UNIVERSITÀ DI BOLOGNA

MASTER'S DEGREE IN BIOINFORMATICS

MASTER'S THESIS

EnGNet 2.0: an improved gene co-expression network inference algorithm based on ensemble strategies and topology optimization

July 2021

Directors:
Fernando M. Delgado-Chaves
fmdelcha@upo.es
Francisco A. Gómez-Vela
fgomez@upo.es

Academic supervisor:
Samuele Bovo
samuele.bovo@unibo.it

Student:
Tommaso Becchi
tommaso.becchi@studio.unibo.it

Abstract

The role of Gene Co-expression Networks (GCNs) is becoming more and more relevant in the modeling of gene expression-based genetic relationships. Current methods for GCNs reconstruction often use a single co-expression measure. However, the combination of co-expression measures is considered complementary when modeling the underlying biological reality, whereas single co-expression measures on their own only detect partial relationships. Moreover, GCNs should exhibit certain topological properties as scale-freeness and sparseness because these features reflect a topological reality and they allow to identify hubs, the main regulators of the networks.

In this work, EnGNet 2.0 algorithm is presented. EnGNet 2.0 is a novel two-step method for the reconstruction of networks, combining multiple co-expression measures at the same time and optimizing the topological properties. First, EnGNet 2.0 uses an ensemble strategy that combines correlation methods and mutual information-based approaches. Second, an heuristic pruning algorithm optimizes both the size and the topological features of the network with a special focus on the hubs.

EnGNet 2.0 returns networks that significantly improve statistical and topological features compared to other methods. Moreover, the usefulness of this algorithm is proven by an application to a human dataset with samples affected by Atopic Dermatitis, revealing an innate immunity-mediated response to this pathology.

Keywords: Gene Co-expression Networks; scale-free topology; ensemble strategy; Systems Biology; disease gene prediction; network analysis;

Contents

1 Introduction

Nowadays, biological studies are often focused on modeling and discovering emergent properties of cells, tissues and organisms functioning as a system. For this reason, Systems Biology investigates complex biological systems from a computational and a mathematical point of view [1]. In this context, complex biological systems can be described at different levels, such as genomics, transcriptomics, proteomics and metabolomics [2]. Among these, transcriptomics studies are increasingly relevant, leading to a new understanding of the genes or pathways associated with specific cell types o specific pathological conditions [3]. In the last few years, micro-arrays experiments and RNA-seq experiments have produced an increasing quantity of gene expression data that need to be processed and interpreted. In particular, NGS-based technologies, such as RNA-seq [4], allow to obtain high-throughput results that lead to a better understanding of the biological complexity.

In this context, Gene Co-expression Networks (GCNs) have become a popular method for analyzing expression data. These networks represent the relationships between genes by means of a graph composed of nodes and edges where nodes correspond to genes and edges represent the relationship between them, as shown in Figure 1. GCNs are undirected networks because they represent mutual relationships and each edge has a weight that represents the value of the correlation between a specific pair of genes.

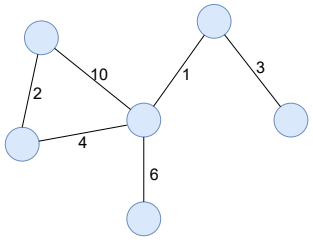


Figure 1: Simple GCN topology in which nodes represent genes and edges represent the relationships between them. These networks have undirected edges and each of them has an own weight that indicates the value of that specific correlation.

GCNs aim to represent a biological reality, consequently they need to satisfy cer-

tain properties. In particular, they are characterized by a scale-free topology, a property related to the degree distribution of the network. In biological networks, most genes are sparsely connected whereas a few are highly connected, resulting in a distribution of node degrees that tends to follow a power law [5]. Highly-connected nodes are called hubs, which have a central role in the regulation of the network processes [6].

A second topological properties, also related to scale-freeness, is the sparseness. GCNs are sparse because genes are usually regulated by a small and limited number of other genes, resulting in a limited number of regulatory inputs per node [7]. Higher-connected networks bring several difficulties, while sparse networks allow to find a compromise between model quality and model complexity [8, 9].

Consequently, GCNs are hierarchical networks in which highly-connected genes are at the top level of the hierarchy [10].

GCNs can be inferred using two main types of approaches: correlation-based and mutual information-based [11]. Both these methods are a measure of the mutual dependence between two variables, however each of them is more suitable for a specific type of interaction [12, 13] Several algorithms for GCN reconstruction have been implemented using different measures and combining them in different ways. Among them, we can highlight REVEAL [14], RELEVANCE [15], ARACNE [16], CLR [17], MRNET [18] or FyNE [19].

GCNs also differ in topology since they can be hyper-connected or optimized. Hyper-connected networks are obtained using all the interactions between the genes, while the use of thresholds allows to generate optimized networks that show only the relevant ones [20]. The use of thresholds also allows to increase the modularity of the networks, a topological feature that is related to scale-freeness [21]. Genes are divided in cluster that represent community structure in which the hubs play a central role.

GCNs properties described above have made them increasingly important in many fields of research. For example, they are applied in synthetic biology, biomarkers discovery or personalized medicine because they can be used for various purposes such as regulatory genes identification and functional gene an-

notation [22, 23]. In particular, they are a useful tool for the discovery of new interactions between genes. Consequently, poorly annotated genes can be studied in a more detailed way when they are connected with well-studied ones. This results in a better characterization of both the genes and their transcripts, allowing new biomarkers identification. For example, in Yan et al. [24], gene network analysis allowed the identification of molecular biomarkers for monitoring cancer progression and treatment. GCN's reliability was proved by different studies in which many predicted interactions have been confirmed experimentally [25].

However, the predictive power of single co-expression measures is limited because different co-expression measures return different results and each of them is more suited for a specific type of interaction. Relationships can be linear or non-linear and there is no one absolute approach that is able to detect the whole range of possibilities. Consequently, an ensemble approach is often used to better detect true biological connections. This method combines different scores, resulting in a wider range of identified interactions [26].

Gómez-Vela et al. [27] combined an ensemble strategy with a greedy-based optimization algorithm to generate GCNs, as shown in Figure 2. The combination of these two approaches is the basis of the development of EnGNet, a novel two-step method for gene networks inference. First, Spearman, Kendall and Normalized Mutual Information (NMI) were used for evaluating every gene pair relationship. Then, for each measure, a significance threshold was used to determine whether or not the relationship is considered valid by a specific measure. A first optimization was carried out through a voting system and a relationship was confirmed if it was considered significant by at least two measures. After this first optimization, the topological features of the network were optimized again by means of a pruning step. In particular, a modification of the Kruskal's minimum spanning tree (MST) was used and most significant edges were selected, until all nodes are connected with no cycles. This reduction in edges significantly improves the sparseness of the network. In the end, hubs were identified and for each of them, its linking edges that were removed in the ensemble network are again evaluated using a threshold. Each individual edge was added to the network if its weight exceeded this threshold, increasing the scale-free property of the network.

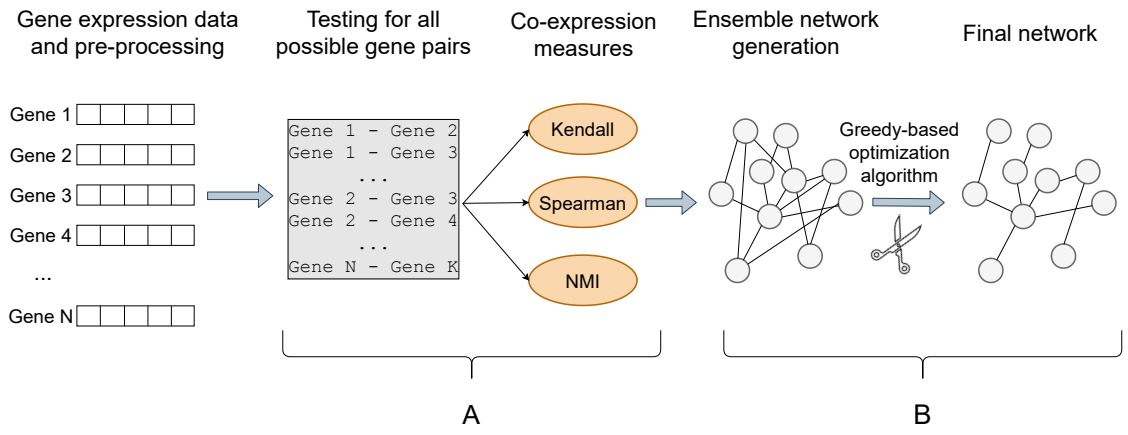


Figure 2: Global workflow of EnGNet for GCNs reconstruction, adapted from Gómez-Vela et al. [27]. The two main steps are represented. (A) First an ensemble strategy was used for network inference, (B) then a greedy-based approach was performed for the final optimization.

EnGNet has demonstrated its usefulness in the modeling of diseases like PTSD and COVID-19 [28, 27], highlighting the ability of the method to generate networks strongly related to their corresponding pathology. Moreover, EnGNet achieved good results against other state of the art methods [29]. Inferred-networks were smaller in size regarding the number of edges. Sparseness and scale-free topology arise as a major convenience of EnGNet networks.

Nevertheless, the method can be improved since some limitations were present. First, using two linear (Kendall, Spearman) and only one non-linear (NMI) correlation measure reduced the possibility to detect non-linear interactions when major voting was applied. A wider range of interactions can be detected using more correlation measures. Second, the method did not automatically selected thresholds for the intermediate co-expression networks, using predefined criteria. Consequently, results were not specific for each dataset and the algorithm can be modified so that both ensemble step and thresholding return optimal values. Moreover, also hubs identification can be improved using better criteria and their removed interactions can be added without fixed thresholds, but using the values that return optimal results for each dataset.

This work describes the implementation of EnGNet 2.0, a new algorithm for GCN

reconstruction. Starting from original EnGNet idea, EnGNet 2.0 algorithm also consisted of two main-steps, namely an ensemble strategy followed by an heuristic pruning. EnGNet 2.0 aims to detect a wide range of interactions using co-expression measures that can identify both linear and non-linear ones. Moreover, EnGNet 2.0 has the goal to obtain results that are specific for each input dataset. Unlike original EnGNet, that used fixed thresholds, EnGNet 2.0 aims to increase the ability of the method to detect true biological interactions by selecting thresholds that maximize the results. The purpose of the algorithm is to generate sparse and scale-free networks in which both the nodes and the edges are biologically significant.

EnGNet 2.0 results were compared to the ones obtained with other methods to verify its competitiveness. This comparison highlights the usefulness of the proposed method and the goodness of its results becomes evident in a biological application with a specific human disease. In this study, the pathological condition was Atopic Dermatitis (AD), a skin inflammation related with a strong immune response [30] with evidence for autoimmune mechanisms [31]. T-cell and leukocyte activation were described as relevant processes in AD [32, 33], resulting in mechanisms that involve cytokine [34]. Selected data were obtained from the study by Möbus et al. [35] which aimed to a deeper understanding of the AD skin transcriptome. This dataset contained gene expression levels from normal and pathological samples, allowing to obtain two final comparable networks. The comparison between normal and pathological results can provide more information about the role of specific genes in the molecular mechanisms in AD.

2 Materials and methods

In the following subsections the main procedures and dataset that were used in this work are described. First, subsection 2.1 describes the original dataset used for the biological validation of the proposed method and subsection 2.2 explains how these data were pre-processed. Then, subsection 2.3 describes how the differentially expressed genes in the comparison between normal and pathological conditions were identified. After this, subsection 2.4 clarifies how different validation steps were performed and subsection 2.5 describes the implementation of

the algorithm. At the end, subsection 2.6 explains the final analyses performed on the outputs.

2.1 Original dataset

The original dataset by Möbus et al. [35] was selected for the biological application of *EnGNet 2.0*. This work aimed to a deeper understanding of the atopic dermatitis (AD) skin transcriptome in *Homo sapiens*, in particular they were focused on the effects of systemic treatment with dupilumab and cyclosporine over gene expression in AD. For this purpose, they conducted a gene expression study of AD using mRNA-Seq data generated by high throughput sequencing with Illumina HiSeq 3000 platform. Intrapersonal lesional (AL) and non-lesional (AN) skin biopsies (4 mm) were collected from 57 patients, AN samples were taken at least 5 cm from the active lesion. Data were collected both prior to the initiation of a systemic therapy (time 0) and 3 months after the initiation of a therapy with dupilumab and cyclosporine.

The original RNA-Seq count table and additional files were retrieved from the Gene Expression Omnibus (GEO) database (ID: GSE157194). Data were obtained from GEO using the *GEOquery* R package [36]. Only data collected at time 0 were used in this project, resulting in a dataset with 111 samples and 43323 genes for each of them.

2.2 Data pre-processing

Exploratory analyses were performed to see a priory differences between samples according to the gene expression pattern for AL and AN. These analyses included unsupervised methods as hierarchical clustering, PCA and MDS plot since they are a good way to get an understanding of the sources of variation in the data. In particular, MDS (Multi Dimensional Scaling) calculates the distances between data and these values are used to map the points in a space with lower dimensions than the original dataset [37].

After exploratory analyses, low-expression genes were removed because RNA-Seq experiments usually screened the whole genome and expression levels are obtained also for those genes that are low-expressed or completely silent. These

genes were not biologically relevant because, if they were not expressed at a biologically meaningful level in any condition, they were not of interest [38]. They were not relevant also from a statistical point of view since their removal reduced the number of statistical tests that need to be carried out in downstream analyses. For these reasons, low-expression genes were identified and removed using the pipeline suggested by Law et al. [39], using the *edgeR* R package [40].

Once low-expression data were removed, data normalization was performed. Data need to be normalized since they belong to different samples and external factors often affect their expression during the experimental procedures. Given this variability, normalization is required to ensure that the expression distributions of each sample are similar across the entire experiment. In this project, CQN (Conditional Quantile Normalization) normalization was used. CQN is a recommended method for RNA-seq data normalization because it is able to take into account GC-content. This is relevant because guanine-cytosine content has a strong sample-specific effect on gene expression measurements that, if left uncorrected, leads to false positives in downstream results [41].

2.3 Differential expression analysis

Differentially expressed genes (DEGs) are those genes whose expression was more different between pathological and normal conditions. DEGs were supposed to play a major role in the involved processes [42], and their identification also allowed to obtain a handy dataset that simplified subsequent analyses.

Consequently, genes that were differentially expressed between AL and AN conditions were identified and only DEGs were taken into account for subsequent steps. Their identification was performed using the *Limma* R package, a well-known package that is commonly used for this purpose. As suggested in the original *Limma* article by Ritchie et al. [43], *Voom* method was used to detect DEGs using a minimum log2FC of 2 and an adjusted p-value of 0.05 as the discriminant thresholds [44].

2.4 Validation

The algorithm implementation presented some steps during which different networks had to be validated and compared. Consequently, Gold Standard (GS) was necessary for these validation steps. The algorithm needed a dataset of verified interactions to be compared with the results to find which network is the one that better identifies true relationships as the discriminant thresholds vary.

STRING database was used for this purpose since it is a biological database of known and predicted protein–protein interactions which contains information from numerous sources, including experimental data, computational prediction methods and public text collections [45]. GS was obtained keeping only the co-expression interactions for *Homo sapiens* that contain those genes that were identified as DEGs in the previous step (Subsection 2.3).

Before proceeding with the comparison, ENGS entries were converted in the correspondent ENSP ones since STRING entries were annotated with this code. The *STRINGdb* R package [46] was used for this purpose.

Once both GS and the output data were annotated with the same code, Receiver Operating Characteristic (ROC) curves were used for the validation. ROC curve describes the ability of a binary classifier system when its discrimination threshold varies. ROC curve plots sensitiviy against specificity at various threshold settings. These parameters are calculated as:

$$\text{sensitivity} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \frac{\text{TN}}{\text{N}} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Sensitivity (recall) represents the number of true positives over the total number of positives (true positives and false negatives). In this project, it calculates the fraction of of interactions in the experimental data that were also present in the GS over the total number of interactions with a weight greater than the threshold. Specificity indicates the number of true negatives over the total number of negatives (true negatives ans false positives).

ROC curves were used to estimate a statistical parameter called AUC (Area Under the Curve) which represents the area under the ROC curve. This value describes how much the model is capable of distinguishing between classes and higher AUC correspond to better models.

In this project, interactions that need to be validated presented two main attributes: the value of the correlation and a TRUE/FALSE classifier that represent the presence/absence of this interaction in the GS. Using these data, the *pRoc* R package [47] was used to create ROC curve and to calculate AUC values. At each validation step the network that returned the higher AUC values was selected as the best choice.

ROC curve was also used to compare final AN and AL networks with results obtained with different methods to test EnGNet 2.0 performance. For this comparison, networks were created using four single co-expression measures and EnGNet 1.0 algorithm [27]. The thresholds used to generate these networks were the same calculated with log-log method during the algorithm implementation (Section 2.5). At the end, five AL networks and five AN networks were used to validate EnGNet 2.0 results.

2.5 Inference rationale

The main part of the project is the implementation of a new algorithm for GCN reconstruction. Starting from the original EnGNet idea, EnGNet 2.0 algorithm also consisted of two main-steps, as shown in Figure 3. First, an ensemble strategy with a major voting technique was adopted to increase the range of detected interactions, combining linear and non-linear measures. Then, an optimization steps aimed to obtain scale-free networks removing all those interactions that were not fundamental for the overall connection.

2.5.1 Ensemble strategy

First optimization step, described in Figure 4, started using CQN normalized data as input. Normalized data were used to calculate four different correlations values for each pair of DEGs: Kendall, Spearman, Normalized Mutual Information (NMI)

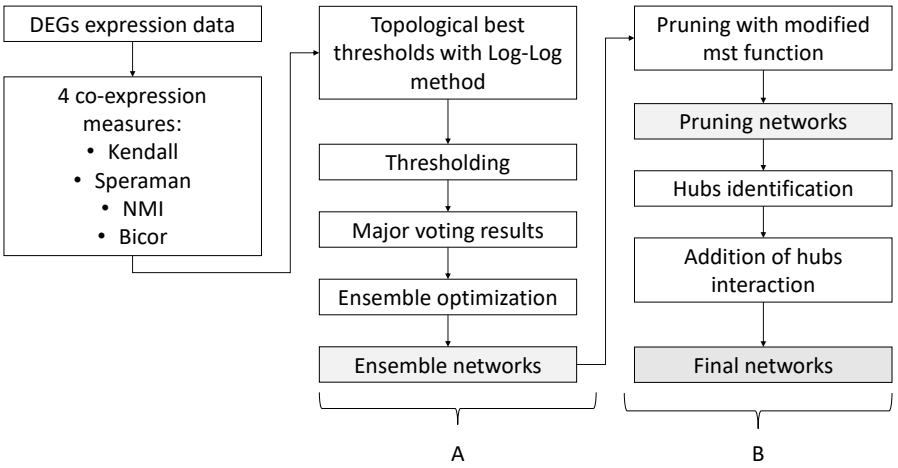


Figure 3: Pipeline of EnGNet 2.0 algorithm. This scheme show the two main EnGNet 2.0 steps. (A) First an ensemble strategy was used, (B) then an heuristic pruning was performed for the final optimization.

and Biweight Midcorrelation (Bicor).

Kendall and Spearman are well studied methods that use ranking-based approaches for linear correlations identification [12, 48]. The *psych* R package [49] was used for their estimation.

Kendall's coefficient is defined as:

$$\tau = \frac{n_c - n_d}{\binom{n}{2}} \tag{1}$$

where n_c and n_d are the number of concordant and discordant pairs between samples after ordering.

Spearman is calculated as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2}$$

in which d_i is the difference between the ranks of corresponding variables.

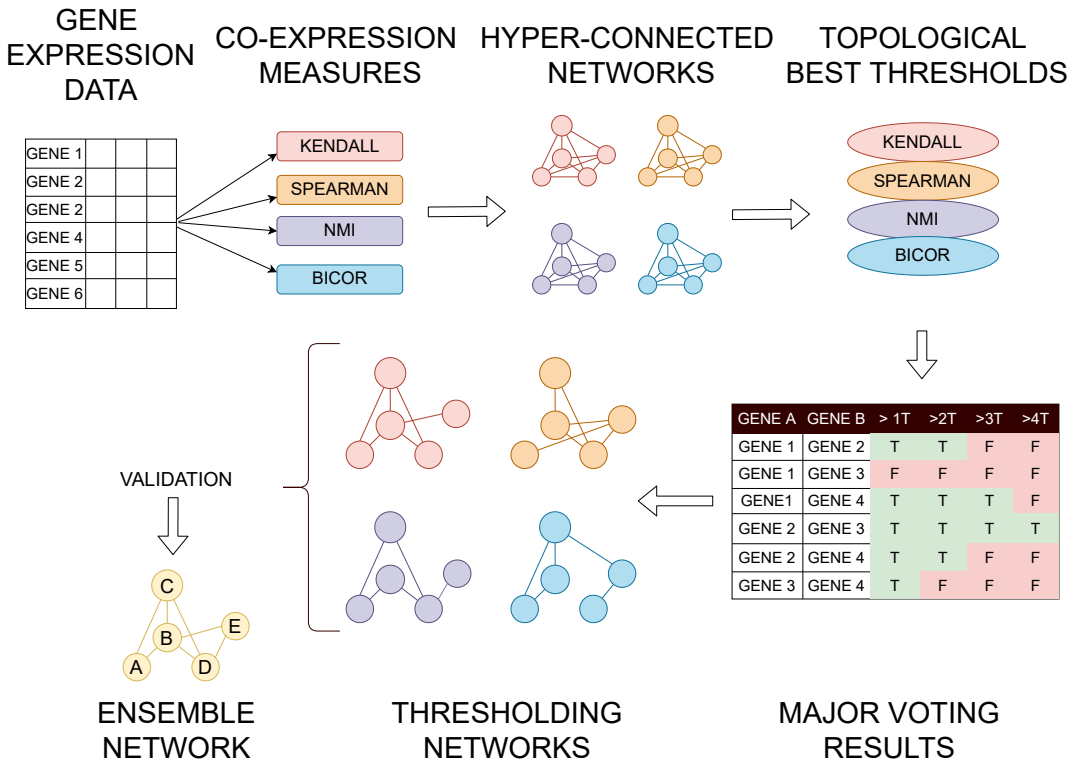


Figure 4: This figure explains how the first part of the algorithm works. Gene expression data were the starting input of the method. Four correlation scores were calculated for each pair of DEGs and hyper-connected networks were obtained. Networks needed to be optimized and the first optimization used thresholds based on topological criteria. Major voting technique was employed to obtain thresholding networks. Final validation with GS allowed to identify the best choice for this first optimization returning an optimized *ensemble network*.

NMI and Bicor were used to detect non-linear dependencies between genes since their usefulness in GCN reconstruction was already tested [15, 50]. They were applied using the *WGCNA* pacakge [51].

Mutual information is defined as:

$$I(X;Y) = D_{KL}(P_{(X,Y)} \| P_X \otimes P_Y) \tag{3}$$

where $P_{(X,Y)}$ is the joint distribution of the variables, P_X and P_Y are the marginal distributions and D_{KL} is the Kullback–Leibler divergence [52].

Biweight midcorrelation is calculated as:

$$\text{bicor}(x, y) = \sum_{i=1}^{m} \tilde{x}_i \tilde{y}_i \tag{4}$$

where $\tilde{x}_i$ and $\tilde{y}_i$ are the weight normalized using median and median absolute deviation.

After the calculation of the correlation values, thresholds were selected to obtain the optimal topological features (scale-freeness) of the networks. Logarithmic method was used to identify those thresholds that allowed to obtain networks whose degree distribution was more similar to a power-law distribution [53]. For this reason, an iterative algorithm screened the range 0-1 every 0.1 and, at each value, a network with all the interactions with an absolute value of the correlation greater than the threshold was created. As shown in Figure 5, power-law functions become linear when the logarithm of both axes is plotted. Consequently, the algorithm checked the degree distribution of the resulting networks and a linear model was fitted against the *log10(degree)* vs *log10(relative frequencies)*. At each step, adjusted R^2 was calculated. This parameter represents the goodness of the model and the network that returned the higher value was selected as the best. Four best thresholds were obtained repeating this process for each correlation and four networks were generated after thresholding.

A validation step was necessary at this point since the algorithm did not know *a priori* how many correlation values above the thresholds were necessary to mark an interaction as relevant. For this reason, all the possible T/F combinations were validated as explained in subsection 2.4. Results were grouped according to the number of TRUE in the combination, then AL and AN data were combined to identify the group that returned the higher AUC. Best result was selected for subsequent analyses, returning two new datasets (one for AL and one for AN samples) with a decreased number of interactions. These results were used to create the *ensemble networks*. In these networks, the weight of the edges was calculated as the average between the values greater than the thresholds.

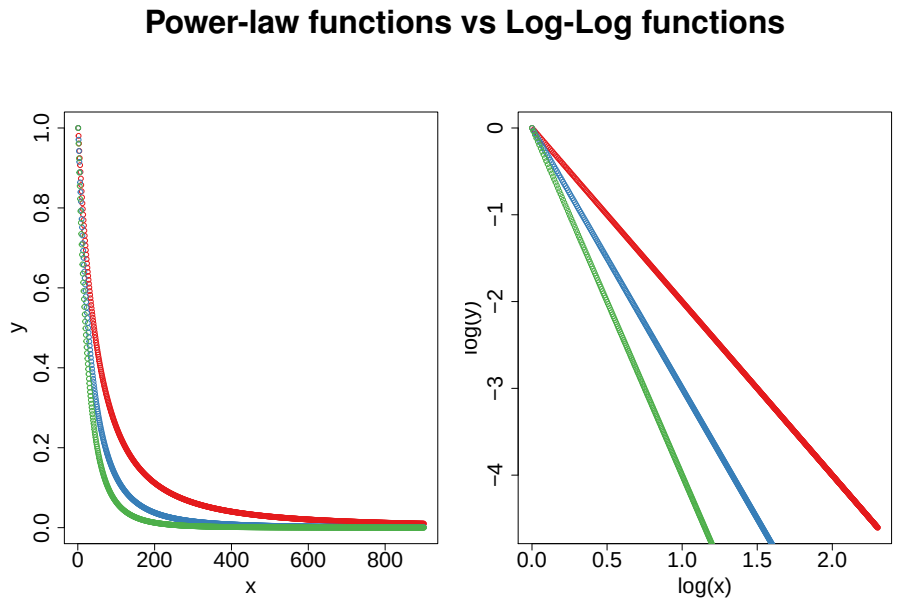


Figure 5: This figure shows on the left how different power-low functions look like. The graph on the right shows the same functions when log10(x) and log10(y) are plotted

2.5.2 Pruning optimization and hubs analysis

The second main step of the algorithm, which is depicted in Figure 6, was the pruning that aimed to obtain scale-free network. The optimization idea was to iteratively remove the edges from the one with the lowest weight to one with the highest. When an edge was removed, the network was screened to find if a node was disconnected or if subgraphs were created. If both these condition were false, the edge was removed, otherwise it was added again.

The *igraph* R package [54] was used for automatically optimize the networks since it provides a function that calculates the *minimum spanning tree* (MST) [55]. Original MST algorithm was modified because it generates the network with the minimum total weight while the EnGNet 2.0 needed to maximize it. Consequently, *pruned networks* were the ones that maximized the total weight of the edges with the minimum number of interactions that allowed to preserve the overall connection.

These networks had a scale-free topology in which hubs had a central role because they were highly-connected nodes, consequently they controlled network

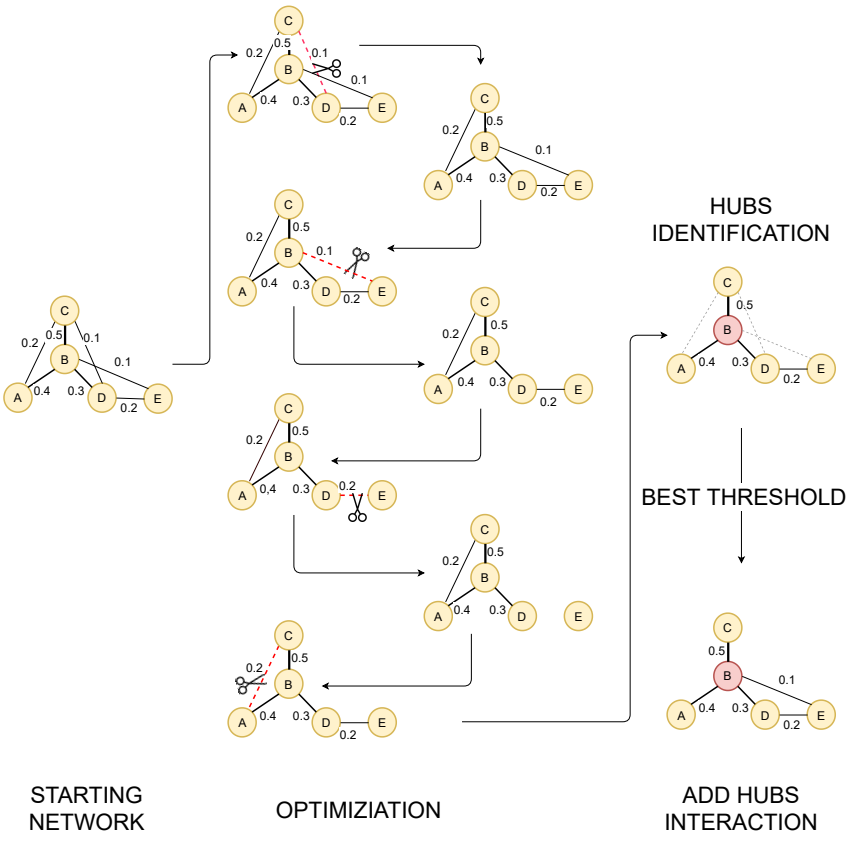


Figure 6: This figure explains how the second optimization step works. The edges of *ensemble networks* were first ordered from the minimum to the maximum weight. Then, they were removed following the previous order and after each removal the networks was checked to find if a node was disconnected or if subgraphs appeared. The edge was re-added if one of these conditions was true. After the pruning, hubs were identified and removed hubs interactions greater than the best threshold were re-added.

processes [56]. For this reason, their interactions were supposed to be more relevant and the algorithm tried to reintroduce removed connections that involved these nodes [6]. First, hubs were identified as those nodes whose degree was higher than the third quartile in the degree distribution of the network. Then, the range from 0.5 to the maximum weight in *ensemble networks* was screened every 0.01. At each value, those removed interactions that involved the hubs and whose weight was greater than the threshold were re-added. Networks obtained with all the thresholds in the selected range were validated as explained in subsection

2.4 and the one that returned the highest AUC was selected. After this last step two *final networks* were obtained, one for AL and one for AN data.

Finally, the complete algorithm was able to take as input AL and AN datasets with normalized DEGs counts and two validated networks were returned as final output.

2.6 Final networks analyses

In this work, two final networks were obtained using EnGNet 2.0, one for AL data and one one for AN data.

These networks were analyzed with cluster and Gene Ontology (GO) analyses to highlight the main biological differences between them. Cytoscape [57] was used to identify clusters in these networks. GLay is a community clustering algorithm implemented on Cytoscape that uses the weight of the edges to associate each node in the network to a cluster [58]. Genes that belonged to the biggest cluster were selected for enrichment analysis since they were the ones more involved in the network processes [59].

Enrichment analysis was performed using Cytoscape's plugins ClueGO [60] which shows over-represented GO-terms in a set of genes. The results provided information about the biological processes represented by the identified clusters in AL and AN networks.

3 Results

In the following subsections, the results obtained in the biological application with the original dataset by Möbus et al. [35] are described. First, subsection 3.1 shows data characteristics highlighted by exploratory analyses and data modifications after pre-processing. After this, the results for DEGs identification are analyzed in subsection 3.2. Ensemble and pruning steps are described in subsection 3.3, then subsection 3.4 show the results obtained after comparison with other methods. In the end, the characteristics of the final outputs of the algorithm are described in subsection 3.5.

3.1 Data pre-processing

Exploratory analyses were performed on raw data without pre-processing to highlight *a priori* differences between AL and AN samples.

First, MDS plot was obtained and the results shows that samples belonging to different groups had different characteristics, as highlighted by Figure 7. In this figure, blue points represent AN samples and they are quite homogeneous. In reverse, red points indicates AL samples and they show more variability. Observed differences between normal and pathological groups confirmed the reliability of the method because in subsequent steps, the algorithm separated the original dataset according to the AL/AN condition, aiming to obtain two different networks.

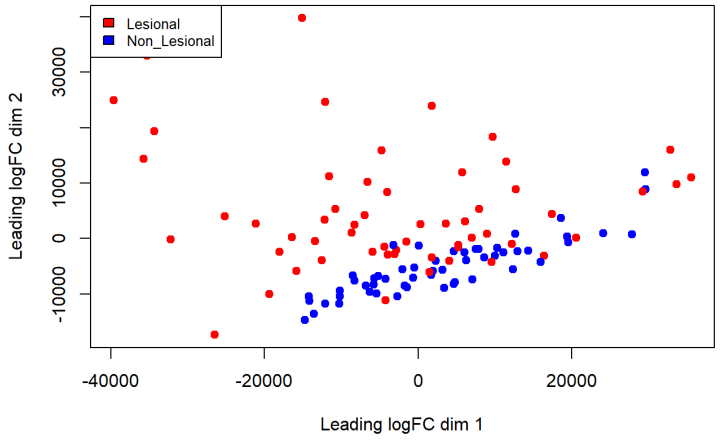


Figure 7: This figure shows the MDS plot obtained with the *limma* R package [43]. The points are colored according to AL/AN condition. It is possible to observe that Non-Lesional samples are quite homogeneous, while the Lesional ones show more variability.

After exploratory analyses, low-expressed genes were removed. Group membership of each sample was taken into account because this project focused on the differences between AL and AN groups. Consequently, genes low-expressed in a group but with normal expression in the other were not removed since they were relevant for the purpose of this work. Before removing genes, count-per-million (CPM) normalization was applied to remove differences in library sizes between

samples [39]. Dataset dimensions decreased from 43223 to 17301 genes for each sample.

3.2 Differential expression analysis

Differences between AL and AN samples were highlighted by DE analysis. A p-value of 0.05 and a minimum log2FC of 2 were used as cutoff to identify those genes whose expression changed the most between the two conditions. These parameters ensured the robustness of the results both from a statistical (p-value) and a biological (log2FC) point of view.

Using these values, 680 DEGs were identified. Among these, 498 were up-regulated and 182 were down-regulated. Figure 8 shows these results in a graphical way. Red points represent DEGs, they are in the areas of the plot in which both -log(P) and Log2FC are above the thresholds. Green and blue points represent genes that satisfies only one condition: green points are the ones with a significant log2FC but with not significant p-value, the opposite for blue ones. Black points represent genes with both the parameters under the thresholds.

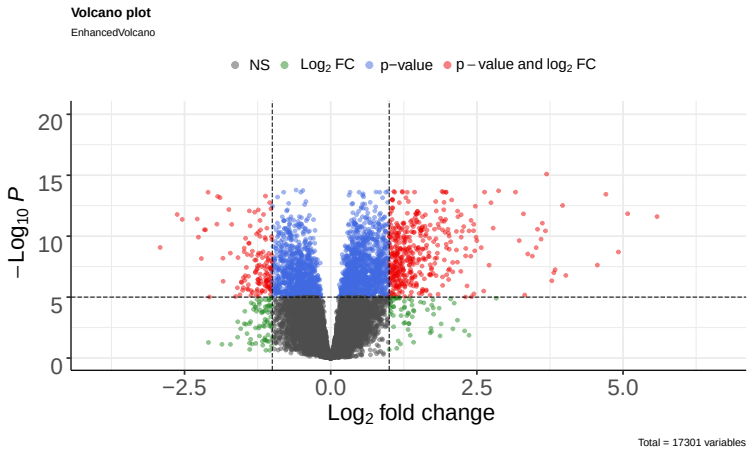


Figure 8: This figure was obtained using the *EnhancedVolcano* R package [61]. Each point in the graph represents a gene. DEGs are the red ones since they are those genes with both log2FC and p-value above the threshold.

3.3 Reconstruction inference

3.3.1 Ensemble strategy

As it was described before, four co-expression measures were calculated for each pair of DEGs. These results created four hyper-connected networks that needed to be optimized.

First, thresholds based on topological criteria were identified for the ensemble step. For each measure, different networks were created using different thresholds in the range 0-1 and the degree distributions of these networks were compared to a power-law distribution. Networks that better preserved scale-free property were selected, they were generated using thresholds summarized in Table 1. NMI values were distributed in a smaller range compared to others measures, this was the reason why NMI best thresholds were lower than the others.

	Kendall	Spearman	NMI	Bicor
Lesional	0.5	0.7	0.3	0.7
Non-Lesional	0.4	0.6	0.3	0.6

Table 1: Best topological thresholds identified with Log-Log method.

Once each measure was processed individually, networks obtained after thresholding needed to be combined with an ensemble technique to identify those interactions that were more relevant in the dataset. Major voting technique was used for this purpose, however the algorithm did not know *a priori* the minimum number of measures above the thresholds that were necessary to select an interaction as relevant A validation step was performed and all the possible T/F combinations were tested. AL and AN results were combined and they were grouped according to the number of TRUE classifiers present in the combination. In the end, the group with the highest AUC median was selected as the best choice. For this dataset, the best choice was to select as relevant the interactions with 4 values above the threshold, as shown in Figure 9.

Ensemble networks were created with this strategy and the weight of the edges was calculated as the average of the measures grater than the thresholds. The weight could be calculated also choosing the maximum between the four val-

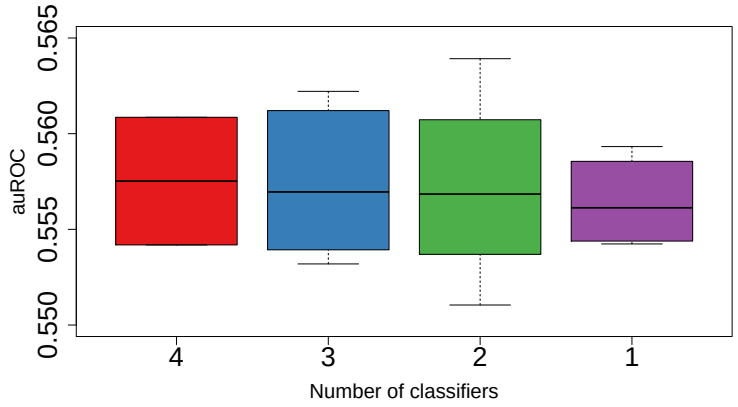


Figure 9: These boxplots show the validation results with GS for all the possible T/F combinations. AL and AN results were combined and they were grouped according to the number of TRUE classifiers. The combination with 4 TRUE (the red one) was the one with the highest median of the AUC values.

ues, however the average method allowed taking into account both linear and non-linear measures because NMI values were smaller than others and their contribute would be lost choosing the maximum. Figure 13 shows how hyper-connected networks were significantly reduced in dimensions with this first optimization.

3.3.2 Pruning optimization and hubs analysis

Ensemble networks were optimized a second time to decrease their dimensions preserving scale-free topology. First, the algorithm sorted all the interactions according to the absolute value of the weights. Then, edges were removed starting from the one with the minimum absolute value. Every time an edges was removed, the network was checked to find if a node was disconnected or if subgraphs were created. If both these condition were false, the edge was removed, otherwise it was added again. This optimization removed the lowest interactions that did non affect the overall connection and *pruned networks* were created.

Most of the interactions were removed with this method, as shown in Figure 13, however Figure 12 highlights that AUC values decreased in AL network after this step. The reason was that some of the removed interactions were biologically

relevant and they needed to be re-added even if they were not fundamental for the overall connection. To overcome this problem, hubs were identified using node degree as discriminant parameter. Degree distribution of *pruned network* was checked and the nodes whose degree was above the third quartile were selected as hubs [62]. 62 out of 296 nodes were identified as hubs in the Lesional network, 43 out of 260 nodes were identified as hubs in the Non-lesional one.

Once hubs were identified, the algorithm screened a range of thresholds to select the one that allowed to add interactions preserving statistical significance. Different networks were created and validated with GS using thresholds in the range from 0.5 to the maximum weight in *ensemble network*. These networks contained all the interactions preserved by the second optimization and the removed hubs interactions with an absolute value of the weight greater than the threshold. For each of them, AUC values were calculated using GS and the results are showed in Figure 10. In both AL and AN networks, the addition of interactions increased the AUC since the curve reached a global maximum. After this maximum, the curve decreased until it reached the same AUC obtained with *pruned networks*. Networks correspondent to the global maximums were selected as the final output of the algorithm to generate *final networks*.

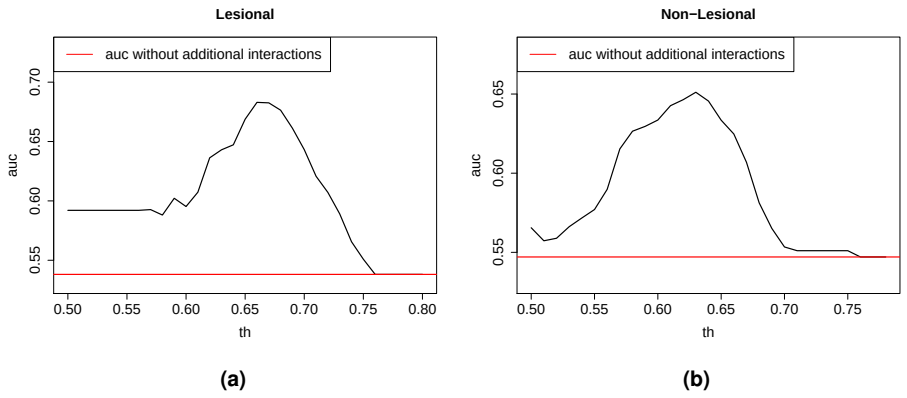


Figure 10: These plots represent the AUC values obtained adding hubs interactions at different thresholds. Lesional (a) and Non-lesional (b) networks show similar results. In both cases the addition of interactions increase the AUC since a global maximum is reached. After this value the curves decrease until they reach the AUC obtained with *pruned networks*

In the end, *final networks* were compared to the ones generated previously to

verify the reliability of the algorithm. As shown in Figure 11 and Figure 12, both the average weight of the edges and the AUC values increased when compared to the previous networks generated during algorithm implementation. Moreover, Figure 13 describes how networks dimensions changed during different steps. All these results confirmed that *final networks* were the ones that returned the best performance and the best topological properties, highlighting the relevance of hubs analysis.

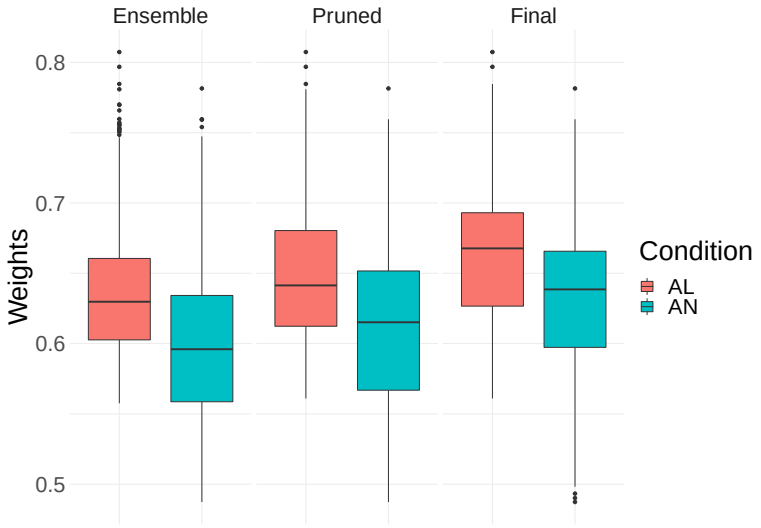


Figure 11: Distributions of the absolute value of the weights in the networks generated during the algorithm implementation. *Final networks* are the ones with higher values for both AL and AN data.

3.4 Performance comparison

Once *final networks* were obtained, they were compared with networks generated by other methods to test their performance. In particular, Five AL and five AN networks were generated for this comparison. Four networks were created using single co-expression measures presented and using the best topological thresholds calculated with Log-Log method (Section 3.3.1). For this purpose, Kendall, Spearman, NMI and Bicor values were calculated for each pair of DEGs and each measure was used individually to create a network that contained the correlations grater than the thresholds. The fifth network was created using EnGNet 1.0 algorithm [27] and the topological thresholds identified with Log-Log method

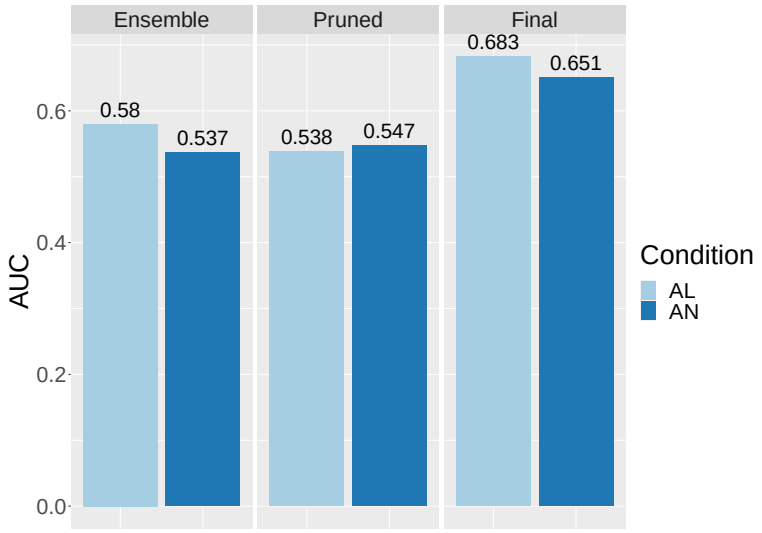


Figure 12: Barplots show AUC values returned by the networks generated during the algorithm implementation. *Final networks* are the ones with higher values for both AL and AN data.

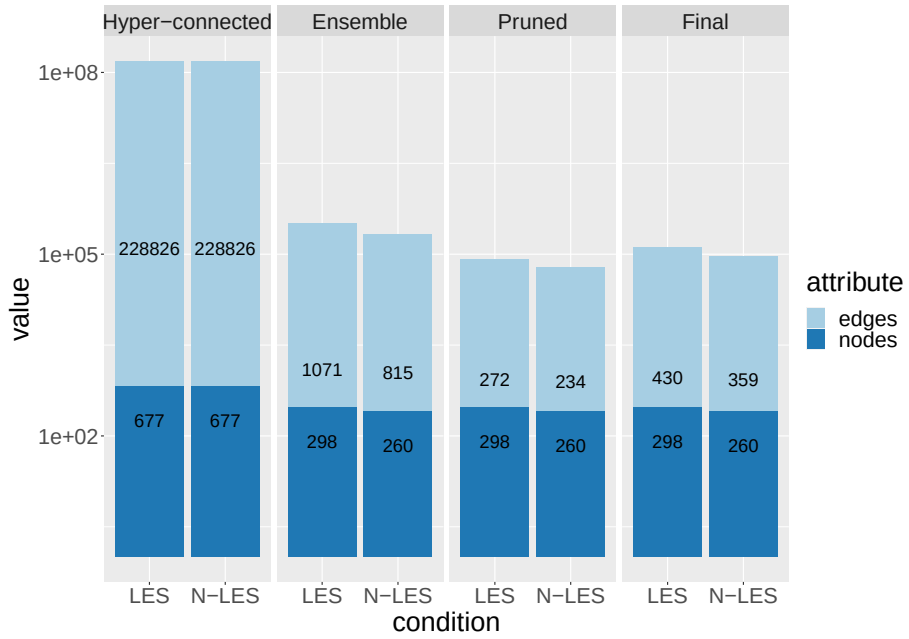


Figure 13: This graph summarizes the number of nodes and edges in the networks generated by the algorithm.

to perform a fair comparison.

Networks were compared according to their ability to detect interactions in a proper way and their topological properties. To do this, four parameters were calculated: number of nodes, number of edges, AUC value resulting from the validation with GS and R^2 value of the fitting with a power-law distribution. The higher value of AUC and R^2, the better is the network since AUC value indicates networks ability to identify true biological interactions and R^2 represent networks similarity with scale-free ones.

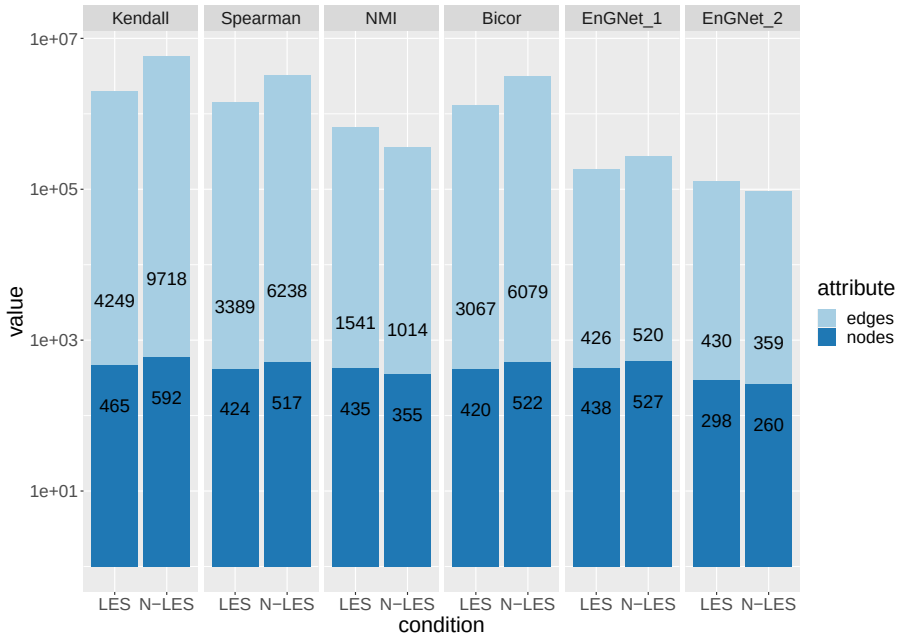


Figure 14: Number of nodes and edges in the networks generated for the comparison.

Figure 14 shows the comparison of the topological properties, highlighting EnGNet 2.0 ability to reduce networks dimensions compared to single-measure ones. When compared to EnGNet 1.0, EnGNet 2.0 networks returned a greater connectivity because the results have a similar number of edges, while EnGNet 2.0 contained less nodes. EnGNet 2.0 networks were more connected because the algorithm did not used a fixed threshold for re-adding hubs interactions, but a threshold was selected to maximize the results for this specific dataset. Consequently, optimal results were reached increasing the number of final interactions compared to EnGNet 1.0.

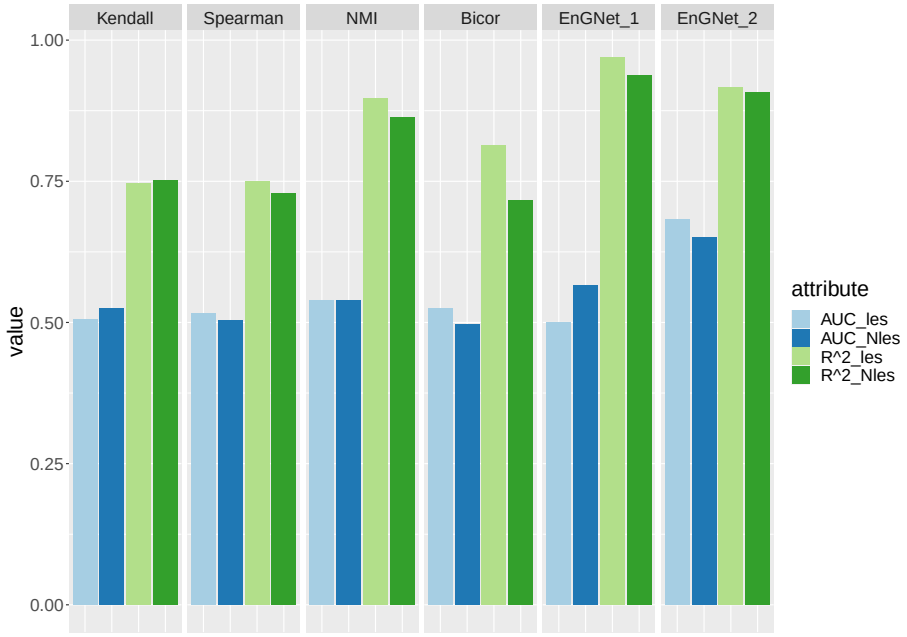
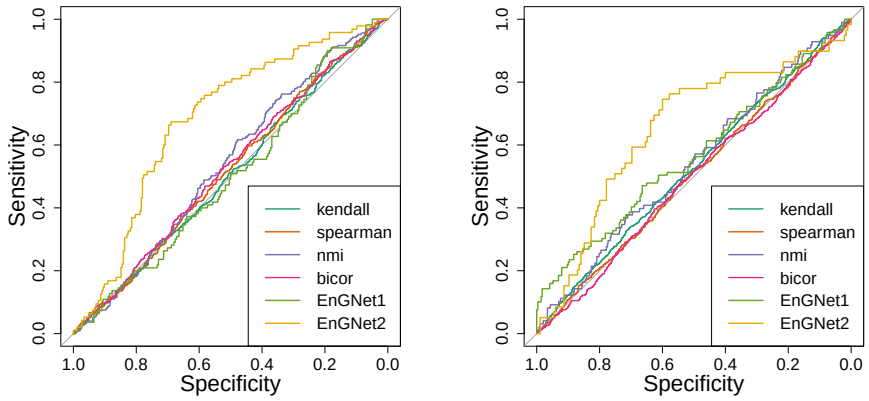


Figure 15: These barplots show AUC and R^2 results obtained from the networks used in comparison. EnGNet 2.0 networks return the higher AUC values for both AL and AN data. R^2 results are quite similar between EnGNet 1.0 and EnGNet 2.0, these methods overcome single-measure ones in the generation of scale-free networks.



(a) Lesional **(b)** Non-Lesional

Figure 16: ROC curves resulting from the comparison of different networks with GS. EnGNet 2.0 (yellow curve) overcomes all the other methods in the ability to discriminate between true and false biological interactions.

After the analysis of the dimensions, the performance of the networks were also compared and Figure 16 shows AUC and R^2 results in a graphical way. As ex-

pected, EnGNet 2.0 results overcome the networks obtained using single co-expression measures in both AUC and R^2. The two versions of EnGNet return similar R^2 results, however EnGNet 2.0 returns much higher AUC values, as highlighted also by ROC curves in Figure 16. In these graphs, curves resulting from single-measure methods and from EnGNet 1.0 follow the diagonal, while EnGNet 2.0 (yellow curve) curve differs from the others, indicating greater ability to detect true biological interactions.

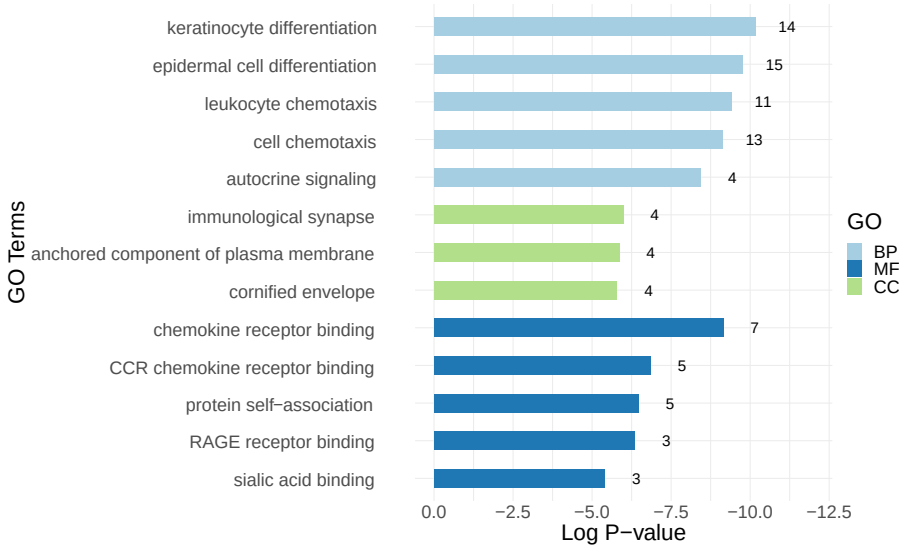
3.5 Final networks analysis

Final networks were analyzed to highlight topological and biological differences between AL and AN results.

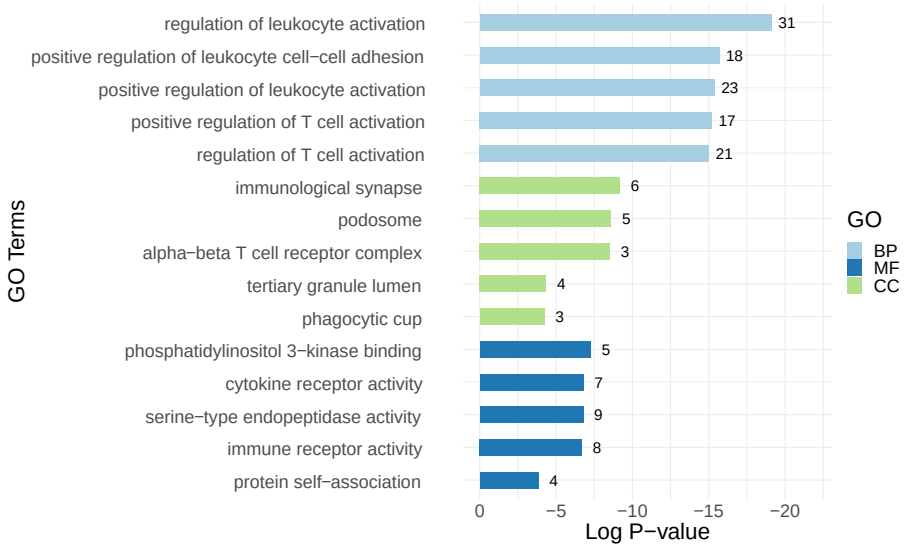
First, cluster analysis was performed to identify the main clusters in both the networks. Genes that belonged to the biggest groups were the ones more involved in the network processes, consequently they were the ones used for subsequent analyses. GLay cluster algorithm implemented in Cytoscape was used in this step as a plugin. Main clusters comprised 226 genes in the AL networks and 203 genes in the AN one.

After this, enrichment analysis was performed on the genes in the main clusters to identify over-expressed GO terms. The results are showed in Figure 17 in which GO terms with the lowest p-value are represented for each category (Biological process, Molecular function, Cellular component)

Different results between AL and AN were expected since AL genes expression levels were related to a pathological condition and the AN ones were associated to a normal situation. Most significant terms show that GO results are consistent with the expectations. Moreover, AL GO terms returned lower p-value and they were characterized by an higher number of genes compared to AN results, indicating that AL results were more specific. The two main BP GO terms associated to AN network are ”Keratinocyte differentiation” and ”Epidermal cell differentiation”. These terms are related to general epidermal functions [63] without a relationship with the pathological condition. In reverse, BP GO terms over-expressed in AL network are ”Regulation of leukocyte activation”, ”Positive regu-



(a) Non-Lesional



(b) Lesional

Figure 17: These bar plots represent the most significant GO terms for the main clusters of AL (a) and AN (b) networks. Each bar represent the Log(p-value) of a GO terms and the number near the bar indicates how many genes are associated to that term.

lation of leukocyte cell–cell adhesion" and "Positive regulation of leukocyte activation". These terms are associated with processes that involved leukocyte activation which is a situation that characterized pathological conditions [64]. "Positive regulation of T cell activation" and "Regulation of T cell activation" were also BP

over-expressed in AL networks, confirming the relationship between results and AD mechanisms [33].

4 Discussion

EnGNet 2.0 aimed to reconstruct Gene Co-expression Networks (GCN) starting from gene expression data. Gene expression levels obtained from normal tissues (AN) and tissues affected by Atopic Dermatitis (AL) were selected as input for the algorithm implementation [35].

In particular, Atopic Dermatitis (AD) is a common disease, characterized by skin inflammation related with a strong immune response [30]. A study by Mittermann et al. [31] describes evidence for autoimmune mechanism and Oyoshi et al. [65] explain the contributions of the innate and adaptive immune system to the pathogenesis of AD, highlighting the role of different cell types. AD skin lesions are characterized by a dermal infiltrate consisting of keratinocyte-derived cytokines, activated T cells and different types of leukocytes such as eosinophils, mast cells, and macrophages that may be coated with IgE.

4.1 Exploratory analyses revealed differences between AL and AN samples, with large genetic up-regulation in Lesional genes

Exploratory analyses were performed on raw data to observe differences between normal and pathological samples. MDS and PCA analyses returned similar results. Both these methods highlighted differences between AL and AN groups before data pre-processing as shown in Figure 7. As expected, samples in the normal condition were quite homogeneous, while the pathological ones show major differences inside the pattern, indicating that AL dataset was characterized by AD pathology at different levels.

After exploratory analyses, the removal of low-expressed genes reduced data dimensions. The presence of noisy, low-expressed genes decreases the sensitivity of detecting DEGs. Thus, identification and filtering of these genes improve DEGs detection [66]. DEGs identification returned 680 genes. Among them, 498 were up-regulated and 182 were down-regulated, highlighting genes activation induced

upon AD. This biological situation was consistent with the study by Nedoszytko et al. [32], which describes epigenetic changes in AD that cause genes activation.

Once DEGs were identified, they were selected as the starting input for the algorithm implementation. GCNs were reconstructed using exclusively DEGs because the algorithm aimed to obtain results that were specific for the selected disease. DEGs identification for GCN reconstruction is a well-known method in the literature [28, 67]. This approach enabled the modeling of the genetic relationships that are considered of relevance in the AL and AN comparison. Those genes whose expression was more different between normal and pathological conditions were supposed to play a major role in the involved processes.

4.2 The novel algorithm EnGNet 2.0, designed and applied to this dataset, yielded the best results for GS comparison and topological properties

Input data contained 680 genes for both AL and AN datasets. The algorithm aimed to generate networks with a reduced number of genes and interactions, while preserving real biological relationships. Hereby, reconstruction of AL network is discussed, the same was performed for control samples.

4.2.1 Ensemble strategy

First, Kendall, Spearman, NMI and Bicor values were calculated for each pair of DEGs. All the results were in the range -1/+1, however different measures had different distributions. In particular, NMI values were distributed in a smaller range compared to Kendall, Spearman and Bicor ones. These differences had to be taken into account to avoid penalizing non-linear correlations. As long as different co-expression measures were considered individually, this unbalance did non affect the results. When different co-expression results were combined, the algorithm overcome these differences by using the average of the four values as the final weight.

Once co-expression values were calculated, topological thresholds based on the degree distribution of the networks were used to generate the *ensemble network*. Since the method started from hyper-connected network, this optimization aimed

to reduce dimensions removing all those interactions that were clearly not relevant. However, the algorithm had to avoid too stringent criteria because a more rigorous pruning will be performed later. For these reasons, a logarithmic method was used even if linear fit on the log-log scale is sometimes biased [68] because there are some limitations in the fitting with the tail of the distribution. Despite log-log limitations, this method identified those thresholds that allowed to obtain networks whose degree distribution was more similar to a power-law distribution [53].

After thresholding, each interaction was characterized by a T/F combination of four values. Consequently, 16 (2^4) different T/F combinations were possible and each of them characterized a different situation. Combinations in which both Kendall and Spearman were TRUE characterized linear correlations, while the ones in which both NMI and Bicor were TRUE identified non-linear correlations. However, there was not a unique way to chose which combinations identified relevant interactions and different data can give different results. The minimum number of TRUE necessary to mark an interaction as relevant was identified with a validation step. This validation compared statistical results of each combination for this specific dataset. AUC values returned by each combination were grouped by the number of TRUE, then AL and AN results were combined. Ensemble method performed on this dataset returned that the best choice for the major voting technique was to select only the interactions with four co-expression measures greater than the thresholds. For this dataset, a lower number of values above the thresholds corresponds to lower statistical results (Figure 9). This situation was coherent with the purpose to detect both linear and non-linear interactions.

Ensemble network was generated using the interactions that satisfied major voting condition. The weight of the edges was calculated as the average of the measures grater than the thresholds. Another possibility for the weight calculation was to choose the maximum between the four values, however the average method avoided penalizing non-linear correlations. This was because NMI values were smaller than the others and the average method kept their contribution into account.

4.2.2 Pruning optimization and hubs analysis

Network dimensions were reduced again to generate *pruned network* with a second optimization that preserved the number of nodes while removing edged that did not affect the overall connection. A modified version of *minimum spanning tree* algorithm was used for this purpose and *pruned networks* were generated as the ones that maximized the total weight of the edges while maintaining the minimum number of interactions that preserve the overall connection.

This step improved scale-freeness of the network, however statistical results decreased by removing biologically significant edges as shown in Figure 12. To avoid losing these interactions, hubs were identified and their removed connections were re-added to obtain *final network*.

This approach was selected to improve the results because *pruned networks* had a scale-free topology in which hubs had a central role. They controlled network processes, consequently their interactions were supposed to be more relevant and the algorithm tried to reintroduce them [6]. In *pruned networks* most of the nodes had one or two connections, consequently both the average and the median of the degree distribution were near to 2. Choosing the average or the median as the cutoff for hubs identification would have returned too many nodes as hubs. To avoid too stringent or too permissive criteria, the third quartile was chosen as cutoff for hubs identification. This method was specific for this set of data, a future version of the algorithm will allow the user to manually select the cutoff to distinguish hubs.

Before re-adding interactions, hubs were analyzed to verify if they were consistent with the disease. Hubs were ordered according to their degree and genes with highest degree in AL *pruned networks* were studied. CHP2 was the gene with the highest degree (10) and its role in the psoriasis mechanisms was already verified by Chen et al. [69]. KLHL16 was the second highest degree node (9) and also for this gene the literature confirmed its involvement in epidermal diseases, as described by Büchau et al. [70]. Then, different genes had a degree equal to 6. Among them, we highlight ABHD12B and HS3ST6, two genes that were already described as up-regulated genes in epidermal diseases [71, 72].

Compared to AL results, AN hubs were more related to common epidermal functions. The main hubs was BTC, a gene that belongs to the epidermal growth factor (EGF) family as described by Dunbar and Goddard [73]. Another hub was IL37 and the expression of this gene has a specific role to regulate the homeostasis of the epidermis [74]. Then, LCE5A was also identified as hub in AN network, De Koning et al. [75] described the role of this gene in the expression of cornified envelope structural proteins and keratinocyte differentiation-regulating proteins.

Consequently, differences between hubs in the two networks and the relationship between AL hubs and AD mechanisms confirmed the reliability of hubs identification to improve the results.

After hubs identification, their removed interactions were re-added to verify if this approach increased statistical results. EnGNet 1.0 re-adds all the hubs interactions greater than a fixed threshold, however this approach is not specific for each dataset. A range of possible thresholds was screened to maximize the performance of the final results. In particular, values from 0.5 to the maximum weight in *ensemble network* were screened every 0.01 units. This approach generated networks with a significant increase in statistical results (Figure 10), confirming the biological relevance of hubs interactions.

Properties of the different networks generated during algorithm implementation are showed in Figure 11 and in Figure 12. As expected, *final network* selected edges with higher weights. The algorithm aimed to identify most relevant interactions and these results were consistent with this condition. Regarding AUC results, *pruned network* returned lower values than *ensemble network*. A lot of interactions were removed during the second optimization, however, many of them were biologically relevant even if they were not fundamental for the overall connection. The re-addition of hubs interaction allowed to clearly increase statistical results. From a topological point of view, the comparison between different networks generated during algorithm implementation in Figure 13 highlighted the ability of the algorithm to generate sparse networks. These results were consistent with the known GCN properties [76].

4.2.3 Validation: EnGNet 2.0 outperformed individual co-expression measures and EnGNet 1.0

Final EnGNet 2.0 network was compared to the ones obtained using single co-expression measures and using EnGNet 1.0. The parameters used for this purpose were the number of nodes and edges, the AUC returned by comparison with GS and the R value returned by log-log fitting of the degree distribution.

EnGNet 2.0 network returned the best results for both AUC and topological parameters. Using this dataset, single-measure and EnGNet 1.0 networks returned a ROC curve that follows the diagonal of the plot (Figure 16). This means that these methods were not able to discriminate in a proper way between verified and not verified interactions. For this specific dataset, EnGNet 2.0 generated network that clearly exceeded single-measure ones both in the goodness of the identified interactions and in the capacity to create small networks with scale-free topology as shown in Figure 14 and in Figure 15. These results highlighted the advantages of using ensemble technique instead of single measure, allowing to detect in a proper way a wider range of interactions.

Figure 14 also shows that EnGNet 2.0 networks had a similar number of edges compared to networks generated by EnGNet 1.0. However, EnGNet 1.0 results contained less nodes, because the automatized process for additional interactions resulted in a greater connectivity for EnGNet 2.0 results. Moreover, the two versions returned similar values also in the comparison with power-law distribution. The main advantages of EnGNet 2.0 compared to EnGNet 1.0, were the possibility to detect both linear and non-linear co-expressions and the screening steps that allowed to avoid fixed thresholds. Consequently, EnGNet 2.0 was able to return results that maximized the performances for this specific dataset and AUC values were significant higher than EnGNet 1.0, as highlighted by both Figure 15 and Figure 16

It is important to take into account that these results are specific for the dataset used in this project. In the future, the method will be tested with other dataset to verify its reliability with different types of data. Moreover, comparison with results obtained with other GCN reconstruction methods will be performed in the future

to better assess EnGNet 2.0 performance.

In the end, the main limitation of the method is described. The algorithm was implemented using Atopic Dermatitis as pathological condition, however this disease does not have a specific database with verified interactions. Consequently, the algorithm needed a comprehensive database that covers a large variety of different conditions and STRING database was selected for this purpose. STRING database allowed to minimizes the loss of information since its co-expression values were based on gene-by-gene correlation tests across a large number of gene expression datasets [45]. This is a standard approach for this type of studies [77, 78] and its validity as reference was already tested [79]. However, we need to take into account that interactions were inserted in STRING using unsupervised methods, consequently the usage of this databsae could introduce some bias in the analysis.

4.3 GO-enrichment revealed the role of a strong immune response activation in Lesional samples

After statistical comparison with other methods, final results obtained with EnGNet 2.0 were analyzed to highlight biological differences between AL and AN networks.

This analysis aimed to verify if EnGNet 2.0 was able to generate networks with interactions related to the biological problem. Main clusters were identified with GLay algorithm, implemented in Cytoscape, resulting in two major component for both AL and AN networks. AL main cluster contained 226 out of 298 genes, AN main cluster contained 203 out of 260 genes.

Selected genes were the more connected ones, consequently they were supposed to be the most important in the regulation of the network processes. ClueGO was used to perform GO analysis on the main AN and AL clusters and the results are showed in Figure 17. GO Terms over-expressed in AL cluster showed lower p-valued and a greater number of associated genes when compared to AN results. This means that genes in AL network were strongly related to specific functions and processes.

In particular, the results highlighted the role of a strong immune response activation in the genes identified in the final AL network. BP comparison between AL and AN GO terms highlighted a strong relationship with processes related to immune response for AL genes, while AN results were related to tissue-specific processes. This is consistent with the biological knowledge about this pathological condition. Nedoszytko et al. [32] reveled the activation of genes affecting the regulation of immune response and inflammatory processes in AD and the role of T-cell activation in AD was studied by Akdis et al. [33]. Regarding MF results, a study by Wang et al. [34] highlights the role of mechanism that involve cytokine in AD and MF terms show a correspondence with the genes in AL cluster. These results confirmed that EnGNet 2.0 was able to generate networks with dimensions that make them easy to interpret and both the final nodes and edges were biologically significant.

5 Conclusions

This project introduces EnGNet 2.0 algorithm, an ensemble-based novel method for the inference of large gene co-expression networks.

First, EnGNet 2.0 applies an ensemble approach that combines topological thresholds and major voting technique. Second, an iterative pruning strategy optimizes both the size and topological features of the final networks.

When compared with networks generated using single co-expression measures and networks generated by EnGNet 1.0, EnGNet 2.0 final outputs were smaller in size regarding the number of both nodes and edges. Consequently, sparseness and scale-free topology are to be highlighted as a major convenience of this approach. Moreover, these features enable an easier interpretation and hypothesis-making by life scientists. In addition, EnGNet 2.0 networks also return the best results in terms of performance after comparison with Gold Standard. This means that the method is able to properly distinguish between true and false biological interactions.

At the end, the biological relevance of EnGNet 2.0 was successfully tested in the application to a specific human disease. EnGNet 2.0 final networks show

a strong relationship with the molecular functions and biological processes that regulate Atopic Dermatitis. In particular, leukocyte activation, T cells activation and cytokine activity emerge as the main processes regulated by the network obtained with pathological data.

These results demonstrate ability of EnGNet 2.0 to generate networks with both biological and topological optimal properties.

Acknowledgments

I would like to thank my supervisors, Professor Fernando M. Delgado-Chaves and Professor Francisco Gomez-Vela, for the time and the expertise they made available to help me during this project.

References

[1] Michael Baitaluk. System biology of gene regulation. In *Biomedical Informatics*, pages 55–87. Springer, 2009.

[2] Wei Tong. Analyzing the biology on the system level. *Genomics, proteomics & bioinformatics*, 2(1):6–14, 2004.

[3] Zheng Wang, Aditya Gudibanda, Ugochukwu Ugwuowo, Frances Trail, and Jeffrey P Townsend. Using evolutionary genomics, transcriptomics, and systems biology to reveal gene networks underlying fungal development. *Fungal Biology Reviews*, 32 (4):249–264, 2018.

[4] Paul A McGettigan. Transcriptomics in the rna-seq era. *Current opinion in chemical biology*, 17(1):4–11, 2013.

[5] Raya Khanin and Ernst Wit. How scale-free are biological networks. *Journal of computational biology*, 13(3):810–818, 2006.

[6] Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLoS Genet*, 2(6):e88, 2006.

[7] B. Jia and X. Wang. Regularized em algorithm for sparse parameter estimation in nonlinear dynamic systems with application to gene regulatory network inference. *Eurasip Journal on Bioinformatics and Systems Biology*, 2014(1), 2014. URL `www.scopus.com`. Cited By :3.

[8] Shupeng Gui, Andrew P Rice, Rui Chen, Liang Wu, Ji Liu, and Hongyu Miao. A scalable algorithm for structure identification of complex gene regulatory network from temporal expression data. *BMC bioinformatics*, 18(1):1–13, 2017.

[9] Bin Jia and Xiaodong Wang. Regularized em algorithm for sparse parameter estimation in nonlinear dynamic systems with application to gene regulatory network inference. *EURASIP Journal on Bioinformatics and Systems Biology*, 2014(1):1–15, 2014.

[10] Mina Moradi Kordmahalleh, Mohammad Gorji Sefidmazgi, Scott H Harrison, and Abdollah Homaifar. Identifying time-delayed gene regulatory networks via an evolvable hierarchical recurrent neural network. *BioData mining*, 10(1):1–25, 2017.

[11] Wentian Li. Mutual information functions versus correlation functions. *Journal of statistical physics*, 60(5):823–837, 1990.

[12] Sapna Kumari, Jeff Nie, Huann-Sheng Chen, Hao Ma, Ron Stewart, Xiang Li, Meng-Zhu Lu, William M Taylor, and Hairong Wei. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. *PloS one*, 7(11):e50411, 2012.

[13] Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, 13(1):1–21, 2012.

[14] Shoudan Liang, Stefanie Fuhrman, Roland Somogyi, et al. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific symposium on biocomputing*, volume 3, pages 18–29, 1998.

[15] Atul J Butte and Isaac S Kohane. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In *Biocomputing 2000*, pages 418–429. World Scientific, 1999.

[16] Adam A Margolin, Ilya Nemenman, Katia Basso, Chris Wiggins, Gustavo Stolovitzky, Riccardo Dalla Favera, and Andrea Califano. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics*, volume 7, pages 1–15. Springer, 2006.

[17] Aviv Madar, Alex Greenfield, Eric Vanden-Eijnden, and Richard Bonneau. Dream3: network inference using dynamic context likelihood of relatedness and the inferelator. *PloS one*, 5(3):e9803, 2010.

[18] Catharina Olsen, Patrick E Meyer, and Gianluca Bontempi. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009:1–9, 2008.

[19] Francisco Gómez-Vela, Carlos D Barranco, and Norberto Díaz-Díaz. Incorporating biological knowledge for construction of fuzzy networks of gene associations. *Applied Soft Computing*, 42:144–155, 2016.

[20] Narsis A Kiani, Hector Zenil, Jakub Olczak, and Jesper Tegnér. Evaluating network inference methods in terms of their ability to preserve the topology and complexity

of genetic networks. In *Seminars in cell & developmental biology*, volume 51, pages 44–52. Elsevier, 2016.

[21] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

[22] Sipko Van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in bioinformatics*, 19(4):575–592, 2018.

[23] Fernando M Delgado-Chaves, Francisco Gómez-Vela, Miguel García-Torres, Federico Divina, and José Luis Vázquez Noguera. Computational inference of gene co-expression networks for the identification of lung carcinoma biomarkers: An ensemble approach. *Genes*, 10(12):962, 2019.

[24] Wenying Yan, Wenjin Xue, Jiajia Chen, and Guang Hu. Biological networks for cancer candidate biomarkers discovery. *Cancer informatics*, 15:CIN–S39458, 2016.

[25] Renping Huang, Yang He, Bei Sun, and Bing Liu. Bioinformatic analysis identifies three potentially key differentially expressed genes in peripheral blood mononuclear cells of patients with takayasu's arteritis. *Cell Journal (Yakhteh)*, 19(4):647, 2018.

[26] Rui Zhong, Jeffrey D Allen, Guanghua Xiao, and Yang Xie. Ensemble-based network aggregation improves the accuracy of gene network reconstruction. *PloS one*, 9(11): e106319, 2014.

[27] Francisco Gómez-Vela, Fernando M Delgado-Chaves, Domingo S Rodríguez-Baena, Miguel García-Torres, and Federico Divina. Ensemble and greedy approach for the reconstruction of large gene co-expression networks. *Entropy*, 21(12):1139, 2019.

[28] Fernando M Delgado-Chaves, Francisco Gómez-Vela, Federico Divina, Miguel García-Torres, and Domingo S Rodriguez-Baena. Computational analysis of the global effects of ly6e in the immune response to coronavirus infection using gene networks. *Genes*, 11(7):831, 2020.

[29] Fernando M Delgado and Francisco Gómez-Vela. Computational methods for gene regulatory networks reconstruction and analysis: A review. *Artificial intelligence in medicine*, 95:133–145, 2019.

[30] Finn Schultz Larsen and Jon M Hanifin. Epidemiology of atopic dermatitis. *Immunology and Allergy Clinics*, 22(1):1–24, 2002.

[31] Irene Mittermann, Karl J Aichberger, Robert Bünder, Nadine Mothes, Harald Renz, and Rudolf Valenta. Autoimmunity and atopic dermatitis. *Current opinion in allergy and clinical immunology*, 4(5):367–371, 2004.

[32] Bogusław Nedoszytko, Edyta Reszka, Danuta Gutowska-Owsiak, Magdalena Trzeciak, Magdalena Lange, Justyna Jarczak, Marek Niedoszytko, Ewa Jablonska, Jan

Romantowski, Dominik Strapagiel, et al. Genetic and epigenetic aspects of atopic dermatitis. *International Journal of Molecular Sciences*, 21(18):6484, 2020.

[33] Cezmi A Akdis, Mübeccel Akdis, Dagmar Simon, Birgit Dibbert, Martina Weber, Stephanie Gratzl, Oliver Kreyden, Rainer Disch, Brunello Wüthrich, Kurt Blaser, et al. T cells and t cell-derived cytokines as pathogenic factors in the nonallergic form of atopic dermatitis. *Journal of investigative dermatology*, 113(4):628–634, 1999.

[34] Jing J Wang, Barbara JS Sanderson, and He Wang. Cyto-and genotoxicity of ultrafine tio2 particles in cultured human lymphoblastoid cells. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 628(2):99–106, 2007.

[35] Lena Möbus, Elke Rodriguez, Inken Harder, Dora Stölzl, Nicole Boraczynski, Sascha Gerdes, Andreas Kleinheinz, Susanne Abraham, Annice Heratizadeh, Christiane Handrick, et al. Atopic dermatitis displays stable and dynamic skin transcriptome signatures. *Journal of Allergy and Clinical Immunology*, 147(1):213–223, 2021.

[36] Sean Davis and Paul S Meltzer. Geoquery: a bridge between the gene expression omnibus (geo) and bioconductor. *Bioinformatics*, 23(14):1846–1847, 2007.

[37] Samuel Kaski, Janne Nikkilä, Merja Oja, Jarkko Venna, Petri Törönen, and Eero Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC bioinformatics*, 4(1):1–13, 2003.

[38] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J Gaffney, Laura L Elo, Xuegong Zhang, et al. A survey of best practices for rna-seq data analysis. *Genome biology*, 17(1):1–19, 2016.

[39] Charity W Law, Monther Alhamdoosh, Shian Su, Xueyi Dong, Luyi Tian, Gordon K Smyth, and Matthew E Ritchie. Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger. *F1000Research*, 5, 2016.

[40] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[41] Kasper D Hansen, Rafael A Irizarry, and Zhijin Wu. Removing technical variability in rna-seq data using conditional quantile normalization. *Biostatistics*, 13(2):204–216, 2012.

[42] Limin Zhang, Lijun Sun, Bin Zhang, and Lihong Chen. Identification of differentially expressed genes (degs) relevant to prognosis of ovarian cancer by use of integrated bioinformatics analysis and validation by immunohistochemistry assay. *Medical science monitor: international medical journal of experimental and clinical research*, 25:9902, 2019.

[43] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[44] Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

[45] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, 47(D1):D607–D613, 2019.

[46] Szklarczyk D. et al. String v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research (Database issue)*, 48, 2019.

[47] X Robin, N Turck, A Hainard, N Tiberti, F Lisacek, JC Sanchez, M Müller, et al. proc: Display and analyze roc curves. r package version 1.10. 0, 2017.

[48] Suzana de Siqueira Santos, Daniel Yasumasa Takahashi, Asuka Nakata, and André Fujita. A comparative study of statistical methods used to identify dependencies between gene expression signals. *Briefings in bioinformatics*, 15(6):906–918, 2014.

[49] William Revelle. psych: Procedures for psychological, psychometric, and personality research. *R package version*, 1(10), 2018.

[50] Lin Yuan, Wen Sha, Zhan-Li Sun, and Chun-Hou Zheng. Biweight midcorrelation-based gene differential coexpression analysis and its application to type ii diabetes. In *International Conference on Intelligent Computing*, pages 81–87. Springer, 2013.

[51] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, (1):559, 2008. URL `https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-559`.

[52] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[53] Xiao Xiao, Ethan P White, Mevin B Hooten, and Susan L Durham. On the use of log-transformation vs. nonlinear regression for analyzing biological power laws. *Ecology*, 92(10):1887–1894, 2011.

[54] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `https://igraph.org`.

[55] Seth Pettie and Vijaya Ramachandran. An optimal minimum spanning tree algorithm. *Journal of the ACM (JACM)*, 49(1):16–34, 2002.

[56] Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007.

[57] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11):2498–2504, 2003.

[58] Gang Su, Allan Kuchinsky, John H Morris, David J States, and Fan Meng. Glay: community structure analysis of biological networks. *Bioinformatics*, 26(24):3135–3137, 2010.

[59] Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[60] Gabriela Bindea, Bernhard Mlecnik, Hubert Hackl, Pornpimol Charoentong, Marie Tosolini, Amos Kirilovsky, Wolf-Herman Fridman, Franck Pagès, Zlatko Trajanoski, and Jérôme Galon. Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–1093, 2009.

[61] Kevin Blighe, Sharmila Rana, and Myles Lewis. *EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling*, 2020. URL `https://github.com/kevinblighe/EnhancedVolcano`. R package version 1.6.0.

[62] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[63] Richard Leon Eckert and Ellen Anne Rorke. Molecular biology of keratinocyte differentiation. *Environmental health perspectives*, 80:109–116, 1989.

[64] Thomas Werfel. The role of leukocytes, keratinocytes, and allergen-specific ige in the development of atopic dermatitis. *Journal of Investigative Dermatology*, 129(8): 1878–1891, 2009.

[65] Michiko K Oyoshi, Rui He, Lalit Kumar, Juhan Yoon, and Raif S Geha. Cellular and molecular mechanisms in atopic dermatitis. *Advances in immunology*, 102:135–226, 2009.

[66] Ying Sha, John H Phan, and May D Wang. Effect of low-expression gene filtering on detection of differentially expressed genes in rna-seq data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 6461–6464. IEEE, 2015.

[67] Shunian Xiang, Zhi Huang, Tianfu Wang, Zhi Han, Y Yu Christina, Dong Ni, Kun Huang, and Jie Zhang. Condition-specific gene co-expression network mining iden-

tifies key pathways and regulators in the brain tissue of alzheimer's disease patients. *BMC medical genomics*, 11(6):39–51, 2018.

[68] Michel L Goldstein, Steven A Morris, and Gary G Yen. Problems with fitting to the power-law distribution. *The European Physical Journal B-Condensed Matter and Complex Systems*, 41(2):255–258, 2004.

[69] Wei Chen, Kuixia Xie, Xinhua Liu, and Hong Chen. Identification of key pathways and genes in psoriasis via gene microarray analysis. *Molecular medicine reports*, 13(3):2327–2337, 2016.

[70] Fanny Büchau, Christina Munz, Cristina Has, Robert Lehmann, and Thomas Michael Magin. Klhl16 degrades epidermal keratins. *The Journal of investigative dermatology*, 138(8):1871–1873, 2018.

[71] Johann E Gudjonsson, Jun Ding, Andrew Johnston, Trilokraj Tejasvi, Andrew M Guzman, Rajan P Nair, John J Voorhees, Goncalo R Abecasis, and James T Elder. Assessment of the psoriatic transcriptome in a large sample: additional regulated genes and comparisons with in vitro models. *Journal of Investigative Dermatology*, 130(7):1829–1840, 2010.

[72] David Adrian Ewald. Molecular characterization of atopic dermatitis.

[73] Andrew J Dunbar and Chris Goddard. Structure-function and biological role of betacellulin. *The international journal of biochemistry & cell biology*, 32(8):805–815, 2000.

[74] Julia Lachner, Veronika Mlitz, Erwin Tschachler, and Leopold Eckhart. Epidermal cornification is preceded by the expression of a keratinocyte-specific set of pyroptosis-related genes. *Scientific reports*, 7(1):1–11, 2017.

[75] HD De Koning, EH Van Den Bogaard, JGM Bergboer, M Kamsteeg, IMJJ van Vlijmen-Willems, K Hitomi, J Henry, M Simon, N Takashita, Akemi Ishida-Yamamoto, et al. Expression profile of cornified envelope structural proteins and keratinocyte differentiation-regulating proteins during skin barrier repair. *British Journal of Dermatology*, 166(6):1245–1254, 2012.

[76] Wynand Winterbach, Piet Van Mieghem, Marcel Reinders, Huijuan Wang, and Dick de Ridder. Topology of molecular interaction networks. *BMC systems biology*, 7(1): 1–15, 2013.

[77] Andrea Franceschini, Jianyi Lin, Christian von Mering, and Lars Juhl Jensen. Svdphy: improved prediction of protein functional associations through singular value decomposition of phylogenetic profiles. *Bioinformatics*, 32(7):1085–1087, 2016.

[78] Qian Zhao, Yan Zhang, Shichun Shao, Yeqing Sun, and Zhengkui Lin. Identification of hub genes and biological pathways in hepatocellular carcinoma by integrated bioinformatics analysis. *PeerJ*, 9:e10594, 2021.

[79] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: a database of predicted functional associations between proteins. *Nucleic acids research*, 31(1):258–261, 2003.

A Supplementary material

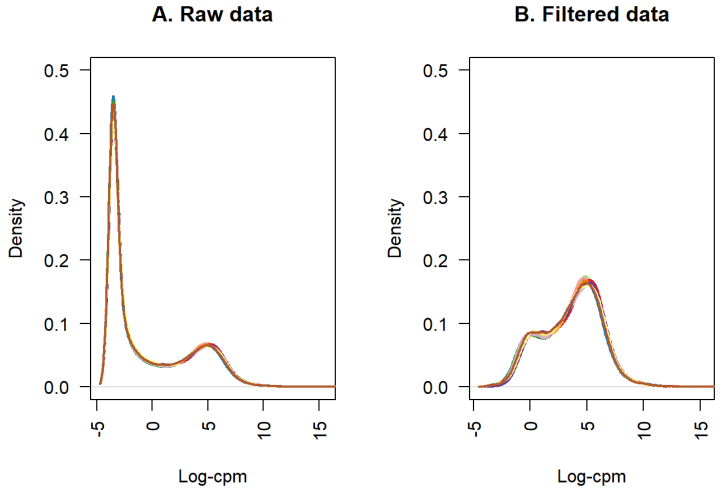


Figure 1: These density plots show how data distribution changes before and after removing low-expressed genes. Plot on the left show raw data without pre-processing and it highlights the presence of a peak at low expression values. This peak is not present in the plot on the right, the one obtained after removing low-expressed genes.

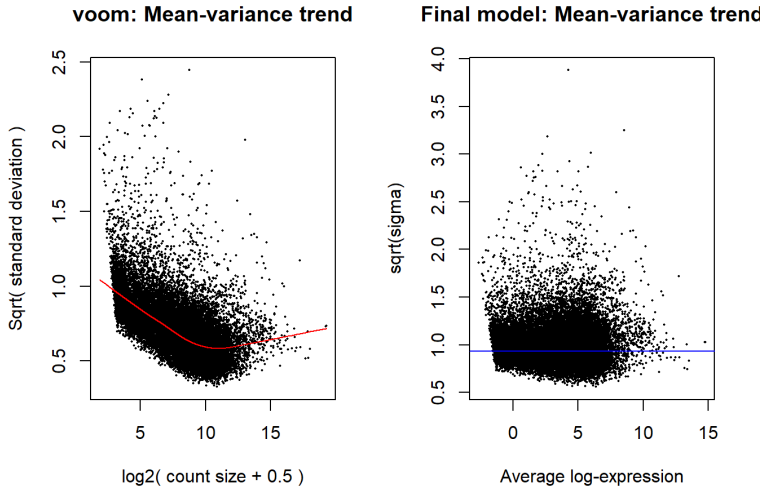


Figure 2: These graphs show how mean-variance trend changes after Voom transformation on the data. Plot on the lrft describes data before applying Voom and it highlights an irregular distribution. Plot on the right shows how the distribution was more regular after Voom transformation.

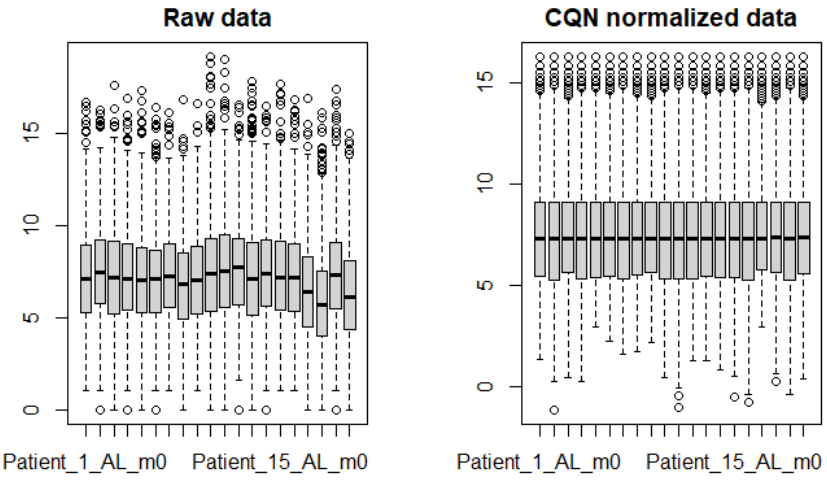


Figure 3: This figure shows how data distribution changes before and after data normalization with CQN. Variability between samples was removed to ensure that the expression distribution of each sample were similar across the entire experiment.

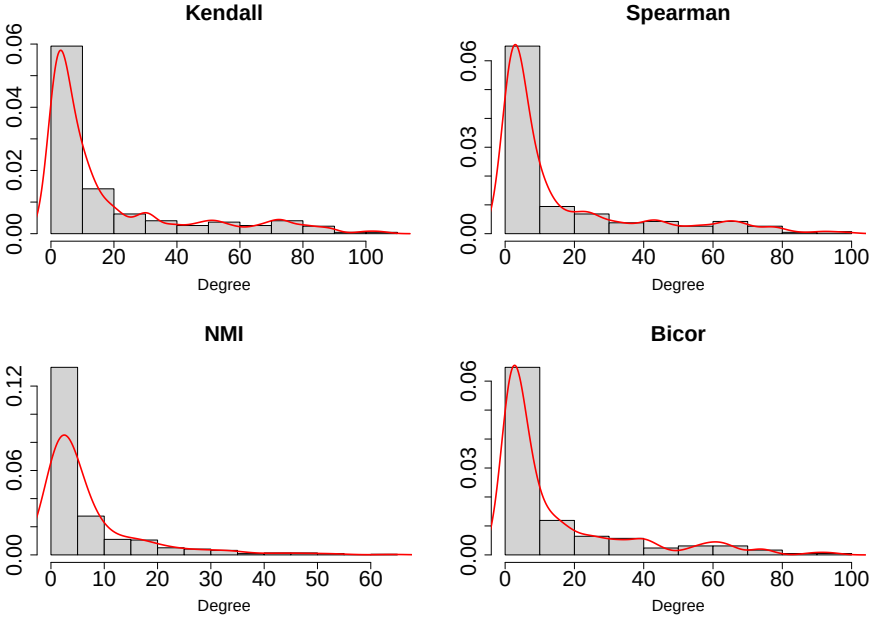
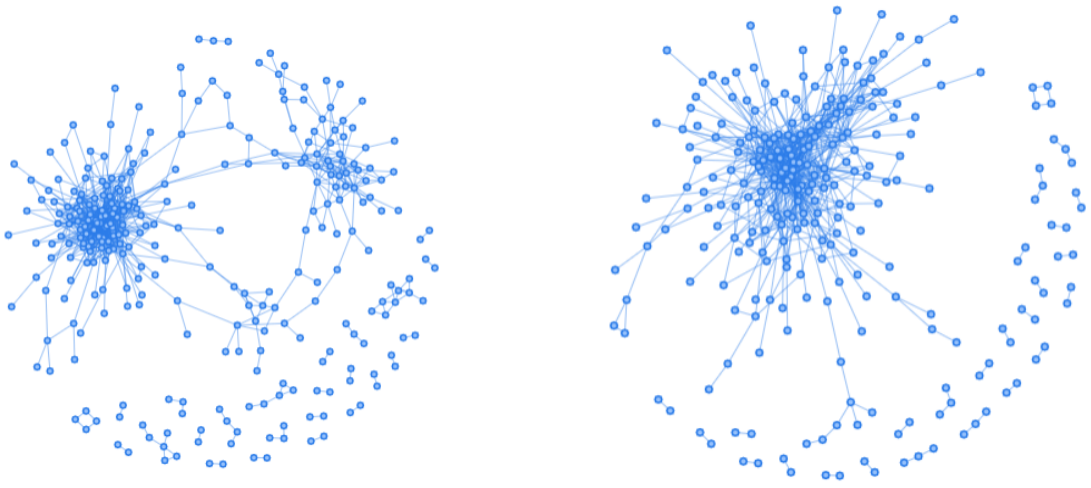
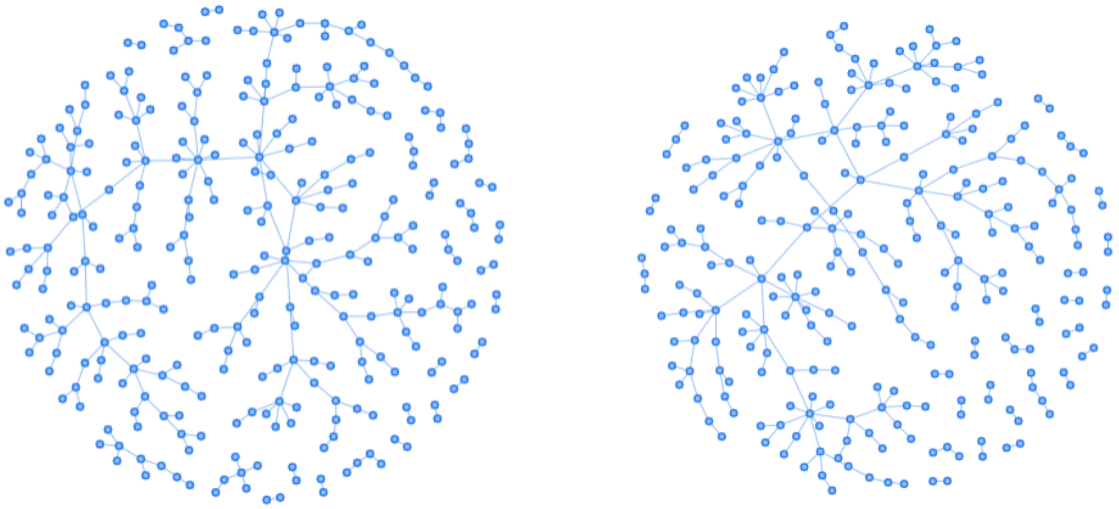


Figure 4: This figure show the distribution of the node degree in the networks obtained using log-log thresholds. They highlight the ability of these thresholds to generate networks with a power-law distribution of the degree.



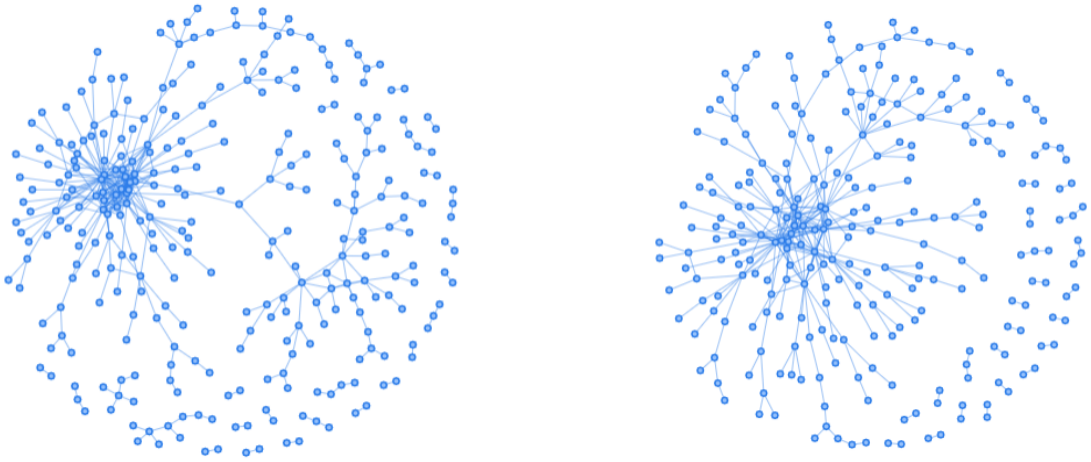
(a) Lesional **(b)** Non-Lesional

Figure 5: *Ensemble networks* obtained after major voting and ensemble strategy are showed in this figure.



(a) Lesional **(b)** Non-Lesional

Figure 6: *Pruned networks* obtained after heuristic pruning are showed in this figure. Compared to *ensemble networks* is is evident that most of the edges have been removed. Edges are selected to maximize the total weight with the minimum number of interactions that preserves the overall connection.



(a) Lesional

(b) Non-Lesional

Figure 7: Networks obtained after the last step are showed in this figure. They are the ones that returned the best AUC values in the range screened for the addition of hubs interactions. Scale-freeness and sparseness are the main advantages of these networks.

Gene name	Gene description	Degree
CHP2	calcineurin like EF-hand protein 2	10
KLHL16	gigaxonin	9
ABHD12B	abhydrolase domain containing 12B	6
H63ST6	heparan sulfate-glucosamine 3-sulfotransferase 6	6
CLDN1	claudin 1	6
TRPV6	transient receptor potential cation channel	6
SFTPC	surfactant protein C	6
LCK	LCK proto-oncogene, Src family tyrosine kinase	6
LAMB2P1	laminin subunit beta 2 pseudogene 1	6
KRT6B	keratin 6B	6

Table 1: Hubs genes with highest degree in AL *pruned network*.

Gene name	Gene description	Degree
BTC	betacellulin	9
ACTRT3	actin related protein T3	9
IL37	interleukin 37	8
TNNC1	troponin C1, slow skeletal and cardiac type	7
FALEC	focally amplified long non-coding RNA	7
S100A9	S100 calcium binding protein A9	6
ID4	inhibitor of DNA binding 4, HLH protein	6
LCE5A	late cornified envelope 5A	6
GPRASP1	G protein-coupled receptor associated protein 1	6
HOXC-AS1	HOXC cluster antisense RNA 1	6

Table 2: Hubs genes with highest degree in AN *pruned network*.