BGSSN: Breast Cancer-Associated Genes Prediction Based on Weighted Sample-Specific Networks of Cancer Subtypes

Qian Liu[1], Yuanyuan Zhang[1,a], Haoyu Zheng[1], and Shudong Wang[2]

[1]School of information and control engineering, Qingdao University of Technology, Qingdao, China;

[2]College of Computer science and Technology, China University of Petroleum (East China), Qingdao.

[a] yyzhang1217@163.com

Abstract. Breast cancer exhibits a notable degree of heterogeneity in its occurrence and progression, encompassing diverse clinical patterns and outcomes among patients even with identical clinical pathological stages. Genetic mutations in different subtypes of breast cancer may lead to different types of disease and have different clinical implications. Therefore, molecular typing based on the characteristics of breast cancer heterogeneity and the screening of associated genes for different subtypes of breast cancer may be able to more accurately determine the pathogenic genes of breast cancer. In this paper, we propose a weighted sample-specific network based on breast cancer subtypes to predict associated genes, named BGSSN. To better reflect the individual characteristics of patients and the importance of patient samples in different subtypes, the weight of samples is added when constructing the sample-specific network. The random walk with restart method is then utilized to predict new breast cancer-associated genes within the constructed network. By leveraging this method, the network structure can be effectively explored to identify potential gene candidates.

Keywords: cancer subtypes; genes prediction; weighted sample-specific networks; random walk with restart.

1. Introduction

Breast cancer has become the most common malignant tumor, ranking the first among female malignant tumors and its incidence is still on the rise, which seriously threatens women's health. The occurrence and development of breast cancer is the result of multiple factors, with high heterogeneity, leading to different prognosis and treatment response, and there are problems of diagnosis difficulties or biases. The causes of tumor heterogeneity include genomic differences within cancer cells, transcriptome differences, and epigenetic modification differences, which require us to further refine the disease-associated genes in order to better guide the treatment of cancer[1, 2]. Therefore, identifying genes associated with different subtypes of breast cancer is an important goal for accurate diagnosis, treatment, and prevention of breast cancer.

At present, many computational methods have been proposed for disease-gene associations prediction. These methods can be roughly divided into three categories: methods using graph theory algorithms; methods using machine learning algorithms; and methods to combine graph theory and machine learning techniques[3]. Methods that use graph theory algorithms to predict disease-gene associations include RWR[4], RWRH[5], PRINCE[6], DADA[7], RWr-MH[8], PhenoRank[9], and NetCore[10]. These methods usually obtain new weights of genes through random walks in the context of PPI networks, which can simply reorder genes to identify new disease genes. However, these methods preselect seed genes from PPI networks with known associations, and not all genes are directly connected in PPI networks, which has certain limitations. Methods that use machine learning algorithms to predict disease-gene associations include CIPHER[11], CrossRankStar[12], pBRIT[13], Scuba[14]. They can efficiently handle large numbers of candidate genes and any number of data sources, as well as severe imbalances with few known disease genes that need to be predicted on a large scale, but they require a large memory footprint. Methods that use a combination of graph theory and machine learning techniques to predict disease-gene associations include IDLP[15] and HerGePred[16]. These methods are based on the input of disease gene heterogeneity networks, and they can predict the genes associated with diseases that are not associated with PPI networks. But they predict a lower degree of disease-associated genes in the PPI network. These methods can predict highly associated genes only if the disease-associated genes are unknown. Most of the above methods are used to predict disease-gene association for multiple diseases at the same time, and few methods to predict associated genes based on disease heterogeneity.

To address this problem, a weighted sample-specific network called BGSSN is proposed for predicting breast cancer-associated genes based on cancer subtypes. This approach assigns different weights to tumor samples of various subtypes, allowing for the construction of weighted sample-specific networks that capture individual patient characteristics. The random walk with restart method is employed to predict new breast-associated genes within these networks. Experiments have proved that BGSSN can more accurately predict breast cancer genes and assist tumor targeted therapy. An overview of the BGSSN is shown in Fig. 1.

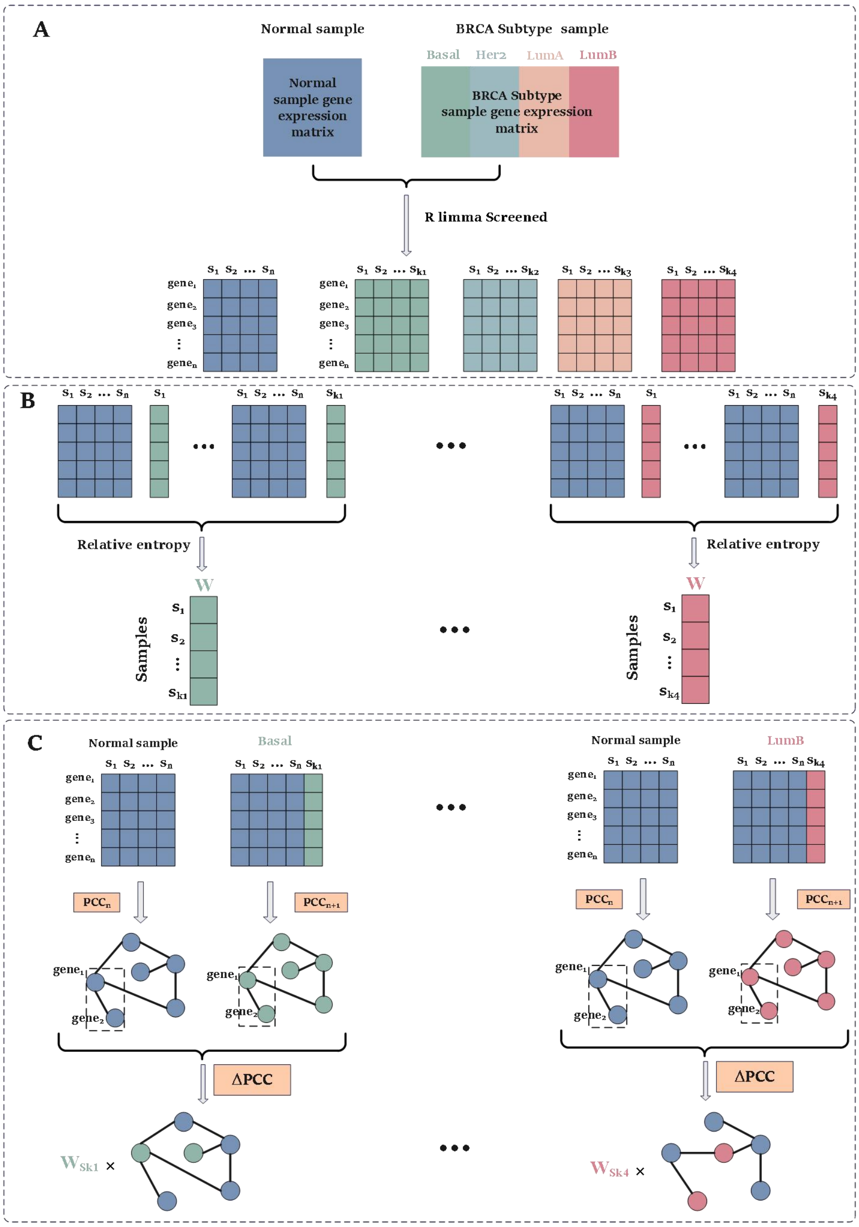

Fig. 1 Overall overview of BGSSN. For breast cancer subtypes, firstly, differentially expressed genes are screened using the R package, secondly, assign different weights to different subtypes of tumor samples. finally, construct weighted sample-specific network and the random walk with restart algorithm is used to predict new breast cancer-associated genes

2. Materials and Methods

2.1 Data Preprocessing

We download BRCA datasets from the UCSC Xena browser (https://xenabrowser.net/), which is divided into four subtypes uses PAM50 assay (Basal, Her2, LumA and LumB). For the data of gene expression, The R package limma is utilized to analyze gene expression data and identify differentially expressed genes among various breast cancer subtypes. As shown in Table 1. For the selection of known breast cancer-associated genes, we refer to the DISEASES resource[17] and choose 50 genes as references (See supplementary materials Table 1).

Table 1. Breast cancer dataset used in this study

Breast cancer Subtype	Sample	Gene expression	The gene after screening
Normal	114	16582	1000
Basal	112	16582	1000
Her2	53	16582	1000
LumA	248	16582	1000
LumB	98	16582	1000

2.2 Methods

2.2.1 Sample importance assessment

In order to address the issue of small data sample size and accurately capture the importance of each sample, the weight of each sample i is determined using the relative entropy between the normal sample and the subtype sample i . Relative entropy, also known as Kullback-Leibler divergence, is a way of describing the difference between two probability distributions. The calculation formula is as follows.

$$P_{g^{(Si)}} = \frac{g^{(Si)}}{\sum_{g} g^{(Si)}}, \tag{1}$$

$$Q_{g^{n}} = \frac{g^{n}}{\sum_{g} g^{n}}, \tag{2}$$

where $g^{(Si)}$ is the expression value of gene g in the sample i of subtype S , $P_{g^{(Si)}}$ is the probability of the total gene distribution of gene g in the sample i of subtype S , and $\sum_{g} P_{g^{(Si)}} = 1$. g^{n} is the expression value of gene g in the normal sample n , $Q_{g^{n}}$ is the probability of the total gene distribution of gene g in the normal sample n and $\sum_{g} Q_{g^{n}} = 1$.

$$w^{(Si)} = \frac{\sum_{n}\sum_{g} P_{g^{(Si)}} \log(\frac{P_{g^{(Si)}}}{Q_{g^{n}}})}{n}, \tag{3}$$

where $w^{(Si)}$ is the weight of the sample i of subtype S , n is the total number of normal samples.

2.2.2 Construction of weighted Sample-specific network

Gene-gene networks reveal how genes interact with each other, in the process of gene network, it is necessary to count and calculate the correlation information between different samples, so it is necessary to establish a network with multiple sample data. However, this network only contains the common regulatory information among various samples, and ignores the specific regulatory abnormal information of each sample, so we propose a weighted sample-specific network. Sample-specific network in different subtypes S_{th} (S=1,...,s) is calculated, using gene expression data from normal samples as reference. First, the marginal Pearson correlation coefficient PCC_n is calculated for each pair of genes in the normal sample, then a single sample i is added to the normal sample to obtain the Pearson correlation coefficient $PCC_{n+1}^{(Si)}$ for each pair of genes in the new sample. Because of the specificity of a single sample, different samples have different differences in the same background network. Then, difference $\Delta PCC^{(Si)}$ between all normal samples and all additional single samples and the weights of the single samples as the gene edge score network $G^{(S)}$.

$$G^{(S)} = \sum_{i} \Delta PCC^{(Si)} = w^{(Si)} \times \frac{PCC_{n+1}^{(Si)} - PCC_n}{\left(1-PCC_n^2\right) \big/ (n-1)} \quad . \tag{4}$$

For genes g and g'. If $G_{gg'}^{(S)} > 0$, if there is a correlation between gene g and g', and the score is used as their correlation weight, where n is the total number of reference samples.

3. Results and Discussion

3.1 Analysis of the breast cancer-associated genes predicted by BGSSN

BGSSN separately predicts new breast cancer-associated genes for different subtypes (See supplementary materials Table 2). Interestingly, it is observed that up to more than 90% of the same genes are predicted in breast cancer subtypes (Fig. 2a). It is not uncommon for different cancer subtypes to share certain genetic characteristics. In this case, the presence of the common gene in different subtypes implies a potential similarity in the underlying molecular mechanisms or pathways involved in these subtypes. Additionally, unique genes are found in breast cancer subtypes. For instance, gene PIGT, CCNI in subtype Basal, YWHAH, APEX1 in subtype LumA, and KHK-A phosphorylation of YWHAH are clinically associated with breast cancer metastasis[18]. The pathogenesis of different subtypes can affect the expression and function of genes, so the study of diseases associated genes is great significance.

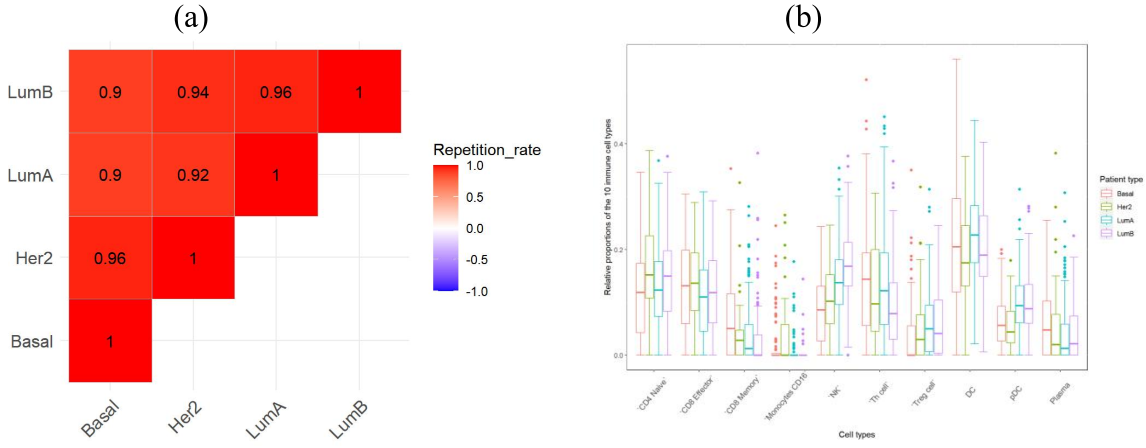

Fig. 2 (a)BGSSN predicts the duplication rate of genes associated with different subtypes of breast cancer (b) The proportion of ten immune cells in different of breast cancer subtypes.

3.2 Evaluation of infiltrating immune cells by tumor immunophenotype

From a clinical point of view, the prognostic effects of different immune cell populations in different subtypes or different stages of tumors are helpful in the design of effective personalized immunotherapies[19]. The expression value of the associated genes predicted by BGSSN is inputted into the TIP tool[20], which outputs the number of tumor immune cells. To visualize the different distributions between subtypes, we use boxplots to show the distribution of the estimated proportion of immune cell types for each subtype (Fig. 2b). It is observed that the proportion of immune cell types in different subtypes is significantly different, which provides a basis for further exploration of their prognostic value.

3.3 Ablation experiments

To demonstrate the performance of BGSSN in predicting breast cancer-associated genes, the accuracy of SSN, LIONESS, and BGSSN is compared for predicting known associated genes in different subtypes of breast cancer. SSN is a differential analysis method to assess SINs based on the statistical perturbation measurement of a single sample against a group of control samples[21]. LIONESS is a method to reverse engineer SINs from aggregate networks without the need for control samples[22]. Of the 50 known breast cancer genes, 20 are randomly selected as seeds to predict the accuracy of the remaining 30 known genes. As shown in Fig. 3a, BGSSN is superior to other methods in predicting genes associated with breast cancer. This shows that BGSSN has better performance in gene prediction than the current classical algorithms. Additionally, the accuracy of BGSSN is compared for predicting associated genes without differentiating subtypes and constructing a single-sample network without adding weights (Fig. 3b). The results demonstrate that BGSSN is better than the method without subtype and without weight in predicting associated genes. This finding implies that by subtyping the disease may lead to more effective identification of breast cancer genes.

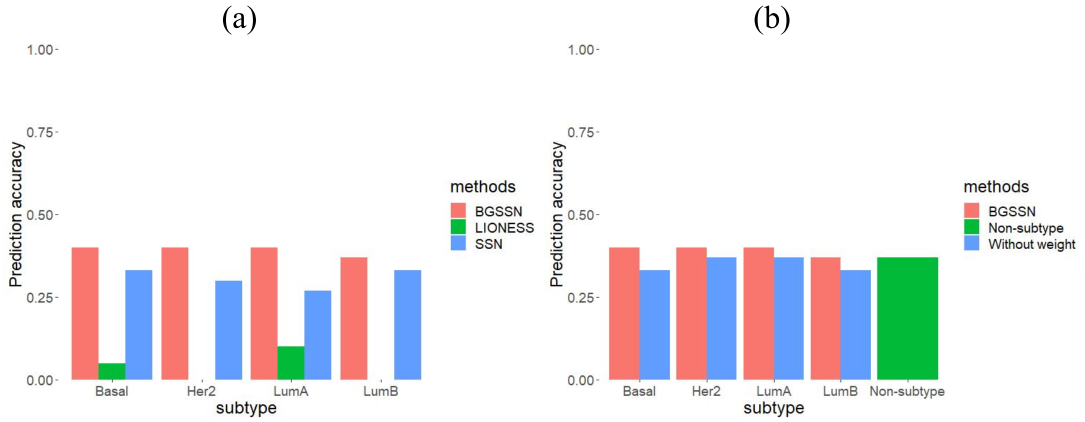

Fig. 3 (a) Accuracy of BGSSN, LIONESS, and SSN in predicting associated genes in different subtypes of breast cancer (b) Accuracy of BGSSN without distinguish subtypes and constructs a single sample network without adding weights to predict breast associated genes

4. Conclusions

Based on the heterogeneity of breast cancer, in order to accurately identify breast cancer-associated genes, we propose a weighted sample-specific method based on breast cancer subtypes to predict associated genes. By constructing a single weighted sample-specific network, the characteristics of the sample network are analyzed to further predict the associated genes of breast cancer. The results of prognostic analysis are helpful to design effective personalized immunotherapy for breast cancer, Ablation experiments demonstrate the performance of breast cancer-associated genes prediction algorithms. In conclusion, the genes predicted by BGSSN are highly likely to be breast cancer-associated genes, and provide new horizons and new ideas for finding the pathogenic genes of cancer.

In this work, only one type data on gene expression of omics data is used to predict breast cancer genes, future research should explore the incorporation of additional types of data, such as methylation, or fuse multiple omics data to identify new breast cancer-associated genes.

Availability of data and materials: Supplementary materials and BGSSN methods are available at https://github.com/qianliu2022/BGSSN-master.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China [Nos.61902430,61873281].

References

[1] Kavitha E S and Kumari E Nanda. CA: A Cancer Journal for Clinicians: A Bibliometrics Study. Library Philosophy and Practice, 2021, 61(2):69-90

[2] G Kann Maricel. Advances in translational bioinformatics: computational approaches for the hunting of disease genes. Briefings in bioinformatics, 2010, 11(1):96-110

[3] Yoonbee Kim, Jonghoon Park, and Youngrae Cho. Network-Based Approaches for Disease-Gene Association Prediction Using Protein-Protein Interaction Networks. International Journal of Molecular Sciences, 2022, 23(13):7411

[4] Köhler Sebastian, Bauer Sebastian, Horn Denise, et al. Walking the Interactome for Prioritization of Candidate Disease Genes. The American Journal of Human Genetics, 2008, 82(4):949-958

[5] Li Yongjin and C Patra Jagdish. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics, 2010, 26(9):1219-1224

[6] Vanunu Oron, Magger Oded, Ruppin Eytan, et al. Associating genes and protein complexes with disease via network propagation. PLoS Computational Biology, 2010, 6(1):e1000641

[7] Erten Sinan, Bebek Gurkan, Ewing Rob M, et al. DADA: Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. BioData mining, 2011, 4(1):1-20

[8] Alberto Valdeolivas, Laurent Tichit, Claire Navarro, et al. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics, 2019, 35(3):497-505

[9] J Cornish Alex, Alessia David, and E Sternberg Michael J. PhenoRank: reducing study bias in gene prioritization through simulations. Bioinformatic, 2018, 34(12):2087-2095

[10] Gal Barel and Ralf Herwig. NetCore: a network propagation approach using node coreness. Nucleic acids research, 2020, 48(17):e98

[11] Wu Xuebing, Jiang Rui, Zhang Michael Q, et al. Network-based global inference of human disease genes. Molecular systems biology, 2008, 4(1):189

[12] Ni Jingchao, Mehmet Koyuturk, Tong Hanghang, et al. Disease gene prioritization by integrating tissue-specific molecular networks using a robust multi-network model. BMC bioinformatics, 2016, 17(1):1-13

[13] Anand Kumar Ajay, Lut Van Laer, Maaike Alaerts, et al. pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion. Bioinformatics, 2018, 34(13):2254-2262

[14] Guido Zampieri, Van Tran Dinh, Michele Donini, et al. Scuba: scalable kernel-based gene prioritization. BMC bioinformatics, 2018, 19(1):1-12

[15] Zhang Yaogong, Liu Jiahui, Liu Xiaohu, et al. Prioritizing disease genes with an improved dual label propagation framework. BMC bioinformatics, 2018, 19(1),

[16] Yang Kuo, Wang Ruyu, Liu Guangming, et al. HerGePred: Heterogeneous Network Embedding Representation for Disease Gene Prediction. IEEE journal of biomedical and health informatics, 2019, 23(4):1805-1815

[17] Pletscher-Frankild Sune, Pallejà Albert, Tsafou Kalliopi, et al. DISEASES: Text mining and data integration of disease–gene associations. Methods, 2015, 74:83-89

[18] Jiyoung Kim, Jengmin Kang, Lim Kang Ye, et al. Ketohexokinase-A acts as a nuclear protein kinase that mediates fructose-induced metastasis in breast cancer. Nature communications, 2020, 11(1):5436

[19] Liu Shuhui, Zhang Yupei, Shang Xuequn, et al. ProTICS reveals prognostic impact of tumor infiltrating immune cells in different molecular subtypes. Briefings in bioinformatics, 2021, 22(6):bbab164

[20] Xu Liwen, Deng Chunyu, Pang Bo, et al. TIP: A Web Server for Resolving Tumor Immunophenotype Profiling. Cancer Research, 2018, 78(23):6575-6580

[21] Liu Xiaoping, Wang Yuetong, Ji Hongbin, et al. Personalized characterization of diseases using sample-specific networks. Nucleic acids research, 2016, 44(22):e164

[22] Kuijjer Marieke Lydia, Tung Matthew George, Yuan Guocheng, et al. Estimating Sample-Specific Regulatory Networks. iScience, 2019, 14:226-240