



LEIDEN UNIVERSITY MEDICAL CENTER

Next generation sequence data analysis

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics



NGS data analysis

In this practical, we will do two types of analyses:

- Exome sequencing data analysis.
- CAGE expression analysis.

NGS data analysis

In this practical, we will do two types of analyses:

- Exome sequencing data analysis.
- CAGE expression analysis.

Cap analysis gene expression (CAGE):

- Isolate mRNA.
- Extract a small part of the beginning of the transcript.
- Sequence.
- Analysis consist mainly of counting mapped sequences in a specific region.

Region files

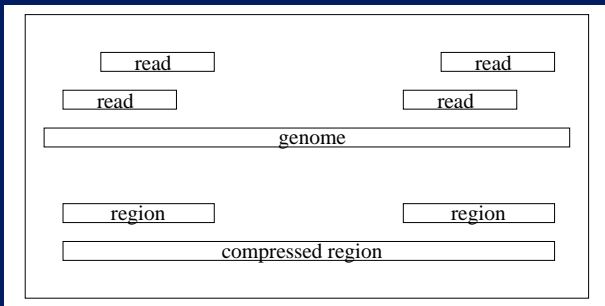


Figure 1 : Region files explained.

Terminology:

- Overlapping reads are combined to a “region”.
- Close by regions are combined to a “compressed region”.

Galaxy

We will use the Galaxy graphical user interface:

- 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>

The screenshot displays the Galaxy web interface. At the top, there is a navigation bar with the 'Galaxy' logo and several menu items: 'Analyze Data', 'Workflow', 'Data Libraries', 'Admin', 'Help', and 'User'. On the left side, there is a 'Tools' sidebar with a search filter and a list of tool categories including 'Filter and Sort', 'Statistics', 'Genomics', and 'NGS: QC and manipulation'. The main area shows the configuration for the 'GAPSS - FASTA to FASTQ' tool. It includes a 'FASTA File to convert:' dropdown menu, a 'score:' input field with a value of '0.8', and an 'EXECUTE' button. Below the configuration, there is a 'What it does' section with explanatory text and a URL: <http://www.igt.hiGAPSS/>. On the right side, there is a 'History' panel with an 'Options' dropdown and a message: 'Your history is empty. Click 'Get Data' on the left pane to start.'

Figure 2 : Galaxy main user interface

<http://galaxy.nbic.nl>

Galaxy

MPileup

Compute genotype likelihoods:
 True False
Compute genotype likelihoods and output them in the binary call format (BCF).

Output uncompressed BCF:
 True False
Similar to the Genotype parameter, except that the output is uncompressed BCF, which is preferred for piping.

Input :

Generate BCF or pileup for one or multiple BAM files. Alignment records are grouped by sample identifiers in @RG header lines. If sample identifiers are absent, each input file is regarded as one sample.

Generated By:
LUMC Interface Generator (0.1)
2011-09-03T14:29:36.793452Z

Based On:
RDF Definition of "MPileup"
2011-09-02T16:17:29.010890Z

Figure 3 : User friendly interface with Galaxy

Galaxy icons

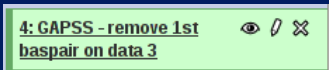


Figure 4 : Default view.

The top icons:

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

Click on the title for a more detailed view.

Galaxy icons



Figure 5 : Detailed view.

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

Outline of the practical

1. Create a SNP file from a FASTQ file.
2. CAGE analysis.
3. Workflows.
 - Rerun the SNP or CAGE analysis with no effort.

See the handouts for the practical.

https://humgenprojects.lumc.nl/trac/humgenprojects/wiki/NGS_workshop