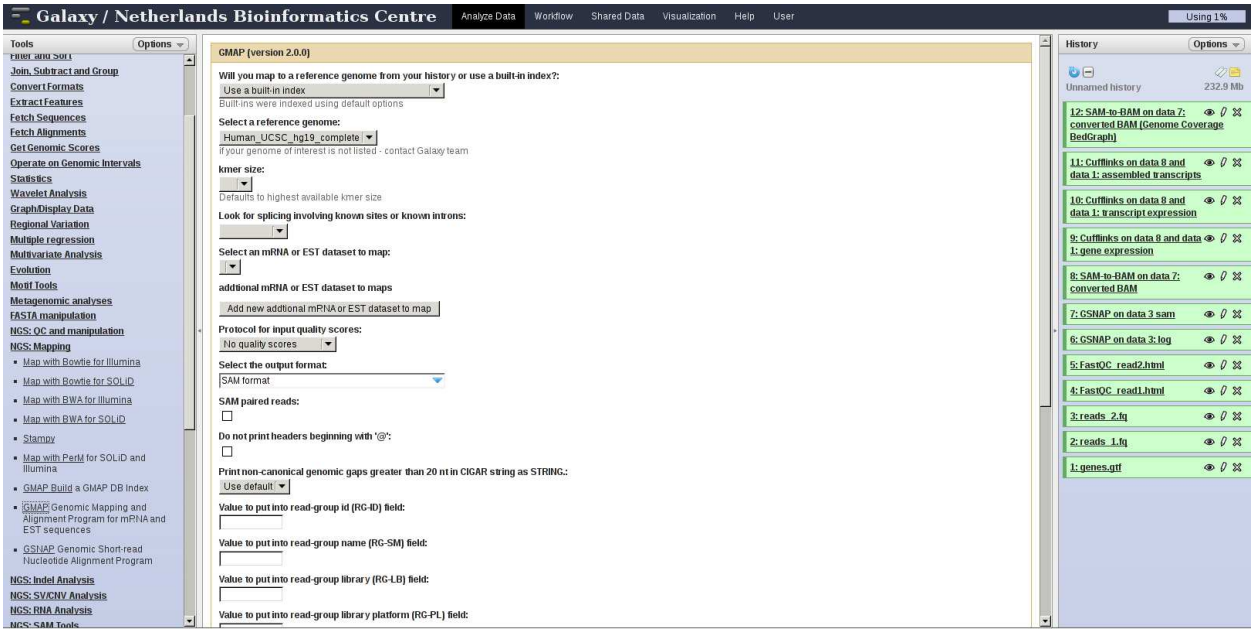
NGS Introduction Course.
Hands on workshop: Next generation sequence data analysis.
Instructors: Martijn Vermaat, Leon Mei, Frank Sleutels, Rutger Brouwer, Jeroen Laros.
Leiden Genome Technology Center
Leiden University Medical Center, The Netherlands

Introduction In this workshop we will first show you a typical analysis done by a bioinformatician. In the first session, we used the Linux command line executables to align to a known reference genome and call SNPs, reporting as a tab-delimited file. We will now show how to do this same analysis with a more biologist friendly tool: Penn State's Galaxy (Blankenberg et al. 2007, PMID 17568012). We will then show a second application in Galaxy: CAGE (expression) analysis reported as a tab-delimited file and viewed in the UCSC Genome Browser.

Galaxy Penn State's Galaxy is a useful way of wrapping many command line modules together in a user-friendly GUI. Galaxy is a web-based system so that you do not need to install any client side application. What you need is just to open your favourite webbrowser (firefox, IE, etc.) and access the galaxy server hosted at page (`http://galaxy.nbic.nl/`). 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: this is a configuration interface of the tool you select from the tool panel
- History panel: here you can have an overview about the data analysis steps you have performed. And you can also extract a workflow from your history and share it with other registered users on the Galaxy server.

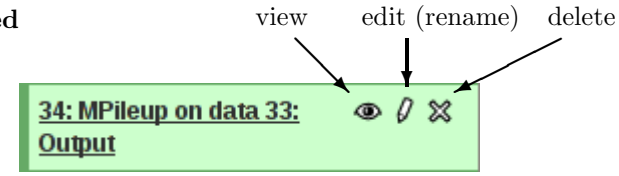


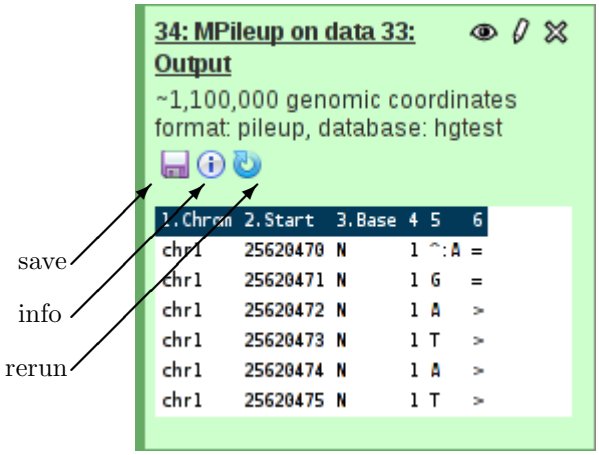
Tool panel Interface panel History panel

Availability and examples The tools used in these exercises are all free for download, including Galaxy itself (`http://galaxy.psu.edu/`), GMAP/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





Preparations.

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.
   - Choose a unique account name e.g., your e-mail address.
   - Choose a password.
   - Choose a public name (or leave blank).

Exercise 1: expression analysis.

The input data is a small selection of reads that should align to the human chromosome 11. After alignment, you can call SNPs and small indels.

Import the data we will use:

- In the "Shared Data" tab click on "Data Libraries".
- Click on "Practical_var".
- Select "reads_1.fq" and "reads_2.fq" and click "Go".

Click on "Analyze Data" to start the analysis.

Do quality control on the input files:

- *NGS: QC and manipulation: Fastqc: Fastqc QC*: run on the `reads_1` data. Choose "FastQC on reads 1" as title.
- Repeat for `reads_2`.

Check the FASTQ file format and align to the reference sequence:

- *NGS: QC and manipulation: FASTQ Groomer*: run on the `reads_1.fq` data. Choose "Sanger" for the quality scores type. (Question: Did you retain all sequences?).
- Repeat for `reads_2`.
- *NGS: Mapping: Stampy*: Choose "Paired-end" and use the groomed FASTQ data sets ("FASTQ Groomer on data 1" as Forward, "FASTQ Groomer on data 2" as Reverse. Align to `hg19` – otherwise leave defaults (Question: How many sequences were aligned?).

Use SAMtools to call SNPs:

- *NGS: SAM Tools: SAM-to-BAM*: input is your Stampy output.
- *NGS Taskforce: LUMC - GAPSS v3: MPileup*: input is the sorted BAM data, choose "hg19" as reference..
- *NGS Taskforce: LUMC - GAPSS v3: BCFVariantCalling*: input is the MPileup Output data (be careful not to use the Status data).
- *NGS Taskforce: LUMC - GAPSS v3: BCFToVCF*: input is the BCF Output data.
- *NGS Taskforce: LUMC - GAPSS v3: VCFUtilsVarFilter*: input is the VCF data.

Lets take this a step further and also annotate your variants with SeattleSeq:

- *NGS Taskforce: LUMC - GAPSS v3: Seattle-seq Annotation*: input is the VCF file. Enter your e-mail address.
- *NGS Taskforce: LUMC - GAPSS v3: Seattle-seq Annotation*: input is the VCF file. Select "InDel" as type of variants. Enter your e-mail address.

Lets save this for future use and look at the data later:

- Click the "save" button to save the SeattleSeq outputs (will save by default to your desktop).
- Open the file with Excel.

Exercise 2: CAGE (Cap Analysis of Gene Expression) analysis
CAGE is a laboratory technique to sequence the 5' end of RNAs. This practical will use a small test data set from a mouse CAGE project. You will convert it to a sanger quality FASTQ file, trim the first basepair (lower quality), align to the full mouse genome, and view this data in a tab-delimited format and in the UCSC genome browser.

Note: (first clean history, under "Options" select "Delete").

Upload all the data we will use:

- Click on "Data Libraries" in the "Shared Data" tab.
- Click on "Practical_CAGE".
- Select "small_CAGE_test_data.scarf" and click "Go".

Click on "Analyze Data" to start the analysis.

First convert the input to FASTQ:

- NGS Taskforce: LUMC - GAPSS v2: GAPSS - SCARF to FASTQ: run on the input.

Check the FASTQ file format:

- NGS: QC and manipulation: FASTQ Groomer: run on the new FASTQ file.

Clean up the data

- NGS Taskforce: LUMC - GAPSS v2: GAPSS Remove 1st bp.
  - Click on the eye to view data.

  - This program has a bug, it lost the data format: tell Galaxy this file is in fastqsanger format by clicking on the pencil and under "Change data type" select "fastqsanger" and save.

Map to the mouse genome build 9.

- NGS Taskforce: LUMC - GAPSS v2: Map with Bowtie for Illumina: use as input your edited FASTQ data, align to `mm9`, deselect the output in SAM format, otherwise leave defaults.

Convert to an in-house alignment format called IGF:

- NGS Taskforce: LUMC - GAPSS v2: GAPSS Bowtie to IGF.
- Rename as `CAGE_IGF` by clicking on the pencil icon.

Make a tab delimited report file:

- NGS Taskforce: LUMC - GAPSS v2: GAPSS Make regions, input is the IGF file.
- To eliminate gaps of 100bp lets run NGS Taskforce: LUMC - GAPSS v2: GAPSS Compress regions, gap size 100.
- Save the compressed regions file to your desktop.
- Open with Excel.
- Sort on the column "#_tags_in_region" (under options when sorting indicate range has column labels) to find the most significant region (i.e. with the most number of tags in a region).

Lets view the data in UCSC:

- NGS Taskforce: LUMC - GAPSS v2: GAPSS IGF to WIG, make sure to use the file `CAGE_IGF`, use Cutoff size 2.
- Save this file to your desktop as `wiggle.gz`.
- Go to the UCSC genome browser.
- Click "Genome Browser".
- Select the mouse genome, build `mm9`.
- Click "add custom tracks" and select the file `wiggle.gz` from your desktop.
- Check out the most significant region from your sorted Excel data (question: does this make sense? (i.e. does it align to the 5' end of a gene?) What about the second region?).

Exercise 3: Workflows

Workflows can be extracted from a history and saved in order to re-run an analysis.

- First, clear the history again.
- In the "Shared Data" tab, select "Published Workflows".
- Click on the "Practical_var" workflow, click "Import workflow".
- Repeat for the "Practical_SAGE" workflow.
- Select one of the Data Libraries, as explained in Exercise 1 and 2.
- Click on the workflow button and select the appropriate workflow. Click "Run".
- Now click "Run workflow" to execute the workflow.