This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI: 10.1109/TIP.2019.2920514, IEEE/ACM Transactions on Computational Biology and Bioinformatics

An effective mixed-model for screening differentially expressed genes of breast cancer based on LR-RF

Mengmeng Sun[1],Tao Ding[1], Xu-Qing Tang[1] and Keming Yu[2]

1. School of Science, Jiangnan University, Wuxi, China

2. Department of Mathematics, Brunel University London, London, UK

Corresponding Author: Xu-Qing Tang, E-mail: txq5139@jiangnan.edu.cn.

Abstract: **Objective** To screen differentially expressed genes quickly and efficiently in breast cancer. **Methods** Two gene microarray datasets of breast cancer, GSE15852and GSE45255, were downloaded from GEO. Microarrays with 182 breast cancer patients and 43 normal breast tissues, containing 22215 genes samples, were used for screening differentially expressed genes. Based on the Bonferroni test of FWER error measure, by combining the Logistic Regression and Random Forest model, we proposed a novel method Logistic Regression-Random Forest (LR-RF) to select differentially expressed genes on microarray data. **Results** After analyzing large samples of microarray data, we have identified 40 differentially expressed genes in breast cancer tissues. The further verified by the experiments illustrate that the average prediction accuracy rate of the LR-RF method is higher than Logistic Regression and Random Forest, and up to 93.11%. What's more, the variance of the LR-RF is as low as 0.00045, which shows the method we proposed has great facility in selecting differentially expressed genes. Besides, through analyzing the gene interaction networks of the top 20 genes we have selected, most of them have been involved in the development of breast cancer. **Conclusion** All of these results demonstrate the reliability and efficiency of LR-RF, and it is anticipated that LR-RF will be useful in providing new knowledge for biologists, medical scientists, and cognitive computing researchers to focus on finding a disease-related gene of breast cancer.

Index Terms: breast cancer, differentially expressed genes, Logistic Regression-Random Forest, Bonferroni test, gene interaction networks.

1 Introduction

Breast cancer is the most common cancer in women worldwide. Across the globe,

breast cancer is the second most common type of cancer and the second leading cause of cancer death in women. According to the latest statistics, every 26 seconds, there is a woman diagnosed with breast cancer. New breast cancer worldwide each year is up to 1.2 million, with an average annual increase in 500,000.[1]So to a certain extent, finding the cancer-causing genes is urgently needed in many ways including biomedical, bioinformatics for the prevention of breast cancer, early diagnosis and related treatment.

With the rapid development of sequencing technologies, a large amount of biological information has been stored in the gene expression data. It is crucial to analyze these gigantic data and extract the useful information from it. Gene chip, also known as DNA microarray, is one of the most important technologies in the field of life science research.[2-4] In fact, based on the extensive application of gene chip technology, the network of public databases in the growing gene chip expression data can provide an enormous powerful tool for breast cancer gene expression analysis. At the same time, the abnormal expression of polygene is the crucial biological factor of the occurrence and progression of breast cancer. Through the analysis of differential expression genes of breast cancer and its interaction networks, it is of practical significance to study the pathogenesis of breast cancer deeply, guide the individual treatment and improve the prognosis of breast cancer patients.

Actually, one of the important tasks of gene chip profiling data analysis is to screen for differentially expressed genes. For example, comparing the differences in gene transcription and expression between normal and disease states as to study the pathogenesis of the disease, doctors can conduct early diagnosis and treatment of the disease, and even predict the prognosis for patients.

The methods of screening differentially expressed genes for gene expression profiles are SAM (significance analysis of microarrays)[5], two-sample t-test[6], and so on. All of them are suitable for different study design and data type gene expression profiles to screen differentially expressed genes. However, false positive of the differential expression genes screened out by the SAM and two samples of t-test methods is too high. In fact, earlier researchers have tried to select differentially expressed genes by Logistic Regression (LR)[7-9], or Random Forest (RF).[10-12] Although LR is one of the classical methods and has been widely used for classification, the traditional LR model employs all (or most) variables for predicting and the response Y is a binary variable taking only two possible values that may influence the screening of differentially expressed genes. Meanwhile, some researchers applied RF to gene selection and classification from microarray data. Because of the gigantic gene data, the model did not pre-select genes to reduce the dataset dimensions and it would cause over fitting.

To find differentially expressed genes related to breast cancer, this paper proposes a methodology for finding such differentially expressed genes from

microarray data. We pre-select genes by LR and get a series of differentially expressed genes based on the Bonferroni test[13] of the Family Wise Error Rate(FWER)[14]error measure firstly. After that, best-related genes of differentially expressed of breast cancer will be identified by RF algorithm. Further studies by analyzing gene interaction networks eventually verified the veracity and flexibility of modeling that LR-RF has an excellent performance for identifying differentially expressed genes.

2 Materials and methods

2.1 Materials

In this paper, we have downloaded two sets of breast cancer datasets from Gene Expression Omnibus (GEO).[15]The accession numbers were GSE15852 and GSE45255, and the chip platform was GPL96. DatasetGSE15852included 43 paired normal tissues and breast cancer tumors. DatasetGSE45255 consisted of 139 breast cancer tumors. By integrating data, there were 182 cases of breast cancer tumors and 43 cases of normal breast tissues for subsequent analysis and each of them contains 22215 genes samples.

It is generally known that each gene has a different expression level, and gene expression value depends on the microarray data due to differences in the experiments. Thus, we normalize the microarray data by MAPMINMAX function in the MATLAB. The MAPMINMAX processes matrices by normalizing the minimum and maximum values of each row to [$y_{\min}$, $y_{\max}$].The formula is:

$$y = (y_{\max} - y_{\min}) \times \frac{x - x_{\min}}{x_{\max} - x_{\min}} + y_{\min} \tag{1}$$

In this study, we set $y_{\min} = 0$ and $y_{\max} = 1$. In other words, we standardize the data to (0, 1).

2.2 Methods

In this section, the Bonferroni test of the Family Wise Error Rate (FWER) measure is briefly reviewed firstly. Further, we introduce the typical LR model and RF model. Lastly, a novel method LR-RF is proposed to screen differentially expressed genes. For all of the microarray data, we pre-select genes by LR method to reduce the dataset dimensions and then use a RF classifier to identify cancer-causing genes.

2.2.1 Bonferroni test of FWER error measure

In the process of the regression analysis, when one assumption is made, the reliability of the results is higher. When the continuous tests are made, the reliability of the results will be greatly reduced, resulting in the increase of the false positives.

When testing multiple hypotheses, the situation becomes much more complicated. Now each test has type I and type II errors, and it becomes unclear how we should measure the overall error rate. The first measure to be suggested was the (FWER), which is the probability of making one or more type I errors among all the hypotheses. In this paper, the Bonferroni test based on the FWER measure was used to screen the differentially expressed genes.

The Bonferroni inequality[16] is often used when conducting multiple tests of significance to set an upper bound on the overall significance level α .If $T_1,\ldots,T_n$ is a set of n statistics with corresponding p-values $P_1,\ldots,P_n$ for testing hypothesis $H_1,\ldots,H_n$, the classical Bonferroni multiple test procedure is usually performed by rejecting $H_0=\{H_1,\ldots,H_n\}$ if any p-values is less than α/n .Furthermore the specific hypothesis H_i is rejected for each $P_i\le\alpha/n$ ($i=1,\ldots,n$).The Bonferroni inequality is in the following.

$$\begin{aligned} FWER &= \Pr(V\ge 1)=\Pr\left(\bigcup_i\left\{H_i=0,P_i<\frac{\alpha}{n}\right\}\right) \\ &\le\sum_{i=1}^{n}\Pr\left(P_i<\frac{\alpha}{n}\mid H_i=0\right)\Pr(H_i=0)\le n\cdot\frac{\alpha}{n}=\alpha \end{aligned} \tag{2}$$

It is ensure that the probability of rejecting at least one hypothesis when all true is no greater than α .

2.2.2 Logistic Regression model

Logistic Regression (LR)[17] models are perhaps the most widely used models in the generalized linear models family. A LR model is used when the response Y is a binary variable taking only two possible values (say, 0 or 1), which is suitable for the study of this article's normal and tumor the two classifications.

The form of the logistic function is

$$f(x)=\frac{e^x}{1+e^x} \tag{3}$$

where $f(x)$ is a continuous curve limited to the [0, 1] interval. Because the dependent variable Y only takes 0, 1 two discrete values, it is not suitable for the

regression model as a dependent variable. The basic idea of LR is that it does not regress to Y directly, but rather defines a probability function:

$$\pi = \Pr(Y = 1 \mid X_1 = x_1, X_2 = x_2, \cdots, X_i = x_i) \tag{4}$$

where there requires $0 \le \pi \le 1$. Directly seeking the expression of π is a very difficult thing, so, we consider

$$\frac{1-\pi}{\pi} = \frac{P(Y \neq 1)}{P(Y = 1)} = k \tag{5}$$

In general, $0 < k < +\infty$ .Then, let

$$\pi = \Pr(Y = 1 \mid X_1 = x_1, X_2 = x_2, \cdots, X_i = x_i) = \frac{1}{1 + a \cdot e^{-b_1 X_1 - \cdots - b_n X_i}} \tag{6}$$

$$(a > 0, b_n \ge 0)$$

π is a Logistic type of function. Then, we deform it and get a new function as follow:

$$\lg(\frac{1-\pi}{\pi}) = b_0 - b_1 x_1 - \cdots - b_n x_i \tag{7}$$

In this article, the regression is relatively column by column, so the probability equation is:

$$\pi = \Pr(Y = 1 \mid X_1 = x_i) = \frac{1}{1 + a \cdot e^{-b_1 X_1}} \ (a > 0, b_1 \ge 0) \tag{8}$$

where X_i is the *i-th* gene value information, and π is the probability of being sick.

2.2.3 Random Forest model

Random Forest(RF) is an algorithm for classification developed by Leo Breiman[18] that uses an ensemble of classification trees. Each of the classification trees is built using a bootstrap sample of the data, and each split the candidate set of variables is a random subset of the variables. Thus, RF uses both bagging (bootstrap aggregation), a successful approach for combining unstable learners, and random variable selection for tree building. Each tree is unpruned (grown fully) to obtain low-bias trees. At the same time, bagging and random variable selection result in the low correlation of the individual trees. The algorithm yields an ensemble that can achieve both low bias and low variance.

2.2.4 Logistic Regression-Random Forest model

The Logistic Regression-Random Forest (LR-RF) model we proposed is divided into two steps. Firstly, we pre-select genes by LR, based on the Bonferroni test of FWER error measure, and we get a series of differentially expressed genes roughly. Then, we use the RF algorithm for the second screening and get the top potential

genes related to breast cancer. As a matter of fact, RF can identify which genes were important in building a forest of trees; we can get the genes' importance score ranking. The importance of a gene is determined whether it is used in the model. So setting a threshold and determining the importance, we delete any genes with an importance below the threshold. Finally, we get the differentially expressed genes in breast cancer.

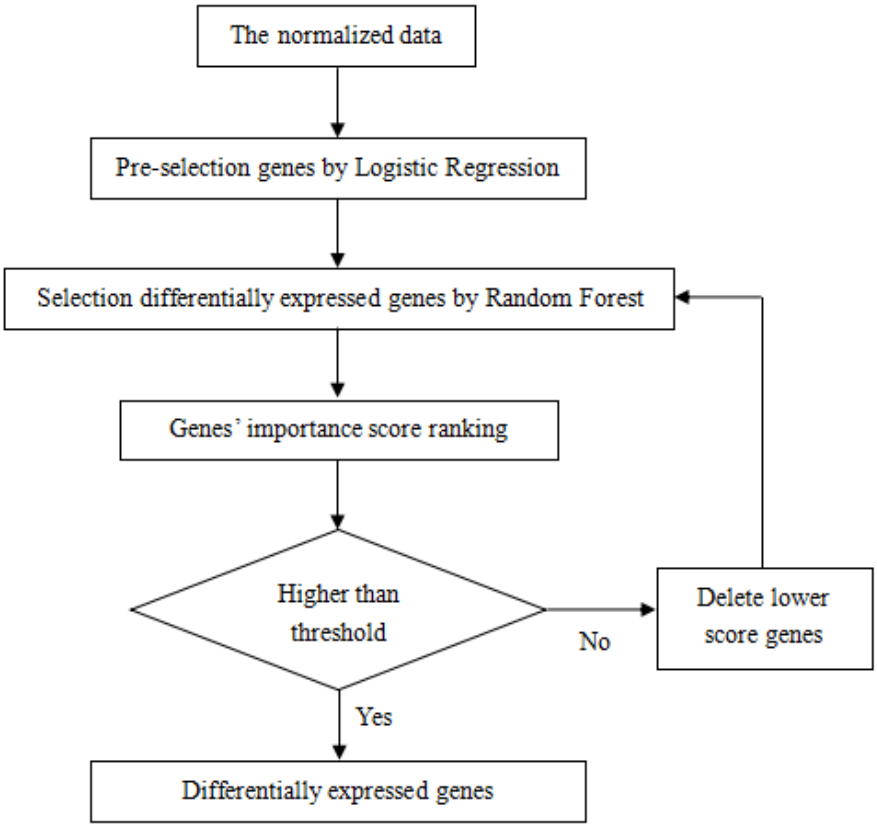

Fig1 the process for LR-RF model

3 Results and analysis

3.1 Differentially expressed genes

For these three methods, we conducted 10 random tests and each test took 80% of the data as a training set, and the remaining 20% data as a test set.

By screening, each method yielded a different number of differentially expressed genes, as showed in the following table. In this study, we set 0.1 as the threshold when use the Random forest method to select genes.

Table1 the number of differentially expressed genes selected by the three methods

Method	1	2	3	4	5	6	7	8	9	10
LR	102	240	372	143	371	629	197	257	233	288

RF	22	33	19	24	27	14	17	25	23	20
LR-RF	102	188	158	133	143	107	174	142	163	137

The table clearly shows that a large number of genes selected by LR method are useless genes. Although the RF method can screen out few differentially expressed genes, model may cause over fitting and then loss quite important genes due to the high dimensionality of the microarray data,. In contrast, the LR-RF method can select important genes form the cancer-causing genes that have been pre-selected by the LR method, and guarantee the assurance of the veracity of identifying differential expression genes according to RF model.

3.2 Stability analysis of the three methods

We use 80% of the data as a training set, and 20% of the data as the test set. Using the differential expression genes screened to predict whether 20% of the samples are breast cancer patients. In this paper, the Rand index[19] is applied to calculate the accuracy rate.

a：The patient is predicted to be a patient.

b：The patient is predicted to be normal.

c：The normal person is predicted to be a patient.

d：The normal person is predicted to be normal.

$$\text{The accuracy rate}=\frac{a+d}{a+b+c+d}.$$

Through validating and comparing the models, we can get the accuracy of the three methods and evaluate the stability of the method by using the variance of accuracy. The smaller the variance, the more stable the method.

Table2 the accuracy rate and stability of the three methods

Method	1	2	3	4	5	6	7	8	9	10	average	variance
LR	90.2	91.9	89.2	93.0	96.3	85.4	96.2	96.0	93.3	92.2	92.3	0.0011
RF	84.4	82.2	91.1	88.9	88.9	82.2	91.1	86.7	84.4	80.0	85.9	0.0014
LR-RF	95.60	93.3	91.1	91.1	93.3	95.6	93.3	93.3	95.6	88.9	93.1	0.00045

Seen from the above table, the average prediction accuracy rate of the LR-RF method is 93.11%, which is higher than LR and RF methods. The variance of the LR-RF method is 0.00045, and the variances of the LR and RF are 0.0011, 0.0014, respectively. LR-RF method's variance is smaller than the other methods' variances. Obviously, comparing with the other two methods, the method LR-RF we proposed is more stable on the premise that selecting differentially expressed genes effectively.

3.3 Hierarchical clustering analysis

We took the union of the differentially expressed genes selected byLR-RF methods ten times, and the clustering analysis was performed by using R software. The hierarchical clustering chart was as follows:

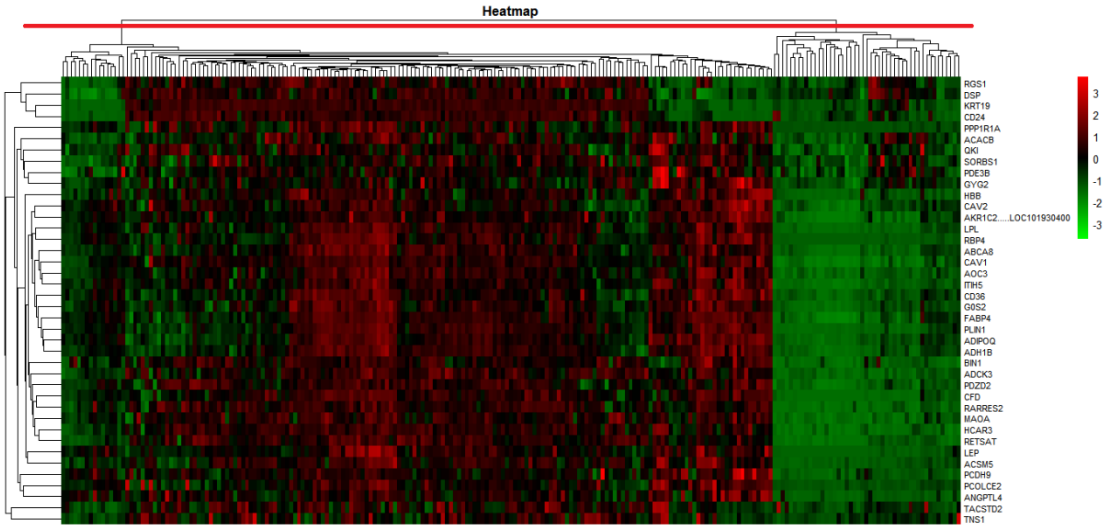

Fig2 The heat map of cluster analysis of differentially expressed genes of 139 breast cancer and 43 normal breast specimens.

The 225 samples are clustered into two groups using differentially expressed genes selected by the LR-RF method. The 139 samples ahead of the figure are breast cancer patients and the remaining samples are 43 normal people. It can be seen from the figure that a significant difference between the two groups. This indicates these genes can distinguish the normal samples from and patient ones.

3.4 The analysis of genes interaction networks

Top 20 differentially expressed genes related to breast cancer selected by our method is listed in Table3. The most exciting results from the table, majority of genes had been described in accumulating papers of breast cancer.

Table 3 The top 20 differentially expressed genes in breast cancer

No.	Gene ID	Gene Symbol	Average Importance	No.	Gene ID	Gene Symbol	Average Importance
1	201650	KRT19	1.155279813	11	218168	ADCK3	0.399596245
2	209493	PDZD2	0.641707676	12	216379	CD24	0.383872561
3	209763	CHRDL1	0.57484195	13	215695	GYG2	0.37417914
4	206488	CD36	0.567017651	14	214439	BIN1	0.374115711
5	211696	HBB	0.53992312	15	43427	ACACB	0.360580478
6	207092	LEP	0.489604283	16	218723	RGCC	0.34986191
7	205478	PPP1R1A	0.471361386	17	210201	BIN1	0.342833509
8	203548	LPL	0.436124485	18	219140	RBP4	0.332710044

9	203853	GAB2	0.420885256	19	221009	ANGPTL4	0.31555432
10	209699	AKR1C2	0.408219183	20	204894	AOC3	0.297631023

Previous research has identified a highly deregulated gene Keratin 19(KRT19) in breast cancer. Furthermore, KRT19 expression was associated with breast tumor subtyping, among estrogen receptor(ER) and Luminal B, and KRT19 expression correlated with poor overall survival.[20-24]PDZ domain containing 2(PDZD2) does not possess an intrinsic enzymatic activity. It functions through a number of direct and indirect interactions with breast tumor suppressors. It inhibits the activities of P and PDZ proteins, or enhances the activity of telomerase.[25] Other studies suggest that the CD36 gene[26] is located on the chromosome 7q11.2, and its encoded protein CD36 molecule is a transmembrane glycoprotein expressed on the surface of platelets and a variety of tumor cells. Seewaldt et al.[27] found that inhibition of CD36 gene expression in normal breast cells can lead to a decrease in adipocyte surrounding cells and an increase in extracellular matrix collagen deposition, which is a key factor in increased mammalian gland density, and the lack of CD36 gene infection may be an important event in the early development of breast cancer.

To further demonstrate the predictive ability of LR-RF, we annotated the differentially expressed genes screened by LR-RF method to Gene MANIA database and found that most of these genes were significantly enriched in pathways related to the breast cancer, such as pathways in cancer, adipocytokine-signaling pathway, neurotrophin signaling pathway. This suggests that they play important roles in the cancers.

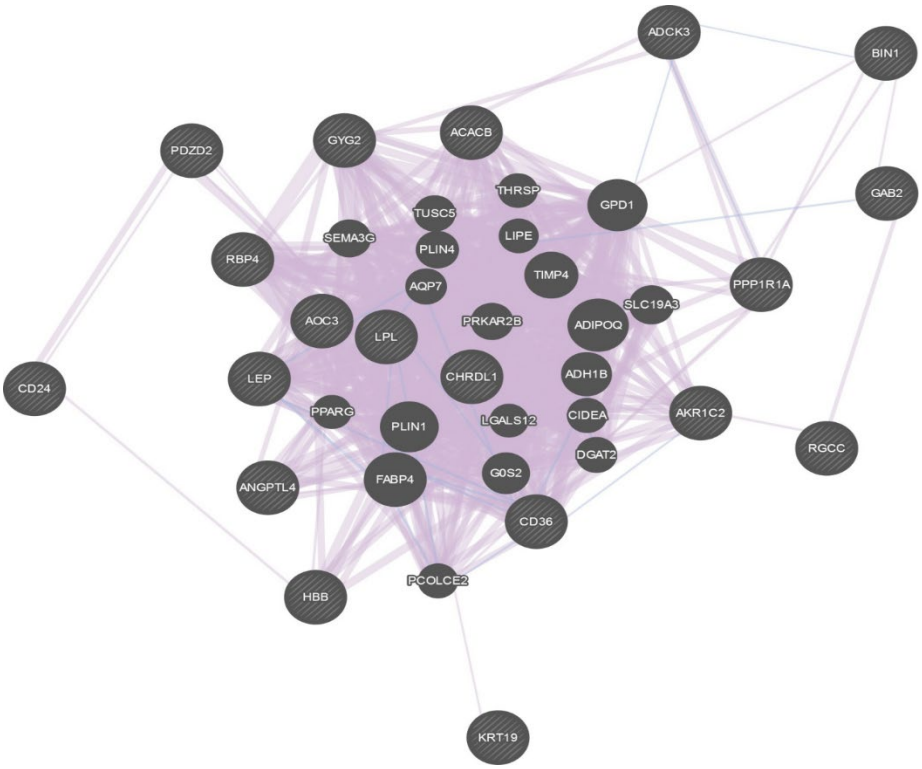

Fig3 Network of the differentially expressed genes constructed by Gene MANIA

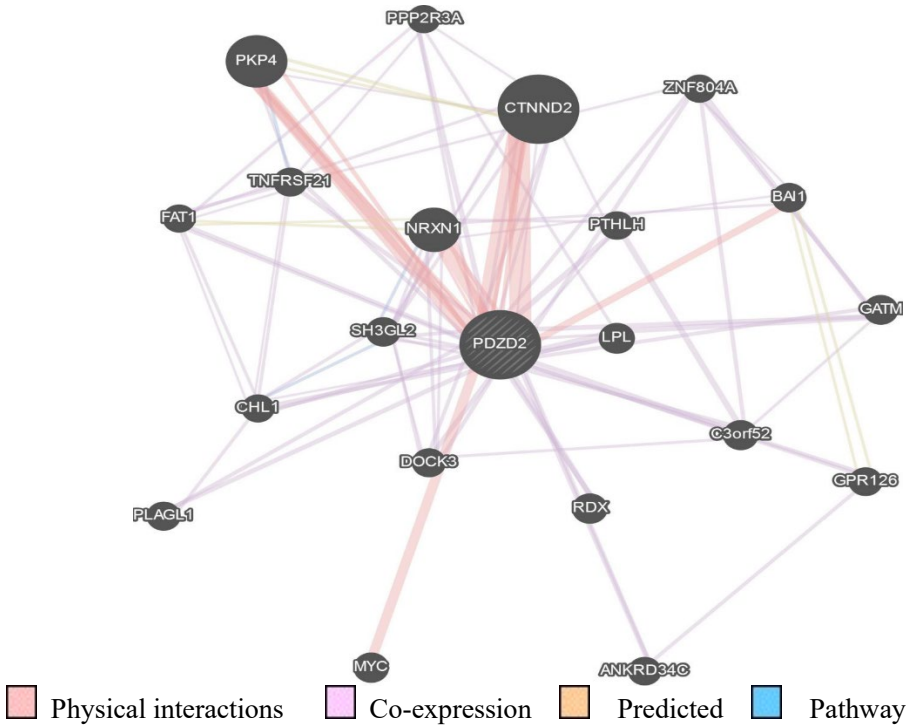

Fig4 Network of the PDZD2 gene constructed by Gene MANIA

From the network fig4 of the differentially expressed genes, the differentially expressed genes are involved in many known pathways and harbor many physical interactions. From the two figures we can see, these differentially expressed genes are densely connected which several, such as PDZD2, LPL and CD36 have been confirmed to be closely related to breast cancer. Fig4 also shows that the interaction network of PDZD2 gene with other genes can be seen clearly.

4 Conclusions

It is well known that many cancer-causing genes of breast cancer are unknown yet. With the availability of huge DNA microarray data, the selection of differentially expressed genes by bioinformatics methods is especially important. We have found candidate genes related to breast cancer based on microarray data and proposed a method for screening differentially expressed genes of breast cancer by combining LR and RF as a machine-learning technique. From 22215 genes samples of breast cancer, we pre-select a series of differentially expressed genes, and then screen breast cancer genes again. Logistic Regression and Random Forest have been trained before and after screening differentially expressed genes. This method greatly improves not only the accuracy of the screening of cancer-causing genes but also the speed.

By ten times random experiments we used two microarray datasets related to breast cancer to measure the stability of methods. We used variance to check the

stability of the method. The analyses show that our proposed method can produce **almost** similar pattern of **results for all** selections considered, and success in selecting the differentially expressed genes. We focused on the top 20 differentially expressed genes ranked by the results.

Although the results include genes that have not been identified, the proposed mixed model LR-RF make it possible to screen differentially expressed genes related to breast cancer efficiently in short time. It is anticipated that LR-RF will be useful in providing new knowledge for biologists, medical scientists, and cognitive computing researchers to focus on finding a disease-related gene of breast cancer.

Acknowledgements

This work was supported by two projects from the National Natural Science Foundation of China (Grand No 11371174 and 11271163).

Reference

1 Torre, L. A.*et al.* Global cancer statistics, 2012. *Ca A Cancer Journal for Clinicians***65**, 87-108 (2015).

2 Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science***270**, 467-470 (1995).

3 Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics***20**, 307-315 (2004).

4 Irizarry, R. A.*et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research***31**, 15 (2003).

5 Grace, C. & Nacheva, E. P. Significance Analysis of Microarrays (SAM) Offers Clues to Differences Between the Genomes of Adult Philadelphia Positive ALL and the Lymphoid Blast Transformation of CML. *Cancer Informatics***11**, 173-183 (2012).

6 Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences***99**, 6567-6572 (2002).

7 Shevade, S. K. & Keerthi, S. S. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics***19**, 2246-2253 (2003).

8 Zhu, J. & Hastie, T. Classification of gene microarrays by penalized logistic regression. *Biostatistics***5**, 427 (2004).

9 Liang, Y.*et al.* Sparse logistic regression with a L 1/2 penalty for gene selection in cancer classification. *BMC Bioinformatics***14**, 198 (2013).

10 Deng, H. & Runger, G. Gene selection with guided regularized random forest ☆.*Pattern Recognition***46**, 3483-3489 (2013).

11 Anaissi, A., Kennedy, P. J., Goyal, M. & Catchpoole, D. R. A balanced iterative random forest for gene selection from microarray data. *BMC Bioinformatics***14**, 261 (2013).

12 Nishiwaki, K., Kanamori, K. & Ohwada, H. in*IEEE International Conference on Cognitive Informatics & Cognitive Computing.* 542-546.

13 Glickman, M. E., Rao, S. R. & Schultz, M. R. False discovery rate control is a

recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology***67**, 850-857 (2014).

14 Peña, E. A., Habiger, J. D. & Wu, W. Classes of multiple decision functions strongly controlling FWER and FDR. *Metrika***78**, 563 (2015).

15 T, B.*et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research***41**, 991-995 (2013).

16 Hommel, G. Hommel, G.: A stagewise rejective multiple test procedure on a modified Bonferroni test. Biometrika 75, 383-386. *Biometrika***75**, 383-386 (1988).

17 Budimir, M. E. A., Atkinson, P. M. & Lewis, H. G. A systematic review of landslide probability mapping using logistic regression. *Landslides***12**, 419-436 (2015).

18 Breiman, L. Random Forests. *Machine Learning***45**, 5-32 (2001).

19 Steinley, D. Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods***9**, 386-396 (2004).

20 Prat, A.*et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Research***12**, R68 (2010).

21 Lehmann, B. D.*et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation***121**, 2750-2767 (2011).

22 Kabir, N. N., Rönnstrand, L. & Kazi, J. U. Keratin 19 expression correlates with poor prognosis in breast cancer. *Molecular Biology Reports***41**, 7729 (2014).

23 Dai, X., Li, Y., Bai, Z. & Tang, X. Q. Molecular portraits revealing the heterogeneity of breast tumor subtypes defined using immunohistochemistry markers. *Scientific Reports***5**, 14499 (2015).

24 Li, Y., Tang, X. Q., Bai, Z. & Dai, X. Exploring the intrinsic differences among breast tumor subtypes defined using immunohistochemistry markers based on the decision tree. *Scientific Reports***6**, 35773 (2016).

25 Tam, C. W., Liu, V. W., Leung, W. Y., Yao, K. M. & Shiu, S. Y. The autocrine human secreted PDZ domain-containing protein 2 (sPDZD2) induces senescence or quiescence of prostate, breast and liver cancer cells via transcriptional activation of p53. *Cancer Letters***271**, 64 (2008).

26 Armstrong, L. C. & Bornstein, P. Thrombospondins 1 and 2 function as inhibitors of angiogenesis. *Matrix Biology***22**, 63 (2003).

27 Koch, M.*et al.* CD36-mediated activation of endothelial cell apoptosis by an N-terminal recombinant fragment of thrombospondin-2 inhibits breast cancer growth and metastasis in vivo. *Breast Cancer Research and Treatment***128**, 337-346 (2011).