PLOS ONE

RESEARCH ARTICLE

Disease-related gene module detection based on a multi-label propagation clustering algorithm

Xue Jiang[1,2], Han Zhang[1,2], Xiongwen Quan[1,2]*, Zhandong Liu[3], Yanbin Yin[4]

1 College of Computer and Control Engineering, Nankai University, Tianjin 300350, China, **2** Tianjin Key Laboratory of Intelligent Robotics, Nankai University, Tianjin 300350, China, **3** Department of Pediatrics-Neurology, Baylor College of Medicine, Houston, TX 77030, United States of America, **4** Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115, United States of America

* quanxw@nankai.edu.cn

Check for updates

OPEN ACCESS

Citation: Jiang X, Zhang H, Quan X, Liu Z, Yin Y (2017) Disease-related gene module detection based on a multi-label propagation clustering algorithm. PLoS ONE 12(5): e0178006. https://doi.org/10.1371/journal.pone.0178006

Editor: Quan Zou, Tianjin University, CHINA

Received: December 3, 2016

Accepted: May 6, 2017

Published: May 19, 2017

Copyright: © 2017 Jiang et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The gene expression data and the training set used in this study were uploaded to the stable and public accessible repository Figshare, thus the datasets can be freely accessed by other researchers. The gene expression data are available through https://figshare.com/s/48a9acbbc90b23eedaad and the training set are available through https://figshare.com/s/13fdc5c17d736142dcd0.

Funding: This work was supported by 15JCYBJC18900, http://www.tstc.gov.cn/zxbs/asxh/kjjh/yyjc/ (H. Z. is the receiver of the above funding), and 61403213, http://www.nsfc.gov.cn/

Abstract

Detecting disease-related gene modules by analyzing gene expression data is of great significance. It is helpful for exploratory analysis of the interaction mechanisms of genes under complex disease phenotypes. The multi-label propagation algorithm (MLPA) has been widely used in module detection for its fast and easy implementation. The accuracy of MLPA greatly depends on the connections between nodes, and most existing research focuses on measuring the similarity between nodes. However, MLPA does not perform well with loose connections between disease-related genes. Moreover, the biological significance of modules obtained by MLPA has not been demonstrated. To solve these problems, we designed a double label propagation clustering algorithm (DLPCA) based on MLPA to study Huntington's disease. In DLPCA, in addition to category labels, we introduced pathogenic labels to supervise the process of multi-label propagation clustering. The pathogenic labels contain pathogenic information about disease genes and the hierarchical structure of gene expression data. Experimental results demonstrated the superior performance of DLPCA compared with other conventional gene-clustering algorithms.

Introduction

High throughput biotechnologies have been routinely used in biological and biomedical research. As a result, tremendous amounts of large-scale omics data have been generated, providing not only great opportunities but also challenges for understanding the molecular mechanisms of complex diseases [1]. Detecting disease-related gene modules by analyzing gene expression data represents one of these opportunities and challenges. Genes with similar expression patterns, as well as those with similar functions, are more likely to be regulated via the same mechanisms [2]. Therefore, we can extract disease-related molecular mechanisms through gene co-expression analysis if the genes involved in the mechanism form a significant co-expression gene module that contains known disease genes [3, 4]. The essence of such co-expression analysis is a clustering problem.

publish/portal0/tab351/ (X. Q. is the receiver of this funding). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Gene expression data usually share characteristics such as small sample size, high dimensionality, and large amounts of noise. Generally, dimensionality reduction approaches and genome-wide biological network analysis methods have been widely studied for analyzing these data. To understand the interaction mechanisms of genes under complex disease phenotypes, biological network analysis is more appropriate [5]. A gene co-expression network (GCN) is usually constructed by measuring gene expression similarity, which represents the co-expression relationship between genes [6]. Each node in the network represents a single gene, and an edge connecting two genes indicates the co-expression [6].

Label propagation algorithms have been shown to be fast and easy to implement for analyzing large-scale complex networks [7]. Thus, these algorithms have been widely applied in text information retrieval [8, 9], multimedia annotation [10, 11], and community discovery [12–15]. A label propagation algorithm is a semi-supervised learning method based on a graph, which uses labels of some nodes to propagate and mark unlabeled nodes in the network [16]. If the number of label categories for one node exceeds two, the multi-label propagation algorithm (MLPA) is widely used. When the labels of nodes in the network are stable, nodes with the same label will be grouped into one specific category. MLPA is known to be fast and efficient for clustering [17]. The accuracy of MLPA depends heavily on the similarity measure between nodes. Most existing methods focus on developing better similarity measures to improve the performance of MLPA [18]. Cheng [19] measured the similarities between nodes with a sparsity induced similarity measure and conducted classification based on the label propagation results. Wang [20] studied label propagation between heterogeneous networks and proposed a strategy to propagate label information in a disorder-disease gene network. Tian [21] reconstructed a similarity matrix based on a weighted linear combination method. These methods improved the accuracy of MLPA from the similarity measure between nodes, though the biological significance of gene sets obtained by MLPA has not been demonstrated, and the hierarchical structures of gene expression data have not been fully used. In addition, the significance of a disease-related gene module performs poorly with loose connections between disease genes when using conventional gene clustering algorithms [22].

In the past few years, several network-based analysis methods to identify disease-related genes [23, 24] or disease-related microRNAs [25] have been proposed. A new local enrichment analysis method for disease-related genes identification has also been proposed [26]. These methods select the top genes of a ranking list as the most likely disease genes and have improved the accuracy of disease gene prediction. Disease-related genes are selected one by one by using these methods. Considering the complex characteristics of complex diseases [27, 28] and the fact that different molecules often work together to play their roles effectively, it is better to detect disease-related modules, which is helpful for understanding the modular mechanisms during disease progression.

Because biological experiments are time consuming, only a small amount of labeled data is present in biological databases. It is particularly urgent to develop efficient and effective computational methods that make full use of the label information for the small number of samples. Therefore, we developed a double label propagation clustering algorithm (DLPCA) for disease-related gene module detection. Compared with MLPA, DLPCA fully uses pathogenic information for sample genes and the hierarchical structure of biological networks while maintaining a fast running speed. In DLPCA, we used pathogenic labels, which represent pathogenic information for genes and the hierarchical structure of the gene co-expression network to supervise the process of category label propagation clustering. Because the DLPCA contains a semi-supervised pathogenic label propagation step, the clustering results have a clear biological meaning. Moreover, to accelerate convergence speed and improve the robustness of the clustering results, we also proposed a seed node selection method based on the local

topological structure of a gene co-expression network. Experimental results demonstrated the feasibility and effectiveness of DLPCA as well as the superior performance of DLPCA compared with other conventional gene clustering algorithms.

The rest of this study is organized as follows: Materials used in our study and methods proposed in this paper are presented in Section 2. Experiments that analyze the performance of DLPCA and the overall discussion of DLPCA are reported in Section 3. Conclusions, along with some suggestions for future research, are presented in Section 4.

Materials and methods

In this section, first, the gene expression data used in our study are described. Next, the construction of the gene co-expression network is briefly introduced. Then, we present the seed selection method based on local topological information. Finally, we describe the DLPCA.

Gene expression data

The gene expression data used in our study were RNA-seq data downloaded from http://www.hdinhd.org. The data were obtained from the striatum tissue of 6-month-old Huntington's disease (HD) mice. The gene expression data contain 4 genotypes, including polyQ 92, polyQ 111, polyQ 140, and polyQ 175. Each genotype has 8 replications. Thus, the gene expression data comprise 32 samples in total. The gene expression data contain 23,351 genes. After removal of genes with insignificant expression changes, 9578 genes remain for further consideration. The data on modifier genes were from Langfelder [29], which contain 520 genes in the training set, including 89 disease genes and 431 non-disease genes.

HD is a type of neurodegenerative diseases that is reported to be caused by a triplet repeat elongation in the Huntington gene (IT15), which leads to neuronal malfunction and degeneration through numerous interactions between genes and a number of different molecular pathways. The course of the disease is a constant progression of symptoms lasting 15 to 20 years after diagnosis and eventually leading to death. Several molecular mechanisms are involved in HD that lead to neuronal dysfunction. Genes with similar expression patterns are usually regulated via the same mechanism, forming modules in the gene co-expression network. Accordingly, if a module contains a relatively large number of disease genes, the biological function of the module may be highly relevant to the disease. This explains why we seek to extract disease-related modules from the gene co-expression network of HD.

Construction of the gene co-expression network

To conduct the multi-label propagation algorithm, we must construct a gene co-expression network using gene expression data. The gene co-expression network used in our study was constructed using the WGCNA software package [30, 31]. As a scale-free network largely corresponds with biological networks, we used the WGCNA software package in our study to ensure that the gene co-expression network is scale-free [32]. Let x_i denote the expression profile of gene i and x_j denote the expression profile of gene j. The weight of the connection between gene i and gene j is w_{ij}, where $w_{ij} = |cor(x_i, x_j)|^{\beta}$. The parameter β is a soft threshold, which is set as the minimal positive integer that ensures the scale-free topology fit of the gene co-expression network is no more than 0.8. It should be noted that the stronger the Pearson correlation, the larger the weight [30, 31]. In the co-expression network $G = (V, E)$, V is the set of nodes, where one node corresponds to a gene. E is the set of edges, showing the mutual interactions between genes. $W = [w_{ij}]$ is the weight matrix of the gene co-expression network. The adjacency matrix is $A = [a_{ij}]$, where a_{ij} represents the interactions between node i and j.

The calculation of a_{ij} is given by

$$a_{ij} = \begin{cases} 1, & \text{if} \quad w_{ij} \neq 0, \\ 0, & \text{else.} \end{cases} \tag{1}$$

The transition probability matrix is $P = [p_{ij}]$, where p_{ij} denotes the probability of transition from node i to node j. In fact, P is a normalized matrix of W along the row vector. The calculation of p_{ij} is given by

$$p_{ij} = \begin{cases} \frac{w_{ij}}{\sum_{j \in N_i} w_{ij}}, & \text{if} \quad N_i \neq \emptyset, \\ 0, & \text{else,} \end{cases} \tag{2}$$

where N_i is the set of neighboring nodes of node i in the gene co-expression network.

Selection of seed nodes

Gene co-expression networks have been shown to exhibit a modular structure. Good seed nodes are helpful for module detection [33]. According to the local topological structure of the gene co-expression network, we selected seed nodes to accelerate the convergence speed and improve the cluster robustness of MLPA [34]. Since nodes with large clustering coefficients and large degrees can spread information quickly and easily, we selected seed nodes based on degree and clustering coefficient. The details for seed nodes selection are shown below.

Step 1. Compute the clustering coefficient of node i, $c_i = \frac{2\sum_{j,k \in N_i} a_{jk}}{d_i \cdot (d_i - 1)}$, where d_i represents the degree of node i. Then, rank all the nodes in descending order according to the clustering coefficient c_i. R_{c_i} represents the ranking of node i in the ranked list.

Step 2. Compute the degree of node i, $d_i = \sum_{j \in N_i} a_{ij}$. Then, rank all the nodes in descending order according to their degrees. R_{d_i} represents the ranking of node i in the rank list.

Step 3. The rank-product strategy [35] yields the comprehensive ranking of node i, $R_i = (R_{c_i} \cdot R_{d_i})^{\frac{1}{2}}$.

Step 4. Rank R_i, $i \in V$, in ascending order and select the first m nodes as seeds. We denote the seed set as S, while the category label of seed node i is f_i, $i \in S$.

It should be clarified that the category labels of seeds are used to extract modules from the gene co-expression network. In the MLPA results nodes with the same category label are considered a module.

Double label propagation clustering algorithm

To make full use of some genes with pathogenic information and improve the biological meaning of the clusters, we take the pathogenic information of genes into consideration during category label propagation. The initial pathogenic label of a gene is given by Eq (3).

$$l_i^0 = \begin{cases} 1, & \text{if } i \text{ is a disease gene,} \\ -1, & \text{if } i \text{ is a non-disease gene,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

We conduct semi-supervised pathogenic label propagation using the known pathogenic information of some genes to supervise the multi-label propagation clustering, thus obtaining the most likely disease-related modules.

Definition 1 Category label update rule. When multi-label propagation is used to detect functional modules, the following update rule for the category labels is used during the label propagation.

The category label of node i is

$$f(i) = \underset{f_n}{\operatorname{argmax}}(\lambda_1 \sum_{j \in N_i^{f_n}} w_{ij} + \lambda_2 \sum_{j \in N_i^{f_n}} a_{ij} + \lambda_3 \sum_{j \in N_i^{f_n}} l_j^{t-1} \cdot l_i^{t-1}) \tag{4}$$

where $N_i^{f_n}$ represents the neighboring nodes of node i with the category label f_n, $n \in S$. $\lambda_1, \lambda_2, \lambda_3$ are parameters. λ_1 controls the effects caused by weighted connectivity. λ_2 controls the effects caused by the number of neighboring nodes. λ_3 controls the effects caused by the pathogenic information of the neighboring nodes. We assumed that the weighted connectivity, the degree and the pathogenic information have equal influence on the category label of a gene.

Definition 2 Pathogenic label update rule. Based on the topological structure of the gene co-expression network, update the pathogenic label of other nodes in the network by using the small amounts of genes with known pathogenic information.

The pathogenic label of node i is

$$l_i^t = \beta_1 \sum_{j \in N_i^{f(i)}} p_{ij} l_j^{t-1} + \beta_2 \sum_{j \in N_i^{-f(i)}} p_{ij} l_j^{t-1} + \beta_3 l_i^{t-1} \tag{5}$$

where l_i^t is the pathogenic label of node i at the tth iteration, $N_i^{f(i)}$ represents the neighboring nodes of node i whose category label is the same as node i. $N_i^{-f(i)}$ represents the neighboring nodes of node i whose category label is different from node i. The symbols $\beta_1, \beta_2, \beta_3$ are parameters. The parameter β_1 regulates the pathogenic effects caused by the nodes in $N_i^{f(i)}$. The parameter β_2 regulates the pathogenic effects caused by the nodes in $N_i^{-f(i)}$. The $\beta_3 l_i^{t-1}$ ensures that the pathogenic label of node i is stable during the pathogenic label updating process. In addition, $\beta_1 + \beta_2 + \beta_3 = 1$ ensures that the pathogenic label is ultimately convergent [20].

Definition 3 Conditions for termination of iteration. The conditions for termination of DLPCA are that the category labels of the nodes in the network stop changing or that the change of the pathogenic information for any node is less than the threshold. In this study, the threshold is 0.1.

The DLPCA procedure is summarized in Algorithm 1.

Algorithm 1: DLPCA

Input: gene expression data, parameters: λ_1, λ_2, λ_3, β_1, β_2, β_3
Input: pathogenic labels of some genes
Output: gene category label
1: Construct the gene co-expression network. Compute the weight matrix, the adjacency matrix, and the transition probability matrix
2: Select the seed nodes
3: **repeat**
4: Update the gene category label according to Eq (4)
5: Update the gene pathogenic label according to Eq (5)
6: **until** conditions for termination are satisfied
7: **return** gene category label

Table 1. Topological information of the gene co-expression network.

Node number	9587
Average weight	0.291
Average weighted connectivity	18.97
Average degree	65.23
Scatters	536

https://doi.org/10.1371/journal.pone.0178006.t001

Results and discussion

In this section, the selection of parameters is first described. Next, we compare the DLPCA with other conventional methods to demonstrate the superior performance of DLPCA. Second, we analyze the time complexity of DLPCA and MLPA. Third, we conduct an enrichment analysis of the modules obtained using DLPCA. Finally, we present an overall discussion to clearly illustrate the purpose of this study and demonstrate the key point of the algorithm.

Parameter selection

Topological information of the co-expression network is shown in Table 1. Fig 1 indicates the biological reasonability of the gene co-expression network. As shown in Fig 2, the degree and the weighted connectivity exhibit a near-linear correlation. The scatter represents the isolated node in the gene co-expression network.

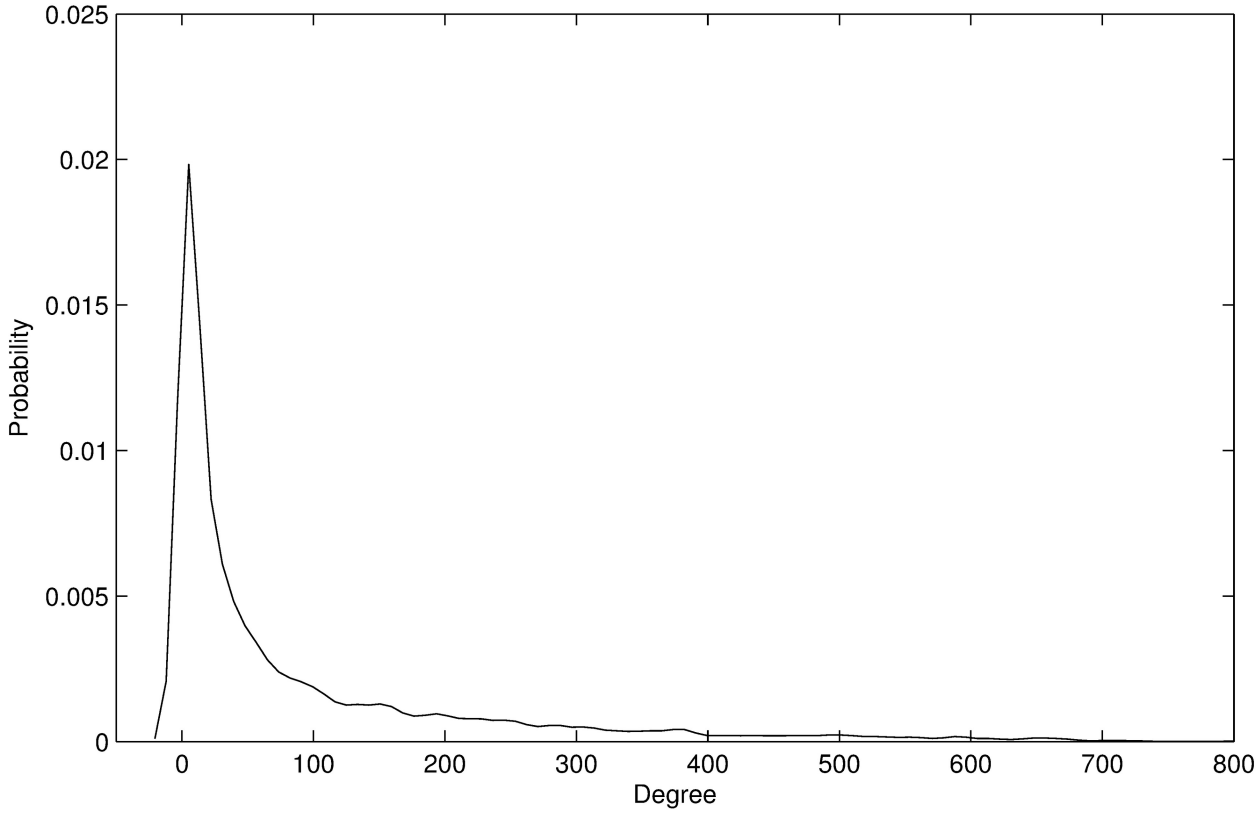

Fig 1. The probability density distribution of weighted connectivity in the co-expression network. The probability density distribution obeys a power-law distribution, showing the biological reasonability of the co-expression network.

https://doi.org/10.1371/journal.pone.0178006.g001

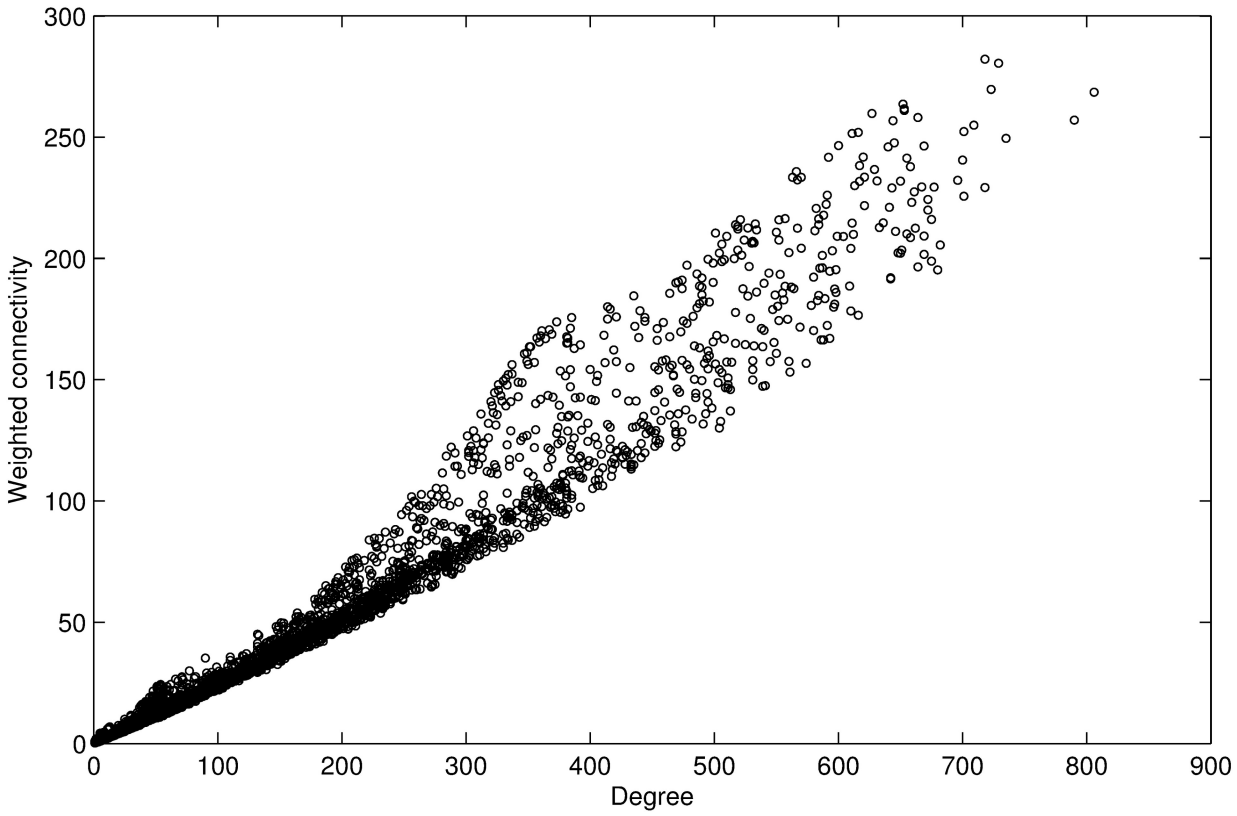

Fig 2. The relationship between degree and weighted connectivity. The scatterplot shows a near-linear correlation between the degree and the weighted connectivity.

https://doi.org/10.1371/journal.pone.0178006.g002

According to Table 1 (the average degree and the average weighted connectivity of the co-expression network) and Fig 2 (the near-linear correlation between the degree and the weighted connectivity), we know that the correlation coefficient is roughly equal to the ratio of the average degree to the average weighted connectivity. Considering that the weighted connectivity, degree, and pathogenic information have equal influence on the category label of a node in the present study, we obtained λ_1: λ_2 = 65.23: 18.97. Following the semi-supervised pathogenic label propagation, the average pathogenic information of all nodes is 0.236. We then obtained λ_2: $\lambda_3 \approx 1$: 0.236 × 0.236. Therefore, we set $\lambda_1 = 3.44$, $\lambda_2 = 1.0$, $\lambda_3 = 20.0$ for computational convenience. It should be noted that traditional MLPA only considers weighted connectivity in category label propagation.

For pathogenic label updating, different parameter combinations may yield different clustering results. To analyze the impact of different parameter combinations on the clustering results, we defined two groups of parameters. Group I is $\beta_1 = \beta_2 = 0.15$, $\beta_3 = 0.7$. Group II is $\beta_1 = 0.2$, $\beta_2 = 0.1$, $\beta_3 = 0.7$.

In addition, to analyze the impact of parameter m on the clustering results, we selected 350, 500, and 750 seed nodes to conduct category label propagation for the traditional MLPA method and the DLPCA method, respectively.

Performance comparison between DLPCA and MLPA

To evaluate the performance of different clustering algorithms, the following criteria were used: the coverage, the significance of the disease-related module, and the significance of

Table 2. The clustering results of each experiment.

Method	Seed num	Module num	Coverage	Avg module size	Avg module significance	Scatter significance	Disease modules		
							num	Avg size	Avg significance
DLPCA $\beta_1 = \beta_2$	350	30	0.750	239.7	0.4240	0.1067	9	539.2	0.8952
	500	31	0.796	246.2	0.4492	0.1218	10	412.5	0.8085
	650	40	0.750	179.8	0.3518	0.1053	10	595.7	0.8090
DLPCA $\beta_1 > \beta_2$	350	25	0.749	276.3	0.1825	0.1491	6	608.8	0.3949
	500	30	0.750	239.7	0.2294	0.1447	8	787.8	0.4015
	650	37	0.752	194.8	0.1595	0.1389	7	710.7	0.4191
MLPA	350	39	0.723	177.8	0.1279	0.1791	6	841.9	0.2923
	500	41	0.724	169.3	0.1465	0.2246	8	839.8	0.3113
	650	116	0.655	80.0	0.1757	0.1803	13	390.5	0.4001

https://doi.org/10.1371/journal.pone.0178006.t002

scatters. The coverage is defined as the ratio of genes in modules to all genes in the network. The significance of the disease-related module is defined as the ratio of disease genes to genes in the training set included in the module. The significance of scatters is defined as the ratio of disease genes to genes in the training set included in the scatters. The clustering results are improved along with increased significance of disease-related modules and decreased significance of scatters. The clustering results of MLPA and DLPCA are shown in Table 2.

Figs 3 and 4 present the clustering results of these methods. As illustrated in Fig 3, with the same seed numbers, the average significance of disease-related modules obtained using

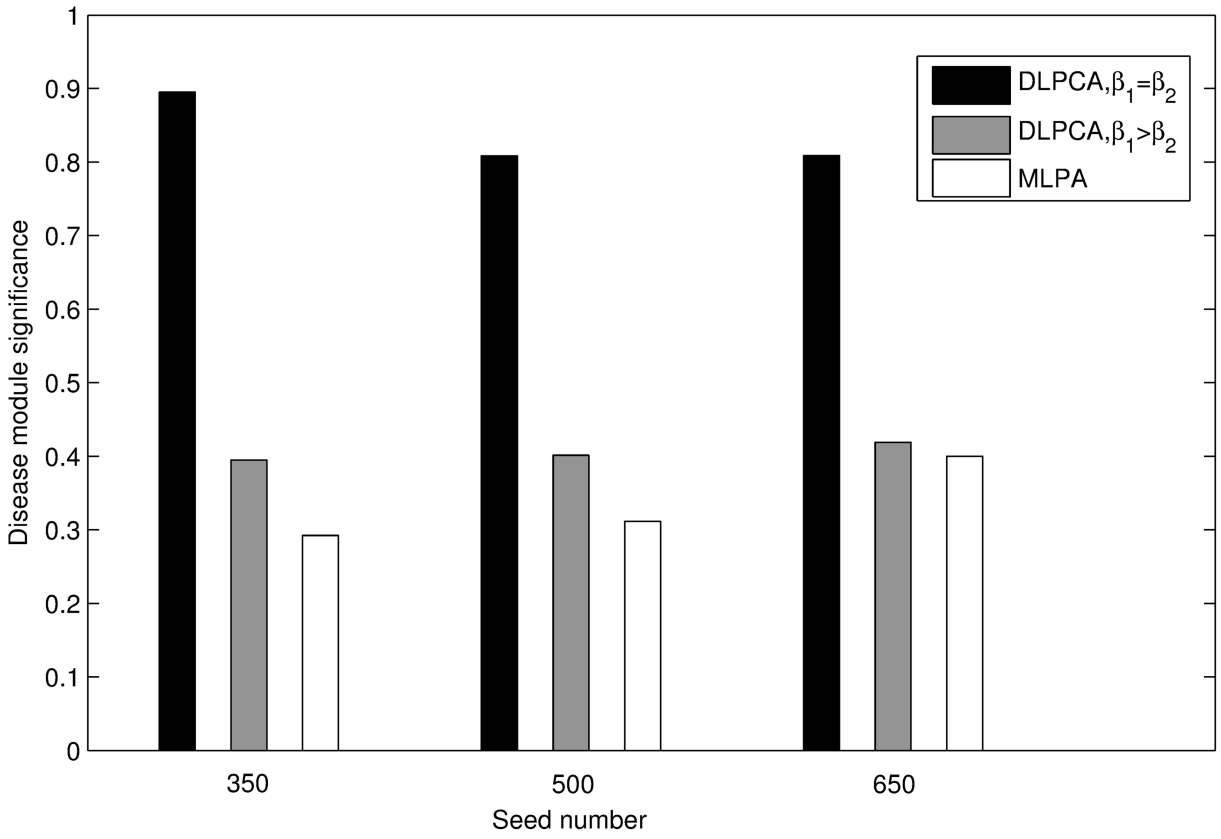

Fig 3. Comparison of the average significance of disease-related modules obtained by DLPCA with $\beta_1 = \beta_2$, DLPCA with $\beta_1 > \beta_2$, and MLPCA. Each grouped bar chart represents the results of different approaches with the same numbers of seeds.

https://doi.org/10.1371/journal.pone.0178006.g003

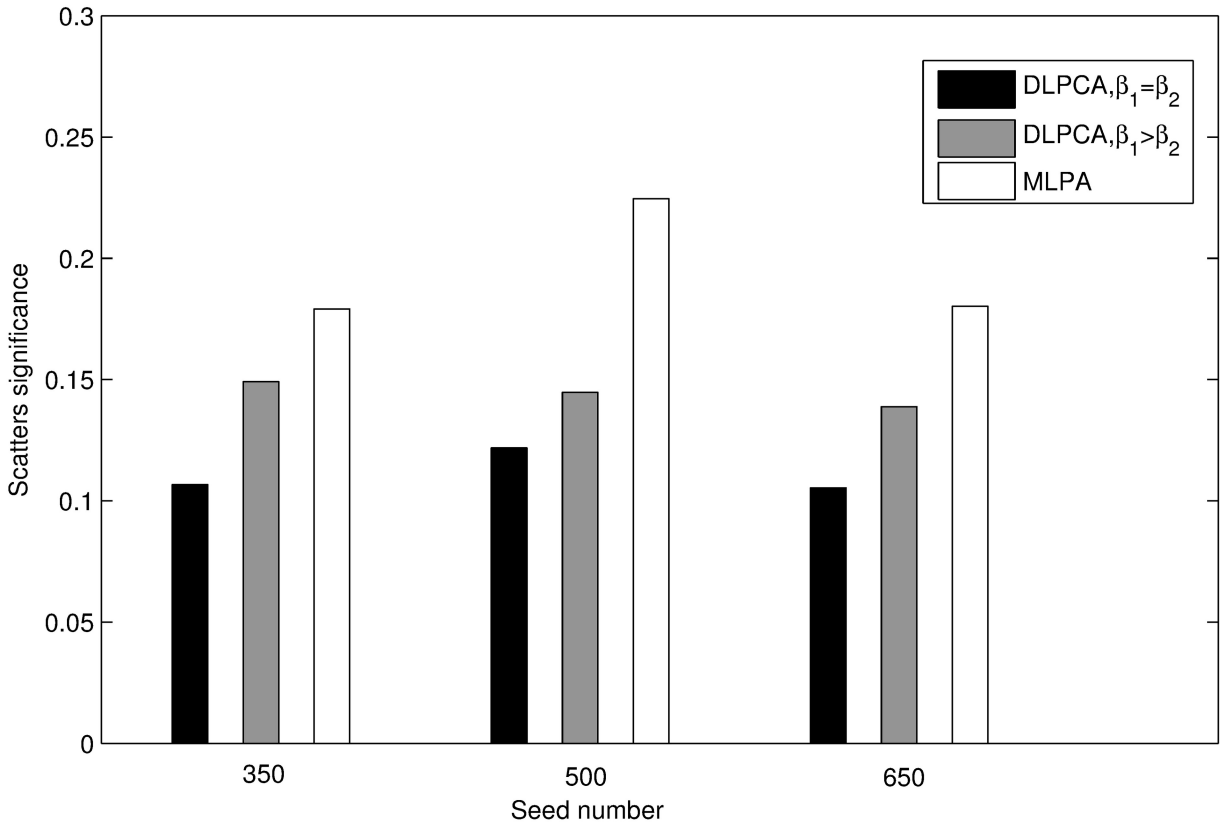

Fig 4. Comparison of the significance of scatters obtained by DLPCA with $\beta_1 = \beta_2$, DLPCA with $\beta_1 > \beta_2$, and MLPCA. Each grouped bar chart represents the results of different approaches with the same numbers of seeds.

https://doi.org/10.1371/journal.pone.0178006.g004

DLPCA with $\beta_1 = \beta_2$ (the average significances of disease modules is 0.837) is higher than that of the other experiments. As shown in Fig 4, with the same number of seeds, the significance of scatters obtained using DLPCA with $\beta_1 = \beta_2$ (the average significances of scatters is 0.111) is lower than that of the other experiments. It is also clear that the average significance of disease-related modules with different numbers of seeds obtained using DLPCA are similar (Fig 3). The significance of scatters obtained using different numbers of seeds in DLPCA are also similar (Fig 4). These results suggest that the clustering results of DLPCA are insensitive to seed number.

Figs 3 and 4 also show that the clustering results of DLPCA with $\beta_1 = \beta_2$ are much better than that of DLPCA with $\beta_1 > \beta_2$, indicating that different parameter combinations significantly impact the clustering results. When the coefficients of the two category labels are equal, i.e., $\beta_1 = \beta_2$ in DLPCA, the average significance of the disease-related modules (the average significance of the disease modules is 0.837) and the significance of the scatters (the significances of the scatters is 0.111) are the best. It demonstrates that DLPCA with $\beta_1 = \beta_2$ can separate disease genes from non-disease genes very well during the clustering process. When the coefficients of the two category labels are not equal, generally, the neighboring nodes whose category labels are the same as that of node i have a greater impact on the pathogenic label of node i, i.e., $\beta_1 > \beta_2$ in DLPCA. Affected by the interaction of the category label and the pathogenic label, DLPCA with $\beta_1 > \beta_2$ may easily fall into local optimization. This situation could be prevented by setting $\beta_1 = \beta_2$, ensuring that category label updating is immune to pathogenic label updating.

Table 3. The clustering results of dynamic cut-off clustering tree algorithm.

DCOTCA	Module num	Coverage	Avg module size	Avg module significance	Scatter significance	Disease modules		
						num	avg size	avg significance
Experiment1	18	0.560	297.4	0.1613	0.1462	14	334.2	0.2064
Experiment2	22	0.568	247.4	0.1572	0.1473	16	286.3	0.2161
Experiment3	34	0.565	159.2	0.1454	0.1521	19	187.6	0.2603
Experiment4	49	0.603	117.9	0.1672	0.1534	26	133.3	0.3030
Experiment5	87	0.606	66.8	0.1950	0.1518	32	82.2	0.3891

https://doi.org/10.1371/journal.pone.0178006.t003

Furthermore, clusters obtained by MLPA have often been shown to contain few genes, which is also confirmed in this study (the average module size of MLPA is shown in Table 2). The experimental results also suggest that MLPA fails to effectively separate disease genes from non-disease genes. However, DLPCA contains a semi-supervised pathogenic label propagation step, which is very helpful for separating disease genes from non-disease genes. DLPCA greatly improves the average significance of disease-related modules compared with MLPA. In the DLPCA results, the sizes of disease-related modules are between 20 and 300 except for two large modules whose sizes are larger than 1000. In summary, DLPCA can effectively improve the performance of clustering results by selecting the appropriate parameters as suggested in our study.

Performance comparison between DLPCA and DCOTCA

To compare the performance of DLPCA with other algorithms, we conducted experiments using the dynamic cut-off tree clustering algorithm (DCOTCA). The clustering results are illustrated in Table 3.

Fig 5 shows that the average significance of disease-related modules using DLPCA (0.837) is much higher than that of DCOTCA (0.275). Fig 6 shows that the average significance of scatters using DLPCA (0.111) is lower than that of DCOTCA (0.150). From Fig 7, we can see that DLPCA also provides much better coverage than other experiments. To summarize, the clustering results of DLPCA are better than those of DCOTCA (Figs 5, 6 and 7).

Time complexity analysis of DLPCA and MLPA

When the number of nodes in the gene co-expression network is n and the number of seed nodes is m, the time complexity of MLPA in each iteration is $O(m \cdot n^2)$. Approximately 10 iterations in each MLPA experiment are needed to reach convergence (Table 4). Given the interaction of the category label and pathogenic label in DLPCA with $\beta_1 > \beta_2$, fewer iteration times are needed relative to DLPCA with $\beta_1 = \beta_2$ to reach convergence. During the category label propagation process, traditional MLPA only considers the impact of weighted connectivity on category label according to a static network; thus, the iteration process of MLPA is the fastest.

The time of per iteration varies. Generally, increasing the seed number increases the time. The average time of each iteration is displayed in Table 5. Note that we used a server with the Linux operating system, 100 GB memory, and an Intel (R) Xeon (R) E5-2603 v3 @1.60GHZ CPU for the data analysis. The algorithm was run on Java 1.7.0_17.

Enrichment analysis

In addition, we conducted enrichment analysis using the DAVID [36] to determine the biological function of the modules obtained using the DLPCA. The clustering results of DLPCA with

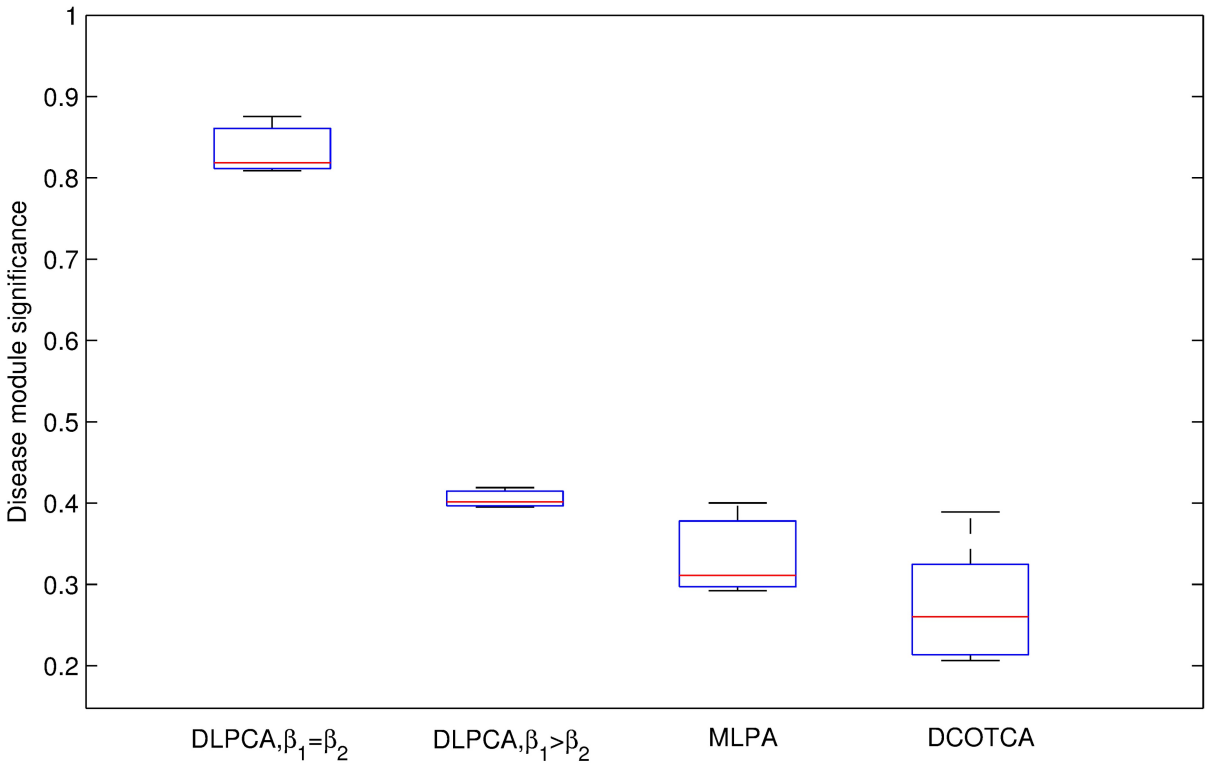

Fig 5. Comparison of the average significance of disease-related modules obtained by DLPCA with $\beta_1 = \beta_2$, DLPCA with $\beta_1 > \beta_2$, and MLPCA. Each box shows the average significance of disease-related modules using an approach with different numbers of seeds.

https://doi.org/10.1371/journal.pone.0178006.g005

$\beta_1 = \beta_2$ and 350 seed genes were used in the enrichment analysis. We listed annotation clusters with high enrichment scores (ES) for the 9 disease-related modules. We also investigated the enrichment annotations for another two modules that are not associated with the disease to analyze the factors that are not effected by or do not contribute to the disease. The detailed annotations of these modules are shown in Table 6.

For module 1 (disease-related module with 91 genes, including 19 disease genes and 1 non-disease genes), identified annotation clusters include metal-binding (cluster 1 with enrichment score 2.69), sequence repeat (cluster 2 with enrichment score 2.54), calcium icon binding (cluster 3 with enrichment score 2.42) and ribosome (cluster 4 with enrichment score 2.14). The significance of the module is 0.95. The annotations for the module suggest that HD maybe associated with these functional annotations above. In fact, HD is caused by the excessive repetition of CAG in the fourth chromosome, which corresponds to the functional annotation, i.e., sequence repeat, of the disease-related module. On the other hand, for module 10 (the non-disease-related module with 623 genes, including 0 disease genes and 137 non-disease genes), annotation clusters such as lysosome (cluster 1 with enrichment score 8.04), cilium (cluster 2 with enrichment score 7.36), Glycoprotein (cluster 3 with enrichment score 5.95) and extracellular matrix (cluster 4 with enrichment score 5.48) were identified. Since the module contains no disease genes, the above functions are most likely not affected by the disease.

The pathology of Huntington disease is very complex and many factors are involved in the disease progression, including inflammation, impaired metabolic pathways, protein mis-

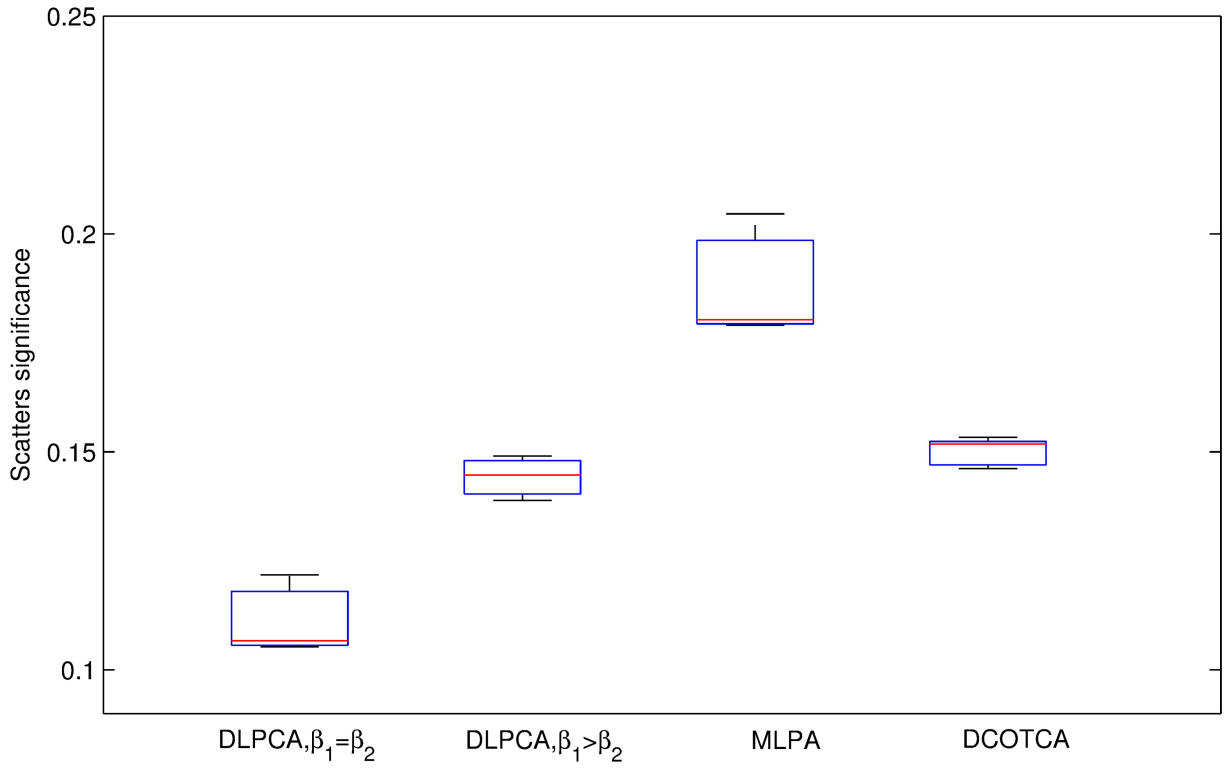

Fig 6. Comparison of the significance for scatters obtained by DLPCA with $\beta_1 = \beta_2$, DLPCA with $\beta_1 > \beta_2$, and MLPCA. Each box shows the significance of scatters using an approach with different numbers of seeds.

https://doi.org/10.1371/journal.pone.0178006.g006

folding [37–39], etc. The enrichment analysis results demonstrate that a disease-related module often contains many functional annotations that could reflect complicated pathologies and also verifies the effectiveness and reasonability of the DLPCA.

An overall discussion

Although tremendous amounts of omics data are being collected along with the rapid development of high-throughput technology, only a small amount of data contain clearly biological annotations, e.g., pathogenic information on genes for specific complex diseases. The challenge is how to fully utilize the small amounts of labeled data to discover effective knowledge from the genome-wide data.

The DLPCA designed in this study aims to mine the most likely disease-related modules from gene expression data by making full use of the pathogenic information of a small number of genes. In addition, DLPCA also makes full use of the hierarchical structures in the network, including the structures represented by the category labels and those represented by the pathogenic labels. This computational method can improve the efficiency and effectiveness of downstream biological experimental analysis. To clarify the main idea of this study and the key point of the DLPCA, we have drawn Fig 8 to clearly demonstrate the properties of DLPCA compared with MLPA. DLPCA is helpful for classifying genes with similar biological properties into one module. Compared with MLPA, DLPCA effectively improves the biological significance of the gene clusters.

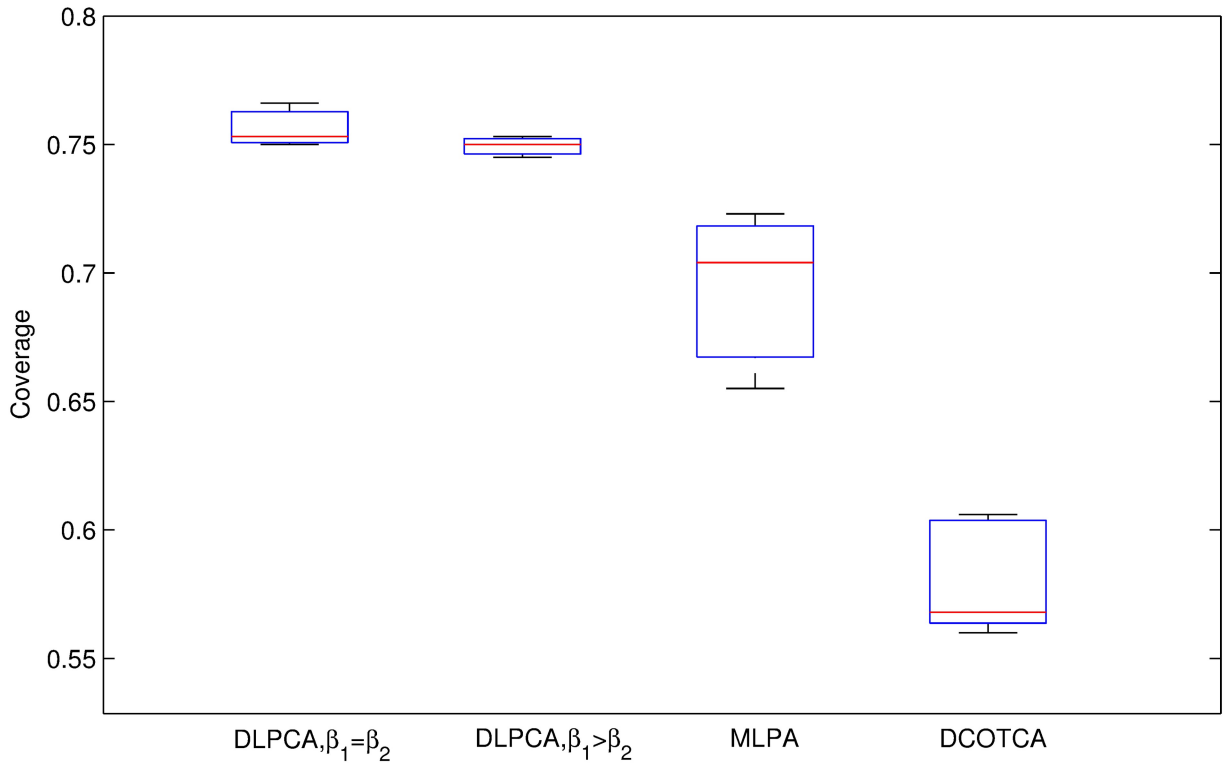

Fig 7. Comparison of the coverage of DLPCA with $\beta_1 = \beta_2$, DLPCA with $\beta_1 > \beta_2$, and MLPCA. Each box shows the coverage using an approach with different numbers of seeds.

https://doi.org/10.1371/journal.pone.0178006.g007

Table 4. The iteration times in each experiment.

Method	DLPCA, $\beta_1 = \beta_2$			DLPCA, $\beta_1 > \beta_2$			MLPA		
Seed num	350	500	650	350	500	650	350	500	650
Iteration times	12	10	14	8	10	11	7	11	8

https://doi.org/10.1371/journal.pone.0178006.t004

Table 5. The average time per iteration in each experiment.

Method	DLPCA, $\beta_1 = \beta_2$			DLPCA, $\beta_1 > \beta_2$			MLPA		
Seed num	350	500	650	350	500	650	350	500	650
Time (hours)	3.1	5.8	6.3	2.9	4.9	6.4	2.6	3.9	5.6

https://doi.org/10.1371/journal.pone.0178006.t005

Conclusions

In this study, we designed a double label propagation clustering algorithm for detecting disease-related modules. This algorithm takes the pathogenic information of genes as a property of nodes in the gene co-expression network. During the clustering process of MLPA, DLPCA not only considers the topological structures of the network but also the biological properties of the nodes in the network. In addition, to accelerate convergence and improve cluster robustness, we also proposed a seed selection strategy according to the local topological structure of the gene co-expression network. Compared with the aforementioned conventional methods, DLPCA effectively improves the accuracy of disease-related module identification.

Table 6. Functional annotation clustering for the modules obtained using DLPCA with $\beta_1 = \beta_2$ and 350 seed nodes.

DLPCA	Module size	Disease genes num	Non-disease genes num	Module sig	Functional annotation clustering	
					Annotation cluster	ES
Module1	91	19	1	0.95	Metal-binding	2.69
					Sequence repeat	2.54
					Calcium icon binding	2.42
					Ribosome	2.14
Module2	109	12	0	1.0	Cytoskeleton	4.05
					Cell junction	3.67
					Calcium icon transport	2.57
					Oxytocin signaling pathway	2.04
Module3	81	9	0	1.0	Golgi apparatus	1.99
Module4	28	2	0	1.0	Ubl conjugation pathway	2.12
Module5	31	2	0	1.0	Nucleotide-binding	3.67
Module6	104	3	0	1.0	Postsynaptic density	3.07
					Endoplasmic reticulum	2.04
Module7	71	5	0	1.0	Retrograde endocannabinoid signaling	3.67
					Membrane	3.01
Module8	1907	18	1	0.95	Nucleotide-binding	9.75
					Chaperone	6.58
					DnaJ domain	6.04
					F-box domain	5.69
					Microtubule	4.94
Module9	2431	15	79	0.16	Zinc, metal-binding	41.41
					Protein transport	10.1
					Ligase	9.87
					Transcription regulation	9.52
					Zinc figure	9.13
Module10	623	0	137	–	Lysosome	8.04
					Cilium	7.36
					Glycoprotein	5.95
					Extracellular matrix	5.48
Module11	1524	0	256	–	Synapse	28.21
					Ion transport	10.12
					Glycoprotein	6.83
					Fatty acid	5.74

https://doi.org/10.1371/journal.pone.0178006.t006

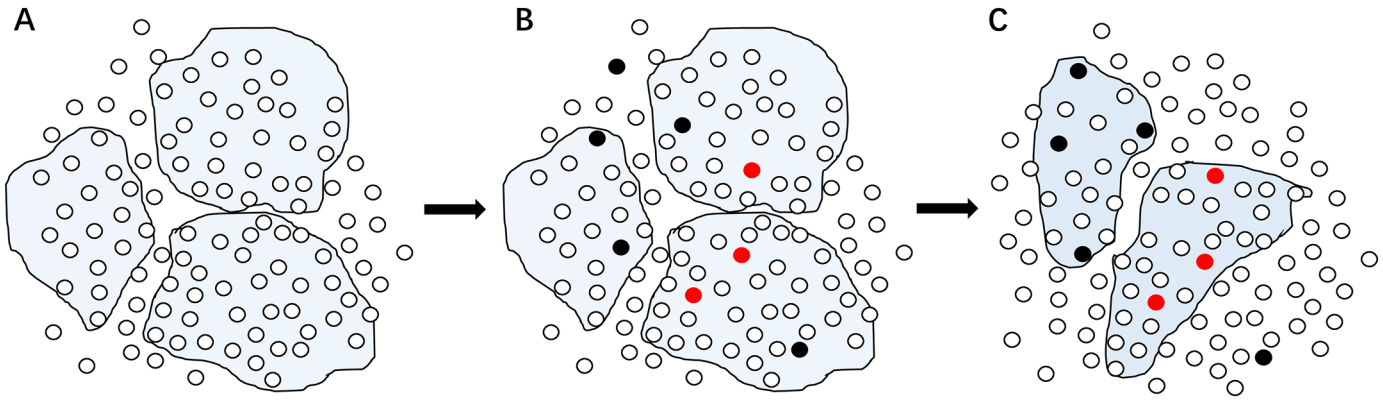

Fig 8. Illustration of DLPCA compared with MLPA. (A) The modules in the gene co-expression network obtained using MLPA. (B) Introduction of the pathogenic information of some genes. Here, red nodes represent disease genes and black nodes represent non-disease genes. (C) The new modular structures in the gene co-expression network obtained using DLPCA.

https://doi.org/10.1371/journal.pone.0178006.g008

However, it should be stated that DLPCA could be applied equally well to other biological networks and genomic data.

Recently, new module detection methods integrating different network structures have been proposed [40]. Generally, the accuracy of disease module detection may be further improved by integrating other biological data as well as gene expression data, especially for gene expression data characterized by large amounts of noise. Therefore, our future efforts will focus on integrating multi-source biological data to further improve the accuracy of disease-related modules.

Acknowledgments

The authors would like to thank the editor and the reviewers for their comments and suggestions, which helped improve the manuscript greatly. We are grateful to associate professor Feng Duan of Nankai University for his careful correction of the English writing. We also thank Mr. Haibin Sun for assistance and discussions.

Author Contributions

Conceptualization: XQ HZ XJ.

Data curation: ZL.

Formal analysis: XJ YY.

Funding acquisition: XQ HZ.

Investigation: XJ YY.

Methodology: XJ.

Project administration: XQ HZ.

Resources: ZL.

Software: XJ.

Supervision: HZ.

Validation: XJ.

Visualization: XJ.

Writing – original draft: XJ HZ.

Writing – review & editing: XJ HZ.

References

1. Tang CY, Hung CL, Zheng H, Lin CY, Jiang H. Novel computational technologies for next-generation sequencing data analysis and their applications. Magnetic Resonance in Medicine. 2015; 3: 1–3. https://doi.org/10.1155/2015/254685 PMID: 26576413
2. Allocco DJ, Kohane IS, Butte AJ. Quantifying the relationship between co-expression, co-regulation and gene function. BMC Bioinformatics. 2004; 5: 1–10. https://doi.org/10.1186/1471-2105-5-18 PMID: 15053845
3. Goh KI. The human disease network. Proceedings of the National Academy of Sciences. 2007; 104: 8685–8690. https://doi.org/10.1073/pnas.0701361104
4. Dong J, Horvath S. Understanding network concepts in modules. BMC systems biology. 2007; 1: 1–20. https://doi.org/10.1186/1752-0509-1-24

5. Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature. 2009; 461: 218–223. https://doi.org/10.1038/nature08454 PMID: 19741703
6. Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. Nature Reviews Genetics. 2004; 5: 101–113. https://doi.org/10.1038/nrg1272 PMID: 14735121
7. Leung IXY, Pan H, Lio P, Crowcroft J. Towards real-time community detection in large networks. Physical Review E Statistical Nonlinear and Soft Matter Physics. 2009; 79: 853–857. https://doi.org/10.1103/PhysRevE.79.066107 PMID: 19658564
8. Yang L, Ji D, Nie Y. Information retrieval using label propagation based ranking. Proceedings of NTCIR6 Workshop Meeting. 2007; 140–144.
9. Speriosu M, Sudan N, Upadhyay S, Baldridge J. Twitter polarity classification with label propagation over lexical links and the follower graph. In The Workshop on Unsupervised Learning in Nlp. 2011; 23: 53–63.
10. Tang J, Hua XS, Qi GJ, Song Y, Wu X. Video annotation based on kernel linear neighborhood propagation. IEEE Transactions on Multimedia. 2008; 10: 620–628.
11. Ismail B, Maher M. Image annotation and retrieval based on multi-modal feature clustering and similarity propagation. University of Louisville, Louisville, KY, USA. 2011.
12. Raghavan UN, Albert R, Kumara S. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E Statistical Nonlinear and Soft Matter Physics. 2007; 76: 036106. https://doi.org/10.1103/PhysRevE.76.036106 PMID: 17930305
13. Gregory S. Finding overlapping communities in networks by label propagation. New Journal of Physics. 2009; 12: 2011–2024. https://doi.org/10.1088/1367-2630/12/10/103018
14. Barber MJ, Clark JW. Detecting network communities by propagating labels under constraints. Physical Review E Statistical Nonlinear and Soft Matter Physics. 2009; 80: 283–289. https://doi.org/10.1103/PhysRevE.80.026129 PMID: 19792222
15. Chen YZ, Shi S, Chen GL, ZhiYong YU. Overlapping community discovery based on node hierarchy and label propagation gain. Moshi Shibie Yu Rengong Zhineng/pattern Recognition and Artificial Intelligence. 2015; 28: 289–298.
16. Zhu X, Ghahramani Z. Learning from labeled and unlabeled data with label propagation. Technical Report CMUCALD-02-107. 2003.
17. Pfeiffer M, Pfeiffer M. Clustering by passing messages between data points. Science. 2007; 315: 972–976.
18. Lei Q, Yu HP, Wu M. Active semi-supervised affinity propagation clustering algorithm. Mathematical Modeling and Its Applications. 2015; 28: 961–968.
19. Cheng H, Liu Z, Yang J. Sparsity induced similarity measure for label propagation. In IEEE International Conference on Computer Vision. 2009; 30: 317–324.
20. Hwang T, Kuang R. A heterogeneous label propagation algorithm for disease gene discovery. SIAM International Conference on Data Mining. 2010; 583–594.
21. Tian Z, Kuang R. Global linear neighborhoods for efficient label propagation. SIAM International Conference on Data Mining. 2012; 863–872.
22. Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: An R package for affinity propagation clustering. Bioinformatics. 2011; 27: 2463–2464. https://doi.org/10.1093/bioinformatics/btr406 PMID: 21737437
23. Zeng X, Liao Y, Liu Y, Zou Q. Prediction and validation of disease genes using HeteSim Scores. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2016; 1–9. https://doi.org/10.1109/TCBB.2016.2520947 PMID: 26890920
24. Liu Y, Zeng X, He Z, Zou Q. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2016; 1–11. https://doi.org/10.1109/TCBB.2016.2550432 PMID: 27076459
25. Zeng X, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. Briefings in Bioinformatics. 2016; 17: 193–203. https://doi.org/10.1093/bib/bbv033 PMID: 26059461
26. Gwinner F, Boulday G, Vandiedonck C, Arnould M, Cardoso C, Nikolayeva I, et al. Network-based analysis of omics data: The LEAN method. Bioinformatics, 2016; 33: 1–9. https://doi.org/10.1093/bioinformatics/btw676 PMID: 27797778
27. Knox SS. From'omics' to complex disease: A systems biology approach to gene-environment interactions in cancer. Cancer Cell International. 2010; 10: 1–11. https://doi.org/10.1186/1475-2867-10-11 PMID: 20420667

28. Romanoski CE, Lee S, Kim MJ, Ingram-Drake L, Plaisier CL, Yordanova R, et al. Systems genetics analysis of gene-by-environment interactions in human cells. American Journal of Human Genetics. 2010; 86: 399–410. https://doi.org/10.1016/j.ajhg.2010.02.002 PMID: 20170901

29. Langfelder P, Cantle JP, Chatzopoulou D, Wang N, Gao F, Al-Ramahi I, et al. Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. Nature Neuroscience. 2016; 19: 623–633. https://doi.org/10.1038/nn.4256 PMID: 26900923

30. Langfelder P, Horvath S. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9: 1–13. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008

31. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology. 2005; 4: 1–45. https://doi.org/10.2202/1544-6115.1128 PMID: 16646834

32. Barabasi AL. Scale-free networks: A decade and beyond. Science. 2009; 325: 412–413. https://doi.org/10.1126/science.1173299 PMID: 19628854

33. Moradi F, Olovsson T, Tsigas P. A local seed selection algorithm for overlapping community detection. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. 2014; 1–8.

34. Subelj L, Bajec M. Robust network community detection using balanced propagation. Physics of Condensed Matter. 2011; 81: 353–362.

35. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics. 2008; 24: 374–382. https://doi.org/10.1093/bioinformatics/btm620 PMID: 18204063

36. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009; 4: 44–57. https://doi.org/10.1038/nprot.2008.211 PMID: 19131956

37. Shynrye L, Kim HJ. Prion-like mechanism in amyotrophic lateral sclerosis: are protein aggregates the key?. Experimental Neurobiology. 2015; 24: 1–7. https://doi.org/10.5607/en.2015.24.1.1 PMID: 25792864

38. Lim J, Yue Z. Neuronal aggregates: formation, clearance, and spreading. Developmental Cell. 2015; 32: 491–514. https://doi.org/10.1016/j.devcel.2015.02.002 PMID: 25710535

39. Wang X, Huang T, Bu G, Xu H. Dysregulation of protein trafficking in neurodegeneration. Molecular Neurodegeneration. 2014; 9: 1–9.

40. Ding Z, Zhang X, Sun D, Luo B. Overlapping community detection based on network decomposition. Scientific Reports. 2016; 6: 24115. https://doi.org/10.1038/srep24115 PMID: 27066904