Short Read Sequencing Analysis Workshop
Day 7 Homework Answers

1. In total Tophat found 16706546 or 83.13% of reads to be properly paired. Your script should have the following commands:

   ```
   TOPHAT="/mnt/scratch_nobackup/<USERNAME>/Workshop/RNA-seq/TopHat"
   SCRATCH="/mnt/scratch_nobackup/<USERNAME>/Workshop"

   ### Load the tophat2 module and its dependencies
   module load bowtie_bowtie2-2.0.2
   module load samtools_0.1.18
   module load tophat_2.0.6

   ### map reads with tophat
   tophat2 --b2-fast -p 18 -r 325 --mate-std-dev 150 --microexon-search
   \--library-type fr-firststrand --rg-id EliRepA --rg-sample Human \
   --no-novel-juncs -o $TOPHAT -G $SCRATCH/Genomes/Hg38/Hg38.genes.gtf \ 
   $SCRATCH/Genomes/Bowtie2Index/Hg38 $SCRATCH/FASTQ/Hg_RNA_R1.fastq \
   $SCRATCH/FASTQ/Hg_RNA_R2.fastq

   ### get alignment stats
   samtools flagstat $TOPHAT/accepted_hits.bam > $TOPHAT/accepted_hits.alignment_stats.txt

   ### create an index for accepted_hits.bam
   samtools index $TOPHAT/accepted_hits.bam
   ```

2.) There are 1562842 ambiguously mapped reads according to HTSeq counts
Your script should have the following commands:

   ```
   SCRATCH="/mnt/scratch_nobackup/verajm/Workshop"
   TOPHAT="/mnt/scratch_nobackup/verajm/Workshop/RNA-seq/TopHat/chr21"

   ### Load modules
   module load htseq_0.6.1
   module load python_2.7.3
   module load numpy_1.9.2
   module load pysam_0.8.4
   module load samtools_0.1.18

   ### Run my commands, or whatever else IN SCRATCH
   ### sort BAM output and sort by name
   samtools sort -n $TOPHAT/accepted_hits.bam
   $TOPHAT/accepted_hits.sortedByName
### Usage: htseq-count [options] <alignment_file> <gff_file>

htseq-count -f bam -r name -s reverse -t exon -m union -i gene_id \
$TOPHAT/accepted_hits.sortedByName.bam \
$SCRATCH/Genomes/Hg38/Hg38.genes.gtf \
> $TOPHAT/Hg38.genes.counts.txt