aviesan
alliance nationale
pour les sciences de la vie et de la santé

How to deal with your RNA-seq data ?

Pierre Pericard, Claire Tofano-Nioche, & the RNA-seq team

École de Bioinformatique AVIESAN-IFB-Inserm 2023

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

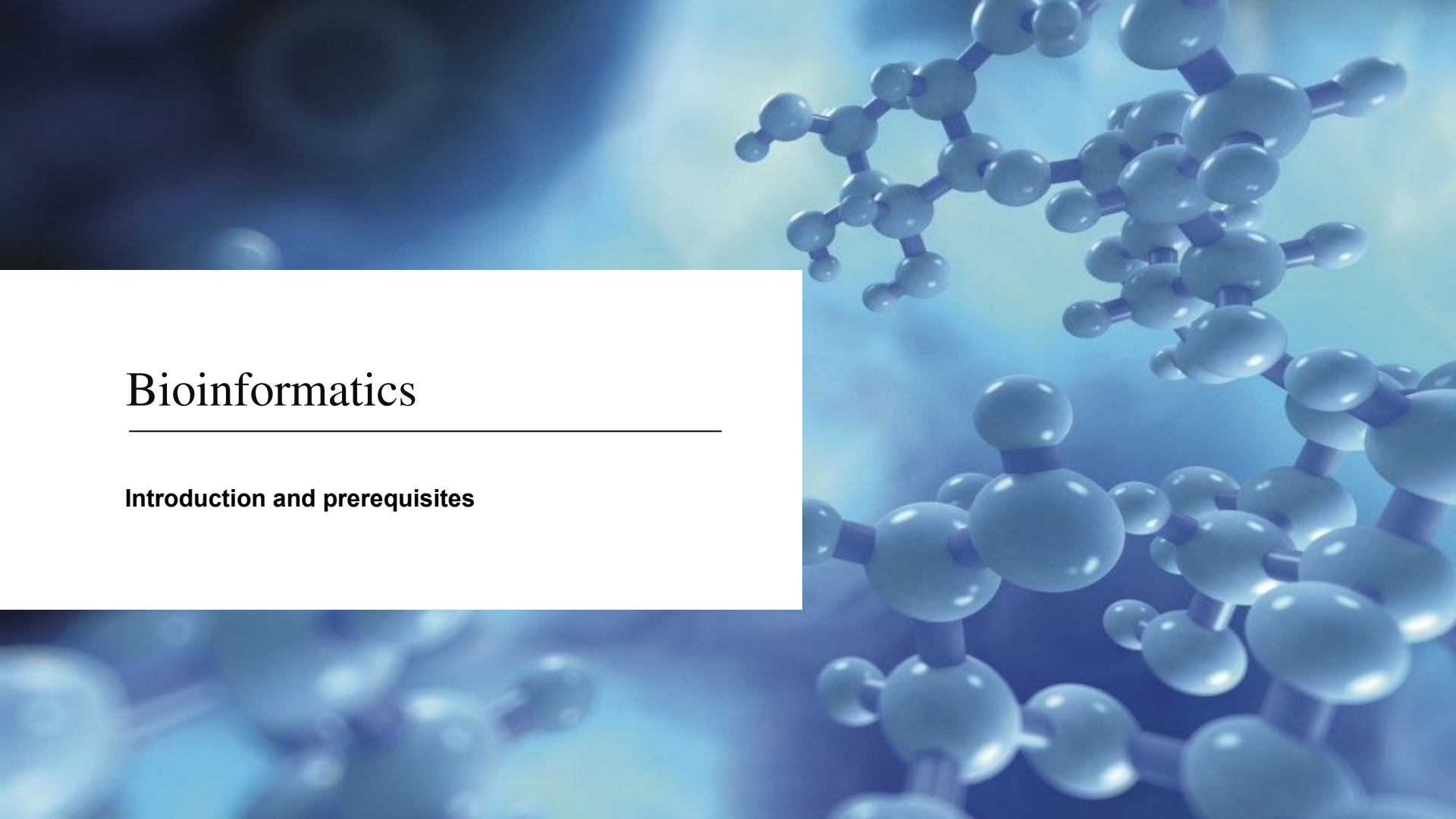
Advanced practice

Gene Sets Analysis methods

06

Bioinformatics

Transcriptome *de novo* assembly

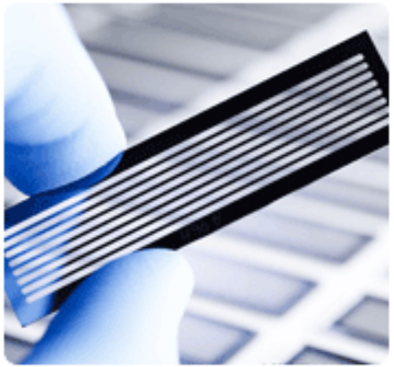


Bioinformatics

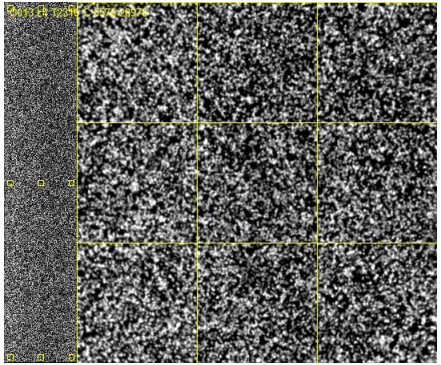
Introduction and prerequisites

Raw NGS data

Instrument



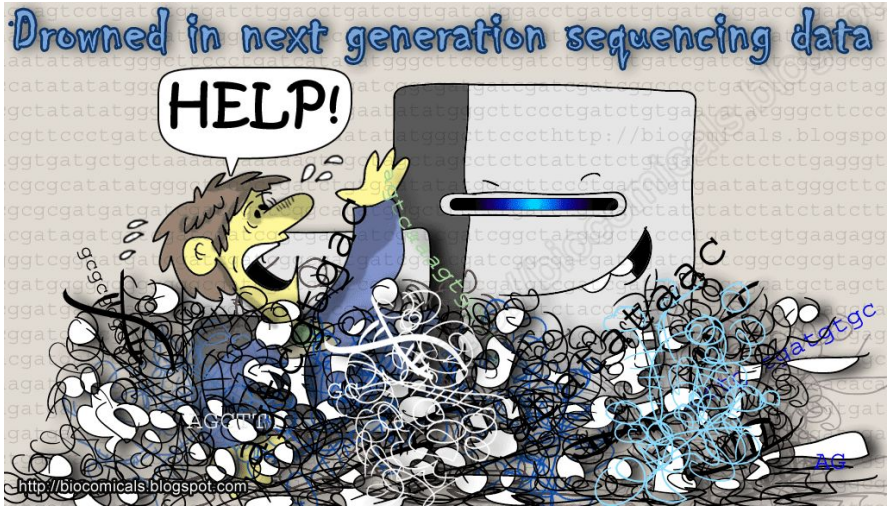
Flowcell



Intensities

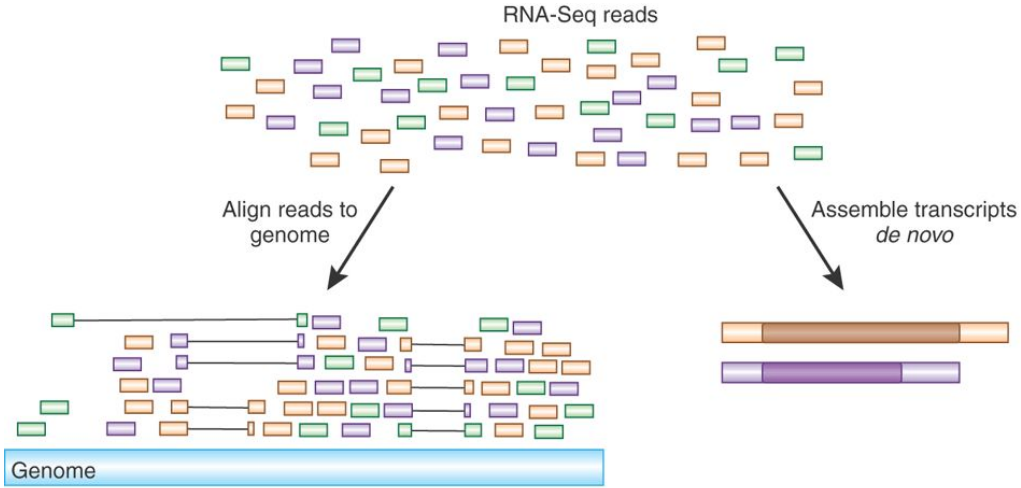
Data storage: NovaSeq6000

- Text file with size between 80Gb to 3Tb (in single flowcell mode)
- Let’s compare : War and peace by Léon Tolstoï
  - 1817 pages
  - 6 cm width
  - 4 Mb
- 1 run :
  - 750 000 times "war and peace”
  - 1350 Millions pages
  - 45 km (138 Eiffel towers)
- x2 for dual flowcell mode



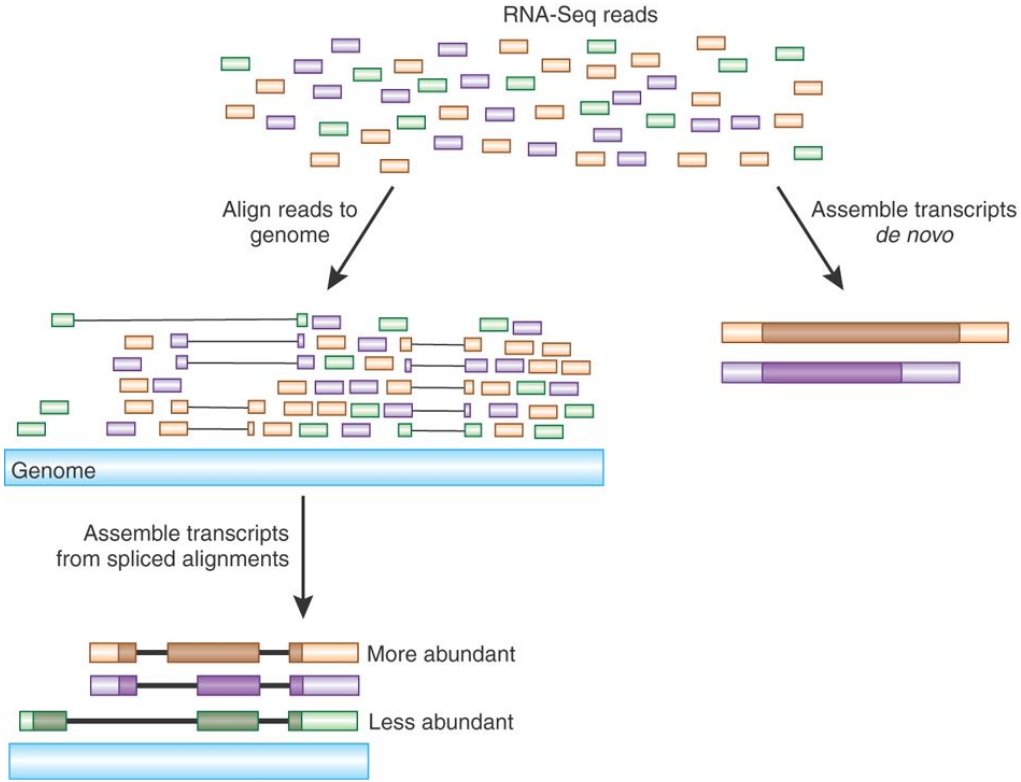
RNA-seq applications

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



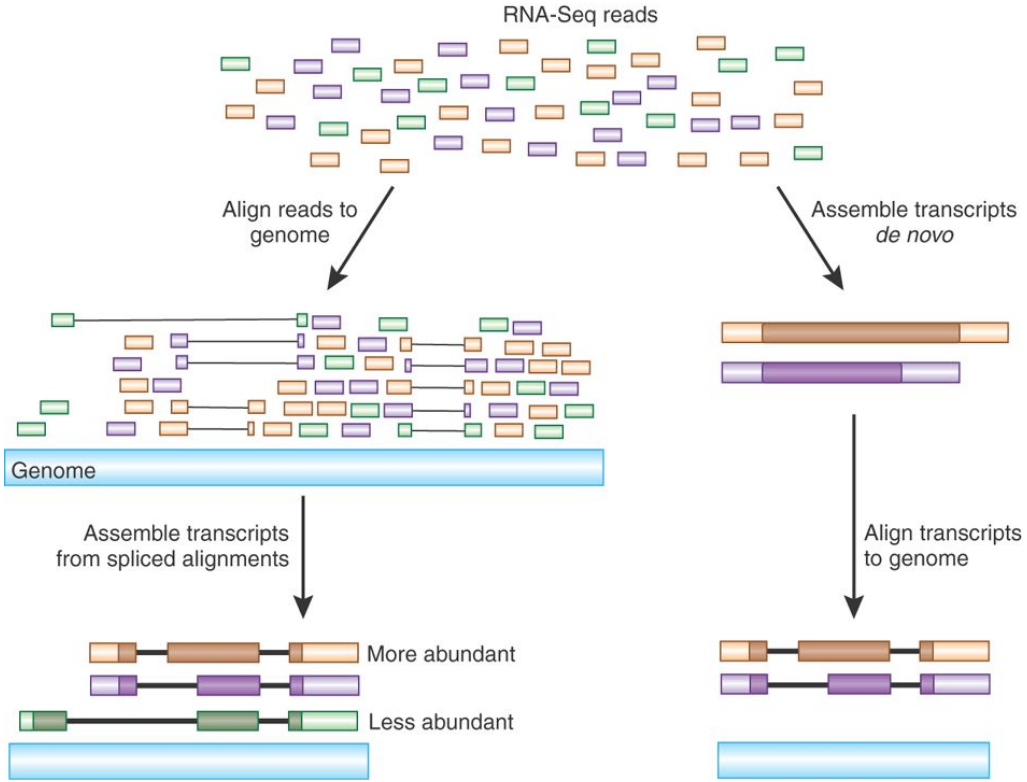
RNA-seq applications

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



RNA-seq applications

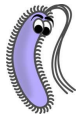
« Transcriptome analysis provides information about the identity and quantity of all RNA molecules in one cell or a population of cells »



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

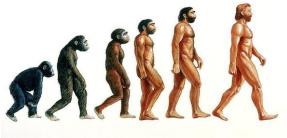
I want to identify antisense RNA:

➔ 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 interesting in small RNA profiling:

➔ Use specific protocol

I'm interesting 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

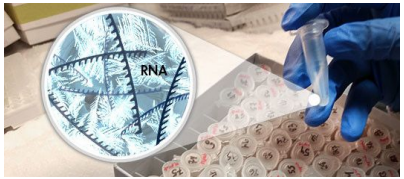
RNA-seq: Why ? How

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:

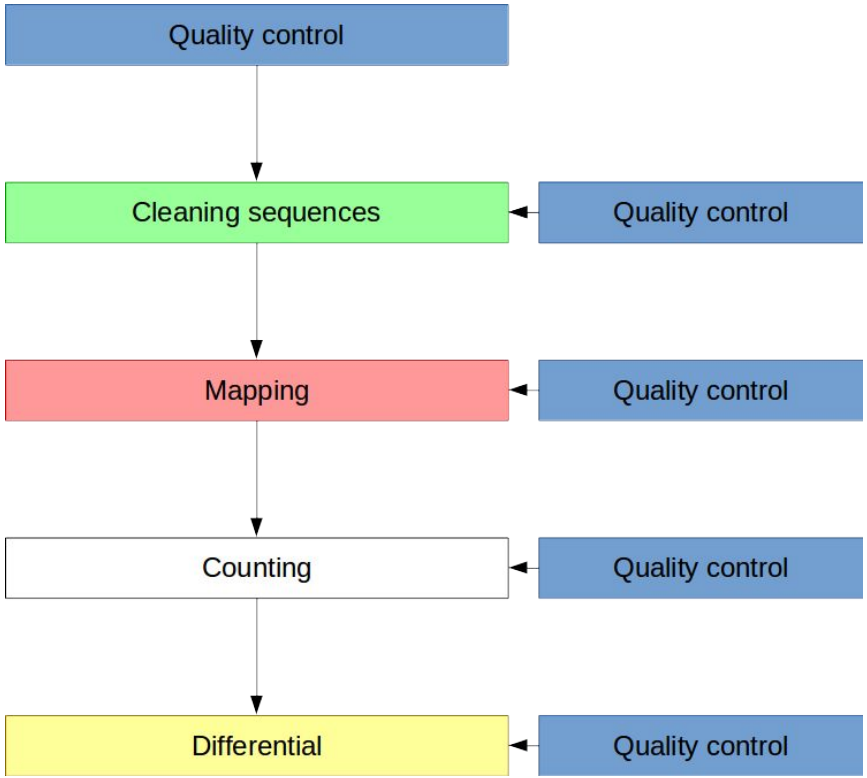
All features (genes, CDS, intron, UTR) of genome in **GTF**/GFF format

Where find the genome and the annotation ?

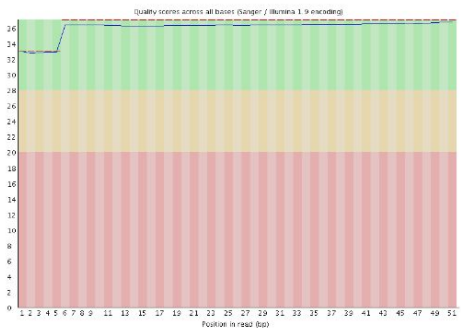
Common databases

Specific databases

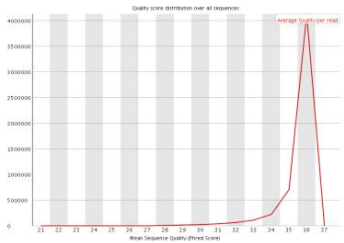
Keep control on your datas



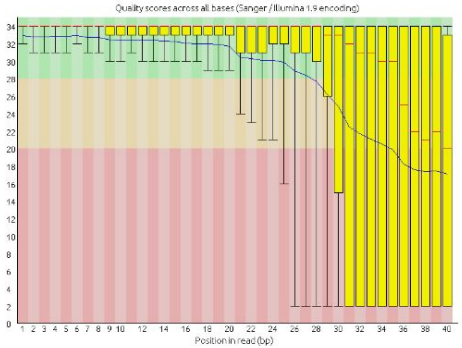
FASTQC: explore quality scores



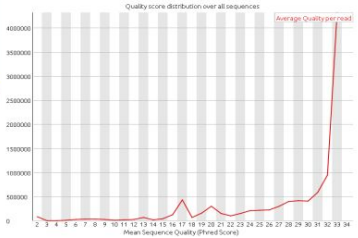
Illumina HISEQ2500



The per base sequence quality are very high along sequence

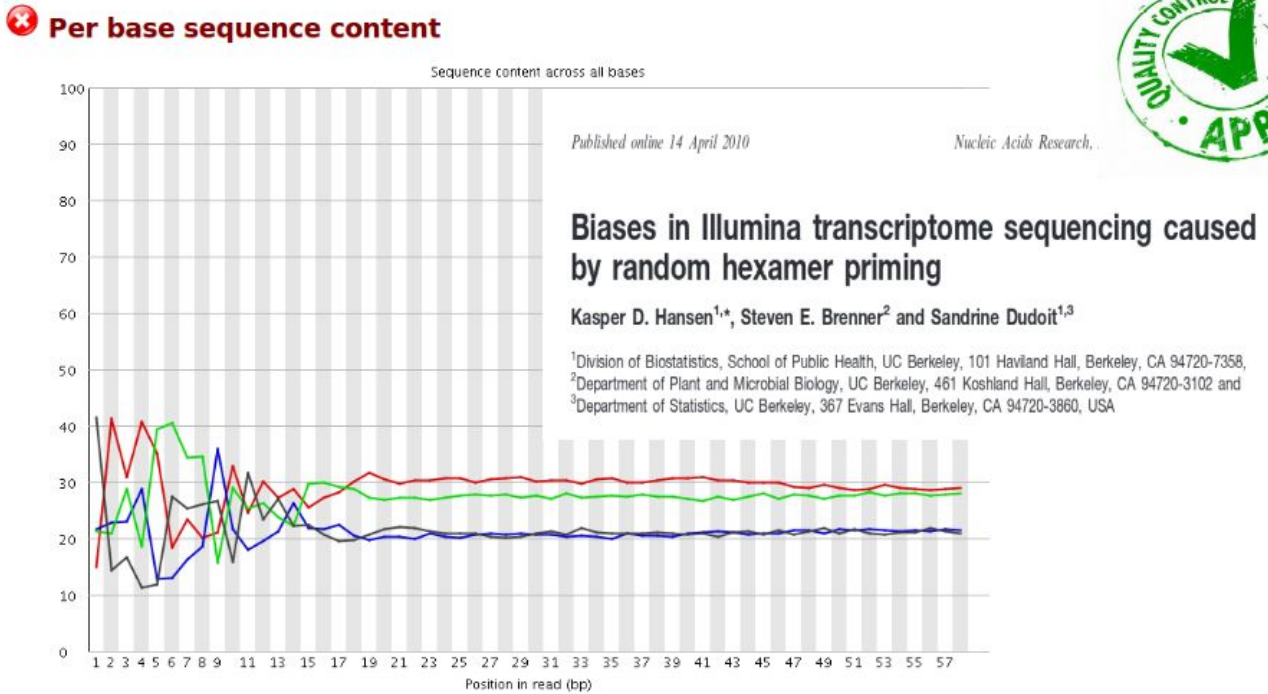


Illumina HISEQ2000

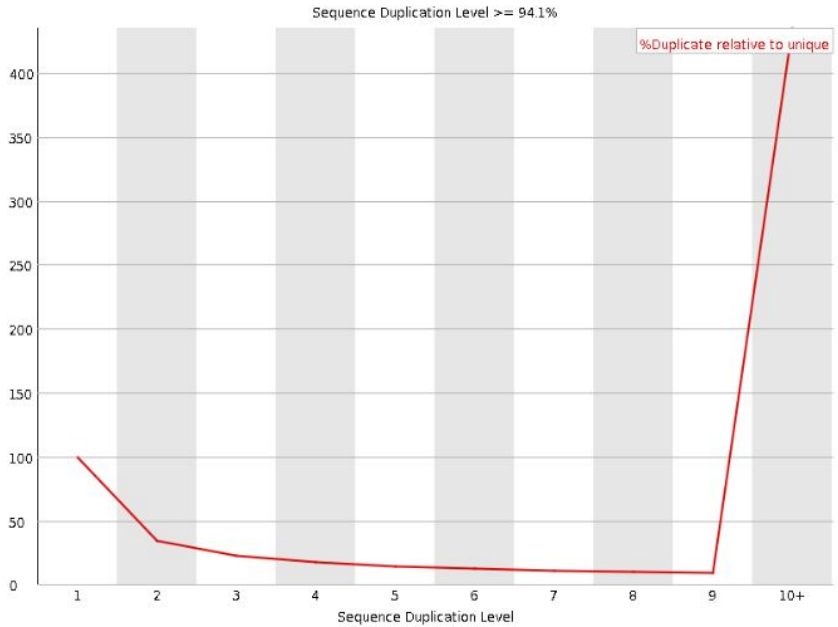


The per base sequence quality are very low towards the end

FASTQC: explore quality scores

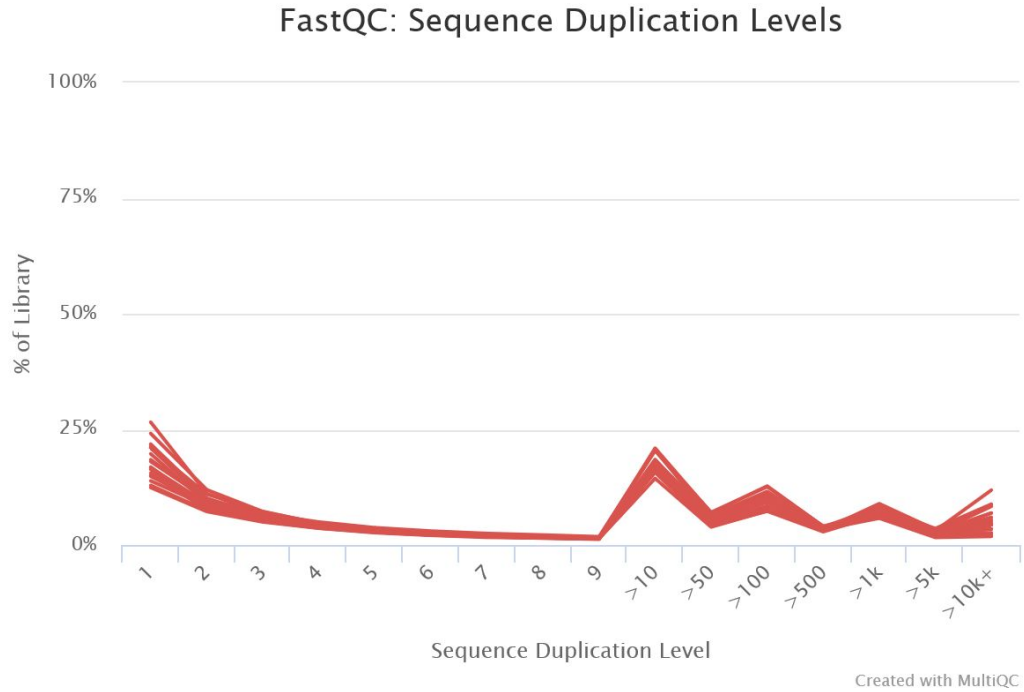


FASTQC: explore quality scores



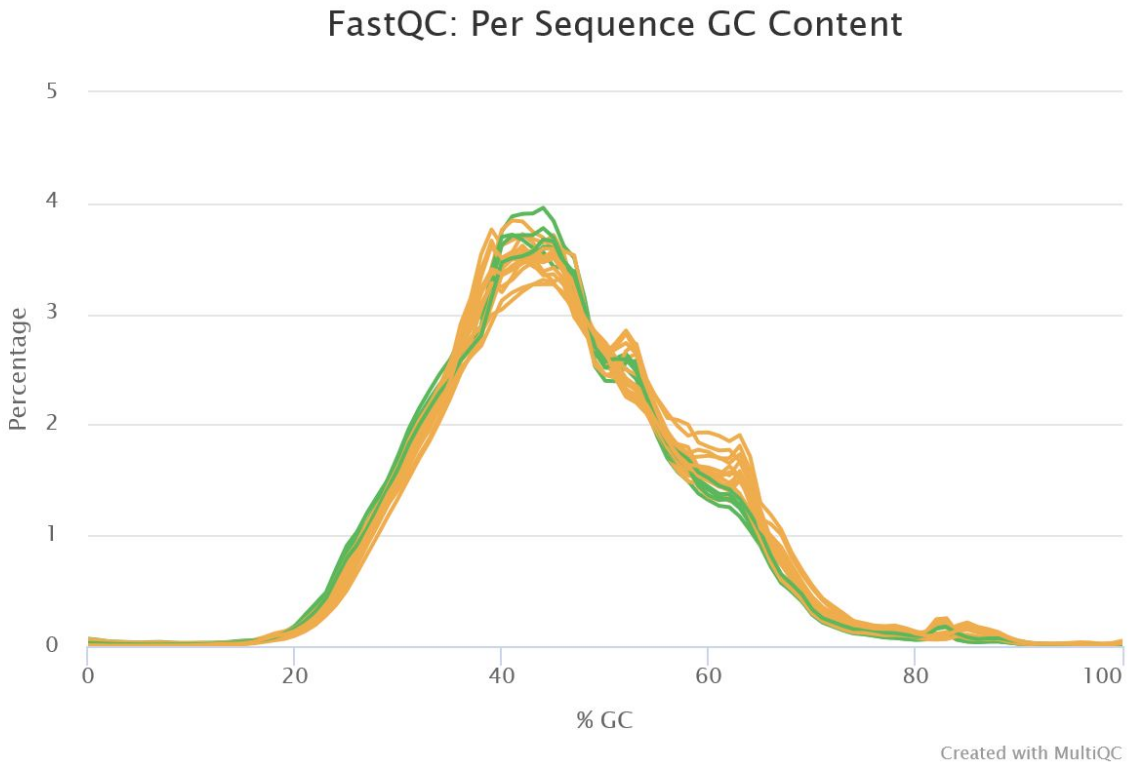
Systematic high duplication level in RNA-seq, why ?

FASTQC: explore quality scores



Systematic high duplication level in RNA-seq, why ?

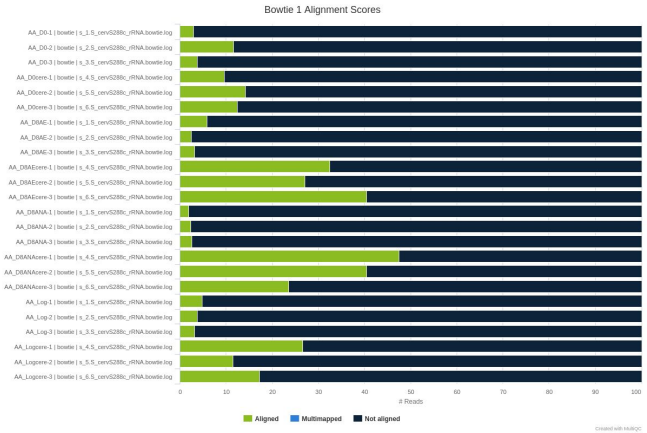
FASTQC: explore quality scores

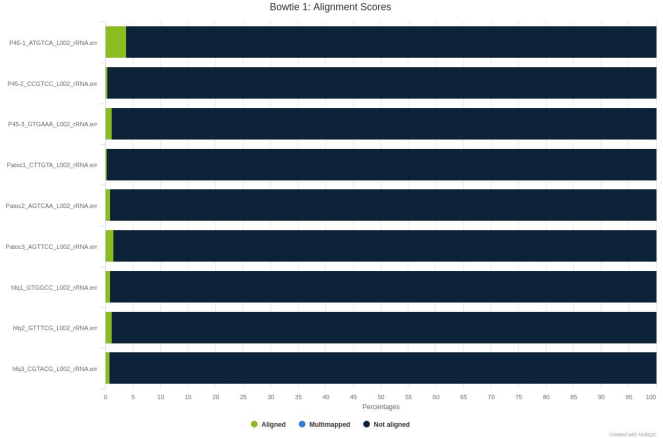


How to screen contaminations ?

Different levels:

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

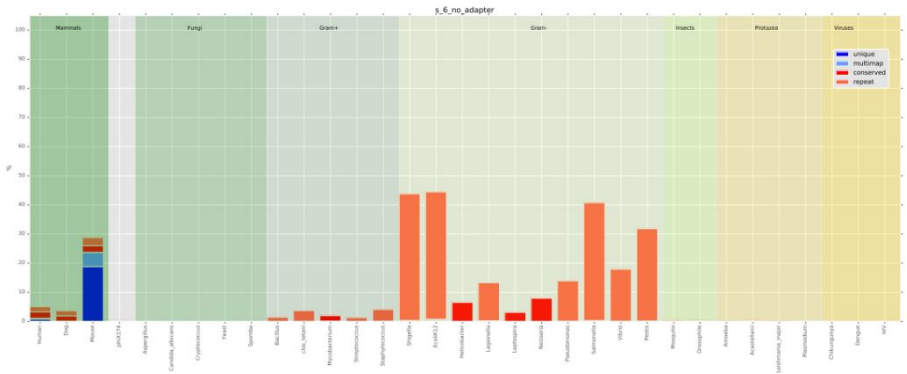


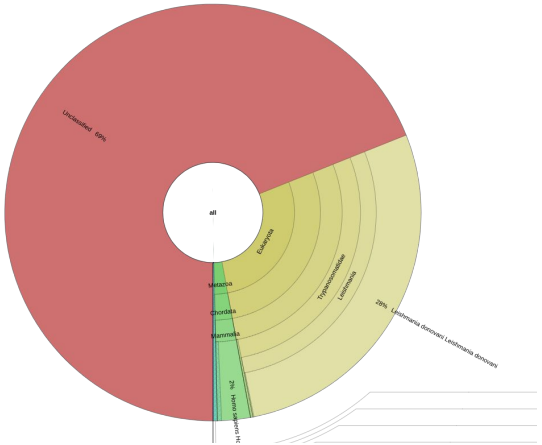


How to screen contaminations ?

Different levels:

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

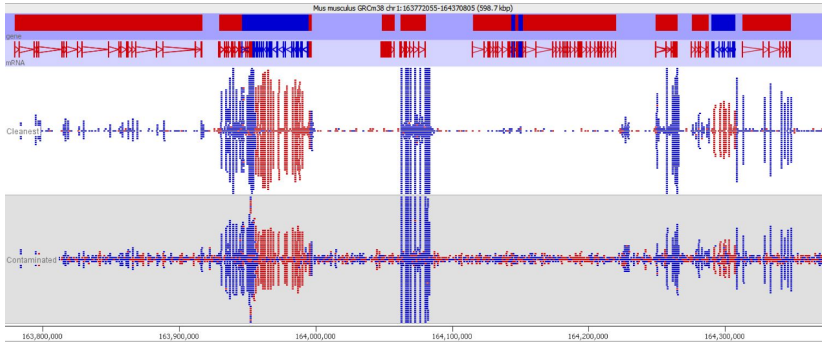


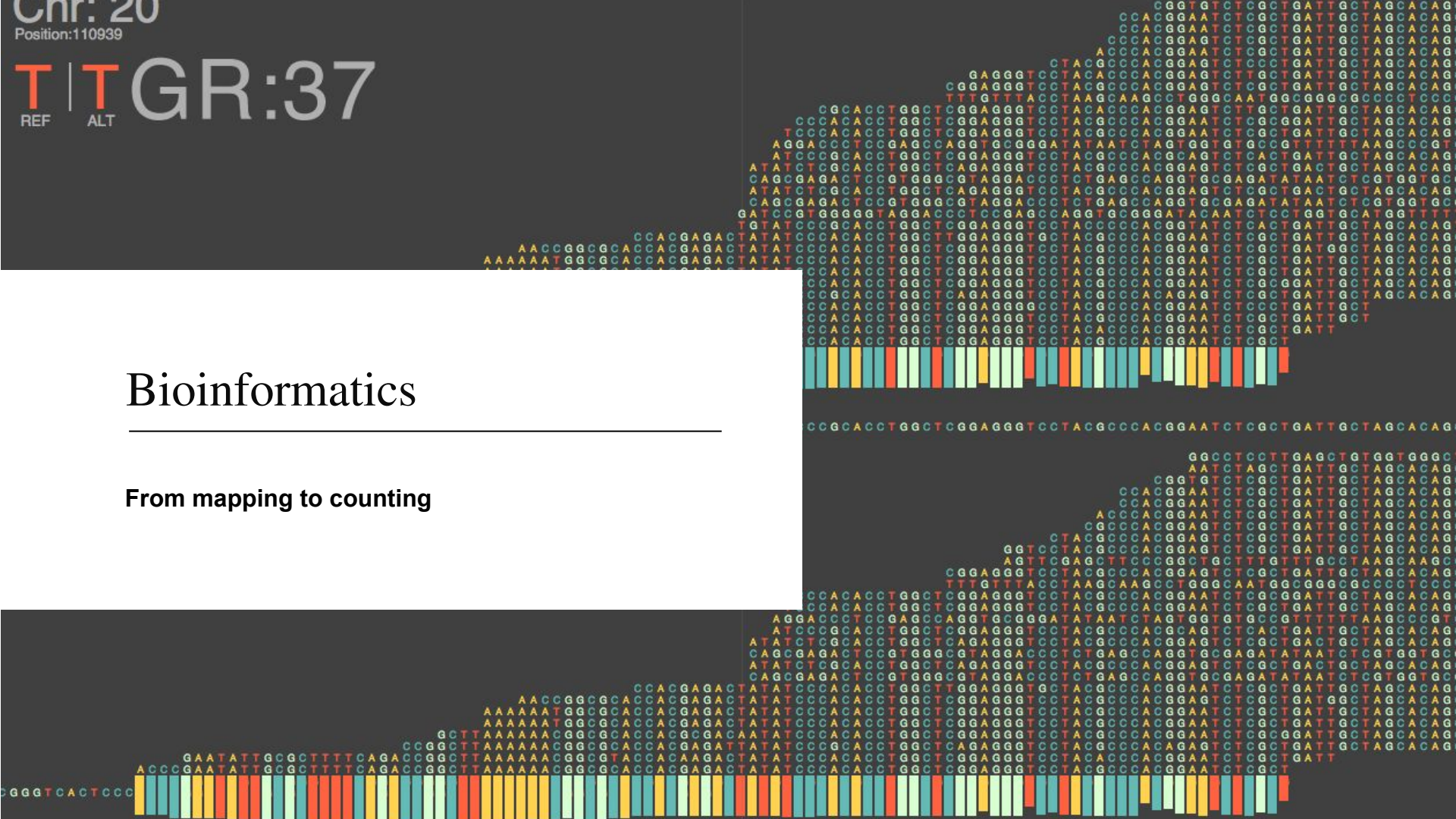


How to screen contaminations ?

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





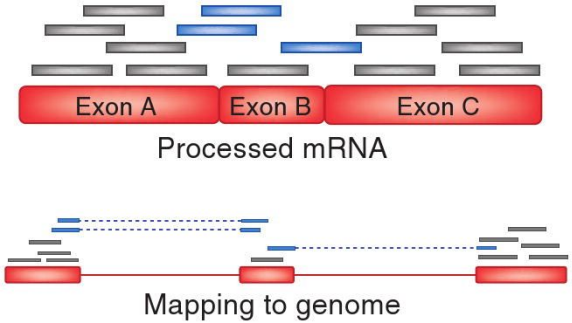
Bioinformatics

From mapping to counting

RNA-seq mapping specificity

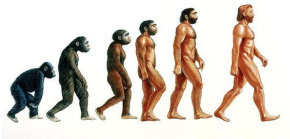
- ★ Mapping on genome or transcriptome ?
  - the transcriptome is currently not well characterised enough to serve as a suitable reference for RNA-Seq
  - get more gene isoforms information through mapping it to the genome

- ★ Take account to reads that come from exon-exon junctions

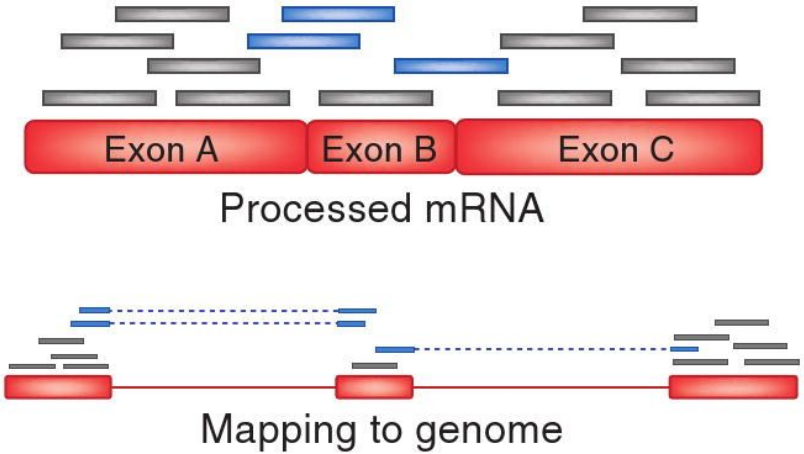


Cole Trapnell & Steven L Salzberg.Nature Biotechnology 27, 455 - 457 (2009)

RNA-seq mapping specificity

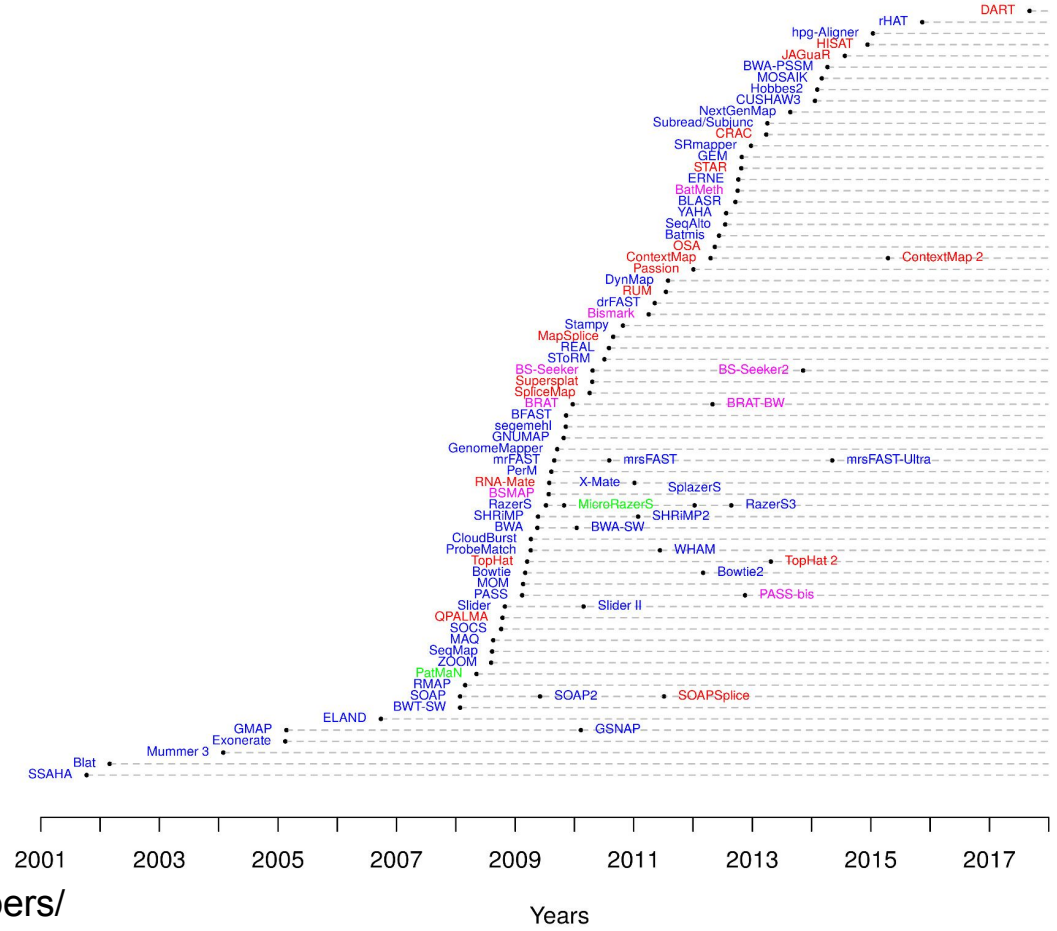


★ Take account to reads that come from exon-exon junctions



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

Mapping timeline



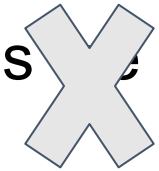
From https://www.ebi.ac.uk/~nf/hts_mappers/

Choose the good mapper

Which one is the best mapper ?

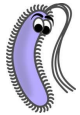
Choose the good mapper

Which one is the best mapper ?



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

Choose the good mapper

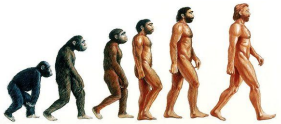
Prokaryotes 

For short reads in 99% of the use-cases:

➔ Bowtie2, BWA

For dual-RNAseq (pathogens + host):

➔ see Eukaryotes

Eukaryotes 

Need a tool able to detect splicing events !

For short reads in 99% of the use-cases:

➔ STAR, Hisat2

For long reads:

➔ Minimap 2

For very small RNA (e.g. miRNA-seq):

➔ BWA, Bowtie

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

Many people uses TopHat2

(> 10,623 citations in Scholar, > 1,000 citations in 2021 only)

but don't

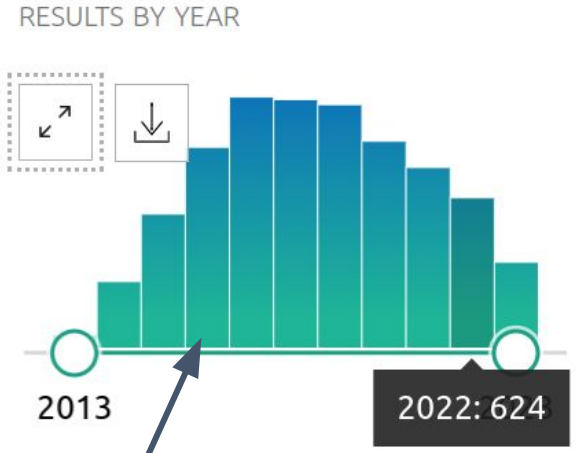
On TopHat2 website (since Feb 2016)

TopHat2 « *is now largely superseded by HISAT2 which provides the same core functionality (i.e. spliced alignment of RNA-Seq reads), in a more accurate and* ***much more efficient*** *way* » .

from Mikaël Salson

Many people uses TopHat2

(> 10,623 citations in Scholar, > 1,000 citations in 2021 only)



but don't

On TopHat2 website (since Feb 2016)

TopHat2 « *is now largely superseded by HISAT2 which provides the same core functionality (i.e. spliced alignment of RNA-Seq reads), in a more accurate and* ***much more efficient*** *way* » .

from Mikaël Salson

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

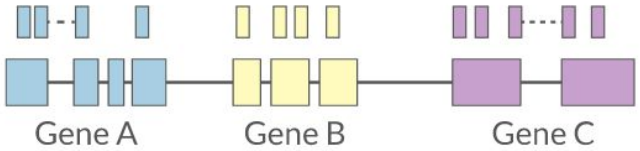
- ★ Percentage of mapped reads along genome
  - Human/Mouse: 70 to 90 %
  - Prokaryotic: more to 90 %
- ★ Uniformity of read coverage on exons and the mapped strand.
- ★ Low rate of multiple mapping
- ★ Low rate of ribosomal RNA

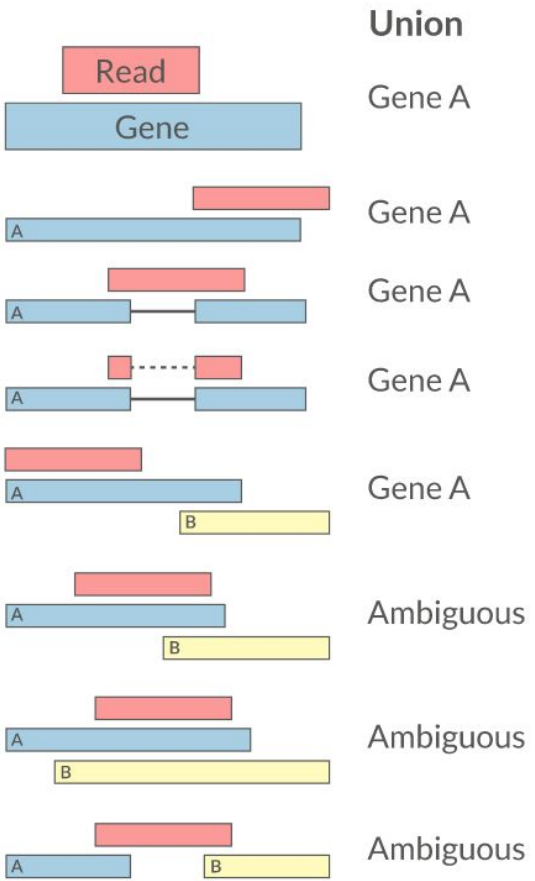
Mapping QC on RNA-seq

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

Quantification / Count

- Read counts ~ gene expression
- Reads can be quantified on any feature (gene, transcript, exon etc)
- Intersection on gene models
- Gene/Transcript level

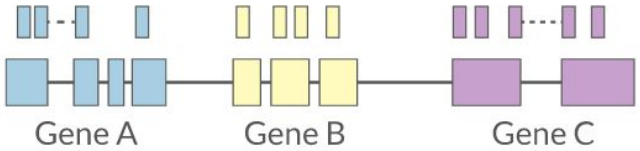




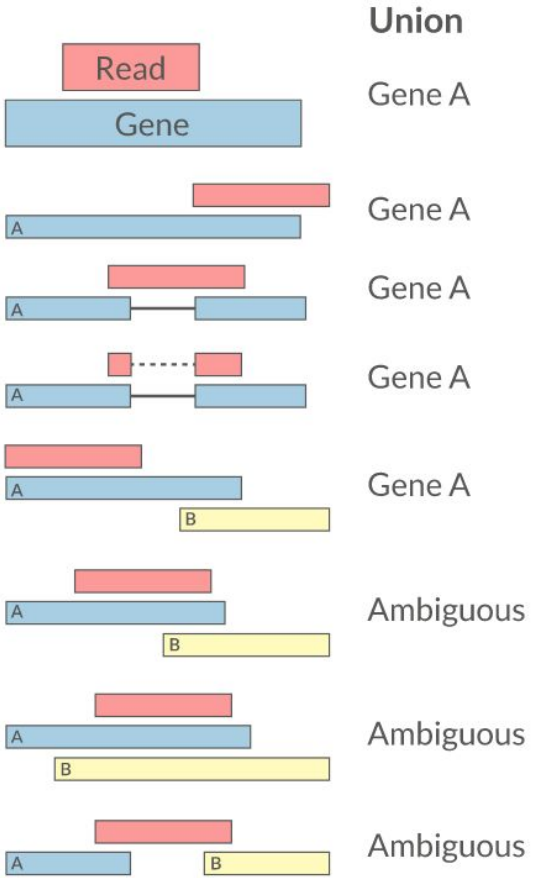
credit: SciLife lab https://scilifelab.github.io/courses/ngsintro/1905/slides/rnaseq/presentation.html#33

Quantification / Count

- Read counts ~ gene expression
- Reads can be quantified on any feature (gene, transcript, exon etc)
- Intersection on gene models
- Gene/Transcript level



ambiguous or **multi-mapping** reads



credit: SciLife lab https://scilifelab.github.io/courses/ngsintro/1905/slides/rnaseq/presentation.html#33

Counting gene expression from alignments

A

B

Approach to handle multireads	Read distribution representation	Counts
Ignore		G1: 10 reads G2: 6 reads
Count once per alignment		G1: 18 reads G2: 14 reads
Split them equally		G1: 14 reads G2: 10 reads
Rescue based on uniquely mapped reads		G1: 15 reads G2: 9 reads
Expectation-maximization	(n)	G1: 15 reads G2: 9 reads
Read coverage based methods		G1: 15 reads G2: 9 reads
Cluster methods		G1:10 reads G2:6 reads Cluster G1/G2: 8 reads

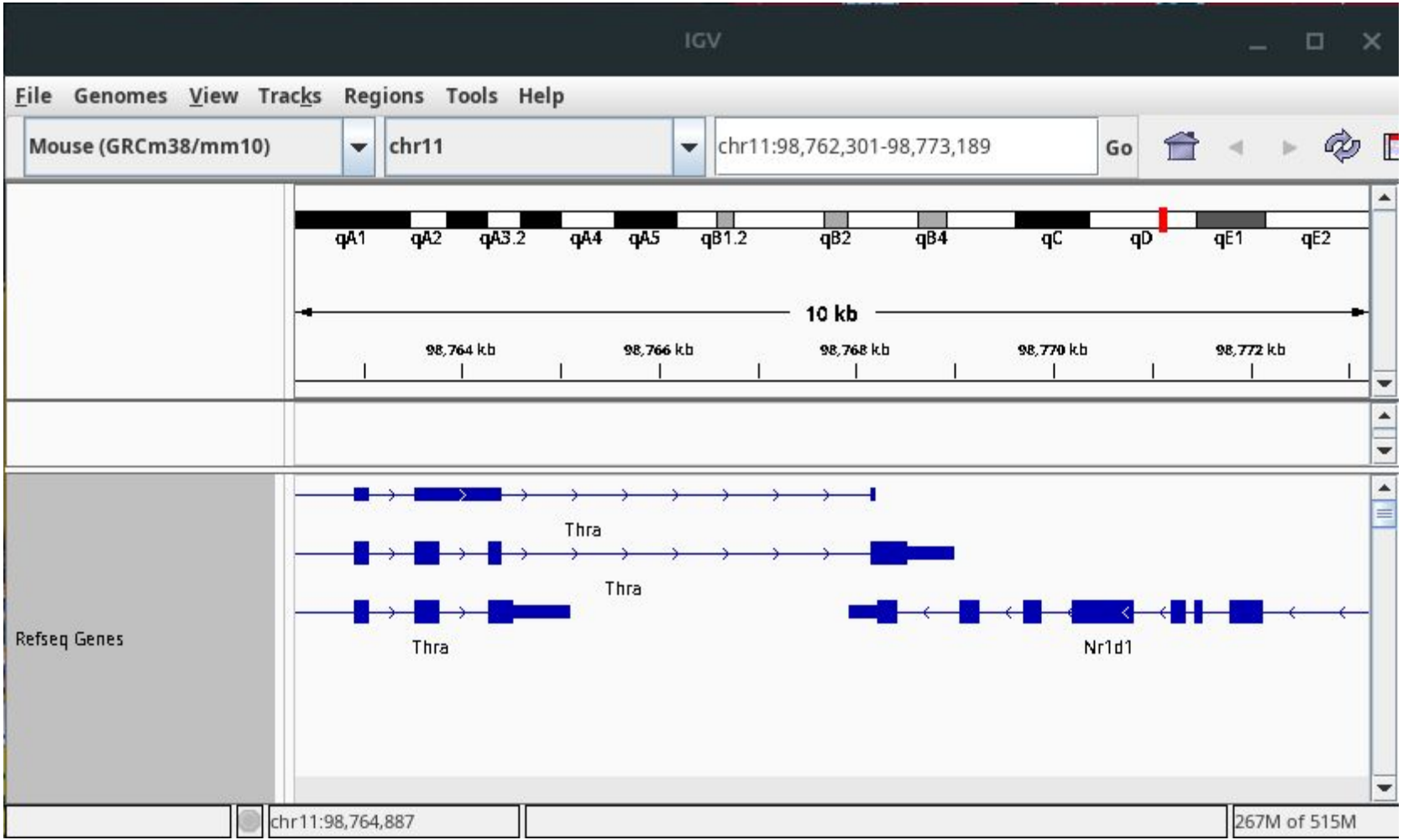
EM: counts are not bound to be integers!

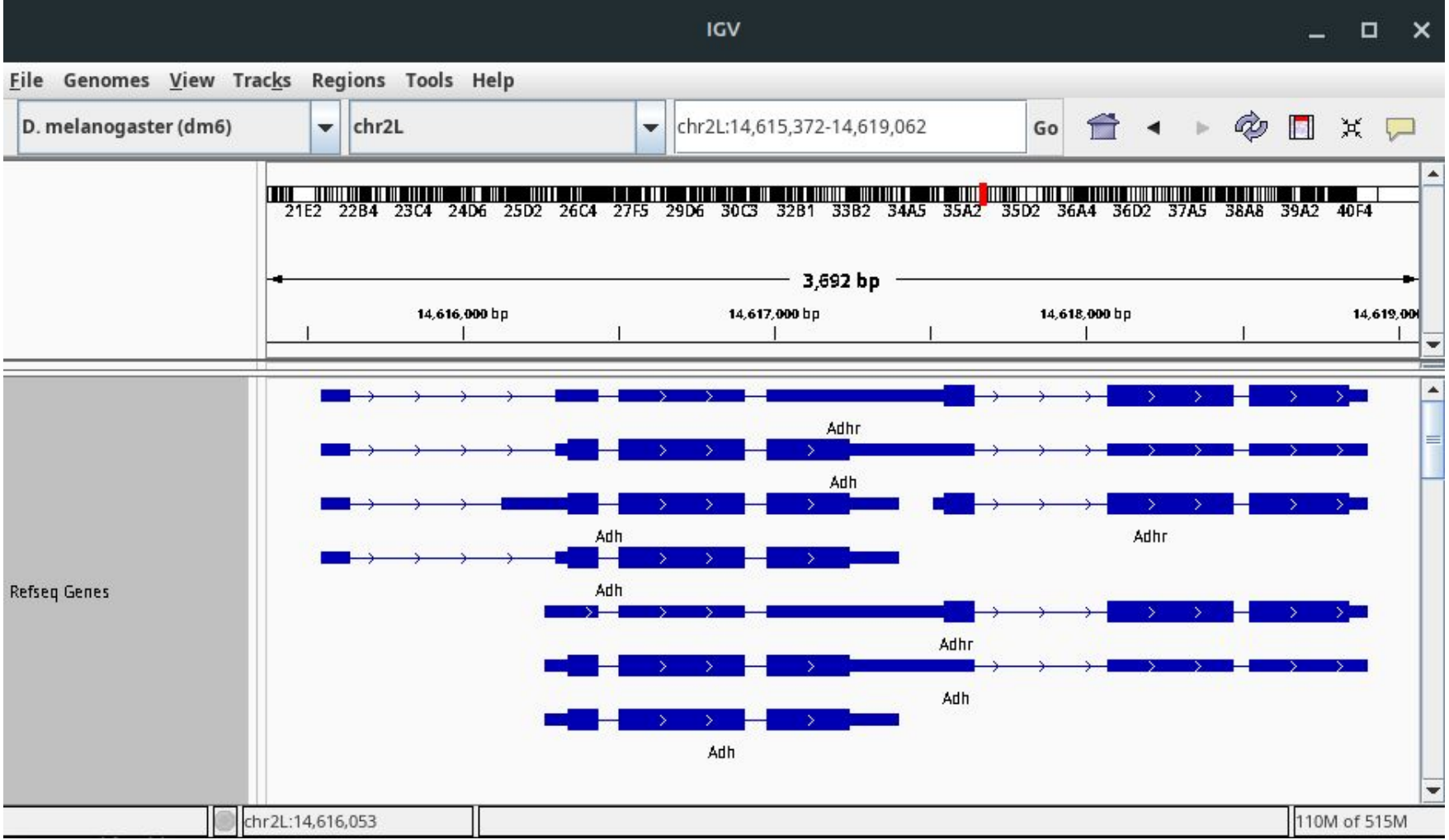
Deschamps-Francoeur, et al. 2020. doi:10.1016/j.csbj.2020.06.014

Table 1
Computational strategies and methods that handle multi-mapped reads.

Tool	Quantification level	Input	Strandedness can be specified	Count type	Strategy	Paired end	Confidence level	Focus
HTSeq-count	Gene	BAM	Y	Counts	Ignore	Y	N	Long RNA
STAR geneCounts	Gene	Fastq	Y	Counts	Ignore	Y	N	Long RNA
Cufflinks	Transcript	BAM	Y	RPKM	Split equally, Rescue	Y	N	Long RNA
featureCounts	Gene	BAM	Y	Counts	Ignore, count all, split equally	Y	N	Long RNA
CoCo	Gene	BAM	Y	Counts, CPM, TPM	Rescue	Y	N	Small RNA Long RNA
ERANGE	Transcript	BAM	N	RPKM	Rescue	Y	N	Long RNA
EMASE	Transcript	BAM	N	Counts, TPM	EM	Y	N	Long RNA
IsoEM2	Both	SAM	Y	FPKM, TPM	EM	Y	Confidence intervals	Long RNA
Kallisto	Transcript	Fastq	Y	TPM	EM	Y	Bootstrap values	Long RNA
RSEM	Both	Fastq, BAM	Y	Counts, TPM, FPKM	EM	Y	95% credibility intervals	Long RNA
Salmon	Transcript	Fastq	Y	Counts, TPM	EM	Y	Bootstrap values	Long RNA
MMR	N/A	BAM	Y	N/A	Read coverage	Y	N/A	Long RNA
MuMRescueLite	Genomic loci	Custom format	N	Counts	Read coverage	N	N	Short sequence tags
Rcount	Gene	BAM	Y	Counts	Read coverage	N	N	Long RNA
ShortStack	Gene	Fastq, BAM	N	Counts, RPM	Read coverage	N	N	Small RNA
mmquant	Gene	BAM	Y	Counts	Gene Clustering	Y	N	Small RNA Long RNA
SeqCluster	Gene	BAM	N	Counts	Gene clustering	N	N	Small RNA
Fuzzy method	Gene	Custom format	N	Fuzzy counts	Fuzzy sets	N	Fuzzy counts	Small RNA Long RNA
geneQC	Gene	SAM	Y	NA	ML	Y	Mapping uncertainty level	Small RNA Long RNA

Deschamps-Francoeur, et al. 2020. doi:10.1016/j.csbj.2020.06.014


IGV
File Genomes View Tracks Regions Tools Help
Mouse (GRCm38/mm10)
chr11
chr11:98,762,301-98,773,189
Go
qA1 qA2 qA3.2 qA4 qA5 qB1.2 qB2 qB4 qC qD qE1 qE2
10 kb
98,764 kb
98,766 kb
98,768 kb
98,770 kb
98,772 kb
Refseq Genes
Thra
Thra
Thra
Nr1d1
chr11:98,764,887
267M of 515M

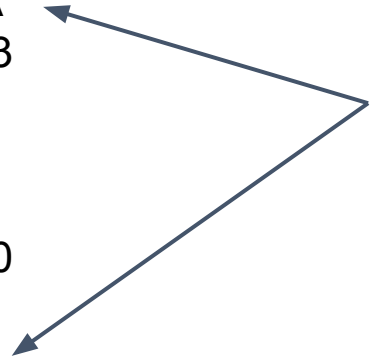

IGV
File Genomes View Tracks Regions Tools Help
D. melanogaster (dm6)
chr2L
chr2L:14,615,372-14,619,062
Go
21E2 22B4 23C4 24D6 25D2 26C4 27F5 29D6 30C3 32B1 33B2 34A5 35A2 35D2 36A4 36D2 37A5 38A8 39A2 40F4
3,692 bp
14,616,000 bp
14,617,000 bp
14,618,000 bp
Adhr
Adh
Adh
Adhr
Refseq Genes
Adh
Adhr
Adh
Adh
chr2L:14,616,053
110M of 515M

Alignment:
STAR
+
Count:
Salmon
or
featureCounts

Gene	**Salmon**	**featureCounts**
FBgn0000055	548	NA
FBgn0000286	308.999	313
FBgn0020270	82	82
FBgn0032343	15.985	13
FBgn0032358	39	35
FBgn0032633	190	190
FBgn0032634	16	16
FBgn0032901	606.999	19
FBgn0053129	479	480
FBgn0085753	201.883	NA
FBgn0261983	105.015	102
FBgn0266369	135	135
FBgn0267511	367.864	NA
FBgn0267519	361.51	NA
FBgn0267522	229.983	NA

Alignment:
STAR
+
Count:
Salmon
or
featureCounts

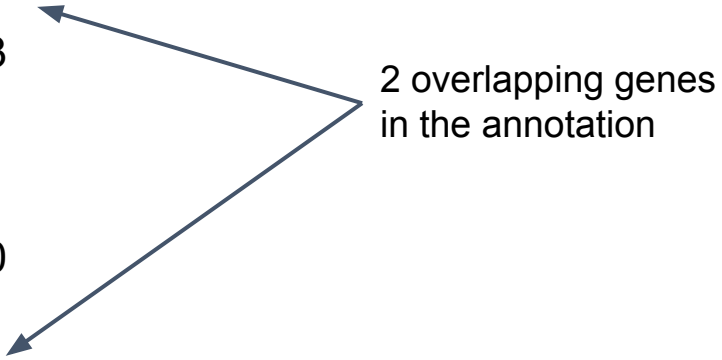
Gene	**Salmon**	**featureCounts**
FBgn0000055	**548**	**NA**
FBgn0000286	308.999	313
FBgn0020270	82	82
FBgn0032343	15.985	13
FBgn0032358	39	35
FBgn0032633	190	190
FBgn0032634	16	16
FBgn0032901	**606.999**	**19**
FBgn0053129	479	480
FBgn0085753	201.883	NA
FBgn0261983	105.015	102
FBgn0266369	135	135
FBgn0267511	367.864	NA
FBgn0267519	361.51	NA
FBgn0267522	229.983	NA

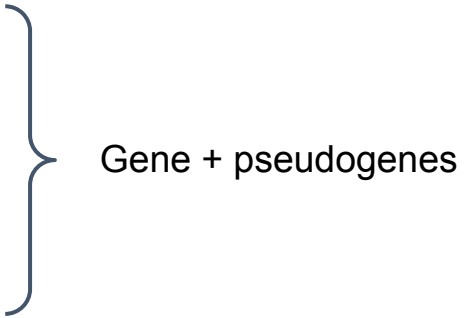


2 overlapping genes
in the annotation

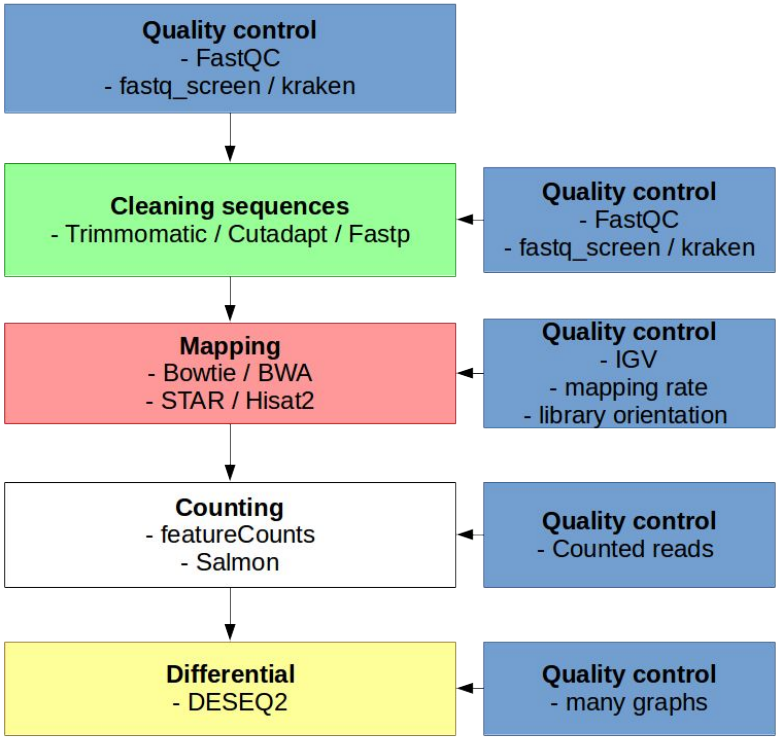
Alignment:
STAR
+
Count:
Salmon
or
featureCounts

	Gene	**Salmon**	**featureCounts**
	FBgn0000055	548	NA
	FBgn0000286	308.999	313
	FBgn0020270	82	82
	FBgn0032343	15.985	13
	FBgn0032358	39	35
	FBgn0032633	190	190
	FBgn0032634	16	16
	FBgn0032901	606.999	19
	FBgn0053129	479	480
28S rRNA gene	FBgn0085753	**201.883**	**NA**
	FBgn0261983	105.015	102
	FBgn0266369	135	135
28S rRNA pseudo	FBgn0267511	**367.864**	**NA**
28S rRNA pseudo	FBgn0267519	**361.51**	**NA**
28S rRNA pseudo	FBgn0267522	**229.983**	**NA**

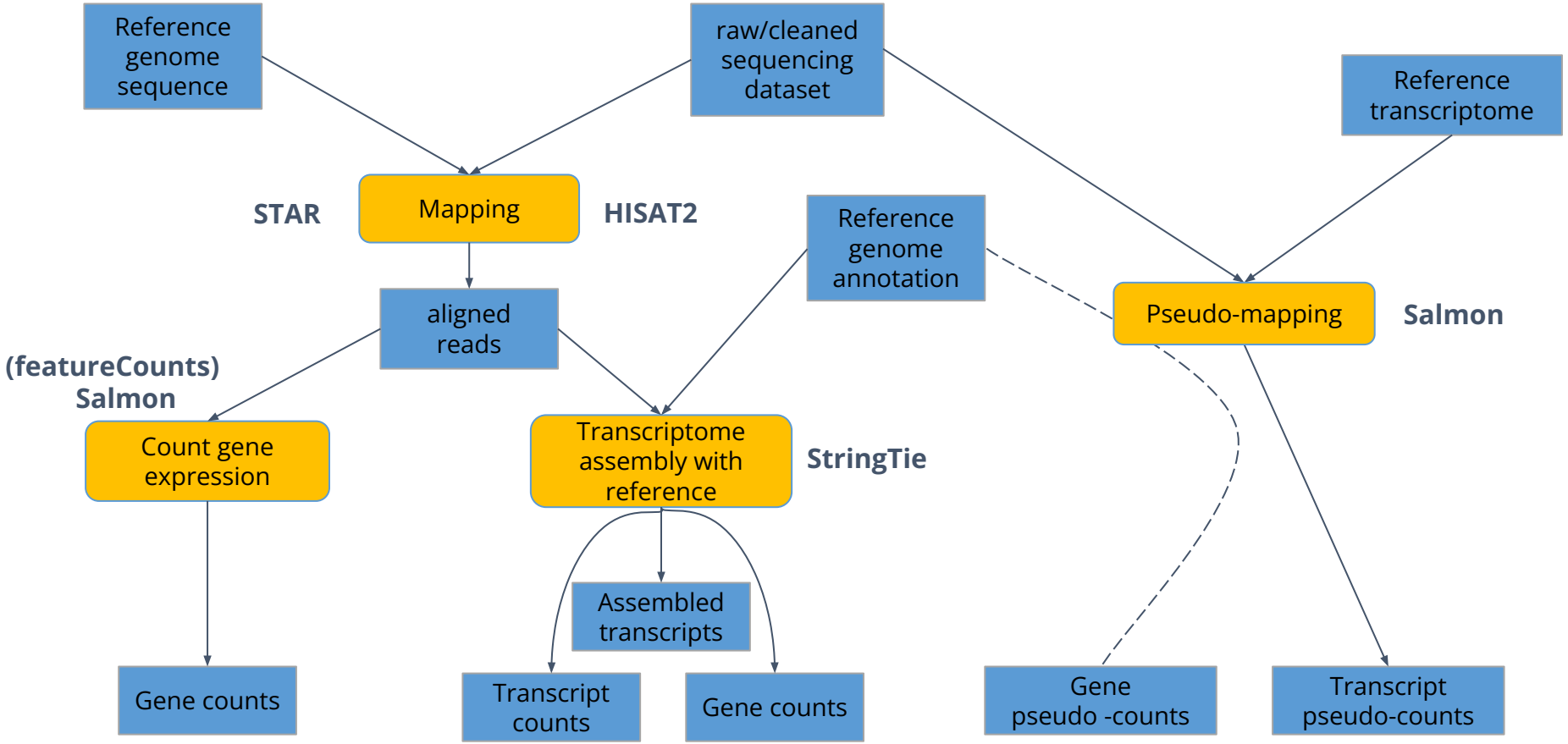




Pipeline

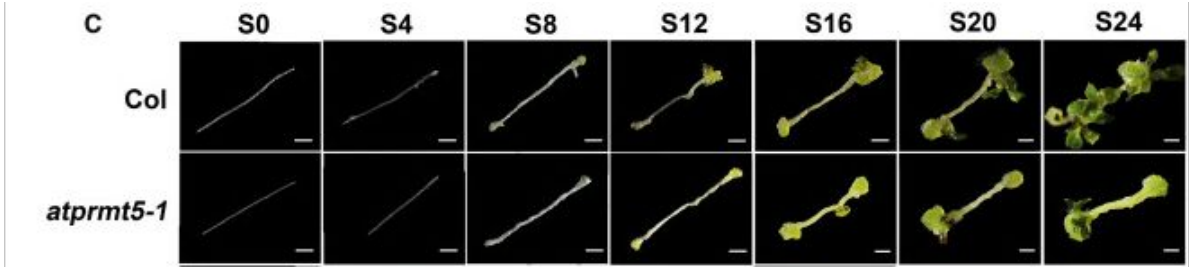

Quality control
- FastQC
- fastq_screen / kraken
Cleaning sequences
- Trimmomatic / Cutadapt / Fastp
Quality control
- FastQC
- fastq_screen / kraken
Mapping
- Bowtie / BWA
- STAR / Hisat2
Quality control
- IGV
- mapping rate
- library orientation
Counting
- featureCounts
- Salmon
Quality control
- Counted reads
Differential
- DESEQ2
Quality control
- many graphs

RNA-seq with reference (Eukaryotes)



RNA-seq experiment

Functional characterization of the protein arginine methyltransferase AtPRMT5 during de novo shoot regeneration in Arabidopsis (knock-out of AtPRMT5, T-DNA insertion)



https://doi.org/10.1016/j.molp.2016.10.010

Organism: *Arabidopsis thaliana*, plant and model organism
Genome & annotation: TAIR, the arabidopsis database
The **A**rabidopsis **I**nformation **R**esource, v. 10.1, GCF_000001735.4

Dataset: 2 conditions (WT *vs*. KO, S16), 3 biological replicates, TruSeq Stranded mRNA Library Prep Kit, paired-end sequencing (R1, R2)

Practice

- **jupyterhub with 4 CPUs & 8 GB RAM : open a terminal**
- **Change directory:**

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

- **Create a new directory:**

```
mkdir TP_rnaseq
```

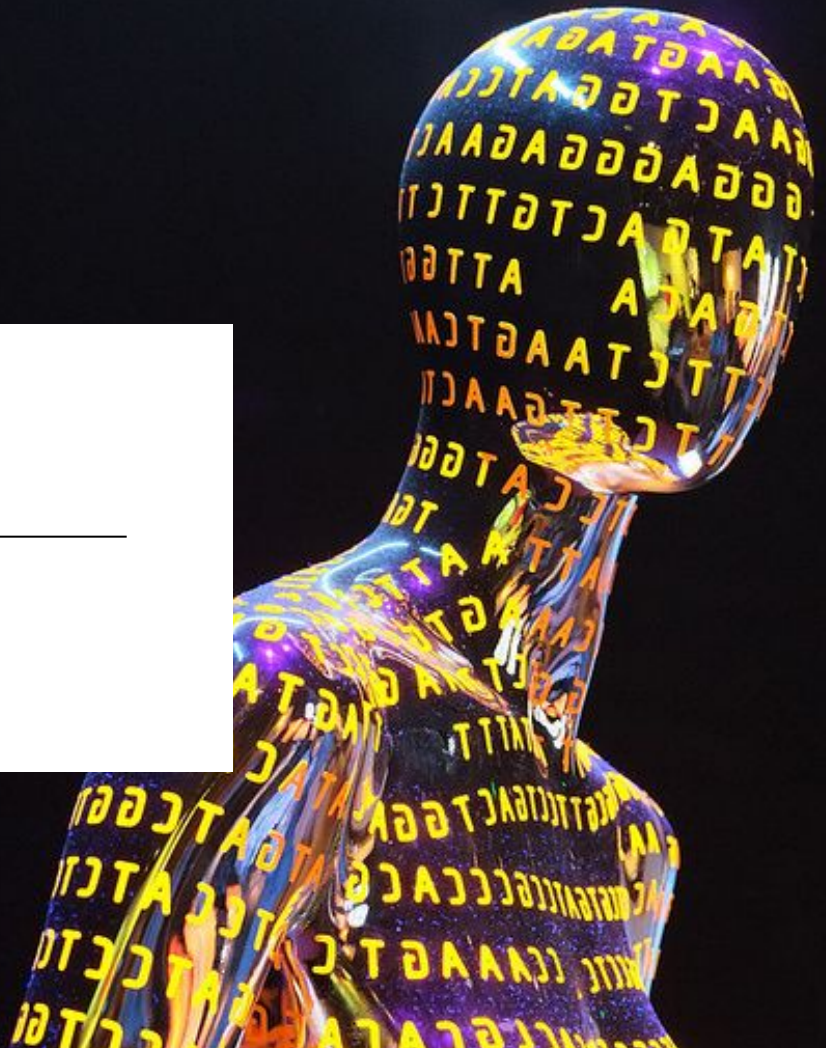
- **Copy the script template in your home:**

```
cp /shared/projects/2325_ebaii/atelier_rnaseq/01-Bioinfo/runme.sh TP_rnaseq
```

- **Follow the commands on the** `runme.sh`

Bioinformatics

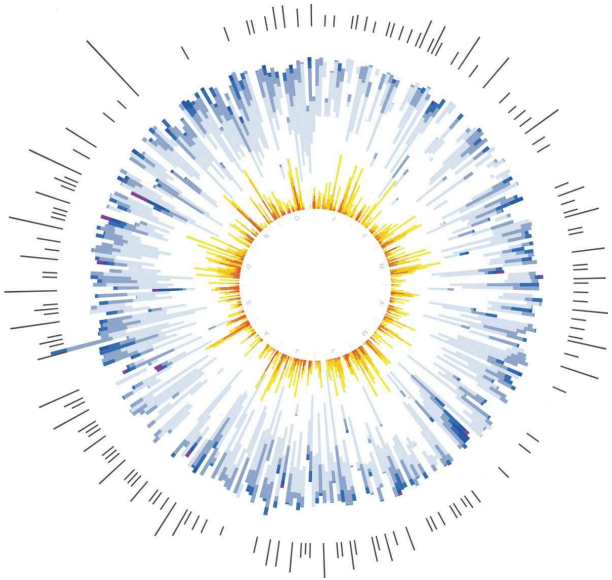
Visualize your data



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
(AtPRM5)