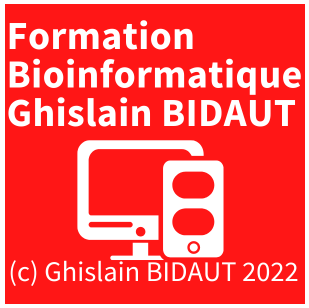


Analyse de données RNA-seq

Ghislain Bidaut, Plateforme Cibi, CRCM, Aix-Marseille Université

24/03/2022

Introduction

Qu'est ce que la bioinformatique ?

- **Apparition en 1970**: B Hesper et P Hogeweg, « Bioinformatica: een werkconcept », *Kameleon*, vol. 1, no 6, 1970, p. 28–29

La bio-informatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation informatique de l'information biologique. Plusieurs champs d'application ou sous-disciplines de la bio-informatique se sont constitués (*Wikipedia*):

- **La bio-informatique des séquences**
- **La bio-informatique structurale**
- **La bio informatique des réseaux**
- **La bio-informatique statistique et des populations**

Projet du séquençage du Génome Humain

- **Idée** lancée en 1985 par 3 scientifiques, **Renatto Dulbecco**, **Robert Sinsheimer** (directeur de UCSC) et **Charles DeLisi**, qui financera le projet (Directeur dept. de biologie du *Département de l'Energie US*)
- **Séquençage** lancé en 1988 par le *National Research Council*. En suisse est créé HUGO (*Human Genome Organisation*) pour la coordination.
- En 1998, Craig Venter, crée *Celera Genomics* avec pour objectif le **séquençage en 3 ans** par séquençage *Shotgun* et le brevetage du génome (!).
- En 2000, la complétion du séquençage est annoncé pour le Consortium et Celera Genomics (match nul) par le président B. Clinton. Coût: $3B.
- *Celera Genomics* avouera avoir utilisé les données du Consortium pour son propre assemblage, mais reproduira un séquençage *de Novo* 3 ans plus tard…
- Publication des séquences brutes en 2001 et des séquences finales en 2004.

Différentes technologies de séquençage

https://planet-vie.ens.fr/thematiques/manipulations-en-laboratoire/la-revolution-de-la-genomique-les-nouvelles-methodes-de

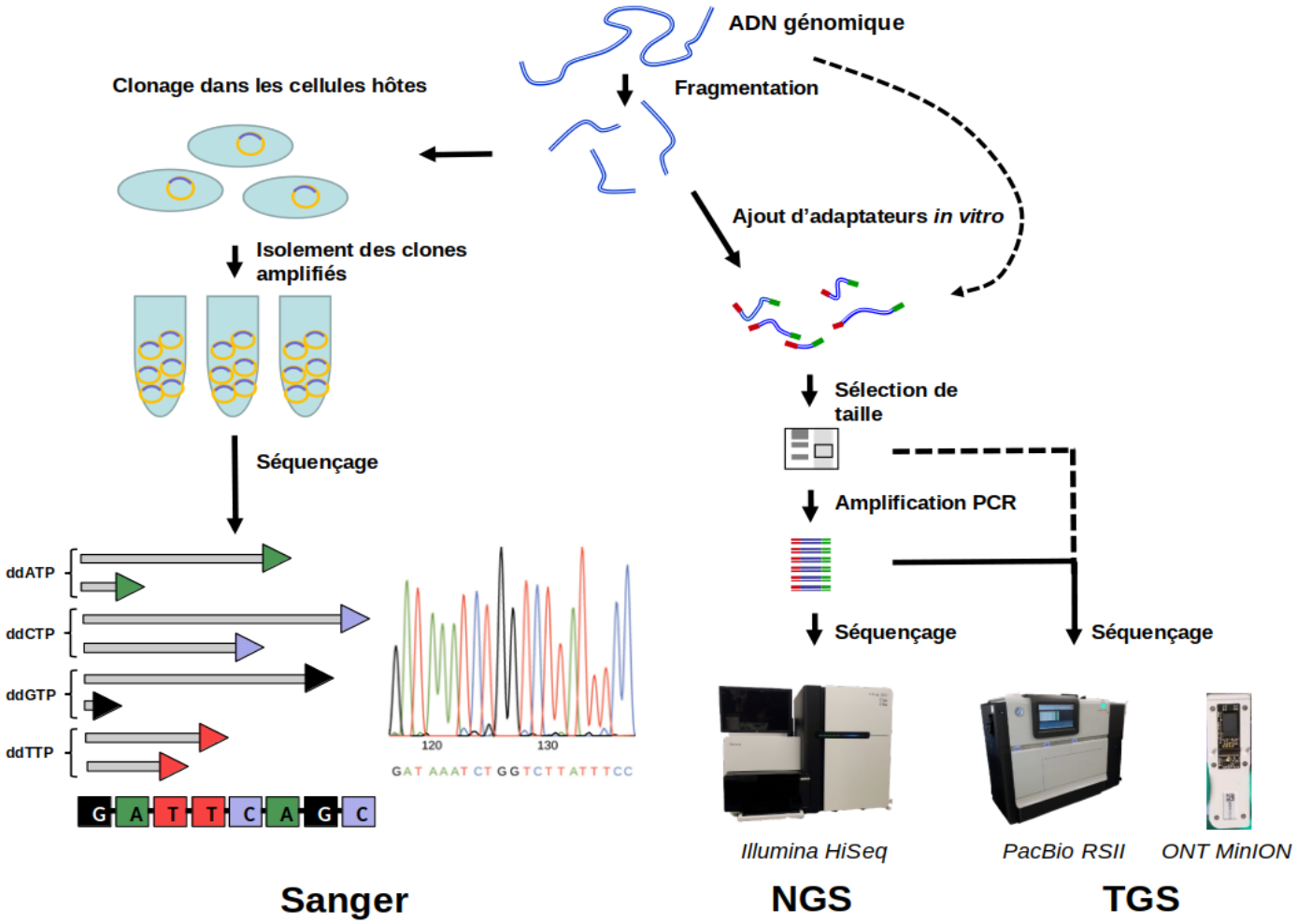
- **Séquençage Sanger**: faible débit, utilisée pour le séquençage du génome humain
- **Séquençage de nouvelle génération**: adapté au séquençage massif d'un grand nombre de génome pour étudier **les variations génétiques** (GWAS). Le leader du marché aujourd'hui est *Illumina*.

Déroulé d'un séquençage

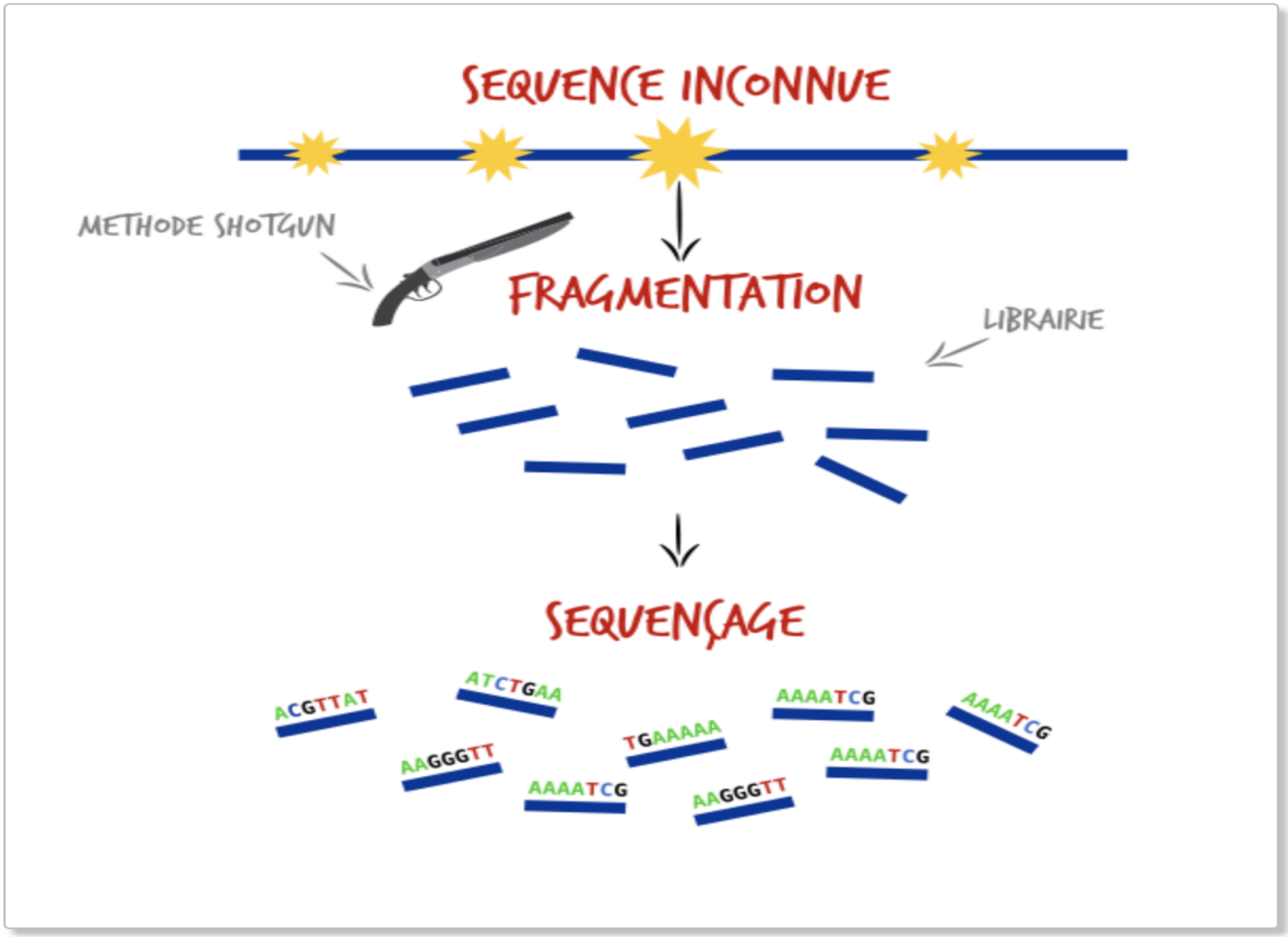
Le principe d'un séquençage NGS consiste à:

- **Créer une *librairie* de fragments d'ADN** (par fragmentation enzymatique et mécanique de l'ADN génomique).
- **Relier ces fragments à des *adaptateurs***, des petites molécules d'ADN de séquences connues nécessaires au séquençeur.
- **Une sélection de la taille minimum et maximum** des fragments est effectuée pour des raisons techniques, notamment l'amplification PCR.
- **Faire une amplification PCR** des fragments.
- **Faire le séquençage**.

Principe général du séquençage

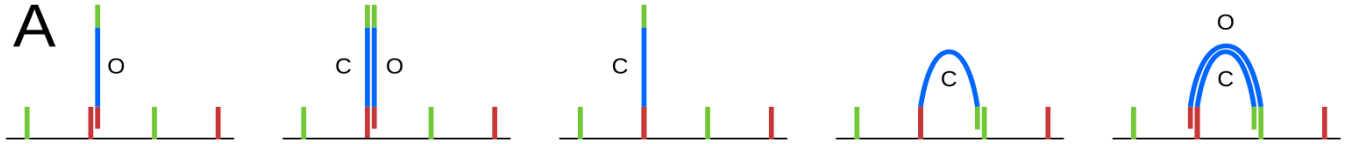


Séquençage Shotgun



Génération des séquences (Illumina)

A



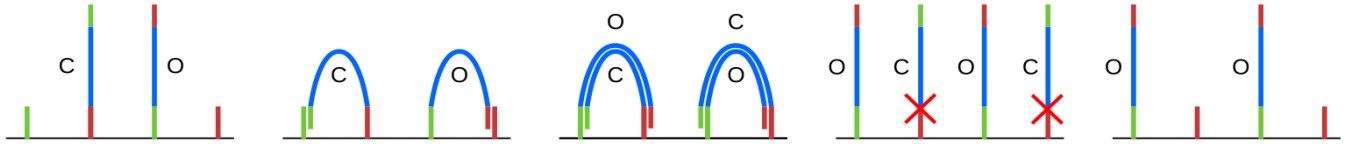
① Une molécule simple brin s'hybride à un oligo-nucléotide fixé à la cellule de flux.

② Un brin complémentaire (C) est synthétisé.

③ La molécule d'origine (O) est enlevée.

④ L'extrémité libre s'hybride.

⑤ Un brin complémentaire est synthétisé.



⑥ Les deux brins, attachés de façon covalente à la cellule de flux, sont séparés par dénaturation.

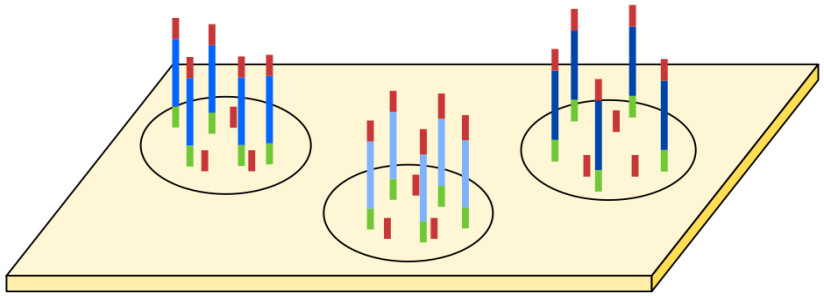
⑦ Les extrémités libres s'hybrident.

⑧ Nouvelle synthèse des brins complémentaires.

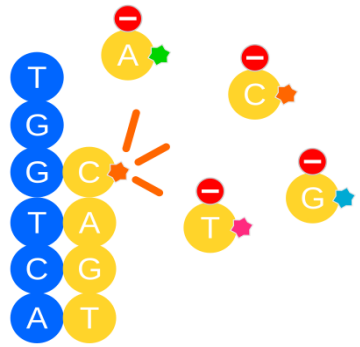
⑨ Dénaturation suivie par l'enlèvement des brins complémentaires à la molécule d'origine.

⑩ Seuls les brins identiques à la molécule d'origine restent et seront séquencés.

B



C



Séquençage par Illumina - principe

- Hybridation d'un brin sur un oligonucléotide attaché à la *FlowCell*
- Un brin complémentaire est synthétisé.
- La molécule d'origine est enlevée et la molécule libre s'hybride *en pont*.
- Un brin supplémentaire est synthétisé de nouveau.
- Les brins complémentaires à la cellule d'origine sont lavés et il ne reste que plusieurs copies d'une même brun (*clusters*)
- Il reste à séquencer les brins présents: Lors de cette étape, le nucléotide incorporé est identifié grâce à un *groupe fluorescent* identifié par laser, permettant d'enregistrer la séquence de manière informatique.
- https://www.youtube.com/watch?v=CZeN-IgjYCo
- https://www.youtube.com/watch?v=WneZp3fSJIk

Figure B: Schéma représentant les *clusters* sur une *FlowCell*

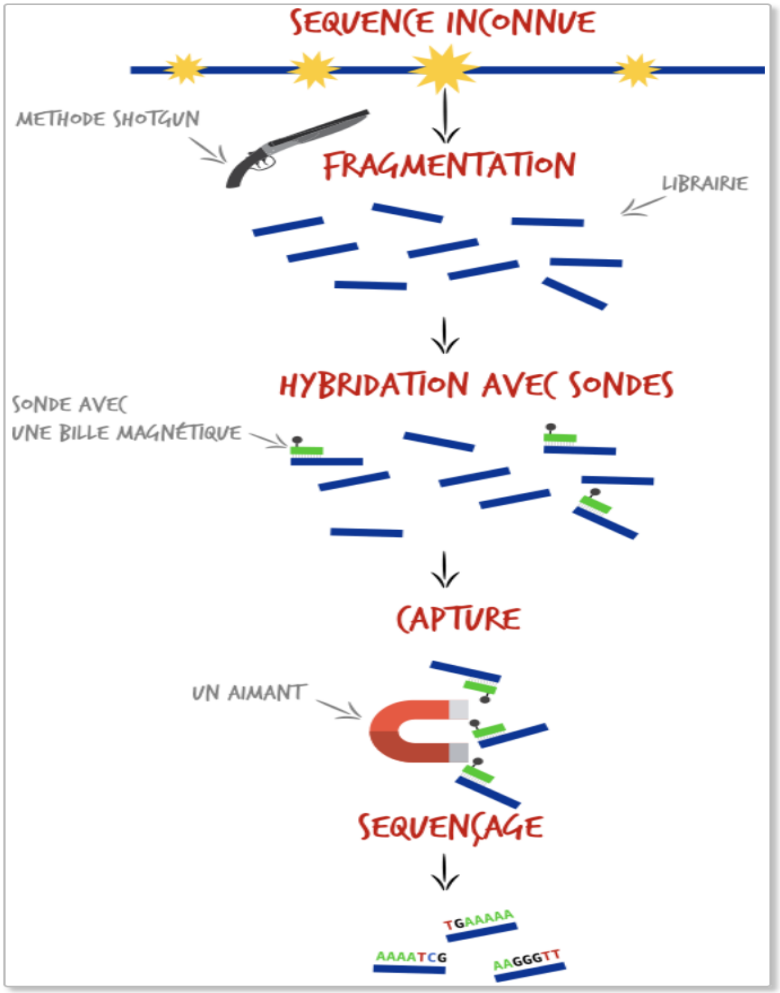
Figure C: Réaction de séquençage

Enrichissement

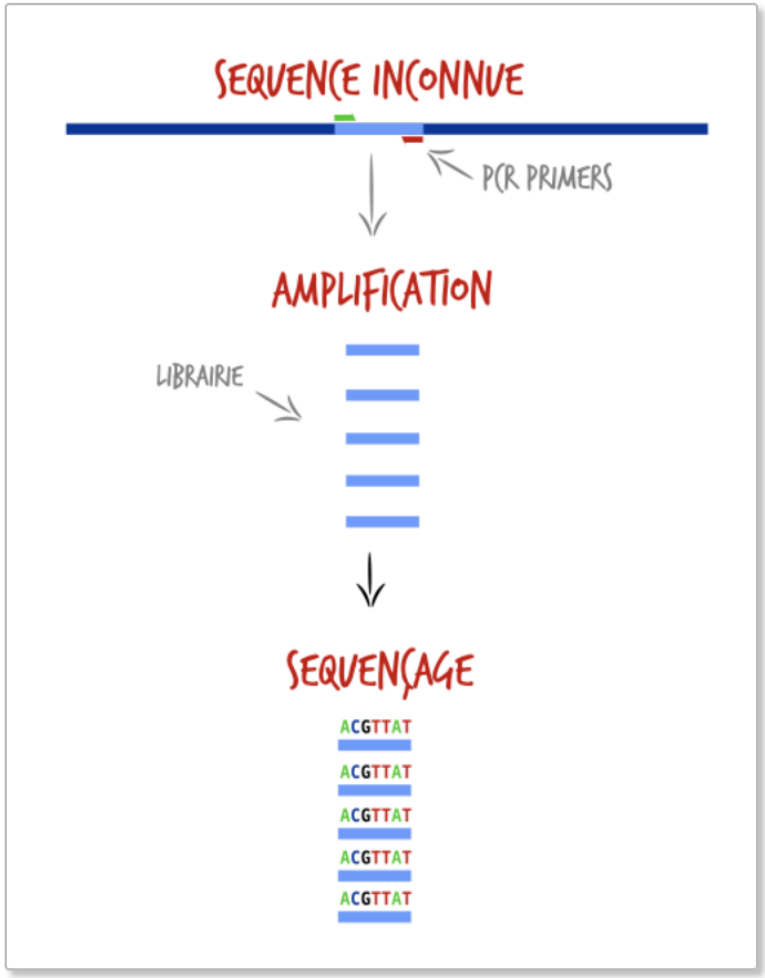
Il est possible de créer une librairie **enrichie en régions d'intérêt**, par exemple pour séquencer uniquement les régions codantes du génome:

- **Par capture**: Il est possible de concevoir des sondes qui vont s'attacher à l'ADN d'intérêt, elle même étant liées à des molécules de *biotines* attachées à des billes magnétiques, qui sont conservées après lavage.
- **par amplicon**: Il est possible de faire une amplification sélectionnée par des amorces PCR choisies.

Séquençage par capture



Séquençage par amplicon



Séquençage d'ARN

Pour le séquençage d'ARN (**RNA-seq**), il existe des étapes supplémentaire à la réalisation de la librairie:

- **Sélection d'ARN non ribosomal** par enrichissement
- **Transformation en ADN**, pour fonctionner avec les séquenceurs Illumina**

Différence Pair-End (PE) et Single-End (SE)

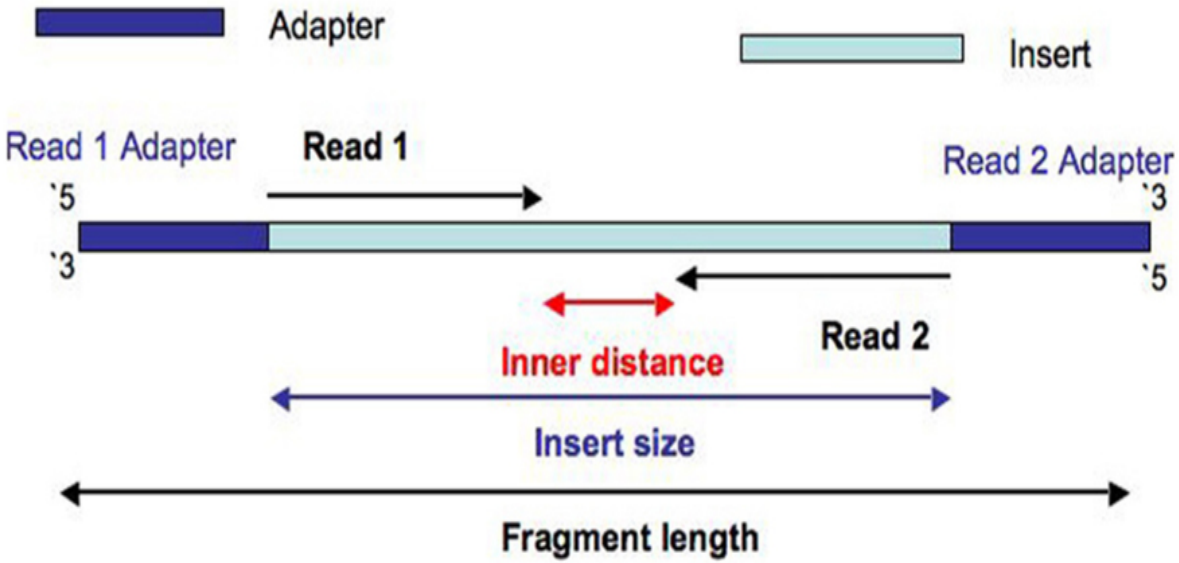
Il est avantageux d'obtenir des fragments de lecture les plus longs possible pour un alignement le plus fiable possible.

- **Lors d'un séquençage simple** (*Single-end*), les brins sont séquencés en partant d'un unique adaptateur. Un *read* correspond donc ensuite à un fragment.
- **Lors d'un séquençage *appairé*** (*Paired-end*), les brins sons séquencés à partir de leur deux extrémités. Les fragment résultats sont appelés **R1** et **R2** et sont liés, qu'ils soient *recouvrant* ou pas.

Un site expliquant cela de manière intéressante:

http://thegenomefactory.blogspot.com/2013/08/paired-end-read-confusion-library.html

Pair-End (PE) et Single-End (SE)



Analyse RNA-seq

Analyse RNA-seq - principe

Analyse de l'expression des gènes = le Transcriptome. C'est une grandeur dynamique.

- Technologie à haut débit précédente: les **Puces à ADN**.
- Technologie basée sur le NGS: Le **RNA-seq**.

Beaucoup d'outils sont communs à ces technologies.

Le NGS appliqué à l'analyse du transcriptome permet:

- Une meilleure concordance entre plateformes
- Une forte sensibilité et meilleure dynamique
- Toutes espèces, toutes régions transcrites
- Une variété d'applications (Analyse diférentielle, analyse de l'épissage alternatif, analyse des gènes de fusion, séquençage *de novo*)

Mais...: Complexité et coût calculatoire accrus = **pipeline bioinformatique plus complexe** par rapport aux microarrays.

L'analyse bioinformatique fait partie intégrante du processus de traitement.

Analyses possibles par le RNA-seq:

- Assemblage de transcriptome *de novo*
- Analyse de l'épissage alternatif
- Découverte des gènes de fusion
- Découverte de nouvelles classes (exemple: sous type moléculaire de tumeurs) (**analyse non supervisée**)
- **Analyse différentielle**: C'est cette application que nous allons décrire ici.

Déroulement d'une analyse RNA-seq différentielle

- **Séquençage**
- **Contrôle qualité**
- **Alignement** sur génome de référence
- **Quantification des valeurs d'expression** (comptage) et normalisation
- **Analyse différentielle** entre conditions expérimentales
- **Visualisation des données** ("diagrammes de chaleur", "volcano plots").
- **Analyse fonctionnelle** des gènes ("Enrichment Analysis")

Pipeline Bioinformatique

- **Contrôle qualité**: FASTQC
- **Trimming**: Trimomatic
- **Alignement**: STAR
- **Quantification des valeurs d'expression** (comptage): featurecount
- **Analyse différentielle** entre conditions expérimentales: EdgeR
- **Visualisation des données**: R+Bioconductor.
- **Analyse fonctionnelle** des gènes: EnrichR, GSEA, …

Départ: les fichiers issus du séquenceur (Fichiers FASTQ)

Ils contiennent les ***reads***: petite séquence d'un fragment d'ADN de longueurs plus ou moins fixe.

- **Single-end** : Chaque read est indépendant et l'ensemble des reads sont stockés dans un fichier *Fastq*.
- **Paired-end**: Dans ce cas, les *reads* sont organisés par paires dans deux fichiers *Fastq*.

```
@HWI-ST865:166:D0C4KACXX:2:1101:1042:1954 1:Y:0:
CNANAAATNAANNNNGNNNNNNNNNANNNNNAAANNNTNNNNNNNNNNTNNTGNNNNTTGTTTNNTTGTGGGTTTCTCTGTCCCCN
+
##########################################################################################
@HWI-ST865:166:D0C4KACXX:2:1101:1241:1970 1:N:0:
CCAGCGACACTTGCAGCTTAGGGGCAAGAGGCTCCCACAACACCCTGTGCGATCGGAAGAGCGGTTCAGCAGGGATGCCGCGGCC
+
GFFIGIIIFGEHHIJJJIIGGGHIIBD=BFG?EDECC@FGCHC?BCCBB)53(;;B;?8299?####################
```

Mesure et encodage qualité: le Phred

Quelques définitions

Qualités:

Qualité exprimée en QPhred

QPhred= probabilité p d'erreur de mauvaise identification de la base

$QPhred=-10\log_{10}p$

Exemple:

Q20 correspond à une probabilité d'erreur de 1%

Q30 correspond à une probabilité d'erreur de 0.1%

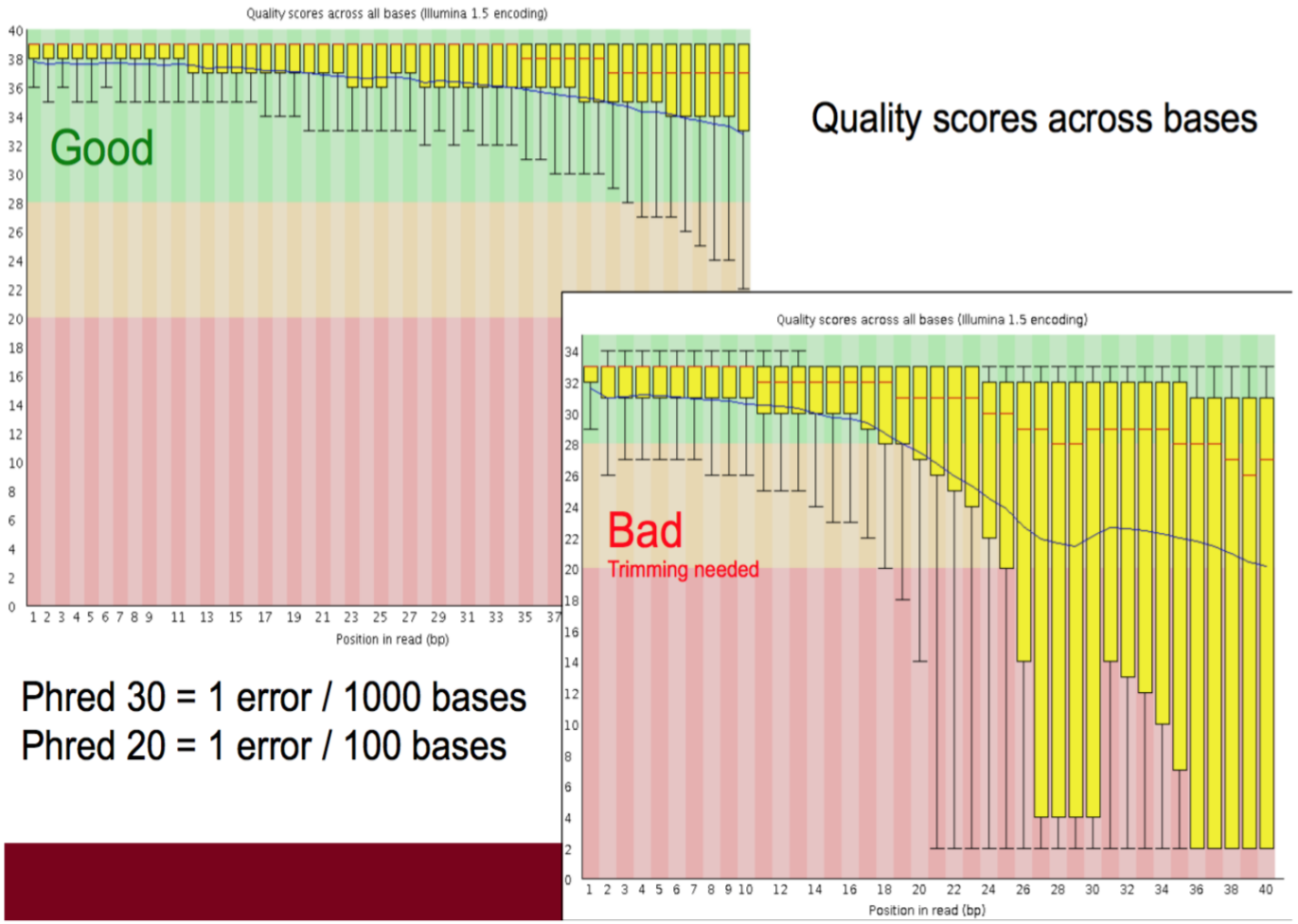
Parenthèse: Infrastructure pour le stockage et le traitement des données

Il est important de stocker les données dans un emplacement fiable, avec beaucoup de mémoire vive et d'espace disponible, et accessible par un réseau rapide (Bref, un **serveur de stockage**).

Pour le traitement des données, un **cluster de calcul** sous Linux est idéal.

Il existe également des logiciels commerciaux sous Windows.

Contrôle Qualité par FastQC



Visualisation par IGV

IGV (*Integrated Genomics Viewer*) est un **explorateur de génome**. Il permet de visualiser de l'information génomique le long du ou des chromosomes de l'organisme que nous sommes en train d'étudier.

- La première étape est de choisir un **génome de référence** (Exemple: hg19 ou mm10).
- Une boite permet de situer **la position actuellement visualisée** sur le chromosome.
- Les informations correspondant à chaque échantillon sont présentées **sous forme de pistes** (*tracks*).
- Les gènes sont **affichés dans leur propre piste** (annotation) avec les exons et introns. Il est possible d'afficher les transcrits individuellement.
- Plusieurs **types de signaux** peuvent être affichés (fragments de lecture, segments CGH, couverture, etc…).

Alignement par STAR: Principe

STAR est un aligneur extrêmement précis qui peut dépasser certains aligneurs par un facteur 50 de vitesse. Par contre, il est très **gourmand en mémoire**.

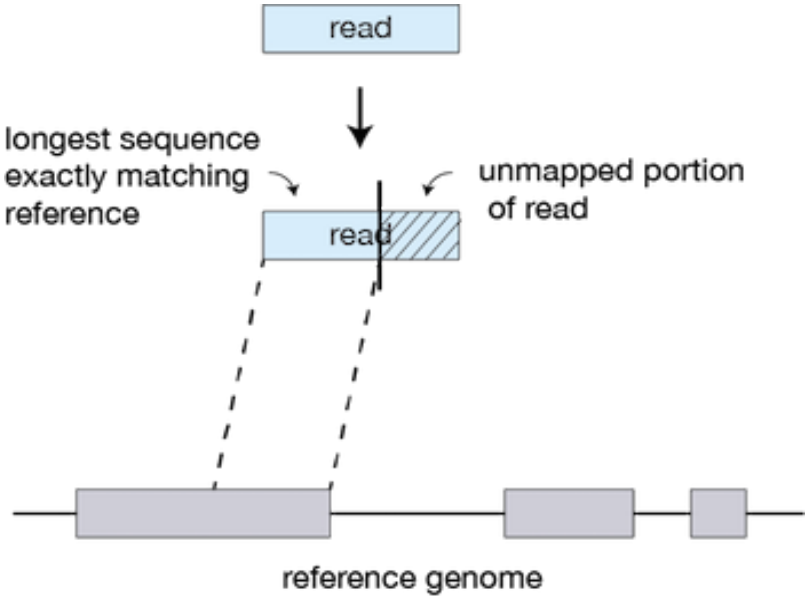
Son principe de fonctionnement est basé sur deux étapes principales:

- **Recherche de graines** (*seeds*)
- **Groupement** (*clustering*), agrafage et calcul de score

Source: https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

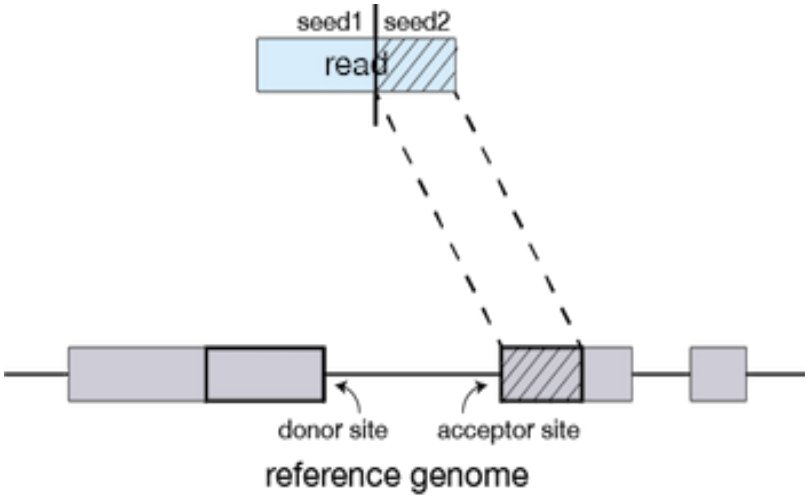
Principe de STAR

Etape 1: Recherche d'alignement **exact maximum** pour chaque read (Maximal Mappable Prefixes - MMP) pour obtenir une liste de graines (*seed1*)



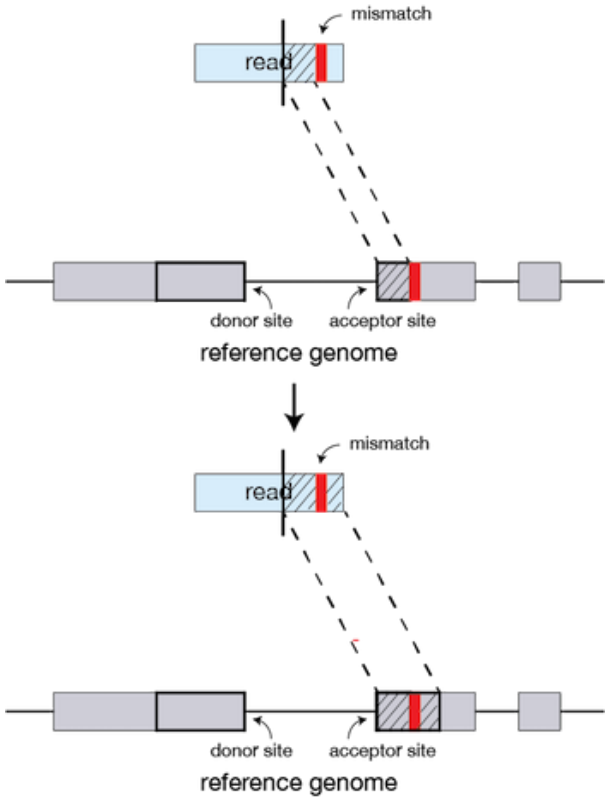
Principe de STAR

Etape 2: STAR va ensuite rechercher un alignement **exact** pour les portions non alignées des graines (appelées *seed2*).



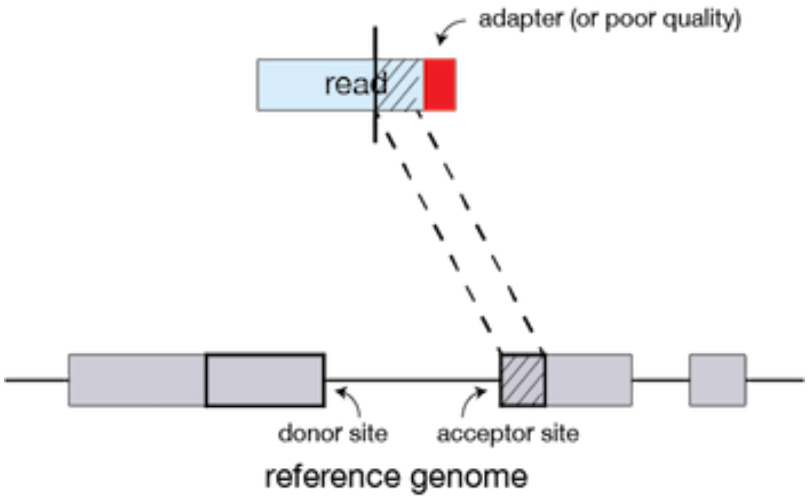
Principe de STAR

Etape 3: Si l'alignement exact n'est pas possible, STAR va **étendre** la séquence recherchée.



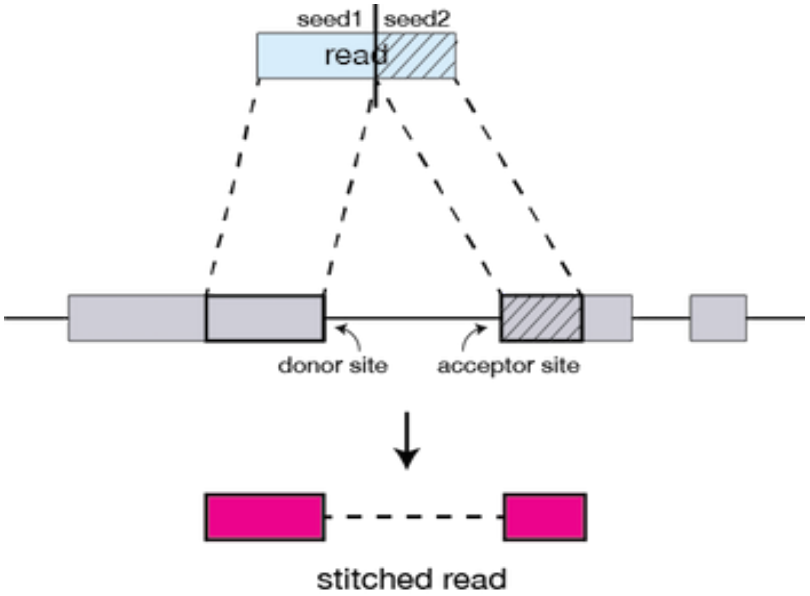
Principe de STAR

Etape 4: Si l'alignement n'est toujours pas possible, STAR va **couper les séquences adaptatrices** ou contaminantes ou de **faible qualité**.

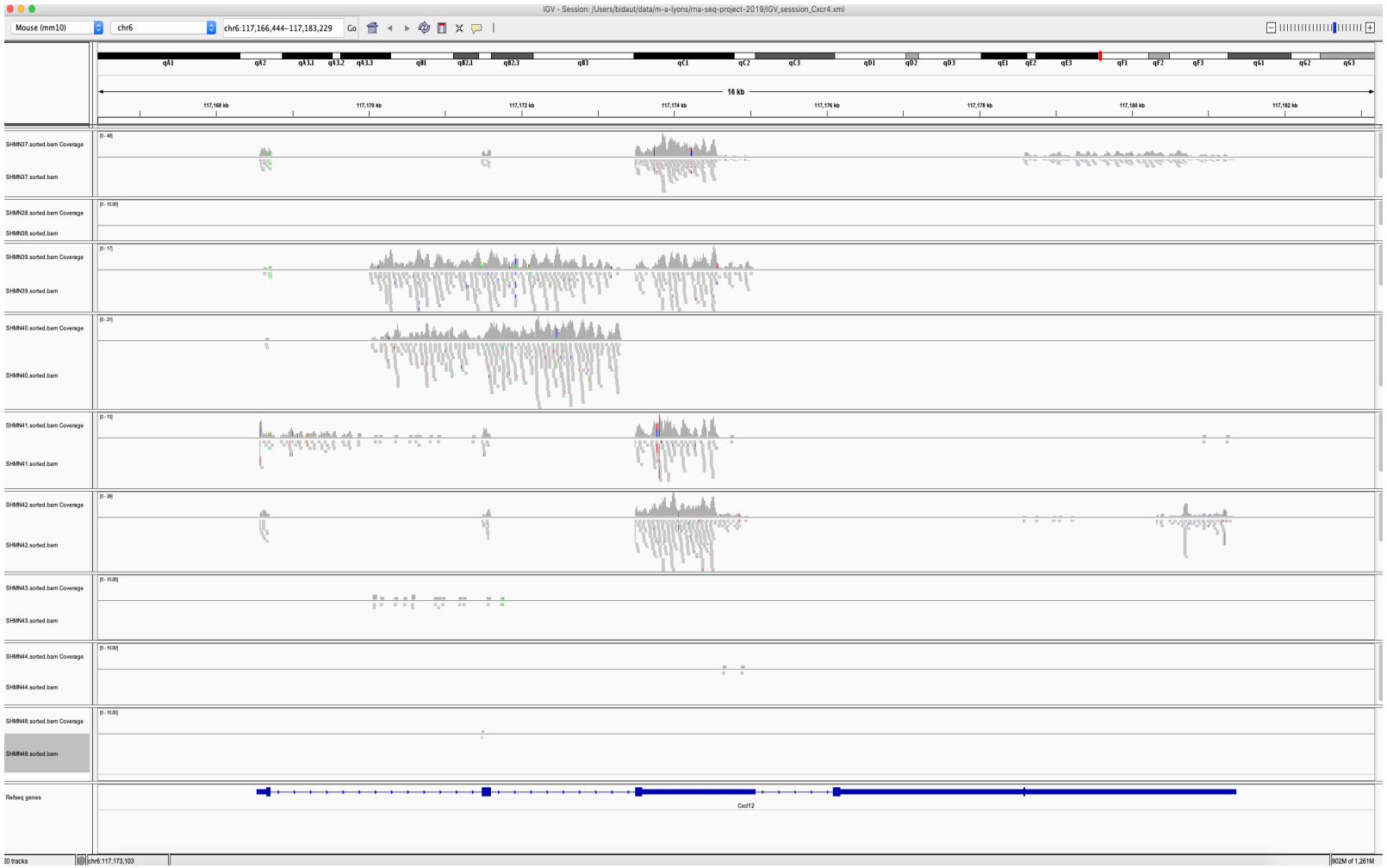


Principe de STAR

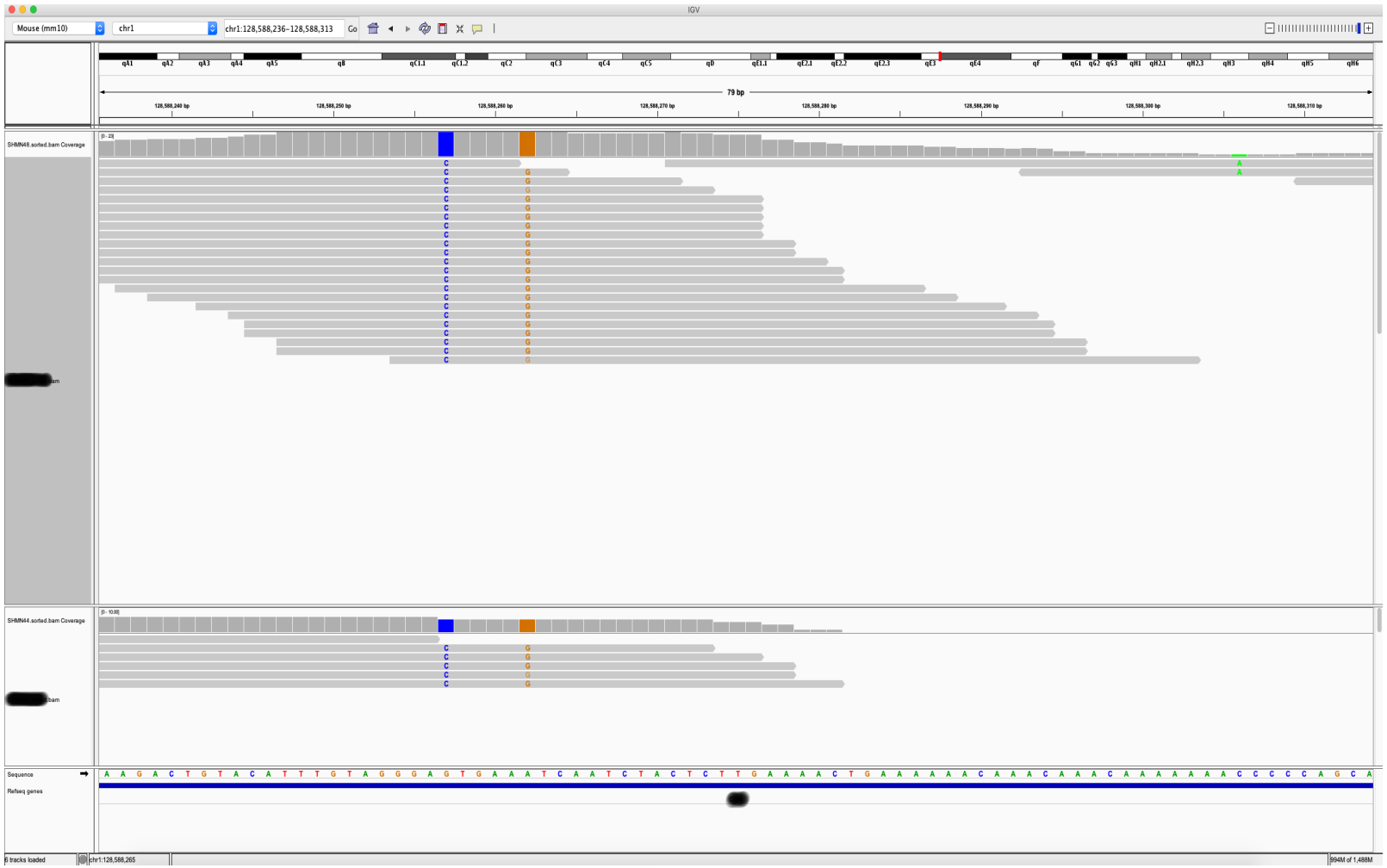
Etape 5: Groupement, Agrafage et Calcul de score. Les graines séparées sont groupées et agrafées sur la base sur leur proximité avec un site donneur pour créer un read complet.



Exemple d'alignement (*Mapping*) RNA-seq sur génome de référence



Analyse de séquence avec IGV



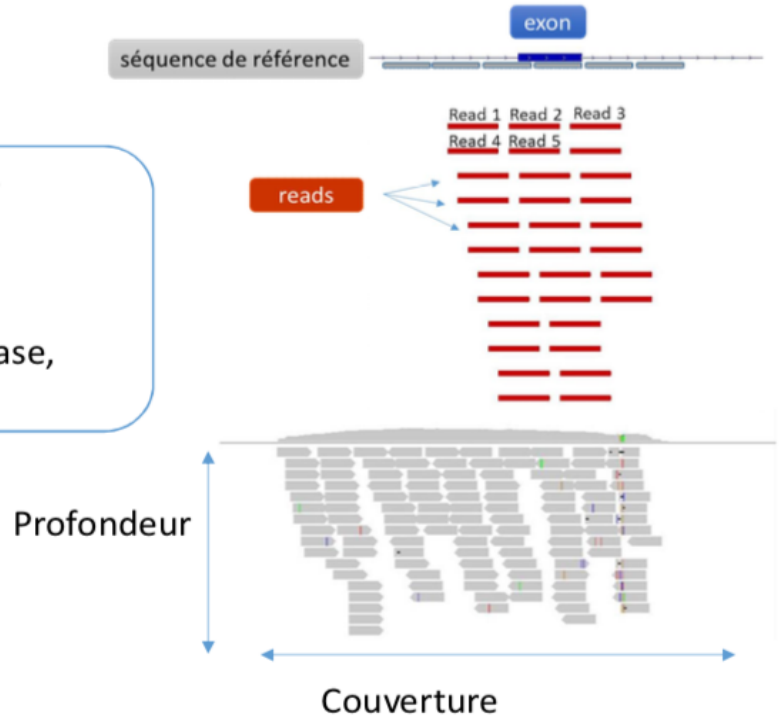
Quelques définitions: Couverture et profondeur

Quelques définitions

Profondeur-Couverture:

Couverture: zone couverte par au moins une lecture, exprimée en %

Profondeur: nombre de lecture de chaque base, exprimée en X



Annotation du génome de référence

Les annotations du génome de référence sont disponibles sous forme de fichiers **GFF/GTF** ou **BED** auprès de Ensembl (BioMart - https://www.ensembl.org/info/data/ftp/index.html) ou NCBI (https://www.ncbi.nlm.nih.gov/refseq/).

```
#!genome-build GRCh38.p13
#!genome-version GRCh38
#!genome-date 2013-12
#!genome-build-accession NCBI:GCA_000001405.28
#!genebuild-last-updated 2019-06
1   havana  gene    11869   14409   .   +   .   gene_id "ENSG00000223972"; gene_version "5"; gene_name "DDX11L1
1   havana  transcript  11869   14409   .   +   .   gene_id "ENSG00000223972"; gene_version "5"; transcript_id
```

Comptage

- Sous l'hypothèse que le nombre de reads venant d'un certain gène est **proportionnel à l'abondance de son ARN dans la cellule**, on compte les reads venant de chaque gène, transcrit ou exon du génome.
- Il est possible de faire un script 'maison' mais il existe maintenant un grand nombre de programmes pour faire cette fonction, notamment **featureCount** de la suite logicielle **SubRead**.
- Les localisations génomiques des objets génomiques (*genome features*) sont données en entrée du programme de comptage (Fichier d'annotation **GTF/GFF**), permettant d'assigner les comptages à chaque transcrit, gene, ou exon.

Comptage par featureCounts

featureCounts permet de compter les fragments de lecture correspondant à chaque gène et d'obtenir un tableau récapitulatif pour l'ensemble de nos échantillons.

featureCounts du package **subRead** permet de faire ce comptage à partir des fichiers BAM résultants de l'alignement de séquences par **STAR** et un fichier d'annotation de génome au format GFF.

Il est possible de travailler en mode PE (Paired-End) ou SE (Single-End). Il n'est pas obligatoire de trier les BAM avant de les donner à **featureCounts**.

Assignation des reads

Les fragments de lecture sont assignés à des attributs (*feature*) ou des meta-attributs (*meta-features*). Il y a donc 2 modes.

- Les attributs (*features*) correspondent à une zone contiguë de l'ADN, par exemple un exon.
- Les meta-attributs (*meta-features*) correspondent à un objet biologique résultant de l'assemblage de plusieurs attributs (les gènes)

featureCounts peut compter au niveau **des exons** ou **des gènes**. Il est recommandé d'utiliser des **identifiants uniques** pour les meta-attributs, comme par exemple les *geneID* du NCBI, et d'éviter les noms de gènes.

Alignements multiples et recouvrants

Pour les **alignements multiples**, **featureCounts** peut les traiter de 3 manières:

- Les ignorer (**Défaut pour la version *bash***)
- Les compter plusieurs fois
- Les compter plusieurs fois mais avec une contribution calculée en fraction ($1/n$, n étant le nombre de fois ou le read apparaît).

Pour les **alignements recouvrants** plusieurs attributs, **featureCounts** peut les traiter de 2 manières:

- Les ignorer (**Recommandé pour le RNA-seq**)
- Les compter pour chaque attribut

Obtention du comptage pour les transcrits

featureCounts permet de compter les reads au niveau des exons (*features*) ou des gènes (*meta-features*). Par contre, Il n'est pas possible d'obtenir les reads pour les transcrits, car il ne peut déterminer à quel transcrit les assigner.

Ce type d'analyse est considérablement plus complexe. Pour ce faire, il faut utiliser un pipeline spécifique basé sur l'aligneur **Salmon**, qui permet d'assigner les reads aux transcrits sur une base de probabilité.

Voir la documentation pour l'aligneur **Salmon**: https://salmon.readthedocs.io/en/latest/salmon.html.

Normalisation RPKM

Nous avons **différentes tailles** de libraires, des biais de séquences dus à la PCR, ou un contenu RNA qui peut différer en échantillons. Il est donc nécessaire d'appliquer une **normalisation pour rendre les échantillons comparables**.

Méthode la plus utilisée: **RPKM: *Reads Per Kilobase per Million***. Soir une valeur de comptage $read(G)$ pouor un gène:

- On divise la taille de la librairie par 10^6. Cela nous donne le **facteur par million**.
- On divise la valeur de comptage $read(G)$ par ce facteur pour obtenir **des reads per millions (RPM).**
- On divise la valeur RPM par la longueur du gène en *Kilobases*, pour obtenir des **RPKM**.

$$RPKM(G) = \frac{[\sum read(G)]}{[\sum Read].longueur(G)} . 10^6 . 10^3$$

Variante: **FPKM** pour le Paired-end. Dans ce cas, deux reads peuvent appartenir au même fragment et ne sont pas comptés 2 fois.

Limitations de la Normalisation RPKM

On a également un problème de normalisation liée à la taille du gène.

Exemple de comptage RPKM pour deux échantillons différents:

- Transcrit **20kb**, comptage **400**, taille de librairie **20Millions** de reads:

$$RPKM(t) = (400/20)/20 = 1$$

- Transcrit **0.5kb**, comptage **10**, taille de librairie **20Millions** de reads:

$$RPKM(t) = (10/20)/0.5 = 1$$

Pas de détection en expression différentielle si on utilise cette transformation !.

De plus, elle est **inadaptée à la comparaison entre échantillons** car on ne peut pas comparer des pourcentages sur des librairies de taille différentes!

Normalisation TPM

Méthode **TPM:** ***Transcripts Per Million***

C'est une variante sur la méthode RPKM mais l'ordre des opérations change:

- On divise la valeur de comptage $read(G)$ par la longueur du gène en *Kilobases*, pour obtenir des **reads per kilobase (RPK)**.
- On somme les **RPK** pour l'échantillon et on divise par 10^6. Cela nous donne le **facteur par million**.
- On divise la valeur RPK par ce facteur pour obtenir **des reads per millions (RPM)**.

$$RPK = \frac{\sum read(G)}{longueur(G)}.10^3$$ (***Read Per Kilobase***)

$$TPM = \frac{RPK(G)}{\sum RPK}.10^6$$

Plus "acceptable" pour la comparaison d'échantillons car la somme totale des TPM est identique pour tous les échantillons.

Effet des gènes très ou peu exprimés

Il faut voir une expérience RNA-seq comme un **échantillonnage de l'espace des transcrits** par la librairie.

Une expérimentation RNA-seq donne donc la **proportion** des transcrits pour une **taille de librairie donnée**.

Si un gène est fortement exprimé dans un échantillon et peu exprimé dans l'autre, on aura une **augmentation artificielle** de l'expression de l'ensemble des gènes (!).

Gène	Longueur	Count Ech1	Count Ech2	RPKM Ech1	RPKM Ech2
Taille librairie		20	20		
Gene 1	2	8	0	0.2	0
Gene 2	2	6	10	0.15	0.25
Gene 3	2	6	10	0.15	0.25

Normalisation TMM

TMM - Trimmed mean of M-values. C'est l'une des **normalisations recommandées** qui permet d'éviter l'effet des gènes fortement ou faiblement exprimés (Robinson & Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* volume 11, Article number: R25, 2010).

Au lieu de faire une mise à l'échelle propre à une librairie, un facteur de normalisation global est calculé en assumant que la majorité des gènes n'est pas différentiellement exprimée et en **ne tenant pas compte des valeurs extrêmes de comptages**.

Ce facteur est calculé à partir de l'**abondance relative** de l'expression de chaque échantillon, soit **un fold change global** (Ce que l'on appelle les *M-Values*).

Un échantillon est choisi comme référence. L'abondance relative moyenne est calculée en "trimmant" les distributions de logFC (d'où le *Trimmmed Mean*), ce qui donne un facteur correctif pour chaque échantillon.

Normalisation TMM

Il est calculable par le package d'analyse **edgeR** sous **R/Bioconductor**.

Autre normalisation: celle proposée par **DESeq2**.

Dillies *et al*. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform . 2013 Nov;14(6):671-83.

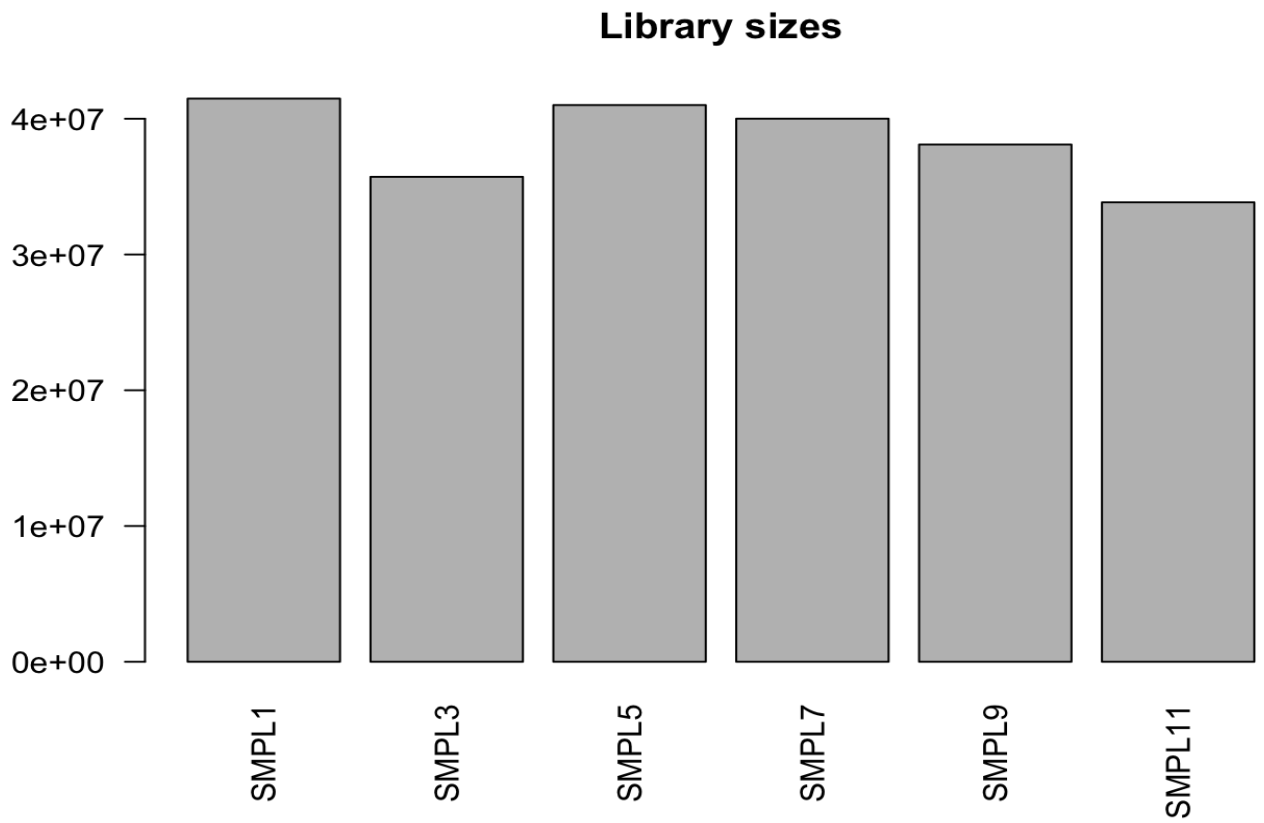
Normalisation (Suite)

La Normalisation type TMM ou autre n'est nécessaire **que pour la visualisation** ou **l'exploration globale** des données de comptage (par exemple sous forme de **carte de chaleur** (*heatmap*)).

Les logiciels d'analyse différentielles (**edgeR** et **DESeq2**) ont leur propre normalisation *intégrée* à la méthode et il n'y a **pas besoin d'appliquer de normalisation** aux comptages **avant** un appel à l'une de ces méthodes.

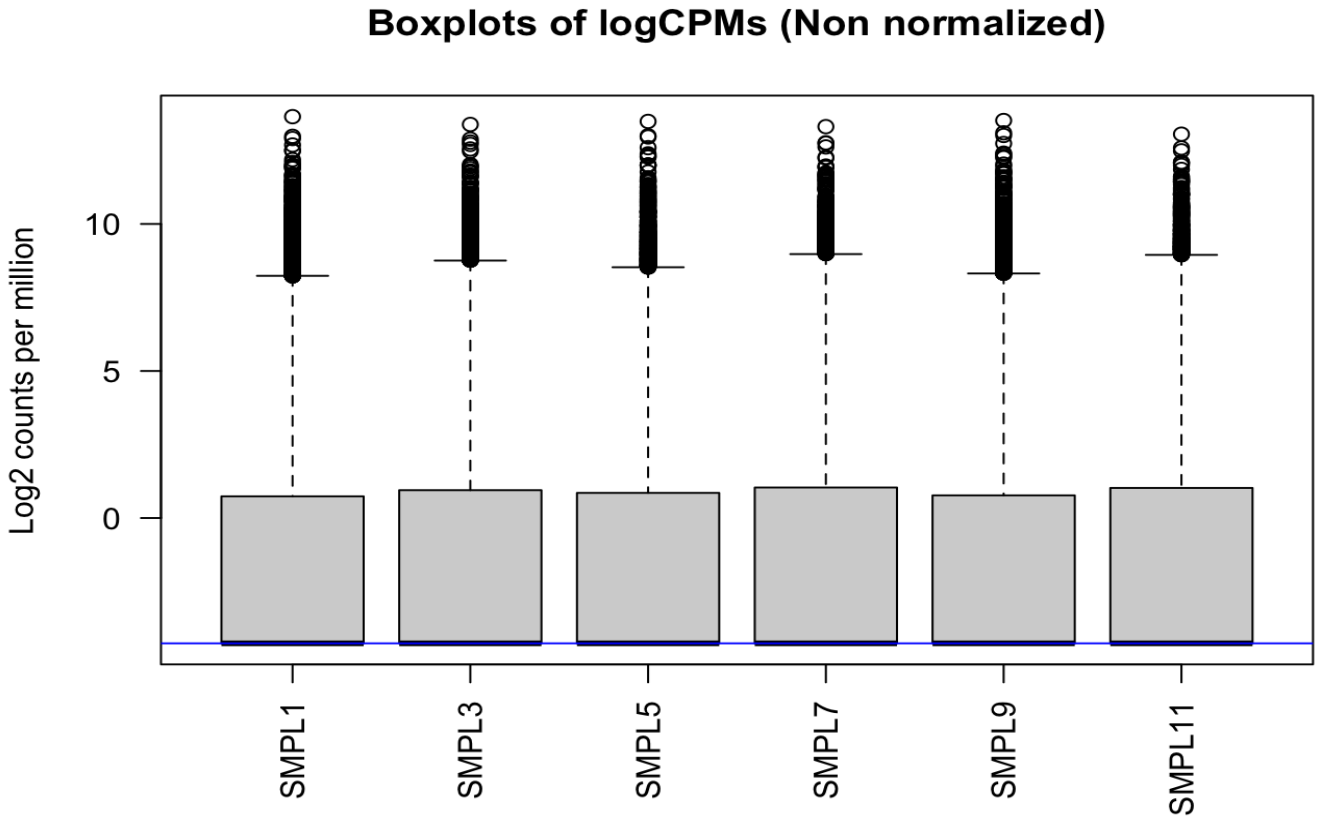
QC (Quality Control)

Comparaison de la taille de librairies



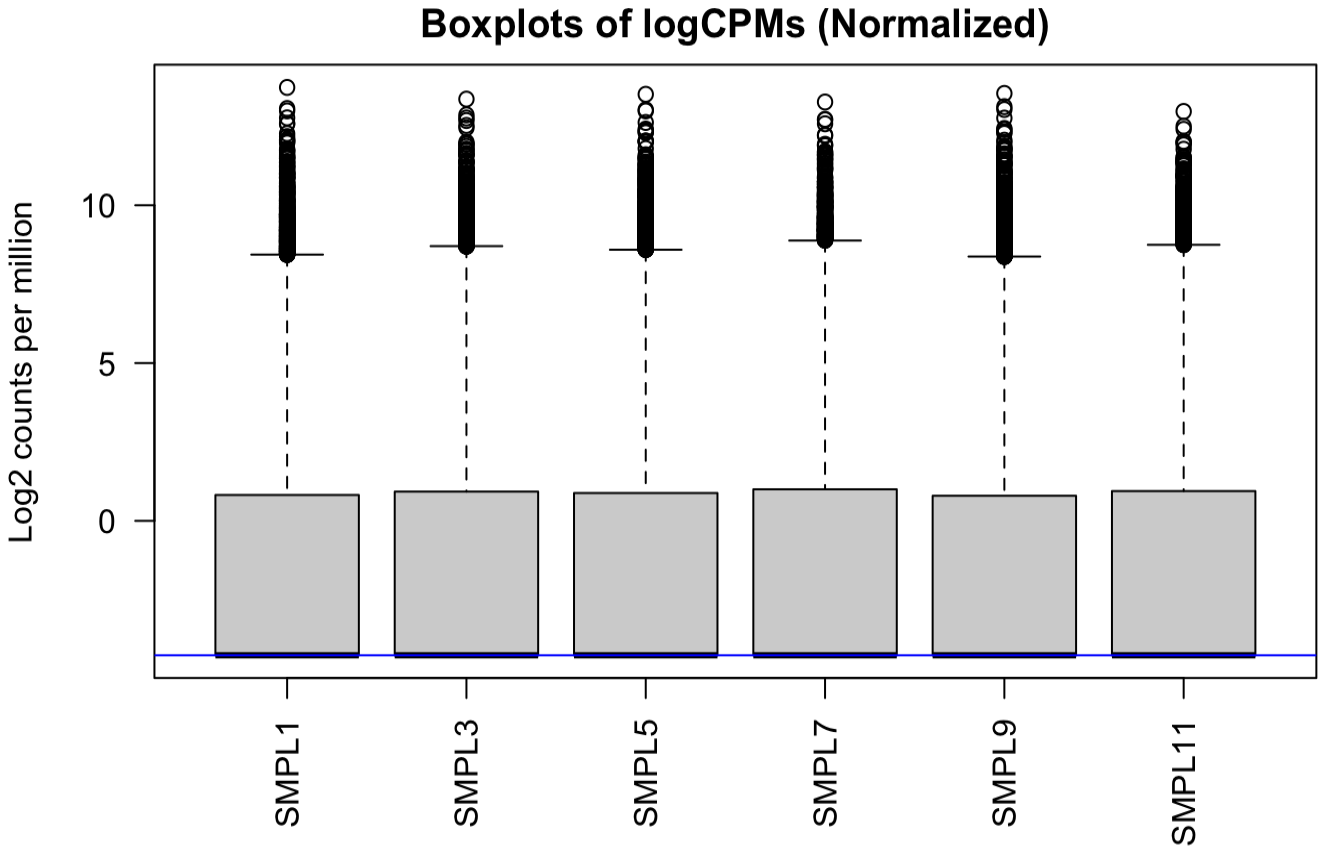
QC (Quality Control)

Comparaison des distributions avant normalisation TMM



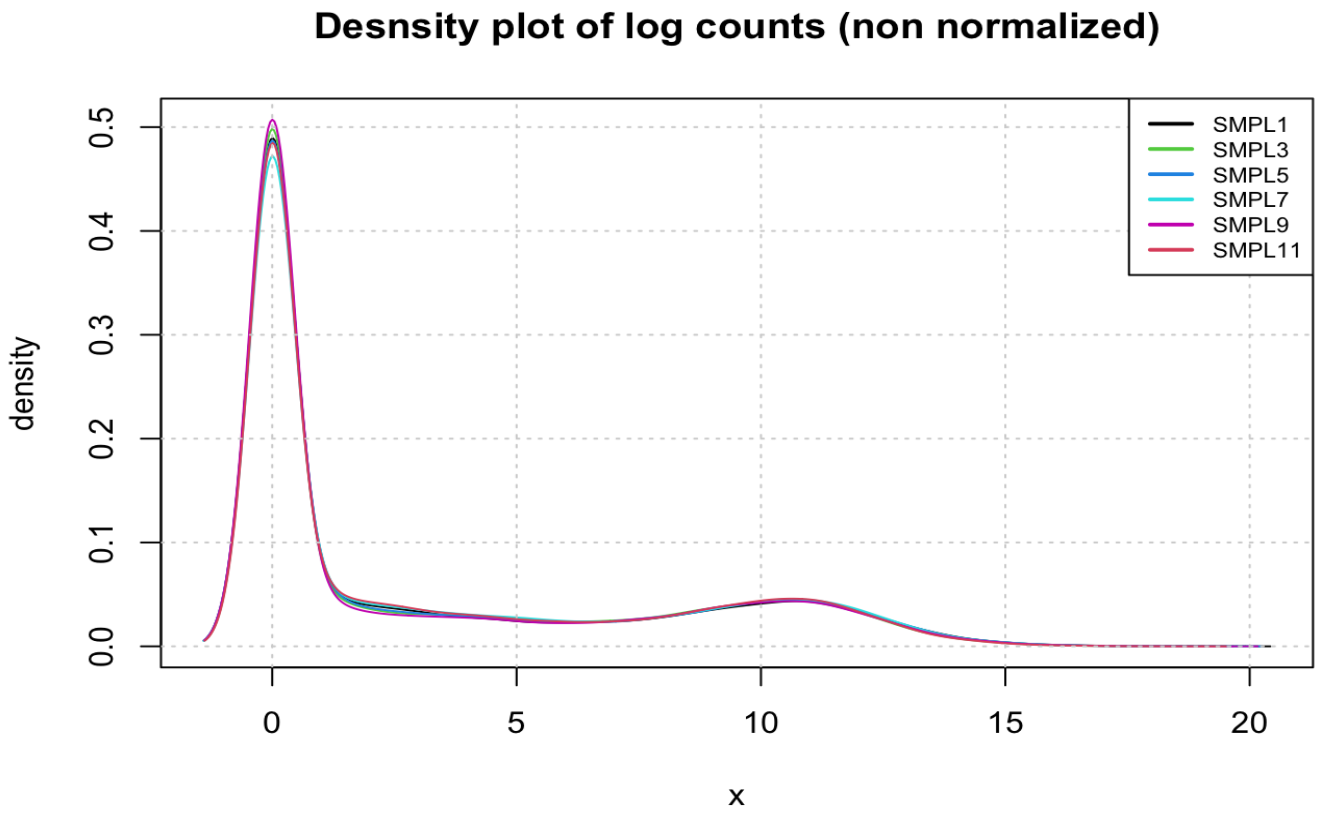
QC (Quality Control)

Comparaison des distributions après normalisation TMM



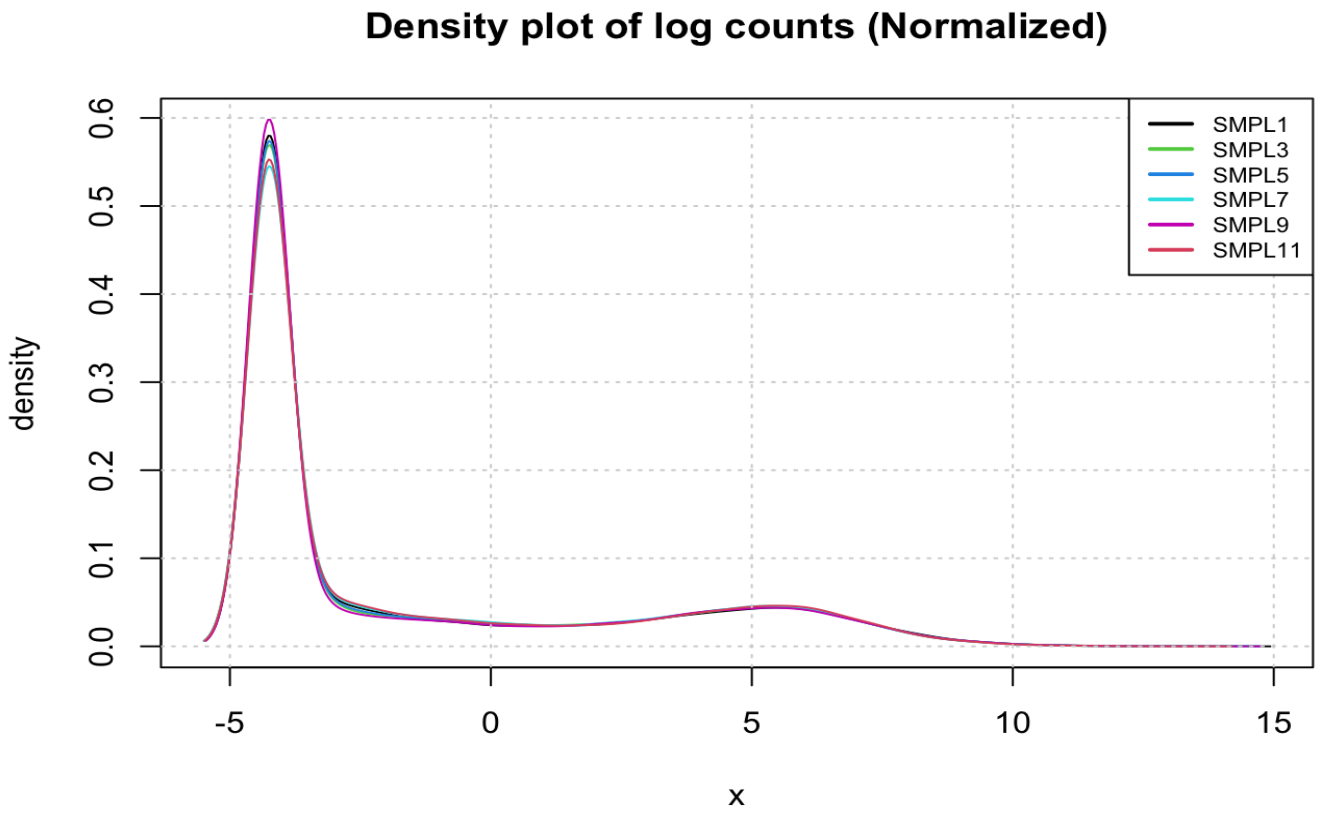
QC (Quality Control)

Comparaison des distributions avant normalisation TMM



QC (Quality Control)

Comparaison des distributions après normalisation TMM



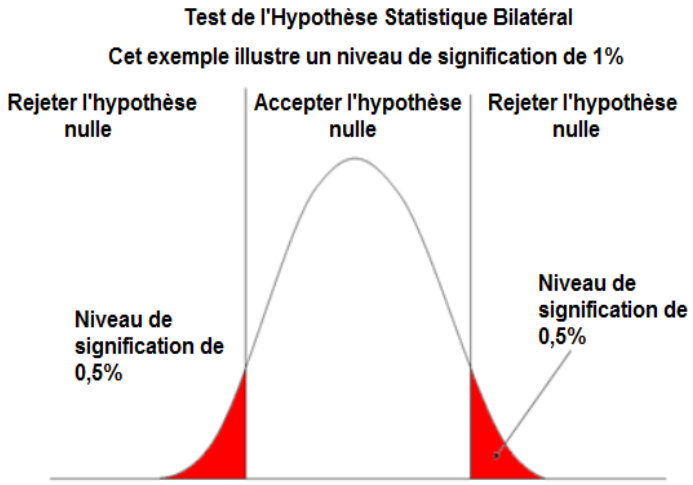
Analyse Différentielle

Analyse différentielle

L'objectif est d'établir quels sont les gènes **différentiellement exprimés** entre plusieurs conditions expérimentales, par exemple, un *contrôle* et un *traitement*. C'est une analyse **supervisée**.

Pour cela, un **test statistique** est utilisé. Rappel: un test statistique est la vérification d'une hypothèse nulle *H0*.

Exemple d'une Hypothèse nulle: Les niveaux d'expression d'un gène en *contrôle* et *traitement* sont égaux. **p-valeur**: probabilité d'obtenir la même valeur du test ou plus extrême si l'hypothèse nulle était vraie.



Analyse différentielle

Exemple d'un design expérimental:

Sample	Batch	Drug	CellType
SMPL1	RNA1	TREATM	Type1
SMPL3	RNA1	CTRL	Type1
SMPL5	RNA2	TREATM	Type1
SMPL7	RNA2	CTRL	Type1
SMPL9	RNA3	TREATM	Type1
SMPL11	RNA3	CTRL	Type1

Dans cet exemple, nous cherchons à établir la liste des gènes différentiellement exprimés entre un contrôle (non traité) et des cellules traitées.

Package Bioconductor edgeR

edgeR est un logiciel spécialement conçu pour l'analyse différentielle d'expression à partir de données obtenues par RNA-seq.

edgeR implémente de nouvelles méthodes statistiques basées sur les distributions binomiales négatives.

Cela le rend adapté aux analyses

- avec de nombreux facteur expérimentaux
- avec de petites tailles d'échantillons.

Robinson, MD. *et al* . (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, 26 (1) 139-140.

Modèles Linéaire Généralisés (GLM)

Objectif: Modéliser l'expression de chaque gène comme une combinaison linéaire de facteurs multiples. Cela permet de tenir **compte de plusieurs facteurs**, tels que le **groupe expérimental**, le **lot (effet *batch*)**, etc…

$$y = a + b.\,group + c.\,patient + d.\,lot + e$$

- y = Expression du gène.
- a, b, c et d = paramètres du modèle à estimer
- a = intercept (valeur quand tous les paramètres sont à un niveau de référence)
- e = terme d'erreur

edgeR modèle les données en utilisant une **distribution binomiale négative**.

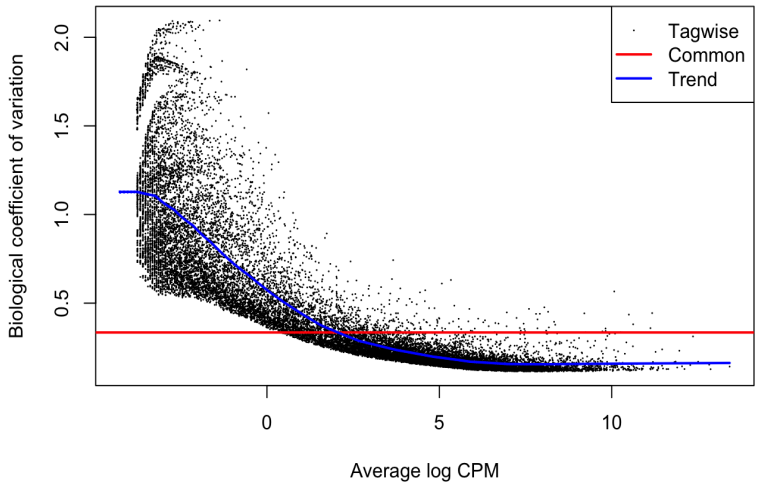
- Comparaison entre 2 groupes sur un facteur unique: **test Simple**
- Comparaisons avec plusieurs facteurs : **Utilisation des GLMs**.

Dispersion

Avant une analyse différentielle, il est important d'estimer les niveaux de variation d'expression des transcrits intra-groupes.

- $Dispersion = BCV^2$
- BCV= *Biological Coefficient of Variation*

Exemple: si l'expression d'un transcrit change de 20% entre réplicats, son BCV est de 0.2 et sa dispersion de 0.4.



Analyse edgeR avec statistiques

Show 10 entries Search:

entrezgene	mgi_symbol	logFC	PValue
271375	Cd200r2	-21.4666085589919	0
68279	Mcoln2	-19.9269428303882	0
74525	Fam234b	-18.1660878970867	1.2352e-319
102294	Cyp4v3	-20.7416046890166	6.43179499960022e-304
108116	Slco3a1	-19.952714528279	1.16133784773855e-289
259277	Klk8	-20.3541976141754	5.73237567946313e-281
215707	Ccdc92	-20.149936347831	1.98964624903405e-280
14168	Fgf13	-20.9523613291614	2.17269496085313e-279
16431	Itm2a	-20.177328304306	2.79298795922538e-269
58233	Dnaja4	-21.0072843699612	4.39665510248675e-264

Showing 1 to 10 of 1,000 entries Previous 1 2 3 4 5 … 100 Next

Correction pour tests multiples

Problème des tests multiples: Dans le cas d'un très grand nombre de tests, le nombre de faux positifs peut devenir très grand.

Il est donc nécessaire d'ajuster l'ensemble des p-valeurs par **une correction pour tests multiples**.

- **Correction de Bonferroni**: on divise les p-valeurs obtenues par le nombre de tests (très stringeant)
- **Benjamini-Hotchberg**: On contrôle la proportion attendue de faux positifs parmis les positifs: (*False Discovery Rate*)

Methode Benjamini-Hotchberg

L'idée est de contrôler la proportion de faux positifs parmis les positifs.

Pour cela, on transforme les p-valeurs de cette manière:

D'abord on tri les p-valeurs par valeur croissante.

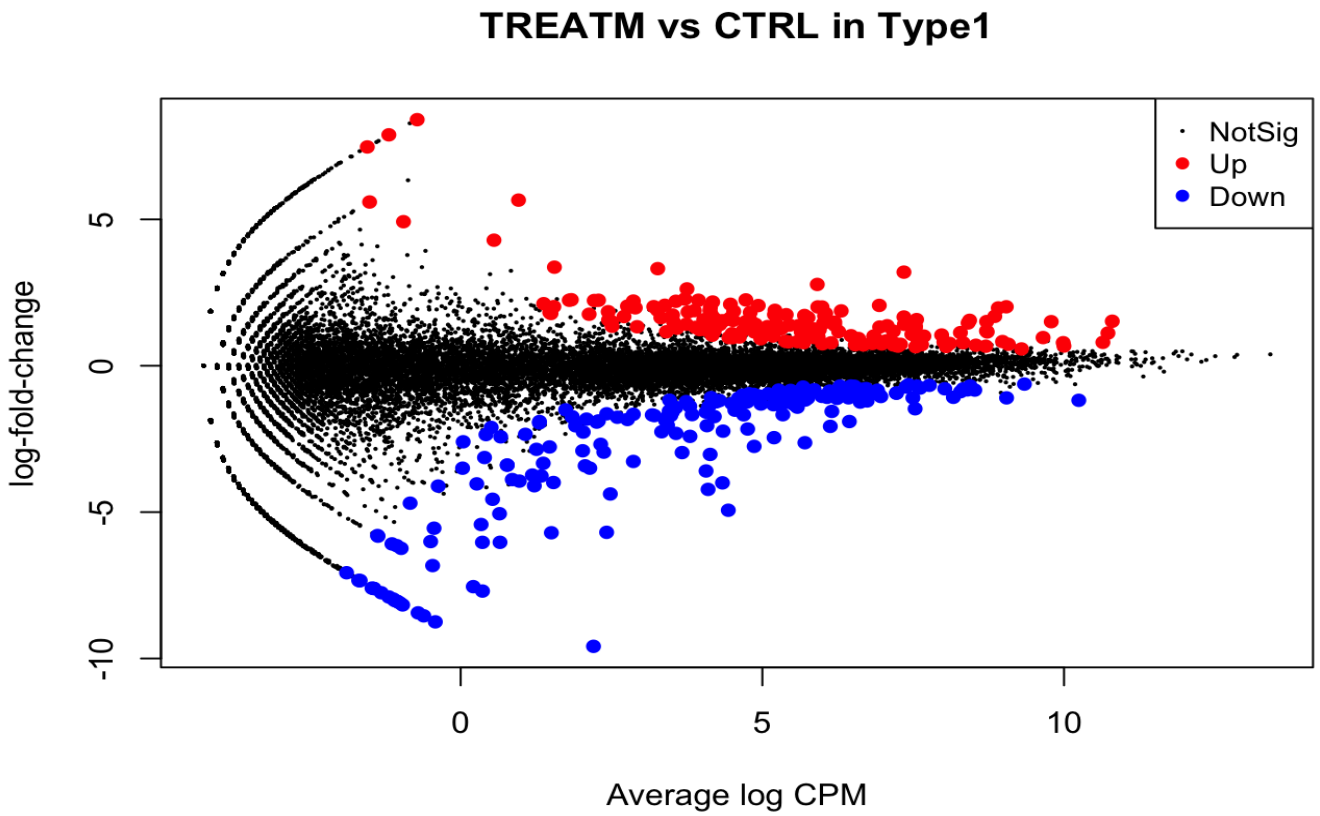
Ensuite, on applique une transformation comme suit:

- La p-valeur la plus large n'est pas corrigée
- la seconde $p = (p.\, n).\, (n - 1)$
- la troisième $p = (p.\, n).\, (n - 2)$ ...
- la plus petite: $p = (p.\, n).\, (n - n + 1) = p.\, n$

La p-valeur ajustée est le FDR.

MA-Plot

MA-Plot: log(Ratio) vs average



Diagrammes thermiques

- Appelé également **Heatmap**
- Normalisation par **ligne** ou **globale** pour l'affichage (**Standardisation**)
- Représentation d'une matrice de **manière graphique**, chaque valeur étant représentée par une **cellule colorée**.
- Les colonnes ou les lignes peuvent être ordonnée de plusieurs façons, soit manuellement, soit par une méthode de type **clustering**.

Visualisation des résultats sous forme de diagramme thermique

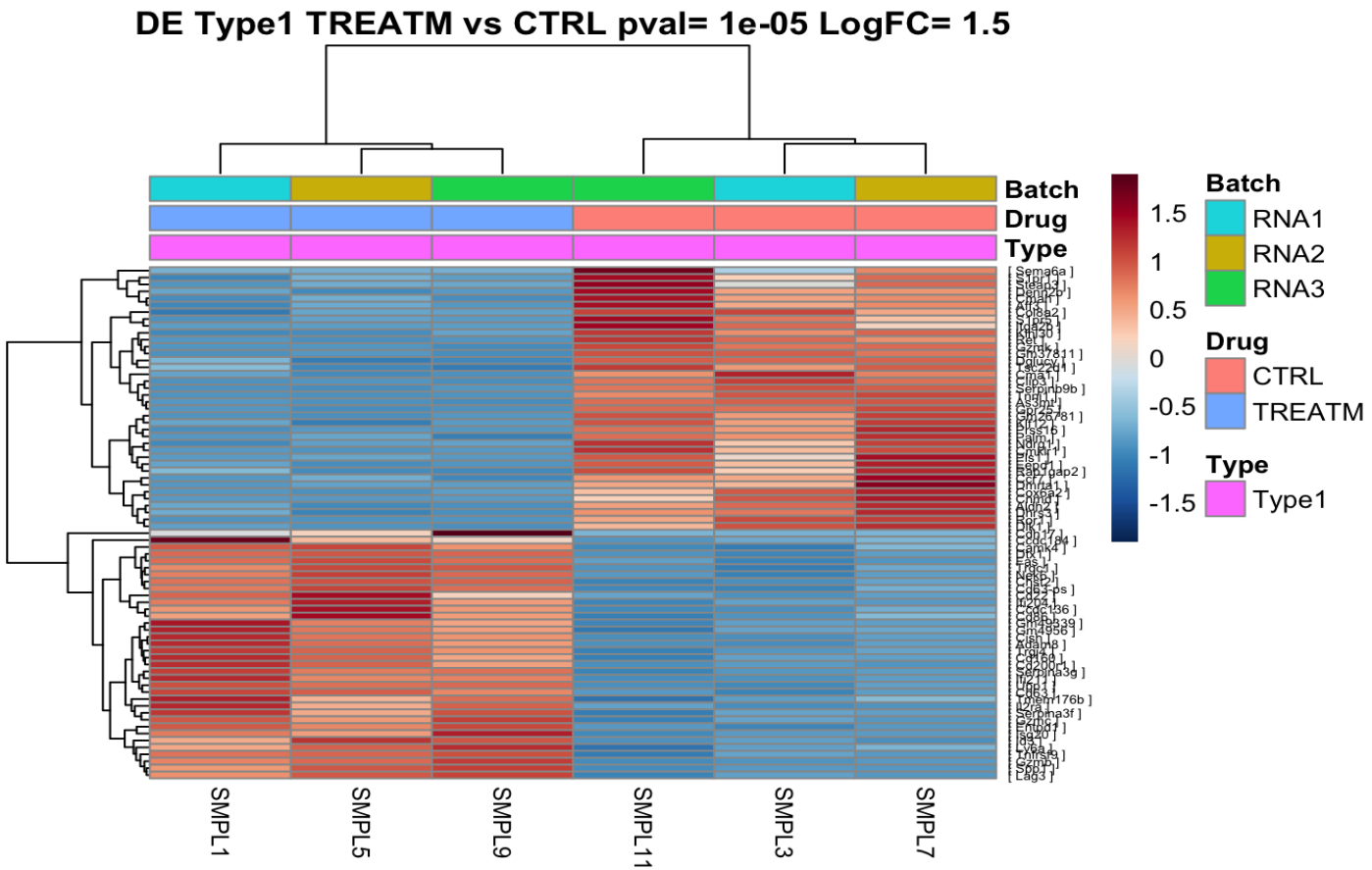
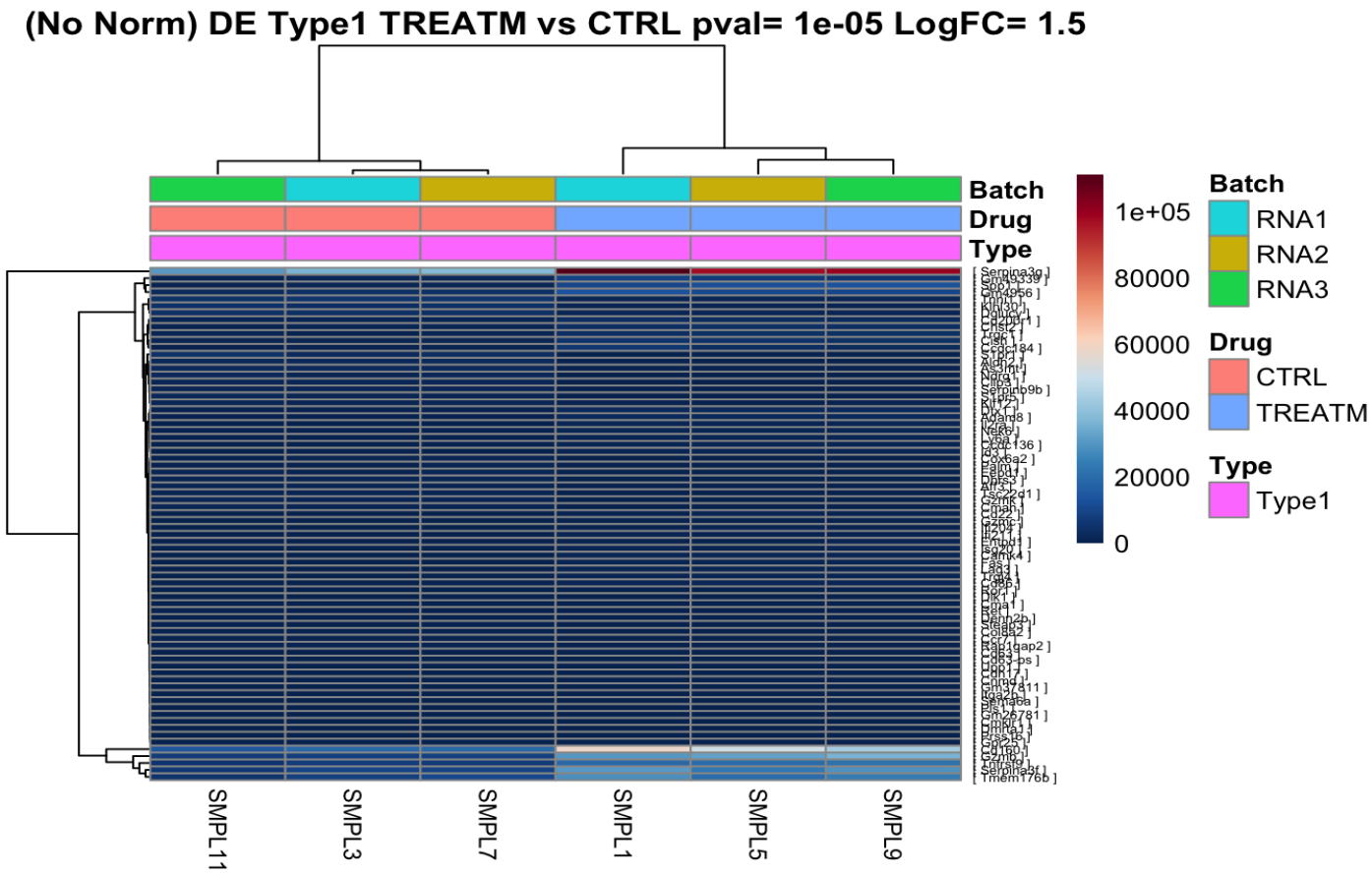


Diagramme thermique (Sans normalisation d'affichage)



Diagrammes en volcan (*volcano plots*)

Lorsque l'on utilise un programme de modélisation statistique pour extraire les gènes différentiellement exprimés, on va s'intéresser **prioritairement aux gènes ayant la p-valeur la plus basse**. Or, il s'avère que des gènes faiblement différentiellement exprimés peuvent être associée à une faible p-valeur si leur expression est caractérisée par une faible variance intra groupes.

Il est donc important de considérer **à la fois** les gènes ayant une **faible p-value** et pour lesquelles on obtient un effet biologique intéressant, c'est à dire un **logFC élevé**.

Un diagramme en volcan est un nuage de point combinant l'effet statistique sur l'axe des ordonnées (p-valeur) et l'effet biologique (logFC) sur l'axe des abscisses.

Il est souvent représenté avec les seuils sur le **logFC** et de **p-valeur** utilisés pour sélectionner les gènes.

Diagramme en volcan

- Seuil utilisé: $LogFC = 1.5, pval = 10^{-3}$

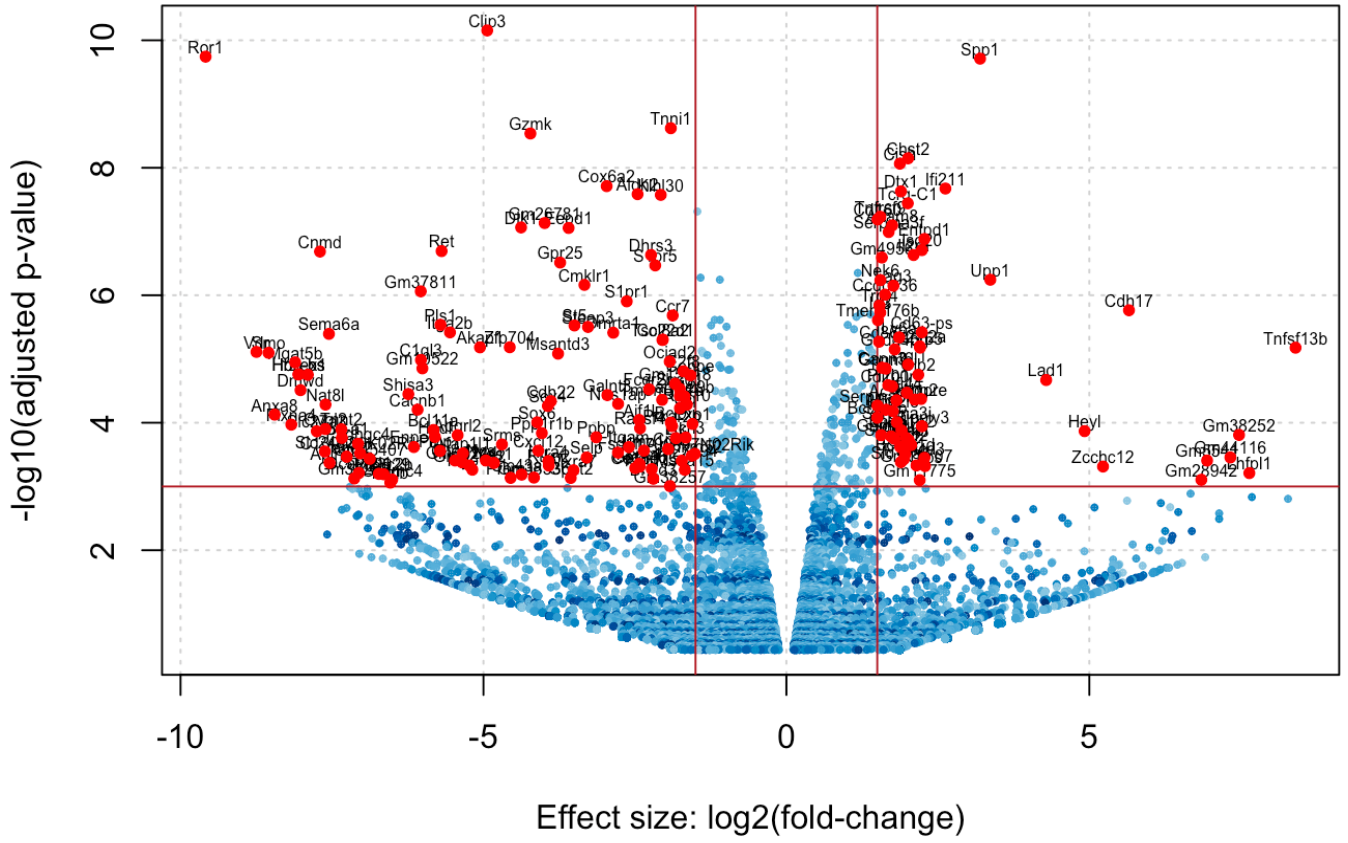
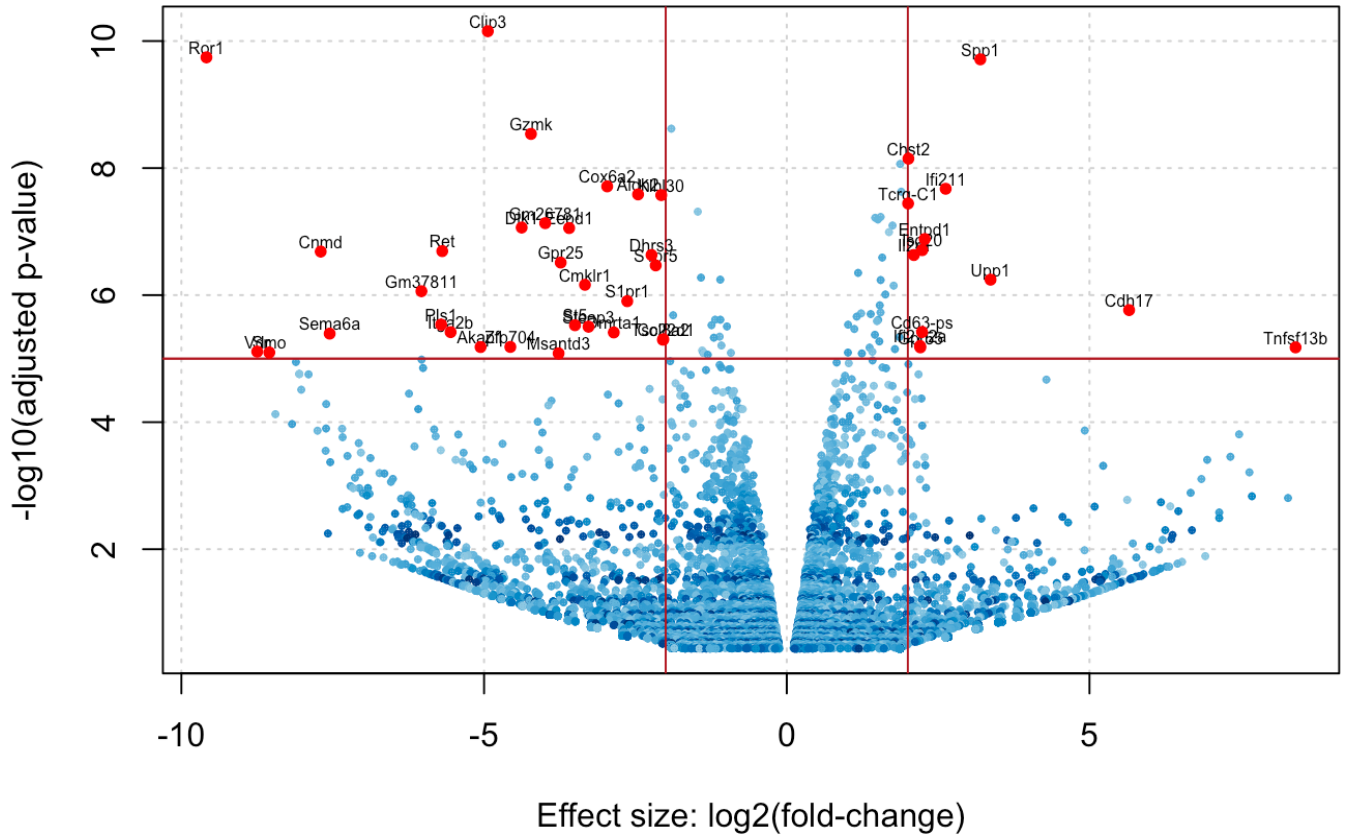


Diagramme en volcan

- Seuil utilisé: $LogFC = 2.0, pval = 10^{-5}$



Rappels sur les extensions de fichiers

- Fichiers de séquences brutes: `.Fastq` (Compressé: `.Fastq.gz`)
- Fichiers de séquences alignées `.bam`
- Index de fichiers de séquences alignées: `.bai`
- Génome complet au format FASTA: `.fa`
- Fichier d'expression: Délimité par des tabulations: `.txt`, `.tdf` ou `.csv`.

Rappel des étapes bioinformatiques

- Contrôle Qualité (**FASTQC**)
- Trimming (**Trimmomatic**)
- Alignement sur le génome de référence (**TopHat** (Spécifique RNA-seq), **SubRead** (Générique), **STAR** (Spécifique RNA-seq))

  ou

- Alignement sur transcriptome de référence(**Salmon**)
- Comptage (**FeatureCount**)
- Analyse différentielle (**edgeR** ou **DESeq**)
- Visualisation des données (**Morpheus**, **Bioconductor**)
- Analyse des données (**Gene Set Enrichment Analysis**)

Point statistiques importants pour une analyse différentielle

- **Il faut des réplicats biologiques (!)**
- La normalisation est nécessaire pour comparer l'expression entre échantillons
  - Différentes tailles de librairies
  - Biais de séquences
- Les programmes d'analyse d'expression différentielles (**edgeR**, **DESeq2**) ont besoin des **comptages bruts** (**Pas les RPKM** ou autre)
- Il faut corriger **pour les tests multiples**.

Rappel sur les différentes normalisations

- Il existe une normalisation sur les données analysées (les *comptages*), appelée RPKM, TMM, ou autre…
- Il existe également une “normalisation” qui consiste à faire une standardisation de l’*affichage* des heatmaps (affichage centré-réduit par ligne sous forme de *z-score*).

Petit Quizz

- En analyse NGS, la bioinformatique et l'infrastructure bioinformatique ne sont que peu ou pas importantes face au séquençage lui-même: (**Vrai/Faux**)
- **Remettre dans l'ordre les phases d'analyses RNA-seq suivantes**:
  - Alignement
  - Comptage des *reads*
  - Visualisation
  - Analyse Différentielle
  - Contrôle Qualité
- **Faire correspondre les logiciels suivants à ces différentes étapes**:
  - STAR
  - IGV
  - FeatureCounts
  - FastQC
  - edgeR

Enrichissement fonctionnel

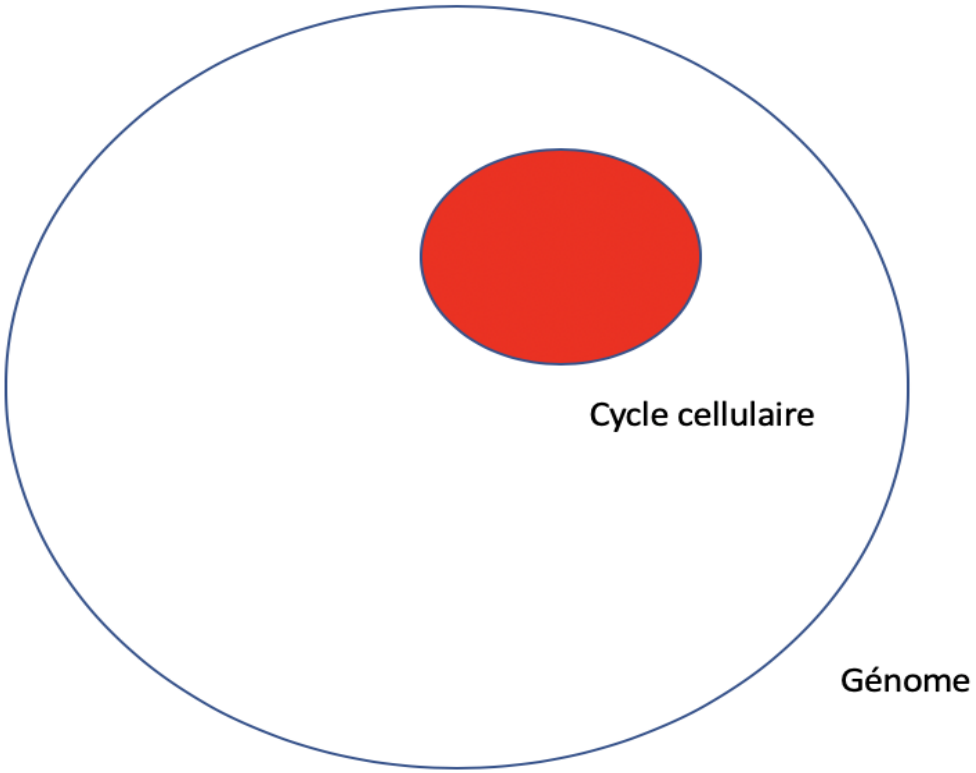
Analyse fonctionnelle

Après avoir identifié une liste de gènes, nous cherchons à obtenir la fonction biologique des gènes présents dans cette liste.

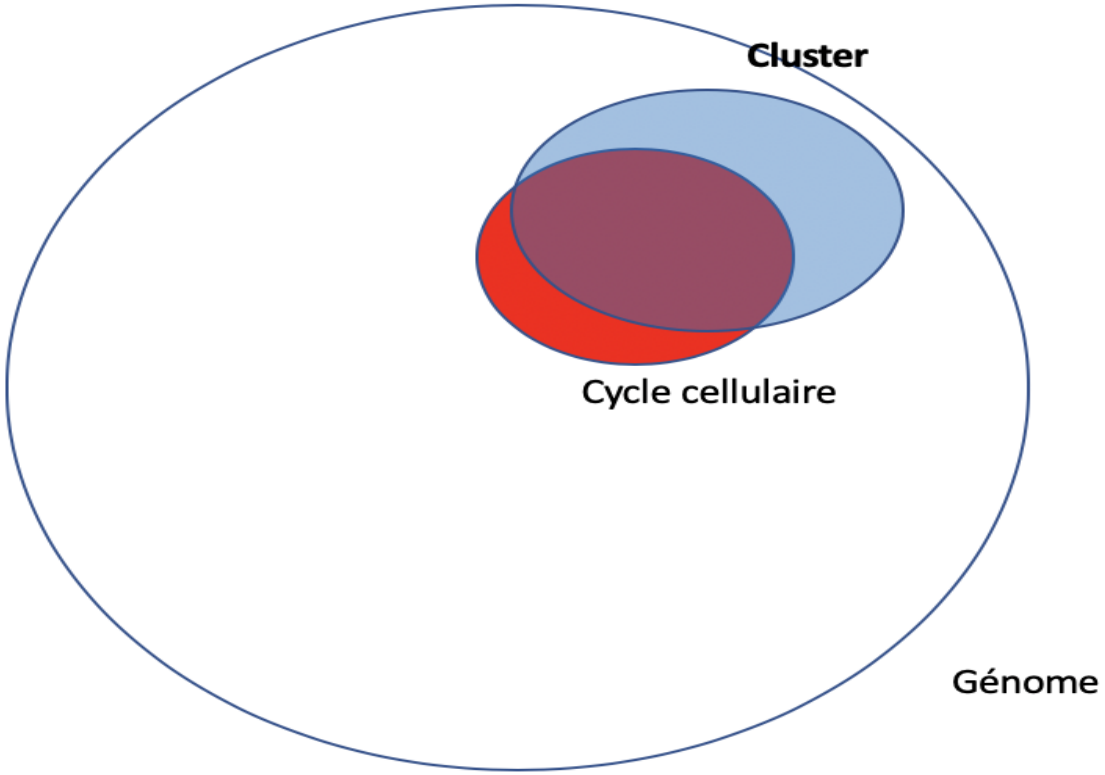
Deux manières de procéder:

- Parcourir la liste et faire une recherche dans la littérature…
- Utiliser les annotations des gènes pour trouver les fonctions moléculaires ou pathways représentés par ces gènes: -> Faire une **Analyse d'enrichissement fonctionnelle**.

Analyse fonctionnelle par enrichissement: principe (1)



Analyse fonctionnelle par enrichissement: principe (2)



Analyse fonctionnelle par enrichissement GO

Elles sont basées sur deux composantes:

- Elles utilisent une **Ontologie** (vocabulaire **contrôlé et stable** mis en place par le ***Gene Ontology Consortium*** ou autre (***Kyoto Encyclopedia of Genes and Genomes, KEGG***)).
- D'où l'utilisation fréquente du raccourci (Enrichissement **GO**)
- Elles sont basées sur un **Enrichissement Fonctionnel** associé à une **validation statistique par Test Hypergéométrique**.

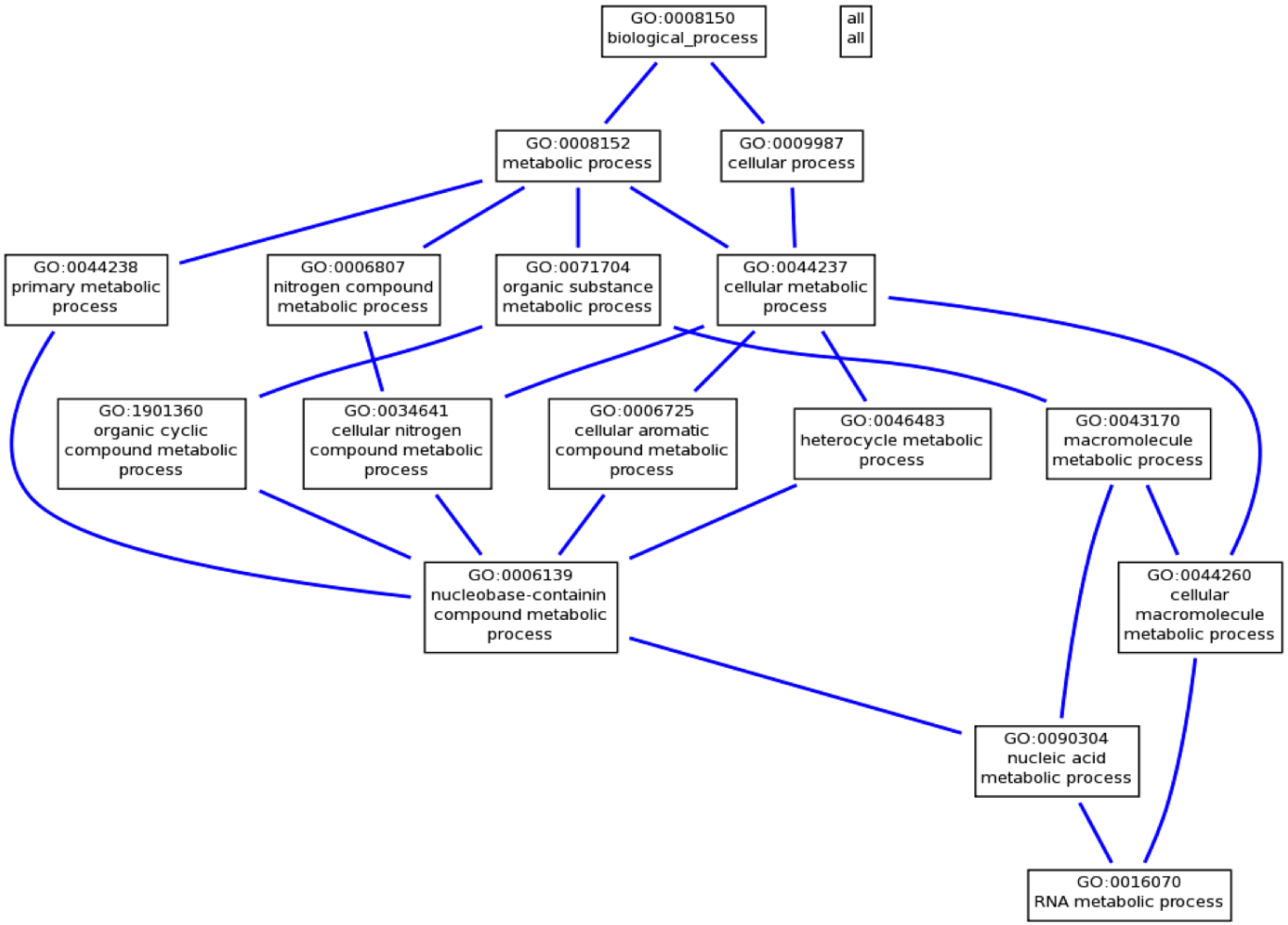
Qu'est ce qu'une ontologie ?

Une ontologie est l'**ensemble structuré des termes et concepts représentant le sens d'un champ d'informations**, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances.

Application au génome: **Gene Ontology** (*Gene Ontology Consortium*http://www.amigo.org). 3 ontologies ont été définies.

- **Biological Process**
- **Cellular Component**
- **Molecular Function**

Graphe GO


GO:0008150
biological_process
all
all
GO:0008152
metabolic process
GO:0009987
cellular process
GO:0044238
primary metabolic
process
GO:0006807
nitrogen compound
metabolic process
GO:0071704
organic substance
metabolic process
GO:0044237
cellular metabolic
process
GO:1901360
organic cyclic
compound metabolic
process
GO:0034641
cellular nitrogen
compound metabolic
process
GO:0006725
cellular aromatic
compound metabolic
process
GO:0046483
heterocycle metabolic
process
GO:0043170
macromolecule
metabolic process
GO:0006139
nucleobase-containin
compound metabolic
process
GO:0044260
cellular
macromolecule
metabolic process
GO:0090304
nucleic acid
metabolic process
GO:0016070
RNA metabolic process

Exemple d'annotation

RPL35A

- GO:0000049 - tRNA binding
- GO:0000184 - nuclear-transcribed mRNA catabolic process, nonsense-mediated decay
- GO:0003735 - structural constituent of ribosome
- GO:0005829 – cytosol
- GO:0006364 - rRNA processing
- …

Annotations par un **vocabulaire contrôlé**.

Test d'enrichissement GO

Une catégorie de gènes regroupe n gènes sur le total de N présents sur la puce. La fréquence de départ de cette catégorie est $F = n/N$.

Ayant obtenus k gènes significativement exprimés ou sous-exprimés, dont p appartiennent à la catégorie C, la fréquence de la catégorie C dans ces gènes est $f = k/p$.

L'enrichissement est défini comme f/F.

Le test d'enrichissement doit répondre à la question: L'enrichissement est-il **statistiquement significatif** par rapport à un tirage au hasard? On le fait par **Test hypergéométrique**.

Exemple de résultat

Term	Ont	N	Up	Down	P.Up	P.Down
peroxisome proliferator activated receptor signaling pathway	BP	13	3	1	0.000236579799839291	0.216785532737776
negative regulation of appetite by leptin-mediated signaling pathway	BP	4	2	0	0.000552257854781615	1
regulation of appetite	BP	20	3	1	0.000897500375237737	0.313423278407892
aggrephagy	BP	5	2	0	0.000914593320360469	1
positive regulation of peroxisome proliferator activated receptor signaling pathway	BP	5	2	1	0.000914593320360469	0.0896690865222862
ribosomal protein import into nucleus	BP	6	2	0	0.00136319724960807	1
leptin-mediated signaling pathway	BP	8	2	0	0.00251252533538245	1
regulation of peroxisome proliferator activated receptor signaling pathway	BP	8	2	1	0.00251252533538245	0.1395805393931
regulation of endopeptidase activity	BP	302	9	4	0.00264226454547343	0.817937433495033
cartilage development	BP	147	6	2	0.00303840822963679	0.762621255839882

Showing 1 to 10 of 30 entries

Previous 1 2 3 Next

Licence