March 2016

An Introduction to Deep-Sequencing Data Analysis

Exercise #1

Dena Leshkowitz and Ester Feldmesser

Introduction

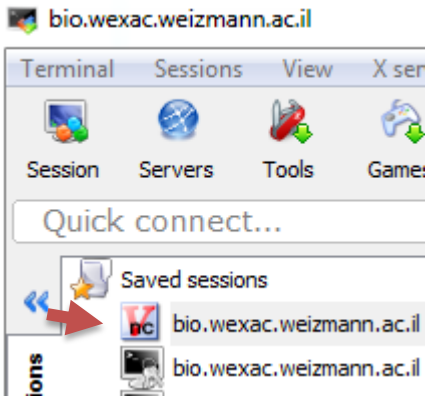
In this workshop we will learn how to evaluate the sequence quality and how to map the reads to a reference genome. The data set in this workshop is a collection of RNA-Seq data from mRNAs extracted from acute lymphoblastic leukemia (ALL) precursor B cell line.

We will use Chipster (Kallio et al. BMC Genomics 2011, 12:507), a user-friendly analysis software for high-throughput data. Its intuitive graphical user interface enables biologists to access a powerful collection of data analysis and integration tools, and to visualize data interactively. Users can collaborate by sharing analysis sessions and workflows. Chipster is an open source, and the server installation package is available for free.

Instructions

1. Accessing the data

a. Find the icon MobaXterm on your desktop and click to open.
b. You will see on the left under "Saved sessions" the following:



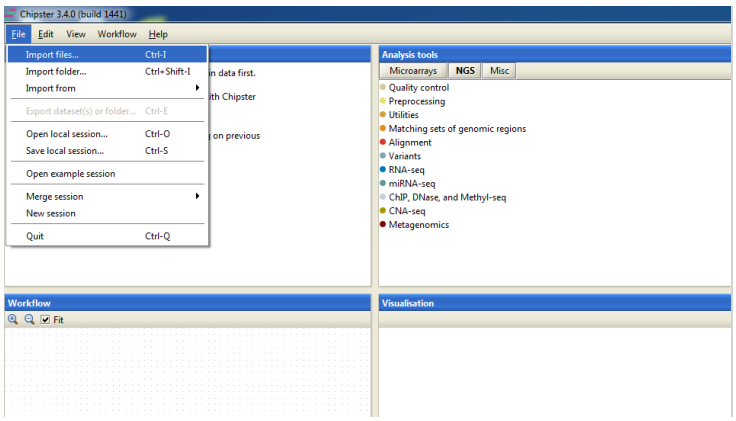
Click on the VNC session (red arrow), it can take a while until you are prompted to insert the required password (ask the instructor), this depends on your user ID. The userID is class (see blue arrow below) and a certain number, insert the password and press enter in your keyboard. The user ID and password that you are now using is needed also to open Chipster.

c. Open the Chipster application found on the VNC session.

Enter your userID and password (the same as before).

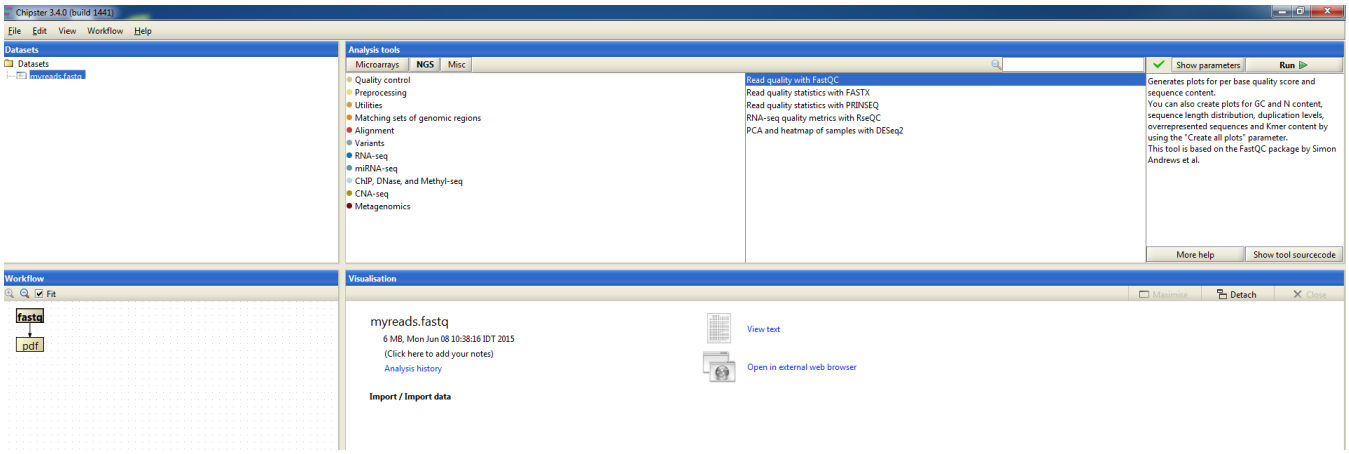
2. Running FASTQC on Chipster

a. Import the fastq file (File->importFile->select the folder files_for_course and in it myreads.fastq file)



b. Select the imported file from the "Workflow" left panel by clicking on it.

c. In the "Visualization" pane (bottom right), click on View text:

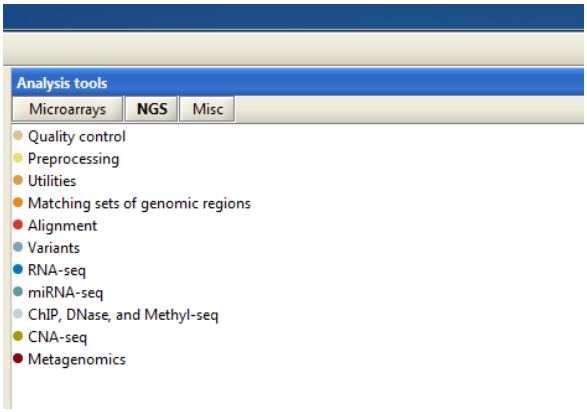


d. Look at the quality value of the first base from the first sequence. Convert the character into a numeric value based on the Supplementary converter below. Our sequences comply with the Sanger convention.

Question 1: What is the probability of an error in the first base?

e. To create a QC report for myreads.fastq with the fastqc program, select the imported file from the left panel by clicking on it.

f. In the "Analysis" tools pane, click on **NGS tab**, and then on "Quality control" and then on "Read quality with FastQC".



g. Click on the **"Run"** button at the right part. Wait for the results, an html result will appear in the left window called Workflow.
Select the html result and from the visualization pane select "Open in an external browser"

Question 2:

(a) How many sequences does the fastq file contain?

(b) Is the base quality the same for all the cycles?

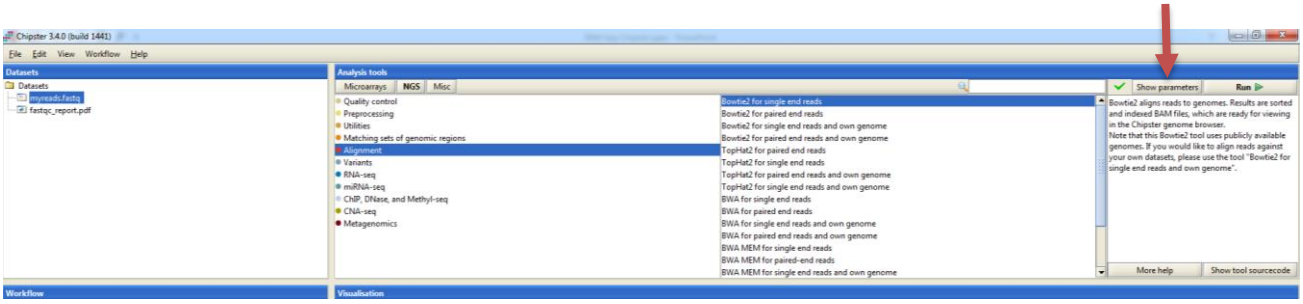
(c) Do all the cycles have an equal base content?

The sequences are from a RNA-Seq experiment. During the course we will discuss the reason for the unequal base content in the beginning of the sequences.

3. Running Bowtie and understanding the output

We are going to run bowtie:

a. Select the imported fastq file from the left panel by clicking on it.
b. In the Analysis tools, click on **NGS tab**, then on "Alignment" and then on "Bowtie2 for single end reads".
c. Click on "Show parameters" on the right (Red arrow below). Be sure to use the most recent version of the human genome: the GRCh38 version.

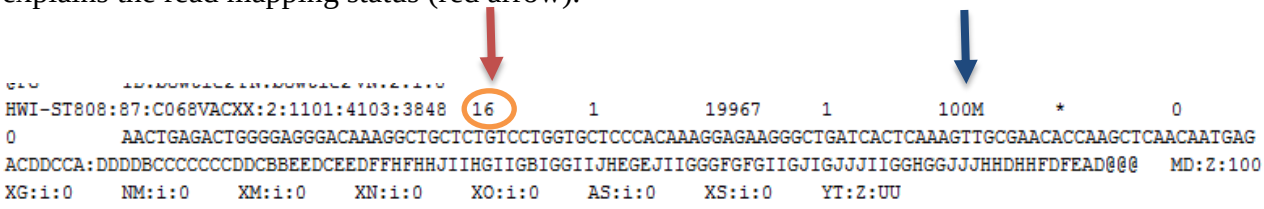


d. Click on the **"Run"** button at the right part. Wait for the results. Open the log file (view text in visualization pane)
   Question 3:
   What percentage of reads was mapped to the genome? Why are there reads not mapped?
e. Select the bam file in order to view it with the BAM viewer. The BAM file has a binary format, the bai file contains an index to the bam file to allow quick visualization. The BAM viewer converts the binary file into text in a format called SAM.

To understand the SAM format look in:
https://samtools.github.io/hts-specs/SAMv1.pdf (page 4) or below in Supplementary.

All the lines that start with @ are headings, after that there is one lane for each read and location. Reads mapped uniquely will also appear in only one line, and reads mapped to multiple locations in the genome will also appear in multiple lines.

In the lines after the headings, the first field is the read name and the second one is a flag that explains the read mapping status (red arrow).



To help you understand this flag go to the utility below. This utility explains SAM flags in plain English:

https://broadinstitute.github.io/picard/explain-flags.html

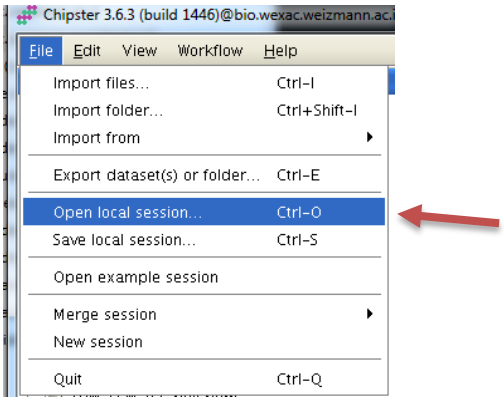
Looking at the second read information this flag is 0.
However this 0 flag is missing from the above picard utility web page. The flag 0 represents a read mapped to the forward strand that is unpaired.

Question 4:
(a) Observe the first and second mapped reads, to which chromosome and location are they mapped?
(b) Do they have mismatches and if yes what is/are it/they?
Hint: Look for the CIGAR (blue arrow; see CIGAR explanation in Supplementary below)

4. Running Tophat

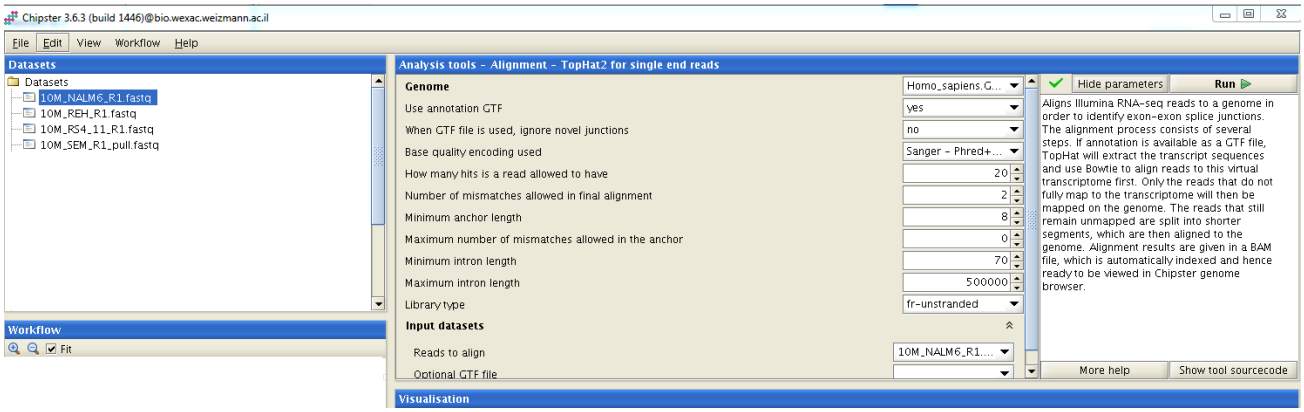
We will now run the alignments you will need for tomorrows exercise in order to analyse transcriptome (RNA-Seq) data. The aligner in this case is Tophat, this aligner utilizes bowtie but in addition it knows to map the reads that map to exon-exon junctions.

Open the local session (see below; File->Open local session) named Exercise-fastq-session.zip found in folder files_for_course.



Select the first fastq file from the "Database" pane and align it with "Tophat for single end reads" against the Homo_sapiens.GRCh38.81 genome, the other parameters should be as shown below. We will use the option "use annotation GTF" yes. Tomorrow we will discuss what this option means.

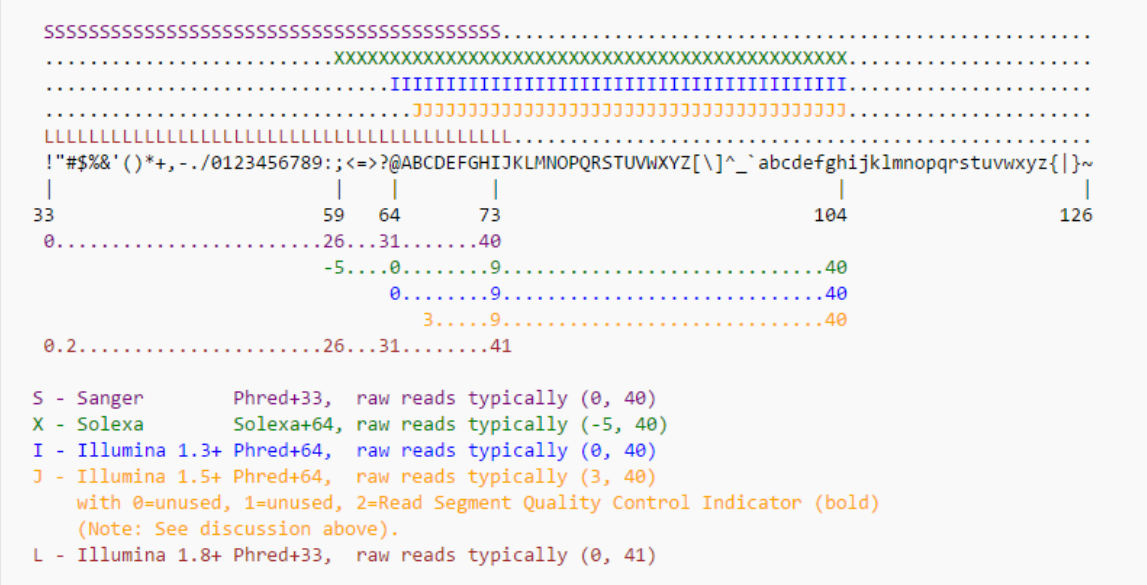
Repeat this for all 4 fastq files. You should see in the bottom of the Chipster window that all 4 jobs are running. This will take more than an hour, so you can close the VNC using the disconnect button in top of the mobaXterm screen and close the computer, the jobs will still run.



Good job. You completed this exercise.

Supplementary

Converting ASCII Characters to quality values



SAM file format in short

SAM file is aTAB-delimited version of the BAM which is a binary file. Apart from the header lines, which are started with the '@' symbol, each alignment line consists of:

Column Fields Description

1. `QNAME` Query template/pair NAME
2. `FLAG` bitwise FLAG
3. `RNAME` Reference sequence NAME
4. `POS` 1-based leftmost POSition/coordinate of clipped sequence
5. `MAPQ` MAPping Quality (Phred-scaled)
6. `CIGAR` extended CIGAR string
7. `MRNM` Mate Reference sequence NaMe ('=' if same as RNAME)
8. `MPOS` 1-based Mate POSistion
9. `LEN` inferred Template LENgth (insert size)
10. `SEQ` query SEQuence on the same strand as the reference
11. `QUAL` query QUALity (ASCII-33 gives the Phred base quality)
12. `OPT` variable OPTional fields in the format TAG:VTYPE:VALUE

CIGAR: CIGAR string.

The CIGAR operations are given in the following table:

M alignment match

I insertion to the reference

D deletion from the reference

N skipped region from the reference

S soft clipping (clipped sequences present in SEQ(

H hard clipping (clipped sequences NOT present in SEQ(

P padding (silent deletion from padded reference(

= sequence match

X sequence mismatch

THE END.