Fourth MGC Next-Generation Sequencing course.
Hands on workshop: Next generation sequence data analysis.
Instructors: Michiel van Galen, Jeroen Laros.
Leiden Genome Technology Center, Leiden University Medical Center, The Netherlands

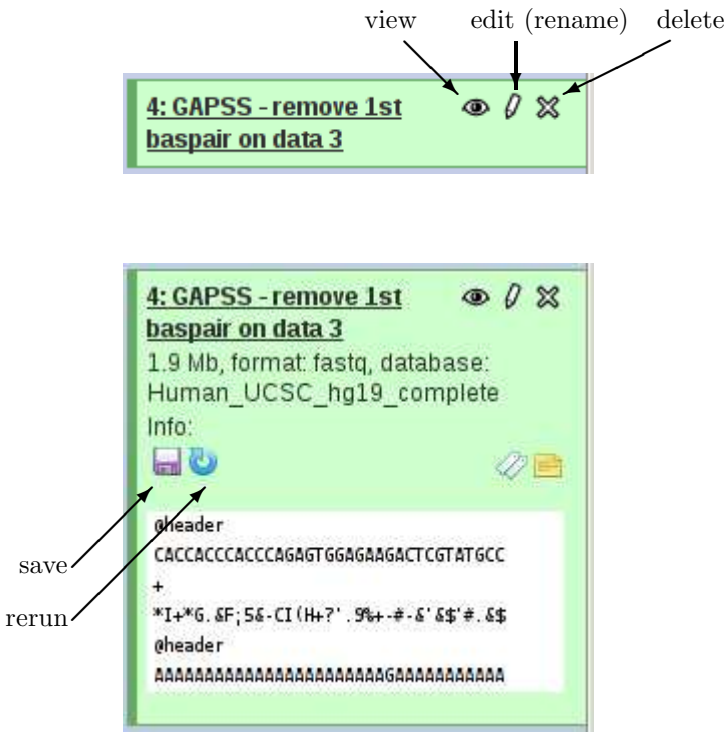
Introduction In this workshop we will first show you a typical analysis done by a bioinformatician. This involves using Linux command line executables to align to a known reference genome and call SNPs, reporting as a tab-delimited file. We will then show how to do this same analysis with a more biologist 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. 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 The tools used in these exercises are all free for download, including Galaxy itself (`http://galaxy.psu.edu`), Bowtie for alignment, SAMtools and Varscan for calling SNPs. The Leiden GAPSS scripts are at `http://www.lgtc.nl/GAPSS`. Examples of analysis with such scripts are in Hestand et al. 2010 (PMID 20615900) for CAGE and SAGE analysis and Ramos et al. 2010 (PMID 20435671) for ChIP-seq analysis.

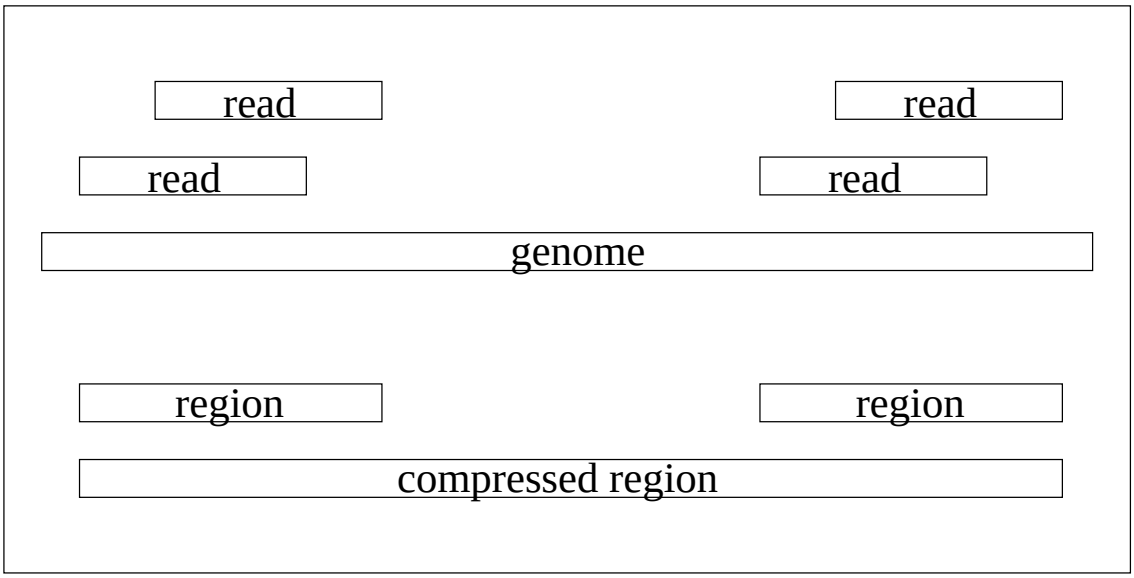
Note on test data Data used in this practical is test data and not full size files. We also only align to a few chromosomes and not full genomes. This is to reduce the time needed to run each step and make this analysis possible within the time permitted.

Some Galaxy icons explained

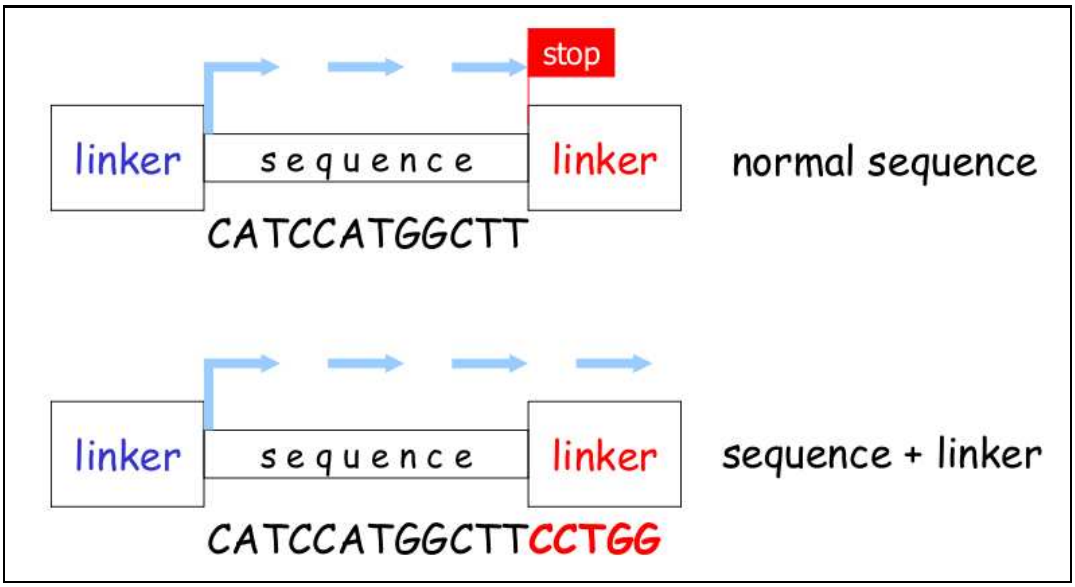


Additional Explanation on GAPSS

Region files are created by combining overlapping aligned reads into one region. A compressed region uses a sliding window to concatenate any regions within this window size into one larger region:



Linkers can be found in sequences if sequencing more cycles than the sequence length. When aligning to a reference genome this can then cause the sequence not to align since it may contain too many mismatches.



Practical 1: create a SNP file from a FASTQ file.
The input data is a small selection of reads that should align to the human BRCA1 gene (located on chromosome 17). After alignment, you can call SNPs. You will first do this with the more traditional command line method, and then repeat this using a much simpler system: Galaxy.

In command line:

1. Open a terminal (click on the TV looking icon on the top taskbar).
2. Go to the SNP directory by typing (*Linux is case sensitive): `cd Desktop/SNP_practical/`
3. View what is in the directory by typing: `ls`
4. Align the file `BRCA1reads.fq` with bowtie to human chr17 by typing:
   - `bowtie --sam /data/bowtie-genome-indexes/UCSC_hg19_Chr17 BRCA1reads.fq > BRCA1reads.sam`
5. Look at the output by typing: `head BRCA1reads.sam`
6. Use SAMtools and Varscan to call SNPs by typing:
   - `samtools faidx chr17.fa`
   - `samtools view -bt chr17.fa -o BRCA1reads.bam BRCA1reads.sam`
   - `samtools sort BRCA1reads.bam BRCA1reads.sorted`
   - `samtools pileup -f chr17.fa BRCA1reads.sorted.bam > BRCA1reads.pileup`
   - `java -jar /usr/local/VarScan/VarScan.v2.1.jar pileup2snp BRCA1reads.pileup > BRCA1reads.SNP`
7. View the data by right clicking on the file `BRCA1reads.SNPs` and open with Openoffice Spreadsheet (right click, "Open with").

In Galaxy: Open Firefox and click on the Galaxy shortcut

Upload all the data we will use:

- Get Data: upload file: `BRCA1reads.fq` (in `Desktop/SNP_practical`).
- Get Data: upload file: `chr17.fa` (in `Desktop/SNP_practical`, this will be needed for SAMtools).

Check the FASTQ file format and align to chr17:

- NGS: QC and manipulation: FASTQ Groomer: run on the `BRCA1reads.fq` Sanger quality (Question: Did you retain all sequences?).
- NGS: Mapping: Map with Bowtie for Illumina: use as input your FASTQ groomed data, align to `UCSC_hg19_Chr17` – otherwise leave defaults (Question: How many reported alignments were there? (note, this was selected data for testing)).

Use SAMtools and Varscan to call SNPs:

- NGS: SAM Tools: SAM-to-BAM: input is your Bowtie output, Reference is "History" (`chr17.fa`).
- NGS: SAM Tools: Generate Pileup: input is the BAM file you just created, For the index (first option) select history and your `chr17.fa` file.
- NGS: Snip Detection: Varscan – pileup2snp: run on your pileup file (Question: How many SNPs have you predicted?).

Lets take this a step further than before and also annotate your SNPs with Ensembl:

- NGS: Snip Detection: GAPSS – Ensembl_SNP: as input use your Varscan SNP file (note: this may take a minute to run).

Lets save this 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).
- open the file with OpenOffice Spreadsheet (right click, "Open with").

Practical 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 (`CAGE_example.fastq`) from a human CAGE project. You will convert it to a sanger quality FASTQ file, trim the first basepair (lower quality), trim for linkers (when we sequence more cycles than the sequence fragment length, we end up with the reverse linker in our sequence), align to the full human genome, and view this data in a tab-delimited format and in the UCSC genome browser.

In Galaxy: (first clean history, under option select "Delete").
Upload all the data we will use:

- Get Data: upload file: `CAGE_example.fastq` (in `Desktop/CAGE_practical`).
- Look at how this data looks.

Convert to sanger quality (this is illumina 1.0) FASTQ:

- NGS: QC and manipulation: FASTQ Groomer: run on the new FASTQ file, quality type solexa 1.0.

Clean up the data

- NGS: Tools LUMC: GAPSS Remove 1st bp.
- NGS: Tools LUMC: GAPSS Edit for linkers (edit on the 3' end the sequence TCGTATGCCGTCTTCTGCTTG).
  - 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 human build 19.

- NGS: Mapping: Map with Bowtie for Illumina: use as input your edited FASTQ data, align to `Human_UCSC_hg19`, deselect the output in SAM format, otherwise leave defaults.

Convert to an in-house alignment format called IGF:

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

Make a tab delimited report file:

- NGS: Tools LUMC: GAPSS Make regions, input is the IGF file.
- To eliminate gaps of 100bp lets run NGS: Tools LUMC: GAPSS Compress regions, gap size 100.
- Save the compressed regions file to your desktop.
- Open with Openoffice Spreadsheet.
- 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: Tools LUMC: GAPSS IGF to WIG, make sure to use the file `CAGE_IGF`.
- Save this file to your desktop.
- Run the command `./convert.sh Galaxy9-*` (needed because of a small bug in one of the tools).
- Go to the UCSC genome browser.
- Click "Genome Browser".
- Click "add custom tracks" and select the file `wiggle.gz` from your desktop.
- Check out the most significant region from your sorted OpenOffice spreadsheet data (question: does this make sense? (i.e. does it align to the 5' end of a gene?)).

Practical 3: Workflows
Workflows can be extracted from a history and saved in order to re-run an analysis.

- First, clear the history again.
- Log in (under User, select Login) Email address: `test@lumc.nl`, Password: `test123`
- Upload either one of the input datasets (`CAGE_practical/CAGE_example.fastq` or `SNP_practical/BRCA1reads.fq`).
  - This time, make sure you select the genome (Human UCSC hg19), this is needed for the rest of the workflow.
- Click on the workflow button and select the appropriate workflow. Click "Run".
- Now click "Run workflow" to execute the workflow.