LEIDEN UNIVERSITY MEDICAL CENTER

# Combining tools into a pipeline

**Jeroen F.J. Laros**
**Leiden Genome Technology Center**
**Department of Human Genetics**
**Center for Human and Clinical Genetics**

# Introduction

## *Pipelines*

Figure 1: A real-life pipeline.

# Introduction

## *Pipelines*

Figure 2: Scene from “Modern times”.

# Introduction

## *Pipelines*

Combining tools:

- The output of one tool can serve as the input for another.
- Not necessarily linear.
- ...

# Introduction

***Pipelines***

Combining tools:

- The output of one tool can serve as the input for another.
- Not necessarily linear.
- ...

Running various different tools:

- Two or three different aligners.
- A couple of variant callers.
- ...

# Introduction

***Running example: Exome sequencing***

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

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

# Introduction

***Running example: Exome sequencing***

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

***Running example: Exome sequencing***

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*.

# Introduction

## *Illumina*

Figure 3: HiSeq 2500.

Characteristics:

- High throughput.
- Paired end.
- High accuracy.
- Read length $2 \times 125$bp.
- Relatively long run time (6 days).
- Relatively expensive.

# Introduction

## *Life Technologies*

Figure 4: Ion proton.

Characteristics:

- Moderate throughput.
- Single end (for now).
- High accuracy.
- Read length $\pm$200bp.
- Short run time.
- Cheap runs.

# Introduction

## *Data analysis*

Resequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.
2. Alignment.
    - Post-alignment quality control.
3. Variant calling.
4. Filtering.
    - Post-variant calling quality control.
5. Annotation.

# Pre-alignment

## *Trimming*

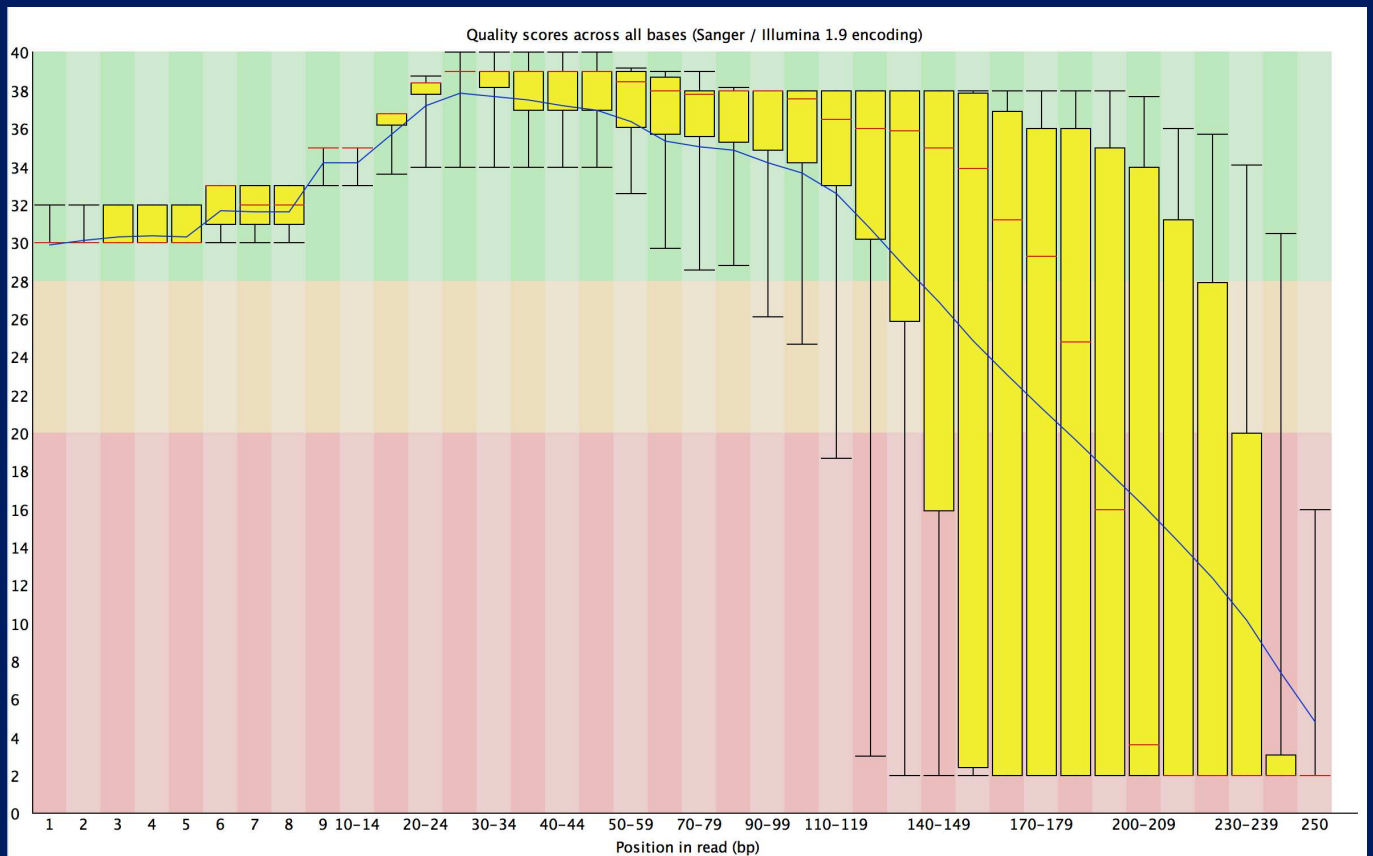


Figure 5: Quality score per position.

# Pre-alignment

## *Clipping*

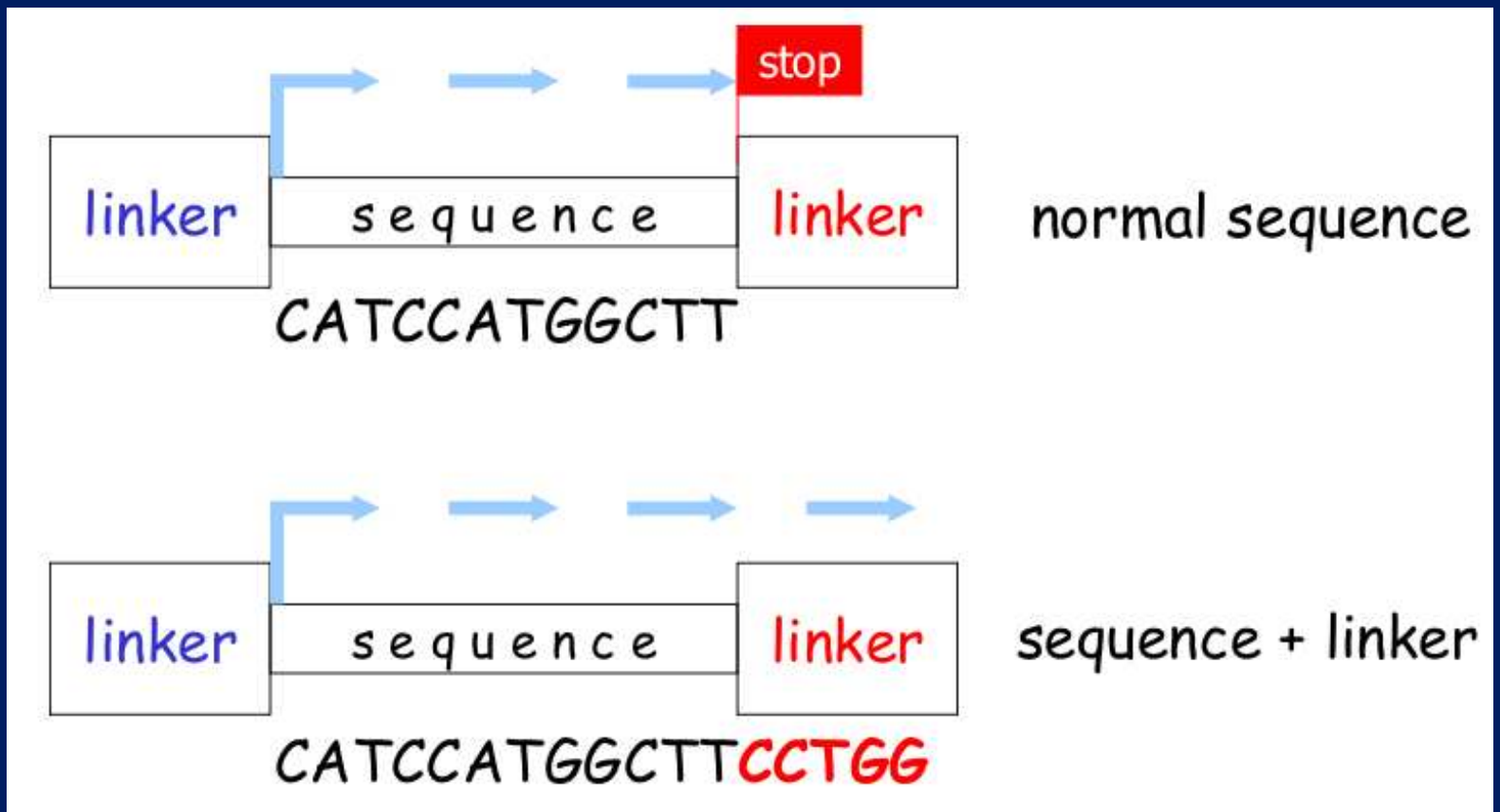


Figure 6: Sequencing linkers.

# Pre-alignment

## *Data cleaning and QC*

Depending on the sequencing platform, parts of the reads need to be removed.

- Remove linker sequences (*Cutadapt*, *FASTX toolkit*).
- Trim low quality reads at the end of the read (*Sickle*, *Trimmomatic*, *FASTX toolkit*).
- Length filtering (*Fastools*).

# Pre-alignment

## *Data cleaning and QC*

Depending on the sequencing platform, parts of the reads need to be removed.

- Remove linker sequences (*Cutadapt*, *FASTX toolkit*).
- Trim low quality reads at the end of the read (*Sickle*, *Trimmomatic*, *FASTX toolkit*).
- Length filtering (*Fastools*).

The *FastQC toolkit* can be used for quality control (both before and after the data cleaning step).

- Positional nucleotide content.
- GC distribution.
- Sequence quality distribution.
- ...

# Pre-alignment

## *Example QC output*

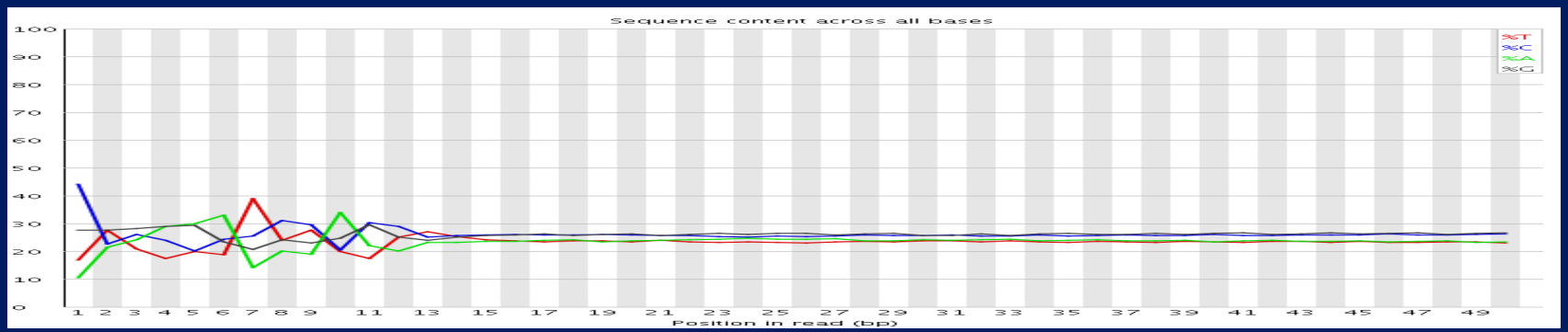


Figure 7: Positional nucleotide content.

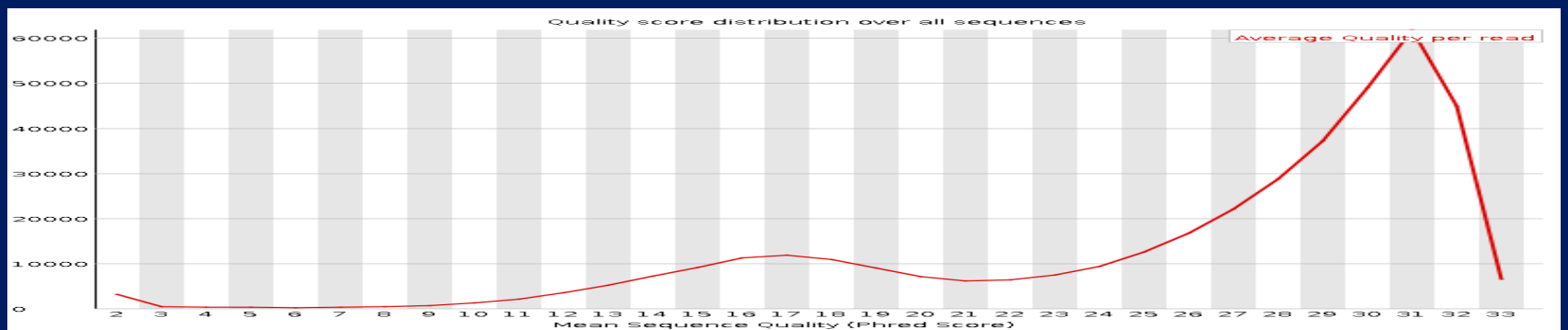


Figure 8: Sequence quality distribution.

# Alignment

## *Choose an aligner*

Alignment needs to be fault-tolerant.

# Alignment

***Choose an aligner***

Alignment needs to be fault-tolerant.

Not all aligners can deal with indels.

- Older aligners only allowed substitutions.

# Alignment

***Choose an aligner***

Alignment needs to be fault-tolerant.

Not all aligners can deal with indels.

- Older aligners only allowed substitutions.

Few aligners can work with large deletions.

- Spliced RNA.
  - *GMAP / GSNAP*.
  - *Tophat*.
- *BWA-MEM*.

# Alignment

## *Choose an aligner*

Alignment needs to be fault-tolerant.

Not all aligners can deal with indels.

- Older aligners only allowed substitutions.

Few aligners can work with large deletions.

- Spliced RNA.
  - *GMAP / GSNAP*.
  - *Tophat*.
- *BWA-MEM*.

The choice of aligner may be restricted by the sequencer.

- For the Ion Torrent / Proton: *Tmap*.
- For the PacBio: *BLASR*.

# Variant calling

## *Pileup*

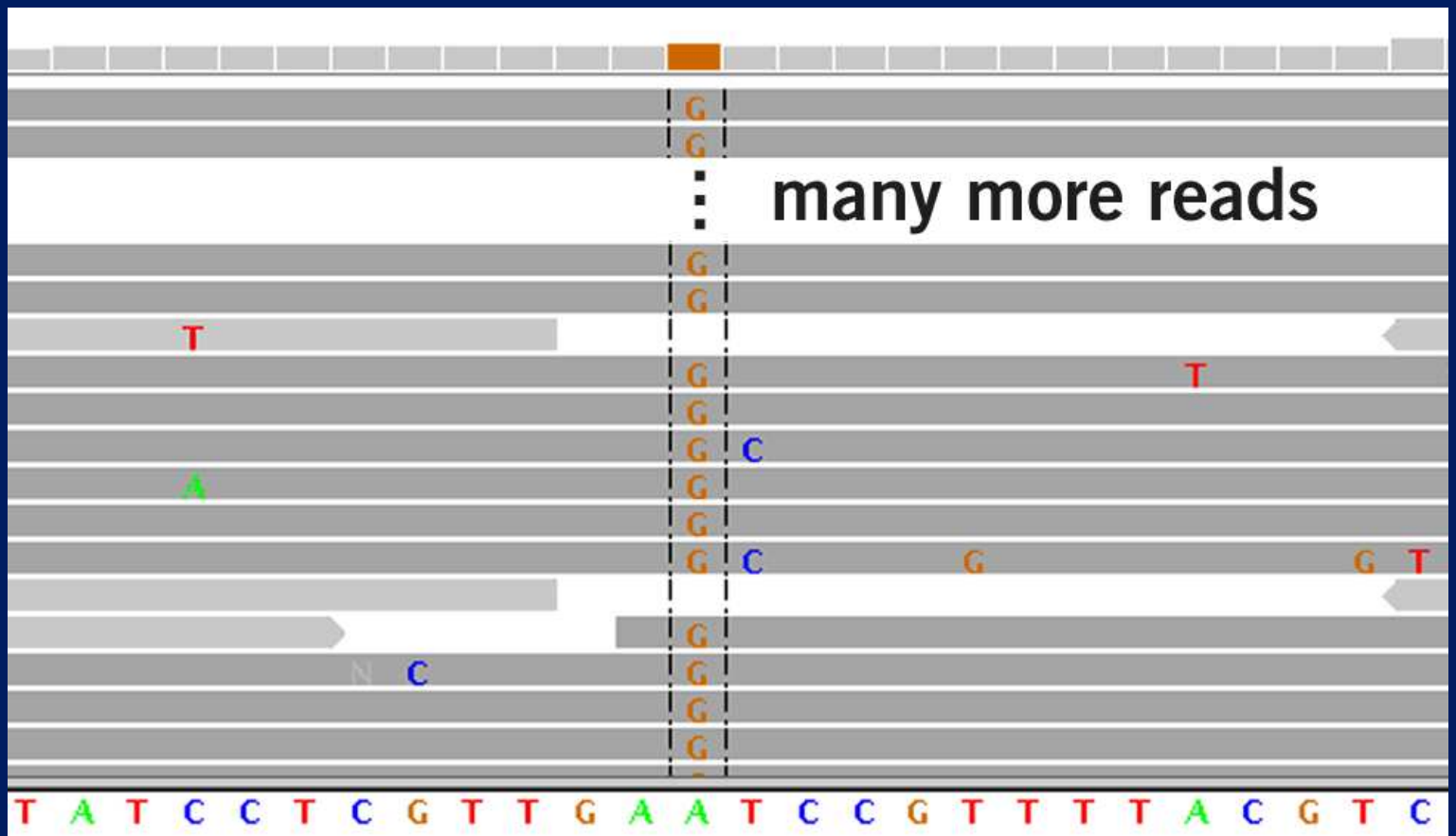


Figure 9: Result of an alignment.

# Variant calling

## *Some considerations*

Things a variant caller might take into account:

- Strand balance.
- Base quality.
- Mapping quality.
  - Distribution within the reads.
- Ploidity of the organism in question.

# Variant calling

***Some considerations***

Things a variant caller might take into account:

- Strand balance.
- Base quality.
- Mapping quality.
    - Distribution within the reads.
- Ploidity of the organism in question.

Complicating factors:

- Pooled samples.

# Variant calling

## *Some considerations*

Things a variant caller might take into account:

- Strand balance.
- Base quality.
- Mapping quality.
  - Distribution within the reads.
- Ploidity of the organism in question.

Complicating factors:

- Pooled samples.
- RNA.
  - Allele specific expression.
  - RNA editing.

# Variant calling

***Some considerations***

Things a variant caller might take into account:

- Strand balance.
- Base quality.
- Mapping quality.
  - Distribution within the reads.
- Ploidity of the organism in question.

Complicating factors:

- Pooled samples.
- RNA.
  - Allele specific expression.
  - RNA editing.
- Strand specific sampleprep.

# Variant calling

## *Choice of variant caller*

Rules of thumb:

- Well known organism and experiment: Statistical model.
- Use a simpler variant caller otherwise.

# Variant calling

## *Choice of variant caller*

Rules of thumb:

- Well known organism and experiment: Statistical model.
- Use a simpler variant caller otherwise.

Popular variant callers:

- *Samtools*.
- *GATK*.
- *VarScan*.

# Variant filtering

## *Filtering on coverage*

We can set some thresholds:

- Minimum.
- Maximum.

# Variant filtering

***Filtering on coverage***

We can set some thresholds:

- Minimum.
- Maximum.

We filter for a maximum coverage because of copy number variation.

# Variant filtering

***Filtering on coverage***

We can set some thresholds:

- Minimum.
- Maximum.

We filter for a maximum coverage because of copy number variation.

A reasonable way to calculate the maximum:

- 2.5 times the average coverage of the targeted regions.

# Variant filtering

***Filtering on coverage***

We can set some thresholds:

- Minimum.
- Maximum.

We filter for a maximum coverage because of copy number variation.

A reasonable way to calculate the maximum:

- 2.5 times the average coverage of the targeted regions.

A better way:

- Do CNV calling and use this as a filter.

# Annotation

***Effect prediction***

In most cases we have too many variants.

Variant annotation.

- Frequency within a population.
- Location of the variant.
  - Gene panels.
  - Location within a gene.
- Conservation.

# Annotation

## *Variant Effect Predictor*

A selection of VEP annotation:

- Affected genes and transcripts.
- Location of the variant.
  - Upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory region.
- Consequence on the protein sequence.
  - Stop gained, missense, stop lost, frameshift.
- Minor allele frequencies from the 1000 Genomes Project.
- SIFT and PolyPhen scores for changes to protein sequence.

`http://www.ensembl.org/info/docs/tools/vep/index.html`

# Pipelines

## *Combining tools*

```
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 tools*

```
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.

# Graphical interfaces

## *Galaxy*

Galaxy: a graphical user interface:

- Wrapper for command line utilities.
- User friendly.
- Point and click.

`http://galaxy.psu.edu/`

# Graphical interfaces

## *Galaxy*

Galaxy: a graphical user interface:

- 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/`

# Graphical interfaces

## *Galaxy*

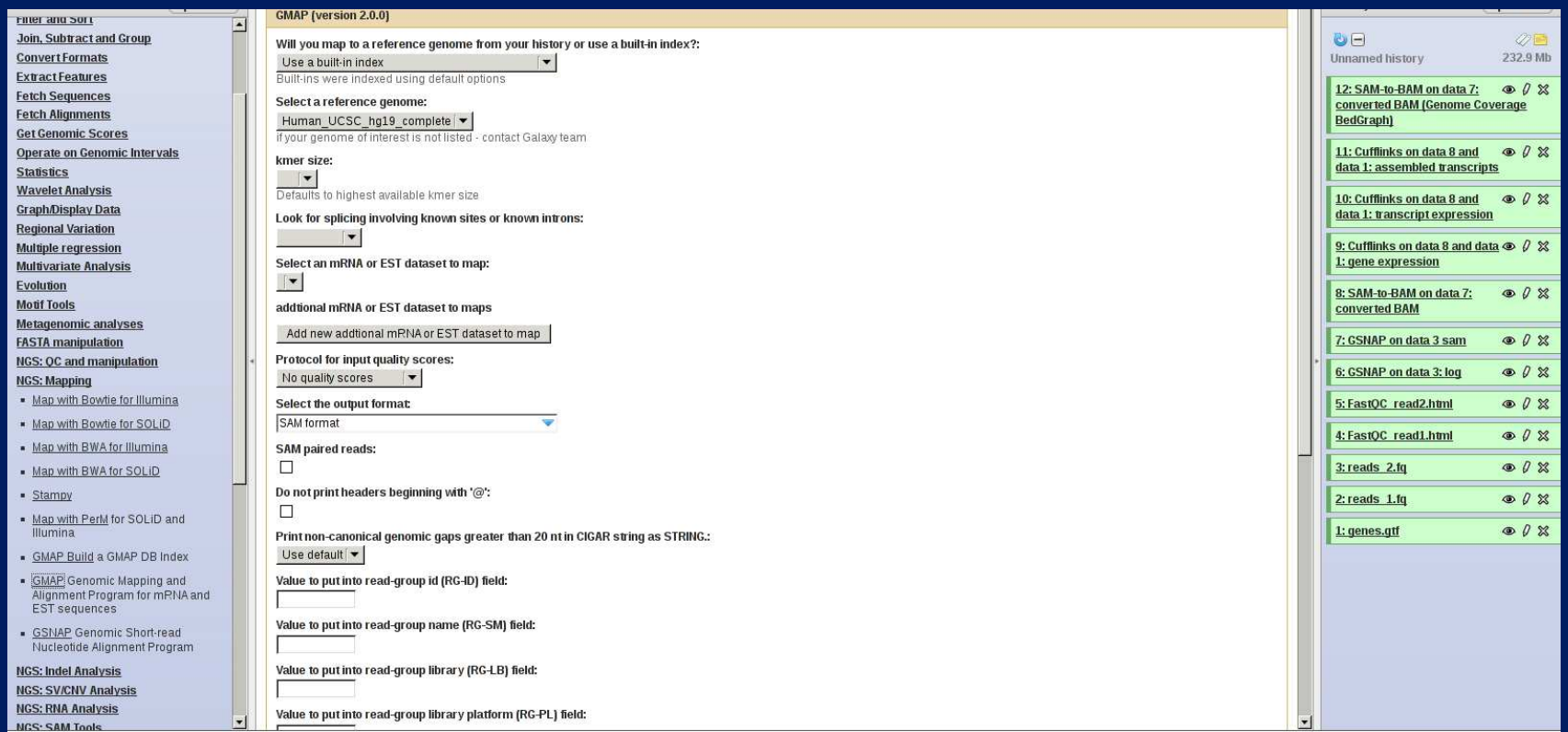


Figure 10: Galaxy main user interface.

# Graphical interfaces

## *Galaxy*

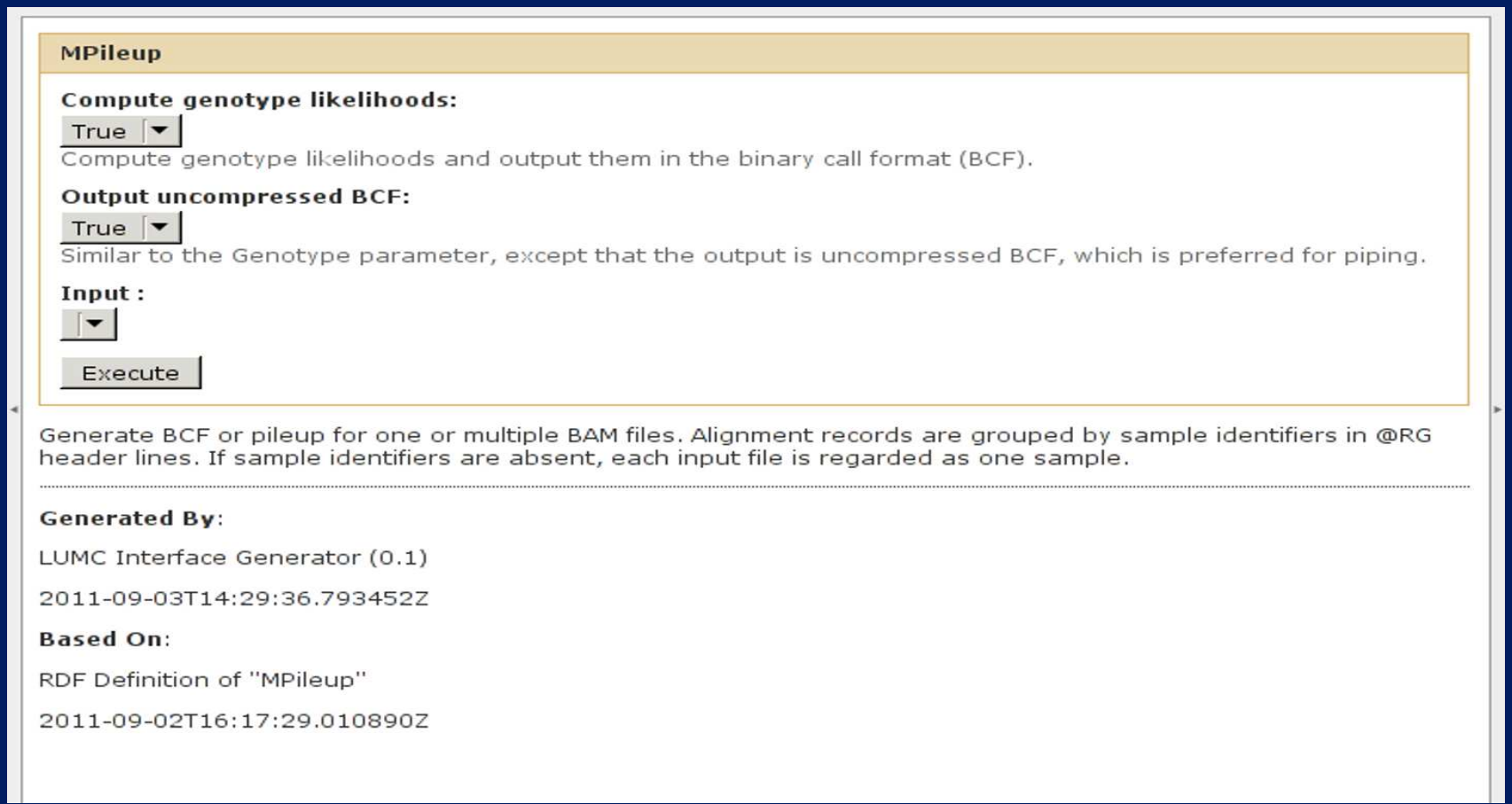


Figure 11: User friendly interface with Galaxy.

# Graphical interfaces

## *Workflow of a parallel pipeline*

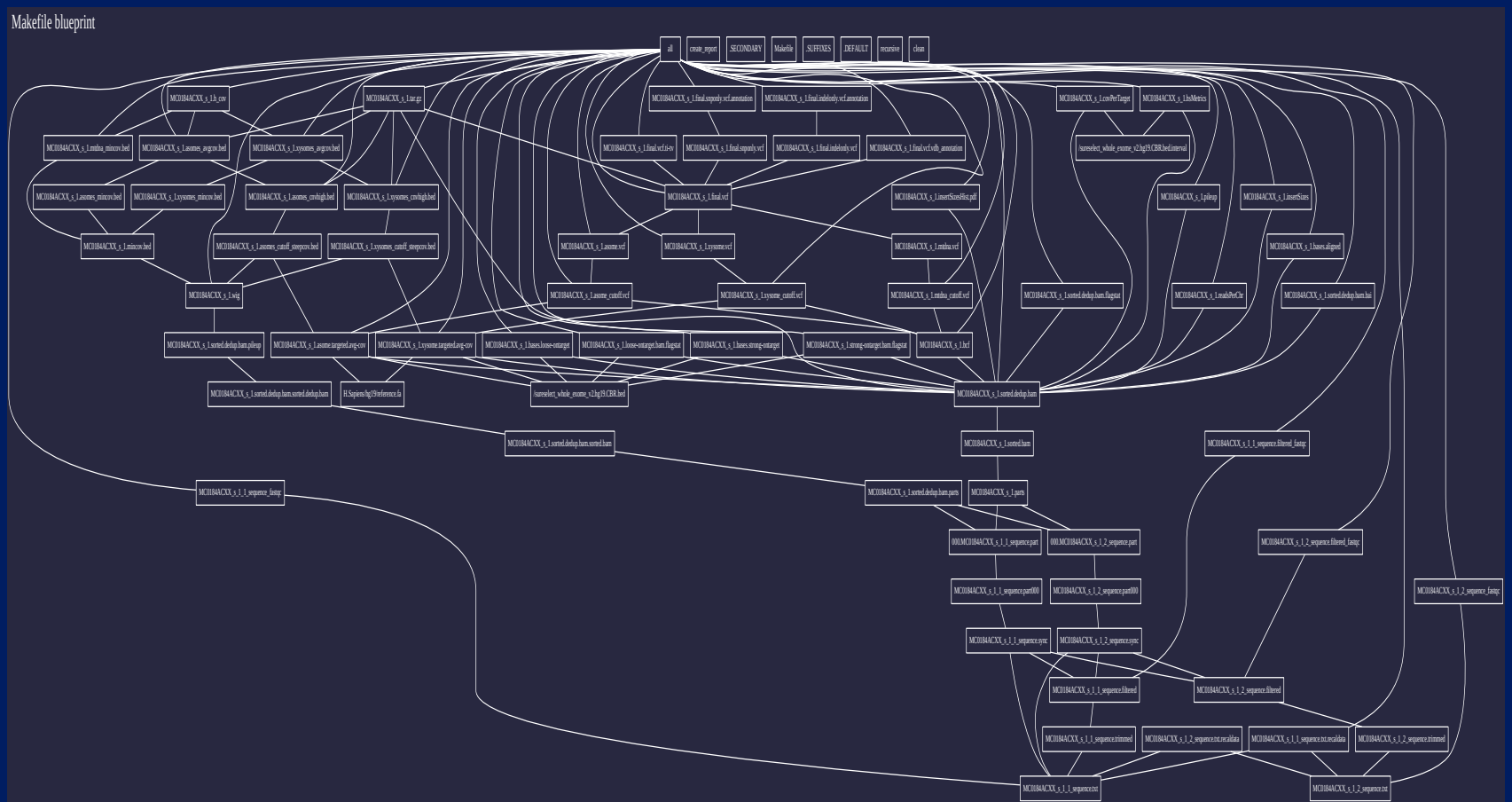


Figure 12: Dependency diagram.

# Questions?

Acknowledgements:

Martijn Vermaat
Michiel van Galen
Johan den Dunnen