

How to deal with your RNA-seq data ?

Rachel Legendre & the RNA-seq team

Summary

01

Bioinformatics

Quality control, Mapping, Counting

02

Statistics

Experimental design, Exploratory data analysis

03

Statistics

Normalization, modelisation and troubleshooting

04

Practice

Differential analysis with SARTools

05

Advanced practice

Gene Sets Analysis methods

06

Bioinformatics

Transcriptome *de novo* assembly



Bioinformatics

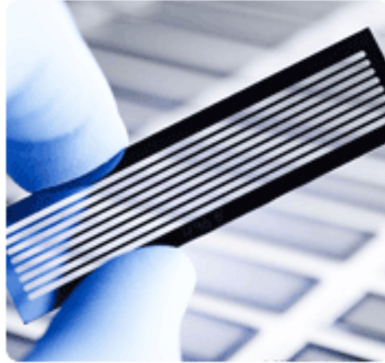
Introduction and prerequisites



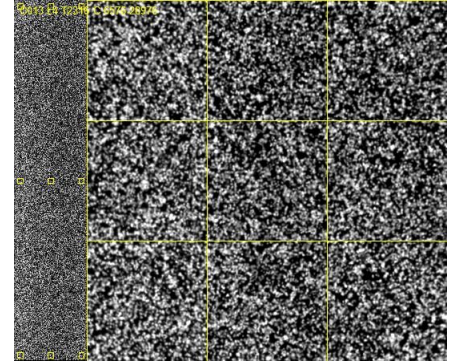
Raw NGS data



Instrument



Flowcell



Intensities

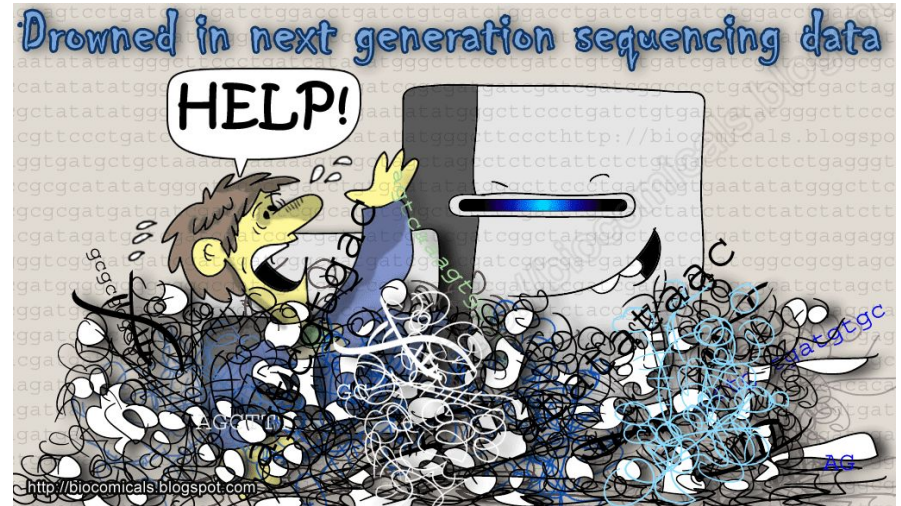
Data storage: Hiseq2500

- Text file with size between 100 to 150 Gb by lane
- Let's compare : War and peace by Léon Tolstoï
 - 1817 pages
 - 6 cm width
 - 4 Mb
- 1 lane :
 - 25 000 times "war and peace"
 - 45 Millions pages
 - 1.5 km (5 Eiffel towers)
- 8 lane by flow cell => 1 Tb of raw data/ week / sequencer
- Times 2 for paired-end



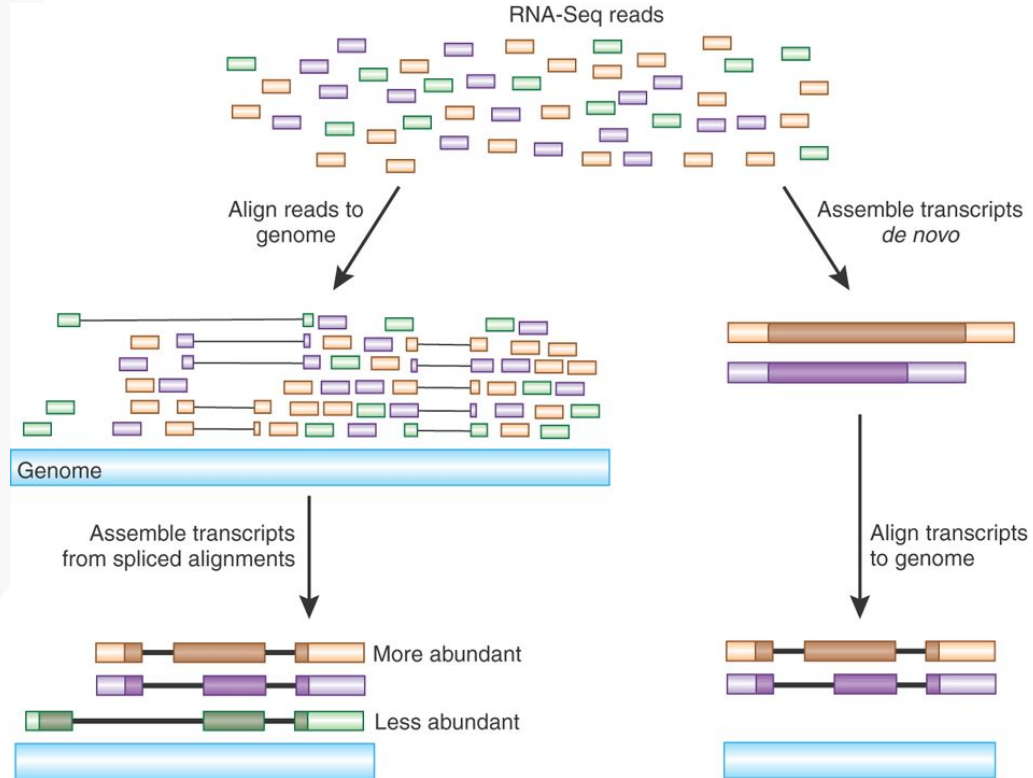
Data storage: NovaSeq6000

- Text file with size between 80Gb to 3Tb (in single flowcell mode)
- Let's compare : War and peace by Léon Tolstoï
 - 1817 pages 4250
 - 6 cm width
 - 4 Mb
- 1 run :
 - 750 000 times "war and peace"
 - 1350 Millions pages
 - 45 km (138 Eiffel towers)
- Times 2 for dual flowcell mode



RNA-seq applications

« Transcriptome analysis provides information about the identity and quantity of all RNA molecules in one cell or a population of cells »



RNA-seq: Why ? How



Ask right question before libraries preparation and sequencing:

Prokaryotes



I don't find a ribo-depletion kit for my organism:

→ Design yourself the oligos

I want to identify antisense RNA:

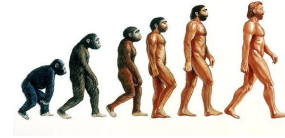
→ Directional protocol (standard)

I'm interested in transposons:

→ Longer read sequencing

→ Paired-end sequencing

Eukaryotes



I want coding genes only:

→ PolyA strategy

I want non-coding genes also:

→ Ribo Depletion

I'm interested in small RNA profiling:

→ Use specific protocole

I'm interested in isoforms:

→ Paired-end sequencing

→ Long read technologies

RNA-seq: Why ? How

Regardless of your organism:

- Complexity of your genome and the biological question paired end or single end, length of reads ?
- Sequencing depth (multiplexing rate)
- More biological replicates than more sequencing depth
- Stranded RNA-seq protocol to assigned reads to a particular strand

RNA-seq: Why ? How

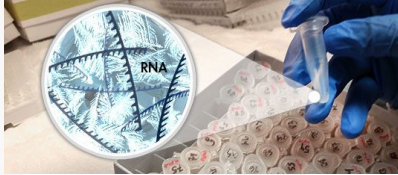
Regardless of your organism:

- Complexity of your genome and the biological question paired end or single end, length of reads ?
- Sequencing depth (multiplexing rate)
- More biological replicates than more sequencing depth
- Stranded RNA-seq protocol to assigned reads to a particular strand

For a successful experiment, it's imperative to include bioinformaticians and biostatistician before the beginning of the RNA extraction



Prerequisites



RNA sample:

- DNase treatment
- Quantity (adapted protocole)
- Quality (RNA integrity number > 7)
- Stocked at -80°C



Reference genome:

Complete genomic sequence in fasta format



Annotation file:

All features (genes, CDS, intron, UTR) of genome in GFF/GTF format

Where find the genome and the annotation ?

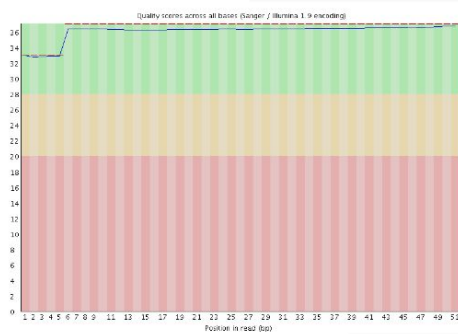
Common databases



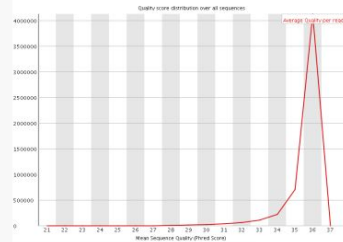
Specific databases



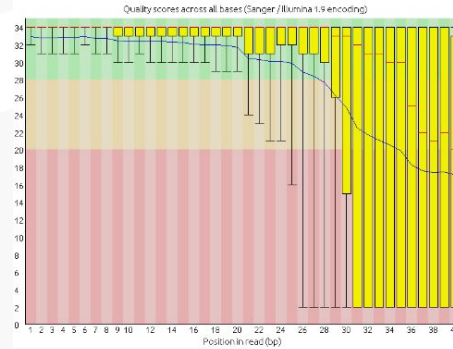
FASTQC: explore quality scores



Illumina HISEQ2500

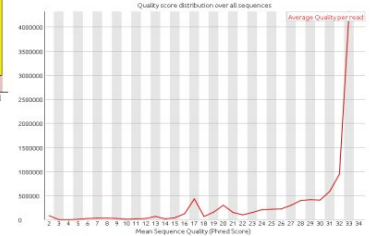


✓ The per base sequence quality are very high along sequence



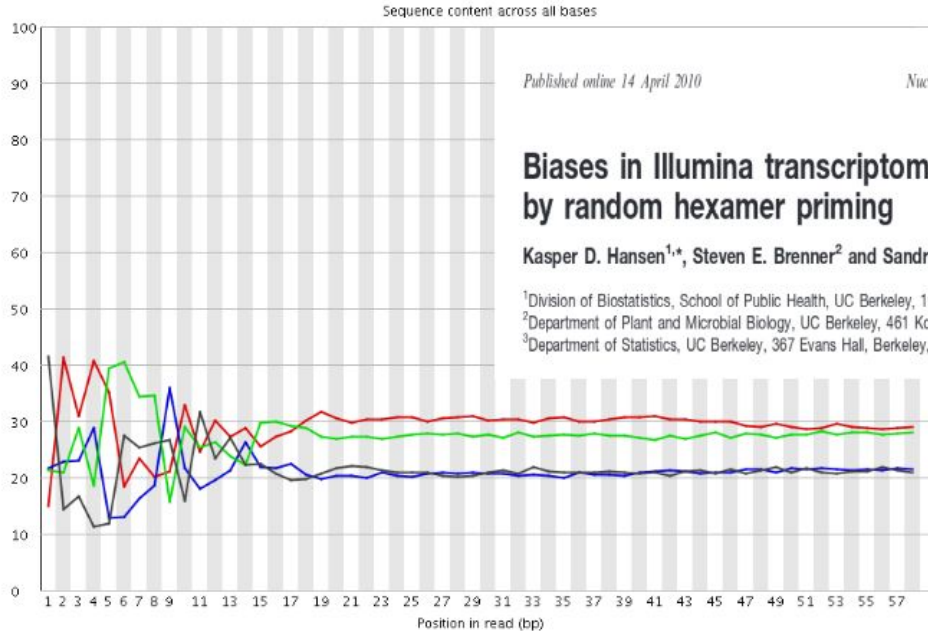
Illumina HISEQ2000

✗ The per base sequence quality are very low towards the end

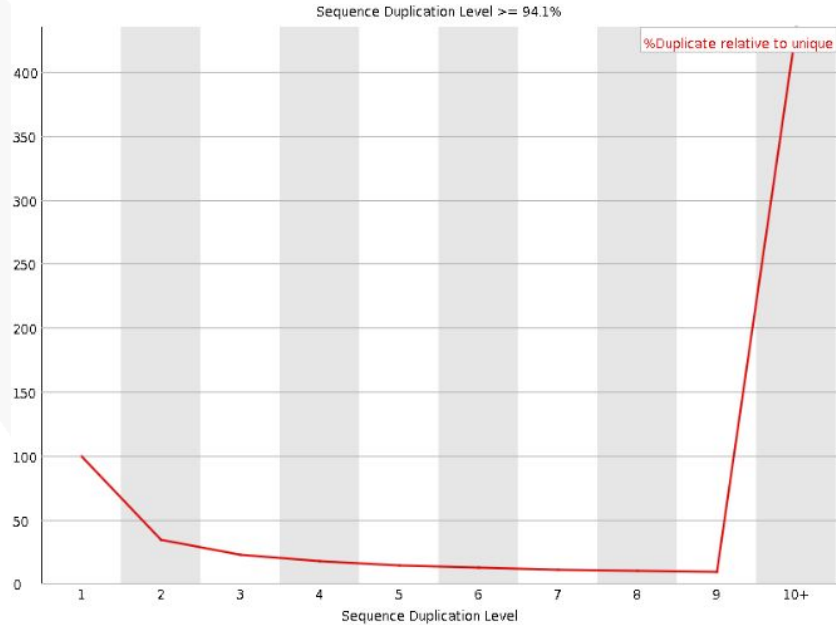


FASTQC: explore quality scores

❌ Per base sequence content



FASTQC: explore quality scores

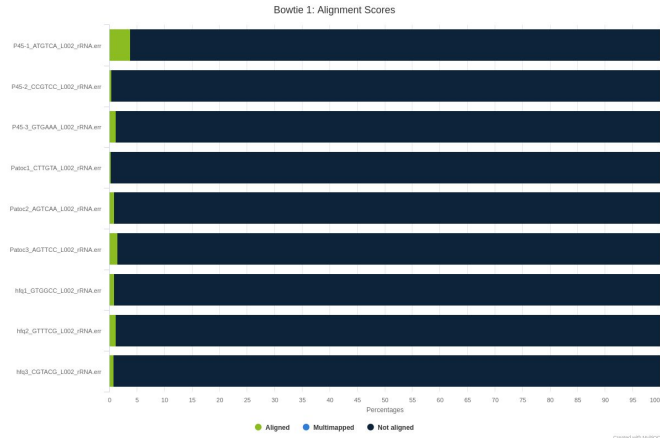
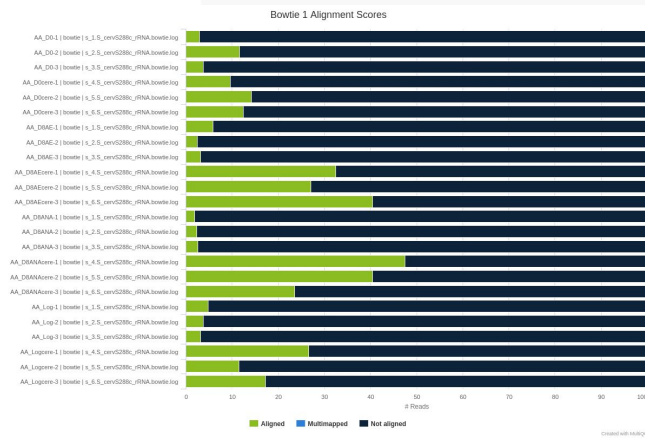


Systematic high duplication level in RNA-seq, why ?

How to screen contaminations ?

Different levels:

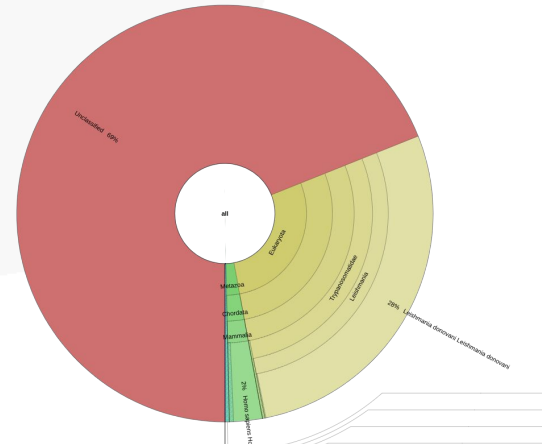
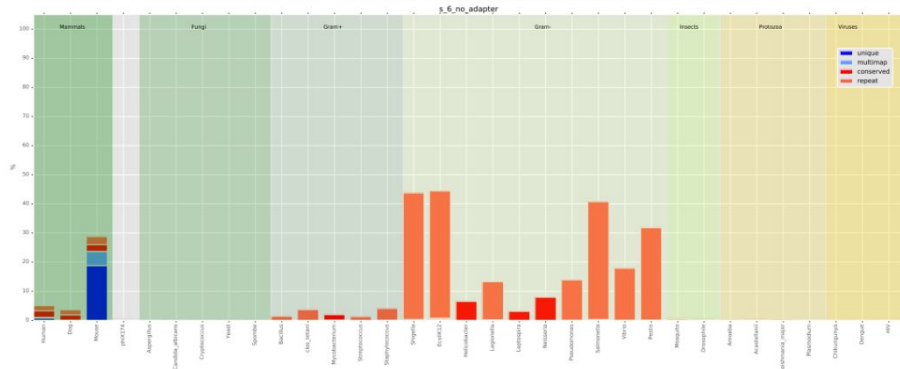
- Ribosomal contamination from same organism
 - Align reads against the ribosomal genome with a dedicated mapper



How to screen contaminations ?

Different levels:

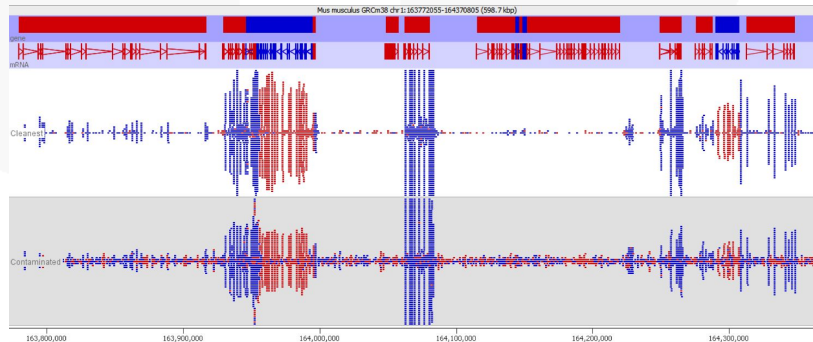
- Ribosomal contamination from same organism
- RNA contamination from other organism
 - Use dedicated or derived tools such as fastq_screen or kraken



How to screen contaminations ?

Different levels:

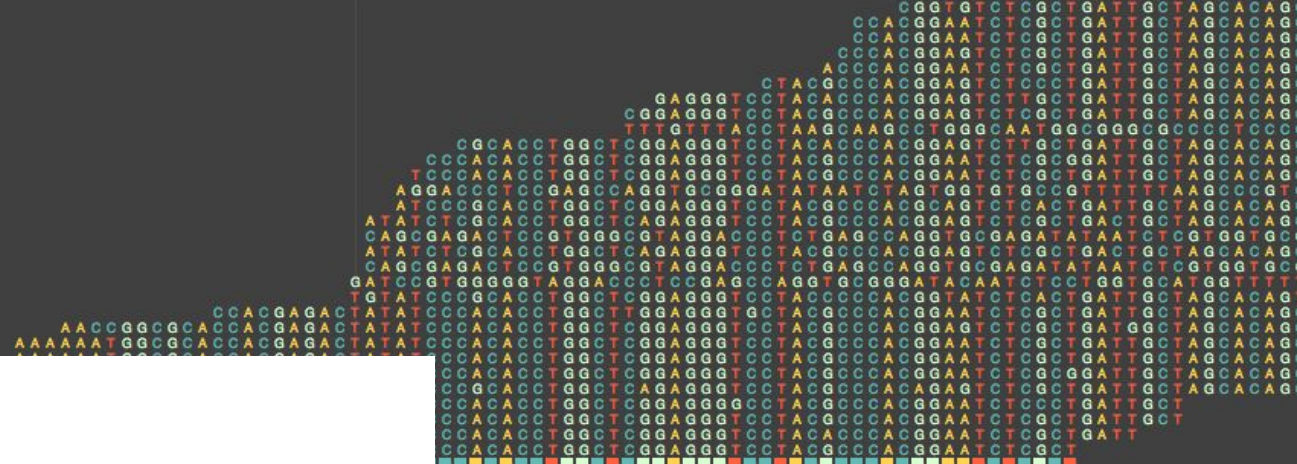
- Ribosomal contamination from same organism
- RNA contamination from other organism
- DNA contamination
 - DNase treatment could be ineffective and for DNA to make it through into the final library.
As soon as you visualise your reads against an annotated genome the presence of DNA is normally fairly apparent as a consistent background of reads over the whole genome



Chr: 20

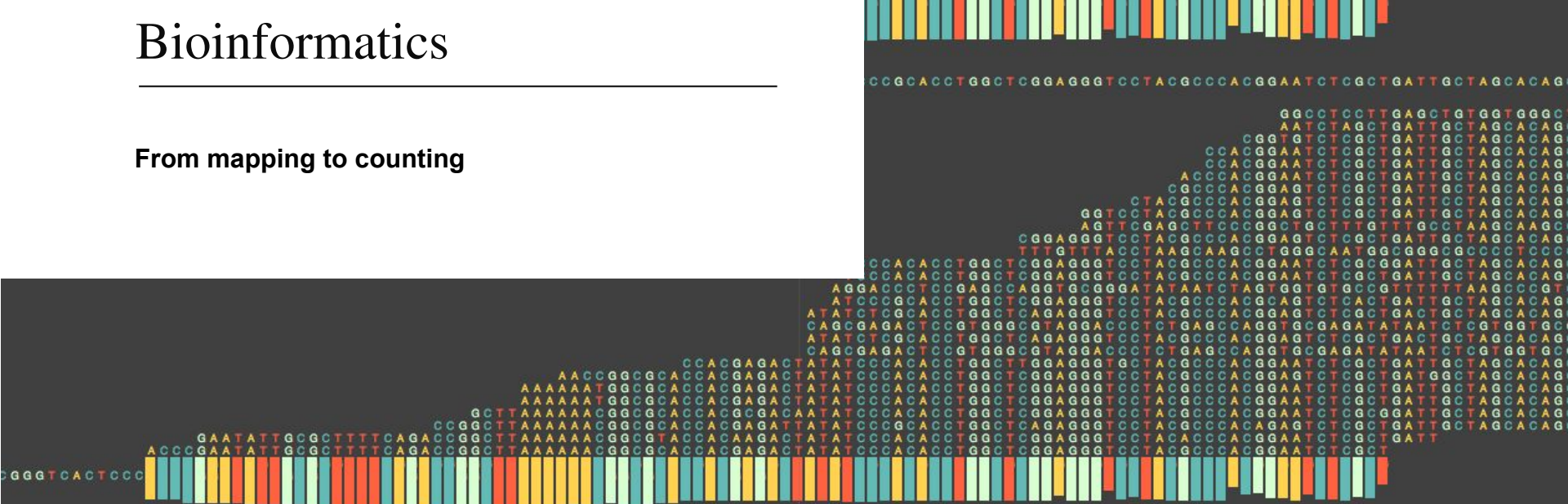
Position: 110939

T | T GR:37
REF ALT



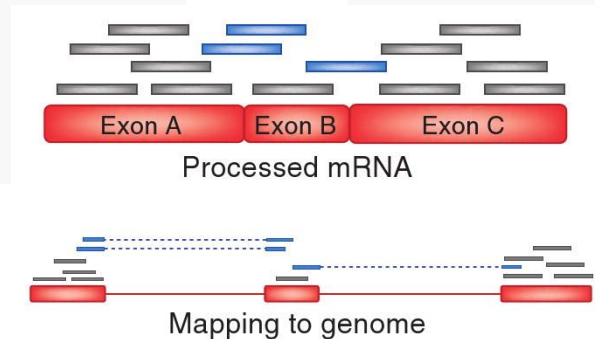
Bioinformatics

From mapping to counting



RNA-seq mapping specificity

- ★ Mapping on genome or transcriptome ?
 - the transcriptome is currently not well characterised enough to serve as a suitable reference for RNA-Seq
 - mapping to a genome is more objective and repeatable
 - get more gene isoforms information through mapping it to the genome
- ★ Take account to reads that come from exon-exon junctions



Cole Trapnell & Steven L Salzberg. Nature
Biotechnology 27, 455 - 457 (2009)

Mapping timeline



From https://www.ebi.ac.uk/~nf/hts_mappers/

Choose the good mapper

Which one is the best mapper ?



Choose the good mapper

Which one is ~~X~~ the best mapper ?

Which mapper should I use based on my data and my analysis ?

Choose the good mapper

Depends on:

- Detection of splicing events
- Length of reads:
 - Very short read (<50) :
 - Up to 1000kb :
 - Long reads :
- Allow gap on alignment

STAR, minimap2, Hisat2

Bowtie1

BWA-SW, bowtie2

Minimap2

STAR, BWA, Bowtie2

Common situations: choose a mapper widely-used and well maintained

Known biases in RNA-seq



Intron coverage: if many reads align to introns, this is indicative of incomplete poly(A) enrichment or abundant presence of immature transcripts.

Intergenic reads: if a significant portion of reads is aligned outside of annotated gene sequences, this may suggest genomic DNA contamination (or abundant non-coding transcripts).

3' bias: over-representation of 3' portions of transcripts indicates RNA degradation.

Mapping QC on RNA-seq

- ★ Percentage of mapped reads along genome
 - Human/Mouse: 70 to 90 %
 - Prokaryotic: more to 90 %
- ★ Uniformity of read coverage on exons and the mapped strand.
- ★ Low rate of multiple mapping
- ★ Low rate of ribosomal RNA



Mapping QC on RNA-seq

- ★ Common :
 - Samtools (flagstats)
 - Bamtools (stats)
 - Picardtools (CollectRNASeqMetrics)
 - RseQC
- ★ Human and mouse :
 - RNAseQC
 - Qualimap



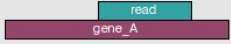



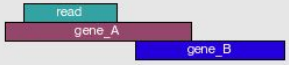
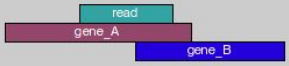

Quantify number of reads on each gene

When counting reads, make sure you know how the program handles the following:

- ❑ overlap size (full read vs. partial overlap)
- ❑ multimapping reads
- ❑ reads overlapping multiple genomic features of the same kind
- ❑ reads overlapping introns

Two popular tools :

- Htseq-count
- featureCounts

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

RNA-seq experiment

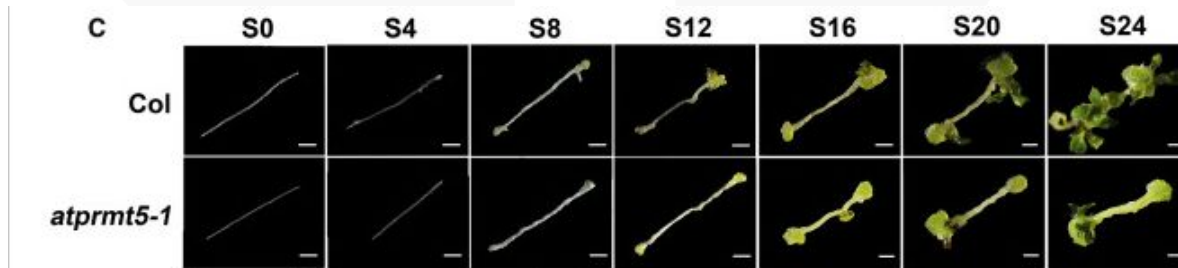
Organism: *Arabidopsis thaliana*, plant and model organism.

Genome and annotation available in TAIR10, the arabidopsis database



Dataset: 3 biological replicates, paired-end sequencing.

Characterization of the function of the protein arginine methyltransferase AtPRMT5 during de novo shoot regeneration in *Arabidopsis* by a knocking-out of AtPRMT5.



Practice

- **Connexion to cluster:**

```
ssh <LOGIN>@core.cluster.france-bioinformatique.fr
```

- **Change directory:**

```
cd /shared/projects/<PROJECT>
```

- **Create a new directory:**

```
mkdir TP_rnaseq
```

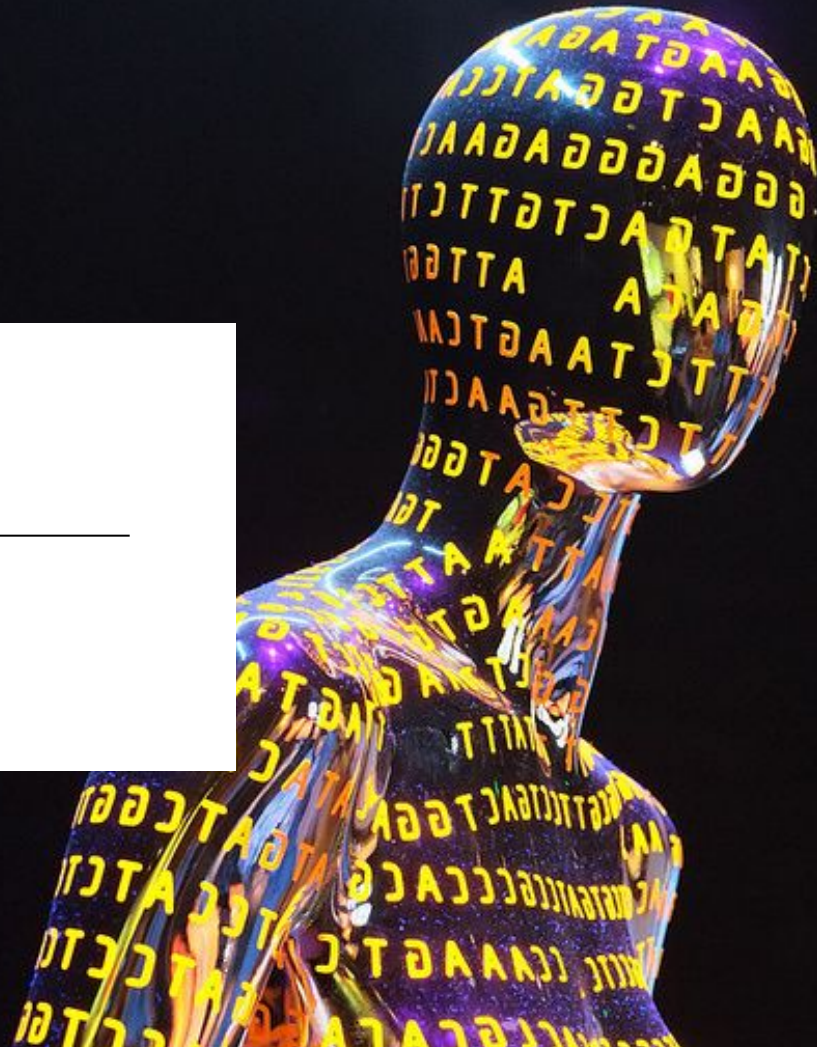
- **Copy the script template in your home:**

```
cp /shared/projects/ebaii2020/atelier_rnaseq/01-Bioinfo/runme.sh TP_rnaseq
```

- **Follow the commands on the runme**

Bioinformatics

Visualize your data



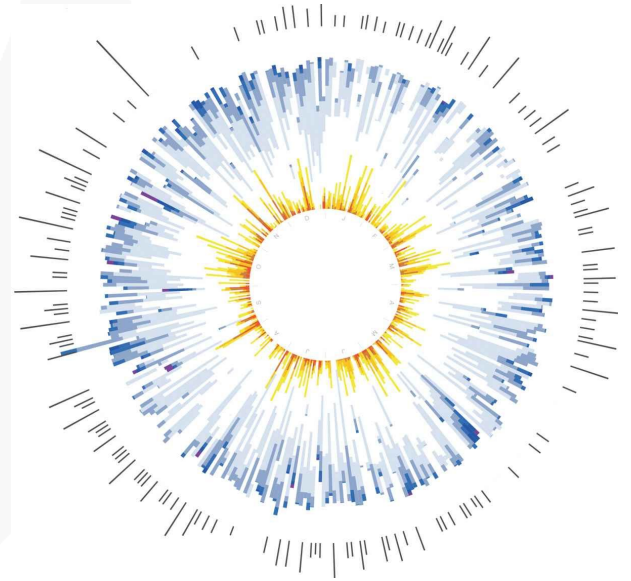
Visualize alignments

Which format ?

- ❖ BAM
- ❖ BigWig, BedGraph (base-by-base scores)
- ❖ BED, GFF (feature-by-feature data)

Which tools ?

- ❖ Browser : IGV, Artemis, UCSC Genome browser, SeqMonk...
- ❖ Snapshots : Deeptools, ngs.plot,...



Visualize alignments

Go to AT4G31120

