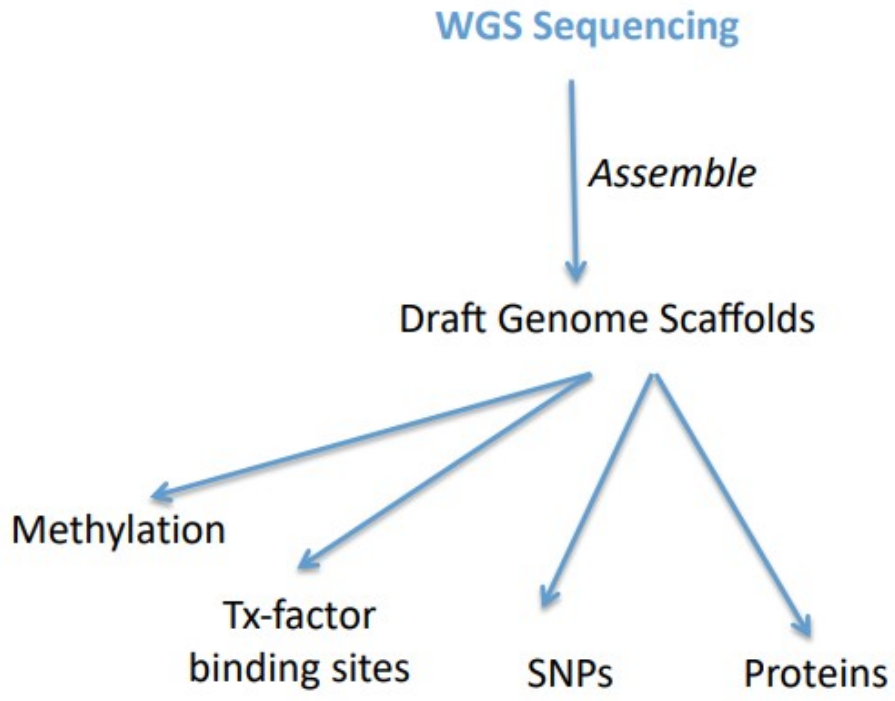
# Initiation aux analyses de données transcriptomiques

**Ouagadougou - Burkina Faso**
**7 au 11 Octobre 2018**

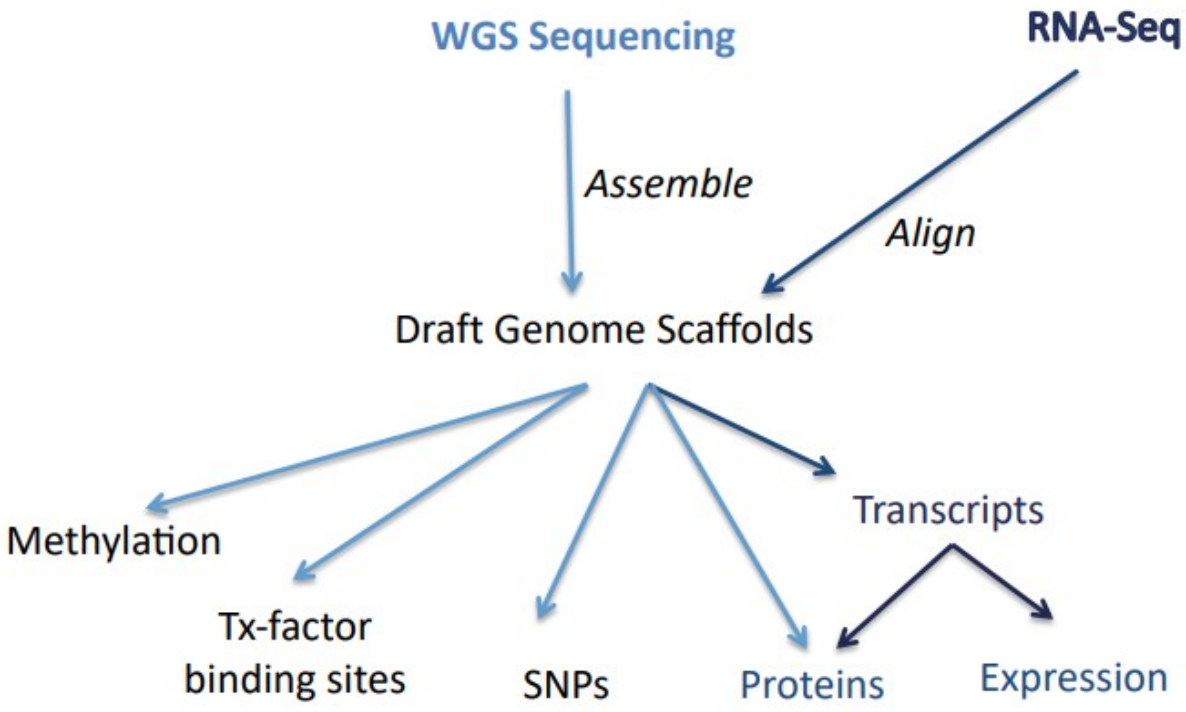
www.southgreen.fr

https://southgreenplatform.github.io/trainings

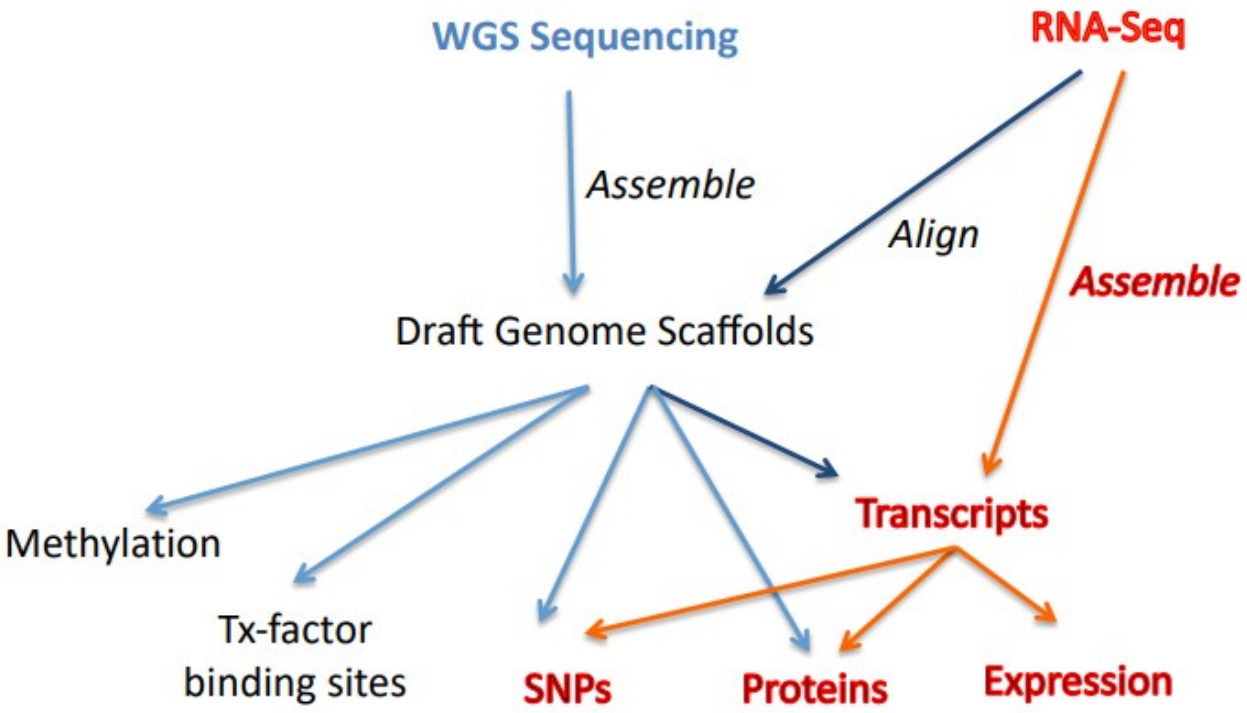
# A paradigm for transcriptome research



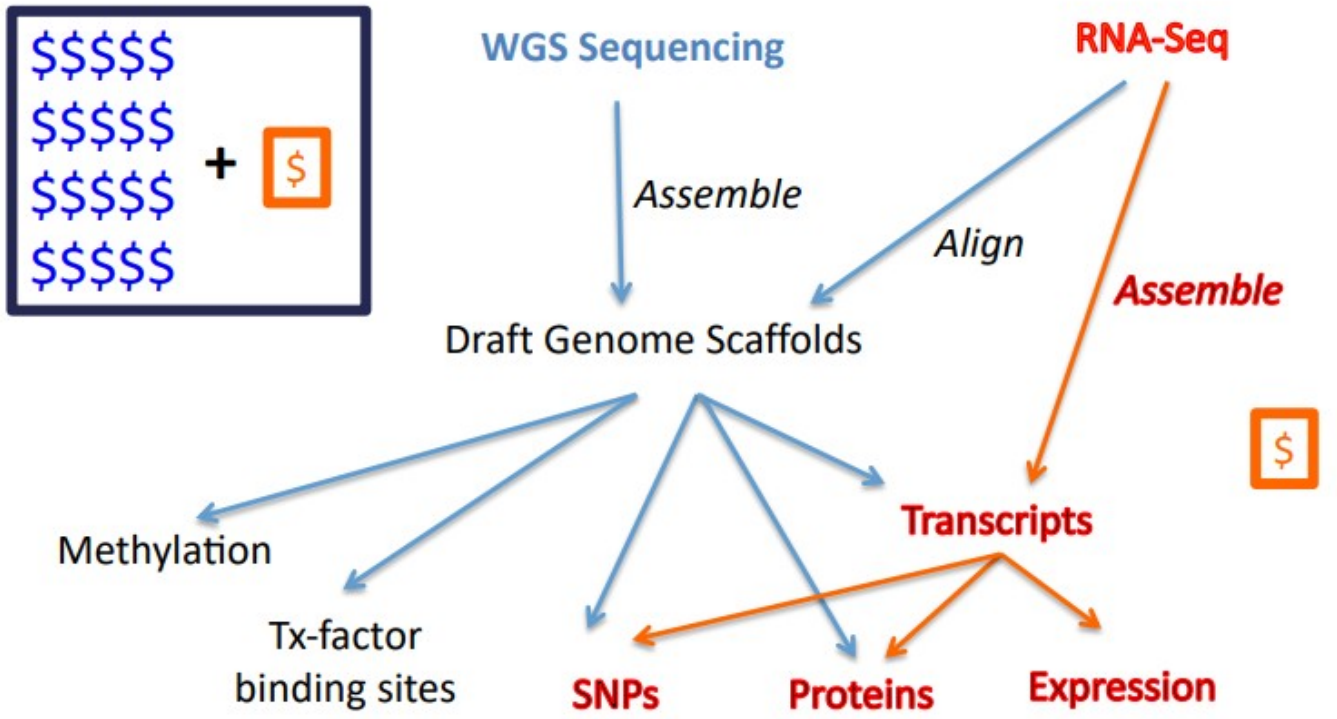
# A paradigm for transcriptome research



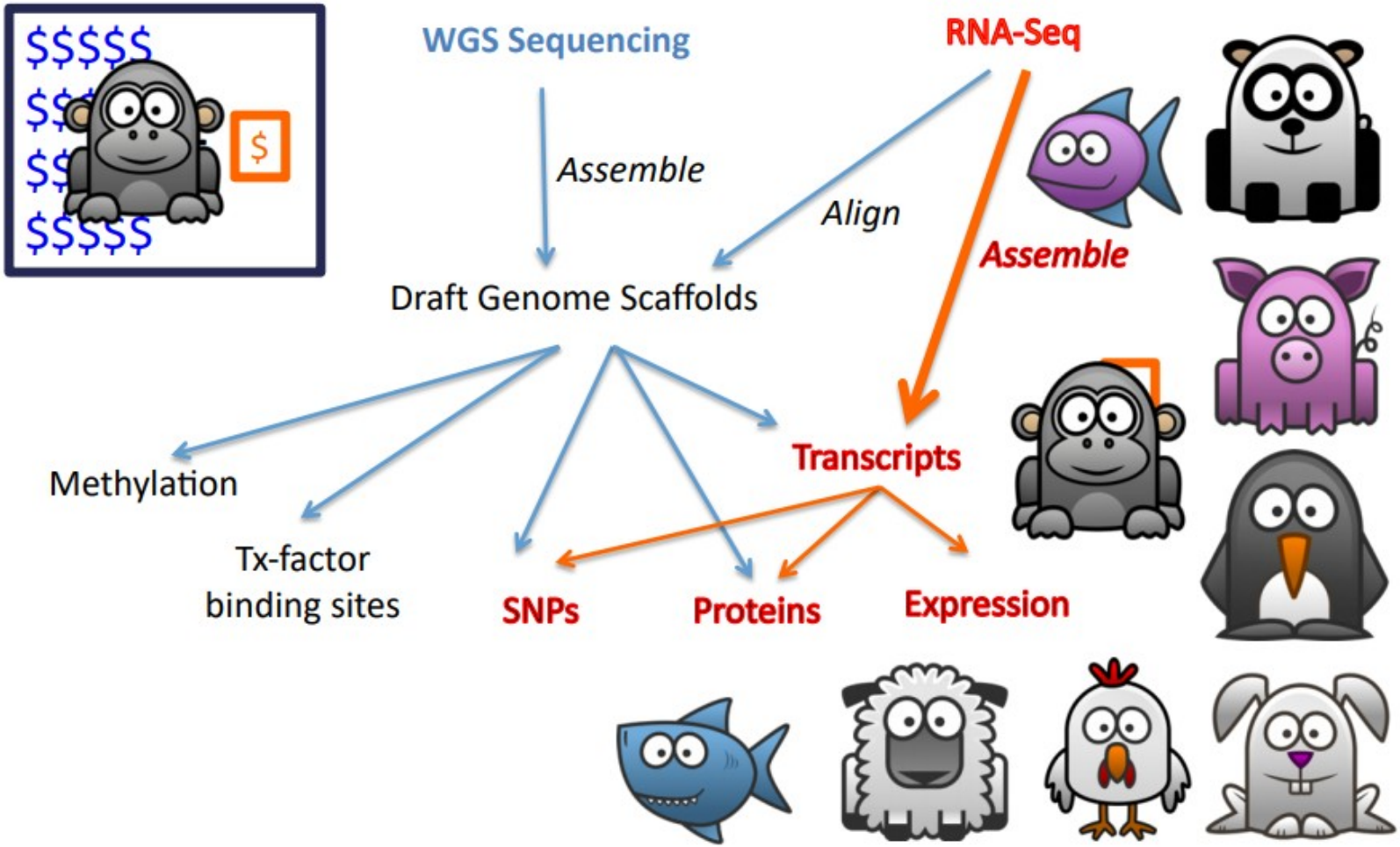
# A paradigm for transcriptome research



# A paradigm for transcriptome research



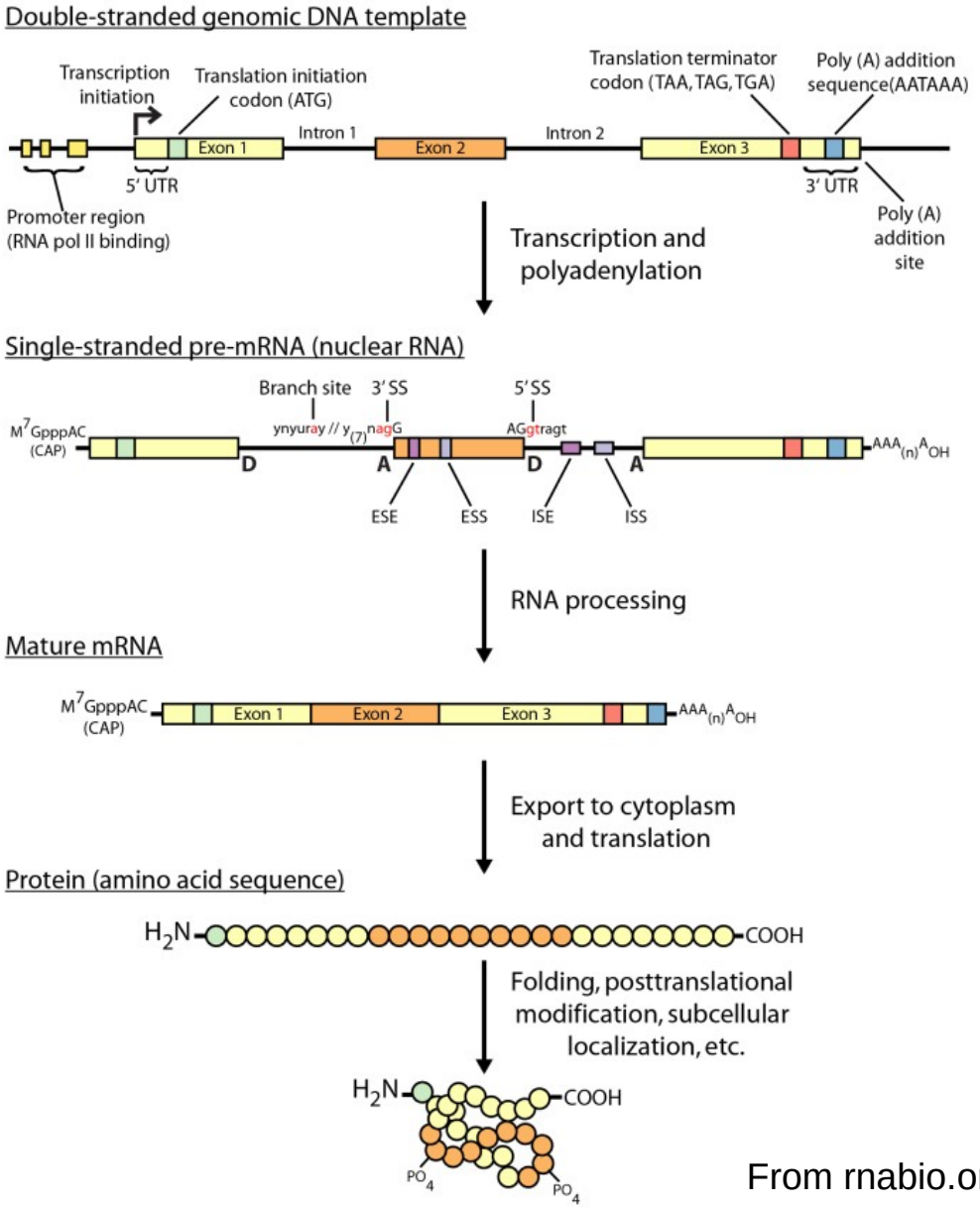
# A paradigm for transcriptome research




From SG-ABiMS RNAseq Training Sept 2019

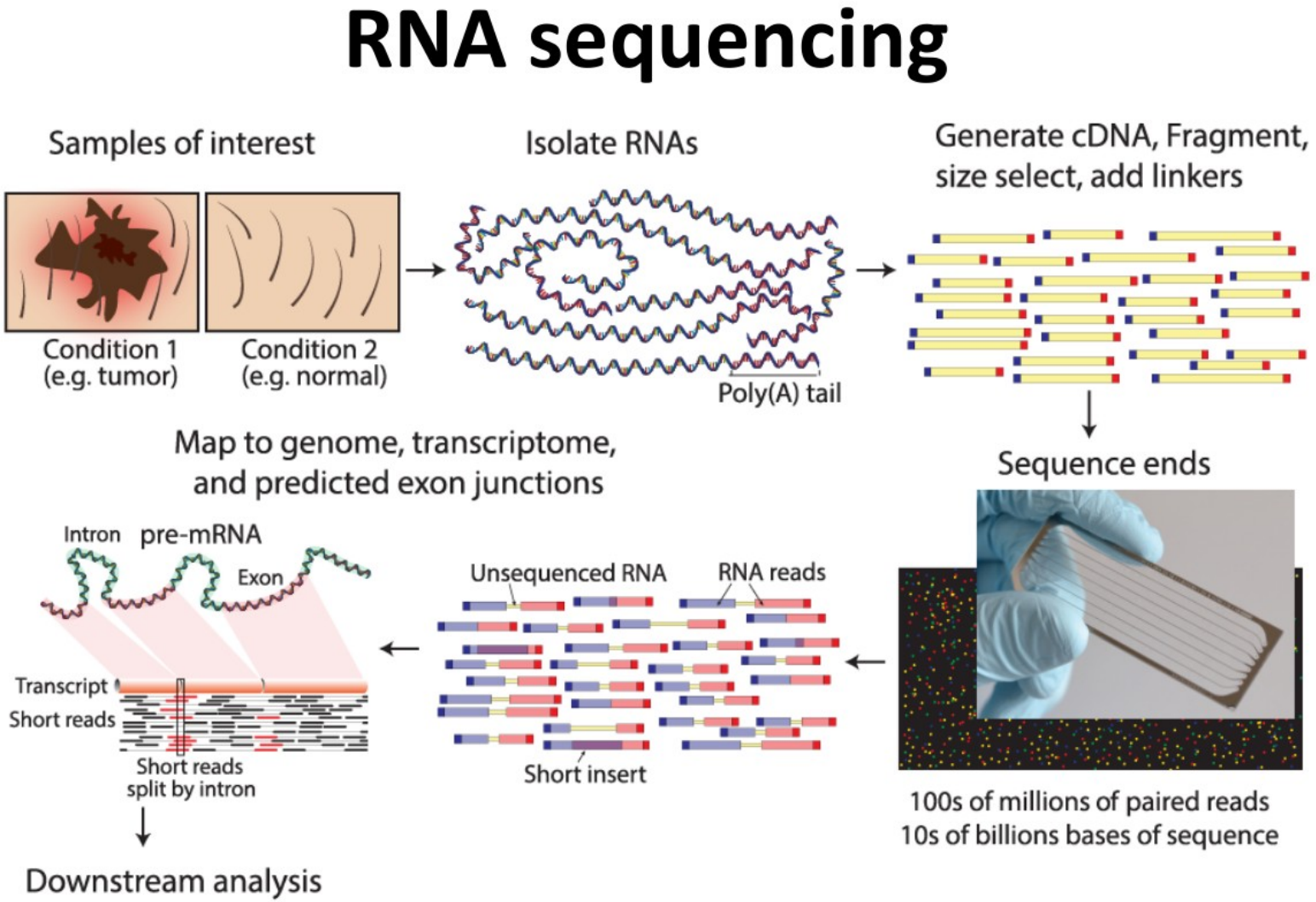
# Pourquoi faire du RNA-Seq

## Gene expression




From rnabio.org 2019

# Pourquoi faire du RNA-Seq




From rnabio.org 2019

# Pourquoi faire du RNA-Seq

**L'accès aux séquences d'ARN permet de :**

- Annoter un génome
- Réaliser un catalogue de gènes exprimés
- Identifier des nouveaux gènes
- identifier des transcripts alternatifs
- Quantifier l'expression des gènes et comparer entre différentes conditions expérimentales
- Identifier des petits ARNs (Regulation de l'expression, silencing …)

**Le choix technologique (déplétion/enrichissement, démultiplexage, séquençage directionnel) dépendra de la question biologique**

# Construction des librairies

## There are many RNA-seq library construction strategies

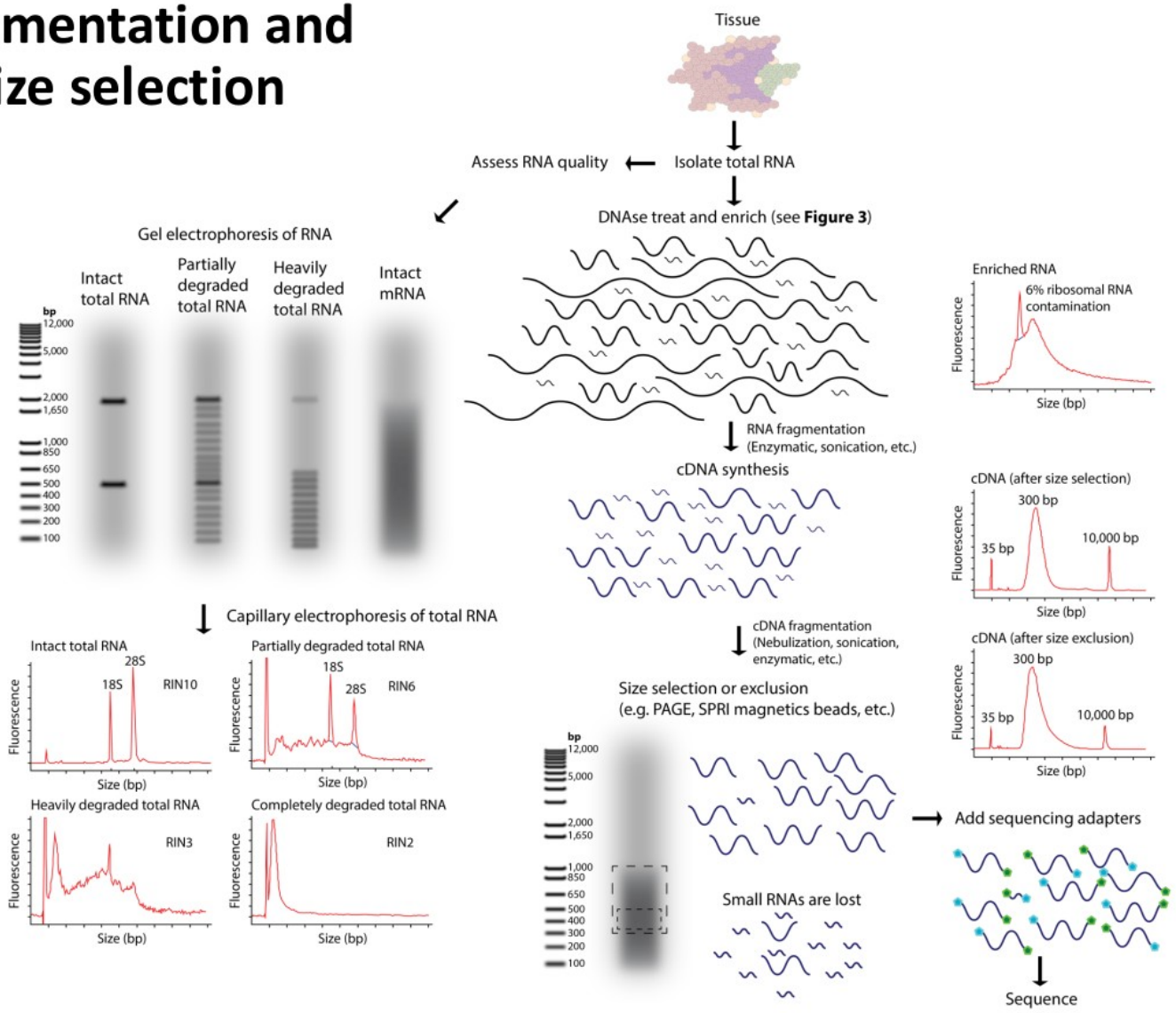
- Total RNA versus polyA+ RNA?
- Ribo-reduction?
- Size selection (before and/or after cDNA synthesis)
  - Small RNAs (microRNAs) vs. large RNAs?
  - A narrow fragment size distribution vs. a broad one?
- Linear amplification?
- Stranded vs. un-stranded libraries
- Exome captured vs. un-captured
- Library normalization?

- These details can affect analysis strategy
  - Especially comparisons between libraries

From rnabio.org 2019

# Construction des librairies

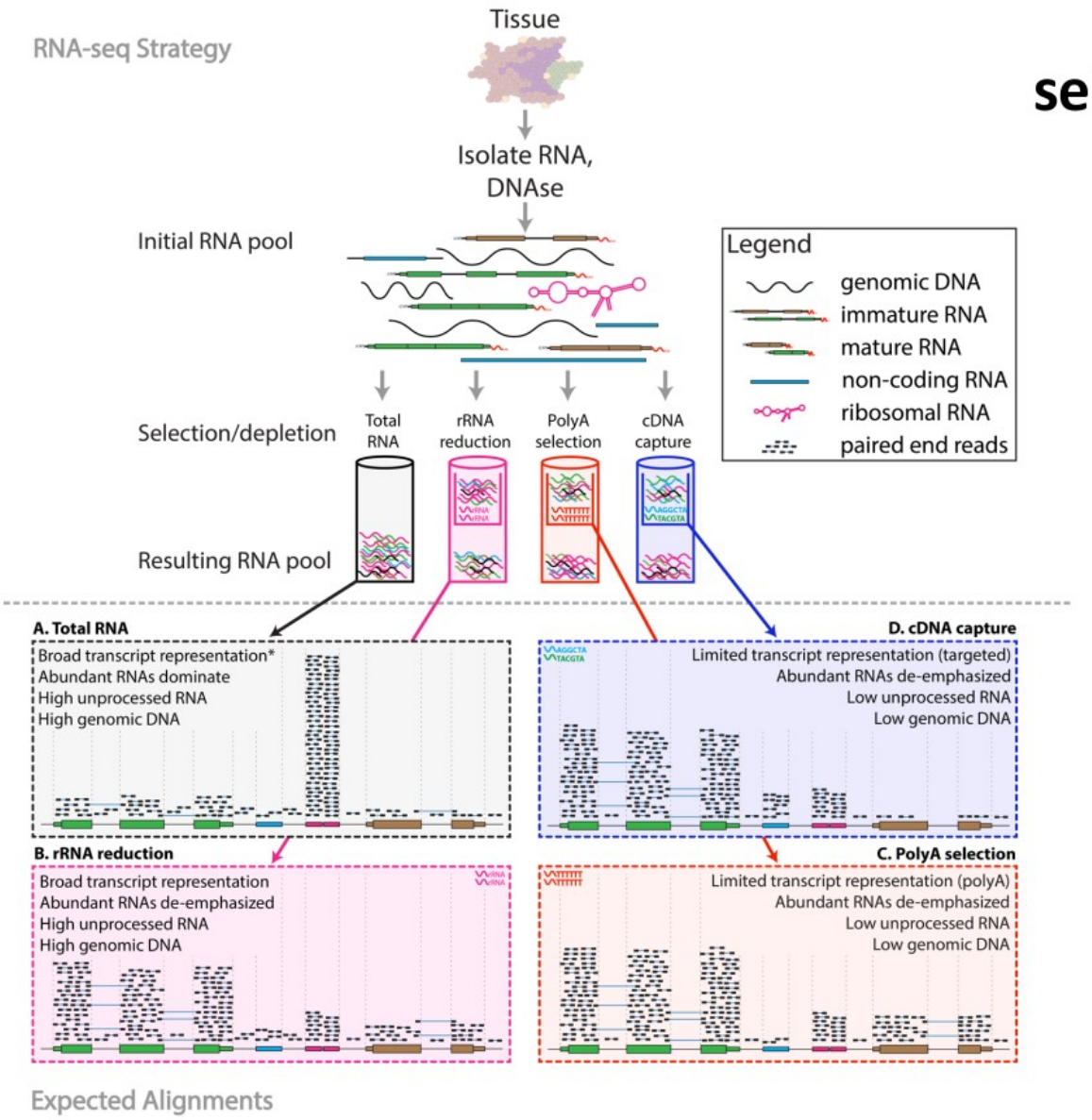
## Fragmentation and size selection




From rnabio.org 2019

# Construction des librairies

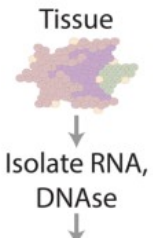
## RNA sequence selection/depletion




From rnabio.org 2019

# Construction des librairies

RNA-seq Strategy




## RNA sequence selection/depletion

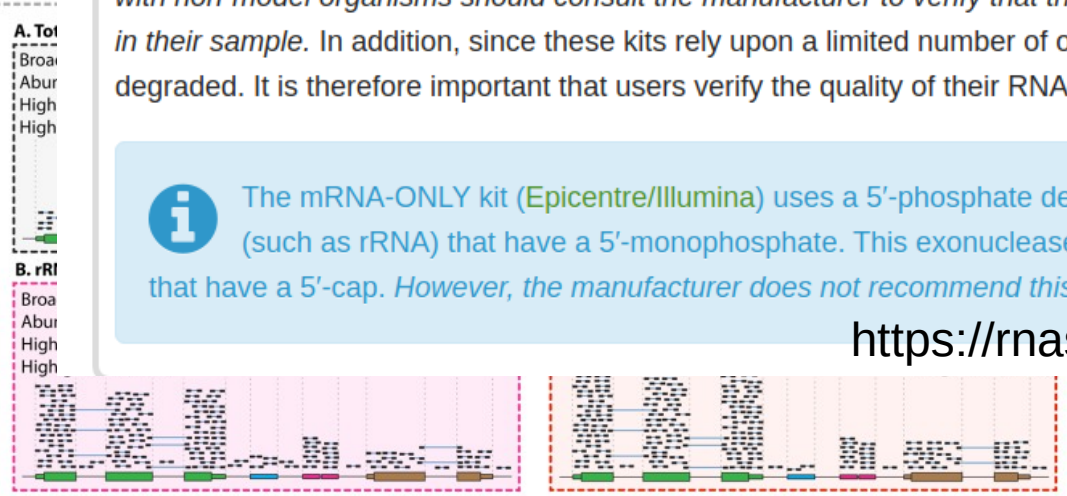
More on rRNA depletion by hybridization

The oligo/rRNA complex is removed from the solution via binding to beads. Different kits use different technologies to capture the bound complex. The capture oligos in the Ribominus (Invitrogen/Life Technologies) and Ribo-Zero (Epicentre/Illumina) kits have a biotin tag, that can be captured using streptavidin coated magnetic beads. The GeneRead kit (Qiagen) uses antibodies that specifically recognize the rRNA/oligo complex.

All of these kits are capable of removing the majority of rRNA from a total RNA sample. *However, users that are working with non-model organisms should consult the manufacturer to verify that the capture oligos are compatible with the rRNA in their sample.* In addition, since these kits rely upon a limited number of oligos they only work well if the input RNA is not degraded. It is therefore important that users verify the quality of their RNA before proceeding.

The mRNA-ONLY kit (Epicentre/Illumina) uses a 5′-phosphate dependent exonuclease to degrade RNAs (such as rRNA) that have a 5′-monophosphate. This exonuclease will not degrade intact and mature mRNAs that have a 5′-cap. *However, the manufacturer does not recommend this kit for RNA-seq.*

https://rnaseq.uoregon.edu/#library-prep



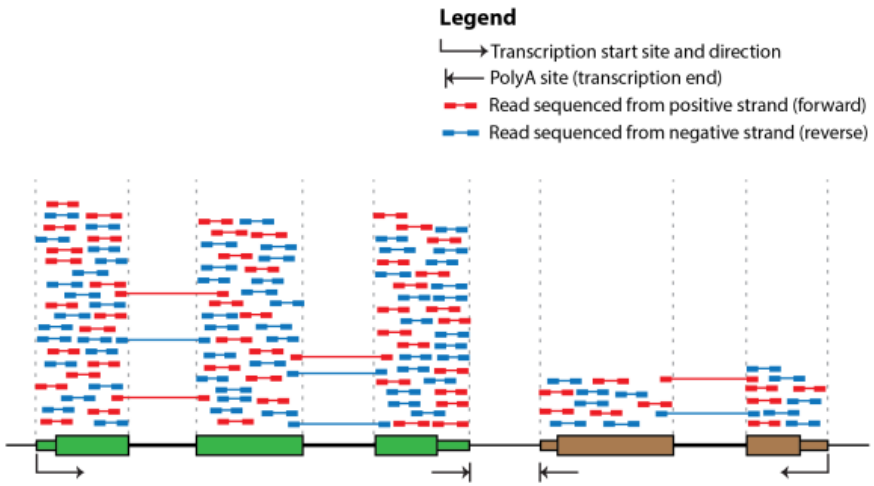
Expected Alignments

From rnabio.org 2019

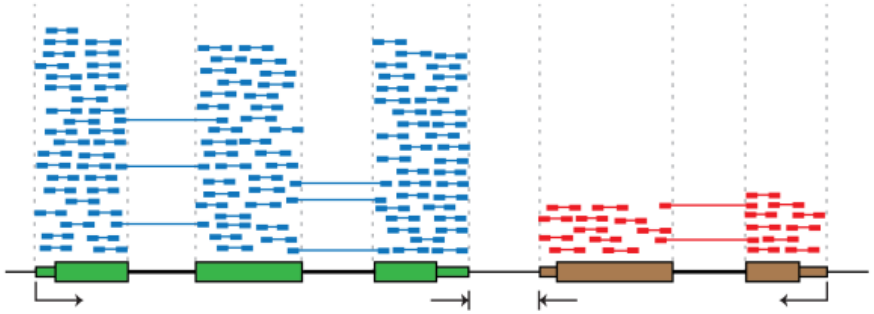
# Construction des librairies

## Stranded vs. unstranded

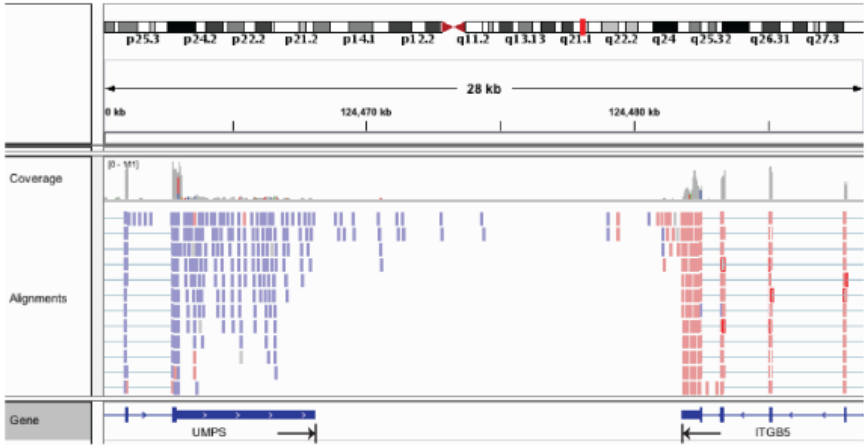
**A. Depiction of cDNA fragments from an unstranded library**




**B. Depiction of cDNA fragments from an stranded library**



**C. Viewing strand of aligned reads in IGV**




From rnabio.org 2019

# Design expérimental

# Experimental design

Build an experimental design

- to control the variability during the experiment in order to address the biological question:
  1. What is the biological question?
  2. How to estimate the associated biological variabilities?
  3. How to control the technical variabilities (day, lane, run, etc.)?

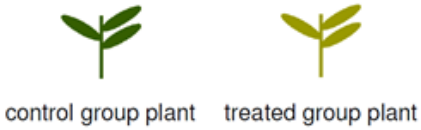
Biological or technical uncontrolled effects could:

- Hide/cancel the biological effect of interest
- Wrongly increase the biological effect of interest

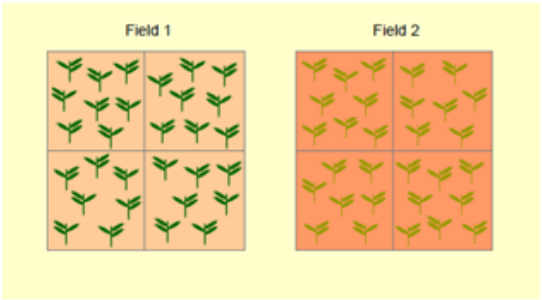
From rnabio.org 2019

# Experimental design

Basic experiment : trouver les différences entre condition contrôle/traitée

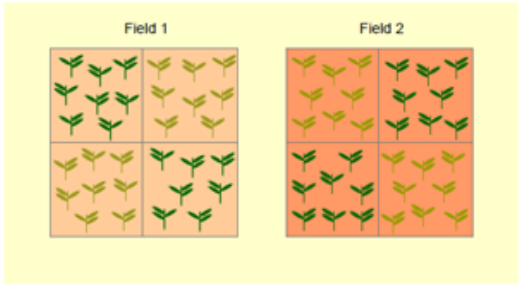



Mauvais plan expérimental : les plantes traitées sont dans un champs et les contrôles dans un autre.



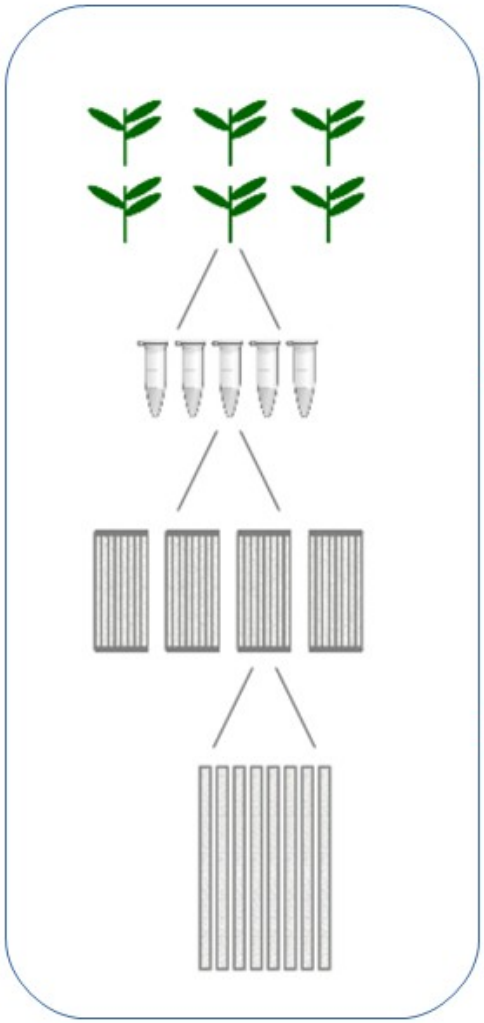

Pas de possibilité de différencier l'effet champs de l'effet traitement

Bon plan expérimental : la moitié des plantes traitées poussent avec un contrôle dans un même champs et l'autre moitié dans un autre champs




Possibilité de différencier l'effet champs de l'effet traitement.

# Experimental design



collect

1 – Variations biologiques : variations individuelles dues aux effets génétiques, de l'environnement

Sample préparation

2 – Variations techniques : effet de la préparation des librairies

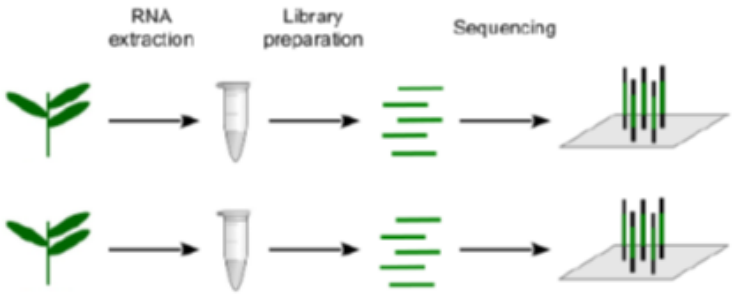
cDNA on lane of flowcell

3 – Variation techniques : effet des lane et des flowcell

Effet lane < Effet Flowcell <Effet de la préparation de la librairie < < Effet biologique

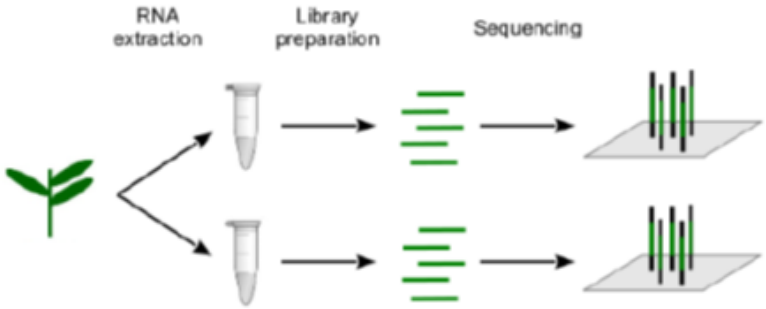
# Experimental design

**Réplicat biologique** : Différents échantillons biologiques, répétés plusieurs fois séparément (au moins 3 fois).




**Réplicat Technique** : Même matériel biologique, répété plusieurs fois indépendamment des étapes techniques.

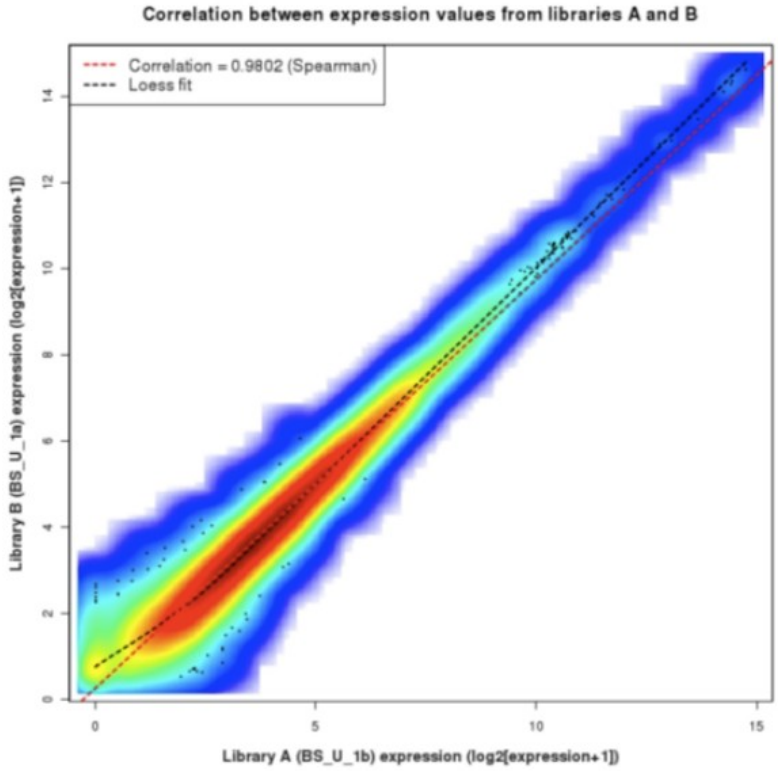
- Plusieurs extractions d'une même échantillon
- Plusieurs séquençages d'une même librairie



# Experimental design

## Replicates

- Technical Replicate
  - Multiple instances of sequence generation
    - Flow Cells, Lanes, Indexes
- Biological Replicate
  - Multiple isolations of cells showing the same phenotype, stage or other experimental condition
  - Some example concerns/challenges:
    - Environmental Factors, Growth Conditions, Time
  - Correlation Coefficient 0.92-0.98




From rnabio.org 2019

# Sequencing depth

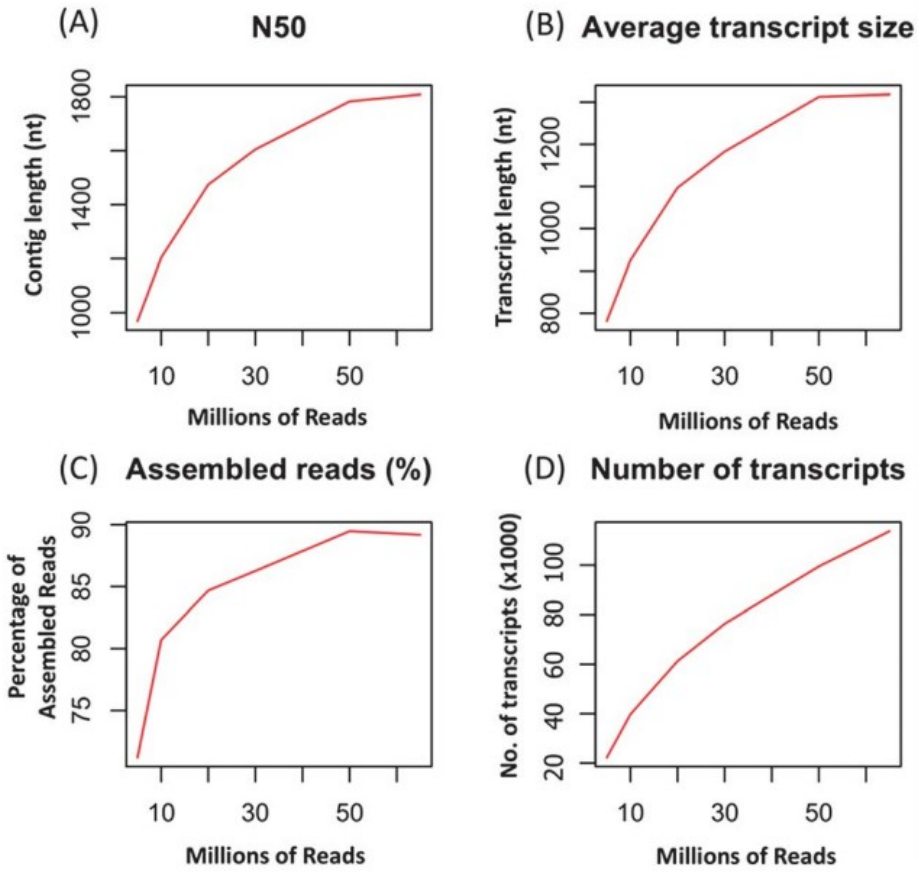



**Fig. 6** Effect of sequencing depth on a transcriptome assembly. Four Paired-End assemblies using 5, 10, 20, 30, 50 and 65 million reads were generated using Oases.[37] The N50 contig size (A), average transcript size (B), percentage of reads used in the assembly (C), and number of transcripts (D) *versus* number of reads used in the assembly are shown.

~ 30 millions de reads par réplique par échantillon

# Sequencing depth

On a besoin de 50 millions de reads par échantillon

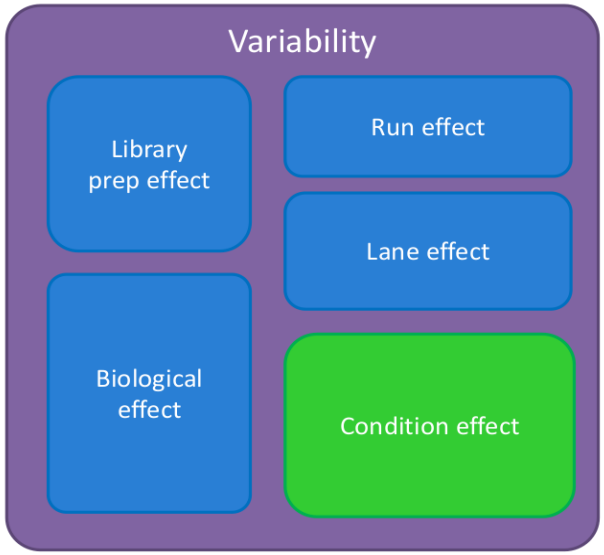
Une lane de Hiseq4000 est capable de générer 2.5 billions des reads (25 000 000 000 000 )

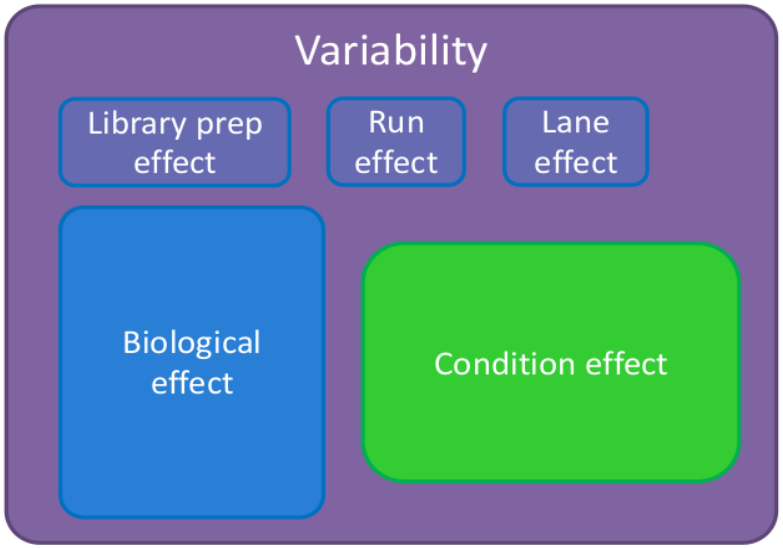
50 échantillons peuvent etre séquencées dans una lane => ~ 16 échantillons X3 repliques biologiques

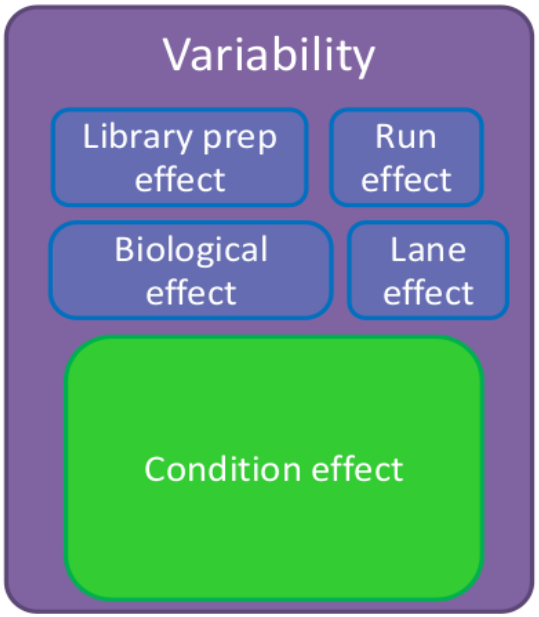
Table 1: Performance Parameters of the HiSeq 3000/4000 Systems.[a]

| | HiSeq 3000 System | HiSeq 4000 System |
|---|---|---|
| Number of Flow Cells per Run | 1 | 1 or 2 |
| Output[b] | | |
| 1 × 50 bp | 105–125 Gb | 210–250 Gb |
| 2 × 75 bp | 325–375 Gb | 650–750 Gb |
| 2 × 150 bp | 650–750 Gb | 1300–1500 Gb |
| Clusters Passing Filter (Single Reads) | 2.1–2.5 billion | 4.3–5 billion |
| Quality Scores | | |
| 2 × 50 bp | ≥ 85% of bases above Q30 | |
| 2 × 75 bp | ≥ 80% of bases above Q30 | |
| 2 × 150 bp | ≥ 75% of bases above Q30 | |
| Daily Throughput | > 200 Gb | > 400 Gb |
| Run Time | < 1–3.5 days | < 1–3.5 days |
| Human Genomes per Run[c] | up to 6 | up to 12 |
| Exomes per Run[d] | up to 48 | up to 96 |
| Transcriptomes per Run[e] | up to 50 | up to 100 |

# Variability




Technical replicates
+ normalization
+ statistics

+

Biological replicates
+ statitics

From Abims RNAseq Training oct 2018

# Experimental design

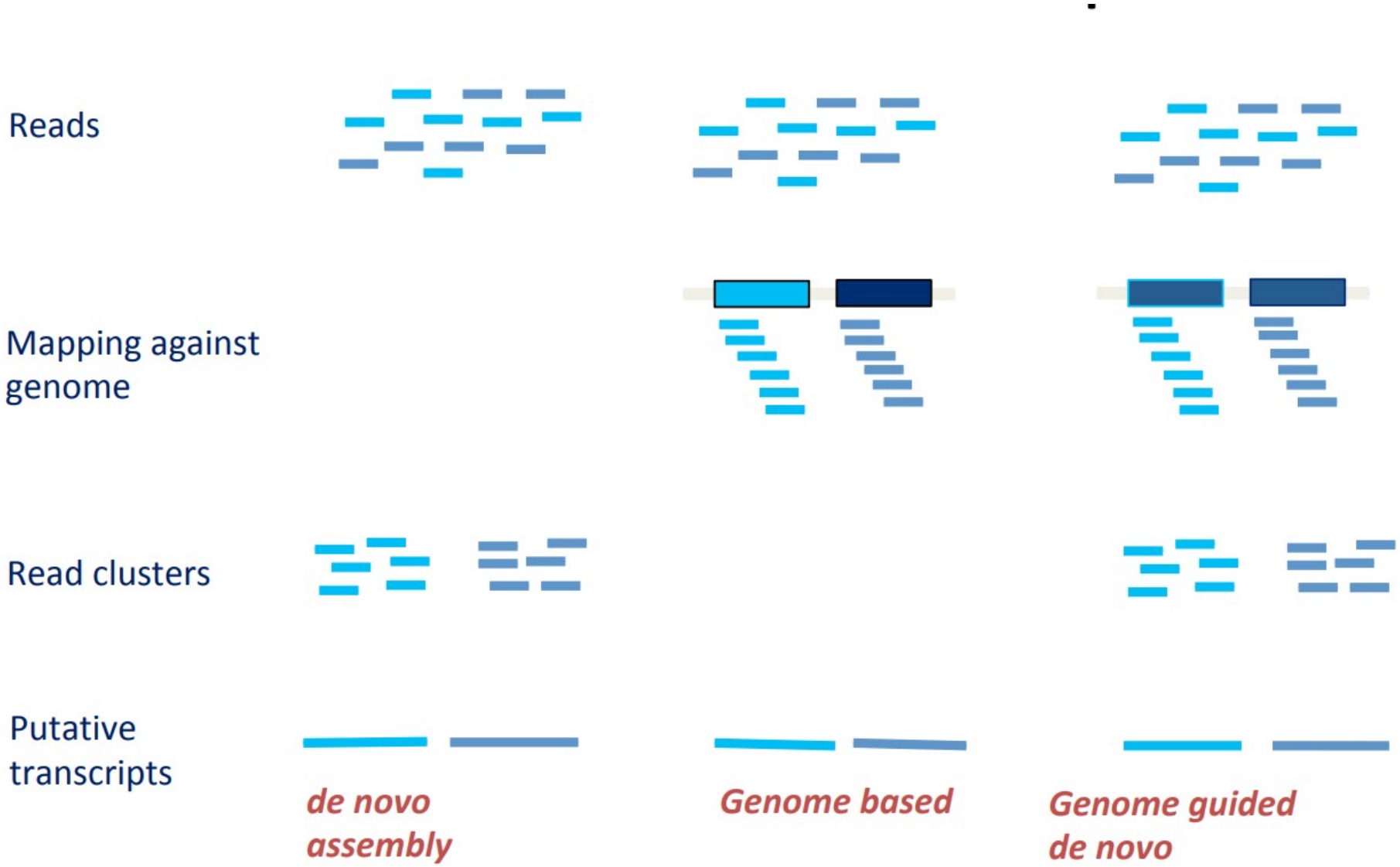
It's up to you! (Haas et al., 2012, Liu Y. et al 2013)

- Detection of differential transcripts:
  - (+) biological replicates
- Construction / transcriptome annotation:
  - (+) depth & (+) conditions
- Search variants:
  - (+) biological replicates & (+) depth

From rnabio.org 2019

# Bioinformatic strategies

# Des reads aux transcripts




From http://www.bioinformatics.nl/courses/RNAseq/1c%20Assembly.pdf

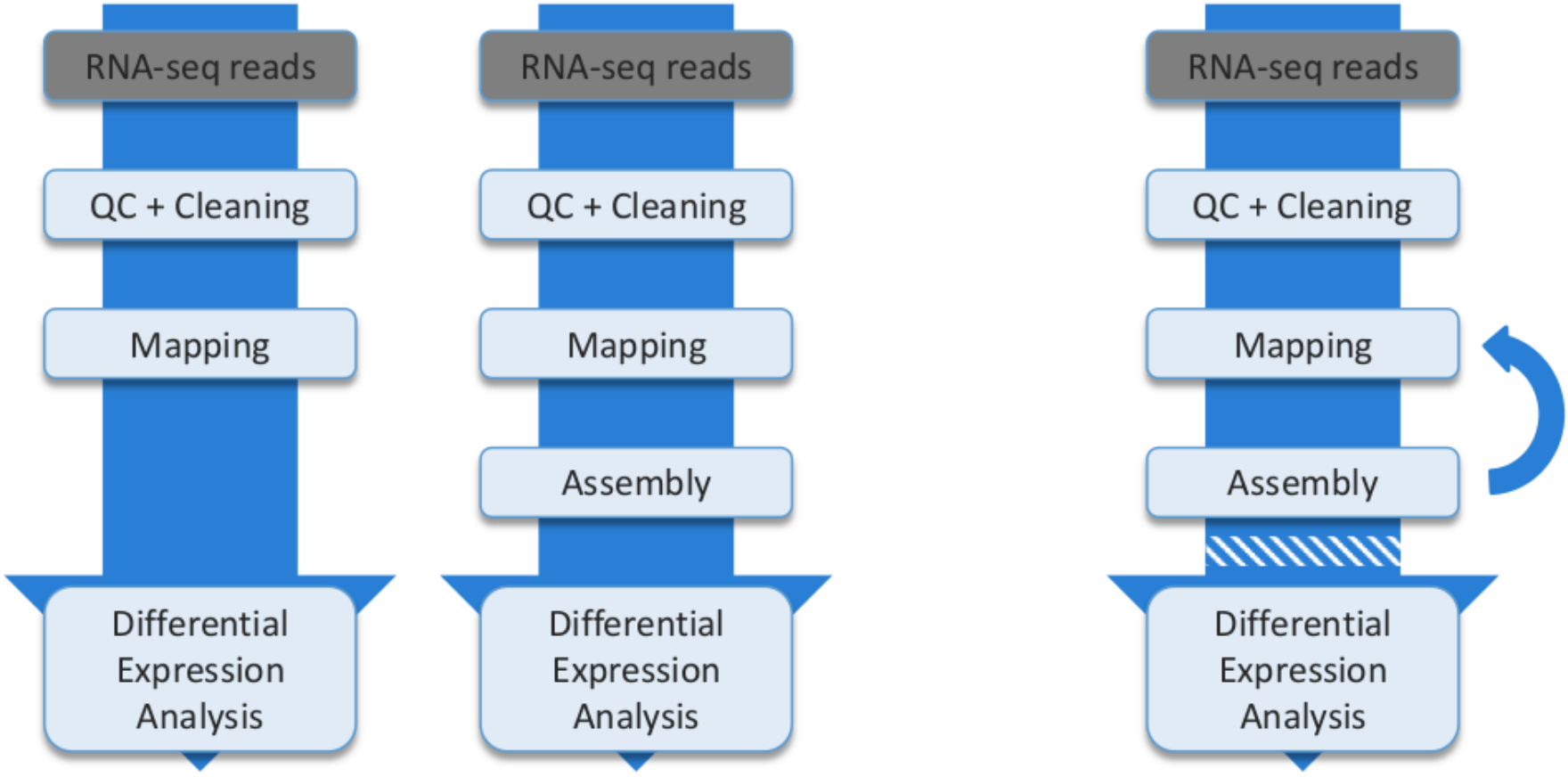
# Vision Global of RNAseq Analysis

- Reference genome
- Reference transcriptome

- Reference genome
- No reference transcriptome

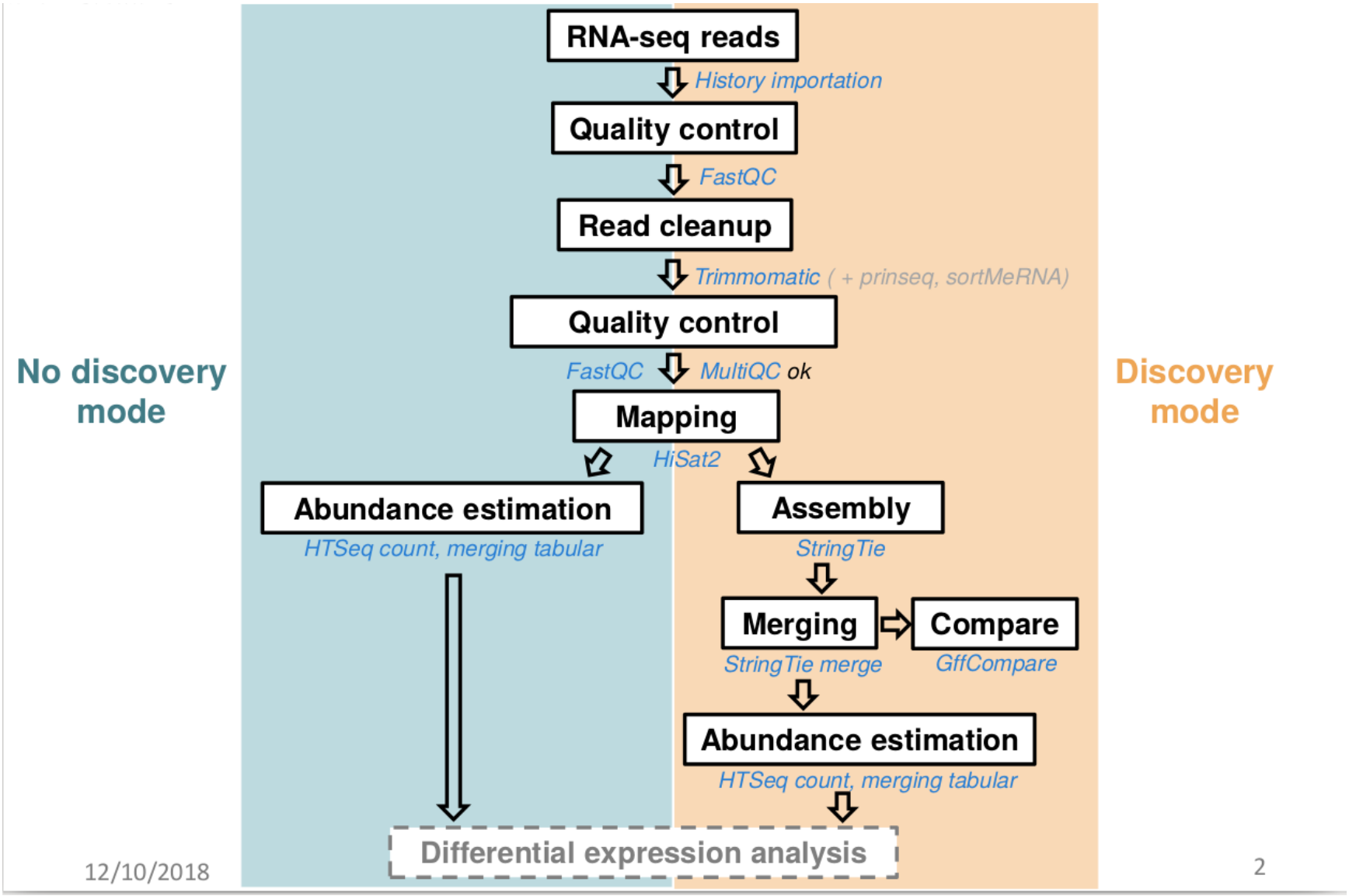
Non discovery mode

Discovery mode




From Abims RNAseq Training oct 2018

# Vision Global of RNAseq Analysis



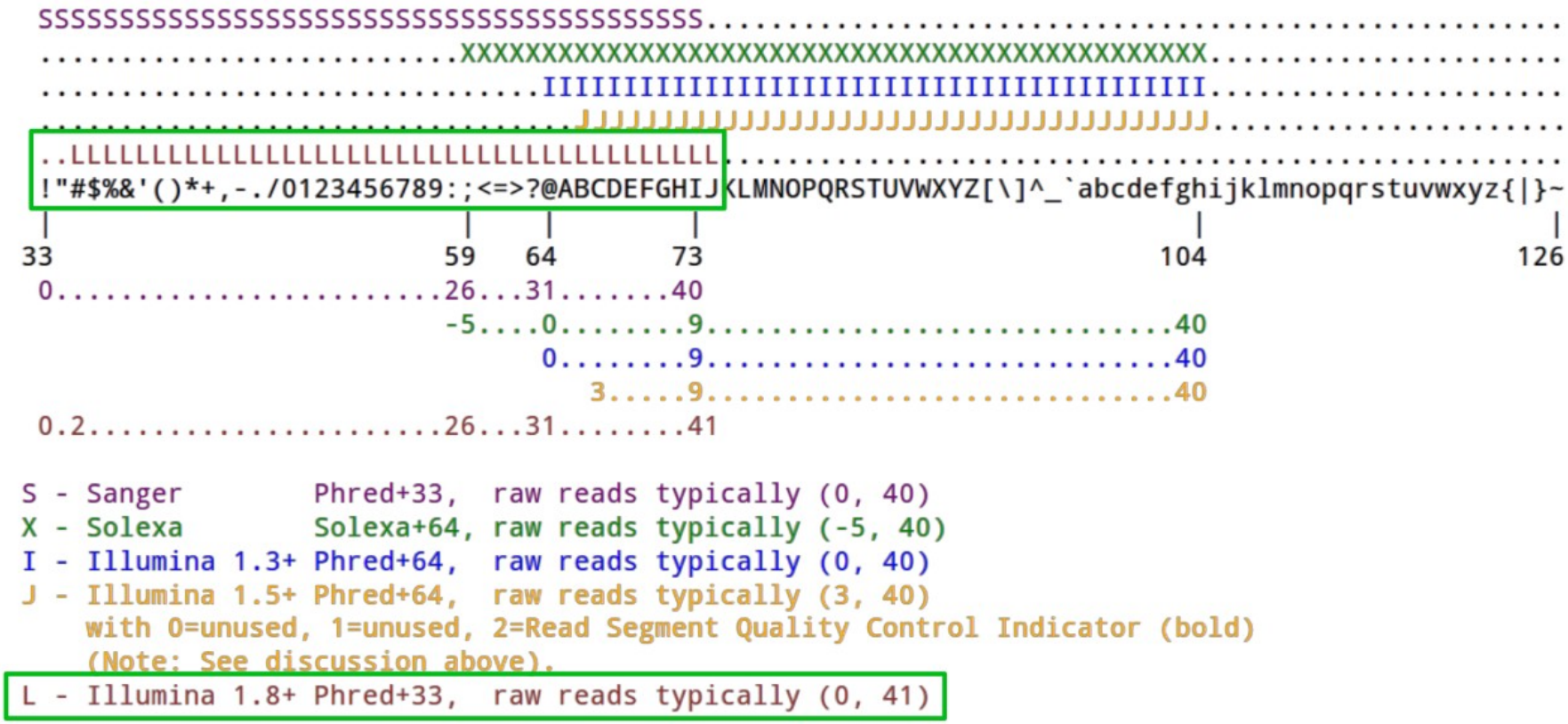

From Abims RNAseq Training oct 2018

# File formats to RNAseq Analysis

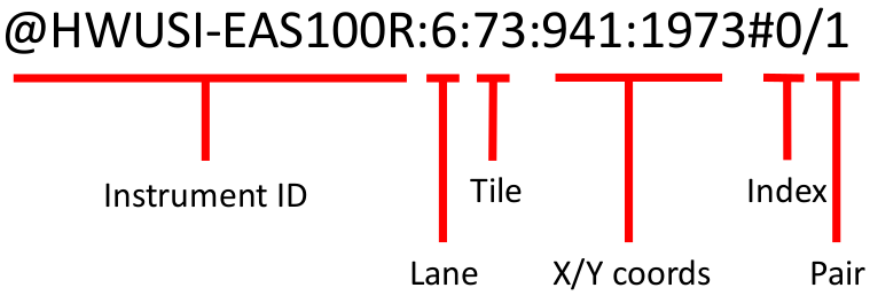
# Format : FASTQ

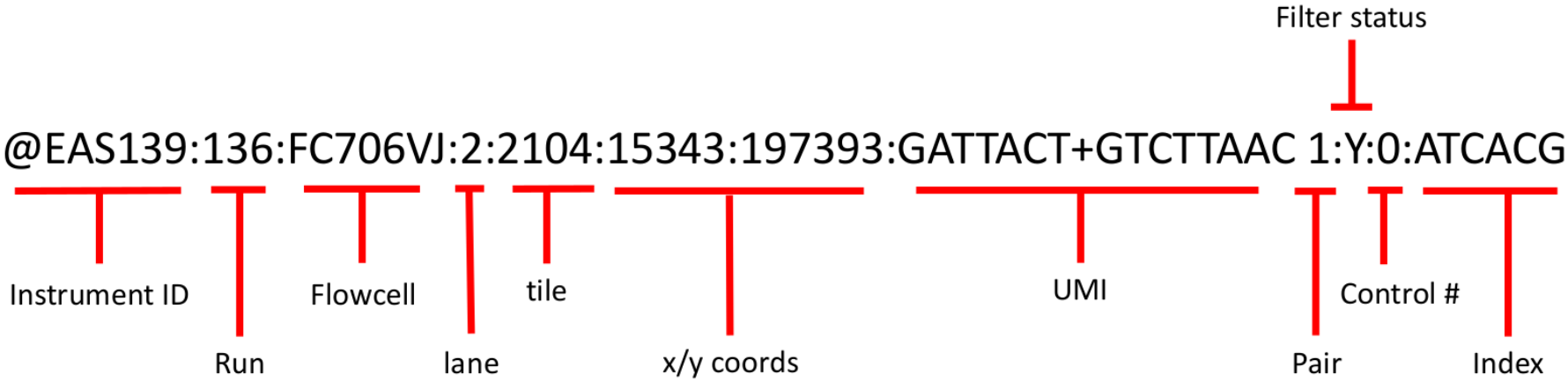

From E.Corre Abims RNAseq Training sept 2019

# Phred score


SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
33 59 64 73 104 126
0.....26...31.....40
-5....0.....9.....40
0.....9.....40
3.....9.....40
0.2.....26...31.....41
S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

# Read naming conventions



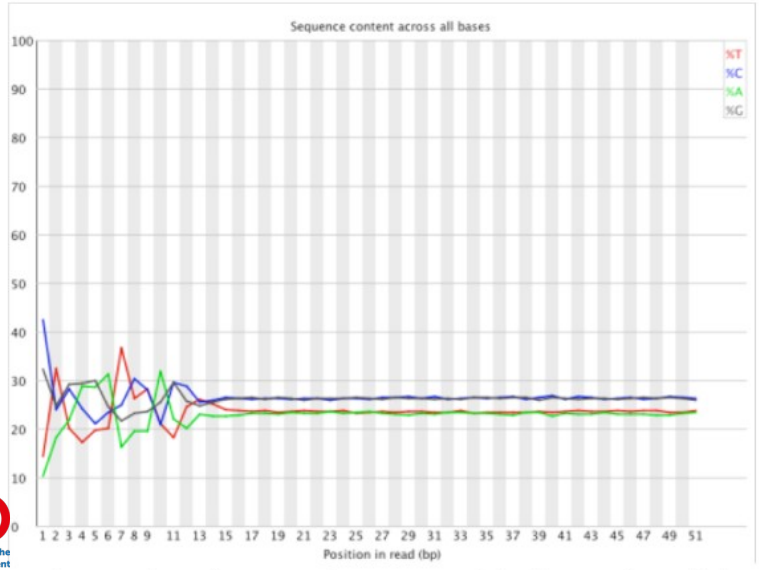

From rnabio.org 2019

# Cleaning

# Sequencing biases

- Unknown nucleotides (Ns)
- Bad quality nucleotides
- Hexamers biases (random priming) ? (Illumina. Now corrected ?)

- Why do we need to correct those ?
  - To remove a lot of sequencing errors (detrimental to the vast majority of assemblers)
  - Because most de-bruijn graph based assemblers can't handle unknown nucleotides




From E.Corre Abims RNAseq Training sept 2019

# Séquences d'adaptateurs et d'amorces

- Can be found in 3' end if insert size is too short
- Why do we need to remove those ?
  - Because they can lead to "bridges" (links) between unrelated sequences (eg. 2 genes) and generate chimeras

Normal case:
insert size > sequencing length

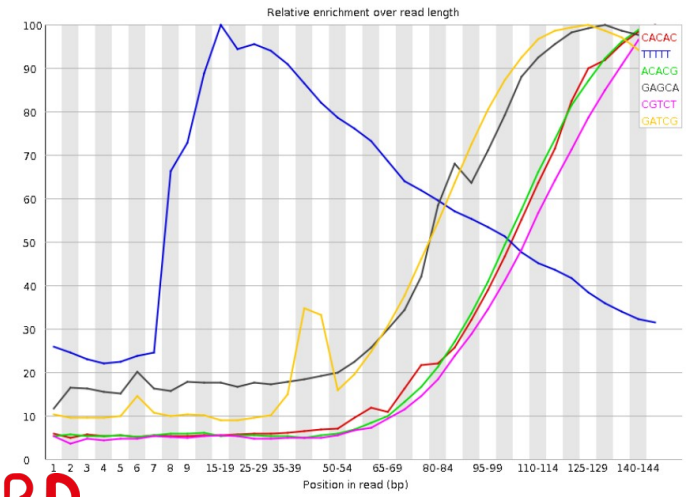
Abnormal case:
insert size < sequencing length

# Poly A tails , low complexity reads

- Some poly A/T tails can be left during library preparation
- Poly A/T or low complexity sequences can also lead to “bridges” between unrelated sequences and generate chimeras

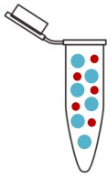
```
>
ACGTAGCTACTAGCTGACGATTCCCGTAGATCATCGGATAAAAAAAAAAAAAAAAAAAAAA
>
TTTTTTTTTTTTTTTTTTTTTTTTTTTTACTGCGTAGCACATGGCTATTATTTCGGCCATCAA
>
CGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG

TGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT
```



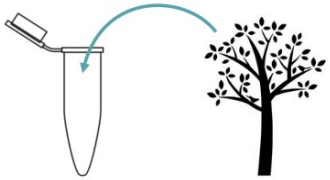
# Contaminations

- Most RNA-seq libraries comprise ribosomal RNA that you may want to remove
- Contaminations can also occur with foreign RNA/DNA (PhiX, Bacteria, …)



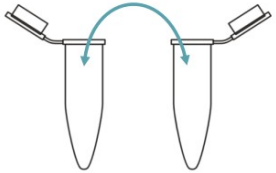
in-contamination

for ex. rRNA



third-party contamination

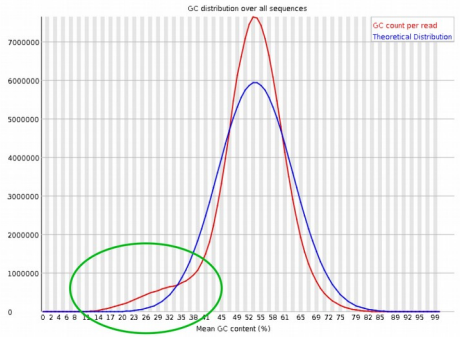
for ex. food - parasite



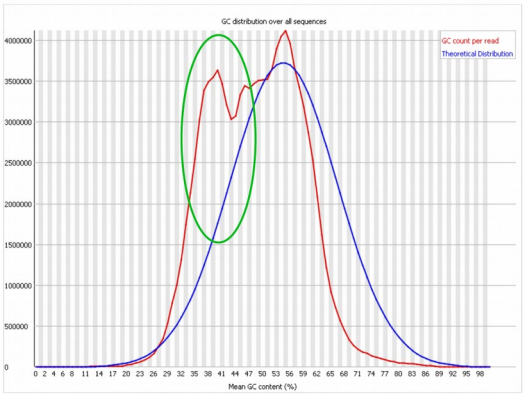
cross-contamination

for ex. experiment

- A contamination ?




**Can this be fixed ?** Maybe…

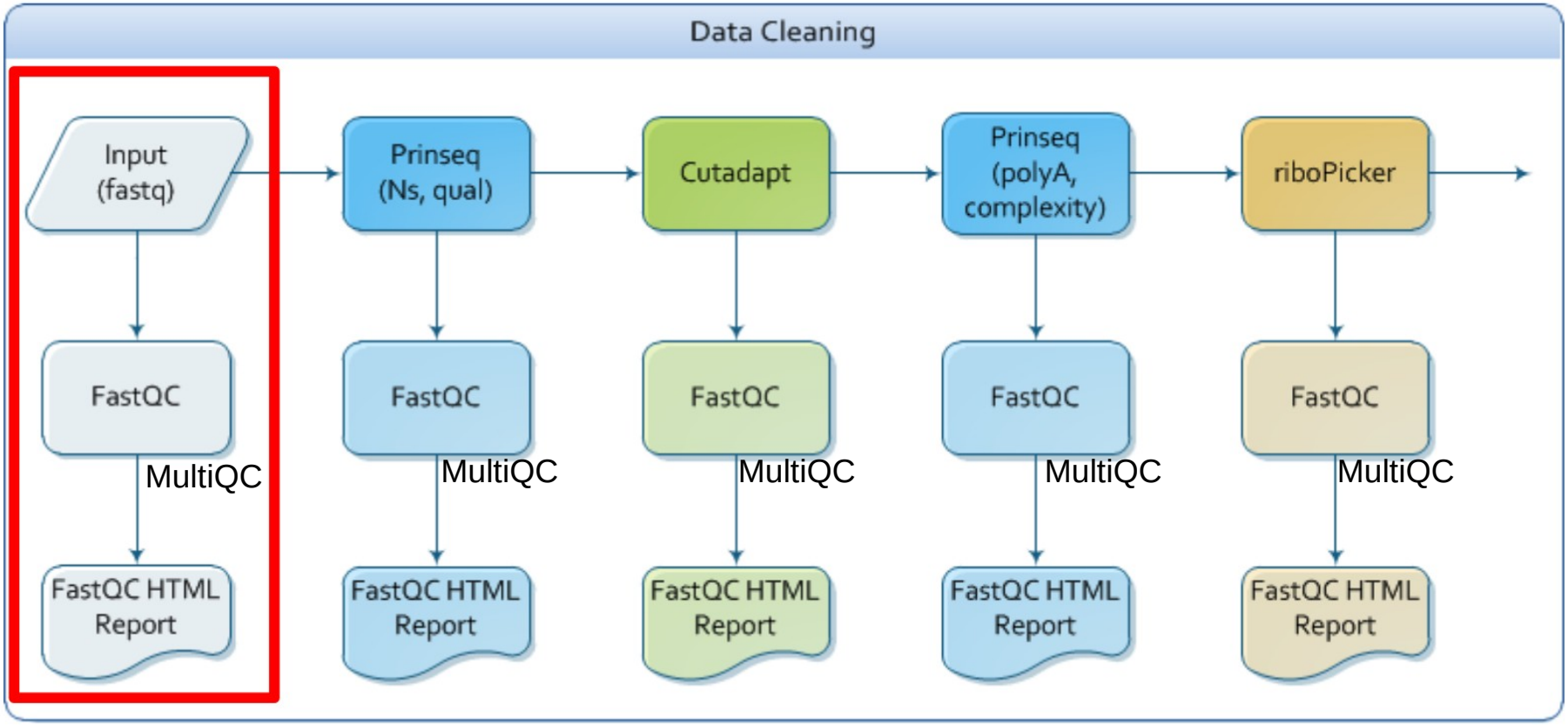



RNAmmer

# Vérification de la qualité des données NGS

| | Problème | Pourquoi les éliminer? | Outils |
|---|---|---|---|
| **Sequences biases** | Ns, mauvaise qualité des nucléotides, biases hexamères (random priming) | Pour éliminer des erreurs de sequencing.<br><br>Désastreux pour la plupart des assembleurs | PRINSEQ2<br>FASTX Toolkit<br>*Trimmomatic* |
| **Adaptors and primers** | Peuvent être trouvés dans le 3' final d'un insert très court | Des ponts entre séquences sans relation aucune: Chimères | *Trimmomatic*, cutadapt, far, btrim, SeqTrim, TagCleaner, solexaQA |
| **Poly A/T tails, low complexity reads** | Des queus poly A/T peuvent etre laissés pendant la préparation de la librairie | Des ponts entre séquences sans relation aucune: Chimères | *PRINSEQ2* |
| **Contaminations** | RNA Ribosomal<br>RNA/DNA etrangere (PhiX, Bacteria, ...) | | SortMeRNA, riboPicker, DeconSeq |

# Nettoyage des reads



# Practice

1

*Go to CLEANING PRACTICE on our github*

https://southgreenplatform.github.io/trainings/ouaga-NGS/rnaseqPractice/#practice-0
https://southgreenplatform.github.io/trainings/ouaga-NGS/rnaseqPractice/#practice-1

# What about references ? …

# Format : FASTA

## Fasta – format for representing nucleic acid or amino acid sequences

```
>AY274119.3 Severe acute respiratory syndrome-related coronavirus
isolate Tor2, complete genome

ATATTAGGTTTTTACCTACCCAGGAAAAGCCAACCAACCTCGATCTCTTGTAGATCTGTTCTCTAAACGA
ACTTTAAAATCTGTGTAGCTGTCGCTCGGCTGCATGCCTAGTGCACCTACGCAGTATAAACAATAATAAA
TTTTACTGTCGTTGACAAGAAACGAGTAACTCGTCCCTCTTCTGCAGACTGCTTACGGTTTCGTCCGTGT
TGCAGTCGATCATCAGCATACCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAAGATGGAGAGCCTTGTTC
TTGGTGTCAACGAGAAAACACACGTCCAACTCAGTTTGCCTGTCCTTCAGGTTAGAGACGTGCTAGTGCG
TGGCTTCGGGGACTCTGTGGAAGAGGCCCTATCGGAGGCACGTGAACACCTCAAAAATGGCACTTGTGGT
…

>FJ882960.1 SARS coronavirus ExoN1 isolate P3pp34, complete genome
CGATCTCTTGTAGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTAGCTGTCGCTCGGCTGCATGCCTA
GTGCACCTACGCAGTATAAACAATAATAAATTTTACTGTCGTTGACAAGAAACGAGTAACTCGTCCCTCT
TCTGCAGACTGCTTACGGTTTCGTCCGTGTTGCAGTCGATCATCAGCATACCTAGGTTTCGTCCGGGTGT
…
```

First line starts with “>” header or ”Comment”; used as a summary/description, often starting with unique accession/identifier

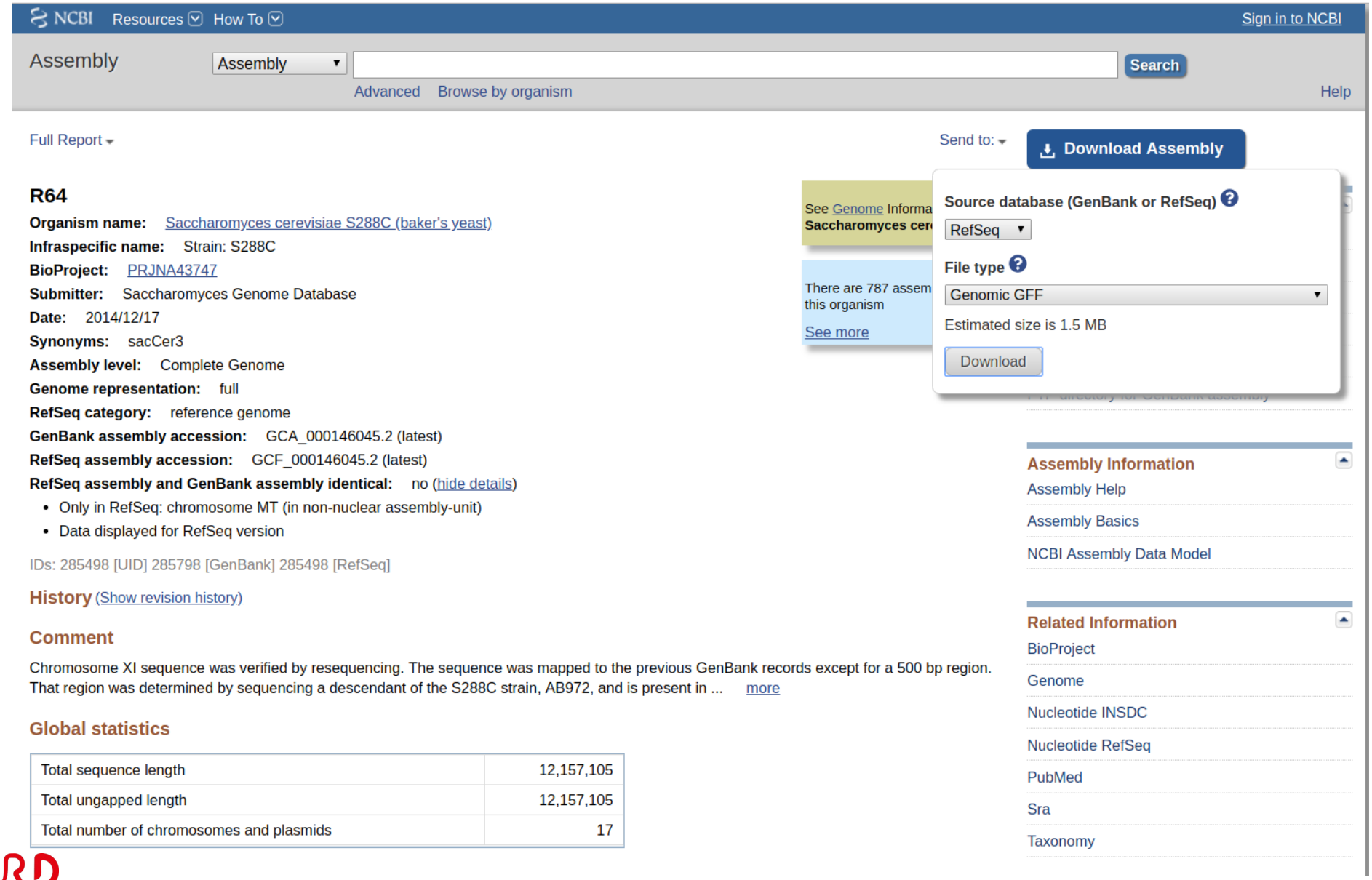
Subsequent lines contain sequence

- Interleaved: sequence broken into multiple lines of characters
- Sequential: entire sequence on a single line

Multiple sequence FASTA obtained by simply concatenating multiple FASTA records together

From rnabio.org 2019

# Format : FASTA



# GFF/GTF - representing sequence features

- GFF – General/Generic Feature Format; Gene Finding Format
  - Two versions in wide use
    - GFF2 (see also GTF)
    - GFF3
      - Added formal support for multiple levels (and direction) of hierarchy (e.g., gene -> transcript -> exon)
- GTF – Gene Transfer Format
  - An extension of GFF2
- GFF2, GFF3 and GTF are all tab-separate files with 9 fields
  - Differing content in 9th column

From rnabio.org 2019

# gtf/gff Format

GFF (general feature format) is a file format used for describing genes and other features of DNA, RNA and protein sequences.

**gff3**

| Seqname | Source | Feature | Start | End | Score | Strand | Frame | Attribute |
|---|---|---|---|---|---|---|---|---|
| chr22 | protein_coding | gene | 19701987 | 19712295 | . | + | . | ID=ENSG00000184702;Name=SEPT5 |
| chr22 | protein_coding | mRNA | 19707711 | 19708397 | . | + | . | ID=ENST00000413258;Name=SEPT5-016;Parent=ENSG00000184702 |
| chr22 | protein_coding | protein | 19707711 | 19708397 | . | + | . | ID=ENSP00000404673;Name=SEPT5-016;Parent=ENST00000413258 |
| chr22 | protein_coding | CDS | 19707711 | 19707761 | . | + | 1 | Name=CDS:SEPT5;Parent=ENST00000413258 |
| chr22 | protein_coding | CDS | 19707843 | 19707977 | . | + | 1 | Name=CDS:SEPT5;Parent=ENST00000413258 |
| chr22 | protein_coding | CDS | 19708165 | 19708189 | . | + | 1 | Name=CDS:SEPT5;Parent=ENST00000413258 |
| chr22 | protein_coding | CDS | 19708291 | 19708397 | . | + | 0 | Name=CDS:SEPT5;Parent=ENST00000413258 |
| chr22 | protein_coding | exon | 19707711 | 19707761 | . | + | . | Parent=ENST00000413258 |
| chr22 | protein_coding | exon | 19707843 | 19707977 | . | + | . | Parent=ENST00000413258 |
| chr22 | protein_coding | exon | 19708165 | 19708189 | . | + | . | Parent=ENST00000413258 |
| chr22 | protein_coding | exon | 19708291 | 19708397 | . | + | . | Parent=ENST00000413258 |

source: Abims RNAseq formation 2018

# Format : GFF/GTF

## GFF/GTF – general structure

**General GFF structure**

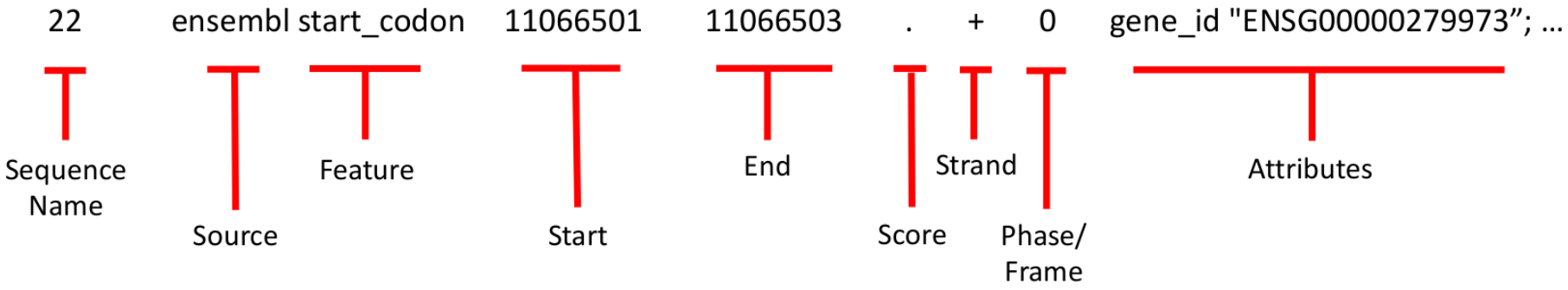
| Position index | Position name | Description |
|---|---|---|
| 1 | sequence | The name of the sequence where the feature is located. |
| 2 | source | Keyword identifying the source of the feature, like a program (e.g. Augustus or RepeatMasker) or an organization (like TAIR). |
| 3 | feature | The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow their parents in a single block (so all exons of a transcript are put after their parent "transcript" feature line and before any other parent transcript line). In GFF3, all features and their relationships should be compatible with the standards released by the Sequence Ontology Project. |
| 4 | start | Genomic start of the feature, with a **1-base offset**. This is in contrast with other 0-offset half-open sequence formats, like BED files. |
| 5 | end | Genomic end of the feature, with a **1-base offset**. This is the same end coordinate as it is in 0-offset half-open sequence formats, like BED files.[*citation needed*] |
| 6 | score | Numeric value that generally indicates the confidence of the source on the annotated feature. A value of "." (a dot) is used to define a null value. |
| 7 | strand | Single character that indicates the Sense (molecular biology) strand of the feature; it can assume the values of "+" (positive, or 5'->3'), "-", (negative, or 3'->5'), "." (undetermined). |
| 8 | phase | phase of CDS features; it can be either one of 0, 1, 2 (for CDS features) or "." (for everything else). See the section below for a detailed explanation. |
| 9 | Attributes. | All the other information pertaining to this feature. The format, structure and content of this field is the one which varies the most between the three competing file formats. |

https://en.wikipedia.org/wiki/General_feature_format

From rnabio.org 2019

# Format : GFF/GTF

## Ensembl GTF example record




**Example of attributes string:**

gene_id "ENSG00000279973"; gene_version "1"; transcript_id "ENST00000624155"; transcript_version "1"; exon_number "1"; gene_name "BAGE5"; gene_source "ensembl"; gene_biotype "protein_coding"; transcript_name "BAGE5-201"; transcript_source "ensembl"; transcript_biotype "protein_coding"; tag "basic"; transcript_support_level "1";

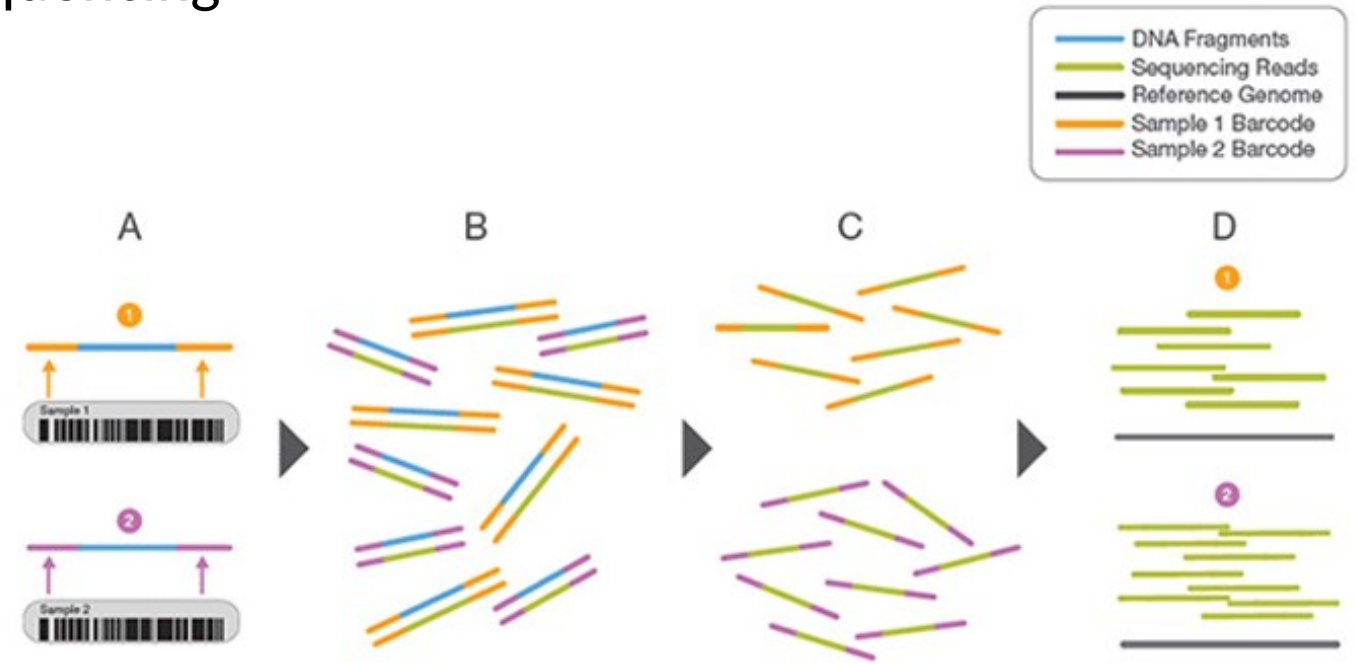
Note: there will be many GTF records/rows per transcript per gene (UTRs, start_codon, exons, etc)

From rnabio.org 2019

# Indexation

Index” has many different meanings :

• Indexes can refer to unique barcodes used for multiplexing DNA before sequencing




A. Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
B. Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
C. Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
D. Each set of reads is aligned to the reference sequence.

From rnabio.org 2019

# Indexation

## Indexing in bioinformatics/CS enables rapid access

- Indexing is a recurring theme in genome analysis
- Files are *big* - scanning through them can take a long time
- Indexing builds a table-of-contents so that we can jump directly to specific positions

- Indexing may require significant compute/time but typically only occurs once
- Each application may require a different indexing strategy

From rnabio.org 2019

# Indexation

## Example index applications and associated files

| Source file | Indexed file | Indexing tool | Use case |
|---|---|---|---|
| .bam | .bai | samtools index | Visualize bam in IGV |
| .fasta | .fai | faidx | Extract specific sequences from ref genome |
| .vcf | vcf.gz.tbi | bgzip/tabix | Pull out specific variants |
| .bed | .bed.gz.tbi | bgzip/tabix | extract specific genomic regions |

From rnabio.org 2019

# Introduction to the BED format

- When working with BAM files, it is very common to want to examine a focused subset of the reference genome
  - e.g. the exons of a gene
- These subsets are commonly specified in 'BED' files
  - https://genome.ucsc.edu/FAQ/FAQformat.html#format1
- Many BAM manipulation tools accept regions of interest in BED format
- Basic BED format (tab separated):
  - Chromosome name, start position, end position
  - Coordinates in BED format are 0 based

From rnabio.org 2019

# Manipulation of SAM/BAM and BED files

- Several tools are used ubiquitously in sequence analysis to manipulate these files

- SAM/BAM files
  - samtools
  - bamtools
  - Picard

- BED files
  - bedtools
  - bedops


From rnabio.org 2019

# Practice

2

*Go to INDEXING PRACTICE on our github*

https://southgreenplatform.github.io/trainings/ouaga-NGS/rnaseqPractice/#practice-2.1

# Trois stratégies d'analyse

# Which alignment strategy is best?

- De novo assembly
  - If a reference genome does not exist for the species being studied
  - If complex polymorphisms/mutations/haplotypes might be missed by comparing to the reference genome
- Align to transcriptome
  - If you have short reads (< 50bp)
- Align to reference genome
  - All other cases
- Each strategy involves different alignment/assembly tools

From rnabio.org 2019

# Stratégies

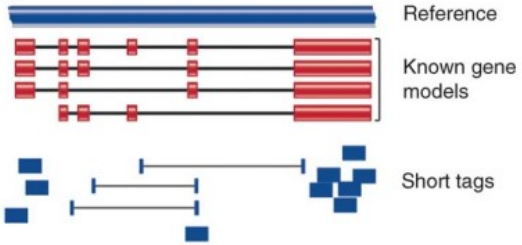
## Three RNA-seq mapping strategies

### De novo assembly

Assemble transcripts from overlapping tags

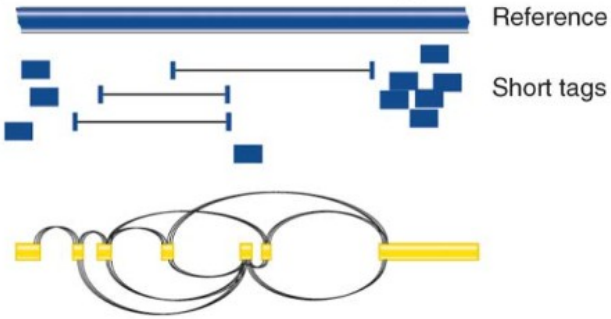

Optional: align to genome to get exon structure

### Align to transcriptome




Use known and/or predicted gene models to examine individual features

### Align to reference genome

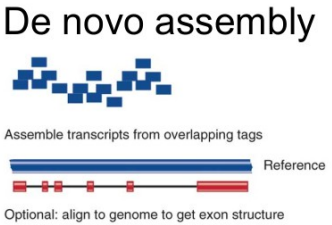



Infer possible transcripts and abundance

Diagrams from Cloonan & Grimmond, Nature Methods 2010

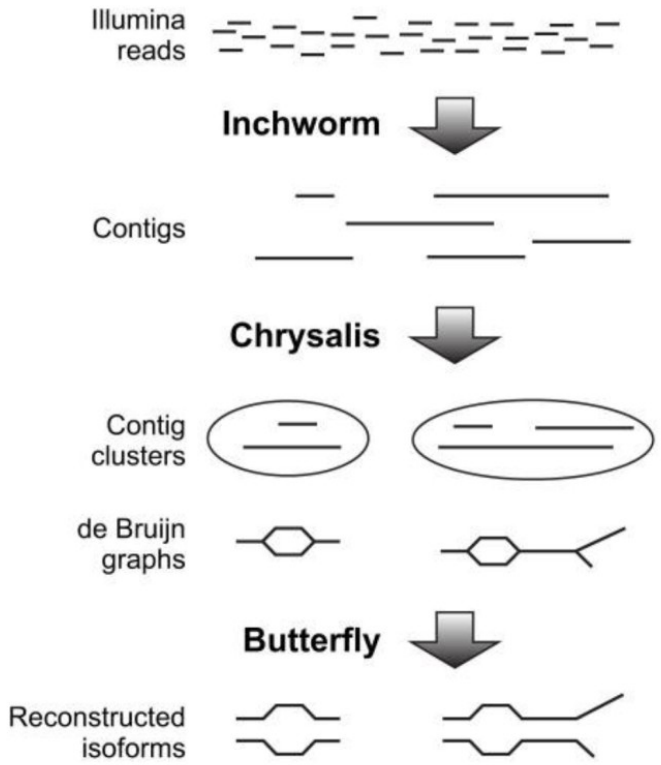
From rnabio.org 2019

# *De novo* Assembly




## Assembly

- Infer transcript structure directly from the data
- Useful when you do not have a reference sequence
- Other uses – highly rearranged genomes (some cancers)
- Computationally expensive
- Tools: Trinity, Velvet, SPAdes

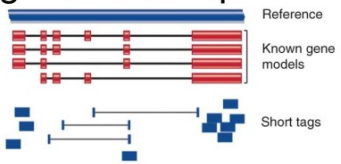



Haas, et al (2013) doi: 10.1038/nprot.2013.084

https://southgreenplatform.github.io/trainings/trinity/

From rnabio.org 2019

Align to transcriptome




Use known and/or predicted gene models to examine individual features

# Pseudoalign to transcriptome

# Pseudoalignement

Quantifying abundances of transcripts by ***pseudoalignment*** for rapidly determining the compatibility of reads with targets, without the need for alignment.

# What is a k-mer?

- A fixed sized (***K***) sequence
- A string of length ***N*** contains ***N-K+1*** k-mers

1-mer

| |
|---|
| A |
| C |
| G |
| T |

2-mer

| | | | |
|---|---|---|---|
| AA | AC | AG | AT |
| CA | CC | CG | CT |
| GA | GC | GG | GT |
| TA | TC | TG | TT |

ATTCGACAGTAGCCATGACTGG

...

- One can build *K*-mer index to represent a string

| 7-mer | iD | N |
|---|---|---|
| ATTCGAC | 1 | 1 |
| TTCGACA | 2 | 1 |
| TCGACAG | 3 | 1 |
| ... | | |

Sailfish: Alignment-free Isoform Quantification from RNA-seq Reads using Lightweight Algorithms Rob Patro, Stephen M. Mount, and Carl Kingsford. *Manuscript Submitted* (2013) http://www.cs.cmu.edu/~ckingsf/class/02714-f13/Lec05-sailfish.pdf

https://www.slideshare.net/duruofei/cmsc702-project-final-presentation

From rnabio.org 2019

# Alignment free approaches for transcript abundance

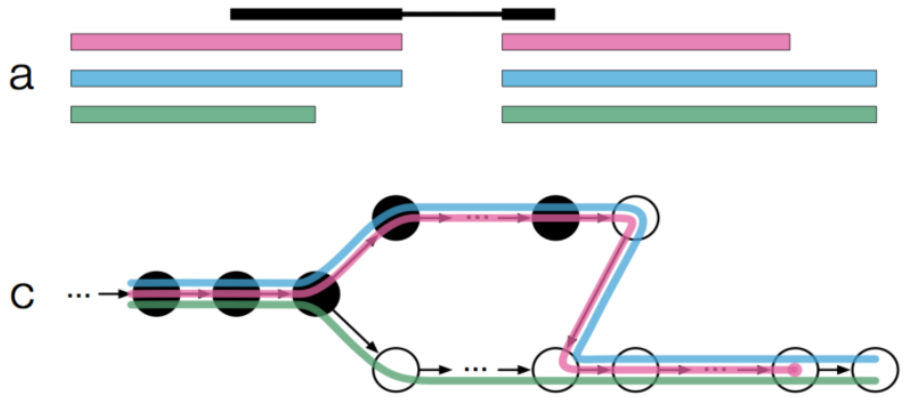
1. Obtain reference transcript sequences
   - e.g. Ensembl, Refseq, or GENCODE)

2. Build a **k-mer index** of all of the k-mers in each transcript sequence
   - Store each k-mer and its position within the transcript. “hashing”

From rnabio.org 2019

# Alignment free approaches for transcript abundance

3. Count number of times each k-mer occurs within each RNAseq read
   - Model relationship between RNA-seq read k-mers and the transcript k-mer index.
   - What transcript is the most likely source for each read?
   - Called “pseudoalignment” , “quasi-mapping”, etc.




Bray, 2016 doi:10.1038/nbt.3519

https://tinyheero.github.io/2015/09/02/pseudoalignments-kallisto.html

4. Handle sequencing errors, isoforms, ambiguity, and determine abundance estimates
   - Transcriptome de Bruijn graphs, likelihood function, expectation maximization, etc.

From rnabio.org 2019

# Advantages/disadvantages of alignment free approaches

- Advantages
  - Very fast and efficient
    - Similar accuracy to alignment based approach but with much, much shorter run time.
  - Do not need a reference genome, only a reference transcriptome

- Disadvantages
  - You don't get a proper BAM file (though a pseudo-bam can be created)
  - Information in reads with sequence errors may be ignored
  - Limited potential for transcript discovery, variant calling, fusion detection, etc.

From rnabio.org 2019

# Common alignment free tools

- Sailfish
  - "Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms." 2014
  - https://www.ncbi.nlm.nih.gov/pubmed/24752080
- RNA-Skim
  - "RNA-Skim: a rapid method for RNA-Seq quantification at transcript level." 2014
  - https://www.ncbi.nlm.nih.gov/pubmed/24931995
- Kallisto
  - "Near-optimal probabilistic RNA-seq quantification." 2016
  - https://www.ncbi.nlm.nih.gov/pubmed/27043002
- Salmon
  - "Salmon provides fast and bias-aware quantification of transcript expression." 2017
  - https://www.ncbi.nlm.nih.gov/pubmed/28263959

From rnabio.org 2019

# Practice

**3.1**

*Go to KALLISTO PRACTICE on our github*

https://southgreenplatform.github.io/trainings/ouaga-NGS/rnaseqPractice/#practice-3.1

# Correlation analysis : PtR Perl-to-R

Raw count matrix -> counts-per-million (CPM) data transformation followed by a log2 transform

## Compare replicate

```
trinityrnaseq-2.8.4/Analysis/DifferentialExpression/PtR --matrix
salmon.isoform.counts.matrix --samples ../sample_qc.txt --log2 -
-min_rowSums 10 --compare_replicates
```

## Correlation matrix

```
trinityrnaseq-2.8.4/Analysis/DifferentialExpression/PtR --matrix
salmon.isoform.counts.matrix --samples ../sample_qc.txt --log2 -
-min_rowSums 10 --CPM --sample_cor_matrix
```

## Principal composant analysis

```
trinityrnaseq-2.8.4/Analysis/DifferentialExpression/PtR --matrix
salmon.isoform.counts.matrix --samples ../sample_qc.txt --log2 -
-min_rowSums 10 --CPM --center_rows --prin_comp 3
```

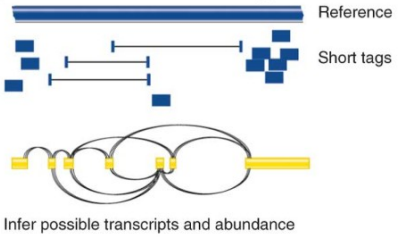
# Practice

3.2

*Go to Examine data before DE PRACTICE on our github*

https://southgreenplatform.github.io/trainings/ouaga-NGS/rnaseqPractice/#practice-3.2

Align to reference genome




# Align to reference genome

# Alignment - How does it work?


- Alignment is about fitting individual pieces (reads) into the correct part of the puzzle
- The human genome project gave us the picture on the box cover (the reference genome)
- Imperfections in how the pieces fit can indicate changes to a copy of the picture

Reference: AGCCTGAGACCGTAAAAAAGTCAAG

A read sequence: GAGACCGTAAAAACGTC

A variant!

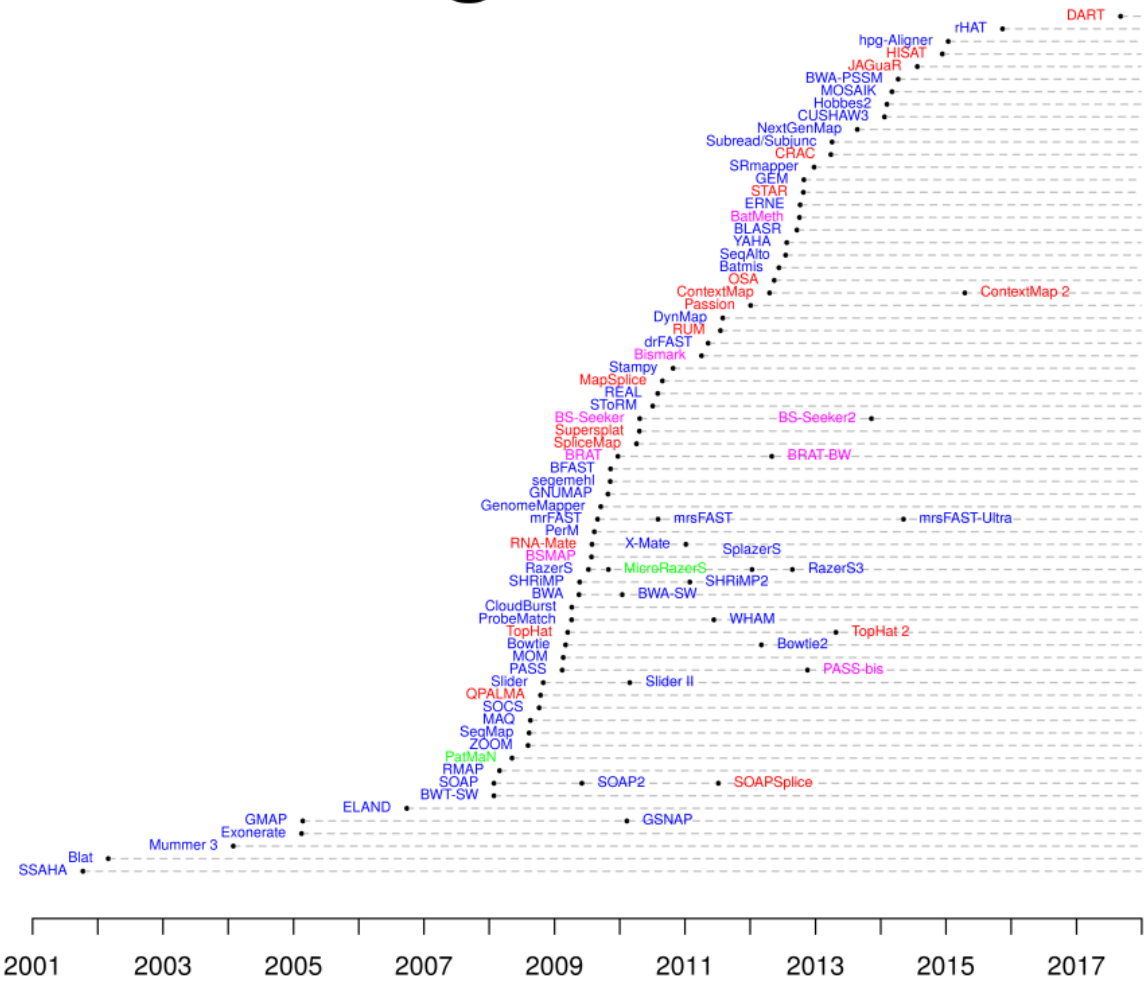
From rnabio.org 2019

# RNA-seq alignment challenges

- Computational cost
  - 100's of millions of reads
- Introns!
  - Spliced vs. unspliced alignments
- Can I just align my data once using one approach and be done with it?
  - Unfortunately probably not

From rnabio.org 2019

# Which read aligner should I use?



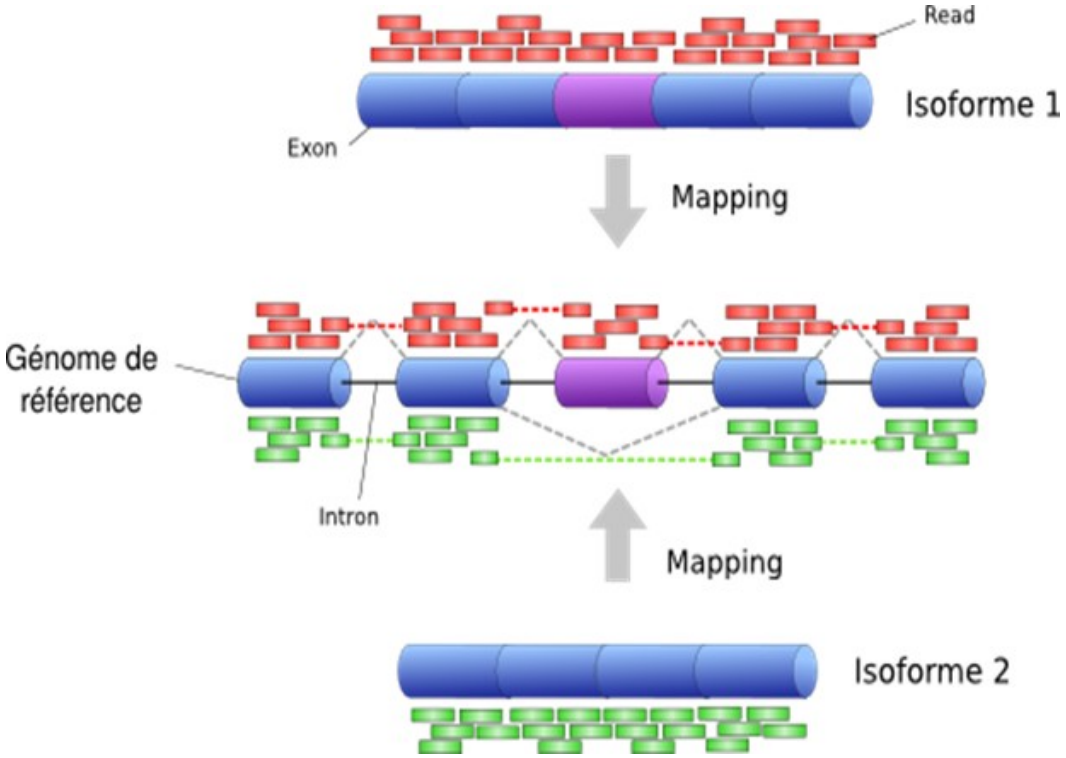

RNA
Bisulfite
DNA
microRNA

http://wwwdev.ebi.ac.uk/fg/hts_mappers/

From rnabio.org 2019

# Mapping sur référence génomique

- Permet la mise en évidence d'isoformes
- Aide à l'annotation structurale du génome



# What is the output of HISAT2?

- A SAM/BAM file
  - SAM stands for Sequence Alignment/Map format
  - BAM is the binary version of a SAM file

- Remember, compressed files require special handling compared to plain text files

- How can I convert BAM to SAM?
  - http://www.biostars.org/p/1701/

- Is HISAT2 the only mapper to consider for RNA-seq data?
  - http://www.biostars.org/p/60478/

From rnabio.org 2019

# SAM format

```
@SQ     SN:sctg_997     LN:6840
@SQ     SN:sctg_998     LN:6818
@SQ     SN:sctg_999     LN:6999

@PG     ID:TopHatVN:2.0.3

HWI-ST132_0435:3:63:3889:100528#GATCAG

0
sctg_0
476
3
73M
*
0
0

CCCGCCGCTCCATGATCTCCAAGAGGCGCAGCTCTCGCAAGGCTTCCGCCAAGGTGGTGGCTT
ggggggggggggggggggggggggggggeeeggeeggYb^ce^bbbc_cac[ddacaa_c

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:73 YT:Z:UU XS:A:+ NH:i:2
CC:Z:sctg_861 CP:i:9032  HI:i:0
```

Reference sequence dictionary

Program

Sequence ID

Flag
Reference sequence name
Leftmost mapping position
Mapping quality
CIGAR string
Reference of the mate/next read
Mapping position of mate/next read
Insert size

Sequence ID
Quality

Mapper tag

htpp://samtools.sourceforge.net/SAMv1.pdf

# Introduction to the SAM/BAM format

- The specification
  - http://samtools.sourceforge.net/SAM1.pdf
- SAM is uncompressed text data
- BAM is a compressed version of SAM
  - lossless BGZF format
- BAM files are usually 'indexed'
  - A '.bai' file will be found beside the '.bam' file
  - Indexing provides fast retrieval of alignments overlapping a specified region without going through all alignments.
  - BAM must be sorted by the reference ID and then the leftmost coordinate before indexing

From rnabio.org 2019

# SAM/BAM header section

- Used to describe source of data, reference sequence, method of alignment, etc.

- Each section begins with character '@' followed by a two-letter record type code. These are followed by two-letter tags and values:

  - @HD The header line
    - VN: format version
    - SO: Sorting order of alignments
  - @SQ Reference sequence dictionary
    - SN: reference sequence name
    - LN: reference sequence length
    - SP: species
  - @RG Read group
    - ID: read group identifier
    - CN: name of sequencing center
    - SM: sample name
  - @PG Program
    - PN: program name
    - VN: program version

```
orjuela@MPLCLTLP0157:~/Documents/2019/BURKINA_FORMATION/TP-OUAGA/HISAT-STRINGTIE/OUT/output/SRR453566/4_samToolsSort$ samtools view -H SRR453566.SAMTOOLSSORT.bam
@HD     VN:1.0  SO:coordinate
@SQ     SN:NC_001133.9  LN:230218
@RG     ID:SRR453566    SM:SRR453566
@PG     ID:hisat2       PN:hisat2       VN:2.0.0-beta
```

From rnabio.org 2019

# SAM/BAM alignment section

| Col | Field | Type | Regexp/Range | Brief description |
|---|---|---|---|---|
| 1 | QNAME | String | [!-?A-~]{1,255} | Query template NAME |
| ★ 2 | FLAG | Int | [0,$2^{16}$-1] | bitwise FLAG |
| 3 | RNAME | String | *\|[!-()+-<>-~][!-~]* | Reference sequence NAME |
| 4 | POS | Int | [0,$2^{29}$-1] | 1-based leftmost mapping POSition |
| 5 | MAPQ | Int | [0,$2^{8}$-1] | MAPping Quality |
| ★ 6 | CIGAR | String | *\|([0-9]+[MIDNSHPX=])+ | CIGAR string |
| 7 | RNEXT | String | *\|=\|[!-()+-<>-~][!-~]* | Ref. name of the mate/next segment |
| 8 | PNEXT | Int | [0,$2^{29}$-1] | Position of the mate/next segment |
| 9 | TLEN | Int | [-$2^{29}$+1,$2^{29}$-1] | observed Template LENgth |
| 10 | SEQ | String | *\|[A-Za-z=.]+ | segment SEQuence |
| 11 | QUAL | String | [!-~]+ | ASCII of Phred-scaled base QUALity+33 |

## Example values

```
1     QNAME  e.g.  HWI-ST495_129147882:1:2302:10269:12362
2     FLAG   e.g.  99
3     RNAME  e.g.  1
4     POS    e.g.  11623
5     MAPQ   e.g.  3
6     CIGAR  e.g.  100M
7     RNEXT  e.g.  =
8     PNEXT  e.g.  11740
9     TLEN   e.g.  217
10    SEQ    e.g.  CCTGTTTCTCCACAAAGTGTTTACTTTTGGATTTTTGCCAGTCTAACAGGTGAAGCCCTGGAGATTCTTATTAGTGATTTGGGCTGGGGCCTGGCCATGT
11    QUAL   e.g.  CCCFFFFFHHHHHJJIJFIJJJJJJJJJJHIJJJJJJJIJJJJJGGHIJHIJJJJJJJJJGHGGIJJJJJJIJEEHHHHFFFFCDCDDDDDDDB@ACDD
```

```
HWI-ST167:7:1208:18540:159854   163     NC_001133.9     69183   255     97M4S   =       69223   141     CGCCGCTACCGCCAATCTTGCCCTTTCTAACCATAGCATTGTACA
TCTGTTGGGGAGAAGGATACTCCCAGAACTCGTTACTGTCTGGCCTCTTGGGAATG        BCCFFFFFGDDHHIGIHJIIGJIIIIIJJJJIIGIIJIJJIDIGGDIJIIJJ<FGGHG@9?DEC@A@AB?>A3=CCBB?<?CC@@C
A<(2@<4:@A@DAAC   AS:i:-10        ZS:i:-17        XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:52T35A8    YS:i:-10    YT:Z:CP RG:Z:SRR453566  NH:i:1
```

From rnabio.org 2019

# SAM/BAM Flags explained

- 12 bitwise flags describing the alignment
- Stored as a binary string of length 11 instead of 11 columns of data
- Value of '1' indicates the flag is set. e.g. 00100000000
- All combinations can be represented as a number from 1 to 2048 (i.e. $2^{11}$-1). This number is used in the BAM/SAM file.
- You can specify 'required' or 'filter' flags in samtools view using the '-f' and '-F' options respectively

| Bit | | Description |
|---|---|---|
| 1 | 0x1 | template having multiple segments in sequencing |
| 2 | 0x2 | each segment properly aligned according to the aligner |
| 4 | 0x4 | segment unmapped |
| 8 | 0x8 | next segment in the template unmapped |
| 16 | 0x10 | SEQ being reverse complemented |
| 32 | 0x20 | SEQ of the next segment in the template being reverse complemented |
| 64 | 0x40 | the first segment in the template |
| 128 | 0x80 | the last segment in the template |
| 256 | 0x100 | secondary alignment |
| 512 | 0x200 | not passing filters, such as platform/vendor quality controls |
| 1024 | 0x400 | PCR or optical duplicate |
| 2048 | 0x800 | supplementary alignment |

Note that to maximize confusion, each bit is described in the SAM specification using its hexadecimal representation (i.e., '0x10' = 16 and '0x40' = 64).

http://broadinstitute.github.io/picard/explain-flags.html

From rnabio.org 2019

# CIGAR strings explained

- The CIGAR string is a sequence of base lengths and associated 'operations' indicating which bases align to the reference (either a match or mismatch), are deleted, are inserted, represent introns, etc.

| Op | BAM | Description |
|---|---|---|
| M | 0 | alignment match (can be a sequence match or mismatch) |
| I | 1 | insertion to the reference |
| D | 2 | deletion from the reference |
| N | 3 | skipped region from the reference |
| S | 4 | soft clipping (clipped sequences present in SEQ) |
| H | 5 | hard clipping (clipped sequences NOT present in SEQ) |
| P | 6 | padding (silent deletion from padded reference) |
| = | 7 | sequence match |
| X | 8 | sequence mismatch |

' that

- e.g. 81M859N19M
  - A 100 bp read consists of: 81 bases of alignment to reference, 859 bases skipped (an intron), 19 bases of alignment

From rnabio.org 2019

# Mapping Quality Check : Flagstats

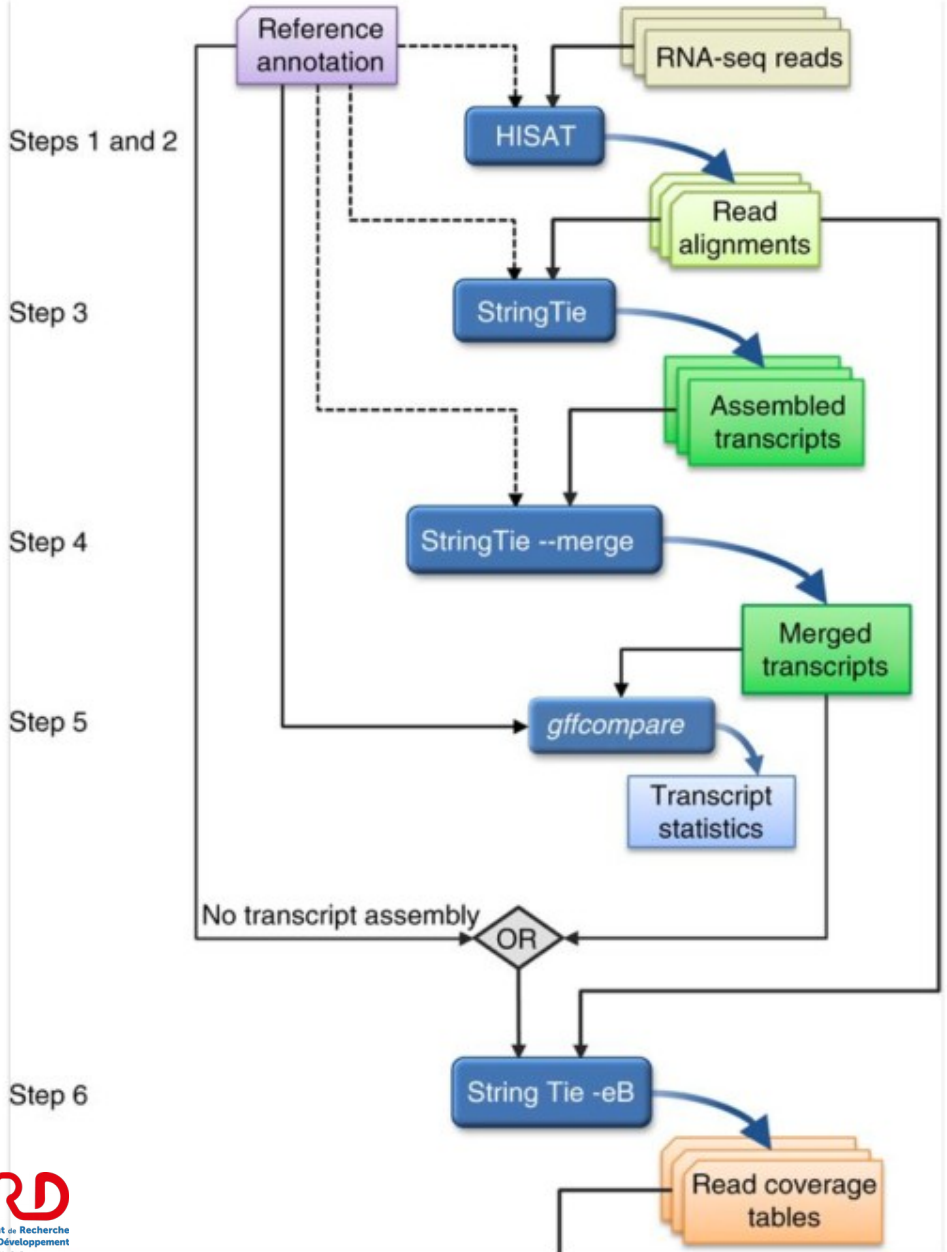
```
168 375 631 + 0 in total (QCpassed reads + QCfailed reads)
133 873 423 + 0 duplicates
143 893 516 + 0 mapped (85.46%:nan%)
168 375 631 + 0 paired in sequencing
84 186 587 + 0 read1
84 189 044 + 0 read2
143 722 660 + 0 properly paired (85.36%:nan%)
143 722 660 + 0 with itself and mate mapped
170 856 + 0 singletons (0.10%:nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

Properly paired reads =>
- Insert size ≤ max_insert_size
- R1 / R2 mapped on same chromosome
- R1 ===> <=== R2

https://broadinstitute.github.io/picard/explain-flags.htm

source: Abims RNAseq formation 2018

# New Tuxedo Workflow




Mapping

Assemblage des nouveau transcrits

Comparaison avec l'annotation

Obtention des tableaux de comptages

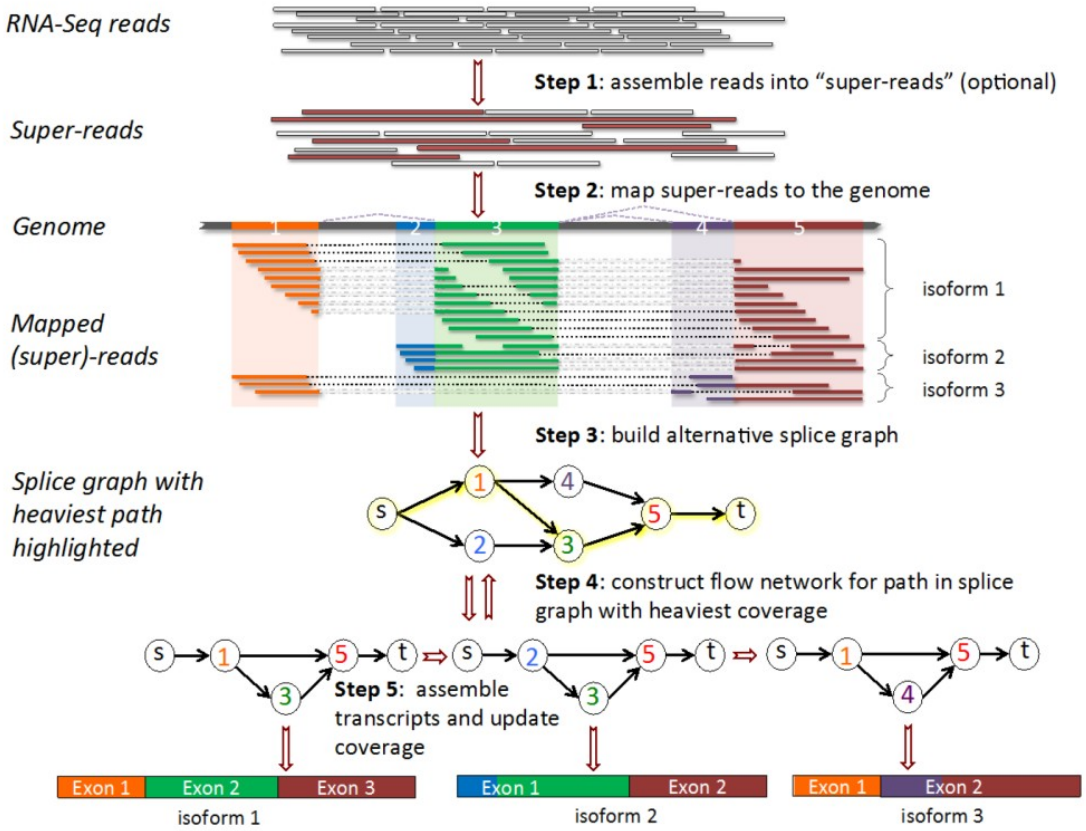
# StringTie

## How does StringTie work?

Map reads to the genome

Infer isoforms:

- iteratively extract the heaviest path from a splice graph
- construct a flow network
- compute maximum flow to estimate abundance
- update the splice graph by removing reads that were assigned by the flow algorithm
- This process repeats until all reads have been assigned.

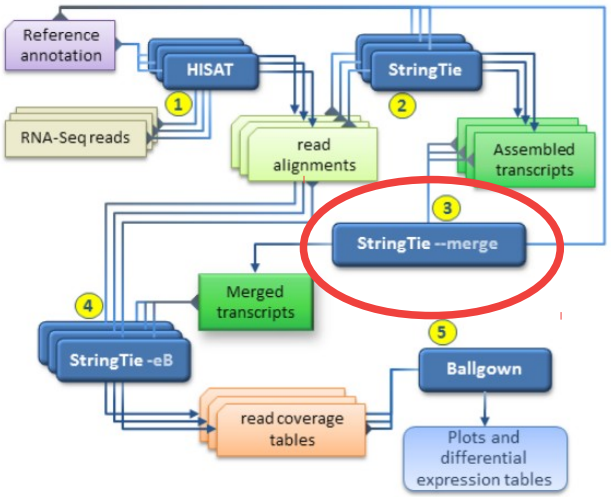



Pertea et al. Nature Biotechnology, 2015

From rnabio.org 2019

# StringTie -merge

- Merge together all gene structures from all samples
  - Some samples may only partially represent a gene structure
- Incorporates known transcripts with assembled, potentially novel transcripts
- For de novo or reference guided mode, we will rerun StringTie with the merged transcript assembly.

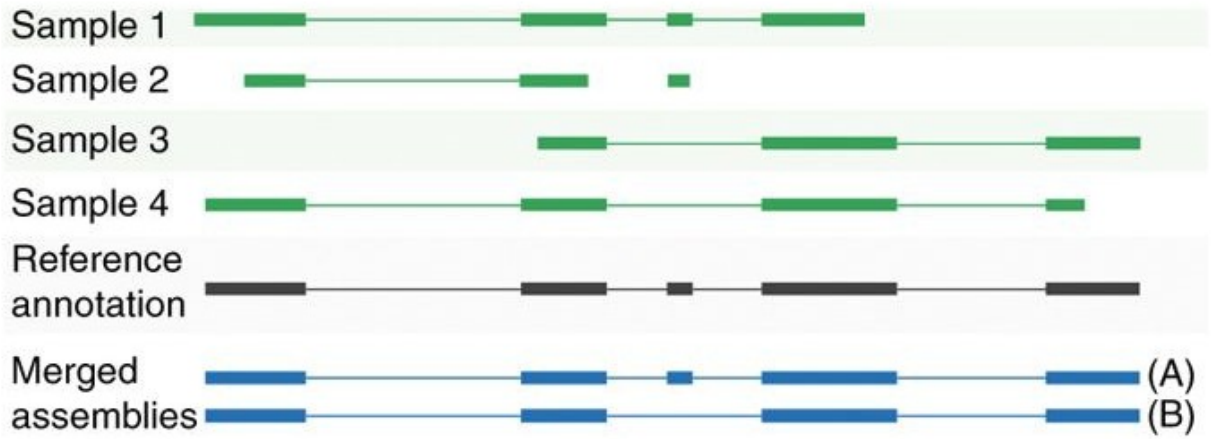



Pertea et al. Nature Protocols, 2016

# StringTie

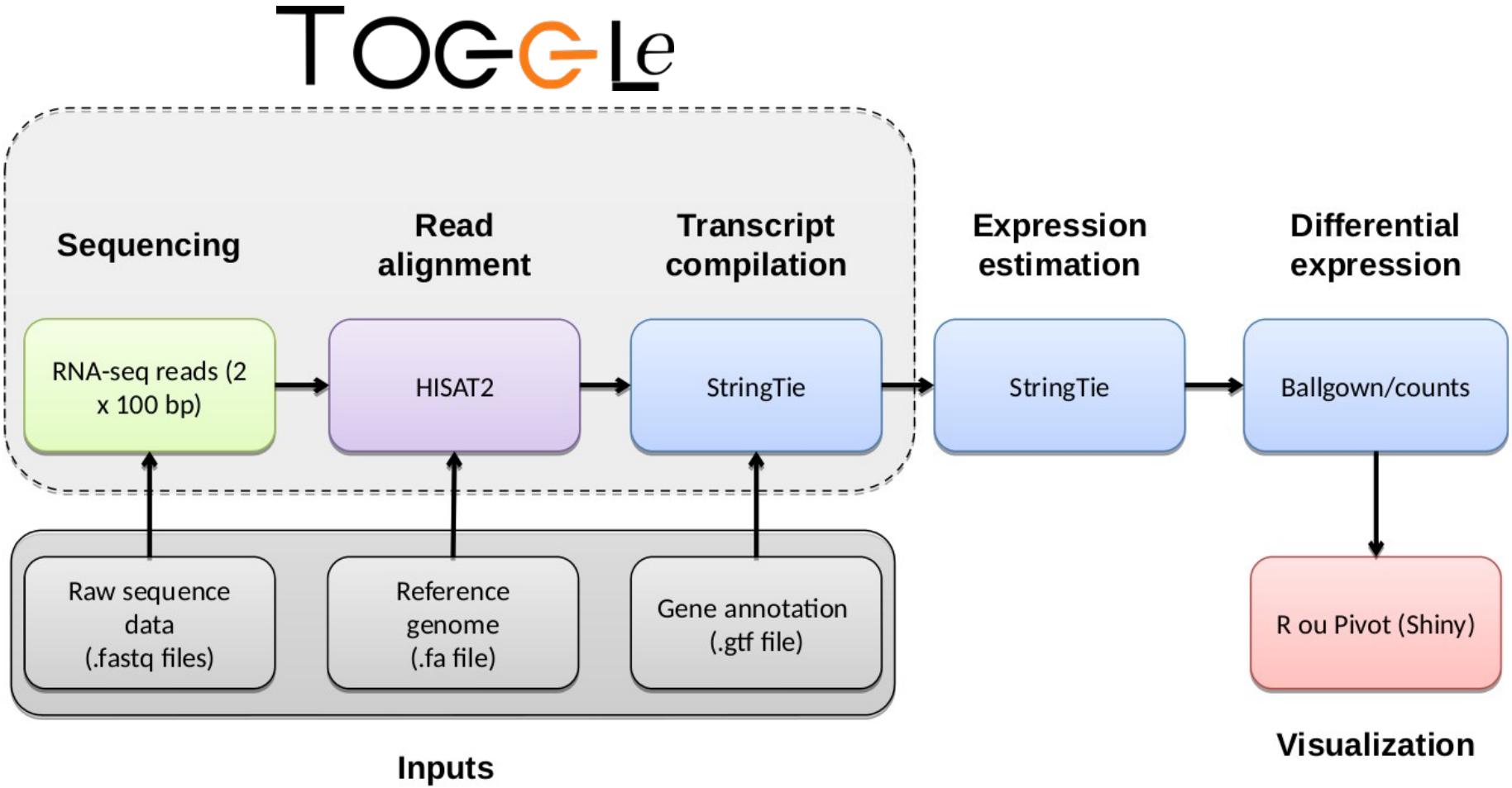
## Figure 2 : Merging transcript assemblies using StringTie's merge function.

From: Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown




In this example, four partial assemblies from four different samples are merged into two transcripts A and B. Samples 1 and 2 are both consistent with the reference annotation, which is used here to merge and extend them to create transcript A. Samples 3 and 4 are consistent with each other but not with the annotation, and these are merged to create transcript B.

# Workflow RNAseq with with Reference




Adapted from rnabio.org 2019

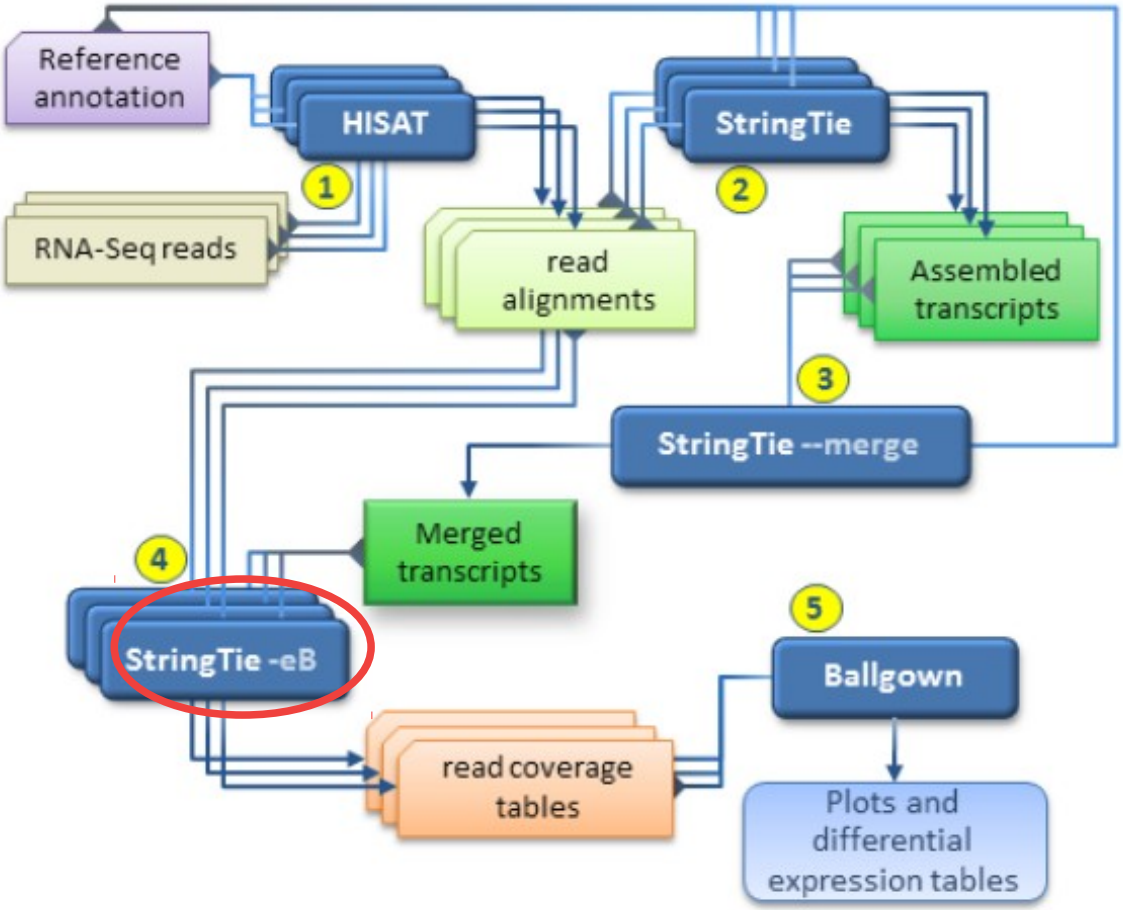
# Practice

4.1

*Go to TOGGLe RNASEQ PRACTICE on our github*

https://southgreenplatform.github.io/trainings/ouaga-NGS/rnaseqPractice/#practice-4.1

# StringTie -e -B




Estimate transcript abundances (-e) and generate read coverage tables (-B)

Pertea et al. Nature Protocols, 2016

# Practice

4.2

*Go to STRINGTIE -e -B PRACTICE on our github*

https://southgreenplatform.github.io/trainings/ouaga-NGS/rnaseqPractice/#practice-4.2

# gffcompare

The following table shows the code used by Cufflinks to classify the transcripts in comparison with the reference annotation

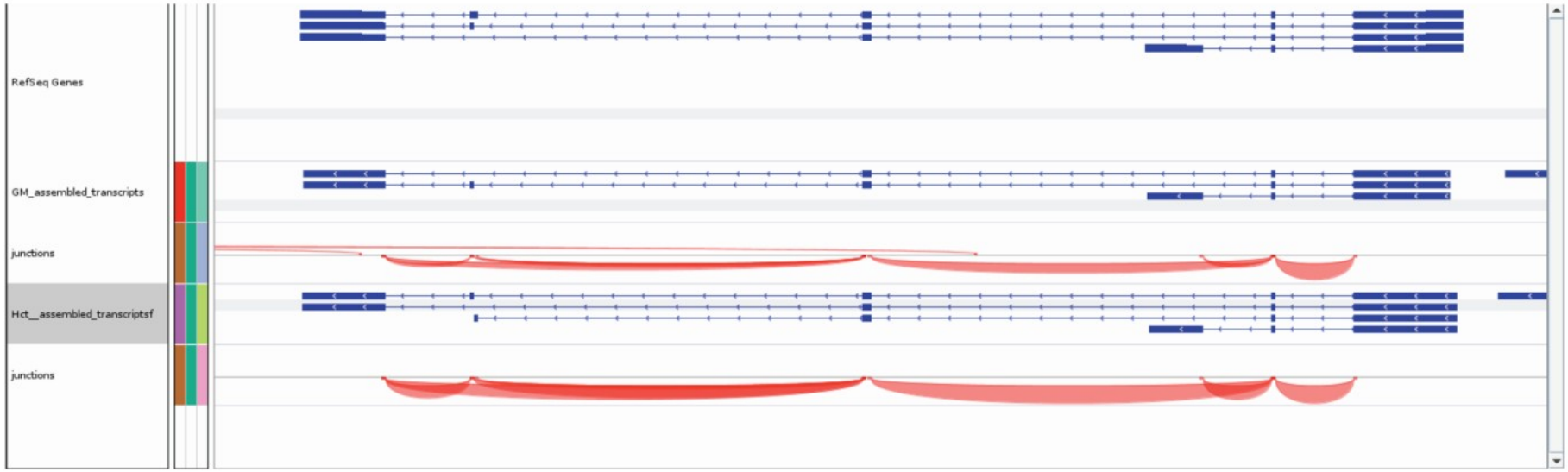
| Priority | Code | Description |
|---|---|---|
| **1** | **=** | **Complete match of intron chain** |
| 2 | c | Contained |
| **3** | **j** | **Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript** |
| 4 | e | Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment |
| 5 | i | A transfrag falling entirely within a reference intron |
| 6 | o | Generic exonic overlap with a reference transcript |
| 7 | p | Possible polymerase run-on fragment (within 2Kbases of a reference transcript) |
| 8 | r | Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case |
| **9** | **u** | **Unknown, intergenic transcript** |
| **10** | **x** | **Exonic overlap with reference on the opposite strand** |
| 11 | s | An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors) |
| 12 | . | (.tracking file only, indicates multiple classifications) |

http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html#cuffcompare-output-files

source: Abims RNAseq formation 2018

# gff compare

=

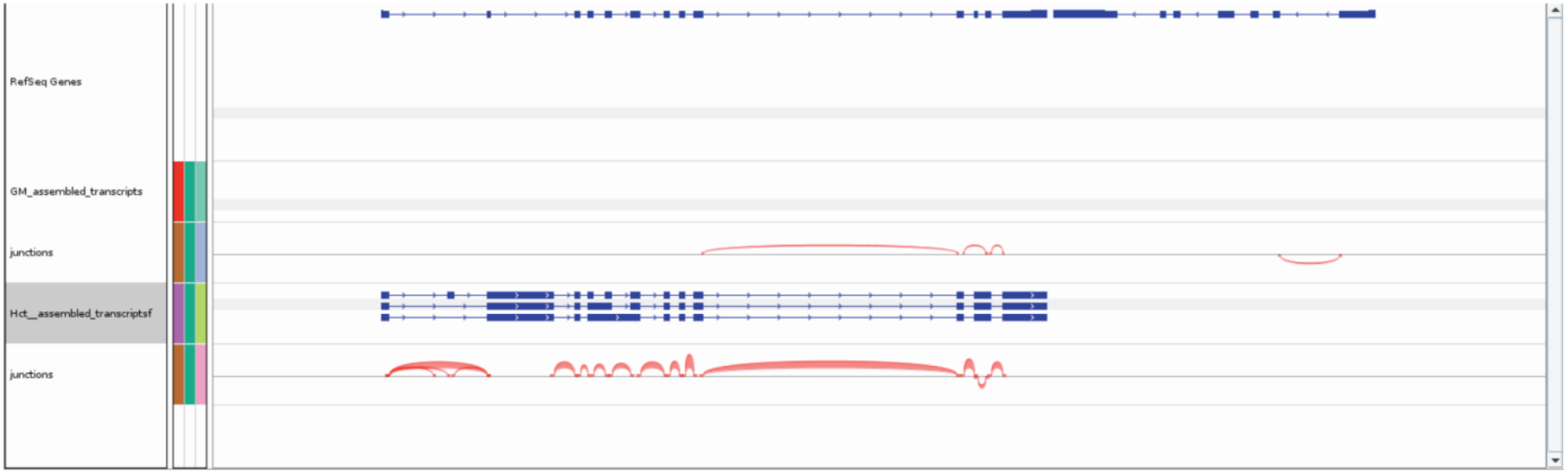



**Complete match of intron chain**

source: Abims RNAseq formation 2018

# gff compare

J

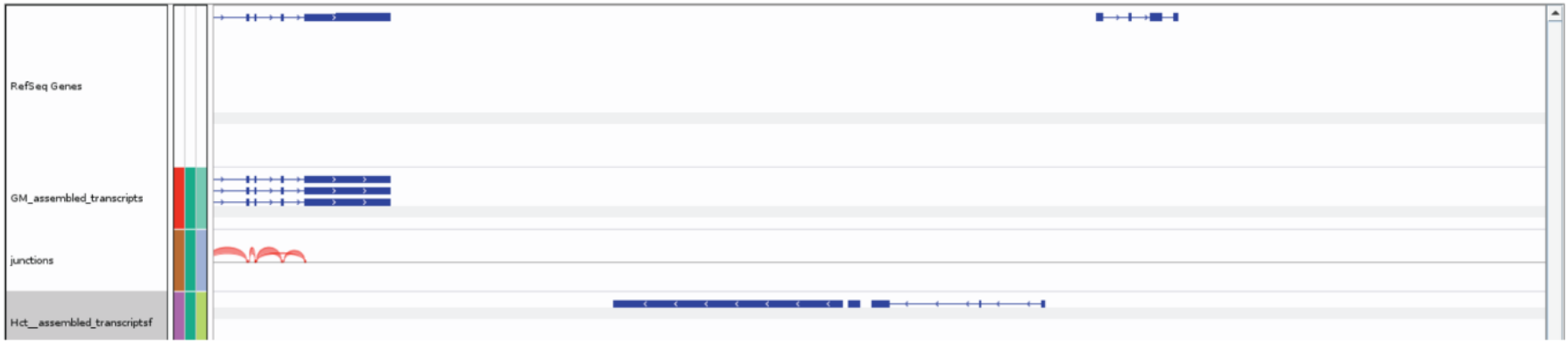



**Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript**

source: Abims RNAseq formation 2018

# gff compare

U




**Unknown, intergenic transcript**

source: Abims RNAseq formation 2018

# Practice

4.3

*Go to GFFCOMPARE PRACTICE on our github*

https://southgreenplatform.github.io/trainings/ouaga-NGS/rnaseqPractice/#practice-4.3

# Practice

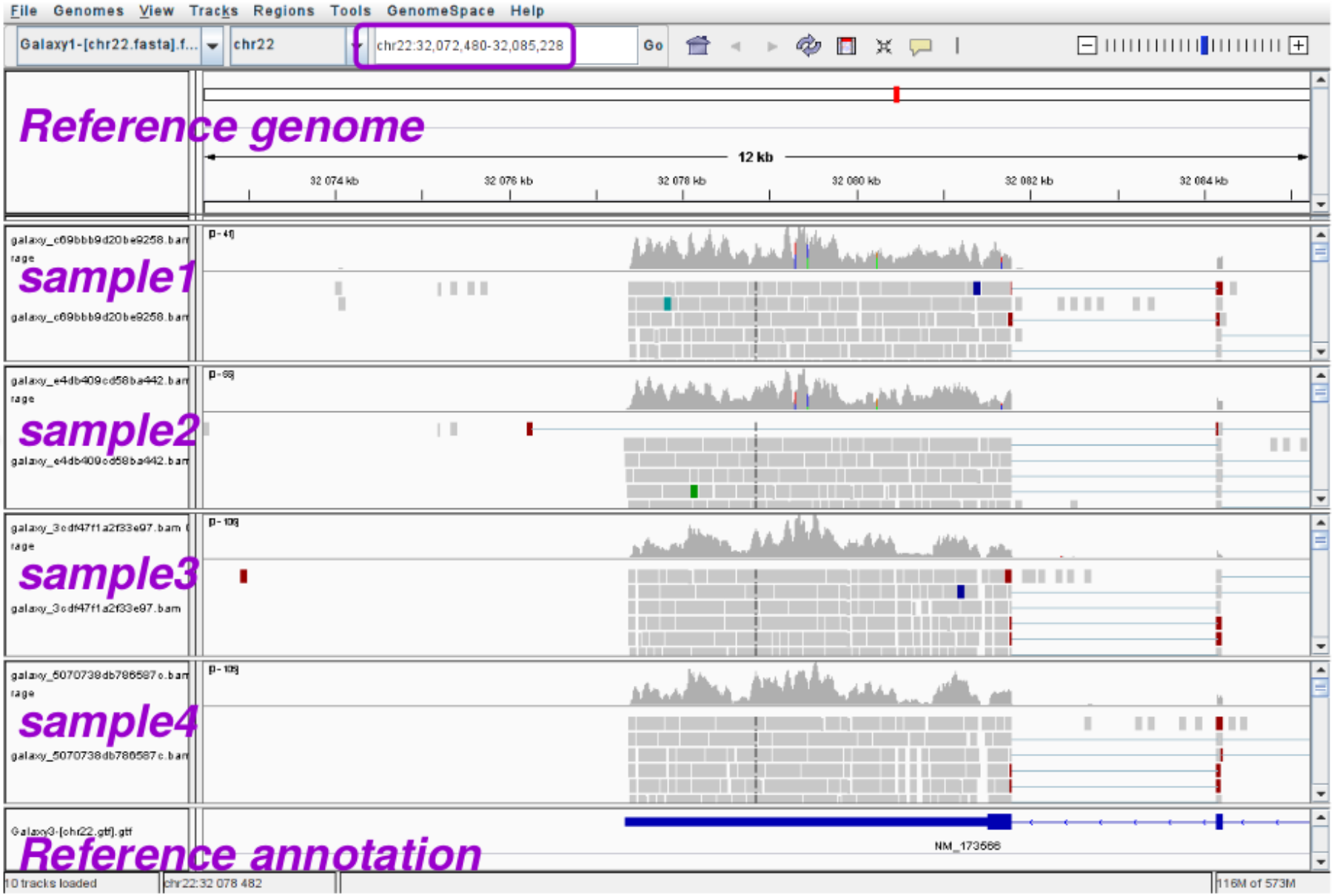
4.4

*Go to Obtainig Counts PRACTICE on our github*

https://southgreenplatform.github.io/trainings/ouaga-NGS/rnaseqPractice/#practice-4.4

# Visualization

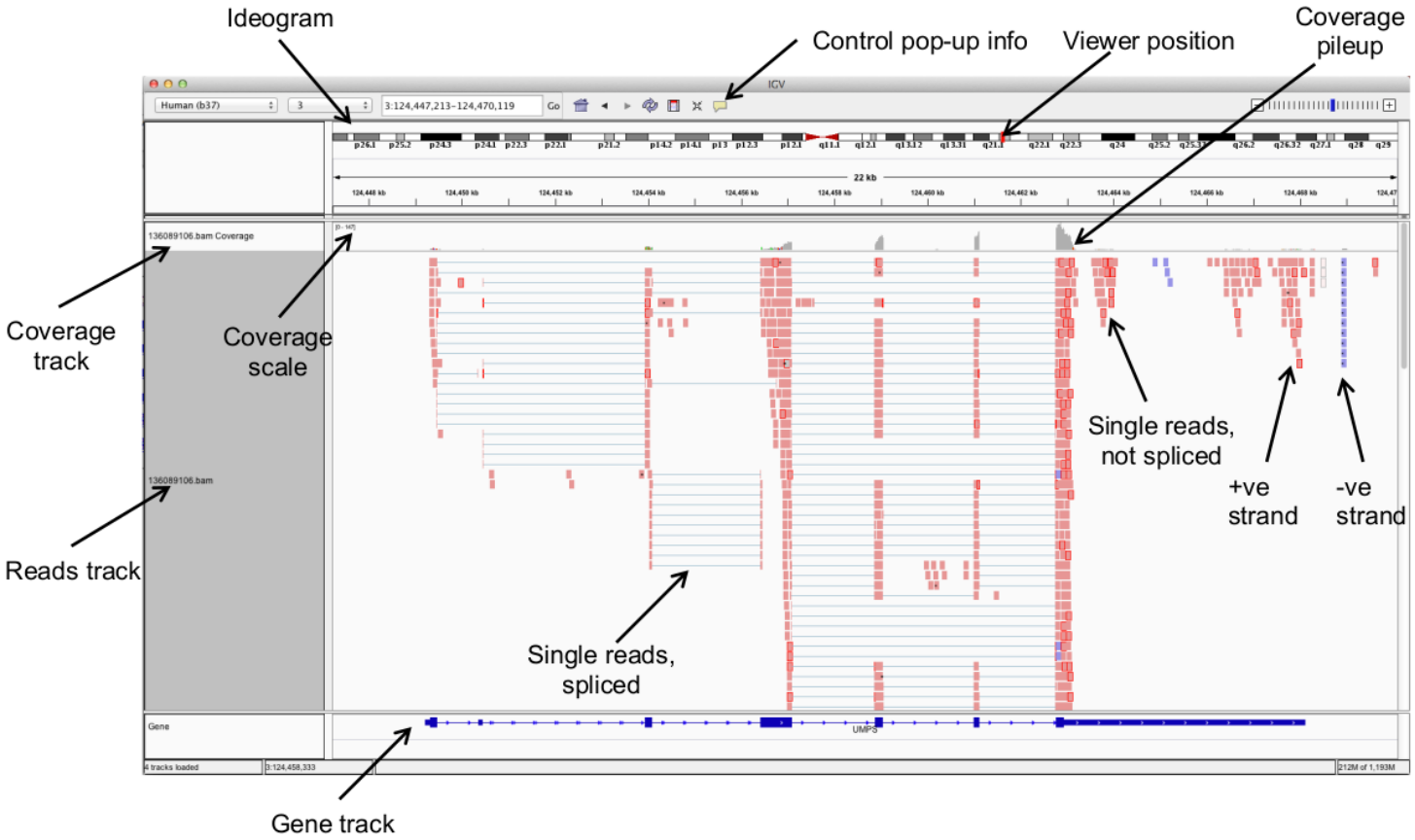
# Local Genome Browser (IGV)




source: Abims RNAseq formation 2018

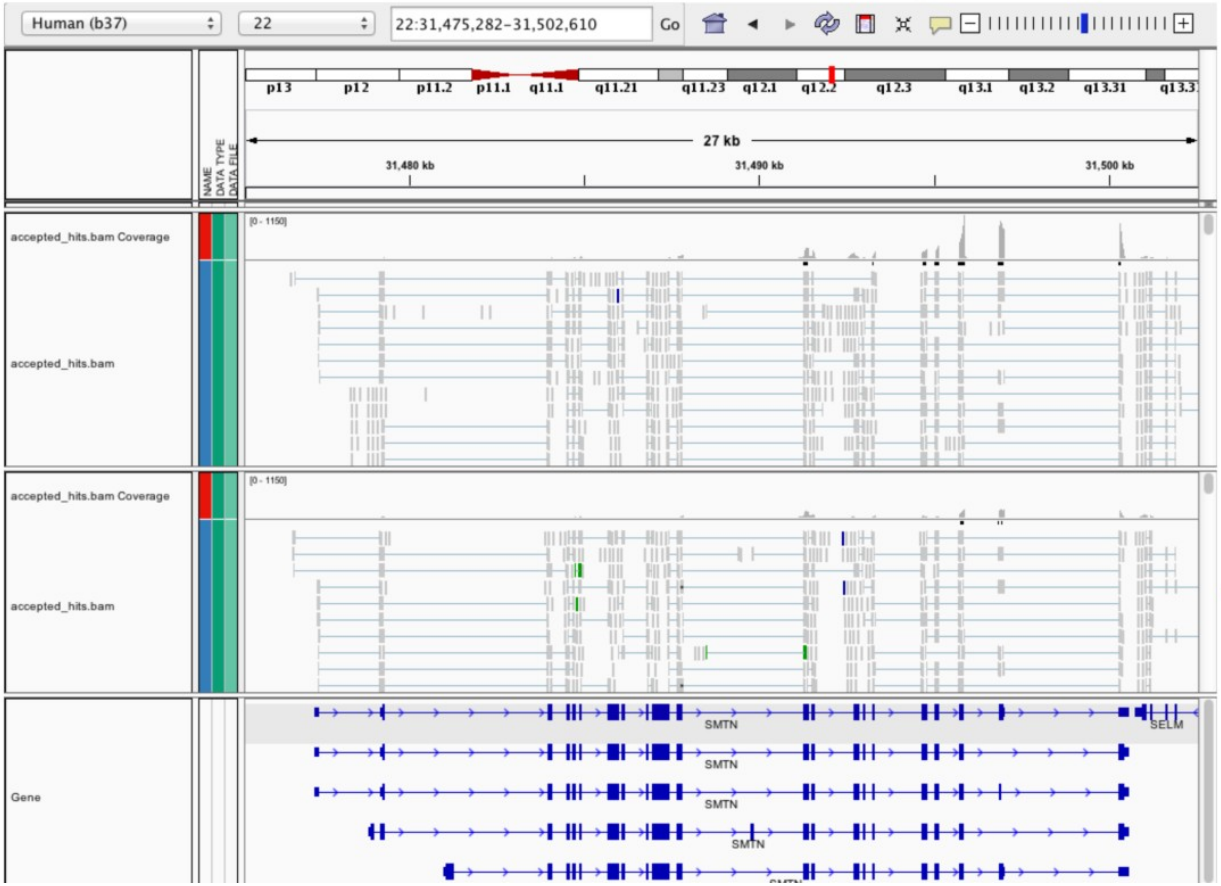
# IGV

## Visualization of RNA-seq alignments in IGV browser




From rnabio.org 2019

# Expression estimation for known genes and transcripts

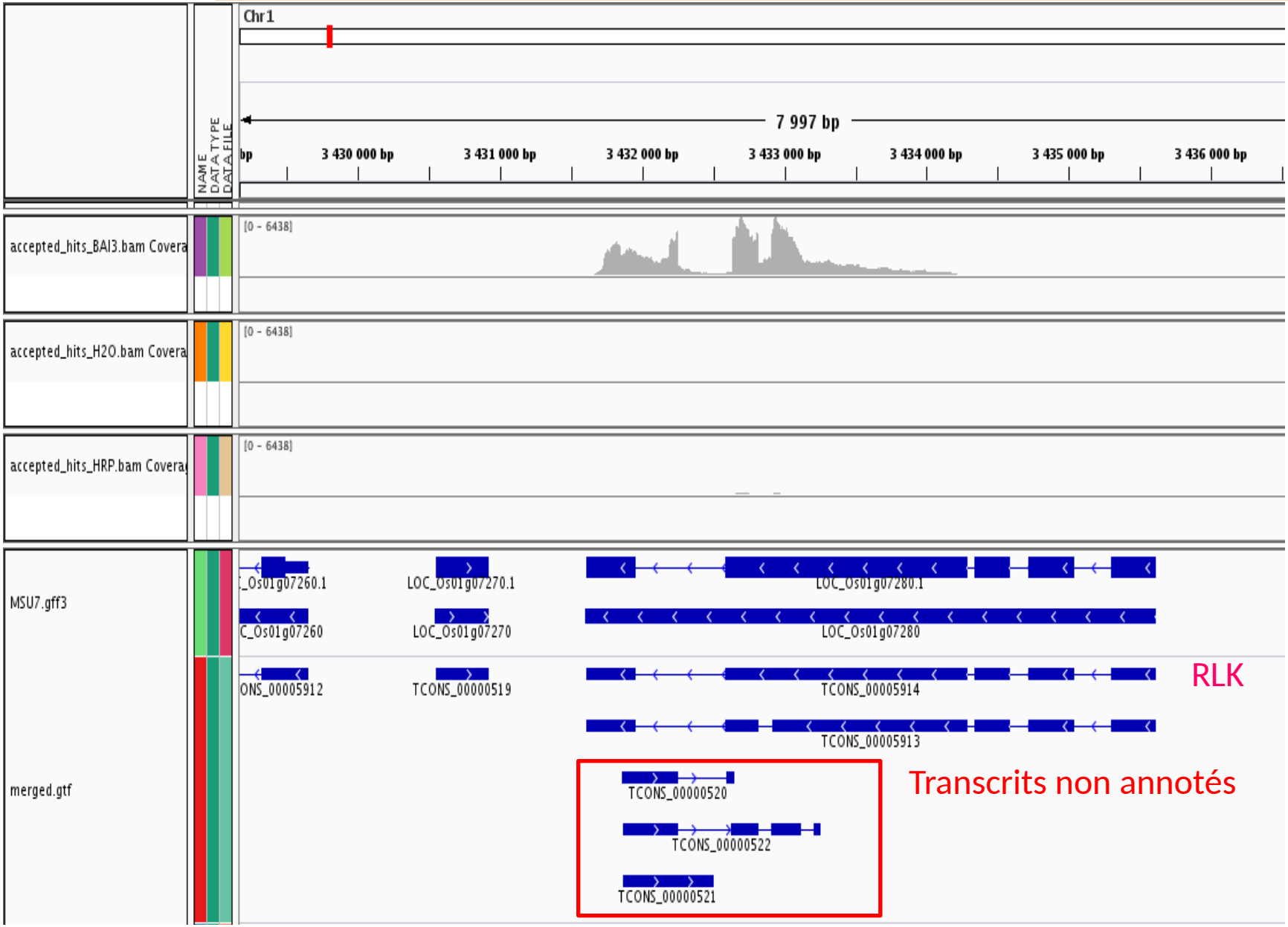



3' bias →

↓ Down-regulated

From rnabio.org 2019

# Local Genome Browser (IGV)



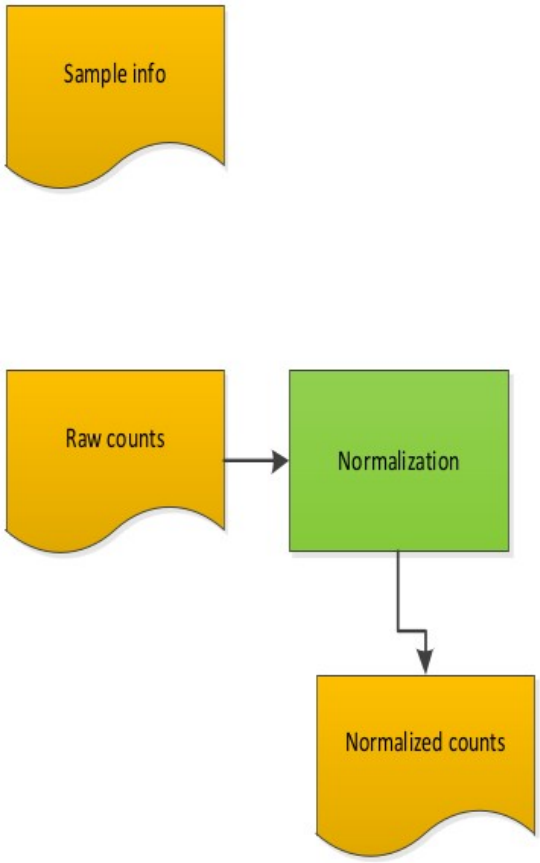
# Practice

5

Go to IGV VISUALIZATION PRACTICE on our github

# Differential Expression Chapter

# Normalisation des données



# Quantifying expression levels

## Factors need to be taken into account before making comparisons

- Library size (i.e. sequencing depth) varies between samples coming from different lanes of the flow cell of the sequencing machine.

- Longer genes will have higher number of reads.

- Library composition (i.e. relative size of the studied transcriptome) can be different in two different biological conditions.

- GC content biases across different samples may lead to a biased sampling of genes (Risso et al. 2011).

- Read coverage of a transcript can be biased and non-uniformly distributed along the transcript (Mortazavi et al. 2008).

# Normalisation des données

- Identifier et corriger les biais techniques dus au séquençage, pour les  rendre comparable

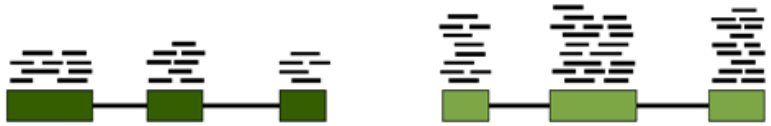
- Types de Normalisation : intra-echantillons (meme sequençage) et inter-echantillons (deux sequençages)

Taille de la banque
Longueur de gènes
Composition en GC des gènes

# Normalisation des données

| | sample 1 | sample 2 | sample 3 |
|---|---|---|---|
| gene A | 752 | 615 | 1203 |
| gene B | 1507 | 1225 | 2455 |

Le comptage du gène B est deux fois plus important que pour le gène A, pourquoi ?



gene A gene B

Le nombre de transcrits pour le gène B est deux fois plus important que pour le gène A

gene A gene B

Les deux gènes ont le même nombre de transcrits,
mais le gène B est deux fois plus long que le gène A.

- Permettre la comparaison de gènes pour un même échantillons.
- Les sources de variabilités : longueur du gène et composition en GC.

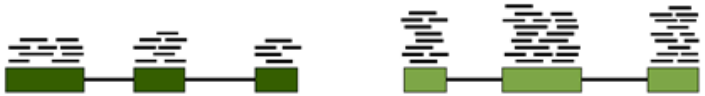
# Normalisation des données

| | sample 1 | sample 2 | sample 3 |
|---|---|---|---|
| gene A | 752 | 615 | 1203 |
| gene B | 1507 | 1225 | 2455 |

Le comptage dans l'échantillon 3 est plus important que dans l'échantillon 2.

gene A in sample 2 gene A in sample 3

Le gène A est plus exprimé dans l'échantillon 3 que dans le 2.



gene A in sample 2 gene A in sample 3

Le gène A est exprimé dans les échantillons 2 et 3, mais la profondeur de séquençage est plus importante dans l'échantillon 3 que dans le 2 (différences de taille des librairies).

- Permettre la comparaison de gènes pour différents échantillons.
- Les sources de variabilités : taille des librairies

# Normalisation des données

## Why performing normalisation ?

- Between-sample → compare a gene in different sample
  - Depth of sequencing == library size
  - Sampling bias during the libraries construction == batch effect
  - Presence of majority fragments == saturation
  - Sequence composition du to PCR-amplification step (GC content)

- Within-sample → compare genes in a sample
  - Gene length
  - Sequence composition (GC content)

# Counts per million CPM

- The most basic normalization approaches address the sequencing depth bias. Such procedures normalize the read counts per gene by dividing each gene's read count by a certain value and multiplying it by 10^6. These normalized values are usually referred to as **CPM (counts per million reads):**

- Total Counts Normalization (divide counts by the **sum** of all counts)
Upper Quartile Normalization (divide counts by the **upper quartile** value of the counts)
Median Normalization (divide counts by the **median** of all counts)

# TPM

Popular metrics that improve upon CPM are RPKM/FPKM (reads/fragments per kilobase of million reads) and TPM (transcripts per million).

- RPKM is obtained by dividing the CPM value by another factor, which is the **length of the gene per kilobases**. FPKM is the same as RPKM, but is used for paired-end reads. Thus, **RPKM/FPKM methods account for, firstly the library size, and secondly the gene lengths**.

- **TPM** also controls for both the library size and the gene lengths, however, with the TPM method, **the read counts are first normalized by the gene length** (per kilobase), **and then gene-length normalized values** are divided by the sum of the gene-length normalized values and multiplied by 10^6. Thus, the sum of normalized values for TPM will always be equal to 10^6 for each library, while the sum of RPKM/FPKM values do not sum to 10^6.

- Therefore, it is easier to interpret TPM values than RPKM/FPKM values.

# FPKM (RPKM)

- RPKM: **Reads** Per Kilobase of transcript per Million mapped reads.
- FPKM: **Fragments** Per Kilobase of transcript per Million mapped reads.

- No essential difference - Just a terminology change to better describe paired-end reads!

From rnabio.org 2019

# Quantifying expression levels

## How do FPKM and TPM differ?

- TPM: Transcript per Kilobase Million
- The difference is in the order of operations:

**FPKM**

1) Determine total fragment count, divide by 1,000,000 (per Million)

2) Divide each gene/transcript fragment count by #1 (Fragments Per Million)

3) Divide each FPM by length of each gene/transcript in kilobases (FPKM)

**TPM**

1) Divide each gene/transcript fragment count by length of the transcript in kilobases (Fragments Per Kilobase)

2) Sum all FPK values for the sample and divide by 1,000,000 (per Million)

3) Divide #1 by #2 (TPM)

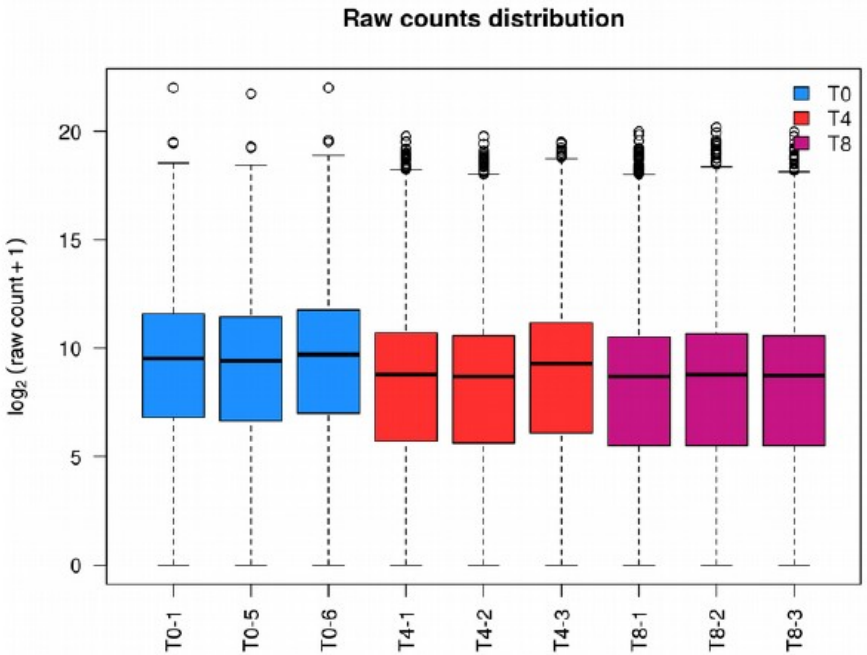
- The sum of all TPMs in each sample is the same. Easier to compare across samples!

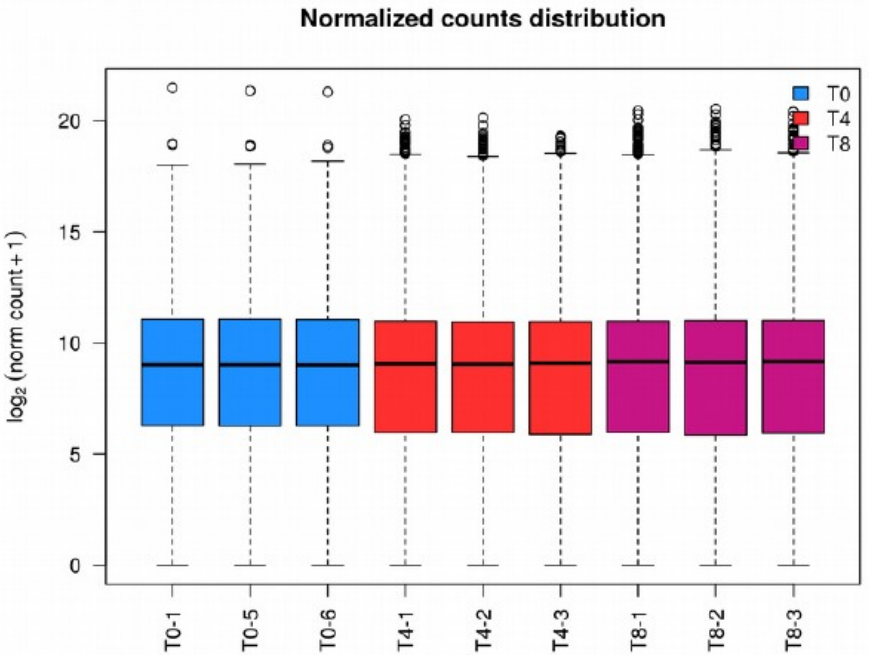
- http://www.rna-seqblog.com/rpkm-fpkm-and-tpm-clearly-explained/
- https://www.ncbi.nlm.nih.gov/pubmed/22872506

From rnabio.org 2019

# Normalisation des données

## Effet de la normalisation :
## Variance des banques RNAseq avant et après normalisation



# Méthodes basées sur les normalisations interbanques

| Normalization Technique | Name authors | Description | Software |
|---|---|---|---|
| UQ | Upper Quartile Ref : Bullard et al., 2010 (Upper) Quartile normalization | Les comptages par gène sont divises par le 3e quartile des comptages non nuls de l'échantillon, puis multiplies par la moyenne des 3e quartiles de tous les échantillons. | EdgeR |
| TC | Total read count adjustment Ref : Mortazavi et al., 2008 | Chaque nombre reads est divisé par le nombre total de reads (taille de la banque), puis multiplier par le nombre total moyen de reads des librairies. | |
| RPKM | Reads Per Kilobase per Million | La normalisation RPKM (Reads Per Kilobase per Million) a été introduite initialement pour faciliter les comparaisons entre gènes d'un même échantillon ; elle combine donc une normalisation inter et intra échantillons. Ainsi, les comptages sont corriges pour prendre en compte la taille de la librairie et la longueur des gènes. Cependant, il a été montre que la correction de la longueur des gènes a pour effet d'introduire un biais dans la variance par gène, en particulier pour les gènes faiblement exprimes. Cette méthode reste toutefois très populaire dans de nombreuses applications. | EdgeR |
| RLE | Relative Log Expression Ref : Anders and Huber, 2010. | La normalisation RLE (Relative Log Expression) a été développée dans le package Bioconductor DESeq. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différentiellement exprimés. Le facteur de normalisation pour un échantillon est obtenu en calculant pour chaque gène la médiane des ratio de ses comptages par rapport a sa moyenne géométrique entre les différents échantillons. L'idée sous-jacente est que les gènes non différentiellement exprimes doivent avoir des comptages similaires entre différents échantillons, et donc un ratio proche de 1. Si l'on suppose que la plupart des gènes ne sont pas différentiellement exprimes, la médiane des ratio constitue une estimation du facteur correctif qui doit être applique a l'ensemble des comptages. | DESeq, DESeq2, EdgeR |
| TMM | Trimmed Mean of M-values Ref : Robinson, M. and Oshlack, A. (2010). | La normalisation TMM (Trimmed Mean of M-values ) est implementée dans le package Bioconductor edgeR. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différentiellement exprimes. Le facteur TMM est calcule pour chaque échantillon, l'un d'eux étant considère comme l'échantillon de référence et les autres comme des échantillons test. Pour chaque échantillon test, le facteur TMM est la moyenne pondérée des log-ratios entre ce test et la référence, après exclusion des gènes les plus exprimes et des gènes ayant les plus forts log-ratios. D'après l'hypothèse selon laquelle il y a peu de gènes différentiellement exprimes, le facteur TMM doit être proche de 1. S'il ne l'est pas, sa valeur donne une estimation du facteur correctif à appliquer aux tailles des librairies (et pas aux comptages bruts) afin de rendre l'hypothèse vraie. | EdgeR |

From E.Corre ABiMS RNAseq Training sept 2019

# Normalisation des données

"Only the DESeq and TMM normalization methods are robust to the presence of different library sizes and widely different library compositions..." Dillies et al. 2013.

The Effective Library Size concept : TMM (edgeR) and DESeq

- Motivation: Different biological conditions express different RNA repertoires, leading to different total amounts of RNA

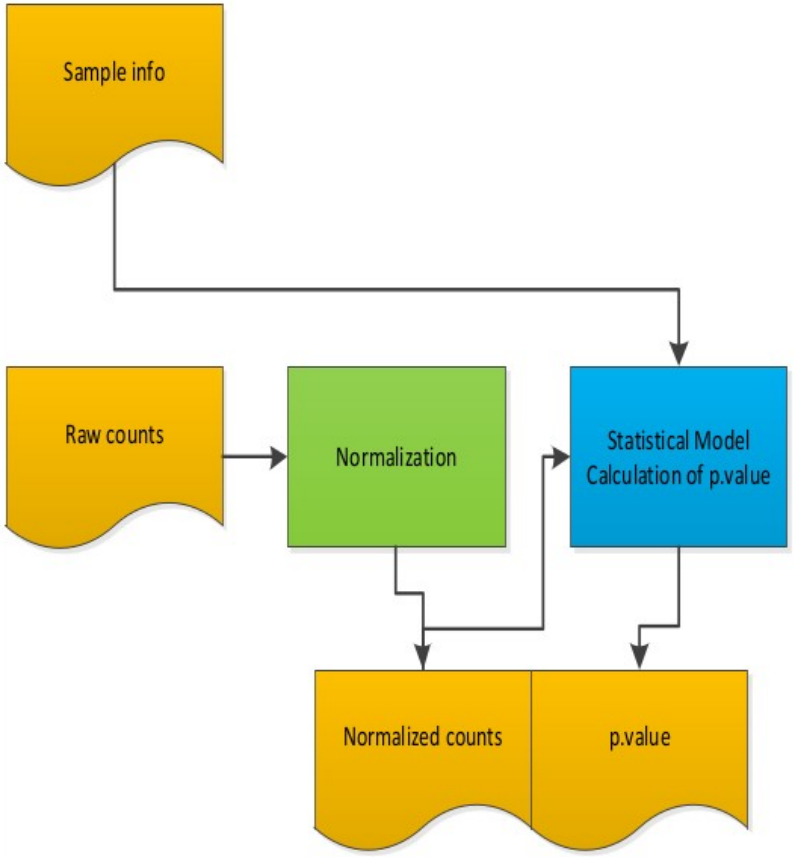
- Assumption: A majority of transcripts is not differentially expressed
  As many down- as up-regulated genes

- Method: Minimizing effect of (very) majority sequences

- Problem: ?

From Julie AUBERT, CNRS

# Statistical Model




From E.Corre ABiMS RNAseq Training sept 2019

# Statistical Model

- **limma** (i.e., voom+limma and vst+limma)
  - unaffected by outliers
  - but they required at least 3 samples per condition
- **SAMseq, ShrinkSeq** (The non-parametric)
  - top performing methods for data sets with large sample sizes
  - required at least 4-5 samples per condition
  - fold change required for statistical significance was lower → compromise the biological significance
  - Small sample sizes inaccuracies in the estimation of the mean and dispersion parameters
- **TSPM**
  - most affected by the sample size
- **DESeq, edgeR and NBPSeq**
  - showed, overall, relatively similar accuracy with respect to gene ranking
  - recommended parameters well chosen and often provide the best results
  - pre-specified FDR threshold varied considerably between the methods
  - **DESeq** : overly conservative
  - **edgeR**, **NBPSeq** : too liberal and called a larger number of false (and true) DE genes.
  - **edgeR**, **DESeq :** varying the parameters of can have large effects on the results
- **EBSeq, baySeq and ShrinkSeq** (posterior probability)
  - **baySeq** performed well under some conditions ; results were highly variable, especially when all DE genes were upregulated in one condition
  - **EBSeq** In the presence of outliers, found a lower fraction of false positives for large sample sizes not fot small sample sizes
  - **baySeq** In the presence of outliers, found a lower fraction of false positives true for small sample sizes not fot large sample sizes

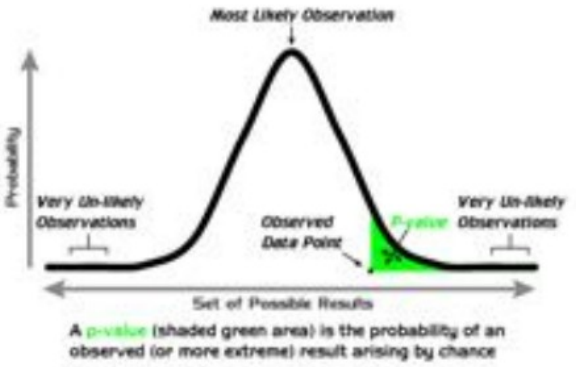
A comparison of methods for differential expression analysis of RNA-seq data. Charlotte Soneson1 and Mauro Delorenzi

# Statistical Model

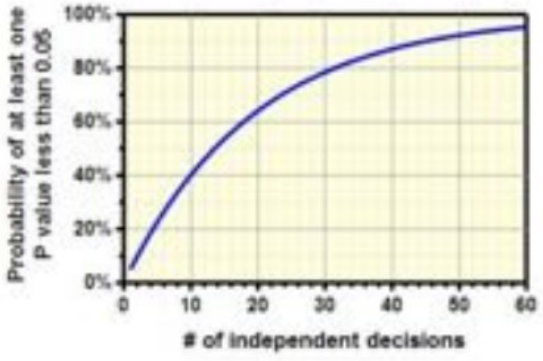
## The results

- p.value

  A P-value is a method that can be used to reject or not reject the null hypothesis. The smaller the p-value, the more unlikely the null hypothesis.




- adjusted p.value / False Discovery Rate
  - Used in multiple hypothesis testing
  - Corrections
    - Bonferroni
    - Benjamini-Hochberg (BH)

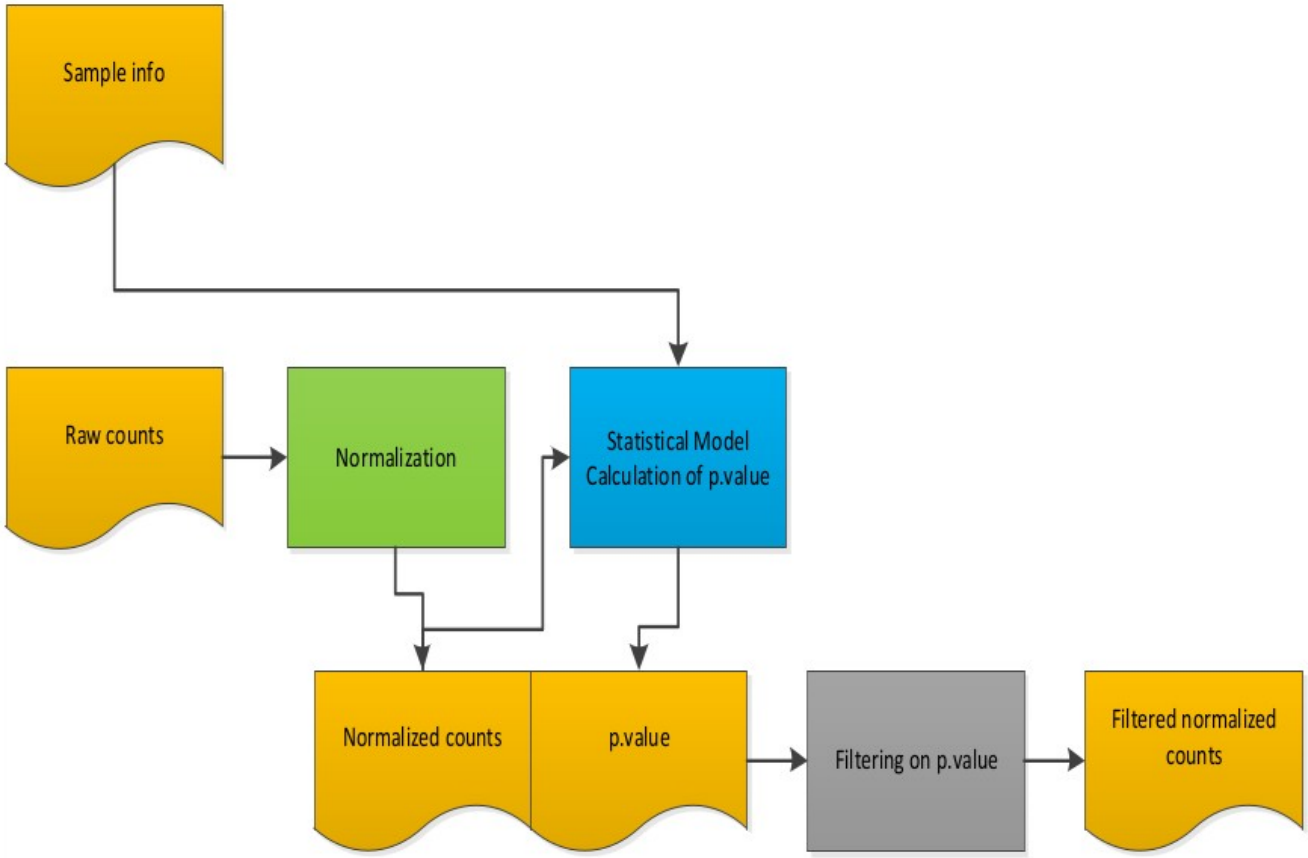



An **FDR adjusted p-value** (or q-**value**) of 0.05 implies that 5% of significant tests will result in false positives



From E.Corre ABiMS RNAseq Training sept 2019

# Statistics Models



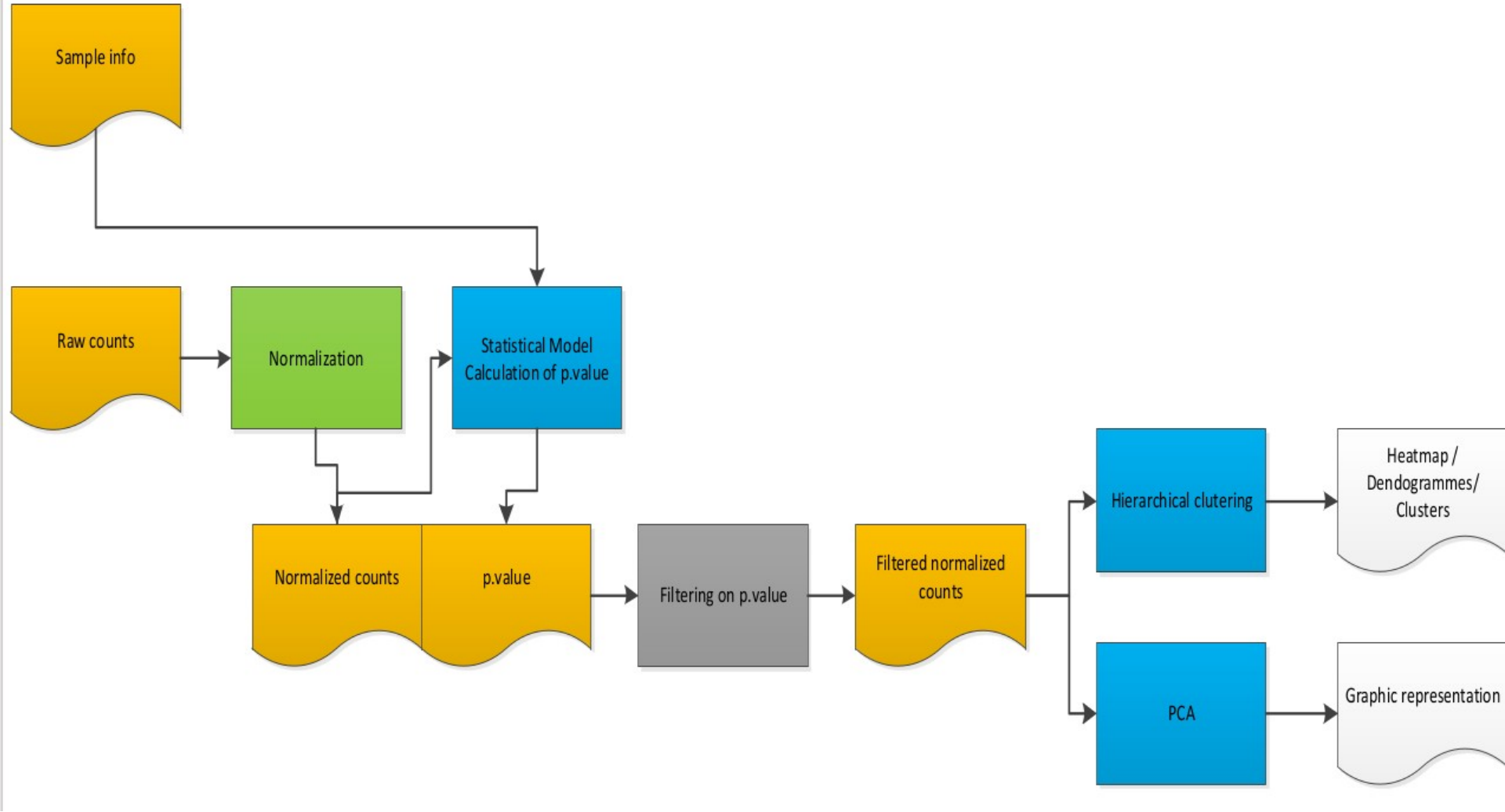

From E.Corre ABiMS RNAseq Training sept 2019

# Statistical Model : filtering by p-value and FDR

## Filtering = alpha risk

- The number alpha is the threshold value that we measure p-values against. It tells us how extreme observed results must be in order to reject the null hypothesis of a significance test.
- Must be set in advance ! 
- Ex:
  - For results with a 90% level of confidence, the value of alpha is 1 - 0.90 = 0.10.
  - For results with a 95% level of confidence, the value of alpha is 1 - 0.95 = 0.05.
  - For results with a 99% level of confidence, the value of alpha is 1 - 0.99 = 0.01.
- So:
  - alpha > pvalue → H0 is rejected → 

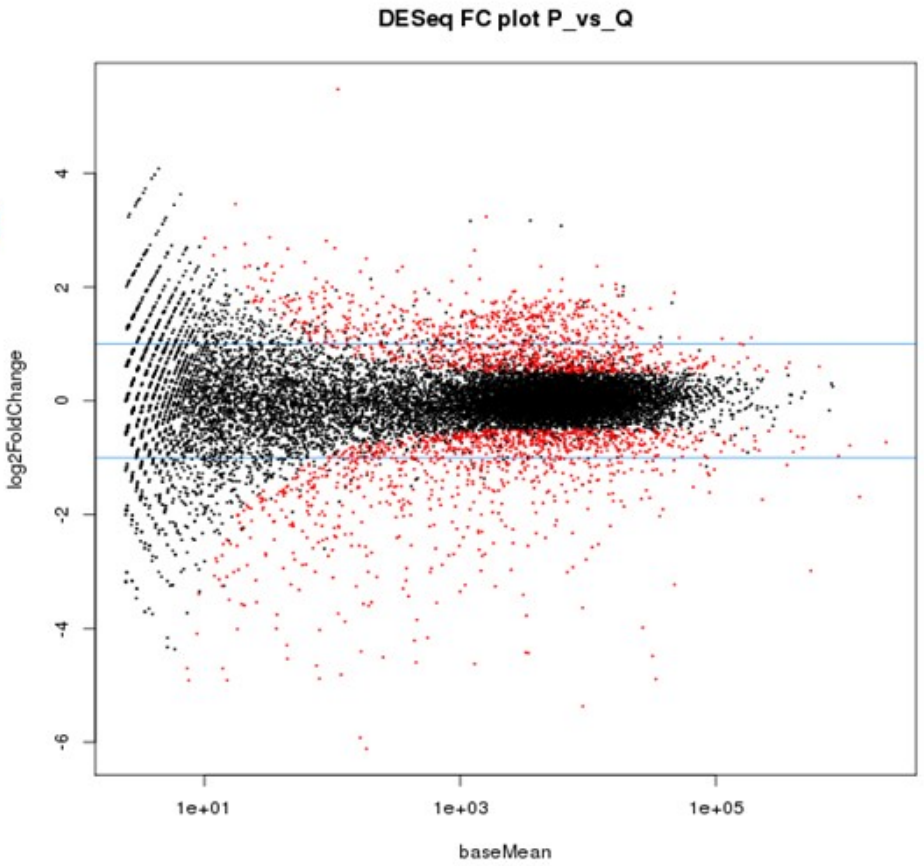
# Statistics tools




From E.Corre ABiMS RNAseq Training sept 2019

# Smear plot

Smear plot / MA plot
Pvalue adj < 0.05



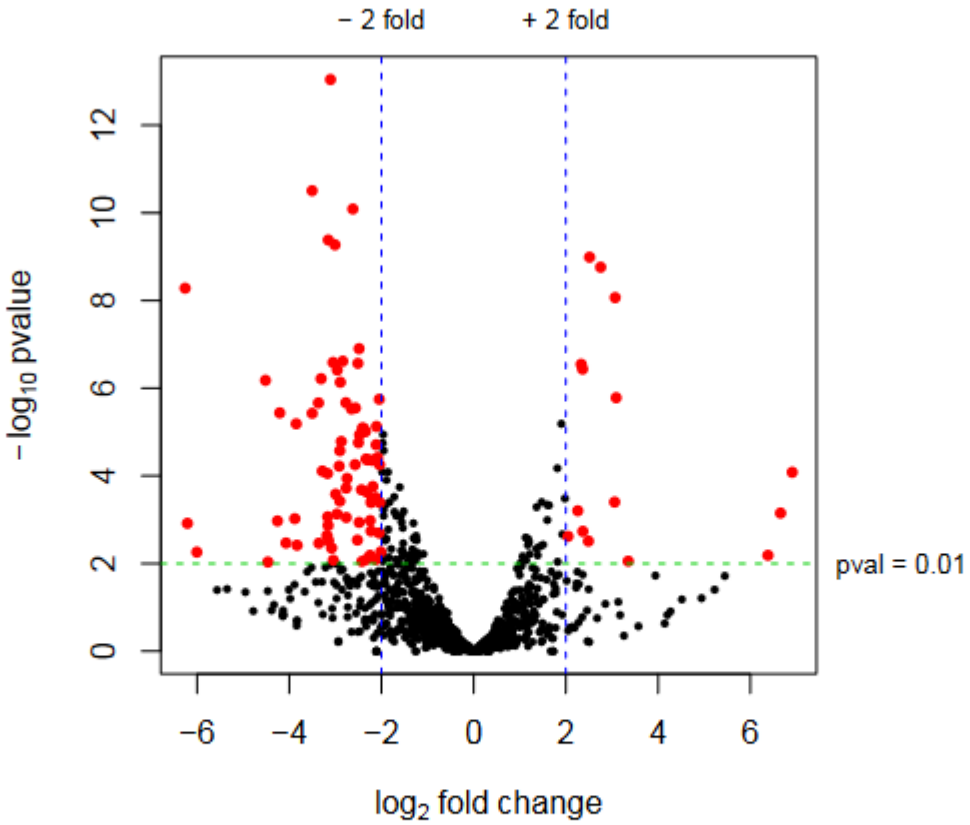

## MA plot

Le MA plot est un graphe qui était initialement utilise dans les analyses de puce a ADN. C'est un nuage de points représentant en abscisse l'expression moyenne du gène à travers les différents échantillons, et en ordonnée le log-ratio des expressions moyennes d'une condition par rapport a l'autre. En RNA-Seq, après normalisation, on s'attend a ce que les points soient repartis symétriquement autour de 0 en ordonnée (c'est-a-dire un ratio de 1).

# Volcano plot

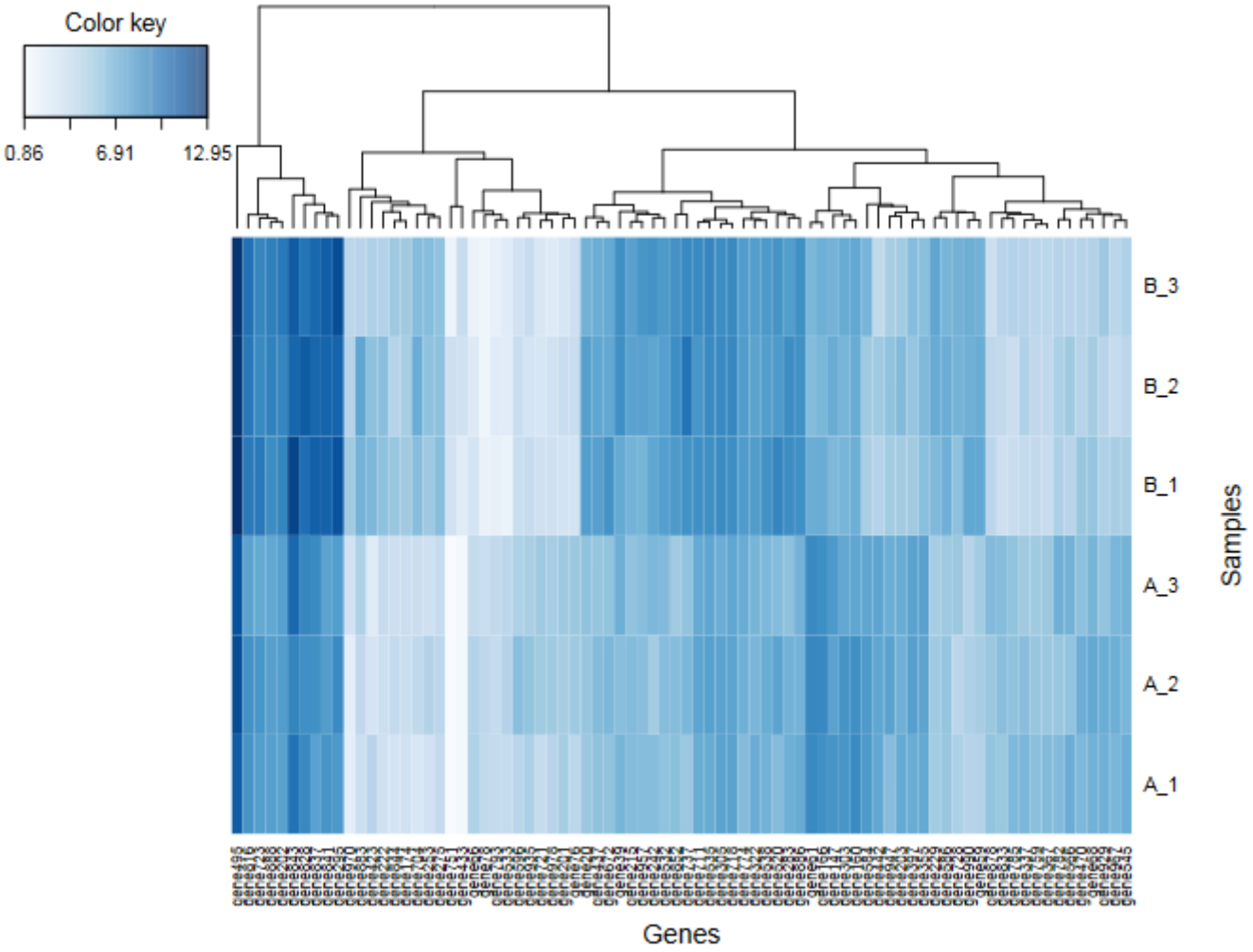
Volcano plot

Pvalue adj < 0.01




Tutorial: http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

# Hierarchical clustering / Heatmap



# Practice



Go to DE PRACTICE on our github

# Practice



Go to PIVOT on our github

# Contributeurs pour cette formation

Alexis Dereeper

Sebastien Ravel

**Christine Tranchant-Dubreuil**

Sebastien Cunnac

Gautier Sarah

**Julie Orjuela-Bouniol**

Catherine Breton

Aurore Compte

Erwan Corre

# Merci pour votre attention !