RNA-seq practical 1: Read mapping

In this first practical we will not actually map sequenced reads, but instead have a closer look at the 'raw' reads as they come from the sequencer and at reads mapped to the Arabidopsis genome. Mapping reads is typically done on a compute server using software that runs on a Linux command line. This is quite straightforward, but requires some basic training on how to use Linux, which is outside of the scope of this course.

The raw reads

If you send your samples to a sequencing facility, you usually get a portable hard disk with so called FASTQ files that contain the sequences.
https://en.wikipedia.org/wiki/FASTQ_format

A FASTQ file can contain millions of sequences, and each sequence is accompanied by information on how reliable the base calling was. The FASTQ format looks like this:

```
@SRR5304927.1 1 length=76
CTATTNTACTTAAAGGATTAATCTAATTGCTCTTAATTACAATGCACAACCTGTCAAATAGATATAGCCTTGTTGA
+SRR5304927.1 1 length=76
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE/EEEEEEEEEEEEEEEEEE
@SRR5304927.2 2 length=73
GGCGANTACGTCGATGTGAAGGTGAATGGTGCGATCCACAAGGGTATGCCTCACAAGTTCTACCATGATCGTA
+SRR5304927.2 2 length=73
AAAA6#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEEEEEEEEA6EEEEEEEEEA/EEEEE
```

The file is organized in blocks of 4 lines per sequence. The second line of each block contains the sequence; the fourth line contains a quality score for each nucleotide encoded by a letter (see the FASTQ Wikipedia page). The quality score runs from 1-41 (higher is better quality).

Note that there is no information in the FASTQ file about the gene/transcript the read was derived from.

Download this file to your computer:
http://www.bioinformatics.nl/courses/RNAseq/ST_1.fastq

To get a summary of the quality of the reads, we can use the program FastQC. You can download it from:
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.7.zip

Download it to the D: drive and unzip it to D:\fastqc_v0.11.7.zip.

Start FastQC by double clicking **run_fastqc.bat** and open de ST_1.fastq file. The quality of the reads is reported in several categories. **Per base sequence quality** shows boxplots for the quality score at each position of the reads. Note that the average base quality goes down towards the end of the reads.
Due to technical reasons the first ~5 nucleotides of the reads are not completely random, which is reflected in the **Per base sequence content** statistic. The

really bad quality nucleotides of some of reads have already been trimmed off, which explains why not all reads have the same length in the **Sequence length distribution**.

Now have a look at this FASTQ file with FastQC:
http://www.bioinformatics.nl/courses/RNAseq/SRR5304927_1.fastq

These RNA-seq reads were sequenced from plants that were grown in the International Space Station. The data were downloaded from the Sequence Read Archive: https://www.ncbi.nlm.nih.gov/sra/SRR5304927

This file has some seriously looking quality issues, but we will see that it can still be used for further analysis.

The Spaceflight reads were mapped to the Arabidopsis reference genome using the Hisat2 software. The mapped reads can be downloaded via this zip file:
http://www.bioinformatics.nl/courses/RNAseq/SRR5304927.zip

The zip contains a **BAM** file (**B**inary **A**lignment **F**ile) with its index file (bai). BAM is a compressed form of the SAM file format:
https://en.wikipedia.org/wiki/SAM_(file_format)

Unzip the **SRR5304927.zip** file to the D: drive.

To view the mapped reads, we can use the "Integrative Genomics Viewer".
https://software.broadinstitute.org/software/igv/download

Select "Launch with **750 MB**" to startup IGV. You might have to first download the program to the D: drive and then run it.

From the Genomes menu choose 'Load genome from server' and select the *Arabidopsis thaliana* TAIR10 genome. Then from the File menu, select "Load from File…" to load the **SRR5304927.bam** file (that was in the downloaded zip file). Now you can select regions on the Arabidopsis genome to view the mapped reads.

In the text box left of the **Go** button type: **AT3G46030** and click **Go** to zoom into the position of the Histone H2B.7 gene.

You can see the mapped reads in the SRR5304927.bam track. The color of a read indicates the strand to which it maps, the arrow indicates the direction. Some reads have vertical lines that indicate a mismatching nucleotide with the reference genome. If you click with the right mouse button in the track you can select "View as pairs" to have read pairs that were derived from the same mRNA fragment connected by thin horizontal lines (or overlapping).

The coverage track shows the number of reads that cover a certain position. In the Gene track you can see mRNA and proteins. For the H2B.7 gene the mapped reads nicely confirm the annotated mRNA.

Some of the annotated genes have no reads mapped to them, like **AT2G01500**, and some genes only have a few reads mapped, like **AT2G01460**, and even reads mapped to unannotated regions.

In the **AT3G46040** gene the coverage track shows gaps that correspond with the annotated introns. Can you spot reads that span two exons?

Looking in this detail at mapped reads is not part of a normal RNA-seq analysis, but can sometimes help to spot problems or understand the data better.

For a first introduction to the Linux command line see:
http://rik.smith-unna.com/command_line_bootcamp/

For a protocol to map reads to a genome see:
https://www.nature.com/articles/nprot.2016.095

The steps to download the FASTQ file and map the reads were as follows (the TAIR11 genome index was created before):

```
fastq-dump --split-spot --split-3 SRR5304927
hisat2 -x TAIR11 -1 SRR5304927_1.fastq -2 SRR5304927_2.fastq -S SRR5304927.sam
samtools sort -o SRR5304927.bam SRR5304927.sam
samtools index SRR5304927.bam
```