Coding for Align data against human genome (Tophat2) and transcriptome (kallisto):
#Everything is done in a directory named HW2, path: /gpfs/scratch/as18818/Practicum5/HW2

copying Kallisto bash script to the working directory:
cp /gpfs/scratch/as18818/Practicum5/HW/Kallisto-HW.sh /gpfs/scratch/as18818/Practicum5/HW2
#Run an interactive session:
srun -c1 -t12:00:00 --mem=16000 --pty /bin/bash

#edit the Kallisto script, change email, cpu_medium, directories path etc:

#!/bin/bash
#SBATCH --job-name=Kallisto # Job name
#SBATCH --mail-type=END,FAIL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=Apoorva.Sharma@nyulangone.org # Where to send mail
#SBATCH --ntasks=4 # Run on a single CPU
#SBATCH --mem=32gb # Job memory request
#SBATCH --time=12:00:00 # Time limit hrs:min:sec
#SBATCH -p cpu_medium

#make sure you cd into your directory and change all paths to match your own paths
cd /gpfs/scratch/as18818/Practicum5/HW2

#this allows us to read in the filename prefix from the file list.txt and subtitute it anywhere you see ${sample} and run array jobs,
sample=$(awk "NR==${SLURM_ARRAY_TASK_ID} {print \$1}" /gpfs/data/courses/bminga3004/2023/Practicum5/Assignment/list2.txt)

module load kallisto/0.44.0

Indexing needed for Kallisto, point to the one in course directory or create your own
#kallisto index -i /gpfs/data/courses/bminga3004/2023/Practicum5/HomoSapiens /gpfs/data/courses/bminga3004/2023/Practicum5/Homo_sapiens.GRCh38.cdna.all.fa

#run kallisto
kallisto quant -i /gpfs/data/courses/bminga3004/2023/Practicum5/HomoSapiens -o /gpfs/scratch/as18818/Practicum5/HW2/${sample}_kallisto -b 100 --bias /gpfs/data/courses/bminga3004/2023/Practicum5/Assignment/${sample}_R1.fastq.gz /gpfs/data/courses/bminga3004/2023/Practicum5/Assignment/${sample}_R2.fastq.gz

```
### Save changes and submit the job.
Sbatch –array=1-6 Kallisto-HW.sh

#Resulting files contains
abundance.h5 , abundance.tsv,  run_info.json for each sequence.
```

Generate transcript counts from transcriptome alignment + turn into gene counts:

```
### LOAD REQUIRED LIBRARIES but first install if they are not installed

library(biomaRt)
library(tximport)
library(rhdf5)

### SET WORKING DIRECTORY ### You will need to edit this and direct it your downloaded kallisto folder
setwd("C:/Users/Apoorva Sharma/Downloads/Practicum5")

### IMPORT ENSEMBl ANNOTATIONS FOR HUMAN GENOME & GENERATE TWO COLUMN FILE LINKING TRANSCRIPT AND GENE IDS
mart <- biomaRt::useMart(biomart = "ensembl", dataset =  "hsapiens_gene_ensembl")
t2g <- biomaRt::getBM(attributes = c("ensembl_transcript_id", "transcript_version", "ensembl_gene_id", "external_gene_name", "description", "transcript_biotype", "refseq_mrna", "refseq_ncrna"), mart = mart)
t2g$target_id <- paste(t2g$ensembl_transcript_id, t2g$transcript_version, sep=".") # append version number to the transcript ID
t2g[,c("ensembl_transcript_id","transcript_version")] <- list(NULL) # delete the ensembl transcript ID and transcript version columns
t2g <- dplyr::rename( t2g, gene_symbol = external_gene_name, full_name = description, biotype = transcript_biotype )
t2g<-t2g[,c(ncol(t2g),1:(ncol(t2g)-1))]

### GENERATE ADDITIONAL OBJECT CONTAINING ONLY PROTEIN CODING GENES
gb <- getBM(attributes=c("ensembl_gene_id","gene_biotype"), mart=mart)
gb_coding<-subset(gb, gb$gene_biotype=="protein_coding")
genes<-gb_coding$ensembl_gene_id

### USE TXIMPORT TO SUMMARIZE TRANSCRIPT COUNTS INTO GENE COUNTS
## For multiple samples, each named as a folder in the kallisto directory (can be abundance.h5 or abundance.tsv file)
#accessions <- list.dirs(full.names=FALSE)[-c(1:2)]
# Assuming your working directory is already set to the kallisto folder
# Get all directories
all_dirs <- list.dirs(path = ".", full.names = FALSE, recursive = FALSE)

# Filter directories based on a pattern (e.g., containing "kallisto")
accessions <- grep("kallisto", all_dirs, value = TRUE)

# Print the accessions to check if LT34_Kallisto is included
print(accessions)
kallisto.dir<-paste0(accessions)
```

```
kallisto.files<-file.path(kallisto.dir,"abundance.tsv") #can also be abundance.tsv
names(kallisto.files)<- accessions
tx.kallisto <- tximport(kallisto.files, type = 'kallisto', tx2gene = t2g, countsFromAbundance ="no")

### GENERATE TWO COLUMN OUTPUT FORMAT
counts<-as.data.frame(tx.kallisto$counts[row.names(tx.kallisto$counts) %in% genes, ])
len <- as.data.frame(tx.kallisto$len[row.names(tx.kallisto$len) %in% genes, ])
ids<-rownames(counts)

### ROUND VALUES (DESEQ2 DOES NOT LIKE FRACTIONS), AND WRITE TO OUTPUT FILE
write.table(round(counts),paste("output",".txt",sep=""), row.names=ids, quote=F, col.names=T, sep="\t")
```

For Tophat:

```
#!/bin/bash
#SBATCH --job-name=Tophat2 # Job name
#SBATCH --mail-type=END,FAIL # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=Apoorva.Sharma@nyulangone.org # Where to send mail
#SBATCH --ntasks=4
#SBATCH --mem=32gb # Job memory request
#SBATCH --time=24:00:00 # Time limit hrs:min:sec
#SBATCH -p cpu_medium


module load trimgalore/0.5.0
module load python/cpu/2.7.15-ES
module load samtools/1.3
module load tophat/2.1.1
module load bowtie2/2.3.4.1
module load subread/1.6.3
module load igenome

#make sure you cd into your directory and change all paths to match your own paths
cd /gpfs/scratch/as18818/Practicum5/HW2

#this allows us to read in the filename prefix from the file list.txt and subtitute it anywhere you see ${sample} and
run array jobs,
sample=$(awk "NR==${SLURM_ARRAY_TASK_ID} {print \$1}"
/gpfs/data/courses/bminga3004/2023/Practicum5/Assignment/list2.txt)



#First trim raw fastq files
trim_galore --paired --length 30 -o /gpfs/scratch/as18818/Practicum5/HW2
/gpfs/data/courses/bminga3004/2023/Practicum5/Assignment/${sample}_R1.fastq.gz
/gpfs/data/courses/bminga3004/2023/Practicum5/Assignment/${sample}_R2.fastq.gz
```

#Map using tophat2 (STAR aligner is associated with faster run times)

tophat2 -o /gpfs/scratch/as18818/Practicum5/HW2/${sample} -G
/gpfs/data/courses/bminga3004/2023/Practicum5/genes.gtf -p 8 --library-type fr-firststrand
$IGENOMES_ROOT/Homo_sapiens/UCSC/hg38/Sequence/Bowtie2Index/genome
/gpfs/scratch/as18818/Practicum5/HW2/${sample}_R1_val_1.fq.gz
/gpfs/scratch/as18818/Practicum5/HW2/${sample}_R2_val_2.fq.gz

samtools sort -o /gpfs/scratch/as18818/Practicum5/HW2/${sample}.sorted.bam ${sample}/accepted_hits.bam

samtools index /gpfs/scratch/as18818/Practicum5/HW2/${sample}.sorted.bam

featureCounts -s 2 -p -B -a /gpfs/data/courses/bminga3004/2023/Practicum5/genes.gtf -o
/gpfs/scratch/as18818/Practicum5/HW2/${sample}_FeatCount
/gpfs/scratch/as18818/Practicum5/HW2/${sample}.sorted.bam

#Generate gene counts from genome alignment using featurecounts

```
        ==========     _____ _    _ ____  _____  ______          _____
        =====         / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
         =====       | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
          ====        \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
           ====       ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
        ==========   |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
                v1.6.3

//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 1 BAM file                                       ||
||                           P LT36.sorted.bam                                ||
||                                                                            ||
||             Output file : LT36_FeatCount                                   ||
||                 Summary : LT36_FeatCount.summary                           ||
||              Annotation : genes.gtf (GTF)                                  ||
||      Dir for temp files : /gpfs/scratch/as18818/Practicum5/HW2             ||
||                                                                            ||
||                Threads : 1                                                 ||
||                  Level : meta-feature level                                ||
||              Paired-end : yes                                              ||
||      Multimapping reads : not counted                                      ||
|| Multi-overlapping reads : not counted                                      ||
||   Min overlapping bases : 1                                                ||
||                                                                            ||
||          Chimeric reads : counted                                          ||
||        Both ends mapped : required                                         ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//

//================================== Running =================================\\
||                                                                            ||
|| Load annotation file genes.gtf ...                                         ||
||    Features : 1271434                                                      ||
||    Meta-features : 66023                                                   ||
||    Chromosomes/contigs : 331                                               ||
||                                                                            ||
|| Process BAM file LT36.sorted.bam...                                        ||
||    Paired-end reads are included.                                          ||
||    Strand specific : reversely stranded                                    ||
||    Assign alignments (paired-end) to features...                           ||
||                                                                            ||
||    WARNING: reads from the same pair were found not adjacent to each       ||
||             other in the input (due to read sorting by location or         ||
||             reporting of multi-mapping read pairs).                        ||
||                                                                            ||
||    Pairing up the read pairs.                                              ||
||                                                                            ||
||    Total alignments : 28505558                                             ||
||    Successfully assigned alignments : 12507604 (43.9%)                     ||
||    Running time : 3.43 minutes                                             ||
||                                                                            ||
||                                                                            ||
|| Summary of counting results can be found in file "/gpfs/scratch/as18818/P  ||
|| racticum5/HW2/LT36_FeatCount.summary"                                      ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//
```

```
      ==========     _____ _    _ ____  _____  ______          _____
       =====        / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
         =====     | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
           ====     \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
             ====   ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
       ==========  |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
                 v1.6.3

//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 1 BAM file                                       ||
||                           P LT34.sorted.bam                                ||
||                                                                            ||
||             Output file : LT34_FeatCount                                   ||
||                 Summary : LT34_FeatCount.summary                           ||
||              Annotation : genes.gtf (GTF)                                  ||
||      Dir for temp files : /gpfs/scratch/as18818/Practicum5/HW2             ||
||                                                                            ||
||                 Threads : 1                                                ||
||                   Level : meta-feature level                               ||
||               Paired-end : yes                                             ||
||       Multimapping reads : not counted                                     ||
|| Multi-overlapping reads : not counted                                      ||
||   Min overlapping bases : 1                                                ||
||                                                                            ||
||           Chimeric reads : counted                                         ||
||         Both ends mapped : required                                        ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//

//================================== Running =================================\\
||                                                                            ||
|| Load annotation file genes.gtf ...                                         ||
||    Features : 1271434                                                      ||
||    Meta-features : 66023                                                   ||
||    Chromosomes/contigs : 331                                               ||
||                                                                            ||
|| Process BAM file LT34.sorted.bam...                                        ||
||    Paired-end reads are included.                                          ||
||    Strand specific : reversely stranded                                    ||
||    Assign alignments (paired-end) to features...                           ||
||                                                                            ||
||    WARNING: reads from the same pair were found not adjacent to each       ||
||             other in the input (due to read sorting by location or         ||
||             reporting of multi-mapping read pairs).                        ||
||                                                                            ||
||    Pairing up the read pairs.                                              ||
||                                                                            ||
||    Total alignments : 26829397                                             ||
||    Successfully assigned alignments : 10451387 (39.0%)                     ||
||    Running time : 3.37 minutes                                             ||
||                                                                            ||
||                                                                            ||
|| Summary of counting results can be found in file "/gpfs/scratch/as18818/P  ||
|| racticum5/HW2/LT34_FeatCount.summary"                                      ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================// ==========     _____ _    _ ____  _____
_______         _____
```

```
      =====       / ____| | | |  _ \|  __ \|  ____|  /\  |  __ \
       =====     | (___ | | | |_) | |__) | |__    /  \ | |  | |
        ====      \___ \| | |  _ <|  _  /|  __|  / /\ \| |  | |
         ====     ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
      ==========  |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
                  v1.6.3

//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 1 BAM file                                       ||
||                           P LT35.sorted.bam                                ||
||                                                                            ||
||             Output file : LT35_FeatCount                                   ||
||                 Summary : LT35_FeatCount.summary                           ||
||              Annotation : genes.gtf (GTF)                                  ||
||      Dir for temp files : /gpfs/scratch/as18818/Practicum5/HW2             ||
||                                                                            ||
||                 Threads : 1                                                ||
||                   Level : meta-feature level                               ||
||               Paired-end : yes                                             ||
||       Multimapping reads : not counted                                     ||
|| Multi-overlapping reads : not counted                                      ||
||   Min overlapping bases : 1                                                ||
||                                                                            ||
||           Chimeric reads : counted                                         ||
||         Both ends mapped : required                                        ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//

//================================== Running =================================\\
||                                                                            ||
|| Load annotation file genes.gtf ...                                         ||
||     Features : 1271434                                                     ||
||     Meta-features : 66023                                                  ||
||     Chromosomes/contigs : 331                                              ||
||                                                                            ||
|| Process BAM file LT35.sorted.bam...                                        ||
||     Paired-end reads are included.                                         ||
||     Strand specific : reversely stranded                                   ||
||     Assign alignments (paired-end) to features...                          ||
||                                                                            ||
||     WARNING: reads from the same pair were found not adjacent to each      ||
||              other in the input (due to read sorting by location or        ||
||              reporting of multi-mapping read pairs).                       ||
||                                                                            ||
||     Pairing up the read pairs.                                             ||
||                                                                            ||
||     Total alignments : 29459069                                            ||
||     Successfully assigned alignments : 13267525 (45.0%)                    ||
||     Running time : 3.57 minutes                                            ||
||                                                                            ||
||                                                                            ||
|| Summary of counting results can be found in file "/gpfs/scratch/as18818/P  ||
|| racticum5/HW2/LT35_FeatCount.summary"                                      ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//
```

```
 ==========     _____ _    _ ____  _____  ______          _____
      =====       / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
       =====     | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
        ====      \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
         ====     ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
      ==========  |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
            v1.6.3

//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 1 BAM file                                       ||
||                           P LT46.sorted.bam                                ||
||                                                                            ||
||             Output file : LT46_FeatCount                                   ||
||                 Summary : LT46_FeatCount.summary                           ||
||              Annotation : genes.gtf (GTF)                                  ||
||      Dir for temp files : /gpfs/scratch/as18818/Practicum5/HW2             ||
||                                                                            ||
||                 Threads : 1                                                ||
||                   Level : meta-feature level                               ||
||              Paired-end : yes                                              ||
||      Multimapping reads : not counted                                      ||
|| Multi-overlapping reads : not counted                                      ||
||   Min overlapping bases : 1                                                ||
||                                                                            ||
||           Chimeric reads : counted                                         ||
||         Both ends mapped : required                                        ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//

//================================== Running =================================\\
||                                                                            ||
|| Load annotation file genes.gtf ...                                         ||
||    Features : 1271434                                                      ||
||    Meta-features : 66023                                                   ||
||    Chromosomes/contigs : 331                                               ||
||                                                                            ||
|| Process BAM file LT46.sorted.bam...                                        ||
||    Paired-end reads are included.                                          ||
||    Strand specific : reversely stranded                                    ||
||    Assign alignments (paired-end) to features...                           ||
||                                                                            ||
||    WARNING: reads from the same pair were found not adjacent to each       ||
||             other in the input (due to read sorting by location or         ||
||             reporting of multi-mapping read pairs).                        ||
||                                                                            ||
||    Pairing up the read pairs.                                              ||
||                                                                            ||
||    Total alignments : 22352430                                             ||
||    Successfully assigned alignments : 6794392 (30.4%)                      ||
||    Running time : 3.13 minutes                                             ||
||                                                                            ||
||                                                                            ||
|| Summary of counting results can be found in file "/gpfs/scratch/as18818/P  ||
|| racticum5/HW2/LT46_FeatCount.summary"                                      ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//
```

```
  ==========     _____ _    _ ____  _____  ______          _____
    =====       / ____| |  | |  _ \|  __ \|  ____|   /\   |  __ \
     =====     | (___ | |  | | |_) | |__) | |__     /  \  | |  | |
      ====      \___ \| |  | |  _ <|  _  /|  __|   / /\ \ | |  | |
       ====     ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
    ==========  |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
            v1.6.3

//========================= featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 1 BAM file                                       ||
||                           P LT47.sorted.bam                                ||
||                                                                            ||
||           Output file : LT47_FeatCount                                     ||
||                Summary : LT47_FeatCount.summary                            ||
||             Annotation : genes.gtf (GTF)                                   ||
||      Dir for temp files : /gpfs/scratch/as18818/Practicum5/HW2             ||
||                                                                            ||
||                Threads : 1                                                 ||
||                  Level : meta-feature level                                ||
||              Paired-end : yes                                              ||
||      Multimapping reads : not counted                                      ||
|| Multi-overlapping reads : not counted                                      ||
||   Min overlapping bases : 1                                                ||
||                                                                            ||
||          Chimeric reads : counted                                          ||
||        Both ends mapped : required                                         ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//

//================================== Running =================================\\
||                                                                            ||
|| Load annotation file genes.gtf ...                                         ||
||    Features : 1271434                                                      ||
||    Meta-features : 66023                                                   ||
||    Chromosomes/contigs : 331                                               ||
||                                                                            ||
|| Process BAM file LT47.sorted.bam...                                        ||
||    Paired-end reads are included.                                          ||
||    Strand specific : reversely stranded                                    ||
||    Assign alignments (paired-end) to features...                           ||
||                                                                            ||
||    WARNING: reads from the same pair were found not adjacent to each       ||
||             other in the input (due to read sorting by location or         ||
||             reporting of multi-mapping read pairs).                        ||
||                                                                            ||
||    Pairing up the read pairs.                                              ||
||                                                                            ||
||    Total alignments : 44074277                                             ||
||    Successfully assigned alignments : 15729391 (35.7%)                     ||
||    Running time : 10.63 minutes                                            ||
||                                                                            ||
||                                                                            ||
|| Summary of counting results can be found in file "/gpfs/scratch/as18818/P  ||
|| racticum5/HW2/LT47_FeatCount.summary"                                      ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//



  ==========     _____ _    _ ____  _____  ______          _____
```

```
      =====       / ____|| | | | _ \|  __ \|  ____|  /\  |  __ \
       =====     | (___ | | | | |_) | | |__)  | |__    /  \ | | | |
        ====      \___ \| | | |  _ <|  _  /|  __|  / /\ \| | | |
         ====     ____) | |__| | |_) | | \ \| |____ / ____ \| |__| |
      ==========  |_____/ \____/|____/|_|  \_\______/_/    \_\_____/
            v1.6.3

//========================== featureCounts setting ===========================\\
||                                                                            ||
||            Input files : 1 BAM file                                        ||
||                          P LT48.sorted.bam                                 ||
||                                                                            ||
||          Output file : LT48_FeatCount                                      ||
||                Summary : LT48_FeatCount.summary                            ||
||             Annotation : genes.gtf (GTF)                                   ||
||      Dir for temp files : /gpfs/scratch/as18818/Practicum5/HW2             ||
||                                                                            ||
||                Threads : 1                                                 ||
||                  Level : meta-feature level                                ||
||              Paired-end : yes                                              ||
||      Multimapping reads : not counted                                      ||
|| Multi-overlapping reads : not counted                                      ||
||   Min overlapping bases : 1                                                ||
||                                                                            ||
||          Chimeric reads : counted                                          ||
||        Both ends mapped : required                                         ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//

//================================== Running ==================================\\
||                                                                            ||
|| Load annotation file genes.gtf ...                                         ||
||    Features : 1271434                                                      ||
||    Meta-features : 66023                                                   ||
||    Chromosomes/contigs : 331                                               ||
||                                                                            ||
|| Process BAM file LT48.sorted.bam...                                        ||
||    Paired-end reads are included.                                          ||
||    Strand specific : reversely stranded                                    ||
||    Assign alignments (paired-end) to features...                           ||
||                                                                            ||
||    WARNING: reads from the same pair were found not adjacent to each       ||
||             other in the input (due to read sorting by location or         ||
||             reporting of multi-mapping read pairs).                        ||
||                                                                            ||
||    Pairing up the read pairs.                                              ||
||                                                                            ||
||    Total alignments : 43684206                                             ||
||    Successfully assigned alignments : 15888427 (36.4%)                     ||
||    Running time : 5.86 minutes                                             ||
||                                                                            ||
||                                                                            ||
|| Summary of counting results can be found in file "/gpfs/scratch/as18818/P  ||
|| racticum5/HW2/LT48_FeatCount.summary"                                      ||
||                                                                            ||
\\====================== http://subread.sourceforge.net/ ======================//
```

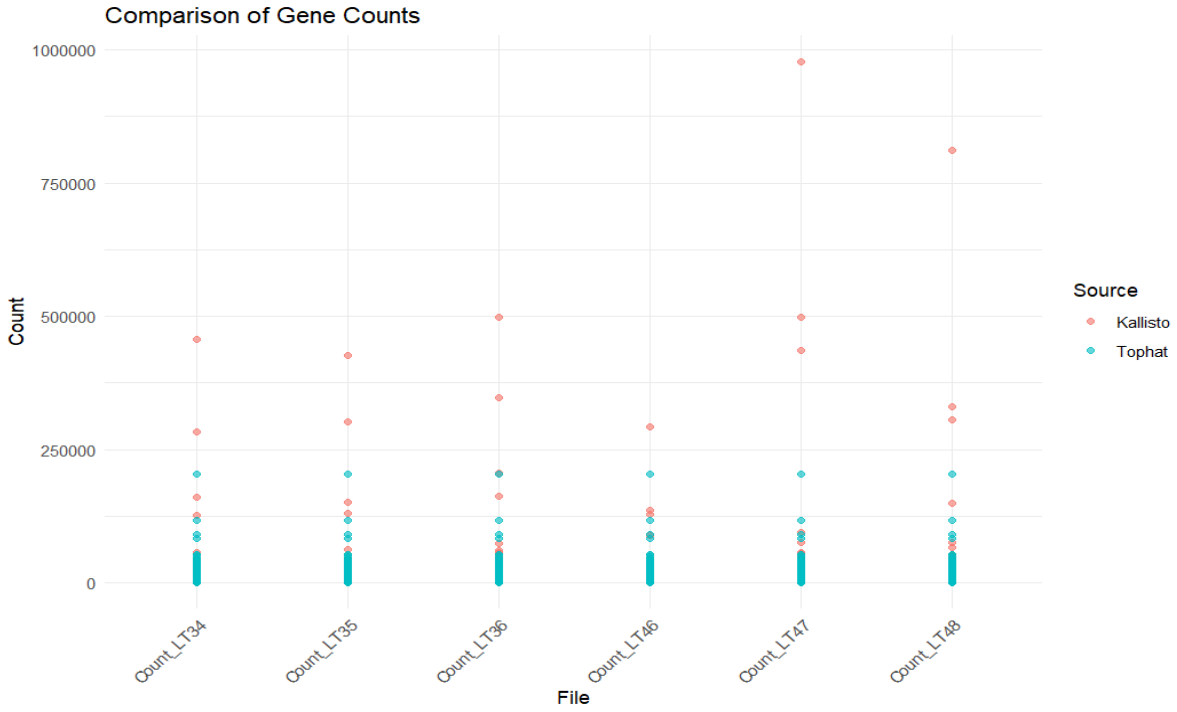
Code to get gene_counts of each sequence from the output of Tophat and combining into an excel sheet:

```
#DO THIS FOR ALL THE SEQUENCE:
featurecounts_file_LT48 <- "LT48_FeatCount"
featurecounts_data_LT48 <- read.table(featurecounts_file_LT48, header=TRUE)
# Extract gene names and counts
genes <- featurecounts_data_LT48$Geneid
counts <- featurecounts_data_LT48$Length
# Remove decimal points and numbers following them from gene IDs
cleaned_genes <- gsub("\\..*", "", genes)

# Create a data frame with cleaned gene names and counts
gene_counts_LT48 <- data.frame(Gene = cleaned_genes, Count = counts)
# Create a data frame with gene names and counts

#gene_counts_LT48 <-  data.frame(Gene = genes, Count = counts)
# View the extracted data
head(gene_counts_LT48)
# Save the data to a new file if needed
write.csv(gene_counts_LT48, file = "gene_counts_LT48.csv", row.names = FALSE)


# Load the required library
library(dplyr)

# Define file names
file_names <- c("LT34", "LT35", "LT36", "LT46", "LT47", "LT48")

# List to store data frames
count_dfs <- list()

# Read and process each file
for (file_name in file_names) {
  # Read the CSV file
  file_path <- paste0("gene_counts_", file_name, ".csv")
  count_df <- read.csv(file_path)

  # Rename the count column to include the file name
  colnames(count_df)[2] <- paste0("Count_", file_name)

  # Add to the list
  count_dfs[[file_name]] <- count_df
}

# Merge data frames based on the gene column
combined_df <- count_dfs[[1]] # Initialize with the first dataframe
for (i in 2:length(count_dfs)) {
  combined_df <- merge(combined_df, count_dfs[[i]], by = "Gene", all = TRUE)
}

# Write the combined dataframe to a CSV file
write.csv(combined_df, file = "combined_gene_counts.csv", row.names = FALSE)
```

Generate scatter plots of gene counts (genome) vs. gene counts (transcriptome) and calculate correlation.

•Are the results similar? If not, can you experiment with featurecounts/htseq to improve this?

The results are not similar for sure.




```
# Load the required libraries
library(ggplot2)
library(tidyr)
library(readxl)
setwd("C:/Users/Apoorva Sharma/Downloads/Practicum5")
# Read the data from the Excel files
output <- read_excel("output.xls")
combined_gene_counts <- read_excel("combined_gene_counts.xls")

# Define the count columns
count_columns_output <- grep("^Count_", names(output))
count_columns_combined <- grep("^Count_", names(combined_gene_counts))

# Melt the data frames to long format for plotting
output_long <- pivot_longer(output, cols = count_columns_output, names_to = "File", values_to = "Count")
combined_gene_counts_long <- pivot_longer(combined_gene_counts, cols = count_columns_combined, names_to
= "File", values_to = "Count")

# Add a column indicating the source of data
output_long$Source <- "Kallisto"
combined_gene_counts_long$Source <- "Tophat"

# Combine the data frames
```

```
combined_data <- rbind(output_long, combined_gene_counts_long)

# Plot the scatter plot
scatter_plot <- ggplot(combined_data, aes(x = File, y = Count, color = Source)) +
  geom_point(alpha = 0.6) +
  labs(x = "File", y = "Count", title = "Comparison of Gene Counts") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Display the plot
print(scatter_plot)


# Remove missing values
kallisto_counts <- kallisto_counts[!is.na(kallisto_counts)]
tophat_counts <- tophat_counts[!is.na(tophat_counts)]

# Ensure both vectors have the same length
min_length <- min(length(kallisto_counts), length(tophat_counts))
kallisto_counts <- kallisto_counts[1:min_length]
tophat_counts <- tophat_counts[1:min_length]

# Calculate correlation
correlation <- cor(kallisto_counts, tophat_counts)

# Print correlation
cat("Correlation between Kallisto and Tophat counts:", correlation, "\n")
```

```
> # Print correlation
> cat("Correlation between Kallisto and Tophat counts:", correlation, "\n")
Correlation between Kallisto and Tophat counts: 0.01019092
>
```

The gene counts derived from Kallisto are consistently higher compared to those obtained from Tophat for the same samples. This suggests that there are differences in the gene count estimates between these two tools, which is not uncommon given that they use different algorithms for read mapping and quantification. Tophat is an older spliced read mapper that aligns RNA-Seq reads to a genome, while Kallisto is a newer tool that uses pseudoalignment for rapid transcript quantification. Kallisto's approach does not rely on the genome itself but rather on a reference transcriptome, which may result in different gene counts.

To potentially improve the alignment and get a more consistent comparison between genome and transcriptome gene counts we can adjust the parameters used in Tophat and Kallisto to ensure they are optimized. Ensure that the genome annotation and the transcriptome reference are consistent with each other.