

# Analyse de données RNASeq

## Etude d'expression différentielle de gènes

## Quelques définitions

- Séquençage: déterminer la succession linéaire des bases A,T,C,G de l'ADN, la lecture de cette séquence permet d'étudier l'information biologique contenue par celle-ci.
- Séquençage Nouvelle Génération (NGS): Séquençage à très haut débit, génération d'un très grand nombre de séquences simultanément
- RNA-seq: transcriptome sequencing. Informations sur les ARNs via le séquençage de l'ADN complémentaire (cDNA)
- Re-séquençage: séquençage d'un génome qui pourra être comparé à une séquence de référence connue (le génome de l'espèce a déjà été séquencé)
- Séquençage *de-novo*: séquençage d'un génome pour lequel il n'existe pas de génome de référence, détermination d'une séquence inconnue

## Pourquoi faire du RNA-seq?

L'accès aux séquences des ARN permet de:

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

...

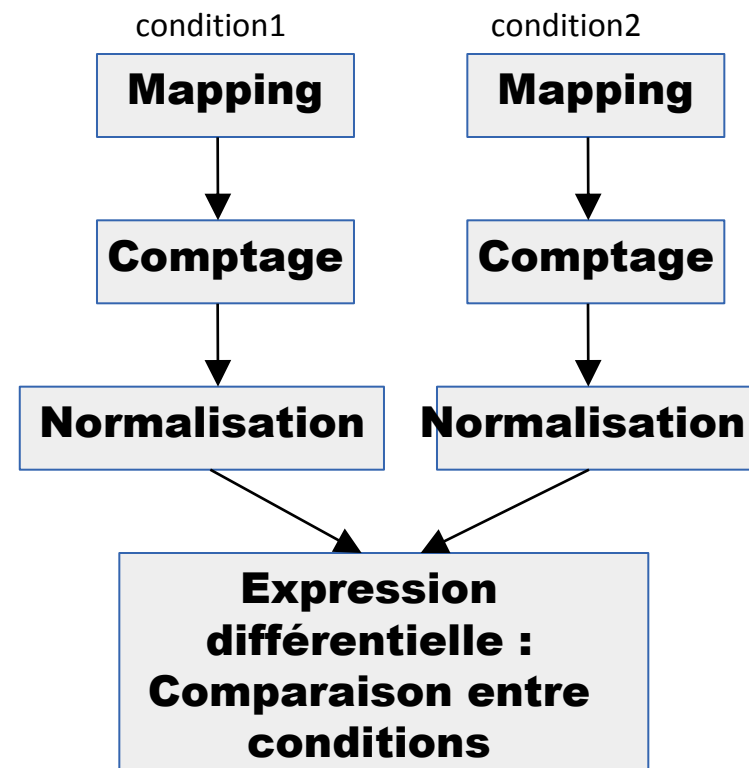
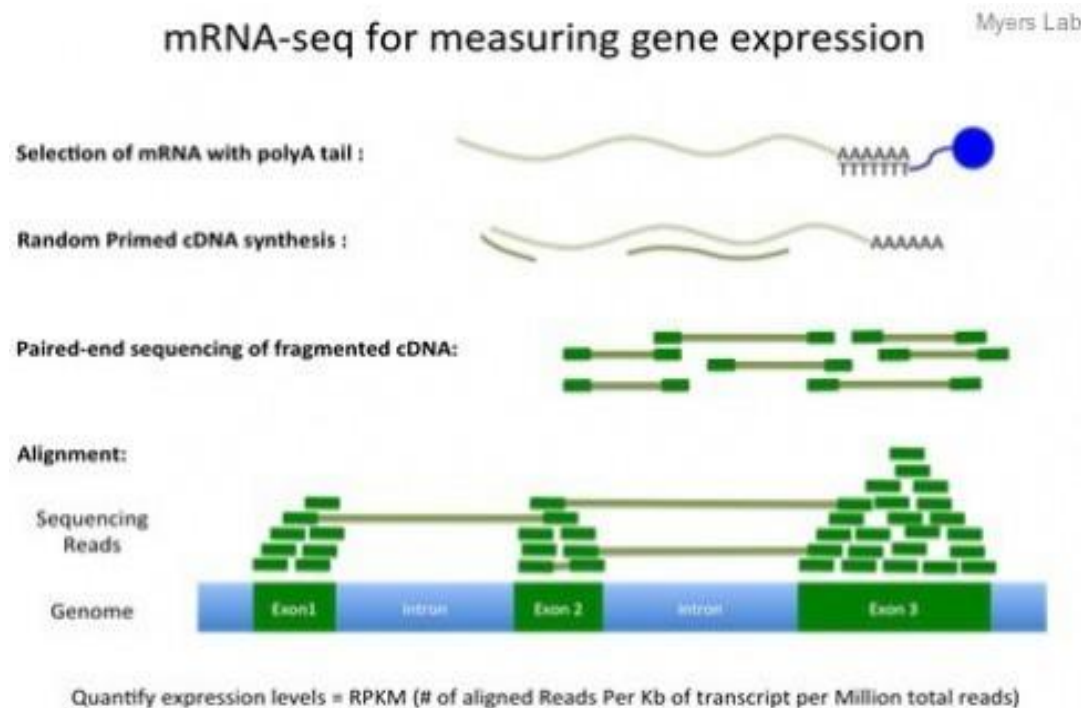
## Avantages RNA-seq / microarray

- Plus précis et plus sensible: permet de faire plus de découvertes
- RNA-seq permet la détection d'épissage alternatif
- Possibilité d'étudier des transcrits faiblement exprimés
- Pas besoin de génome de référence  
(pour le microarray, c'est nécessaire pour dessiner les sondes)

# But du TP

- 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
- Comparer les résultats obtenus entre deux approches/outils différents. Comprendre les différences

# Principe général basé sur le comptage des reads



# 1) Mapping

# Choix de l'outil de mapping

## 1) Si on dispose d'un génome de référence

**Utilisation de « splice junction mapper »  
(ex : TopHat2, CRAC, MapSplice)**

### 1.1) Si on a une annotation

**=> Optimise l'alignement en considérant l'annotation GFF**

**=> Permet d'aller rechercher de nouveaux gènes**

### 1.2) Si on n'a pas d'annotation

**=> Aide à l'annotation structurale (mise en évidence de gènes)**

## 3) Si on dispose d'un transcriptome de référence

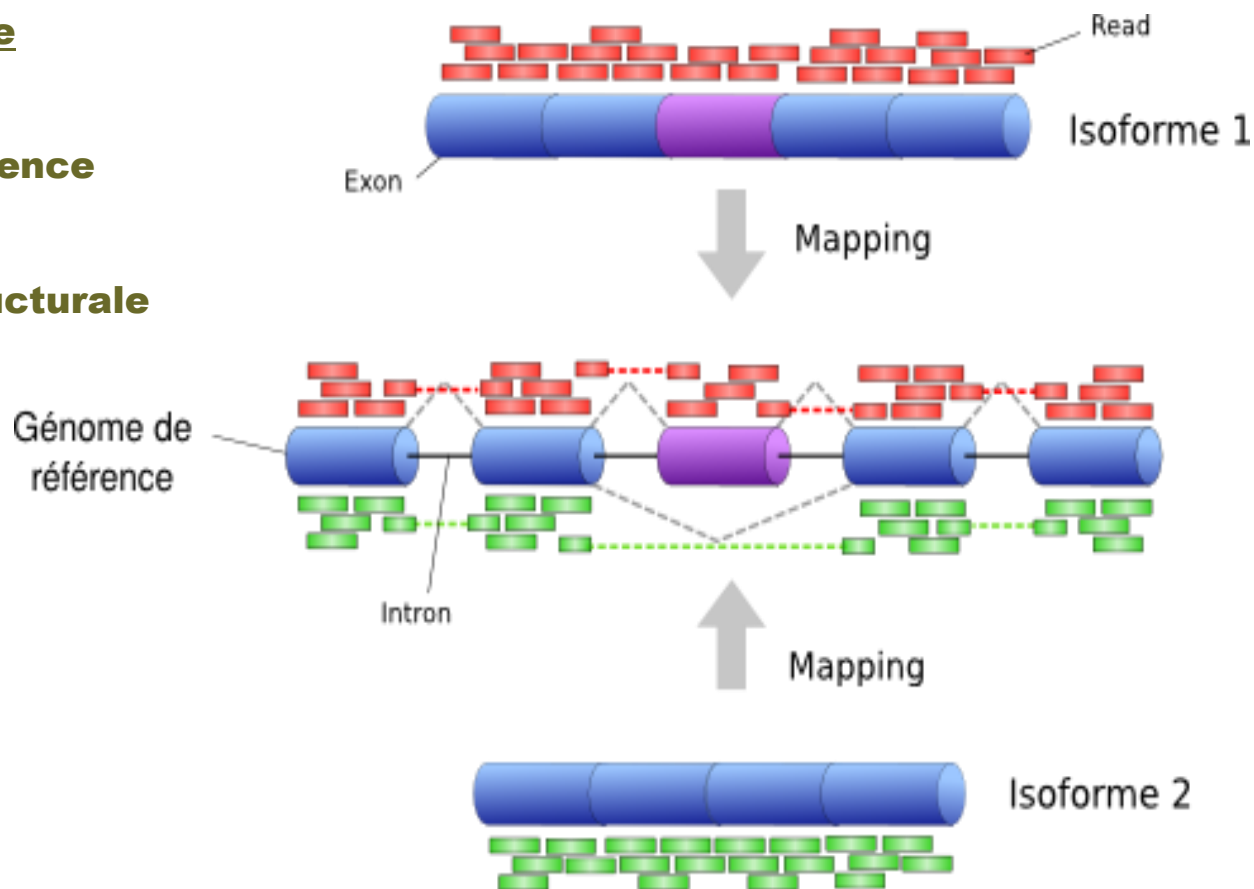
**Utilisation de mapper classique (ex : BWA, bowtie)**



## Mapping sur une référence génomique

=> permet la mise en évidence  
d'isoformes

=> aide à l'annotation structurale  
du génome

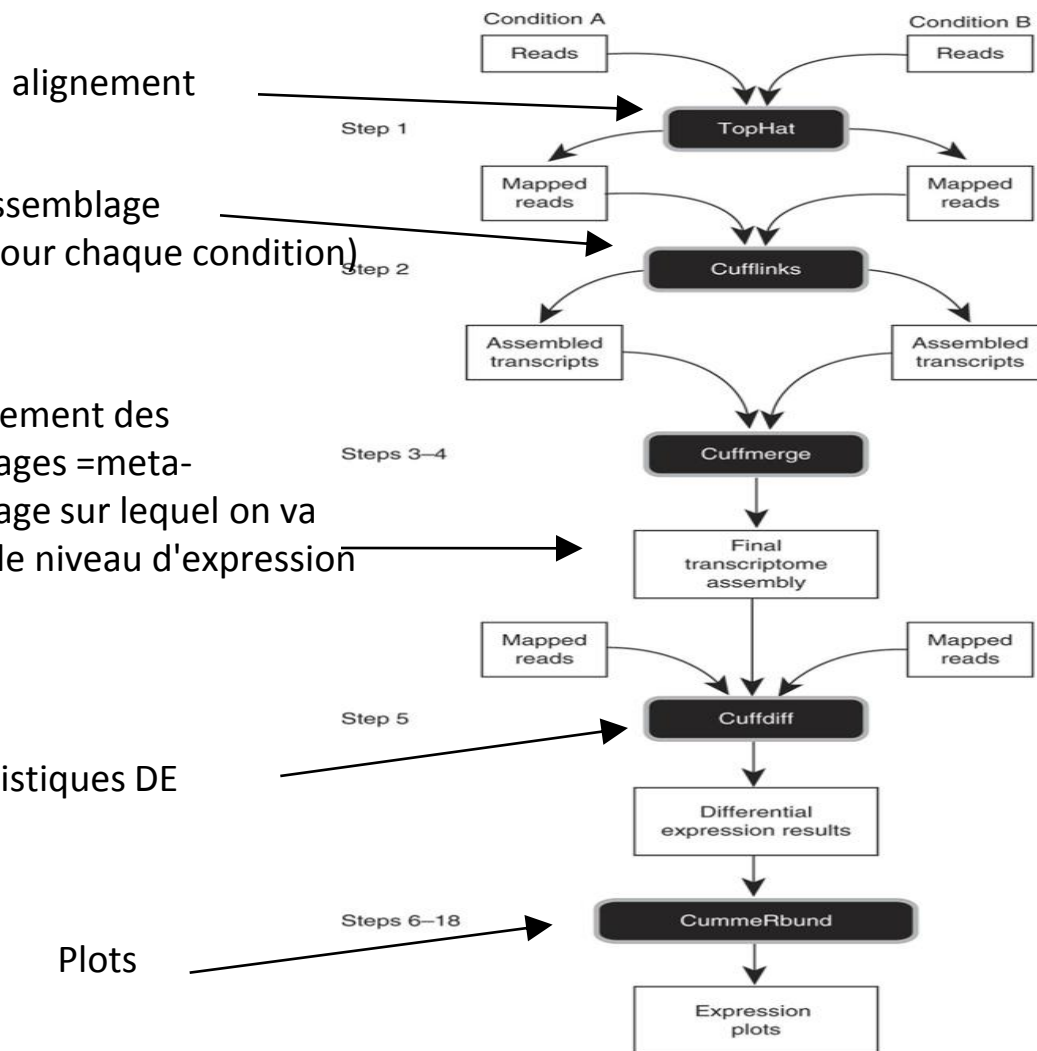


## Suite TopHat / Cufflinks

Regroupement des  
assemblages = meta-  
assemblage sur lequel on va  
calculer le niveau d'expression

Statistiques DE

Plots



## 2) Comptage

## 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

### 3) Normalisation des données

Objectif : permettre de comparer les valeurs obtenues pour différents échantillons

Erreur à ne pas commettre : croire que les données de RNA-seq sont plus stables que celles de puces à ADN et que la normalisation n'est pas nécessaire

« One particularly powerful advantage of RNA-seq is that it can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets » (Wang et al., Nat. Rev. Genet., 2009)

En réalité, les biais existent bien mais sont différents

=> Nécessité de réaliser des méthodes de normalisation spécifiques

Principaux biais actuellement identifiés :

- taille de la banque (= profondeur de séquençage)
- longueur des gènes
- composition en GC des gènes

Effet de la taille de la banque :

Illustration très simple :

On considère deux échantillons ayant la même composition en ARN

On réalise une banque pour chaque échantillon

On a obtenu 2 781 315 reads pour la banque A et 2 254 901 reads pour la banque B

=> on aura « artificiellement » 1.2334 fois plus de chaque ARN dans la banque A alors que les quantités « réelles » sont identiques

## Effet de la longueur des gènes :

Pour un même niveau d'expression, un long transcrit aura plus de chances d'être séquencé et donc plus de reads qu'un transcrit plus court

=> plus facilement mis en évidence comme DE

=> nécessaire de corriger ce biais



## Méthodes de normalisation :

### 1) Méthodes de normalisation inter-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 total moyen de reads à travers les banques

- Upper Quartile (UQ) : dans la méthode TC on remplace le nombre total de reads par le 3ème quartile des comptes différents de 0

=> normalisation moins sensible aux valeurs extrêmes

= normalisation plus robuste, notamment dans le cas où un ou plusieurs gènes très abondants sont différentiellement exprimés

- RLE (Relative log expression)

- TMM (Trimmed Means of M-Values)

<http://biorxiv.org/content/biorxiv/early/2015/09/03/026062.full.pdf>

## Méthodes de normalisation :

### 2) Reads Per Kilobase per Million (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 de normalisation inter et intra-banque

=> permet de comparer des gènes entre eux mais pas forcément nécessaire pour comparer 2 conditions sur un même gène

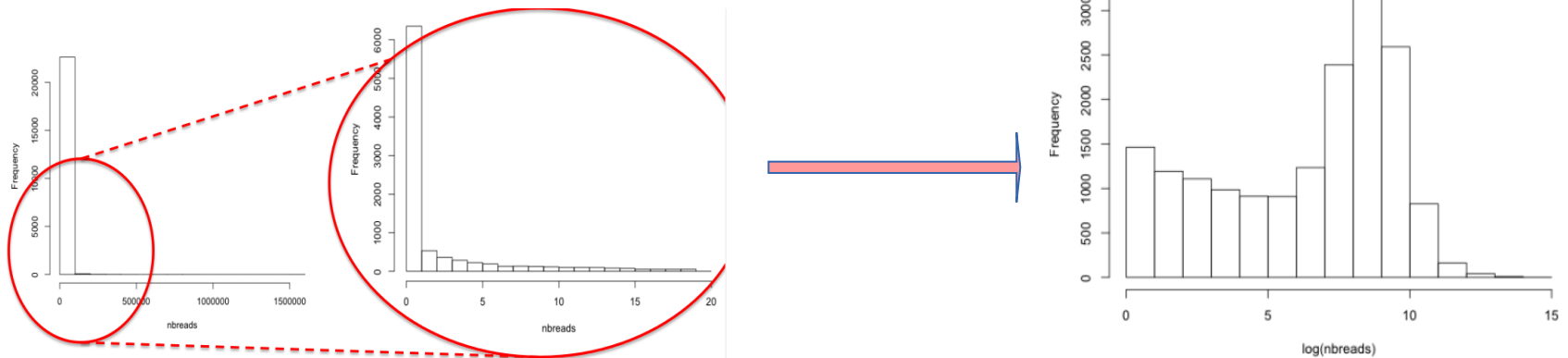
### 3) Normalisation prenant en compte le biais associé à la composition en GC

- méthode du Total Count peu efficace (pas de prise en compte des différences possibles entre les compositions en ARN des conditions)
- méthode RPKM peu efficace (même dans les cas où un biais lié à la longueur des gènes existe, l'utilisation du RPKM ne permet pas de le corriger complètement)
- méthodes à privilégier : Upper-Quartile, RLE, TMM

## 4) La recherche de gènes différentiellement exprimés

# Modélisation des données

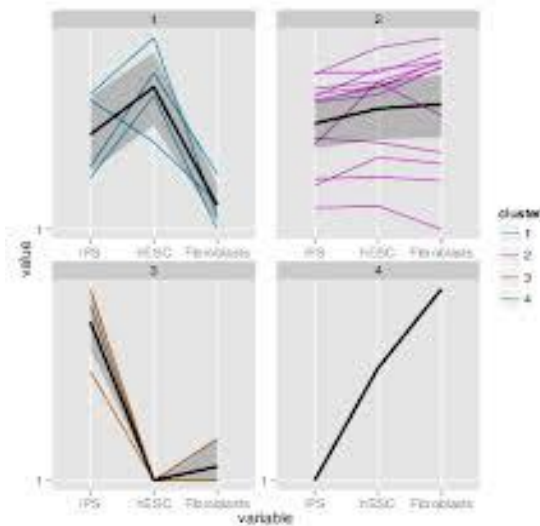
- Utilisation non pas du nombre de reads mais de  $\log(\text{nb reads})$  pour que cela suive une loi statistique
- + nécessité de transformer les « 0 »
- => loi binomiale négative



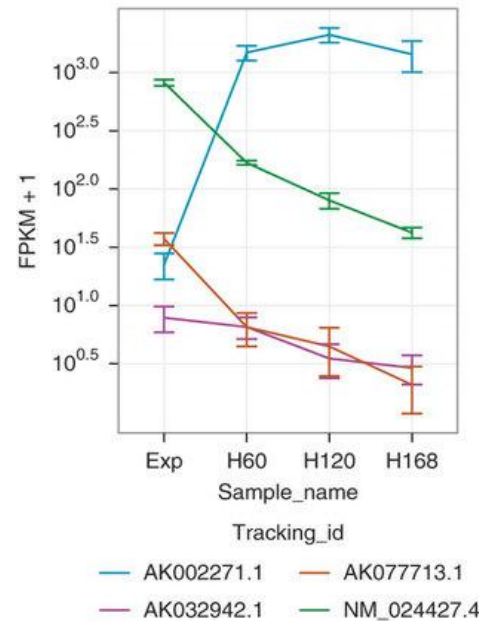
- Utilisation du  $\log(\text{FoldChange})$
- Fold Change** = ratio entre les 2 niveaux d'expression
- = ratio de la valeur finale sur la valeur initiale

# Méthodes basés sur les RPKM (Cuffdiff)

# Cuffdiff - CummeRbund

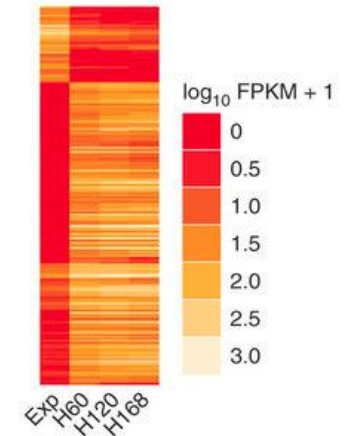


**a** `expressionPlot(isoforms(tpn1), logMode=T)`

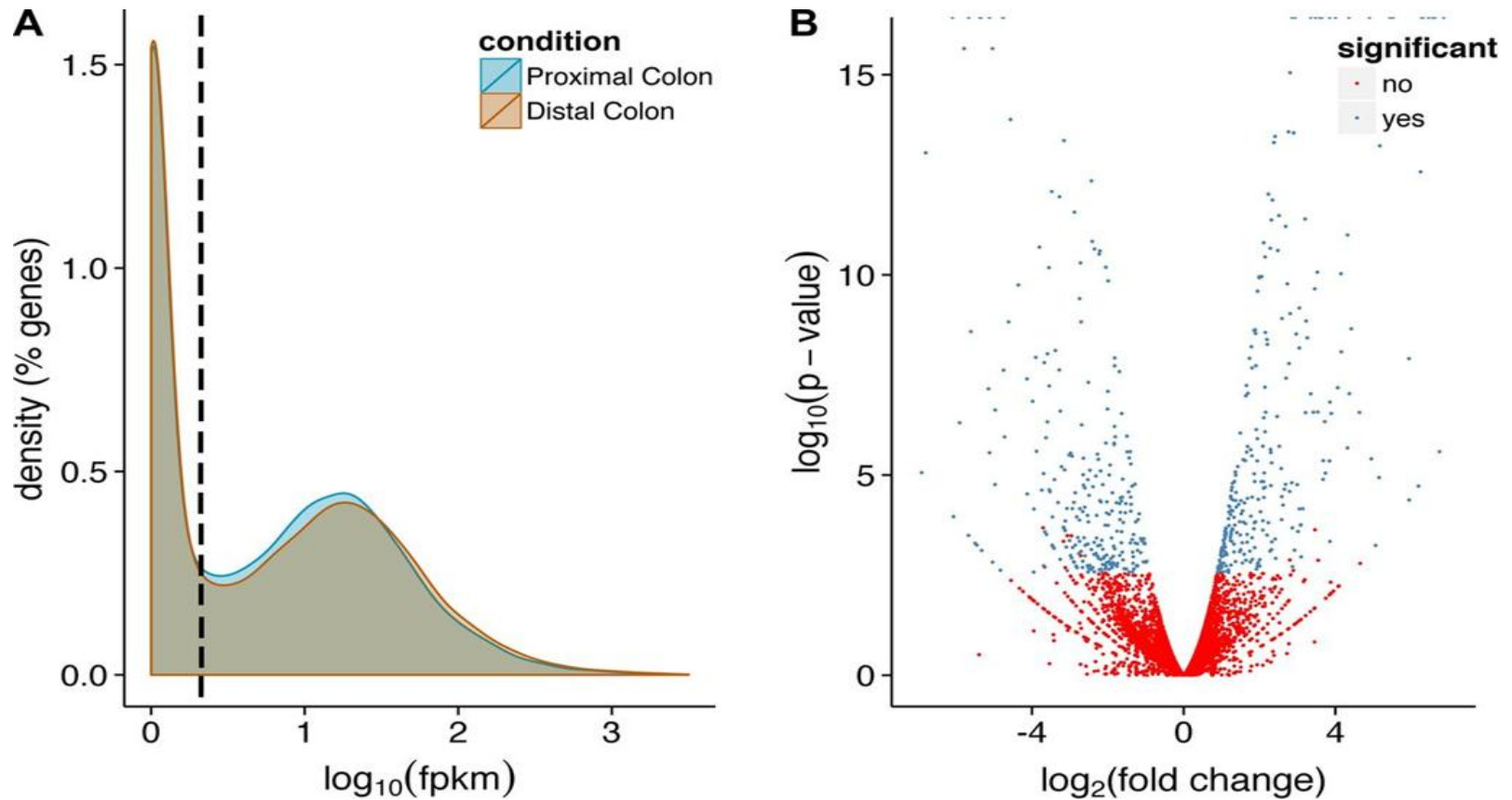


**b**

```
sig_genes <- getGenes(cd, geneIdList)
csHeatmap(sig_genes,
           clustering="row",
           labRow=F)
```



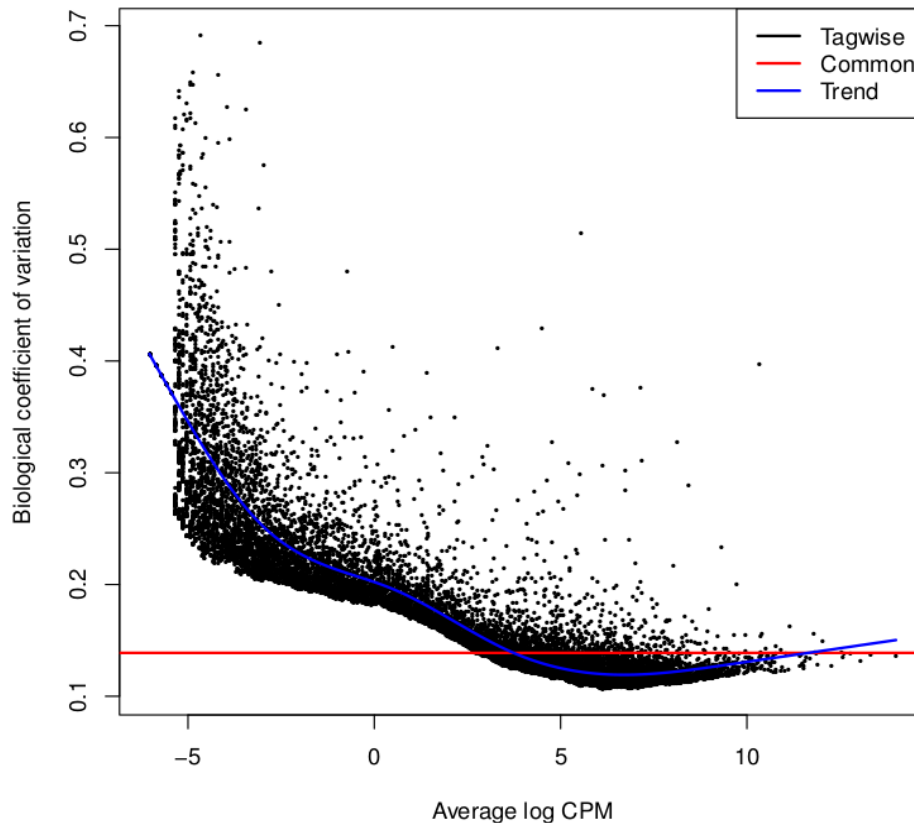
## Cuffdiff - CummeRbund



# Méthodes basées sur les normalisation inter-banques (RLE, TMM, Upper-Quartile) (EdgeR et DESeq)



## EdgeR



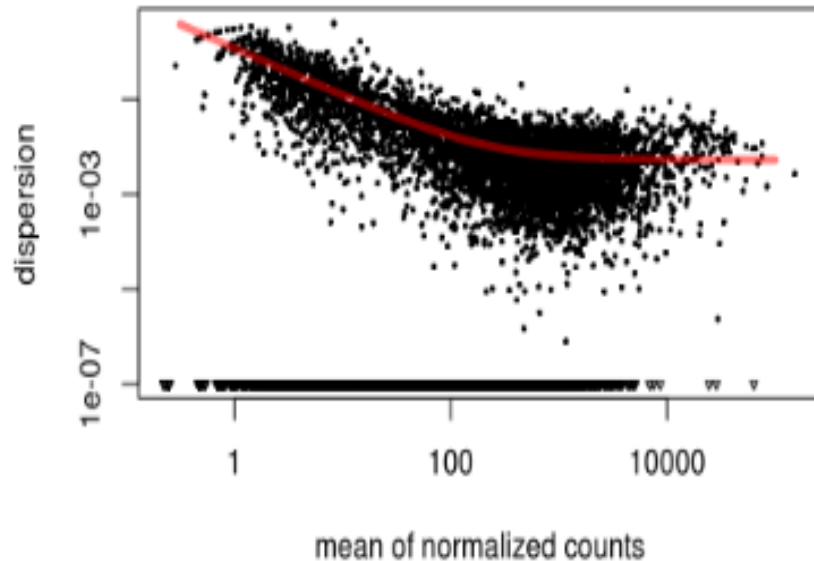
Estimation de la dispersion  
(entre réplicats biologiques)

- utilisation de la valeur individuelle (« tagwise ») ou de la valeur ajustée « trend » ou « common »

- 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

## Comparaison des outils

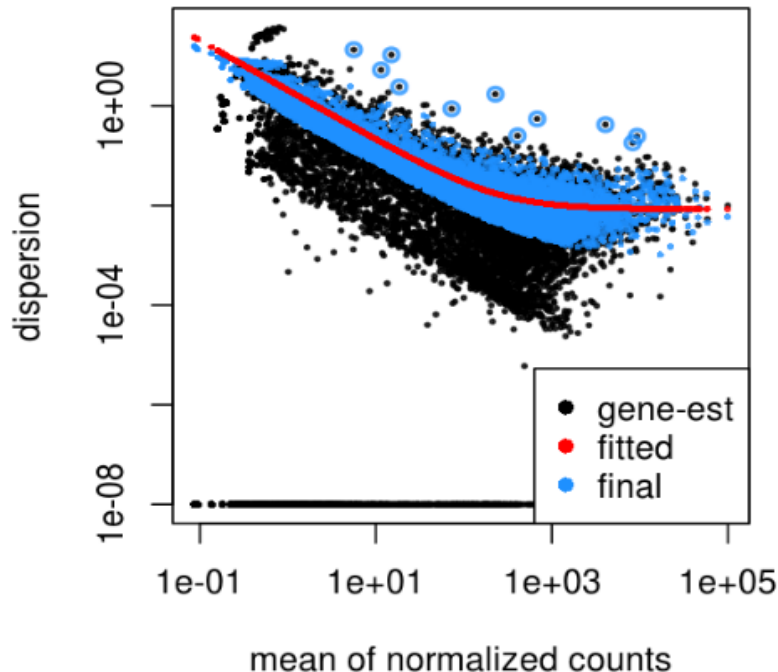
DESeq utilise une estimation de la variance qui la rend moins permissive pour les grandes variabilités entre conditions. Dès qu'au moins l'une des conditions présente une variabilité importante, la méthode ne fait pas confiance à ce gène et ne va pas le considérer comme différentiellement exprimé, même s'il y a une grande différence entre conditions.

A l'opposé, quand la variabilité intra-condition 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 à privilégier pour des expérimentations très répétables

DESeq2 plus souple, sera moins stringent et détectera plus de gènes différentiellement exprimés.

## DESeq2

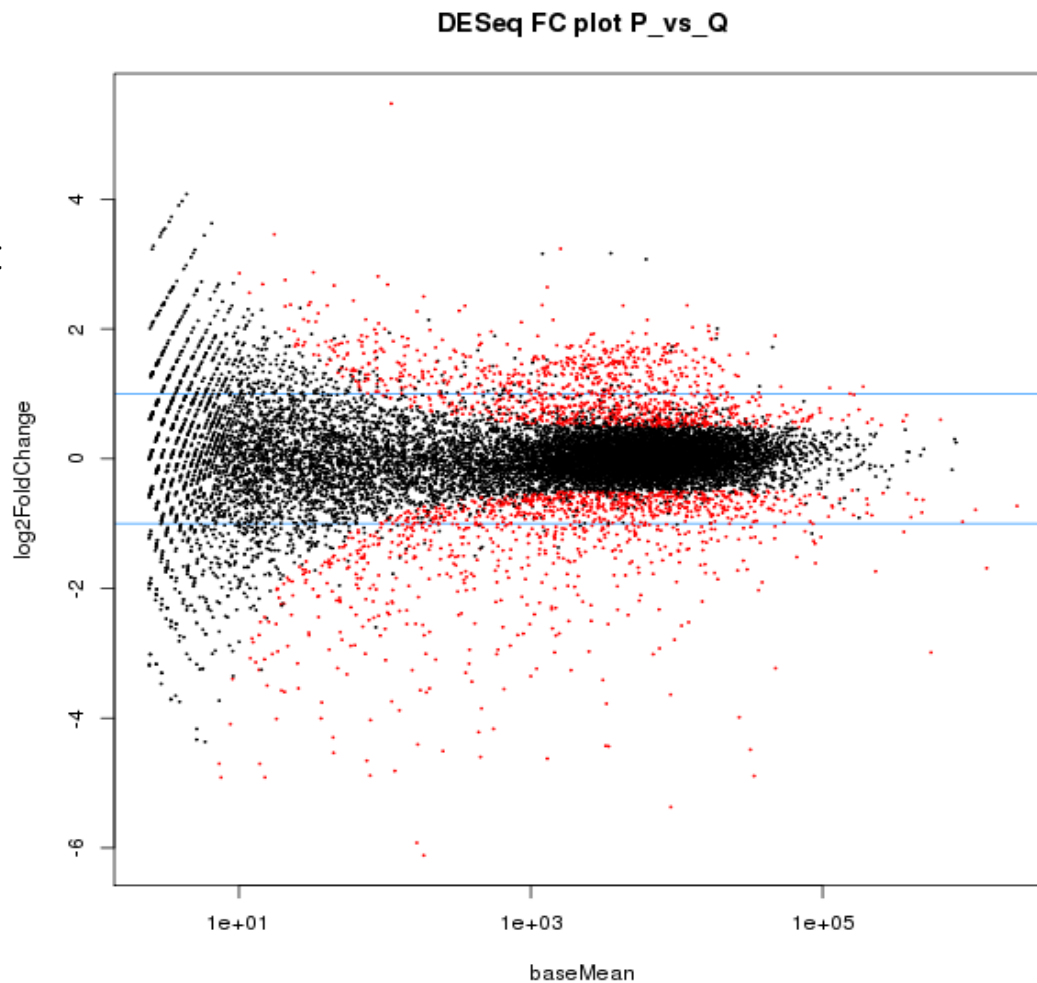


### Estimation de la dispersion

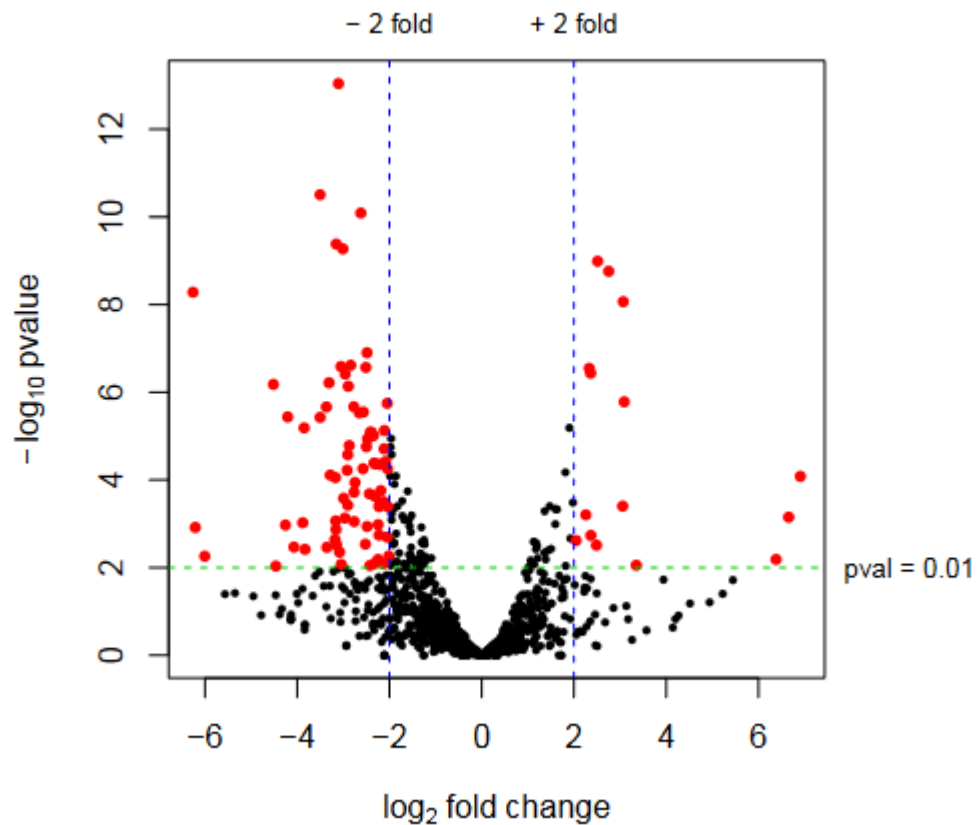
- utilisation d'une valeur intermédiaire (en bleu) entre la dispersion individuelle (en noir) et la dispersion ajustée (en rouge)
- utilisation de la dispersion individuelle si celle-ci est considérée comme extrême par rapport à la distribution globale (points entourés de bleu)

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

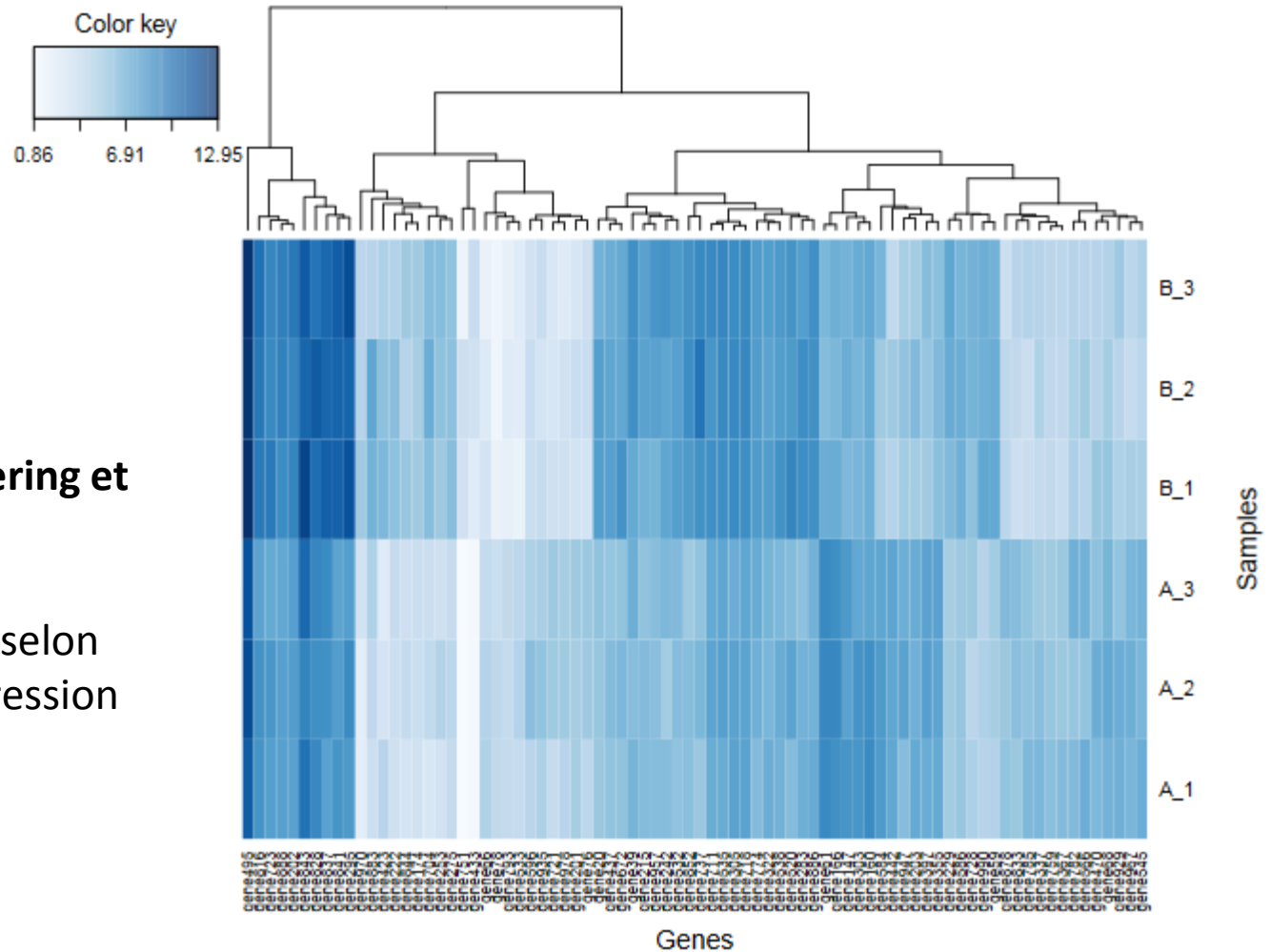
Smear plot / MA plot  
Pvalue adj < 0.05



Volcano plot  
Pvalue adj < 0.01



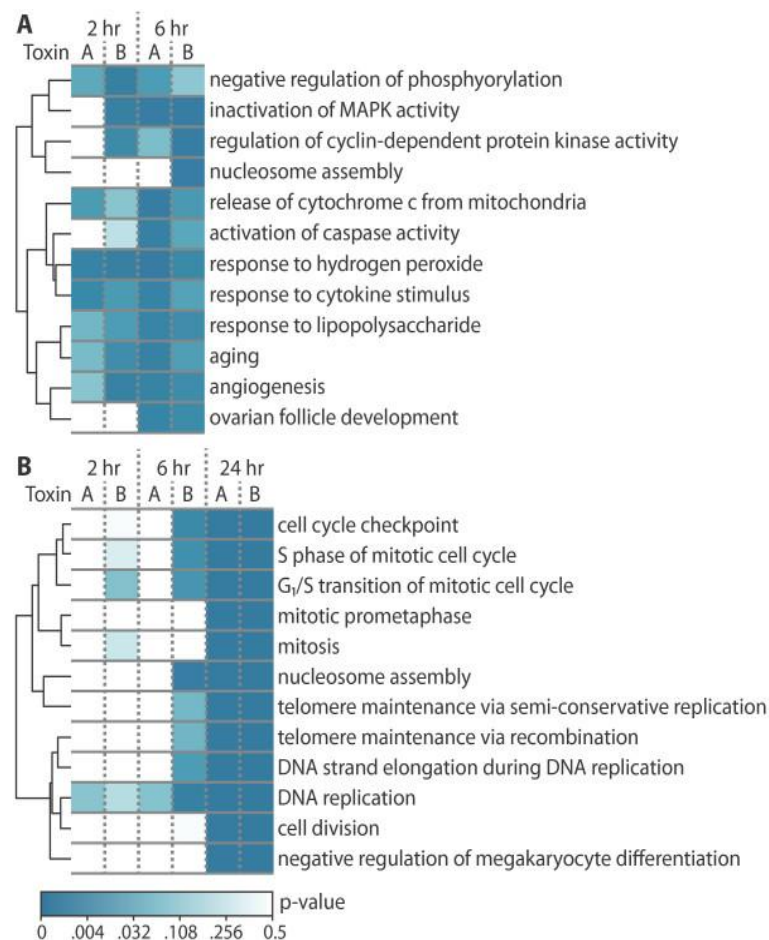
Tutorial: <http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf>



## Hierarchical Clustering et Heatmap

=> Regroupement selon les patterns d'expression





# TopGO : Etude de l'enrichissement de termes Gene Ontology

**Nécessite de disposer d'une annotation fonctionnelle GO des transcrits**

**=> Teste s'il existe des enrichissements significatifs de fonctions GO entre les gènes DE et les gènes non-DE (entre 2 conditions)**



DiffExDB (Differential Expression Database) Quick search Login

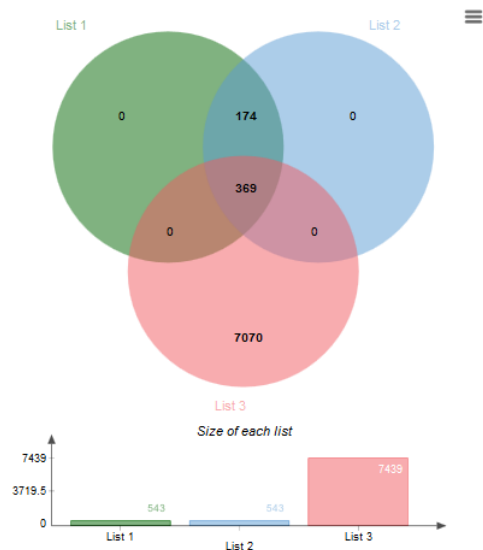
Choose a species:

Oryza sativa

Project	Experiment1	Experiment2	Min p-value	Min logFC	Max logFC
Compare: Response to M.graminicola (Petitot et al, 2016)	O. sativa.nipponbare 0dpi	vs O. sativa.nipponbare 2dpi	0.001	-20	20
<input checked="" type="checkbox"/> intersect Response to M.graminicola (Petitot et al, 2016)	O. sativa.nipponbare 0dpi	vs O. sativa.nipponbare 2dpi	0.001	-20	20
<input checked="" type="checkbox"/> intersect Response to M.graminicola (Petitot et al, 2016)	O. sativa.nipponbare 0dpi	vs O. sativa.nipponbare 8dpi	0.001	-20	20
<input type="checkbox"/> intersect Response to M.graminicola (Petitot et al, 2016)	O. sativa.nipponbare 0dpi	vs O. sativa.nipponbare 2dpi	0.001	-20	20
<input type="checkbox"/> intersect Response to M.graminicola (Petitot et al, 2016)	O. sativa.nipponbare 0dpi	vs O. sativa.nipponbare 2dpi	0.001	-20	20

Filter by genes: enter a list of genes:

submit



Click on a venn diagram figure to display the linked elements:

Common elements in List 1 List 2 List 3 :

LOC\_Os10g25060  
LOC\_Os05g47950  
LOC\_Os10g20450  
LOC\_Os07g48460  
LOC\_Os03g61280  
LOC\_Os02g51040  
LOC\_Os10g42030  
LOC\_Os01g22249  
LOC\_Os02g18450  
LOC\_Os08g29570

# DiffExDB

Application web pour  
l'exploration de  
données d'expression  
différentielle:

- Croisement entre comparaisons
- Carte d'expression Heatmap

<http://bioinfo-web.mpl.ird.fr/cgi-bin2/microarray/public/diffexdb.cgi>

# ShortStack: Traitement de small RNA

## BIOINFORMATICS

### ShortStack: Comprehensive annotation and quantification of small RNA genes

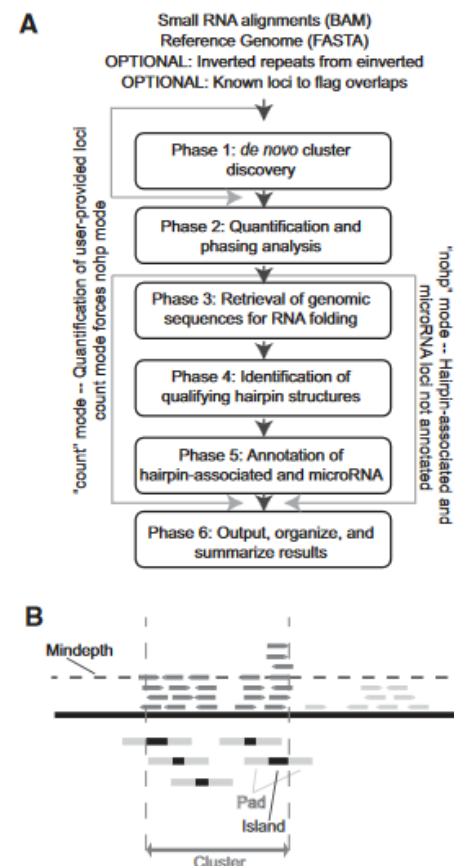
MICHAEL J. AXTELL<sup>1</sup>

Department of Biology, and Huck Institutes of the Life Sciences, Penn State University, University Park, Pennsylvania 16802, USA

#### ABSTRACT

Small RNA sequencing allows genome-wide discovery, categorization, and quantification of genes producing regulatory small RNAs. Many tools have been described for annotation and quantification of microRNA loci (*MIRNAs*) from small RNA-seq data. However, in many organisms and tissue types, *MIRNA* genes comprise only a small fraction of all small RNA-producing genes. ShortStack is a stand-alone application that analyzes reference-aligned small RNA-seq data and performs comprehensive de novo annotation and quantification of the inferred small RNA genes. ShortStack's output reports multiple parameters of direct relevance to small RNA gene annotation, including RNA size distributions, repetitiveness, strandedness, hairpin-association, *MIRNA* annotation, and phasing. In this study, ShortStack is demonstrated to perform accurate annotations and useful descriptions of diverse small RNA genes from four plants (*Arabidopsis*, tomato, rice, and maize) and three animals (*Drosophila*, mice, and humans). ShortStack efficiently processes very large small RNA-seq data sets using modest computational resources, and its performance compares favorably to previously described tools. Annotation of *MIRNA* loci by ShortStack is highly specific in both plants and animals. ShortStack is freely available under a GNU General Public License.

Keywords: microRNA; small RNA; siRNA; software; bioinformatics; next-generation sequencing



## Exercice:

- 1) Réaliser un comptage de reads par gène à partir d'un fichier BAM à l'aide de l'outil samtools idxstats.
- 1) Dans Galaxy, importer un jeu de données complet qui sera utilisé pour l'analyse d'expression différentielle  
*Shared data => Data libraries => Formation => RNASeq*
- 1) Filtrer les séquences pour ne garder celles ayant au moins 10 reads sur l'ensemble des conditions. Combien de gènes avez-vous filtré?  
*Il n'est pas possible de réaliser des tests fiables sur de faibles comptages. Ceci permet de limiter le nombre de tests statistiques et ainsi diminuer l'effet des corrections pour tests multiples*
- 1) Réaliser une étude d'expression différentielle avec les outils EdgeR. Observer les sorties graphiques du logiciel. En fixant un seuil de p-value à 0.01, combien de gènes sont retrouvés DE?