BMC Bioinformatics

RESEARCH

CrossMark

A network-based pathway-expanding approach for pathway analysis

Qiaosheng Zhang[1,2], Jie Li[1*], Haozhe Xie[1], Hanqing Xue[1] and Yadong Wang[1]

From The 27th International Conference on Genome Informatics
Shanghai, China. 3–5 October 2016

Abstract

Background: Pathway analysis combining multiple types of high-throughput data, such as genomics and proteomics, has become the first choice to gain insights into the pathogenesis of complex diseases. Currently, several pathway analysis methods have been developed to study complex diseases. However, these methods did not take into account the interaction between internal and external genes of the pathway and between pathways. Hence, these approaches still face some challenges. Here, we propose a network-based pathway-expanding approach that takes the topological structures of biological networks into account.

Results: First, two weighted gene-gene interaction networks (tumor and normal) are constructed integrating protein-protein interaction(PPI) information, gene expression data and pathway databases. Then, they are used to identify significant pathways through testing the difference of topological structures of expanded pathways in the two weighted networks. The proposed method is employed to analyze two breast cancer data. As a result, the top 15 pathways identified using the proposed method are supported by biological knowledge from the published literatures and other methods. In addition, the proposed method is also compared with other methods, such as GSEA and SPIA, and estimated using the classification performance of the top 15 expanded pathways.

Conclusions: A novel network-based pathway-expanding approach is proposed to avoid the limitations of existing pathway analysis approaches. Experimental results indicate that the proposed method can accurately and reliably identify significant pathways which are related to the corresponding disease.

Keywords: Pathway analysis, Network-based, Significant pathway, Protein-protein interaction

Background

Complex diseases are likely to be associated with the effects of multiple genes, proteins and biological pathways [1]. Pathway analysis methods that combine multiple types of high-throughput data, such as genomics and proteomics, have become the first choice to gain insights into the pathogenesis of complex diseases. A biological pathway that reduces data involving thousands of altered genes and proteins into a smaller and more interpretable set of altered processes and combines multiple types of high-throughput data plays an important role in understanding the mechanisms of complex diseases, improving clinical treatment, and discovering drug targets and biomarkers [2].

The most commonly employed traditional pathway analysis methods use classical pathway databases (i.e., KEGG [3], MSigDB [4], Reactome [5], BioCyc [6], Meta-Cyc [7], RegulonDB [8], PantherDB [9] and Gene Ontology [10]) to analyse gene expression profile data. These analyses use statistical methods to identify significant pathways in a particular biological process, such as GSEA [11], PAGE [12], GAGE [13] and MeanAbs [14]. A limitation of this class of algorithms is their ignorance of interactions between genes and proteins because neither network topology nor dynamics is taken into account [15]. These limitations are addressed by network-based pathway analyses. Accordingly, several pathway analysis models that

*Correspondence: jieli@hit.edu.cn
[1]School of Computer Science and Technology, Harbin Institute of Technology, West Da-Zhi Street, Harbin, China
Full list of author information is available at the end of the article

reflect the laws of life activities and employ network topology information have been proposed [16], such as SPIA [17], PARADIGM [18], PathOlogist [19], Active Modules [20], AMBIENT [21], GIGA [22] and GANPA [23]. Although the above methods utilize network topology information, they only consider the topological structure of the pathway itself and do not take into account the information of pathway external genes in biological networks; thus, they do not fully mine pathway information. For example, only the pathway internal topology is utilized by the SPIA method, whereas the PathOlogist model only computes the probability of an interaction of pathway internal genes being active when it is consistent with the known regulatory logic of the pathway. Hence, how to take the interactions between internal and external genes of the pathway and between pathways into account in the pathway analysis method is the main problem addressed in this paper.

To that end, we proposed a novel network-based pathway analysis method. First, we integrated protein-protein interaction (PPI) information, gene expression profile data and pathway databases into the pathway analysis and constructed two whole-genome level gene-gene interaction networks. Then, we expanded pathways based on the k-walks algorithm [24, 25] to two small networks in two weighted networks (tumor and normal). Finally, we scored the pathways corresponding to the gene expression profile data based on the correlations of these two small networks to identify significant pathways (see Fig. 1).

Methods

Construction of a weighted gene-gene interaction network
PPI network provides a valuable framework to elucidate the functional organization of the proteome. However, existing PPI networks cannot accurately describe the interactions between proteins in specific conditions and have different degrees of false positive and false negative results because most large-scale PPI networks are obtained in different experimental conditions, predicted/extracted using different algorithms [26, 27]. Additionally, the interaction or the intensity between proteins varies in different cells or tissues.

The gene co-expression network (GCN) is an undirected graph where each node corresponds to a gene and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them [28]. Using gene expression profiles obtained from a number of genes for several samples or experimental conditions, a gene co-expression network can be constructed by looking for pairs of genes that show a similar expression pattern across samples. In this study, the weight of each pair of genes is calculated by the Pearson's correlation coefficient. Pearson's correlation coefficient was selected as the co-expression measurement because it was the most popular

co-expression measurement used in the construction of gene co-expression networks. The absolute values correspond to an interaction mechanism where the intensity of one gene is related to its co-expressed gene. However, a gene co-expression network does not guarantee the existence of a real interaction between the corresponding proteins; instead, it only suggests that there may be an interaction between the proteins.

To accurately describe the change in gene interactions for several samples or experimental conditions, here we constructed two weighted gene-gene interaction networks (tumor and normal) with PPI and GCN (see Fig. 2).

Pathway-based extension of the sub-network
The gene-gene interaction of pathway is different in different tissues or samples. These differences may be caused by changes in the interactions between internal genes of the pathway or between pathway and neighbor genes. To assess the significance of the pathway in different phenotypic data, we expanded the pathway based on the k-walk algorithm [25] by considering all of the above factors in two weighted gene-gene interaction networks separately. The pathway-based extension of the sub-network was constructed as follows:

Let $G = (V, E)$ comprise a set V of genes and a set E of edges denote the weighted gene-gene interaction network with $E \subseteq V \times V$. Let $n = |V|$ reflect the number of genes. Symmetric matrix A represents the weighted $n \times n$ adjacency matrix of G, where a_{ij} denotes the weight of the edge connecting gene i to gene j. Let $d_i (i \in 1 \cdots n)$ represent the weighted degree of each gene node i where $d_i = \sum_{j=1}^{n} a_{ij}$. Then, a_{ij} is calculated by Pearson's correlation coefficient as:

$$a_{ij} = \begin{cases} |cor(x_i, x_j)|^\beta & x_i, x_j \text{ is expression} \\ & \text{data of gene } V_i, V_j \\ 0 & \text{otherwise} \end{cases} \quad (1)$$
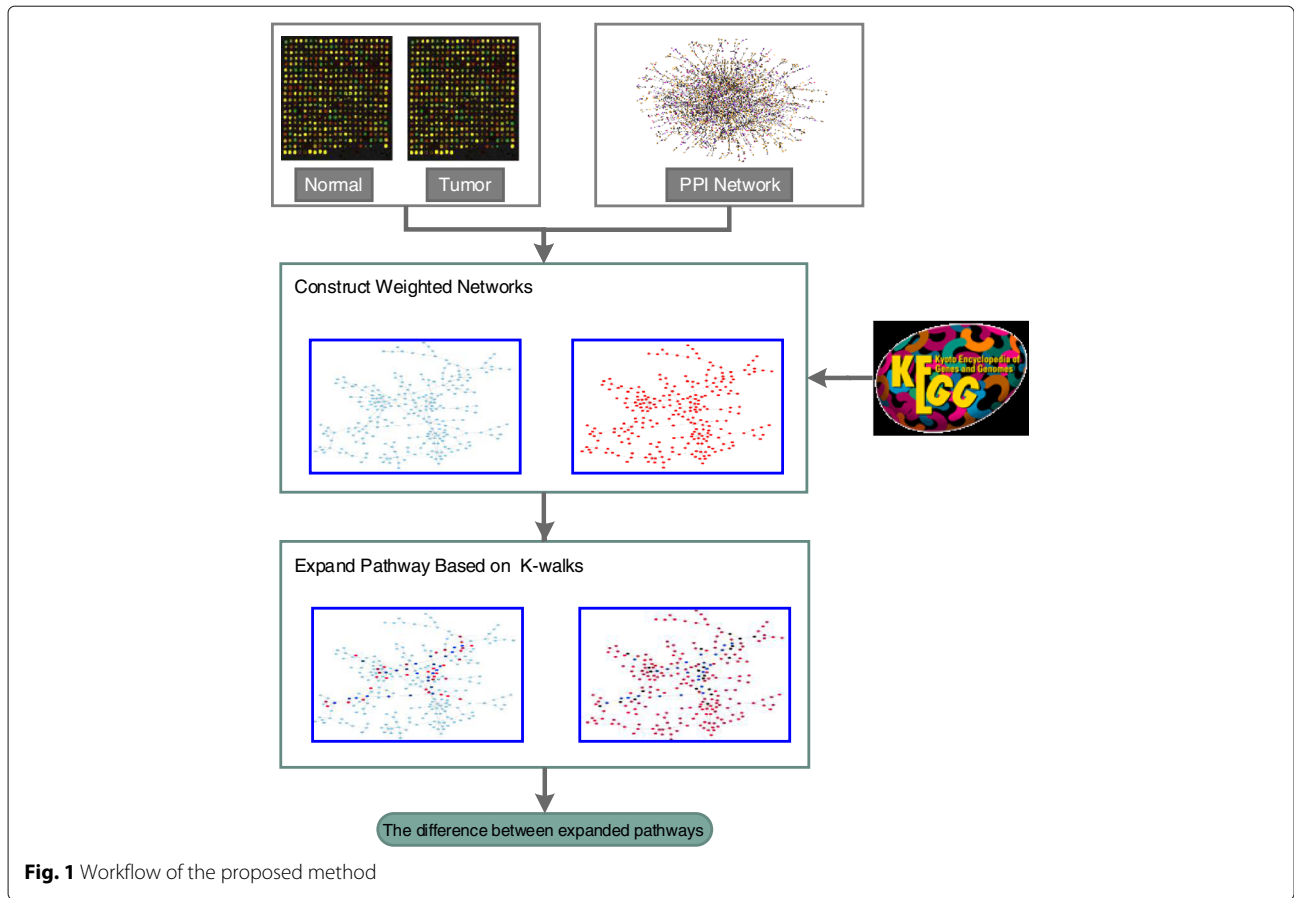
where $\beta = 1$.

Given a gene set S ($|S| \geq 2$) of a pathway belonging to a subset of G, we formally define an edge relevance function $ER{:}E \to \mathbb{R}^+$ that maps any edge to its relevance. The extended process of a gene set of a pathway simulates random walks on a graph by the Markov Chain model. The possibility of transiting from gene i to gene j is calculated as:

$$P_{ij} = \frac{a_{ij}}{d_i} \quad (2)$$

Here, a gene set S is a set of absorbing states of the Markov chain. If the random walk starts from gene x, the modified transition will be:

$$^x P_{ij} = \begin{cases} 1 & i \in S \backslash \{x\} \text{ and } i = j \\ 0 & i \in S \backslash \{x\} \text{ and } i \neq j \\ P_{ij} & \text{otherwise} \end{cases} \quad (3)$$

Fig. 1 Workflow of the proposed method

Then, the transition matrix is described as follow:

$$^xP = \begin{pmatrix} ^xQ & ^xR \\ 0 & I \end{pmatrix} \tag{4}$$

where ^xQ is a matrix that denotes transient states, ^xR is a matrix that denotes the transition probability from transient states to absorbing states, and I is the identity matrix. After k steps, the transition matrix becomes $(^xQ)^k$.

Given that the walk started in state x, the joint probability of visiting the edge $E(i,j)$ between step k and $k+1$ is calculated as follow:

$$P\left[X_k = i, X_{k+1} = j, L|X_0 = x\right] =$$
$$\begin{cases} f(i,j,k,x,r,L) & j \text{ is a transient state} \\ \left[(^xQ)^{L-1}\right]_{xi}\left[(^xR)\right]_{ij} & j \text{ is an absorbing state} \end{cases} \tag{5}$$

where L is a total walk length, $f(i,j,k,x,r,L) = \sum\limits_{r\in S\backslash\{x\}}$ $\left[(^xQ)^k\right]_{xi}\left[(^xQ)\right]_{ij}\left[(^xQ)^{L-k-2}(^xR)\right]_{jr}$, $\left[(^xQ)^k\right]_{xi}$ is the probability of transiting from x to i in k steps.

The probability of a walk of length L starting in x is calculated as follow:

$$P[L|X_0 = x] = \sum\limits_{r\in S\backslash\{x\}}\left[(^xQ)^{L-1}(^xR)\right]_{xr} \tag{6}$$

The $e(x,i,j)$ is defined as the number of times a random walk starts in x using the transition from i to j. Given that the walk length is L, the conditional expectations of $e(x,i,j)$ is given by:

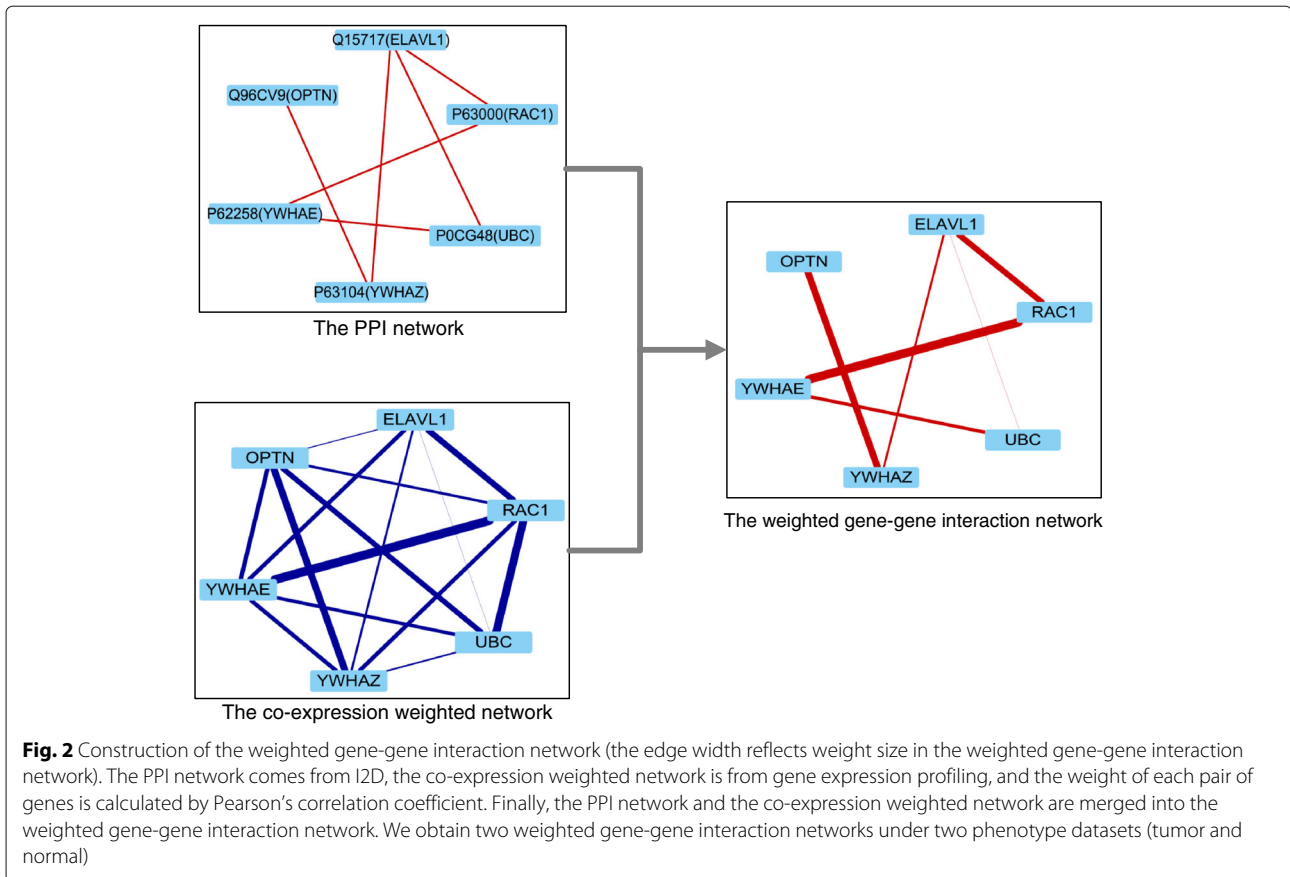$$E[e(x,i,j)|L] = \sum\limits_{k=0}^{L-1}\frac{P[X_k = i, X_{k+1} = j, L|X_0 = x]}{P[L|X_0 = x]} \tag{7}$$

Let L_{max} denote a maximal walk length. Then:

$$E[e(x,j,i)|L \leq L_{max}] = \sum\limits_{L=1}^{L_{max}} E[e(x,i,j)|L] \tag{8}$$

Finally, the edge relevance ER is given by:

$$ER(i,j) =$$

$$\sum\limits_{r\in S} l_x \big| E\left[e(x,i,j)|L \leq L_{max}\right] - E\left[e(x,j,i)|L \leq L_{max}\right]\big| \,\forall (i,j) \in E \tag{9}$$

where the vector l_x represents an initial probability distribution. Here, the maximal relevance score that can lead to a connected subgraph is chose as the threshold θ. Finally, a subnetwork is obtained by keeping only edges with rele-

Fig. 2 Construction of the weighted gene-gene interaction network (the edge width reflects weight size in the weighted gene-gene interaction network). The PPI network comes from I2D, the co-expression weighted network is from gene expression profiling, and the weight of each pair of genes is calculated by Pearson's correlation coefficient. Finally, the PPI network and the co-expression weighted network are merged into the weighted gene-gene interaction network. We obtain two weighted gene-gene interaction networks under two phenotype datasets (tumor and normal)

vance scores $ER(i, j)$ above a threshold value θ (see Fig. 3). In the paper, we set L_{max} to 50 by default.

Identification of significant pathways

For a given pathway i, the pathway was expanded in two weighted gene-gene interaction networks (Tumor_Net and Normal_Net)for two phenotypic datasets separately. The union genes of two expanded pathway play a role in performing a similar function in normal or tumor tissues. Moreover, genes in expanded pathway from tumor and normal tissues are almost different. The union genes construct sub-networks in two weighted networks. These sub-networks weights' differences can describe the change of pathways between different phenotypes. Accordingly, we calculate the difference between the two pathway-based sub-networks reflects the change of the given pathway for the two phenotypic datasets through the union of the two sub-networks.

Let Union_Pathway[i] denote the union of two sub-networks (T_Ex_Pathway[i] and N_Ex_Pathway[i]) that are the expansion of pathway i in two weighted gene-gene interaction networks (Tumor_Net and Normal_Net). Then, we mapped Union_Pathway[i] into the two weighted gene-gene interaction networks (T_subnet[i] and N_subnet[i]) and obtained two edge weight vectors

T_w[i] and N_w[i]. Pearson's correlation coefficient was calculated as:

$$Corr_i(T\_w[i], N\_w[i]) =$$

$$\frac{\sum_{k=1}^{n}(T\_w[i]_k - \overline{T\_w[i]})(N\_w[i]_k - \overline{N\_w[i]})}{\sqrt{\sum_{k=1}^{n}(T\_w[i]_k - \overline{T\_w[i]})^2 \sum_{k=1}^{n}(N\_w[i]_k - \overline{N\_w[i]})^2}} \quad (10)$$

where n is the dimension of the vector and $\overline{T\_w[i]} = \frac{1}{n}\sum_{k=1}^{n} T\_w[i]_k$, $\overline{N\_w[i]} = \frac{1}{n}\sum_{k=1}^{n} N\_w[i]_k$. Finally, we calculated Score[i], which depicts the difference in pathway i for two phenotypic datasets as follow:
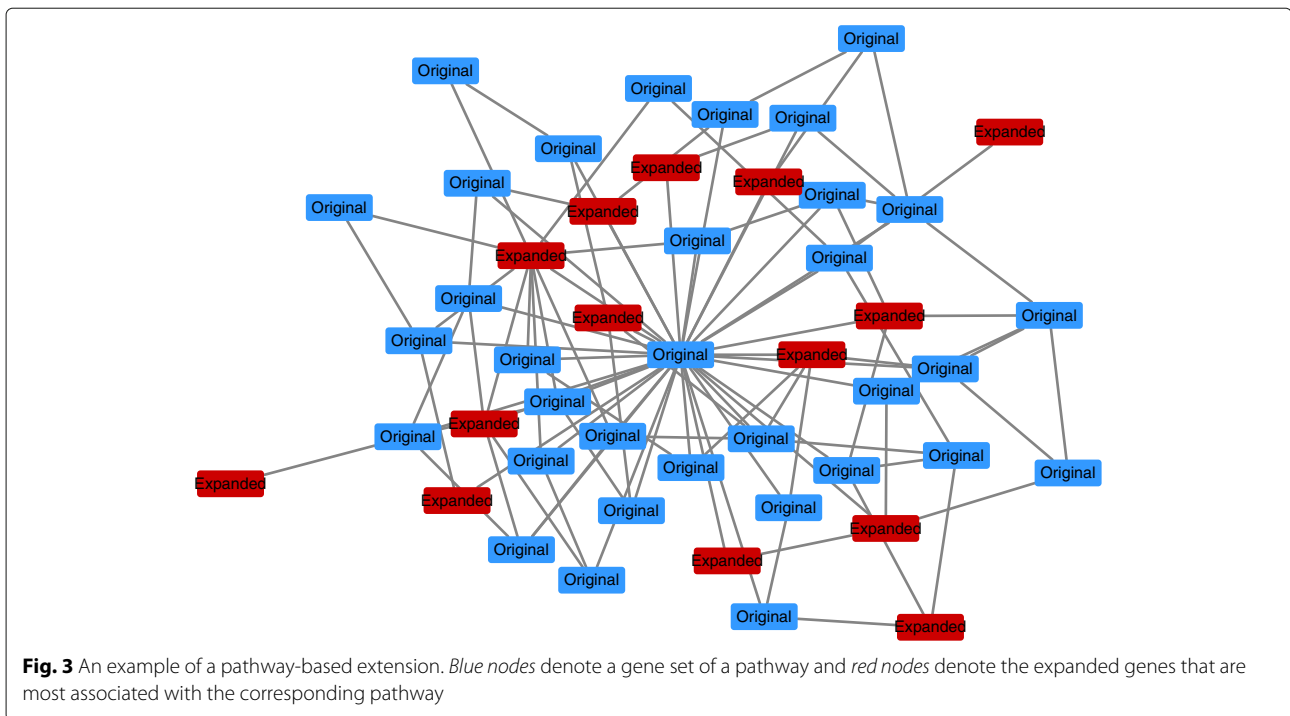
$$Score[i] = 1 - |Corr_i(T\_w[i], N\_w[i])| \quad (11)$$

Here, Score[i] is a measure depicting the relevance degree between pathway i and the corresponding disease (for the pseudo-code see Algorithm 1).

Results and discussion

Data

The breast invasive carcinoma (BRCA) dataset was downloaded from the TCGA (The Cancer Genome Atlas) website (http://cancergenome.nih.gov/). The BRCA dataset consists of 590 samples obtained from comparing 529 breast cancer samples with 61 normal samples using the Agilent platform. The second dataset

Fig. 3 An example of a pathway-based extension. *Blue nodes* denote a gene set of a pathway and *red nodes* denote the expanded genes that are most associated with the corresponding pathway

was available via the Gene Expression Omnibus (ID= GSE25066). This dataset compared 99 pathologic complete response (PCR) samples and 389 residual disease (RD) samples [29] (http://www.ncbi.nlm.nih.gov/ geo/query/acc.cgi?acc=GSE25066). PPI network (version 2.9) was obtained from the Interologous Interaction Database (I2D) website [26] (http://ophid.utoronto.ca/ ophidv2.204/downloads.jsp). PPI was mapped into the gene-gene interaction (GGI) data through the UniProt website (www.uniprot.org/). Finally, 234,524 unique gene pairs were selected for BRCA, and 204,772 unique gene pairs were selected for GSE25066 by data pretreatment. The KEGG pathways were downloaded from the ConsensusPathDB website (http://consensuspathdb.org/). We selected 280 pathways related to humans by screening; only genes of pathways belonging to the BRCA gene set were used in the downstream analysis. The breast cancer gene set was downloaded from website(http://mlg.hit.edu. cn/SIDD/).

To identify the significance of the given pathway, first, we dealt with the PPI data. The PPI network was mapped into the gene-gene interaction (GGI) network in which the weight of each pair of genes was calculated using high-throughput gene expression profiling data. Finally, we obtained two weighted gene-gene interaction networks for the two phenotypic datasets. The weighted gene-gene interaction network has 15,129 vertices and 234,524 edges for BRCA.

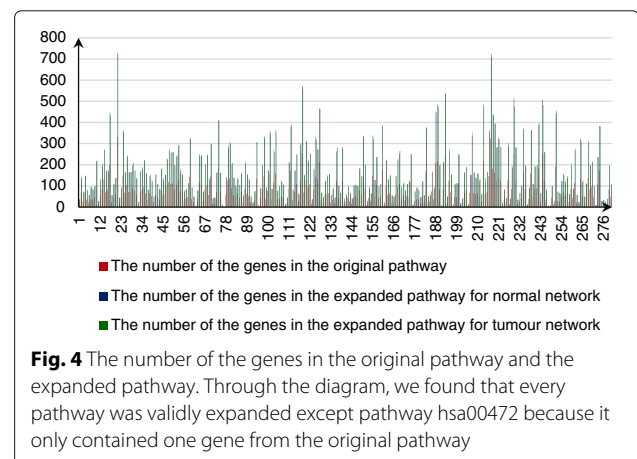Based on the above algorithm, we expanded a gene set of the given pathway based on the k-walks algorithm into two sub-networks in two weighted networks (tumor and normal). Then, we compared the number of the genes in the original pathway and the expanded pathway (see Fig. 4). Finally, the union of the two sub-networks served as the ultimate expansion of the given pathway.

Next, we ran the proposed approach using the BRCA dataset.

To provide a more comprehensive understanding of the proposed method, we discuss the method from the following aspects separately.

The pathway score
Based on the background mentioned above, each pathway score depicts the degree of relevance between the given pathway and the corresponding disease. All scores were

Fig. 4 The number of the genes in the original pathway and the expanded pathway. Through the diagram, we found that every pathway was validly expanded except pathway hsa00472 because it only contained one gene from the original pathway

---

Algorithm 1: Compute Score

Input : *Pathways, Tumor_Net, Normal_Net*
Output: *Score*
for *i = 1 to The total number of pathways* **do**

> 1.Pathway[i] maps into Tumor_Net;
> 2.T_Ex_Pathway[i]= Expand Pathway[i] based on Limited K-walks algorithm in Tumor_Net;
> 3.Pathway[i] maps into Normal_Net;
> 4.N_Ex_Pathway[i]=Expand Pathway[i] based on Limited K-walks algorithm in Normal_Net;
> 5.Union_Pathway[i]=T_Ex_Pathway[i] ∪ N_Ex_Pathway[i] ;
> 6.T_subnet[i]=Union_Pathway[i] maps into Tumor_Net;
> 7.T_w[i]= The weight of the T_subnet[i] in Tumor_Net;
> 8.N_subnet[i]= Union_Pathway[i] maps into Normal_Net;
> 9.N_w[i]= The weight of the N_subnet[i] in Normal_Net;
> 10.$Corr_i$= The correlation coefficient of T_w[i] and N_w[i];
> 11.Score[i]=$1 - |Corr_i|$;

end

---

calculated using Algorithm 1 (Additional file 1: Table S1). The top 15 pathways were tabulated based on their scores (see Table 1).

Top pathways identified by the proposed method should be significantly associated with the breast cancer risk. To test the idea, we compared the intersections of the breast cancer gene set and pathway gene sets before and after expansion (see Fig. 5). The breast cancer gene set comes from SIDD which integrates 22 disease gene knowledge sources. We found that more genes associated with breast cancer were expanded to the original pathway gene set through pathway expansion. This result demonstrates that the proposed method can expand genes associated with the corresponding disease.

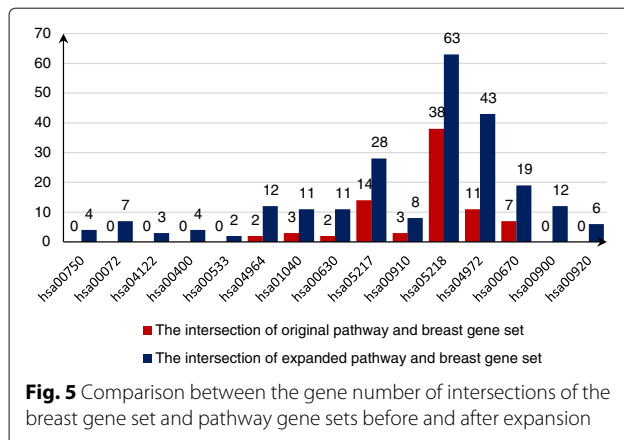**Analysis of the top 15 BRCA pathways**

In order to prove that the pathways identified by the proposed method are associated with the breast cancer risk, we need to look for the supports of biological knowledge and other methods. Table 1 shows that top 15 pathways identified from BRCA by the proposed method are significantly associated with the breast cancer risk through reference. Here we give the supports of biological knowledge and other methods for the top 15 BRCA pathways.

The number 1 ranked significant pathway identified by our method was vitamin B6 metabolism (hsa00750). Growing evidence suggests that the lack of several trace elements, such as vitamin B6 and folate, can induce DNA damage (e.g., single or double-stranded breaks or fusion), eventually leading to tumors, cancers and a variety of degenerative diseases [30]. There is a significant negative correlation between the plasma B6 level and different types of cancer. Vitamin B6 can reduce the homocysteine and pyridoxal phosphate levels, which have potential biological effects on tumors. Vitamin B6 deficiency leads to lower serine hydroxymethyltransferase activity, lower generation of 5,10-methylenetetrahydrofolate and the

Table 1 Top 15 pathways identified from BRCA

Rank	Entry	Name	Score	SPIA	GSEA	Proof
1	hsa00750	Vitamin B6 metabolism	0.997735	No	No	[30–33]
2	hsa00072	Synthesis and degradation of ketone bodies	0.940425	No	Yes	[34–37]
3	hsa04122	Sulphur relay system	0.855753	No	No	[38, 39]
4	hsa00400	Phenylalanine,tyrosine and tryptophan biosynthesis	0.850563	No	No	[40]
5	hsa00533	Glycosaminoglycan biosynthesis	0.836469	No	No	[41–43]
6	hsa04964	Proximal tubule bicarbonate reclamation	0.803311	No	Yes	[46]
7	hsa01040	Biosynthesis of unsaturated fatty acids	0.799334	No	No	[47]
8	hsa00630	Glyoxylate and dicarboxylate metabolism	0.785954	No	Yes	[48]
9	hsa05217	Basal cell carcinoma	0.779876	No	No	[49, 50]
10	hsa00910	Nitrogen metabolism	0.77962	No	Yes	[51, 52]
11	hsa05218	Melanoma	0.758975	Yes	Yes	[53]
12	hsa04972	Pancreatic secretion	0.754263	No	Yes	[54, 55]
13	hsa00670	One carbon pool by folate	0.7452	No	No	[56]
14	hsa00900	Terpenoid backbone biosynthesis	0.736641	No	No	[57, 58]
15	hsa00920	Sulphur metabolism	0.733627	No	No	[59]

Note: Yes if the pathway was also ranked in the SPIA or GSEA top 15; No if otherwise

Fig. 5 Comparison between the gene number of intersections of the breast gene set and pathway gene sets before and after expansion

generation of a dUMP instead of a dTMP mismatch to DNA, which is more likely to lead to a chromosome chain break and /or impair DNA excision repair. The reduced generation of 5,10-methylenetetrahydrofolate may lead to DNA hypomethylation. Abnormal methylation of DNA has been found in different tumor types [31, 32]. Vitamin B6 deficiency can increase the sensitivity of the steroid hormone,which may lead to breast cancer or colon cancer [33]. These findings demonstrate that the proper intake of vitamin B6 can reduce the risk of breast cancer; therefore, this pathway is significantly associated with the breast cancer risk.

The number 2 ranked significant pathway identified by the proposed method was Synthesis and degradation of ketone bodies (hsa00072). Ketone bodies (i.e., 3-hydroxy-butyrate and/or butanediol) are sufficient to drive mitochondrial biogenesis in human breast cancer cells [34, 35]. Carcinoma-associated fibroblasts (CAF) produce "mitochondrial fuels", including lactic acid, ketones, fatty acids, and glutamine, that provide a "eutrophication" microenvironment for tumor cells and promote tumor cell proliferation when metabolized; these fuels are the major cellular components of the breast cancer stroma [36]. It was reported that CAF reduced the apoptosis of human breast cancer MCF7 cells induced by tamoxifen and fulvestrant by 4.4-fold and 2.5-fold, respectively [37]. Lactic acid and ketones are sufficient to induce tamoxifen resistance in breast cancer MCF7 cells. Metformin and arsenic trioxide can overcome CAF-induced drug resistance in MCF7 cells. These findings indicate that this pathway is also significantly associated with the breast cancer risk.

The proposed method ranked the Sulphur relay system (hsa04122) in 3rd place. Sulphur enables the transport of oxygen across cell membranes. Oxygen is necessary for healthy cellular regeneration in mammals. Therefore, sulphur deficiencies may promote sickness and disease. Sulphur is commonly used as an herbal medicine to treat inflammation and cancer and organic sulphur has been

studied in several types of cancers and found to have remarkable anti-cancer benefits. Methylsulfonylmethane (MSM) is an organic sulphur-containing natural compound without any toxicity. It was found that MSM substantially decreased the viability of human breast cancer cells in a dose-dependent manner and recommended the use of MSM as a trial drug for the treatment of all types of breast cancers [38]. Leimkühler et al. pointed out that sulphur not only prevented but also helped reverse cancer [39]. Hence, the sulphur relay system is significantly associated with the breast cancer risk to some extent.

Phenylalanine, tyrosine and tryptophan biosynthesis (hsa00400) was ranked 4th in the list of proposed methods. ENO1 in phenylalanine tyrosine and tryptophan biosynthesis was significantly overexpressed in HER-2/neu positive breast tumors [40]. This finding indicates that this pathway is associated with breast cancer to some extent; however, the clear relationship between this pathway and breast cancer re-quires further verification.

The 5th ranked pathway was Glycosaminoglycan biosynthesis (hsa00533). Abnormal glycosaminoglycan (GAG) concentrations have been reported for various types of tumors, suggesting that they may play a role in neoplasia. Recently, cell biology studies revealed that glycosaminoglycans were among the key macromolecules that affected cell properties and functions by acting directly on cell receptors or via interactions with growth factors. The interactions of GAGs with growth factors, cytokines and growth factor receptors have been implicated in cancer growth and progression. GAGs are involved in signalling cascades that regulate the angiogenesis, invasion and metastasis of malignant cells. Investigations of the fine structures and specific biological roles of GAGs has led to novel therapeutic approaches [41–43]. The above references denote that glycosaminoglycan biosynthesis and breast cancer have a certain degree of correlation.

The top 6–15 pathways are also associated with human breast cancer (see Table 1). Based on the Table 1, one can argue that the proposed method is very efficient in identifying significant pathways of the corresponding complex disease.

Classification performance using the original genes and expanded genes of the pathway

To estimate the classification performance of the top 15 expanded pathways, we firstly prepared our data set consisted of 60 normal and 60 tumor samples randomly derived from the BRCA dataset. The original genes and expanded genes of the pathway were selected classification features and SVM is employed to classify the selected samples. Next, a 10-fold cross validation was used to train and test SVM. The above experiment was repeated 100 times and the average accuracy of SVM is shown in Fig. 6.
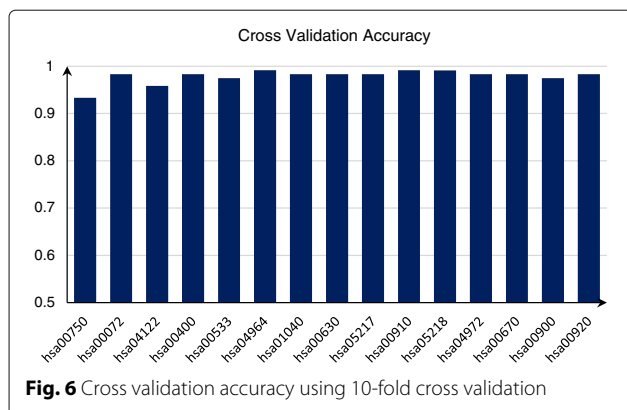
Fig. 6 Cross validation accuracy using 10-fold cross validation

In results, the lowest accuracy was 0.9333 and the highest accuracy was 0.9917. The experimental results suggested that the union of the original genes and expanded genes of the pathway had a good classification ability and that the top 15 pathways were significantly different between the two phenotypic data sets.

Analysis using alternative methods

To assess the validity of the proposed approach, we analysed the same data using GSEA and SPIA. The GSEA approach searches for gene sets that are enriched at the top or bottom of the ranked list of all genes. This method is a typical representative of the gene set enrichment analysis methods. The SPIA method scores a gene product as highly impactful if it points to other impactful gene products in the network diagram. This method is a representative of the network-based pathway analysis approaches. Therefore, we compared our method with GSEA and SPIA. It was interesting to examine pathways ranked at the top by our method but not by GSEA and SPIA, which reflected the validity of our method.

In GSEA, the analysis performed 1000 permutations with an FDR cutoff of 25%. Then, 115 pathways were identified (Additional file 2: Table S2) of which 6 were found among the top 15 pathways identified using the proposed method (see Table 1).

In SPIA, a significant threshold of 5% was used on the FDR corrected P-values to infer pathway significance. Then, 3 pathways were identified (Additional file 3: Table S3) of which one was identified by the proposed method (see Table 1). The SPIA did not identify any of the top 5 pathways identified using the proposed method.

Validation of the alternative data set

To test the effectiveness of the proposed method, we ran the proposed approach on GSE25066. The data were obtained from response and survival following Taxane-Anthracycline chemotherapy for newly diagnosed invasive breast cancer. Anthracyclines and taxanes are the two most active classes of cytotoxic agents for early and advanced stage breast cancer and thus are commonly used as a component of either adjuvant or neoadjuvant therapy and/or in patients with metastatic breast cancer (MBC) [44]. Finally, we also obtained two weighted gene-gene interaction networks for the two phenotypic datasets. The weighted gene-gene interaction network has 10,856 vertices and 204,772 edges for GSE25066. Our intention was to identify significant pathways for breast cancer patients before and after Taxane-Anthracycline use and to evaluate the pharmacological mechanism of Taxane-Anthracycline. Among the top 15 pathways identified using the proposed method, which were significant pathways for Taxane-Anthracycline except for collecting duct acid secretion pathway (hsa04966), the GSEA and SPIA did not identify any. The relationship between collecting duct acid secretion pathway and Taxane-Anthracycline and/or breast cancer requires further verification. The results (Additional file 4: Table S4) showed that our approach discovered significant pathways for Taxane-Anthracycline. The top 15 pathways are shown in Table 2.

The significantly impacted pathways identified by the proposed method in the corresponding conditions were mostly consistent with the known biological processes. Accordingly, the novel proposed method is of methodological and biological significance for future research.

Conclusions

Pathway analysis not only reduces data involving thousands of altered genes and proteins to a smaller and more interpretable set of altered processes but can also combine multiple types of high-throughput data. The analysis results play an important role in elucidating the mechanisms of complex diseases, improving clinical treatment, and discovering drug targets and biomarkers. Therefore, pathway-based analysis of complex diseases has become a research hotspot. To date, these methods have entered the third stage [45]: 1) Pathway-based gene set enrichment analysis; 2) Pathway-based functional class clustering and scoring approaches; and 3) Network-based pathway approaches.

Unlike existing pathway analysis approaches that do not take into account the interaction between internal and external genes of the pathway and between pathways, we propose a novel approach that addresses the above-mentioned limitations by expanding a pathway based on the k-walk algorithm to two small networks in two weighted networks (tumor and normal). Finally, our approach effectively identified significant pathways that corresponded to a complex disease through a series of verification steps. It is undeniable that the pathways identified by GSEA and SPIA but not by our method are mostly significantly associated with the breast cancer

Table 2 Top 15 pathways identified from GSE25066

Rank	Entry	Name	Score	SPIA	GSEA	Proof
1	hsa05033	Nicotine addiction	0.430656	No	No	[60]
2	hsa05217	Basal cell carcinoma	0.402534	No	No	[49, 50]
3	hsa04740	Olfactory transduction	0.398952	No	No	[61, 62]
4	hsa04742	Taste transduction	0.393069	No	No	[63, 64]
5	hsa04340	Hedgehog signaling pathway	0.376035	No	No	[65, 66]
6	hsa04727	GABAergic synapse	0.362031	No	No	[67]
7	hsa04713	Circadian entrainment	0.356687	No	No	[68, 69]
8	hsa00053	Ascorbate and aldarate metabolism	0.35363	No	No	[70]
9	hsa04723	Retrograde endocannabinoid signaling	0.343314	No	No	[71–73]
10	hsa04978	Mineral absorption	0.342461	No	No	[74]
11	hsa04961	Endocrine and other factor-regulated calcium reabsorption	0.341614	No	No	[75–77]
12	hsa00140	Steroid hormone biosynthesis	0.337664	No	No	[78, 79]
13	hsa04966	Collecting duct acid secretion	0.336179	No	No	Not Found
14	hsa04330	Notch signaling pathway	0.333964	No	No	[80, 81]
15	hsa04614	Renin-angiotensin system	0.332994	No	No	[82, 83]

Note: Yes if the pathway was also ranked in SPIA or GSEA top 15; No if otherwise

risk. Based on the above analysis, our method combined with GSEA may produce better results. Hence, we will combine our method with GSEA in future studies. This study provides a new research direction for the pathway-based analysis of complex diseases. We will employ more datasets to assess the validity of our approach in future research.

Additional files

Additional file 1: Table S1. The scores of pathways from BRCA. (PDF 234 kb)

Additional file 2: Table S2. The results of GSEA from BRCA. (PDF 321 kb)

Additional file 3: Table S3. The results of SPIA from BRCA (Top 20). (PDF 110 kb)

Additional file 4: Table S4. The scores of pathways from GSE25066. (PDF 272 kb)

Availability of data and material
The data supporting the findings of this work are contained within the manuscript.

Author's contributions
JL designed the method, QZ and JL performed simulations, analyses and wrote the manuscript. HX, HX and YW participated in the preparation of the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]School of Computer Science and Technology, Harbin Institute of Technology, West Da-Zhi Street, Harbin, China. [2]College of Science, Heilongjiang Bayi Agricultural University, Xinfeng Road, Daqing, China.

Published: 23 December 2016

References
1. Jin W, Qin P, Lou H, Jin L, Xu S. A systematic characterization of genes underlying both complex and mendelian diseases. Human Mol Genet. 2012;21(7):1611–24.
2. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Leiserson MD, Niu B, McLellan MD, Uzunangelov V, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929–44.
3. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. Kegg for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2010;38(suppl 1):355–60.
4. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (msigdb) 3.0. Bioinformatics. 2011;27(12):1739–40.
5. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath GR, Wu GR, Matthews L, et al. Reactome: a

knowledgebase of biological pathways. Nucleic Acids Res. 2005;33(suppl 1):D428–32.

6. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M, Paley S, Rhee SY, Shearer AG, Tissier C, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res. 2008;36(suppl 1):623–31.

7. Karp PD, Riley M, Paley SM, Pellegrini-Toole A. The metacyc database. Nucleic Acids Res. 2002;30(1):59–61.

8. Huerta AM, Salgado H, Thieffry D, Collado-Vides J. Regulondb: a database on transcriptional regulation in escherichia coli. Nucleic Acids Res. 1998;26(1):55–9.

9. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. Panther: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129–41.

10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.

11. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Nat Acad Sci. 2005;102(43):15545–50.

12. Kim SY, Volsky DJ. Page: parametric analysis of gene set enrichment. BMC Bioinformatics. 2005;6(1):1.

13. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. Gage: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics. 2009;10(1):1.

14. Efron B, Tibshirani R, et al. On testing the significance of sets of genes. Ann Appl Stat. 2007;1(1):107–29.

15. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):1002375.

16. Creixell P, Reimand J, Haider S, Wu G, Shibata T, Vazquez M, Mustonen V, Gonzalez-Perez A, Pearson J, Sander C, et al. Pathway and network analysis of cancer genomes. Nat Methods. 2015;12(7):615.

17. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-S, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis. Bioinformatics. 2009;25(1):75–82.

18. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D, Stuart JM. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. Bioinformatics. 2010;26(12):237–45.

19. Greenblum SI, Efroni S, Schaefer CF, Buetow KH. The pathologist: an automated tool for pathway-centric analysis. BMC Bioinformatics. 2011;12(1):1.

20. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002;18(suppl 1):233–40.

21. Bryant WA, Sternberg MJ, Pinney JW. Ambient: active modules for bipartite networks-using high-throughput transcriptomic data to dissect metabolic response. BMC Syst Biol. 2013;7(1):1.

22. Breitling R, Amtmann A, Herzyk P. Graph-based iterative group analysis enhances microarray interpretation. BMC Bioinformatics. 2004;5(1):1.

23. Fang Z, Tian W, Ji H. A network-based gene-weighting approach for pathway analysis. Cell Res. 2012;22(3):565–80.

24. Zheng S, Zhao Z. Genrev: exploring functional relevance of genes in molecular networks. Genomics. 2012;99(3):183–8.

25. Dupont P, Callut J, Dooms G, Monette J-N, Deville Y, Sainte BP. Relevant subgraph extraction from random walks in a graph. Res Report RR. 2006.

26. Brown KR, Jurisica I. Unequal evolutionary conservation of human protein interactions in interologous networks. Genome Biol. 2007;8(5):1.

27. Hakes L, Pinney JW, Robertson DL, Lovell SC. Protein-protein interaction networks and biology-what's the connection? Nat Biotechnol. 2008;26(1):69–72.

28. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302(5643):249–55.

29. Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, et al. A genomic predictor of response

and survival following taxane-anthracycline chemotherapy for invasive breast cancer. Jama. 2011;305(18):1873–81.

30. Wang TC, Song YS, Wang H, Zhang J, Yu SF, Gu YE, Chen T, Wang Y, Shen HQ, Jia G. Oxidative dna damage and global dna hypomethylation are related to folate deficiency in chromate manufacturing workers. J Hazardous Mater. 2012;213:440–6.

31. Wu XY, Ni J, Xu WJ, Zhou T, Wang X. Interactions between mthfr c677t-a1298c variants and folic acid deficiency affect breast cancer risk in a chinese population. Asian Pac J Cancer Prevention. 2012;13(5):2199–206.

32. Ito K, Nakahara I, Sakamoto Y. Studies on vitamin b6 metabolism of cancer cells and tumor-bearing rat liver. ii. uptake of pyridoxine derivatives by tumor cells and the liver of tumor-bearing rats. Gann= Gan. 1964;55:379–85.

33. Sujol G, Docquier A, Boulahtouf A, Castet-Nicolas A, Cavaillès V. Vitamin b6 and cancer: from clinical data to molecularly mechanisms. Bulletin du cancer. 2011;98(10):1201–8.

34. Martinez-Outschoorn UE, Lin Z, Whitaker-Menezes D, Howell A, Sotgia F, Lisanti MP. Ketone body utilization drives tumor growth and metastasis. Cell Cycle. 2012;11(21):3964–71.

35. Martinez-Outschoorn UE, Lin Z, Whitaker-Menezes D, Howell A, Lisanti MP, Sotgia F. Ketone bodies and two-compartment tumor metabolism: stromal ketone production fuels mitochondrial biogenesis in epithelial cancer cells. Cell Cycle. 2012;11(21):3956–63.

36. Martinez-Outschoorn UE, Lisanti MP, Sotgia F. Catabolic cancer-associated fibroblasts transfer energy and biomass to anabolic cancer cells, fueling tumor growth. In: Seminars in Cancer Biology. Elsevier; 2014. p. 47–60.

37. Martinez-Outschoorn UE, Goldberg AF, Lin Z, Ko YH, Flomenberg N, Wang C, Pavlides S, Pestell RG, Howell A, Sotgia F, et al. Anti-estrogen resistance in breast cancer is induced by the tumor microenvironment and can be overcome by inhibiting mitochondrial function in epithelial cancer cells. Cancer Biol Therapy. 2011;12(10):924–38.

38. Lim EJ, Hong DY, Park JH, Joung YH, Darvin P, Kim SY, Na YM, Hwang TS, Ye SK, Moon ES, et al. Methylsulfonylmethane suppresses breast cancer growth by down-regulating stat3 and stat5b pathways. PloS One. 2012;7(4):33361.

39. Mendel RR, Leimkühler S. The biosynthesis of the molybdenum cofactors. JBIC J Biol Inorg Chem. 2015;20(2):337–47.

40. Zhang D, Tai LK, Wong LL, Chiu LL, Sethi SK, Koay ES. Proteomic study reveals that proteins involved in metabolic and detoxification pathways are highly expressed in her-2/neu-positive breast cancer. Mol Cell Proteomics. 2005;4(11):1686–96.

41. De Ceuninck F, Gaufillier S, Bonnaud A, Sabatini M, Lesur C, Pastoureau P. Ykl-40 (cartilage gp-39) induces proliferative events in cultured chondrocytes and synoviocytes and increases glycosaminoglycan synthesis in chondrocytes. Biochem Biophys Res Commun. 2001;285(4):926–31.

42. Olsen EB, Trier K, Eldov K, Ammitzbøll T. Glycosaminoglycans in human breast cancer. Acta obstetricia et gynecologica Scandinavica. 1988;67(6):539–42.

43. Afratis N, Gialeli C, Nikitovic D, Tsegenidis T, Karousou E, Theocharis AD, Pavão MS, Tzanakakis GN, Karamanos NK. Glycosaminoglycans: key players in cancer cell biology and treatment. Febs J. 2012;279(7):1177–97.

44. Andreopoulou E, Sparano JA. Chemotherapy in patients with anthracycline and taxane-pretreated metastatic breast cancer: an overview. Curr Breast Cancer Rep. 2013;5(1):42–50.

45. Zhang Q, Li J, Xue H, Kong L, Wang Y. Network-based methods for identifying critical pathways of complex diseases: a survey. Mol BioSyst. 2016;12(4):1082–9.

46. Magdeldin S, Yoshida Y, Li H, Maeda Y, Yokoyama M, Enany S, Zhang Y, Xu B, Fujinaka H, Yaoita E, et al. Murine colon proteome and characterization of the protein pathways. BioData Mining. 2012;5(1):1.

47. Johanning GL. Modulation of breast cancer cell adhesion by unsaturated fatty acids. Nutrition. 1996;12(11):810–6.

48. Chen WY, Wu F, You ZY, Zhang ZM, Guo YL, Zhong LX. Analyzing the differentially expressed genes and pathway cross-talk in aggressive breast cancer. J Obstetrics Gynaecol Res. 2015;41(1):132–40.

49. Ferguson MS, Nouraei SR, Davies BJ, McLean N. Basal cell carcinoma of the nipple–areola complex. Dermatologic Surgery. 2009;35(11):1771–5.

50. Frisch M, Hjalgrim H, Olsen JH, Melbye M. Risk for subsequent cancer after diagnosis of basal-cell carcinoma: a population-based, epidemiologic study. Ann Internal Med. 1996;125(10):815–21.

51. Blackburn GL, Maini BS, Bistrian BR, McDermott WV. The effect of cancer on nitrogen, electrolyte, and mineral metabolism. Cancer Res. 1977;37 (7 Part 2):2348–53.

52. Hines JR, Williams JS. Nitrogen mustard as adjunctive chemotherapy for breast carcinoma. Br J Surg. 1975;62(6):497–500.

53. Goggins W, Gao W, Tsao H. Association between female breast cancer and cutaneous melanoma. Int J Cancer. 2004;111(5):792–4.

54. Mocci E, Milne RL, Méndez-Villamil EY, Hopper JL, John EM, Andrulis IL, Chung WK, Daly M, Buys SS, Malats N, et al. Risk of pancreatic cancer in breast cancer families from the breast cancer family registry. Cancer Epidemiol Biomarkers Prevention. 2013;22(5):803–11.

55. Hruban R, Petersen GM, Ha P, Kern S. Genetics of pancreatic cancer. from genes to families. Surgical Oncol Clinics North America. 1998;7(1):1–23.

56. Xu X, Chen J. One-carbon metabolism and breast cancer: an epidemiological perspective. J Genet Genomics. 2009;36(4):203–14.

57. Rabi T, Bishayee A. Terpenoids and breast cancer chemoprevention. Breast Cancer Res Treatment. 2009;115(2):223–39.

58. Yang H, Ping Dou Q. Targeting apoptosis pathway with natural terpenoids: implications for treatment of breast and prostate cancer. Current Drug Targets. 2010;11(6):733–44.

59. Ip C, Ganther H. Comparison of selenium and sulfur analogs in cancer prevention. Carcinogenesis. 1992;13(7):1167–70.

60. Guha P, Bandyopadhyaya G, Polumuri SK, Chumsri S, Gade P, Kalvakolanu DV, Ahmed H. Nicotine promotes apoptosis resistance of breast cancer cells and enrichment of side population cells with cancer stem cell-like properties via a signaling cascade involving galectin-3, α9 nicotinic acetylcholine receptor and stat3. Breast Cancer Res Treatment. 2014;145(1):5–22.

61. Sanz G, Leray I, Dewaele A, Sobilo J, Lerondel S, Bouet S, Grébert D, Monnerie R, Pajot-Augy E, Mir LM. Promotion of cancer cell invasiveness and metastasis emergence caused by olfactory receptor stimulation. PloS One. 2014;9(1):85110.

62. Wang S, Xiao Y-Q, Liu Z-Q, Yang X-H, Xiong X-L, Zhu W-F, Luo D-Y, et al. Network-guided genetic screening for metastasis-related microrna-200c in breast cancer. Tumor. 2013;33(2):111–8.

63. Berteretche M, Dalix A, d'Ornano AC, Bellisle F, Khayat D, Faurion A. Decreased taste sensitivity in cancer patients under chemotherapy. Support Care Cancer. 2004;12(8):571–6.

64. Hong JH, Omur-Ozbek P, Stanek BT, Dietrich AM, Duncan SE, Lee Y, Lesser G. Taste and odor abnormalities in cancer patients. J Support Oncol. 2009;7(2):58–65.

65. Kubo M, Nakamura M, Tasaki A, Yamanaka N, Nakashima H, Nomura M, Kuroki S, Katano M. Hedgehog signaling pathway is a new therapeutic target for patients with breast cancer. Cancer Res. 2004;64(17):6071–4.

66. Tanaka H, Nakamura M, Kameda C, Kubo M, Sato N, Kuroki S, Tanaka M, Katano M. The hedgehog signaling pathway plays an essential role in maintaining the cd44+ cd24-/low subpopulation and the side population of breast cancer cells. Anticancer Res. 2009;29(6):2147–57.

67. Watanabe M, Maemura K, Oki K, Shiraishi N, Shibayama Y, Katsu K. Gamma-aminobutyric acid (GABA) and cell proliferation: focus on cancer cells. Histology and histopathology. 2006;21(10):1135.

68. Berger AM, Farr LA, Kuhn BR, Fischer P, Agrawal S. Values of sleep/wake, activity/rest, circadian rhythms, and fatigue prior to adjuvant breast cancer chemotherapy. J Pain Symptom Management. 2007;33(4):398–409.

69. Schernhammer ES, Kroenke CH, Laden F, Hankinson SE. Night work and risk of breast cancer. Epidemiology. 2006;17(1):108–11.

70. Aricò A., Ferraresso S, Bresolin S, Marconato L, Comazzi S, Te Kronnie G, Aresu L. Array-based comparative genomic hybridization analysis reveals chromosomal copy number aberrations associated with clinical outcome in canine diffuse large b-cell lymphoma. PloS One. 2014;9(11):111817.

71. Melck D, De Petrocellis L, Orlando P, Bisogno T, Laezza C, Bifulco M, Di Marzo V. Suppression of nerve growth factor trk receptors and prolactin receptors by endocannabinoids leads to inhibition of human breast and prostate cancer cell proliferation 1. Endocrinology. 2000;141(1):118–26.

72. Maccarrone M, Finazzi-Agro A. The endocannabinoid system, anandamide and the regulation of mammalian cell apoptosis. Cell Death & Differentiation. 2003;10(9):946–55.

73. DI MARZO V. Targeting the endocannabinoid system in cancer therapy: a call for further research. Nat Med. 2002;8(6):547.

74. Swenson KK, Henly SJ, Shapiro AC, Schroeder LM. Interventions to prevent loss of bone mineral density in women receiving chemotherapy for breast cancer. Clin J Oncol Nursing. 2005;9(2):177.

75. Harvey JM, Clark GM, Osborne CK, Allred DC. Estrogen receptor status by immunohistochemistry is superior to the ligand-binding assay for predicting response to adjuvant endocrine therapy in breast cancer. J Clin Oncol. 1999;17(5):1474–1474.

76. Lin J, Manson JE, Lee IM, Cook NR, Buring JE, Zhang SM. Intakes of calcium and vitamin d and breast cancer risk in women. Arch Int Med. 2007;167(10):1050–9.

77. Cui Y, Rohan TE. Vitamin d, calcium, and breast cancer risk: a review. Cancer Epidemiol Biomarkers Prevention. 2006;15(8):1427–37.

78. Claus EB, Risch N, Thompson WD. Genetic analysis of breast cancer in the cancer and steroid hormone study. Am J Human Genet. 1991;48(2):232.

79. Konecny G, Pauletti G, Pegram M, Untch M, Dandekar S, Aguilar Z, Wilson C, Rong HM, Bauerfeind I, Felber M, et al. Quantitative association between her-2/neu and steroid hormone receptors in hormone receptor-positive primary breast cancer. J Nat Cancer Institute. 2003;95(2):142–53.

80. Stylianou S, Clarke RB, Brennan K. Aberrant activation of notch signaling in human breast cancer. Cancer Res. 2006;66(3):1517–25.

81. Farnie G, Clarke RB. Mammary stem cells and breast cancer-role of notch signalling. Stem Cell Rev. 2007;3(2):169–75.

82. Herr D, Rodewald M, Fraser H, Hack G, Konrad R, Kreienberg R, Wulff C. Potential role of renin-angiotensin-system for tumor angiogenesis in receptor negative breast cancer. Gynecologic Oncol. 2008;109(3):418–25.

83. Koh WP, Yuan JM, Sun CL, van den Berg D, Seow A, Lee HP, Mimi CY. Angiotensin i-converting enzyme (ace) gene polymorphism and breast cancer risk among chinese women in singapore. Cancer Res. 2003;63(3):573–8.