Triad of Coronary Heart Disease, Hyperlipidemia and Diabetes - A Genomics Approach

A dissertation submitted to Pondicherry University in partial fulfilment of the requirement for the award of the degree of

MASTER OF SCIENCE
IN
BIOINFORMATICS

BY
MEHAK CHOPRA
(Regd. No. 19378018)

Under the guidance of

Dr. A. MURALI

ASSISTANT PROFESSOR

CENTRE FOR BIOINFORMATICS

SCHOOL OF LIFE SCIENCES

PONDICHERRY UNIVERSITY

PUDUCHERRY – 605 014

JULY, 2021

BONAFIDE CERTIFICATE

This is to certify that this project entitled *“Triad of Coronary Heart Disease, Hyperlipidemia and Diabetes – A Genomics Approach*” is a bonafide record work done by **MEHAK CHOPRA**, REGD. NO. **19378018**, **MASTER OF SCIENCE IN BIOINFORMATICS** at **PONDICHERRY UNIVERSITY**, Pondicherry and has been carried out under direct supervision.

Dr. A. MURALI	**Dr. A. DINAKARA RAO**
Assistant Professor	**Professor & Centre Head**
Centre for Bioinformatics	**Centre for Bioinformatics**
Pondicherry University	**Pondicherry University**

Submitted for the M.Sc. Project held on **July 12, 2021** at Centre for Bioinformatics, Pondicherry University, Puducherry.

EXAMINERS

DECLARATION

I hereby declare that the dissertation entitled "*Triad of Coronary Heart Disease, Hyperlipidemia and Diabetes – A Genomics Approach*" is submitted by me, for the award of the degree of Master of Science in Bioinformatics to Pondicherry University is a record of bonafide research work carried out under the supervision of Dr. A. Murali, Assistant Professor, Centre for Bioinformatics, Pondicherry University.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or full, for the award of any other degree or diploma in this institute or of any other institute or university.

Place: Puducherry
Date**: July 12, 2021** **Mehak Chopra**

ACKNOWLEDGMENT

I owe my deepest gratitude to my guide and mentor **Dr. A. Murali, Assistant Professor, Centre for Bioinformatics, Pondicherry University** for providing me with the guidance, all the suggestions and encouragement and steering me in the right direction without which I would not have been able to complete this project.

I would also like to thank **Mr. Lukkani Laxman Kumar**, **Mr. Satya Ranjan, Miss. Jyotilipsa Mohanty** and **Miss. Ahana Roy** for supporting me throughout my M.Sc. project and their help, assistance in the successful completion of this project.

I extend my earnest gratitude to **Dr. A. Dinakara Rao**, Centre Head, Centre for Bioinformatics and all the faculty of Centre for bioinformatics for their constant support.

I thank almighty god, my parents, my family members and friends for all their support through the academic period and in project work.

INDEX

LIST OF FIGURES

LIST OF TABLES

1. Abstract

Cardiovascular disease, the major cause of premature disabilities in an ageing population and deaths accounts for 30-50% deaths in developed nations. One of the cardiovascular illnesses, Heart Failure, whose hospitalisation and fatality rates have risen steadily over the last 25 years. Many people who suffer heart failure also have coronary artery disease (CAD). Hyperlipidemia, like hyperglycemia, is hypothesised to be the cause of impeachment and is expected to cause impairments in a range of cell types as well as the development of diabetic complications. These ideas emphasise the progressive nature of Type 2 Diabetes Mellitus (T2DM) and accompanying cardiovascular risk, which poses unique issues at various phases of a diabetic's life. To reduce the high surgical risk in innumerable cardiovascular patients, ischemic disease is an area of active research. The study helps in identifying the Differentially Expressed Genes (DEGs) using the RNA-seq raw data of patients suffering from Coronary Heart Disease (CHD) and Hyperlipidemia (HYP). Top 10 hubgenes in each catgory were identified and gene set enrichment analysis provided the functional analysis of the genes DEG's involved. 13 and 18 genes in CHD & FCHD, HYP & FHYP were found to be intersecting, respectively. In Gene Hub analysis, CWC22 (CWC22 spliceosome associated protein homolog) and RPS5 (ribosomal protein S5) were found to be the common hubs. Gene Ontology Analysis helped in characterising the functions of the DEG's involved in CHD and HYP like Hydrolase activity, acting on ester bonds; Positive regulation of hydrolase activity; Response to extracellular stimulus and response to oxygen containing compound. EIF2AK3 (eukaryotic translation initiation factor 2 alpha kinase 3), F13A1 (coagulation factor XIII A chain) and RSU1 (Ras suppressor protein 1) were found to have common association in between Coronary Heart Disease, Hyperlipidemia and Diabetes using KEGG and Genome Wide Association Studies (GWAS) Catalog. The correlation was also observed using the gene-disease interaction networks and gene rank bar plots. Identified key hub genes and pathways shed light on the molecular mechanism that may contribute to the discovery of novel therapeutic targets and development of new strategies for the Coronary Heart Disease and Hyperlipidemia. Persons with T2DM are at such a high risk of having CHD, therefore, it's critical to uncover factors that may influence their beliefs of the danger of acquiring CHD. The analysis helps to uncover the association and finds the correlation between Coronary Heart Disease, Hyperlipidemia and Diabetes.

2. Introduction

2.1 Cardiovascular Disease:

Cardiovascular diseases, also known as Coronary Artery Disease, Coronary Microvascular Disease, Coronary Syndrome X, Ischemic Heart Disease, Nonobstructive Coronary Artery Disease, Obstructive Coronary Artery Disease the major cause of premature disabilities in an ageing population and deaths accounts (DarshanDoshi, Ori Ben-Yehuda, Machaon Bonafede et al., 2016) for 30-50% deaths in developed nations. Plaque, a waxy substance that forms inside the lining of major coronary arteries, is a common cause of coronary heart disease. The deposit might partially or completely obstruct blood flow in the heart's major arteries. Some varieties of this issue can be caused by an illness or injury that affects the way the heart's arteries work. Another kind of CHD is coronary microvascular disease. It is getting more normal in countries recently devastated by history or ongoing conditions, as expectations for everyday comforts improve (Handerson, 1996).

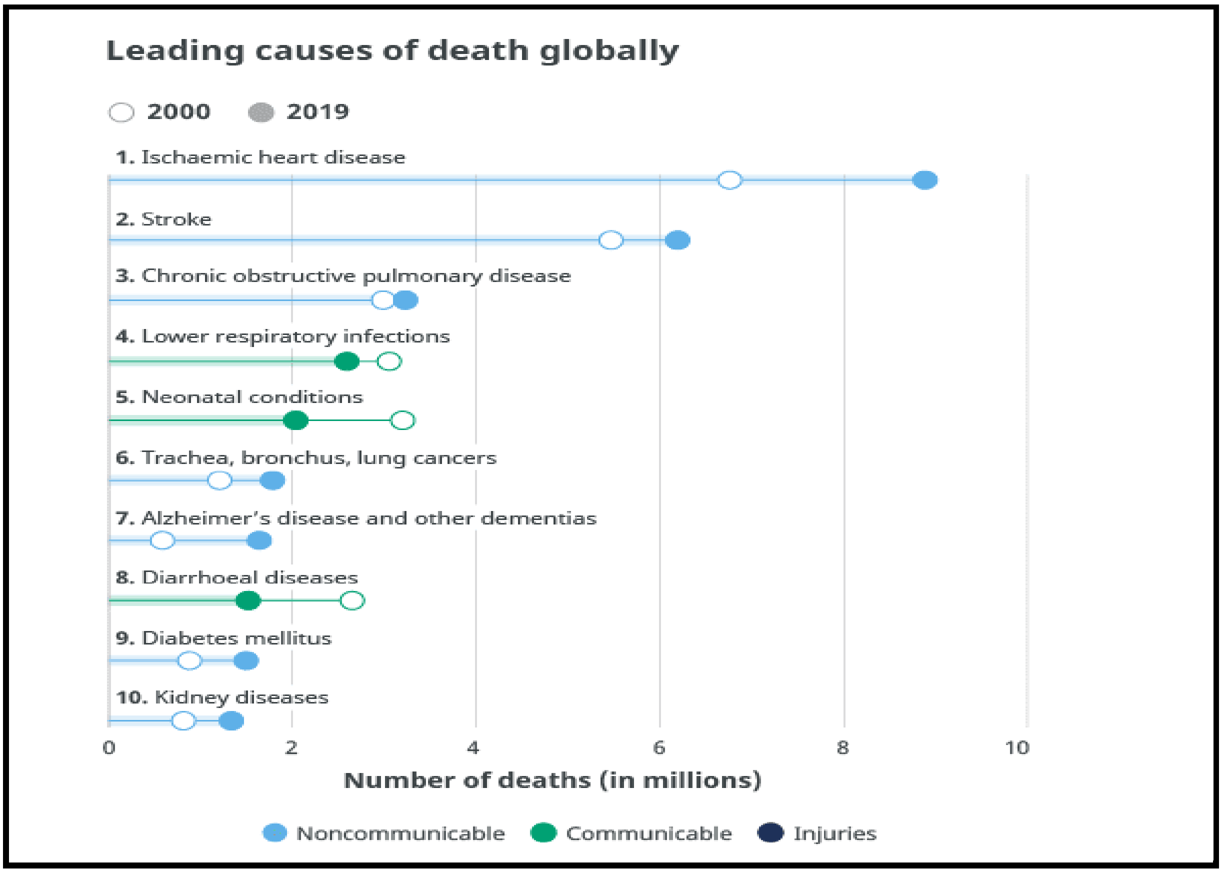


Fig. 1: *WHO Report* Shows the Leading Causes of Death Globally *(December, 2020)*

Heart Failure (HF) is a clinical syndrome caused by anatomical and functional myocardial abnormalities that induce ventricular filling and/or ejection dysfunction (Pagliaro et al., 2020). Hereditary inclination and natural variables like smoking history, poor diet and sedentary lifestyle lead to the growth of atherosclerotic plaques in the vessel dividers of the coronary veins bringing about diminished myocardial perfusion. To reduce the high surgical risk in innumerable cardiovascular patients, ischemic disease is an area of active research.

Insulin resistance, hyperlipidemia, hypertension, hyperglycemia, and obesity are common comorbidities in patients with ischemic heart disease. Individually, these comorbidities act as proatherogenic stimuli which include inflammation and alter platelet functions leading to abnormalities in the endothelial barrier which leads to increased vascular permeability, and together they form the metabolic syndrome disease. Inflammation and altered platelet function are caused by proatherogenic stimuli, which lead to anomalies in the endothelium barrier, which leads to increased vascular permeability. This inflammation causes cellular wall instability, platelet and fat-laden macrophage buildup, and the formation of an atherosclerotic plaque in the artery wall, as well as diminished perfusion to the heart (Brittany A Potz, Anshul B Parulkar, Ruhul M Abid, Neel R Sodha, 2018).

One of the cardiovascular illnesses, Heart Failure, whose hospitalisation and fatality rates have risen steadily over the last 25 years. Many people who suffer heart failure also have coronary artery disease (CAD). Not only in individuals with HF with reduced ejection fraction (HFrEF), but also in those with HF with intact ejection fraction, CAD is frequented. Epidemiological data also show that ischemic CAD is the most common cause of HF, rather than hypertension or valvular heart disease. This is significant not only because ischemic CAD is a potentially curable (or reversible) cause of HF, but also because the presence of CAD is linked to worsening long-term outcomes in a synergistic and independent manner. (DarshanDoshi, Ori Ben-Yehuda, Machaon Bonafede et al., 2016). Despite advancements in CAD diagnosis and treatment, gender disparities persist, with ischemic heart disease mortality in women being high. Women are more prone to experience anginal equivalents such as fatigue, dyspnea, indigestion, or jaw pain, but most individuals with ACS present with usual symptoms such as central chest pain or pressure (Perdoncin & Duvernoy, 2017). Small molecule, cytokine, endothelial progenitor cell, stem cell, gene, and mechanical therapies have all shown promise in promoting collateral blood vessel formation and thereby minimising myocardial ischemia (Brittany A Potz, Anshul B Parulkar, Ruhul M Abid, Neel R Sodha, 2018).

In terms of cardiovascular comorbidities in study of (DarshanDoshi, Ori Ben-Yehuda, Machaon Bonafede et al., 2016), 83.3 percent of patients had hypertension, while about half had hyperlipidemia, diabetes, or baseline CAD.

2.2 Hyperlipidemia and its association with Coronary Artery Disease:

Lipids are carried in the bloodstream by lipoprotein particles and Hyperlipidemia, a complex disorder, a condition in which serum lipid levels are abnormally high, caused by dietary disorders, obesity, genetic diseases such as familial hypercholesterolemia (FH) or other diseases such as diabetes (Yao et al., 2020). It involves a cholesterol imbalance in the blood, with low-density lipoprotein cholesterol (LDL-C) and high-density lipoprotein cholesterol (HDL-C). The amount of cholesterol in the body is regulated by LDL-C and HDL-C, and an imbalance might raise the risk of cardiovascular events such as myocardial infarction and stroke (Karr, 2017). Because extra lipids in the blood accumulate in the walls of arteries; hyperlipidemia is a risk factor for atherosclerosis. LDL-C is now well recognised as a key player in the course of atherosclerosis and related consequences. Cardiovascular events are responsible for a significant share of overall mortality. As a result, high LDL-C is the focus of the majority of hyperlipidaemia treatment efforts (Bułdak et al., 2019). When low-density lipoprotein (LDL) is oxidised, oxidised (ox) LDL is formed, which is a heterogeneous mixture of oxidised lipids and proteins. Hyperlipidemia is related to increased TLR4 expression on circulating monocytes, and OxLDL promotes TLR4 expression in macrophages. Studies have demonstrated that circulating monocytes from hyperlipidemic individuals have higher levels of tissue factor (TF) compared with healthy controls. Also, acute coronary syndrome patients have elevated levels of both circulating monocyte-derived microparticles (MPs) as well as TF+ MPs (Owens et al., 2014).

Microvascular function is regulated by a range of lipids, including triglycerides (TG) and total cholesterol (TC), as well as high and low density lipoproteins (HDL, LDL). Hypercholesterolemia reduces coronary blood flow reserve and capillary density, causes capillary endothelial cells to die, and causes left ventricular (LV) dysfunction. Hyperlipidemia patients are nearly twice as likely to suffer cardiovascular disease. Hypercholesterolemia is thought to affect the lipid bilayer of the membrane, the control of intracellular calcium ions, and the isoform expression patterns of myosin heavy chain, rendering the heart more vulnerable to external injury (Yao et al., 2020).

2.3 Hyperlipidemia, Coronary Heart Disease and Diabetes Mellitus:

Hyperglycemia is a symptom of diabetes mellitus, a set of metabolic illnesses characterised by insulin production, insulin action, or both. With 371 million cases estimated worldwide in 2012, and anticipated to climb to 552 million by 2030, Diabetes is a severe global public health burden. Cardiovascular disease, neuropathy, nephropathy, retinopathy, bunions, osteoporosis, Alzheimer's disease, and cancer are among the complications of diabetes mellitus. Out of two forms of Diabetes, in adults 90-95% of cases have been observed of Type-2 Diabetes, a combination of insulin resistance and impaired insulin resistance (Zhou et al., 2015). Because insulin deficiency inhibits 6-desaturase enzyme activity, the conversion of omega-6 polyunsaturated fatty acids to active metabolites is hindered in diabetes patients. Less active metabolites, which are important components of cell membrane structure, impact cellular processes. Hyperlipidemia, like hyperglycemia, is hypothesised to be the cause of impeachment and is expected to cause impairments in a range of cell types as well as the development of diabetic complications (Hulbert et al., 2005; Hagve, 1988).

In three earlier population-based investigations, hyperinsulinemia has been linked to an elevated incidence of coronary heart disease and cardiovascular disease death (Martín-Timón, 2014; Ginsberg, 2000; Lakka et al., 2000). Diabetic vascular disease increases the risk of coronary artery disease (CAD) and stroke by two-four times, and reduces the risk of heart failure by two-eight times. Patients with T2DM and no prior history of CAD had the same risk of cardiac events as those who have had a previous myocardial infarction, according to research. Visceral obesity, insulin resistance (IR), and variations in the levels of a variety of circulating variables are all linked to vascular dysfunction. The observation of vascular risk clustering in association with IR has led to the conclusion that cardiovascular risk appears early, prior to the onset of T2DM, whereas the strong interactions between hyperglycemia and microvascular disease suggest that this risk does not appear until frank hyperglycemia. These ideas emphasise the progressive nature of T2DM and accompanying cardiovascular risk, which poses unique issues at various phases of a diabetic's life. (Martín-Timón, 2014).

2.4 RNA-seq Approach:

RNA sequencing is an approach of Next generation sequencing which is used for mapping and quantifying transcriptomes. Data from RNA seq can help in the identification of biomarkers and thus provide insight for better therapeutic approaches. The differential gene expression is performed to find the upregulated and downregulated genes which are involved in GO and pathways. Based on the network analysis the hub genes are identified, which helps in a better understanding of the mechanism of Coronary Heart Disease and Hyperlipidemia. The Protocol includes downstream computational analysis including quality control, quantification of gene expression, and differential expression.

3. Objectives of the Study

To uncover the association between Coronary Heart Disease, Hyperlipidemia and Diabetes by-

Objective -1 Identification of Differentially Expressed genes between Coronary Heart Disease, Familial Coronary Heart Disease, Hyperlipidemia, Familial Hyperlipidemia and Normal using RNA-seq data.

Objective -2 Gene Enrichment and Pathway Analysis.

Objective -3 Identification of Hub Genes using Cytohubba.

Objective -4 Prefetching of diabetes related genes and Identifying the intersection between Coronary Heart Disease, Hyperlipidemia genes using GWAS Catalog.

Objective -5 Association and Visualization of Diabetes related genes with Coronary Heart Disease and Hyperlipidemia by Phenolyzer.

4. Materials and Methods

4.1 Materials:

4.1.1 NCBI BioProject and Sequence Read Archive:

Replaced by NCBI's Genome Project Database, Bioproject (https://www.ncbi.nlm.nih.gov/bioproject/) organizes metadata of large volumes of data and encircles biological data derived from a single organization or from a cohort of coordinating organizations associated with a single initiative. Thereby, it serves as a central portal for the submitted data representing higher order organisation, description and classification across several NCBI archival databases (Barrett et al., 2012). SRA is an international public archival resource for next generation sequence data. The largest individual global project generating next-generation sequence is the 1000 genome project which has half of all data in the SRA.

4.1.2 Sequence Retrieved Archive Toolkit:

The SRA Toolkit (https://trace.ncbi.nlm.nih.gov/Traces/sra/)allows us to use, view, and search the data within NCBI: SRA and convert it from SRA format to a required format (fastq/fasta/sam). An International Public Archival Resource for next-generation sequence data called Sequence Read Archive (SRA) was initiated by International Nucleotide Sequence Database Collaboration (INSDC) with the goal to preserve public-domain sequencing data and to provide freely, permanent access to the data (Leinonen, Sugawara, & Shumway, 2011). It has various tools to get a subset of large files, fetch specific sequences. (fastq_dump version ".0.10.9") To download the sequence, a prefetch tool is used and fastq-dump is used to convert the data format.

4.1.3 FastQC:

FastQC tool (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)is used to check the quality of the retrieved sequence. Quality of the raw-reads was examined using FastQC which is amalgamation of a command line interface (CLI) written in Python 3, depends on FastQC for processing FASTQ files, a frontend website written in JavaScript, HTML, and CSS, that utilizes Highcharts (http://www.highcharts.com) and D3 (https://d3js.org) Javascript libraries for plotting, and Bootstrap.js (https://getbootstrap.com) for styling and interactivity. With the advanced features, it provides reads quality by position, information of the presence of adaptor sequences, reports on tetramer frequencies and other characteristics of raw-reads (Brown, Pirrung, &Mccue, 2017). It has a set of analyses that check for any problem in the data which should be known before doing any further analysis. It takes input files in BAM/SAM or Fastq format and exports the results to an HTML based report which has the summary graphs and tables to assess the data.

4.1.4 Trimmomatic:

Trimmomatic(http://www.usadellab.org/cms/?page=trimmomatic) (version "0.39") is a flexible, pair-aware and efficient preprocessing tool, enhanced for illumina NGS data which includes a variety of processing steps for read trimming and filtering and identification of adaptor sequences and quality filtering. The poor quality of raw-reads can interfere negatively with the downstream analysis of the Next Generation Sequencing data (Bolger, Lohse, &Usadel, 2014). To avoid the presence of overrepresented sequences or adaptors in paired-read sequences, trimmomatic was used to trim the sequences below a threshold value.

4.1.5 HISAT2:

HISAT2 (https://ccb.jhu.edu/software/hisat2) is a fast and sensitive alignment program for mapping the next generation sequence read against the general human population. HISAT2 is a novel genome indexing scheme that implements a Graph FM (GFM) index to capture a wide representation of genetic variants and align raw sequencing reads to a graph. It intends to provide higher alignment accuracy that captures the entire human genome with very low memory requirements (Kim, Paggi, Park, Bennett, & Salzberg, 2019). HISAT2 is developed based on the HISAT and Bowtie2 implementations. The output is in SAM format, which can be interoperated with other tools such as SAMtools, GATK that use SAM.

4.1.6 SAMTools:

SAMtools(http://www.htslib.org/) (version "1.1") is a set of programs for using high-throughput sequencing data. SAMtools, a library and software package for parsing and modifying alignments in the SAM/BAM format, was used. Convert alignments from other formats, sort and merge alignments, delete PCR duplicates, create per-position information in the pileup format, call SNPs and short indel variants, and display alignments in a text-based viewer are all features of SAMtools. BAM, the corresponding binary representation, is small and allows for fast retrieval of alignments in specific regions. Applications can perform stream-based processing on particular genomic regions using positional sorting and indexing instead of loading the entire file into memory. The SAM/BAM format, in combination with

SAMtools, allows a gene to be aligned independently of downstream analyses (Li et al., 2009). It consists of reading, writing, sorting, editing, indexing, and viewing the SAM/BAM file formats. SAMtools helps in converting the SAM format to the BAM file format which is compressed and more efficient to work with. It also summarizes SNP and short indel sequences.

4.1.7 Ensembl:

Ensembl(https://www.ensembl.org) is a method for producing and disseminating genome annotation such as genes, variation, control, and comparative genomics for vertebrates and main model species (Yates et al., 2020). The Ensembl human genome-103 transcripts has been taken from Ensembl.

4.1.8 StringTie:

StringTie (https://ccb.jhu.edu/software/stringtie/) is a fast and efficient assembler of RNA - Seq alignments into potential transcripts. It assembles and quantifies full-length transcripts representing multiple splice variants for each gene locus. StringTie is a transcript assembler that determines gene expression levels using the highest flow optimization technique in a specially constructed flow network while simultaneously assembling each isoform of a gene. It also includes alignment to a genome as well as de novo read assembly, unlike other transcript assemblers (Pertea et al., 2015). The differentially expressed genes between experiments can be identified using StringTie’s output which can be processed by software like DESeq2, edgeR, Ballgown, or cuffdiff.

4.1.9 DeSeq2:

The DESeq2 (https://bioconductor.org/packages/release/bioc/html/DESeq2.html) method, (in a R/Bioconductor package) detects and corrects low dispersion estimates by projecting the dispersion's reliance on the average expression intensity across all samples. The most common method for comparing transcriptomics data is to test the null hypothesis that the logarithmic fold shift (LFC) for a gene's expression between treatment and control is exactly zero, implying that the gene is unaffected by the treatment. Differential analysis is often used to generate a list of genes that pass multiple-test adjustment and are ranked by P value. Significant adjustments, however, may not be the most interesting candidates for further research, even if statistically valid. The benefits of DESeq2's new features by defining a variety of applications that can be done with shrunken fold changes and their standard error estimates, such as improved gene ranking and visualisation, hypothesis tests above and below a threshold, and the regularised logarithm transformation for quality evaluation and clustering of overdispersed count info (Love, Huber, & Anders, 2014).

4.1.10 Cluster Profiler:

One of the most extensively utilised strategies for evaluating gene lists or genome-wide areas of interest (ROIs) produced from various high throughput investigations is functional enrichment analysis (Feng et al., 2021). The cluster profiler (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) in R (version "4.1"), released under the Bioconductor project, is used for Statistical analysis and visualization of functional profiles for genes and gene clusters. The maps of the whole GO and KEGG corpus are obtained from GO.db and KEGG.db (Yu et al., 2012). Instead of providing species-specific GO annotation, clusterProfiler uses the Bioconductor project's genome-wide annotation packages (OrgDb) (Feng et al., 2021). In addition, clusterProfiler includes a function called compareCluster that automatically calculates enriched functional categories for each gene cluster and offers numerous viewing options. Also, it can compare gene–disease relationships among gene clusters in collaboration with the R package DOSE (Yu et al., 2012).

4.1.11 String Database:

STRING (https://string-db.org/) is a biological database that holds information about known and predicted protein-protein interactions. It also serves to highlight functional enrichment in the user-provided list of proteins, using several functional classification systems (von Mering et al., 2003).

4.1.12 Cytoscape:

Cytoscape (https://cytoscape.org/) is an open-source platform for visualizing molecular and interaction networks and integrating with gene expression profiles and annotations. Also, Cytoscape has a rich collection of Plugins for network and molecular profile analysis (Paul Shannon et al., 1971). CytoHubba is a new Cytoscape plugin for ranking nodes in a network based on network attributes. Degree, Edge Percolated Component, Maximum Neighborhood Component, Density of Maximum Neighborhood Component, Maximal Clique Centrality, and six centralities (Bottleneck, EcCentricity, Closeness, Radiality, Betweenness, and Stress) are among the 11 topological analysis methods available in CytoHubba (Chin et al., 2014).

4.1.13 Phenolyzer:

Phenolyzer (http://phenolyzer.usc.edu) is a tool to map user-supplied phenotypes to related diseases, a resource that integrates existing knowledge on known disease genes, an algorithm to predict previously unknown disease genes, a machine learning model that scores and prioritises all candidate genes, and a network visualisation tool to examine gene-gene and gene-disease relationships. It translates user-supplied disease or phenotypic phrases into related disease names, which are then used to search and score relevant seed genes in precompiled databases (Yang et al., 2015).

4.1.14 KEGG Database:

KEGG (https://www.genome.jp/kegg/disease/) is an eighteen-database resource that has been manually curated and is organised into systems, genomics, chemicals, and health information. It also includes KEGG mapping tools, which help researchers comprehend cellular and organism-level activities by analysing genome sequences and other molecular data. Based on the notion of functional orthologs, KEGG mapping is a prediction approach of recreating molecular network systems from molecular building blocks. Various diseases have been linked to network variations, which disrupt molecular networks caused by human gene variants, viruses, other pathogens, and environmental factors, since the launch of the KEGG NETWORK database (Furumichi et al., 2021).

4.1.15 GWAS Catalog:

The NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/) is a database of published human GWAS that is open to the public. Expert scientists manually curate each publication to ensure that the Catalog has correct and structured metadata for publishing, study design, sample and trait data, and the most important published results. The GWAS Catalog provides a high-quality curated database of all published genome-wide association studies, allowing researchers to find causal variations, understand disease mechanisms, and find new therapeutic targets. It contains 284 datasets with comprehensive P-value summary statistics for genome-wide and new targeted array studies, totaling 6 x 109 unique variant-trait statistics (Buniello et al., 2019).

4.2 Methodology

4.2.1 Extraction of Raw Reads:

From NCBI Bioproject, the raw reads of the samples of healthy, coronary heart disease, hyperlipidemia, familial coronary heart disease and familial hyperlipidemia with accession number of Bioproject dataset “PRJNA663423” were extracted. The prefetch tool of the SRA toolkit is used to download the data which is in SRA format. Next, the fastq-dump was used to convert the SRA data to a fastq format. As the reads were paired end, split files function was used to dump each sequence as separate forward and reverse files. With the help of SRA toolkit, the .sra files of the 15 biosamples were retrieved.

4.2.2 Quality Assurance - I:

Using the .gzip format files, an .html file report of fastQC was generated. With the help of the .html result file, Quality assurance of base sequence was done by observing the whisker plot with stretched outs peaks symbolizing the mean of quality. The presence or absence of the adaptor sequences was visualised.

4.2.3 Trimming of Raw-reads:

Using Trimmomatic, the overrepresented sequences and adaptors were trimmed. It provided 4 files, 2 paired and 2 unpaired whereas unpaired files gave information of the sequences which had been trimmed. Hence, the focus is only on the paired sequences.

4.2.4 Quality Assurance – II:

Using the fastQC, the quality of the trimmed sequences were checked to get assurance of the quality to be falling in the region above >20 on the Precision scoring matrix PHRED scale.

- >25 = absolutely good (Green zone)
- 20-25 = good (yellow zone)
- <20 = relatively poor (pink zone)

4.2.5 Indexing and Sequence Alignment:

To align the raw reads with the human genome, the transcript genome index of Human Genome “grch38_tran.tar.gz” was downloaded directly from the freely available web-page of HISAT2. The Hisat2 was used to align the reads to the human reference genome.

The alignment score of each raw sequence with the human transcript genome index was noted. The output is in .sam format. The SAM format consists of one header section which starts with character ‘@’ and one alignment section. All the lines in sam format are TAB delimited.

4.2.6 Conversion of .sam to .bam files:

The .sam files were converted to .bam files using SAMtools. Later, the sorted BAM files were generated which were further used for the quantification purposes. A BAM file (. bam) is the binary version of a SAM file. BAM file contains information about the entire file, such as sample name, sample length, and alignment method. Alignments in the alignments section are associated with specific information in the header section.

4.2.7 Assembly of the transcripts:

StringTie tool was used to assemble the transcripts for each sample, where the BAM file was used as the input file. The file generated the assembly file for each dataset and using option – merge, all the assemblies were run to create a single merged transcriptome annotation. With StringTie, the merged transcriptome assembly along with the BAM files from the previous hisat2 for each replicate using option –e and –B.

Once the mapping of the transcripts was done, the output came in .gtf file format. Merging of the mapped transcripts was accomplished by creating a text file with the paths of all the output of samples in the gtf file.

The output stringtie_merged.gtf gave the information of mapped transcripts. Further, the head file and the numbers of transcripts were checked using the following command lines

To further analyse the transcript expression level, refers to how abundant a transcript, i.e. if a certain mRNA (a transcript) exist in many or few copies (expression level), using stringtie, the abundance of the each sample was done. Furthermore, the available python script (http://ccb.jhu.edu/software/stringtie/dl/prepDE.py) on stringtie webpage was used to generate the read counts that were used to find the differential expressed genes.

4.2.8 Library Normalisation:

The DESeq2 package was loaded into the R environment. A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data were presented as a table which reported, for each sample, the number of sequence fragments that have been assigned to each gene. The distribution of expression levels of each sample was plotted and the expression of each gene was generated using a plot.
The comparison analysis was done with respect to -

- Healthy VS. Coronary Heart Disease samples
- Healthy VS. Hyperlipidemia samples
- Healthy VS. Familial Hyperlipidemia samples
- Healthy VS. Familial Coronary Heart Disease samples.

4.2.9 Differential Expressed Gene Analysis:

The goal of DESeq2 is to calculate a scaling factor for each sample. Scaling factors take read depth and library composition into account. By considering the p – value < 0.05 and Log2 Fold Change (FC)| ≥ +2 as up-regulated genes and | Log2 Fold Change (FC)| ≤ -2 as downregulated genes were evaluated.

4.2.10 Gene Enrichment Analysis:

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether a predefined set of genes (ex: those belonging to a specific GO term or KEGG pathway) shows statistically significant, concordant differences between two biological states. Using Cluster Profiler in R, specific genes, functions were compared with CHD, FCHD, HYP, and FHYP.

Dot Plots, Enrichment Maps and Ridge plots were generated using the Cluster Profiler in the R environment.

4.2.11 Phenotype Gene analysis:

For Phenotype based gene analysation and to find the genes associated with Diabetes, Phenolyzer was used by entering gene ids and disease name. Network plot showing protein-protein interaction, transcription interaction, genes having same family and in same biosystem was generated. Bar Plot showing the expression of the particular gene with gene enrichment value was visualized.

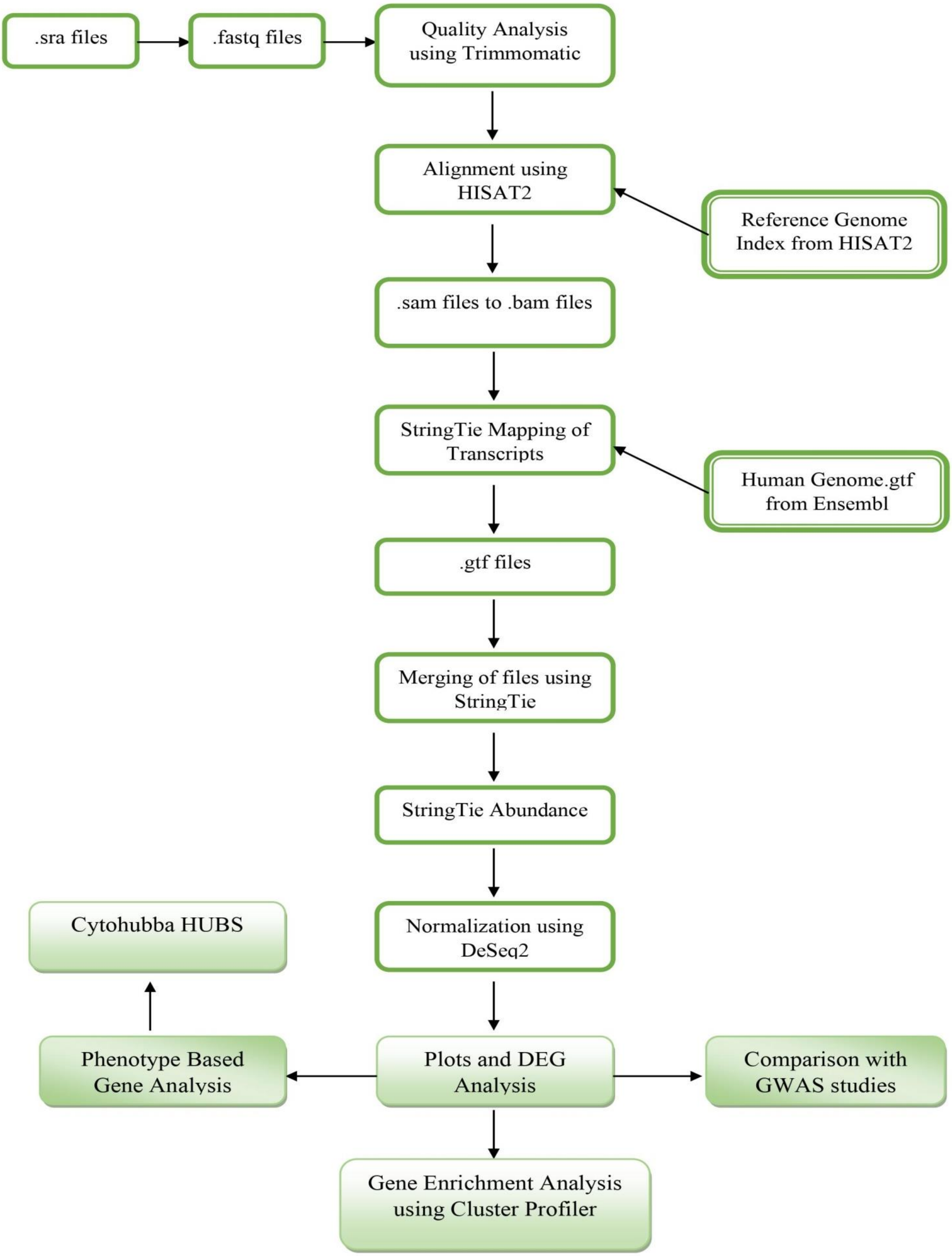
4.2.12 Identification of Hubs:

The protein-protein interaction network was generated for the differentially expressed genes using the STRING database. In this database, the association between the proteins is based on co-occurrence, co-expression, gene-fusion, literature mining, and some computational predictions. The network was constructed for the upregulated genes and down regulated genes. The constructed network was analysed using Cytoscape based on factors such as degree,betweenness, and closeness. The genes which have a large number of connecting nodes (highest degree) were considered as hub proteins. These hub proteins were further analysed.

4.2.13 Comparative analysis with KEGG db and GWAS Catalog:

KEGG database, specifically, KEGG Disease database and Genome Wide Association Studies Catalog were used to find the genes associated with Diabetes comparable to the Differential Expressed genes of Coronary Heart Disease and Hyperlipidemia. The genes having positive association with Diabetes, Coronary Heart Disease and Hyperlipidemia were further analysed with Literature.

5. **Workflow**


.sra files
.fastq files
Quality Analysis using Trimmomatic
Alignment using HISAT2
Reference Genome Index from HISAT2
.sam files to .bam files
StringTie Mapping of Transcripts
Human Genome.gtf from Ensembl
.gtf files
Merging of files using StringTie
StringTie Abundance
Normalization using DeSeq2
Cytohubba HUBS
Phenotype Based Gene Analysis
Plots and DEG Analysis
Comparison with GWAS studies
Gene Enrichment Analysis using Cluster Profiler

6. Results

6.1 Raw Read Samples:

From the Bio-project dataset Accession ID: PRJNA663423, the metadata of 3 Raw-read samples of H, CHD, FCHD, HYP and FHYP were taken into consideration for RNA-seq data analysis. All the samples belong to male category of different age groups (Table 1).

Table 1: Shows the age of all the male samples

H - Healthy		
Sample 1	Sample 2	Sample 3
Age - 23 y	Age - 19 y	Age - 36 y
CHD - Coronary Heart Disease		
Sample 1	Sample 2	Sample 3
Age - 41 y	Age - 52 y	Age - 56 y
HYP - Hyperlipidemia		
Sample 1	Sample 2	Sample 3
Age - 65 y	Age - 50 y	Age - 55 y
FCHD - Familial Coronary Heart Disease		
Sample 1	Sample 2	Sample 3
Age - 79 y	Age - 45 y	Age - 45 y
FHYP - Familial Hyperlipidemia		
Sample 1	Sample 2	Sample 3
Age - 50 y	Age - 45 y	Age - 20 y

6.2 Per Base Sequence Quality:

The fastqc tool was used to check the quality of the reads, it produces an HTML report with summary graphs. It has many different analyses that are performed on the sequence data. A graph displaying the estimated quality score (shown in Fig. 2) at each position of the reads in all the sequence was produced and the mean quality score in all the base calls for each sequencewas calculated. Fastqc also gave statistics of the GC content, sequence duplication, adapter content,per base N content, sequence length distribution.

Gentle Quality trimming and adaptor clipping is a crucial step, which was performed with the paired end raw data sequence. The over-expressed sequences, adaptors were removed. In the fig., the Per base sequence quality of one raw sequence before trimming and after trimming, can be visualized. Quality assurance was done using the Precision Scoring matrix Phred scale, where all the reads belong to more than 26, on Phred scale.

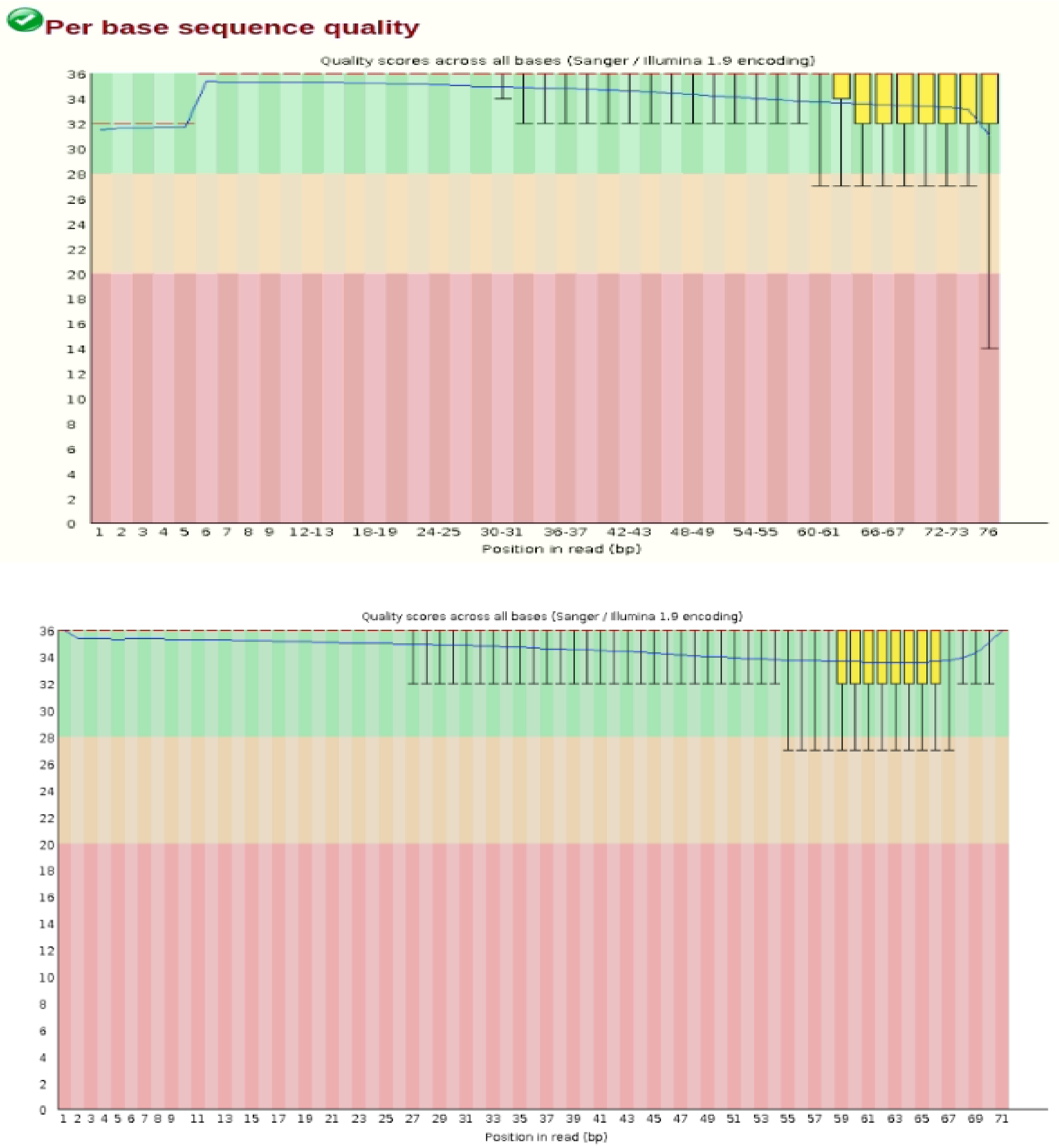


Fig. 2: Per base sequence quality of the raw sequence before and after trimming

6.3 Alignment of Raw Reads:

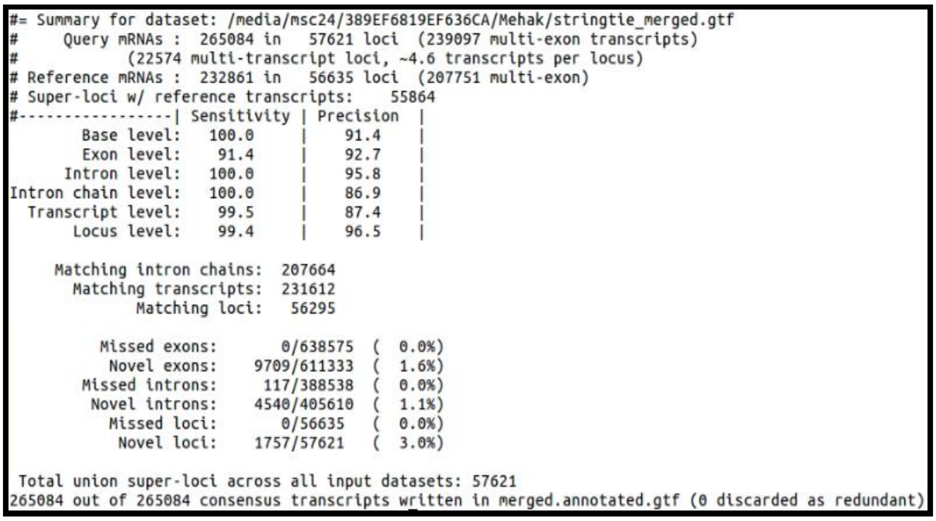
Pre-processed reads were aligned to the human reference genome using HISAT2. It produced a set of SAM files which contained the read alignment for each RNA library. The computed best Alignment score of all the reads against the Human Genome Transcript Index was noted. The alignment score was more than 95% for each of the sequence which makes the alignment finest and reliable (Table 2)

Table 2: Shows the exact alignment rate of each sample fetched using HISAT2

Samples **(total = 15)**	**Healthy**	**Coronary Heart Disease**	**Familial Coronary Heart Disease**	**Hyperlipidemia**	**Familial Hyperlipidemia**
Sample 1	96.86%	95.99%	96.91%	96.50%	97.13%
Sample 2	96.77%	97.11%	96.88%	96.86%	96.35%
Sample 3	97.03%	96.89%	96.67%	96.67%	97.19%

6.4 Transcript Estimation:

Assembled matched transcripts and details of matched/mismatched Exon and Intron against Ensembl Human Genome using Stringtie were evaluated. The overall total number of transcripts was found to be 265084, out of which 231612 transcripts were toned.



```
#= Summary for dataset: /media/msc24/389EF6819EF636CA/Mehak/stringtie_merged.gtf
#     Query mRNAs :  265084 in   57621 loci  (239097 multi-exon transcripts)
#            (22574 multi-transcript loci, ~4.6 transcripts per locus)
# Reference mRNAs :  232861 in   56635 loci  (207751 multi-exon)
# Super-loci w/ reference transcripts:    55864
#-----------------| Sensitivity | Precision  |
        Base level:   100.0     |    91.4    |
        Exon level:    91.4     |    92.7    |
      Intron level:   100.0     |    95.8    |
Intron chain level:   100.0     |    86.9    |
  Transcript level:    99.5     |    87.4    |
       Locus level:    99.4     |    96.5    |

     Matching intron chains:  207664
       Matching transcripts:  231612
              Matching loci:   56295

          Missed exons:       0/638575  (  0.0%)
           Novel exons:    9709/611333  (  1.6%)
         Missed introns:    117/388538  (  0.0%)
          Novel introns:   4540/405610  (  1.1%)
           Missed loci:       0/56635   (  0.0%)
            Novel loci:    1757/57621   (  3.0%)

 Total union super-loci across all input datasets: 57621
265084 out of 265084 consensus transcripts written in merged.annotated.gtf (0 discarded as redundant)
```

Fig. 3: Shows the details of the matched transcripts.

6.5 Differential Expressed Genes:

DeSeq2 identifies Differential Expressed genes by estimating variance-mean dependence in count data from high-throughput sequencing assays and tests for differential expression based on a model using the negative binomial distribution. The obtained reads were assembled using StringTie to measure the gene expression and differentially expressed genes using the DESeq2 package in R. As input, the DESeq2 package required count data from RNA seq in the form of a matrix. DESeqDataSet was the object class used to store the read counts. Using the feature counts function of Rsubread package, the count matrix was produced, thus the function DESeqDataSetFromMatrix was used. By keeping the value of LogFC ± 2 (more than +2 = upregulated and less than -2 = downregulated) and Padj. < 0.05, more stringent and statistically significant value, the upregulated and downregulated genes were found.

In Fig. 4, the common Differentially Expressed Genes were found using the tool called Multiple List Comparator. Venn Diagrams were generated in between all the four categories. In between HYP and FHYP, intersection of 18 DEG's were found whereas in CHD and FHYP, 13 DEG's were found in common. In table 3, the common genes between the CHD & FCHD, HYP & FHYP are listed. In table 4, 61 genes with at least occurrence of 2 are listed.

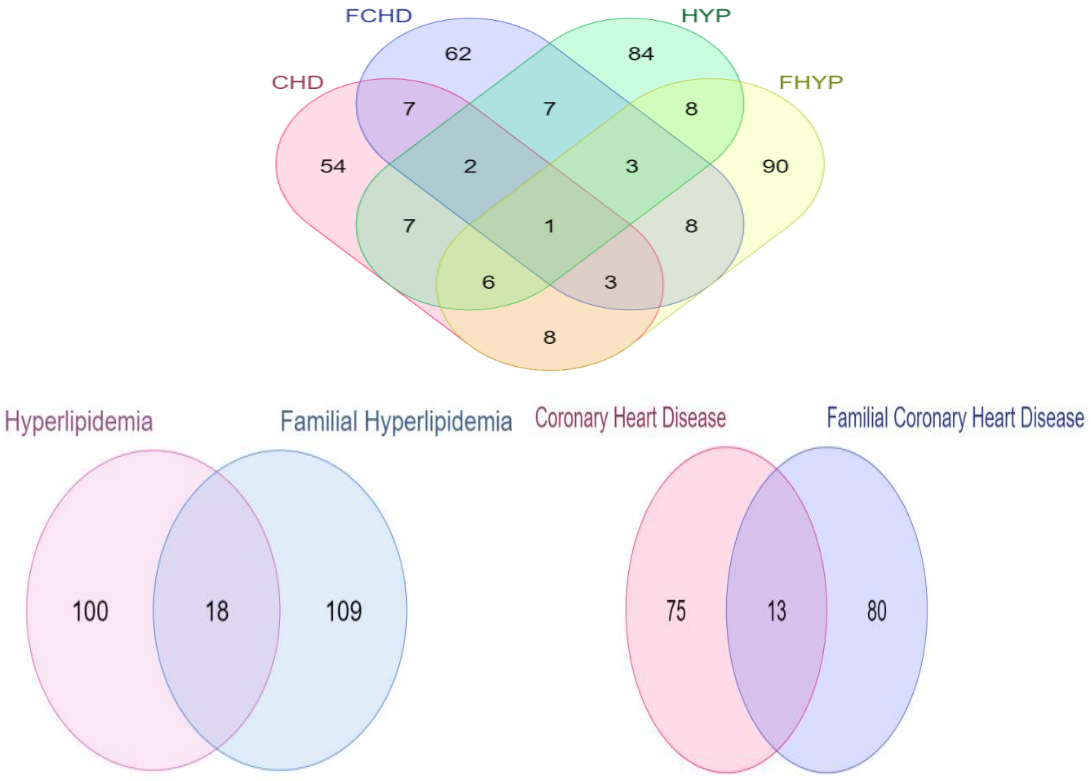


Fig. 4: Venn Diagrams showing common genes among the particular category

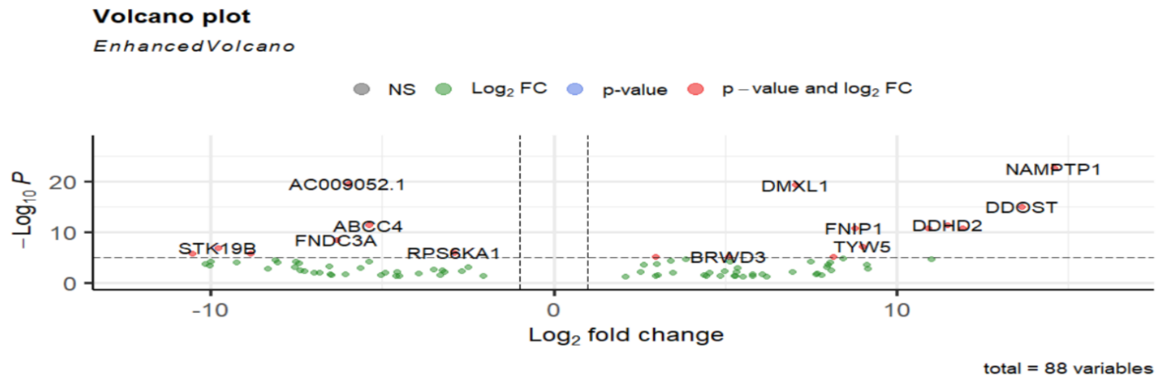
Table 3: Listed Common Genes in between CHD & FCHD, HYP & FHYP

13 Common Genes in between CHD and FCHD	Gene Names	18 Common Genes in between HYP and FHYP	Gene Names
AC004223.3	--	AC093512.2	--
AC113935.1	--	AC113935.1	--
AL157871.6	--	ACOX1	acyl-CoA oxidase 1
CHST11	Carbohydrate sulfotransferase 11	AL157895.1	--
DHX57	DExH-box helicase 57	CES4A	carboxylesterase 4A
EIF3CL	eukaryotic translation initiation factor 3 subunit C-like	CWC22	CWC22 spliceosome associated protein homolog
F13A1	coagulation factor XIII A chain	DSP	desmoplakin
FBXL14	F-box and leucine rich repeat protein 14	FNDC3A	fibronectin type III domain containing 3A
GRIN3A	glutamate ionotropic receptor NMDA type subunit 3A	GFI1B	growth factor independent 1B transcriptional repressor
IP6K1	inositol hexakisphosphate kinase 1	HLA-V	major histocompatibility complex, class I, V (pseudogene)
PTER	phosphotriesterase related	HRAT92	heart tissue-associated transcript 92
QKI	QKI, KH domain containing RNA binding	MTHFSD	Methenyl tetrahydrofolate synthetase domain containing
USP24	ubiquitin specific peptidase 24	OXSM	3-oxoacyl-ACP synthase, mitochondrial
		RPS5	ribosomal protein S5
		SNED1	sushi, nidogen and EGF like domains 1
		STK19B	serine/threonine kinase 19B (pseudogene)
		TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor
		ZNF701	zinc finger protein 701

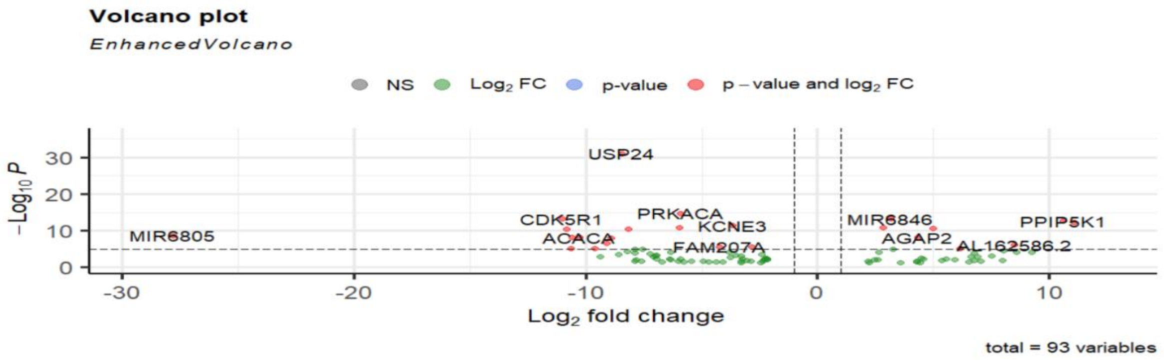
6.6 Volcano Plots:

A volcano plot is a type of scatterplot that shows statistical significance (P value) versus magnitude of change (fold change). It enables quick visual identification of genes with large fold changes that are also statistically significant. These may be the most biologically significant genes. In a volcano plot, the most upregulated genes are towards the right, the most downregulated genes are towards the left, and the most statistically significant genes are towards the top. The Differential Expressed Genes file from DESeq2 was taken in consideration and the volcano plots were generated for the each category (Fig. 5), i.e., Coronary Heart Disease, Familial Coronary Heart Disease, Hyperlipidemia and Familial Hyperlipidemia.

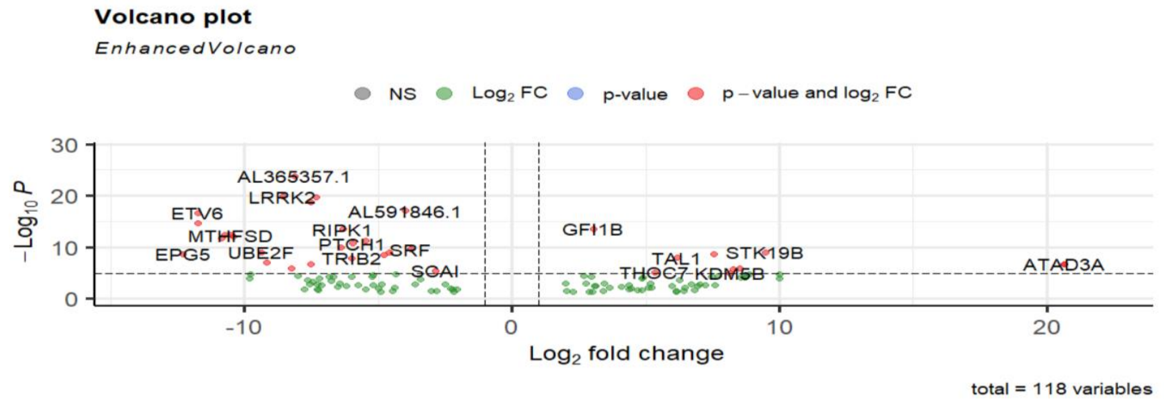
A) Coronary Heart Disease



B) Familial Coronary Heart Disease



C) Hyperlipidemia



(d) Familial Hyperlipidemia

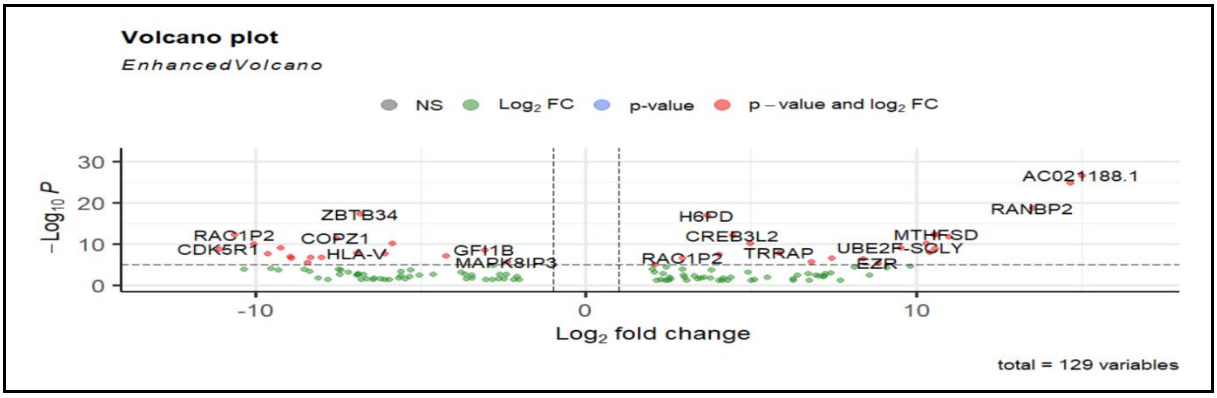


Fig. 5: Volcano Plot showing Upregulated and Downregulated genes against LogFC and Padj value for A) Coronary Heart Disease, B) Familial Coronary Heart Disease, C) Hyperlipidemia and D) Familial Hyperlipidemia.

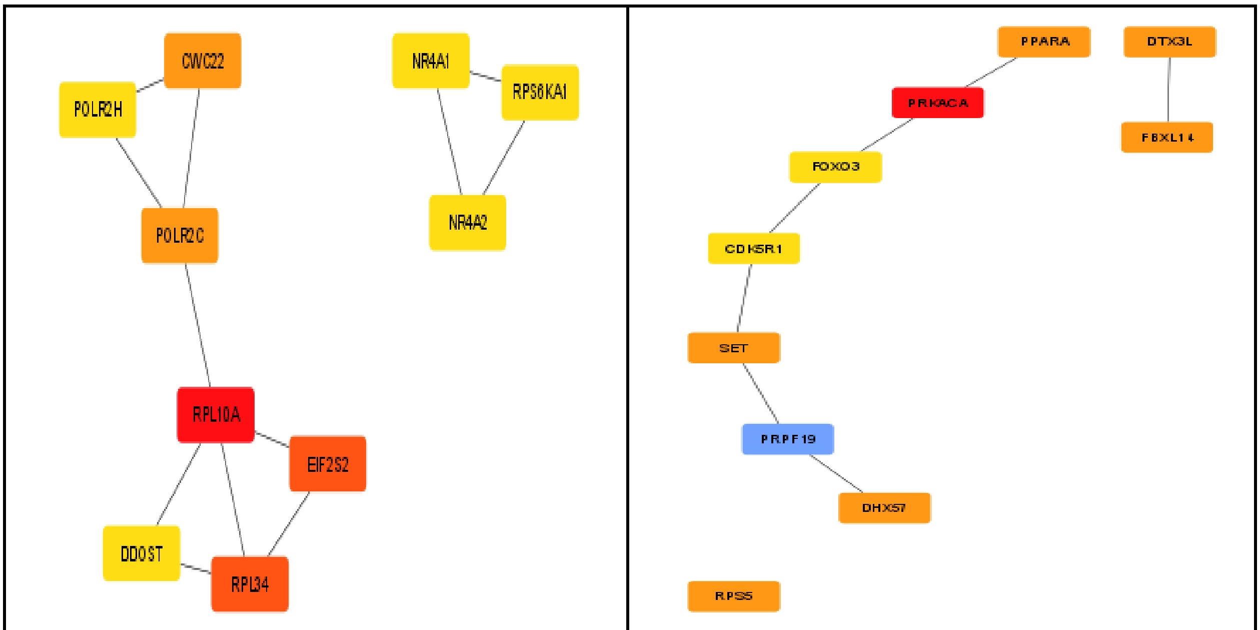
Table 4: Overall comparison between the Differential genes in between CHD, FCHD, HYP, FHYP

Item	Occurrences	Present In	Item	Occurrences	Present In
AC113935.1	4	CHD, FCHD, HYP, FHYP	DAPK3	2	CHD,HYP
ACOX1	3	FCHD, HYP, FHYP	DHX57	2	CHD, FCHD
CWC22	3	CHD, HYP, FHYP	DSP	2	HYP, FHYP
EIF3CL	3	CHD, FCHD, FHYP	EIF2AK3	2	CHD, FHYP
FNDC3A	3	CHD, HYP, FHYP	F13A1	2	CHD, FCHD
GRIN3A	3	CHD, FCHD, HYP	FBXL14	2	CHD, FCHD
HLA-V	3	FCHD, HYP, FHYP	FEN1	2	FCHD, FHYP
HRAT92	3	CHD, HYP, FHYP	FNIP1	2	CHD, HYP
OXSM	3	CHD, HYP, FHYP	FOXO3	2	FCHD, HYP
PTER	3	CHD, FCHD, HYP	FOXO3B	2	CHD, FHYP
QKI	3	CHD, FCHD, FHYP	GABPB2	2	CHD, HYP
RPS5	3	FCHD, HYP, FHYP	GFI1B	2	HYP, FHYP
SNED1	3	CHD, HYP, FHYP	HIPK3	2	FCHD, FHYP
STK19B	3	CHD, HYP, FHYP	IP6K1	2	CHD, FCHD
USP24	3	CHD, FCHD, FHYP	MED28	2	FCHD,HYP
ABCC4	2	CHD, HYP	MTHFSD	2	HYP, FHYP
AC004223.3	2	CHD, FCHD	NAMPTP1	2	CHD, FHYP
AC067968.1	2	FCHD, FHYP	PHYHD1	2	CHD, HYP
AC087854.1	2	CHD, FHYP	PPIP5K1	2	FCHD, FHYP
AC093512.2	2	HYP, FHYP	RN7SKP241 PRTFDC1	2	FCHD, HYP
AC139769.2	2	FCHD, HYP	RPL10A	2	CHD, HYP
AL157871.6	2	CHD, FCHD	RSU1	2	CHD, FHYP
AL157895.1	2	HYP, FHYP	SACM1L	2	FCHD, HYP
B4GALT5	2	FCHD, HYP	SPIN2B	2	CHD, FHYP
BRD3OS	2	FCHD, FHYP	TAL1	2	HYP, FHYP
CCL3	2	FCHD, FHYP	TMEM184B	2	CHD, FHYP
CDK5R1	2	FCHD, FHYP	ZAP70	2	CHD, HYP
CES4A	2	HYP, FHYP	ZBTB34	2	FCHD, FHYP
CHST11	2	CHD, FCHD	ZNF549	2	FCHD, HYP
CPAMD8	2	CHD, FHYP	ZNF701	2	HYP, FHYP
DAPK3	2	CHD, HYP			

6.7 Cytoscape - Cytohubba HUBS:

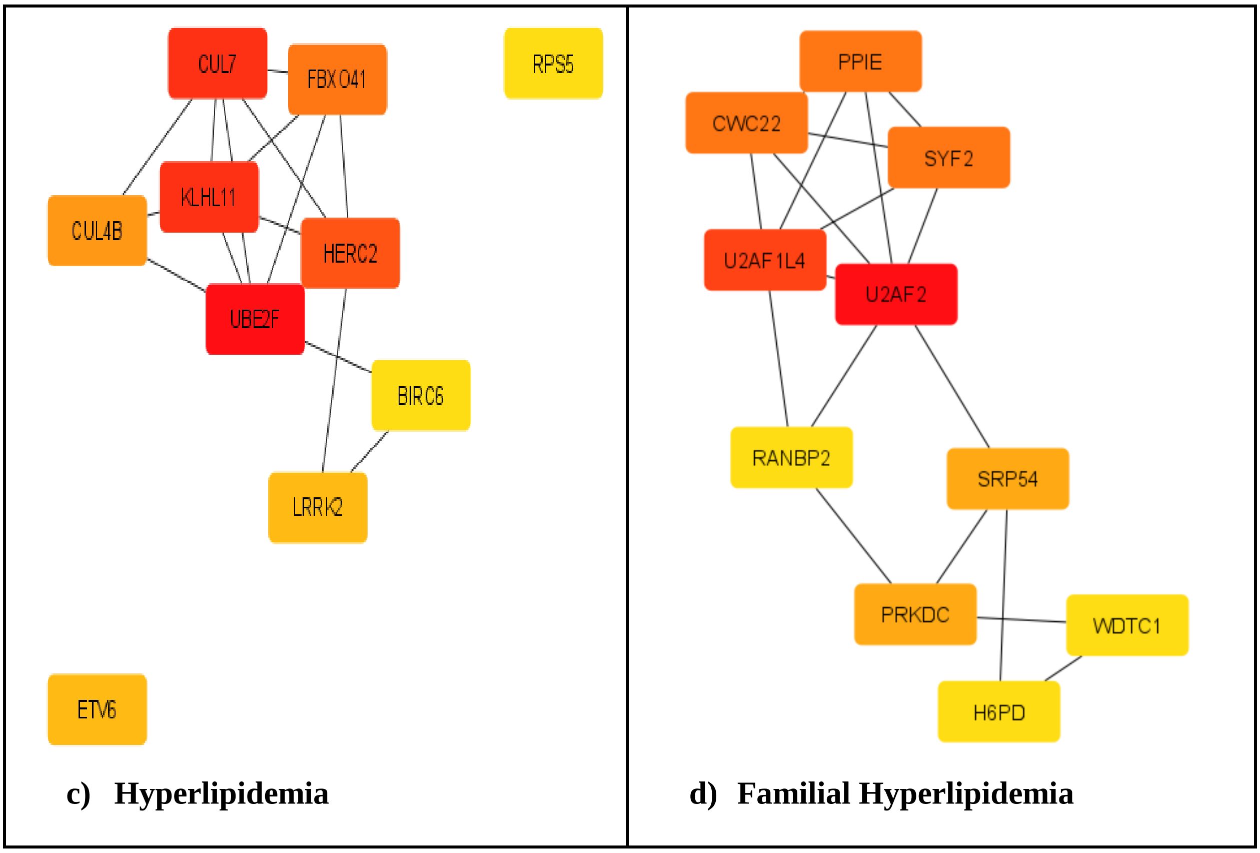
The differentially expressed genes of both upregulated and downregulated genes were used to construct the protein-protein interaction network with significant interaction with p value < 0.05. The association such as co-expression, co-occurrence, gene fusion, database, experiments were considered for network construction. The PPI network was constructed using the STRING database which was then analysed in Cytoscape to find the hub proteins.

Based on highest connectivity, closeness, and betweenness - the hub proteins using Cytoscape's plug-in Cytohubba are identified which are shown in fig. 6.



a) **Coronary Heart Disease**

b) **Familial Coronary Heart Disease**



c) **Hyperlipidemia**

d) **Familial Hyperlipidemia**

Fig. 6: Shows identified hubs in (a) Coronary Heart Disease (b) Familial Coronary Heart Disease (c) Hyperlipidemia (d) Familial Hyperlipidemia using Cytohubba in Cytoscape

In the four categories (Refer Fig. 6), Two Hub genes, CWC22 and RPS5 were found to be present in CHD, FHYP and FCHD, HYP respectively (Table 5).

Table 5: Common Hubs

Gene name	Gene ID	Occurrences	Category
CWC22 spliceosome associated protein homolog	CWC22	2	CHD , FHYP
ribosomal protein S5	RPS5	2	FCHD , HYP

6.8 Gene Enrichment Analysis:

The GO Gene Set Enrichment Analysis was done using a Bioconductor package called Cluster Profiler. It considers genes, their logfc values, p-values and compares the genes, their functions. Later, it gives the best gene ontology functions by performing numerous permutations with Gene Ontology ID, Genes involved and their functions. In Fig. 7-10, the gene ontology with gene enrichment analysis can be visualised from the Dot Plot, Enrichment Map and Ridge Plot.

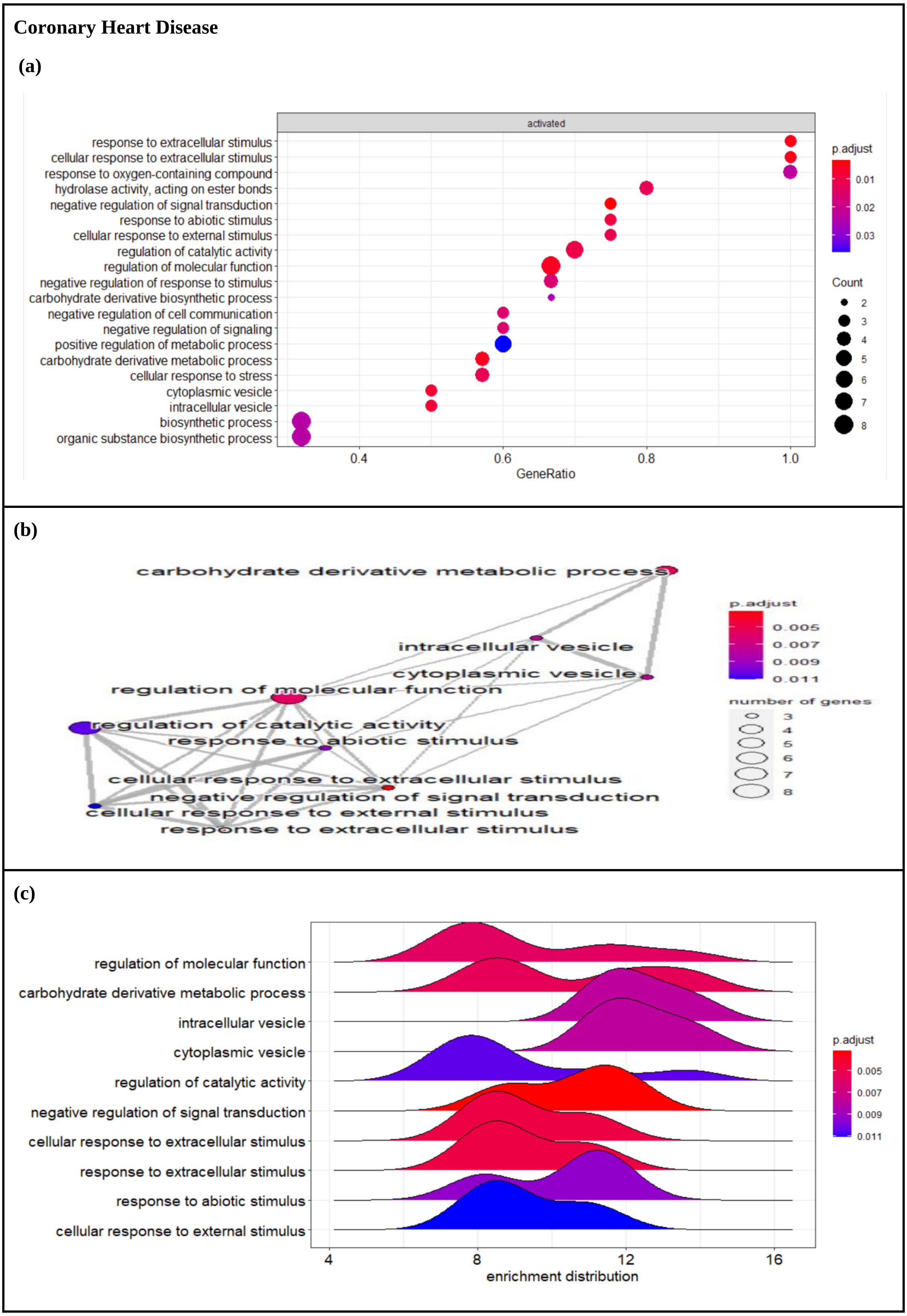


Fig. 7: GO-GSEA Plots (a) Dot Plot (b) Enrichment Map (c) Ridge Plot of Coronary Heart Disease

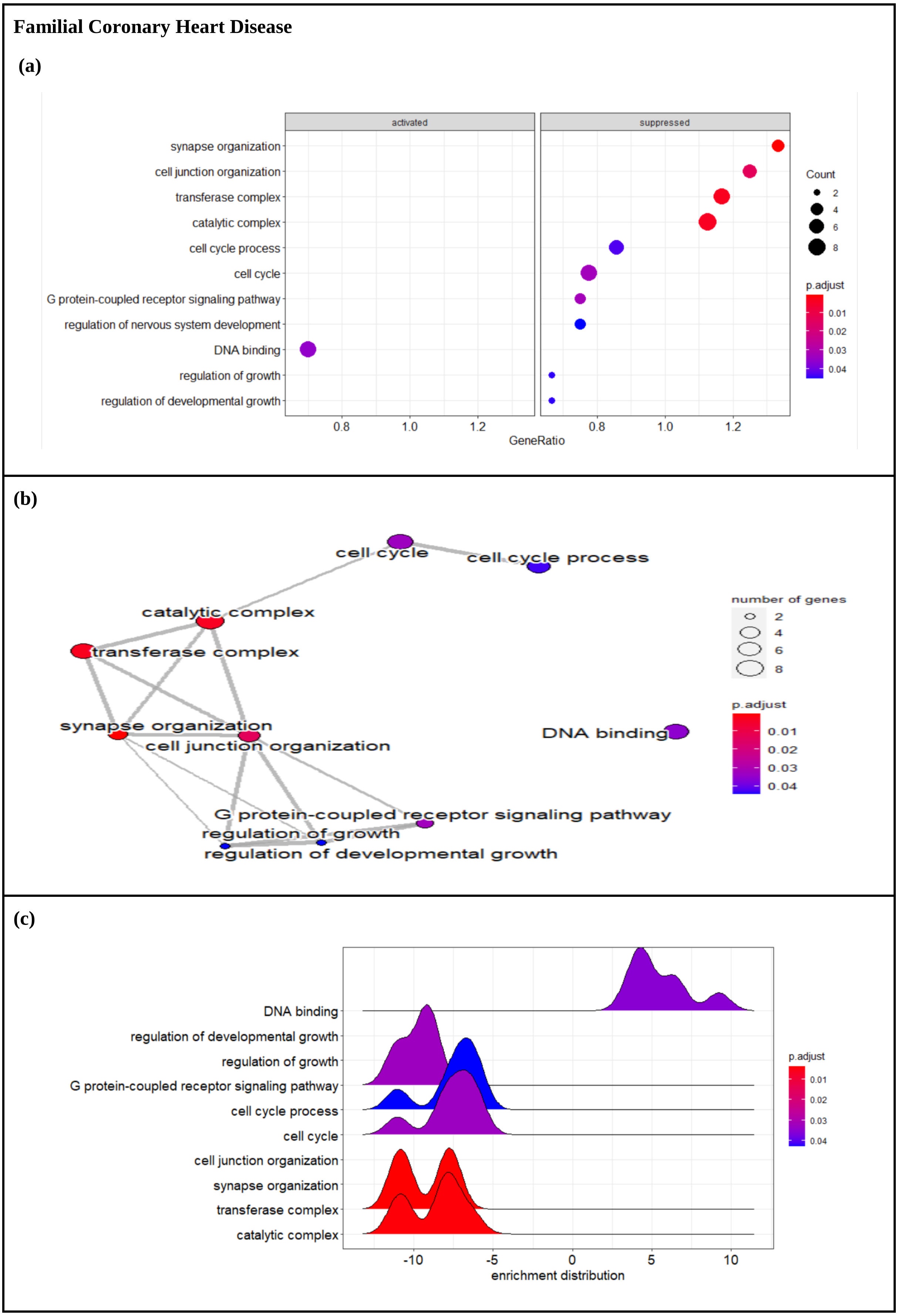


Fig. 8: GO-GSEA Plots (a) Dot Plot (b) Enrichment Map (c) Ridge Plot of Familial Coronary Heart Disease

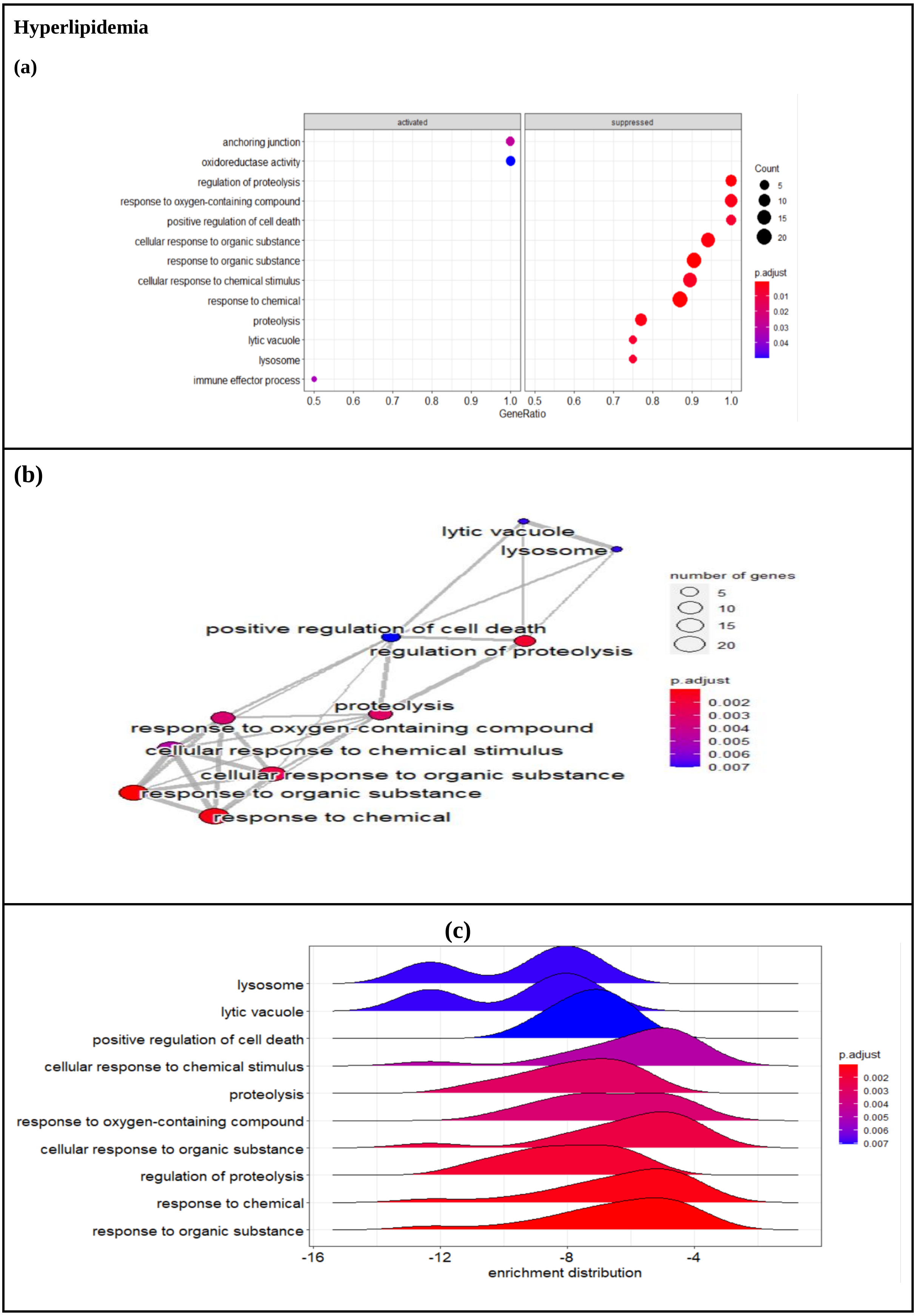


Fig. 9: GO-GSEA Plots (a) Dot Plot (b) Enrichment Map (c) Ridge Plot of Hyperlipidemia

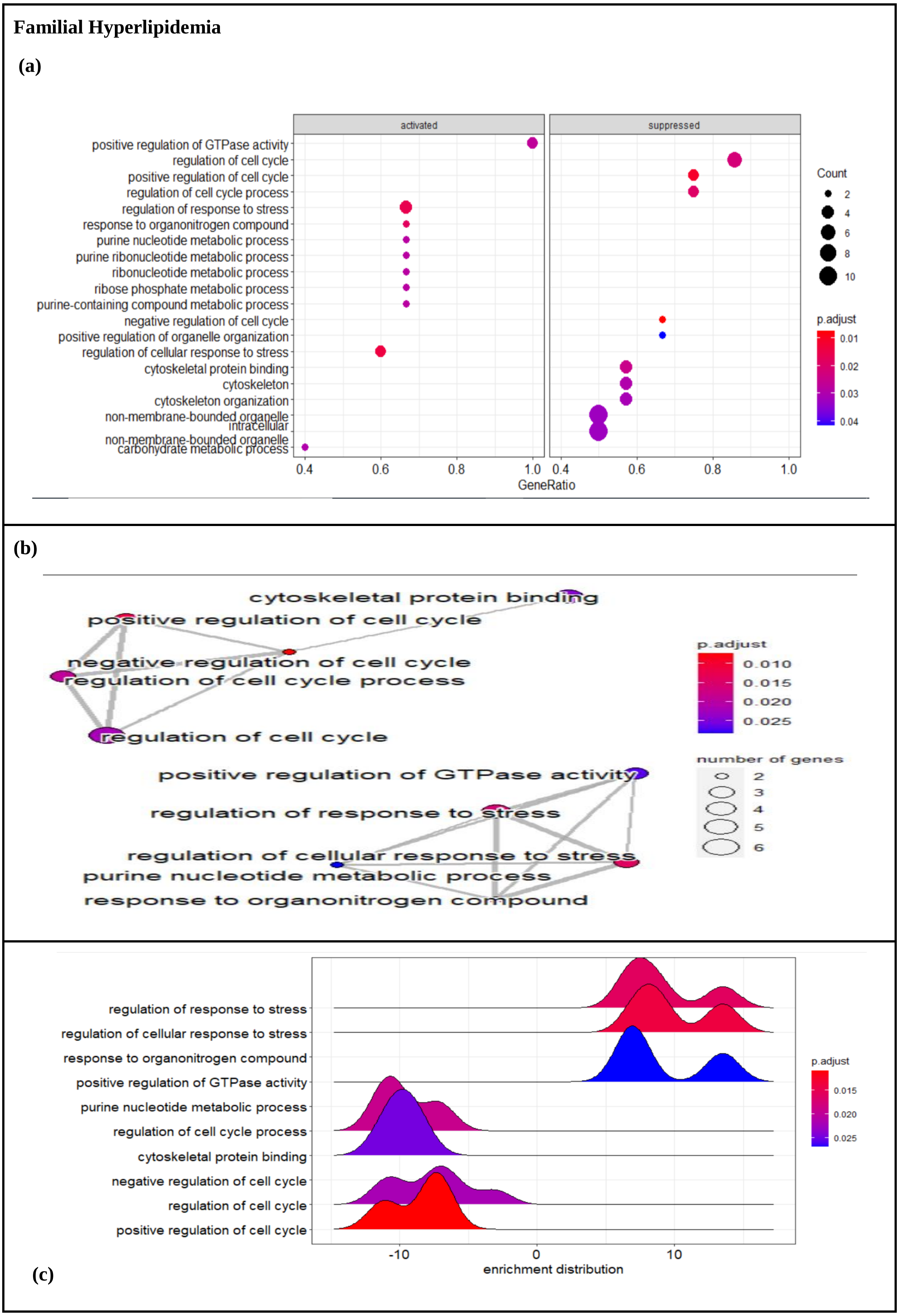


Fig. 10: GO-GSEA Plots (a) Dot Plot (b) Enrichment Map (c) Ridge Plot of Familial Hyperlipidemia

Table 6: Shows common Gene Ontology

Item	Present In	GO ID
Hydrolase activity, acting on ester bonds	CHD, HYP	GO:0016788
positive regulation of hydrolase activity	HYP, FHYP	GO:0051345
response to extracellular stimulus	CHD, HYP	GO:0009991
response to oxygen-containing compound	CHD, HYP	GO:1901700

Table 6 shows the common Gene Ontology functions in between Coronary Heart Disease and Hyperlipidemia. Hydrolase activity, acting on ester bonds, positive regulation of hydrolase activity, response to extracellular stimulus and response to oxygen-containing compounds were found to be common functions in between the Coronary Heart Disease and Hyperlipidemia with the occurrence of at least 2 of the particular category.

7. Discussions

T2DM is the most common type of diabetes, accounting for the vast majority of cases and T2DM develops when the amount of insulin in the blood is insufficient to sustain normal glucose levels. Insulin resistance and insufficient insulin secretion are the main flaws. Which of these is the most important is still up for debate, and it may differ from patient to patient. Many diabetes loci revealed using GWAS appear to affect insulin secretion or pancreatic beta-cell development, which suggests that loss of insulin secretion, is the main event that causes T2D. While T2D's microvascular consequences are a significant cause of morbidity, the macrovascular complications of coronary heart disease (CHD) and stroke are the leading causes of diabetes mortality. Diabetes has long been known as a risk factor for cardiovascular disease (CVD), with a 2- to 4-fold increased risk of death from CVD (Goodarzi & Rotter, 2020). Insulin resistance, according to Reaven, could be the cause of hypertension, dyslipidemia (particularly high triglyceride and low high density lipoprotein levels), poor glucose tolerance, and coronary heart disease(Reaven, 1997).

Knowledge of a disease's possible risk may be a vital first step toward taking action to reduce the disease's occurrence. Because persons with T2DM are at such a high risk of having CHD, it's critical to uncover factors that may influence their beliefs of the danger of acquiring CHD (Ammouri et al., 2018).

7.1 Phenotype Based Gene Analyser:

Based on illness or phenotype phrases input as free text, Phenolyzer outperforms rival approaches for prioritising Mendelian and complicated disease genes. To find the common genes associated with Coronary Heart Disease and Diabetes (Type-2 Diabetes Mellitus), the gene-disease-term interaction network of seed genes and predicted genes showing protein-protein interaction, transcription interaction, genes having same family or same Biosystem were evaluated (Fig. 11).

Phenolyzer finds and interprets the score relevant seed genes and a bar plot (Fig. 12) of the most highly ranked genes with normalized scores was obtained which shows that some of the high ranked genes have direct or indirect relationship with positive association of the causal of diseases called Coronary Heart Disease and Diabetes.

It also shows that some of the genes have relation with phenotypes like obesity, which has connection with Hyperlipidemia and that might become the risk factors for the Coronary Heart Disease.

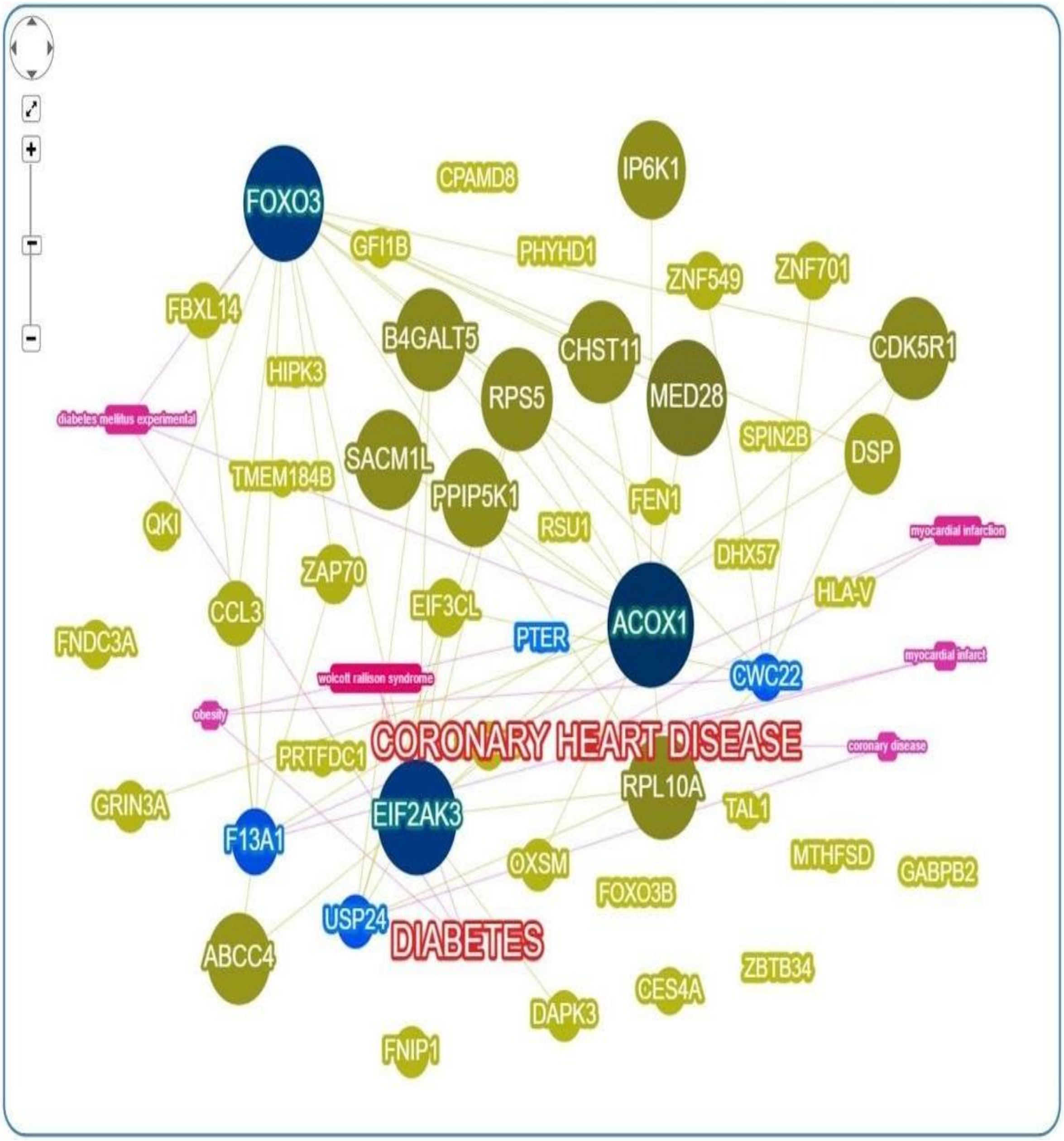
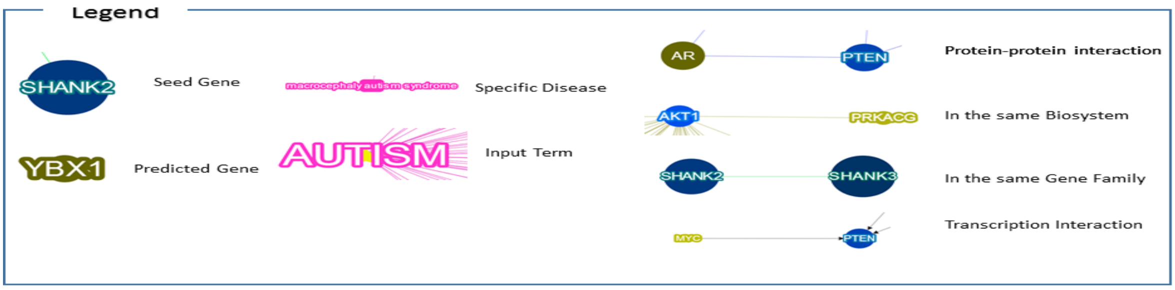


Fig. 11:Gene-Disease-Term Interaction Network

To understand the network, refer below:



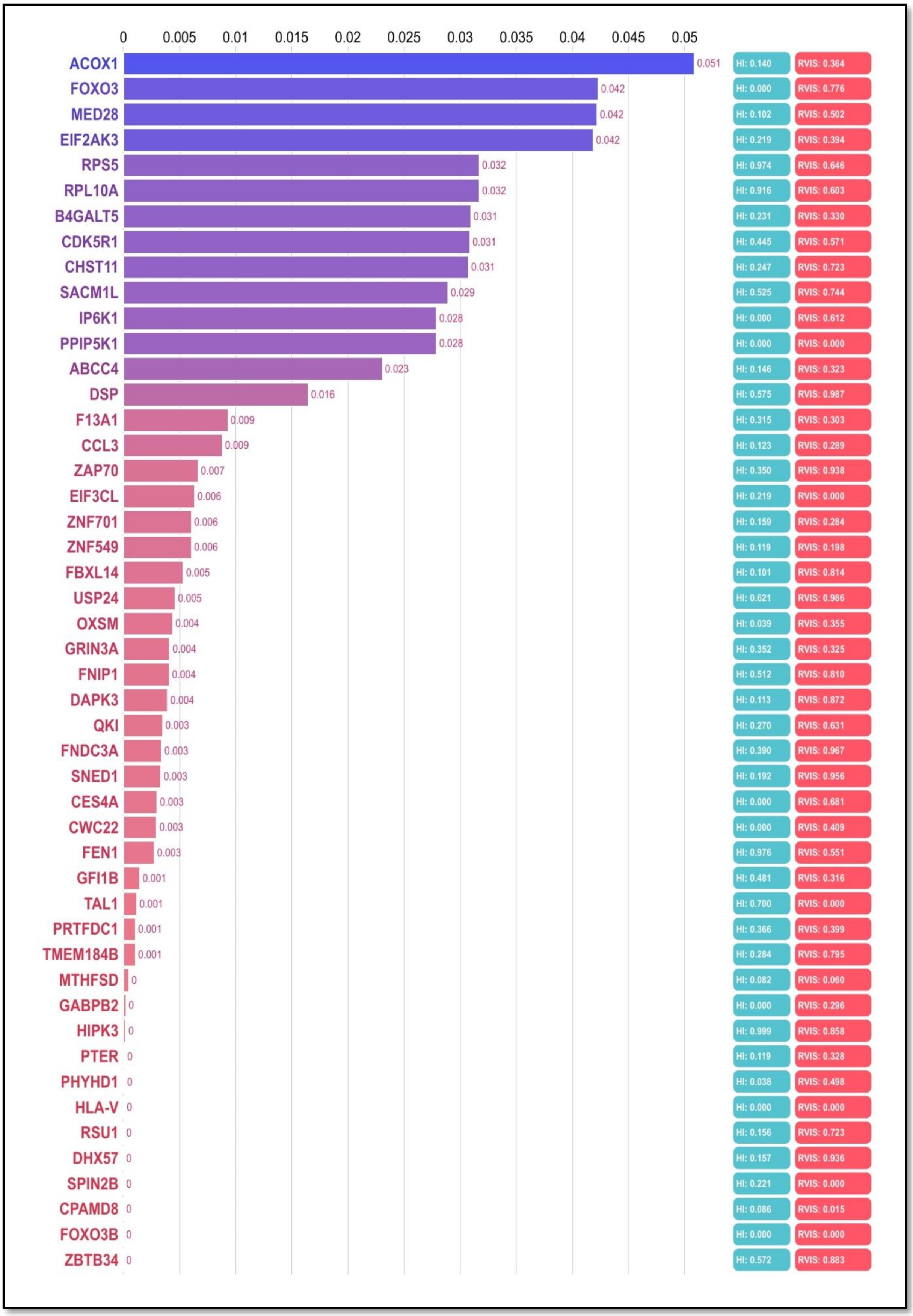


Fig. 12:Prioritizing candidate genes using the Phenolyzer.

Table 7: Genes having association with Diabetes, Coronary Heart Disease and Hyperlipidemia

Gene Name	Using KEGG Analysis		Study Accession in GWAS
eukaryotic translation initiation factor 2 alpha kinase 3	EIF2AK3	(CHD & FH - experimental) & T2D (KEGG - Disease)	
	Genome Wide Association Studies **(Comparison with the genes having occurrence of at least 2)**		
coagulation factor XIII A chain	F13A1	(CHD & FCHD - experimental) & T2D (GWAS)	GCST005166
Ras suppressor protein 1	RSU1	(CHD & FHYP - experimental) & T2D (GWAS)	GCST005173

From KEGG Disease database, EIF2AK3 was found to be linked with Diabetes (Polak & Cavé, 2007) which was also showing positive association with disease in the patients having Coronary Heart Disease and Familial Hyperlipidemia.

From Genome Wide Association Studies, it was noted that F13A1 and RSUI genes have interrelationship with Diabetes Mellitus (Almgren et al., 2017; Divers et al., 2017). Out of Differential Expressed genes found in experimental analysis of Coronary Heart Disease, Familial Coronary Heart Disease, Hyperlipidemia and Familial Hyperlipidemia comparable to GWAS studies, BCL9L, F13A1, FBF1, FGF1, FGD6, GRIN3A, NUS1, CFAP161, EPG5, GRIN3A, HIVEP2, MAPK81P3, SEC63 were found to have positive association with the causal of CHD which supports our analysis. The analysis shows that FI3A1, EIG2AK3 and RSUI might have the common gene association with the occurrence of CHD, Hyperlipidemia and Type-2 Diabetes Mellitus.

8. Conclusions

Overall 61 common Differentially Expressed Genes by keeping p-value < 0.05 and logfc ± 2, with the occurrence of at least two, were found using RNA-seq analyses of the patients suffering from Coronary Heart Disease, Familial Coronary Heart Disease, Hyperlipidemia and Familial Hyperlipidemia. 13 and 18 genes in CHD & FCHD, HYP & FHYP were found to be intersecting, respectively. In Gene Hub analysis, CWC22 and RPS5 were found to be the common hubs. Gene Ontology Gene Set Enrichment Analysis helped in characterising the functions of the DEG’s involved in CHD and HYP like Hydrolase activity, acting on ester bonds; Positive regulation of hydrolase activity; Response to extracellular stimulus and response to oxygen containing compound. Gene-Disease term interaction network and bar plot helped distinguish the genes having interrelationship between CHD, Diabetes and other phenotypes like Obesity. While comparing the experimental results with Diabetic genes using KEGG disease database and Genome Wide Association Studies, EIF2AK3, F13A1 and RSU1 were found to have common association in between Coronary Heart Disease, Hyperlipidemia and Diabetes. Hence, integrative RNA-seq data analysis and network biology provide some valuable insight into the mechanism of CHD, HYP and Diabetes, which in turn, may contribute to the development of novel drug targets or diagnoses. However, further validation is required in terms of molecular biology and in-depth studies for better understanding to elucidate the risk of coronary heart disease to treat more effectively and to find the drug therapeutic active sites of the common associated genes.

9. References

Almgren, P., Lindqvist, A., Krus, U., Hakaste, L., Ottosson-Laakso, E., Asplund, O., Sonestedt, E., Prasad, R. B., Laurila, E., Orho-Melander, M., Melander, O., Tuomi, T., Holst, J. J., Nilsson, P. M., Wierup, N., Groop, L., & Ahlqvist, E. (2017). Genetic determinants of circulating GIP and GLP-1 concentrations. *JCI Insight*, *2*(21). https://doi.org/10.1172/jci.insight.93306

Ammouri, A. A., Abu Raddaha, A. H., Natarajan, J., & D'Souza, M. S. (2018). Perceptions of risk of coronary heart disease among people living with type 2 diabetes mellitus. *International Journal of Nursing Practice*, *24*(1), 1–9. https://doi.org/10.1111/ijn.12610

Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., …Ostell, J. (2012). BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata. *Nucleic Acids Research*, *40*(D1), 57–63.https://doi.org/10.1093/nar/gkr1163

Bolger, A. M., Lohse, M., &Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Brittany A Potz, Anshul B Parulkar, Ruhul M Abid, Neel R Sodha, and F. W. S. (2018). Novel Molecular Targets for Coronary Angiogenesis and Ischemic Heart Disease. *Coronary Artery Dis.*, *28*(7), 605–613. https://doi.org/10.1097/MCA.0000000000000516.Novel

Brown, J., Pirrung, M., &Mccue, L. A. (2017). FQC Dashboard: Integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics*, *33*(19), 3137–3139. https://doi.org/10.1093/bioinformatics/btx373

Bułdak, Ł., Marek, B., Kajdaniuk, D., Urbanek, A., Janyga, S., Bołdys, A., Basiak, M., Maligłówka, M., &Okopień, B. (2019). Endocrine diseases as causes of secondary hyperlipidaemia. *EndokrynologiaPolska*, *70*(6), 511–519. https://doi.org/10.5603/EP.a2019.0041

Buniello, A., Macarthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research, 47(D1), D1005–D1012. https://doi.org/10.1093/nar/gky1120

Chin, C. H., Chen, S. H., Wu, H. H., Ho, C. W., Ko, M. T., & Lin, C. Y. (2014). cytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology*, *8*(4), S11. https://doi.org/10.1186/1752-0509-8-S4-S11

DarshanDoshi, Ori Ben-Yehuda, Machaon Bonafede, N. J., Dimitri Karmpaliotis, Manish A. Parikh, Jeffrey W. Moses, G. W. S., & Martin B. Leon, Allan Schwartz, A. J. K. (2016). Underutilization of Coronary Artery Disease Testing Among Patients Hospitalized With New-Onset Heart Failure. *JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY*, *68*(5). https://doi.org/10.1016/j.jacc.2016.05.060

Divers, J., Palmer, N. D., Langefeld, C. D., Brown, W. M., Lu, L., Hicks, P. J., Smith, S. C., Xu, J., Terry, J. G., Register, T. C., Wagenknecht, L. E., Parks, J. S., Ma, L., Chan, G. C., Buxbaum, S. G., Correa, A., Musani, S., Wilson, J. G., Taylor, H. A., … Freedman, B. I. (2017). Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes. *BMC Genetics*, *18*(1), 1–13. https://doi.org/10.1186/s12863-017-0572-9

Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., & Liu, S. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 100141. https://doi.org/10.1016/j.xinn.2021.100141

Furumichi, M., Sato, Y., & Ishiguro-watanabe, M. (2021). KEGG : integrating viruses and cellular organisms. 49(October 2020), 545–551. https://doi.org/10.1093/nar/gkaa970

Ginsberg, H. N. (2000). Insulin resistance and cardiovascular disease. *Journal of Clinical Investigation*, *106*(4), 453–458. https://doi.org/10.1172/JCI10762

Goodarzi, M. O., & Rotter, J. I. (2020). Genetics Insights in the Relationship between Type 2 Diabetes and Coronary Heart Disease. *Circulation Research*, *126*(11), 139–148. https://doi.org/https://doi.org/10.1161/CIRCRESAHA.119.316065

Hagve, T. A. (1988). Effects of unsaturated fatty acids on cell membrane functions. *Scandinavian Journal of Clinical and Laboratory Investigation*, *48*(5), 381–388. https://doi.org/10.1080/00365518809085746

Handerson, A. (1996). Coronary heart disease- overview.*The Lancet.*

Hulbert, A. J., Turner, N., Storlien, L. H., & Else, P. L. (2005). Dietary fats and membrane function: Implications for metabolism and disease. *Biological Reviews of the Cambridge Philosophical Society*, *80*(1), 155–169. https://doi.org/10.1017/S1464793104006578

Karr, S. (2017). Epidemiology and management of hyperlipidemia. *The American Journal of Managed Care*, *23*(9), S139–S148.

Kim, D., Paggi, J. M., Park, C., Bennett, C., &Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Lakka, H. M., Lakka, T. A., Tuomilehto, J., Sivenius, J., &Salonen, J. T. (2000). Hyperinsulinemia and the risk of cardiovascular death and acute coronary and cerebrovascular events in men: The Kuopio Ischaemic Heart Disease Risk Factor Study. *Archives of Internal Medicine*, *160*(8), 1160–1168. https://doi.org/10.1001/archinte.160.8.1160

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(SUPPL. 1), 2010–2012. https://doi.org/10.1093/nar/gkq1019

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/s13059-014-0550-8

Martín-Timón, I. (2014). Type 2 diabetes and cardiovascular disease: Have all risk factors the same strength? *World Journal of Diabetes*, *5*(4), 444. https://doi.org/10.4239/wjd.v5.i4.444

Owens, A. P., Byrnes, J. R., &Mackman, N. (2014). Hyperlipidemia, Tissue Factor, Coagulation and Simvastatin. *Trends Cardiovasc. Med.*, *24*(3), 95–98. https://doi.org/doi:10.1016/j.tcm.2013.07.003

Pagliaro, B. R., Cannata, F., Stefanini, G. G., & Bolognese, L. (2020). Myocardial ischemia and coronary disease in heart failure. *Heart Failure Reviews, 25*(1), 53–65. https://doi.org/10.1007/s10741-019-09831-z

Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski and Trey Ideker. (1971). Cytoscape: A Software Environment for Integrated Models. *Genome Research*, *13*(22), 426. https://doi.org/10.1101/gr.1239303.metabolite

Perdoncin, E., &Duvernoy, C. (2017). *Treatment of Coronary Artery Disease in Women.pdf* (pp. 201–208). Methodist DebakeyCardiovasc J.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., &Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/nbt.3122

Polak, M., & Cavé, H. (2007). Neonatal diabetes mellitus: A disease linked to multiple mechanisms. *Orphanet Journal of Rare Diseases, 2*(1), 1–11. https://doi.org/10.1186/1750-1172-2-12

Reaven, G. M. (1997). Role of insulin resistance in human disease. *Nutrition*, *13*(1), 64. https://doi.org/10.1016/s0899-9007(96)00380-2

vonMering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., &Snel, B. (2003). STRING: A database of predicted functional associations between proteins. Nucleic Acids Research, 31(1), 258–261. https://doi.org/10.1093/nar/gkg034

Yang, H., Robinson, P. N., & Wang, K. (2015). Phenolyzer: Phenotype-based prioritization of candidate genes for human diseases. *Nature Methods, 12*(9), 841–843. https://doi.org/10.1038/nmeth.3484

Yao, Y. S., Li, T. Di, & Zeng, Z. H. (2020). Mechanisms underlying direct actions of hyperlipidemia on myocardium: An updated review. *Lipids in Health and Disease*, *19*(1), 1–6. https://doi.org/10.1186/s12944-019-1171-8

Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, *16*(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., …Flicek, P. (2020). Ensembl 2020. *Nucleic Acids Research*, *48*(D1), D682–D688. https://doi.org/10.1093/nar/gkz966

Zhou, X., Zhang, W., Liu, X., Zhang, W., & Li, Y. (2015). Interrelationship between diabetes and periodontitis: Role of hyperlipidemia. *Archives of Oral Biology*, *60*(4), 667–674. https://doi.org/10.1016/j.archoralbio.2014.11.008