Additional Exercises

Check out the differences in measured expression between the two aligners:

```
cts <- read.delim("counts.txt", comment.char = "#", row.names = 1)
View(cts)
# here the 6th column contains the counts of bowtie2, the 7th of hisat2:
bt2 <- cts[,6]
hs2 <- cts[,7]
plot(bt2, hs2)
# this looks better with a log transformation:
plot(log2(bt2+1), log2(hs2+1))
```

There is a probably a difference between the total number of counts between the aligners because the alignment rates differed. It would be nice to correct for the total number of counts. The most simple way to do that is to transform them into CPM values. Have a look over here, and generate a table with CPM values in stead of counts.

Look up genes with specifically large differences in log transformed counts and/or CPM values and check them in IGV. You can e.g. do that like this:

```
diff <- abs(log2(bt2+1) - log2(hs2+1))
names(diff) <- rownames(cts)
top10 <- diff[order(diff, decreasing = TRUE)][1:10]
View(cts[names(top10), c(6,7)])
```

Project 1: you can use the built-in genome and gtf in IGV
Project 3: load in addition to you bam file also the gtf you have used for the counting in IGV

Can you think of reasons for these differences? If the counts are low in one and high in the other, do the reads end up somehwere else?

You can look up specific reads in a bam file like this:

```
# change <read name> to a read name you want to look up (remove the <>)
samtools view SRR7822040.chr5.hs2.bam | grep <read name> | cut -f 1-5
```

In these genes that differ a lot between `bowtie2` and `hisat2` there might be some reads with a (very) low MAPQ. A mapping quality of 1 means basically total uncertainty on where a read belongs. Check out the option `-Q` in `featureCounts`. Does having a threshold in MAPQ affect counting?

Try to figure out whether your data is stranded, e.g. by looking into the original publication (you can find it at the group work page). If unsure, you can check it with `rseqc`:

Download `gtf2bed.pl`

```
wget https://raw.githubusercontent.com/timothyjlaurent/GenomicsTools/master/gtf2bed.pl
```

Generate a bed file from the gtf:

```
perl ./gtf2bed.pl Mus_musculus.GRCm38.102.chr5.gtf > Mus_musculus.GRCm38.102.chr5.bed
```

Install `reseqc`

```
conda install -c bioconda rseqc
```

Run `infer_experiment.py`:

```
infer_experiment.py \
-i alignments/SRR7822040.chr5.hs2.bam \
-r reference/Mus_musculus.GRCm38.102.chr5.bed
```

If your library is stranded, does it have an effect on the counts if you specify it at `hisat2` and/or `featureCounts`? How can specifying strandedness help in quantification of gene expression?