# INDENTIFICATION OF DIFFERENTIALLY REGULATED GENES BETWEEN NORMAL AND TNF INDUCED IN HUMAN TRANSCRIPTOME

A MAJOR PROJECT DISSERTATION SUBMITTED

IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE DEGREE

OF

**MASTERS OF TECHNOLOGY**

IN

**BIOINFORMATICS**

SUBMITTED BY

**ABHISHEK GUPTA**

**(2K19/BIO/02)**

UNDER THE GUIDANCE OF

**PROF. YASHA HASIJA**

**DEPARTMENT OF BIOTECHNOLOGY**

**DELHI TECHNOLOGICAL UNIVERSITY**

Shahbad Daulatpur, Main Bawana Road,

Delhi- 110042, India

# CERTIFICATE

This is to certify that the M.Tech. Major report entitled “**INDENTIFICATION OF DIFFERENTIALLY REGULATED GENES BETWEEN NORMAL AND TNF INDUCED IN HUMAN TRANSCRIPTOME**” submitted by Abhishek Gupta (2K19/BIO/02) in partial fulfilment of the requirement for the award of the degree of Master of Technology from Delhi Technological University, is an authentic record of the candidate’s own work carried out by him under my guidance. To best of my knowledge this work has not been submitted in part and full for any Degree or Diploma to this University or elsewhere.

Date: 2nd September, 2021

02-09-2021

Prof. Yasha Hasija
Department of Biotechnology
Delhi Technological University

15.09.2021

PROFESSOR PRAVIR KUMAR
HEAD OF DEPARTMENT (BIOTECHNOLOGY)
DELHI TECHNOLOGICAL UNIVERSITY

# DECLARATION

I am Abhishek Gupta (2K19/BIO/02) student of M.Tech Bioinformatics, hereby declare that the project entitled Dissertation titled "**INDENTIFICATION OF DIFFERENTIALLY REGULATED GENES BETWEEN NORMAL AND TNF INDUCED IN HUMAN TRANSCRIPTOME**" which is submitted by me to Department of Biotechnology, Delhi Technological University, Delhi in the partial fulfilment of the requirement for the award of the degree of Master of Technology, is original and not copied from any source without proper citation. This work has not previously formed for the basis for the award of any degree, Diploma Associateship, Fellowship or other similar title or recognition.

Date: 30th AUGUST, 2021

ABHISHEK GUPTA
(2K19/BIO/02)

DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY

# ACKNOWLEDGEMENT

At the time of submission of my M.Tech. Dissertation, I would first like to thank GOD for giving me patience, strength, capability, and willpower to complete my work. Apart from our efforts, the success of this project depends largely on the encouragement and guidelines of many others. I, therefore, take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

My initial thank is addressed to my mentor **Prof. Yasha Hasija**, Department of Biotechnology, Delhi Technological University, who gave me this opportunity to work in a project under her. It was his enigmatic supervision, constant encouragement and expert guidance which have enabled me to complete this work. I humbly seize this opportunity to express my gratitude to him.

I would also like to extend my sincere gratitude to **Prof. Pravir Kumar** for providing basic infrastructure and facilities.

I would like to extend my sincere gratitude to **Mr. Rajkumar Chakraborty** for his keen observation and continuous advice. Her thoughtful inputs on the project issues have been invaluable for the productive progress of the project. I extend my thanks to technical staff Mr. Jitender Singh and Mr. C.B. Singh who had been an aid whenever required. Lastly, I wish to extend my thanks to my family and friends who have supported me through the entire process.

ABHISHEK GUPTA

(2K19/BIO/02)

## CONTENTS

## ABSTRACT:

According to Human Genome Project humans contains nearly 25,000 genes. Humans differs from each other just because of 0.1% of DNA. Since all the cells of an individual contains similar genetic material still they differ among one another in function and this is because of the differential gene expression. Since cells respond differently to different stimulus it is interesting to note their response. Gene expression is the main reason how cell gone a respond to different stimuli. In this study we have taken sequenced mRNA from human cell lines which was treated with TNF-alpha for some time period and try to identify all the differential expressed genes using RNA-Seq. We will extend our study to find out the pathways in which these genes have involved.

In this study we have used DESeq2 package for normalization, statistical analysis and for visualization of dataset. Genes obtained at the end of the analysis can be act as biomarkers for the cancer treatment.

## INTRODUCTION:

NF-kappaB: It is basically a transcription factor and like most of the transcription factors are protein, NF-kappaB is also protein and basically, it's a complex of protein. It involves in many functions such as regulating the response to stimuli and immune response. It also helps in the cytokines expression[1]. There are many genes whose expression depends on the NF-kappaB and many study suggest that genes whose expression depends upon this transcription factor are basically involved in cancer[2]. So, this transcription factor is present in the genes which will code for the proteins that are involved in apoptosis and other pathways which are part of cell cycle regulation[3]. So, this is the up regulation of this transcription factor that contributes to the resistance in the anticancer treatment.

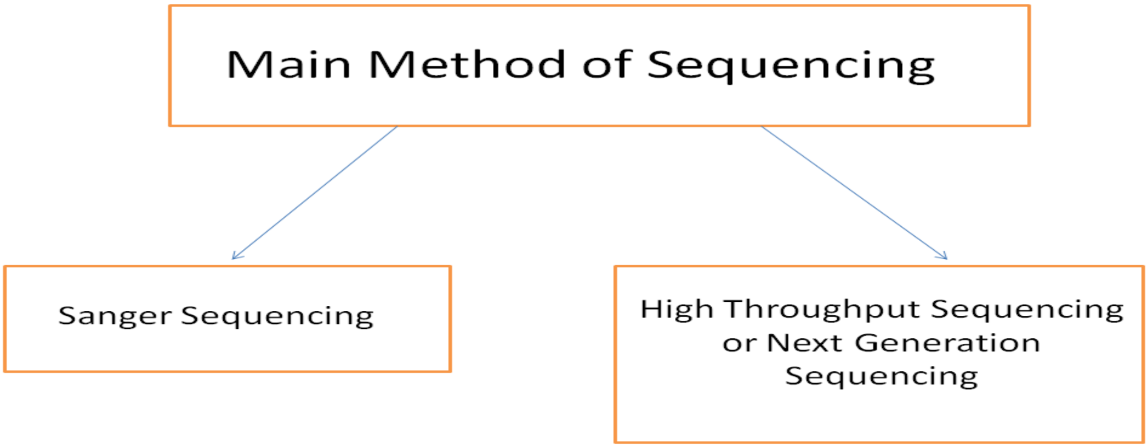
NF-κB as described above is a transcription factor and it's a dimer. NF-κB is formed by multigene NF-κB/Rel family member. When it comes to the humans there are only 5 proteins of this family out of which 3 are from Rel family and 2 are from NF-κB family and these are RelA which is also known as p65, RelB and the third protein of Rel family which is c-Rel. Proteins of NF-κB derived from precursors p105 and p100 and these small proteins are NF-κB1 also known as p50 and NF-κB2 which is also known as p52. For the activity of transcription subunits of NF-κB require to form dimer. These dimers can be formed by same unit i.e heterodimerization or by different unit of NF-κB homodimerization. Homodimers are generally not favorable because many studies showed that due to homodimers expression of gene was inhibited in some particular cases. In the general scenarios of cells i.e in resting cells NF-κB dimers wont able to bind with the DNA due to the presence of IκB and this inhibitor protein bind with the NF-κB dimers. Expression of this Inhibitor protein is controlled by responsive promoter of NF-κB. This NF-κB responsive promoter along with the TNFAIP3 gene (this TNFAIP3 is NF-κB dependent target) forms a protein which is known as A20 protein. This A20 protein is the crucial for the generation of protein complex that helps in NF-κB signaling. Now due to cellular stress or signals such as Pro-inflammatory extracellular signals is the main cause due to which IKK kinase activation took place. Due to the kinase activation phosphorylation of IκB also takes place. This IKK kinase have a regulatory subunit. Due to the Phosphorylation of inhibitor protein IκB this regulatory unit NEMO/IKK gamma of IKK is degraded because of ubiquitination. And now, NF-κB can able to bind to DNA regulatory elements because this NF-κB is free and can be translocated anywhere in the cell.

To find the gene expression in the whole genome RNA-Sequencing is the best method because of its advantage over other method which we had already seen. But RNA sequencing itself is a very tedious task and the researcher had to make many choices i.e., platform to be used, preparation of library protocol and this can also involve which kit to buy and all other decisions. Most famous kit for this method is TrueSeq (Illumina, catalog no. RS-122-2103). This method is popular because of its advantage over others such as it can start with the very less quantity of RNA (0.1-1 µg) and the other benefit it offers is that it can maintain the specificity of strand. As we all are aware that information of strand is crucial for analysis. Many a times some genes when undergo transcription they did it in reverse manner i.e., in anti-sense manner. This way their regulatory purpose could be fulfilled but this can also be associated with diseases. When it comes to the standard experiment which involves control and treatment, anti-sense strand can play a important there because most of the signals are coming from there only. So, if there is no annotation of sense and anti-sense then all the reads under study can be taken as “sense” signal. Also note that this is very common to find genes having anti-sense signal than sense signal.

## THEORY

### Nucleic Acid Sequencing

Sequencing of Nucleic acids is the most frequent technique used in the field of the Biotechnology. There are two types of nucleic acid sequencing possible as there are only two types of nucleic acids are present. So, the first one is the DNA sequencing and the other one is the RNA sequencing. Both of these sequencing is performed daily in different labs based on the purpose of sequencing. Both the nucleic acid has their own importance and in fact RNA can be obtained from DNA by the process of transcription and cDNA can also be converted to RNA in-vitro by the process of reverse transcription[4]. When one performs DNA sequencing it actually uses RNA. Extracting the RNA from the target and then converts it to the cDNA using the process of the reverse transcription and the then sequencing can be done using the different methods available. Sequencing refers to finding the sequence of nucleotides i.e. A, T, G, C in a molecule of DNA. There are different methods of sequencing present but there are 2 main methods of sequencing which are popularly used[5].


Main Method of Sequencing
Sanger Sequencing
High Throughput Sequencing or Next Generation Sequencing

Only a certain length of a nucleic acid can be sequenced in any method. For the sequencing of larger sequence, nucleotide need to split into the smaller fragments for sequencing and then these fragments be ensemble using different methods. Like Nucleic acid, protein sequencing can also be done using some methods[6].

Sanger Sequencing:

The method of Sanger sequencing follows the normal synthesis process that it contains all the raw materials needed in a reaction i.e., the Primer, Nucleotides, DNA polymerase, DNA template and the special type of nucleotides will also be added in very small concentrations and these are known as Dideoxyribonucleotides etc. Primer is a small piece of nucleic acid that is used by the DNA polymerase to start the process of the DNA synthesis because this enzyme called DNA polymerase can only add the new nucleotide to the 3' end of its previous nucleotides so that's why primer is required. These Dideoxyribonucleotides are special in a sense that these do not have any 3' end oxygen atom due to which no other nucleotide is able to form covalent bond with it and it would ultimately terminate the reaction and this is the main principle of the Sanger Sequencing[7].

Sanger Method of Sequencing:

In 1950s Sanger for the very first time sequenced the insulin protein by separating its chains and then overlapped the fragments to find its complete sequence[8].

Sanger's sequencing method involved three steps: -

1) PCR with the Dedeoxyribonucleotides (dNTPs)

In Sanger’s method of sequencing reaction mixture was prepared using the above mentioned ingredients and then primer will attach to the template of DNA and then DNA polymerase helps in synthesizing the phosphodiester bonds between existing nucleotides and in the newly joined nucleotides. Once the dideoxyribonucleotide will attach to the newly formed DNA chain it will terminate the whole process. This happens because dNTPs lack the 3’ Oxygen atom which involved in the phosphodiester bond formation. This results in many copies of the DNA terminated in between and then these copies helps in determining the original sequence of the DNA of interest[8].

Gel Electrophoresis:

In this step all the DNA molecules are loaded on a Gel and electric current is passed due to which DNA start moving from top to bottom as DNA carries negative charge. Since all the DNA molecules carries same mass to charge ratio so they will separate on the basis of the size. Smaller the size of the molecule the more it move from upper end to the lower end. This property become the foundation for our next step i.e., the determination of the original sequence[9].

Determination of Sequence using Gel analysis

As we all aware of the property of DNA polymerase that it can only synthesize new strand of DNA in 5’-3’ direction only when primer is attached to the template strand as the chain termination took place due to the dNTPs. These dNTPs correspond to a particular nucleotide in the original sequence due to which different lengths of molecules are present. The smallest molecule will only have 1 nucleotide which is dNTP and the second smallest is of length 2 having 1 nucleotide and another dNTP. In this way running all the molecules on the gel gave different bands at different position and by reading these bands determination of the original sequence can be done.

The only problem in the sequencing is that one cannot able to sequence more 1200-1300 nucleotide by this method and for sequencing larger genome it would take many years and even decade to complete. So, for sequencing of the larger DNA molecule or whole genome High throughput Sequencing is used which I will discuss in the subsequent section[10].

## LITERATURE SURVEY

### HISTORY OF DNA SEQUENCING:

In 1900s century, sequencing of 10-20 nucleotides is itself a very difficult task. In 1968 Bacteriophage lambda's 12 bases have been sequenced by Wu and this can be done with the help of the primer extension method and then 5 years later Gilbert and Maxam determine the 24 bases of the lactose-repressor binding site and to achieve this task Gilbert and maxam worked for 2 years or we can say that they can able to determine information of base at each month[11].

Then in 1976s Sanger develop a method of sequencing which is also known as the chain termination reaction and method was developed by Gilbert and both of these methods can able to sequence hundreds of nucleotides in a day. Both the method involves gel electrophoresis analysis to determine the original sequence. These both methods were the breakthrough of that time and had transformed the field of sequencing completely. So, these methods were accepted very quickly at that time by the researchers[11], [12].

### FILE FORMATS:

### FASTQ FORMAT:

FASTQ is a special type of file format that contains the information of the sequence and quality of these sequences in a single file. The quality can be measured in the form of a Phred score. Phred score can be calculated by identifying the ASCII character present in these files. This format is widely accepted as this is type of the format which is returned by the sequencer. This format contains information which is far greater than the FASTA format.

This format contains 4 lines: -

1) In the first line it has sequence identifier and it usually begin from "@" anything which is written after this "@" to the very first blank space is considered as the sequence identifier and anything written after the blank space is called as the sequence description.
2) The second line contains the sequence.
3) In the third line also, there is a same sequence identifier as that of the first line and it usually starts with "+" sign and it may also contain sequence description.
4) The Fourth line contains the ascii characters which are the measures of the Phred Score.

These Files are generally generated by Sequencer and these formats are compatible with many software and if they are not compatible these can be converted to some other file formats such as FASTA and others and then can be used for further analysis. This file format can also be generating through other file format such as BAM but they are also a kind of FASTQ files[13].

**SAM FORMAT:**

This is the first file format which was introduced for sequence alignment reports. SAM stands for Sequence Alignment Map. Although this File format requires lot of space than the file formats we have so if one face the problems related to memory then he generally avoid this file format. This file format contains the information related to the alignment, quality scores. Moreover, this file format also contains the information of the reads present in the original sample and it also contains information related to paired end, sample and many other information. This is reason for occupying large space by SAM files.

SAM format considered as the most powerful format since it can be easily read by humans and also this format can also be generated by every algorithm related to alignment that exists so far. It consists of lot of information for every read present in the sample i.e., it has header which is a row consists of 11 fields which contain information about each read present in the sample.[14]

**BAM FORMAT:**

This file format is generated from SAM format and the biggest motivation behind this format is to reduce the size of Alignment files. In this file format same information is encoded in binary due to which reduction in size took place. The biggest disadvantage of this format is that it’s not human readable due to which its need to convert to SAM format if it has to make readable. There are many different tools available which converts the BAM format into the SAM format[15].

**GTF FORMAT:**

GTF is a file format we will use in our study. GTF basically stands for Gene Transfer Format and as its name suggest it has relation with Gene Structure. So, this is a file format which contains information related to gene and its structure and this format can also be validated by the gene structure in general. This format basically derived from GFF which stands for general feature format through some modifications[16].

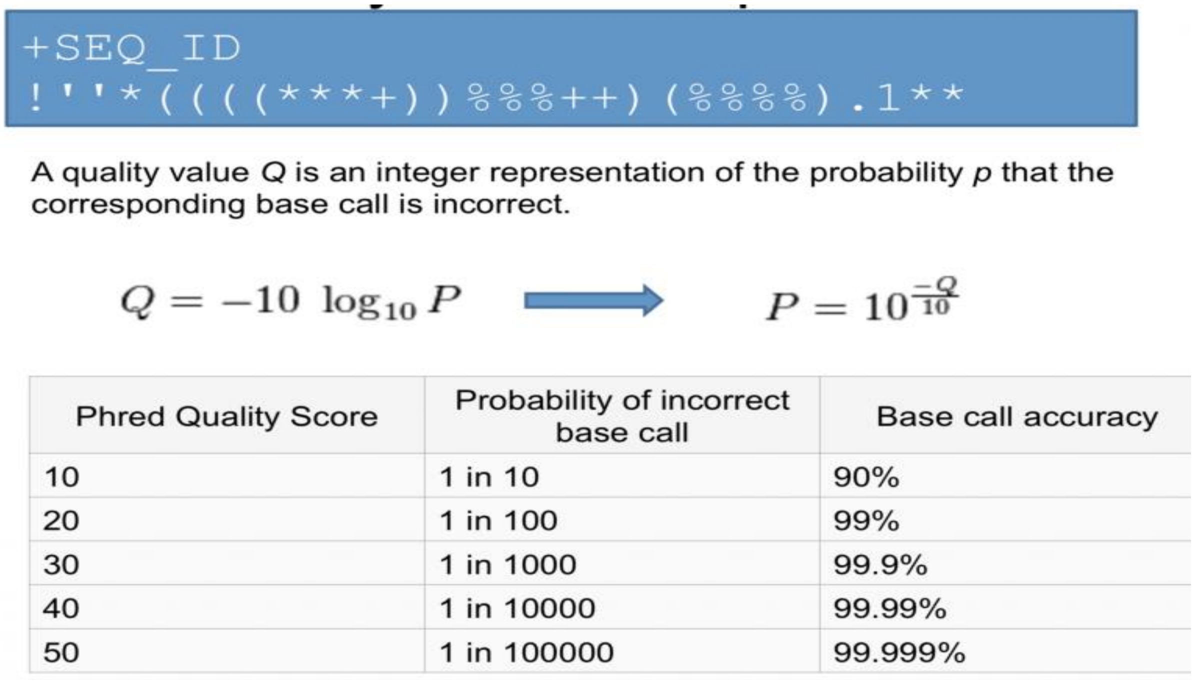
## Quality Score:

This quality score will be given to each and every base which is present inside the read. Different sequencer has different method of encoding but usually Phred-33 is most common and widely accepted by many sequencers. These quality score indicate the probability of the incorrect base at that particular position.

Phred quality scores is started from 0 and it would end at 40 but these are not used as such as encoding made working with these formats easy and it also reduce the size of the file. So encoding is the method of converting these values into other forms for example: - 2 is encoded by # and 40 by I and these characters are known as ASCII characters and they have certain probabilities attached along with them[17].

We can use the formula mentioned below to find the quality of the reads.

Q = -10 log10 P



```
+SEQ_ID
!''*((((***+))%%%++)(%%%%).1**
```

A quality value *Q* is an integer representation of the probability *p* that the corresponding base call is incorrect.

$$Q = -10 \ \log_{10} P \quad \Longrightarrow \quad P = 10^{\frac{-Q}{10}}$$

| Phred Quality Score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10000 | 99.99% |
| 50 | 1 in 100000 | 99.999% |

Any Phred Quality Score greater than 30 is acceptable.

### NORMALIZATION:

It is a process of much interest and importance when someone is dealing with large amount of data because it helps in the better visualization of data and it is one of the crucial steps of exploratory data Analysis. In our study also we are dealing with large amount of data so this becomes the very first step of the Differential gene expression analysis. There are two factors which are governing the expression of gene one is called interesting factors and another one is called uninteresting factors. So, in the process of normalization scaling of raw count values is performed with regard to uninteresting factor. So, the comparison between samples or within the samples become easy. Normalization also removes biasness to a great extent[18].

There are many factors which lead to the process of Normalization[19].

1) Sequencing Depth: This process is very much important when we are comparing expression of genes among the different samples. In some samples expression of genes is more than the others which results in the more sequencing depth in the sample.
2) Length of Genes: This factor is taken in consideration when we are comparing expression of genes within a sample. Some genes may have same level of expression but they differ in the length and the gene that has greater length may have more reads mapped to it.
3) Composition of RNA: This is also one of the important factor due to which the process of normalization become important. Sometimes contamination is present in the sample and sometimes some sample contains gene which are highly differentially expressed due to which normalization methods become less effective.

**Types of Normalization Methods:**

1) Counts per Million (CPM): It a method which is suitable when samples are presents as replicates as we have in our study. This method is basically effective where there is a sequencing depth factor. In this method counts will be scaled based on the total number of reads. This method is not applicable where comparisons will be made within the samples[19].
2) Transcripts per Kilobase Million: It is method of normalization where comparisons of gene counts are made and these comparisons can be made within a sample or they can also be made among the samples. Gene length and sequence depth are main factors for using this normalization method[18].
3) DESeq2's median of ratios: It is one of the best methods which is used for the normalization of Differential expression analysis. The main factors due to which this method is used when sequencing depth and RNA composition problem are there which we discussed above. This method uses statistical tools to scales the count[18].

Given below are few steps for Median of ratios method[19]:

1) In the first step we need to create a pseudo reference sample that helps in the determination of geometrical mean for example if have 2 samples than we calculate the mean by multiplying the number of reads of both the sample for that particular gene and then taking their square root.
2) In the next step we will find the ratio of each sample with respect to this pseudo reference sample. The ratio would change for only very few genes since majority of genes will not be differentially expressed.
3) Now we will calculate the normalization factor samples we have and to do so we need to find the median value of all the ratios be taken in the above step. Since each sample have different ratios, due which different samples have different median or we can say different normalization factor.
4) In the final step we would use our normalization factor to calculate the normal count values. As we have calculated the normalization factor of each sample so we divide each value in that particular sample through this normalization factor and the results we obtained are the normalized count values.

This method of normalization can be performed in 1 step in the DESeq2 package of R studio and the main algorithm behind this method is one we discussed above.

## DIFFERENTIAL GENE EXPRESSION:

We will generate feature count matrix before differential gene expression analysis. So, that data is very useful because it will be used in the analysis of differential gene expression. This data basically represents the count of sequence reads associated with the particular gene. It is very common to interpret the fact that the higher count means the more reads are associated with that gene and it ultimately represent the higher level of expression of that particular gene in that sample[20].

With the help of the differential gene expression here we are trying to find out the change in the gene expression level of 2 groups i.e., case vs control.

This correlation of the expression between two groups can be linked with some clinical outcomes as well as it can be helpful in finding out the biomarkers of genes.

For the purpose of the differential gene expression analysis of RNA -seq data, many software and packages are present and developed over the time. Among these many new and old tools developed so far, the most recommended and used tools are DESeq2 and EdgeR[21]. Both these methods not only give the same results but also used the same algorithms and methods i.e., negative binomial model. Both of these methods are very precise and both these methods are not sensitive but they are also very specific because they reduced false positive and false negative results[20].

These 2 are not the only tools for differential gene expression analysis but there is some other method such as Limma-Voom[22] is also very frequently used but this method is not very sensitive in case of small sample sizes. So that's why this method is used when we have large data i.e., when the count of replicate per group is greater than 20.

According to Soneson and Dleorenzi, 2013[23] these methods are not perfect in every condition i.e., these methods perform different in different conditions. So, these methods show great amount of variation even after sharing the lot of similarity among one another.

## DIFFERENTIAL GENE EXPRESSION USING DESEQ2:

We have worked with very large dataset in this study and it was not feasible to show all the data here directly. The only possible solution of this problem is to represent this huge dataset in the form of graphs. There are some specialized graphs that can able to represent large dataset very easily for example heatmap, volcano curve etc. Some of these plots can also be useful to represent other kind of dataset as well i.e., they are not unique to differential gene expression analysis[24].

To initialize the DESeq2 analysis in R-Studio version 3.6.2 we need to download and install some package of studio such as DESeq2, ggplot2 and others.

We also need to create Metadata for the purpose of analysis of dataset.

We had also created normalized data for every gene.

### Count normalization of dataset using DESeq2:

As we have already seen the theory of count normalization process in the earlier section. But now to implement the same logic there is a need to perform few steps apriori. These are mention below[25].

1) To begin with this step firstly we need to check the row names in metadata dataframe were in same order as the columns name in the counts dataframe.

2) There a need to create an object of the DESeqDataSet.
3) Finally, we generate the normalized count of this dataset.

**Match the metadata and counts data:**

As mentioned above it is mandatory to check the row and column name in the metadata and count dataframe respectively. If this requirement doesn't meet then the DESeq2 will show the error. We can check with the help of the below code snippet[25].

```
### Check that sample names match in both files
all(colnames(data) %in% rownames(meta))
all(colnames(data) == rownames(meta))
```

It is also possible to use match() function in case the data doesn't match.

**Creating DESeq2 Object:**

R-Studio has different kind of software in it and different packages are used for different kind of analysis for example for the purpose of biological data analysis we generally used Bioconductor. This specialized package has many facilities like user can define and can also use the custom class. With the help of this facility user can store data which is provided in the form of input and data which is obtained after initial analysis i.e., the intermediate data and the final result which is obtained after completing complete analysis. These data structure resemble with the list because this help used to store different datatypes within themselves. But there are some properties which make them different from the list. i.e., these data structures have some predefined data slots, these have some particular type of classes of data. Now to access the data stored in these data structure user needs to define some function which are specific to that particular package[26].

Now we are in the situation of creating our own DESeqDataSet object. The basic requirement to create this object is the metadata that we create earlier and the feature count matrix because both of these things act as a input for the purpose of this object creation. Now to create this object there should be a clear understanding of Design formula. This design formula contains the information about the specified column in the metadata table and how these columns will be used in the further analysis. When it comes to dataset, we have only one column (sampletype) that is of interest. This column is important because it has 3 factor levels and these levels contained some information and this information will be used by DESeq2 to evaluate gene expression level for each gene in terms of these 3 levels.

## GENERATING THE NORMALIZED COUNT: -

To make the fair gene comparisons among the samples generation of normalize count data is essential.

While discussing the process of normalization I have discussed about the median of ratios method of normalization and that method is quite lengthy and it need lot of mathematical calculations. But fortunately DESeq2 has a inbuilt function estimateSizefactors(). This function calling is enough to generate size factors for user. In case of RNA-seq analysis it is not mandatory to especially focus on this step because it will generally be performed by function of DESeq2[27].

```
dds <- estimateSizeFactors(dds)
```

The output of this function should be obtained in some object and to assign back the value of function to same object is important to complete the computation.

### Quality Control:

QC is very important step of the DESeq2 analysis, QC involves the steps of genes and sample level on the count data matrix which ensures that the data we have is good or not[28].

### Sample-level Quality Control:

First of all, it is mandatory to know what are the similarity we have in the samples.

- In our dataset we have lot of samples and due to high number of samples might be possible that some samples are similar and others are different. So, we need to find these samples.
- We also need to check whether these similarities fit to the expectation from the experiment's design?
- Another challenge is to find out the point of variation in our dataset.

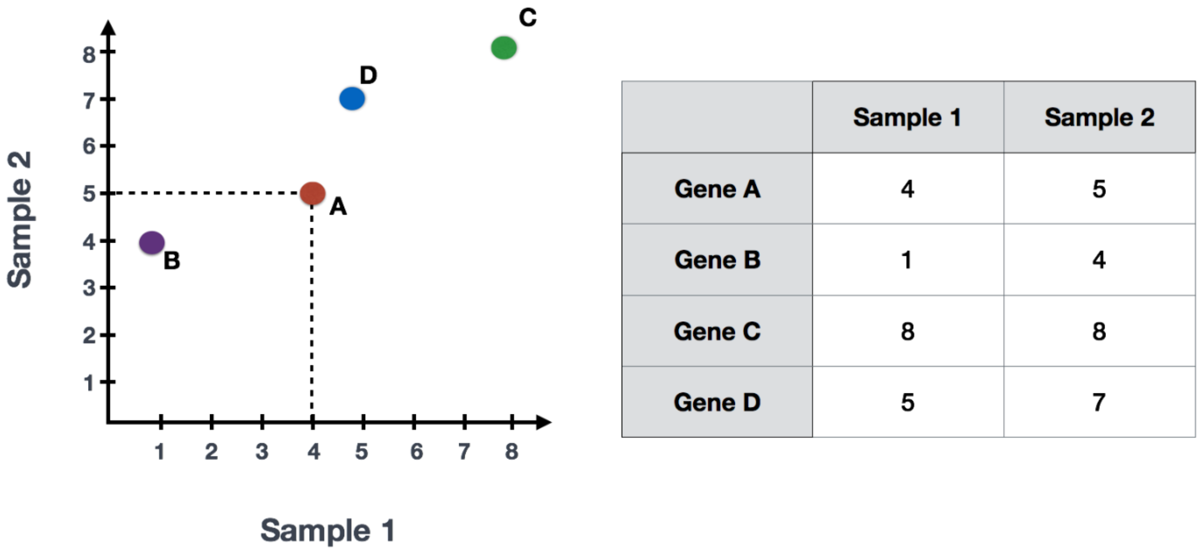
Now to find out the similarity in our sample there are two methods which can be useful i.e., Principal Component Analysis (PCA) and another one is the machine learning based method which is known as hierarchical clustering methods. As we all are aware about clustering it is very useful method in finding the similarity because all the samples which are similar can be put under one cluster. Clustering also helps in finding whether the experimental condition represent the point of variation in the data. This QC analysis also helpful in finding out the outliers if present in our data. Decision whether these outliers will be included or excluded will be taken based on our experimental conditions and if these outliers need to be removed, they will be removed from the dataset prior to the DE analysis[29].

In this process of unsupervised clustering there is a special method of normalization is used which is known as the log2 -transformation. This method improves the visualization by improving the clustering of the normalized count. By default, in sample level QC DESeq2 used a regularized log transform of the normalized counts due to its advantage. This transformation basically improves the clustering by managing the variance across the mean.

**Principal Component Analysis:**

As mentioned above one of the techniques used in sample level QC analysis is the Principal Component Analysis (PCA) because this technique basically focusses on the variation and due to which it can able find the pattern in the dataset and these pattern helped us in finding the similarity[30].

Let’s understand the theory of PCA with the help of the example. Let suppose we have some dataset which contains 2 samples and 4 genes. As discussed PCA help us in finding the similarity so our next aim is to find the similarity between these 2 samples. To find out these similarities we can able to plot the counts of both these samples using x and y axis of graph. Let suppose we will plot sample 1 on the x-axis and sample 2 on the y-axis.



| | Sample 1 | Sample 2 |
|---|---|---|
| Gene A | 4 | 5 |
| Gene B | 1 | 4 |
| Gene C | 8 | 8 |
| Gene D | 5 | 7 |

*Figure 1 source: GITHUB* https://hbctraining.github.io/DGE_workshop/lessons/03_DGE_QC_analysis.html#mov10-quality-assessment-and-exploratory-analysis-using-deseq2

Now to perform PCA analysis, the initial step is to find the best fitted line that can be drawn through this data. If we noticed carefully all the variation in this graph is through the diagonal. This means that most of the points of variations are present in between these 2 endpoints of the line. And this particular plot is known as the first principal component or it can also be termed as the PC1. Gene B and Gene C are present on the endpoints and they are basically deciding the direction of this line.

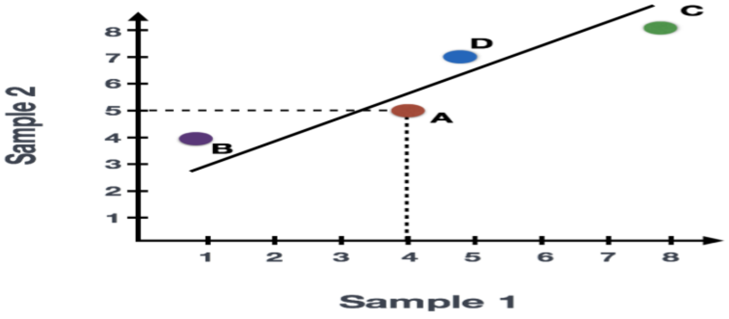
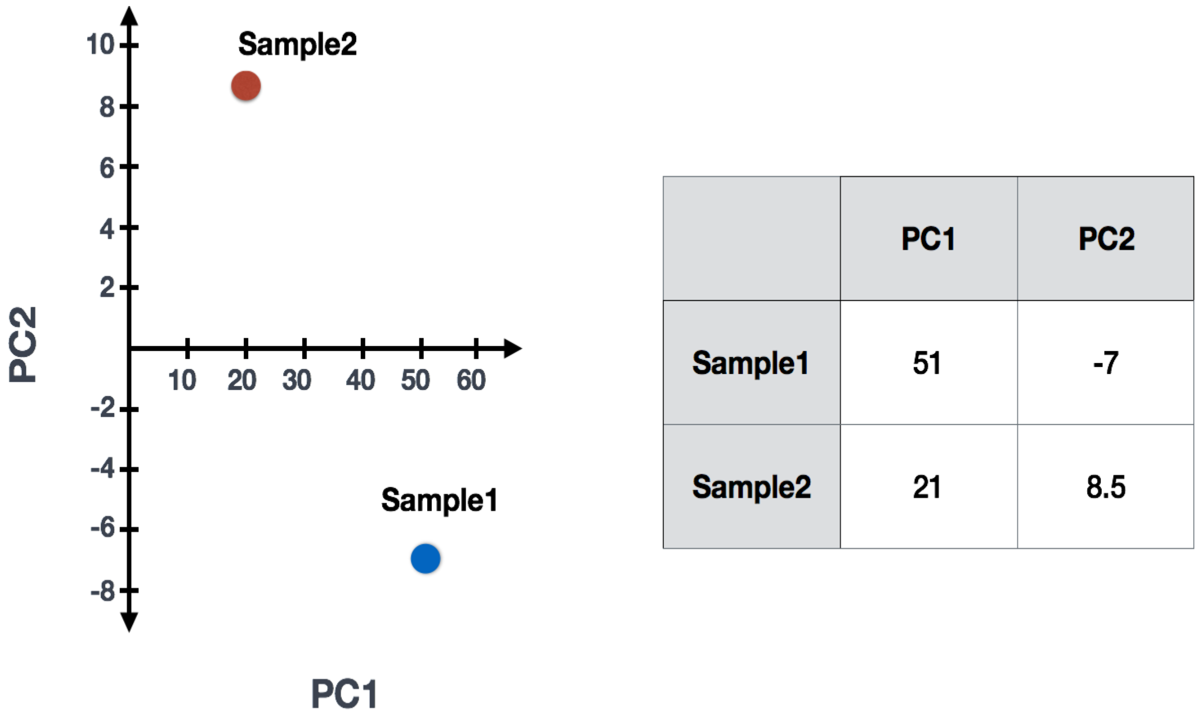


*Figure 2 source: GITHUB [https://hbctraining.github.io/DGE_workshop/lessons/03_DGE_QC_analysis.html#mov10-quality-assessment-and-exploratory-analysis-using-deseq2](https://hbctraining.github.io/DGE_workshop/lessons/03_DGE_QC_analysis.html#mov10-quality-assessment-and-exploratory-analysis-using-deseq2)*

Now we have already drawn the line and also observed the influence per gene its time to move forward on the next step of the PCA i.e., in next step PCA will calculate the per sample score and to calculate this PC1 score is very simple and this can be done by multiplying the normalized read count to the influence and then sum all this score for all the genes present in the sample. Now it is also possible to plot another line the data which have 2$^{nd}$ most variation in the data and then we could able to compute scores and this plot is known as the PC2. In similar way it is possible to plot the third line and the compute the scores and plot is named as PC3. This can be done repeatedly until we will hit all the samples present in our dataset.

```
Sample1 PC1 score = (read count Gene A * influence Gene A) + (read count Gene B *
influence Gene B) + .. for all genes
```

But our dataset is very large because it contains 8-9 samples and thousands of genes. Now remember the first step where we have plotted the sample 1 on x-axis and sample 2 on the y-axis. So accordingly, our sample will be plot in n-dimensional space and we have 9 samples so it would be represented in 9-dimensional space. But fortunately, our end result will be in the 2-dimensional space i.e., the columns representing the samples and columns representing the scores for each of the PC. But we want the final plot of Principal Component Analysis and it can be obtained be plotting all the PCs against each other.



| | PC1 | PC2 |
|---|---|---|
| Sample1 | 51 | -7 |
| Sample2 | 21 | 8.5 |

*Figure 3 source: GITHUB* *https://hbctraining.github.io/DGE_workshop/lessons/03_DGE_QC_analysis.html#mov10-quality-assessment-and-exploratory-analysis-using-deseq2*

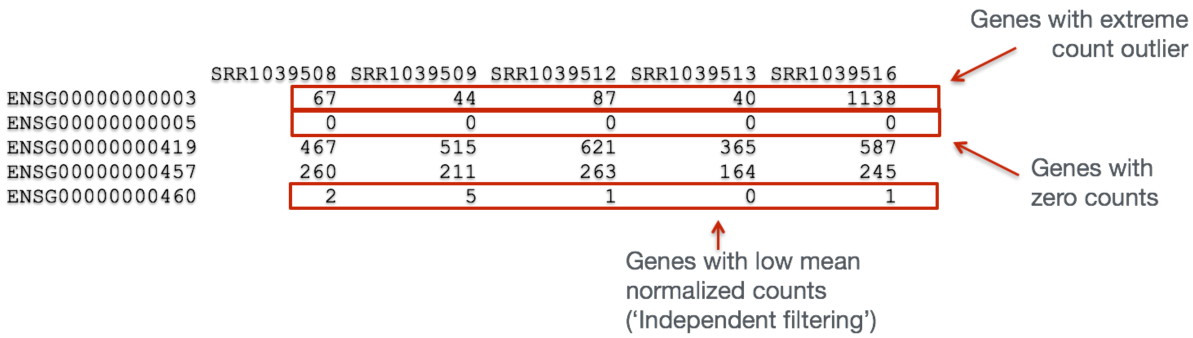
Samples that have same or similar levels of expression for genes and if these genes are also contributing in the variation which was represented by PC1, then they will be plotted together in the PC1. This is the reason why biological replicates will have same score and they will be clustered together on Principal Component. Similarly, samples which are from different group or in different treatment group will have different score and they will be plot away from each other. This can be better understood by some examples.

GENE LEVEL QUALITY CONTROL:

In the previous step we have successfully checked the Sample level QC with the help of the PCA. Now it is also important to check QC at Gene level because there might be chance some of the genes might not expressed differentially and there are some others which will have very low expression. Removing these genes from the dataset can improve the overall accuracy as well time of computing will also be reduced[31].

There are three types of genes which we want to remove from our dataset.

- Genes that have no (0) counts in all the samples.
- Genes with a very low mean normalized count.
- Genes which act as outliers.



| | SRR1039508 | SRR1039509 | SRR1039512 | SRR1039513 | SRR1039516 |
|---|---|---|---|---|---|
| ENSG00000000003 | 67 | 44 | 87 | 40 | 1138 |
| ENSG00000000005 | 0 | 0 | 0 | 0 | 0 |
| ENSG00000000419 | 467 | 515 | 621 | 365 | 587 |
| ENSG00000000457 | 260 | 211 | 263 | 164 | 245 |
| ENSG00000000460 | 2 | 5 | 1 | 0 | 1 |

Genes with extreme count outlier

Genes with zero counts

Genes with low mean normalized counts ('Independent filtering')

*Figure 4 source: GITHUB https://hbctraining.github.io/DGE_workshop/lessons/03_DGE_QC_analysis.html#mov10-quality-assessment-and-exploratory-analysis-using-deseq2*

DESeq2 will perform these filtering automatically and it is crucial step in case one is not using DESeq2 tool for analysis.

## DIFFERENTIAL EXPRESSION OF GENE USING DESeq2:

This is the final and the most important step of this whole analysis because we have so much of pre-processing starting from normalization to eliminating the genes which were not contributing in this study. So, in this step we will input this final normalized data to the NB models. Along with fitting this data we will perform certain statistical test that is essential for differentially expressed genes. These statistical tests will help us in knowing whether these results that is obtained by mean expression levels of different sample groups are statistically significant or not.

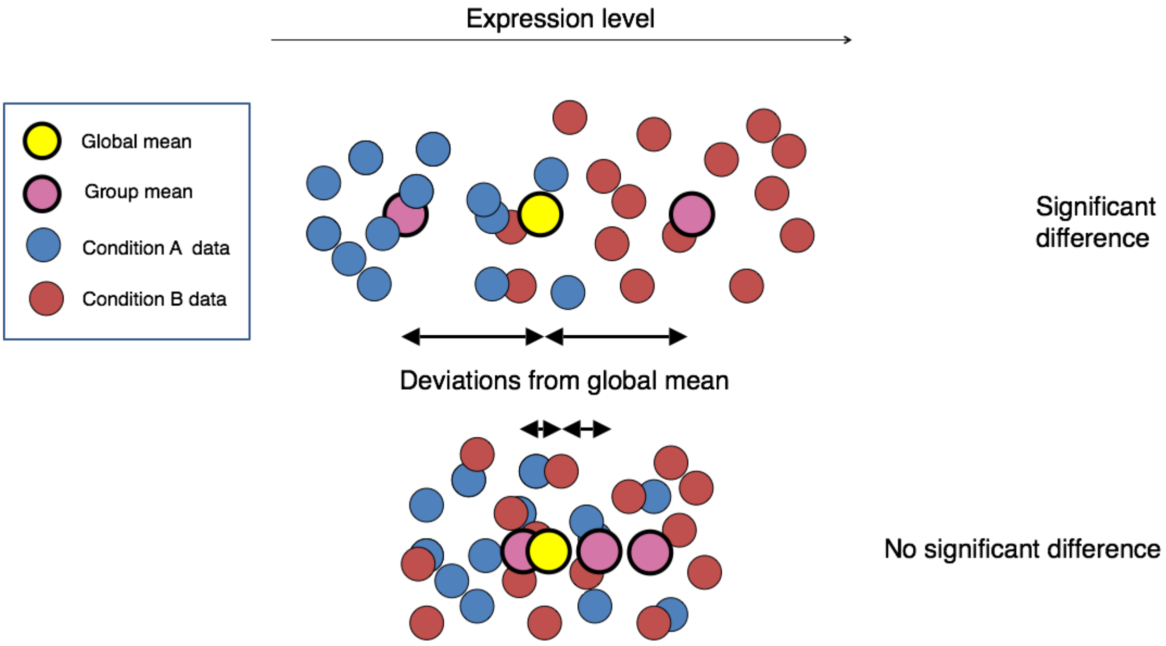


*Figure 5 Image credit: Paul Pavlidis, UBC*

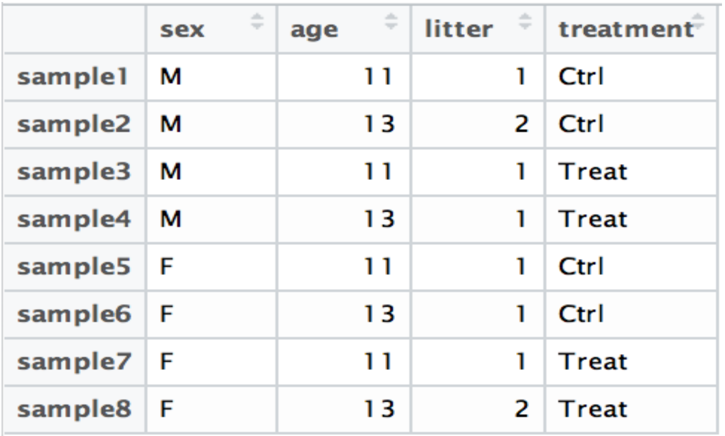
DESeq2 is not very old tool people came to know about it in 2014 after its paper was published, but this tool is continuously updated by the developers and it has made available to R users through Bioconducter. The idea behind this tool is to have better dispersion estimation method than the existing tools. This method also uses Generalized Linear Models from other methods.

When we use differential gene expression with the help of DESeq2 it is not a single step process. It's a multistep process and involved many tedious processes which can be performed by this tool by just calling the function. DESeq2 basically calculates the difference in the library depth by just modelling the raw counts we have with the help of the normalization. Along with this it also finds the dispersion gene-wise. After these steps it basically narrow down these estimates of dispersions and it will try to find out the more accurate dispersions that is able to model the counts. After all these steps DESeq2 can able to fit the models along with the hypothesis testing. For the hypothesis testing it can either use Wald test or Likelihood Ratio Test.

We have already performed Quality Control both at the gene level and the sample and it basically found the source of variation in our dataset and it's always recommended to know the source of variation in advance by any mean. Having good knowledge of the domain along with the dataset can also help in finding out the source of variation. Again, I mentioning that finding the source of variation can provide some advantage i.e., we can remove these sources of variation prior to our analysis or we can control them by adding these sources of variation in our design formula[32].

DESIGN FORMULA:

Design Formula is nothing but an information one pass to the statistical software. This information can be factor of interest we want to test or it can be source of variation that we are interested in controlling. Let suppose our dataset have 2 major sources of variation i.e., gender and age. So, we include both of them in our model. In case we have more two sources of variation then one can make use of metadata and include all the source of variation. The factor of interest should be entered at the last in our design formula. Let say we have the following data in our metadata[33].



|  | sex | age | litter | treatment |
|---|---|---|---|---|
| sample1 | M | 11 | 1 | Ctrl |
| sample2 | M | 13 | 2 | Ctrl |
| sample3 | M | 11 | 1 | Treat |
| sample4 | M | 13 | 1 | Treat |
| sample5 | F | 11 | 1 | Ctrl |
| sample6 | F | 13 | 1 | Ctrl |
| sample7 | F | 11 | 1 | Treat |
| sample8 | F | 13 | 2 | Treat |

*Table 1 shows the example of major source of variation, source: GITHUB https://hbctraining.github.io/DGE_workshop/lessons/03_DGE_QC_analysis.html#mov10-quality-assessment-and-exploratory-analysis-using-deseq2*

In order to evaluate the expression differences among treatments and we know the major source of variation are age and sex then we can write the expression of design formula by:

This (~) sign be used to prior to the factors which are source of variation in our metadata and in also indicates DESeq2 to use this information for the design formula. The column name in the metadata should match factors included in the design formula.

There is a two line of code for the purpose of performing Differential Expression analysis we will write the code along with other information further in the material and method section. But here we will discuss all the steps that 2 lines of codes will perform for us.

So, the function we used further is DESeq() function this function alone can perform many steps for us. These steps are:

- Estimating size factors
- Estimating dispersions
- Gene-wise dispersion estimates.
- Mean-dispersion relationship
- Final dispersion estimates
- Fitting model and testing

DESeq2 also provide the individual function to carry out these steps individually but in our study we use DESeq() function. But we will understand these steps individually for better understanding.

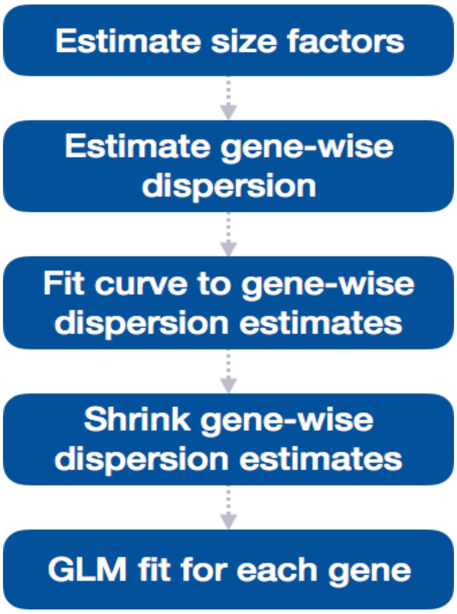

Estimate size factors
Estimate gene-wise dispersion
Fit curve to gene-wise dispersion estimates
Shrink gene-wise dispersion estimates
GLM fit for each gene

*Figure 6 source: GITHUB https://hbctraining.github.io/DGE_workshop/lessons/03_DGE_QC_analysis.html#mov10-quality-assessment-and-exploratory-analysis-using-deseq2*

The above diagram shows the workflow of the DE analysis and we will discuss them one by one.

1) **Estimate size factors: -** As mentioned in the workflow this is the initial step of our study. This will be performed in a similar way we did for the normalize the raw counts. As mentioned earlier DESeq2 will do this step by its own but in case if perform this step initially for the purpose of normalization then DESeq2 will not calculate these values again but will use these values only[34].

2) **Estimate gene wise dispersion: -** Dispersion refers to the measure of spread in the data. There are few methods to find out the dispersion these includes variance, standard deviation, Interquartile range and others but DESeq2 has its own way to calculate the dispersion by using mean and variance.

```
Var = μ + α*μ^2 here μ is mean and α is dispersion
```

Genes having very high-count values, value of the dispersion square will be equal to the (var/μ). So, 0.02 dispersion can be interpreted as 20% variation around the mean which is expected across all the biological replicates.

As we have seen in the formula that dispersion estimates are directly proportional to the variance and it is inversely proportional to the mean. So, we can conclude dispersion based on these two quantities.

3) **Fit Curve to gene wise dispersion estimates: -** So, this is our next step after finding the dispersion. This step is also crucial for purpose of visualization because in this step we will fit a curve to the dispersion estimates for each and every gene taken in study. As we know that every gene varies from each other so they also have dissimilar biological variability and this in case of large number of genes there is a distribution which defines the estimates of dispersion[33].

4) **Shrink gene wise dispersion estimates: -** this is the next step of the workflow and again as the name suggest in this step, we will shrink the values of dispersion estimates values obtained in the last step to the value of dispersion that is expected[35].

   When the sample size is small then this curve helps in the identification of DE genes more accurately. We can't shrink the value of genes according to our wish but it depends on certain factor such as:

   - Closeness of each gene dispersion estimate from the curve we will plot.
   - How much sample we have (lesser the sample more is the shrinkage).

Shrinkage offers great advantage to our analysis because it can eliminate many false positives from the dataset. Genes that have very low dispersion estimate can be easily shrunken to the curve when compare to the genes that have high dispersion estimates. If any dispersion estimate is slightly far from the curve, then that can be easily shrunk towards the curve than the genes having high dispersion estimates. Because some genes are not able to modelling assumptions that’s why they have higher variability. Shrinking forcefully reduces the accuracy of the model and can only increase the number of false positive.

Last step involves lot of mathematics and difficult to understand so will not going to discuss that.

## Materials and Methods

### 1) Dataset:

We used mRNA which was sequenced from U2OS cells (wild type) and the cells which are transfected for very short time with siRNA specific for RELA gene or control siRNA. There were 3 controls or can also be named as untreated cells and there are cells which are stimulated using the TNFalpha cytokines for some specific period of time. We have taken sample which is treated with TNFalpha cytokines for 30 mins and its 2 replicate and another sample which is treated with TNFalpha cytokines for 4 hours and its replicates.

| S.No | Sample Name | Accession No | Type of Sample | Total Bases |
|---|---|---|---|---|
| 1. | GSM2990358: WT ctr rep1; Homo sapiens; RNA-Seq | SRR6701565 | Control | 48388 |
| 2. | GSM2990359: WT ctr rep2; Homo sapiens; RNA-Seq | SRR6701566 | Control | 15688 |
| 3. | GSM2990360: WT ctr rep3; Homo sapiens; RNA-Seq | SRR6701567 | Control | 26311 |
| 4. | GSM2990361: WT TNF30min rep1; Homo sapiens; RNA-Seq | SRR6701568 | Treated | 23926 |
| 5. | GSM2990362: WT TNF30min rep2; Homo sapiens; RNA-Seq | SRR6701569 | Treated | 22982 |
| 6. | GSM2990363: WT TNF30min rep3; Homo sapiens; RNA-Seq | SRR6701570 | Treated | 24703 |
| 7. | GSM2990364: WT TNF4h rep1; Homo sapiens; RNA-Seq | SRR6701571 | Treated | 15884 |
| 8. | GSM2990365: WT TNF4h rep2; Homo sapiens; RNA-Seq | SRR6701572 | Treated | 30204 |
| 9. | GSM2990366: WT TNF4h rep3; Homo sapiens; RNA-Seq | SRR6701573 | Treated | 16440 |

Above mentioned samples are controls and treated samples for our study. The first 3 samples are controls and the other samples are treated with TNFalpha cytokines for 30 minutes and 4 hours respectively. All these samples are downloaded from Sequence Read Archive (SRA) Genbank[36].

These samples are downloaded using the following steps:

i) Searched for SRA explorer in the Broswer.
ii) Opened the SRA Explorer in the Browser and then searched for SRP132529 and then these sequences can be easily retrieved from there in a FASTQ format.

2) **Quality Check:**

**A) FA STQC:**

For the purpose of checking the quality of the raw data coming from NGS pipeline FASTQC is a very crucial step. We performed FASTQC on the data we extracted from SRA explorer and this step helped to locate the problem in the data. In this step we check for some important parameters such as Phred Score, GC content and some other important parameters. By performing this step, it would be clear whether the samples are free from contamination or not.

We used some Linux command to run the FASTQC non-interactively. In this way we specified the list of FASTQ files on the command line. By running the FASTQ files non-interactively we obtained the HTML files for all the samples and these files contain all the information along with graphs to analyse the data easily[37].

**B) MULTIQC:**

MultiQC is a tool very similar to FastQC and it can only be performed to summarize the results of FastQC for all the samples. This is done to summarize the result at one place. MultiQC is performed on the files generated by FastQC. Like FastQC we also performed MultiQC in a non-interactively method using the commands of linux. In this step we mentioned the name of FastQC files on the command line and then run the MultiQC command on command line and then we obtained a output file[37].

3) **Downloading Data of Reference Genome:**

After the checking the quality of data, it’s now become important to perform the mapping of the dataset to the reference genome. In our study we will map our dataset on the Human genome. So, to map our dataset we need to download our reference dataset and then we will perform mapping. To download the reference dataset opened the web link (http://daehwankimlab.github.io/hisat2/download/#h-sapiens) and then copied the link of GRCh38 index files or we can also directly downloads these files[38].

We have also downloaded the gtf files from the weblink given below.

(ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_37/gencode.v37.annotation.gtf.gz）.

4) **Mapping of Dataset to Reference Genome using HISAT-2:**

In earlier step we have downloaded the index and gtf files of reference genome. Now to map our dataset on the reference genome there is a tool called called HISAT-2 which we had used for study. This tool takes the dataset and reference genome as input and provides BAM or SAM files as output. It depends upon the requirement which file it gave as output. Since SAM files are very bigger in size, we have converted them into the BAM files and then used them as for the further analysis.

This step of mapping is very crucial since we don't know where exactly these genes are present. We performed this step of mapping on each sample and then used this information further. Since this step required high computing power so this step took of time[39].

5) **Generation of Feature Count Matrix:**

In order to perform differential gene expression analysis, we need to create feature count matrix for all these samples and this matrix will be used further. Here features count matrix represents how many reads of particular sample align to the genes under study (these genes are treated as the features)[40]. For example, let say we have 4 Samples A, B, C, D and we have 5 genes i.e., Gene1, Gene2, Gene3, Gene4, Gene5. So the feature matrix can be generated as given below.

| | A | B | C | D |
|---|---|---|---|---|
| Gene1 | 102 | 88 | 65 | 130 |
| Gene2 | 76 | 54 | 10 | 32 |
| Gene3 | 109 | 77 | 98 | 21 |
| Gene4 | 43 | 83 | 78 | 142 |
| Gene5 | 204 | 92 | 98 | 6 |

So, this is how features matrix will be generated for our samples and we will use this matrix for further analysis.

## 6) Statistical Analysis and normalization using DESeq2:

- **Transformation normalized counts through rlog transformation:**

In order to improve the clustering and distances in case of PCA and other visualization methods, there is a need to use this rlog transformation method.

This method is only crucial in case of visualization of the QC assessment and this method we won't use anywhere else.

```
### Transform counts for data visualization
 rld <- rlog(dds, blind=TRUE)
```

This is the simple function of R script which is enough to perform this step. Here we have an argument "blind=TRUE" basically this argument helps in the transformation which doesn't contain biasness due to sample condition information.

This rlog function provide us the return value in the form of object (DESeqTransform), this object is also specific to the DESeq2. We will use this object as such while plotting PCA and other methods that we will use in quality assessment[26].

We can also vst() function instead of rlog() function in case we have more than 20 samples. Because vst() function is faster than the rlog() function.

- **Generating PCA plot:**

The biggest advantage of using DESeq2 is that it has may built-in function and calling them enough to generate plots and matrices. Interestingly, we have built-in function for PCA as well and it make use of the ggplot2 package. The function name is plotPCA() and by just passing parameters in this function we are done with generating PCA plot. This function accepts that rlog object that we had obtained in the previous step as input[29]. So, this method not only reduce time but also prevent us from writing long codes to extract information form rlog object.

This method can only take two arguments as input:

- rlog() object that we discussed above.
- intgroup (this basically contains the name of the columns of interest in our metadata).

```
### Plot PCA
plotPCA(rld, intgroup="sampletype")
```

- **Hierarchical Clustering:**

To generate the heatmap we have to perform few tasks manually since there is no built-in function for it in DESeq2. But in our study we used pheatmap() function which is available in the pheatmap package. But unfortunately, it doesn’t accept rlog object as input instead it needs numeric values dataframe as input parameters[41]. For obtaining these dataframes information one should retrieved information from that object.

```
### Extract the rlog matrix from the object
rld_mat <- assay(rld)   ## assay() is function from the "SummarizedExperiment" package that was loaded when you loaded DESeq2
```

After running this code, we need to calculate the pairwise correlation values for all the samples. It can be achieved through cor() function.

```
### Compute pairwise correlation values
rld_cor <- cor(rld_mat)    ## cor() is a base R function

head(rld_cor)   ## check the output of cor(), make note of the rownames and colnames
```

Now we can use these correlation values to generate heatmap:

```
### Plot heatmap
pheatmap(rld_cor)
```

Results of above 2 processes will decide whether the data is of good quality or not and if the data is of good quality, then we can proceed to next step.

- **Differential Expression Analysis using DESeq2: -**

As we have discussed in the theory of differential Expression analysis that this is a multi-step process and does require significant amount of computing power and time but thanks to DESeq2 because with just 2 lines of code we will get the result of our DE but we have all the steps involved in it. But since we discussed lot about in theory now, we can run this pipeline on our raw counts.

First step involve in this pipeline is creating a DESeqDataSet in a same way we did it in normalization and then we mention the location of our raw data counts and metadata as input along with the design formula[33].

```
## Create DESeq object
dds <- DESeqDataSetFromMatrix(countData = data, colData = meta, design = ~ sampletype)
```

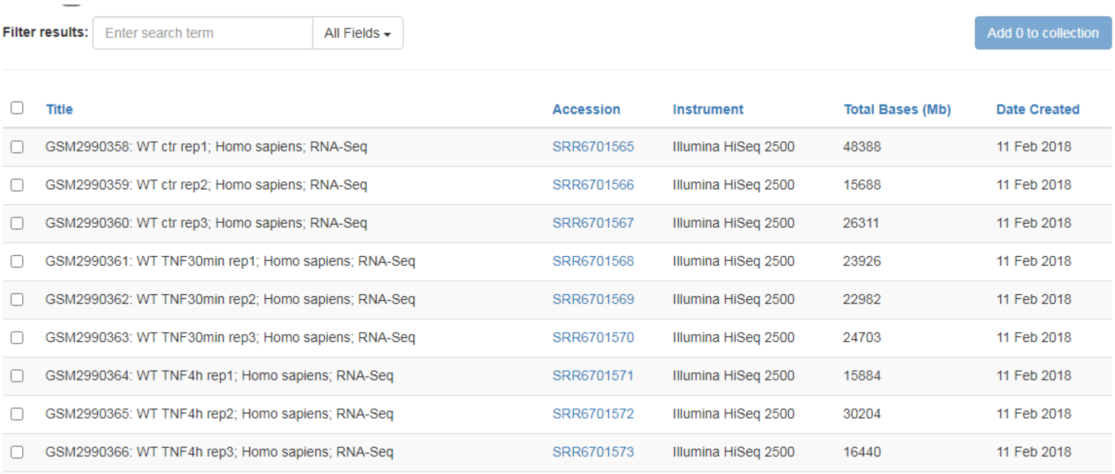
now the second step just involves the calling of DESeq() function to run the actual DE analysis.

```
## Run analysis
dds <- DESeq(dds)
```

7) **Pathway enrichment Analysis:** After completion of normalization and statistical analysis now we are interested in knowing the functions of these genes which are obtained in the study. Because knowing their function and the pathways in which they are involved basically help us to gain more knowledge about effect of TNF treatment on human genes. Some of these genes can act as a biomarker for any disease. To perform the pathway analysis, we have used a tool known as CPDB (http://cpdb.molgen.mpg.de/). This tool also provides the information about the Gene Ontology and some other information and also it is easy to use[42].

## Result and Discussions

1) **DATASET:** - Dataset was successfully downloaded from the link mentioned in the Materials and Methods. One should be careful while downloading the dataset as incomplete downloading can affect the results completely and that could lead the truncation or may also affect the quality of dataset. And these errors are very common but can be identified in the step of Quality Check i.e., in the case of FASTQC and MultiQC.

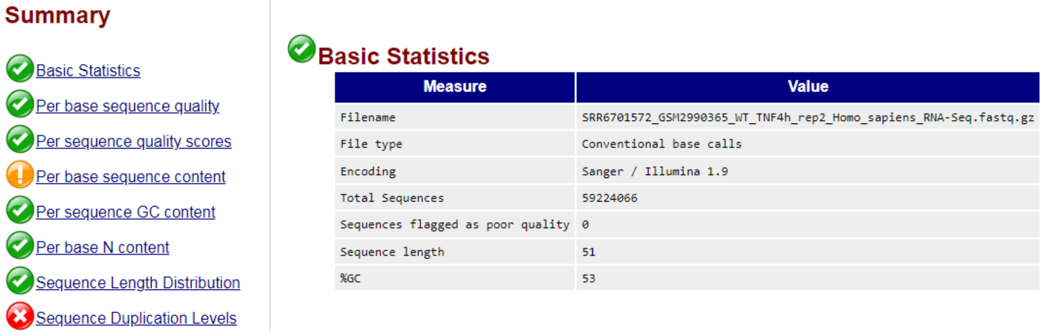


Filter results: Enter search term | All Fields ▾ | Add 0 to collection

| □ | Title | Accession | Instrument | Total Bases (Mb) | Date Created |
|---|---|---|---|---|---|
| □ | GSM2990358: WT ctr rep1; Homo sapiens; RNA-Seq | SRR6701565 | Illumina HiSeq 2500 | 48388 | 11 Feb 2018 |
| □ | GSM2990359: WT ctr rep2; Homo sapiens; RNA-Seq | SRR6701566 | Illumina HiSeq 2500 | 15688 | 11 Feb 2018 |
| □ | GSM2990360: WT ctr rep3; Homo sapiens; RNA-Seq | SRR6701567 | Illumina HiSeq 2500 | 26311 | 11 Feb 2018 |
| □ | GSM2990361: WT TNF30min rep1; Homo sapiens; RNA-Seq | SRR6701568 | Illumina HiSeq 2500 | 23926 | 11 Feb 2018 |
| □ | GSM2990362: WT TNF30min rep2; Homo sapiens; RNA-Seq | SRR6701569 | Illumina HiSeq 2500 | 22982 | 11 Feb 2018 |
| □ | GSM2990363: WT TNF30min rep3; Homo sapiens; RNA-Seq | SRR6701570 | Illumina HiSeq 2500 | 24703 | 11 Feb 2018 |
| □ | GSM2990364: WT TNF4h rep1; Homo sapiens; RNA-Seq | SRR6701571 | Illumina HiSeq 2500 | 15884 | 11 Feb 2018 |
| □ | GSM2990365: WT TNF4h rep2; Homo sapiens; RNA-Seq | SRR6701572 | Illumina HiSeq 2500 | 30204 | 11 Feb 2018 |
| □ | GSM2990366: WT TNF4h rep3; Homo sapiens; RNA-Seq | SRR6701573 | Illumina HiSeq 2500 | 16440 | 11 Feb 2018 |

*Figure 7 shows the table of all the samples used in our study out of these first three are controls and rest are treated Samples*

2) FASTQC: - After successfully completed the FastQC the software has generated two files for the user. Out of which we used html file to check the quality of all the samples. After analyzing these html files. It was concluded that the Quality of all the samples are more than the benchmark we set for various parameters such as Phred Score, GC content and various other parameters. All the samples have phred score more than 30 which is good enough to move further on our next analysis.



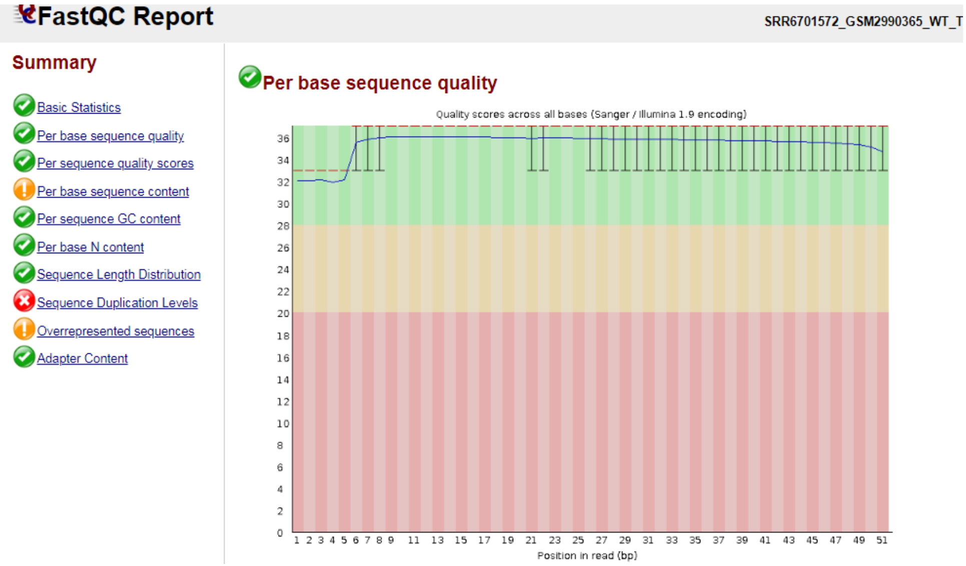
Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels

✓ Basic Statistics

| Measure | Value |
|---|---|
| Filename | SRR6701572_GSM2990365_WT_TNF4h_rep2_Homo_sapiens_RNA-Seq.fastq.gz |
| File type | Conventional base calls |
| Encoding | Sanger / Illumina 1.9 |
| Total Sequences | 59224066 |
| Sequences flagged as poor quality | 0 |
| Sequence length | 51 |
| %GC | 53 |

*Figure 8 shows the basics Stastics of Sample SRR06701572 after performing the FASTQC*



FastQC Report SRR6701572_GSM2990365_WT_T

Summary

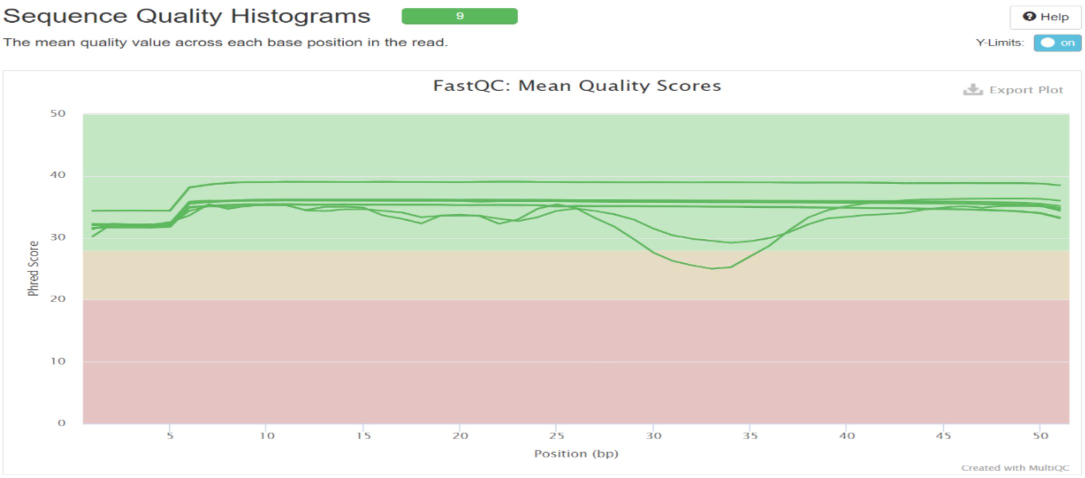
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ! Overrepresented sequences
- ✓ Adapter Content

✓ Per base sequence quality

*Figure 9 Shows the per base quality of Sample SRR067015722*

For every Sample html file contains lot of information out which we have shown in above figure what are the basic Statistics and Phred Score graph. We have these files for every sample and they do contain all these information.

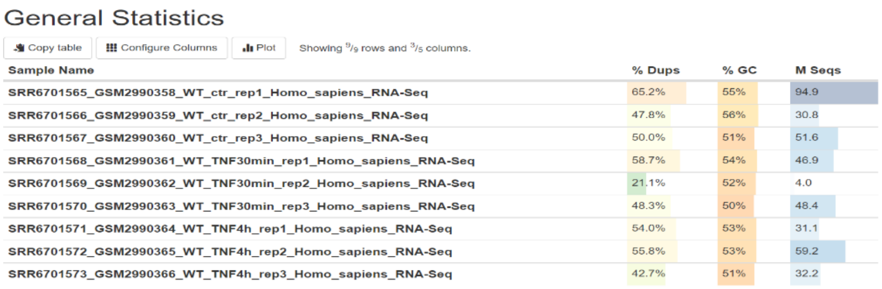
**MultiQC:-** It is similar to what we have done in FastQC but the difference is that the in FastQC results are generated for each and every sample separately but in case of MultiQC all the samples along with the controls was used together to check the quality and it is easy to compare the quality among the samples together. As mentioned in FastQC. The quality of all the samples is above the required range in different parameters. After checking the result of MultiQC same conclusion can be drawn here as well.



Sequence Quality Histograms

The mean quality value across each base position in the read.

*Figure 10 shows the mean quality score of the all sample used under study*



General Statistics

Showing $^{9}/_{9}$ rows and $^{3}/_{5}$ columns.

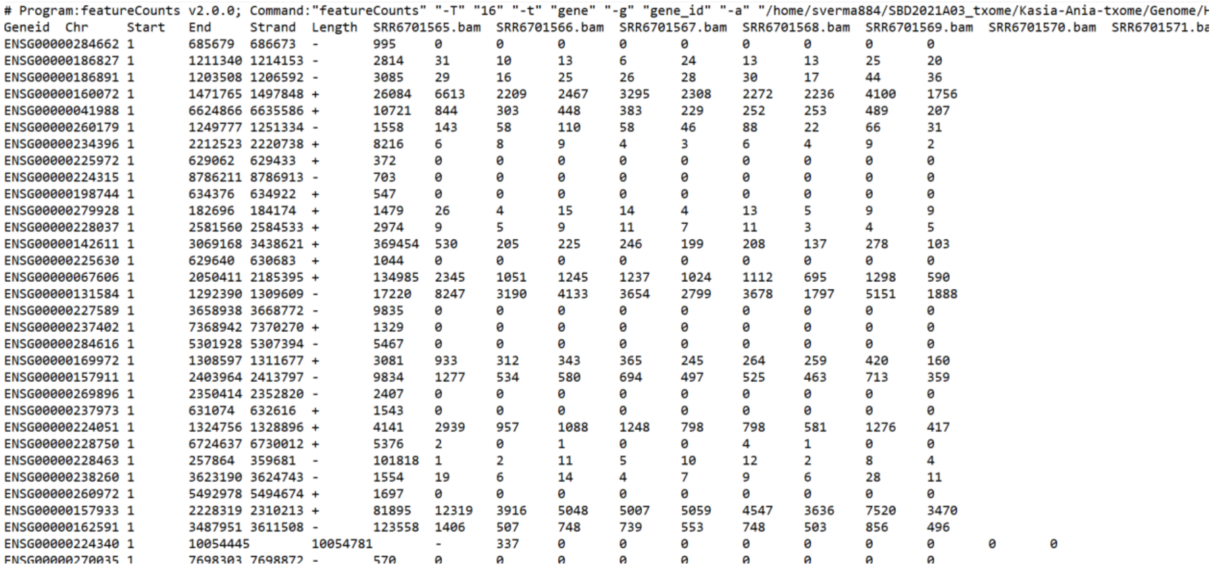
| Sample Name | % Dups | % GC | M Seqs |
|---|---|---|---|
| SRR6701565_GSM2990358_WT_ctr_rep1_Homo_sapiens_RNA-Seq | 65.2% | 55% | 94.9 |
| SRR6701566_GSM2990359_WT_ctr_rep2_Homo_sapiens_RNA-Seq | 47.8% | 56% | 30.8 |
| SRR6701567_GSM2990360_WT_ctr_rep3_Homo_sapiens_RNA-Seq | 50.0% | 51% | 51.6 |
| SRR6701568_GSM2990361_WT_TNF30min_rep1_Homo_sapiens_RNA-Seq | 58.7% | 54% | 46.9 |
| SRR6701569_GSM2990362_WT_TNF30min_rep2_Homo_sapiens_RNA-Seq | 21.1% | 52% | 4.0 |
| SRR6701570_GSM2990363_WT_TNF30min_rep3_Homo_sapiens_RNA-Seq | 48.3% | 50% | 48.4 |
| SRR6701571_GSM2990364_WT_TNF4h_rep1_Homo_sapiens_RNA-Seq | 54.0% | 53% | 31.1 |
| SRR6701572_GSM2990365_WT_TNF4h_rep2_Homo_sapiens_RNA-Seq | 55.8% | 53% | 59.2 |
| SRR6701573_GSM2990366_WT_TNF4h_rep3_Homo_sapiens_RNA-Seq | 42.7% | 51% | 32.2 |

*Figure 11 shows the general statistics of all the samples under study after MultiQC Step*

3) **Reference Genome:** It a very important to download the reference genome to know the position of reads on the human genome. Where exactly these sample dataset present on the human chromosome or more specifically at which genes. These step important in the sense that because without doing this step the whole study will go in vain. So it is this step which will going to help us to find which genes are going to affect by the treatment on the samples versus on the control.

4) **Genome Mapping Using Hisat2:** It is the step for which we have downloaded the above GTF, GFF and Index files. In this step we had mapped the reads on the reference genome for each sample. This step generate one summary file which is a text file and it has given the information that how much alignment is performed, how many reads were mapped single time, how many reads didn't map at all and other things. This step had also generated bam file for each sample and these bam files were used in the further analysis especially in the Feature matrix generation.

5) **Feature Extraction:** Using all the bam files generated above. It is possible to generate the feature matrix which contained the information regarding the genes affected in control and treated samples. Out of 60,000 features mostly are unaffected by the treatment but few hundred are over-expressed or expressed very low when compare to the control samples. This feature matrix is the input for the DESeq2 package of R-Studio and this package helps in the further analysis.



```
# Program:featureCounts v2.0.0; Command:"featureCounts" "-T" "16" "-t" "gene" "-g" "gene_id" "-a" "/home/sverma884/SBD2021A03_txome/Kasia-Ania-txome/Genome/H
Geneid  Chr     Start   End     Strand  Length  SRR6701565.bam  SRR6701566.bam  SRR6701567.bam  SRR6701568.bam  SRR6701569.bam  SRR6701570.bam  SRR6701571.ba
ENSG00000284662 1       685679  686673  -       995     0       0       0       0       0       0       0       0       0
ENSG00000186827 1       1211340 1214153 -       2814    31      10      13      6       24      13      13      25      20
ENSG00000186891 1       1203508 1206592 -       3085    29      16      25      26      28      30      17      44      36
ENSG00000160072 1       1471765 1497848 +       26084   6613    2209    2467    3295    2308    2272    2236    4100    1756
ENSG00000041988 1       6624866 6635586 +       10721   844     303     448     383     229     252     253     489     207
ENSG00000260179 1       1249777 1251334 -       1558    143     58      110     58      46      88      22      66      31
ENSG00000234396 1       2212523 2220738 +       8216    6       8       9       4       3       6       4       9       2
ENSG00000225972 1       629062  629433  +       372     0       0       0       0       0       0       0       0       0
ENSG00000224315 1       8786211 8786913 -       703     0       0       0       0       0       0       0       0       0
ENSG00000198744 1       634376  634922  +       547     0       0       0       0       0       0       0       0       0
ENSG00000279928 1       182696  184174  +       1479    26      4       15      14      4       13      5       9       9
ENSG00000228037 1       2581560 2584533 +       2974    9       5       9       11      7       11      3       4       5
ENSG00000142611 1       3069168 3438621 +       369454  530     205     225     246     199     208     137     278     103
ENSG00000225630 1       629640  630683  +       1044    0       0       0       0       0       0       0       0       0
ENSG00000067606 1       2050411 2185395 +       134985  2345    1051    1245    1237    1024    1112    695     1298    590
ENSG00000131584 1       1292390 1309609 -       17220   8247    3190    4133    3654    2799    3678    1797    5151    1888
ENSG00000227589 1       3658938 3668772 -       9835    0       0       0       0       0       0       0       0       0
ENSG00000237402 1       7368942 7370270 +       1329    0       0       0       0       0       0       0       0       0
ENSG00000284616 1       5301928 5307394 -       5467    0       0       0       0       0       0       0       0       0
ENSG00000169972 1       1308597 1311677 +       3081    933     312     343     365     245     264     259     420     160
ENSG00000157911 1       2403964 2413797 -       9834    1277    534     580     694     497     525     463     713     359
ENSG00000269896 1       2350414 2352820 -       2407    0       0       0       0       0       0       0       0       0
ENSG00000237973 1       631074  632616  +       1543    0       0       0       0       0       0       0       0       0
ENSG00000224051 1       1324756 1328896 +       4141    2939    957     1088    1248    798     798     581     1276    417
ENSG00000228750 1       6724637 6730012 +       5376    2       0       1       0       0       4       1       0       0
ENSG00000228463 1       257864  359681  -       101818  1       2       11      5       10      12      2       8       4
ENSG00000238260 1       3623190 3624743 -       1554    19      6       14      4       7       9       6       28      11
ENSG00000260972 1       5492978 5494674 +       1697    0       0       0       0       0       0       0       0       0
ENSG00000157933 1       2228319 2310213 +       81895   12319   3916    5048    5007    5059    4547    3636    7520    3470
ENSG00000162591 1       3487951 3611508 -       123558  1406    507     748     739     553     748     503     856     496
ENSG00000224340 1       10054445        10054781        -       337     0       0       0       0       0       0       0       0       0
ENSG00000270035 1       7698303 7698872 -       570     0       0       0       0       0       0       0       0       0
```

*Figure 12 shows the image of feature count matrix generated for all the samples under study*

6) **Differential Gene Expression Analysis Using DESeq2:** This step is very crucial step for the analysis of differential gene expression because this step involves normalization and statistical analysis because we obtained around 60,000 features in the last step. So, to deal with such a huge amount of data was not easy so we performed this step. After this step we got the list of genes which are differentially expressed both in the case of 30 min TNF treatment and 4 hours of TNF treatment. We had set our own parameters to get the list of genes which are upregulated and which are downregulated and I will mention the parameters while mentioning the list of those genes but before that we have obtained many plots both in the case of 30 minutes of TNF treatment and in case of 4 hours of TNF treatment, we are going to interpret these results because these help in the normalization as well in the statistical analysis.

- **Principal Component Analysis (PCA):** It is a method used in the quality control and it was already discussed above with a example here we discuss our plot that we got after the DESeq2 analysis.

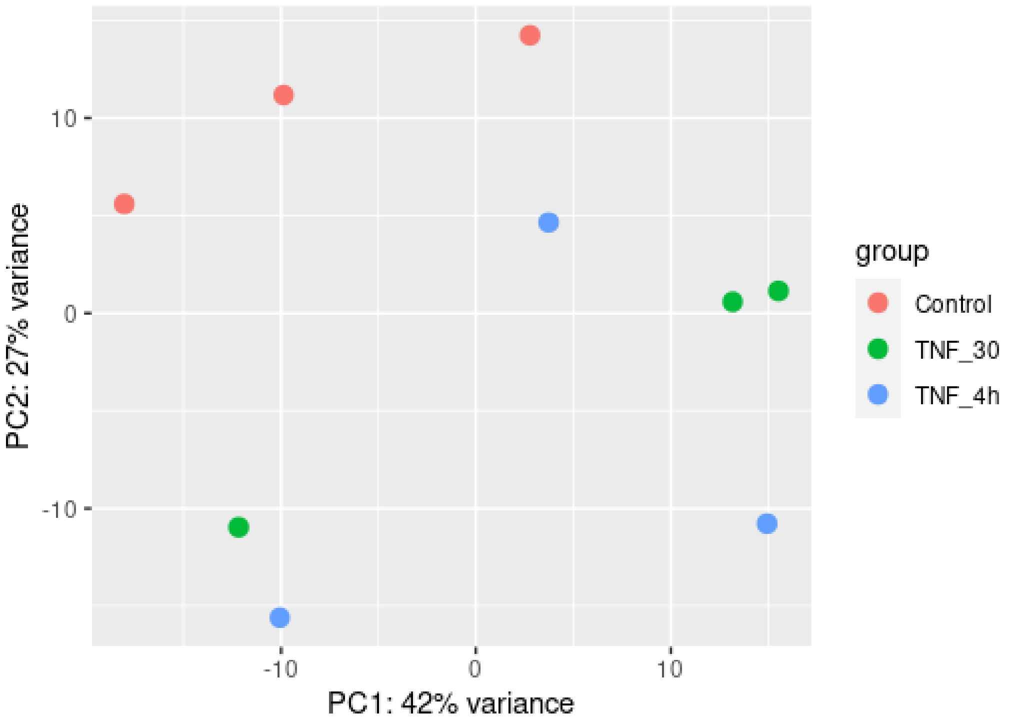


*Figure 13 shows the PCA plot of control vs treated samples.*

Here we can observe that our control group (orange) is completely clustered away from the samples (green and blues) which represent that the quality of our data is good. This plot also showing that our dataset doesn't have any outliers. These plots are very good and we there is no concern in this plot. We have also plotted graph for both the samples individually which is present below.

The plot representing in figure-14 is a very good representation of control vs treated samples and they separated very well in the plot. Also, this is representing that our data doesn't have any outlier. So, this PC is showing that the quality of our data is good. Also, if see the values of variance they are also very good in this plot. Likewise, we have also created plot for treated sample for 4 hours.

In case of the control vs treated 4 hours it has also somewhat similar to the previous plot, but if we observe they are similar samples just differing in the TNF treatment time. Also, these samples are clustered very well i.e., control samples are clustering together and treated samples are clustering together. So we shouldn't worry about exploring further PCs because these are already giving us very good results. Theory and process of these plots are already mentioned above.

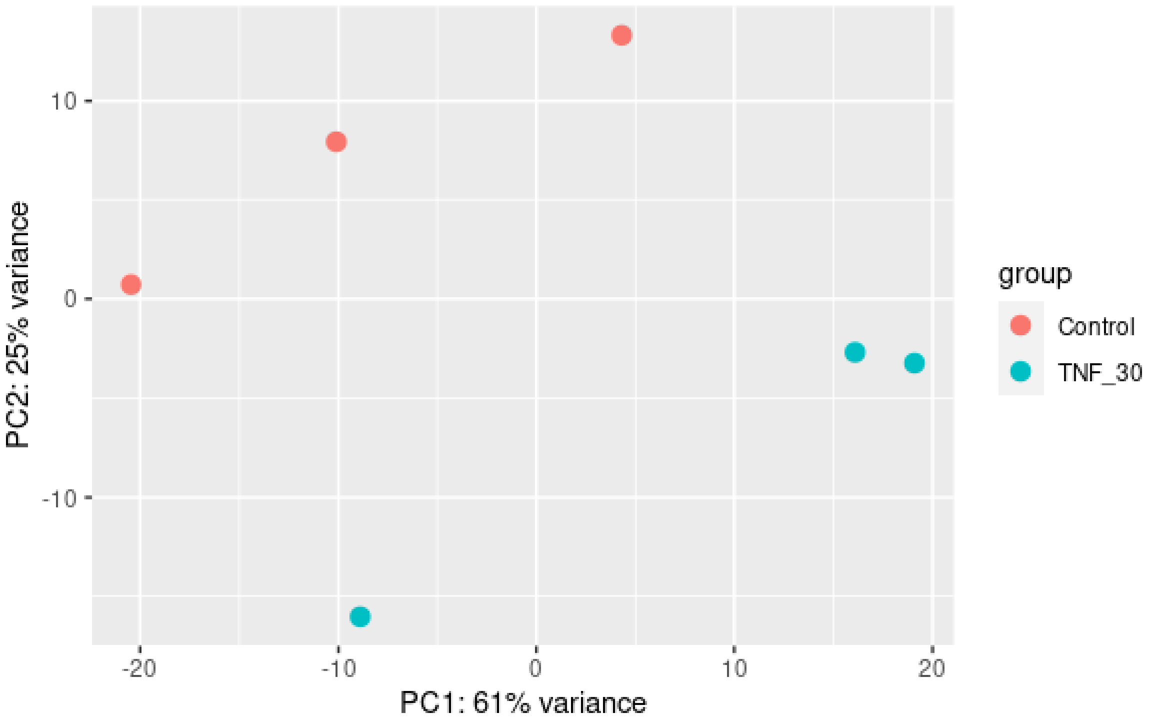


*Figure 14 shows the PCA plot of the control vs treated (30 mins).*

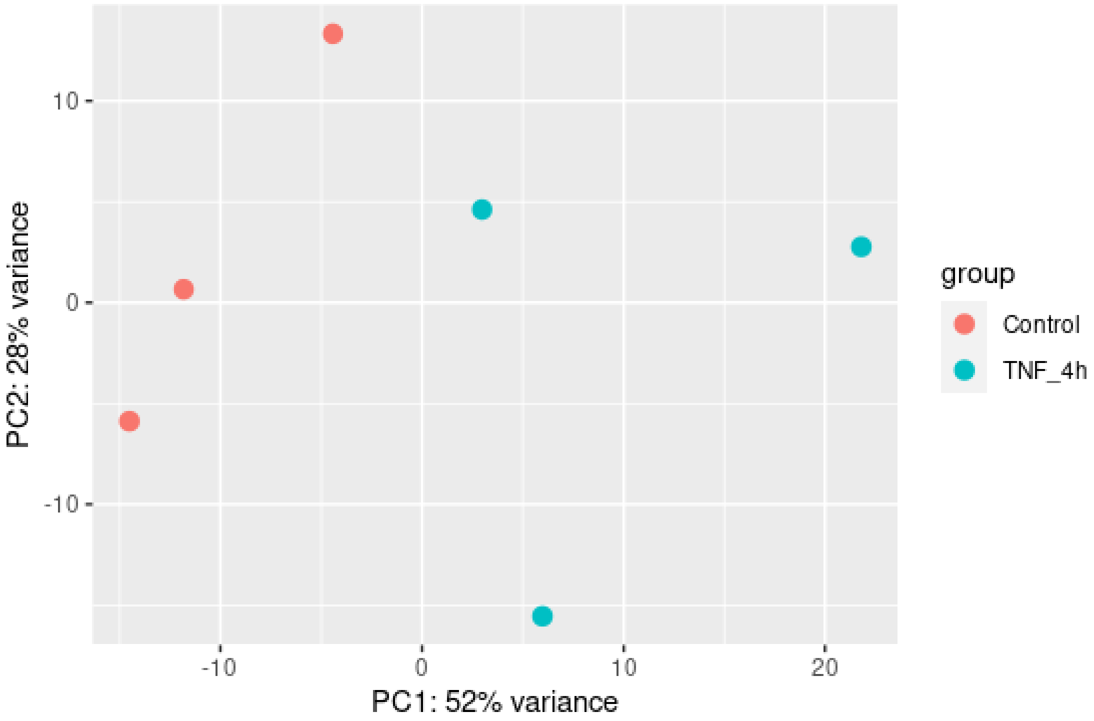


*Figure 15 showing the PCA plot of control vs 4 hours.*

- **Hierarchical Clustering:** It is similar to PCA and have the same purpose to find the patterns and outliers in the sample. It basically displaying correlation of gene expression for all pairwise combination of samples in dataset. It's a well known to us that all genes were not differentially expressed that's why samples show high correlation with each other. We have taken the value 0.8 as cut-off i.e., the values greater than 0.8 are not outliers while values below 0.8 are outliers. If we look these graphs, we can observe that controls are clustering together while treated samples are clustering together and that's what the purpose of generating heatmap because we want to observe which samples are similar and this similarity is based on the normalized gene expression value. Since both the treated samples are clustering together and differing from controls, this is again showing that quality control is very good in our dataset.

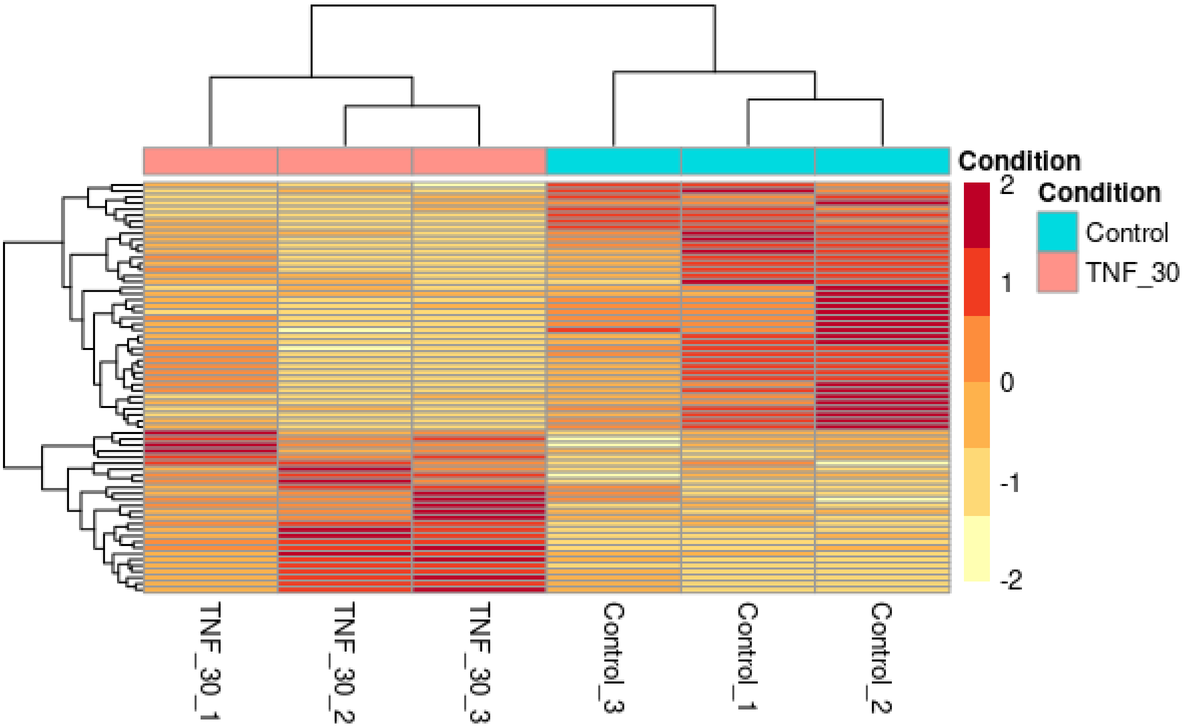


*Figure 16 showing the heatmap of control vs 30 minutes.*

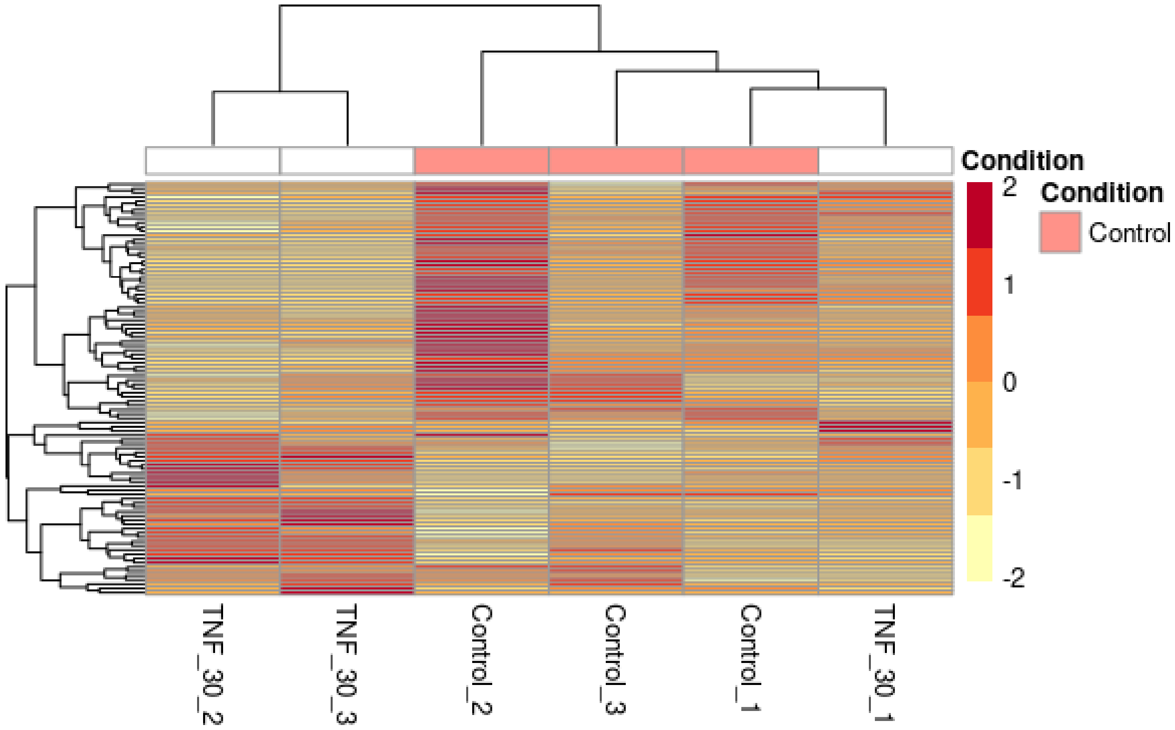


*Figure 17 showing the heatmap control vs treated 4 hours*

- **Dispersion Estimates: -** We have already discussion what dispersion represent and how it was calculated by the DESeq2 method. We also discussed that dispersion is inversely proportional to mean and directly proportional to variance. So basically, genes having same values for mean can vary only when their variance value differ.

  Let us summarize the graph present below we can see that there were three colour coding used here 1) is black which shows the gene dispersion estimates. 2) is blue these were the shrink values of gene dispersion estimates and also called the final value. 3) is the red colour, this basically represent the best fitted line in this data set. Black dot represents each gene.

  Both the plots represented below are very good plots and showing that our data is very good fit for the DESeq2 model. If we look carefully in our plots that the genes are scattered in the whole plot which is also ensuring that our data has no contamination. Because in case of contamination we can observe that the data is clustered at one point only. And process of shrinkage also helps us lot in getting the better dispersion estimates. We had already discussed the process of shrinkage in the previous articles.

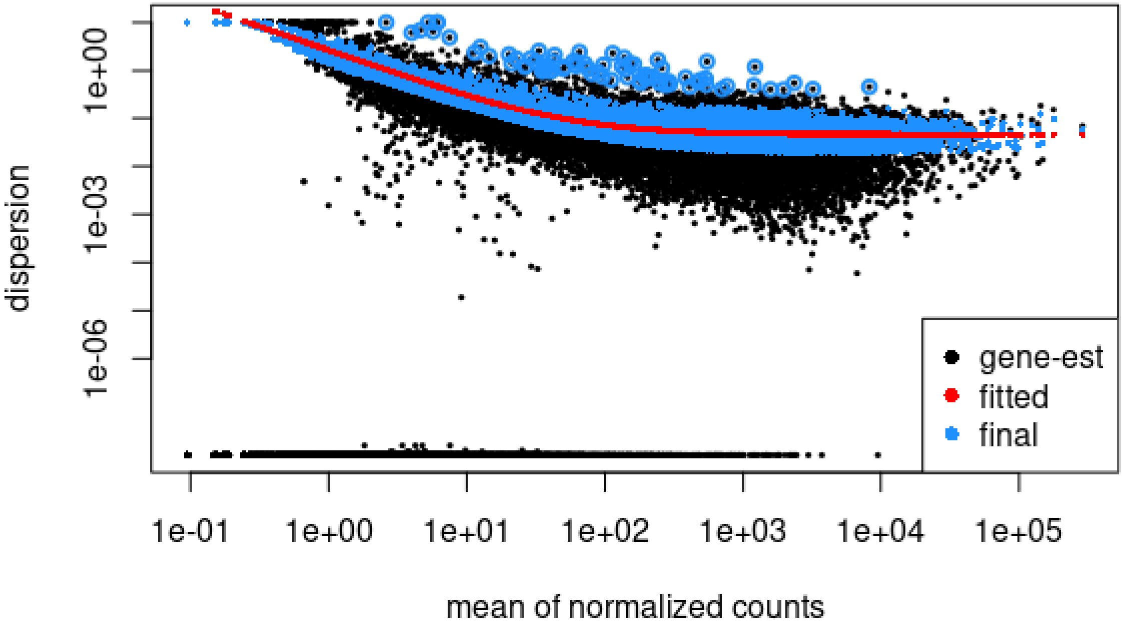


*Figure 18 showing the dispersion curve for samples treated with TNF after 30 minutes*

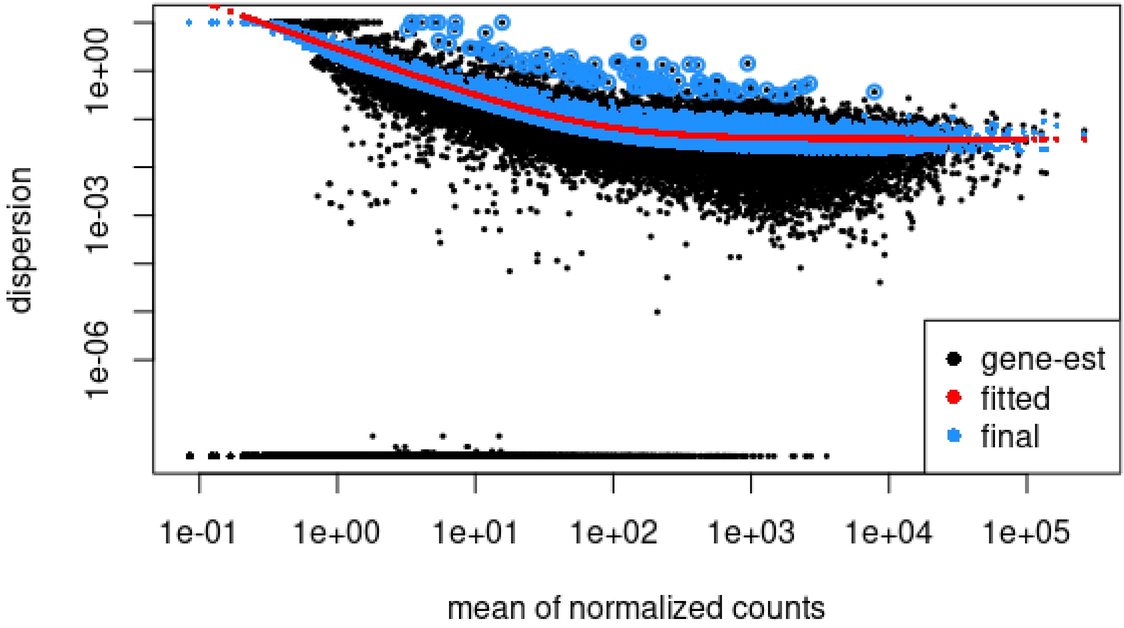
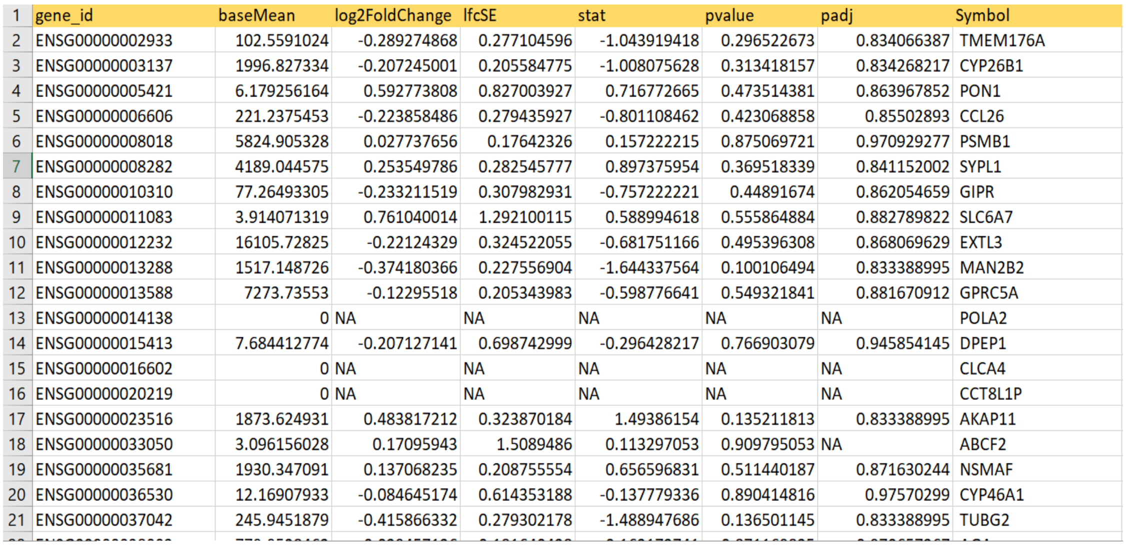


*Figure 19 showing the dispersion curve for samples treated with TNF after 4 hours minutes*

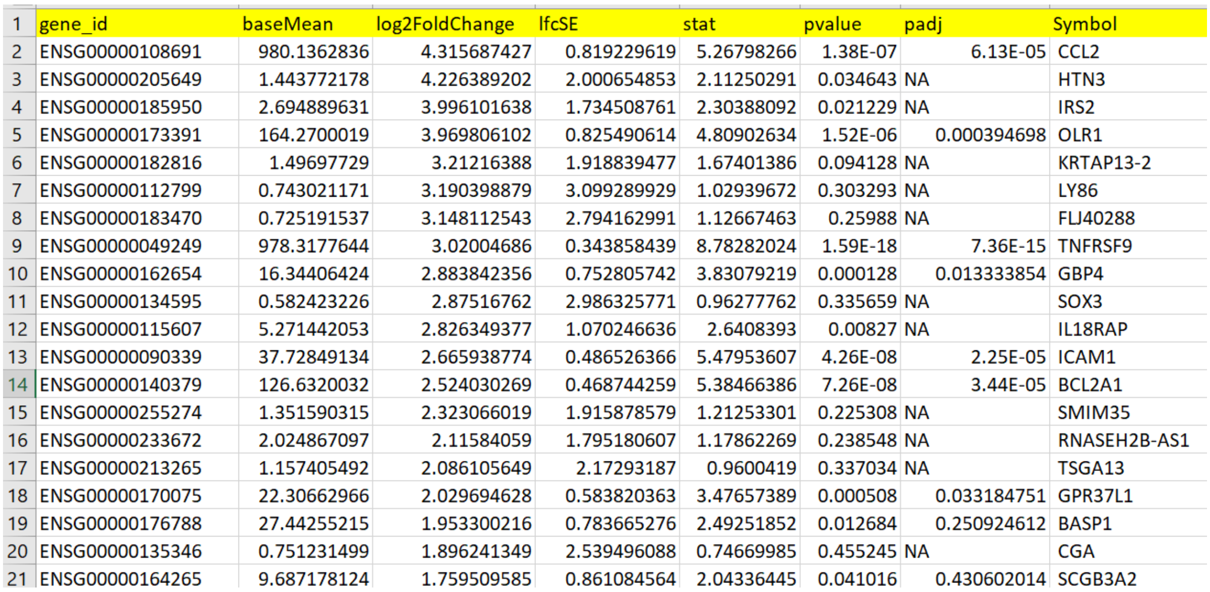
- **Expression Matrix:**

After all these steps of DESeq2 we obtained the normalized expression matrix for both the treated samples and this matrix contain more than 1000 rows which is good number but we are not interested in all these rows because we have to sort out these lists based on the p-value and log2Foldchange values which we will do in next step. This expression matrix is most desirable output of output of study and its basically the foundation on which we will able to identify our up-regulated and down-regulated genes.



| 1 | gene_id | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | Symbol |
|---|---|---|---|---|---|---|---|---|
| 2 | ENSG00000002933 | 102.5591024 | -0.289274868 | 0.277104596 | -1.043919418 | 0.296522673 | 0.834066387 | TMEM176A |
| 3 | ENSG00000003137 | 1996.827334 | -0.207245001 | 0.205584775 | -1.008075628 | 0.313418157 | 0.834268217 | CYP26B1 |
| 4 | ENSG00000005421 | 6.179256164 | 0.592773808 | 0.827003927 | 0.716772665 | 0.473514381 | 0.863967852 | PON1 |
| 5 | ENSG00000006606 | 221.2375453 | -0.223858486 | 0.279435927 | -0.801108462 | 0.423068858 | 0.85502893 | CCL26 |
| 6 | ENSG00000008018 | 5824.905328 | 0.027737656 | 0.17642326 | 0.157222215 | 0.875069721 | 0.970929277 | PSMB1 |
| 7 | ENSG00000008282 | 4189.044575 | 0.253549786 | 0.282545777 | 0.897375954 | 0.369518339 | 0.841152002 | SYPL1 |
| 8 | ENSG00000010310 | 77.26493305 | -0.233211519 | 0.307982931 | -0.757222221 | 0.44891674 | 0.862054659 | GIPR |
| 9 | ENSG00000011083 | 3.914071319 | 0.761040014 | 1.292100115 | 0.588994618 | 0.555864884 | 0.882789822 | SLC6A7 |
| 10 | ENSG00000012232 | 16105.72825 | -0.22124329 | 0.324522055 | -0.681751166 | 0.495396308 | 0.868069629 | EXTL3 |
| 11 | ENSG00000013288 | 1517.148726 | -0.374180366 | 0.227556904 | -1.644337564 | 0.100106494 | 0.833388995 | MAN2B2 |
| 12 | ENSG00000013588 | 7273.73553 | -0.12295518 | 0.205343983 | -0.598776641 | 0.549321841 | 0.881670912 | GPRC5A |
| 13 | ENSG00000014138 | 0 | NA | NA | NA | NA | NA | POLA2 |
| 14 | ENSG00000015413 | 7.684412774 | -0.207127141 | 0.698742999 | -0.296428217 | 0.766903079 | 0.945854145 | DPEP1 |
| 15 | ENSG00000016602 | 0 | NA | NA | NA | NA | NA | CLCA4 |
| 16 | ENSG00000020219 | 0 | NA | NA | NA | NA | NA | CCT8L1P |
| 17 | ENSG00000023516 | 1873.624931 | 0.483817212 | 0.323870184 | 1.49386154 | 0.135211813 | 0.833388995 | AKAP11 |
| 18 | ENSG00000033050 | 3.096156028 | 0.17095943 | 1.5089486 | 0.113297053 | 0.909795053 | NA | ABCF2 |
| 19 | ENSG00000035681 | 1930.347091 | 0.137068235 | 0.208755554 | 0.656596831 | 0.511440187 | 0.871630244 | NSMAF |
| 20 | ENSG00000036530 | 12.16907933 | -0.084645174 | 0.614353188 | -0.137779336 | 0.890414816 | 0.97570299 | CYP46A1 |
| 21 | ENSG00000037042 | 245.9451879 | -0.415866332 | 0.279302178 | -1.488947686 | 0.136501145 | 0.833388995 | TUBG2 |

*Figure 20 showing the screenshot of excel sheet which contain the information about the expression matrix of treated sample (30 min)*

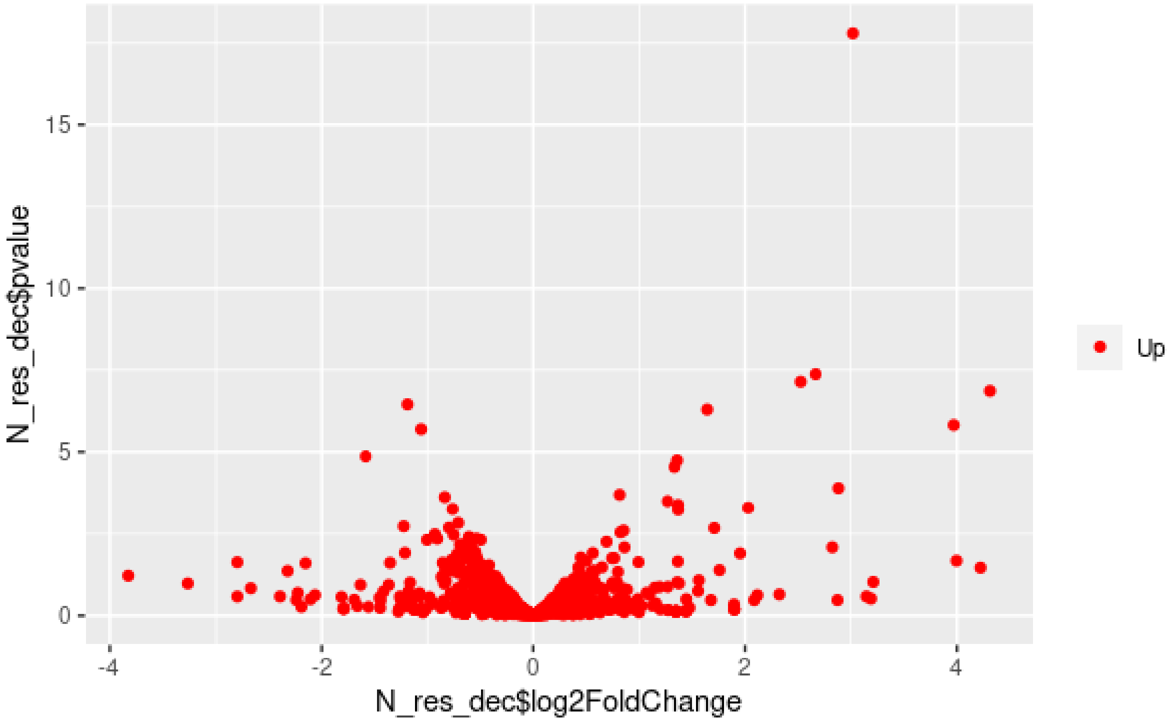


| 1 | gene_id | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | Symbol |
|---|---|---|---|---|---|---|---|---|
| 2 | ENSG00000108691 | 980.1362836 | 4.315687427 | 0.819229619 | 5.26798266 | 1.38E-07 | 6.13E-05 | CCL2 |
| 3 | ENSG00000205649 | 1.443772178 | 4.226389202 | 2.000654853 | 2.11250291 | 0.034643 | NA | HTN3 |
| 4 | ENSG00000185950 | 2.694889631 | 3.996101638 | 1.734508761 | 2.30388092 | 0.021229 | NA | IRS2 |
| 5 | ENSG00000173391 | 164.2700019 | 3.969806102 | 0.825490614 | 4.80902634 | 1.52E-06 | 0.000394698 | OLR1 |
| 6 | ENSG00000182816 | 1.49697729 | 3.21216388 | 1.918839477 | 1.67401386 | 0.094128 | NA | KRTAP13-2 |
| 7 | ENSG00000112799 | 0.743021171 | 3.190398879 | 3.099289929 | 1.02939672 | 0.303293 | NA | LY86 |
| 8 | ENSG00000183470 | 0.725191537 | 3.148112543 | 2.794162991 | 1.12667463 | 0.25988 | NA | FLJ40288 |
| 9 | ENSG00000049249 | 978.3177644 | 3.02004686 | 0.343858439 | 8.78282024 | 1.59E-18 | 7.36E-15 | TNFRSF9 |
| 10 | ENSG00000162654 | 16.34406424 | 2.883842356 | 0.752805742 | 3.83079219 | 0.000128 | 0.013333854 | GBP4 |
| 11 | ENSG00000134595 | 0.582423226 | 2.87516762 | 2.986325771 | 0.96277762 | 0.335659 | NA | SOX3 |
| 12 | ENSG00000115607 | 5.271442053 | 2.826349377 | 1.070246636 | 2.6408393 | 0.00827 | NA | IL18RAP |
| 13 | ENSG00000090339 | 37.72849134 | 2.665938774 | 0.486526366 | 5.47953607 | 4.26E-08 | 2.25E-05 | ICAM1 |
| 14 | ENSG00000140379 | 126.6320032 | 2.524030269 | 0.468744259 | 5.38466386 | 7.26E-08 | 3.44E-05 | BCL2A1 |
| 15 | ENSG00000255274 | 1.351590315 | 2.323066019 | 1.915878579 | 1.21253301 | 0.225308 | NA | SMIM35 |
| 16 | ENSG00000233672 | 2.024867097 | 2.11584059 | 1.795180607 | 1.17862269 | 0.238548 | NA | RNASEH2B-AS1 |
| 17 | ENSG00000213265 | 1.157405492 | 2.086105649 | 2.17293187 | 0.9600419 | 0.337034 | NA | TSGA13 |
| 18 | ENSG00000170075 | 22.30662966 | 2.029694628 | 0.583820363 | 3.47657389 | 0.000508 | 0.033184751 | GPR37L1 |
| 19 | ENSG00000176788 | 27.44255215 | 1.953300216 | 0.783665276 | 2.49251852 | 0.012684 | 0.250924612 | BASP1 |
| 20 | ENSG00000135346 | 0.751231499 | 1.896241349 | 2.539496088 | 0.74669985 | 0.455245 | NA | CGA |
| 21 | ENSG00000164265 | 9.687178124 | 1.759509585 | 0.861084564 | 2.04336445 | 0.041016 | 0.430602014 | SCGB3A2 |

*Figure 21 showing the screenshot of excel sheet which contain the information about the expression matrix of treated sample (4 hours)*

- **Volcano Plot:**

  This plot between p-value (y-axis) and log2FoldChange (x-axis) is known as volcano plot because of its shape. This we plotted by using the values we obtained in the expression matrix. Now if we carefully observe most of the data is clustered between 1 and -1 and this information is very important and it will used while deciding the parameters for up-regulated and down-regulated genes.



- **Counting of Differential expressed genes:**

  After all these steps we obtained a csv file (expression matrix) containing the list of genes with many parameters like pvalue, padj, log2Foldchange and etc. But we are interested in some of these parameters. So, this file contains the list which have undergo even the slightest change but we don’t want all these genes. To obtain the genes which are downregulated and upregulated we had use pvalue should be less than 0.5 and log2Foldchange value should be greater than 0.7 for the upregulation and less than 0.7 for the downregulation and we get some genes which is present in the form of Venn diagram. So, this Venn diagram showing the number of genes which are upregulated and downregulated after 30 mins of TNF treatment vs 4 hour of TNF treatment. It also the shows the number of genes which were common in both the cases. So, there are 49 genes which remain affected in both 30 minutes and 4 hours and there are 115 genes which were differentially expressed in 30 minutes and there were 137 genes which were differentially expressed in case of 4 hours after the TNF treatment.

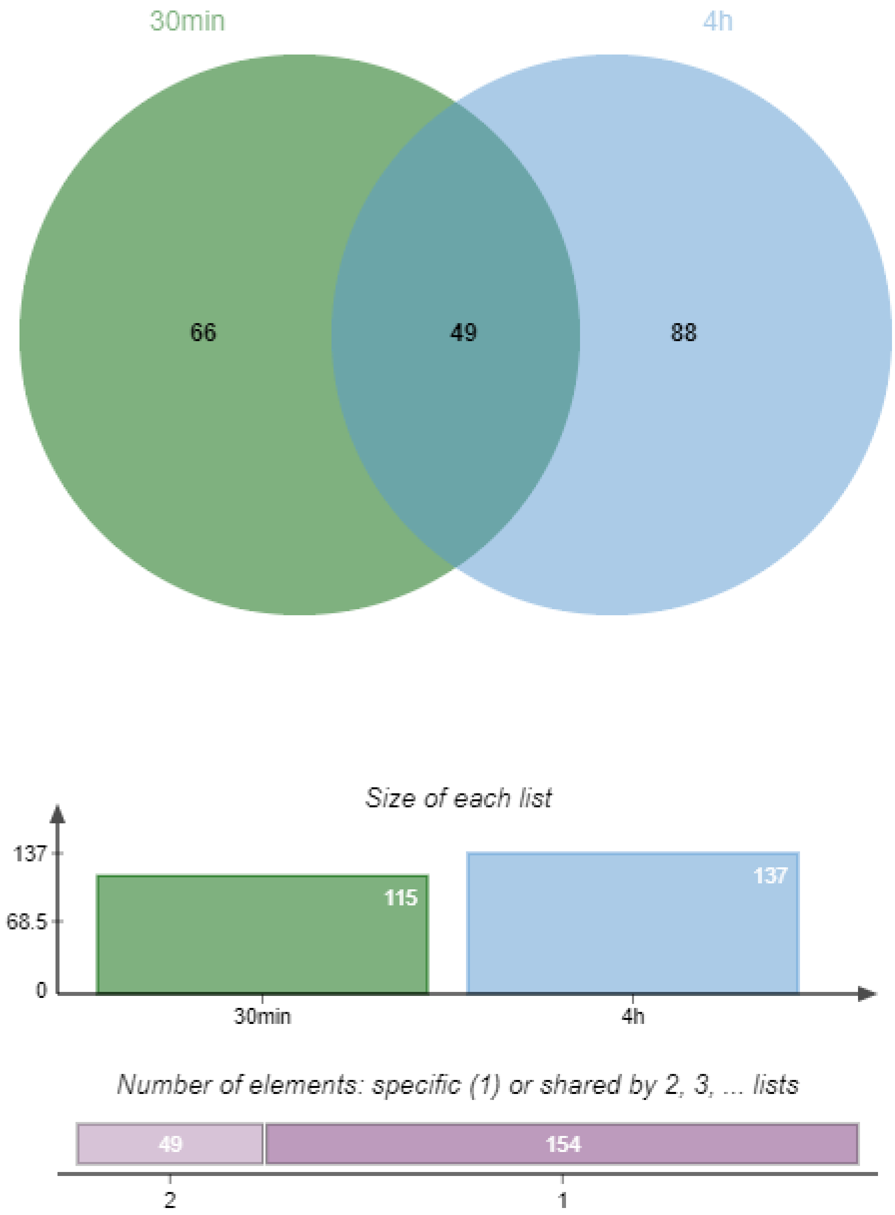


*Figure 22 shows the Venn diagram displaying number of genes differentially expressed after 30 minutes of TNF treatment vs 4 hours of TNF treatment*

Apart from this we have also created the Venn diagram of number of genes upregulated in 30 minutes vs in 4 hours and same we did for downregulated genes. In Venn diagram of downregulated genes, we observed that 15 genes are downregulated in both the experimental condition while 30 minutes has only 38 genes which was downregulated and in case of 4 hour 68 genes are downregulated.

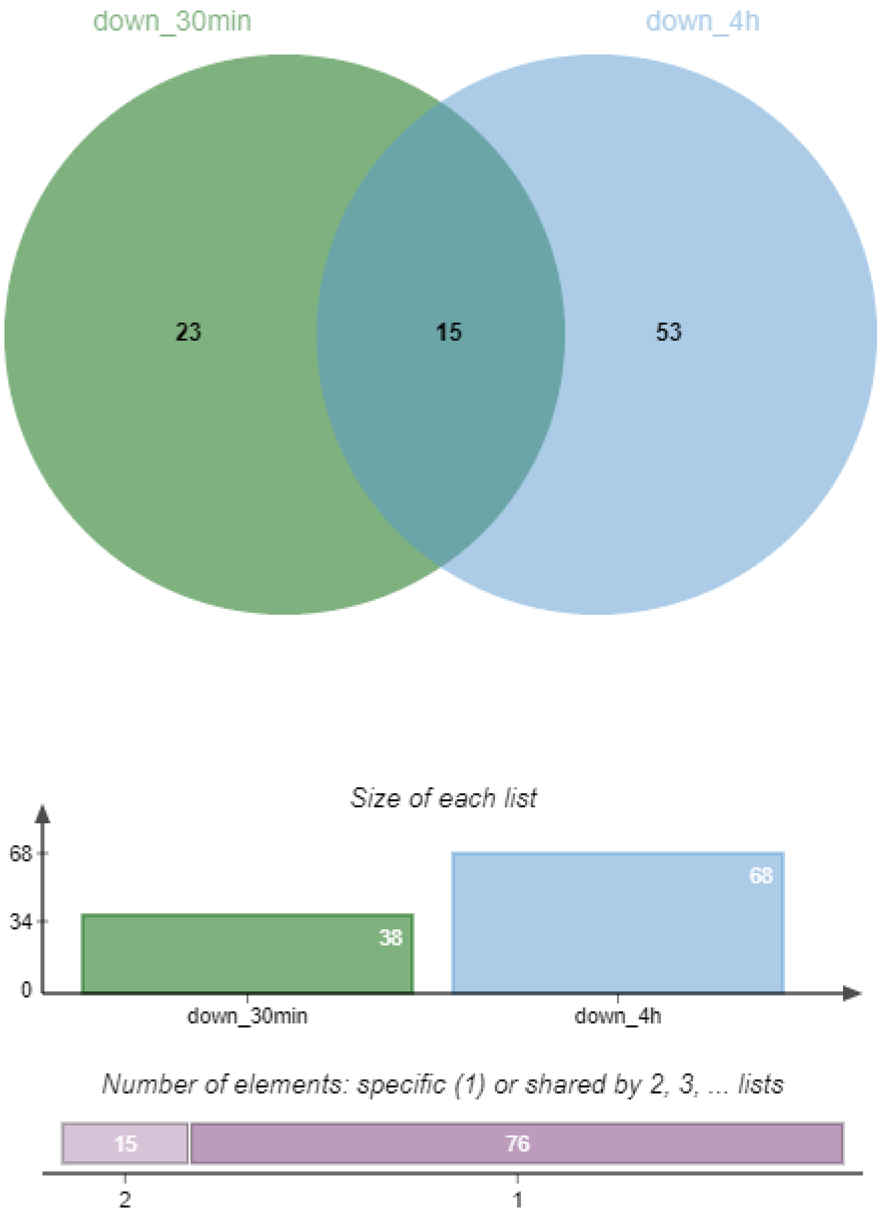


*Figure 23 shows the Venn diagram displaying number of genes down-regulated after 30 minutes of TNF treatment vs 4 hours of TNF treatment*

In Venn diagram of up-regulated genes, we observed that 31 genes were downregulated in both the experimental condition while 30 minutes experiment has only 77 genes which was up-regulated and in case of 4 hour 69 genes are downregulated.

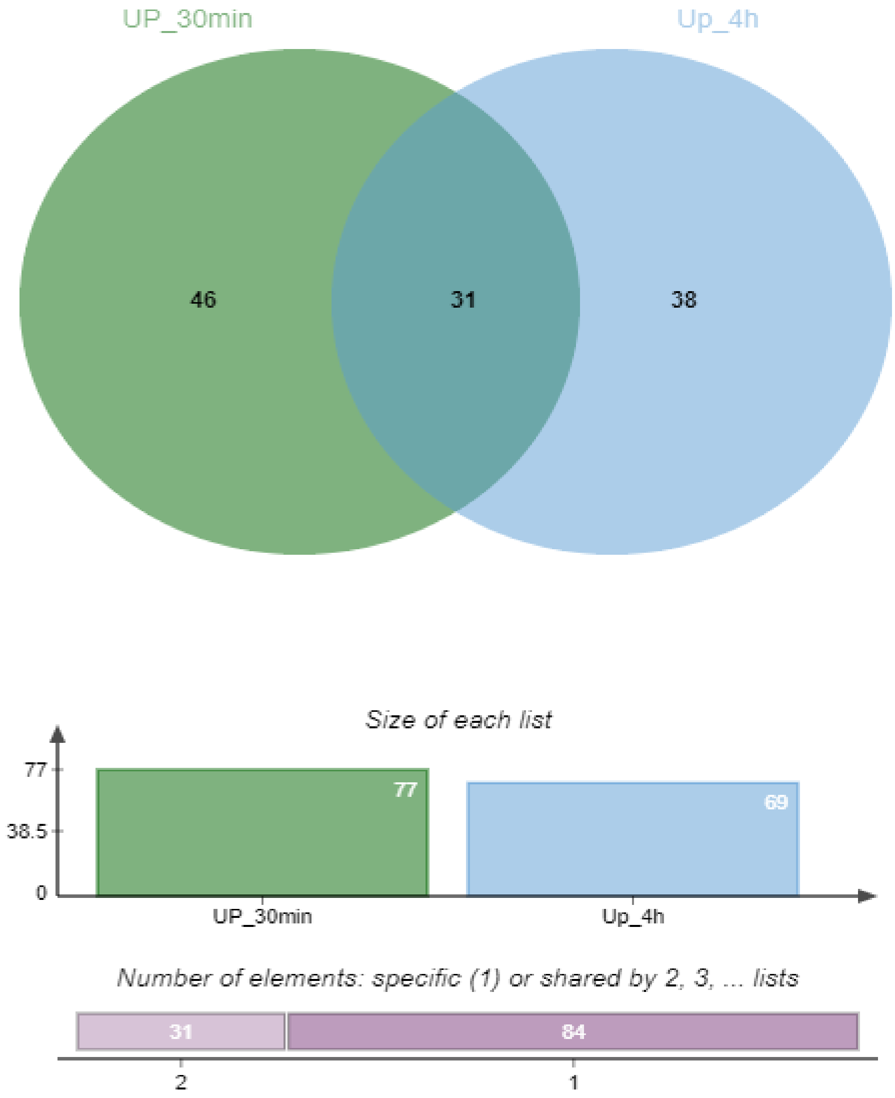
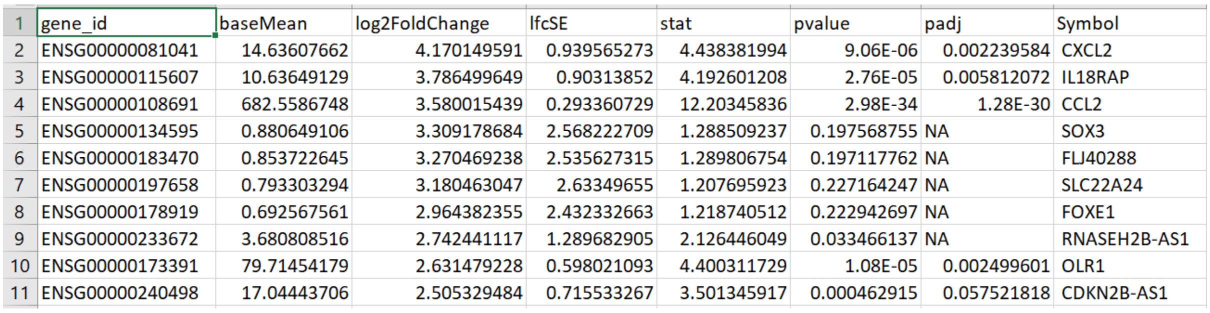


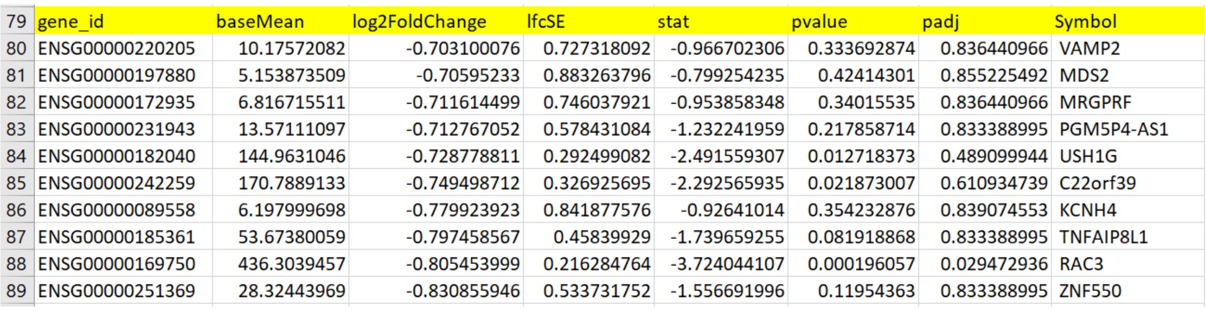
*Figure 24 shows the Venn diagram displaying number of genes up-regulated after 30 minutes of TNF treatment vs 4 hours of TNF treatment*

- **Differentially Expressed Genes:** The ultimate aim of our study is to find the differentially expressed genes and we have taken two treated samples. So here we are showing the name of the top 10 upregulated and downregulated genes of both the samples. We have also found two genes which were upregulated during 30 minutes of TNF treatment but down-regulated after 4 hours of TNF treatment.



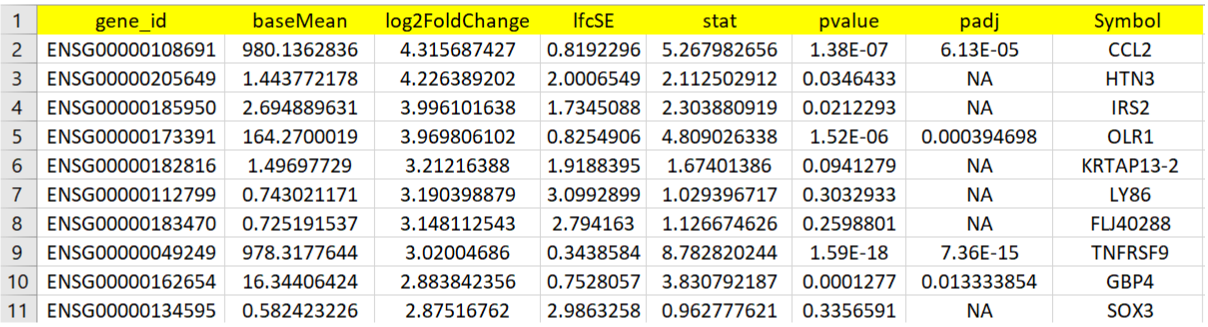
| 1 | gene_id | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | Symbol |
|---|---|---|---|---|---|---|---|---|
| 2 | ENSG00000081041 | 14.63607662 | 4.170149591 | 0.939565273 | 4.438381994 | 9.06E-06 | 0.002239584 | CXCL2 |
| 3 | ENSG00000115607 | 10.63649129 | 3.786499649 | 0.90313852 | 4.192601208 | 2.76E-05 | 0.005812072 | IL18RAP |
| 4 | ENSG00000108691 | 682.5586748 | 3.580015439 | 0.293360729 | 12.20345836 | 2.98E-34 | 1.28E-30 | CCL2 |
| 5 | ENSG00000134595 | 0.880649106 | 3.309178684 | 2.568222709 | 1.288509237 | 0.197568755 | NA | SOX3 |
| 6 | ENSG00000183470 | 0.853722645 | 3.270469238 | 2.535627315 | 1.289806754 | 0.197117762 | NA | FLJ40288 |
| 7 | ENSG00000197658 | 0.793303294 | 3.180463047 | 2.63349655 | 1.207695923 | 0.227164247 | NA | SLC22A24 |
| 8 | ENSG00000178919 | 0.692567561 | 2.964382355 | 2.432332663 | 1.218740512 | 0.222942697 | NA | FOXE1 |
| 9 | ENSG00000233672 | 3.680808516 | 2.742441117 | 1.289682905 | 2.126446049 | 0.033466137 | NA | RNASEH2B-AS1 |
| 10 | ENSG00000173391 | 79.71454179 | 2.631479228 | 0.598021093 | 4.400311729 | 1.08E-05 | 0.002499601 | OLR1 |
| 11 | ENSG00000240498 | 17.04443706 | 2.505329484 | 0.715533267 | 3.501345917 | 0.000462915 | 0.057521818 | CDKN2B-AS1 |

*Figure 25 showing the list of the top-10 genes which are up-regulated after 30 minutes of TNF treatment*



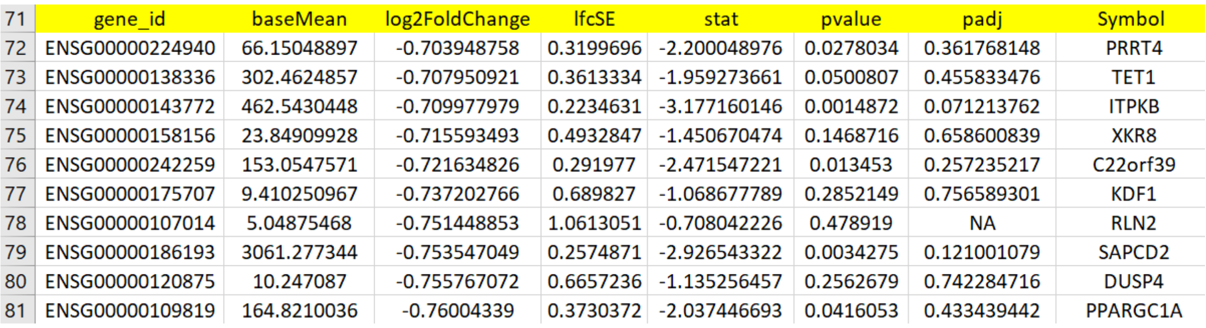
| 79 | gene_id | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | Symbol |
|---|---|---|---|---|---|---|---|---|
| 80 | ENSG00000220205 | 10.17572082 | -0.703100076 | 0.727318092 | -0.966702306 | 0.333692874 | 0.836440966 | VAMP2 |
| 81 | ENSG00000197880 | 5.153873509 | -0.70595233 | 0.883263796 | -0.799254235 | 0.42414301 | 0.855225492 | MDS2 |
| 82 | ENSG00000172935 | 6.816715511 | -0.711614499 | 0.746037921 | -0.953858348 | 0.34015535 | 0.836440966 | MRGPRF |
| 83 | ENSG00000231943 | 13.57111097 | -0.712767052 | 0.578431084 | -1.232241959 | 0.217858714 | 0.833388995 | PGM5P4-AS1 |
| 84 | ENSG00000182040 | 144.9631046 | -0.728778811 | 0.292499082 | -2.491559307 | 0.012718373 | 0.489099944 | USH1G |
| 85 | ENSG00000242259 | 170.7889133 | -0.749498712 | 0.326925695 | -2.292565935 | 0.021873007 | 0.610934739 | C22orf39 |
| 86 | ENSG00000089558 | 6.197999698 | -0.779923923 | 0.841877576 | -0.92641014 | 0.354232876 | 0.839074553 | KCNH4 |
| 87 | ENSG00000185361 | 53.67380059 | -0.797458567 | 0.45839929 | -1.739659255 | 0.081918868 | 0.833388995 | TNFAIP8L1 |
| 88 | ENSG00000169750 | 436.3039457 | -0.805453999 | 0.216284764 | -3.724044107 | 0.000196057 | 0.029472936 | RAC3 |
| 89 | ENSG00000251369 | 28.32443969 | -0.830855946 | 0.533731752 | -1.556691996 | 0.11954363 | 0.833388995 | ZNF550 |

*Figure 26 showing the list of the top-10 genes which are down-regulated after 30 minutes of TNF treatment*



| 1 | gene_id | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | Symbol |
|---|---|---|---|---|---|---|---|---|
| 2 | ENSG00000108691 | 980.1362836 | 4.315687427 | 0.8192296 | 5.267982656 | 1.38E-07 | 6.13E-05 | CCL2 |
| 3 | ENSG00000205649 | 1.443772178 | 4.226389202 | 2.0006549 | 2.112502912 | 0.0346433 | NA | HTN3 |
| 4 | ENSG00000185950 | 2.694889631 | 3.996101638 | 1.7345088 | 2.303880919 | 0.0212293 | NA | IRS2 |
| 5 | ENSG00000173391 | 164.2700019 | 3.969806102 | 0.8254906 | 4.809026338 | 1.52E-06 | 0.000394698 | OLR1 |
| 6 | ENSG00000182816 | 1.49697729 | 3.21216388 | 1.9188395 | 1.67401386 | 0.0941279 | NA | KRTAP13-2 |
| 7 | ENSG00000112799 | 0.743021171 | 3.190398879 | 3.0992899 | 1.029396717 | 0.3032933 | NA | LY86 |
| 8 | ENSG00000183470 | 0.725191537 | 3.148112543 | 2.794163 | 1.126674626 | 0.2598801 | NA | FLJ40288 |
| 9 | ENSG00000049249 | 978.3177644 | 3.02004686 | 0.3438584 | 8.782820244 | 1.59E-18 | 7.36E-15 | TNFRSF9 |
| 10 | ENSG00000162654 | 16.34406424 | 2.883842356 | 0.7528057 | 3.830792187 | 0.0001277 | 0.013333854 | GBP4 |
| 11 | ENSG00000134595 | 0.582423226 | 2.87516762 | 2.9863258 | 0.962777621 | 0.3356591 | NA | SOX3 |

*Figure 27 showing the list of the top-10 genes which are up-regulated after 4 hours of TNF treatment*



| 71 | gene_id | baseMean | log2FoldChange | lfcSE | stat | pvalue | padj | Symbol |
|---|---|---|---|---|---|---|---|---|
| 72 | ENSG00000224940 | 66.15048897 | -0.703948758 | 0.3199696 | -2.200048976 | 0.0278034 | 0.361768148 | PRRT4 |
| 73 | ENSG00000138336 | 302.4624857 | -0.707950921 | 0.3613334 | -1.959273661 | 0.0500807 | 0.455833476 | TET1 |
| 74 | ENSG00000143772 | 462.5430448 | -0.709977979 | 0.2234631 | -3.177160146 | 0.0014872 | 0.071213762 | ITPKB |
| 75 | ENSG00000158156 | 23.84909928 | -0.715593493 | 0.4932847 | -1.450670474 | 0.1468716 | 0.658600839 | XKR8 |
| 76 | ENSG00000242259 | 153.0547571 | -0.721634826 | 0.291977 | -2.471547221 | 0.013453 | 0.257235217 | C22orf39 |
| 77 | ENSG00000175707 | 9.410250967 | -0.737202766 | 0.689827 | -1.068677789 | 0.2852149 | 0.756589301 | KDF1 |
| 78 | ENSG00000107014 | 5.04875468 | -0.751448853 | 1.0613051 | -0.708042226 | 0.478919 | NA | RLN2 |
| 79 | ENSG00000186193 | 3061.277344 | -0.753547049 | 0.2574871 | -2.926543322 | 0.0034275 | 0.121001079 | SAPCD2 |
| 80 | ENSG00000120875 | 10.247087 | -0.755767072 | 0.6657236 | -1.135256457 | 0.2562679 | 0.742284716 | DUSP4 |
| 81 | ENSG00000109819 | 164.8210036 | -0.76004339 | 0.3730372 | -2.037446693 | 0.0416053 | 0.433439442 | PPARGC1A |

*Figure 28 showing the list of the top-10 genes which are down-regulated after 4-hours of TNF treatment*

- **Common Genes:** As mentioned in the above step we have analyzed the data very carefully and found that there are 2 genes which were up-regulated in the 30 minutes of TNF treatment but down-regulated in the 4-hours of TNF treatment. So, we can say these types of genes are affected the most. These 2 genes are ‘CXCL2’ and ‘SYNDIG1L’. SYNDIG1L is associated with the Huntington disease[43] and CXCL2 is associated with wound healing, cancer metastasis and angiogenesis[44].

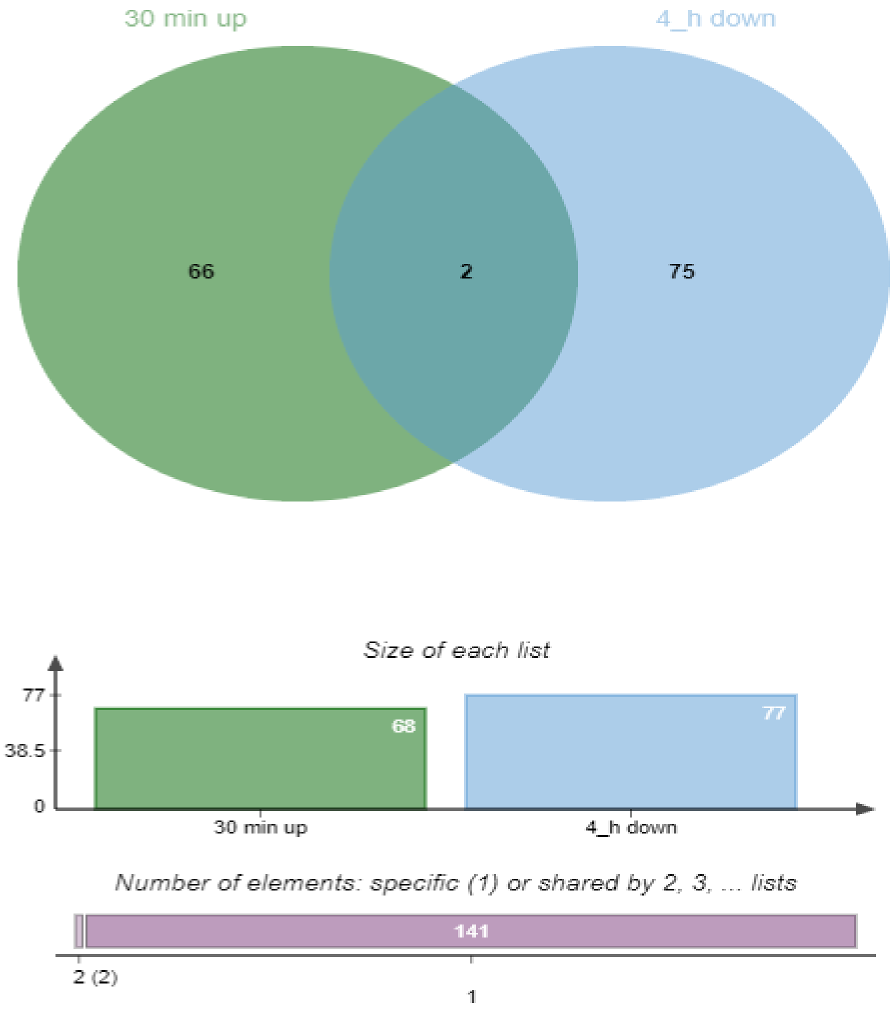
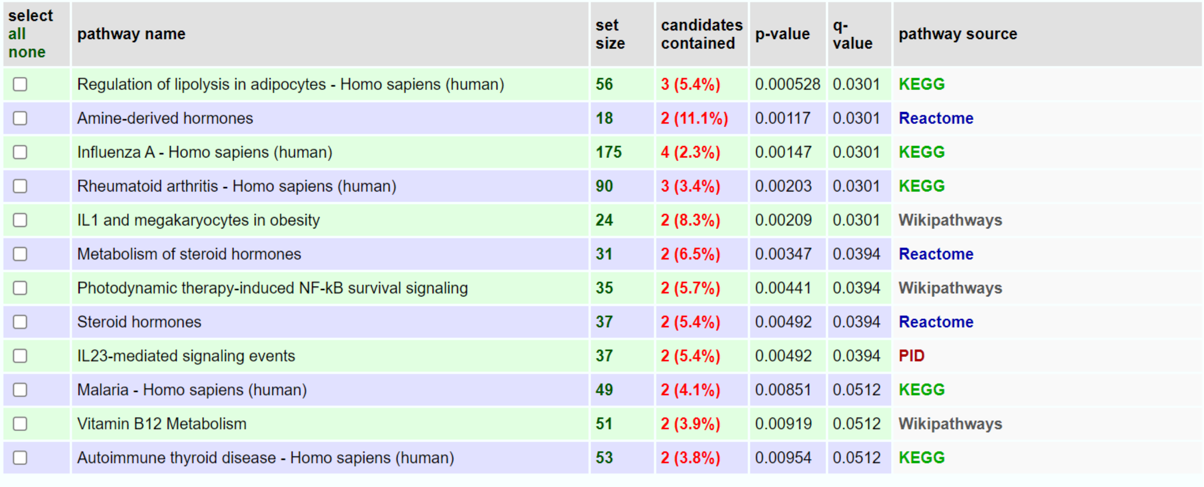


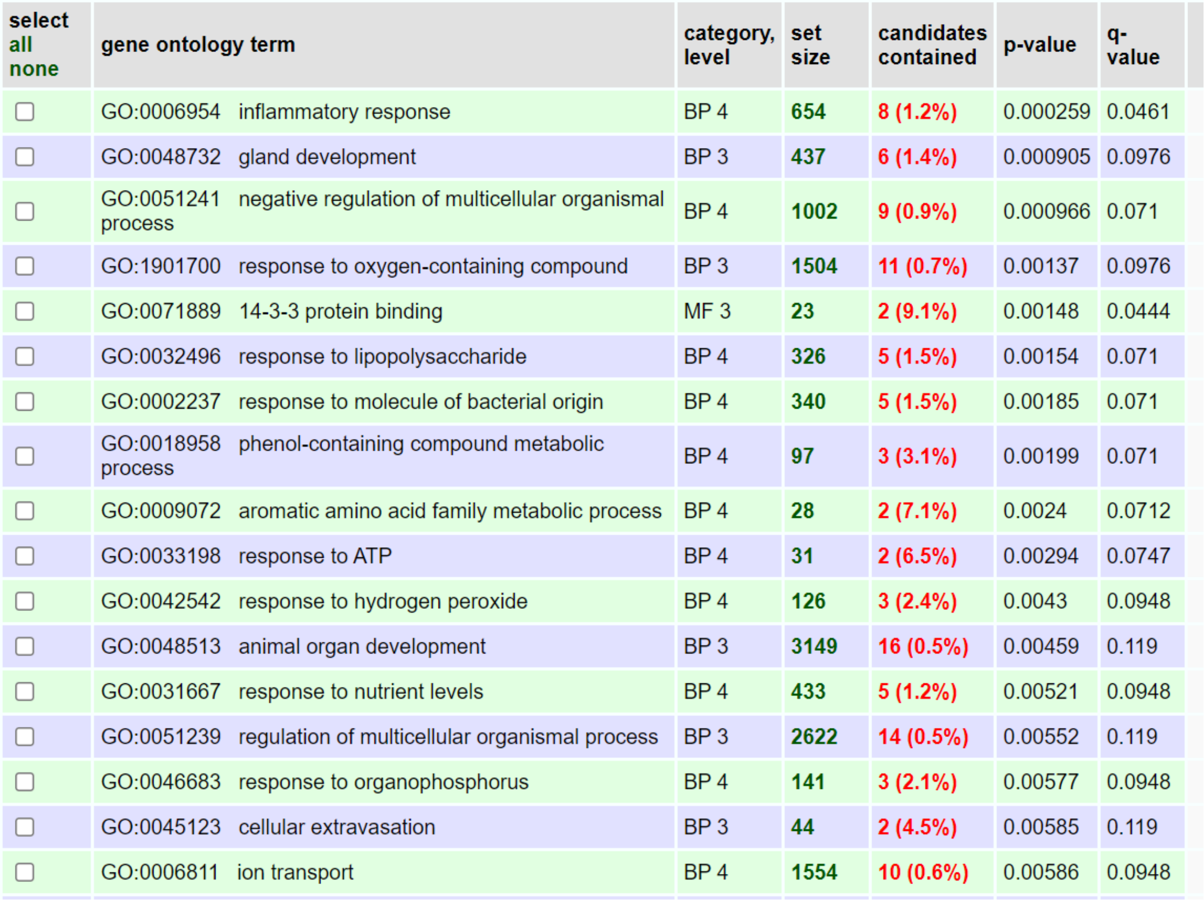
*Figure 29 showing the Venn Diagram of 30 mins up_regulated sample vs 4_hours downregulated genes sample.*

- **Pathway Enrichment Analysis:** We got the differentially expressed genes in the last step and to know the function of these genes and pathways in which they involved pathway analysis step was performed. Below figures showing that these genes are involved in variety of pathways and some genes are also responsible in NF-κB and this transcription factor is responsible for anti-cancer treatment likewise there are many other genes responsible for many other diseases. We have also mentioned the Gene ontology of all these genes in the below figures.



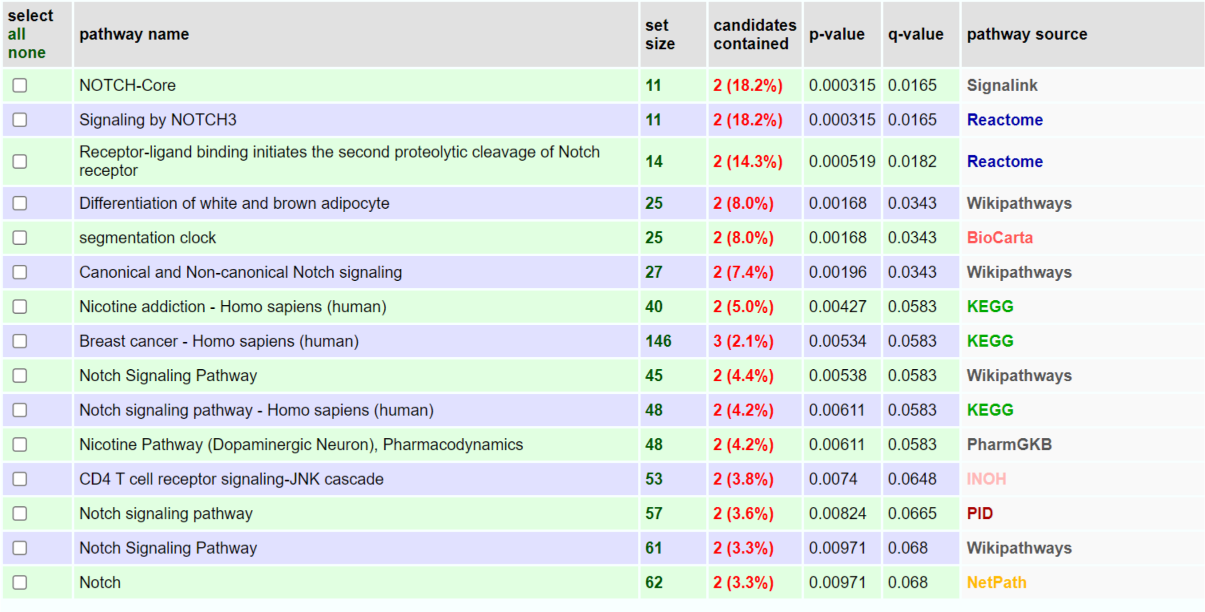
| select<br>all<br>none | pathway name | set size | candidates contained | p-value | q-value | pathway source |
|---|---|---|---|---|---|---|
| ☐ | Regulation of lipolysis in adipocytes - Homo sapiens (human) | 56 | 3 (5.4%) | 0.000528 | 0.0301 | KEGG |
| ☐ | Amine-derived hormones | 18 | 2 (11.1%) | 0.00117 | 0.0301 | Reactome |
| ☐ | Influenza A - Homo sapiens (human) | 175 | 4 (2.3%) | 0.00147 | 0.0301 | KEGG |
| ☐ | Rheumatoid arthritis - Homo sapiens (human) | 90 | 3 (3.4%) | 0.00203 | 0.0301 | KEGG |
| ☐ | IL1 and megakaryocytes in obesity | 24 | 2 (8.3%) | 0.00209 | 0.0301 | Wikipathways |
| ☐ | Metabolism of steroid hormones | 31 | 2 (6.5%) | 0.00347 | 0.0394 | Reactome |
| ☐ | Photodynamic therapy-induced NF-kB survival signaling | 35 | 2 (5.7%) | 0.00441 | 0.0394 | Wikipathways |
| ☐ | Steroid hormones | 37 | 2 (5.4%) | 0.00492 | 0.0394 | Reactome |
| ☐ | IL23-mediated signaling events | 37 | 2 (5.4%) | 0.00492 | 0.0394 | PID |
| ☐ | Malaria - Homo sapiens (human) | 49 | 2 (4.1%) | 0.00851 | 0.0512 | KEGG |
| ☐ | Vitamin B12 Metabolism | 51 | 2 (3.9%) | 0.00919 | 0.0512 | Wikipathways |
| ☐ | Autoimmune thyroid disease - Homo sapiens (human) | 53 | 2 (3.8%) | 0.00954 | 0.0512 | KEGG |

*Figure 30 shows the list of pathways in which the upregulated genes of treated sample are involved (4 hours)*



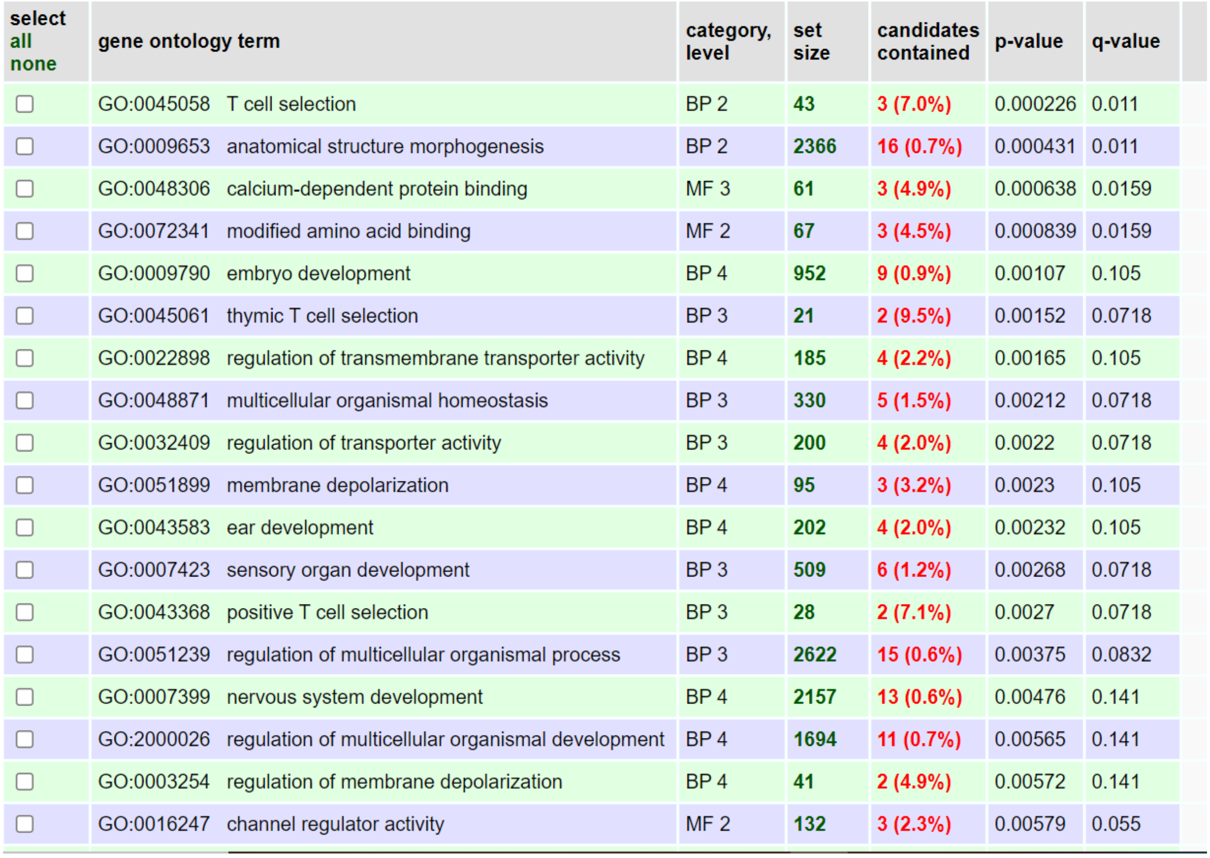
| select<br>all<br>none | gene ontology term | category, level | set size | candidates contained | p-value | q-value |
|---|---|---|---|---|---|---|
| ☐ | GO:0006954 inflammatory response | BP 4 | 654 | 8 (1.2%) | 0.000259 | 0.0461 |
| ☐ | GO:0048732 gland development | BP 3 | 437 | 6 (1.4%) | 0.000905 | 0.0976 |
| ☐ | GO:0051241 negative regulation of multicellular organismal process | BP 4 | 1002 | 9 (0.9%) | 0.000966 | 0.071 |
| ☐ | GO:1901700 response to oxygen-containing compound | BP 3 | 1504 | 11 (0.7%) | 0.00137 | 0.0976 |
| ☐ | GO:0071889 14-3-3 protein binding | MF 3 | 23 | 2 (9.1%) | 0.00148 | 0.0444 |
| ☐ | GO:0032496 response to lipopolysaccharide | BP 4 | 326 | 5 (1.5%) | 0.00154 | 0.071 |
| ☐ | GO:0002237 response to molecule of bacterial origin | BP 4 | 340 | 5 (1.5%) | 0.00185 | 0.071 |
| ☐ | GO:0018958 phenol-containing compound metabolic process | BP 4 | 97 | 3 (3.1%) | 0.00199 | 0.071 |
| ☐ | GO:0009072 aromatic amino acid family metabolic process | BP 4 | 28 | 2 (7.1%) | 0.0024 | 0.0712 |
| ☐ | GO:0033198 response to ATP | BP 4 | 31 | 2 (6.5%) | 0.00294 | 0.0747 |
| ☐ | GO:0042542 response to hydrogen peroxide | BP 4 | 126 | 3 (2.4%) | 0.0043 | 0.0948 |
| ☐ | GO:0048513 animal organ development | BP 3 | 3149 | 16 (0.5%) | 0.00459 | 0.119 |
| ☐ | GO:0031667 response to nutrient levels | BP 4 | 433 | 5 (1.2%) | 0.00521 | 0.0948 |
| ☐ | GO:0051239 regulation of multicellular organismal process | BP 3 | 2622 | 14 (0.5%) | 0.00552 | 0.119 |
| ☐ | GO:0046683 response to organophosphorus | BP 4 | 141 | 3 (2.1%) | 0.00577 | 0.0948 |
| ☐ | GO:0045123 cellular extravasation | BP 3 | 44 | 2 (4.5%) | 0.00585 | 0.119 |
| ☐ | GO:0006811 ion transport | BP 4 | 1554 | 10 (0.6%) | 0.00586 | 0.0948 |

*Figure 31 shows the list of Gene Ontology of upregulated genes of 4 hour treated sample*



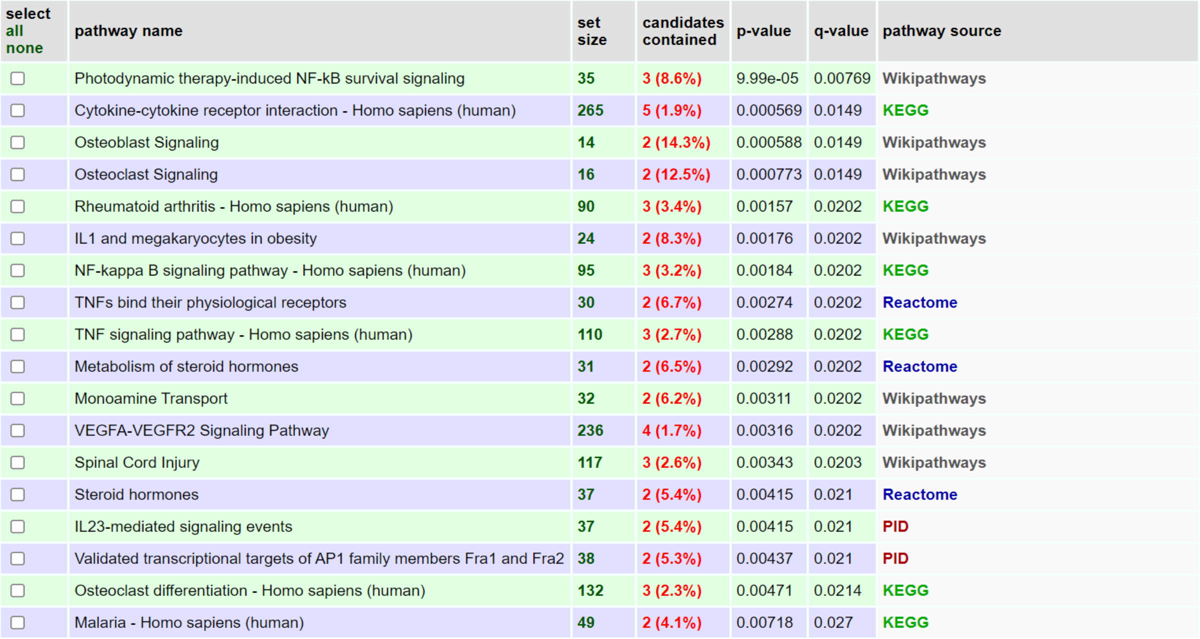
| select<br>all<br>none | pathway name | set size | candidates contained | p-value | q-value | pathway source |
|---|---|---|---|---|---|---|
| □ | NOTCH-Core | 11 | 2 (18.2%) | 0.000315 | 0.0165 | Signalink |
| □ | Signaling by NOTCH3 | 11 | 2 (18.2%) | 0.000315 | 0.0165 | Reactome |
| □ | Receptor-ligand binding initiates the second proteolytic cleavage of Notch receptor | 14 | 2 (14.3%) | 0.000519 | 0.0182 | Reactome |
| □ | Differentiation of white and brown adipocyte | 25 | 2 (8.0%) | 0.00168 | 0.0343 | Wikipathways |
| □ | segmentation clock | 25 | 2 (8.0%) | 0.00168 | 0.0343 | BioCarta |
| □ | Canonical and Non-canonical Notch signaling | 27 | 2 (7.4%) | 0.00196 | 0.0343 | Wikipathways |
| □ | Nicotine addiction - Homo sapiens (human) | 40 | 2 (5.0%) | 0.00427 | 0.0583 | KEGG |
| □ | Breast cancer - Homo sapiens (human) | 146 | 3 (2.1%) | 0.00534 | 0.0583 | KEGG |
| □ | Notch Signaling Pathway | 45 | 2 (4.4%) | 0.00538 | 0.0583 | Wikipathways |
| □ | Notch signaling pathway - Homo sapiens (human) | 48 | 2 (4.2%) | 0.00611 | 0.0583 | KEGG |
| □ | Nicotine Pathway (Dopaminergic Neuron), Pharmacodynamics | 48 | 2 (4.2%) | 0.00611 | 0.0583 | PharmGKB |
| □ | CD4 T cell receptor signaling-JNK cascade | 53 | 2 (3.8%) | 0.0074 | 0.0648 | INOH |
| □ | Notch signaling pathway | 57 | 2 (3.6%) | 0.00824 | 0.0665 | PID |
| □ | Notch Signaling Pathway | 61 | 2 (3.3%) | 0.00971 | 0.068 | Wikipathways |
| □ | Notch | 62 | 2 (3.3%) | 0.00971 | 0.068 | NetPath |

*Figure 32 shows the list of pathways in which the down regulated genes of treated sample are involved (4 hours)*

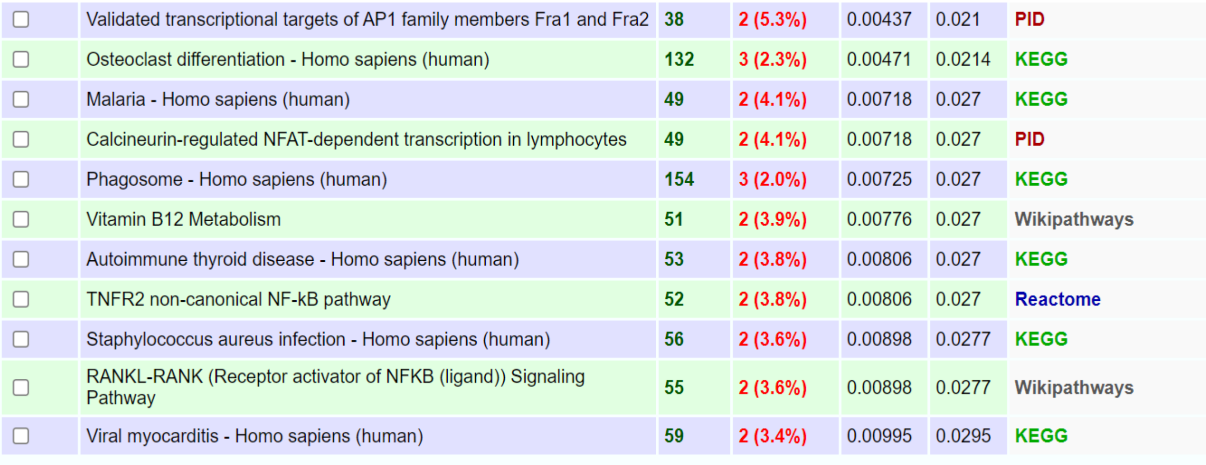


| select<br>all<br>none | gene ontology term | category, level | set size | candidates contained | p-value | q-value |
|---|---|---|---|---|---|---|
| □ | GO:0045058 T cell selection | BP 2 | 43 | 3 (7.0%) | 0.000226 | 0.011 |
| □ | GO:0009653 anatomical structure morphogenesis | BP 2 | 2366 | 16 (0.7%) | 0.000431 | 0.011 |
| □ | GO:0048306 calcium-dependent protein binding | MF 3 | 61 | 3 (4.9%) | 0.000638 | 0.0159 |
| □ | GO:0072341 modified amino acid binding | MF 2 | 67 | 3 (4.5%) | 0.000839 | 0.0159 |
| □ | GO:0009790 embryo development | BP 4 | 952 | 9 (0.9%) | 0.00107 | 0.105 |
| □ | GO:0045061 thymic T cell selection | BP 3 | 21 | 2 (9.5%) | 0.00152 | 0.0718 |
| □ | GO:0022898 regulation of transmembrane transporter activity | BP 4 | 185 | 4 (2.2%) | 0.00165 | 0.105 |
| □ | GO:0048871 multicellular organismal homeostasis | BP 3 | 330 | 5 (1.5%) | 0.00212 | 0.0718 |
| □ | GO:0032409 regulation of transporter activity | BP 3 | 200 | 4 (2.0%) | 0.0022 | 0.0718 |
| □ | GO:0051899 membrane depolarization | BP 4 | 95 | 3 (3.2%) | 0.0023 | 0.105 |
| □ | GO:0043583 ear development | BP 4 | 202 | 4 (2.0%) | 0.00232 | 0.105 |
| □ | GO:0007423 sensory organ development | BP 3 | 509 | 6 (1.2%) | 0.00268 | 0.0718 |
| □ | GO:0043368 positive T cell selection | BP 3 | 28 | 2 (7.1%) | 0.0027 | 0.0718 |
| □ | GO:0051239 regulation of multicellular organismal process | BP 3 | 2622 | 15 (0.6%) | 0.00375 | 0.0832 |
| □ | GO:0007399 nervous system development | BP 4 | 2157 | 13 (0.6%) | 0.00476 | 0.141 |
| □ | GO:2000026 regulation of multicellular organismal development | BP 4 | 1694 | 11 (0.7%) | 0.00565 | 0.141 |
| □ | GO:0003254 regulation of membrane depolarization | BP 4 | 41 | 2 (4.9%) | 0.00572 | 0.141 |
| □ | GO:0016247 channel regulator activity | MF 2 | 132 | 3 (2.3%) | 0.00579 | 0.055 |

*Figure 33 shows the list of Gene Ontology of down regulated genes of 4 hours treated sample*

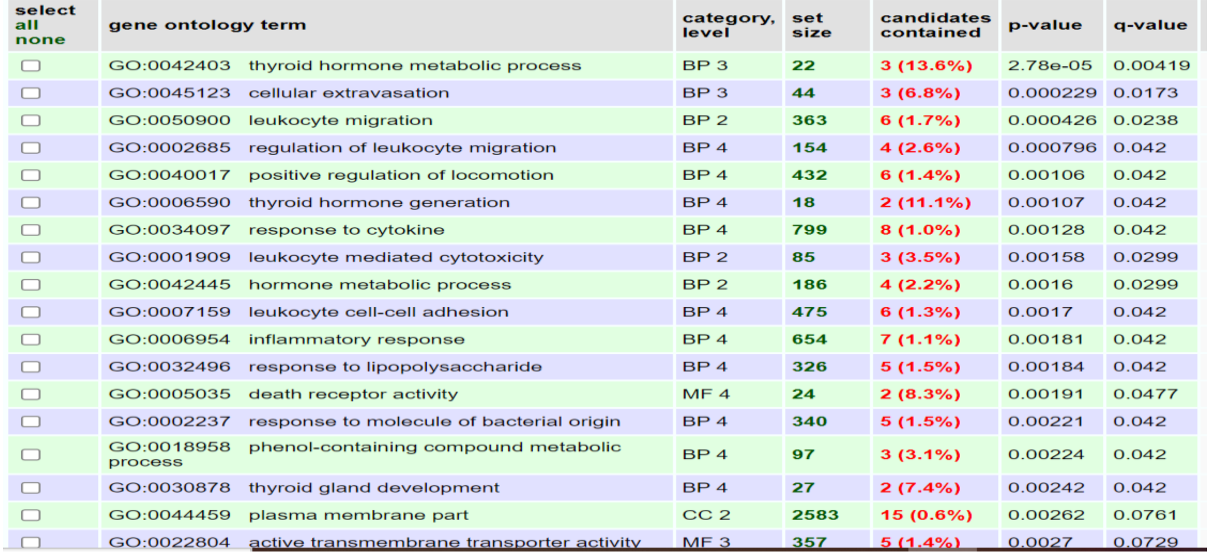


| select<br>all<br>none | pathway name | set size | candidates contained | p-value | q-value | pathway source |
|---|---|---|---|---|---|---|
| ☐ | Photodynamic therapy-induced NF-kB survival signaling | 35 | 3 (8.6%) | 9.99e-05 | 0.00769 | Wikipathways |
| ☐ | Cytokine-cytokine receptor interaction - Homo sapiens (human) | 265 | 5 (1.9%) | 0.000569 | 0.0149 | KEGG |
| ☐ | Osteoblast Signaling | 14 | 2 (14.3%) | 0.000588 | 0.0149 | Wikipathways |
| ☐ | Osteoclast Signaling | 16 | 2 (12.5%) | 0.000773 | 0.0149 | Wikipathways |
| ☐ | Rheumatoid arthritis - Homo sapiens (human) | 90 | 3 (3.4%) | 0.00157 | 0.0202 | KEGG |
| ☐ | IL1 and megakaryocytes in obesity | 24 | 2 (8.3%) | 0.00176 | 0.0202 | Wikipathways |
| ☐ | NF-kappa B signaling pathway - Homo sapiens (human) | 95 | 3 (3.2%) | 0.00184 | 0.0202 | KEGG |
| ☐ | TNFs bind their physiological receptors | 30 | 2 (6.7%) | 0.00274 | 0.0202 | Reactome |
| ☐ | TNF signaling pathway - Homo sapiens (human) | 110 | 3 (2.7%) | 0.00288 | 0.0202 | KEGG |
| ☐ | Metabolism of steroid hormones | 31 | 2 (6.5%) | 0.00292 | 0.0202 | Reactome |
| ☐ | Monoamine Transport | 32 | 2 (6.2%) | 0.00311 | 0.0202 | Wikipathways |
| ☐ | VEGFA-VEGFR2 Signaling Pathway | 236 | 4 (1.7%) | 0.00316 | 0.0202 | Wikipathways |
| ☐ | Spinal Cord Injury | 117 | 3 (2.6%) | 0.00343 | 0.0203 | Wikipathways |
| ☐ | Steroid hormones | 37 | 2 (5.4%) | 0.00415 | 0.021 | Reactome |
| ☐ | IL23-mediated signaling events | 37 | 2 (5.4%) | 0.00415 | 0.021 | PID |
| ☐ | Validated transcriptional targets of AP1 family members Fra1 and Fra2 | 38 | 2 (5.3%) | 0.00437 | 0.021 | PID |
| ☐ | Osteoclast differentiation - Homo sapiens (human) | 132 | 3 (2.3%) | 0.00471 | 0.0214 | KEGG |
| ☐ | Malaria - Homo sapiens (human) | 49 | 2 (4.1%) | 0.00718 | 0.027 | KEGG |



| | | | | | | |
|---|---|---|---|---|---|---|
| ☐ | Validated transcriptional targets of AP1 family members Fra1 and Fra2 | 38 | 2 (5.3%) | 0.00437 | 0.021 | PID |
| ☐ | Osteoclast differentiation - Homo sapiens (human) | 132 | 3 (2.3%) | 0.00471 | 0.0214 | KEGG |
| ☐ | Malaria - Homo sapiens (human) | 49 | 2 (4.1%) | 0.00718 | 0.027 | KEGG |
| ☐ | Calcineurin-regulated NFAT-dependent transcription in lymphocytes | 49 | 2 (4.1%) | 0.00718 | 0.027 | PID |
| ☐ | Phagosome - Homo sapiens (human) | 154 | 3 (2.0%) | 0.00725 | 0.027 | KEGG |
| ☐ | Vitamin B12 Metabolism | 51 | 2 (3.9%) | 0.00776 | 0.027 | Wikipathways |
| ☐ | Autoimmune thyroid disease - Homo sapiens (human) | 53 | 2 (3.8%) | 0.00806 | 0.027 | KEGG |
| ☐ | TNFR2 non-canonical NF-kB pathway | 52 | 2 (3.8%) | 0.00806 | 0.027 | Reactome |
| ☐ | Staphylococcus aureus infection - Homo sapiens (human) | 56 | 2 (3.6%) | 0.00898 | 0.0277 | KEGG |
| ☐ | RANKL-RANK (Receptor activator of NFKB (ligand)) Signaling Pathway | 55 | 2 (3.6%) | 0.00898 | 0.0277 | Wikipathways |
| ☐ | Viral myocarditis - Homo sapiens (human) | 59 | 2 (3.4%) | 0.00995 | 0.0295 | KEGG |

*Figure 34 shows the list of pathways in which the upregulated genes of treated sample are involved (30 mins)*



| select<br>all<br>none | gene ontology term | | category, level | set size | candidates contained | p-value | q-value |
|---|---|---|---|---|---|---|---|
| ☐ | GO:0042403 | thyroid hormone metabolic process | BP 3 | 22 | 3 (13.6%) | 2.78e-05 | 0.00419 |
| ☐ | GO:0045123 | cellular extravasation | BP 3 | 44 | 3 (6.8%) | 0.000229 | 0.0173 |
| ☐ | GO:0050900 | leukocyte migration | BP 2 | 363 | 6 (1.7%) | 0.000426 | 0.0238 |
| ☐ | GO:0002685 | regulation of leukocyte migration | BP 4 | 154 | 4 (2.6%) | 0.000796 | 0.042 |
| ☐ | GO:0040017 | positive regulation of locomotion | BP 4 | 432 | 6 (1.4%) | 0.00106 | 0.042 |
| ☐ | GO:0006590 | thyroid hormone generation | BP 4 | 18 | 2 (11.1%) | 0.00107 | 0.042 |
| ☐ | GO:0034097 | response to cytokine | BP 4 | 799 | 8 (1.0%) | 0.00128 | 0.042 |
| ☐ | GO:0001909 | leukocyte mediated cytotoxicity | BP 2 | 85 | 3 (3.5%) | 0.00158 | 0.0299 |
| ☐ | GO:0042445 | hormone metabolic process | BP 2 | 186 | 4 (2.2%) | 0.0016 | 0.0299 |
| ☐ | GO:0007159 | leukocyte cell-cell adhesion | BP 4 | 475 | 6 (1.3%) | 0.0017 | 0.042 |
| ☐ | GO:0006954 | inflammatory response | BP 4 | 654 | 7 (1.1%) | 0.00181 | 0.042 |
| ☐ | GO:0032496 | response to lipopolysaccharide | BP 4 | 326 | 5 (1.5%) | 0.00184 | 0.042 |
| ☐ | GO:0005035 | death receptor activity | MF 4 | 24 | 2 (8.3%) | 0.00191 | 0.0477 |
| ☐ | GO:0002237 | response to molecule of bacterial origin | BP 4 | 340 | 5 (1.5%) | 0.00221 | 0.042 |
| ☐ | GO:0018958 | phenol-containing compound metabolic process | BP 4 | 97 | 3 (3.1%) | 0.00224 | 0.042 |
| ☐ | GO:0030878 | thyroid gland development | BP 4 | 27 | 2 (7.4%) | 0.00242 | 0.042 |
| ☐ | GO:0044459 | plasma membrane part | CC 2 | 2583 | 15 (0.6%) | 0.00262 | 0.0761 |
| ☐ | GO:0022804 | active transmembrane transporter activity | MF 3 | 357 | 5 (1.4%) | 0.0027 | 0.0729 |

*Figure 35 shows the list of Gene Ontology of down regulated genes of 30 minutes treated sample*

## DISCUSSION AND CONCLUSION:

This study aim was to perform all the differentially expressed genes due to TNF treatment at for different time period. Since this study involves lot of computational thinking along with computational power to perform on the local work station. For the purpose of this study, we have used a high configuration server which help us to complete this study on time.

We have successfully computed all the steps and extract out all the genes which were differentially expressed. Representing all the data manually in the forms of table took more time and is not able possible in some case so, we used graphs and plots to represent wherever possible.

After getting the names of differentially expressed genes using p-value and log2FoldChange as the main perimeter, we tried to identify the pathways in which these genes are involved and we have noticed that these genes were involved in the variety of pathways. Some of these pathways are related to cancer and rare diseases. So, knowing the functions of these genes completes our aim of study but we can extend it further because the secondary analysis of this study still needs to perform which will tell us the more information about these genes.

## FUTURE SCOPE:

We have completed this study up to the pathway enrichment analysis but this is not the end of this study since, many secondary analyses can be performed on this study and we can able to extract more information about these genes but due to time constraint and limited computing power we have we are not proceeding further now. But this study can able to reveal the role of TNF in many diseases and also, we can find the more potent biomarkers.

Although, we have found many biomarkers out of this huge data due to much further steps becomes easy and less time consuming. In this study we have found two genes which behave differently during different time period. But if we noticed there were many genes which wasn't affected by either of the time period. We can basically explore those genes which are associated in cancer and rare diseases. Many of these genes are associated with the pathways which create resistance in the cancer treatment so in these scenario's this study become more important. We can also perform some comparative studies in which find the effect of some stimulus on these genes and if the effect are same then we ca use the one which is suitable maximally.

## REFERENCES:

[1] V. S and K. M, “Regulation and function of NF-kappaB transcription factors in the immune system,” *Annual review of immunology*, vol. 27, pp. 693–733, 2009, doi: 10.1146/ANNUREV.IMMUNOL.021908.132641.

[2] M. S. Hayden and S. Ghosh, “NF-κB, the first quarter-century: remarkable progress and outstanding questions,” *Genes & Development*, vol. 26, no. 3, p. 203, Feb. 2012, doi: 10.1101/GAD.183434.111.

[3] P. ND, “The diverse and complex roles of NF-κB subunits in cancer,” *Nature reviews. Cancer*, vol. 12, no. 2, pp. 121–132, Feb. 2012, doi: 10.1038/NRC3204.

[4] B. Guttman, “Frameshift Mutation,” *Brenner’s Encyclopedia of Genetics: Second Edition*, pp. 110–110, Feb. 2013, doi: 10.1016/B978-0-12-374984-0.00555-6.

[5] S. Neidle, “The Building-Blocks of DNA and RNA,” *Principles of Nucleic Acid Structure*, pp. 20–37, 2008, doi: 10.1016/B978-012369507-9.50003-0.

[6] B. R. Korf, “Introduction to Human Genetics,” *Clinical and Translational Science: Principles of Human Research: Second Edition*, pp. 281–311, 2017, doi: 10.1016/B978-0-12-802101-9.00016-8.

[7] J. M. Heather and B. Chain, “The sequence of sequencers: The history of sequencing DNA,” *Genomics*, vol. 107, no. 1, p. 1, Jan. 2016, doi: 10.1016/J.YGENO.2015.11.003.

[8] F. Sanger, S. Nicklen, and A. R. Coulson, “DNA sequencing with chain-terminating inhibitors,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, no. 12, p. 5463, 1977, doi: 10.1073/PNAS.74.12.5463.

[9] A. Totomoch-Serra, M. F. Marquez, and D. E. Cervantes-Barragán, “Sanger sequencing as a first-line approach for molecular diagnosis of Andersen-Tawil syndrome,” *F1000Research*, vol. 6, 2017, doi: 10.12688/F1000RESEARCH.11610.1.

[10] B. Sikkema-Raddatz *et al.*, “Targeted Next-Generation Sequencing can Replace Sanger Sequencing in Clinical Diagnostics,” *Human Mutation*, vol. 34, no. 7, pp. 1035–1042, Jul. 2013, doi: 10.1002/HUMU.22332.

[11] “Istorijat sekvenciranja DNK.” http://scindeks.ceon.rs/article.aspx?artid=1452-82581304301G (accessed Aug. 30, 2021).

[12] M. Gužvić, “The history of DNA sequencing,” *Journal of Medical Biochemistry*, vol. 32, no. 4, pp. 301–312, Oct. 2013, doi: 10.2478/JOMB-2014-0004.

[13] P. J. A. Cock, C. J. Fields, N. Goto, M. L. Heuer, and P. M. Rice, “The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants,” *Nucleic Acids Research*, vol. 38, no. 6, pp. 1767–1771, Apr. 2010, doi: 10.1093/NAR/GKP1137.

[14] P. J. A. Cock, J. Bonfield, B. Chevreux, and H. Li, “SAM/BAM format v1.5 extensions for de novo assemblies,” *bioRxiv*, p. 020024, May 2015, doi: 10.1101/020024.

[15] G. Tischler and S. Leonard, “biobambam: tools for read pair collation based algorithms on BAM files,” *Source Code for Biology and Medicine 2014 9:1*, vol. 9, no. 1, pp. 1–18, Jun. 2014, doi: 10.1186/1751-0473-9-13.

[16] E. Keibler and M. R. Brent, “Eval: A software package for analysis of genome annotations,” *BMC Bioinformatics 2003 4:1*, vol. 4, no. 1, pp. 1–4, Oct. 2003, doi: 10.1186/1471-2105-4-50.

[17] “PipeMAGI: an integrated and validated workflow for analysis of NGS data for clinical diagnostics”, Accessed: Aug. 30, 2021. [Online]. Available: https://www.bioin-

[18] X. Qiu, A. I. Brooks, L. Klebanov, and A. Yakovlev, “The effects of normalization on the correlation structure of microarray data,” *BMC Bioinformatics 2005 6:1*, vol. 6, no. 1, pp. 1–11, May 2005, doi: 10.1186/1471-2105-6-120.

[19] B. Uragun and R. Rajan, “Developing an appropriate data normalization method,” *Proceedings - 10th International Conference on Machine Learning and Applications, ICMLA 2011*, vol. 2, pp. 195–199, 2011, doi: 10.1109/ICMLA.2011.53.

[20] F. Rapaport *et al.*, “Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data,” *Genome Biology 2013 14:9*, vol. 14, no. 9, pp. 1–13, Sep. 2013, doi: 10.1186/GB-2013-14-9-R95.

[21] J. T. Leek, E. Monsen, A. R. Dabney, and J. D. Storey, “EDGE: extraction and analysis of differential gene expression,” *Bioinformatics*, vol. 22, no. 4, pp. 507–508, Feb. 2006, doi: 10.1093/BIOINFORMATICS/BTK005.

[22] M. E. Ritchie *et al.*, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Research*, vol. 43, no. 7, pp. e47–e47, Apr. 2015, doi: 10.1093/NAR/GKV007.

[23] C. Soneson and M. Delorenzi, “A comparison of methods for differential expression analysis of RNA-seq data,” *BMC Bioinformatics 2013 14:1*, vol. 14, no. 1, pp. 1–18, Mar. 2013, doi: 10.1186/1471-2105-14-91.

[24] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2,” *Genome Biology 2014 15:12*, vol. 15, no. 12, pp. 1–21, Dec. 2014, doi: 10.1186/S13059-014-0550-8.

[25] E. Maza, “In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design,” *Frontiers in Genetics*, vol. 0, no. SEP, p. 164, Sep. 2016, doi: 10.3389/FGENE.2016.00164.

[26] L. Chen, J. Reeve, L. Zhang, S. Huang, X. Wang, and J. Chen, “GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data,” *PeerJ*, vol. 6, no. 4, p. e4600, Apr. 2018, doi: 10.7717/PEERJ.4600.

[27] X. Li *et al.*, “A combined approach with gene-wise normalization improves the analysis of RNA-seq data in human breast cancer subtypes,” *PLOS ONE*, vol. 13, no. 8, p. e0201813, Aug. 2018, doi: 10.1371/JOURNAL.PONE.0201813.

[28] W. A. da Silveira, E. S. Hazard, D. Chung, and G. Hardiman, “Molecular Profiling of RNA Tumors Using High-Throughput RNA Sequencing: From Raw Data to Systems Level Analyses,” *Methods in Molecular Biology*, vol. 1908, pp. 185–204, 2019, doi: 10.1007/978-1-4939-9004-7_13.

[29] M. O. Arowolo, M. Adebiyi, A. Adebiyi, and O. Okesola, “PCA Model for RNA-Seq Malaria Vector Data Classification Using KNN and Decision Tree Algorithm,” *2020 International*

*Conference in Mathematics, Computer Engineering and Computer Science, ICMCECS 2020*, Mar. 2020, doi: 10.1109/ICMCECS47690.2020.240881.

[30] W. Cui *et al.*, “Discovery and characterization of long intergenic non-coding RNAs (lincRNA) module biomarkers in prostate cancer: an integrative analysis of RNA-Seq data,” *BMC Genomics 2015 16:7*, vol. 16, no. 7, pp. 1–10, Jun. 2015, doi: 10.1186/1471-2164-16-S7-S3.

[31] F. Wagner, D. Barkley, and I. Yanai, “Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis,” *bioRxiv*, p. 655365, Jun. 2019, doi: 10.1101/655365.

[32] F. Rapaport *et al.*, “Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data,” *Genome Biology 2013 14:9*, vol. 14, no. 9, pp. 1–13, Sep. 2013, doi: 10.1186/GB-2013-14-9-R95.

[33] M. Love, S. Anders, and W. Huber, “Beginner’s guide to using the DESeq2 package,” 2014, doi: 10.1101/002832.

[34] M. Love, S. Anders, and W. Huber, “Differential analysis of RNA-Seq data at the gene level using the DESeq2 package,” 2013, Accessed: Aug. 30, 2021. [Online]. Available: http://bioconductor.org/packages/release/data/experiment/html/parathyroidSE.

[35] M. J. Ruiz, A. M. Cameán, I. M. Moreno, and Y. Picó, “Determination of microcystins in biological samples by matrix solid-phase dispersion and liquid chromatography–mass spectrometry,” *Journal of Chromatography A*, vol. 1073, no. 1–2, pp. 257–262, May 2005, doi: 10.1016/J.CHROMA.2004.08.128.

[36] S. A. Wolf, L. Epping, S. Andreotti, K. Reinert, and T. Semmler, “SCORE: Smart Consensus Of RNA Expression—a consensus tool for detecting differentially expressed genes in bacteria,” *Bioinformatics*, vol. 37, no. 3, pp. 426–428, Apr. 2021, doi: 10.1093/BIOINFORMATICS/BTAA681.

[37] C. M. Ward, T.-H. To, and S. M. Pederson, “ngsReports: a Bioconductor package for managing FastQC reports and other NGS related log files,” *Bioinformatics*, vol. 36, no. 8, pp. 2587–2588, Apr. 2020, doi: 10.1093/BIOINFORMATICS/BTZ937.

[38] Y. Guo, Y. Dai, H. Yu, S. Zhao, D. C. Samuels, and Y. Shyr, “Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis,” *Genomics*, vol. 109, no. 2, pp. 83–90, Mar. 2017, doi: 10.1016/J.YGENO.2017.01.005.

[39] D. Kim, J. M. Paggi, C. Park, C. Bennett, and S. L. Salzberg, “Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype,” *Nature Biotechnology 2019 37:8*, vol. 37, no. 8, pp. 907–915, Aug. 2019, doi: 10.1038/s41587-019-0201-4.

[40] L. Collado-Torres, A. Nellore, and A. E. Jaffe, “recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor,” *F1000Research*, vol. 6, 2017, doi: 10.12688/F1000RESEARCH.12223.1.

[41] Jr. Cecil C. Bridges, “Hierarchical Cluster Analysis:,” *http://dx.doi.org/10.2466/pr0.1966.18.3.851*, vol. 18, no. 3, pp. 851–854, Aug. 2016, doi: 10.2466/PR0.1966.18.3.851.

[42] A. Kamburov, K. Pentchev, H. Galicka, C. Wierling, H. Lehrach, and R. Herwig, “ConsensusPathDB: toward a more complete picture of cell biology,” *Nucleic Acids Research*, vol. 39, no. suppl_1, pp. D712–D717, Jan. 2011, doi: 10.1093/NAR/GKQ1156.

[43] K. E *et al.*, “SynDIG1: an activity-regulated, AMPA- receptor-interacting transmembrane protein that regulates excitatory synapse development,” *Neuron*, vol. 65, no. 1, pp. 80–93, Jan. 2010, doi: 10.1016/J.NEURON.2009.12.021.

[44] N. Genster, O. Østrup, C. Schjalm, T. Eirik Mollnes, J. B. Cowland, and P. Garred, “Ficolins do not alter host immune responses to lipopolysaccharide-induced inflammation in vivo,” *Scientific Reports 2017 7:1*, vol. 7, no. 1, pp. 1–12, Jun. 2017, doi: 10.1038/s41598-017-04121-w.