



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

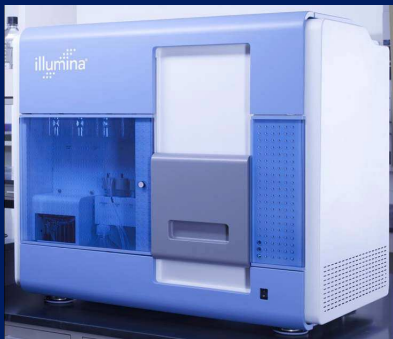
- Matthew S. Hestand^{+1,2}, Michiel van Galen⁺², Michel Villerius⁺¹, Gert-Jan B. van Ommen¹, Johan T. den Dunnen^{1,2}, and Peter A.C. 't Hoen¹

⁺Equal contribution

¹The Center for Human and Clinical Genetics,

²Leiden 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)

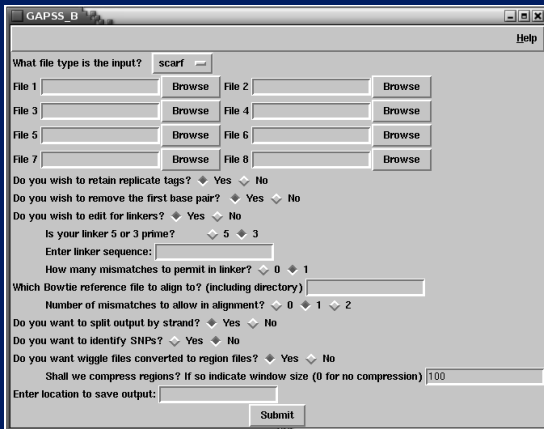
Requirements:

- 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)



The screenshot shows the GAPSS_B application window with the following settings:

- What file type is the input? **scarf**
- File 1, File 2, File 3, File 4, File 5, File 6, File 7, File 8: Each has an empty text box and a **Browse** button.
- Do you wish to retain replicate tags? Yes No
- Do you wish to remove the first base pair? Yes No
- Do you wish to edit for linkers? Yes No
- Is your linker 5 or 3 prime? 5 3
- Enter linker sequence:
- How many mismatches to permit in linker? 0 1
- Which Bowtie reference file to align to? (including directory)
- Number of mismatches to allow in alignment? 0 1 2
- Do you want to split output by strand? Yes No
- Do you want to identify SNPs? Yes No
- Do you want wiggle files converted to region files? Yes No
- Shall we compress regions? If so indicate window size (0 for no compression)
- Enter location to save output:
- Submit** button

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

Galaxy Analyze Data Workflow Data Libraries Admin Help User

Tools

- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Multiple regression
- Evolution
- Metagenomic analyses
- Rig Data
- Rig Simulate
- Rig Visualise
- Rig Model Data
- Regional Variation
- Convert Formats
- FASTA manipulation
- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS Tools LUMC
- GAPSS
 - [GAPSS - FASTA to FASTQ](#)
 - [GAPSS - FASTQ to FASTA](#)
 - [GAPSS - SCARF to FASTQ](#)
 - [GAPSS - remove replicates](#)
 - [GAPSS - remove 1st base](#)
 - [GAPSS - Add 3' sequence](#)
 - [GAPSS - Edit for linkers](#)
 - [GAPSS - Bowtie to IGF](#)
 - [GAPSS - IGF to WIG](#)
 - [GAPSS - Make regions](#)
 - [GAPSS - Compress regions](#)

GAPSS - FASTA to FASTQ

FASTA File to convert:

score:

EXECUTE

Use a if unset.

What it does

This tool converts data from FASTQ format to FASTA format.

run as: perl GAPSS_FASTA2FASTQ.pl "FASTA file" "score to use or blank"

This script converts a FASTA file over to FASTQ (quality scores are by default Sanger, but can be adjusted /notes for user: # best sanger score = # best Solexa score =

Input: FASTA file, and quality score to use #Output: FASTQ file

This tool is based on GAPSS by Matt Hestand, Michiel van Gaen and Michel Villerius.

<http://www.kyb.niigapss>

History Options

refresh | collapse all

Unnamed history 0

Add tags to history

Your history is empty. Click 'Get Data' on the left-pane to start

LGTC install and GAPSS added by Michel Villerius.

<http://galaxy.nbic.nl/>

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

- 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

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

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

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

- Remove exact duplicates from the input.

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

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

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

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

GAPSS / Galaxy

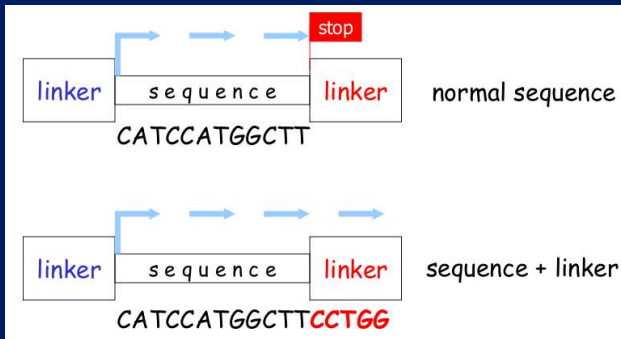
NGS Tools LUMC

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](#)

- 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

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

- Intermediate GAPSS format.
 - Generic alignment format.

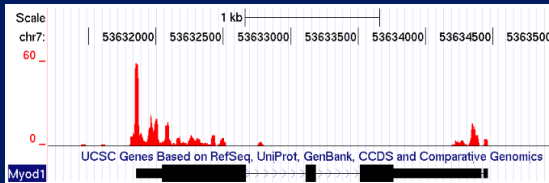
GAPSS / Galaxy

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

NGS Tools LUMC

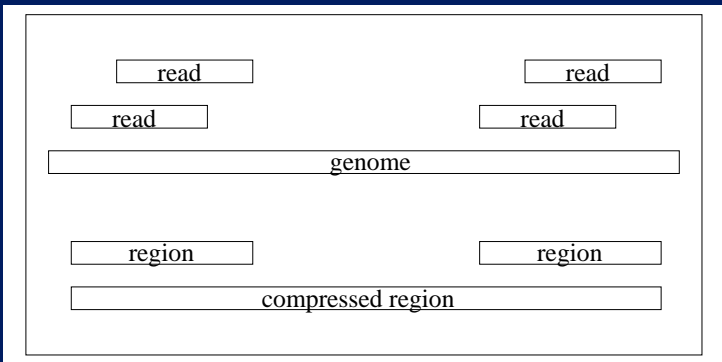
GAPSS

- 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

- Make region files.



Region files



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

Ensembl variant annotation

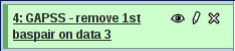
- Is the variant known?
- Does it hit a gene?
 - Is it in an intron?
 - Does it hit a splice site?
 - Is it in the coding region?
 - Is there a gain/loss of a stop codon?
 - Does the variant result in a frameshift?
 - ...
 - Is it in the 5'/3' UTR of a gene?
 - ...
- Is it in a regulatory region?
- ...




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/>

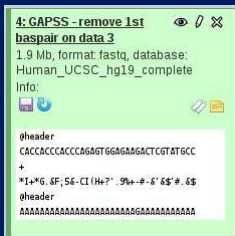
Galaxy icons



4: GAPSS - remove 1st   
baspair on data 3

- 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.



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>