PLOS ONE

RESEARCH ARTICLE

Identifying Cancer Subtypes from miRNA-TF-mRNA Regulatory Networks and Expression Data

Taosheng Xu[1,2☯], Thuc Duy Le[3☯]*, Lin Liu[3], Rujing Wang[1], Bingyu Sun[1], Jiuyong Li[3]*

1 Institute of Intelligent Machines, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, Anhui, China, **2** Department of Automation, University of Science and Technology of China, Hefei, Anhui, China, **3** School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, South Australia, Australia

☯ These authors contributed equally to this work.
* thuc.le@unisa.edu.au (TDL); jiuyong.li@unisa.edu.au (JL)

CrossMark
click for updates

OPEN ACCESS

Citation: Xu T, Le TD, Liu L, Wang R, Sun B, Li J (2016) Identifying Cancer Subtypes from miRNA-TF-mRNA Regulatory Networks and Expression Data. PLoS ONE 11(4): e0152792. doi:10.1371/journal.pone.0152792

Editor: Bibekanand Mallick, National Institute of Technology, Rourkela, INDIA

Received: December 13, 2015

Accepted: March 18, 2016

Published: April 1, 2016

Copyright: © 2016 Xu et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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

Funding: This work has been partially supported by Australian Research Council (http://www.arc.gov.au/) Discovery Project DP130104090 (JL and LL), and the National Natural Science Foundation of China 31371340 (BS), http://www.nsfc.gov.cn/publish/portal1/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Abstract

Background

Identifying cancer subtypes is an important component of the personalised medicine framework. An increasing number of computational methods have been developed to identify cancer subtypes. However, existing methods rarely use information from gene regulatory networks to facilitate the subtype identification. It is widely accepted that gene regulatory networks play crucial roles in understanding the mechanisms of diseases. Different cancer subtypes are likely caused by different regulatory mechanisms. Therefore, there are great opportunities for developing methods that can utilise network information in identifying cancer subtypes.

Results

In this paper, we propose a method, weighted similarity network fusion (WSNF), to utilise the information in the complex miRNA-TF-mRNA regulatory network in identifying cancer subtypes. We firstly build the regulatory network where the nodes represent the features, i.e. the microRNAs (miRNAs), transcription factors (TFs) and messenger RNAs (mRNAs) and the edges indicate the interactions between the features. The interactions are retrieved from various interatomic databases. We then use the network information and the expression data of the miRNAs, TFs and mRNAs to calculate the weight of the features, representing the level of importance of the features. The feature weight is then integrated into a network fusion approach to cluster the samples (patients) and thus to identify cancer subtypes. We applied our method to the TCGA breast invasive carcinoma (BRCA) and glioblastoma multiforme (GBM) datasets. The experimental results show that WSNF performs better than the other commonly used computational methods, and the information from miRNA-TF-mRNA regulatory network contributes to the performance improvement. The WSNF method successfully identified five breast cancer subtypes and three GBM subtypes

which show significantly different survival patterns. We observed that the expression patterns of the features in some miRNA-TF-mRNA sub-networks vary across different identified subtypes. In addition, pathway enrichment analyses show that the top pathways involving the most differentially expressed genes in each of the identified subtypes are different. The results would provide valuable information for understanding the mechanisms characterising different cancer subtypes and assist the design of treatment therapies. All datasets and the R scripts to reproduce the results are available online at the website: http://nugget.unisa.edu.au/Thuc/cancersubtypes/.

Introduction

Rather than being a single disease, cancer involves different subtypes characterised by different sets of molecules [1, 2]. Identifying cancer subtypes is a crucial task for selecting the right treatment for patients, as different cancer subtypes may respond well to different treatment therapies. For example, estrogen receptor (ER) positive breast cancer subtype would respond to hormone therapy, and the human epidermal growth factor receptor 2 (HER2) positive subtype is likely to benefit from chemotherapy. However, our current understanding of the mechanisms controlling each cancer subtype is still far from complete.

Several computational methods have been developed to identify cancer subtypes. These methods fall into three different streams of research. In the first stream, data mining or machine learning models are built to utilise gene expression datasets for clustering samples (patients) into different groups, each corresponding to a cancer subtype [3–7]. However, utilising one genomic data type may not be sufficient to identify cancer subtypes accurately. With the advance of sequencing technologies, multiple data types of cancer patients such as genomic, miRNA and related clinical data are made available nowadays. These wealth of datasets lead to the second stream of research in which researchers analyse different types of data separately for identifying subtypes and the results obtained separately are then integrated to form the final result. Highlights of this approach are [1, 8–10]. However, analysing the different types of data separately may lose the complementary information in the data of the same patients, and there may be conflict in the results obtained using different types of data. The last stream of research focuses on analysing multi-omics data at the same time and has identified some important cancer subtypes recently [11–14].

However, the information from gene regulatory networks is rarely used by the existing computational methods. Gene regulatory networks play an important role in every life process, and understanding the dynamics of these networks help reveal the mechanisms of diseases [15]. Although the importance of network-based information has been addressed in recent works [16, 17], there is still a lack of methods utilising biological information from networks to identify cancer subtypes. Moreover, it remains a great challenge to associate the multi-omics data and network information with cancer subtypes and the outcomes in particular prognosis. Recently, Liu et al. [18] proposed the NCIS (network-assisted co-clustering for the identification of cancer subtypes) method to utilise the expression profiles of mRNAs and the network information of mRNA-mRNA interactions with a bi-clustering method to discover cancer subtypes. However, gene regulatory networks are complex and involve many types of regulators including miRNAs and TFs. It is of interest to utilise the information in the networks that involve miRNAs, TFs, and mRNAs in identifying cancer subtypes. The information may not

only improve the accuracy of the computational models, but also provide insights into the mechanisms (the regulatory networks) regulating each cancer subtype.

In this paper, we propose a method, called weighted similarity network fusion (WSNF), to identify cancer subtypes by making use of both the expression data and network information of miRNAs, TFs and mRNAs. Given a dataset containing the expression profiles of a set of miRNAs, TFs and mRNAs (known as features in the rest of the paper), WSNF firstly retrieves the interactions between these features from different interatomic databases to build the miRNA-TF-mRNA regulatory network. In the network, features are represented by nodes and interactions between features are indicated by edges. We then calculate the weight (i.e. importance) of a feature by utilising the miRNA-TF-mRNA network information and the expression variation of the features. Finally, we modify the similarity network fusion (SNF) approach [11] to take the feature weight into consideration when clustering patients for identifying cancer subtypes.

We apply the WSNF method to the TCGA breast cancer and GBM datasets. The experimental results show that our method has successfully identified five breast cancer subtypes and three GBM subtypes which show significantly different survival patterns. The information from the miRNA-TF-mRNA regulatory network improves the performance of the network fusion approach, as the WSNF method performs better than both SNF [11], the network fusion method without using feature weight and NCIS [18] that uses only mRNA expression data and mRNA-mRNA interactions. We also compare our method with Consensus clustering (CC) [7], a method that is commonly used in TCGA research. The experimental results show that the WSNF method also has better performance with both the breast cancer and GBM datasets. For the breast cancer dataset, we analyse the identified subtypes in detail and report the results in terms of the expression patterns, the differences in the miRNA-TF-mRNA regulatory networks across the different subtypes, and the functional pathways characterising each subtype. The information can be valuable for assisting the treatment design of specific breast cancer subtypes.

Materials and Methods

Method overview

We propose to use the miRNA-TF-mRNA regulatory network to assist the identification of cancer subtypes. There are three main steps in the WSNF method (Fig 1), including: 1) constructing miRNA-TF-mRNA regulatory network, 2) calculating the weight for each feature (miRNA, TF, mRNA), and 3) modifying and applying the similarity network fusion approach [11] to identify cancer subtypes, while taking the feature weight into consideration. We describe the details of each step in the following.

Constructing the miRNA-TF-mRNA regulatory network

In this step, we use a variety of sources to build the miRNA-TF-mRNA interaction networks. The network contains different types of interactions, including those between miRNA-mRNA, miRNA-TF, TF-miRNA, TF-mRNA, TF-TF, and mRNA-mRNA. Fig 2 shows the details of the data sources for retrieving the different type interactions. In the figure, each type of the interactions is represented as a link where the source is the regulator and the arrow end is the target. The data sources are listed next to each type of the interactions.

We firstly get the list of TFs by combining the TFs in the Encyclopedia of DNA Elements (ENCODE) ChIP-seq data, TransmiR [19] and FANTOM5 Human transcription factors which are available at http://fantom.gsc.riken.jp/5/sstar/Browse_Transcription_Factors_hg19. Finally a list of 1679 TFs is obtained (see the S1 File for the list).

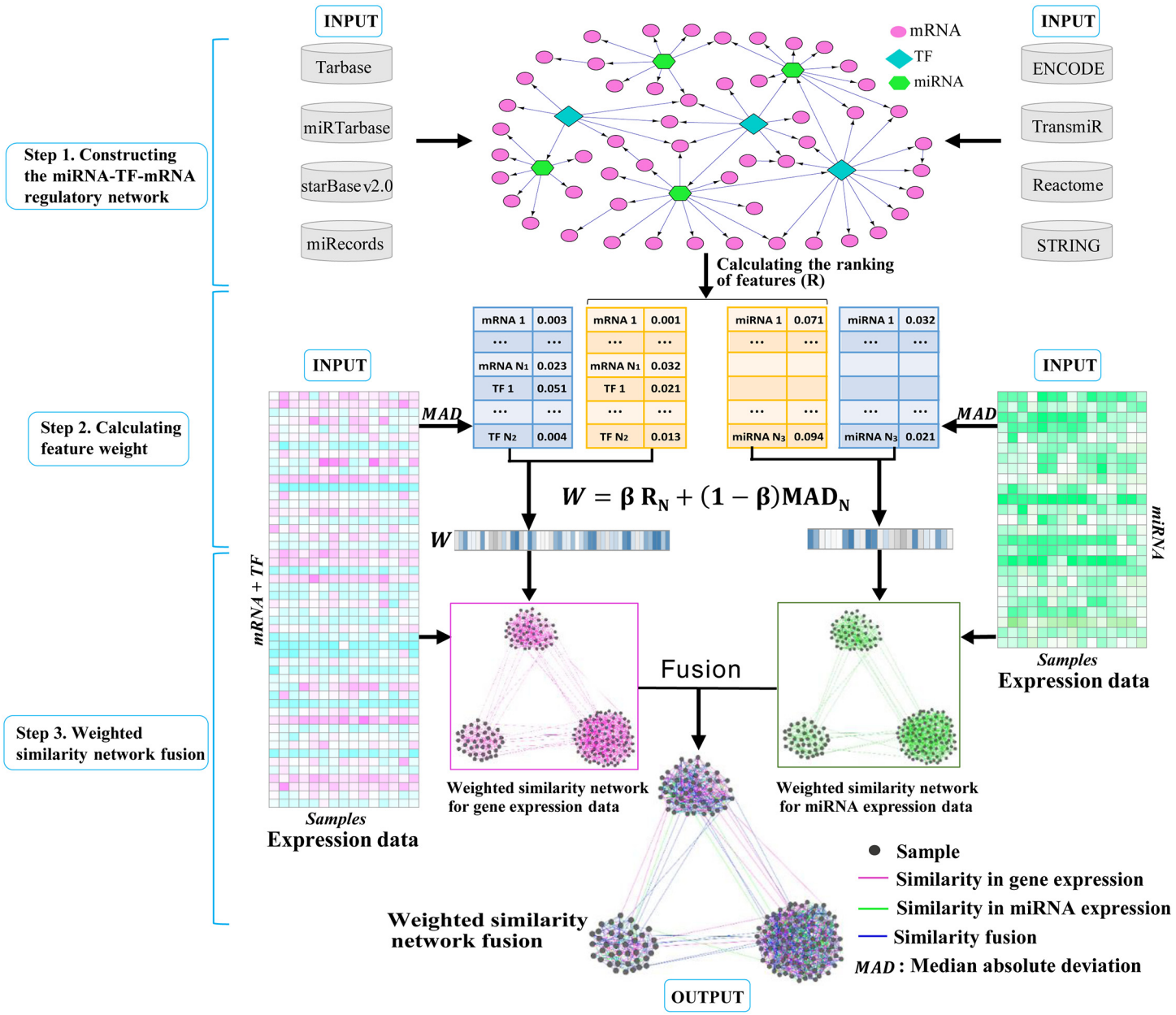

Fig 1. Workflow of WSNF. In step 1, interactions between miRNAs, TFs and mRNAs obtained from the databases are used to construct the miRNA-TF-mRNA regulatory network. In step 2, the ranking of each feature (R) is calculated based on the network information, and gene and miRNA expression data are used to get the feature expression variation (MAD) across all the samples. Then for each feature, its ranking and expression variation are combined to obtain its weight (W). In step 3, the weighted sample similarity networks are obtained from genes (mRNAs, TFs) and miRNAs separately using the weights and expression data of the features, and finally network fusion and clustering are performed to find patient groups that imply cancer subtypes.

doi:10.1371/journal.pone.0152792.g001

As shown in Fig 2, we obtain the miRNA-mRNA and miRNA-TF interactions from experimentally confirmed databases, including Tarbase [20], mirTarbase [21], mirRecords [22], and prediction database starBase v2.0 [23]. Tarbase, mirTarbase and mirRecords include the curated confirmed interactions from the literature. starBase v2.0 contains the union of the sets of miRNA-mRNA interactions predicted by the five miRNA target prediction software programs (TargetScan, PicTar, PITA, miRanda and RNA22). It also tests each of the miRNA-mRNA

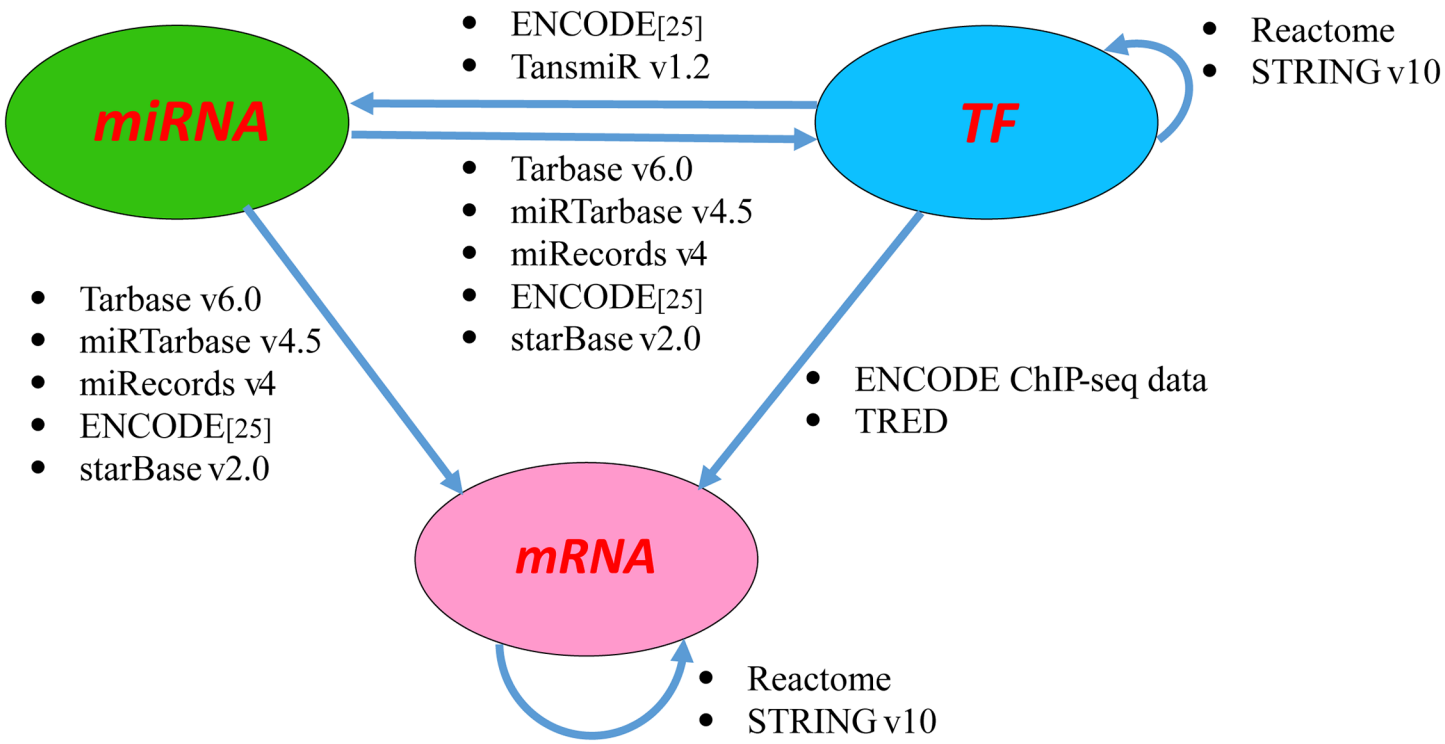

Fig 2. The data sources for constructing the miRNA-TF-mRNA regulatory network.

doi:10.1371/journal.pone.0152792.g002

interaction pairs based on TCGA Pan-cancer [24] expression datasets. The criterion of the validation test is the anti-correlation with negative Pearson correlation coefficient (p-value< 0.05) between a miRNA and its target. In our network, we use the miRNA-mRNA interactions in starBase v2.0 that are supported by at least one TCGA Pan-cancer expression dataset. In addition, the miRNA-mRNA interactions derived from ENCODE data [25] are also used in our work. The interactions are available at: http://encodenets.gersteinlab.org/.

The mRNA-mRNA interactions are retrieved from Reactome [26] and STRING v10.0 [27]. Since contained in the Reactome and STRING are the protein-protein interaction pairs, we use the *org.Hs.eg.db* R package [28] to map the protein-gene annotation to get the corresponding mRNA-mRNA interaction pairs. We choose the score cut-off as 0.9 in STRING v10.0 to select the mRNA-mRNA pairs of high credibility for our network.

For TF regulation, we obtain the interactions between TF-mRNA from the ENCODE ChIP-seq data [29] and Transcriptional Regulatory Element Database (TRED) [30]. ENCODE ChIP-seq data at UCSC Genome Browser are processed using the computational pipeline to generate uniform peaks of TF binding. TRED is an integrated repository for both cis- and trans-regulatory elements. It contains the curated transcriptional regulation information, including the transcription factor binding motifs and experimental evidence. We retrieve the TF-TF interactions from Reactome and STRING, with the protein-gene annotation mapping as that for getting the TF-TF interactions. For our network, TF-miRNA interactions are obtained from two sources: TransmiR [19] and the supplementary data of [25] that is also available at http://encodenets.gersteinlab.org/.

Calculating feature weights

With the proposed WSNF method, we calculate the weight of a feature in two stages. Firstly, we use the information of the miRNA-TF-mRNA network constructed in the previous step to rank the features. Then the expression data is used to find the expression variation of each feature across all the samples in the datasets. At last, the weight of a feature is obtained by combining its ranking and expression variation.

Stage 1: Computing ranking of features using Google PageRank. Google PageRank [31, 32] is an algorithm which was initially used to rank the vast number of webpages by Google Search. It is based on a directed graph $G(V,E)$ where the nodes V represent webpages and the edges E indicate the hyperlinks between the webpages. The basic assumption is that an important webpage is likely to have more inbound links from other webpages. Suppose there are N webpages $\{p_1, p_2, \ldots, p_N\}$. The ranking of a webpage p_i is defined as the following:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \quad (1)$$

where $PR(p_i)$ and $PR(p_j)$ are the rankings of webpages p_i and p_j respectively, with $p_i \leftarrow p_j$; d is the damping factor which is like a click-through probability used to decay the ranking of the webpages with no outgoing links, and $0 < d < 1$; $M(p_i)$ is the set of webpages that are linked to p_i ; and $L(p_j)$ is the number of outbound links from p_j. So a webpage p_i will have a high ranking if it is linked by many other high-ranked webpages p_j. For interested readers, the convergence and computation of the PageRank using the above iterative formula (i.e. Eq 1) are illustrated in [33, 34].

For our case of utilising miRNA-TF-mRNA regulatory network to rank a feature, a molecular regulating many targets is important. In our miRNA-TF-mRNA network, denoted as $G(V, E)$, the nodes V are the features (miRNAs, TFs and mRNAs) and the edges E are the interactions between regulators and their targets. The direction of an edge is from a regulator to its target. An important regulator is analogous to an important webpage in PageRank that many other webpages link to, except that the regulator has many links going out of it to its targets. Suppose there are N features $\{f_1, f_2, \ldots, f_N\}$. The ranking (regulatory importance) of a feature f_i can be defined as follows using a modified PageRank algorithm:

$$R(f_i) = \frac{1-d}{N} + d \sum_{f_j \in T(f_i)} \frac{R(f_j)}{L(f_j)} \quad (2)$$

where $R(f_i)$ and $R(f_j)$ are the rankings of features f_i and f_j respectively, with $f_i \rightarrow f_j$; d is the the damping factor, and $0 < d < 1$; $T(f_i)$ is the set of targets that f_i regulates; and $L(f_j)$ is the number of regulators which regulate f_j.

The R and Matlab scripts of computing the feature ranking from miRNA-TF-mRNA regulatory network is provided in the S2 File.

Stage 2: Integrating feature ranking and feature variation. The expression variation across samples is an important indicator for the research of cancer genomic data. The features (e.g. genes) with higher expression variations are always treated as more important biological marker in cancer mechanisms. We use the median absolute deviation (MAD) to represent the expression variation of a feature. The MAD of a feature f_i is calculated as:

$$MAD(f_i) = median(|X(f_i) - median(X(f_i))|) \quad (3)$$

where $X(f_i)$ is a numeric vector which represents the expression values of feature f_i across all samples (patients).

To integrate the feature variation with feature ranking, NCIS [18] follows the idea of GeneRank [35] to simply replace the part $[\frac{1-d}{N}]$ in Google PageRank algorithm with the MAD to obtain the final weight of a feature. However, we find that the final weight obtained in this way by both GeneRank and NCIS is strongly correlated with the feature weight directly calculated with Eq 2, i.e. without using MAD. The strong correlation implies that the approach taken by the two methods of integrating MAD is not effective as the expression variation information is not reflected by the final weight obtained using their approach. The detailed results on this finding are shown in the S3 File.

To overcome this problem, we adopt a linear model to effectively integrate the feature ranking and the feature variation in this paper. We firstly normalise the feature ranking obtained from the miRNA-TF-mRNA regulatory network and feature variation from expression data as follows:

$$R_N(f_i) = \frac{R(f_i)}{\Sigma_{m=1}^{N} R(f_m)} \tag{4}$$

$$MAD_N(f_i) = \frac{MAD(f_i)}{\Sigma_{m=1}^{N} MAD(f_m)} \tag{5}$$

A linear model is then applied to integrate these two measures to get the final weight for each feature.

$$W(f_i) = \beta R_N(f_i) + (1-\beta) MAD_N(f_i) \tag{6}$$

where β is a tuning parameter for the importance of the miRNA-TF-mRNA regulatory network information. The larger the value of β is the more important role the information of the miRNA-TF-mRNA regulatory network will play in calculating the final weight of the features. In our experiments, we set β to 0.8 to focus more on the network information for the cancer subtype discovery.

Weighted similarity network fusion

We utilise the feature weight information to assist the identification of cancer subtypes from the gene expression data and miRNA expression data. To this end, we modify the similarity network fusion (SNF) method [11] to incorporate the feature weight obtained in the previous step into the process of cancer subtype classification.

SNF is a multi-omics data processing method that constructs a fusion patient similarity network by integrating the patient similarity obtained from each of the genomic data types. SNF calculates the similarity between patients using each single data type separately. The similarities between patients from different data types are then integrated by a cross-network diffusion process to construct the fusion patient similarity matrix. Finally, a clustering method is applied to the fusion patient similarity matrix to cluster patients into different groups, which imply different cancer subtypes.

The key step of SNF is to define the similarity between patients, as we need to stratify similar patients into the same group (subtype). Euclidean distance is used in SNF to measure the similarity between patients in single genomic data type, where, however, all features are treated as equally important. Suppose that there is an expression profile dataset (n patients $\times$ p features),

then the Euclidean distance between patient S_i and patient S_j is:

$$Distance(S_i, S_j) = \sqrt{\sum_{m=1}^{p} (f_m^{S_i} - f_m^{S_j})^2}; \forall i, j \leq n, i \neq j \quad (7)$$

where $f_m^{S_i}$ and $f_m^{S_j}$ are the expression values of f_m in patients S_i and S_j, respectively.

We modify the patient distance formula as follows take the weight of each feature into consideration:

$$Distance(S_i, S_j) = \sqrt{\sum_{m=1}^{p} W(f_m) * (f_m^{S_i} - f_m^{S_j})^2}; \forall i, j \leq n, i \neq j \quad (8)$$

By using the above modified samples distance formula, the proposed WSNF method considers similarity of two patients based on not only the overall difference between the expression levels of all their features, but also the importance (weight) of each of the features. As we make use of the miRNA-TF-mRNA network information in the calculation of feature weight and our method treats different features differently, we will see in the Results and discussion Section that WSNF significantly outperforms the SNF and the other commonly used methods for identifying cancer subtypes.

Results and Discussion

Datasets

In this paper, we use the BRCA and GBM datasets from The Cancer Genome Atlas (TCGA) for our experiments, including the gene (mRNA and TF) expression data, miRNA expression data and clinical data (overall survival time, survival status and some clinical covariates). The Level 3 TCGA tumor samples are downloaded from the Broad GDAC Firehose (timestamp: 2015-04-02). To get the most number of matched samples for both cancers, we use RNASeq and miRNAHiseq data for BRCA and microarray data for GBM.

The genes and miRNAs with very low expression levels and low variations across samples are removed. The different cut-off points are selected based on the distribution characteristics of the BRCA and GBM datasets (see the S3 File). For the BRCA RNASeq and miRNAHiseq datasets, we firstly use the *log*2 transformation to preprocess them, which is commonly used for RNA-sequencing data as introduced in the *DESeq2* [36] R package. We calculate the average value for each feature across samples and remove the 25% genes and 60% miRNAs with low average expression. Then the standard deviation of each gene and miRNA is calculated, and genes and miRNAs with standard deviation less than 0.5 are also removed. For the GBM microarray data, there are some missing observations. We firstly apply the imputation by using the *impute* R pacakage [37]. Then we calculate the standard deviation of each gene and miRNA. The genes with standard deviation less than 0.6 and the miRNAs with standard deviation less than 0.2 are removed. The detailed processing procedure of the datasets are recorded in the S3 File. In the end, there are 587 matched samples in BRCA with 12,233 mRNAs, 1,338 TFs and 361 miRNAs. Meanwhile, for GBM there are 276 matched samples with 10,278 mRNAs, 1,083 TFs and 287 miRNAs (see the S3 File).

Network construction

As mentioned in the Materials and Methods Section, we use several public databases to construct the miRNA-TF-mRNA regulatory network. Table 1 shows the number of interactions

Table 1. The interactions used for constructing the miRNA-TF-mRNA regulatory network for the BRCA dataset.

	Database	Total interactions	Found interactions
	Tarbase v6.0	17,526	12,130
miRNA → TF&	miRTarBase v4.5	37,423	26,847
miRNA →mRNA	miRecords v4	1,707	1,095
	starBase v2.0	320,709	219,088
	ENCODE [25]	117,193	54,603
TF → miRNA	ENCODE [25]	1,648	579
	Transmir v1.2	649	457
TF → mRNA	ENCODE ChIP-Seq	229,486	133,952
	TRED	7,066	4,739
TF → TF&	Reactome	127,452	60,648
mRNA → mRNA	STRING v10.0	250,843	122,938

doi:10.1371/journal.pone.0152792.t001

from the data sources for constructing the regulatory networks for the BRCA dataset. Similar information for the GBM dataset is in the S3 File.

The identified subtypes have significantly different survival patterns

With the constructed networks and the BRCA and GBM expression datasets, WSNF identifies five breast cancer subtypes and three GBM subtypes. The identified cancer subtypes and related clinical information for breast cancer and GBM are given in the S4 and S5 Files. To assess how well our method has performed in identifying cancer subtypes, we conduct survival analysis of the identified cancer subtypes. Figs 3 and 4 show the survival curves of the patients in the five subtypes of BRCA and the three subtypes of GBM, respectively. The *p*-values from the Log-rank tests [38] are 0.00483 for BRCA and 0.00279 for GBM. The p-values suggest that the identified subtypes in both datasets have significantly different survival patterns, indicating different cancer subtypes respectively.

Furthermore, we use the Silhouette width [39] and black-white heatmap to demonstrate the consistency of the samples (patients) in each subtype and the difference across different subtypes, respectively. As shown in Figs 3 and 4, the overall average Silhouette width values are positive for both BRCA and GBM. Note that the Silhouette width value is positive if the samples in each subtype are consistent, and negative otherwise. Meanwhile, the black-white heatmaps are generated from the matrix of sample similarity by arranging the samples according the cluster labels. The block boundaries for all subtypes are very clear. In particular, the third subtype of BRCA has a high Silhouette width value and a clear contrast in the black-white heatmap, which suggests unique characteristics of the patients in this subtype.

The network information improves the identification of cancer subtypes

To investigate whether the information from the miRNA-TF-mRNA regulatory network actually helps improve the identification of cancer subtypes, we compare the WSNF method with the previously proposed methods including NCIS [18], Consensus clustering (CC) [7], and SNF [11]. NCIS utilises gene expression data and the information from mRNA-mRNA interactions. CC is the commonly used clustering method in TCGA research papers [1, 8, 40–42] based on single genomic data type. SNF is the multiple genome data fusion and clustering method but does not use the information from the gene regulatory networks. To make a fair comparison, from our processed datasets (BRCA & GBM) and constructed miRNA-TF-mRNA regulatory

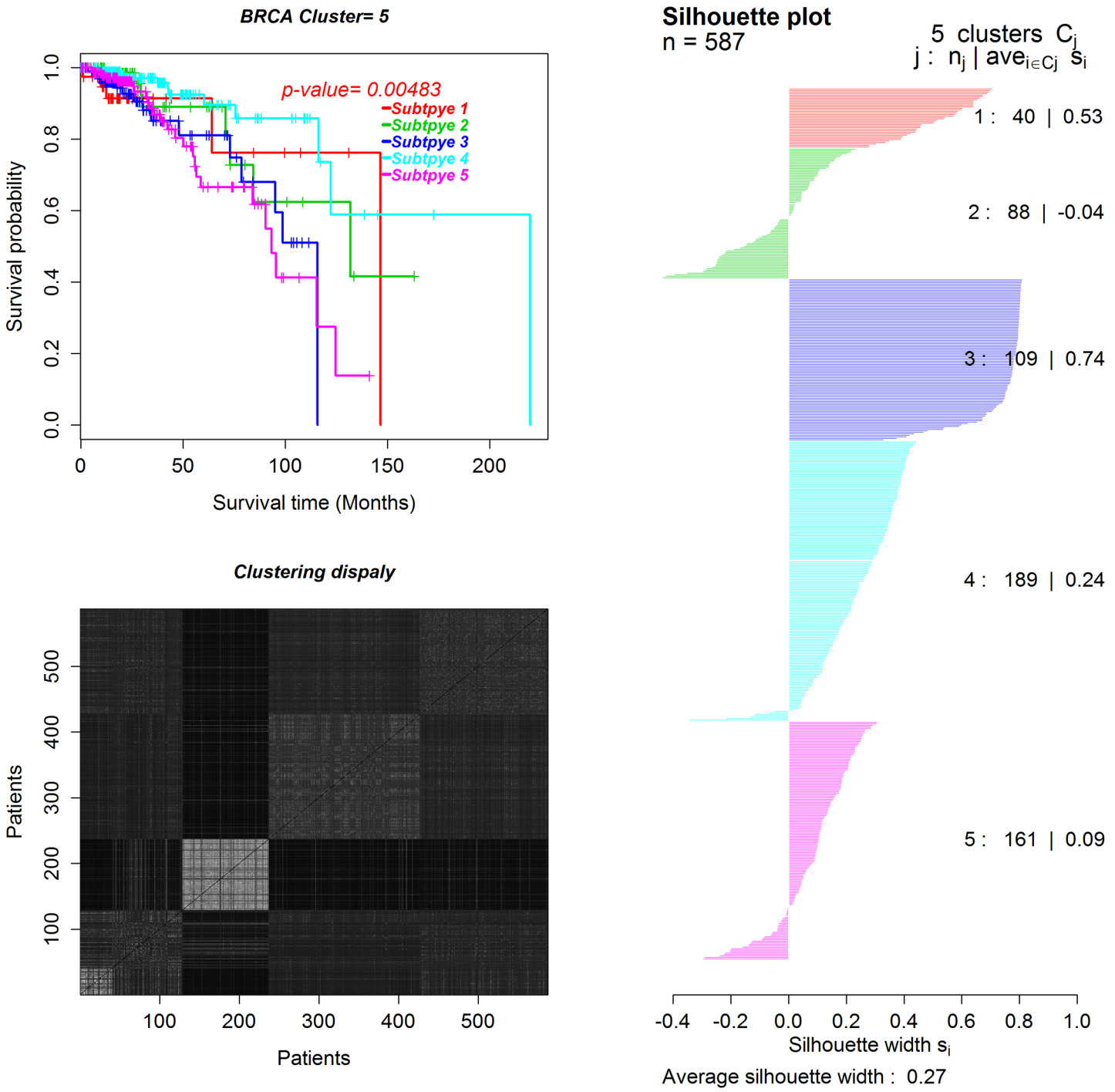

Fig 3. The survival curves and Silhouette plots for the five subtypes of BRCA. j, n_j, s_i in the Silhouette plot are subtype label, the number of patients in the subtype and the Silhouette width for patient i, respectively.

doi:10.1371/journal.pone.0152792.g003

networks, we use the gene expression data and extract mRNA-mRNA interactions as the input for NICS. We concatenate the normalised gene expression data and normalised miRNA expression data for each patient as the input data for CC. The inputs of the SNF are the gene expression data and miRNA expression data. The inputs of our WSNF method are the gene

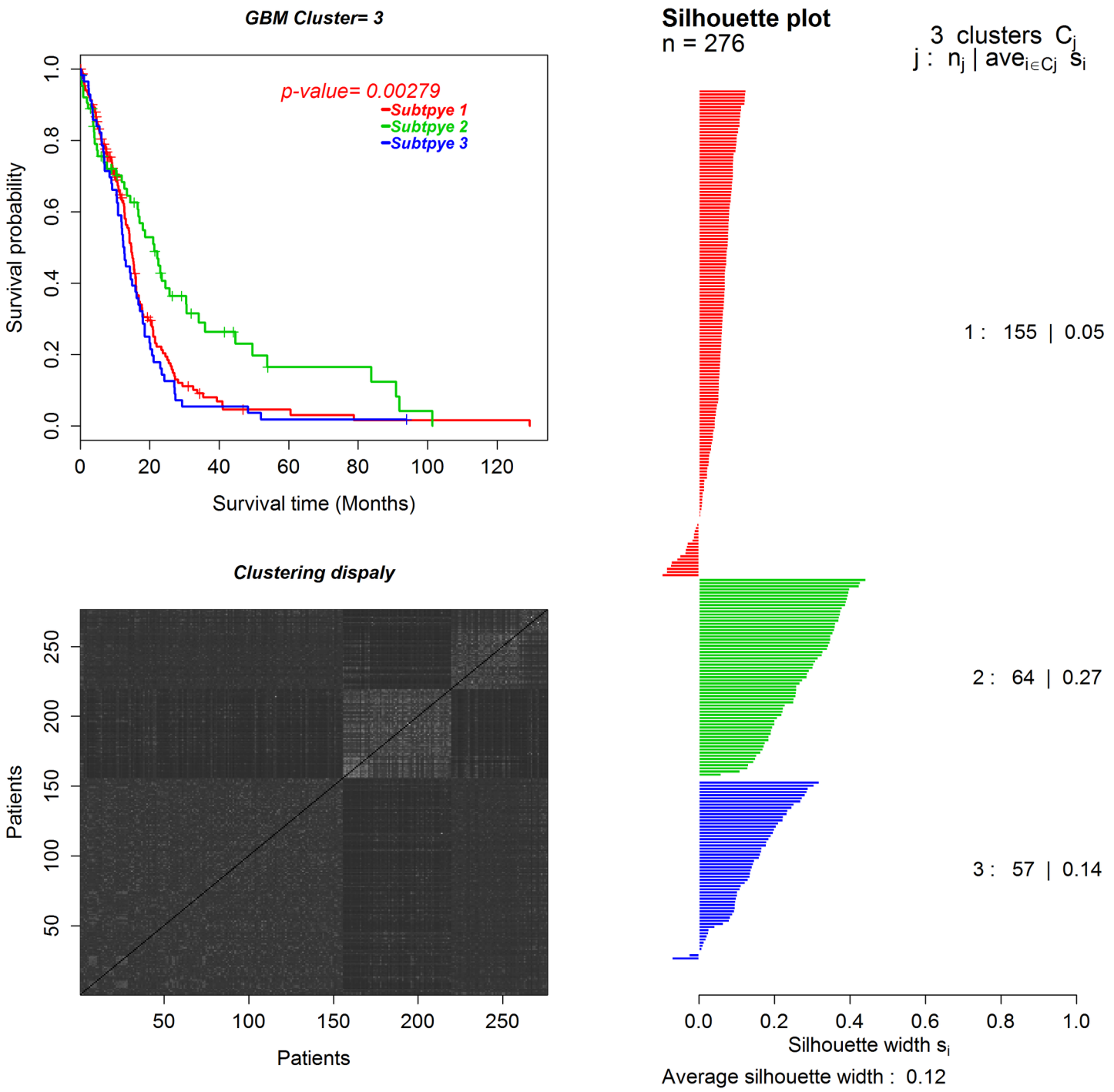

Fig 4. The survival curves and Silhouette plots for the three subtypes of GBM. j, n_j, s_i in the Silhouette plot are subtype label, the number of patients in the subtype and the Silhouette width for patient i, respectively.

doi:10.1371/journal.pone.0152792.g004

expression data, miRNA expression data and the miRNA-TF-mRNA regulatory networks. We conduct the survival analyses for the identified subtypes by each of the methods and compare the p-values of the Log-rank tests [38] to evaluate the significance of the different survival distributions across subtypes.

Table 2. Comparison of the Log-rank tests of cancer subtypes identified by different methods.

Dataset	NCIS	CC	SNF	WSNF(β = 1)	WSNF(β = 0.8)
BRCA	0.374	0.0634	0.0583	0.0277	**0.00483**
GBM	0.091	0.321	0.0107	0.00364	**0.00279**

doi:10.1371/journal.pone.0152792.t002

From Table 2, we see that WSNF has significantly lower p-values than other common methods in both the BRCA and GBM datasets. When β is set to 1, the weight for the features is completely determined by the miRNA-TF-mRNA regulatory network. The results show that the WSNF method is better than the other existing methods, suggesting that the information from the miRNA-TF-mRNA regulatory network helps improve the identification of the subtypes. We observe further that the method performs very well in both datasets when β is 0.8 (which is default value used for β).

Breast cancer subtypes show different expression patterns

In the previous section, we have demonstrated the performance of WSNF using the BRCA and GBM datasets. The results suggest that WSNF is capable of discovering cancer subtypes with distinct survival patterns and our method outperforms the existing cancer subtype identification methods. We investigate the mRNA, TF and miRNA expression patterns across the five different breast cancer subtypes. Similar to [8], we extract the "core samples" which are identified on the basis of their Silhouette width by removing samples with negative Silhouette width values in each subtype. There are 502 samples with positive Silhouette width values across the five subtypes. We also obtain 69 normal samples from TCGA for comparison. The heatmaps for mRNA, TF, and miRNA expression are shown in Fig 5. Taking normal group as the reference, we can see from the figure that the expression profiles between the subtypes are significantly different.

To have a closer look at the expression patterns of genes characterising each subtype, we use the *Voom* [43] method and *Limma* [44] R Package to find the differentially expressed genes (adjusted p-value<0.01) between each subtype and normal samples. We select the top 1500 differentially expressed genes in each subtype for the analysis. Fig 6 shows the overlap of differentially expressed genes across the subtypes. There are 473 common differentially expressed genes for all subtypes. Meanwhile, each subtype has their specific genes (Subtype 1: 271, Subtype 2: 82, Subtype 3: 393, Subtype 4: 291, Subtype 5: 157). The common genes across the five subtypes and the subtype-specific genes are listed in the S6 File. Although there are some common differentially expressed genes for all subtypes, their expression patterns are quite different as shown in Fig 7. In the latter section, we conduct the pathway analysis for the subtype-specific genes to explore their function characteristics in each subtype.

Alterations in regulatory networks across breast cancer subtypes

We extract the TF gene *BCL11A* to show the alterations in the miRNA-TF-mRNA regulatory network across the identified breast cancer subtypes. *BCL11A* is a proto-oncogene that has a significant effect on breast cancer [45]. As shown in Fig 8, *BCL11A* is highly expressed in Subtype 3, but lowly expressed in other subtypes. We map the patients in Subtype 3 to clinical data and find that 73.5% of the patients are in triple-negative class, including ER-, PR- and HER2-. This is consistent with the results in [45], which proved that *BCL11A* is highly expressed in triple-negative breast cancer.

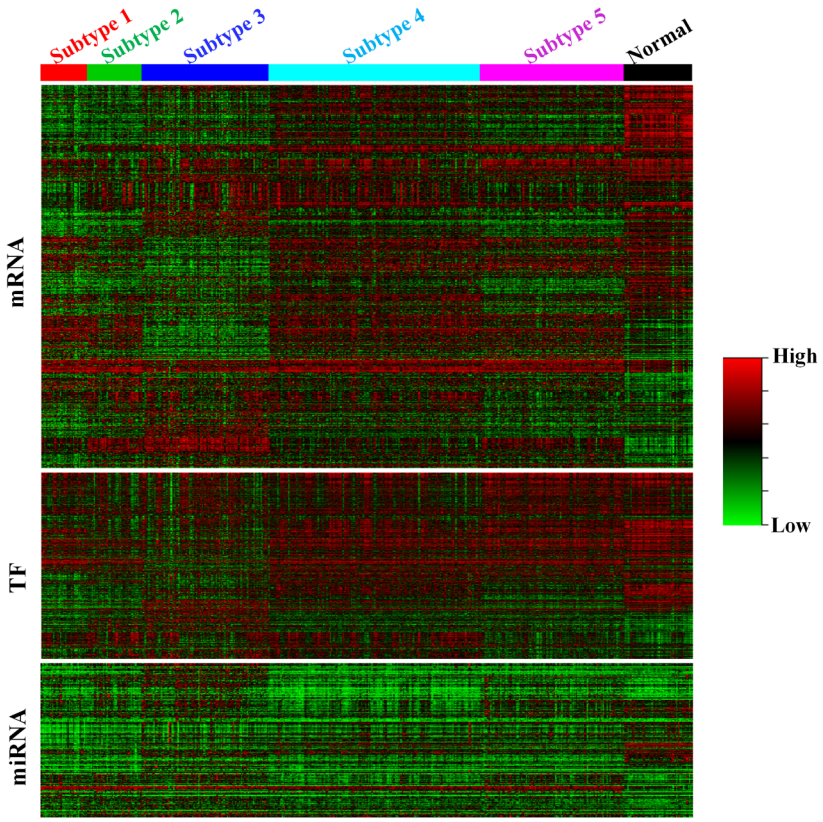

Fig 5. mRNA, TF and miRNA expression heatmap for BRCA dataset.

doi:10.1371/journal.pone.0152792.g005

The target genes of *BCL11A*, including *ANKRPD30B, MAG, BPIFB1* and *GRPR* are lowly expressed in Subtype 3 and this pattern is opposite to those in other subtypes. These different patterns suggest that *BCL11A* down regulates *ANKRPD30B, MAG, BPIFB1* and *GRPR* in breast cancer, and the expression level of *BCL11A* may be a marker of different subtypes. We also observe the co-expression between *BCL11A* and *PTPRZ1* in all subtypes and normal samples. However, the expression level of *BCL11A* and *PTPRZ1* are different across different subtypes and normal samples, suggesting that the co-expression of *BLC11A* and *PTPRZ1* are specific to the subtypes.

To investigate why *BCL11A* has low expression levels in Subtypes 1, 2 and 5 (and not very high in Subtype 4), we observe the changes in the expression levels of its upstream regulators. As in Fig 8, *miR-190b* and *ESR1* have high expression levels in the three subtypes, which is totally opposite to that in Subtype 3. This observation suggests that *miR-190b* and *ESR1* may down regulate *BCL11A* in breast cancer. The level of down regulation may characterise different breast cancer subtypes. We believe that the information of miRNA-TF-mRNA regulatory mechanisms across subtypes would provide insights into the cause of each subtype.

Top enriched pathways in different breast cancer subtypes

To investigate the pathways involved in each subtype, we conduct the pathway analysis on the differentially expressed genes characterising each subtype as shown in Fig 6. We use GeneGO Metacore^{TM} (https://portal.genego.com/) to select the top 5 significant pathways for each subtype. The genes to be analysed are the subtype-specific genes (Subtype 1: 271, Subtype 2: 82,

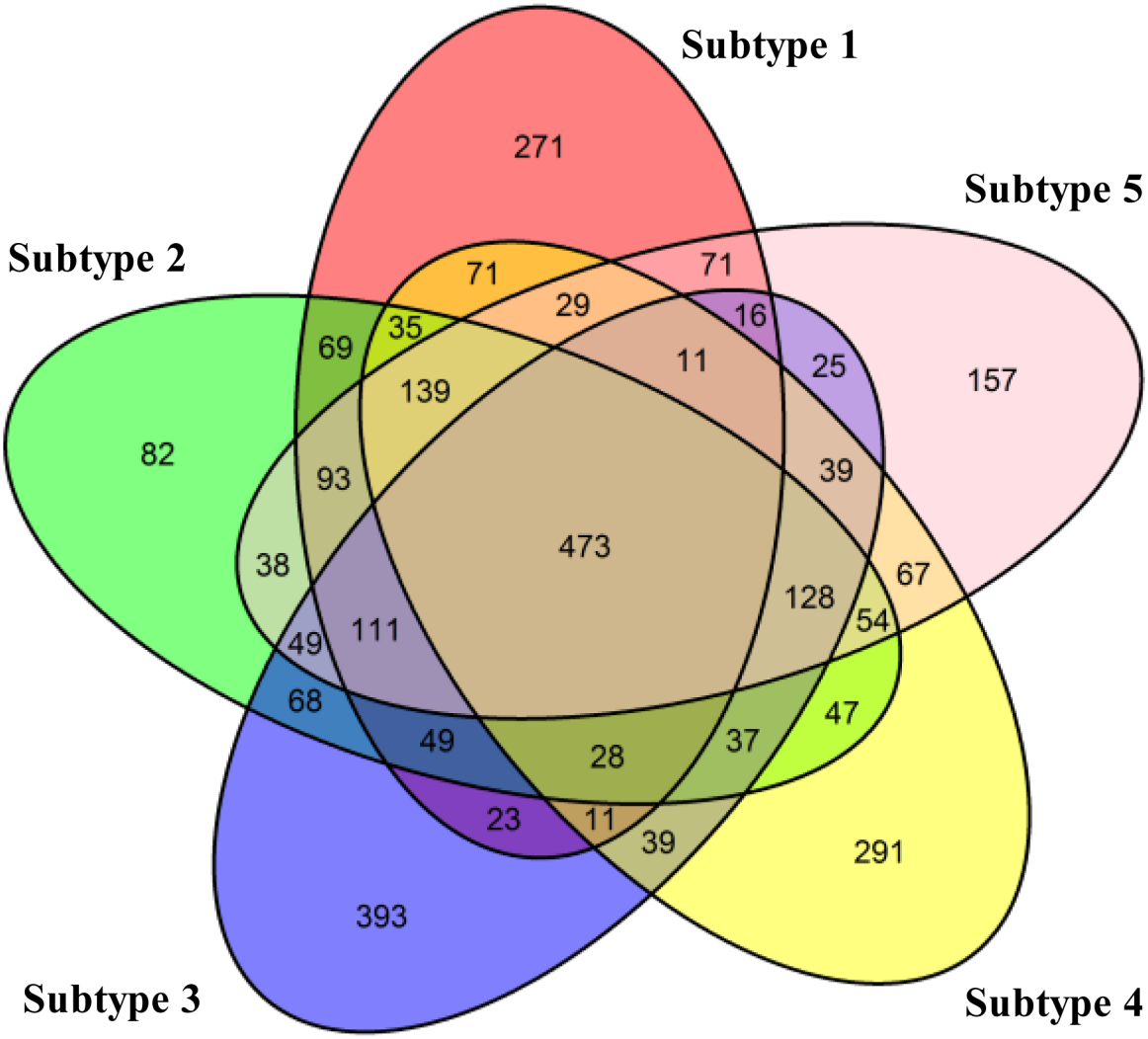

Fig 6. The overlap of the differentially expressed genes across the five subtypes of BRCA.

doi:10.1371/journal.pone.0152792.g006

Subtype 3: 393, Subtype 4: 291, Subtype 5: 157), as we wish to observe different biological pathways for different subtypes. We also conduct the pathway analysis for the 473 common genes in all the subtypes.

Table 3 shows the top 5 enriched pathways of the common genes and subtype-specific genes of each subtype. We can see from the table that the pathways are quite different between different subtypes. The significant pathways of the common genes are related to cell cycle. Pathways in Subtype 1 are related to Epithelia to Mesenchymal Transition (EMT), which implies the progression of breast carcinoma to metastasis [46]. Meanwhile, pathways in Subtype 3 are related to the neurophysiological process, Subtype 4 pathways are about the cytoskeleton remodeling, and Subtype 5 pathways are related to the immune responses. These pathways show that different subtypes have different causes.

Discussion and Conclusion

Identifying cancer subtypes is one of the important components in the personalised medicine framework, as correctly stratifying patients into subtypes will increase the chance to provide

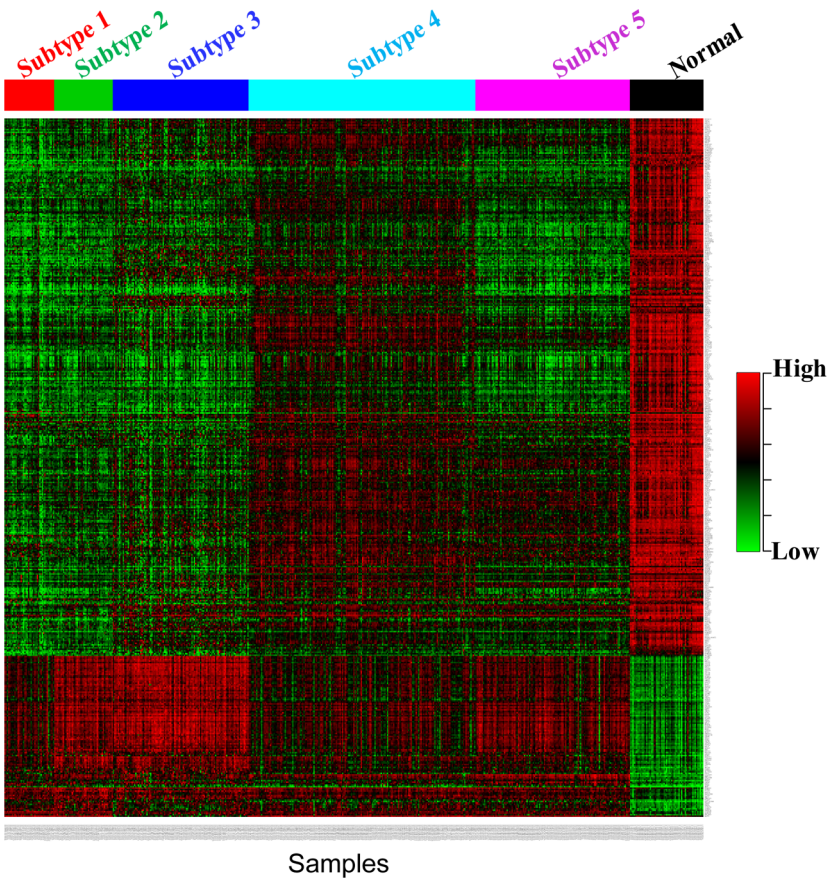

Fig 7. The heatmap of 473 common differentially expressed genes in the five subtypes of BRCA.

doi:10.1371/journal.pone.0152792.g007

the best treatment option. Computational methods have been advanced in the last decade to systematically cluster patients into groups based on their genetic profiles. Especially, SNF is an effective multi-omics data fusion method for stratification of cancer subtypes. Compared with other existing methods, SNF is time efficient and effective in uncovering subtypes with distinct survival patterns. However, similar to other existing methods, SNF is not able to exploit the biological importance of the features in building the model. Therefore, the valuable information from biological networks, such as gene regulatory networks, is not utilised in the procedure of grouping patients into subtypes. However, network information is very important for understanding the mechanisms of cancer development and progression.

In this paper, we have proposed the WSNF method. WSNF is based on SNF, but it makes use of the miRNA-TF-mRNA regulatory network to take the importance of the features into consideration. We applied WSNF to the breast cancer and glioblastoma multiform datasets, and the experimental results have shown that with the assistance of the network information WSNF outperforms the other cancer subtype identification methods that do not use the information. The results suggest that the miRNA-TF-mRNA regulatory network can provide valuable information for clustering cancer subtypes.

The performance of the WSNF method could be improved further. WSNF is based on the public databases of interactions between miRNAs, TFs, and mRNAs. Although the databases are comprehensive, they do not cover all the true interactions. Moreover, the gene regulatory networks may involve other types of molecules such as long non-coding RNAs, and they are

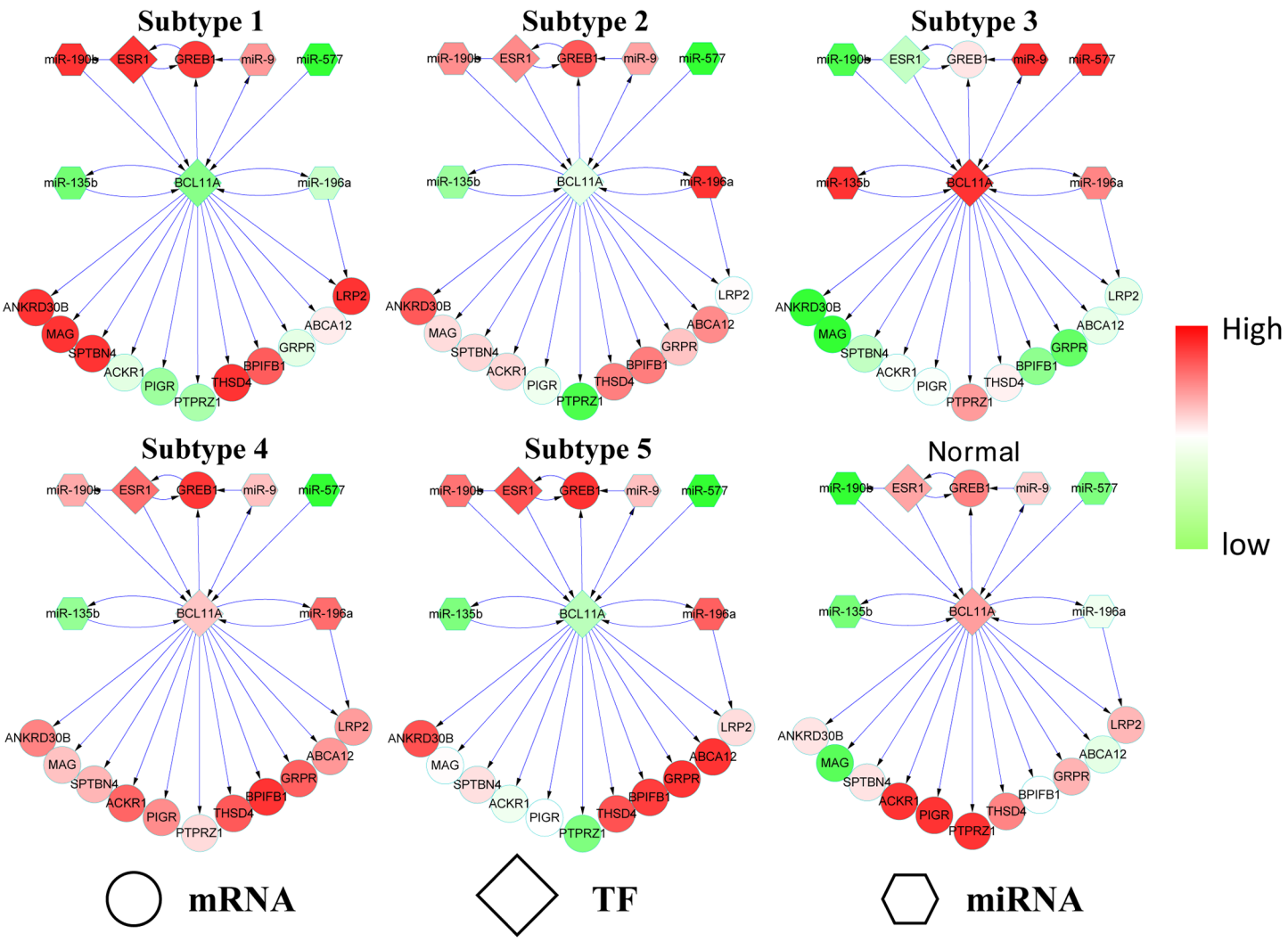

Fig 8. The expression patterns of miRNAs, TFs and mRNAs in the BCL11A network.

doi:10.1371/journal.pone.0152792.g008

not included in the gene regulatory network in this paper. Therefore, a more complete gene regulatory network with multiple types of gene regulators (can be obtained when more data become available) would help improve further the performance of WSNF in identifying significant cancer subtypes.

We have observed that the expression patterns of genes across the identified breast cancer subtypes are very different, suggesting that the expression levels of groups of genes may characterise the cancer subtypes. We have also investigated the expression patterns of genes in the sub-networks around *BCL11A* across different subtypes of breast cancer. The results show that the expression pattern of the genes in Subtype 3, where 73.5% of the patients have triple-negative (ER-, PR- and HER2-), is very different from those in other subtypes. Moreover, functional pathway analysis shows that different pathways involved in different breast cancer subtypes, suggesting that the breast cancer subtypes may be caused by different pathways. These findings are useful for domain experts to design different treatments for different breast cancer subtypes.

Table 3. Top 5 enriched pathways in five subtypes of BRCA. The *p*-values have been adjusted by the Benjamini-Hochberg (BH) method.

Datasets	Top 5 enriched pathways	Adj- *p* -value
Common	Cell cycle The metaphase checkpoint	1.337E-15
	Cell cycle Role of APC in cell cycle regulation	1.118E-10
	Cell cycle Spindle assembly and chromosome separation	5.812E-08
	Reproduction Progesterone-mediated oocyte maturation	3.553E-07
	Cell cycle Chromosome condensation in prometaphase	3.947E-07
Subtype 1	NETosis in SLE	3.209E-06
	Development WNT signaling pathway.Part 2	7.827E-05
	Cell adhesion Cell-matrix glycoconjugates	1.536E-04
	Hypoxia-induced EMT in cancer and fibrosis	1.969E-04
	Immune response IL-12 signaling pathway	2.404E-04
Subtype 2	Immune response IL-6 signaling pathway	4.784E-03
	Neurophysiological process Receptor-mediated axon growth repulsion	9.892E-03
	Signal transduction IP3 signaling	1.165E-02
	Immune response Function of MEF2 in T lymphocytes	1.258E-02
	Development Role of HDAC and calcium	1.403E-02
Subtype 3	Cell adhesion ECM remodeling	8.725E-04
	Breast cancer (general schema)	2.768E-03
	Neurophysiological process Melatonin signaling	3.299E-03
	Neurophysiological process Receptor-mediated axon growth repulsion	3.896E-03
	Action of GSK3 beta in bipolar disorder	4.203E-03
Subtype 4	Cell adhesion Gap junctions	1.212E-05
	Cytoskeleton remodeling Neurofilaments	1.132E-04
	Cytoskeleton remodeling Keratin filaments	4.832E-04
	Cell adhesion Tight junctions	4.832E-04
	Breast cancer (general schema)	7.987E-04
Subtype 5	Development Prolactin receptor signaling	4.736E-05
	Immune response ETV3 affect on CSF1-promoted macrophage differentiation	7.416E-05
	Immune response Human NKG2D signaling	1.303E-04
	Immune response TSLP signalling	1.445E-04
	Immune response Murine NKG2D signaling	1.936E-04

doi:10.1371/journal.pone.0152792.t003

In summary, we have developed a method utilising the information of miRNA-TF-mRNA regulatory network to identify cancer subtypes. The method has successfully identified subtypes in breast cancer and glioblastoma multiforme. The results provide strong indicators for further analysis of the mechanisms of the subtypes. We provide all datasets, results and scripts for readers to reproduce and further analyse the results.

Supporting Information

S1 File. TF list(1679).
(CSV)

S2 File. R scripts for our method.
(R)

S3 File. Supplementary Materials.
(PDF)

S4 File. BRCA subtype results with clinical information.
(CSV)

S5 File. GBM subtype results with clinical information.
(CSV)

S6 File. The significant differentially expressed genes in the five BRCA subtypes.
(CSV)

Author Contributions

Conceived and designed the experiments: TX TDL JL. Performed the experiments: TX TDL. Analyzed the data: TX TDL LL. Contributed reagents/materials/analysis tools: TX TDL RW BS. Wrote the paper: TX TDL LL JL.

References

1. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490(7418):61–70. doi: 10.1038/nature11412 PMID: 23000897
2. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 2014; 513:202–209. doi: 10.1038/nature13480 PMID: 25079317
3. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; 286 (5439):531–537. doi: 10.1126/science.286.5439.531 PMID: 10521349
4. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98(19):10869–10874. doi: 10.1073/pnas.191367098 PMID: 11553815
5. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proceedings of the National Academy of Sciences of the United States of America. 2001; 98(24):13790–13795. doi: 10.1073/pnas.191502998 PMID: 11707567
6. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proceedings of the National Academy of Sciences of the United States of America. 2004; 101(3):811–816. doi: 10.1073/pnas.0304146101 PMID: 14711987
7. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning. 2003; 52(1–2):91–118. doi: 10.1023/A:1023949509487
8. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer Cell. 2010; 17(1):98–110. doi: 10.1016/j.ccr.2009.12.020 PMID: 20129251
9. Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature. 2012; 489(7417):519–525. doi: 10.1038/nature11404 PMID: 22960745
10. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487(7407):330–337. doi: 10.1038/nature11252 PMID: 22810696
11. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods. 2014; 11(3):333–337. doi: 10.1038/nmeth.2810 PMID: 24464287
12. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. Bioinformatics. 2009; 25(22):2906–2912. doi: 10.1093/bioinformatics/btp543 PMID: 19759197
13. Shen R, Mo Q, Schultz N, Seshan VE, Olshen AB, Huse J, et al. Integrative subtype discovery in glioblastoma using iCluster. PLoS One. 2012; 7(4):e35236. doi: 10.1371/journal.pone.0035236 PMID: 22539962

14. Serra A, Fratello M, Fortino V, Raiconi G, Tagliaferri R, Greco D. MVDA: a multi-view genomic data integration methodology. BMC Bioinformatics. 2015; 16(1):261. doi: 10.1186/s12859-015-0680-3 PMID: 26283178
15. Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks. Nature Reviews Molecular Cell Biology. 2008; 9(10):770–780. doi: 10.1038/nrm2503 PMID: 18797474
16. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2011; 12(1):56–68. doi: 10.1038/nrg2918 PMID: 21164525
17. Dutta B, Pusztai L, Qi Y, André F, Lazar V, Bianchini G, et al. A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes. British Journal of Cancer. 2012; 106(6):1107–1116. doi: 10.1038/bjc.2011.584 PMID: 22343619
18. Liu Y, Gu Q, Hou JP, Han J, Ma J. A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. BMC Bioinformatics. 2014; 15(1):37. doi: 10.1186/1471-2105-15-37 PMID: 24491042
19. Wang J, Lu M, Qiu C, Cui Q. TransmiR: a transcription factor–microRNA regulation database. Nucleic Acids Research. 2010; 38(suppl 1):D119–D122. doi: 10.1093/nar/gkp803 PMID: 19786497
20. Vergoulis T, Vlachos IS, Alexiou P, Georgakilas G, Maragkakis M, Reczko M, et al. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. Nucleic Acids Research. 2012; 40(D1):D222–D229. doi: 10.1093/nar/gkr1161 PMID: 22135297
21. Hsu SD, Tseng YT, Shrestha S, Lin YL, Khaleel A, Chou CH, et al. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. Nucleic Acids Research. 2014; 42(D1):D78–D85. doi: 10.1093/nar/gkt1266 PMID: 24304892
22. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Research. 2009; 37(suppl 1):D105–D110. doi: 10.1093/nar/gkn851 PMID: 18996891
23. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Research. 2013; p. gkt1248.
24. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nature Genetics. 2013; 45(10):1113–1120. doi: 10.1038/ng.2764 PMID: 24071849
25. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. 2012; 489(7414):91–100. doi: 10.1038/nature11245 PMID: 22955619
26. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic Acids Research. 2014; 42(D1):D472–D477. doi: 10.1093/nar/gkt1102 PMID: 24243840
27. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research. 2014; p. gku1003.
28. Carlson M, Falcon S, Pages H, Li N. org.Hs.eg.db: Genome wide annotation for Human; 2013.
29. ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. Science. 2004; 306(5696):636–640. doi: 10.1126/science.1105136 PMID: 15499007
30. Jiang C, Xuan Z, Zhao F, Zhang MQ. TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Research. 2007; 35(suppl 1):D137–D140. doi: 10.1093/nar/gkl1041 PMID: 17202159
31. Brin S, Page L. Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer Networks. 2012; 56(18):3825–3833. doi: 10.1016/j.comnet.2012.10.007
32. Page L, Brin S, Motwani R, Winograd T. The PageRank citation ranking: bringing order to the Web. Technical Report. 1999;.
33. Higham DJ, Taylor A. The sleekest link algorithm. Institute of Mathematics and Its Applications (IMA) Mathematics Today. 2003; 39:192–197.
34. Kamvar SD, Haveliwala TH, Manning CD, Golub GH. Extrapolation methods for accelerating PageRank computations. In: Proceedings of the 12th International Conference on World Wide Web. ACM; 2003. p. 261–270.
35. Morrison JL, Breitling R, Higham DJ, Gilbert DR. GeneRank: using search engine technology for the analysis of microarray experiments. BMC Bioinformatics. 2005; 6(1):233. doi: 10.1186/1471-2105-6-233 PMID: 16176585
36. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014; 15(12):550. doi: 10.1186/s13059-014-0550-8 PMID: 25516281

37. Hastie T, Tibshirani R, Narasimhan B, Chu G. impute: Imputation for microarray data. R package version. 2012;1(0).

38. Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports Part 1. 1966; 50(3):163–170. PMID: 5910392

39. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics. 1987; 20:53–65. doi: 10.1016/0377-0427(87)90125-7

40. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, et al. The somatic genomic landscape of glioblastoma. Cell. 2013; 155(2):462–477. doi: 10.1016/j.cell.2013.09.034 PMID: 24120142

41. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474(7353):609–615. doi: 10.1038/nature10166 PMID: 21720365

42. Cancer Genome Atlas Research Network. Integrated genomic characterization of endometrial carcinoma. Nature. 2013; 497(7447):67–73. doi: 10.1038/nature12113 PMID: 23636398

43. Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014; 15(2):R29. doi: 10.1186/gb-2014-15-2-r29 PMID: 24485249

44. Smyth GK. Limma: linear models for microarray data. In: Bioinformatics and computational biology solutions using R and Bioconductor. Springer; 2005. p. 397–420.

45. Khaled WT, Lee SC, Stingl J, Chen X, Ali HR, Rueda OM, et al. BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. Nature Communications. 2015; 6. doi: 10.1038/ncomms6987 PMID: 25574598

46. Wang Y, Zhou BP. Epithelial-mesenchymal transition in breast cancer progression and metastasis. Chinese Journal of Cancer. 2011; 30(9):603. doi: 10.5732/cjc.011.10226 PMID: 21880181