SouthGreen bioinformatics platform

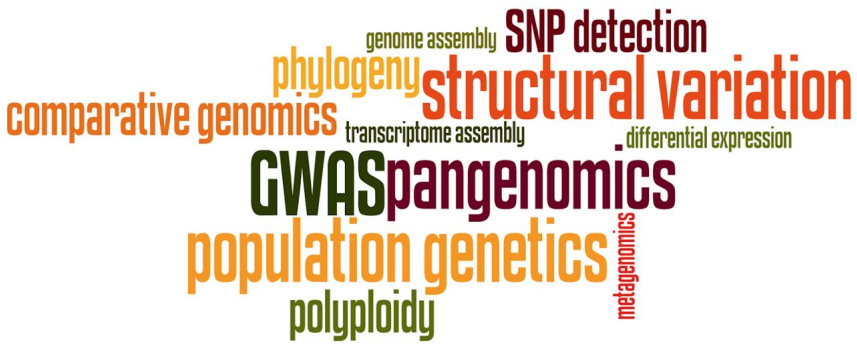
Trainings 2019

agap Bioversity International UMR BGPI Biologie et Génétique des Interactions Plante-Parasite diade IPME

ABiMS ASTRE Territoires - Risques - Ecosystèmes BOREA Biologie des Organismes et Ecosystèmes Aquatiques intertryp Host-vector-parasite-environment interactions in neglected tropical diseases due to trypanosomatids MiVEGEC Maladies Infectieuses et Vecteurs Ecologie, Génétique, Evolution et Contrôle UMR QualiSud

SouthGreen bioinformatics platform

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




Rice

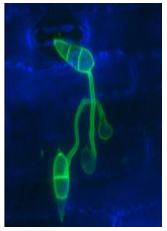

Banana


Palm


Sorghum


Coffee


Cassava


Magnaporthe

www.southgreen.fr

South Green bioinformatics platform

Charte

Larmande Pierre
Sabot François
Tando Ndomassi
Tranchant-Dubreuil Christine

Comte Aurore
Dereeper Alexis

Orjuela-Bouniol Julie

Bocs Stephanie
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

Mahé Frédéric
Ravel Sébastien

Sempere Guilhem

SouthGreen bioinformatics platform

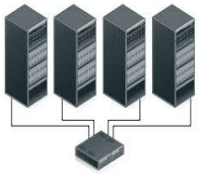
Workflow manager

TOGGLE
Toolbox for generic NGS analyses

SNAKEMAKE

Galaxy

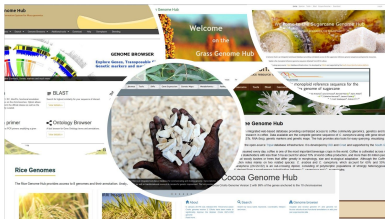
HPC and trainings….



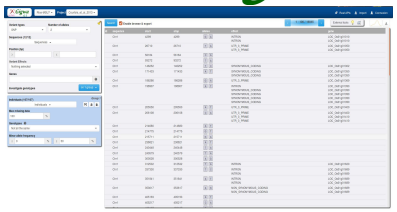
IRD Institut de Recherche pour le Développement FRANCE

cirad

Genome Hubs & Information System

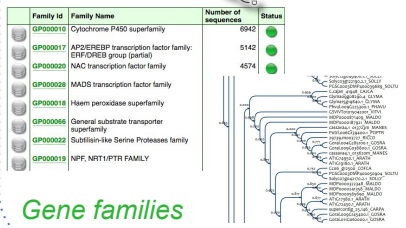


Gigwa



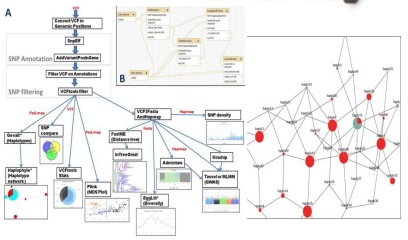
SNPs and Indels

GreenPhyl



Gene families

SNiPlay



https://github.com/SouthGreenPlatform

@green_bioinfo

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

South Green bioinformatics platform

ABiMS — Erwan Corre

IPME — Marie Simonin, Sébastien Cunnac

ASTRE Animal · Santé Territoires · Risques · Écosystèmes — Etienne Loire, Julie Reveillaud

UMR QualiSud — Florentin Constancias

diade — Valentin Klein

MiVEGEC Maladies Infectieuses et Vecteurs Écologie, Génétique, Évolution et Contrôle

i-Trop bioinformatics — Valérie Noël

And more collaborators !

SouthGreen bioinformatics platform

18-19/03	Guide de survie à Linux - IRD
21/03	Initiation à l'utilisation du cluster CIRAD – CIRAD
22/03	Initiation à l'utilisation du cluster itrop - IRD
15-16/04	Initiation au gestionnaires de workflow SG & Gigwa – IRD
18-19/04	Guide du Jedi en Linux & bash - CIRAD
13–16/05	Python - IRD
17/05	Initiation aux analyses de données transcriptomiques – IRD
21/05	Utilisation avancée du cluster IRD – IRD
23-24/05	Initiation aux analyses de données métagénomiques – IRD
6/06	Manipulation de données et figures sous R – CIRAD
25-27/09	Assemblage et annotation de transcriptomes - IRD

Trainings 2019

- **South Green Trainings** :
  https://southgreenplatform.github.io/trainings/
- Slides & Practices : **RNAseq**
- Working environment : **Softwares to install**

Initiation aux analyses de données transcriptomiques

www.southgreen.fr

https://southgreenplatform.github.io/trainings

Objectif du module

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 TOGGLe : Hisat2 + stringtie
- Pseudo-alignment using Galaxy : kallisto
- Differential expression analysis: EdgeR, Deseq2 : package shiny Pivot
- Clustering, co-expression network: R

Pourquoi faire du RNA-Seq

L'accès aux séquences d'ARN permet de :

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

Le choix technologique (depletion/enrichissement, démultiplexage, séquençage directionnel) dépendra de la question biologique

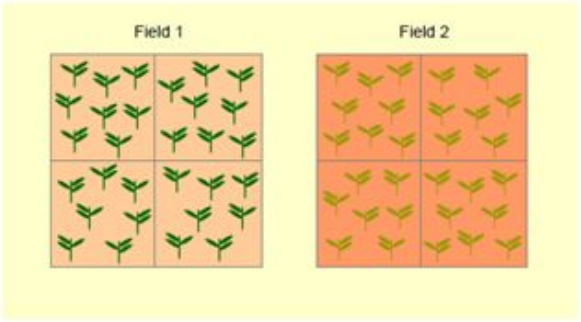
1- Design experimental

Experimental design

Basic experiment : trouver les différences entre condition contrôle/traitée

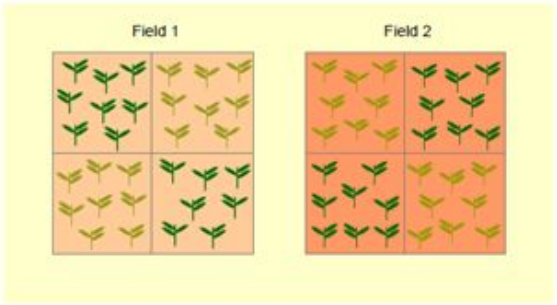
control group plant treated group plant

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



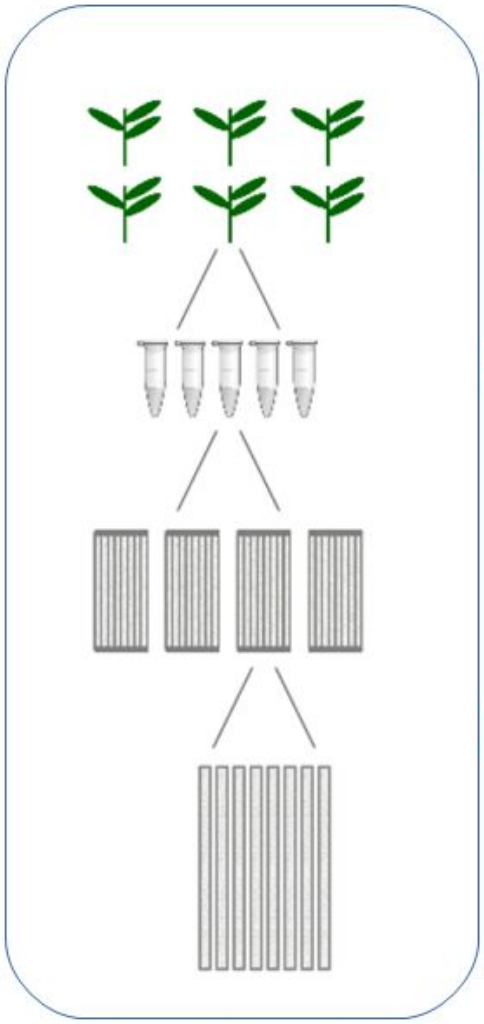
Pas de possibilité de différencier l'effet champs de l'effet traitement

Bon plan expérimental : la moitié des plantes traitées poussent avec un contrôle dans un même champs et l'autre moitié dans un autre champs



Possibilité de différencier l'effet champs de l'effet traitement.

Experimental design



collect?

1 – Variations biologiques : variations individuelles dues aux effets génétiques, de l'environnement

Sample preparation

2 – Variations techniques : effet de la préparation des librairies

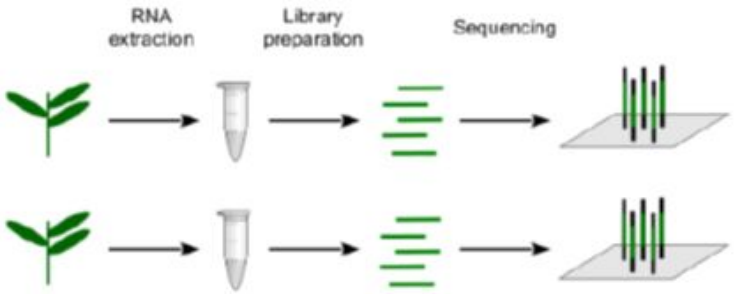
cDNA on lane of flowcell

3 – Variation techniques : effet des lane et des flowcell

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

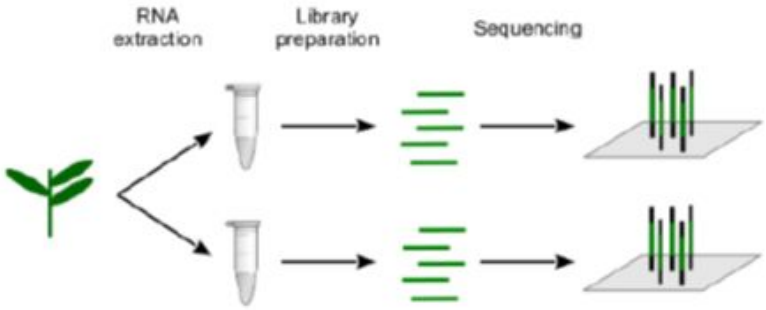
Experimental design

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



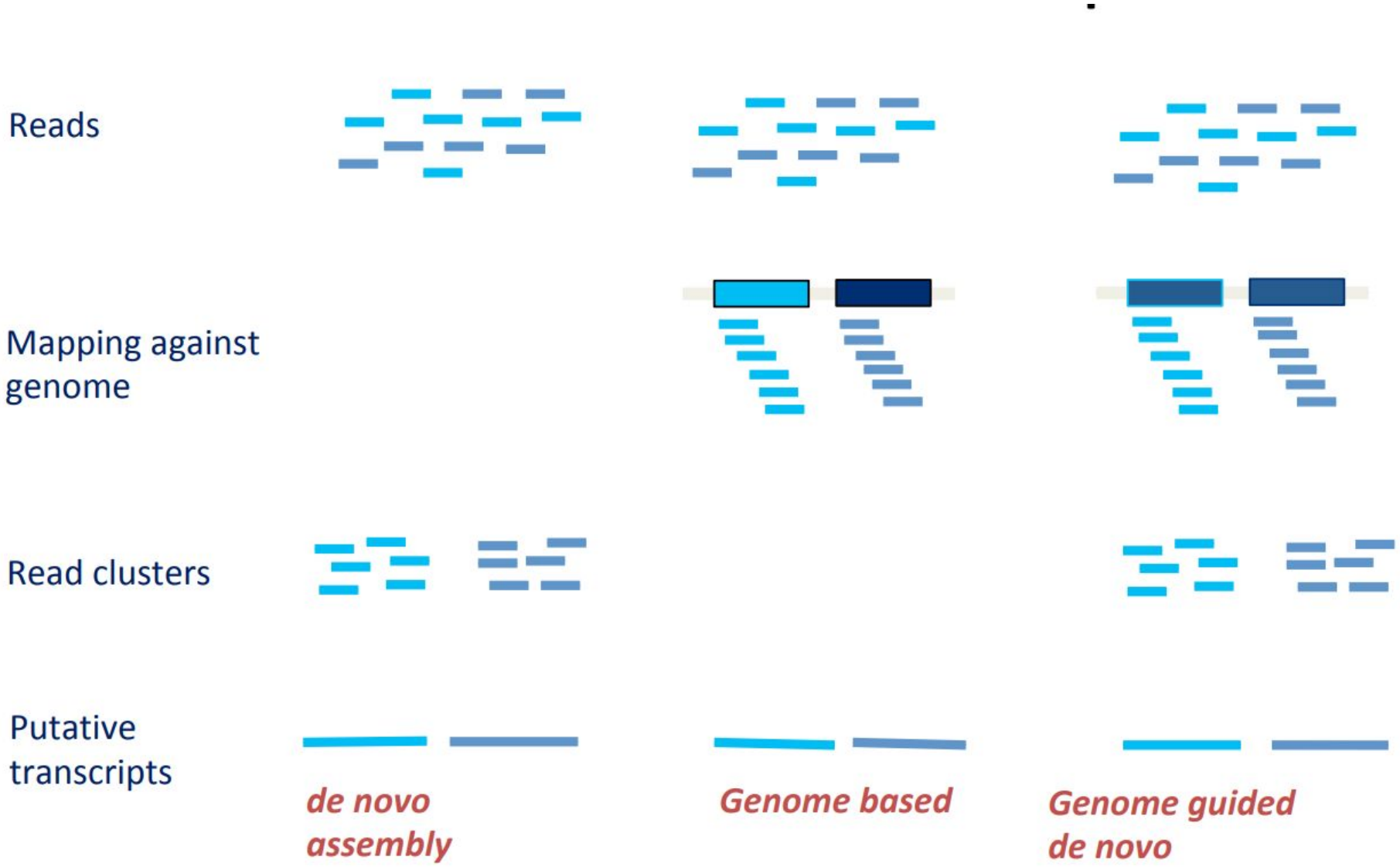
Réplicat Technique : Même matériel biologique, répété plusieurs fois indépendamment des étapes techniques.

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



2- Des reads aux transcripts

Des reads aux transcripts

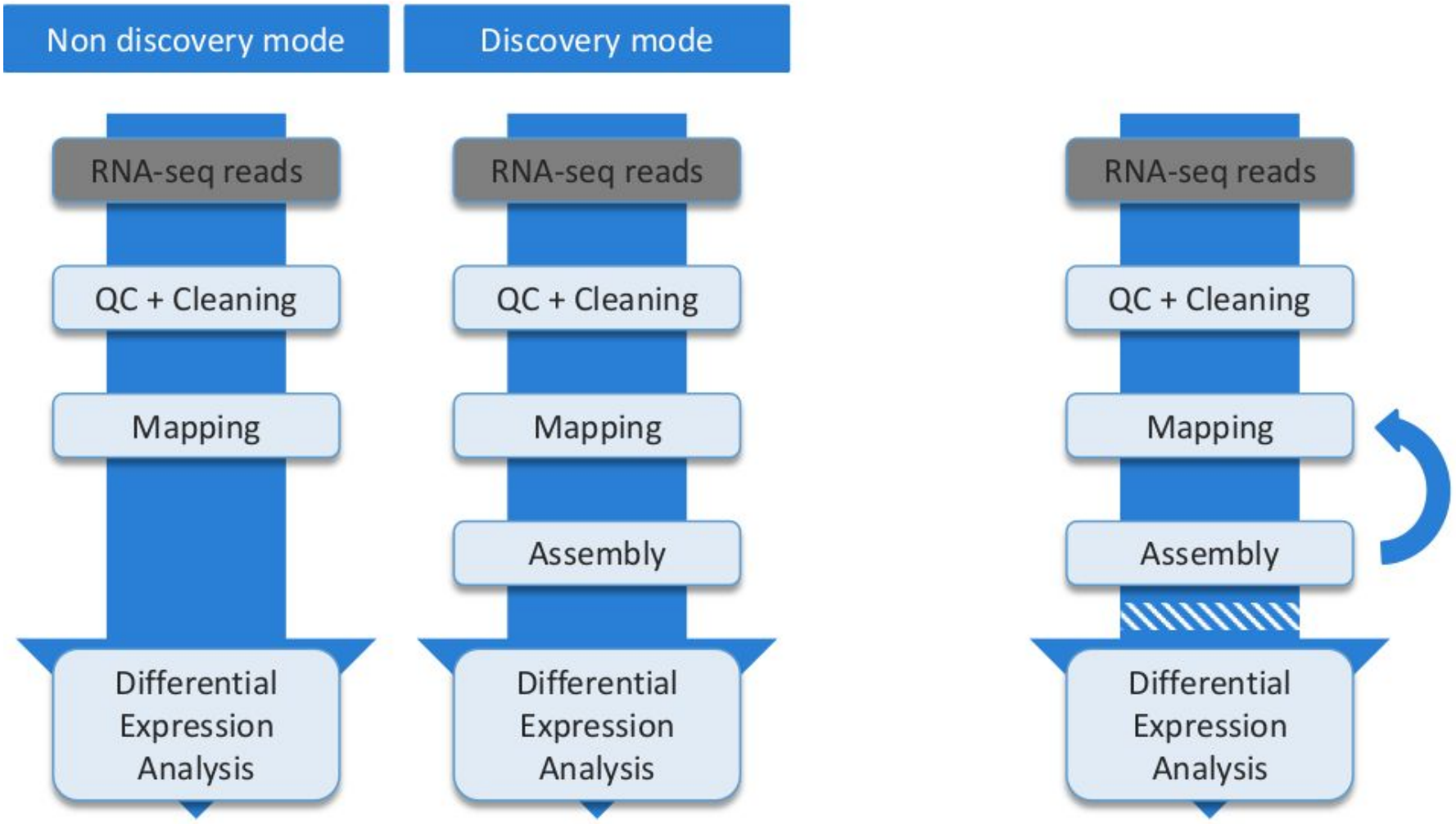


From http://www.bioinformatics.nl/courses/RNAseq/1c%20Assembly.pdf

Vision Global of RNAseq Analysis

- Reference genome
- Reference transcriptome

- Reference genome
- No reference transcriptome



From Abims RNAseq Training oct 2018

Vision Global de l'analyse RNAseq

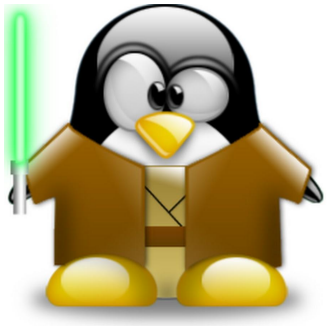
17/05 Initiation aux analyses de données transcriptomiques – IRD

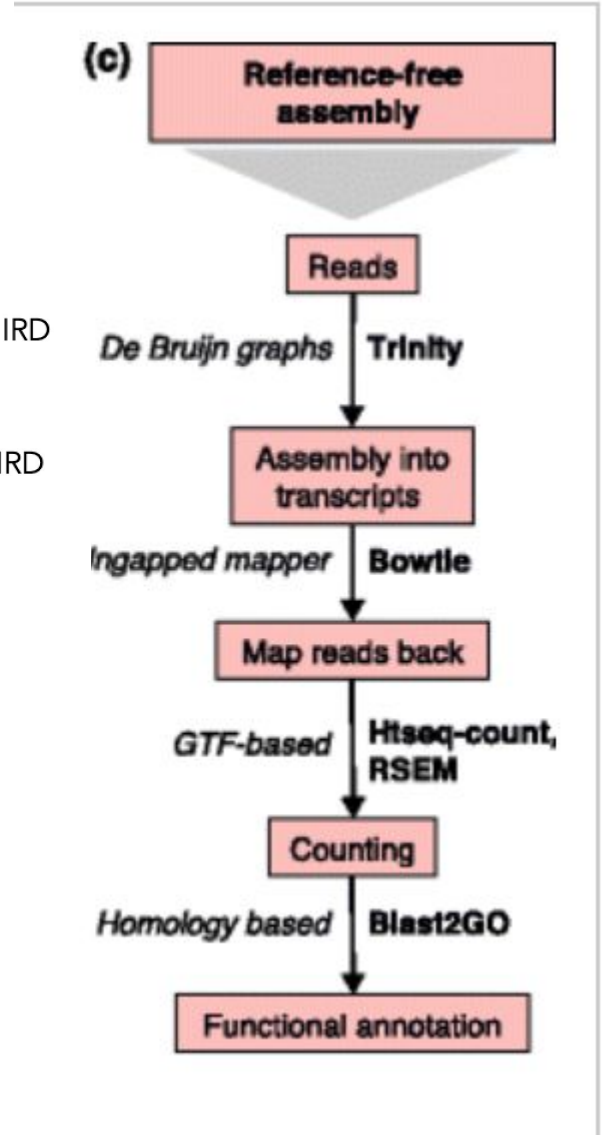
21/05 Utilisation avancée du cluster IRD – IRD

23-24/05 Initiation aux analyses de données métagénomiques – IRD

6/06 Manipulation de données et figures sous R – CIRAD

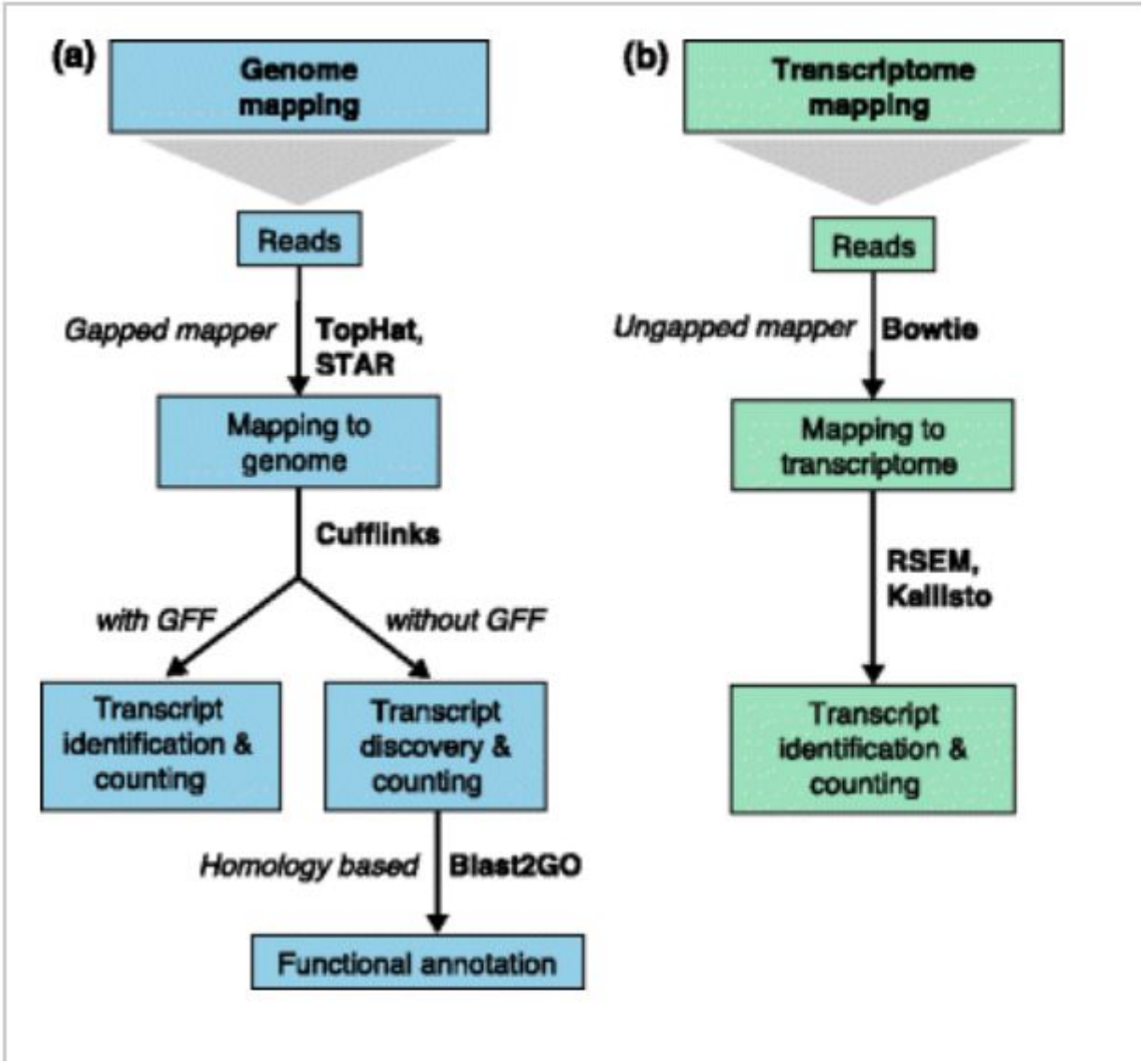
25-27/09 Assemblage et annotation de transcriptomes - IRD





https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Analysis RNAseq avec une référence

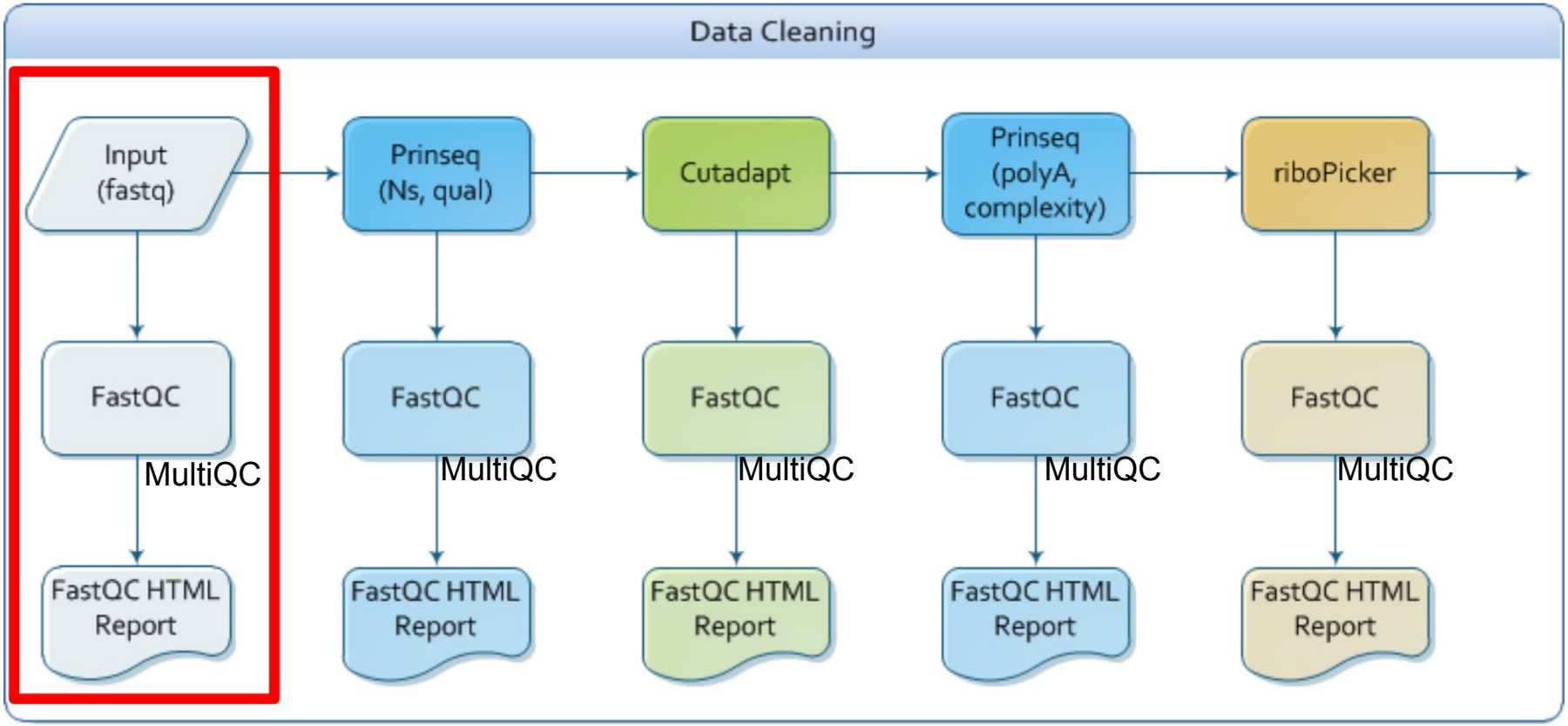


https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

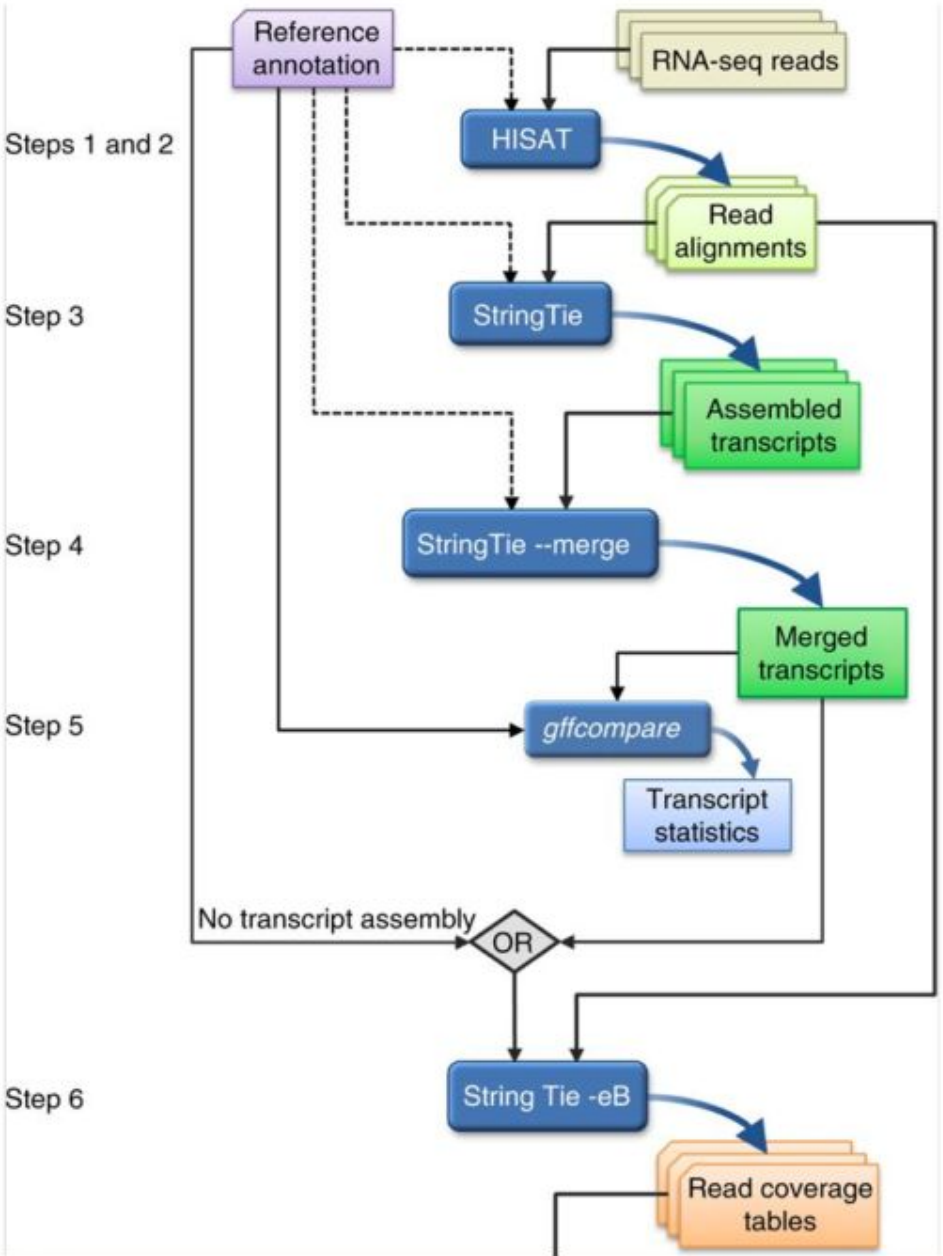
Vérification de la qualité des données NGS

	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 *Trimmomatic*
Adaptors and primers	Peuvent être trouvés dans le 3' final d'un insert très court	Des ponts entre séquences sans relation aucune: Chimères	*Trimmomatic*, cutadapt, far, btrim, SeqTrim, TagCleaner, solexaQA
Poly A/T tails, low complexity reads	Des queus poly A/T peuvent etre laissés pendant la préparation de la librairie	Des ponts entre séquences sans relation aucune: Chimères	*PRINSEQ2*
Contaminations	RNA Ribosomal RNA/DNA etrangere (PhiX, Bacteria, ...)		SortMeRNA, riboPicker, DeconSeq

Nettoyage des reads



New Tuxedo Workflow



Mapping

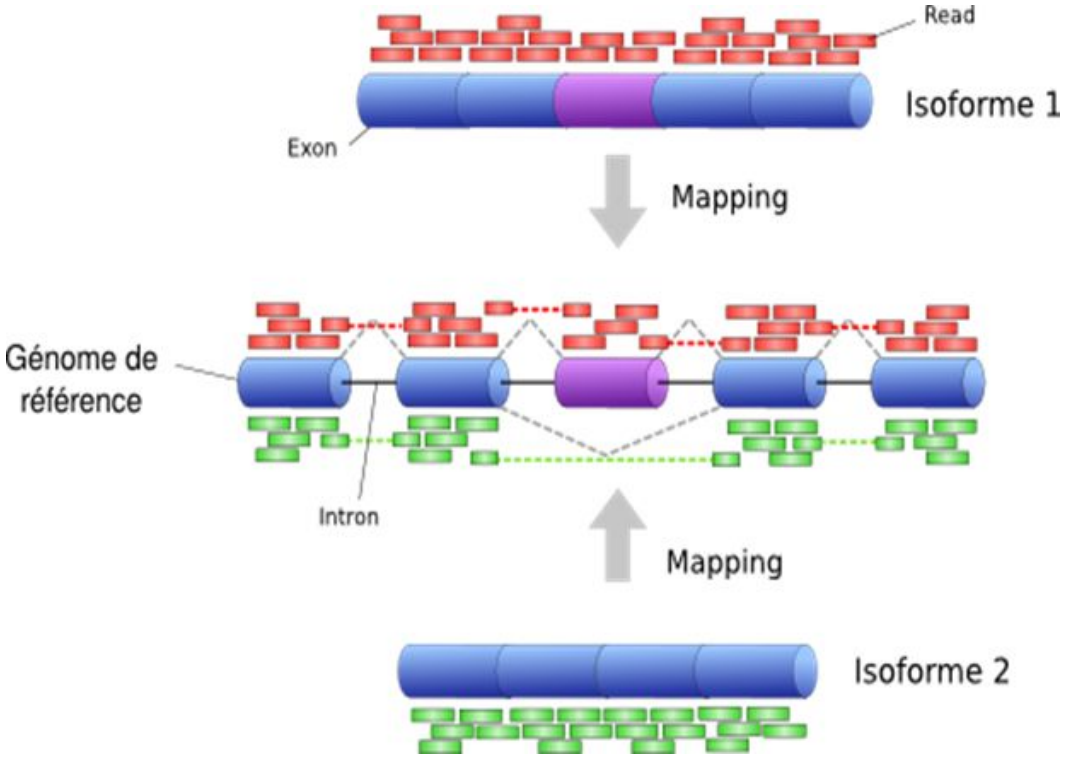
Assemblage des nouveau transcrits

Comparaison avec l'annotation

Obtention des tableaux de comptages

Mapping sur référence génomique

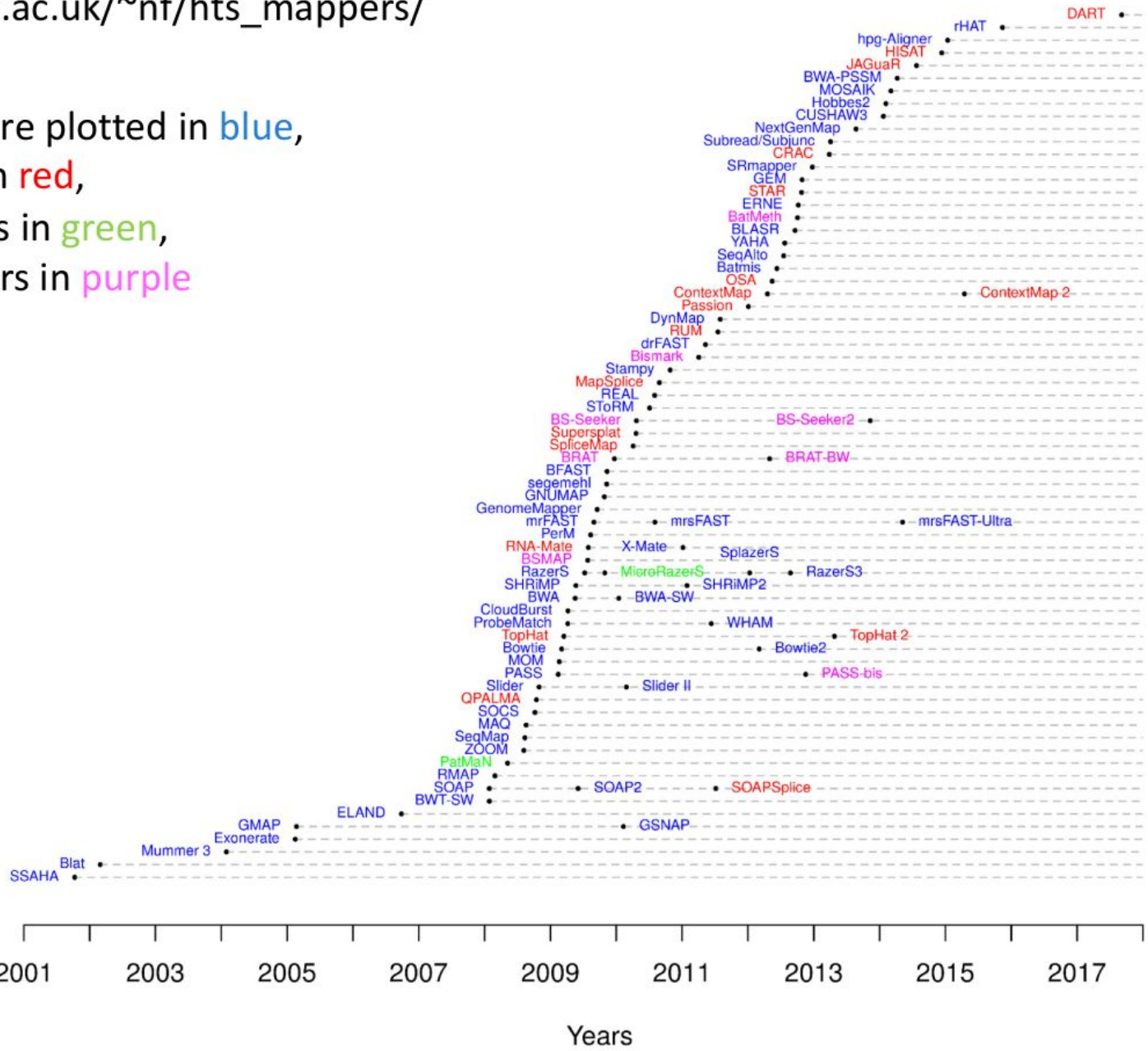
- Permet la mise en évidence d'isoformes
- Aide à l'annotation structurale du génome



Choix de l'outil de mapping

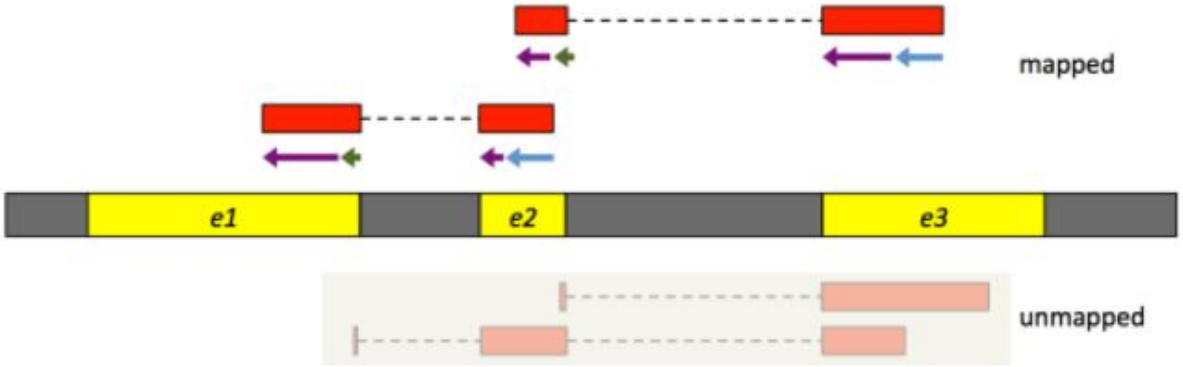
http://www.ebi.ac.uk/~nf/hts_mappers/

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

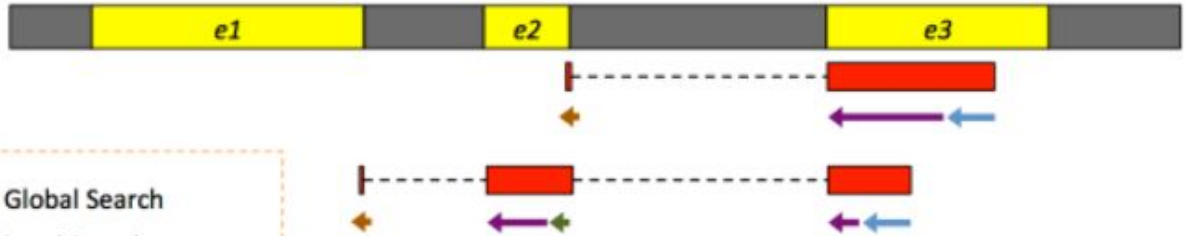


HISAT

1st run of HISAT to discover splice sites



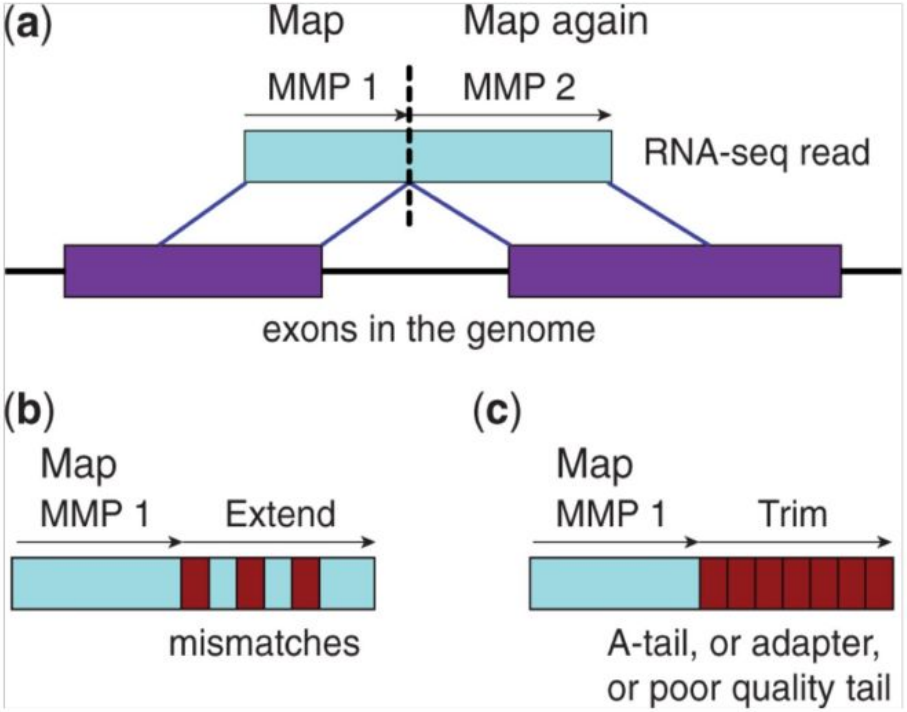
2nd run of HISAT to align reads by making use of the list of splice sites collected above



Kim et al. *Nature Methods* 2015

STAR

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

SAM format

```
@SQ     SN:sctg_997     LN:6840
@SQ     SN:sctg_998     LN:6818
@SQ     SN:sctg_999     LN:6999

@PG     ID:TopHatVN:2.0.3

HWI-ST132_0435:3:63:3889:100528#GATCAG

0
sctg_0
476
3
73M
*
0
0

CCCGCCGCTCCATGATCTCCAAGAGGCGCAGCTCTCGCAAGGCTTCCGCCAAGGTGGTGGCTT
ggggggggggggggggggggggggggggggeeeggeeggYb^ce^bbbc_cac[ddacaa_c

AS:i:0 XN:i:0 XM:i:0 XO:i:0 XG:i:0 NM:i:0 MD:Z:73 YT:Z:UU XS:A:+ NH:i:2
CC:Z:sctg_861 CP:i:9032  HI:i:0
```

Reference sequence dictionary

Program

Sequence ID

Flag
Reference sequence name
Leftmost mapping position
Mapping quality
CIGAR string
Reference of the mate/next read
Mapping position of mate/next read
Insert size

Sequence ID
Quality

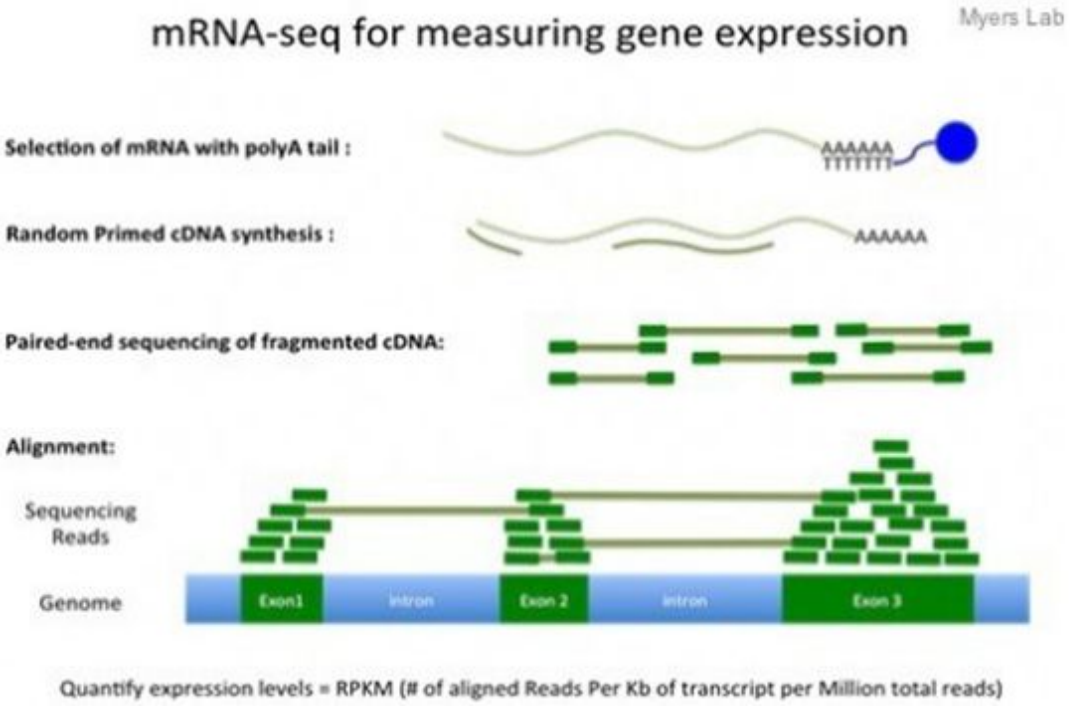
Mapper tag

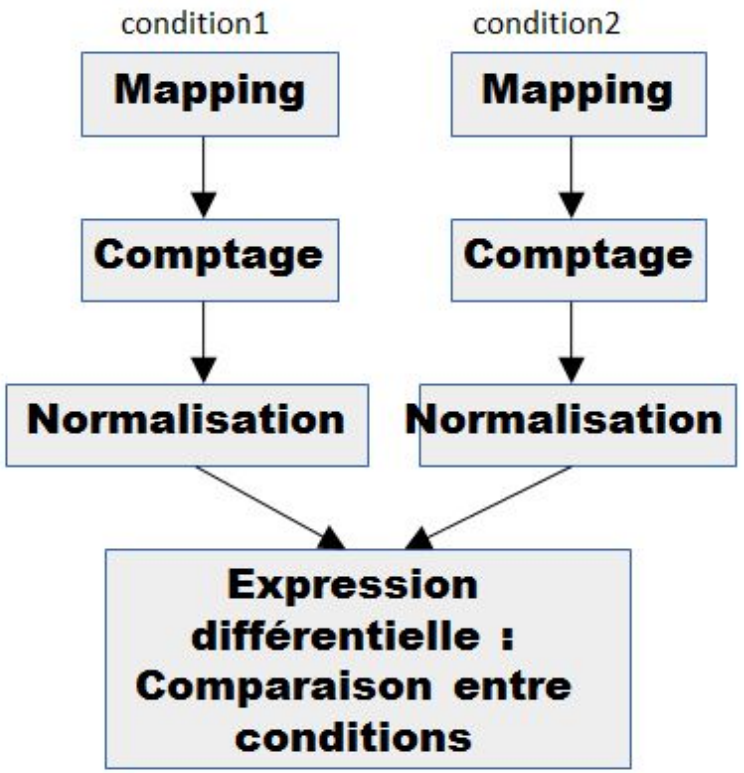
htpp://samtools.sourceforge.net/SAMv1.pdf

source: Abims RNAseq formation 2018

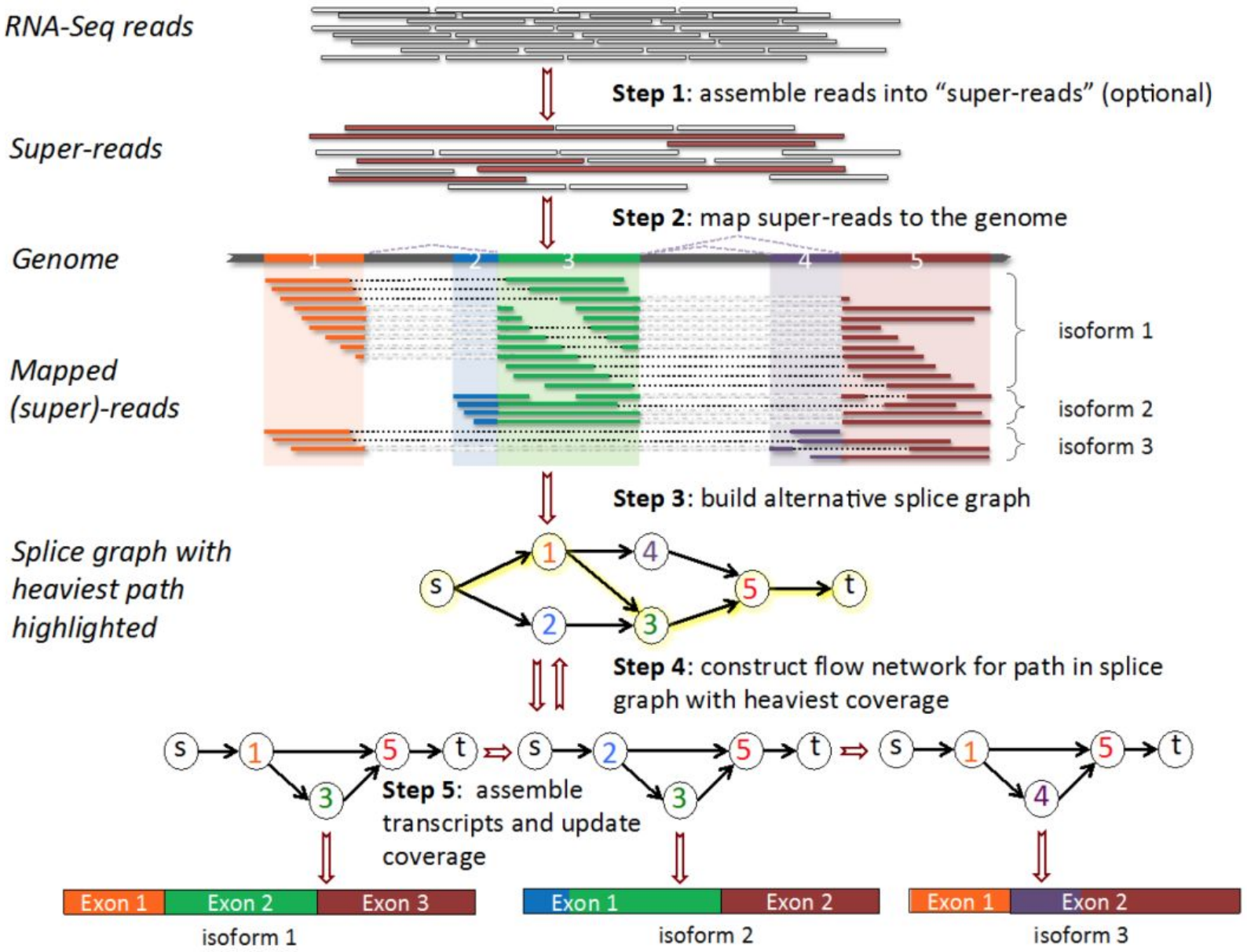
3- Comptage

Principe du comptage des reads





StringTie

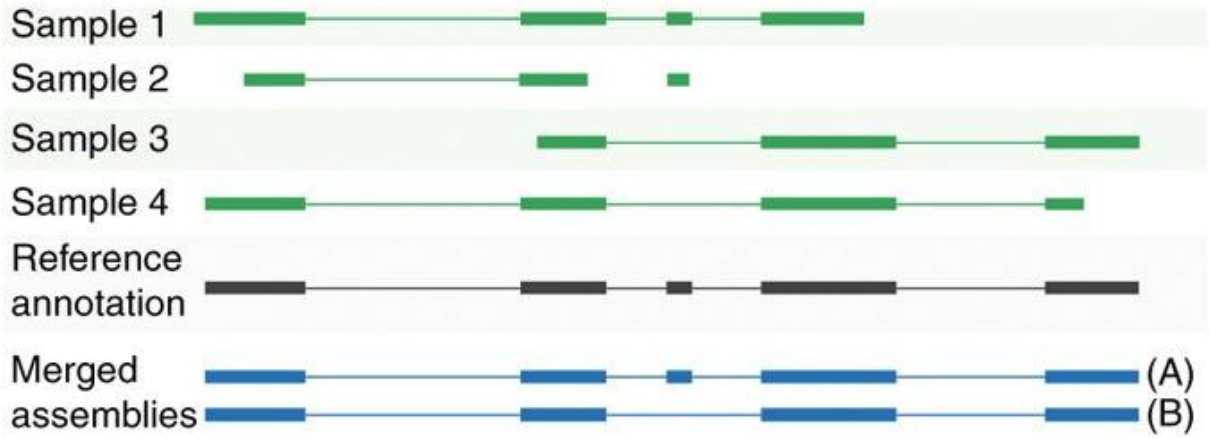


Pertea, et al. (2015). "StringTie enables improved reconstruction of a transcriptome from RNA-seq reads." Nat Biotech **33**(3): 290-295.

StringTie

Figure 2 : Merging transcript assemblies using StringTie's merge function.

From: Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

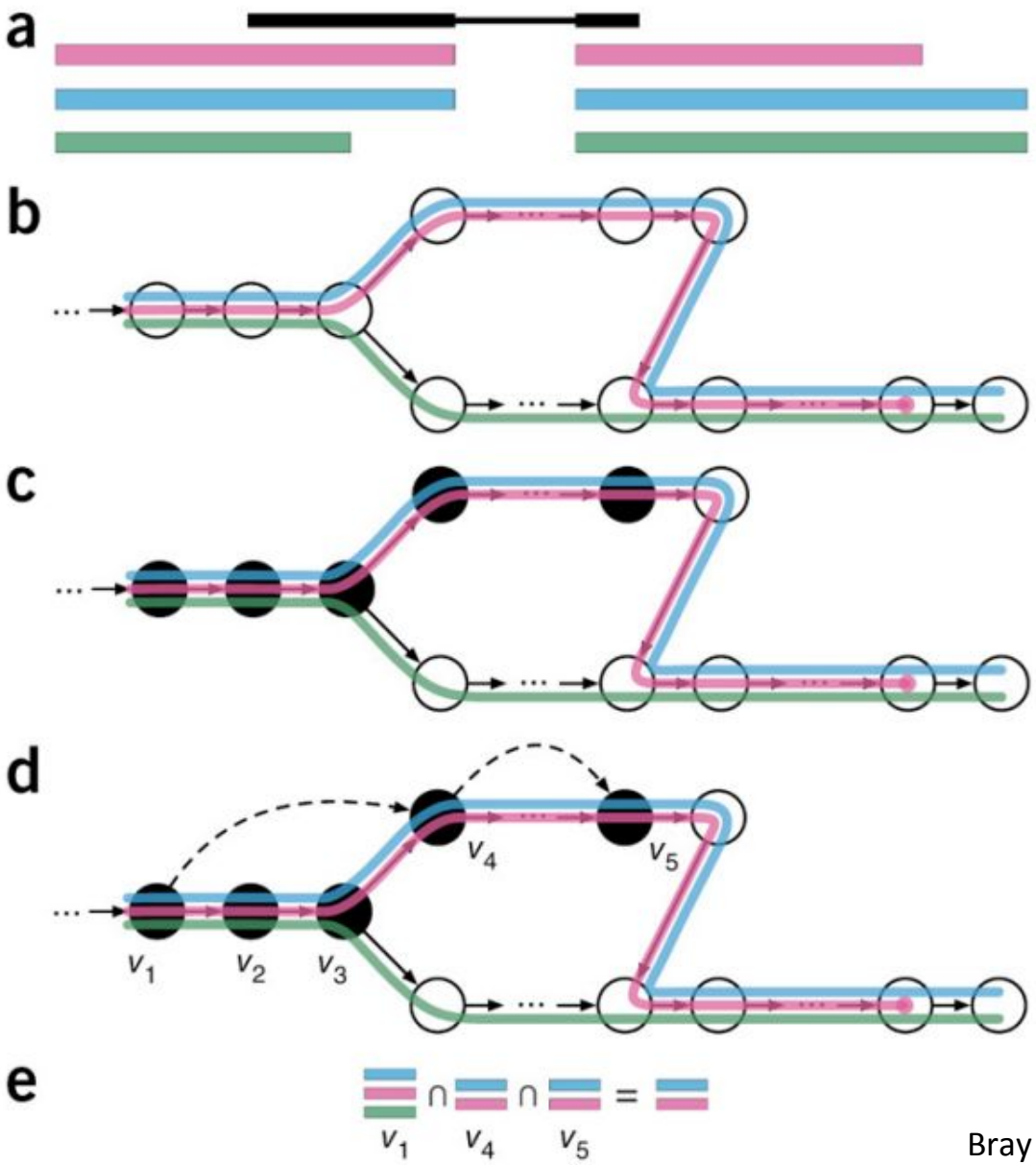


In this example, four partial assemblies from four different samples are merged into two transcripts A and B. Samples 1 and 2 are both consistent with the reference annotation, which is used here to merge and extend them to create transcript A. Samples 3 and 4 are consistent with each other but not with the annotation, and these are merged to create transcript B.

Pseudoalignement

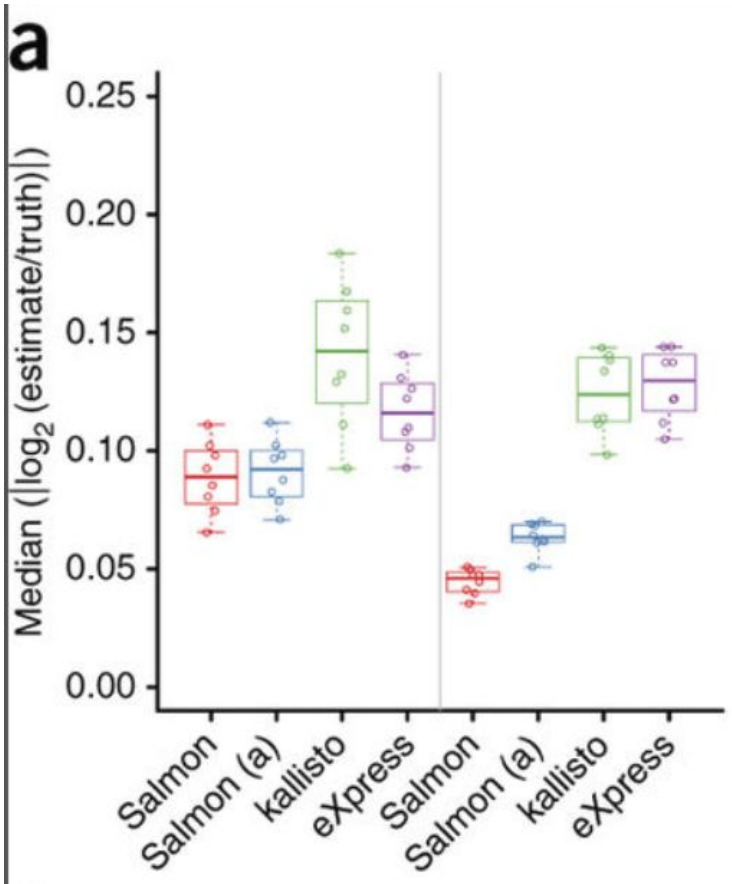
Quantifying abundances of transcripts by ***pseudoalignment*** for rapidly determining the compatibility of reads with targets, without the need for alignment.

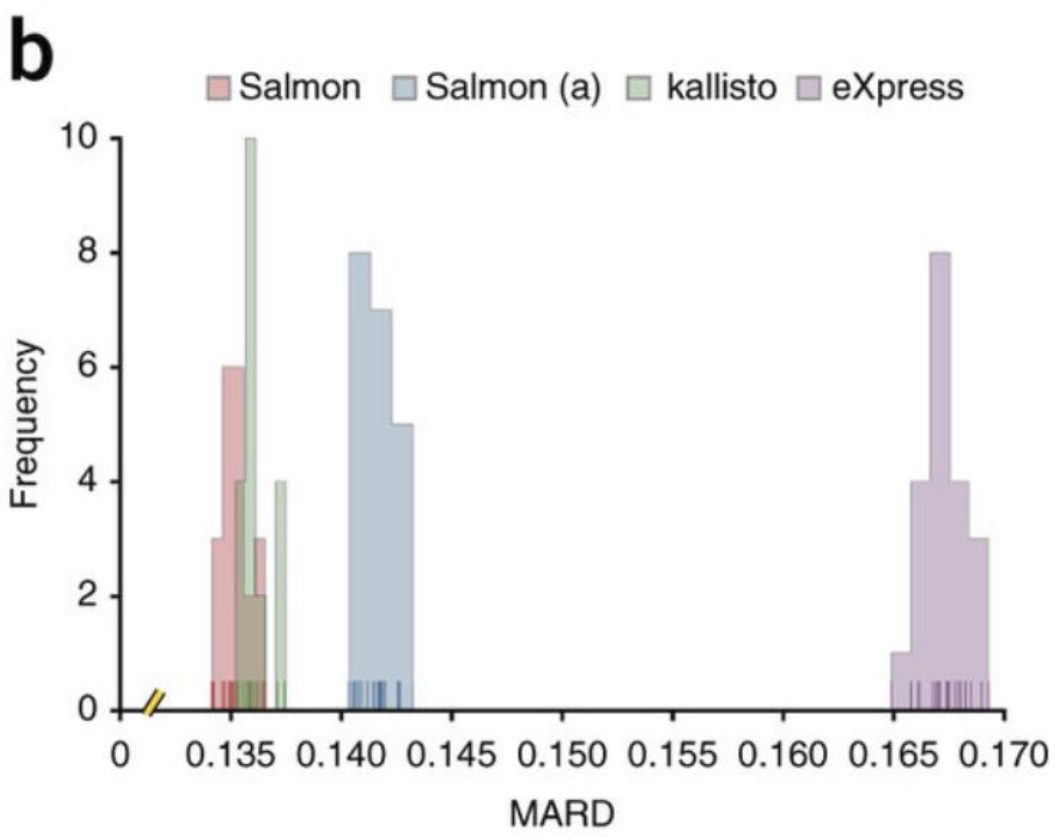
Kallisto



Bray et al. 2016. Nat Biotech

Salmon





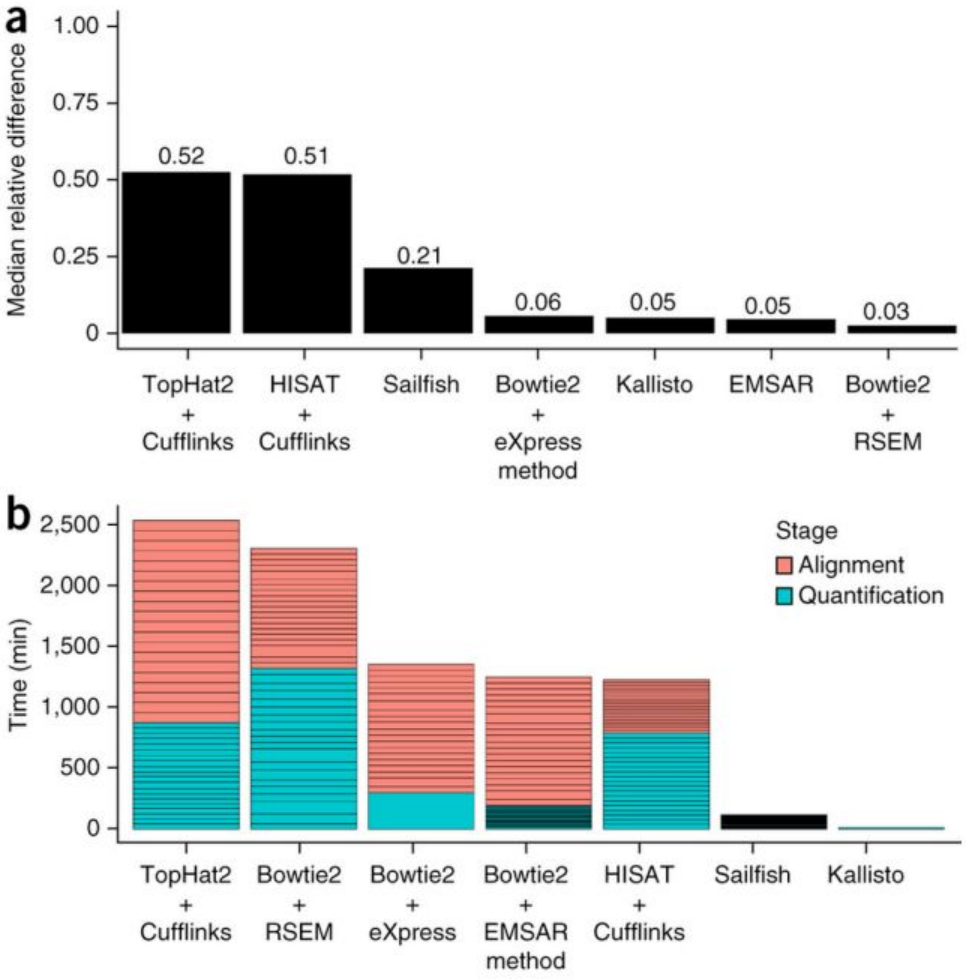
c

	Sensitivity at given FDR			
FDR	Salmon	Salmon (a)	kallisto	eXpress
0.01	0.326	0.233	0.072	0.128
0.05	0.409	0.379	0.248	0.162
0.1	0.454	0.442	0.296	0.211

d

Type	Salmon	Salmon (a)
All transcripts	1,183	1,197
Two-isoform genes	228	244
Type	**kallisto**	**eXpress**
All transcripts	2,620	2,472
Two-isoform genes	545	531

Comparaison des méthodes



Bray, N. L., H. Pimentel, P. Melsted and L. Pachter (2016). "Near-optimal probabilistic RNA-seq quantification." Nat Biotech **34**(5): 525-527.

gtf/gff Format

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

gff3

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attribute
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

source: Abims RNAseq formation 2018

gff compare

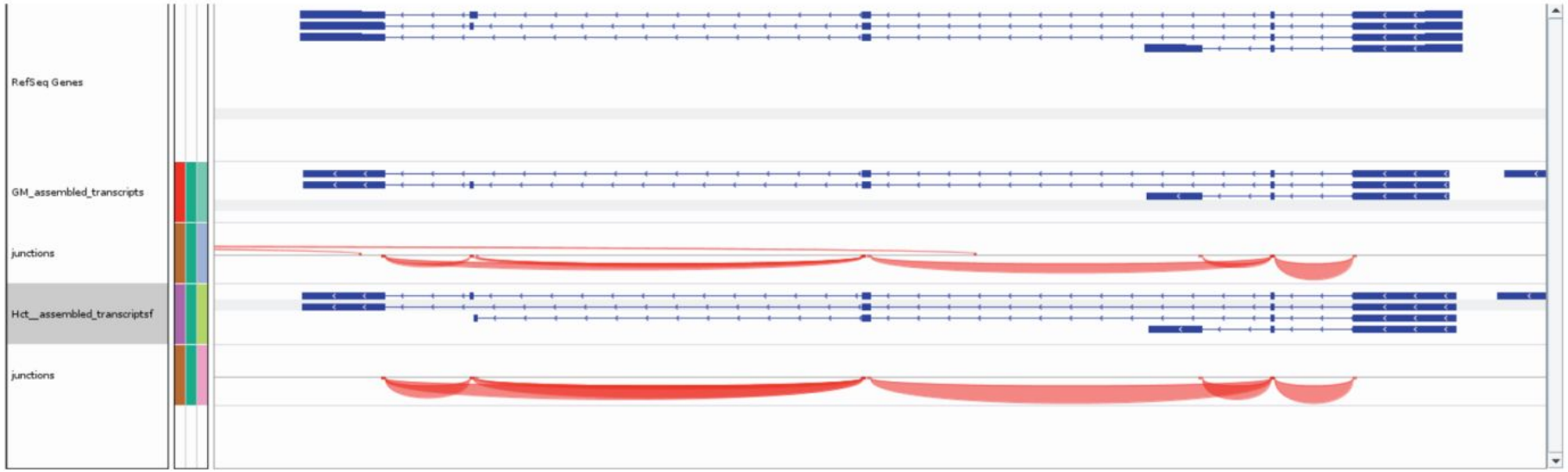
The following table shows the code used by Cufflinks to classify the transcripts in comparison with the reference annotation

Priority	Code	Description
1	**=**	**Complete match of intron chain**
2	c	Contained
3	**j**	**Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript**
4	e	Single exon transfrag overlapping a reference exon and at least 10 bp of a reference intron, indicating a possible pre-mRNA fragment
5	i	A transfrag falling entirely within a reference intron
6	o	Generic exonic overlap with a reference transcript
7	p	Possible polymerase run-on fragment (within 2Kbases of a reference transcript)
8	r	Repeat. Currently determined by looking at the soft-masked reference sequence and applied to transcripts where at least 50% of the bases are lower case
9	**u**	**Unknown, intergenic transcript**
10	**x**	**Exonic overlap with reference on the opposite strand**
11	s	An intron of the transfrag overlaps a reference intron on the opposite strand (likely due to read mapping errors)
12	.	(.tracking file only, indicates multiple classifications)

source: Abims RNAseq formation 2018

gff compare

=

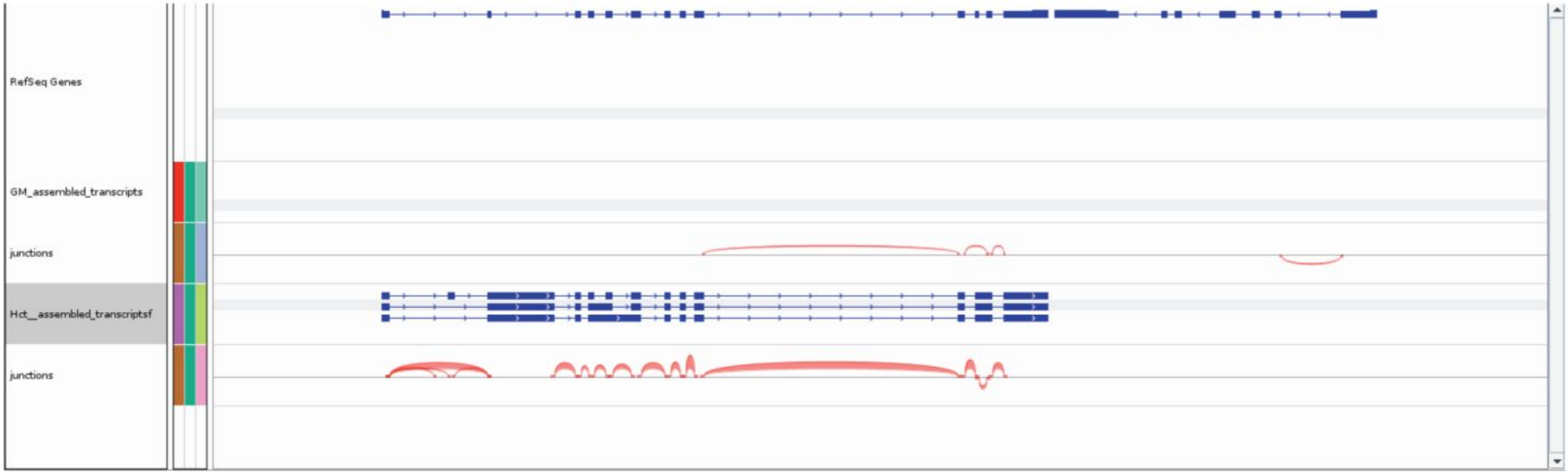


Complete match of intron chain

source: Abims RNAseq formation 2018

gff compare

J

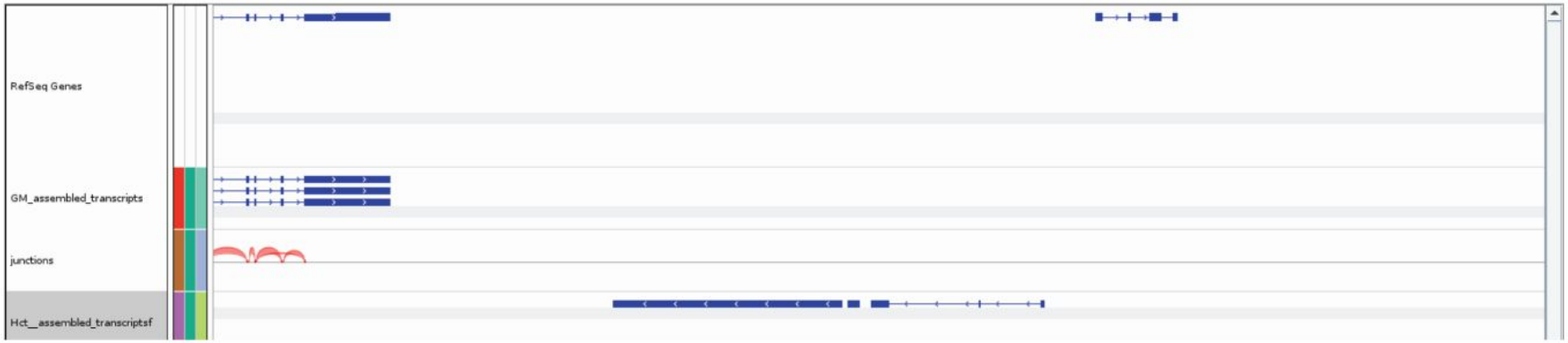


Potentially novel isoform (fragment): at least one splice junction is shared with a reference transcript

source: Abims RNAseq formation 2018

gff compare

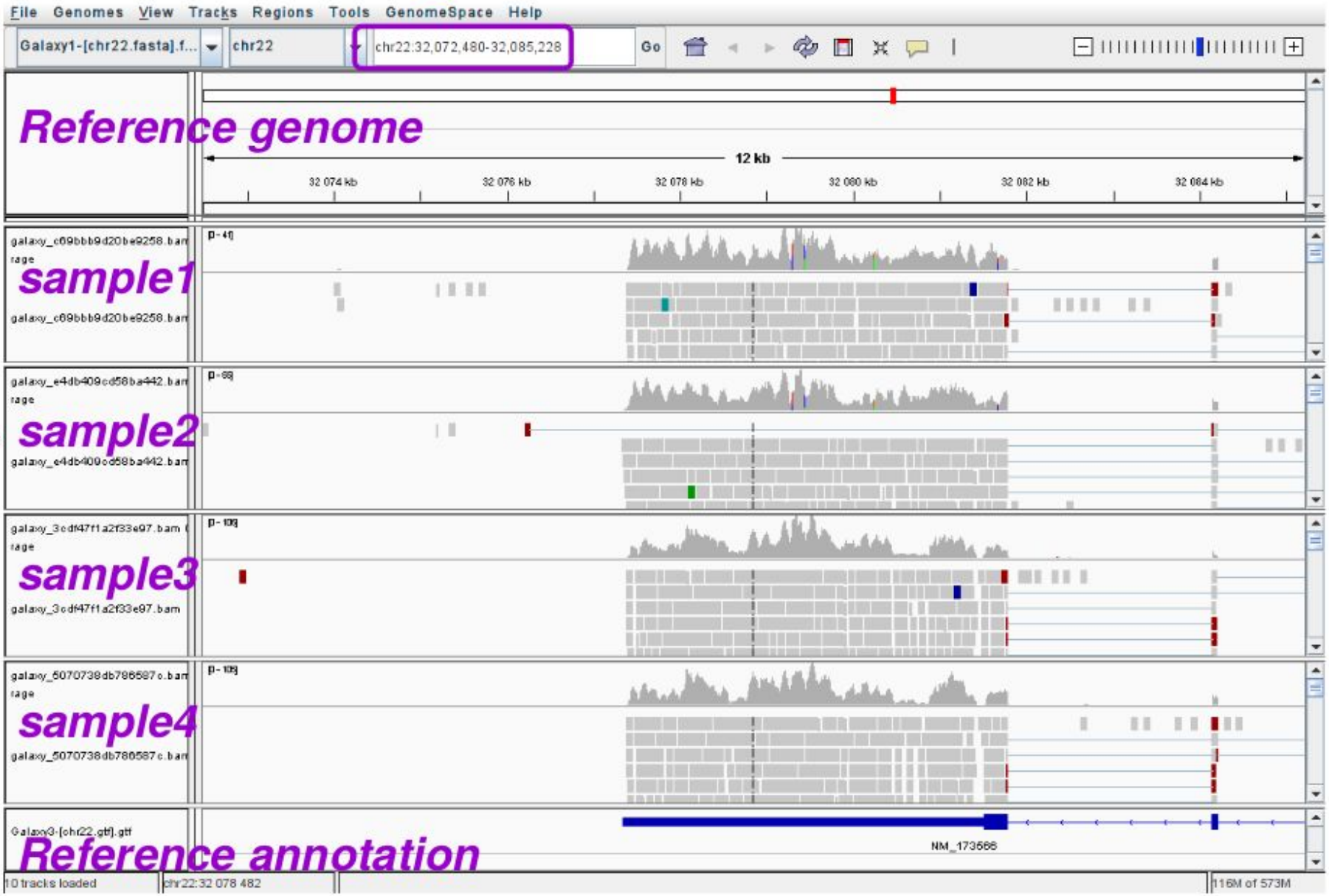
U



Unknown, intergenic transcript

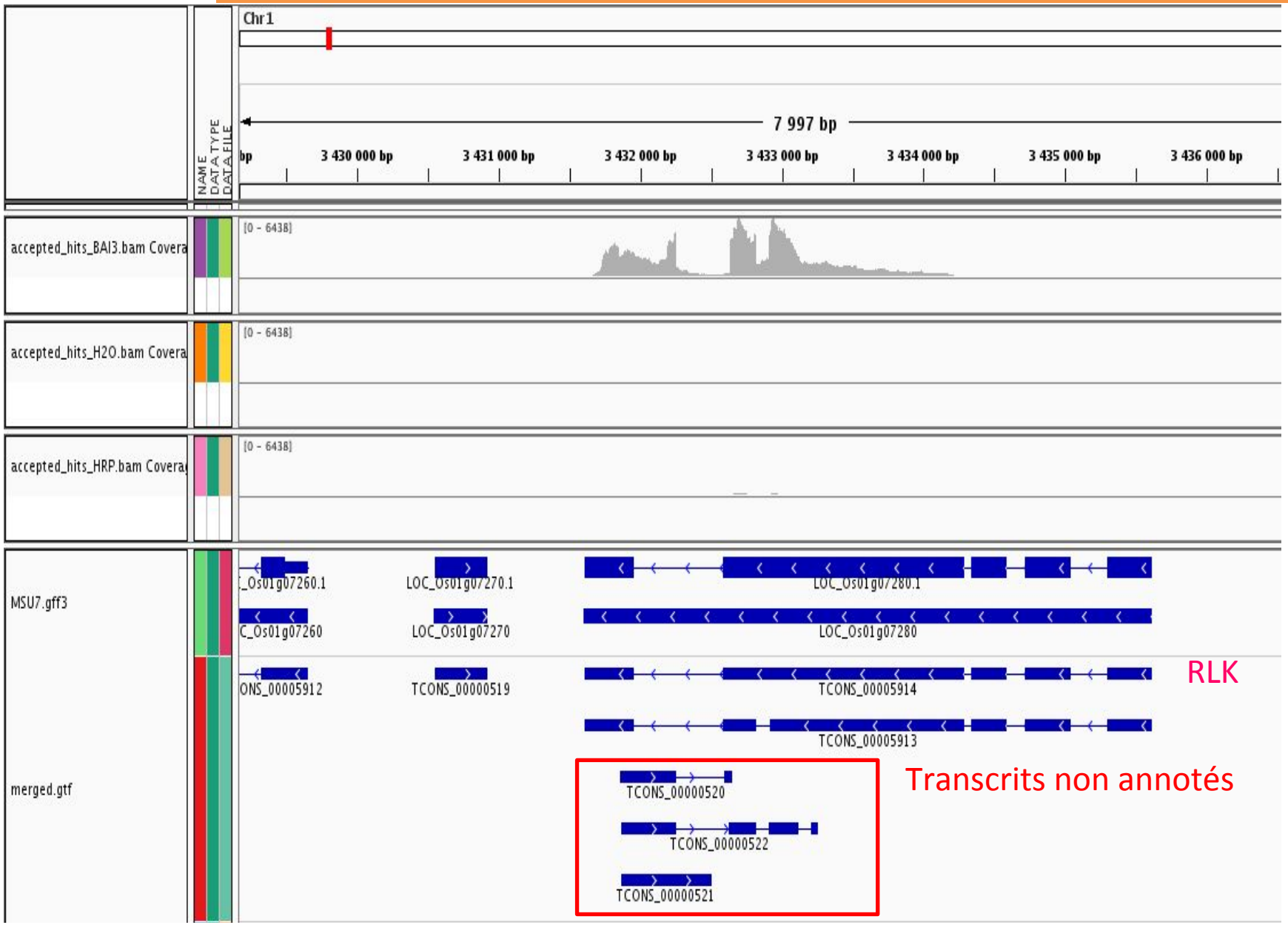
source: Abims RNAseq formation 2018

Local Genome Browser (IGV)



source: Abims RNAseq formation 2018

Local Genome Browser (IGV)



Practice : Galaxy and TOGGLe

TP-Galaxy-TOGGLe

4- Normalisation des données

Normalisation des données

- Identifier et corriger les biais techniques dus au séquençage, pour les rendre comparable

- Types de Normalisation : intra-echantillons (meme sequençage) et inter-echantillons (deux sequençages)

Taille de la banque
Longueur de gènes
Composition en GC des gènes

Normalisation des données

	sample 1	sample 2	sample 3
gene A	752	615	1203
gene B	1507	1225	2455

Le comptage du gène B est deux fois plus important que pour le gène A, pourquoi ?

gene A gene B

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

gene A gene B

Les deux gènes ont le même nombre de transcrits,
mais le gène B est deux fois plus long que le gène A.

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

Normalisation des données

	sample 1	sample 2	sample 3
gene A	752	615	1203
gene B	1507	1225	2455

Le comptage dans l'échantillon 3 est plus important que dans l'échantillon 2.

gene A in sample 2 gene A in sample 3

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

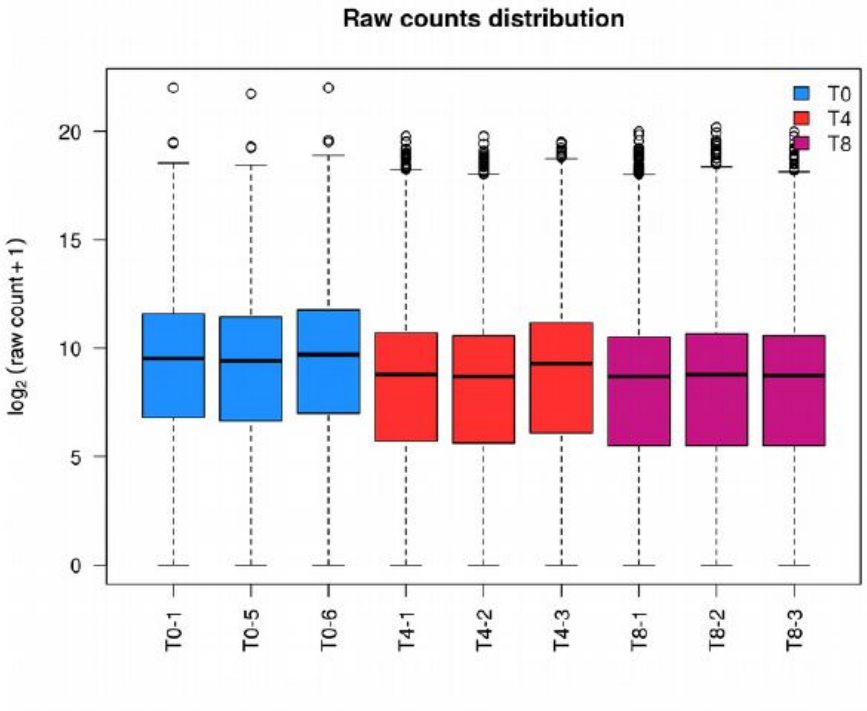
gene A in sample 2 gene A in sample 3

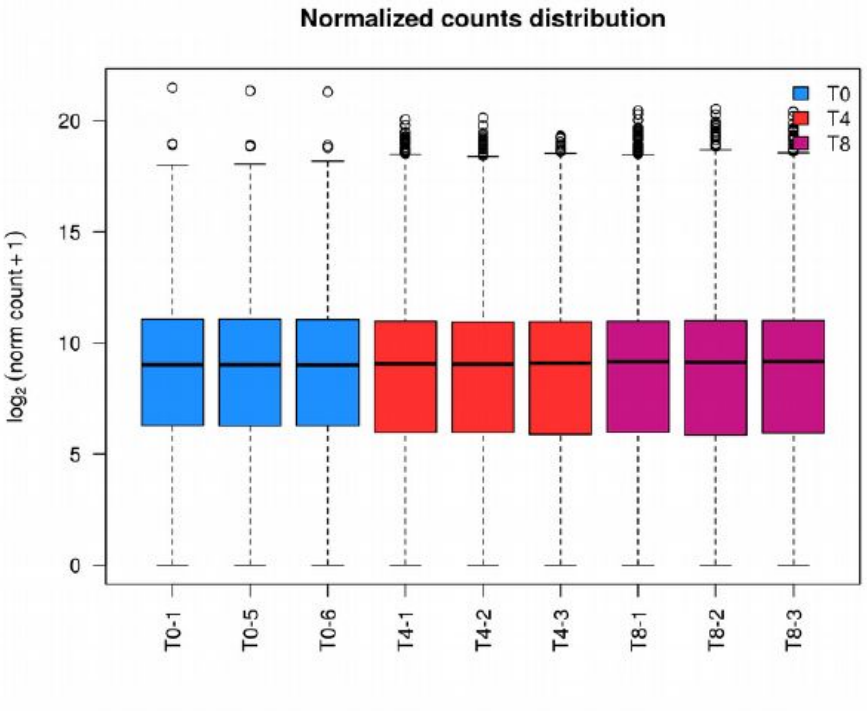
Le gène A est exprimé dans les échantillons 2 et 3, mais la profondeur de séquençage est plus importante dans l'échantillon 3 que dans le 2 (différences de taille des librairies).

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

Normalisation des données

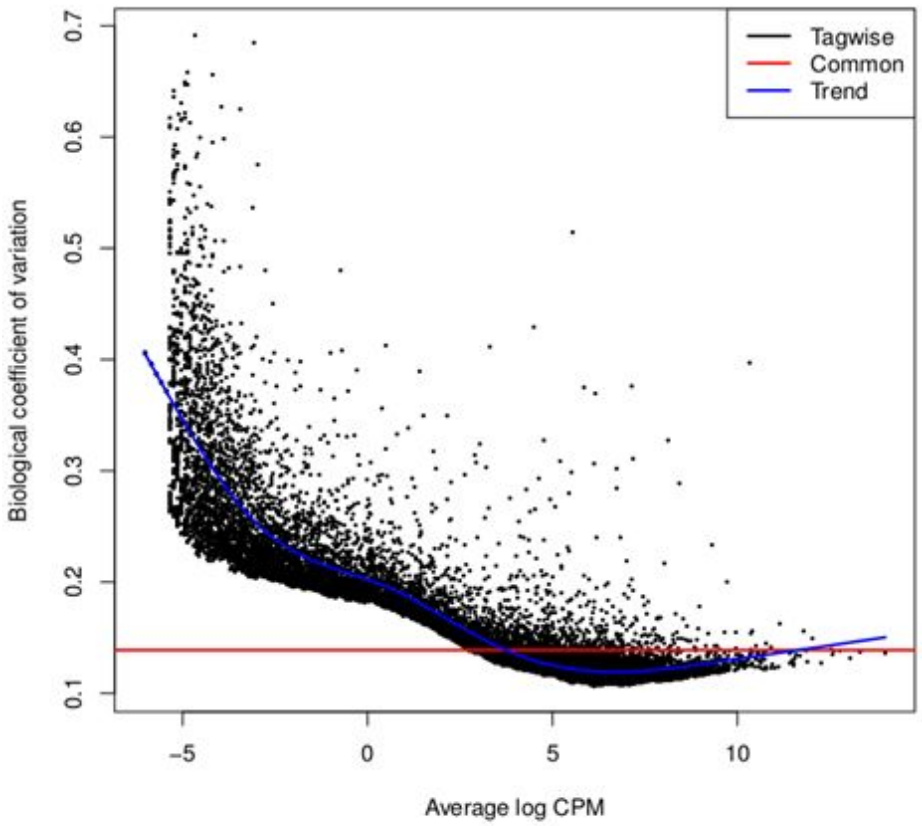
Effet de la normalisation :
Variance des banques RNAseq avant et après normalisation





Méthodes basées sur les normalisations interbanques

EdgeR



Estimation de la dispersion
(entre réplicats biologiques)

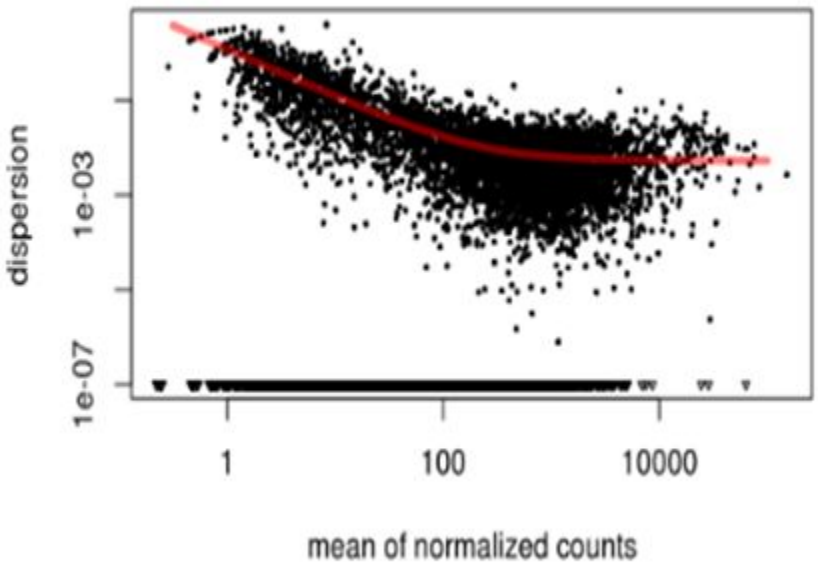
- utilisation de la valeur individuelle (« tagwise ») ou de la valeur ajustée « trend » ou « common » pour le calcul des tests statistiques de DE

- 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

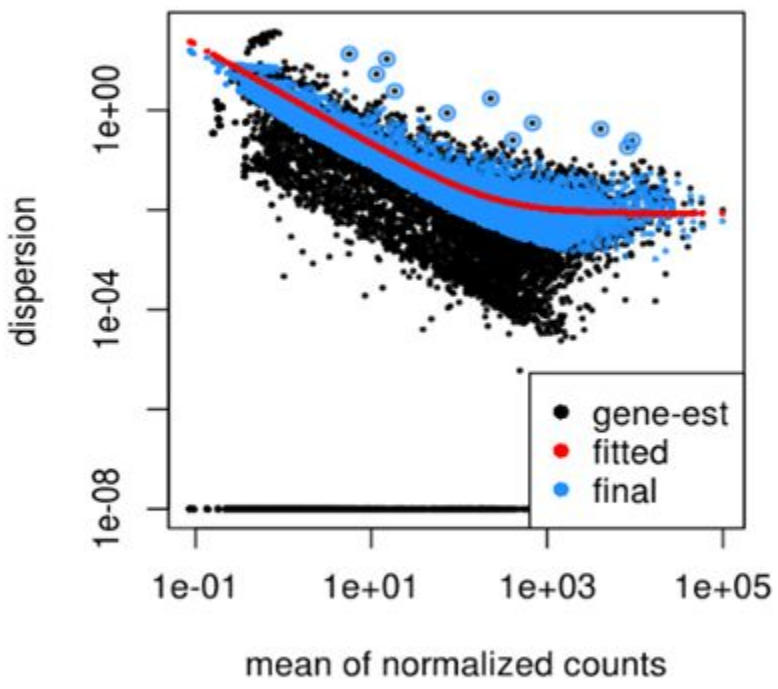
- utilisation de la valeur ajustée pour les transcrits dont l'estimateur individuel (en noir) inférieur à la valeur ajustée

- 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

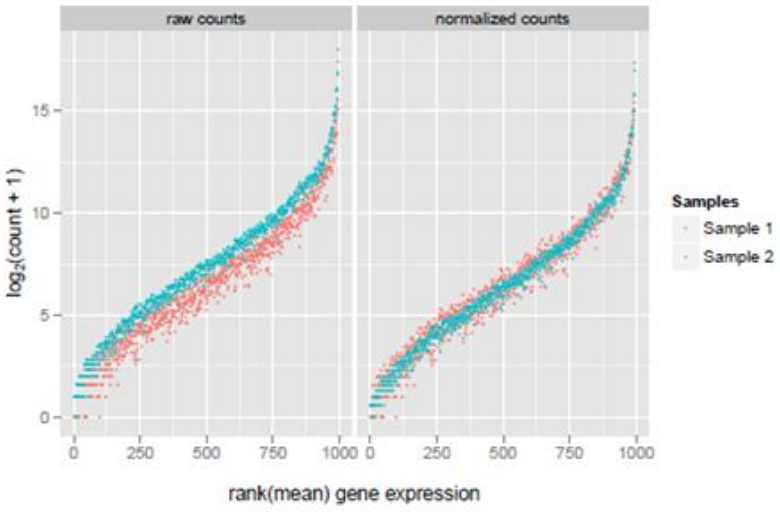
DESeq2 plus souple, sera moins stringent et détectera plus de gènes différentiellement exprimés.

Méthodes basées sur les normalisations interbanques

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



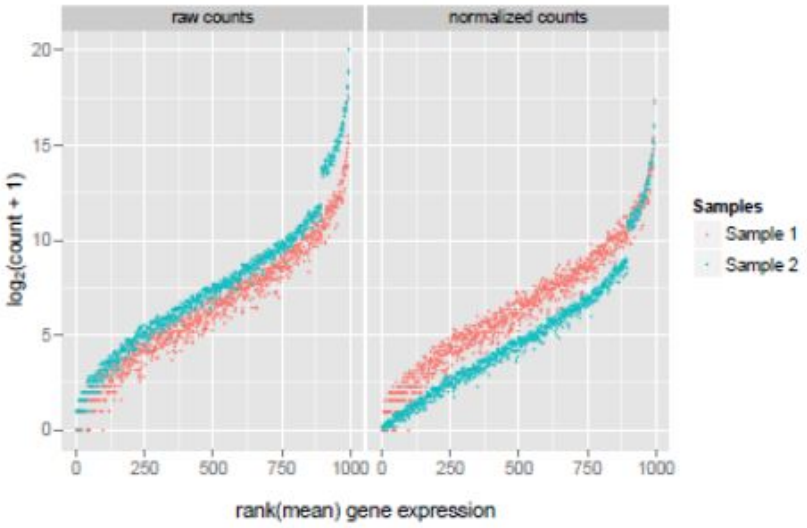

```
## With edgeR
cpm(...,
normalized.lib.sizes=TRUE)
```

Méthodes basées sur les normalisations interbanques

Upper Quartile

Les comptages par gène sont divises par le 3e quartile des comptages non nuls de l'échantillon, puis multiplies 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)
```

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

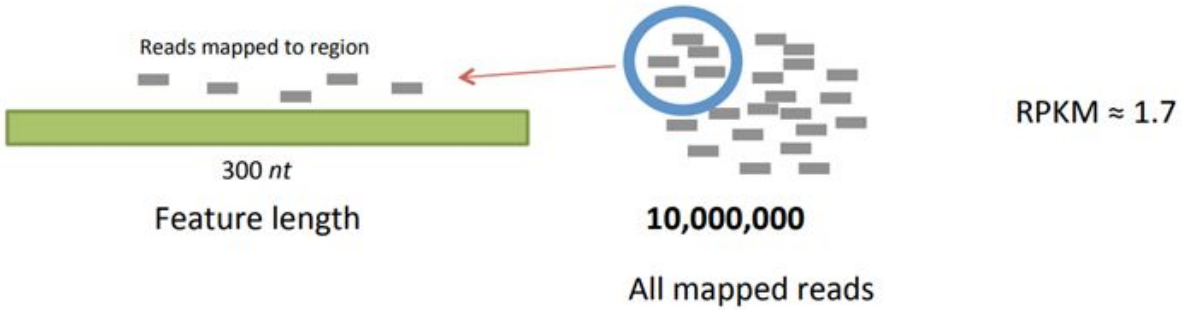
- méthode du Total Count peu efficace (pas de prise en compte des différences possibles entre les compositions en ARN des conditions)
- méthode RPKM peu efficace (même dans les cas où un biais lié à la longueur des gènes existe, l'utilisation du RPKM ne permet pas de le corriger complètement)
- méthodes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

Normalisation des données

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



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

RPKM: *R*eads *P*er *K*ilo base of transcript per *M*illion reads

Méthodes basées sur les normalisations interbanques

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 exprimes. Le facteur de normalisation pour un échantillon est obtenu en calculant pour chaque gène la médiane des ratio de ses comptages par rapport a sa moyenne géométrique entre les différents échantillons. L'idée sous-jacente est que les gènes non différentiellement exprimes 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 exprimes, la médiane des ratio constitue une estimation du facteur correctif qui doit être applique a l'ensemble des comptages.

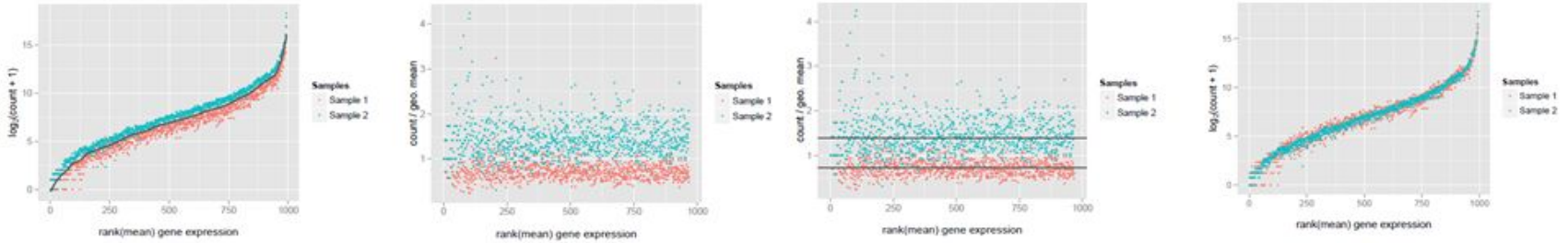
Ref : Anders and Huber, 2010. Dans edgeR, DESeq – DESeq2

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(…, mrthod=“ RLE”)
## with DESeq
estimateSizeFactors(…)
```



Méthodes basées sur les normalisations interbanques

TMM

La normalisation TMM (Trimmed Mean of M-values) est implementée dans le package Bioconductor edgeR. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différentiellement exprimes. Le facteur TMM est calcule pour chaque échantillon, l'un d'eux étant considère 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 exprimes et des gènes ayant les plus forts log-ratios. D'après l'hypothèse selon laquelle il y a peu de gènes différentiellement exprimes, le facteur TMM doit être proche de 1. S'il ne l'est pas, sa valeur donne une estimation du facteur correctif à appliquer aux tailles des librairies (et pas aux comptages bruts) afin de rendre l'hypothèse vraie.

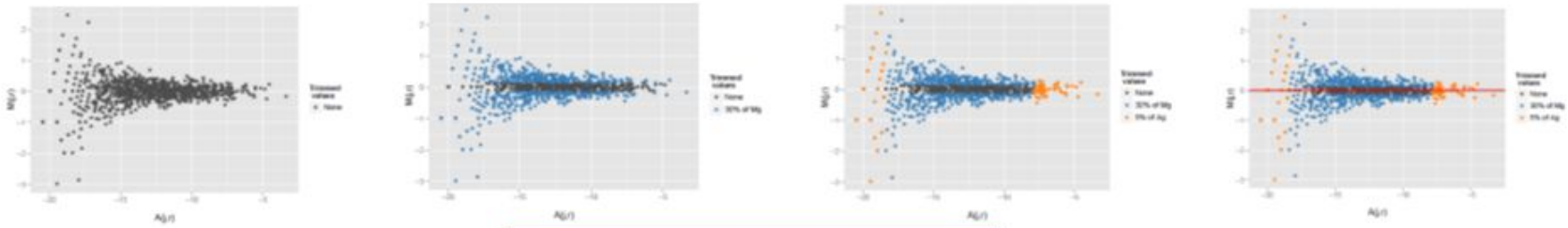
Ref : Robinson, M. and Oshlack, A. (2010). Dans edgeR.

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

2 – Trim 30% on M-values

3 – Trim 5% on A-values

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



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

Statistics tools

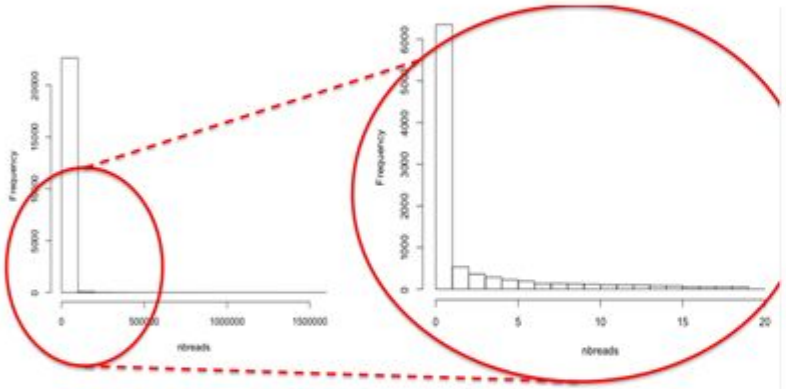
- **limma** (i.e., voom+limma and vst+limma)
  - unaffected by outliers
  - but they required at least 3 samples per condition
- **SAMseq, ShrinkSeq** (The non-parametric)
  - top performing methods for data sets with large sample sizes
  - required at least 4-5 samples per condition
  - fold change required for statistical significance was lower → compromise the biological significance
  - Small sample sizes inaccuracies in the estimation of the mean and dispersion parameters
- **TSPM**
  - most affected by the sample size
- **DESeq, edgeR and NBPSeq**
  - showed, overall, relatively similar accuracy with respect to gene ranking
  - recommended parameters well chosen and often provide the best results
  - pre-specified FDR threshold varied considerably between the methods
  - **DESeq** : overly conservative
  - **edgeR**, **NBPSeq** : too liberal and called a larger number of false (and true) DE genes.
  - **edgeR**, **DESeq :** varying the parameters 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 fot small sample sizes
  - **baySeq** In the presence of outliers, found a lower fraction of false positives true for small sample sizes not fot large sample sizes

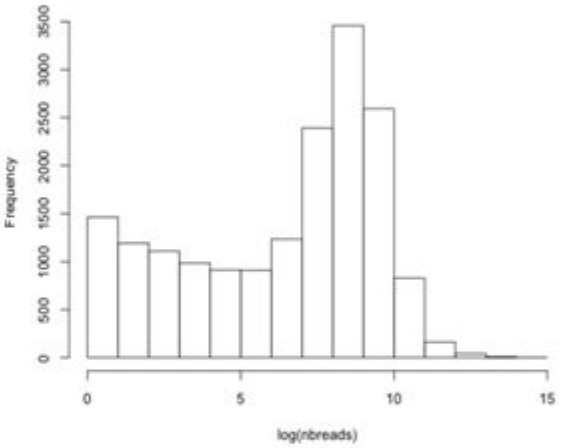
A comparison of methods for differential expression analysis of RNA-seq data. Charlotte Soneson1 and Mauro Delorenzi

5- 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

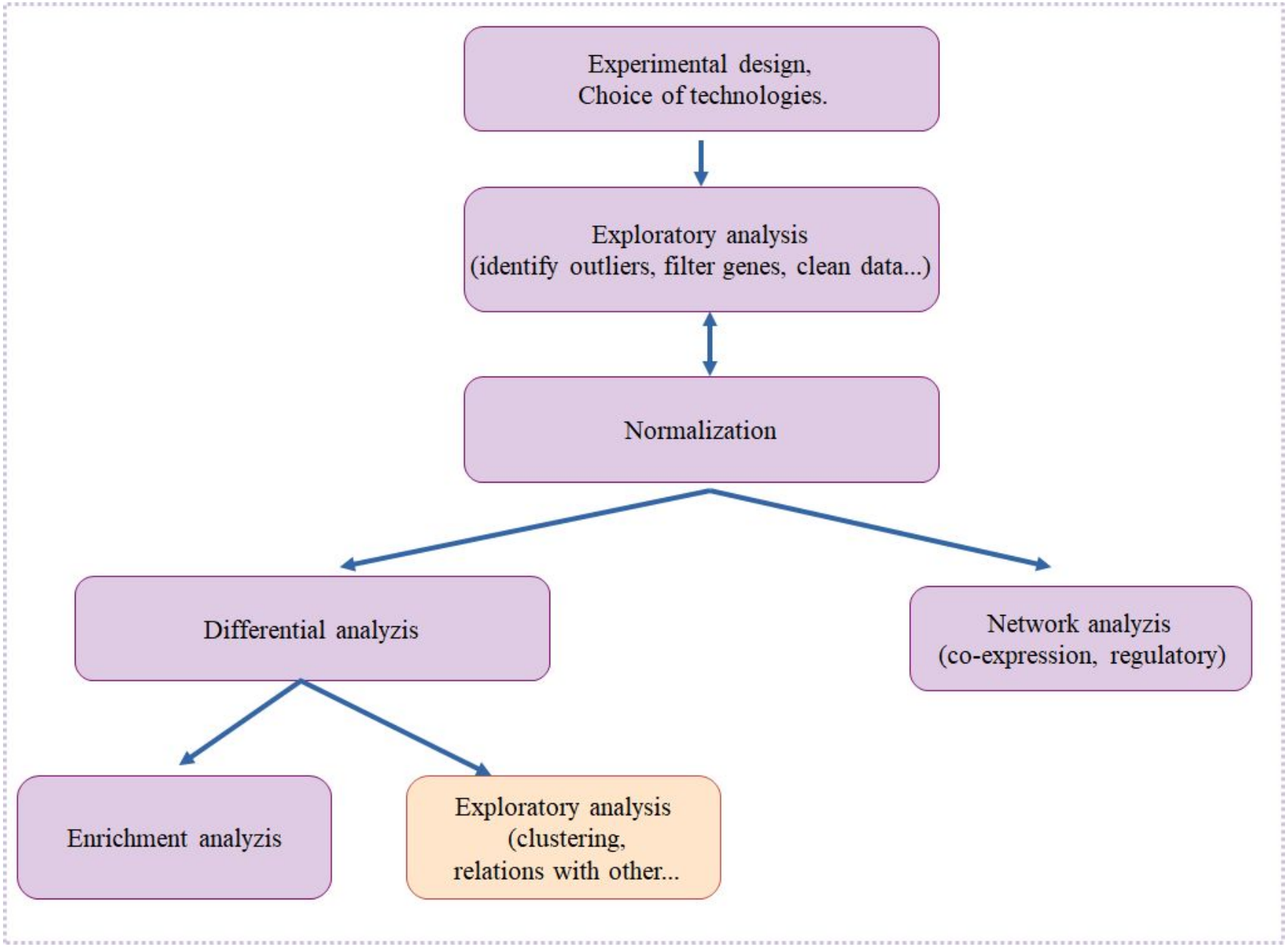
=> besoin de correction et utiliser une pvalue ajustée adaptée aux tests multiples:

- 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)

6- Plots and Graphical Représentations

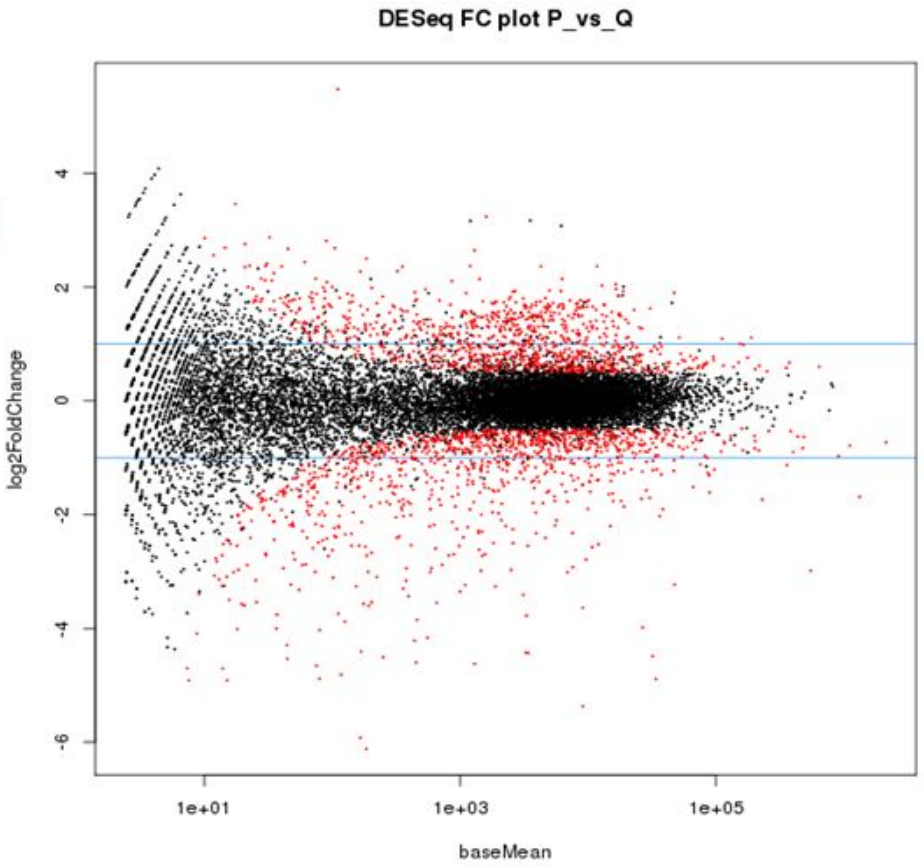
Analyses exploratoires



Smear plot

Smear plot / MA plot

Pvalue adj < 0.05



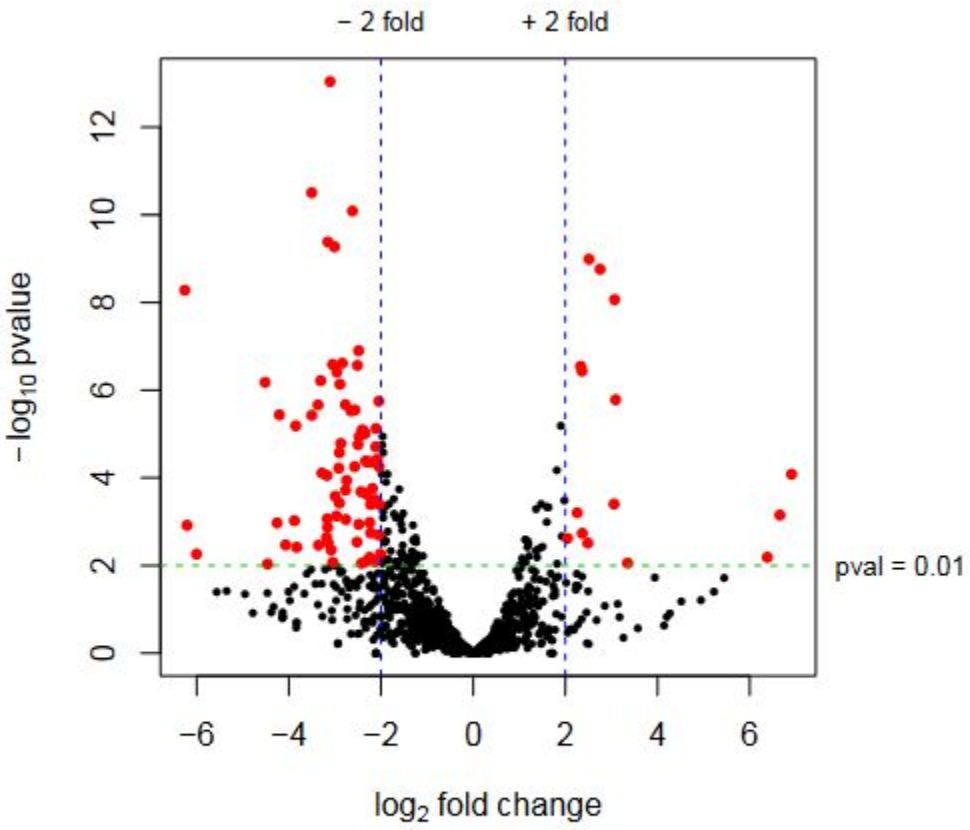
MA plot

Le MA plot est un graphe qui était initialement utilise dans les analyses de puce a 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 a l'autre. En RNA-Seq, après normalisation, on s'attend a ce que les points soient repartis symétriquement autour de 0 en ordonnée (c'est-a-dire un ratio de 1).

Volcano plot

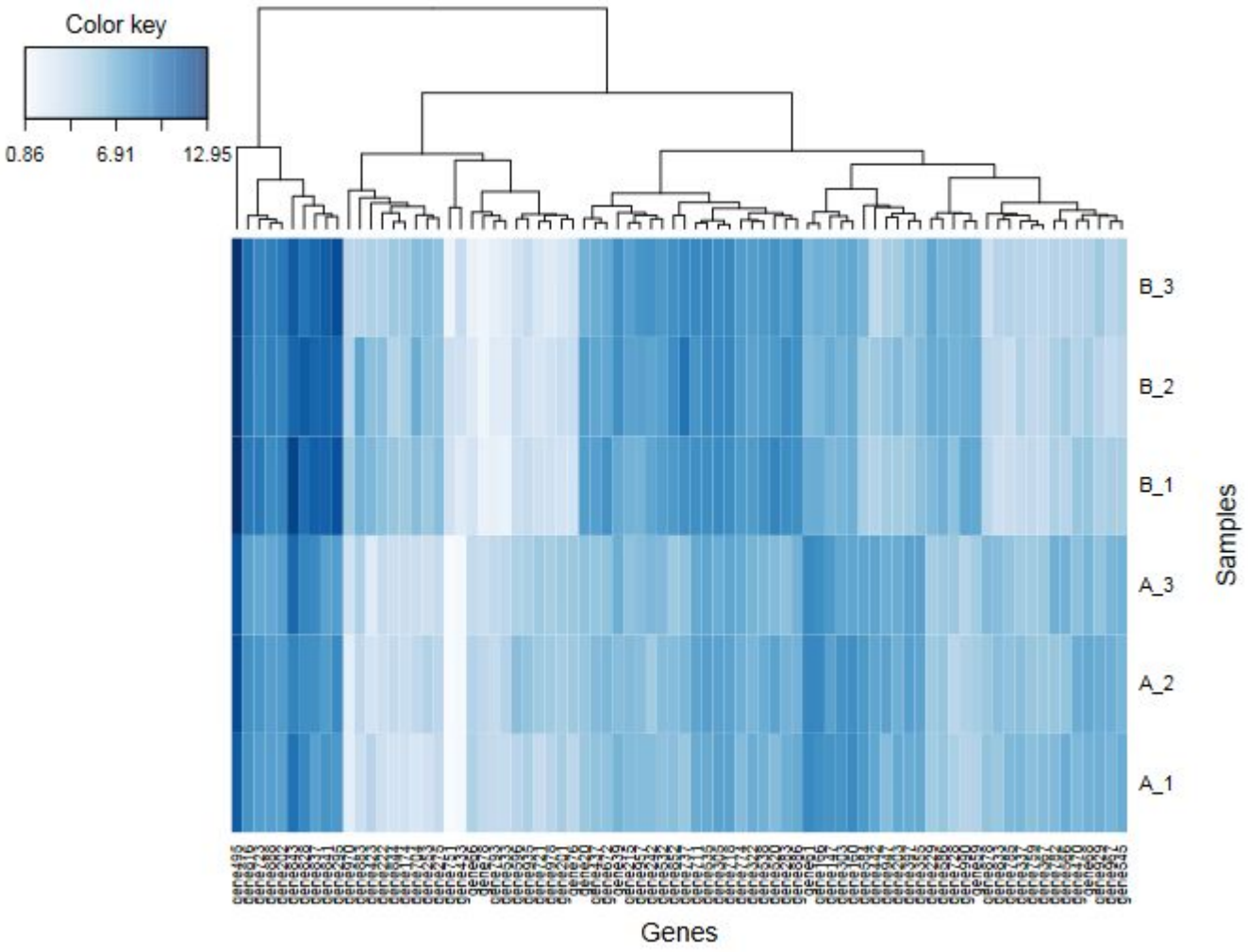
Volcano plot

Pvalue adj < 0.01



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

Hierarchical clustering / Heatmap



Practice : R et PIVOT

TP Pivot

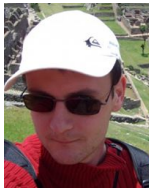
Contributeurs pour cette formation

Alexis Dereeper

Sebastien Ravel

Christine Tranchant-Dubreuil

Sebastien Cunnac



Gautier Sarah

Julie Orjuela-Bouniol

Catherine Breton

Aurore Compte

Merci pour votre attention !