UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE DEL FARMACO

CORSO DI LAUREA MAGISTRALE IN PHARMACEUTICAL BIOTECHNOLOGIES - BIOTECNOLOGIE FARMACEUTICHE

MASTER THESIS

RNA-SEQ analysis of differential gene expression patterns in subcutaneous adipose tissue biopsies from people with low vs normal birth weight – implications for risk of developing type 2 diabetes.

RELATORE: Prof. Stefano Comai

CORRELATORE: Dr. Allan Vaag

Mrs. Sofie Olund Villumsen

Prof. Rashmi Prasad B

LAUREANDO: PRABHUDEVA THUMMALA

ANNO ACCADEMICO: 2022-23

Table of Contents

Preface and acknowledgements

This Thesis Project is a part of my Master's degree in Pharmaceutical Biotechnologies at the University of Padova, Italy. I worked on this project, conducted at the Steno Diabetes Center Copenhagen (SDCC) within the group of Translational Type 2 Diabetes (T2D) Research. This is to analyse data from the high carbohydrate overfeeding (HCOF) study managed by Allan Vaag and Charlotte Brøns at SDCC, Copenhagen. I am very grateful for this opportunity to perform the RNA-Sequencing (RNA-seq) analysis on adipose tissue from the cohort study. The cohort consists of normal and low birth weight individuals prone to developing T2D later in life and matched controls. This thesis focuses on the baseline RNA-seq data from the HCOF Study. During the period of this project, I developed my skills within Linux, RStudio, edgeR (RStudio package), Reactome and the David-data base for data wrangling and analysis. I improved my skills to work in the research collaborations and strengthened my knowledge in bioinformatical analyses. I gained much more knowledge within metabolic research in T2D.

I would like to give a great thanks to my external supervisor Allan Vaag for the professional guidance and for welcoming me so kindly in the research group. Heartful Thanks to our research collaborative co-supervisor Rashmi Prasad and Sofie Olund Villumsen for the training and support. I also would like to thank my internal supervisor Stefano Comai, co-supervisor Giovanni Minervini for the kind supervision and guidance throughout the project. Finally, I extend my thanks to Charlotte Brøns, Sidsel Seide Gertsen and Line Ohrt-Elingaard for providing further support and a very welcoming environment at the SDCC. I am truly thankful for the supportive people around me and everything I have learned during the project period.

Abstract

Background:

Fetal development and programming has lifelong implications for heath and risk of disease, and both overnutrition and undernutrition is known to cause fetal adaptations and developmental changes via epigenetic mechanism. Such adaptations can play important roles in perspective to metabolic disorders including risk of type 2 diabetes (T2D). Low birthweight (LBW) may also result in reduced adult height, increased abdominal obesity and various metabolic risk factors including non-alcoholic fatty liver disease (NAFLD), which is on the path to development of T2D. Metabolic changes in subcutaneous adipose tissue (SAT) of LBW individuals has taken a vivid role in causing the variation of metabolic traits. In this study, SAT gene expression patterns were compared between age- and BMI matched LBW and normal birthweight (NBW) aged 37 years.

Objective:

To compare expression levels of RNA sequencing from SAT biopsies between LBW and NBW subjects, and thereby to understand the molecular mechanisms underlying increased risk of T2D in people born with LBW.

Methods:

A total of 133 samples were analysed via RNA sequencing, which includes 85 adipose tissue samples (i.e., from baseline, overfeeding and randomization) and 48 preadipocyte samples. For my thesis, the analysis of only the adipose tissue samples from baseline biopsies was included. Non-stranded and polyA-selected mRNA library preparation has been done on all samples, followed by PE100 sequencing resulting Fastq files. The pipeline included FastQC tool for quality check, STAR for Alignment, Featurecounts for quantifying, edgeR for the differential expression analysis. Pathway analysis was done using Reactome and David database.

Results:

FastQC reports were generated and the data was in good quality and met the standards. After the alignment and quantification, GeneCounts file with a total of 60483 genes was obtained. Among the groups of LBW (n=17) vs NBW (n=12) we found 50 significantly different gene expressions p-value < 0.05 (without adjusting for multiple testing), when all features (all *genes, non-codingRNA, small RNA, pseudogenes)* are considered. In contrast, 31 significant (p-value < 0.05) differential gene expression levels were found when only protein coding genes were considered. Gene ontology results were obtained for both downregulated and upregulated genes in LBW compared to NBW. Pathway analysis identified significant differences to involve metallothionein bind metals, response to metal ions regulation of complement cascade and peptide ligand-binding receptors. Network analysis of these results shows the genetic interactions within the areas of signal transduction, metabolism, gene expression in developmental biology and associated networks.

Conclusion:

Differential SAT gene expression levels were identified between LBW at increased risk of T2D compared with matched NBW controls, which however did not persist after FDR. Interestingly, genes involved the processes of ion homeostasis, apoptotic process, cellular response to stimuli and stress were found among the significant positive log fold change (logFC) - upregulated genes. Whereas, significant negative logFC genes – downregulated were seen in the pathways related to lipid metabolic process, cholesterol homeostasis, steroid and glycoprotein metabolic pathways. These differences may play a role for the increased risk of T2D in LBW subjects.

Abbreviations

BW Birth Weight

BMI Body Mass Index

CPM Count Per Million

DM Diabetes Mellitus

DE Differential Expression

DKD Diabetes Kidney Disease

FDR False Discovery Rate

FC FeatureCounts

GO Gene Ontology

GC Guanine - Cytosine

HCOF High Carbohydrate Over Feeding

LRT Likelihood Ratio Tests

NBW Normal Birthweight

logFC Log Fold Change

LBW Low Birthweight

NB Negative Binomial

NGS Next-Generation Sequencing

NAFLD Non-Alcoholic Fatty Liver Disease

QL Quasi Likelihood

RNA-seq RNA-Sequencing

STAR Spliced Transcripts Alignment to a Reference

SAT Subcutaneous Adipose Tissue

T2D Type 2 Diabetes

1 Introduction

Currently, there are over 537 million adults (aged 20-79 years) living with diabetes, which corresponds to approximately 1 in 10 adults globally. It is predicted that this number may rise to 643 million (1 in 9 adults) by 2030 and further increase to 784 million (1 in 8 adults) by 2045 (1).

T2D is the most common form of diabetes, accounting 90% of cases. In addition to genetic factors, various other factors such as obesity, dietary composition, lifestyle choices, physical inactivity, and exposure to an adverse fetal environment contribute to the development of T2D (2,3). Birth weight (BW) serves as an indicator of fetal growth and has a profound influence on subsequent phenotypical changes, including height, size, muscle mass, fat deposition, and metabolic and skeletal alterations (2-5). It is well documented that there is strong association between the BW and T2D (2-8). LBW individuals exhibit distinct physical changes such as increased abdominal fat (6) and reduced insulin secretion (7). Moreover, they display altered expression of insulin signalling proteins in muscle and fat tissues (8). Consequently, these biological changes in LBW individuals are closely linked to the development of conditions such as T2D, hypertension, and cardiovascular disease (6-11). Similarly, the presence of LBW family history suggests that it may have a potential hereditary influence on the likelihood of experiencing LBW in future generations (12).

Understanding the long-term causal effect of LBW to determine various disease risks has importance in assessing the individual risk factors for T2D. This presents an opportunity to implement early nutritional interventions that can mitigate the risk of disease burden in the future (13). In addition to the findings from published studies, investigating the link between environmental factors such as rapid changes in the diet and patterns of gene changes offers valuable insights into understanding the epigenetic determinants and their influence on BW outcomes. The early development process is influenced by multiple factors and can be affected by an unfavourable fetal environment (14). Individuals born with LBW have shown an imbalanced pattern in their hormonal responses (15), which in turn increases their susceptibility to conditions such as obesity, T2D and other diseases. The changes in genes expression and methylation patterns play a crucial role in regulating metabolism through the central nervous system (15).

During intrauterine (IU) development, regulatory mechanisms work to maintain homeostasis, but they can be compromised by factors such as aging, obesity, or other influences. Any abnormal modifications to these regulatory mechanisms can lead to disruptions in insulin physiology and contribute to the development of insulin resistance. Sometimes, these changes can have long-lasting effects on the physiology and metabolism of offspring.

BW of an individual is strongly associated with i) environmental influences that contribute to phenotypic associations and induce epigenetic modifications in the genome ii) indirect effect of maternal genotypes and shared genetic effects between mother and offspring (16). By exploring and understanding the potential connections between particular genetic variants and levels of gene expression, we can gain valuable insights into molecular mechanisms in the progression of T2D. Establishing a comprehensive profile of significant correlation between gene expression levels and their regulation pattern requires specific research studies and analysis (17). This study is to identify the distinct genes and their expression patterns within SAT that differentiate individuals with LBW from those with NBW. Identification of such gene expression differences helps to understand the behaviour of these fat cells within the two groups. The idea is also to find whether the metabolism of individuals with LBW can return to its original state after an overfeeding followed by an exercise intervention. The underlying hypothesis is that the metabolism of LBW deviates from that of individuals with NBW. This distinct metabolism in LBW individuals may make them more vulnerable to developing T2D later in life compared to those with NBW.

My thesis focuses on differences in gene expression patterns in SAT biopsies obtained from NBW and LBW individuals. Bulk RNA has been collected and total RNA isolated and sequenced by BGI sequencing. The sequenced reads were further proceeded for differential analysis. And finally, we performed pathway analysis of the significant genes.

2. Objective

The impact of low birth weight on metabolic and genetical traits associated with risk of developing T2D when exposed to an affluent lifestyle paves the way of the study design. The primary objective of this project design is to examine whether HCOF have distinct negative metabolic effects on LBW subjects compared to those with NBW as control group. Additionally, the project aims to assess if LBW individuals exhibit reduced expandability of their SAT and an increased potential for preadipocyte proliferation and/or differentiation. This project also focuses if exercise can revert and/or minimize the deleterious cardiometabolic effects of HCOF in individuals with or without increased risk of T2D.

Overall, the project of this study contains a large amount of data and involves many sub groups.

For my thesis, I aim to analyse the RNA-seq data from the baseline samples of both LBW and NBW groups. This includes the following aims:

1) To study the differential gene expression of adipose tissue at baseline level between the groups LBW and NBW. Further, to make a pathway analysis and connect the function of significant genes to their metabolic biological processes.
2) Understand the differences of the significant genes, their biological processes and related pathways.

3. Background

3. 1 Type 2 diabetes

Development of T2D is mainly due to i) inadequate secretion of insulin by pancreatic β-cells or ii) decreased response from tissues to insulin. As a consequence, the progression of T2D disrupts the regulation of glucose level in the body resulting in high blood sugar level, known as hyperglycaemia. The presence of obesity, particularly central visceral adiposity plays a crucial role in development of T2D. Persistent elevation in the level of blood glucose or post-meal hyperglycaemia following carbohydrate intake are characteristic features of T2D (17). Endogenously, three different hormones glucagon, epinephrine, and cortisol are known to increase glucose levels by promoting biological processes such as glycogenolysis and gluconeogenesis. Also, dietary carbohydrate intake is an important exogenous factor that increase blood glucose levels (17). Previous studies have demonstrated that adopting a regular exercise routine and maintaining a healthy diet can effectively reduce the risk of developing T2D (20). On the other hand, certain non-modifiable risk factors like ethnicity, family history, genetic pre-disposition have strong genetic basis in T2D. These factors are largely determined by an individual's genetic makeup and are not easily influenced by external factors or lifestyle changes. Understanding both gene changes and non-genetic risk factors which could influence the risk of developing T2D can help take appropriate preventive measures when necessary.

Etiology of T2D is influenced by combination of genetic factors, the metabolic processes and the environmental factors. T2D has strong hereditary connections and T2D susceptibility genes are more common in the general population, which limits the explanation on total estimated heritability of T2D. This suggests that there may be additional unidentified T2D susceptibility genes with a greater influence on the risk of developing T2D in the general population (18). In recent years, extensive genome-wide association studies have provided evidence for the polygenic nature of T2D. Insulin resistance in T2D is linked to the malfunctioning of adipose tissue and the generation of free fatty acids within it. Patients with T2D have been found to have downregulation of genes involved in oxidative metabolism (20). In previous studies, several genes including TCF7L2, PPARG, FTO, KCNJ11, NOTCH2, WFS1, CDKAL1, IGF2BP2, SLC30A8, JAZF1, HHEX along with more than 600 single nucleotide polymorphisms (SNP’s) were discovered to be more significant in individuals having T2D (19). For instance, KCNJ11 gene, which is involved in the normal functioning of pancreatic beta cells responsible for insulin production and release, and the TCF7L2 gene, which plays a role in glucose metabolism and the production of glucagon-like peptide-1.

3.2 Low birth weight and health risks

Deviations from healthy birth weight involve wide range of subsequent adverse outcomes and traits. In general, BW less than 2500 grams irrespective of gestational age is considered as LBW as per WHO. There are so many etiological causes of LBW such as intrauterine growth retardation (IUGR), preterm birth, fetal inadequate nutrition, congenital anomalies and many other fetal, maternal conditions (15). Infants with LBW are more likely to develop complications and have risks of cardiovascular disorders, metabolic disorders, cognitive deficits, motor delays, cerebral palsy and others (15, 27). The limited supporting evidence on BW as a reliable marker for assessing the intrauterine environment in relation to subsequent health and disease has given an extensive scope for further investigation into its potential implications.

The reduction in the weight of fetus or infant is due to several mechanisms. How does being born with LBW increase matters in the risk of developing certain diseases fifty or more years later? Many of the changes that occur during developmental stages have direct effects on physiological conditions later in life. For example, inadequate nutrition or overnutrition during early development can lead to metabolic alterations. Inadequate nutrition in the womb, resulting in restricted fetal growth and development is associated with LBW. This nutritional insufficiency can lead to metabolic alterations and long-term changes in the body's physiology. LBW individuals often exhibit metabolic adaptations including alterations in insulin sensitivity, glucose metabolism, and lipid metabolism which can contribute to an increased risk of developing metabolic disorders, including T2D. Until this date, many evidence based studies links epigenetic factors with human diseases and these epigenetic factors mediates activation, repression or silencing of genetic transcription (15).

Studies by Plagemann et al. done in animal models states that over-nutrition in pre-and/or neonatal period can lead to alterations in DNA methylation patters of genes which are involved with regulation of appetite, body weight and metabolism. This causes the neonates to acquire adipogenic and diabetogenic phenotypes (15).

3.3 Thrifty Phenotype Hypothesis

Many years ago, Hales and Barker proposed the thrifty phenotype hypothesis (31). Poor fetal and infant nutrition are the base for pathological changes associated with the risk of glucose intolerance and insulin resistance in later life. Maternal malnutrition can cause poor fetal growth and infant developmental changes. The impact of other maternal and placental abnormalities influences the fetal growth as well. As per the thrifty phenotype hypothesis, poor fetal nutrition leads to an improper growth of the pancreatic β-cell mass or decrease in the islet of Langerhans' function (causing the impairment in insulin secretion). Progressively, this results in the glucose intolerance and the insulin resistance accounting for T2D followed by metabolic syndrome. These changes are also depending on risk factors like obesity, physical inactivity and other comorbid conditions.

3.4 Low Birth weight and gene changes in Diabetes

So far, we discussed that weight at birth play prominent role in the development of adult disease risk. The studies of utero genes and the pathways related to the birth weight, obesity are important to understand the long-term health outcomes. The genes mediating the mechanisms of controlling these clinical outcomes and their associated pathways are often complex. It is easier to drive the process of understanding this complexity from the fetal tissues compared to adult tissues. Transcriptomic data from various tissues in utero development helps to identify these genes and integrate with the genetic predisposition for various traits. Transcriptional level analysis of these genes also plays an important role in linking the disease with the profile of these genetic variants and polymorphisms. Genome Wide Association Studies have challenges to detect the causal effects of underlying genes in different tissue types and developmental stages (34).

The causal effects of these genetic risks are involved in controlling the utero expression patterns. Defects in any of these genes responsible for the pancreatic development and the insulin supports the development of T2D. This triggers the need to identify the genetic regulators of the weight at birth. Expression patterns of genes in fetus have shown influence in many metabolic mechanisms. For example, the polymorphism in G protein 3 subunit gene is linked to LBW in pregnancy (33). Genes encoding IGF-I, IGF-II, insulin, and their respective receptors could relate to BW (33). ADCY5 gene associated both with BW and

T2D. Genes like TCF7L2 has now been shown to modulate pancreatic islet function (35). T2D risk alleles at genes; HHEX-IDE and KCNQ1 show similar effects to ADCY5 and CDKAL1 in being associated with LBW. Identification and analysis of such genes is very critical in predicting the maintenance of glucose homeostasis, pancreatic beta cell function, early onset of T2DM and to reveal mutational effects. PPARG and KCNJ11 encodes a protein that acts as a target for classes of therapeutic agents widely used in diabetes management (35). This information could be used to intervention studies for developing and improving rational therapeutical targets. In this study, we want to corelate the gene expression levels and their functions towards the associated diabetic risks in LBW compared to NBW.

3.5 RNA-Sequencing

Overview:

RNA-seq involves the quantification of RNA in a biological sample such as from the adipose tissue at the cellular transcriptome level using next-generation sequencing (NGS). After NGS, the sequenced raw reads are checked for the quality. This quality control step is done using FastQC or FastQScreen or FASTX. If any criteria exist that impact the quality of reads, this should be removed. Tools like Skewer, Cutadapt or Trimmomatic are useful to cut adaptors/primer and trim reads with low quality. Finest quality trimmed reads are processed for mapping and then aligned to the genome. This step involves either mapping against reference genome or to the transcriptome. For the genome, the Splice-aware aligners like STAR, Tophat2, HISAT2 are used while the aligners (Bowtie2, BWA, GEM) and Quasimappers (Salmon or Kallisto) are used for the transcriptome.

After the alignment, next step is quantification using gene annotation file by the applications like FeatureCounts, RSEM, eXpress to produce the counts. Further, the counts obtained are used for differential expression and functional analysis. Thus, we can derive the analysis of biological process and pathways corresponding to the sequenced reads.

RNA-seq allows to study the expression of genetic changes in different stages of development and understand the biological pathways resulting in disease progression. From this process, the genes that cause various interesting differences could be detected. Comparing the expression patterns of adipose tissues from LBW and NBW individuals can help us identify and understand the genes which behave differently between the groups.

4 Methods

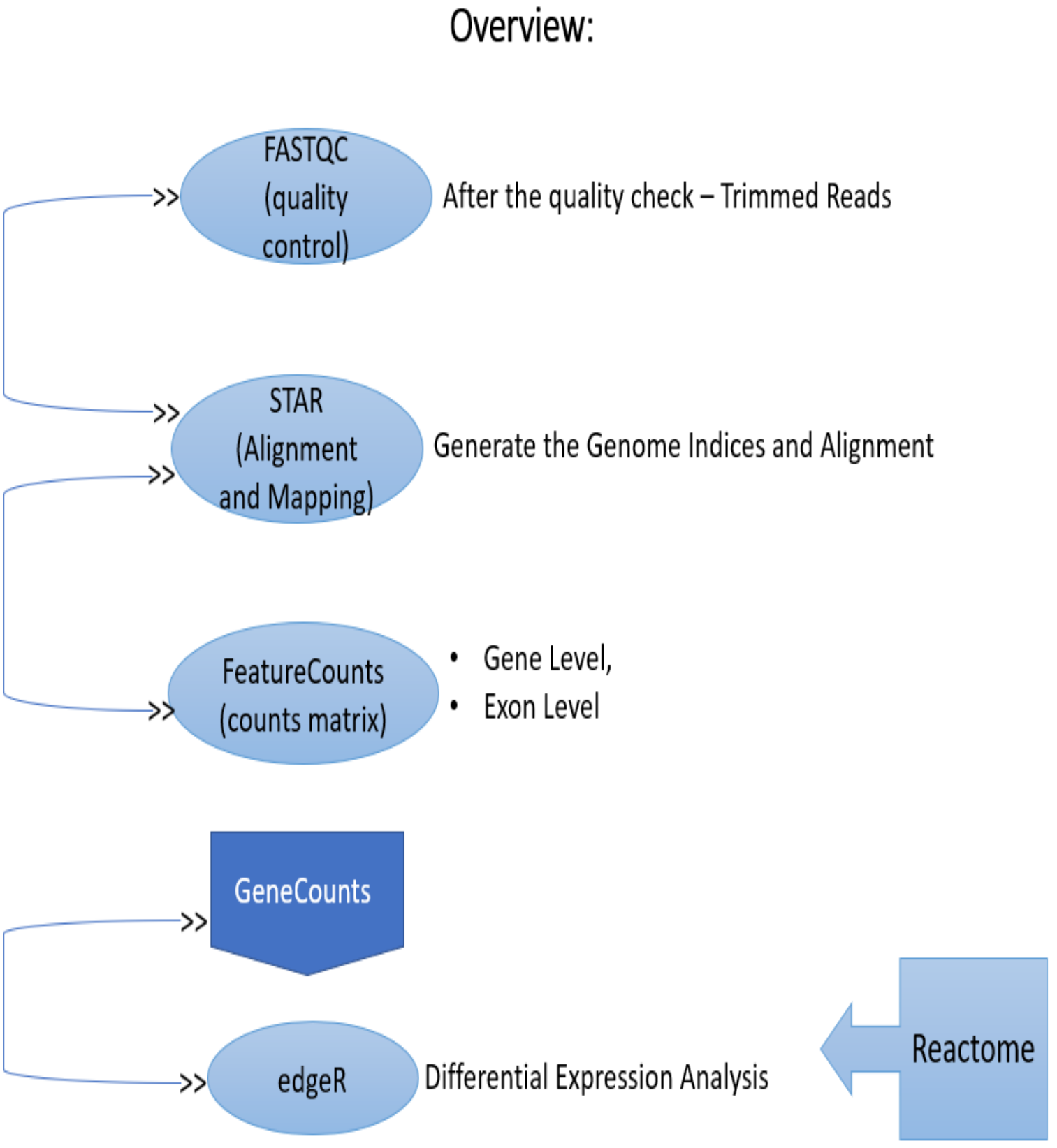


Fig: 4 (i) Overview of RNA-seq analysis pipeline

The methodology to analyse RNA-seq data in this study involves various levels like, library preparations, sequencing of reads and, pre-processing of data, aligning them to a reference genome and quantification to counts. Finally, further analysis of these counts can be done by differential expression analyses. Each step has a different selection of tools as shown in fig 4. (i) and each step should be considered carefully when setting up the pipeline for data analysis of RNA-seq.

4.1 Study Outline:

This study includes healthy Caucasian males born at term (weeks 39-41) in 1979-1980 with LBW (birth weight < 10^{th} percentile) and BMI, age-matched NBW control individuals (BW of 50 to 90^{th} percentile). Subjects with a family history of diabetes and/or a self-reported high physical activity level (>10hrs /week) were

excluded. Also, those who have lost/gained more than 3 kg within the past 6 months or those who consume alcohol (drink more than general recommendations) or substance abusers were excluded. All subjects were screened for current and previous health status to ensure eligibility. Blood samples, blood pressure, and electrocardiogram were obtained to ensure good health of the participants. All participants report to Rigshospitalet, Copenhagen -Denmark, where different tests were performed. Tissue biopsies at the baseline state were obtained from the abdominal.

4.2 Sample preparation and RNA sequencing at BGI:

Total of 29 adipose tissue samples i.e., 17 from LBW and 12 from NBW were analysed via RNA-seq. Total RNA was isolated from all sample preparations (amount ≥ 200ng, concentration 1000ng/microL≥c≥10 ng/microlitre, quality RIN/RQN value ≥ 7.0). mRNA enrichment and purification: Oligo dT Selection to enrich the mRNA or rRNA depletion (For total RNA extracted from whole blood, globin mRNA is depleted). The experimental pipeline shown in fig 4. (ii) was used:

- RNA fragment and reverse transcription (For stranded specific mRNA libraries second-strand cDNA synthesis with dUTP instead of dTTP)
- End repair, add A and adaptor ligation
- PCR
- Single strand separation and cyclization
- DNA nano ball synthesis
- Sequencing on DNB-seq platform

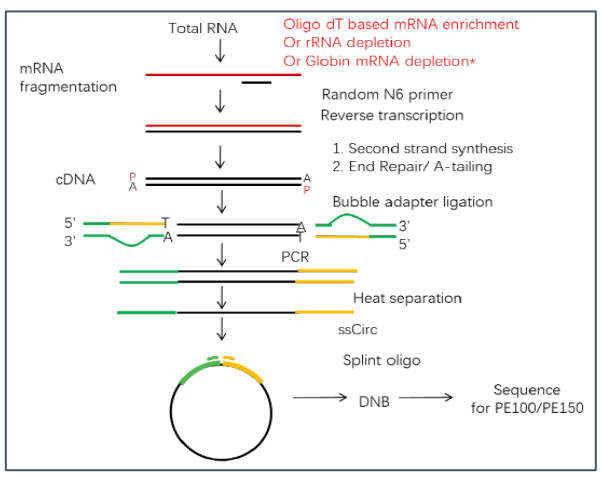


Fig: 4.2.(i) Experimental pipeline of Transcriptome.

Non-stranded and polyA-selected mRNA library preparation has been done on all samples, followed by PE100 sequencing with 4GB clean data per sample on DNBSEQ. After sequencing raw reads were filtered, which includes removing adaptor sequences, contamination, and low-quality reads from raw reads (replicate their results). Resulting fastq files from paired end reads, were received on a hard drive.

4.3 RNA-seq pipeline

Setting up RNA-seq pipeline

We used Linux for analysing the RNA-seq data. Initially, Linux commands and deep study of various literature was done to be able to set up the pipeline. Working on the Linux server and choosing the right selection of tools were most important among these tasks. Learning the issues on accessing the server and working with data on server is much crucial to be able to deal with the data with no harm to other files. Proper training was given to me on how to access the

servers and run the commands on Linux. Next step was to gain knowledge of the R language to use the package edgeR for differential analysis.

The systemic approach for the RNA-seq pipeline in this analysis was using the following tools: i) Spliced Transcripts Alignment to a Reference (STAR) was used for mapping and alignment ii) Subread Package (FeatureCounts) for quantification iii) edgeR for differential analysis iv) Reactome and David Database for the network and pathway analysis.

Setup of Pipeline in Bash file: Usually, this file consists of script in bash format (in file of .sh). It requires the following steps, i) Commands for making required directories for various output files in each step ii) Create the genome indices iii) Run the Alignment of reads with reference genome iv) Run the Feature Counts with the aligned reads. Path to files of input should be exact; otherwise, the errors will be reflected. The bash script in this analysis was developed based on numerous online literary sources and references. Two Pipelines was set up

a) Pipeline for STAR and Counts b) Pipeline for edgeR.

This script was checked and corrected by my supervisor at each step if any errors. After set up of this script, it was tested for one or two samples to check for results and some changes were done wherever applicable. Following this step, the scripts were run on the server.

Condor submission of job files (scripts on servers) is required since there will be many jobs ongoing on the server. This plays an important role also when it is required to use more than one CPU and to request for memory. Scripts for the condor submission is shown in supplementary file (sup) 9.A. (i). After a clear review of the scripts, the condor and pipeline files were uploaded on server. Lastly, give permissions to the files before submission of the condor.

FASTQC – Quality Check

FastQC is a commonly used tool for quality control analysis of RNA-seq data. This tool is used to assess the quality of the raw sequencing reads generated from the RNA-seq experiment. This report is used to identify any potential issues with the sequencing data, such as poor-quality reads, overrepresented sequences, or adapter contamination. By using FastQC, researchers can ensure that the RNA-seq data is of high quality, which is critical for downstream analysis such as mapping and differential expression analysis.

FastQC generates a report that provides information about various quality metrics as mentioned below. **Basic statistics** like read length (total seq), poor

quality seq, sequence length and % of GC content. **Per base sequence quality** with the distribution of quality scores ranging from low, medium to high with colour bands; this is a graphical representation in which x-axis denotes the position of base in read and y-axis denotes the quality scores. **Per sequence quality score** have average quality score on x-axis plotted against number of sequences on y-axis; the peak should be >20 with no bumps. **Per base sequence quality** in a random library should have equal amounts of each nucleotide (~25% of each nucleotide). **Per Sequence GC content** is important to consider for central peak matching the theoretical distribution, usually sharp peaks will be observed in case of any over represented sequences and broad peaks appear in contamination of samples. **Per Base N content** helps to filter lot of N content in reads if it exists. **Sequence duplication levels** gives an idea on reads represented more than once; low level of duplication may indicate high coverage of target sequence and a high level is more likely to indicate some kind of enrichment bias e.g., PCR over amplification. Over Represented Sequences identifies the contamination level such as vector or adapter sequences and these are important to be removed.

Mapping and Alignment using STAR:

To determine where the RNA-seq reads originated from, these reads should be aligned to the reference genome using STAR (38). This tool has high mapping speed and accuracy than other aligner methods. It works based on algorithm of finding the Maximal Mappable Prefix (MMP) hits between reads (or read pairs) and the genome, using a suffix array index. STAR algorithm consists of two major steps: seed searching step and clustering/stitching/scoring step. It uses a novel strategy for spliced alignments and address many challenges of RNA-seq data mapping. STAR also performs local alignment, automatically soft clipping ends of reads with high mismatches.

STAR tool has two steps

a) **Creating genome indices** (sometimes already available on individual institution server).

   Usually, the reference genome sequences (FASTA format) and annotation files (GTF format) from NCBI, ENSEMBL or GENCODE. The reference genome should be from the same species that the analysing sample belongs to, this is important because sometimes taking the other species may give false results. From these files, STAR uses the script with standard format

generates the genome indices which should be saved into separate folder. Script for generating the genome indices was provided under sup 9.A. (ii).

b) Mapping Reads to reference Genome.

In this step, STAR tool maps RNA-seq reads in the form of FASTA or FASTQ files to the genome files generated in the previous step. The mapping script has various input parameters that run the mapping job and gives the output files of alignments in the form of SAM/BAM format.

The scripts and guidelines for STAR were followed from the STAR manual version 2.7.0a. STAR uses the standard script shown in sup 9.A. (iii) with all required commands to perform the mapping.

Quantification with FeatureCounts

After the alignment, the next step is to measure how many reads have mapped to each genomic features such as genes, exon, promoter, genomic bins and chromosomal locations. BAM files (output from the STAR) are input to FeatureCounts (FC). This tool is more accurate, fast and easy to use.

It works by counting the reads that map to a single location which is called a uniquely mapping. FC also consider if data is stranded or not. Our data is paired-end and counting tools takes only proper paired reads into account and each read pair is counted only once as single "fragment" (39). The output from FC is of 2 files i) Count matrix with samples in columns and genes in rows. ii) Summary file that shows how many reads were assigned and not assigned.

FC quantification can be done in two levels a) Gene level – which summarizes the expression level of a gene but don't distinguish between the isoforms when multiple transcripts are being expressed from same gene.
b) Exon level – counting the reads that are overlapping at each annotated exon. This approach tests splicing between the experimental conditions. FC supports both single and multithreaded processing, very useful for summarizing data generated in large sequencing studies (40). Script used for FC is mentioned under sup 9.A. (iv).

Differential Expression Analysis by using edgeR

The analysis of baseline study LBW vs NBW included the subjects as shown in table 1 below:

	LBW	NBW	Total Number
Analysis	17	12	29

Table: 1 Showing the subjects from both LBW and NBW groups

After obtaining the gene counts from both the groups, it is essential to assess the changes in gene expression levels between different groups, typically control and testing samples. This analysis can be performed using two types of RNA-seq: a) which measures expression cell by cell and conditions between cell types, and b) which measures changes in gene expression levels at the tissue level. To perform differential expression (DE) analysis of RNA-seq data, we used the edgeR tool in the R programming language. This software is designed to identify changes between two or more groups when at least one group has replicated measurements, using a table of read counts where rows correspond to genes and columns to independent samples.

The script for the edgeR pipeline as shown in section sup 9.A. (v) was developed using the latest version (28-OCT-2022) of the edgeR user's guide (41). This script was then customized by making some corrections as per the requirement which is suitable to perform the following steps. The first step in the differential analysis is to read these counts into an R session for which edgeR has separate functions. Various other steps involved in the differential expression analysis using edgeR include designing a matrix, filtering data to remove low counts, normalizing library size, estimating dispersion size and testing for differentially expressed genes.

Pathway and Network Analysis using edgeR

The key aspect of analysing differential gene expression data is interpreting the results in terms of biological processes and pathways. Gene Ontology (GO) databases are specifically designed to annotate genes with possible GO terms. Counting of DE genes that are annotated to possible GO terms, gives the way to interpret the results. GO terms that occur more frequently in list of DE genes are said to be over-represented or enriched. This helps to identify the enriched pathways and with the help of these identified hits, network analysis is carried out to explore and visualize the functional interactions between the genes. In our study, we used the Reactome(42), David database (43) for pathway and network analysis.

5 Results

5. 1 Fast-QC Results

Fine quality data was obtained with the results as follows.

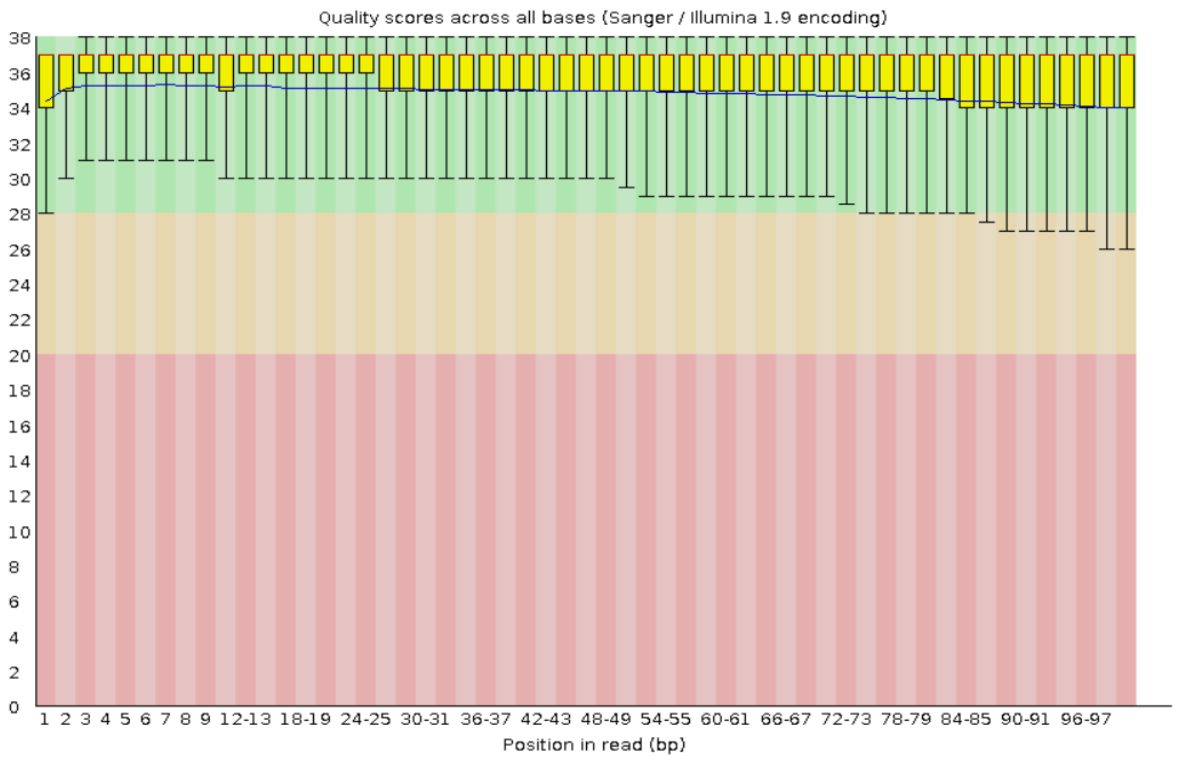
i. **Basic statistics:**

Basic Statistics

Measure	Value
Filename	55_1.fq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	25789545
Sequences flagged as poor quality	0
Sequence length	100
%GC	48

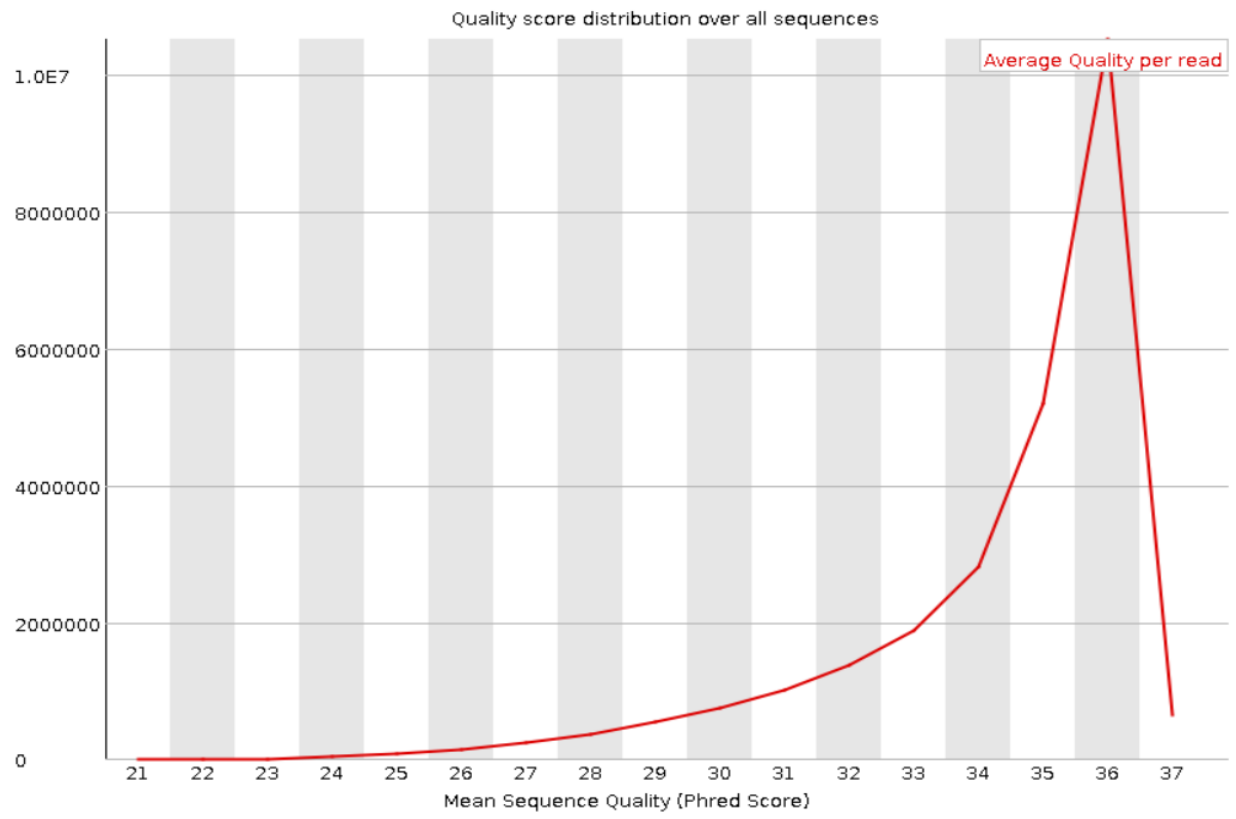
One of the examples, showing the base statistics. Overall, most of sequences are having similar properties like GC content more than 47, sequence length of 100 in all and no poor-quality sequences were flagged among all samples.

ii. **Per base sequence quality:**



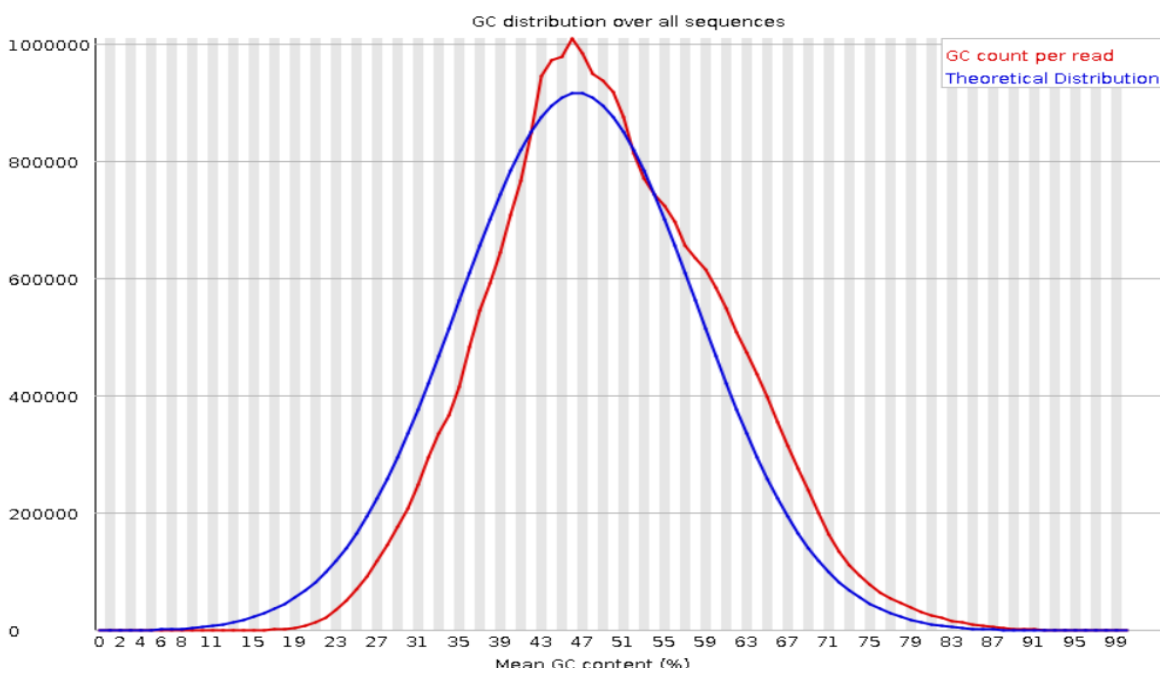
The yellow box represents the 25th and 75th percentiles, with the red line as the median. The whiskers are the 10th and 90th percentiles. Blue line gives the average quality score. In all the samples, average quality Score was 34-36 indicating high quality; in the end few extending to 26 and 28.

iii. **Per sequence quality scores**



All the samples have good quality score ranging the peak from 34 to 36, no bumps were noted.

iv. **Per sequence GC content**



GC content is almost 48 in all, 49 in few and 50 rarely. Peak is observed above the theoretical distribution.

v. Other parameters like **Per base sequence content** were almost 25 for all, uniformly distributed. **Sequence Length Distribution** was same in all 100. No per base N content and adapter content was removed. All the reads were good in quality.

5.2 Gene Counts data

A master file with the counts of all samples of analysis was obtained post FeatureCounts step. This file is loaded into Rstudios (44) and show the read count for all of the genes for each sample. In this table, the row names are

gene identification numbers and the columns represents reads from each sample. Since, my thesis is focused on base line of NBW vs LBW, the samples that belongs to this part were filtered out and separated. The following table shown in table 2 depicts the row names of geneid and columns of reads in each sample.

	Geneid	Sample_1	Sample_5	Sample_8	Sample_9	Sample_13	Sample_23	Sample_25	Sample_29
	<chr>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	ENSG00000223972	1	0	1	0	0	2	0	0
2	ENSG00000227232	115	129	68	49	66	84	82	78
3	ENSG00000278267	17	32	17	13	12	9	7	13
4	ENSG00000243485	0	0	0	0	0	1	0	0
5	ENSG00000274890	0	0	0	0	0	0	0	0
6	ENSG00000237613	0	0	0	0	0	0	0	0

Table 2: showing sample of geneid in rows and sample in columns

Phenotype file:

Each sample have its BGI sequence number and all the details of it as shown in below phenotype file from table 3.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Visit	ID	BW	...4	Amount (mg)	Total conc. n	260/280 ratio	Total conc. n	ng/ul (Bioana	RIN	28S/18S ratio	For BGI	Save for validation	BGI_ID
2	A	69	LBW	BAS	140	3519.1	1.87	3237	249	8.1	1.2	Aliquote 9 ul + 6ul H2O	4ul + 6ul H2O	1
3	A	26	LBW	BAS	105	2077.4	1.85	1768	136	8.2	1.3	Aliquote 9 ul + 6ul H2O	4ul + 6ul H2O	5
4	A	49	LBW	BAS	67	1539.2	1.84	1365	105	7.4	1	Aliquote 9 ul + 6ul H2O	4ul + 6ul H2O	9
5	A	54	LBW	BAS	45	4370.6	1.82	1248	95	6.8	0.8	Aliquote 9 ul + 6ul H2O	4ul + 6ul H2O	13
6	A	32	LBW	BAS	52	850.2	1.81	1131	87	7.1	0.9	Aliquote 9 ul + 6ul H2O	4ul + 6ul H2O	23
7	A	31	LBW	BAS	36	1149.2	1.8	1118	86	7.8	1.1	Aliquote 9 ul + 6ul H2O	4ul + 6ul H2O	25
8	A	70	LBW	BAS	36	694.2	1.75	767	59	8.2	1.3	13 µl	0 µl	36
9	A	23	LBW	BAS	32	738.4	1.75	611	47	8.6	1.1	13 µl	0 µl	42
10	A	16	LBW	BAS	27	795.6	1.84	533	41	8.7	1.3	13 µl	0 µl	50
11	A	58	LBW	BAS	40	663	1.7	403	31	8.3	1.1	13 µl	0 µl	63
12	A	66	LBW	BAS	26	405.6	1.77	299	23	8.6	1.2	13 µl	0 µl	73
13	A	41	LBW	BAS	56	336.7	1.71	169	13	8.2	0.8	13 µl	0 µl	81
14	A	13	LBW	BAS	44	871	1.74	962	74	8.1	1.1	13 µl	0 µl	142
15	A	55	LBW	BAS	54	1232.4	1.78	832	64	8.4	1.3	13 µl	0 µl	32
16	A	22	LBW	BAS	31	949	1.57	507	39	8.6	1.3	13 µl	0 µl	53

Table 3: showing sample of BGI id and its related experimental data

5.3 edgeR results

edgeR works with Limma package in R. edgeR stores the data in DGEList. The DGEList serves as an input for various functions provided by edgeR package to perform different steps of the analysis workflow. The DGElist is created using 'DGEList' function, it can be further processed for the functions like normalization methods, dispersion estimation, statistical modelling, and hypothesis testing. Grouping is required to identify the samples from each group and it can be done by giving group command. Normalization of counts by trimmed mean of M values (TMM) can be performed by calcNormFactors function. Normalizing the counts data is important to eliminate the composition biases between the libraries.

As per the edgeR user guide, a gene is required to have a count of 5-10 in a library to be considered expressed in that library. Hence, filtration with count-per-million (CPM) was done. After cpm, the genes that are lowly expressed are filtered out. The results of normalized counts were transformed into log-counts-per-million (lcpm).

Both Linear modelling and differential expression analysis in edgeR requires the matrix. We can create the matrix with the treatment conditions applied to each sample (in our analysis the matrix should be composed of two conditions NBW and LBW).

Dispersion estimation in edgeR is obtained from estimateDisp function where the dispersion of a gene can be predicted from its abundance. We used the common dispersion in one run. The Dispersion estimation uses negative binomial (NB) model and quasi-likelihood (QL) F-test provides more robust and reliable error rate control when the number of replicates is small. QL dispersion estimation and hypothesis testing was done using the function glmQLFit (). We selected the coefficient 2 in this analysis.

QL F-tests gives strict error rate control over likelihood ratio tests (LRT). From this step, the top DE genes can be viewed by function topTags (lists the top DE genes ranked by pvalue).

From this step and in further results, a positive log2-fold change (logFC) will indicate a gene up-regulated in the NBW relative to the LBW, whereas negative logFC represent the gene more highly expressed in LBW.

Gene Ontology:

In edgeR, GO analyses can be performed using goana function. Alternatively, we used different approach by downloading the CSV or Text format file available in Biomart/ENSEMBL database and then merging with the output file obtained in the above step. In the first phase the annotation file (downloaded from the Biomart) and output file were loaded and tabulated into R. In the next phase, these two files were merged using the merge command and resulted table was saved into separate file.

When all features are considered:

The results from this are shown in below table. Generally, the scores of p-value and False Discovery Rate (FDR) are used to determine the significant genes from the list. When all the features (*genes, non-coding RNA, small RNA, pseudogenes, etc*.) were considered no significant hits with FDR<0.05 after correcting for multiple testing. But there were 50 significant hits of DE genes with threshold set to (p<0.05, logFC>1 & logFC<-1) among the list and shown in table 4.

	logFC	logCPM	PValue	FDR	regulation
NA_ENSG00000276171	5.230925	3.27533097	2.367863e-06	0.04764614	Up
NECAB1_ENSG00000123119	-1.311647	3.61601394	2.587067e-05	0.19758029	Down
PTX3_ENSG00000163661	-2.351934	-0.13371925	2.945735e-05	0.19758029	Down
SCGB1B2P_ENSG00000268751	2.124440	0.75747234	5.701075e-05	0.28679260	Up
MT1A_ENSG00000205362	-1.522866	1.34043619	3.537644e-04	0.99999999	Down
KRT18_ENSG00000111057	-1.947336	1.42441914	4.111566e-04	0.99999999	Down
NA_ENSG00000274735	-2.245594	0.84528475	9.461371e-04	0.99999999	Down
NA_ENSG00000277105	-1.872662	0.91119424	1.249080e-03	0.99999999	Down
C4A_ENSG00000244731	-1.355848	-0.34542848	1.746520e-03	0.99999999	Down
RPL10P9_ENSG00000233913	1.619403	2.49738575	2.729407e-03	0.99999999	Up
SAA4_ENSG00000148965	-1.577378	1.33953261	2.880259e-03	0.99999999	Down
KIF19_ENSG00000196169	-1.401836	1.08893285	4.650422e-03	0.99999999	Down
_ENSG00000280302	-1.018195	0.04561200	5.052574e-03	0.99999999	Down
CH25H_ENSG00000138135	-1.083701	-0.65915965	6.134231e-03	0.99999999	Down
NA_ENSG00000279400	-1.374783	0.81635416	6.191707e-03	0.99999999	Down
FAIM2_ENSG00000135472	1.021857	0.90640547	6.245206e-03	0.99999999	Up
SAA1_ENSG00000173432	-1.173497	9.75608792	7.390011e-03	0.99999999	Down
URAD_ENSG00000183463	-2.022015	-0.56199640	8.836361e-03	0.99999999	Down
RNA5SP334_ENSG00000201695	-1.178836	3.37026607	9.091746e-03	0.99999999	Down
TRDN_ENSG00000186439	-1.473846	2.70920174	9.570442e-03	0.99999999	Down
DES_ENSG00000175084	-1.860730	-0.24235196	1.370755e-02	0.99999999	Down
GATD3_ENSG00000160221	1.893395	1.88066586	1.584947e-02	0.99999999	Up
SAA2_ENSG00000134339	-1.376499	6.07256177	1.614562e-02	0.99999999	Down
NA_ENSG00000275530	-1.399399	-0.10577288	1.632543e-02	0.99999999	Down
FOSB_ENSG00000125740	1.592947	0.05643344	1.779137e-02	0.99999999	Up
COBL_ENSG00000106078	-1.062931	1.78466027	1.811021e-02	0.99999999	Down
SAA2-SAA4_ENSG00000255071	-1.313371	5.63788761	1.973152e-02	0.99999999	Down
ANKRD20A11P_ENSG00000215559	-1.082475	0.25340810	2.157082e-02	0.99999999	Down
PWP2_ENSG00000241945	1.470350	-0.10733212	2.201030e-02	0.99999999	Up
CCL13_ENSG00000181374	1.147230	0.98746416	2.351973e-02	0.99999999	Up
PRND_ENSG00000171864	1.220850	0.46029366	2.492551e-02	0.99999999	Up
SCUBE1_ENSG00000159307	-1.413943	-0.09829337	2.494576e-02	0.99999999	Down
CNTN6_ENSG00000134115	1.597171	-0.06778820	2.555964e-02	0.99999999	Up
DPYSL4_ENSG00000151640	-1.486726	0.08357348	2.561735e-02	0.99999999	Down
USP6_ENSG00000129204	1.109888	0.72468778	2.808001e-02	0.99999999	Up
CSN1S1_ENSG00000126545	-1.326223	1.94013177	2.877296e-02	0.99999999	Down
APOL4_ENSG00000100336	-1.166132	4.11150976	2.934135e-02	0.99999999	Down
BBOX1_ENSG00000129151	1.013095	-0.04327008	3.029868e-02	0.99999999	Up
RNA5SP333_ENSG00000200336	-1.354830	-0.22857204	3.039224e-02	0.99999999	Down
SYNDIG1_ENSG00000101463	-1.052977	0.73200327	3.242888e-02	0.99999999	Down
GOLGA8O_ENSG00000206127	1.229046	-0.84633967	3.340493e-02	0.99999999	Up
WNT3_ENSG00000108379	-1.315031	-0.91100613	3.395421e-02	0.99999999	Down
IGKV3-11_ENSG00000241351	1.296481	2.12113121	3.956438e-02	0.99999999	Up
IGHG1_ENSG00000211896	1.385028	1.26455746	4.096980e-02	0.99999999	Up
_ENSG00000281383	-1.127534	-0.98786547	4.320683e-02	0.99999999	Down
NA_ENSG00000280102	-1.196011	4.24948214	4.444301e-02	0.99999999	Down
NA_ENSG00000280602	-1.195202	4.60962994	4.487282e-02	0.99999999	Down
RN7SL2_ENSG00000274012	-1.190753	4.61965079	4.532947e-02	0.99999999	Down
IGKV1-8_ENSG00000240671	1.250353	2.05737664	4.576723e-02	0.99999999	Up
RPPH1_ENSG00000277209	-1.293477	1.64540085	4.759673e-02	0.99999999	Down

Table 4: showing 50 significant DE genes when all features are considered

Volcano plot of results when all features were considered:

The results of table 2 were plotted in a volcano plot with log fold change on x-axis and -log10(pvalue) on y axis as shown in fig 5.3. (i) Downregulated genes in this list were labelled with blue color while the red ones indicate upregulated. Grey labelling's denote the genes with no significance.

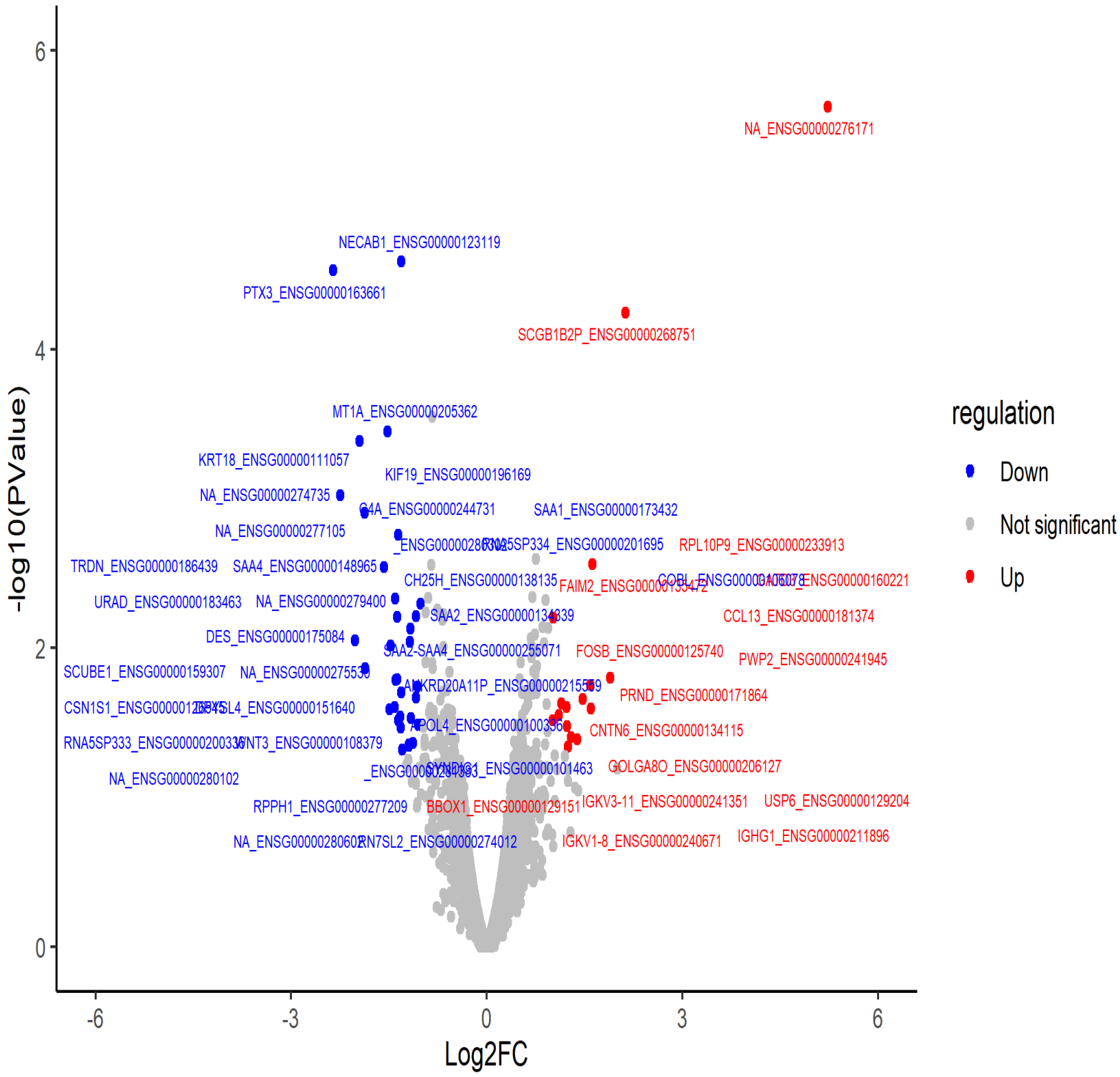


Fig: 5.3.(i) volcano plot showing 50 significant DE genes when all features are considered. Blue color represents the downregulated, red color with upregulated and grey color with non-significant.

With only protein coding genes:

Interestingly, when we considered only the protein coding genes from the list of total 22357 genes, the results showed significant hits of 31 DE protein coding genes ($p<0.05$, $logFC>1$ & $logFC<-1$).The results are shown in table 5.

	logFC	logCPM	PValue	FDR	regulation
NECAB1_ENSG00000123119	-1.301994	3.759775417	2.568009e-05	0.2075144	Down
PTX3_ENSG00000163661	-2.347542	0.008777034	2.852431e-05	0.2075144	Down
MT1A_ENSG00000205362	-1.522301	1.484980412	3.392300e-04	1.0000000	Down
KRT18_ENSG00000111057	-1.939414	1.570653408	4.218436e-04	1.0000000	Down
C4A_ENSG00000244731	-1.367317	-0.198481047	1.653766e-03	1.0000000	Down
SAA4_ENSG00000148965	-1.575895	1.484104657	2.793236e-03	1.0000000	Down
KIF19_ENSG00000196169	-1.411538	1.239743878	4.471208e-03	1.0000000	Down
CH25H_ENSG00000138135	-1.088145	-0.512772893	6.064934e-03	1.0000000	Down
FAIM2_ENSG00000135472	1.023941	1.052451873	6.133317e-03	1.0000000	Up
SAA1_ENSG00000173432	-1.168357	9.899951064	7.349277e-03	1.0000000	Down
URAD_ENSG00000183463	-2.011879	-0.416682317	8.902027e-03	1.0000000	Down
TRDN_ENSG00000186439	-1.462583	2.853218737	9.753954e-03	1.0000000	Down
DES_ENSG00000175084	-1.847959	-0.096500526	1.406884e-02	1.0000000	Down
GATD3_ENSG00000160221	1.913853	2.025243567	1.419811e-02	1.0000000	Up
SAA2_ENSG00000134339	-1.373538	6.214732866	1.547410e-02	1.0000000	Down
FOSB_ENSG00000125740	1.591989	0.197161649	1.732177e-02	1.0000000	Up
SAA2-SAA4_ENSG00000255071	-1.310159	5.780277637	1.898806e-02	1.0000000	Down
COBL_ENSG00000106078	-1.052852	1.931227317	1.899064e-02	1.0000000	Down
PWP2_ENSG00000241945	1.487576	0.039184592	2.023360e-02	1.0000000	Up
CCL13_ENSG00000181374	1.151050	1.129321145	2.218327e-02	1.0000000	Up
CNTN6_ENSG00000134115	1.604023	0.076977155	2.435438e-02	1.0000000	Up
PRND_ENSG00000171864	1.223573	0.612146877	2.476256e-02	1.0000000	Up
DPYSL4_ENSG00000151640	-1.483154	0.230537222	2.561400e-02	1.0000000	Down
SCUBE1_ENSG00000159307	-1.400200	0.048050648	2.606671e-02	1.0000000	Down
USP6_ENSG00000129204	1.118145	0.871688300	2.674996e-02	1.0000000	Up
APOL4_ENSG00000100336	-1.176471	4.261483158	2.719684e-02	1.0000000	Down
CSN1S1_ENSG00000126545	-1.321114	2.081302096	2.840861e-02	1.0000000	Down
BBOX1_ENSG00000129151	1.017865	0.103307459	2.963750e-02	1.0000000	Up
SYNDIG1_ENSG00000101463	-1.057899	0.882224227	3.210881e-02	1.0000000	Down
GOLGA8O_ENSG00000206127	1.235979	-0.698248838	3.277330e-02	1.0000000	Up
WNT3_ENSG00000108379	-1.315795	-0.763345794	3.414899e-02	1.0000000	Down

Table 5: showing 31 significant DE genes when only protein coding is considered

Volcano plot when only protein coding genes were considered:

The results of table 3 were plotted in a volcano plot with log fold change on x-axis and -log10(Pvalue) on y axis as shown in fig 5.3. (ii).

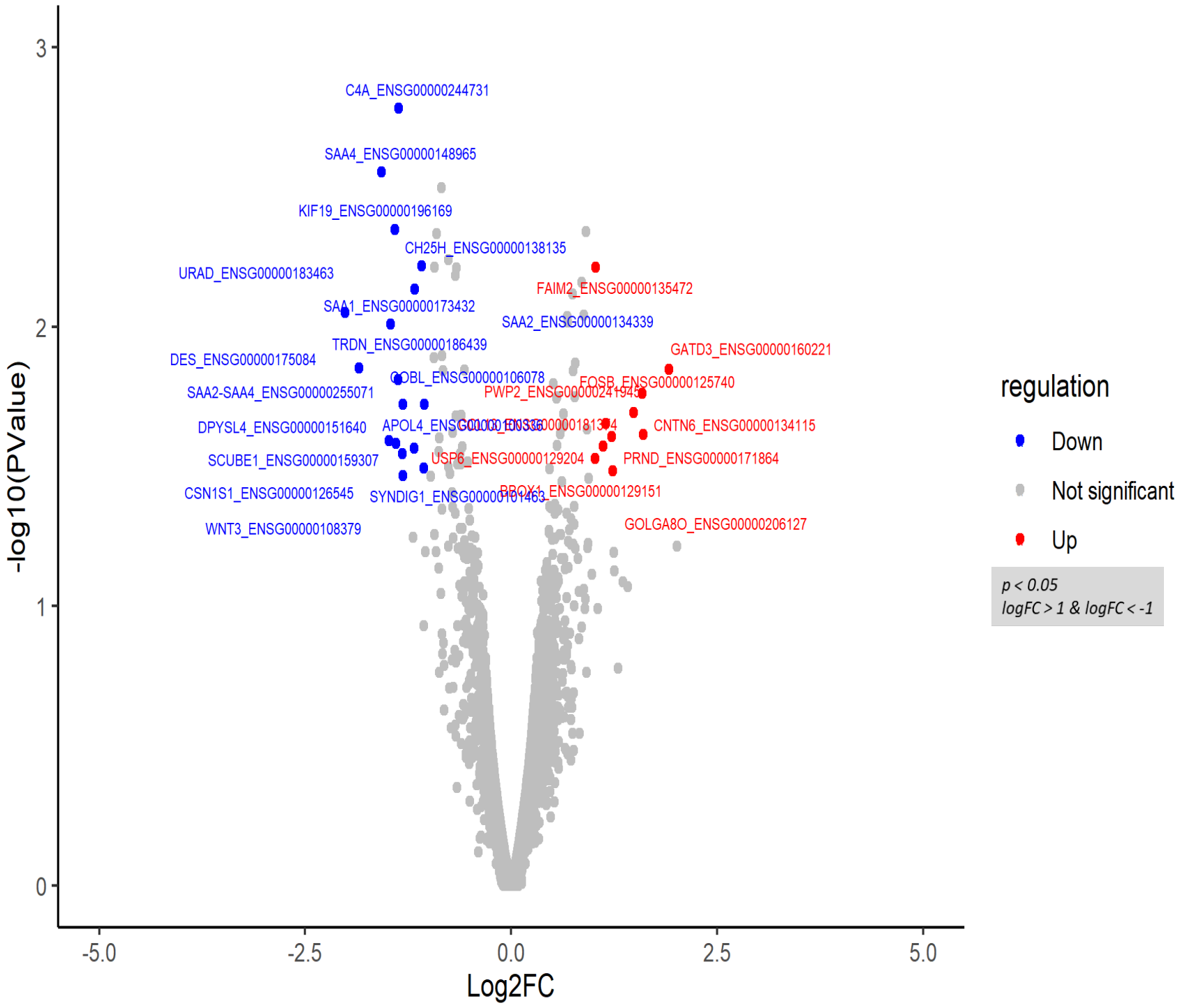


Fig: 5.3.(ii) volcano plot showing up and down regulated DE genes among protein coding genes with the p value <0.05. Blue color represents the downregulated, red color with upregulated and grey color with non-significant

DE genes:

31 significant DE when only protein coding genes were considered are as shown in table 6 below:

Downregulated	NECAB1, PTX3, MT1A, KRT18, C4A, SAA4, KIF19, CH25H, SAA1, URAD, RNA5SP334, TRDN, DES, SAA2, COBL, SAA2-SAA4, ANKRD20A11P, SCUBE1, DPYSL4, CSN1S1, APOL4, RNA5SP333, SYNDIG1, WNT3, RN7SL2, RPPH1
Upregulated	SCGB1B2P, RPL10P9, FAIM2, GATD3, FOSB, PWP2, CCL13, PRND, CNTN6, USP6, BBOX1, GOLGA8O, IGKV3-11, IGHG1, IGKV1-8

Table 6: showing downregulated and upregulated gene list

5.4 Gene Ontology results

Gene ontology results are presented in sup 9.B. Total 50 genes in the significant list when all features are considered are shown in table 9.B. (i). Downregulated among LBW in comparison to NBW (only protein coding) are mentioned in the table 9.B. (ii) along with the biological processes and molecular functions associated with each gene. Table 9.B. (iii) shows the list of upregulated genes.

5.5 Pathway and Network analysis

Pathway results from Reactome are presented in sup 9.C. with the columns of pathway associated, number of genes from the list involved in the pathway (entities found), total number of genes associated with pathway in general (entities in total), significance (entities pvalue), FDR, number of genes related to reactions involved in the pathway and total reactions associated with the pathway. Downregulated among LBW in comparison to NBW (only protein coding) are mentioned in the table 9.C. (i) whereas list of upregulated genes is as mentioned in table 9.C. (ii).
Network analysis of downregulated genes are in table 9.D. (i) and upregulated genes is shown in table 9.D. (ii).

6. Discussion

In this study, the investigation is focused on the differential expression of genes between LBW vs NBW. By elucidating these gene expression patterns, we aim to contribute to the understanding of the molecular mechanisms underlying the association between LBW and the development of T2D. Identifying such specific gene expression changes in LBW can also provide clues about the biological processes that contribute to the long-term health consequences. We identified several genes that were differentially expressed between LBW and NBW, including genes involved in immune response, metabolic pathways including lipid and glucose metabolism, and cell cycle regulation. 50 genes were found significant among LBW vs NBW when all the features were considered whereas 31 significant DE genes were identified among the protein coding genes.

Downregulated in LBW compared with NBW:

In the list shown in table 6, serum amyloid A1(SAA1), serum amyloid A2(SAA2), serum amyloid A4 (SAA4), cholesterol 25-hydroxylase (CH25H), apolipoprotein L4(APOL4) and metallothionein 1A(MT1A) are identified as very important related to T2D as below: SAA1, SAA2, SAA4 genes encode for proteins known as serum amyloid A, which are acute-phase proteins involved in several biological processes such as cholesterol transport, tissue repair, and immune response. There have been numerous studies investigating the association between genetic variations in the A1(SAA1), serum amyloid A2(SAA2), serum amyloid A4(SAA4) genes and the risk of T2D. The genetic variations in SAA1 were associated with an increased risk of T2D (45). In the cohort study of 264 patients with T2D and 275 non-diabetic controls, SAA was increased in T2D patients with incipient or overt nephropathy, and SAA was associated with impairment of cholesterol transporters, scavenger receptor class B type I (SR-BI) mediated cholesterol efflux to serum (45, 46). Variations of SAA2-SAA4 were associated with increased insulin resistance and a higher risk of developing type T2D. However, a meta-analysis of several studies found no significant association between SAA4 gene variations and T2D (47).

CH25H gene is important in regulating cholesterol metabolism and immune response. Regarding the CH25H gene, this gene has a certain role in insulin resistance (48). Recent study showed that overexpression of this gene improves the insulin sensitivity in mice (49). APOL4 is involved in the transport of lipids, particularly cholesterol, and may also have a role in innate immunity. MT1A involved in the regulation of cellular metal ion homeostasis, detoxification of heavy metals, and protection against oxidative stress. MT1A gene, was significantly related to the prevalence of T2D and also significantly associated with the low activity of serum superoxide dismutase (SOD) in T2DM (50).

These genes accounts to the majority of pathways among the downregulated and include the pathways of metabolism of lipids, cellular response to stimuli, developmental biology and Ion homeostasis. Also, the other genes like triadin

(TRDN) essential for muscle contraction, dihydropyrimidinase like 4(DPYSL4) which regulates the neuronal development and complement C4A (Rodger's blood group) (C4A) – key component of the immune system were important findings. The pathways involved by downregulated genes mostly contribute to the networks of immune system, cellular response to stimuli, metabolism and extracellular matrix organization. Overall, while there is some evidence suggesting an association between certain genetic variations in the A1(SAA1), SAA2, CH25H and MT1A genes and T2D risk, further studies are needed to confirm these findings and explore the underlying mechanisms.

Upregulated in LBW compared to NBW:

Among the upregulated genes shown in table 6, two genes namely immunoglobulin heavy constant gamma 1 (G1m marker) (IGHG1) involved in antibody-dependent cellular cytotoxicity and immunoglobulin kappa variable 3-11(IGKV3-11) involved in immune response were predominantly significant. These two genes IGHG1, IGKV3-11 play a vital role and contributes to the majority of the pathways in the upregulated genes. IGHG1 gene, a gene located on chromosome 14 has an important role in the immune response. It is found to be upregulated in the T2D (51). Other important genes are C-C motif chemokine ligand 13(CCL13) with the function of inflammatory response, immunoglobulin kappa variable 1-8(IGKV1-8) with the immune response and gamma-butyrobetaine hydroxylase 1(BBOX1) of cartinine synthesis. CCL13, also known as monocyte chemotactic protein-4 (MCP-4), is a member of the CC chemokine family, and it has been implicated in the pathogenesis of various inflammatory diseases. There are several studies that showed the strong association of this gene with diabetes (52,53). Studies have shown that CCL13 levels are increased in the islets and serum of individuals with T2D, and thus it may contribute to insulin resistance and impaired glucose tolerance (52-54). In many studies BBOX1 is associated with the diabetes kidney disease (DKD)and in diabetic nephropathy (55). Urinary BBOX1 (uBBOX1) levels were significantly upregulated in the urine of patients with DKD (56).

As mentioned earlier in the results section 5.4, the pathways include chemokine receptors bind chemokines, regulation of complement cascade and role of phospholipids in phagocytosis. Likewise, network analysis (as shown in section 5.5) of these upregulated genes showed the networks of immune system, signal transduction, metabolism of proteins, vesicle mediated transport and homeostasis. There is no strong evidence in the scientific literature linking the other upregulated genes of this study to T2D. However, it is important to note that the genetics of diabetes is complex, and multiple genes and environmental factors can contribute to the development of the disease.

The findings suggest that the differences in gene expression levels may contribute to the increased risk of T2D in LBW individuals. It is well established that LBW is associated with an increased risk of developing metabolic disorders, including T2D. However, the mechanisms underlying this association are not fully

understood. The present study shows importance on the potential role of gene expression differences in SAT in the development of T2D in LBW individuals. The differences are not significant after FDR adjustment. The pathways related to lipid metabolism, cholesterol homeostasis, steroid and glycoprotein metabolism are known to play a crucial role in maintaining metabolic homeostasis. The differences in these pathways among LBW individuals may indicate impaired metabolic function, which could contribute to the development of T2D. On the other hand, this may further exacerbate the risk of T2D.

From this study identification of DE genes by RNA-seq can provide insights into the molecular mechanisms underlying LBW individuals risk of developing T2D later in life. Through RNA-seq, it is possible to detect even low abundant transcripts and quantify the gene expression. The differences in gene expression presented in this study depicts the strong association of LBW contributing towards the development of T2D. Adipose tissue is relevant to metabolic disorders among LBW, but it is important to consider that gene expression patterns may vary across the tissue. The findings from adipose tissue may require additional integration with multi-tissue studies. The other factors like gestational age, maternal health, environmental exposures and lifestyle factors may contribute to LBW and influence the gene expression. Controlling these factors is essential to isolate the specific effects of LBW. Validating the functional significance of DE genes through in vitro experiments to assess their impact on relevant cellular processes, such as glucose metabolism, insulin signalling, or inflammation may give better understanding of underlying mechanisms. Further research in genetic association studies in larger population-based studies is important with DE genes. Exploring the epigenetic modifications associated with the identified DEGs can provide additional insights into their regulation and potential impact on diabetes. Some of the above studies are already in the pipeline of our study group to investigate BW as a key component in risk of developing T2D.

7. Conclusion

As per the hypothesis of study, there is a unique difference in the expression patterns of SAT genes between LBW individuals and their matched NBW. The study observed that certain genes related to cellular ion homeostasis, apoptotic process, cellular response to stimuli and stress showed increased expression (positive logFC) while genes associated with lipid metabolic process, cholesterol homeostasis, steroid and glycoprotein metabolic pathways showed decreased expression (negative logFC) in LBW compared to NBW individuals.

To conclude, the present study co-relates the differential expression of SAT genes in LBW individuals with the increased risk of T2D in comparison with the NBW controls. The findings suggest that these differences in gene expression levels may contribute to the increased risk of T2D in LBW individuals and could provide an opportunity for future research on the underlying mechanisms of metabolic disorders.

8. References

1. International Diabetes Federation. IDF Diabetes Atlas | Tenth Edition [Internet]. 2021 [cited 2022 Aug 11]. Available from: https://diabetesatlas.org/atlas/tenth-edition/
2. Barker, DJP, JG Eriksson, T Forsén, and C Osmond. 2002. “Fetal Origins of Adult Disease: Strength of Effects and Biological Basis.” *International Journal of Epidemiology* 31 (6): 1235–39. https://doi.org/10.1093/ije/31.6.1235.
3. Young, T. Kue, Patricia J. Martens, Shayne P. Taback, Elizabeth A. C. Sellers, Heather J. Dean, Mary Cheang, and Bertha Flett. 2002. “Type 2 Diabetes Mellitus in Children.” *Archives of Pediatrics & Adolescent Medicine* 156 (7): 651. https://doi.org/10.1001/archpedi.156.7.651.
4. Carlsson, S, P G Persson, M Alvarsson, S Efendic, A Norman, L Svanström, C G Ostenson, and V Grill. 1999. “Low Birth Weight, Family History of Diabetes, and Glucose Intolerance in Swedish Middle-Aged Men.” *Diabetes Care* 22 (7): 1043–47. https://doi.org/10.2337/diacare.22.7.1043.
5. Wei, Jung-Nan, Fung-Chang Sung, Chung-Yi Li, Chia-Hsuin Chang, Ruey-Shiung Lin, Chau-Ching Lin, Chuan-Chi Chiang, and Lee-Ming Chuang. 2003. “Low Birth Weight and High Birth Weight Infants Are Both at an Increased Risk to Have Type 2 Diabetes Among Schoolchildren in Taiwan.” *Diabetes Care* 26 (2): 343–48. https://doi.org/10.2337/diacare.26.2.343.
6. Rasmussen, Eva Lind, Charlotte Malis, Christine Bjørn Jensen, Jens-Erik Beck Jensen, Heidi Storgaard, Pernille Poulsen, Kasper Pilgaard, et al. 2005. “Altered Fat Tissue Distribution in Young Adult Men Who Had Low Birth Weight.” *Diabetes Care* 28 (1): 151–53. https://doi.org/10.2337/diacare.28.1.151.
7. Jensen, Christine B., Heidi Storgaard, Flemming Dela, Jens Juul Holst, Sten Madsbad, and Allan A. Vaag. 2002. “Early Differential Defects of Insulin Secretion and Action in 19-Year-Old Caucasian Men Who Had Low Birth Weight.” *Diabetes* 51 (4): 1271–80. https://doi.org/10.2337/diabetes.51.4.1271.
8. Ozanne, S. E., C. B. Jensen, K. J. Tingey, H. Storgaard, S. Madsbad, and A. A. Vaag. 2005. “Low Birthweight Is Associated with Specific Changes in Muscle Insulin-Signalling Protein Expression.” *Diabetologia* 48 (3): 547–52. https://doi.org/10.1007/s00125-005-1669-7.
9. Kaijser, Magnus, Anna-Karin Edstedt Bonamy, Olof Akre, Sven Cnattingius, Fredrik Granath, Mikael Norman, and Anders Ekbom. 2009. “Perinatal Risk Factors for Diabetes in Later Life.” *Diabetes* 58 (3): 523–26. https://doi.org/10.2337/db08-0558.

10. Olokoba, Abdulfatai B, Olusegun A Obateru, and Lateefat B Olokoba. 2012. “Type 2 Diabetes Mellitus: A Review of Current Trends.” *Oman Medical Journal* 27 (4): 269–73. https://doi.org/10.5001/omj.2012.68.
11. Rich-Edwards, J W, G A Colditz, M J Stampfer, W C Willett, M W Gillman, C H Hennekens, F E Speizer, and J E Manson. 1999. “Birthweight and the Risk for Type 2 Diabetes Mellitus in Adult Women.” *Annals of Internal Medicine* 130 (4 Pt 1): 278–84. https://doi.org/10.7326/0003-4819-130-4_part_1-199902160-00005.
12. Eriksson, J G, T Forsén, J Tuomilehto, C Osmond, and D J P Barker. 2003. “Early Adiposity Rebound in Childhood and Risk of Type 2 Diabetes in Adult Life.” *Diabetologia* 46 (2): 190–94. https://doi.org/10.1007/s00125-002-1012-5.
13. Zeng, Ping, and Xiang Zhou. 2019. “Causal Association Between Birth Weight and Adult Diseases: Evidence From a Mendelian Randomization Analysis.” *Frontiers in Genetics* 10: 618. https://doi.org/10.3389/fgene.2019.00618.
14. Beck Jensen, Rikke Bodin, Marla Chellakooty, Signe Vielwerth, Allan Vaag, Torben Larsen, Gorm Greisen, Niels E. Skakkebaek, Thomas Scheike, and Anders Juul. 2003. “Intrauterine Growth Retardation and Consequences for Endocrine and Cardiovascular Diseases in Adult Life: Does Insulin-Like Growth Factor-I Play a Role?” *Hormone Research in Paediatrics* 60 (Suppl. 3): 136–48. https://doi.org/10.1159/000074515.
15. Negrato, Carlos Antonio, and Marilia Brito Gomes. 2013. “Low Birth Weight: Causes and Consequences.” *Diabetology & Metabolic Syndrome* 5 (1): 49. https://doi.org/10.1186/1758-5996-5-49.
16. Warrington, Nicole M., Robin N. Beaumont, Momoko Horikoshi, Felix R. Day, Øyvind Helgeland, Charles Laurin, Jonas Bacelis, et al. 2019. “Maternal and Fetal Genetic Effects on Birth Weight and Their Relevance to Cardio-Metabolic Risk Factors.” *Nature Genetics* 51 (5): 804–14. https://doi.org/10.1038/s41588-019-0403-1.
17. Peng, Shouneng, Maya A. Deyssenroth, Antonio F. Di Narzo, Haoxiang Cheng, Zhongyang Zhang, Luca Lambertini, Arno Ruusalepp, et al. 2018a. “Genetic Regulation of the Placental Transcriptome Underlies Birth Weight and Risk of Childhood Obesity.” *PLOS Genetics* 14 (12): e1007799. https://doi.org/10.1371/journal.pgen.1007799.
18. Vaag A, Brøns C, Gillberg L, Hansen NS, Hjort L, Arora GP, Thomas N, Broholm C, Ribel-Madsen R, Grunnet LG. Genetic, nongenetic and epigenetic risk determinants in developmental programming of type 2 diabetes. Acta Obstet

Gynecol Scand. 2014 Nov;93(11):1099-108. doi: 10.1111/aogs.12494. Epub 2014 Sep 30. PMID: 25179736.

19. McCarthy, Mark I. 2010a. “Genomics, Type 2 Diabetes, and Obesity.” *New England Journal of Medicine* 363 (24): 2339–50. https://doi.org/10.1056/NEJMra0906948.
20. Galicia-Garcia, Unai, Asier Benito-Vicente, Shifa Jebari, Asier Larrea-Sebal, Haziq Siddiqi, Kepa B. Uribe, Helena Ostolaza, and César Martín. 2020. “Pathophysiology of Type 2 Diabetes Mellitus.” *International Journal of Molecular Sciences* 21 (17): 6275. https://doi.org/10.3390/ijms21176275.
21. Bellou, Vanesa, Lazaros Belbasis, Ioanna Tzoulaki, and Evangelos Evangelou. 2018. “Risk Factors for Type 2 Diabetes Mellitus: An Exposure-Wide Umbrella Review of Meta-Analyses.” *PLOS ONE* 13 (3): e0194127. https://doi.org/10.1371/journal.pone.0194127.
22. McCarthy, Mark I. 2010a. “Genomics, Type 2 Diabetes, and Obesity.” *New England Journal of Medicine* 363 (24): 2339–50. https://doi.org/10.1056/NEJMra0906948.
23. Shulman, Gerald I. 2000. “Cellular Mechanisms of Insulin Resistance.” *Journal of Clinical Investigation* 106 (2): 171–76. https://doi.org/10.1172/JCI10583.
24. Yaribeygi, Habib, Farin Rashid Farrokhi, Alexandra E Butler, and Amirhossein Sahebkar. 2019. “Insulin Resistance: Review of the Underlying Molecular Mechanisms.” *Journal of Cellular Physiology* 234 (6): 8152–61. https://doi.org/10.1002/jcp.27603.
25. Eckel, Robert H, Scott M Grundy, and Paul Z Zimmet. n.d. “The Metabolic Syndrome.” *Lancet (London, England)* 365 (9468): 1415–28. https://doi.org/10.1016/S0140-6736(05)66378-7.
26. Westman, Eric C. 2021. “Type 2 Diabetes Mellitus: A Pathophysiologic Perspective.” *Frontiers in Nutrition* 8 (August). https://doi.org/10.3389/fnut.2021.707371.
27. K. C., Anil, Prem Lal Basel, and Sarswoti Singh. 2020. “Low Birth Weight and Its Associated Risk Factors: Health Facility-Based Case-Control Study.” *PLOS ONE* 15 (6): e0234907. https://doi.org/10.1371/journal.pone.0234907.
28. Belbasis, Lazaros, Makrina D. Savvidou, Chidimma Kanu, Evangelos Evangelou, and Ioanna Tzoulaki. 2016. “Birth Weight in Relation to Health and Disease in Later Life: An Umbrella Review of Systematic Reviews and Meta-Analyses.” *BMC Medicine* 14 (1): 147. https://doi.org/10.1186/s12916-016-0692-5.

29. Bernhardsen, Guro Pauck, Trine Stensrud, Bjørge Herman Hansen, Jostein Steene-Johannesen, Elin Kolle, Wenche Nystad, Sigmund Alfred Anderssen, et al. 2020. “Birth Weight, Cardiometabolic Risk Factors and Effect Modification of Physical Activity in Children and Adolescents: Pooled Data from 12 International Studies.” *International Journal of Obesity* 44 (10): 2052–63. https://doi.org/10.1038/s41366-020-0612-9.
30. Sjöholm, Pauline, Katja Pahkala, Belinda Davison, Harri Niinikoski, Olli Raitakari, Markus Juonala, and Gurmeet R. Singh. 2021. “Birth Weight for Gestational Age and Later Cardiovascular Health: A Comparison between Longitudinal Finnish and Indigenous Australian Cohorts.” *Annals of Medicine* 53 (1): 2060–71. https://doi.org/10.1080/07853890.2021.1999491.
31. Hales, C Nicholas, and David J P Barker. 2001. “The Thrifty Phenotype Hypothesis.” *British Medical Bulletin* 60 (1): 5–20. https://doi.org/10.1093/bmb/60.1.5.
32. Horikoshi, Momoko, Hanieh Yaghootkar, Dennis O Mook-Kanamori, Ulla Sovio, H Rob Taal, Branwen J Hennig, Jonathan P Bradfield, et al. 2013. “New Loci Associated with Birth Weight Identify Genetic Links between Intrauterine Growth and Adult Height and Metabolism.” *Nature Genetics* 45 (1): 76–82. https://doi.org/10.1038/ng.2477.
33. Dunger, David B., Clive J. Petry, and Ken K. Ong. 2007. “Genetics of Size at Birth.” *Diabetes Care* 30 (Supplement_2): S150–55. https://doi.org/10.2337/dc07-s208.
34. Peng, Shouneng, Maya A. Deyssenroth, Antonio F. Di Narzo, Haoxiang Cheng, Zhongyang Zhang, Luca Lambertini, Arno Ruusalepp, et al. 2018a. “Genetic Regulation of the Placental Transcriptome Underlies Birth Weight and Risk of Childhood Obesity.” *PLOS Genetics* 14 (12): e1007799. https://doi.org/10.1371/journal.pgen.1007799.
35. McCarthy, Mark I. 2010a. “Genomics, Type 2 Diabetes, and Obesity.” *New England Journal of Medicine* 363 (24): 2339–50. https://doi.org/10.1056/NEJMra0906948.
36. RNA-Seq: Basics, Applications and Protocol. Available from the link https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461#D1
37. Your Genome. Available from: https://www.yourgenome.org/facts/what-is-rna-sequencing/
38. Alignment with STAR | Introduction to RNA-Seq using high-performance computing. Available from the link https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

39. Introduction to RNA-seq using high-performance computing (HPC). Available from the link https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/05_counting_reads.html

40. Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30. https://doi.org/10.1093/bioinformatics/btt656.

41. Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “EdgeR : A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

42. The Reactome Pathway Knowledgebase. Version 84 released on 23/march/2023. Available from: https://reactome.org/

43. DAVID: database for annotation, visualization, and integrated discovery. Version DAVID 2021 (Dec. 2021). Available from the link https://david.ncifcrf.gov/

44. RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/ .

45. Tsun, J. G. S., Shiu, S. W. M., Wong, Y., Yung, S., Chan, T. M., & Tan, K. C. B. (2013). Impact of serum amyloid A on cellular cholesterol efflux to serum in type 2 diabetes mellitus. Atherosclerosis, 231(2), 405–410. https://doi.org/10.1016/j.atherosclerosis.2013.10.008.

46. Xie, Xiang, Yi-Tong Ma, Yi-Ning Yang, Xiao-Mei Li, and Yi-Tong Ma. 2012. “HIGH PREVALENCE OF TYPE-2 DIABETES IN CHINESE OIL WORKERS: INTERACTION BETWEEN SAA1 GENE AND WORK STRESS.” Heart 98 (Suppl 2): E40.1-E40. https://doi.org/10.1136/heartjnl-2012-302920a.92.

47. Yang, Meng-Ting, Wei-Hung Chang, Tien-Fen Kuo, Ming-Yi Shen, Chu-Wen Yang, Yin-Jing Tien, Bun-Yueh Lai, Yet-Ran Chen, Yi-Cheng Chang, and Wen-Chin Yang. 2021. “Identification of Novel Biomarkers for Pre-Diabetic Diagnosis Using a Combinational Approach.” Frontiers in Endocrinology 12 (April). https://doi.org/10.3389/fendo.2021.641336.

48. Russo, Lucia, Lindsey Muir, Lynn Geletka, Jennifer Delproposto, Nicki Baker, Carmen Flesher, Robert O’Rourke, and Carey N. Lumeng. 2020. “Cholesterol 25-Hydroxylase (CH25H) as a Promoter of Adipose Tissue Inflammation in Obesity and Diabetes.” Molecular Metabolism 39 (September): 100983. https://doi.org/10.1016/j.molmet.2020.100983.

49. Noebauer, Britta, Alexander Jais, Jelena Todoric, Klaus Gossens, Hedwig Sutterlüty-Fall, and Elisa Einwallner. 2017. “Hepatic Cholesterol-25-Hydroxylase Overexpression Improves Systemic Insulin Sensitivity in Mice.” Journal of Diabetes Research 2017: 1–8. https://doi.org/10.1155/2017/4108768.
50. Yang, Lina, Hongyan Li, Ting Yu, Haijun Zhao, M. George Cherian, Lu Cai, and Ya Liu. 2008. “Polymorphisms in Metallothionein-1 and -2 Genes Associated with the Risk of Type 2 Diabetes Mellitus and Its Complications.” American Journal of Physiology-Endocrinology and Metabolism 294 (5): E987–92. https://doi.org/10.1152/ajpendo.90234.2008
51. Donath, Marc Y., and Steven E. Shoelson. 2011. “Type 2 Diabetes as an Inflammatory Disease.” Nature Reviews Immunology 11 (2): 98–107. https://doi.org/10.1038/nri2925.
52. Igoillo-Esteve, M., L. Marselli, D. A. Cunha, L. Ladrière, F. Ortis, F. A. Grieco, F. Dotta, et al. 2010. “Palmitate Induces a Pro-Inflammatory Response in Human Pancreatic Islets That Mimics CCL2 Expression by Beta Cells in Type 2 Diabetes.” Diabetologia 53 (7): 1395–1405. https://doi.org/10.1007/s00125-010-1707-y.
53. Taylor-Fishwick, D. A., J. R. Weaver, W. Grzesik, S. Chakrabarti, S. Green-Mitchell, Y. Imai, N. Kuhn, and J. L. Nadler. 2013. “Production and Function of IL-12 in Islets and Beta Cells.” Diabetologia 56 (1): 126–35. https://doi.org/10.1007/s00125-012-2732-9.
54. Butcher, Matthew J., Daniel Hallinger, Eden Garcia, Yui Machida, Swarup Chakrabarti, Jerry Nadler, Elena V. Galkina, and Yumi Imai. 2014. “Association of Proinflammatory Cytokines and Islet Resident Leucocytes with Islet Dysfunction in Type 2 Diabetes.” Diabetologia 57 (3): 491–501. https://doi.org/10.1007/s00125-013-3116-5.
55. Chowdhury, Utpala Nanda, M. Babul Islam, Shamim Ahmad, and Mohammad Ali Moni. 2020. “Network-Based Identification of Genetic Factors in Ageing, Lifestyle and Type 2 Diabetes That Influence to the Progression of Alzheimer’s Disease.” Informatics in Medicine Unlocked 19: 100309. https://doi.org/10.1016/j.imu.2020.100309.
56. Zhou, Le-Ting, Lin-Li Lv, Shen Qiu, Qing Yin, Zuo-Lin Li, Tao-Tao Tang, Li-Hua Ni, et al. 2019. “Bioinformatics-Based Discovery of the Urinary BBOX1 MRNA as a Potential Biomarker of Diabetic Kidney Disease.” Journal of Translational Medicine 17 (1): 59. https://doi.org/10.1186/s12967-019-1818-2.

9. Supplementary

A. SCRIPTS

i. Condor script

```
Universe = vanilla
Executable =
/ludc/Active_Projects/LBWHCOFSATBRNAS/Private/run1//Condor/Executives/2.sh
Arguments =
Output = /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/run1//Condor/Log/2.out
getenv = TRUE
Log = /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/run1//Condor/Log/2.log
Error = /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/run1//Condor/Log/2.error
request_cpus = 6
request_memory = 60000
notify_user = prabhudeva.thummala@regionh.dk
notification = always
queue 1
```

ii. Script for generating genome indices in STAR

```
#!/bin/bash

      cd /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files

mkdir /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files/STAR_INDEX


 # STEP 1: Generate genome indices

       #generate the genome indices

 /ludc/Tools/Software/STAR/2.7.1a/bin/STAR
--runThreadN 10 --runMode genomeGenerate --genomeDir /ludc/Active_Projects
--genomeFastaFiles /ludc/Reference_Data/Public/Human/STAR_Genomes/GrcH38/G
--sjdbGTFfile /ludc/Reference_Data/Public/Human/STAR_Genomes/GrcH38/gencod

echo "finished running!"
```

iii. Script for mapping and alignment in STAR

```
#!/bin/bash
cd /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files
mkdir /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files/STAR_Alignment
mkdir /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files/FeatureCounts
mkdir
/ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files/STAR_Alignment/117/
mkdir /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files/FeatureCounts/117/

# STEP 1: Run STAR
# run alignment
/ludc/Tools/Software/STAR/2.7.1a/bin/STAR \
--runThreadN 10 --readFilesIn \
```

```
/ludc/Raw_Data_Archive/Sequencing/Rna_Seq/LBWHCOFSATBRNAS/Raw_Content/20220824_FastQ/F21FTSEUHT0020-01_HUMyrfE/Clean//117/117_1.fq.gz \
/ludc/Raw_Data_Archive/Sequencing/Rna_Seq/LBWHCOFSATBRNAS/Raw_Content/20220824_FastQ/F21FTSEUHT0020-01_HUMyrfE/Clean//117/117_2.fq.gz \
--genomeDir /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/STAR_INDEX \
--sjdbGTFfile /ludc/Reference_Data/Public/Human/STAR_Genomes/hg38_Gencode22/gencode.v22.annotation.gtf \
--outFilterMismatchNmax 10 --outFilterType BySJout --outReadsUnmapped Fastx --readFilesCommand zcat --outSAMtype BAM SortedByCoordinate \
--outSAMstrandField intronMotif –outFileNamePrefix \
/ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files/STAR_Alignment/117/

    echo "STAR finished running!"
```

iv. Script for FeatureCounts

```
#run featureCounts
/ludc/Tools/Software/Subread/1.6.4/bin/featureCounts \
-p -s 0 -T 10 -a /ludc/Reference_Data/Public/Human/STAR_Genomes/hg38_Gencode22/gencode.v22.annotation.gtf \
/ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files/STAR_Alignment/117/Aligned.sortedByCoord.out.bam \
-o /ludc/Active_Projects/LBWHCOFSATBRNAS/Private/Prab_Files/FeatureCounts/117/117.genecounts.txt
echo "featureCounts finished running!"
echo "analysis complete"
```

v. Script for edgeR

```
getwd()
setwd("/home/prabhudev_t/LUDC/")
counts=read.table(file="/home/prabhudev_t/LUDC/GeneCounts.txt", sep="\t", header = T)
head(counts)
dim(counts)
x=counts[,c(1,7:10)]
row.names(x)=x$Geneid
x=x[,-1]
head(x)
group <- factor(c(1,1,2,2))

library(edgeR)
y<-DGEList(counts=x,group=group)
y <- calcNormFactors(y)
```

```
cpm <- cpm(y, log = False, normalized.lib.sizes=TRUE)
keep <- rowSums(cpm(y)>1) >=2
y <- y[keep,]
dim(y)
write.table(cpm, "/home/prabhudev_t/LUDC/cpm_Results", row.names=TRUE)

logcpm <- cpm(y,log=TRUE)
write.results(logcpm, row.names=TRUE, col.names=TRUE, file =
"/home/prabhudev_t/LUDC/logcpm_Results", sep="\t")

design <- model.matrix(~group)
design
y <- estimateDisp(y,design)
fit <- glmFit(y,design)
lrt <- glmLRT(fit,coef=2)
topTags(lrt)

##########################################
#here to write our results:
#write.table(topTags(lrt, 35000L), file = "output.txt", row.names = TRUE, col.names=TRUE)
##########################################


##########################################
#For annotation, I downloaded a new csv or text file from Biomart/ensembl, and then
merged it by using a common header ENS like below:

annotation<-read.table("/mart_exportcopy.txt", header=TRUE)
head(annotation)
results<-read.table("output.txt", header=TRUE)
head(results)

annotated_results<-merge(results,annotation,by = "ENS", all=TRUE)
#annotated_results$FDR<-p.adjust(annotated_results$PValue,"fdr")
head(annotated_results)
write.table(annotated_results,file='output_annotated.txt',col.names=TRUE,row.names=FA
LSE,sep='\t')
##########################################
```

B. Gene Ontology Results

i. Gene list from the table of 50 significant (when all features considered)

Ensembl gene id	Gene name	Gene symbol
ENSG00000215559	ankyrin repeat domain 20 family member A11, pseudogene	ANKRD20A11P
ENSG00000100336	apolipoprotein L4	APOL4
ENSG00000129151	gamma-butyrobetaine hydroxylase 1	BBOX1
ENSG00000244731	complement C4A (Rodgers blood group)	C4A
ENSG00000181374	C-C motif chemokine ligand 13	CCL13
ENSG00000138135	cholesterol 25-hydroxylase	CH25H
ENSG00000134115	contactin 6	CNTN6
ENSG00000106078	cordon-bleu WH2 repeat protein	COBL
ENSG00000126545	casein alpha s1	CSN1S1
ENSG00000175084	desmin	DES
ENSG00000151640	dihydropyrimidinase like 4	DPYSL4
ENSG00000135472	Fas apoptotic inhibitory molecule 2	FAIM2
ENSG00000125740	FosB proto-oncogene, AP-1 transcription factor subunit	FOSB
ENSG00000160221	glutamine amidotransferase class 1 domain containing 3	GATD3
ENSG00000206127	golgin A8 family member O	GOLGA8O
ENSG00000211896	immunoglobulin heavy constant gamma 1 (G1m marker)	IGHG1
ENSG00000240671	immunoglobulin kappa variable 1-8	IGKV1-8
ENSG00000241351	immunoglobulin kappa variable 3-11	IGKV3-11
ENSG00000196169	kinesin family member 19	KIF19
ENSG00000111057	keratin 18	KRT18
ENSG00000205362	metallothionein 1A	MT1A
ENSG00000123119	N-terminal EF-hand calcium binding protein 1	NECAB1
ENSG00000171864	prion like protein doppel	PRND
ENSG00000163661	pentraxin 3	PTX3
ENSG00000241945	PWP2 small subunit processome component	PWP2
ENSG00000274012	RNA component of signal recognition particle 7SL2	RN7SL2
ENSG00000200336	RNA, 5S ribosomal pseudogene 333	RNA5SP333
ENSG00000201695	RNA, 5S ribosomal pseudogene 334	RNA5SP334
ENSG00000233913	ribosomal protein L10 pseudogene 9	RPL10P9
ENSG00000277209	ribonuclease P RNA component H1	RPPH1
ENSG00000173432	serum amyloid A1	SAA1
ENSG00000134339	serum amyloid A2	SAA2
ENSG00000255071	SAA2-SAA4 readthrough	SAA2-SAA4
ENSG00000148965	serum amyloid A4, constitutive	SAA4
ENSG00000268751	secretoglobin family 1B member 2, pseudogene	SCGB1B2P
ENSG00000159307	signal peptide, CUB domain and EGF like domain containing 1	SCUBE1
ENSG00000101463	synapse differentiation inducing 1	SYNDIG1
ENSG00000186439	triadin	TRDN
ENSG00000183463	ureidoimidazoline (2-oxo-4-hydroxy-4-carboxy-5-) decarboxylase	URAD
ENSG00000129204	ubiquitin specific peptidase 6	USP6
ENSG00000108379	Wnt family member 3	WNT3

ii. Downregulated genes in LBW compared to NBW (only protein coding)

Gene Name	GOTERM - Biological Process	GOTERM - Molecular Function
N-terminal EF-hand calcium——binding protein 1(NECAB1)	GO:0001835~blastocyst hatching,GO:0042984~regulation of amyloid precursor protein biosynthetic process,	GO:0005509~calcium ion binding,GO:0005515~protein binding,GO:0042802~identical protein binding,
SAA2-SAA4 readthrough(SAA2-SAA4)	GO:0006953~acute-phase response,	
Wnt family member 3(WNT3)	GO:0000902~cell morphogenesis,GO:0001707~mesoderm formation,GO:0007276~gamete generation,GO:0007411~axon guidance,GO:0009948~anterior/posterior axis specification, GO:0030177~positive regulation of Wnt signaling pathway,GO:0030182~neuron differentiation,GO:0035115~embryonic forelimb morphogenesis,GO:0035116~embryonic hindlimb morphogenesis,GO:0044338~canonical Wnt signaling pathway involved in mesenchymal stem cell differentiation,GO:0044339~canonical Wnt signaling pathway involved in osteoblast differentiation	GO:0005109~frizzled binding,GO:0005125~cytokine activity,GO:0005515~protein binding, GO:0048018~receptor agonist activity,
apolipoprotein L4(APOL4)	GO:0006629~lipid metabolic process,GO:0006869~lipid transport,GO:0042157~lipoprotein metabolic process,	GO:0008289~lipid binding,
casein alpha s1(CSN1S1)	GO:0032355~response to estradiol,GO:0032570~response to progesterone,GO:1903494~response to dehydroepiandrosterone,GO:1903496~response to 11-deoxycorticosterone,	GO:0005515~protein binding,
cholesterol 25-hydroxylase(CH25H)	GO:0006629~lipid metabolic process,GO:0008203~cholesterol metabolic process,GO:0016126~sterol biosynthetic process,GO:0034340~response to type I interferon,GO:0035754~B cell chemotaxis,GO:1903914~negative regulation of fusion of virus membrane with host plasma membrane,	GO:0000254~C-4 methylsterol oxidase activity,GO:0001567~cholesterol 25-hydroxylase activity, GO:0008395~steroid hydroxylase activity,
complement C4A (Rodgers blood group)(C4A)	GO:0006954~inflammatory response,GO:0006956~complement activation,GO:0006958~complement activation, classical pathway,GO:0045087~innate immune response,GO:2000427~positive regulation of apoptotic cell clearance,	GO:0001849~complement component C1q binding,GO:0004866~endopeptidase inhibitor activity,
cordon-bleu WH2 repeat protein(COBL)	GO:0000578~embryonic axis specification,GO:0001757~somite specification,GO:0001843~neural tube closure,GO:0001889~liver development,GO:0030041~actin filament polymerization,GO:0030903~notochord development,GO:0033504~floor plate development,GO:0048565~digestive tract development, GO:1900006~positive regulation of dendrite development,GO:1900029~positive regulation of ruffle assembly,	GO:0003785~actin monomer binding,GO:0005515~protein binding,
desmin(DES)	GO:0006936~muscle contraction,GO:0007010~cytoskeleton organization,GO:0008016~regulation of heart contraction,GO:0045109~intermediate filament organization,GO:0060538~skeletal muscle organ development,	GO:0008092~cytoskeletal protein binding,GO:0042802~identical protein binding,
dihydropyrimidinase like 4(DPYSL4)	GO:0007399~nervous system development,GO:0070997~neuron death,GO:0097485~neuron projection guidance,	GO:0016810~hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds,GO:0016812~hydrolase activity
keratin 18(KRT18)	GO:0007049~cell cycle,GO:0009653~anatomical structure morphogenesis,GO:0033209~tumor necrosis factor-mediated signaling pathway,GO:0043000~Golgi to plasma membrane CFTR protein transport,GO:0043001~Golgi to plasma membrane protein transport,GO:0043066~negative regulation of apoptotic process,GO:0045104~intermediate filament cytoskeleton organization,GO:0097191~extrinsic apoptotic signaling pathway,GO:0097284~hepatocyte apoptotic process,GO:0098609~cell-cell adhesion,	GO:0003723~RNA binding, GO:0097110~scaffold protein binding,GO:0098641~cadherin binding involved in cell-cell adhesion,
kinesin family member 19(KIF19)	GO:0007018~microtubule-based movement,GO:0060404~axonemal microtubule depolymerization,GO:0070462~plus-end specific microtubule depolymerization,	GO:0008574~ATP-dependent microtubule motor activity, plus-end-directed,GO:0016887~ATPase activity,

metallothionein 1A(MT1A)	GO:0006882~cellular zinc ion homeostasis,GO:0010273~detoxification of copper ion,GO:0045926~negative regulation of growth,GO:0071276~cellular response to cadmium ion,GO:0071280~cellular response to copper ion,GO:0071294~cellular response to zinc ion,	GO:0005515~protein binding,GO:0008270~zinc ion binding,GO:0046872~metal ion binding,
pentraxin 3(PTX3)	GO:0001550~ovarian cumulus expansion,GO:0001878~response to yeast,GO:0006954~inflammatory response,GO:0008228~opsonization,GO:0030198~extracellular matrix organization,GO:0044793~negative regulation by host of viral process,GO:0044869~negative regulation by host of viral exo-alpha-sialidase activity ,GO:1903019~negative regulation of glycoprotein metabolic process,	GO:0001849~complement component C1q binding,GO:0001872~(1->3)-beta-D-glucan binding,
serum amyloid A1(SAA1)	GO:0001819~positive regulation of cytokine production,GO:0006953~acute-phase response,GO:0007204~positive regulation of cytosolic calcium ion concentration GO:0048247~lymphocyte chemotaxis,GO:0050708~regulation of protein secretion,GO:0050728~negative regulation of inflammatory response,	GO:0001664~G-protein coupled receptor binding,GO:0008201~heparin binding,
serum amyloid A2(SAA2)	GO:0006953~acute-phase response,	GO:0005515~protein binding,
serum amyloid A4, constitutive(SAA4)	GO:0006953~acute-phase response,	GO:0005515~protein binding,
synapse differentiation inducing 1(SYNDIG1)	GO:0006886~intracellular protein transport,GO:0051965~positive regulation of synapse assembly,GO:0097091~synaptic vesicle clustering,	GO:0005515~protein binding,GO:0035254~glutamate receptor binding,GO:0042803~protein homodimerization activity,
triadin(TRDN)	GO:0006874~cellular calcium ion homeostasis,GO:0006936~muscle contraction,GO:0009617~response to bacterium,GO:0010649~regulation of cell communication by electrical coupling ,GO:0014808~release of sequestered calcium ion into cytosol by sarcoplasmic reticulum, GO:0060047~heart contraction, GO:0086036~regulation of cardiac muscle cell membrane potential,GO:0090158~endoplasmic reticulum membrane organization,	GO:0030674~protein binding, bridging,GO:0044325~ion channel binding,
ureidoimidazoline (2-oxo-4-hydroxy-4-carboxy-5-) decarboxylase(URAD)	GO:0000255~allantoin metabolic process,GO:0006144~purine nucleobase metabolic process,GO:0019628~urate catabolic process,	GO:0016831~carboxy-lyase activity,GO:0051997~2-oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline decarboxylase activity,

iii. Upregulated genes in LBW compared to NBW (only protein coding)

Gene Name	GOTERM-Biological Process	GOTERM- Molecular Function
C-C motif chemokine ligand 13(CCL13)	GO:0002548~monocyte chemotaxis, GO:0006874~cellular calcium ion homeostasis, GO:0006935~chemotaxis,GO:0006954~inflammatory response,GO:0006955~immune response, , GO:0070374~positive regulation of ERK1 and ERK2 cascade,GO:0071346~cellular response to interferon-gamma,GO:0071347~cellular response to interleukin-1.	GO:0005102~receptor binding,GO:0005515~protein binding,GO:0008009~chemokine activity,GO:0048020~CCR chemokine receptor binding,
Fas apoptotic inhibitory molecule 2(FAIM2)	GO:0002931~response to ischemia, GO:0006915~apoptotic process, GO:0021549~cerebellum development, GO:0051402~neuron apoptotic process, GO:0097190~apoptotic signalling pathway, GO:1902042~negative regulation of extrinsic apoptotic signalling pathway via death domain receptors, GO:2001234~negative regulation of apoptotic signalling pathway,	GO:0005515~protein binding,
FosB proto-oncogene, AP-1 transcription factor subunit(FOSB)	GO:0000122~negative regulation of transcription from RNA polymerase II promoter, GO:0007565~female pregnancy, GO:0009410~response to xenobiotic stimulus, GO:0032870~cellular response to hormone stimulus, GO:0043278~response to morphine	GO:0000978~RNA polymerase II core promoter proximal region sequence-specific DNA binding, GO:0000981~RNA polymerase II transcription factor activity, sequence-specific DNA binding
PWP2 small subunit processome component(PWP2)	GO:0000028~ribosomal small subunit assembly, GO:0000462~maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA),	GO:0003723~RNA binding,
contactin 6(CNTN6)	GO:0007155~cell adhesion,GO:0007156~homophilic cell adhesion via plasma membrane adhesion molecules, GO:0007417~central nervous system development,GO:0045747~positive regulation of Notch signalling pathway,GO:0070593~dendrite self-avoidance,	GO:0098632~protein binding involved in cell-cell adhesion,
gamma-butyrobetaine hydroxylase 1(BBOX1)	GO:0045329~carnitine biosynthetic process,	GO:0005506~iron ion binding, GO:0008270~zinc ion binding, GO:0008336~gamma-butyrobetaine dioxygenase activity, GO:0046872~metal ion binding,
golgin A8 family member O(GOLGA8O)	GO:0007030~Golgi organization,	GO:0005515~protein binding,
prion like protein doppel(PRND)	GO:0006878~cellular copper ion homeostasis,GO:0007340~acrosome reaction,GO:0051260~protein homooligomerization,	GO:0005507~copper ion binding, GO:0005515~protein binding,
ubiquitin specific peptidase 6(USP6)	GO:0006464~cellular protein modification process,GO:0006511~ubiquitin-dependent protein catabolic process,GO:0006886~intracellular protein transport,GO:0016579~protein deubiquitination,GO:0036211~protein modification process,GO:0060627~regulation of vesicle-mediated transport,GO:0090630~activation of GTPase activity,	GO:0003676~nucleic acid binding, GO:0004197~cysteine-type endopeptidase activity, GO:0004843~thiol-dependent ubiquitin-specific protease activity, GO:0005096~GTPase activator activity,

C.Reactome Results

i. Downregulated in LBW compared to NBW (only protein coding genes)

Pathway name	#Entities found	#Entities total	Entities ratio	Entities pValue	Entities FDR	#Reactions found	#Reactions total
Activation of C3 and C5	2	7	4.61E-04	5.35E-05	0.005297	3	4
Metallothionein's bind metals	2	16	0.001055	2.77E-04	0.013583	6	27
Response to metal ions	2	21	0.001384	4.75E-04	0.015686	6	31
Initial triggering of complement	2	120	0.007909	0.014173	0.168192	3	21
Formyl peptide receptors bind formyl peptides and many other ligands	1	11	7.25E-04	0.016542	0.168192	1	3
Regulation of Complement cascade	2	139	0.009161	0.018688	0.168192	14	42
Complement cascade	2	156	0.010281	0.023176	0.168192	20	72
Advanced glycosylation end product receptor signalling	1	16	0.001055	0.023974	0.168192	2	4
CRMPs in Sema3A signalling	1	18	0.001186	0.026932	0.168192	3	5
Scavenging by Class B Receptors	1	21	0.001384	0.031353	0.168192	2	5
Interleukin-4 and Interleukin-13 signalling	2	211	0.013906	0.040328	0.168192	1	47
WNT ligand biogenesis and trafficking	1	28	0.001845	0.041593	0.168192	8	12
TRAF6 mediated NF-kB activation	1	30	0.001977	0.0445	0.168192	1	4
Muscle contraction	2	232	0.01529	0.047834	0.168192	5	53
Miscellaneous transport and binding events	1	36	0.002373	0.05317	0.168192	1	13
Striated Muscle Contraction	1	40	0.002636	0.058908	0.168192	4	4
TAK1-dependent IKK and NF-kappa-B activation	1	55	0.003625	0.08013	0.168192	1	17
Ion homeostasis	1	64	0.004218	0.092643	0.168192	1	17
Kinesins	1	68	0.004482	0.098152	0.168192	2	14
Semaphorin interactions	1	71	0.004679	0.102263	0.168192	3	41
Amyloid fiber formation	1	89	0.005866	0.126553	0.168192	2	33
Class B/2 (Secretin family receptors)	1	99	0.006525	0.139775	0.168192	1	24
Innate Immune System	4	1340	0.088315	0.140026	0.168192	25	725
DDX58/IFIH1-mediated induction of interferon-alpha/beta	1	104	0.006854	0.146313	0.168192	1	53
COPI-dependent Golgi-to-ER retrograde traffic	1	107	0.007052	0.150214	0.168192	2	11
Post-translational protein phosphorylation	1	109	0.007184	0.152805	0.168192	1	1
MyD88 cascade initiated on plasma membrane	1	109	0.007184	0.152805	0.168192	1	70
Toll Like Receptor 5 (TLR5) Cascade	1	109	0.007184	0.152805	0.168192	1	71
Toll Like Receptor 10 (TLR10) Cascade	1	109	0.007184	0.152805	0.168192	1	71
Synthesis of bile acids and bile salts	1	113	0.007447	0.157964	0.168192	1	83
TRAF6 mediated induction of NFkB and MAP kinases upon TLR7/8 or 9 activation	1	116	0.007645	0.161813	0.168192	1	60
Toll Like Receptor 3 (TLR3) Cascade	1	116	0.007645	0.161813	0.168192	1	73
MyD88 dependent cascade initiated on endosome	1	117	0.007711	0.163093	0.168192	1	75
Toll Like Receptor 7/8 (TLR7/8) Cascade	1	118	0.007777	0.16437	0.168192	1	79
Stimuli-sensing channels	1	120	0.007909	0.16692	0.168192	1	33
TRIF(TICAM1)-mediated TLR4 signalling	1	121	0.007975	0.168192	0.168192	1	70
MyD88-independent TLR4 cascade	1	121	0.007975	0.168192	0.168192	1	72
Toll Like Receptor 9 (TLR9) Cascade	1	121	0.007975	0.168192	0.168192	1	80
Interleukin-1 signalling	1	125	0.008238	0.173261	0.173261	1	59
Bile acid and bile salt metabolism	1	125	0.008238	0.173261	0.173261	1	99
Regulation of Insulin-like Growth Factor (IGF) transport and uptake by Insulin-like Growth Factor Binding Proteins (IGFBPs)	1	127	0.00837	0.175784	0.175784	1	14
MyD88:MAL(TIRAP) cascade initiated on plasma membrane	1	133	0.008766	0.183311	0.183311	1	76
Toll Like Receptor TLR6:TLR2 Cascade	1	133	0.008766	0.183311	0.183311	1	78
Toll Like Receptor TLR1:TLR2 Cascade	1	136	0.008963	0.18705	0.18705	1	78
Toll Like Receptor 2 (TLR2) Cascade	1	136	0.008963	0.18705	0.18705	1	80
Formation of the cornified envelope	1	138	0.009095	0.189533	0.189533	8	27

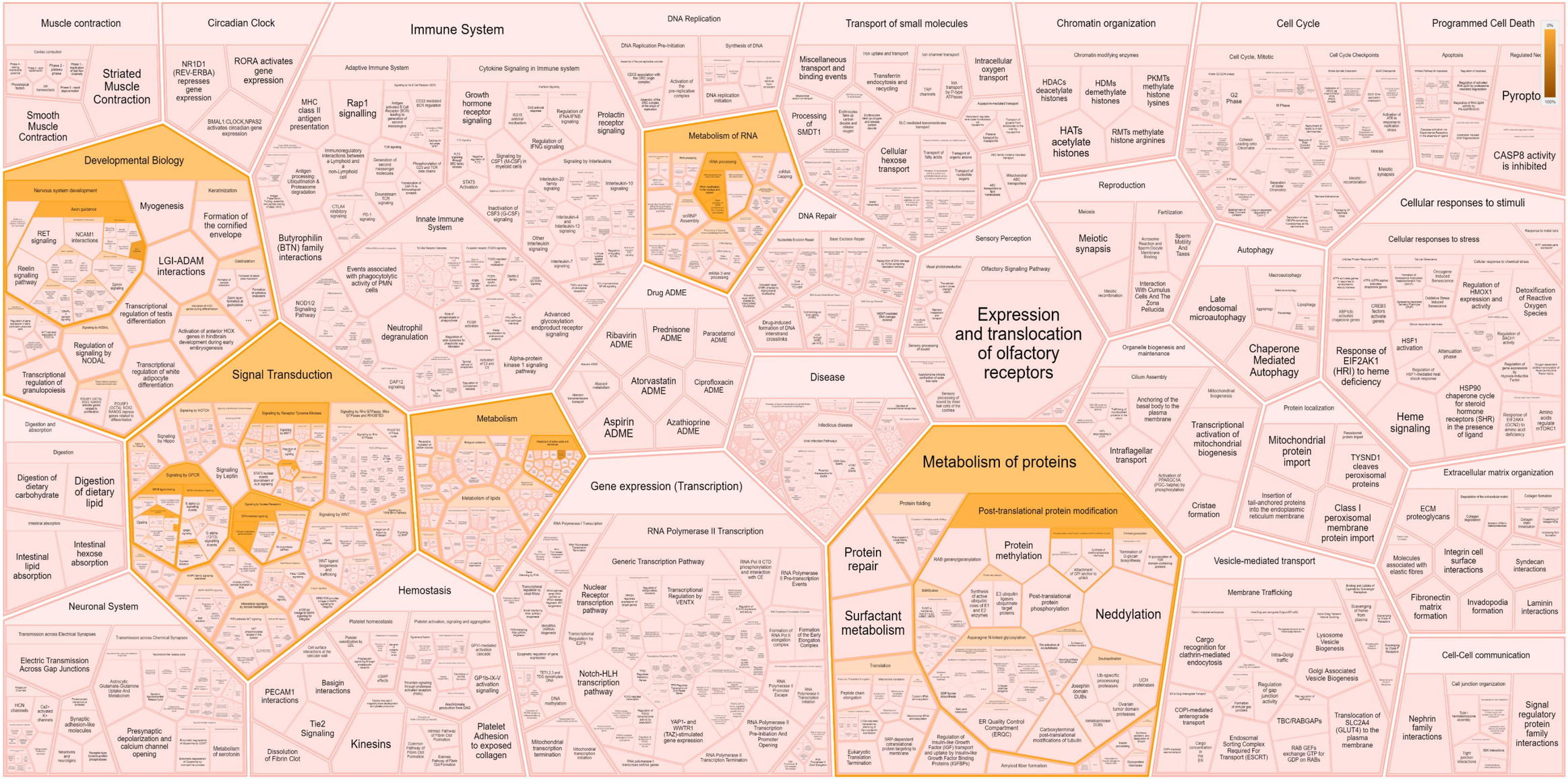
Immune System	6	2624	0.172939	0.195106	0.195106	27	1659
Cardiac conduction	1	147	0.009688	0.200618	0.200618	1	33
Golgi-to-ER retrograde transport	1	148	0.009754	0.201841	0.201841	2	18
Cytokine Signalling in Immune system	3	1036	0.068279	0.204926	0.204926	3	740
Toll Like Receptor 4 (TLR4) Cascade	1	165	0.010875	0.222355	0.222355	1	107
Binding and Uptake of Ligands by Scavenger Receptors	1	168	0.011072	0.225922	0.225922	2	33
Interferon gamma signalling	1	173	0.011402	0.231833	0.231833	1	18
GPCR ligand binding	2	609	0.040137	0.235361	0.235361	2	217
Interleukin-1 family signalling	1	183	0.012061	0.243526	0.243526	1	92
Factors involved in megakaryocyte development and platelet production	1	194	0.012786	0.256191	0.256191	2	43
Signalling by Interleukins	2	658	0.043367	0.263214	0.263214	2	505
Toll-like Receptor Cascades	1	202	0.013313	0.265275	0.265275	1	198
Peptide ligand-binding receptors	1	203	0.013379	0.266402	0.266402	1	83
Ion channel transport	1	208	0.013709	0.272017	0.272017	1	51
TCF dependent signalling in response to WNT	1	215	0.01417	0.279809	0.279809	1	71
Intra-Golgi and retrograde Golgi-to-ER traffic	1	219	0.014434	0.284226	0.284226	2	48
Keratinization	1	226	0.014895	0.291892	0.291892	15	34
G alpha (q) signalling events	1	285	0.018783	0.353464	0.353464	3	35
Vesicle-mediated transport	2	828	0.054571	0.359717	0.359717	4	252
Interferon Signalling	1	318	0.020958	0.385633	0.385633	1	74
Signalling by GPCR	2	876	0.057734	0.386421	0.386421	8	392
Signalling by WNT	1	331	0.021815	0.397881	0.397881	9	157
Metabolism of steroids	1	331	0.021815	0.397881	0.397881	1	250
Transport of small molecules	2	969	0.063863	0.436911	0.436911	2	454
Cellular responses to stimuli	2	1025	0.067554	0.466353	0.466353	6	481
Class A/1 (Rhodopsin-like receptors)	1	414	0.027285	0.470744	0.470744	1	185
G alpha (i) signalling events	1	426	0.028076	0.480553	0.480553	3	74
Neutrophil degranulation	1	478	0.031503	0.521087	0.521087	2	10
Axon guidance	1	585	0.038555	0.595183	0.595183	3	298
Developmental Biology	2	1313	0.086535	0.603556	0.603556	18	607
Nervous system development	1	621	0.040928	0.617547	0.617547	3	324
Membrane Trafficking	1	668	0.044026	0.644971	0.644971	2	219

ii. Upregulated in LBW compared to NBW (only protein coding genes)

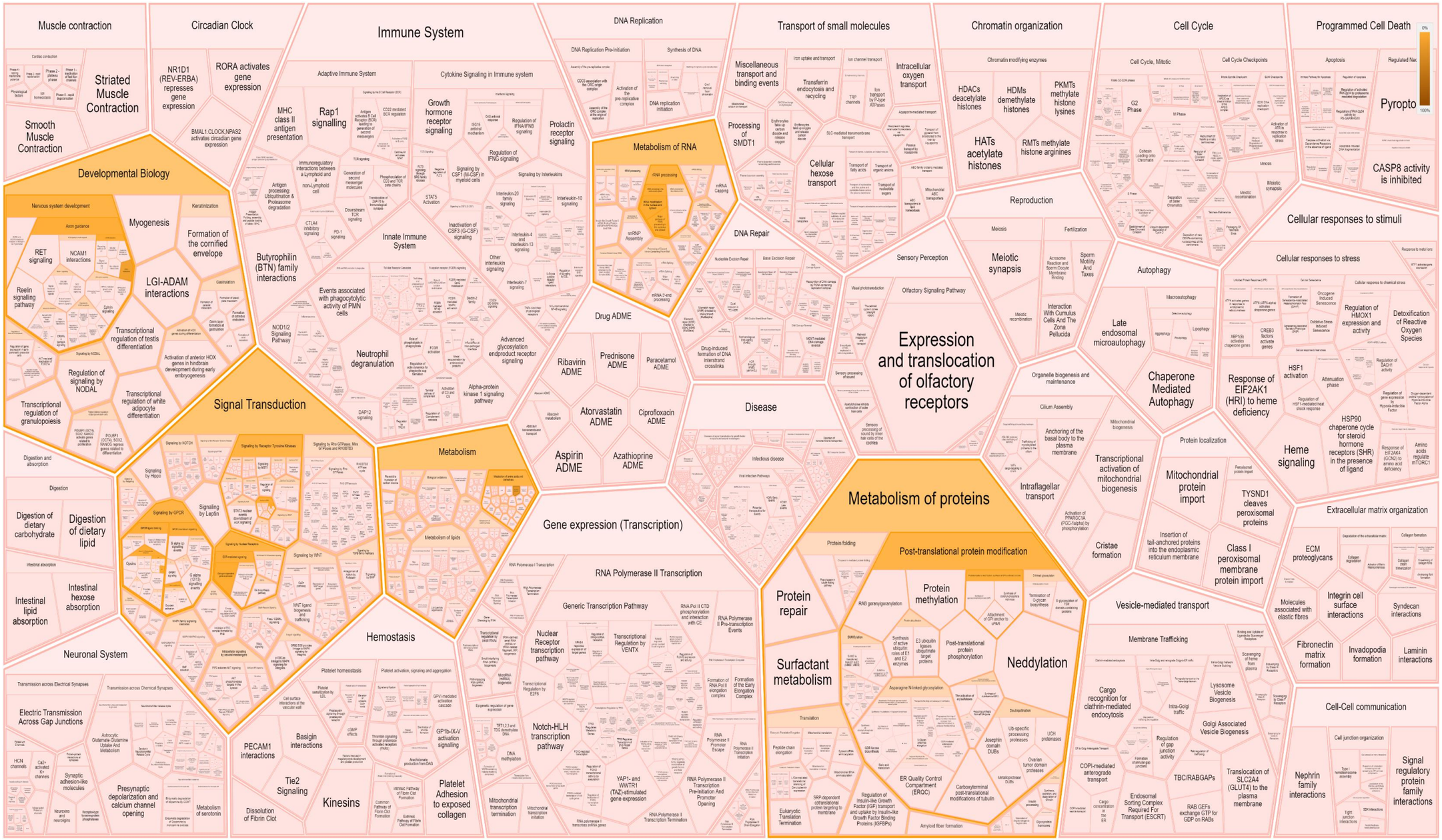
Pathway name	Entities found	Entities total	Entities pValue	Entities FDR	Reactions found	Reactions total
Chemokine receptors bind chemokines	2	57	0.00129051	0.041296313	2	19
Carnitine synthesis	1	4	0.0037882	0.060611206	1	4
CHL1 interactions	1	9	0.008505104	0.085051036	1	5
Peptide ligand-binding receptors	2	198	0.014476378	0.115811025	2	83
NGF-stimulated transcription	1	39	0.036382714	0.168594065	2	37
Class A/1 (Rhodopsin-like receptors)	2	333	0.038185573	0.168594065	2	185
rRNA modification in the nucleus and cytosol	1	60	0.055470982	0.168594065	1	8
Nuclear Events (kinase and transcription factor activation)	1	61	0.056371313	0.168594065	2	48
GPCR ligand binding	2	470	0.070870437	0.168594065	2	217
Post-translational modification: synthesis of GPI-anchored proteins	1	93	0.084773233	0.168594065	1	14
signalling by NTRK1 (TRKA)	1	117	0.105562206	0.168594065	2	102
Estrogen-dependent gene expression	1	119	0.107275091	0.168594065	2	64
L1CAM interactions	1	121	0.108984994	0.168594065	1	54
signalling by NTRKs	1	139	0.124240672	0.168594065	2	164
signalling by GPCR	2	713	0.143887655	0.168594065	5	392
Major pathway of rRNA processing in the nucleolus and cytosol	1	183	0.160537868	0.168594065	4	7
rRNA processing in the nucleus and cytosol	1	193	0.168594065	0.168594065	5	15
ESR-mediated signalling	1	195	0.170196845	0.170196845	2	111
rRNA processing	1	203	0.176579914	0.176579914	5	21
signalling by Nuclear Receptors	1	272	0.229809251	0.229809251	2	193
G alpha (i) signalling events	1	317	0.262818021	0.262818021	3	74
Metabolism of amino acids and derivatives	1	376	0.304147254	0.304147254	1	248
signalling by Receptor Tyrosine Kinases	1	543	0.40997372	0.40997372	2	746
Axon guidance	1	558	0.418722938	0.418722938	1	297
Nervous system development	1	584	0.43360931	0.43360931	1	323
Signal Transduction	3	2598	0.462070525	0.462070525	9	2530
GPCR downstream signalling	1	638	0.463425542	0.463425542	3	175
Metabolism of RNA	1	719	0.505481088	0.505481088	5	189
Developmental Biology	1	1138	0.679003789	0.679003789	1	606
Post-translational protein modification	1	1429	0.764672394	0.764672394	1	526
Metabolism of proteins	1	1949	0.867918634	0.867918634	1	795
Metabolism	1	2145	0.894616744	0.894616744	1	2031

D. Network analysis results

i. Downregulated genes in LBW compared to NBW (only protein coding)



ii. Upregulated in LBW compared to NBW (only protein coding)



END