RESEARCH ARTICLE

Multilayer network analysis of miRNA and protein expression profiles in breast cancer patients

Yang Zhang[1], Jiannan Chen[1], Yu Wang[1], Dehua Wang[1], Weihui Cong[1], Bo Shiun Lai[2], Yi Zhao[1]*

1 Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong, China, **2** Johns Hopkins University School of Medicine, Baltimore, Maryland, United States

* zhao.yi@hit.edu.cn

Abstract

MiRNAs and proteins play important roles in different stages of breast tumor development and serve as biomarkers for the early diagnosis of breast cancer. A new algorithm that combines machine learning algorithms and multilayer complex network analysis is hereby proposed to explore the potential diagnostic values of miRNAs and proteins. XGBoost and random forest algorithms were employed to screen the most important miRNAs and proteins. Maximal information coefficient was applied to assess intralayer and interlayer connection. A multilayer complex network was constructed to identify miRNAs and proteins that could serve as biomarkers for breast cancer. Proteins and miRNAs that are nodes in the network were subsequently categorized into two network layers considering their distinct functions. The betweenness centrality was used as the first measurement of the importance of the nodes within each single layer. The degree of the nodes was chosen as the second measurement to map their signalling pathways. By combining these two measurements into one score and comparing the difference of the same candidate between normal tissue and cancer tissue, this novel multilayer network analysis could be applied to successfully identify molecules associated with breast cancer.

Introduction

Breast cancer is the second leading cause of cancer death among women and results in millions of new cases every year [1]. Often assuming regulatory roles in eukaryotic cells, miRNAs are small, non-coding RNAs of roughly 20~22 nucleotides that can bind to and inhibit protein coding mRNAs [2]. The expression profiles of miRNAs are correlated with cancer type, stage, and other clinical variables [3]. Therefore, miRNA expression profiling could be a useful tool for cancer diagnosis and prognosis. MiRNAs play important roles in almost all aspects of cancer biology, including proliferation, apoptosis, tissue invasion, metastasis, and angiogenesis [4]. miRNAs also play important roles in toxicogenomics and may explain the relationship between toxicant exposure and tumorigenesis. Previous work has identified 63 miRNA genes shown to be epigenetically regulated in association with 21 diseases, including 11 cancer types [4]. Many

OPEN ACCESS

Citation: Zhang Y, Chen J, Wang Y, Wang D, Cong W, Lai BS, et al. (2019) Multilayer network analysis of miRNA and protein expression profiles in breast cancer patients. PLoS ONE 14(4): e0202311. https://doi.org/10.1371/journal.pone.0202311

Editor: Irene Sendiña-Nadal, Universidad Rey Juan Carlos, SPAIN

Received: July 26, 2018

Accepted: March 19, 2019

Published: April 4, 2019

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The project is supported by the National Natural Nature Science Foundation of China under Project No. 61573119 to Y Zhao and a Fundamental Research Project of Shenzhen under Project Nos. JCYJ20170307150444573 to Y Zhang, JCYJ20180306172131515 to Y Zhang, JCYJ20170307151313122153 to Y Zhao, and JCYJ20150625142543468 to Y Zhao.

Competing interests: The authors have declared that no competing interests exist.

proteins have known oncogenic properties that contribute to tumorigenesis. Therefore, proteomics data could also be used to study the characteristics and observe the presence of cancer [5].

In recent years, there is growing interest to investigate the role of mircoRNA (miRNA) in normal and malignant cells. The expression profiling of miRNAs has already entered into cancer clinics as diagnostic and prognostic biomarkers to assess tumor initiation, progression and response to treatment in cancer patients[6][7][8]. The peer-reviewed scientific literatures on miRNAs in cancer are huge and their role in cancer is very diverse both in terms of the disease and experimental approaches used by the investigators.

Increasing understanding of the molecular misregulation underlying carcinogenesis had created opportunities to use miRNAs as diagnostic and prognostic indicators. Many signature miRNAs have been identified and investigated in clinical trials. Such as miR-10b in Glioma[9]; miR-29 in Head and neck squamous cell carcinoma[10]; circulating miRNAs in Ovarian cancer[11], etc.

Binding of miRNAs to mRNAs leads to destabilization or translational repression of the target mRNA, which in turn regulates the expression of protein. Multiple miRNAs and proteins known to be involved in different signalling pathways are deregulated in breast cancer. For example, by targeting the NF-κB and TGF-β pathways, miR-520/373 family is a tumor suppressor in estrogen receptor negative breast cancer [12]. Overexpression of miR-221/222 also leads to deregulation of multiple oncogenic signalling pathways [13]. Claudin-5 is involved in breast cancer cell motility through the N-WASP and ROCK signalling pathways [14]. Finally, Piwil2 plays an important role in pathways involved in proliferation and anti-apoptosis in the breast cancer stem cells [15].

MicroRNAs (miRNA) and proteins, frequently dysregulated in cancers, could serve as biomarkers for breast cancer. Therefore, identifying these miRNAs and proteins could facilitate early diagnosis of breast cancer. Machine learning plays increasingly important roles in cancer diagnosis [5]. Algorithms such as Bayes, decision tree, and support vector machine, are widely used in the classification of breast cancer [16]. Deep learning methods like convolutional neural network are also prevalent in the analysis of biopsy images of cancer [17]. Previous studies have compared the performance of various statistical methods in classifying cancer based on Mass Spectrometry (MS) spectra. These methods encompass linear discriminant analysis, quadratic discriminant analysis, k-nearest neighbour classifier, bagging and boosting classification trees, support vector machine, and random forest (RF). It has been demonstrated that RF outperforms other methods in the analysis of MS data [18]. As a result, the RF algorithm was adopted for filtering miRNAs and proteins, thereby retaining the most relevant biomarkers. Furthermore, to reduce the chance of missing important biomarkers, two established ensemble learning methods—random forests and XGBoost, were employed to complete the feature selection result.

Multilayer network, incorporating multiple channels of connectivity in a system, has been studied extensively in multiple disciplines [19][20]. However, applying multilayer network to study the biology complex systems is a recent development. Multilayer network analysis technique could integrate different layers of genomic information [21], facilitate understanding of cancer complexome [22], and find the node that plays the most central roles in the whole structure [23][24]. To improve understanding of interaction between miRNAs and cancer protein, multilayer networks consisting of protein and miRNA expression was constructed in this study. In the multilayer network, miRNAs and proteins are regarded as nodes in each layer. Both the Maximal information coefficient (MIC) values between nodes within each layer and between two separate layers were computed to determine whether there exists intralayer and interlayer edges between any two nodes under a specific threshold of MIC. This model consists of multiple subsystems and multiple connectivity layers, allowing different dynamic processes to be coupled and improving our visual understanding of multilayer systems. In particular, this biological multilayer network model exhibits the interrelationship between the miRNA and protein, thereby studying their combined action on cancer at different scales and levels.

To better understand their roles in the context of biological networks, miRNA and protein expression profile networks were constructed for both normal and breast tissues. Furthermore, due to the large number of miRNAs and proteins, in order to prevent analysis process from interference of unrelated variables, random forest model and XGBoost were applied to filter miRNAs and proteins before establishing a multilayer network. The filtered molecules were used as nodes in the network. Both threshold and MIC values between every two nodes determined the final structure of the multilayer network. Comparing the betweenness centrality of the node between health control and patient samples could lead to the novel finding of miRNAs and proteins related to cancer.

Materials and methods

Data

Experimental data were collected from the Cancer Genome Atlas/TCGA (https://portal.gdc.cancer.gov/projects/TCGA-BRCA). The cohort of TCGA study consists of 1097 patients. Among then, 1085 are females and 12 are males. 757 are white, 183 are black or African American, 61 are Asians, 1 is American Indian or Alaska native and 95 are unreported. miRNA expression data consists of 1182 tissues samples exploring expression level of 1881 different miRNAs. Among them, 1078 cases are tumor tissues and 104 cases are paracancerous normal tissues. Protein expression data consists of 925 tissues samples investigating expression level of 285 candidate proteins. 882 cases are tumor tissues and 43 cases are paracancerous normal tissues.

The miRNA expression data and protein expression data were obtained from the same patient. By considering the bias raised by different studies, it is highly recommended to use the data from the same study for the analysis. Other physiological factors in different individuals will affect the accuracy of the analysis results. In order to avoid this effect, we selected paracancerous normal tissues of the same individual as controls.

Expression data of primary solid tumor and normal tissues were categorized. For both proteins and miRNAs, candidates with expression level with a value of zero are considered as noise and then filtered in the further analysis. After filtering, the miRNA and protein expression data dimensions ultimately used for analysis were 1182×320 and 925×147 (the number of miRNA tissue samples was 1182, and the number of miRNA species was 320; the number of protein tissue samples was 925, and the number of protein species was 147).

Process overview

Schematic representation of data processing and analysis is shown (Fig 1). XGBoost and random forest algorithms were employed for feature selection, and the results were used for the subsequent processing step. Subsequently, MIC value for any two nodes was calculated, so that the weight network of expression data can be obtained. By setting specific threshold, the MIC values are then converted into Boolean variables, resulting in a complex network without weights. Finally, score related to breast cancer of each node was computed and nodes were ranked by scores.

Random forest algorithm and feature selection

Since random forest performed well on Mass Spectrometry spectra data, the same method was used for miRNA and protein expression profiles data [18]. Random forest [25], as one kind of ensemble learning method, in which each learning algorithm is a decision tree. Unlike in an

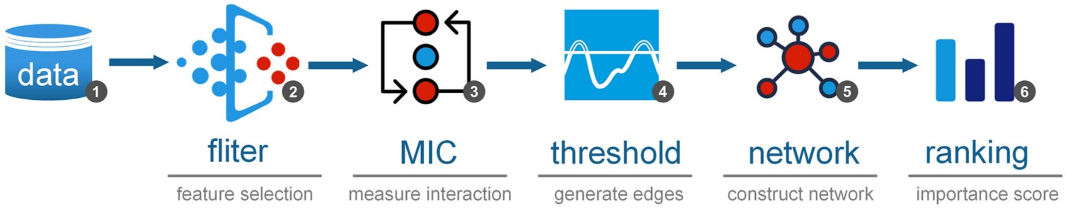

Fig 1. Schematic representation of data processing and analysis. Each icon denotes an analytical process. Icon 1 denotes data containing miRNAs and proteins. Icon 2 shows the feature selection process, which includes XGBoost and random forest algorithms. Icon 3 indicates calculation of the MIC value for every two nodes in the network, which represents the interaction between nodes. Icon 4 is the process that generates edges in the network by setting a specific threshold of MIC. Icon 5 represents the construction of a multilayer network. Icon 6 is the final step of analysis process that gives each node an importance score related to breast cancer.

https://doi.org/10.1371/journal.pone.0202311.g001

ordinary decision tree, k attributes are first selected as candidate attributes, one of which is selected to divide the tree node. Given the number of miRNAs and proteins is large, to reduce computing costs, feature selection, one of the commonly data dimension reduction methods, was applied. It is based on a criterion that selects parts of original features that can best separate different types of samples. According to the feature evaluation strategy[26], feature selection algorithm can be divided into Filter and Wrapper which are two complementary methods that were combined to characterize the molecular expression levels of normal tissues and tumor tissues. The Filter method is independent of the machine learning algorithm that was subsequently adopted. This method calculates for each feature a statistic that can represent how well a feature has distinguished the sample. On the other hand, the Wrapper method randomly selects a subset of the feature set as a temporary feature set for the random forest model, wherein the set with the smallest prediction error and fewer feature numbers serves as the final feature set.

XGBoost algorithm

XGBoost [27] is similar to Boosting for accurate classification through gradient iterations of weak classifiers. In order to efficiently retrieve the best segmentation, the training data sets are sorted before training. As both miRNA data and protein data are labeled, it belongs to supervised learning model. In XGBoost, the best model is selected by applying the accuracy rate (or error rate) and the logistic loss as evaluation criterion. Then the best prediction is achieved for the known training data set and the test data set under the optimized evaluation criterion.

Similar to random forests, each classifier is also a decision tree. Consider this similarity between random forest and XGBoost, to avoid missing important cancer-associated molecules, the results of feature selection of these two algorithms were merged into one set as the final feature set. The construction of the decision tree in the random forest algorithm is independent, however, in the XGBoost algorithm, classifiers are not independent to each other, every latter classifer is optimized based on the classifer result of the previous one. Formally, the mathematical model of XGBoost can be presented as the following formula:

$$\begin{aligned}
\hat{y}_i^{(0)} &= 0\\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i)\\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i)\\
&\cdots\\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i),
\end{aligned} \tag{1}$$

where $\hat{y}_i^{(j)}$ represents the classification result of the first j-th classifier. The XGBoost algorithm adds a new function to the original model in each iteration. The reason to add a new function is to minimize the loss of the objective function. By minimizing the following objective function $Obj(\Theta)^{(t)}$ as follow:

$$\begin{aligned} Obj(\Theta)^{(t)} &= L(\Theta) + \Omega(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^t) + \sum_{i=1}^{t} \Omega(f_i) \\ &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C \end{aligned} \tag{2}$$

where $L(\Theta)$ is the loss function to compute error of training set and $\Omega(\Theta)$ is the regularization term to control complexity of the base classifiers. We used the method of Taylor series expansion to approximate the objective function:

$$\begin{aligned} Obj(\Theta)^{(t)} &= \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \Omega(f_t) + C \\ &= \sum_{i=1}^{n} (2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t^2(x_i)) + \Omega(f_t) + C \\ &\approx \sum_{i=1}^{n} \begin{pmatrix} l(y_i,\ \hat{y}_i^{(t-1)}) + \partial_{\hat{y}_i^{(t-1)}} l(y_i,\ \hat{y}_i^{(t-1)}) f_t(x_i) \\ +0.5 \times \partial^2_{\hat{y}_i^{(t-1)}} l(y_i,\ \hat{y}_i^{(t-1)}) f_t^2(x_i) \end{pmatrix} + \Omega(f_t) + C \\ &= \sum_{i=1}^{n} (l(y_i,\ \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + 0.5 \times h_i f_t^2(x_i)) + \Omega(f_t) + C \end{aligned} \tag{3}$$

Where g_i represents the first derivative of the function $l(y_i, \hat{y}_i^{(t-1)})$, and h_i represents the second derivative of the function $l(y_i, \hat{y}_i^{(t-1)})$.

The breast cancer data used in this paper are all classified as two categories, one is tumor tissue data and the other one is paracancerous normal tissue data. Logistic function was selected as the loss function for the model. For tree structure splitting in the training, each miRNA and protein representing a leaf node will split. The training samples on each leaf node will get a probability value that is fed back through the model. The loss function is applied in the training, so the effective splitting features and optimal splitting points can be obtained by the loss function before and after the leaf nodes are splitted. As long as the sample cumulative loss function on each leaf node is minimized, the loss function of the overall sample set is minimized. At the end, the final model with the smallest loss function can be obtained.

Maximal information coefficient

Mutual information has been widely used to find non-linear relationships between two variables. Reshef [28] proposed the method of Maximal Information Coefficient (MIC) based on mutual information. The primary advantage of MIC is that a broad correlation analysis can be captured on a sufficient number of statistical samples. $M(\boldsymbol{X},\boldsymbol{Y})$ represents the population feature matrix of $\boldsymbol{X},\boldsymbol{Y}$.

$$M(\boldsymbol{X}, \boldsymbol{Y})_{s,t} = \frac{I^*((\boldsymbol{X}, \boldsymbol{Y}), s, t)}{\log\ min\{s, t\}} \quad s, t > 1 \tag{4}$$

where $I(\boldsymbol{X},\boldsymbol{Y})$ is interactive information of $\boldsymbol{X}$ and $\boldsymbol{Y}$, s,t are the number of divisions on the horizontal and vertical axes, $s\ t < n^{0.6}$ (empirical values), and n is the number of samples.

For the miRNAs and proteins screened by the algorithm above, each miRNA or protein is treated as a node, and its expression level in different patients is the attribute of the node. In this study, to measure the correlation between any two molecules in the network, the MIC values between any two nodes were calculated. The greater the MIC is, the stronger the correlation is.

Multilayer network

Multilayer network is denoted by $M = (G,C)$, where $G = \{G_\alpha;\alpha \in \{1,2,...,m\}\}$ is a set of single layer networks which is denoted as $G_\alpha = (X_\alpha,E_\alpha)$. X_α and E_α is the set of nodes and edges belongs to the layer G_α, respectively. $C = \{E_{\alpha\beta} \subseteq X_\alpha \times X_\beta;\alpha, \beta \in \{1,2,...,m\}, \alpha \neq \beta\}$ represents the set of edges that connects the nodes in different layers. Elements in C are called cross-layer connected edges. Element in E_α is called the intra-layer node connection of M. The set of nodes in layer G_α is denoted as: $X_\alpha = \{\mathrm{x}_1^\alpha, \ldots, x_{N_\alpha}^\alpha\}$, and the adjacency matrix in layer G_α is denoted as:

$$A^{[\alpha]} = (a_{ij}^\alpha) \in \mathbb{R}^{N_\alpha \times N_\alpha}, a_{ij}^\alpha = \begin{cases} 1, & (x_i^\alpha, x_j^\alpha) \in E_\alpha \\ 0, & (x_i^\alpha, x_j^\alpha) \notin E_\alpha \end{cases} \quad 1 \leq i,j \leq N_\alpha,\ 1 \leq \alpha \leq m. \tag{5}$$

The adjacency matrix in cross-layer $E_{\alpha\beta}$ is denoted as:

$$A^{[\alpha,\beta]} = (a_{ij}^{\alpha\beta}) \in \mathbb{R}^{N_\alpha \times N_\beta}, a_{ij}^\alpha = \begin{cases} 1, & (x_i^\alpha, x_j^\beta) \in E_{\alpha\beta} \\ 0, & (x_i^\alpha, x_j^\beta) \notin E_{\alpha\beta} \end{cases}. \tag{6}$$

In this study, a single layer network was established between miRNAs, and another single layer network was composed of proteins, which together constituted a two-tier multilayer network. The structure of a multi-layered network can sort out the internal interactions of the same kind of molecules while also taking into account the interactions of different kinds of molecules. Thanks to multilayer structure, in the process of cancer-associated biomarker recognition, the identification of a molecule will no longer be limited to the interaction of the same kind of molecules.

Betweenness centrality

The betweenness centrality [29] can measure the importance of nodes in the network. If the two network nodes, v_i and v_j are two non-adjacent nodes, the shortest path between them will pass through some nodes. If the certains nodes exists in many of these paths, one can infer that the node is relatively important. The betweenness centrality of node B_k is represented as:

$$\mathrm{I_{ij}}(v_k) = \left\{ \begin{array}{l} 1, v_\mathrm{k} \text{ appears in the shortest path of } v_i \text{ and } v_j \\ 0, \text{ other} \end{array} \right\}, \tag{7}$$

$$B_k = \frac{\sum\limits_{k \neq i \neq j} I_{ij}(v_k)}{N}, \tag{8}$$

where N represents the number of shortest paths. The betweenness centrality reflects the role of the node in the entire network and has a strong practical significance. In different networks, if the betweenness centrality of the same molecule is distinctly different, thereby indicating that this molecule (miRNA or protein) has played a significant role in the breast cancer. In this study we adopted the centrality function of MatLab to calculate the betweenness centrality of the nodes.

Importance score of nodes

To determine the importance score of nodes in the miRNA layer related to breast cancer, $B_{normal}^{miRNA} = \{B_{normal,k}^{miRNA}\}$ and $B_{\mathrm{cancer}}^{miRNA} = \{B_{\mathrm{cancer},k}^{miRNA}\}$ were used to represent betweenness centrality of nodes of normal tissue and cancer tissue, respectively. In miRNA layer, difference in betweenness centrality of the same miRNA belongs to different tissues is taken as importance score of nodes related to breast cancer. We standardize B_k^{miRNA} as follows:

$$\mathrm{B}_{\mathrm{k}}^{*\mathrm{miRNA}} = \frac{B_k^{miRNA} - E(B^{miRNA})}{\sigma(B^{miRNA})}, \tag{9}$$

Difference in betweenness centrality, denoted as $B^{miRNA} = \{|B_{normal,k}^{miRNA} - B_{cancer,k}^{miRNA}|\} = \{B_k^{miRNA}\}$, $\sigma(B^{miRNA})$ represents the standard deviation of the B^{miRNA} set, and $E(B^{miRNA})$ represents the mean of the B^{miRNA} set. $D_{normal,k}^{miRNA}$ represents the degree of node k in the miRNA network of normal tissue. Note that the calculation of degree here only considers cross-layer connected edges:

$$\mathrm{D}_{\mathrm{normal,k}}^{\mathrm{miRNA}} = \sum_{\mathrm{k}\neq\mathrm{j}} \mathrm{a}_{kj}^{miRNA\times\mathrm{protein}}, \mathrm{a}_{kj}^{miRNA\times\mathrm{protein}} \in \mathrm{A}^{[\mathrm{miRNA,protein}]}. \tag{10}$$

Similarly, $D_{\mathrm{cancer},k}^{miRNA}$ represents the same indicator of cancer tissue. Absolute value of difference of degree is $D^{miRNA} = \{|D_{normal,k}^{miRNA} - D_{cancer,k}^{miRNA}|\} = \{D_k^{miRNA}\}$. We standardize D_k^{miRNA} as follows:

$$\mathrm{D}_{\mathrm{k}}^{*\mathrm{miRNA}} = \frac{\mathrm{D}_k^{miRNA} - E(\mathrm{D}^{miRNA})}{\sigma(\mathrm{D}^{miRNA})}, \mathrm{D}^{miRNA} = \{\mathrm{D}_{\mathrm{k}}^{\mathrm{miRNA}}\}. \tag{11}$$

Degree distribution is often power law distribution. However, by the Jarque-Bera test, $\{B_{normal,k}^{miRNA} - B_{cancer,k}^{miRNA}\}$ can be considered as normal distribution when the MIC threshold is 0.35 (alpha = 0.05, p = 0.1592). Finally, the importance score of node k in the miRNA network is:

$$\mathrm{S}_{\mathrm{k}}^{miRNA} = |\mathrm{B}_{\mathrm{k}}^{*\mathrm{miRNA}}| + |\mathrm{D}_{\mathrm{k}}^{*\mathrm{miRNA}}|. \tag{12}$$

In the same way, calculate the score of the protein molecule as $S_k^{protein}$.

Results

Feature selection

XGBoost for feature selection. Because of the imbalance between positive and negative samples of miRNA and protein expression data, up-sampling was used to amplify positive samples. The leave-one-out method was used to train and validate the datasets. The error rate, logic loss, Root Mean Squared Error (RMSE) of the training and testing datasets (Fig 2a) gradually decrease in the model, and the Area under the Curve of ROC (AUC) (Fig 2a) gradually increases and stabilizes after 35 iterations. The error rates are 0.0005, 0.005; logical loss values 0.0098, 0.0274; the AUC values close to 1, 1; and the RMSE values 0.0293, 0.0709 in the training and testing datasets, respectively. Similarly, the XGBoost algorithm has a high accuracy for classification on miRNA expression data. Computation of importance scores (Fig 2b) through the use of XGBoost algorithm suggests that *mir.139*, *mir.21*, *mir.183*, *mir.96*, *mir.190b* and *mir.6507* are significantly associated with breast cancer.

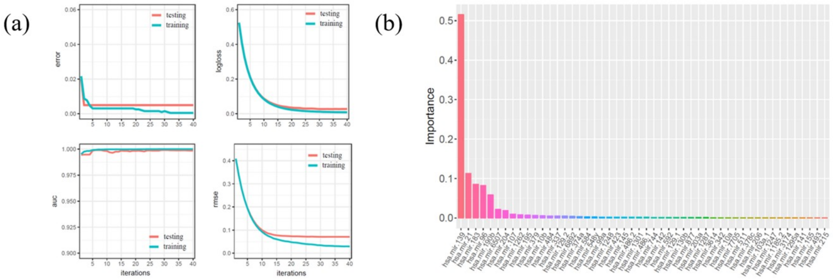

Fig 2. Analysis results of miRNAs based on XGBoost algorithm. (a) The trend of error rate, logistic loss, AUC and RMSE in the training and testing of miRNA expression data. Iteration steps (x-axis) as well as error rate, logistic loss, AUC and RMSE (y-axis in each of the four panels). Each panel has two lines representing the training set (blue) and test set (red). **(b)** Ranking of important miRNA candidates. miRNA candidates (x-axis) and the importance score of each miRNA candidate (y-axis), as determined through XGBoost algorithm, are shown.

https://doi.org/10.1371/journal.pone.0202311.g002

Error rate logic loss and RMSE of the training and testing datasets decrease in the model, whereas the AUC gradually increases then stabilizes after 30 iterations (Fig 3a). The error rates are 0, 0.0118 in the training and testing datasets, respectively; logical loss values are 0.0074, 0.04; AUC values are 1, 0.999; and the RMSE values are 0.0126, 0.0941. Similarly, the XGBoost algorithm is accurate in classifying protein expression data. Importance score as calculated by XGBoost shows that *Bax*, *GSK3.alpha.beta*, *E-cadherin*, *Rab11*, *Caveolin.1* and *Collagen_VI* contribute to the high classification accuracy of tumor and normal tissue in breast cancer (Fig 3b).

XGBoost classification algorithm further shows that some of the classified miRNA (*mir.139* [30], *mir.21* [31], *mir.96* [32], *mir.183* [33]), and protein (*Bax* [34], *GSK3* [35] and *mTOR* [35], *E-cadherin* [36], *Rab11* [37], *caveolin.1* [38]) functions are related to breast cancer.

Random forest for feature selection. To estimate the accuracy of the classification, 10-fold cross-validation method was used to assess the classification model (Table 1). When the number of selected miRNAs is 50 in the breast cancer dataset, the cross-validation accuracy rate is 98.50%.

The accuracy coefficient measures the correct rate of sample classification, and the Kappa [39] coefficient is used for checking consistency and could also measure the effect of classification accuracy. As accuracy and Kappa coefficients increase, their standard deviations decrease (Table 1).

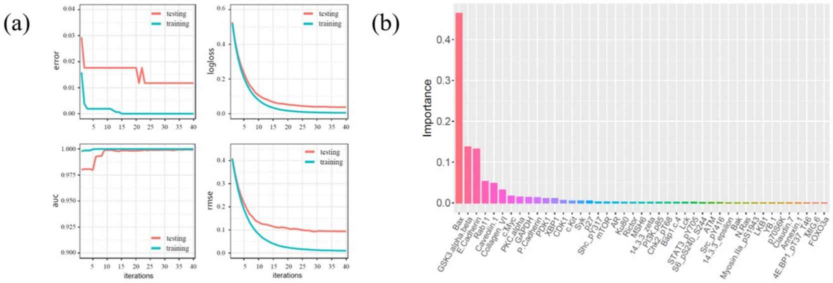

Fig 3. Analysis results of proteins based on XGBoost algorithm. (a) The trend of error rate, logistic loss, AUC and RMSE in the training and testing of protein expression data. The x-axes represent iteration steps and y-axes represent value of error rate, logistic loss, AUC and RMSE, respectively. Every subgraph has two lines represent the training set and test set, respectively. **(b)** The ranking of important variables of protein. The x-axis represents the protein molecules and y-axis represents the importance score of proteins computed by XGBoost algorithm which is different from the score at the end in this study.

https://doi.org/10.1371/journal.pone.0202311.g003

Table 1. miRNA classification results by random forest algorithm.

Number of miRNAs	Accuracy coefficient	Kappa coefficient	Accuracy coefficient SD	Kappa coefficient SD
10	0.9605	0.9208	0.0314	0.06298
20	0.9800	0.9600	0.02582	0.05164
30	0.9755	0.9508	0.02582	0.05164
50	0.9850	0.9700	0.02415	0.04830
60	0.9850	0.9700	0.02415	0.04830
70	0.9800	0.9600	0.03496	0.06992
80	0.9850	0.9700	0.02415	0.04830
100	0.9850	0.9700	0.02415	0.04830

https://doi.org/10.1371/journal.pone.0202311.t001

In this study, four cancer-associated miRNAs were screened by XGBoost algorithm, and three of them, namely *mir.21*, *mir.96*, and *mir.183*, were screened out by random forests. A comparison of the two miRNA datasets indicated that 28% of the feature selections are consistent. Similar to the analysis miRNA datasets, a 10-fold cross-validation method was used to assess the classification model to obtain protein classification (Table 2).

For breast cancer datasets, when the number selected proteins is 10, the cross-validation accuracy is 94.76%. In the 10 selected proteins, *Bax* [34], *GSK3* [35], *E-cadherin* [36], *caveolin-1* [38], *PI3K* [40], *Collagen* [41], *XBP1* [42], *Syk* [43] were found to be significantly associated with breast cancer.

Summary of feature selection. After obtaining two miRNA candidate sets and two protein candidate sets selected by two algorithms, the union of the two miRNA sets was taken as the final miRNA candidate set, in the same manner, the final proteins candidate set was obtained. The number of selected miRNA sets is 86, and the number of selected protein sets is 30.

Calculate MIC and threshold setting

As the MIC increases, the number of nodes and edges decreases. If the selected MIC threshold is so small that the number of nodes and edges in both network becomes too large, identification of nodes that have significant differences becomes more difficult. If the selected MIC threshold is so large that the network becomes too sparse, many connections are missed, which is not conducive to analyze the relationship between the nodes. MIC was calculated between any two candidates in the miRNA and protein datasets obtained through feature selection, and the threshold was set to 0.2, 0.35, and 0.5.

Under the MIC threshold of 0.5, miRNA network of cancer tissue and normal tissue was plotted (Fig 4). The cancer network (Fig 4a) is sparser than the normal network (Fig 4b). This finding indicates that the interaction of miRNA networks differs significantly by cell type and

Table 2. Protein classification results by random forest algorithm.

Number of Proteins	Accuracy	Kappa	Accuracy SD	Kappa SD
10	0.9476	0.8952	0.09024	0.1770
20	0.9342	0.8691	0.09068	0.1789
30	0.9342	0.8691	0.09068	0.1789
50	0.9231	0.8448	0.10284	0.2081
60	0.9231	0.8448	0.10284	0.2081
70	0.9231	0.8448	0.10284	0.2081
78	0.9231	0.8448	0.10284	0.2081

https://doi.org/10.1371/journal.pone.0202311.t002

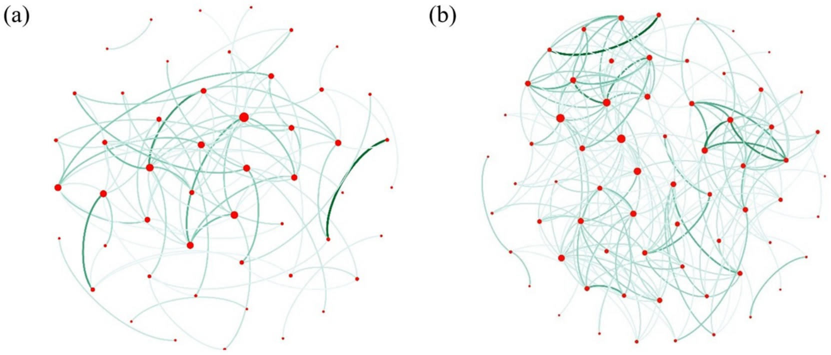

Fig 4. miRNA network of cancer tissue and normal tissue. **(a)** miRNA network of cancer tissue containing candidates with MIC greater than 0.5; **(b)** miRNA network of normal tissue containing MIC greater than 0.5. The size of the node represents node degree which is the number of connections it has to other nodes and the color darkness of the edge represents the size of the MIC value.

https://doi.org/10.1371/journal.pone.0202311.g004

supports the use of complex networks for breast cancer analysis. Similarly, under the MIC threshold of 0.5, protein network of cancer tissue and normal tissue was also plotted.

The protein network also shows the same characteristics as miRNAs, that the network of cancerous tissue is much sparser than that of normal tissue (Fig 5). Because the miRNA network of cancer cells has a small number of nodes at an MIC threshold of 0.5 and may miss some important proteins, we decided not to use this MIC threshold to construct a complex network.

Figure analysis was applied to determine which MIC threshold should be adopted. While the number of nodes varies inappreciably when MIC threshold is set to 0.35, the number of connections between nodes is significantly reduced, suggesting that these complex networks are distinct.

Several principles were considered when selecting a threshold. Firstly, the MIC threshold selected must not lead to the loss of too many nodes. For instance, in the analytical process described, less than 5% of nodes are lost. Secondly, the number of edges of the network could not be too small, and the number of edges of the two networks must be significantly different. Number of edges in the miRNA network of normal tissue (Fig 6b) is about 1.59 times that in the miRNA network of cancer tissue (Fig 6a).

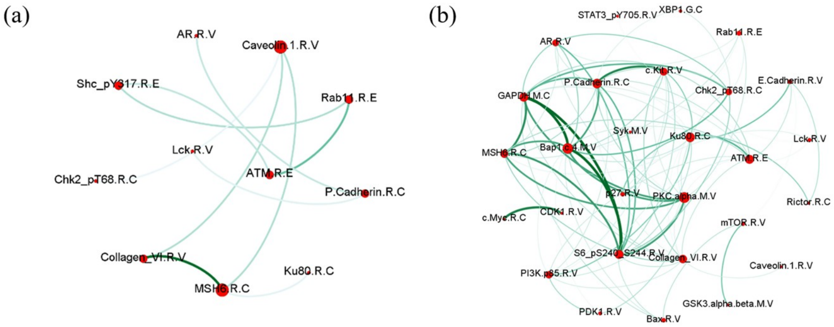

Fig 5. Protein network of cancer tissue and normal tissue. **(a)** Protein network of cancer tissue containing candidates with MIC greater than 0.5; **(b)** Protein network of normal tissue containing candidates with MIC greater than 0.5. The size of the node represents the size of the degree, and the color depth of the edge represents the size of the MIC value.

https://doi.org/10.1371/journal.pone.0202311.g005

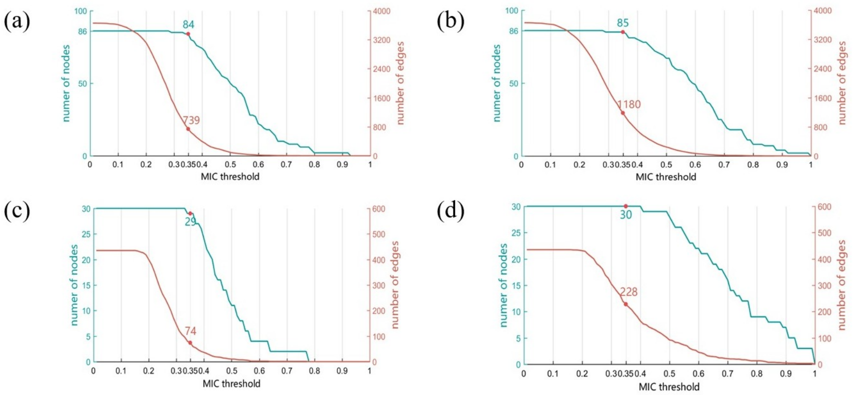

Fig 6. Relationship between MIC threshold and the number of nodes and edges. Analyses of miRNA networks of cancer tissue (a) and normal tissue (b), as well as protein networks of cancer tissue (c) and normal tissue (d) are shown. The x-axes represent threshold of MIC, y-axis to the left of each figure represents number of nodes and y-axis to the right represents number of edges in the network. Each figure shows the number of nodes or edges corresponding to a threshold of 0.35.

https://doi.org/10.1371/journal.pone.0202311.g006

Through observing the difference in structure of network under different thresholds (Fig 7), it was be found that when the MIC threshold is 0.35, the decrease in the number of network nodes is not obvious, effectively fulfilling the first principle of threshold selection. Difference in the number of edges between the two networks is also kept at a relatively high level; that is, this MIC threshold could effectively differentiate the two networks, which is in accordance with the second principle of threshold selection. Therefore, we selected 0.35 as the MIC threshold for analysis.

Multilayer network

Multilayer networks were generated after calculating MIC between nodes and setting MIC thresholds to 0.2, 0.35, or 0.5 (Fig 8).

A multilayer network with a threshold of 0.2 had more edges than other multilayer network with higher threshold, which causes the impact of key connections in the network become

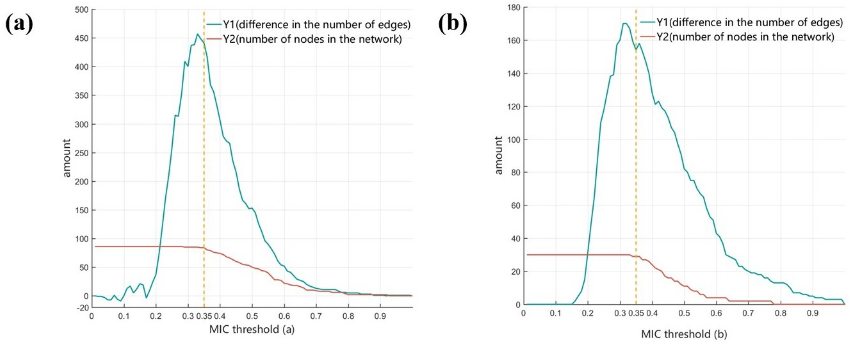

Fig 7. Relationship between MIC threshold and structure differences within the (a) miRNA network and (b) protein network. Curve Y1, which is obtained by subtracting the number of edges of the normal tissue and the cancer tissue network, represents the structural difference between the two networks. Curve Y2, which is obtained by selecting the smaller values of the number of nodes of the normal tissue and the cancer tissue network, represents the richness of the nodes of both networks. The x-axis of each figure denotes threshold of MIC and the y-axis indicates the value corresponding to Y1 and Y2.

https://doi.org/10.1371/journal.pone.0202311.g007

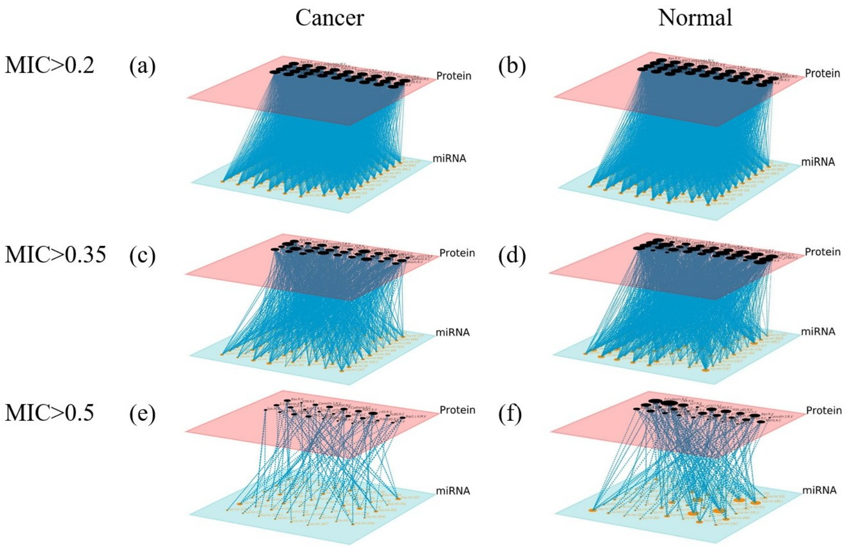

Fig 8. The multilayer network of cancer tissue and normal tissue. The multilayer networks were constructed for cancer tissue (a, c, e) and normal tissue (b, d, f) with MIC greater than 0.2 (a & b), MIC greater than 0.35 (c & d), and MIC greater than 0.5 (e & f). The red layer is the protein layer and the blue layer is the miRNA layer.

https://doi.org/10.1371/journal.pone.0202311.g008

smaller. Hence, under the threshold of 0.2, two types of cells is difficult to be distinguished well with this threshold. When the threshold is 0.5, there are obvious differences between the two multi-layer networks, but the number of edges is sparse and some important relationships may be mistakenly omitted. These problems are averted when the threshold of 0.35 is used, which affirms the use of this threshold.

Node ranking

To better understand the details of the networks, the nodes representing different candidates were ranked according to betweenness centrality and node degree (Figs 9 and 10).

Among the top 15 miRNAs, the relationships of 11 miRNAs with breast cancer in previously published studies were confirmed (Table 3).

Among the selected proteins, we were able to confirm the relationships of the top 10 proteins with breast cancer using published literature (Table 4).

Discussion

The multilayer network analysis proposed helps identify miRNAs and proteins that could be associated with breast cancer. While biomarkers were previously selected using machine learning, this study is novel in that a combination of machine learning and multilayer network methods was used.

This combinatorial approach to identify cancer biomarkers could prevent missing critical miRNA or protein candidates and ensure a more robust analysis. For example, the final ranking of nodes generated from multilayer network method in combination with machine learning differs from that using machine learning alone. This finding suggests that the combinatorial effect of multilayer network analysis and machine learning yields more comprehensive information. It also shows that the multilayer network analysis method could facilitate the discovery of novel molecular candidates.

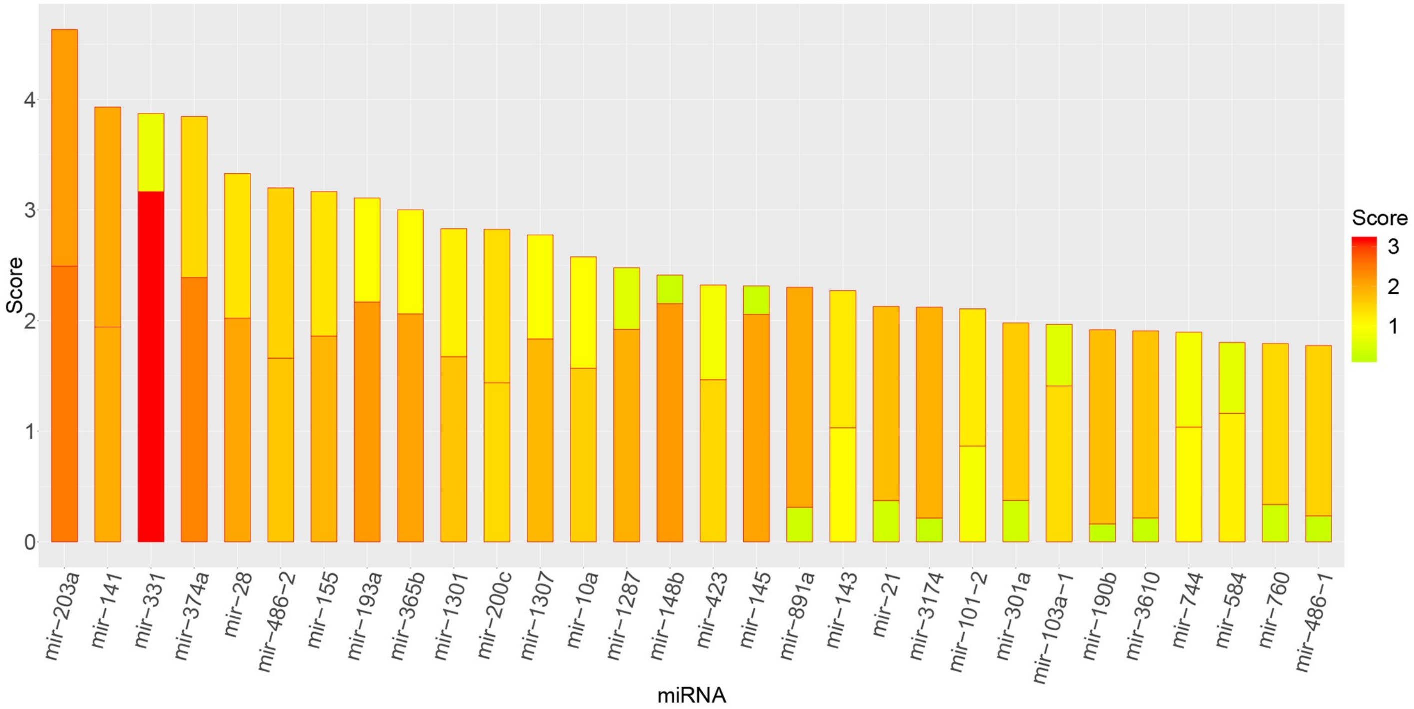

Fig 9. Score ranking of miRNAs. The horizontal axis represents each molecule (top 30 scores), and the vertical axis represents the magnitude of the importance score. The x-axis represents miRNA molecule, and y-axis represents score of miRNAs. The scores of the intra-layer edges (node betweenness) are displayed in the lower parts of the bars and the scores of the inter-layer edges (node degree) are displayed in the upper parts of the bars.

https://doi.org/10.1371/journal.pone.0202311.g009

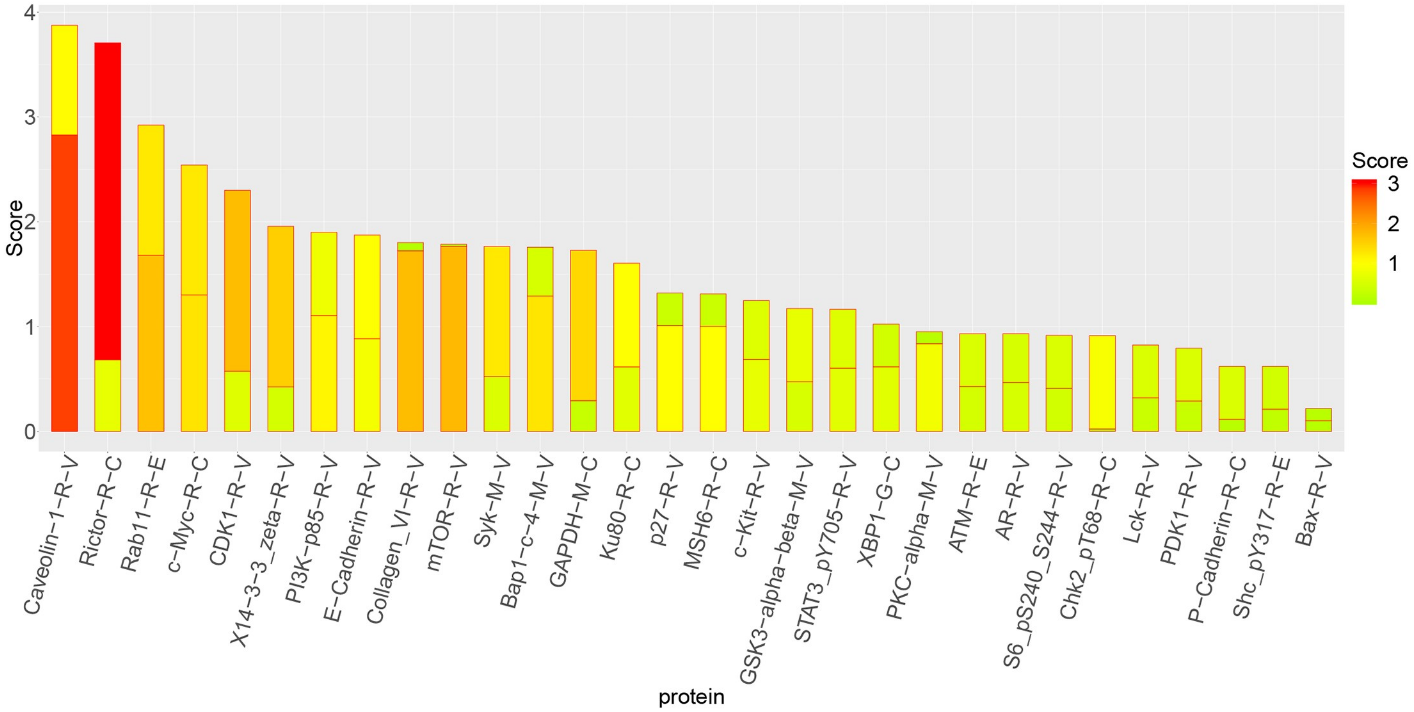

Fig 10. Score ranking of protein. The horizontal axis represents each molecule, and the vertical axis represents the magnitude of the importance score. The x-axis represents miRNA molecule and y-axes represents score of proteins. The scores of the intra-layer edges (node betweenness) are displayed in the lower parts of the bars and the scores of the inter-layer edges (node degree) are displayed in the upper parts of the bars.

https://doi.org/10.1371/journal.pone.0202311.g010

Table 3. Cancer-related miRNAs.

miRNA Candidate	Description
mir-203a [44]	Reconstitution of *Runx2* in *MDA-MB-231-luc* cells delivered with *miR-203* reverses the inhibitory effect of the miRNAs on tumor growth and metastasis.
mir-141 [45]	*miR-141* has distinct profiles in *EGF-dependent* breast cancer cell invasion, proliferation, and cell cycle progression.
mir-374a [46]	The *Wnt/β-catenin* signaling is hyperactivated in metastatic breast cancer cells that express *miR-374a*.
mir-28 [47]	In breast cancer cells, *miR-28* regulates *Nrf2* expression at the posttranscriptional level by binding to the *3′UTR* of *Nrf2* mRNA and resulting in *Nrf2* mRNA degradation.
mir-155 [48]	*miR-155* expression is upregulated in breast cancer cells, which reduces the levels of *RAD51* and affects the cellular response to ionizing radiation.
mir-193a [49]	*miR-193a* expression is downregulated in breast cancer cell lines and tissues when compared with the adjacent non-tumor tissues.
mir-365b [50]	*miR-365* expression levels are significantly higher in breast cancer tissues when compared with adjacent non-tumor tissues.
mir-1301 [51]	*miR-1301* is overexpressed in breast cancer tissues and cell lines and cell tissues, whereas downregulation of *miR-1301* inhibits the proliferation of breast cancer cells *in vitro*.
mir-200c [52]	For the claudin-low breast cancer, *miR-200c* has therapeutic effects in an *in vivo* model.
mir-10a [53]	The median expression levels of *miR-10b* in tumor tissue when compared with adjacent non-tumor tissue are significantly higher in relapsed patients than in relapse-free patients.
mir-148b [54]	*miR148b* is a major coordinator of breast cancer progression in a relapse-associated microRNA signature.

https://doi.org/10.1371/journal.pone.0202311.t003

Although published work has described miRNA or protein networks of expression profiles separately, interrelationships between two networks have not been thoroughly investigated. To address this knowledge gap, the interrelationship between miRNA and protein networks was studied through the MIC. The most suitable MIC threshold was determined by analyzing the

Table 4. Cancer-related Proteins.

Protein Candidate	Description
E-Cadherin-R-V [36]	*E-cadherin* is regulated epigenetically via methylation of the promoter in most intraductal breast carcinomas.
Caveolin-1-R-V [38]	*Caveolin-1* expression is significantly decreased in breast cancer-associated fibroblasts compared to normal fibroblasts and is associated with increased invasion-promoting capacity.
PI3K-p85-R-V [40]	The inhibition of *PI3K* promotes ER activity, as manifested by increases in ER binding to target promoters and ER target gene expression.
Collagen_VI-R-V [41]	*Collagen VI* are upregulated in breast cancer, generating a microenvironment that promotes tumour progression and metastasis.
Rictor-R-C [55]	*Rictor* expression is upregulated significantly as compared with nonmalignant tissues in invasive breast cancer specimens.
Rab11-R-E [56]	Rab coupling protein (*FIP1C*), an effector of the *Rab11* GTPases, is amplified and overexpressed in 10% to 25% of primary breast cancers.
c-Myc-R-C [57]	In the *AhR/HDAC6/c-Myc* signaling pathway, phthalates induce proliferation and invasiveness of estrogen receptor-negative breast cancer.
CDK1-R-V [58]	Combined inhibition of *Cdk1* and *PARP* in BRCA–wild-type cancer cells (breast cancer–associated cell) results in reduced colony formation, delayed growth of human tumor xenografts.
14-3-3_zeta-R-V [59]	*ErbB2* and *14-3-3* overexpression promotes cell migration and antagonizes cell adhesion.
mTOR-R-V [60]	The *PI3K/Akt/mTOR* pathway results in cell growth and tumor proliferation, and it plays a significant role in endocrine resistance in breast cancer.

https://doi.org/10.1371/journal.pone.0202311.t004 interrelationship between the MIC threshold and the number of nodes and edges of the network. By using the most optimized MIC threshold to construct the multilayer networks, miRNAs and proteins associated with breast cancer were identified. Although the top-ranked candidates for protein biomarkers were previously identified, the combinatorial approach proposed reveals potentially novel miRNAs associated with breast cancer, such as mir-331, mir-486-2, mir-1307 and mir-1287. Roles of these miRNA candidates in breast cancer will be confirmed through molecular means.

A minor drawback associated with the multilayer network analysis is that the optimized threshold value must be determined through analysis of the distribution map (Figs 6 and 7) measuring the number of edges and nodes in network under different thresholds. The approximate range can be selected but the optimal value cannot be obtained automatically. This shortcoming could be overcome by establishing algorithms that facilitate selection of optimized threshold values.

Because single network analysis provides limited information, the proposed combinatorial approach will allow for a deeper understanding of multiple networks and signaling pathway in cancer. The regulatory architecture of miRNA and protein in breast cancer patients analyzed in multiple network-wide will potentially enable novel cancer biomarker discovery.

Supporting information

S1 File. miRNA expression data of breast cancer. The first column of the data is the name of miRNAs, and first row of the data is the code of the corresponding cases of samples in TCGA, where the fields ending with '-11' is normal samples and the rest are cancer tissues.
(CSV)

S2 File. Protein expression data of breast cancer. The first column of the data represent cell type of the miRNAs, where '0' means cancer samples and '1' means normal tissues. The second column of the data is the code of the corresponding cases of samples in TCGA, and first row of the data is the name of proteins.
(CSV)

Author Contributions

Conceptualization: Yang Zhang, Yi Zhao.

Data curation: Dehua Wang, Weihui Cong.

Formal analysis: Yang Zhang, Jiannan Chen, Dehua Wang, Weihui Cong.

Funding acquisition: Yang Zhang, Yi Zhao.

Investigation: Yang Zhang.

Methodology: Jiannan Chen, Dehua Wang, Weihui Cong.

Project administration: Yang Zhang, Yi Zhao.

Software: Jiannan Chen, Dehua Wang, Weihui Cong.

Supervision: Yang Zhang, Yi Zhao.

Validation: Yang Zhang, Yu Wang.

Visualization: Yang Zhang, Jiannan Chen, Yu Wang, Dehua Wang, Weihui Cong.

Writing – original draft: Yang Zhang, Jiannan Chen.

Writing – review & editing: Yang Zhang, Yu Wang, Bo Shiun Lai, Yi Zhao.

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International journal of cancer. 2010 Dec 15; 127(12):2893–917. https://doi.org/10.1002/ijc.25516 PMID: 21351269
2. Lin S, Gregory RI. MicroRNA biogenesis pathways in cancer. Nature reviews cancer. 2015 Jun; 15 (6):321. https://doi.org/10.1038/nrc3932 PMID: 25998712
3. Chu A, Robertson G, Brooks D, Mungall AJ, Birol I, Coope R, et al. Large-scale profiling of microRNAs for the cancer genome atlas. Nucleic acids research. 2015 Aug 13; 44(1):e3-. https://doi.org/10.1093/nar/gkv808 PMID: 26271990
4. Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, et al. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. Cell reports. 2018 Apr 3; 23(1):313–26. https://doi.org/10.1016/j.celrep.2018.03.075 PMID: 29617669
5. Wisniewski JR, Ostasiewicz P, Mann M. High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. Journal of proteome research. 2011 May 18; 10(7):3040–9. https://doi.org/10.1021/pr200019m PMID: 21526778
6. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005; 435(7043):834–8. https://doi.org/10.1038/nature03702 PMID: 15944708
7. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res. 2005; 65(16):7065–70. https://doi.org/10.1158/0008-5472.CAN-05-1783 PMID: 16103053
8. Volinia S, Galasso M, Sana ME, Wise TF, Palatini J, Huebner K, et al. Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. Proc Natl Acad Sci U S A. 2012; 109 (8):3024–9. https://doi.org/10.1073/pnas.1200010109 PMID: 22315424
9. Sun L, Yan W, Wang Y, Sun G, Luo H, Zhang J, et al. MicroRNA-10b induces glioma cell invasion by modulating MMP-14 and uPAR expression via HOXD10[J]. Brain research, 2011, 1389: 9–18. https://doi.org/10.1016/j.brainres.2011.03.013 PMID: 21419107
10. Kinoshita T, Nohata N, Hanazawa T, Kikkawa N, Yamamoto N, Yoshino H, et al. Tumour-suppressive microRNA-29s inhibit cancer cell migration and invasion by targeting laminin–integrin signalling in head and neck squamous cell carcinoma[J]. British journal of cancer, 2013, 109(10): 2636. https://doi.org/10.1038/bjc.2013.607 PMID: 24091622
11. Taylor DD, Gercel-Taylor C. MicroRNA signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer[J]. Gynecologic oncology, 2008, 110(1): 13–21. https://doi.org/10.1016/j.ygyno.2008.04.033 PMID: 18589210
12. Keklikoglou I, Koerner C, Schmidt C, Zhang JD, Heckmann D, Shavinskaya A, et al. MicroRNA-520/373 family functions as a tumor suppressor in estrogen receptor negative breast cancer by targeting NF-κB and TGF-β signaling pathways. Oncogene. 2012 Sep; 31(37):4150. https://doi.org/10.1038/onc.2011.571 PMID: 22158050
13. Rao X, Di Leva G, Li M, Fang F, Devlin C, Hartman-Frey C, et al. MicroRNA-221/222 confers breast cancer fulvestrant resistance by regulating multiple signaling pathways. Oncogene. 2011 Mar; 30 (9):1082. https://doi.org/10.1038/onc.2010.487 PMID: 21057537
14. Escudero-Esparza A, Jiang WG, Martin TA. Claudin-5 is involved in breast cancer cell motility through the N-WASP and ROCK signaling pathways. Journal of Experimental & Clinical Cancer Research. 2012 Dec; 31(1):43.
15. Lee JH, Jung C, Javadian-Elyaderani P, Schweyer S, Schütte D, Shoukier M, et al. Pathways of proliferation and antiapoptosis driven in breast cancer stem cells by stem cell protein piwil2. Cancer research. 2010 May 11:0008–5472.
16. Leung MK, Delong A, Alipanahi B, Frey BJ. Machine learning in genomic medicine: a review of computational problems and data sets. Proceedings of the IEEE. 2016 Jan; 104(1):176–97.
17. Wang J, Yang X, Cai H, Tan W, Jin C, Li L. Discrimination of breast cancer with microcalcifications on mammography by deep learning. Scientific reports. 2016 Jun 7; 6:27327. https://doi.org/10.1038/srep27327 PMID: 27273294
18. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, et al. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics. 2003 Sep 1; 19 (13):1636–43. PMID: 12967959

19. De Domenico M, Solé-Ribalta A, Cozzo E, Kivelä M, Moreno Y, Porter MA, et al. Mathematical formulation of multilayer networks. Physical Review X. 2013 Dec 4; 3(4):041022.

20. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. Multilayer networks. Journal of complex networks. 2014 Sep 1; 2(3):203–71.

21. Cantini L, Medico E, Fortunato S, Caselle M. Detection of gene communities in multi-networks reveals cancer drivers. Scientific reports. 2015 Dec 7; 5:17386. https://doi.org/10.1038/srep17386 PMID: 26639632

22. Rai A, Pradhan P, Nagraj J, Lohitesh K, Chowdhury R, Jalan S. Understanding cancer complexome using networks, spectral graph theory and multilayer framework. Scientific reports. 2017 Feb 3; 7:41676. https://doi.org/10.1038/srep41676 PMID: 28155908

23. De Domenico M, Solé-Ribalta A, Omodei E, Gómez S, Arenas A. Ranking in interconnected multilayer networks reveals versatile nodes. Nature communications. 2015 Apr 23; 6:6868. https://doi.org/10.1038/ncomms7868 PMID: 25904405

24. De Domenico M, Nicosia V, Arenas A, Latora V. Structural reducibility of multilayer networks. Nature communications. 2015 Apr 23; 6:6864. https://doi.org/10.1038/ncomms7864 PMID: 25904309

25. Biau G, Scornet E. A random forest guided tour. Test. 2016 Jun 1; 25(2):197–227.

26. Inza I, Larrañaga P, Blanco R, Cerrolaza AJ. Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial intelligence in medicine. 2004 Jun 1; 31(2):91–103. https://doi.org/10.1016/j.artmed.2004.01.007 PMID: 15219288

27. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785–794). ACM.

28. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting novel associations in large data sets. science. 2011 Dec 16; 334(6062):1518–24. https://doi.org/10.1126/science.1205438 PMID: 22174245

29. Borgatti SP. Centrality and network flow. Social networks. 2005 Jan 1; 27(1):55–71.

30. Krishnan K, Steptoe AL, Martin HC, Pattabiraman DR, Nones K, Waddell N, et al. miR-139-5p is a regulator of metastatic pathways in breast cancer. Rna. 2013 Dec 1; 19(12):1767–80. https://doi.org/10.1261/rna.042143.113 PMID: 24158791

31. Yan LX, Huang XF, Shao Q, Huang MY, Deng L, Wu QL, et al. MicroRNA miR-21 overexpression in human breast cancer is associated with advanced clinical stage, lymph node metastasis and patient poor prognosis. Rna. 2008 Nov 1; 14(11):2348–60. https://doi.org/10.1261/rna.1034808 PMID: 18812439

32. Hong Y, Liang H, Wang Y, Zhang W, Zhou Y, Yu M, et al. miR-96 promotes cell proliferation, migration and invasion by targeting PTPN9 in breast cancer. Scientific reports. 2016 Nov 18; 6:37421. https://doi.org/10.1038/srep37421 PMID: 27857177

33. Macedo T, Silva-Oliveira RJ, Silva VA, Vidal DO, Evangelista AF, Marques M. Overexpression of mir-183 and mir-494 promotes proliferation and migration in human breast cancer cell lines. Oncology letters. 2017 Jul 1; 14(1):1054–60. https://doi.org/10.3892/ol.2017.6265 PMID: 28693273

34. Kholoussi NM, El-Nabi SE, Esmaiel NN, Abd El-Bary NM, El-Kased AF. Evaluation of Bax and Bak gene mutations and expression in breast cancer. BioMed research international. 2014; 2014.

35. Azoulay-Alfaguter I, Elya R, Avrahami L, Katz A, Eldar-Finkelman H. Combined regulation of mTORC1 and lysosomal acidification by GSK-3 suppresses autophagy and contributes to cancer cell growth. Oncogene. 2015 Aug; 34(35):4613. https://doi.org/10.1038/onc.2014.390 PMID: 25500539

36. Chao YL, Shepard CR, Wells A. Breast carcinoma cells re-express E-cadherin during mesenchymal to epithelial reverting transition. Molecular cancer. 2010 Dec; 9(1):179.

37. Boulay PL, Mitchell L, Turpin J, Huot-Marchand JÉ, Lavoie C, Sanguin-Gendreau V, et al. Rab11-FIP1C is a critical negative regulator in ErbB2-mediated mammary tumor progression. Cancer research. 2016 May 1; 76(9):2662–74. https://doi.org/10.1158/0008-5472.CAN-15-2782 PMID: 26933086

38. Simpkins SA, Hanby AM, Holliday DL, Speirs V. Clinical and functional significance of loss of caveolin-1 expression in breast cancer-associated fibroblasts. The Journal of pathology. 2012 Aug 1; 227(4):490–8. https://doi.org/10.1002/path.4034 PMID: 22488553

39. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005 May 1; 37(5):360–3. PMID: 15883903

40. Bosch A, Li Z, Bergamaschi A, Ellis H, Toska E, Prat A, et al. PI3K inhibition results in enhanced estrogen receptor function and dependence in hormone receptor–positive breast cancer. Science translational medicine. 2015 Apr 15; 7(283):283ra51-. https://doi.org/10.1126/scitranslmed.aaa4442 PMID: 25877889

41. Karousou E, D'Angelo ML, Kouvidi K, Vigetti D, Viola M, Nikitovic D, et al. Collagen VI and hyaluronan: the common role in breast cancer. BioMed research international. 2014; 2014.

42. Chen X, Iliopoulos D, Zhang Q, Tang Q, Greenblatt MB, Hatziapostolou M, et al. XBP1 promotes triple-negative breast cancer by controlling the HIF1α pathway. Nature. 2014 Apr; 508(7494):103. https://doi.org/10.1038/nature13119 PMID: 24670641

43. Hardy SD, Geahlen RL. Investigating the role of Syk in TGF-β induced P-bodies and breast cancer metastasis.

44. Taipaleenmäki H, Browne G, Akech J, Zustin J, Van Wijnen AJ, Stein JL,et al. Targeting of Runx2 by miR-135 and miR-203 impairs progression of breast cancer and metastatic bone disease. Cancer research. 2015 Apr 1; 75(7):1433–44. https://doi.org/10.1158/0008-5472.CAN-14-1026 PMID: 25634212

45. Uhlmann S, Zhang JD, Schwäger A, Mannsperger H, Riazalhosseini Y, Burmester S, et al. miR-200bc/429 cluster targets PLCγ1 and differentially regulates proliferation and EGF-driven invasion than miR-200a/141 in breast cancer. Oncogene. 2010 Jul; 29(30):4297. https://doi.org/10.1038/onc.2010.201 PMID: 20514023

46. Cai J, Guan H, Fang L, Yang Y, Zhu X, Yuan J, et al. MicroRNA-374a activates Wnt/β-catenin signaling to promote breast cancer metastasis. The Journal of clinical investigation. 2013 Jan 16; 123(2).

47. Yang M, Yao Y, Eades G, Zhang Y, Zhou Q. MiR-28 regulates Nrf2 expression through a Keap1-independent mechanism. Breast cancer research and treatment. 2011 Oct 1; 129(3):983–91. https://doi.org/10.1007/s10549-011-1604-1 PMID: 21638050

48. Gasparini P, Lovat F, Fassan M, Casadei L, Cascione L, Jacob NK, et al. Protective role of miR-155 in breast cancer through RAD51 targeting impairs homologous recombination after irradiation. Proceedings of the National Academy of Sciences. 2014 Mar 25; 111(12):4536–41.

49. Xie F, Hosany S, Zhong S, Jiang Y, Zhang F, Lin L, et al. MicroRNA-193a inhibits breast cancer proliferation and metastasis by downregulating WT1. PloS one. 2017 Oct 10; 12(10):e0185565. https://doi.org/10.1371/journal.pone.0185565 PMID: 29016617

50. Li M, Liu L, Zang W, Wang Y, Du Y, Chen X, et al. miR-365 overexpression promotes cell proliferation and invasion by targeting ADAMTS-1 in breast cancer. International journal of oncology. 2015 Jul 1; 47 (1):296–302. https://doi.org/10.3892/ijo.2015.3015 PMID: 25998153

51. Lin WH, Li J, Zhang B, Liu LS, Zou Y, Tan JF, et al. MicroRNA-1301 induces cell proliferation by down-regulating ICAT expression in breast cancer. Biomedicine & Pharmacotherapy. 2016 Oct 1; 83:177–85.

52. Knezevic J, Pfefferle AD, Petrovic I, Greene SB, Perou CM, Rosen JM. Expression of miR-200c in claudin-low breast cancer alters stem cell functionality, enhances chemosensitivity and reduces metastatic potential. Oncogene. 2015 Dec; 34(49):5997. https://doi.org/10.1038/onc.2015.48 PMID: 25746005

53. Chang CH, Fan TC, Yu JC, Liao GS, Lin YC, Shih AC, et al. The prognostic significance of RUNX2 and miR-10a/10b and their inter-relationship in breast cancer. Journal of translational medicine. 2014 Dec; 12(1):257.

54. Zhang JG, Shi Y, Hong DF, Song M, Huang D, Wang CY, et al. MiR-148b suppresses cell proliferation and invasion in hepatocellular carcinoma by targeting WNT1/β-catenin pathway. Scientific reports. 2015 Jan 28; 5:8087. https://doi.org/10.1038/srep08087 PMID: 25627001

55. Joly MM, Hicks DJ, Jones B, Sanchez V, Estrada MV, Young C, et al. Rictor/mTORC2 drives progression and therapeutic resistance of HER2-amplified breast cancers. Cancer research. 2016 Aug 15; 76 (16):4752–64. https://doi.org/10.1158/0008-5472.CAN-15-3393 PMID: 27197158

56. Boulay PL, Mitchell L, Turpin J, Huot-Marchand JÉ, Lavoie C, Sanguin-Gendreau V, et al. Rab11-FIP1C is a critical negative regulator in ErbB2-mediated mammary tumor progression. Cancer research. 2016 May 1; 76(9):2662–74. https://doi.org/10.1158/0008-5472.CAN-15-2782 PMID: 26933086

57. Hsieh TH, Tsai CF, Hsu CY, Kuo PL, Lee JN, Chai CY, et al. Phthalates induce proliferation and invasiveness of estrogen receptor-negative breast cancer through the AhR/HDAC6/c-Myc signaling pathway. The FASEB Journal. 2012 Feb 1; 26(2):778–87. https://doi.org/10.1096/fj.11-191742 PMID: 22049059

58. Johnson N, Li YC, Walton ZE, Cheng KA, Li D, Rodig SJ, et al. Compromised CDK1 activity sensitizes BRCA-proficient cancers to PARP inhibition. Nature medicine. 2011 Jul; 17(7):875. https://doi.org/10.1038/nm.2377 PMID: 21706030

59. Lu J, Guo H, Treekitkarnmongkol W, Li P, Zhang J, Shi B, et al. 14-3-3ζ cooperates with ErbB2 to promote ductal carcinoma in situ progression to invasive breast cancer by inducing epithelial-mesenchymal transition. Cancer cell. 2009 Sep 8; 16(3):195–207. https://doi.org/10.1016/j.ccr.2009.08.010 PMID: 19732720

60. Paplomata E, O'Regan R. The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. Therapeutic advances in medical oncology. 2014 Jul; 6(4):154–66. https://doi.org/10.1177/1758834014530023 PMID: 25057302