

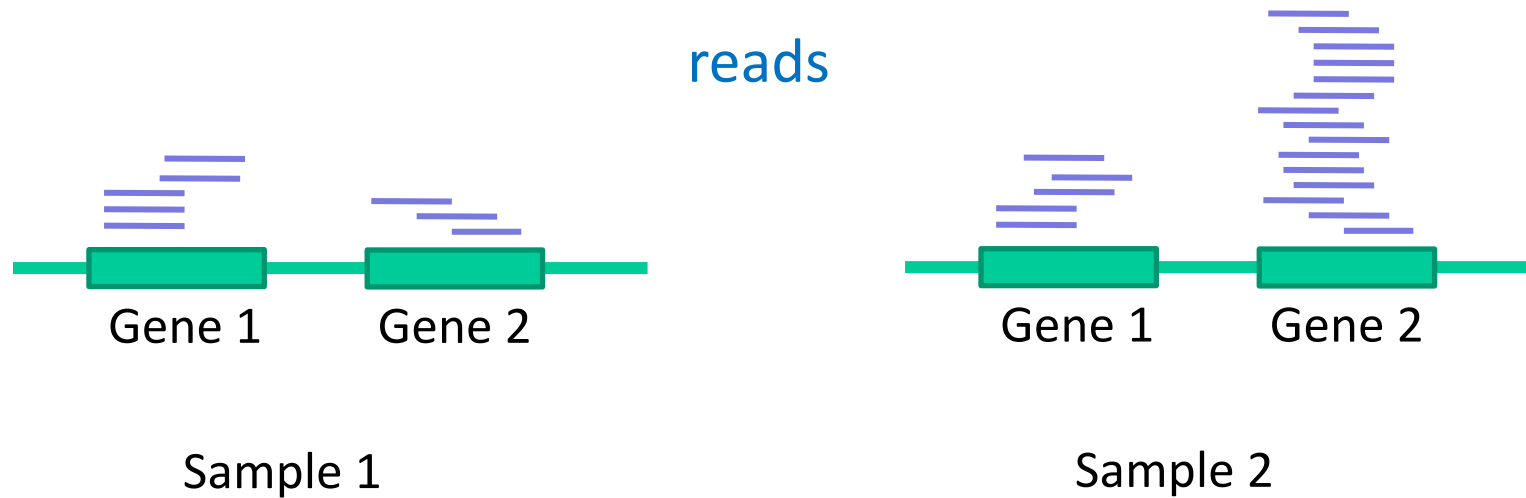
IFSBM « Big Data »

Introduction à l'analyse de transcriptome par RNA-seq

Daniel Gautheret

Avec des diapos de Gaëlle Lelandais, U. Paris-Saclay

Le RNA-seq sert d'abord à mesurer l'expression

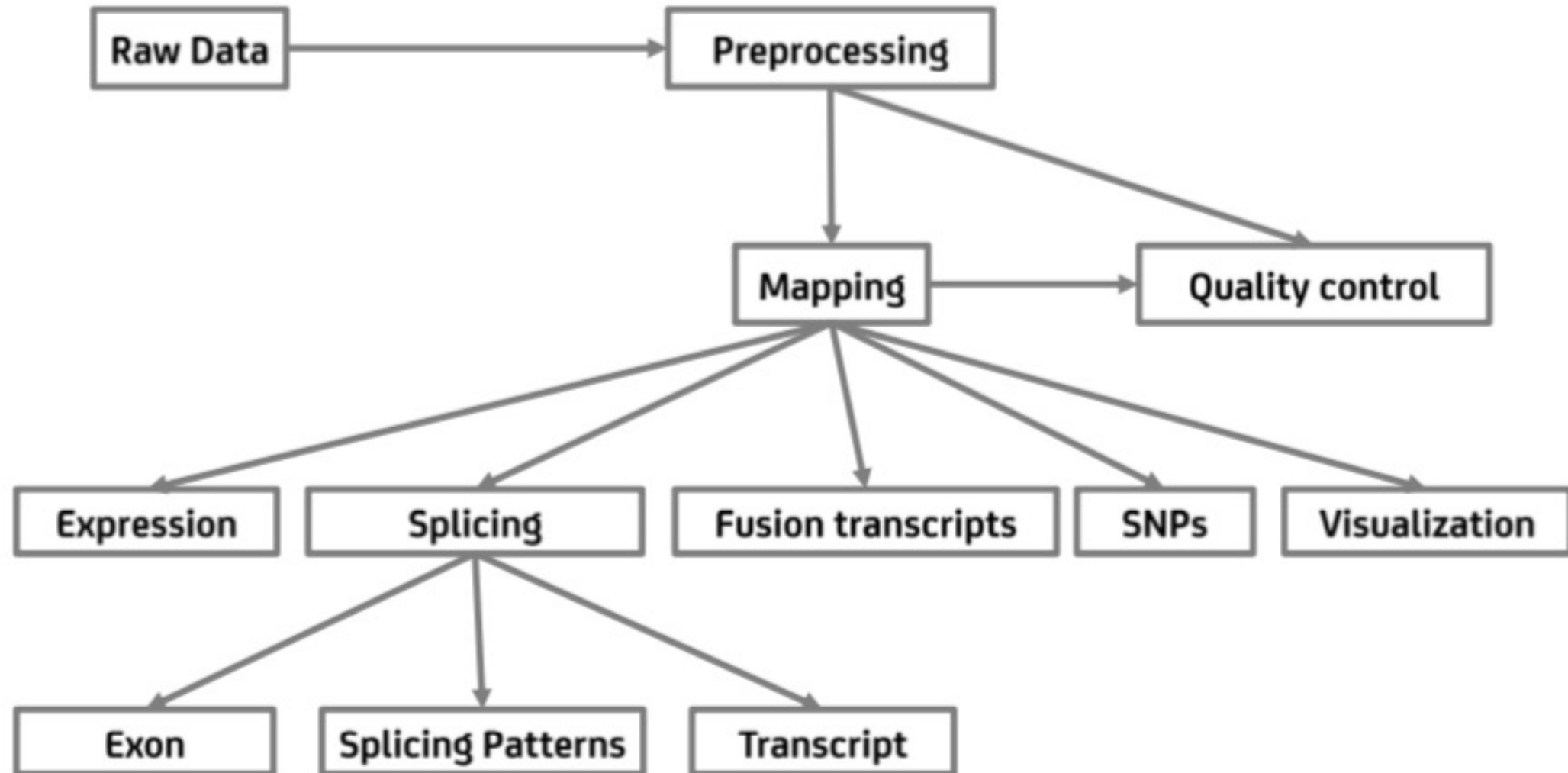


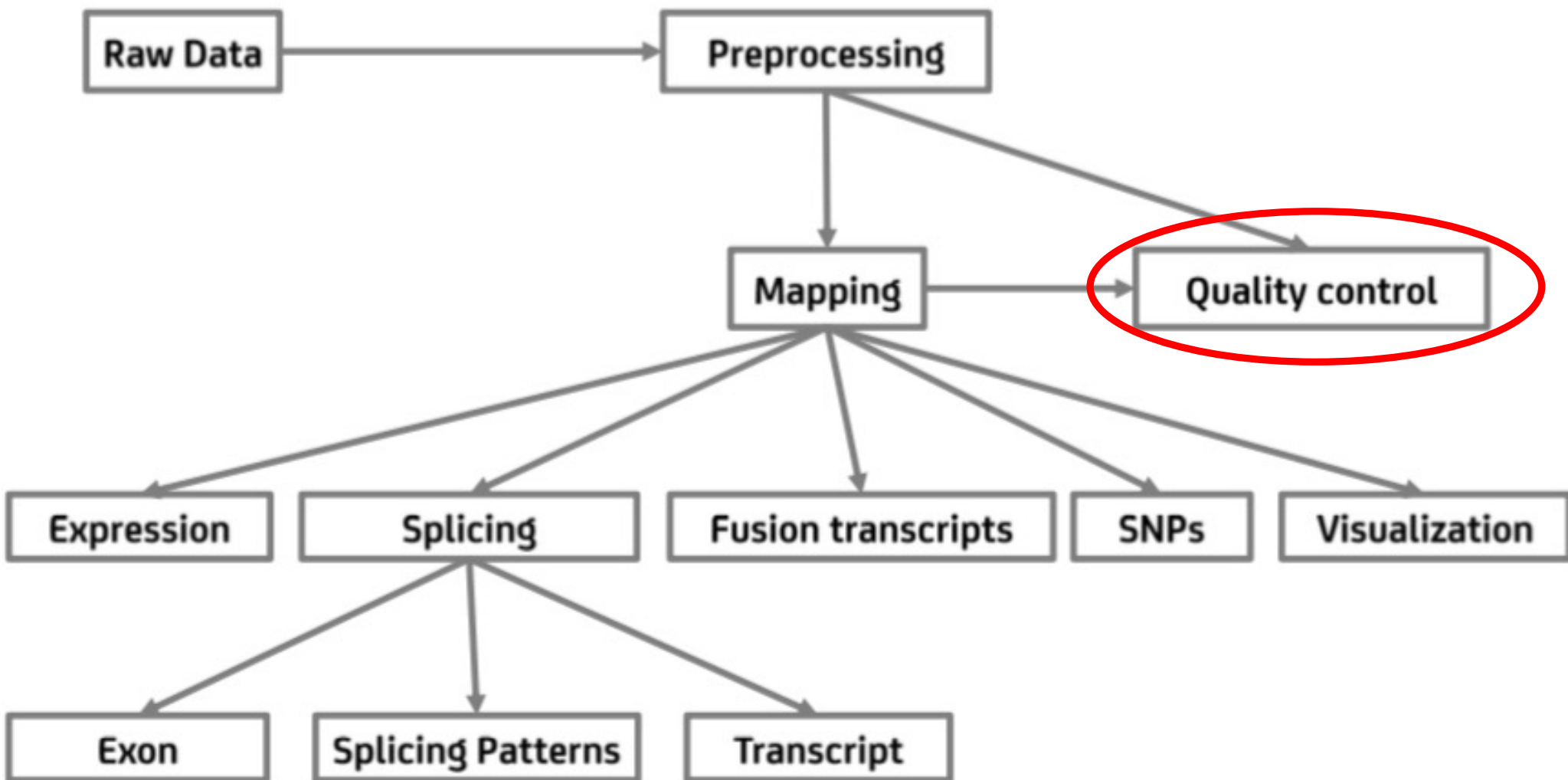
Mais pas que

- Mesurer l'expression des gènes
- Mesurer l'épissage alternatif
- Détecter les mutations exprimées
- Annoter les gènes: nouveaux exons
- Détecter les transcrits de fusion



Un pipeline d'analyse RNA-seq





FASTQ Format

CCCGAGCTCAGCCCCGACAGATCAGGACCGTGAGGATGGGGTCATGGCGTCTCTCCCCTGCCCTGCTCTGTG
+
AAAA6AEAEFAEEAEEEEEEEEEEEEEEEE6EEE6/AEEE/EEEEEEEEE6EEEEEEEEEEEEEEEEEEEE6EEEE/E
@NB501949:31:H2NWHBGX3:1:11101:7216:7526 1:N:0:AGTTCC
GTTTGTTTGTTTTTGAACAGGGTATTGCTCTGTCTCATCCAGGCCAGAGTGTAGTGGCGTGATCACCACTCACTGC
+
AAAAAEE/EA/EEEE
@NB501949:31:H2NWHBGX3:1:11101:14260:7526 1:N:0:AGTTCC
GCCAGGCATAGGCTACCCAGTGGTTCTCAAAGTGTCTCTTGGATCAGCAGCAGCAGCATACCGGGGATGGA
+
AAAAAEE
@NB501949:31:H2NWHBGX3:1:11101:22341:7527 1:N:0:AGTTCC
CCCACCACCAGAAATGAACAAAAGCATTTTACCTAAAAATACACCAGCAAATGTACTCAGCTTCAATCACAAAT
+
AAAAAEE
@NB501949:31:H2NWHBGX3:1:11101:22098:7527 1:N:0:AGTTCC
GCCGAAGCCACTCCACTGTCTCAGCATTTCATTGACTTGAAAAAGTCCTTGTTGCTCCAGACCTCCGTGTTAGCC
+
AAAAAEE
@NB501949:31:H2NWHBGX3:1:11101:8707:7528 1:N:0:AGTTCC
GTCTTGAGGACCTCTGTGTATTTGTCAATTTCTTCTCCACGTTCTTCTCGGCTGTTTCCGTAGCCTCATGAGCT
+
AAAAAEE
@NB501949:31:H2NWHBGX3:1:11101:4370:7528 1:N:0:AGTTCC
GGCCACTGCACCCAGCTTTTATCGTGTTTTGTGCACTACTGTAAACCTTGAATAACACCATGGGGCCCATACGA
+
AAAAAEEEE6EE

Sequence quality encoding: Phred score

Phred scores Q : Q scores are defined as a property that is logarithmically related to the base-calling error probabilities (P).

$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Encodage de la qualité transformation des scores en lettres

@SEQ_ID1
GATTTGGGGT
+
42,40,134,47,36 35 40 40 40



@SEQ_ID1
GATTTGGGGT
+
*(å/\$#((((

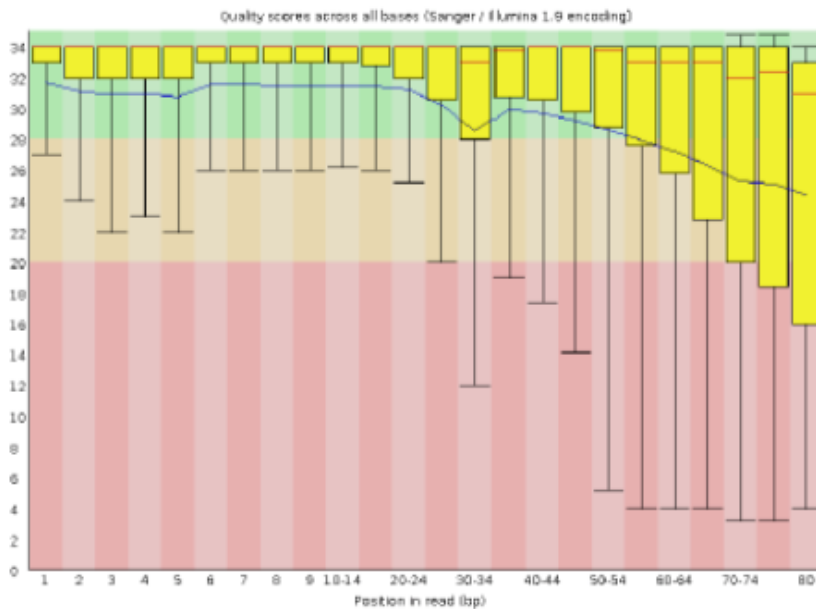


0	32	64	96	128	160	192	224
1	33	65	97	129	161	193	225
2	34	66	98	130	162	194	226
3	35	67	99	131	163	195	227
4	36	68	100	132	164	196	228
5	37	69	101	133	165	197	229
6	38	70	102	134	166	198	230
	39	71	103	135	167	199	231
	40	72	104	136	168	200	232
	41	73	105	137	169	201	233
	42	74	106	138	170	202	234
11	43	75	107	139	171	203	235
12	44	76	108	140	172	204	236
13	45	77	109	141	173	205	237
14	46	78	110	142	174	206	238
15	47	79	111	143	175	207	239
16	48	80	112	144	176	208	240
17	49	81	113	145	177	209	241
18	50	82	114	146	178	210	242
19	51	83	115	147	179	211	243
20	52	84	116	148	180	212	244
21	53	85	117	149	181	213	245
22	54	86	118	150	182	214	246
23	55	87	119	151	183	215	247
24	56	88	120	152	184	216	248
25	57	89	121	153	185	217	249
26	58	90	122	154	186	218	250
27	59	91	123	155	187	219	251
28	60	92	124	156	188	220	252
29	61	93	125	157	189	221	253
30	62	94	126	158	190	222	254
31	63	95	127	159	191	223	255

FASTQC: Pass or Fail?

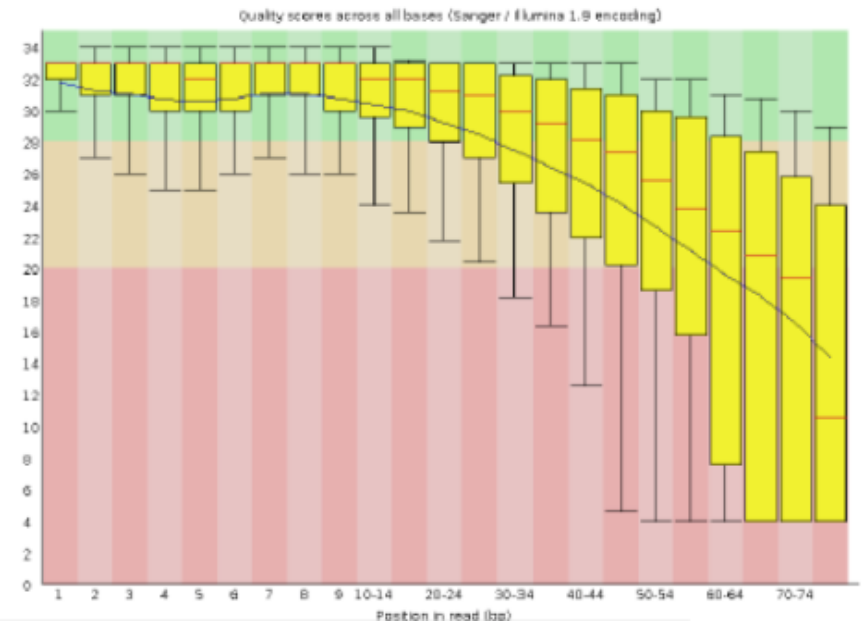
CG10128_RNAi_2

PASS

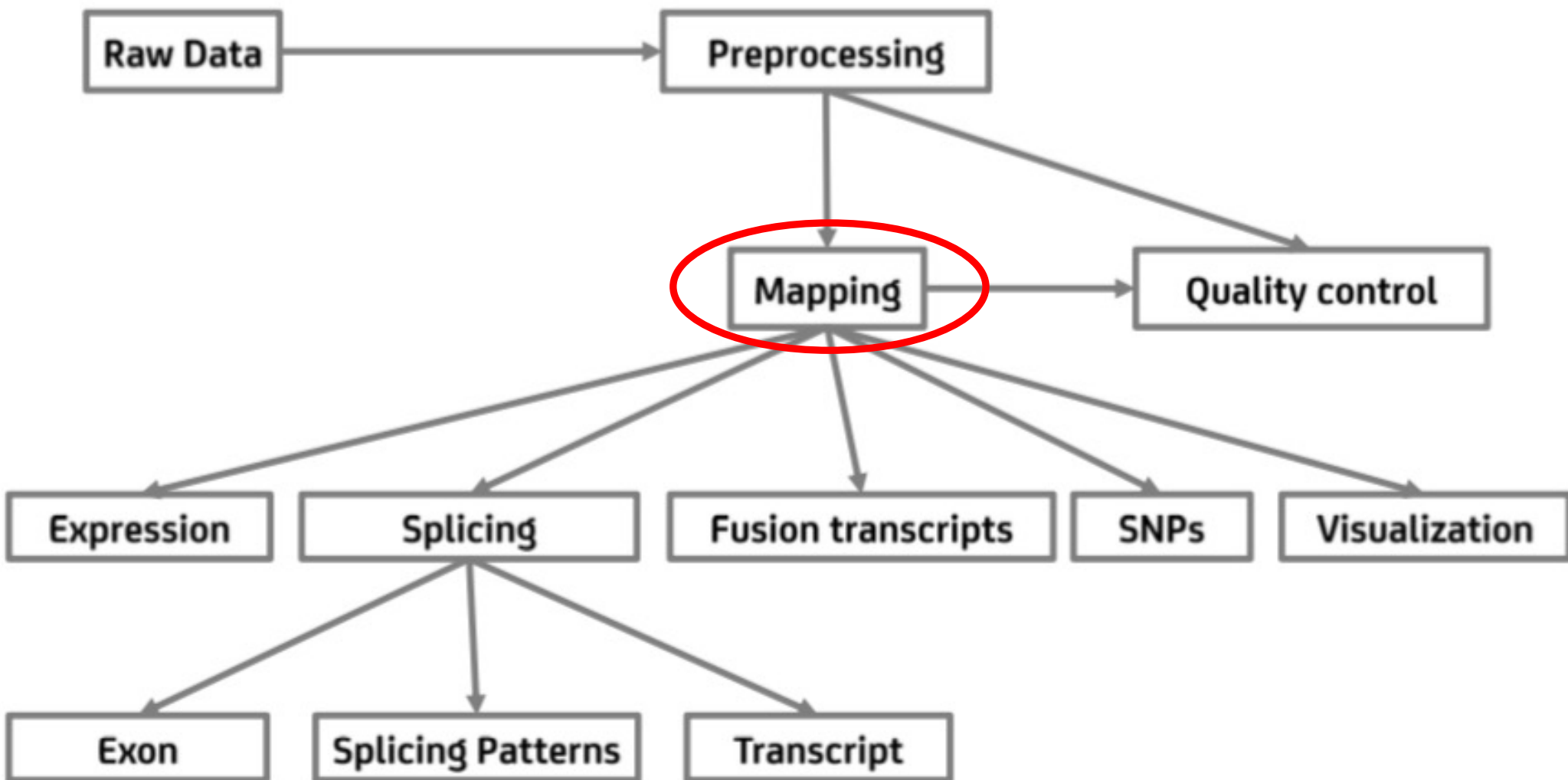


CG10203_RNAi_2

FAIL

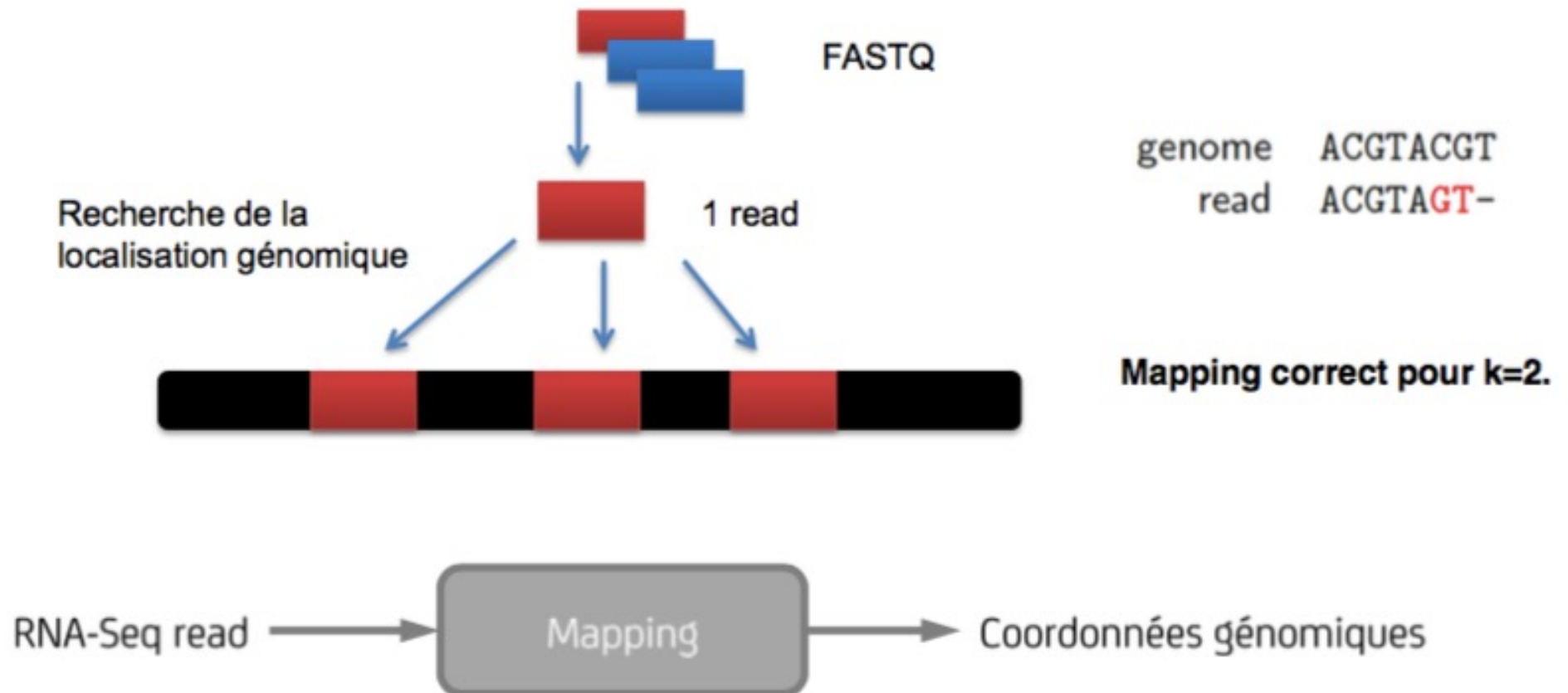


Alternative: trimming low quality bases: Cutadapt, FastqTrimmer, Trimmomatic



Mapping

Mapper=trouver tous les loci où le read est présent à k erreurs près.



Algorithme en $O(mn)$

ACGTTACCGAATCGATCAAGTCTGA
TAC

Algorithme en $O(mn)$

ACGTTACCGAATCGATCAAGTCTGA
TAC

Algorithme en $O(mn)$

ACGTTACCGAATCGATCAAGTCTGA
TAC

Algorithme en $O(mn)$

ACGTTACCGAATCGATCAAGTCGA
TAC

Algorithme en $O(mn)$

ACGTTACCGAATCGATCAAGTCTGA
TAC



OK pour 1 read: $O(3 \cdot 10^9 \times 100)$
Mais pour 1^8 reads???

Indexed Matching

- Créer à l'avance un index de la base à interroger
- Rechercher la chaîne dans l'index

L'algorithme de BLAST

- Index de k-mots de la référence
- Recherche des k-mots de la query dans l'index
- Extension autour des k-mots par Smith-Waterman

Gestion problématique des mismatches dans les k-mers
Effet important de la taille de k



Suffix array

“GOOGOL”

Tableau trié de tous les suffixes
d’une chaîne de caractères

0 GOOGOL\$

1 OOGOL\$

2 OGOL\$

3 GOL\$

4 OL\$

5 L\$

6 \$



6 \$

3 GOL\$

0 GOOGOL\$

5 L\$

2 OGOL\$

4 OL\$

1 OOGOL\$



(6,3,0,5,2,4,1)

(index relativement peu
encombrant)

Suffix array trié

Toutes les occurrences d’une même chaîne sont regroupées.

Suffix arrays

Exemple: trouver la chaîne **GO**

6 \$
3 **GO**L\$
0 **GO**OGOL\$
5 L\$
2 OGOL\$
4 OL\$
1 OOGOL\$

Suffix array trié

Burrows-Wheeler Transform (BWT)

- Permet de compresser efficacement une séquence en maintenant les propriétés d'un suffix array pour la recherche

Voir videos de Ben Langmead pour

1. Création d'une BWT (et décodage)
2. Utilisation d'une BWT comme index de recherche (FM index)

1: https://www.youtube.com/watch?v=6BJbEWyO_N0

2: <https://www.youtube.com/watch?v=kvVGj5V65io>

Les principaux logiciels de mapping

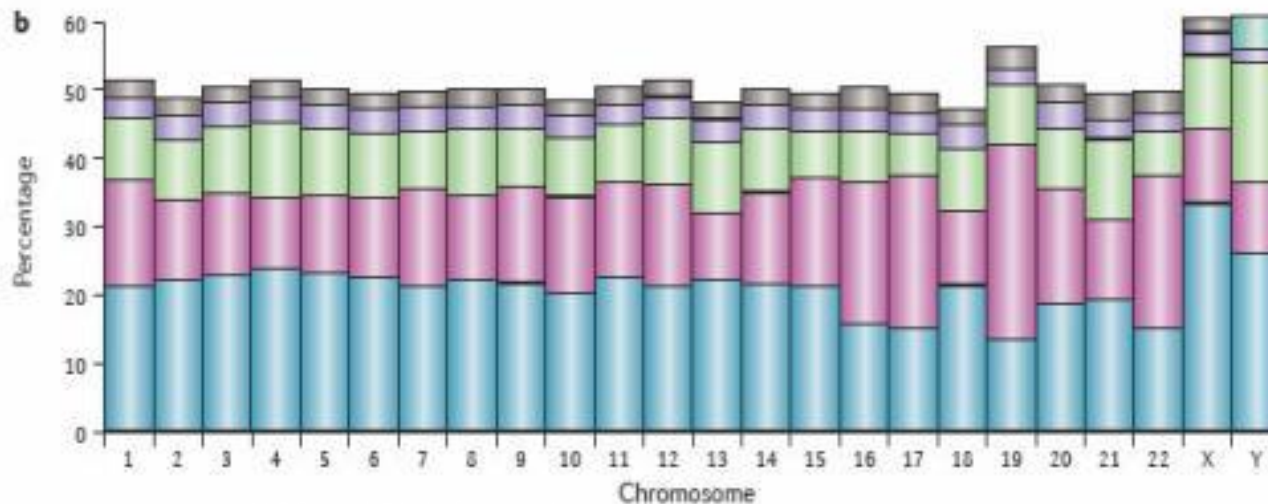
- Génomique
 - BOWTIE
 - BWA
- Transcriptomique (avec introns)
 - HiSAT
 - STAR

Le problème des répétitions

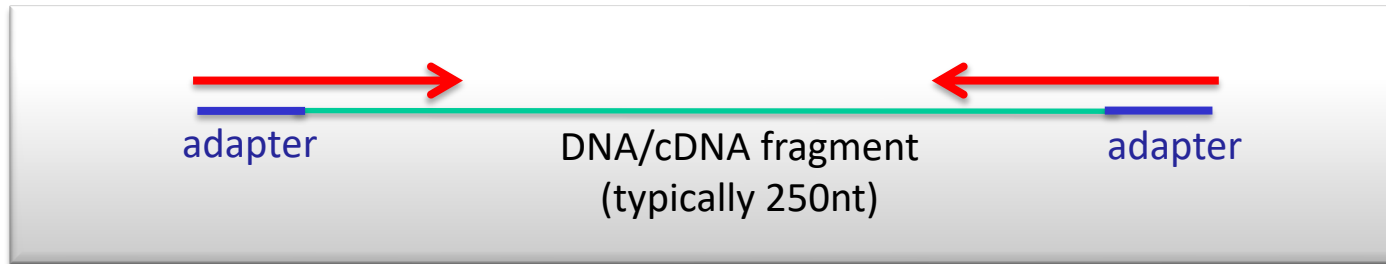
Approximately **50%** of the human genome is comprised of repeats

a

Repeat class	Repeat type	Number (hg19)	Cvg	Length (bp)
Minisatellite, microsatellite or satellite	Tandem	426,918	3%	2–100
SINE	Interspersed	1,797,575	15%	100–300
DNA transposon	Interspersed	463,776	3%	200–2,000
LTR retrotransposon	Interspersed	718,125	0%	200–5,000
LINE	Interspersed	1,506,845	21%	500–8,000
rDNA (16S, 18S, 5.8S and 28S)	Tandem	698	0.01%	2,000–43,000
Segmental duplications and other classes	Tandem or interspersed	2,270	0.20%	1,000–100,000



Séquençage paired-end vs. single end.



Benefits of paired-end sequencing

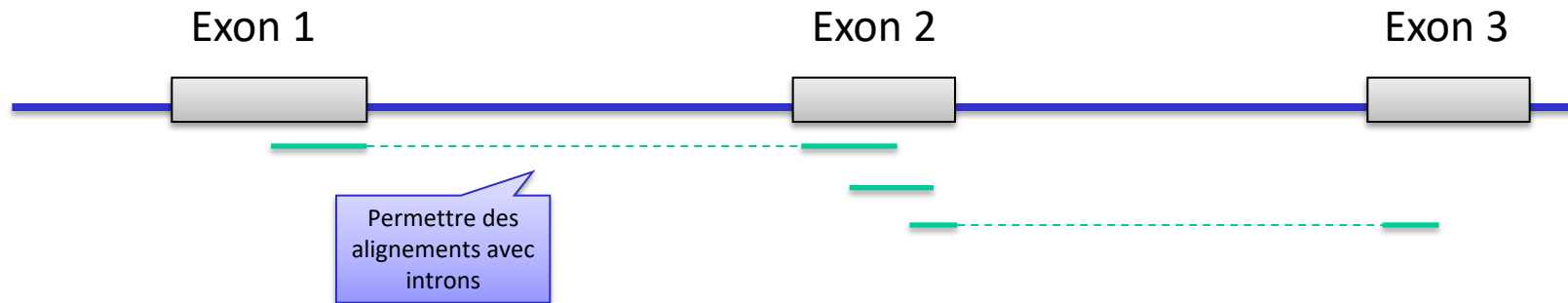
- Single-end alignment – repeated sequence



- Paired-end alignment – unique sequence

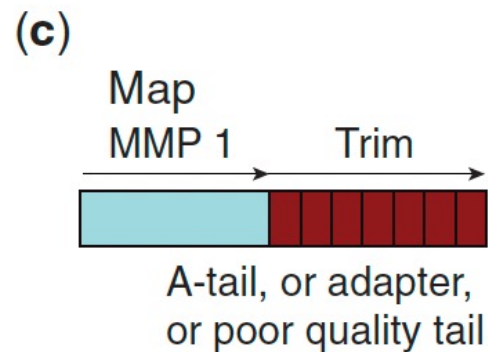
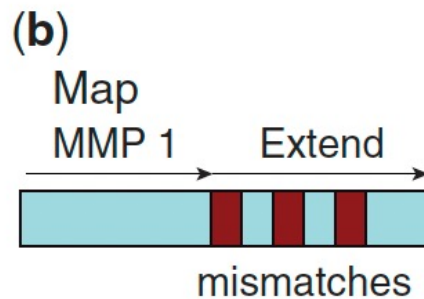
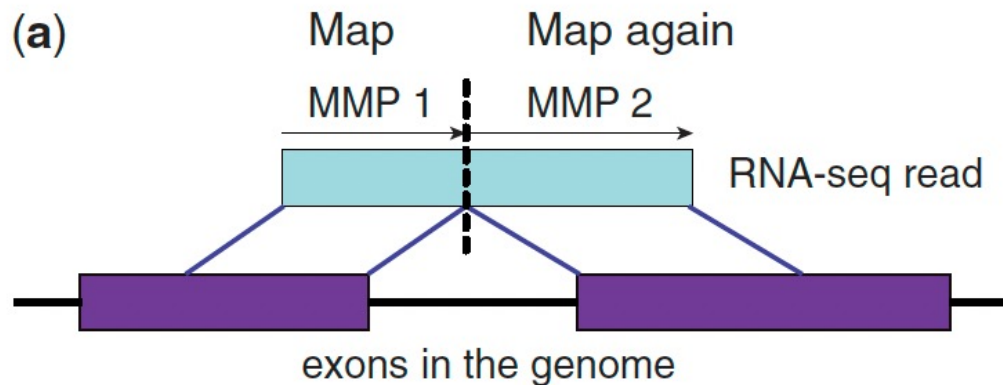


La spécificité du mapping RNA-seq



Le programme STAR

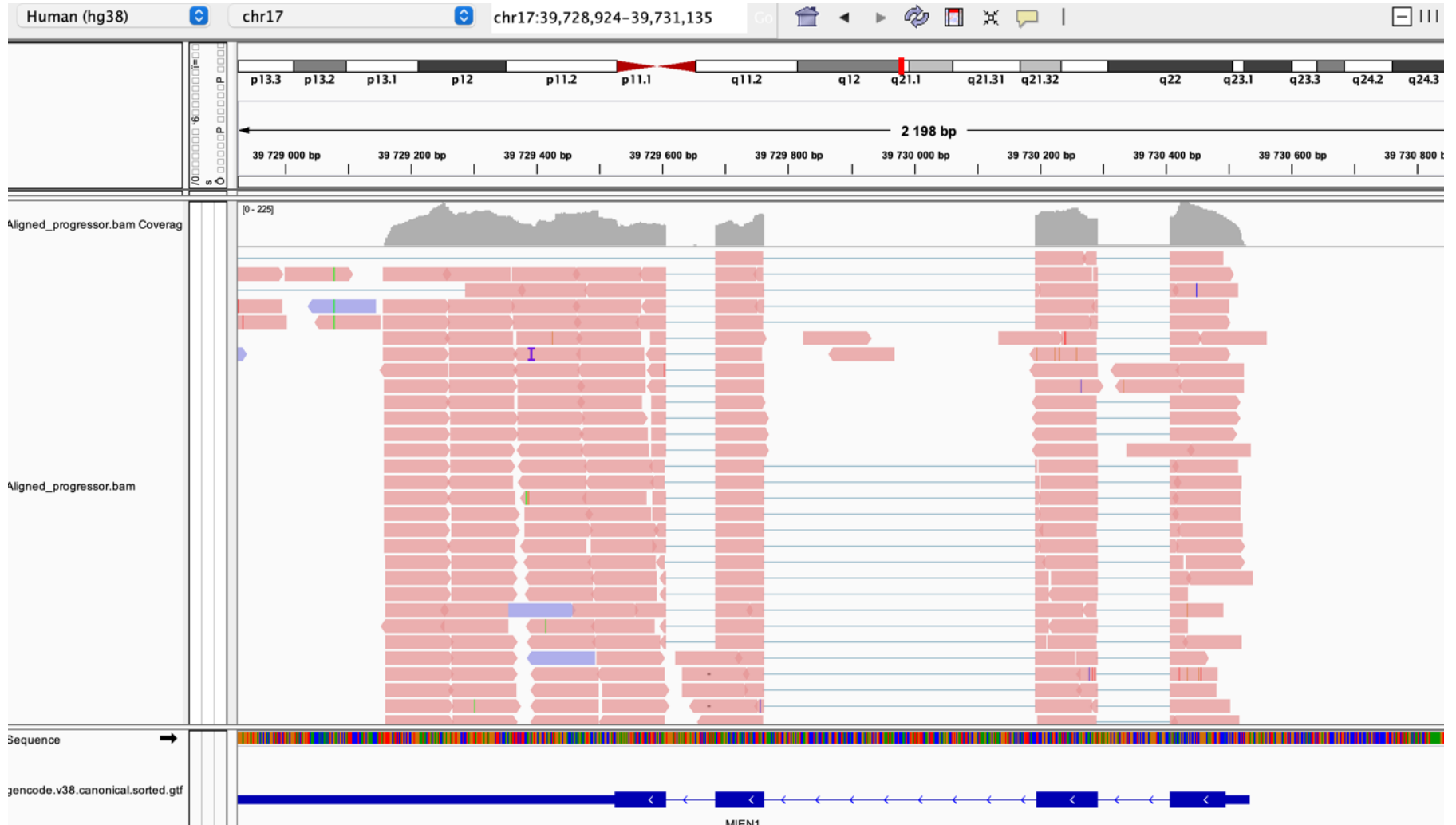
Dobin et al.
Bioinformatics, 2013



MMP=maximal
perfect match
(implemented
through
uncompressed suffix
arrays)

a,b,c: 3 cas possibles pour les reads alignés de manière incomplète.

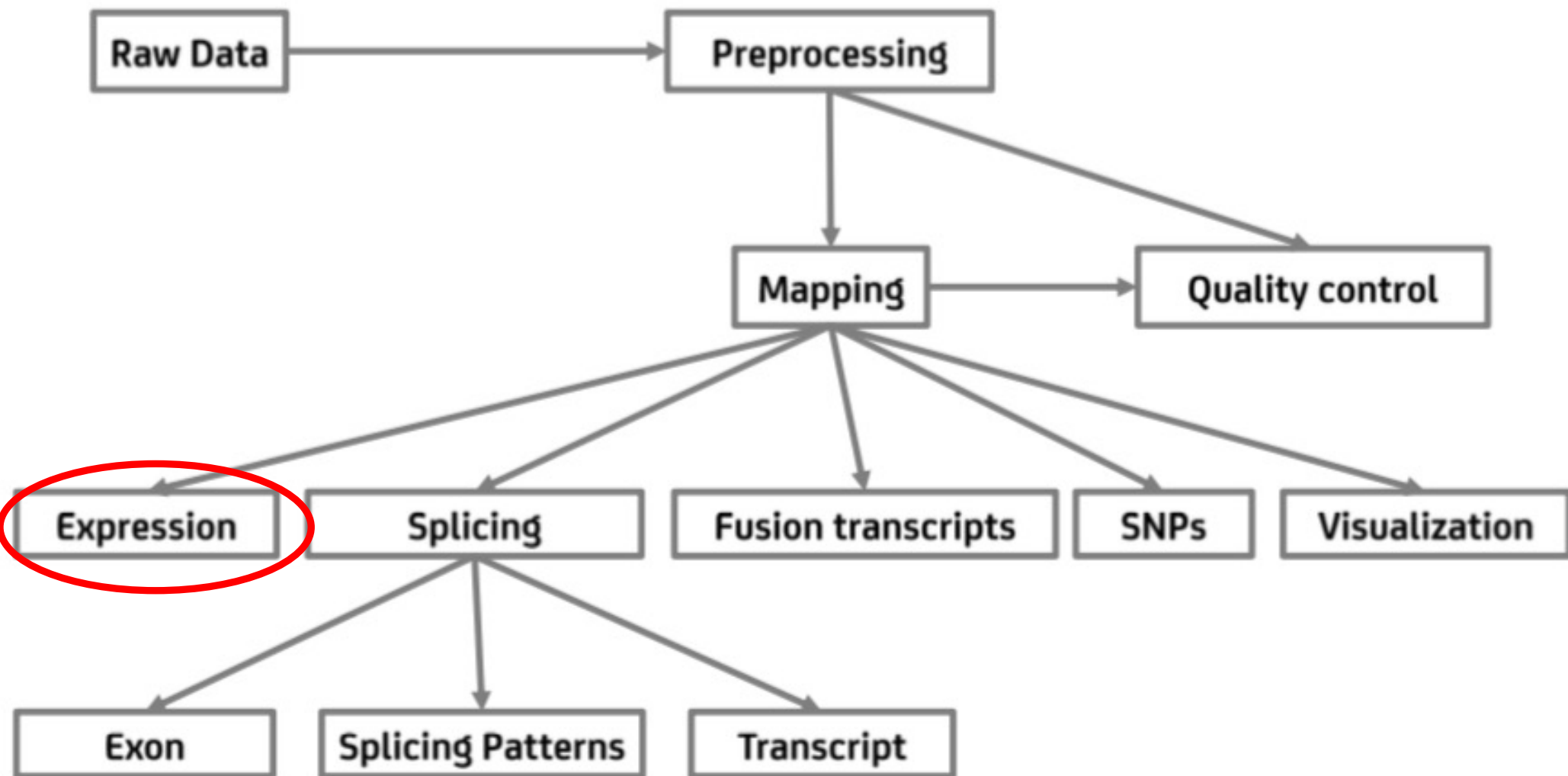
Reads alignés



Alignment formats

- SAM
- BAM
- PileUp

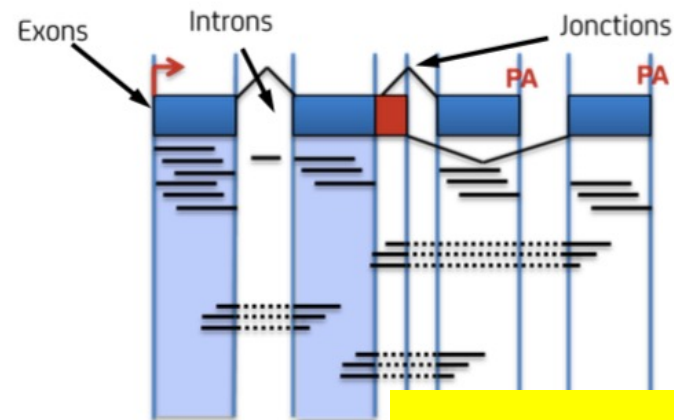
SAM and BAM are now the standard for aligned data
PileUp is used for variant calling



Les méthodes de comptage

Mesure d'expression avec featureCounts

featureCounts takes as input SAM/BAM files and an annotation file including chromosomal coordinates of features. It outputs numbers of reads assigned to features (or meta-features).



Liao Y, Smyth GK, Shi W.
Bioinformatics. 2014

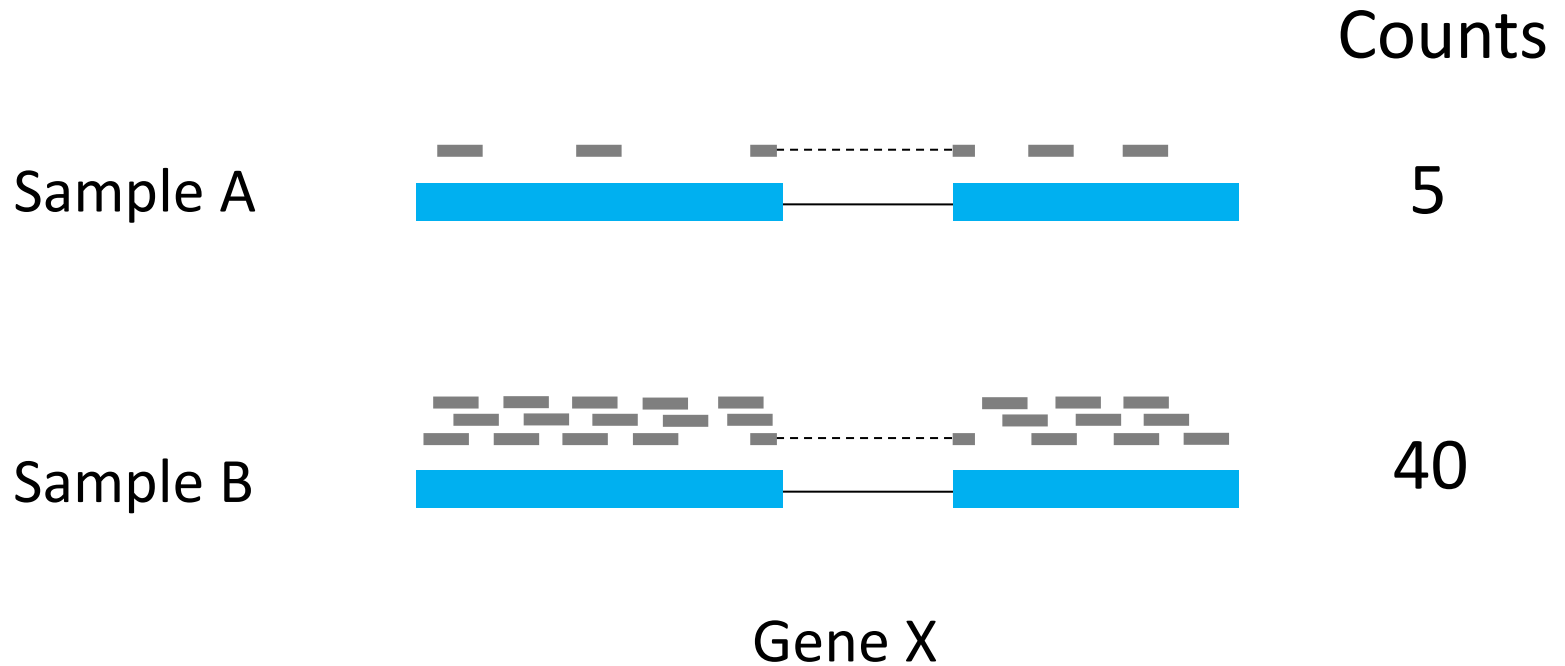
Voir format GTF

Indexer les fichiers BAM

- Pour connaître les reads alignés sur une région donnée, il faut indexer le fichier BAM
- Sans index, il faudrait parcourir tout le fichier pour répondre
- Indexation= tri par position + création d'une table des positions
- Produit un fichier **.BAI**

```
samtools sort sample.bam -o sample_sorted.bam  
samtools index sample_sorted.bam
```

Expression: résumé!



Main software for counting

Software	Reference	EM
Htseq-count	genome	-
FeatureCounts	genome	-
Cufflinks	genome	yes
RSEM	transcriptome	yes
Kallisto	transcriptome (pseudo mapping)	yes
Sailfish/Salmon	transcriptome (pseudo mapping)	yes

Analyse différentielle

Problématique de normalisation (1/2)

- Le nombre de sequences (*reads*) depend de la longueur des gènes ...

A1BG	4
A1CF	41
A2M	1
A2ML1	3
A2MP1	3
A3GALT2	1
A4GALT	420
A4GNT	1
AA06	0
AAAS	2452
AACS	3234
AACSP1	1544



8 *reads*
alignés



4 *reads*
alignés

Problématique de normalisation (2/2)

- Le nombre de sequences (*reads*) dépend de la “taille des librairies”* ...

A1BG	4	7
A1CF	41	32
A2M	1	4
A2ML1	3	6
A2MP1	3	1
A3GALT2	1	3
A4GALT	420	327
A4GNT	1	1
AA06	0	0
AAAS	2452	2054
AACS	3234	1678
AACSP1	1544	1926

Somme 1 Somme 2

Biais systématique ?

(si par exemple Somme 1 = 40 000
et Somme 2 = 30 000)

Table de comptage normalisée

* La somme des comptages par expérience est souvent nommé *library size* dans les articles.

Common expression units for genes and transcripts

- Raw counts: no normalization
- RPM,CPM (read/million, count/million)
 - Normalized by library size
- RPKM/FPKM (read/kb/million, fragment/kb/million)
 - Normalized by library size and gene-size
- TPM (transcript/million)
 - Normalized by transcript size and by millions of transcripts

Quantifier l'expression différentielle

➤ Ratio

$$R = \frac{Q_A}{Q_B} = \frac{\text{Valeur}(s) \text{ de comptage Condition A}}{\text{Valeur}(s) \text{ de comptage Condition B}}$$

➤ Fold Change

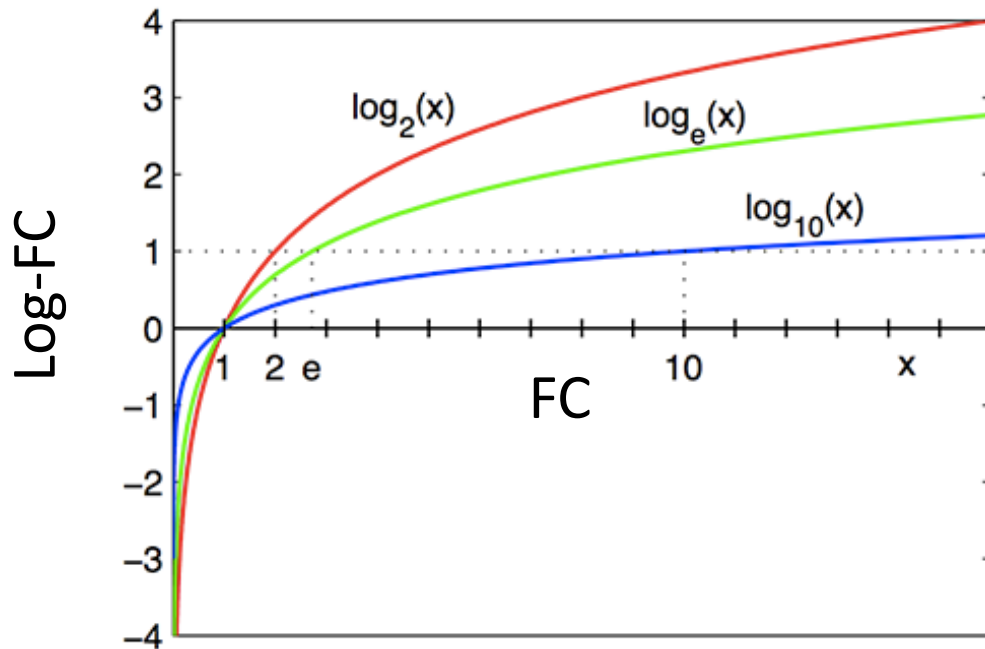
$$FC = \begin{cases} R & \text{si } R \geq 1 \\ \frac{1}{R} & \text{si } R < 1 \end{cases}$$

➤ Log fold change

$$\log FC = \log_2(R) = \log_2(Q_A) - \log_2(Q_B)$$

Interprétation du logFC (taille d'effet)

- Les valeurs du logFC peuvent être positives et négatives. L'utilisation de la base 2 permet de traduire un doublement par une unité de variation (+/-).



<https://fr.wikipedia.org/wiki/Logarithme>

$$\log FC_g > 0 \Leftrightarrow Q1_g > Q2_g$$

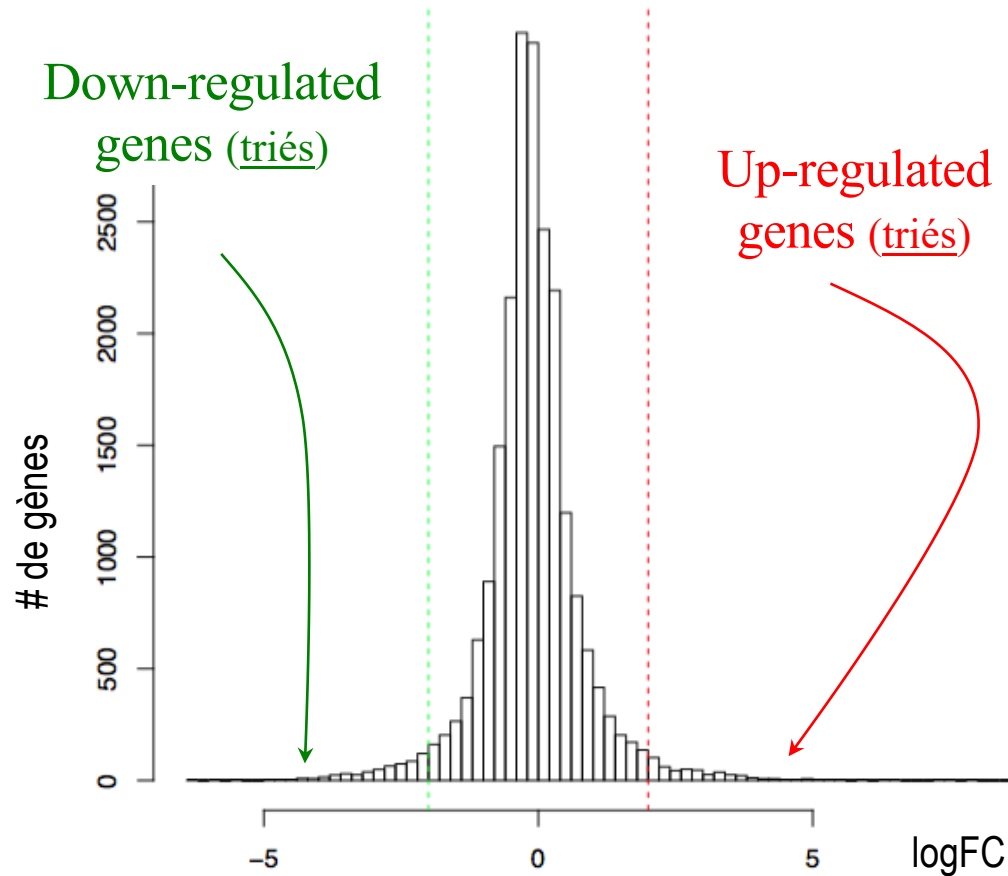
↳ “Up-regulated gene”

$$\log FC_g < 0 \Leftrightarrow Q1_g < Q2_g$$

↳ “Down-regulated gene”

Notion de taille d'effet

Utiliser la taille d'effet



Source des données : Yang et al., Mol Cell Biol (2016)

Liste de gènes, candidats
pour être différentiellement
exprimés

Tenir compte de la reproductibilité

(dispersion)

- Dans la pratique, les expériences sont répétées. La taille d'effet mesurée est fondée sur plusieurs observations.

$$\text{Gène 1: } \log FC_{G1} = \frac{1}{3}(0.5 + 2.5 + 6) = 3$$

$$++++ \quad \text{Gène 2: } \log FC_{G2} = \frac{1}{3}(3 + 2.8 + 3.2) = 3$$

Meilleure cohérence entre les observations

Notion de dispersion

Quantifier la dispersion*

➤ Variance (estimateur)

$$Var(X) = \frac{1}{n-1} \sum (x_i - m)^2$$

➤ Ecart type

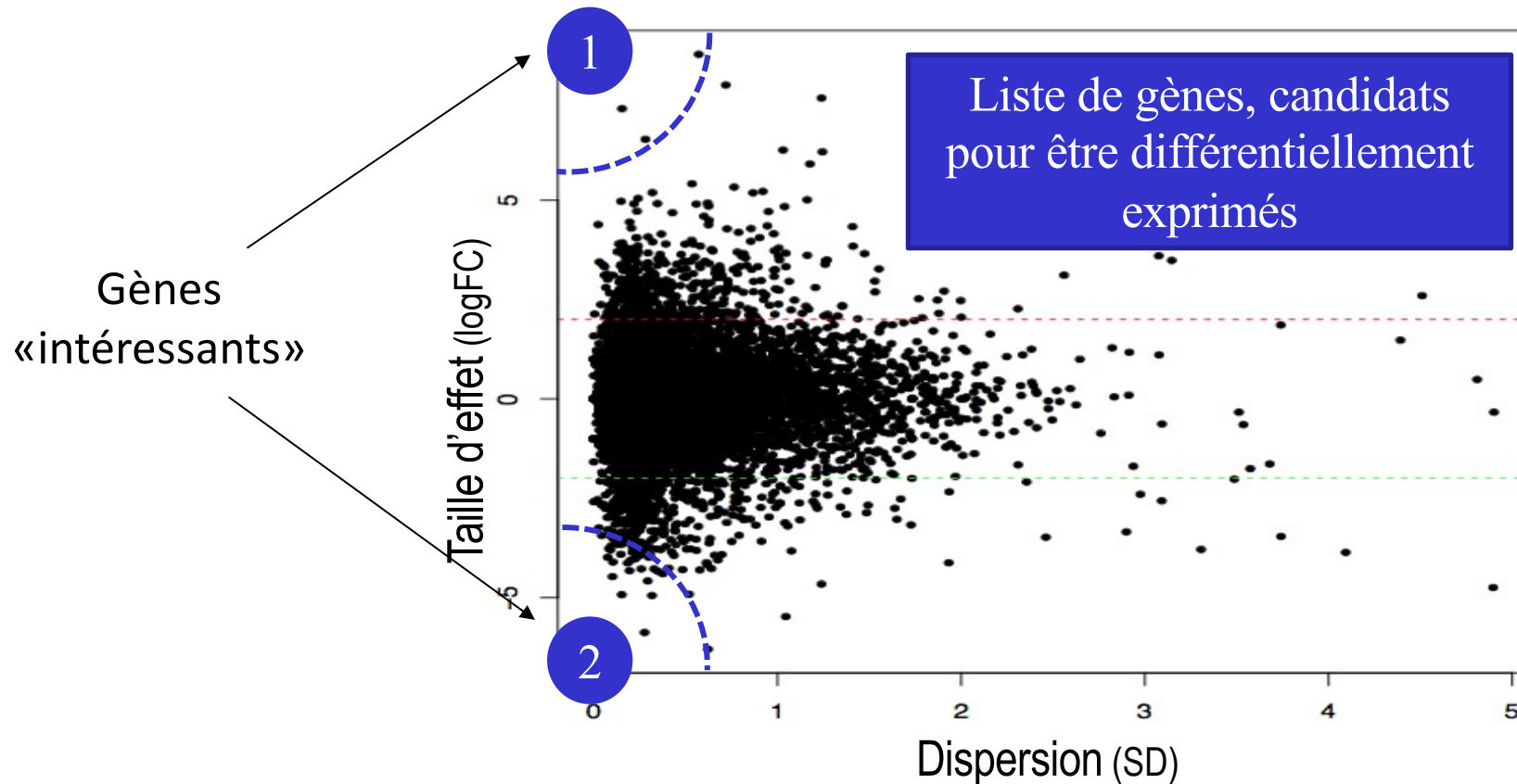
$$SD = \sqrt{Var(X)}$$

➤ Exemple :

	logFC	Variance	SD
Gène 1	3	7.75	2.78
→ Gène 2	3	0.04	0.20

* Mais aussi le coefficient de variation biologique, l'erreur standard, etc.

Combiner taille d'effet et dispersion

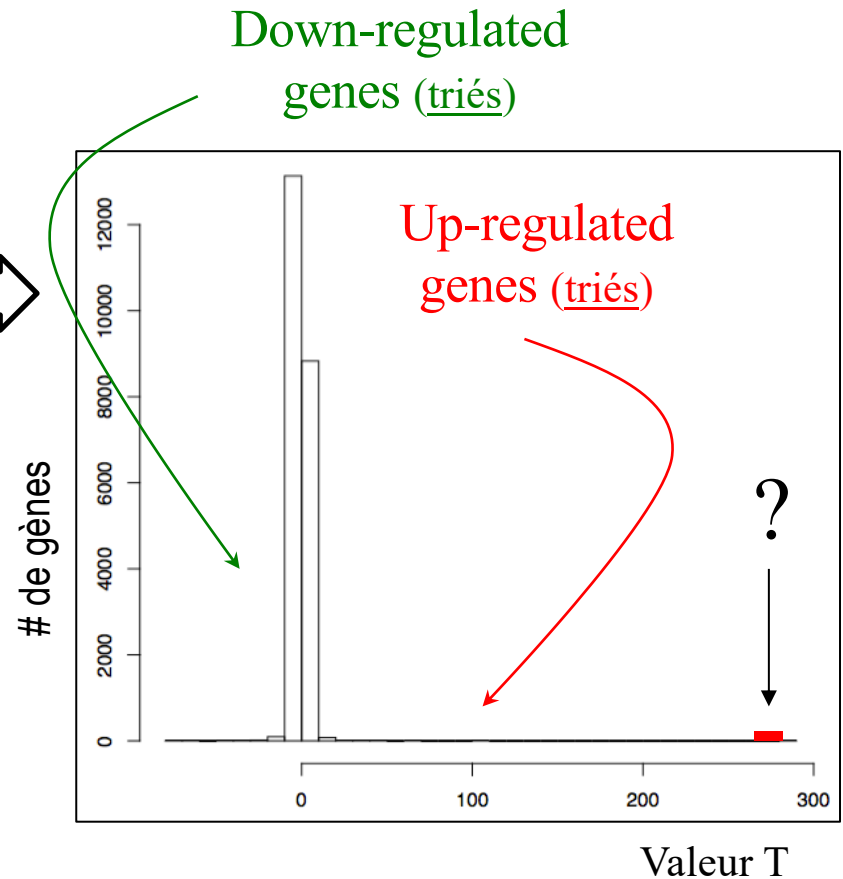
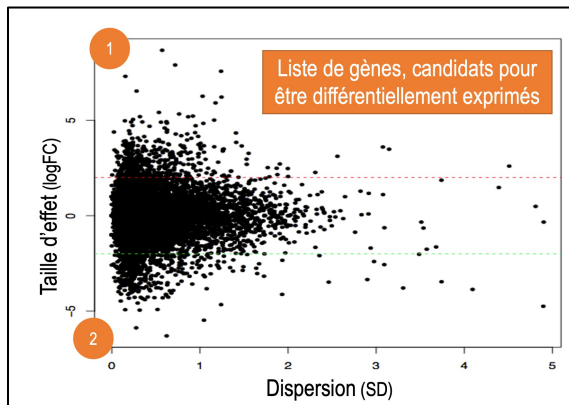


Source des données : Yang et al., Mol Cell Biol (2016)

Statistique de Student

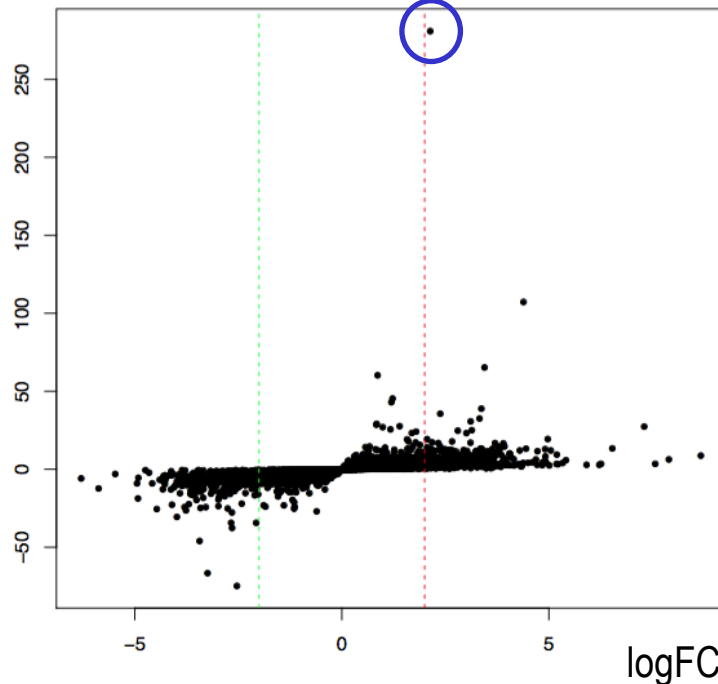
- Taille d'effet, dispersion, nombre d'observations :

$$t_g = \frac{\log FC_g}{\frac{SD_g}{\sqrt{n}}}$$



... certains gènes ont des valeur T artificiellement gonflées!

Valeur T



Importance trop grande donnée au paramètre de dispersion ($n = 3$)

- Estimer correctement la variabilité associée aux observations de comptage d'un gène est un **défi statistique**.

A1BG	4	6	2	7	6	7
A1CF	41	33	42	32	42	32
A2M	1	3	1	4	3	7
A2ML1	3	2	2	6	7	3
A2MP1	3	2	2	1	1	0
A3GALT2	1	4	4	3	2	1
A4GALT	420	344	291	327	360	371
A4GNT	1	1	2	1	3	3
AA06	0	0	0	0	0	0
AAAS	2452	2192	1977	2054	2134	2100
AACS	3234	2804	2609	1678	1670	1742
AACSP1	1544	1369	1300	1926	2015	1963

Source des données : Yang et al., Mol Cell Biol (2016)

Tenir compte du nombre de séquences alignées (force d'expression)

➤ Une même taille d'effet peut être observée pour des données de comptages différentes

Gene 1:

$$\log FC_{G1} = \log 2(8000 / 1000) = 3$$

Gene 2:

$$\log FC_{G2} = \log 2(8 / 1) = 3$$

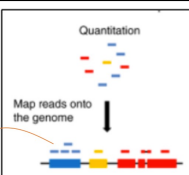
Variations aléatoires ?

La notion de “force d'expression” est intéressante à prendre en compte pour établir une liste de gène candidats, avec une meilleure fiabilité.

* Cette information est nommée *expression strength* dans les articles statistiques.

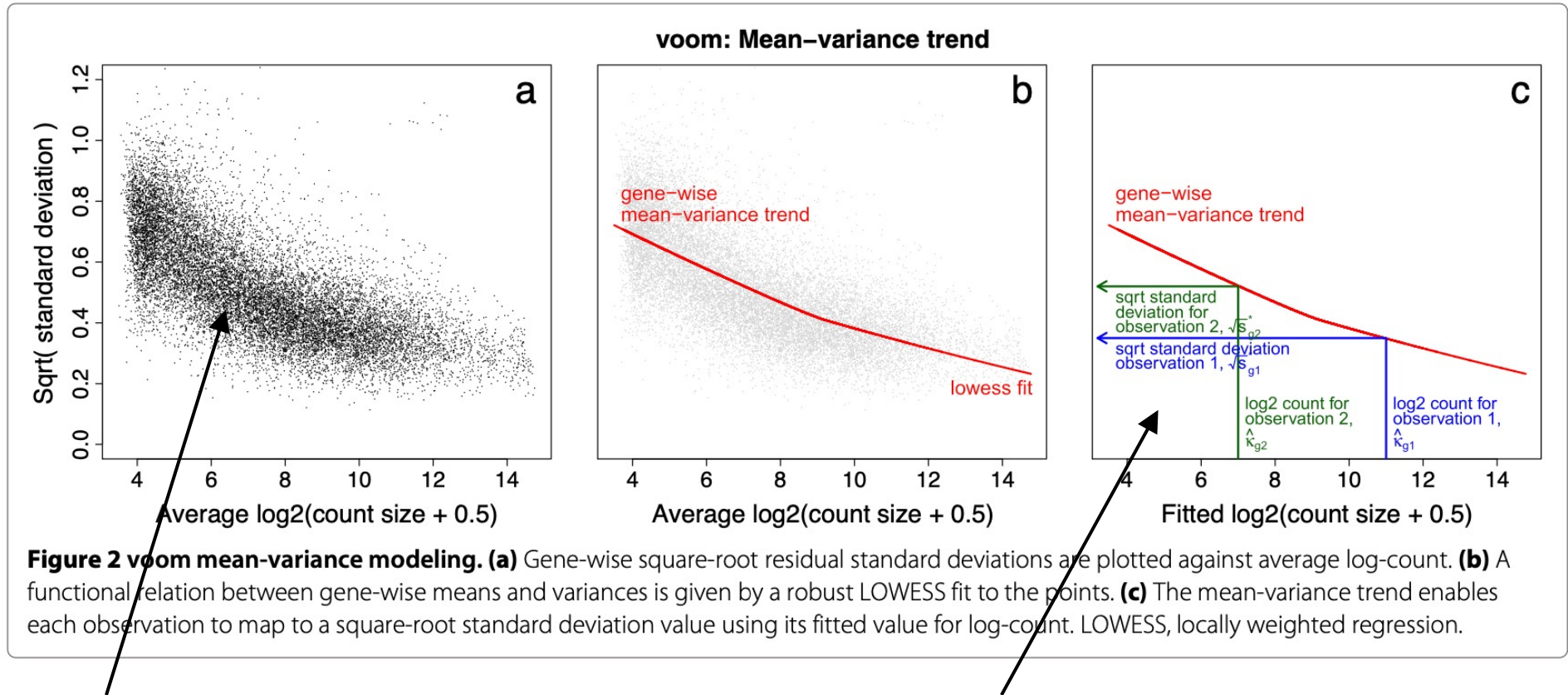
Les données analysées

➤ Table de comptages (counts) :



	Condition A			Condition B		
	R1	R2	R3	R1	R2	R3
A1BG	4	6	2	7	6	7
A1CF	41	33	42	32	42	32
A2M	1	3	1	4	3	7
A2ML1	3	2	2	6	7	3
A2MP1	3	2	2	1	1	0
A3GALT2	1	4	4	3	2	1
A4GALT	420	344	291	327	360	371
A4GNT	1	1	2	1	3	3
AA06	0	0	0	0	0	0
AAAS	2452	2192	1977	2054	2134	2100
AACS	3234	2804	2609	1678	1670	1742
AACSP1	1544	1369	1300	1926	2015	1963

Limma: the « Voom » transformation



Observation de la variance en fonction de l'expression moyenne des gènes

La régression permet d'estimer la variance attendue en fonction du log₂ (count)

Bilan

- L'analyse différentielle est un problème difficile qui nécessite l'utilisation de méthodologies d'analyses complexes.

DESeq et
DESeq2 →
edgeR →
LIMMA+Voom →

TABLE 1. RNA-seq differential gene expression tools and statistical tests

Name	Assumed distribution	Normalization	Description	Version	Citations ^d	Reference
<i>t</i> -test	Normal	DESeq ^a	Two-sample <i>t</i> -test for equal variances	–	–	–
log <i>t</i> -test	Log-normal	DESeq ^a	Log-ratio <i>t</i> -test	–	–	–
Mann-Whitney	None	DESeq ^a	Mann-Whitney test	–	–	Mann and Whitney (1947)
Permutation	None	DESeq ^a	Permutation test	–	–	Efron and Tibshirani (1993a)
Bootstrap	Normal	DESeq ^a	Bootstrap test	–	–	Efron and Tibshirani (1993a)
<i>baySeq</i> ^c	Negative binomial	Internal	Empirical Bayesian estimate of posterior likelihood	2.2.0	159	Hardcastle and Kelly (2010)
<i>Cuffdiff</i>	Negative binomial	Internal	Unknown	2.1.1	918	Trapnell et al. (2012)
<i>DEGseq</i> ^c	Binomial	None	Random sampling model using Fisher's exact test and the likelihood ratio test	1.22.0	325	Wang et al. (2010)
<i>DESeq</i> ^c	Negative binomial	DESeq ^a	Shrinkage variance	1.20.0	1889	Anders and Huber (2010)
<i>DESeq2</i> ^c	Negative binomial	DESeq ^a	Shrinkage variance with variance based and Cook's distance pre-filtering	1.8.2	197	Love et al. (2014)
<i>EBSeq</i> ^c	Negative binomial	DESeq ^a (median)	Empirical Bayesian estimate of posterior likelihood	1.8.0	80	Leng et al. (2013)
<i>edgeR</i> ^c	Negative binomial	TMM ^b	Empirical Bayes estimation and either an exact test analogous to Fisher's exact test but adapted to over-dispersed data or a generalized linear model	3.10.5	1483	Robinson et al. (2010)
<i>Limma</i> ^c	Log-normal	TMM ^b	Generalized linear model	3.24.15	97	Law et al. (2014)
<i>NOISeq</i> ^c	None	RPKM	Nonparametric test based on signal-to-noise ratio	2.14.0	177	Tarazona et al. (2011)
<i>PoissonSeq</i> ^c	Poisson log-linear model	Internal	Score statistic	1.1.2	37	Li et al. (2012)
<i>SAMSeq</i> ^c	None	Internal	Mann-Whitney test with Poisson resampling	2.0	54	Li and Tibshirani (2013)

^aSee Anders and Huber (2010).
^bSee Robinson and Oshlack (2010).
^cR (v3.2.2) and bioconductor (v3.1).
^dAs reported by PubMed Central articles that reference the listed reference (December 21, 2015).

Taille
d'effet

Dispersion

Niveau
expression

<http://www.rnajournal.org/cgi/doi/10.1261/rna.053959.115>

ANNEXES

Format bed

obligatoire			name	score	strand	Thick start	Thick end	color
chr7	127471196	127472363	Pos1	0	+	127471196	127472363	255,0,0
chr7	127472363	127473530	Pos2	0	+	127472363	127473530	255,0,0
chr7	127473530	127474697	Pos3	0	+	127473530	127474697	0,255,0
chr7	127474697	127475864	Pos4	0	+	127474697	127475864	255,0,255

Attention

Le premier nucléotide est numéroté 0.

end - start = taille de la séquence



Format GFF

Permet de décrire les features et leur position

1. **seqname** - The name of the sequence (chromosome/scaffold)
2. **source** - The program that generated this feature
3. **feature** - Type of feature ("CDS", "start_codon", "stop_codon", "exon")
4. **start** - Starting position of the feature in the sequence (starts at 1)
5. **end** - Ending position of the feature (inclusive).
6. **score** - Score between 0 and 1000 (or "." if no value)
7. **strand** - '+', '-', or '.'
8. **frame** - If coding exon, *frame* should be 0-2: reading frame of the first base.
9. **group** - All lines with the same group are linked together into a single item.

Format GTF

=format GFF avec extension du champ 9

```
chr1 HAVANA transcript 2423739 2424697 . - . gene_id "ENSG00000224387.1"; transcript_id "ENST00000424657.1"; gene_type "antisense"; gene_s
chr1 HAVANA exon 2424540 2424697 . - . gene_id "ENSG00000224387.1"; transcript_id "ENST00000424657.1"; gene_type "antisense"; gene_status "K
chr1 HAVANA exon 2423739 2424035 . - . gene_id "ENSG00000224387.1"; transcript_id "ENST00000424657.1"; gene_type "antisense"; gene_status "K
chr1 HAVANA gene 2424876 2425918 . - . gene_id "ENSG00000229393.1"; gene_type "antisense"; gene_status "KNOWN"; gene_name "RP3-395M20.3"; lev
chr1 HAVANA transcript 2424876 2425918 . - . gene_id "ENSG00000229393.1"; transcript_id "ENST00000442305.1"; gene_type "antisense"; gene_s
chr1 HAVANA exon 2425822 2425918 . - . gene_id "ENSG00000229393.1"; transcript_id "ENST00000442305.1"; gene_type "antisense"; gene_status "K
chr1 HAVANA exon 2424876 2425292 . - . gene_id "ENSG00000229393.1"; transcript_id "ENST00000442305.1"; gene_type "antisense"; gene_status "K
chr1 HAVANA gene 2439972 2458067 . - . gene_id "ENSG00000157881.13"; gene_type "protein_coding"; gene_status "KNOWN"; gene_name "PANK4"; lev
chr1 HAVANA transcript 2439972 2458067 . - . gene_id "ENSG00000157881.13"; transcript_id "ENST00000378466.7"; gene_type "protein_coding"; g
chr1 HAVANA exon 2457903 2458067 . - . gene_id "ENSG00000157881.13"; transcript_id "ENST00000378466.7"; gene_type "protein_coding"; gene_sta
chr1 HAVANA CDS 2457903 2458050 . - 0 gene_id "ENSG00000157881.13"; transcript_id "ENST00000378466.7"; gene_type "protein_coding"; gene_sta
chr1 HAVANA start_codon 2458048 2458050 . - 0 gene_id "ENSG00000157881.13"; transcript_id "ENST00000378466.7"; gene_type "protein_coding"; g
chr1 HAVANA exon 2453157 2453239 . - . gene_id "ENSG00000157881.13"; transcript_id "ENST00000378466.7"; gene_type "protein_coding"; gene_sta
chr1 HAVANA CDS 2453157 2453239 . - 2 gene_id "ENSG00000157881.13"; transcript_id "ENST00000378466.7"; gene_type "protein_coding"; gene_sta
chr1 HAVANA exon 2452540 2452754 . - . gene_id "ENSG00000157881.13"; transcript_id "ENST00000378466.7"; gene_type "protein_coding"; gene_sta
chr1 HAVANA CDS 2452540 2452754 . - 0 gene_id "ENSG00000157881.13"; transcript_id "ENST00000378466.7"; gene_type "protein_coding"; gene_sta
```

position

#1 à #8

annotation

#9