

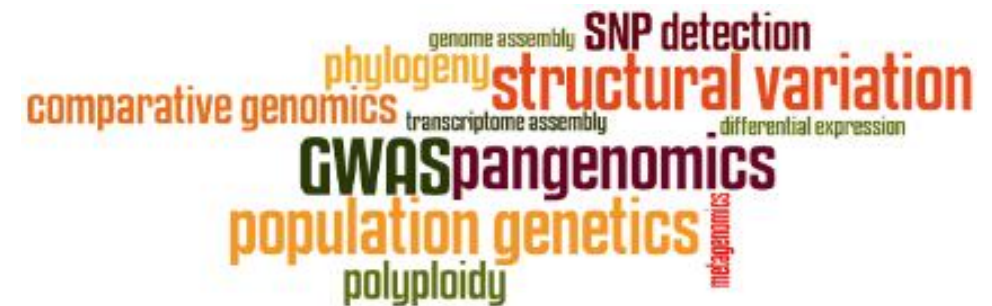
Session de formation 2023



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



Mutualisation



Cacao



Banana



Coffee



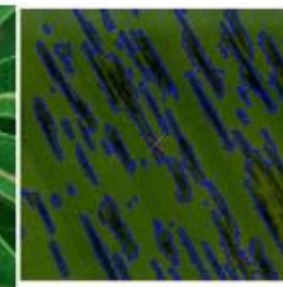
Rice



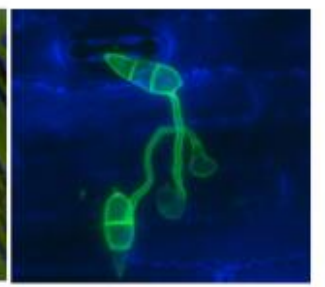
Palm



Cassava



Pseudocercospora



Magnaporthe

South Green

bioinformatics platform



4 institutes



3 research units



25+



Storage and computing
resources

Tools

Trainings



400+



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



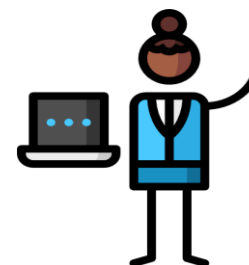
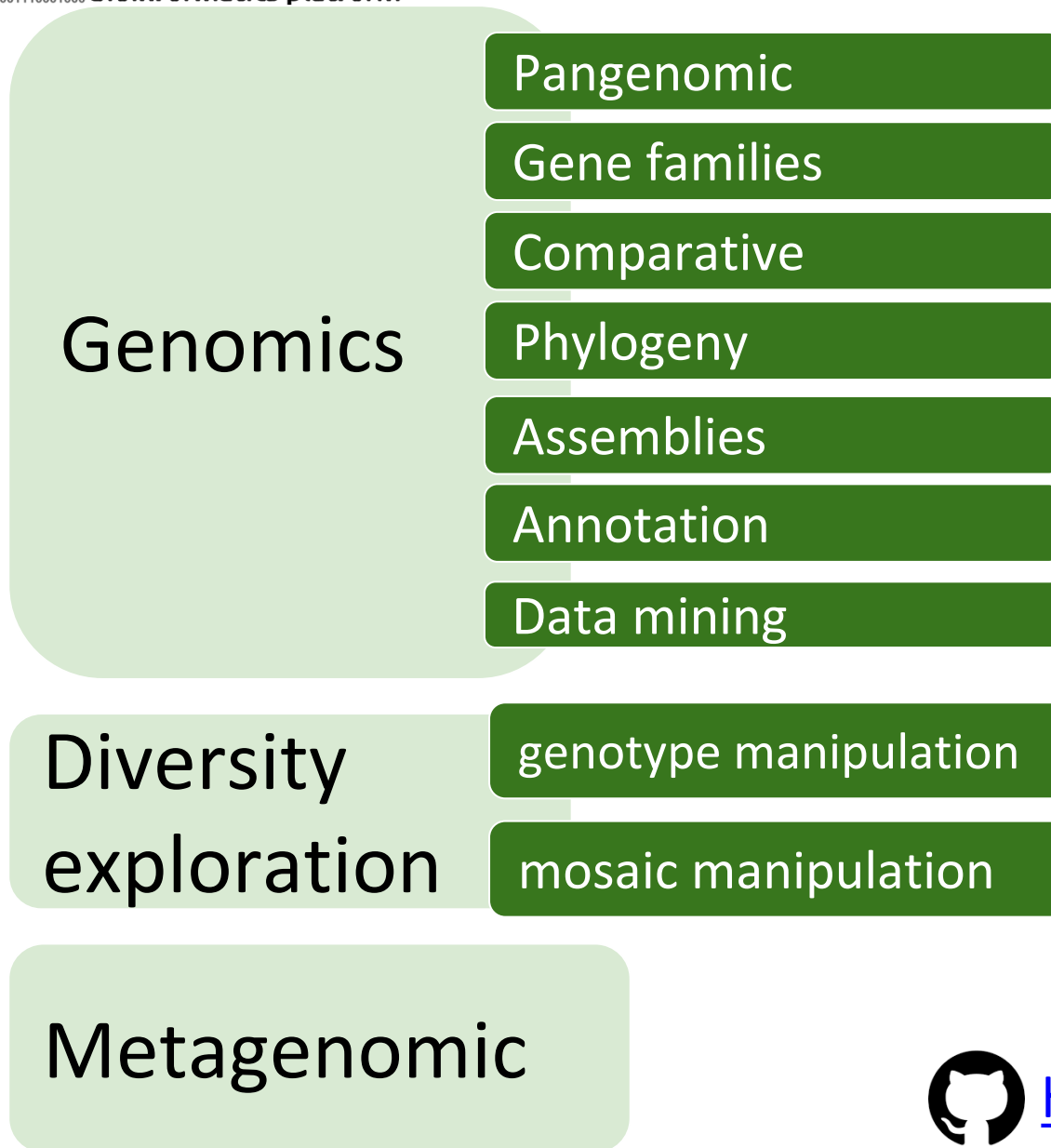
400+



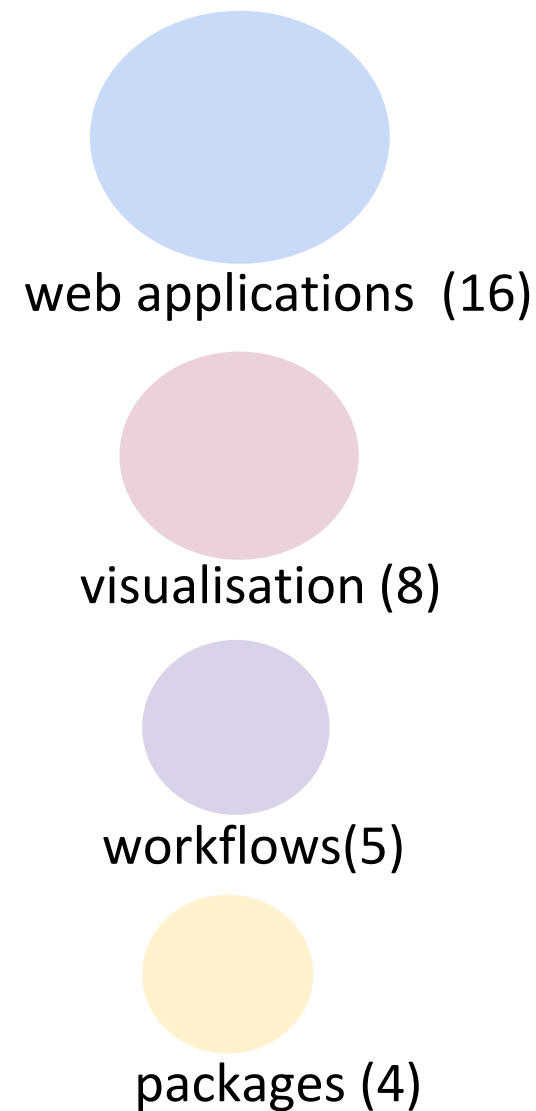
600+ tools

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

Collaborative development of tools



**+20
tools**



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

I-Trop

Plant & Health Bioinformatics Platform

IRD Institut de Recherche
pour le Développement
FRANCE

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



AUORE
COMTE



JACQUES
DAINAT



ALEXIS
DEREEPER



BRUNO
GRANOILLAC



JULIE
ORJUELA-



NDOMASSI
TANDO



CHRISTINE
TRANCHANT

bioinfo@ird.fr



[@ItropBioinfo](https://twitter.com/ItropBioinfo)



Florian Charriat
Antoni Exbrayat



Guilhem Sempere



Bruno Granouillac
Jacques Dainat



Nicolas Fernandez



Thomas Denecker

And more collaborators !



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



Comte Aurore
Dereeper Alexis
Ravel Sébastien



Bocs Stephanie
Boizet Alice
De Lamotte Frédéric
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

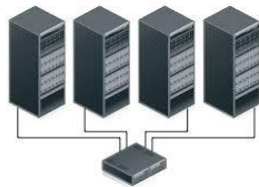
Workflow manager

TOGGLE
Toolbox for generic NGS analyses

SNAKEMAKE

Galaxy

HPC and trainings....



IRD
Institut de Recherche
pour le Développement

cirad

37 courses organized last 7 years



Genome Hubs & Information System



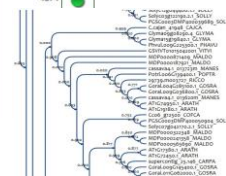
Gigwa

SNPs and Indels

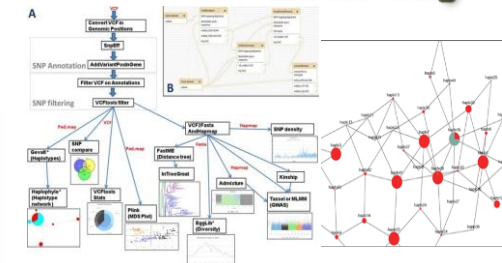
GreenPhyl

Family Id	Family Name	Number of sequences	Status
GP000015	Cytochrome P450 superfamily	6942	●
GP000017	AP2/ERF transcription factor family: ERF-ERF3 group (partial)	5142	●
GP000020	NAC transcription factor family	4574	●
GP000028	MADS transcription factor family		
GP000018	Haem peroxidase superfamily		
GP000065	General substrate transporter superfamily		
GP000022	Subtilisin-like Serine Proteases family		
GP000019	NPF, NRT1/PTR FAMILY		

Gene families



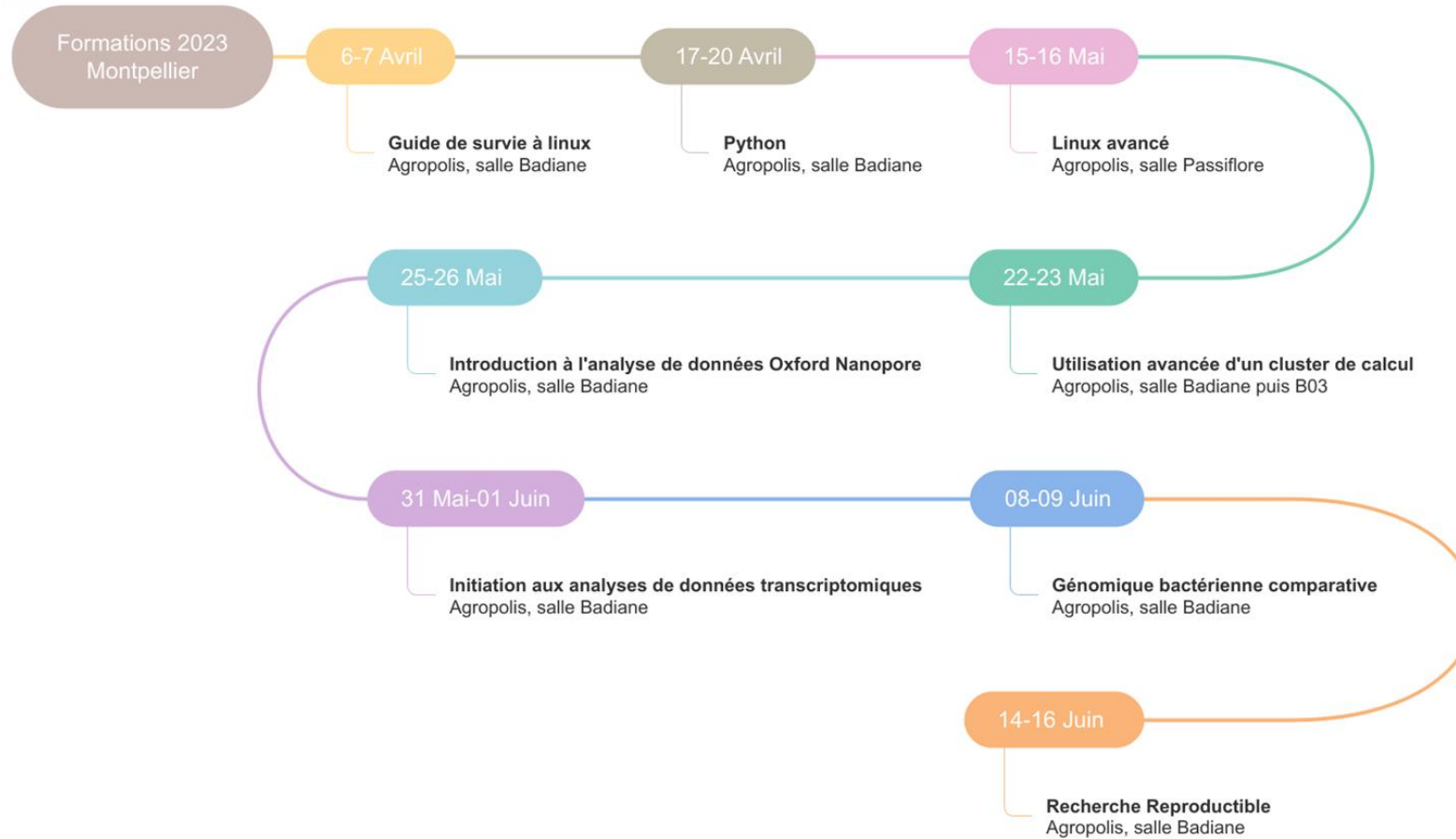
SNiPlay



<https://github.com/SouthGreenPlatform>



@green_bioinfo



Modules de formation 2023

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

Lancement des VM

- Sur Biosphere:
<https://biosphere.france-bioinformatique.fr/catalogue/>
- VM RNASeq_SG
- Configurer
- Groupe:RNA-Seq SG
- Gavarit Xlarge (4 vCPUs)

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
- Détecter les gènes différentiellement exprimés entre 2 conditions

Applications

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

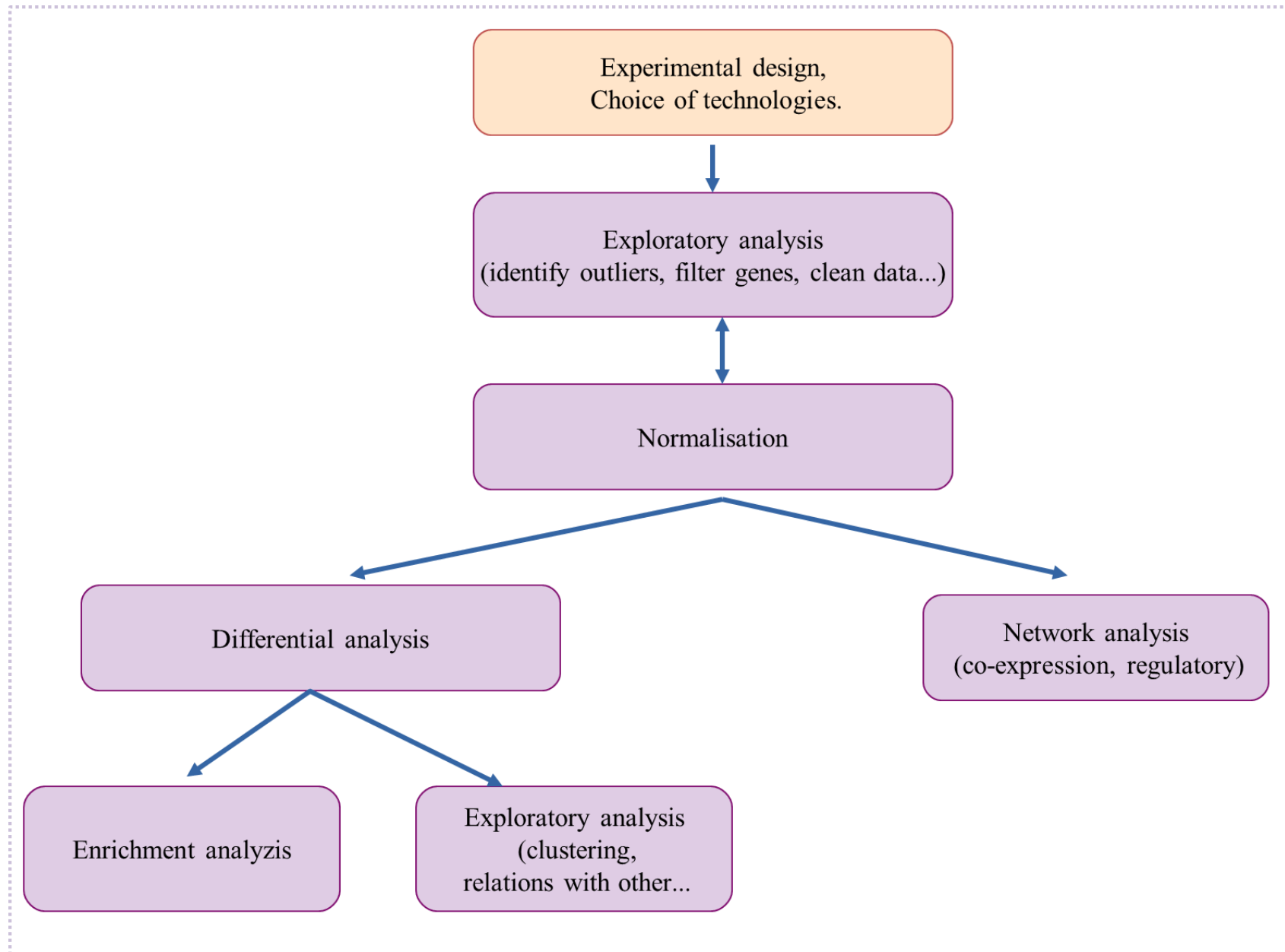
Pourquoi faire du RNAseq ?

- **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



Basic experiment : trouver les différences entre condition
contrôle/traitée

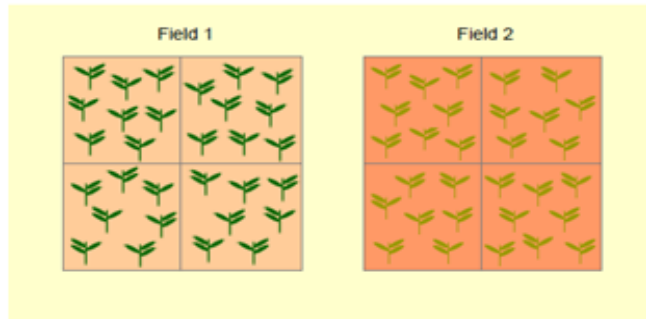


control group plant



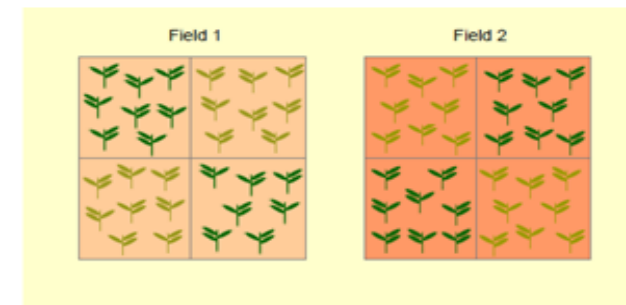
treated group plant

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



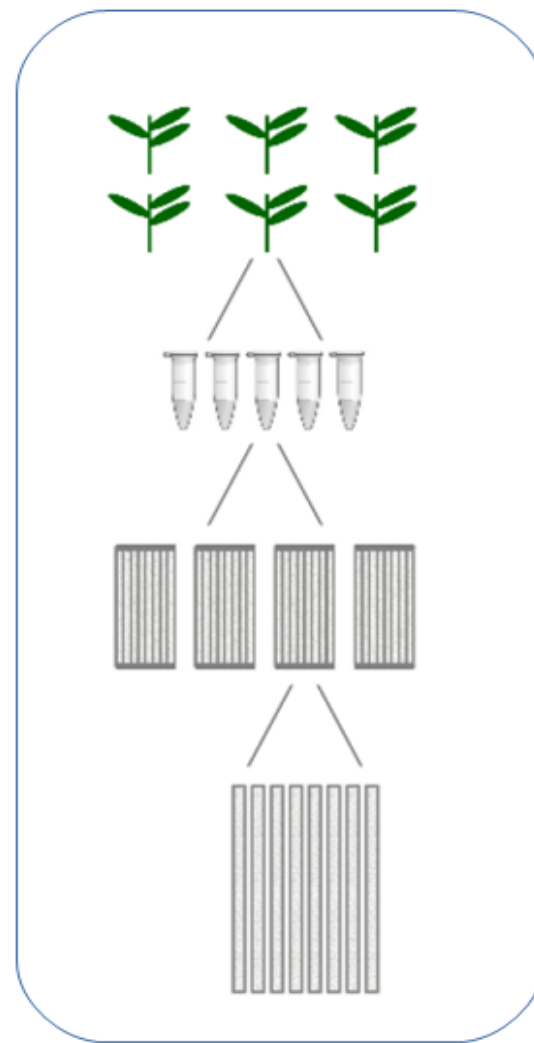
Pas de possibilité de différencier
l'effet champs de l'effet traitement

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

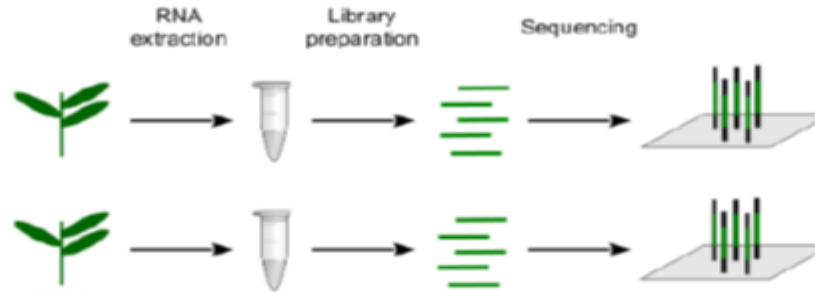
2 – Variations techniques :
effet de la préparation
des librairies

cDNA on lane of flowcell

3 – Variation techniques : effet des
lane et des flowcell

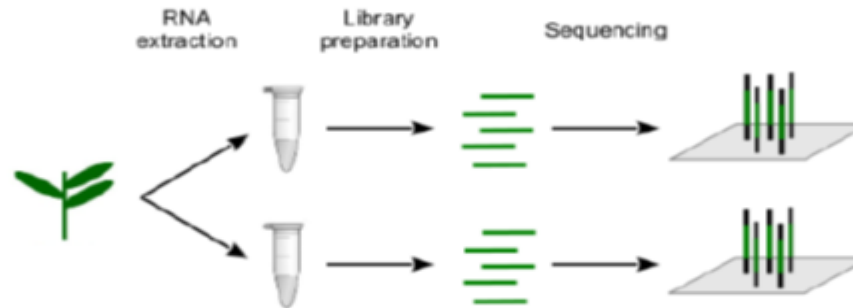
Effet lane < Effet Flowcell < Effet de la préparation de la librairie < < Effet biologique

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



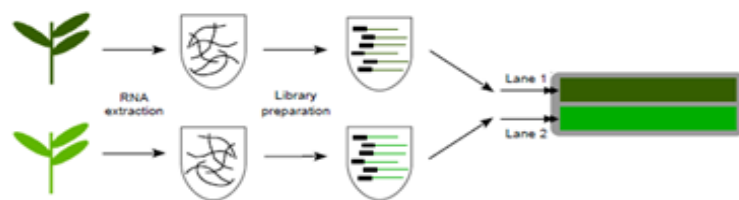
Réplicat Technique : Même matériel biologique, répété plusieurs fois indépendamment des étapes techniques.

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



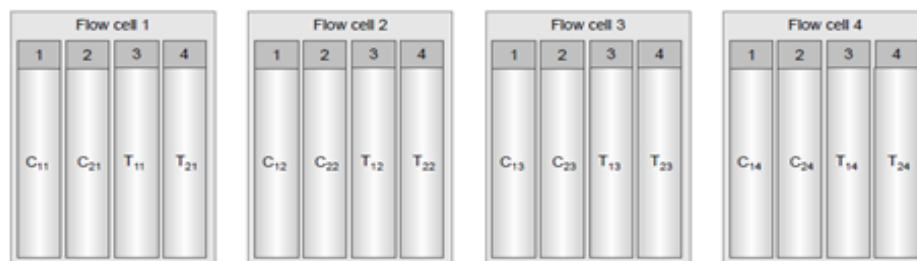
Échantillons séquencés sur deux lanes différentes.

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



Exemple :

2 réplicats biologiques par condition et
4 réplicats techniques par réplicats
biologiques



1

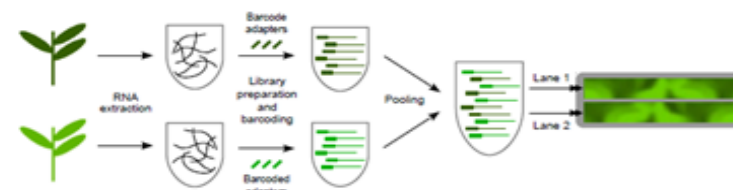
2

3

4

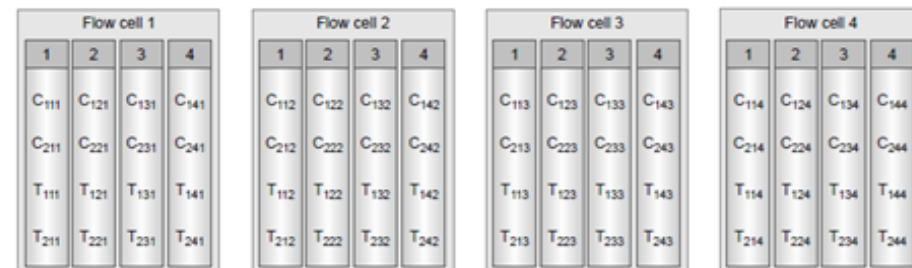
Multiplexes RNAseq plan d'expérimentation :

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



Exemple :

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

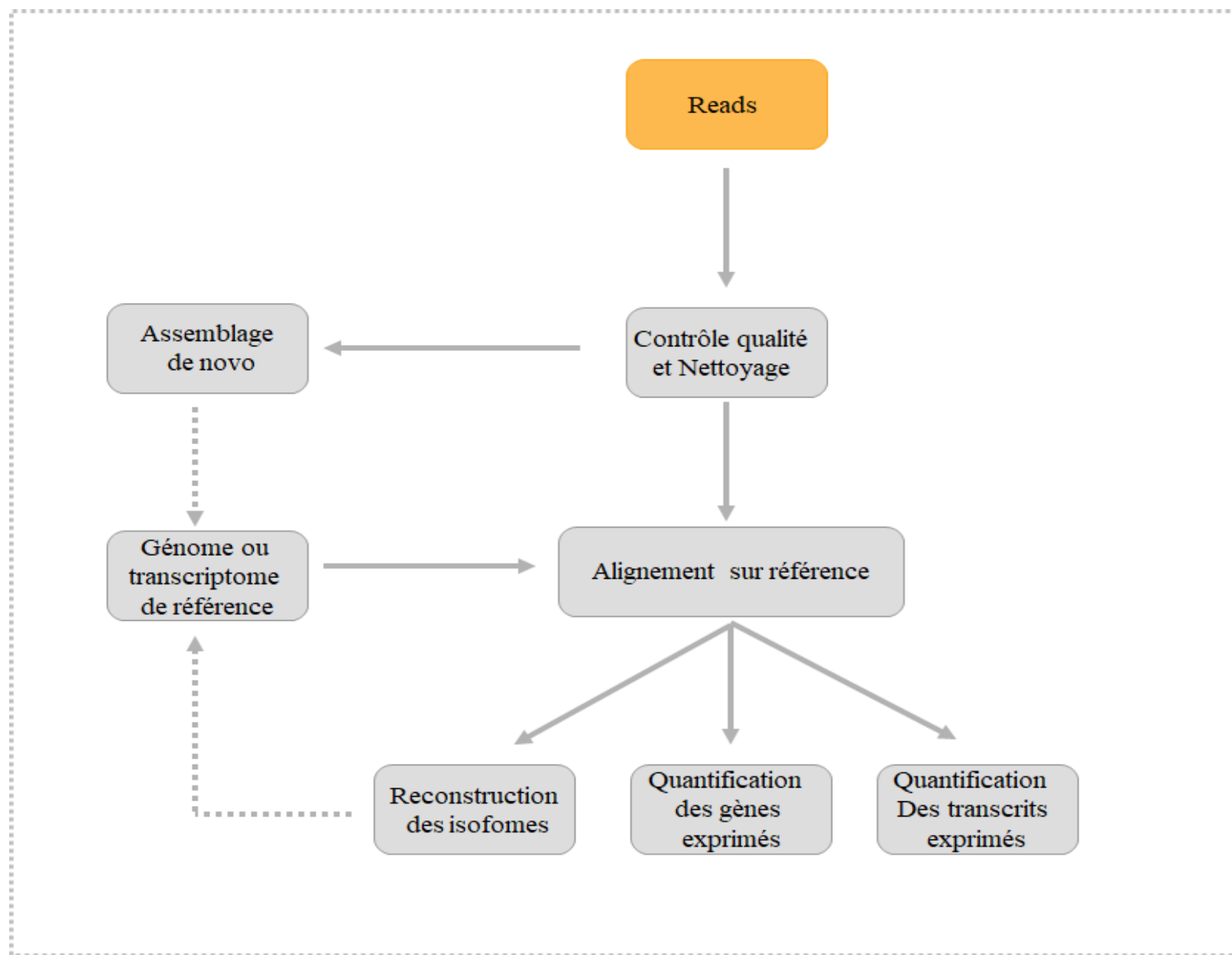


1

2

3

4



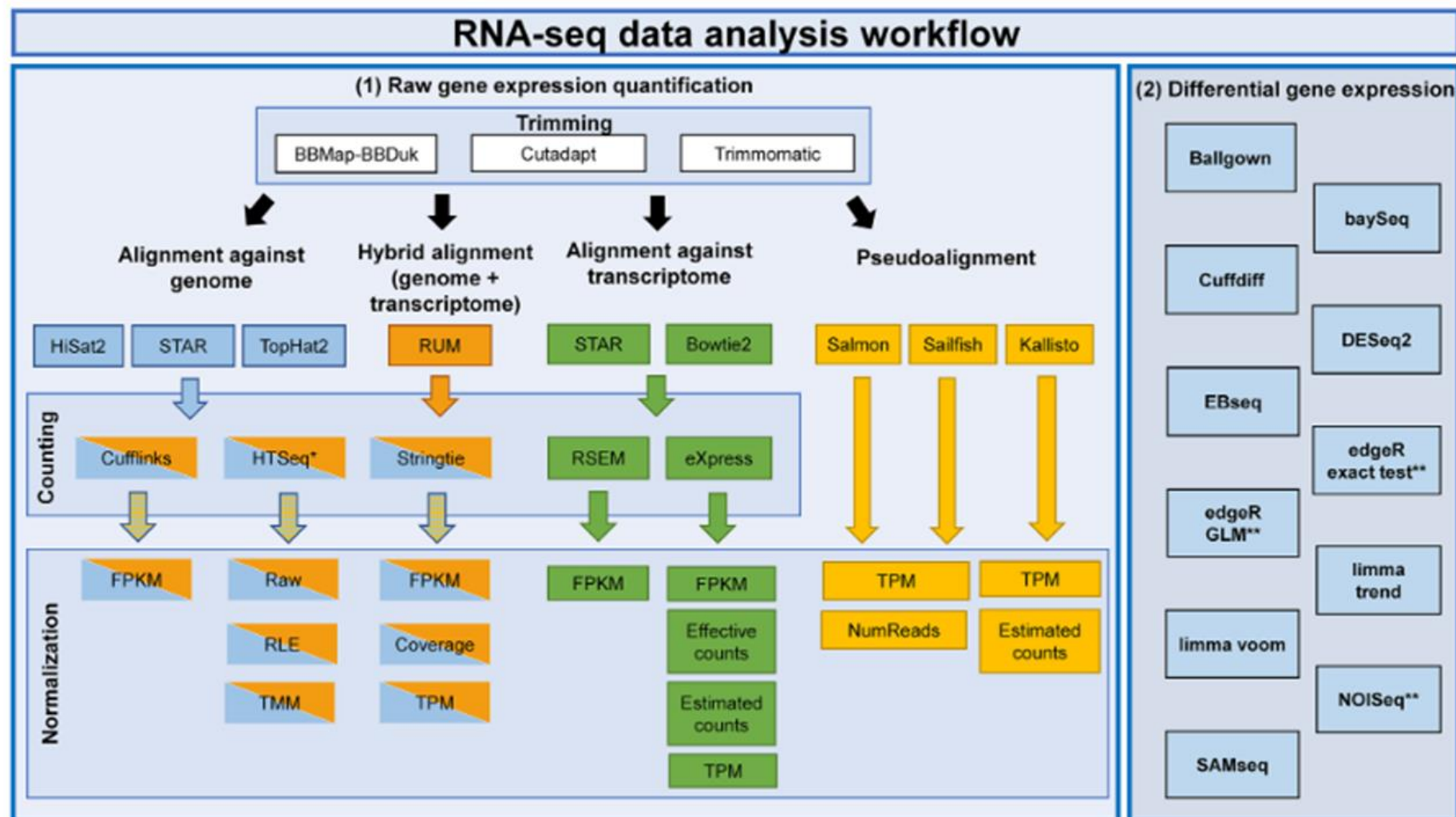


Figure 1. RNA-seq analysis workflow. Left panel (1) represents the raw gene expression quantification workflow. Every box contains the algorithms and methods used for the RNA-seq analysis at trimming, alignment, counting, normalization and pseudoalignment levels. The right panel (2) represents the algorithms 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

Reads



Mapping against
genome



Read clusters



Putative
transcripts



*de novo
assembly*

Genome based

*Genome guided
de novo*

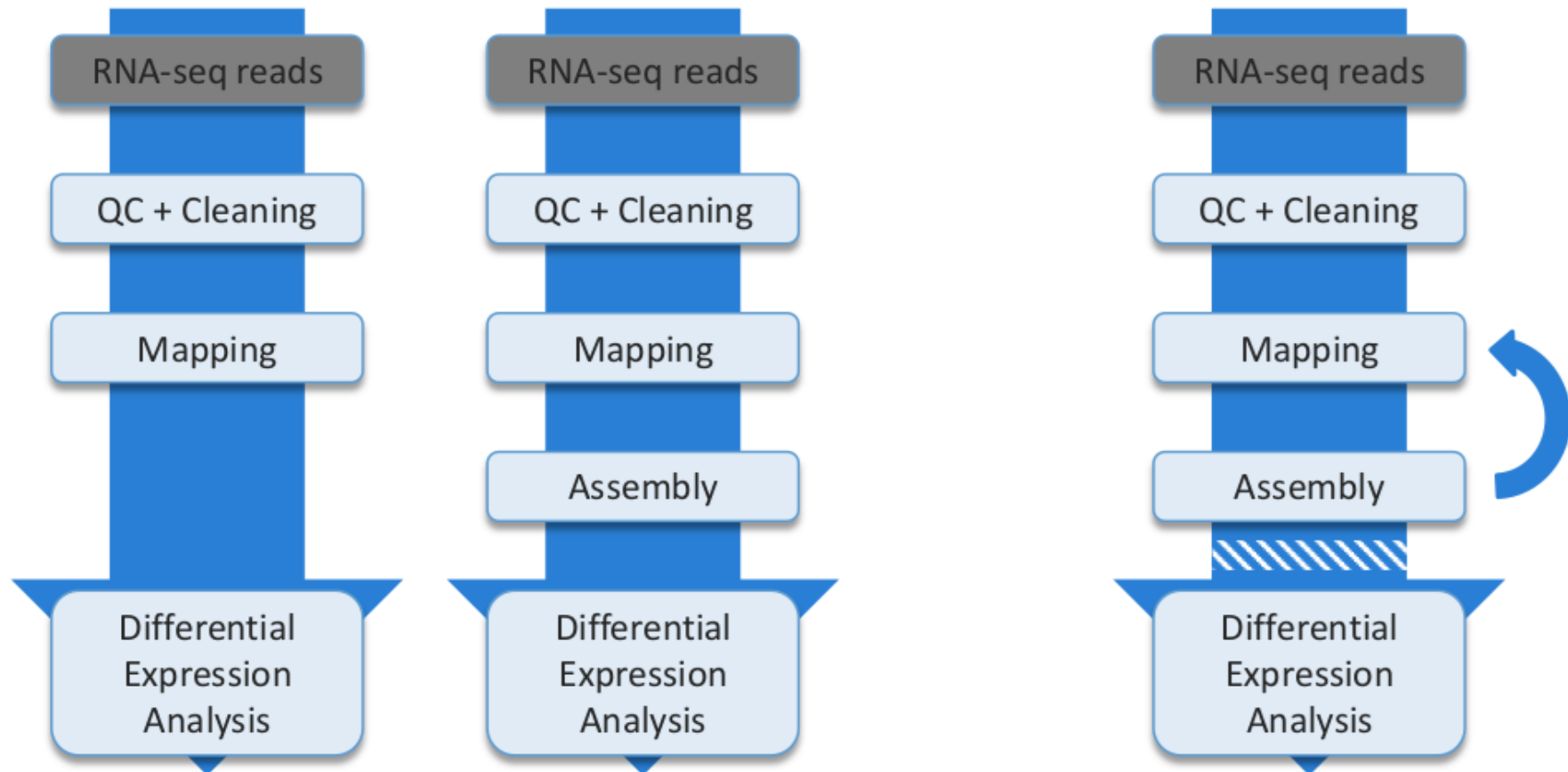
Vision Global of RNAseq Analysis

- Reference genome
- Reference transcriptome

- Reference genome
- No reference transcriptome

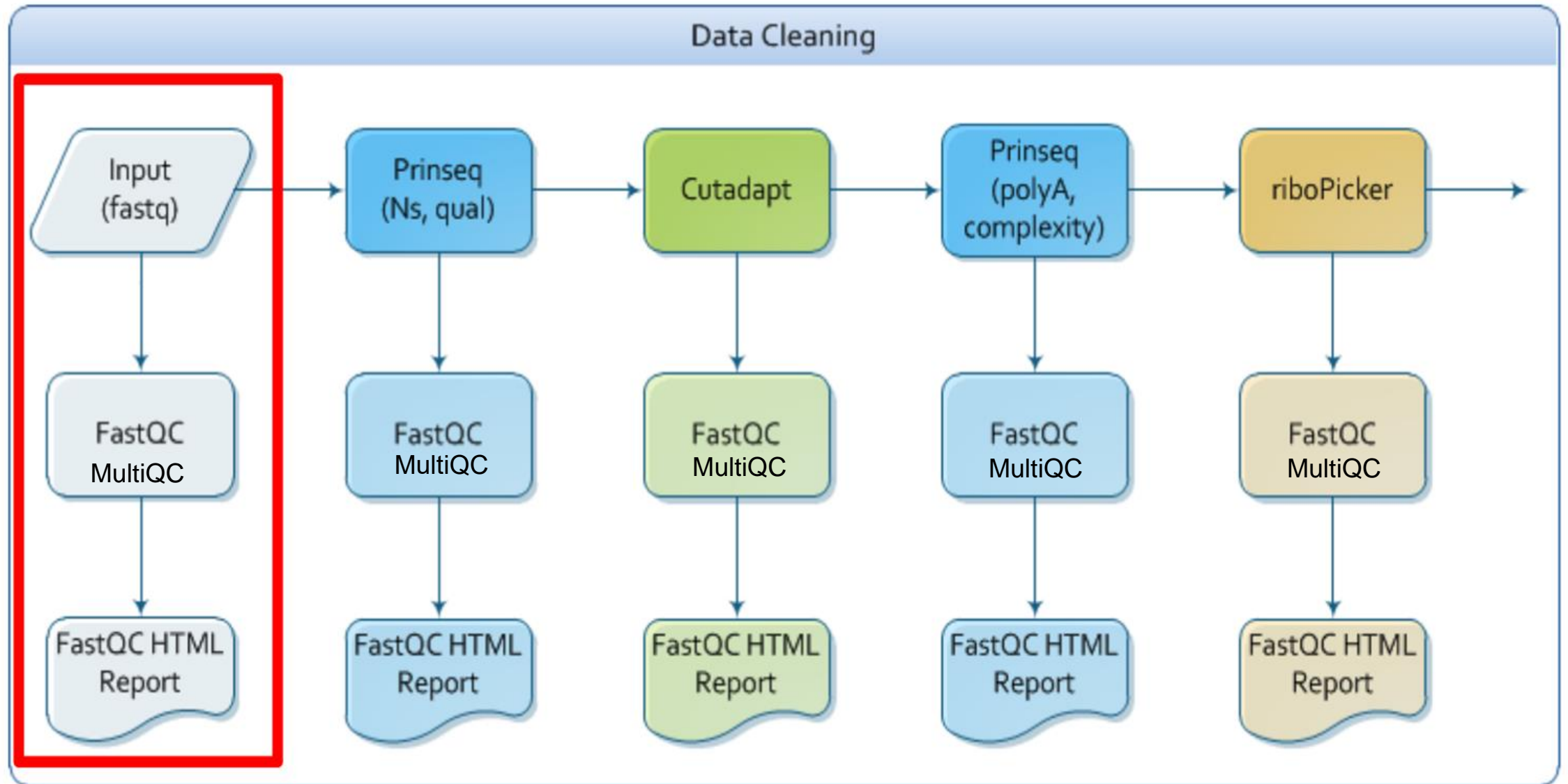
Non discovery mode

Discovery mode

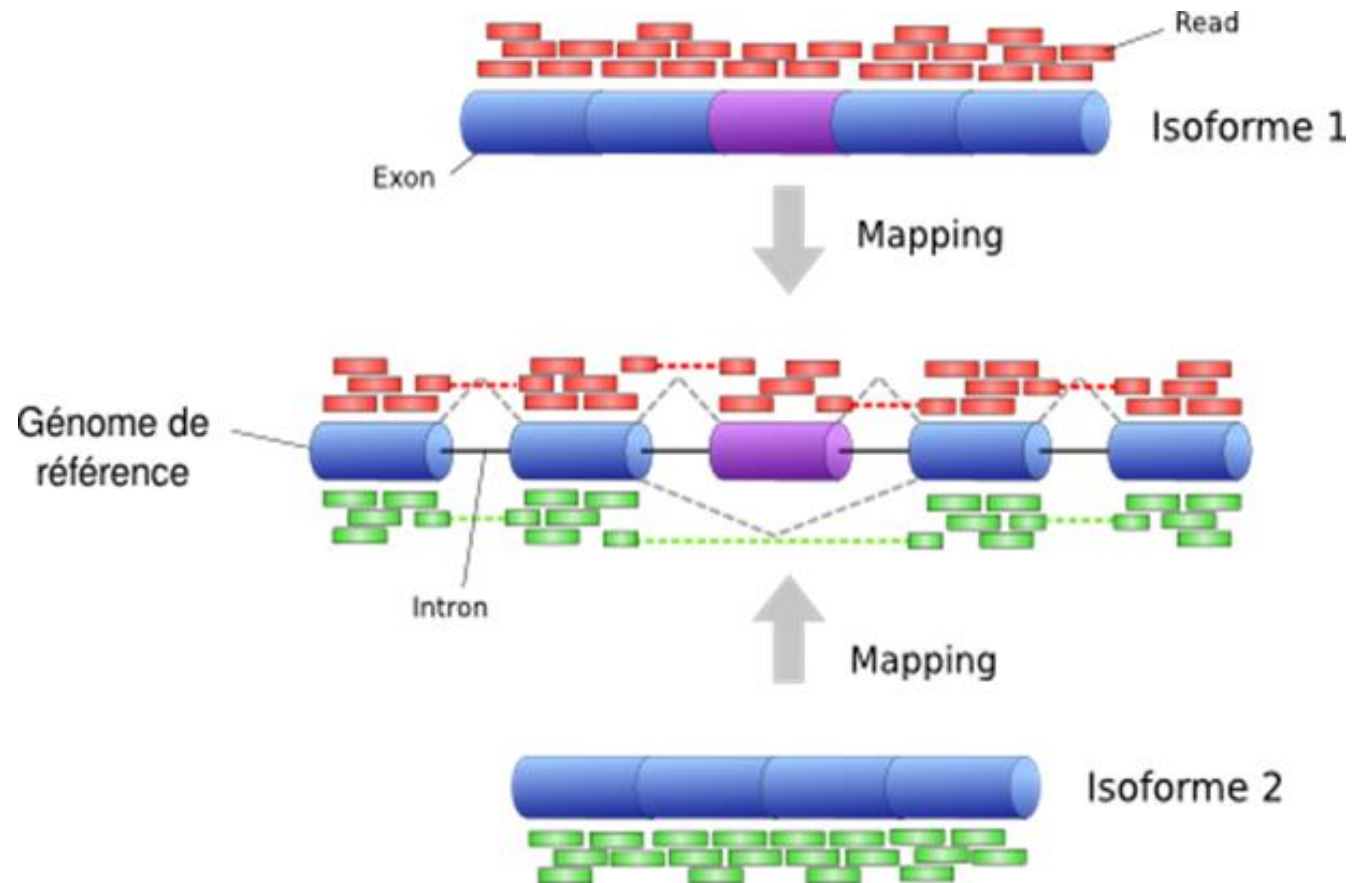


Vérification de la qualité des données NGS

	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 laissés pendant la préparation de la librairie	Des ponts entre séquences sans relation aucune: Chimères	<i>PRINSEQ2</i>
Contaminations	RNA Ribosomal RNA/DNA étrangère (PhiX, Bacteria, ...)		SortMeRNA, riboPicker, DeconSeq



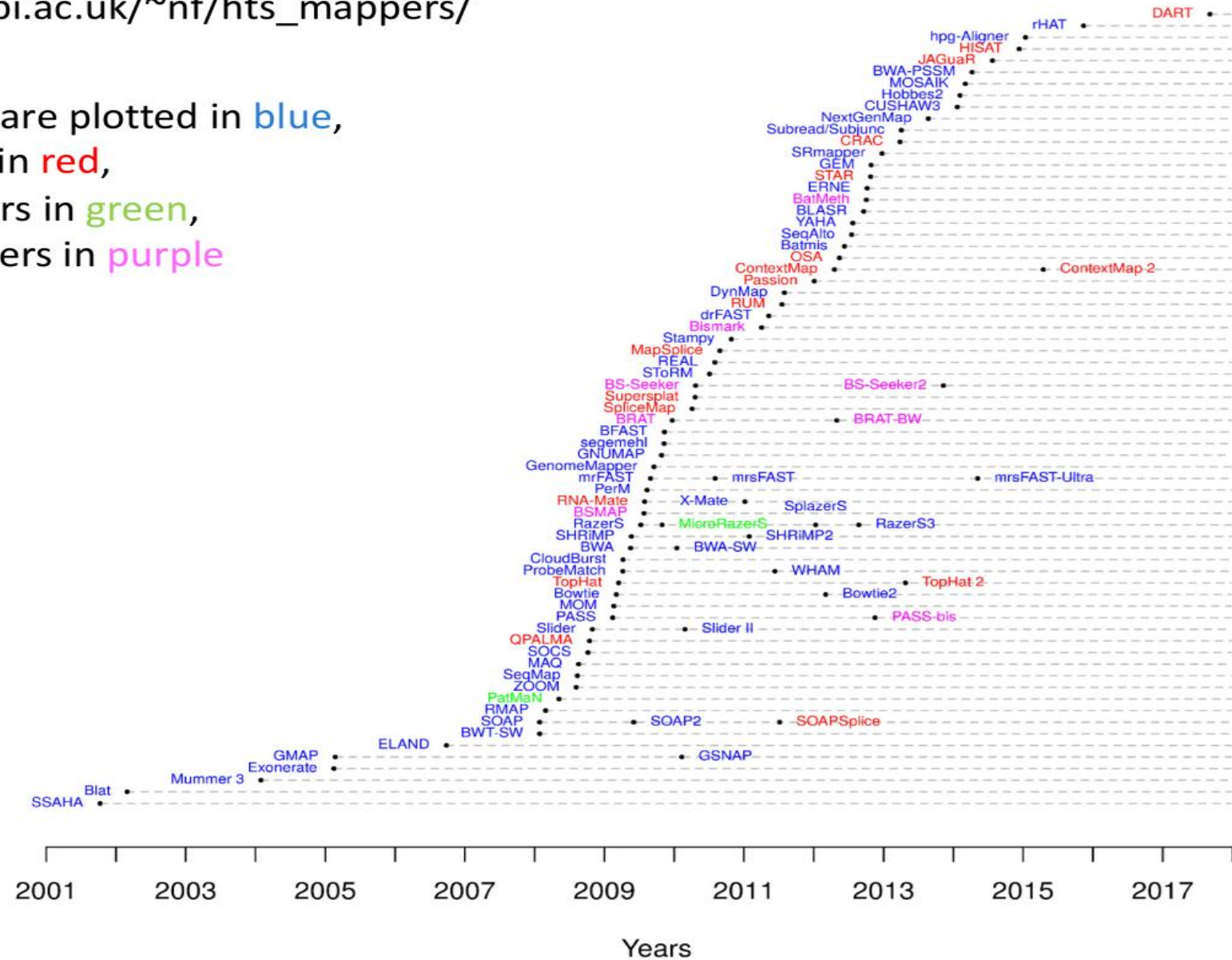
- Permet la mise en évidence d'isoformes
- Aide à l'annotation structurale du génome



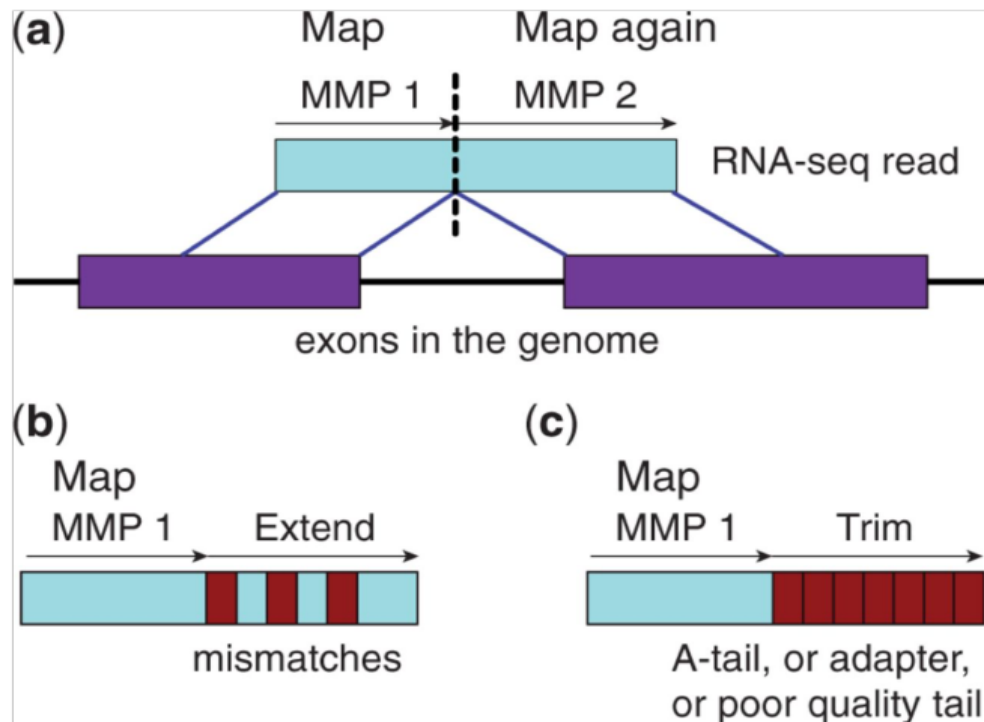
Choix de l'outil de mapping

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.

A. Dobin & *al.*, *Bioinformatics*, 2013

- Mapper tag

3- Comptage

mRNA-seq for measuring gene expression

Myers Lab

Selection of mRNA with polyA tail :



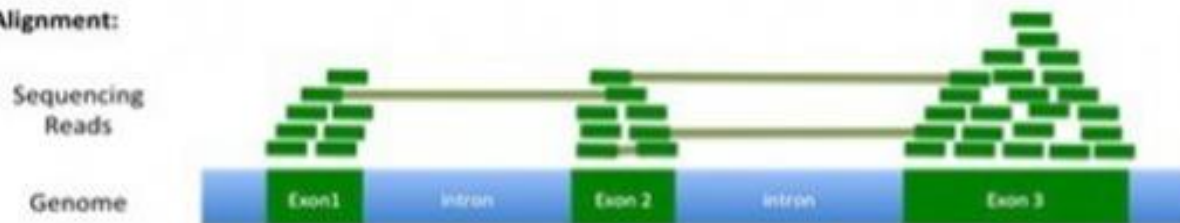
Random Primed cDNA synthesis :



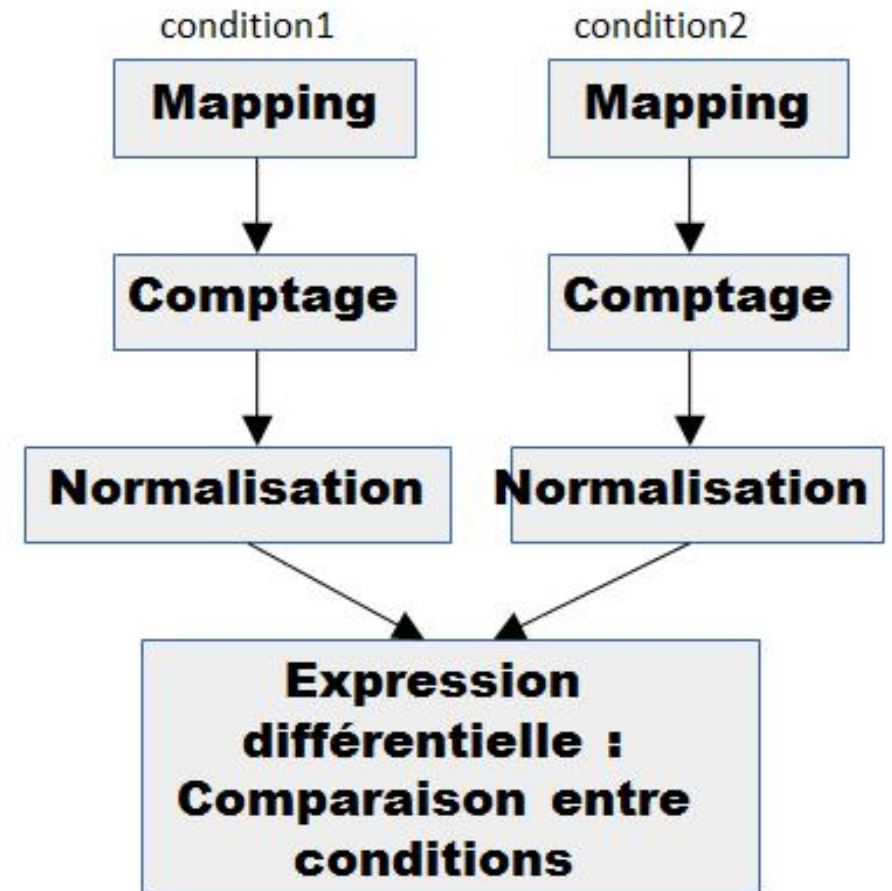
Paired-end sequencing of fragmented cDNA:



Alignment:



Quantify expression levels = RPKM (# of aligned Reads Per Kb of transcript per Million total reads)



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

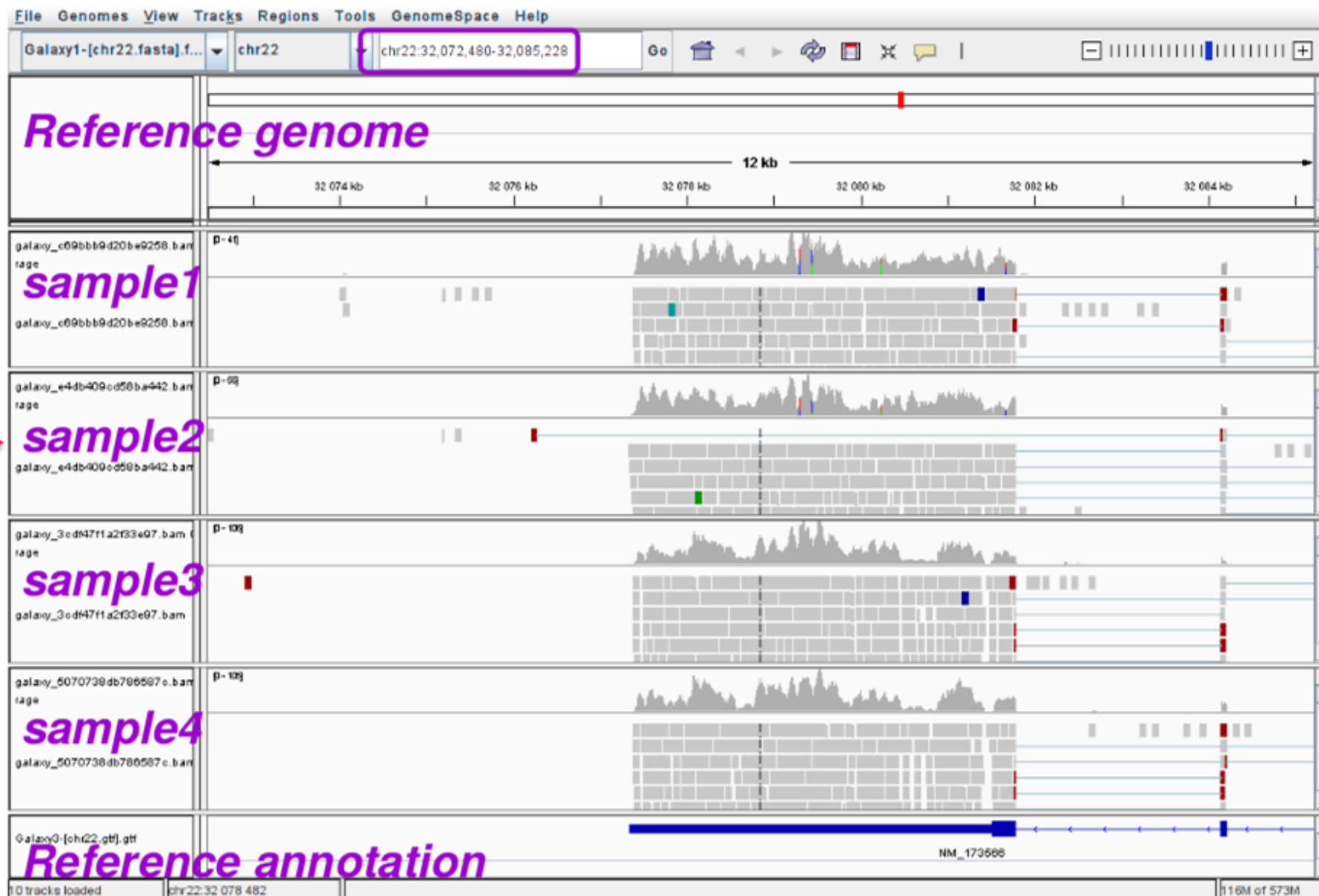
=> samtools idxstats

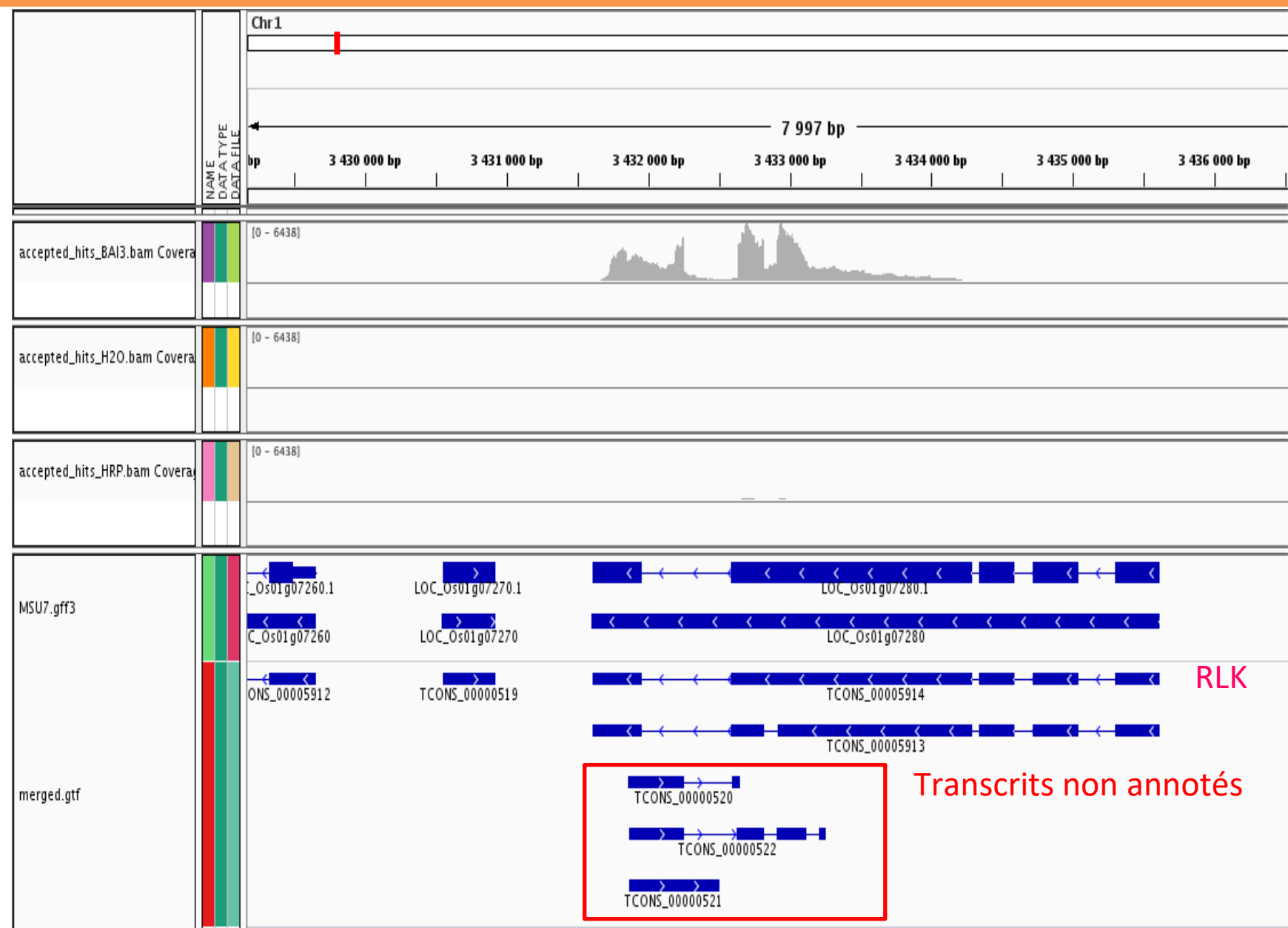
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

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

gff3

Seqname	Source			Score	Strand	Frame		
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



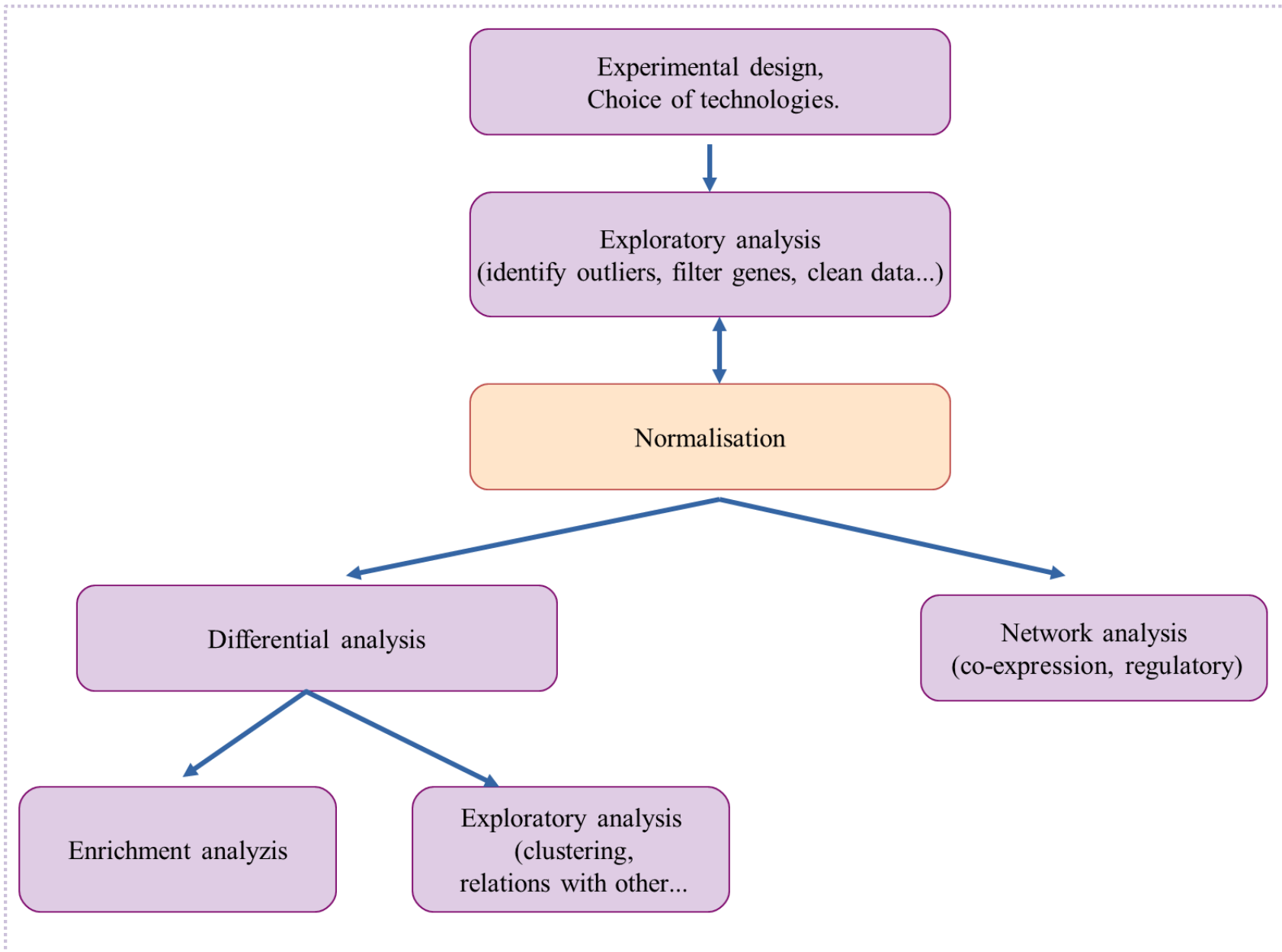


Recherche de gènes différentiellement exprimés

- 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



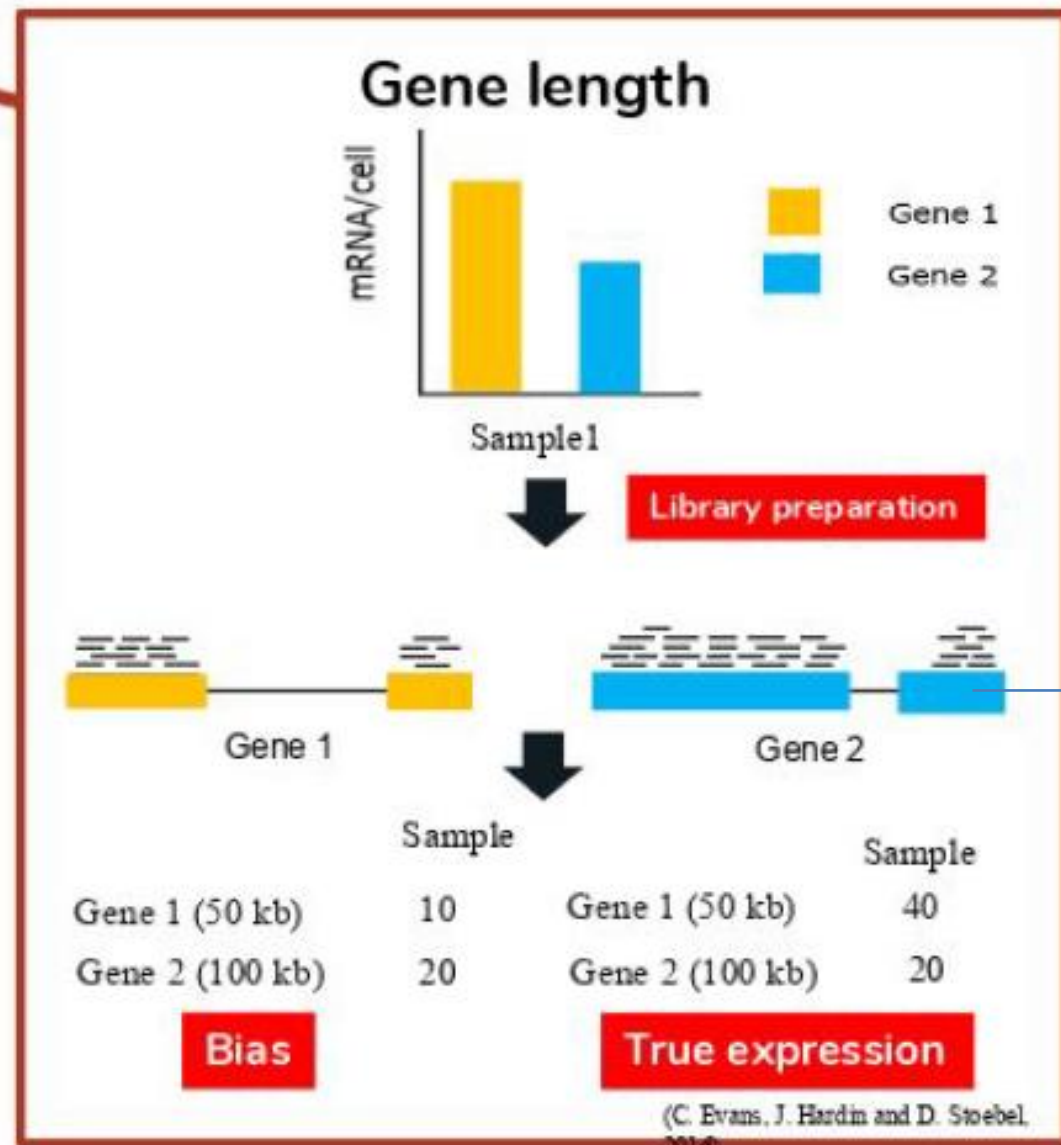
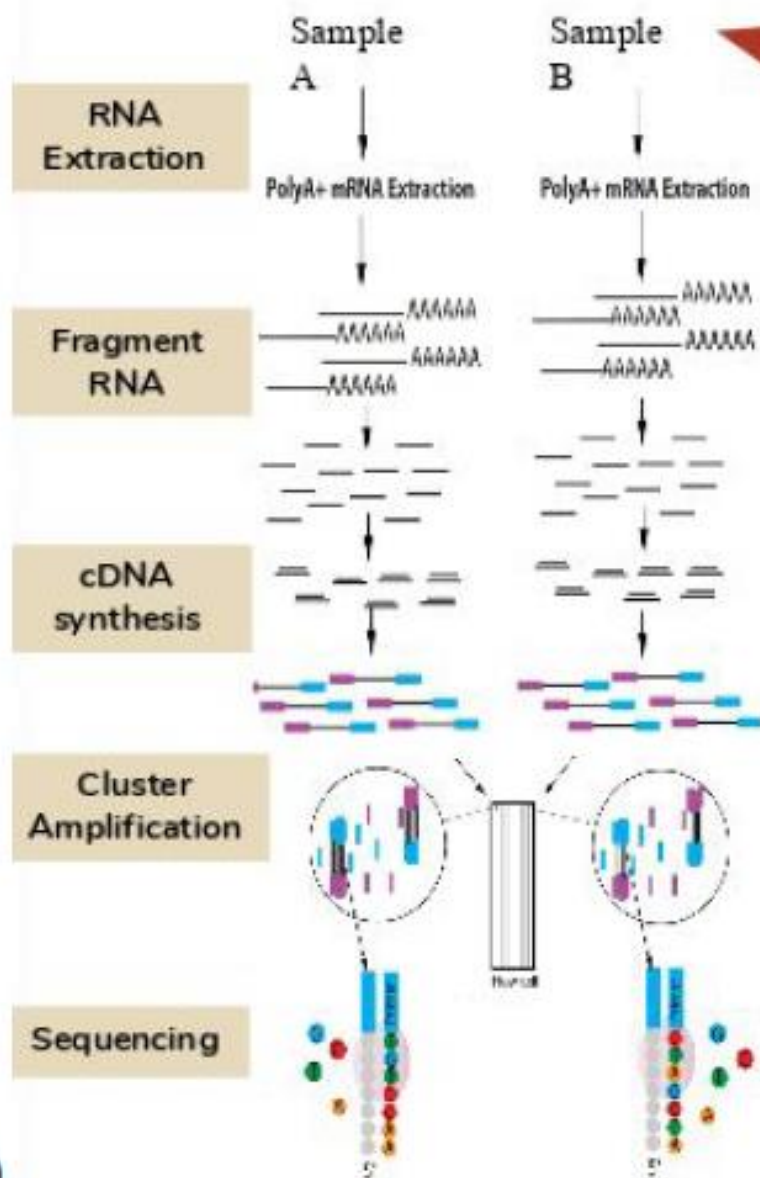
- **Identifier et corriger** les biais techniques dus au séquençage, pour les rendre comparable
 - S'assurer que les données sont exploitables
 - Réduire les biais techniques expérimentaux
 - De pouvoir comparer les données des différentes conditions entre elles
 - De s'approcher des hypothèses favorables pour l'analyse différentielle (distribution gaussienne des 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 ?**

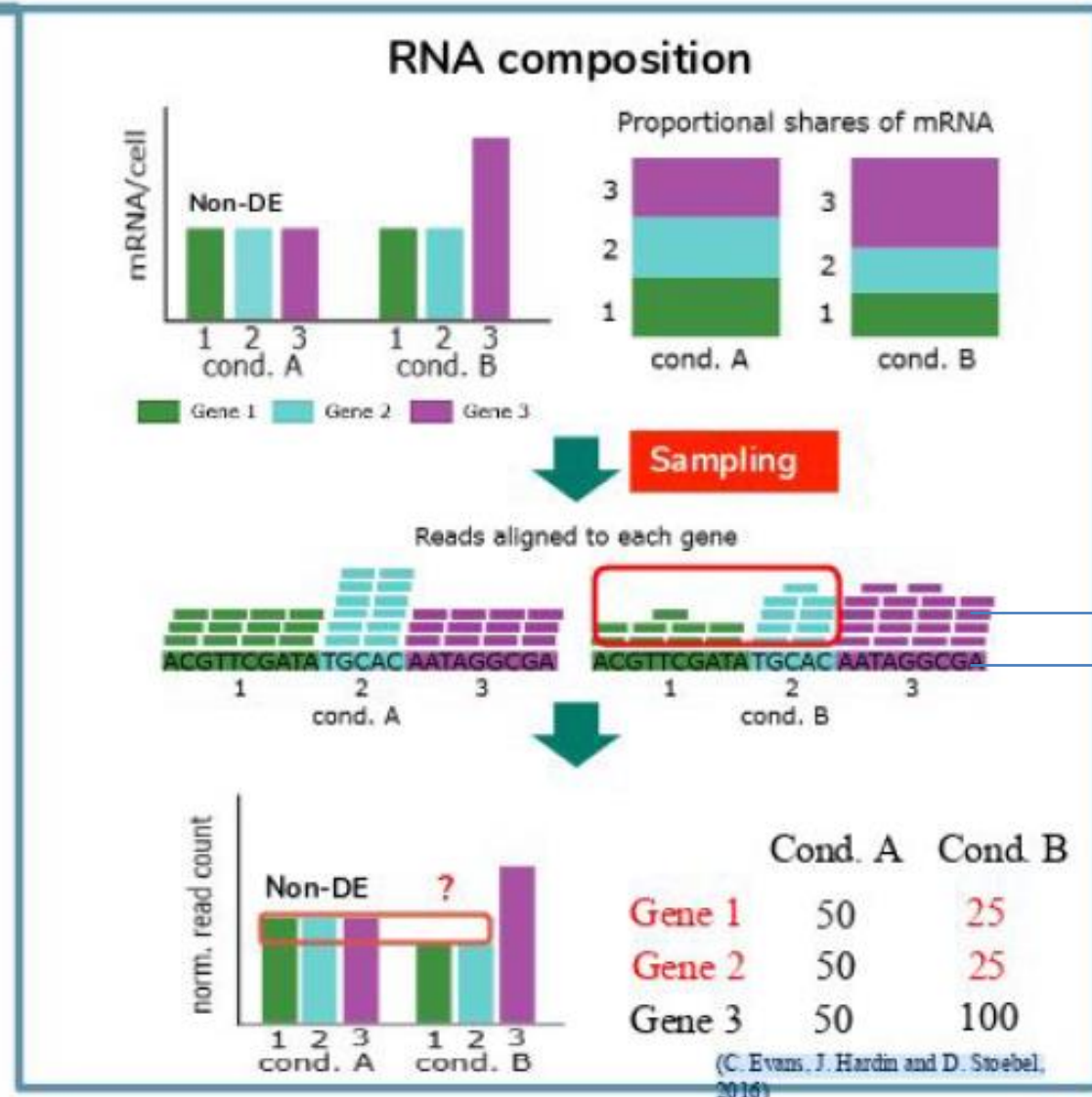
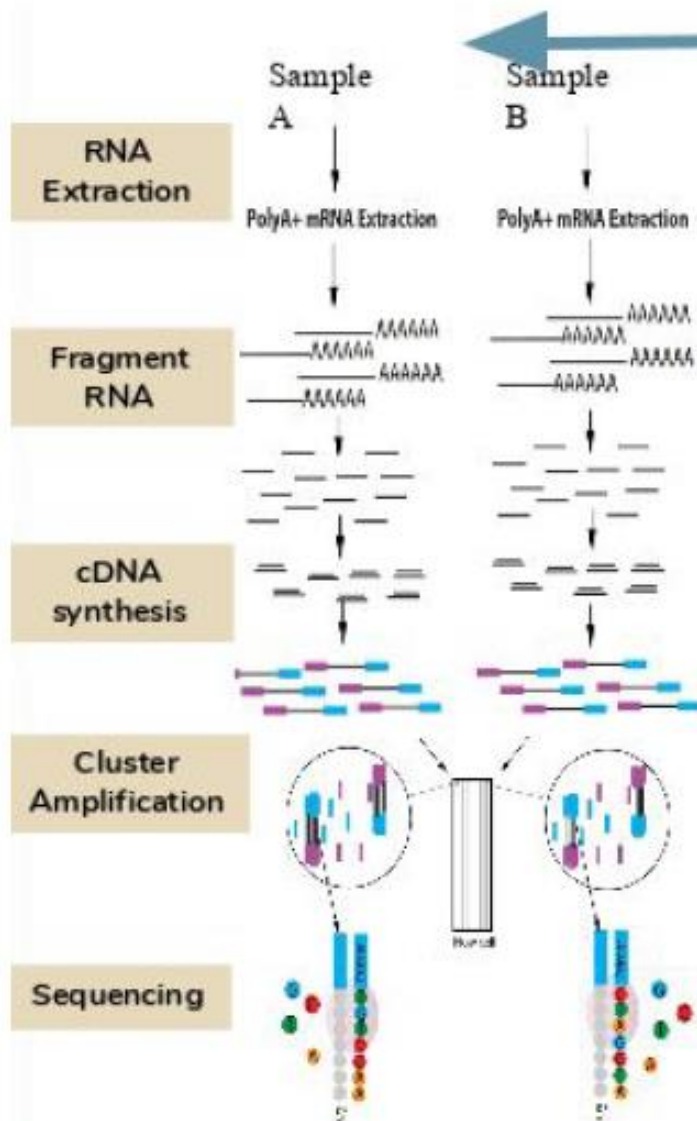
- Entre échantillon -> Comparer le niveau d'expression d'un gène entre différents échantillons
 - Profondeur du séquençage == taille de la banque
 - Biais d'échantillonnage durant la construction de la banque == effet batch
 - Présence de fragments majoritaires == saturation
 - Composition de la séquence due 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)



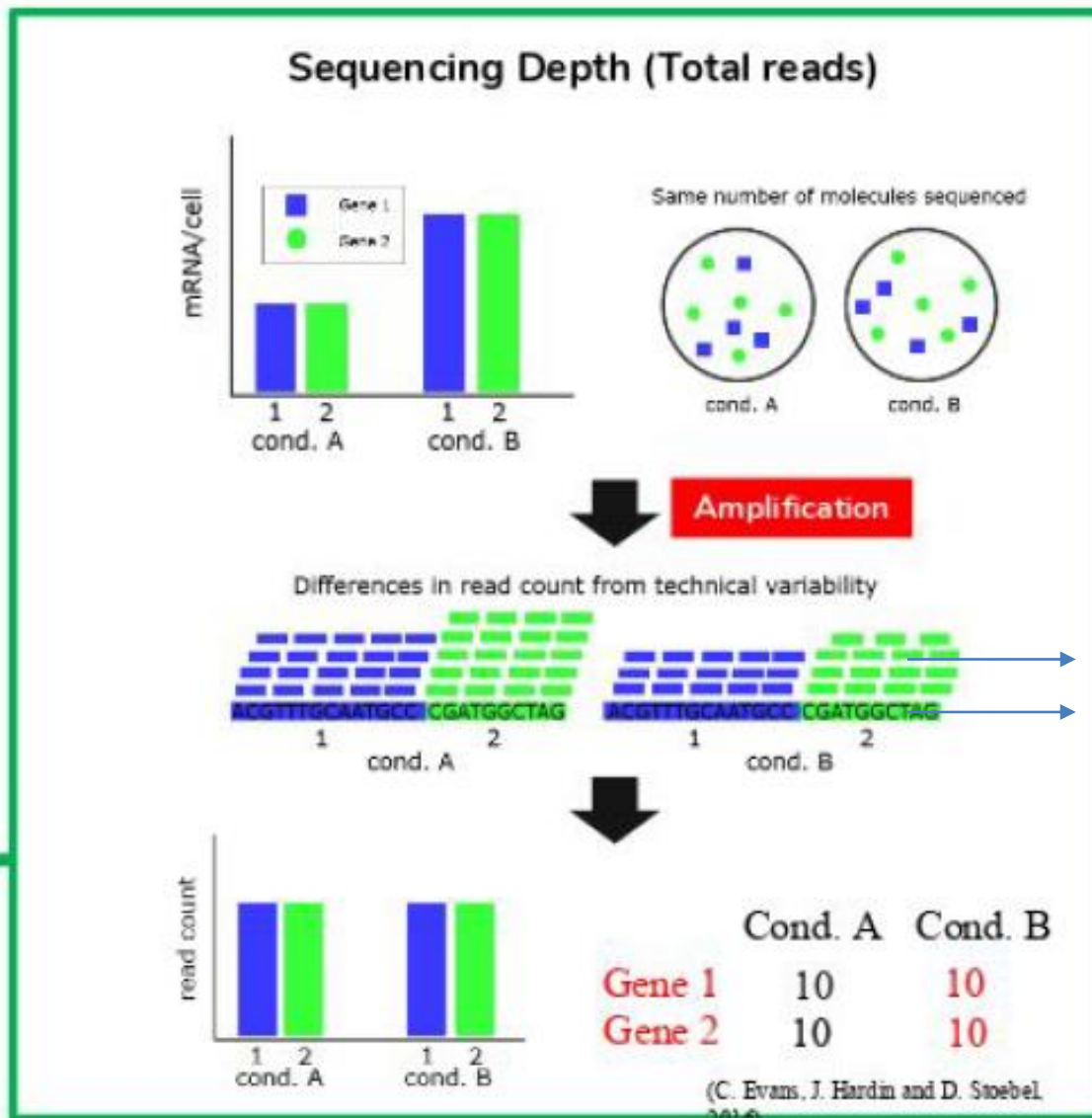
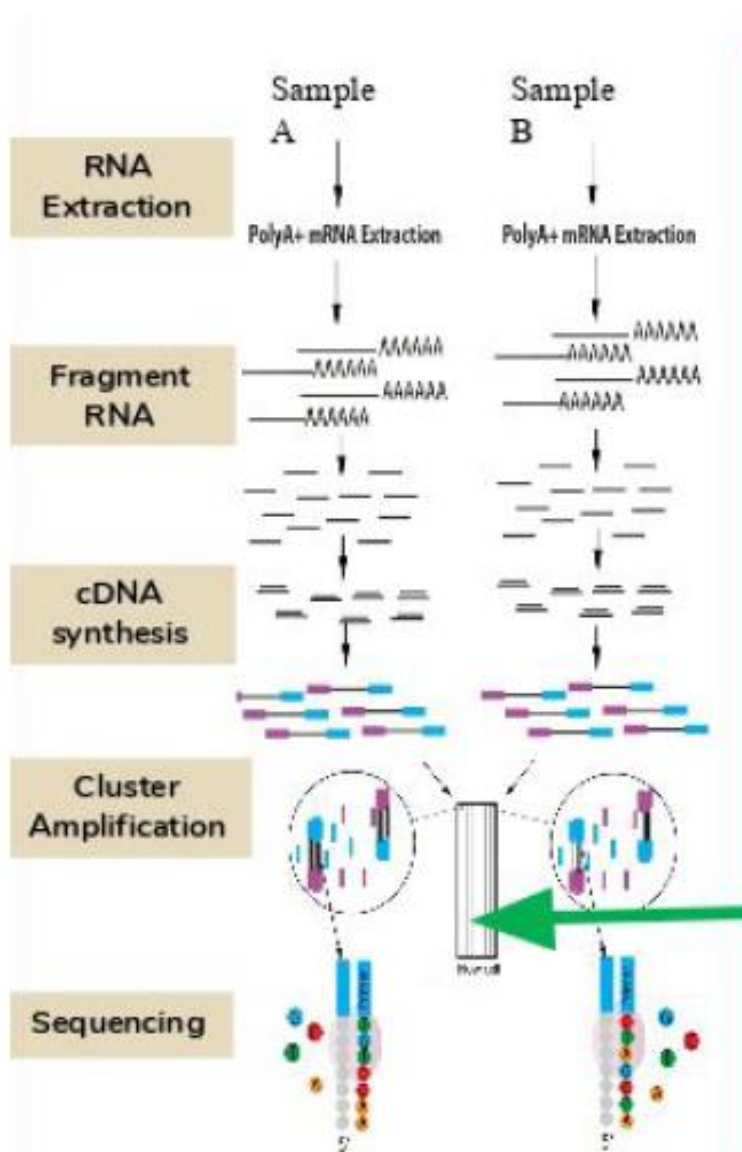


(B)

(C)



Reads
mRNAs



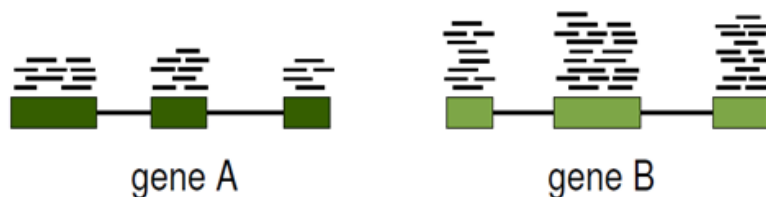
(A)

- **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)

	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.



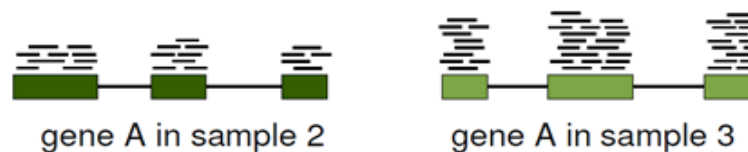
- Permettre la comparaison de gènes pour un même échantillon.
- Les sources de variabilités : longueur du gène et composition en GC.

	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.

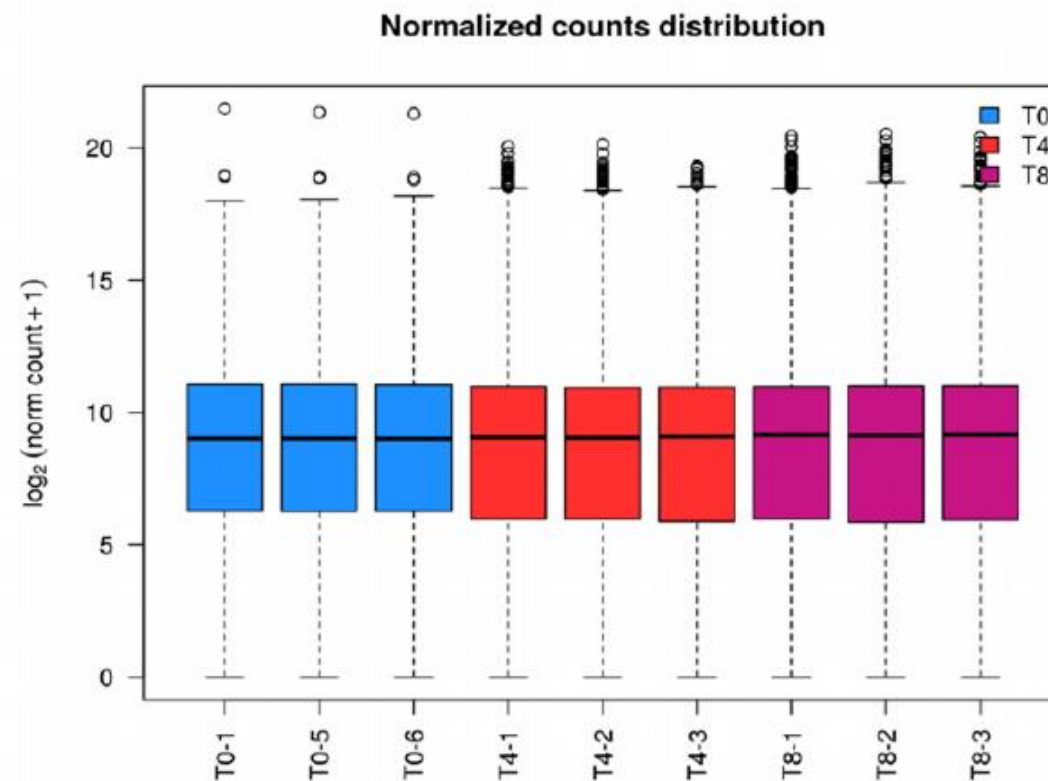
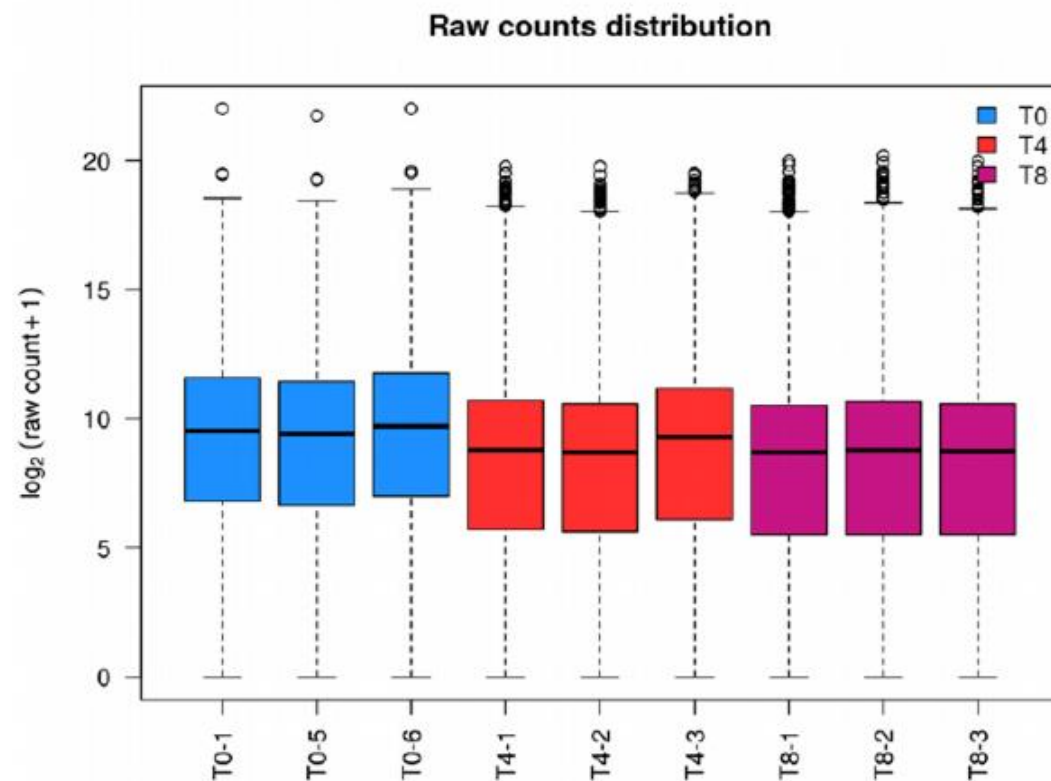


Le gène A est exprimé dans les échantillons 2 et 3, mais la profondeur de séquençage est plus importante dans l'échantillon 3 que dans le 2 (différences de taille des librairies).



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

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

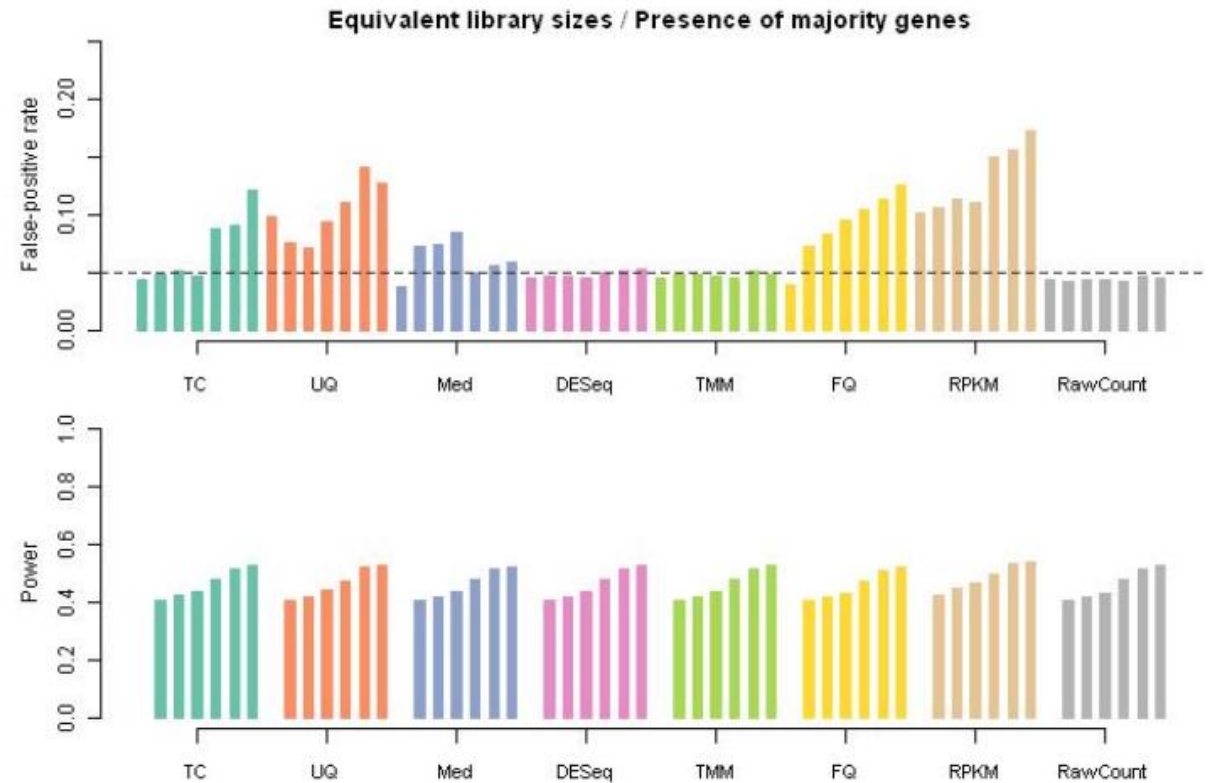
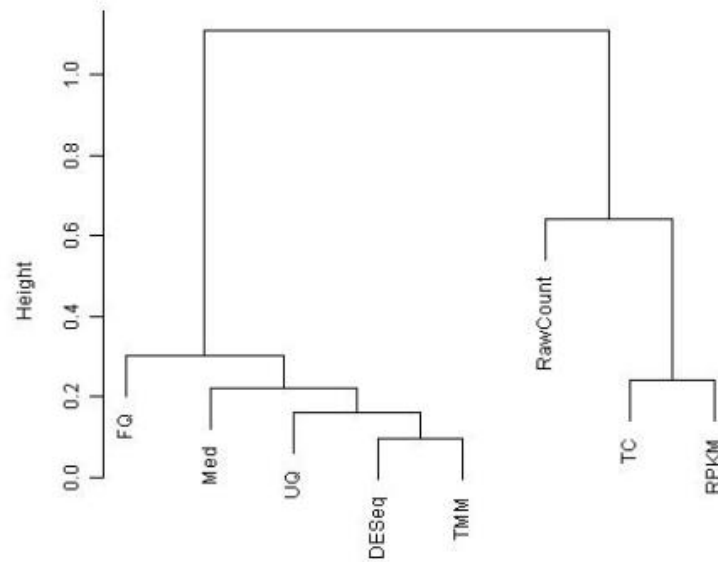
- **Total Count (TC)** : On divise chaque nombre de reads par le nombre total de reads (c'est-à-dire la taille de la banque) et on multiplie par le nombre moyen de reads des banques.
- **Upper Quartile (UQ)** :
 - **Objectif** : On applique la méthode TC en remplaçant le nombre total de reads par le 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 différenciellement 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 normalisation 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 et Inter-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 genes ne sont pas différentiellement 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.

Effect on differential analysis (DESeq v. 1.6):



Inflated FPR for all methods except for TMM and DESeq (RLE).

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**

Méthodes de Normalisation Intra-banques

Les biais 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 evidence en DE



- Corriger le Bias

5- Choix de la méthode de normalization

Méthodes de Normalisation

RLE , TMM, Upper-
Quartile

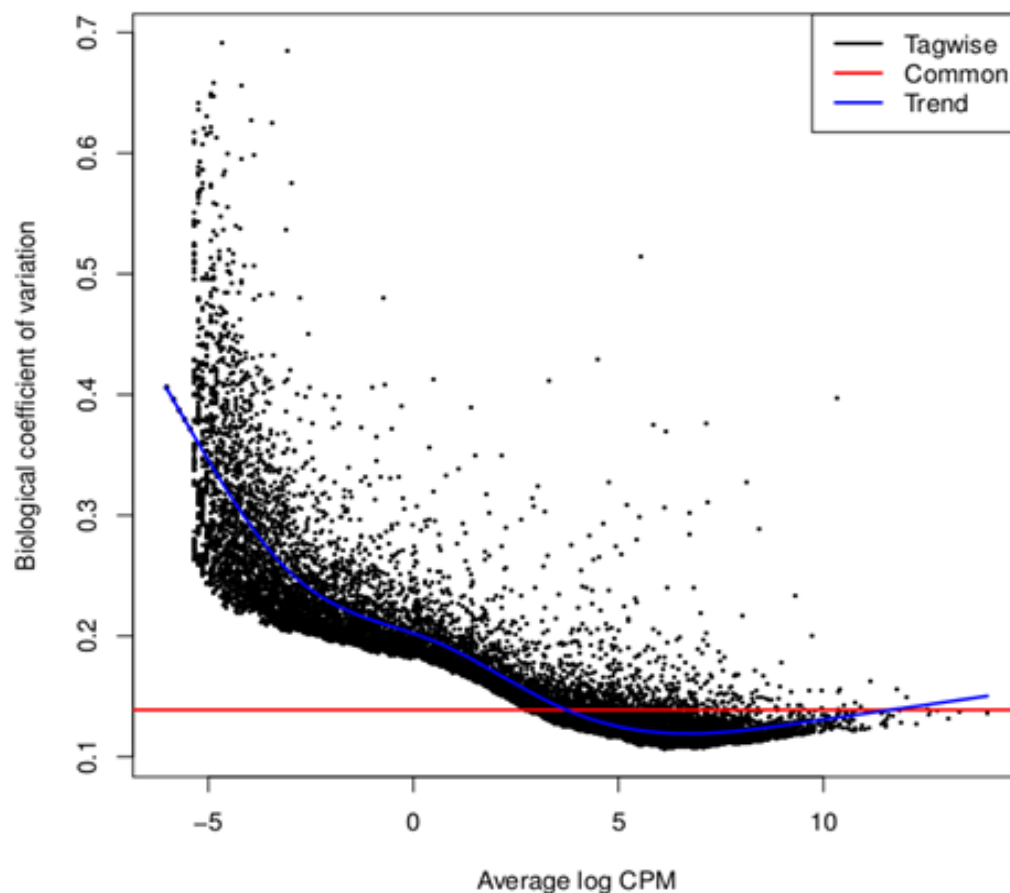
Correction par la variance

Loi binomiale négative

Edge R , DESeq



EdgeR



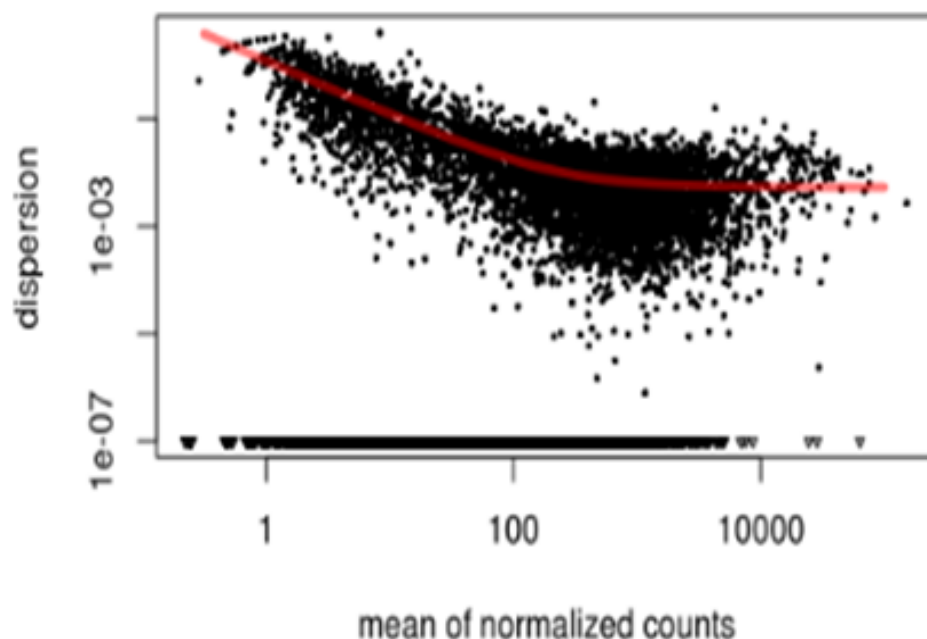
Estimation de la dispersion (entre réplicats biologiques)

- 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



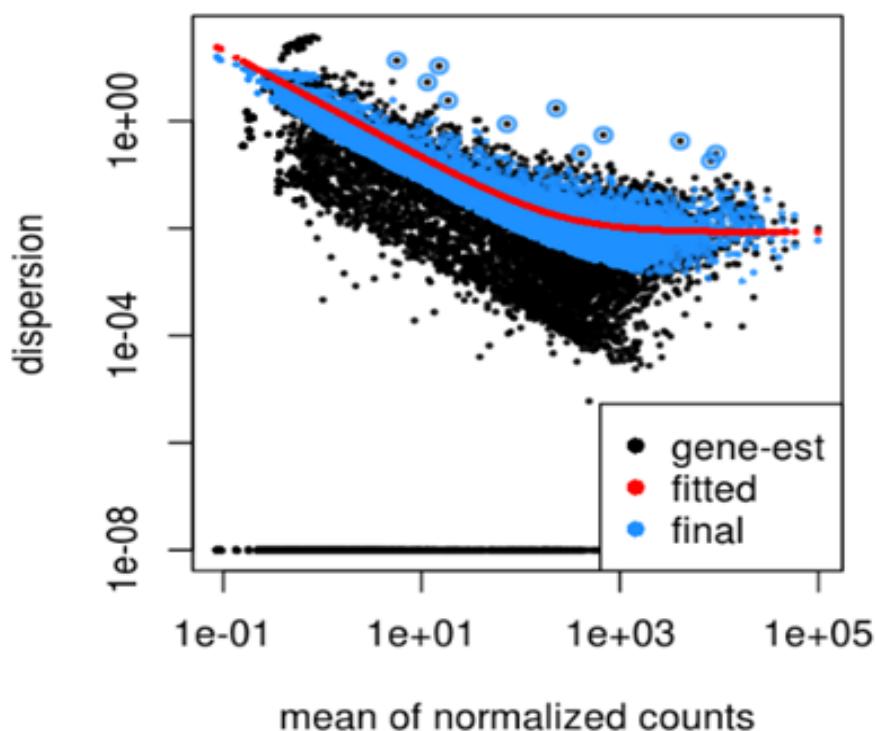
Estimation de la dispersion

- 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'outil 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 stringente et il détectera plus de gènes différentiellement exprimés.

Practice : DIANE

[TP Diane](#)

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)

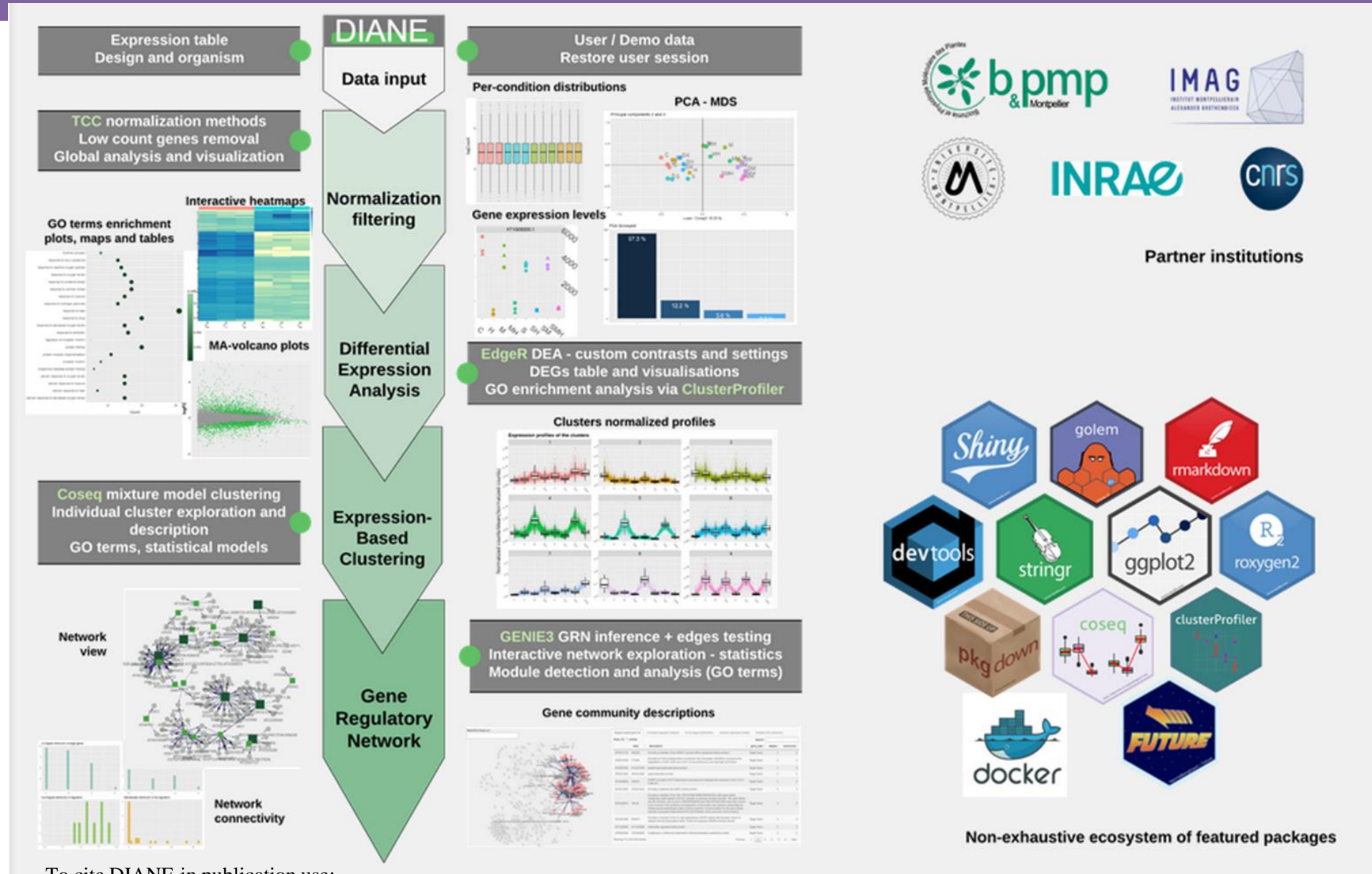
(https://www.yeastgenome.org/strain/S288C#genome_sequence)

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>

Practice : DIANE Import Data

DIANE

Context

Data import ← ad expression data and experimental design

Normalisation

Exploratory analysis

Differential Expression

Expression based clustering

Run a clustering

Explore clusters

Gene Regulatory Network

Network inference

Network analysis

Ready to upload datasets

Legal mentions

Software versions

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 :
☐ Comma ☐ Semicolon ☒ Tab

Choose CSV/TXT expression file ?
Browse... count_table.txt

Separator :
☒ Tab

Choose CSV/TXT gene information file (optional) ?
Browse... No file selected

Other ORGANISM DATABASE

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

CHANGE SEED SET SEED

Preview of the expression matrix

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

Design and gene information files

Separator :
☒ Comma ☐ Semicolon ☐ Tab

Choose CSV/TXT design file (optional)
Browse... No file selected

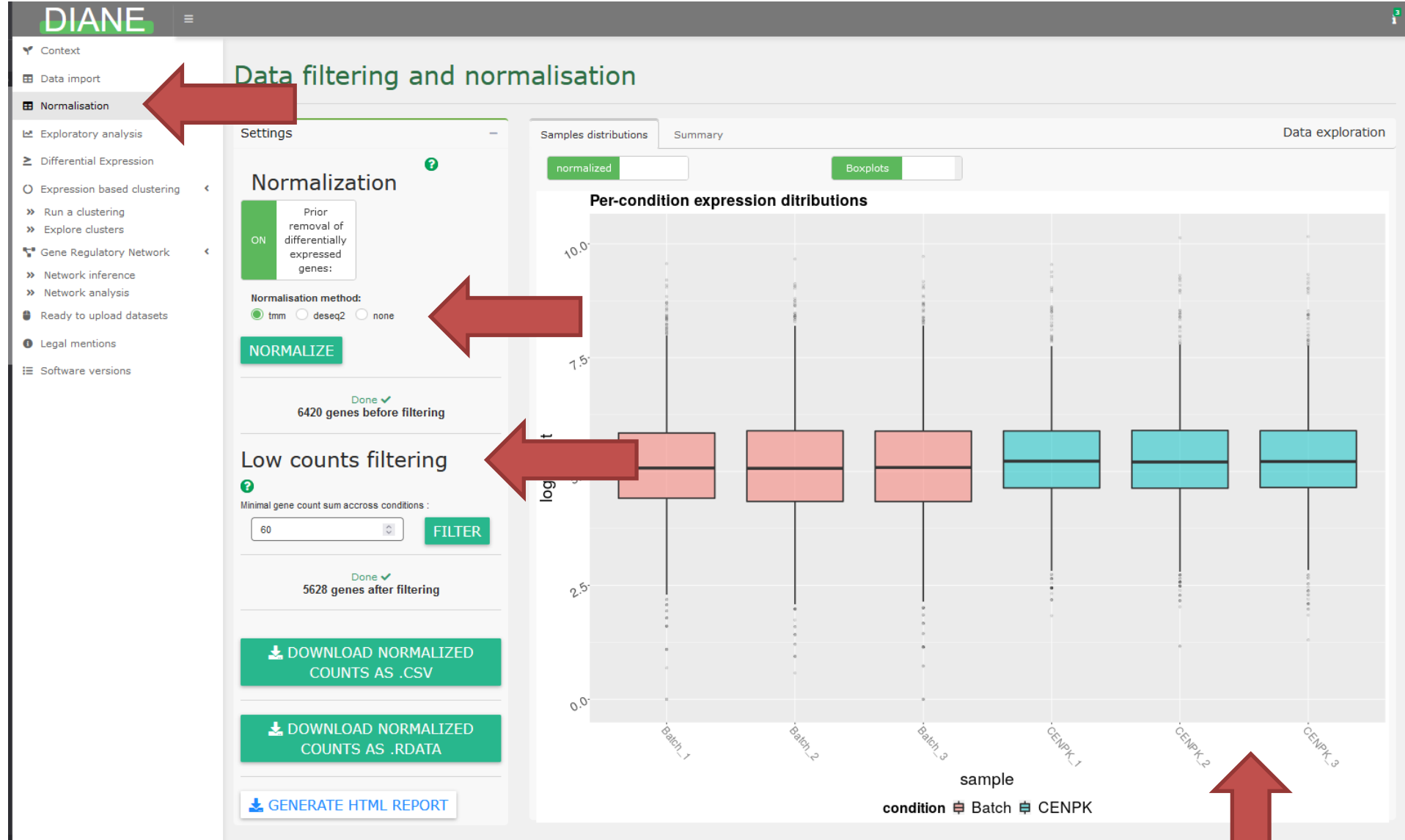
Describe the levels of each factors for your conditions

Reads_count.csv

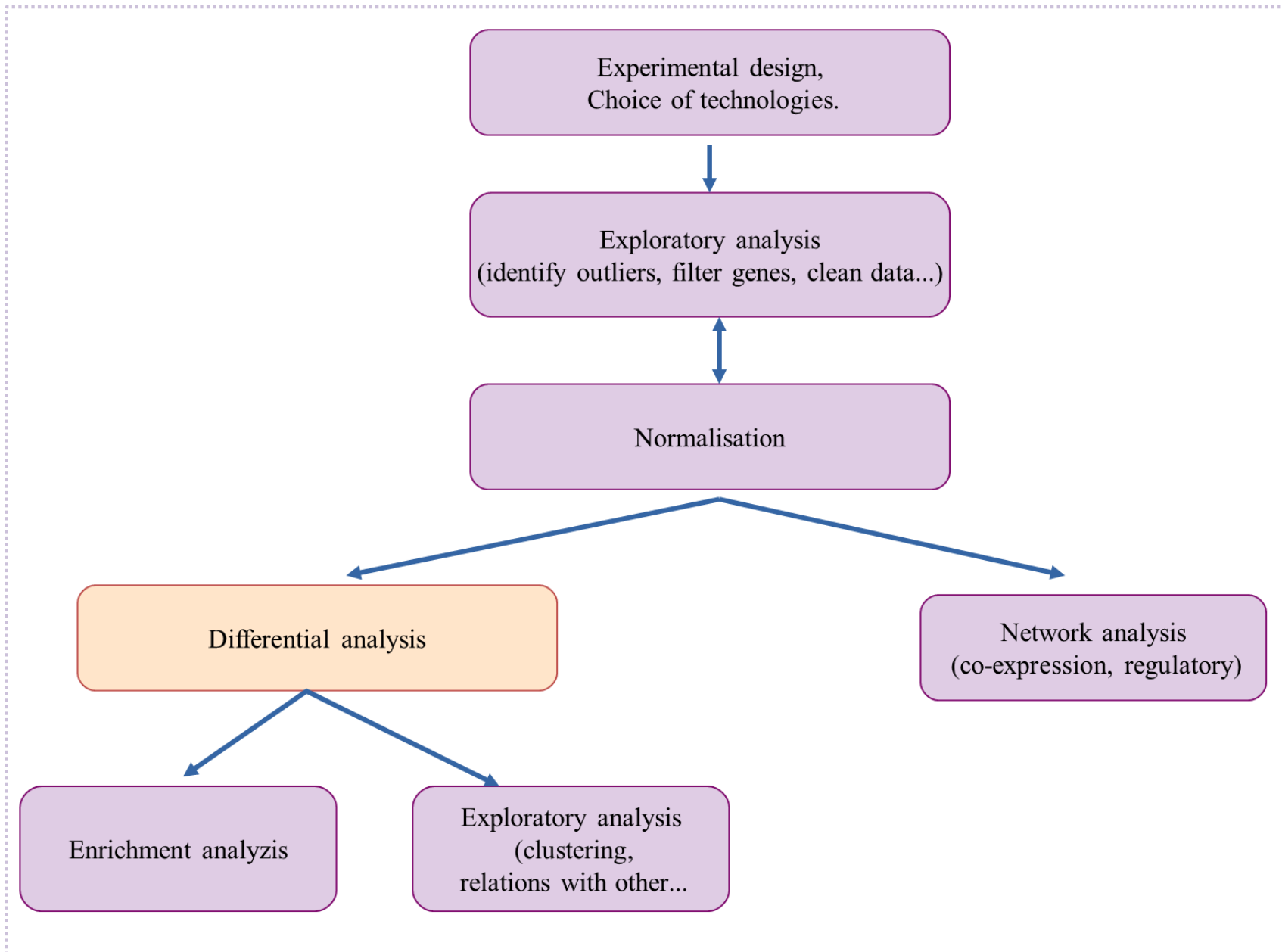
GeneInformation

Designed.csv

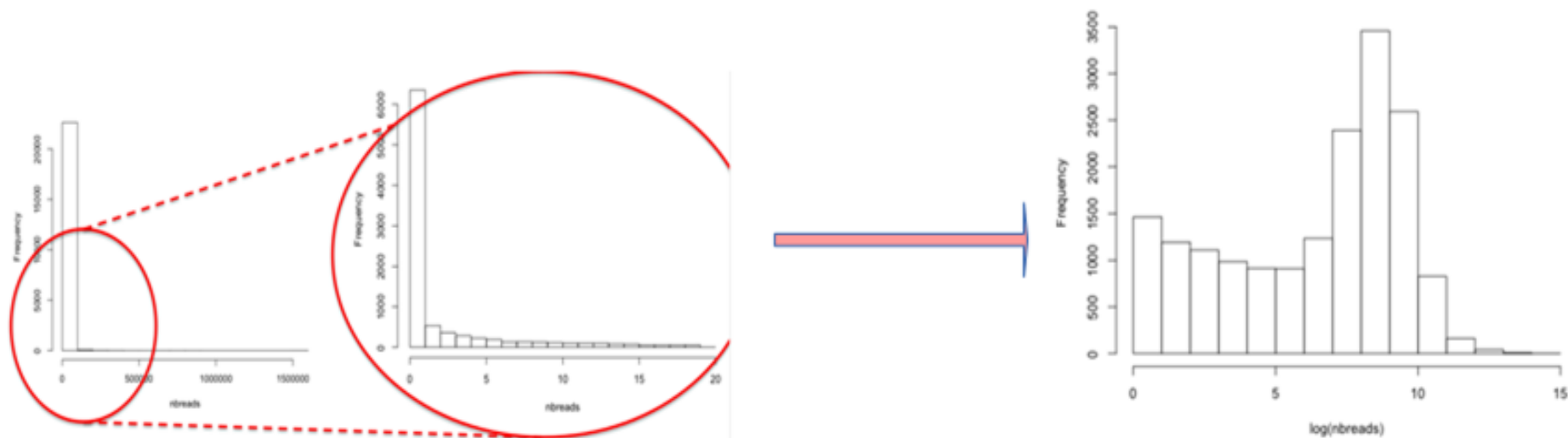
DIANE Normalisation Filtration



6- Recherche de gènes différentiellement exprimés



- **Note** : Utilisation du $\log(\text{nbr de reads})$ pour que les données suivent une loi statistique, et nécessité de transformer les « 0 ».
 - \Rightarrow Loi binomiale négative



- **Note** : Utilisation du $\log_2(\text{FoldChange})$
 - Fold Change = ratio entre 2 niveaux d'expression
 - Ratio de la valeur finale / valeur initiale

Rappel : définition du Fold-Change :

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

[illegible]

Rappel : définition du Fold-Change :

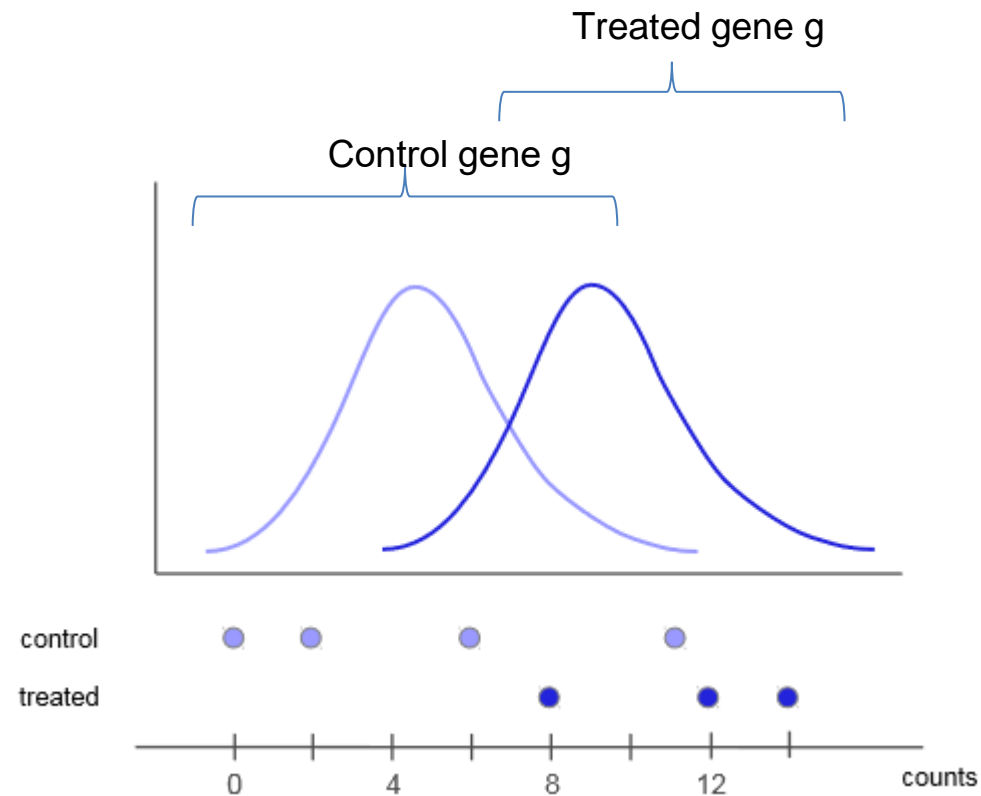
$$FC = \frac{\text{expression condition } V}{\text{expression condition } G}$$

Gene	V1	V2	V3	G1	G2	G3	FC	<i>p</i> -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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

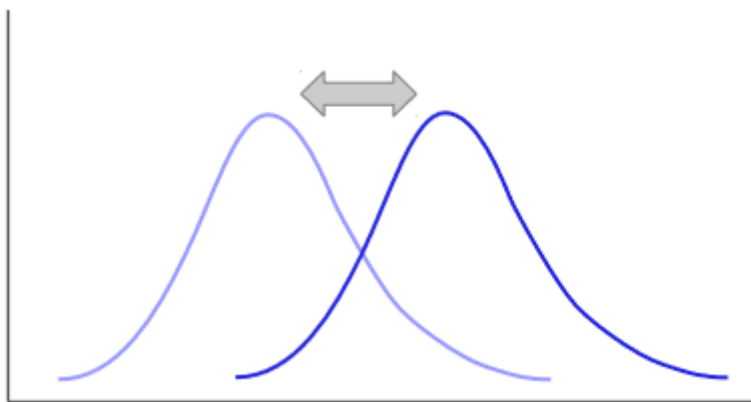
H_0 : le gène g a un niveau d'expression constant

H_1 : hypothèse alternative : le gène g est différentiellement exprimé

	control				treated		
Gene 1	5	1	0	0	4	0	0
Gene 2	0	2	1	2	1	0	0
Gene 3	92	161	76	70	140	88	70
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
Gene g	0	11	2	6	12	8	14
:	:	:	:	:	:	:	:
:	:	:	:	:	:	:	:
Gene G	15	25	9	5	20	14	17



Les tests statistiques permettent de quantifier cette différence



H₀ : le gène g a un niveau d'expression constant

H₁ : hypothèse alternative : le gène g est différentiellement exprimé

Détermine si l'hypothèse H₀ est vérifiée

- **Les tests statistiques ne donnent pas une réponse simple**

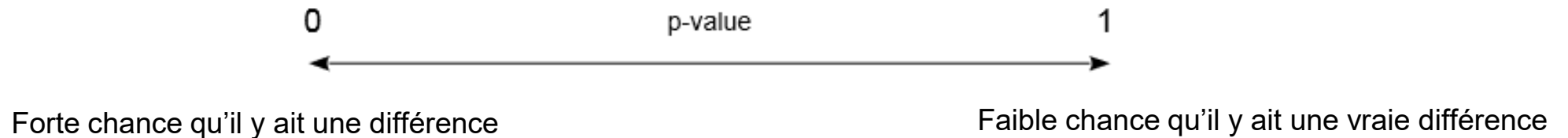
- Un test statistique détermine la probabilité que H₀ est vraie
- Ici H₀ : le gène g n'est pas exprimé de manière différentielle entre les conditions
- **On applique le calcul de la statistique qui est convertie en p-value : quantifier la différence**
- **Si p est faible alors H₀ est rejetée en faveur de l'Hypothèse alternative H₁**

H₀ : le gène g a un niveau d'expression constant Faux

DONC le gène g est un gène Différentiellement exprimé

P-value

- La valeur p est la probabilité de voir un résultat aussi extrême ou plus extrême que les données observées, lorsque l'hypothèse nulle est vraie.
- Attention : Il ne s'agit pas de la probabilité que l'hypothèse nulle soit vraie.



- La convention en biologie : seuil de la p -value 0,05 (5%)

- **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):**

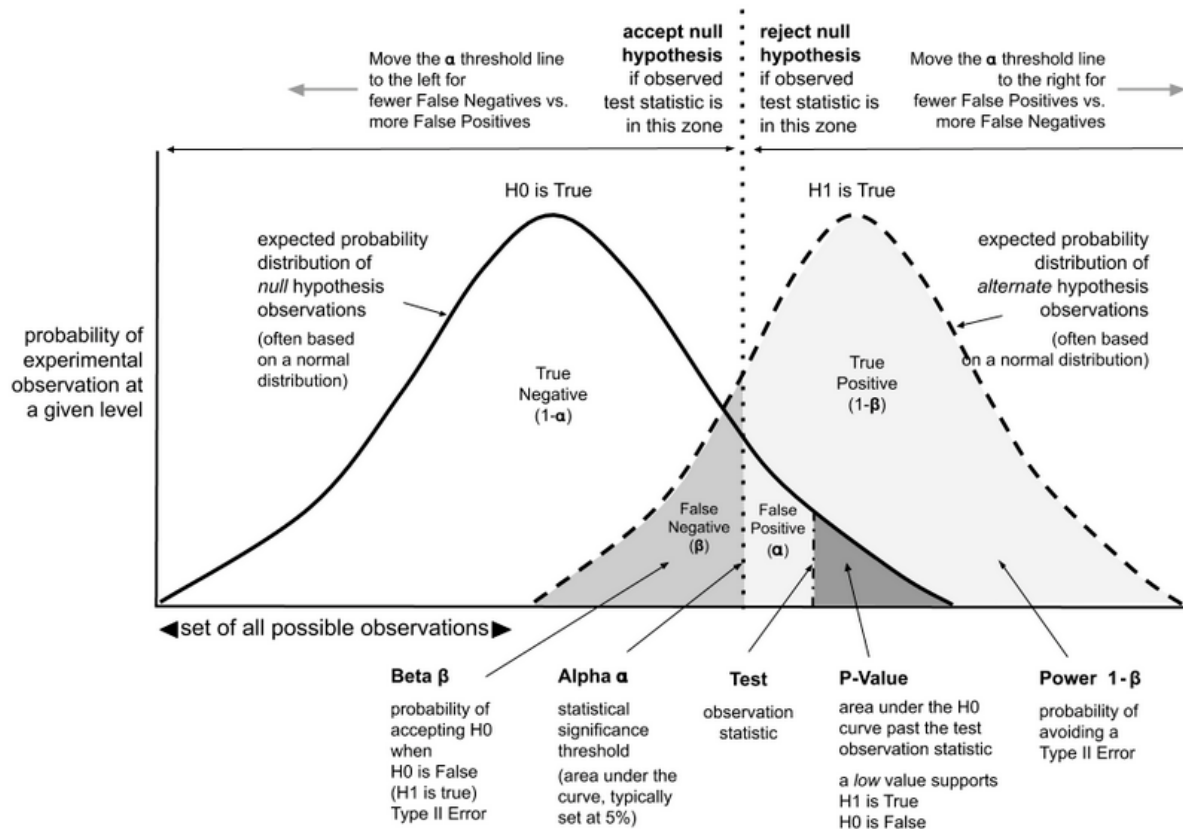
- Probabilité β de rejeter H_0 alors qu'elle est fausse
- Probabilité de décider qu'un gène n'est pas différentiellement exprimé alors qu'il l'est.
- **Faux négatif**

- **Conséquence :**

- En testant les 20000 gènes avec $\alpha = 5\%$
- **On s'attend à obtenir 20000 x 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	$1-\beta$

Test p value Error



		Actual situation "truth"	
		H_0 true	H_0 false
Decision	Do not reject H_0	Correct decision	Incorrect decision type II error
	Reject H_0	Incorrect decision type I error	Correct decision

Table 1: α = P(type I error), β = P(type II error) and power = $1 - \beta$.

- **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 $1 - 0,99 = 0,01$

=> $\alpha > p\text{-value}$ Ho 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 $P_k \leq \frac{k}{m}\alpha$ et déclarer les gènes 1.....k différentiellement 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 différentiellement exprimé pour $\alpha = 0,05$

Practice : DIANE

[TP Diane](#)

Practice Analyse différentielle DIANE

DIANE

- Context
- Data import
- Normalisation
- Exploratory analysis
- Differential Expression**
- Expression based clustering
 - Run a clustering
 - Explore clusters
- Gene Regulatory Network
 - Network inference
 - Network analysis
- Ready to upload datasets
- Legal mentions
- Software versions

Differential expression analysis

Settings

Conditions to compare for differential analysis :

Reference: ☒ Batch, ☐ CENPK

Perturbation: ☐ Batch, ☒ CENPK

Adjusted pvalue (FDR) :

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

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

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

Show entries

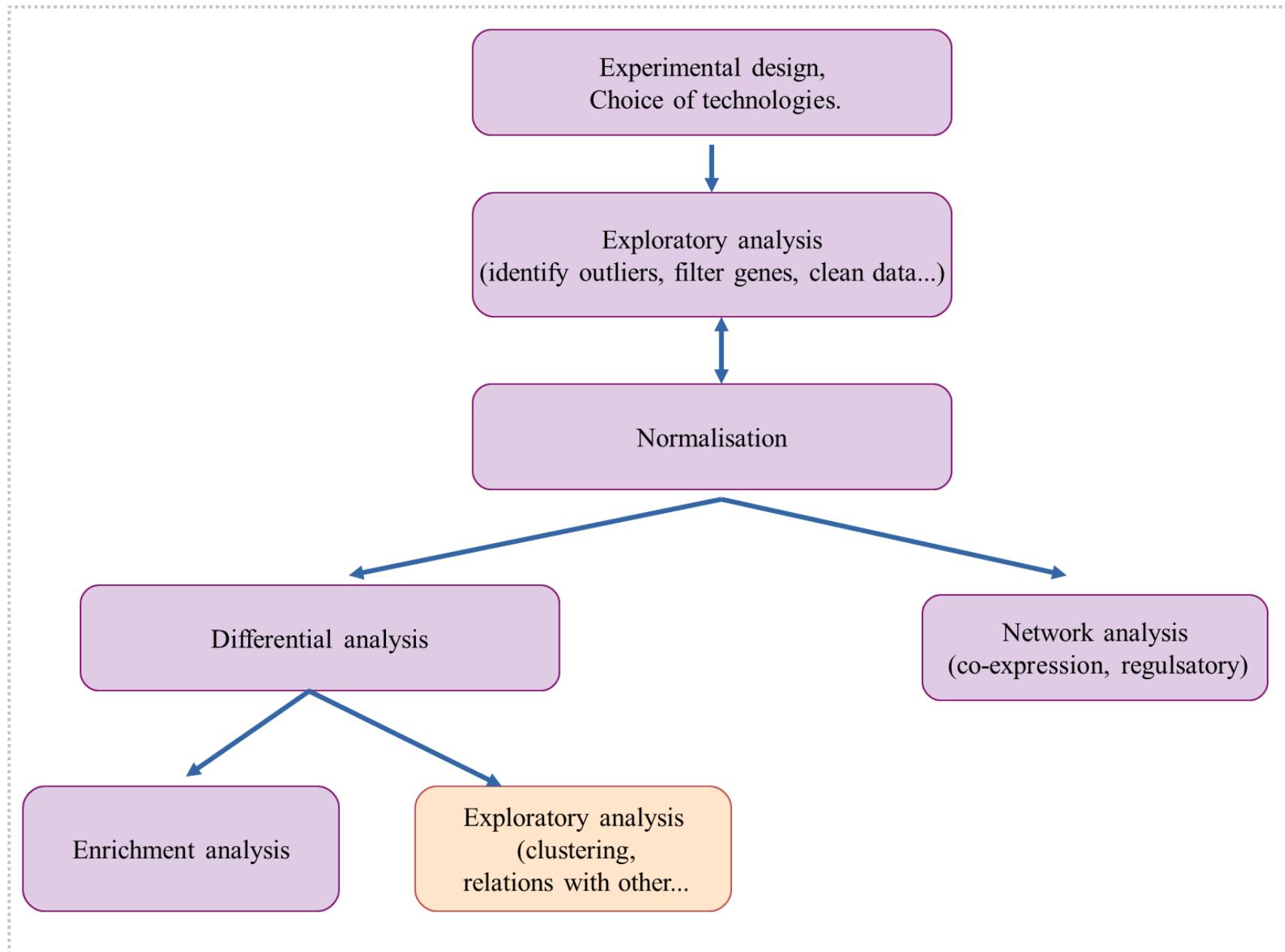
Search:

	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
YMR303C	8.407	7.856	0.000	Up
YCR010C	7.902	10.69	0.000	Up
YGL205W	7.094	8.400	0.000	Up
YHR096C	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 2 3 4 5 ... 141 Next

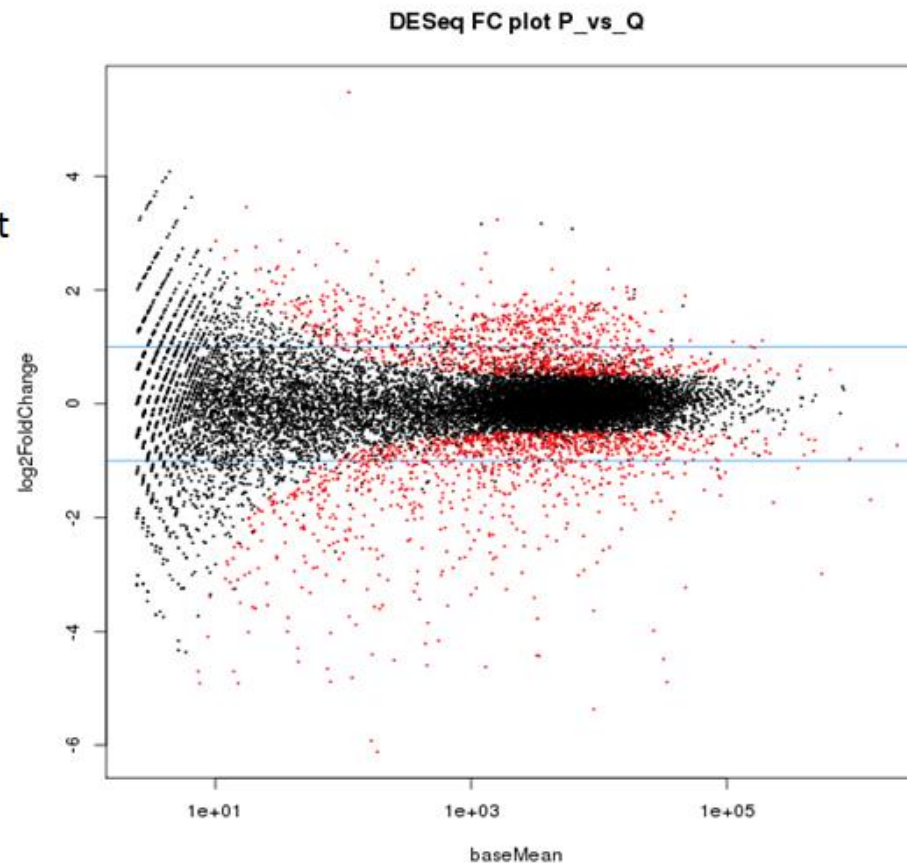
6- Plots and Graphical Représentations



Smear plot / MA plot
Pvalue adj < 0.05

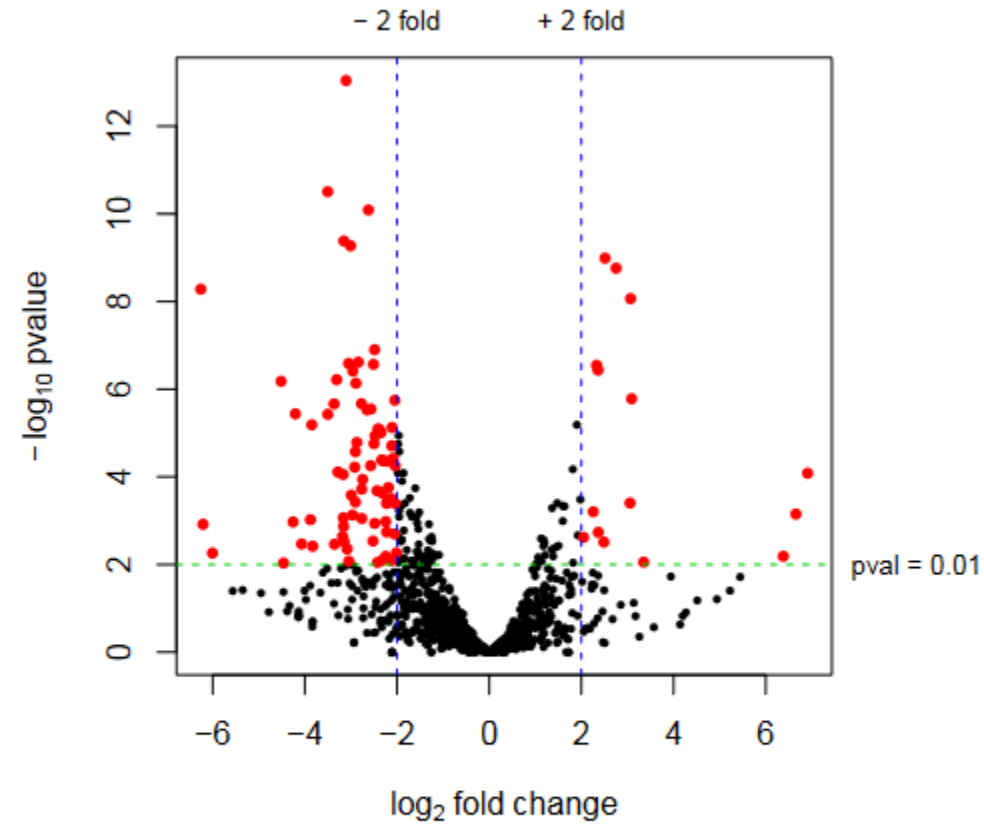
MA plot

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

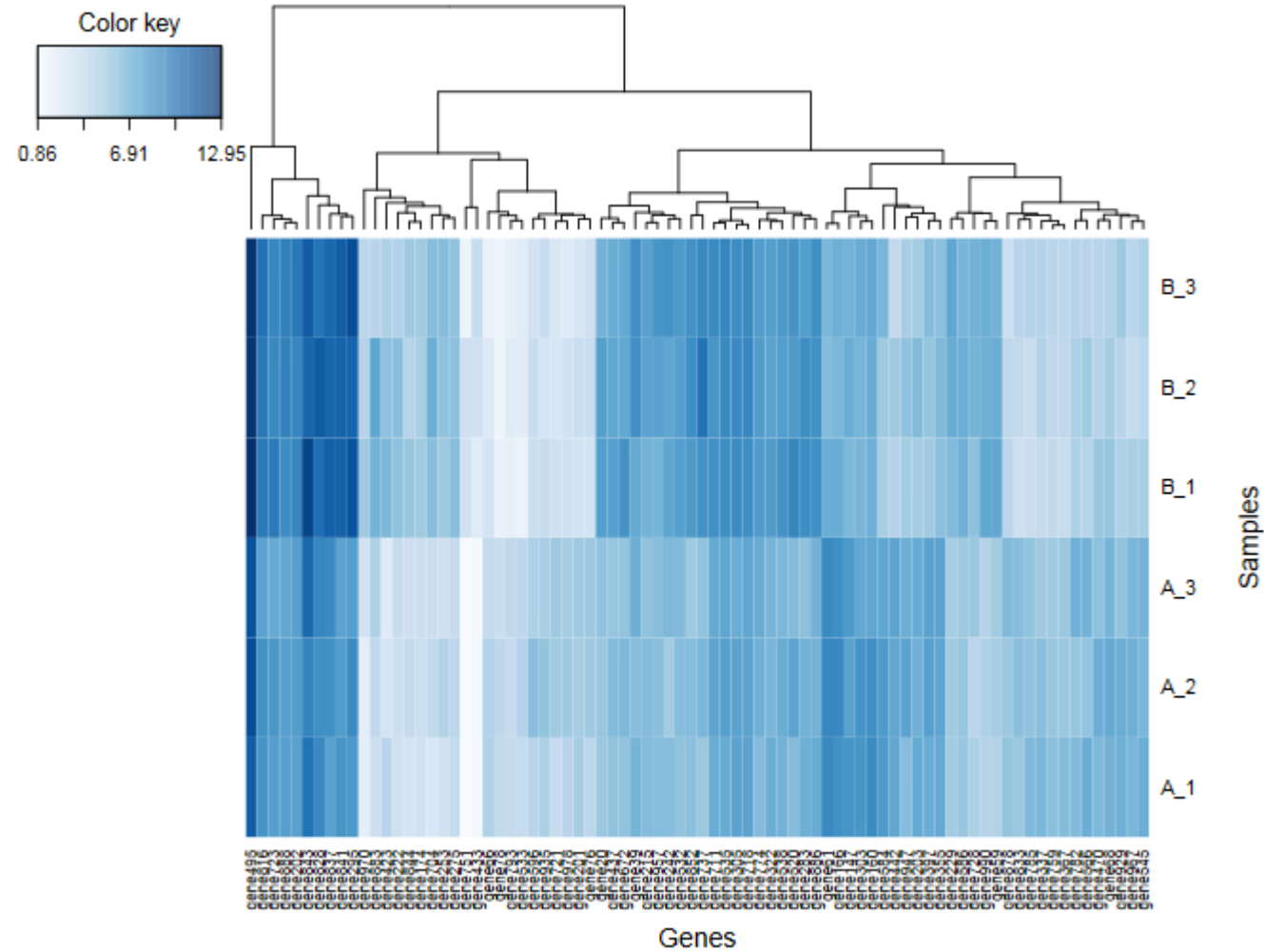
Volcano plot
Pvalue adj < 0.01



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

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

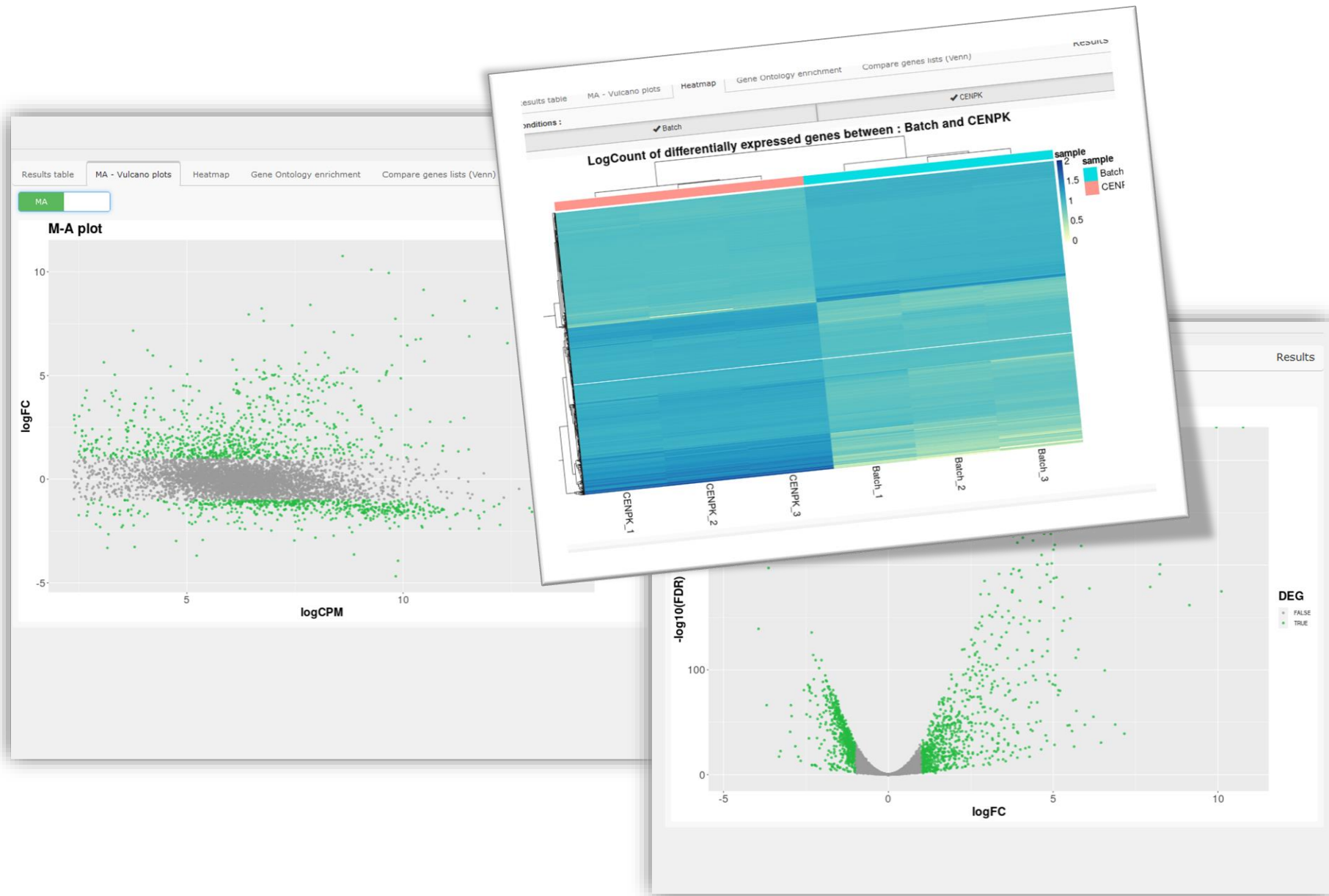
Hierarchical clustering / Heatmap



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



Sebastien Ravel



Christine Tranchant-Dubreuil



Sebastien 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](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**

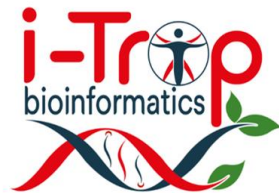
<https://itrop-survey.ird.fr/index.php/515725?lang=fr>



SUIVEZ NOUS SUR TWITTER !



South Green : [@green_bioinfo](https://twitter.com/green_bioinfo)



I-Trop : [@ltropBioinfo](https://twitter.com/ltropBioinfo)



N'oubliez pas de nous citer !

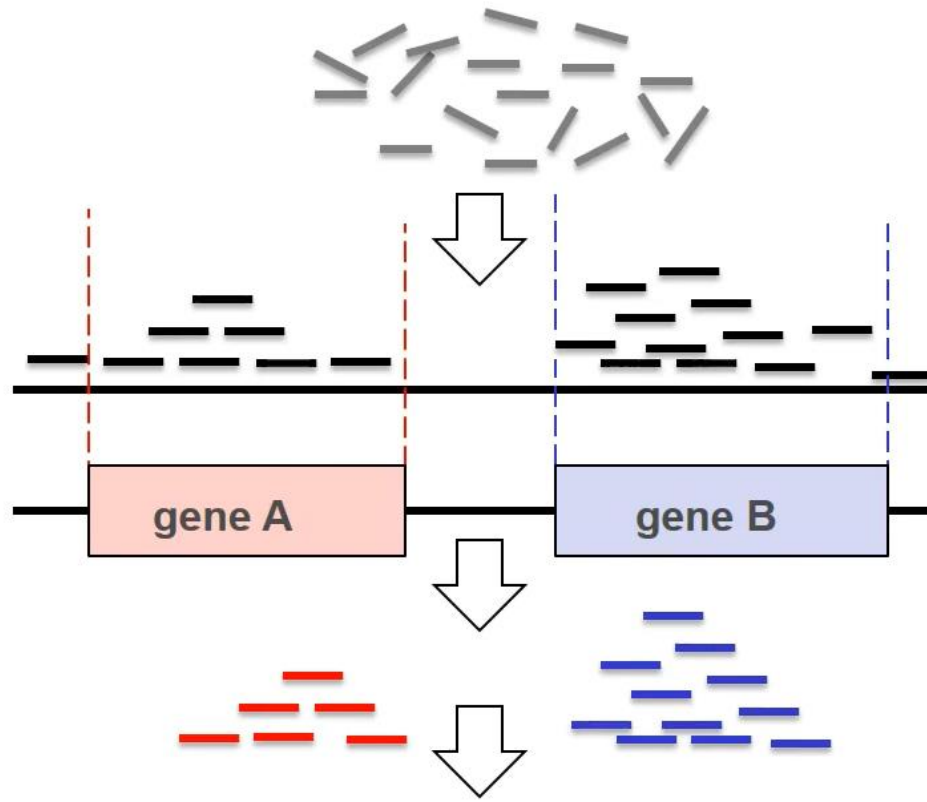
Comment citer les clusters?

"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 NBPSeg**
 - 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, NBPSeg : 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 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



	Control 1	Control 2	Control 3	Sample 1	Sample 2	Sample 3
Gene A	6	5	7	170	100	110
Gene B	11	11	10	3	4	2
Gene C	200	150	355	50	1	3
Gene D	0	1	0	2	0	1

STEP	TOOL	FILE
Quality control	FastQC	FASTQ
Pre-processing	Trimmo-matic	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:

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

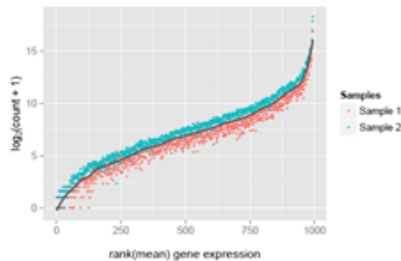
Normalization Technique	Name authors	Description	Software
UQ	Upper Quartile Ref : Bullard et al., 2010 (Upper) Quartile normalization	Les comptages par gène sont divisés par le 3e quartile des comptages non nuls de l'échantillon, puis multipliés par la moyenne des 3e quartiles de tous les échantillons.	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 multiplié 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 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.	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 ratios de ses comptages par rapport à sa moyenne géométrique entre les différents échantillons. L'idée sous-jacente est que les gènes non différentiellement 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 différentiellement exprimés, la médiane des ratios constitue une estimation du facteur correctif qui doit être appliqué 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 implémentée dans le package Bioconductor edgeR. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différentiellement 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 différentiellement 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

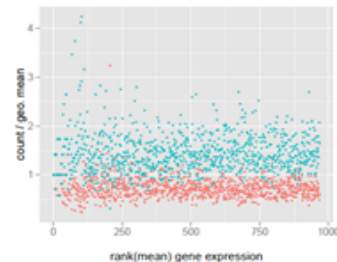
La normalisation RLE (Relative Log Expression) a été développée dans le package Bioconductor DESeq. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différentiellement exprimés. Le facteur de normalisation pour un échantillon est obtenu en calculant pour chaque gène la médiane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

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

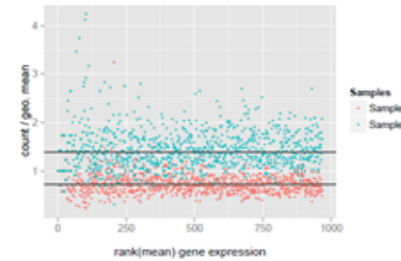
1 – Calcule une pseudo référence



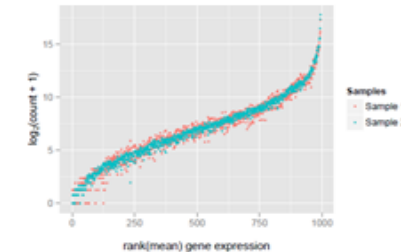
2 – Centre les échantillons comparés à la référence



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



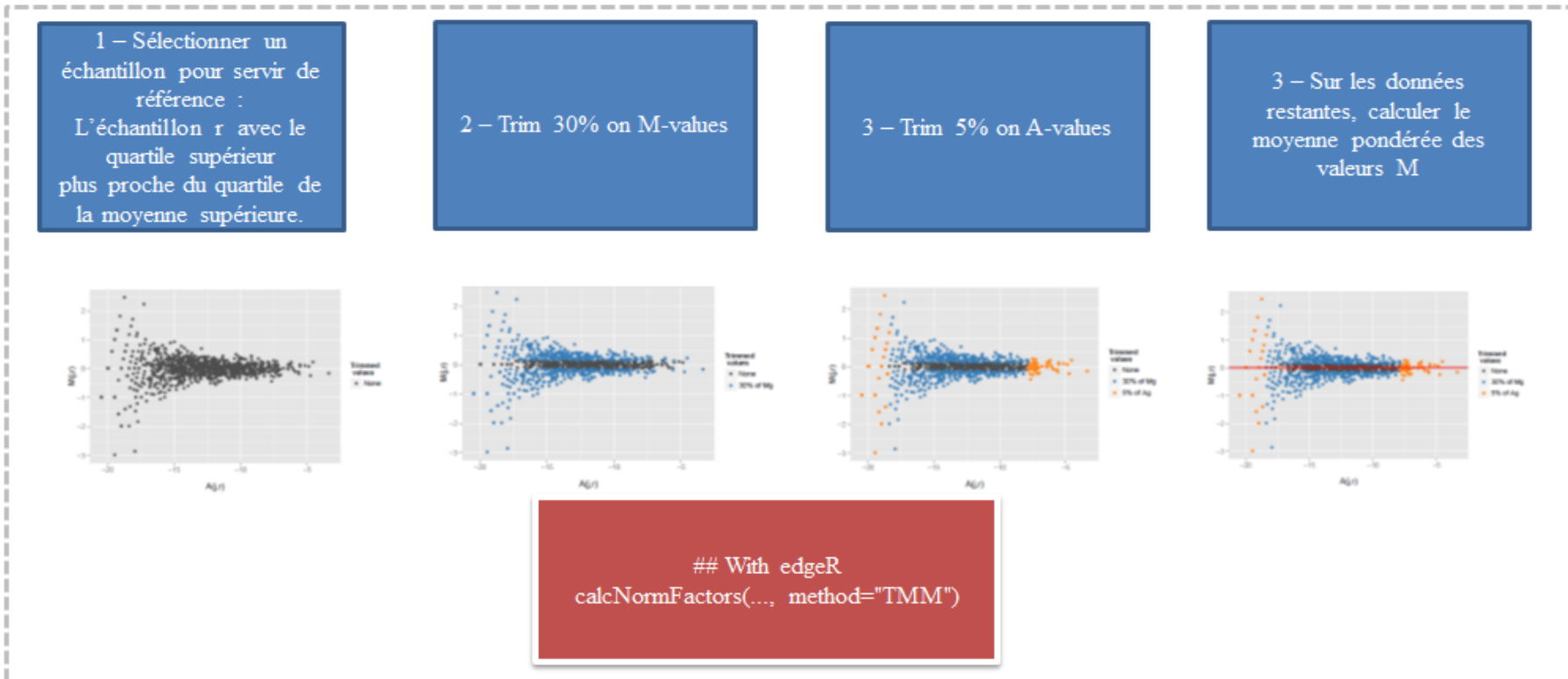
```
## With edgeR
calcNormFactors(..., method="RLE")
## with DESeq
estimateSizeFactors(...)
```



TMM

La normalisation TMM (Trimmed Mean of M-values) est implémentée dans le package Bioconductor edgeR. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différentiellement 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 différentiellement 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 bibliothèques (et pas aux comptages bruts) afin de rendre l'hypothèse vraie.

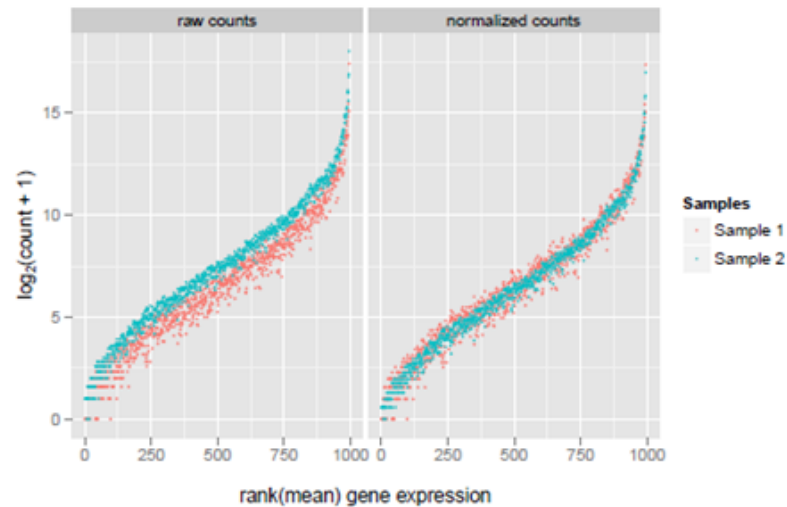
Ref : Robinson, M. and Oshlack, A. (2010). Dans edgeR.



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

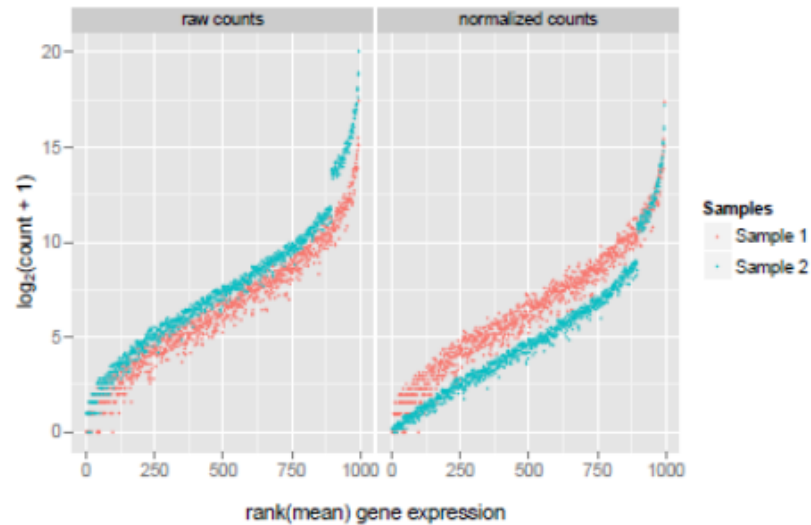


```
## With edgeR
cpm(...,
normalized.lib.sizes=TRUE)
```

Upper Quartile

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

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



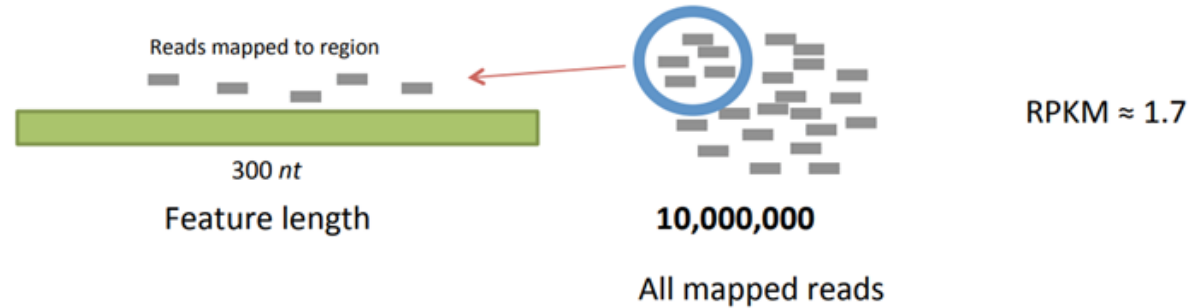
1 – dans lequel $Q(p)_j$ est un quantile donné (généralement le 3e quartile) de la distribution des comptes dans l'échantillon j .

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

Correcting for **transcript length** and **total number of reads**

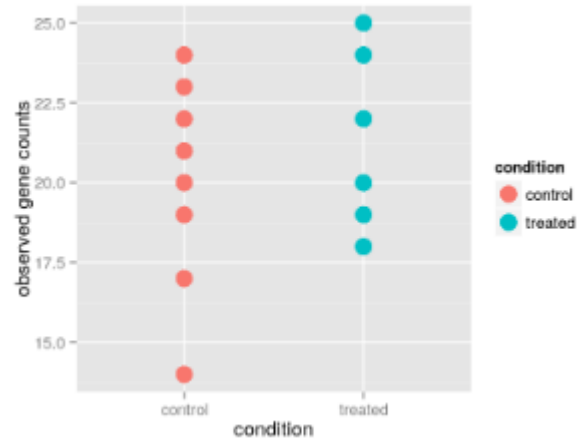
RPKM

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.



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

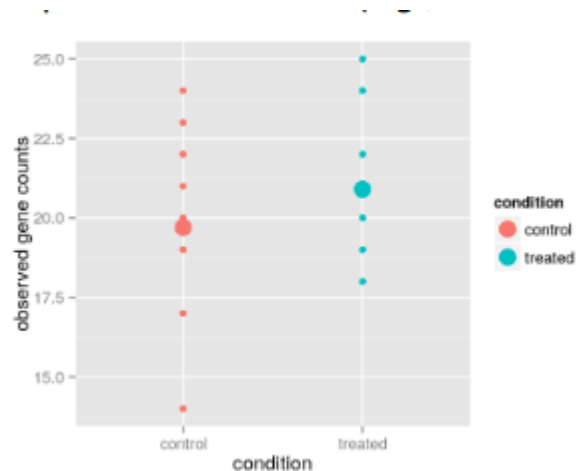
RPKM: Reads Per Kilo base of transcript per Million reads



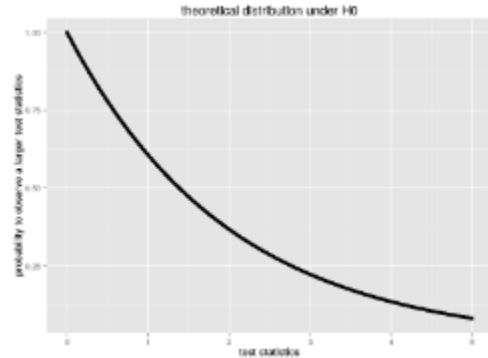
1 : formuler l' Hypothèse

H_0 : le nombre de reads counts moyen de gènes g dans les échantillons de contrôle est le même que le nombre de reads counts moyen dans les échantillons traités, qui est testée contre une alternative

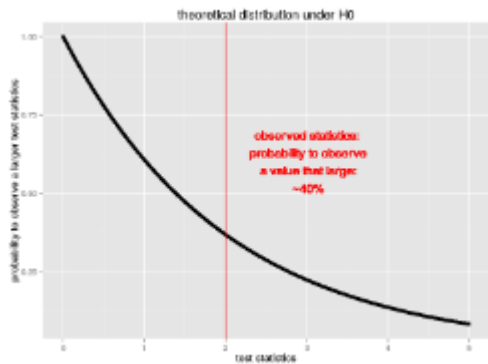
H_1 : le nombre de reads counts moyen de gènes g dans les échantillons de contrôle est différent le nombre de reads counts moyen dans les échantillons traités.



2 : à partir des observations (reads counts), appliquer un test statistique,

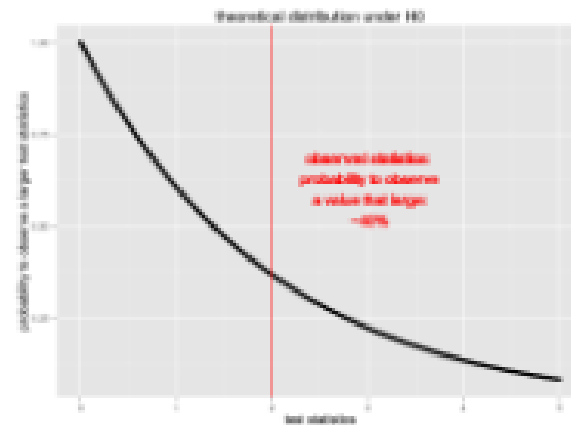
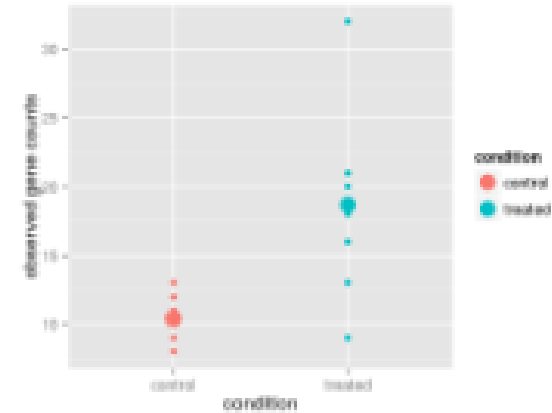
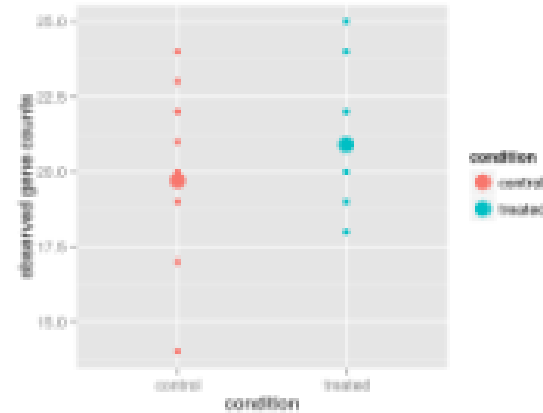


3 Trouver la distribution théorique du test statistique sous H_0

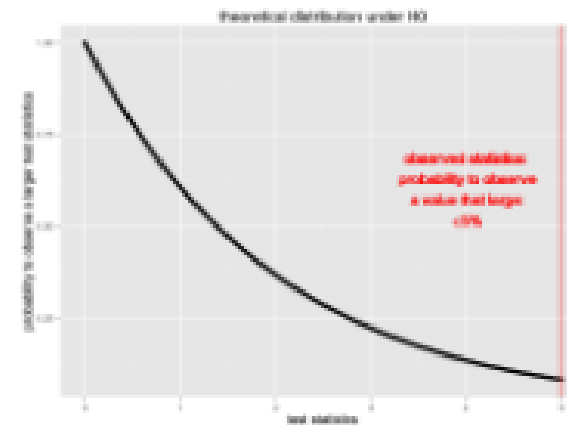


4 déduire la probabilité que les observations se produisent sous H_0 : c'est ce qu'on appelle la valeur p

5 conclure : si la valeur p est faible (généralement inférieure à $\alpha = 5\%$ par convention),
 H_0 est peu probable : " H_0 est rejetée". improbable :
 Nous avons donc : $\alpha = P(H_0 \text{ est rejetée})$



Do not reject H_0



Reject H_0