Analysis Report (Part3): Counting the Reads of Genes

(Last updated: 12/05/21)

1. Quantifying the Abundance of Genes

In order to quantify the abundance of genes in each sample, the number of reads corresponding to each gene in the genome in that sample is counted using **HTSeq.0.11.1**. The file *Homo_sapiens.GRCh38.103.gtf* provides hg38 genome annotations used in the counting. Gene annotations describe the structure of transcripts (model) expressed from those gene loci. A transcript model consists of the coordinates of the exons of a transcript on a reference genome. The following Bash script was employed:

```bash
#!/bin/bash

ANNOT=/data/Homo_sapiens.GRCh38.103.gtf

# FETAL SAMPLES

printf "\nn***** Assembling the transcripts for run SRR2071348 ...\n\n"
htseq-count -f bam -s no -m union --nonunique all -r name  /data/align/SRR2071348.bam \
  $ANNOT > htseq/SRR2071348_counts.txt

printf "\nn***** Assembling the transcripts for run SRR2071349 ...\n\n"
htseq-count -f bam -s no -m union --nonunique all -r name  /data/align/SRR2071349.bam \
  $ANNOT > htseq/SRR2071349_counts.txt

printf "\nn***** Assembling the transcripts for run SRR2071352 ...\n\n"
htseq-count -f bam -s no -m union --nonunique all -r name  /data/align/SRR2071352.bam \
  $ANNOT > htseq/SRR2071352_counts.txt

# ADULT SAMPLES

printf "\nn***** Assembling the transcripts for run SRR2071346 ...\n\n"
htseq-count -f bam -s no -m union --nonunique all -r name  /data/align/SRR2071346.bam \
  $ANNOT > htseq/SRR2071346_counts.txt

printf "\nn***** Assembling the transcripts for run SRR2071347  ...\n\n"
htseq-count -f bam -s no -m union --nonunique all -r name  /data/align/SRR2071347.bam \
  $ANNOT > htseq/SRR2071347_counts.txt

printf "\nn***** Assembling the transcripts for run SRR2071350 ...\n\n"
htseq-count -f bam -s no -m union --nonunique all -r name  /data/align/SRR2071350.bam \
  $ANNOT > htseq/SRR2071350_counts.txt
```

After saving the above script as *htseq-count.sh* in *bash-scripts* directory, the following Bash command was used for running it:

```
nohup sh bash-scripts/htseq-count.sh > htseq-count.out &
```

2. Merging the Count Files

The following R script was employed to merge the counts files into one file *merged_counts-v2.txt*:

```
# Set the working directory
setwd("~/brain/brain-zip/htseq-v2")

# Load fetal samples
SRX683795 <- read.table("SRR2071348_counts.txt", header = FALSE)
SRX683796 <- read.table("SRR2071349_counts.txt", header = FALSE)
SRX683799 <- read.table("SRR2071352_counts.txt", header = FALSE)

# Load adult samples
SRX683793 <- read.table("SRR2071346_counts.txt", header = FALSE)
SRX683794 <- read.table("SRR2071347_counts.txt", header = FALSE)
SRX683797 <- read.table("SRR2071350_counts.txt", header = FALSE)

# check to see if all elements of the first column (transcripts) are
# the same across all 6 samples
all(SRX683797[,1] == SRX683794[,1])
all(SRX683793[,1] == SRX683794[,1])
all(SRX683799[,1] == SRX683793[,1])
all(SRX683797[,1] == SRX683794[,1])
all(SRX683795[,1] == SRX683797[,1])

# create a merged_counts table
merged_counts <- data.frame(row.names = SRX683795[,1], SRX683795 = SRX683795[,2],
                            SRX683796 = SRX683796[,2], SRX683799 = SRX683799[,2],
                            SRX683793 = SRX683793[,2], SRX683794 = SRX683794[,2],
                            SRX683797 = SRX683797[,2])
write.table(merged_counts, "../../merged_counts-v2.tsv", quote = FALSE, sep = '\t')
```