LEIDEN UNIVERSITY MEDICAL CENTER

Galaxy & GAPSS

(General annotation Pipeline for Second-generation Sequencers)

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

GAPSS

General annotation Pipeline for Second-generation Sequencers

- Matthew S. Hestand[+1,2], Michiel van Galen[+2], Michel Villerius[+1], Gert-Jan B. van Ommen[1], Johan T. den Dunnen[1,2], and Peter A.C. 't Hoen[1]

[+]Equal contribution
[1]The Center for Human and Clinical Genetics,
[2]Leiden Genome Technology Center,
Leiden University Medical Center, Leiden, The Netherlands.

History

Illumina Genome Analyzer.

- Our first data 2007.

Few analysis tools.

- Many custom Perl scripts.

In 2008 start “patching together” scripts.

- Creation of GAPSS (v1).

GAPSS (v1)

Frequently used for:

- CAGE / SAGE.
- ChIP-seq.
- (Exome) resequencing.

Input: FASTA / FASTQ / SCARF.

Output:

- Summary file (tags in, tags aligned, editing details).
- UCSC wiggle files.
- Region files.
- SNP report.

GAPSS (v1)

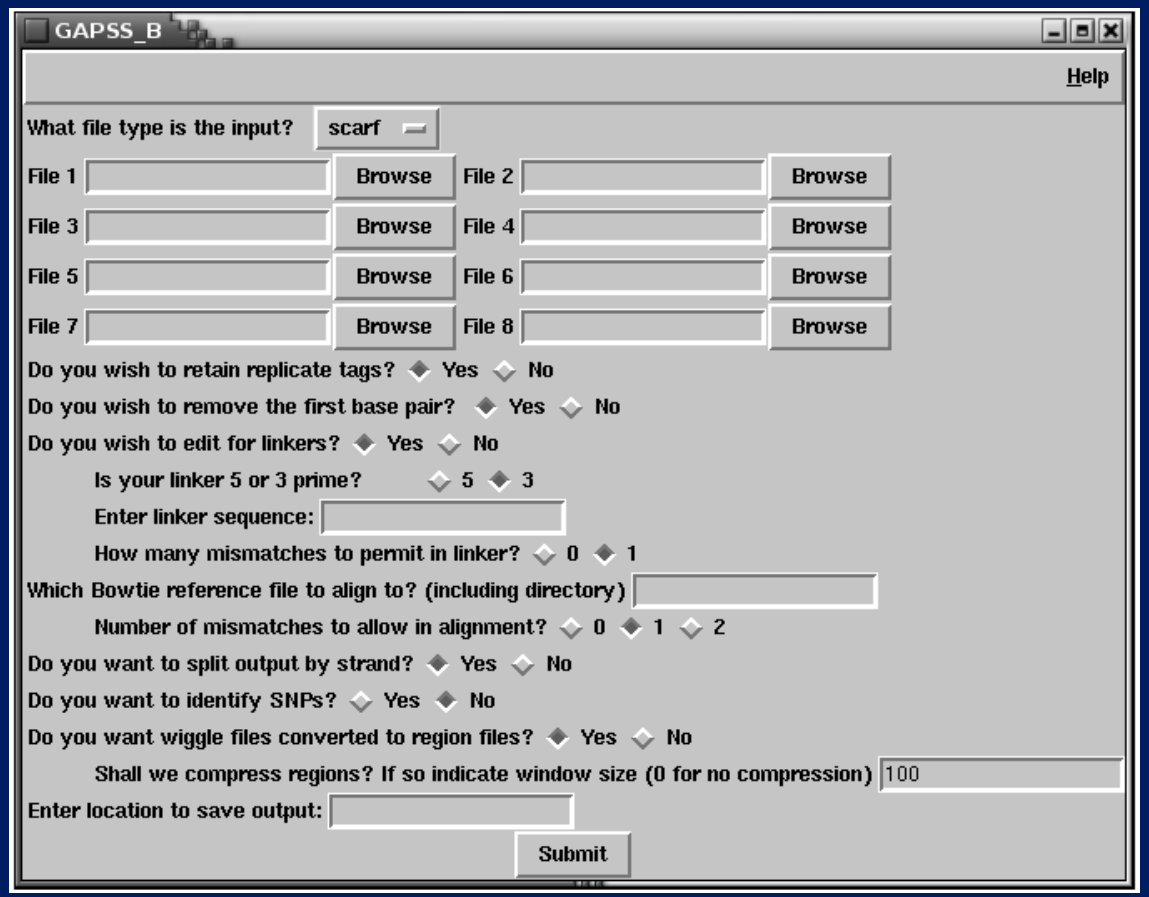
Requirements:

- Alignment tool.
  - Rmap (Smith et al. 2008).
  - Bowtie (Langmead et al. 2009).
- Perl (BioPerl).
- Linux.

Runtime (with Bowtie):

- For two full lanes: ±30 minutes (four processors, 6.5 GB memory).

GAPSS (v1)



There is also a command line version available.

GAPSS (v2)

Improvements:

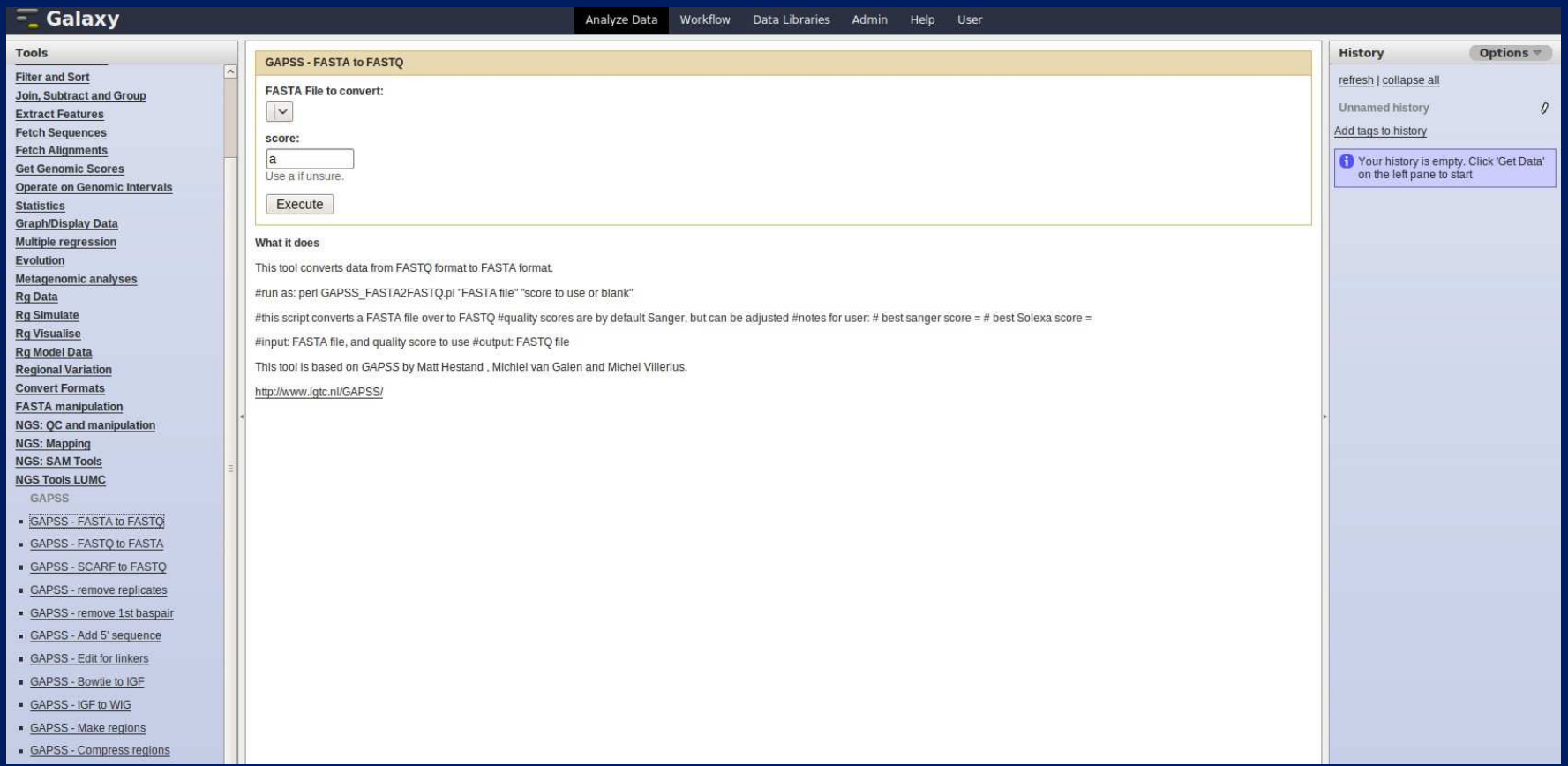
- Exact number of tags in regions.
- Use Varscan (Koboldt et al. 2009) for variant calling.
- Annotate SNPs.

This version uses Galaxy as a graphical interface.

- Gives the user the opportunity to make their own workflows.
- Workflows can be shared between users.

http://main.g2.bx.psu.edu/

New GAPSS interface: Galaxy



LGTC install and GAPSS added by Michel Villerius.

http://galaxy.nbic.nl/

GAPSS / Galaxy

- You can add your own tools to a Galaxy installation.
- A total of twelve tools were added.
- Users can make their own combinations.

GAPSS / Galaxy

- Convert several formats to FASTQ.
  - FASTA to FASTQ.
  - SCARF to FASTQ.

GAPSS / Galaxy

- Remove exact duplicates from the input.

GAPSS / Galaxy

- Remove the first basepair from all reads.
  - The first basepair frequently has a poor quality.

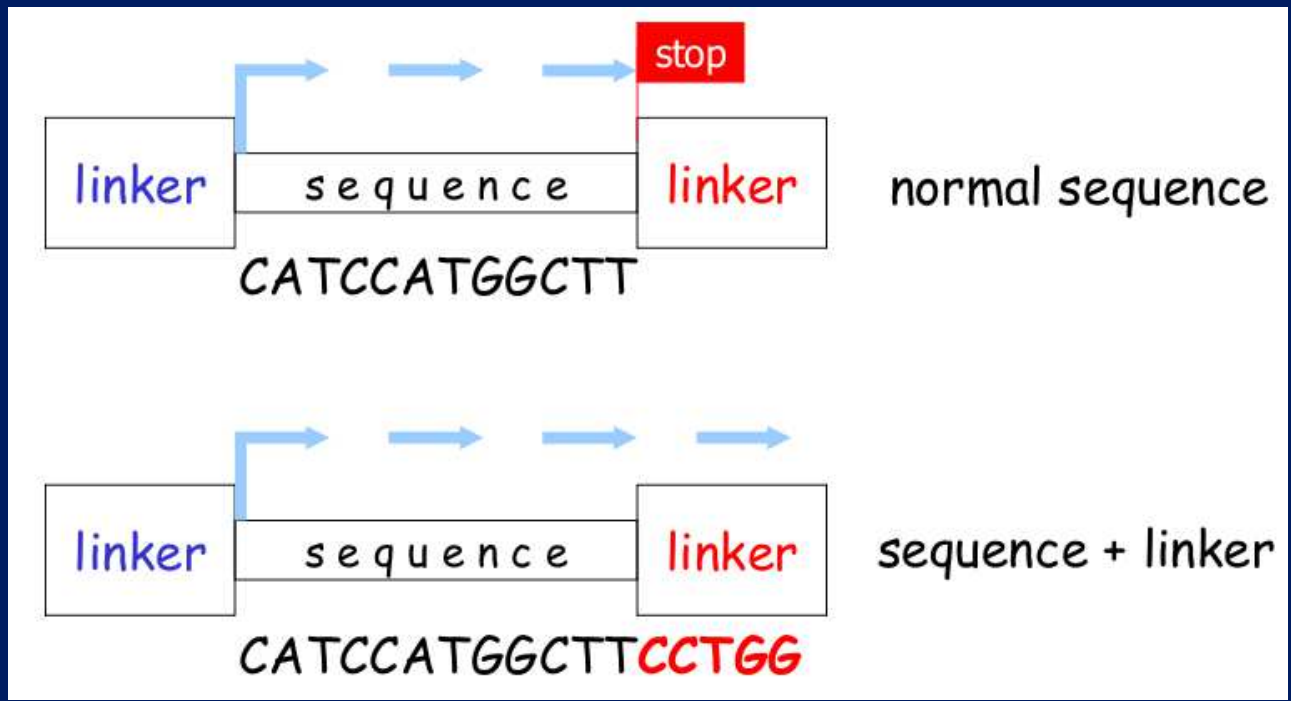
GAPSS / Galaxy

- Add a sequence to the start of each read.
  - A restriction site used to sheer the DNA.
  - DeepSAGE.

GAPSS / Galaxy

- Remove (parts of) a linker sequence from each read.
  - The linker sequence can be on either side of the read.

GAPSS: Edit for linkers



- A linker sequence can be found if there were more sequencing cycles than the length of the sequence.
- Linkers must be removed to align properly.

GAPSS / Galaxy

- Intermediate GAPSS format.
  - Generic alignment format.

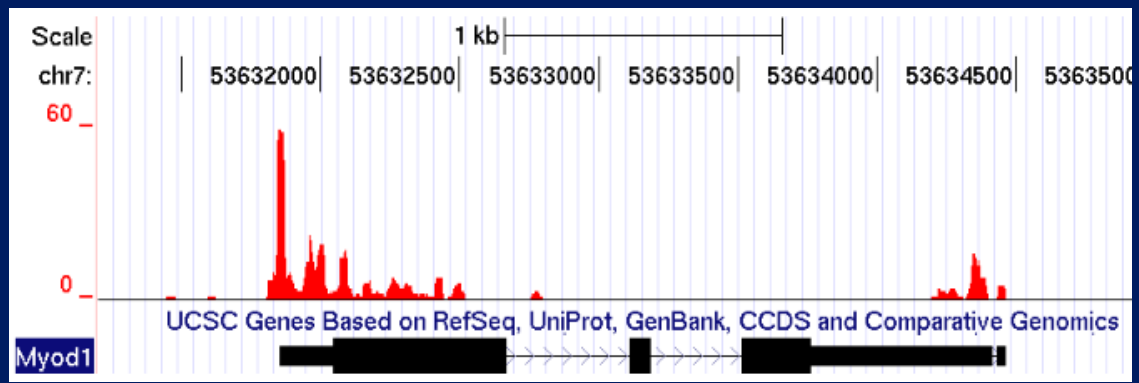
GAPSS / Galaxy

NGS Tools LUMC

GAPSS

- GAPSS - FASTA to FASTQ
- GAPSS - FASTQ to FASTA
- GAPSS - SCARF to FASTQ
- GAPSS - remove replicates
- GAPSS - remove 1st baspair
- GAPSS - Add 5' sequence
- GAPSS - Edit for linkers
- GAPSS - Bowtie to IGF
- GAPSS - IGF to WIG
- GAPSS - Make regions
- GAPSS - Compress regions

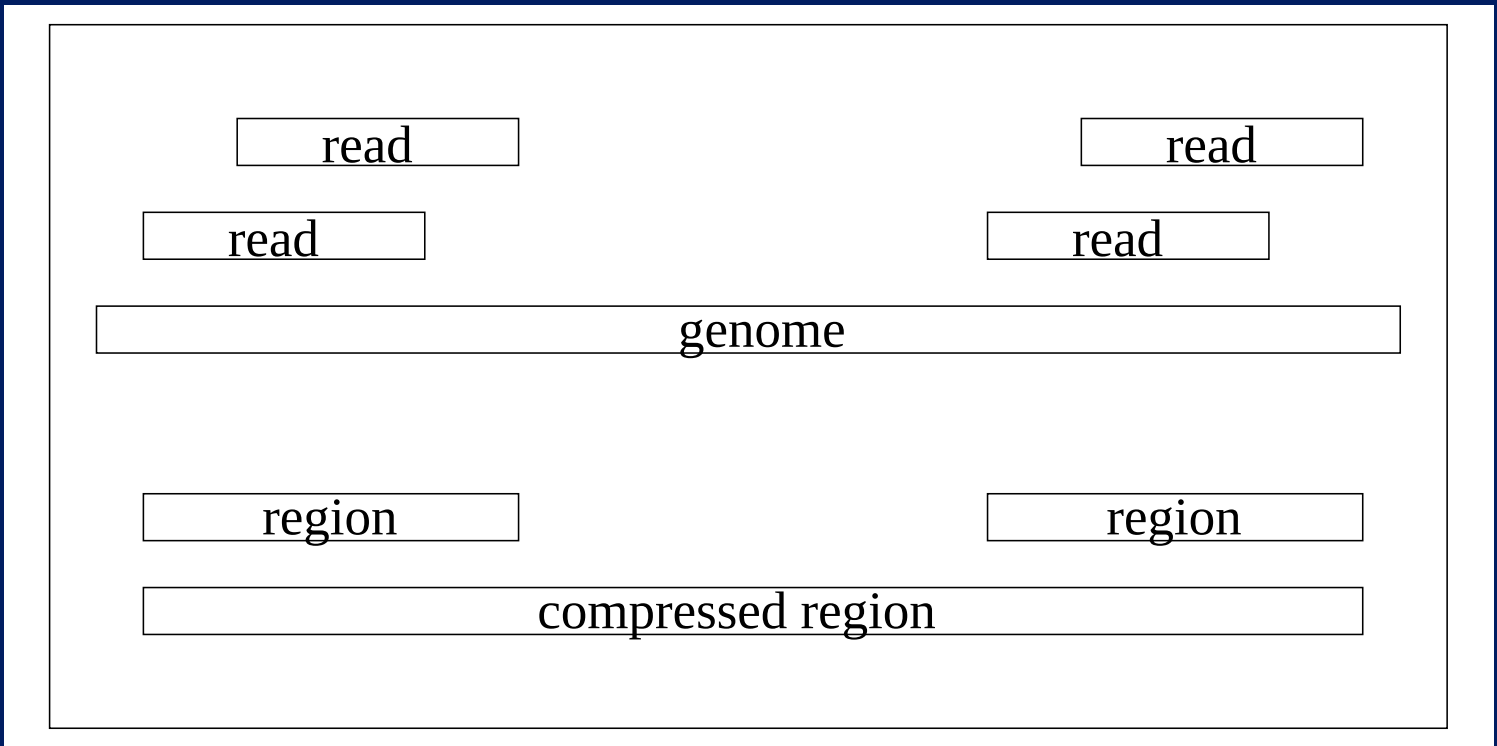
- Create a wiggle file.



GAPSS / Galaxy

- Make region files.

Region files



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

Ensembl variant annotation

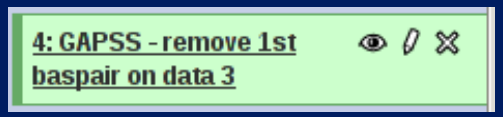
- Is the variant known?
- Does it hit a gene?
  - Is it in an intron?
    - Does it hit a splice site?
  - Is it in the coding region?
    - Is there a gain/loss of a stop codon?
    - Does the variant result in a frameshift?
    - …
  - Is it in the 5'/3' UTR of a gene?
  - …
- Is it in a regulatory region?
- …

Galaxy

- 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 http://galaxy.nbic.nl/

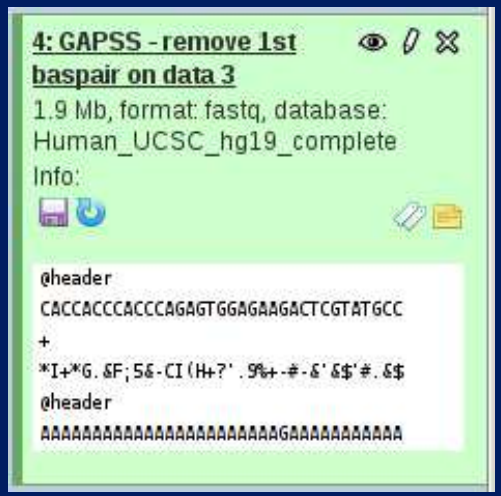
Galaxy icons



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

- Click on the title for a more detailed view.

Galaxy icons



- 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.

Questions?

Acknowledgements:

Matthew Hestand
Michiel van Galen
Michel Villerius
Gert-Jan van Ommen
Johan den Dunnen
Peter-Bram ’t Hoen

Ivo Fokkema
Yavuz Ariyurek

http://www.lgtc.nl/GAPSS

https://humgenprojects.lumc.nl/trac/GAPSS3/wiki/course