Session de formation 2018

12 Mars	Guide de survie à Linux : les commandes de base pour débuter sur un serveur linux
13 Mars	Linux avancé : manipuler et filtrer des fichiers sans connaissance de programmation
15 Mars	Initiation à l'utilisation du cluster bioinformatique itrop
22 Mars	Initiation à git
23 Mars	Initiation aux gestionnaires de workflow South Green: Galaxy ou TOGGLe
26 Mars	**Initiation aux analyses de données transcriptomiques**

Bioversity International

Institut de Recherche pour le Développement

diade

IPME

SouthGreen bioinformatics platform

agap

UMR BGPI Biologie et Génétique des Interactions Plante-Parasite

LSTM Laboratoire des Symbioses Tropicales et Méditerranéennes

INRA SCIENCE & IMPACT

cirad

Bioversity International

Institut de Recherche pour le Développement

SouthGreen bioinformatics platform

diade IPME UMI233 TransVIH-MI MiVEGEC

plateau i-trop

agap UMR BGPI LST&M

INRA SCIENCE & IMPACT cirad

www.southgreen.fr

https://github.com/SouthGreenPlatform

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

Session de formation 2018

- Toutes nos formations :
  https://southgreenplatform.github.io/trainings/
- Topo & TP : **Analyse RNASeq**
- Environnement de travail : **Logiciels à installer**

Initiation aux analyses de données transcriptomiques

www.southgreen.fr

https://southgreenplatform.github.io/trainings

Objectif du module

Objectif!

- 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
- Connaître

Applications

- Mapping and counting using TOGGLe : topHat + HTSeq-count
- Mapping and counting using Galaxy : kallisto
- Differential expression analysis: EdgeR, limma
- Clustering, co-expression network: WGCNA, R, pivot

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...)

...

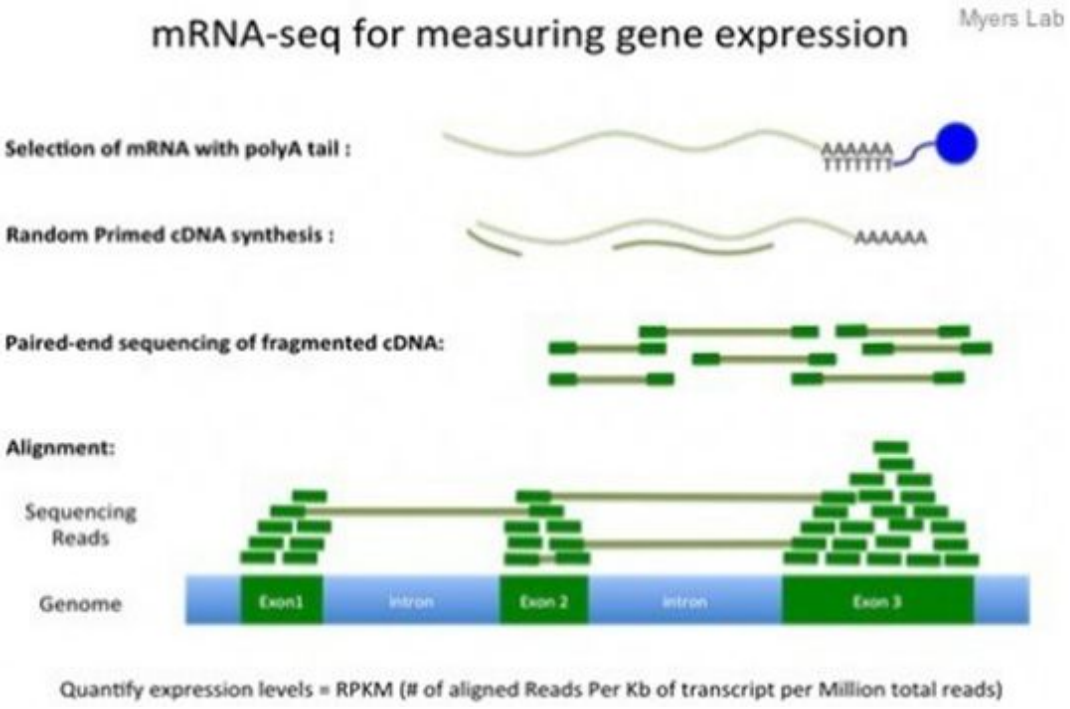
Avantages RNA-Seq / Microarray

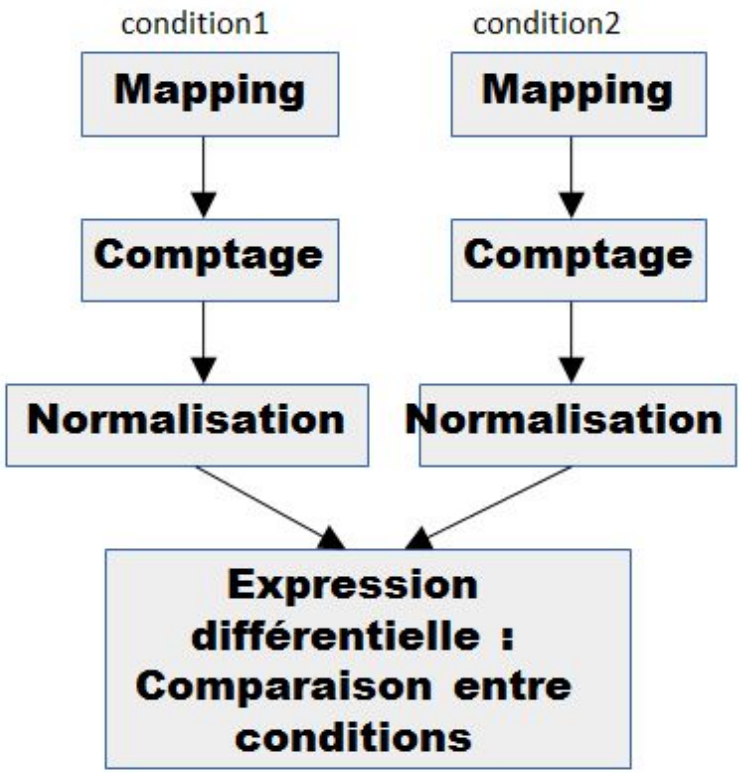
L'accès aux séquences des ARN permet de:

- 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...)

...

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)

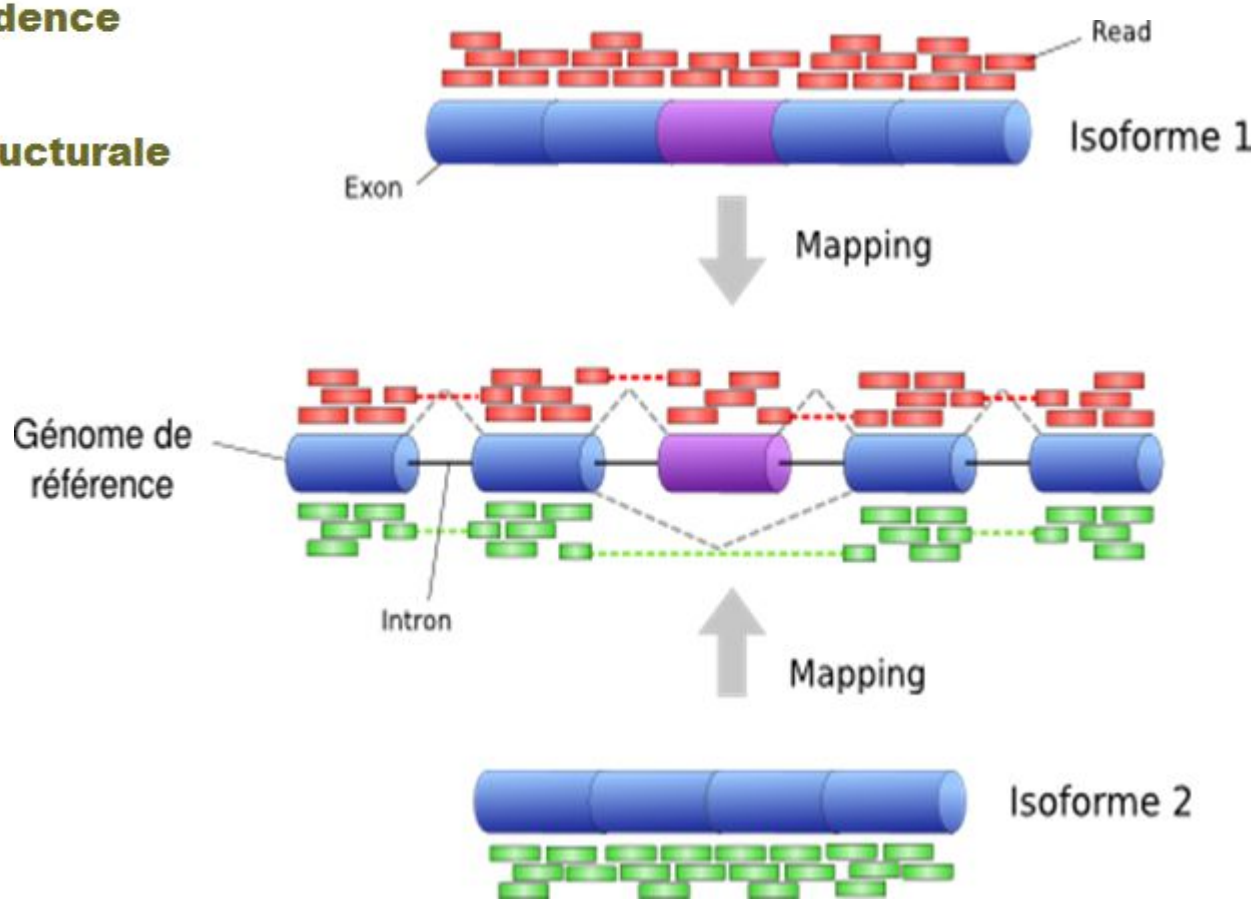
3) Si on dispose d'un transcriptome de référence
Utilisation de mapper classique (ex : BWA, bowtie)

Mapping sur référence génomique

Mapping sur une référence génomique

=> permet la mise en évidence d'isoformes

=> aide à l'annotation structurale du génome



Suite TopHat/Cufflinks

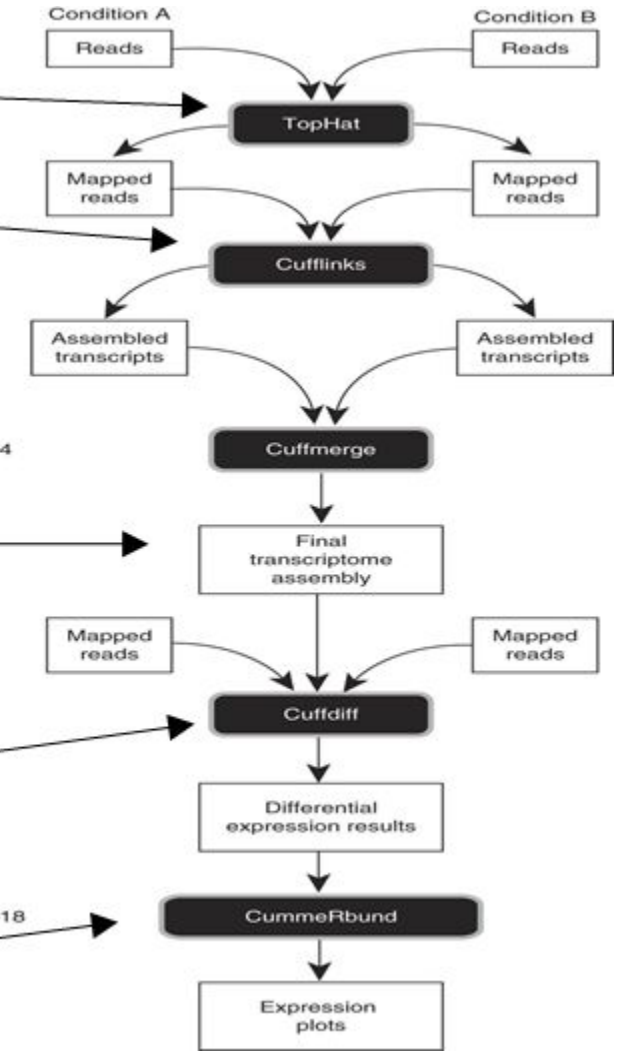
alignement

Assemblage
(pour chaque condition)

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

Statistiques DE

Plots



2- Comptage

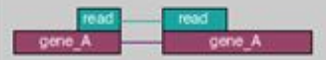
Choix de l'outil de comptage

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

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

Normalisation des données

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

Normalisation des données

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

Normalisation des données

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

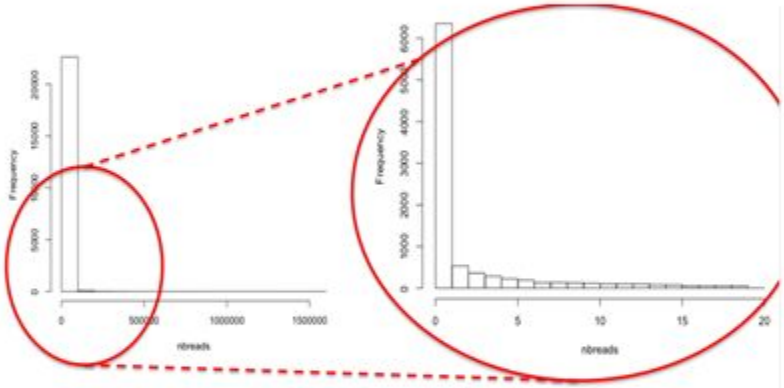
3) Normalisation prenant en compte le biais associé à la composition en GC

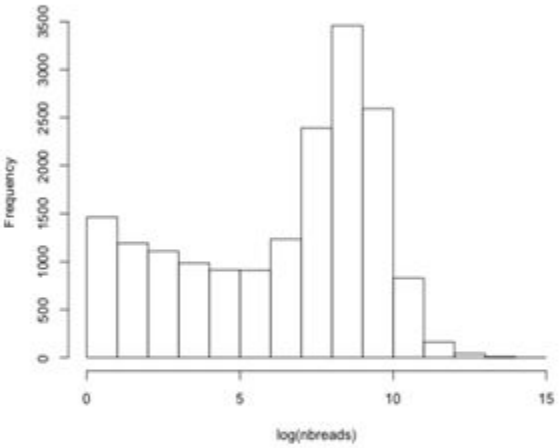
- méthode du Total Count peu efficace (pas de prise en compte des différences possibles entre les compositions en ARN des conditions)
- méthode RPKM peu efficace (même dans les cas où un biais lié à la longueur des gènes existe, l'utilisation du RPKM ne permet pas de le corriger complètement)
- méthodes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

4) Recherche de gènes différentiellement exprimés

Modélisation des données

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





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

pvalue and FDR

pvalue= risque/probabilité de déclarer un gène différentiellement exprimé alors qu'il ne l'est pas

pvalue 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

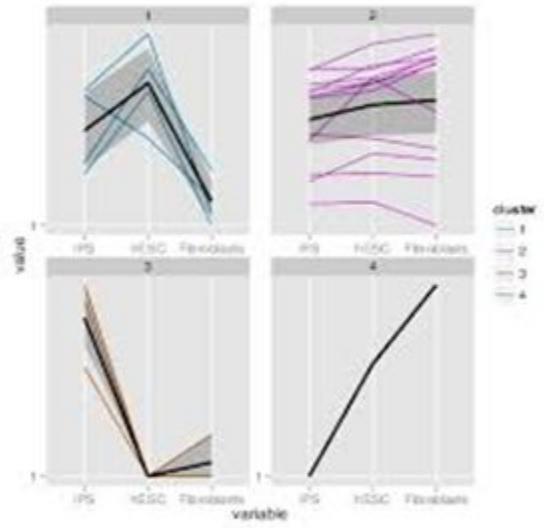
=> besoin de correction et utiliser une pvalue ajustée adaptée aux tests multiples:

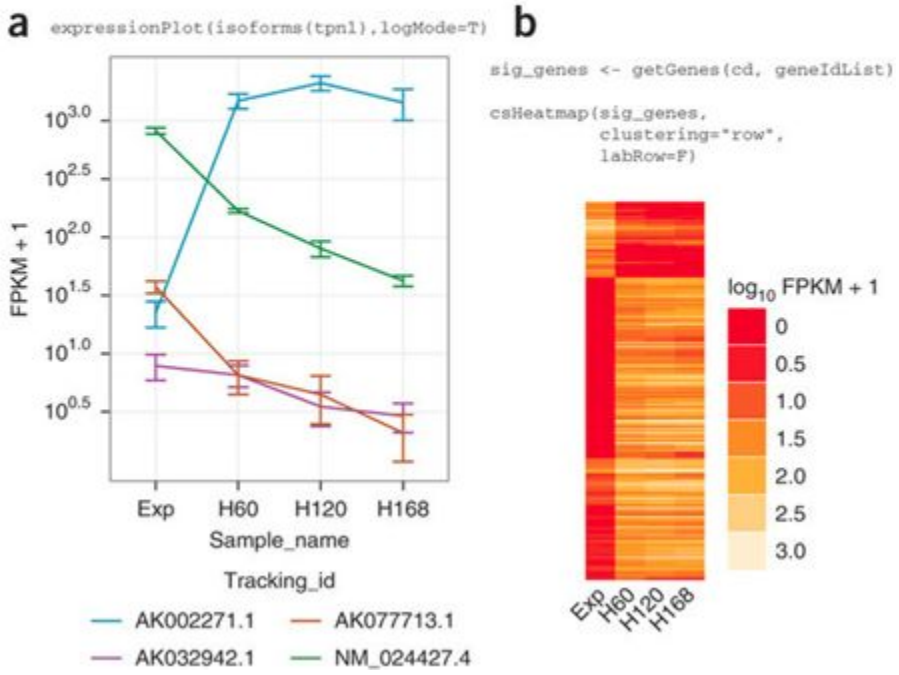
- procédure de Benjamini-Hochberg (BH) qui consiste à contrôler le False Discovery Rate (**FDR**), c'est à dire la proportion de faux positifs dans les gènes déclarés différentiellement exprimés

- procédure de Bonferroni (+ stringent)

Méthodes basées sur les RPKM (cuffdiff)

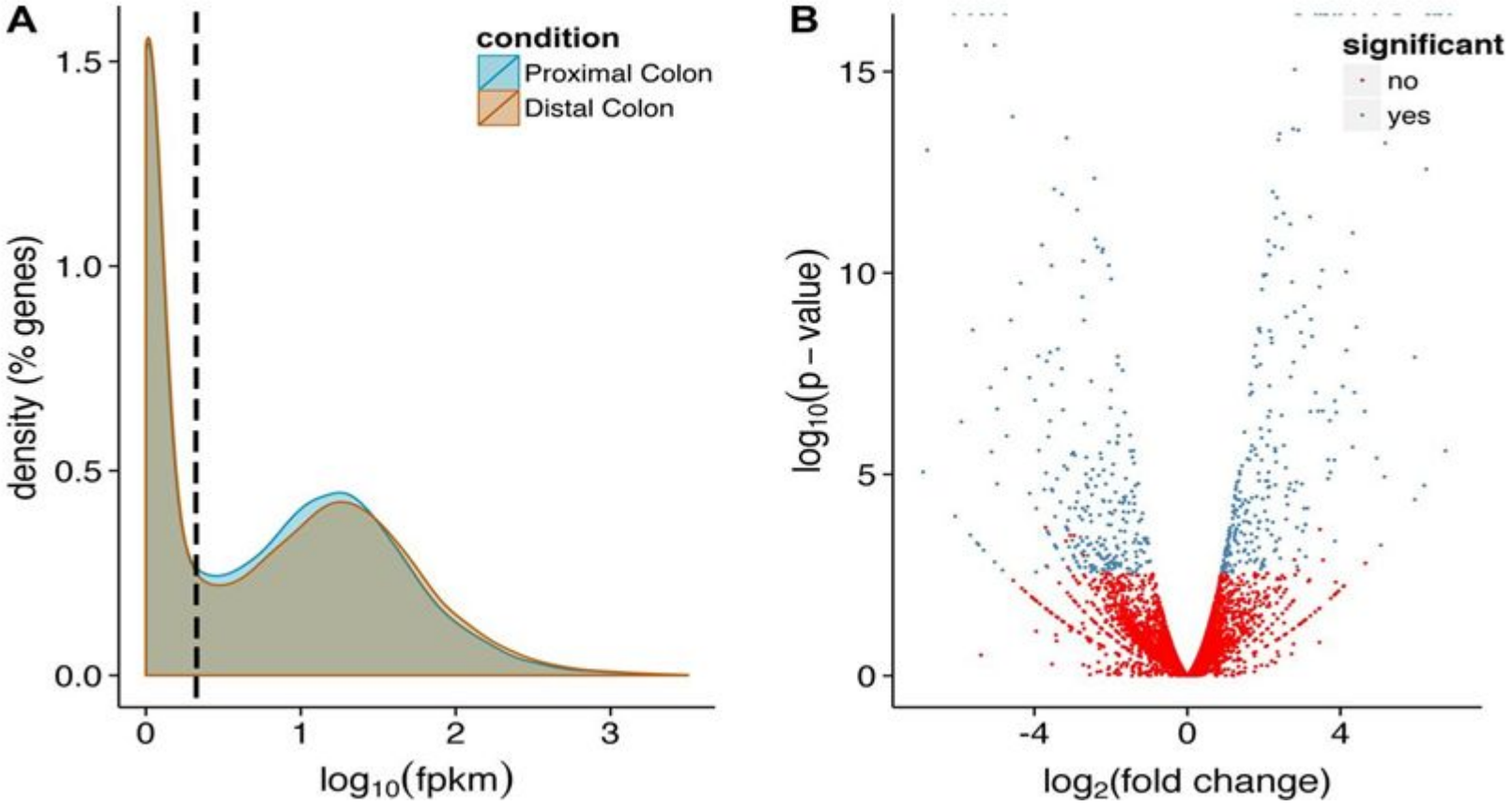
Cuffdiff - CummeRbund





Méthodes basées sur les RPKM (cuffdiff)

Cuffdiff - CummeRbund



Méthodes basées sur les normalisations interbanques

Méthodes basées sur les normalisation inter-banques
(RLE, TMM, Upper-Quartile)

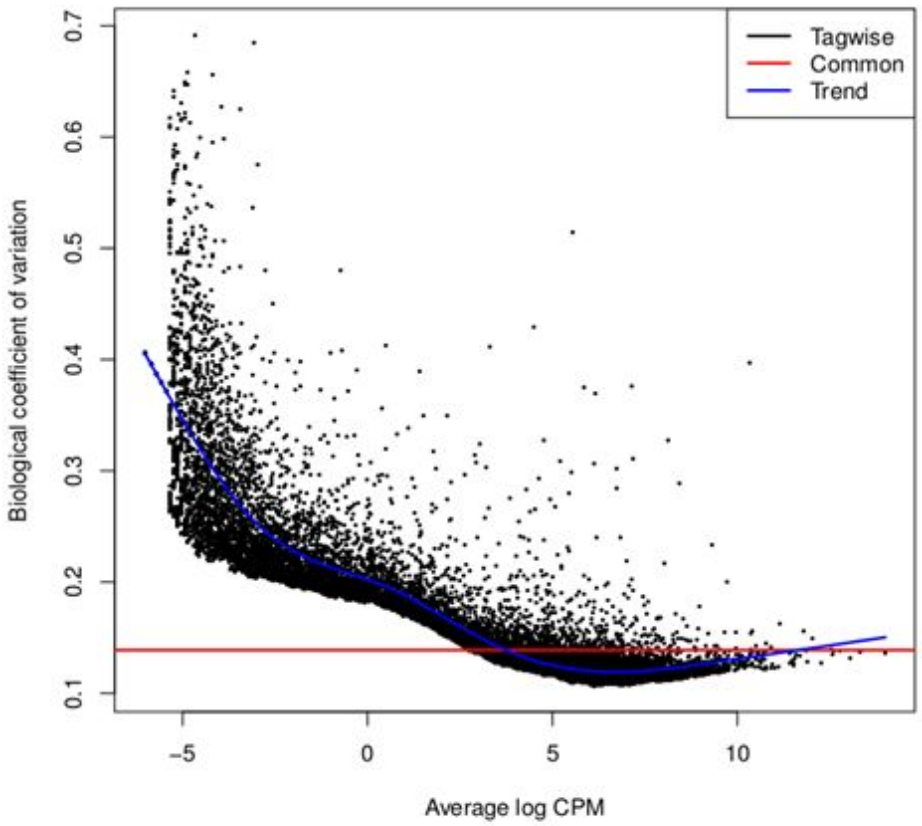
Avec une correction par la variance

(basées sur une loi binomiale négative)

(EdgeR et DESeq)

Méthodes basées sur les normalisations interbanques

EdgeR



Estimation de la dispersion
(entre réplicats biologiques)

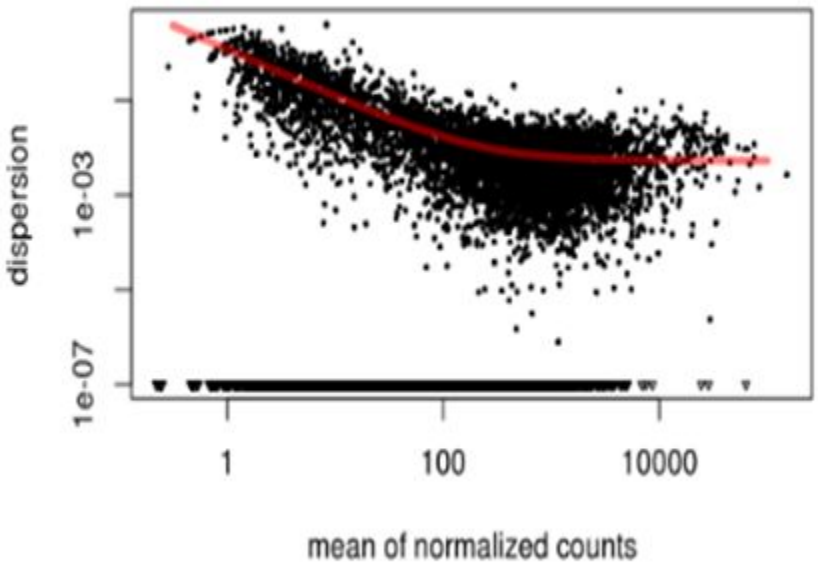
- utilisation de la valeur individuelle (« tagwise ») ou de la valeur ajustée « trend » ou « common » pour le calcul des tests statistiques de DE

- 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)

Méthodes basées sur les normalisations interbanques

DESeq



Estimation de la dispersion

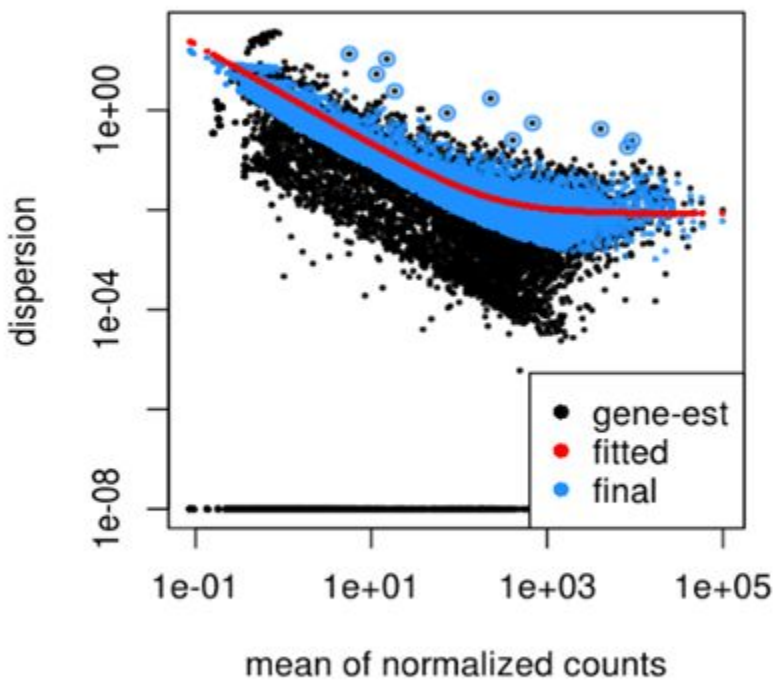
- utilisation de la valeur ajustée pour les transcrits dont l'estimateur individuel (en noir) inférieur à la valeur ajustée

- 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

Méthodes basées sur les normalisations interbanques

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

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

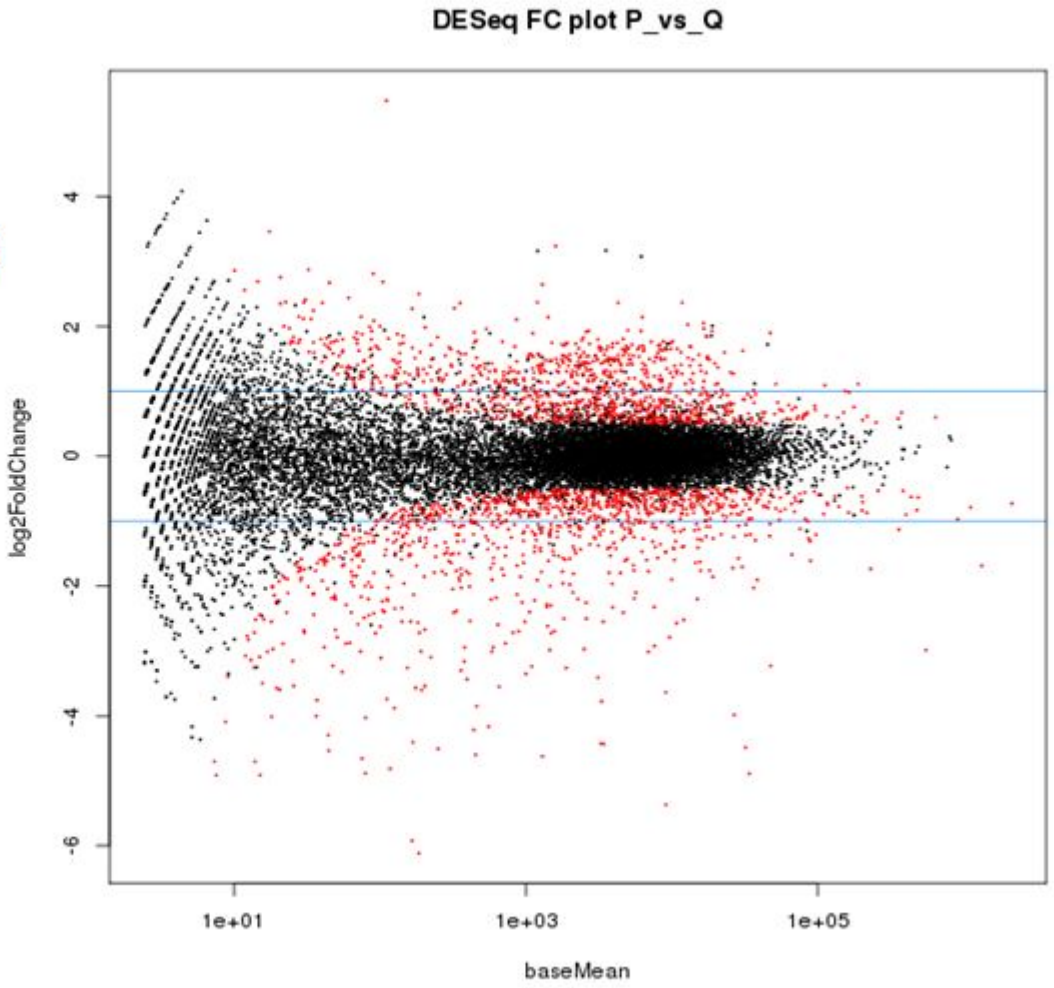
DESeq2 plus souple, sera moins stringent et détectera plus de gènes différentiellement exprimés.

5) Plots and Graphical Représentations

Smear plot

Smear plot / MA plot

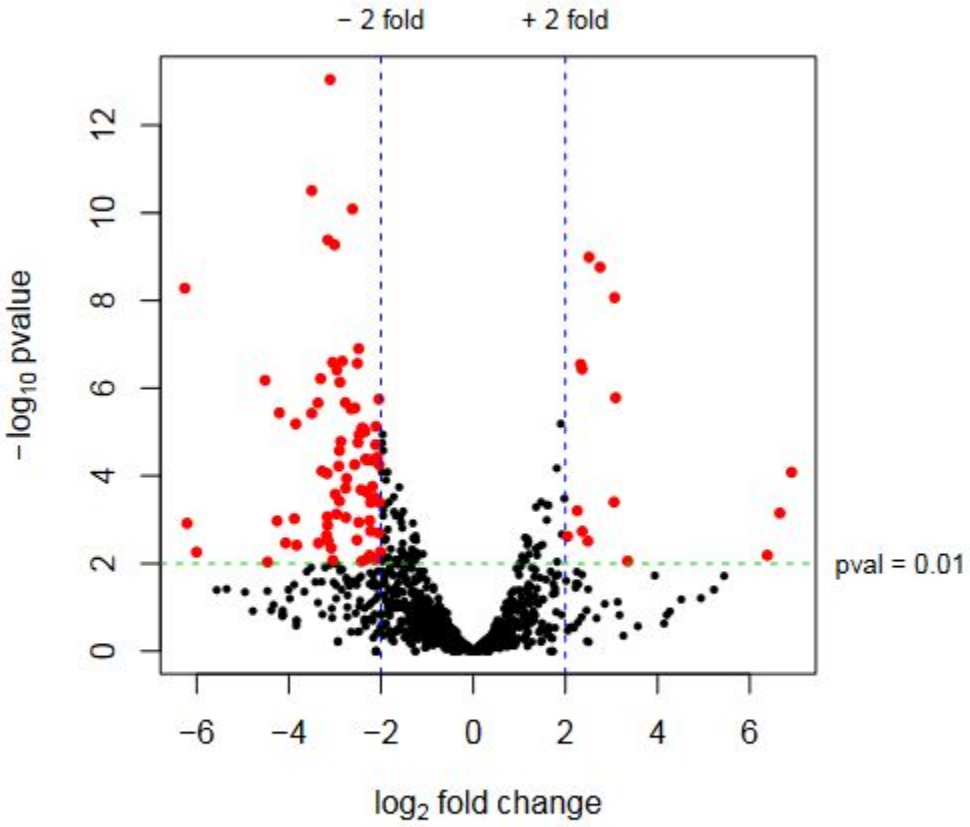
Pvalue adj < 0.05



Volcano plot

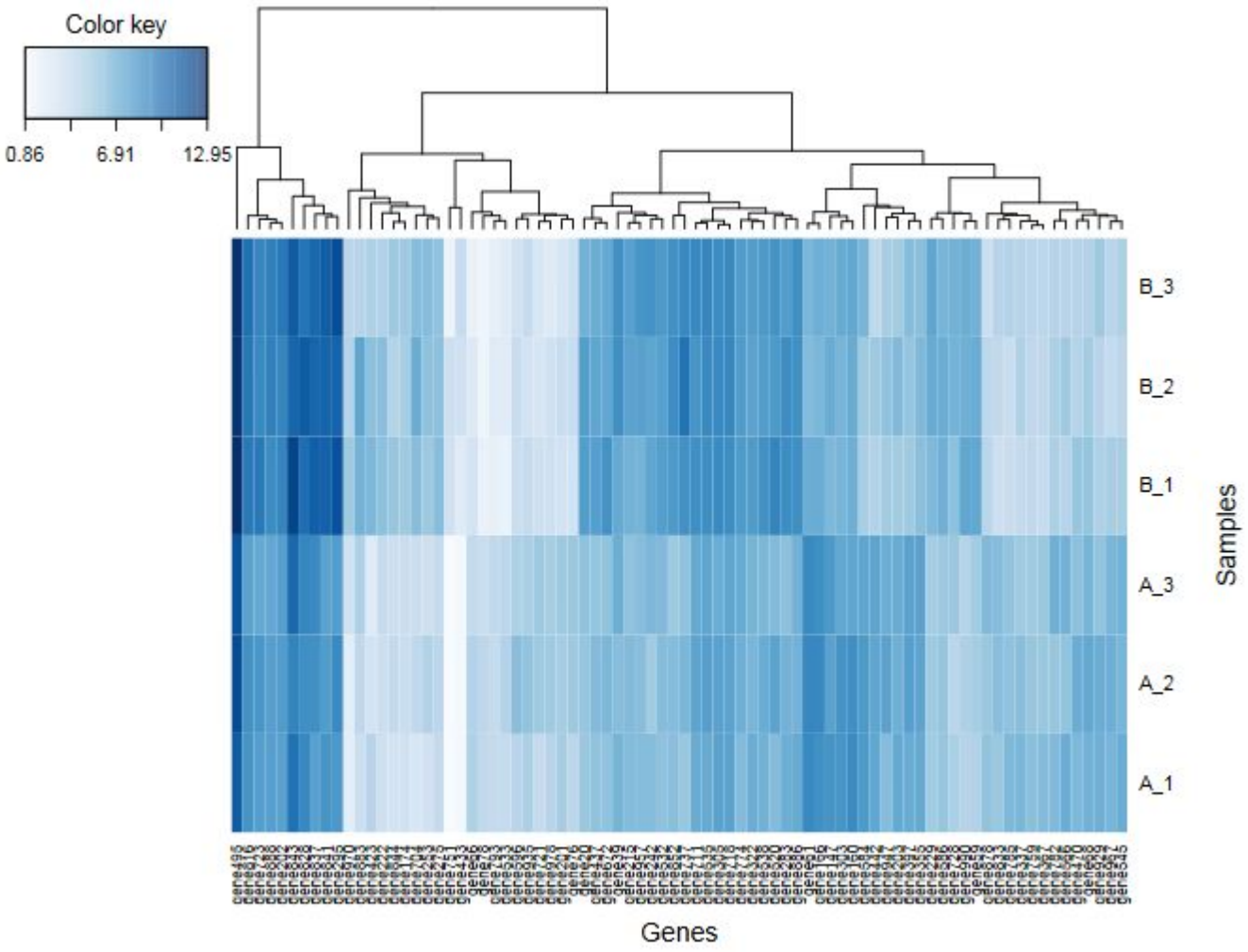
Volcano plot

Pvalue adj < 0.01

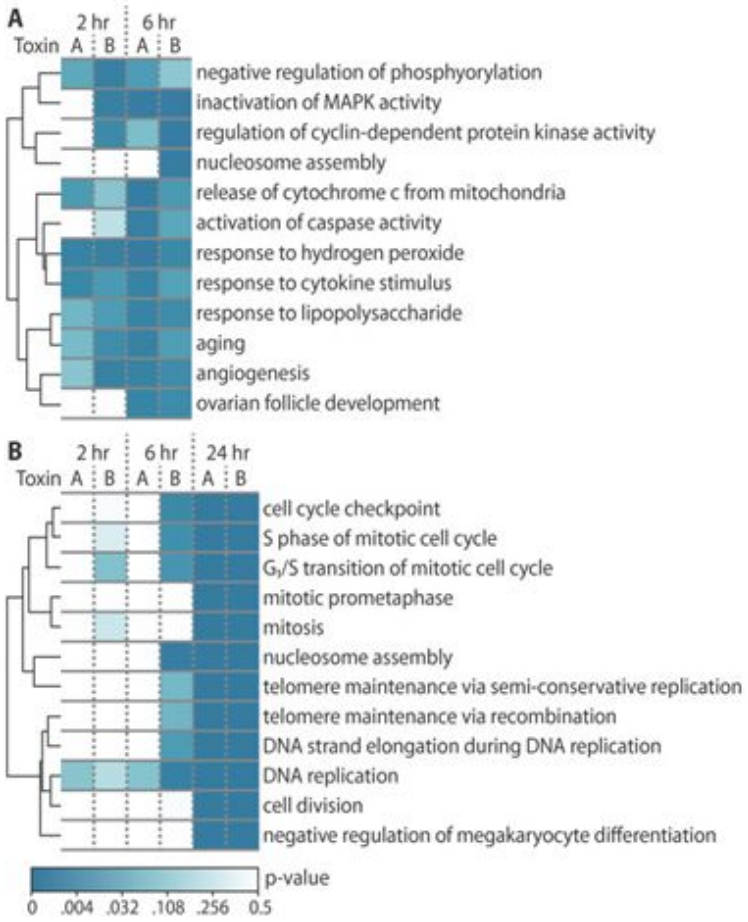


Tutorial: http://www.nathalievilla.org/doc/pdf/tutorial-rnaseq.pdf

Hierarchical clustering / Heatmap



Enrichissement Gene Ontology

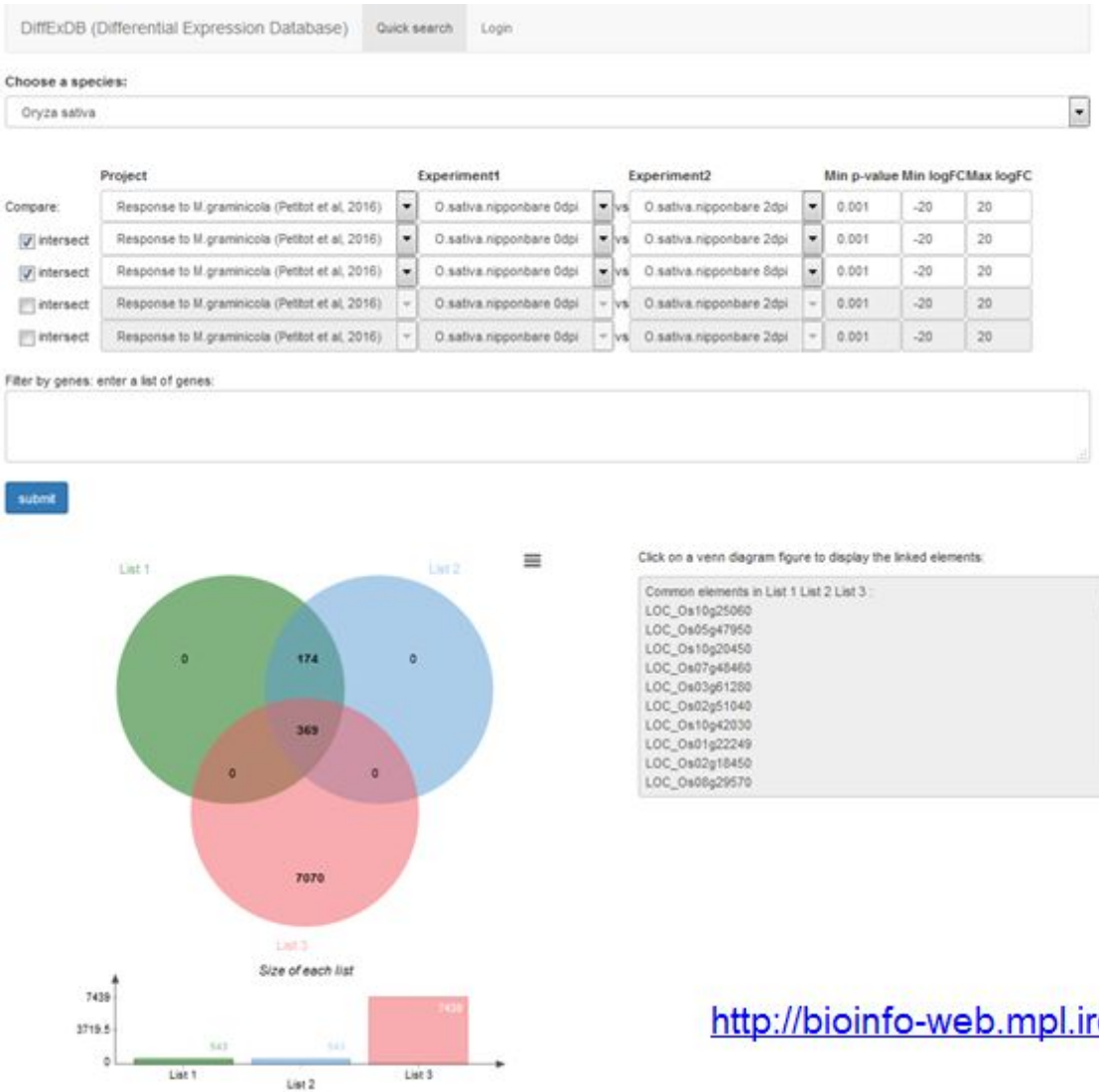


TopGO : Etude de l'enrichissement de termes Gene Ontology

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 : base de données d'expression



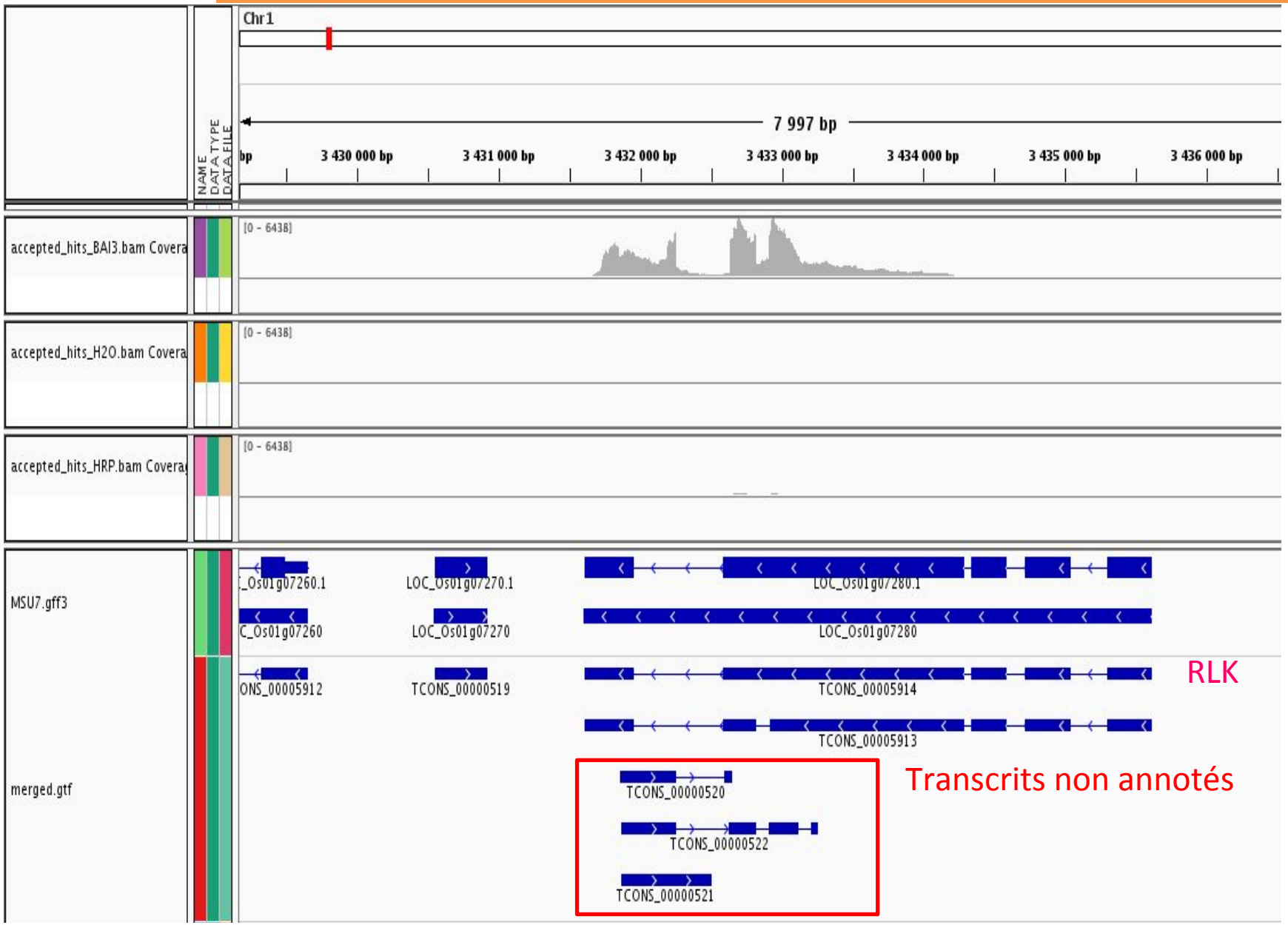
DiffExDB

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

Local Genome Browser (IGV)



Local Genome Browser (IGB)