



Session de formation 2023



bioinformatics platform dedicated to the genetics and genomics of tropical and Mediterranean plants and their pathogens

réseaux
génomique
plantes
sud
Réseau
ressources
montpelliérain
Infrastructure
internationale
orienté
développement
service
calcul
multi-instituts
comptences
végétale plateforme d'analyses
communauté outils multi-instituts
s'appuie mutualisation partage



SNP detection
phylogeny structural variation
comparative genomics transcriptome assembly differential expression
GWAS pangenomics
population genetics
polypliody

Mutualisation



Cacao

Banana

Coffee

Rice

Palm

Cassava

Pseudocercospora

Magnaporthe

South Green

bioinformatics platform



4 institutes



25+



3 research units



Tools

Storage and computing
resources



400+

Trainings



Meso@LR au CINES

1090 threads :

35 standard nodes

2 bigmem nodes

1 GPU node

500 To of replicated storage

CINES

1130 threads:

30 standard node

1 supermem node

1 GPU node

150 To on 3 NAS + 210 To scratch



400+



600+ tools

Resources mutualised at Meso@LR through the
Mudis4Ls project (purchase/storage/data)

Collaborative development of tools

Genomics

- # Pangenomic

- # Gene families

- # Comparative

- ## Phylogeny

- # Assemblies

- # Annotation

- # Data mining

Diversity exploration

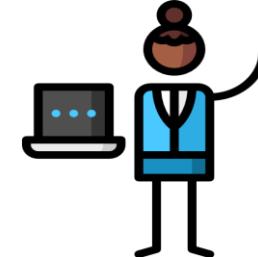
- # genotype manipulation

- ## mosaic manipulation

Metagenomic



<https://github.com/SouthGreenPlatform/>



+20
tools

web applications (16)

visualisation (8)

workflows(5)

packages (4)



Plant & Health Bioinformatics Platform



<https://bioinfo.ird.fr/>



AURORE
COMTE

JACQUES
DAINAT

ALEXIS
DEREEPER

BRUNO
GRANOUILLAC

JULIE
ORJUELA-

NDOMASSI
TANDO

CHRISTINE
TRANCHANT

bioinfo@ird.fr



[@I tropBioinfo](https://twitter.com/I tropBioinfo)



Florian Charriat
Antoni Exbrayat



Guilhem Sempere



Bruno Granouillac
Jacques Dainat



Nicolas Fernandez



Thomas Denecker

And more collaborators !

South Green

bioinformatics platform



Larmande Pierre
Orjuela-Bouniol Julie
Sabot François
Tando Ndomassi
Tranchant-Dubreuil Christine

Comte Aurore
Dereeper Alexis
Ravel Sébastien



Bocs Stephanie
Boizet Alice
De Lamotte Fredéric
Droc Gaetan
Dufayard Jean-François
Hamelin Chantal
Martin Guillaume
Pitollat Bertrand
Ruiz Manuel
Sarah Gautier
Summo Marilyne



Rouard Mathieu
Guignon Valentin
Catherine Breton



Sempere Guilhem

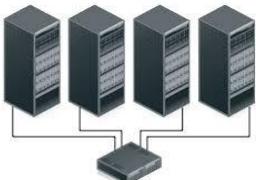


South Green bioinformatics platform

Workflow manager



HPC and trainings....



<https://github.com/SouthGreenPlatform>

Genome Hubs & Information System

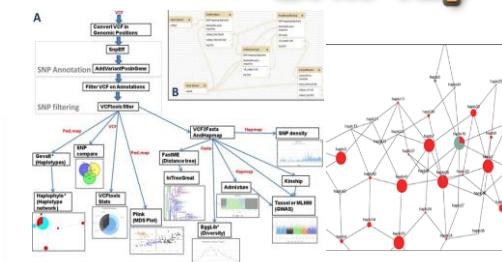


SNPs and Indels



Family Id	Family Name	Number of sequences	Status
GP000016	Cytochrome P450 superfamily	6942	green
GP000017	AP2/EREBP transcription factor family: ERF/DREB group (partial)	5142	green
GP000020	NAC transcription factor family	4574	green
GP000028	MADS transcription factor family		
GP000018	Haem peroxidase superfamily		
GP000066	General substrate transporter superfamily		
GP000022	Sulfatase-like Serine Proteases family		
GP000019	NPF/NRT1/PTR FAMILY		

Gene families



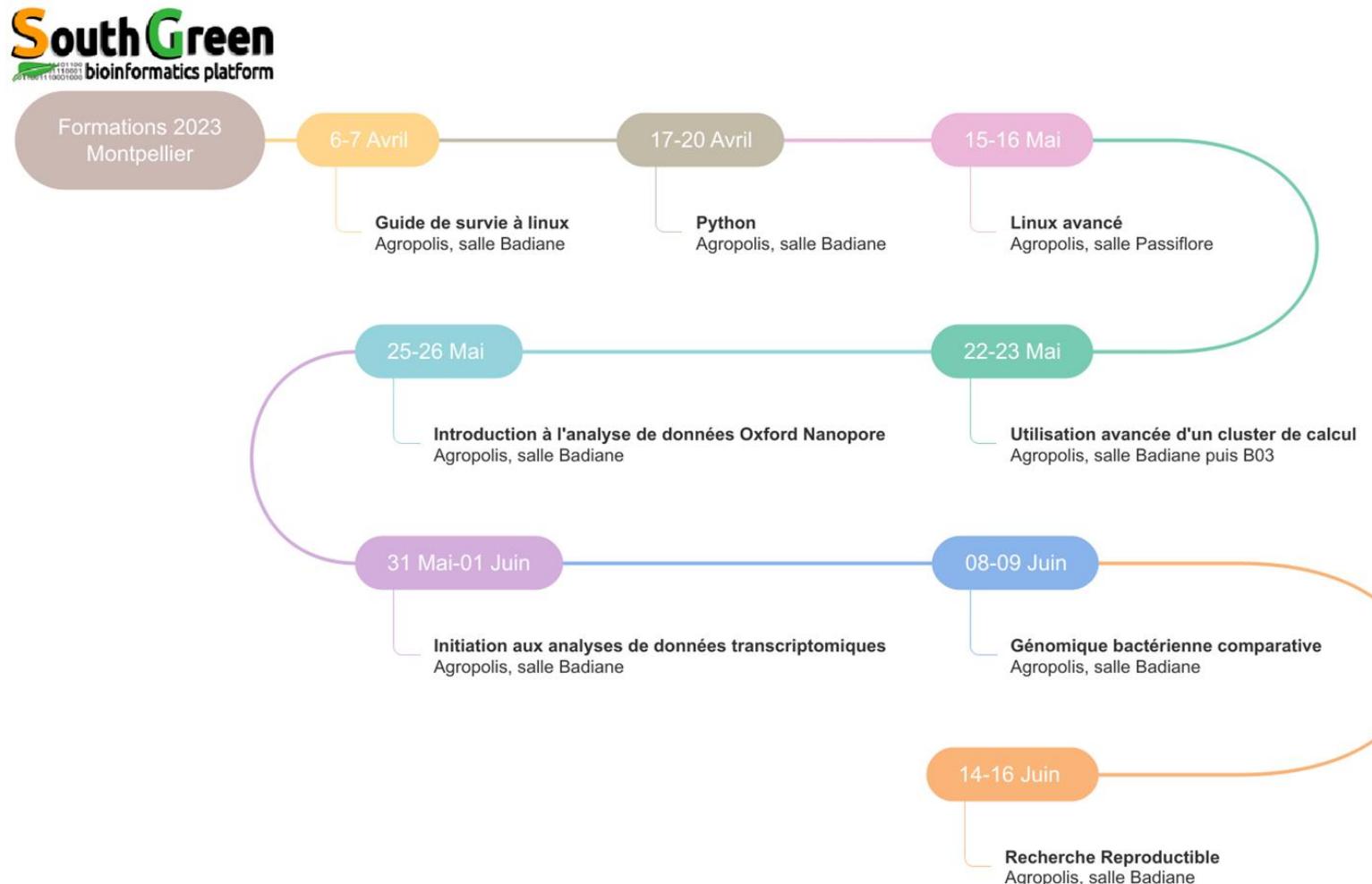
SNiPlay



The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics, Current Plant Biology, 2016

South Green

bioinformatics platform



Modules de formation 2023

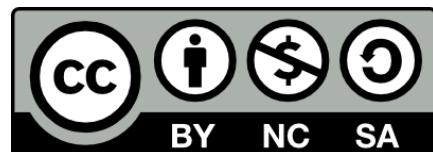
- Toutes nos formations :
<https://southgreenplatform.github.io/trainings/>
- Topo & TP : [RNAseq](#)
- Environnement de travail : [Logiciels à installer](#)



Initiation aux analyses de données transcriptomiques

www.southgreen.fr

<https://southgreenplatform.github.io/trainings>



Objectifs

- Connaître et manipuler des packages/outils disponibles pour la recherche de gènes différentiellement exprimés
- Réfléchir sur les différentes techniques de normalisation des données
- Déetecter les gènes différentiellement exprimés entre 2 conditions

Applications

- Mapping and counting using STAR , HTCseq Count
- Differential expression analysis: EdgeR, Deseq2 : DIANE

Pourquoi faire du RNAseq ?

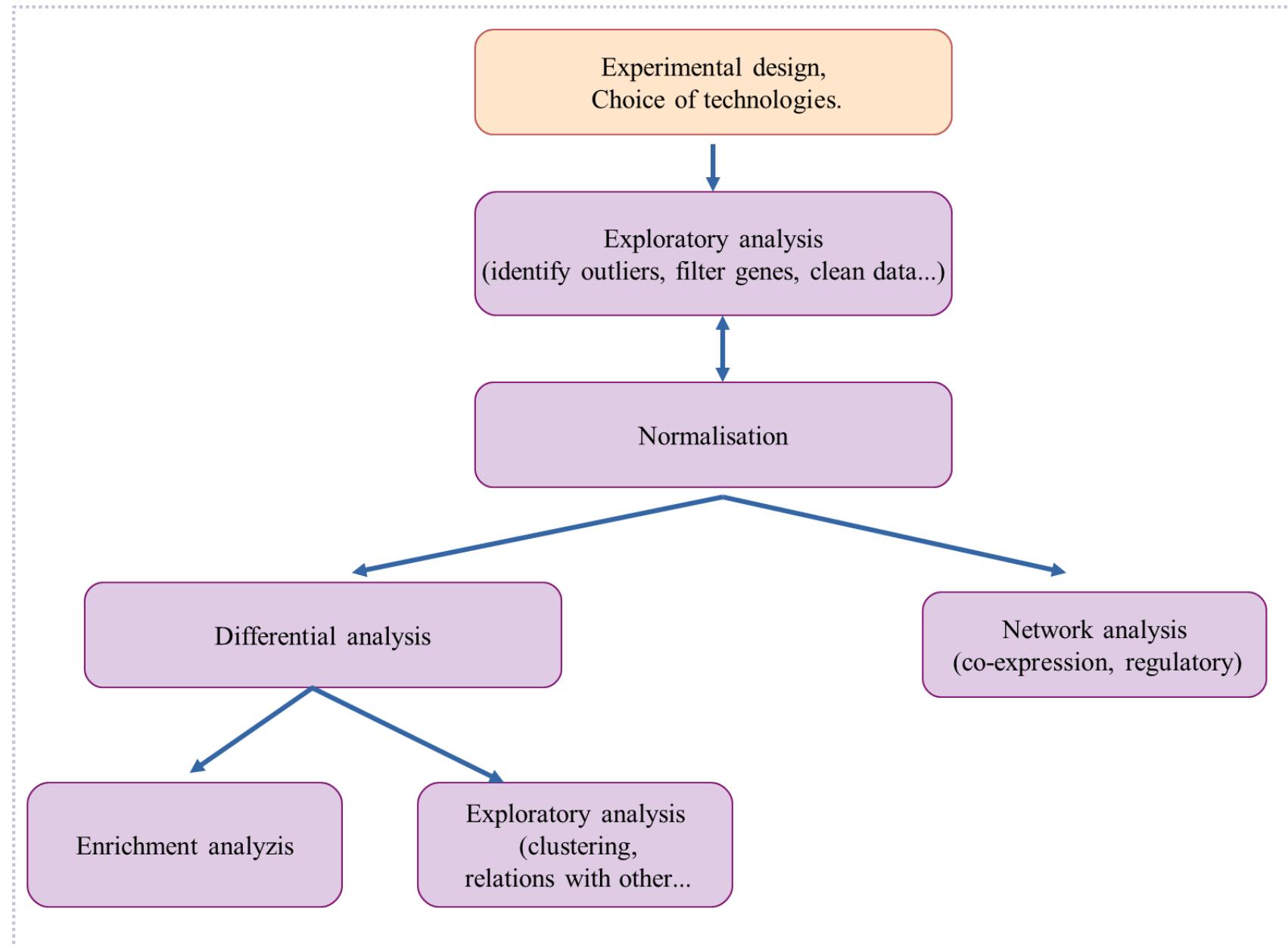
- **L'analyse d'expression différentielle** (différence d'expression dans des conditions précises) au niveau transcriptomique.
- Etude de **l'épissage alternatif** (isoformes) et recherche de nouveaux transcrits.
- **Recherche d'allèles spécifiques** et quantification de leur expression.
- **Construction d'un transcriptome** de novo pour les organismes non modèles.

Choix technologiques

- **Déplétion / enrichissement :**
 - Déplétion des ARNr (eucaryote ou procaryote)
 - Sélection des transcrits poly-A (eucaryotes)
- **Séquençage directionnel**
 - Dans le cas des études ARN anti-sens
- **Multiplexage**
 - Ajouts de séquences tags afin de grouper plusieurs échantillons à séquencer sur une même piste de flowcell.

1- Design experimental

Experimental design



Experimental design

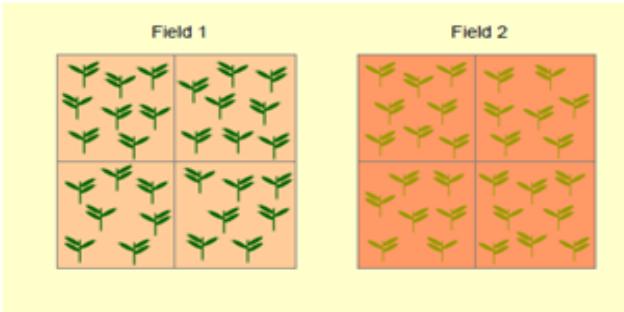
Basic experiment : trouver les différences entre condition contrôle/traitée



control group plant treated group plant

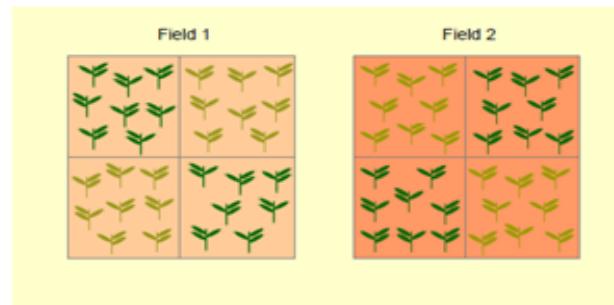


Mauvais plan expérimental : les plantes traitées sont dans un champs et les contrôles dans un autre.



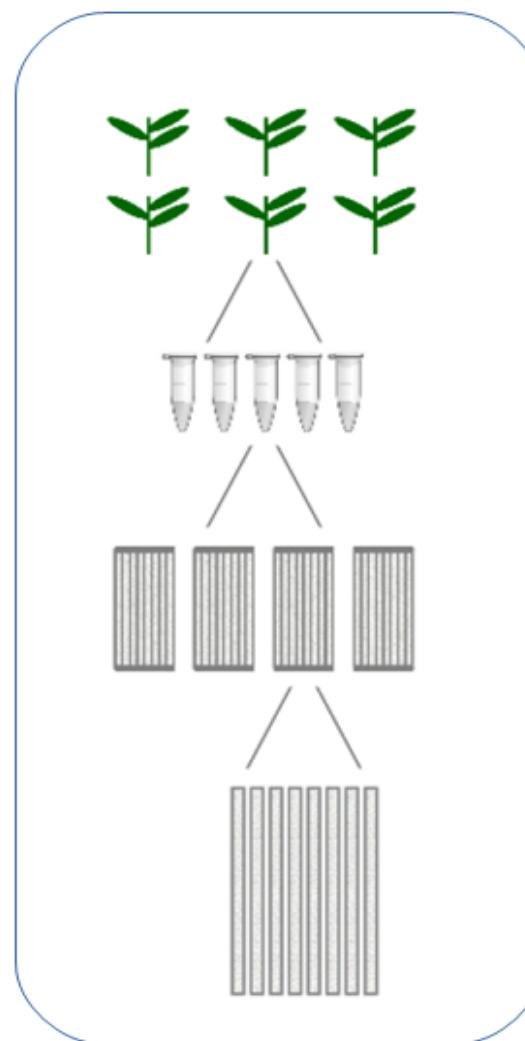
Pas de possibilité de différencier l'effet champs de l'effet traitement

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 preparation

2 – Variations techniques :
effet de la préparation
des librairies

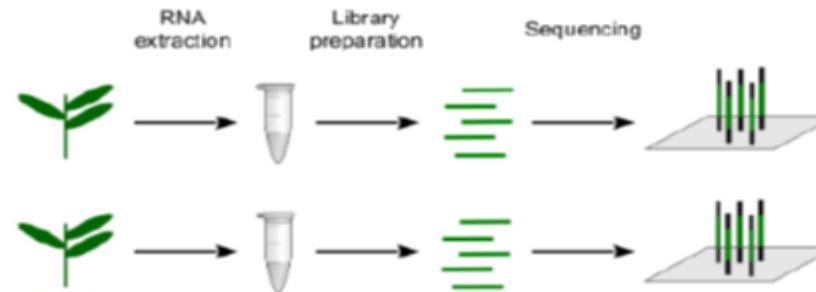
cDNA on lane of flowcell

3 – Variation techniques : effet des
lane et des flowcell

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

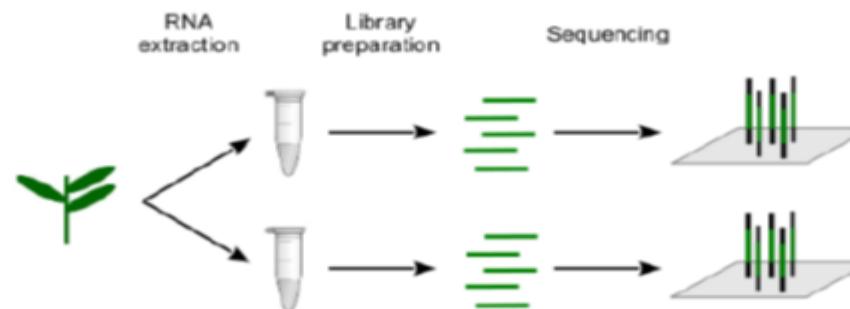
Experimental design

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.

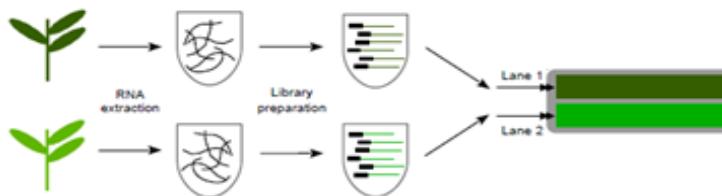
- Plusieurs extractions d'une même échantillon
- Plusieurs séquençages d'une même librairie



Experimental design

Échantillons séquencés sur deux lanes différentes.

- l'effet lane ne peut pas être mesuré mais la comparaison entre échantillon est préservée.



Exemple :

2 répliquats biologiques par condition et
4 répliquats techniques par répliquats
biologiques

Flow cell 1				Flow cell 2				Flow cell 3				Flow cell 4			
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
C ₁₁	C ₂₁	T ₁₁	T ₂₁	C ₁₂	C ₂₂	T ₁₂	T ₂₂	C ₁₃	C ₂₃	T ₁₃	T ₂₃	C ₁₄	C ₂₄	T ₁₄	T ₂₄

1

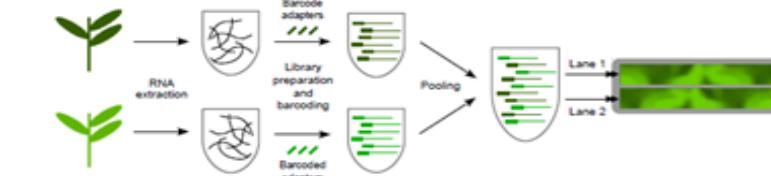
2

3

4

Multiplexes RNAseq plan d'expérimentation :

- Les fragments d'ADN sont barcodés, donc plusieurs échantillons sont séquencés sur la même lane



Exemple :

2 répliquats biologiques par condition et
4 répliquats techniques par répliquats
biologiques répartis sur 4 flowcell

Flow cell 1				Flow cell 2				Flow cell 3				Flow cell 4			
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
C ₁₁₁	C ₁₂₁	C ₁₃₁	C ₁₄₁	C ₂₁₁	C ₂₂₁	C ₂₃₁	C ₂₄₁	T ₁₁₁	T ₁₂₁	T ₁₃₁	T ₁₄₁	T ₂₁₁	T ₂₂₁	T ₂₃₁	T ₂₄₁
T ₂₁₂	T ₂₂₂	T ₂₃₂	T ₂₄₂	T ₁₁₂	T ₁₂₂	T ₁₃₂	T ₁₄₂	T ₂₁₂	T ₂₂₂	T ₂₃₂	T ₂₄₂	T ₁₁₃	T ₁₂₃	T ₁₃₃	T ₁₄₃
C ₂₁₃	C ₂₂₃	C ₂₃₃	C ₂₄₃	T ₁₁₃	T ₁₂₃	T ₁₃₃	T ₁₄₃	T ₂₁₃	T ₂₂₃	T ₂₃₃	T ₂₄₃	T ₁₁₄	T ₁₂₄	T ₁₃₄	T ₁₄₄
T ₂₁₄	T ₂₂₄	T ₂₃₄	T ₂₄₄	T ₁₁₄	T ₁₂₄	T ₁₃₄	T ₁₄₄	T ₂₁₄	T ₂₂₄	T ₂₃₄	T ₂₄₄				

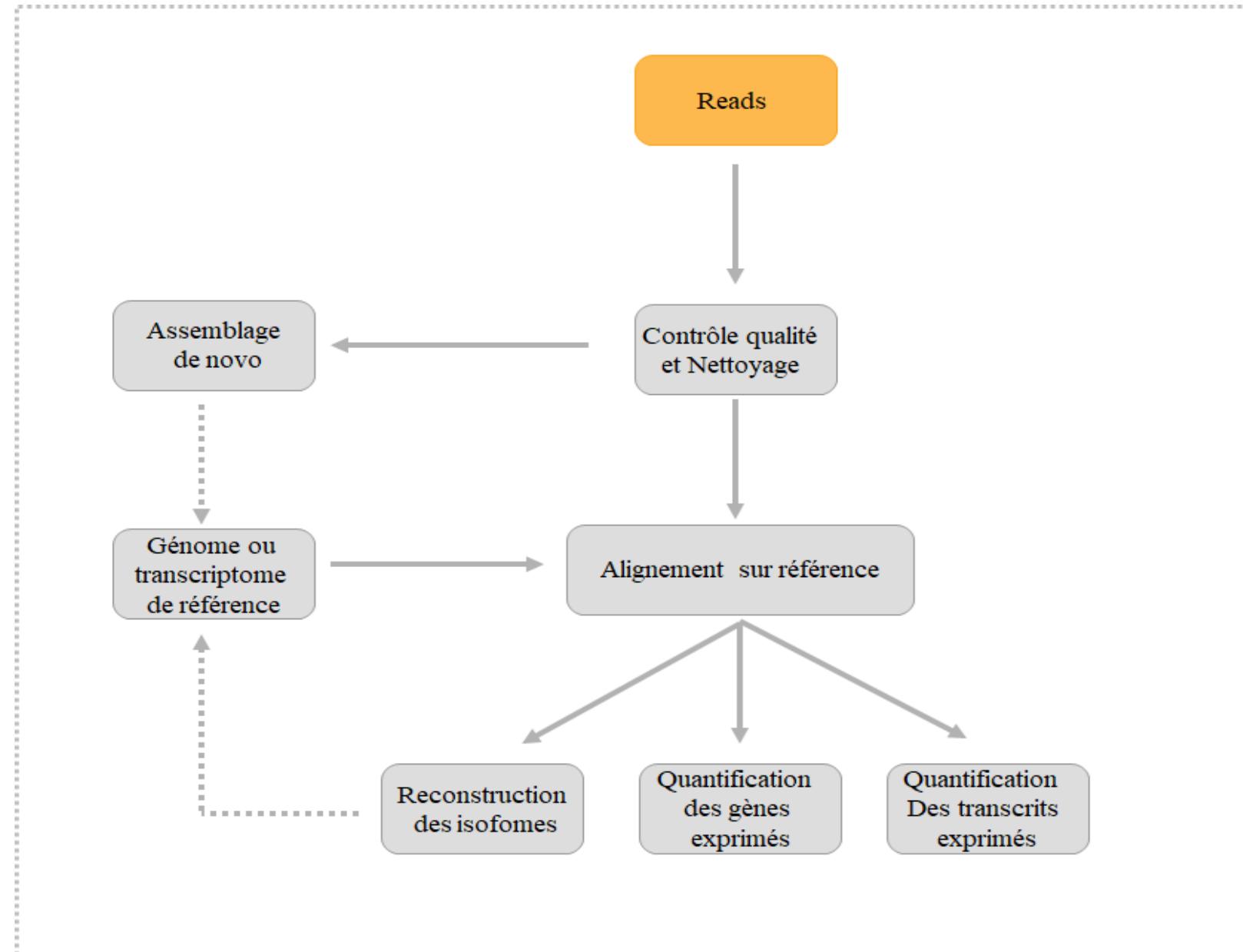
1

2

3

4

Experimental design



RNA-seq data analysis workflow

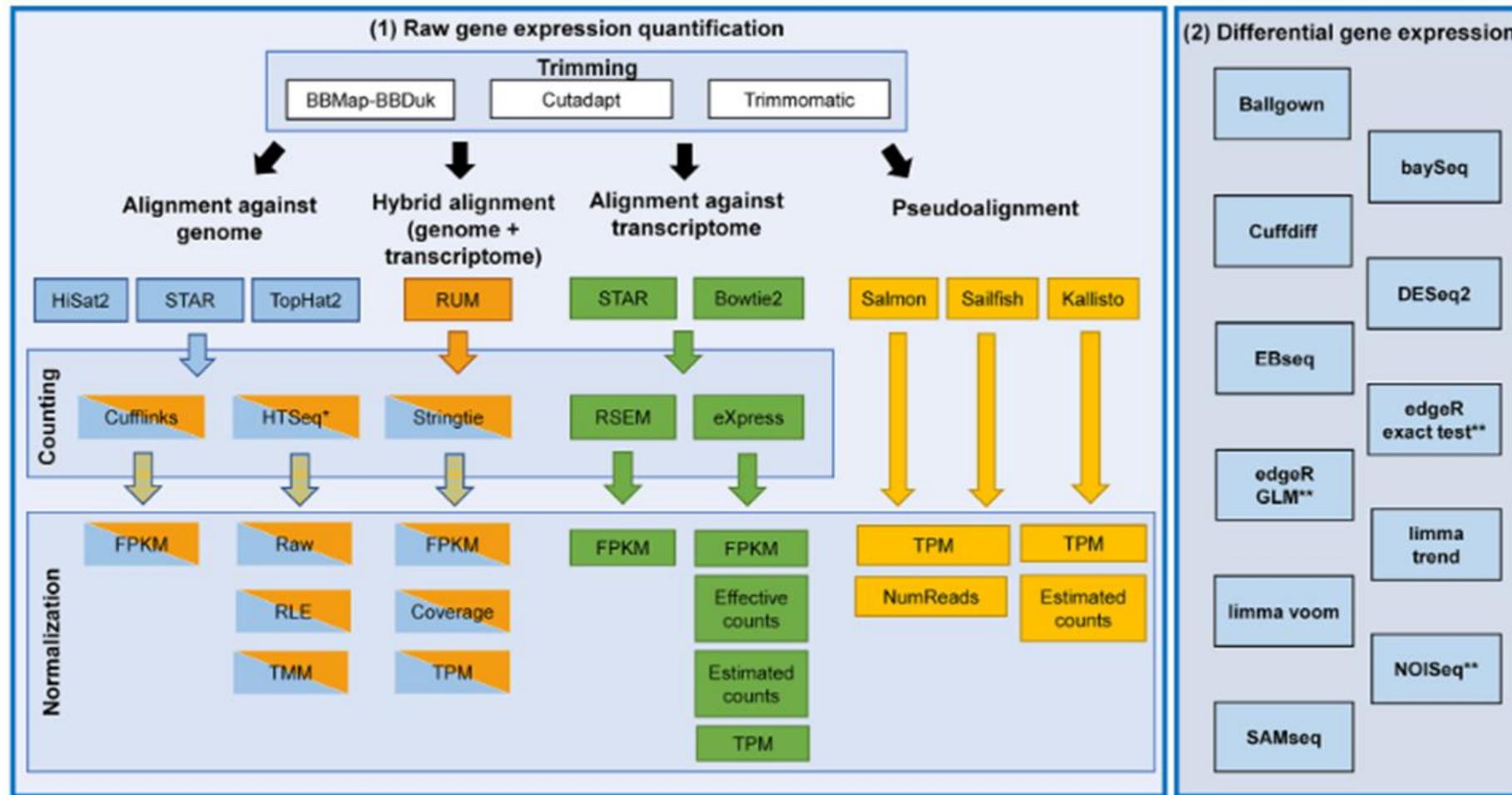
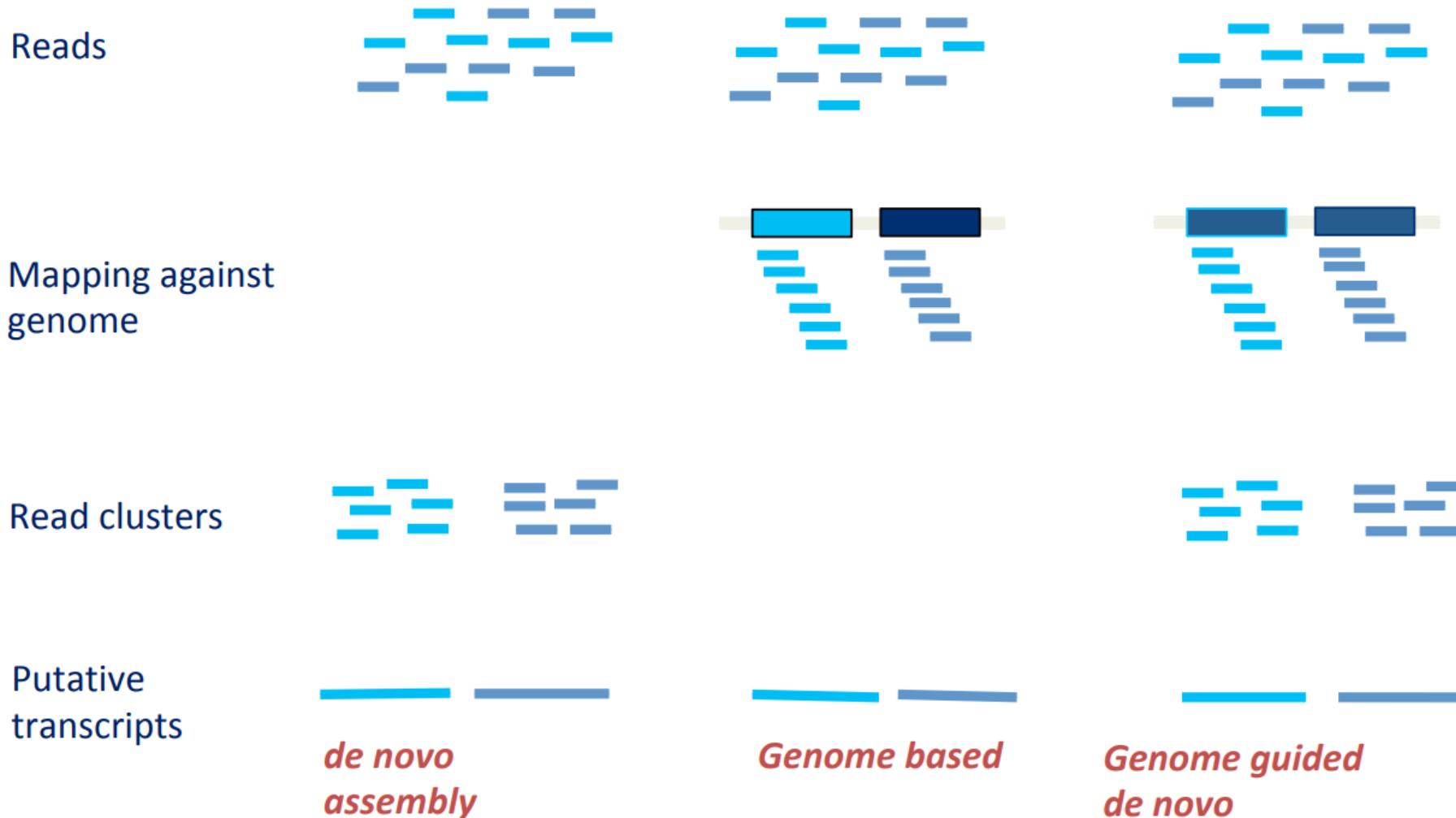


Figure 1. RNA-seq analysis workflow. Left panel (1) represents the raw gene expression quantification workflow. Every box contains the algorithms and methods used for the RNA-seq analysis at trimming, alignment, counting, normalization and pseudoalignment levels. The right panel (2) represents the algorithms used for the differential gene expression quantification. *HTSeq was performed in two modes: union and intersection-strict. **EdgeR exact test, edgeR GLM and NOISeq have internally three normalization techniques that were evaluated separately.

2- Des reads aux transcripts

Des reads aux transcripts



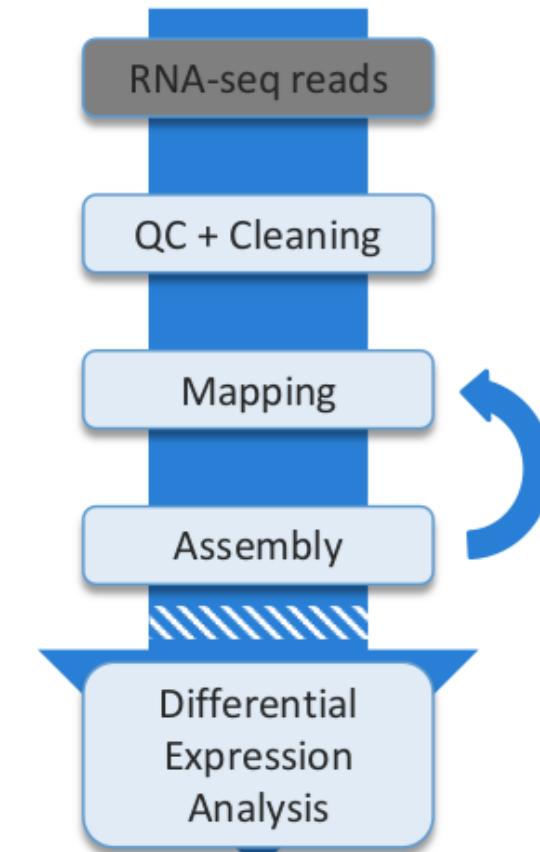
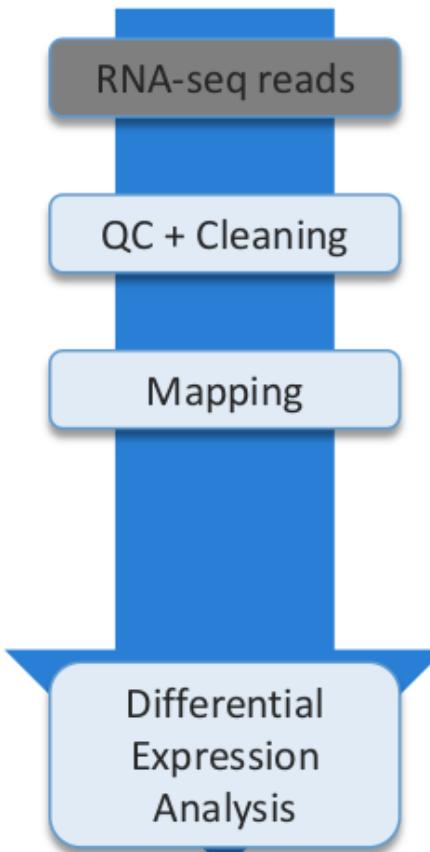
Vision Global of RNAseq Analysis

- Reference genome
- Reference transcriptome

- Reference genome
- No reference transcriptome

Non discovery mode

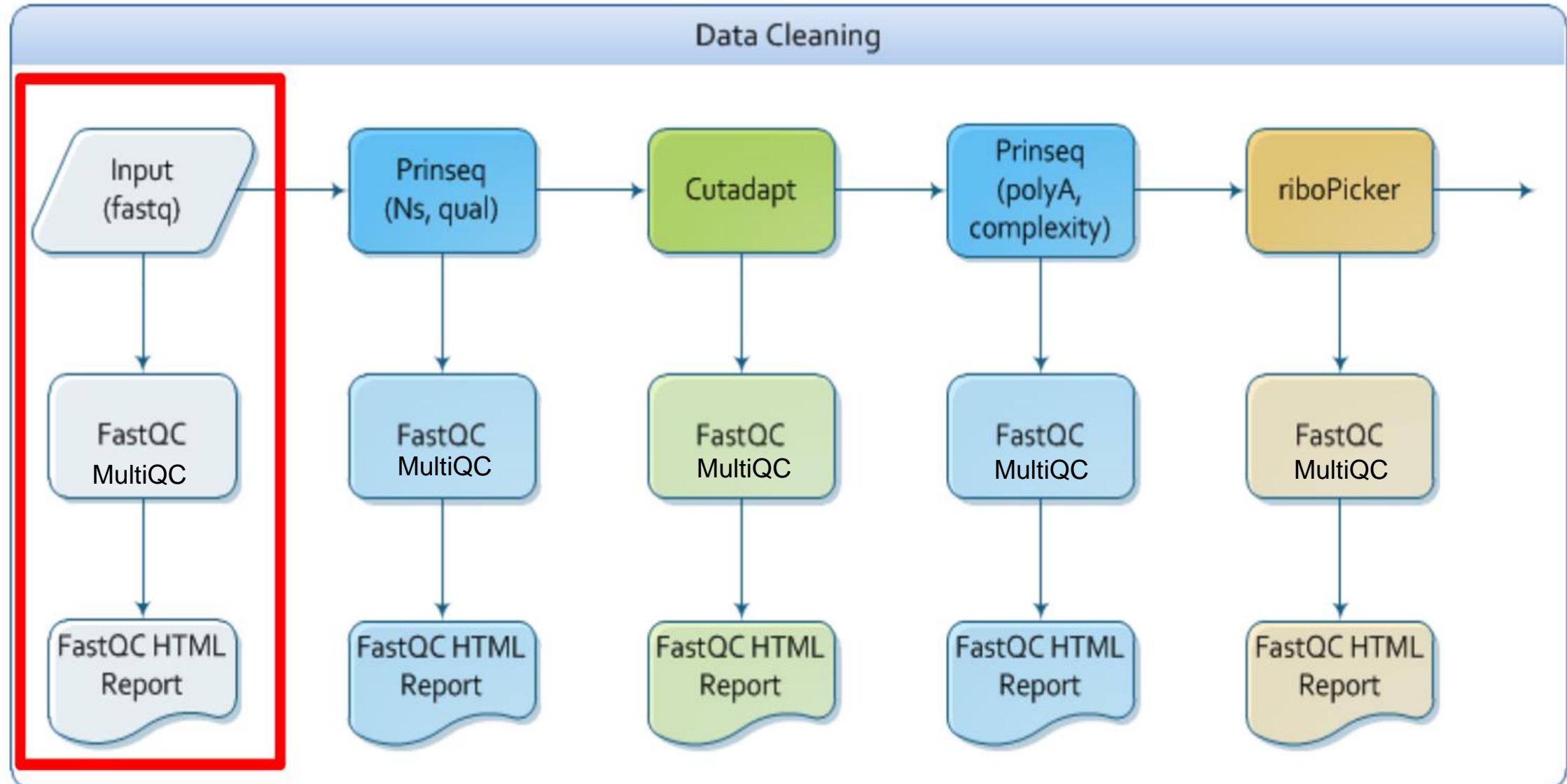
Discovery mode



Vérification de la qualité des données NGS

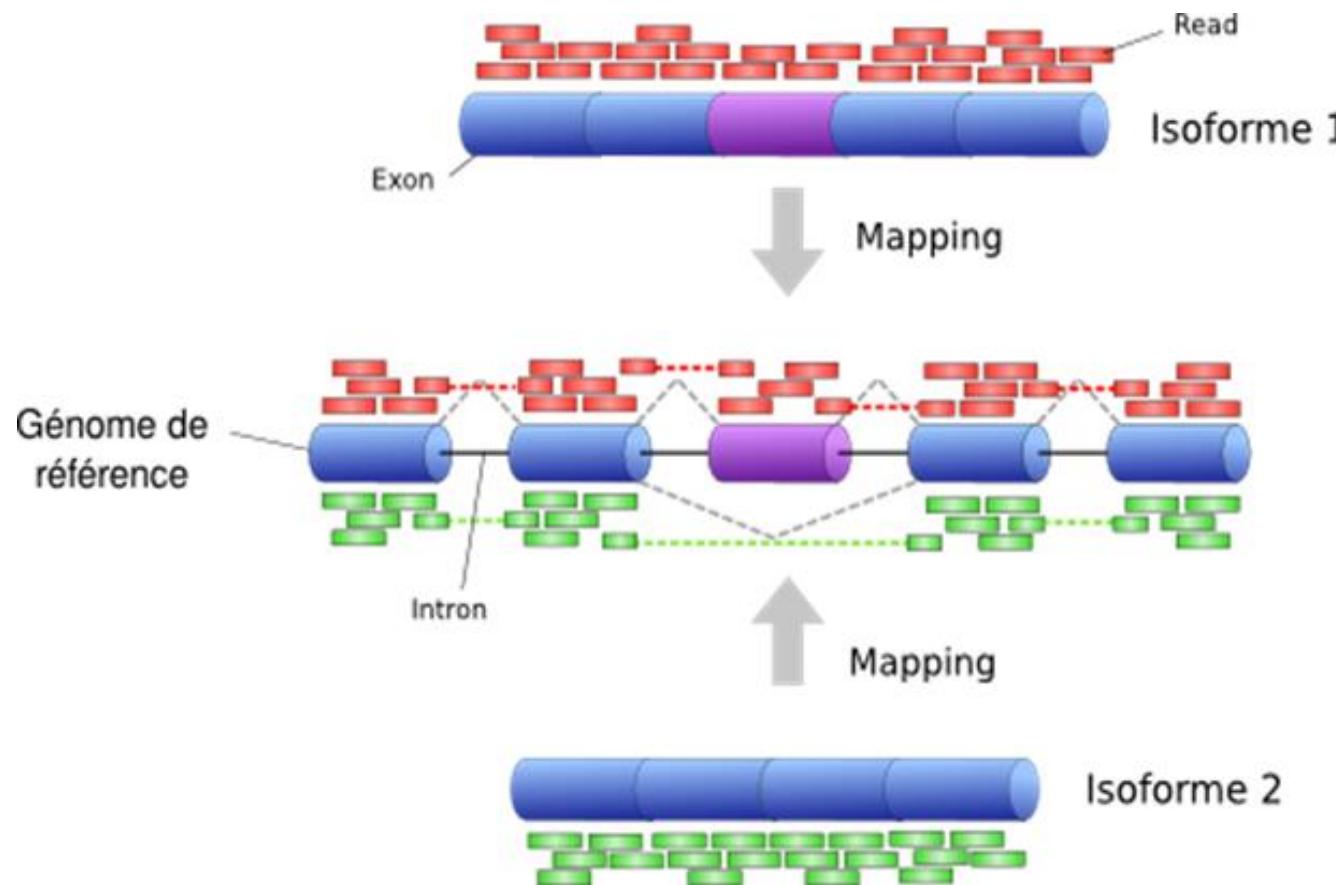
	Problème	Pourquoi les éliminer?	Outils
Sequences biases	Ns, mauvaise qualité des nucléotides, biases hexamères (random priming)	Pour éliminer des erreurs de sequencing. Désastreux pour la plupart des assembleurs	PRINSEQ2 FASTX Toolkit <i>Trimmomatic</i>
Adaptors and primers	Peuvent être trouvés dans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

Nettoyage des reads



Mapping sur référence génomique

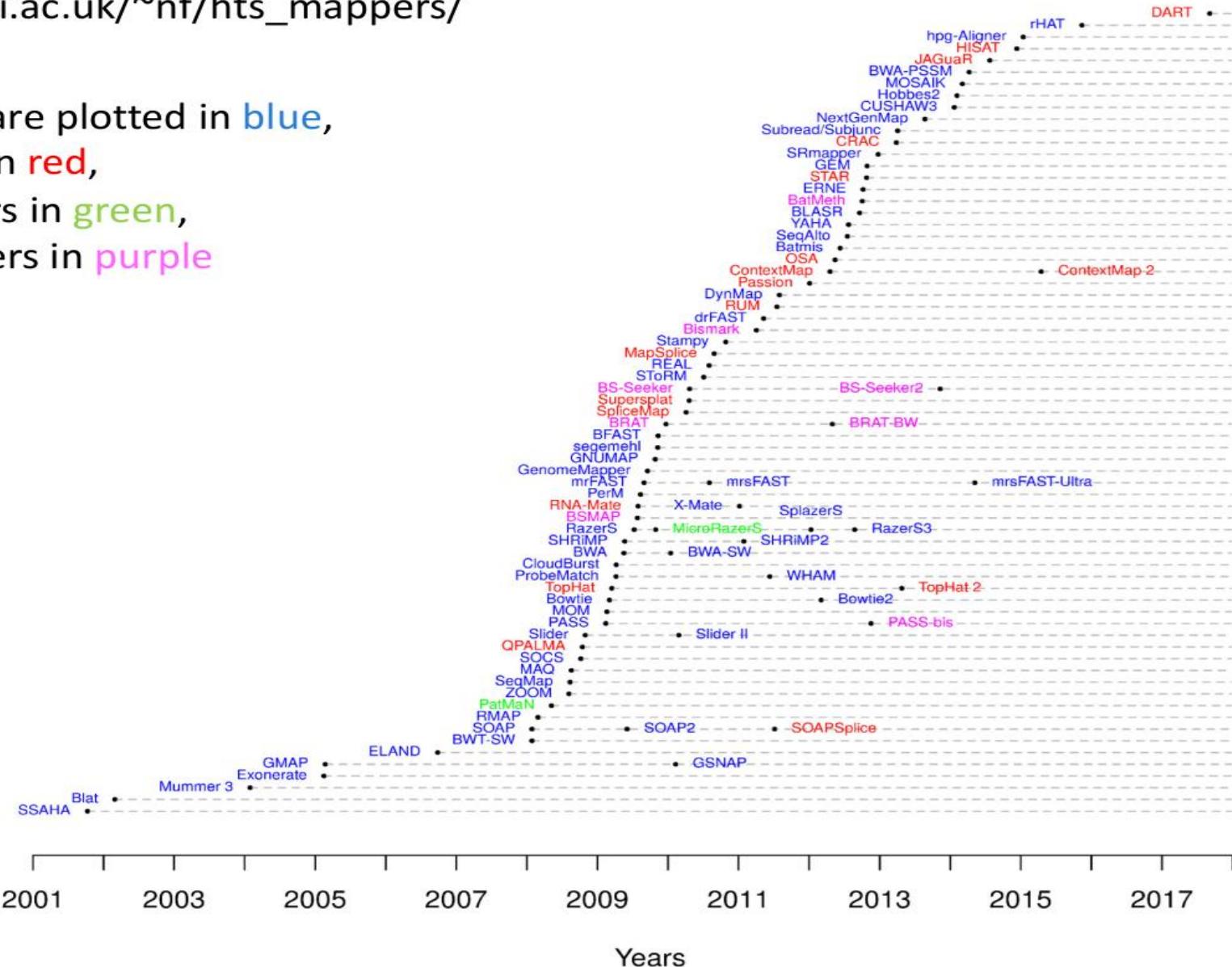
- Permet la mise en évidence d'isoformes
- Aide à l'annotation structurale du génome



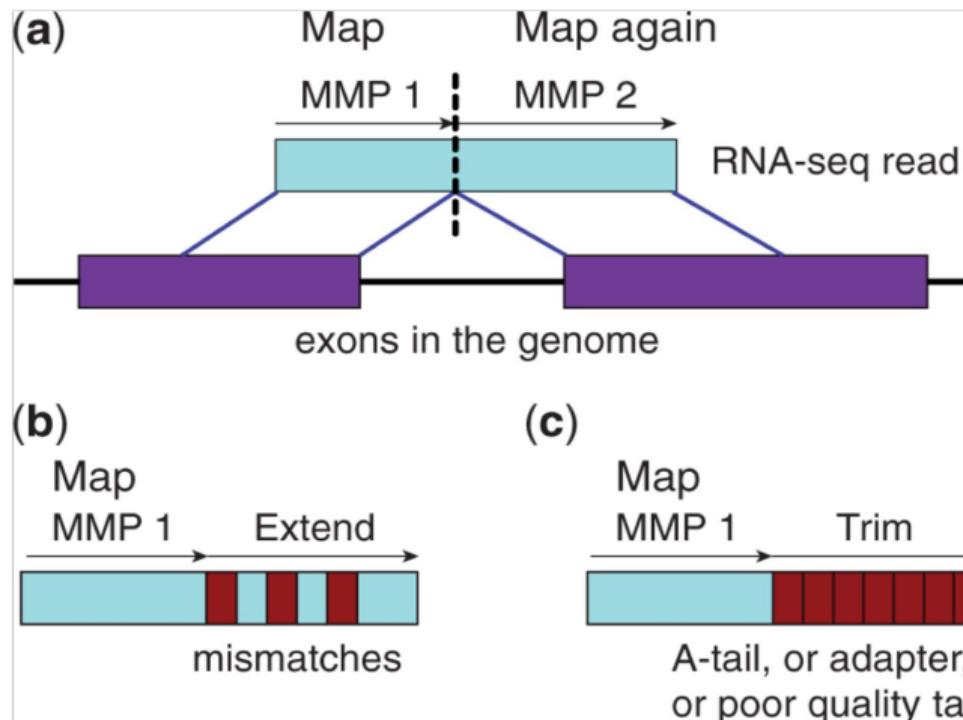
Choix de l'outil de mapping

http://www.ebi.ac.uk/~nf/hts_mappers/

DNA mappers are plotted in blue
RNA mappers in red,
miRNA mappers in green,
bisulfite mappers in purple



1st step Maximum Mapability Prefix search
-> no mismatches



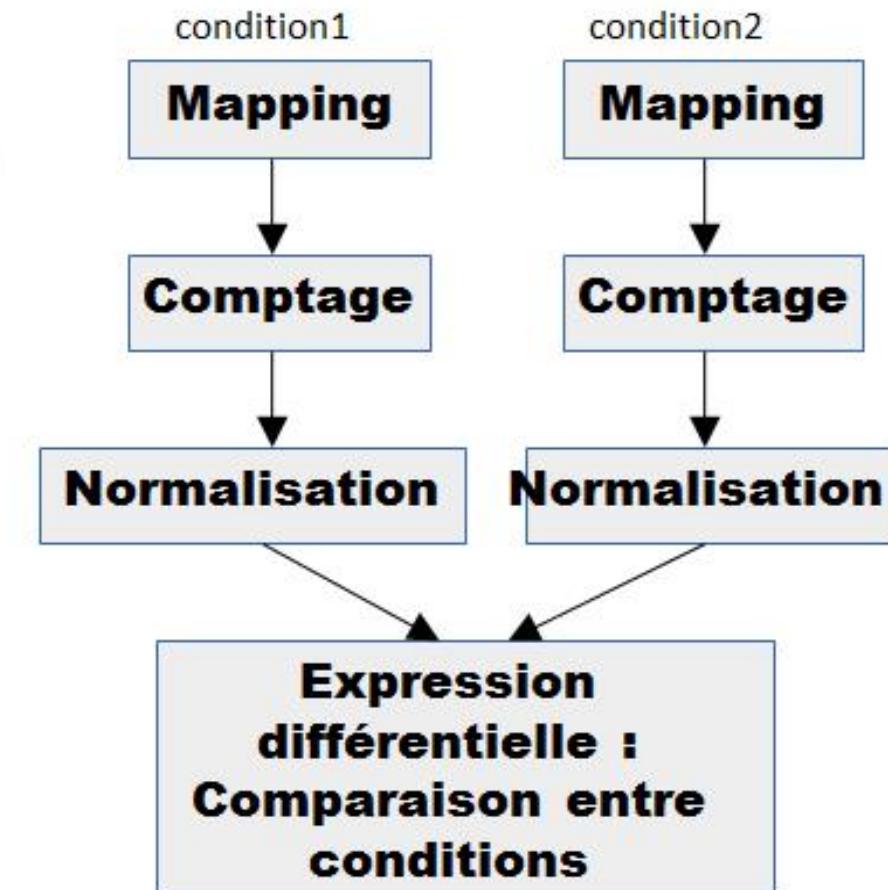
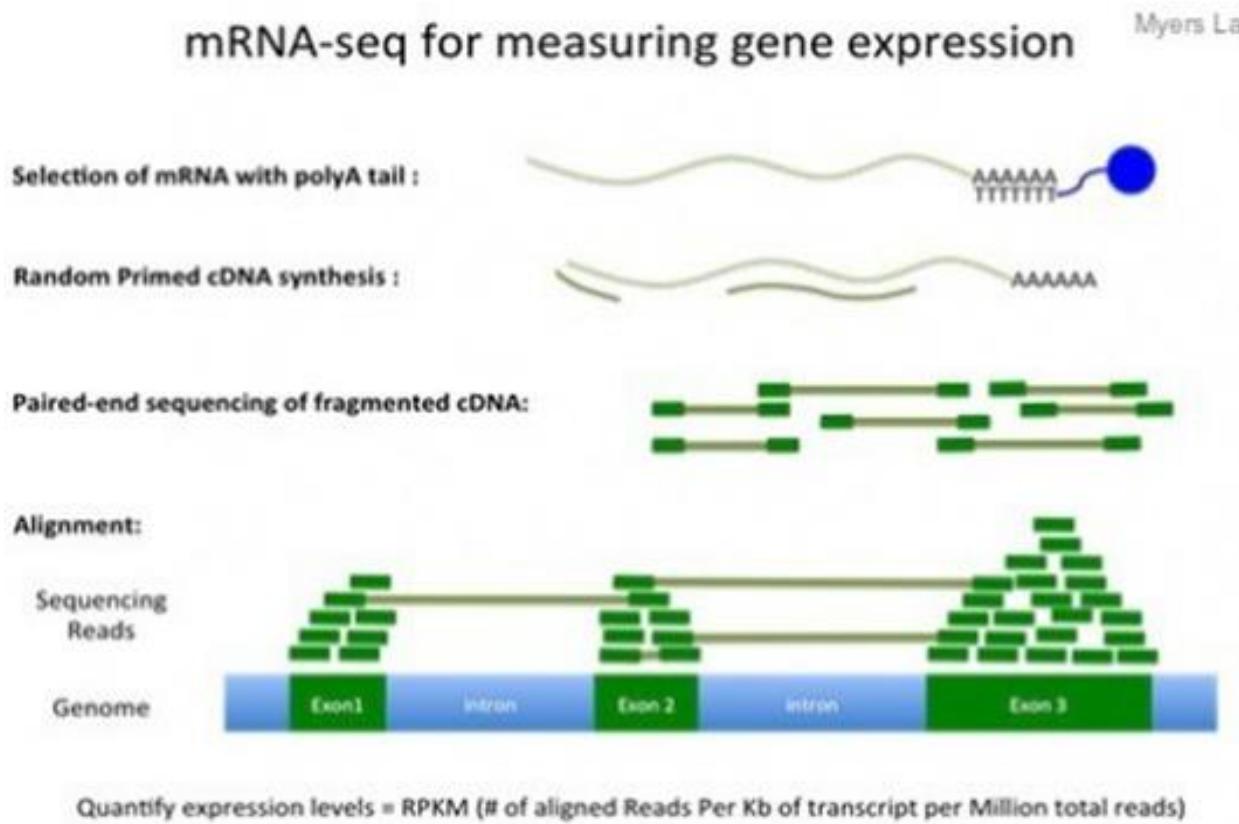
2nd step :
At the second stage STAR switches
MMPs to generate read-level
alignments that (contrary to MMPs)
can contain mismatches and indels.

STAR is extremely fast but requires
a substantial amount of RAM to run
efficiently.

SAM format

3- Comptage

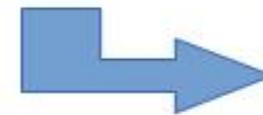
Principe du comptage des reads



Choix de l'outil de comptage

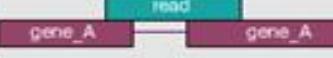
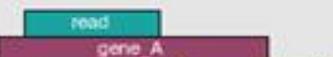
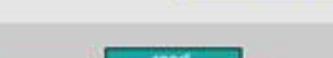
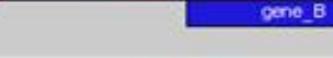
1) Si le mapping a été fait sur un génome de référence annoté

=> Utilisation de HTSeq-count
(prend en entrée l'annotation GFF)



2) Si le mapping a été fait sur un transcriptome de référence

=> samtools idxstats

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

gtf/gff Format

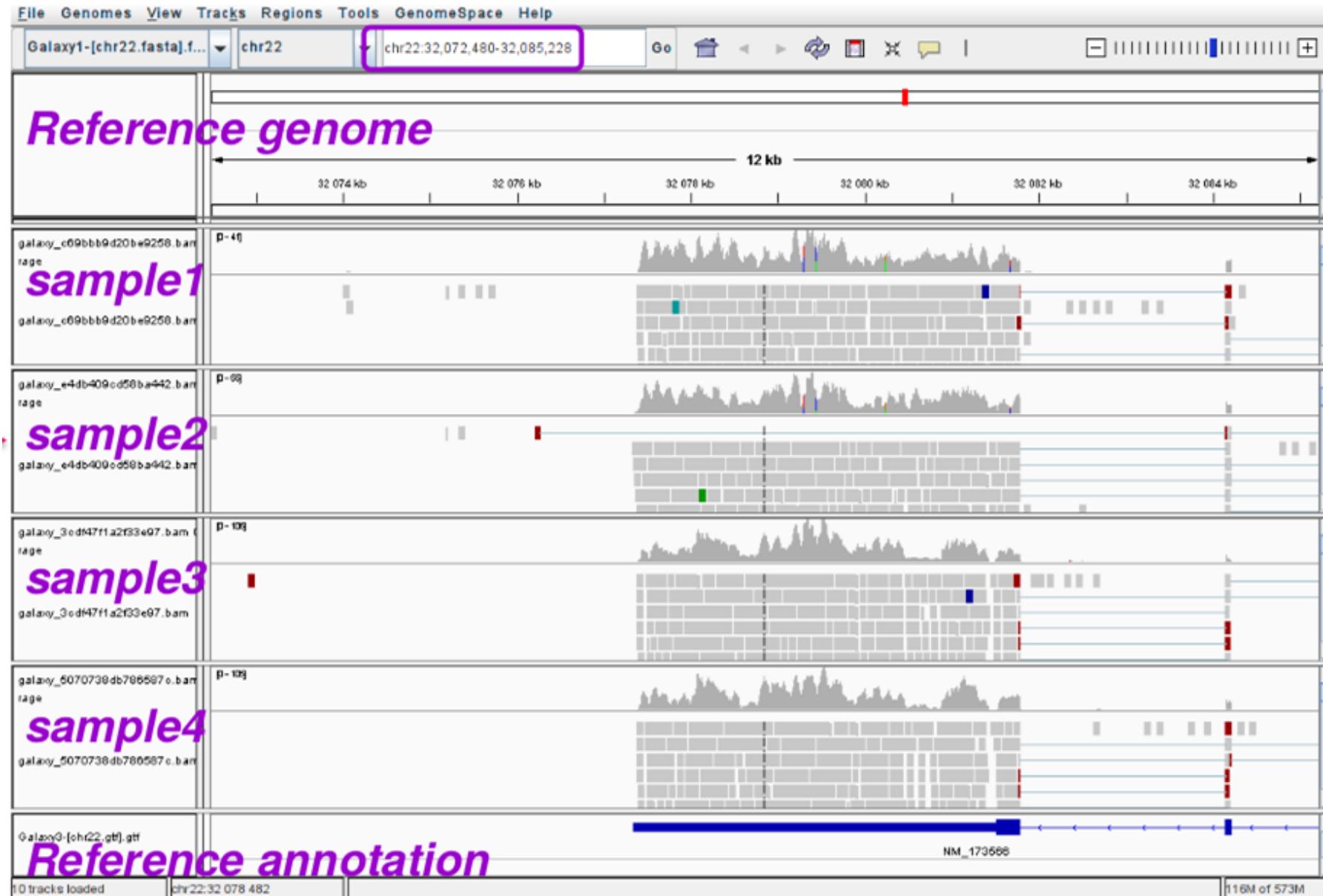
GFF (general feature format) is a file format used for describing genes and other features of DNA, RNA and protein sequences.

gff3

Seqname	Source	Feature	Start	End
chr22	protein_coding	gene	19701987	19712295
chr22	protein_coding	mRNA	19707711	19708397
chr22	protein_coding	protein	19707711	19708397
chr22	protein_coding	CDS	19707711	19707761
chr22	protein_coding	CDS	19707843	19707977
chr22	protein_coding	CDS	19708165	19708189
chr22	protein_coding	CDS	19708291	19708397
chr22	protein_coding	exon	19707711	19707761
chr22	protein_coding	exon	19707843	19707977
chr22	protein_coding	exon	19708165	19708189
chr22	protein_coding	exon	19708291	19708397

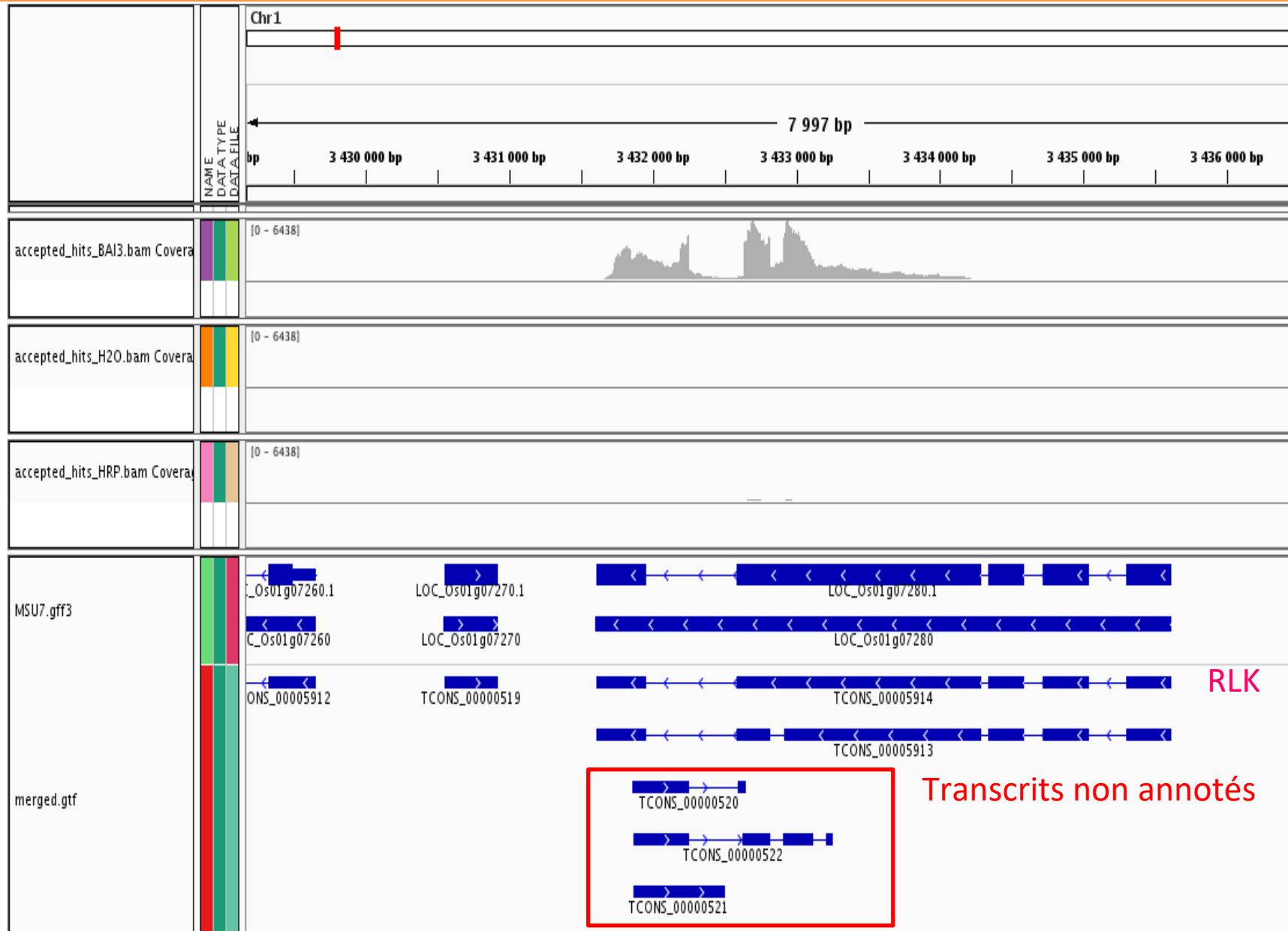
Score	Strand	Frame	Attribute
.	+	.	ID=ENSG00000184702;Name=SEPT5
.	+	.	ID=ENST00000413258;Name=SEPT5-016;Parent=ENSG00000184702
.	+	.	ID=ENSP00000404673;Name=SEPT5-016;Parent=ENST00000413258
.	+	1	Name=CDS:SEPT5;Parent=ENST00000413258
.	+	1	Name=CDS:SEPT5;Parent=ENST00000413258
.	+	1	Name=CDS:SEPT5;Parent=ENST00000413258
.	+	0	Name=CDS:SEPT5;Parent=ENST00000413258
.	+	.	Parent=ENST00000413258
.	+	.	Parent=ENST00000413258
.	+	.	Parent=ENST00000413258
.	+	.	Parent=ENST00000413258

Local Genome Browser (IGV)



source: Abims RNAseq formation 2018

Local Genome Browser (IGV)



Practice Présentation des données

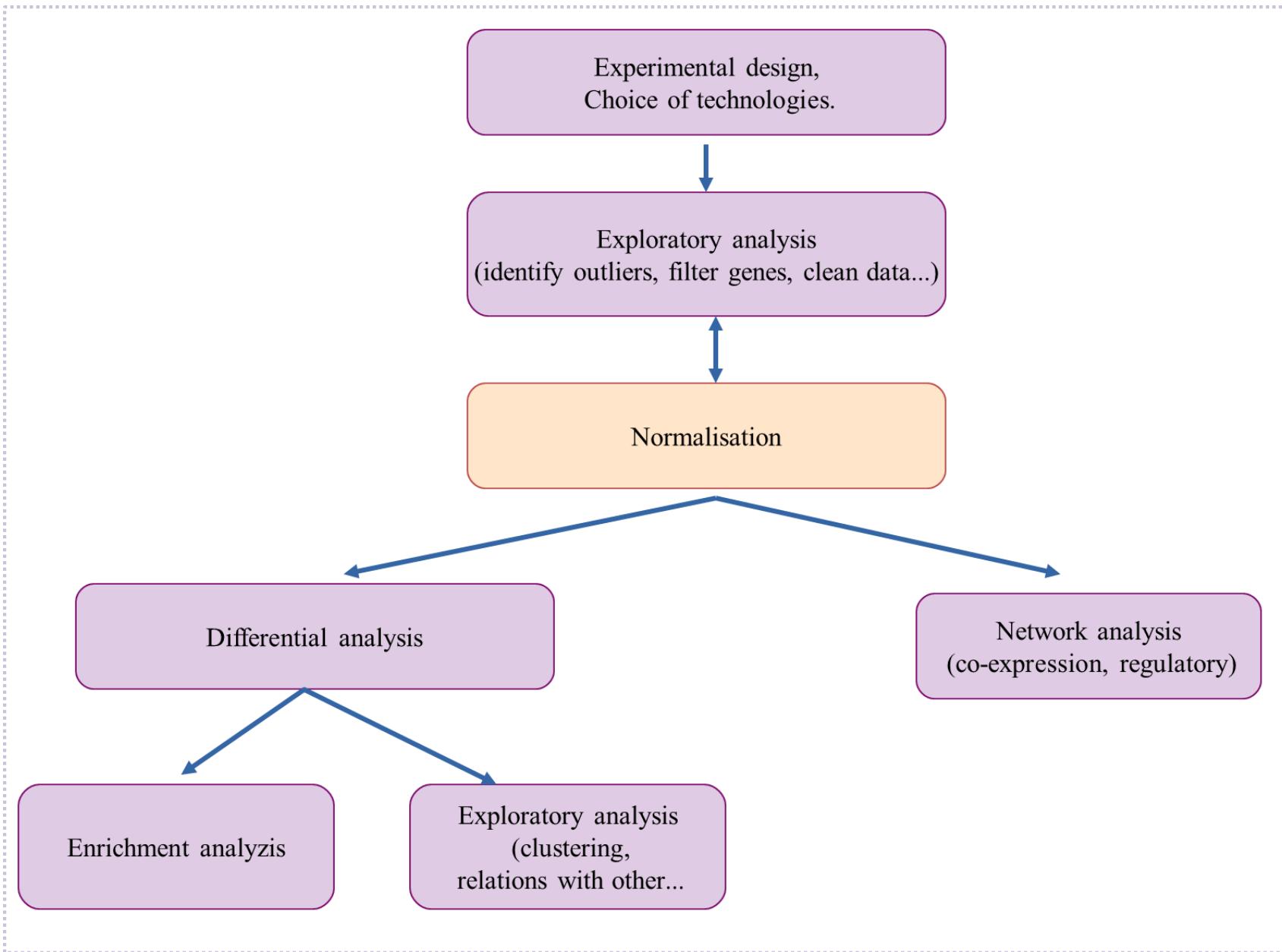
Recherche de gènes différentiellement exprimés

- Un gène est déclaré **différentiellement exprimé** (DE) entre 2 conditions si la différence d'expression observée est **statistiquement significative** i.e plus grande qu'une variation naturelle aléatoire.
 - Besoin d'un test statistique
 - Les principaux étapes de l'analyse :
 - Design experimental
 - Normalization
 - Analyse différentielle
 - Tests statistiques multiples



4- Normalisation des données

Normalisation des données



- **Identifier et corriger** les biais techniques dus au séquençage, pour les rendre comparable
 - S'assurer que les données sont exploitables
 - Réduire les bias techniques expérimentaux
 - De pouvoir comparer les données des différentes conditions entre elles
 - De s'approcher des hypothèses favorables pour l'analyse différentielle (distribution gaussiennedes données)
- **Types de Normalisation :**
 - Intra-échantillon (même séquençage)
 - Inter-échantillon (deux séquençage)
- Ce qui **influence** la normalisation:
 - Taille de la banque
 - Longueur de gènes
 - Composition en GC

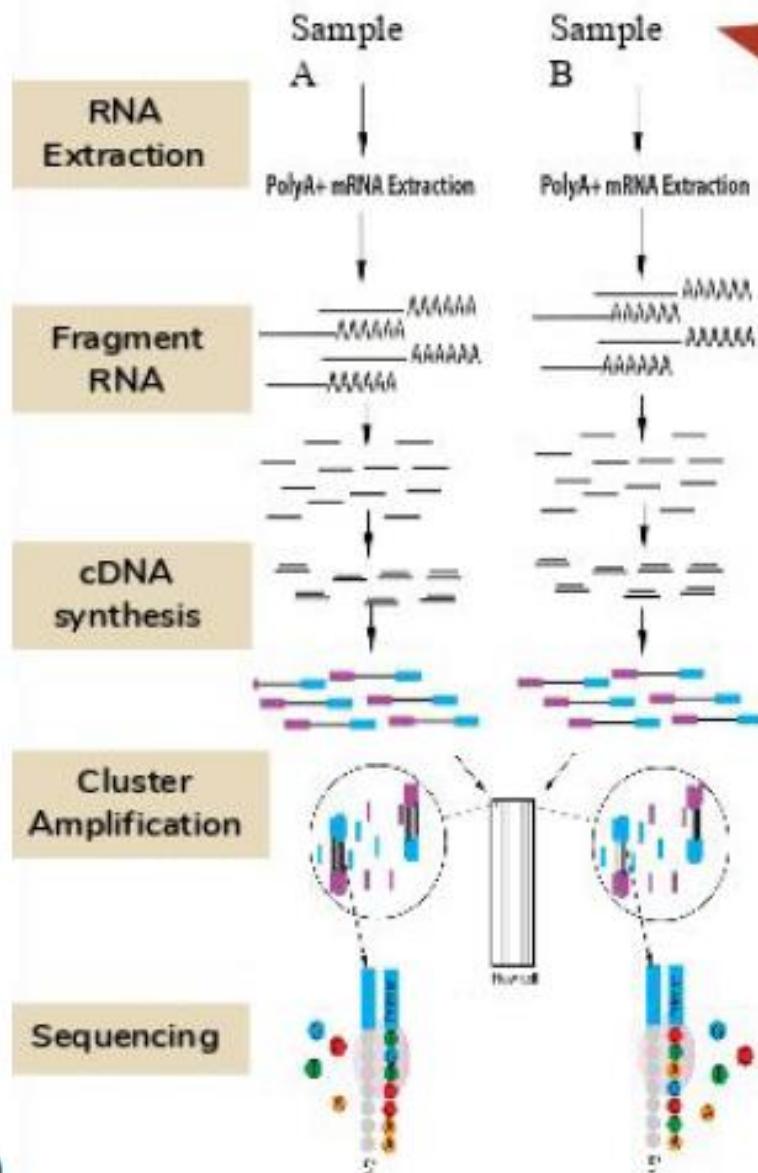


- **Pourquoi réaliser une normalization ?**

- Entre échantillon -> Comparer le niveau d'expression d'un gène entre différent échantillons
 - Profondeur du séquençage == taille de la banque
 - Biais d'échantillonnage durant la construction de la banque == effet batch
 - Présence de fragments majoritaires == saturation
 - Composition de la séquence dûe à l'étape d'amplification PCR (composition en GC)
- Parmi les échantillons -> comparer les gènes dans un échantillon
 - Longueur des gènes
 - Composition de la séquence (GC content)



Normalisation des données



Gene length



Sample 1

Library preparation



Gene 1



Gene 2

Sample

Gene 1 (50 kb)

10

Gene 1 (50 kb)

40

Gene 2 (100 kb)

20

Gene 2 (100 kb)

20

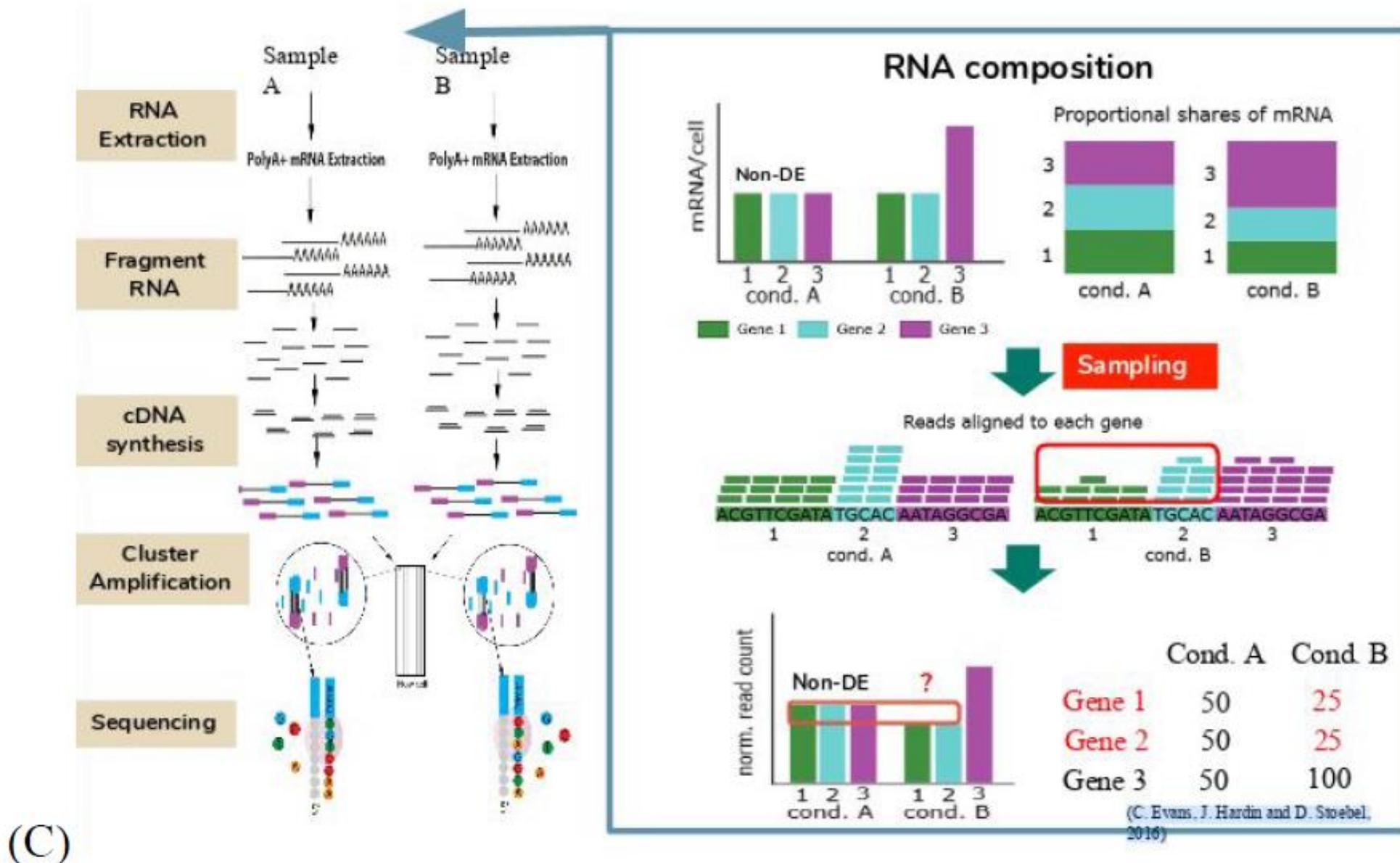
Bias

True expression

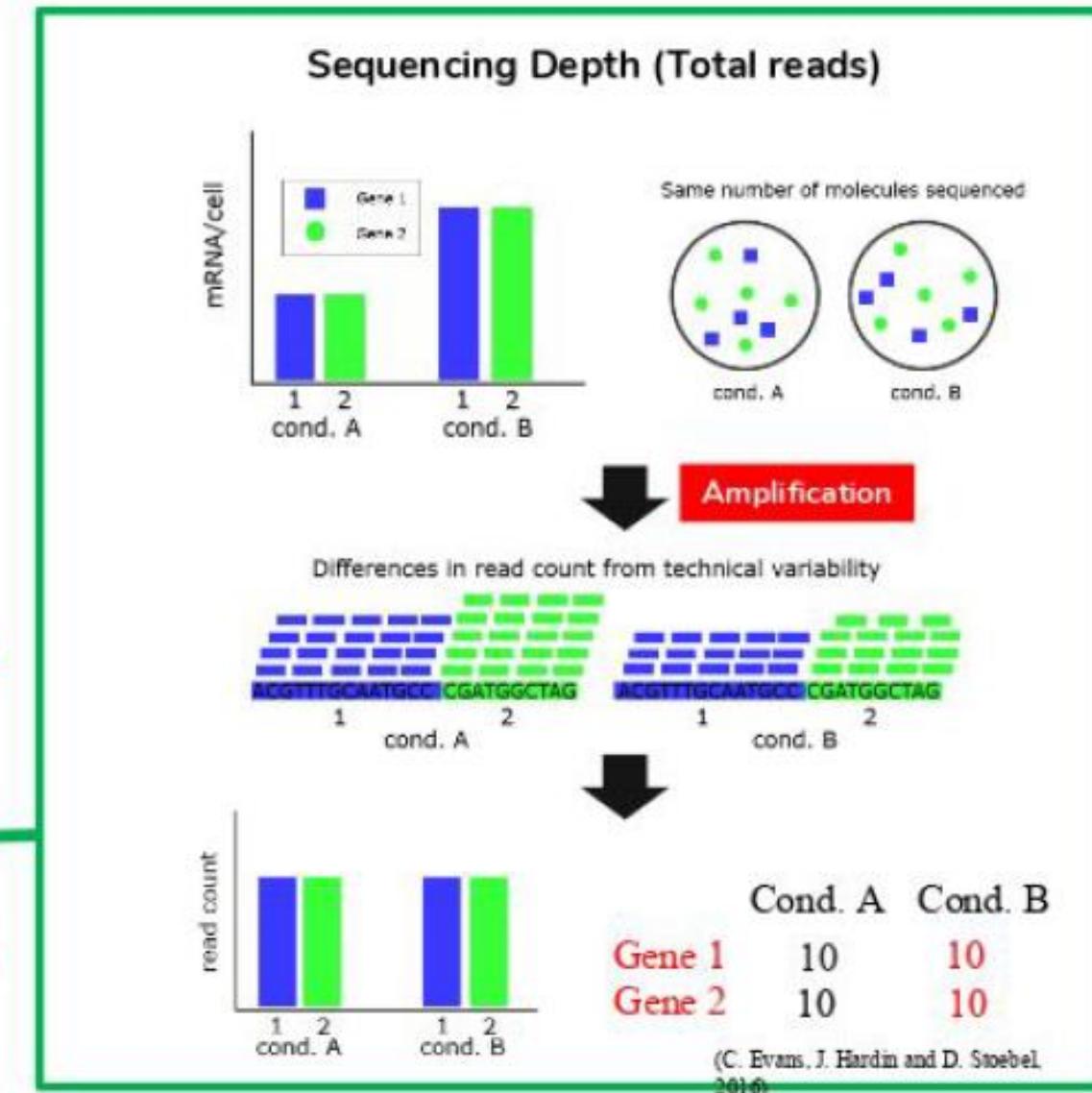
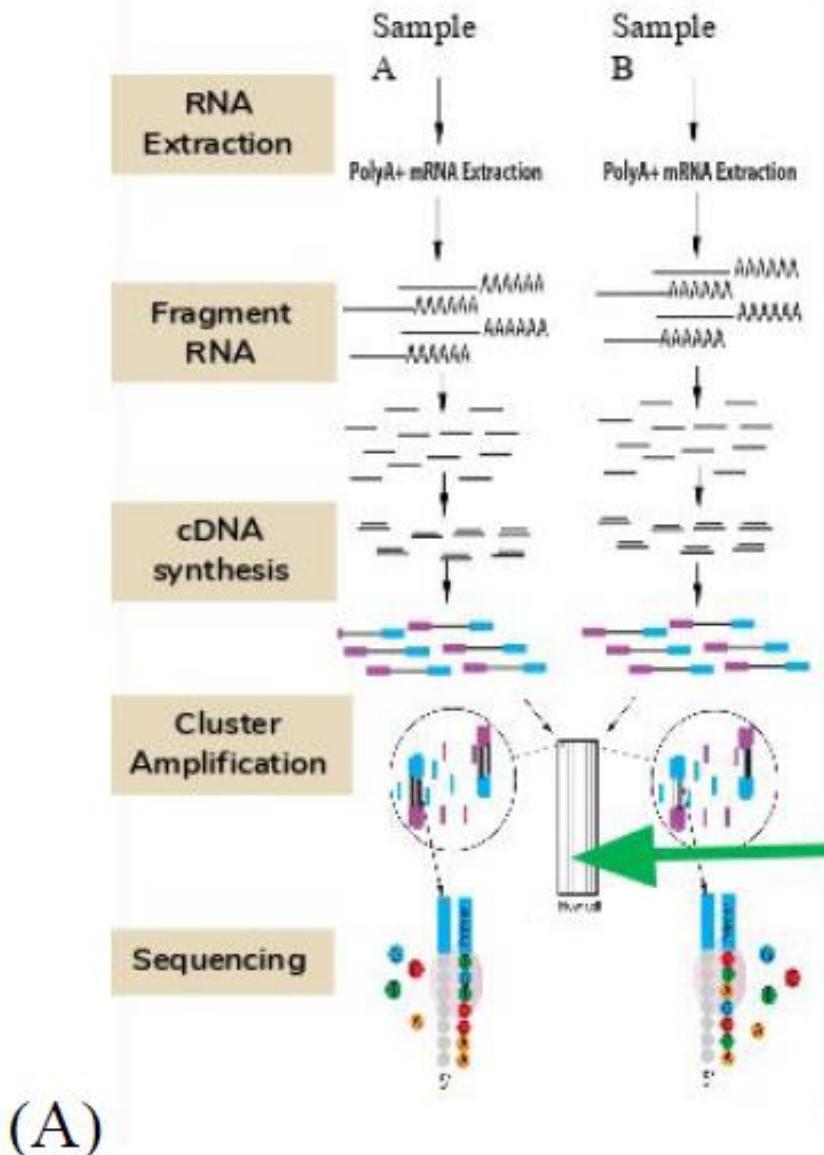
(C. Evans, J. Hardin and D. Stoebel,
2010)

(B)

Normalisation des données



Normalisation des données



- **Facteurs à prendre en compte avant la comparaison des conditions :**

- Taille de la banque (i.e. profondeur de séquençage) qui varie entre échantillons venant de différentes lanes de la flow cell du séquenceur.
- Longueur des gènes ayant nombre important de séquences
- Composition de la banque (taille relative du transcriptome) peut être différente entre deux conditions biologiques
- Composition en GC parmi différent échantillons peut conduire à un biais d'échantillonnage des gènes (Risso et al, 2011)
- La couverture des séquences des transcrits peut être biaisé et non uniformément distribué le long du transcrit (Mortazavi et al, 2008)

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 ?



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



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.



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



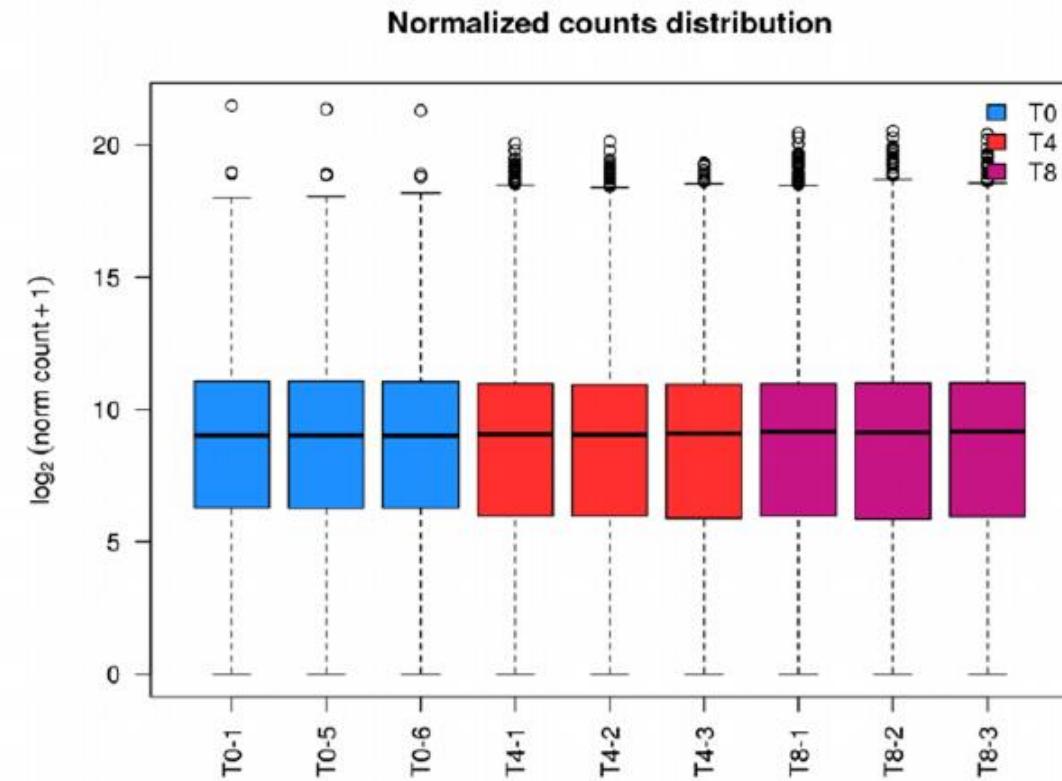
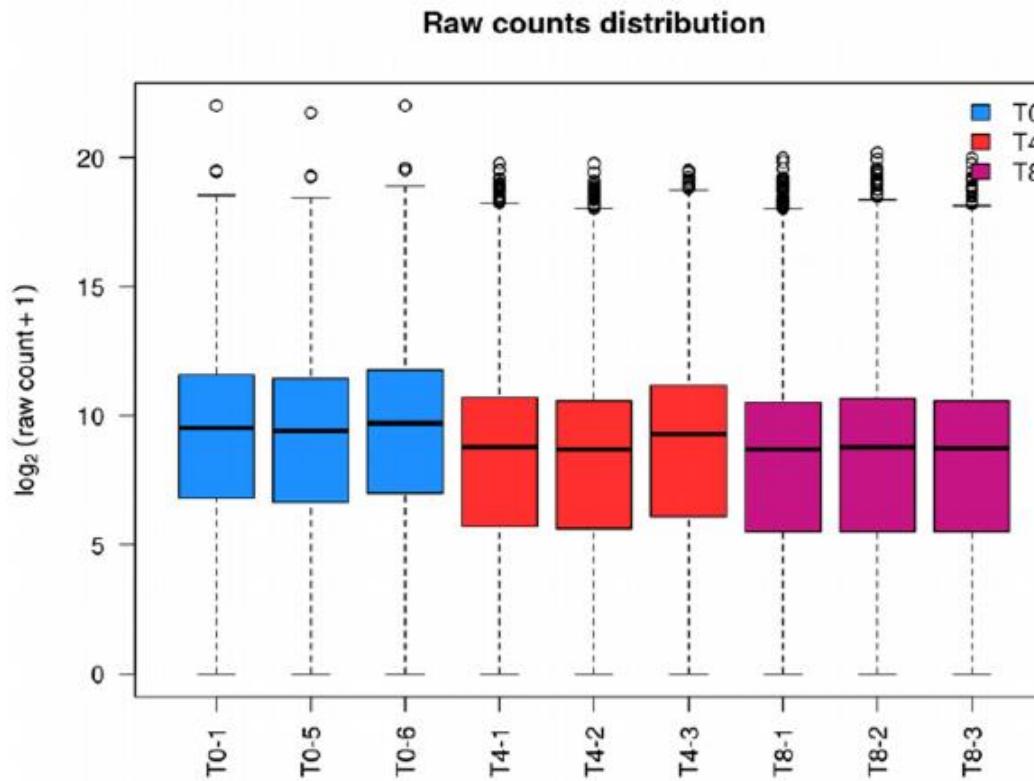
Le gène A est exprimé dans les échantillons 2 et 3, mais la profondeur de séquençage est plus importante dans l'échantillon 3 que dans le 2 (différences de taille des librairies).



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

Normalisation des données

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



Méthodes de Normalisation Intra-banques

Objectif : calculer un facteur d'échelle appliqué à chaque banque

- **Total Count (TC)** : On divise chaque nombre de reads par le nombre total de reads (c'est-à-dire la taille de la banque) et on multiplie par le nombre moyen de reads des banques.
- **Upper Quartile (UQ)** :
 - **Objectif** : On applique la méthode TC en remplaçant le nombre total de reads par le 3^{ième} quartile des comptes différents de 0.
 - Normalisation moins sensible aux valeurs extrêmes
 - Normalisation plus robuste, notamment dans le cas où plusieurs gènes très abondants sont différemment exprimés.
- **Reads Per Kilobase per Millions (RPKM)** :
 - **Objectif** : réaliser une normalisation qui tient compte de la taille de la banque (par une méthode de type Total Count) et de la longueur des gènes.
 - Mélange normalization inter et intra banque
 - Permet de comparer les gènes entre eux mais inadaptée pour comparer 2 conditions sur un même gène.

Méthodes de Normalisation Intra-banques

Objectif : calculer un facteur d'échelle appliqué à chaque banque

- **TMM** : Trimmed Mean of M-values
 - TMM normalization method Considère que la plupart des gènes ne sont pas différemment exprimés.
 - TMM normalise l'output totale d'ARN parmi les échantillons et **ne tient pas compte de la longueur du gène** ou de la taille de la bibliothèque pour la normalisation.
 - Le TMM sera un bon choix pour éliminer les effets de lot tout en comparant les échantillons de différents tissus ou génotypes ou dans les cas où la population d'ARN serait significativement différente parmi les échantillons.
 - TMM est implémenté dans edgeR et permet une meilleure comparaison entre échantillon
 - **edgeR ne prend pas en compte la longueur des gènes** pour la normalisation, car il suppose que la longueur des gènes est constante entre les deux normalisation
- **RLE** : Relative Log Expression (RLE) : Semblable à TMM, cette méthode de normalisation est basée sur l'hypothèse que la plupart des gènes ne sont pas DE. Pour un échantillon donné, le facteur d'échelle RLE est calculé comme la médiane du rapport, pour chaque gène, de son nombre de lectures sur sa moyenne géométrique sur tous les échantillons.

Méthodes de Normalisation Intra-banques

Objectif : calculer un facteur d'échelle appliqué à chaque banque

- **Méthode Total Count (TC) => Peu efficaces**
 - Pas de prise en compte des différences possibles entre les compositions en ARN des conditions.
- **Méthode RPKM => Peu efficaces**
 - Même dans le cas où un biais lié à la longueur des gènes existe, l'utilisation du RPKM ne permet pas de le corriger complètement.
- **Méthode à Privilégier => Upper-Quatile, RLE, TMM**
 - Même dans le cas où un biais lié à la longueur des gènes existe, l'utilisation du RPKM ne permet pas de le corriger complètement

Méthodes de Normalisation Intra-banques

Les bais techniques

- **Effet de la Taille de la Banque :**
 - Deux échantillons de même composition en ARN.
 - Une banque pour chaque échantillon
 - Banque A : 2 781 315 reads
 - Banque B : 2 254 901 reads
 - • On aura donc artificiellement 1,2334 fois plus de reads dans A que dans B
 - Pourtant les quantités “réelles” sont identiques
 - Bias de la Taille de la banque
- **Effet de la Longueur des Gènes:**
 - Pour un même niveau d’expression.
 - Un long transcrits est plus facilement séquencé
 - Donc plus de reads
 - Il sera plus facilement mis en évidence en DE
 - • Corriger le Bias

5- Choix de la méthode de normalization

Méthodes de Normalisation intra – banque

RLE , TMM, Upper-
Quartile

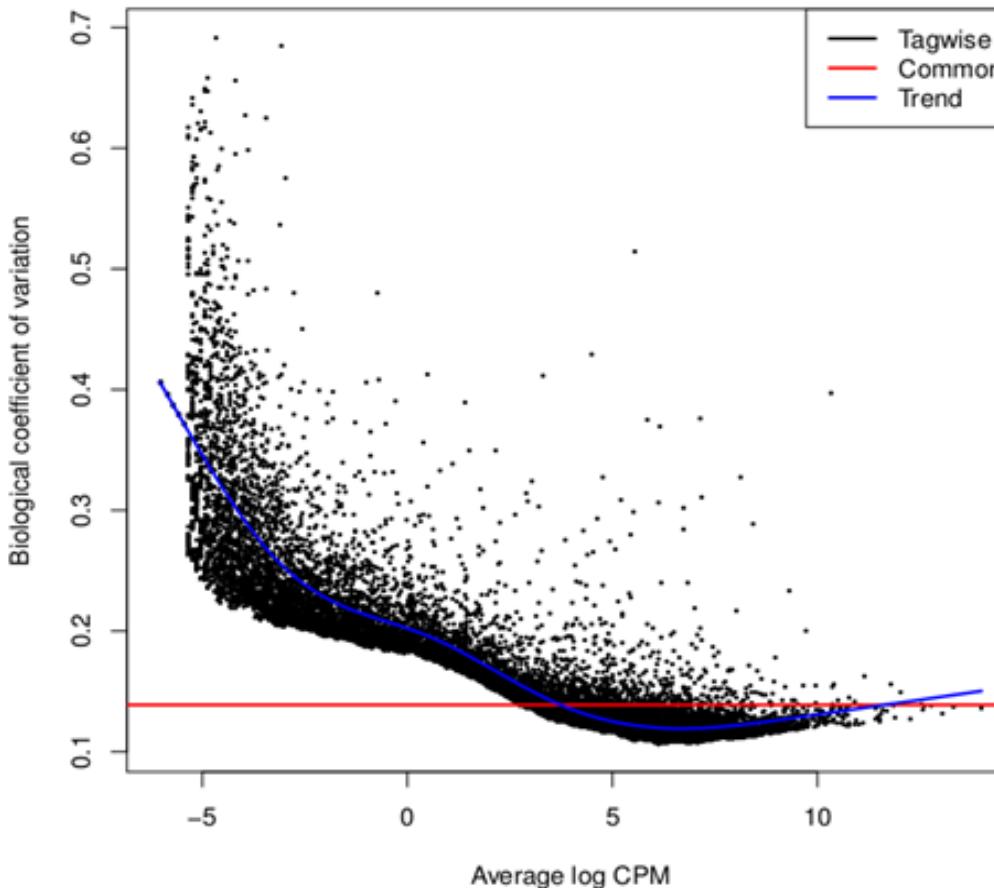
Correction par la variance

Loi binomiale négative

Edge R , DESeq



EdgeR



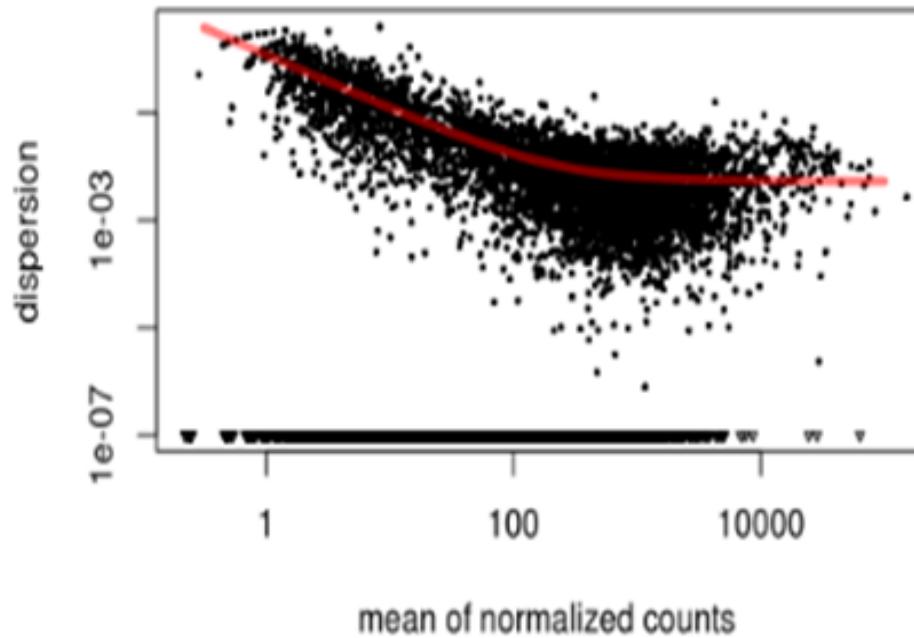
Estimation de la dispersion (entre réplicats biologiques)

- utilisation de la valeur individuelle (« tagwise ») ou de la valeur ajustée « trend » ou « common » pour le calcul des tests statistiques de DE

- Utiliser la méthode « tagwise » lorsqu'on a au moins 4 réplicats
- Utiliser la méthode commune lorsqu'on a peu de réplicats (2 ou 3)

=> Utilisation de ces valeurs de dispersion pour le calcul des tests statistiques de DE (p-value)

DESeq



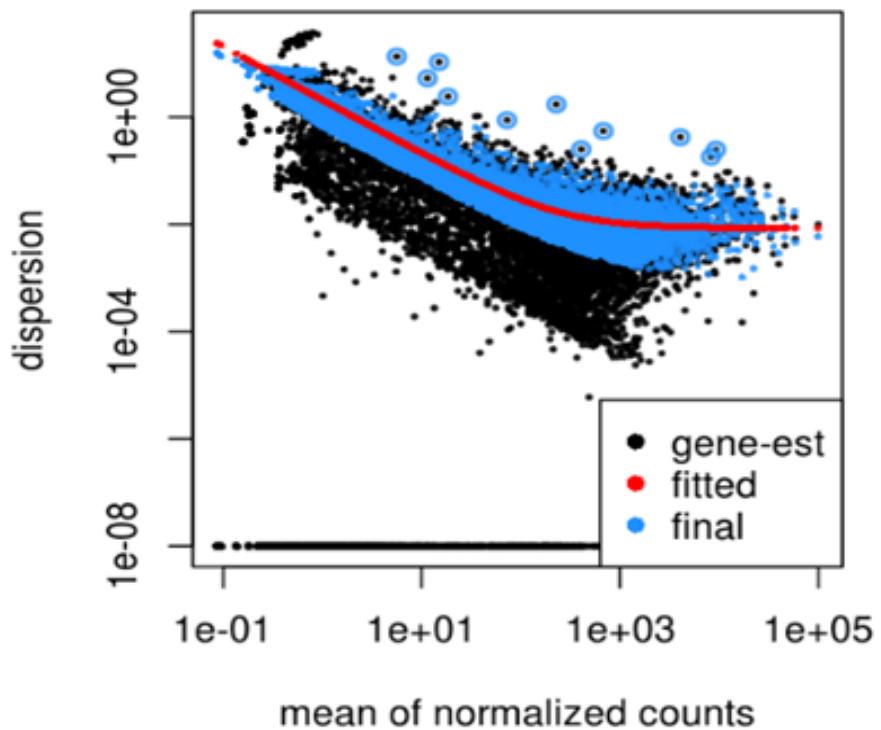
Estimation de la dispersion

- utilisation de la valeur ajustée pour les transcrits dont l'estimateur individuel (en noir) inférieur à la valeur ajustée
- utilisation de la valeur individuelle pour les transcrits dont l'estimateur individuel est supérieur à la valeur ajustée

=> Utilisation de ces valeurs de dispersion pour le calcul des tests statistiques de DE (p-value)

=> Plus sensible à la dispersion des données

DESeq2



Estimation de la dispersion

- utilisation d'une valeur intermédiaire (en bleu) entre la dispersion individuelle (en noir) et la dispersion ajustée (en rouge)
 - utilisation de la dispersion individuelle si celle-ci est considérée comme extrême par rapport à la distribution globale (points entourés de bleu)
- => Utilisation de ces valeurs de dispersion pour le calcul des tests statistiques de DE (p-value)

Comparaison des outils

Objectif : Choisir l'outils le plus adapté

- **DESeq** utilise une estimation de la variance qui la rend moins permissive pour les grandes variabilités entre conditions.
 - Dès qu'au moins l'une des conditions présente une variabilité importante, la méthode ne fait pas confiance à ce gène et ne va pas le considérer comme différentiellement exprimé.
 - En revanche, quand la variabilité intra-conditions est plus faible, DESeq fait plus confiance et sélectionne même les gènes qui ont un fold-Change plus faible que ceux sélectionnés par EdgeR.
 - DESeq est à privilégier pour des expérimentations très répétables.
- **DESeq2** est plus souple , moins stringent et il détectera plus de genes différentiellement exprimés.

Practice : DIANE

TP Diane

DIANE : Exercice *Saccharomyces cerevisiae*

Data:

* ref : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488244/>

* data : NCBI SRA database under accession number SRS307298 _*S. cerevisiae*_.

Genome size of _*S. cerevisiae*_ : 12M (12.157.105)

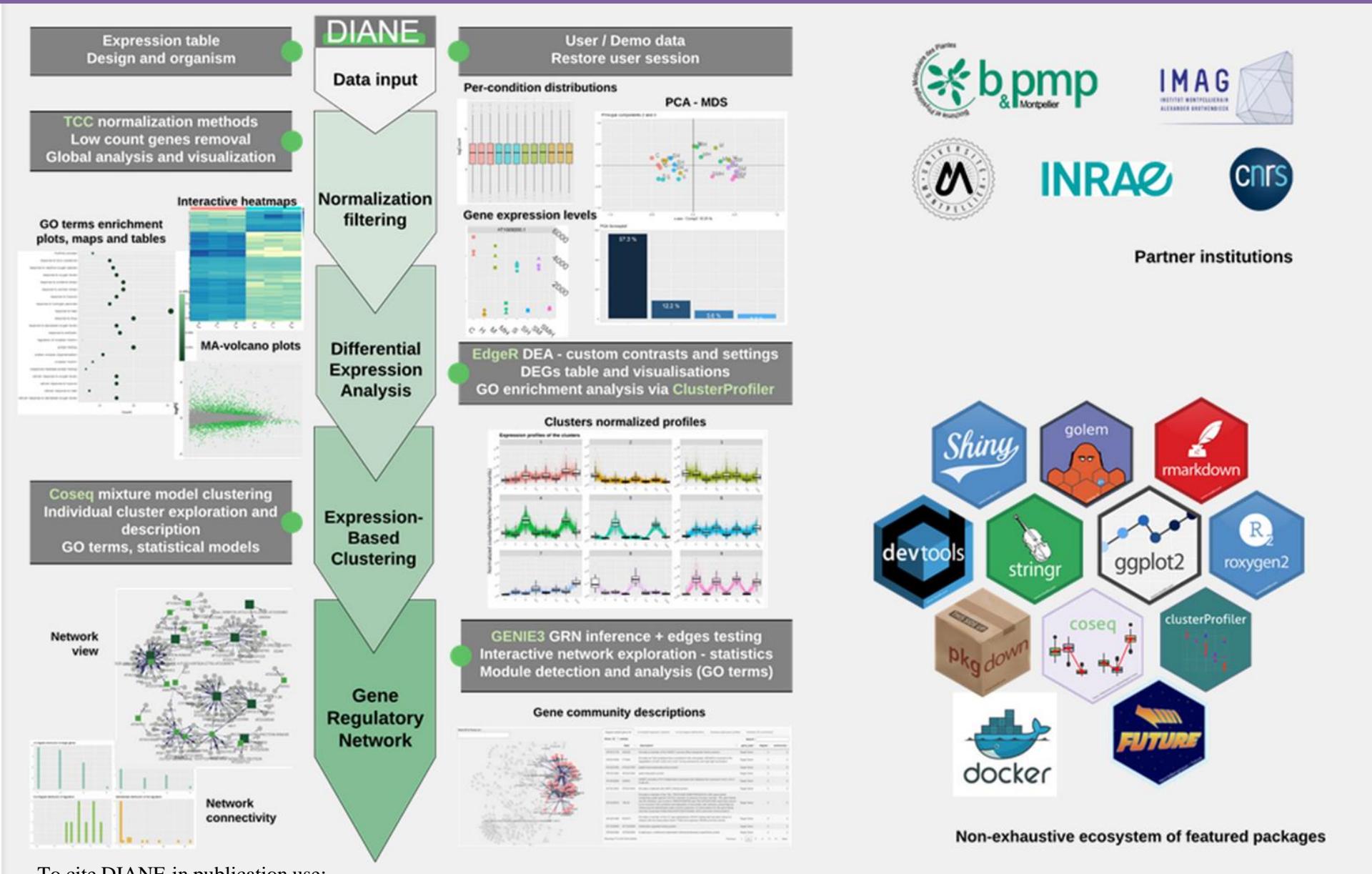
(https://www.yeastgenome.org/strain/S288C#genome_sequence)

Outils :

DIANE : (<https://diane.bpmp.inrae.fr/>)

- Après avoir aligné les séquences contre la référence de *Saccharomyces cerevisiae* déterminer une liste de gènes différentiellement exprimés entre les conditions batch et CENPK.
 - Vous avez à votre disposition la matrice reads counts.
 - Etape 1 Normalisation :
 - Quelle est l'influence de la méthode de normalisation ,
 - tmm,
 - Deseq2
 - aucune normalization.
 - Etape 2 Analyse différentielle avec DIANE
 - Etape 3 Générer une liste et des graph, MA plot , volcano plot et Heatmap.

Practice Normalisation des données DIANE



To cite DIANE in publication use:

Cassan, O., Lèbre, S. & Martin, A. Inferring and analyzing gene regulatory networks from multi-factorial expression data: a complete and interactive suite. BMC Genomics 22, 387 (2021). <https://doi.org/10.1186/s12864-021-07659-2>



Partner institutions



Non-exhaustive ecosystem of featured packages

Practice : DIANE Import Data

Import expression data and experimental design

Expression file upload

Demo Arabidopsis data

Toggle to import your data

Expected gene IDs are in the form
No gene ID requirement
FOR OTHER

Your organism : Other

Separator : Tab

Choose CSV/TXT expression file : count_table.txt

Choose CSV/TXT gene information file (optional) : No file selected

Organism : Other

Database : Other

Seed ensuring reproducibility (optional, can be left as default value) : 34

CHANGE SEED SET SEED

Preview of the expression matrix

Design and gene information files

Separator : Comma

Choose CSV/TXT design file (optional) : No file selected

Describe the levels of each factors for your conditions

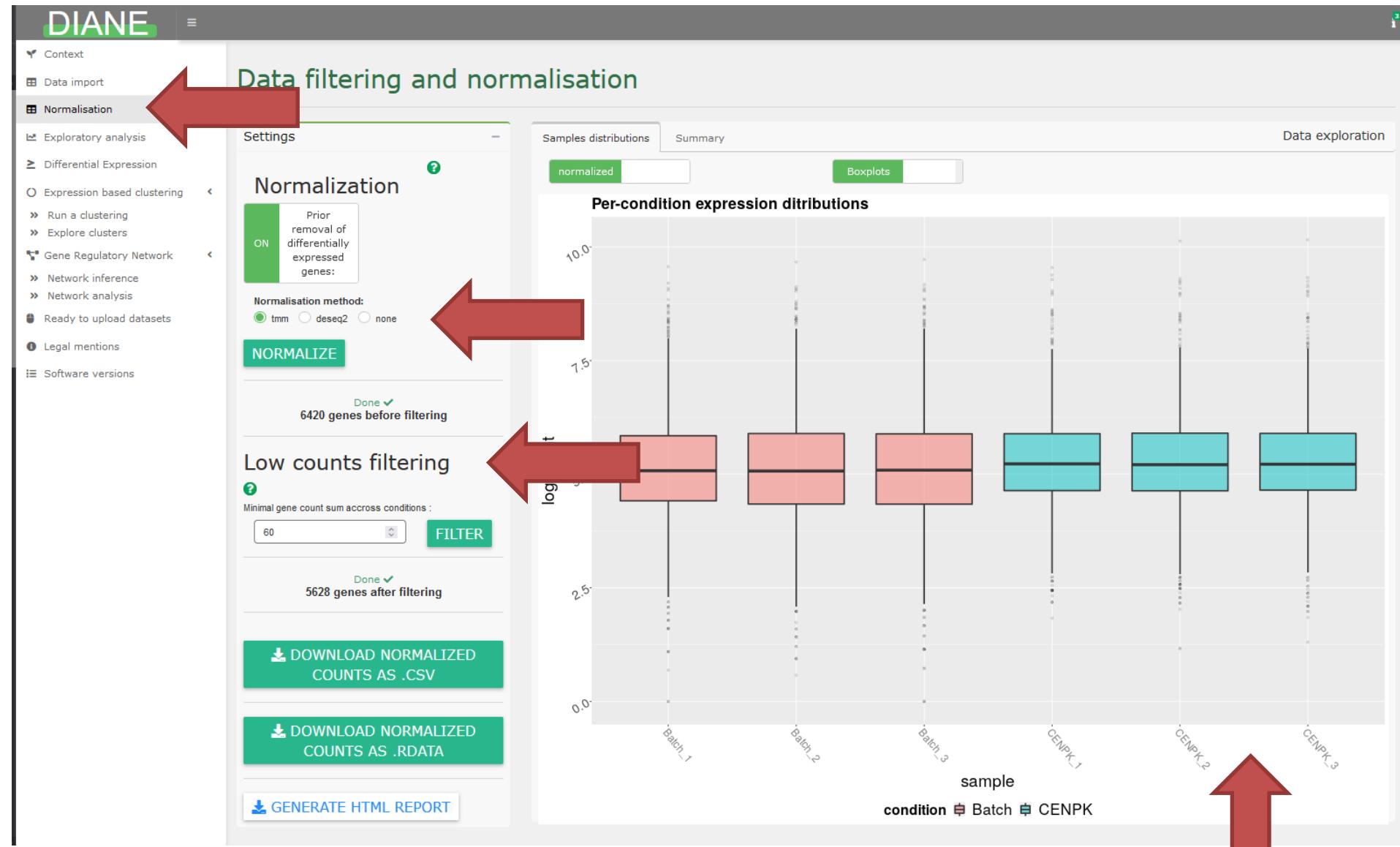
Reads_count.csv

GeneInformation

Designed.csv

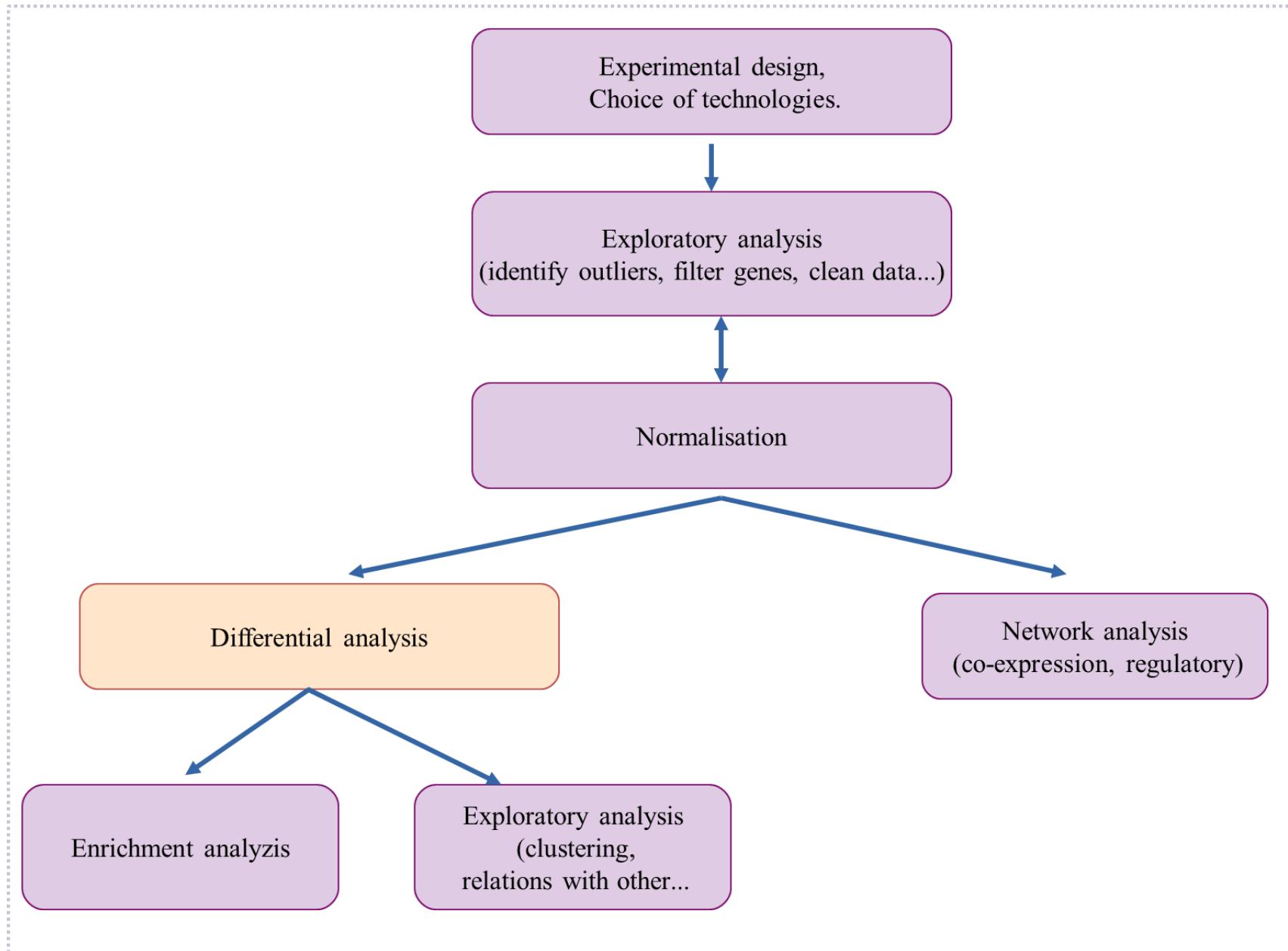
This might help you visualize the general aspect of the data and different sequencing depths of your conditions.

DIANE Normalisation Filtration



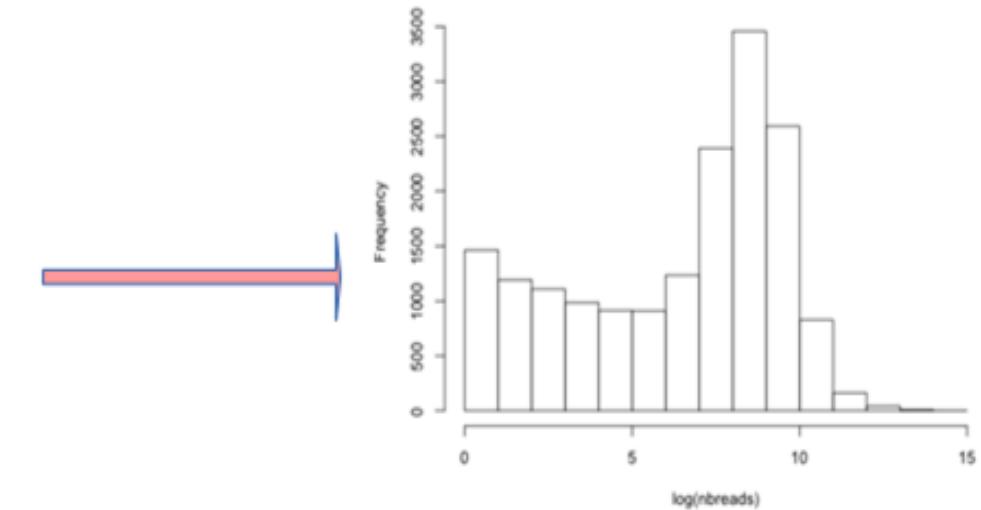
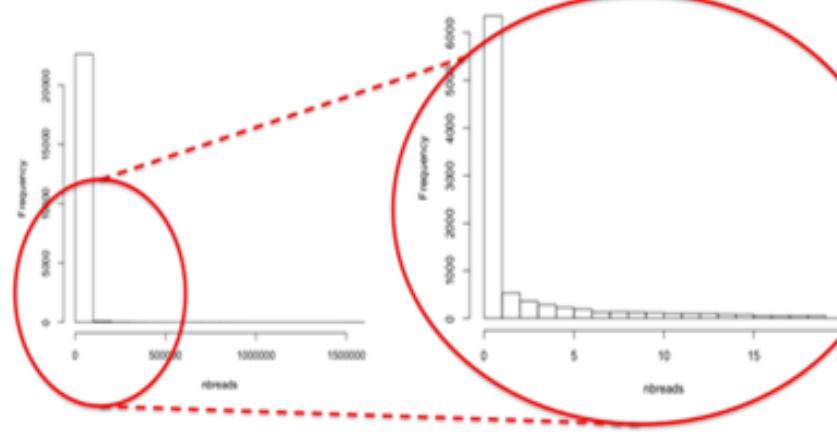
6- Recherche de gènes différentiellement exprimés

Modélisation des données



Modélisation des données

- **Note** : Utilisation du log(nbr de reads) pour que les données suivent une loi statistique , et nécessité de transformer les « 0 ».
 - => Loi binomiale négative



- **Note** : Utilisation du log(FoldChange)
 - Fold Change = ratio entre 2 niveaux d'expression
 - Ratio de la valeur finale / valeur initiale

Rappel : définition du Fold-Change

$$FC = \frac{\text{expression condition } V}{\text{expression condition } G}$$

Rappel : définition du Fold-Change :

$$FC = \frac{\text{expression condition } V}{\text{expression condition } G}$$

Gene	V1	V2	V3	G1	G2	G3	FC	p-valeur
Gene1	5	7	6	2	2	2	3	0.06
Gene2	800	1000	900	350	250	200	3	0.03
Gene3	700	900	1100	350	200	250	3	0.10
Gene4	900	500	1300	200	550	50	3	0.06
:	:	:	:	:	:	:	:	:

- **p-value** : risque/probabilité de déclarer un gène différentiellement exprimé alors qu'il ne l'est pas.
 - **H_0 : le gène g a un niveau d'expression constant**
 - P-value de 0.05= on autorise qu'un gène ait 5% de risque d'être appelé DE alors qu'il ne l'est pas
 - **Problème des tests multiples**: si on teste 10000 gènes non différentiellement exprimés, on autorise 500 gènes à être déclarés DE alors qu'ils ne le sont pas

=> **nécessité de filtrer au préalable les gènes** (ex: niveau total d'expression < 10) pour limiter le nombre de tests

=> **Plus on augmente le nombre de tests , plus on a de chance de décider qu'un gène est différentiellement exprimé alors qu'il ne l'est pas.**

=> **besoin de correction** et utiliser une p-value ajustée adaptée aux tests multiples:

- procédure de Benjamini-Hochberg (BH) qui consiste à contrôler le False Discovery Rate (**FDR**), c'est à dire la proportion de faux positifs dans les gènes déclarés différentiellement exprimés
- procédure de Bonferroni (+ stringent)

- **Erreur de 1ere espèce (Type 1 error):**

- Probabilité α de rejeter H_0 alors qu'elle est vraie
- Probabilité de décider qu'un gène est différentiellement exprimé alors qu'il ne l'est pas.
- **Faux positif**

- **Erreur de 2ere espèce (Type 2 error):**

- Probabilité β de rejeter H_0 alors qu'elle est fausse
- Probabilité de décider qu'un gène n'est pas différentiellement exprimé alors qu'il l'est.
- **Faux négatif**

- **Conséquence :**

- En testant les 20000 gènes avec $\alpha = 5\%$
- **On s'attend à obtenir $20000 \times 0,05$ faux positifs soit 1000 gènes qui ne sont en réalité pas différentiellement exprimés.**

Situation	Décision	
	accepter H_0	rejeter H_0
H_0 vraie	$1-\alpha$	α
H_0 fausse (diff. expr.)	β	$1-\beta$

- **Filtre = risque**

- Le nombre est la valeur seuil que nous mesurons contre la p-value.
- Elle indique à quel point les résultats observés doivent être extrêmes pour rejeter l'hypothèse nulle d'un test significatif
- Ex

=> Résultat avec un niveau de 90% de niveau de confiance , alpha est $1 - 0,90 = 0,10$

=> Résultat avec un niveau de 95% de niveau de confiance , alpha est $1 - 0,95 = 0,05$

⇒ Résultat avec un niveau de 99% de niveau de confiance , alpha est $1 - 0,99 = 0,01$

⇒ $\alpha > p\text{-value}$ H_0 Rejetée

- **False Discovery Rate (FDR)**

- **Principe** : ajuster le seuil α en fonction des résultats observés (p-value obtenues)
- M tests ayant des p-value p_1, \dots, p_m triées par ordre croissant
- Pour un seuil α trouver le plus grand k tel que $Pk \leq \frac{k}{m}\alpha$ et déclarer les gènes $1, \dots, k$ différemment exprimés.

	R_1	R_2	G_1	G_2	p-value	$\alpha * k/m$
267628_at	441.8	431.5	347.2	375.2	0.036937	0.01
267629_at	226.5	205.6	185.2	175.9	0.090013	0.02
267630_at	1142.6	1080.7	1019.8	1018.6	0.096209	0.03
267627_at	57	6	45.5	38.6	0.721558	0.04
267631_at	77.7	58	84.4	57.4	0.872008	0.05

- Ici **aucun gène** n'est déclaré différemment exprimé pour $\alpha = 0,05$

Practice : DIANE

[TP Diane](#)

Practice Analyse différentielle DIANE

Differential expression analysis

Results table MA - Vulcano plots Heatmap Gene Ontology enrichment Compare genes lists (Venn) Results

Show 10 entries Search:

	logFC	logCPM	FDR	Regulation
YIL057C	10.76	8.602	0.000	Up
YMR107W	9.953	9.671	0.000	Up
YJR095W	8.602	11.42	0.000	Up
YJR095W	8.407	7.856	0.000	Up
YMR303C	7.902	10.69	0.000	Up
YCR010C	7.094	8.400	0.000	Up
YGL205W	6.950	8.828	0.000	Up
YHR096C	6.890	9.951	0.000	Up
YAL054C	6.768	10.41	0.000	Up
YBR067C	6.734	10.29	0.000	Up

Showing 1 to 10 of 1,403 entries Previous 1 2 3 4 5 ... 141 Next

Conditions to compare for differential analysis :

Reference: ✓ Batch CENPK Perturbation: Batch ✓ CENPK

Adjusted pvalue (FDR) : 0.05

Absolute Log Fold Change (Log2 (Perturbation / Reference)) : 1

DETECT DIFFERENTIALLY EXPRESSED GENES

Done ✓ See plots and tables for more details

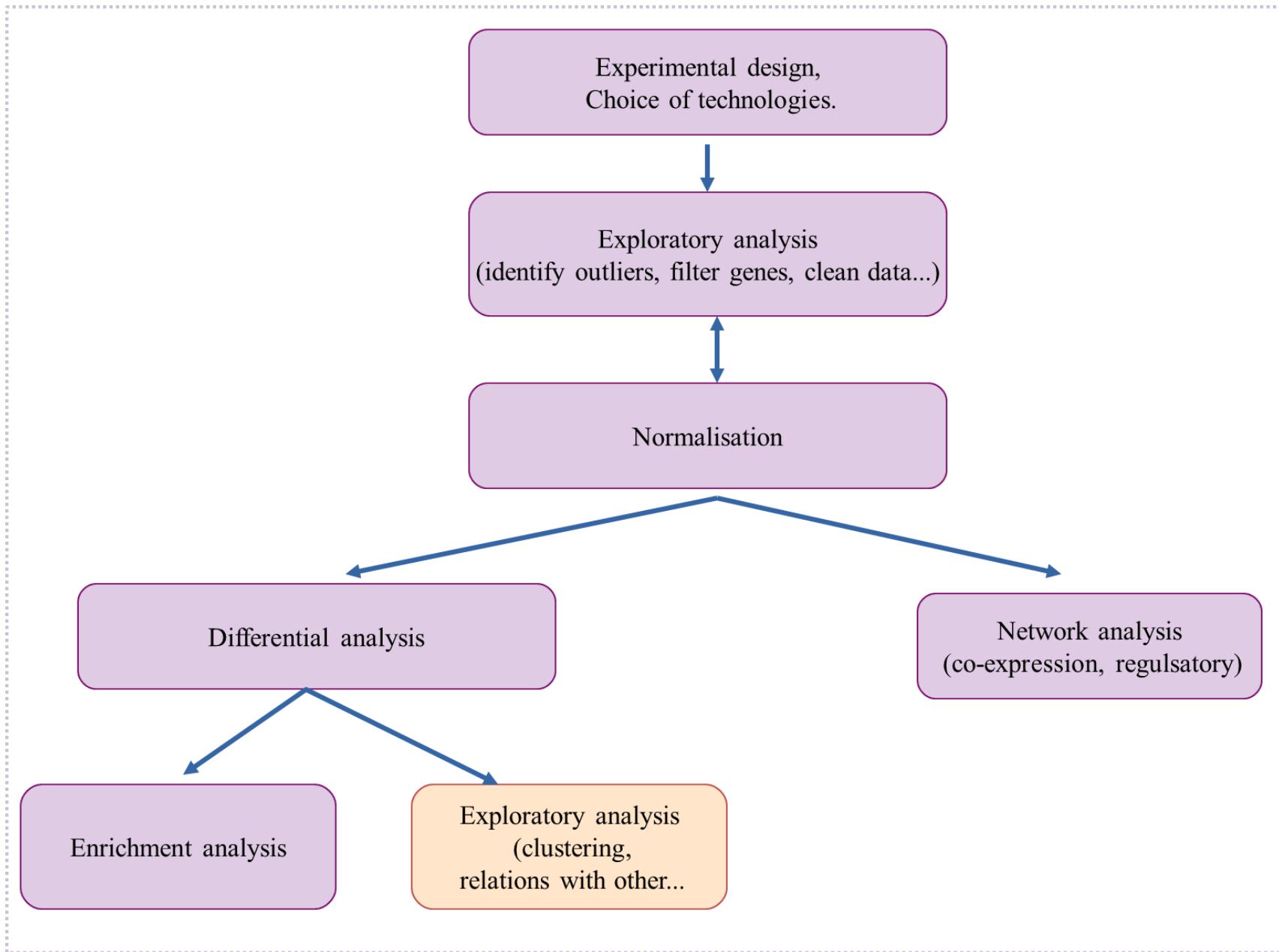
820 up regulated GENES
583 down-regulated GENES

DOWNLOAD RESULT TABLE AS .TSV

GENERATE HTML REPORT

6- Plots and Graphical Représentations

Analyses exploratoires

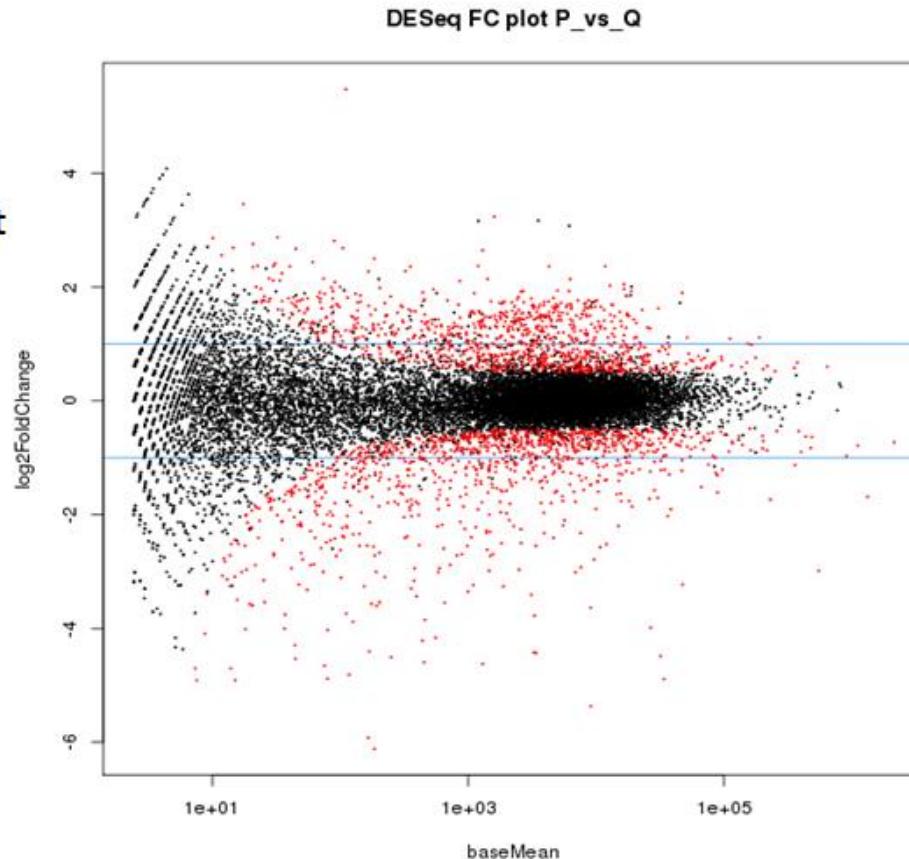


Smear plot

Smear plot / MA plot
Pvalue adj < 0.05

MA plot

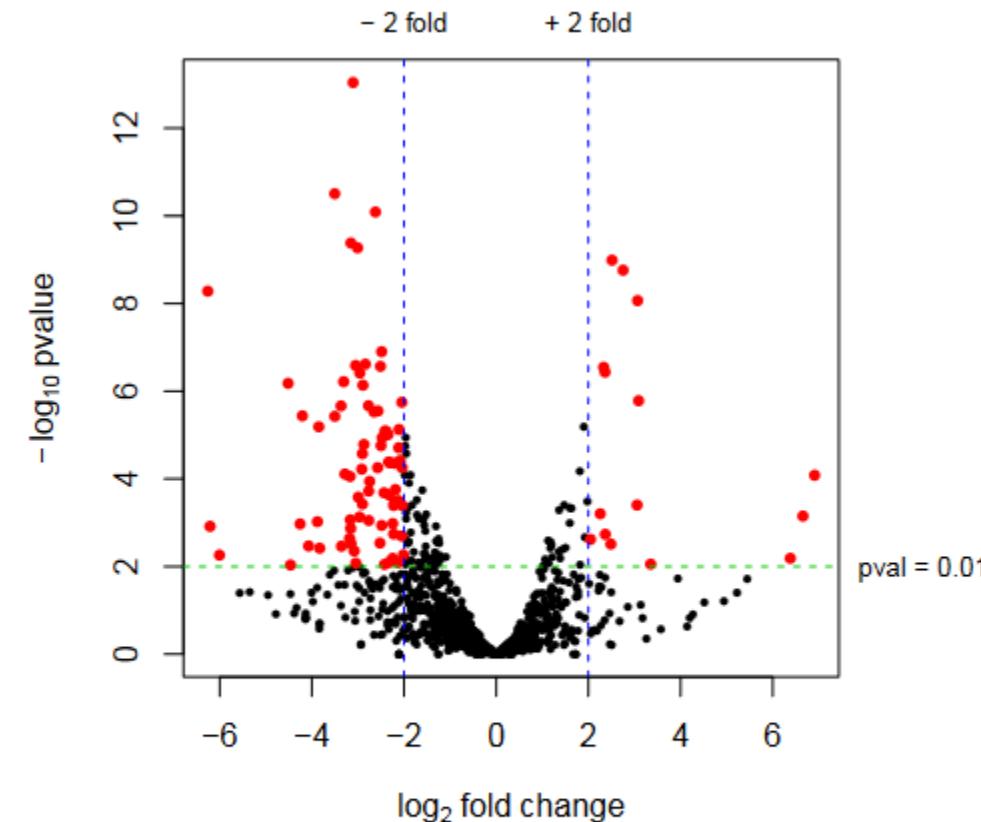
Le MA plot est un graphe qui était initialement utilisé dans les analyses de puce à ADN. C'est un nuage de points représentant en abscisse l'expression moyenne du gène à travers les différents échantillons, et en ordonnée le log-ratio des expressions moyennes d'une condition par rapport à l'autre. En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



M : ordonnées, ratios des intensités. $\log_2 R - \log_2 G = \log_2(R/G)$
A : abscisses, moyenne des intensités du spot. $\frac{1}{2} (\log_2 R + \log_2 G)$

Volcano plot

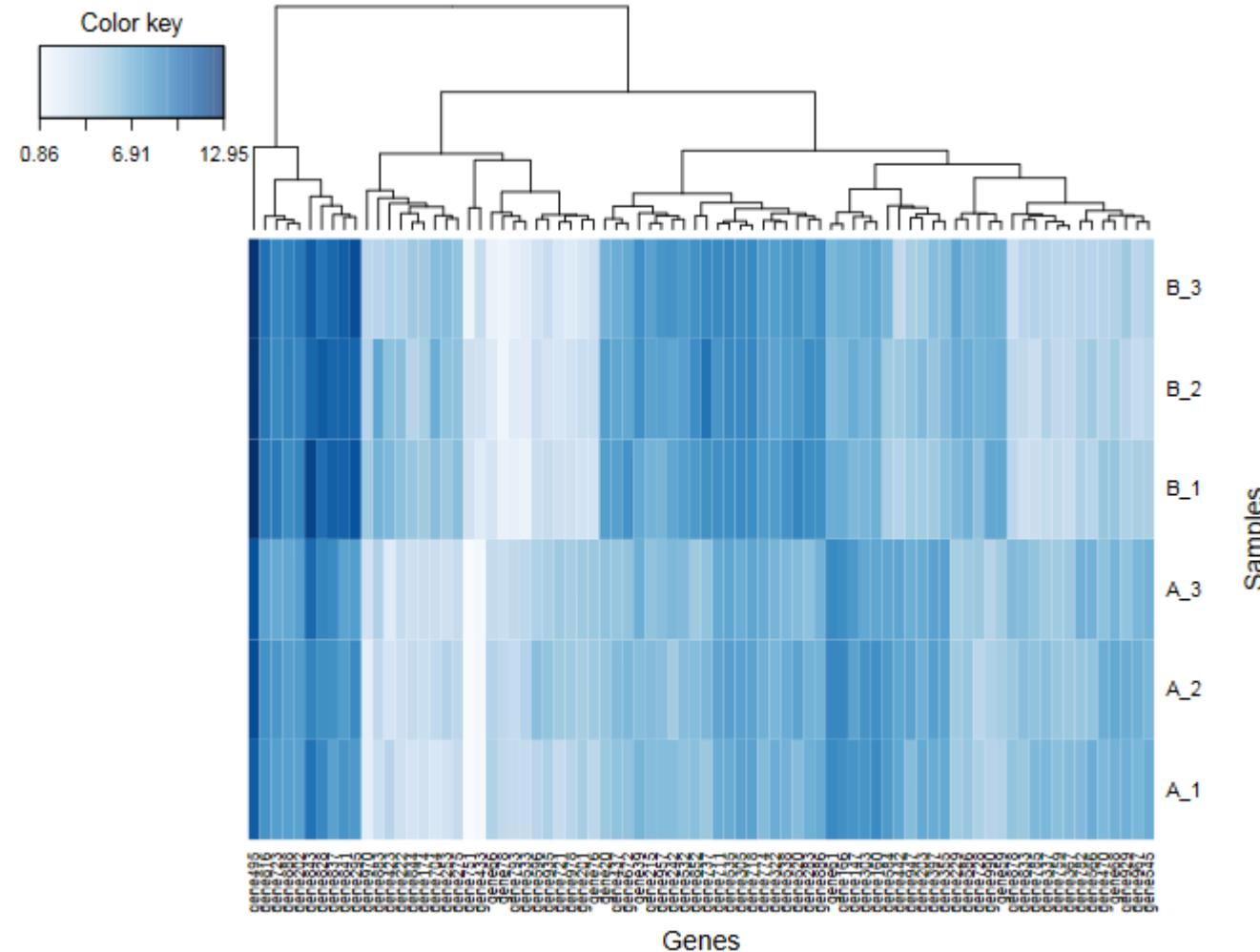
Volcano plot
Pvalue adj < 0.01



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

Fold change vs. P-valeur (t-test ou autre)

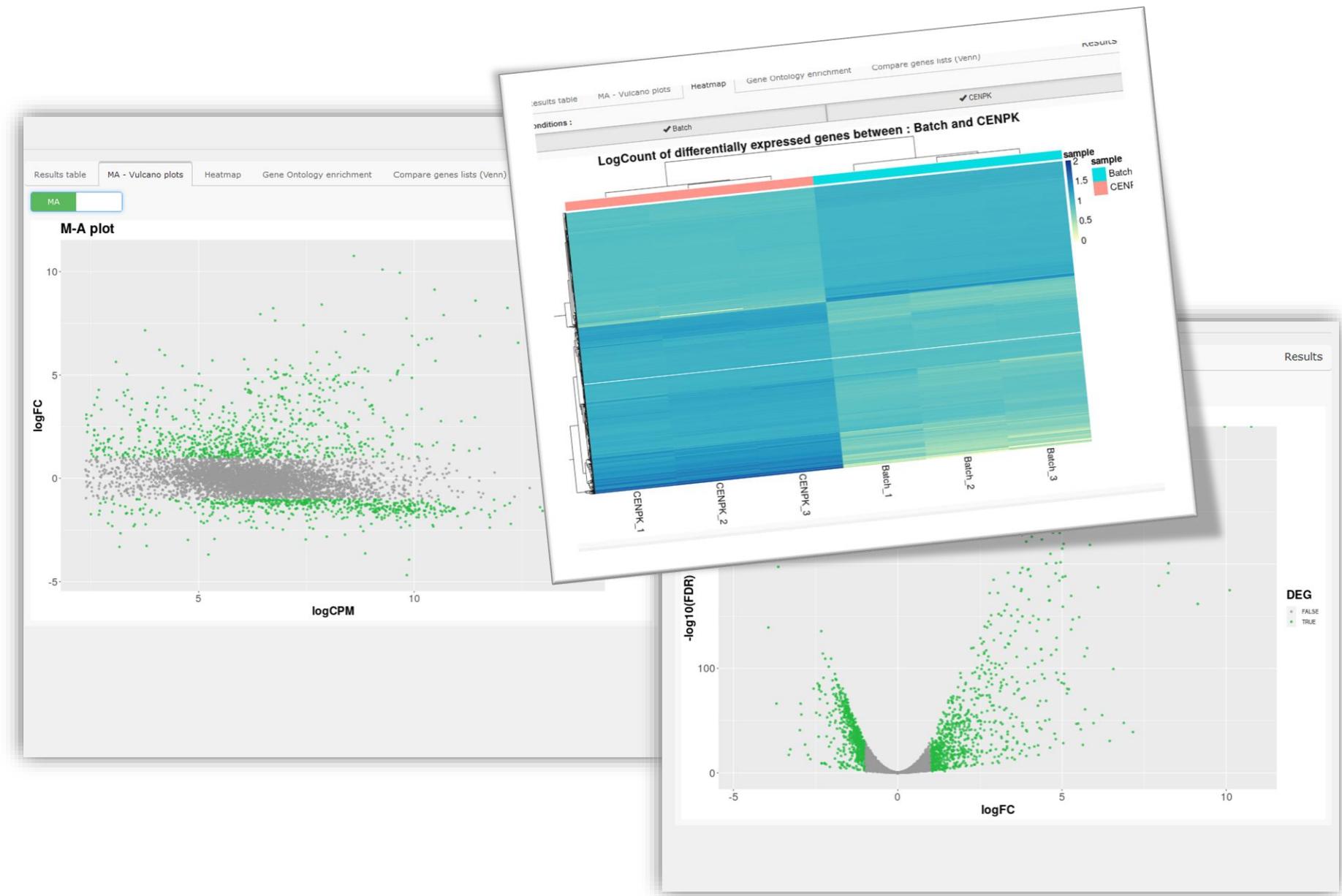
Hierarchical clustering / Heatmap



Practice : DIANE

TP Diane

Practice Représentation graphique



DIANE - Differential Expression Analysis report

Code ▾

Dashboard for the Inference and Analysis of Networks from Expression data

This report was automatically generated by [DIANE](#) to improve research reproducibility.

It contains the main settings and results for the DEA tab of the application, reporting the last transcriptome comparison that was performed.



Your settings

Normalization method :

Hide

```
print(the_r$norm_method)
```

```
## [1] "deseq2"
```

Reference and perturbation condition :

Hide

```
paste(r$ref, r$trt)
```

```
## [1] "Batch CENPK"
```

Threshold adjusted p-value and minimal expected absolute log fold change :

Hide

```
paste("FDR =", r$fdr, "LFC = ", r$lfc)
```

```
## [1] "FDR = 0.05 LFC = 1"
```

Contributeurs pour cette formation



Alexis Dereeper



Sébastien Ravel



Christine Tranchant-Dubreuil



Sébastien Cunnac



Gautier Sarah



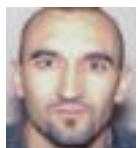
Julie Orjuela-Bouniol



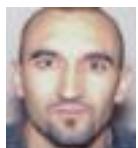
Catherine Breton



Aurore Compte



Alexandre Soriano



Guilhem Sempéré



Links

Related courses : <https://www.nathalievialaneix.eu/>

Related courses : <https://southgreenplatform.github.io/trainings/linuxJedi/>

Tutorial RNAseq : <http://nathalievilla.org/doc/pdf/slides-rnaseq.pdf>

Book : <http://compgenomr.github.io/book/>

Degust : <http://degust.erc.monash.edu/>

MeV: <http://mev.tm4.org/>

MicroScope: <http://microscopebioinformatics.org/>

Comparison of methods for differential expression:

https://southgreenplatform.github.io/trainings//files/Comparison_of_methods_for_differential_gene_expression_using_RNA-seq_data.pdf

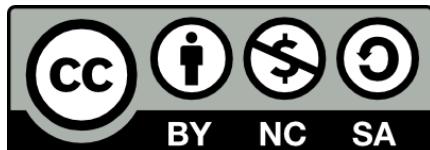
PIVOT: <https://github.com/qinzhu/PIVOT/>

DESeq2: <http://www.bioconductor.org/packages/release/bioc/html/DESeq2.html>

EdgR: <https://bioconductor.org/packages/release/bioc/html/edgeR.html>

DIANE: <https://diane.bpmp.inrae.fr/>

Merci pour votre attention !



Le matériel pédagogique utilisé pour ces enseignements est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions (BY-NC-SA) 4.0 International:
<http://creativecommons.org/licenses/by-nc-sa/4.0/>



**Merci de prendre 5 min pour remplir
l'enquête**

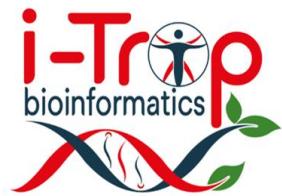
<https://itrop-survey.ird.fr/index.php/515725?lang=fr>



SUIVEZ NOUS SUR TWITTER !



South Green : [@green_bioinfo](#)



i-Trop : [@ItropBioinfo](#)



N'oubliez pas de nous citer !

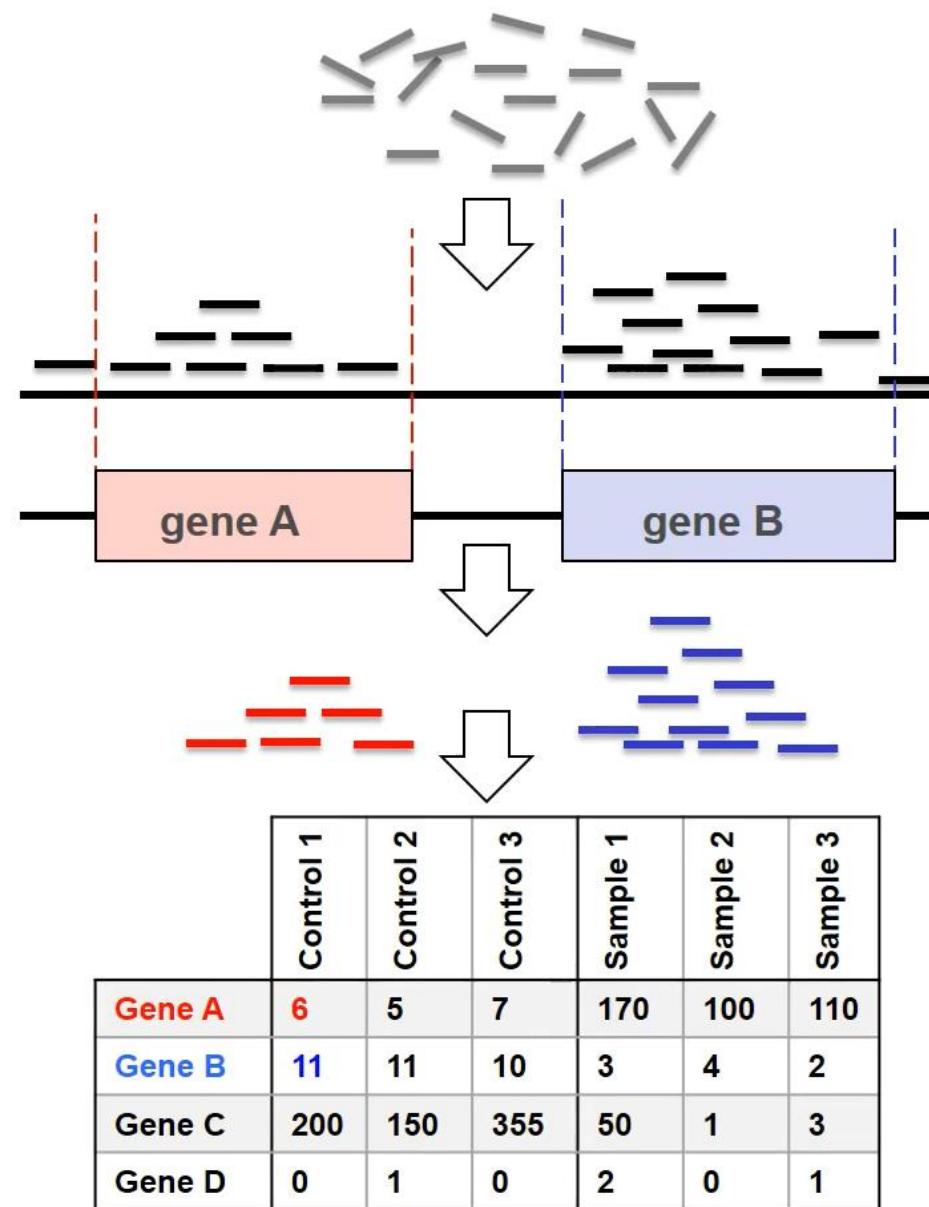
Comment citer les clusters?

"The authors acknowledge the IRD i-Trop HPC at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL: <http://bioinfo.ird.fr/> "

“The authors acknowledge the CIRAD UMR-AGAP HPC (South Green Platform) at CIRAD montpellier for providing HPC resources that have contributed to the research results reported within this paper. URL:
<http://www.southgreen.fr>”

- **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 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 for small sample sizes
 - **baySeq** In the presence of outliers, found a lower fraction of false positives true for small sample sizes not for large sample sizes

RNA-seq data analysis: steps, tools and files



STEP	TOOL	FILE
Quality control	FastQC	FASTQ
Pre-processing	Trimmomatic	FASTQ
Alignment	HISAT2	BAM
Quality control	RSeQC	
Quantitation	HTSeq	Read count file (TSV)
Combine count files to table	Define NGS experiment	Read count table (TSV)
Quality control	PCA, clustering	
Differential expression analysis	DESeq2, edgeR	Gene lists (TSV)

Omics data: multiple testing issue

Context:

We perform a large number N of statistical tests for which we reject or not H_0 .

Possible conclusions:

		Decisions	
		Non rejects of H_0	Rejects of H_0
Unknown truths	H_0 true	TN	FP
	H_0 false	FN	TP

Among all the genes told differentially expressed, the False Discovery Rate (FDR) is:

$$\frac{FP}{FP + TP}$$

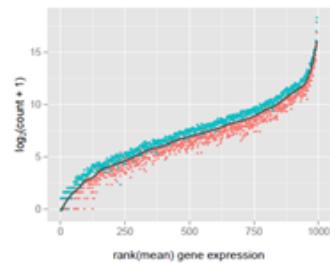
Normalization Technique	Name authors	Description	Software
UQ	Upper Quartile Ref : Bullard et al., 2010 (Upper) Quartile normalization	Les comptages par gène sont divisés par le 3e quartile des comptages non nuls de l'échantillon, puis multipliées par la moyenne des 3e quartiles de tous les échantillons.	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	<p>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.</p> <p>Ainsi, les comptages sont corrigés pour prendre en compte la taille de la librairie et la longueur des gènes. Cependant, il a été montré 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 exprimés. Cette méthode reste toutefois très populaire dans de nombreuses applications.</p>	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 différemment exprimés. Le facteur de normalisation pour un échantillon est obtenu en calculant pour chaque gène la médiane des ratios de ses comptages par rapport à sa moyenne géométrique entre les différents échantillons. L'idée sous-jacente est que les gènes non différemment exprimés 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 différemment exprimés, la médiane des ratios constitue une estimation du facteur correctif qui doit être appliquée à 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 différemment exprimés. Le facteur TMM est calculé pour chaque échantillon, l'un d'eux étant considéré 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 exprimés et des gènes ayant les plus forts log-ratios. D'après l'hypothèse selon laquelle il y a peu de gènes différemment exprimés, 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

RLE

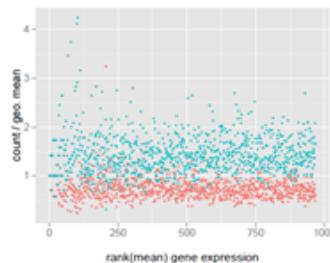
La normalisation RLE (Relative Log Expression) a été développée dans le package Bioconductor DESeq. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différemment exprimés. Le facteur de normalisation pour un échantillon est obtenu en calculant pour chaque gène la médiane des ratio de ses comptages par rapport à sa moyenne géométrique entre les différents échantillons. L'idée sous-jacente est que les gènes non différemment exprimés 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 différemment exprimés, la médiane des ratio constitue une estimation du facteur correctif qui doit être appliquée à l'ensemble des comptages.

Ref : Anders and Huber, 2010. Dans edgeR, DESeq – DESeq2

1 – Calcule une pseudo référence

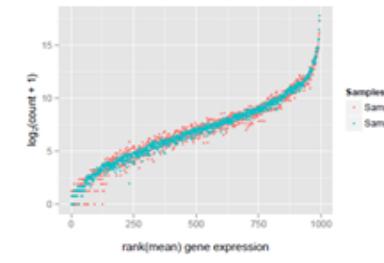
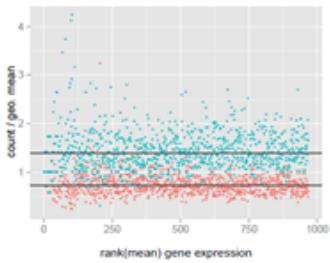


2 – Centre les échantillons comparés à la référence



3 – Calcule un facteur de normalisation : médiane des ratio de ses comptages par rapport à sa moyenne géométrique entre les différents échantillons

```
## With edgeR
calcNormFactors(..., method = "RLE")
## with DESeq
estimateSizeFactors(...)
```

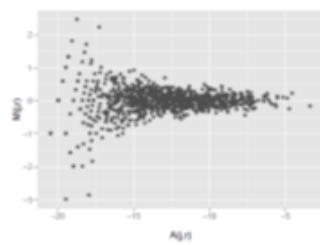


TMM

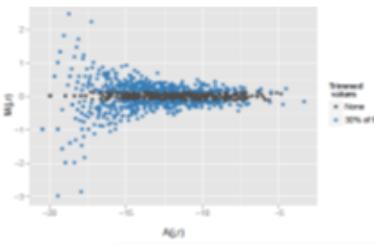
La normalisation TMM (Trimmed Mean of M-values) est implementée dans le package Bioconductor edgeR. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différemment exprimés. Le facteur TMM est calculé pour chaque échantillon, l'un d'eux étant considéré comme l'échantillon de 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 exprimés et des gènes ayant les plus forts log-ratios. D'après l'hypothèse selon laquelle il y a peu de gènes différemment exprimés, 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.

Ref : Robinson, M. and Oshlack, A. (2010). Dans edgeR.

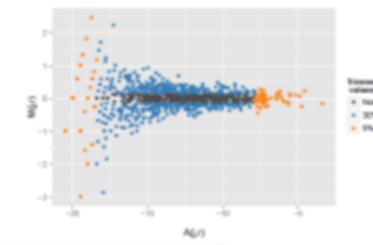
1 – Sélectionner un échantillon pour servir de référence :
L'échantillon r avec le quartile supérieur plus proche du quartile de la moyenne supérieure.



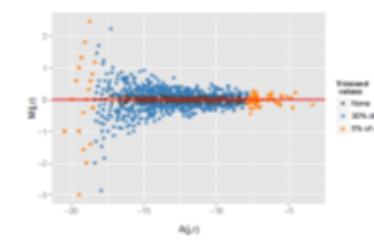
2 – Trim 30% on M-values



3 – Trim 5% on A-values



3 – Sur les données restantes, calculer la moyenne pondérée des valeurs M

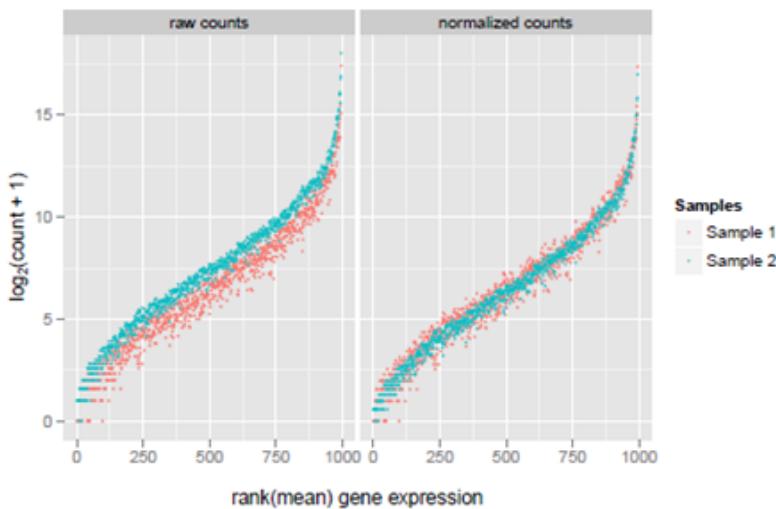


```
## With edgeR  
calcNormFactors(..., method="TMM")
```

Total read count adjustment

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.

Ref : Mortazavi et al., 2008

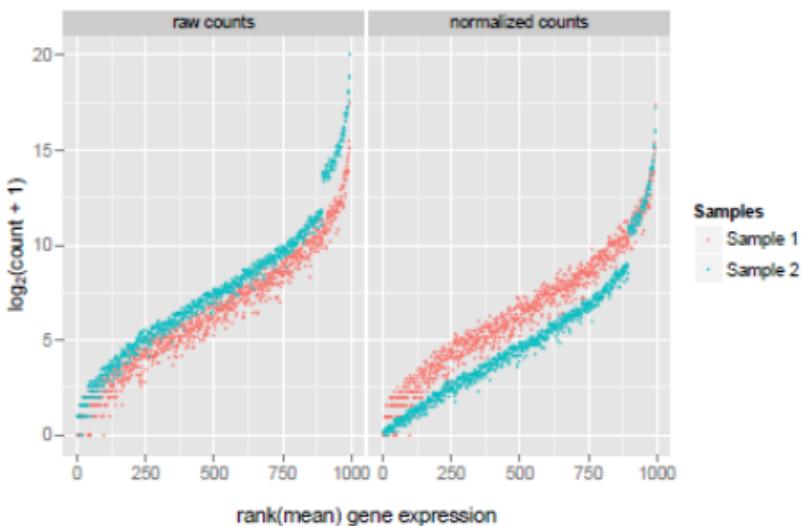


```
## With edgeR  
cpm(...,  
normalized.lib.sizes=TRUE)
```

Upper Quartile

Les comptages par gène sont divisés par le 3e quartile des comptages non nuls de l'échantillon, puis multipliées par la moyenne des 3e quartiles de tous les échantillons.

Ref Bullard et al., 2010 (Upper) Quartile normalization



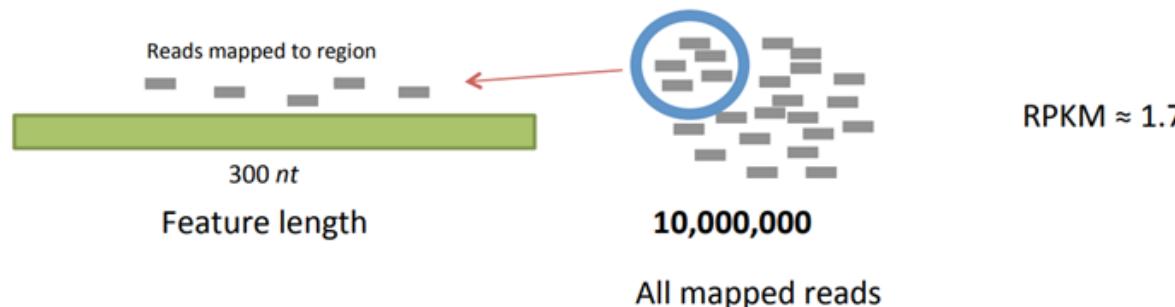
1 – dans lequel $Q(p)j$ est un quantile donné (généralement le 3e quartile) de la distribution des comptes dans l'échantillon j.

```
## With edgeR  
calcNormFactors(..., method = "upperquartile",  
                 p = 0.75)
```

Correcting for transcript length and total number of reads

RPKM

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 corrigés pour prendre en compte la taille de la librairie et la longueur des gènes. Cependant, il a été montré 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 exprimés. Cette méthode reste toutefois très populaire dans de nombreuses applications.



$$RPKM = 10^9 \times \frac{\text{Number of reads mapped to a region}}{\text{Total reads} \times \text{region length}}$$

RPKM: Reads Per Kilo base of transcript per Million reads