Big data: pipeline tools

Instructors: Martijn, Dijkstra, Pieter Neerincx, Freerk van Dijk
Genomics Coordination Center, University Medical Center Groningen, The Netherlands

This workshop is based on a course developed by: Michiel van Galen, Jeroen Laros, and Netherlands Bioinformatics Center.
Leiden Genome Technology Center, Leiden University Medical Center, The Netherlands

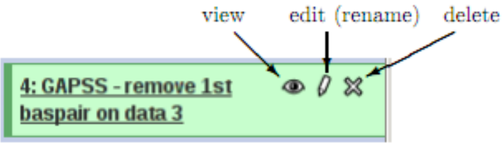
Introduction In this workshop we will show you a typical analysis done by a bioinformatician.
We will do this same analysis with a biologist friendly pipeline tool: Penn State's Galaxy (Blankenberg et al. 2007, http://www.ncbi.nlm.nih.gov/pubmed/17568012).

Galaxy Penn State's Galaxy is a useful way of wrapping many command line modules together in a userfriendly GUI. 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.

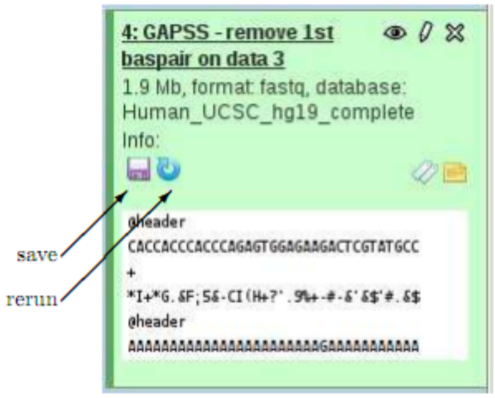
Availability and examples Galaxy is freely available for download: http://galaxyproject.org/. Likewise are the tools that you will use in this practical (use google.com to find them): BWA for alignment, FastQC and Picard for quality control, SAMtools and Varscan for calling SNPs. For this practical, we will use the Galaxy installation at NBIC: http://galaxy.nbic.nl/.

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



4: GAPSS - remove 1st baspair on data 3



4: GAPSS - remove 1st baspair on data 3
1.9 Mb, format: fastq, database: Human_UCSC_hg19_complete
Info:

```
@header
CACCACCCACCCAGAGTGGAGAAGACTCGTATGCC
+
*I+*G.&F;5&-CI(H+?'.9%+-#-&'&$'#.&$
@header
AAAAAAAAAAAAAAAAAAAAAAAAGAAAAAAAAAAA
```

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.
   - Use your names, with @gcc.rug.nl as Email address (**that is so we can see your progress afterwards**).
   - Choose a password (same as your account name).
   - Make the public name the same as your account name.
3. Go to http://galaxy.nbic.nl/workflow/list_published
4. Click on "GCC_ngs_course".
5. Click on "Import workflow".
6. Repeat step 3–5 for "GCC_ngs_course_including_mark_duplicates".

Exercise 1: create a SNP file from a FASTQ file.

The input data is a small selection of reads that should align to a part of the TTN gene (located on chromosome 2). After alignment, you can call SNPs.

Import the data.

- Click on "Shared Data" -> "Data Libraries".
- Click on "GCC course".
- Select "Reads_indv1.fq.gz" and click "Go".

Click on "Workflow" to start the analysis.

- Click "GCC_ ngs_course" followed by "run"
- Run the workflow (This may take some time to start)

Questions:

After completion (blocks coloured green on the right) look at the produced output:

- *NGS: QC and manipulation: Fastqc: Fastqc_QC:* this tool checks the basic quality of the NGS dataset before alignment. Download (click the blue save button next to the data) the produced output and unzip it, afterwards open the picture named "per_base_quality.png".
  (Question: Given that a base quality score of 16 is sufficient, are most of the bases in the reads of high enough quality to be aligned?)
- *NGS: QC and manipulation: FASTQ Groomer:* this process checked the quality scores offset of the NGS data
  (Question: Did you retain all sequences?).
- *NGS: Mapping: Map with BWA for Illumina:* here the ngs data is aligned to human genome build 19 with default settings
  (Questions: How many reported alignments to chromosome 2 were there? Where there alignments to other chromosomes?)
- *NGS: SAM Tools: SAM-to-BAM:* here the produced SAM file is converted to a binary format to reduce the filesize.
- *NGS: SAM Tools: Generate Pileup:* input is the BAM file you just created to produce a pileup file on which algorithms can perform SNP calling.
- *NGS: NGS Taskforce: LUMC – GAPSS v2: VarScan – pileup2snp:* this software called SNPs on the produced pileup file.
  (Question: How many SNPs have you predicted?).
- *NGS: NGS Taskforce: LUMC – GAPSS v2: GAPSS – Ensembl SNP:* here your SNPs are annotated with information from the ENSEMBL database.

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

- Click the "save" button to save the Ensembl SNP output (will save by default to your desktop).
- You can now open the file with Excel.

To start the next exercise with a new clean history do the following:

- In the history panel on the right click on "Unnamed history" and give the history a name. Your history is now saved and available for future use.
- Click the grey "wheel" in the right of the history panel and select "Create New"

Exercise 2: Create a SNP file using additional filtering during analysis
The input data is a the same data as used in exercise 1, this time the workflow we use consists of an additional procedure.

Import the data.

- Click on "Shared Data" -> "Data Libraries".
- Click on "GCC course".
- Select "Reads_indv1.fq.gz" and click "Go".

Click on "Analyze Data" to start the analysis using the previously selected workflow.

- Click "Workflow"
- Click "GCC_ngs_course_including_mark_duplicates" followed by "run"
- Run the workflow (This may take some time to start)

As you might have noticed an additional analysis step named MarkDups is performed. Here we will look into the output of two specific steps in detail.

Questions:
After completion (blocks coloured green on the right) look at the produced output:

- *NGS: Picard: Mark duplicate reads:* This step marks duplicate reads. Download the output and open it.
  (Question: how many duplicate reads were detected?)
- *NGS: NGS Taskforce: LUMC – GAPSS v2: GAPSS – Ensembl SNP:* Download the ENSEMBL annotated SNPs to your desktop. Compare the output to the SNP file produced in exercise 1.
  (Questions: How many SNPs were detected?
  Given that there's a lab procedure before generating the NGS data, what kind of reads did you correct for using the Mark Duplicates step?
  How does not removing duplicate reads affect your downstream analysis such as SNP calling?)

Exercise 3: Identifying mutations in a family
In this exercise we will compare the SNP lists of a trio, father, mother and child. The analyzed in exercise 2 belongs to the father. To analyze the data of the mother and child respectively select "Reads_indv2.fq.gz" and "Reads_indv3.fq.gz" from the data library and analyze them as described in the procedure from exercise 2.

Questions:
(Compare the SNPs generated for all three individuals, which SNP is most likely to be the causal variant?)