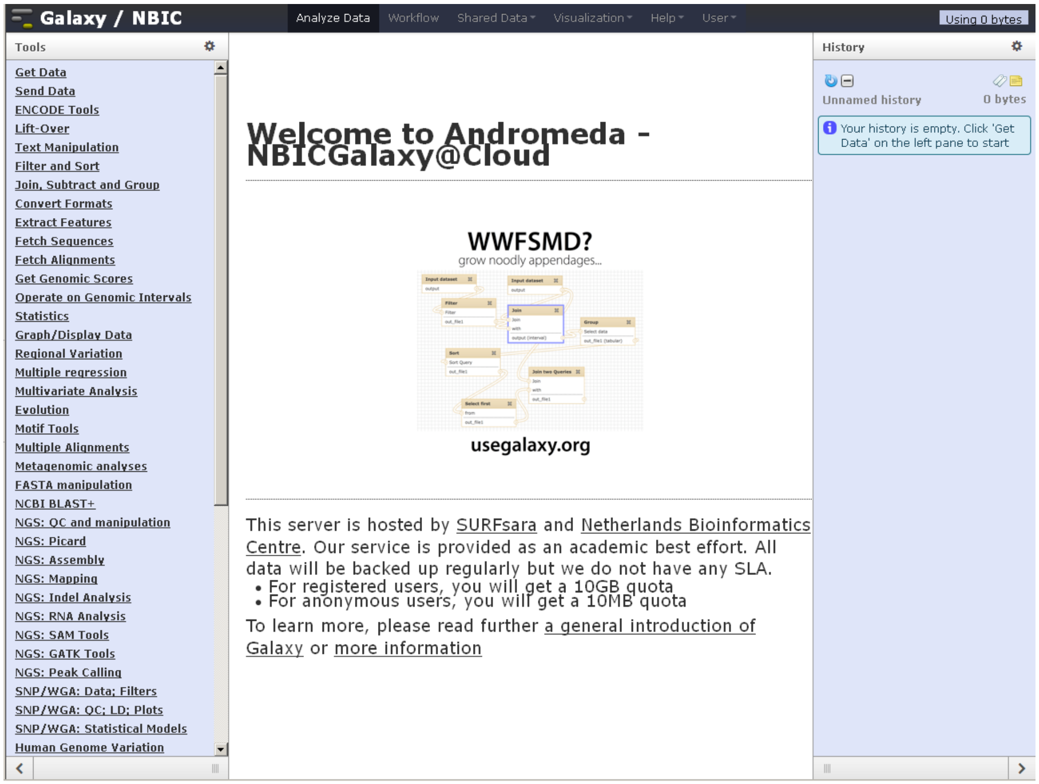
FOS course
RNA-seq using Galaxy, Practical assisted by Eleonora de Klerk
Workflow generated by Jeroen F.J. Laros, edited by Eleonora de Klerk

Galaxy Penn State's Galaxy is a useful way of wrapping many command line modules together in a user-friendly interface. Galaxy is a web-based system, therefore you do not need to install any application.
What you need is just to open your web-browser (firefox, IE, etc.) and access the galaxy server hosted at page (http://galaxy.nbic.nl/).

Log in
1. Open a browser and go to http://galaxy.nbic.nl/
2. Register to gain access to data libraries and workflows.
• Click on "User", then on "Register" in the top bar.

When logged in, you can save your workflow and execute the entire workflow on a new dataset without manually executing each individual step. You can also easily share these workflows with others.

When you open the Galaxy page, you will see three panels as shown in the figure below.
• Tool panel: here you find a list of tools provided by Galaxy
• Interface panel: a configuration interface of the tool you select from the tool panel
• History panel: overview about the data analysis steps you have performed.



Your history is empty. Click 'Get Data' on the left pane to start

Welcome to Andromeda - NBICGalaxy@Cloud

This server is hosted by SURFsara and Netherlands Bioinformatics Centre. Our service is provided as an academic best effort. All data will be backed up regularly but we do not have any SLA.
- For registered users, you will get a 10GB quota
- For anonymous users, you will get a 10MB quota

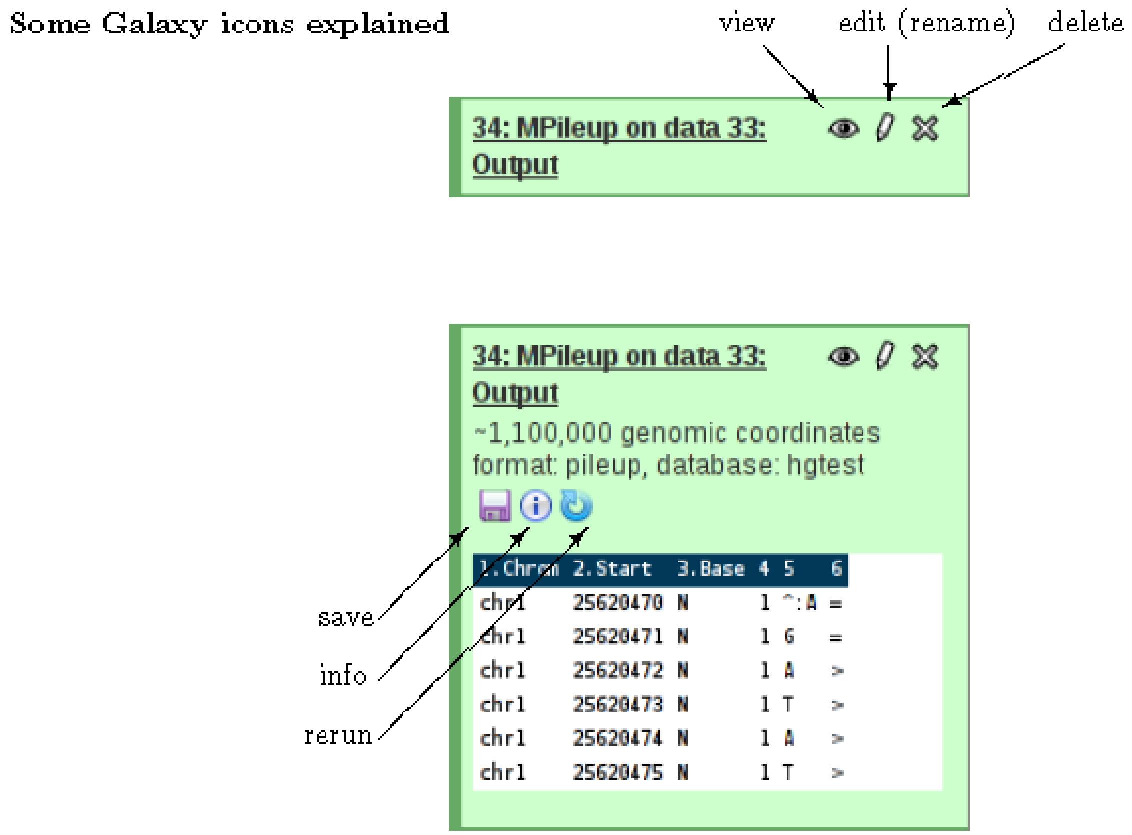
To learn more, please read further a general introduction of Galaxy or more information

Tool panel **Interface panel** **History panel**

Availability and examples The tools used in these exercises are all free for download, including Galaxy itself (http://galaxy.psu.edu/), GSNAP for alignment, SAMtools and Cufflinks for expression analysis.

Note on test data Data used in this practical is test data and not full size files. This is to reduce the time needed to run each step and make this analysis possible within the time permitted.

Some Galaxy icons explained.



34: MPileup on data 33: Output

34: MPileup on data 33: Output

~1,100,000 genomic coordinates
format: pileup, database: hgtest

1.Chrom	2.Start	3.Base	4	5	6
chr1	25620470	N	1	^:A	=
chr1	25620471	N	1	G	=
chr1	25620472	N	1	A	>
chr1	25620473	N	1	T	>
chr1	25620474	N	1	A	>
chr1	25620475	N	1	T	>

Exercise 1: Expression analysis.

The input data is a small selection of reads that should align mostly to a small region on the human genome.
After alignment, you can do expression analysis and visualize the data on the genome.
The data we will use are shared data that have been uploaded by other users.
To import the data in your history:
• Click on "Shared Data", then on "Data Libraries".
• Click on "GAPSS3 RNA".
• Select all datasets and click "Go".
Click on "Analyze Data" to start the analysis.

1- Rename the history in the history panel as you prefer, by clicking on ''unnamed history''.
2- Look at the files you have just imported. What are the files named reads1 and reads2? What is the difference between a reference genome and the file named genes.gtf?

3- Look at one of the FASTQ files (fq) (by clicking the file name you will see the first lines, by clicking on the eye-icon you can see the first megabyte on the interface panel). Each read is represented by four lines: a header, the read itself, a + and the quality scores.
If you would have a file with 6.000.000 lines, how many raw reads would you have?

Try to answer these questions before going on! If you can't, ask me.

Do some standard QC on the FASTQ files:
• NGS: QC and manipulation: Fastqc:ReadQC
Run on reads 1.fq, use "FastQC read1" as title.
• NGS: QC and manipulation: Fastqc:ReadQC
Run on reads 2.fq, use "FastQC read2" as title.

View the output data (eye-icon).

When looking at the output of the QC steps, you will notice a lot of warnings. These warning arise partially from the fact that we work with a very small artificial dataset.

1. Which base-position has the lowest quality score. Do you know why?
2. What is the average quality per read for the majority of the reads?
3. What is the read length?
4. What is the total number of sequences?

Align the reads to the human reference genome:
• Go to NGS:RNA analysis -> GSNAP

Select Fastq as input format, reads 1.fq as input dataset, select “Paired Reads” and use reads 2.fq as the second dataset, and keep the Illumina default for read orientation. For this practical we will use default settings and the human reference genome. Mapping might take some time. Be patient (you need to wait for the outputfile to proceed).

- Investigate the output of GSNAP (the SAM file).
  If you scroll to the right, you will notice that all the information from the input files (the FASTQ files) are still present in the sam file.

  If you want to know what is in every field, you can look at: http://samtools.sourceforge.net/SAM1.pdf

Question: To which chromosome are most of the reads mapped?

• Convert the output file (SAM) into BAM.
Go to NGS: SAM Tools: SAM-to-BAM: Use the output of GSNAP (SAM) for input.

The BAM file is a compressed version of the SAM file. As all original data is present and the file is rather small compared to the input, this is an ideal dataset to store for later use.

Question: Why you cannot view the BAM file?
Question: How large is the BAM file and how large is the SAM file?

Expression analysis:
Perform the expression analysis running cufflinks on the mapped reads.
• NGS: RNA Analysis:→Cufflinks
Use the BAM file for input instead of the SAM file, select “Use Reference Annotation” and use genes.gtf for annotation. You should get two output files.

Edit the attributes of the transcript expression dataset to change the data type from tabular to xls (hint: do not use ‘Convert format’ but Datatype). If the job is not finished after few minutes, refresh the history panel.
Download the transcript expression in .xls data and open it with Excel.

Question: How many genes are expressed in your data?
Question: Which gene has the highest coverage?

Data visualization:

Create now a wig file. To create a wiggle file we need to convert the bam file into a pileup file first. After that we can convert the pileup into wiggle.

- NGS:SAM Tools:Generate pileup from BAM (keep default settings).
- Convert Format: mpileup2wig

Look at the wiggle file on UCSC. Look for the most expressed genes in your cufflink list. Practice zooming in in different exons.

Question: In which part of the gene (Cryab) do you see coverage? Is the coverage equal in the last exon compared to the middle exon? How many reads are present in the mostly covered exon?

Exercise 2: Novel transcript detection

Rerun the Cufflinks step, but this time select "Use reference annotation as guide". Before you can do that, consider that when visualizing more than one dataset in the UCSC genome browser, each dataset must have an identifier.
To do this, we need to add a header to the assembled transcript files. These headers (reference.txt and assisted.txt) are provided in the data library.
Add the headers to the corresponding datasets:

• Text Manipulation: Concatenate datasets: Select reference.txt and add the assembled transcripts dataset from the first Cufflinks analysis. Now rename the output file just created in order to easily remember what kind of data it is (ex. Assembled_transcript_reference).
• Text Manipulation: Concatenate datasets: Select assisted.txt and add the assembled transcripts dataset from the second Cufflinks analysis. Change name as before (ex. Assembled_transcript_assisted).

You can now upload the datasets to the UCSC genome browser.
First download the files on your Desktop, then upload them into UCSC genome brower. Go to the same session where the wiggle file is uploaded, and enter the first file. The file is a .gtf file of ~80MB, uploading the file will take some time. Once you upload the first file, change name and description by clicking on the field Name – User Track. Both files have as default User Track and User Supplied Track as name and description. Therefore you need to change the name and the description before you upload the second file, or it will be overwritten.

Go now to the genome browser and look for the following specific genes:

Question: What do you see in the SIK2 gene?
Change the display type of the two custom tracks from "dense" to "pack". You can now see the separate transcripts in more detail.
Question: What is the difference between transcript "NM015191" and "NM015191ext"?

Question: In the LAYN gene, what does “CUFF.303.2” represent?

Exercise 3: Workflows

Workflows can be extracted from a history and saved in order to re-run an analysis. In this case we will delete our history and run a pre-existing workflow with the same analysis we just did.

• First, clear history.
– In the history panel, click “Options”, then “Delete”.
• Select the Data Library, as explained in Exercise 1.
Import a workflow.
• Click on “Shared Data”, then on “Published Workflows”.
• Select “GAPSS3 RNA”.
• Click on “Import workflow”.
• Click on the workflow button and select the imported “GAPSS3 RNA” workflow. Click “Run”.
• Now click “Run workflow” to execute the workflow