Universiteit Leiden
Opleiding Informatica

Developing a computational framework to identify muscle-specific gene expression modules in RNA-seq data

Name: Min Zhu

Date: 15/08/2019

1st supervisor: Thomas Bäck
2nd supervisor: Vered Raz

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Contents

Abstract

Gene co-expression network is a statistical framework that can identify clusters of functionally related genes, named modules. Weighted Gene Correlation Network Analysis (WGNCA) is widely applied in biological conditions with high contrast. In this work, it was explored whether specific gene modules can be identified between closely related biological tissues with mild differences. In human, dierent skeletal muscles may have different physiological and biomechanical function. Molecular differences between skeletal muscles are predominantly unknown. In the group, a pilot study generated RNA-sequencing data from 3 subjects and 6 muscles. In that study dierences between skeletal muscles were smaller than the inter-individual dierences. The aim of this project was to investigate a better method to identify muscle-specific modules. To reach a statistical significance, simulated data was generated from the pilot with dierent degrees of inter-individual dierences. Using the simulated data, muscle-specific modules are compared that were identified with the classical WGCNA to a consensus network. This work shows that the consensus network performs better. The advantages and disadvantages for each method will be discussed as well.

1 Introduction

Skeletal muscles are of one origin and hence are considered as groups of tissues with similar molecular properties. Numerous muscles work together to achieve specific bio-mechanical functions[1]. This suggests that there is specificity among muscles. Indeed, different patterns of muscles involvement are recognised between muscular dystrophy(MD) and aging, where some muscles are initially affected, and others are only later involved or spared[2]. Those differences between muscles may implicate the existence of muscle-specific gene groups.
The molecular mechanism underlying this differential muscle involvement is yet unclear. With the development of high-throughput sequencing technologies (Next Generation Sequencing, NGS), it is possible to obtain dense data of gene expression levels, from which gene groups could be identified. Considerable research efforts have been devoted to the analysis of genome-wide gene expression datasets using NGS technologies. There are quite some studies on skeletal muscles. Nonetheless, only a few studies have focused on understanding the differences between muscle types. With this study, I would like to explore methods to detect gene co-expression networks and to assess this methodology to detect muscle-specific gene modules.

1.1 RNA Sequencing

RNA sequencing (RNA-Seq) is an approach taking advantage of deep sequencing technologies. As shown in Figure 1, the first step in RNA-seq is the generation of cDNA fragments from mRNA molecules. After the sequencing adaptors (blue strings) are added to all cDNA fragments, and short sequence reads are generated by the NGS instrument. The resulting reads of nucleotide sequences are aligned to the genome. The resulting sequence reads are then classified into exonic reads (reads are wholly contained in exon(the part of genes are encoded in the final mature RNA) defined by the library), junction reads (reads between two exons joined together in the final RNA) and poly(A) end-reads (the tail parts of the RNA sequences). With these three sorts of reads, one can generate a base-resolution expression profile for each gene.

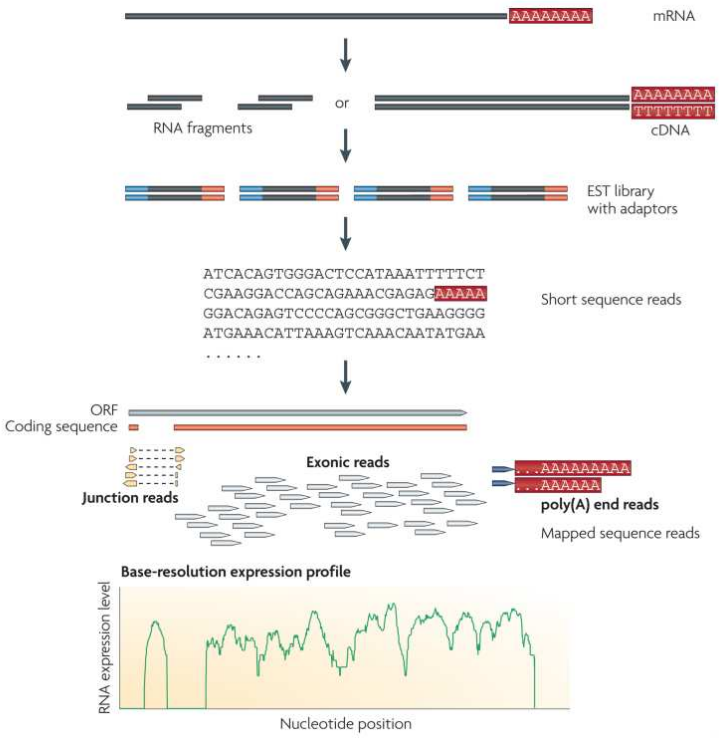


Figure 1: RNA-seq workflow[3]

Compared with other methods quantifying the transcriptomics, RNA-seq shows remarkable performance. It has a lower background noise than tiling microarray and requires less amount of material than cDNA sequencing[3]. Besides that, it is capable of distinguishing different transcript isoforms and allelic expression and identify all possible transcripts.

1.2 Network

A network is a set of nodes and edges connecting the nodes[4]. For one network, the nodes and edges may have various properties; for instance, the edges can be either directed or undirected; they can be assigned with weights to illustrate the emphasis of the connection. One needs to take all the features of the network into account when analysing the network. The social network, the Internet or traffic network are good examples of the form of network.
Since Euler figured out seven Bridges of Königsberg problem in 1735, for decades, the study of networks or graph theory has been pursued in many fields. As computer power improved, far more data can be obtained and analysed than previous. People now emphasise on the large-scale statistical properties of graphs instead of the features of a single node. For a network with thousands of nodes,

the primary aim becomes to reveal and understand the statistical properties as well as the structure of the entire system.
The idea of the network has been widely applied in various domains. With these studies, people concluded some common properties observed in different kinds of networks. These properties are believed to be practical to understand the network.
The degree of a node is the number of edges between the node and other nodes. Thus, the portion of nodes in the network of which degree is k is defined as p_k; it is the probability that a randomly chosen node has degree k. The degree distribution of a network can be constructed by the degrees of all nodes. Alternatively, we can use the probability $P_k$that the degree of one node is greater than or equal to k to represent the degree data.

$$P_k = \sum_{k'=k}^{\infty} p_{k'} \tag{1}$$

Ideally, for the simplest network model, the degree distribution of a random graph is found binomial, or Poisson in the limit of large graph size. However, in real life, most of the distributions of networks are highly right-skewed. Many of them follow a power-law distribution in their tails.

$$P_k \sim \sum_{k'=k}^{\infty} k'^{-\alpha} \sim k^{-(\alpha-1)} \tag{2}$$

With the increase of the degree k, the form of P_k decays slowly. In this kind of networks, most nodes only connect to a few nodes; at the same time, a little number of nodes are highly connected. These key nodes are called hub nodes. This topology shows remarkable tolerance against error, which attributed to the robustness. Networks with power-law degree distributions, as known as scale-free networks, are of great interest, are frequently found in biological systemsas well.

1.3 Gene co-expression network

The notion of gene co-expression network is a widely used and meaningful application of complex network. It is increasingly applied to study the function of genes from a system level[5]. The nodes represent genes, and the edges mean that the corresponding genes are (significantly) co-expressed according to the given samples. It must also be mentioned that many works show that many co-expression networks only approximately share the scale-free property. We may need to modify the distribution so that it can fit scale-free topology better. In the network, there are groups of nodes with high topological overlap, which may be functionally related. These groups are defined as modules. Thus, one can explore the gene data by detecting the modules from gene co-expression network and studying the relationship between these modules and the trait

of interest. Instead of using the raw count data from the RNA-seq experiments, usually one would transform the count data properly, for instance, using variance-stabilising transformation or a simple log transformation.

1.4 Weighted gene co-expression network analysis

In general, (Pearson) correlation coefficient is used to calculate the similarity of the pairs of observations. It is usually also applied in gene expression cluster analysis. Thus, the co-expression network is usually undirected, and the edges do not carry weights so that the pairs of nodes are either connected or not. A common strategy is to pick a number as a 'hard' threshold so that a gene co-expression network is constructed based on whether the correlation is larger than the threshold. However, we cannot ignore the disadvantage of a hard threshold. Using a number as a threshold may lead to loss of information. For instance, when taking 0.8 as the threshold, the nodes pair with correlation value 0.79 will be regarded as unconnected, which may miss an essential edge in the graph[5]. This poses another issue that the result is very sensitive to the choice of the threshold: when reducing the threshold to 0.75, the nodes pairs with correlation value 0.79 will be 'connected', this will introduce a number of edges compared with the graph with threshold 0.8. In addition, the gene co-expression is a binary property because of the hard threshold. Whether it has biological meaning remains a problem.
Thus, 'soft' thresholding was proposed: it assigns a weight to each edge with a number in [0,1] instead of using 1 and 0 to represent connected and unconnected. Weighted gene co-expression network analysis (WGCNA) was then introduced in 2005 by Steve Horvath[5], and it is now widely applied in various biological contexts.
Sometimes there are a few genes that far away from all the other genes. Unlike some network systems in real life that all the nodes should belong to a cluster to make sure each node can be reached, these "outlying" nodes tend to be not functionally related to the closest clusters in biological networks. In this case, it is not wise to merge these nodes and the clusters. Thus a grey module is defined for those outlying genes that do not belong to any modules. A grey module is a group for unassigned genes which do not share functional relation.

1.5 Research question

As we know, gene co-expression network is a very helpful tool to identify clusters of functionally related genes. It is essential to study and understand the function of genes. The remarkable performance of Weighted Gene Correlation Network Analysis (WGNCA) appeals to us, and WGCNA became the preferred method. However, from the result of pilot RNA-seq data, we noticed that the differences between related skeletal muscles are smaller than the inter-individual differences between humans using WGCNA while all the samples are taken into account regardless of individuals. This shows the differences between individuals have a leading effect instead of muscle types.

Therefore, we are interested in answering the following question: how can we avoid the effect of non-relevant factors in clustering?
If the network is generated by the samples from the same individual, it is supposed to be homogeneous in the individual aspect. A consensus network strategy is then proposed to apply to avoid the variance among individuals. The idea is to create co-expression networks per individual followed by the generation of a consensus network across individuals.
By simulating RNA-seq data with different degrees of inter-individual differences, we could create the gene co-expression network per individuals or among all the samples. Therefore we would be able to compare this method with the original WGCNA method and check if this method enables us to detect more modules that can tell the difference between muscles.
This thesis is structured as follows: in Section 2, we first introduce several clustering algorithms for gene expression data set. The used materials and applied methods in this project are described and illustrated in Section 3. The experiment results in the Section 4 shows the potential of our proposed method. At the end We discuss the performance of proposed method and the further steps in the future work.

2 Related work

Different from other network analysis, the result of biological networks should be biologically meaningful as well. The clusters of the networks are not only based on statistics, but also have some biological meaning. In order to analysis the gene co-expression network generated by the gene expression data set, a number of module detecting methods have been proposed.

2.1 K-means

The k-means algorithm is one of the oldest clustering algorithms, but still very popular. It was proposed in 1979 by Hartigan *et al.* [6]. This method is not specifically aimed at gene co-expression networks; it is widely applied in various fields.
The main idea of the k-means clustering method is to divide M points into K clusters so that the value of the within-cluster sum of squares (WCSS) reaches its minimal[6]. To achieve this goal, first K cluster centres are generated. By assigning the data points to their closest clusters, the centres of cluster keep moving. When the difference between original centres and new centres is smaller than a certain threshold, the position of the centres tends to be stable. One can then take these clusters as the final clustering result.
To use the k-means algorithm, one must set the number of clusters K in advance, which is time-consuming to figure out what is the best parameter. In addition, the performance heavily depends on the initial centres (some initial centres may lead to local optimal instead of global optimal).

2.2 Co-clustering method

In 2002, Daniel *et al.* [7] proposed a distance function using additional information from the available biological networks. In this case, metabolic networks with a set of chemical reactions are used. With the help of Kyoto Encyclopedia of Genes and Genomes (KEGG) database, they first obtain a subnetwork based on the given gene expression data, in which the nodes are molecules, and the edges are weighted. They took correlation measure as distance function and combined biological network nodes and genes. The clustering methodology they applied is hierarchical average linkage clustering.
The results showed that this method managed to detect the relationship on the gene expression data and biological network (like metabolic networks). People can also choose various clustering methods to improve performance. However, the disadvantage of co-clustering is very obvious. This method is limited to the available biological network data.

2.3 Independent Component Analysis

There are two main steps in decomposition methods: decompose the expression matrix (RNA reads file) into multiple matrices and then extract module from

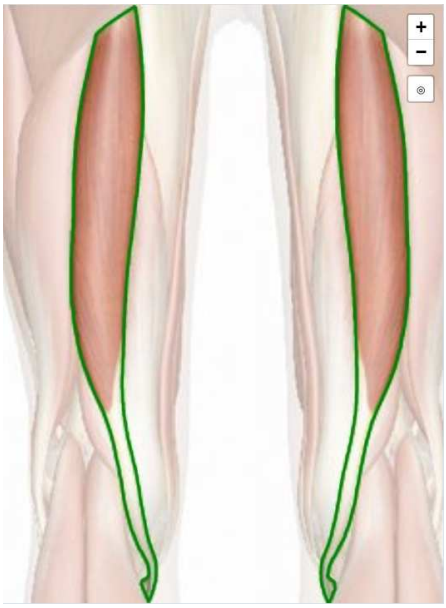
every component. Independent Component Analysis (ICA) was developed to find a linear representation of data to capture the features of the data[8]. The observation is defined as a linear mixture of n independent components (signals) for the observation is generated by a series of latent variables, which can not be observed directly. By using an iterative process, n independent signals are detected with randomly initialised weights. After that, the post-processing step is carried on to obtain modules. One can choose from a false-discovery rate (FDR) and z-scores. Whether a gene will be assigned to a module depends on the cutoff which denotes the compactness of the module.
Overall, this method performs well not only when there is overlap between modules but also when there is no overlap. However, it requires the number of modules before the analysis and it is sensitive to the number of samples in the data set [9]. The decrease of samples affect the performance of ICA a lot. In addition, Moreover, this method has several parameters, which need to be tuned on every single data set. This will affect the biological interpretation. The lack of external information would affect the performance as well.

3 Materials and method

3.1 Pilot data

The original RNA-seq gene expression data set was generated from seventeen human skeletal muscle samples. These samples came from six different leg muscles from three young, healthy, male individuals. Their ages range from 18 to 30. In total, there are 58,051 genes in the data set. The muscles are gracilis (G), semitendinosus (ST), vastus medialis (VM), gastrocnemius lateralis (GCL), rectus femoris (RF) and vastus lateralis (VL). For muscle semitendinosus, distal (relatively away from heart) and middle samples were obtained and sequenced separately. These two samples are annotated as ST_D and ST_M.

(a) Gracilis (G)

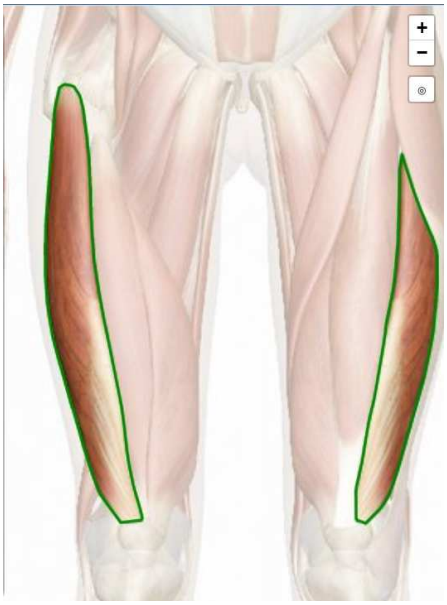


(b) Semitendinosus (ST)

(c) Vastus medialis (VM)

(d) Gastrocnemius lateralis (GCL)

(e) Rectus femoris (RF)



(f) Vastus lateralis (VL)

Figure 2: The positions of muscles involved in this project[10]

There is only one sample of vastus lateralis (VL). If we apply the sample VL

as input, we would not know the source of difference is the muscle effect or the individual effect. To avoid bias, we removed this sample from the whole data set and did not take it into account. In addition, genes that are low expressed among all the samples are filtered, which did not affect the analysis.

3.2 Data simulation

In order to explore the inter-individual differences and to correct for the individual effect, we would like to have more RNA-seq data from more people. Due to the limited existing data, we propose to simulate a data set of twenty individuals based on the original data set.
A linear model is a commonly-used tool for regression. By fitting the model, one could generate new data with the model. Linear mixed-effect model(LME) is an extension of simple linear models[11]. Compared with the linear model, LME allows fixed and random effect factors. Since our interest is the effect of various muscles instead of the difference between individuals, we apply the linear mixed-effect model to fit the pilot data and further simulate new data.
In general, the linear mixed-effect model can be represented as the following formula.

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \varepsilon \tag{3}$$

where
$\mathbf{y}$ is a $N \times 1$ vector of observations;
$\mathbf{X}$ is a $N \times p$ matrix of p fixed effect factors;
β is a $p \times 1$ column vector of the fixed effect regression coefficients;
$\mathbf{Z}$ is a $N \times q$ matrix for the q random effect factors ;
$\mathbf{u}$ is a $q \times 1$ column vector of the random effect regression coefficients;
and ε is a $N \times 1$ vector of the residuals.
When fitting the linear mixed-effect model to the data points and estimating the coefficients, instead of estimating u, we usually assume that u follows normal distribution[11], with mean $\mu = 0$ and standard deviation $\sigma = \mathbf{G}$ the variance-covariance matrix of the random effects.

$$u \sim \mathcal{N}(0, \mathbf{G}) \tag{4}$$

Although it is not always the case that the random effects are normally distributed, stratified analysis and the confidence intervals plot can be used to improve the assumption if a good deal of data is available. In our study, there are only 16 data points, around 5 per group; we need to assume a normal distribution to have sufficient power.

3.3 Network generation

Before starting to generate the gene-coexpression network, the gene expression data frame is first transformed. Then the output is used to calculate the absolute value of the correlation between each pair of genes, construct the similarity matrix. The $n \times n$ similarity matrix $\mathbf{S} = [s_{ij}]$ will then be transformed into

an $n \times n$ adjacency matrix $\mathbf{A} = [a_{ij}]$, of which the component a_{ij} shows the strength of connection between gene i and j. Instead of implementing 'hard' threshold, two 'soft' adjacency functions were applied in WGCNA to avoid losing information. According to Horvath [5], the results of these two functions are very similar when using the scale-free topology criterion. The default function is the power function in WGCNA.
After that, the adjacency matrix is used to calculate the distance (dissimilarity) between nodes. Here, the topological overlap dissimilarity measure is implemented to reduce the noise[12]. It was proved to come to modules with biological meaning[5]. The topological overlap between two nodes can measure the similarity between each other in the topological level.
With this, one could carry on clustering and find the modules. The default method applied in WGCNA is hierarchical clustering[13]. It is a "bottom-up" approach. The results of hierarchical clustering are usually illustrated in the form of the dendrogram. The discrete branches of the clustering dendrogram correspond to modules. With dynamic branch cut function, similar modules can be merged.

3.3.1 Correlation measure

Correlation measures are used to illustrate how similar each pair of genes is among all the samples. In general, the Pearson correlation is a commonly used similarity measure.

$$cor_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{5}$$

where X and Y are the variables; $\bar{X}$ and $\bar{Y}$ are the mean of X and Y respectively. Nevertheless, it is very sensitive to outliers. Besides the Pearson correlation, the biweight midcorrelation (bicor) is proposed by Wilcox[14]. Compared with the Pearson correlation, biweight midcorrelation is proved to be more robust[15]. To calculate the biweight midcorrelation of vector $\mathbf{x}, \mathbf{y}$, the quantities u_a, v_a is first defined

$$u_a = \frac{x_a - med(\mathbf{x})}{9 \cdot mad(\mathbf{x})} \tag{6}$$

$$v_a = \frac{y_a - med(\mathbf{y})}{9 \cdot mad(\mathbf{y})} \tag{7}$$

where $med(\mathbf{x})$ and $mad(\mathbf{x})$ are the median of $\mathbf{x}$ and the median absolute deviation of $\mathbf{x}$ respectively. The weights $w_a^{(X)}$ can be obtained by

$$w_a^{(X)} = (1 - u_a^2)^2 I(1 - |u_a|) \tag{8}$$

If $1 - |u_a| > 0$, $I(1 - |u_a|)$ equals to 1, otherwise it equals to 0. The value of weights indicated the difference between x_a and $med(\mathbf{x})$ and between x_a and

$9 \cdot mad(\mathbf{x})$. The biweight midcorrelation of $\mathbf{x}$ and $\mathbf{y}$ can be calculated by the following formula:

$$bicor(\mathbf{x},\mathbf{y}) = \frac{\sum_{a=1}^{m}(x_a - med(\mathbf{x}))w_a^{(a)}(y_a - med(y))w_a^{(y)}}{\sqrt{\sum_{b=1}^{m}\left[(x_b - med(\mathbf{x}))w_b^{(x)}\right]^2}\sqrt{\sum_{c=1}^{m}\left[(y_c - med(y))w_c^{(y)}\right]^2}} \tag{9}$$

After calculating the correlation between each pair of vector, the $n \times n$ similarity matrix $S = [s_{ij}]$ will be transformed into an $n \times n$ adjacency matrix $A = [a_{ij}]$ by using adjacency function. Instead of picking a number as a 'hard' threshold, two 'soft' adjacency functions were applied in WGCNA: the sigmoid function[13]

$$a_{ij} = sigmoid(s_{ij}, \alpha, \tau_0) \equiv \frac{1}{1 + e^{-\alpha(s_{ij} - \tau_0)}} \tag{10}$$

and the power adjacency function

$$a_{ij} = power(s_{ij}, \beta) = |s_{ij}|^{\beta} \tag{11}$$

where the parameters α, τ_0 and β can be adjusted.
According to Horvath [5], the results of these two functions are very similar when using the scale-free topology criterion. We choose the power function in this study so that we have fewer parameters to tune. The idea of selecting a suitable power is to make the correlation more similar to the scale-free topology. So we would calculate for a series of powers and see with which power the network resembles a scale-free graph better. Here scale-free topology fit index is applied to evaluate if power is reasonable. If the value is larger than 0.8, then the power can be used.

3.3.2 Distance measure

It is believed that genes with high topological overlaps tend to share the same neighbourhood, which means they tend to be in the same modules in the network[12]. To obtain the topological overlap value of a pair of genes, one needs to compare all the genes directly connected to these two and check how much is shared. The topological overlap of two genes can be defined as

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}} \text{ if } i \neq j \tag{12}$$

where $l_{ij} = \sum_u a_{iu}a_{uj}$ and $k_i = \sum_{u \neq i} a_{iu}$ is the node connectivity, equals the number of nodes directly connected to node i. The dissimilarity matrix can be obtained by $d_{ij} = 1 - \omega_{ij}$.

3.3.3 Consensus network

As discussed previously, we plan to generate a gene co-expression network per individual and then construct a consensus network among all the networks. There

are several methods proposed to obtain consensus network. There are different methods to create the consensus network[16][17][18][19]. These methods take different measures as the input, as shown in Figure 3. Here, we introduce these methods in the order of appearance in the modified workflow.

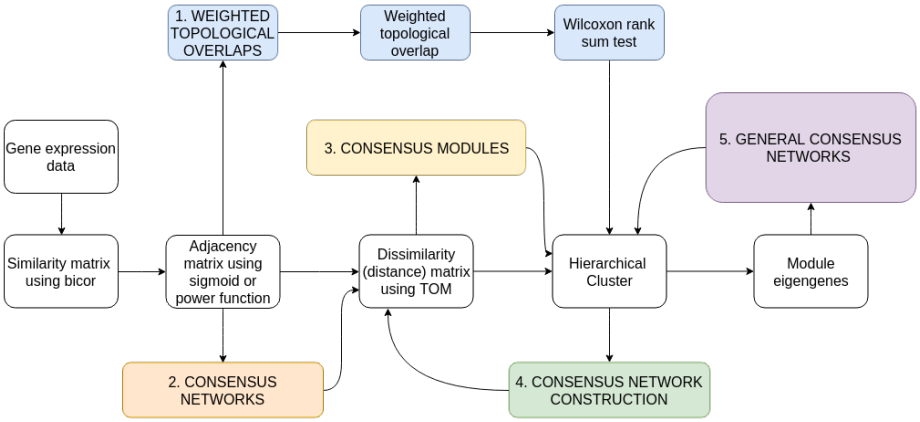


Figure 3: Various methods modifying the WGCNA pipeline

Method One Stefano *et al.* [16] implemented this pipeline to construct the consensus network. Their raw data from microarrays and RNA-Seq is analysed by various methods, respectively. Different from the topological overlap measure used in WGCNA, they calculated weighted topological overlap, which shows both significantly positive and negative correlations[16]. The weighted connectivity of a node i is

$$K_i = \sum_j a_{ij} \tag{13}$$

and the weighted topological overlap (wTO) is calculated as

$$\omega_{ij} = \frac{c_{ij} + a_{ij}}{min(K_i, K_j) + 1 - |a_{ij}|} \tag{14}$$

where $C = A * A^T$. To begin with, the Wilcoxon rank-sum test is applied to check if there is a statistical difference between datasets and to ensure one is able to construct the consensus network with all the datasets. In addition, another Wilcoxon rank-sum test with alternative hypothesis H_1 that the mean $|wTO|$ for a given GRF-GRF (gene regulatory factor) pair is larger than 0.3 is performed to avoid the potential false positive. Then the median of wTO values among all datasets for each pair of genes is obtained as the consensus wTO.

Method Two The method was proposed by Peter Langfelder and Steve Horvath[17] in 2007. The idea is very intuitive: only if the nodes(genes) have

connections among all the input networks, they should be connected in a consensus network. With this notion, we can have

$$Consensus_{q,ij}((\mathbf{A^{(1)}}),(\mathbf{A^{(2)}}),...,(\mathbf{A^{(n)}})) = Quantile_{q,ij}((\mathbf{A^{(1)}}),(\mathbf{A^{(2)}}),...,(\mathbf{A^{(n)}})) \tag{15}$$

where i and j are two genes, A is the adjacency matrix for every data set.
It is more robust than using the minimal of the input adjacency matrices. For this method compares networks directly, it would not work properly when the data sets vary in sample sizes, array platforms or gene expression normalisation methods[17], which will lead to the bias on the result of quantile transformation.

Method Three Similar to the previous method(Eq.15), Peter Langfelder and Steve Horvath defined consensus modules as modules in the consensus network[17]. A consensus gene dissimilarity matrix is obtained by

$$Dissim(Consensus(TOM(\mathbf{A^{(1)}}),TOM(\mathbf{A^{(2)}}),...,TOM(\mathbf{A^{(n)}}))) \tag{16}$$

where TOM is calculated in Eq 12. The result can be used as the input of subsequent hierarchical clustering.
This method shares the same disadvantage in common with Method three for its direct comparison of network.

Method Four The intention of this method is to construct one consensus network from subsampled datasets to obtain higher reliability than the standard WGCNA method[19].
This method was first proposed by Monti *et al.* [18] in 2003. The consensus matrix is introduced to indicate how frequent each pair of genes is clustered together.

$$A_{i,j} = \frac{\text{number of times gene } i \text{ is clustered with gene } j}{\text{number of times gene } i \text{ is subsampled with gene } j} \tag{17}$$

This provides a similarity measure which can be used to obtain a distance matrix as the input of clustering.

Method Five A robust approach was proposed to avoid the limitation of data sets in various aspect[17]. Instead of using TOM (topological overlap matrix), 'compressed' adjacency matrix was applied.
First of all, the pipeline of module detection is applied to every single data set. The corresponding module eigengenes (the first principal component of the expression matrix, TOM matrix in this case of the module) are obtained. Thus, in data set s, the module that gene i belongs to is the module with the highest module eigengene according to gene i denoted as $Module_{(i)}^{(s)}$.

$$Module_{(i)}^{(s)} = argmax_J(|cor(x_i, E_J)|) \tag{18}$$

where $cor(x_i, E_J)$ is the correlation between the expression of gene i and the eigengene E_J.
The compressed adjacency is then defined as

$$a^{(s)}_{compressed,ij} = \frac{1}{2}\left[1 + cor(E^{(s)}_{Module^{(s)}_{(i)}}, E^{(s)}_{Module^{(s)}_{(j)}})\right] \tag{19}$$

Thus one can calculate the gene dissimilarity for clustering.

$$d_{ij} = Dissim(Consensus(a^{(1)}_{compressed}, a^{(2)}_{compressed}, ...)) \tag{20}$$

Among these five methods, the first method introduced the weighted topological overlap; then it took the median as the consensus value. Except for the fourth approach, the idea behind the other approaches is very similar. In general, it takes the median of the input as the consensus value, which is robust against the outlier to some extent. But they start the consensus measure from various steps. For the first three methods, they generate one final dissimilarity matrix for clustering while the fifth performs the module detection on data set separately and then generate the consensus network. This is particularly helpful for the analysis of datasets with different properties. However, this will introduce more parameters which may lead to over-fitting. The fourth approach should be adjusted when implemented. For it initially aimed at consensus network for multiple subsampled data sets. It is notable that the function can only tell how frequent two genes are clustered together, not how close they are in the module. In our case, we are going to use simulated data sets based on the linear mixed model, which will be consistent. There will be no difference in sample sizes, array platforms or normalisation method. Thus, the concern about different properties of data sets does not exist. To avoid over-fitting, we would try the approach that modify the WGCNA pipeline before clustering. Thus we can either calculate consensus adjacency matrix or consensus TOM matrix. Since taking the minimal value is too strict, we plan to use first quartile and the second quartile as known as median as a lower bound threshold.

3.3.4 Module detection

In general, modules are assumed to be groups of genes which are highly related among all the samples. The definition applied to detect modules is modules are groups of nodes with high topological overlap[5]. With the dissimilarity(TOM) matrix, one could carry on clustering and find the modules. The default method applied in WGCNA is hierarchical clustering[13].

Algorithm 1 Hierarchical clustering

Require: Distance Matrix $DissTOM$
  Each gene starts in its own cluster
  while there is more than one cluster **do**
    Calculate the distance between each pair of genes
    Cluster the pair with the lowest distance
    Add cluster to cumulative cluster list
  end while
  return cumulative cluster list

The branches of the clustering dendrogram correspond to modules. After the clustering, according to the expected minimum distance between modules, a further dynamic branch cut function can be applied. The similar modules will then be merged.

3.4 Differential expression analysis

To analysis the RNA-seq expression, one common way is to carry on the differential expression analysis based on the factor information of gene samples. The aim is to abstract the most significant differentially expressed genes. It is noteworthy that the variance between genes is both from biological (biological difference) and technical (mistake during the experiments etc.) nature. McCarthy *et al.* found that the expression data shows a strong mean-variance relationship, which does not work with normal-based analysis[20]. Thus, negative binomial (NB) distribution was proposed to model the expression data[21].

$$Y_{gi} \sim \mathrm{NB}(M_i p_{gj}, \phi_g) \tag{21}$$

where M_i is the total number of reads of sample i, $p_g j$ is the relative abundance of gene g in group j that sample i belongs to. $M_i p_{gj}$ is then the mean. ϕ_g is the dispersion[21] and variance is $\mu_{gi}(1 + \mu_{gi}\phi_g)$.
The relative abundance is vital to differential expression analysis. Generally, ϕ_g means the coefficient of biological variation. This enables to divide the biological variation from the technical one. By using conditional maximum likelihood, the gene-wise dispersion could be estimated[22]. Then these dispersion values are shrunk to a consensus value by empirical Bayes procedure[23]. At last, an exact test similar to Fisher's exact test, which works for overdispersed data is applied to assess the differential expression[24]. The statistical results are essential to detect the most significant differentially expressed genes.

3.5 False discovery rate

In general, the p-value is applied to determine the significance of the results. When the p-value is small enough (usually 0.05), we can reject the null hypothesis H_0. However, multiple comparisons would lead to a highly increased false positive (type I error, which is to reject a true null hypothesis). To avoid

this side effect, a false discovery rate (FDR) was proposed to use in 1995 by Benjamini and Hochberg [25]. Compared with other multiple testing correction methods, FDR tries to restrict the ratio between positive and false positive.

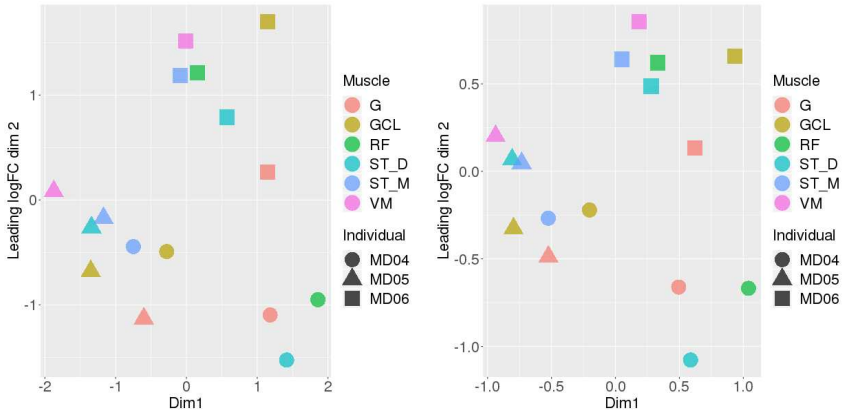
$$FDR = p\frac{m}{k} \tag{22}$$

where p is original p-value, m is the number of hypotheses and k is the rank of this p-value among all the tests. It has been proved to be a desirable control in many applications and now is widely used in many fields.

4 Results

4.1 Original pipeline

With the pilot data set, we first apply the original WGCNA pipeline. Before the analysis, the MDS (Multidimensional Scaling Plot) and PCA (Principal Component Analysis) plots are obtained respectively to have a general idea about the data set. Although a few samples that come from the same muscle are very close to each other, all these plots show that the samples tend to cluster based on the individual instead of muscle, the individual factor is the leading factor of the clusters.



(a) MDS plot with top 500 genes

(b) MDS plot with top 5000 genes

(c) MDS plot with top 10000 genes

(d) MDS plot with top 20000 genes

Figure 4: MDS plots based on pilot RNA-Seq data set with different numbers of top genes used to calculate pairwise distances.

Here we took the various amount of the genes into account when calculating the pairwise distances. Although the distances between a few pairs of muscles are close (GCL from MD04 and MD05), the samples roughly gather based on the individual factor. This agrees with what we observed in the PCA plot (Figure 5). It is notable that although two samples (MD06_GCL and MD04_RF) seems far away from other samples, the distance between them to the rest is not that big in PC1 (first principal component which has the largest possible variance).

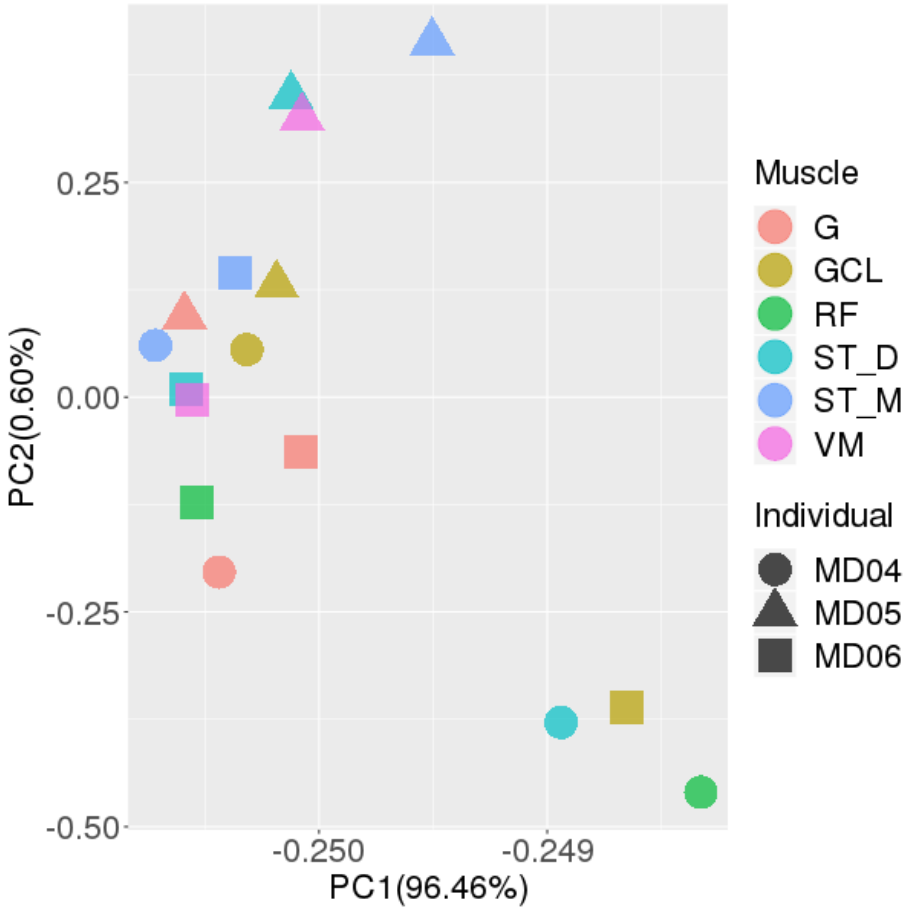


Figure 5: PCA plot for pilot data set

From the sample dendrogram and trait heat map (Figure 06), we can see that the branches are samples from the same individual, which means that the samples cluster largely by individual and not by muscle, as a consequence of the bigger inter-individual variation compared to the inter-muscle variation.

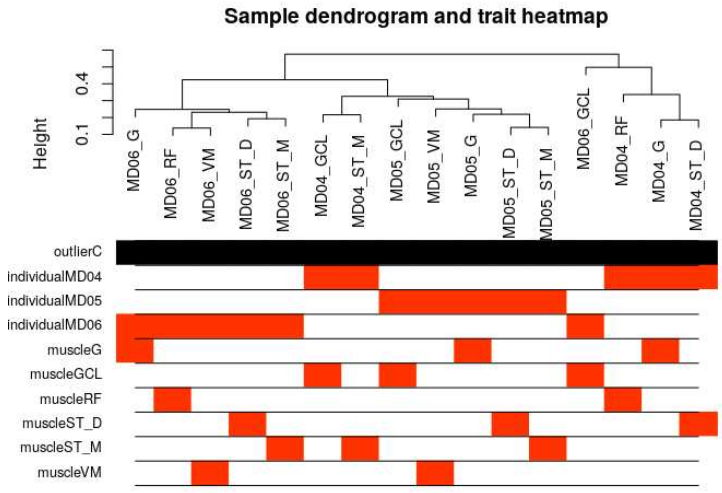


Figure 6: Dendrogram of samples and trait heat map for pilot data set

In order to make the pilot data set fit the scale-free topology better, we tried a list of powers and took 0.8 as the threshold. In this case, 8 is the lowest power for which the scale-free topology fit index curve reaches a high value (over 0.8). In this case, 8 is picked to raise the correlation matrix and continue with further steps.

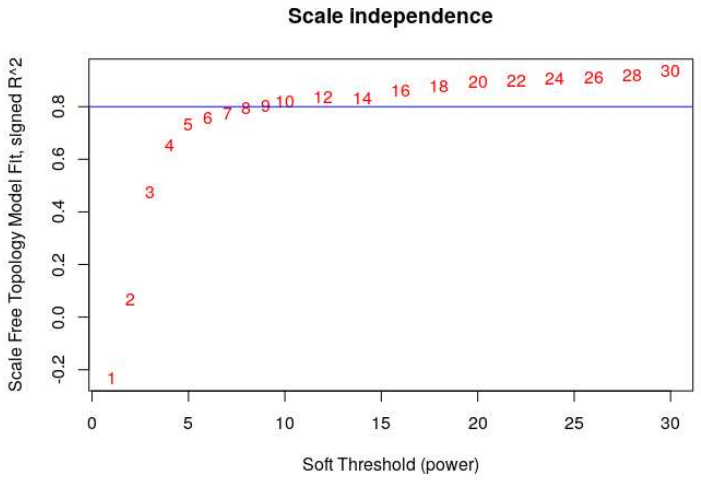
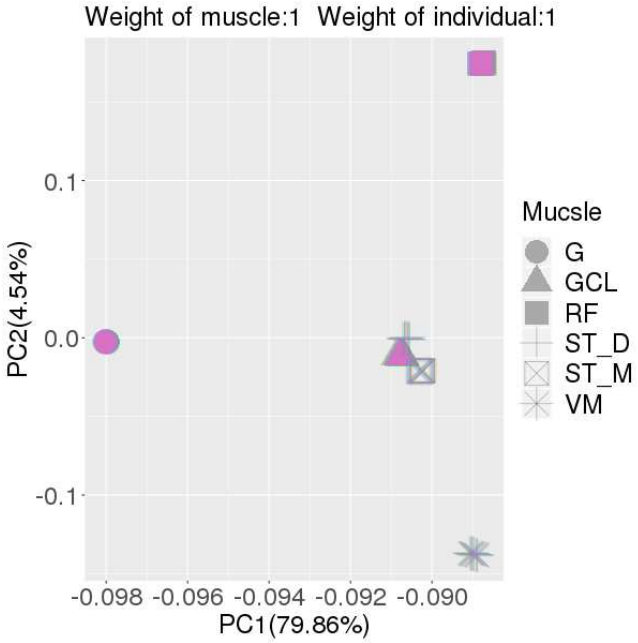


Figure 7: Analysis of network topology for various soft-thresholding powers. The y-axis shows the scale-free fit index and the x-axis refers to the power.

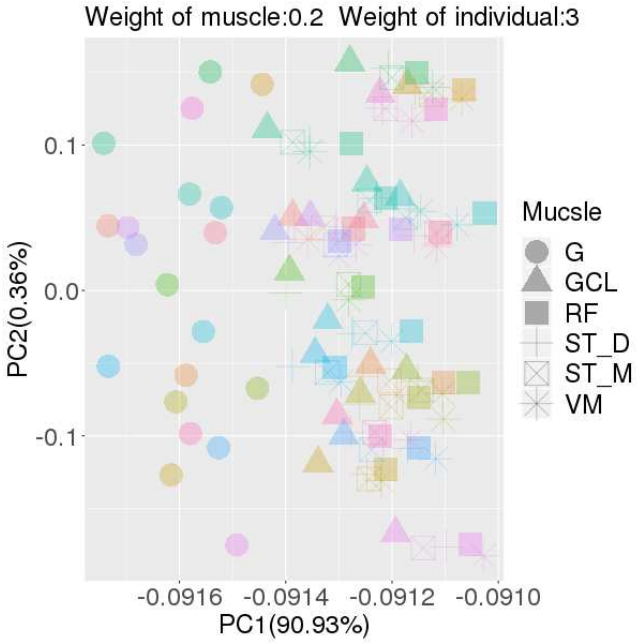
After generating the full gene co-expression network, the modules are detected using a hierarchical clustering algorithm. Among 74 modules clustered based on the pilot data, there are 26 modules found that the module eigengenes have a relationship with at least one trait, by taking False Discovery Rate (FDR) 0.05 as a cut-off. By observing the heat map of the module trait relationship, it is clear that there are many more modules related to the individual effect. This also concurs with what we found in the MDS plot that samples tend to be clustered by individuals. Only 4 modules are related to muscle effect while half of them are relevant to individual effect at the same time.

4.2 Data preparation

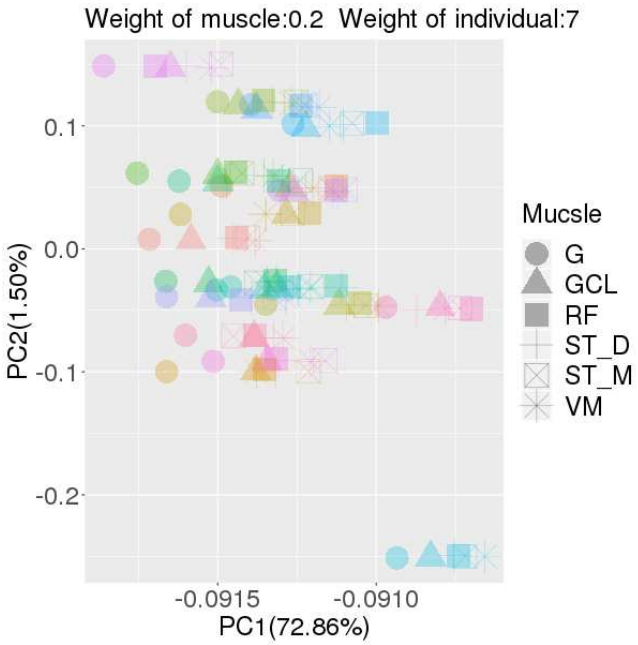
As discussed previously, we are limited by the number of samples we have. It would be powerless if we generate a consensus network based on three individuals. In order to study and compare the performance of the consensus method and the original WGCNA method, we would like to simulate more samples from various individuals. In this project, we simulated samples from 20 individuals. During the simulation, we found that the linear mixed effect model emphasised the muscle effect, which does not agree with the real situation. To generate data sets that are more similar to the real data, we added two multiplier weights to both muscle effect factor and individual effect factor. By decreasing the muscle effect and increasing the individual effect, we obtain simulated data sets in which the individual effect is the major factor.
As we can see in Figure 8, the data set without weights (Figure 8(a))and the one with weight of muscle factor: 0.2 and weight of individual factor:7 (8(c))are two extreme data sets. In Figure 8(a), all the samples from the same muscles clearly gather together. Meanwhile in Figure 8(c), although there are overlaps between samples, the leading factor of clusters is individual. In these two plots, we can see that muscle effect factor and individual effect factor take the lead in the clusters respectively, while the one with weight of muscle factor: 0.2 and weight of individual factor:3 (Figure 8(b)) is a data set in between, that is more similar to the real situation. We keep these two extreme data sets for comparison to evaluate how the methods perform in these different scenarios. For the sake of discussion, we annotated these three data sets with the weights of muscle and individual factor. For instance, the notation of the data set with weight of muscle factor: 0.2 and weight of individual factor:7 is Mus 0.2 Ind 3.



(a) Simulated data set without weights



(b) Simulated data set weight of muscle: 0.2 and weight of individual: 3

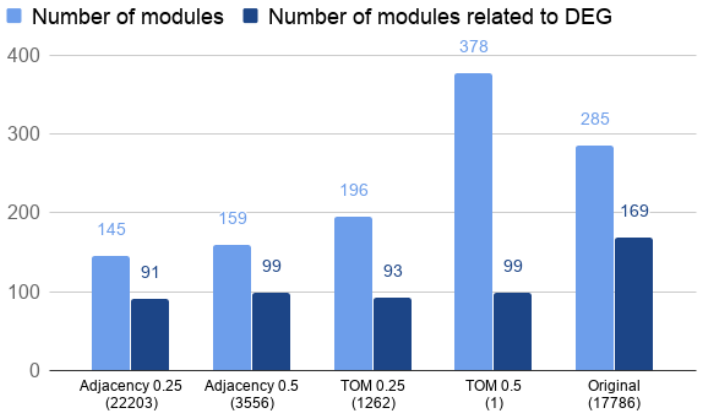


(c) Simulated data set weight of muscle: 0.2 and weight of individual: 7

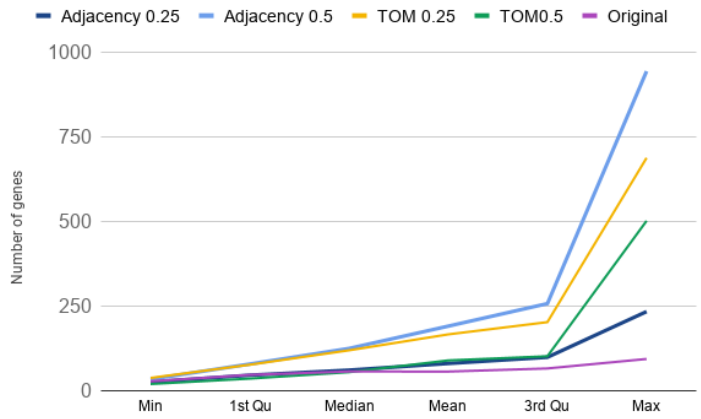
Figure 8: PCA plots for simulated data sets with 3 various weights, for simulated data sets, where various colors represent different individuals.

4.3 Parameter Selection

For the simulated data sets, we would first like to calculate the correlation and TOM matrix separately using data from every single individual, for the following consensus calculation. There are two steps to intercept and two quantiles strategies. Thus, we have a total of 4 combinations. We applied these 4 ways and original WGCNA pipeline to the simulated data set with the weight of muscle factor: 0.2, the weight of individual factor:3.



(a) The number of modules detected by various methods.The number in the brackets is the number of genes in grey module which are in fact unassigned.



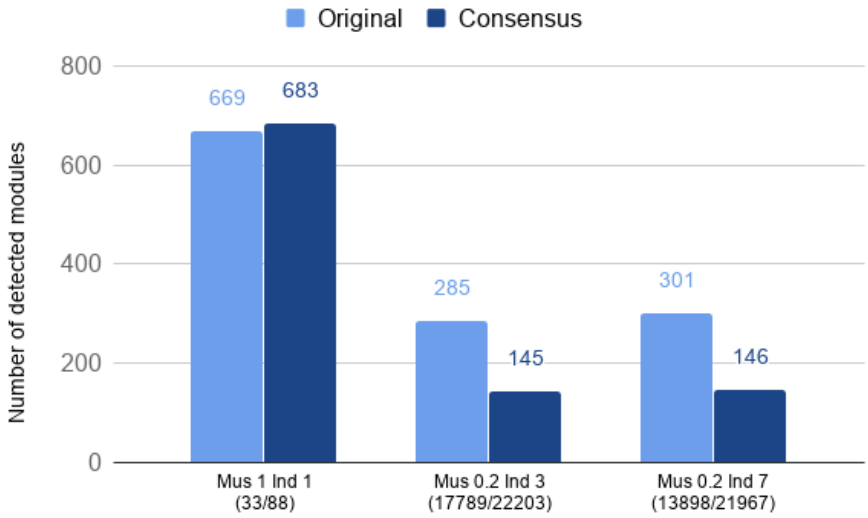
(b) Statistics summary of the size of modules detected by various methods

Figure 9: The comparison between 4 consensus methods and original method in two different point of view. "Adjacency" and "TOM" refer to consensus adjacency matrix method and consensus TOM matrix method respectively.The number after the consensus strategy (0.25 and 0.5) is the percentage of the numbers, which are first quartile and median.

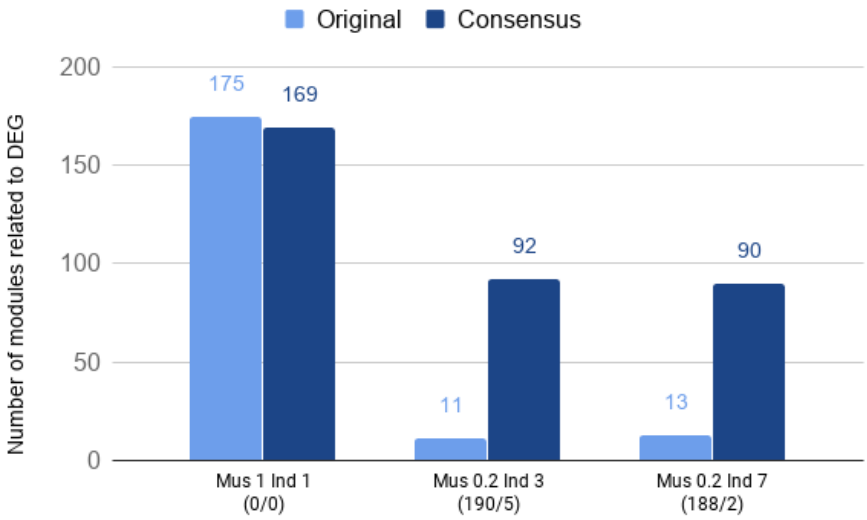
Based on the results of these consensus methods, we can observe that consensus adjacency method using first quantile is most similar to the original. Their number of genes in grey module is comparable and the features of the sizes of modules are also very much alike. Thus we picked the consensus adjacency method using first quantile for the further study.

4.4 Methods comparison

From the following two plots, we can see there are vast differences between the results using two methods when there is a stronger individual effect. For the simulated data sets with the stronger individual effects(Mus 0.2 Ind 3 and Mus0.2 Ind 7), the number of modules detected by the consensus method is lower than in the original pipeline. However, most of the top differential expressed genes are found in the modules using consensus method, while over 180 differential expressed genes are in the grey module according to the original WGCNA pipeline and are, therefore "unassigned" and not forming a module with potential biological interpretation. In the simulated datasets with the most significant muscle effect (Mus 1 Ind 1), the consensus method and the WGCNA on the full matrix perform comparably. In both cases, all differential expressed genes are found in modules, and the total numbers of modules and "unassigned" genes are comparable.



(a) The number of modules detected by original method and consensus method



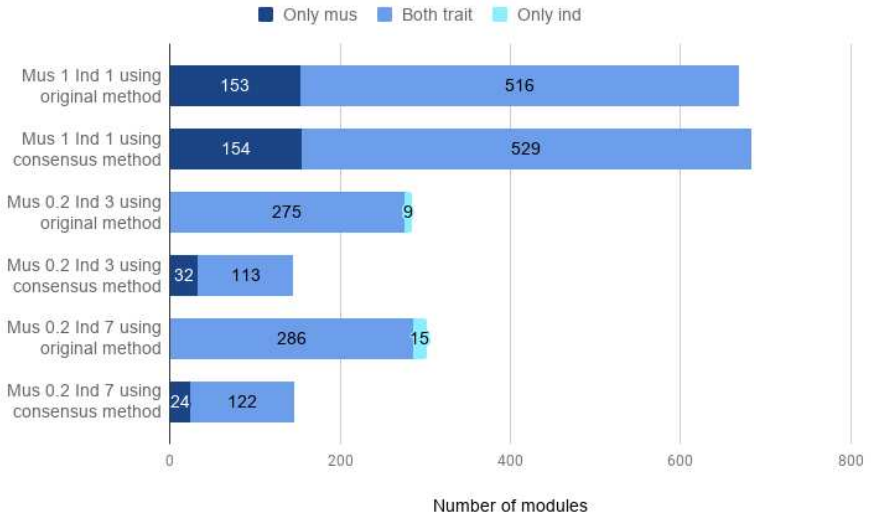
(b) The number of modules relating to top 200 DEG detected by original method and consensus method

Figure 10: The comparison between consensus method and original method. "Mus" and "Ind" leads the weight of muscle effect factor and individual effect factor respectively. The numbers in the brackets are the genes assigned to grey module.

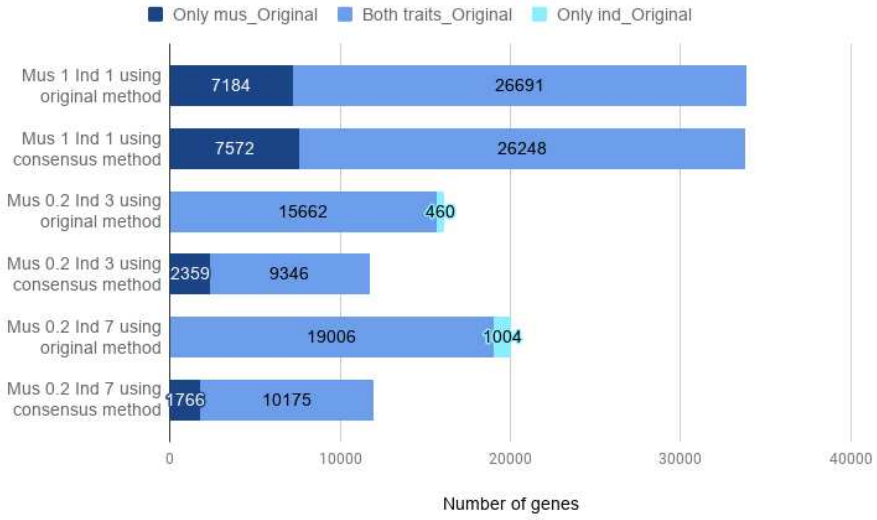
Based on these modules, we carried on module trait correlation analysis. The table below is used to define and classify the modules based on their correlation with two traits.

Table 1: Definition of the classes of modules

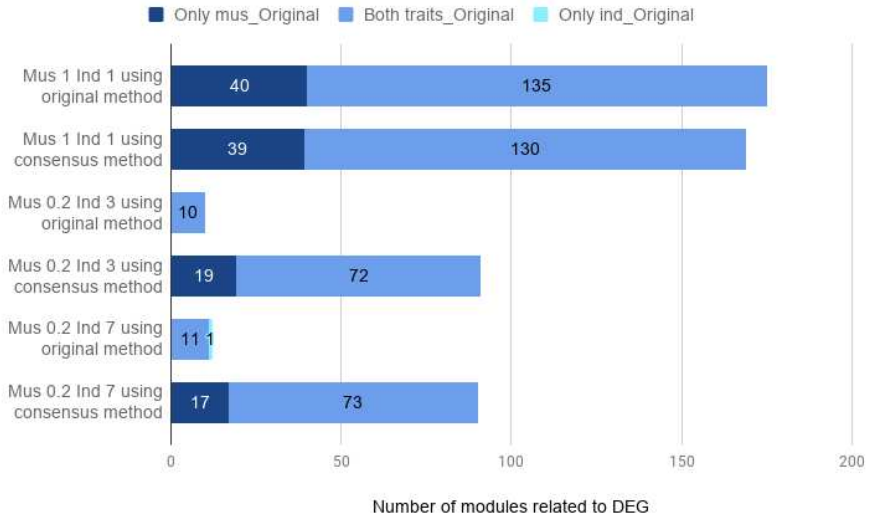
		FDR of Individual	
		<0.05	>0.05
FDR of Muscle	<0.05	Both traits	Only muscle
	>0.05	Only individual	None



(a) The number of modules correlated to various traits detected by original method and consensus method



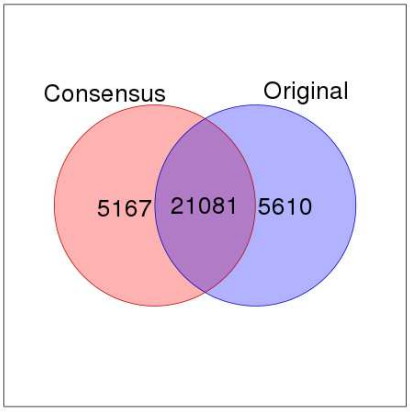
(b) The number of genes correlated to various traits detected by original method and consensus method



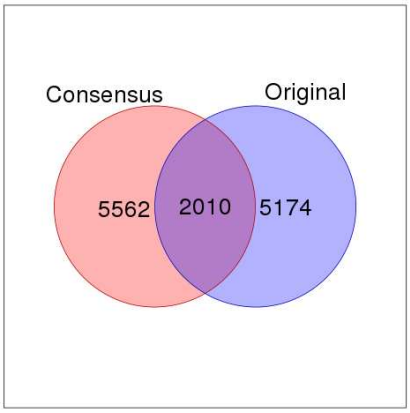
(c) The number of modules contains top 200 DEG correlated to various traits detected by original method and consensus method

Figure 11: The stacked bar charts for the modules/genes correlated to different traits using two methods

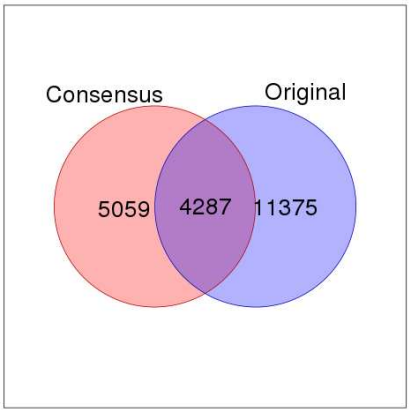
These three plots show that the consensus method detects more modules that are only correlated to muscle effect. It manages to avoid the influence of individual and detect more modules related to muscle effect, especially when there is a strong effect from the individual.
In a further step, we compared the genes in modules related to the same factors using two different methods. As we can see in Figure 12(c), almost 80% (21081/26248 and 21081/26691) of the genes are found in modules related to both muscle and individual factors using two methods for the simulated data set without weights, while around 1/3 of genes are shared in the modules only related to muscle factor (Figure 12(d)).



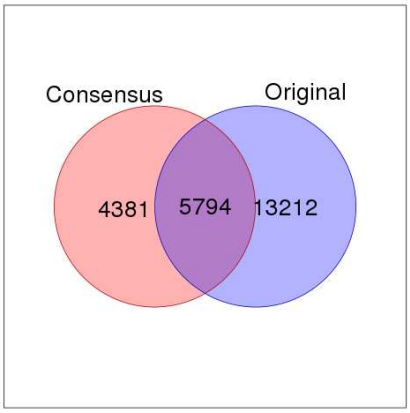
(a) Genes in modules related to both muscle and individual factors in simulated data set(muscle: 1 individual:1)



(b) Genes in modules only related to muscle factor simulated data set(muscle: 1 individual:1)



(c) Genes in modules related to both muscle and individual factors in simulated data set(muscle: 0.2 individual:3)



(d) Genes in modules related to both muscle and individual factors in simulated data set(muscle: 0.2 individual:7)

Figure 12: The Venn diagram of genes in different modules that two methods detect using various simulated data sets, where the pink circle is the genes found by consensus method and the purple one is using original WGCNA pipeline.

5 Discussion

According to the previous study in the group, we already knew that the dierences between skeletal muscles are smaller than the inter-individual dierences. Since we are interested in finding the differences between the muscles instead of individuals. It is important to avoid the effect of individual. Thus, we proposed to apply a consensus method to help detect muscle-specific modules.

By calculating the adjacency matrix within individual, there should be few variance caused by individual for all the samples used in the calculation come from the same individual. If a pair of genes are highly co-expressed, the adjacency value of this pair of gene will be high. This pattern should be extendable to biologically similar individuals, for instance, people with similar gender, ethnicity, age, etc. It is safe to take a relative low adjacency value among all the scores of the pair of the genes to represent the correlation between this pair. With the adjacency matrices, we take the first quartile among all the adjacency values of each pair of genes to generate the consensus adjacency matrix for the further steps.

From the results of experiments using various simulated data sets, we can see that there are much more modules detected in the simulated data set without weight than in the ones with higher individual effect either using original WGCNA method or the consensus method. Besides, over 90% genes are clustered in the module in the non-weight simulated data set while more than a half of the genes are marked as unassigned when the effect of individual increased. The number of top differentially expressed genes are clustered in non-grey modules agrees with this as well: there are quite a few DEG found in grey modules when there is a stronger individual effect.

Overall, the performance of the consensus method is better than the original pipeline when the influence of the individual is stronger. When the muscle effect is the primary factor, the consensus method does not outperform the clustering on the full expression matrix.

As we saw in the pilot data set that the individual effect was stronger than the muscle effect, we should be more confident in the results from the consensus method. However, due to the limitation of the size of data, we can not use the real data alone. Thus, we simulated some data sets for this project. There may be no strong biological meaning behind the data set. The method we proposed here is still a statistical validation without a possibility for biological interpretation which requires real data.

To validate the performance of the consensus method, one needs more real data from more individuals from the wet lab to carry on the experiment. Based on the result, a further gene enrichment analysis (like the Database for Annotation, Visualisation and Integrated Discovery (DAVID)) could be performed using the detected modules so that one can check if the result does have some biological meaning, and the consensus method is helpful.

We hope that this work can inspire people to develop a new algorithm that manages to conduct the gene coexpression network analysis and avoid the influence

of those unwanted traits, which would help people have a further understanding of the functions of genes.

References

[1] Erin E Terry, Xiping Zhang, Christy Hoffmann, Laura D Hughes, Scott A Lewis, Jiajia Li, Matthew J Wallace, Lance A Riley, Collin M Douglas, Miguel A Gutierrez-Monreal, et al. Transcriptional profiling reveals extraordinary diversity among skeletal muscle tissues. *Elife*, 7:e34613, 2018.

[2] Fedik Rahimov, Oliver D King, Doris G Leung, Genila M Bibat, Charles P Emerson, Louis M Kunkel, and Kathryn R Wagner. Transcriptional profiling in facioscapulohumeral muscular dystrophy to identify candidate biomarkers. *Proceedings of the National Academy of Sciences*, 109(40):16234–16239, 2012.

[3] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57, 2009.

[4] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[5] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[6] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

[7] Daniel Hanisch, Alexander Zien, Ralf Zimmer, and Thomas Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(suppl_1):S145–S154, 2002.

[8] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[9]

[10] Tim Barclay. *Muscles of the Leg and Foot*, 2019.

[11] J Bruin. Introduction to generalized linear mixed models @ONLINE, 2019.

[12] Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8(1):22, 2007.

[13] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[14] Rand R Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic press, 2011.

[15] Peter Langfelder and Steve Horvath. Fast r functions for robust correlations and hierarchical clustering. *Journal of statistical software*, 46(11), 2012.

[16] Stefano Berto, Alvaro Perdomo-Sabogal, Daniel Gerighausen, Jing Qin, and Katja Nowick. A consensus network of gene regulatory factors in the human frontal lobe. *Frontiers in genetics*, 7:31, 2016.

[17] Peter Langfelder and Steve Horvath. Eigengene networks for studying the relationships between co-expression modules. *BMC systems biology*, 1(1):54, 2007.

[18] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118, 2003.

[19] Rachel Shahan, Christopher Zawora, Haley Wight, John Sittmann, Wanpeng Wang, Stephen M Mount, and Zhongchi Liu. Consensus coexpression network analysis identifies key regulators of flower and fruit development in wild strawberry. *Plant physiology*, 178(1):202–216, 2018.

[20] Davis J. McCarthy, Yunshun Chen, and Gordon K. Smyth. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297, 01 2012.

[21] 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, 11 2009.

[22] Gordon K Smyth and Arünas P Verbyla. A conditional likelihood approach to residual maximum likelihood estimation in generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(3):565–572, 1996.

[23] Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

[24] Mark D Robinson and Gordon K Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2007.

[25] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[26] Albert-László Barabási et al. *Network science*. Cambridge university press, 2016.

[27] Wouter Saelens, Robrecht Cannoodt, and Yvan Saeys. A comprehensive evaluation of module detection methods for gene expression data. *Nature communications*, 9(1):1090, 2018.

[28] Andy M Yip and Steve Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC bioinformatics*, 8(1):22, 2007.

[29] Saskia Freytag. Simulating and cleaning gene expression data using ruvcorr in the context of inferring gene co-expression. 2015.

[30] J.A. Rice. *Mathematical Statistics and Data Analysis*. Advanced series. Cengage Learning, 2007.

[31] Charles E McCulloch and John M Neuhaus. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical science*, pages 388–402, 2011.

[32] Tom AB Snijders and Roel J Bosker. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. SAGE, 2011.

[33] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[34] Zhenqiang Su, Paweł P Łabaj, Sheng Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, Charles Wang, Gary P Schroth, Robert A Setterquist, John F Thompson, et al. A comprehensive assessment of rna-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature biotechnology*, 32(9):903, 2014.

[35] Bruce Alberts and John Wilson. *Molecular biology of the cell*. Garland Science, 2008.

[36] Wei Vivian Li and Jingyi Jessica Li. Modeling and analysis of rna-seq data: a review from a statistical perspective. *Quantitative Biology*, 6(3):195–209, 2018.