LEIDEN UNIVERSITY MEDICAL CENTER

RNA-seq using Galaxy

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Data generation.

Figure 1: HiSeq 2000.

Characteristics:

- Paired end.
  - 2×150 base pairs.
- High Throughput.
  - 150 Giga bases in 8 days.
- Cheap.

Introduction

General layout of an RNA-seq pipeline.

1. Pre-alignment.
    - QC.
    - Data cleaning.

Introduction

General layout of an RNA-seq pipeline.

1. Pre-alignment.
    - QC.
    - Data cleaning.
2. Alignment.
    - Use a specialised (RNA) aligner.

Introduction

General layout of an RNA-seq pipeline.

1. Pre-alignment.
    - QC.
    - Data cleaning.
2. Alignment.
    - Use a specialised (RNA) aligner.
3. Expression (gene, transcripts) analysis.
    - Known transcripts.

Introduction

General layout of an RNA-seq pipeline.

1. Pre-alignment.
    - QC.
    - Data cleaning.
2. Alignment.
    - Use a specialised (RNA) aligner.
3. Expression (gene, transcripts) analysis.
    - Known transcripts.
4. Transcript assembly.
    - New transcripts, alternative splicing, etc.

Pre-alignment

FastX / FastQC.

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

FastX / FastQC.

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 tool kit is used for quality control (both before and after the data cleaning step).

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

Pre-alignment

FastQC report.

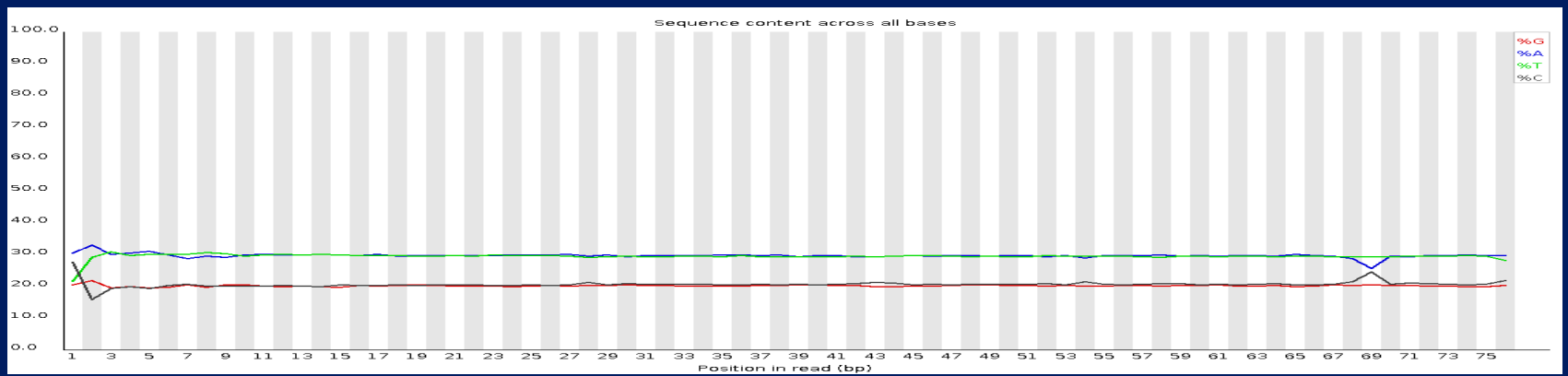

Figure 2: Per base sequence content.

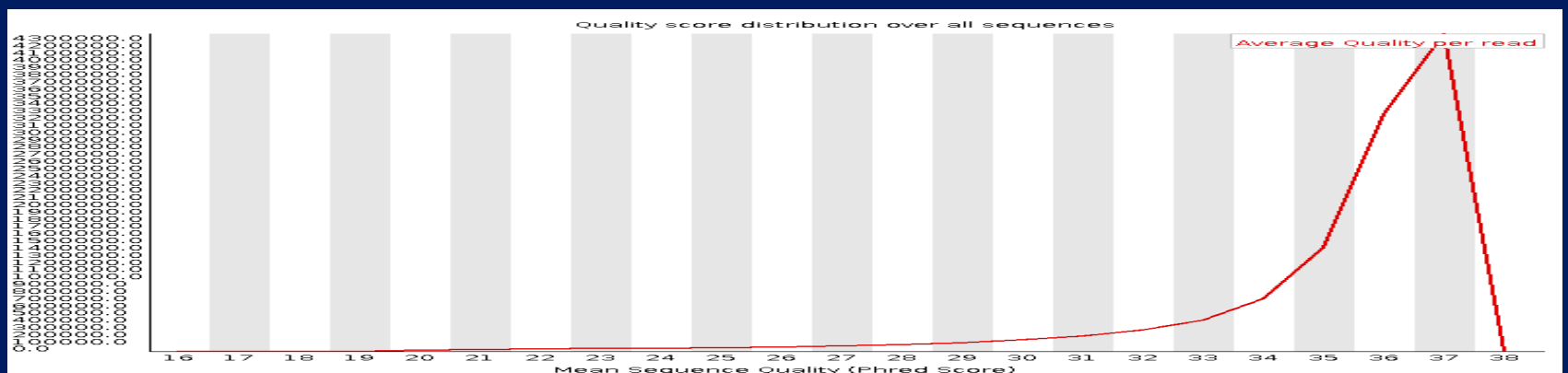

Figure 3: Per sequence quality.

Alignment

Gmap.

Gmap: A Genomic Mapping and Alignment Program for mRNA and EST Sequences.

Gsnap: Genomic Short-read Nucleotide Alignment Program.

http://research-pub.gene.com/gmap/

Alignment

Gmap.

Gmap: A Genomic Mapping and Alignment Program for mRNA and EST Sequences.

Gsnap: Genomic Short-read Nucleotide Alignment Program.

Some features:

- Split read alignment.
  - Split both ends.
  - Split a read into many pieces.
- Fast.
- Memory efficient.
  - No limit on intron size.

http://research-pub.gene.com/gmap/

Expression analysis and transcript assembly

Cufflinks.

Input:

- Aligned reads.
  - Tophat.
  - Gmap / Gsnap.

http://cufflinks.cbcb.umd.edu/

Expression analysis and transcript assembly

Cufflinks.

Input:

- Aligned reads.
  - Tophat.
  - Gmap / Gsnap.

What it can do:

- Assembled transcripts.
- Estimated transcript abundance.

http://cufflinks.cbcb.umd.edu/

Expression analysis and transcript assembly

Cufflinks.

Input:

- Aligned reads.
  - Tophat.
  - Gmap / Gsnap.

What it can do:

- Assembled transcripts.
- Estimated transcript abundance.

Differential expression and regulation (**`cuffcompare`**).

http://cufflinks.cbcb.umd.edu/

Pipelines

Combining 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 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.

Galaxy

Overview.

Data intensive biology for everyone.

- Open source.
- Web based.
    - No installation required.

http://galaxy.psu.edu/

http://galaxy.nbic.nl/

Galaxy

Overview.

Data intensive biology for everyone.

- Open source.
- Web based.
  - No installation required.

- 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

The Galaxy GUI.

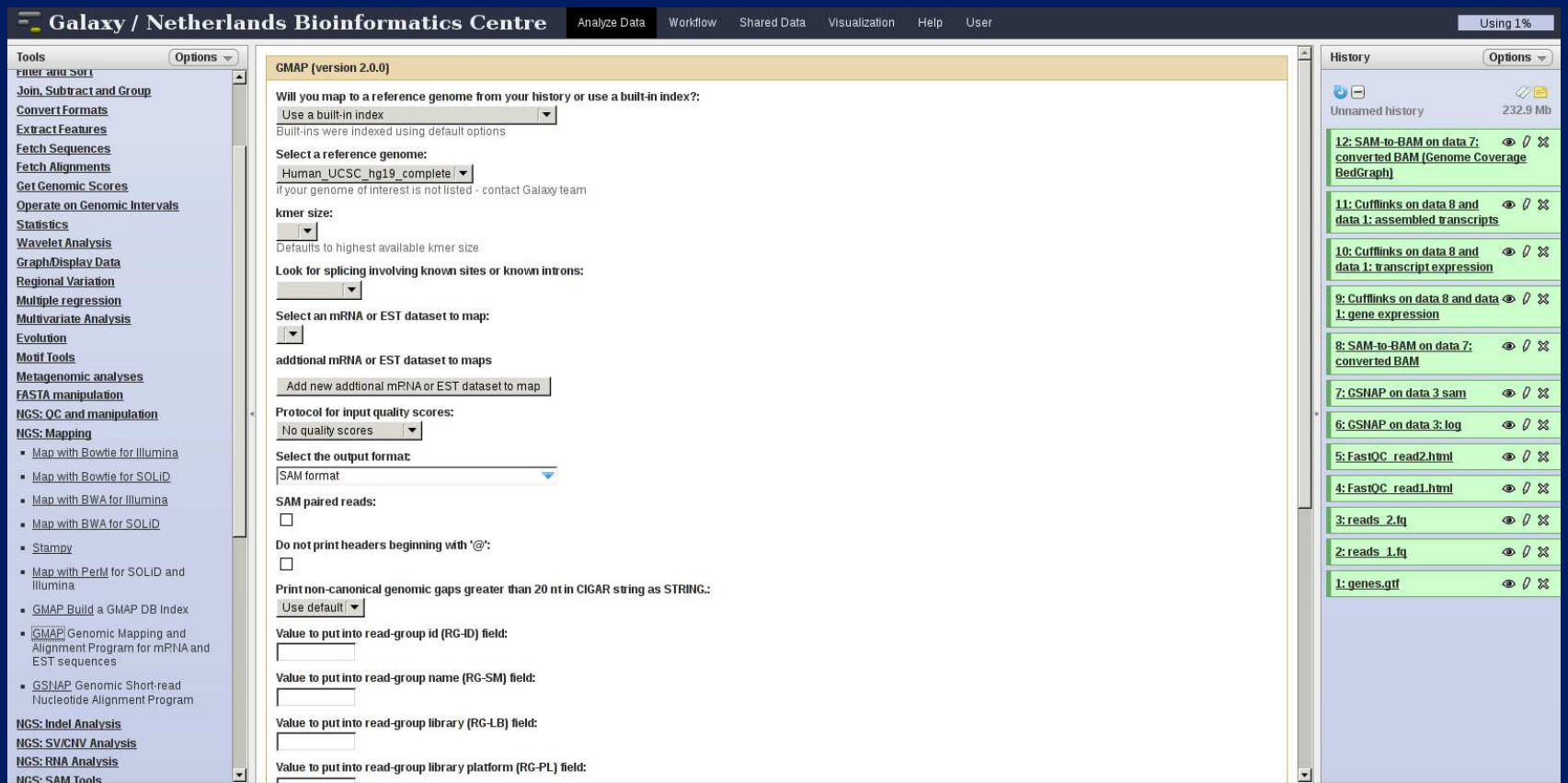

Figure 4: Galaxy panels.

Galaxy

Galaxy icons.

Figure 5: Collapsed history item.

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

- Click on the title for a more detailed view.

Galaxy

Galaxy icons.

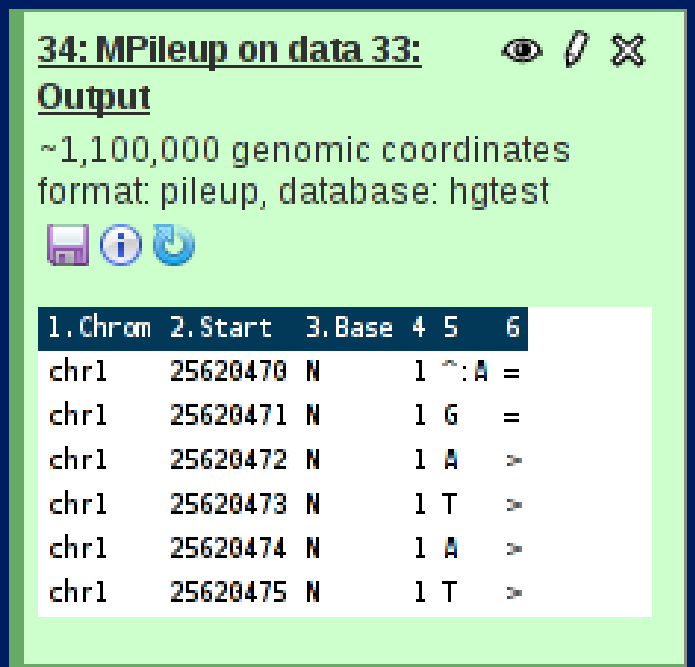

Figure 6: History item.

- Diskette: save.
- Blue looping arrow: rerun.

Galaxy

Outline of the practical

1. Do a typical RNA-seq analysis.
2. Workflows.
    - Rerun the analysis with no effort.

Questions?

Acknowledgements:

Peter-Bram ’t Hoen
Johan den Dunnen