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

Sequencers: HiSeq

Figure 1: HiSeq 2000.

Characteristics:

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

Introduction

Sequencers: Ion Torrent

Figure 2: Ion torrent.

Characteristics:

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

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 part of the read that is left.

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 part of the read that is left.

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.

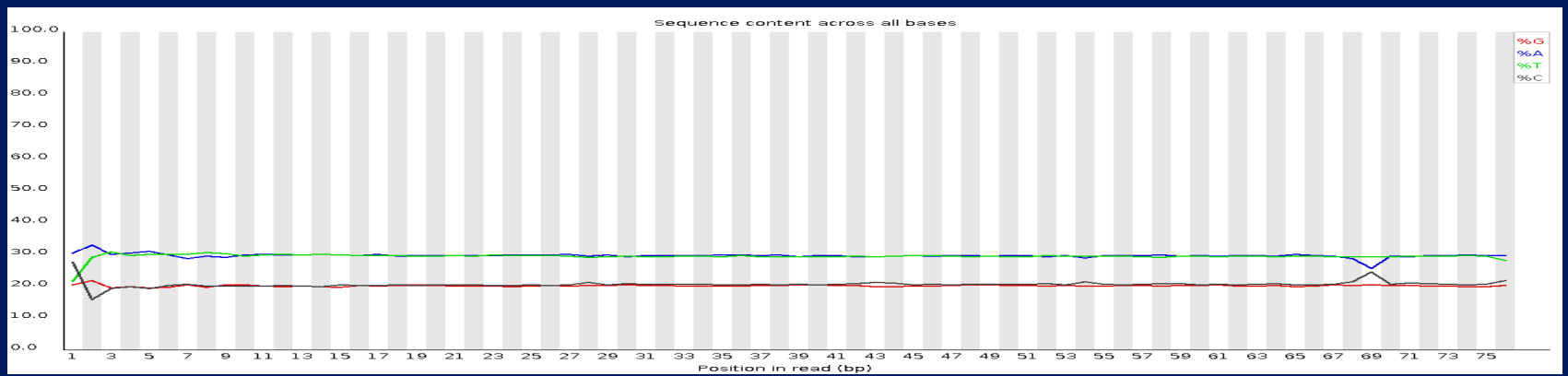

Figure 3: Per base sequence content.

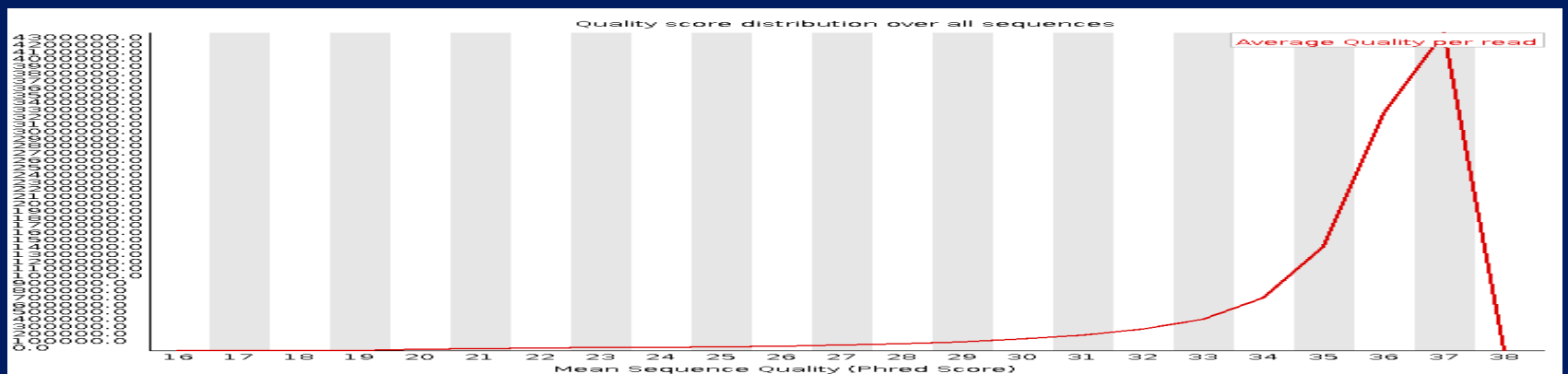

Figure 4: Per sequence quality.

Alignment

RNA aligners.

Difference with DNA:

- Splicing.

Alignment

RNA aligners.

Difference with DNA:

- Splicing.

This affects:

- Insert sizes.
- Mapping of reads that cover an exon-exon boundary.

Alignment

RNA aligners.

Difference with DNA:

- Splicing.

This affects:

- Insert sizes.
- Mapping of reads that cover an exon-exon boundary.

Available tools:

- Tophat.
- Gmap / Gsnap.
- PASSion.
- MapSplice.
- HMMSplicer.
- . . .

Alignment

Choose your aligner carefully.

If you work with pre-mRNA, the options are limited.

- Some tools find exons first, then use this to break up reads.
- Some tools prefer splitting reads over mapping them in an intron.

Alignment

Choose your aligner carefully.

If you work with pre-mRNA, the options are limited.

- Some tools find exons first, then use this to break up reads.
- Some tools prefer splitting reads over mapping them in an intron.

Some tools heavily rely on annotation.

- A list of known splice sites.
- Motives (canonical splice sites).

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/

Expression analysis and transcript assembly

Cufflinks.

Modes of operation:

- Use predefined transcripts.
- Assemble transcripts assisted by known transcripts.
- Assemble transcripts with no prior knowledge.

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

Expression analysis and transcript assembly

Cufflinks.

Modes of operation:

- Use predefined transcripts.
- Assemble transcripts assisted by known transcripts.
- Assemble transcripts with no prior knowledge.

When to use:

- Only interested in expression.
- Alternative splicing.

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

Variant calling

Principle of variant calling

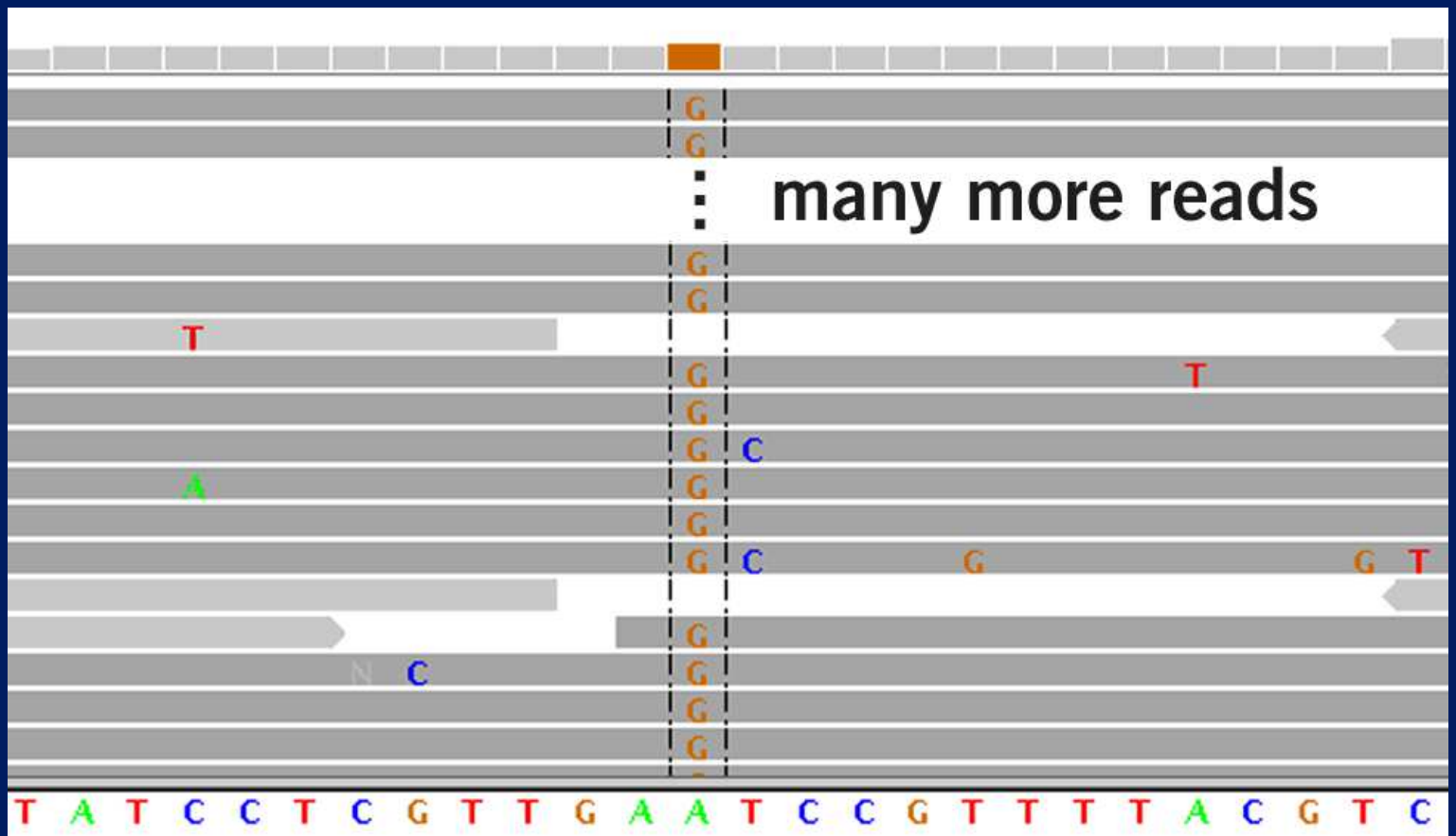

Figure 5: Result of an alignment.

Variant calling

Principle of variant calling

In principle, we call a variant when we are confident we have seen one.

Variant calling

Principle of variant calling

In principle, we call a variant when we are confident we have seen one.

But when are we confident?

- More than x times?
- In more than y percent of the reads covering the variant?

Variant calling

Principle of variant calling

In principle, we call a variant when we are confident we have seen one.

But when are we confident?

- More than x times?
- In more than y percent of the reads covering the variant?

Variant callers can use:

- Fixed settings.
- Statistical models.

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.

Some complications when analysing RNA:

- Allele specific expression.
  - Heterozygosity may not be detected.

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.

Some complications when analysing RNA:

- Allele specific expression.
  - Heterozygosity may not be detected.
- Tissue specific expression.
  - Some variants will be missed completely.

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.

Some complications when analysing RNA:

- Allele specific expression.
  - Heterozygosity may not be detected.
- Tissue specific expression.
  - Some variants will be missed completely.
- RNA editing.
  - Some variants will not be present on DNA.

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.

Some complications when analysing RNA:

- Allele specific expression.
  - Heterozygosity may not be detected.
- Tissue specific expression.
  - Some variants will be missed completely.
- RNA editing.
  - Some variants will not be present on DNA.
- Strand specific sampleprep.

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.

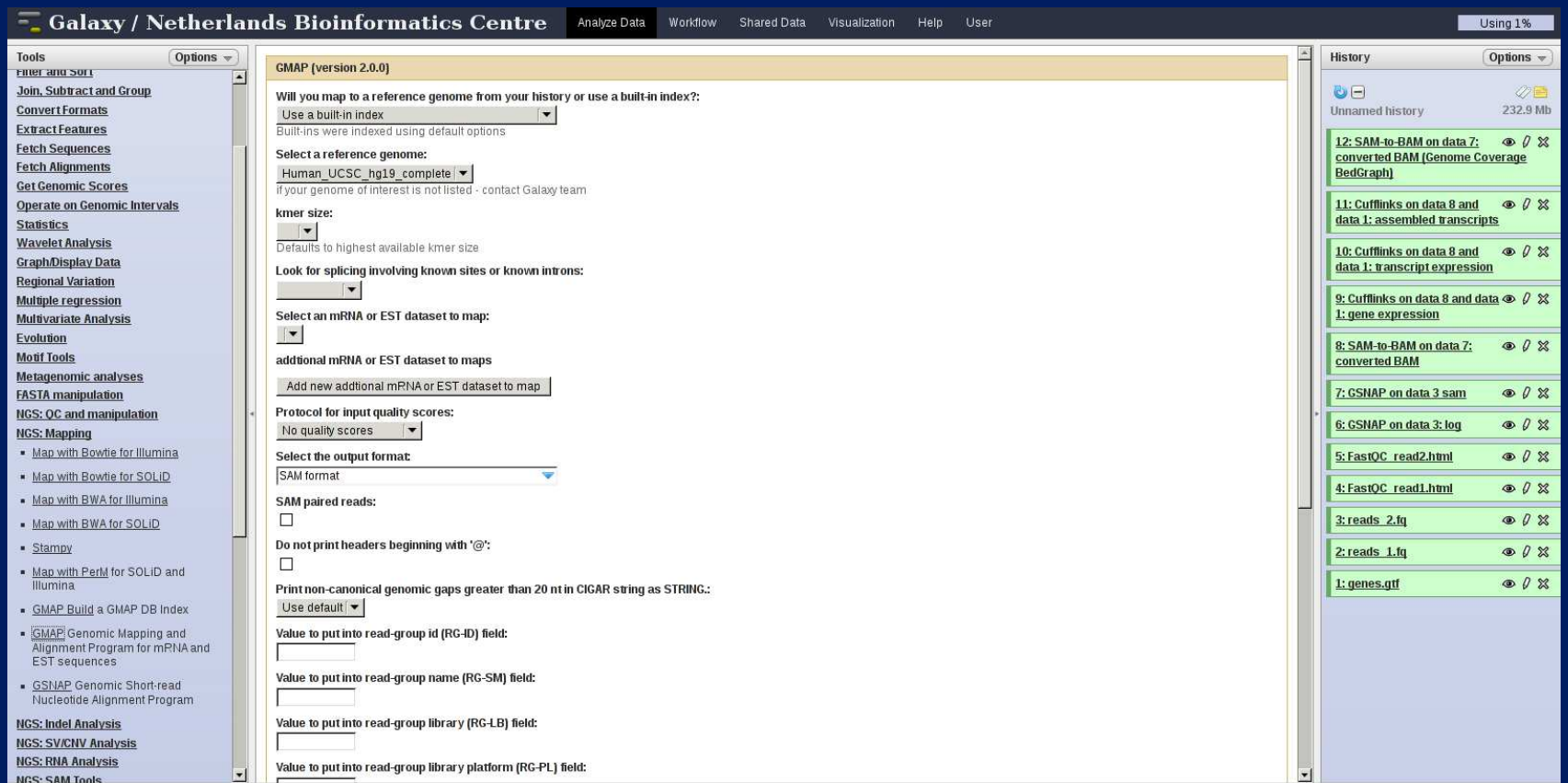

Figure 6: Galaxy panels.

Galaxy

Galaxy icons.

Figure 7: Collapsed history item.

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

- Click on the title for a more detailed view.

Galaxy

Galaxy icons.

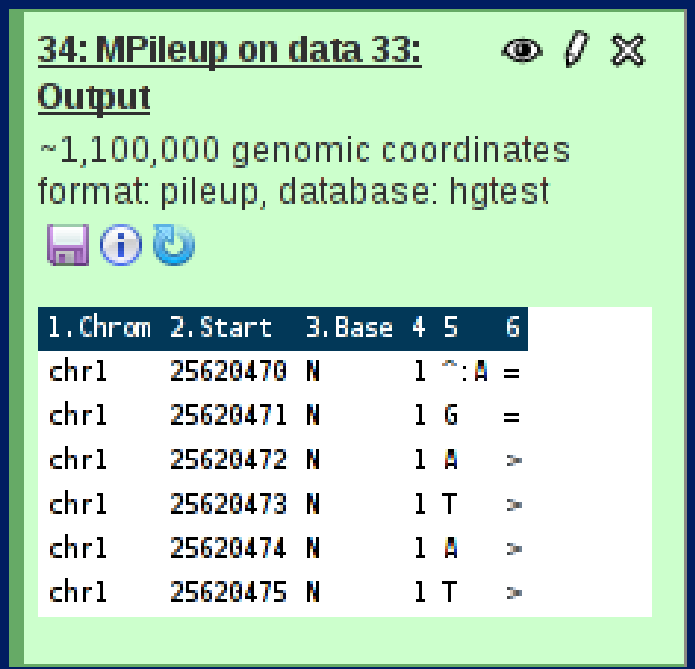

Figure 8: History item.

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

Galaxy

Outline of the practical

1. Do a typical RNA-seq analysis.
    - Expression.
    - Novel transcripts.
2. Variant calling.
3. Workflows.
    - Rerun the analysis with no effort.

Questions?

Acknowledgements:

Hailiang Mei
Michiel van Galen
Martijn Vermaat
Johan den Dunnen