**Correlating Diabetes with Healthy and Unhealthy Heart:**

**Methodology:**

1. Extraction of Raw Reads:

Replaced by NCBI’s Genome Project Database, Bioproject organizes metadata of large volume of data and encircle biological data derived from single organization or from cohort of coordinating organizations associated with single initiative. Thereby, serves as a central portal for the submitted data represents higher order organisation, description and classification across several NCBI archival databases (Barrett et al., 2012). 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 finalized.

2. Sequence Retrieved Archive Toolkit:

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). With the help of SRA toolkit, the .sra files of the 15 biosamples were retrieved using command line terminal.

Commands:

>prefetch.2.10.9 accessionid\_of\_sra\_files #download the .sra files

>fastq\_dump.2.10.9 -v accessionid.sra #convert .sra file to fastq

>fastq\_dump.2.10.9 -v --split\_files accessionid.sra #splitting of the paired end files

>fastq\_dump.2.10.9 –v --split\_files -gzip accessioned.sra #converting .sra files to .gzip compressed format

3. Quality Assurance - I:

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). Using the .gzip format files .html file report of fastQC was generated using the following command –

Command:

>fastqc accessionid.fasta.gzip\_1 accessionid.fasta.gzip\_2 #two gzip files as the paired ends were splitted into forward and reverse strand.

Using the .html result file, Quality assurance of base sequence was done by observing the whisker plot with stretched outs peaks symbolized the mean of quality. The presence or absence of the adaptor sequences was visualised.

4. Trimming of Raw-reads:

Trimmomatic 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 seqences, trimmomatic was used to trim the sequences below a threshold value.

Command:

Java -jar trimmomatic-0.39.jar PE -threads 6 -phred64 path/input\_R1.fastq.gz path/input\_R2.fastq.gz output\_trimmed\_R1\_paired.fastq.gz output\_trimmed\_R1\_unpaired.fastq.gz output\_trimmed\_R2\_paired.fastq.gz output\_trimmed\_R2\_unpaired.fastq.gz LEADING:3 TRAILING:3 MINLEN:36

It provides 4 files, 2 paired and 2 unpaired whereas unpaired files give information of the sequences which has been trimmed. Hence, the focus is only on the paired sequences.

5. 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 Precision scoring matrix FRED scale.

>25 = absolutely good (Green zone)

20-25 = good (yellow zone)

<20 = relatively poor (pink zone)

6. Indexing and Sequence Alignment:

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 entire human genome with very low memory requirements (Kim, Paggi, Park, Bennett, & Salzberg, 2019). 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.

Command:

>./hisat2 –p 4 –x path/to/genome.tran.file/genome\_tran\_file –dta -1 path/of/forward\_paired/forward\_pairedfile -2 path/to/reverse\_paired/reverse\_pairedfile –S path/to/new\_file/newfilename.sam -t

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.

7. Conversion of .sam to .bam files:

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 streambased 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).

Commands:

>samtools view –S –b path/to\_file.sam/file.sam > file.bam #conversion of .SAM to .BAM

>samtools sort -@ 5 path/to\_file.bam/file.bam path/to\_file.sam/file.sam #sorting of BAM files

The sorted BAM files were generated which were further used for the quantification purposes.

8. Assembly of the transcripts:

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). 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) .

Commands:

>stringtie path/to/sorted.bam –l accession\_id –p 8 –G path/human\_genome.gtf –o path/to/file.gtf #mapping of the transcripts

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 gtf file. Later, the file was given name “mergelist.txt”. Note that any name could be given.

Command:

>stringtie -merge -p 5 -G path/to/human\_genome.gtf -o stringtie\_merged.gtf path/to/mergelist.txt #merging of all the transcripts in gtf files with human genome

The output stringtie\_merged.gtf gave the information of mapped transcripts. Further, the head file and the number of transcripts were checked using the following command lines

Commands:

>cat path/to/stringtie\_merged.gtf | head #head file

>cat path/to/stringtie\_merged.gtf | grep –v “^#” | awk ‘$3==”transcript” {print}’ |wc –l #number of transcripts

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.

Command:

Stringtie –e –B –p 5 –G path/to/stringtie\_merged.gtf path/to/sortedfile.bam –o path/to/stringtie\_samplefile.gtf #abundance of the transcripts

Hence, the stringtie provides 6 output files for each sample, i.e., 5 .ctab files and 1 stringtie\_file.gtf.

In the RNA-seq analysis, for the finding of the differential expressed genes, stringtie\_file.gtf was used. Furthermore, using the python script available on stringtie webpage, the CSV files of the transcripts having read counts were generated that are used to find the differential expressed genes. Python script is freely available on-

http://ccb.jhu.edu/software/stringtie/dl/prepDE.py

>python prepDE.py –i sample\_lst.txt

Sample\_list.txt is the file having accession id and paths of all the generated gtf files. For example:

accessionid path/to/file.gtf

9. Library Normalisation:

The DESeq 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).

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 log2FoldChange = +-2, the upregulated and downregulated genes were evaluated.

With the help of the script given below, the comparison analysis was done with respect to Healthy VS. Coronary Heart Disease samples, Healthy VS. Hyperlipidemia samples, Healthy VS. Familial Hyperlipidemia samples and Healthy VS. Familial Coronary Heart Disease samples. PHENO\_DATA.txt is a text format file having Accession ID’s, Category (control vs treatment) and Population type (Healthy, Coronary Heart Disease, Hyperlipidemia, Familial Coronary Heart Disease or Familial Hyperlipidemia).

DESeq2 R analysis commands:

#Import DESeq2 library in R

> library("DESeq2")

#To load gene(/transcript) count matrix and labels

> countData <- as.matrix(read.csv("gene\_count\_matrix.csv", row.names="gene\_id"))  
> colData <- read.csv(“PHENO\_DATA.txt”, sep="\t", row.names=1)

#Check all sample IDs in colData are also in CountData and match their orders

> all(rownames(colData) %in% colnames(countData))  
[1] TRUE  
> countData <- countData[, rownames(colData)]  
> all(rownames(colData) == colnames(countData))  
[1] TRUE

#Create a DESeqDataSet from count matrix and labels

> dds <- DESeqDataSetFromMatrix(countData = countData,  
        colData = colData, design = ~ CHOOSE\_FEATURE) #Population=Feature

#Run the default analysis for DESeq2 and generate results table

> dds <- DESeq(dds)  
> res <- results(dds)

#Sort by adjusted p-value and display

> (resOrdered <- res[order(res$padj), ])

#To write the output in csv file format

>write.csv(as.matrix(resOrdered), “file.csv”)

#To read the csv file

>read.csv(“file.csv”)

**References:**

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

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

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

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

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

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