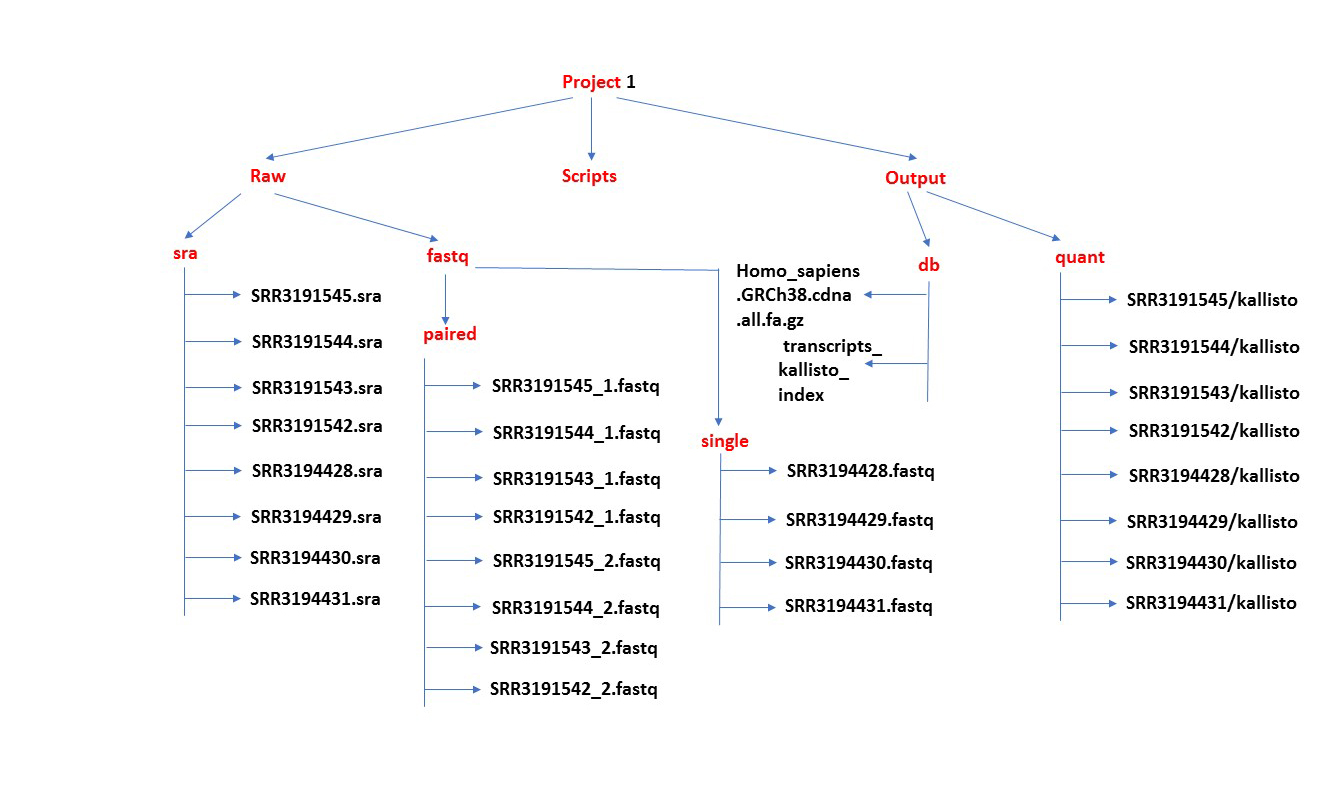
**Bigyan Rimal**

**Research Project**

1. Sample Access



1. **Super Computer Submission Script**

quanah:/lustre/scratch/brimal/project1/raw/sra$ nano wget\_forloop.sh

GNU nano 2.3.1 File: wget\_forloop.sh

#!/bin/bash

#$ -V

#$ -cwd

#$ -S /bin/bash

#$ -N quals

#$ -o $JOB\_NAME.o$JOB\_ID

#$ -e $JOB\_NAME.e$JOB\_ID

#$ -q omni

#$ -pe fill 3

#$ -P quanah

for i in SRR3191542 SRR3191543 SRR3191544 SRR3191545 SRR3194428 SRR3194429 SRR3194430 SRR3194431

do

wget [ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP070/SRP070895/${i}/${i}.sra](ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByStudy/sra/SRP/SRP070/SRP070895/$%7bi%7d/$%7bi%7d.sra)

quanah: qsub wget\_forloop.sh

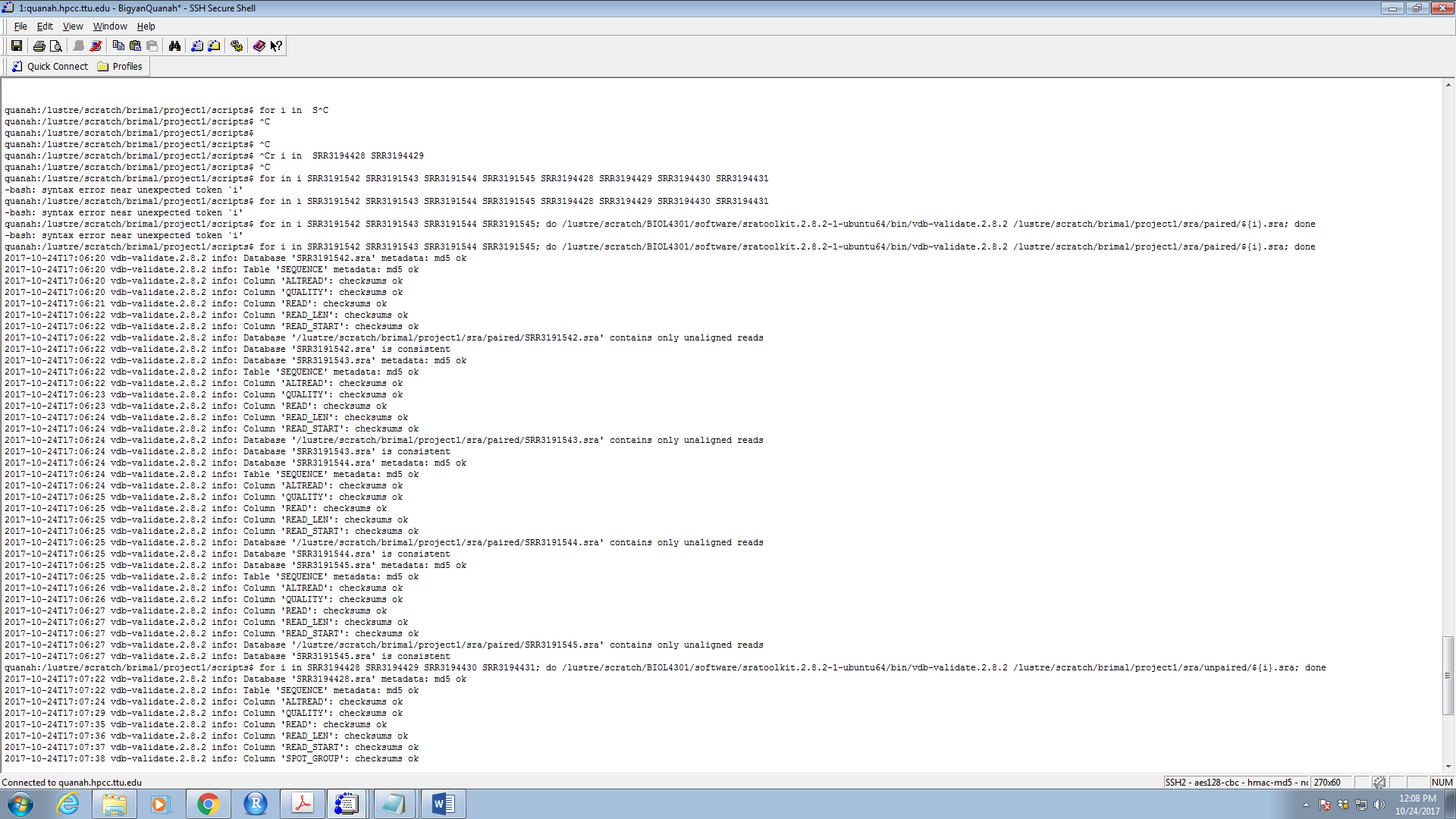
**3. Loop to run certain required features in all the samples**

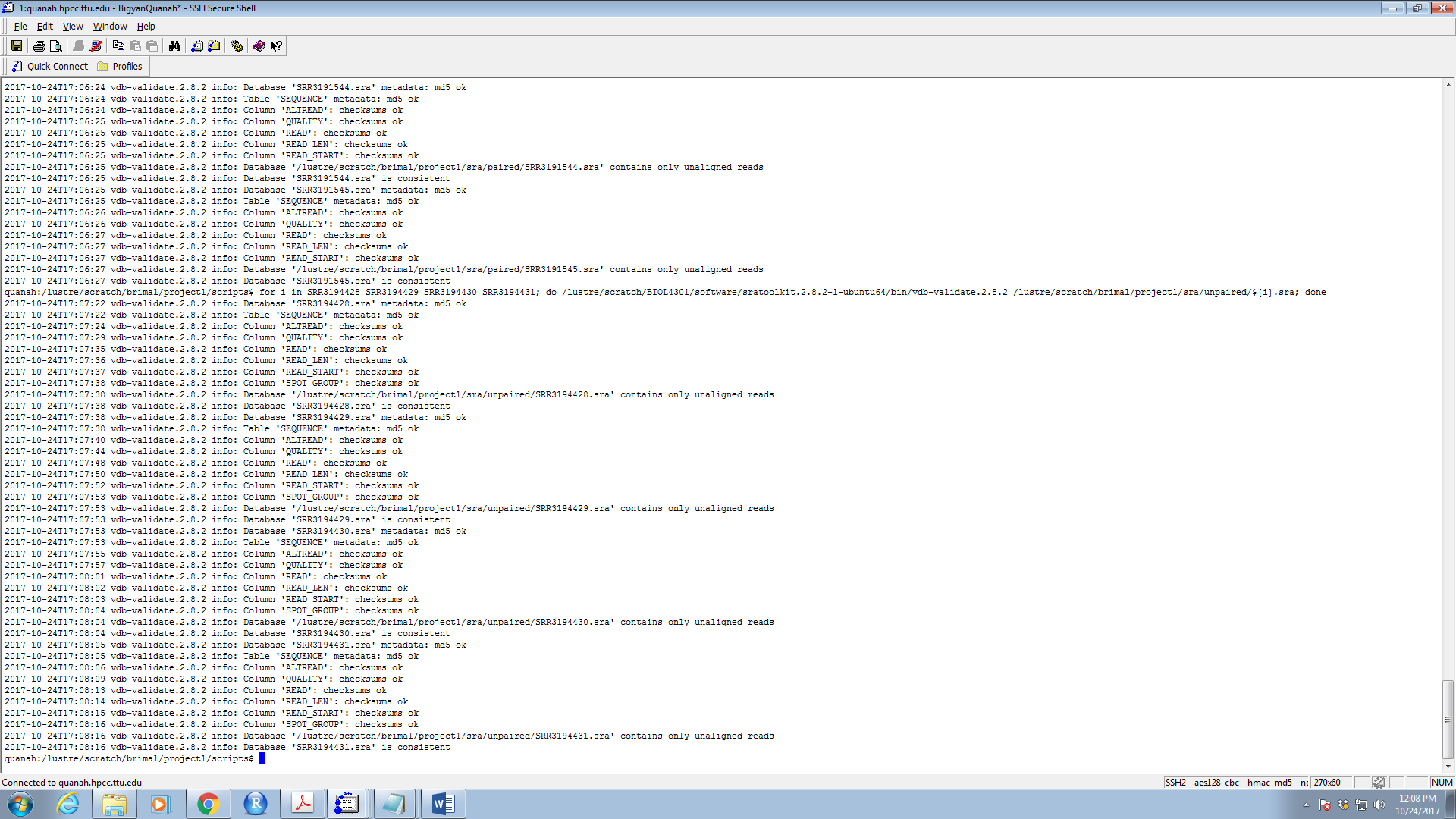
for i in SRR3191542 SRR3191543 SRR3191544 SRR3191545 SRR3194428 SRR3194429 SRR3194430 SRR3194431

do

/lustre/scratch/BIOL4301/software/sratoolkit.2.8.2-1-ubuntu64/bin/vdb-validate.2.8.2 /lustre/scratch/brimal/project1/raw/sra${i}.sra;

done





**4.**

**For Single End Reads**

cd /lustre/scratch/brimal/project1/raw/sra/single

for i in SRR3194428 SRR3194429 SRR3194430 SRR3194431

do

/lustre/scratch/BIOL4301/software/sratoolkit.2.8.2-1-ubuntu64/bin/fastq-dump.2.8.2 ./${i}.sra

done

**For Paired End Reads**

cd /lustre/scratch/brimal/project1/raw/sra/paired

for i in SRR3191545 SRR3191544 SRR3191543 SRR3191542

do

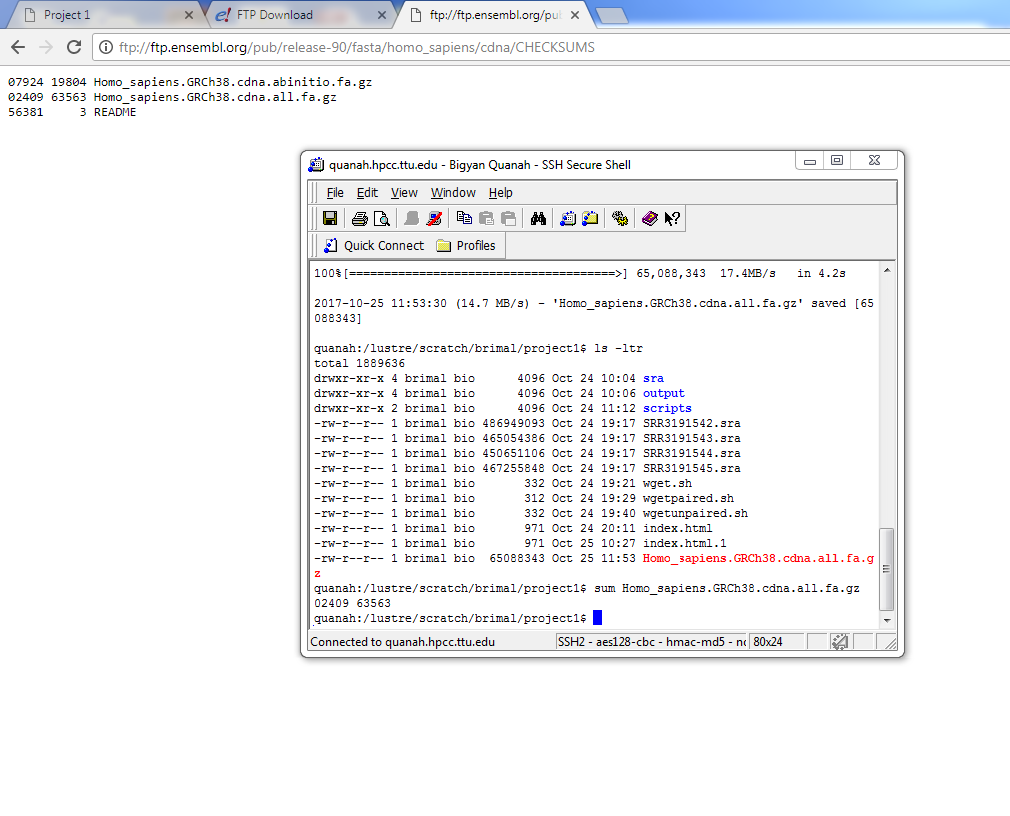
/lustre/scratch/BIOL4301/software/sratoolkit.2.8.2-1-ubuntu64/bin/fastq-dump.2.8.2 --splitfile ./${i}.sra

done

**5.**

quanha: /lustre/scratch/brimal/project1/$ wget <ftp://ftp.ensembl.org/pub/release-90/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz>

quanha: /lustre/scratch/brimal/project1/$ sum Homo\_sapiens.GRCh38.cdna.all.fa.gz



**6.**

quanah:/lustre/scratch/brimal/project1/output/db

/lustre/scratch/BIOL4301/software/kallisto\_linux-v0.43.1/kallisto index -i "transcripts\_kallisto\_index" ./Homo\_sapiens.GRCh38.cdna.all.fa.gz

**7.**

**For Paired End reads**

quanah:/lustre/scratch/brimal/project1/output/db

for i in SRR3191542 SRR3191543 SRR3191544 SRR3191545

do

/lustre/scratch/BIOL4301/software/kallisto\_linux-v0.43.1/kallisto quant -i

/lustre/scratch/brimal/project1/output/db/transcripts\_kallisto\_index –b 100 –o ../quant/${i}/kallisto

**For single end reads**

quanah:/lustre/scratch/brimal/project1/output/db

for i in SRR3194428 SRR3194429 SRR3194430 SRR3194431

do

/lustre/scratch/BIOL4301/software/kallisto\_linux-v0.43.1/kallisto quant -i

/lustre/scratch/brimal/project1/output/db/transcripts\_kallisto\_index --single –b 100 –l 187 –s 70 –o ../quant/${i}/kallisto

**8.**

**Effective Length**

A genomic region with a sequence that can normally appear in RNASeq experiment, which could be a gene or an exon or an isoform, could have multiple start points which could generate multiple fragments of a particular length, which is also referred as Effective length. It ( is actually calculated with the equation below, where, \mu_{FLD} is the mean of the fragment length distribution obtained from the reads that were aligned and is length

\widetilde{l}_i = l_i - \mu_{FLD} + 1

This is important factor to consider because this helps eliminate the bias during reads abundance estimation.

**Estimated Counts**

Counts refers to the number of reads aligning to a particular sequence, denoted as a random variable . are generally expressed as E[], estimated using expectation maximization algorithm, because of the alternative splicing of the genes and are also known as estimated counts. They are important to include the gene expression in all forms i.e. regular exons, alternative splicing etc.

**TPM**

Transcripts per million is the measurement of proportion of transcripts present in the experimental RNA pool. It is the very nice representation of if we are looking for transcript abundance. If we were to sequence a million full length transcripts, TPM is the number of transcript we see of a particular type, for abundance of other transcripts in the experimental sample is known. This parameter is sample dependent, therefore could be one of the most stable unit among experiments.

**9.**

setwd("C:/Users/brimal/Desktop")

install.packages("ggplot2")

install.packages("dplyr")

install.packages("data.table", type="source")

library(mgcv)

source("[http://bioconductor.org/biocLite.R"](http://bioconductor.org/biocLite.R%22))

biocLite("devtools")    # only if devtools not yet installed

biocLite("pachterlab/sleuth")

library('sleuth')

source("[http://bioconductor.org/biocLite.R"](http://bioconductor.org/biocLite.R%22))

biocLite("biomaRt")

help(package = 'sleuth')

base\_dir = "C:/Users/brimal/Desktop"

sample\_id <- dir(file.path(base\_dir,"quant"))

kal\_dirs <- sapply(sample\_id, function(id) file.path(base\_dir, "quant", id, "kallisto"))

kal\_dirs

s2c <- read.table(file.path(base\_dir, "hiseq\_info.txt"), header = TRUE, stringsAsFactors=FALSE)

s2c <- dplyr::select(s2c, sample = run\_accession, condition)

s2c

s2c <- dplyr::mutate(s2c, path = kal\_dirs)

print(s2c)

so <- sleuth\_prep(s2c, ~ condition)

so <- sleuth\_fit(so)

so <- sleuth\_fit(so, ~1, 'reduced')

so <- sleuth\_lrt(so, 'reduced', 'full')

models(so)

options(mc.cores = 4L)

mart <- biomaRt::useMart(biomart = "ENSEMBL\_MART\_ENSEMBL", dataset = "hsapiens\_gene\_ensembl", host = 'ensembl.org')

t2g <- biomaRt::getBM(attributes = c("ensembl\_transcript\_id", "ensembl\_gene\_id", "external\_gene\_name"), mart = mart)

t2g <- dplyr::rename(t2g, target\_id = ensembl\_transcript\_id, ens\_gene = ensembl\_gene\_id, ext\_gene = external\_gene\_name)

so <- sleuth\_prep(s2c, ~ condition, target\_mapping = t2g)

so <- sleuth\_fit(so)

so <- sleuth\_fit(so, ~1, 'reduced')

so <- sleuth\_lrt(so, 'reduced', 'full')

sleuth\_live(so)

results\_table <- sleuth\_results(so, 'reduced:full', test\_type = 'lrt')

plot\_pca(so, color\_by = 'condition')

plot\_group\_density(so, use\_filtered = TRUE, units = "est\_counts",trans = "log", grouping = setdiff(colnames(so$sample\_to\_covariates),"sample"), offset = 1)

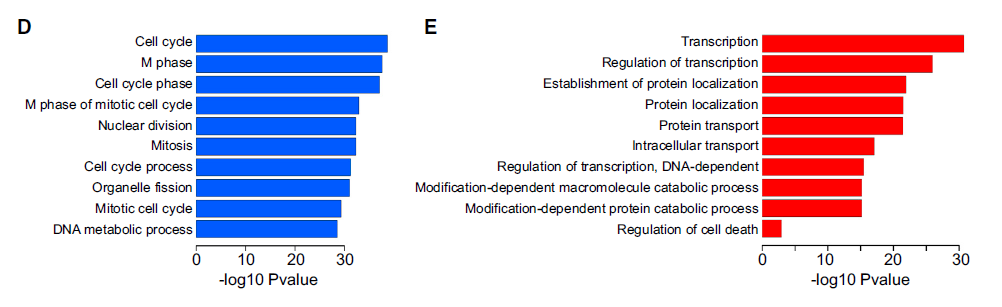
results\_significant <- dplyr::filter(results\_table, qval <= 0.05)

write.csv(results\_table, file = "results\_table.csv")

write.csv(results\_significant, file = "sigs.csv")

**10.**

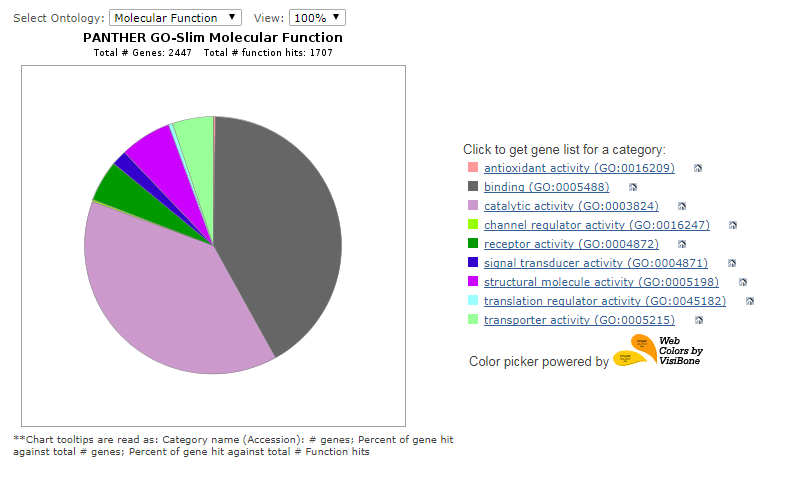
Calvet et. al. 2016 found Zika Xiruns (ZIKV) in amniotic fluid of two pregnant women, whose fetuses were diagnosed with microcephaly. This finding suggested ZIKV can cross the placental barrier as well. Their finding opened a new avenue of ZIKV research and its connection with neural stem-cell development. Mlakar et. al. proposed that neural cells are affected by ZIKV and that led to the arrest of development of the cerebral cortex. Tang et. al. followed on these works and reported how ZIKV infects human nural progenitor cells (hNPCs) derived from stem cells. They found that ZIKV further releases the ZIKV particle after infections increasing the cell death, dysregulation of cell-cycle pathways and attenuated hPNC growth. They were able to establish a tractable experimental model syste, to investigate the impact and mechanism of ZIKV on human brain development. They also found that ZIKV infection led to significant increase in caspase-3 (apoptotic enzyme) activation in hNPCs compared to mock infection suggesting ZIKV leds to programmed neural cell death. As shown in Figure 1. Tang et. al. also presented the data to support their hypotheses on ZIKV’s effect on cell-cycle processes and cell death.



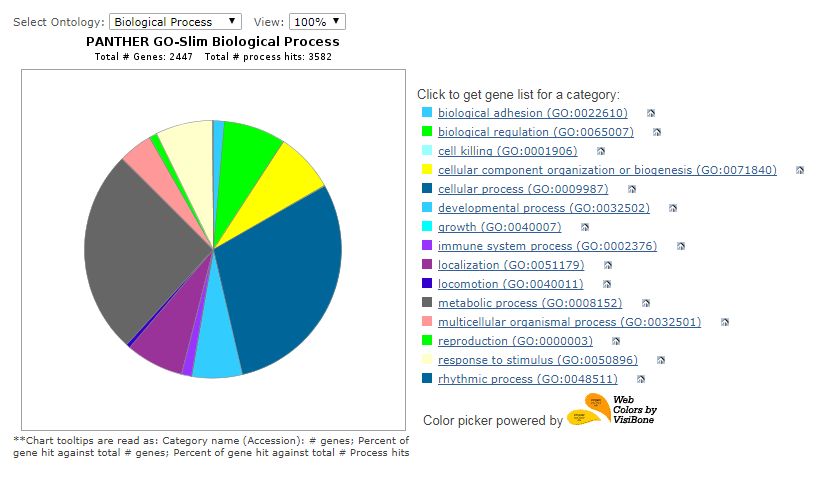
Yi. et. al. followed on this work and came up with more number of differentially expressed genes only by using a different, more intensive and fast gene expression analysis methods including assembly tool ,kallisto and statistical analysis and visualization tool, sleuth. They found some 751 more differentially expressed gene that were not found by cufflink assembly, technique used by Tang et. al and elucidated how ZIKV infection could affect hNPCs transcriptome, refining pathway predictions and isoform-specific dynamics.

I found 2869 significant differentially expressed gene from my sleuth analysis and used Panther to do the functional annotation on those genes. Five different gene ontology parameters provided following set of functional annotation on those genes.

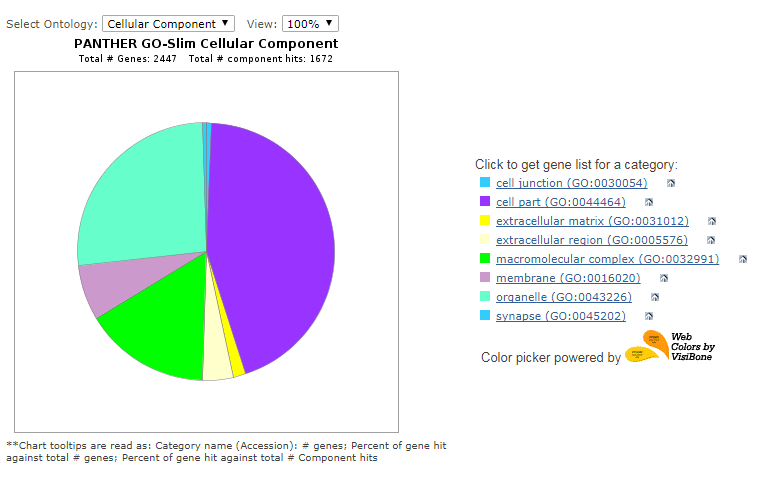
1. Gene Ontology based on Molecular functions



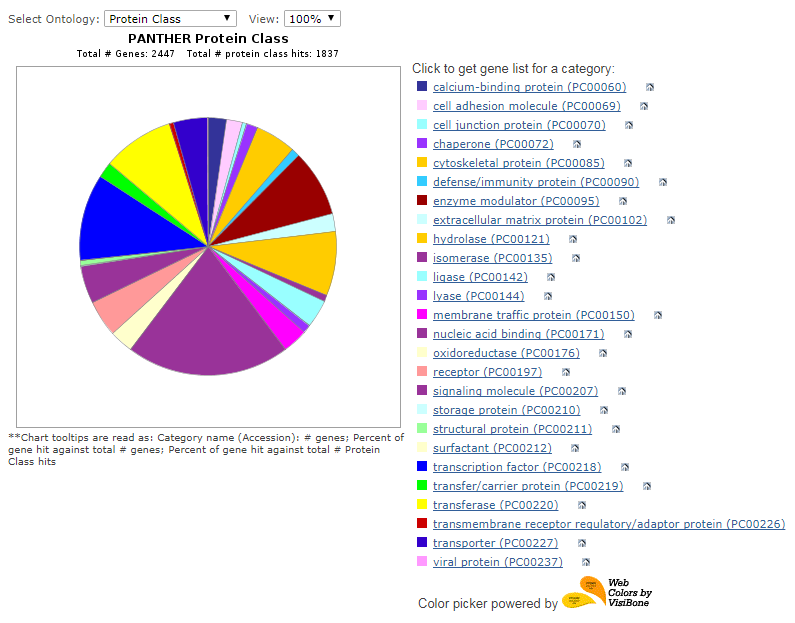
1. Gene ontology based on Biological Processes



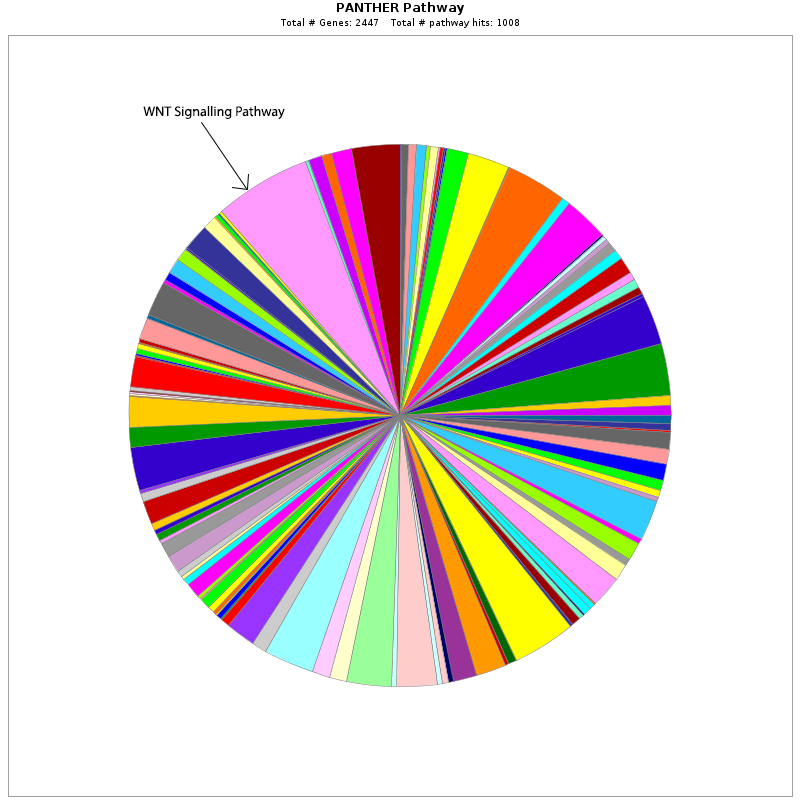
1. Gene Ontology based on Cellular Component



1. Gene ontology based on Protein Class



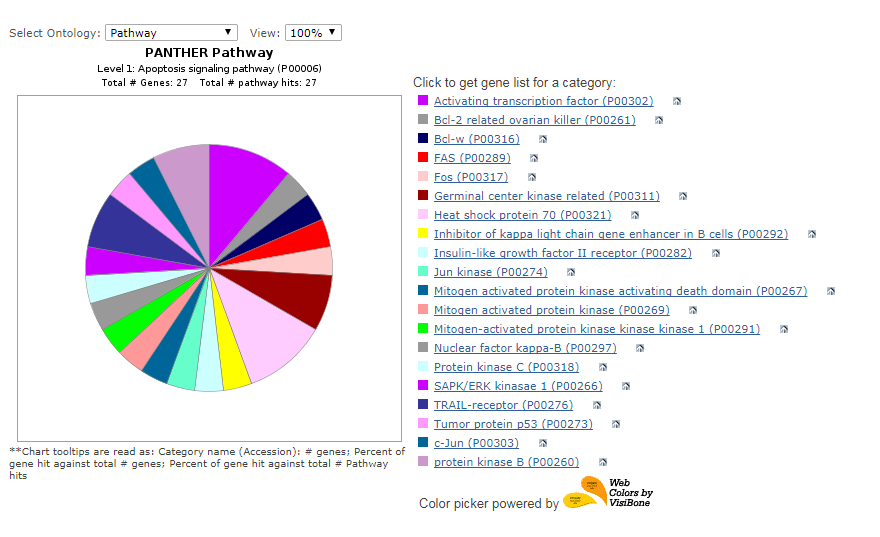
1. Gene ontology based on Pathways

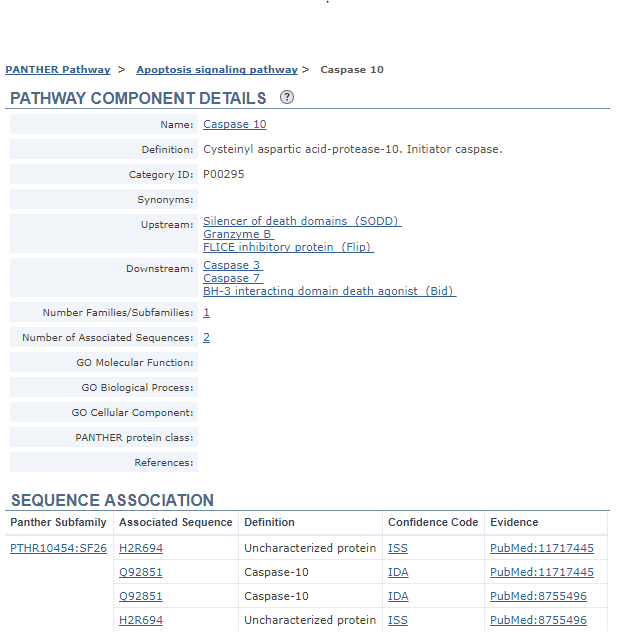


Apoptosis signaling pathway

Most of these annotations correlates with the data presented by Tang et. al. as shown in Figure 1. The cellular process shown in figure 1. are also vividly seen in the Panther annotation from my analysis. However, my interest was mainly attracted by the Apoptosis Signalling Pathway and the higher expression of Caspase enzyme system in hNPCs. Following diagram shows what genes are mainly involved in apoptosis pathway in hNPCs and how caspases are increasingly expressed.

Wnt Signalling pathway incudes highest number of sequences that were differentially expressed in ZIKV infected hNPCs. Interestingly, Nusse et. al., 2008, found that Wnts are the cadidates for neural stem cells’ self-renewal and major factors blocking their differentiation. Also, their study found that Wnt knockout mouse lost the midbrain, further, supporting role of Wnts on brain development. They also found that Wnt3a mutant mouse exhibited underdevelopment of hippocampus. To support this finding they performed in vivo exogenous expression of Wnt3a and found enhanced neurogenesis, which helped them come up with a conclusion that Wnt signaling pathway is major regulator of neurogenesis in the hippocampus.





Above figures show that the TRAIL-receptor pathway, which is the initiator pathway for programmed cell death and Caspase-10 (Initiator caspase), very important enzyme to initiate the caspase cascade for apoptosis includes 7.4% of total sequences differentially expressed relating to Apoptotic pathway.

All these results leads to the conclusion that ZIKV infection mainly targets hNPCs and affect various metabolic pathways and cellular process ranging from cell development to cell death.

**References**

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Nusse, R., Fuerer, C., Ching, W., Harnish, K., Logan, C., Zeng, A., … Kalani, Y. (2008). Wnt signaling and stem cell control. *Cold Spring Harbor Symposia on Quantitative Biology*, *73*, 59–66. https://doi.org/10.1101/sqb.2008.73.035

Tang, H., Hammack, C., Ogden, S. C., Wen, Z., Qian, X., Li, Y., … Ming, G. L. (2016). Zika virus infects human cortical neural progenitors and attenuates their growth. *Cell Stem Cell*, *18*(5), 587–590. https://doi.org/10.1016/j.stem.2016.02.016

The farrago. (2017). *What the FPKM? A review of RNA-Seq expression units*. [online] Available at: https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/ [Accessed 2 Nov. 2017].

Yi, L., Pimentel, H., & Pachter, L. (2017). Zika infection of neural progenitor cells perturbs transcription in neurodevelopmental pathways. *PLoS ONE*, *12*(4), 1–11. https://doi.org/10.1371/journal.pone.0175744