June 2015

An Introduction to Deep-Sequencing Data Analysis

Hands-on #1

Dena Leshkowitz and Esti 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 open source, and the server installation package is freely available.

Instructions

1. Accessing the data

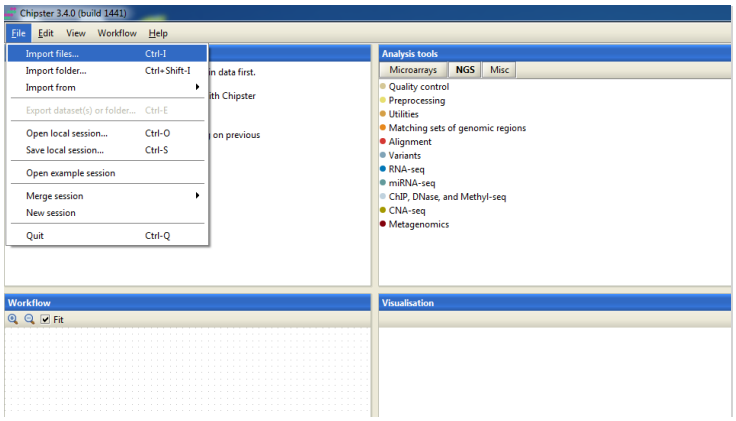
a. Open the Chipster application found on your desktop.
Enter your userID: (you need to ask us!)

Enter your password (you need to ask us!)

b. Let's find the sequence file we will import to Chipster. The file is found on disk **D** under folder "**Course2015**" and in folder "**Course2015-exercise1**", we will use file **myreads.fastq** in this exercise.

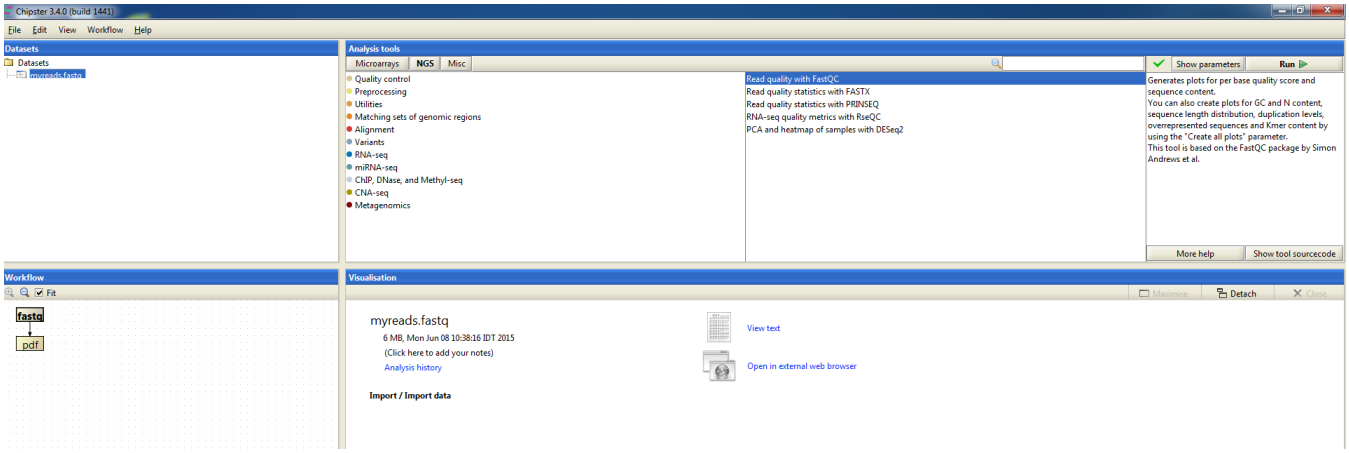
2. Running FASTQC on Chipster

a. Import the fastq file (File->importFIle->select the myreads.fastq file)



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

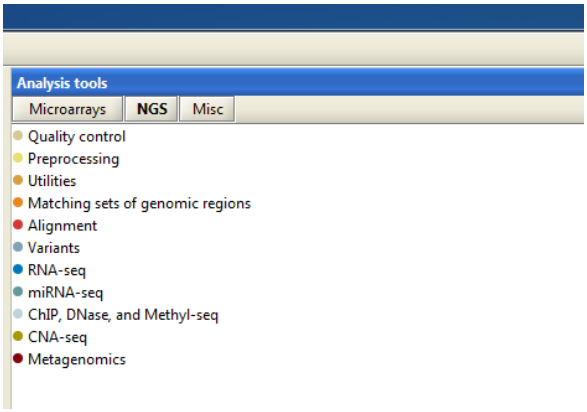
c. Under Visualization panel, click on View text:



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 table below. What is the probability of an error in this 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. Under the Analysis tools, click on **NGS**, 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, a pdf will appear in the left window called Workflow.
h. Double click on the PDF and look at the results.
i. Questions:

   How many sequences does the fastq file contain?

   Is the base quality the same for all the cycles?

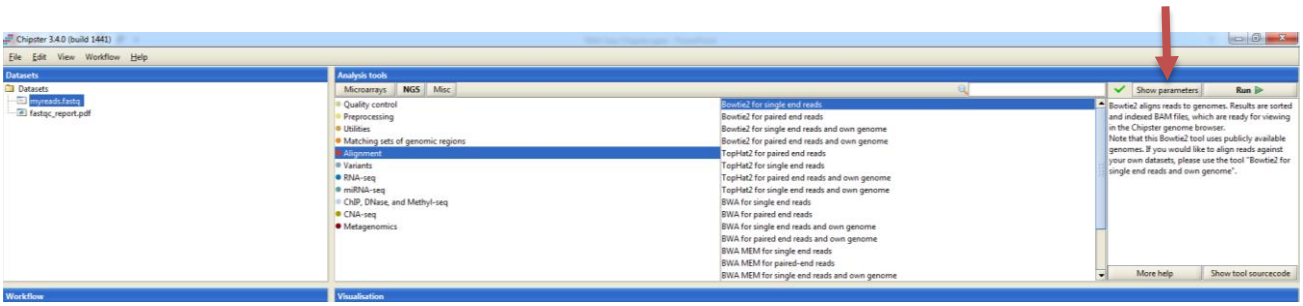
   Do all the cycles have an equal base content?

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 file from the left panel by clicking on it.
b. Under the Analysis tools, click on **NGS**, then on "Alignment" and the 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.

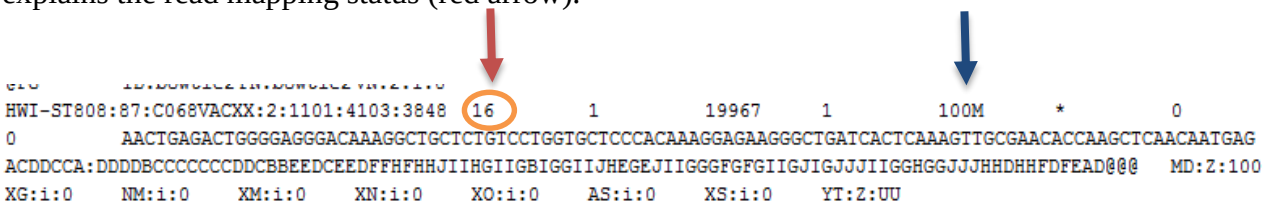


d. Click on the **"Run"** button at the right part. Wait for the results. Open the log file on the left bottom side.
e. Questions:
   What percentage of reads was mapped to the genome? Why are there reads not mapped?
f. Select the bam file in order to view it with the BAM viewer. The BAM file has a binary 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 at:
http://samtools.github.io/hts-specs/SAMv1.pdf , page 4.

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

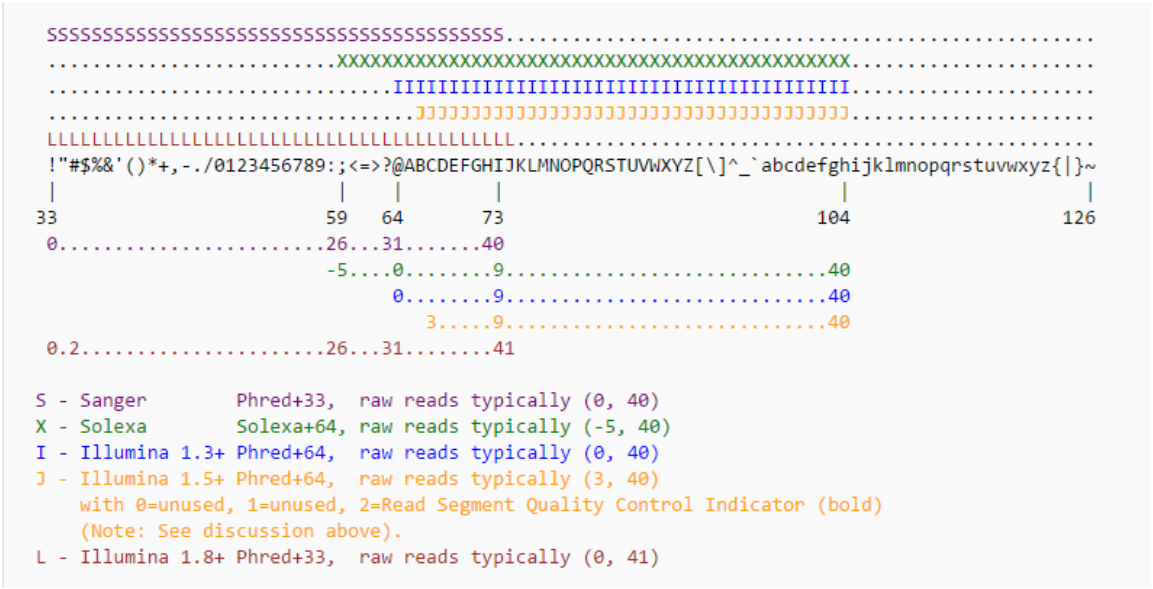
However the 0 flag is missing from the utility. The flag 0 represents unpaired read mapped to the forward strand.

Observe the first and second mapped reads, to which chromosome and location are they mapped?
Do they have mismatches and if yes what is/are it/they?

Hint: Look for the CIGAR (blue arrow).

Supplementary

Converting ASCII Characters to quality values


SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
33 59 64 73 104 126
0.....26...31.....40
-5....0.....9.....40
0.....9.....40
3.....9.....40
0.2.....26...31.....41
S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

THE END.