



LEIDEN UNIVERSITY MEDICAL CENTER

RNA-seq using Galaxy

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics



Sequencers.



Figure 1 : HiSeq 2000.



Figure 2 : Ion proton.

High throughput, accurate,
cheap.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
 - QC.
 - Data cleaning.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
 - QC.
 - Data cleaning.
2. Alignment.
 - Use a specialised (RNA) aligner.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
 - QC.
 - Data cleaning.
2. Alignment.
 - Use a specialised (RNA) aligner.
3. Expression (gene, transcripts) analysis.
 - Known transcripts.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
 - QC.
 - Data cleaning.
2. Alignment.
 - Use a specialised (RNA) aligner.
3. Expression (gene, transcripts) analysis.
 - Known transcripts.
4. Transcript assembly.
 - New transcripts, alternative splicing, etc.

Combining tools in a pipeline.

```
1 bwa aln -t 8 $reference $i > $i.sai
2 bwa samse $reference $i.sai $i > $i.sam
3 samtools view -bt $reference -o $i.bam $i.sam
```

Listing 1 : Shell script.

Combining tools in a pipeline.

```

1  bwa aln -t 8 $reference $i > $i.sai
2  bwa samse $reference $i.sai $i > $i.sam
3  samtools view -bt $reference -o $i.bam $i.sam

```

Listing 1 : Shell script.

```

1  %.sai: %.fq
2    $(BWA) aln -t $(THREADS) $(call MKREF, $@) $< > $@
3
4  %.sam: %.sai %.fq
5    $(BWA) samse $(call MKREF, $@) $^ > $@
6
7  %.bam: %.sam
8    $(SAMTOOLS) view -bt $(call MKREF, $@) -o $@ $<

```

Listing 2 : Makefile.

Overview.

Data intensive biology for everyone.

- Open source.
- Web based.
 - No installation required.

<http://galaxy.psu.edu/>

Overview.

Data intensive biology for everyone.

- Open source.
- Web based.
 - No installation required.

Wrapper for command line utilities.

- User friendly.
 - Point and click.
- Workflows.
 - Save all the steps you did in your analysis.
 - Rerun the entire analysis on a new dataset.
 - Share your workflow with other people.

<http://galaxy.psu.edu/>

Global overview of the GUI.

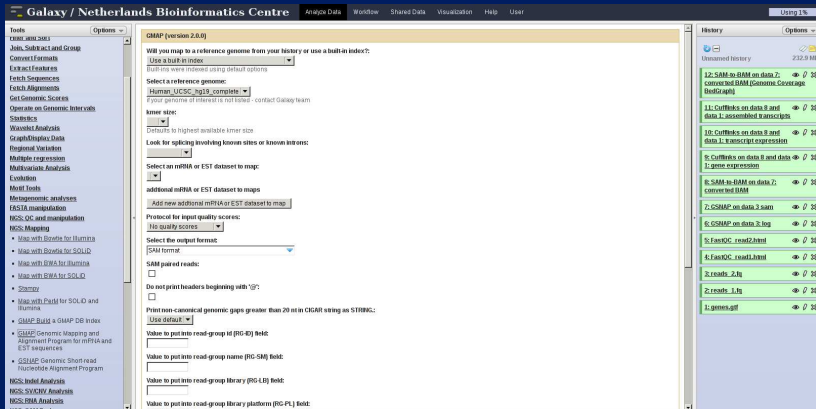


Figure 3 : Galaxy panels.

The Galaxy tool menu.

Lots of tools installed, especially for NGS data analysis.



Figure 4 : Collapsed tool menu.

The Galaxy user interface

The Galaxy tool menu.

Lots of tools installed, especially for NGS data analysis.



Figure 4 : Collapsed tool menu.

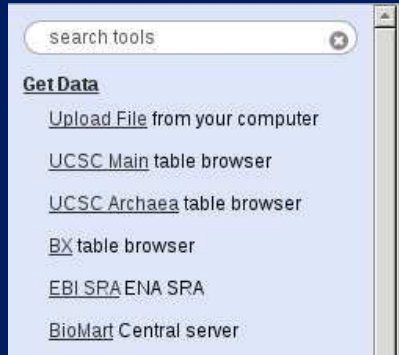


Figure 5 : Tool menu.

Selecting a tool.

The screenshot shows the Galaxy web interface with the 'Upload File' tool selected. The interface includes a top navigation bar with 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Cloud', 'Help', and 'User'. A left sidebar lists various tools under categories like 'Get Data', 'Text Manipulation', and 'Send Data'. The main content area is titled 'Upload File (version 1.3.0)' and contains the following sections:

- File Format:** A dropdown menu set to 'Auto-detect' and a note: 'Which format? See help below'.
- File:** A text input field containing 'local:/data/mouse_galaxy' and a 'Browse...' button. A note below states: 'FTP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator)'.
- URL list:** A large empty text area for specifying a list of URLs.
- Files uploaded via FTP:** A table with columns 'File', 'Size', and 'Date'. Below the table is a note: 'Please create or log into a Galaxy account to view files uploaded via FTP. This Galaxy server allows you to upload files via FTP. To upload some files, log into the FTP server at seegalaxy.org using your Galaxy credentials (email address and password)'.
- Convert spaces to tabs:** A checkbox labeled 'Yes' which is currently unchecked. A note below says: 'Use this option if you see weird characters by hand'.
- Genomic:** A dropdown menu with the text 'Additional Spaces Are Below'.
- Buttons:** A blue 'Upload' button.
- Auto-detect:** A section explaining that the system will attempt to detect file formats like FASTA, FASTQ, BAM, etc. It notes that if a file is not detected properly, it may be a compressed file that needs to be decompressed.
- ABI:** A section explaining that an ABI sequence file is in 'ABI' format with a '.abi' file extension and that the 'File Format' must be manually selected.
- Asi:** A section explaining that an Asi pairwise alignment format consists of three lines per alignment: a summary line and two sequence lines, separated by blank lines.

On the right side of the interface, there is a 'History' panel showing 'Unshared history' with '0 bytes' and a message: 'Your history is empty. Click "Get Data" in the left pane to start'.

Figure 6 : Uploading a file.

The history panel.

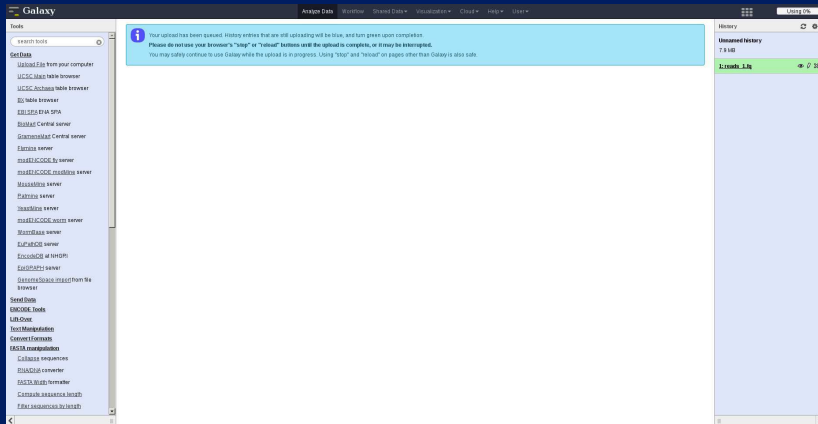


Figure 7 : Uploading completed.

History item icons.

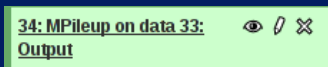


Figure 8 : Collapsed history item.

Every history item has some basic functions:

- Eye: view.
- Pencil: edit (rename).
- Cross: delete.

History item icons.




Figure 8 : Collapsed history item.

Every history item has some basic functions:

- Eye: view.
- Pencil: edit (rename).
- Cross: delete.




Click on the title for a more detailed view.

History item icons.

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	>

Figure 9 : History item.

- Diskette: save.
- Blue looping arrow: rerun.

Viewing the data (eye icon).

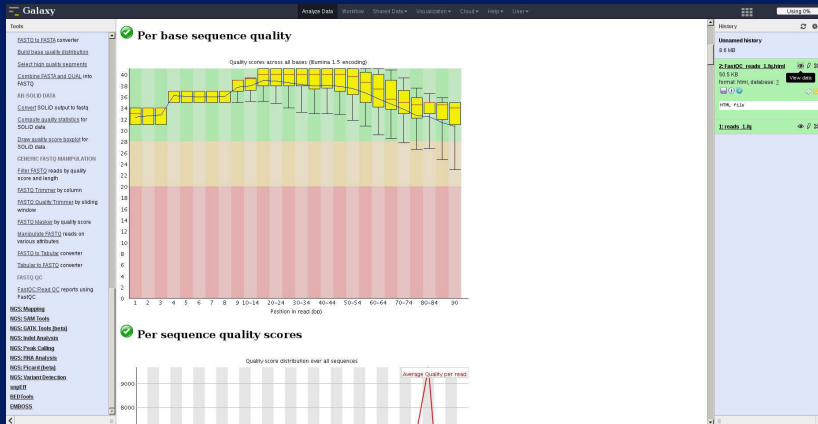


Figure 10 : See your output in the tool panel.

Extracting a workflow.



Figure 11 : Extract workflow.

Once you have finished an analysis, you can save it as a workflow.

- Store for later use.
- Share with others.

Extracting a workflow.



Figure 11 : Extract workflow.

Once you have finished an analysis, you can save it as a workflow.

- Store for later use.
- Share with others.

Click on the “gear” icon in the history panel.

Modifying workflows.

The screenshot displays the Galaxy workflow editor interface. On the left, a 'Tools' sidebar lists various bioinformatics tools, including 'FastQC: Read QC'. The main workspace, titled 'Workflow Canvas | QC', shows a workflow diagram with three 'FastQC: Read QC' tool instances connected in a sequence. The first instance takes an 'input dataset' as input. The second instance takes the output of the first and a 'Library to clip' as input. The third instance takes the output of the second and a 'FASTQ File' as input. A 'Filter FASTQ' tool is also present, which takes a 'FASTQ File' as input and outputs an 'output_file'. The right-hand side of the interface shows the 'Details' panel for the selected 'FastQC: Read QC' tool, including its version (0.52), a description of its input and output, and options to edit step actions and attributes.

Figure 12 : Create or edit your workflows.

Acknowledgements:

Wibowo Arindrarto
Irina Pulyakhina
Mateusz Kuzak
Hailiang Mei
Peter-Bram 't Hoen
Johan den Dunnen

<http://galaxy.nbic.nl/>