Exercise 1.

Non-PCR-based T-DNA insertion analysis (Goes where simple PCRs have never gone before).

In one of projects that I'm involved in I work with an Arabidopsis mutant. This Arabidopsis mutant contains a T-DNA insertion in the gene named AT1G02840 or *SR1*. This gene plays a role in the regulation of pre-mRNA splicing.

As a bioinformatician I am sometimes very naïve when it comes to molecular biology and I actually thought *SR1* should not be expressed. After performing standard analysis involving read-mapping, gene-structure prediction and quantification I noticed that *SR1* was still expressed. I therefore decided to have a closer look at this gene.

I reasoned as follows: If the T-DNA disrupts the promoter of the gene, the gene might not be expressed in the first place. Given the fact that the gene is expressed, the T-DNA is more likely to be located in the transcribed part of the gene. Therefore, the T-DNA is likely to be present in the RNA sample fused to the SR1 transcript fragments. In this exercise you will learn how I determined whether 1. The T-DNA was expressed in the RNA sample and 2. It has been inserted in the SR1 gene. Note, that due to time-restrictions, the analysis you will perform is slightly different from the original analysis I performed. Nevertheless the same principles apply.

As the T-DNA is unlikely to be found by mapping the reads to the reference genome of Arabidopsis thaliana, a *de novo* assembly is required.

For the de novo assembly you will use the files: exercise1_1.fastq and exercise1_2.fastq which correspond to the forward and reverse Illumina reads that correspond to the Arabidopsis mutant SR1 gene. The de novo assembly will be performed using Trinity program.

Detail:

Upload the files exercise1_1.fastq and exercise1_2.fastq by performing twice the following steps:

1. From the option Get Data in the main menu, select the sub-item upload file to upload the files to be used in this exercise.

2. Select 'fastqsanger' as file format.

3. In the URL/Text box enter:

http://www.bioinformatics.nl/courses/RNAseq/exercise1_1.fastq

4. Press execute.

5. Repeat steps 1 and 2 and this time in the URL/Text box enter:

http://www.bioinformatics.nl/courses/RNAseq/exercise1_2.fastq

6. Press execute.

Perform the Trinity de novo assembly

1. Select NGS:Assembly at the main menu, and choose the sub-item Trinity.

2. Make sure that under the option left/forward strand reads is the file exercise_1.fastq and right/reverse strand reads is the file exercise_2.fastq, respectively.

3. ***(one word of caution! DO NOT SET THE ADDITIONAL PARAMETER OPTION TO YES!)***

4. Leave the rest of the parameters at their default values and press execute.

The de novo assembly of the reads should now be running. Be a little bit patient please. When the assembly is done (green) please have a look at the resulting output which is a multi-sequence fasta file. So tell me, do you also think that performing a standard de novo assembly is too easy to be true? Of course you do ……. unless something went wrong!

It is very likely that the assembly program has produced several contigs. Let's see how they are related to each other. To this end, you will have to execute the following steps:

a. *Go to the following link http://www.ebi.ac.uk/Tools/msa/mafft/*
b. *View the contigs that have been produced on the Galaxy webpage, select them all and copy.*
c. *Paste these contigs*
d. *Select Nucleic acid as the sequence type and choose ClustalW as output format.*
e. *Press submit.*

Do you think all the sequences are from the same "locus"?

Let's inspect the results a bit further: How many loci have been created by the Trinity? Tip: Trinity names it contigs as using the following format: compX_cY_seqZ where X, Y and Z can vary. When two contigs have the same X and Y values, they are considered to be from the same locus.

In the next step we are going to check whether one or more of the assembled contigs are in fact fusions between the T-DNA and the *SR1* transcript sequences. It makes sense to try to match the contigs. In order to make your exercise more pleasant, I took the time to collect these sequences. You will use the blastn program for matching the T-DNA and *SR1* sequences against the contigs.

For this part the following things have to be done.

1. *Upload the sequences I collected for you.*
    a. *As you did on the previous exercise, upload a file from this URL: http://www.bioinformatics.nl/courses/RNAseq/exercise_1_collected_sequences.fasta*
        i. *Set the file type to fasta and not fastq (good-practice)*
2. *Create a blast-database from your assembled sequences. You will search the sequences I collected against the blast database.*

a. *Select NCBI BLAST+ at the main menu, and choose the sub-item NCBI BLAST+ makeblastdb.*
b. *Select nucleotide as the molecule type of input.*
c. *Select the fasta-file corresponding to the assembled sequences (in Galaxy, 4: Trinity on data 1 and data 2: Assembled Transcripts).*
d. *Set a title for the blast-database.*
e. *Press execute.*

3. *Now you are ready to run blastn. So select now the sub-item NCBI BLAST+ blastn (because you will search a nucleotide sequence against a nucleotide database).*
    a. *For the Nucleotide query sequence, select the uploaded collected sequences file.*
    b. *Select 'BLAST database from your history' under Subject Database.*
    c. *Select your blast database under nucleotide BLAST database (in Galaxy, 6: nucleotide BLAST database from data 4)*
    d. *Type of BLAST: blastn.*
    e. *Select pairwise HTML under output format.*
    f. *Press execute.*

At this point your blast-search should be executed. Please wait, you will be able to inspect the results in a very short time.

Inspect the blast results and try to identify those contigs that are fusions between T-DNA and SR1 transcript sequences. To this end you can execute the following steps.

a. *Take a pen and paper.*
b. *Draw a straight line on scale from 1 to the length of the longest contig.*
c. *For each blast hit (not the very short ones) draw a line from the start to end position for this selected contig.*
d. *Check how the lines corresponding to SR1-blast-hits are orientated compared to the line for the T-DNA.*

Exercise 2.

From reads to annotated contigs:

As mentioned in the classes, a basic RNAseq analysis contains the following steps: Transcript reconstruction, annotation of putative transcripts and if possible, quantification and differential expression of the transcripts. In this exercise we will focus on the first two steps. You will work on data from the species *Miscanthus senensis.* This species is of particular interest in Europe as a potential crop for biofuel production. The transcriptome of *M. senensis* has recently been sequenced and published, providing a rich source of information to whomever is interested in this species.

I recently performed a *de novo* assembly and annotation for the available RNAseq data of this species. The goal of this exercise is to give you an idea of how I did that by doing it yourselves. Although the original tools that I used are different versions of the tools that you will be using, the general principle is the same. As *de novo* assembly of a large number of reads can take days and requires the access to substantial computational power, you are going to work with a small fraction of all the reads. The sub-selection of reads that I'm using is not random. It will be your job to find out why this is not the case.

Ok enough information; let's get to work.

The *de novo* assembly:

1. **The locations of the forward- and reverse fastq files that you will need to upload are http://www.bioinformatics.nl/courses/RNAseq/exercise2_left.fastq and http://www.bioinformatics.nl/courses/RNAseq/exercise2_right.fastq , respectively.**
2. **Perform a *de novo* assembly using Trinity in the same way as you did in exercise 1.**
3. **Download the assembled contigs to your computer.**

Annotation of the contigs:

As the number of contigs you will get is very small, we can use the Blast2Go version with a graphical interface for annotating the contigs. For really large datasets, which are common nowadays, there are more efficient ways of performing some of these steps (using local computer clusters and databases).

Starting blast2go:

1. Open Internet Explorer or Firefox and go to this page: http://www.blast2go.com/b2glaunch/start-blast2go
2. Set the select the amount of java memory option to 500mb.
3. Press on the Please click here link.
4. Load your fasta file by clicking on File in the main menu and then on Load Sequences (e.g ..)

Performing blastx search against nr, mapping:

The first thing you will need to do is to search your sequences against a database of "known" sequences. You will do this using the blast interface of Blast2GO. You will need to run blastx because you are searching nucleotide sequences against a protein database.

1. Select Blast on the main menu and then Run blast step
2. Press the play icon (from the old tape recorders)
3. Give a name for the blast output and wait a while for the blast step to finish.
4. Select Mapping from the main menu and then Run GO-mapping step and wait a while for the mapping to finish.
5. Select Annotation on the main menu and the Run annotation step
6. Select Annotation on the main menu, then select Enzyme code and KEGG and finally Load pathway-mappings from KEGG

After you have completed all these steps, you have a basic annotation of your contigs. Now, look at the pathways …. For which pathway do you think I selected the reads corresponding to its members?

Exercise 3.

A typical differential expression analysis for a species with a sequenced genome.

In this final exercise you will work with the well-known model species *Arabidopsis thaliana*. I found an experiment online in which RNAseq data was generated for plants grown under either low or high Mg^{2+} concentrations. One thing that I'm missing is RNAseq data for plants grown under normal Mg^{2+} conditions. But hey ….. for the sake of this exercise, I couldn't care less about the missing control sample.

Ok, let's get to the point…… The purpose of this exercise is to get a first impression of the famous Tuxedo pipeline/work-flow. This pipeline/workflow involves the following steps: Mapping reads to a genome, reconstructing transcripts, quantifying the transcripts and performing differential expression analysis.

Mapping reads to the Arabidopsis genome

For the sake of time we will use a very small subset of the RNAseq data.

First we need load some shared data.

1. Upload the following fastq file: http://www.bioinformatics.nl/courses/RNAseq/SRR631034_sub.fastq
2. Select Shared Data
    a. Data libraries
    b. WUR RNAseq course, June 5-7, 2013

i. Mark TAIR10_GFF3_genes.gff3

c. Select import to current history and press GO

d. Go back to Analyze Data

Running tophat

1. Now select NGS: RNA Analysis from the main menu and then Tophat for Illumina.
2. Select SRR631034_sub.fastq file under the option RNA-Seq FASTQ file:
3. Select use a built-in genome
4. Select Arabidopsis thaliana TAIR10 under reference genome
5. Select full parameter list under TopHat settings to use:
6. Select Yes under Use Own Junctions and Use Gene Annotation Model
7. Select TAIR10_GFF3_genes.gff3 under Gene Model Annotation
8. Press execute

Note! Step 6 and 7 ensure that you use the existing TAIR10 annotation as guidance for the mapper. That is for finding introns. The output files which I usually consider important are the accepted_hits.bam which contains the read alignments and splice junctions which contain identified splice-junctions.

Optionally the accepted_hits.bam file (with its index file .bai) can be visualized using the Tablet program: http://bioinf.scri.ac.uk/tablet/

- *Install Tablet on the D:\ drive.*
- *With Open Assembly you can open the bam file (not the bai file, but make sure that file is in the same folder).*
- *With Import Features you can load the gff3 file with the genome annotation.*

Transcript assembly using cufflinks

In this section you will learn how to perform transcriptome assembly using cufflinks. There are 3 modes in which you can run cufflinks. 1. You use an existing annotation and just quantify the transcripts. 2. Perform a completely new assembly using only the aligned reads and 3. You use the existing annotation as a starting guide but also allow the construction of new models.

In this section we are going to run cufflinks in the 3rd mode. As running cufflinks in this mode can take a lot of time +- 30 minutes you will only run it for one of the samples. So let's get started.

First we need load some shared data.

1. Select Shared Data
    a. Data libraries
    b. WUR RNAseq course, June 5-7, 2013
    c. Mark exercise3_34_mapped.bam (This file contains the aligned reads from the low Mg2+ sample).

d. Mark exercise3_35_mapped.bam (This file contains the aligned reads from the high Mg2+ sample).
e. Select import to current history and press GO

Optionally you can visualize both bam files side by side in the Integrative Genomics Viewer (IGV) http://www.broadinstitute.org/igv/ (Launch with 1.2 GB, cancel the Windows Security Alert)
First from Genomes select Load Genome From Server, download TAIR10. Then from File | Load From File, load both bam files (first download them to your computer, including their index files). Now for instance compare the reads mapped to the gene: AT2G36255

Running cufflinks

1. Select NGS: RNA Analysis
2. Select Cufflinks
3. Select exercise3_34_mapped.bam as the SAM or BAM file of aligned RNA-Seq reads
4. Under Use Reference annotation select Use reference annotation as guide (This corresponds to the 3rd run mode mentioned above).
5. Select TAIR10_GFF3_genes.gff3 under Reference_Annotation
6. Select Yes under use multi-read correct:
7. Press Execute (Note: This takes a while)

There are 3 output files. First you have the assembled transcripts which is a gtf file containing the exons of all the transcripts that have been found or existed in the annotation. Next you have the transcript and gene expression files which contain FPKM values for individual transcripts and genes, respectively.

Please verify that the expression level of a gene is the sum of the expression levels of its transcripts.

Merging assemblies using cuffmerge

You can imagine that each RNAseq sample will lead to a different transcript assembly. When you do a differential expression analysis you have to compare the same transcriptome under different conditions. Cuffmerge enables you to merge several assemblies into a single overall assembly. Let's see how that works. We are going to merge the existing TAIR10 annotation together with assemblies that I generated for the low- and high- Mg2+ samples.

First we need load some shared data.

1. Select Shared Data
    a. Data libraries
    b. WUR RNAseq course, June 5-7, 2013
    c. Mark cufflinks_transcripts_35.gtf (Assembly high Mg2+)
    d. Mark cufflinks_transcripts_34.gtf (Assembly Low Mg2+)
    e. Select import to current history and press GO

Merging

1. Select NGS: RNA Analysis
2. Select Cuffmerge
3. Select cufflinks_transcripts_35.gtf under GTF file produced by Cufflinks (Adds first GTF)
4. Press Add new Additional GTF input Files
5. Select cufflinks_transcripts_34.gtf under GTF file produced by Cufflinks (Adds second GTF)
6. Select Yes under Use Reference Annotation
7. Select TAIR10_GFF3_genes.gff3 under Reference Annotation
8. Press Execute

A file named merged.gtf will be created.

Differential expression

Finally, you have come to the last stage of the pipeline/workflow. In this stage you are going to perform a differential expression analysis. For differential expression you should have biological replicates. However, it is possible to do the analysis without replicates. In that case, the program estimates the gene/transcript expression variation by first threating the samples as if they were replicates. This approach assumes that the expression of the majority of genes / transcripts is not different between the conditions. Ok, let's wrap up.

Running the test

1. Select NGS: RNA Analysis
2. Select Cuffdiff
3. Select the result of Cuffmerge under Transcripts:
4. Select exercise3_34_mapped.bam under the first SAM or BAM file of aligned RNA-Seq reads:
5. Select exercise3_35_mapped.bam under the second SAM or BAM file of aligned RNA-Seq reads:
6. Select Yes under Use multi-read correct
7. Press execute

The cuffdiff run will take some time. For a full explanation of the output, please have a look at this page.
http://cufflinks.cbcb.umd.edu/manual.html

The tables produced can directly be downloaded and imported into excel. The most interesting ones are gene differential expression testing and transcript differential expression testing. Before downloading you could first (in Galaxy) filter the table for significant genes (column 14). Aalt-Jan van Dijk used these genes for the GO analysis in his presentation this morning. You could try to redo this on the DAVID website:
http://david.abcc.ncifcrf.gov/summary.jsp