How to deal with your RNA-seq data ?

Claire Toffano-Nioche, Sarah Farhat
& the RNA-seq teamS (present & past)

École de Bioinformatique IFB-Inserm 2024

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

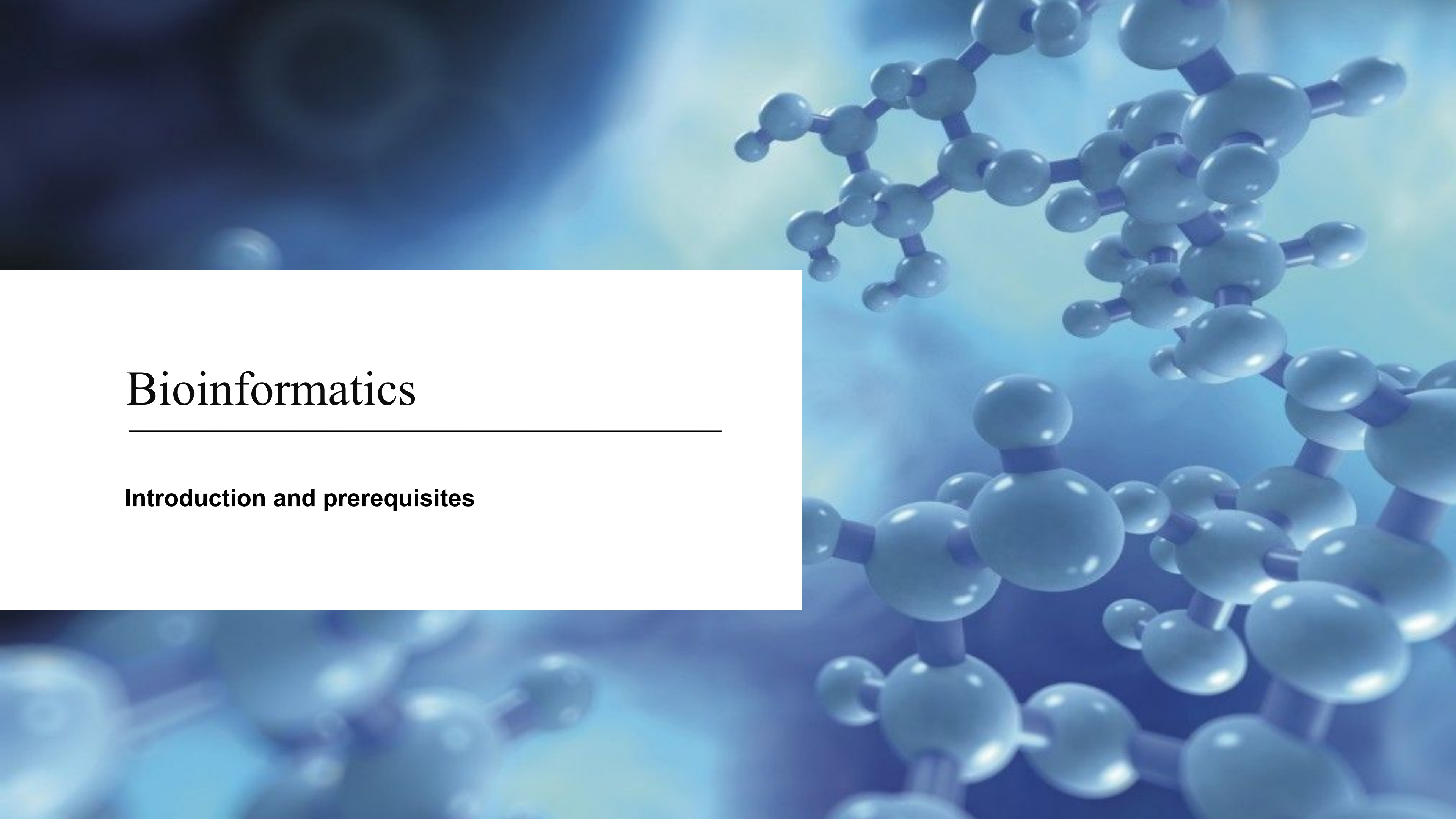
Bioinformatics

Transcriptome *de novo* assembly

07

Workflow

Automation, Reproducibility,
and Scalability

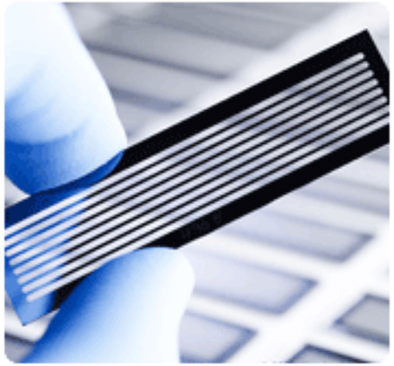


Bioinformatics

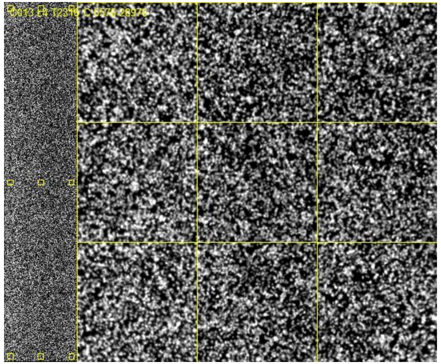
Introduction and prerequisites

Raw NGS data

Instrument



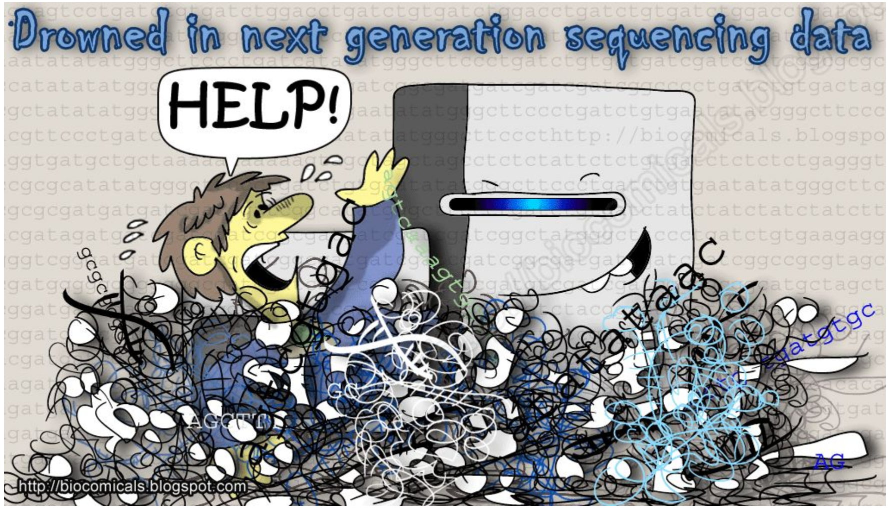
Flowcell



Intensities

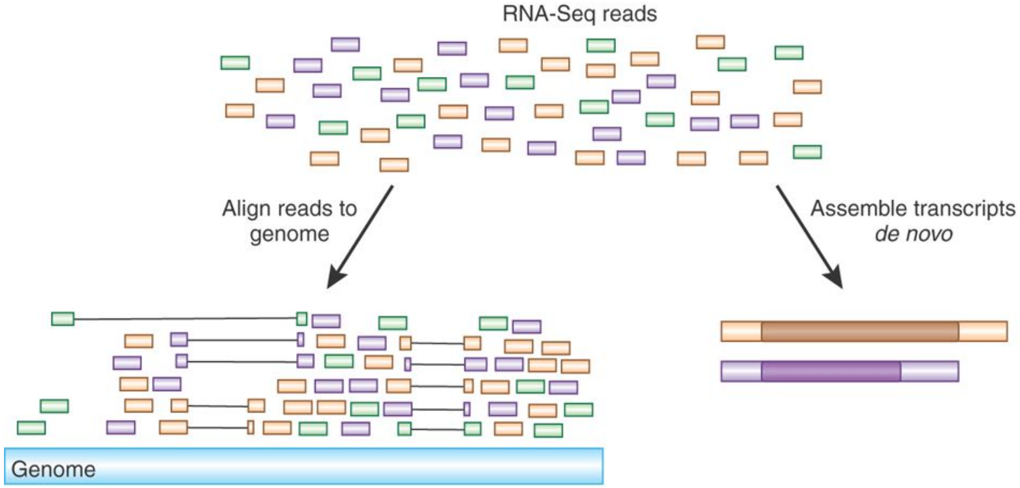
Data storage: NovaSeq6000

- 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


Drowned in next generation sequencing data
HELP!
http://biocomicals.blogspot.com

RNA-seq applications

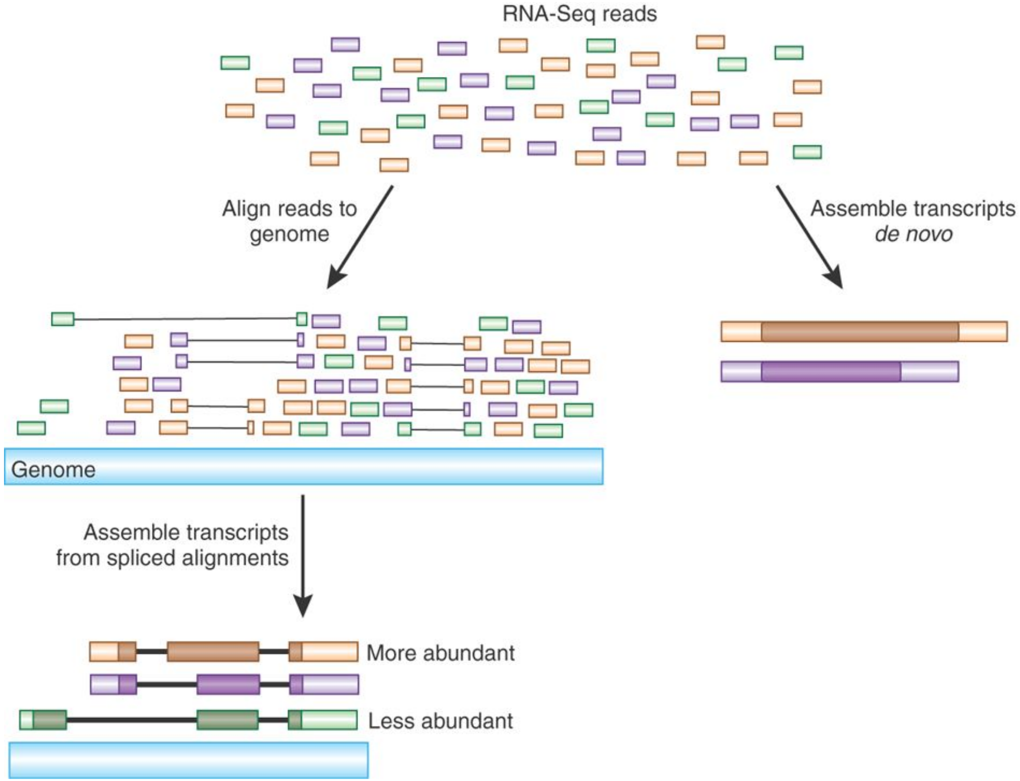
« Transcriptome analysis provides information about the identity and quantity of all RNA molecules in one cell or a population of cells »



Haas, B., Zody, M. Advancing RNA-Seq analysis. *Nat Biotechnol* 28, 421–423 (2010). https://doi.org/10.1038/nbt0510-421

RNA-seq applications

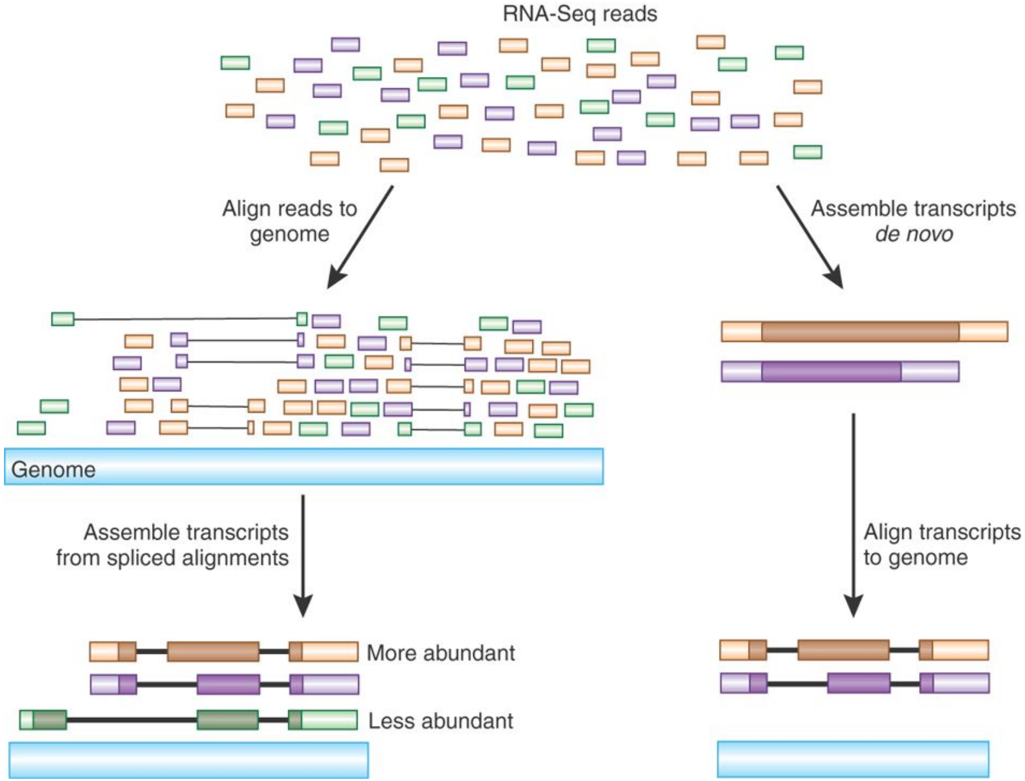
« Transcriptome analysis provides information about the identity and quantity of all RNA molecules in one cell or a population of cells »


RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
More abundant
Less abundant

Haas, B., Zody, M. Advancing RNA-Seq analysis. *Nat Biotechnol* 28, 421–423 (2010). https://doi.org/10.1038/nbt0510-421

RNA-seq applications

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

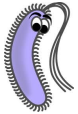

RNA-Seq reads
Align reads to genome
Assemble transcripts *de novo*
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Haas, B., Zody, M. Advancing RNA-Seq analysis. *Nat Biotechnol* 28, 421–423 (2010). https://doi.org/10.1038/nbt0510-421

RNA-seq: Why ? How

ight 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:

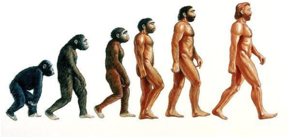
➔ Directional protocol (standard)

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

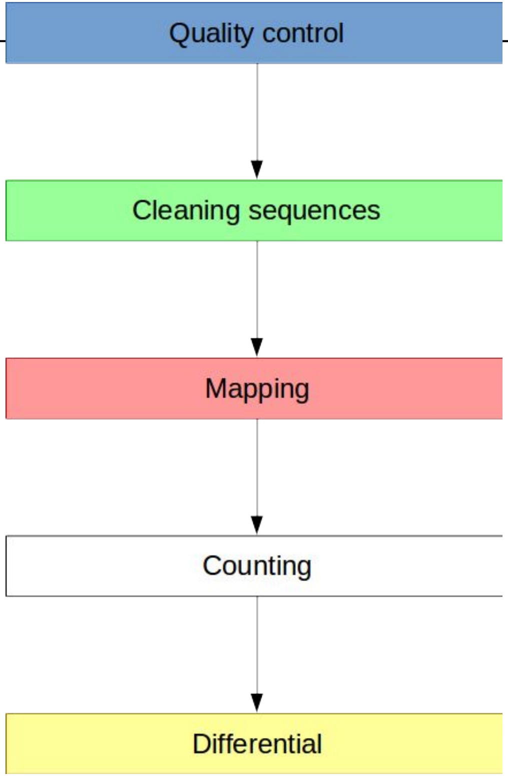
For a successful experiment, it's imperative to include bioinformatician and biostatistician before the beginning of the RNA extraction

Prerequisites

RNA sample:

□ DNAse treatment
□ Quantity (adapted protocole)
□ Quality (RNA Integrity Number, RIN > 7)
□ Stocked at -80°C

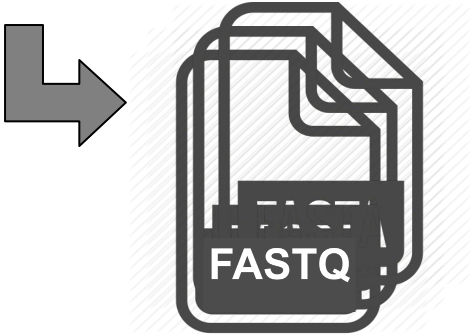
RNA-seq management:


Quality control
Cleaning sequences
Mapping
Counting
Differential

Prerequisites

RNA sample:

- □ DNAse treatment
- □ Quantity (adapted protocole)
- □ Quality (RNA Integrity Number, RIN > 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 and with positions given by base pair numbering

Where find the genome and the annotation ?

Common databases

Ensembl

UCSC

NCBI National Center for Biotechnology Information

Specific databases

FlyBase

EcoCyc™

MicroScope

GENCODE

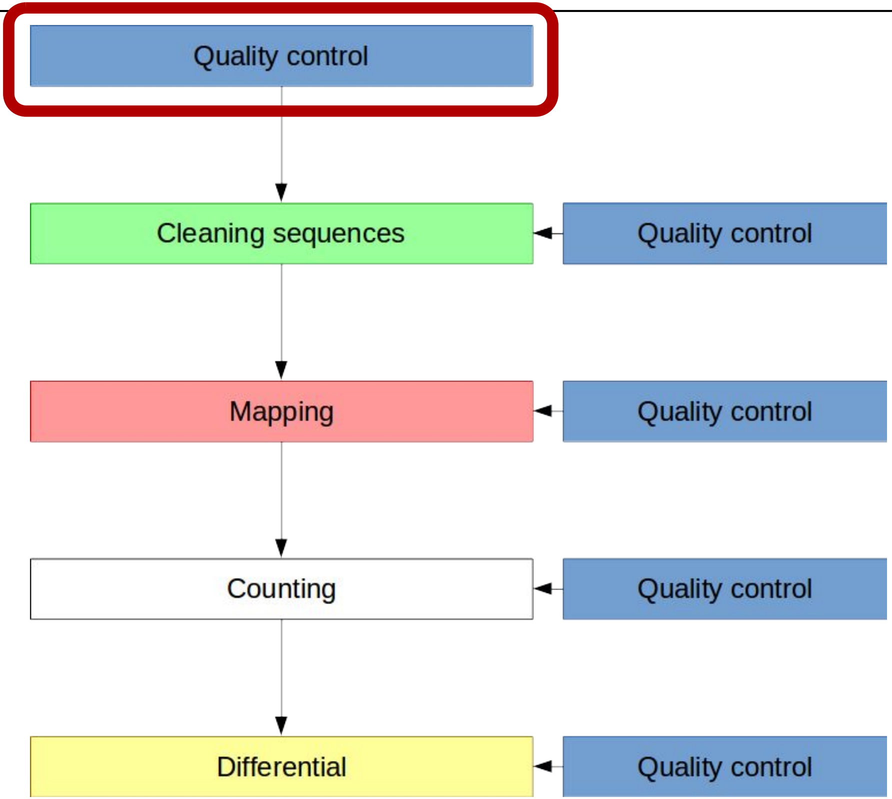
tair

SGD Saccharomyces GENOME DATABASE

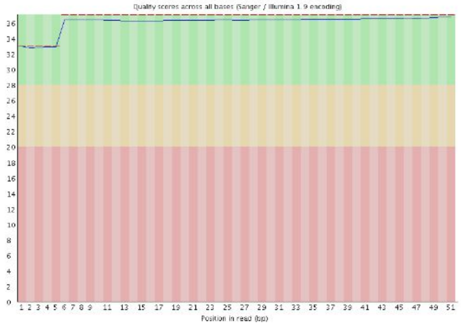
AmoebaDB Amoeba Genomics Resource

VectorBase Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Keep control on your data

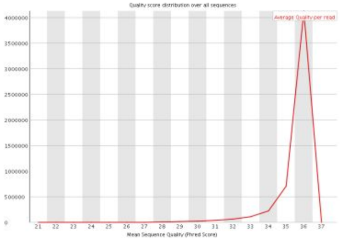


FASTQC: explore quality scores

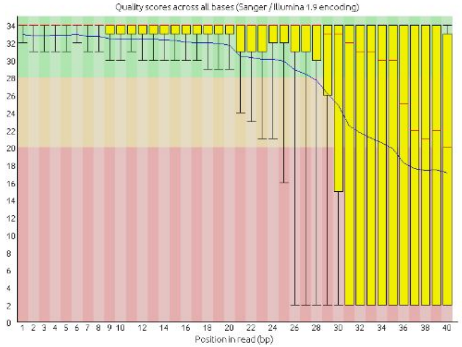


QUALITY CONTROL APPROVED

Illumina HISEQ2500

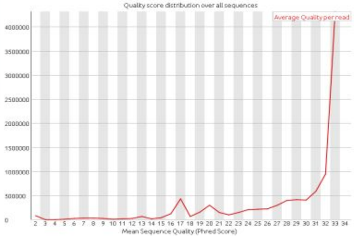


✓ The per base sequence quality are very high along sequence



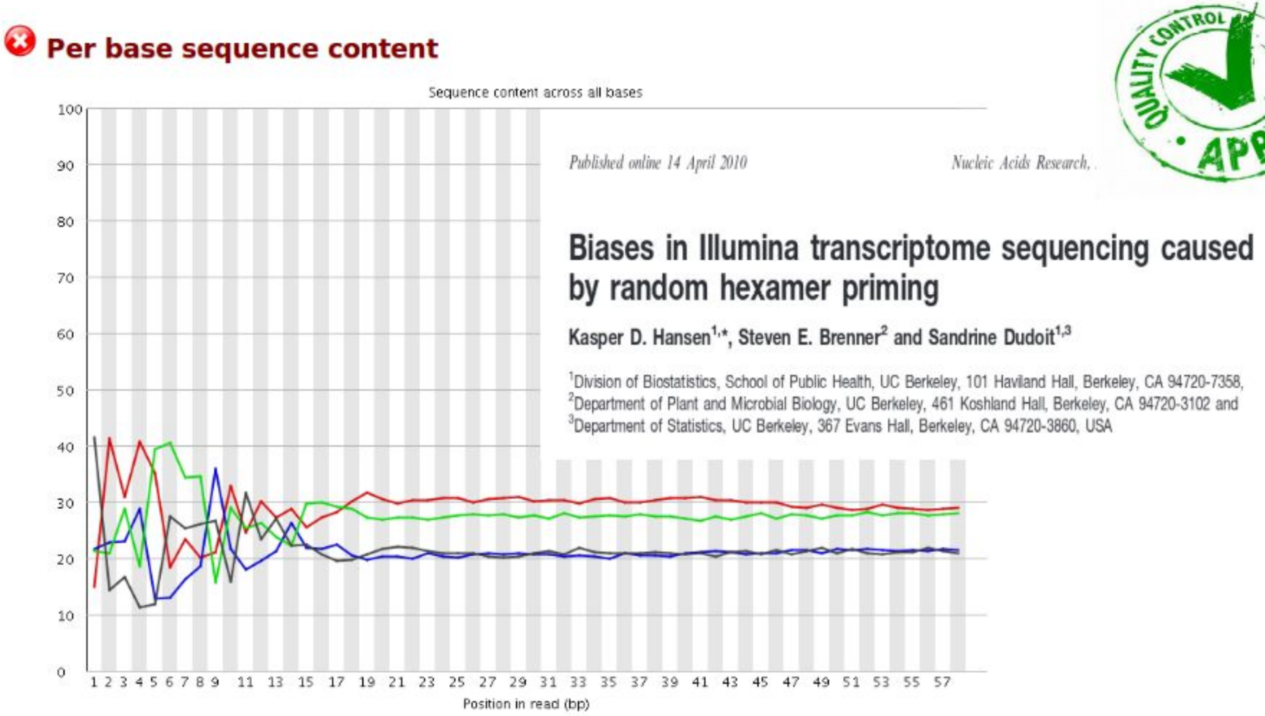
QUALITY CONTROL • REJECTED •

Illumina HISEQ2000



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

FASTQC: explore quality scores



Published online 14 April 2010 *Nucleic Acids Research,*

Biases in Illumina transcriptome sequencing caused by random hexamer priming

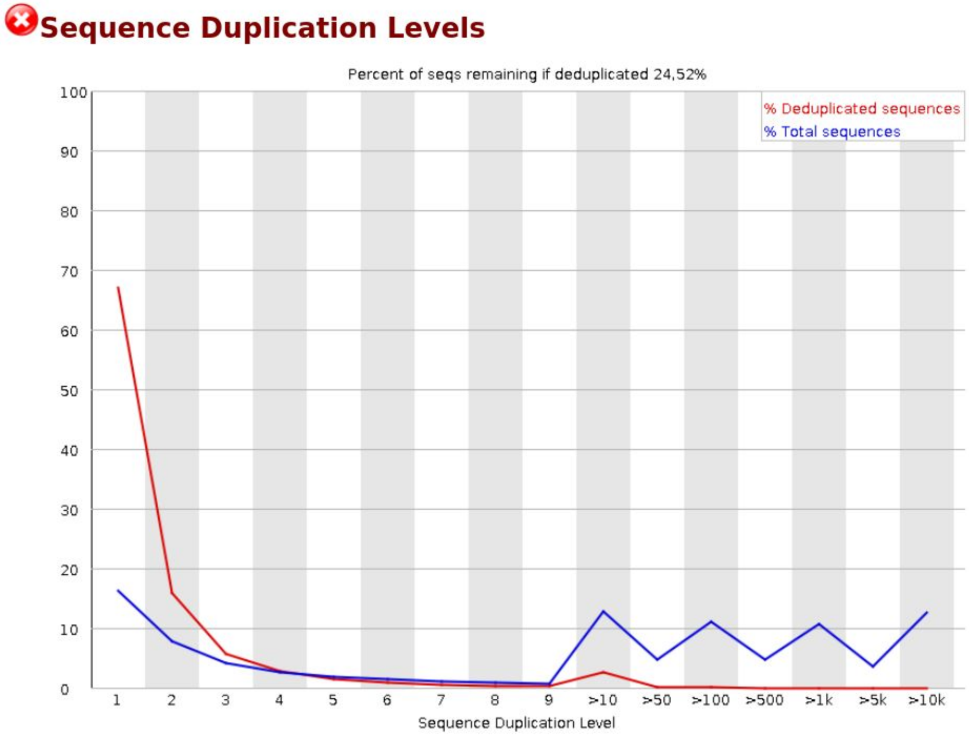
Kasper D. Hansen[1,*], Steven E. Brenner[2] and Sandrine Dudoit[1,3]

[1]Division of Biostatistics, School of Public Health, UC Berkeley, 101 Haviland Hall, Berkeley, CA 94720-7358, [2]Department of Plant and Microbial Biology, UC Berkeley, 461 Koshland Hall, Berkeley, CA 94720-3102 and [3]Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA

QUALITY CONTROL APPROVED

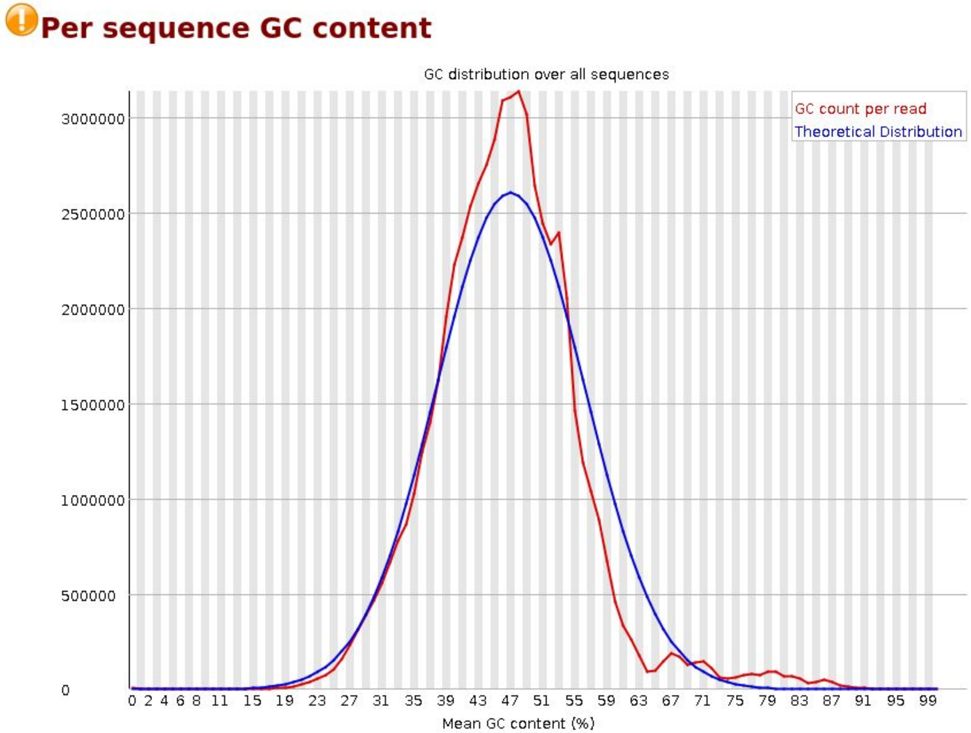
FASTQC: explore quality scores

Systematic high duplication level in RNA-seq, why ?



QUALITY CONTROL APPROVED

FASTQC: explore quality scores

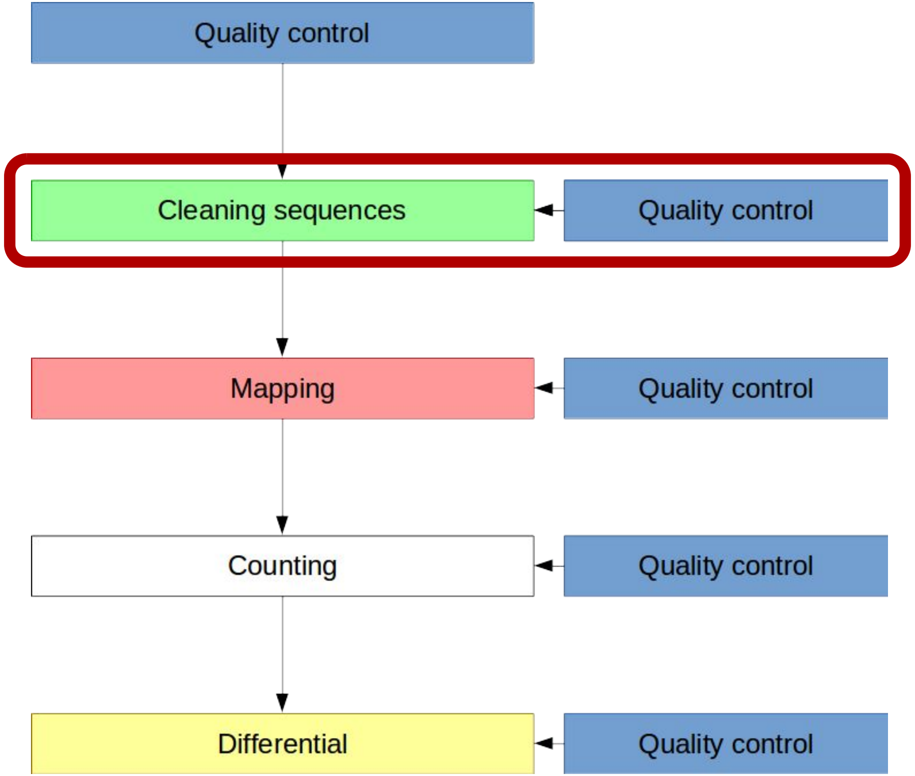


e.g. mouse RNAseq sample

Within the coding, intron and flanking DNA functional compartments of largely single copy genes in mouse, GC content is 47.36%

https://bionumbers.hms.harvard.edu/bionumber.aspx?id=102409

Pipeline: cleaning step (if needed)



Quality control

Cleaning sequences

Quality control

Mapping

Quality control

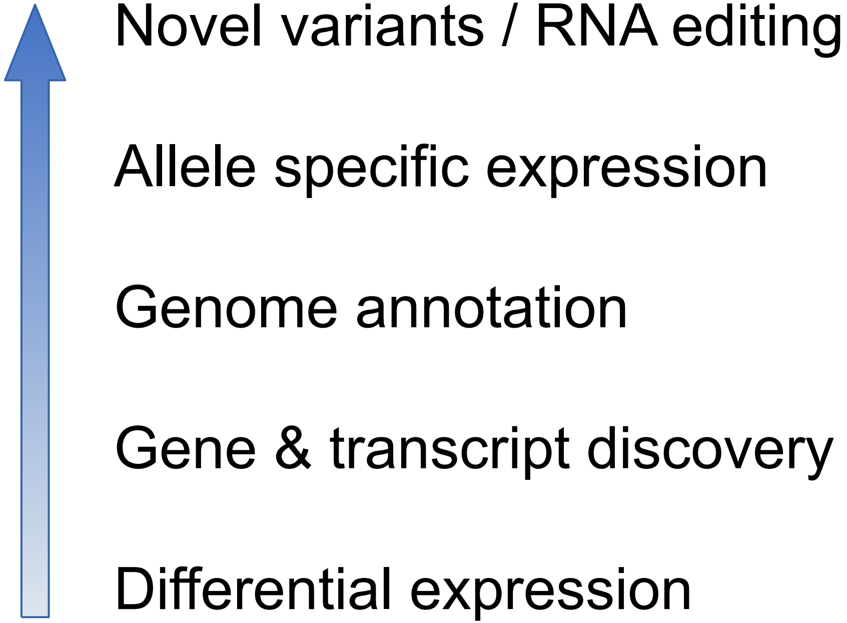
Counting

Quality control

Differential

Quality control

Clean to get the best quality possible?



Novel variants / RNA editing

Allele specific expression

Genome annotation

Gene & transcript discovery

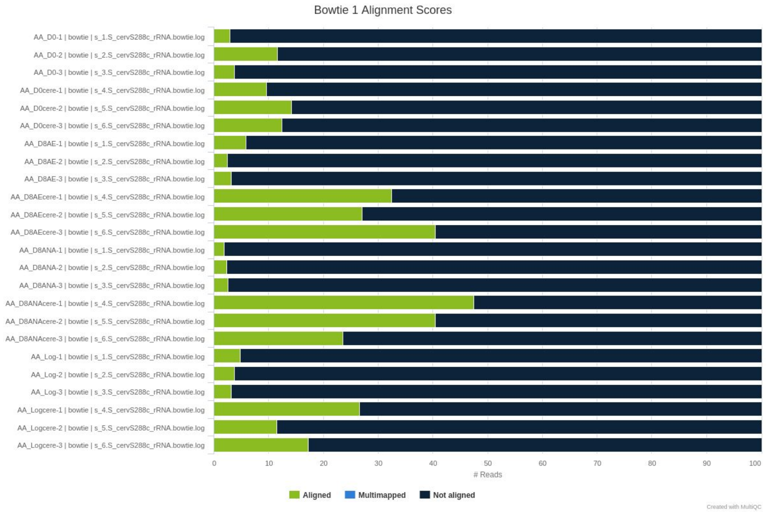
Differential expression

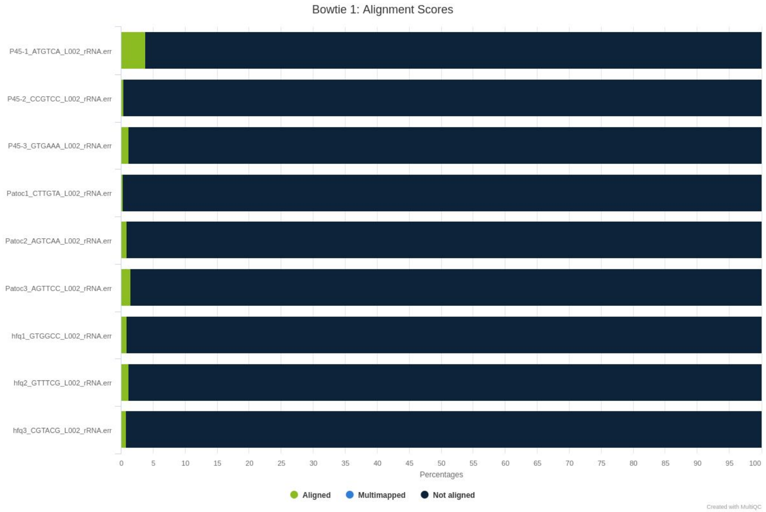
How to screen contaminations ?

Different levels:

🟩 Aligned 🟦 Multimapped ⬛ Not aligned

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

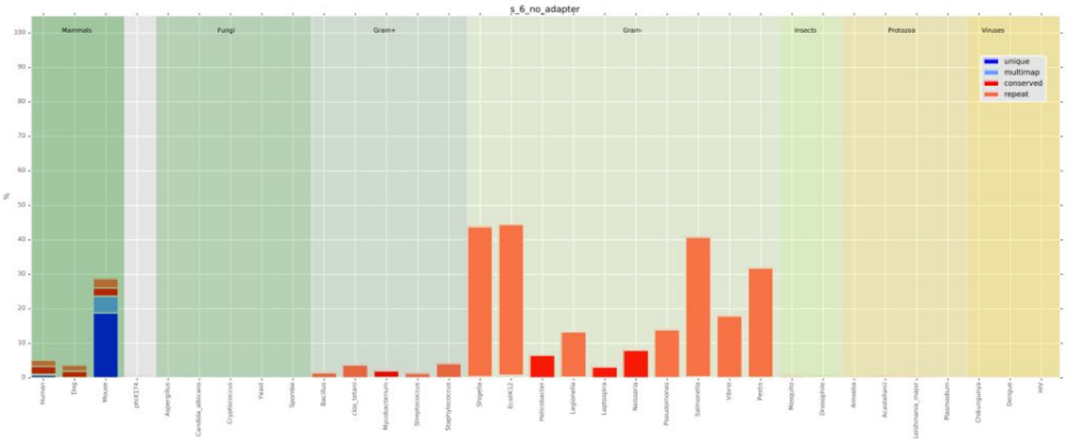


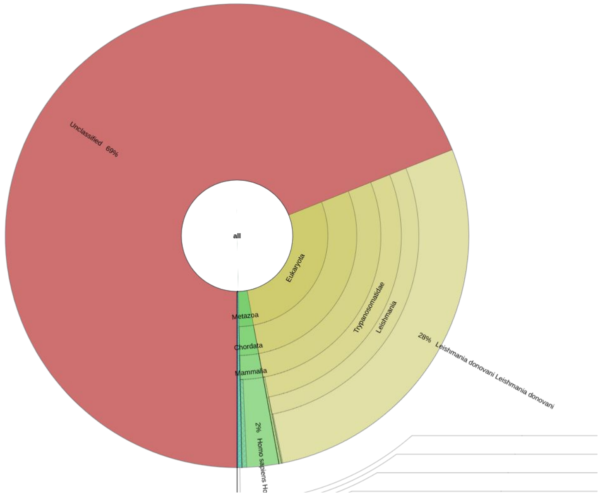


How to screen contaminations ?

Different levels:

- Ribosomal contamination from same organism
- RNA contamination from other organism
    - Use dedicated or derived tools such as fastq_screen or kraken2

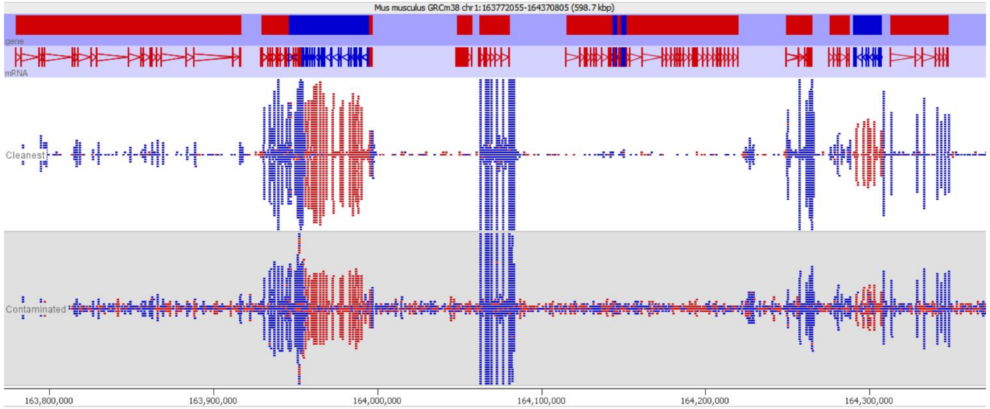




How to screen contaminations ?

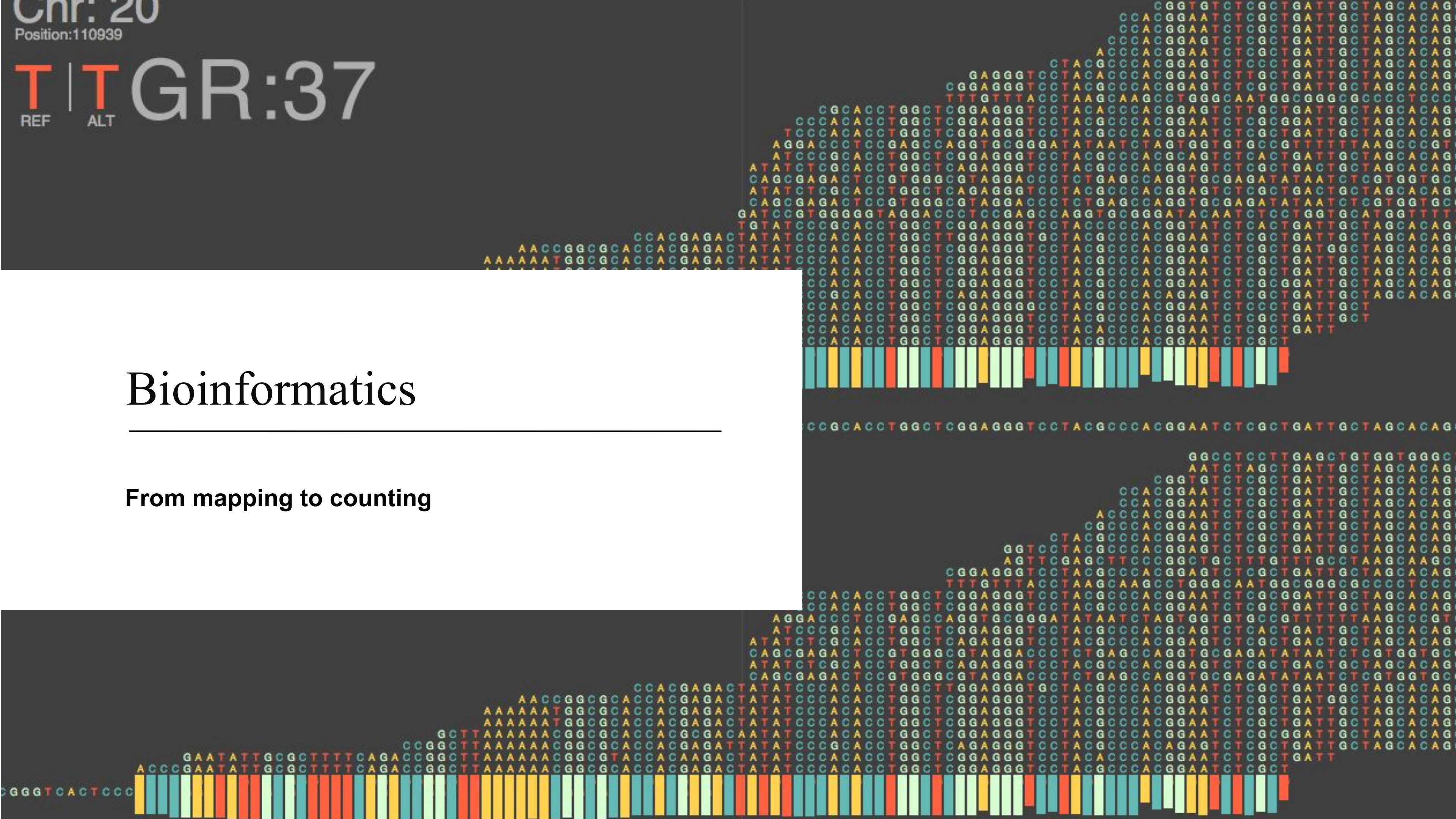
Different levels:

- 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

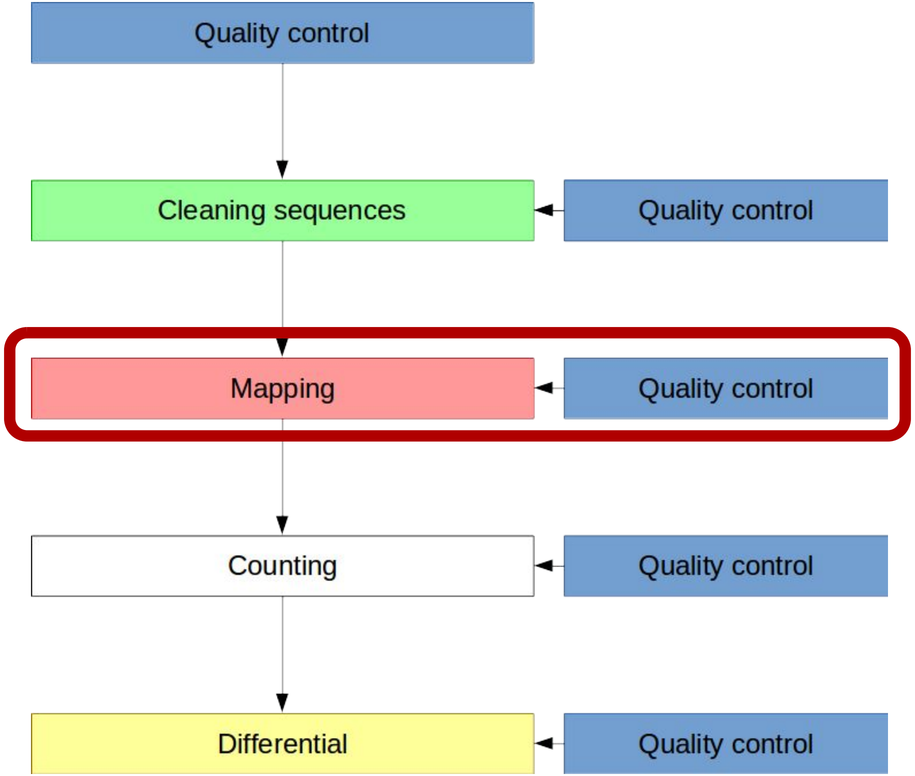


Bioinformatics

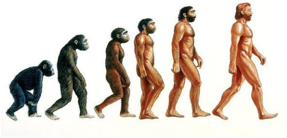
From mapping to counting



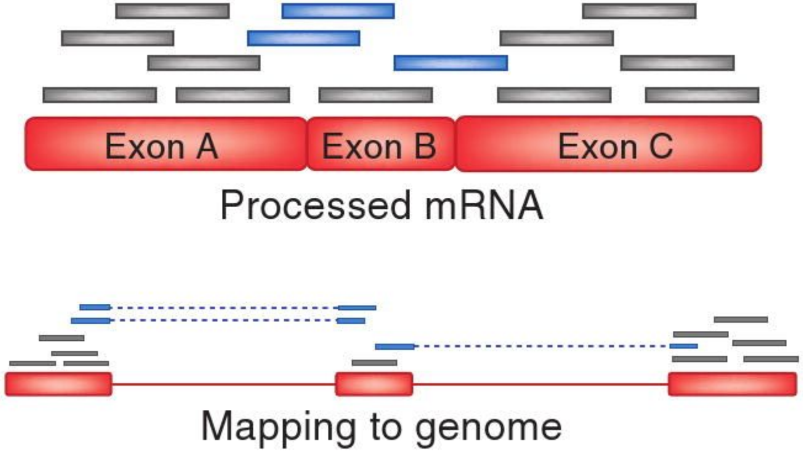
Pipeline: mapping step



RNA-seq mapping specificity

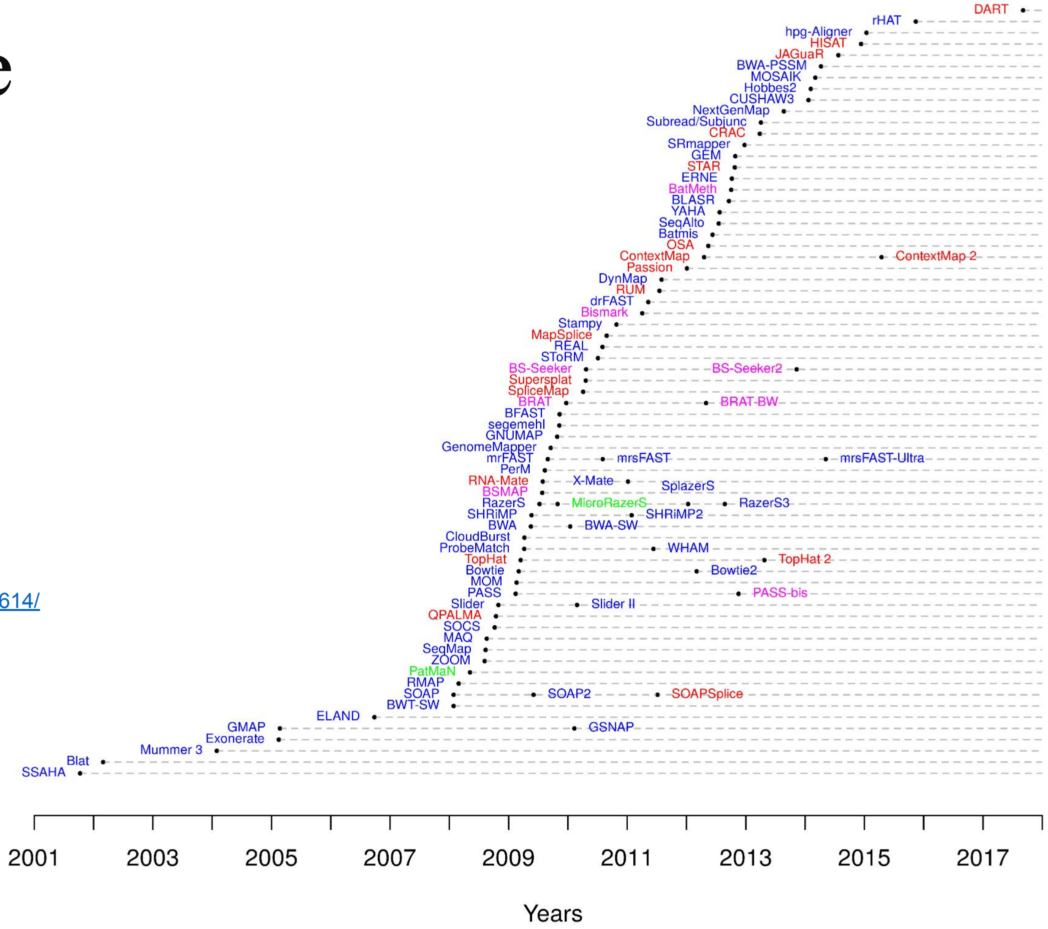


★ Take account to reads that come from exon-exon junctions



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

Mapping timeline



From https://pubmed.ncbi.nlm.nih.gov/23060614/

Choose the good mapper

Which one is the best mapper ?

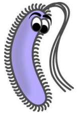
Choose the good mapper

~~Which one is the best mapper ?~~

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

Choose the good mapper

Prokaryotes



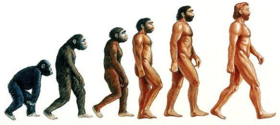
For short reads in 99% of the use-cases:

➔ 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

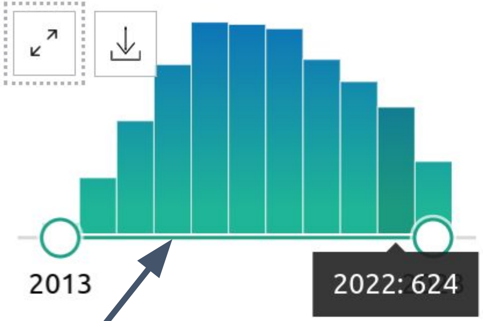
For very small RNA (e.g. miRNA-seq):

➔ BWA, Bowtie

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

Choose the good mapper

Many people uses TopHat2
(>10K citations in Scolar, > 1K 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* ».

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.

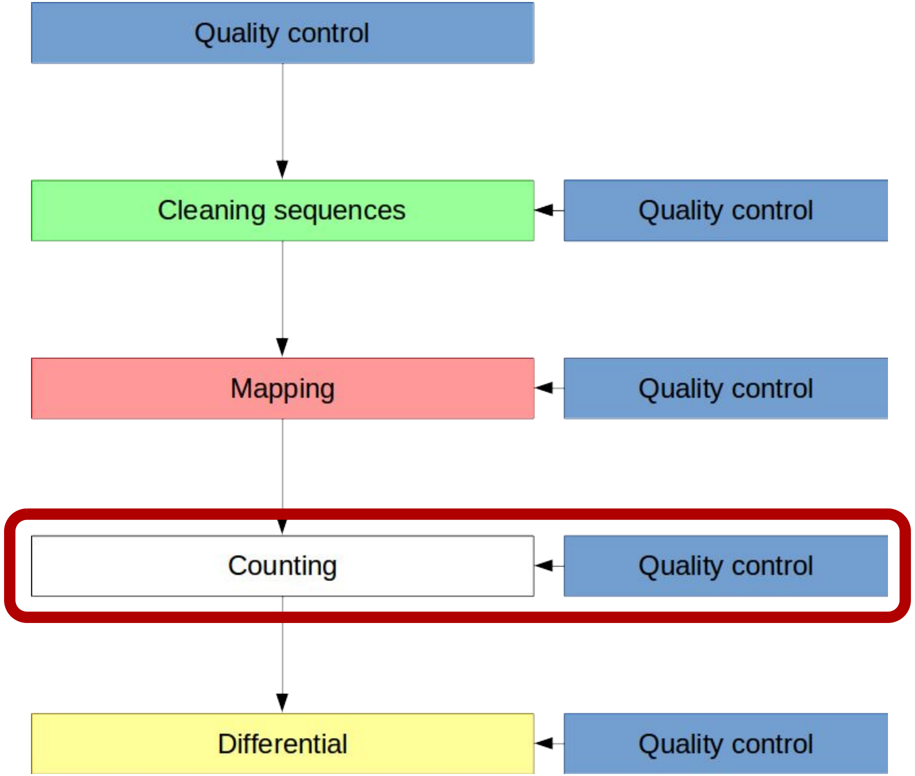
Mapping QC (Quality Check) on RNA-seq

- ★ Percentage of mapped reads along genome
  - ○ Human/Mouse: 70 to 90 %
  - ○ Prokaryotic: > 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, tools

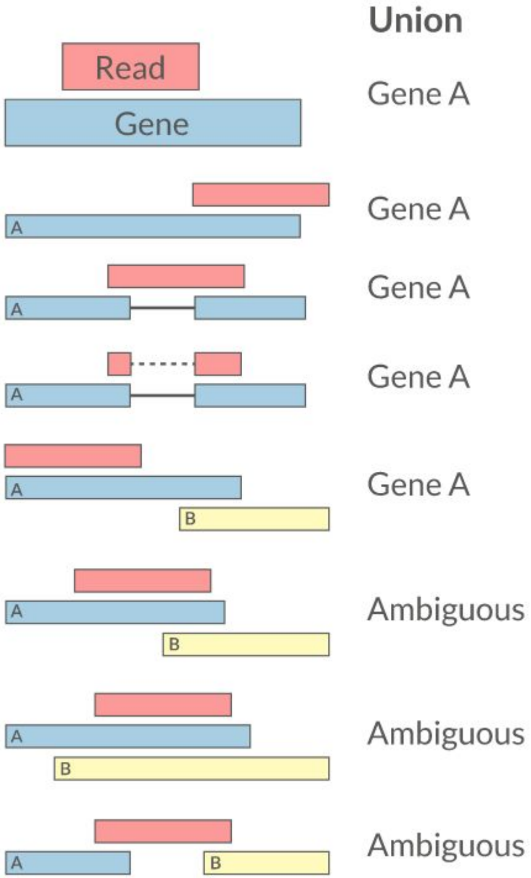
- ★ Common :
  - ○ Samtools (flagstats)
  - ○ Bamtools (stats)
  - ○ Picardtools (CollectRnaSeqMetrics)
  - ○ RseQC
- ★ Human and mouse :
  - ○ RNAseQC
  - ○ Qualimap

Pipeline: counting step



Quantification / Count

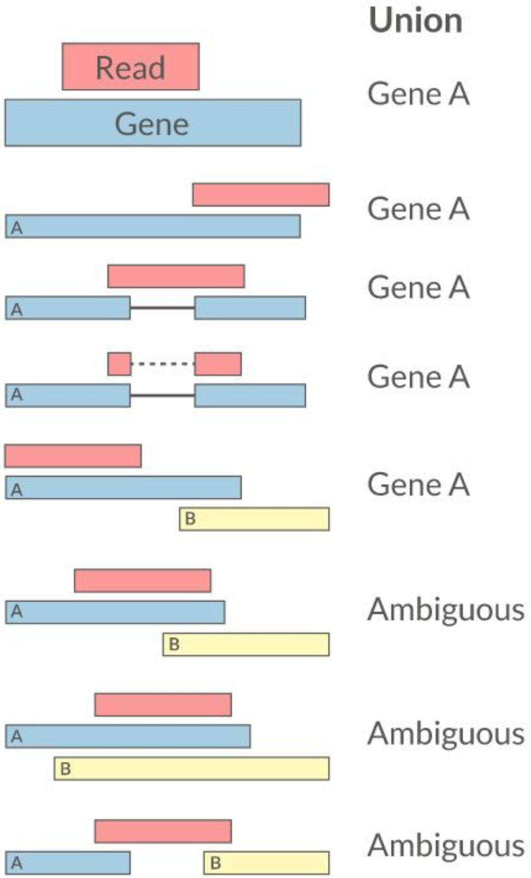
- ★ Reads count ~ gene expression
- ★ Reads can be quantified on any features (gene, transcript, exon, etc)
- ★ Manage:
  - ○ intersection on gene models
  - ○ gene / transcript level



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

Quantification / Count

- ★ Reads count ~ gene expression
- ★ Reads can be quantified on any features (gene, transcript, exon, etc)
- ★ Manage:
  - ○ intersection on gene models
  - ○ gene / transcript level

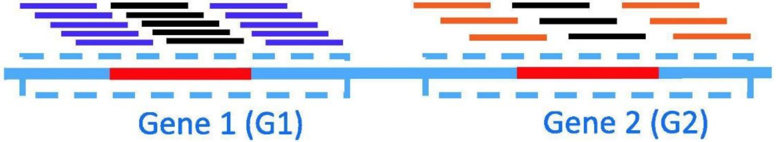


ambiguous or **multi-mapped** reads

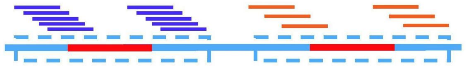
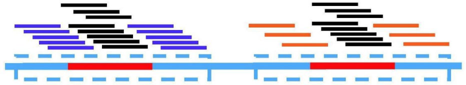
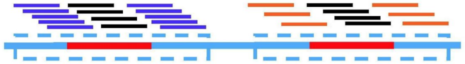
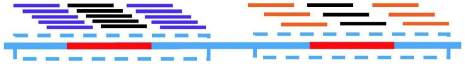
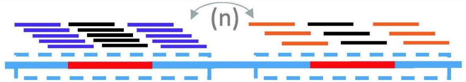
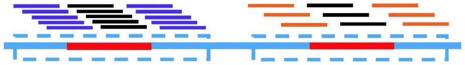
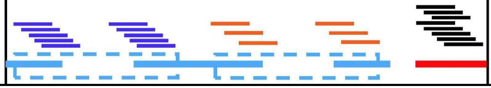
credit: SciLife lab https://scilifelab.github.io/courses/ngsintro/1905/slides/rnaseq/presentation.html#33

Quantification / Count

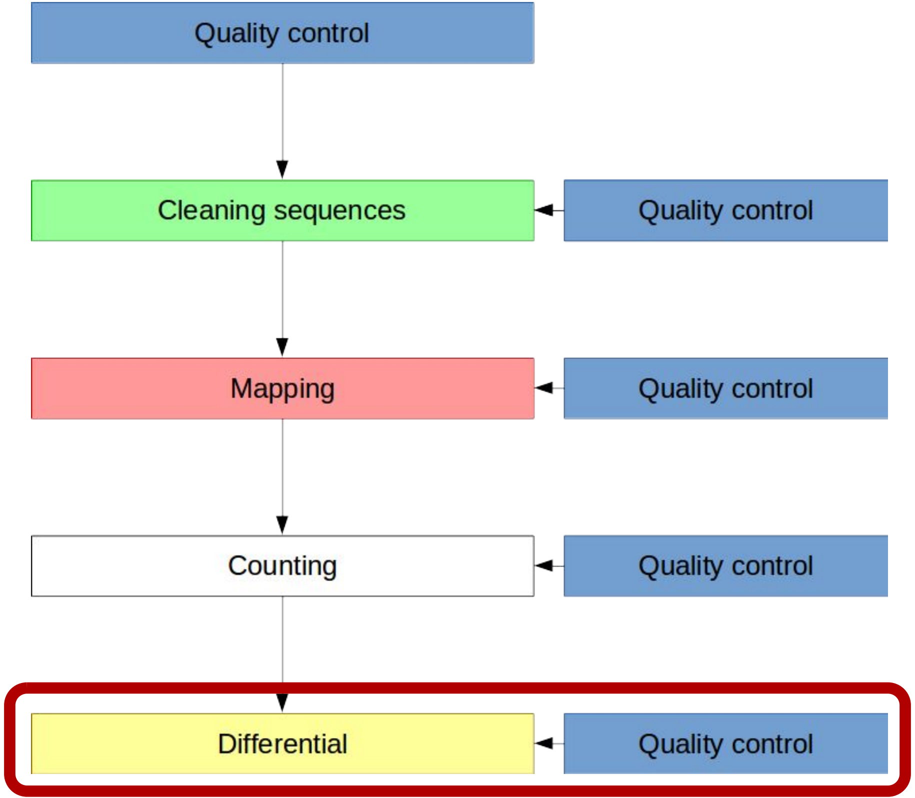
How to handle « multi-mapped » reads?



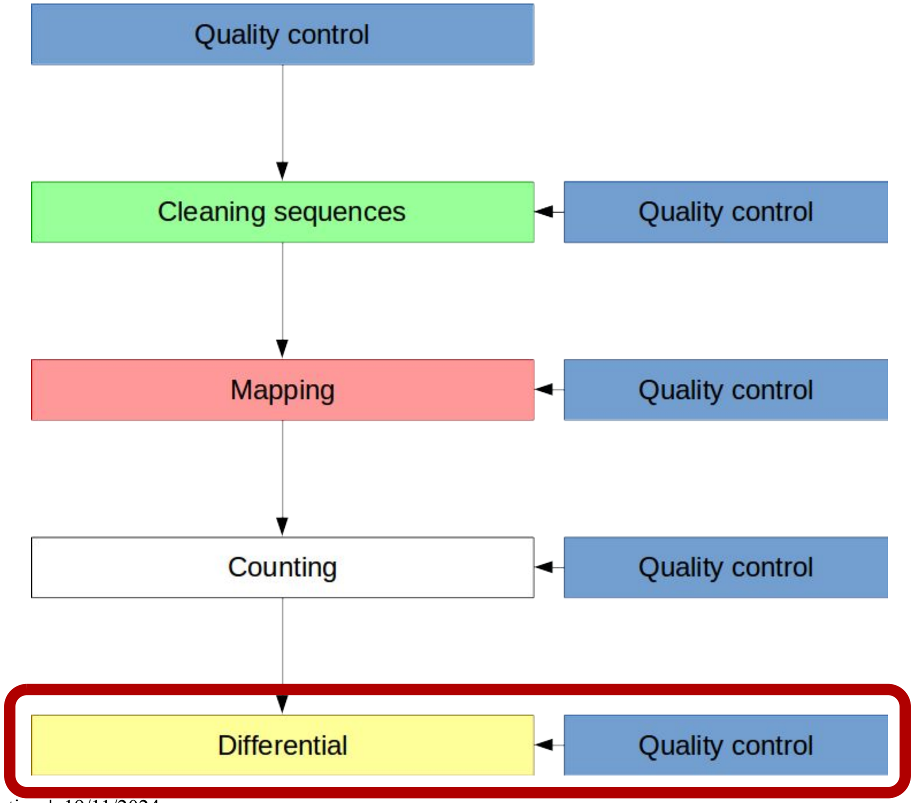
Deschamps-Francoeur, et al. 2020. doi:10.1016/j.csbj.2020.06.014

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		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

Pipeline: biostatistical step



Pipeline: biostatistical step



Quality control

Cleaning sequences ← Quality control

Mapping ← Quality control

Counting ← Quality control

Differential ← Quality control

02

Statistics

Experimental design,
Exploratory data analysis

03

Statistics

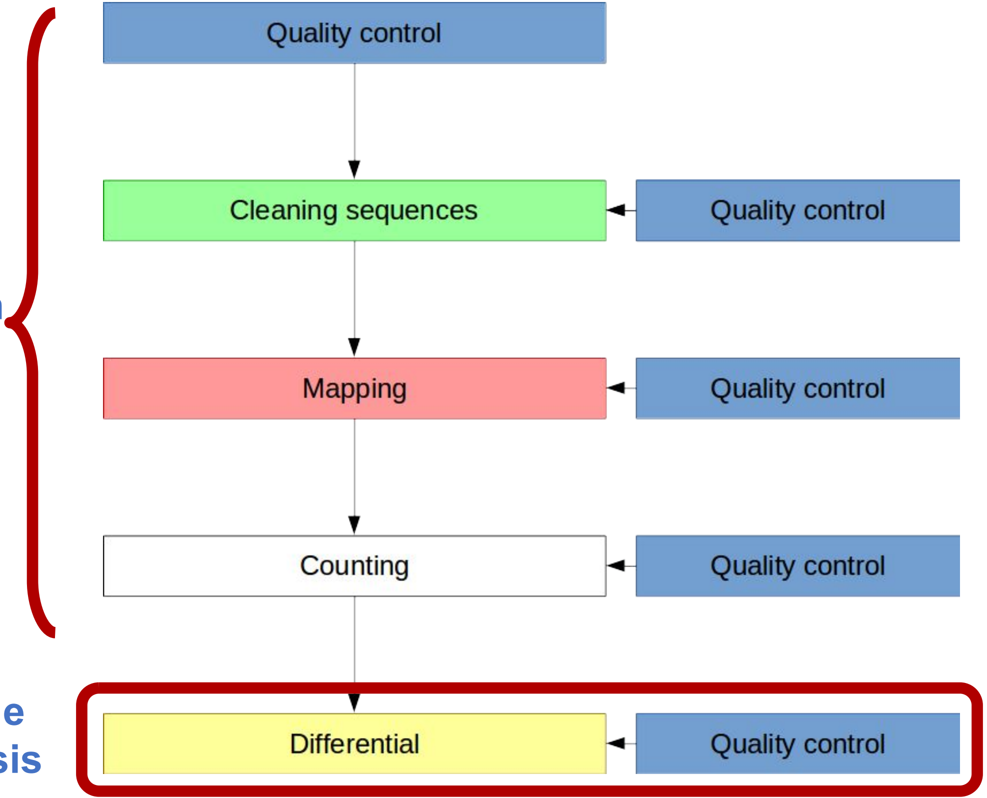
Normalization, modelisation
and troubleshooting

04

Practice

Differential analysis
with SARTools

Pipeline: input files



Quality control

Cleaning sequences ← Quality control

Mapping ← Quality control

Counting ← Quality control

Differential ← Quality control

To do for each sample

Associate all sample counts in one analysis

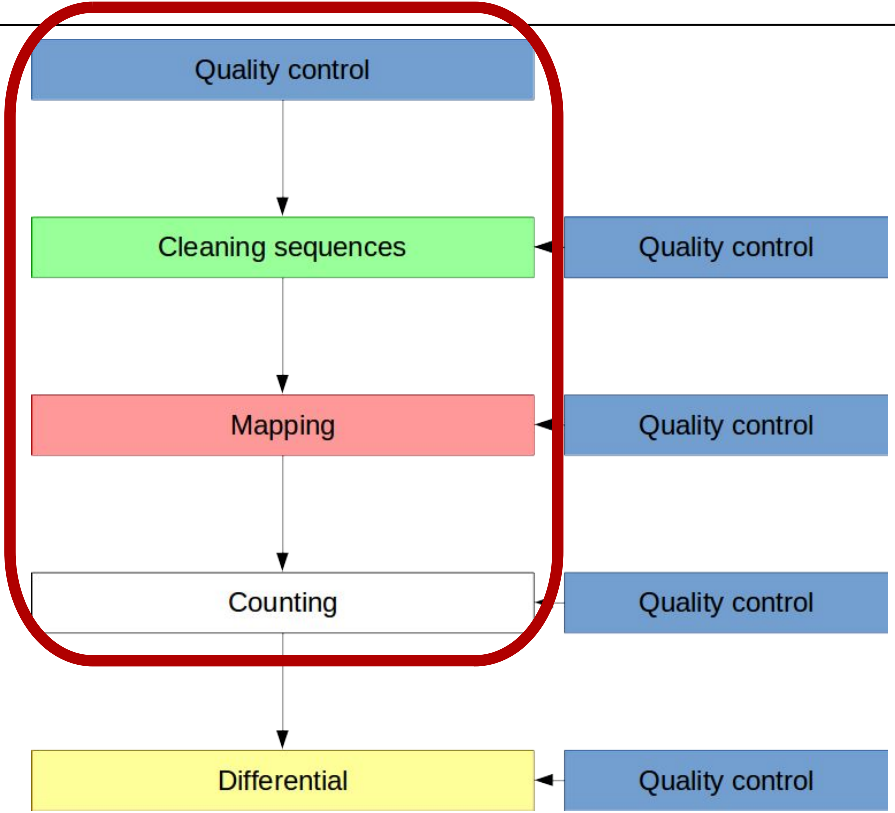
Pipeline: computing steps

FastQC

FastP

STAR

featureCounts



Quality control

Cleaning sequences

Quality control

Mapping

Quality control

Counting

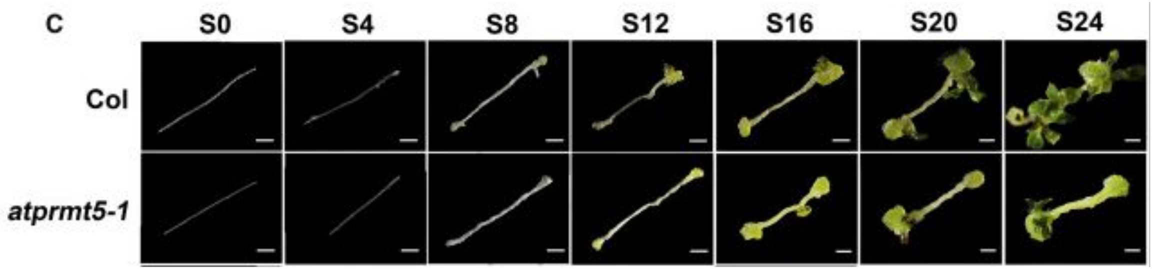
Quality control

Differential

Quality control

RNA-seq experiment

Functional characterization of the protein arginine methyltransferase AtPRMT5 during *de novo* shoot regeneration in Arabidopsis (6 days in callus-induction medium+Sx=x days in shoot-induction medium). *atprmt5-1*: knock-out of AtPRMT5 by T-DNA insertion



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

Organism: *Arabidopsis thaliana*, plant and model organism

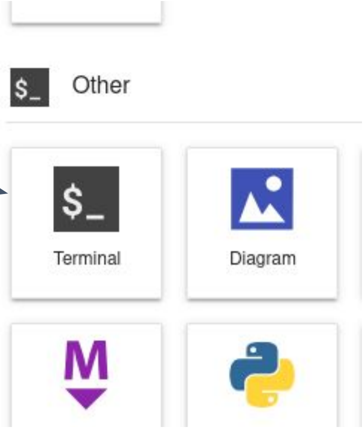
Genome & annotation: **T**he **A**rabidopsis **I**nformation **R**esource, TAIR v. 10.1, GCF_000001735.4

Dataset: 2 conditions (WT *vs.* KO *atprmt5-1* S16, 3 biological replicates, TruSeq Stranded mRNA Library Prep Kit, paired-end sequencing (R1, R2)

https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-5044/sdrf

Practice

- **jupyterlab `with 4 CPUs & 8 GB RAM : open a terminal`**



- **If needed change directory and go to your project space:**

```
cd /shared/projects/<YOUR_PROJECT_NAME>
```

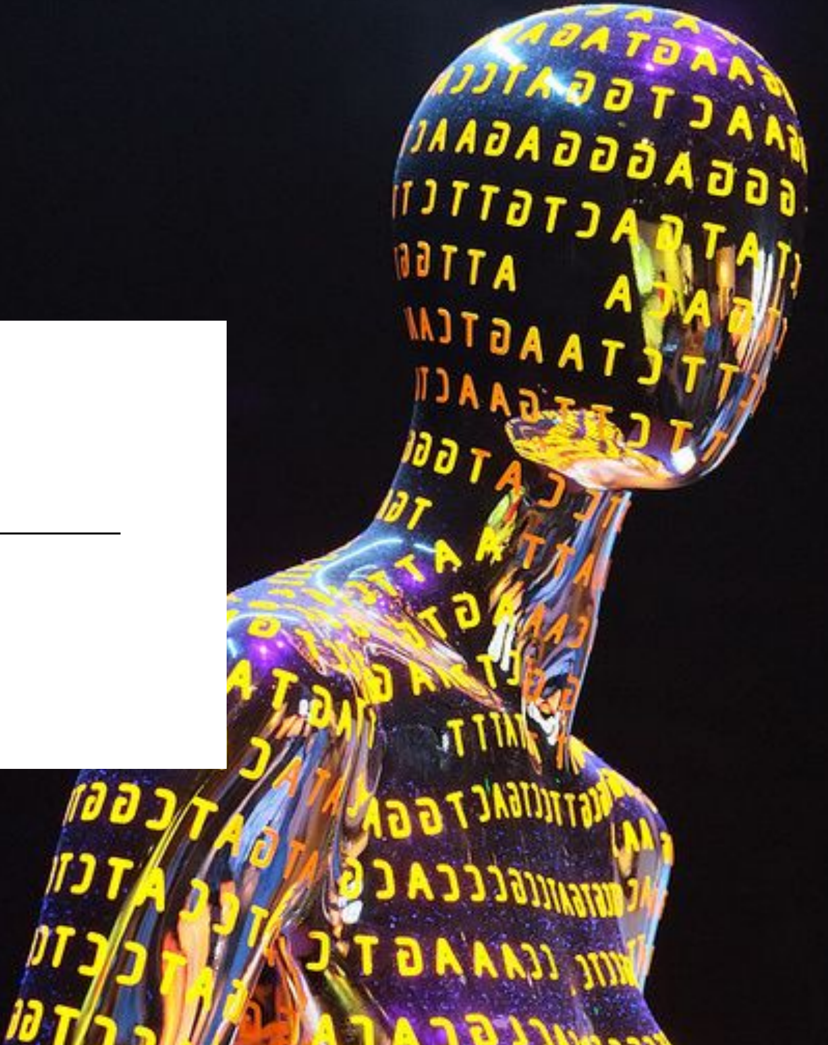
- **Copy the repository in your projet space:**

```
cp -r /shared/projects/2422_ebaii_n1/atelier_rnaseq/01-Bioinfo/ .
```

- **Open the runme.ipynb file**

Bioinformatics

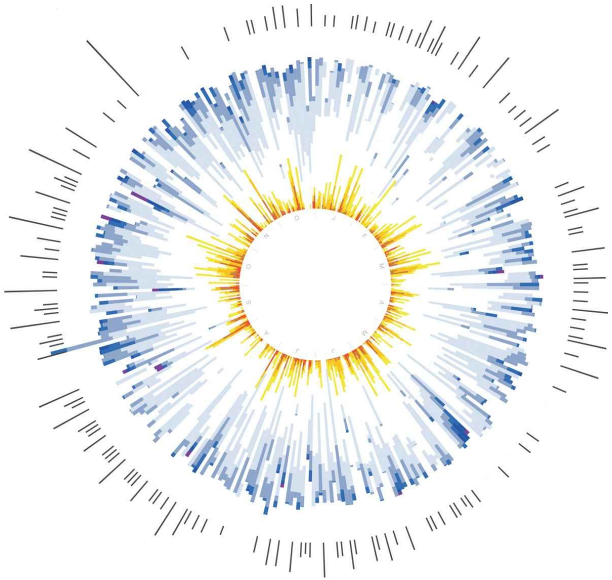
Visualize your data



Visualize alignments

Which format ?

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

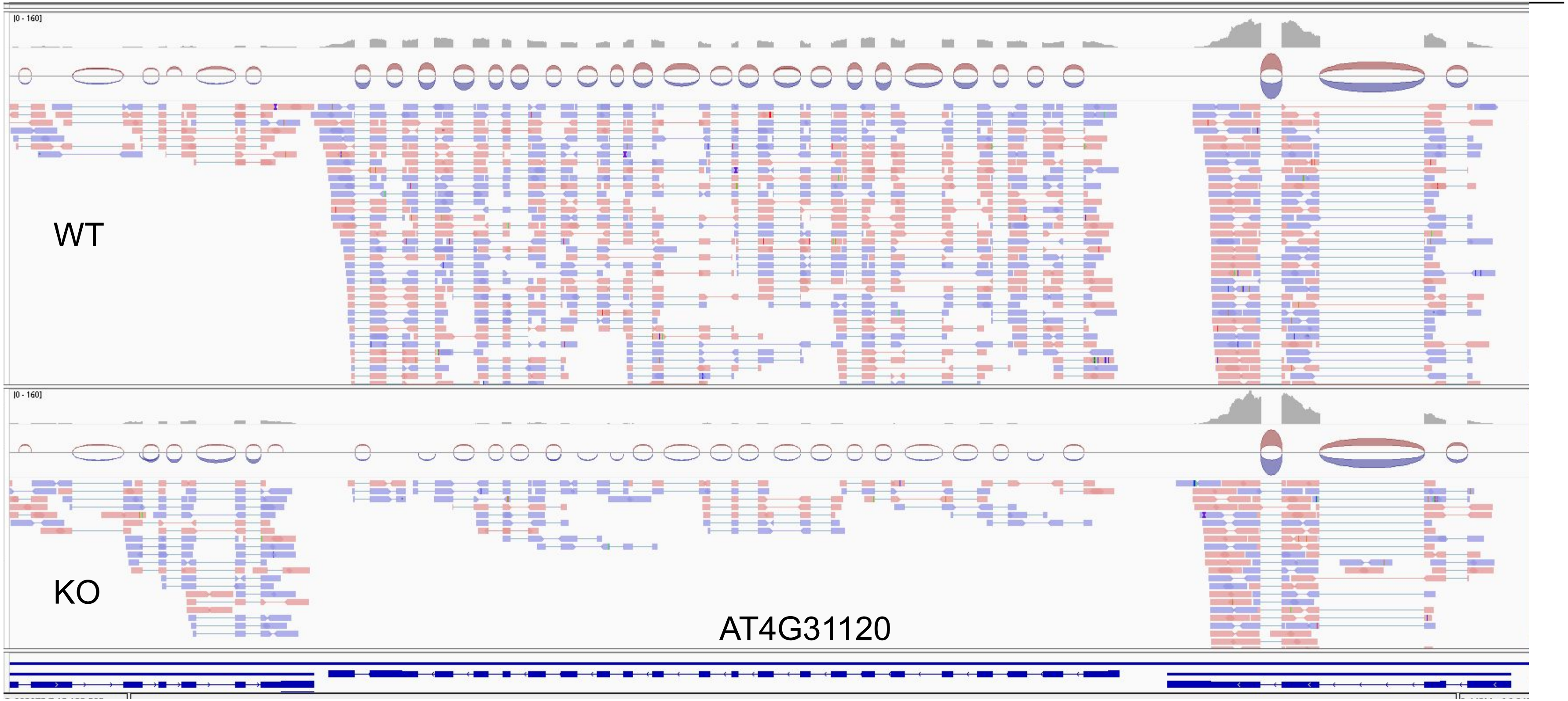


Which tools ?

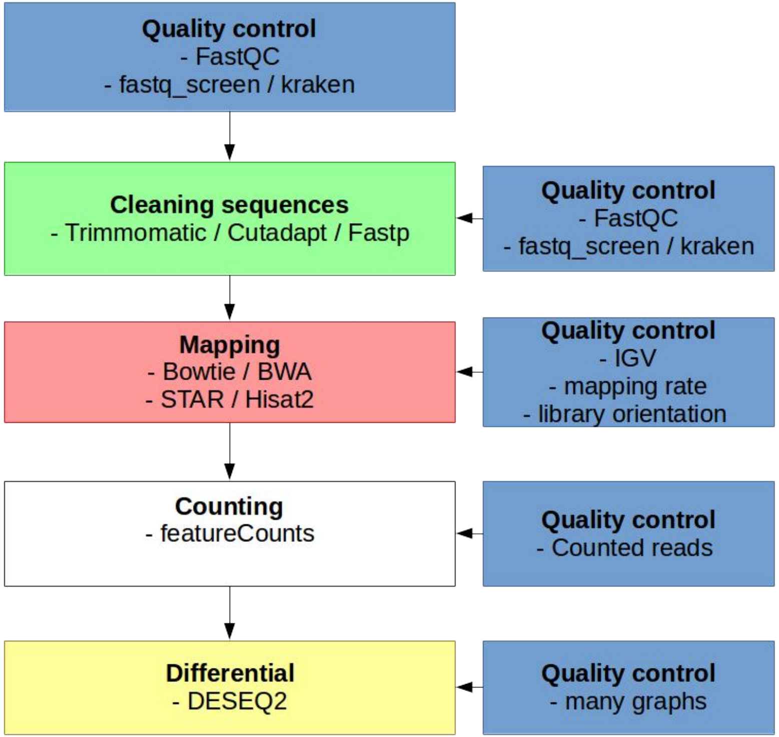
- ❖ Browser : IGV, Artemis, UCSC Genome browser, SeqMonk…
- ❖ Snapshots : Deeptools, ngs.plot,...

Visualize alignments

Go to AT4G31120 (AtPRM5)



Pipeline: toolS



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

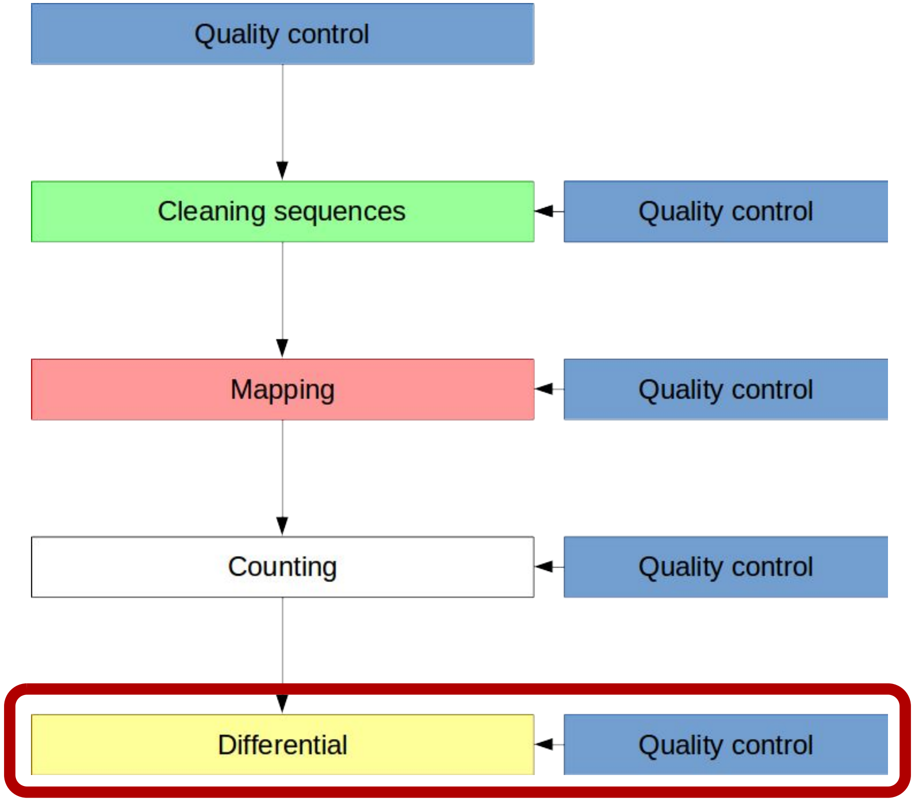
Counting
- featureCounts

Quality control
- Counted reads

Differential
- DESEQ2

Quality control
- many graphs

Next: Differential Gene Expression



02

Statistics

Experimental design,
Exploratory data analysis

03

Statistics

Normalization, modelisation
and troubleshooting

04

Practice

Differential analysis
with SARTools

The End