Analysing Transcriptomics data

Stranded vs Unstranded RNASeq data

Check the papers:

https://www.nature.com/articles/nmeth.1491.pdf

* https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0026426

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3552703/

We will be working with RNASeq data from Arabidopsis thaliana, so we will need to download the genome assembly, annotation and transcripts from https://www.araport.org/. Use guest credentials to get the following data:

Genome assembly: TAIR10_Chr.all.fasta.gz

The latest annotation release in GTF format (Araport11_GFF3_genes_transposons.201606.gtf.gz) and the predicted transcripts in fasta format (cDNA: Araport11_genes.201606.cdna.fasta.gz).

Mapping reads to the genome

You will need the software hisat2 (https://ccb.jhu.edu/software/hisat2/index.shtml), download a binary for your platform form the authors website, Also please install samtools, e.g., sudo apt install samtools

```
wget ftp://ftp.ccb.jhu.edu/pub/infphilo/hisat2/downloads/hisat2-2.1.0-Linux_x86_64.zip
```

Before mapping you must created an index of your genome sequence. Follow the steps (in the folwder where you have the Arabidopsis thaliana genome):

```
gunzip TAIR10_Chr.all.fasta.gz
```

```
gunzip Araport11_GFF3_genes_transposons.201606.gtf.gz
```

```
./hisat2-2.1.0/hisat2_extract_splice_sites.py Araport11_GFF3_genes_transposons.201606.gtf  > Araport11_genes_splicesites.ss
```

```
./hisat2-2.1.0/hisat2_extract_exons.py Araport11_GFF3_genes_transposons.201606.gtf > Araport11_genes_exons.ex
```

```
hisat2-build --exon Araport11_genes_exons.ex --ss Araport11_genes_splicesites.ss TAIR10_Chr.all.fasta TAIR10_Chr.all #This will take some time
```

Make sure to install the latest version of hte SRA toolkit (https://www.ncbi.nlm.nih.gov/sra/docs/toolkitsoft/), and use faster-dump to get the following SRA accessions: DRR01612 and SRR6914596

```
./sratoolkit.2.9.6-1-ubuntu64/bin/fasterq-dump DRR016126 -p -split-files
```

```
./sratoolkit.2.9.6-1-ubuntu64/bin/fasterq-dump SRR6914596  -p -split-files
```

Map these reads to the Arabidopsis genome, using hisat2:

```
./hisat2-2.1.0/hisat2 --threads 4 -x TAIR10_Chr.all -1 DRR016126_1.fastq -2 DRR016126_2.fastq | samtools view -b| samtools sort -o DRR016126.ATHA.sorted.bam

samtools index DRR016126.ATHA.sorted.bam

./hisat2-2.1.0/hisat2 --threads 4 -x TAIR10_Chr.all -1 SRR6914596_1.fastq -2 SRR6914596_2.fastq | samtools view -b| samtools sort -o SRR6914596.ATHA.sorted.bam

samtools index SRR6914596.ATHA.sorted.bam
```

Visualize the BAM file in the IGV. Look for the genes: AT4G18670, AT4G00710

Make sure to activate the visualization of splice junctions.

Run hisat2 again, but this time specify that the library is stranded for the proper lib (only one of them, which one?)