Alzheimer Disease Analysis by Using RNA Single Transcriptome Data

Omnia Saeed- Youmna Ramadan- Menna Abdelsattar- Mennatullah Mahmoud

*Abstract* Alzheimer's disease is a progressive neurologic disorder that causes the brain to shrink (atrophy) and brain cells to die. It is the most common cause of dementia which is a continuous decline in thinking, behavioral and social skills that affects a person's ability to function independently. Also, it is a general term for memory loss and other cognitive abilities serious. This disease accounts for 60-80% of dementia cases. Accordingly, we need to determine the enriched GO terms between healthy and Late-onset Alzheimer disease (LOAD) in the following transcriptomic single data PRJNA714081. The aim of our project is to identify the deferential expression genes (up and down regulation of genes) that compares cases to controls genes using deseq2 using tximport.r script, also to determine the functional enrichment analysis that defines the pathways and functions of these genes and which genes from data samples contributed to the pathway leads to the occurrence of the disease. To reach our aim we need to make first quality preprocessing and Pseudoalignment by Kallisto against chromosome 22 before any step. Finally, we made interpretation for analysis to prove the truth of our results. Also we made bash script by command line (using R programming language in linux for running the project).

# Introduction

Alzheimer’s disease is a brain disorder that slowly destroys memory and thinking skills and, eventually, the ability to carry out the simplest tasks. It is a progressive neurologic disorder that causes the brain to shrink (atrophy) and brain cells to die. Also, it is the most common cause of dementia which is a continuous decline in thinking, behavioral and social skills that affects a person's ability to function independently. So, it is considered as a general term for memory loss and other cognitive abilities serious. In most people with the disease those with the late-onset type symptoms first appear in their mid-60s. Early-onset Alzheimer’s occurs between a person’s 30s and mid-60s and is very rare. This disease accounts for 60-80% of dementia cases. Accordingly, we need to determine the enriched GO terms between healthy and Late-onset Alzheimer disease (LOAD) in the following transcriptomic single data PRJNA714081. RNA-Seq is a sequencing technique which uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment, analyzing the continuously changing cellular transcriptome. Single-read sequencing involves sequencing RNA from only one end. The Gene Ontology (GO) is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. More specifically, it aims to: 1) maintain and develop its controlled vocabulary of gene and gene product attributes; 2) annotate genes and gene products, and assimilate and disseminate annotation data; and 3) provide tools for easy access to all aspects of the data provided by the project, and to enable functional interpretation of experimental data using the GO, for example via enrichment analysis. The Gene Ontology (GO) knowledgebase is the world’s largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research. In our project we downloaded our data from NCBI which is The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information. We made filtrations to the data as there are 15 samples of transcriptomic single data PRJNA714081 each one contains about 2 million reads and this is computationally expensive (needs high computational power). So, we filtered this data to get 15 samples in which each sample contains only 10000 genes to deal with the RAM of our laptops. The aim of our project is to identify the deferential expression of genes (up and down regulation of genes) that compares cases to controls using deseq2 using tximport.r script by command line in Linux, also to determine the functional enrichment analysis that defines the pathways and functions of these genes and which genes from data samples contributed to the pathway leads to the occurrence of Alzaheimer’s disease. The goal of differential expression testing is to determine which genes are expressed at different levels between conditions. These genes can offer biological insight into the processes affected by the conditions of interest. Differential expression analysis means taking the normalized read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups. For example, we use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, and whether it is greater than what would be expected just due to natural random variation. Comparison of differential gene expression patterns in these conditions has enabled the identification of common elements that are significantly enriched in gene classes with particular functions such as protein synthesis, hormone delivery, and morphological plasticity. It is also used for determining whether the gene is upregulated, downregulated, turned on or off with respect to cases and compare the results to controls. There are different methods for differential expression analysis such as edgeR and DESeq based on negative binomial (NB) distributions or baySeq and EBSeq which are Bayesian approaches based on a negative binomial model. In our project, we used DESeq. It is important to consider the experimental design when choosing an analysis method. DESeq can perform multiple comparisons. One of the main uses of the GO is to perform enrichment analysis on gene sets. For example, given a set of genes that are up-regulated under certain conditions, an enrichment analysis will find which GO terms are over-represented (or under-represented) using annotations for that gene set. Gene set enrichment analysis (GSEA) (also functional enrichment analysis) is a method to identify classes of genes or proteins that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes. The method uses statistical approaches to identify significantly enriched or depleted groups of genes. Transcriptomic technologies and proteomics results often identify thousands of genes which are used for the analysis. Researchers performing high-throughput experiments that yield sets of genes (for example, genes that are differentially expressed under different conditions) often want to retrieve a functional profile of that gene set, in order to better understand the underlying biological processes. This can be done by comparing the input gene set to each of the bins (terms) in the gene ontology – a statistical test can be performed for each bin to see if it is enriched for the input genes. So, functional enrichment analysis is to compare the enrichment of any type of biologically relevant labels such as gene ontology terms in these genes to that in the background. Hypergeometric distribution or binomial distribution is usually applied to give the p-value for whether the enrichment is significantly different in the given gene cluster and in the background. So the choice of background and the cutoff value are important for the reporting result. To reach the aim of our project we need to make first quality preprocessing step by apply the quality step on the minimized data using FastQC after converting the data into fastq files using SRA tools. Then we decide if we need to shuffle the data or not using seqkit tool. In our case all the data lies between 20-30 quality score in the yellow region so when we made shuffle nothing changes as the data was not large enough. Also trimming the data didn’t make any change for the same reason. After the quality step we need to perform Pseudoalignment by Kallisto against chromosome 22. A pseudoalignment of a read is simply a set of target sequences that the read is compatible with. Contrary to normal read alignment, where we specify where the read aligns and how, for example how the nucleotides in the read match up with the nucleotides in the target sequence. a new tool called kallisto is considered to the bioinformatics community which marks a huge advancement in how RNA-seq analysis is done. Kallisto is a “lightweight algorithm” that is super-fast at quantifying the abundance of transcripts from RNA-seq data with high accuracy. The speed of the program can be attributed to the usage of “psuedoalignments” which aims to determine, for each read, not where in each transcript it aligns, but rather which transcripts it is compatible with. As such, it’s NOT necessary to do a full alignment of the reads to the genome which is often the slowest step in sequencing analysis. Instead, the raw sequence reads (e.g. fastq) are directly compared to transcript sequences and then used to quantify transcript abundance. In order to analyze data with Kallisto we need several inputs: We need the FastQ files from the RNA-Seq experiment; we usually start with reads that have already been trimmed/ filtered then we need a reference transcriptome. This is a file that has the sequences for all the known expressed genes. Reference transcriptomes are usually available from repositories like Ensemble and NCBI. We will be using the human reference transcriptome of chromosome 22. Unlike a genome, the transcriptome is only coding genes. finally, we made interpretation for analysis to prove the truth of our results. As data analysis and interpretation which is the process of assigning meaning to the collected information and determining the conclusions, significance, and implications of the findings. Also we made bash script by command line (by R) in linux for running the project).

# Methodology

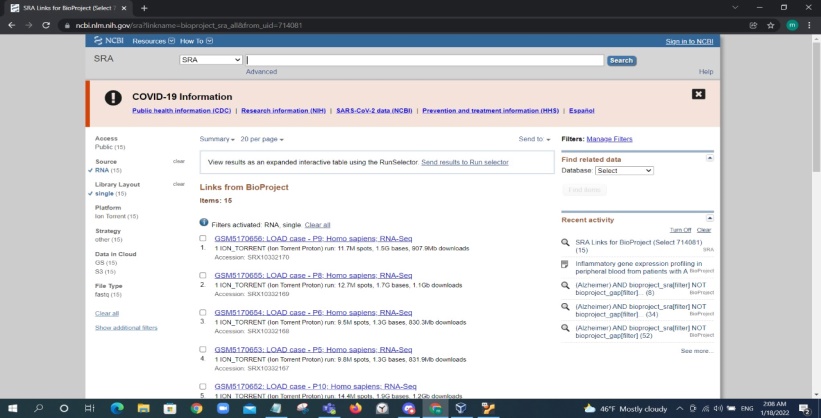
## Download Data

First downloaded the project data through these steps:

1. Download sra-tools by command line: $ conda install -c bioconda sra-tools.
2. Download entrez-direct by command line: $ conda install -c bioconda entrez-direct.
3. Download metadata by command line: $ Download single data using: esearch -db sra -query [PRJNA714081] | efetch -format runinfo | cut -d "," -f 1 | grep SRR | xargs fastq-dump -X 10000 -- skip-technical --read-filter pass –dumpbase –gzip.

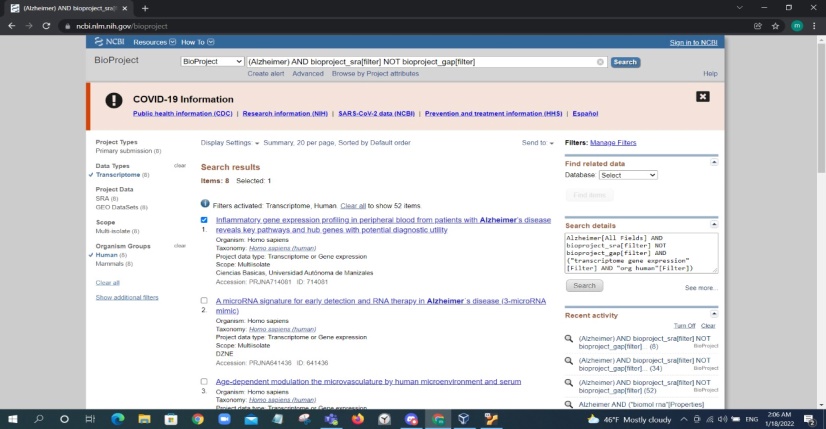
We did some filtrations to reach these data:

1. We choose SRA Database in NCBI and searched for Alzheimer, then we filtered data types into transcriptome, and then choose Human from organism groups, then we finally choose the first project.



Fig(1)

To get the data of this project we have searched for SRA run selector and send our data to Run selector, download Meta data and use the above command lines to download 1000 read from each sample that contains 20 M reads to work.



Fig(2)

## Quality Score

Second, we got the quality score, we used the below command line to get the quality score: $ for f in ~/workdir/fqData/\*.fq.gz;do fastqc –t 1 –f fastq –noextract $f;done.

* If the quality is low, we checked any of the two steps below:

1. Shuffle by command line: $ for f in ~/workdir/\*.fq.gz;do seqkit shuffle $f >; done.
2. Extract the data from the middle by command line: $ esearch –db sra –query PRJNA714081 | efetch –format runinfo | cut –d “,” –f 1 | grep SRR | xargs fastq-dump –N 40000-X 50000 –skip-technical –read-filter pass –gzip.

## Pseudoalignment using Kalisto

Third, we do Pseudoalignment using Kalisto to align out read against chromosome 22 by the following steps:

1. Download reference file “ Chromosome 22 “

* Download human transcriptome by command lines:

$ wget –c ftp://ftp.ebi.ac.uk/pub/database/gencode/Gencode\_human/rlease\_29/gencode.v29.pc\_transcripts.fa.gz.$ gunzip gencode.v29.pc\_transcripts.fa.gz.

* Download the Transcriptome Annotation File by command lines:

$ wget –c ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode\_human/release\_29/gencode.v29.annotation.gtf.gz.$ gunzip gencode.v29.annotation.gtf.gz

* Select the transcripts of Chr22 by command line: $ cd ~/workdir/sample\_data/

$conda activate ngs1

$ READS=$(grep “^chr22” gencode.v29.annotation.gtf | awk –F’\t’ ‘{print $9}’ | awk –F’;’ ‘{print $1}’ | awk –F’ ‘ ‘{print $2}’ | awk –F’”’ ‘{print $2}’ | sort | uniq)

for value in $READS

do

echo “Processing: $value”

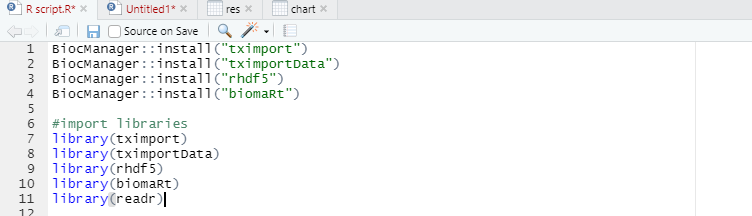
seqkit grep –r –p ${value} gencode.v29.pc\_transcripts.fa | awk –F’|’ ‘{print $1}’ >> gencode.v29.pc\_transcripts.chr22.simplified.fa

done.

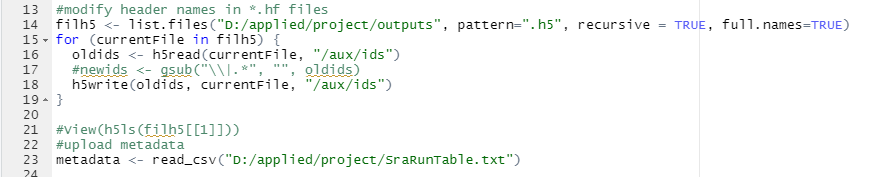
* After that perform steps of Kalisto by the following steps:
* Download Kalisto by command line: $ conda install –c bioconda –y kallisto.
* Indexing by command line: $ kallisto index –I transcriptome.idx –k 25 gencode.v29.pc\_transcripts.chr22.simplified.fa
* kalisto pseudoalignment by command line: $ for I in \*.fastq.gz; do kallisto quant –I transcriptome.idx –single –l 200 -s 20 -o outputs/$i aligned "$i" --pseudobam; done.

## Deferential expression using deseq2

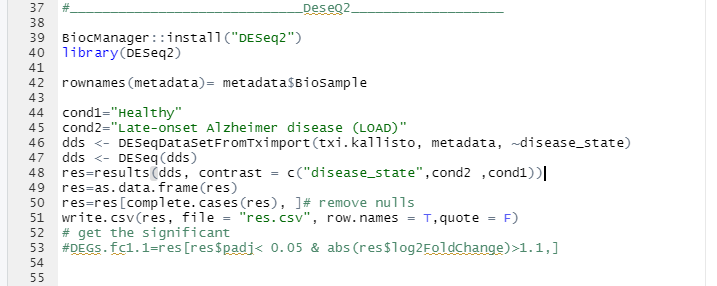
* First download libraries using for deseq2 then import them like:



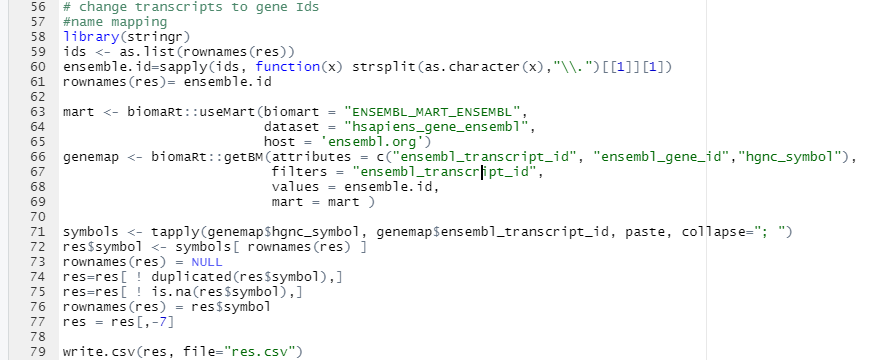
* Second create list of files then pick up files ended by .h5 from my samples of kallisto results then modify their header names files to appear the transcripts to be easily enter to deseq function and download the metadata of my samples:



* Third change the names of filh5 files’ names to their bio samples names then using the outputs of the kallisto to input for deseq2 by tximport library, then put the outputs exits to deseq2 library functions, in deseq2 and compare between two conditions (“healthy” & “Late-onset Alzheimer disease (LOAD)”):



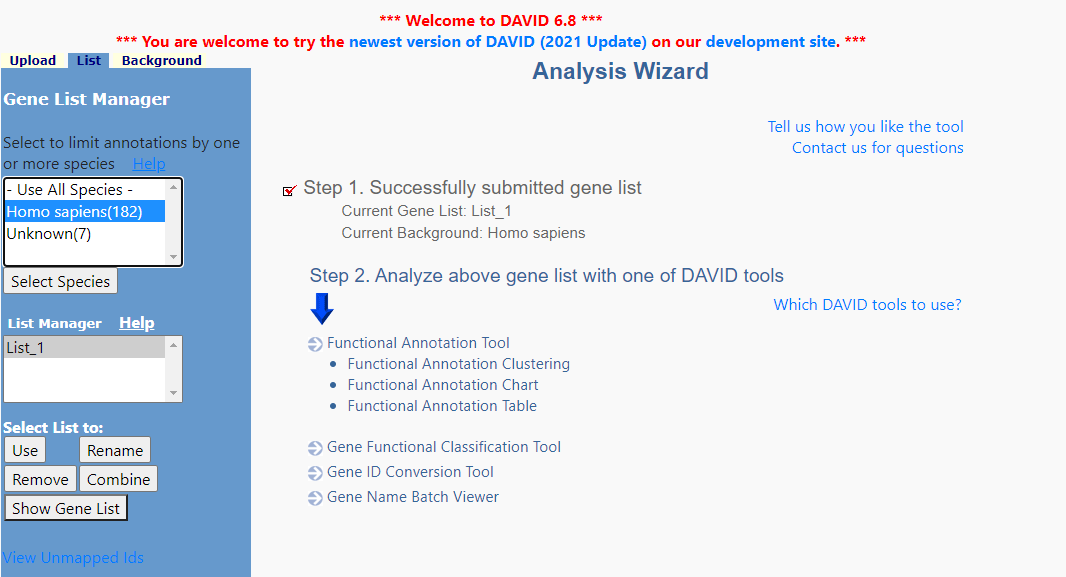
* Fourth change transcripts to gene IDs and name mapping of the ensemble IDs and save the results in res.csv file.

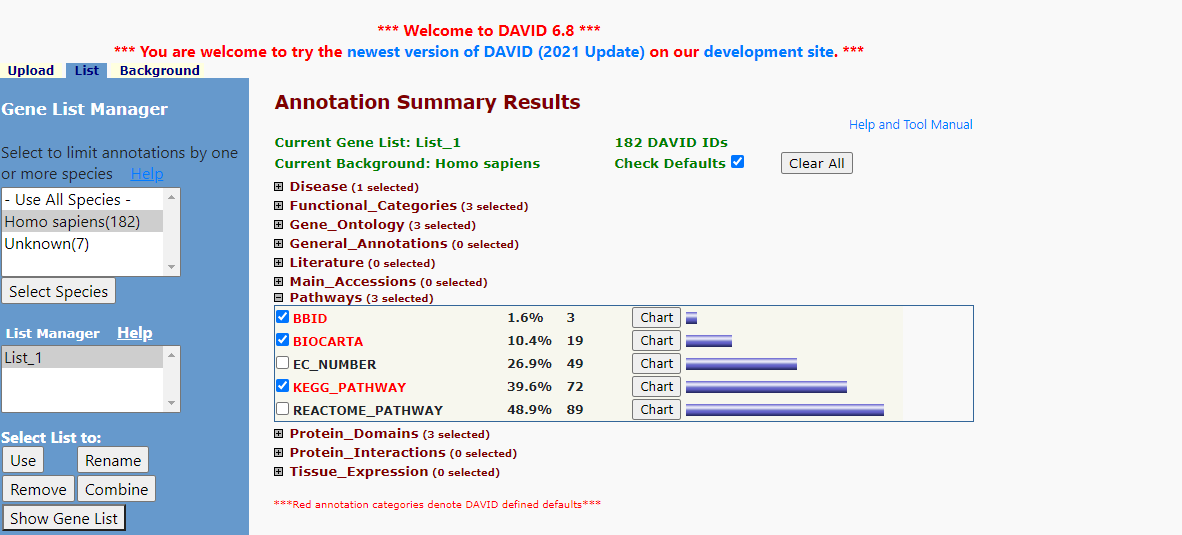


Fifth in David functional enrichment analysis:

1. Make start analysis then take a copy of genes.
2. Choose official gene symbol.
3. Homosapiens.
4. Gene list.
5. Then submit file.
6. Choose functional annotation chart.
7. Download KEGG\_pathways chart.





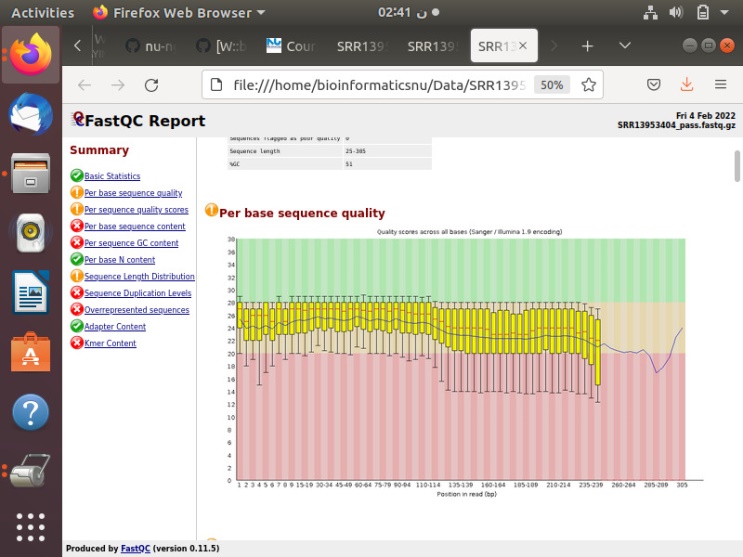


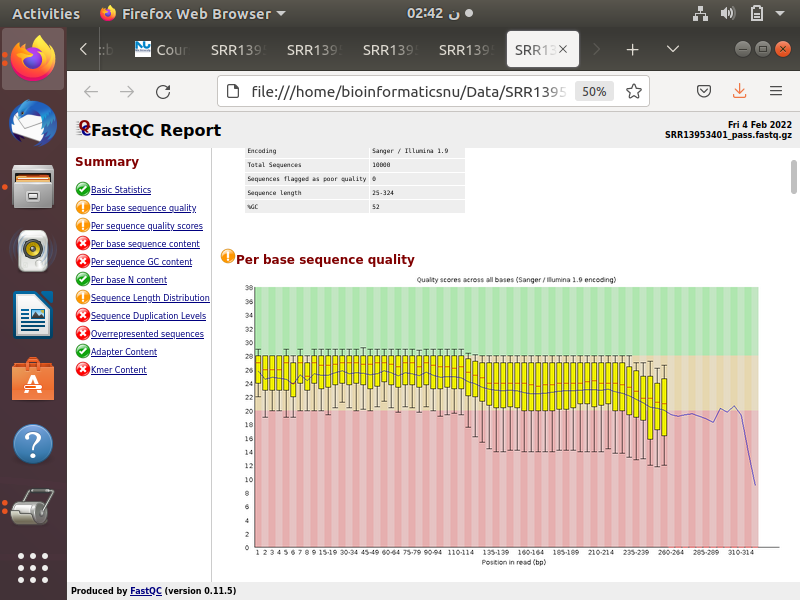
* Finally using GOplot package using R script concentrates on the visualization of biological data. More precisely, the package will help combine and integrate expression data with the results of a functional analysis. It used for visualization purpose only.

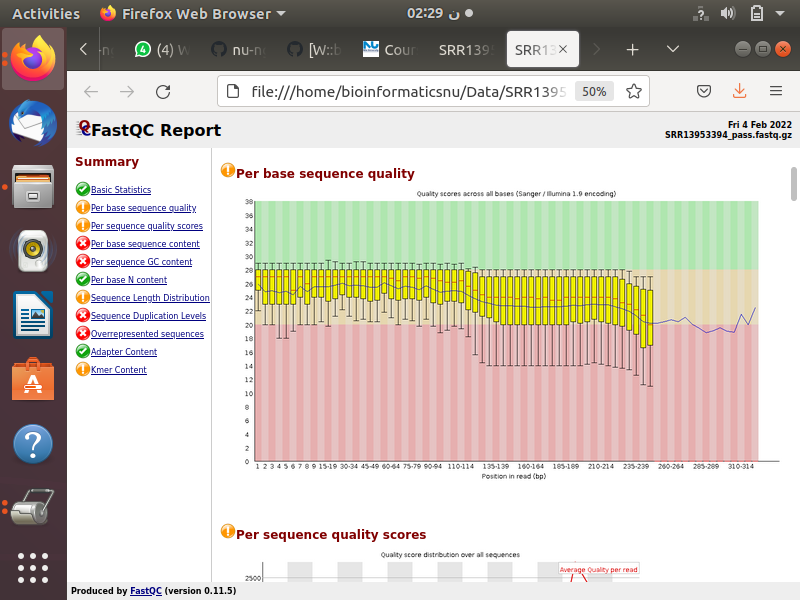
# Results and discussion

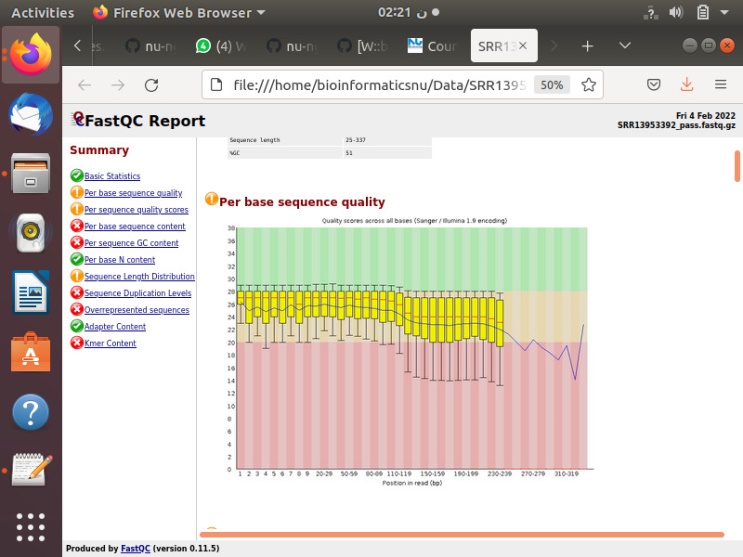
## Quality Results

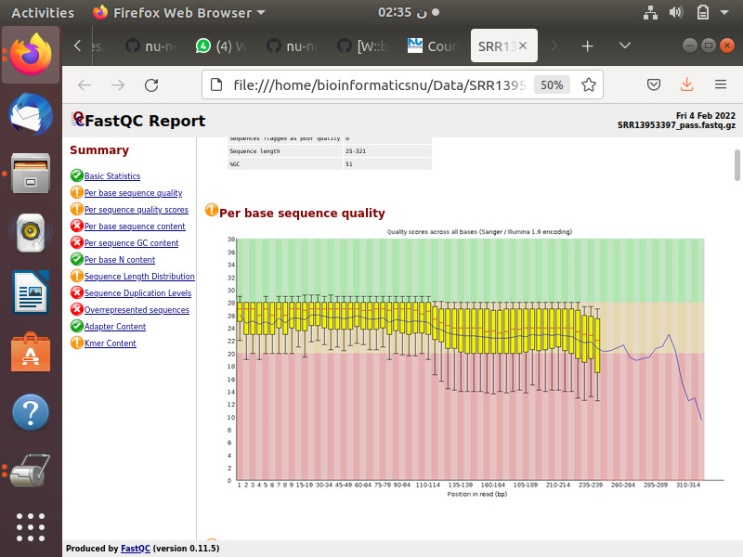
It's clear that the quality score ranges between 20 -28 in most of our data which is moderate quality score.



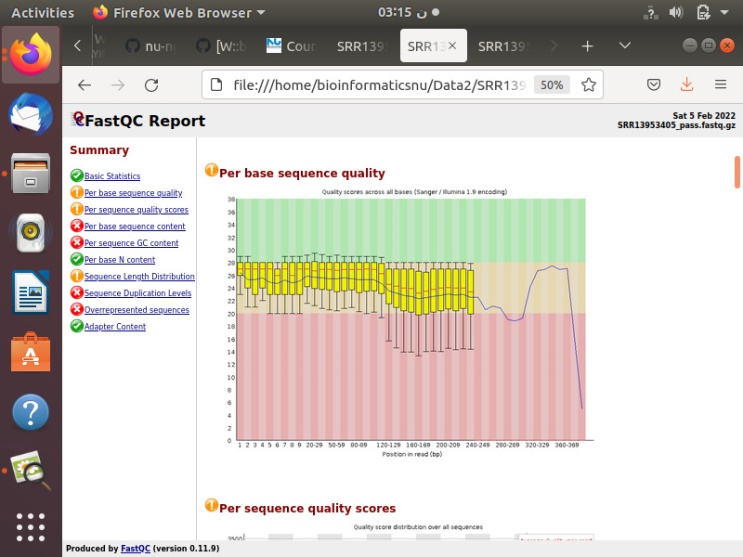


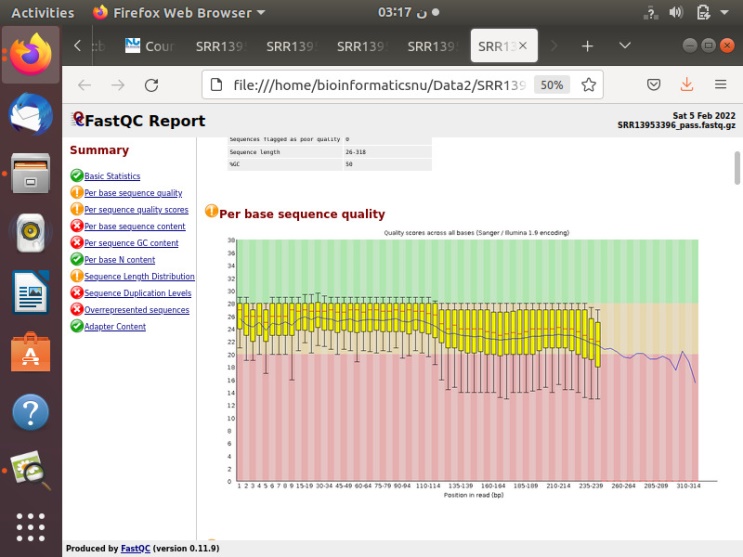


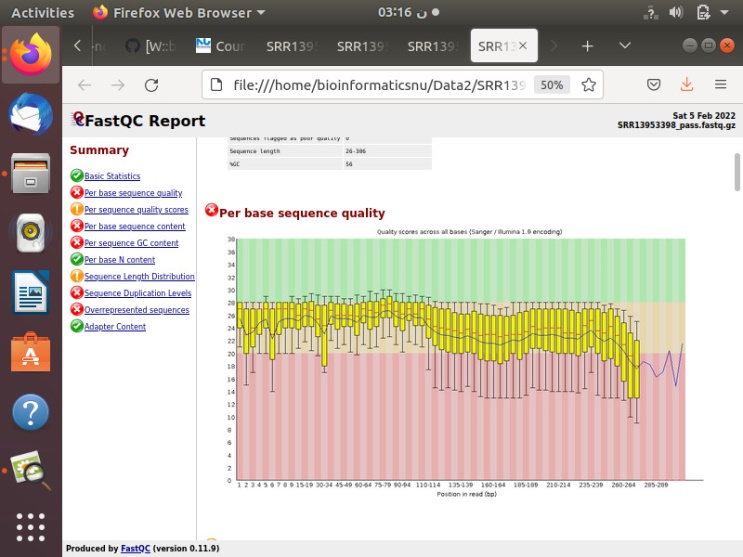




In order to improve this quality, we have tried to shuffle the data but unfortunately the data didn't change. We have tried to pick up the reads from the middle of our data and get its quality score seems to get worse as the majority lies between (20-29) and some are below 20.



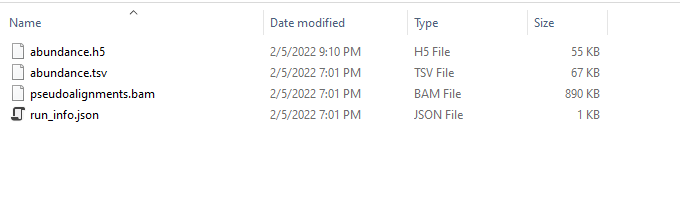




## Kallisto Results

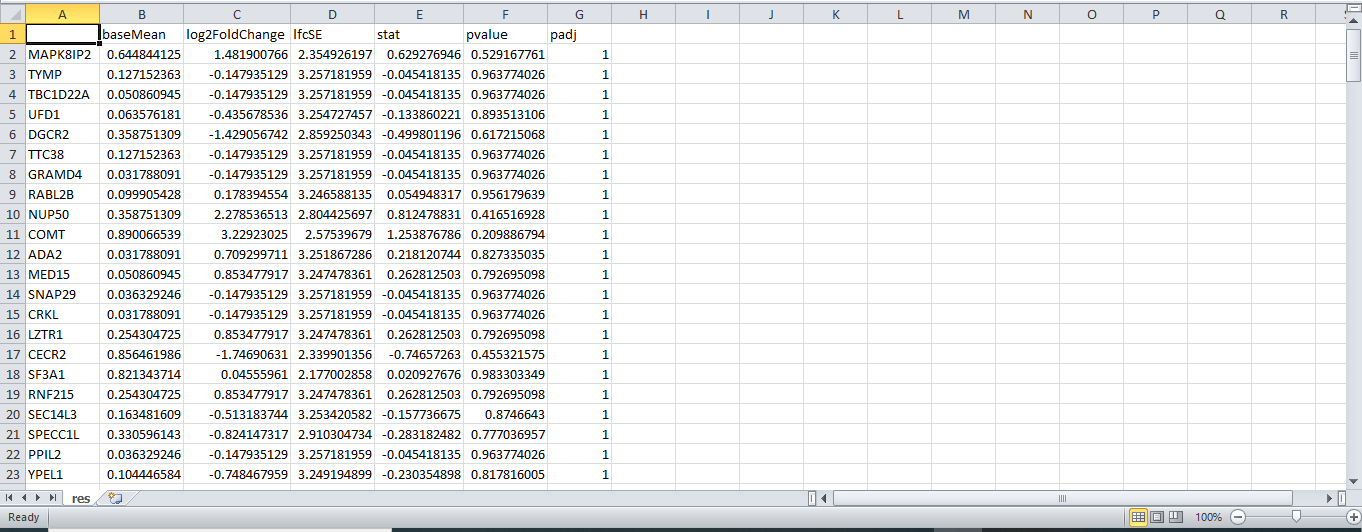
* We have got 15 output files for each read, each output file consists of 4 files, which are:

abundance. H5, abundance.tsv, run\_info.json , pseudoalignment.bam and abundance. h5 will be used in Deseq2.



## Deseq2 output explanation

From deseq2 steps, the res.csv file is uploaded; we have six columns of information reported for each gene: the names of genes, baseMean, log2FoldChange, IfcSE, stat, pvalue, padj.



We mainly concentrated on log2FoldChange, pvalue, padj for each gene:

a) log2FoldChange: is the effect size estimate. It tells us how much the gene's expression seems to have changed due to treatment with treated samples in comparison to untreated samples, +ve values indicate to upregulated, -ve values indicate to down regulated genes.

b) Pvalue: If the p-value is small we reject the null hypothesis and state that there is evidence against the null (i.e. the gene is differentially expressed), Calculating a p-value involves calculating a test statistic from the data and then comparing it to the distribution of that statistic under the null hypothesis, therefore, if we have, 10,000 genes, then you expect that the p<0.05 or p=0.1 just by chance that make that significant, but in our data we have more than 0.1, it ranges between 0.5&0.9

c) Padj: DESeq2 adjust the p value from wald test using Benjamini method (BH-adjusted p values), which is presented in the column of padj in the res.csv file, but it gives 1 for all genes this means that our data is not good enough.

Note:

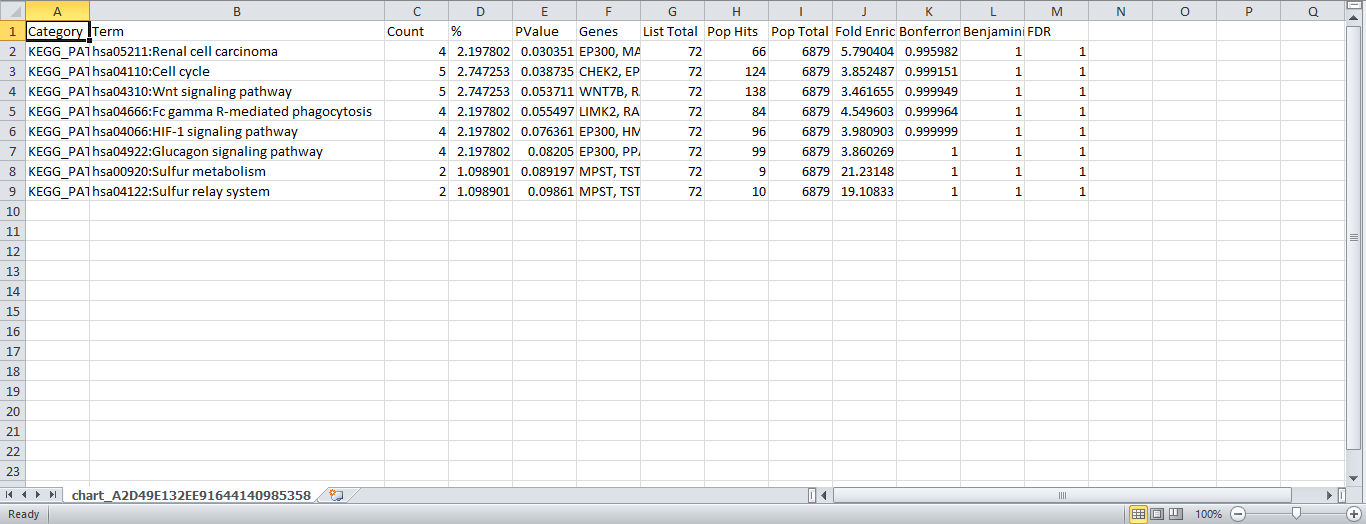
- To solve these problems in pvalue and padj we have 2 options:

a) Take more than 10,000 reads (to 25,000) to be more significant.

b) If also give us the same values after we enlarge the number of reads, we can make trimming on the fastQC to enhance the quality then make deseq2 steps.

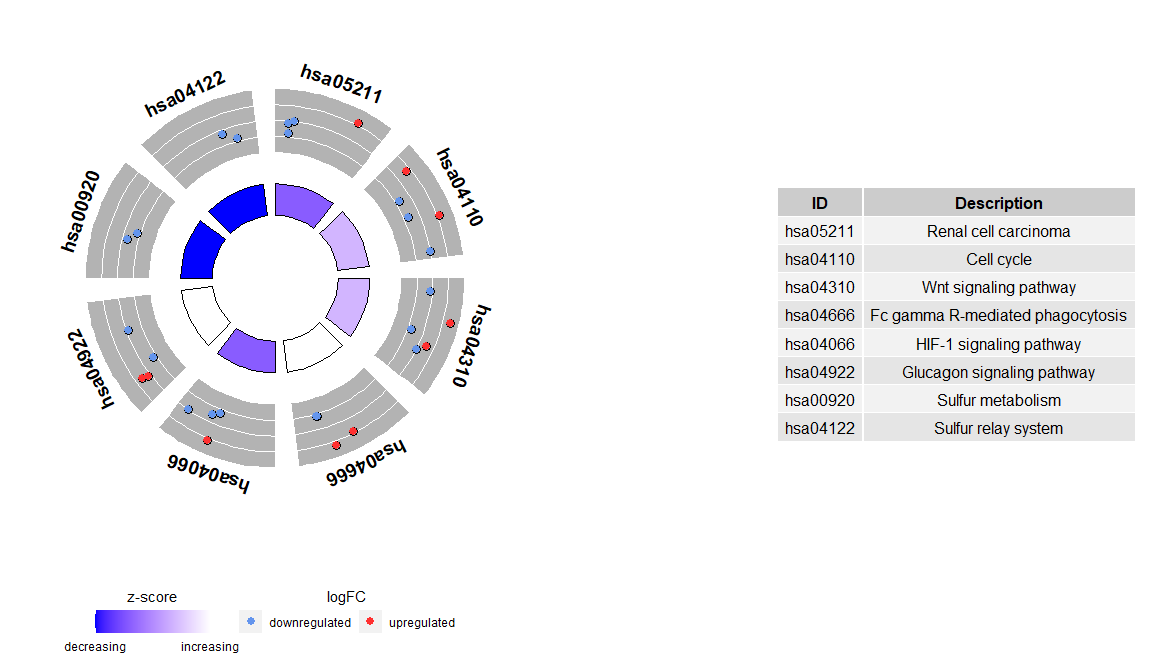
## David output explanation

* From pathways we selected pathaway\_KEGG, the chart file is uploaded, we have 13 columns of information reported for each KEGG\_PATHWAY: term, count, %, Pvalue, Genes, List Total, Pop Hits, Pop Total, Fold enrichment, Bonferron, Benjamin, FDR.

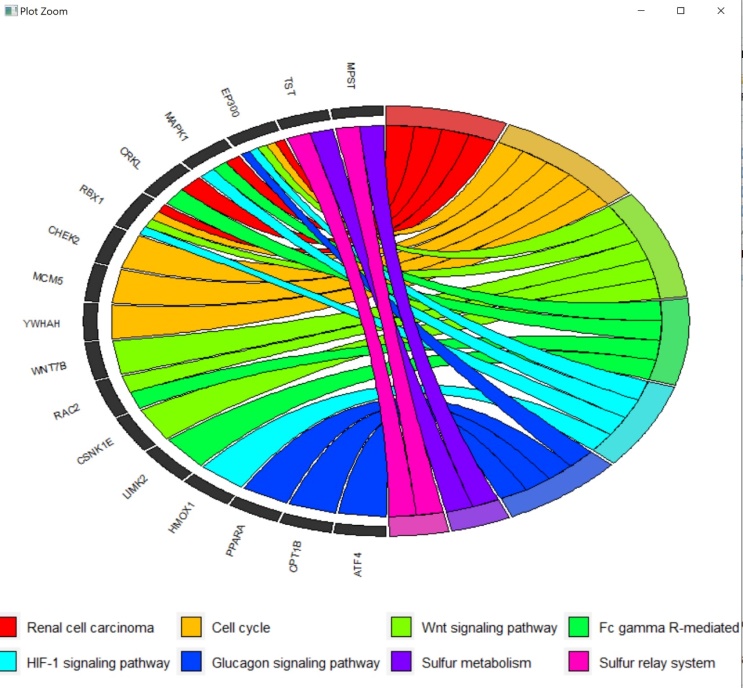


We mainly concentrated on term of pathway, count of genes that participate in this pathway, the names of these genes, Pvalue, Bonferron, Benjamin, and FDR.

## GOplot Results



This Fig shows that the outer circle shows a scatter plot for each term of the logFC of the assigned genes. Red circles display up- regulation and blue ones down- regulation by default. Therefore, it is easier to understand, why in some cases highly significant terms have a z-score close to zero.



This plot indicates the space argument defines the space between the colored rectangles representing the logFC. Also the font size of the gene labels (gene.size) and the space (gene.space) between them was changed. The genes were ordered according to their logFC values setting gene.order to ‘logFC’.



This heatmap indicates to genes are grouped together based on their expression patterns, thus clusters are likely to contain sets of co-regulated or functionally related genes.

## Interpretation for the our analysis

We pick 3 pathways and see how these pathways effect on Al Alzheimer disease:

a) Hypoxia Inducible Factor-1 as a Target for Neurodegenerative Diseases.

In this paper, we found that, the activity of HIF-1 is controlled by post-translational modifications on different amino acid residues of its subunits, mainly the alpha subunit. Besides in ischemic stroke, emerging evidence has revealed that HIF-1 activity and expression of its down-stream genes, such as vascular endothelial growth factor and erythropoietin, are altered in a range of neurodegenerative diseases. At the same time, experimental and clinical evidence has demonstrated that regulating HIF-1 might ameliorate the cellular and tissue damage in the neurodegenerative diseases [1].

b) Alzheimer’s and Hyperglycemia: Role of the Insulin Signaling Pathway and GSK-3 Inhibition in Paving a Path to Dementia

In this paper, we found that, there is a relation between Al Alzheimer Disease and the insulin signaling pathway. Insulin signaling pathway impairment leads to hyperphosphorylation of Tau protein, which plays a vital role in Al Alzheimer Disease pathology. The effect of insulin on cognition is bidirectional; the intranasal route of insulin showed to have a promising effect on cognition improvement. Subcutaneous and intravenous insulin can increase the risk of dementia. Further studies are encouraged to use a specific anti-diabetic medication that can reduce the progression of Al Alzheimer Disease [2].

c) Dysregulated Fc gamma receptor-mediated phagocytosis pathway in Alzheimer's disease: network-based gene expression analysis

In this paper we found that, gene-based association analysis identified PRKCD in the Fc gamma receptor-mediated phagocytosis pathway as being significantly associated with cognitive function and cerebrospinal fluid biomarkers. The identification of the Fc gamma receptor-mediated phagocytosis pathway implicates the peripheral innate immune system in the pathophysiology of AD. PRKCD is known to be related to neurodegeneration induced by amyloid-β [3].

##### References

1. Zhang, Z., Yan, J., Chang, Y., ShiDu Yan, S., & Shi, H. (2011). Hypoxia inducible factor-1 as a target for neurodegenerative diseases. Current medicinal chemistry, 18(28), 4335–4343. <https://doi.org/10.2174/092986711797200426>
2. Aljanabi, N. M., Mamtani, S., Al-Ghuraibawi, M., Yadav, S., & Nasr, L. (2020). Alzheimer's and Hyperglycemia: Role of the Insulin Signaling Pathway and GSK-3 Inhibition in Paving a Path to Dementia. Cureus, 12(2), e6885. <https://doi.org/10.7759/cureus.6885>
3. Park, Y. H., Hodges, A., Risacher, S. L., Lin, K., Jang, J. W., Ahn, S., Kim, S., Lovestone, S., Simmons, A., Weiner, M. W., Saykin, A. J., Nho, K., & AddNeuroMed consortium and the Alzheimer's Disease Neuroimaging Initiative (2020). Dysregulated Fc gamma receptor-mediated phagocytosis pathway in Alzheimer's disease: network-based gene expression analysis. Neurobiology of aging, 88, 24–32. <https://doi.org/10.1016/j.neurobiolaging.2019.12.001>
4. “Alzheimer's disease,” Mayo Clinic, 26-Jun-2021. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447#:~:text=Alzheimer's%20disease%20is%20a%20progressive,person's%20ability%20to%20function%20independently. [Accessed: 07-Feb-2022].
5. “Gene ontology resource,” Gene Ontology Resource. [Online]. Available: http://geneontology.org/. [Accessed: 07-Feb-2022].
6. National Center for Biotechnology Information. [Online]. Available: https://www.ncbi.nlm.nih.gov/. [Accessed: 07-Feb-2022].

Embl-Ebi, “Differential gene expression analysis,” Differential gene expression analysis | Functional genomics II. [Online]. Available: https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/rna-sequencing/performing-a-rna-seq-experiment/data-analysis/differential-gene-expression-analysis/#:~:text=Differential%20expression%20analysis%20means%20taking,expression%20levels%20between%20experimental%20groups. [Accessed: 07-Feb-2022].