

Research Paper

# Predicting Potential Drugs for Breast Cancer based on miRNA and Tissue Specificity

Liang Yu✉, Jin Zhao and Lin Gao

School of Computer Science and Technology, Xidian University, Xi'an, 710071, P.R. China.

✉ Corresponding author: lyu@xidian.edu.cn

© Ivyspring International Publisher. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY-NC) license (<https://creativecommons.org/licenses/by-nc/4.0/>). See <http://ivyspring.com/terms> for full terms and conditions.

Received: 2017.10.16; Accepted: 2017.12.14; Published: 2018.05.22

## Abstract

Network-based computational method, with the emphasis on biomolecular interactions and biological data integration, has succeeded in drug development and created new directions, such as drug repositioning and drug combination. Drug repositioning, that is finding new uses for existing drugs to treat more patients, offers time, cost and efficiency benefits in drug development, especially when in silico techniques are used. MicroRNAs (miRNAs) play important roles in multiple biological processes and have attracted much scientific attention recently. Moreover, cumulative studies demonstrate that the mature miRNAs as well as their precursors can be targeted by small molecular drugs. At the same time, human diseases result from the disordered interplay of tissue- and cell lineage-specific processes. However, few computational researches predict drug-disease potential relationships based on miRNA data and tissue specificity. Therefore, based on miRNA data and the tissue specificity of diseases, we propose a new method named as miTS to predict the potential treatments for diseases. Firstly, based on miRNAs data, target genes and information of FDA (Food and Drug Administration) approved drugs, we evaluate the relationships between miRNAs and drugs in the tissue-specific PPI (protein-protein) network. Then, we construct a tripartite network: drug-miRNA-disease. Finally, we obtain the potential drug-disease associations based on the tripartite network. In this paper, we take breast cancer as case study and focus on the top-30 predicted drugs. 25 of them (83.3%) are found having known connections with breast cancer in CTD (Comparative Toxicogenomics Database) benchmark and the other 5 drugs are potential drugs for breast cancer. We further evaluate the 5 newly predicted drugs from clinical records, literature mining, KEGG pathways enrichment analysis and overlapping genes between enriched pathways. For each of the 5 new drugs, strongly supported evidences can be found in three or more aspects. In particular, Regorafenib (DB08896) has 15 overlapping KEGG pathways with breast cancer and their p-values are all very small. In addition, whether in the literature curation or clinical validation, Regorafenib has a strong correlation with breast cancer. All the facts show that Regorafenib is likely to be a truly effective drug, worthy of our further study. It further follows that our method miTS is effective and practical for predicting new drug indications, which will provide potential values for treatments of complex diseases.

Key words: drug repositioning, miRNAs, tissue specificity, module distance

## Introduction

The identification of therapeutic approaches for the treatment of cancer is an arduous, costly, and often inefficient process. By conservative estimates, it now takes over 15 years and \$800 million to \$1 billion to bring a new drug to market [1]. Drug repositioning, which is the discovery of new indications for existing

drugs, is an increasingly attractive mode of therapeutic discovery. A repositioned drug does not need the initial six to nine years required for the development of new drugs, but instead goes directly to preclinical testing and clinical trials, thus reducing risk and costs [2]. Repositioning drugs has been

implemented in several ways. One of the well-known examples is sildenafil citrate, which was repositioned from a hypertension drug to a therapy for erectile dysfunction [3]. Drugs treat diseases by targeting the proteins related to the phenotypes arising from the disease. However, drug development does not accurately follow the “one gene, one drug, one disease” principle, which has been challenged in many cases [4] and the traditional drug repositioning methods by accident makes it hard to satisfy medical needs by successfully repositioning a large number of existing drugs. Computational methods are able to solve this question by high-level integration of available biological data and elucidation of unknown mechanisms.

In recent years, systems biology continues to make important progress to solve fundamental problems in biology and leading to practical applications in medicine and drug discovery [5]. Network-based computational systems biology emphasizes the interactions among biomolecules and highlights the network concept. Typically, a network comprises a set of nodes and edges, and is described by graph theory in a mathematical manner [6]. A node can be a biological molecule, for example, gene, RNA, protein, metabolite, and pathway. A node can also be at the phenotype level such as disease and drug. An edge can represent the complex interaction between two nodes such as protein-protein interaction, drug-disease therapeutic relationship, drug-protein target relationship, and so on. The accumulation of different high-throughput biology data, such as gene expression data, miRNA expression data and drug-target data, has made the reconstruction of biomolecular and cellular networks possible. There are many network-based methods to predict the new indications of drugs [7,8,9,10,11]. Cheng et al. built a bipartite graph composed of the approved drugs and proteins linked by drug target binary associations and relied on a supervised network-based inference method to predict drug-target interactions [12]. Chen et al. constructed a general heterogeneous network which comprised drugs and proteins linked by protein-protein sequence similarity, drug-drug chemical similarity, and the known drug-target interaction [13]. Yeh et al. [14] developed a network flow approach for identifying potential target proteins, which have a strong influence on disease genes in the context of biomolecular networks. The biomolecular networks are weighted by degree of co-expression of interacting protein pair.

More recently, many studies have demonstrated that drugs can regulate microRNA (miRNAs) expression and mature miRNAs as well as their precursors can be targeted by small molecular drugs

[15,16,17,18]. For example, Miravirsen (SPC3649) is the first miRNA-targeted drug in clinical trials, which can successfully inhibit miR-122 expression that is required by hepatitis C virus replication [19]. The expression levels of 32 miRNAs (significant up-regulation of 22 miRNAs and down-regulation of 10 miRNAs) were changed after the treatment of trichostatin A in human breast cancer cell lines [20]. miRNAs are non-coding small RNAs (~23 nucleotides) that downregulate gene expression at the post transcriptional level by inhibiting translation or initiating mRNA degradation and are dysregulated in most of human cancers [21]. Increasingly evidences have demonstrated that miRNAs play significant roles in many important biological processes, such as cell growth [22,23,24], cellular signaling [25], tissue development [26] and disease process [27,28,29,30,31,32]. Although only approximately 2000 miRNAs exist in humans, they regulate 30% of all genes [33]. miRNAs have been identified to play a crucial role in various human disease, especially in cancers. Therefore, targeting miRNAs with drugs will provide a new type of therapy for complex diseases [34] and a new direction for drug repositioning. However, few computational researches predict drug-disease relationships based on miRNA data. Moreover, many genes with tissue-specific expression and function are expected to underlie many human diseases [35,36,37].

Therefore, in this study, we propose a new method based on miRNA data and tissue specificity of diseases, named as *miTS*, to predict potential drugs for diseases. The framework of *miTS* is shown in Figure 1. Firstly, we download miRNA expression data of diseases from TCGA [38], miRNA-target gene relationship data from three experimentally validated databases: miRecords [39], miRTarbase [40] and TarBase [41], and the drug-target gene data from Drugbank [42] and KEGG [43]. Secondly, we select differentially expressed miRNAs of diseases based on a threshold and preprocess the target information of FDA approved drugs. Thirdly, we evaluate the relationships between miRNAs and drugs in the tissue-specific PPI network. And then, we construct a tripartite network: drug-miRNA-disease. Finally, we obtain the potential drug-disease associations based on the tripartite network. In this paper, we take breast cancer as case study and evaluate the results from CTD (Comparative Toxicogenomics Database) benchmark, clinical records, literature mining, KEGG pathways enrichment analysis and overlapping genes between enriched pathways. In the top-30 drugs, we find 5 new drugs for breast cancer. In particular, Regorafenib (DB08896) has 15 overlapping KEGG pathways with breast cancer and their p-values are all very small. In addition, whether in the literature

curation or clinical validation, Regorafenib has a strong correlation with breast cancer. All the facts show that Regorafenib is likely to be a truly effective drug, worthy of our further study.

## Methods

### Datasets

#### Drug-target data

FDA-approved drugs of human and their corresponding targets are downloaded from KEGG database and Drugbank. We merge the two datasets and get 1,732 drugs, 1,714 targets and 12,361 drug-target pairs.

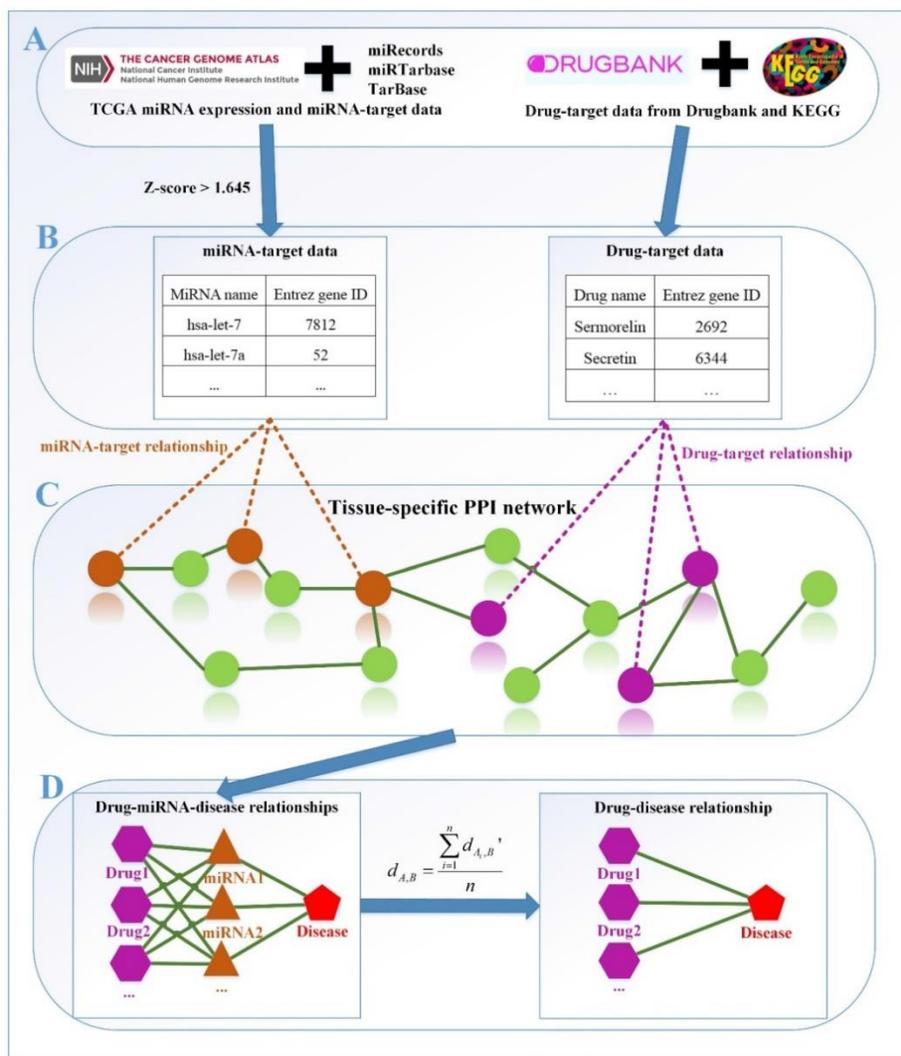
#### miRNA-target data

The target genes of miRNAs are downloaded

from miRecords, miRTarbase, and TarBase databases. We merge the three datasets and get 340 miRNAs, 2,028 targets and 3,652 miRNA-target pairs.

#### miRNA-disease data

The miRNA-disease curated relationships are downloaded from HMDD (the Human MicroRNA Disease Database) [44]. HMDD presents more detailed and comprehensive annotations to the human miRNA-disease association data, including miRNA-disease association data from the evidence of genetics, epigenetics, circulating miRNAs, and miRNA-target interactions. Finally, we get 578 miRNAs, 383 diseases and 6,448 miRNA-disease relationships.



**Figure 1.** The framework of our method *miTS*. (A) Data preparation: miRNA expression data of breast cancer got from TCGA, miRNA-target gene data got from miRecords, miRTarbase and TarBase, and drug-target gene data got from Drugbank and KEGG. (B) Data preprocessing: we use Z-score to obtain the differentially expressed miRNAs for diseases and preprocess the target information of drugs. (C) In the tissue-specific PPI network, the targets of drug and miRNA are mapped to the PPI network. Orange nodes represent the target genes of miRNAs. Purple nodes represent the target genes of drugs. Green nodes represent the background genes. (D) Based on the module distance algorithm, we construct a drug-miRNA-disease tripartite network, and then based on the tripartite network, we get potential drugs for diseases.  $d_{A,B}$  represents the association score between a drug and a disease.

### miRNA expression data

Taking breast cancer as case study, we download the miRNAs expression data related with breast cancer from TCGA and get a matrix of 503 rows and 1,189 columns, row representing miRNA, column representing cancer sample, and the values in the matrix representing the RPKM (Reads Per Kilobase per Million mapped reads) for the miRNAs. We take the mean value of the RPKM values for 1,189 samples as the final value.

### Disease-gene data

The genes related with breast cancer are downloaded from OMIM [45] database.

### Tissue-specific PPI Interaction network

We download the mammary tissue-specific PPI network marked as "Top Edges" from GIANT (Genome-scale Integrated Analysis of gene Networks in Tissues) database [46] (<http://giant.princeton.edu/>) (2017 version). GIANT proposes a tissue-specific benchmark to automatically up-weight datasets relevant to a tissue from a large data of different tissues and cell-types. Finally, we get 15,269 proteins and 883,071 protein-protein interactions. The weights on the edges are proportional to the relationships between nodes. In order to apply module distance algorithm [47] (details shown in Figure 2) to calculate the relationships between drugs and miRNAs, we use the Gaussian kernel  $e^{-w^2}$  to transfer protein-protein closeness  $w$  to protein-protein distance  $w'$ , as shown in formula (1).

$$w' = e^{-w^2} \tag{1}$$

### Screening differentially expressed miRNAs

In order to obtain the differentially expressed miRNAs of breast cancer, we first filter the miRNAs expression data downloaded from TCGA. For a miRNA  $r$ , we use formula (2) to calculate its  $Z$  - score.

$$Z - score = \frac{R - mean(r)}{\sigma(r)} \tag{2}$$

Where  $R$  is the RPKM value of miRNA  $r$ ;  $mean(r)$  and  $\sigma(r)$  represent mean value and standard deviation of  $r$ , respectively. Then we choose  $Z - score = 1.645$  (p-value = 0.05) as threshold to screen differentially expressed miRNAs. Finally, we get a total of 40 differentially expressed miRNAs of breast cancer (see Table 1). In Table 1, the miRNAs marked by "\*" represent they have connections with breast cancer in HMDD. We find 34 of 40 (85%) differentially expressed miRNAs are related with breast cancer, which indicates that miRNAs associated with breast

cancer tend to be highly expressed in breast cancer patients. Then, we choose the 34 miRNAs marked by "\*" in Table 1 for further study.

**Table 1.** Differentially expressed miRNAs of breast cancer

miRNA name	Z-score	miRNA name	Z-score	miRNA name	Z-score
hsa-mir-21*	3.32	hsa-mir-375*	2.32	hsa-mir-23a*	1.91
hsa-mir-22*	2.94	hsa-mir-101-1*	2.29	hsa-mir-199a-2*	1.90
hsa-mir-10b*	2.93	hsa-mir-200c*	2.28	hsa-mir-126*	1.90
hsa-mir-30a*	2.85	hsa-mir-25*	2.27	hsa-mir-100*	1.86
hsa-mir-148a*	2.77	hsa-let-7a-3*	2.21	hsa-let-7c*	1.79
hsa-mir-99b	2.73	hsa-let-7a-1*	2.21	hsa-mir-151	1.78
hsa-mir-143*	2.73	hsa-mir-30d*	2.19	hsa-mir-199a-1*	1.73
hsa-mir-182*	2.72	hsa-mir-92a-2*	2.18	hsa-mir-26a-2*	1.72
hsa-let-7b*	2.61	hsa-let-7f-2*	2.12	hsa-mir-142	1.72
hsa-mir-10a*	2.56	hsa-mir-93*	2.03	hsa-mir-29c*	1.70
hsa-mir-103-1	2.50	hsa-mir-29a*	2.03	hsa-mir-181a-1*	1.69
hsa-let-7a-2*	2.44	hsa-mir-28	2.00	hsa-mir-141*	1.66
hsa-mir-30e	2.38	hsa-mir-199b*	1.98		
hsa-mir-183*	2.37	hsa-mir-203*	1.94		

The miRNAs marked by "\*" represent they have relationship with breast cancer in HMDD.

### Construct drug-miRNA-disease tripartite network

The relationship between a miRNA and a drug is derived by measuring the correlation between their target sets. Because miRNA target genes, drug target genes and protein-protein interaction (PPI) networks remain largely incomplete, we calculate the distance between two modules based on the shortest path in incomplete networks [44].

Figure 2 gives an example to calculate the distance between miRNA  $A$  and drug  $B$  in a weighted tissue-specific PPI network. As shown in Figure 2, miRNA  $A$  has three target genes, marked as  $a, b, c$  and drug  $B$  has four targets, marked as  $c, d, e, f$ . For the node  $a$ , its distance to targets  $\{c, d, e, f\}$  of drug  $B$  are 0.8, 1.0, 1.1 and 1.9 respectively, so its shortest distance to drug  $B$  is 0.8. In this way, we can obtain the distances between each node in gene set  $\{a, b, c\}$  and drug  $B$ , and the distances between each node in target set  $\{c, d, e, f\}$  and miRNA  $A$ , shown in Figure 2. Finally, the distance between miRNA  $A$  and drug  $B$ ,  $d_{A,B}'$ , is equals to the sum of all the distances divided by the total number of nodes related to miRNA  $A$  and drug  $B$ . Here, the total number is 7.

Through the above calculation process, we get 1,017 drugs, 25 miRNAs and 25,425 drug-miRNA relationships. Combining the drug-miRNA relations with the miRNA-breast cancer information, we construct a drug-miRNA-breast cancer tripartite network.

### Predicting potential drugs for breast cancer

Based on the drug-miRNA-breast cancer

tripartite network, we predicting potential drugs for breast cancer. If a drug and breast cancer have common miRNA neighbors, there will be a connection between them. Finally, there are 25 common miRNAs between drugs and breast cancer. We use formula (3) to calculate the average distance between the 25 miRNAs related to breast cancer and drugs as the drug-breast cancer relationship distance score,  $d_{C,B}$ .

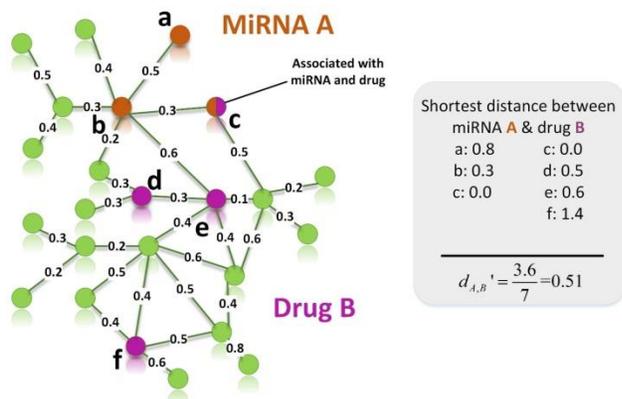
$$d_{C,B} = \frac{\sum_{i=1}^n d_{C_i,B}'}{n} \tag{3}$$

Where  $d_{C_i,B}'$  represents the distance between the  $i$ -th miRNA of disease  $C$  and the drug  $B$  (the calculation of  $d_{C_i,B}'$  as shown in Figure 2);  $n$  represents the number of miRNAs corresponding to disease  $C$ . Here,  $C$ = breast cancer and  $n$ = 25.

In order to make the drug-disease distances be proportional to their direct correlations, we use formula (4) to normalize  $d_{C,B}$  as  $S_{C,B}$ :

$$S_{C,B} = \frac{Max_d - d_{C,B}}{Max_d - Min_d} \tag{4}$$

Where  $Max_d$  and  $Min_d$  represent the maximum and the minimum of all the drug-disease distances, respectively;  $d_{C,B}$  represents the distance between disease  $C$  and drug  $B$ ;  $S_{C,B}$  represents the direct association between disease  $C$  and drug  $B$ .



**Figure 2.** An example for calculating the distance between target set of miRNA A and target set of drug B. Orange and purple nodes represent genes related to miRNA A and drug B, respectively. Node  $c$  is a shared node, so it is marked by two colors.

## Results

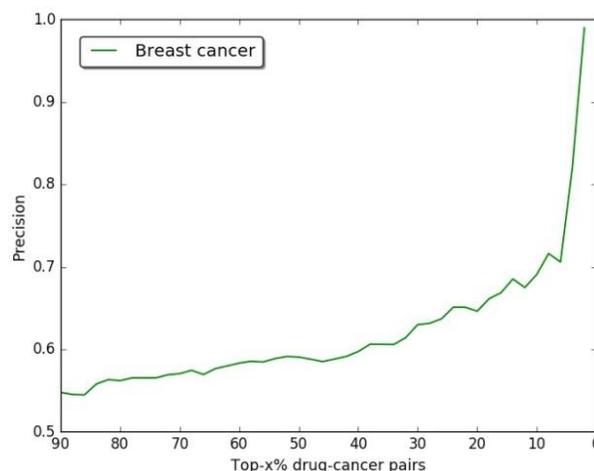
### CTD benchmark verification

In our study, we choose breast cancer as case, the drug-breast cancer associations are ranked in descending order according to their scores. In order to

verify the accuracy of our results, we use the drug-breast cancer relationships data in Comparative Toxicogenomics Database (CTD) [48] as benchmark. As shown in Figure 3, we give the precision curves of predicted drug-breast cancer relationship results. For each given threshold, the precision of our method is calculated by formula (5).

$$precision = \frac{P_{CTD}}{P} \tag{5}$$

Where  $P$  represents the number of predicted drug-disease pairs;  $P_{CTD}$  represents the number of drug-disease pairs, which can be found in CTD database.



**Figure 3.** The precision of our predictions at different top-x% drug-breast cancer pairs.

In Figure 3, we give the precision curves of predicted drug-breast cancer pairs at different top-x%. From the figure, we find the higher the associations ranking, the higher the accuracy. Hence, for the breast cancer, we choose top 30 drugs for further analysis. The top 30 drugs related to breast cancer are shown in Table 2. We validate the 30 drugs by CTD database and find 11 (36.7%) of them are marked as “therapeutic (T)”, which means that they have a highly correlation with breast cancer. In addition, we find 14 of the rest 19 drugs also have connections with breast cancer in CTD database with inference score [49] over 0 and they are marked as “Ref” in Table 2. The inference score [46] reflects the degree of similarity between CTD chemical-gene-disease networks and a similar scale-free random network, which is computed as shown below:

$$Y = -\ln[P\{G \text{ associated with both } C \text{ and } D | k, n_G\} P(\text{no other } G \text{ connects } C \text{ and } D | k, n_G)] \tag{6}$$

where  $Y$  represents the inference score;  $P$  represents the probability that a vertex in a large

network interacts with another vertex decays according to a power law [50];  $G$ ,  $C$ , and  $D$  represent a gene, chemical, and disease respectively;  $k$  represents the number of connection between  $G$ ,  $C$ , or  $D$ ;  $n_G$  represents a gene set. The higher the  $Y$  score, the more likely the inference network has atypical connectivity [51]. That is to say, there are 83.3% (25/30) drugs can be found in the CTD database and we predict five potential drugs for breast cancer (DB08871, DB00031, DB08813, DB08896, and DB06813, marked as boldface in Table 2).

**Table 2.** The top 30 drugs related to breast cancer

Rank	Drugbank ID	Drug name	Marker	Inference Score	Similarity Score
1	DB08818	Hyaluronic acid	Ref	61.1	1.00000
2	DB00570	Vinblastine	T	40.6	0.97977
3	DB00642	Pemetrexed	T	12.38	0.97666
4	DB01169	Arsenic trioxide	T	212.32	0.96998
5	DB00242	Cladribine	Ref	14.83	0.96343
6	DB04967	Lucanthone	Ref	45.07	0.96120
7	DB09073	Palbociclib	T	53.75	0.96083
8	DB02701	Nicotinamide	Ref	63.26	0.96013
9	DB01005	Hydroxyurea	Ref	32.57	0.95909
10	DB01204	Mitoxantrone	T	25.1	0.95847
11	DB00309	Vindesine sulfate	T	2.54	0.95388
12	DB00361	Vinorelbine	T	4.36	0.95388
13	<b>DB08871</b>	<b>Eribulin mesylate</b>	<b>None</b>	<b>None</b>	<b>0.95388</b>
14	DB01394	Colchicine	Ref	50.29	0.95213
15	DB01229	Paclitaxel	T	111.41	0.95050
16	DB01248	Docetaxel	T	72.35	0.95050
17	DB00440	Trimethoprim	Ref	6.82	0.94971
18	DB01179	Podofilox	Ref	2.87	0.94918
19	DB05260	Gallium nitrate	Ref	19.88	0.94671
20	DB00441	Gemcitabine	T	112.67	0.94582
21	<b>DB00031</b>	<b>Tenecteplase</b>	<b>None</b>	<b>None</b>	<b>0.94357</b>
22	<b>DB08813</b>	<b>Nadroparin</b>	<b>None</b>	<b>None</b>	<b>0.94325</b>
23	DB00432	Trifluridine	Ref	12.49	0.94205
24	DB01073	Fludarabine	Ref	59.43	0.94184
25	DB00694	Daunorubicin	Ref	85.05	0.94044
26	DB00970	Dactinomycin	Ref	98.94	0.93988
27	<b>DB08896</b>	<b>Regorafenib</b>	<b>None</b>	<b>None</b>	<b>0.93853</b>
28	<b>DB06813</b>	<b>Pralatrexate</b>	<b>None</b>	<b>None</b>	<b>0.93853</b>
29	DB00563	Methotrexate	T	123.36	0.93799
30	DB00615	Rifabutin	Ref	2.89	0.93433

Ranked by drug-breast cancer similarity score. Marker has three values: T(therapeutic), Ref (inferred by genes) and None (no record in CTD database). Inference Score represents the score for the inference based on the topology of the network consisting of the chemical, disease, and one or more genes used to make the inference.

## Clinical evaluation

For the five predicted drugs, we further analyze them based on the ClinicalTrials.gov (<https://clinicaltrials.gov/>). ClinicalTrials.gov is a registry and results database of publicly and privately supported clinical studies of human participants conducted around the world. Currently, it lists

242,537 studies with locations in all 50 states and in 198 countries (April 25, 2017). From the ClinicalTrials.gov, we can find 84 records for drug Eribulin mesylate (DB08871) treat breast cancer. For example, "Eribulin Mesylate Phase IV Clinical Trial in Korean Patients with Metastatic or Locally Advanced Breast Cancer (ESKIMO) (NCT01961544)", the purpose is assessing the safety of Eribulin which is approved for the treatment of the patients in Korea with locally advanced or metastatic breast cancer; "Eribulin with Trastuzumab as First-line Therapy for Locally Recurrent or Metastatic HER2 Positive Breast Cancer (NCT01269346)", the purpose is evaluating the safety and efficacy of Eribulin mesylate in combination with trastuzumab as first line treatment in female subjects with locally recurrent or metastatic human epidermal growth factor receptor (HER2) positive breast cancer; "Eribulin Mesylate in Treating Patients with Previously Treated Metastatic Breast Cancer (NCT01908101)", and so on. For drug Nadroparin (DB08813), we find one record: "Prevention of Venous and Arterial Thromboembolism, in Cancer Patients Undergoing Chemotherapy, With a Low Molecular Weight Heparin (Nadroparin Calcium) (NCT00951574)", 1200 patients with lung, breast, gastrointestinal (stomach, colon-rectum, pancreas), ovarian or head and neck cancer undergoing chemotherapy will be randomly assigned in a 2:1 ratio and in double-blind conditions to a treatment with subcutaneous low molecular weight heparin (nadroparin calcium, one injection/day) or placebo for the overall duration of chemotherapy or up to a maximum of 4 months. For drug Regorafenib (DB08896), we find three records related with breast cancer, "Refametinib in Combination with Regorafenib in Patients with Advanced or Metastatic Cancer (NCT02168777)", "Effect of Regorafenib on Digoxin and Rosuvastatin in Patients with Advanced Solid Malignant Tumors (NCT02106845)", and so on. For drug Pralatrexate (DB06813), we find a clinical study of Pralatrexate in 22 female patients with previously-treated breast cancer (NCT01118624). Only one drug, Tenecteplase (DB00031) was not found in ClinicalTrials.gov.

## Literature curation

In the above section, the top 30 drugs related with breast cancer are validated by CTD database and Clinical database. After our analysis, we obtain five potential drugs (Eribulin mesylate, Tenecteplase, Nadroparin, Regorafenib, Pralatrexate) for breast cancer. In ClinicalTrials.gov database, only one drug, Tenecteplase, cannot be found its corresponding record. In this section, we will analyze the five potential drugs for breast cancer by literature mining.

The summarized results are shown in Table 3.

Eribulin mesylate (DB08871) is an anticancer drug marketed by Eisai Co. under the trade name Halaven, which was approved by the U.S. Food and Drug Administration (FDA) on November 15, 2010, to treat patients with metastatic breast cancer [52]. In 2016, Kurebayashi J et al. investigated the combined effects of Eribulin and antiestrogens. They used a panel of eight breast cancer cell lines, including five estrogen receptors (ER)-positive and three ER-negative cell lines. The results of this study demonstrate that Eribulin had potent antitumor effects on estrogen-stimulated ER-positive breast cancer cells [53].

**Table 3.** Literature verification for the five potential drugs

Drug names	Drugbank ID	Original indications	Supported references for breast cancer
Eribulin mesylate	DB08871	anticancer drug	[49] [50]
Nadroparin	DB08813	anticoagulan	[51]
Regorafenib	DB08896	oral multi-kinase inhibitor	[52][53][54]
Pralatrexate	DB06813	anticancer drug	[55][56]
Tenecteplase	DB00031	tissue plasminogen activator	[57][58][59]

Nadroparin (DB08813) is an anticoagulant belonging to a class of drugs called low molecular weight heparins (LMWHs), which is used in general and orthopedic surgery to prevent thromboembolic disorders. In 2015, Sun Y et al. [54] used the MTT test to observe the effect of different concentrations of nadroparin on the growth capacity of breast cancer cells MDA-MB-231. The purpose was to study the effect of nadroparin in the migration of breast cancer cells MDA-MB-231 and its action mechanism. The results show that nadroparin can inhibit the growth capacity of breast cancer cells MDA-MB-231 and the migration and invasion of breast cancer cells MDA-MB-231. Its mechanism is to down-regulate MMP-2 and MMP-9 expressions after combining with Integrin  $\beta$ 3.

Regorafenib (DB08896) is an oral multi-kinase inhibitor developed by Bayer which targets angiogenic, stromal and oncogenic receptor tyrosine kinase (RTK). Regorafenib has been demonstrated to increase the overall survival of patients with metastatic colorectal cancer [55]. Stalker L et al. using regorafenib in mammary tumor cell lines, the results show regorafenib may prove clinically useful in inhibiting breast cancer cell migration and metastasis [56]. Su J C et al. investigated the potential of regorafenib to suppress metastasis of triple-negative breast cancer (TNBC) cells through targeting SHP-1/p-STAT3/VEGF-A axis and found a significant correlation between cancer cell migration

and SHP-1/p-STAT3/VEGF-A expression in human TNBC cells [57].

Pralatrexate (DB06813) is an anti-cancer drug. It is the first drug approved as a treatment for patients with relapsed T-cell lymphoma [58]. Pralatrexate results in increased activity of CASP3 protein, which has been found to be necessary for normal brain development as well as its typical role in apoptosis, where it is responsible for chromatin condensation and DNA fragmentation [59].

Tenecteplase (DB00031) is a tissue plasminogen activator (tPA) produced by recombinant DNA technology using an established mammalian cell line and used as a thrombolytic drug. Nielsen VG et al. [60] to study whether tissue-type plasminogen activator (tPA) in plasma obtained from patients with breast cancer, lung cancer, pancreatic cancer and colon cancer is less than that obtained from normal individuals. The results show that tissue-type plasminogen activator-induced fibrinolysis in breast cancer, lung cancer, pancreatic cancer and colon cancer patients is enhanced. Sumiyoshi K et al. [61] found that the increase in levels of plasminogen activator and type-1 plasminogen activator inhibitor in human breast cancer may play a role in tumor progression and metastasis. Although we have not found the relationship between tenecteplase (DB00031) and breast cancer through the literatures, the drug had the similar effects as nadroparin [62]. Therefore, we infer that tenecteplase is likely to have effect on breast cancer.

### KEGG pathway functional enrichment analysis

In this section, we will further make KEGG pathway enrichment analysis on five potential drugs and their associated disease. KEGG (<http://www.kegg.jp/> or <http://www.genome.jp/kegg/>) is an encyclopedia of genes and genomes [63]. Its primary goal is to assign functional meanings to genes and genomes both at the molecular and higher levels. Thus, drugs or diseases can be associated with certain pathways through their related genes. If a drug has overlapping KEGG pathways with a disease, the drug and the disease may have great relevance. That is, the drug may treat or cause the disease by acting on the overlapping pathways.

We use DAVID [64,65] functional annotation tool for KEGG pathway enrichment analysis. DAVID provides a comprehensive set of functional annotation tools for investigators to understand biological significance of a large number of genes. For any given gene list, DAVID is able to visualize genes on BioCarta & KEGG pathway maps, identify enriched biological themes, especially GO terms, and so on. Therefore, we use DAVID to identify overlapping

KEGG pathways between potential drugs and breast cancer. The  $p$ -value is set to be less than 0.05.

We find Nadroparin and Regorafenib have 4 and 15 overlapping KEGG pathways with breast cancer, respectively. The details are shown in Table 4. From Table 4, we can find their corresponding  $p$ -values are very small.

**Table 4.** Overlapping KEGG pathways between potential drugs and breast cancer

Drug Name	Overlapping enriched pathways	p-value
Nadroparin	hsa05210: Colorectal cancer	0.00897
	hsa05161: Hepatitis B	0.02098
	hsa05166: HTLV-I infection	0.03704
	hsa05200: Pathways in cancer	0.04687
Regorafenib	hsa04015: Rap1 signaling pathway	3.41E-14
	hsa04014: Ras signaling pathway	7.18E-14
	hsa04151: PI3K-Akt signaling pathway	3.36E-10
	hsa05230: Central carbon metabolism in cancer	4.38E-08
	hsa05215: Prostate cancer	2.21E-07
	hsa05200: Pathways in cancer	1.13E-06
	hsa05218: Melanoma	4.76E-06
	hsa05214: Glioma	1.65E-04
	hsa05221: Acute myeloid leukemia	0.00404
	hsa05205: Proteoglycans in cancer	0.00433
	hsa05220: Chronic myeloid leukemia	0.00660
	hsa04012: ErbB signaling pathway	0.00952
	hsa05206: MicroRNAs in cancer	0.01157
	hsa05231: Choline metabolism in cancer	0.01269
	hsa04722: Neurotrophin signaling pathway	0.01761

Although the drug Eribulin mesylate has not overlapping functional pathways with the breast cancer at present, it can be enriched to “hsa04540: Gap junction”. In fact, protein connexin 43 (Cx43), a part of intercellular gap junctions, is frequently down-regulated in tumors [66]. Studies have demonstrated that gap junctions (GJs) composed of connexin (Cx) proteins have the potential to modulate drug chemosensitivity in multiple tumor cells [67].

For drug Tenecteplase, we find one function enrichment pathway: “hsa04610: Complement and coagulation cascades”. In 2016, based on the microarray data of GSE3467 from Gene Expression Omnibus(GEO) database, Yu J et al. [68] identified the differentially expressed genes (DEGs) between 9 PTC samples and 9 normal controls. The purpose was predicted key genes and pathways in papillary thyroid carcinoma. Their results showed that the highly expressed genes in papillary thyroid carcinoma were mainly enriched on the “hsa04610: Complement and coagulation cascades” functional pathway. As for Pralatrexate, because it has only two targets: DHFR and TYMS, it has few related KEGG pathways. That is the main reason that Pralatrexate has no overlapping KEGG pathways with breast cancer at present.

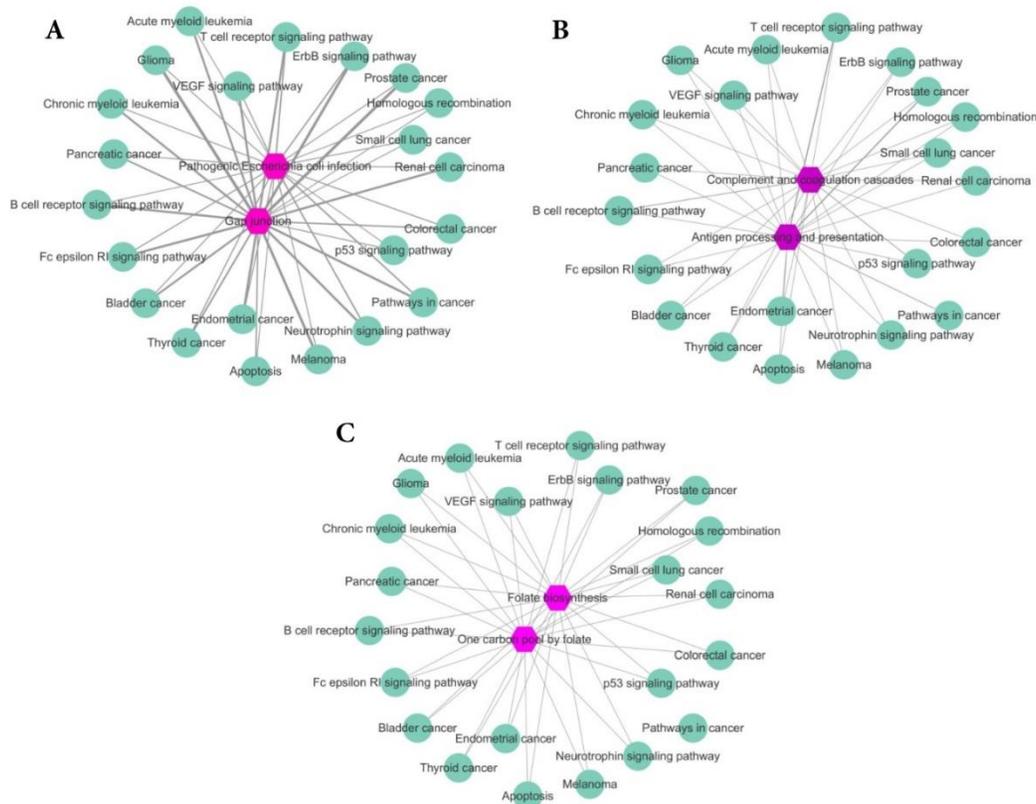
## Overlapping genes between enriched KEGG pathways

To further analyze our results, for Eribulin mesylate, Tenecteplase and Pralatrexate, we count the common genes between enriched pathways of each drug and those of breast cancer. The more common genes, the stronger relationship between the drug and disease. The results are shown in Figure 4A-C, respectively. The purple hexagon nodes represent the enriched pathways of a drug. The light green circular nodes represent breast cancer enriched pathways. The width of edges represents the number of common genes between two pathway sets. The wider the edge, the more the number of common genes. From Figure 4, we can find the three drugs Eribulin mesylate, Tenecteplase and Pralatrexate all have strong connection with breast cancer, which further imply the three drugs are likely to be the potential treatments of breast cancer.

## Discussions and conclusions

At present, “undruggable” proteins can be targeted via their miRNA gene regulators, enabling the treatment of diseases that seem impossible to cure. Human diseases result from the disordered interplay of tissue- and cell lineage-specific processes. Therefore, here we propose a new method *miTS* to predict new indications of drugs based on miRNA data and the tissue specificities of diseases. Taking breast cancer as case study, we predict five potential drugs and analyze them from five aspects: CTD benchmark, clinical records, literature curation, KEGG pathway functional enrichment analysis and overlapping genes between enriched KEGG pathways. We find five new drugs which are supported at least in three ways. In particular, Regorafenib (DB08896) has 15 overlapping KEGG pathways with breast cancer and their  $p$ -values are all very small. In addition, whether in the literature curation or clinical validation, Regorafenib has a strong correlation with breast cancer. All the evidence shows Regorafenib is likely to be a truly effective drug, worthy of our further study. The results have demonstrated the performance of our model and the feasibility of drug repositioning based on miRNA data and tissue specificity.

Due to the incompleteness of data, there may be some biases in our method. With the continuous improvement of data, our method *miTS* will find more effective drugs for disease treatment. All in all, our research reveals a promising perspective to predict drug-disease relationships and seeks new opportunities for drug repositioning.



**Figure 4.** The common genes between enriched pathway sets of drugs and breast cancer. The purple hexagon nodes represent the enriched pathways of a drug. The light green circular nodes represent breast cancer enriched pathways. The width of edges represents the number of common genes between two pathway sets. The wider the edge, the more the number of common genes. A. The common genes between enriched pathway sets of Eribulin mesylate and breast cancer. B. The common genes between enriched pathway sets of Tenecteplase and breast cancer. C. The common genes between enriched pathway sets of Pralatrexate and breast cancer.

## Abbreviations

miTS, miRNA and Tissue Specificity; CTD, Comparative Toxicogenomics Database; KEGG, Kyoto Encyclopedia of Genes and Genomes; DAVID, the Database for Annotation, Visualization and Integrated Discovery; FDA, Food and Drug Administration; GEO, Gene Expression Omnibus.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 61672406), the Fundamental Research Funds for the Central Universities (No. JB180307), the National Natural Science Foundation of China (Nos. 61532014, 91530113, 61502363), the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2016JQ6057).

## Competing Interests

The authors have declared that no competing interest exists.

## References

- Adams C P, Brantner V V. Estimating the cost of new drug development: is it really \$802 million? *Health affairs*, 2006; 25(2): 420-428.
- Ashburn T T, Thor K B. Drug repositioning: identifying and developing new uses for existing drugs. *Nature reviews Drug discovery*, 2004; 3(8): 673-683.
- Novac N. Challenges and opportunities of drug repositioning. *Trends in pharmacological sciences*, 2013; 34(5): 267-272.
- Hopkins A L. Network pharmacology: the next paradigm in drug discovery. *Nature chemical biology*, 2008; 4(11): 682-690.
- Wang Y, Zhang X S, Chen L. Modelling biological systems from molecules to dynamical networks. *BMC systems biology*, 2012; 6(1): S1.
- Barabasi A L, Oltvai Z N. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 2004; 5(2): 101-113.
- Liang Y, Xiaoke M, Long Z, et al. Prediction of new drug indications based on clinical data and network modularity. *Scientific Reports*, 2016; 6:32530.
- Liang Y, Ruidan S, Bingbo W, et al. Prediction of novel drugs for hepatocellular carcinoma based on multi-source random walk. *IEEE/ACM Trans Comput Biol Bioinform*. 2017; 14(4):966-977.
- Liang Y, Bingbo W, Xiake M, et al. The extraction of drug-disease correlations based on module distance in incomplete human interactome. *BMC System Biology*. 2016; 10(Suppl 4):111.
- Liang Y, Jianbin H, Zhixin M, et al. Inferring drug-disease associations based on known protein complexes. *BMC medical genomics*, 2015; 8: S2.
- Liang Y, Jin Z, Lin G. Drug repositioning based on triangularly balanced structure for tissue-specific diseases in incomplete interactome. *Artificial Intelligence in Medicine*. 2017;77: 53-63.
- Cheng F, Liu C, Jiang J, et al. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol*, 2012; 8(5): e1002503.
- Chen X, Liu M X, Yan G Y. Drug-target interaction prediction by random walk on the heterogeneous network. *Molecular BioSystems*, 2012; 8(7): 1970-1978.
- Yeh S H, Yeh H Y, Soo V W. A network flow approach to predict drug targets from microarray data, disease genes and interactome network-case study on prostate cancer. *Journal of clinical bioinformatics*, 2012; 2(1): 1.
- Bose D, Jayaraj G, Suryawanshi H, et al. The Tuberculosis Drug Streptomycin as a Potential Cancer Therapeutic: Inhibition of miR-21 Function by Directly Targeting Its Precursor. *Angewandte Chemie International Edition*, 2012; 51(4): 1019-1023.

16. Xu S. miRNAs as Novel Therapeutic Targets and Diagnostic Biomarkers for Parkinsons Disease: U.S. Patent Application 14/416,932[P]. 2013; :7-24.
17. Liu Z, Sall A, Yang D. MicroRNA: an emerging therapeutic target and intervention tool. *International journal of molecular sciences*, 2008; 9(6): 978-999.
18. Srinivasan S, Selvan S T, Archunan G, et al. MicroRNAs: the next generation therapeutic targets in human diseases. *Theranostics*. 2013; 3(12):930-42.
19. Lanford R E, Hildebrandt-Eriksen E S, Petri A, et al. Therapeutic silencing of microRNA-122 in primates with chronic hepatitis C virus infection. *Science*, 2010; 327(5962): 198-201.
20. Rhodes L V, Nitschke A M, Segar H C, et al. The histone deacetylase inhibitor trichostatin A alters microRNA expression profiles in apoptosis-resistant breast cancer cells. *Oncology reports*, 2012; 27(1): 10.
21. Bartel D P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 2004; 116(2): 281-297.
22. Esquela-Kerscher A, Slack F J. Oncomirs – microRNAs with a role in cancer. *Nature Reviews Cancer*, 2006; 6(4): 259-269.
23. He Z, Wang Y, Huang G, et al. The lncRNA UCA1 interacts with miR-182 to modulate glioma proliferation and migration by targeting iASPP. *Arch Biochem Biophys*. 2017; 623-624:1-8.
24. Feng J, Xue S, Pang Q, et al. miR-141-3p inhibits fibroblast proliferation and migration by targeting GAB1 in keloids. *Biochem Biophys Res Commun*. 2017; 490(2):302-308.
25. Cui Q, Yu Z, Purisima E O, et al. Principles of microRNA regulation of a human cellular signaling network. *Molecular systems biology*, 2006; 2(1): 46.
26. Krichevsky A M, King K S, Donahue C P, et al. A microRNA array reveals extensive regulation of microRNAs during brain development. *Rna*, 2003; 9(10): 1274-1281.
27. Calin G A, Croce C M. MicroRNA signatures in human cancers. *Nature reviews cancer*, 2006; 6(11): 857-866.
28. Xiangxiang Z, Xuan Z, Quan Z. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Briefings in Bioinformatics*. 2016; 17(2):193-203.
29. Quan Z, Jinjin L, Li S, et al. Similarity computation strategies in the microRNA-disease network: A Survey. *Briefings in Functional Genomics*. 2016; 15(1): 55-64
30. Xiangxiang Z, Yuanlu L, Yuansheng Liu, et al. Prediction and validation of disease genes using HeteSim Scores. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2017; 14(3):687-695.
31. Yuansheng L, Xiangxiang Z, Zengyou H, et al. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2017; 14(4): 905-915.
32. Wu R, He Q, Chen H, et al. MicroRNA-448 promotes multiple sclerosis development through induction of Th17 response through targeting protein tyrosine phosphatase non-receptor type 2 (PTPN2). *Biochem Biophys Res Commun*. 2017; 486(3):759-766.
33. Marco F. Schmidt. Drug target miRNAs: chances and challenges. *Trends in Biotechnology*. 2014; 32(11): 578-585.
34. Zhang S, Chen L, Jung E J, et al. Targeting microRNAs with small molecules: Between Dream and Reality. *Clinical pharmacology and therapeutics*, 2010; 87(6): 754.
35. Cai, J.J. & Petrov, D.A. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol. Evol.* 2010; 2: 393-409.
36. Winter, E.E., Goodstadt, L. & Ponting, C.P. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome Res*. 2004; 14: 54-61.
37. Wei T, Shixiang W, Zhen Y, Andrew E. et al. Tumor Origin Detection with Tissue-Specific miRNA and DNA methylation Markers. *Bioinformatics*. 2017; Doi: 10.1093/bioinformatics/btx622.
38. Weinstein J N, Collisson E A, Mills G B, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 2013; 45(10): 1113-1120.
39. Xiao F, Zuo Z, Cai G, et al. miRecords: an integrated resource for microRNA-target interactions. *Nucleic acids research*, 2009; 37(suppl 1): D105-D110.
40. Chou C H, Chang N W, Shrestha S, et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic acids research*, 2016; 44(D1): D239-D247.
41. Vlachos I S, Paraskevopoulou M D, Karagkouni D, et al. DIANA-TarBase v7. 0: indexing more than half a million experimentally supported miRNA: mRNA interactions. *Nucleic acids research*, 2015; 43(D1): D153-D159.
42. Wishart D S, Knox C, Guo A C, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 2006; 34(suppl 1): D668-D672.
43. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 2000; 28(1): 27-30.
44. Li Y, Qiu C, Tu J, et al. HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic acids research*, 2013; gkt1023.
45. Hamosh A, Scott A F, Amberger J S, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 2005; 33(suppl 1): D514-D517.
46. Greene C S, Krishnan A, Wong A K, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 2015; 47(6): 569-576.
47. Menche J, Sharma A, Kitsak M, Ghiassian S, Vidal M, Loscalzo J, et al. Uncovering disease-disease relationships through the incomplete interactome. *Science* 2015; 347(6224):1257601.
48. Davis A P, Grondin C J, Lennon-Hopkins K, et al. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic acids research*, 2015; 43(D1): D914-D920.
49. Li, H., Liang S. Local network topology in human protein interaction data predicts functional association. *PLoS ONE* 2009; 4(e):6410.
50. Barabasi, A.L., Albert, R. Emergence of scaling in random networks. *Science* 1999; 286: 509-512.
51. King, B.L., Davis, A.P., Rosenstein, M.C., Wiegers, T.C. & Mattingly, C.J. Ranking transitive chemical-disease inferences using local network topology in the comparative toxicogenomics database. *PLoS One* 2012; 7(e):46524.
52. US Food and Drug Administration. FDA approves new treatment option for late-stage breast cancer. Press release, 2013.
53. Kurebayashi J, Kanomata N, Yamashita T, et al. Antitumor and anticancer stem cell activities of eribulin mesylate and antiestrogens in breast cancer cells. *Breast Cancer*, 2016; 23(3): 425-436.
54. Sun Y, Gu J. Study on effect of naringenin in inhibiting migration and invasion of breast cancer cells and its molecular mechanism. *Zhongguo Zhong Yao Za Zhi*. 2015; 40(6): 1144-1150.
55. Van Cutsem E, Sobrero A F, Siena S, et al. Phase III CORRECT trial of regorafenib in metastatic colorectal cancer (mCRC). *Journal of Clinical Oncology* 2012; 30(15\_suppl): 3502-3502.
56. Stalker L, Pemberton J, Moorehead R A. Inhibition of proliferation and migration of luminal and claudin-low breast cancer cells by PDGFR inhibitors. *Cancer cell international*, 2014; 14(1): 89.
57. Su J C, Mar A C, Wu S H, et al. Disrupting VEGF-A paracrine and autocrine loops by targeting SHP-1 suppresses triple negative breast cancer metastasis. *Scientific Reports*, 2016; 6.
58. O'connor O A. Pralatrexate: an emerging new agent with activity in T-cell lymphomas. *Current opinion in oncology*, 2006; 18(6): 591-597.
59. Porter A G, Jänicke R U. Emerging roles of caspase-3 in apoptosis. *Cell death and differentiation*, 1999; 6(2): 99-104.
60. Nielsen V G, Matika R W, Ley M L B, et al. Tissue-type plasminogen activator-induced fibrinolysis is enhanced in patients with breast, lung, pancreas and colon cancer. *Blood Coagulation & Fibrinolysis*, 2014; 25(3): 248-253.
61. Sumiyoshi K, Baba S, Sakaguchi S, et al. Increase in levels of plasminogen activator and type-1 plasminogen activator inhibitor in human breast cancer: possible roles in tumor progression and metastasis. *Thrombosis research*, 1991; 63(1): 59-71.
62. De Luca G, Marino P. Adjunctive benefits from low-molecular-weight heparins as compared to unfractionated heparin among patients with ST-segment elevation myocardial infarction treated with thrombolysis. A meta-analysis of the randomized trials. *American heart journal*. 2007; 154(e6): 1081-1085.
63. Minoru Kanehisa1, Miho Furumichi1 Mao Tanabe, Yoko Sato and Kanea Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucl. Acids Res*. 2017; 45(D1): D353-D361.
64. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*. 2008; 4: 44.
65. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*. 2009; 37:1-13.
66. Radić J, Krušlin B, Šamija M, et al. Connexin 43 Expression in Primary Colorectal Carcinomas in Patients with Stage III and IV Disease. *Anticancer research*, 2016; 36(5): 2189-2196.
67. Yang Y, Zhu J, Zhang N, et al. Impaired gap junctions in human hepatocellular carcinoma limit intrinsic oxaliplatin chemosensitivity: A key role of connexin 26. *International journal of oncology*, 2016; 48(2): 703-713.
68. Yu J, Mai W, Cui Y, et al. Key genes and pathways predicted in papillary thyroid carcinoma based on bioinformatics analysis. *Journal of endocrinological investigation*, 2016; 39(11): 1285-1293.