BIO609: Part 2, exercise solutions

Exercise 5: handling RNA-seq data

```
#!/bin/bash
for i in {1..10}
do
   wget https://bioinfo.evolution.uzh.ch/share/bio609/dicty/sample${i}.fastq.gz
done
```

Exercise 6: map the RNA-seq data to the reference genome

We will now try to use STAR and map our 10 samples of RNA-seq data to the reference genome. First, you need to create an "index" for the dd.fasta reference genome. You can do this with:

```
mkdir istar # folder for genome index
STAR --runMode genomeGenerate --genomeDir istar --genomeFastaFiles dd.fasta
```

The command should take around 1-2 minutes to finish. The index is now stored in the folder **istar**.

Then you can write a script to map the RNA-seq data to the reference genome using the newly created index above.

An example command to map sample1 would be:

```
#!/bin/bash
for i in {1..10}
do
   STAR --genomeDir istar --readFilesIn sample${i}.fastq.gz --readFilesCommand zcat
   mv Aligned.out.sam sample${i}.sam
   samtools view -S -b sample${i}.sam > sample${i}.bam
   samtools sort sample${i}.bam sample${i}
   samtools index sample${i}.bam
   rm sample${i}.sam
done
```

Exercise 7: download genome annotation in GFF format and count reads aligned to genes

```
#!/bin/bash
for i in {1..10}
do
   htseq-count -f bam sample${i}.bam dd.gtf > sample${i}.tab
done
```

* Exercise 8: combine gene expression tables into one single table

```
awk 'NF > 1{ a[$1] = a[$1]"\t"$2} END {for( i in a ) print i a[i]}' *.tab >
samples.tab
```