Functional clusters analysis and research based on differential coexpression networks

Shuai Meng, Guixia Liu, Lingtao Su, Liyan Sun, Di Wu, Lingwei Wang and Zhao Zheng

College of Computer Science and Technology, Jilin University, Changchun, PR China; Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, PR China

ABSTRACT
Differential coexpression analysis has gradually become an important approach to improve the conventional method of analysing differentially expressed genes. With this approach, it is possible to discover disease mechanisms and underlying regulatory dynamics which remain obscure in differential expression analysis. The detection of differential coexpression links and functional clusters between different disease states is a demanding task. Nevertheless, there is no gold standard for detecting differential coexpression links and functional clusters. Consequently, we developed a novel fusion algorithm FDvDe (Fusion of differential vertex and differential edge) to detect differential coexpression links by aggregating the set of ‘differential vertex’ and ‘differential edge’. Then, we constructed differential coexpression networks between normal and tumour states by integrating the differential coexpression links. With this approach, we identified 1823 genes and 29370 links. Then, we developed the algorithms GTHC (GO term hierarchical clusters) to identify functional modules. The distance matrix used in the hierarchical process was formed by the GO semantic similarity. Furthermore, we aggregated the densities among clusters describing the connectivity among clusters and topological property analysis to discover the hub genes and hub pathways which play an important role in disease mechanism. In this paper, we showed that our approach worked well on a data set of breast cancer samples (68 tumour samples) and normal samples (73 normal samples), and revealed the crucial role and biological significance of the modules and hub genes found in this approach.

Introduction
In the last 18 years, microarrays have gradually become a common technological means for discovery of disease mechanisms based on datasets of global gene expression. Along with the emergence and development of molecular and cellular biology, we have altered the traditional way of studying disease mechanisms through differential expression analysis to differential coexpression analysis of genes participating together in a distinct functional group or a complex biological pathway. Therefore, a large number of coexpression networks include physical underlying protein–protein interactions and metabolic reactions, as well as functional associations such as epistasis of single nucleotide polymorphisms (SNP), synthetic lethality relationships and coexpression of genes [1–3]. These coexpression networks have been widely applied to answer and solve the most common biological questions including identifying cancer-related genes and drug discovery [3–5].

Coexpression networks, as a type of molecular networks, are constructed using the correlation coefficient of the whole gene set, pair by pair. However, similar to other static molecular networks, most of the algorithms based on coexpression network only reveal the gene regulatory mechanism under one state. Through experimental analysis, the molecular interactions or functional modules vary corresponding to different conditions including genetic variations and changes in the environment. Consequently, the conventional analysis based on one condition has to be modified to adapt to a variety of conditions. Only with this approach, can we understand the underlying cellular dynamics mechanism. Up till now, the differential coexpression network analysis has been developed and has gradually become one of the most important methods for discovery of the underlying mechanisms of pathogenesis, such as the Weighted Gene Coexpression Network Analysis (WGCNA) [6], Differential Coexpression profile (DCp) [7], Differential Coexpression enrichment (DCE) [7], Differential Coexpressed Links (DCELs) [3], Differential Coexpression Network (DCEN) [3], Detection of global changes of the network topology (DCglob) [8] and Detection of local changes of...
the network (DCloc) [8]. Differential coexpression analysis has been applied to successfully identify the differential coexpression modules, a group of genes strongly correlated under one condition but not under other, and differentially expressed genes [9–12].

Although various methods to construct differential coexpression networks have been developed, there is no gold strategy of selecting differentially coexpressed links. Based on the different angles of constructing differential networks, some authors, e.g. Hsu et al. [3] chose the differential coexpressed links by calculating the different value of the pair of genes under different conditions and obtained the pair of genes with P-values lower than a threshold t. Others, for example Ideker and Krogan [13] constructed the differential coexpression network by integrating the edges obtained in the process of edge-wise subtraction of the coexpressed gene pairs between the different disease conditions of A and B. With this approach, it is possible to discover some underlying regulation relationships that are not detected in specific networks. Meanwhile, it will allow us to discard the correlations that do not differentiate one disease state of interest from another [13]. Still others obtained sparse group-specific networks and compared the network topologies, such as modularities between groups and degrees [14,15].

Although most of the above-mentioned methods have revealed some underlying biological mechanisms of pathogenesis and addressed a number of important questions, these approaches are limited for lack of criteria of selecting differentially coexpressed links and the scientific interpretation of the hub genes. In this paper, we developed a novel method, FDvDe (Fusion of Differential vertexes and Differential edges), to select differentially coexpressed links and construct a differential coexpression network by integrating these links. With this method, we considered both the differences of gene pair correlation coefficient and the topological property of the same vertex under the two different conditions, A and B. Next, we developed the algorithm GTHC (gene ontology (GO) term hierarchical clusters) to obtain 64 clusters which are relevant to breast cancer. Afterwards, we studied the correlation among these clusters, and identified a number of hub genes that are crucial to the mechanism of disease. Finally, we found that the genes identified through these approaches occupied an important position in the topological structure of the differential coexpression network by quantitatively calculating the significance of these genes.

Methods

The algorithms FDvDe and GTHC were implemented using the statistical programming language R [16]. While the former was applied to choose the differential coexpressed edges, the latter was applied to detect the functional modules of the differential coexpression network. The flow chart of these two algorithms is illustrated in Figure 1.

Gene expression data pre-processing

The original normalized gene expression datasets of 20 362 genes across breast cancer samples (68 samples) and normal samples (73 samples) were obtained from The Cancer Genome Atlas (TCGA) website. The GO terms could be obtained from the GO website (http://purl.obolibrary.org/obo/go.obo), and we downloaded the whole GO annotations for human genes from the National Center for Biotechnology Information (NCBI) Entrez Gene database. Then we applied some steps of the method proposed by Hsu et al. [3] to conduct the data pre-processing. First, we used the process of log normal distribution to transfer the raw gene expression data to satisfy the normal distribution and then used a quantile normalization method to normalize to make sure that all the arrays were in a similar distribution. Second, according to the circumstance that there were some probes with an NA value, we discarded the probes with more than 20% of NA values and applied the KNN (k-nearest neighbours) algorithm embedded in the R package ‘impute’ to impute the missing data; the parameter k was set to 10 [3]. Third, the expression of genes with multiple expression arrays was calculated by the average of these expressions. Finally, the genes without GO annotations were discarded, as the information provided by GO terms of the gene is needed to obtain the semantic similarity matrix.

Constructing coexpression network

After the data pre-processing step, we obtained the final gene expression data from normal samples denoted by condition A and tumour samples denoted by condition B. We used the Pearson correlations to calculate the correlation coefficient of all pairs of genes separately for condition A and condition B, denoted by $p_{ij}^{c} = \text{pearson}(\text{gene}_i, \text{gene}_j)$, $c = A$ or $B$. With this approach, we obtained the correlation coefficient matrix $A$ and matrix $B$, respectively, for condition $A$ and condition $B$, which were implemented in the next steps.

Differential vertex

The principle of selecting the differential vertexes was based on the topological property. For the coefficient matrix $A$ and $B$, we chose the pairs of genes whose
correlation coefficient values $p$ were over the threshold $T = 0.9$; consequently, we obtained coexpression networks $N_A$ and $N_B$ corresponding to condition $A$ and $B$. For a given gene, we applied the method proposed by Bockmayr et al. [8] to calculate the differential vertex based on the neighbourhood of the given gene, let $V_A$ denote the set of genes which are its neighbours in $N_A$, whereas $V_B$ in $N_B$:

$$d_v = 1 - \frac{V_A \cap V_B}{V_A \cup V_B}$$

(1)
According to this definition, we know that the greater the number of common neighbours of the given gene in both networks, the smaller the value of \( d_i \).

**Differential edge**

After defining \( d_i \), we modified the method proposed by Wally et al. [12] and defined the formula of differential edge \( d_e \) as follows:

\[
d_e = \left| \left( p_{i,j}^A - p_{i,j}^B \right)^\beta \right| \max \left( p_{i,j}^A, p_{i,j}^B \right)
\]

(2)

Where \( p_{i,j}^A, p_{i,j}^B \) denoted the Pearson correlation coefficient of \( \text{gene}_i \) and \( \text{gene}_j \) in both networks. The parameter \( \beta \) was set to 0.8 if the signs of \( p_{i,j}^A \) and \( p_{i,j}^B \) are not the same, e.g., \( p_{i,j}^A \) has a negative sign, while \( p_{i,j}^B \) a positive one; on the contrary, if the sign is the same, \( \beta \) was set to 1. The modification was derived from the hypothesis that the alteration of the sign is more important than the invariability of the sign if the values are equal; for example, the difference of \(-0.2\) and \(-0.2\) is more significant than that of \(-0.3\) and \(-0.7\). The latter part of this formula aimed to solve the following situation: 0.0 to 0.5 and 0.5 to 1.0, in which we considered that the changes of the latter had more importance.

In this equation, we used the method of absolute value to make sure that the result is positive. Meanwhile, we used permutation testing to guarantee the significance of the result. After 1000 random samplings, we obtained the \( P \)-values by calculating the count of the differential edges detected in the permutation testing that were lower than that detected in the original data. Then, we adjusted the \( P \)-values using the Benjamini–Hochberg (BH) procedure [17]. Finally, we chose the differential edges (gene pairs) with \( P \)-value smaller than 2.5 \times 10^{-4} to prepare for the next step.

**FDvDe**

After the above two parallel procedures, we obtained two sets of gene vertices, \( X \) and \( Y \), where \( X \) stands for the vertices obtained in the procedure of differential vertex, and \( Y \) stands for the gene vertices existing in the differential edges (pair of gene vertex) obtained from the procedure of differential edge. The detailed algorithm FDvDe (fusion of differential vertex and edge) using pseudo-code is displayed in Figure 2.

With this procedure, we deleted the edges (gene pairs) if both of the two vertices of an edge are not in the set of \( V \) (set of genes in \( X \cap Y \)) simultaneously. Then we constructed the differential coexpression network with these reserved edges.

**Topological properties**

To analyse the differential coexpression network and explore the biological significance and correlations of the hierarchical clusters and hub genes in the point of structure properties, the following four network topologies were quantified.

1. Degree distributions of differential coexpression network. The distributions were defined to describe the fraction of vertices in the differential network having degree \( k \). As we know, most complicated networks satisfy the property scale-free and power law, expressed as \( P(k) \sim k^{-\gamma} \), where we usually set \( \gamma \) between 2 and 3 [18,19].

2. Average path lengths for differential coexpression networks. The average path length is a concept in network topology that is defined as the average number of steps along the shortest paths for all possible pairs of network nodes; it is a measure of the efficiency of information or mass transport in a network.

3. The significance of seed nodes. We defined this concept to describe the significance of the seed nodes in network topology based on the these two correlation coefficients ECC [20] and \( p_{ij} \):

\[
\text{ECC}_{ij} = \frac{E_{ij}}{\min(k_i, k_j)}
\]

(3)

where \( E_{ij} \) represents the number of triangles that include the edge \( (i, j) \) and \( \min(k_i, k_j) \), the minimum of the degree of \( \text{gene}_i \) and \( \text{gene}_j \), which represents the maximal possible number of triangles including the edge \( (i, j) \). We applied the Pearson correlation coefficient \( p_{ij} = \text{pearson} (\text{gene}_i, \text{gene}_j) \), Where \( p_{ij} \) represents the

---

**Figure 2.** Detailed algorithm of FDvDe.
Pearson correlation coefficient of gene\textsubscript{i} and gene\textsubscript{j}. Then we defined the final correlation of gene\textsubscript{i} and gene\textsubscript{j}, \( F_{ij} \) as follows:

\[
F_{ij} = ECC_{ij} \times p_{ij}
\]  

(4)

And the significance of gene\textsubscript{i}, \( S_i \) was defined as

\[
S_i = \sum_{(ij) \in E} F_{ij}
\]

(5)

where \( j \) represents all of the neighbours of gene\textsubscript{i} in the network. Then we sorted the whole \( S_i \); the more significant the gene was, the greater the number of positions of the gene were.

(4) Connectivity densities

We applied the connectivity densities proposed by Hsu et al. \cite{3} to describe the correlation among the clusters to discover the hub genes that regulate more than one functional module and understand the mechanism of pathogenesis. It was defined as follows:

\[
D_{ij} = \begin{cases} 
\frac{E}{C^2_{ij}}, & \text{if } (i = j) \\
\frac{E}{n_i + n_j}, & \text{if } (i \neq j)
\end{cases}
\]

(6)

where \( E \) represents the number of edges between the two clusters or within the cluster; \( i = j \) denotes that the

**Input**: \( GO^A \), the GO terms of gene A; \( GO^B \), the GO terms of gene B.

**Main step**:

**Step 1**: count: the number of the common terms of \( GO^A \) and \( GO^B \).

**Step 2**: count\textsubscript{a}: the number of GO terms in gene A having the common ancestors with GO terms in gene B. count\textsubscript{b}: the number of GO terms in gene B having the common ancestors with GO terms in gene A. and set the initial value of count\textsubscript{a} and count\textsubscript{b} to count.

**Step 3**: Defined \( GO^{A-B}_i \) to represent the i-th terms in \( GO^A \) and not in \( GO^B \), and defined \( GO^{B-A}_j \) to represent the j-th terms in \( GO^B \) and not in \( GO^A \).

FOR (every terms in \( GO^{A-B} \) )

{ 
    IF \( (GO^{A-B}_i \) have the common ancestor with terms in \( GO^{B-A} \) )
        count\textsubscript{a}=count\textsubscript{a}+1;
    
}

FOR (every terms in \( GO^{B-A} \) )

{ 
    IF \( (GO^{B-A}_j \) have the common ancestor with terms in \( GO^{A-B} \) )
        count\textsubscript{b}=count\textsubscript{b}+1;
    
}

**Step 4**: Defined \( G_{a^2} \) to represent the semantic similarity of gene A and B:

\[
G_{a^2} = \frac{count_{a} \cdot count_{b}}{2}
\]

Figure 3. Details of the algorithm GTHC.
two clusters are the same, whereas $i \neq j$ that the two clusters are not the same. Then we used the z-score to transfer the value $D_{ij}$. Last, we chose a value greater than 4.57 for further research, and we can hypothesize that the chosen clusters have the strongest relations with breast cancer.

**GO term hierarchical clusters**

In this step, we developed a novel algorithm GTHC (GO term hierarchical clusters) to calculate the functional semantic similarity matrix of the whole genes and replace the Pearson correlation coefficient distance matrix with the GO semantic similarity matrix. The algorithm was modified from the method proposed by Schlicker et al. [21]. To calculate the functional similarity, Schlicker combined Resnik’s method, which developed the concept of ‘information content’ [22], and Lin and Dekang’s idea [23] to calculate the semantic similarity between two terms using the proportion of the common ancestors of the terms and the information to describe the terms. After the pre-processing procedure, the genes without annotation of GO term were deleted; consequently, the retained genes were all annotated by a GO term. For the given two genes, $A$ and $B$, annotated by GO terms, let $GO_A^i$ denote the $i$-th GO term of gene $A$ in the whole $N$ terms and $GO_B^j$ denote the $j$-th GO term of gene $B$ in the whole $M$ terms. The algorithm GTHC is displayed in Figure 3.

**Results and discussion**

After the gene expression data pre-processing, we obtained 15,673 genes in normal samples and 15,602 genes in tumour samples for further research. The two algorithms, FDvDe and GTHC, were developed to choose the differential coexpressed links and identify the functional similarity clusters.

**FDvDe**

The FDvDe algorithm was based on two paratactic methods of detecting differential vertexes and differential edges. The result of the dissimilarity degree of vertex under different conditions is shown in Figure 4. We discovered that the number of dissimilarity degrees was close to 1, indicating that most of the genes were topologically dissimilar. Then, we chose the vertexes (genes) having a differential value greater than 0.85 for further research and deleted the other vertexes, which were topologically similar between the two coexpression networks, $A$ and $B$.

Concurrently, we used matrix $A$ and $B$ obtained from the process of constructing the coexpression network to detect the differential edges. Some of the results are demonstrated in Table 1.

**Table 1. Results for differential edges and the corresponding $P$-values.**

| Gene A | Gene B | Differential value | $P$-value | Gene A | Gene B | Differential value | $P$-value |
|--------|--------|-------------------|-----------|--------|--------|-------------------|-----------|
| A1BG   | OSBPL7 | 0.863083 128      | 0.000 221 | GPRIN2 | RNF43  | 1.241349 883      | 0.000 039 |
| AAGAB  | DIS3L  | 0.869059 107      | 0.000 216 | TPD52  | VANGL2 | 1.247799 966      | 0.000 032 |
| AARD   | ANO1   | 0.87049 225       | 0.000 203 | GRHL1  | LMX1B  | 1.24870 512       | 0.000 028 |
| AARD   | ATP1B1 | 0.896587 053      | 0.000 197 | BCL11A | LOC254559 | 1.251137 346   | 0.000 024 |
| AARD   | IL20RA | 0.912219 665      | 0.000 186 | CDH3   | KIAA1324 | 1.262955 169   | 0.000 019 |
| ABCA6  | NGFR   | 0.9415 286        | 0.000 173 | FOXA1  | GABRP  | 1.264008 929      | 0.000 017 |
| ABCA8  | CzfrH0 | 0.941831 906      | 0.000 172 | FAM26E | ST6GAL2 | 1.267190 614      | 0.000 016 |
| ABCB1  | FREM1  | 0.973572 982      | 0.000 159 | SPARC  | ST6GAL2 | 1.273932 556      | 0.000 013 |
| ABCD2  | KATNAL2| 1.051671 666      | 0.000 132 | CDH1   | PROM1  | 1.279989 599      | 0.000 012 |
| ABHD15 | AVP11  | 1.060975 917      | 0.000 127 | EHF    | TFAP2A | 1.260409 093      | 0.000 011 |
| ABHD15 | CTPB2  | 1.06636 586       | 0.000 118 | SDCC4  | TAMPSS13 | 1.281732 049     | 0.000 001 |
| ABHD15 | TRAF4  | 1.16111 159       | 0.000 098 | CHMP4C | IRF6   | 1.301079 329      | <0.000 001 |
| ACACB  | GPT2   | 1.207879 814      | 0.000 056 | PVRIL4 | RNF43  | 1.320567 473      | <0.000 001 |
| AGR2   | GYTLT1B| 1.211047 164      | 0.000 049 | EPCAM  | LOC642587 | 1.352511 554      | <0.000 001 |
| AGR2   | BCL11A | 1.230493 906      | 0.000 043 | ANKRd40 | UBE2Z  | 1.363658 447      | <0.000 001 |

Note: The table shows the edges (Gene A and Gene B), their differential value and the corresponding $P$-value.
After the above procedure, we obtained two datasets, $X$ and $Y$, which were the necessary input in algorithm FDvDe, and the number of gene vertices in $X$ and $Y$ is displayed in Figure 5. Next, we used the FDvDe algorithm to fuse these datasets to obtain the final differential edges. Finally, we used these edges to construct a differential network of 1823 genes and 29,370 edges (pairs of genes).

**Topological properties**

Through the analysis of the differential coexpression network, we discovered that the network satisfied the properties of scale-free and power law which are common in a complicated network, indicating that our method of constructing the network was rational. The degree distribution and the shortest path lengths are described in Figure 6. The top 50 genes with significant topological...
property in the set of seed nodes are displayed in Table 2. The connectivity densities among these 64 clusters are displayed in Figure 7(B).

**GO term hierarchical clusters and connectivity among them**

In the differential coexpression network, all the 1823 genes annotated by GO terms were classified into 64 clusters shown in the heatmap of Figure 7(A). The quantitative distribution of these clusters is shown in Figure 8. The clusters classified by the GTHC algorithm were significantly highly rich in the biological themes. We performed the gene functional enrichment analysis using the bioinformatics tool DAVID [24,25], and the most important results are presented in Table 3. We performed further analysis to study the connectivity among these 64 clusters. The connectivity densities are displayed in Figure 7(B), in which we could discover that clusters 44 and 63 had the closest connection with other clusters. In order to simplify the research and detect the clusters 44 and 63 had the closest connection with other genes significantly. In order to simplify the research and detect the clusters 44 and 63 had the closest connection with other genes significantly. In order to simplify the research and detect the clusters 44 and 63 had the closest connection with other genes significantly. In order to simplify the research and detect the clusters 44 and 63 had the closest connection with other genes significantly. In order to simplify the research and detect the clusters 44 and 63 had the closest connection with other genes significantly. In order to simplify the research and detect the clusters 44 and 63 had the closest connection with other genes significantly. In order to simplify the research and detect the clusters 44 and 63 had the closest connection with other genes significantly.

**Table 2. Top 50 genes in the set of seed nodes.**

| Position | Gene         |
|----------|--------------|
| 1774     | TRPV6        |
| 1775     | CXADR        |
| 1776     | PROM1        |
| 1777     | GRAM2D       |
| 1778     | SOX10        |
| 1779     | GRPIN2       |
| 1780     | VANGL2       |
| 1781     | GRHL1        |
| 1782     | BCL11A       |
| 1783     | MRKOBH       |
| 1784     | 1785         |
| 1786     | 1787         |
| 1788     | 1789         |
| 1790     | 1791         |
| 1792     | 1793         |
| 1794     | 1795         |
| 1796     | 1797         |
| 1798     | 1799         |
| 1800     | 1801         |
| 1802     | 1803         |
| 1804     | 1805         |
| 1806     | 1807         |
| 1808     | 1809         |
| 1810     | 1811         |
| 1812     | 1813         |
| 1814     | 1815         |
| 1816     | 1817         |
| 1818     | 1819         |
| 1820     | 1821         |
| 1822     | 1823         |

In these 64 clusters, there were a large number of genes significantly specifically expressed in the differential coexpression network. Cluster 9 included SPDEF, FOXA1, XBP1 and TFE3, which were differentially expressed in breast cancer [28]. We also found that FOXN2, FOXN3 and FOXP2 were included in the family of FOXA1 and the ARID3A having the same function annotation with ARID (AT-rich interaction domain). Cluster 13 included PAPSS1, AK3, AK5, CK8, HSPA12B, HYOU1, TK2 and UCK2 belonging to the same function term of adenosine triphosphate (ATP)-binding, in which PAPSS1 has been reported to further enhance MTA1s-mediated repression of ligand-ER transactivation function [29]. Meanwhile, cluster 42 included ACCAA2, ACACB, ACSM5, ACSS2 and ACSS3, five genes 100% enriched in the functional term of ATP-binding. Some observations suggest ACSS2 to be expressed to a significant extent in particular tumour types, including triple-negative breast cancers [30]. Cluster 50 included three genes, USP19, USP43 and USP54, regulating the same pathway ‘thiol-dependent ubiquitinyl hydrolase activity.’ Cluster 56 included ITIH5, SERPIN3, SERPINB5, SERPIN1, SPINT1, SPINT2, TIMP2, TIMP4 and WDFC2 belonging to the same term shown in Table 3, in which the SERPINA family genes were studied by de Ronde et al. [31]. We also discovered seventeen genes SLC11A2, SLC25A16 etc belonging to the SCL family, which is significantly expressed in breast cancer [31]. Cluster 63 included PK1B, SFN and YWHAG, in which SFN may account for some of the unexplained multiple-case breast cancer families [32].

Recently, genetic interactions have been discovered to include the between-pathway mode, in which the correlation of gene pairs regulates two different pathways, and the within-pathway mode, in which the connections of gene pairs operate within a single pathway [33]. A high proportion of the interactions existing in the differential coexpression network are of the between-pathway mode [34]. Consequently, we studied the correlation among the 64 clusters, and we chose the clusters among which the connectivity densities are greater than the threshold, and constructed the clusters correlation network shown in Figure 7(C). With this approach, we detected some hub genes that regulate the network by the between-pathway mode. For example, YWHAG, was identified only in cluster 63, and linked to the other 11 clusters of the 14 ones which have interactions with cluster 63. The YWHAG gene has been testified to be important in understanding the breast cancer mechanism [35]. The BTD in cluster 44 was shown to link to 10 of all the
11 clusters which have interactions with cluster 63, and it was verified to be a potential breast cancer biomarker in plasma [36]. FAAH2 widely existing in the same cluster needs to be researched. Recently, CTNNB1, which was included in cluster 52, was researched by Tanabe et al. [37] in cancer and stem cells. When we mapped the hub genes to the set of seed nodes, as expected, most of the hub genes obtained by the employed approach had a significant meaning of the topological properties in the differential coexpression network.

Identification of hub genes based on the GO term and the topological properties

With the approaches described above, we identified a number of genes that existed in the clusters that contribute to different biological processes and we mapped the obtained genes to the set of seed nodes. In the analysis of topological property, we had all the 1823 items in the set of seed nodes (the bigger the number is, the more important the vertex is). Expectedly, most of the hub
genes detected in the clusters had an important position in the seed nodes. For example, SCNN1A (sodium channel epithelial 1 alpha subunit), which has been previously studied [38], existed in cluster 58 and its location was 1819 of the whole 1823 items. The PPAP2C gene (phosphatidic acid phosphatase type 2C) existed in cluster 39 and its location was 1815. The PPAPDC1B gene (phosphatidic acid phosphatase type 2 domain containing 1B) having the same function as PPAP2C, has been reported to be specifically expressed in HER2+ tumours [39]. Overall, these results verified that the genes identified by the method proposed in this paper were related to breast cancer and have very important topological properties in the differential coexpression network.

Conclusions

Although in the last ten years, a number of various methods and algorithms have been proposed to construct differential coexpression networks and identify the functional clusters and hub genes, there is still no gold standard. In this paper, we developed two novel algorithms, FDvDe and GTHC. The former one aimed to obtain the final set of differentially coexpressed edges (pairs of genes) considering both the differences in the correlation of gene expression and the neighbours of the same vertex under different conditions, whereas the latter one was proposed to detect the functional clusters based on the GO term semantic similarity matrix. Then, we conducted integration analysis of the GO term semantic similarity and topological properties in the differential coexpression network. With these approaches, we developed a rational and effective means to construct the differential coexpression network, while the research frame reflected in this experiment to detect the hub genes and testify their statistical significance and topological properties aimed to present a novel research idea to analyse the mechanism of pathogenesis of breast cancer. We believe that the analysis of differential coexpression networks will provide data for future studies on the mechanisms of cancer.

| Cluster | Term | Category          | Count | %     | P-value |
|---------|------|-------------------|-------|-------|---------|
| 1       | Protein binding | GOTERM\_MF\_DIRECT | 88    | 72.1  | <0.001  |
| 2       | Oxidoreductase  | UP\_KEYWORDS      | 22    | 75.9  | <0.001  |
| 8       | Membrane       | UP\_KEYWORDS      | 68    | 98.6  | <0.001  |
| 13      | ATP binding    | GOTERM\_MF\_DIRECT | 8     | 80    | <0.001  |
| 16      | Small GTPase mediated signal transduction | GOTERM\_BP\_DIRECT | 16   | 84.2  | <0.001  |
| 17      | Signal transduction | GOTERM\_BP\_DIRECT | 18  | 75     | <0.001  |
| 24      | RNA-binding    | UP\_KEYWORDS      | 12    | 63.2  | <0.001  |
| 28      | Translational initiation | GOTERM\_BP\_DIRECT | 48   | 84.2  | <0.001  |
| 31      | Structural molecule activity | GOTERM\_MF\_DIRECT | 11   | 91.7  | <0.001  |
| 34      | Signal peptide | UP\_SEQ\_FEATURE  | 17    | 100    | <0.001  |
| 35      | Alternative splicing | UP\_KEYWORDS      | 13    | 92.9  | 0.0023  |
| 43      | Transmembrane helix | KEGG\_PATHWAY    | 13    | 94.1  | <0.001  |
| 50      | Thiol-dependent ubiquitinyl hydrolase activity | GOTERM\_MF\_DIRECT | 3   | 100    | <0.001  |
| 56      | Negative regulation of endopeptidase activity | GOTERM\_BP\_DIRECT | 9   | 100    | <0.001  |
| 58      | Transport      | UP\_KEYWORDS      | 26    | 96.3  | <0.001  |
| 59      | Rho guanyl-nucleotide exchange factor activity | GOTERM\_MF\_DIRECT | 9   | 100    | <0.001  |
| 60      | Phosphoprotein | UP\_KEYWORDS      | 6     | 100    | 0.0102  |
| 61      | Regulation of cyclin-dependent protein serine | GOTERM\_BP\_DIRECT | 2   | 100    | 0.0023  |
| 63      | Protein kinase C inhibitor activity | GOTERM\_MF\_DIRECT | 3   | 100    | <0.001  |
| 64      | Cytoskeleton   | GOTERM\_CC\_DIRECT | 2   | 100    | 0.02    |

Note: The clusters are in column 'Cluster'; column 'Term' represents the function or pathway that the cluster mainly regulates belonging to the biological category shown in the column 'Category'; column 'Count' represents the count of genes in the corresponding term, while '%' represents the count divided by the number of all the genes in the cluster. The 'P-value' reflects the statistical significance.
Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Nature Science Foundation of China [grant number 61373 051], [grant number 61502 343]; Science and Technology Development Program of Jilin Province [grant number 2014020404G]; Science Research Funds for the Guangxi Universities [grant number KY2015ZD122]; Science Research Funds for the Wuzhou University [grant number 2014A002]; Project of Science and Technology Innovation Platform of Computing and Software Science (985 Engineering); Key Laboratory for Symbol Computation and Knowledge Engineering of the National Education Ministry of China; Fundamental Research Funds for the Central.

References

[1] Vidal M, Cusick ME, Barabasi AL. Interactome networks and human disease. Cell. 2011;144(6):986–998.
[2] Mitra K, Carvunis AR, Ramesh SK, et al. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013;14(10):719–732.
[3] Hsu CL, Juan HF, Huang HC. Functional analysis and characterization of differential coexpression networks. SCI REP-UK. 2015 [cited 2015 Aug 18];5(2):13295. DOI:10.1038/srep13295
[4] Harrold JM, Ramanathan M, Mager DE. Network-based approaches in drug discovery and early development. Clin Pharmacol Ther. 2013;94(6):651–658.
[5] Robin X, Croixell P, Radetskaya O, et al. Personalized network-based treatments in oncology. Clin Pharmacol Ther. 2013;94(6):646–650.
[6] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005 [cited 2005 Aug 12];4(1):17. DOI:10.2202/1544-6115.1128
[7] Yu H, Liu BH, Ye ZQ, et al. Link-based quantitative methods to identify differentially coexpressed genes and gene pairs. BMC Bioinform. 2011 [cited 2011 Aug 2];12(1):315. DOI:10.1186/1471-2105-12-315
[8] Bockmayr M, Frederick K, Balazs G, et al. New network topology approaches reveal differential correlation patterns in breast cancer. BMC Syst Biol. 2013 [cited 2017 Mar 1];7:78. DOI:10.1186/1752-0509-7-78
[9] Fuente ADL. From ‘differential expression’ to ‘differential networking’ – identification of dysfunctional regulatory networks in diseases. Trends Genet Tigr. 2010;26(7):326–333.
[10] Southworth LK, Owen AB, Kim SK. Aging mice show a decreasing correlation of gene expression within genetic modules. PLoS Genet. 2009;5(12):61–82.
[11] van Nas A, Guhathakuria D, Wang SS, et al. Elucidating the role of gonadal modules in sexually dimorphic gene coexpression networks. Endocrinology. 2009;150(3):1235–1249.
[12] Walley AJ, Jacobson P, Falchi M, et al. Differential coexpression analysis of obesity-associated networks in human subcutaneous adipose tissue. Int J Obesity. 2012;36(1):137–147.
[13] Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012 [cited 2012 Jan 17];8(1):565. DOI:10.1038/msb.2011.99
[14] Reverter A, Ingham P, Lehnert SA, et al. Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. Bioinformatics. 2006;22(19):2396–2404.
[15] Zhang B, Li H, Riggins RB, et al. Differential dependency network analysis to identify condition-specific topological changes in biological networks. Bioinformatics. 2009;25(4):526–532.
[16] Index R, Development TR, Team C, et al. R: a language and environment for statistical computing. Computing. 2011;1:12–21.
[17] Yick CY, Zwinderman AH, Kunst PW, et al. Transcriptome sequencing (RNA-Seq) of human endobronchial biopsies: asthma versus controls. Eur Respir J. 2013;42(3):662–670.
[18] Choromanski K, Matuszak M, Miękisz J. Scale-free graph with preferential attachment and evolving internal vertex structure. J Stat Phys. 2013;151(6):1175–1183.
[19] Oonela JP, Saramäki J, Hyvönen J, et al. Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA. 2007;104(18):7332–7336.
[20] Radicchi F, Castellano C, Cecconi F, et al. Defining and identifying communities in networks. Proc Natl Acad Sci USA. 2004;101(9):2658–2663.
[21] Schlicker A, Domingues FS, Rahnenführer J, et al. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinform. 2006 [cited 2006 Jun 15];7(1):302. DOI:10.1186/1471-2105-7-302
[22] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Mellish CS, editor. Proceedings of the 14th International Joint Conference on Artificial Intelligence; 1995 Aug 20–25; Montreal, Canada. San Francisco (CA): Morgan Kaufmann Publishers Inc; 1995. p. 448–453.
[23] Lin D. An information-theoretic definition of similarity. In: Shavlik JW, editor. Proceedings of the Fifteenth International Conference on Machine Learning; 1998 Jul 24–27; Madison (WI). San Francisco (CA): Morgan Kaufmann Publishers Inc; 1998. p. 296–304.
[24] Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008;4(1):44–57.
[25] Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
[26] Stark C, Breitkreutz BJ, Reguly T, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006;34(Suppl 1):535–539.
[27] Tucker CL, Fields S. Lethal combinations. Nat Genet. 2003;35(3):204–205.
[28] Doane AS, Danso M, Lal P, et al. An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. Oncogene. 2006;25(28):3994–4008.
[29] Meng Q, Kumar R. Dual targeting of estrogen sulfation and nuclear steroid receptor signaling by MTA1s. Cancer Res. 2006;66:78–78.
[30] Comerford SA, Huang Z, Du X, et al. Acetate dependence of tumors. Cell. 2014;159(7):1591–1602.

[31] de Ronde JJ, Lipset EH, Mulder L, et al. SERPINA6, BEX1, AGTR1, SLC26A3, and LAPTM4B are markers of resistance to neoadjuvant chemotherapy in HER2-negative breast cancer. Breast Cancer Res Tr. 2013;137(1):213–223.

[32] Lewis AG, James F, Anna M, et al. Mutation analysis of FANCD2, BRIP1/BACH1, LMO4 and SFN in familial breast cancer. Breast Cancer Res. 2005;7(6):R1005–R1016.

[33] Costanzo M, Baryshnikova A, Myers CL, et al. Charting the genetic interaction map of a cell. Curr Opin Biotech. 2011;22(1):66–74.

[34] Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks. Nat Biotechnol. 2005;23(5):561–566.

[35] Yoo JO, Kwak SY, An HJ, et al. miR-181b-3p promotes epithelial-mesenchymal transition in breast cancer cells through Snail stabilization by directly targeting YWHAG. Biochim Biophys Acta. 2016;1863(7):1601–1611.

[36] Kang UB, Younghoe A, Won LJ, et al. Differential profiling of breast cancer plasma proteome by isotope-coded affinity tagging method reveals biotinidase as a breast cancer biomarker. BMC Cancer. 2010 [cited 2017 Mar 01];10:114. DOI:10.1186/1471-2407-10-114

[37] Tanabe S, Kawabata T, Aoyagi K, et al. Gene expression and pathway analysis of CTNNB1 in cancer and stem cells. World J Stem Cells. 2016;8(11):384–395.

[38] Varley KE, Gertz J, Roberts BS, et al. Recurrent read-through fusion transcripts in breast cancer. Breast Cancer Res Tr. 2014;146(2):287–297.

[39] Haverty PM, Fridlyand J, Li L, et al. High-resolution genomic and expression analyses of copy number alterations in breast tumors. Gene Chromosome Canc. 2008;47 (6):530–542.