
Article:
Cheng, CW orcid.org/0000-0002-2873-0828, Beech, DJ orcid.org/0000-0002-7683-9422 and Wheatcroft, SB orcid.org/0000-0002-6741-9012 (2020) Advantages of CEMiTool for gene co-expression analysis of RNA-seq data. Computers in Biology and Medicine, 125. 103975. ISSN 0010-4825

https://doi.org/10.1016/j.compbiomed.2020.103975

Advantages of CEMiTool for gene co-expression analysis of RNA-seq data

Chew Weng Cheng* PhD, David J. Beech PhD, Stephen B. Wheatcroft MD PhD

Discovery and Translational Science Department, Leeds Institute of Cardiovascular and Metabolic Medicine, Faculty of Medicine and Health, University of Leeds, Leeds, LS2 9JT.

Email: C.W.Cheng@leeds.ac.uk; D.J.Beech@leeds.ac.uk; S.B.Wheatcroft@leeds.ac.uk

***Corresponding author:**
Chew Weng Cheng, PhD
Discovery and Translational Science Department,
Leeds Institute of Cardiovascular and Metabolic Medicine,
Faculty of Medicine and Health,
University of Leeds, Leeds, LS2 9JT.
Email: C.W.Cheng@leeds.ac.uk

Abstract

Gene co-expression analysis is widely applied to transcriptomics data to associate clusters of genes with biological functions or identify therapeutic targets in diseases. Recently, the emergence of high-throughput technologies for gene expression analyses allows researchers to establish connections through gene co-expression analysis to identify clinical disease markers. However, gene co-expression analysis is complex and may be a daunting task. Here, we evaluate three co-expression analysis packages (*WGCNA, CEMiTool,* and *coseq*) using published RNA-seq datasets derived from ischemic cardiomyopathy and chronic obstructive pulmonary disease. Results show that the packages produced consensus co-expression clusters using default parameters. *CEMiTool* package outperformed the other two packages and required less computational resource and bioinformatics experience. This evaluation provides a basis on which data analysts can select bioinformatics tools for gene co-expression analysis.

Keywords: RNA-seq; co-expression; ischemic cardiomyopathy; Chronic obstructive pulmonary disease; WGCNA; CEMiTool; coseq

1.0 Introduction

With the emergence of next-generation sequencing technologies, many researchers have needed to interrogate gene expression patterns to answer their research questions. However, many genes have unknown functions, which makes interpretation challenging. **Co-expression analysis** (see Box 1 for an explanation of all special terms used in gene co-expression analysis) performed systematically can help alleviate this challenge. Co-expression analysis uses global gene expression levels to cluster genes into several modules based on the correlation estimation. The derived modules allow researchers to understand how genes are interacting with one another and to predict their possible roles, because genes with similar biological functions tend to exhibit a strong correlation in expression levels [1, 2]. Gene expression levels can be derived from RNA-seq or microarray data sets. Following identification of gene modules, gene set enrichment analysis is usually performed to uncover if the gene set is enriched for biological pathways or functions. In addition, the gene modules identified can be further used to correlate with clinical traits to reveal potential associations. Finally, gene co-expression analysis is a robust approach to investigate the functional roles of a gene.

A fundamental objective of RNA-seq analysis is to identify genes that are significantly up- or down-regulated between conditions. However, differentially expressed genes with unknown functions or a low number of hits are challenging to interpret. Thus, gene co-expression analysis is thought to be a plausible approach to RNA-seq data to identify functional genes. However, it is unclear whether all available packages have equivalent performance to derive biological conclusions. Besides, co-expression analysis considers the variability in gene expression levels and that improving the accuracy of the co-expression networks using RNA-seq data which may be confounded by technical variation [3]. Gene co-expression analysis has been used for many studies, including antigen discovery [4], identification of regulatory genes [5-9], and functional classification of genes [10, 11].

Numerous gene co-expression analysis packages are available to researchers, including Weighted gene co-expression network analysis (*WGCNA*) [12], co-expression modules identification tool (*CEMiTool*) [13], co-expression of RNA-seq data (*coseq*) [14, 15], *petal* [16], Co-expressed biological processes (*CoP*) [17], and *CoXpress* [18]. Currently, a comprehensive study in which these packages are compared is lacking. As the workflow of co-expression analysis is complex and time-consuming, it can appear as a daunting approach to users with little bioinformatics experience. In this study, three main Bioconductor packages for gene co-expression analysis are discussed: *WGCNA*, *coseq*, and *CEMiTool.* These packages are freely available to the research community, are user-friendly and well-maintained. Three real RNA-seq data sets of ischemic cardiomyopathy (ICM) and chronic obstructive pulmonary disease (COPD) were downloaded to evaluate the performance of these three packages. The ICM dataset contained 28 healthy controls and 29 ICM patients whereas the COPD dataset consisted 91 normal spirometry controls and 98 COPD patients.

Gene co-expression analysis begins with the identification of the association between genes through correlation estimation. The association represents the similarity between expression levels of the genes across samples. Two correlation measurements are commonly applied to gene co-expression construction, *Pearson's* and *Spearman's* correlations. Subsequently, groups of co-expressed genes are clustered based on several methods, including hierarchical clustering and *K*-means clustering, which are discussed in detail elsewhere [19]. *WGCNA* and *CEMiTool* are both based on hierarchical clustering, while *coseq* uses *K*-means clustering. *WGCNA* and *CEMiTool* packages are similar in principle, but the latter provides an automated pipeline. *CEMiTool* is more efficient in co-expression analysis with its automated pipeline - and users do not require extensive bioinformatics experience. On the contrary, *WGCNA* involves complex workflows, requiring users to have extensive bioinformatics skills and higher computational power. *coseq* proposes to offer flexibility in identifying groups of co-expressed genes due to its clustering and transformation models [14]. Currently, however, the consensus outputs by these clustering methods remain to be discussed.

Although multiple gene co-expression analysis packages have been developed, there remains a lack of independent comparison of these packages; especially those available in R. Russo and colleagues [13] performed an extensive evaluation between *WGCNA* and *CEMiTool* with at least 1000 sample size but did not compare the performance with other clustering methods, such as *K*-means clustering in *coseq*. The *K*-means algorithm is suggested to produce highly specific modular genes [15]. To this end, we applied real data from ICM and COPD patients to compare three main packages, *WGCNA*, *CEMiTool*, and *coseq*, for gene co-expression analysis and fill the gap in choosing an appropriate package based on the research aim and resources available. This study will aid researchers in choosing an appropriate package for gene co-expression analysis, taking into consideration the research aims, computational resources, and bioinformatics skills.

Box 1. Explanation of all special terms used in gene co-expression analysis.

Batch effect: This is an artificial source of variation introduced by different sequencing platforms, sample source, and experimental design.

Beta value (b): This is a number represents the power at which the gene co-expression modules achieve scale-free topology. This is also known as the soft-threshold value.

Centered Log Ratio (CLR): It is a transformation method for genes expression profiles in *coseq*. This transformation considers the geometric mean of the genes and therefore, able to identify the small differences in genes with homogeneous expression levels across conditions.

Co-expression analysis: This is an approach to cluster genes with highly correlated expression levels into multiple modules.

Connectivity: This value indicates the connection strength between a gene and other genes in the network.

Differentially expressed genes (DEGs): These are the genes that achieved difference statistically in their expression levels between two conditions. The level of significance is usually determined by p-value below 0.05.

Eigengene: The is a weighted mean value of the expression levels of all genes within a gene co-expression module.

Gene Ontology (GO): This is an algorithm to annotate the biological functions of a group of genes.

Hierarchical clustering: This is a method that uses an agglomerative approach where genes with similar expression levels are merged into clades.

***K*-means clustering:** This is a clustering method where genes are assigned to the cluster based on the minimum distance to the cluster mean (centroid). The number of clusters is determined by *k* value.

Module: Modules consist of clusters with many interconnected genes.

Scale-free topology (SFT): This is a degree distribution of which the network follows power-law.

Slope heuristics approach: This is a model selection algorithm to determine the number of co-expressed gene clusters in *coseq*. The algorithm uses the asymptotic penalised likelihood criterion for model selection, and the lowest penalty score defines the best possible number of clusters.

Topological overlap matrix (TOM): This defines how well the genes are connected and derive the weighted networks. TOM considers the adjacency and strength of connectedness between genes and their neighbours.

Variance-stabilizing transformation (VST): This is a method to fit the mean-variance relationships to the transformed data, making expression levels the more homoscedastic.

2.0 Methods

2.1 *Ischemic cardiomyopathy RNA-seq dataset*

For a practical comparison, two real ICM RNA-seq datasets were used. RNA-seq studies of ischemic cardiomyopathy (ICM) were downloaded from the Gene Expression Omnibus (GEO) database (GSE48166 and GSE116250). In total, 28 ICM samples and 29 non-heart failure (NF) samples were used to compare three gene co-expression analysis packages. ICM samples were obtained from the left ventricle of patients undergoing heart cardiac transplantation, whilst patients with no major cardiac history served as NF group [20, 21].

2.2 *Chronic obstructive pulmonary disease dataset*

To increase robustness of the study, we downloaded another RNA-seq dataset derived from lung tissues of chronic obstructive pulmonary disease (COPD) patients. The dataset was available from GEO with accession number GSE57148. The dataset included 91 control individuals and 98 COPD patients from a Korean population [22]. The total RNA was isolated from fresh frozen lung tissue that was remote from the lung cancer.

2.3 *Bioinformatics processing*

Raw fastq files obtained from GEO were subjected to initial quality assessment using *FASTQC* [23]. All the outputs from *FASTQC* were merged using *MULTIQC* [24]. The raw fastqc files were trimmed and filtered using *TrimGalore* [25], raw reads below 20 bases long and bases quality below 20 were removed. Subsequently, all the cleaned reads were aligned to the *Homo sapiens* genome Ensembl GRCh38.p13 using *STAR* version 2.7 [26] with the following alignment parameters: --sjdbOverhang 100 --outFilterScoreMinOverLread 0 --outFilterMatchNminOverLread 0 --outFilterMatchNmin 0. Reads alignment were performed on

a single-end mode. The mapped reads were quantified using *featureCounts* release 1.6.5 [27]. Gene expression levels were expressed as raw read counts.

2.4 *Differential gene expression analysis*

Raw read counts were processed using *DESeq2* version 3.9 [28]. Pre-filtering was applied to the data sets to filter low read counts; genes with less than 10 reads were removed from the analysis. Since the RNA-seq data were downloaded from two independent studies, batch effect correction was applied prior to read counts normalisation in *DESeq2*. The read counts across the samples were normalised using the DESeq method, based on median ratio of gene counts. Subsequently, variance stabilizing transformation estimated from the fitted dispersion-mean relation was applied to obtain normalised gene expression values. Principal component analysis (PCA) plots were used to assess the quality of the data. For the ICM dataset, two PCA plots were constructed before and after batch effect removal as this dataset contained two independent studies. Differentially expressed genes (DEGs) were identified between control and diseased groups in ICM and COPD dataset. *P*-value for each DEG was adjusted for multiple testing using the Benjamini-Hocberg correction [29]. DEGs with a *p*-adjusted threshold below 0.05 were deemed significantly differentially expressed.

2.5 *Gene co-expression analysis*

The summary of the co-expression analysis packages is presented in Table 1. A more detailed explanation of these packages can be found within the developers' publications. All the workflows in the present study are in accordance with the recommendations provided by the developers. In order to standardise the input data and assess the performance of the three packages, top 5,000 most variable genes were selected from *DES*eq2 (supplementary file 1 for ICM; supplementary file 2 for COPD). The selection criteria of the most variable genes were based on the transformed expression profiles using variance stabilizing transformation.

The selected genes were used to construct gene co-expression analysis in *WGCNA*, *CEMiTool*, and *coseq*. The parameters and codes used to construct the analysis are provided as supplementary code.

Table 1. ***Software packages for gene co-expression analysis***

Method	Version	Clustering Algorithm	Publish Year	Reference
WGCNA	1.68	Hierarchical clustering	2008	[12]
CEMiTool	1.8.3	Hierarchical clustering	2018	[13]
coseq	1.8.0	*K*-means	2018	[14, 15]

2.6 *WGCNA*

WGCNA was used to construct signed weighted gene co-expression modules from the top 5,000 variable genes. *WGCNA* identifies modules of genes based on correlation estimates. Two types of correlations can be performed on *WGCNA* to cluster genes into several modules: signed and unsigned adjacency matrix correlations. These two types of adjacency matrix correlations are defined as:

$$signed\ a_{ij} = \left|(1 + cor(x_i, x_j))/2\right|^{\beta}$$

$$unsigned\ a_{ij} = \left|cor(x_i, x_j)\right|^{\beta}.$$

Where a_{ij} is the network adjacency between gene expression profiles x_i and x_j. In the signed adjacency matrix, the correlation interval is scaled into 0 and 1. Signed correlations cluster

positively and negatively correlated genes into modules. Previous studies suggest that signed correlations provide more insight into the enrichment of functional groups [30]. Using unsigned correlations, positively and negatively correlated genes are clustered into the same modules, which are difficult to discriminate in the analysis. Module detection was based on the hierarchical clustering of adjacencies given by the topological overlap measure. A soft thresholding power β -value was chosen using "picksoftThreshold" function. A series of β -values (ranging between 1 and 30) was screened to evaluate the average connectivity degrees of different modules. A β -value was selected by plotting the R^2 against soft threshold β. In the ICM dataset, β=9 was selected because it represents the lowest power for which the scale-free topology index. For the COPD dataset, β=11 was selected. Subsequently, the adjacency matrix was transformed into a topological overlap matrix (TOM) to measure the connectivity of genes within the network [31]. The connectivity of genes is defined as the sum of its adjacency in relation to all other genes in the network. The TOM is measured between a value of 0 and 1. Based on TOM, a higher value (towards 1) indicates the set of genes are highly connected, and therefore, the strong interconnectivity will create a meaningful co-expression association. Contrary, when the TOM value is closer to 0, it signifies no connections between genes. A minimum module size was set to 40, highly correlated modules were merged by setting merging modules threshold to 0.2, and each of the modules was assigned a unique colour. Based on the recommended parameters, *WGCNA* generated 13 co-expressed modules in the ICM dataset while 14 modules in the COPD dataset.

2.7 *CEMiTool*

CEMiTool is similar to *WGCNA* but runs on an automated pipeline. The automated pipeline could improve the reproducibility of the results because the analysis parameters are standardized in the pipeline. The package also incorporates functional enrichment analysis to identify potential biological functions. Moreover, *CEMiTool* implements automated gene

filtration on gene expression profiles based on the inverse gamma distribution. To account for potentially inconsistent results in relation to the two other packages due to the difference in the number of input genes, the automated gene filtration function was disabled by setting 'filter=FALSE' in *CEMiTool* command. The number of modules was determined using the same algorithm as *WGCNA* by computing the soft thresholding power β. Based on the automated pipeline*, CEMiTool* derived 14 modules in the ICM dataset and 19 modules in the COPD dataset. The package provides a user-friendly and automated pipeline, outputs publication-ready figures, and generates a report in HTML format.

2.8 *coseq*

coseq is another Bioconductor package in R to generate gene co-expression analysis based on *K*-means clustering [15]. Prior to the clustering algorithm, the expression levels of the genes will be normalised. The normalisation method is represented as:

$$p_{ij} = \frac{y_{ij}/s_j+1}{\sum_{j'} y_{ij'}/s_{j'}+1},$$

where p_{ij} represents the proportion of normalised reads observed for gene *i* across all samples, s_j are normalization scaling factors correspond to library sizes and y_{ij} indicates the raw read count for gene *i* in sample *j*. *coseq* uses a novel transformation strategy called Log Centered Log Ratio (logCLR) to transform RNA-seq expression data follow by *K*-means clustering algorithm. The logCLR is defined by

$$logCLR\left(x_j\right) \coloneqq \begin{cases} -[\ln\left(1 - \ln[\frac{x_j}{g(x)}]\right)]^2 & if\ x_j/g(x) \leq 1, \\ (\ln[\frac{x_j}{g(x)}])^2 & otherwise, \end{cases}$$

where the logCLR is defined by $logCLR\left(x_j\right)$ and $g\ (x)$is the geometric mean of x. Therefore, *K*-means clustering algorithm based on logCLR is represented by

$$SSE_{logCLR}\left(C^{(K)}\right) \coloneqq \sum_{K=1}^{K} \sum_{i \in C_k} \left|\left|logCLR(X_i) - \mu_{k,logCLR}\right|\right|_2^2,$$

where SSE refers to sum of squared errors and $\mu_{k,logCLR}\left|\right|_2^2$ is the arithmetic mean of the transformed data. Then, the number of clusters is identified through slope heuristics approach. It is defined as

$$\widehat{K} := \underset{K \ \leq n}{\arg\min}\ \text{crit}\ (K).$$

The logCLR, together with *K*-means clustering was shown to produce tights and distinct clusters of genes [15]. The read counts of the top 5,000 variable genes identified through *DESeq2* were used as the input for *coseq*. The gene expression matrix was normalized using trimmed means of M values and transformed by logCLR. *K*-means clustering algorithm was fit to the transformed data, computed *K*= 2, …., 40 clusters, and iterated for 1000 cycles. Based on the slope heuristics approach for model selection in ICM dataset, the *K*=17 was selected and derived 17 clusters, each of the clusters contained a different number of genes. For the COPD dataset, the *K*=12 was selected and generated 12 co-expressed clusters. *coseq* pipeline is easy to use and does not require high computational power. However, users are advised to repeat the process several times to ensure reproducible results. Results from one run to another might differ due to their dependency on the initialization point.

2.9 *Clinical traits association*

WGCNA and *CEMiTool* have the function to correlate clinical parameters to the gene modules. Module-trait associations were estimated from the correlation between module eigengenes (ME) and clinical parameters. The ME of a module implies the first principal component of the module. This process allows easy identification of a module that is strongly correlated to clinical parameters. In the present study, disease status (ICM or COPD) served as a clinical parameter and estimated the correlations against the gene expression profiles. *coseq* does

not provide an algorithm to study the association of clinical parameters to gene expression profiles.

2.10 *Functional enrichment analysis*

Each of the gene modules derived from ICM and COPD dataset was subjected to functional enrichment analysis to determine the biological functions. *clusterProfiler* version 3.9 [32] revealed the potential functions in each module, and Gene Ontology (GO) terms with false discovery rate (FDR) threshold below 0.05 were considered statistically significant.

3.0 Results

3.1 Differentially expressed genes analysis

3.1.1 Ischemic cardiomyopathy RNA-seq dataset

The present study utilised two RNA-seq data sets available on public domain (GSE48166 and GSE116250) [20, 21]. In total, 57 samples were retrieved from two studies, 28 ICM samples and 29 non-heart failure samples. First, PCA plots were used to assess the quality and variation of two data sets, using normalised read counts of 21,878 identified genes. In Figure 1A, samples were clustered according to two library preparation batches indicating technical artefacts. On the left side of the plot in Figure 1A, samples were generated from GSE48166 while on the right they were from GSE116250. The technical artefacts could be introduced by library preparation and sequencing machines, which may explain the segregation of samples on the PCA plot from two different sources rather than experimental design (Figure 1A). As failure to correct batch effects will compromise the co-expression analysis, batch effects were adjusted in *DESeq2* package. The sources of the samples were flagged as "~Batch" in the design function and the estimated batch effect factors were used to model the expression

values in the regression step. After batch effects correction, the PCA plot (Figure 1B) showed visible separation between control and ICM patients based on global expression of 25,897 transcripts. Samples were clustered according to control and ICM status (Figure 1B; red and turquoise shapes). The first principal component on the X-axis contributes majorly to the PCA plot, which explained over 19% of the variance and separated samples into two groups.

Differential expression analysis was performed to identify genes that are up- or down-regulated in relation to ICM. Negative-Binomial-distribution was used to model the differential expression analysis and calculated the dispersion estimate for each gene (Figure 1C). The dispersion estimates account for the variance in gene expression and generate more homogenous expression levels across genes. *DESeq2* identified a total of 4985 differentially expressed genes (*padj* <0.05) (Figure 1D). Of these, 2281 genes were downregulated in ICM patients while 2704 were upregulated (supplementary file 3). After batch effects correction, 5000 genes were selected based on the highest variance to construct co-expression analysis using the three packages. The aim was to test the efficiency and reproducibility of each package. In each of the co-expression packages, default parameters recommended by the developers were used. Using the 5,000 most variable genes, *WGCNA* identified 13 co-expression modules while *CEMiTool* detected 14. Conversely, *coseq* identified 17 co-expression modules - the highest number of modules among three packages.

3.1.2 Chronic obstructive pulmonary disease dataset

To demonstrate the consistency of the results in the present study, we downloaded another RNA-seq dataset from the public repository (GSE57148) [22]. The COPD dataset consisted of 91 controls and 98 COPD patients. The quality of the dataset was first assessed using principal component analysis. The PCA was generated using normalised read counts of 34,187 genes. In Figure 2A, the PCA plot did not show apparent separation between two groups of individuals. PCA only considers the variance of the data, and the direction of the

highest variance may not necessarily represent a true expression profile. Therefore, co-expression analysis is useful in this case, where it reduces the dimensionality of the data by groups genes with similar expression levels.

Subsequently, differential expression analysis was performed to identify differentially regulated genes in this COPD dataset. *DESeq2* detected a total of 8437 differentially expressed genes with the threshold at *padj* <0.05 and consisted 3905 upregulated genes and 4532 downregulated genes (Figure 2B and supplementary file 4). For co-expression analysis, top 5000 genes with highest variances were selected. Using default parameters in three co-expression tools, *WGCNA* identified 14 modules, *CEMiTool* detected 19 modules and *coseq* derived 12 modules.

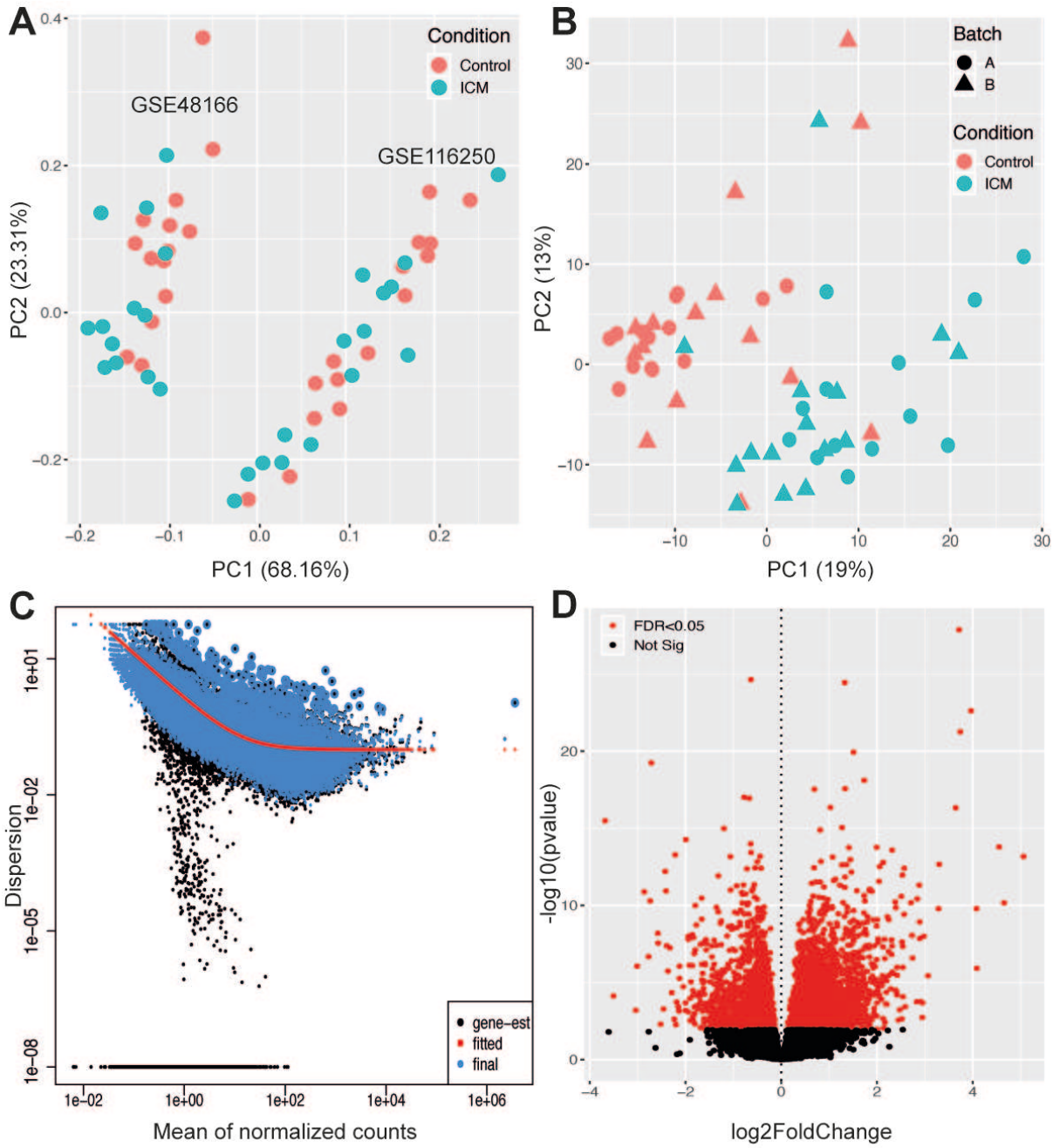

Figure 1. Differential expression analysis using *DESeq2* for ICM dataset. 57 samples were derived from GSE48166 and GSE116250. PCA plots were used to identify batch effects, before **(A)** and after **(B)** batch effect correction. Circle and triangle symbols represent batch 1 and 2, while red and turquoise colour shapes indicate control and ICM, respectively. **(C)** dispersion plot derived from *DESeq2*. Red dots represent the trend line for all samples, black dots indicate outliers, and blue dots imply corrected dispersion value. **(D)** volcano plot of differentially expressed genes between control and ICM patients. Each dot represents one gene and red colour dots denote significant differentially expressed genes that achieved adjusted *P*-value <0.05.

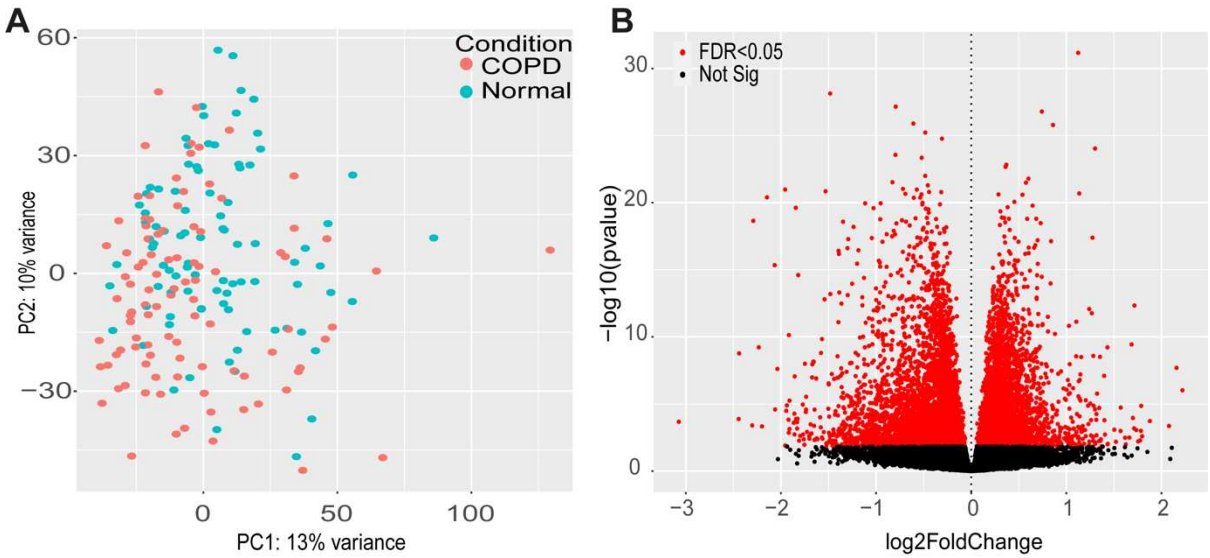

Figure 2. Differential expression analysis using *DESeq2* for COPD dataset. 189 samples were derived from GSE57148. **(A)** PCA plot was used to assess the quality of the RNA-seq dataset. Red and turquoise colour circles represent COPD and control samples. **(B)** volcano plot of differentially expressed genes between control and COPD patients. Each dot represents one gene and red colour dots denote significant differentially expressed genes that achieved adjusted *P*-value <0.05.

3.2 *WGCNA*

3.2.1 Ischemic cardiomyopathy RNA-seq dataset

WGCNA was used to construct co-expression analysis using the 5,000 most variant genes. First, hierarchical clustering was applied to identify the clustering of the samples. From the dendrogram, 57 samples were clustered according to the disease status (Figure 3A). We found two big clusters with ICM patients (red) and healthy patients (white). Several soft-thresholding powers were tested to identify the relative balanced scale independence and mean connectivity for co-expression network construction. In Figure 3B, power 9 was found to be the lowest power at which the scale-free topology fit index R^2 flattens out after reaching power 9 (Figure 3B). Therefore, power 9 was chosen to generate a hierarchical clustering

tree. The highly similar modules were merged by cutting the dendrogram at the height of 0.2, which corresponds to 0.8 correlation (Figure 3C). This improves the intra-connectedness of the modular genes and derived 13 modules (supplementary file 5). Each co-expressed module was allocated a specific colour. Among the 13 modules, the grey module contained the highest number of genes (2637 genes) followed by red module (413 genes) while the cyan module contained the lowest number of genes (81 genes). The relationship of the modules to disease status was evaluated (Figure 3D). Six modules were found to be significantly associated with disease status (p-value< 0.05). The cyan module shows the strongest association with ICM (r=0.77, p-value=2×10^{-12}) followed by the grey module (r=0.74, p-value=3×10^{-11}), blue module (r=0.63, p-value=1×10^{-7}), green module (r=0.50, p-value=8×10^{-5}), black module (r=0.34, p-value=0.01), and salmon module (r=0.32, p-value=0.02). Although the grey module appears to be associated with disease status and formed the largest module, careful investigation is needed as this module contained a high number of unassigned genes.

3.2.2 Chronic obstructive pulmonary disease dataset

For the COPD dataset, *WGCNA* was applied to top 5,000 genes to construct co-expression modules. In Figure 4A, the dendrogram shows the clustering of 189 samples according to the sample condition. The soft-power threshold 11 was selected to construct co-expression modules. In Figure 4B, power 11 was found to be the lowest power at which the scale-free topology fit index R^2 flattens out. Further, modules with high similarity were merged by setting a threshold at the height of 0.2 (Figure 4C). After merging modules with high similarity, the final 14 modules were derived (supplementary file 6). Each co-expressed module was allocated a specific colour. Among the 14 co-expression modules, the turquoise module contained the highest number of genes (1153 genes) followed by blue module (583 genes) while the tan module contained the lowest number of genes (58 genes). The association of the modules to disease status was assessed (Figure 4D). Of the 14 modules, five modules were found to be significantly associated with disease status (p-value< 0.05). The magenta

module shows the strongest association with COPD (r=0.56, p-value=8×10^{-17}) followed by the brown module (r=0.29, p-value=4×10^{-5}), black module (r=0.17, p-value=0.02), yellow module (r=-0.35, p-value=1×10-6) and greenyellow module (r=-0.17, p-value=×0.02).

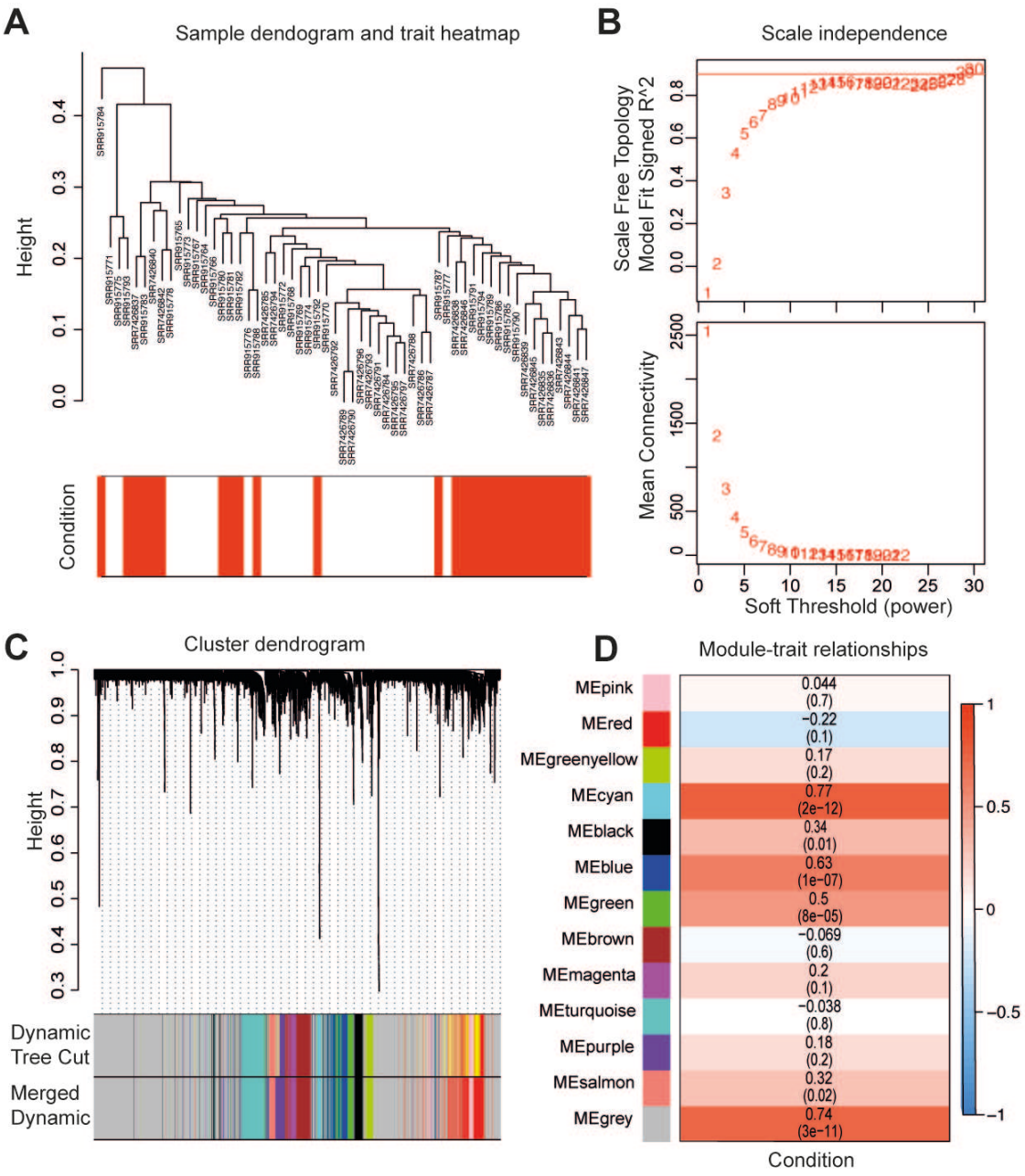

Figure 3. *WGCNA* outputs for ICM dataset. (A) Sample clustering for the 5,000 genes using expression profiles between control and ICM samples. The red colour bar indicates ICM patients while white colour bar denotes controls. **(B)** Determination of network topology using various soft-thresholding power β. The bottom panel shows the mean connectivity for various soft-thresholding power β **(C)** Cluster dendrogram of 13 co-expressed genes modules. The dendrogram was constructed with unsupervised hierarchical method and each colour bar below the dendrogram represents a module. **(D)** Module-condition correlations with the significance level. Each row corresponds to a module eigengene and each cell contains the correlation coefficient and *p*-value in parentheses. The colour of the table represents the strength of the correlation from positively correlated (red) to negatively correlated (blue).

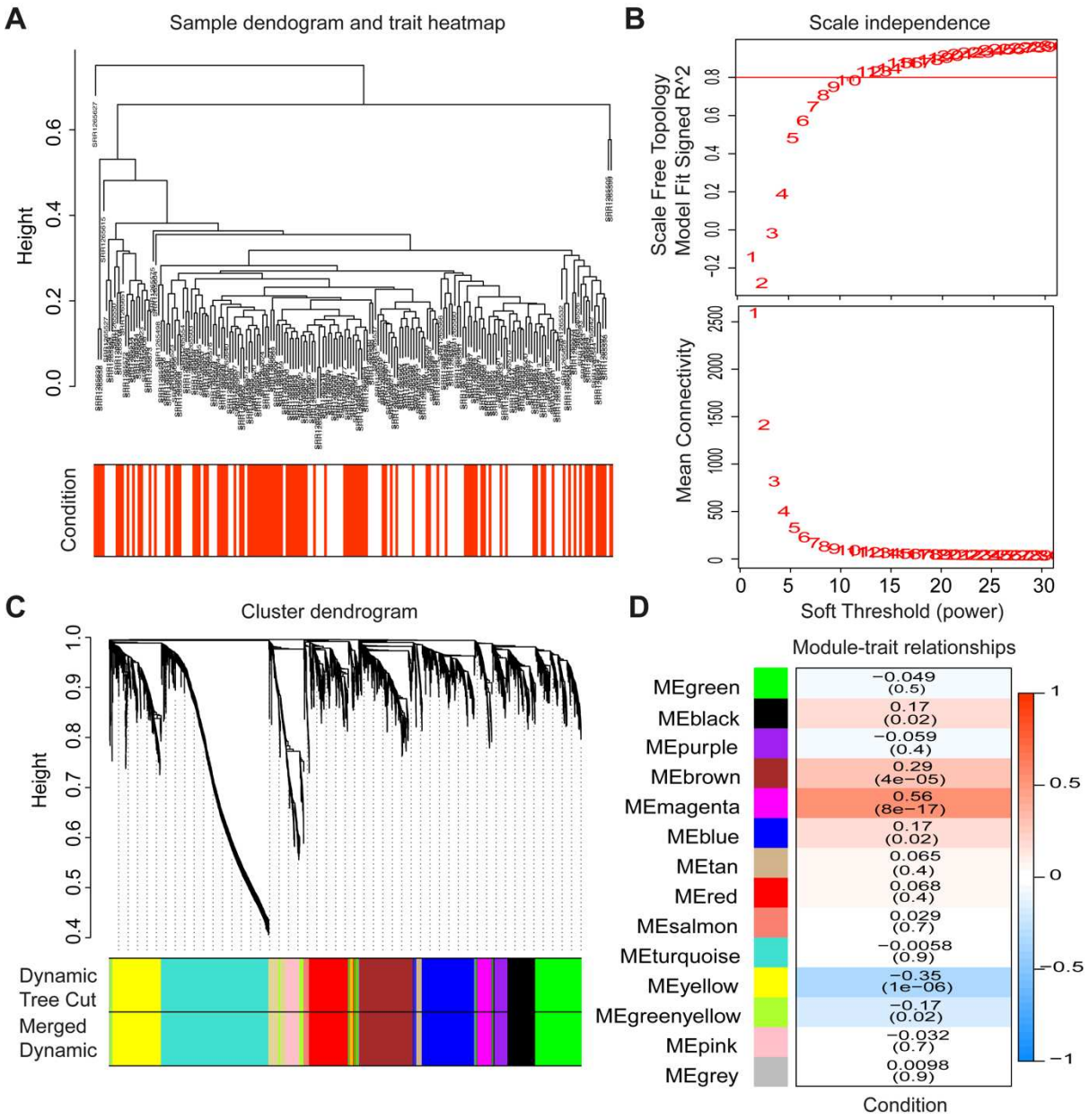

Figure 4. *WGCNA* outputs for COPD dataset. (A) Sample clustering for the 5,000 genes using expression profiles between control and COPD samples. The red colour bar indicates COPD patients while white colour bar denotes controls. **(B)** Determination of network topology using various soft-thresholding power β. The bottom panel shows the mean connectivity for various soft-thresholding power β **(C)** Cluster dendrogram of 14 co-expressed genes modules. The dendrogram was constructed with unsupervised hierarchical method and each colour bar below the dendrogram represents a module. **(D)** Module-condition correlations with the significance level. Each row corresponds to a module eigengene and each cell contains the correlation coefficient and *p*-value in parentheses. The colour of the table represents the strength of the correlation from positively correlated (red) to negatively correlated (blue).

3.3 *CEMiTool*

3.3.1 Ischemic cardiomyopathy RNA-seq dataset

CEMiTool operates on similar principles as *WGCNA,* but the pipeline is automated with pre-defined parameters. Using the 5,000 high variable genes, a dendrogram was generated, showing the clustering of 57 samples (Figure 5A). The dendrogram derived from *CEMiTool* is consistent with *WGCNA*, where samples were clustered according to the disease status. Two large clusters belonging to ICM patients (turquoise) and healthy patients (red) were generated. In the default settings, the scale-free topology fit index was pre-set at 0.80, and soft-threshold value of 14 was selected (Figure 5B). The dissimilarity threshold of 0.8 was used as a cut-off on hierarchical clustering, which identified 14 co-expression modules. Gene set enrichment analysis was performed to identify the module activity between ICM and healthy patients (Figure 5C). Among the 14 co-expressed modules, 11 modules showed significant module activity (p-value< 0.05), such as M1, M2, M3, M4, M5, M6, M8, M9, M10, M11, and M12 (supplementary file 7). Of those, the largest module contained 361 co-expressed genes (M1), while the smallest module contained 35 genes (M13). Figure 5D shows the profile plots of four co-expressed modules. Each line represents the expression level of each gene in a module, and the black line denotes the mean expression of all genes. In Figure 5D, the first plot shows the gene expression levels in the M5 module. The expression levels were higher in ICM patients (light blue background) with several peaks. In contrast, M8 and M12 had homogenous expression levels across all samples with several low peaks.

3.3.2 Chronic obstructive pulmonary disease dataset

For the COPD dataset, the 5,000 high variable genes were applied to *CEMiTool.* In Figure 6A, the dendrogram shows the clustering of 189 samples. The dendrogram is consistent with *WGCNA*, where the samples clustered loosely according to the disease status. Using the

default parameters, the soft-threshold value of 5 was selected (Figure 6B) which generated 19 co-expressed modules. *CEMiTool* performs gene set enrichment analysis automatically to identify module activity between COPD and control subjects (Figure 6C). Of those 19 co-expression modules, 13 modules showed significant module activity (p-value< 0.05), such as M11, M17, M2, M3, M7, M12, M13, M14, M8, M9, M4, M18 and M5 (supplementary file 8). The M1 module contained the highest number of genes (1116 genes), while M18 module contained the smallest number of genes (45 genes). Figure 6D shows the profile plots of three co-expressed modules. Each line represents the expression level of each gene in a module, and the black line denotes the mean expression of all genes.

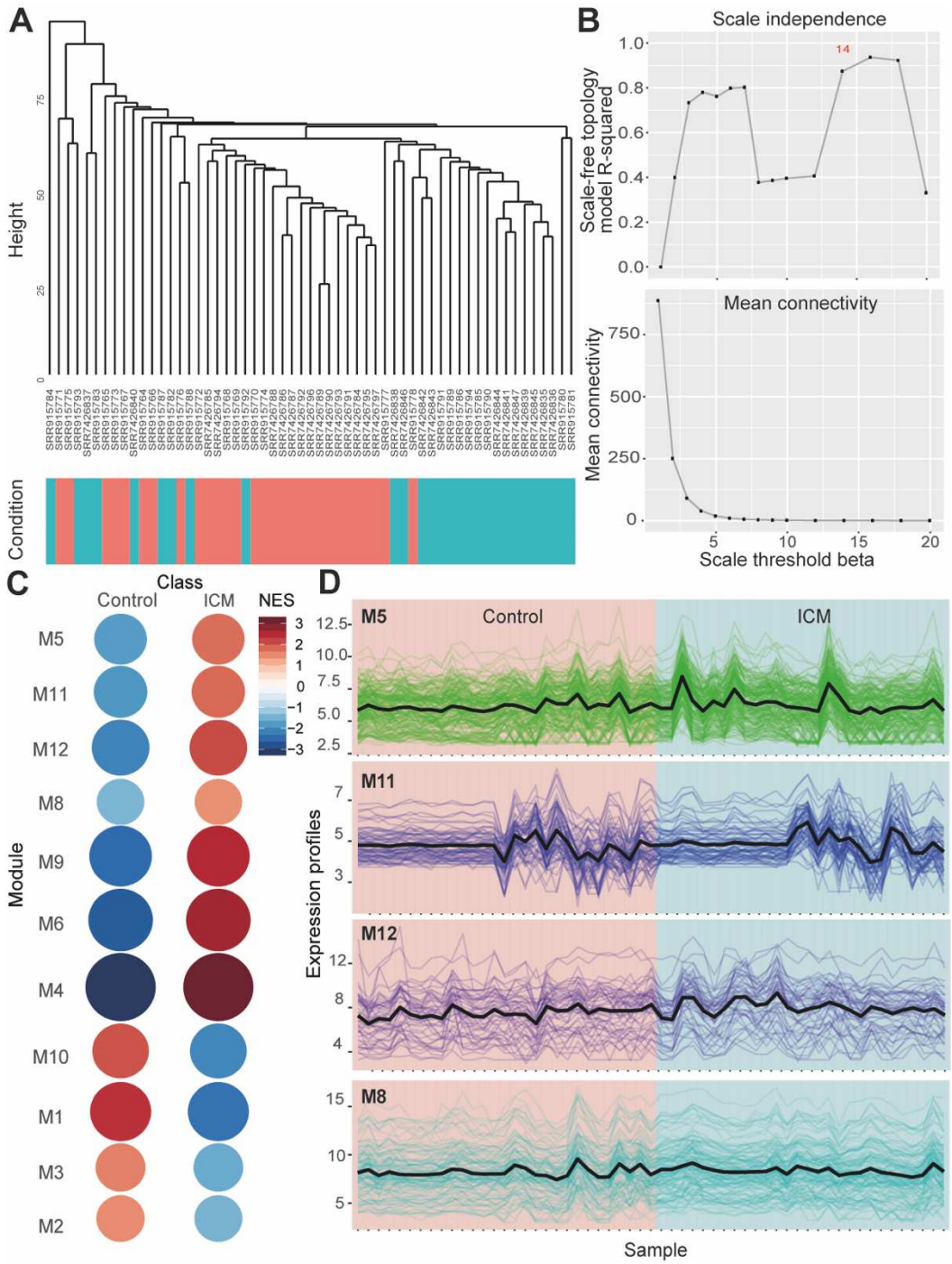

Figure 5. *CEMiTool* outputs for ICM dataset. (A) Clustering dendrogram for 5,000 genes based on the expression profiles. Green colour represents ICM patients while red colour denotes healthy controls. **(B)** Scale-free topology index and mean connectivity to identify power β between 1 and 20. From the plot, scale-free topology can be achieved at the soft-thresholding power of 14, above 0.8 scale free topology threshold based on the default settings by *CEMiTool.* **(C)** Gene enrichment analysis showing the module activity on control and ICM group. The size of the circle and colour corresponding to the normalised enrichment score. **(D)** Expression profiles of four co-expressed genes modules in control and ICM

patients. Each line in the plot represents each gene. The dark bold line indicates the mean expression value.

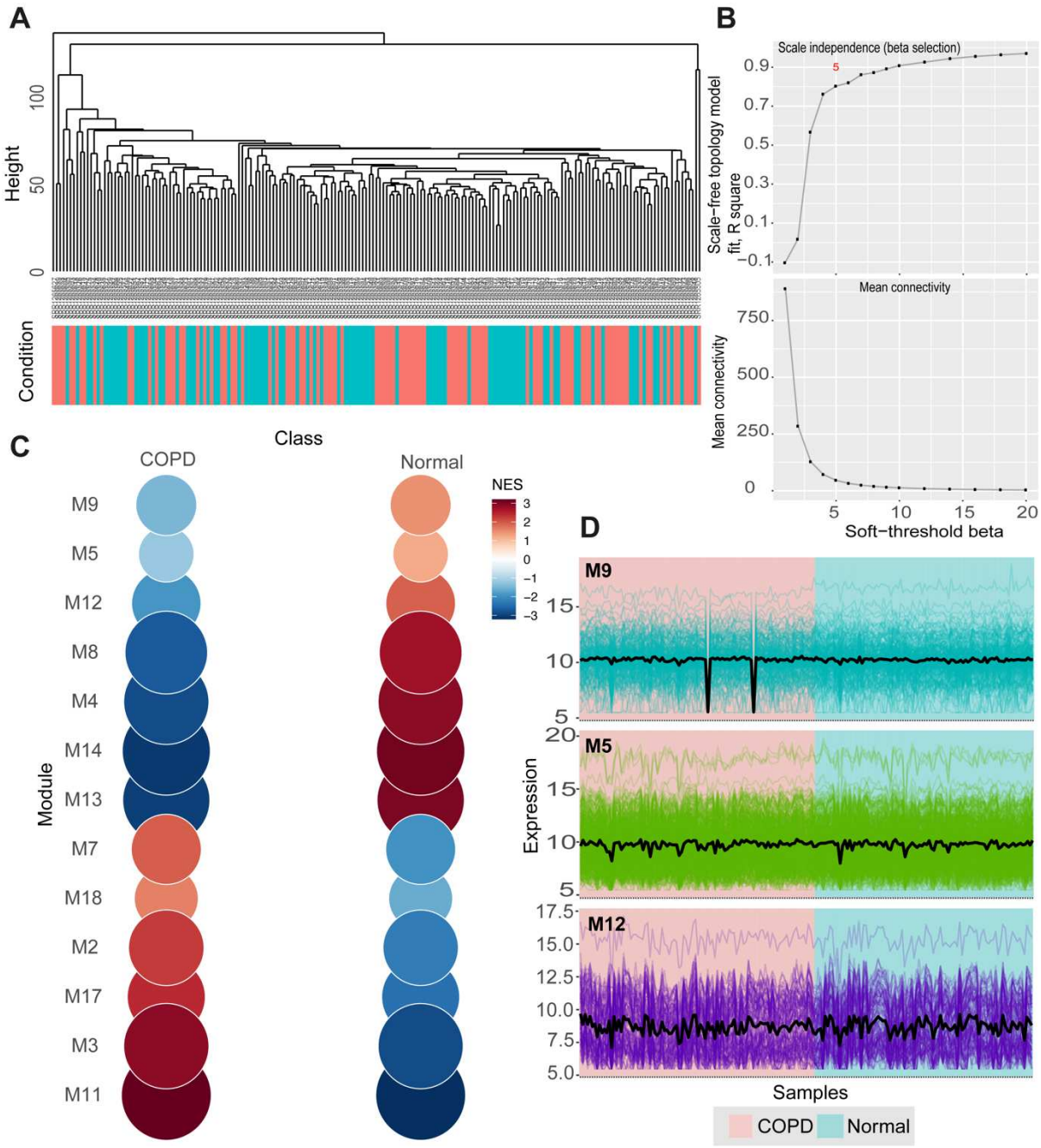

Figure 6. *CEMiTool* outputs for COPD dataset. (A) Clustering dendrogram for 5,000 genes based on the expression profiles. Green colour represents controls while red colour denotes COPD individuals. **(B)** Scale-free topology index and mean connectivity to identify power β between 1 and 20. From the plot, scale-free topology can be achieved at the soft-thresholding power of 5, above 0.8 scale free topology threshold based on the default settings by *CEMiTool.* **(C)** Gene enrichment analysis showing the module activity on control and COPD group. The size of the circle and colour corresponding to the normalised enrichment score. **(D)** Expression

profiles of three co-expressed genes modules in control and COPD patients. Each line in the plot represents each gene. The dark bold line indicates the mean expression value.

3.4 *coseq*

3.4.1 Ischemic cardiomyopathy RNA-seq dataset

coseq is a co-expression analysis package based on the *K*-means approach. In order to compare the results to *WGCNA* and *CEMiTool*, the top 5,000 variable genes were used as the input. Users should also be aware that *coseq* outputs depend on the initialization points. Therefore, the number of co-expressed modules may vary from one computer to another and one run to another. The initialization points may hinder the reproducibility of the results, however, in the current analysis, 1,000 iterations were used to avoid problems arising from initialization points. The results were consistent between each run when the iterations were set to 1,000. The *K*-means algorithm identified 17 co-expressed modules (Figure 7A). Of those, nine modules have higher mean expression levels in ICM patients compared to healthy patients. The six boxplots in Figure 7A (3, 4, 5, 7, 12, and 14) suggested module-specific profiles, where the mean expression levels are higher in ICM patients. Module 10 has the largest number of co-expressed genes (1452 genes), while module 2 has only one gene (supplementary file 9). The maximum conditional probabilities for each module are presented in Figure 7B. It represents the number of co-expressed genes in each module and maximum conditional probabilities. In this case, the maximum conditional probabilities threshold was fixed at 0.80. Modules contained a high number of genes with maximum conditional probabilities above 0.80, suggesting the module assignment is high confidence. Module 10, 11, and 15 have higher number of co-expressed genes with maximum conditional probabilities above 0.80, and therefore, the membership for the genes assigned to the modules has a greater degree of certainty.

3.4.2 Chronic obstructive pulmonary disease dataset

For the COPD dataset, the 5,000 variable genes were applied to *coseq*. At the start of the analysis, 1,000 iterations were specified to avoid errors in the results due to initialization points and to ensure reproducibility. Through the *K*-means algorithm implements in *coseq*, 12 co-expressed modules were generated (Figure 8A). All the co-expression modules showed higher mean expression profiles in COPD subjects compared to controls. Of the 12 modules, module 7 contained the largest number of genes (2140 genes), while module 2 and module 3 contained the lowest number of genes (4 genes) (supplementary file 10). The maximum conditional probabilities for each module are presented in Figure 8B. It represents the number of co-expressed genes in each module and maximum conditional probabilities. In this analysis, the maximum conditional probabilities threshold was fixed at 0.80. Modules contained a high number of genes with maximum conditional probabilities above 0.80, suggesting the module assignment is high confidence. Module 7, 8, and 12 have higher number of co-expressed genes with maximum conditional probabilities above 0.80, and therefore, the membership for the genes assigned to the modules are more accurate.

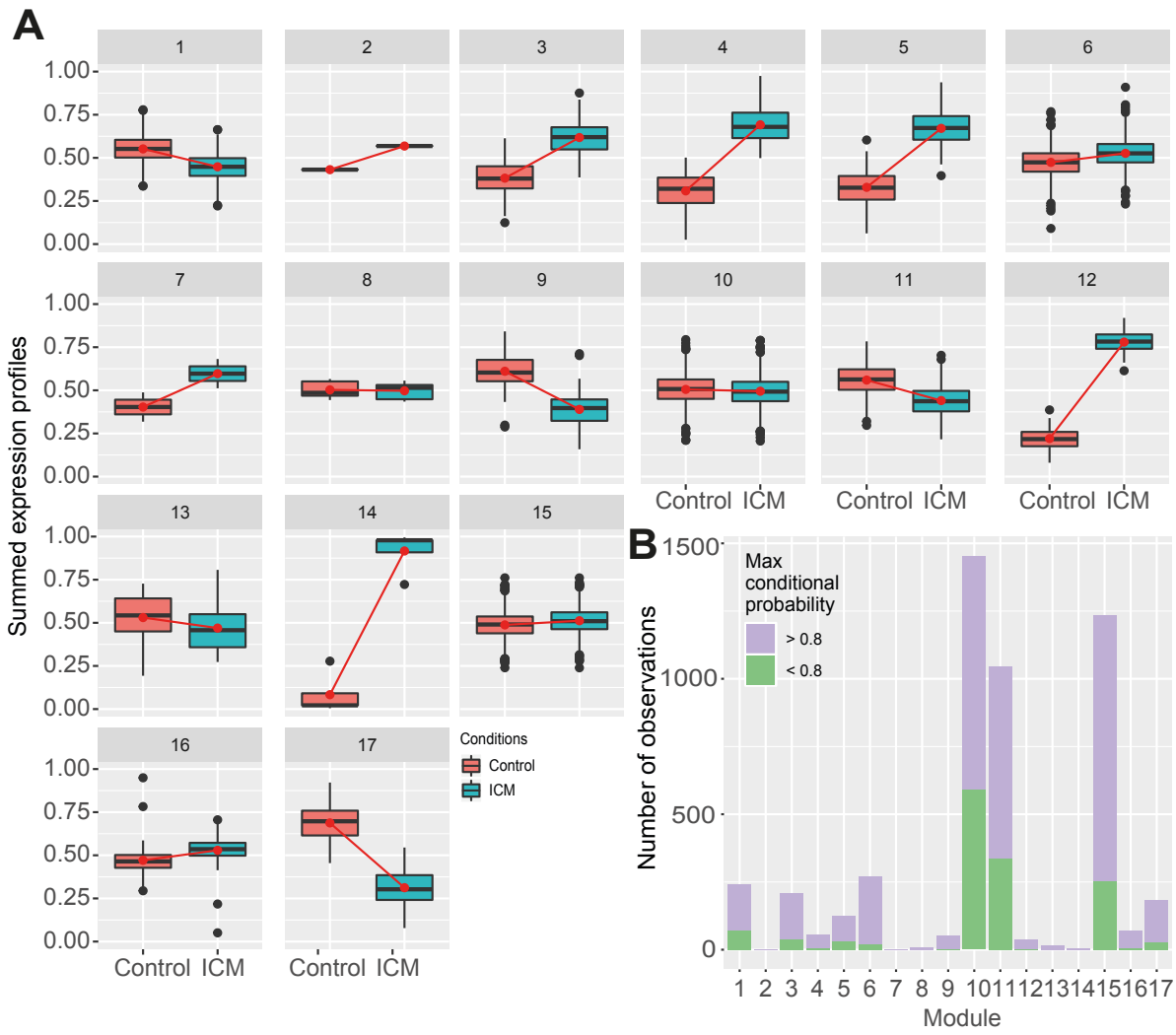

Figure 7. *coseq* outputs for ICM dataset. (A) 17 co-expression clusters derived using *K*-means approach. Normalized expression profiles were used to construct the clusters. The red and green boxplots represent expression profiles in control and ICM patients, respectively. The connected red lines indicate the mean expression profiles for each group. **(B)** Number of observations with a maximum conditional probability in which the genes assigned to each cluster. The threshold was set at 0.8 per cluster.

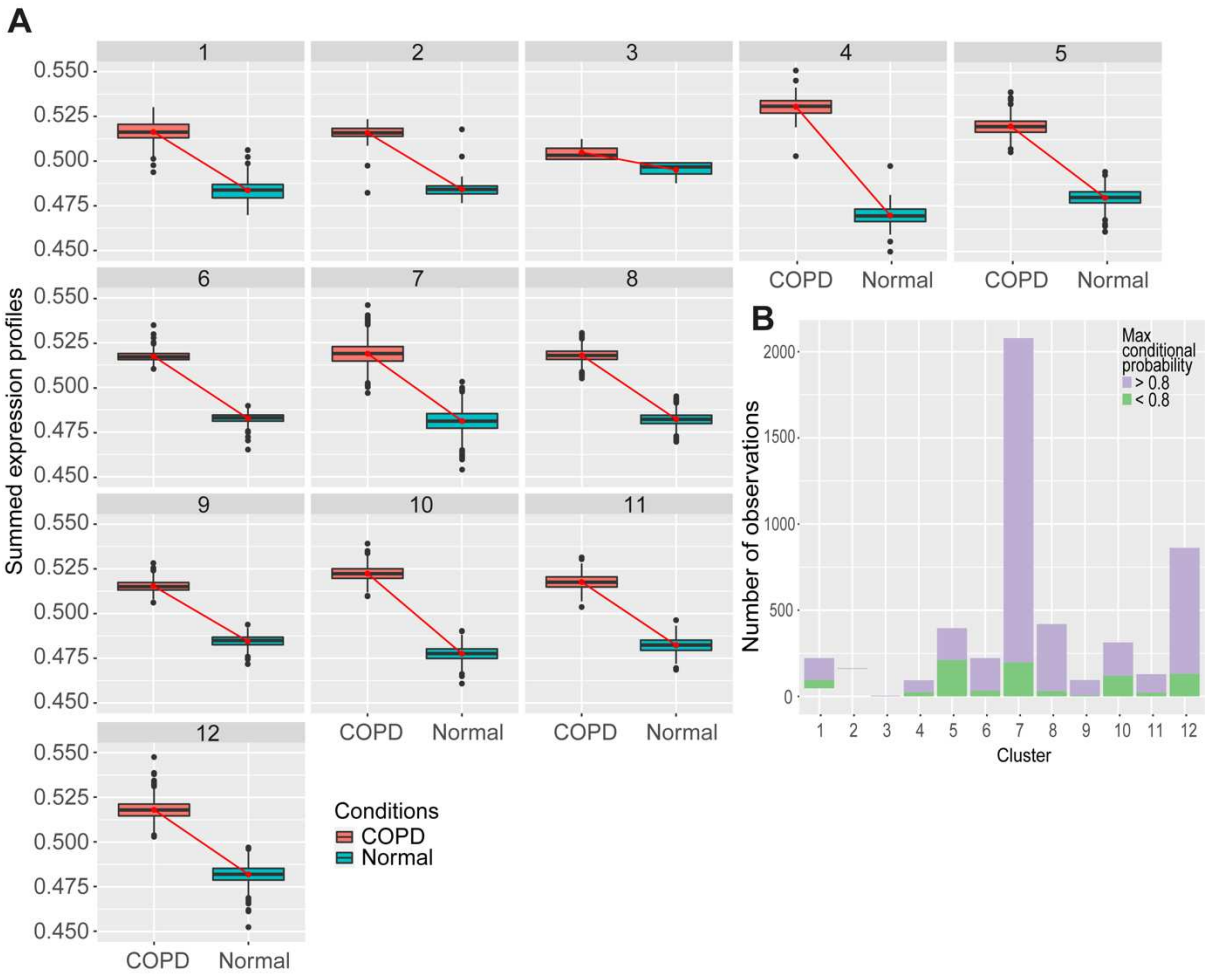

Figure 8. *coseq* outputs for COPD dataset. (A) 12 co-expression clusters derived using *K*-means approach. Normalized expression profiles were used to construct the clusters. The red and green boxplots represent expression profiles in COPD patients and controls, respectively. The connected red lines indicate the mean expression profiles for each group. **(B)** Number of observations with a maximum conditional probability in which the genes assigned to each cluster. The threshold was set at 0.8 per cluster.

3.5 GO enrichment analysis

3.5.1 Ischemic cardiomyopathy RNA-seq dataset

To compare the outputs across three packages, functional enrichment analysis was applied to each module to determine the underlying biological processes (supplementary file 11, 12 and 13). *WGCNA* produced 13 co-expressed modules, *CEMiTool* generated 14 modules, and *coseq* identified 17 modules (Figure 9A). The most representative module of genes was reported here such as green module in *WGCNA,* M6 module in *CEMiTool,* and module 3 in *coseq*. Between these three modules generated from three packages, there are 32 consensus co-expressed genes that contribute to similar GO terms (Figure 9B – D). These modules enriched for extracellular structure organization (GO:0043062), extracellular matrix organization (GO:0030198), and collagen fibril organization (GO:0030199).

3.5.2 Chronic obstructive pulmonary disease dataset

For the COPD dataset, we compare the functional enrichment analysis of the co-expressed gene modules derived from three co-expression tools. The aim was to identify the underlying biological processes (supplementary file 14, 15 and 16). *WGCNA* produced 14 co-expressed modules, *CEMiTool* generated 19 modules, and *coseq* identified 12 modules (Figure 10A). The most representative module from each co-expression package was chosen to demonstrate the consensus findings. The most representative module of genes was reported here such as blue module in *WGCNA*, M2 module in *CEMiTool* and module 7 in *coseq*. Intriguingly, there are 235 similar GO terms found between these three modules (Figure 10B – D). These modules enriched for extracellular structure organization (GO:0043062), extracellular matrix organization (GO:0030198), connective tissue development (GO:0061448), ossification (GO:001503) and cartilage development (GO:0051216).

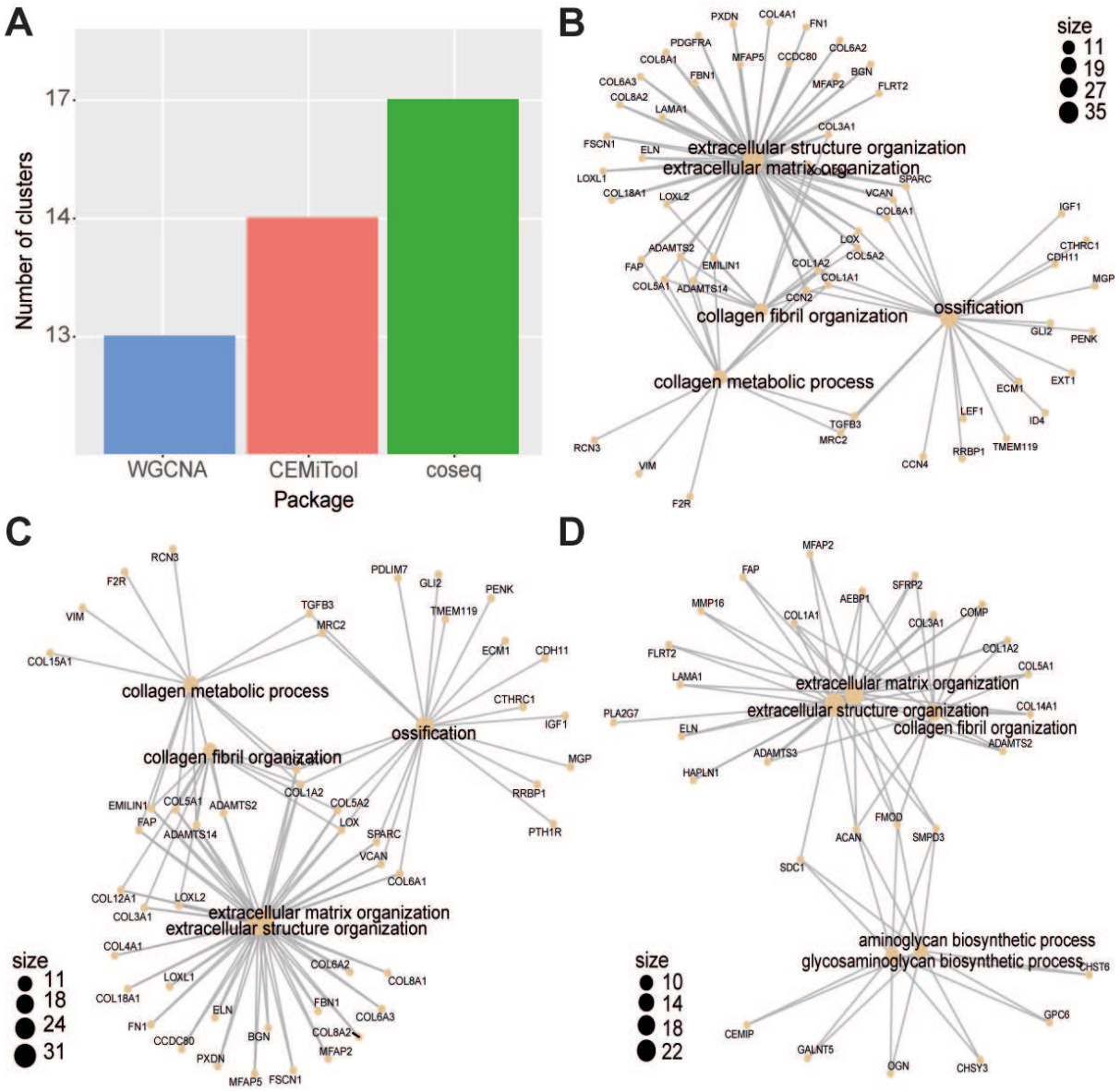

Figure 9. Number of clusters generated from each package and enrichment analysis for ICM dataset. (A) Number of co-expressed clusters produced from each package; *WGCNA*, *CEMiTool*, and *coseq*. Y-axis shows the number of co-expression modules generated by each package. Gene enrichment analysis for a representative module from each package. **(B)** *WGCNA* **(C)** *CEMiTool*, and **(D)** *coseq*. Each GO term achieves FDR< 0.05 and the size of the circle represents the number of co-expressed genes involved in the GO term.

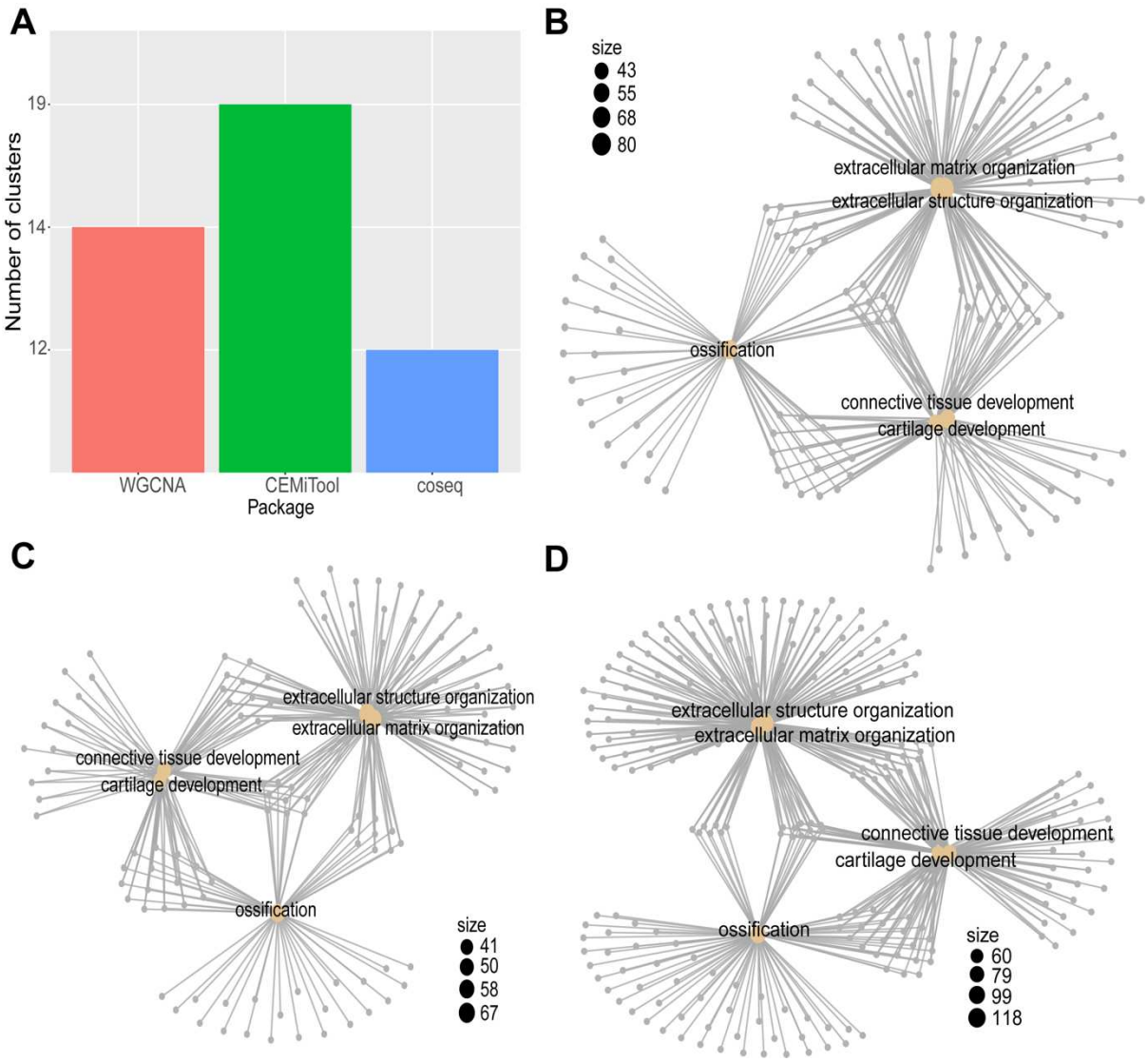

Figure 10. Number of clusters generated from each package and enrichment analysis for COPD dataset. (A) Number of co-expressed clusters produced from each package; *WGCNA*, *CEMiTool*, and *coseq*. Y-axis shows the number of co-expression modules generated by each package. Gene enrichment analysis for a representative module from each package. **(B)** *WGCNA* **(C)** *CEMiTool*, and **(D)** *coseq*. Each GO term achieves FDR< 0.05 and the size of the circle represents the number of co-expressed genes involved in the GO term.

4.0 Discussion

In this study, we tested three main co-expression analysis packages available on R: *WGCNA, CEMiTool,* and *coseq.* The main objective for this study was to evaluate these packages in terms of their reliability, reproducibility, and ease of use. Two publicly available RNA-seq data of ischaemic cardiomyopathy (GSE48166 and GSE116250) and one chronic obstructive pulmonary disease (GSE57148) were used to assess these three packages. We observed a degree of consensus outputs from three packages, despite a different number of co-expressed modules derived from each package. Together, the current study presents an overview of the three co-expression analysis packages, allowing researchers to choose a suitable co-expression analysis package for their studies.

WGCNA is a novel package for co-expression analysis. The algorithm is built on hierarchical clustering where module construction is based on the gene correlations. Substantial biomedical studies have applied *WGCNA* to identify clinically important therapeutic targets [33-35]. Furthermore, *WGCNA* can provide insights into the module-trait relationship, which can identify gene targets. However, the pipeline in *WGCNA* is complex and users are expected to have extensive bioinformatics experience. Multiple steps in *WCGNA* required empirical judgement such as the selection of a soft-thresholding value which may vary between scientists. As discussed later, this value impacts the reproducibility of the results. Although users are required to specify the parameters, WGCNA offers flexibility to optimise the parameters based on the dataset. *WGCNA* requires high computer resources to run the analysis. In the present analysis, *WGCNA* required around 30 minutes of computing time on an 8GB memory machine to generate the results. This does not consider the time required to understand the whole pipeline and optimise the parameters. Notably, however, *WGCNA* has extensive tutorials available and a well-documented pipeline.

CEMiTool is a relatively new package for co-expression analysis. The pipeline is adapted from the novel principle in *WGCNA.* The package is easy to use without extensive bioinformatics experience required because the algorithm has been automated. The package also produces publication-ready figures and an HTML report (supplementary file 17 for ICM dataset; supplementary file 18 for COPD dataset). *CEMiTool* also provides a function to determine gene set enrichment analysis to study the association of co-expressed modules with clinical traits. Its simplicity of use and free availability on R facilitates generation of biologically sound findings and high levels of adoption of this methodology are expected in the future. Despite the advantages of the automated pipeline, the pre-defined parameters might not be optimised for all studies. Therefore, careful consideration is needed when using this automated pipeline. The whole pipeline was completed with five command lines, and results were generated within ten minutes using 5,000 variable genes.

Contrasting with the other two packages, *coseq* is a robust alternative tool. It fits *K*-means for co-expression analysis and constructs modules. Additionally, *coseq* is built in with Gaussian mixture models for co-expression analysis [14]. *coseq* is relatively simple to use to construct co-expression modules with comprehensive documents available for users. The package does not require high computer resources compared to the other two packages; the results were produced within three minutes. However, compared to other packages, *coseq* is unable to determine the module-trait association. This might hinder the ability to identify gene targets for biomarker discovery studies. Moreover, *coseq* depends on the initialization point to construct co-expression modules. So, the findings may vary from one run to another, and this will jeopardise the reproducibility of the results. In the analysis, we set iterations to 1,000 prior to the co-expression analysis, to ensure the reproducibility of the results. We observed reliable outputs from each run after setting iterations to 1,000.

Soft-thresholding value is used in *WGCNA* and *CEMiTool* to construct modules of co-expressed genes. The soft-thresholding value plays a critical role in determining the number

of co-expressed modules, and thus has a significant impact on the findings. In the current analysis, we observed a lower power value generated by *CEMiTool* than *WGCNA.* This is consistent with the result obtained by Russo and colleagues [13]. *CEMiTool* pre-set the scale-free topology fit index (R^2) to 0.80, whereas *WGCNA* is defined by the user. A pre-defined R^2 will, therefore, ensure the reproducibility of the results as the optimum R^2 is rather arbitrary. The R^2 value may differ between scientists based on their judgement, some may prefer a higher value because it yields a more scale-free network. However, *CEMiTool* has high reproducibility compare to *WGCNA*, owning to the pre-defined R^2 value.

Here, 5,000 most variable genes were used to construct co-expression modules in three packages. It should be noted that to effectively compare the three packages using the same number of genes, the filtering parameters in each package were disabled before proceeding with co-expression analysis. Three packages yielded a different number of co-expression module. In the ICM dataset, *WGCNA* produced 13 co-expression modules, 14 modules in *CEMiTool*, and 17 modules in *coseq*. In the COPD dataset, *WGCNA* produced 14 co-expression modules, 19 modules in *CEMiTool*, and 12 modules in *coseq*. Although we observed a different number of modules generated, there are some overlapping modular genes between packages. In the ICM dataset, *coseq* produced the highest number of co-expression modules, however, on closer scrutiny, some modules contained fewer genes for which the gene assignment was ambiguous - such as module 2, 7, 8, and 14.

Functional enrichment analysis was performed on each module to identify similarities and differences on the outputs across three novel packages. One of the most representative modules from each package was chosen to show the biological functions for which they were enriched. In both the ICM and COPD dataset, it is clear that all the three co-expression packages were able to generate similar GO terms. For the ICM dataset, there was a module consistently enriched for extracellular structure organization (GO:0043062), extracellular matrix organization (GO:0030198), and collagen fibril organization (GO:0030199).

Interestingly, there are 32 consensus co-expressed genes present in three packages that contribute to the GO terms. For the COPD dataset, there was a module enrichment for extracellular structure organization (GO:0043062), extracellular matrix organization (GO:0030198), connective tissue development (GO:0061448), ossification (GO:001503) and cartilage development (GO:0051216). Of those GO terms, there are 235 overlapping GO terms between these modules. These results suggest that the co-expression analysis produced by the three packages are comparable, and the findings are biologically sound. In literature review, extracellular matrix plays a critical role in cardiac homeostasis by providing structural support and transferring molecular signals [36-38]. Frangogiannis 2019 reported that ischemic injury drives the changes in the cardiac extracellular matrix and essentially regulates the cardiac inflammation and repair. This process may be implicated in the pathogenesis of cardiac remodelling [37, 38]. Furthermore, the extracellular matrix has been described to cross-link with fibrillar collagens in ischemic cardiomyopathy [38]. Fibrillar collagens provides mechanical support and acts on the cell surface receptors to transfer molecular signals [39]. Multiple studies have reported an increase level of fibrillar collagens through activating cardiac fibroblasts in ischemic heart [38]. As expected, these modules consistently enriched for collagen fibril organization. Consistently, extracellular matrix components in the lungs are associated significantly with COPD [40, 41]. The changes in the expression of extracellular matrix proteins is one of the predictive markers characterising the disease stages [42]. Based on functional enrichment analysis, three packages produced biologically relevant findings and provided insights into ischemic cardiomyopathy and chronic obstructive pulmonary disease. In the current analysis, we have not studied the co-expressed genes in a greater resolution, further work should emphasize the unifying co-expressed modules to identify therapeutic targets for ischemic cardiomyopathy and chronic obstructive pulmonary disease.

5.0 Conclusion

In this work, we compared the performance of three main co-expression packages available on R using real-world RNA-seq datasets. This serves as a guide to allow researchers to select a suitable tool for co-expression analysis. In general, our results showed that three packages are comparable based on functional enrichment analysis, although each package has its own advantages and disadvantages. *CEMiTool* has high reproducibility and the automated pipeline offers the ease of use for all researchers regardless of bioinformatics experience. In contrast, *WGCNA* presents a more complex pipeline, but offers flexibility to optimise the parameters according to the dataset. *coseq* requires less computer resource for analysis and produces comparable results to other packages, but lacks the ability to identify gene targets. We acknowledge that care should always be taken when interpreting co-expression modules. To improve the accuracy of results, it is advisable to compare the results with multiple packages. Therefore, considering the availability of documentation, simplicity of use, running time, and computational resources, we consider that *CEMiTool* outperforms other tools. We hope that this comparison of these three packages will assist researchers in selecting the most suitable tool for their investigation.

6.0 Reference

1. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences, 1998. **95**(25): p. 14863-14868.
2. Jiang, D., C. Tang, and A. Zhang, *Cluster analysis for gene expression data: a survey.* IEEE Transactions on Knowledge & Data Engineering, 2004(11): p. 1370-1386.
3. Ballouz, S., W. Verleyen, and J. Gillis, *Guidance for RNA-seq co-expression network construction and analysis: safety in numbers.* Bioinformatics, 2015. **31**(13): p. 2123-2130.
4. Cheng, C.W., et al., *Clinical expression and antigenic profiles of a Plasmodium vivax vaccine candidate: merozoite surface protein 7 (PvMSP-7).* Malaria journal, 2019. **18**(1): p. 197.
5. McDermott-Roe, C., et al., *Transcriptome-wide co-expression analysis identifies LRRC2 as a novel mediator of mitochondrial and cardiac function.* PloS one, 2017. **12**(2): p. e0170458.
6. Chang, Y.-M., et al., *Three TF co-expression modules regulate pressure-overload cardiac hypertrophy in male mice.* Scientific reports, 2017. **7**(1): p. 7560.
7. Li, K., et al., *Network-based transcriptomic analysis reveals novel melatonin-sensitive genes in cardiovascular system.* Endocrine, 2019. **64**(2): p. 414-419.

8. Tang, Y., et al., *Co-expression analysis reveals key gene modules and pathway of human coronary heart disease.* Journal of cellular biochemistry, 2018. **119**(2): p. 2102-2109.
9. Deshpande, V., et al., *Understanding the progression of atherosclerosis through gene profiling and co-expression network analysis in Apobtm2SgyLdlrtm1Her double knockout mice.* Genomics, 2016. **107**(6): p. 239-247.
10. Ma, X., et al., *Co-expression gene network analysis and functional module identification in bamboo growth and development.* Frontiers in genetics, 2018. **9**: p. 574.
11. Aoki, K., Y. Ogata, and D. Shibata, *Approaches for extracting practical information from gene co-expression networks in plant biology.* Plant and Cell Physiology, 2007. **48**(3): p. 381-390.
12. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC bioinformatics, 2008. **9**(1): p. 559.
13. Russo, P.S., et al., *CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses.* BMC bioinformatics, 2018. **19**(1): p. 56.
14. Rau, A. and C. Maugis-Rabusseau, *Transformation and model choice for RNA-seq co-expression analysis.* Briefings in bioinformatics, 2017. **19**(3): p. 425-436.
15. Godichon-Baggioni, A., C. Maugis-Rabusseau, and A. Rau, *Clustering transformed compositional data using K-means, with applications in gene expression and bicycle sharing system data.* Journal of Applied Statistics, 2019. **46**(1): p. 47-65.
16. Petereit, J., et al., *petal: Co-expression network modelling in R.* BMC systems biology, 2016. **10**(2): p. 51.
17. Ogata, Y., et al., *CoP: a database for characterizing co-expressed gene modules with biological information in plants.* Bioinformatics, 2010. **26**(9): p. 1267-1268.
18. Watson, M., *CoXpress: differential co-expression in gene expression data.* BMC bioinformatics, 2006. **7**(1): p. 509.
19. Gan, G., C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. Vol. 20. 2007: Siam.
20. Sweet, M.E., et al., *Transcriptome analysis of human heart failure reveals dysregulated cell adhesion in dilated cardiomyopathy and activated immune pathways in ischemic heart failure.* BMC genomics, 2018. **19**(1): p. 812.
21. Li, W., et al., *Identification of potential genes for human ischemic cardiomyopathy based on RNA-Seq data.* Oncotarget, 2016. **7**(50): p. 82063.
22. Kim, W.J., et al., *Comprehensive Analysis of Transcriptome Sequencing Data in the Lung Tissues of COPD Subjects.* Int J Genomics, 2015. **2015**: p. 206937.
23. Andrews, S., *FastQC: a quality control tool for high throughput sequence data*. 2010, Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
24. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-3048.
25. Krueger, F., *Trim galore.* A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, 2015.
26. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
27. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2013. **30**(7): p. 923-930.
28. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550.
29. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal statistical society: series B (Methodological), 1995. **57**(1): p. 289-300.
30. Horvath, S., *Weighted network analysis: applications in genomics and systems biology*. 2011: Springer Science & Business Media.

31. Yip, A.M. and S. Horvath, *Gene network interconnectedness and the generalized topological overlap measure.* BMC bioinformatics, 2007. **8**(1): p. 22.
32. Yu, G., et al., *clusterProfiler: an R package for comparing biological themes among gene clusters.* Omics: a journal of integrative biology, 2012. **16**(5): p. 284-287.
33. Yang, Q., et al., *Candidate biomarkers and molecular mechanism investigation for glioblastoma multiforme utilizing WGCNA.* BioMed research international, 2018. **2018**.
34. Liu, Z., et al., *Identification of surrogate prognostic biomarkers for allergic asthma in nasal epithelial brushing samples by WGCNA.* Journal of cellular biochemistry, 2019. **120**(4): p. 5137-5150.
35. Giulietti, M., et al., *Identification of candidate miRNA biomarkers for pancreatic ductal adenocarcinoma by weighted gene co-expression network analysis.* Cellular Oncology, 2017. **40**(2): p. 181-192.
36. Hughes, C. and J. Jacobs, *Dissecting the role of the extracellular matrix in heart disease: Lessons from the Drosophila genetic model.* Veterinary sciences, 2017. **4**(2): p. 24.
37. Frangogiannis, N.G., *The extracellular matrix in myocardial injury, repair, and remodeling.* The Journal of clinical investigation, 2017. **127**(5): p. 1600-1612.
38. Frangogiannis, N.G., *The Extracellular Matrix in Ischemic and Nonischemic Heart Failure.* Circulation Research, 2019. **125**(1): p. 117-146.
39. Bella, J. and D.J. Hulmes, *Fibrillar collagens*, in *Fibrous proteins: structures and mechanisms*. 2017, Springer. p. 457-490.
40. Bidan, C.M., et al., *Airway and Extracellular Matrix Mechanics in COPD.* Front Physiol, 2015. **6**: p. 346.
41. Ito, J.T., et al., *Extracellular Matrix Component Remodeling in Respiratory Diseases: What Has Been Found in Clinical and Experimental Studies?* Cells, 2019. **8**(4).
42. Bihlet, A.R., et al., *Biomarkers of extracellular matrix turnover are associated with emphysema and eosinophilic-bronchitis in COPD.* Respir Res, 2017. **18**(1): p. 22.

Supplementary files

Supplementary codes: Commands and parameters used in generating the results from *WGCNA*, *CEMiTool*, and *coseq*. The RNA-seq data were mapped to reference genome with *STAR* aligner and read counts were quantified with *featureCounts.* Raw read counts and differentially expressed genes were processed with *DESeq2*.

Supplementary file 1: List of 5,000 most variable genes derived from *DESeq2* for ICM dataset. Variance stabilising transformed normalised read counts for 5,000 genes derived from *DESeq2* after mapping RNA-seq data to human reference genome.

Supplementary file 2: List of 5,000 most variable genes derived from *DESeq2* for COPD dataset. Variance stabilising transformed normalised read counts for 5,000 genes derived from *DESeq2* after mapping RNA-seq data to human reference genome.

Supplementary file 3: Significant differential expressed genes in pairwise comparison between ICM and control group. Fold-change is log2- transformed. Genes with an adjusted *p*-value below the threshold of 0.05 were considered significantly regulated.

Supplementary file 4: Significant differential expressed genes in pairwise comparison between COPD and control group. Fold-change is log2- transformed. Genes with an adjusted *p*-value below the threshold of 0.05 were considered significantly regulated.

Supplementary file 5: 13 gene co-expressed clusters derived from *WGCNA* for ICM dataset. Each cluster is labelled with different colour code.

Supplementary file 6: 14 gene co-expressed clusters derived from *WGCNA* for COPD dataset. Each cluster is labelled with different colour code.

Supplementary file 7: 14 co-expressed clusters derived from *CEMiTool* for ICM dataset. Each cluster is labelled with a prefix "M".

Supplementary file 8: 19 co-expressed clusters derived from *CEMiTool* for COPD dataset. Each cluster is labelled with a prefix "M".

Supplementary file 9: 17 co-expressed clusters derived from *coseq* for ICM dataset. Each cluster is labelled numerically.

Supplementary file 10: 12 co-expressed clusters derived from *coseq* for COPD dataset. Each cluster is labelled numerically.

Supplementary file 11: Functional enrichment analysis for each cluster in *WGCNA* for ICM dataset. The analysis was carried out with *clusterProfiler* and significant GO terms are defined as having false discovery rate below 0.01.

Supplementary file 12: Functional enrichment analysis for clusters derived in *CEMiTool* for ICM dataset. The analysis was performed using *clusterProfiler* and significant GO terms are defined as having false discovery rate below 0.01.

Supplementary file 13: Functional enrichment analysis for clusters derived in *coseq* for ICM dataset. The analysis was carried out with *clusterProfiler* and significant GO terms are defined as having false discovery rate below 0.01.

Supplementary file 14: Functional enrichment analysis for each cluster in *WGCNA* for COPD dataset. The analysis was carried out with *clusterProfiler* and significant GO terms are defined as having false discovery rate below 0.01.

Supplementary file 15: Functional enrichment analysis for clusters derived in *CEMiTool* for COPD dataset. The analysis was performed using *clusterProfiler* and significant GO terms are defined as having false discovery rate below 0.01.

Supplementary file 16: Functional enrichment analysis for clusters derived in *coseq* for COPD dataset. The analysis was carried out with *clusterProfiler* and significant GO terms are defined as having false discovery rate below 0.01.

Supplementary file 17: Html report generated by *CEMiTool* for ICM dataset.

Supplementary file 18: Html report generated by *CEMiTool* for COPD dataset.

Acknowledgments

This work was undertaken on ARC4, part of the High Performance Computing facilities at the University of Leeds, UK.

Funding

This work was funded by Leeds Cardiovascular Endowment, Leeds Institute of Cardiovascular and Metabolic Medicine, University of Leeds.