National Technical University of Athens
School of Mechanical Engineering
Department of Mechanical Design and automatic Control

Application of computational methods for discovering the pathological mechanisms of osteoarthritis disease

Diploma Thesis

Smaragda Dimitrakopoulou

Supervised by

Associate Prof. Dr. Leonidas . Alexopoulos

Athens,2019

Abstract

Osteoarthritis (OA) is the most common form of joint disease and a major cause of disability, particularly in the second half of life. It is a whole joint disease involving several tissues: cartilage, synovium, meniscus and subchondral bone. Until today, many disease-modifying treatments, focusing on one tissue pathogenesis, have been studied. However they had limited success. Thus, the integration of all tissues might be necessary to increase an understanding of the disease. In this diploma thesis pathway and network-based approaches were used to develop a functional description of this multi-tissue disease.

Differential Expression Analysis and Weighted Gene Co-Expression Network Analysis were implemented using four microarray datasets corresponding to the four tissues. Through WGCNA, a set of modules were identified. Highly correlated modules were merged creating meta-modules. The genes included in each meta-module were used for the identification of important pathways through Pathway Analysis. As a final step, candidate drugs for OA were evaluated.

The Pathway Analysis underlined the correlation of the constructed meta-modules with specific biological functions. Pathways related to the immune system and the extracellular matrix were identified as common pathological mechanisms between the tissues. Furthermore, many signalling pathways known for their connection with OA pathogenesis were identified. The main targets of 11 drugs were detected. The drug evaluation consisted of checking if the drug's targets were included in any meta-module. 3 out of 11, Sorafenib, Raloxifene and PD169316 satisfied this criterion. First experimental results on drug screening independently identified Sorafenib and PD169316 as the most promising compounds from a similar library of 9 drugs.

In summary, statistical and correlation based approaches could be applied in parallel to identify molecular mechanisms involved in multiple tissues in OA. Further work will try to incorporate into drug evaluation the results of the Pathway Analysis and the network properties.

Περίληψη

Η οστεοαρθρίτιδα είναι μία από τις πιο συνήθεις ασθένειες των αρθρώσεων και μία από τις πιο σημαντικές αιτίες εμφάνισης αναπηρίας στη μέση ηλικία. Έχει πλέον αναγνωριστεί ως ασθένεια όλης της άρθρωσης και περιλαμβάνει αλλαγές στο χόνδρο, στην αρθρική μεμβράνη, στη κνήμη και στο μηνίσκο. Μέχρι σήμερα, δεν έχει βρεθεί φάρμακο κατα της οστεοαρθρίτιδας καθώς οι έρευνες επικεντρώνονται στην παθογένεια μόνο ενός από τους 4 ιστούς/οστά. Συνεπώς, σε αυτή την διπλωματική εργασία η ενσωμάτωση δεδομένων από όλους τους εμπλεκόμενους στην οστεοαρθρίτιδα ιστούς/οστά πραγματοποιείται προκειμένου να κατανοηθούν πληρέστερα οι μηχανισμοί παθογένειας της ασθένειας.

Σε αυτή τη διπλωματική εργασία πραγματοποιήθηκε γονιδιακή ανάλυση και ανάλυση δικτύων συνεκφραζόμενων γονιδίων, χρησιμοποιώντας δεδομένα από τέσσερα πειράματα, ένα για κάθε ιστό/οστό. Μέσω της ανάλυσης δικτύων συνεκφραζόμενων γονιδίων βρέθηκαν σύνολα που περιλάμβαναν συσχετιζόμενα σε όλα τα δείγματα γονίδια. Τα αλληλοσυσχετιζόμενα σύνολα ενώθηκαν δημιουργώντας μετα-σύνολα. Τα γονίδια του κάθε μετα-συνόλου χρησιμοποιήθηκαν για την εύρεση σημαντικών βιολογικών μονοπατιών μέσω της ανίστοιχης ανάλυσης. Τέλος, υποψήφια φάρμακα κατα της οστεοαρθρίτιδας αξιολογήθηκαν.

Η ανάλυση βιολογικών μονοπατιών υπογράμμισε τη συσχέτιση των μετα-συνόλων με συγκεκριμένες βιολογικές λειτουργίες. Βιολογικά μονοπάτια που σχετίζονται με το ανοσοποιητικό σύστημα και την εξωκυτάρια μήτρα αποδείχθηκαν σημαντικά και στους 4 ιστούς/όστα. Επίσης βρέθηκαν σηματοδοτικά βιολογικά μονοπάτια, γνωστά για το ρόλο τους στην οστεοαρθρίτιδα. 11 φάρμακα αξιολογήθηκαν σύμφωνα με το αν περιλαμβάνονταν οι βασικοί στόχοι τους σε κάποιο μετα-σύνολο. 3 από τα 11, τα Sorafenib, Raloxifene και PD169316 ικανοποιούσαν αυτό το κριτήριο. Μία αρχική ανεξάρτητη πειραματική έρευνα αξιολόγησε 9 φάρμακα και βρήκε τα Sorafenib και PD169316 ως πιο ελπιδοφόρα φάρμακα κατα της οστεοαρθρίτιδας.

Εν συντομία, μέθοδοι που βασίζονται στη στατιστική και στην αλληλοσυσχέτιση δεδομένων χρησιμοποιήθηκαν μαζί προκειμένου να ανακαλύψουν σημαντικούς μηχανισμούς παθογένειας της οστεοαρθρίτιδας. Μελλοντικά θα γίνει προσπάθεια να συμπεριληφθούν τόσο τα δίκτυα συνεκφραζόμενων γονιδίων όσο και τα βιολογικά μονοπάτια στην αξιολόγηση των φαρμάκων.

Acknowledgments

I would like to thank my supervisor Associate. Prof. Dr. Leonidas Alexopoulos for his support and guidance during my diploma thesis. Furthermore I would like to thank all the members of the Biomedical Systems Laboratory for their support and their willingness to help me.

I would like to deeply thank Michael Neidlin for the opportunity he gave me to learn, his willingness to help me in everything I asked for and his extensive hours of advice and guidance throughout the integration of this diploma thesis.

Also, I would like to thank my friends for always being there when I need them. Finally, I would like to thank my parents,my sister and my partner who love and support me in everything I do and they are indeed the greatest inspiration of all.

Contents

CHAPTER 1

Introduction

1.1 Osteoarthritis

Osteoarthritis (OA) is the most common degenerative joint disorder that affects one or several diarthrodial joints, including small joints (such as those in the hand) and large joints (such as the knee and hip joints). The incidence of OA increases with age and by 65 years approximately 80% of the population has some radiographic evidence of disease. Primary signs include pain, transient morning stiffness and crepitus on joint motion (a grating sound or sensation produced in the joint) that lead to instability and physical disability.[1]

OA can be classified as primary (or idiopathic) and secondary. Primary OA results from a combination of risk factors, with increasing age and obesity being the most prominent. Other risk factors include sex, joint biomechanics and genetic factors.[2] Secondary OA is based on the attribution to recognized causative factors, such as trauma, surgery on the joint structures and abnormal joints at birth.

Osteoarthritis is now considered a disease of the whole joint [3], including alterations in the articular cartilage, subchondral bone, ligaments, capsule and synovial membrane, ultimately leading to joint failure as shown in figure 1.1. Among the structural damages to the joint are loss of cartilage, osteophyte formation, subchondral bone changes and meniscal alterations.

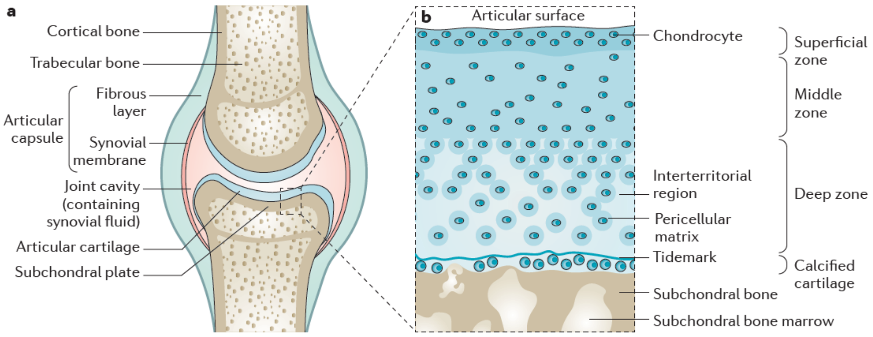


Figure 1.1: (a)Diarthrodial joints join two adjacent bones that are covered by a layer of specialized articular cartilage and are encased in a connective tissue capsule lined by a synovial membrane, consisting of a thin cell layer of macrophages and fibroblasts [3],(b)Cross-section of the articular surface of a diathrodial joint illustrating schematically

Cartilage Degeneration

The articular cartilage is composed of water ($\geq 70\%$) and organic extracellular matrix components, mainly type II collagen and aggrecan or other proteoglycans. The cartilage matrix is avascular and aneural and is populated by a single cell type: the chondrocyte. In OA, the cartilage matrix undergoes striking changes in its composition and structure. Initially, surface fibrillations appear and, as the pathological process continues, deep fissures associated with exfoliation of cartilage fragments develop, ultimately leading to delamination and exposure of the underlying calcified cartilage and bone. As the disease progresses, the proteoglycans become depleted followed by the erosion of the collagen network, which marks irreversible progression.

Periarticular Bone

The bone beneath the articular cartilage is organized into a plate-like layer of cortical bone and a contiguous region of cancellous bone. OA is accompanied by increases in the volume, thickness and contour of the cortical plate, alterations in the state of bone mineralization and material properties, changes in the subchondral trabecular bone architecture and bone mass, the formation of bone cysts, and the appearance of bone marrow lesions and osteophytes. Bone may also undergo direct physical damage that results in the formation of microcracks or fissures within the cortical or trabecular bone. Subchondral bone cysts (fluid-filled holes) are a common feature of advanced OA.

Synovium

The synovium includes the synovial membrane and the fluid. In OA, the role of synovial inflammation in the pathophysiology of OA is now widely accepted. Synovitis has been considered secondary to the cartilage changes yet findings indicate that synovial inflammation could be a component of the early events leading to the clinical stage of OA. Synovial inflammation leads to the production and release of pro-inflammatory cytokines and several other inflammatory mediators.

1.2 Drugs

Topical, oral and injectable pharmacological treatments are available for individuals with OA. Current therapies are at best moderately effective pain relievers, and it is worth noting that studies report that most people with OA have persistent pain despite taking all their prescribed therapies. First-line therapies include topical NSAIDs and oral paracetamol. Until today,a number of potential disease-modifying pharmacological therapies have been investigated with disappointing results.[4] However the chance of successful OA drug development may improve in the future as a consequence of shift of focus into system-oriented studies.[5]

1.3 Systems Biology

Systems orientated research offers the possibility of identifying novel therapeutic targets and relevant diagnostic markers for complex diseases such as osteoarthritis. Systems orientated studies treat cells and tissues as biological systems. A biological system is a set of elements (e.g., genes, proteins, and metabolites) with multiple and diverse functions; these elements interact in a specific and non linear manner to produce coherent behaviours over time. Interaction networks may be generated from the elements of a biological system which can facilitate an understanding of the architecture,activity, and key players in that system.

Network-based Systems orientated studies make use of known or inferred functional and physical interactions between the elements of a system or can be developed from statistical associations (e.g., correlations between expression values). Data are often collected from disparate sources and organized into a coherent structure that can be interrogated by graph theory or logical (probabilistic) approaches. Network medicine postulates a ”disease module” hypothesis, where disease-associated genes or proteins are likely share the same topological neighbourhood in a network. Defining communities of network elements (genes, proteins) is a useful way to identify elements that have a close relationship, shared functionality, or disease association.

A systems biology approach to comprehending OA is founded on the hypothesis that OA is a multi-system disorder resulting from the dysfunction of a number of networks that, together, alter the homeostatic balance of the joint. Therefore, comprehensive and multisystem approaches are necessary to understand the complexity of OA and direct the development of innovative treatment strategies.

To date most studies, pertaining to using a systems approach in OA research, are principally based on interrogation of a single ”omics” survey in a single tissue at a single time point. Goldring et al [6] emphasised the importance of cytokines in initiation and progression of OA cartilage. Melas et al [7] re-established important genes to OA pathogenesis in cartilage like IL1B, TNF,IL6, as well as discovered some new key-players. Mariani et al [8] underlined the importance of MAPK pathway and Wnt signalling pathways in OA cartilage. Brophy et al [9] studied the meniscus of OA and non-OA patients. He identified key players in meniscus osteoarthritis pathogenesis. Park et al [10] studied the gene expression profile in osteoarthritis synoviym, identified important genes and pathways and compared the findings with the ones derived from studies focusing on osteoarthritis cartilage. Interestingly, his study concluded that OA cartilage and OA sunovium are driven by different pathological mechanisms.

The need for a different approach has emerged, for the integration of all tissues related to OA.[11] To that end, this diploma thesis uses co-expression networks in order to identify biological functions and elements important to all tissues involved in OA. Weighted Gene Co-expression Network Analysis was chosen for the construction of the co-expression networks, as it has provided important insights in the OA mechanisms, common across the species [5]

CHAPTER 2

Theory

In this chapter the theory of the methods used in this diploma thesis is presented.

Firstly, background correction and normalisation of the microarray data will be discussed. These two procedures are of high importance, as they ensure that any analysis afterwards provides significant biological insights. In the next section, Differential Expression Analysis will be presented, one of the most basic analyses done in every microarray data in order to distinguish whether there is a difference in the expression levels of the genes between samples belonging to different conditions. Then Pathway Analysis will be presented. Pathway analysis uses as input the results of the Differential Expression Analysis and provides more biologically meaningful results like altered biological processes among samples of the different conditions. The next section will refer to a different way of analysing the microarray data,the construction of a network taking into account the data's topological properties. Specifically, Weighted Gene Co-Expression Network Analysis will be discussed. Finally methods to ensure the stability of such networks will be presented.

2.1 Background Correction and Normalisation

The goal of most microarray experiments is to survey patterns of gene expression by assaying the expression levels of thousands to tens of thousands of genes in a single assay.

Typically,DNA or RNA is first isolated from different tissues, developmental stages, disease states or samples subjected to appropriate treatments. The DNA or RNA is then labelled and hybridized to the arrays using an experimental strategy that allows expression to be assayed and compared between appropriate sample pairs. Common strategies include the use of a single label and independent arrays for each sample (single-Dye), or a single array with distinguishable fluorescent dye labels for the individual RNAs (Dual-Dye). Regardless of the approach chosen, the arrays are scanned after hybridization and independent grayscale images, typically 16−bit TIFF (Tagged Information File Format) images, are generated for each pair of samples to be compared. These images must then be analysed to identify the arrayed spots and to measure the relative fluorescence intensities for each element.

The hypothesis underlying microarray analysis is that the measured intensities for each arrayed gene represent its relative expression level. Biologically relevant patterns of expression are typically identified by comparing measured expression levels between different states on a gene−by-gene basis. But before the levels can be compared appropriately, a number of transformations must be carried out on the data to eliminate questionable or low−quality measurements and to adjust the measured intensities to facilitate comparisons between classes of samples.

2.1.1 Background Correction

Background Correction is the process of removing background noise from the signal intensities. Background noise is the measurement of signal intensity that is caused by either autofluorescense of the array surface or by non-specific binding. In other words, background noise is the proportion of the measured intensities that do not reflect the expression of the genes.

In two-channel studies the data extracted from the image processing consist of a measure for the spot intensity and its local background, for each spot on the array and for its color. Similarly, in one channel experiments, the data extracted from the image processing consist of a measure for the spot intensity and its local background, for each spot on the array.[12] One of the first methods used in the bibliography for correcting the spot intensities is by simply subtracting the background intensities. If $TrueIntensity(i)$ denotes to the true signal intensity of the gene i, then $TrueIntensity(i)$ is calculated as shown in equation 2.1.

$$TrueIntensity(i) = PM(i) - MM(i) \tag{2.1}$$

where $PM(i)$ and $MM(i)$ denotes the measured spot intensity of the gene i and its local background, respectively.

One of the most common problems of simple subtraction as a background method is that it can result to some values being negative. Therefore, when the background corrected data are log2-transformed, missing values are obtained. This leads not only to loss of information but also to bias. Therefore, many methods trying to eliminate this phenomenon were introduced, like replacing every negative value after background correction with a standard small positive value. However, simple subtraction has many other disadvantages as it simplifies the background noise existing in the data.

These disadvantages that methods depending on the simple subtraction of the background intensities have, led to the appearance of a new approach regarding the background correction of the data. For a given probe p, let θ_p denote the expression level of p, that is, the concentration of RNA transcripts homologous to probe p in the unlabelled sample, and i_{ps} the background-corrected spot intensity as measured on array s. A simple model relating i_{ps} to θ_p is

$$i_{ps} \sim k_s * a_p * \theta_p \ , \ p = 1, \ldots, P, \ s = 1, \ldots, S \tag{2.2}$$

where k_s is an array$-$specific constant of proportionality and a_p is a probe$-$specific constant. There are P probes and S arrays.[12] The model states that for a given array, the intensities i_{ps} are proportional to the expression levels θ_p with coefficients that vary for the different probes.

Rewriting in the log$-$scale (using, by convention, base 2 logarithms) and introducing an error term, the above equation becomes as follows

$$\log_2 i_{ps} = \log_2 k_s + \log_2 a_p + \log_2 \theta_p + e_{ps}, p = 1, \ldots, P, s = 1, \ldots, S \tag{2.3}$$

For the identification of the parameters a_p,k_s and e_{ps} many methods exist in the bibliography,[13],[14] with the most popular one being RMA by Irizary et al [15]

2.1.2 Normalisation

Single$-$Dye cDNA microarrays

The backgroundcorrected spot intensities should reflect the abundance of the corresponding target genes in the samples. However, often the relation is not that of simple proportionality: the true signals may be distorted in various ways. One of these is spatial bias. Spatial bias is the presence of regions with overall higher or lower intensity levels across the samples, as shown in figure 2.1.

Another form of distortion may appear when data from replicate arrays are compared graphically and various forms of systematic departure from the identity line are observed. This phenomenon is termed relative intensity bias according to David Edwards [12] and can be connected to array effects, systematic errors or difference in the average relative expression levels across samples. Some of the distortion that may appear in the data plots are shown in figure 2.2

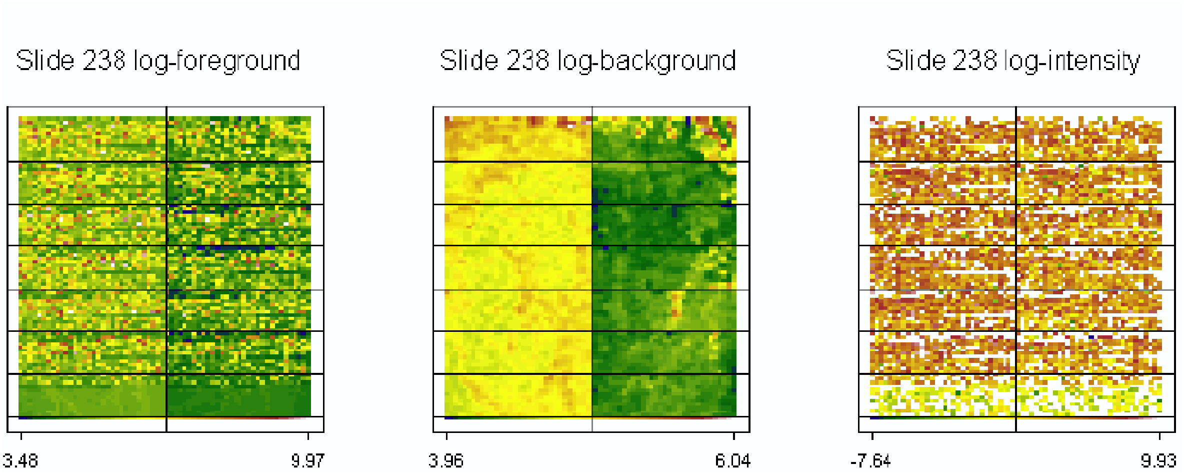


Figure 2.1: Spatial Background Bias [12]

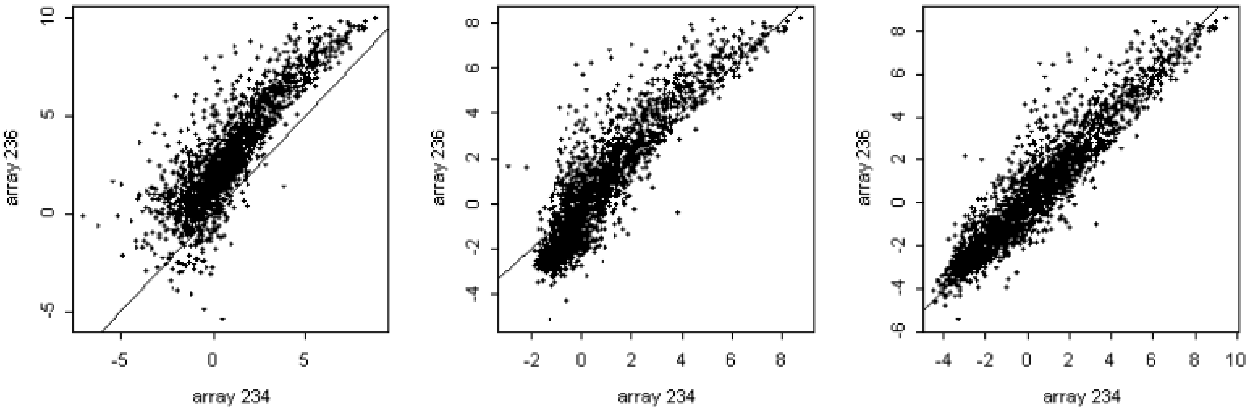


Figure 2.2: Examples of relative intensity bias [12]

Dual-Dye cDNA microarrays

For Dual−dye cDNA microarrays, the purpose of dye normalization is to balance the fluorescence intensities of the two dyes used (green Cy3 and red Cy5 dye) as well as to allow the comparison of expression levels across experiments (slides). Dye bias can be most obviously seen in an experiment where two identical mRNA samples are labeled with different dyes and subsequently hybridized to the same slide. In this situation, it is rare to have the dye intensities equal on average and often the intensities are higher for the green dye.

This bias can stem from a variety of factors including physical properties of the dyes (heat and light sensitivity, relative half−life), efficiency of dye incorporation, experimental variability in probe coupling and processing procedures, and scanner settings at the data collection step.

Furthermore, the relative gene expression levels (as measured by log ratios) from replicate experiments may have different spreads due to differences in experimental conditions, including unequal quantities of starting RNA or systematic bias. Some scale adjustment may then be required so that the relative expression levels from one particular experiment do not dominate the average relative expression levels across replicate experiments.

There are various methods available in the bibliography for the normalisation of both the dual−dye and the single-dye cDNA microarrays. One of the most frequently used normalisation methods is the " quantile normalisation ". The goal of quantile normalization is to make the distribution of probe intensities the same for arrays $i = 1,\ldots,I$. The normalization maps probe level data from all arrays,$i = 1,\ldots,I$, so that an I -dimensional quantile– quantile plot follows the I -dimensional identity line.

2.1.3 Methods

In the figure 2.3 many of the available methods of background correction and normalisation can be seen.

2.1.4 MA Plots

Almost always, biological comparisons made on microarrays are very specific in nature, i.e. only a small proportion of genes are expected to be differentially expressed. Therefore, the remaining genes are expected to have constant expression and so can be used as indicators of the whether the background correction and normalisation method used is appropriate for the data.

Dual-Dye cDNA microarrays

For a spot j, $j = 1,\ldots p$, let R_j and G_j denote the measured fluorescence intensities (after background correction and normalisation) for the red and green dyes, respectively. In order to test if

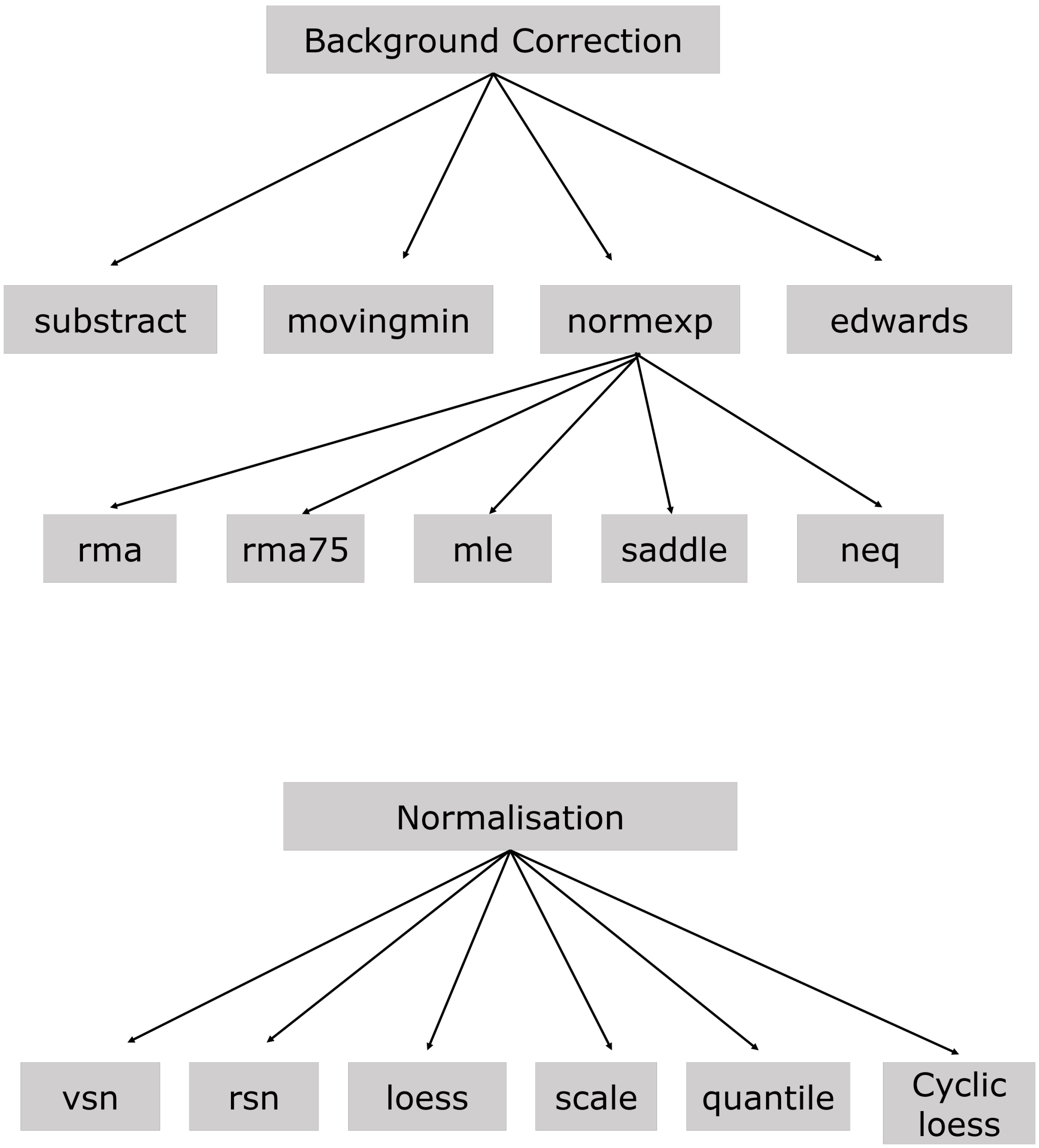


Figure 2.3: Different methods of background correction and normalisation. For further information regarding the background correction methods look into [13],[16],[14],[12],[17],[18] For further information regarding the normalisation methods look into [19],[20],[21],[22], [14]

bias and background noise is successfully removed from the data, it is useful to plot the log intensity ratio $M = log_2\frac{R}{G}$ vs. the mean log-intensity $A = log_2\sqrt{RG}$ as described by Yang et al [21] An M vs. A plot amounts to a 45o counter-clockwise rotation of the (log_2G, log_2R)-coordinate system, followed by scaling of the coordinates.

In the M vs. A plot the points should be scattered around the zero line, with only a few of them having great M values. If the points are scattered around a line parallel to the zero line or a line that displays a gradient, the data have remaining bias and some systematic errors have not been successfully removed. This is also indicated if the points in the plot display curvature.

Single-Dye cDNA microarrays

In single-Dye cDNA data, it is straightforward to adapt the same approach proposed by Yang et al. [20] for the Dual-Dye cDNA data in order to check if spatial and intensity bias are successfully removed from the data. The method as changed for adapting to one-channel data can be described as follows.[12]

If $V^1 = (V_1^1, \ldots, V_P^1)$ and $V^2 = (V_1^2, \ldots, V_P^2)$ encodes the log_2, background corrected and normalised intensities of the genes' expressions of the samples belonging to either group 1 or 2 respectively:

$$M = V^1 - V^2 \tag{2.4}$$

$$A = V^1 + V^2 \tag{2.5}$$

As in the Dual-Dye microarrays the points in the M vs. A plot should be scattered around the zero line. Anything else indicates that there is still bias in the data.

2.2 Differential Expression Analysis

One of the reasons to carry out a microarray experiment is to monitor the expression level of genes at a genome scale. Patterns could be derived from analysing the gene expression data of the genes, and new insights could be gained into the underlying biology.

Fundamental to the task of analysing gene expression data is the need to identify genes whose patterns of expression differ according to phenotype or experimental condition. Such an analysis called Differential Expression Analysis and most times focuses on identifying the genes which change their expression among samples belonging to one of two conditions, eg. healthy and diseased.

In order to identify if one gene is differentially expressed in the two conditions, two measures should be calculated, a measure of magnitude and a measure of significance. In a Differential Expression Analysis, these two measures are calculated for each and every gene, separately.

2.2.1 Measure of magnitude

Assume that an experiment population consists of N samples. From N samples, m are control samples (eg. normal) and N-m are case samples (eg. diseased). Furthermore, assume that $E_g(i)$ encodes the expression level of the gene g of the i^{th} sample. The calculation of the measure of magnitude aims to clarify if there is an important and constant difference in the expression of the gene g among the two groups. In other words, it answers the biological question of whether the expression of the gene g is importantly higher or lower in the case population compared to the control population (eg. disease vs normal samples). Usually, as a measure of magnitude is used the $\log_2 FoldChange$ which is is calculated as shown in equation 2.6.

$$\overline{E}_{control} = \frac{\sum_{iecontrol} E_g(i)}{m} \tag{2.6}$$

$$\overline{E}_{case} = \frac{\sum_{jecase} E_g(j)}{N-m} \tag{2.7}$$

$$= \log_2 \frac{\overline{E}_{case}}{\overline{E}_{control}} \tag{2.8}$$

$$\tag{2.9}$$

In the $\log_2 FoldChange$, the mean value of the genes' expression levels of control and case samples, $\overline{E}_{control}$ and $\overline{E}_{case}$ respectively, are used. In this way, it is ensured that the observed difference in the expression between the groups is constant and not found by chance. The $2-fold$ logarithm is used in order to easily interpret the results. For example, a $\log_2 FoldChange$ of 1 means that the average expression of the gene in the case samples is doubled compared to the one of the control samples.

2.2.2 Measure of Significance

The measure of significance answers the following biological question:

" *Is the observed difference in the expression level of the gene g between the control and the case samples statistically significant*? "

In order to answer this question a statistical hypothesis is tested and a p-value is calculated. The p-value is the probability that the null hypothesis is true. In Differential Expression Analysis the null hypothesis is that the difference in the expression levels of the gene depends on the change of the condition.

Assume that $x(i)$ and $y(i)$ with $(i = 1, N)$ are the values of two continuous variables, X and Y respectively. In statistical analysis the dependence of the variable Y from the variable X can be tested by fitting best a linear regression through the points $(x(i), y(i))$ and calculating the linear coefficient R^2. Mathematically, a simple linear regression can be written as shown in equation 2.10

$$Y = b_0 + b_1 X \tag{2.10}$$

In equation 2.10, b_0 and b_1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, b_0 and b_1 are known as the model coefficients or parameters. The goal is to estimate b_0 and b_1 in order the line to fit best the available data. In other words, b_0 and b_1 should be calculated so that the line is as close as possible to the given points. There are a number of ways of measuring closeness. However the most common approach involves minimizing the least square criterion. Let $\hat{y}(i) = b_0 + b_1 x(i)$ be the prediction of the linear model for Y based on the i_{th} value of X. Then

$$e_i = y(i) - \hat{y}(i) \tag{2.11}$$

represents the i_{th} residual,which is the difference between the i_{th} observed response and the i_{th} predicted by the linear model value.

The residual sum of squares (SS) is calculated as shown in equation 2.12

$$SS = e_1^2 + e_1^2 + e_2^2 + \ldots + e_N^2 \tag{2.12}$$

The least square approach chooses the b_0 and b_1 in order to minimize the SS. An example of best fit using the least square approach can be seen in the figure 2.4

After the identification of the line that best fits the available data, R^2 can be calculated. The R^2 statistic provides an alternative measure of fit. It takes the form of proportion− the proportion of variance explained− and so it always takes a value between 0 and 1. R^2 can be calculated as shown in the equation 2.13

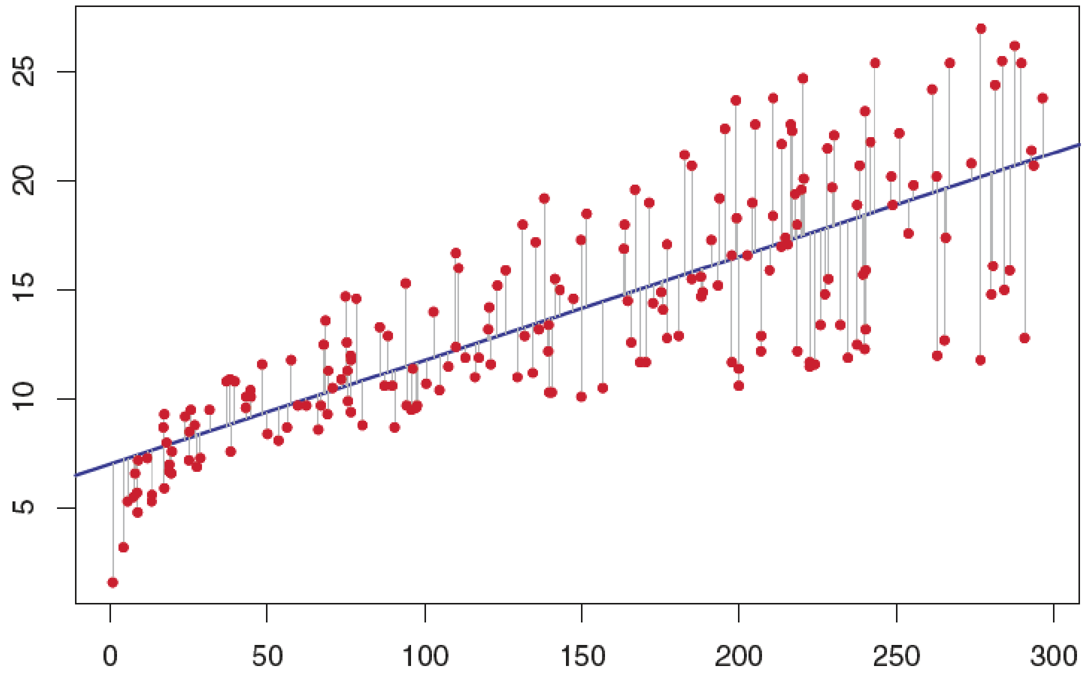


Figure 2.4: Best fit Line

$$R^2 = \frac{TSS - RSS}{TSS} \tag{2.13}$$

TSS measures the total variance in the response of Y and can be thought of as the amount of variability inherent in the response before the regression is performed. It is calculated as shown in equation 2.14

$$TSS = \frac{SS(Y_{mean})}{N} \tag{2.14}$$

In contrast, RSS measures the amount of variability that is left unexplained after performing the regression and can be calculated as shown in equation 2.15

$$RSS = \frac{SS(fit)}{N} \tag{2.15}$$

Hence, $TSS - RSS$ measures the amount of variability in the response that is explained (or removed) by performing the regression, and R^2 measures the proportion of variability in Y that can be explained using X. An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression, therefore that the variable X and Y are greatly dependent.

Finally in order to establish that R^2 is statistically important, and therefore the correlation between the two variables, a p−value is calculated using F−test. For the calculation of the p−value, the F variable is defined according to the equation 2.16

$$F = \frac{SS(mean) - SS(fit)}{SS(fit)} * \frac{N - p_{fit}}{p_{fit} - p_{mean}} \tag{2.16}$$

where p_{fit} encodes the degrees of freedom of the linear fit and is equal to the model parameters, hence 2. Similarly, p_{mean} encodes the degrees of freedom if there was no linear relationship between the variables X and Y. In that case the slope parameter would be equal to zero and therefore the model parameters would be reduced to 1. The p-value is defined by checking the corresponding to the sample size and degrees of freedom F-distribution, as shown in figure 2.5.

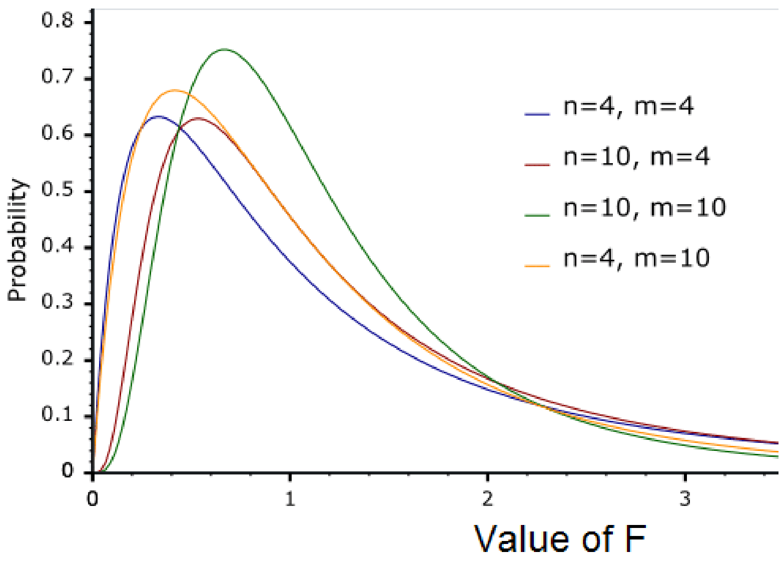


Figure 2.5: F-Distribution

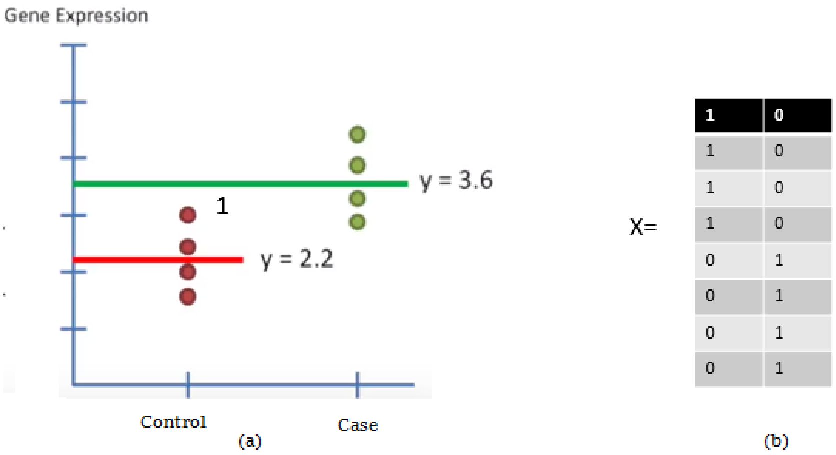
In the differential expression analysis, the variable X is categorical, as it corresponds to the state of the samples (control or case study), and therefore it can take only two values, 0 or 1 . On the other hand, the variable Y is continuous, as it corresponds to the gene expression levels of the gene g. That is:

$$Y(i) = E_g(i) \tag{2.17}$$

In order to find in this case if the correlation of the variables X and Y is significant, the same procedure, as explained for the two continuous variables, can be followed. In other words, a linear regression that best fits the data should be defined and then the statistical R^2 should be calculated and tested for its significance by an F−test. In this case, though, the linear regression that best fits the data is not defined by the least square approach, but is calculated by the equation 2.18

$$Y = X * [\overline{E}_{control}\overline{E}_{case}]^T + residuals \tag{2.18}$$

where $\overline{E}_{control}$ and $\overline{E}_{case}$ are defined as in the measure of magnitude. X is a $Nx2$ matrix which plays the role of a switch between the two conditions(control and the case Samples.)



1	0
1	0
1	0
1	0
0	1
0	1
0	1
0	1

Figure 2.6: (a)Best linear fit in the differential expression analysis,(b)The matrix X

Usually differentially expressed genes are the ones in the genome that have:

$$log_2 FoldChange \geq 1.5 \tag{2.19}$$
$$p_{value} \leq 0.05 \tag{2.20}$$

However these values are no solid and it is up to the researcher to decide.

2.3 Gene Set Analysis

The analysis of genome-wide expression data involves the task of compiling a list of statistical significance of all genes over multiple conditions, enabling the identification of differentially expressed genes. Many approaches have been developed for that matter. However, arriving at a list of significant genes does not alone necessarily facilitate the biological problem. To overcome this, methods based on statistical hypothesis tests have been developed that shift the analysis from individual genes to sets of genes.[23] Gene Set Analysis (GSA) has the advantage of incorporating existing biological knowledge into the expression analysis. As an example of a common approach, Gene Ontology (GO) terms can be used to define gene sets, thus enabling the identification of, e.g. statistically significant biological processes, through the use of GSA. Gene sets are not restricted to GO terms, as they can be defined in an unlimited number of ways, correlating to anything from metabolic or signalling pathways to transcription factors and chromosomal positions. There are currently two major types of GSA procedure for incorporating biological knowledge into Differential Expression Analysis. The first type can be referred to as the over-representation approaches and the second type as the aggregate score approaches.[24]

Over-representation analysis can be summarized as follows : First, form a list of candidate genes, that is genes which are considered differentially expressed (Differential Expression Analysis). Then, for each gene set, we create a two−by−two table comparing the number of candidate genes that are members of the category to those that are not members. The significance of over-representation can be assessed, for example, using the hypergeometric distribution or its binomial approximation. A limitation of the over-representation approach is that it ignores all the genes that did not make the list of candidate genes. Therefore, the results will be highly dependent on the cut-off used in constructing this list.

The aggregate score approach, does not have this limitation. The methods that depend on the aggregate score approach takes a list of gene-level statistics as an input and, based on these statistics, calculate a gene set statistic for each gene set being analysed. In the list, all genes of the original expression dataset can be included, thus these methods do not require any a priori significance cut-off. Due to this advantage they have gained scientific focus over the past few years.

The GSA work-flow as described by Leif Varemo et al [23] starts with calculating the gene set statistics with various methods, followed by the significance estimation. Finally the significance of its gene set is estimated by implementing the consensus scoring approach. This procedure can be seen in figure 2.7.

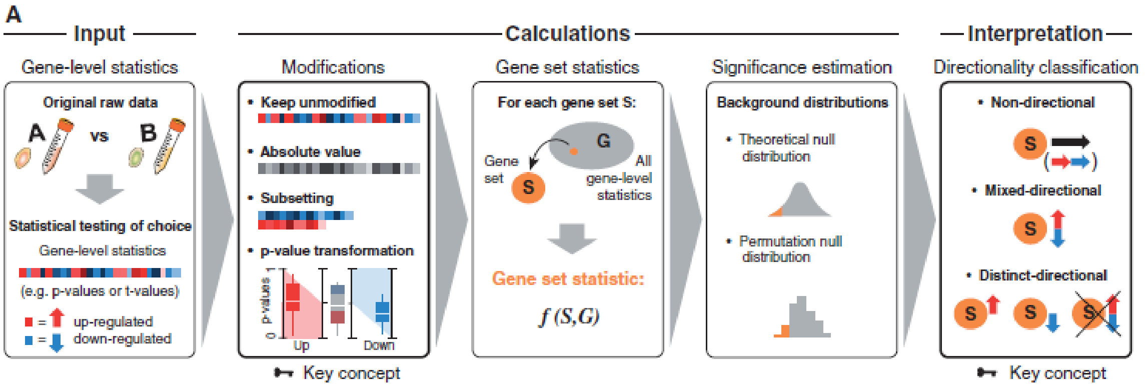


Figure 2.7: Gene Set Analysis Workflow [23]

2.3.1 Gene Set Statistics

There are many ways to calculate the gene set statistics. Some of them, take as input a list of the genes p-values,whereas some of them take as input a list of the genes t-values. In other words, some conclude in the calculation of the gene set statistics the sign of the genes regulation (t−values), whereas other do not (p−values). Methods that frequently are used for the calculation of the gene set statistics are shown in the table **??**

Methods used for the calculation of Gene Set Statistic
Fisher's combined probability test
PAGE
Stouffer's method
Reporter features
Tail Strength
Wilcoxon rank-sum test
Mean
Median
Sum
Gene Set Enrichment Analysis (GSEA)

Of these methods GSEA is by far the most popular and most frequently used in the GSA. [25] For that reason, will be analysed further.

Gene Set Enrichment Analysis(GSEA)

GSEA takes as input a list of genes and their measure of significance, specifically their t-values. The genes can be ordered in a ranked list L, according to their differential expression between the classes. Given an a priori defined set of genes S, the goal of GSEA is to determine whether the members of S are randomly distributed throughout L or primarily found at the top or bottom. For that reason, an enrichment score (ES) is calculated that reflects the degree to which a set S is overrepresented at the extremes (top or bottom) of the entire ranked list L . The score is calculated by walking down the list L, increasing a running-sum statistic when a gene in S is encountered S and decreasing it when a gene not in S is encountered. In particular, if "i" is denoted to a gene of the list "L" the enrichment score (ES) is calculated as follows:

$$ES(S) = max(P_{hit} - Pmiss) \tag{2.21}$$

$$P_{hit}(S,i) = \sum_{ieS, j \leq i} \frac{| r_j |^p}{N_R} \quad with \tag{2.22}$$

$$N_R = \sum_{ieS} | r_j{}^p \quad and \tag{2.23}$$

$$P_{miss}(S,i) = \sum_{i \notin S, j \leq i} \frac{1}{N - N_H} \quad with \tag{2.24}$$

$$N_H = \sum_{i \notin S} | r_j |^p \quad and \tag{2.25}$$

$$N = N_H + N_R \tag{2.26}$$

2.3.2 Assessing Gene Set Significance

Each gene set statistic can be converted into a p-value that estimates the statistical significance of that gene set. By definition, the p-value of a gene set statistic is the probability to observe a new gene set statistic that is equal to or more extreme than the given gene set statistic.

This probability can be estimated provided a null distribution, i.e. the probability distribution of the gene set statistic. For 5 of the 11 gene set statistics, theoretical null distributions, defined by continuous functions, can be used to estimate the p-values. In all 11 cases, the null distributions can also be estimated by a permutation approach. This approach can be performed in two ways, either by randomizing the genes, referred to as gene sampling, or by randomizing the sample labels, referred to as sample permutation. Gene sampling is carried out for each gene set by randomly taking a sample of genes (of the same number as in the gene set) and recalculating the gene set statistic. This is repeated a large number of times (e.g. 10 000 times) to give a discrete null distribution. The gene set p-value is simply the fraction of random gene set statistics that are equal to or more extreme (in general larger) than the original gene set statistic. Sample permutation is similar to gene sampling; however, in this case, the original sample labels are randomized, and all the gene-level statistics, and subsequently all the gene set statistics, are recalculated based on the new labelling. This procedure is also repeated a large number of times, and the p-values are calculated in the same way as described for the gene sampling. The choice of permutation approach is tightly connected to the underlying null hypothesis. When using gene sampling, the association of a gene set with the phenotype is compared with the association of the rest of the genes to the phenotype. On the other hand, by using sample permutation, the association of a gene set to the phenotype is compared with its association to random phenotypes.

As a final step of the analysis, the gene set's p-values are adjusted for multiple testing with one of the known methods. The result of a GSA is a list of gene set adjusted p-values, indicating the significance of each gene set. Usually, as a cut-off threshold for identifying the most significant gene sets is:

$$adj.p_{value} \leq 0.05 \tag{2.27}$$

2.3.3 Consensus Scoring of gene sets

A further step of the analysis is to assign the resulting gene set p-values to the appropriate directionality class which depends on the statistical test.

Three directionality classes can be defined: the non-directionality, the mixed-directionality and the distinct directionality. The non-directional class contains gene set p-values where the information about direction of differential expression is omitted, so that significant gene sets can be interpreted as affected by differential expression in general. For the mixed-directional class, a gene set can be significantly affected by differentially expressed genes in either or both directions. Finally, the distinct-directional class aims to identify gene sets that are significantly affected by regulation in a distinct direction.

By using various combinations of gene-level statistics, gene set statistics, significance estimation methods and directionality classes, different unique GSA runs can be performed. Each run will produce a list of gene set p-values for some or all of the directionality classes. To achieve a consensus result, the different gene set p-value vectors belonging to the same class are aggregated to produce a consensus score for each gene set and class. The aggregation is based on ranking the gene sets according to their P-value and using rank aggregation approaches to yield a consensus score for each gene set. Two simple approaches are to use either the mean or the median of the ranks of a given gene set as the consensus score.

As the p-values of the gene sets depends highly on the method used for calculating the gene set statistic and for defining the null distribution, it is more robust to use a consensus score for the identification of the most important gene sets. For example, one can choose the gene sets that were ranked in the top five of one of the categories as the most important.

2.4 Weighted Gene Co-Expression Network Analysis

Networks can be used to describe the pairwise relationships between n nodes (which are sometimes referred to as vertices). For example, networks can be used to describe the relationships between n genes. A correlation network is a network whose adjacency matrix is constructed on the basis of pairwise correlations between numeric vectors. The numeric vectors may represent observed quantitative measurements of variables. For example, the gene expression levels (transcript abundances) across different samples can be represented by a numeric vector. The aim of Weighted Gene Co-Expression Network Analysis (WGCNA) is to identify clusters of genes highly co-expressed, as well as, the "key" genes in these clusters.[26],[27]

2.4.1 Network Construction

Methods for defining an adjacency matrix are also known as network construction. The network adjacency matrix can be defined by transforming a similarity or dissimilarity matrix, a symmetric matrix, or even a general square matrix. Multiple similarity matrices can be combined into a single consensus network, which allows one to define consensus modules. For the construction of the WGCNA network the dissimilarity topological overlap matrix is used.

Pearson's Correlation

The Pearson correlation (also known as sample correlation) between two vectors x and y is defined as follows:

$$cor(x,y) = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} \tag{2.28}$$

$$cov(x,y) = \frac{\sum_u^N (x_u - mean(x))(y_u - mean(y))}{N-1} \tag{2.29}$$

The first step for constructing the WGCNA network is calculating the pearson correlation matrix. The pearson correlation matrix is a square symmetric matrix with number of rows and columns equivalent to the number of genes. If x_i and x_j $(i, j = 1 \dots N)$ are the vectors containing the gene expression levels across the different samples of genes x and y respectively, the matrix's elements can be described:

$$DatX_{ij} = cor(x_i,x_j) \tag{2.30}$$

Adjacency matrix

A correlation network adjacency matrix is constructed on the basis of the pairwise correlations $cor(x_i,x_j)$. The adjacency matrix can either be weighted or unweighted. The unweighted adjacency matrix results in a network where only highly correlated nodes are connected, and therefore a sparse network. The weighted adjacency matrix results to a network where all nodes are connected and the correlation between the nodes is shown by the thickness of the line.

A weighted adjacency matrix can be either signed or unsigned. The signed weighted network takes into consideration the sign of the correlation whereas the unsigned treats the negative and the positive correlation the same. An unsigned weighted adjacency matrix can be defined as follows:

$$A_{ij} =\mid DatX_{ij} \mid^{\beta} \tag{2.31}$$

A signed weighted adjacency matrix can be defined as follows:

$$A_{ij} = (0.5 + 0.5DatX_{ij})^{\beta} \tag{2.32}$$

where in both cases the power parameter is required to satisfy $\beta \geq 1$.

An unweighted correlation network can be defined by thresholding the absolute values of the correlation matrix, i.e.,

$$A_{ij} = \begin{cases} 1, & if \mid DatX_{ij} \mid \geq \phi \\ 0, & if \mid DatX_{ij} \mid \leq \phi \end{cases} \tag{2.33}$$

A basic disadvantage of the unweighted network is that whether two nodes are connected or not, depends exclusively on the cut-off value ϕ of the step function. For example, two nodes whose correlation is $\phi + \Delta\phi$ are considered correlated and are connected in the network, whereas two nodes whose correlation is $\phi - \Delta\phi$ are considered not correlated and are not connected in the network. Therefore, the resulting network depends highly on the cut-off value ϕ. The weighted network tries to eliminate this problem by smoothing the step function. This is achieved through the introduction of the parameter β. Higher the β, closer the power adjacency function to the step function, as shown in figure 2.8.

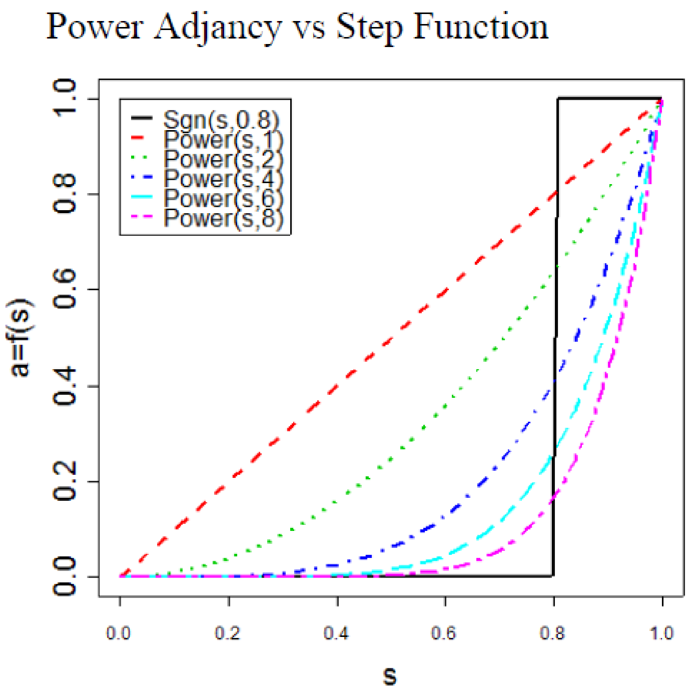


Figure 2.8: Step function vs Adjacency power function for various powers β

As shown in figure 2.8 a weighted correlation network constructed with the power adjacency function depends on the parameter β. Since each parameter value of an adjacency function leads to a network with different topological properties, the choice of the parameter value can be guided by desirable topological properties.

In WGCNA the parameter β of the constructed weighted correlation network is chosen in order for the network to exhibit approximate scale-free topology. Scale-free topology networks consists of a few genes with great connectivity and usually larger heterogeneity and cluster separability. Furthermore, many biological networks, like the metabolic, have been found to exhibit approximate scale-free topology properties. In order to find if the choice of a parameter β has lead to an approximate scale-free topology network, the general connectivity of each gene is calculated:

$$GCon_i = \sum a_{ij} \tag{2.34}$$

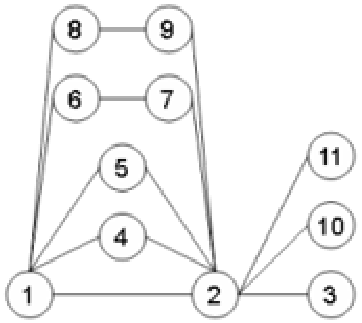
as well as the frequency of each connectivity $GCon_i$. Then both the connectivity and the frequency are log10 transformed. In order for the network to have approximate scale-free topology the plot of the connectivity$\sim$ frequency should be linear with a scale-free index R^2 close to 1. In practice, an $R^2 \geq$ 0.8 is enough to consider that the network exhibits approximate scale-free topology properties.

Dissimilarity overlap matrix

In practice, an original network adjacency matrix $A^{original}$ is often transformed into new network adjacency matrix denoted by A. For example, a transformation can be used to change the topological properties of a network. An adjacency function AF is defined as a matrix valued function that maps an n×n dimensional adjacency matrix $A^{original}$ onto a new n×n dimensional network adjacency. The topological Overlap Matrix(TOM)-based adjacency function replaces the adjacencies of the original adjacency matrix $A^{original}$ by a measure of interconnected adjacencies that are based on shared neighbours.

$$TOM_{ij} = \frac{(\sum_u (a_{iu}a_{uj})) + a_{ij}}{min(GCon_i, GCon_j) + 1 - a_{ij}} \tag{2.35}$$

Basically, the TOM_{ij} is a measure of the common neighbours that the genes x_i and x_j share. The topological overlap measure can serve as a filter that decreases the effect of spurious or weak connections, and it can lead to more robust networks. Since the $A^{original}$ adjacency matrix is sparse (many zeroes) and susceptible to noise, in most cases it is advantageous to use TOM_{ij} for the construction of the network.



$$TOM(i,j) = \frac{|\,N_1(i) \cap N_1(j)\,| + a_{ij}}{\min(|\,N_1(i)\,|, |\,N_1(j)\,|) + 1 - a_{ij}}$$

Figure 2.9: TOM is a measure of the common neighbours that two nodes share

TOM_{ij} matrix,similarly with the adjacency matrix A_{ij} are based on the similarities between the genes. However, the WGCNA uses dissimilarity measures in order to find the clusters of the co-expressed genes. Consequently, the dissimilarity topological overlap matrix is calculated, as a final step and the basis for the construction of the network in WGCNA.

$$DISTOM_{i}j = 1 - TOM_{i}j \tag{2.36}$$

2.4.2 Module Identification

Detecting clusters (also referred to as groups or modules) of closely related objects is an important problem in data mining in general. Network modules are often defined as clusters. Partitioning-around-medoids (PAM) clustering and hierarchical clustering are often used in network applications. Partitioning-around medoids (aka. k-medoid clustering) leads to relatively robust clusters but requires that the user specifies the number k of clusters. Hierarchical clustering is attractive in network applications since (a) it does not require the specification of the number of clusters and (b) it works well when there are many singleton clusters and when cluster sizes vary greatly. However, hierarchical clustering requires the user to determine how to cut branches of the resulting cluster tree. Toward this end, in WGCNA one can use the dynamic hybrid method which combines the advantages of both hierarchical clustering and partitioning-around-medoids clustering.[28]

Partitioning-around-medoids

Partitioning-around-medoids (PAM) is a clustering procedure that implements an iterative algorithm for minimizing the within-cluster scatter. Assume that $Cl(i)$ encodes a cluster assignment, i.e., $Cl(i) = q$ if the i^{th} object is in the q^{th} cluster (where q = 1, . . . ,k indexes the k clusters). Then the within cluster scatter with regard to the dissimilarity matrix D_{ij}, $WithinScatter(Cl, D_{ij})$ is defined as follows:

$$WithinScatter(Cl, D_{i}j) = \frac{1}{2} \sum_{q=1}^{k} \sum_{Cl(i)=q} \sum_{Cl(j)=q} d_{ij} \tag{2.37}$$

A medoid can be defined as that object of a cluster, whose average dissimilarity to all the objects in the cluster is minimal, i.e., it is the most centrally located object inside a given cluster.

PAM is a classical partitioning technique of clustering that clusters the data set of n objects into k clusters where the integer k is a user-supplied parameter and is generally applicable since it can input any dissimilarity measure.

Hierarchical Clustering

Hierarchical algorithms find successive clusters based on previously defined clusters. These algorithms can be either bottom-up (agglomerative) or top-down (divisive).

Agglomerative hierarchical clustering treats objects as separate clusters and merges them into successively larger clusters. Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure called a dendrogram or cluster tree. The root of the tree consists of a single cluster containing all objects, and the leaves correspond to individual objects. Agglomerative hierarchical clustering has two inputs: (a) a pairwise dissimilarity measure (eg. $DissTOM$ matrix)and (b) a method for constructing an inter-cluster dissimilarity measure. The inter-cluster dissimilarity measure(also known as linkage method or agglomeration method) is based on the pairwise dissimilarities between objects inside the clusters.

In WGCNA, the inter-cluster dissimilarity between two clusters $clust_{q1}$ and $clust_{q2}$ is defined as the average dissimilarity between the objects of each cluster.(average linkage hierarchical clustering)

$$d_{average}(clust_{q1}, clust_{q2}) = \frac{\sum_{ieclust_{q1}} \sum_{jeclust_{q2}} d_{ij}}{\mid clust_{q1} \mid\mid clust_{q2} \mid} \tag{2.38}$$

where d_{ij} denotes the pairwise dissimilarity between objects i and j, $\mid clust_{q1} \mid$ and $\mid clust_{q2} \mid$ denotes the number of objects in $clust_{q1}$ and $clust_{q2}$, accordingly.

For the identification of the clusters of genes, one can choose a threshold dissimilarity height to cut the dendrogram. Then clusters are defined as the separate branches below the cut height. This algorithm is also known as Static cut Algorithm.

Another way of identifying the clusters is by implementing an algorithm that respects the morphology of the dendrogram. This algorithm is also known as Dynamic Cut Algorithm. The algorithm implements an adaptive, iterative process of cluster decomposition and combination and stops when the number of clusters becomes stable. It starts by obtaining a few large clusters by the static tree cut. The joining heights of each cluster are analyzed for a characteristic pattern of fluctuations indicating a sub-cluster structure; clusters exhibiting this pattern are recursively split. To avoid over-splitting, very small clusters are joined to their neighbouring major clusters.

Hybrid Method

The Hybrid cut tree method is a bottom-up algorithm that uses both a cluster tree and a dissimilarity measure as an input. As a hybrid between hierarchical clustering and partitioning-around-medoids (PAM) clustering it may improve the detection of outlying members of each cluster. The hybrid cluster detection proceeds along two steps. First, branches that satisfy specific criteria for being clusters are detected. Next, all previously unassigned objects are tested for sufficient proximity to clusters detected in the first step; if the nearest cluster is close enough, the object is assigned to that cluster.

2.4.3 Hub-genes

Co-expression modules identified by clustering are often large, and so, it is important to identify which genes in each module best explains its behaviour. A widely used approach is to identify highly connected genes in a co-expression network. These genes are called hub genes. Hub genes are frequently more relevant to the functionality of the networks than other nodes. This is also the case in biological networks, although mathematical derivations show that this is only the case for intra-modular hub genes (as opposed to inter-modular hub genes). Intra-modular hubs are central to specific modules in the network, while intermodular hubs are central to the entire network. To identify hub genes, centrality measures, mainly ”betweenness centrality”, are often used. Genes with high betweenness centrality are important as shortest-path connectors through a network. Betweenness centrality can be measured as shown in equation 2.39

$$BetweenessCentrality(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{2.39}$$

where σ_{st} is the total number of shortest paths from the node s to the node t and $\sigma_{st}(v)$ is the total number of those that pass through the node v.

2.4.4 Eigengenes

Identified co-expression modules usually form a biologically meaningful meta-network that reveals a higher-order organisation of the transcriptome. The analysis for constructing the meta-network can be viewed as a network reduction scheme that reduces a gene co-expression network involving thousands of genes to orders of magnitude smaller meta-network involving module representatives (one eigengene per module).

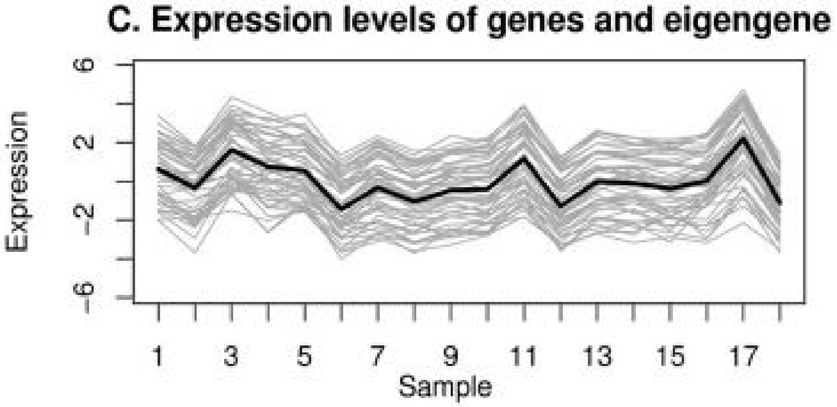


Figure 2.10: Expression levels (y-axis) of module genes (grey lines) and the eigengene (black line) across microarray samples (x-axis).

A module eigengene is defined as the first principal component of the expression matrix of the corresponding module.[29] It is always more robust than the modules. For that reason, it is advantageous to use the eigengenes of the modules for defining the correlation of the modules with the disease. Furthermore, a module's genes that have a high correlation with its eigenegene are usually the ones that are the hub-genes.

Eigengenes of different modules often exhibit high correlations. If the eigengenes are indexed by capital letters, for example E_J denotes the eigengene of the J^{th} module, then the connection strength (adjacency) between two eigenegenes E_J and E_I can be defined as:

$$a_{eigen_{IJ}} = \frac{1 + cor(E_I, E_J)}{2} \tag{2.40}$$

These connection strength can be used to define the meta-network, also known as eigengene network. Since eigengenes form a network, same module detection procedures like the ones mentioned before can be used to identify modules comprised of eigengenes, the meta-modules. In WGCNA, meta-modules are detected as branches of the resulting cluster tree by using average hierarchical clustering. The resulting meta-modules are sets of positively correlated eigengenes. Meta-modules may reveal a higher order organization among gene co-expression modules. Furthermore, meta-modules,as well as the eigengenes, are highly robust to noise.

Both the eigenegenes and the meta-modules can be used for further meta-analysis like differential eigengene network analysis or meta-module preservation analysis in different datasets(different tissues).

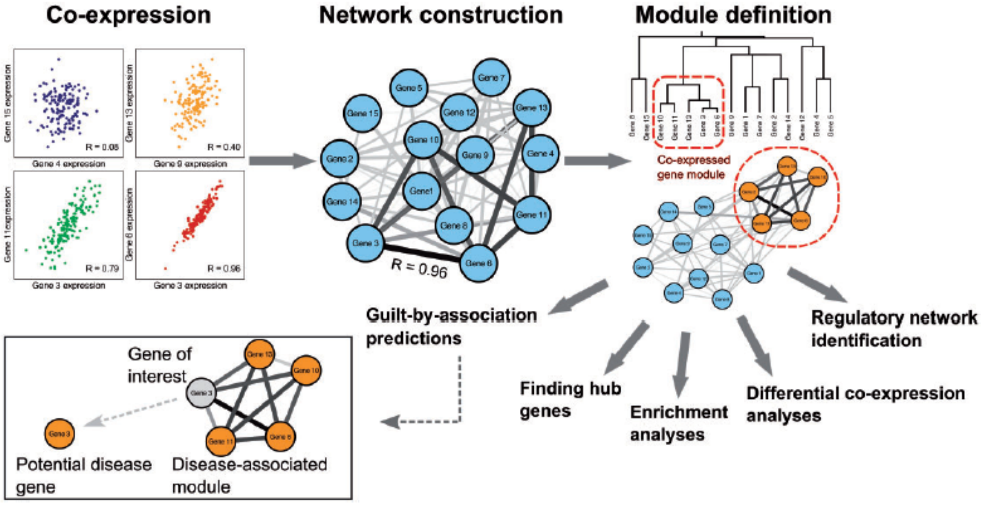


Figure 2.11: Workflow of the Weighted gene Co-Expression Analysis(WGCNA)

2.5 Module Stability Analysis

Many module detection methods, like WGCNA, identify groups of genes whose expression profiles are highly correlated. For such modules to be biological meaningful, their stability is of high importance. However, the modules identified by cluster analysis can be strongly affected by noise and outlying observations. Further, many clustering methods can be considered non-robust in the sense that a small change in the underlying network adjacency can lead to a large change in the resulting clustering (for example, previously separate clusters may merge or a cluster may be split).

There are many ways that the robustness of the clustering method and the stability of the identified modules can be examined. For example, one may perform the network analysis and module identification repeatedly on artificial data sets derived from the original data or add varying amounts of random noise to the data.

Let's assume the construction of N artificial datasets. The network analysis and module identification of each artificial dataset will lead to N new sets of modules. Hypothetically, if $clust_i$ is denoted to a module of the original network , then in order for that module $clust_i$ to be considered stable it should be included to every new set of modules. In reality, this is rarely the case. For that reason, the $clust_i$ is considered stable if there is a module $clust_j$ in each new set of modules that contain a great percentage of its genes. Larger the average percentage, more stable the original $clust_i$. If the above-mentioned hypothesis is true for either all or the great majority of the modules of the original dataset, then the clustering method can be considered robust, and the constructed original network with its clusters biologically meaningful.

There are two major methods to construct artificial expression datasets. The first method of defining an artificial expression dataset is by removing randomly a small percentage of the Samples, usually less than 10%. The second method is by re-sampling the original dataset, allowing for the multiple repetition of a sample. In other words,the new dataset will consist of randomly chosen samples of the original dataset, and one sample can be chosen multiple times.

Implementing the first method one can check if the clustering of the genes will be affected by the loss of information. The second method concludes to almost entirely different datasets and it is a more demanding stability test. However, in small datasets (small number of samples) there is a possibility that the new artificial dataset consists of a very small proportion of the original samples. This is a consequence of allowing the repetition of samples multiple times in the new dataset and lead to not informative results.

CHAPTER 3

Results

This chapter provides an overview of the diploma thesis. The application of the methods analysed in the theory chapter will be presented and the results will be discussed. Each section corresponds to one part of the analysis done in this diploma thesis and depends on the results of the previously presented sections.

3.1 Datasets

The purpose of this diploma thesis, as described in the Introduction, is to detect common pathological mechanisms among the tissues involved in knee-osteoarthritis: cartilage, synovium,meniscus and subchondral bone. In this respect, four microarray datasets where used, one microarray dataset for each tissue affected, as shown in table 3.1. All the microarray datasets derive from Gene Expression Omnibus (GEO). GEO is a database repository of high throughput gene expression data and microarrays.

Synovium

GSE55235 is a genome-wide transcriptomic dataset from 79 individuals. It includes 20 healthy controls, 26 osteoarthritis patients and 33 rheumatoid arthritis patients. The Affymetrix tech-

Microarray datasets used	GEO Accession
Synovium	GSE55235
Cartilage	GSE117999
Subchondral Bone	GSE51588
Meniscus	GSE98918

Table 3.1: The GEO Accession of the datasets used in the analysis

nology (specifically Affymetrix Human Genome U133A Array) was used to identify gene transcripts that were expressed differently in the synovial tissues of the joints of the samples.

For the purpose of this diploma thesis, the 33 rheumatoid arthritis patients were excluded from the analysis.

Cartilage

GSE117999 is a genome-wide transcriptomic dataset that includes 24 Samples. 12 Patients undergoing arthroscopic partial meniscectomy (APM) without any evidence of OA and 12 patients undergoing total knee arthroplasty (TKA) due to end-stage OA were consented. Cartilage was garnered from the non-weight bearing site of the medial intercondylar notch. The Agilent technology (specifically Agilent-072363 SurePrint G3 Human GE v3 8x60K Microarray) was used to probe differentially expressed transcripts between the healthy and the OA cartilage.

Subchondral Bone

GSE51588 is a genome-wide transcriptomic dataset. For its creation total RNA from regions of interest from human OA (n=20) and non-OA (n=5) knee lateral and medial tibial plateaus (LT and MT) were isolated. The Agilent technology (specifically Agilent-026652 Whole Human Genome Microarray 4x44K v2) was used to probe differentially expressed transcripts between the healthy and the OA subchondral bone regions.

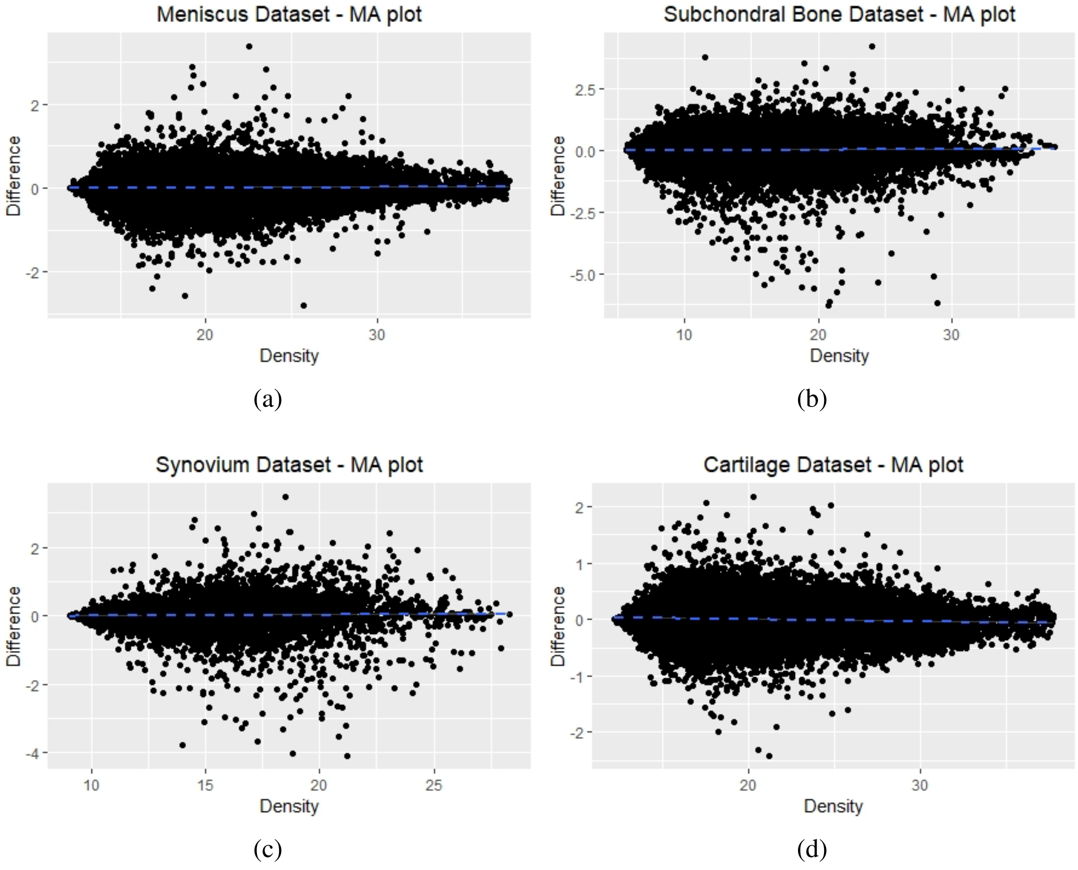
For the purpose of this diploma thesis, only the 20 OA samples and the 5 normal samples corresponding to knee medial tibial plateau were used. This decision stemed from 2 reasons. Firstly, an important difference in the gene expression levels was observed between the knee lateral and medial plateaus of the same samples. Subsequently, the use of both LT and MT samples would conclude in loss of biological information. Secondly, comparing the gene expression levels between LT and MT it was obvious that the knee medial tibial palteaus was more affected by osteoarthritis than the LT.

Meniscus

GSE98918 is a genome-wide transcriptomic dataset from 24 individuals. It includes 12 healthy controls and 12 osteoarthritis patients. The Agilent technology (Agilent-072363 SurePrint G3 Human GE v3 8x60K Microarray) was used to identify gene transcripts that were expressed differently in meniscus tissues obtained from the OA and non-OA samples.

3.2 Background Correction and Normalisation

For background correction and normalisation, RMA and quantile normalisation were chosen in all datasets. The combination of these methods was chosen because it resulted in MA plots scattered around the zero line, which could not be achieved with any other method. This can be seen in figure 3.1 where the MA plot of each dataset is presented. The package " limma " in the programming language R was used for the application of the methods.[14]



(a)

(b)

(c)

(d)

Figure 3.1: MA plots of the four datasets after background correction and normalisation

3.3 Differential Expression Analysis

Before finding the differentially expressed genes in each tissue, the samples of each dataset were divided into two groups,normal and diseased. Then, both normal and diseased samples of each dataset were hierarchically clustered according to their genes' expression, as shown in figures 3.2 : 3.5. From the resulting dendrograms, possible outliers in the microarray datasets were detected and removed.

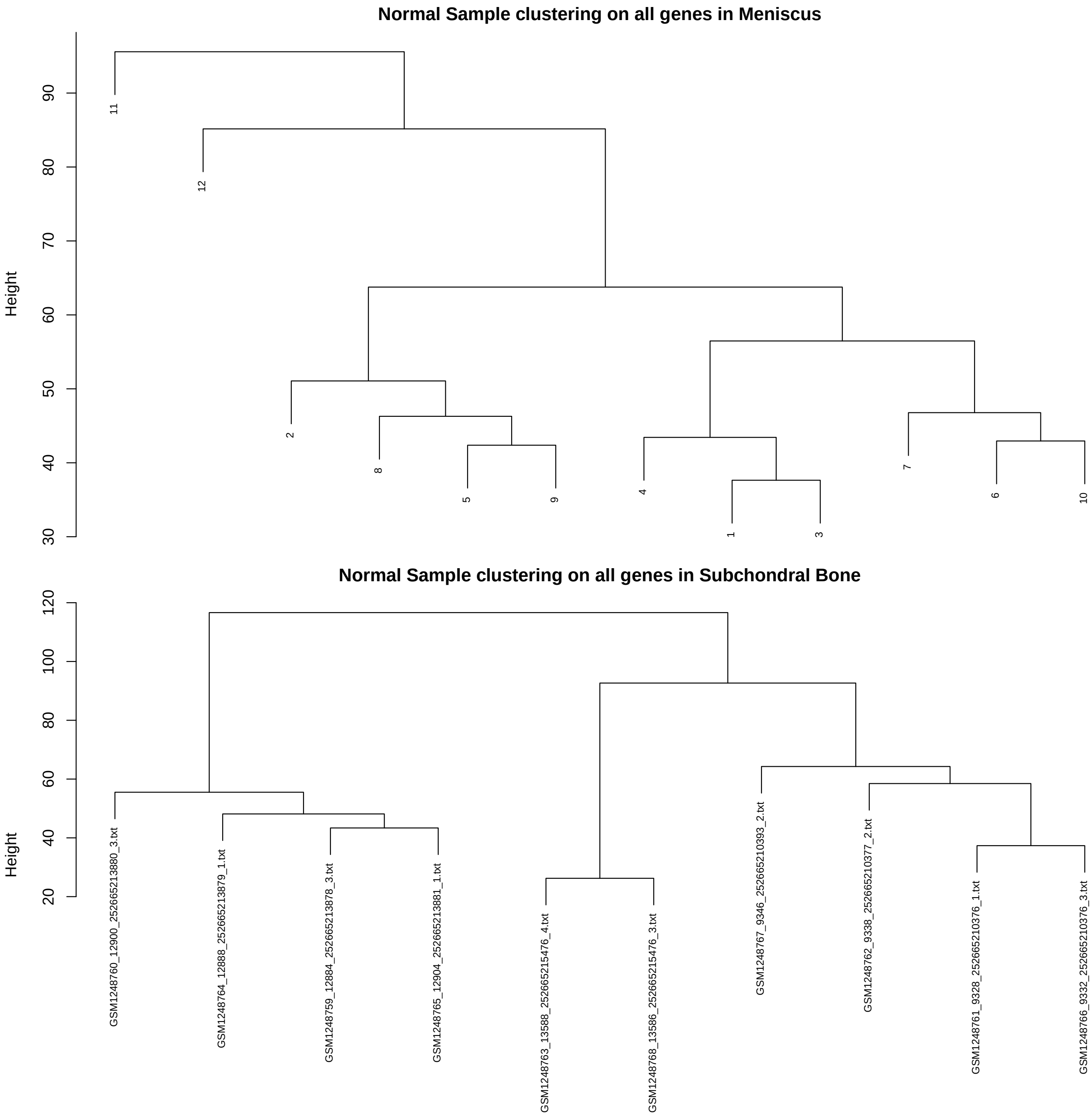
Specifically, as shown in figure 3.2 there were two outliers in the normal samples of the meniscus dataset, the 11_{th} and the 12_{th} control sample, whereas no outliers could be detected among the normal samples of the subchondral bone dataset. Furthermore, it is clear from figure 3.3 that no outliers could be detected in the normal samples of the synovium dataset. On the other hand, the 11_{th} normal sample in the cartilage dataset was greatly different from the rest of the samples and therefore it was considered an outlier and removed. Regarding the diseased samples, as can be seen in figures 3.4 and 3.5, the microarray datasets corresponding to meniscus, subchondral bone and synovium had no outliers. On the other hand, in the cartilage dataset the 15_{th} and 19_{th} diseased samples were outliers and therefore, they were removed.

Once the outliers in each dataset were removed, the differentially expressed genes (DEGs) in each dataset were detected. For the detection of the DEGs the package ”limma” in the programming language R was used.[14] In order for a gene to be considered differentially expressed, the two conditions described in equation 3.1 should be true:

$$log_2 FoldChange \geq 1.5 \tag{3.1}$$

$$adj.p_{value} \leq 0.05 \tag{3.2}$$

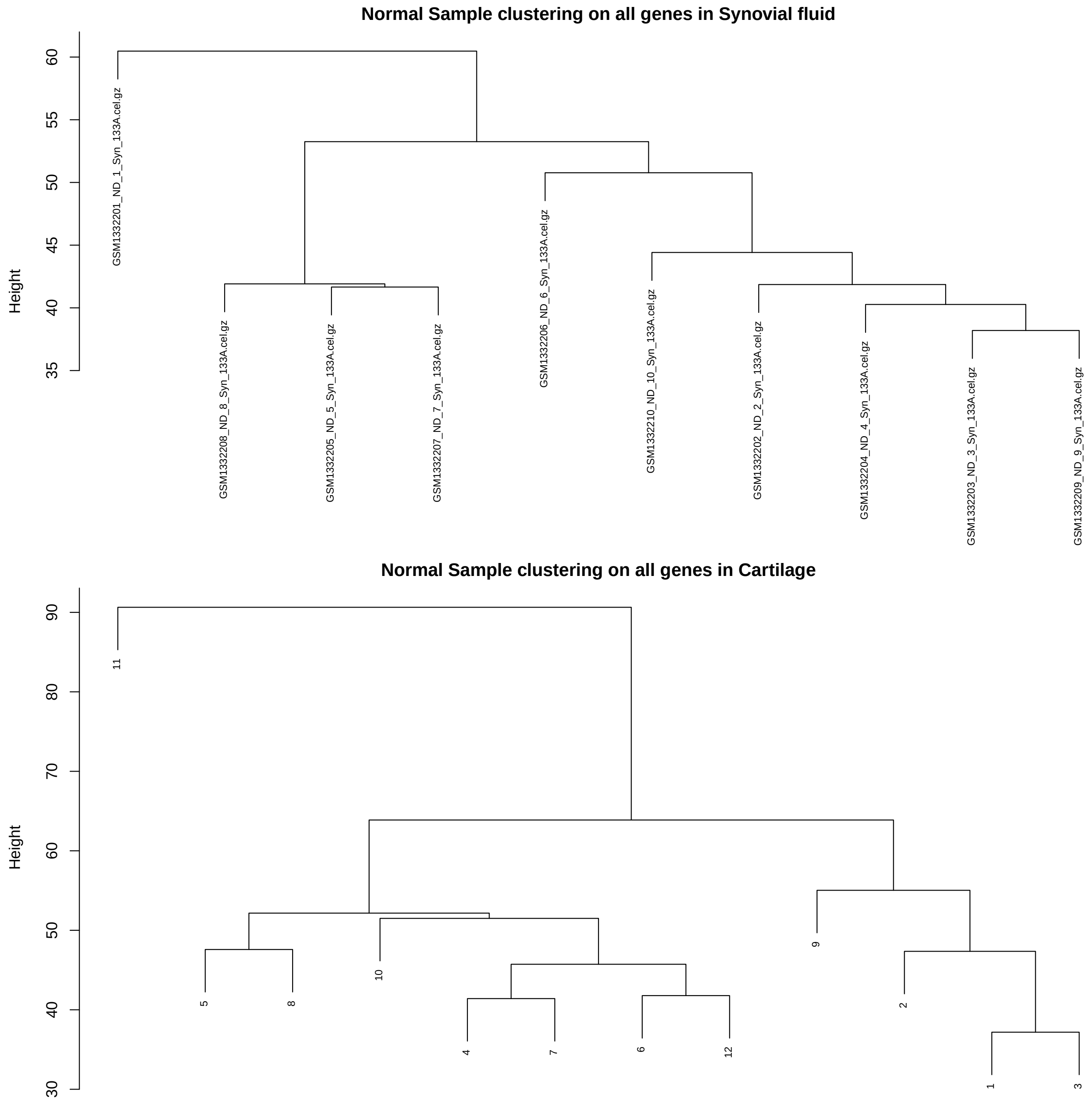
From figure 3.6 to figure 3.9 the volcano plots of the four datasets are presented. The points colored with orange correspond to the DEGs of each dataset. Red color was used to emphasise the genes that had $\mid log_2 FoldChange \mid \geq 2$. In other words, the points with red color correspond to genes whose expression between the normal and the OA samples had a 4-fold increase or decrease. Finally, genes that had $\mid log_2 FoldChange \mid \geq 2.5$ were named in the volcano plots.



Normal Sample clustering on all genes in Meniscus

Normal Sample clustering on all genes in Subchondral Bone

Figure 3.2: Hierarchical clustering of the Normal Samples in Mensicus and Subchondral Bone Dataset



Normal Sample clustering on all genes in Synovial fluid

Normal Sample clustering on all genes in Cartilage

Figure 3.3: Hierarchical clustering of the Normal Samples in Synovium and Cartilage Dataset

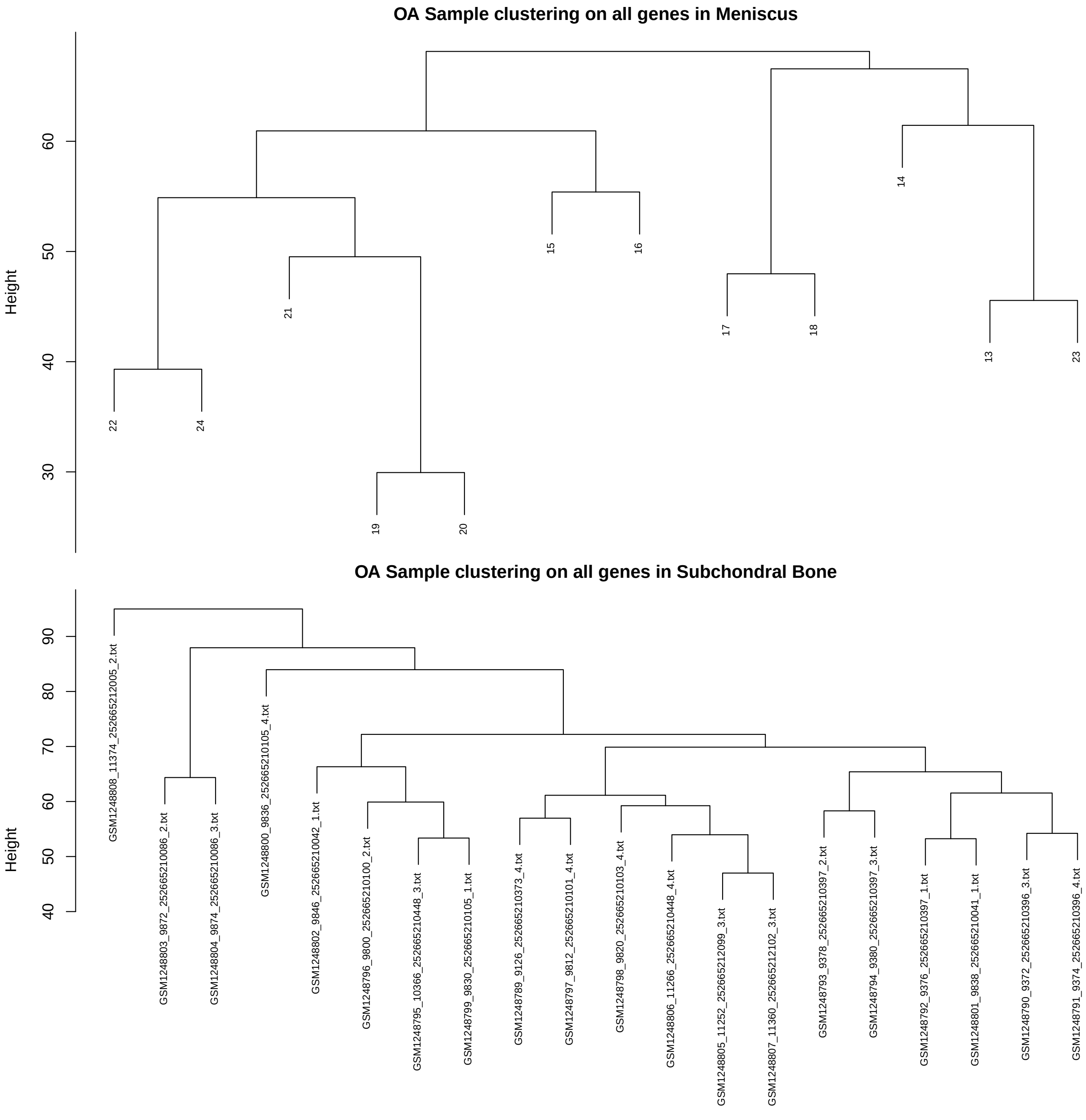


Figure 3.4: Hierarchical clustering of the OA Samples in Mensicus and Subchondral Bone Dataset

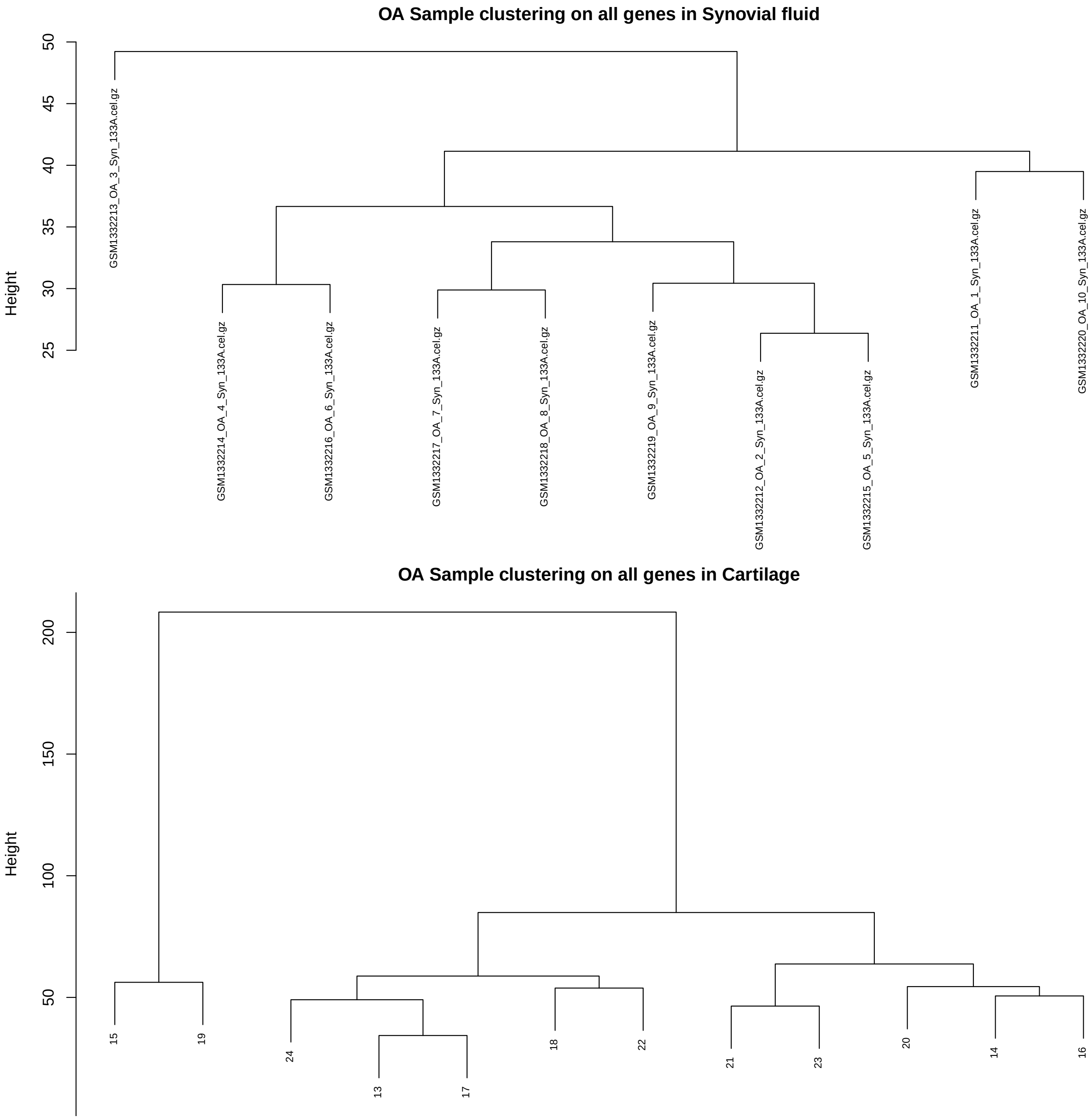


Figure 3.5: Hierarchical clustering of the OA Samples in Synovium and Cartilage Dataset

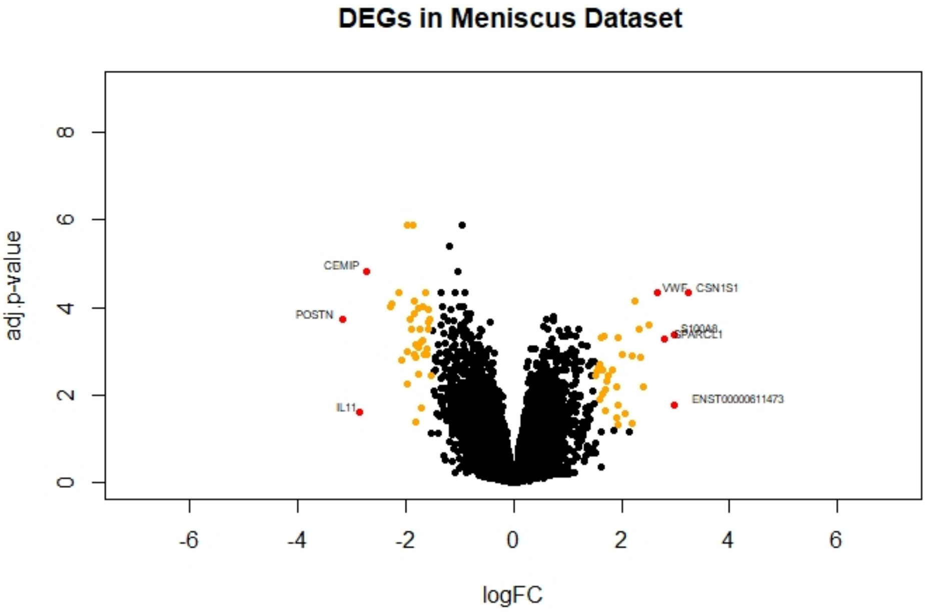


Figure 3.6: Volcano plot of the Meniscus Dataset

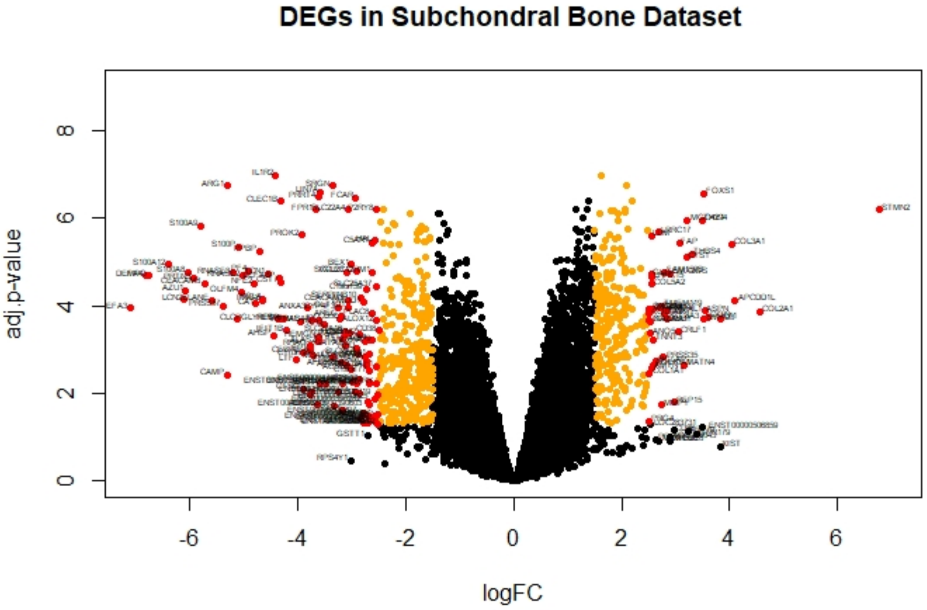


Figure 3.7: Volcano plot of the Subchondral Bone Dataset

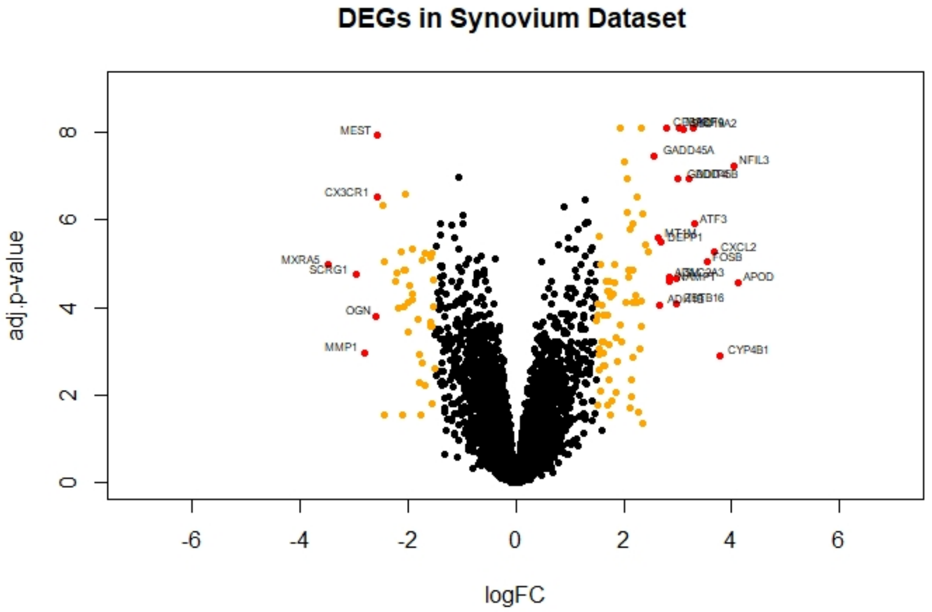


Figure 3.8: Volcano plot of the Synovium Dataset

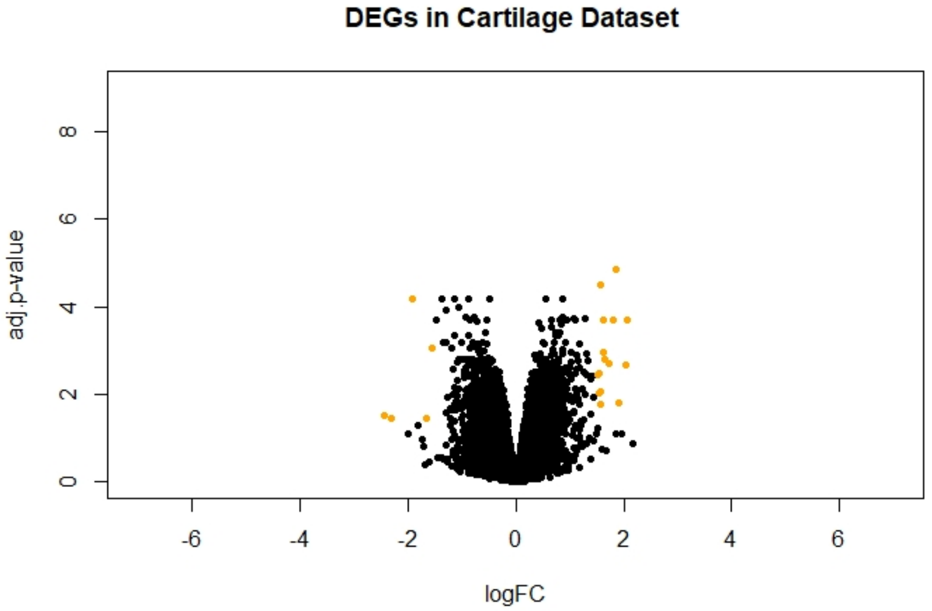


Figure 3.9: Volcano plot of the Cartilage Dataset

As depicted in the volcano plots, the Subchondral Bone dataset had the greatest number of DEGs. In particular, 712 genes were differentially expressed between the normal and the diseased samples, while 112 genes had $\mid log_2FoldChange \mid \geq 2$. It is worth mentioning that it was the only dataset having genes with $\mid log_2FoldChange \mid \geq 4$.

In the synovium dataset, 126 genes were differentially expressed between the normal and the diseased samples with 47 of them having $\mid log_2FoldChange \mid \geq 2$. Respectively, in the meniscus dataset, 69 genes were DEGs and just 14 genes had $\mid log_2FoldChange \mid \geq 2$.

Finally, the cartilage dataset had the least number of DEGs identified. Specifically, 32 genes were expressed differently between the normal and the OA samples with no DEG having $\mid log_2FoldChange \mid \geq 2$.

The Venn Diagram of the four datasets is shown in figure 3.10. Not one gene was identified as differentially expressed in all four datasets. The genes that were differentially expressed in three out of four datasets, as well as their functions are presented in table 3.2.

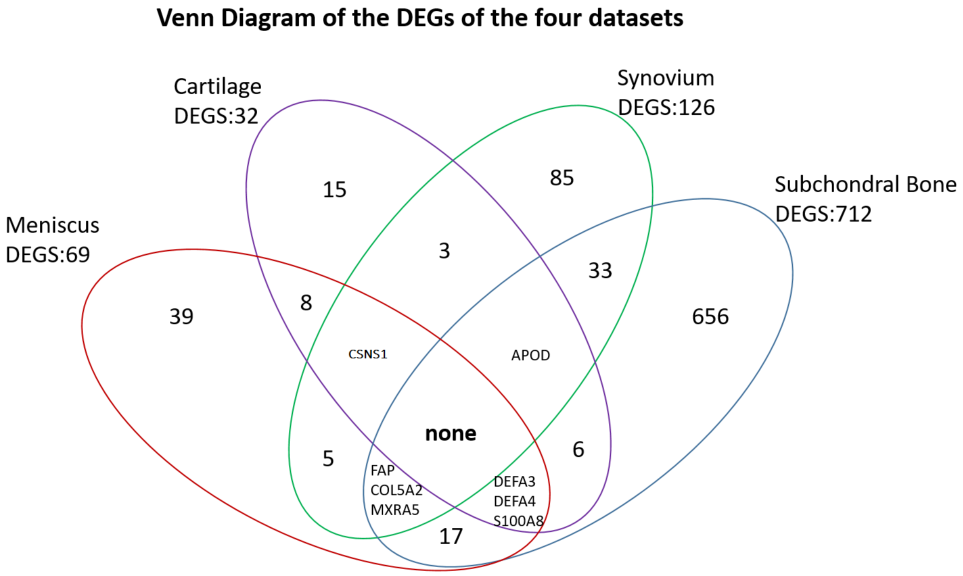


Figure 3.10: Venn Diagram of the four datasets regarding their DEGs

DEGs	***Function***
FAP	encodes protein(homodimeric integral membrane gelatinase) involved in the control of fibroblast growth or epithelial-mesenchymal interactions during development, tissue repair, and epithelial carcinogenesis
COL5A2	encodes an alpha chain for one of the low abundance fibrillar collagens, closely related to type XI collagen
MXRA5	encodes protein involved in extracellular matrix remodelling
APOD	encodes a component of high density lipoprotein, closely related with the enzyme lecithin involved in lipoprotein metabolism
DEFA3	encodes a protein (defensin, alpha 3) found in the microbicidal granules of neutrophils and likely plays a role in phagocyte-mediated host defense
S100A8	encodes a protein, involved in the regulation of a number of cellular processes such as cell cycle progression and differentiation. This protein may function in the inhibition of casein kinase and as a cytokine.
DEFA4	encodes protein (defensin, alpha 4) which is found in the neutrophils; it exhibits corticostatic activity and inhibits corticotropin stimulated corticosterone production
CSN1S1	encodes protein (casein alpha s1)

Table 3.2: The common differentially expressed genes in 3 out of 4 datasets

3.4 Weighted Gene Co-Expression Network Analysis

For the Weighted Gene Co-expression Analysis (WGCNA), only the diseased samples of the four microarray datasets were used. Consequently, in the analysis were included 12 OA Samples of the meniscus dataset, 12 OA samples of the synovium dataset, 25 OA samples of the subchondral bone dataset and 10 samples of the cartilage dataset. The construction of a weighted signed network was chosen, as such a network would provide more meaningful biological insights.

The aim of WGCNA is to group together genes that are highly co-expressed across the samples of a dataset, forming clusters. For the identification of the clusters, the co-expression network is constructed first, using the adjacency matrix as a basis. In this case, we had four datasets and we aimed to find modules that were common to all four datasets. In order this to be achieved, the co-expression network of each dataset was constructed. Then genes that were highly co-expressed consistently were grouped together. In other words, gene A and gene B would be clustered together if they were highly co-expressed in all four datasets.

Before constructing the networks and identifying the clusters, the power β was calculated in order the network to exhibit approximate scale-free topology properties. As can be seen in figure 3.11, in all datasets but synovium dataset, a power β equal to 12 would result in networks with scale free topology $R^2 = 0.8$. However, the network constructed from the synovium dataset if a power $\beta = 12$ was chosen, would have an $R^2 \sim 0.5$. This would affect the consensus modules identified by WGCNA and would result to modules highly driven from the synovium dataset, and therefore not important biologically.

For that reason, a power $\beta = 20$ was chosen, as with this power the network constructed from synovium dataset was closer to exhibit approximate scale free topology properties. The constructed consensus network can be seen in figure 3.12. The different colors correspond to different modules identified.

From the 11461 genes included in the analysis, 2850 were clustered in 37 modules. The number of genes included in each module can be seen in the table 3.3. In the grey module were assigned all the genes that were not considered members of any other module. In other words, the grey module contains all the genes that could not be clustered in co-expression modules. For the application of the WGCNA the package ”WGCNA” in the programming language R was used.[26],[28]

Module Color	Genes	Module Color	Genes	Module Color	Genes	Module Color	Genes
grey	8611	greenyellow	68	royalblue	53	steelblue	37
turquoise	585	purple	68	darked	51	paleturquoise	36
blue	278	tan	67	darkgreen	46	violet	36
brown	144	cyan	66	darkturquoise	45	darkmagneta	35
yellow	115	salmon	66	darkgrey	44	darkolivegreen	35
green	91	midnightblue	64	orange	44	sienna3	34
red	80	lightcyan	62	darkorange	42	yellowgreen	33
black	73	grey60	56	white	40	skyblue3	33
pink	70	lightgreen	55	skyblue	39		
magenta	69	lightyellow	55	saddlebrown	37		

Table 3.3: The common differentially expressed genes in 3 out of 4 datasets

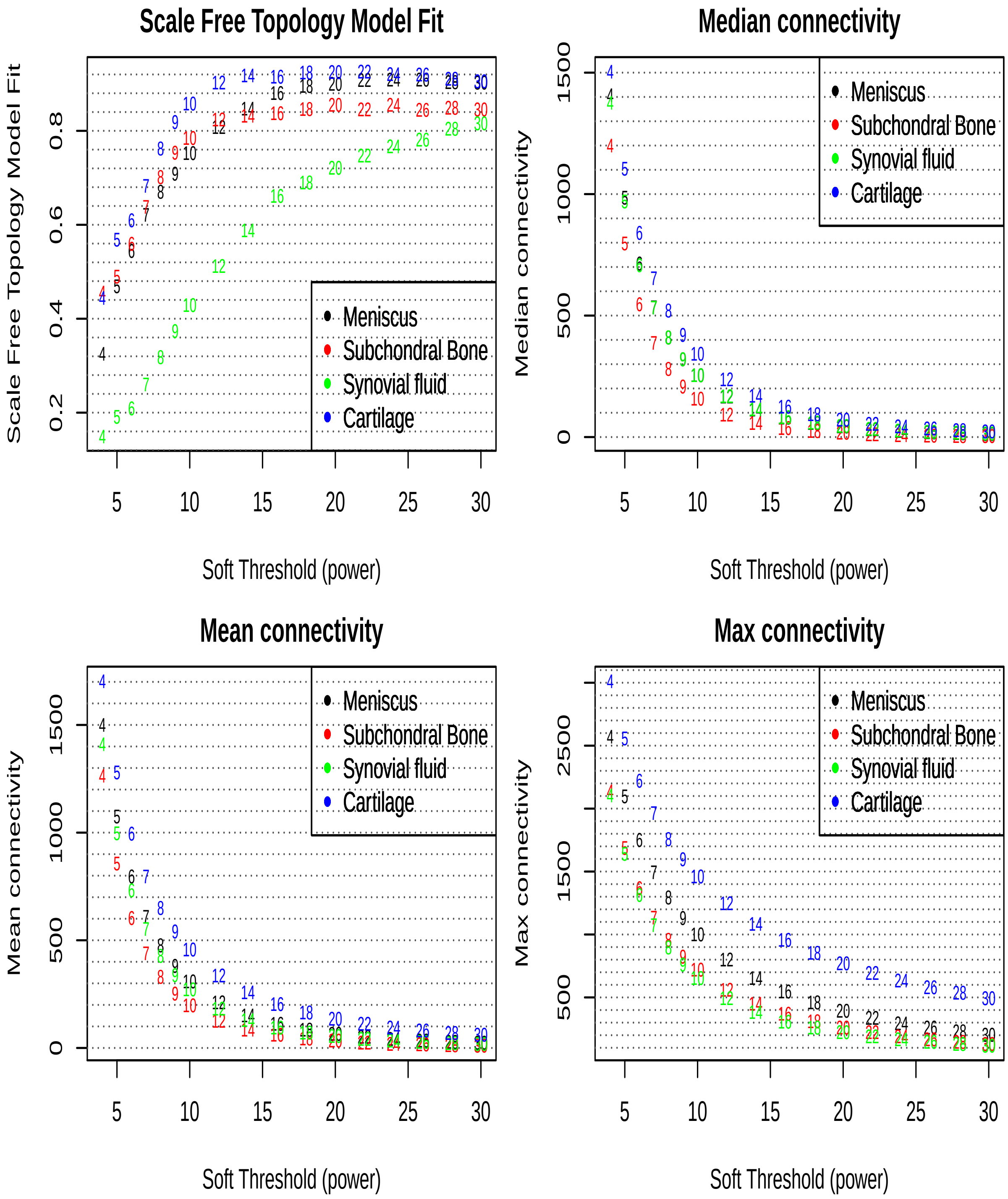
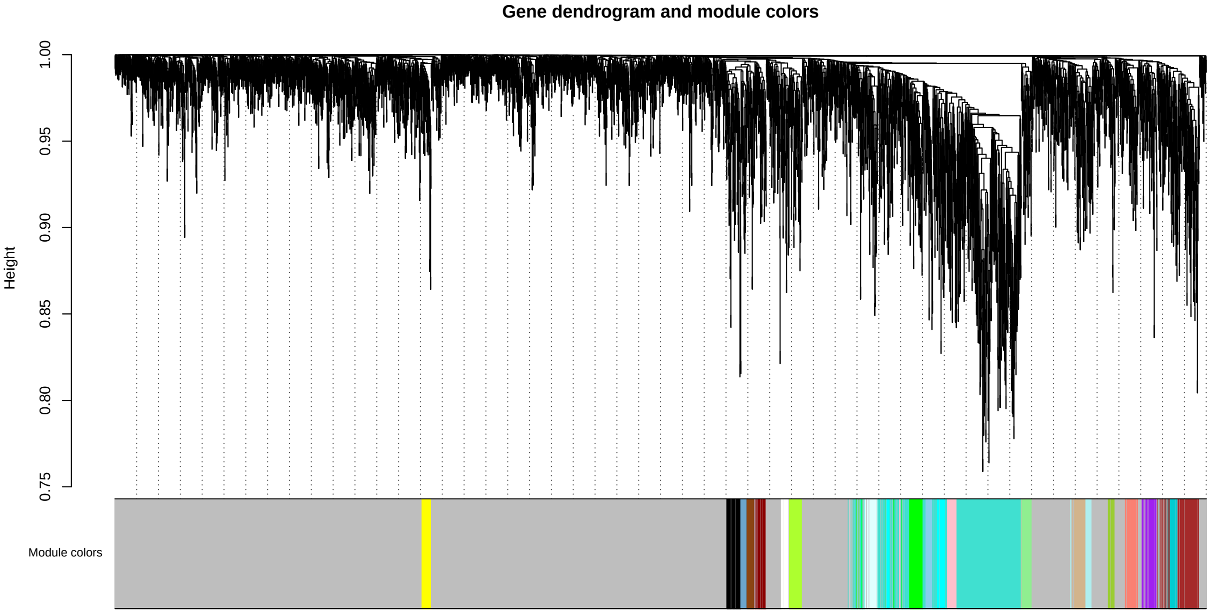


Figure 3.11: (a) Plot of Scale free topology Model Fit index $\sim$ Soft threshold power β.(b) Plot of Median Connectivity $\sim$ Soft threshold power β.(c) Plot of Mean Connectivity $\sim$ Soft threshold power β. (d) Plot of Max Connectivity $\sim$ Soft threshold power β.



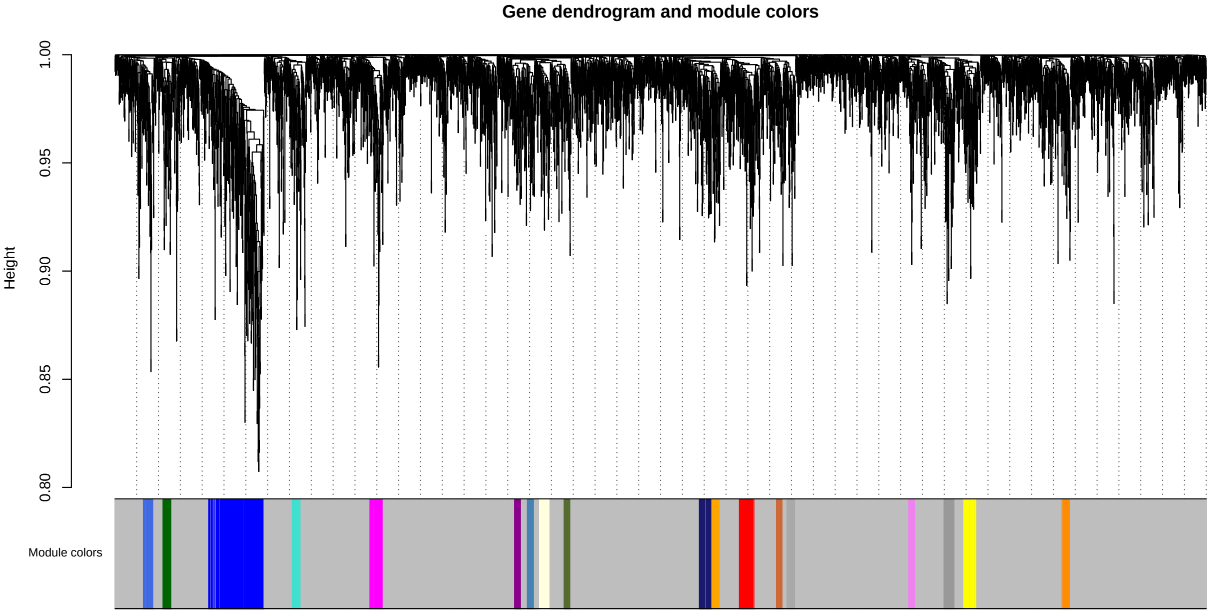


Figure 3.12: The consensus network that coonstructed from the four datasets and the modules that were identified. Different colours correspond to different modules.

3.5 Meta-Modules

Meta-modules are clusters of highly co-expressed modules. As shown in figures 3.13 : 3.16 some of the identified modules are highly correlated, and therefore can be merged together, creating meta-modules. The meta-modules can be detected as brances of the dendrograms or as the red boxes in the diagonal of the heat maps.

As shown in figures 3.13 : 3.16, the modules identified by WGCNA are clustered differently in each dataset. For example, in the meniscus dataset all modules can be clustered to 3 meta-modules as cutting the dendrogram to a height between 0.4 and 1.2 will result to three distinct brances. On the other hand, in the synovium dataset the number of the possibly identified meta-modules depend on the chosen cutting height in the dendrogram. Respecting the shape of the dendrogram in figure 3.15, it is safe to assume that the modules can be clustered into 6 meta-modules. The modules are clustered according to their gene's correlation in a specific dataset and therefore the identified meta-modules highly depend on the dataset used. For instance, the genes belonging in two modules may be highly co-expressed in the synovium dataset and not co-expressed in the meniscus dataset. Consequently, the respective modules will be clustered together only in the synovium dataset.

The most important conclusion of the meta-modules identification in the different datasets was that most of the identified by WGCNA modules were correlated and therefore they should be merged before a meta-analysis could be performed. Clustering the modules as shown in the heatmaps of the cartilage and meniscus dataset would conclude to three huge meta-modules and it could lead to very general results in the meta-analysis. On the other hand , clustering the modules as shown in the heatmap of the subchondral bone dataset would result to many small meta-modules and could lead to lack of results. Consequently, the modules were clustered as in the synovium dataset, creating 6 meta-modules.

In order to have a better picture of the meta-modules, the co-expression networks of the 2850 genes belonging to the meta-modules were constructed, as shown in figures 3.17 : 3.20. The different colours were used to distinguish which genes belong to each meta-module. As shown in figure 3.19 all meta-modules despite dark green are distinct from each other. This is normal, considering that the meta-modules were identified using the synovium Dataset. The genes of the dark green meta-module are scattered in all the constructed networks. Furthermore, it can be easily seen that in all other networks than synovium the purple and pink meta-module,as well as, the light green and the light cyan are merged topologically. For the construction of the co-expression networks of the meta-modules genes the program Cytoscape was used.

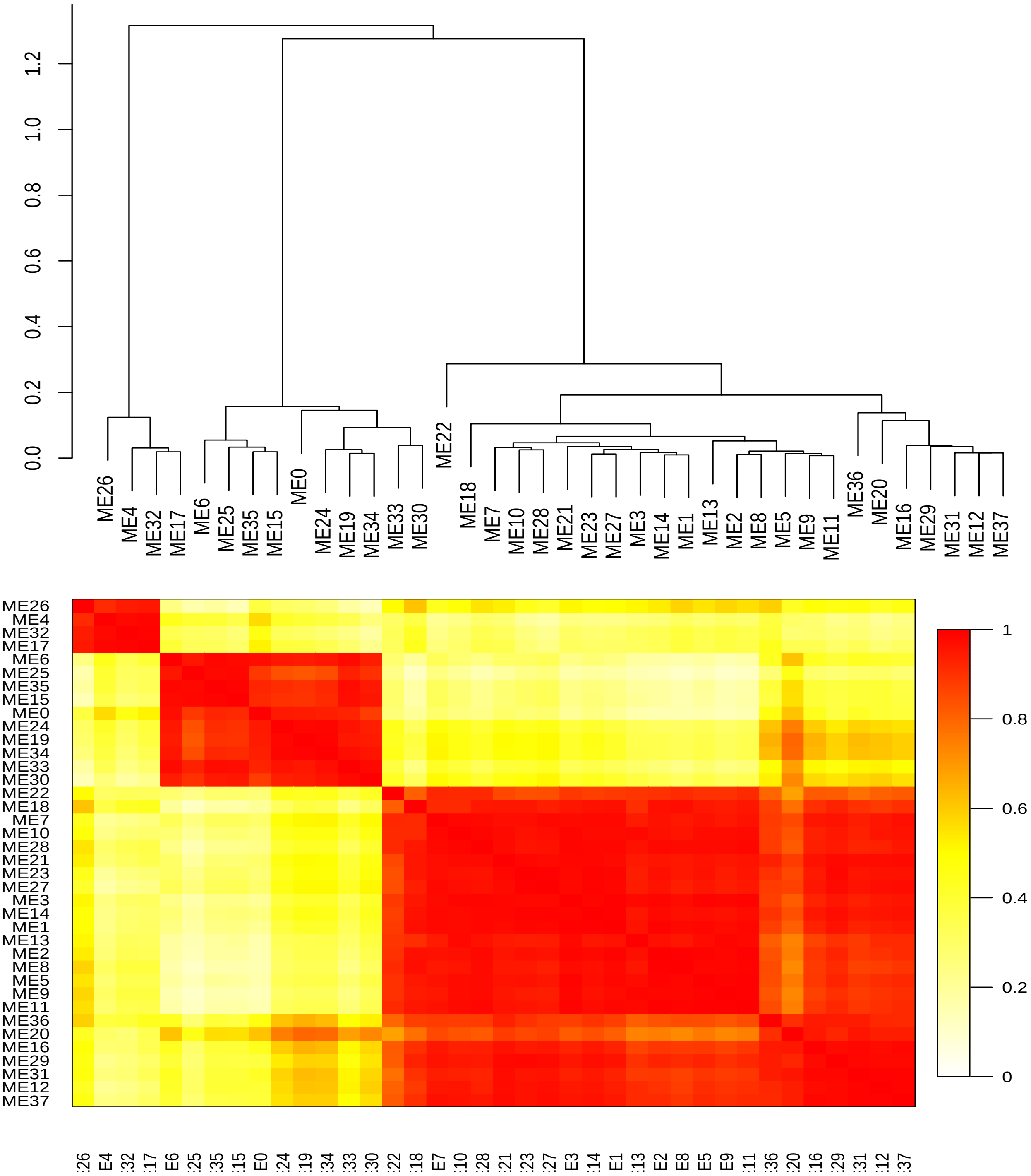


Figure 3.13: Hierarchical Clustering and Heatmap of the modules in the Meniscus Dataset.Each " MEXX " with XX being a number corresponds to a module.

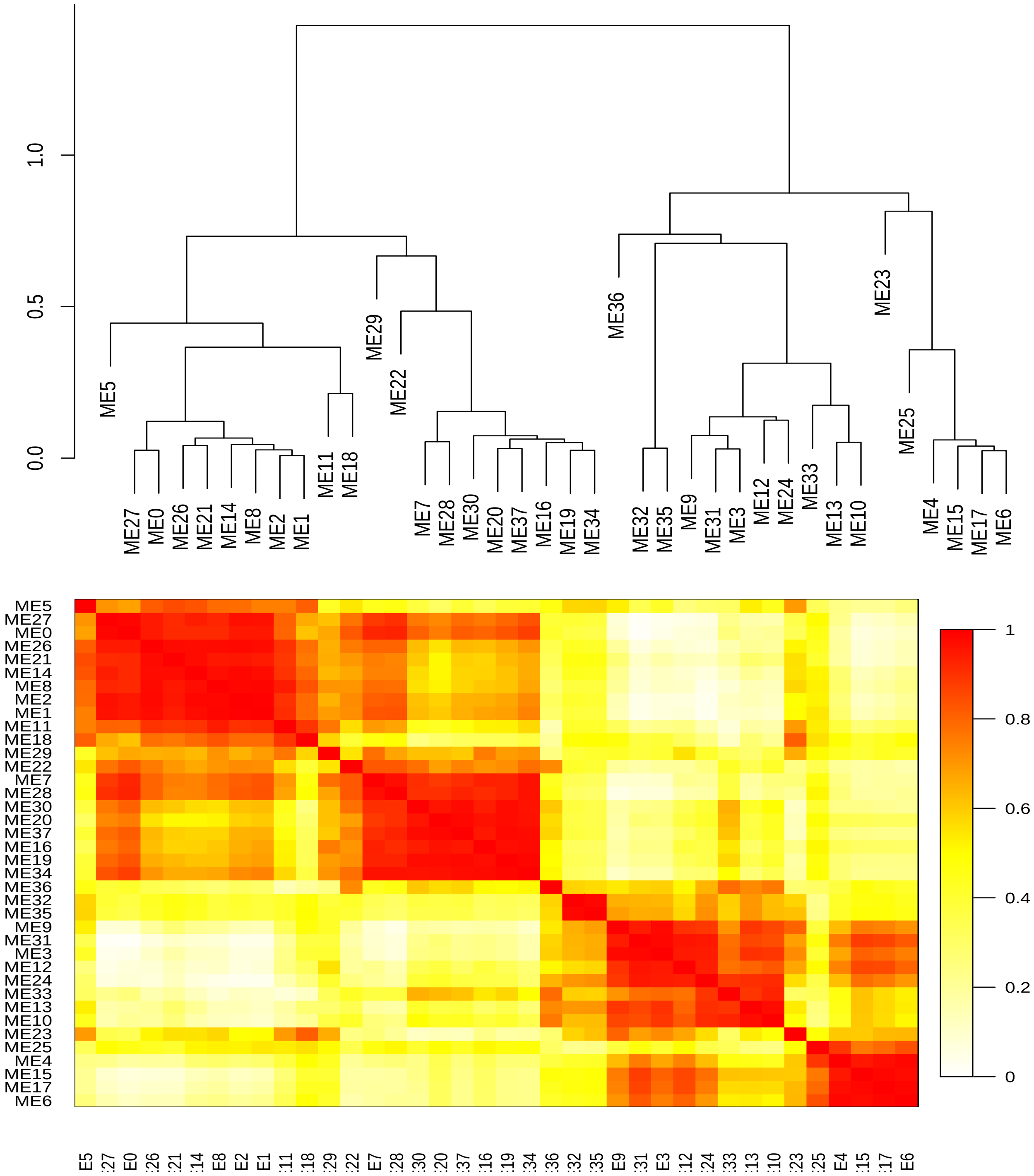


Figure 3.14: Hierarchical Clustering and Heatmap of the modules in the Subchondral Bone Dataset.Each " MEXX " with XX being a number corresponds to a module.

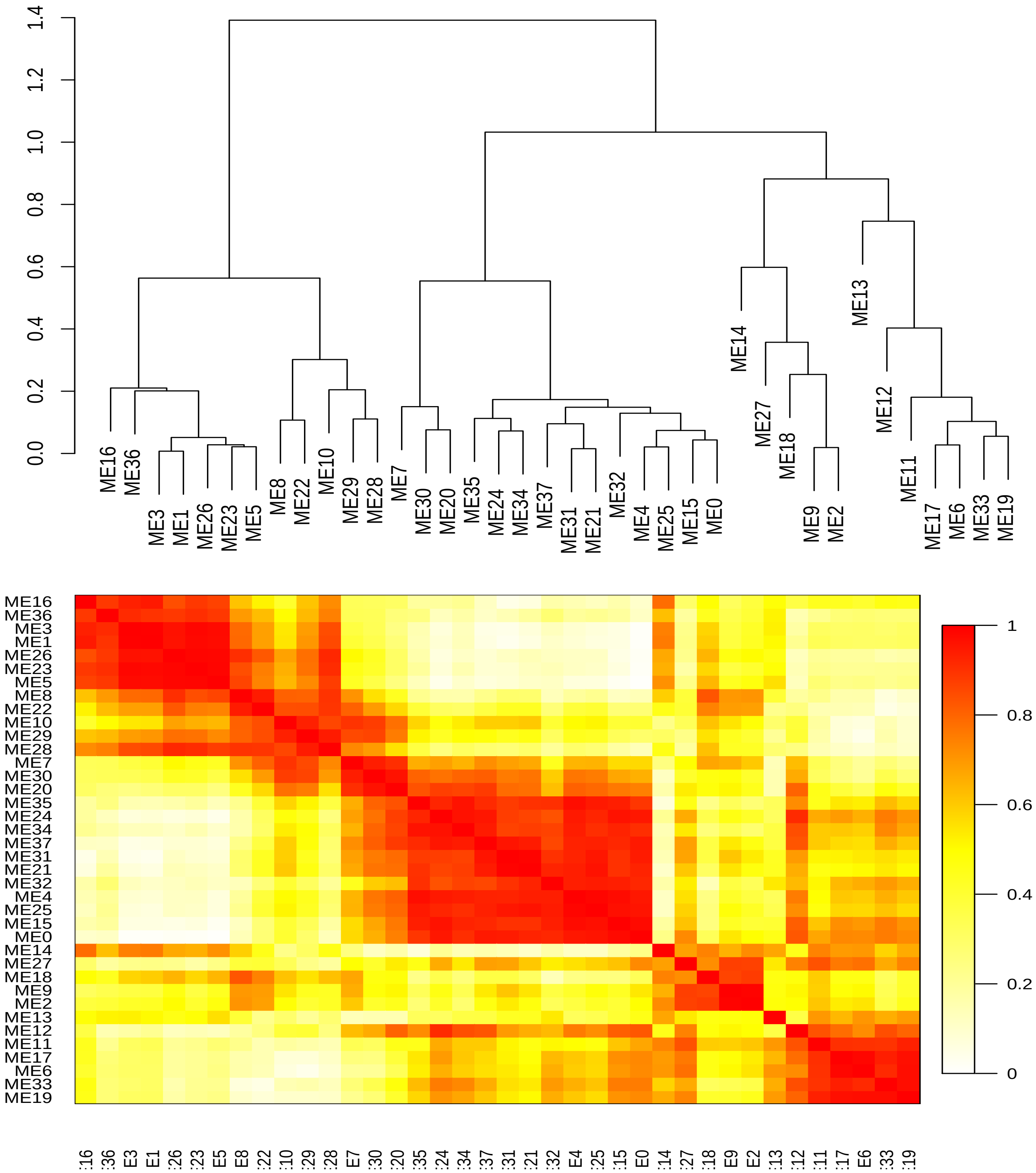


Figure 3.15: Hierarchical Clustering and Heatmap of the modules in the Synovium Dataset.Each " MEXX " with XX being a number corresponds to a module.

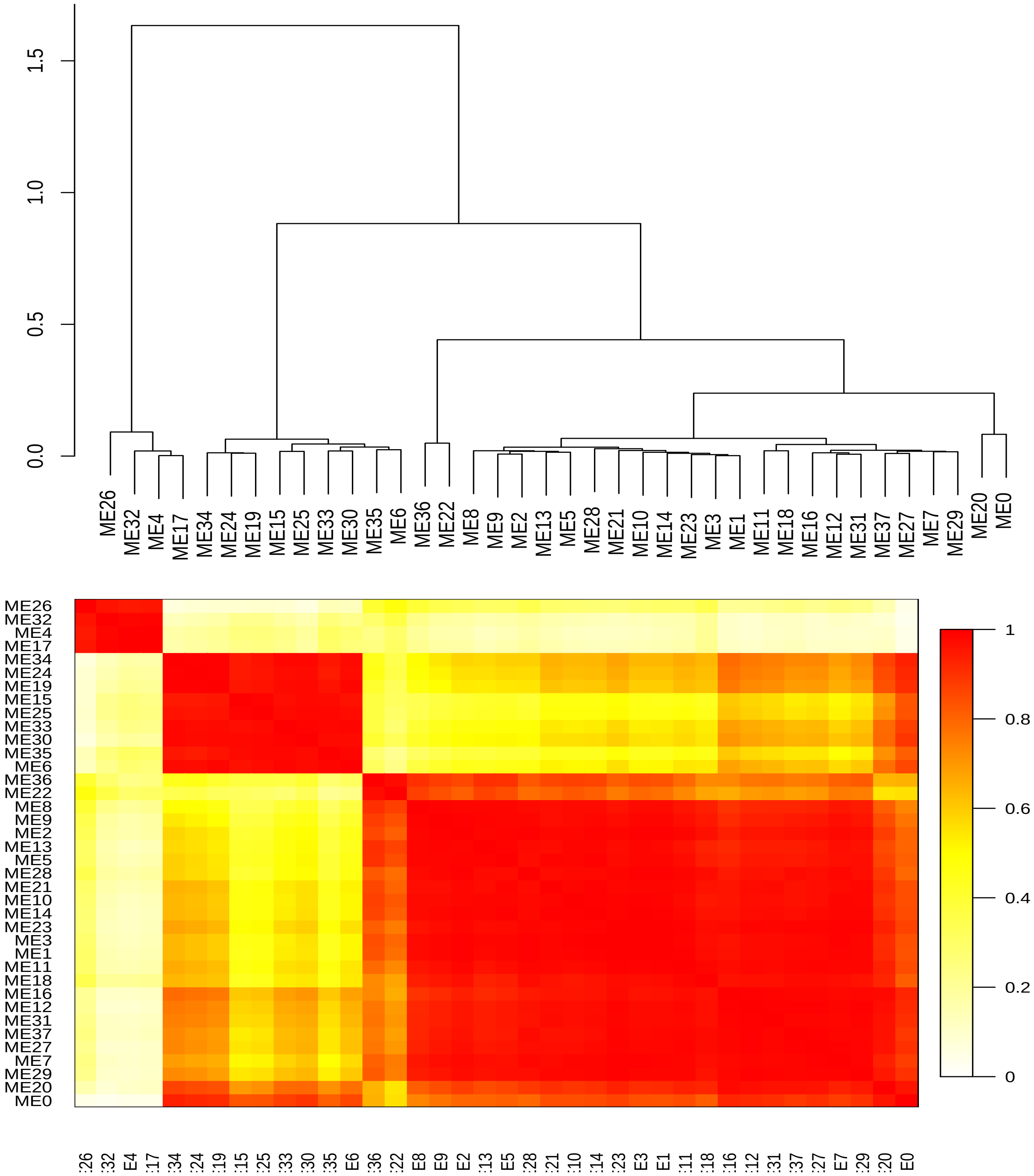


Figure 3.16: Hierarchical Clustering and Heatmap of the modules in the Cartilage Dataset.Each " MEXX " with XX being a number corresponds to a module.

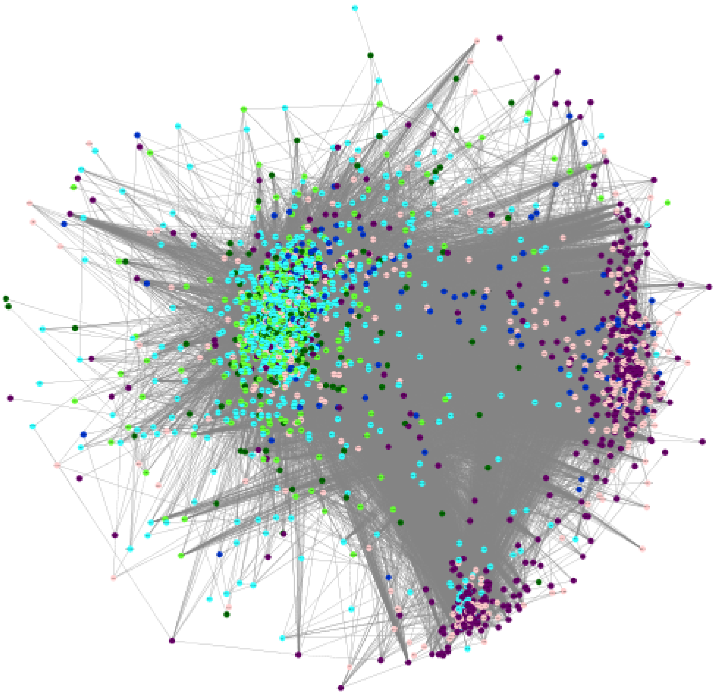


Figure 3.17: Network construction of the module's genes in the Meniscus Dataset. Each color correspond to one of the 6 meta-modules.

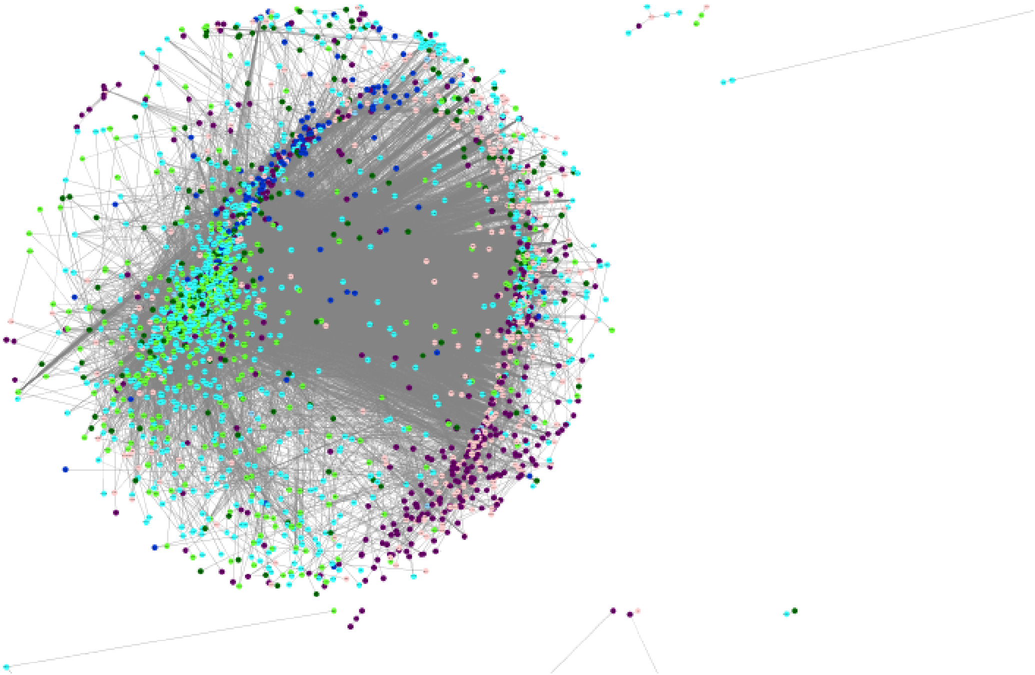


Figure 3.18: Network construction of the module's genes in the Subchondral Bone Dataset. Each color correspond to one of the 6 meta-modules.

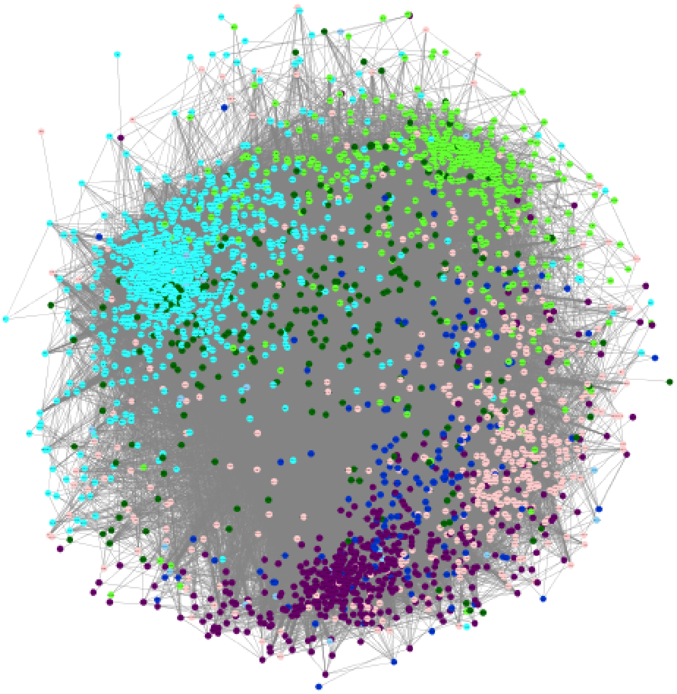


Figure 3.19: Network construction of the module's genes in the Synovium Dataset. Each color correspond to one of the 6 meta-modules.

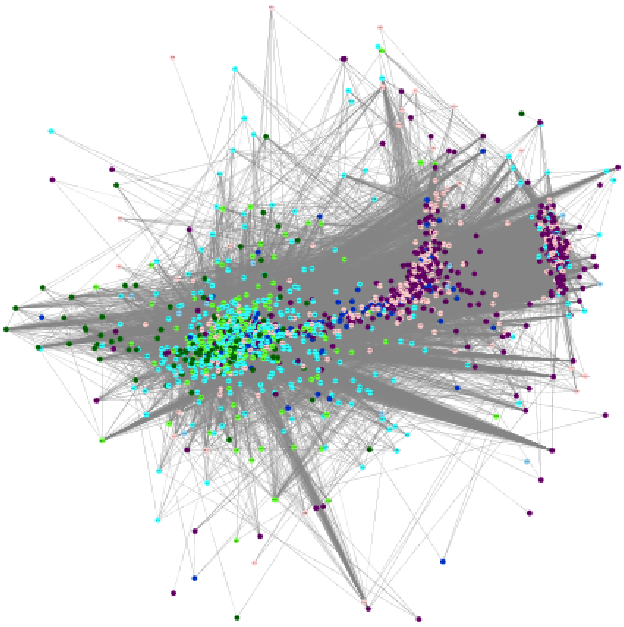


Figure 3.20: Network construction of the module's genes in the Cartilage Dataset. Each color correspond to one of the 6 meta-modules.

3.6 Pathway Analysis

Until this point of the analysis, genes whose expression levels were highly correlated between the samples of the four datasets were clustered together into modules. Then, modules that were highly correlated according to the synovium dataset were clustered together, creating 6 meta-modules.

The basis of a co-expression network analysis is that genes which are highly correlated, usually serve a common biological purpose. Therefore, each meta-module corresponds to biological functions and pathways that are common between the four tissues. In other words, the identification of the meta-modules assures that there are some common pathological mechanisms among the tissues involved in knee-osteoarthritis.

Pathway analysis was performed to each meta-module, in order to find its related pathways, using the package ” piano” [23] in the programming language R. For the Pathway analysis, the method ”Gene Set Enrichment Analysis(GSEA)” was used. GSEA takes as input a vector of gene’s t-values. As the meta-modules were identified using the synovium dataset, it was chosen to use the synovium dataset gene’s t-values for the Pathway Analysis. The identified pathways in each cluster can be seen in the tables 3.4 : 3.9

Pathways that have been studied for their importance in OA pathogenesis are presented in table 3.10, as well as a reference of the respective studies. In table 3.11 some of the identified pathways that are related to other diseases are presented. For the treatment of the presented diseases, drugs are studied originally used to treat arthritis. In the table 3.11 a reference of respective studies are presented also.

Purple meta−module
REACTOME SIGNALING BY RHO GTPASES
REACTOME TRANSCRIPTION
REACTOMEPOST_TRANSLATIONAL_PROTEIN_MODIFICATION

Table 3.4: The pathways identified in the purple meta-module.

Royal blue meta-module
none

Table 3.5: The pathways identified in the royal blue meta-module.

Pink meta-module
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION
NABA_COLLAGENS

Table 3.6: The pathways identified in the pink meta-module.

Dark green meta-module
KEGG_CELL_ADHESION_MOLECULES_CAMS
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION
KEGG_INTESTINAL_IMMUNE_NETWORK_FOR_IGA_PRODUCTION
KEGG_TYPE_I_DIABETES_MELLITUS
KEGG_LEISHMANIA_INFECTION
KEGG_ASTHMA
KEGG_AUTOIMMUNE_THYROID_DISEASE
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS
KEGG_ALLOGRAFT_REJECTION
KEGG_GRAFT_VERSUS_HOST_DISEASE
KEGG_VIRAL_MYOCARDITIS
REACTOME_MHC_CLASS_II_ANTIGEN_PRESENTATION
REACTOME_IMMUNE_SYSTEM
REACTOME_ADAPTIVE_IMMUNE_SYSTEM

Table 3.7: The pathways identified in the dark green meta-module.

Light cyan meta-module
KEGG_SPLICEOSOME
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON
ST_JNK_MAPK_PATHWAY
PID_TAP63_PATHWAY
REACTOME_TCA_CYCLE_AND_RESPIRATORY_ELECTRON_TRANSPORT
REACTOME_TRANSPORT_OF_MATURE_TRANSCRIPT_TO_CYTOPLASM
REACTOME_INTEGRIN_CELL_SURFACE_INTERACTIONS
REACTOME_MRNA_PROCESSING
REACTOME_CLEAVAGE_OF_GROWING_TRANSCRIPT_IN_THE_TERMINATION_REGION_

Table 3.8: The pathways identified in the light cyan meta-module.

Light green meta-module
REACTOME_3_UTR_MEDIATED_TRANSLATIONAL_REGULATION
NABA_MATRISOME_ASSOCIATED
NABA_MATRISOME

Table 3.9: The pathways identified in the light green meta-module.

Pathways	Reference
REACTOME_SIGNALING_BY_RHO_GTPASES	Mengxi et al [30]
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON	Gardiner et al [31]
ST_JNK_MAPK_PATHWAY	Loeser et al [32]
PID_TAP63_PATHWAY	Taniguchi et al [33]
REACTOME_TCA_CYCLE_AND _RESPIRATORY_ELECTRON_TRANSPORT	Mobasheri et al [34]

Table 3.10: Pathways that were studied for their connection with Osteoarthritis

Pathways	Reference
KEGG_TYPE_I_DIABETES_MELLITUS	Zhang et al [35]
KEGG_LEISHMANIA_INFECTION	Roder et al [36]
KEGG_ASTHMA	Huo et al [37] Kruse et al [38]
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	Zhang et al [39]
KEGG_GRAFT_VERSUS_HOST_DISEASE	Blatt et al [40]

Table 3.11: References of studies using drug repositioning from arthritis for dealing with the presented diseases

3.7 Stability Analysis

In order for the modules identified by WGCNA to be biologically meaningful, they should be stable. In other words, they should not be affected by small changes in the co-expression matrix. In this diploma thesis, the stability of the constructed networks and therefore the stability of the identified modules was established with two different procedures. Each procedure was repeated 50 times.

The first procedure consisted of randomly removing 10 % of the samples of each microarray dataset. Consequently, four new datasets were created. These datasets were used for conducting WGCNA in the same way with the original ones. Specifically, new networks were constructed and a new set of common modules were identified. Finally, for the stability analysis, the new set of modules was compared with the original one. In the second procedure, we used re-sampling of the samples for defining the new datasets. Afterwards, as in the first procedure, new networks were constructed and a new set of modules were identified and compared with the original set.

In figures 3.22 and 3.23 the modules identified in each run of the first and second procedure respectively are depicted, as well as the original modules. It can be easily seen that most of the original modules were constantly identified. As shown in figure 3.21 $\sim$ 75 % on average of the genes belonging to the original modules could be identified in the modules of each run.

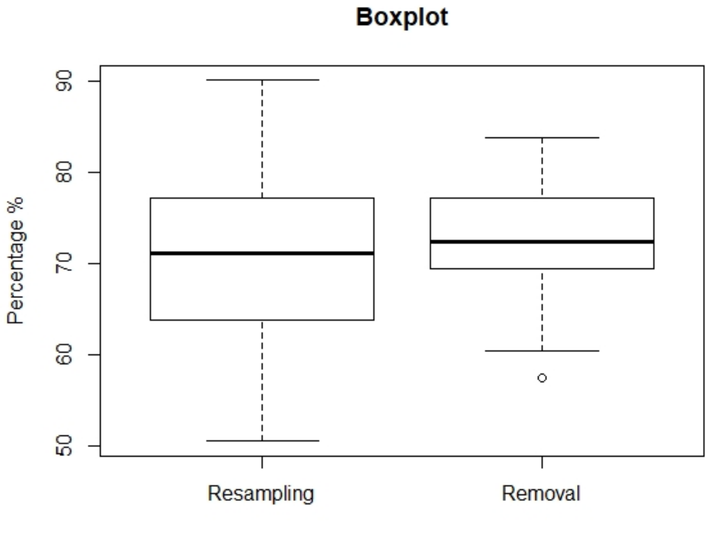
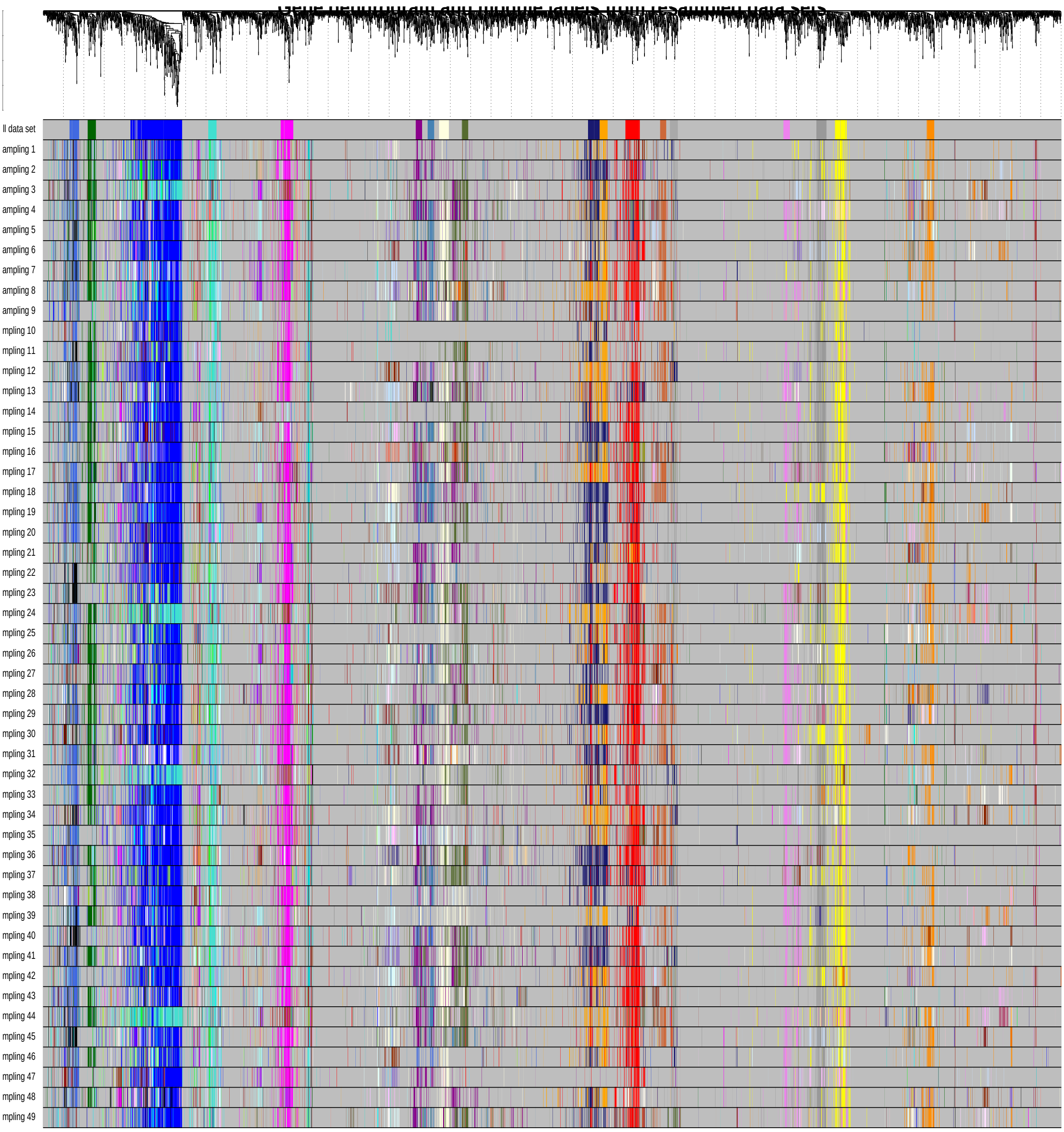


Figure 3.21: Boxplot of the percentage of the genes belonging to the original modules that were identified in the modules of each run


Gene dendrogram and module labels from resampled data sets
ll data set
ampling 1
ampling 2
ampling 3
ampling 4
ampling 5
ampling 6
ampling 7
ampling 8
ampling 9
mpling 10
mpling 11
mpling 12
mpling 13
mpling 14
mpling 15
mpling 16
mpling 17
mpling 18
mpling 19
mpling 20
mpling 21
mpling 22
mpling 23
mpling 24
mpling 25
mpling 26
mpling 27
mpling 28
mpling 29
mpling 30
mpling 31
mpling 32
mpling 33
mpling 34
mpling 35
mpling 36
mpling 37
mpling 38
mpling 39
mpling 40
mpling 41
mpling 42
mpling 43
mpling 44
mpling 45
mpling 46
mpling 47
mpling 48
mpling 49

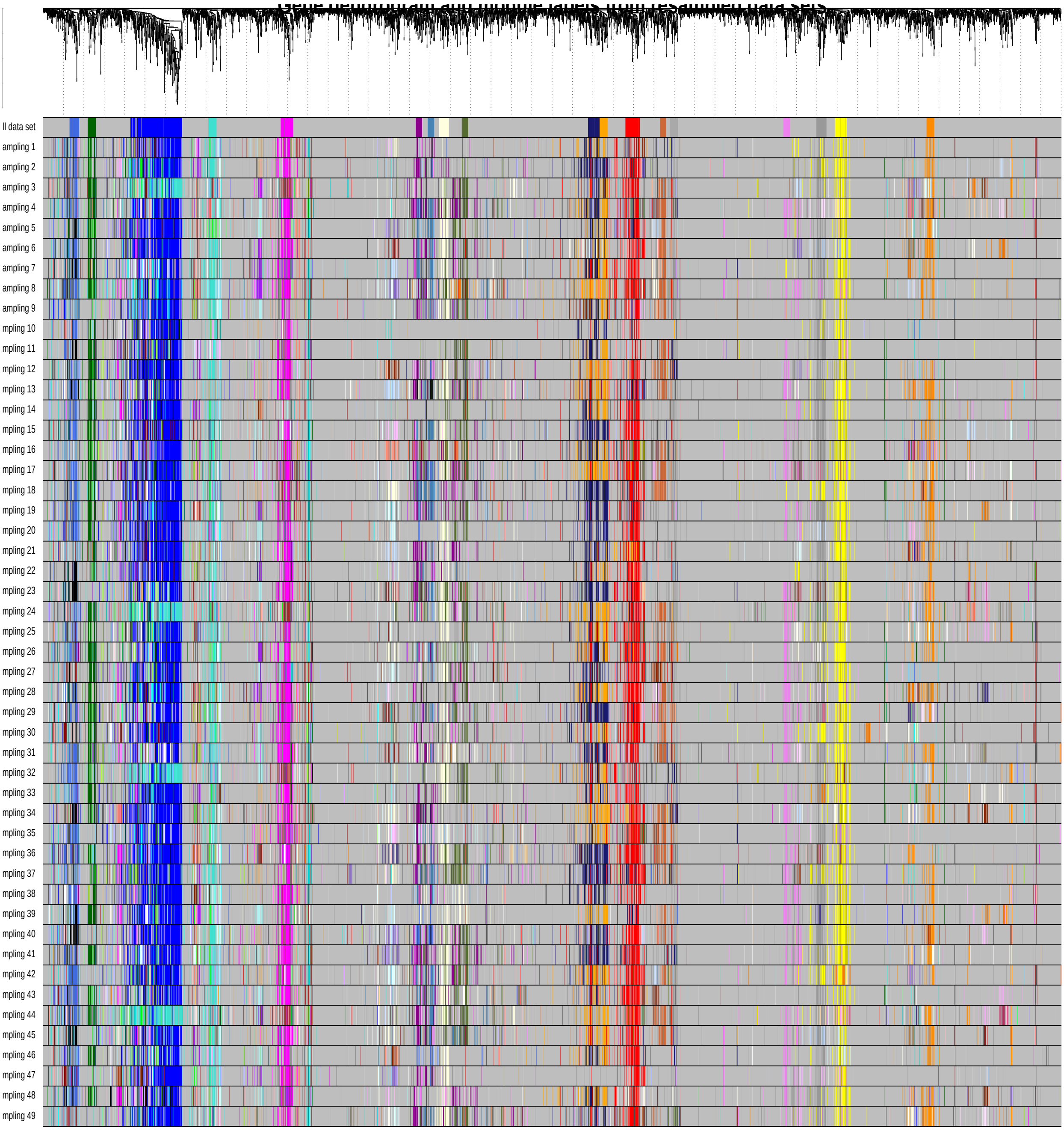
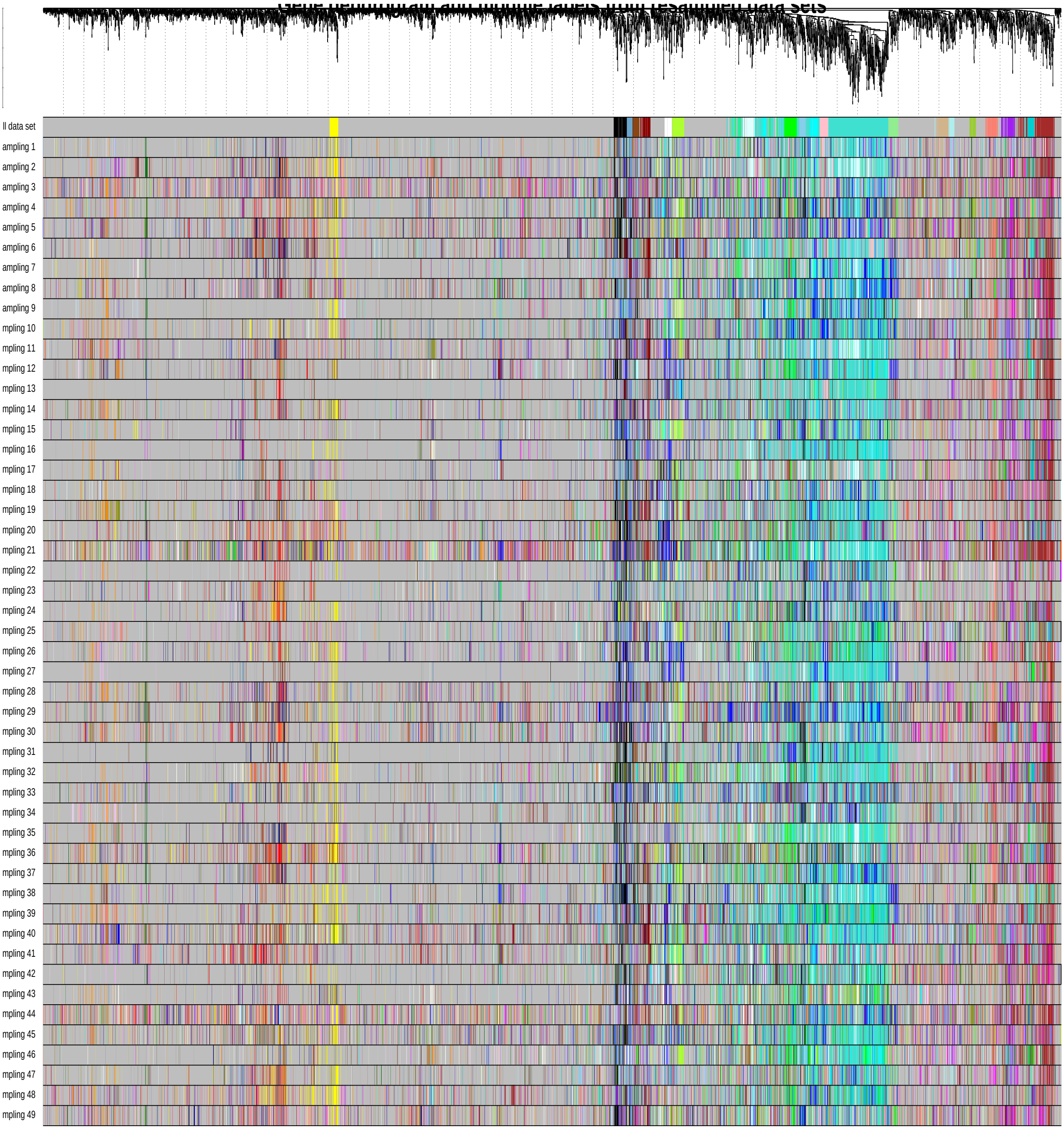


Figure 3.22: Module Identification in 50 new datasets derived from the original datasets with removing randomly 10 % of the Samples


Gene dendrogram and module labels from resampled data sets
ll data set
ampling 1
ampling 2
ampling 3
ampling 4
ampling 5
ampling 6
ampling 7
ampling 8
ampling 9
mpling 10
mpling 11
mpling 12
mpling 13
mpling 14
mpling 15
mpling 16
mpling 17
mpling 18
mpling 19
mpling 20
mpling 21
mpling 22
mpling 23
mpling 24
mpling 25
mpling 26
mpling 27
mpling 28
mpling 29
mpling 30
mpling 31
mpling 32
mpling 33
mpling 34
mpling 35
mpling 36
mpling 37
mpling 38
mpling 39
mpling 40
mpling 41
mpling 42
mpling 43
mpling 44
mpling 45
mpling 46
mpling 47
mpling 48
mpling 49

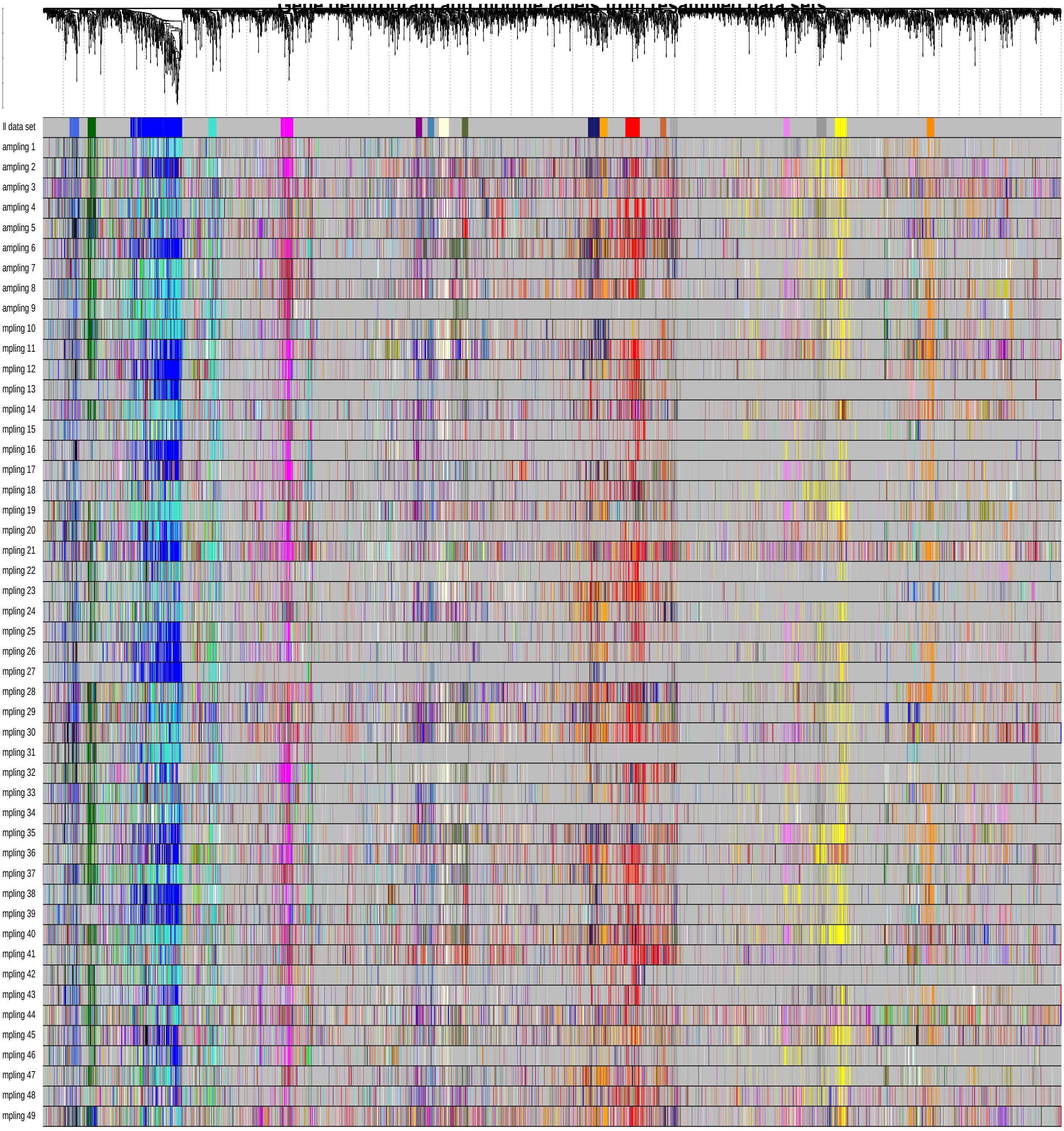


Figure 3.23: Module Identification in 50 new datasets derived from the original datasets with Re-Sampling

3.8 Drugs

As the final step of the diploma thesis a list of drugs either available in the laboratory or found in previous studies were evaluated. Firstly, their main targets were identified according to the on-line database DrugBank. Then it was checked if any of their identified targets was gene of the meta-modules. The drugs evaluated, as well as their main targets can be seen in the table 3.12

Drugs	Targets
Canakinumab	IL1B
Adalimumab	TNF, IL1A
Celecoxib	PTGS2, PDPK1, CA2, CA3, ABCB5, ABCG2, ABCB1
Carprofen	PTGS2, PTGS1
SP600125	MAPK inhibitor
Raloxifene	ESR1, ESR2, SEPRINB9, TFF1
PD169316	TGFB, TGFA
SB202190	TBRI
Sorafenib	BRAF, RAF1,FLT4 KDR, FLT3, PDGFR, KIT, FGFR1, RET, FLT1
Clomifene	ER1, SHBG
Imperatorin	ERK

Table 3.12: The drugs evaluated and their main targets.

From the drugs evaluated, Raloxifene, PD169316 and Sorafenib had targets that were genes of meta-modules.

3.9 Comparison with in vitro experiments

A setup as presented by Neidlin et al [41] has been used to experimentally evaluate the efficacy of drugs from table 3.12. Specifically, Imperatorin, SP600125, SB202190, PD169316, Regorafenib, Sorafenib, Clomifene, Rapamycin, Raloxifene were tested in this experiment. In brief, healthy and degrading (OA) cartilage tissue explants were treated with each of the drugs for 24h. After stimulation for 24h the supernatant was retrieved and analysed with a bead-based sandwich ELISA assay (Luminex xMAP) for a set of 26 cytokines that included major players involved in inflammation, cartilage degeneration and osteoarthritis: (PEDF, CXCL11, IL13, ZG16, IL4, GROA, IFNG, CYTC, IL8, IL17F, IL12A, TNFA, IL1A, TFF3, IL6, ICAM1, IL10, FST, S100A6, CXCL10, PROK1, CCL5, IL20, TNFSF12, FGF2, MMP9). The response was used as a measure to evaluate drug efficacy. In other words, degrading tissue treated with drug

A should produce a response as close as possible to healthy untreated tissue. Figure 3.24 shows the relative absolute distance (fluorescence intensities of treated disc - fluorescence intensities of healthy disc/fluorescence intensities of healthy disc) for drug treatments with concentrations of 1µM and 10 µM respectively.

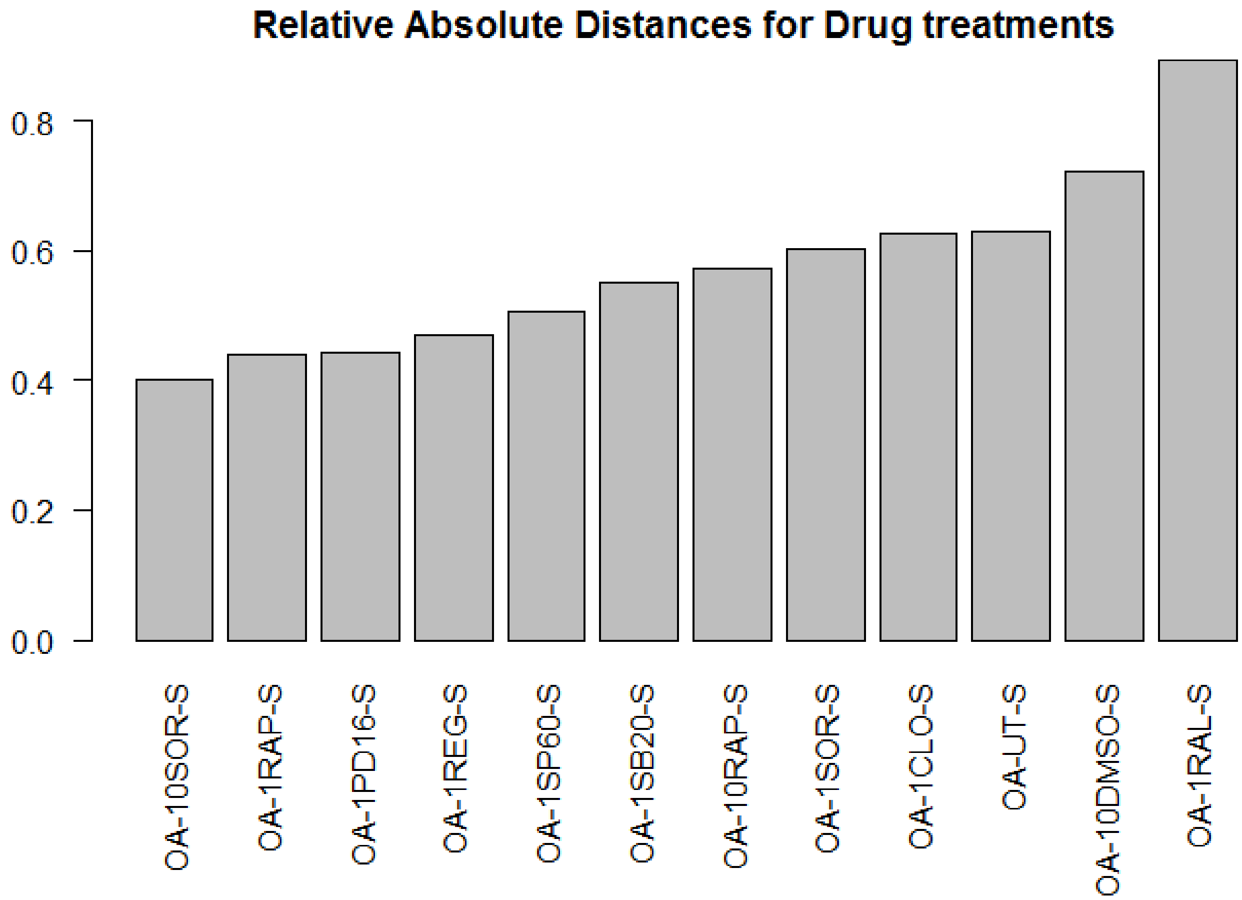


Figure 3.24: The relative absolute distance between healthy and treated with the drugs of table 3.12 cartilage tissue

The above independent experimental drug screening study identified also Sorafenib and PD169316 as promising compounds.

CHAPTER 4

Discussion

OA is a multi-tissue disease, including cartilage degradation, meniscus and subchondral bone alterations and synovium inflammation. The purpose of this diploma thesis was to identify common pathological mechanisms among the four tissues. To that end, Differential Expression Analysis and Weighted Gene Co-Expression Network Analysis were implemented using four microarray datasets corresponding to the four tissues. Through WGCNA, 37 modules were identified including 2850 genes. Highly correlated modules were merged creating six meta-modules. The genes included in each meta-module were used for the identification of important pathways through Pathway Analysis. As a final step, candidate drugs for OA were evaluated.

Through the above-mentioned analysis we have identified meta-modules closely related to specific functions. The pink and the light green meta-module are related to extracellular matrix. This is an interesting finding considering that biological functions related to extracellular matrix has been only reported to cartilage OA. The dark green meta-module includes many pathways related to the immune system and inflammation. Furthermore, it includes pathways related to other major diseases. Interestingly, drug repositioning from arthritis have been studied for treating a great number of these diseases as shown in table 3.11. Finally, the light cyan meta-module includes many signalling pathways with a great number of them being reported in recent studies for their correlation with OA, as shown in table 3.10.

The drug evaluation done in this diploma thesis, has concluded to Sorafenib, Raloxifene and PD169316 as the most promising drugs for OA treatment. An experimental drug screening study testing 9 drugs identified independently Sorafenib and PD169316 as promising compounds.

Choosing the Weighted Gene Co-Expression Analysis to be the chore of this diploma thesis stems from many reasons. WGCNA allows the integration of multiple heterogeneous datasets and therefore provides a better understanding of complex diseases like osteoarthritis, while Differential Expression Analysis could only be implemented in each dataset separately. Furthermore, DEGs depend on multiple statistical hypothesis and the cut-off threshold chosen by the researcher for the measures of magnitude and significance. Consequently, there is a statistical uncertainty of whether the DEGs are correlated with the disease or not. On the other hand WGCNA does not involve multiple hypothesis testing and therefore limits the discovery of false positives and true negatives. Moreover, it has not a strict threshold, but encounters the relations of all genes in order to find the co-expression modules.

However, as with any method available in the bibliography, it has its limitations. First of all, as the adjacency matrix depends on the original microarray dataset, the constructed network and therefore the identified modules are influenced by the machine used to generate the data as well as the background correction and normalisation methods used in the data. Furthermore, the WGCNA results depend on the sample size, as small sample size means that the network could not exhibit approximate scale-free topology. Another disadvantage of WGCNA concerns the integration of multiple datasets that correspond to heterogeneous tissues, states etc. Tissue-specific or condition specific co-expression modules may not be detectable in a co-expression network constructed from multiple tissues or conditions because the correlation signal of the tissue or condition-specific modules is weakened by a lack of correlation in other tissues/conditions.

As far as the implementation of WGCNA in this diploma thesis is concerned, one limitation would be that the synovium dataset was generated by Affymetrix technology whereas the other datasets were generated with Agilent technology. Furthermore, it should be beneficial to check what proportion of the genes in each meta-module were responsible for the enriched pathways found. For instance, it may be the case that the pathways identified in the dark green meta-module conclude only 50 % of the genes included in that meta-module. In this way , we could identify biological function-specific genes in the meta-modules. Moreover, it would be beneficial to check if the modules are preserved in normal tissues and if the genes belonging to the meta-modules have a different expression in the normal samples.

In conclusion, there are many studies published that try to identify important functions in OA using ”omics” data of only one tissue.[8],[7],[9] However, OA is a disease of the whole joint and therefore the integration of datasets of all tissues is needed so as to understand OA pathological mechanisms. This diploma thesis, is the first attempt to our knowledge to integrate all four tissues with promising results. We have managed to detect pathways, already established for their relation to OA, like signalling by rho GTPases.[30] Furthermore, we identify new pathways that play an important role to all four datasets and may be important to understand OA pathogenesis. Finally, drugs showing possible OA adverse properties were found both in this analysis and an independent experimental study.

As further work, we consider expanding the list of drugs for evaluation. Furthermore, we would try to incorporate into the drug evaluation the results of the Pathway Analysis and the network properties. Finally, new experiments will be performed.

Bibliography

[1] Lane N. Osteoarthritis year in review 2016: clinical ; 2017.

[2] Martel-Pelletier J, Wildi LM, Pelletier JP. Future therapeutics for osteoarthritis ; 2012.

[3] Martel-Pelletier1 J, , Andrew J Barr2 , , Flavia M Cicuttini4, Philip G Conaghan2 , et al. Osteoarthritis. Annals of Indian Academy of Neurology. 2016;.

[4] Karsdal MA, Michaelis M, Ladel C, et al. Disease-modifying treatments for osteoarthritis (DMOADs) of the knee and hip: lessons learned from failures and opportunities for the future ; 2016.

[5] Mueller AJ, Canty-Laird EG, Clegg PD, et al. Cross-species gene modules emerge from a systems biology approach to osteoarthritis. npj Systems Biology and Applications. 2017;.

[6] Goldring MB. Osteoarthritis and cartilage: the role of cytokines. ; 2000.

[7] Melas IN, Chairakaki AD, Chatzopoulou EI, et al. Modeling of signaling pathways in chondrocytes based on phosphoproteomic and cytokine release data. Osteoarthritis and Cartilage. 2014;.

[8] Mariani E, Pulsatelli L, Facchini A. Signaling pathways in cartilage repair. International Journal of Molecular Sciences. 2014;.

[9] Brophy RH, Zhang B, Cai L, et al. Transcriptome comparison of meniscus from patients with and without osteoarthritis. Osteoarthritis and Cartilage. 2018;.

[10] Park R, Ji JD. Unique gene expression profile in osteoarthritis synovium compared with cartilage: analysis of publicly accessible microarray datasets. Rheumatology International. 2016;.

[11] Felson DT. Osteoarthritis: Priorities for osteoarthritis research: much to be done. Nature Reviews Rheumatology. 2014;.

[12] Edwards D. Non-linear normalization and background correction in one-channel cDNA microarray studies. Bioinformatics. 2003;.

[13] Silver JD, Ritchie ME, Smyth GK. Microarray background correction: Maximum likelihood estimation for the normal-exponential convolution. Biostatistics. 2009;.

[14] Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47.

[15] Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data; 2003. 2. Available from: https://academic.oup.com/biostatistics/article-abstract/4/2/249/245074.

[16] Ritchie ME, Silver J, Oshlack A, et al. A comparison of background correction methods for two-colour microarrays. Bioinformatics. 2007;23(20):2700–2707.

[17] McGee M, Chen Z. Parameter estimation for the exponential-normal convolution model for background correction of affymetrix GeneChip data. Statistical Applications in Genetics and Molecular Biology. 2006;5(1).

[18] Shi W, Oshlack A, Smyth GK. Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips. Nucleic Acids Research. 2010;38(22).

[19] Bolstad BM, Irizarry RA, Astrand M, et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias; 2003. 2. Available from: http://www.bioconductor.org.

[20] Yang. Normalisation for cDNA Microarray Data; 2001. Available from: http://spiedl.org/terms.

[21] Yang YH, Dudoit S, Luu P, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation; 2002. 4.

[22] Yang YH, Thorne NP. Normalization for two-color cDNA microarray data. 2003;:403–418Available from: http://projecteuclid.org/euclid.lnms/1215091155.

[23] Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Research. 2013;.

[24] Irizarry RA, Wang C, Zhou Y, et al. Gene set enrichment analysis made simple. Statistical Methods in Medical Research. 2009;.

[25] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;.

[26] Langfelder H. WGCNA: an R package for weighted correlation network analysis ; 2008.

[27] STEVE H. Weighted Network Analysis_ Applications in Genomics and Systems Biology-Springer-Verlag New York (2011). 2011;.

[28] Langfelder P, Horvath S. Fast R Functions for Robust Correlations and Hierarchical Clustering; 2012.

[29] Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC systems biology. 2007;.

[30] Lv M, Zhou Y, Polson SW, et al. Identification of Chondrocyte Genes and Signaling Pathways in Response to Acute Joint Inflammation. Scientific Reports. 2019;9(1):93. Available from: http://www.nature.com/articles/s41598-018-36500-2.

[31] Gardiner MD, Vincent TL, Driscoll C, et al. Transcriptional analysis of micro-dissected articular cartilage in post-traumatic murine osteoarthritis. Osteoarthritis and Cartilage. 2015; 23(4):616–628. Available from: http://dx.doi.org/10.1016/j.joca.2014.12.014.

[32] Loeser RF, Erickson EA, Long DL. Mitogen-activated protein kinases as therapeutic targets in osteoarthritis. Current Opinion in Rheumatology. 2008;20(5):581–586.

[33] Taniguchi Y, Kawata M, Ho Chang S, et al. Regulation of Chondrocyte Survival in Mouse Articular Cartilage by p63. Arthritis and Rheumatology. 2017;69(3):598–609.

[34] Mobasheri A, Rayman MP, Gualillo O, et al. The role of metabolism in the pathogenesis of osteoarthritis. Nature Reviews Rheumatology. 2017;13(5):302–311. Available from: http://dx.doi.org/10.1038/nrrheum.2017.50.

[35] Zhang P, Wang F, Hu J. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. AMIA Annual Symposium proceedings AMIA Symposium. 2014;2014:1258–67. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25954437{%}0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4419869.

[36] Roder C, Thomson MJ. Auranofin: Repurposing an Old Drug for a Golden New Age. Drugs in R and D. 2015;15(1):13–20.

[37] Huo Y, Zhang HY. Genetic mechanisms of asthma and the implications for drug repositioning. Genes. 2018;9(5).

[38] Kruse RL, Vanijcharoenkarn K. Drug repurposing to treat asthma and allergic disorders: Progress and prospects. Allergy: European Journal of Allergy and Clinical Immunology. 2018;73(2):313–322.

[39] Zhang M, Luo H, Xi Z, et al. Drug repositioning for diabetes based on 'omics' data mining. PLoS ONE. 2015;10(5):1–13.

[40] Blatt J, Corey SJ. Drug repurposing in pediatrics and pediatric hematology oncology. Drug Discovery Today. 2013;18(1-2):4–10. Available from: http://dx.doi.org/10.1016/j.drudis.2012.07.009.

[41] Neidlin M, Chantzi E, Macheras G, et al. Investigations of cytokine interplay with an in vitro model of osteoarthritis. Osteoarthritis and Cartilage. 2018;26(2018):S111. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1063458418303431.