FUNMarker: Fusion network-based method to identify prognostic and heterogeneous breast cancer biomarkers

Xingyi Li[1], Ju Xiang[1,2], Jianxin Wang[1], Jinyan Li[3], Fang-Xiang Wu[4] and Min Li[1, *]

Abstract—Breast cancer is a heterogeneous disease with many clinically distinguishable molecular subtypes each corresponding to a cluster of patients. Identification of prognostic and heterogeneous biomarkers for breast cancer is to detect cluster-specific gene biomarkers which can be used for accurate survival prediction of breast cancer outcomes. In this paper, we proposed a FUsion Network-based method (FUNMarker) to identify prognostic and heterogeneous breast cancer biomarkers by considering the heterogeneity of patients' samples and biological information from multiple sources. To reduce the affect of heterogeneity of patients, samples were first clustered using the K-means algorithm based on the principal components of gene expression. For each cluster, to comprehensively evaluate the influence of genes on breast cancer, genes were weighted from three aspects: biological function, prognostic ability and correlation with known disease genes. Then they were ranked via a label propagation model on a fusion network that combined physical protein interactions from seven types of networks and thus can reduce the impact of incompleteness of interactome. We compared FUNMarker with three state-of-the-art methods and the results show that biomarkers identified by FUNMarker have stronger discriminative power than the existing methods in differentiating patients with different prognostic outcomes and have biological interpretability.

Index Terms—Prognostic and heterogeneous biomarker, Label propagation, Fusion network

1 INTRODUCTION

Breast cancer is a malignant tumor that is caused by multiple factors and is highly heterogeneous with many clinically distinguishable molecular subtypes each corresponding to a cluster of patients [1-3]. Meanwhile, The prognosis of breast cancer with the same pathological type and clinical stage is quite different even after the same treatment, which indicates the high heterogeneity in breast cancer [4-6]. Therefore, the study of the prognosis and heterogeneity of breast cancer has profound significance. With the development of molecular biology, some molecular biomarkers have been found to be associated with the prognosis of breast cancer, which makes it possible to more accurately and effectively evaluate the prognosis of breast cancer patients [7-10].

At present, using biological networks as the analytical means is a useful way to discover prognostic biomarkers, because biomolecules do not perform functions individually, while they function together. Therefore, biological networks contain large amounts of biological information [11-16]. A number of methods identify prognostic biomarkers to predict outcomes of patients using gene expression data and single biological network, such as protein-protein interaction (PPI) networks, transcription factor-target networks or miRNA-target gene networks [17-20].

However, despite the impressive progress in high-throughput technology, existing interactions remain incomplete and false-positives, and single network contains only the single biological information, which prompts us to think about how to use existing network data to minimize the inaccuracy of follow-up analysis caused by these problems. Previous study has shown that integrating several sources of protein interactions can be helpful for disease research, such as mining disease modules [21]. Based on this, we consider that the use of a highly reliable fusion network can also contribute to the identification of biomarkers.

Some biomarker discovery methods are based on networks extracted from certain databases [22, 23], such as the Reactome FI network which includes pathways, PPIs, gene co-expression and Gene Ontology (GO) annotations. However, these methods do not comprehensively analyze the network data from multiple biological perspectives, which may miss some important information.

Meanwhile, the network-based computational methods typically identify prognostic biomarkers based on the ranking of genes in the networks. Liu et al. [24] classify the cancer samples by the directed random walk. Winter et al. [17] adopt the NetRank algorithm, which is similar to PageRank, to predict the outcome of cancer patients. These methods are usually affected by the heterogeneity of cancer samples [25], resulting in the low prediction accuracy. It is well known that the heterogeneity is mainly caused by the genomic instability [26, 27], and the poor prediction is most likely due to the difficulty of identifying prognostic biomarkers for specific cancer samples. Some methods have been proposed to reduce the impact of heterogeneity by analyzing multi-omics data [28, 29]. Some methods cluster cancer samples to eliminate the heterogeneity of samples [23].

In this paper, the survival time regression problem is

- *X. Li, J. Xiang, J. Wang and M. Li are with the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China, 410083. Email: xingyili@csu.edu.cn, xiangju@csu.edu.cn, jxwang@mail.csu.edu.cn, limin@mail.csu.edu.cn.*
- *J. Xiang is also with the Neuroscience Research Center & Department of Basic Medical Sciences, Changsha Medical University, Changsha, Hunan, P.R. China, 410219. Email: xiangju@csu.edu.cn.*
- *J. Li is with the Advanced Analytics Institute, Faculty of Engineering & IT, University of Technology Sydney, Australia. Email: Jinyan.Li@uts.edu.au*
- *F.X. Wu is with the Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SKS7N5A9, Canada. E-mail: faw341@mail.usask.ca.*

converted to a classification problem. Then, a FUsion Network-based method (FUNMarker) is proposed to identify prognostic and heterogeneous biomarkers for accurate survival prediction of breast cancer outcomes and it is different from the global biomarker analysis which focuses on representative genes across all the subtypes of patients. We fist cluster the breast cancer patients through the principal components of gene expression data to minimize the effects of the heterogeneity of breast cancer samples. For each cluster, we adopt a fusion network to identify the prognostic biomarkers of breast cancer, because the fusion network derived from multiple network sources can minimize the impact of incomplete network data and better reveal the molecular mechanism of diseases [21]. Meanwhile, due to the fact that the fusion network contains more sources, the importance of a gene in the network needs to be evaluated from multiple respects combined with a variety of biological information. Therefore, we score genes in the fusion network from the perspective of biological functions, prognostic ability and correlation with known disease-associated genes (DAGs), and a label propagation model is applied to evaluate the importance rankings of genes based on the weighted fusion network. Top genes are used as the prognostic biomarkers. Finally, a random forest classifier is applied to evaluate the outcome of cancer samples according to the gene expression of biomarkers.

Because the effective prognostic biomarkers should have the high discrimination ability to distinguish patients with good prognosis from those with poor prognosis, as well as the high biological interpretability. Therefore, we evaluated identified biomarkers from both the classification accuracy and the functional interpretability. We compare FUNMarker with several network-based biomarker identification methods (such as CPR, NetRank and stSVM) on six datasets. The results show that biomarkers identified by FUNMarker have stronger ability to differentiate patients with different prognostic outcomes compared with other methods. Then, we compare the performance of our proposed framework in every single biological network and the fusion network, and it is verified that the fusion network can help to improve the accuracy of classification on each dataset better than any single network. We also comprehensively evaluate the biological significance of biomarkers found by FUNMarker, such as the reproducibility, the survival analysis, the biological interpretability and dysregulation of biomarkers. All these show the superiority of FUNMarker in identifying prognostic and heterogeneous breast cancer biomarkers. The MATLAB-package for our algorithm is freely available from https://github.com/CSUBioGroup/FUNMarker.

2 MATERIALS AND METHODS

2.1 Dataset and preprocessing

We collected a high-throughput sequencing data from the UCSC Cancer Genomics Browser (https://genome-cancer.ucsc.edu), which is derived from files downloaded from The Cancer Genome Atlas (TCGA) data. The dataset shows the gene-level transcription estimates, as in log2(x+1) transformed RSEM (RNA-seq by expectation maximization) normalized count. The survival information of patients was collected from the TCGA database (https://portal.gdc.cancer.gov/). We also downloaded four gene expression profiles were downloaded from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) [30], namely GSE1456 [31], GSE2034 [32], GSE3494 [33], GSE4922 [34], and 131 breast cancer patients from van de Vijver dataset [35]. For BRCA, NKI, GSE2034, GSE3494 and GSE4922, samples that survived more than ten years were labeled as good prognosis (five years for GSE1456), and samples that survived less than five years were labeled as poor prognosis. Probes corresponding to multiple genes were discarded, and when multiple probes are mapped to the same gene, the median value was used to eliminate the influence of measurement errors. The gene expression data were normalized by the Z-score. Meanwhile, due to the certain number of zero values in the RNA-seq data, we filter out the genes whose number of zero values are more than 10% of the total number of samples so as to reduce the impact of noise. Table 1 shows the details of the gene expression datasets.

The data of human GO annotations was collected from Gene Ontology Consortium [36, 37] (http://www.geneontology.org/) . Known disease-associated genes (DAGs) were downloaded from DisGeNET (http://www.disgenet.org/) [38].

The fusion network contains the following sources: (1) regulatory interactions; (2) binary interactions from several yeast two-hybrid high-throughput and literature-curated datasets; (3) literature-curated interactions derived mostly from low-throughput experiments; (4) metabolic enzyme-coupled interactions; (5) protein complexes; (6) kinase-substrate pairs and (7) signaling interactions. The network data was downloaded from [21] and only physical protein interactions were retained. There are 13,460 proteins and 141,296 interactions in the network.

- Binary interactions: Several yeast-two-hybrid high-throughput datasets [39-43] with binary interactions from IntAct [44] and MINT databases [45] were combined, resulting in 28,653 interactions between 8,120 proteins.
- Literature curated interactions: the interactions obtained by low-throughput experiments from the literature were collected from IntAct, MINT, BioGRID [46] and HPRD datasets [47], resulting in 88,349 interactions between 11,798 proteins.
- Regulatory interactions: The TRANSFAC database [48] preserved the regulatory interactions and there were 271 transcription factors regulating 564 genes via 1,335 interactions.
- Metabolic enzyme-coupled interactions: The metabolic interactions were obtained from Lee, et al. [49] and there are 5,325 such metabolic links between 921 enzymes.
- Protein complexes: the CORUM database [50] was used to obtain protein complexes, resulting in 2,837 complexes with 2,069 proteins connected by 31,276 interactions.
- Kinase-substrate pairs: The network of peptides that can be bound by kinases can be obtained from PhosphositePlus [51] , which included 6,066 interactions between 1,843 kinases and substrates.
- Signaling interactions: The dataset from Vinayagam, et al. [52] listed 32,706 interactions between 6,339 proteins that integrated several sources, both high-throughput and literature curation, into a directed network in which cellular signals were

transmitted by PPIs.

2.2 Prognostic and heterogeneous biomarker identification

The prediction and diagnosis of cancer patients are difficult due to the heterogeneity of samples. Inspired by Choi et al. [23], cancer samples were clustered by using the principal components of gene expression and k-means to minimize the effects of the heterogeneity of samples. We utilized the maximal silhouette score to determine the number of clusters. The principal component analysis (PCA) plots were shown in Fig.1.

TABLE 1 SUMMARY OF THE GENE EXPRESSION DATASETS

Name	Good samples	Poor samples	Total samples	Total genes	Characteristic for label	Reference
BRCA	31	92	123	14160	days_to_death	https://genome-cancer.ucsc.edu
GSE1456	123	22	145	12432	SURV_RELAPSE	[31]
GSE2034	44	93	137	12432	Time to relapse or last follow-up	[32]
GSE3494	123	36	159	12432	Disease-Specific Survival Time	[33]
GSE4922	107	69	176	12432	Disease Free Survival Time	[34]
NKI	83	48	131	10703	TIMEsurvival	[35]

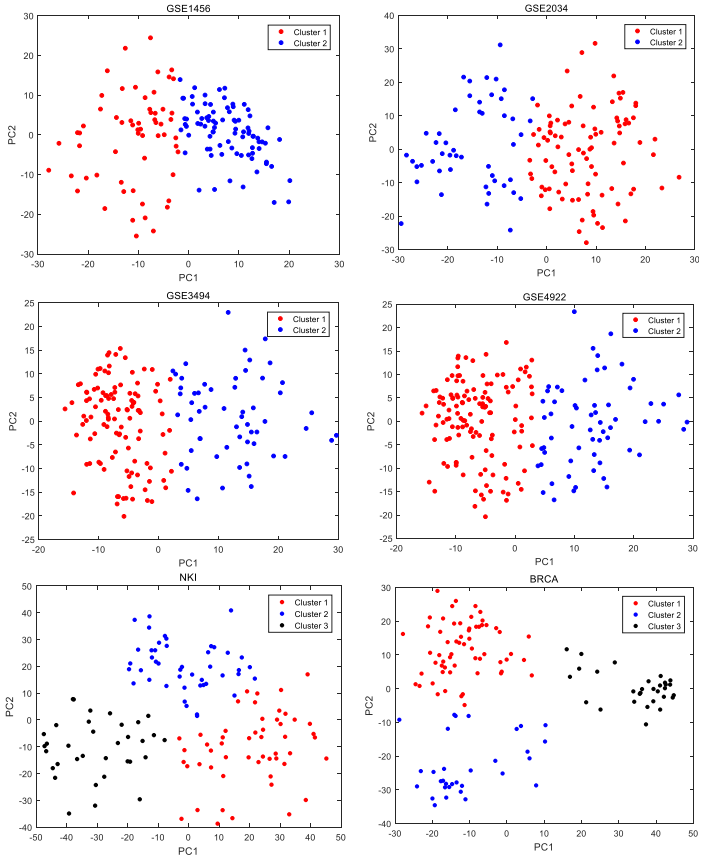

Fig.1 PCA plots. PCA plots using two highest principal components (PC) for each dataset.

To comprehensively evaluate the effect of genes on breast cancer, genes in the network were scored from the perspective of biological functions, prognostic ability and correlation with DAGs for each cluster of samples. The scores of DAGs which represent the relevance of breast cancer were used as the values of these genes in the network. The prognostic ability of a gene was evaluated through its t-score, which is calculated as follows,

$$s_t(i) = \frac{\overline{x_i} - \overline{y_i}}{\sqrt{\frac{s_{x_i}^2}{n} + \frac{s_{y_i}^2}{m}}} \quad (1)$$

where $\overline{x_i}$ and $\overline{y_i}$ are the means of gene i in positive samples and negative samples. s_{x_i} and s_{y_i} are the sample standard deviations of gene i in positive samples and negative samples, respectively. n and m are the number of positive and negative samples, respectively.

The score of biological functions for each gene is defined as the frequency of GO terms containing the gene within the set of all GO terms,

$$s_{GO}(i) = \frac{N_i}{N} \quad (2)$$

where N_i is the number of GO terms contained in gene i, and N denotes the number of all GO terms.

For each scoring strategy, we introduced the label propagation model to rank genes in the fusion network with the initial weight of gene nodes. With the given fusion network $W_{(n\times n)}$ and the weighted genes, the label propagation model is to reassign a score to each gene according to the structure of the network and the initial weights of genes. Let $\hat{g}$ be the initial weight of genes in the fusion network which represents the prior information constructed by different scoring methods, and g be the score vector to be learnt. The label propagation considers that two connected nodes in the network should be assigned similar scores [53], and the objective function can be expressed as follows.

$$\psi(\mathrm{g}) = \sum_{i,j}^{n} W_{ij}\left(\frac{g_i}{\sqrt{D_{ii}}} - \frac{g_j}{\sqrt{D_{jj}}}\right)^2 + \frac{1-\alpha}{\alpha}\sum_{i}^{n}(g_i - \hat{g}_i)^2 \quad (3)$$

where $\hat{g}_i$ is the i-*th* element of vector $\hat{g}$ while g_i is the i-*th* element of vector g. D is a diagonal matrix and D_{ii} is the sum of row i of W. α ($0<\alpha<1$) is used to control the relative importance of the prior information, and the value is set as 0.6. The parameter analysis is shown in Fig.S1. In Eq. (3), the first term is the

Laplacian graph constraint, and it is used to make the connected genes to have similar label scores. The second term is the regularization term, which ensures the label scores of genes to close the initial scores.

After constructing the normalized matrix $W' = D^{-1/2}WD^{-1/2}$, i.e., $W'_{ij} = W_{ij} / \sqrt{D_{ii}D_{jj}}$, Eq. (3) is equivalent to,

$$\psi(g) = g^T(I - W')g + \frac{1-\alpha}{\alpha}(g-\hat{g})^T(g-\hat{g}) \quad (4)$$

For each cluster, three sets of scores of genes are obtained by the label propagation model and the three scoring strategies. Then, the comprehensive scores of genes are obtained by,

$$Score(i) = \frac{1}{3}\sum_{m=1}^{3}\psi_m(i) \quad (5)$$

where m represents the three scoring strategies: biological functions, prognostic ability and correlation with DAGs, and $\Psi_m(i)$ is the score of gene i in the m-*th* scoring strategy.

Finally, we computed the average rank of each gene in all clusters and selected high ranking genes (in the study, the top-ranking value is set as 100) as biomarkers. Thus, genes which are more important topologically and functionally, related to breast cancer and significantly differentially expressed are ranked higher. The details are stated in Algorithm FUNMarker.

Algorithm FUNMarker Algorithm of biomarker discovery

Input: Genes expression data and fusion network

Output: Biomarkers

```
Data_norm ← normalize the gene expression data
Obtain PC1 and PC2 from Data_norm based on PCA
Cluster the samples using K-means
for c = 1;:::;K do
    for i = 1;2;3 (three scoring strategies) do
        weighted_network_i ← i scoring strategy
        score_i ← Label propagation (weighted_network_i)
    end for
    Score_c = (1/3) Σ_{m=1}^{3} score_m
end for
Compute the average rank of each gene in all clusters
Biomarkers ← Top-ranking genes
```

3 RESULTS AND DISCUSSION

3.1 Prognostic effectiveness of biomarker identification methods

To evaluate the performance of our method, a comprehensive scheme was built to verify the effectiveness of our method. Three network-based biomarker identification methods were compared to our method, namely NetRank [17], stSVM [18] and CPR [23]. NetRank is a biomarker discovery method which is alike to Pagerank. NetRank sorts genes based on their prognostic relevance using both expression and network information. stSVM identifies biomarkers by smoothing t-statistics of individual genes over the structure of network. CPR clusters samples and uses modified PageRank to score and rank genes and extracts effective prognostic and heterogeneous features.

A random forest classifier and five-fold cross-validation were used to evaluate the performance of the methods in the classification. For the unbiased evaluation, we repeated these experiments for 100 times for entire datasets.

The ROC curves and AUC were shown in Fig.2 and Table 2. The results have shown that the ROC and AUC resulted from FUNMarker is better than other methods, which indicates that the biomarkers found by FUNMarker have stronger discriminative power than those by other methods for breast cancer patients with different prognosis. Meanwhile, the method CPR shows better classification performance than the other two methods in general. This may be attributed to the fact that CPR reduced the influence of heterogeneity of samples through sample clustering. However, CPR does not consider enough biological information in view of the complexity of cancer samples. This may lead to its relatively poorer classification ability than FUNMarker.

TABLE 2 AUC OF THE FOUR METHODS ON SIX DATASETS

Methods	GSE1456	GSE2034	GSE3494	GSE4922	NKI	BRCA
FUNMarker	0.87	0.73	0.85	0.76	0.88	0.83
CPR	0.70	0.64	0.65	0.63	0.78	0.64
NetRank	0.67	0.63	0.62	0.59	0.74	0.61
stSVM	0.68	0.58	0.63	0.60	0.74	0.60

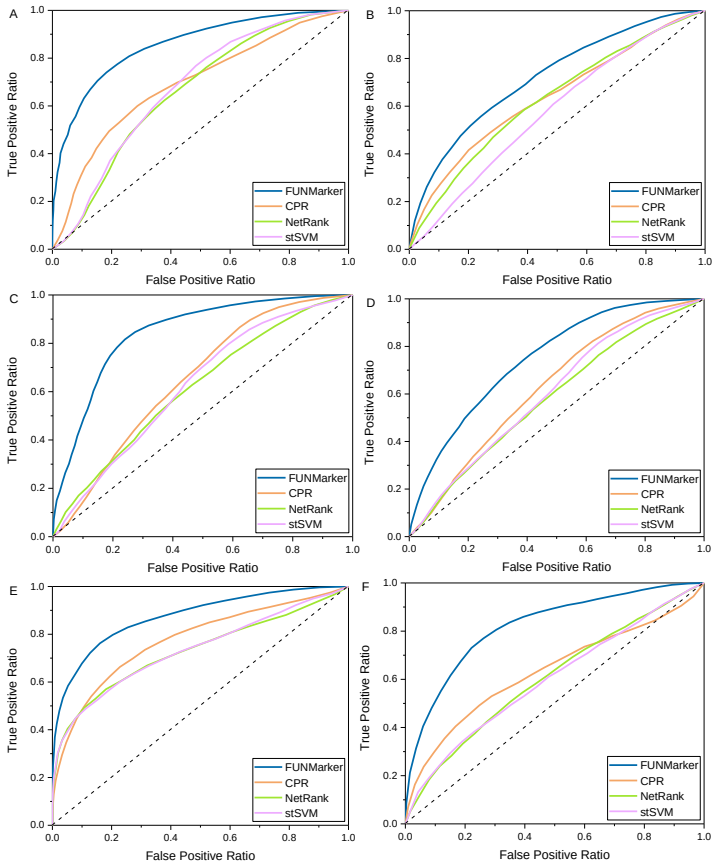

Fig.2 ROC curves. (A) GSE1456, (B) GSE2034, (C) GSE3494, (D) GSE4922, (E) NKI, (F) BRCA.

3.2 Analysis of the effectiveness of fusion network

To verify the effectiveness of the fusion network in the cancer prognosis analysis, we applied our proposed framework to every single biological network, and then we compared the AUC of every single network and fusion network. The results were shown in Fig.3. For every dataset, the AUC of our method using the fusion network is higher than those using single networks, which indicates that, compared with any single network, the fusion network can improve the accuracy of classification on each dataset.

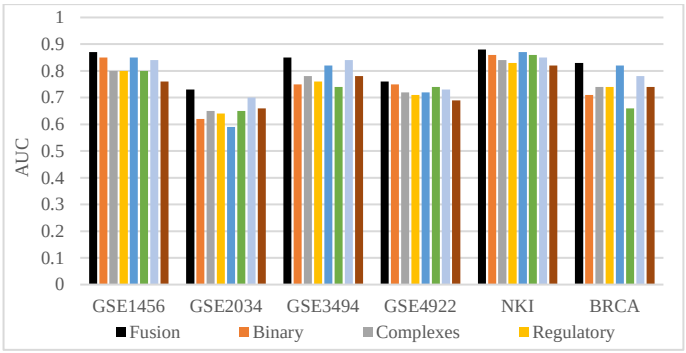

Fig.3 Comparison of the fusion network and single networks.

3.3 Reproducibility and Kaplan–Meier survival analysis

Reproducibility is an important consideration for the effectiveness of biomarker identification methods. Therefore, we analyzed the overlapped prognostic biomarkers among the six datasets. Fig.4 shows that there are overlapped biomarkers appearing in *n* datasets, which indicates that FUNMarker can capture biomarkers with reproducibility. Meanwhile, biomarkers identified by FUNMarker include some famous disease genes related to breast cancer, such as TP53, AKT1, BRCA1, CHEK2, EP300, MLH1, CTNNB1, ESR1, ATF4, MDM2, STK11, IKBKG, and TP53, IKBKG, EP300 appear in at least two datasets.

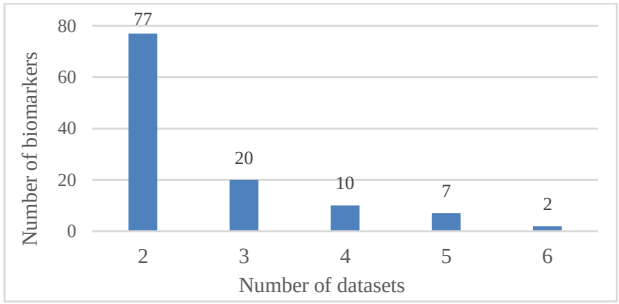

Fig.4 Reproducibility of biomarker prediction. The bars show the number of prognostic biomarkers appearing in *n* datasets.

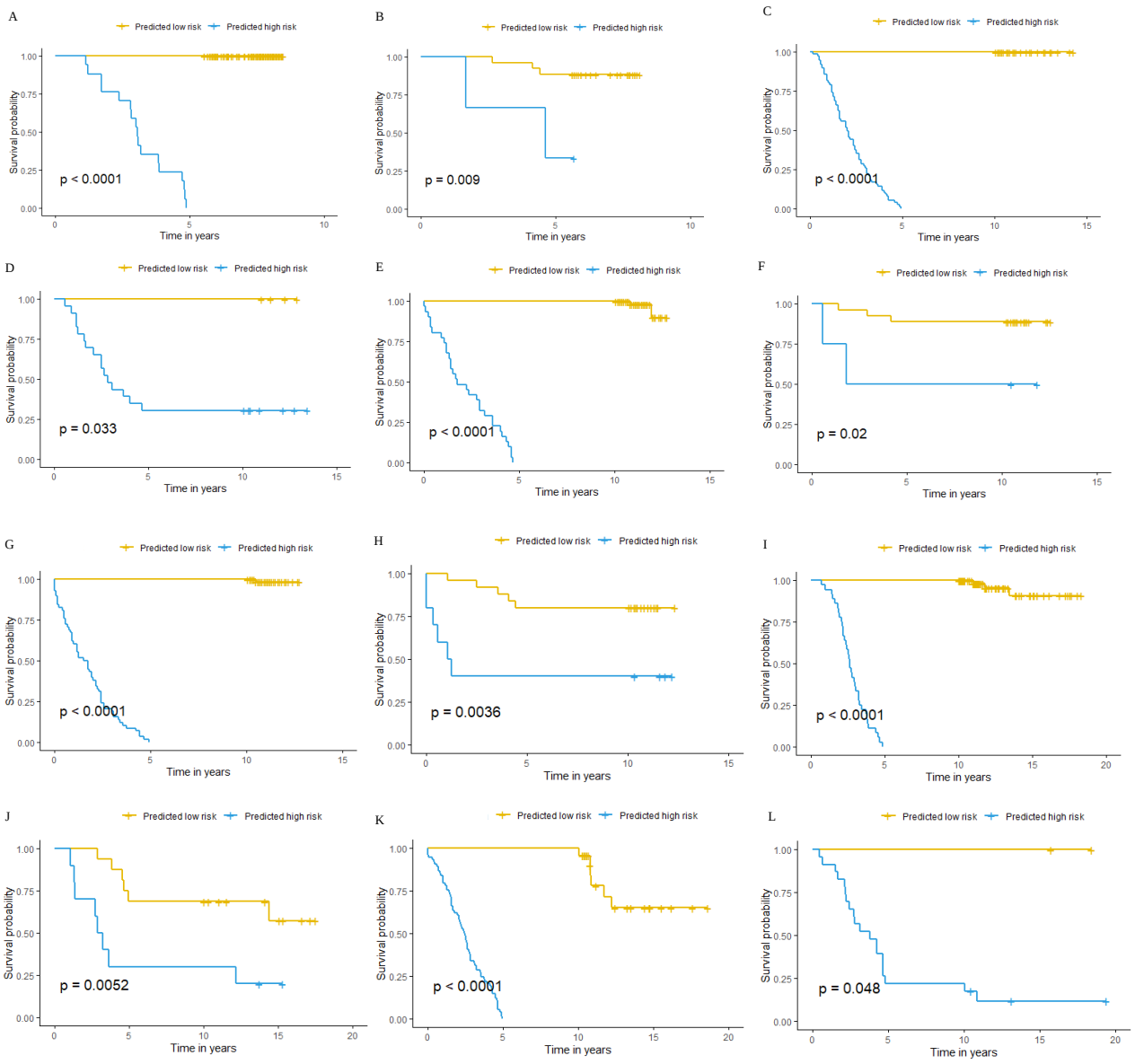

Fig.5 The Kaplan-Meier survival curves for the prognosis data of breast cancer in training and test cohorts. (A) GSE1456 in training cohorts. (B) GSE1456 in test cohorts. (C) GSE2034 in training cohorts. (D) GSE2034 in test cohorts. (E) GSE3494 in training cohorts. (F) GSE3494 in test cohorts. (G) GSE4922 in training cohorts. (H) GSE4922 in test cohorts. (I) NKI in training cohorts. (J) NKI in test cohorts. (K) BRCA in training cohorts. (L) BRCA in test cohorts.

Next, we estimated the advantage of converting the survival time regression to the classification problem, which divides the patients into two classes based on the survival time, and the effectiveness of prognostic biomarkers identified by FUNMarker. The Kaplan–Meier survival analysis is implemented by R package 'survival'.

To investigate the Kaplan-Meier analysis for biomarkers captured by FUNMarker, we chose the cutoff score according to the votes of random forest and then samples in both the training and test cohorts were classified into two classes: good and poor prognosis by using biomarkers resulted from FUNMarker. The Kaplan-Meier survival estimation is shown in Fig.5. The Kaplan-Meier survival curves show that p-value<0.0001 in training cohorts (Fig.5A, Fig.5C, Fig.5E, Fig.5G, Fig.5I, Fig.5K) and *p*-values in test cohorts are significant on all the six datasets (Fig.5B, Fig.5D, Fig.5F, Fig.5H, Fig.5J, Fig.5L), which indicates that FUNMarker performs well for the breast cancer.

3.4 Functional analysis of biomarkers

To verify the relationship between biomarkers identified by FUNMarker and breast cancer, we first analyzed the GO annotations for the identified biomarkers. The R package 'clusterProfiler' was used for GO enrichment analysis at three levels: biological process (BP), cellular component (CC), and molecular function (MF). The results of BRCA dataset were shown in Fig.6 and Fig.S2-Fig.S6. It can be seen that most of biomarkers are mapped well onto breast cancer-related processes or biological factors, which reveals the functional importance of the identified biomarkers.

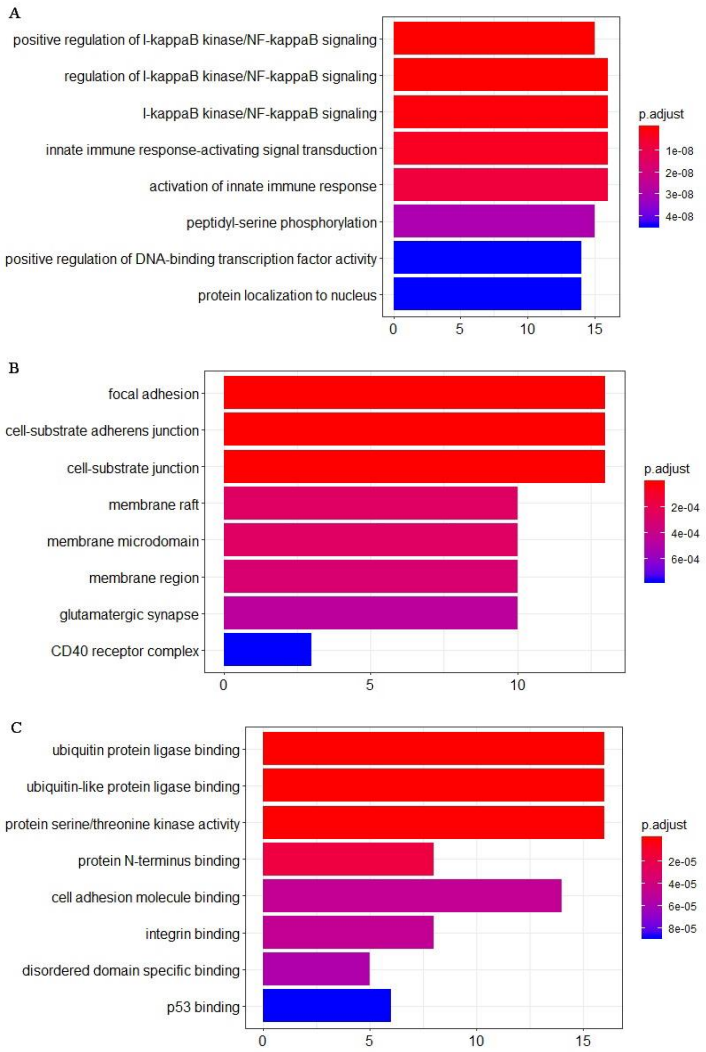

Fig.6 GO annotations for the identified biomarkers in BRCA dataset. (A) BP. (B) CC. (C) MF.

In order to verify the interpretability of biomarkers found by FUNMarker, we analyzed the enrichment of DAGs in biomarkers. DAGs can directly reveal the causes of diseases in the biological sense, Thus, the enrichment of DAGs in biomarkers were calculated to indicate the functional interpretability power of biomarkers. The hypergeometric test was used to calculate the p-value of the enrichment of the DAGs.

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \tag{6}$$

where N is the number of all genes in the gene expression data, M is the number of DAGs enriched in all genes, n is the number of biomarkers, m is the number of DAGs enriched in the biomarkers.

We transformed the p-value to $-log_{10}(p-value)$. The significance of DAG ratio of biomarkers captured by FUNMarker in six datasets were shown in Fig.7. The results show that the method incorporating the functional information and inferring features from multiple levels is more enriched with DAGs.

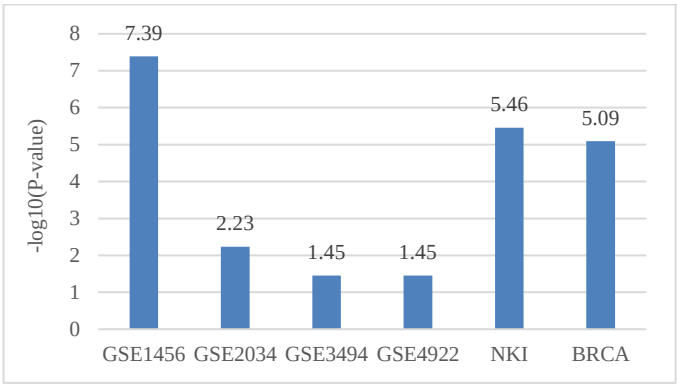

Fig.7 Significance of DAG ratio for biomarkers identified by FUNMarker in six datasets was evaluated by hypergeometric test.

Differentially expressed genes (DEGs) can effectively assess the extent of dysfunction and may contain more prognosis information. Therefore, we also analyzed the level of significance of dysregulation in the biomarkers. DEGs were obtained by t-test and hypergeometric test was used to calculate the p-value of the enrichment of the DEGs. The significance of DEG ratio is shown in Fig.8. Obviously, FUNMarker has the powerful ability of identifying dysfunction-explainable biomarkers.

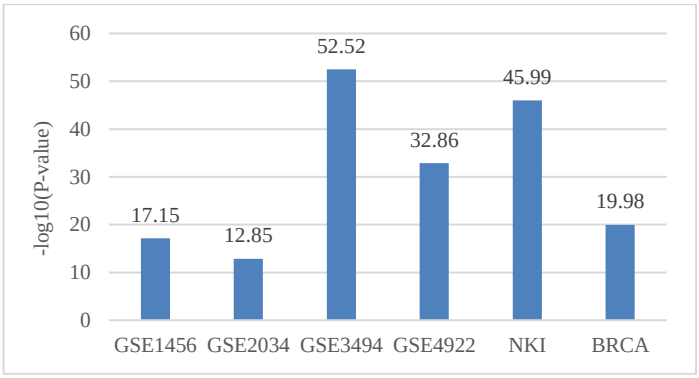

Fig.8 Significance of DEGs ratio for biomarkers identified by FUNMarker in six datasets was evaluated by hypergeometric test.

4 CONCLUSION

As a highly heterogeneous disease, breast cancer lacks effective prognostic biomarkers that can effectively predict the outcome of breast cancer patients. Conventional network-based methods either only utilize the network with single attribute and gene expression profiles to identify the prognostic biomarkers, or use network with multiple attributes without considering the integration of multiple biological information. The former may lead to the inaccurate results due to the incomplete interactome, false-positives and false-negatives of the networks, and the latter may result in that important information may not be used effectively.

In this study, we proposed a novel computational framework, which identified the prognostic breast cancer

biomarkers based on multiple network fusion and multiple scoring strategies. The biomarkers captured by our approach have both strong discriminative power in differentiating patients with different prognostic outcomes and better biological interpretability.

To validate the performance of our computational framework, we have compared our approach with some state-of-the-art approaches, such as CPR, NetRank and stSVM. The results have shown that the biomarkers can classify patients more accurately than other methods. We also analyzed the effectiveness of the fusion network in the cancer prognosis analysis, and the results showed that the fusion network can improve the accuracy of classification on each dataset, compared to any single network. Then, reproducibility and Kaplan–Meier survival analysis were investigated. Moreover, as the features should be biologically meaningful, we also analyzed the functional interpretability and dysregulation of biomarkers. Since DAGs can directly reflect the correlation of diseases, DEGs can effectively evaluate the extent of dysfunction and may capture more prognosis information. We analyzed the biological sense of biomarkers according to the enrichment of DAGs and DEGs. As expected, the results indicated that the biomarkers obtained by FUNMarker have meaningful biological sense.

At present, we only analyze the biomarkers related to the prognosis of breast cancer. In the future, we would expand our work to pan-cancer analysis. Meanwhile, a smaller but powerful set of biomarkers are more practical in the clinical application, thus we would optimize FUNMarker to achieve the optimal classification ability with the minimum number of biomarkers.

ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China (60832019, 61772552), the 111 Project (No. B18059), the Hunan Provincial Science and Technology Program (2018WK4001), and the Hunan Provincial Innovation Foundation For Postgraduate (CX20190123). Parts of this paper appeared at the 2019 International Conference on Intelligent Computing (ICIC2019) [54].

REFERENCES

[1] B. Weigelt, J. L. Peterse, and L. J. Van't Veer, "Breast cancer metastasis: markers and models," *Nature reviews cancer,* vol. 5, no. 8, pp. 591, 2005.

[2] D. J. Slamon, G. M. Clark, S. G. Wong, W. J. Levin, A. Ullrich, and W. L. McGuire, "Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene," *science,* vol. 235, no. 4785, pp. 177-182, 1987.

[3] T. J. Key, P. K. Verkasalo, and E. Banks, "Epidemiology of breast cancer," *The lancet oncology,* vol. 2, no. 3, pp. 133-140, 2001.

[4] H. G. Russnes, N. Navin, J. Hicks, and A.-L. Borresen-Dale, "Insight into the heterogeneity of breast cancer through next-generation sequencing," *The Journal of clinical investigation,* vol. 121, no. 10, pp. 3810-3818, 2011.

[5] B. D. Lehmann, J. A. Bauer, X. Chen, M. E. Sanders, A. B. Chakravarthy, Y. Shyr, and J. A. Pietenpol, "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *The Journal of clinical investigation,* vol. 121, no. 7, pp. 2750-2767, 2011.

[6] L. J. Van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. Van Der Kooy, M. J. Marton, and A. T. Witteveen, "Gene expression profiling predicts clinical outcome of breast cancer," *nature,* vol. 415, no. 6871, pp. 530, 2002.

[7] C. L. Sawyers, "The cancer biomarker problem," *Nature,* vol. 452, no. 7187, pp. 548, 2008.

[8] M. C. Cheang, D. Voduc, C. Bajdik, S. Leung, S. McKinney, S. K. Chia, C. M. Perou, and T. O. Nielsen, "Basal-like breast cancer defined by five biomarkers has superior prognostic value than triple-negative phenotype," *Clinical cancer research,* vol. 14, no. 5, pp. 1368-1376, 2008.

[9] M. T. Weigel, and M. Dowsett, "Current and emerging biomarkers in breast cancer: prognosis and prediction," *Endocrine-related cancer,* vol. 17, no. 4, pp. R245-R262, 2010.

[10] B. Győrffy, A. Lanczky, A. C. Eklund, C. Denkert, J. Budczies, Q. Li, and Z. Szallasi, "An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients," *Breast cancer research and treatment,* vol. 123, no. 3, pp. 725-731, 2010.

[11] X. Li, W. Li, M. Zeng, R. Zheng, and M. Li, "Network-based methods for predicting essential genes or proteins: a survey," *Briefings in bioinformatics*, doi: 10.1093/bib/bbz017 [Epub ahead of print], 2019.

[12] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature reviews genetics,* vol. 12, no. 1, pp. 56, 2011.

[13] M. Li, R. Zheng, Y. Li, F. X. Wu, and J. Wang, "MGT-SM: A Method for Constructing Cellular Signal Transduction Networks," *IEEE/ACM Transactions on Computational Biology & Bioinformatics,* vol. PP, no. 99, pp. 1-1, 2017.

[14] X. Chen, M. Li, R. Zheng, S. Zhao, F.-X. Wu, Y. Li, and J. Wang, "A novel method of gene regulatory network structure inference from gene knock-out expression data," *Tsinghua Science and Technology,* vol. 24, no. 4, pp. 446-455, 2019.

[15] S. L. Wong, L. V. Zhang, A. H. Tong, Z. Li, D. S. Goldberg, O. D. King, G. Lesage, M. Vidal, B. Andrews, and H. Bussey, "Combining biological networks to predict genetic interactions," *Proceedings of the National Academy of Sciences,* vol. 101, no. 44, pp. 15682-15687, 2004.

[16] D. D. Licatalosi, and R. B. Darnell, "RNA processing and its regulation: global insights into biological networks," *Nature Reviews Genetics,* vol. 11, no. 1, pp. 75, 2010.

[17] C. Winter, G. Kristiansen, S. Kersting, J. Roy, D. Aust, T. Knösel, P. Rümmele, B. Jahnke, V. Hentrich, and F. Rückert, "Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes," *PLoS computational biology,* vol. 8, no. 5, pp. e1002511, 2012.

[18] Y. Cun, and H. Fröhlich, "Network and data integration for biomarker signature discovery via network smoothed t-statistics," *PloS one,* vol. 8, no. 9, pp. e73074, 2013.

[19] R. Liu, X. Wang, K. Aihara, and L. Chen, "Early diagnosis of complex diseases by molecular biomarkers, network biomarkers, and dynamical network biomarkers," *Medicinal research reviews,* vol. 34, no. 3, pp. 455-478, 2014.

[20] J. Zhang, Y. Xiang, L. Ding, T. B. Borlawsky, H. G. Ozer, R. Jin, P. Payne, and K. Huang, "Using gene co-expression network analysis to predict biomarkers for chronic lymphocytic leukemia." p. S5.

[21] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási, "Uncovering disease-disease relationships through the incomplete interactome," *Science,* vol. 347, no. 6224, pp. 1257601, 2015.

[22] X. Wang, S.-s. Wang, L. Zhou, L. Yu, and L.-m. Zhang, "A network-pathway based module identification for predicting the prognosis of ovarian cancer patients," *Journal of ovarian research,* vol. 9, no. 1, pp. 73, 2016.

[23] J. Choi, S. Park, Y. Yoon, and J. Ahn, "Improved prediction of breast cancer outcome by identifying heterogeneous biomarkers," *Bioinformatics,* vol. 33, no. 22, pp. 3619-3626, 2017.

[24] W. Liu, C. Li, Y. Xu, H. Yang, Q. Yao, J. Han, D. Shang, C. Zhang, F. Su, and X. Li, "Topologically inferring risk-active pathways toward precise cancer classification by directed random walk," *Bioinformatics,* vol. 29, no. 17, pp. 2169-2177, 2013.

[25] K. Polyak, "Heterogeneity in breast cancer," *The Journal of clinical investigation,* vol. 121, no. 10, pp. 3786-3788, 2011.

[26] R. A. Burrell, N. McGranahan, J. Bartek, and C. Swanton, "The causes and consequences of genetic heterogeneity in cancer evolution," *Nature,* vol. 501, no. 7467, pp. 338, 2013.

[27] E. C. de Bruin, N. McGranahan, R. Mitter, M. Salm, D. C. Wedge, L. Yates, M. Jamal-Hanjani, S. Shafi, N. Murugaesu, and A. J. Rowan, "Spatial and temporal diversity in genomic instability processes defines lung cancer evolution," *Science,* vol. 346, no. 6206, pp. 251-256, 2014.

[28] K. Ovaska, M. Laakso, S. Haapa-Paananen, R. Louhimo, P. Chen, V. Aittomäki, E. Valo, J. Núñez-Fontarnau, V. Rantanen, and S. Karinen, "Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme," *Genome medicine,* vol. 2, no. 9, pp. 65, 2010.

[29] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: recent progress in multi-omics data integration methods," *Frontiers in genetics,* vol. 8, pp. 84, 2017.

[30] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic acids research,* vol. 30, no. 1, pp. 207-210, 2002.

[31] Y. Pawitan, J. Bjöhle, L. Amler, A.-L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, and S. Klaar, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts," *Breast cancer research,* vol. 7, no. 6, pp. R953, 2005.

[32] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, and J. Yu, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet,* vol. 365, no. 9460, pp. 671-679, 2005.

[33] L. D. Miller, J. Smeds, J. George, V. B. Vega, L. Vergara, A. Ploner, Y. Pawitan, P. Hall, S. Klaar, and E. T. Liu, "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," *Proceedings of the National Academy of Sciences,* vol. 102, no. 38, pp. 13550-13555, 2005.

[34] A. V. Ivshina, J. George, O. Senko, B. Mow, T. C. Putti, J. Smeds, T. Lindahl, Y. Pawitan, P. Hall, and H. Nordgren, "Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer," *Cancer research,* vol. 66, no. 21, pp. 10292-10301, 2006.

[35] M. J. Van De Vijver, Y. D. He, L. J. Van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, and M. J. Marton, "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine,* vol. 347, no. 25, pp. 1999-2009, 2002.

[36] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, and J. T. Eppig, "Gene ontology: tool for the unification of biology," *Nature genetics,* vol. 25, no. 1, pp. 25, 2000.

[37] G. O. Consortium, "Expansion of the Gene Ontology knowledgebase and resources," *Nucleic acids research,* vol. 45, no. D1, pp. D331-D338, 2016.

[38] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic acids research*, pp. gkw943, 2016.

[39] T. Rolland, M. Taşan, B. Charloteaux, S. J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, and R. Mosca, "A proteome-scale map of the human interactome network," *Cell,* vol. 159, no. 5, pp. 1212-1226, 2014.

[40] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, and N. Ayivi-Guedehoussou, "Towards a proteome-scale map of the human protein–protein interaction network," *Nature,* vol. 437, no. 7062, pp. 1173, 2005.

[41] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, and S. Koeppen, "A human protein-protein interaction network: a resource for annotating the proteome," *Cell,* vol. 122, no. 6, pp. 957-968, 2005.

[42] K. Venkatesan, J.-F. Rual, A. Vazquez, U. Stelzl, I. Lemmens, T. Hirozane-Kishikawa, T. Hao, M. Zenkner, X. Xin, and K.-I. Goh, "An empirical framework for binary interactome mapping," *Nature methods,* vol. 6, no. 1, pp. 83, 2009.

[43] H. Yu, L. Tardivo, S. Tam, E. Weiner, F. Gebreab, C. Fan, N. Svrzikapa, T. Hirozane-Kishikawa, E. Rietman, and X. Yang, "Next-generation sequencing to generate interactome datasets," *Nature methods,* vol. 8, no. 6, pp. 478, 2011.

[44] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, and U. Hinz, "The IntAct molecular interaction database in 2012," *Nucleic acids research,* vol. 40, no. D1, pp. D841-D846, 2011.

[45] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A. P. Nardozza, and E. Santonico, "MINT, the molecular interaction database: 2012 update," *Nucleic acids research,* vol. 40, no. D1, pp. D857-D861, 2011.

[46] C. Stark, B.-J. Breitkreutz, A. Chatr-Aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, and X. Shi, "The BioGRID interaction database: 2011 update," *Nucleic acids research,* vol. 39, no. suppl_1, pp. D698-D704, 2010.

[47] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, and A. Venugopal, "Human protein reference database—2009 update," *Nucleic acids research,* vol. 37, no. suppl_1, pp. D767-D772, 2008.

[48] V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, and O. V. Kel-Margoulis, "TRANSFAC®: transcriptional regulation, from patterns to profiles," *Nucleic acids research,* vol. 31, no. 1, pp. 374-378, 2003.

[49] D.-S. Lee, J. Park, K. Kay, N. A. Christakis, Z. Oltvai, and A.-L. Barabási, "The implications of human metabolic network topology for disease comorbidity," *Proceedings of the National Academy of Sciences,* vol. 105, no. 29, pp. 9880-9885, 2008.

[50] A. Ruepp, B. Waegele, M. Lechner, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and H.-W. Mewes, "CORUM: the comprehensive resource of mammalian protein complexes—2009," *Nucleic acids research,* vol. 38, no. suppl_1, pp. D497-D501, 2009.

[51] P. V. Hornbeck, J. M. Kornhauser, S. Tkachev, B. Zhang, E. Skrzypek, B. Murray, V. Latham, and M. Sullivan, "PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse," *Nucleic acids research,* vol. 40, no. D1, pp. D261-D270, 2011.

[52] A. Vinayagam, U. Stelzl, R. Foulle, S. Plassmann, M. Zenkner, J. Timm, H. E. Assmus, M. A. Andrade-Navarro, and E. E. Wanker, "A directed protein interaction network for investigating intracellular signal transduction," *Sci. Signal.,* vol. 4, no. 189, pp. rs8-rs8, 2011.

[53] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLoS computational biology,* vol. 6, no. 1, pp. e1000641, 2010.

[54] X. Li, J. Xiang, J. Wang, F.-X. Wu, and M. Li, "Identification of Prognostic and Heterogeneous Breast Cancer Biomarkers Based on Fusion Network and Multiple Scoring Strategies." pp. 529-534.

Xingyi Li received the B.S. degree in communication engineering from Central South University, China, in 2015. Currently, she is working toward the Ph.D. degree in the School of Computer Science and Engineering, Central South University, Changsha, China. Her current research interests include bioinformatics and systems biology.

Ju Xiang received the B.S. and M.S. degrees from Xiangtan University in 2005 and 2008, separately. He is currently working toward the PhD degree in the School of Computer Science and Engineering, Central South University, China. His research interests include complex networks, bioinformatics, machine learning and deep learning.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the Dean and a Professor of the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, Bioinformatics and computer network. He is a senior member of the IEEE.

Jinyan Li is a Professor of Data Science and Program Leader of Bioinformatics at the Advanced Analytics Institute, Faculty of Engineering & IT, University of Technology Sydney, Australia. He has been actively working on data mining and bioinformatics for 20 years. He has published 220 papers, including 120 papers in the prestigious journals of data mining, machine learning, and computational biology. He is widely known for his pioneering research on the theories and algorithms of Emerging Patterns (EPs). One of these papers has received 1230 Google Scholar citations. Jinyan has a Bachelor degree of Science (Applied Mathematics) from National University of Defense Technology (China), a Master degree of Engineering (Computer Engineering) from Hebei University of Technology (China), and a PhD degree (Computer Science) from the University of Melbourne (Australia).

Fang-Xiang Wu (M'06-SM'11) received the B.Sc.degree and the M.Sc. degree in applied mathematics, both from Dalian University of Technology, Dalian, China, in 1990 and 1993, respectively, the first Ph.D. degree in control theory and its applications from Northwestern Polytechnical University, Xi'an, China, in 1998, and the second Ph.D. degree in biomedical engineering from University of Saskatchewan (U of S), Saskatoon, Canada, in 2004. During 2004-2005, he worked as a Postdoctoral Fellow in the Laval University Medical Research Center (CHUL), Quebec City, Canada. He is currently a Professor of the Division of Biomedical Engineering and the Department of Mechanical Engineering at the U of S. His current research interests include Artificial Intelligence, Machine Learning/Deep learning, Computational Biology and Bioinformatics, Medical Image Analytics, Complex Network Analytics. Dr. Wu is serving as the editorial board member of five international journals, the guest editor of several international journals, and as the program committee chair or member of several international conferences. He has also reviewed papers for many international journals. He is a senior member of IEEE.

Min Li received the PhD degree in Computer Science from Central South University, China, in 2008. She is currently a Professor and vice Dean of the School of Computer Science and Engineering, Central South University, Changsha, Hunan, P.R. China. Her main research interests include bioinformatics and systems biology.