3.4. MGMIN Robustly Finds Informative Differential Methylation Probes Associated With HPV Status

The goal of a bias correction approach is to reduce the technical variation and identify the biological informative signals at the same time. We used a strategy similar to Teschendorff et al. (2012) to compare the result between MGMIN and BMIQ in identifying the differential methylation probes (DMPs) associated with HPV status. First, Dataset 2 consisting of two HPV+ and three HPV− fresh frozen HNC samples were used as the training set to obtain the DMPs associated with HPV status by the *limma* method (Smyth, 2005) and an FDR threshold 0.35 which was as same as (Teschendorff et al., 2012). Both Dataset 3 and Dataset 4 described in the methods section were used as test set. We reanalyzed Dataset 2 and got similar numbers of DMPs to those reported in Teschendorff et al. (2012) with no normalization method (Raw) or BMIQ method (shown in **Table 1**). The results in **Table 1** shows that the positive predictive value (PPV) of MGMIN is slightly less than BMIQ in terms of GSE38266 (Dataset 3) whereas MGMIN outperforms BMIQ in GSE95036 (Dataset 4). The reason for MGMIN slightly inferior to BMIQ in Dataset 3 may be the entire low quality of the dataset (see **Figure 6**) which is that the ratio of samples passing filters is <0.9 (r = 0.88) under the least restrictive condition. Let τ_p represent the p-value threshold for bad probes and τ_r

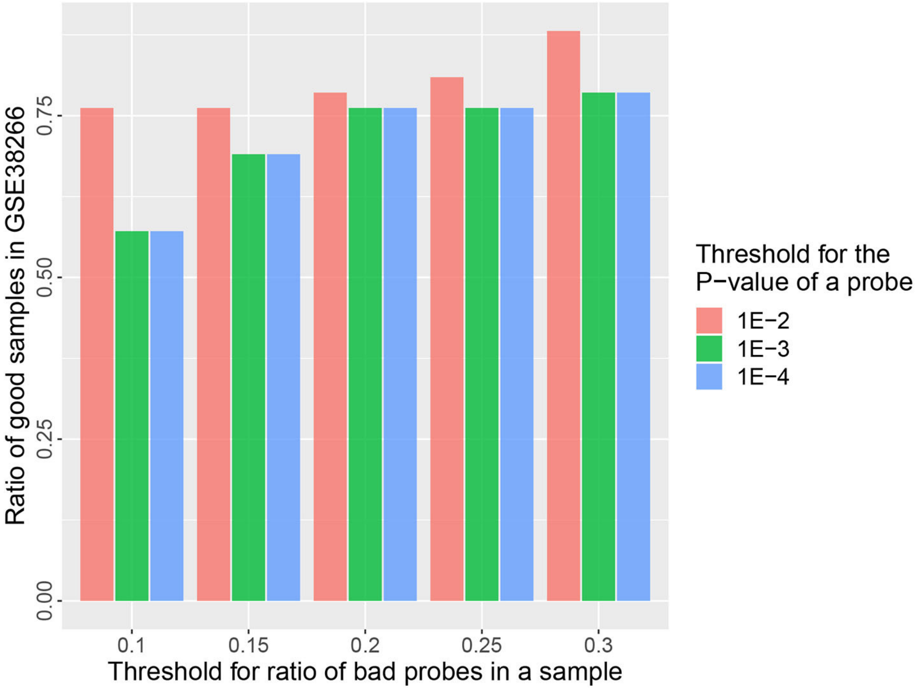


FIGURE 6 | Barplots of the ratio of good samples in GSE38266 under different quality control options (τ_p&τ_r).

represent the threshold for the ratio of bad probes in a sample. The maximum value of τ_r is set to 0.3 here in our opinion because a sample with more than 30% bad probes is vulnerable. We can get the same test dataset from GSE38266 with the one described in Teschendorff et al. (2012) which consists of 18 HPV+ and 14 HPV− samples under the following conditions: (i) $\tau_p = 1e - 4$ or $1e - 3$ and $\tau_r = 0.2$ or 0.25, (ii) $\tau_p = 1e - 2$ and $\tau_r = 0.1$ or 0.15. Overall, MGMIN identified more true positive features than BMIQ.

4. DISCUSSIONS

In this paper, we have proposed a method called MGMIN for correcting the probe design bias of type II probes in Illumina Infinium 450K BeadChips, which can reduce the technical variation and improve the ability to find biologically differential methylation signals. We have shown that MGMIN outperforms BMIQ on multiple evaluation datasets in correcting the type II design bias and improving the data quality.

Similar to BMIQ, MGMIN uses quantile information to correct the *M*-values of type II probes while leaving the *M*-values of type I probes unchanged. The three-state beta-mixture distribution model in BMIQ sets two default breakpoints (0.2, 0.75) to divide the β-values into three classes: hypomethylated, hemimethylated, and hypermethylated, which works well for most cases. However, the result curves of BMIQ show obviously inconsistent in some samples with high heterogeneity. We set 3 or 4 classes for probes depending on the result of $\mu_T^F - \sigma_T^F$ to ensure that the fitted hypermethylated component of type II probes can be located in the left of the hypermethylated component of type I probes, which can partly eliminate the effects of the heterogeneity of samples.

Based on the results of Dataset 3, we think the high quality of dataset is the base of normalization, in other words, there is no meaning to correct the samples with low quality. It should be pointed out that the parameter estimation of MGMIN is slower than that of BMIQ (about 1.5 times), which can be relieved by reducing the number of iterations.

MGMIN can be used in the 450K methylation data preprocessing with other methods to normalize the values of the two type probes and improve the data quality.

DATA AVAILABILITY STATEMENT

The datasets for this study can be found in GEO: GSE29290, GSE38268, GSE38266, and GSE95036.

AUTHOR CONTRIBUTIONS

ZW performed the experiments and wrote the manuscript. All authors read and revised the final manuscript.

FUNDING

This work has been supported by the National Key Research and Development Program of China (Nos: 2017YFC1201201 and 2017YFC0907503).

REFERENCES

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., et al. (2011). High density DNA methylation array with single CPG site resolution. *Genomics* 98, 288–295. doi: 10.1016/j.ygeno.2011.07.007

Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., and Fuks, F. (2011). Evaluation of the infinium methylation 450K technology. *Epigenomics* 3, 771–784. doi: 10.2217/epi.11.105

Degli Esposti, D., Sklias, A., Lima, S. C., Beghelli-de la Forest Divonne, S., Cahais, V., Fernandez-Jimenez, N., et al. (2017). Unique DNA methylation signature in HPV-positive head and neck squamous cell carcinomas. *Genome Med*. 9:33. doi: 10.1186/s13073-017-0419-z

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., et al. (2010). Comparison of beta-value and *M*-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* 11:587. doi: 10.1186/1471-2105-11-587

Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., et al. (2009). The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific cpg island shores. *Nat. Genet*. 41, 178–186. doi: 10.1038/ng.298

Lechner, M., Fenton, T., West, J., Wilson, G., Feber, A., Henderson, S., et al. (2013). Identification and functional validation of HPV-mediated hypermethylation in head and neck squamous cell carcinoma. *Genome Med*. 5:15. doi: 10.1186/gm419

Maksimovic, J., Gordon, L., and Oshlack, A. (2012). Swan: subset-quantile within array normalization for illumina infinium humanmethylation450 beadchips. *Genome Biol*. 13:R44. doi: 10.1186/gb-2012-13-6-r44

McLachlan, G. J., Bean, R., and Jones, L. B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* 22, 1608–1615. doi: 10.1093/bioinformatics/btl148

Pan, W., Lin, J., and Le, C. T. (2002). Model-based cluster analysis of microarray gene-expression data. *Genome Biol*. 3:research0009-1. doi: 10.1186/gb-2002-3-2-research0009

Paul, D. S., Teschendorff, A. E., Dang, M. A., Lowe, R., Hawa, M. I., Ecker, S., et al. (2016). Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nat. Commun*. 7:13555. doi: 10.1038/ncomms13555

Smyth, G. K. (2005). "Limma: linear models for microarray data," in Bioinformatics sand Computational Biology Solutions Using R and Bioconductor, eds R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit (New York, NY: Springer), 397–420. doi: 10.1007/0-387-29362-0_23

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., et al. (2012). A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450K DNA methylation data. *Bioinformatics* 29, 189–196. doi: 10.1093/bioinformatics/bts680

Touleimat, N., and Tost, J. (2012). Complete pipeline for infinium® human methylation 450K beadchip data processing using subset quantile normalization for accurate dna methylation estimation. *Epigenomics* 4, 325–341. doi: 10.2217/epi.12.21

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977–987. doi: 10.1093/bioinformatics/17.10.977

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

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership