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

Sequencers: HiSeq

Figure 3 : HiSeq 2000.

Characteristics:

- High throughput.
- Paired end.
- High accuracy.
- Read length 2×150bp.
- Relatively long run time.
- Relatively expensive.

Introduction

Sequencers: Ion Torrent

Figure 4 : Ion torrent.

Characteristics:

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

Introduction

Data analysis

Resequencing pipelines can roughly be divided in five steps.

Introduction

Data analysis

Resequencing pipelines can roughly be divided in five steps.

1. Pre-alignment.
    - Quality control.
    - Data cleaning.

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.

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.

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.

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

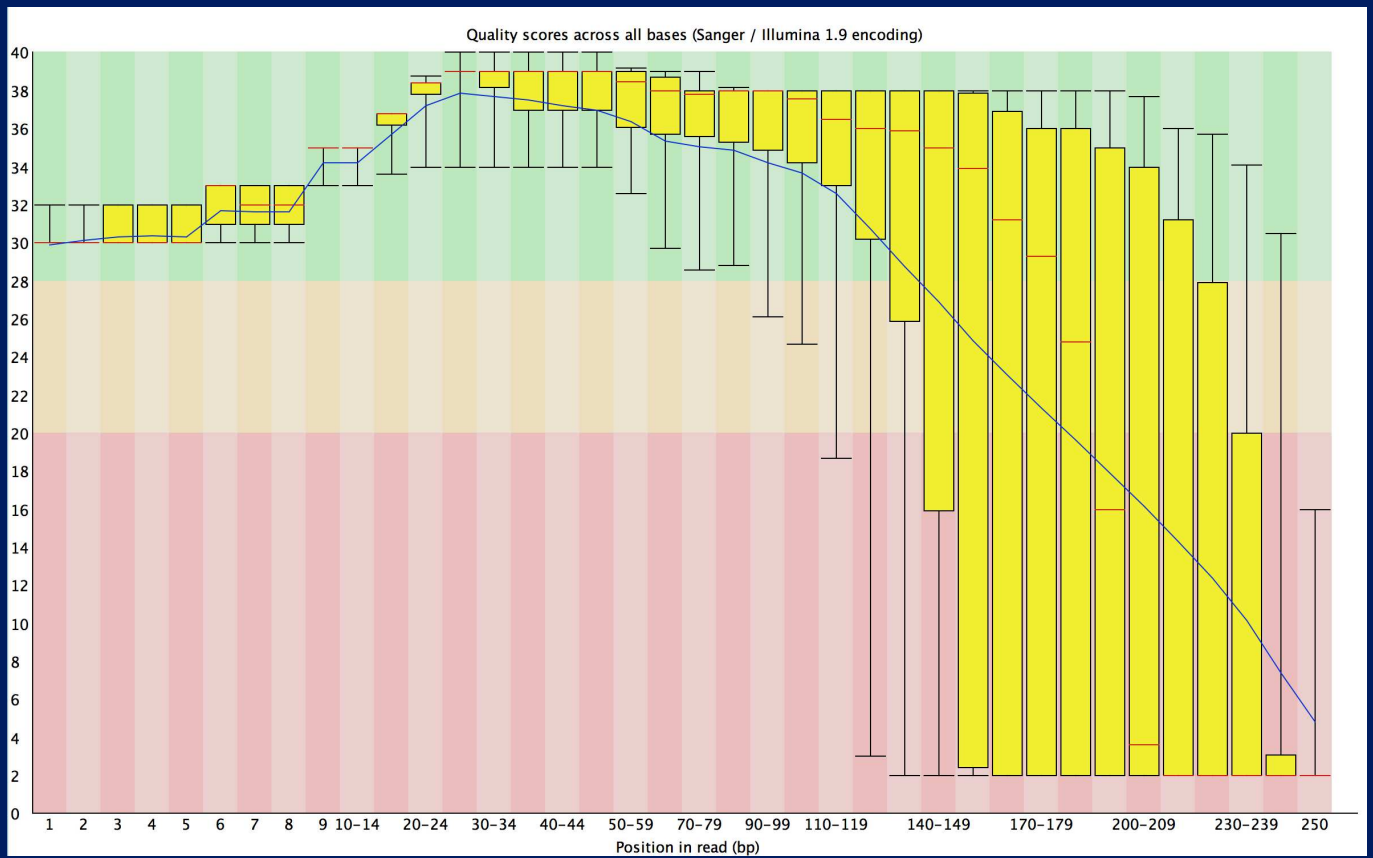

Figure 5 : Quality score per position.

Pre-alignment

Clipping

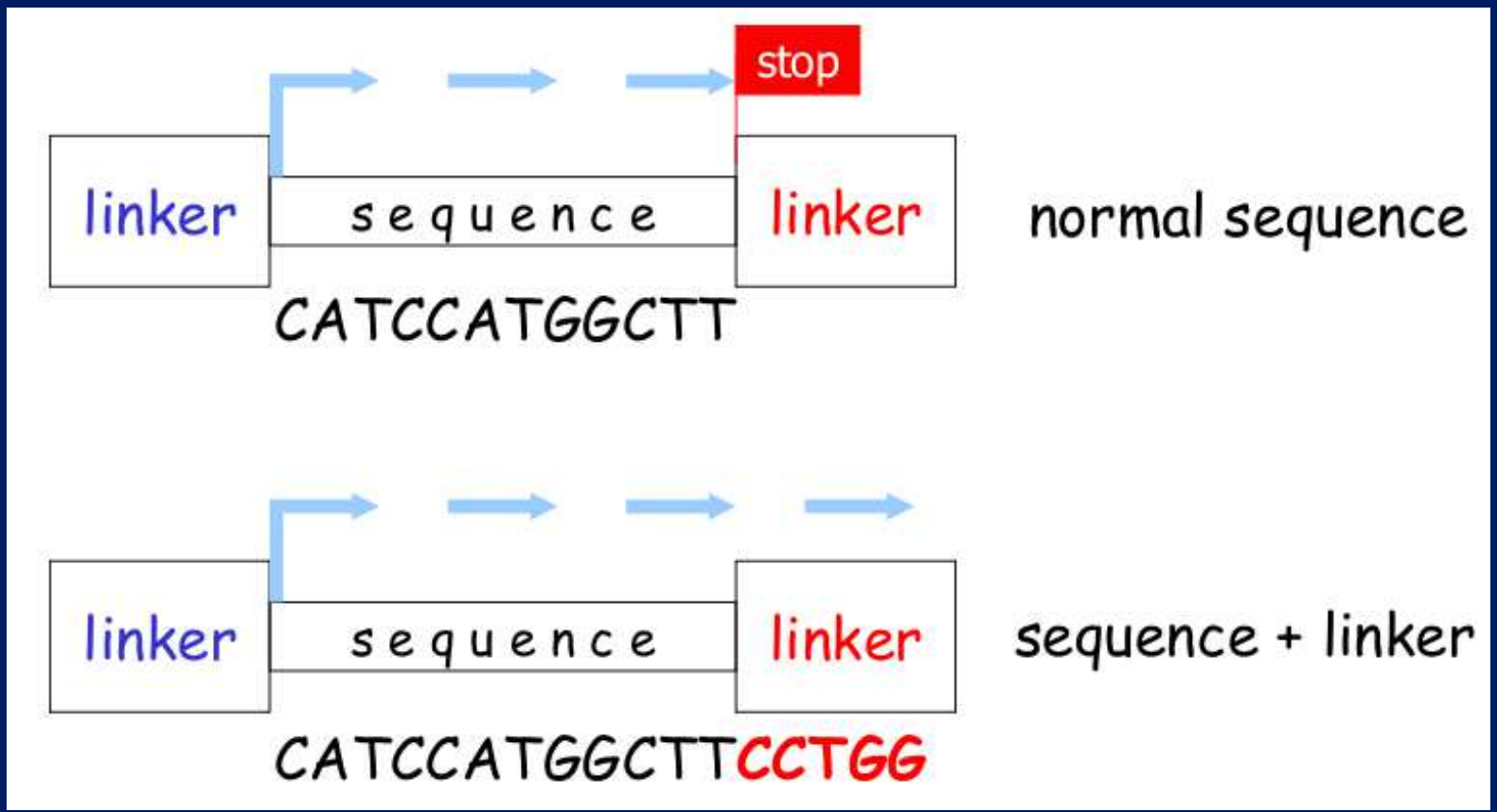

Figure 6 : Sequencing linkers.

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*).
- Clip 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*).
- Clip 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).

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

Pre-alignment

Example QC output

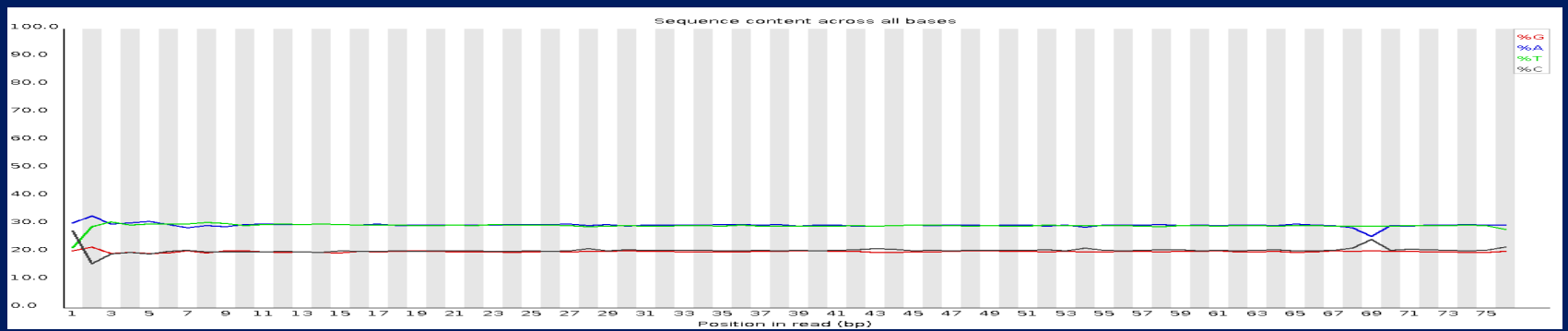

Figure 7 : Per base sequence content.

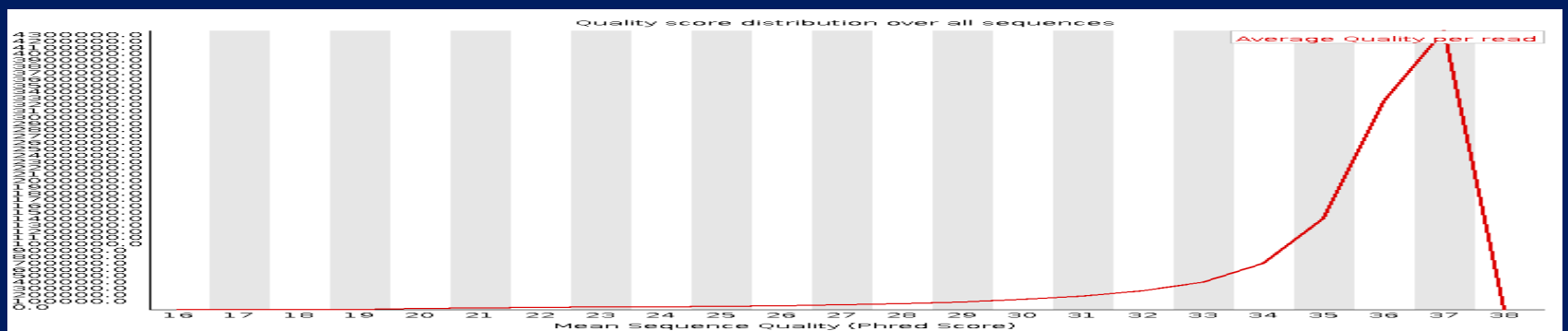

Figure 8 : Per sequence quality.

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.

- Only a couple of years ago, only SNPs were considered.
  - *Bowtie*.

Alignment

Choose an aligner

Alignment needs to be fault-tolerant.

Not all aligners can deal with indels.

- Only a couple of years ago, only SNPs were considered.
  - *Bowtie*.

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.

- Only a couple of years ago, only SNPs were considered.
  - *Bowtie*.

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: *Tmap*.
- For the PacBio: *BLASR*.

Variant calling

Pileup

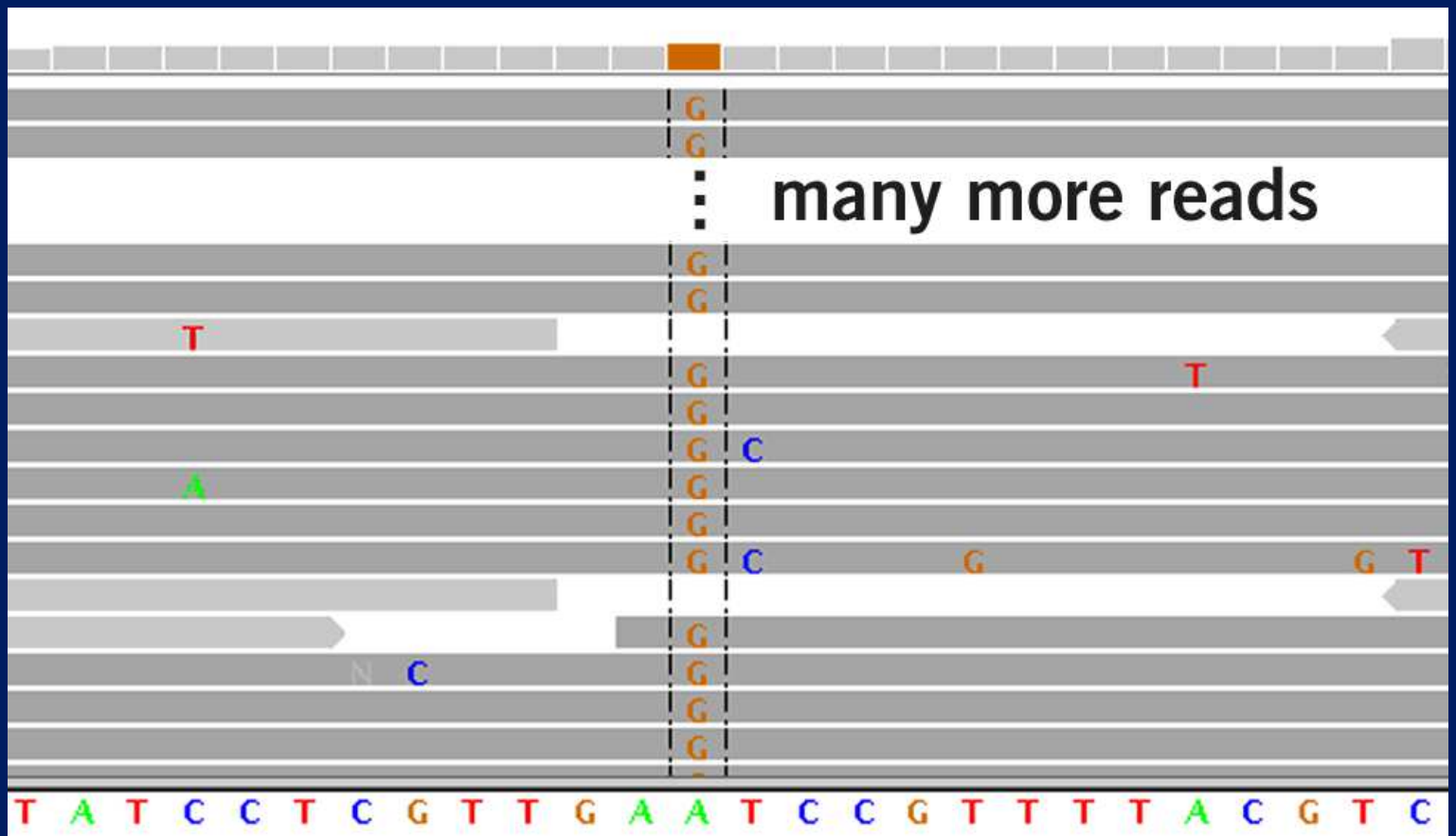

Figure 9 : Result of an alignment.

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.
  - More or less the same as ploidity.

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.
  - More or less the same as ploidity.
- RNA.
  - Allele specific expression.
  - Tissue 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.
  - More or less the same as ploidity.
- RNA.
  - Allele specific expression.
  - Tissue 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 good way to calculate the maximum:

- Calculate the mean coverage.
    - Only of the covered (targeted) regions.
- Multiply this number with a reasonable factor e.g., 2.5.

Annotation

What is already known about a variant

A selection of SeattleSeq annotation:

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

Annotation

What is already known about a variant

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

What is already known about a variant

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

What is already known about a variant

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

What is already known about a variant

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 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

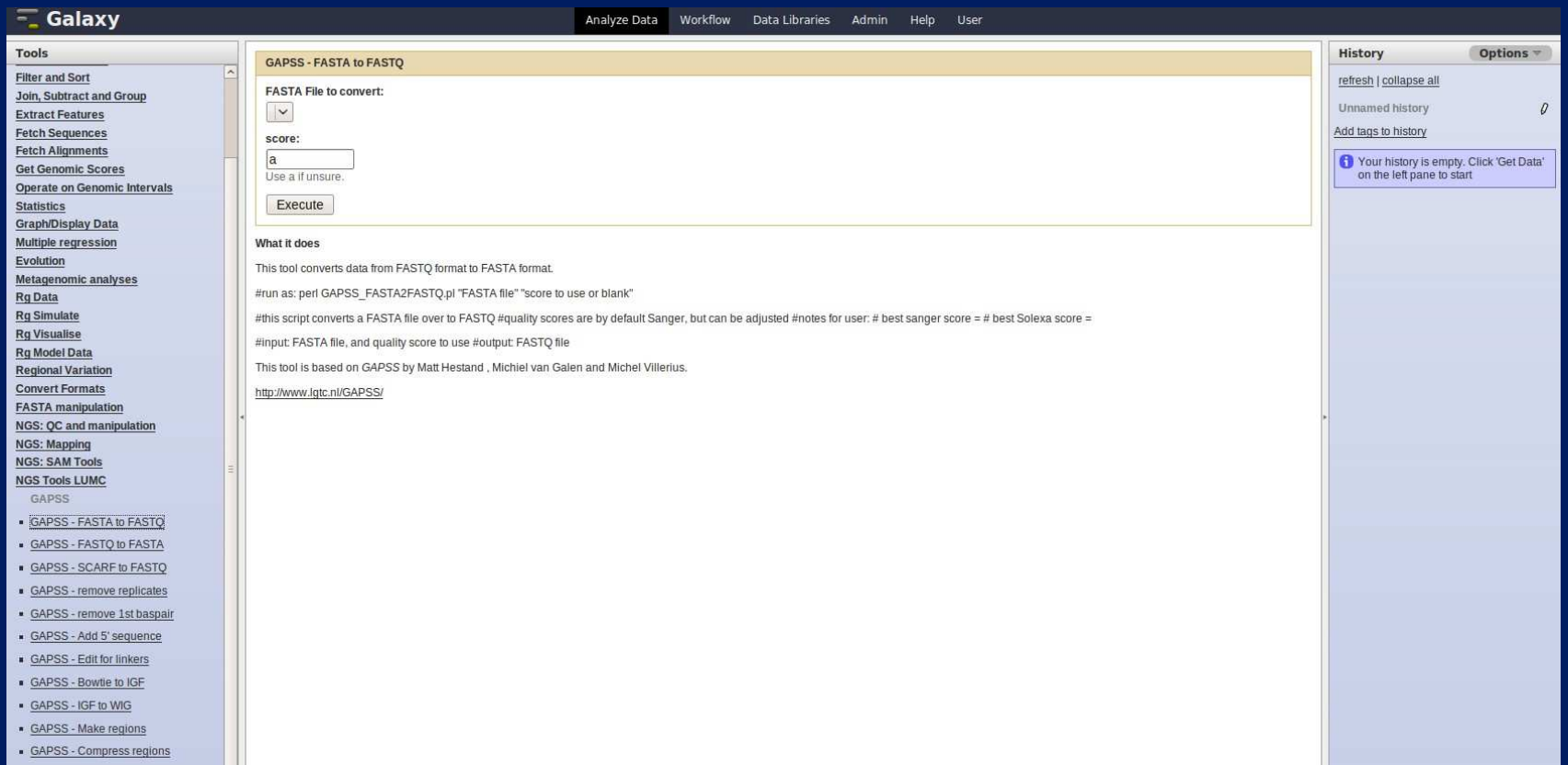

Figure 10 : Galaxy main user interface

http://galaxy.nbic.nl/

Graphical interfaces

Galaxy

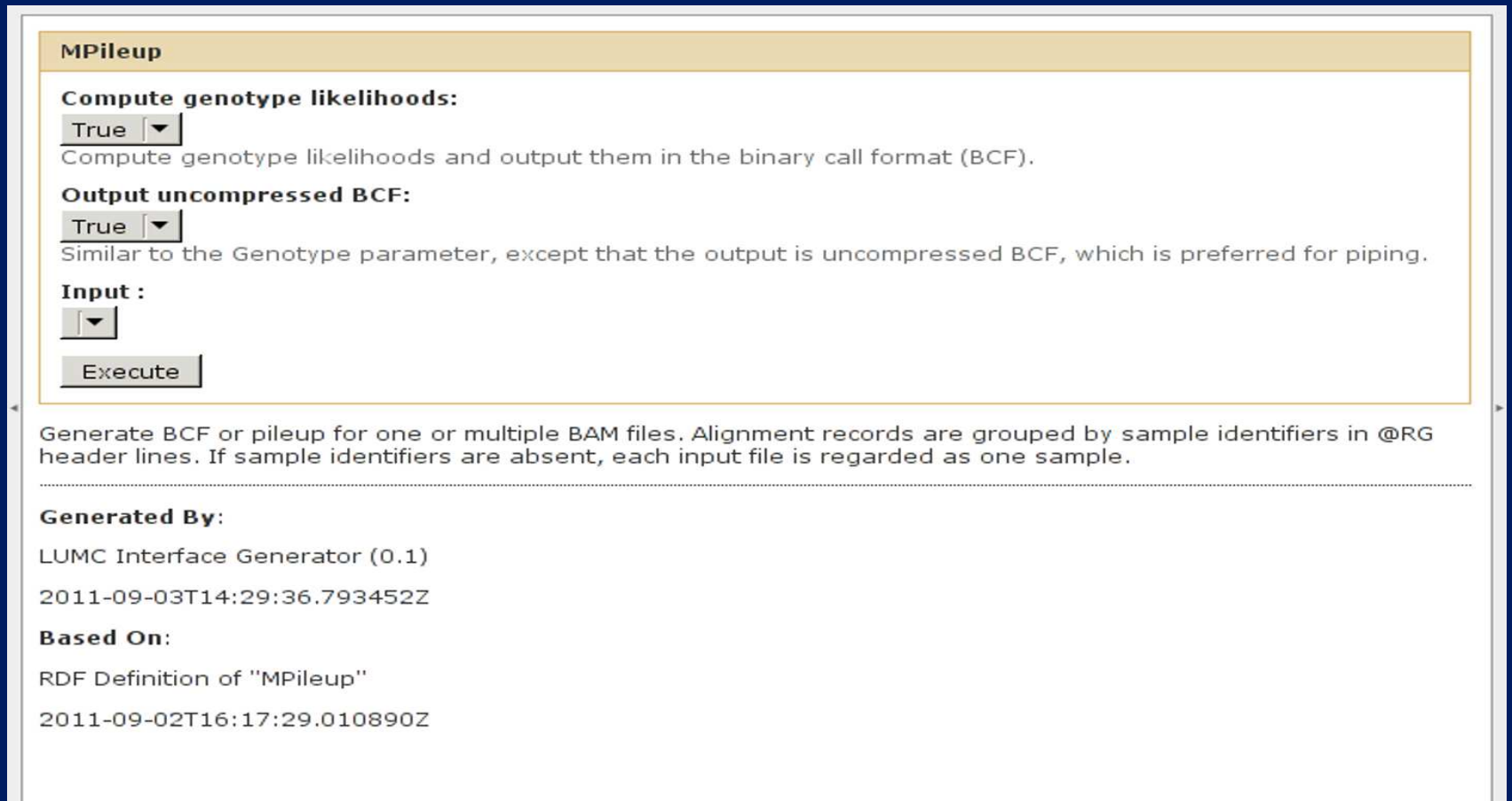

Figure 11 : User friendly interface with Galaxy

Graphical interfaces

Workflow of a parallel pipeline

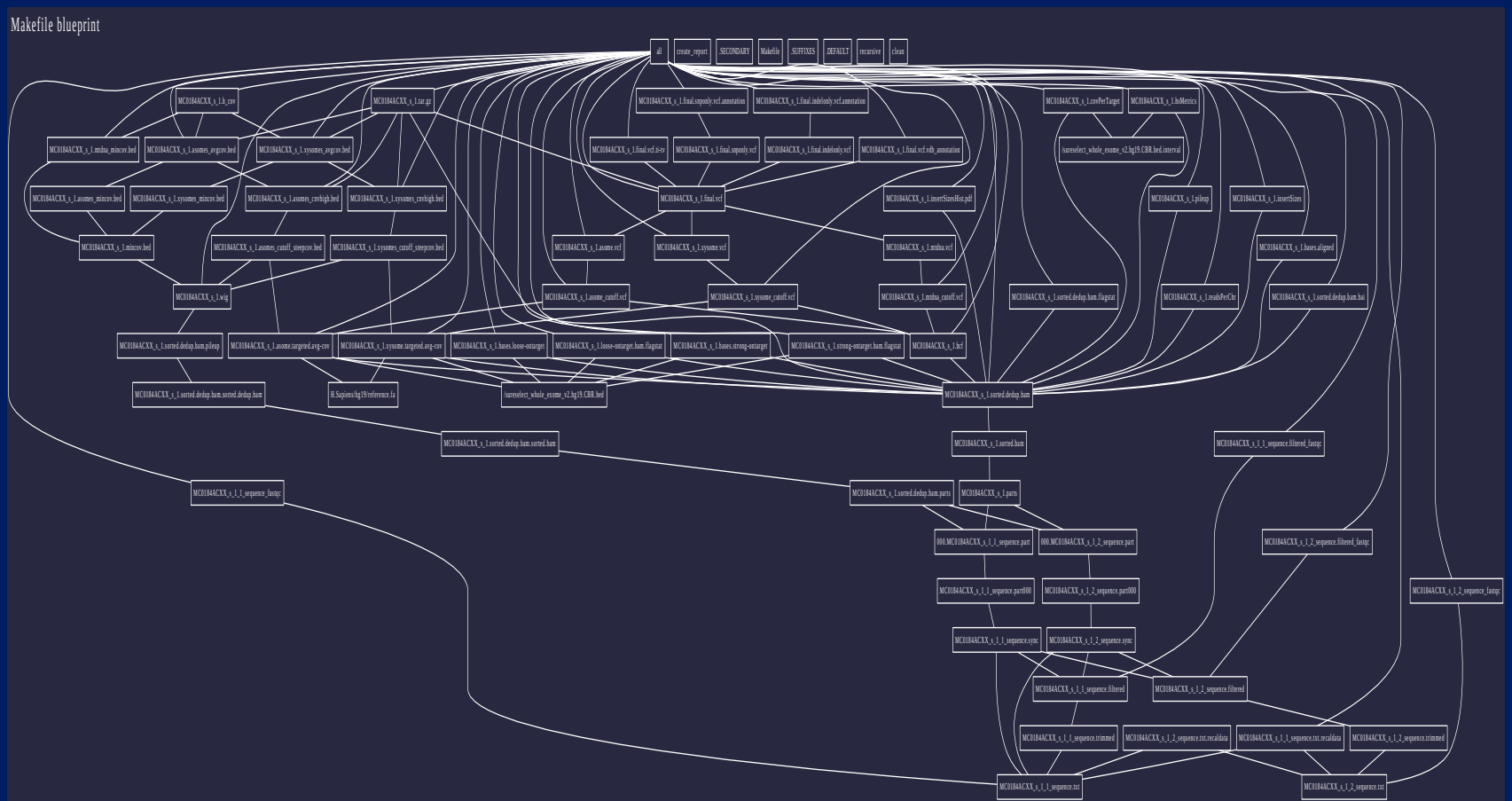

Figure 12 : Dependency diagram.

Graphical interfaces

Workflow of a parallel pipeline

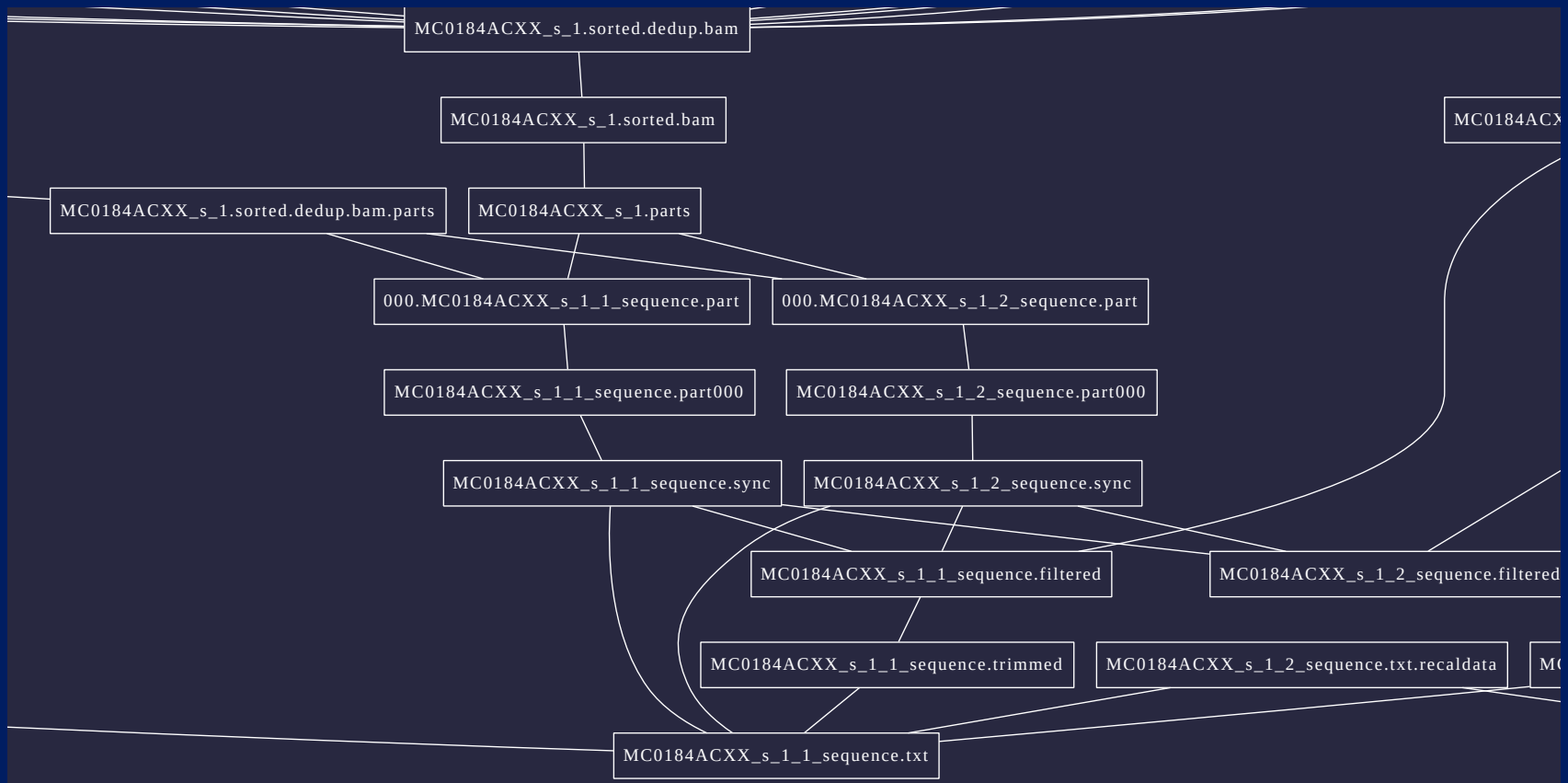

Figure 13 : Zoomed in.

Questions?

Acknowledgements:

Michiel van Galen
Martijn Vermaat
Johan den Dunnen