March, 2016

Learn How to Detect Differentially Expressed Genes from RNA-Seq Data with Chipster

Dena Leshkowitz

Introduction

This workshop is an introduction to the basic principles and knowhow for analyzing RNA-Seq in order to detect differentially expressed genes. We will be using Chipster, an intuitive graphical user interface, to align the reads to a genome (Tophat), quantify the genes (HTSeq) and detect differentially expressed genes (DESeq2).

We will use RNA-Seq sequences derived from acute lymphoblastic leukemia (ALL) precursor B cell lines carrying a chromosome translocation, cells- RS4;11 and SEM, and compare to two precursor B cell lines that lack this translocation NALM6 and REH. This data is private do not distribute.

We will analyse the data with 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.

Following is a list of articles and links that are relevant to the workshop:

http://www.illumina.com/content/illumina-marketing/us/en/techniques/sequencing/rna-sequencing.html

1. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8(9):1765-86. doi: 10.1038/nprot.2013.099. PubMed PMID: 23975260.
2. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105-11. doi: 10.1093/bioinformatics/btp120. PubMed PMID: 19289445; PubMed Central PMCID: PMCPMC2672628.

3. Anders S, Pyl PT, Huber W. HTSeq-a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9. doi: 10.1093/bioinformatics/btu638. PubMed PMID: 25260700.
4. Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, et al. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics. 2011;12:507. doi: 10.1186/1471-2164-12-507. PubMed PMID: 21999641; PubMed Central PMCID: PMCPMC3215701.

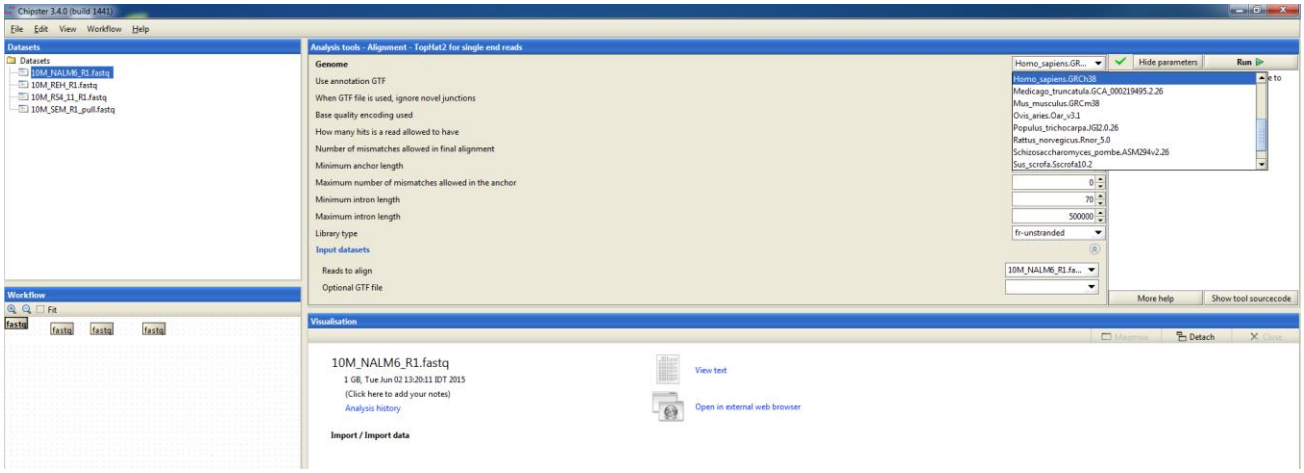
Instructions

1. Accessing the data

Open Chipster through VNC. The Tophat run you started yesterday should be completed. If not call an instructor and ask for help.

2. Understanding Tophat outputs

Each Tophat run produced 4 output files, therefore in total there are 16 files produced.



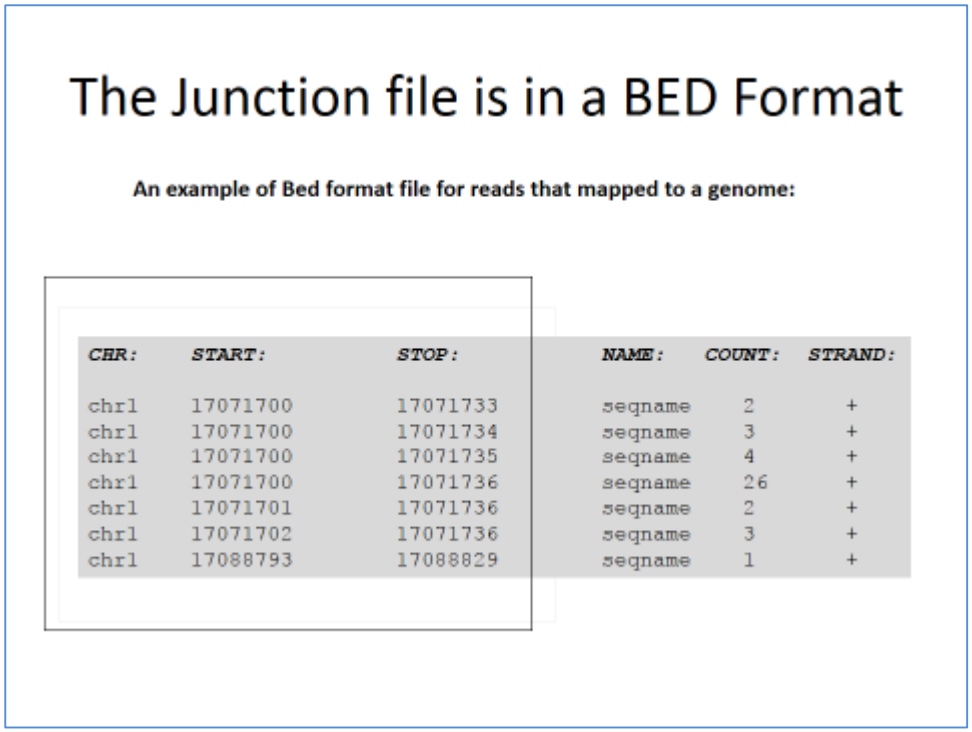
a. Select the tophat-summary.txt that belongs to NALM6 sample (use the workflow arrows to understand which file belongs to which initial fastq file) and view it (in Visualization pane, click on View text).

   Q1: How many reads were in the input fastq file and how many were mapped?

b. Select the juctions.bed relevant to NALM6 sample and view it with the **spreadsheet** option in Visualization pane. See the explanation below for this bed file.

Q2: How many reads support the first junction in the file (JUNC00000001)?

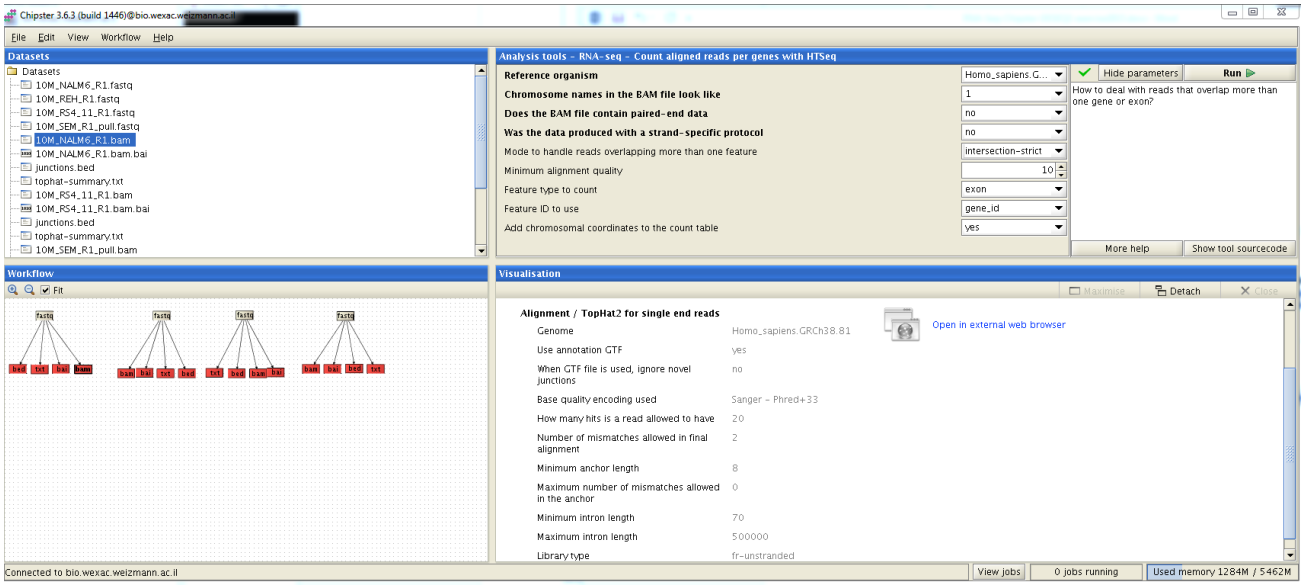
Q3: How many junctions were identified (hint each junction in the file is represented in a separate row)?



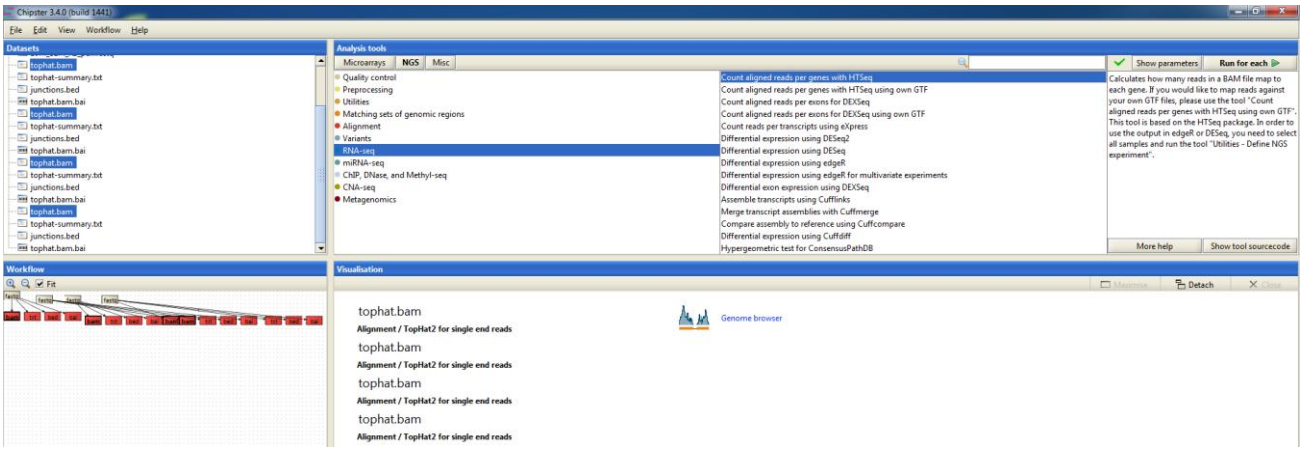
3. Counting the number of reads on genes (HTSeq)

In order to count the number of reads on genes we will run HTSeq.
http://www-huber.embl.de/users/anders/HTSeq/doc/count.html

We want to be strict in our count and consider reads that are uniquely mapped to one gene only and therefore we'll run the program with the option "intersect-strict". To run HTSeq, first select one of the BAM output files, from the Analysis Tools window select the RNA-Seq and the option "count aligned reads per genes with HTSeq". Make sure the genome is Homo_sapiens.GRCh38.81 same as that used for mapping. Repeat this process for all 4 BAM files.



4. Understanding the HTSeq output



View 10M_NALM6_R1.tsv which contains the gene counts for NALM6 sample with the spreadsheet option.

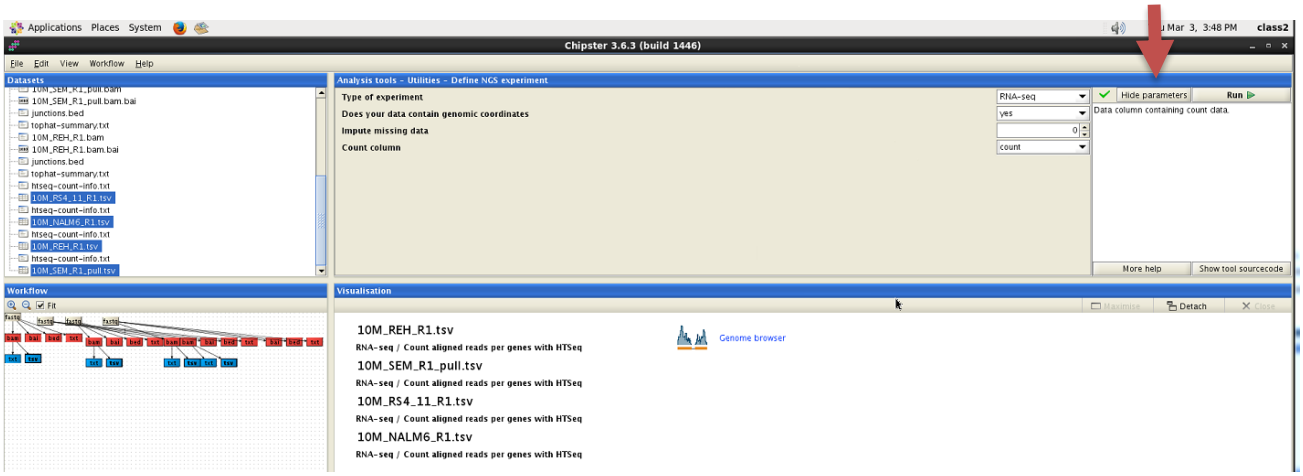
Q4: How many genes are represented in this file?

Note: The Ensemble accessions in the file represent mRNA and sRNA. Since this analysis was for polyA containing transcripts, there are many genes in this file that are not relevant. One should therefore consider using a different annotation file or prefiltering this file.

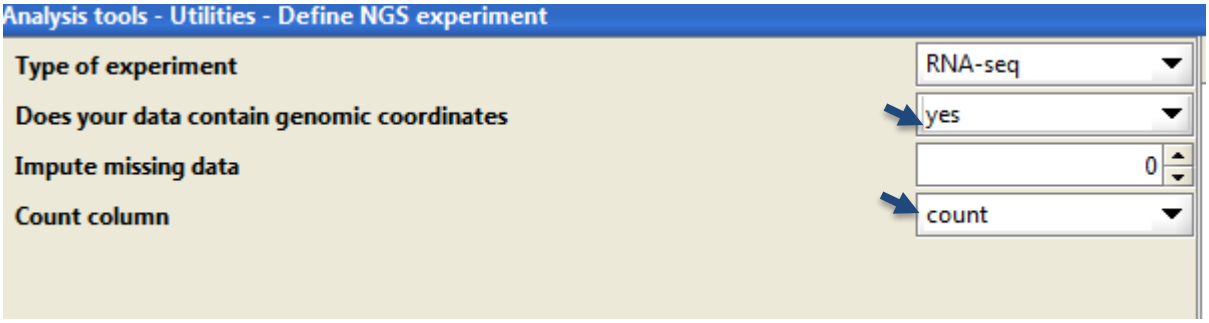
Q5: How many counts (reads) does the first gene have?

5. Merging HTSeq counts to one file

Up to this point in the analysis each sample was analysed separately. We now want to create one matrix containing the gene counts from all the samples. Select the four count files (ending with .tsv) files and from the Analysis tools pane, choose utilities and select the program “Define NGS experiment” found under “utilities”. Click on the “show parameters” (red arrow below).



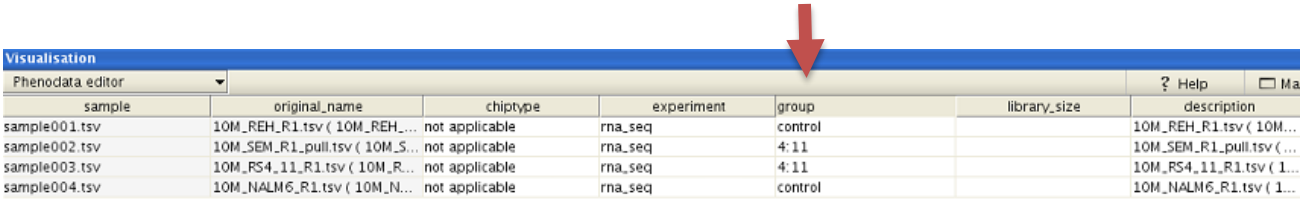
Change the “does your data contain coordinates” to “yes” (from “no”) and the “Count column” to “count”.



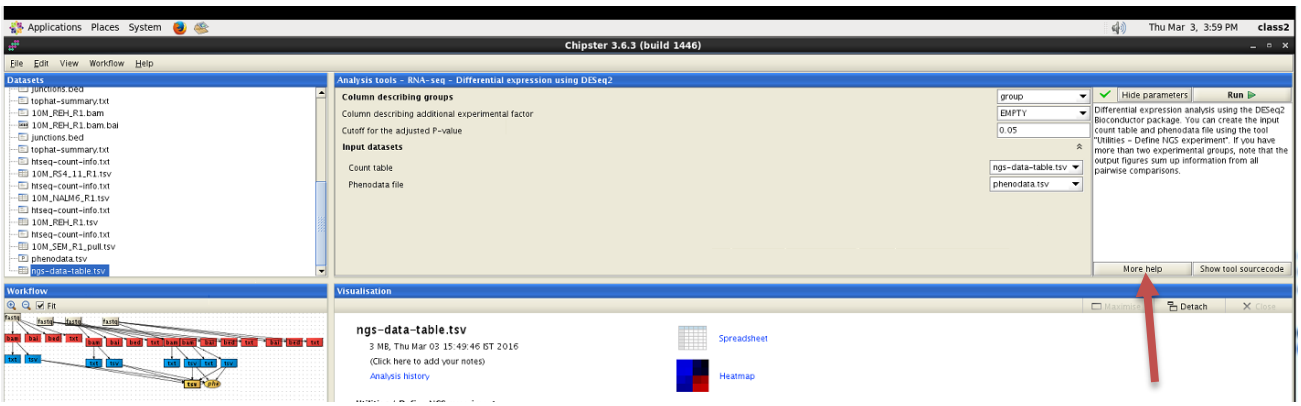
Click on the Run button. Two files should be created. View the merged matrix file - ngs-data-table.tsv with spreadsheet.

6. Running DESeq2

a. In this DESeq2 analysis we would like to identify genes which are statistically significantly differentially expressed between the two sample groups – with and without 4; 11 translocation. Towards this aim we need to define to which group each sample belongs to. Start with editing the the phenodata.tsv file using the "Phenodata editor" from the Visualization window. We need to assign the samples to two groups by editing the "group" column (see below red arrow). Close the file to save your changes.

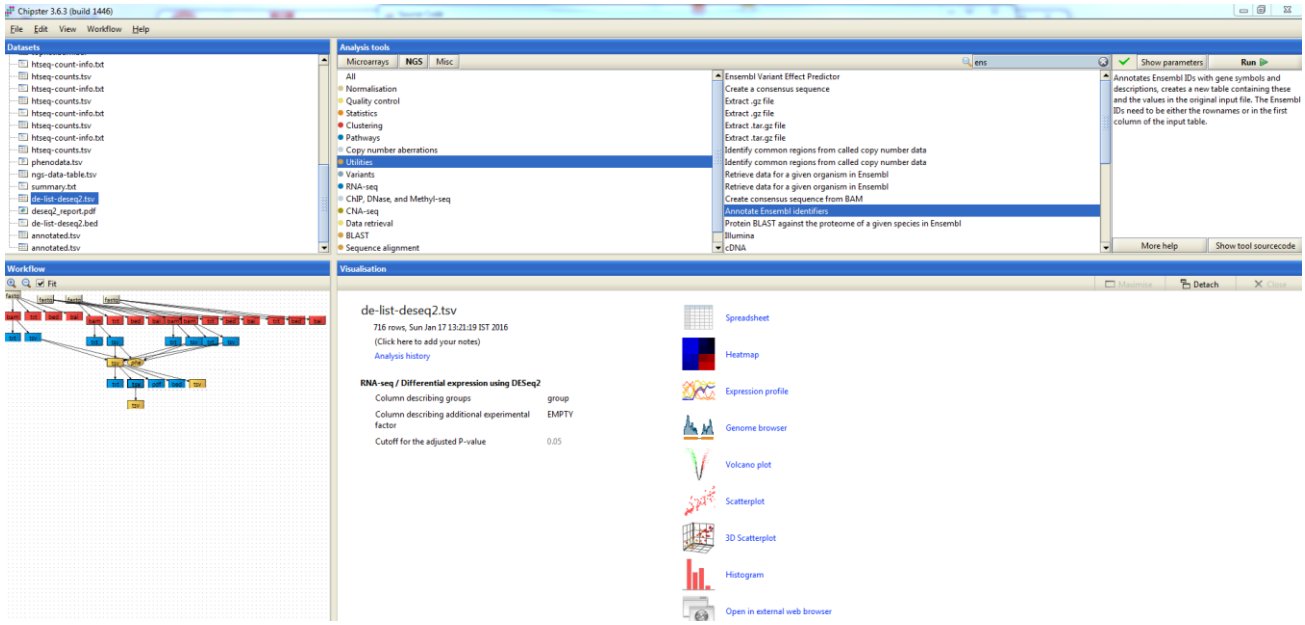


b. We are now ready to run DESeq2. Select the ngs-data-table.tsv file , from the "Analysis tools" pane, choose RNA-Seq and select the "Differential expression using DESeq2". Open the parameters and make sure the phenodata.tsv is used as Phenodata file. Click the run button.



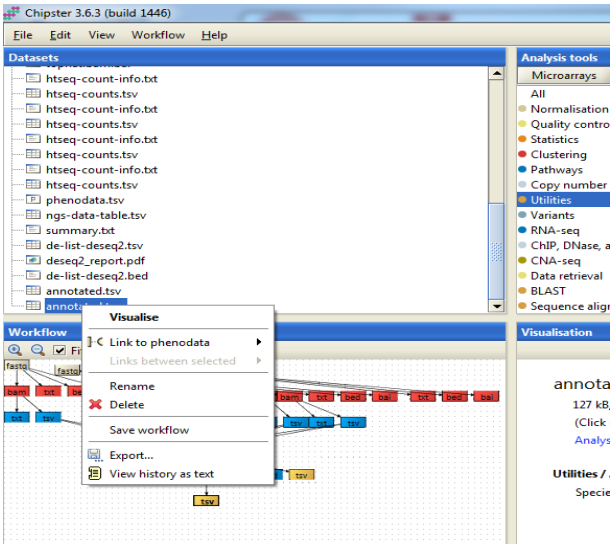
Observe the four DESeq2 outputs. To learn more about the outputs select the "more help" on the DESeq2 tool (red arrow).

c. Adding annotation – gene information to the Ensembl accession number.
Select the file de-list-deseq2.tsv and from the Analysis tools choose utilities and select the "Annotate Ensembl identifiers" program. Click the "Run" button.
A new file named annotated.tsv has been created. Open it with the spreadsheet option from the Visualization window.



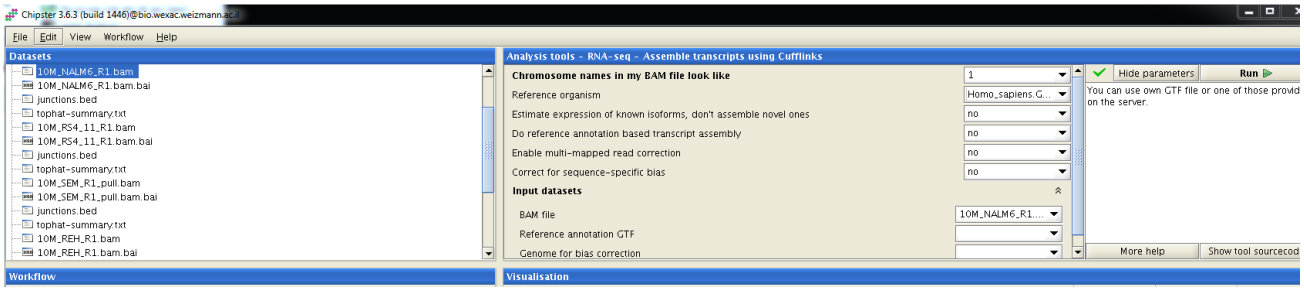
Also open the summary file and answer this question:

Q6: How many genes are differentially expressed? How many are up in 4:11 and how many are down?



7. Running Cufflinks for tomorrows exercise

Cufflinks is a program that allows to assemble the reads to transcripts. We will discuss this program in more detail tomorrow. Since this program takes some time to run we will run it now so we can view the results tomorrow.



Select the NALM bam file, in the Analysis window choose RNA-Seq and select the program "Assemble transcripts with cufflinks". Make sure the genome is appropriate one. We will run Cufflinks with the option that provides cufflinks with an annotation file. Therefore, change the option "Do reference annotation based transcript assembly" to "yes". In addition run Cufflinks without reference annotation. Two jobs should be running now.

Good work!

THE END