12-2019

Detecting Differentially Co-Expressed Gene Modules Via The Edge-Count Test

Anne Gratius Lin
University of Arkansas, Fayetteville

Citation

Lin, A. G. (2019). Detecting Differentially Co-Expressed Gene Modules Via The Edge-Count Test. *Graduate Theses and Dissertations* Retrieved from https://scholarworks.uark.edu/etd/3475

Detecting Differentially Co-Expressed Gene Modules Via The Edge-Count Test

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analytics

by

Anne Gratius Lin
University of Western Australia
Bachelor of Science in Biochemistry, 2010
Webster University
Master of Business Administration, 2016

December 2019
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Qingyang Zhang, Ph.D.
Thesis Director

Giovanni Petris, Ph.D.
Committee Member

Mark Arnold, Ph.D.
Committee Member

Jyotishka Datta, Ph.D.
Committee Member

Abstract

Background

Gene expression profiling by microarray has been used to uncover molecular variations in many different diseases. Complementary to conventional differential expression analysis, differential co-expression analysis can identify gene markers from the systematic and granular level. There are three aspects for differential co-expression network analysis, including the network global topological comparison, differential co-expression cluster identification, and differential co-expressed genes and gene pair identification. To date, most of the methods available still rely on Pearson's correlation coefficient despite its nonlinear insensitivity.

Results

Here we present an approach that is robust to nonlinearity by using the edge-count test for differential co-expression analysis. The performance of the new approach was tested with synthetic data and found to have significant results. For real data, we used a human cervical cancer data set prepared from 29 pairs of cervical tumor and matched normal tissue samples. Hierarchical cluster analysis resulted in the identification of clusters containing differentially co-expressed genes associated with the regulation of cervical cancer.

Conclusion

The proposed approach targets all different types of differential co-expression and it is sensitive to nonlinear relations. It is easy to implement and can be applied to any sequencing data to identify gene co-expression differences between multiple conditions.

Table of Contents

Chapter 1 - Introduction

Networks provide a straightforward depiction of interactions between the nodes. Intuitive network concepts (e.g. connectivity and module) have been found useful for analyzing complex interactions such as gene co-expression networks (Gov & Arga, 2017). Gene co-expression networks can be reconstructed from gene expression data using pair-wise correlation metrics that identify sets of genes that are expressed in a coordinated fashion. Altered co-expression patterns of genes between two states (for instance, healthy vs. tumor) may indicate rewiring of transcriptional networks in response to disease or adaptation to different environments. Analysis of such an alteration, also called differential co-expression analysis (de la Fuente, 2010), represents significant potential to identify gene clusters affected by state transition, and provide valuable insights on elucidation of the disease mechanisms and identification of molecular signatures of the disease (Farahbod & Pavlidis, 2018).

In general, the first step in differential co-expression analysis requires defining individual connections between genes based on correlation measures or mutual information between each pair of genes (van der Graaf, Franke, Võsa, van Dam, & de Magalhães, 2017) and a correlation cut-off given to filter the low-correlation pairs. This usually involves the assumption that genes are jointly normally distributed, i.e., there exists a linear correlation between genes, such that the hypothesis test can be written as:

$$H_0: \rho_1 = \rho_2 \text{ vs. } H_\alpha: \rho_1 \neq \rho_2$$

where ρ_1 and ρ_2 represent the true correlation coefficients between gene A and gene B in two phenotypes. Fisher's z-transformation is then employed to stabilize the variance of sample correlation coefficients in each condition, and serve as a normalizing transformation (McKenzie, Katsyv, Song, Wang, & Zhang, 2016):

$$z_1 = \frac{1}{2} log \frac{1 + \mathrm{r}_1}{1 - \mathrm{r}_1} \rightarrow N\left(\frac{1}{2} log \frac{1 + \rho_1}{1 - \rho_1}, \frac{1}{\sqrt{n_1 - 3}}\right),$$

$$z_2 = \frac{1}{2} log \frac{1 + \mathrm{r}_2}{1 - \mathrm{r}_2} \rightarrow N\left(\frac{1}{2} log \frac{1 + \rho_2}{1 - \rho_2}, \frac{1}{\sqrt{n_2 - 3}}\right),$$

where r_1 and $r_2$are the sample correlation coefficients, and n_1 and n_2 stand for the sample sizes of two phenotypes. A two-sided *p*-value is calculated using the standard normal distribution:

$$p - value = 2P\left(Z > \frac{|z_1 - z_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}} \right),$$

where *Z* represents a standard normal random variable. The false discovery rate is controlled using Benjamini-Hochberg's procedure (Benjamini & Hochberg, 1995) with an adjusted p-value threshold to determine the statistical significance of differential co-expression.

Chapter 2 – Materials & Methods

Numerous methods have been developed to detect and measure the differential co-expression of genes including methods to identify differentially co-expressed gene clusters which use newly detected gene sets or predefined set of genes. Most of these methods use Pearson's product-moment correlation which implicitly assumes linear correlation and joint normal distribution in the previously described framework (Ihmels, Bergmann, Berman, & Barkai, 2005; Kendziorski & Choi, 2009; Southworth, Owen, & Kim, 2009). However, the assumption of joint normality is over simplistic as gene expression data can strongly deviate from normality. It has been suggested that the sampling distribution of Pearson's product-moment correlation is insensitive to the effects of non-normality and is underpowered in detecting nonlinear changes in gene co-expression (Bishara & Hittner, 2012). For instance, nonlinear transformations such as taking the square root or the reciprocal of variable *x* (increases or decreases) linear relationships between variables, changes the correlation between variables and handicaps the performance of Pearson's correlation.

With the exception of Southworth et al. (2009), who used Spearman's correlation to build weighted co-expression networks, most existing methods use a targeted approach. For example, Kendziorski et al. (Kendziorski & Choi, 2009) proposed a targeted approach that focused on the analysis of modules based on known gene annotations, and used dispersion indices to test for the significance of resulting gene set co-expression changes. Though the

method has the advantage of not requiring strong correlations between gene sets, it relies on the study of known functional gene sets and is not able to identify novel, non-annotated modules or modules that would only partially match annotated categories.
On the other hand, DCA (differential clustering analysis) (Ihmels et al., 2005) is an example of a "semi-targeted" approach which uses the modules defined in one condition as a reference for the second condition. In order to avoid bias towards one of the conditions, Ihmels et al. suggested doing a reciprocal analysis, switching the reference and target conditions, while Southworth et al. used a third dataset as reference whilst applying hierarchical clustering using the difference in pair-wise correlations between both conditions as a similarity metric for two genes (Southworth et al., 2009). A drawback of this approach is the neglect of weak but significant condition-dependent correlation structures between groups of genes that might otherwise belong to distinct, strongly co-expressed and conserved clusters.

Other methods (Hyojin, Junehawk, & Seokjong, 2016; Liesecke et al., 2018) attempt to correct for Pearson's limitation by ranking correlation coefficients. For example, (Hyojin et al., 2016) successfully applied Differential Co-Expression Networks (DCENs) to identify dynamic changes in gene regulatory networks through graphical representation of the differences of co-expression correlation changes of gene pairs between conditions. Though the co-expression networks were generated using Pearson's correlation, the meta-analysis was conducted using rank-based methods. (Liesecke et al., 2018) ranked Pearson's correlation coefficients and compared them with Spearman's, and found that ranking partially corrects for the range restriction effect though the correlations were robust for high variance genes only.

Above all, weighted gene co-expression network analysis (WGCNA) is the most commonly used tool to detect differentially co-expressed clusters (Langfelder & Horvath, 2008). It constructs the co-expression network using Pearson's correlation and defines a dissimilarity measure for gene nodes. Then, average linkage hierarchical clustering is applied, coupled with the dissimilarity matrix to identify differentially co-expressed clusters. A preservation analysis can be applied to test whether clusters detected in one condition are preserved in another condition.

A comparable method DiffCoEx (Tesson, Breitling, & Jansen, 2010) provides two types of differential co-expression: within-cluster differential co-expression and cluster-to-cluster differential co-expression. They build an adjacency matrix using Pearson's correlation coefficient, and calculate the topological overlap measure to identify genes that share similar neighbors. The clusters are identified by the dissimilarity matrix. The statistical significance of differential co-expression is assessed using a statistical measure. This method can be extended to studies of more than two conditions.

As discussed, previous methods focused on identification of differentially co-expressed gene pairs, revealing many insightful biological hypotheses. However, these methods rely on Pearson's correlation which is insensitive to nonlinear changes. Therefore, the field of differential co-expression analysis would benefit from a nonlinear sensitive method for identifying differentially correlated modules. Here we present an approach for differential co-expression analysis by incorporating the edge-count test by Chen and Friedman (Chen & Friedman, 2017). We first describe the algorithm and then, to illustrate the method's effectiveness, we perform a simulation study comparing it to Pearson's correlation before presenting the results of an analysis performed on a dataset gathered from a public functional genomics data repository, the NCBI Gene Expression Omnibus (GEO).

Methods

To maneuver around the assumption of normality, we reformulate the search for differentially co-expressed genes as a nonparametric comparison between two joint distributions and use the following hypothesis:

$$H_0\colon F_1 = F_2 \; vs \; H_\alpha\colon F_1 \neq F_2 \,,$$

where F_1 and F_2 represent the joint distributions of genes A and B in two phenotypes after quantile normalization. Quantile normalization is a global adjustment method that minimizes any non-biological differences by controlling the marginal distributions and removing inter-dataset variations. Bolstad proposed a reliable non-linear method that quickly can quickly normalize within a set without choosing either a baseline to which all samples are normalized or working in a pairwise manner (Bolstad, 2001). Given *N* datasets with *p* variables (genes), form a matrix *X* of dimension *p* X *N* where each dataset is a column. We first set $d = \left(\frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}\right)$, sort each column of *X* to give X_{sort} , and then project each row of X_{sort} onto *d* to get X'_{sort} . Finally, we can get the normalized version of *X* by rearranging each column of X'_{sort} to have the same ordering as original *X*.

$$proj_{\boldsymbol{d}}\boldsymbol{q}_i = \frac{\boldsymbol{q}_i \cdot \boldsymbol{d}}{\boldsymbol{d} \cdot \boldsymbol{d}}\boldsymbol{d} = \frac{1}{\sqrt{N}}\sum_{j=1}^{N} q_{ij}\boldsymbol{d} = (\frac{1}{N}\sum_{j=1}^{N} q_{ij}\,, \ldots, \frac{1}{N}\sum_{j=1}^{N} q_{ij})$$

where $q_i = (q_{i1}, \ldots, q_{iN})$ is a row in X_{sort} and X'_{sort} is given by $q'_i = proj_{\boldsymbol{d}}\boldsymbol{q}_i$.

Normalization is achieved by taking the average of each quantile in a particular row and substituting the average value for each of the individual elements in that row (Bolstad, 2001).

One can then be sure that any significant difference between F_1 and F_2 is attributed to differential co-expression in two phenotypes (Q. Zhang, 2018).

Edge-count test

Identification of co-expression clusters starts with an adjacency matrix defined between all the genes under consideration, based on pair-wise correlations using Chen and Friedman's modified edge-count test (Chen & Friedman, 2017). Like other edge-count tests, the modified version requires a similarity graph such as the minimum spanning tree (MST) that has been constructed over the pooled samples from different groups (Q. Zhang, 2018). The reasoning is if two groups are of different distributions, samples from the same group would be inclined to be closer than those from the other group. Therefore, edges in the MST would be more likely to connect samples from the same group. The test rejects the null if the number of between-group edges is significantly less than expected. To compare two multivariate distributions, $x_1, \ldots, x_n \sim_{iid} F_X,\ y_1, \ldots, y_m \sim_{iid} F_Y$, we test H_0:

$$H_0\!: F_1 = F_2 \ vs\ H_\alpha\!: F_1 \neq F_2\,.$$

Let G be the MST constructed on pooled samples from two groups,$1,2,\dots,N=n+m$, using Kruskal's algorithm with $|G|$ denoting the number of edges, e, in G, and $|\cdot|$ representing the number of elements in the set. As illustrated from Chen and Friedman, let $g_i=0$ if sample i is from group X and $g_i=1$ otherwise. For $e=(i,j)$, define:

$$J_e=\begin{cases}0, & g_i\neq g_j\\ 1, & g_i=g_j=0\\ 2, & g_i=g_j=1\end{cases},$$

$$R_k=\sum_{e\in G} I_{J_e=k}, k=0,1,2\ ,$$

where R_0 is the number of between-group edges (which is the test statistic for the edge-count test), and $R_1$and R_2 are the numbers of edges connecting samples both from their respective groups. The new test statistic is defined as follows:

$$S=(R_1-\mu_1,R_2-\mu_2)\Sigma^{-1}\begin{pmatrix}R_1-\mu_1\\ R_2-\mu_2\end{pmatrix},$$

where $\mu_1=E(R_1)$, $\mu_2=E(R_2)$, and $\Sigma=V(R_1,R_2)^{'}$ is shown in the following lemma:

Lemma 1

$$\mu_1=|G|\frac{n(n-1)}{N(N-1)},$$

$$\mu_2=|G|\frac{m(m-1)}{N(N-1)},$$

$$\Sigma_{11}=\mu_1(1-\mu_1)+2C\frac{n(n-1)(n-2)}{N(N-1)(N-2)}+(|G|(|G|-1)-2C)\frac{n(n-1)(n-2)(n-3)}{N(N-1)(N-2)(N-3)},$$

$$\Sigma_{22}=\mu_2(1-\mu_2)+2C\frac{m(m-1)(m-2)}{N(N-1)(N-2)}+(|G|(|G|-1)-2C)\frac{m(m-1)(m-2)(m-3)}{N(N-1)(N-2)(N-3)},$$

$$\Sigma_{12}=\Sigma_{21}=(|G|(|G|-1)-2C)\frac{nm(n-1)(m-1)}{N(N-1)(N-2)(N-3)}-\mu_1\mu_2\ .$$

where $C=\frac{1}{2}\sum_{k=1}^{N}|G_k|^2-|G|$, with G_k being the subgraph in G that includes all edge(s) that connect to node k.

This algorithm can easily be extended to the study of differential co-expression over more than two conditions. The only required change is in the edge-count test, where in the case of multiple groups, a sequence of pair-wise comparisons needs to be conducted. Recently, (Q.

Zhang, 2018) extended Chen and Friedman’s test to a multiple-group case and proposed an overall test to compare more than two groups simultaneously. In the report (Zhang et al., 2017), it was proven that the test statistic for *p* groups asymptotically follows a Chi-square distribution with p degrees of freedom as N goes to infinity and the bootstrap null becomes a multivariate normal distribution. For an edge $e\ \in G,$ let

$$A_e = \{e\} \cup \{e' \in G : e' and\ e\ share\ a\ node\},$$

$$B_e = A_e \cup \{e'' \in G : \exists e' \in A_e,\ such\ that\ e'' and\ e'\ share\ a\ node\}.$$

If $|G| = O(N), \sum_{k=1}^{N} |G_k|^2 - 4|G|^2/N = O(N),\ \sum_{e\in G} |A_e|\,|B_e|$=o($N^{1.5}$)$, lim_{N\to\infty} n_i/N$=$\lambda i \in (0,1),\ then\ S \to \chi^{2}_{p}$ under the permutation null.

Chen and Friedman has proven that these conditions can be satisfied by k-MST based on Euclidean distance where the topology of G completely determines the permutation distribution of the test statistic. This facilitates the application of the test for large multi-group sample sizes where permutation computation can become progressively intensive. For small sample sizes, $min(n, m) = 20$, direct *p*-value approximation is feasible (Q. Zhang, 2018). According to Chen and Friedman (Chen & Friedman, 2017), the power of the modified edge-count test can be increased (at a computational cost) when the similarity graph becomes slightly denser for k-MST, where $k \in \{1,2,\ldots,5\}$. To the other extreme, if the similarity graph becomes too dense, it becomes difficult to distinguish edges that have similarity and edges that do not provide any “similarity” or counter information. This reduces the power of the test. For practicality, 3-MST is a reasonable initial choice as our sample sizes are in the hundreds. Our computational pipeline consists of the following steps:

Minimum Spanning Tree

Expanding on the idea of the minimum spanning tree (MST), a similarity graph such as a minimum spanning tree (MST) defines a measure of similarity between the gene expression profiles (B. Zhang & Horvath, 2005).

Let $D = \{d_i\}$ be a set of expression data with each $d_i = (e_i^1, \ldots, e_i^t)$ representing the expression levels at time 1 through time *t* of gene i . Given an edge-weighted (undirected) and connected graph $G(D) = (V, E)$ with the vertex set $V = \{d_i | d_i \in D\}$and the edge set $E = \{(d_i, d_j) | for\ d_i, d_j \in D\ and\ i \neq j\}$, a spanning tree of the graph $G(D)$ is a tree that spans $G(D)$(i.e., it includes every vertex of $G(D)$) and is a sub-graph of $G(D)$ (every edge in the tree belongs to $G(D)$). Each edge $(u, v) \in E$ has a weight that represents the distance (or dissimilarity), $\rho(u, v)$ between u and v, which could be defined as the Euclidean distance, the correlation coefficient, or some other distance measures. The cost of the spanning tree is the sum of the weights of all the edges in the tree (of which there can be many). A minimum spanning tree is the spanning tree where the cost is minimized among all spanning trees. A *k*-MST has exactly *k* vertices and forms a sub-graph of a larger graph (Knecht & Jungnickel, 2016).

Adjacency Matrix

The MST is transformed into an adjacency matrix and the associated edge list to create the un-weighted network object needed by the edge-count test. The adjacency matrix encodes the connection strength between each pair of nodes by which the column and row names are the nodes of the network. It is defined such that an entry of 1 indicates a connection between the nodes, and a 0 indicates no connection. The edge list contains all of the information necessary to create network objects, and has a minimum of two columns, one column of nodes that are the source of a connection and another column of nodes that are the target of the connection.

The false discovery rate is then controlled using Benjamini-Hochberg's method (Benjamini & Hochberg, 1995) with an adjusted p-value threshold to assess for statistical significance in differential co-expression.

Agglomerative Hierarchical Clustering

Hierarchical clustering is an unsupervised method that proceeds either by iteratively merging small clusters into larger ones (agglomerative algorithm), or by splitting large clusters into smaller ones (divisive algorithm) (Grira, Crucianu, & Boujemaa, 2004).

The agglomerative algorithm requires formation of a similarity-dissimilarity matrix in which each cell of the matrix describes the degree of similarity between the two entities. From this matrix, clusters are built by the addition of similar entities into the same cluster. At each step of the procedure, the similarity-dissimilarity matrix is recalculated in order to compute the relationship of the new clusters with the remaining entities. This generates solutions which can be graphically presented as a hierarchy of clusters or a dendrogram which can be partitioned by cutting at a desired level (K. Blashfield, 1976).

The four most popular methods of agglomerative hierarchical clustering are single linkage, complete linkage, average linkage, and the minimum variance method. It has been argued that the only difference between these methods concerns the step in which the similarity-dissimilarity relation between the new cluster and the remaining entities is computed (Johnson, 1967).

The general equation used to compute this relation is:

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \delta |d_{hi} - d_{hj}|$$

where d_{ij} refers to the euclidean distance between the entities i and j which have been joined to form the new cluster k. The relation expressed as euclidean distance between the new cluster k and the remaining entities h is denoted by d_{hk}. The designations α_j, β, and δ are parameters whose values are specified by the agglomerative hierarchical clustering procedure (K. Blashfield, 1976). We will focus on the complete linkage method as that is the method used in the hierarchical clustering procedure.

In complete linkage cluster analysis (also known as the ‘maximum method’ (Johnson, 1967) since the proximity between two clusters is the proximity between their two most distant members), a cluster is defined as a group of entities in which each member is more similar to all members of the same cluster than it is to all members of any other cluster (K. Blashfield, 1976). The values of the parameters of the general equation above are (K. Blashfield, 1976):

$$\alpha_i = \alpha_j = \frac{1}{2};\ \beta = 0; \delta = \frac{1}{2}$$

Complete linkage also has the property of being invariant under monotonic transformations of the similarity-dissimilarity matrix (K. Blashfield, 1976). While this solves the problem of chaining (where several clusters are joined together simply because one of their members is within close proximity of a member from a separate cluster), one criticism is that it is a *space-diluting* method (Johnson, 1967). This lies in the fact that an entity cannot join a cluster until it obtains a given similarity level with all members of a cluster, the probability of a cluster obtaining a new member becomes smaller as the size of the cluster increases. This increases the effective distance between the cluster and some non-member and prevents similar clusters from merging together, thus *diluting* the multi-dimensional space (K. Blashfield, 1976).

Since agglomerative hierarchical clustering does not yield a discrete number of clusters, we use the Gap statistic to determine the optimal number of clusters (or where to cut the tree). The Gap statistic is the only proposed automated method (Tibshirani, Walther, & Hastie, 2000) that is capable of accurately estimating single clusters.

Gap statistic

The Gap Statistic is constructed from the within-cluster distances and comparing their sum to the expected value under a null distribution. As noted by Tibshirani (Tibshirani et al., 2000) we have, for r clusters C_r:

$$Gap_n(k) = E_n^*[logW_k] - [logW_k],$$

where E_n^* denotes expectation under a sample size n, and, with $n_r = |C_r|$ and D_r being the sum of the pair-wise distances for all points in cluster r,

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r = \sum_{r=1}^{k} \frac{1}{2n_r} \sum_{i,i' \in C_r} d_{i,i'}$$

where distance d as the squared Euclidean distance, and W_k is the pooled within-cluster sum of squares around the cluster means. Computationally, the estimate of the optimal number of clusters is found to be:

$$\hat{k}_G = smallest\ k | Gap(k) \geq (k+1) - s_{k+1}$$

where s_k is the standard error from the estimation of $Gap(k)$. Tibshirani considers both a uniform distribution approach and a principal component construction (Tibshirani et al., 2000). In many cases, the uniform distribution performs better since it is the most likely to produce spurious clusters by the gap test (Tibshirani et al., 2000). The estimate of the optimal clusters will then be the value that maximizes the gap statistic.

Assessing true cluster structure

To ensure clusters represent true structure in the data, we use bootstrap resampling (the clusterboot() function of the package fpc in R v2.1-11.1) with the hclust() implementation (Hennig, 2007).

clusterboot‘s algorithm uses the *Jaccard coefficient*, a similarity measure between sets. The Jaccard similarity between two sets A and B is the ratio of the number of elements in the intersection of A and B over the number of elements in the union of A and B. The basic general strategy is as follows:

1. Cluster the data as usual.
2. Draw a new dataset by re-sampling the original dataset with replacement. Cluster the new dataset.
3. For every cluster in the original clustering, find the most similar cluster in the new clustering giving the maximum Jaccard coefficient. If the maximum Jaccard coefficient is less than 0.5, the original cluster is considered to be ‘dissolved’ and not a real cluster. Repeat steps 2-3 several times.

The cluster stability of each cluster in the original clustering is the mean value of its Jaccard coefficient over all the bootstrap iterations. As a rule of thumb, clusters with a stability value less than 0.6 should be considered unstable. Values between 0.6 and 0.75 indicate that the cluster is measuring a pattern in the data, but there isn’t high certainty about which points should be clustered together. Clusters with stability values above about 0.85 can be considered highly stable and are likely to be real clusters.

Materials

Simulation studies

A simulation study was conducted for two purposes:

(i) to prove superiority of the edge-count test over Pearson’s correlation,

(ii) to show that the clustering method to be used works well with the real data in determining the optimal number of clusters.

(i) edge-count test versus Pearson's Correlation

A simulation study was performed to empirically compare the edge-count test with Pearson's Correlation in a linear and nonlinear setting, where *X* and *Y* represent the expression levels of two genes and subscripts "1" and "2" stand for two conditions:

Linear setting:

$$(X_1, Y_1)^T \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right], (X_2, Y_2)^T \sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho+\Delta\\ \rho+\Delta & 1\end{pmatrix}\right],$$

where $\rho = 0$, $\Delta \in \{0.3, 0.4, 0.6, 0.8\}$.

Non-linear setting:

$$X_i \sim Unif(0,1), Y_i = \frac{2}{X_i} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_i^2), i = 1,2,$$

where $\sigma_1 = 0, \sigma_2 = \sigma_1 + \Delta, \Delta \in \{0.3, 0.4, 0.6, 0.8\}$.

For each setting, we generated data sets with sample sizes $n_1 = n_2 = 150$ and two approaches were applied to test for the difference between two joint distributions. For the edge-count test, we took 3-MST based on Euclidean distance and computed the *p*-value using the R package gTests (https://cran.r-project.org/web/packages/gTests). To evaluate the significance of Pearson's Correlation, we performed a Fisher's z-transformation introduced in Zhang (X. Zhang et al., 2012). This stabilizes the variance of sample correlation coefficients in each condition and serves as a normalizing transformation (McKenzie et al., 2016) as the transformed z_i approaches a standard normal distribution with variance $\frac{1}{n_i - 3}$:

$$z_1 = \frac{1}{2} log \frac{1+\mathrm{r}_1}{1-\mathrm{r}_1} \rightarrow N\left(\frac{1}{2} log \frac{1+\rho_1}{1-\rho_1}, \frac{1}{\sqrt{n_1 - 3}}\right),$$

$$z_2 = \frac{1}{2} log \frac{1+\mathrm{r}_2}{1-\mathrm{r}_2} \rightarrow N\left(\frac{1}{2} log \frac{1+\rho_2}{1-\rho_2}, \frac{1}{\sqrt{n_2 - 3}}\right),$$

where r_1 and r_2 are the sample correlation coefficients of (X_i, Y_i), and n_1 and n_2 stand for the sample sizes.

P-values were determined through a two-sample z-test:

$$p - value = 2P\left(Z > \frac{|z_1 - z_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}\right),$$

and adjusted via Benjamini-Hochberg's method (Benjamini & Hochberg, 1995) where a threshold of < 0.05 was used to determine the statistical significance of results:

$$H_0\text{: } F_1(X, Y) = \cdots = F_p(X, Y)\ vs\ H_\alpha\text{: } F_i(X, Y) \neq\ F_j(X, Y)\ for\ some\ i\ and\ j$$

where F_i represents the joint cumulative distribution of (X_i, Y_i). The true positive rate (accuracy) of each method is summarized in Fig 1 in Results section.

The hierarchical clustering and cluster assignment was performed using the Gap Statistic and misclassification analysis was performed on the resulting clusters using the variation of information (Meilă, 2007).

The detailed algorithm and R code used in this algorithm are given in Additional File 1.

Misclassification Error – the Variation of Information

The variation of information (VI) metric (Meilă, 2007) was used to measure the agreement of the observed cluster solutions with the actual classification determined by the procedure for generating the simulations. The VI is obtained by measuring the distance between two clusters (observed and predetermined) by obeying the triangle inequality. It is defined by:

$$VI(X, Y) = -\sum_{i,j} r_{ij}\left[\log\left(\frac{r_{ij}}{p_i}\right) + \log\left(\frac{r_{ij}}{q_j}\right)\right],$$

where X and Y are disjoint subsets of a set A, $X = \{X_1, \dots, X_k\}$ and $Y = \{Y_1, \dots, Y_l\}$, and

$$n = \sum_i |X| = \sum_j |Y| = |A|, p_i = \frac{|X_i|}{n}, q_j = \frac{|Y_j|}{n}, r_{ij} = \frac{|X_i \cap Y_j|}{n}.$$

$VI(X, Y)$ is always non-negative and runs from 0 to 1, with $VI(X, Y) = 0$ *if and only if* $X = Y$. For each cluster construction k = 2,3,4,5, we compute the VI metric for each correlation setting ($\sigma_X = 0.3, 0.4, 0.6, 0.8$) of 150 samples, and for correlation setting, $\sigma_X = 0.6$, with sample sizes of $n = 50, 75, 100, 200$. These are represented in Figure 2. in Results.

Real Data - **Gene expression dataset**

A human cervical cancer data set prepared from 29 pairs of cervical tumor and matched normal tissue was used to illustrate our proposed computational pipeline. Of the twenty-nine cases with paired specimens, 21 patients had a diagnosis of squamous cell carcinoma, six had adenocarcinoma and two had an intermediate diagnosis of adenosquamous cell carcinoma. The data were restricted to a set of the 375 most varying genes, and log-transformed for further processing. The raw data was obtained from the NCBI Gene Expression Omnibus (GEO) (Witten, Tibshirani, Gu, Fire, & Lui, 2010)(Tibshirani), a public functional genomics data repository.

Quantile normalization was performed (via normalize.quantiles() of the package preprocessCore in R v1.34.0) for each group so as to match the marginal distributions of the genes across groups (Figure 3). The purpose of quantile normalization is to avoid the rejection of H_0 due to marginal difference (differential expression) instead of different dependency patterns (differential co-expression).

For each gene pair, the inter-sample distances and an edge-count test with 3-MST was implemented, followed by a BH procedure with FDR<0.05 for multiplicity adjustment. With the edge list from the gene pairs forming a network, we use the Gap statistic to determine the optimal number of clusters Figure 4. Figure 5 depicts the overall network clusters.

Chapter 3 - Results

We present results of simulations for a range of correlations to test for the true positive rate of each method under both settings.

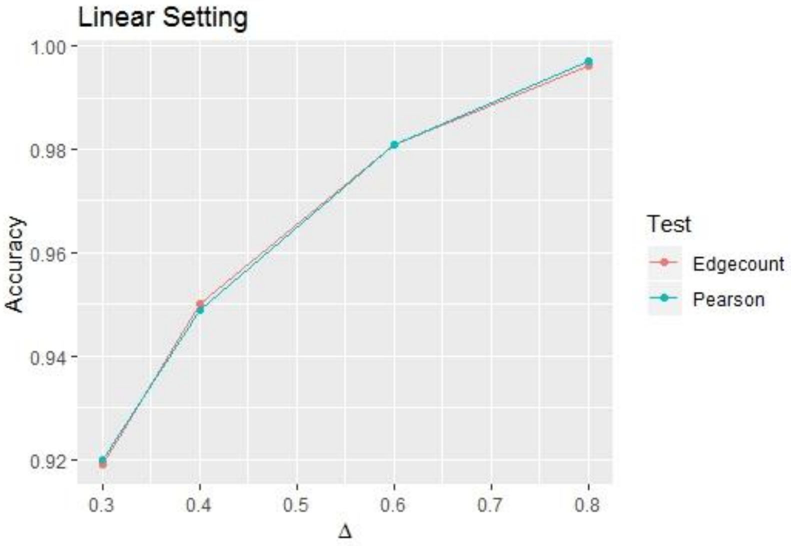

Figure 1A. Performance comparison of two methods under a linear setting. The x-axis is the value of Δ, and y-axis is the true positive rate.

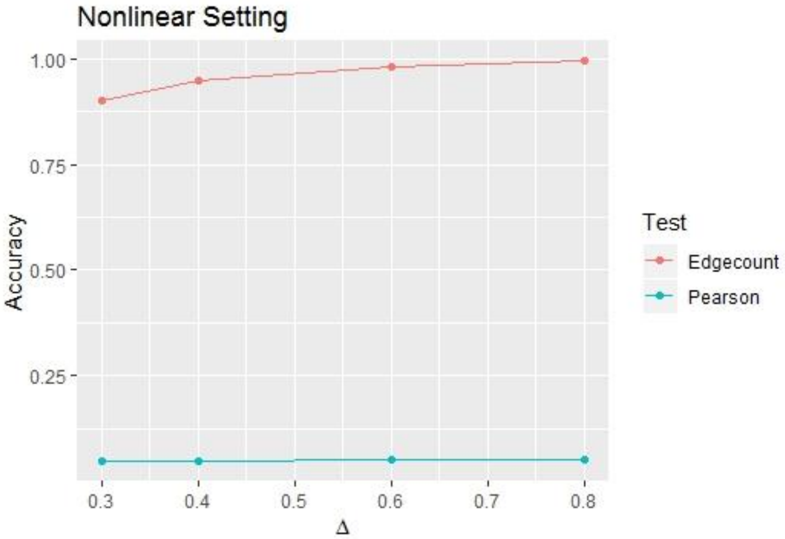

Figure 1B. Performance comparison of two methods under a nonlinear setting. The x-axis is the value of Δ, and y-axis is the true positive rate.

As seen from Fig 1, the two methods achieved comparable accuracy in the linear setting. For the non-linear (inverse transformation) setting, the edge-count test substantially outperforms the Pearson's method which fails to identify any differences. This underlines the strength of the edge-count test, not only in capturing linear changes but demonstrating significantly better sensitivity for nonlinear settings.

Comparing Clusters

We evaluate the misclassification error of the clustering algorithm by using the information metric: the variation of information (Figure 2). The results are presented on a $k = 2, 3, 4, 5$ cluster construction with each cluster comprised of $n = 150$ samples for the correlation setting $\{0.3, 0.4, 0.6, 0.8\}$, and $n = 50, 75, 100, 200$ for the correlation setting 0.6.

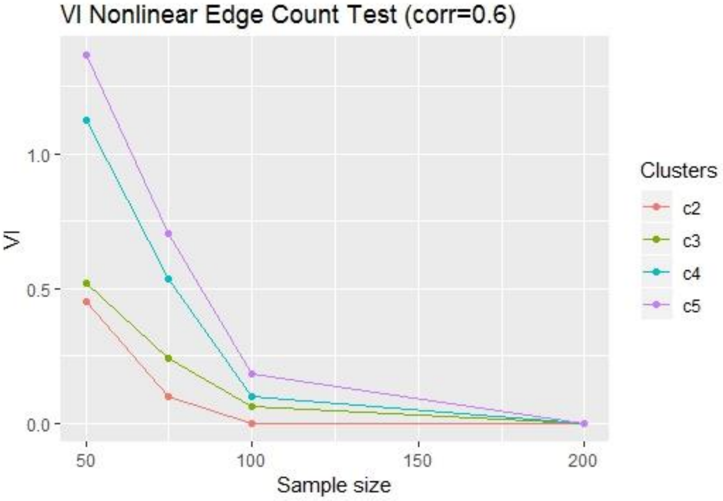

Figure 2A.Variation of Information for Nonlinear edge-count Test with fixed correlation 0.6.

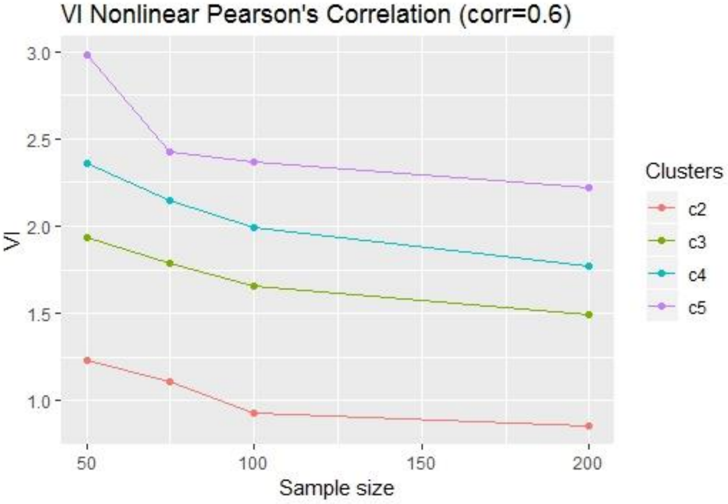

Figure 2B. Variation of Information for Nonlinear Pearson's Correlation with fixed correlation 0.6.

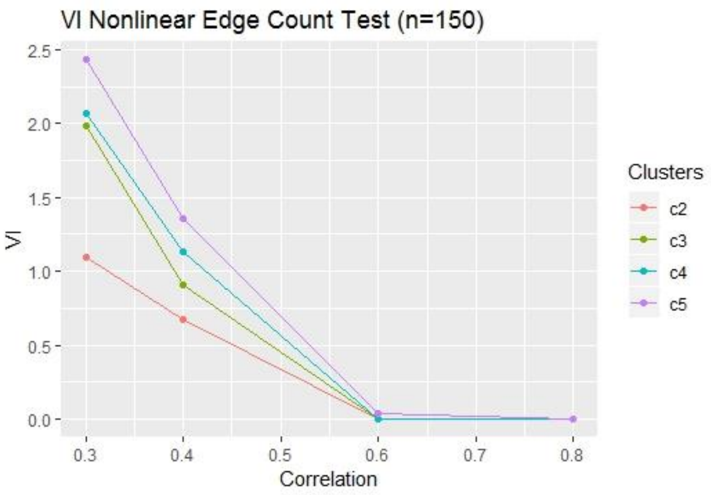

Figure 2C. Variation of Information for Nonlinear edge-count Test with fixed sample size 150.

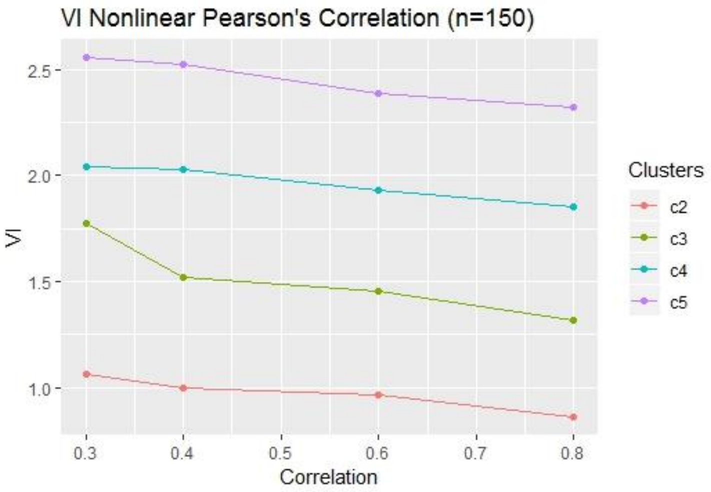

Figure 2D. Variation of Information for Pearson's correlation with fixed sample size 150.

As we can see in Figure 2, the VI metric agrees with the nonlinear sensitivity of the edge-count test over Pearson's correlation in Figure 1. The misclassification error for the edge-count test has the largest decrease for five clusters by 1.367 for fixed correlation of 0.6, and a decrease of 2.433 for fixed sample size of 150. The misclassification error for the Pearson's correlation has the largest decrease for five clusters by 0.756 for fixed correlation of 0.6, and a decrease of 0.456 for three clusters for fixed sample size of 150.

Real Data

Agglomerative hierarchical clustering is used to group genes that have a similar expression pattern in multiple samples. The resulting modules often represent biological processes and can be phenotype specific. In order to assess the difference between normal and tumor samples, we performed an unsupervised hierarchical clustering of the samples using the complete linkage method (previously describe in Chapter 2. The results of the analysis are summarized in Figure 5. We identified a total of 4 differentially co-expressed clusters comprising a total of 10135 gene pairs. From the dendrogram, we can see there are clusters that are significantly more correlated than one would expect due to chance. This suggests that these genes may be controlled by common regulatory factor(s). The data are consistent with several previous findings (Witten et al., 2010).

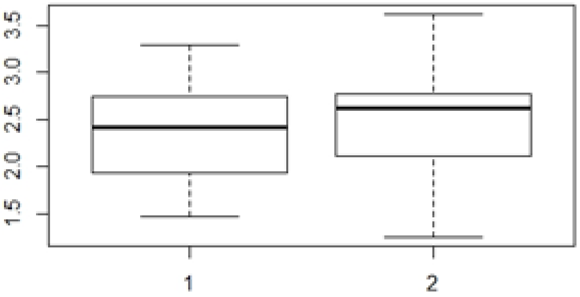

Figure 3A. Distributions before Quantile Normalization. "1" = Normal, "2" = Tumor.

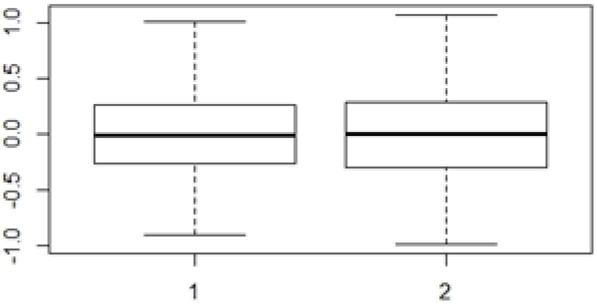

Figure 3B. Distributions after Quantile Normalizaton. "1" = Normal, "2" = Tumor.

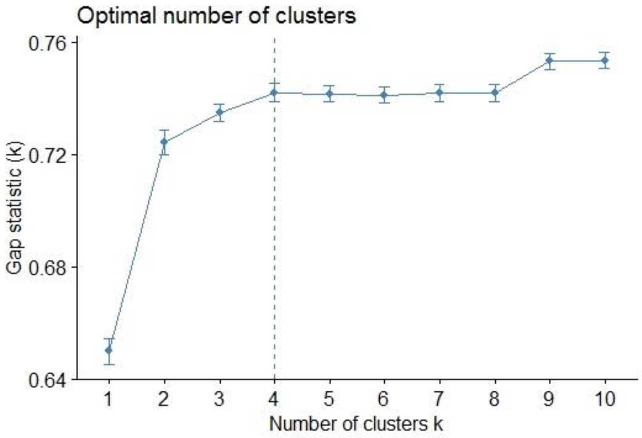

Figure 4. Optimal number of clusters via the Gap Statistic

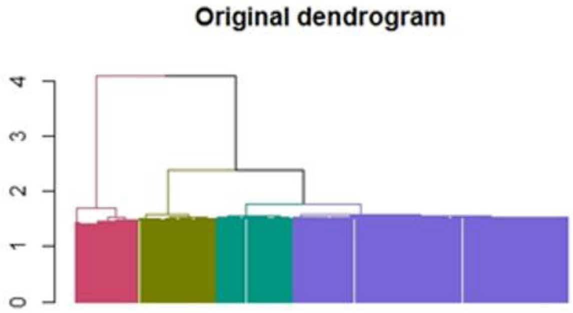

Figure 5A. Clustering analyses of differentially co-expressed genes – Dendrogram.

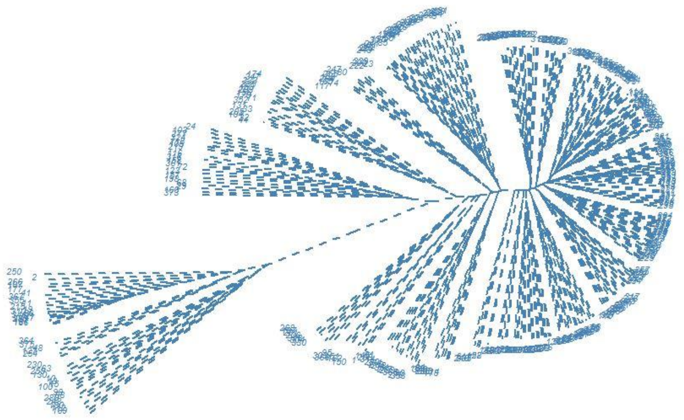

Figure 5B. Clustering analyses of differentially co-expressed genes – Unrooted Dendrogram

As shown in Figure 5B, the clustering analysis resulted in the identification of two major subgroups that show an almost perfect separation between genes in normal and tumor samples. The Gap Statistic optimized the number of clusters to be 4 (Figure 4) with cluster sizes of 72, 32, 250 and 21 with Jaccard coefficient of 0.7105, 0.6346, 0.9202 and 0.8058 respectively, indicating high stability of cluster formation.

Of these 4 clusters exhibiting significant differential co-expression ($P < 0.05$), *miR-200b~429*, *miR-34b~34c*, *miR-503~424*, *miR-29c~29b*, *miR-15b~16*, *miR-200c~141, miR-99b~125a* and *miR-25~106b* were identified. Seventeen of the sub-clusters were associated with gene up-regulation in cervical cancer and 13 sub-clusters were associated with gene down-regulation in cancer. Interestingly, the two clusters that are most associated with the cervical cancer versus normal class labels both belong to the *miR-200* family.

Chapter 4 - Discussion

Hierarchical clustering, a classic clustering method commonly used by clinical researchers, was used primarily due to its consistency of results in the simulation data and its ease of use, as it requires the setting of few parameters. Hierarchical clustering is also available in standard gene expression databases, such as the Gene Expression Omnibus from which the data was obtained (de Souto, Costa, de Araujo, Ludermir, & Schliep, 2008).

Other clustering methods such as *k*-means (McQueen, 1967), mixture of multivariate Gaussians (McLachlan, Bean, & Peel, 2002), spectral clustering (Ng, Jordan, & Weiss, 2002) and nearest neighbor-based methods (Ertoz, Steinbach, & Kumar, 2002) have also been used in analyzing gene expression data. However, common limitations of these new methods include the requirement of using particular programming environments and the specification of a number of different parameters, which makes their implementation difficult for non-expert users.

Differential co-expression via the edge-count test provides information that would be missed using classical methods focusing on the identification of differentially expressed genes. The algorithm presented has the advantage of comparing two (or more) datasets in a global, unbiased and unsupervised manner. It represents a major improvement over earlier comparisons due to its nonlinear sensitivity. We demonstrate an example in the simulation study where differential co-expression patterns were uncovered using the edge-count test but that were missed by Pearson's correlation. As seen from Figure 2B & 2C, non-linear transformations away from normality greatly reduce the absolute magnitude of Pearson's correlation and inflate the error rates.

A fundamental advantage of using the edge-count test is that it requires no model assumptions and is an efficient approach. This is useful in our case as differential co-expression may be caused by different biological mechanisms. For example, a group of genes may be under the control of a common regulator (e.g. a transcription factor or epigenetic modification) that is active in one condition, but absent in the other condition. In such a case, the correlation structure induced by variation in the common regulator would only be present in the first condition. Another possible interpretation relates to the presence or absence of variation in some factors driving a gene cluster. To observe correlation of a group of genes responding to a common factor, this factor needs to vary. In the absence of variation of the driving factor, no correlation can be observed, even though the actual biological links that form the network are not altered. It is therefore important to ensure that the perturbations which give rise to variation within each condition are: (i) biologically relevant (as opposed to batch effects, for example) and (ii) comparable in nature and amplitude.

However, a drawback of the edge-count test is that it can be computationally and time intensive due to the calculation of the minimum spanning tree. Moreover, it only works well with genes that are highly differentially correlated or with large sample sizes. We see evidence of this in Figure 2 where there are sharp increases in misclassification error for samples less than 100 (Fig. 2.a.) and for correlation less than 0.5 (Fig. 2.b.). Therefore, more research is required to adjust for this.

Our algorithm constitutes a valuable tool of broad applicability in studying gene regulatory networks or performing exploratory data analysis. This approach illustrates the high value of the proposed test not only in quantitative analysis of DCE genes but is also broadly applicable to the analysis of any large scale gene expression data. Future possibilities for this algorithm include implementing the new pipeline into software tools such as R package, giving a competitive edge over other algorithms such as WGCNA (Langfelder & Horvath, 2008).

Chapter 5 - Conclusion

Differential co-expression may be caused by different biological mechanisms. For example, a group of genes may be under the control of a common regulator (e.g. a transcription factor or epigenetic modification) that is active in one condition, but absent in the other condition. In such a case, the correlation structure induced by variation in the common regulator would only be present in the first condition. Another possible interpretation relates to the presence or absence of variation in some factors driving a gene module. To observe correlation of a group of genes responding to a common factor, this factor needs to vary. In the absence of variation of the driving factor, no correlation can be observed, even though the actual biological links that form the network are not altered. It is therefore important to ensure that the perturbations which give rise to variation within each condition are biologically relevant and comparable in nature and amplitude.

Motivated by the fact that these perturbations are generally non linear, the algorithm presented provides a tailored approach in studying how different sample groups respond. A fundamental advantage of this algorithm is that it requires no model assumptions and though it requires a series of intermediate steps, is a tailored approach through a powerful graph-based test. This constitutes a valuable tool of broad applicability in gene regulatory network analysis.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Bishara, A. J., & Hittner, J. B. (2012). Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychol Methods, 17*(3), 399-417. doi:10.1037/a0028087

Bolstad, B. (2001). Probe level quantile normalization of high density oligonucleotide array data. *Unpublished manuscript*.

Chen, H., & Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association, 112*(517), 397-409.

de la Fuente, A. (2010). From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends Genet, 26*(7), 326-333. doi:10.1016/j.tig.2010.05.001

de Souto, M. C., Costa, I. G., de Araujo, D. S., Ludermir, T. B., & Schliep, A. (2008). Clustering cancer gene expression data: a comparative study. *BMC Bioinformatics, 9*(1), 497. doi:10.1186/1471-2105-9-497

Ertoz, L., Steinbach, M., & Kumar, V. (2002). *A new shared nearest neighbor clustering algorithm and its applications.* Paper presented at the Workshop on clustering high dimensional data and its applications at 2nd SIAM international conference on data mining.

Farahbod, M., & Pavlidis, P. (2018). Differential coexpression in human tissues and the confounding effect of mean expression levels. *Bioinformatics, 35*(1), 55-61. doi:10.1093/bioinformatics/bty538

Gov, E., & Arga, K. Y. (2017). Differential co-expression analysis reveals a novel prognostic gene module in ovarian cancer. *Scientific Reports, 7*(1), 4996. doi:10.1038/s41598-017-05298-w

Grira, N., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content, 1*, 9-16.

Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis, 52*(1), 258-271.

Hyojin, K., Junehawk, L., & Seokjong, Y. (2016, 15-18 Dec. 2016). *Differential Co-Expression Networks using RNA-seq and microarrays in Alzheimer's disease.* Paper presented at the 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM).

Ihmels, J., Bergmann, S., Berman, J., & Barkai, N. (2005). Comparative gene expression analysis by a differential clustering approach: application to the Candida albicans transcription program. *PLoS genetics, 1*(3), e39.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*(3), 241-254.

K. Blashfield, R. (1976). *Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods* (Vol. 83).

Kendziorski, C., & Choi, Y. (2009). Statistical methods for gene set co-expression analysis. *Bioinformatics, 25*(21), 2780-2786. doi:10.1093/bioinformatics/btp502

Knecht, T., & Jungnickel, D. (2016). A note on the k-minimum spanning tree problem on circles. *Operations Research Letters, 44*(2), 199-201. doi:https://doi.org/10.1016/j.orl.2015.12.019

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics, 9*(1), 559. doi:10.1186/1471-2105-9-559

Liesecke, F., Daudu, D., Dugé de Bernonville, R., Besseau, S., Clastre, M., Courdavault, V., . . . Dugé de Bernonville, T. (2018). Ranking genome-wide correlation measurements improves microarray and RNA-seq based global and targeted co-expression networks. *Scientific Reports, 8*(1), 10885. doi:10.1038/s41598-018-29077-3

McKenzie, A. T., Katsyv, I., Song, W.-M., Wang, M., & Zhang, B. (2016). DGCA: a comprehensive R package for differential gene correlation analysis. *BMC systems biology, 10*(1), 106.

McLachlan, G. J., Bean, R. W., & Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics, 18*(3), 413-422.

McQueen, J. (1967). Some methods for classification and analysis of multivariate observations, paper presented at the 5th Berkeley Symposium on Mathematics, Statistics, and Probability. *Univ. of Calif., Berkeley*.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis, 98*(5), 873-895. doi:https://doi.org/10.1016/j.jmva.2006.11.013

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). *On spectral clustering: Analysis and an algorithm.* Paper presented at the Advances in neural information processing systems.

Southworth, L. K., Owen, A. B., & Kim, S. K. (2009). Aging mice show a decreasing correlation of gene expression within genetic modules. *PLoS genetics, 5*(12), e1000776-e1000776. doi:10.1371/journal.pgen.1000776

Tesson, B. M., Breitling, R., & Jansen, R. C. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. *BMC Bioinformatics, 11*(1), 497. doi:10.1186/1471-2105-11-497

Tibshirani, R., Walther, G., & Hastie, T. (2000). {E}stimating the {N}umber of {C}lusters in a {D}ataset via the {G}ap {S}tatistic. *Journal of the Royal Statistical Society, Series B, 63*, 411-423. doi:citeulike-article-id:3989914

van der Graaf, A., Franke, L., Võsa, U., van Dam, S., & de Magalhães, J. P. (2017). Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics, 19*(4), 575-592. doi:10.1093/bib/bbw139

Witten, D., Tibshirani, R., Gu, S. G., Fire, A., & Lui, W.-O. (2010). Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biology, 8*(1), 58. doi:10.1186/1741-7007-8-58

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology, 4*(1).

Zhang, Q. (2018). A powerful nonparametric method for detecting differentially co-expressed genes: distance correlation screening and edge-count test. *BMC systems biology, 12*(1), 58.

Zhang, X., Zhao, X. M., He, K., Lu, L., Cao, Y., Liu, J., . . . Chen, L. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics, 28*(1), 98-104. doi:10.1093/bioinformatics/btr626