Islamic University of Technology

Department of Computer Science and Engineering

RNA-Seq Analysis: Consensus analysis with the WGCNA package in R

Melengo Evehe Habib Davy 154447
Toussie Oussoumanou 154448

Under the supervision of: Tareque Mohmud Chowdhury

November 2019

CERTIFICATE OF SUCCESSFUL COMPLETION OF THE THESIS/PROJECT

All thanks to the Almighty Allah for his mercy and guidance on us. Within the academic year 2018-2019 we have successfully completed our thesis work on RNA-Seq Analysis: Consensus analysis with the Weighted Correlation Nectwork Analysis (WGCNA) package in R. This thesis/project was undertaken by students of the Islamic University of Technology a subsidiary organ of the Organization of Islamic Cooperation undet the supervision of Tareque Mohmud Chowdhury assistant professor in the department of Computer Science and Engineering. This report is submitted as a partial fulfillment for the award of a bachelor of science degree in Computer Science and Engineering.

By:

Melengo Evehe Habib Davy N^o 154447

Toussie Oussoumanou N^o 154448

Supervised by:

Tareque Mohmud Chowdhury

Assistant Professor, Department of Computer Science and Engineering

Prof. Dr. Muhammad Mahbub Alam

Head of the Department of Computer Science and Engineering

Date ..../......../.........

Abstract

By making use of the WGCNA package in R as principal tool for experimentation, we perform a consensus analysis on two data sets containing the female mice and male mice liver expression data. Performing a consensus analysis on two data sets containing the female mice and male mice liver expression data, with the help of the weighted correlation network analysis (WGCNA) package in R, gives us the results that can be seen by observing the plots and other figures produced through the implementation of the experiment, which are joined and commented into the present document.

Acknowledgment

First and foremost we offer our sincerest gratitude and thanks to the Almighty Allah who gives us the abilities to successfully perform this task. And no doubt without the help of the Almighty Allah this report could not been completed or written.

It would not have been possible to write this bachelor thesis without the help and support of the kind people around us, to only some of whom it is possible to give particular mention here.

Above all, we would like to acknowledge the advice and guidance of our supervisor Tareque Mohmud Chowdhury, assistant professor department of computer science and engineering (CSE), Islamic University of Technology (IUT), Organization of the Islamic Cooperation (OIC). This thesis report would not have been possible without his help, support and patience.

We would like also to acknowledge the support and efforts of Prof. Dr. Muhammad Mahbub Alam, head of the department of computer Science and engineering (CSE), Islamic University of Technology (IUT), Organization of the Islamic Cooperation (OIC).

Last, but by no means the least, we thank our beloved parents who worked a lot from our birth till today for our lives and success.

For any errors or inadequacies that may remain in this work, the responsibility is entirely ours.

Contents

List of Figures

Chapter 1

Introduction

By making use of the WGCNA package in R as principal tool for experimentation, we perform a consensus analysis on two data sets containing the female mice and male mice liver expression data. Here, we concentrate on the parts of the analysis that illustrate the idea behind consensus analysis even though it parallels the data expression analysis of the female mice. The processes behind the execution of our experiment consist in a data input, cleaning and preprocessing, directly followed by a network construction and consensus module detection (made of three steps that are the automatic one-step network construction and consensus module detection, the step-by-step network construction and module detection, including scaling of topological overlap matrices and to finish, dealing with large datasets or block-wise network construction and consensus module detection, including comparing the block-wise approach to the standard single-block method), that ends up on relating consensus modules to modules in individual sets and relating the modules to external traits, to study the relationships among traits and modules using eigengene networks. The results can be seen by observing the plots and other figures produced through the implementation of the experiment, which are joined and commented into the present document. But before going though the expension and explication of the implementation processes, we would like to disclame the ownership of the original research paper and tutoriels (see the references section). In order to enter in contact with the Weighted Correlation network analysis with the WGCNA package in R, we replicated the this part of the original work and have prefered to keep the explanations of the implementation processes almost as same as in the tutorials provided by the owners of the original work.

Chapter 2

Related Works

At the beginning of our journey through the study of the weighted correlation network analysis with the WGCNA package in R, we went through the network analysis of liver expression data from female mice, to ifnd modules related to body weight. Going through that analysis of a single empirical gene expression data set, forced us into the following processes:

- the input and cleaning of the data - the network construction and module detection - relating modules to esternal clinical traits an identifying important genes - the interfacing network analysis with other data such as functional annotation and gene ontology - the network visualization using WGCNA functions - and the exportation of networks to external software

This primary experiment prepared us to the consensus analysis that it parallels closely and where some sections are repeated.

Chapter 3

Data input and cleaning, including re-formatting the data for consensus analysis

3.1 Loading expression data

The expression data is contained in two files: "LiverFemale3600.csv" and "LiverMale3600.csv". After starting an R session, we check that the current directory is appropriately set, and load the requisite pack-ages and the data. In addition to expression data, the data files contain extra information about the surveyed probes we do not need. The data sets contain roughly 130 samples each. Each row corresponds to a gene and column to sample or auxiliary information. We remove the auxiliary data and put the expression data into a multi-set format suitable for consensus analysis.

3.2 Rudimentary data cleaning and outlier removal

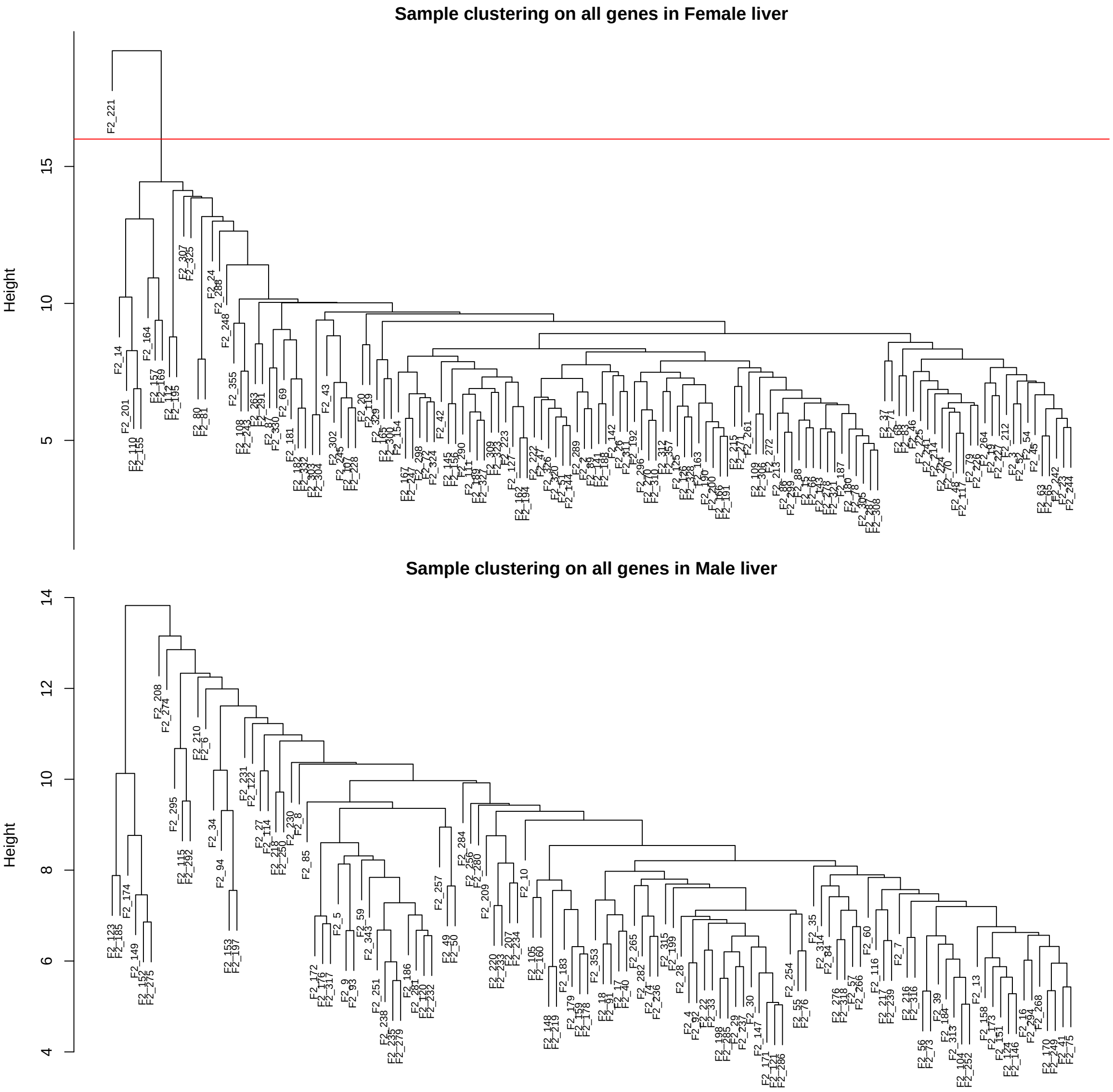
We first identify genes and samples with excessive numbers of missing samples. These can be identified using the function goodSamplesGenesMS. If the last statement returns TRUE , all genes and samples have passed the cuts. If it returns FALSE , the following code removes the offending samples and genes. Then we cluster the samples on their Euclidean distance, separately in each set. (The easiest way to see the two dendrograms at the same time is to plot both into a pdf file that can be viewed using standard pdf viewers.)

By inspection, there seems to be one outlier in the female data set, and no obvious outliers in the male set. We remove the female outlier using a semi-automatic code that only requires a choice of a height cut. We first re-plot the two sample trees with the cut lines included (as it is shown in the figure accompanying this section) and then we perform the actual outlier removal.

3.3 Loading clinical trait data

We now read in the trait data and match the samples for which they were measured to the expression samples.

We have the expression data for both sets in the variable multiExpr, and the corresponding clinical traits in the variable Traits. The last step is to save the relevant data for use in the subsequent analysis.



Sample clustering on all genes in Female liver

Sample clustering on all genes in Male liver

Figure 3.1: SampleClustering

Chapter 4

Network construction and consensus module detection

This step is the bedrock of all network analyses using the WGCNA methodology. We present three different ways of constructing a network and identifying modules:
a. Using a convenient 1-step function for network construction and detection of consensus modules, suitable for users wishing to arrive at the result with minimum effort;
b. Step-by-step network construction and module detection for users who would like to experiment with customized or alternate methods;
c. An automatic block-wise network construction and module detection method for users who wish to analyze data sets too large to be analyzed all in one.
In this section, we illustrate the 1-step, automatic multiple set network construction and detection of consensus modules. We note that while the actual network construction and module detection is executed in a single function call, a preliminary step of choosing a suitable soft-thresholding power must be performed first.

4.1 Choosing the soft-thresholding power: analysis of network topology

Constructing a weighted gene network entails the choice of the soft thresholding power β to which co-expression similarity is raised to calculate adjacency. The authors of "A general framework for weighted gene co-expression network analysis" (Statistical Applications in Genetics and Molecular Biology, 4(1):Article 17, 2005) have proposed to choose the soft thresholding power based on the criterion of approximate scale-free topology. We illustrate the use of the function "pickSoftThreshold" that performs the analysis of network topology and choose a proper soft-thresholding power. We chooses a set of candidate powers (the function provides suitable default values, but we choose the power 6 for both sets), and the function returns a set of network indices that should be inspected. The result is shown in the next figure.

4.2 Network construction and module detection

Here we make use of the function *blockwiseConsensusModules*. We have chosen the soft thresholding power 6, minimum module size 30, the module detection sensitivity *deepSplit* 2, cut height for merging of modules 0.20 implying that modules whose eigengenes are correlated above $1 - 0.2 = 0.8$ will be merged, we requested that the function return numeric module labels rather than color labels, we have effectively turned off reassigning genes based on their module eigengene-based connectivity K_{ME} , and we have instructed the code to save the calculated consensus topological overlap.

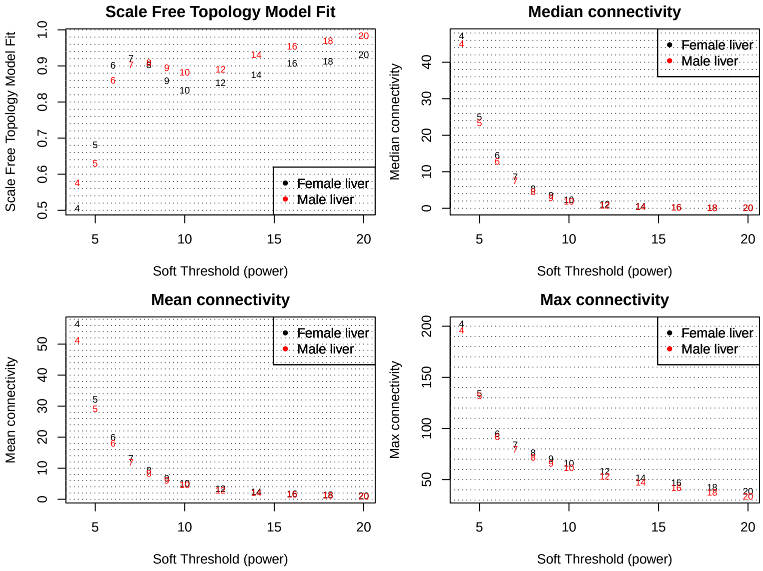


Figure 4.1: scaleFreeAnalysis

The above figure show the summary network indices (y-axes) as functions of the soft thresholding power (x-axes). Numbers in the plots indicate the corresponding soft thresholding powers. The plots indicate that approximate scale-free topology is attained around the soft-thresholding power of 6 for both sets. Because the summary connectivity measures decline steeply with increasing soft-thresholding power, it is advantageous to choose the lowest power that satisfies the approximate scale-free topology criterion.

The result net has several components:

```
> names(net)
[1] "colors"  "unmergedColors"
[3] "multiMEs" "goodSamples"
[5] "goodGenes" "dendrograms"
[7] "TOMFiles" "blockGenes"
[9] "blocks" "originCount"
[11] "networkCalibrationSamples""individualTOMInfo"
[13] "consensusTOMInfo" "consensusQuantile"
```

For now we only need the module labels contained in the component colors , the module eigengenes for each data set contained in multiMEs , and the gene dendrogram (clustering tree) in dendrograms. A quick way to take a look at the results is to plot the gene dendrogram and the corresponding module colors. The resulting plot is shown in figure below. The above figure represent the Gene dendrogram obtained by clustering the dissimilarity based on consensus Topological Overlap with the corresponding module colors indicated by the color row.

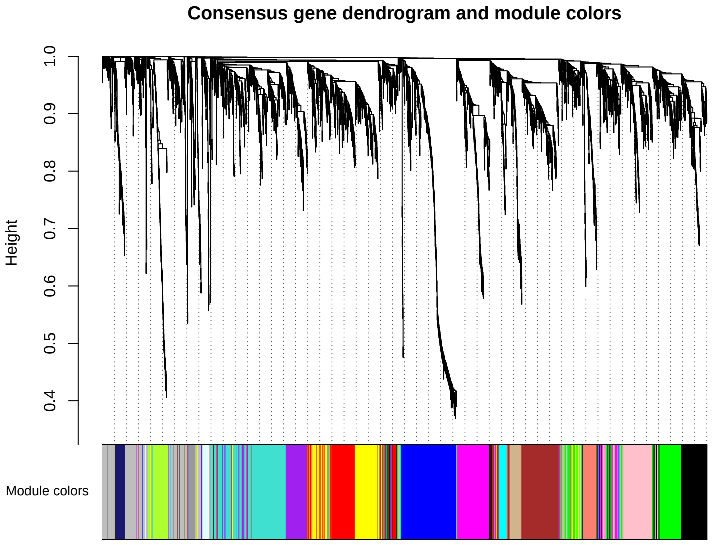


Figure 4.2: ConsensusDendrogram-auto

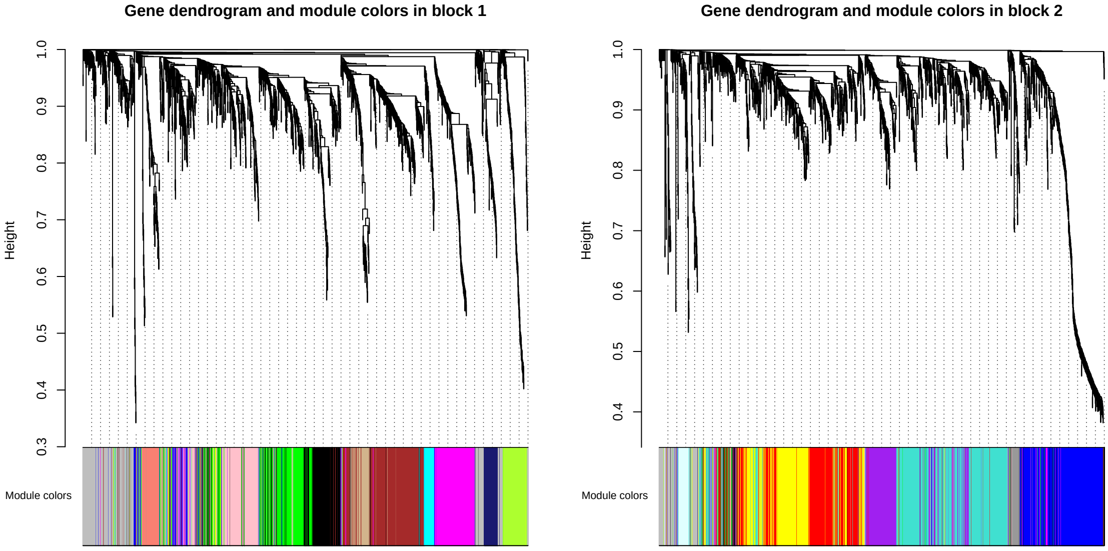


Figure 4.3: BlockwiseGeneDendrosAndColors

The following figure represent the Gene dendrograms obtained by block-wise clustering the dissimilarity based on consensus Topological Overlap with the corresponding module colors indicated by the color row. And the figure below, the last of this section represent the Gene dendrogram obtained in the single-block analysis together with the corresponding single-blockmodule colors and the module colors obtained by block-wise clustering. The plot indicates a strong correspondence between the standard and the block-wise modules.

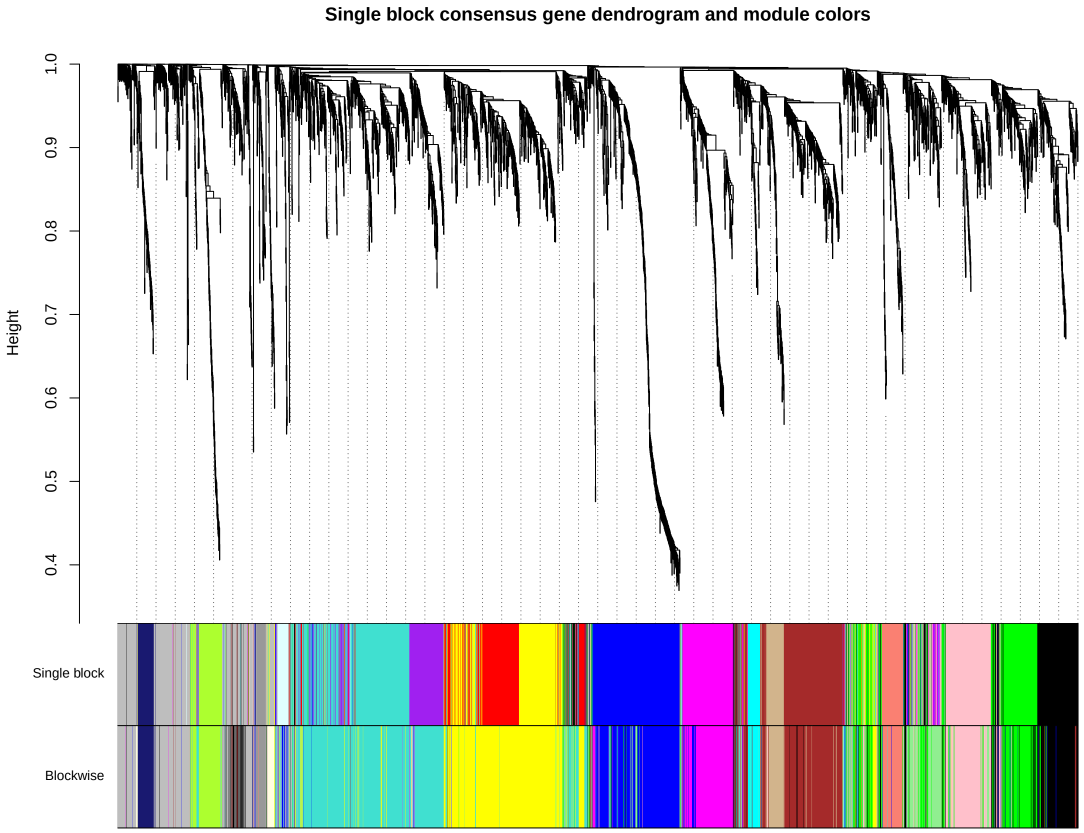


Figure 4.4: SingleDendro-BWColors

Chapter 5

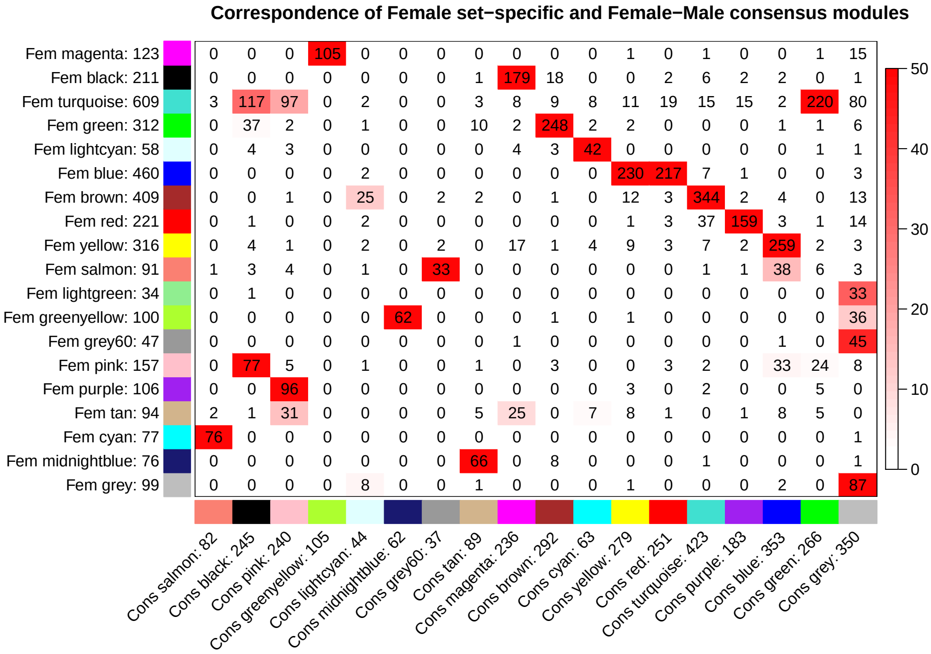
Relating the consensus modules to female set-specific modules

After having worked through the one-step network construction and module detection in the female mouse liver data, we first loaded the femaledata and renamed them so that they do not conflict with the consensus data and we loaded the results of the consensus module identification.

The consensus network analysis results are represented by the variables consMEs , moduleLabels , moduleColors , and consTree. We had then to relate the female modules to the consensus modules. We calculated the overlaps of each pair of female-consensus modules, and used the Fisher's exact test (also known as hypergeometric test) to assign a p-value to each of the pairwise overlaps. To display the p-value and count tables in an informative way, we created a color-coded table of the intersection counts. The colors indicate the p-value significance.

The resulting color-coded table is shown in Fig. 1. The table indicates that most female set-specific modules have a consensus counterpart. This indirectly shows that the module structure in the male liver expression data is very similar to the female data. Interestingly, there are two female modules (labeled by grey60 and lightgreen colors) that have no consensus counterpart; almost all genes in the two female modules are labeled “grey”, that is unassigned, in the consensus network.

The following figure show the Correspondence of female set-specific modules and the female-male consensus modules. Each row of the table corresponds to one female set-specific module (labeled by color as well as text), and each column corresponds to one consensus module. Numbers in the table indicate gene counts in the intersection of the corresponding modules. Coloring of the table encodes - log(p), with p being the Fisher's exact test p-value for the overlap of the two modules. The stronger the red color, the more significant the overlap is. The table indicates that most female set-specific modules have a consensus counterpart.



	Cons salmon: 82	Cons black: 245	Cons pink: 240	Cons greenyellow: 105	Cons lightcyan: 44	Cons midnightblue: 62	Cons grey60: 37	Cons tan: 89	Cons magenta: 236	Cons brown: 292	Cons cyan: 63	Cons yellow: 279	Cons red: 251	Cons turquoise: 423	Cons purple: 183	Cons blue: 353	Cons green: 266	Cons grey: 350
Fem magenta: 123	0	0	0	105	0	0	0	0	0	0	0	1	0	1	0	0	1	15
Fem black: 211	0	0	0	0	0	0	0	1	179	18	0	0	2	6	2	2	0	1
Fem turquoise: 609	3	117	97	0	2	0	0	3	8	9	8	11	19	15	15	2	220	80
Fem green: 312	0	37	2	0	1	0	0	10	2	248	2	2	0	0	0	1	1	6
Fem lightcyan: 58	0	4	3	0	0	0	0	0	4	3	42	0	0	0	0	0	1	1
Fem blue: 460	0	0	0	0	2	0	0	0	0	0	0	230	217	7	1	0	0	3
Fem brown: 409	0	0	1	0	25	0	2	2	0	1	0	12	3	344	2	4	0	13
Fem red: 221	0	1	0	0	2	0	0	0	0	0	0	1	3	37	159	3	1	14
Fem yellow: 316	0	4	1	0	2	0	2	0	17	1	4	9	3	7	2	259	2	3
Fem salmon: 91	1	3	4	0	1	0	33	0	0	0	0	0	0	1	1	38	6	3
Fem lightgreen: 34	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33
Fem greenyellow: 100	0	0	0	0	0	62	0	0	0	1	0	1	0	0	0	0	0	36
Fem grey60: 47	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	45
Fem pink: 157	0	77	5	0	1	0	0	1	0	3	0	0	3	2	0	33	24	8
Fem purple: 106	0	0	96	0	0	0	0	0	0	0	0	3	0	2	0	0	5	0
Fem tan: 94	2	1	31	0	0	0	0	5	25	0	7	8	1	0	1	8	5	0
Fem cyan: 77	76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Fem midnightblue: 76	0	0	0	0	0	0	0	66	0	8	0	0	0	1	0	0	0	1
Fem grey: 99	0	0	0	0	8	0	0	1	0	0	0	1	0	0	0	2	0	87

Figure 5.1: ConsensusVsFemaleModules

Chapter 6

Relating consensus module to external microarray sample traits and exporting the results of network analysis

In this section we illustrate the use of module eigengenes to relate consensus modules to external microarray sample information such as classical clinical traits. In this analysis we have available several clinical traits. We relate the traits to consensus module eigengenes in each of the two sets. It is important to keep in mind that while the consensus modules is a single module assignment for all genes, the module eigengenes represent the modules in each of the two sets. In other words, we have a single module assignment for each gene, but we have two sets of consensus module eigengenes, because a given module (set of genes) has a particular expression profile in the female mice, and a different expression profile in the male mice.Similarly, we have the trait data separately for the female and for the male mice. We display the module-trait relationships using a color-coded table and print the correlations and the corresponding p-values, and color-code the entris by the p-value significance. The two tables are shown in the following figures. The two module-trait relationship tables look similar but they are not the same. For example, they both identify the turquoise, purple and green modules as highly related to weight, although the actual correlations and p-values differ slightly. There are several ways of forming a measure of module-trait relationships that summarizes the two sets into one measure. We form a very conservative one (for each module-trait pair we take the correlation that has the lower absolute value in the two sets if the two correlations have the same sign, and zero relationship if the two correlations have opposite signs). The consensus module–trait relationships are displayed again using a color-coded table.

The table is shown in third figure. The advantage of the consensus relationship table is that it isolates the module-trait relationships that are present in both sets, and hence may be in a sense considered validated. For example, we confirm that the turquoise, purple, and green modules are highly related to weight in both sets; the brown module is highly related to insulin levels etc.

Among the following figures: The first shows the Relationships of consensus module eigengenes and clinical traits in the female data. Each row in the table corresponds to a consensus module, and each column to a trait. Numbers in the table report the correlations of the corresponding module eigengenes and traits, with the p-values printed below the correlations in parentheses.The table is color coded by correlation according to the color legend.

The second shows the Relationships of consensus module eigengenes and clinical traits in the male data. Each row in the table corresponds to a consensus module, and each column to a trait. Numbers in the table report the correlations of the corresponding module eigengenes and traits, with the p-values printed below the correlations in parentheses. The table is color coded by correlation according to the color legend.

And the third, the Consensus relationships of consensus module eigengenes and clinical traits across the female and male data. Each row in the table corresponds to a consensus module, and each column to a trait. Numbers in the table report the consensus correlations of the corresponding module eigengenes and traits, with the p-values printed below the correlations in parentheses. The table is color coded by correlation according to the color legend. Missing (NA) entries indicate that the correlations in the male and female data sets have opposite signs and no consensus can be formed.

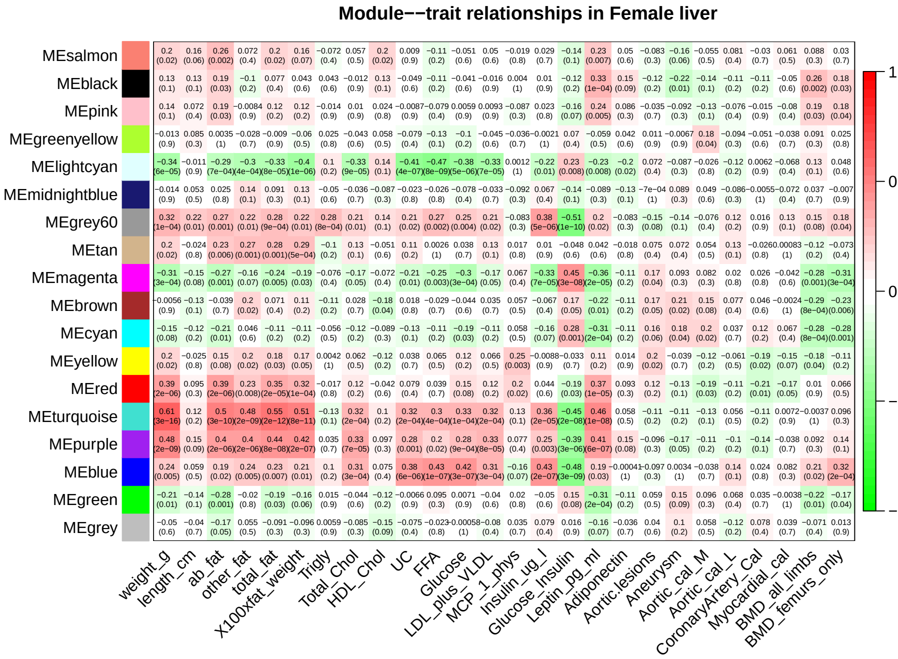


Figure 6.1: images/ModuleTraitRelationships-female

6.0.1 Exporting results of the network analysis

We now put together a data frame that summarizes the results of network analysis, namely the gene significances (GS) and module memberships (also known as kME) of all probes. We start by loading the gene annotation table.

We next (re-)calculate the module eigengenes in the ”alphabetic” order and calculate the gene significances and module memberships in each data set.

We perform a very simple ”meta-analysis” by combining the Z scores of correlations from each set to form a meta-Z score and the corresponding p-value.

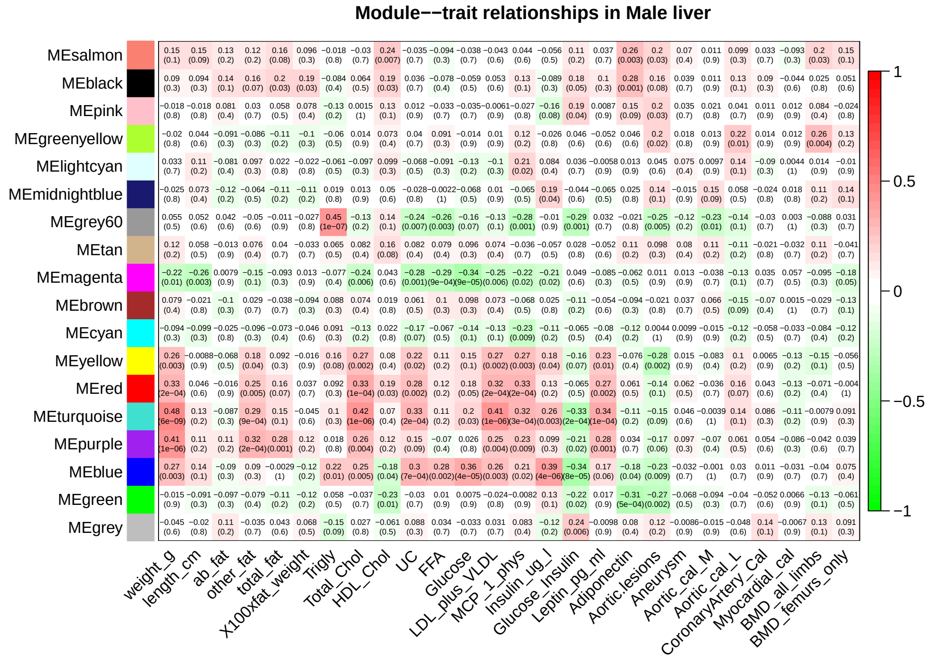


Figure 6.2: ModuleTraitRelationships-male

Next we form matrices holding the GS and kME. We use a simple re-shaping trick to put the values and the associated p-values and meta-analysis results next to one another.

Finally we put together the full information data frame and write it into a plain text CSV file that can be read by standard spreadsheet programs. And we should note that the probes are not sorted in any particular way.

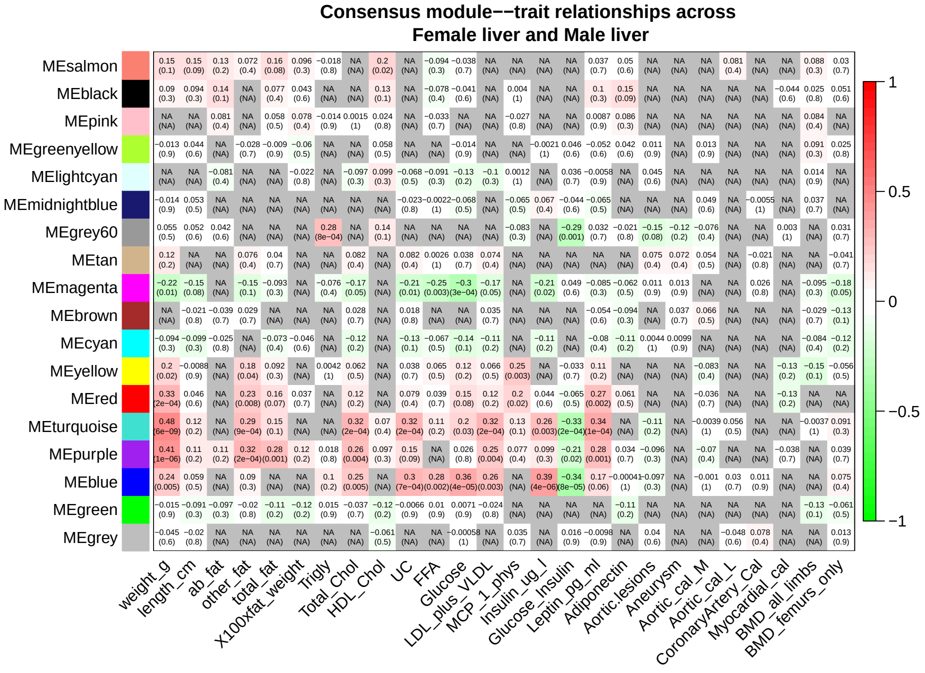


Figure 6.3: ModuleTraitRelationships-consensus

Chapter 7

Studying and comparing the relationships among modules and traits between the two data sets, including the visualization of consensus eigengene networks and the results of the differential analysis

In this section the consensus eigengene networks in the female and male data sets are compared(such comparison is often called differential analysis). Consensus eigengene networks capture the relationships among consensus modules; the relationships are quantified by eigengene correlations. We first extend the consensus eigengenes by adding the body weight clinical trait as an additional "eigengene". We now call the function "plotEigengeneNetworks" that performs the differential analysis.

The plot shown in the following figure shows the eigengene dendrograms and eigengene network heatmaps, as well as two plots of network preservation between the two sets, namely a heatmap plot of the preservation adjacency and a barplot of mean preservation of relationships for each eigengene. The overall preservation of the two eigengene networks is 0.94, which is quite high. A visual inspection of the female and male network heatmap plots indicates that the inter-module relationships in the two data sets are indeed very similar. The red blocks along the diagonal in the network heatmaps indicate meta-modules, groups of correlated eigengenes, and these are largely preserved between the two sets. The preservation heatmap and barplots indicate that most relationships are very highly preserved. The message here is that inter-module relationships are strongly preserved across similar data sets and encode biologically meaningful information.

In other words, the figure show a Summary plot of consensus eigengene networks and their differential analysis. The top two panels show the dendrograms (clustering trees) of the consensus module eigengenes in the two sets indicated in the titles. Below, the eigengene networks in the two sets are shown as heatmaps labeled Female liver and Male liver. In the heatmaps, red denotes high adjacency (positive correlation) and green denotes low adjacency (negative correlation). The Preservation heatmap shows the preservation network, defined as one minus the absolute difference of the eigengene networks in the two data sets. The barplot shows the mean preservation of adjacency for each of the eigengenes to all other eigengenes (the barplot depicts the column means of the preservation heatmap).

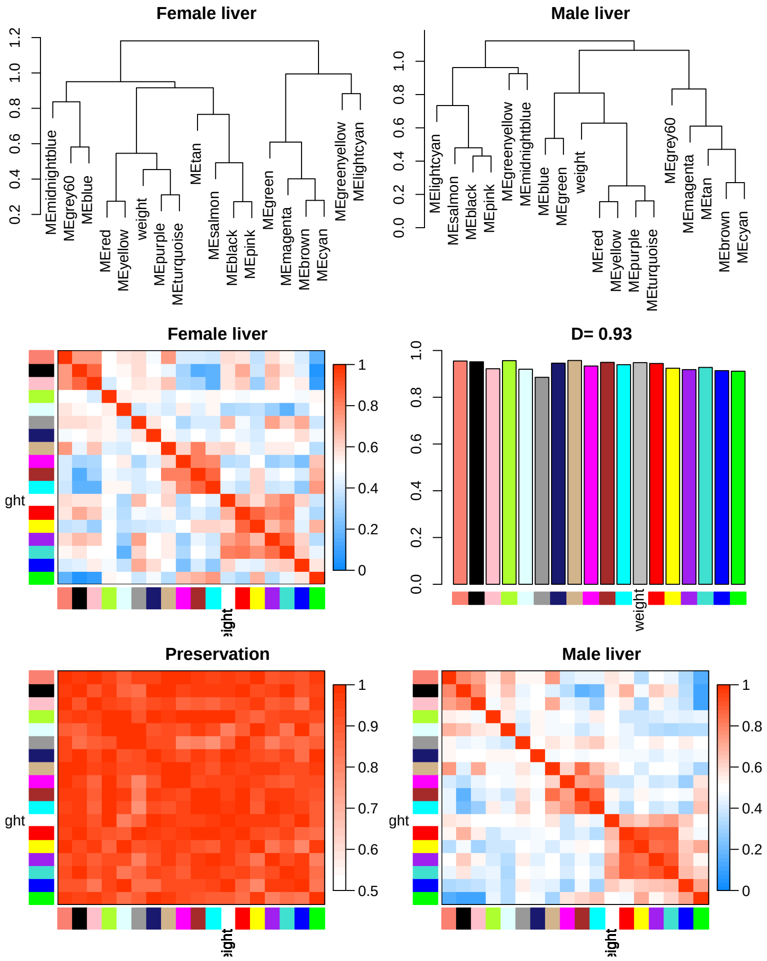


Figure 7.1: EigengeneNetworks

Chapter 8

Conclusion

A weighted network is the result the transformation of the correlation matrix into a matrix of connection strenghts using a power function. The use of weighted networks represents an improvement over unweighted networks that are based on dichotomizing the correlation matrix, because the continuous nature of the gene co-expression information is preserved and the results of weighted network analyses are highly robust, whereas unweighted networks display sensitivity to the choice of the cutoff threshold.

We applied the network construction algorithm to a subset of gene expression data from an F2 intercross between inbred strains C3H/HeJ and C57BL/6J. Liver gene expression data from 135 female mice were used for this analysis.

We were able to identify 12 distinct gene modules or groups of genes with high topological overlap. To distinguish between modules, we have designate each module by an arbitrary color. The number of genes included in the modules ranged from 34 (Light-yellow) to 772 (Red), and their mean overall connectivity (kall) ranged from 6.49 (Salmon) to 27.58 (Brown). We also defined the intramodular connectivity (kin) for each gene based on its Pearson correlation with all other genes in the module.

Chapter 9

Further Work

In further works, we may perform an analysis of simulated data, on gene expression data to evaluate defferent module detection methods and gene screening approaches. And doing so while steel focusing on key concepts of weighted gene co-expression network analysis (WGCNA) in the R environment. And we may even explore the meta-analysis of several data sets.

Chapter 10

References

Anatole Ghazalpour, Sudheer Doss, Bin Zhang, Susanna Wang, Christopher Plaisier, Ruth Castellanos1,Alec Brozell1, Eric E. Schadt, Thomas A. Drake, Aldons J. Lusis, Steve Horvath, Integrating Genetic and NetworkAnalysis to Characterize Genes Related to Mouse Weight, PLoS Genet, 2006, volume2, e13, link to journal.