A Thesis

Entitled

POTENTIAL GENETIC BIOMARKERS FOR DILATED CARDIOMYOPATHY USING GENOMIC DATA

by

Ammar F Eljack

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the

Master of Science in Biomedical Science Degree in

Bioinformatics, Proteomics, and Genomics

_____

Dr. Sadik Khuder, Committee Chair

_____

Dr. Robert Blumenthal, Committee Member

_____

Dr. Jiang Tian, Committee Member

_____

Dr. Cyndee Gruden

College of Graduate Studies

The University of Toledo
May 2020

An Abstract of

POTENTIAL GENETIC BIOMARKERS FOR DILATED CARDIOMYOPATHY USING GENOMIC DATA

by

Ammar F. Eljack

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the

Master of Science in Biomedical Science Degree in

Bioinformatics, Proteomics, and Genomics

The University of Toledo

May 2020

Dilated cardiomyopathy (DCM) belongs to the heterogeneous group of heart muscle disorders called cardiomyopathies, characterize by left ventricular dilatation and reduce myocardial muscle contractility that leads to a reduced ejection fraction of the heart. There are several causes of DCM; genetic, toxic, metabolic, endocrine, infiltrative, and idiopathic disorders. Forty causative genes encoding for a variety of proteins have been identified up to date. The objective of this study is to identify potential biomarkers related to the disease process of DCM.

Microarray and RNA-Seq profiles of cardiac tissue were used to identify differentially expressed genes (DEGs). Four Microarray datasets (GSE 3585, GSE3586, GSE9800, and GSE42955), and three RNA-Seq datasets were selected (GSE55296, GSE65466, and GSE71613) from the Gene Expression Omnibus (GEO) database. Weighted gene co-expression network analysis (WGCNA) was used to explore the hub genes involved in the disease process of DCM.

A total of 51 modules for microarray datasets and 15 modules for RNA-Seq datasets were identified using correlation network analysis. Four common statistically- significant genes were identified among these modules, including AP3M2, ECM2, ERBB2, and ZNF83. AP3M2 gene, which involves protein trafficking to lysosomes and specialized organelles, ECM2 gene, which affects in extracellular matrix protein function, ERBB2 gene, which involves in Erb-B2 Receptor Tyrosine Kinase 2 signaling pathway, and ZNF83 gene which involves in transcriptional regulation in the cell. A total of 9 hub genes that were differentially over-expressed significantly in cardiac tissue from RNA-Seq datasets, including EIF4GA which is related to viral myocarditis, HACD1, MYOM3, PTPN4, and NRBP1 are associated with muscular disorders, CELSR which play an essential role for planar cell polarity, SLC27A6 which is transporter involve in LCFA uptake process, SCMH1 involves in negative regulation of gene expression, and DCAF11 encodes WD repeat-containing protein, that involves in protein modification pathway .

We identified four hub genes from microarray and nine hub genes from RNA-Seq datasets using weighted gene co-expression networks analysis. We propose that these hub genes play an essential role in DCM pathogenesis and disease progression and could be a useful tools as genetic markers for the disease. Further studies and validations of these hub genes are needed to confirm our findings and to improve our understanding of the disease process.

Acknowledgments

First, I would like to thank Dr. Khuder for his time and tremendous efforts to help me to complete my thesis that required a lot of effort and time to be done. Second I would like to thank Dr. David Allison, who exposed me to the field of bioinformatics and helped me to get accepted into the new field of knowledge and practice, which I was not aware of it. Third I would like to thank my committee members Dr. Blumenthal and Dr. and Dr. Tian, for their support and help. I devoted this work to my mother, Mrs. Zienab, and the soul of my father, Dr. Fadlalla. Also, I would like to thank my soulmate Zainab for her continuous encouragement and help during this period. Moreover, I would like to thank my sister Sahar and my brothers for their great support throughout the years.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION:

Cardiomyopathies are a heterogeneous group of disorders that affect the heart muscle[1]. According to American Heart Association, 2006 (AHA) scientific statement proposed a contemporary definition and classification of the cardiomyopathies, cardiomyopathy defines as "a heterogeneous group of diseases of the myocardium associated with mechanical and/or electrical dysfunction that usually (but not invariably) exhibit inappropriate ventricular hypertrophy or dilation and are due to a variety of causes and frequently are genetic in nature"[2]. Moreover, cardiomyopathies were classified according to anatomy and physiology into; dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), restricted cardiomyopathy (RCM), arrhythmogenic right ventricular cardiomyopathy/dysplasia (ARVC/D), stress-induced and unclassified cardiomyopathy[3]. Besides that, the 2006 AHA categorizes cardiomyopathies into two main groups: Primary cardiomyopathy (mainly affecting the heart), and secondary cardiomyopathy (affecting the heart and other organs). The primary cardiomyopathies further subdivided into genetic, mixed, and acquired. The genetic cardiomyopathies include HCM, ARVC/D, left ventricular noncompaction, Protein Kinase AMP-Activated Non-Catalytic Subunit Gamma 2 (PRKAG2) and Danon glycogen storage diseases, conduction defects, mitochondrial myopathies, and ion channel disorders. The Mixed cardiomyopathies include DCM and RCM. The last one has acquired cardiomyopathies which include myocarditis, stress-induced (takotsubo), peripartum, tachycardia-induced, and infants of insulin-dependent diabetic mothers.[2]

The disorder can be inherited or acquired and can be caused by acute coronary syndrome, hypertension, diabetes, alcohol use, metabolic disorders, infiltrative disease, and other causes, although the reason is frequently unknown. Cardiomyopathies can be silent for a long time or

manifested clinically as symptoms of heart failure, arrhythmias, valvular lesions, dizziness or fainting, and even sudden cardiac death. The treatment for cardiomyopathies, in general, includes lifestyle modification, medical therapy, surgery, and medical device implantation[1,4].

DCM is a progressive disorder of the heart muscle, characterized by ventricular wall enlarging and reduction in the contractility function of the heart in the absence of high pressure over follow and absence of the valvular pathology. The DCM is the third most frequent cause of heart failure in the United States after CAD and hypertension. Furthermore, it is the most common type of cardiomyopathies. DCM is also a primary indication for heart transplantation[2]. Dilated cardiomyopathy is more common in adult than children and more common in men than women. The estimated annual incidence of dilated cardiomyopathy in the United States is 2.4 - 8 cases per 100,000[5]. The prevalence is estimated to be 36 cases per 100,000 and accounts for 10,000 deaths and 46,000 hospital admission in the United States annually[6,7].

The exact mechanism of dilated cardiomyopathy development is unknown. However, several genes reported in literature associated with disease pathogenesis. In our study, we are trying to identify hub genes from different microarray and RNA-Seq datasets. Hub genes are a set of genes that are highly connected with other genes and are essential ones in disease process. We proposed that these hub genes play an essential role in the disease pathogenesis.

CHAPTER 2

Background:

2.1 Dilated cardiomyopathy:

Dilated cardiomyopathy (DCM) is a heart muscle disorder that leads to dilatation and impairment of one or both heart ventricles[3,4,8–10]. Patients with dilated cardiomyopathy can present with arrhythmias, systolic dysfunction that lead to overt heart failure, and even sudden cardiac death at a late stage.

2.2 Causes of dilated cardiomyopathy:

2.2.1 Idiopathic:

This is the most common cause of DCM, and this term is applied when no other etiology was identified. In a large cohort study by Felker *et al*., they estimated about 50% of patients (616 of the 1230 patients total) included in the study were idiopathic[11]. Among patients with idiopathic DCM, about 50% of them have a familiar disorder, the most common mode of inheritance is autosomal dominant. However, other patterns are reported, like autosomal recessive, X-lined, and mitochondrial.

2.2.2 Familiar:

During the past two decades, more than 30 genes associated with familial (FDCM). Most FDCM are transmitted through autosomal dominant mode, although other forms are reported too in the literature. In a large study by Mestroni *et al.* in Italy characterized the following subtypes of FDCM: autosomal dominant (56%), autosomal recessive (16%), X-linked FDCM, with different mutations of the dystrophin gene(10%), a novel form of autosomal dominant FDCM with

subclinical skeletal muscle disease (7.7%), FDCM with conduction defects (2.6%), and rare unclassifiable forms (7.7%)[12].

2.2.2.1 Common phenotypes of familial dilated cardiomyopathy (FDCM):

Autosomal dominant FDCM without conduction system disorders:

Up to date the most common genes involved are genes associated with sarcomere proteins[13], namely beta myosin heavy chain (MYH7), alpha myosin heavy chain (MYH6), cardiac troponin T (TNNT2), titin (TTN), alpha-tropomyosin (TPM1), and cardiac troponin C (TNNC1)[14–17]. MYH7 is a gene encoding beta myosin heavy chain that plays a vital role in sarcomere function that leads to early-onset ventricular dilation and diminished contractility and approximately responsible for 0.04 % of all DCM[13,18–20]. TNNT2 mutation of cardiac troponin T is responsible for 0.03% and usually associated with aggressive disease[13,19–21]. TTN, which is the largest human protein and the most common cause of FDCM responsible for 0.15 – 0.20 of an estimated fraction of DCM, moreover it implicated in 25% of familial and 18% of sporadic DCM cases[22–25]. In Herman et al. study, they identified 72 unique mutations (25 nonsense, 23 frameshift, 23 splicing, and one large tandem insertion) that altered full-length titin[23]. Interestingly TTN also implicated in peripartum cardiomyopathy, which belongs to genetic cardiomyopathy[26,27].

Autosomal dominant FDCM with conduction system disorders:

The two most common mutations associated with conduction system disease are SCN5A and LMNA. LMNA is one of the most common genes that caused DCM, and it encodes filament proteins lamin A and C, which together form a scaffold called the nuclear lamina, which provides a structural integrity to the nucleus and involved in chromatin structure and gene expression[28 29]. LMNA mutations associated with 5-10% of FDCM and 2-5 of sporadic

DCM[30,31]. Up to now, more than 160 mutations are found that implicated in the DCM pathogenesis[28]. LMNA DCM can present with different variants of conduction system disease from first-degree heart block to ventricular tachycardia and fibrillation. Moreover, the DCM can manifest clinically at any time of development of the conduction system disease[30,32–37]. On the other hand, SCN5A, which encodes the alpha subunit of the primary cardiac sodium channel $Na_v1.5$ is associated with conduction system disease but usually accompanies ventricular dysfunction in contrast to LMNA cardiomyopathy, which traditionally associated with preserved ventricular function. Sinoatrial (SA) node dysfunction and atrial arrhythmias are a common presentation of the conduction system disease here[38–40]. Moreover, genetic variants of SCN5A are involved in other conduction system disease like Brugada syndrome and long QT syndrome.

X-linked FDCM:

The dystrophin gene is the most common X-linked mutation that causes FDCM[41,42]. Mutations are more often associated with skeletal muscle disorders like Duchenne and Becker Muscular Dystrophy. In X-linked cardiomyopathy, the patient may have rational skeletal muscle dystrophin expression and an isolated absence of cardiac dystrophin[43]. Another X-linked disorder is Bartha syndrome, which is manifested as dilated cardiomyopathy, skeletal myopathy, short stature, and neutropenia. It is caused by a mutation in a gene G4.5 that encodes a protein called tafazzins[44].

Autosomal recessive FDCM:

Mutations that involved ALMS1 can present with Alstorm syndrome, which is manifested as DCM, hearing and ocular impairment, obesity, and diabetes[4546]. Other autosomal recessive genetic mutation that leads to FDCM is cardiac troponin I (TNNI3), through impairment of

myocardial contractility[47].Desminopathy is a group of disorder that transmitted by autosomal recessive mode, characterized by cardiac and skeletal muscle diseases involvement, and it is caused by mutations in desmin (DES) gene[48]. Desmin is a type III intermediate filament protein that integrates the sarcolemma, Z disk, and nuclear membrane in sarcomeres and regulates sarcomere architecture[49].

2.2.3 Ischemic disorders:

Coronary arterial disease (CAD) secondary to atherosclerosis is the most common cause of heart failure and ischemic cardiomyopathy in the United States[50]. Most patient with ischemic cardiomyopathy has known CAD. However, occult disease is not an uncommon cause of DCM. In clinical practice, ischemic cardiomyopathy is usually used for cardiac dysfunction related to myocardial ischemia[11].

2.2.4 Infiltrative diseases:

Which includes cardiac amyloidosis, sarcoidosis, and hemochromatosis. These disorders are leading to DCM and various degrees of conduction system disorders. Infiltrative diseases associated with the worst prognosis of cardiomyopathies[11].

2.2.5 Infectious diseases:

Various infectious organisms can lead to myocarditis and, subsequently, DCM. Viral myocarditis is one of the most common causes of DCM, and the most common viruses that implicated in pathogenesis are parvovirus B19, human herpesvirus 6, coxsackievirus B. Viruses usually lead to damage through two main mechanisms: direct cardiac cytotoxicity and by the development of autoimmune response[51]. Moreover, HIV myocarditis and DCM can also be caused by drug toxicity and secondary infections. Lyme disease also is another cause of infectious DCM, which

is caused by bacteria called *Borrelia burgdorferi* leads to cardiac involvement if the form of conduction system block and myocarditis[52]. Chagas disease is a protozoal infection caused by *Trypanosoma cruzi* and thought to be the leading cause of DCM in South and Central America[53].

2.2.6 Autoimmune disorders:

Several autoimmune antibodies associated with DCM; these autoantibodies are sometimes referred to as anti-heart antibodies (AHAs). These AHAs targeted different cardiac antigens, including Beta-1 adrenoceptor, Alpha-myosin heavy chain, Beta-myosin heavy chain, myosin light chain, and Troponin. One of the possible mechanisms for beta-1 adrenoceptor autoantibodies is by acting as an agonist, and that sustained activation leads to intracellular calcium imbalance, apoptosis, and DCM with heart failure[54]. The systemic lupus erythematosus (SLE) is another autoimmune disorder that leads to DCM.

2.2.7 Endocrinological disorders:

Hyperthyroidism and hypothyroidism are two endocrinological disorders that leads to DCM, and these thyroid hormones have an adrenergic effect that leads to increase heart rate and contractility, sustained effect of these hormones lead to impairment of ventricular contrition, diastolic relaxation, and increased cardiac output[55]. In pheochromocytoma, excess sympathomimetics hormone leads to direct myocardial injury and inflammation[56]. Rarely Cushing's syndrome and acromegaly can cause cardiac dysfunction and then DCM[57].

2.2.8 Toxins/ Medications:

Alcohol and cocaine are the two major toxins that cause DCM. Although the definitive pathogenesis of alcoholic cardiomyopathy is poorly understood, several mechanisms have been reported, one of them is a direct cytotoxic effect of ethanol and its metabolites that leads to

oxidative stress and apoptosis, another one is due to long term alcohol consumptions leads to decrease myocardial protein synthesis and increase protein catabolism[58–60]. Cocaine-induced cardiomyopathy is not well understood, but the various postulated mechanisms have been reported; cocaine can induce direct toxicity to cardiac muscles, cocaine can induce a hyper sympathetic state that leads to cardiac cell necrosis, and lastly, chronic ischemia or myocardial infarction (MI) from cocaine use can lead to cardiomyopathy[61–63].

Several drugs can cause DCM, the two main drugs that extensively reported are anthracycline-induced cardiomyopathy (doxorubicin, daunorubicin, idarubicin, epirubicin, and the anthraquinone mitoxantrone), and Trastuzumab which is a monoclonal antibody against HER/neu receptor that we used for breast cancer treatment. HER2 signaling pathway has a role in cardiac development and play an important role in preventing cardiac injury, and Trastuzumab cause cardiotoxicity through oxidative stress that lead to apoptosis and cell necrosis[64]. Anthracycline induces DCM through reactive oxygen species generation, DNA damage, apoptosis induction, and protein synthesis inhibition[65].

2.2.9 Nutritional deficiencies and electrolyte imbalance:

Selenium, thiamine, and carnitine deficiencies are associated with DCM. Selenium deficiency leads to a decrease in the activity of glutathione peroxide, which results in generation free radicals that are toxic to cardiac muscle[66]. On the other hand, thiamin deficiency leads to high cardiac output failure and, eventually, the development of DCM[67]. On the other hand, patients with end-stage kidney disease with hypocalcemia, hypophosphatemia, and uremia can presents with DCM through unknown mechanism[68].

2.2.10 Miscellaneous:

Peripartum cardiomyopathy, tachycardia-induced cardiomyopathy, obstructive sleep apnea, obesity, heatstroke, hypothermia, and radiation are all associated with DCM[69–71].

2.3 Diagnosis:

The initial diagnostic workup for a patient with DCM is echocardiogram, which reveals left ventricular dilatation >112% corrected to body surface area and age, with reduced ejection function (FS <25% and/or LVEF <45%)[72]. A stress echocardiogram can be used when the patient is not able to exercise or unavailability of the exercise stress test. Failure to an increase in LVEF from rest to peak stress by ≥5% or a percentage change from a baseline of ≥20% associated with poor prognosis[73]. Another diagnostic workup is by searching for underlying conditions. Cardiac magnetic resonance (CMR) is indicated for infiltrative disorders like amyloidosis, hemochromatosis, and sarcoidosis. Moreover, it can be needed for functional and viability assessment, like in myocarditis. Endomyocardial biopsy (EMB) is necessary for patients with clinical symptoms of DCM and when the initial diagnostic modalities failed to confirm the diagnosis[74]. Moreover, EMB should be performed when histological information will affect prognosis or guide specific treatment therapy. Although performing EMB is not routine in the diagnosis of DCM, it does not mean we do not need to be done, in the study by Yasuchika Takeishi and Akiomi Yoshihisa who retrospectively analyzed 378 patients with suspected DCM who underwent EMB, the diagnostic impact of EBM may be relatively high in a patient with hypertrophic cardiomyopathy than those with DCM[75]. Routine genetic testing is recommended only in familial DCM (≥2 affected family members), where it's diagnostic yield is 30-35%[76,77].

2.4 Transcriptomics technologies:

Microarray is an advanced transcriptomics profiling technology to examine and identify the gene expression profile of the sample. Recently RNA-Seq had emerged as a new alternative method for gene expression profiling. The main difference between the two technologies is that the microarray gives an indirect measure of gene expression through profiling predefined transcript/ gene through hybridization. In contrast, RNA-Seq provides a direct measure of gene expression through the full sequencing of the whole transcriptome[78,79]. Moreover, RNA-Seq can identify a novel transcript since it does not require a transcript specific probe and can detect a higher percentage of differentially expressed genes (DEGs)[80,81].

CHAPTER 3

Methods:

3.1 Identification of differentially expressed genes (DEGs)

Microarray and RNA-Seq data were obtained through the Biotechnology Gene Expression Omnibus (GEO dataset) (http://www.ncbi.nlm.nih.gov/geo/), two types of expression were used; expression profiling by the array and expression profiling by high throughput sequencing using the keywords ‘‘Dilated Cardiomyopathy”, “microarray”, and “RNA-seq”. Search results with organisms other than *Homo sapiens* were excluded, and only adult samples were included. Moreover, we used the PubMed database to search for relevant studies on Microarray and RNA-Seq gene expression in dilated cardiomyopathy.

Four microarray datasets were included in our study, A total of 78 microarray samples were included, 34 were controls, and 44 were patients with DCM. Two of the microarray datasets used Affymetrix Human Gene 1.0 ST and U133A Array (GSE42955, and GSE3585), other platforms that used is Agilent-012097 Human 1A Microarray (V2) G4110B for GSE 9800, and Human Unigene3.1 cDNA Array 37.5K v1.0 for GSE 3586. All DCM tissue samples obtained from myocardium at time of cardiac transplant. The list of studies is shown in table 1.

Table 1: showing the characteristic of microarray datasets involved in the study

Gene list	GSE	Authors	Tissue	Platform	Samples
1	GSE3585	Barth AS *et al*., 2006	Left ventricular myocardium	Affymetrix Human Genome U133 Array	5 Controls 7 Patient

2	GSE3586	Barth *et al.*, 2006	Septal myocardium	Human Unigene3.1 cDNA Array 37.5K v1.0	15 Controls 13 Patients
3	GSE9800	NA	Left ventricular myocardium	Agilent-012097 Human 1A Microarray (V2)	9 Controls 12 Patients
4	GSE42955	Molina-Navarro MM et al., 2013	Left ventricular myocardium	Affymetrix Human Gene 1.0 ST Array	5 Controls 12 Patients

The microarray dataset was analyzed using a GEO2R analyzer, which is freely available from NCBI (https://www.ncbi.nlm.nih.gov/geo/geo2r/)[82–84]. Linear models of microarray data (Limma) was used to identify the DEGs between DCM samples and control samples[83,84]. The DEGs were selected using a cut off $p < 0.05$. Then we filtered the probes without known gene symbols, and duplicates were removed.

In addition, three RNA-seq datasets were included. A total of 41 samples were included, 18 were controls, and 23 were patients with DCM. The list of studies is shown in table 2. We analyzed RNA-Seq datasets from left ventricular myocardium tissue for patients with DCM and controls. We identified about 10,000 differentially expressed genes (DEGs) from each dataset, then we selected the top 1000 significant DEGs for each dataset with a level of significance ($p < 0.05$) using student's T test.

Table 2: showing the characteristics of RNA-seq datasets involved in the study

Gene list	GSE	Authors	Tissue	Platform	Samples
1	GSE55296	Tarazón E *et al*., 2014	Left ventricular myocardium	AB 5500xl Genetic Analyzer	10 Controls 13 Patients
2	GSE65446	Gonzalez-Valdes I *et al.*, 2015	Left ventricular myocardium	Illumina Genome Analyzer IIx	4 Controls 6 Patients
3	GSE71613	Schiano C *et al*., 2017	Left ventricular myocardium	Illumina HiSeq 2000	4 Controls 4 Patients

The significantly differentially expressed genes (DEGs) were selected using a cut off $p < 0.05$. Then we filtered the probes without known gene symbols, and duplicates were removed. Finally, the significant genes from the different datasets were analyzed using the R 'WGCNA' package[85].

3.2 Construction of Weighted Gene Co-expression Network (WGCNA)

The significant DEGs were inputted into the WGCNA program. WGCNA generated an output of gene modules, which are a set of genes with topological overlaps. To explain the hierarchical clustering tree using a top-bottom approach, all modules start as one cluster and repeatedly splits as they move down the tree to form branches and leaves. In this case, the branches represent the gene modules, and the leaves signify genes. This method helps us to understand the relationship between these genes and how they interact within the disease[85].

A node represents a cluster of genes, and the adjacency to each node is calculated (usually a score of 1 and 0). The score is assigned based on a certain threshold, and if the connection between the nodes is above the threshold, it is scored 1, and 0 if the connection is below the threshold. Weighted gene co-expression analysis (WGCNA) is used to construct an adjacency matrix using these scores, soft power, and Pearson's correlation (PC) coefficients for gene pairs. Soft power ranges from 10 to 30, which depicts the mean connectivity of the gene network. The lower the soft power, the higher the mean connectivity of the network. Below is the formula for adjacency.

$$\text{Adjacency} = 0.5 \text{ X } (1 + \text{PC})^{\text{soft power}}$$

A scale-free topography is used to select soft thresholding. The adjacency matrix is converted to a topological overlap matrix (TOM) to reduce the impact of noise as much as possible. The principal component analysis was conducted for the co-expressed module to generate the module epigenomes. The genes with high network connectivity in each module selected and referred to as hub genes. An overview of the method used to identify the hub genes is shown in figure 1.

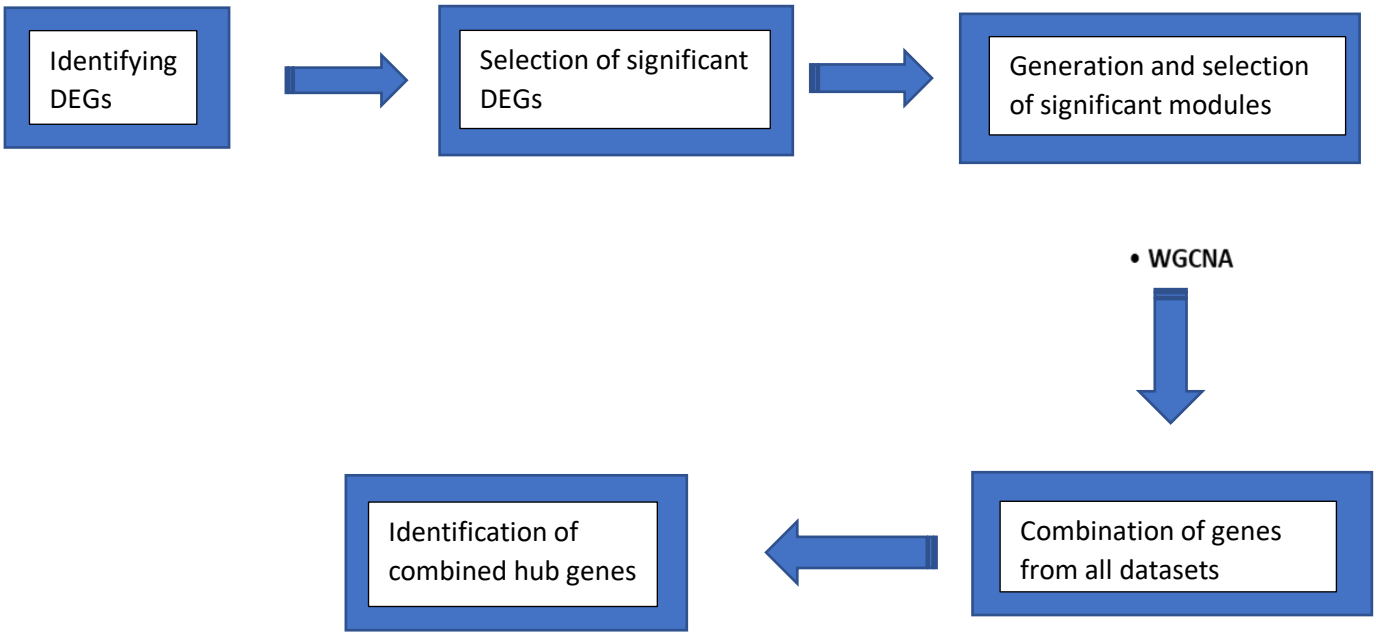

Figure 1: Method used to identify the hub genes using WGCNA.

3.3 Functional annotations modules:

GO analysis can annotate a gene with function involving cellular component (CC), molecular function (MC), and biological process (BP)[86]. To obtained more biological information about the genes created through WGCNA, The Database for Annotation, Visualization, and Integrated Discovery (DAVID version 6.8,https://david.ncifcrf.gov/) was used. Functional annotations charts and tables were used to identify particular GO biological processes, and KEGG pathways involvements, compared to the background list of the human genes, and to calculate enrichment scores of the GO biological process terms[87,88].

3.4 Protein-Protein interactions (PPI):

STRING database version 11.0 (https://string-db.org/) was used to identify the protein-protein interaction networks between observed genes. PPI denotes the generation process of protein complex combined with other protein molecules[89,90]. By logging to the STRING database website, the combined gene lists were submitted, followed by the selection of organism (*Homo sapiens*), then submission was made to get the results.

CHAPTER 4

RESULTS:

4.1 Differentially expressed genes (DEGs)

We identified more than 1000 DEGs for each microarray and RNA seq datasets. Only 1000 DEGs involved in our analysis. Top 20 DEGs for microarray datasets are shown in Table3A-D.

Table 3A: Top 20 DEGs with their gene title and p-value for the dataset GSE3585

Gene Symbol	Gene title	p. value	Adj p. value
PHLDA1	pleckstrin homology like domain family A member 1	6.46E-07	0.007
NPPB	natriuretic peptide B	2.35E-06	0.008
CFH	complement factor H	2.75E-06	0.008
TRMT5	tRNA methyltransferase 5	3.30E-06	0.008
IGFBP3	insulin like growth factor binding protein 3	3.50E-06	0.008
PHLDA1	pleckstrin homology like domain family A member 1	4.91E-06	0.008
ODC1	ornithine decarboxylase 1	5.67E-06	0.008
ETV5	ETS variant 5	6.37E-06	0.008
IDH2	isocitrate dehydrogenase (NADP(+)) 2, mitochondrial	8.01E-06	0.009
SLC30A1	solute carrier family 30 member 1	9.18E-06	0.010
PPDPF	pancreatic progenitor cell differentiation and proliferation factor	1.02E-05	0.010
CBFB	core-binding factor beta subunit	1.09E-05	0.010
NPPA	natriuretic peptide A	1.29E-05	0.010
KLHL3	kelch like family member 3	1.29E-05	0.010
C16orf45	chromosome 16 open reading frame 45	1.39E-05	0.010
IDH2	isocitrate dehydrogenase (NADP(+)) 2, mitochondrial	1.61E-05	0.010
HMGN2	high mobility group nucleosomal binding domain 2	1.80E-05	0.011
ASMTL	acetylserotonin O-methyltransferase-like	2.03E-05	0.012
H2AFZ	H2A histone family member Z	2.84E-05	0.016
CLK1	CDC like kinase 1	3.12E-05	0.017

Table 3B: Top 20 DEGs with their gene title and p-value for the dataset GSE3586

Gene symbol	Gene title	p. value	Adj p-value
DLAT	dihydrolipoamide S-acetyltransferase	2.15E-13	3.28E-09
GJA5	gap junction protein alpha 5	1.22E-12	1.24E-08
CCDC80///ANKH	coiled-coil domain containing 80///ANKH inorganic pyrophosphate transport regulator	4.59E-12	3.06E-08
CDC42EP3	CDC42 effector protein 3	5.07E-12	3.06E-08
SEC31A	SEC31 homolog A, COPII coat complex component	6.87E-12	3.06E-08
DPM1///FPGS	dolichyl-phosphate mannosyltransferase subunit 1, catalytic///folylpolyglutamate synthase	7.02E-12	3.06E-08
PAIP2	poly(A) binding protein interacting protein 2	9.90E-12	3.14E-08
PHYH	phytanoyl-CoA 2-hydroxylase	1.02E-11	3.14E-08
CSDE1	cold shock domain containing E1	1.03E-11	3.14E-08
HTRA1	HtrA serine peptidase 1	1.22E-11	3.39E-08
FGF14	fibroblast growth factor 14	1.38E-11	3.51E-08
CSDE1	cold shock domain containing E1	1.59E-11	3.73E-08
C16orf45	chromosome 16 open reading frame 45	3.26E-11	6.62E-08
IMMT	inner membrane mitochondrial protein	4.58E-11	8.73E-08
IKZF5///YWHAQ	IKAROS family zinc finger 5///tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein theta	5.37E-11	9.52E-08
SUCLA2	succinate-CoA ligase ADP-forming beta subunit	5.78E-11	9.52E-08
ATP6V1D	ATPase H+ transporting V1 subunit D	7.39E-11	1.10E-07
ZNF83	zinc finger protein 83	7.58E-11	1.10E-07
ECHDC1	ethylmalonyl-CoA decarboxylase 1	8.15E-11	1.13E-07
KIFAP3	kinesin associated protein 3	9.21E-11	1.18E-07

Table 3C: Top 20 DEGs with their gene title and p-value for the dataset the GSE9800

Gene symbol	Gene title	p. value	Adj p. value
NLRP10	NLR family pyrin domain containing 10	3.70E-08	0.001
ELN	elastin	4.68E-07	0.004
NDC80	NDC80, kinetochore complex component	5.95E-07	0.004

FOSB	FosB proto-oncogene, AP-1 transcription factor subunit	9.81E-07	0.005
IDH2	isocitrate dehydrogenase (NADP(+)) 2, mitochondrial	1.25E-06	0.006
CA14	carbonic anhydrase 14	1.63E-06	0.006
APOLD1	apolipoprotein L domain containing 1	1.75E-06	0.006
PLCE1	phospholipase C epsilon 1	2.32E-06	0.006
ESM1	endothelial cell specific molecule 1	2.51E-06	0.006
ESM1	endothelial cell specific molecule 1	2.84E-06	0.006
SNCA	synuclein alpha	3.73E-06	0.008
SCG2	secretogranin II	4.66E-06	0.009
FAP	fibroblast activation protein alpha	5.46E-06	0.009
ALDOC	aldolase, fructose-bisphosphate C	5.59E-06	0.009
LOXL2	lysyl oxidase like 2	6.08E-06	0.009
RABL6	RAB, member RAS oncogene family-like 6	6.57E-06	0.009
PMVK	phosphomevalonate kinase	7.47E-06	0.010
DDX54	DEAD-box helicase 54	7.94E-06	0.010
AHSA2	AHA1, activator of heat shock 90kDa protein ATPase homolog 2 (yeast)	1.17E-05	0.013
FBLIM1	filamin binding LIM protein 1	1.20E-05	0.013

Table 3D: Top 20 DEGs with their gene title and p-value for dataset the GSE42955

Gene symbol	Gene title	p. value	Adj p. value
ETS2	ETS proto-oncogene 2, transcription factor	1.27E-06	0.030
SCN2B	sodium voltage-gated channel beta subunit 2	2.35E-06	0.030
AASS	aminoadipate-semialdehyde synthase	2.73E-06	0.030
ELOVL7	ELOVL fatty acid elongase 7	5.18E-06	0.041
GABPB2	GA binding protein transcription factor beta subunit 2	6.16E-06	0.041
PTP4A3	protein tyrosine phosphatase type IVA, member 3	9.02E-06	0.043
PITPNM2	phosphatidylinositol transfer protein membrane associated 2	9.74E-06	0.043
E2F8	E2F transcription factor 8	1.03E-05	0.043
ACKR1	atypical chemokine receptor 1 (Duffy blood group)	1.18E-05	0.043
PER3	period circadian clock 3	1.38E-05	0.046
TRIM8	tripartite motif containing 8	1.59E-05	0.048
SELM///SELM	selenoprotein M///selenoprotein M	1.93E-05	0.050
RHOJ	ras homolog family member J	2.05E-05	0.050
LPCAT4	lysophosphatidylcholine acyltransferase 4	2.10E-05	0.050
TP53INP2	tumor protein p53 inducible nuclear protein 2	2.45E-05	0.050
HMGB2	high mobility group box 2	2.50E-05	0.050

SLC41A1	solute carrier family 41 member 1	3.05E-05	0.050
SNRPB	small nuclear ribonucleoprotein polypeptides B and B1	3.13E-05	0.060
TIMP4	TIMP metallopeptidase inhibitor 4	3.63E-05	0.063
PRKD1	protein kinase D1	3.98E-05	0.064

WGCNA was applied to each microarray datasets from the heart tissue included in the analysis. Overall, 15 gene modules were identified to be significant (p-value <0.05) from a total of 51 modules generated. Figures 2A-D show the modules created from each data set, with the corresponding p-values. They are color-coded to represent their expression levels, with red (1 on the scale) representing overexpression and blue (-1 on the scale) indicating under-expressed genes.

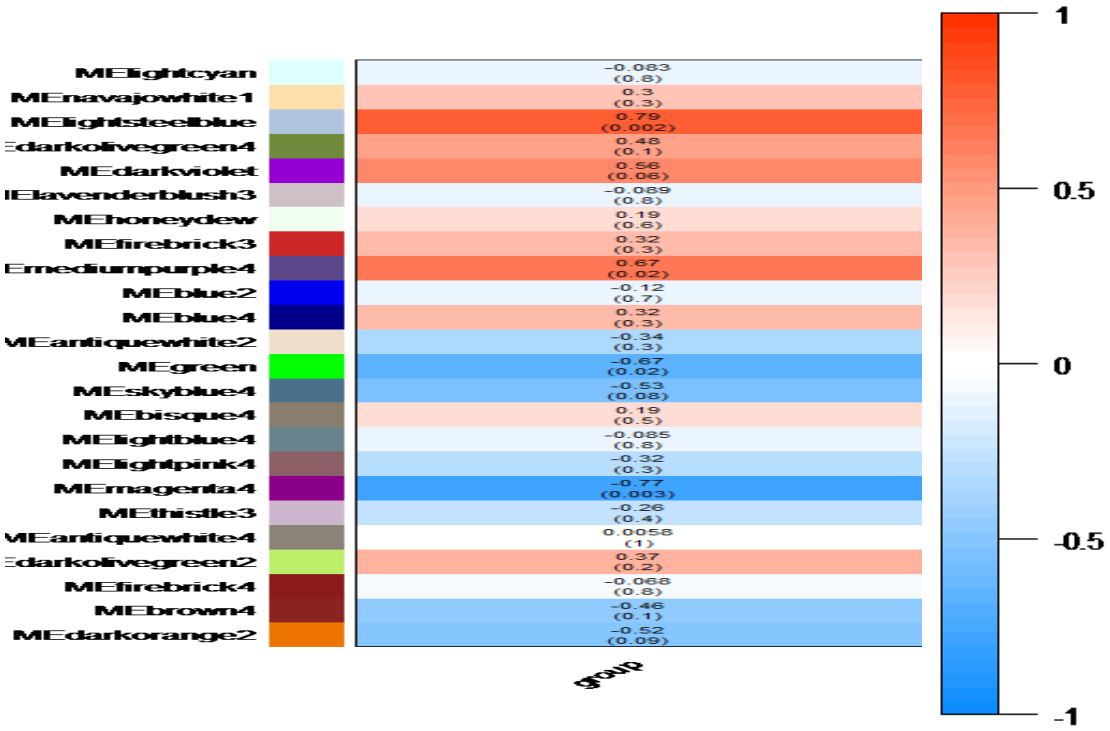

Figure 2A: The modules generated form WGCNA for the GSE3585 with four significant modules (p <0.05).

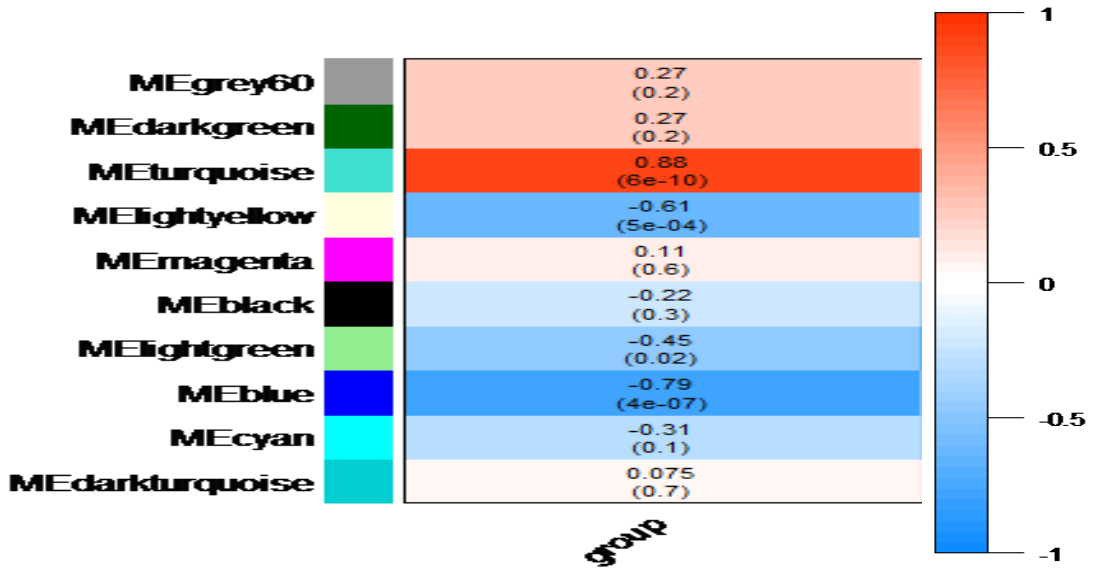

Figure 2B: The modules generated form WGCNA for the GSE3586 with four significant modules ($p < 0.05$)

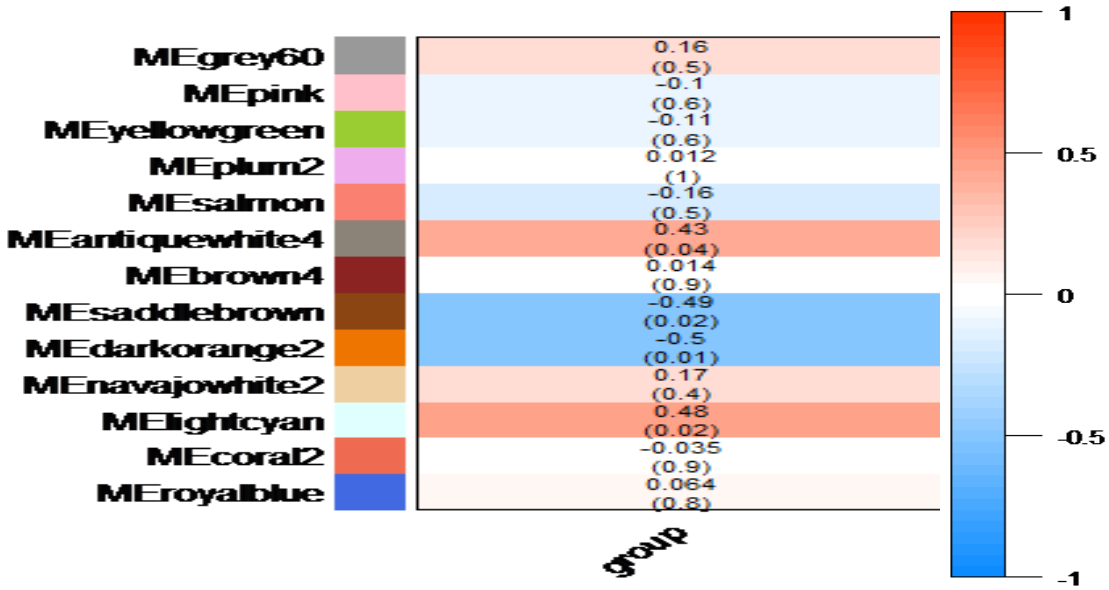

Figure 2C: The modules generated form WGCNA for the GSE9800 with four significant modules ($p < 0.05$)

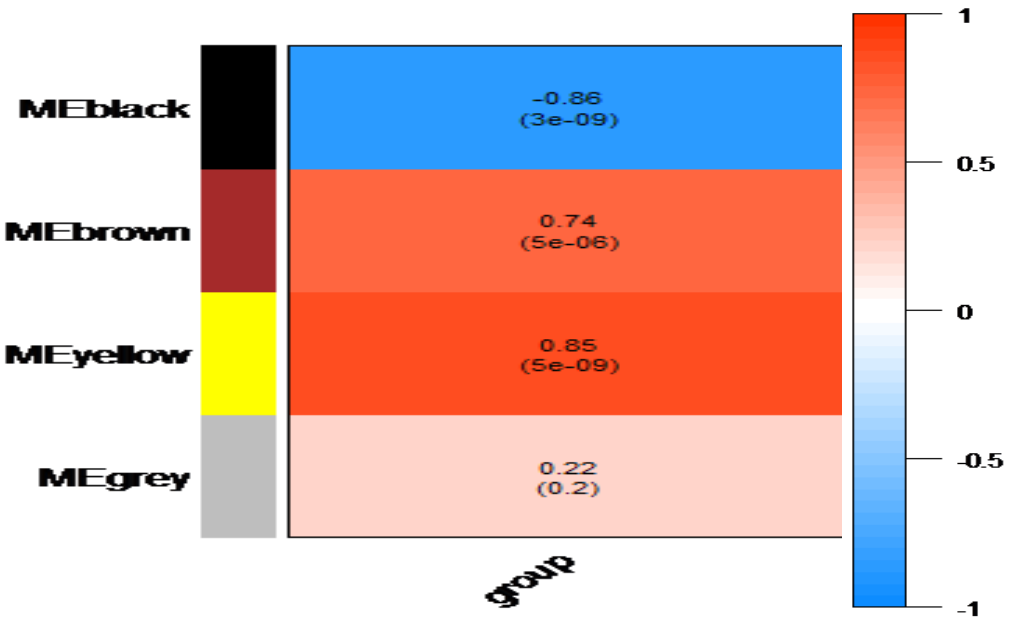

Figure 2D: The modules generated form WGCNA for the GSE42995 with three significant modules ($p < 0.05$).

Moreover, WGCNA was applied to each RNA-seq datasets from the heart tissue included in the analysis. Overall, three gene modules were identified to be significant (p-value <0.05) from a total of 15 modules generated. Figures 3A-C show the modules made from each RNA-seq dataset, with the corresponding p-values.

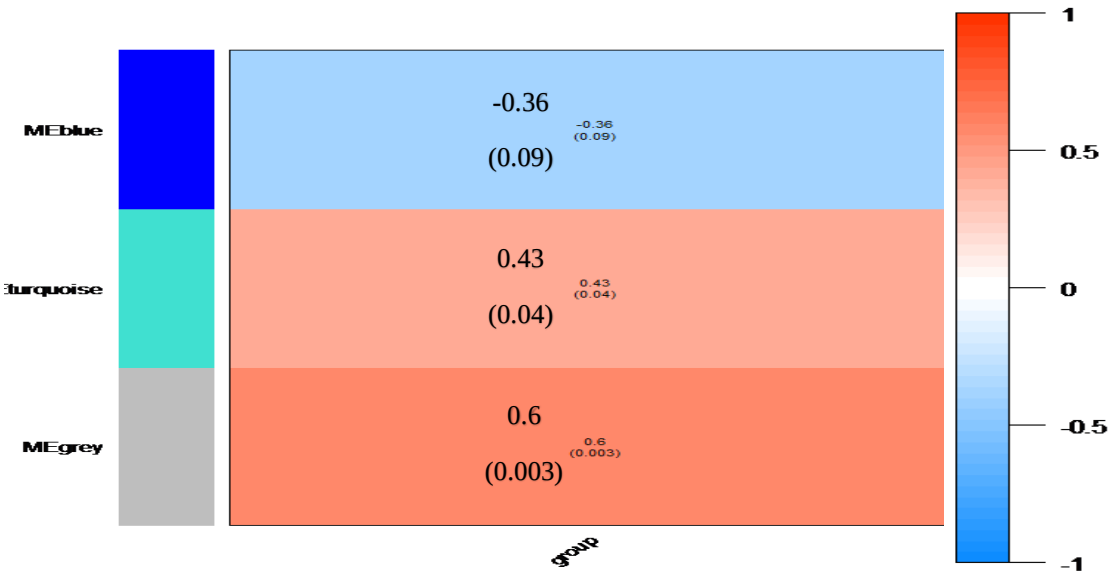

Figure 3A: The modules generated form WGCNA for the GSE55296 with two significant modules ($p < 0.05$)

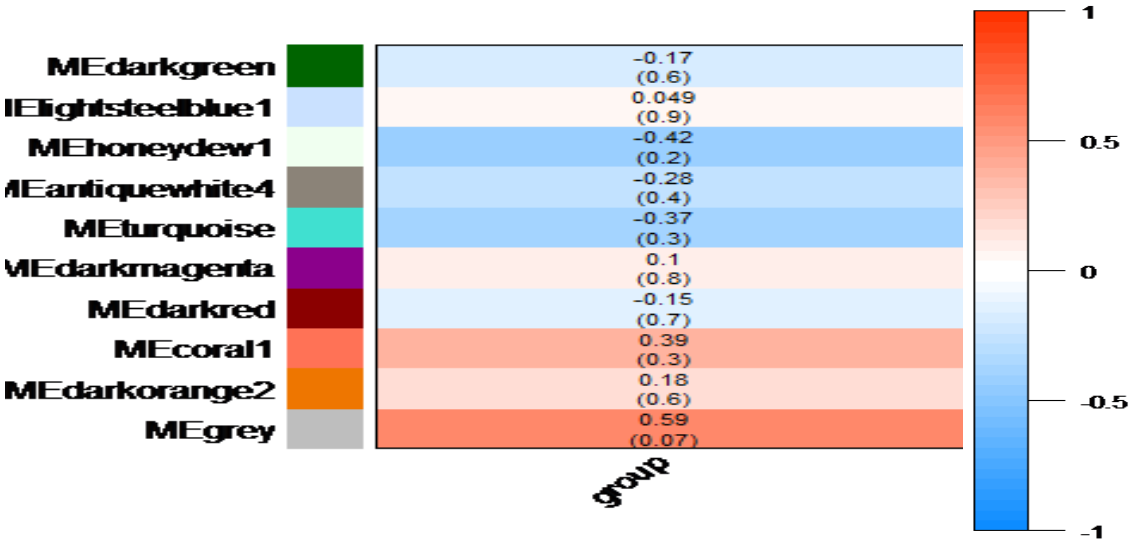

Figure 3B: The modules generated form WGCNA for the GSE65446 with no significant modules ($p < 0.05$)

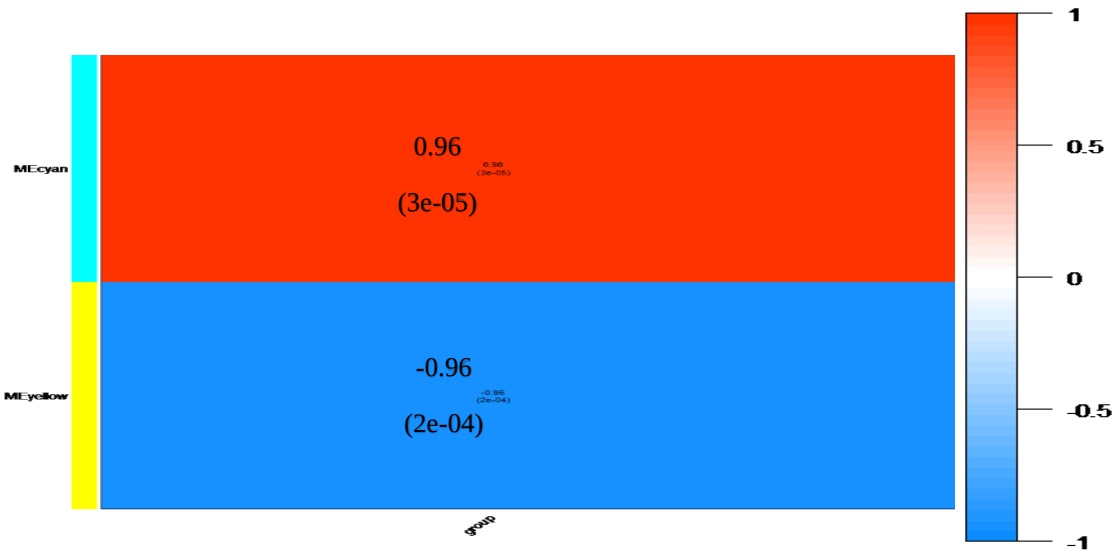

Figure 3C: The modules generated form WGCNA for the GSE71613 with two significant modules ($p < 0.05$)

4.2 Identification of hub gens:

The genes identified from modules further subjected to a cutoff point of $p < 0.05$ to determine their statistical significance. For microarray datasets, we used GSE3585, GSE3586, GSE9800, and GSE42955, all of them from cardiac tissue, and the significant genes from each set were matched for consensus genes. Hub genes were generated by combing the four datasets with significant modules and genes. A total of four hub genes that were differentially expressed in cardiac tissue, including AP3M2 (over-expression), ECM2 (over-expression), ERBB2 (under-expression), and ZNF83 (over-expression). The p-values for the hub genes are reported in Table 5.

Table 5: The Hub genes generated by WGCNA and their p-values across the four microarray datasets.

Gene symbol	p. value 4	p. value 3	p. value 2	p. value 1	Mean p-value	SD
AP3M2	0.0029	6.14E-08	0.0048	0.0006	0.0020	0.0022
ECM2	0.0096	3.81E-05	0.0046	0.0024	0.0041	0.0040
ERBB2	0.0189	5.76E-07	3.61E-06	0.0001	0.0046	0.0090
ZNF83	0.0250	2.49E-09	0.0009	0.0028	0.0070	0.0118

For RNA-seq datasets, we used GSE55296, GSE65446, and GSE71613, all of them from cardiac tissue, all of them were significant genes from each set were matched for consensus genes. Hub genes were generated by combing the two RNA-seq datasets, which contained statically significant genes. Two statistically significant modules (turquoise and grey) were selected from GSE55296 based on statistical significant of the modules and genes. A total of 9 hub genes that were differentially over-expressed significantly in cardiac tissue, including: CELSR1 (p-value 0.011), DCAF11 (p-value 0.046), EIF4G1 (p value0.045), HACD1 (p-value 0.037), MYOM3 (p-value 0.012), NRBP1 (p-value 0.009), PTPN4 (p-value 0.022), SCMH1 (p-value 3.80E-05), and SLC27A6 (p-value 0.047).

4.3 Gene-annotation enrichment analysis of hub genes and protein-protein interactions:

Gene ontology enrichment analysis was performed using The Database for Annotation, Visualization, and Integrated Discovery (DAVID) to identify the cellular component, molecular function, and biological process[87]. From functional annotation charts and tables, the biological

process of each gene was identified; AP3M2 gene which involves in protein trafficking to lysosomes and specialized organelles (GO:0006886), ECM2 gene which involves in extracellular matrix protein function (GO:0030198), and positive regulation of cell-substrate adhesion (GO:0010811), ZNF83 gene which involves in transcriptional regulation in the cell (GO:0006355), and ERBB2 gene which involves positive regulation of protein phosphorylation (GO:0001934) and positive regulation of cell adhesion (GO:0045785). 4 KEGG pathways were identified Lysosome (hsa04142), ErbB signaling pathway(hsa04012), Calcium signaling pathway (hsa04020), and HIF-1 signaling pathway (hsa04066). Protein-Protein Interaction (PPI) from the string database shows 17 nodes, 14 edges, and PPI enrichment p-value of 0.22 (Figure 4).

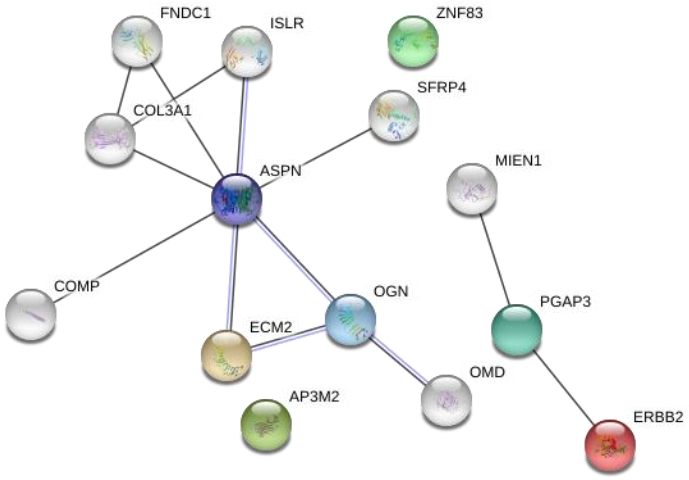

Figure 4: PPI network with current interaction generated from the sting database. The circle represents the gene, and the line represents the PPI between genes.

For RNA- seq datasets; from DAVOD's functional annotations tables and charts, biological processes were identified; MYOM3 which involves in immunoglobulin and fibronectin structures, NRBP1 involves in protein phosphorylation process (GO:0006468), HDAC1 involves

negative regulation of transcription by RNA polymerase II (GO:0000122), and chromatin organization (GO:0006325), DCAF11 encodes WD repeat-containing protein, that involves in protein modification process, PTPN4 involves in protein dephosphorylation process (GO:0006470), CERSR1 involves in cell adhesion (GO:0007155) and G protein-coupled receptor signaling pathway(GO:0007186), and epithelium development (GO:0060429), EIF4G1 involves in regulation of translation (GO:0006417), and regulation of cellular protein metabolic process (GO:0032268), SCMH1 involves in negative regulation of gene expression (GO:0010629), and SLC27A6 involves in the very-long-chain fatty acid metabolic process (GO:0000038). 2 KEGG pathways were identified; SLC27A6 with PRAP signaling pathway (hsa03320), and EIF4G1 with viral myocarditis (hsa05416). PPI network generated form the string database shows 21 nodes, 64 edges, and PPI enrichment p-value of 1.82e-14 (Figure 5).

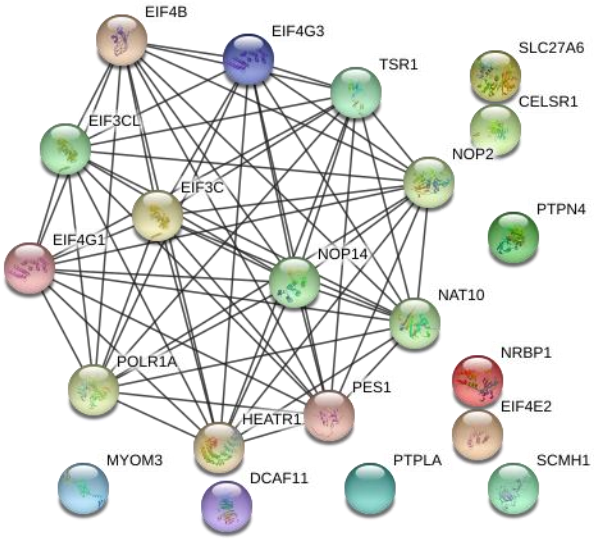

Figure 5: PPI network with current interaction generated from the string database. The circle represents the gene, and the line represents the PPI between genes.

CHAPTER 5

DISCUSSION:

Identification of genetic biomarkers for DCM is a very challenging process given multiple various causes and molecular mechanisms associated with disease development and pathogenesis. The pathogenesis of DCM varies from infectious, metabolic, infiltrative, endourological, cardiotoxin to autoimmune, and genetic. Moreover, the different molecular and pathological mechanisms involved in DCM, like inflammation, stress, myocardial cell injury, oxidative stress, matrix protein remodeling, and neurohormones involvement[91]. Although of these challenges present, recent advances in high throughput 'omics' technologies; transcriptomics, proteomics, and metabolomics, enable quantitative measurements of expression or abundance of biological molecules in a whole biological system, and the rapid expansion in the field of molecular genetics and biomarkers will give us a better understating about the mechanism of the disease[92].

Microarray analysis is one of the transcriptomics technologies used to measure gene expression profiling of thousands of genes at once. Using this technique, we can identify several genes that are statistically significant among the four datasets involved in our study. By comparing differentially expressed genes in DCM and healthy controls, we able to generate four hub genes AP3M2, ECM2, ERBB2, and ZNF83 that may play an important role in the pathogenesis of DCM. One of these genes is ECM2 which involves in extracellular matrix protein function, this gene encoding an extracellular matrix protein expressed predominantly in adipose and female-specific tissues and its chromosomal localization to 9q22.3[93]. In the study by Wang *et al.* demonstrated that mice lacking laminin alpha four which is an important component of

extracellular matrix (ECM) have developed cardiomyopathy[94]. The ErbB2 encodes a receptor tyrosine kinase, which is overexpressed in human tumors. Trastuzumab is a drug used to treat the patient with breast cancer, and dilated cardiomyopathy is a well-known side effect of this medication. In the study by Ozcelik *et al.* demonstrated that mice lacking ErbB2 developed severe dilated cardiomyopathy, they suggested signaling from the ErbB2 receptor is essential for the development of normal adult heart function[95]. Another study by Crone *et al.* postulated the same finding of ErbB2 signaling in cardiomyocyte is crucial to prevent DCM[96]. ZNF83 encodes Zinc finger protein 83, which involves in transcriptional regulation. Zinc finger protein is one of the most abundant proteins the body and associated with the regulation of several cellular processes. The overexpression of zinc finger protein GATA4 showed to prevent cardiac myocyte apoptosis induced by anthracyclines drugs, which are one of the main medications associated with cardiomyopathy[97]. Moreover, data reported the zinc finger protein GATA4 plays an important role in gene transcription associated with cardiac muscle hypertrophy, which is one of the earliest steps of heart failure development[98].

RNA-Seq, on the other hand, is an advanced sequencing technology that uses next-generation sequencing (NGS) to identify presence and quantity of RNA in the biological sample at a specific time. Our study revealed nine statistically significant hub genes ; EIF4G1 encodes eukaryotic translation initiation factor 4 gamma 1. Up to now, two functional homologs are identified EIG4GI and EIF4GII, which both form the EIF4G, which plays an important role in mammalian translation process[99]. Viral myocarditis is one of the most common causes of dilated cardiomyopathy, and Coxsackievirus is the most common virus that caused DCM. Several studies reported the role of Coxsackievirus B3 (CV-B3) protease 2A activity in DCM development[100–102]. In the study by Voigt *et al.* using CV-B3 myocarditis mouse mode, CV-B3

protease 2A-induced cleavage of EIF4G1 and dystrophin results in a complete shutdown of cap-dependent RNA translation, and favors cap-independent protein translation of viral proteins[102]. So, this finding suggested cleavage of EIFG1 is a critical step in viral replication, and inhibition of coxsackievirus-associated EIF4G1 cleavage could be a target for pharmacotherapy to prevent viral cardiomyopathy disease. Moreover, this gene EIF4G involved in alcoholic cardiomyopathy through the direct effect of alcohol-induced dephosphorylation of 4E-BP1 that leads to increase the affinity for EIF4E and as a result the association leads to decrease the affinity of EIF4E to EIF4G which leads to inhibition of cap-dependent RNA translation and protein synthesis[103,104]. For further understanding of physiological and biological processes, the KEGG pathway demonstrated significant enrichment of EIF4G1 with a viral myocarditis pathway (hsa05446 with a false discovery rate of 0.0166).

In our study, several genes related to skeletal muscle development and function are reported, HDACD1, PTPN4 , MYOM3, and NRBP1. HACD1 encodes 3-hydroxyacyl-CoA dehydratase 1, which contains a catalytic motif of protein tyrosine phosphatases (PTP) family, which play an essential role in skeletal muscle development and physiology by regulating cell proliferation, differentiation, growth, migration, and motility[105]. HACD1is mainly expressed in skeletal muscles and heart, and its mutations are associated with congenital myopathies[106]. PTPN4 encodes Protein Tyrosine Phosphatase Non-Receptor Type 4, which belongs to the PTP family. Although of the importance of PTP family in the normal T cell receptor signaling pathway, the experiential study reported animals lacking PTPN4 showed healthy T cell development and T cell receptor interactions, meaning that other family members of PTPs97 may gain the function of PTPN4.

Moreover, a recent study suggested that the PTPN4 is targeted gene for miR-181c-5p induced myocardial ischemia/ reperfusion injury through hypoxia/ reoxygenation cellular injury and activation of cell apoptosis pathway[107]. MYOM3 encodes myomesin 3, which links the intermediate filament cytoskeleton to the M-disk of myofibrils in skeletal muscle. Disruption of the M band affects sarcomere integrity and structure[108]. Up to now, most common genes associated with autosomal dominant DCM involve genes associated with sarcomere proteins. Mutations in the MYOM3 are associated with Alzheimer's Disease and Autosomal Recessive Limb-Girdle Muscular Dystrophy Type 2D[109]. A small clinical study by Shakeel *et al.* reported the importance of MYOM3 in the development of DCM[110]. Moreover, several skeletal muscle disorders like Duchenne and Becker Muscular Dystrophy are associated with X-linked FDCM. NRBP1 encodes Nuclear Receptor Binding Protein 1, and it is associated with Dengue Virus infection and Spinal Muscular Atrophy Type III[111]. In the Gil-Cayuela *et al.* study reported, the altered expression of NRBP2 (has sequence similarity to NRBP1) was related to disturbing ventricular function in a patient with DCM through alteration of intracellular self-renewal and recycling of the components leading to disruption of the autophagy degradation process[112].

The CELSR1 is a member of the cadherin family and encodes Cadherin EGF LAG seven-pass G-type receptor. CELSR1, FRIZZLED, and VANGL2 mediate extracellular adhesive interactions that required for the development and maintenance of planar cell polarity (PCP), which is core cell contact and signaling pathway essential to early development and tissue organization. Multiple CELSR mutations reported in human-associated with neural tube defects and cardiomyopathy[113114,115]. In another study by Yates *et al.* reported CELSR1 and VANGL2 are required for lung development, and mutations in them can lead to lung malformation, pulmonary hypertension, and idiopathic pulmonary fibrosis, which all can lead to cardiac

disorders[116]. SLC27A6 encodes Solute Carrier Family 27 Member 6, which is a member of fatty acid transport protein family (FATP) that involve in long-chain fatty acids (LCFA) uptake process. SLC27A6 gene expressed primarily in the heart muscle. Given these data, FATP6 could play an important role in lipid-induced cardiac diseases[117]. In animal model study observed that rats who developed myocardial infarction had a low level of FATP6 protein and reduced fatty acid oxidation[118].

5.1 LIMITATION:

One of the study limitations in our study is the small sample size; hence it is challenging to generalize among the whole population. Also, the genes generated from our study used a cardiac tissue sample only. Given the small sample size, there is a lack of validations of our DEGs and correlations between them. Further study is needed to get a better understanding of pathogenies of DCM.

5.2 CONCLUSION:

In conclusion, two significant hub genes that generated from combined microarray analysis using different bioinformatics tools; ECM2 which involves in extracellular matrix protein and Erb2 related to Trastuzumab-induced cardiomyopathy, besides seven genes generated from RNA-seq analysis; EIF4GA which is related to viral myocarditis, four genes associated with skeletal muscle disorders HACD1, MYOM3, PTPN4, and NRBP1, CELSR which plays a vital role for planar cell polarity, and SLC27A6 which is transporter involve in LCFA uptake process. These potential genetic biomarkers may serve an essential role in pathogenies and disease process of dilated cardiomyopathy. However, further experimental validations are required to confirm our findings.

References:

1. McKenna, W. J., Maron, B. J. & Thiene, G. Classification, Epidemiology, and Global Burden of Cardiomyopathies. *Circ. Res.* **121**, 722–730 (2017).

2. Maron, B. J. *et al.* Contemporary definitions and classification of the cardiomyopathies: an American Heart Association Scientific Statement from the Council on Clinical Cardiology, Heart Failure and Transplantation Committee; Quality of Care and Outcomes Research and Functio. *Circulation* **113**, 1807–16 (2006).

3. Richardson, P. *et al.* Report of the 1995 World Health Organization/International Society and Federation of Cardiology Task Force on the Definition and Classification of cardiomyopathies. *Circulation* **93**, 841–2 (1996).

4. Elliott, P. Cardiomyopathy. Diagnosis and management of dilated cardiomyopathy. *Heart* **84**, 106–12 (2000).

5. Manolio, T. A. *et al.* Prevalence and etiology of idiopathic dilated cardiomyopathy (summary of a National Heart, Lung, and Blood Institute workshop. *Am. J. Cardiol.* **69**, 1458–66 (1992).

6. Stergiopoulos, K. & Lima, F. V. Peripartum cardiomyopathy-diagnosis, management, and long term implications. *Trends Cardiovasc. Med.* **29**, 164–173 (2019).

7. Masarone, D. *et al.* Epidemiology and Clinical Aspects of Genetic Cardiomyopathies. *Heart Fail. Clin.* **14**, 119–128 (2018).

8. Report of the WHO/ISFC task force on the definition and classification of

cardiomyopathies. *Br. Heart J.* **44**, 672–3 (1980).

9. Dec, G. W. & Fuster, V. Idiopathic dilated cardiomyopathy. *N. Engl. J. Med.* **331**, 1564–75 (1994).

10. Luk, A., Ahn, E., Soor, G. S. & Butany, J. Dilated cardiomyopathy: a review. *J. Clin. Pathol.* **62**, 219–25 (2009).

11. Felker, G. M. *et al.* Underlying causes and long-term survival in patients with initially unexplained cardiomyopathy. *N. Engl. J. Med.* **342**, 1077–84 (2000).

12. Mestroni, L. *et al.* Familial dilated cardiomyopathy: evidence for genetic and phenotypic heterogeneity. Heart Muscle Disease Study Group. *J. Am. Coll. Cardiol.* **34**, 181–90 (1999).

13. Kamisago, M. *et al.* Mutations in sarcomere protein genes as a cause of dilated cardiomyopathy. *N. Engl. J. Med.* **343**, 1688–96 (2000).

14. Olson, T. M., Kishimoto, N. Y., Whitby, F. G. & Michels, V. V. Mutations that alter the surface charge of alpha-tropomyosin are associated with dilated cardiomyopathy. *J. Mol. Cell. Cardiol.* **33**, 723–32 (2001).

15. Li, D. *et al.* Novel cardiac troponin T mutation as a cause of familial dilated cardiomyopathy. *Circulation* **104**, 2188–93 (2001).

16. Mogensen, J. *et al.* Severe disease expression of cardiac troponin C and T mutations in patients with idiopathic dilated cardiomyopathy. *J. Am. Coll. Cardiol.* **44**, 2033–40 (2004).

17. Carniel, E. *et al.* Alpha-myosin heavy chain: a sarcomeric gene associated with dilated and hypertrophic phenotypes of cardiomyopathy. *Circulation* **112**, 54–9 (2005).

18. Daehmlow, S. *et al.* Novel mutations in sarcomeric protein genes in dilated cardiomyopathy. *Biochem. Biophys. Res. Commun.* **298**, 116–20 (2002).

19. Villard, E. *et al.* Mutation screening in dilated cardiomyopathy: prominent role of the beta myosin heavy chain gene. *Eur. Heart J.* **26**, 794–803 (2005).

20. Hershberger, R. E. *et al.* Coding sequence mutations identified in MYH7, TNNT2, SCN5A, CSRP3, LBD3, and TCAP from 313 patients with familial or idiopathic dilated cardiomyopathy. *Clin. Transl. Sci.* **1**, 21–6 (2008).

21. Hershberger, R. E. *et al.* Clinical and functional characterization of TNNT2 mutations identified in patients with dilated cardiomyopathy. *Circ. Cardiovasc. Genet.* **2**, 306–13 (2009).

22. Norton, N. *et al.* Exome sequencing and genome-wide linkage analysis in 17 families illustrate the complex contribution of TTN truncating variants to dilated cardiomyopathy. *Circ. Cardiovasc. Genet.* **6**, 144–53 (2013).

23. Herman, D. S. *et al.* Truncations of titin causing dilated cardiomyopathy. *N. Engl. J. Med.* **366**, 619–28 (2012).

24. Gerull, B. *et al.* Mutations of TTN, encoding the giant muscle filament titin, cause familial dilated cardiomyopathy. *Nat. Genet.* **30**, 201–4 (2002).

25. Schafer, S. *et al.* Titin-truncating variants affect heart function in disease cohorts and the

general population. *Nat. Genet.* **49**, 46–53 (2017).

26. Ware, J. S. *et al.* Shared Genetic Predisposition in Peripartum and Dilated Cardiomyopathies. *N. Engl. J. Med.* **374**, 233–41 (2016).

27. van Spaendonck-Zwarts, K. Y. *et al.* Titin gene mutations are common in families with both peripartum cardiomyopathy and dilated cardiomyopathy. *Eur. Heart J.* **35**, 2165–73 (2014).

28. Tesson, F. *et al.* Lamin A/C mutations in dilated cardiomyopathy. *Cardiol. J.* **21**, 331–42 (2014).

29. Carmosino, M. *et al.* Role of nuclear Lamin A/C in cardiomyocyte functions. *Biol. cell* **106**, 346–58 (2014).

30. Parks, S. B. *et al.* Lamin A/C mutation analysis in a cohort of 324 unrelated patients with idiopathic or familial dilated cardiomyopathy. *Am. Heart J.* **156**, 161–9 (2008).

31. Perrot, A. *et al.* Identification of mutational hot spots in LMNA encoding lamin A/C in patients with familial dilated cardiomyopathy. *Basic Res. Cardiol.* **104**, 90–9 (2009).

32. Peretto, G. *et al.* Cardiac and Neuromuscular Features of Patients with LMNA-Related Cardiomyopathy. *Ann. Intern. Med.* (2019). doi:10.7326/M18-2768

33. Arbustini, E. *et al.* Autosomal dominant dilated cardiomyopathy with atrioventricular block: a lamin A/C defect-related disease. *J. Am. Coll. Cardiol.* **39**, 981–90 (2002).

34. van Tintelen, J. P. *et al.* Severe myocardial fibrosis caused by a deletion of the 5' end of the lamin A/C gene. *J. Am. Coll. Cardiol.* **49**, 2430–9 (2007).

35. MacLeod, H. M., Culley, M. R., Huber, J. M. & McNally, E. M. Lamin A/C truncation in dilated cardiomyopathy with conduction disease. *BMC Med. Genet.* **4**, 4 (2003).

36. Fatkin, D. *et al.* Missense mutations in the rod domain of the lamin A/C gene as causes of dilated cardiomyopathy and conduction-system disease. *N. Engl. J. Med.* **341**, 1715–24 (1999).

37. Pasotti, M. *et al.* Long-term outcome and risk stratification in dilated cardiolaminopathies. *J. Am. Coll. Cardiol.* **52**, 1250–60 (2008).

38. Olson, T. M. *et al.* Sodium channel mutations and susceptibility to heart failure and atrial fibrillation. *JAMA* **293**, 447–54 (2005).

39. McNair, W. P. *et al.* SCN5A mutation associated with dilated cardiomyopathy, conduction disorder, and arrhythmia. *Circulation* **110**, 2163–7 (2004).

40. Olson, T. M. & Keating, M. T. Mapping a cardiomyopathy locus to chromosome 3p22-p25. *J. Clin. Invest.* **97**, 528–32 (1996).

41. Muntoni, F. *et al.* Brief report: deletion of the dystrophin muscle-promoter region associated with X-linked dilated cardiomyopathy. *N. Engl. J. Med.* **329**, 921–5 (1993).

42. Towbin, J. A. *et al.* X-linked dilated cardiomyopathy. Molecular genetic evidence of linkage to the Duchenne muscular dystrophy (dystrophin) gene at the Xp21 locus. *Circulation* **87**, 1854–65 (1993).

43. Nakamura, A. X-Linked Dilated Cardiomyopathy: A Cardiospecific Phenotype of Dystrophinopathy. *Pharmaceuticals (Basel).* **8**, 303–20 (2015).

44. Ichida, F. *et al.* Novel gene mutations in patients with left ventricular noncompaction or Barth syndrome. *Circulation* **103**, 1256–63 (2001).

45. Marshall, J. D. *et al.* New Alström syndrome phenotypes based on the evaluation of 182 cases. *Arch. Intern. Med.* **165**, 675–83 (2005).

46. Hearn, T. *et al.* Mutation of ALMS1, a large gene with a tandem repeat encoding 47 amino acids, causes Alström syndrome. *Nat. Genet.* **31**, 79–83 (2002).

47. Murphy, R. T. *et al.* Novel mutation in cardiac troponin I in recessive idiopathic dilated cardiomyopathy. *Lancet (London, England)* **363**, 371–2 (2004).

48. Taylor, M. R. G. *et al.* Prevalence of desmin mutations in dilated cardiomyopathy. *Circulation* **115**, 1244–51 (2007).

49. Brodehl, A., Gaertner-Rommel, A. & Milting, H. Molecular insights into cardiomyopathies associated with desmin (DES) mutations. *Biophys. Rev.* **10**, 983–1006 (2018).

50. Sekulic, M., Zacharias, M. & Medalion, B. Ischemic Cardiomyopathy and Heart Failure. *Circ. Heart Fail.* **12**, e006006 (2019).

51. Rose, N. R. Viral myocarditis. *Curr. Opin. Rheumatol.* **28**, 383–9 (2016).

52. Fish, A. E., Pride, Y. B. & Pinto, D. S. Lyme carditis. *Infect. Dis. Clin. North Am.* **22**, 275–88, vi (2008).

53. Schofield, C. J. & Dias, J. C. P. A cost-benefit analisys of chagas disease control. *Mem. Inst. Oswaldo Cruz* **86**, 285–295 (1991).

54. Jane-wit, D. *et al.* Beta 1-adrenergic receptor autoantibodies mediate dilated cardiomyopathy by agonistically inducing cardiomyocyte apoptosis. *Circulation* **116**, 399–410 (2007).

55. Klein, I. & Ojamaa, K. Thyroid hormone and the cardiovascular system. *N. Engl. J. Med.* **344**, 501–9 (2001).

56. Sardesai, S. H., Mourant, A. J., Sivathandon, Y., Farrow, R. & Gibbons, D. O. Phaeochromocytoma and catecholamine induced cardiomyopathy presenting as heart failure. *Br. Heart J.* **63**, 234–7 (1990).

57. Damjanovic, S. S. *et al.* High output heart failure in patients with newly diagnosed acromegaly. *Am. J. Med.* **112**, 610–6 (2002).

58. Preedy, V. R., Atkinson, L. M., Richardson, P. J. & Peters, T. J. Mechanisms of ethanol-induced cardiac damage. *Br. Heart J.* **69**, 197–200 (1993).

59. Patel, V. B., Why, H. J., Richardson, P. J. & Preedy, V. R. The effects of alcohol on the heart. *Adverse Drug React. Toxicol. Rev.* **16**, 15–43 (1997).

60. Piano, M. R. & Phillips, S. A. Alcoholic cardiomyopathy: pathophysiologic insights. *Cardiovasc. Toxicol.* **14**, 291–308 (2014).

61. Virmani, R., Robinowitz, M., Smialek, J. E. & Smyth, D. F. Cardiovascular effects of cocaine: an autopsy study of 40 patients. *Am. Heart J.* **115**, 1068–76 (1988).

62. Willens, H. J., Chakko, S. C. & Kessler, K. M. Cardiovascular manifestations of cocaine abuse. A case of recurrent dilated cardiomyopathy. *Chest* **106**, 594–600 (1994).

63. Fineschi, V. *et al.* Myocardial disarray: an architectural disorganization linked with adrenergic stress? *Int. J. Cardiol.* **99**, 277–82 (2005).

64. Cardinale, D. *et al.* Trastuzumab-Induced Cardiotoxicity: Clinical and Prognostic Implications of Troponin I Evaluation. *J. Clin. Oncol.* **28**, 3910–3916 (2010).

65. Tan, T. C., Neilan, T. G., Francis, S., Plana, J. C. & Scherrer-Crosbie, M. Anthracycline-Induced Cardiomyopathy in Adults. *Compr. Physiol.* **5**, 1517–40 (2015).

66. Observations on effect of sodium selenite in prevention of Keshan disease. *Chin. Med. J. (Engl).* **92**, 471–6 (1979).

67. Abelmann, W. H. & Lorell, B. H. The challenge of cardiomyopathy. *J. Am. Coll. Cardiol.* **13**, 1219–39 (1989).

68. Aoki, J. *et al.* Clinical and pathologic characteristics of dilated cardiomyopathy in hemodialysis patients. *Kidney Int.* **67**, 333–40 (2005).

69. Malone, S. *et al.* Obstructive sleep apnoea in patients with dilated cardiomyopathy: effects of continuous positive airway pressure. *Lancet (London, England)* **338**, 1480–4 (1991).

70. Grogan, M., Smith, H. C., Gersh, B. J. & Wood, D. L. Left ventricular dysfunction due to atrial fibrillation in patients initially believed to have idiopathic dilated cardiomyopathy. *Am. J. Cardiol.* **69**, 1570–3 (1992).

71. Sliwa, K. *et al.* Current state of knowledge on aetiology, diagnosis, management, and therapy of peripartum cardiomyopathy: a position statement from the Heart Failure Association of the European Society of Cardiology Working Group on peripartum

cardiomyopathy. *Eur. J. Heart Fail.* **12**, 767–78 (2010).

72. Mathew, T. *et al.* Diagnosis and assessment of dilated cardiomyopathy: a guideline protocol from the British Society of Echocardiography. *Echo Res. Pract.* **4**, G1–G13 (2017).

73. Neskovic, A. N. & Otasevic, P. Stress-echocardiography in idiopathic dilated cardiomyopathy: instructions for use. *Cardiovasc. Ultrasound* **3**, 3 (2005).

74. Cooper, L. T. *et al.* The role of endomyocardial biopsy in the management of cardiovascular disease: a scientific statement from the American Heart Association, the American College of Cardiology, and the European Society of Cardiology. Endorsed by the Heart Failure Society of. *J. Am. Coll. Cardiol.* **50**, 1914–31 (2007).

75. Takeishi, Y. & Yoshihisa, A. Distinct Roles of Myocardial Biopsy in Patients with Suspected Cardiomyopathy. *J. Card. Fail.* **22**, S154 (2016).

76. Hershberger, R. E. *et al.* Genetic Evaluation of Cardiomyopathy—A Heart Failure Society of America Practice Guideline. *J. Card. Fail.* **24**, 281–302 (2018).

77. Hershberger, R. E. *et al.* Genetic evaluation of cardiomyopathy: a clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med.* **20**, 899–909 (2018).

78. Rao, M. S. *et al.* Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. *Front. Genet.* **9**, (2019).

79. Merrick, B. A. *et al.* RNA-Seq Profiling Reveals Novel Hepatic Gene Expression Pattern in Aflatoxin B1 Treated Rats. *PLoS One* **8**, e61768 (2013).

80. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

81. Wang, C. *et al.* The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.* **32**, 926–32 (2014).

82. Davis, S. & Meltzer, P. S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **23**, 1846–7 (2007).

83. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry RA, Huber W, E. No Title Bioinformatics and Computational Biology Solutions using R and Bioconductor. *New York Springer;* 397–420 (2005).

84. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).

85. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

86. Gaudet, P., Škunca, N., Hu, J. C. & Dessimoz, C. Primer on the Gene Ontology. in 25–37 (2017). doi:10.1007/978-1-4939-3743-1_3

87. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

88. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths

toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).

89. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

90. Snel, B. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **28**, 3442–3444 (2000).

91. Dookhun, M. N., Sun, Y., Zou, H., Cao, X. & Lu, X. Classification of New Biomarkers of Dilated Cardiomyopathy Based on Pathogenesis—An Update. *Health (Irvine. Calif).* **10**, 300–312 (2018).

92. Bozkurt, B. *et al.* Current Diagnostic and Treatment Strategies for Specific Dilated Cardiomyopathies: A Scientific Statement From the American Heart Association. *Circulation* **134**, (2016).

93. Nishiu, J., Tanaka, T. & Nakamura, Y. Identification of a novel gene (ECM2) encoding a putative extracellular matrix protein expressed predominantly in adipose and female-specific tissues and its chromosomal localization to 9q22.3. *Genomics* **52**, 378–81 (1998).

94. Wang, J. *et al.* Cardiomyopathy Associated with Microcirculation Dysfunction in Laminin α4 Chain-deficient Mice. *J. Biol. Chem.* **281**, 213–220 (2006).

95. Ozcelik, C. *et al.* Conditional mutation of the ErbB2 (HER2) receptor in cardiomyocytes leads to dilated cardiomyopathy. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8880–5 (2002).

96. Crone, S. A. *et al.* ErbB2 is essential in the prevention of dilated cardiomyopathy. *Nat. Med.* **8**, 459–465 (2002).

97. Kim, Y. *et al.* Anthracycline-induced suppression of GATA-4 transcription factor: implication in the regulation of cardiac myocyte apoptosis. *Mol. Pharmacol.* **63**, 368–77 (2003).

98. Katanasaka, Y., Suzuki, H., Sunagawa, Y., Hasegawa, K. & Morimoto, T. Regulation of Cardiac Transcription Factor GATA4 by Post-Translational Modification in Cardiomyocyte Hypertrophy and Heart Failure. *Int. Heart J.* **57**, 672–675 (2016).

99. Gradi, A. *et al.* A novel functional human eukaryotic translation initiation factor 4G. *Mol. Cell. Biol.* **18**, 334–42 (1998).

100. Lim, B.-K. *et al.* Inhibition of Coxsackievirus-associated dystrophin cleavage prevents cardiomyopathy. *J. Clin. Invest.* **123**, 5146–51 (2013).

101. Bouin, A. *et al.* Enterovirus Persistence in Cardiac Cells of Patients With Idiopathic Dilated Cardiomyopathy Is Linked to 5' Terminal Genomic RNA-Deleted Viral Populations With Viral-Encoded Proteinase Activities. *Circulation* **139**, 2326–2338 (2019).

102. Voigt, A., Becher, M. P., Rahnefeld, A., Klingel, K. & Knobeloch, K.-P. ISGylation exerts a protective function in virus-induced dilated cardiomyopathy. *Eur. Heart J.* **34**, 3409–3409 (2013).

103. Lang, C. H. *et al.* Inhibition of muscle protein synthesis by alcohol is associated with

modulation of eIF2B and eIF4E. *Am. J. Physiol.* **277**, E268-76 (1999).

104. Steiner, J. L. & Lang, C. H. Alcoholic Cardiomyopathy: Disrupted Protein Balance and Impaired Cardiomyocyte Contractility. *Alcohol. Clin. Exp. Res.* **41**, 1392–1401 (2017).

105. Muhammad, E. *et al.* Congenital myopathy is caused by mutation of HACD1. *Hum. Mol. Genet.* **22**, 5229–5236 (2013).

106. Li, D., Gonzalez, O., Bachinski, L. L. & Roberts, R. Human protein tyrosine phosphatase-like gene: expression profile, genomic structure, and mutation analysis in families with ARVD. *Gene* **256**, 237–43 (2000).

107. Ge, L. *et al.* miR-181c-5p Exacerbates Hypoxia/Reoxygenation-Induced Cardiomyocyte Apoptosis via Targeting PTPN4. *Oxid. Med. Cell. Longev.* **2019**, 1–15 (2019).

108. Musa, H. *et al.* Targeted homozygous deletion of M-band titin in cardiomyocytes prevents sarcomere formation. *J. Cell Sci.* **119**, 4322–4331 (2006).

109. Fukuzawa, A. *et al.* Interactions with titin and myomesin target obscurin and obscurin-like 1 to the M-band - implications for hereditary myopathies. *J. Cell Sci.* **121**, 1841–1851 (2008).

110. Shakeel, M., Irfan, M. & Khan, I. A. Rare genetic mutations in Pakistani patients with dilated cardiomyopathy. *Gene* **673**, 134–139 (2018).

111. Qi, W. *et al.* C. elegans DAF-16/FOXO interacts with TGF-ß/BMP signaling to induce germline tumor formation via mTORC1 activation. *PLoS Genet.* **13**, e1006801 (2017).

112. Gil-Cayuela, C. *et al.* The altered expression of autophagy-related genes participates in

heart failure: NRBP2 and CALCOCO2 are associated with left ventricular dysfunction parameters in human dilated cardiomyopathy. *PLoS One* **14**, e0215818 (2019).

113. Stahley, S. Celsr1-mediated planar cell polarity: defining the adhesive interface and mechanisms of asymmetry.

114. Wang, Y. & Nathans, J. Tissue/planar cell polarity in vertebrates: new insights and new questions. *Development* **134**, 647–58 (2007).

115. Simons, M. & Mlodzik, M. Planar cell polarity signaling: from fly development to human disease. *Annu. Rev. Genet.* **42**, 517–40 (2008).

116. Yates, L. L. *et al.* The PCP genes Celsr1 and Vangl2 are required for normal lung branching morphogenesis. *Hum. Mol. Genet.* **19**, 2251–67 (2010).

117. Anderson, C. M. & Stahl, A. SLC27 fatty acid transport proteins. *Mol. Aspects Med.* **34**, 516–28

118. Heather, L. C. *et al.* Fatty acid transporter levels and palmitate oxidation rate correlate with ejection fraction in the infarcted rat heart. *Cardiovasc. Res.* **72**, 430–7 (2006).

APPENDIX

A1: Top 20 DEGs generated from WGCNA for RNA-Seq dataset GSE 55296

Gene symbol	Module color	GS group	P-value of GS group
UBC	blue	-0.67999	0.000357
TEKT3	blue	-0.60623	0.002167
MAP3K13	blue	-0.575	0.004102
FOXO3	blue	-0.56973	0.004541
RGS22	blue	-0.56179	0.005275
HHLA3	blue	-0.54209	0.007538
LTO1	blue	-0.54028	0.007781
TRARG1	blue	-0.54009	0.007807
RBP4	blue	-0.52855	0.009519
ARFIP1	blue	-0.52847	0.009532
CCL27	blue	-0.52075	0.010844
RAB20	blue	-0.51523	0.011869
VIP	blue	-0.51445	0.01202

CNOT6L	blue	-0.51431	0.012046
MAP9	blue	-0.51152	0.012602
DIPK1C	blue	-0.49974	0.015179
RUNDC3A	blue	-0.49805	0.015581
CCDC184	blue	-0.49517	0.016287
ZNF18	blue	-0.49008	0.0176
SREK1IP1	blue	-0.48081	0.020208

A2: Top 20 DEGs generated from WGCNA for RNA-Seq dataset GSE65446

Gene symbol	Module color	GS group	P-value of GS group
SSBP3	antiquewhite4	-0.8864	0.000634
CAMKMT	antiquewhite4	-0.87031	0.001055
NAV2	antiquewhite4	-0.8446	0.002106
BMPR1B	antiquewhite4	-0.79478	0.006007
STK24	antiquewhite4	-0.79158	0.006365
ELAVL1	antiquewhite4	-0.76948	0.009255
TBX5	antiquewhite4	-0.76665	0.009681
MARK4	antiquewhite4	-0.76172	0.010457
SLC7A2	antiquewhite4	-0.7543	0.011707
SAMD12	antiquewhite4	-0.74382	0.013647
GARNL3	antiquewhite4	-0.7438	0.01365
DENND2B	antiquewhite4	-0.74366	0.013678
CLASP2	antiquewhite4	-0.74281	0.013845
GK	antiquewhite4	-0.74251	0.013904

PDK4	antiquewhite4	-0.73586	0.015261
MYOCD	antiquewhite4	-0.73252	0.015977
UCP2	antiquewhite4	-0.73063	0.016392
SEMA3C	antiquewhite4	-0.72746	0.017105
PPM1L	antiquewhite4	-0.72703	0.017203

A3: Top 20 DEGs generated from WGCNA for RNA-Seq dataset GSE 71613

Gene symbol	Module color	GS group	P-value of GS group
NPAS3	cyan	0.98879	3.49E-06
KLHL3	cyan	0.97975	2.04E-05
TLE4	cyan	0.977761	2.70E-05
YPEL3	cyan	0.976502	3.19E-05
MYO1D	cyan	0.975407	3.65E-05
ZNF385B	cyan	0.971731	5.53E-05
FRZB	cyan	0.970667	6.17E-05
PDGFC	cyan	0.967122	8.67E-05
FMNL3	cyan	0.960945	0.000145
CFAP70	cyan	0.960558	0.000149
KMT2A	cyan	0.958575	0.000172
ECT2L	cyan	0.957271	0.000189
ETV5	cyan	0.955048	0.000219
ATP10D	cyan	0.954413	0.000229

ANGPTL7	cyan	0.952792	0.000254
PROX2	cyan	0.950542	0.000291
OGT	cyan	0.948695	0.000325
GAB2	cyan	0.948435	0.00033
SEC31A	cyan	0.948148	0.000335
XAF1	cyan	0.946984	0.000358

R codes for WGCNA

```
d1 <- read.csv("file.csv")

dis  = d1[!duplicated(d1$Gene.symbol),]

rm(d1)

library(WGCNA)

library(flashClust)

options(stringsAsFactors=FALSE)

allowWGCNAThreads()

rownames(dis) <- dis$Gene.symbol

datExpr= as.data.frame(t(dis[, -c(1)]))

names(datExpr)= dis$Gene.symbol

rownames(datExpr)=names(dis)[-c(1)]

dim(datExpr)

#-----Load trait data

traitData= read.csv("target.csv")

dim(traitData)

head(traitData)
```

```
names(traitData)

rownames(traitData) <- traitData$Sample

traitData$Sample <- NULL

datTraits= traitData[,-1]

# Call sample outliers

#-----Sample dendrogram and traits

A=adjacency(t(datExpr),type="signed")

A = as.matrix(A)

#-----Calculate whole network connectivity

k=as.numeric(apply(A,2,sum))-1



#-----Standardized connectivity

Z.k=scale(k)

thresholdZ.k=-3.5

outlierColor=ifelse(Z.k<thresholdZ.k,"red","black")

sampleTree = flashClust(as.dist(1-A), method = "average")

#-----Convert traits to colors
```

```
traitColors=data.frame(numbers2colors(datTraits,signed=TRUE))

dimnames(traitColors)[[2]]=paste(names(datTraits))

datColors=data.frame(outlier=outlierColor,traitColors)

save(datExpr0, datTraits, file="SamplesAndTraits1.RData")

options(stringsAsFactors = FALSE)

lnames= load(file="SamplesAndTraits1.RData")

lnames

dim(datExpr)

dim(datTraits)

#choosing a set of soft-thresholding powers

powers= c(seq(1,10,by=0.5), seq(from =12, to=40, by=2))

#call network topology analysis function

sft = pickSoftThreshold(datExpr, powerVector=powers, verbose =5,networkType="signed")

par(mfrow= c(1,2))

cex1=0.9
```

```
plot(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2], xlab= "Soft Threshold (power)", ylab="Scale Free Topology Model Fit, signed", type= "n", main= paste("Scale independence"))

text(sft$fitIndices[,1], -sign(sft$fitIndices[,3])*sft$fitIndices[,2], labels=powers, cex=cex1, col="red")

abline(h=0.90, col="red")

plot(sft$fitIndices[,1], sft$fitIndices[,5], xlab= "Soft Threshold (power)", ylab="Mean Connectivity", type="n", main = paste("Mean connectivity"))

text(sft$fitIndices[,1], sft$fitIndices[,5], labels=powers, cex=cex1, col="red")

softPower=10

adjacency=adjacency(datExpr0, power=softPower, type="signed")

TOM= TOMsimilarity(adjacency, TOMType="signed")

dissTOM= 1-TOM

geneTree= flashClust(as.dist(dissTOM), method="average")

plot(geneTree, xlab="", sub="", main= "Gene Clustering on TOM-based dissimilarity", labels= FALSE, hang=0.04)

minModuleSize=30
```

```
dynamicMods= cutreeDynamic(dendro= geneTree, distM= dissTOM, deepSplit=2,
pamRespectsDendro= FALSE, minClusterSize= minModuleSize)

table(dynamicMods)

dynamicColors= labels2colors(dynamicMods)

plotDendroAndColors(geneTree, dynamicColors, "Dynamic Tree Cut", dendroLabels= FALSE,
hang=0.03, addGuide= TRUE, guideHang= 0.05, main= "Gene dendrogram and module colors")

#-----Merge modules whose expression profiles are very similar

MEList= moduleEigengenes(datExpr0, colors= dynamicColors)

MEs= MEList$eigengenes

#Calculate dissimilarity of module eigenegenes

MEDiss= 1-cor(MEs)

#Cluster module eigengenes

METree= flashClust(as.dist(MEDiss), method= "average")

plot(METree, main= "Clustering of module eigengenes", xlab= "", sub= "")

MEDissThres= 0.42

abline(h=MEDissThres, col="red")

merge= mergeCloseModules(datExpr0, dynamicColors, cutHeight= MEDissThres, verbose =3)
```

```
mergedColors= merge$colors

mergedMEs= merge$newMEs

plotDendroAndColors(geneTree, cbind(dynamicColors, mergedColors), c("Dynamic Tree Cut",
"Merged dynamic"), dendroLabels= FALSE, hang=0.03, addGuide= TRUE, guideHang=0.05)

moduleColors= mergedColors

colorOrder= c("grey", standardColors(50))

moduleLabels= match(moduleColors, colorOrder)-1

MEs=mergedMEs

save(MEs, moduleLabels, moduleColors, geneTree, file=
"SamplesAndColors_thresh24merge42_signed1.RData")

datt=datExpr0

nGenes = ncol(datt);

nSamples = nrow(datt);

#-----Recalculate MEs with color labels

MEs0 = moduleEigengenes(datt, moduleColors)$eigengenes

MEs = orderMEs(MEs0)

#-----Correlations of genes with eigengenes
```

```
moduleGeneCor=cor(MEs,datt)

moduleGenePvalue = corPvalueStudent(moduleGeneCor, nSamples);

moduleTraitCor = cor(MEs, datTraits, use = "p");

moduleTraitPvalue = corPvalueStudent(moduleTraitCor, nSamples);

textMatrix = paste(signif(moduleTraitCor, 2), "\n(",

                   signif(moduleTraitPvalue, 1), ")", sep = "");

dim(textMatrix) = dim(moduleTraitCor)

par(mar = c(6, 8.5, 3, 3));

# Display the correlation values within a heatmap plot

labeledHeatmap(Matrix = moduleTraitCor,xLabels = names(datTraits),yLabels = names(MEs),

ySymbols = names(MEs),  colorLabels = FALSE, colors = blueWhiteRed(50), textMatrix =

textMatrix, setStdMargins = FALSE, cex.text = 0.7,   zlim = c(-1,1),  main = paste("Module-trait

relationships"))

######----------end----------#######

#----------Gene significance by Module membership scatterplots
```

```
whichTrait="group" #Replace this with the trait of interest



nGenes = ncol(datt);

nSamples = nrow(datt);

selTrait = as.data.frame(datTraits[,whichTrait]);



names(selTrait) = whichTrait

modNames = substring(names(MEs), 3)

geneModuleMembership = as.data.frame(signedKME(datt, MEs));

MMPvalue = as.data.frame(corPvalueStudent(as.matrix(geneModuleMembership), nSamples));

names(geneModuleMembership) = paste("MM", modNames, sep="");

names(MMPvalue) = paste("p.MM", modNames, sep="");

geneTraitSignificance = as.data.frame(cor(datt, selTrait, use = "p"));

GSPvalue = as.data.frame(corPvalueStudent(as.matrix(geneTraitSignificance), nSamples));

names(geneTraitSignificance) = paste("GS.", names(selTrait), sep="");

names(GSPvalue) = paste("p.GS.", names(selTrait), sep="");

par(mfrow=c(2,3))
```

```
counter=0
for(module in modNames[1:length(modNames)]){
  counter=counter+1
  if (counter>6) {
    par(mfrow=c(2,3))
    counter=1
  }
  column = match(module, modNames);
  moduleGenes = moduleColors==module;
  verboseScatterplot(abs(geneModuleMembership[moduleGenes, column]),
                     abs(geneTraitSignificance[moduleGenes, 1]),
                     xlab = paste(module,"module membership"),
                     ylab = paste("GS for", whichTrait),
                     col = module,mgp=c(2.3,1,0))
}
######--------------------end--------------------#######
```

```
#----------------------Eigengene heatmap

which.module="brown" #replace with module of interest

datME=MEs

datExpr=datt



ME=datME[, paste("ME",which.module, sep="")]

par(mfrow=c(2,1), mar=c(0.3, 5.5, 3, 2))

plotMat(t(scale(datExpr[,moduleColors==which.module ]) ),

      nrgcols=30,rlabels=F,rcols=which.module,

      main=which.module, cex.main=2)

par(mar=c(5, 4.2, 0, 0.7))

barplot(ME, col=which.module, main="", names.arg=c(row.names(datt)), cex.names=0.5,

cex.main=2,

      ylab="eigengene expression",xlab="sample")
```

```
#####################################################################


which.module="green" #replace with module of interest

datME=MEs

datExpr=datt


ME=datME[, paste("ME",which.module, sep="")]

par(mfrow=c(2,1), mar=c(0.3, 5.5, 3, 2))

plotMat(t(scale(datExpr[,moduleColors==which.module ]) ),

      nrgcols=30,rlabels=F,rcols=which.module,

      main=which.module, cex.main=2)

par(mar=c(5, 4.2, 0, 0.7))

barplot(ME, col=which.module, main="", names.arg=c(row.names(datt)), cex.names=0.5,

cex.main=2,

      ylab="eigengene expression",xlab="sample")
```

```
#########################################################################
```

```
#----------------------Eigengene heatmap
which.module="red" #replace with module of interest
datME=MEs
datExpr=datt

ME=datME[, paste("ME",which.module, sep="")]
par(mfrow=c(2,1), mar=c(0.3, 5.5, 3, 2))
plotMat(t(scale(datExpr[,moduleColors==which.module ]) ),
        nrgcols=30,rlabels=F,rcols=which.module,
        main=which.module, cex.main=2)
par(mar=c(5, 4.2, 0, 0.7))
```

```
barplot(ME, col=which.module, main="", names.arg=c(row.names(datt)), cex.names=0.5,
cex.main=2,
     ylab="eigengene expression",xlab="sample")



#######################################################################


which.module="grey" #replace with module of interest
datME=MEs
datExpr=datt


ME=datME[, paste("ME",which.module, sep="")]
par(mfrow=c(2,1), mar=c(0.3, 5.5, 3, 2))
plotMat(t(scale(datExpr[,moduleColors==which.module ]) ),
     nrgcols=30,rlabels=F,rcols=which.module,
     main=which.module, cex.main=2)
par(mar=c(5, 4.2, 0, 0.7))
```

```
barplot(ME, col=which.module, main="", names.arg=c(row.names(datt)), cex.names=0.5,
cex.main=2,
    ylab="eigengene expression",xlab="sample")
```