LEIDEN UNIVERSITY MEDICAL CENTER

Exome sequencing with GAPSS3

(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

Introduction

In *exome sequencing*, we select genomic regions of interest using a *target-enrichment strategy*.

- PCR.
- On array capture.
- In-solution capture.

Introduction

In *exome sequencing*, we select genomic regions of interest using a *target-enrichment strategy*.

- PCR.
- On array capture.
- In-solution capture.

Overview of an in-solution capture.

- Fragmentation.
- Size selection.
- Linker ligation.
- Capture.

Introduction

In *exome sequencing*, we select genomic regions of interest using a *target-enrichment strategy*.

- PCR.
- On array capture.
- In-solution capture.

Overview of an in-solution capture.

- Fragmentation.
- Size selection.
- Linker ligation.
- Capture.

These regions are then *sequenced*.

- Illumina Genome Analyzer II (GAII).
- Illumina HiSeq 2000.

Introduction

Figure 1: GAII.

Figure 2: HiSeq 2000.

Paired end, high throughput, cheap.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.
2. Alignment.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.
2. Alignment.
3. Variant calling.
    - Description of difference with the reference.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.
2. Alignment.
3. Variant calling.
    - Description of difference with the reference.
4. Filtering.
    - Variant calling has a high false positive rate.

General layout

Exome sequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Data cleaning.
2. Alignment.
3. Variant calling.
    - Description of difference with the reference.
4. Filtering.
    - Variant calling has a high false positive rate.
5. Annotation.

Pre-alignment

We use the Trimmomatic / FASTX toolkit for data cleaning.

- Remove linker sequences.
- Clip low quality reads at the end of the read.
- Judge the read that is left over.

Pre-alignment

We use the Trimmomatic / FASTX toolkit for data cleaning.

- Remove linker sequences.
- Clip low quality reads at the end of the read.
- Judge the read that is left over.

The FASTQC toolkit is used for quality control (both before and after the data cleaning step).

- GC content.
- GC distribution.
- Quality scores distribution.
- ...

Pre-alignment

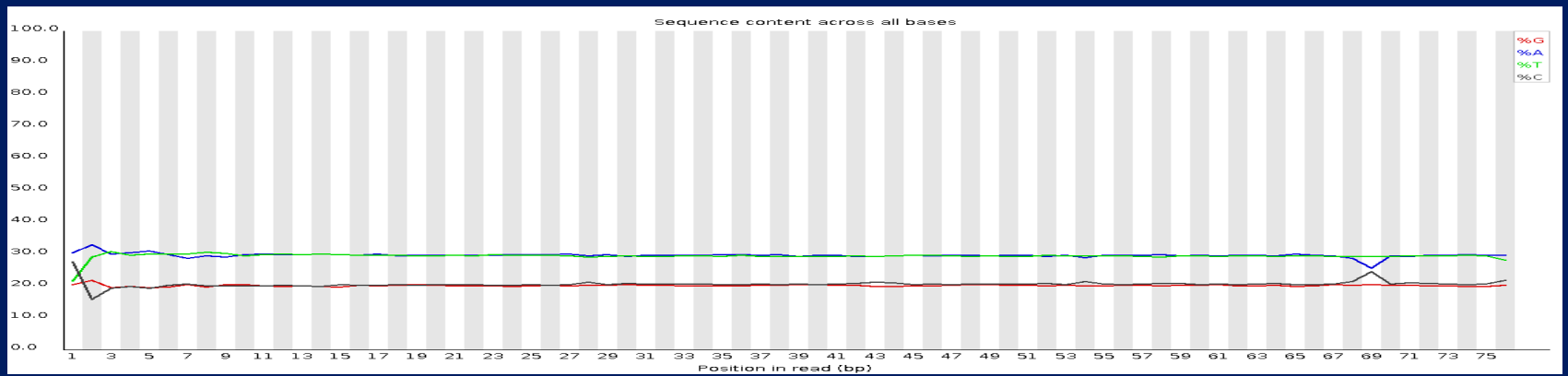


Figure 3: Per base sequence content.

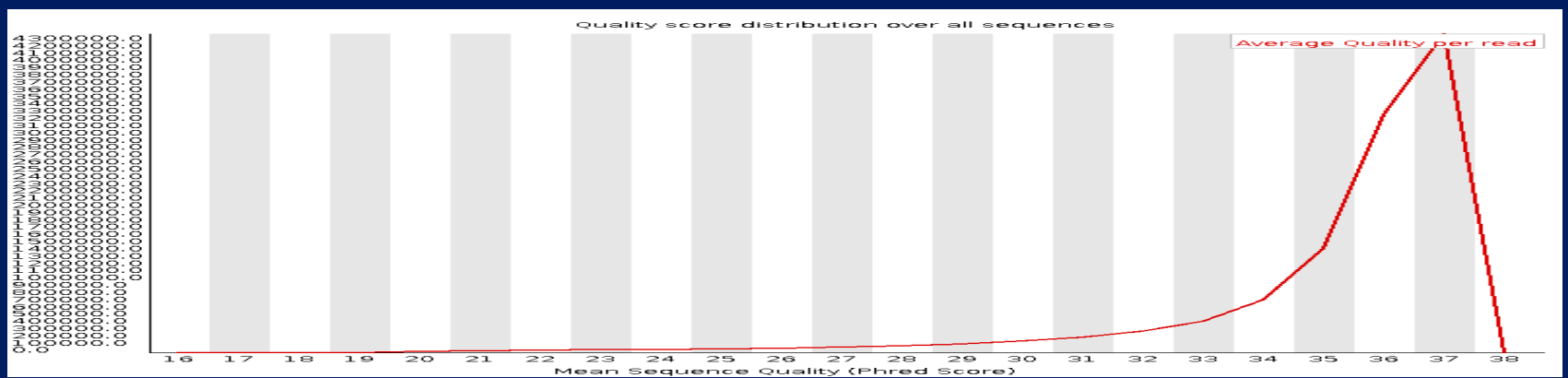


Figure 4: Per sequence quality.

Alignment

Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads.

Alignment

Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads.

Some features:

- Base quality recalibration.
  - First map 1% of the input.
  - Recalibrate the Fastq quality scores.
  - Redo the alignment with the recalibrated scores.

Alignment

Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads.

Some features:

- Base quality recalibration.
  - First map 1% of the input.
  - Recalibrate the Fastq quality scores.
  - Redo the alignment with the recalibrated scores.
- Uses BWA for the hard work.
  - Switches to its accurate built in aligner when BWA fails.

Burrows-Wheeler Aligner (BWA) is a short read aligner that allows small insertions and deletions.

Variant calling

Variant calling is done by Samtools, BCFtools / VCFutils.

The output of most modern aligners is in *Sequence Alignment / Map* (SAM) format.

Variant calling

Variant calling is done by Samtools, BCFtools / VCFutils.

The output of most modern aligners is in *Sequence Alignment / Map* (SAM) format.

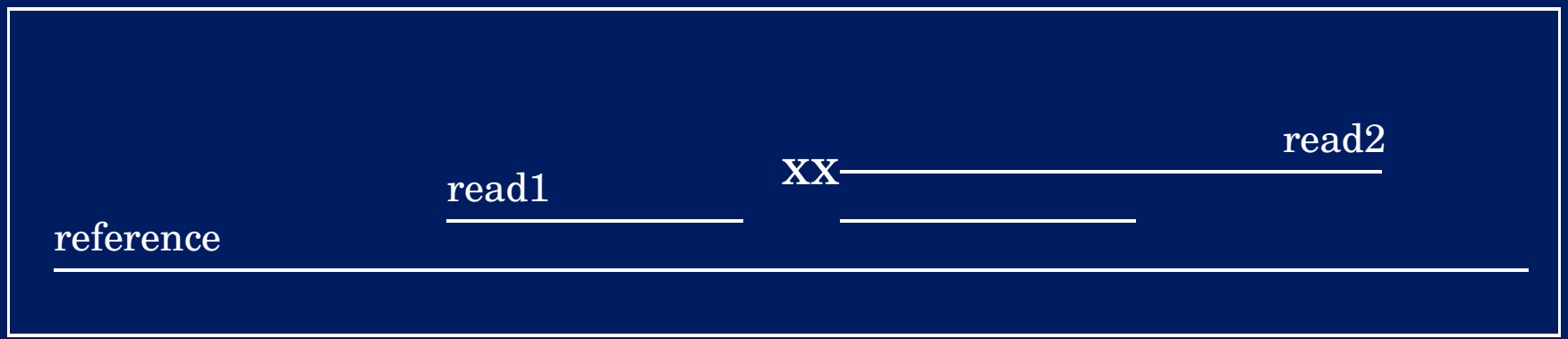
Mainly file format conversions.

- SAM $\rightarrow$ BAM.
- BAM $\rightarrow$ BAM.sorted.
- BAM.sorted $\rightarrow$ BAM.sorted.index.
- BAM.sorted $\rightarrow$ mpileup (BAQ realignment).
- BAM.sorted $\rightarrow$ BCF.
- BCF $\rightarrow$ VCF.

We end up with a list in *Variant Call Format* (VCF).

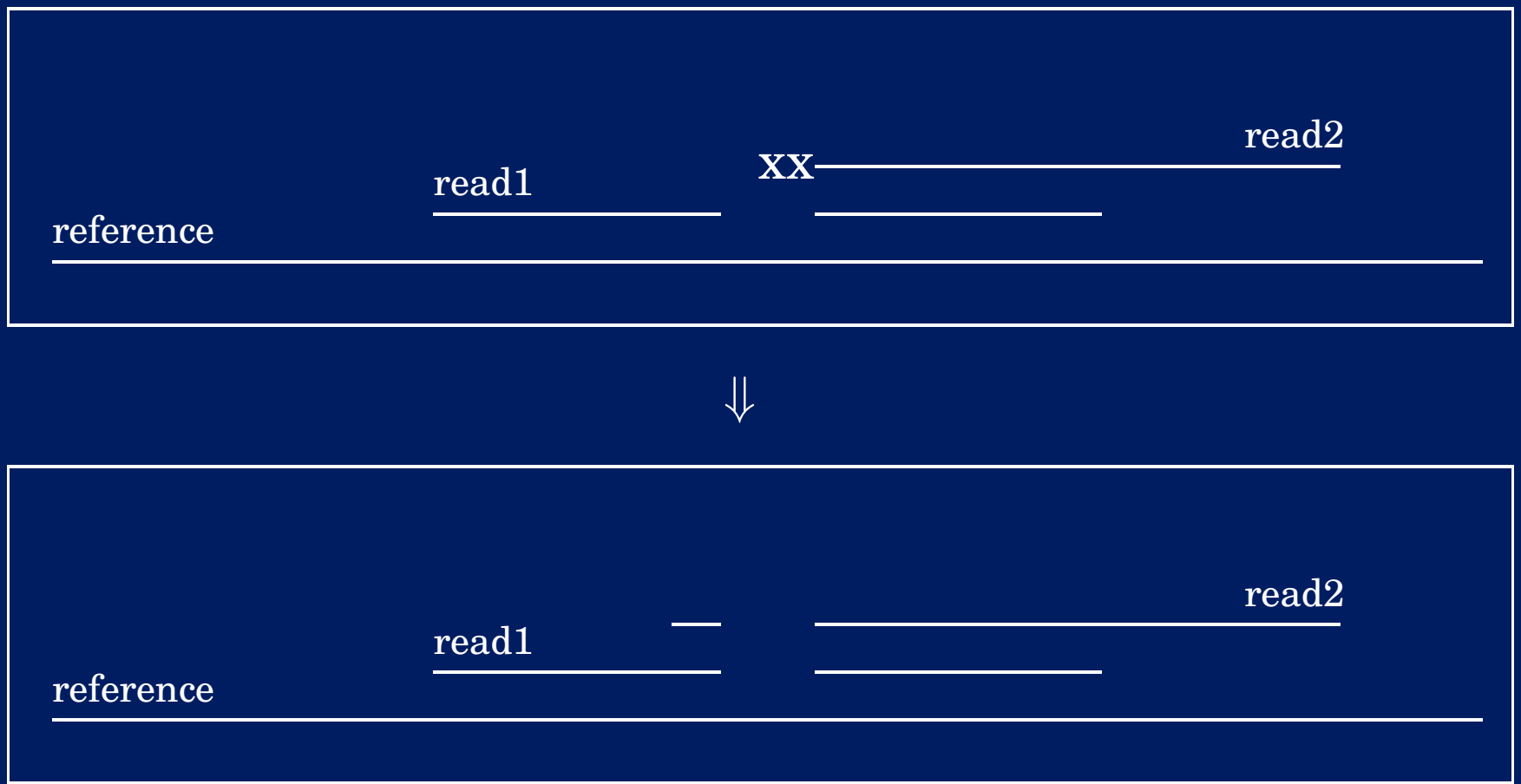
Variant calling

Base Alignment Quality (BAQ) realignment:
Remove SNPs around indels.



Variant calling

Base Alignment Quality (BAQ) realignment:
Remove SNPs around indels.



Annotation

We use different annotation sources and an in-house database.

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq annotation:

- Is the variant known?
- Does it hit a gene?
  - Is it in an intron?
    - Does it hit a splice site?

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq 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?
    - ...

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq 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?
  - ...

Annotation

We use different annotation sources and an in-house database.

A selection of SeattleSeq 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?
- …

Pipelines

Combining all these tools in a pipeline:

```
1  bwa aln -t 8 $reference $i > $i.sai
2  bwa samse $reference $i.sai $i > $i.sam
3  samtools view -bt $reference -o $i.bam $i.sam
```

Listing 1: Shell script

Pipelines

Combining all these tools in a pipeline:

```
1  bwa aln -t 8 $reference $i > $i.sai
2  bwa samse $reference $i.sai $i > $i.sam
3  samtools view -bt $reference -o $i.bam $i.sam
```

Listing 1: Shell script

```
1  %.sai: %.fq
2    $(BWA) aln -t $(THREADS) $(call MKREF, $@) $< > $@
3
4  %.sam: %.sai %.fq
5    $(BWA) samse $(call MKREF, $@) $^ > $@
6
7  %.bam: %.sam
8    $(SAMTOOLS) view -bt $(call MKREF, $@) -o $@ $<
```

Listing 2: Makefile

Pipelines

Name	Manual	Options
bwa	490	±40
samtools	616	±70
fastx	?	±30

Table 1: Selection of used tools

The number of parameters is impressive, but the number of combinations is incredible.

Pipelines

Name	Manual	Options
bwa	490	±40
samtools	616	±70
fastx	?	±30

Table 1: Selection of used tools

The number of parameters is impressive, but the number of combinations is incredible.

Sometimes, people want to tweak these parameters, which is not very practical without a graphical interface.

Graphical interfaces

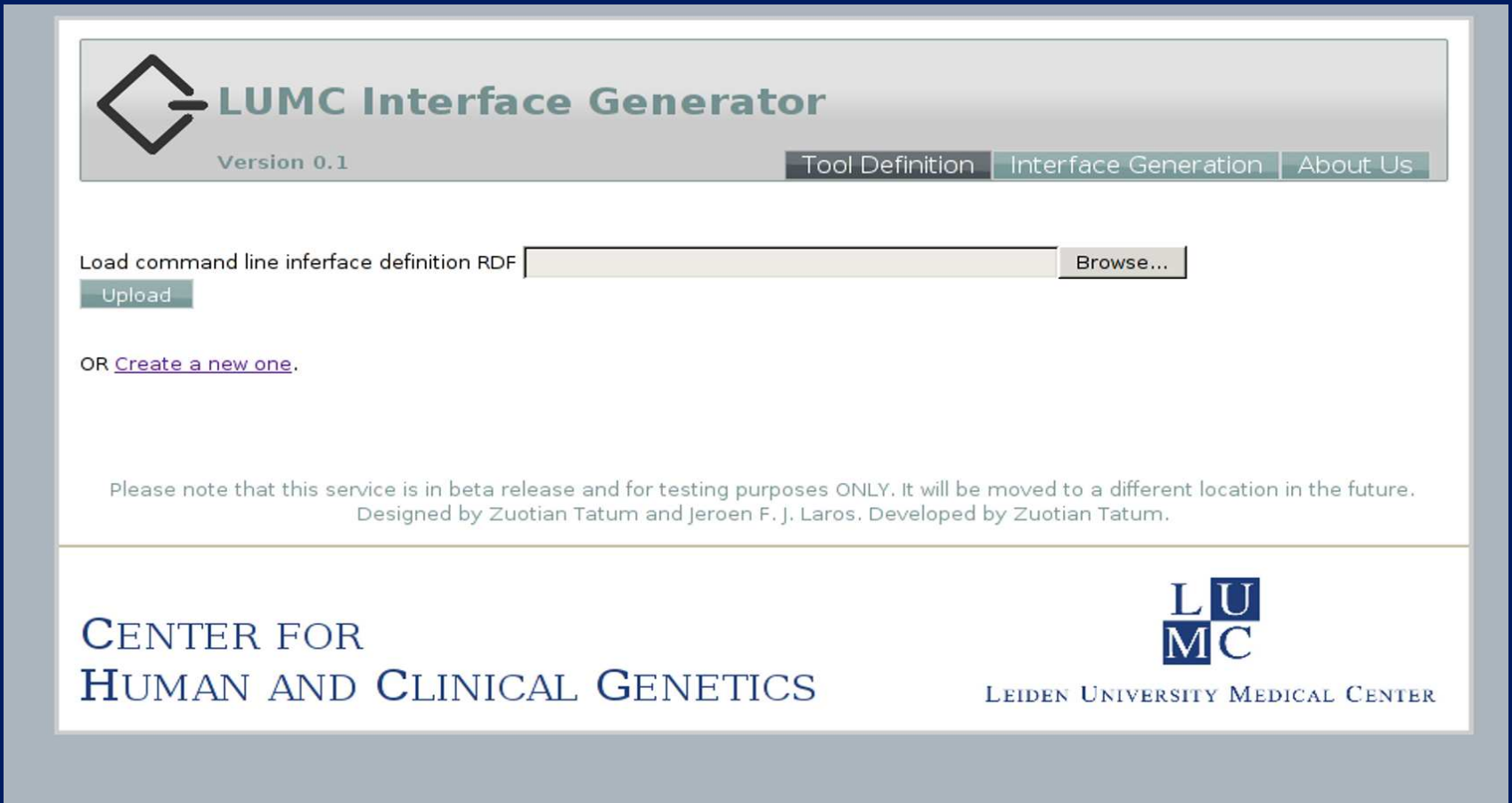


Figure 5: Interface generator

Graphical interfaces

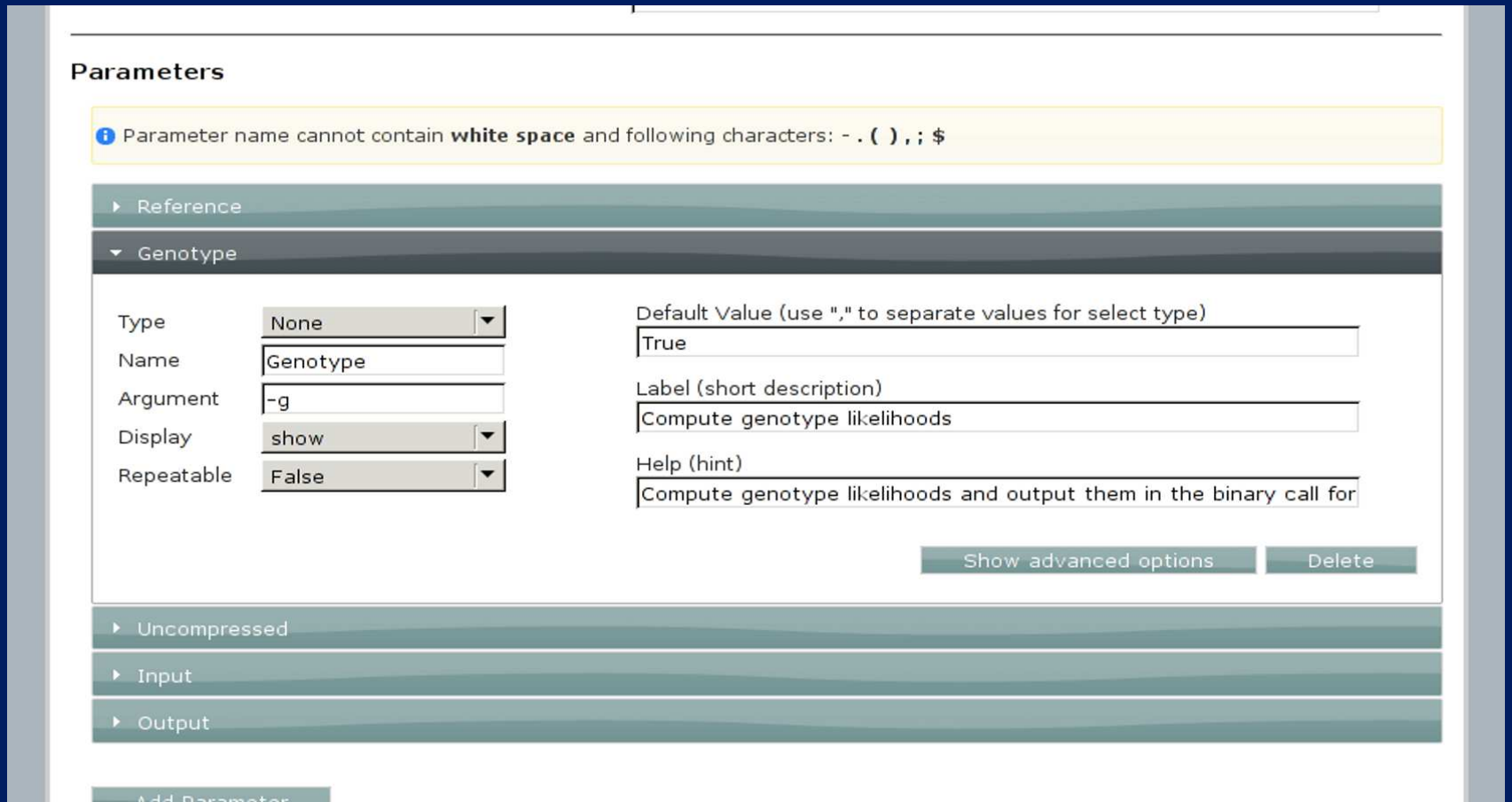


Figure 6: Using the interface generator

Graphical interfaces

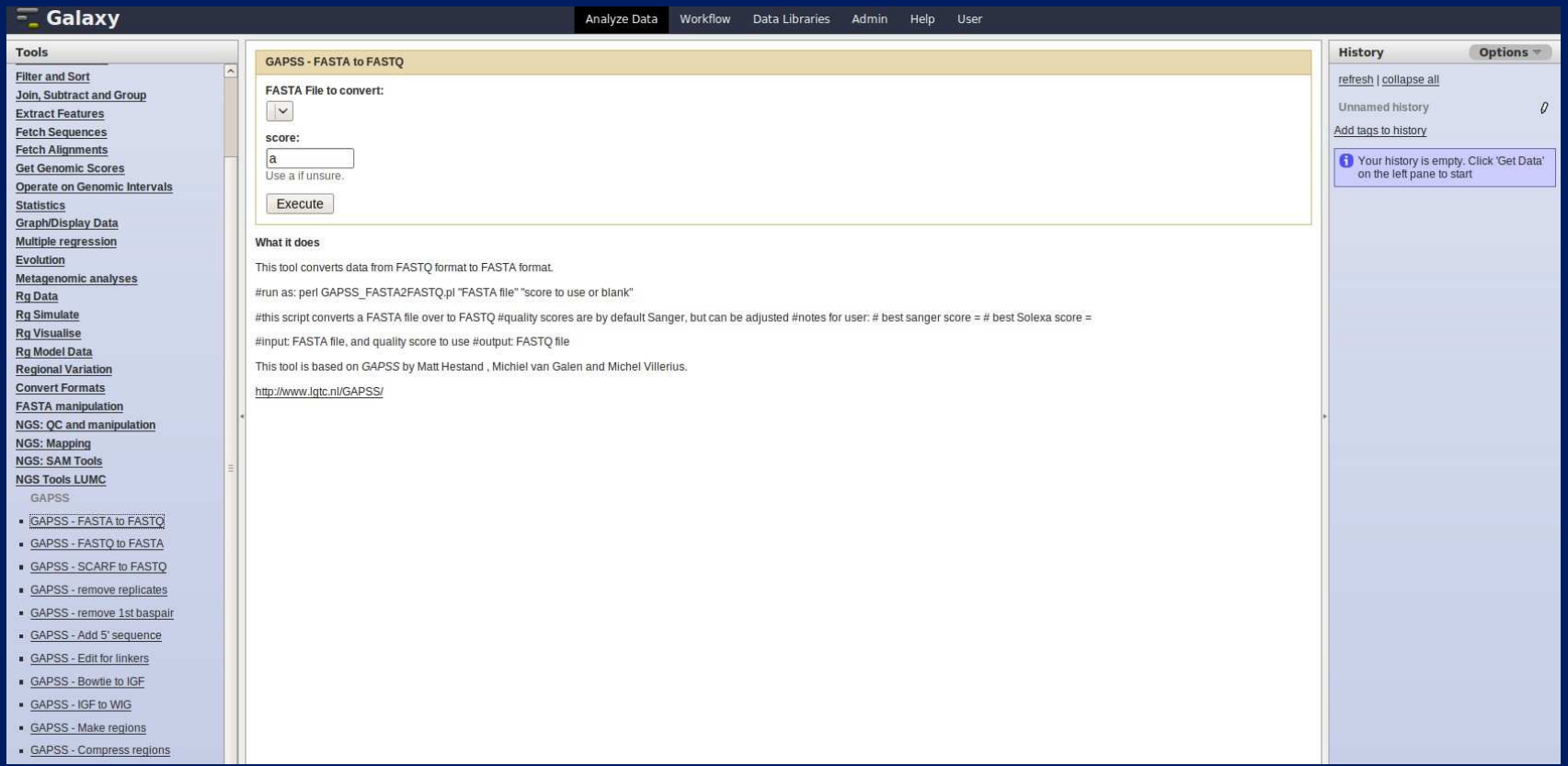


Figure 7: Galaxy

http://galaxy.nbic.nl/

Graphical interfaces

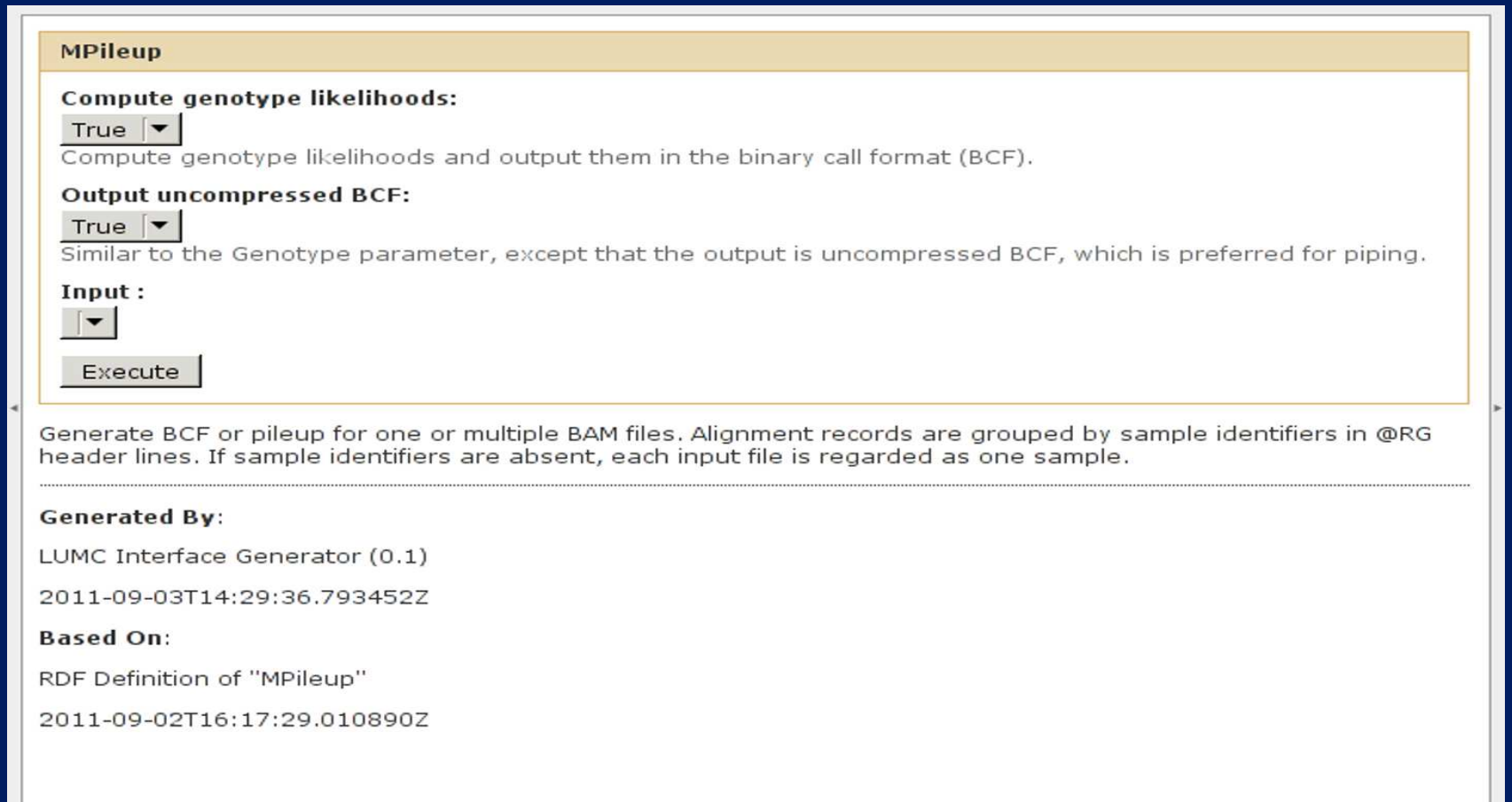


Figure 8: User friendly interface with Galaxy

Questions?

Acknowledgements

Michiel van Galen
Martijn Vermaat
Zuotian Tatum
Yu-Ching Lai
Michel Villerius
Jaap van der Heijden
Bradley ten Broeke
Johan den Dunnen

https://www.mutalyzer.nl/projects/GAPSS3/
http://www.lgtc.nl/