

RESEARCH ARTICLE

MRHCA: a nonparametric statistics based method for hub and co-expression module identification in large gene co-expression network

Yu Zhang^{1,*}, Sha Cao², Jing Zhao³, Burair Alsaihati⁴, Qin Ma^{5,*} and Chi Zhang^{6,*}

¹ Colleges of Computer Science and Technology, Jilin University, Changchun 130012, China

² Department of Biostatistics, Indiana University School of Medicine, IN 46202, USA

³ Center for Health Outcomes and Population Research, Sanford Research, Sioux Falls, SD 57104, USA

⁴ Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, GA 30602, USA

⁵ Bioinformatics and Mathematical Biosciences Lab, Department of Agronomy, Horticulture, and Plant Science, South Dakota State University, Brookings, SD 57007, USA

⁶ Center for Computational Biology and Bioinformatics and Department of Medical and Molecular Genetics, Indiana University School of Medicine, IN 46202, USA

* Correspondence: zy26@jlu.edu.cn, qin.ma@sdstate.edu, czhang87@iu.edu

Received July 30, 2017; Revised October 4, 2017; Accepted October 24, 2017

Background: Gene co-expression and differential co-expression analysis has been increasingly used to study co-functional and co-regulatory biological mechanisms from large scale transcriptomics data sets.

Methods: In this study, we develop a nonparametric approach to identify hub genes and modules in a large co-expression network with low computational and memory cost, namely MRHCA.

Results: We have applied the method to simulated transcriptomics data sets and demonstrated MRHCA can accurately identify hub genes and estimate size of co-expression modules. With applying MRHCA and differential co-expression analysis to *E. coli* and TCGA cancer data, we have identified significant condition specific activated genes in *E. coli* and distinct gene expression regulatory mechanisms between the cancer types with high copy number variation and small somatic mutations.

Conclusion: Our analysis has demonstrated MRHCA can (i) deal with large association networks, (ii) rigorously assess statistical significance for hubs and module sizes, (iii) identify co-expression modules with low associations, (iv) detect small and significant modules, and (v) allow genes to be present in more than one modules, compared with existing methods.

Keywords: gene co-expression network; algorithm for large scale networks analysis; statistical significance of gene co-expression; Mutual Rank

Author summary: MRHCA—a non-parametric co-expression analysis method for large association network is developed in this work. The advantage of MRHCA includes: (i) it can be generally applied to association networks of most assessment methods; (ii) it outputs exact significance level for each identified hub; (iii) the method are with relative small computational consumption that can be applied to large association networks; (iv) it is sensitive to the modules of weak associations; and (v) it enables overlapped modules in its outputs. R codes and a C package of the MRHCA are released by the following GitHub link: <https://github.com/zy26/mrct>.

INTRODUCTION

Co-expression network analysis is increasingly used to systematically identify all combinations of genes sharing similar functional status and how they vary under different environmental conditions. In multiple plant and bacteria studies, the method has mainly been used to study species specific regulatory mechanism [1–3]. A powerful utilization of the method is differential co-expression (DC) analysis, which could identify condition specific co-expression modules by comparing multiple networks. In human disease study, DC analysis gains its popularity with its strength in finding groups of co-functioning genes with discerning disease progression relevance [4]. In fact, applications of DC analysis in pan-cancer level have identified possible drivers of metabolic reprogramming and high lactate production through cancer progression [5,6].

Weighted correlation network analysis (WGCNA) is among the most popular co-expression network analysis methods, which identifies co-expression modules using a hierarchical clustering based network partition [7]. Andy *et al.* has developed a spectral clustering approach for co-expression network partition [8]. Other co-expression network analysis methods partition the genes using algorithm such as k-mean or self-organizing maps [9,10]. These methods can efficiently identify large co-expression modules in a co-expression network and reflect the topological features of the large modules. However, these methods usually miss out modules with small number of genes, as they lack a rigorous thresholding in their network partitioning method. Our past analysis revealed that a large number of small but robust co-expression modules usually exist in a real co-expression network, which collectively enrich target gene set of the same transcriptional regulatory signals [6,11]. In addition, the network partition methods exclusively assign one gene into one module, while a large number of genes may belong to multiple modules, hence are not suitable for DC analysis. Another drawback of existing co-expression methods is their high computational and memory cost. WGCNA works the best with a whole network of less than 5,000 genes; spectral clustering and similar approaches have computational consumption of at least $O(n^3)$, where n being the total number of genes, causing these methods inapplicable for large data sets [12,13].

Our developed Mutual Rank-based Hub and Co-expression Analysis, namely, MRHCA, is a non-parametric approach for identification of hub genes in a co-expression network, as well as the module genes with estimated size associated with each hub. Compared with existing methods, MRHCA could deal with large number of genes, identify co-expression modules with low

association values, and allow genes to be present in more than one modules. We have applied the method to one simulated co-expression data set, seven *E. coli* transcriptomics data sets, and RNA-Seq data sets of 21 TCGA cancer types, and demonstrated the validity and capacity of the method.

RESULTS

Mutual Rank (MR) based formulation of hub property in a co-expression network

Denote a transcriptomic data set of M genes and N samples by $\mathbf{G}_{M \times N}$, and $\mathbf{G}_{M \times M}$ is defined as the absolute value of the pairwise gene correlation matrix for the M genes, with larger values in \mathbf{C} suggesting stronger co-expression association. The Mutual Rank (MR) of two genes i and j is a non-parametric statistics characterizing the level of association between the two genes, defined by

$$MR(i,j) = MR(j,i) = \sqrt{\text{Rank}(i \rightarrow j) \cdot \text{Rank}(j \rightarrow i)},$$

where $\text{Rank}(i \rightarrow j)$ and $\text{Rank}(j \rightarrow i)$ are the ranks of element $\mathbf{C}(i,j)$ in the j -th column of \mathbf{C} , $\mathbf{C}(\cdot, j)$, and the i -th row of \mathbf{C} , $\mathbf{C}(i, \cdot)$, respectively [3,6,14]. Smaller ranks denote higher association values. Particularly, we set $\mathbf{C}(i,i) = 0$ in our analysis, which implies $MR(i,j) = M$, for $i = 1, \dots, M$.

We have previously revealed that the distribution of the MRs can reflect how likely it is for a gene to be a hub in a co-expression network [6,11]. Intuitively, if gene i is a hub of a co-expression network, and the non-hub neighbors of gene i are denoted as $x_k, k = 1, \dots, K$, then these neighbor genes should show stronger associations with gene i comparing to other genes, hence smaller values of $\text{Rank}(i \rightarrow x_k)$. On the other hand, $\text{Rank}(x_k \rightarrow i)$ are also small since i has stronger associations to x_k comparing to others. These indicate in the case when gene i is hub of a co-expression network, $MR(i, \cdot)$ is very small.

Define the growth rate of a gene with the j -th ranked association to gene i with step size P as a vector:

$$\text{Growth}(MR_{i,(j)}, P) = (g_1, g_2, \dots, g_M)$$

where $MR_{i,(j)}, j = 1, \dots, M$ is j -th smallest element of $MR(i, \cdot)$, and

$$g_k = \frac{MR_{i,(j_{k+})} - MR_{i,(j_{k-})}}{j_{k+} - j_{k-} + 1}, k = 1, \dots, M$$

with

$$j_{k-} = \max\left(k - \frac{P}{2}, 0\right), j_{k+} = \min\left(k + \frac{P}{2}, M\right).$$

The discussion above suggests that a gene module of size K with hub gene i should have the first K values,

g_1, g_2, \dots, g_K , in vector $\text{Growth}(\mathbf{MR}_{i,(j)}, \mathbf{P})$ enriched by small values. In our past study, we have demonstrated the use of the statistics, $\text{Growth}(\mathbf{MR}_{i,(j)}, \mathbf{P})$ in hub identification, and developed a permutation test based approach to assess the significance level of the identified hubs [6]. However, the approach was limited to small datasets due to the computational and memory consumption of permutation tests. In this study, we have developed a statistical approximation of the null distribution of \mathbf{MR} , which could be easily used to assess the significance of hubs, and this algorithm is then ready to be applied on much larger co-expression networks.

Approximation of the null distribution of \mathbf{MR}

We define $\tilde{\mathbf{G}}_{M \times N}$ as a randomized gene expression data set generated by randomly shuffling each row of $\mathbf{G}_{M \times N}$ respectively, and $\tilde{\mathbf{C}}_{M \times M}$ as the co-expression matrix associated with $\tilde{\mathbf{G}}_{M \times N}$. In addition, we define $\hat{\mathbf{C}}_{M \times M}$ as a randomly assigned co-expression association matrix of $\tilde{\mathbf{G}}_{M \times N}$ generated by:

$$\hat{C}(i, j) \sim f(x) \text{ i.i.d. for } \forall i \neq j,$$

$$\hat{C}(i, i) = 0, i = 1, \dots, N$$

$$\hat{C}(i, j) = \hat{C}(j, i) > 0, i \neq j, i = 1, \dots, N, j = 1, \dots, N$$

where $f(x)$ is the empirical distribution of $\{\tilde{C}(i, j) | i < j\}$.

Denote the \mathbf{MR} matrices computed from $\tilde{\mathbf{G}}_{M \times N}$ and $\hat{\mathbf{C}}_{M \times M}$ by $\widetilde{\mathbf{MR}}_{M \times M}$ and $\widehat{\mathbf{MR}}_{M \times M}$, respectively.

It is noteworthy that $\widetilde{\mathbf{MR}}_{M \times M}(i, \cdot)$ form an empirical distribution of $\mathbf{MR}_{M \times M}(i, \cdot)$ under the null hypothesis, i.e. i is not presented in any co-expression module. The goal here is to derive an approximation of the distribution of $\widetilde{\mathbf{MR}}_{M \times M}(i, \cdot)$ that demand less computation consumption comparing to permutation. We have first mathematically proved that an empirical null distribution of $\widehat{\mathbf{MR}}_{M \times M}(i, j), i \neq j$, can be derived by generating matrices $\widehat{\mathbf{MR}}(i, j)'$ through the following three steps (see more details in the Section of Methods):

- (i) Sample p_{ij} from the uniform distribution $U(0, 1)$;
- (ii) Sample $X_{i,j,1}, X_{i,j,2}$ from binomial distribution $\text{Binom}(n-2, p_{ij})$ independently
- (iii) Write $\widehat{\mathbf{MR}}(i, j)' = \sqrt{(X_{i,j,1} + 1) \cdot (X_{i,j,2} + 1)}$.
- (iv) Repeating steps (i)–(iii) multiple times, the generated $\widehat{\mathbf{MR}}(i, j)'$ forms an empirical null distribution of $\widehat{\mathbf{MR}}(i, j)$.

It is worth to note that this null distribution of $\widehat{\mathbf{MR}}(i, j)$ is independent to the method that was used to calculate the gene-wise correlations.

While elements in $\hat{\mathbf{C}}_{M \times M}$ are independent of each other, the off-diagonal elements in $\tilde{\mathbf{C}}_{M \times M}$ are usually not independent as a distance measure, and this will cause certain level of deviance of $\widehat{\mathbf{MR}}$ from $\widetilde{\mathbf{MR}}$. However, since $\tilde{\mathbf{C}}_{M \times M}$ is formed by pair-wise associations of randomly reshuffled vectors and the dependence among elements will be weak when M is sufficiently large, we posit the null distribution of $\widehat{\mathbf{MR}}$ can be a good approximation of the distribution of $\widetilde{\mathbf{MR}}$. In the next section, we will test if the distribution of $\widetilde{\mathbf{MR}}(i, \cdot)$ can be well fitted by the null distribution of $\widehat{\mathbf{MR}}$ for the commonly used gene co-expression association assessment methods.

Randomized data based goodness of fit test for the empirical null distribution of $\widehat{\mathbf{MR}}$

We conducted a goodness of fit test to evaluate if the null distribution of $\widehat{\mathbf{MR}}(i, j)$ can be applied to approximate the distribution of $\widetilde{\mathbf{MR}}(i, \cdot)$ when M is sufficiently large, say $M > 100$. The evaluation was conducted on a whole genome expression profile of multiple *E. coli* samples, for nine most commonly used methods of calculating co-expression association matrix, namely Pearson Correlation (PC), Spearman Correlation (SC), Kendall rank Correlation (KC), Weighted rank Correlation (WC), Distance Covariance (DC), Hoeffding's measure of dependence (HD), Polynomial-Regression based dependence (PR), Spline-Regression based dependence (SR), and Mutual Information (MI) [15,16].

Within the *E. coli* expression matrix $\mathbf{G}_{M \times N}$, we select the expressions of the top K highly expressed genes, $\mathbf{G}_{K \times N}$, and generate a randomized gene expression data $\tilde{\mathbf{G}}_{K \times N}$ by shuffling each row of $\mathbf{G}_{K \times N}$ (see more details in the Section of Methods). We compute $\tilde{\mathbf{C}}_{K \times K}^M$ and $\widetilde{\mathbf{MR}}_{K \times K}^M$ by using each of the co-expression association method \mathcal{M} and generate an empirical null distribution of $\widehat{\mathbf{MR}}_{K \times K}$ with 1000 iterations. The goodness of fit of $\widetilde{\mathbf{MR}}_{K \times K}^M(i, \cdot)$

by the empirical null distribution of $\widehat{\mathbf{MR}}_{K \times K}$ is tested by Kolmogorov Smirnov (KS) test (see the section of Methods). We have conducted the tests for $K = 50, 100, 200, 500, 1,000$, and $2,000$. Figure 1 shows the empirical cumulative density functions of $\widetilde{\mathbf{MR}}_{K \times K}^M(i, \cdot)$ versus the empirical null distribution of $\widehat{\mathbf{MR}}_{K \times K}$, as well as

distributions of p values of the KS tests among the distribution of $\widehat{MR}_{K \times K}^M(i, \cdot)$ and simulated $\widetilde{MR}_{K \times K}^M(i, \cdot)$ and the empirical null distribution of $\widehat{MR}_{K \times K}$, when $K = 1,000$. We could see that the derived empirical null distribution of $\widehat{MR}_{K \times K}$ can accurately fit $\widetilde{MR}_{K \times K}^M(i, \cdot)$ for eight out of the nine analyzed co-expression association methods, namely PC, SC, KC, WC, HD, SR, PR, and MI, for all the selected K values. However, for method DC, significant differences between $\widetilde{MR}_{K \times K}^M(i, \cdot)$ and $\widehat{MR}_{K \times K}^M(i, \cdot)$, as well as $\widehat{MR}_{K \times K}^M(i, \cdot)$ and the empirical null distribution of $\widehat{MR}_{K \times K}$, are observed for $K = 1,000$ and $2,000$, suggesting the $\widetilde{MR}_{K \times K}^M(i, \cdot)$ associated with DC cannot be well approximated by its null distribution when K is large. We have also observed significant differences of $\widetilde{MR}_{K \times K}^M(i, \cdot)$ versus $\widehat{MR}_{K \times K}^M(i, \cdot)$ and its empirical null distribution for MI when the number of bins is improperly selected (see Supplementary Figure S1). It is worth to note that the $\widehat{MR}_{K \times K}^M(i, \cdot)$ can be well fitted by its empirical null distribution in all of our tests, which validates the derived empirical null distribution, and further indicates the fitting bias is caused by the difference between $\widetilde{MR}_{K \times K}^M(i, \cdot)$ and $\widehat{MR}_{K \times K}^M(i, \cdot)$ due to the non-negligible dependence among the pairwise co-

expression associations formulated by \mathcal{M} . Empirical cumulative density functions of $\widetilde{MR}_{K \times K}^M(i, \cdot)$ and p values of KS tests for all the tested K and \mathcal{M} are given in Supplementary Figure S2.

Identification of hub genes by the empirical null distribution of \widehat{MR}

After demonstrating that the empirical null distribution of \widehat{MR} can approximate the distribution of $\widetilde{MR}_{K \times K}^M(i, \cdot)$ for eight out of nine co-expression association methods when K is large, we apply this fact into calculating the significance level of observed $\text{Growth}(\mathbf{MR}_{i,(j)}, P)$ under a null hypothesis that the genes are expressed independent of each other. Compared with the permutation test based significance. In detail, an empirical distribution of $\text{Growth}(\mathbf{MR}_{i,(j)}, P)$ under the null hypothesis is approximately generated from the empirical null distribution of \widehat{MR} , and the significance levels of $\text{Growth}(\mathbf{MR}_{i,(j)}, P)$ for all i and j can thus be assessed. As a result, for all pairs of i and j , the significance of gene i being a hub gene with module size j is evaluated by whether there are in total j significantly small values in $\text{Growth}(\mathbf{MR}_{i,(j)}, P)$. Figure 2A and 2B illustrate MR's growth rate of genuine hub genes in two co-expression modules with size = 20 and 300, versus the empirical null distribution of $\text{Growth}(\widehat{MR}_{i,(j)}, P)$ calcu-

Pseudo code of MR based method:

```

Input data: gene expression matrix  $G_{M \times N}$  or co-expression association  $C_{M \times N}$ 
For  $i$  in 1 to  $M$  do
  Compute  $R_{M \times N}$  with  $R(i, j) = \text{Rank}(j \rightarrow i)$ 
For  $i$  in 1 to  $M$ 
  For  $i$  in 1 to  $M$  do
    Compute  $MR_{M \times M}$  by  $MR(i, j) = \sqrt{R(i, j) \cdot R(j, i)}$ 
  Sort  $MR(i, \cdot)$  to  $MR_{i,(j)}$ 
  Compute growth  $(MR_{i,(j)}, P)$ 
For  $i$  in 1 to ROUDS
  For  $i$  in 1 to  $M$  do
    Generate random number  $p$  from  $U(0, 1)$ 
    Generate two independent random number  $X_1, X_2$  from  $\text{Binom}(p, M - 1)$ 
     $MR^E(i, j) = \sqrt{(X_1 + 1)(X_2 + 1)}$  #empirical null distribution of  $\widehat{MR}_{M \times M}$ 
    Sort  $MR^E(i, \cdot)$  to  $MR^E_{i,(j)}$ 

  Compute growth  $(MR^E_{i,(j)}, P)$  #empirical distribution of growth  $(MR^E_{i,(j)}, P)$ 
For  $i$  in 1 to  $M$ 
  Compare growth  $(MR_{i,(j)}, P)$  vs  $\text{Growth}(MR^E, P)$  to assess
  1) Significance level of hub property of  $i$ 
  2) If  $i$  is a significant hub, the size of module centered by  $i$ 

```

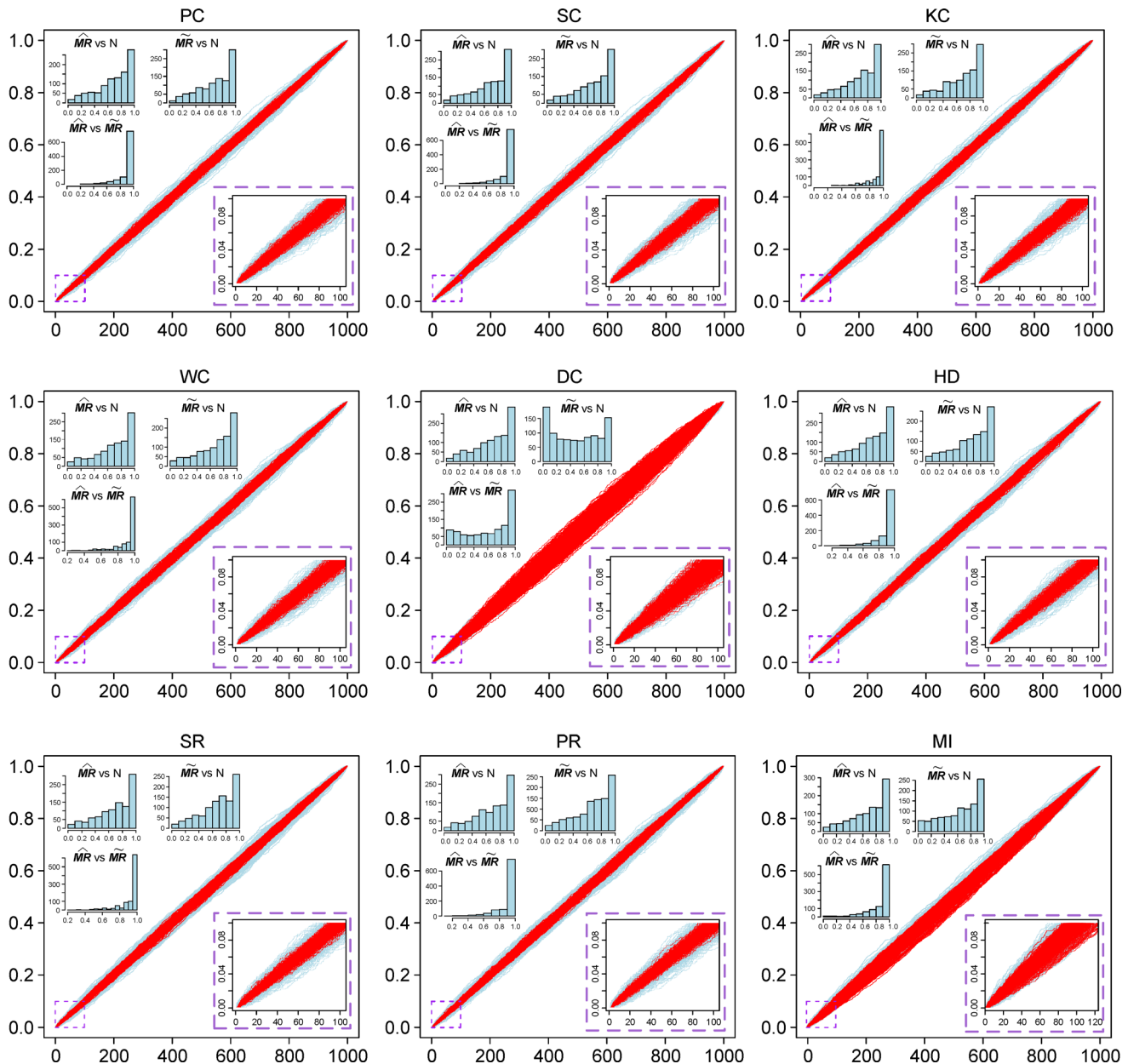


Figure 1. Empirical cumulative density functions of $\widehat{MR}_{K \times K}^M(i, \cdot)$ versus the empirical null distribution of $\widehat{MR}_{K \times K}$ for the nine co-expression association methods. 1,000 randomly generated cumulative density functions of $\widehat{MR}_{K \times K}^M(i, \cdot)$ and empirical null distributions of $\widehat{MR}_{K \times K}$ are colored by red and light blue in each plot, with the 0–0.1 quantile enlarged on the bottom-right part. The x-axis and y-axis represent the MR values in increasing order and the empirical cumulative probability, respectively. p values of the KS tests among the simulated $\widehat{MR}_{K \times K}^M(i, \cdot)$ (\widehat{MR}), $\widehat{MR}_{K \times K}^M(i, \cdot)$ (\widehat{MR}), and the empirical distribution of $\widehat{MR}_{K \times K}(N)$ are shown on the top-left of each plot. The number of bins used for computing the MI is five.

lated based on our method. Pseudo code of the MR based identification of hub genes and co-expression modules is listed below see more details in the section of Method.

Our MR based hub gene and co-expression module identification is more sensitive to co-expressed modules

with low co-expression associations, as it is based on a nonparametric statistics. In addition, since MRHCA constructs co-expression modules from identified hubs, it can more accurately characterize the overlapping characteristics among different modules, compared with

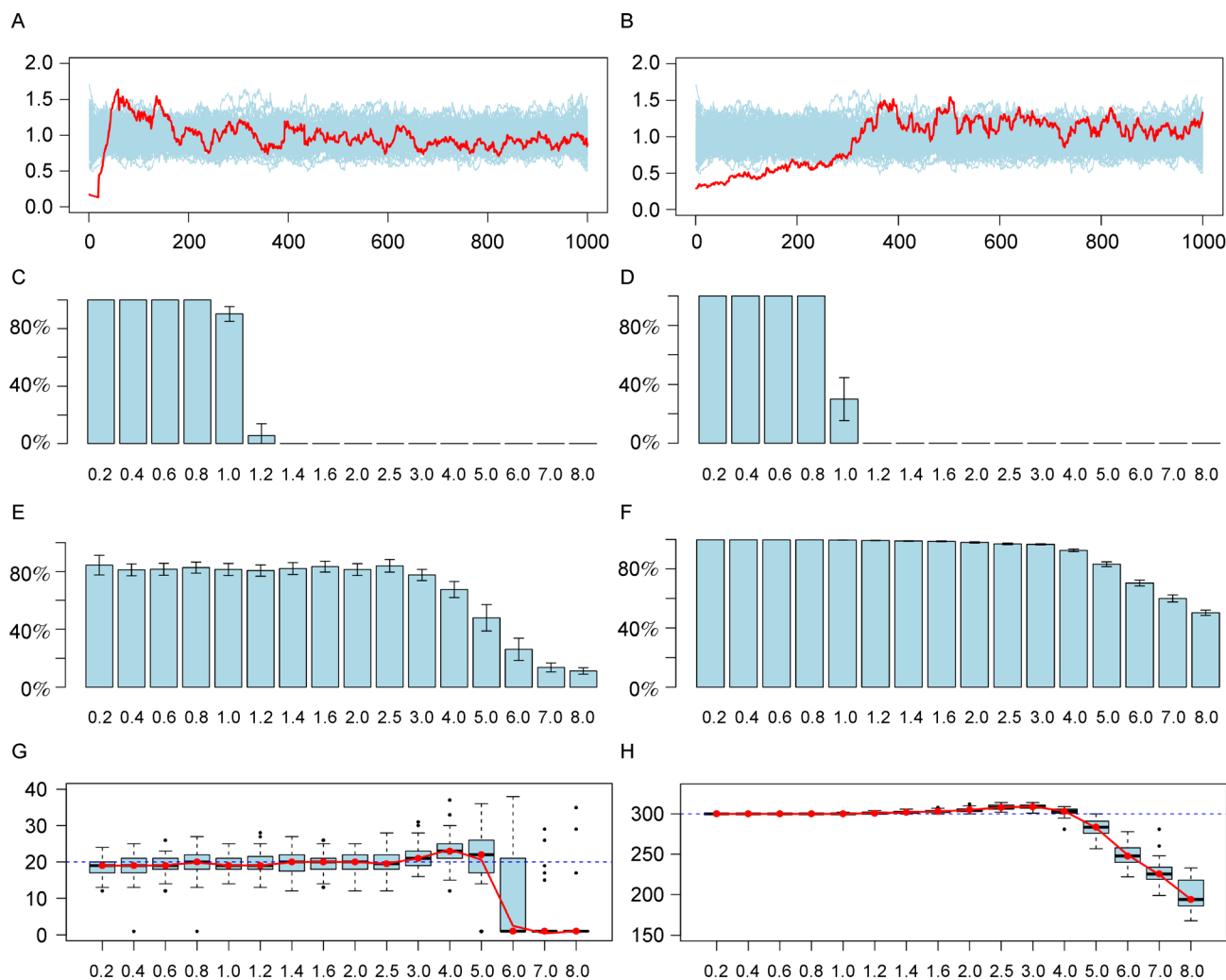


Figure 2. Statistics of the hub genes and co-expression modules identified by MRHCA in simulated data. (A,B) Growth rates of MR of real hub (red) versus empirical null distribution (light blue) when module size is 20 (A) and 300 (B). (C,D) Jaccard Indices between the real module versus WGCNA identified modules with respect to different σ^2 (x-axis). (E,F) Jaccard Indices between the real module versus MRHCA identified modules with respect to different σ^2 (x-axis). (G,H) Average size of the module (y-axis) centered by the top significant hub in 100 rounds of simulation with different σ^2 (x-axis), when the simulated hub and co-expression module with size 20 (G) and 300 (H).

existing partitioning methods that do not allow overlapping genes between modules [17].

Evaluation of the performance of MRHCA on simulated co-expression data

We compare the performances of MRHCA with WGCNA, which is the most popular gene co-expression identification method, on simulated gene expression data sets. We randomly shuffled rows of the gene expression profile of the top 1,000 highly expressed genes in the *E. coli* dataset mentioned above. For each gene, we first normalize its expressions across multiple samples by

dividing each expression value by the standard deviation of the gene across all the samples. Then, we simulate four non-overlapping co-expression modules of size 20, 40, 60 and 300 from four randomly assigned hub genes via $G(X_i) = G(H_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, where $G(H_i)$ and $G(X_i)$ are the expression profile of a hub gene and the rest genes in the module. For each $\sigma^2 = 0.2, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 2.0, 2.5, 3.0, 4.0, 5.0, 6.0, 7.0$, and 8.0 , we simulate 100 gene expression datasets of size 20, 40, 60, 300, and MRHCA and WGCNA are applied to the simulated data for hub gene and co-expression module identification. Of note, there is only one hub and thus one module for each simulated dataset.

We have seen that WGCNA can accurately identify simulated co-expression modules for $\sigma^2 \leq 0.8$ while no module is identified when $\sigma^2 > 1.2$. Figure 2C and 2D shows the Jaccard indices between the WGCNA identified modules and true co-expression modules, with module sizes 20 and 300. In contrast, MRHCA can accurately identify the co-expression modules for $\sigma^2 \leq 3$, as shown by the Jaccard indices in Figure 2E and 2F. In addition, MRHCA achieves more than 50% prediction accuracy for the co-expression module of size 20 when $\sigma^2 \leq 5$, and for module of size 300 when $\sigma^2 \leq 8$. Meanwhile, both methods are with satisfactory low false discover rate. We see no false discoveries made by all the tests for WGCNA, while less than 0.01% false discoveries are made for MRHCA only when $\sigma^2 < 2.5$, as shown in Supplementary Figure S3.

We have further evaluated the sizes of the co-expression modules identified by MRHCA. Of note, all the true hub genes are identified by MRHCA as the top significant hubs. We have also seen that the method

can accurately predict the module size for the true hubs when $\sigma^2 \leq 5$, as shown in Figure 2G and 2H. It is also worth to note that there are more than 5% non-hub genes are also identified as hubs when $\sigma^2 \leq 5$. Further analysis revealed that all of these identified hub genes are in the module, which suggests a low false discover rate of MRHCA regarding to module identification. We posit such an observation is caused by the natural hierarchical structure of gene co-expression network. It is worth to note that some of the transcriptional regulatory signals, which are the true hubs of co-expression modules, cannot be quantified in transcriptomics level, such as post-translational modification of a transcription factor or changes in epi-genomics level [18,19]. MRHCA will comprehensively identify all genes show significant hub property, hence can robustly target the co-expression module when the true hub is missing. Similar results for simulated datasets with modules of sizes 40 and 60 are given in Supplementary Figure S4.

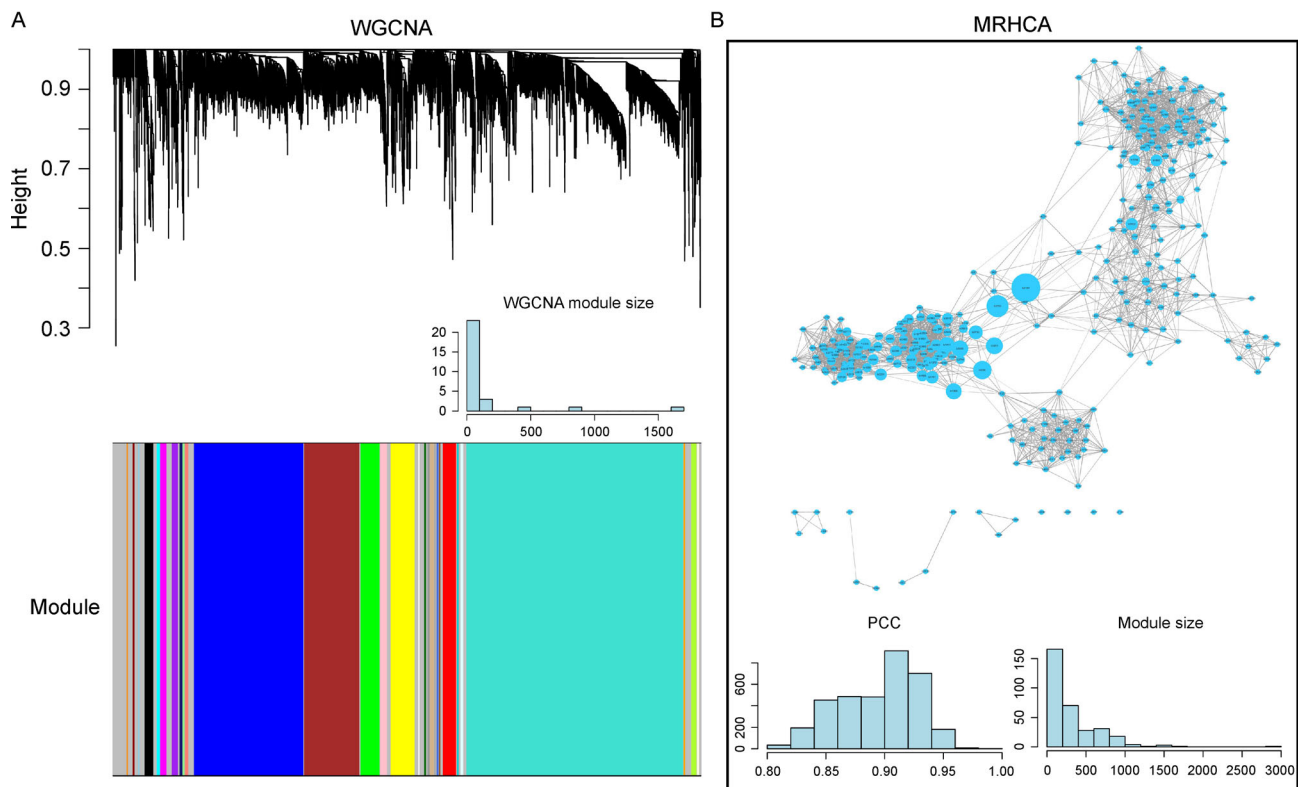


Figure 3. WGCNA and MR based co-expression module identification in *E. coli* RNA-Seq data. (A) The dendrogram shows co-expression association, and color shows the 29 modules identified by WGCNA. The histogram in the middle shows the size distribution of the modules. (B) the network plot illustrates the strong associations among the hub genes identified by MRHCA, in which the size of node reflects the size of module associated with the hub. Figure plotting parameters are given in Supplementary Figure S5. The histograms on the bottom shows: Pearson Correlation Coefficients of co-expression association between the linked hub gene pairs in the network plot (left) and size distribution of modules (right).

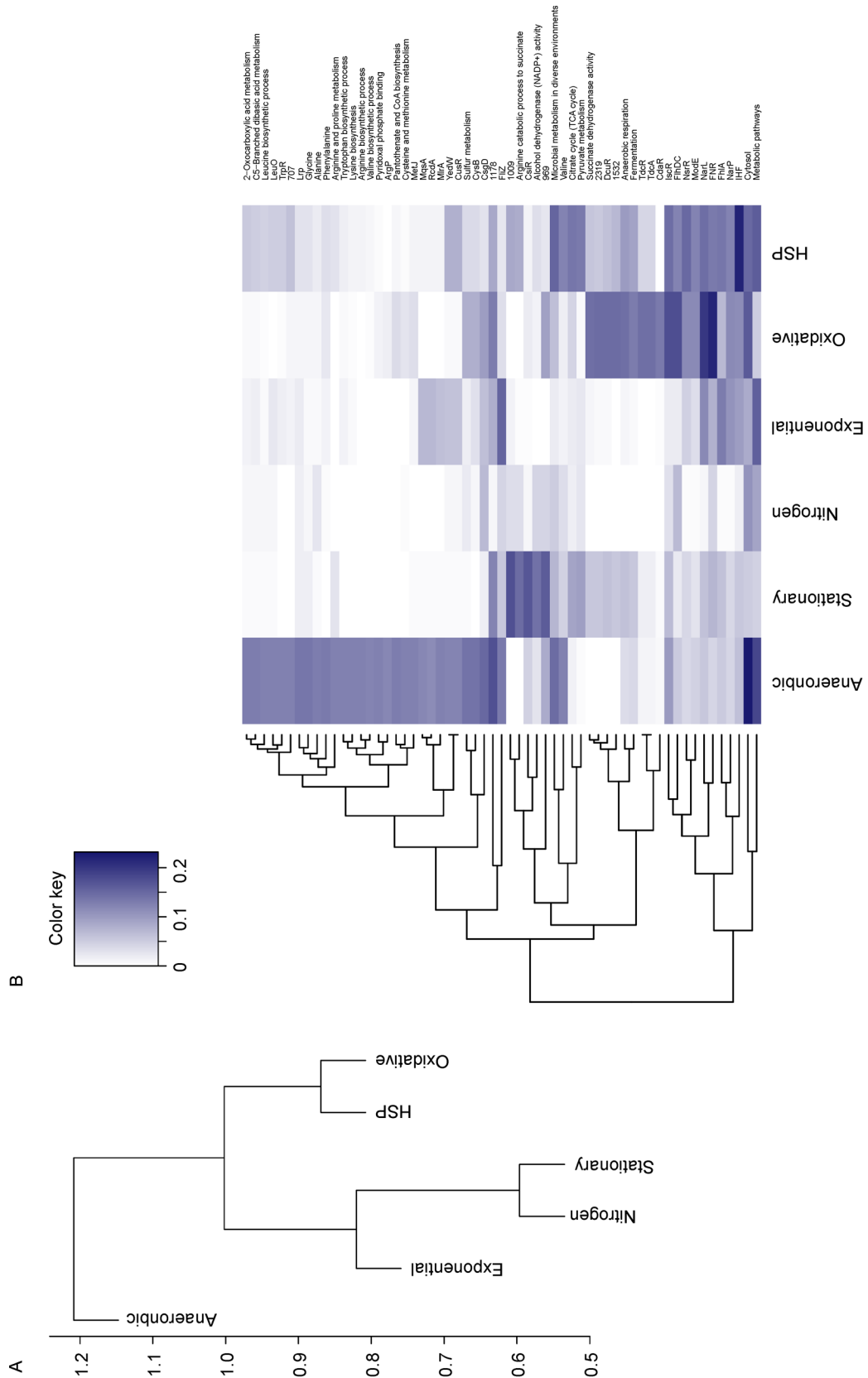


Figure 4. Differentially co-expressed modules and pathways through the six conditions. (A) Hierarchical clustering of the six conditions derived by the rate of pathways enriched in the small modules of each condition. Specifically, the rate of a pathway enriched in small modules of one condition is computed by the number of modules that are significantly enriched by this pathway divided by the total number of small modules. (B) The top 60 differentially co-expressed gene sets.

Application of MRHCA on *E. coli* transcriptomics data reveals highly overlapped co-expression modules and large number of experimental condition associated differentially co-expressed small modules

We have also compared the two methods, MRHCA vs WGCNA, on an *E. coli* RNA-Seq data set to demonstrate the strength of MRHCA in characterizing overlapping co-expression modules. Specifically, we have identified 29 non-overlapping co-expression modules by WGCNA, and 323 genes are identified as significant hubs by MRHCA. Figure 3 illustrates the distribution of the WGCNA and MR identified modules. The size of two WGCNA identified modules are large than 500, three larger than 100, and 21 ranging from 10–100. In contrast, we have seen most of the MR identified hubs are highly co-expressed with each other (see Figure 3), suggesting a significant level of overlapping among the modules centered by the hubs. Since a gene can be regulated by multiple transcriptional regulatory signals, or involved in multiple functional machineries, it is rational to observe such highly connected hubs or overlapping co-expressed modules. MRHCA enables overlapping co-expression module identification could better reflect the true co-expression network structure comparing to network partition based method.

In addition to its applicability in co-expression networks with low association values and allowing overlapping modules, MRHCA could also enable identification of small co-expression modules and their biological implications. We have applied the MR based co-expression module identification on six *E. coli* microarray data sets of different experimental conditions namely anaerobiosis (AN), exponential growth (EG), heat shock (HS), nitrogen limitation (NL), oxidative stress (OS), and stationary growth (SG) [20,21], for differential co-expression analyses. Specifically, we have identified 286 (AN), 297 (EG), 241 (HS), 232 (NL), 299 (OS), and 177 (SG) small co-expression modules with sizes ranging from 10 to 400 and computed pathway enrichments of these modules against 482 GO terms, 136 transcriptional factor regulating gene sets, 92 KEGG pathways and 116 operons, to characterize the biological functions related to each identified module. Here, the pathway enrichment analysis is based on hypergeometric test. Figure 4A shows the hierarchical clustering of the six conditions derived from the pathways enrichment of the small modules of each condition.

We have seen NL, SG and EG, and HS and OS form two different groups and are significantly distinct from AN. Since condition specific pathways are more likely to be activated under similar conditions, we posit this observation is consistent to the metabolic conditions of the six conditions: NL, SG and EG are closer to normal

metabolic condition, HS and OS are with high metabolism rate stimulated by increasing heat or oxidation, and AN is with suppressed central metabolism due to limited oxygen. Figure 4B illustrates the rate of the top 60 differentially co-expressed pathways/gene sets that are enriched in small modules of the six conditions. Particularly, we have observed the amino acid metabolic pathways are specifically enriched in the modules of AN, which is consistent with the fact that increased amino acid metabolism for pyruvate production are needed under hypoxia conditions [22,23]. We have also seen the genes regulated by nitrate/nitrite response regulators, NarL and NarP, are less co-expressed under NL condition, suggesting that nitrate/nitrite response genes are inactivated when nitrogen level is limited [24,25]. The citrate cycle and pyruvate metabolism pathways are specifically co-expressed under HS condition, explained by the fact that heat shock leads to increased central metabolism [26]. Fermentation, anaerobic respiration, succinate dehydrogenase activity and related metabolic pathways are more enriched under OS, HS and AN conditions, since central metabolism are specifically elevated or reprogrammed under these conditions [26,27]. The oxidative stress responsive proteins Ferredoxin-NADP⁽⁺⁾ reductase (FNR) are specifically co-expressed under OS condition [28]. All of these observations suggest that MRHCA could enable differential co-expression analysis with identified small modules, which is potentially a novel approach for identification of condition specific biological mechanisms.

Application of MR on TCGA data reveals distinct regulatory mechanisms of co-expression modules with different size

Considering the fact that those identified hubs by MRHCA always have relatively smaller MR values, thus we have developed a fast algorithm of the method which considers only the top K co-expressed genes with gene i in the computation of $MR(i)$. Although the method can only accurately estimate sizes for the module with sizes less than K , this proves to have $O(M^2 \log(K))$ time consumption and $O(MK)$ memory consumption. See more details in the Section of Methods. Of note, the fast algorithm can identify hub genes for all sized modules as sensitively as the original version.

We have applied the fast MR algorithm on TCGA RNA-Seq data sets of 21 cancer types by setting $K = 500$. Notably, we merge any two modules if the MR value of their hubs is smaller than 50 or the absolute value of the PCC between the two hubs is larger than 0.5. Figure 5A and 5B illustrates the distribution of sizes of the small and medium modules identified in each cancer type. For all analyzed cancer types, we have consistently observed a

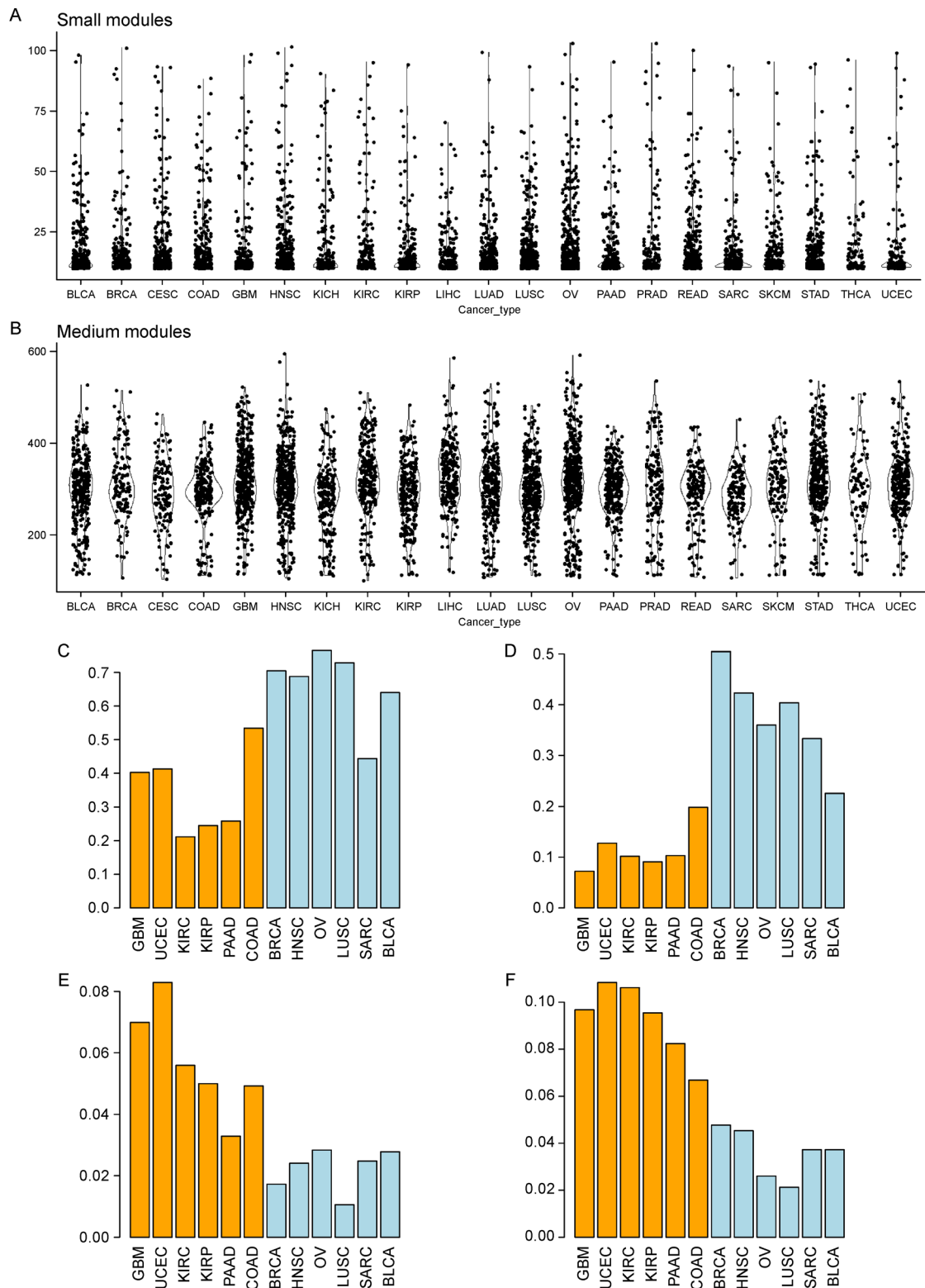


Figure 5. Small and medium sized co-expression modules in TCGA data. (A) Distributions of sizes of non-overlapping small modules. (B) Distributions of sizes of medium modules. (C) Proportions of small modules that enrich positional gene sets. (D) Proportion of medium modules that enrich positional gene sets. (E) Proportion of small modules that enrich transcription factor targets. (F) Proportion of small modules that enrich gene sets of oncogenic signatures and cancer modules.

significant number modules with less than 25 genes, as well as another group of modules with sizes ranging from 100 to 500. We term these two groups of modules as small and medium modules since there are another group of modules with more than 500 genes that cannot be fully identified by the fast algorithm. Interestingly, the identified small modules are almost non-overlapping as small modules will be merged into larger modules if the MR between their hubs is smaller than 50. For medium sized modules, we indeed observe significant levels of overlapping between modules.

In order to understand the biological implications of the small and medium sized modules, we conduct a pathway enrichment analysis for each module against six classes of gene sets retrieved from MsigDB, including positional gene sets, canonical gene sets, transcription factor targets, cancer modules, GO terms, and oncogenic signatures. We adjust false discovery rate with Bonferroni correction, and the significance cutoff is set at 0.05 [29].

Proportions of the small and medium modules enriched by the above six types of gene sets through different cancer types are compared. Interestingly, modules in cancer types with high copy number variations (CNV), namely BRCA, HNSC, OV, LUSC, and SARC, are enriched by distinct types of gene sets comparing to those cancer types with high small somatic mutations (SM), namely, COAD, GBM, KIRC, KRIP, PAAD and UCEC [30,31]. For high CNV cancer types, positional gene sets are enriched by ~70% of the small modules, and ~40% of the medium modules; for high SM cancer types, the positional gene sets are enriched by only 30% of the small modules and 10% of the medium modules (Figure 5C and 5D). In addition, medium modules of the high SM cancers enrich more gene sets of transcriptional factor targets (Figure 5E), and cancer modules and oncogenic signatures (Figure 5F). This suggests distinct transcriptional regulatory mechanisms between the two classes of cancer types. High CNV cancer types have highly varied genomes with large-scale structure variations, thus co-expression modules of these cancer types are highly associated with their genome locations. In contrast, high SM cancer types are usually highly altered in specific oncogenic pathways, hence co-expression modules in these cancer types are more likely to be affected by the downstream transcriptional regulations of the dysregulated oncogenic signals. In addition, the canonical biological pathways and GO terms are enriched more than 80% of the medium modules, while only 15% of the small modules, suggesting the medium modules are generally related to certain biological mechanisms but the small modules, which are also significantly associated with CNV or altered oncogenic signals, are not strongly associated with certain biological pathways.

DISCUSSION AND CONCLUSION

We have developed a non-parametric co-expression analysis method for large association network, namely MRHCA. The advantage of MRHCA includes: (i) it can be generally applied to association networks of most assessment methods; (ii) it outputs exact significance level for each identified hub and module; (iii) the method are with relative small computational and memory consumption, hence can be applied to large association networks; (iv) it is sensitive to the modules of weak associations; and (v) it enables overlapped modules in its outputs. We believe these capacities can bring substantial novel knowledge in gene co-expression study, especially in the field of precision medicine. R codes and a C package of the MRHCA are released by the following GitHub link: <https://github.com/zy26/mrct>.

Multiple types of omics data including genomics, transcriptomics, DNA methylation, proteomics and imaging data collected from same cancer tissues as well as clinical information in TCGA enables an integrative and systematic analysis to elucidate the crosstalk between different levels of alternations and their roles in determining cancer tissue's progression and clinical response. Co-expression analysis, not limited to gene expression data types only, is the ideal tool to identify general cancer related modules and cancer type specific modules that bring substantial mechanistic level interpretation of the dysregulated genes from multiple aspects. For example, one of the biggest challenge in cancer treatment relates to drug resistance. Using MRHCA, co-expression modules of different sizes can be comprehensively identified. By examining those genes that belong to modules enriched by known drug function related genes, we may be able to better understand the drug resistance mechanisms by functional annotation of these co-expressed markers.

However, the number of features for such integrative analyzes can easily reach $\sim 10^6$. Existing methods tend to perform a course partition of the feature space as a preliminary step, which clearly cut out potential connections of genes to modules. MRHCA is capable of detecting all the possible module hubs with statistical significance, and build up the modules form their hubs, which could handle large association networks with greatly reduced computational and memory cost. We have tested MRHCA in an independent study involves 0.8 million features, including CpG site methylation levels and exon expression data. We have successfully executed the fast version of MRHCA with $K_0=5,000$ on a 512 G RAM machine and identified strong correlations between local CpG islands methylation level and exon expressions (unpublished data). Detailed memory and computational

cost of MRHCA and the fast version are given in Supplementary Methods.

Traditional formulations of co-expression methods could be greatly biased by the way of how the strength of associations among multiple features are computed [16]. MRHCA is shown to be robust to different types of associations, which is due to the conversion of the association matrix to a non-parametric mutual rank matrix. In addition, MRHCA could predict the sizes of each hub-associated module with significance control, which does not discriminate small modules, and could allow genes to appear in multiple modules. This is a very attractive feature as one goal of gene co-expression analysis is to recover the transcriptional regulation network, so that each identified module corresponds to a group of genes that are regulated by the same transcriptional regulatory signal. It is very common for a gene to be regulated by multiple factors. In addition, our analysis has revealed strong overlaps among the co-expression modules in both prokaryotic and eukaryotic cells, and the identified small and medium sized modules are usually targets of certain transcriptional regulatory signals. Based on this, differential co-expression analysis can be directly conducted on MRHCA identified modules to detect possible alterations in transcriptional regulation under different conditions.

Our analysis suggests that MRHCA may identify strongly co-expressed genes with strong hub property from one co-expression module. One question remains not fully solved here is that how to reconstruct the module structure including identifying all the genes and sub hierarchical modules of a co-expression module. A second unsolved question is how to formulate differential co-expression analysis by using the modules identified by MRHCA. We have shown the topological structure of a large co-expression network can be visualized by the network of its hub genes and size of the modules centered by each hub (Figure 4B). However, such visualization method does not consider the overlaps among modules and functional enrichment of each module. A visualization method can handle such issues is expected. One possible solution is to use nodes to represent each module and the edges represent the Jaccard Index between each module pair. However, such a method cannot handle the dependence among multiple modules. We fully anticipate future effort can be focused on these questions, which we believe can shed light to new insights in gene co-expression and differential co-expression analysis. In addition, our methods have revealed distinct regulatory mechanisms between the cancer types of high CNV and small somatic mutations. A possible future direction is to fully understand possible regulatory mechanisms for each gene through different tumors of one cancer type, and predict patient specific regulatory signals.

METHODS

Data analyzed in this study

We have utilized seven *E. coli* datasets, one on RNA-Seq and six on microarray platforms, 21 TCGA human cancer transcriptomics data sets with RSEM normalizations. Detailed information of the analyzed data is provided in Table 1. For each dataset, we exclude those genes with 0 expression values in more than 50% samples.

Mathematical derivation of the empirical null

distribution of $\widehat{MR}(i,j)$

We have conducted a mathematical derivation for the empirical null distribution of $\widehat{MR}(i,j)$. For an $M \times M$ randomly assigned co-expression association matrix $\widehat{C}_{M \times M}$, $\widehat{C}_{M \times M}(i,j)$ is independent to $\widehat{C}_{M \times M}(x,j)$ and $\widehat{C}_{M \times M}(i,j)$ for $x \neq i, y \neq j$ by definition, hence $\text{Rank}(i \rightarrow j) \perp \text{Rank}(j \rightarrow i) | \widehat{C}_{M \times M}(i,j)$. Denote the pdf and cdf of the non-diagonal elements in $\widehat{C}_{M \times M}$ as f and F , respectively. We have the joint probability of $\text{Rank}(i \rightarrow j)$ and $\text{Rank}(j \rightarrow i)$ under given $\widehat{C}_{M \times M}(i,j)$:

$$\begin{aligned} P(\text{Rank}(i \rightarrow j) = x, \text{Rank}(j \rightarrow i) = y | \widehat{C}_{M \times M}(i,j)) = \\ \binom{M-1}{x-1} (1-F(\widehat{C}_{M \times M}(i,j)))^{x-1} F(\widehat{C}_{M \times M}(i,j))^{M-x} \\ \cdot \binom{M-1}{y-1} (1-F(\widehat{C}_{M \times M}(i,j)))^{y-1} F(\widehat{C}_{M \times M}(i,j))^{M-y}. \end{aligned}$$

Hence the joint distribution of $\text{Rank}(i \rightarrow j)$ and $\text{Rank}(j \rightarrow i)$ is:

$$\begin{aligned} P(\text{Rank}(i \rightarrow j) = x, \text{Rank}(j \rightarrow i) = y) = \\ \int (\mathbf{C}_{M-1}^{x-1} (1-F(k))^{x-1} F(k)^{M-x}) \\ \cdot (\mathbf{C}_{M-1}^{y-1} (1-F(k))^{y-1} F(k)^{M-y}) \cdot dF(k). \end{aligned}$$

This indicates that we could generate the empirical null distribution of $\widehat{MR}_{M \times M}(i,j)$ for any $i \neq j$ using the following procedures:

- (i) Sample P_{ij} from the uniform distribution $U(0,1)$;
- (ii) Sample $X_{i,j,1}, X_{i,j,2}$ from binomial distribution $\text{Binom}(n-2, p_{ij})$ independently
- (iii) Write $\widehat{MR}(i,j)' = \sqrt{(X_{i,j,1} + 1) \cdot (X_{i,j,2} + 1)}$.
- (iv) Repeating steps (i)–(iii) multiple times, the

Table 1 Detailed information of datasets used in our study

Species	Condition	Datatype	#Genes	#Samples
<i>E. coli</i>	General	RNA-Seq	4464	155
<i>E. coli</i>	Exponential growth	Microarray	4297	37
<i>E. coli</i>	Stationary	Microarray	4297	131
<i>E. coli</i>	Anaerobiosis	Microarray	4297	54
<i>E. coli</i>	Heat shock	Microarray	4297	54
<i>E. coli</i>	Nitrogen limitation	Microarray	4297	43
<i>E. coli</i>	Oxidative stress	Microarray	4297	29
Human	BLCA	RNA-Seq	17478	182
Human	BRCA	RNA-Seq	17631	994
Human	CESC	RNA-Seq	17541	145
Human	COAD	RNA-Seq	17418	233
Human	GBM	RNA-Seq	17642	169
Human	HNSC	RNA-Seq	17721	303
Human	KICH	RNA-Seq	17129	66
Human	KIRC	RNA-Seq	17661	480
Human	KIRP	RNA-Seq	17342	141
Human	LIHC	RNA-Seq	16931	134
Human	LUAD	RNA-Seq	17737	470
Human	LUSC	RNA-Seq	17967	502
Human	OV	RNA-Seq	17910	266
Human	PAAD	RNA-Seq	18012	56
Human	PRAD	RNA-Seq	17692	195
Human	READ	RNA-Seq	17531	85
Human	SARC	RNA-Seq	17000	77
Human	SKCM	RNA-Seq	17289	334
Human	STAD	RNA-Seq	17949	415
Human	THCA	RNA-Seq	17412	494
Human	UCEC	RNA-Seq	17781	370

In particular, the fourth column represents the total number of genes in the datasets after excluding the lowly expressed ones.

generated $\widehat{MR}(i,j)'$ forms an empirical null distribution of $\widehat{MR}(i,j)$.

More detailed mathematical derivation of the joint distribution of Rank($i \rightarrow j$) and Rank($j \rightarrow i$) is given in Supplementary Methods.

Analyzed co-expression association assessment methods and statistical test of the consistency between $\widetilde{MR}(i, \cdot)$ and the empirical null distribution of $\widehat{MR}(i,j)$

We have conducted Kolmogorov Smirnov test to see if the distribution of $\widetilde{MR}_{K \times K}^M(i, \cdot)$ can be accurately approximated by the empirical null distribution of $\widetilde{MR}_{K \times K}^M(i,j)$, for $K = 50, 100, 200, 500, 1,000,$ and $2,000$, for nine commonly used co-expression calculation methods.

We first generate a randomized gene expression data set $\widetilde{G}_{K \times N}$ with K genes and N samples by randomly shuffling each row of the top K highly expressed genes in *E. coli* RNA-Seq data set. Co-expression association matrix $\widetilde{C}_{K \times K}^M$ is then computed using each co-expression calculation method \mathcal{M} . Among the nine methods, DC is computed using “dcovU” function in R package “energy”, HD is computed using “hoeffd” function in R package “Hmisc”, PR and SR are computed using “adjacency.polyReg” and “adjacency.splineReg” in “WGCNA” package, MI is computed using “mi.empirical” function in “entropy” package with bin = 5, and PC, SC, KC and WC are computed using internal functions in R. $\widetilde{MR}_{K \times K}^M$ of $\widetilde{C}_{K \times K}^M$ is then computed, and $\widehat{MR}_{K \times K}^M$ is generated by random sampling from the empirical distribution of $\widetilde{MR}_{K \times K}^M(i,j), i < j$ with $\widetilde{MR}_{K \times K}^M(i,j) =$

Pseudo code of MR based method (fast version):

Input data: gene expression matrix $G_{M \times N}$ or co-expression association $C_{M \times M}$

For i in 1 to M do

Compute $R_{M \times K_0}, R_{M \times K_0}[i, j]$ is the index of the gene, whose co-expression association with i ranks j -th among all genes' co-expression association with i .

For i in 1 to ROUDS

For i in 1 to M do

Generate random number p from $U(0,1)$

Generate two independent random number X_1, X_2 from $\text{Binom}(p, M-1)$

if $(X_1 > K)X_1 = K_0$; if $(X_2 > K)X_2 = K_0$;

$MR^E(i, j) = \sqrt{(X_1 + 1)(X_2 + 1)}$ #empirical null distribution of $\widehat{MR}_{M \times M}$

Sort $MR^E(i, \cdot)$ to $MR^E_{i, (j)}$

Compute $\text{Growth}(MR^E_{i, (j)}, P)$ #empirical distribution of $\text{Growth}(MR^E_{i, (j)}, P)$

For i in 1 to M

For j in 1 to M do

if $(R_{M \times K_0}[i, k] = j) \quad R(i \rightarrow j) = k$; else $R(i \rightarrow j) = K_0$;

if $(R_{M \times K_0}[j, k] = i) \quad R(j \rightarrow i) = k$; else $R(j \rightarrow i) = K_0$;

Compute $MR_{M \times M}$ by $MR(i, j) = \sqrt{R(i \rightarrow j) \cdot R(j \rightarrow i)}$

Sort $MR(i, \cdot)$ to $MR_{i, (j)}$

Compute $\text{Growth}(MR_{i, (j)}, P)$

For i in 1 to M

Compare $\text{Growth}(MR_{i, (j)}, P)$ vs $\text{Growth}(MR^E, P)$ to assess

1) Significance level of hub property of i

$MR_{K \times K}^M(j, i)$ for $i \neq j$ and $\widehat{MR}_{K \times K}^M(i, j) = K, i = 1, \dots, K$. We further generate a $K \times K$ matrix $\widehat{MR}_{K \times K}^E$ with each element generated from the derived empirical null distribution of $\widehat{MR}_{K \times K}^M(i, j)$. KS test is then conducted to compare $\widehat{MR}_{K \times K}^M(i, \cdot)$ with $\widehat{MR}_{K \times K}^M(i, \cdot)$ and $\widehat{MR}_{K \times K}^E(i, \cdot)$ for each i . Distribution of the p values are used to evaluate if the empirical null distribution of $\widehat{MR}_{K \times K}^M$ can be utilized to approximate the distribution of $\widehat{MR}_{K \times K}^M(i, \cdot)$ for method \mathcal{M} .

MR based significance assessment for hub gene and co-expression module

To determine if a gene i is a hub and estimate the size of the module centered by the gene, we develop a significant hits score H_i computed from $MR_{K \times K}^M(i, \cdot)$ and $\widehat{MR}_{R \times K}^E$ using the following procedures:

1. Compute

$$\text{Growth}\left(\widehat{MR}_{k, (j)}^E, P\right) \triangleq \{g_{k, 1}^E, \dots, g_{k, K}^E\}, k = 1, \dots, R$$

2. Compute

$$\text{Growth}(MR_{i, (j)}^M, P) \triangleq \{g_{i, 1}, \dots, g_{i, K}\}, i = 1, \dots, R$$

3. Define

$$r_{i, j} = \begin{cases} 1, & \text{if } \frac{\sum_{k=1, \dots, R} (g_{i, j} < g_{k, j}^E)}{R} < 0.05 \\ -\text{Penalty}, & \text{if } \frac{\sum_{k=1, \dots, R} (g_{i, j} < g_{k, j}^E)}{R} \geq 0.05 \end{cases}$$

$$\text{and } H_{i, j} = \sum_{k=1}^j r_{i, j}$$

4. The significance hits score of gene i is then defined by $H_i = \max(H_{i,j})$

Gene i is determined as a significant hub if $H_i > \text{SizeCutOff}$ and the size of the co-expression module centered by i is estimated by $\text{argmax}_j(H_{i,j})$. In this study,

we set $\text{SizeCutOff} = 10$, $\text{Penalty} = 100$, the step size $P = 10$ and 50 to identify hub genes of small and general co-expression modules, respectively. The co-expression module centered by i is computed by the genes

$$\left\{ j \mid MR(i,j) < MR_{i,(j)}, J = \text{argmax}_j(H_{i,j}) \right\}.$$

Pathway enrichment analysis in *E. coli* and human data

Pathway enrichment analysis of each identified co-expression module is conducted using hypergeometric test with Bonferroni adjusted $p = 0.05$ used as significance cutoff. Modules in *E. coli* data are tested against 482 GO terms, 92 KEGG pathways, and 136 transcriptional factor regulating gene sets and 116 operons extracted from DOOR II database. Modules in human cancer data are tested against MsigDB gene sets (version 6) of six biological characteristics including 325 positional gene sets (c1), 1,329 canonical gene sets (c2), 615 transcription factor targets (c3), 430 cancer modules (c4), 5,916 GO terms (c5), and 188 oncogenic signatures (c6) [29].

A fast version of MR method for large data sets

Our MR method identifies a gene as a significant hub with a co-expression module of size K if all the first K terms in the growth of MR are significantly low comparing to the empirical distribution. It is noteworthy that, without losing much if normaiton, the algorithm's computation and memory consumption can be largely improved, if we are only to identify those modules centered by a hub gene with sizes smaller than K_0 . In this case, to compute the MR and growth of MR of gene i , only genes with top K_0 co-expression associations with i need to be identified, and the computation consumption of which is only $M \cdot \log(K_0)$, M being the total number of genes, and the largest memory consumption is a $M \times K_0$ matrix, which is used to store index of the genes with top K_0 co-expression associations with each gene i . Pseudo code of this fast version of MR method is given below.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at [10.1007/s40484-018-0131-z](https://doi.org/10.1007/s40484-018-0131-z).

ACKNOWLEDGEMENTS

This work was supported by the Showalter Young Investigator Award,

Indiana CTSI; by National Science Foundation/EPSCoR Award (No. IIA-1355423), the State of South Dakota Research Innovation Center, the Agriculture Experiment Station of South Dakota State University, and the National Science Foundation of the United States (No.1546869). This research used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant (No. ACI-1548562).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Yu Zhang, Sha Cao, Jing Zhao, Burair Alsaihati, Qin Ma and Chi Zhang declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

- Serin, E. A., Nijveen, H., Hilhorst, H. W. and Ligterink, W. (2016) Learning from co-expression networks: possibilities and challenges. *Front. Plant Sci.*, 7, 444
- Michalak, P. (2008) Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes. *Genomics*, 91, 243–248
- Obayashi, T. and Kinoshita, K. (2009) Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression. *DNA Res.*, 16, 249–260
- van Dam, S., Vösa, U., van der Graaf, A., Franke, L. and de Magalhães, J. P. (2017) Gene co-expression analysis for functional classification and gene-disease predictions. *Brief. Bioinform.*, bbw139
- Chen, J., Ma, M., Shen, N., Xi, J. J. and Tian, W. (2013) Integration of cancer gene co-expression network and metabolic network to uncover potential cancer drug targets. *J. Proteome Res.*, 12, 2354–2364
- Zhang, C., Liu, C., Cao, S. and Xu, Y. (2015) Elucidation of drivers of high-level production of lactates throughout a cancer development. *J. Mol. Cell Biol.*, 7, 267–279
- Langfelder, P. and Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9, 559
- Perkins, A. D. and Langston, M. A. (2009) Threshold selection in gene co-expression networks using spectral graph theory techniques. *BMC Bioinformatics*, 10, S4
- Ruan, J., Dean, A. K. and Zhang, W. (2010) A general co-expression network-based approach to gene expression analysis: comparison and applications. *BMC Syst. Biol.*, 4, 8
- Li, B., Zhang, Y., Yu, Y., Wang, P., Wang, Y., Wang, Z. and Wang, Y. (2015) Quantitative assessment of gene expression network module-validation methods. *Sci. Rep.*, 5, 15258
- Zhang, C. S.T., Cao, S., Xu, Y. (2016) Autophagy in Cancer Cells vs. Cancer Tissues: Two Different Stories. In *Targeting Autophagy in Cancer Therapy*. Yang, J.-M. Ed. Swedish: Springer
- Song, W. M. and Zhang, B. (2015) Multiscale embedded gene co-expression network analysis. *PLoS Comput. Biol.*, 11, e1004574
- Qin, X., Dai, W., Jiao, P., Wang, W. and Yuan, N. (2016) A multi-similarity spectral clustering method for community detection in

- dynamic networks. *Sci. Rep.*, 6, 31454
14. Okamura, Y., Aoki, Y., Obayashi, T., Tadaka, S., Ito, S., Narise, T. and Kinoshita, K. (2015) COXPRESdb in 2015: coexpression database for animal species by DNA-microarray and RNAseq-based expression data with multiple quality assessment systems. *Nucleic Acids Res.*, 43, D82–D86
 15. Song, L., Langfelder, P. and Horvath, S. (2012) Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13, 328
 16. Kumari, S., Nie, J., Chen, H. S., Ma, H., Stewart, R., Li, X., Lu, M. Z., Taylor, W. M. and Wei, H. (2012) Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PLoS One*, 7, e50411
 17. Ding, Z., Zhang, X., Sun, D. and Luo, B. (2016) Overlapping community detection based on network decomposition. *Sci. Rep.*, 6, 24115
 18. Jaenisch, R. and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, 33, 245–254
 19. Day, D. A. and Tuite, M. F. (1998) Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *J. Endocrinol.*, 157, 361–371
 20. Ma, Q., Yin, Y., Schell, M. A., Zhang, H., Li, G. and Xu, Y. (2013) Computational analyses of transcriptomic data reveal the dynamic organization of the *Escherichia coli* chromosome under different conditions. *Nucleic Acids Res.*, 41, 5594–5603
 21. Faith, J. J., Driscoll, M. E., Fusaro, V. A., Cosgrove, E. J., Hayete, B., Juhn, F. S., Schneider, S. J. and Gardner, T. S. (2008) Many Microbe Microarrays Database: uniformly normalized Affymetrix compendia with structured experimental metadata. *Nucleic Acids Res.*, 36, D866–D870
 22. Gleason, J. E., Corrigan, D. J., Cox, J. E., Reddi, A. R., McGinnis, L. A. and Culotta, V. C. (2011) Analysis of hypoxia and hypoxia-like states through metabolite profiling. *PLoS One*, 6, e24741
 23. Sengupta, S., Park, S. H., Patel, A., Carn, J., Lee, K. and Kaplan, D. L. (2010) Hypoxia and amino acid supplementation synergistically promote the osteogenesis of human mesenchymal stem cells on silk protein scaffolds. *Tissue Eng. Part A*, 16, 3623–3634
 24. Darwin, A. J. and Stewart, V. (1995) Expression of the *narX*, *narL*, *narP*, and *narQ* genes of *Escherichia coli* K-12: regulation of the regulators. *J. Bacteriol.*, 177, 3865–3869
 25. Filenko, N., Spiro, S., Browning, D. F., Squire, D., Overton, T. W., Cole, J. and Constantinidou, C. (2007) The NsrR regulon of *Escherichia coli* K-12 includes genes encoding the hybrid cluster protein and the periplasmic, respiratory nitrite reductase. *J. Bacteriol.*, 189, 4410–4417
 26. Hasan, C. M. and Shimizu, K. (2008) Effect of temperature up-shift on fermentation and metabolic characteristics in view of gene expressions in *Escherichia coli*. *Microb. Cell Fact.*, 7, 35
 27. Vemuri, G. N., Altman, E., Sangurdekar, D. P., Khodursky, A. B. and Eiteman, M. A. (2006) Overflow metabolism in *Escherichia coli* during steady-state growth: transcriptional regulation and effect of the redox ratio. *Appl. Environ. Microbiol.*, 72, 3653–3661
 28. Palatnik, J. F., Valle, E. M. and Carrillo, N. (1997) Oxidative stress causes ferredoxin-NADP⁺ reductase solubilization from the thylakoid membranes in methyl viologen-treated plants. *Plant Physiol.*, 115, 1721–1727
 29. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102, 15545–15550
 30. Zack, T. I., Schumacher, S. E., Carter, S. L., Cherniack, A. D., Saksena, G., Tabak, B., Lawrence, M. S., Zhang, C. Z., Wala, J., Mermel, C. H., *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, 45, 1134–1140
 31. Kandath, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, 502, 333–339