EN5636 - Atelier NGS

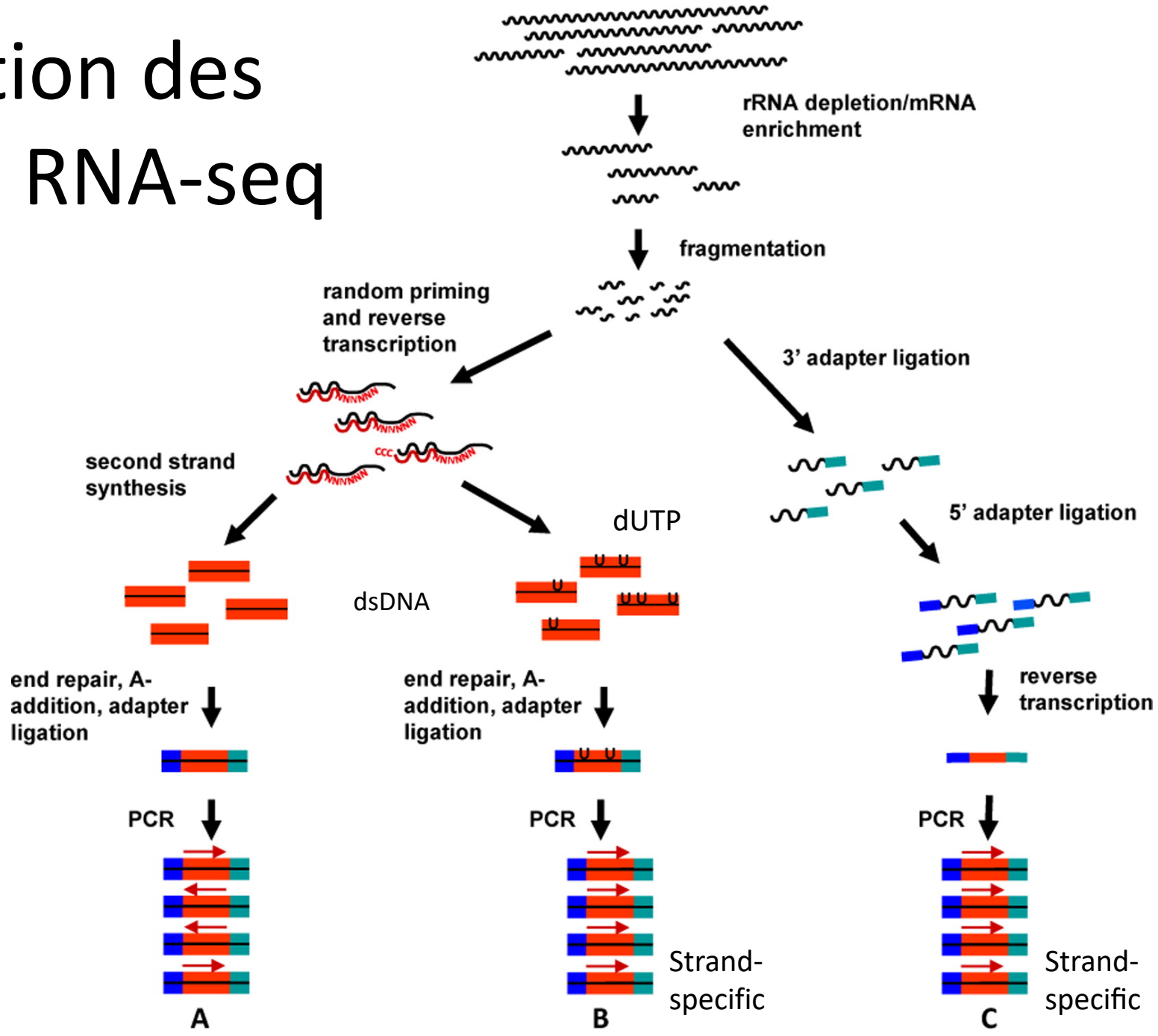
Introduction à l'analyse de transcriptome par RNA-seq

Daniel Gautheret

Avec des diapos de:
Yannick Boursin, IGR
Frédéric Lemoine, Institut Pasteur
Sacha Schutz

UNIVERSITÉ PARIS SUD

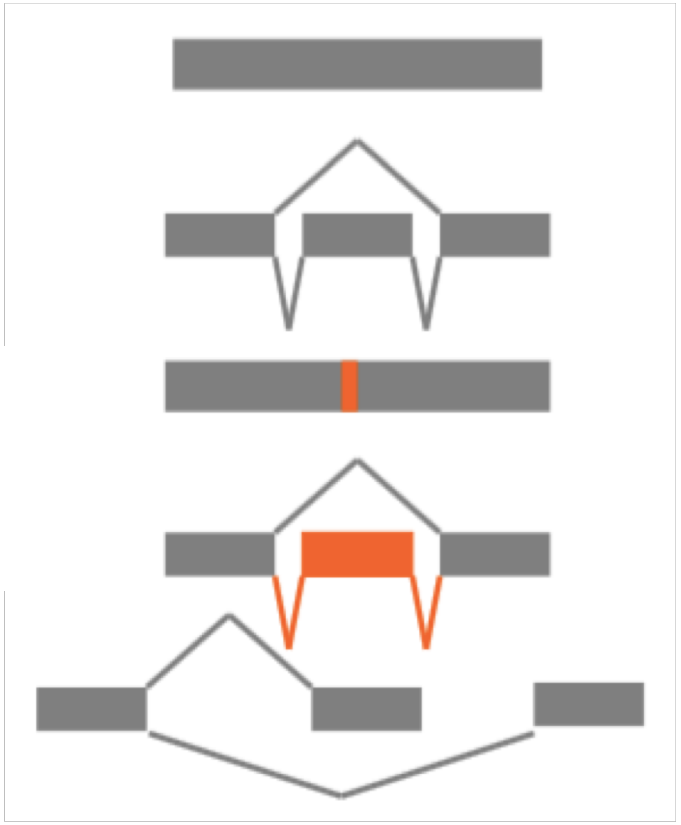
Préparation des banques RNA-seq



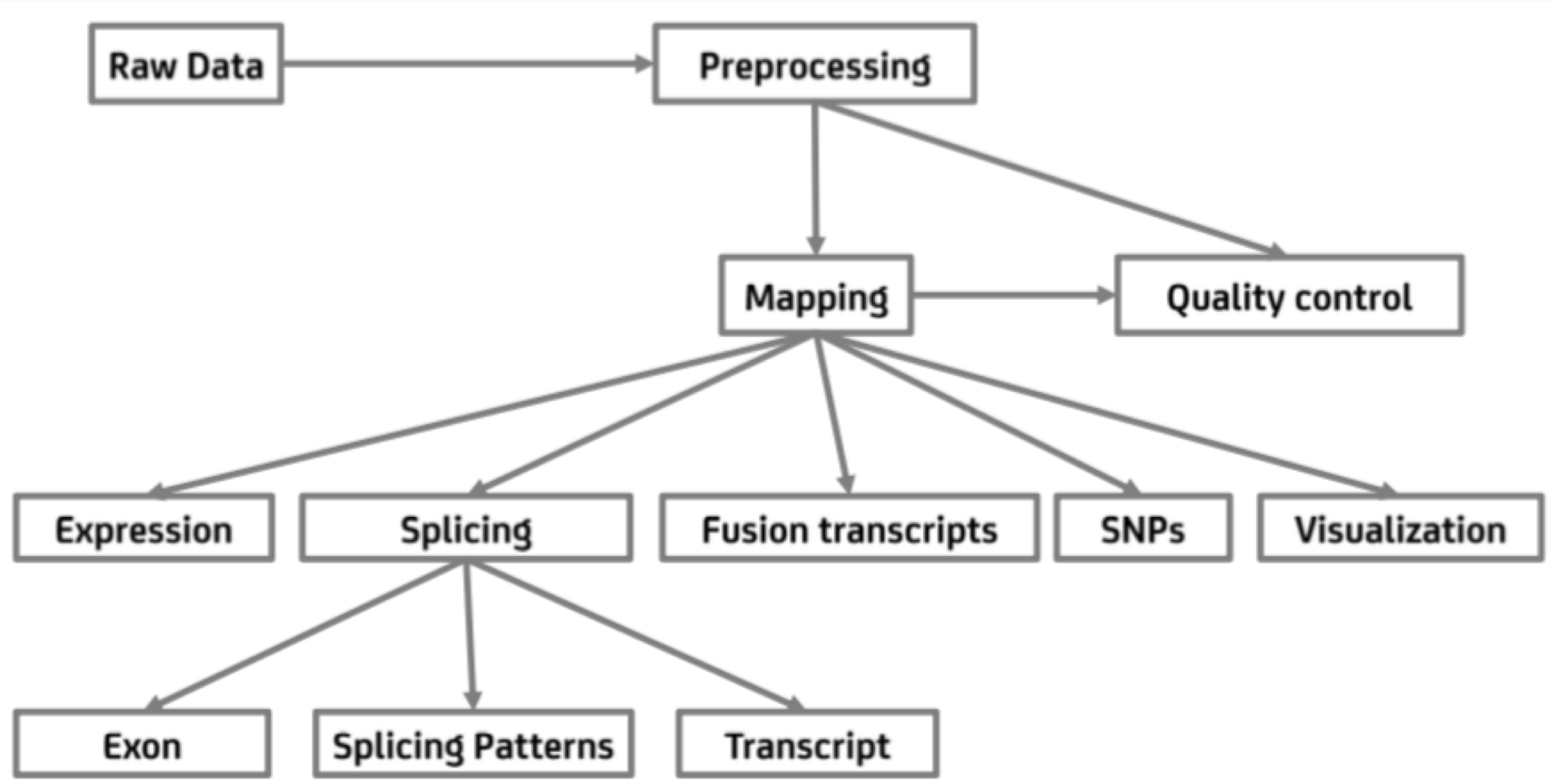
Van Dijk et al.
Exp Cell. Res. 2014

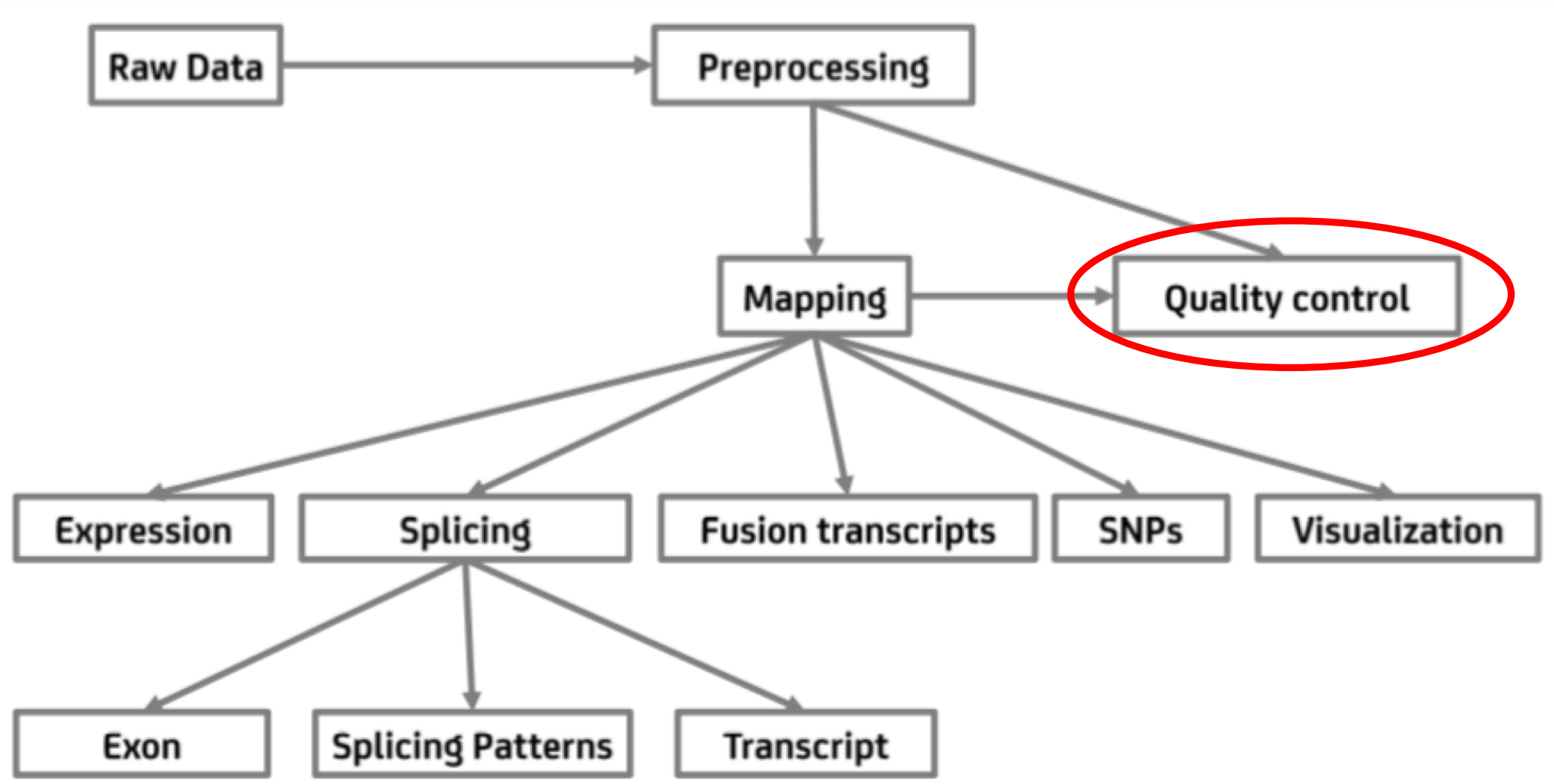
Applications du RNA-seq

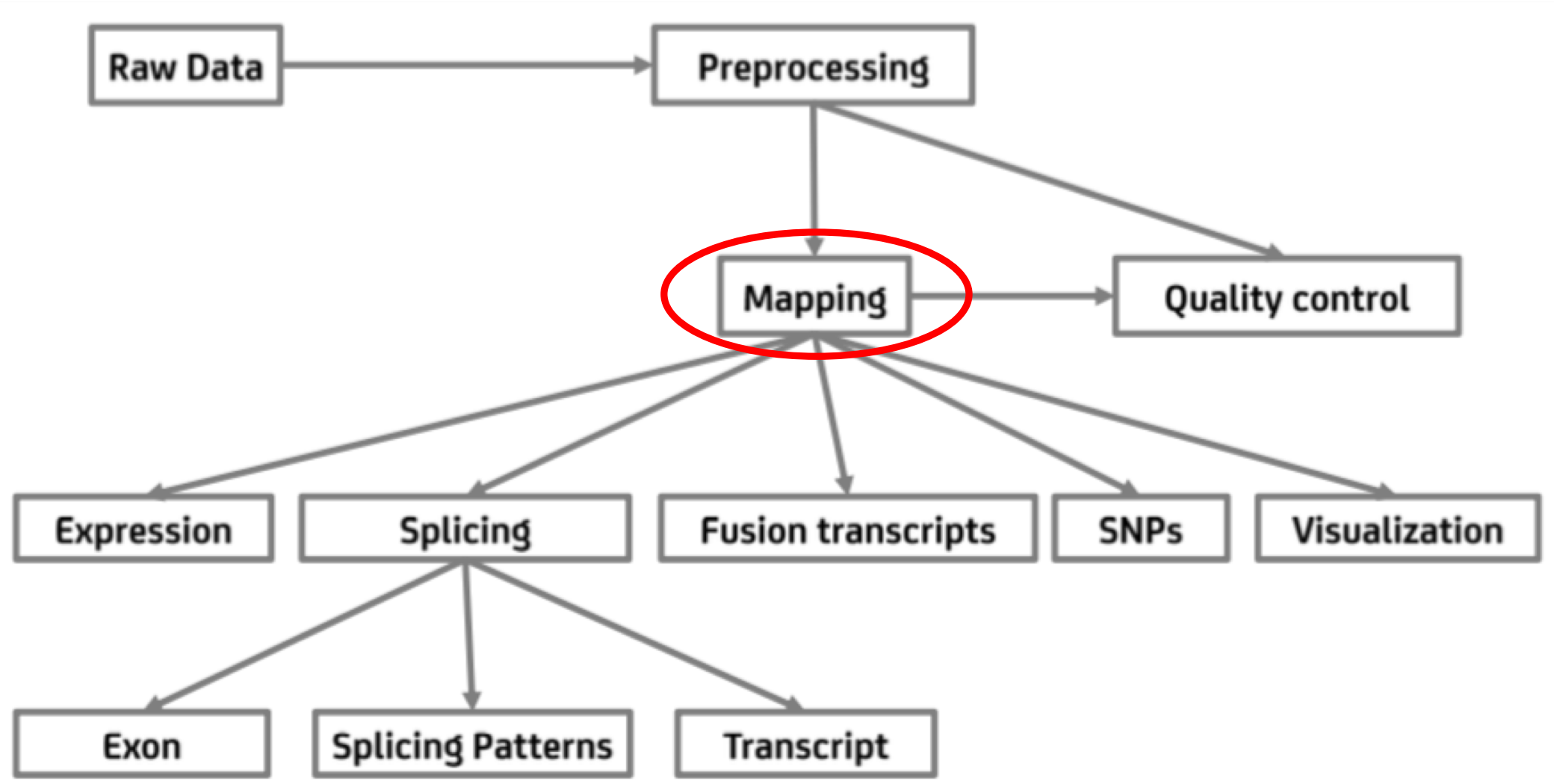
- Mesurer l'expression des gènes
- Mesurer l'épissage alternatif
- Détecter les mutations exprimées
- Annoter les gènes: nouveaux exons
- Détecter les transcrits de fusion



Un pipeline d'analyse RNA-seq

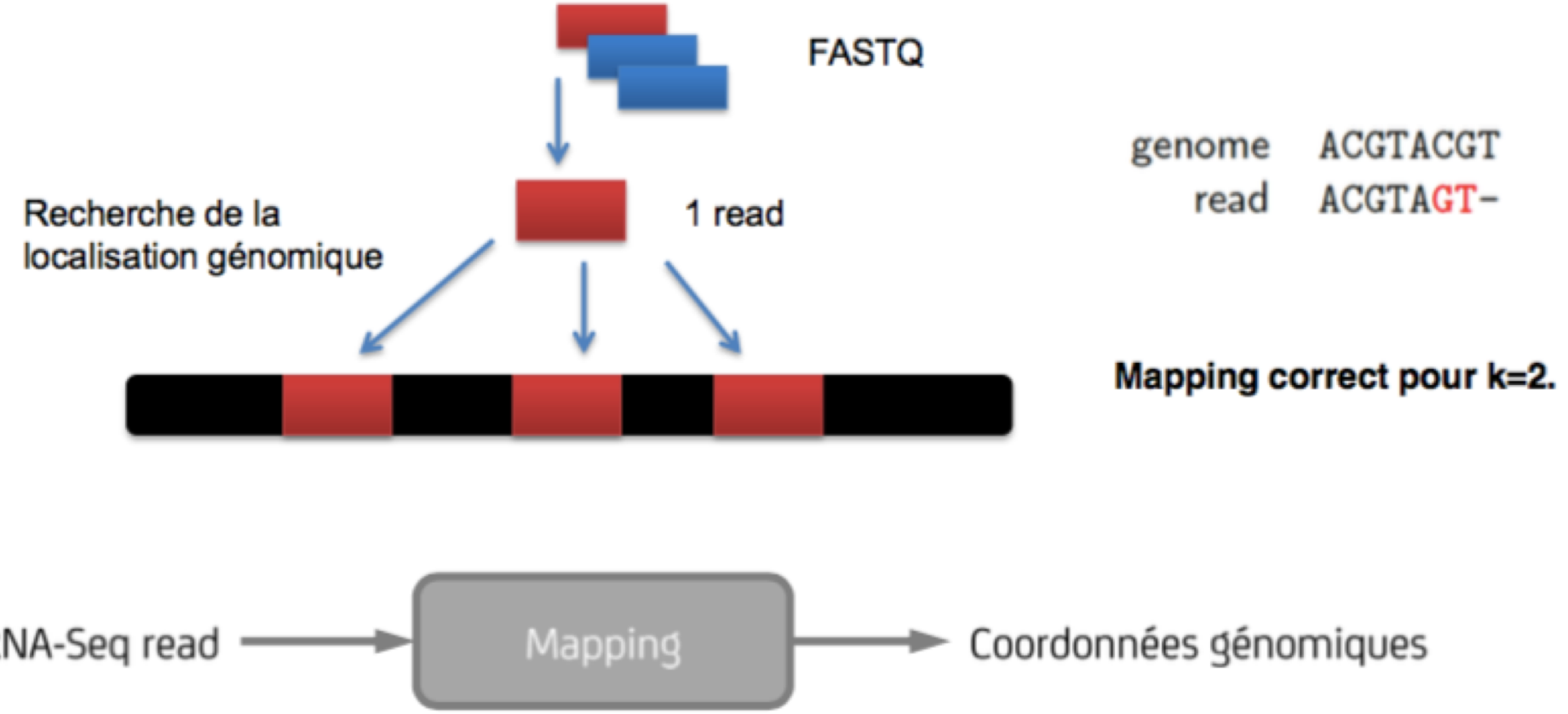



Raw Data
Preprocessing
Mapping
Quality control
Expression
Splicing
Fusion transcripts
SNPs
Visualization
Exon
Splicing Patterns
Transcript


Raw Data
Preprocessing
Mapping
Quality control
Expression
Splicing
Fusion transcripts
SNPs
Visualization
Exon
Splicing Patterns
Transcript

Mapping

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



Mapping to a reference genome: Challenges

Mapping algorithms must address the requirements and characterics of NGS reads:

– Millions of reads per run
– Reads of different size (35bp - 200bp)
– Different types of reads (single-end, paired-end, mate-pair, etc.)
– Base-calling quality factors
– Sequencing errors (~ 1%)
– Repetitive regions
– Sequencing organism vs. reference genome
– Must adjust to evolving sequencing technologies and data formats

Algorithme en **O**(mn)

ACGTTACCGAATCGATCAAGTCGA
TAC

Algorithme en **O**(mn)

ACGTTACCGAATCGATCAAGTCGA

TAC

Algorithme en **O**(mn)

ACGTTACCGAATCGATCAAGTCGA

TAC

Algorithme en **O**(mn)

ACGTTACCGAATCGATCAAGTCGA

TAC

Algorithme en **O**(mn)

ACGTTACCGAATCGATCAAGTCGA

TAC

OK pour 1 read: O(3.10^9 x 100)

Mais pour 1^8 reads???

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

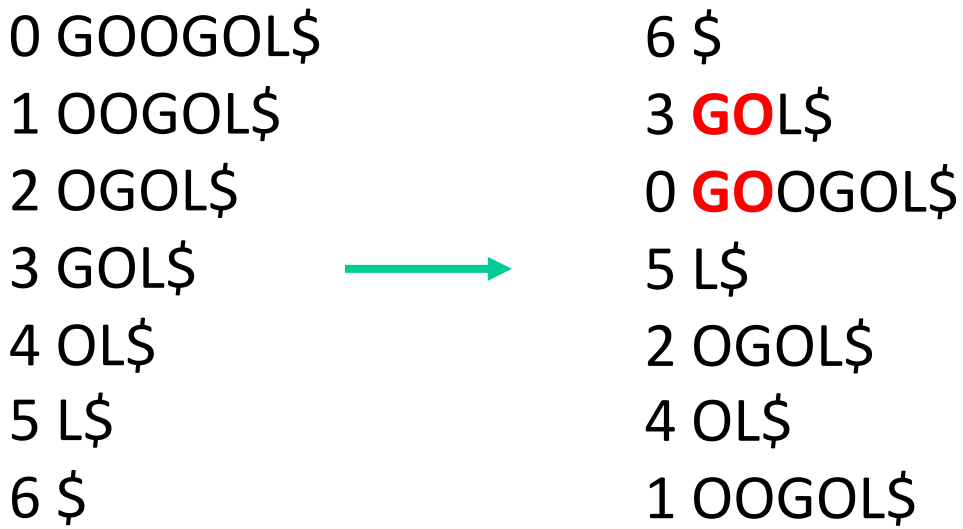
Tableau trié de tous les suffixes d’une chaîne de caractères

0 GOOGOL$		6 $		
1 OOGOL$		3 GOL$		
2 OGOL$		0 GOOGOL$		
3 GOL$	→	5 L$	→	(6,3,0,5,2,4,1)
4 OL$		2 OGOL$		
5 L$		4 OL$		
6 $		1 OOGOL$		

Propriété: toutes les occurrences d’une même chaîne sont regroupées.

Suffix arrays

Exemple: trouver la chaîne **GO**



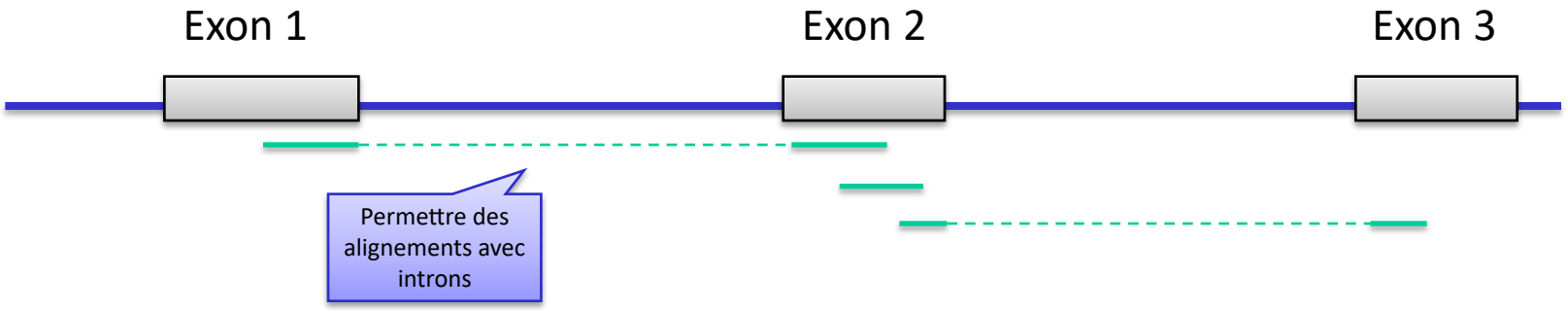
Burrow-Wheeler Transform (BWT)

- Permet de compresser efficacement un Suffix array
- Maintient une capacité de recherche dans la forme compressée

Les algorithmes de mapping

name	seed-and-extend	pigeon hole	spaced seed	q-gram	suf. tree	B.–W.
SSAHA	X					
Blat	X					
MUMmer2					X	
Eland			X			
MAQ		X				
SOAP		X	X			
RMAP		X	X			
SeqMap		X				
QPalma					X	
Mosaik	X					
SOCS		X				
ZOOM			X			
PASS	X					
SOAP2						X
BWA						X
SHRiMP				X		
Bowtie						X
BFAST			X		X	
mrFAST	X					
RazerS				X		
MPScan					X	
PerM			X			
CloudBurst			X			
GNUMap			X			
mrsFAST	X					
novoalign	?	?	?	?		
GASSST			X			
Stampy	X					
SOAP3						X
Bowtie2						X
Saruman				X		

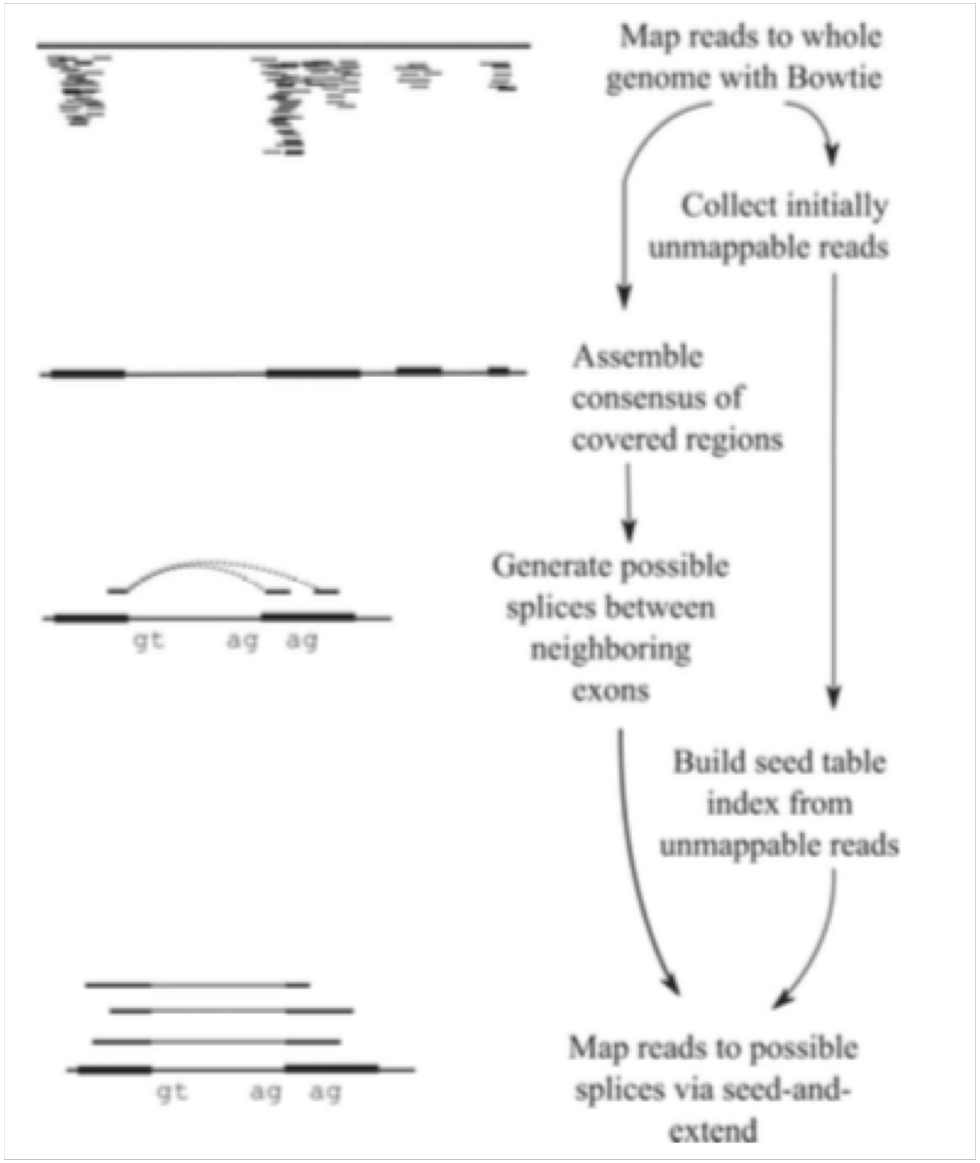
La spécificité du mapping RNA-seq



Le programme TopHat

Successeurs:

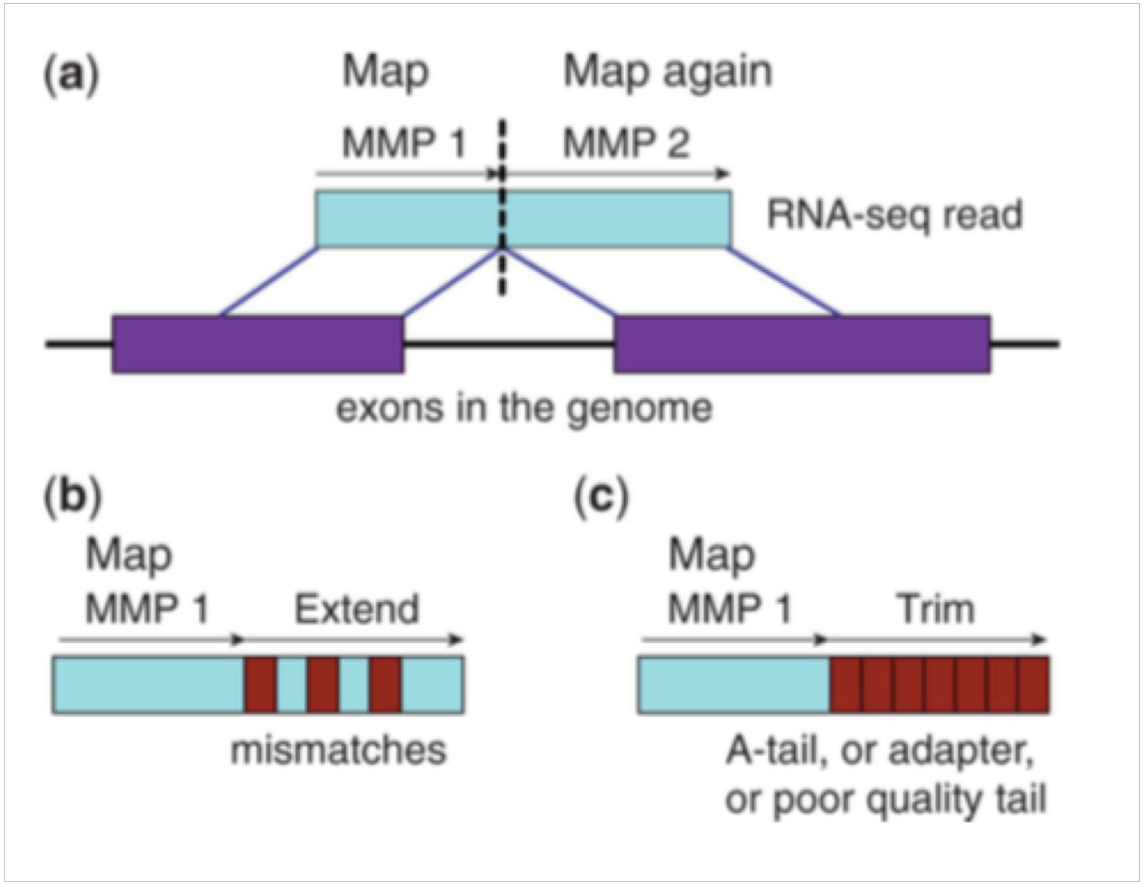
- TopHat2
- HiSat
- HiSat2



Trapnell et. al. Bioinformatics, 2009 (TopHat)
Kim et al. Nature Methods 2015 (HiSat)

Le programme STAR

Dobin et al. Bioinformatics, 2013



STAR utilise les jonctions intron/exon préannotées pour améliorer la détection des évènements a,b,c.

Le mode « 2-Pass mapping » permet de découvrir de nouvelles jonctions

Problèmes avec tous les algorithmes de mapping

La mappabilité: une partie du génome reste invisible

		H.sapiens	*M.musculus*	*D.melanogaster* (dm3)		*C.elegans*
		(hg19)	**(mm9)**	**with het.**	**without het.**	**(ce6)**
Genome size (bp)		3,107,677,273	2,725,765,481	168,736,537	159,454,756	100,281,426
Repeat sequences (bp)		1,406,290,513	1,153,714,659	44,719,009	38,601,028	13,121,257
Proportion of repeats		45.25%	42.33%	26.50%	24.20%	13.08%
LTR		8.05%	10.56%	10.46%	–	10.46%
Non-LTR	SINEs	12.59%	7.39%	0.00%	–	0.09%
	LINEs	19.73%	19.66%	7.08%	–	0.36%
Uniquely mapped positions ($m=0$)						
$k=36$		2,489,885,654	2,178,433,024	119,915,412	116,918,511	92,332,303
		(80.12%)	(79.92%)	(71.07%)	(73,32%)	(92.07%)
$k=50$		2,627,947,484	2,267,226,534	121,732,432	118,368,697	93,775,749
		(84.56%)	(83.18%)	(72.14%)	(74,23%)	(93.51%)
$k=75$		2,729,902,459	2,349,591,487	124,087,375	120,329,119	95,226,461
		(87.84%)	(86.20%)	(73.54%)	(75,46%)	(94.96%)
Uniquely mapped positions ($m=2$)						
$k=36$		2,175,066,863	1,964,593,763	114,889,241	113,088,604	87,385,879
		(69.99%)	(72.07%)	(68.09%)	(70,92%)	(87.14%)
$k=50$		2,380,109,920	2,100,436,231	117,178,560	114,915,550	90,050,144
		(76.59%)	(77.06%)	(69.44%)	(72,06%)	(89.80%)
$k=75$		2,582,297,225	2,225,670,208	119,798,046	116,955,098	92,369,340
		(83.09%)	(81.65%)	(71.00%)	(73,35%)	(92.11%)

Repeat elements have been identified and classified by the RepeatMasker program [37]. The mappability has been computed for $k=36$,50 and 75, with $m=0$ and 2.
doi:10.1371/journal.pone.0030377.t002

Derrien et al. PloS One 2012

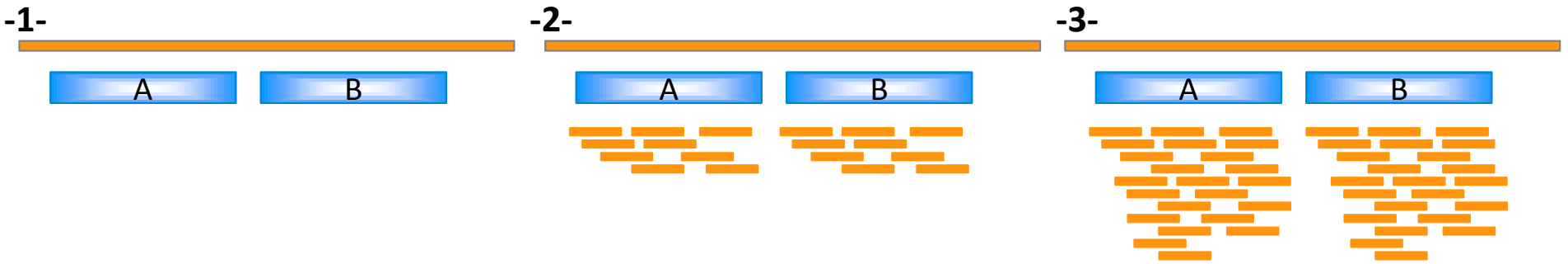
Alignment key parameters: multiple maps

3 strategies:

1- Report only unique alignment

2- Report best alignments and randomly assign reads across equaly good ones

3- Report all (best) alignments



Treangen T.J. and Salzberg S.L. 2012. Nature review Genetics 13, 36-46

Alignment key parameters – Using single or paired-end reads ?

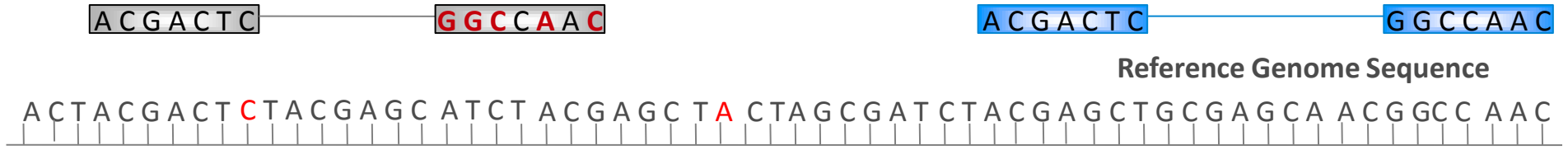
The type of sequencing (i.e. single or paired-end reads) is often driven by the application.

Exemple : Finding large indels, genomic rearrangements, ...

However, in most cases, the pair information can improve mapping specificity

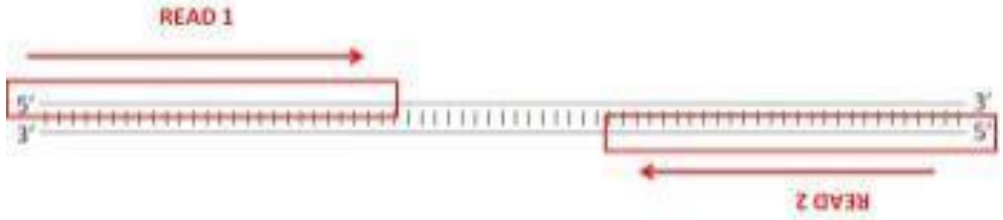
- Single-end alignment – repeated sequence

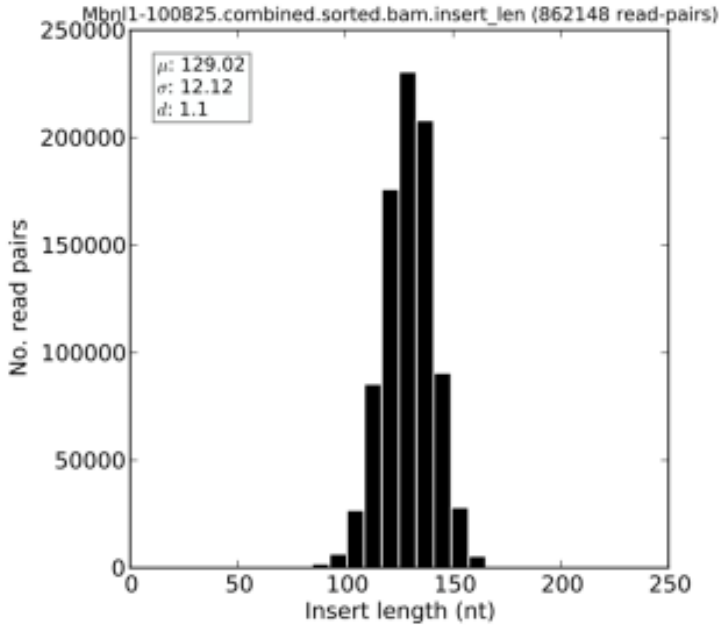
- Paired-end alignment – unique sequence



Paired-end mapping

- Insert-size checking





- % of **"All Good"**= both reads in the pair have aligned
- "the pair is properly aligned" meaning that they mapped within a proper distance from each other
- % of **"All Bad"** = neither the read nor its mate mapped
- % of **Only one read maps =** only one read in a pair is mapped

Limitations of Alignment Tools

Even if we have nice tools to align reads on a reference genome, several issues remain:

- Homopolymer mapping
- Efficiently align small indels
- Alignment on several genomes
- Alignment on repeated sequences
- ...

Mapping - Vocabulary

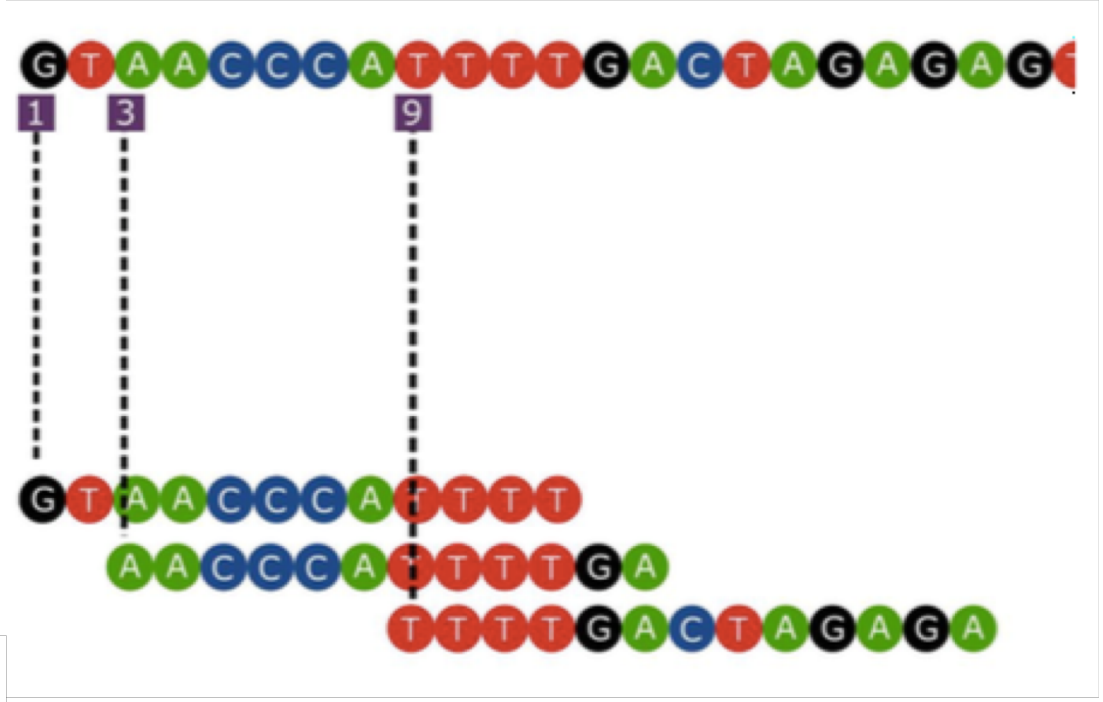
Alignment : **(mapping)** The reads alignment aims at transforming the single reads information in an organized and reduced set of information.
Mismatch : Incoherence between two nucleotides
Reference Genome : The reference genome is a known sequence, supposed to be as close as possible to the input genome, and which is used as an anchor to organize the single reads information.
Gap : Bridge within the read alignment (i.e. small Insertion/deletion)
Mappability : Uniqueness of a region (repeated region = low mappability, unique region = good mappability)
Indels : Insertion/Deletion into the reference genome

Alignment formats

- Many formats exist:
    - SAM
    - BAM
    - ELAND (Illumina specific)
    - MAQ map
    - PileUp

SAM and BAM are now the standard for aligned data

Format SAM
(séquences alignées sur une référence)



Information minimale:

```
chr7 1324324    ACGTGCGTTTGCGT
chr8 1724354    GCGTGATGCGTAAG
chr8 1424324    GTATGTTATATGTA
```

Format SAM

11 champs obligatoires

Sequence ID	Flag	Chr	Position	Map Qual	Cigar	Paired end info		
HWI-ST1136:196:HS113:4:1101:4333:28021	163	chr2	217279469	255	100M	=	217279487	117
HWI-ST1136:196:HS113:4:1101:4333:28021	83	chr2	217279487	255	99M1S	=	217279469	-117
HWI-ST1136:196:HS113:4:1101:4320:28039	163	chr11	65271253	255	100M	=	65271335	182
HWI-ST1136:196:HS113:4:1101:4320:28039	83	chr11	65271335	255	100M	=	65271253	-182
HWI-ST1136:196:HS113:4:1101:4274:28047	99	chr4	763497	255	100M	=	763607	210
HWI-ST1136:196:HS113:4:1101:4274:28047	147	chr4	763607	255	100M	=	763497	-210
HWI-ST1136:196:HS113:4:1101:4333:28054	99	chr17	74433086	255	100M	=	74433100	114
HWI-ST1136:196:HS113:4:1101:4333:28054	147	chr17	74433100	255	100M	=	74433086	-114
HWI-ST1136:196:HS113:4:1101:4353:28065	99	chr11	62293812	255	100M	=	62293909	197
HWI-ST1136:196:HS113:4:1101:4353:28065	147	chr11	62293909	255	100M	=	62293812	-197

...

SAM (fin)

...

Sequence	Base qualities	Optional tags
AGAGAATCGACAAAAGGCTCTGGCCCG	CCCFFFFFHHHHHJJJIJIJJJJJJIJJJB	NH:i:1 HI:i:1 AS:i:197 nM:i:0
TCTGGCCCGCAGAGCTGAGAAGTTATT	DDDDDBDBDCDDDDDEDDDEDDCCAACDEEE	NH:i:1 HI:i:1 AS:i:197 nM:i:0
AACGAATGTAACTTTAAGGCAGGAAAG	CCCFFFFFHHHHHJJJJJJJJJJIJJJIIII	NH:i:1 HI:i:1 AS:i:198 nM:i:0
ATAGAGGCCCTCTAAATAAGGAATAAA	DDDDDDDFFFDDHHHHHJIIGJJJIJIGGCJ	NH:i:1 HI:i:1 AS:i:198 nM:i:0
CCTGAGATGTGCGTAGCCTCCGTGTAA	CCCFFFFFHHHHHJJJJJJJJJIJIJJJJJJ	NH:i:1 HI:i:1 AS:i:198 nM:i:0
ACCCAGCCTTTACCAGCAGCGTACGGC	ADDDDDDCDDDCDDDDDDDDDDDFFFFHHHH	NH:i:1 HI:i:1 AS:i:198 nM:i:0
GCTGGCATGGTGGTGGGCACCCATAAT	CCCFFFFDHHFHHHGIJIJJJJJJJJJJIJJ	NH:i:1 HI:i:1 AS:i:198 nM:i:0
GGGCACCCATAATCCTAGCTGCTCAGG	DDDBCDCDDDDDCDDDDDDEEECCCFFFEHH	NH:i:1 HI:i:1 AS:i:198 nM:i:0
GCCCTTTCAACTTTCCCTCTGGTCCTT	CCCFFFFFHHHHHJJIJJIJJJGIJJJJJJJ	NH:i:1 HI:i:1 AS:i:196 nM:i:1
CACATCCCCATCTGGGCCCTCTCCTTT	DDDDDDDDDCBDDDDDDDDCDEFFFFFHHHH	NH:i:1 HI:i:1 AS:i:196 nM:i:1

Le champ CIGAR

Example: 52M36890N45M3S

All Cigar operations

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Les Flags SAM

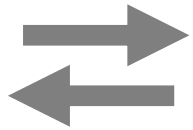
Example:

- **Decimal Flag Value** 83
- **Binary Flag Value**

2048	1024	512	256	128	64	32	16	8	4	2	1
0	0	0	0	0	1	0	1	0	0	1	0

- **To each bit corresponds a meaning**

Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

SAM  BAM

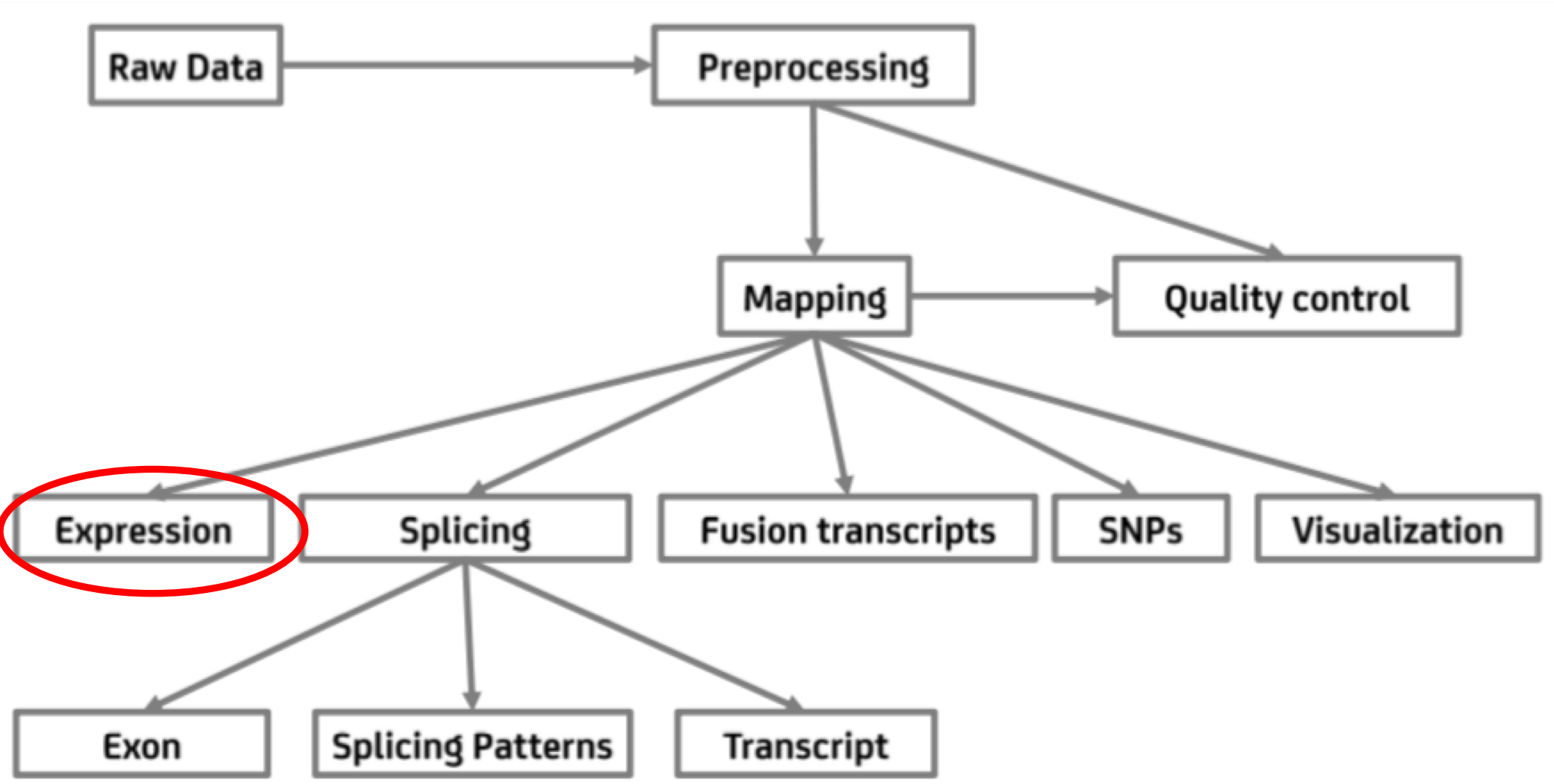
SAMtools

Fichier texte

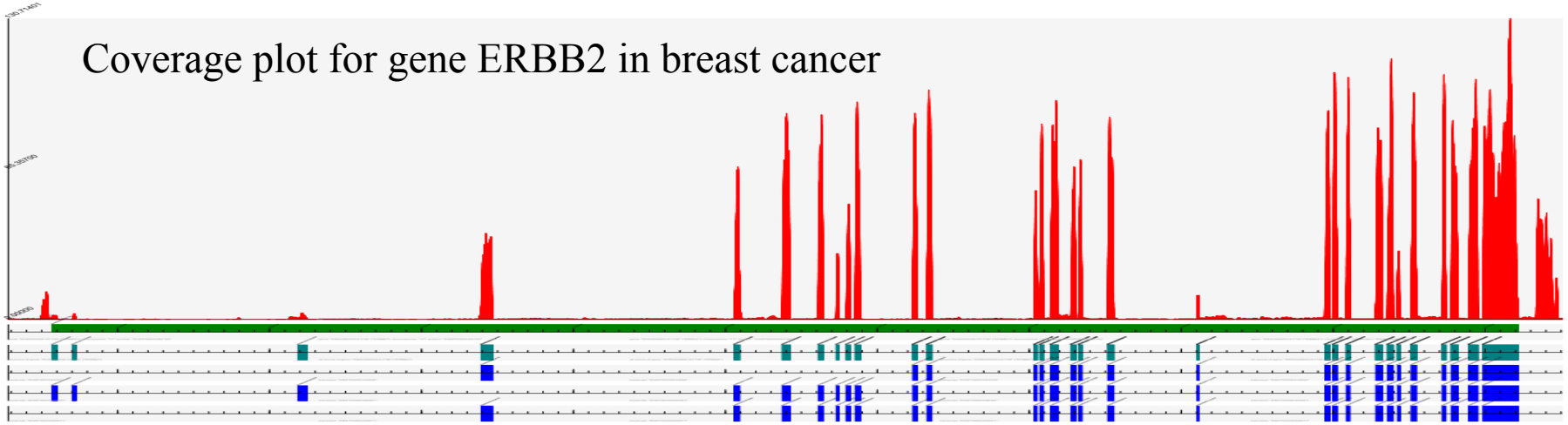
Fichier binaire

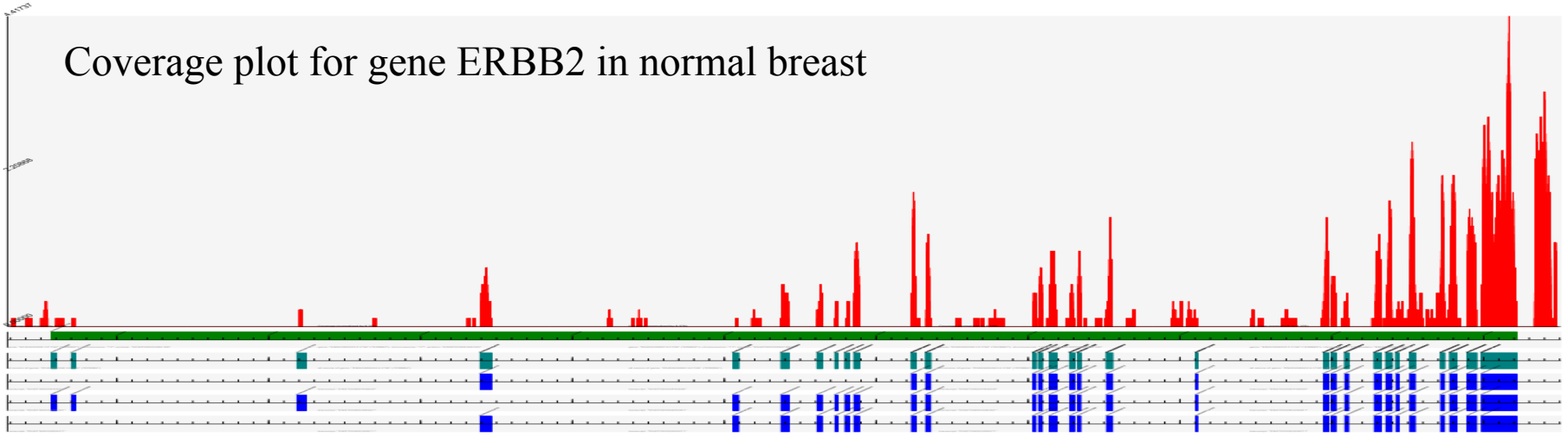
Samtools

- La boîte à outils pour traiter les SAMs sous Unix
  - BAM <-> SAM
  - BAM <-> FASTQ
  - Tri de BAM
  - Indexation du BAM (création fichier .bai)

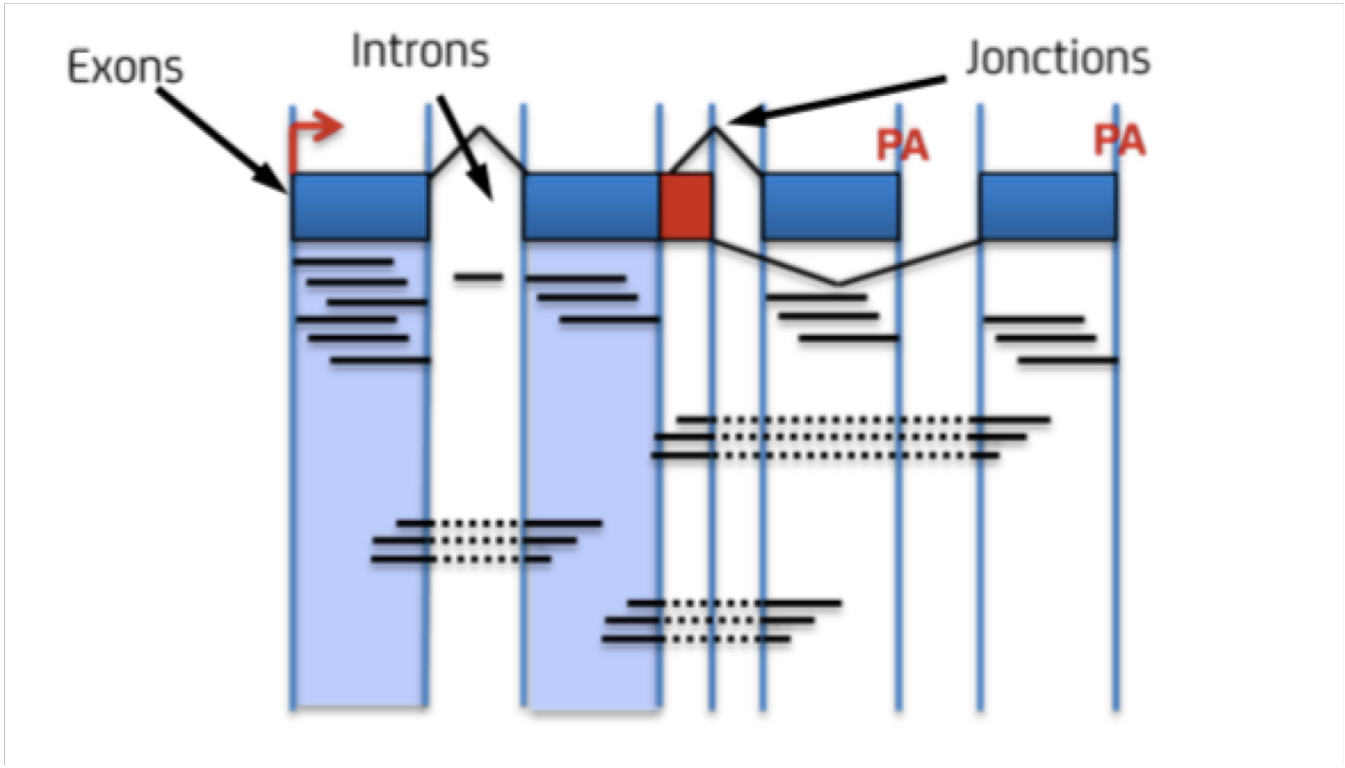

Raw Data
Preprocessing
Mapping
Quality control
Expression
Splicing
Fusion transcripts
SNPs
Visualization
Exon
Splicing Patterns
Transcript

Differential expression





Mesure de l’expression par comptage des reads alignés



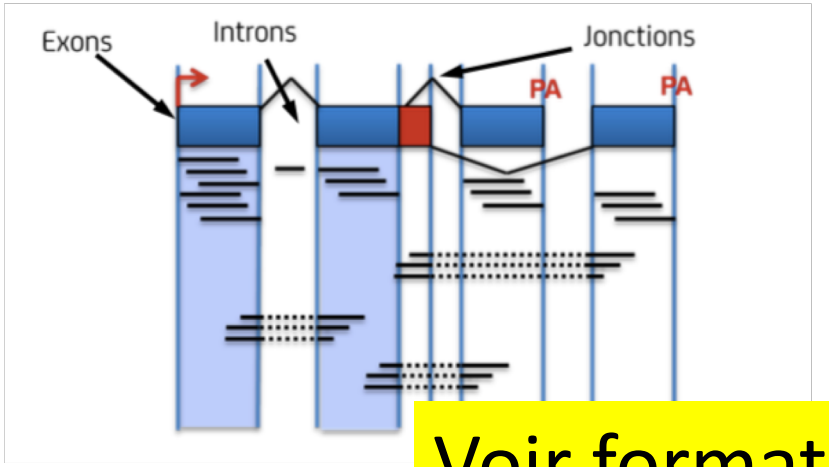
Indexer les fichiers BAM

- 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
```

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

Notre jeu de données « EMT »

AMERICAN SOCIETY FOR MICROBIOLOGY | Molecular and Cellular Biology

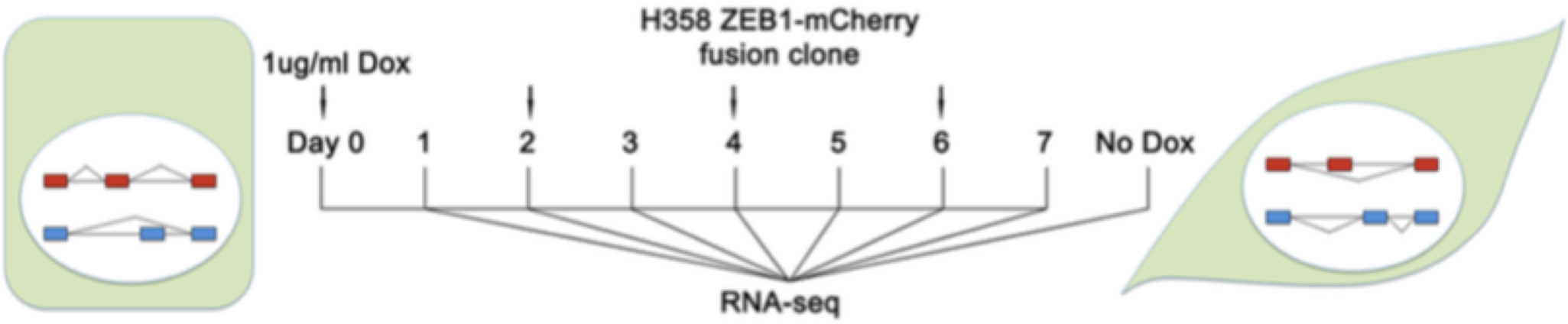
Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition

Yueqin Yang,[a,b] Juw Won Park,[c,d,e] Thomas W. Bebee,[a,b] Claude C. Warzecha,[a,b*] Yang Guo,[c,f] Xuequn Shang,[f] Yi Xing,[c] Russ P. Carstens[a,b]

Departments of Genetics[a] and Medicine,[b] Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA; Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, Los Angeles, California, USA[c]; Department of Computer Engineering and Computer Science[d] and KBRIN Bioinformatics Core,[e] University of Louisville, Louisville, Kentucky, USA; School of Computer Science, Northwestern Polytechnical University, Xi'an, China[f]

E

M



Epithelial cell

Mesenchymal cell

(non-small cell lung cancer (NSCL) cell line H358)

Data

- Sequence libraries are polyA+, pair-end 2x100nt, each in biological triplicate.
- Sequencing is performed on a Illumina HiSeq 2500.
- Fastq files were obtained here: http://www.ncbi.nlm.nih.gov/sra?term=SRP066794

Data Sampling

- Initial fastq files: 72Mx2 reads
- Reads mapping Chr18 (STAR mapping + grep on SAM file) : 685,000 x2 reads
- Sampled by a factor of 0.5 (Samtools) : 343,000 x2 reads

This represents 0.5% of total reads, thus actual runtimes and space requirement would be up to 200 times higher than in our exercices.

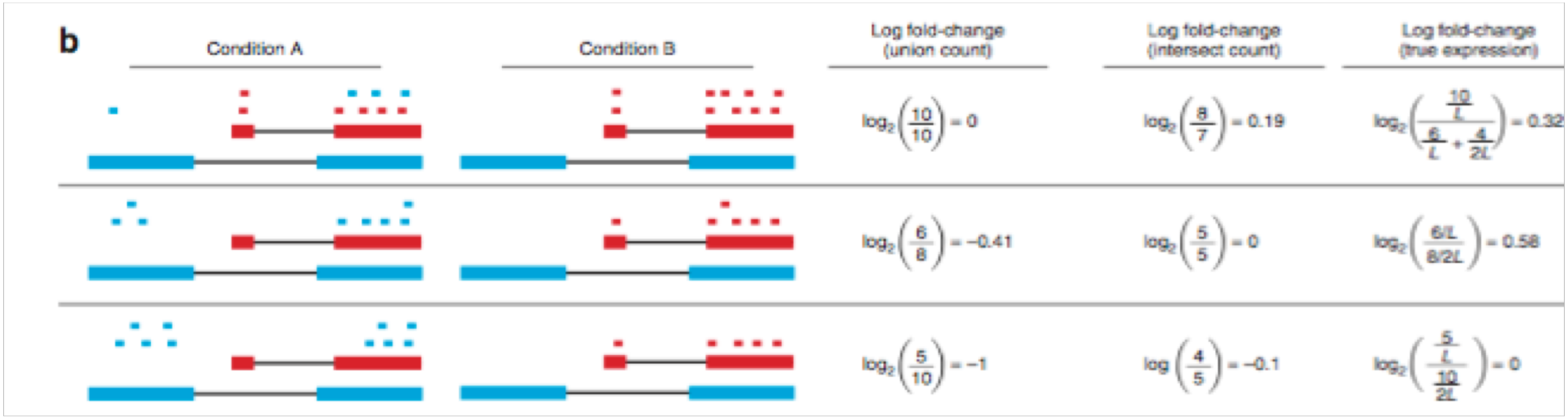
Les méthodes de comptage

Limitation du comptage par gène
(type FeatureCount)

Reality

Bad approximations = FeatureCounts

What we want



Trapnell et al. Cufflinks 2013

1. Use specific part of transcript to estimate mapping on non-specific parts
2. Expectation Maximimization (EM) algorithms are used to assign reads to isoforms

Common normalization methods

- RPM
  - Normalized by library size
- RPKM
  - Normalized by library size and gene-size
- TPM
  - Transcript-level count, normalized by library size

Software

FeatureCount

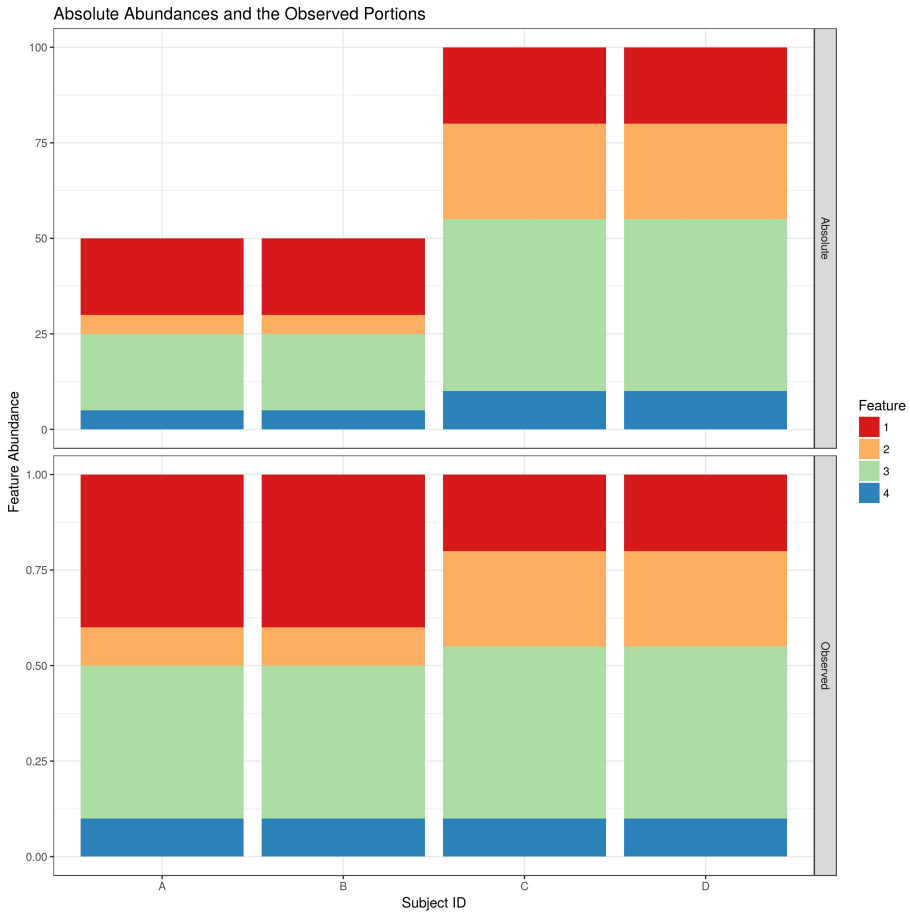
Cufflinks, RSEM, Kallisto, Salmon

Library size scaling is not sufficient

4 conditions

4 genes

4 genes



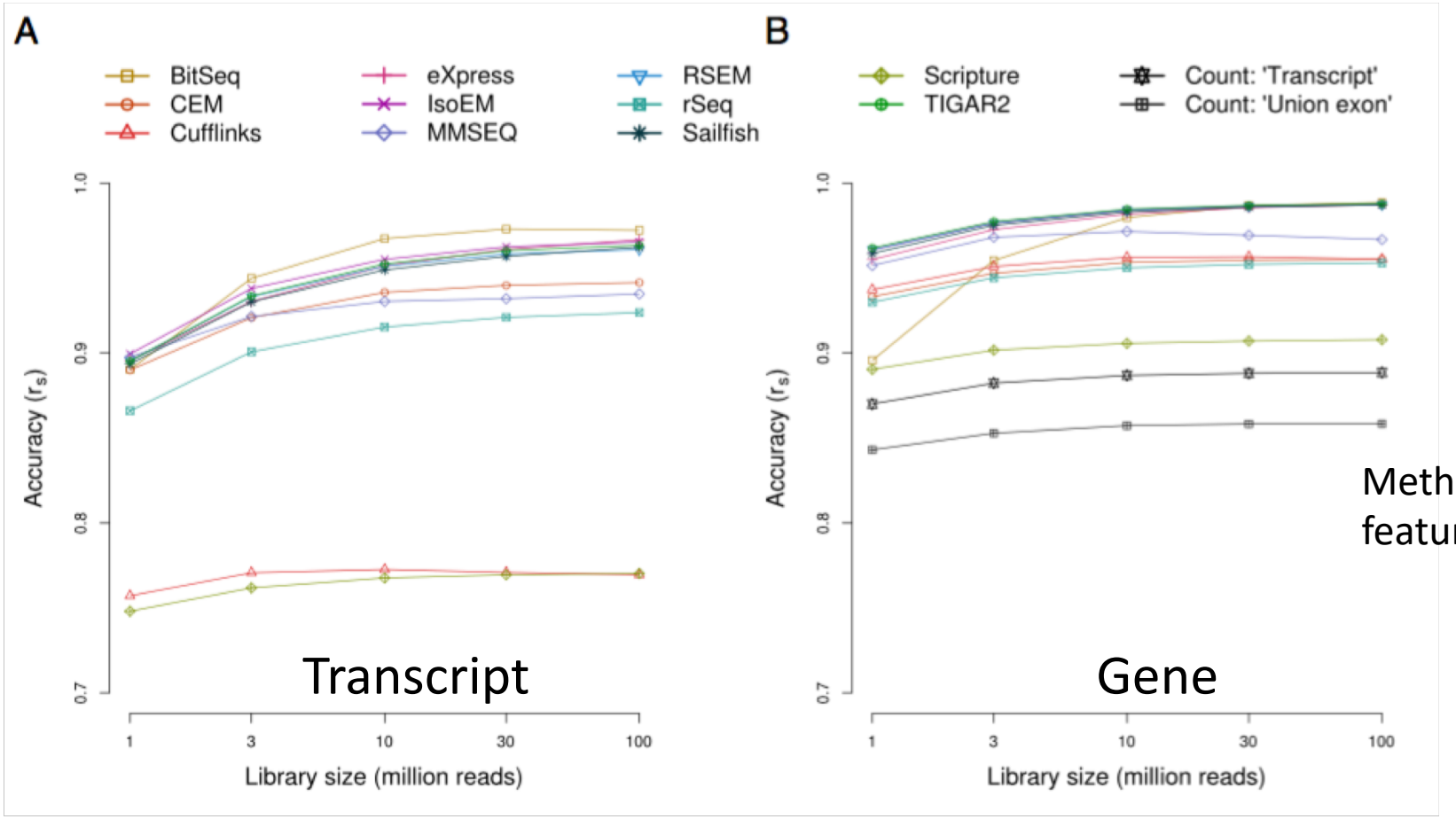
Real RNA abundance

RNA-seq counts (libraries have same size)

Actual changes are lost!

Quinn et al. Bioinformatics, 2018

Précision pour transcrits et gènes

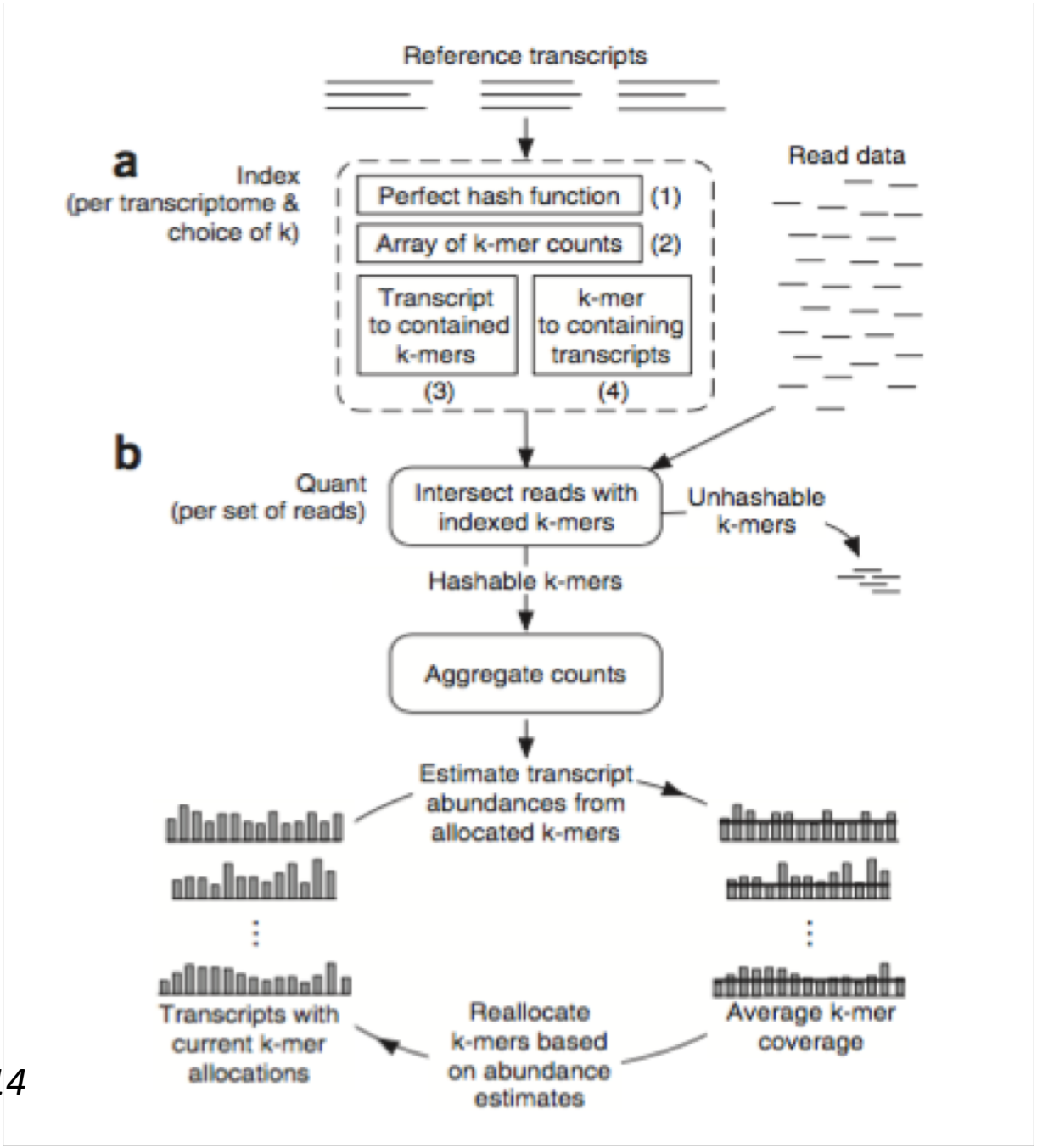


Methodes type featureCounts

Kanitz et al. Genome Biol. 2015

Les nouvelles methodes sans mapping

Sailfish
Salmon
Kallisto

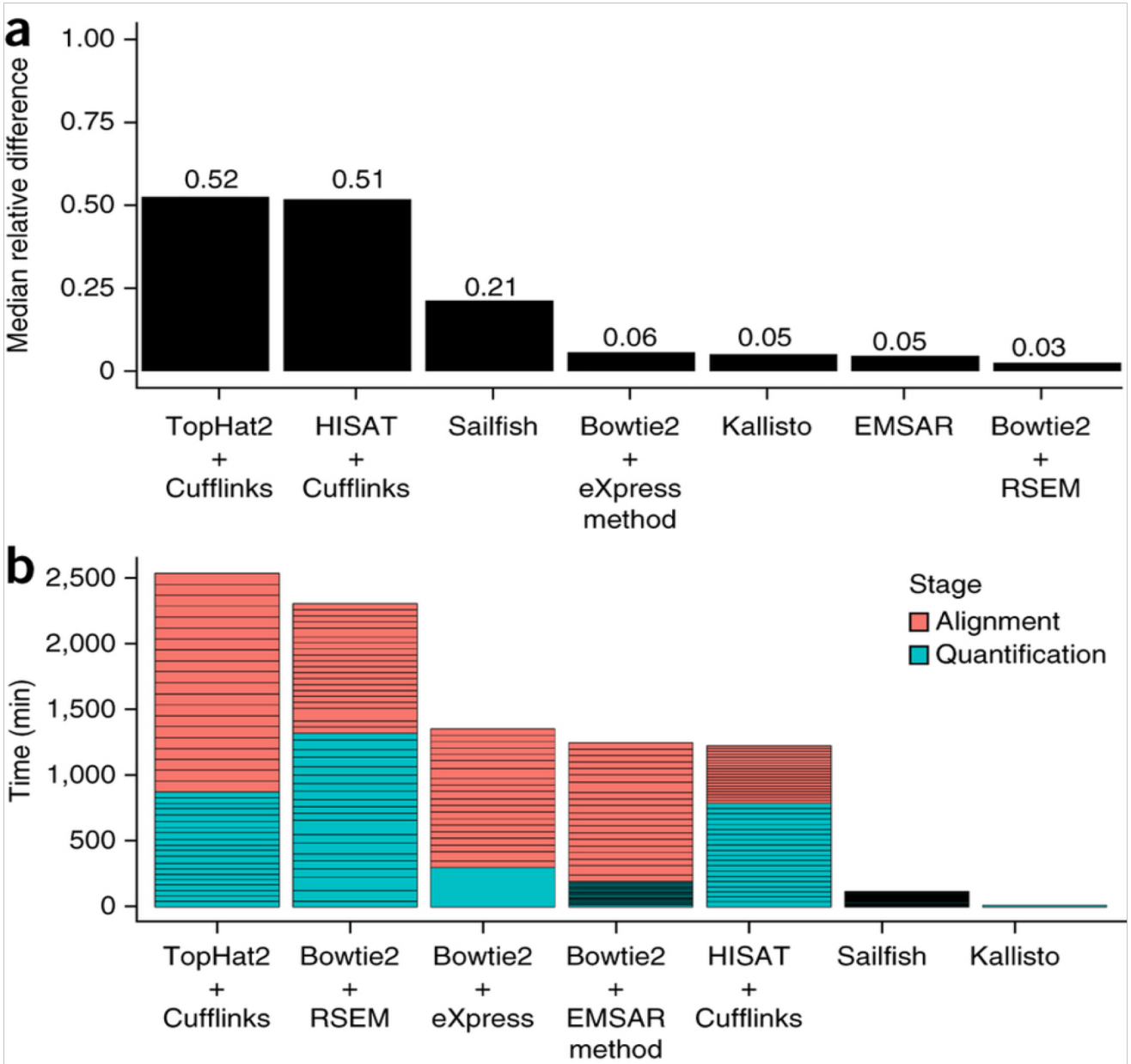


Sailfish. Patro et al. 2014

Performance in gene level prediction

Precision

CPU Times



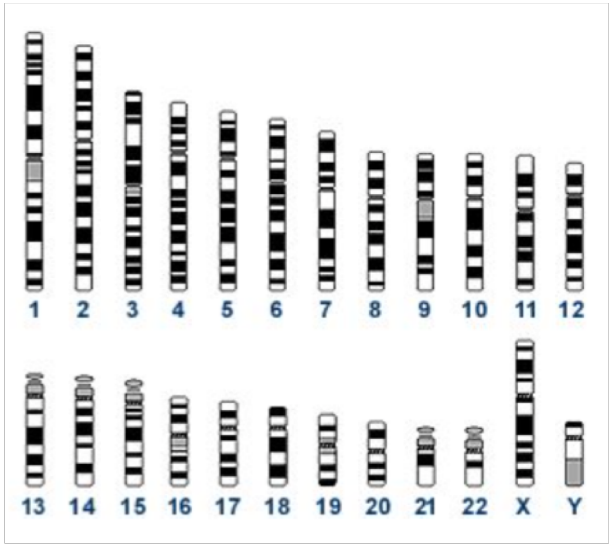
(Bray et al. Kallisto 2016)

Formats de régions

Les régions

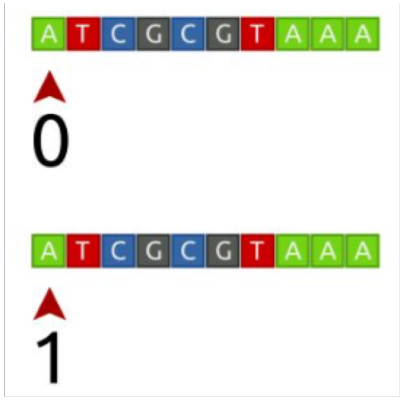
Les coordonnées génomiques permettent de définir une région exacte du génome

<**chromosome**>:<**start**>-<**end**>
`chr7:117465784-117715971`



Formats de fichiers utilisant les régions

- Attention: suivant le format la première base peut être numérotée 0 ou 1



Format/library	Type
BED	0-based
GTF	1-based
GFF	1-based
SAM	1-based
BAM	0-based
VCF	1-based
BCF	0-based
Wiggle	1-based
GenomicRanges	1-based
BLAST	1-based
GenBank/EMBL Feature Table	1-based

Format bed

obligatoire

name score strand Thick start Thick end color

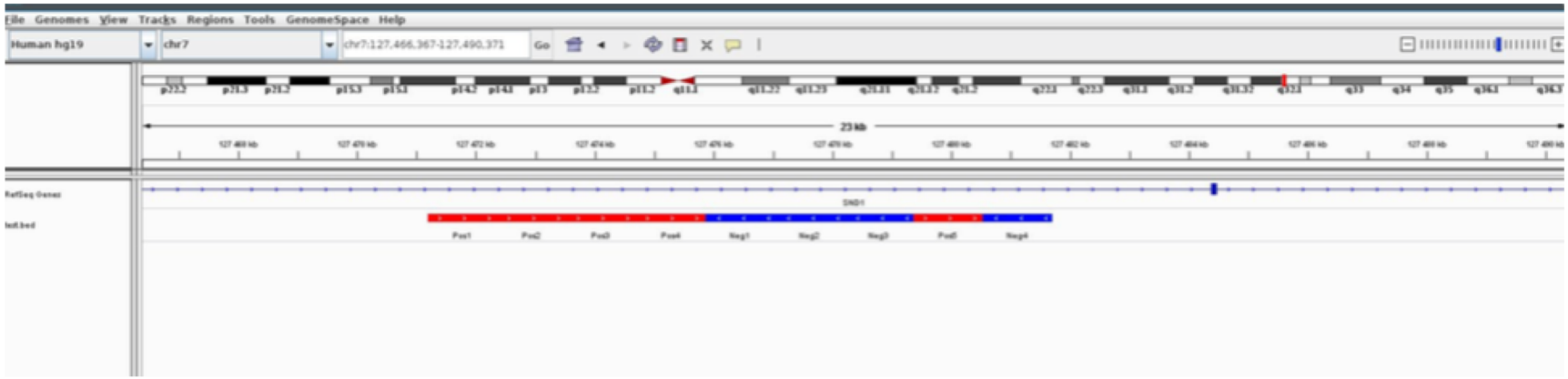
```
chr7 127471196 127472363 Pos1 0 + 127471196 127472363 255,0,0
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 0,255,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864
255,0,255
```

Attention

Le premier nucléotide est numéroté 0.

end - start = taille de la séquence

Le format bed est lisible par les navigateurs de génome



IGV genome browser

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

chr9	hg38_refGene	stop_codon	133255666	133255668	0.000000	-	.	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	CDS	133255669	133256356	0.000000	-	1	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	exon	133255176	133256356	0.000000	-	.	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	CDS	133257409	133257542	0.000000	-	1	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	exon	133257409	133257542	0.000000	-	.	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	CDS	133258097	133258132	0.000000	-	1	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	exon	133258097	133258132	0.000000	-	.	gene_id	NM_		NM_020469;
chr9	hg38_refGene	CDS	133259819	133259866	0.000000	-	1	gene_id	NM_		NM_020469;
chr9	hg38_refGene	exon	133259819	133259866	0.000000	-	.	gene_id	NM_		NM_020469;
chr9	hg38_refGene	CDS	133261318	133261374	0.000000	-	1	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	exon	133261318	133261374	0.000000	-	.	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	CDS	133262099	133262168	0.000000	-	2	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	exon	133262099	133262168	0.000000	-	.	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	CDS	133275162	133275189	0.000000	-	0	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	start_codon	133275187	133275189	0.000000	-	.	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	exon	133275162	133275214	0.000000	-	.	gene_id	NM_020469;	transcript_id	NM_020469;

Features

#9

Récupérer séquences (fasta) et annotations (gtf) du génome humain

```
wget http://hgdownload.soe.ucsc.edu/goldenPath/hg19/chromosomes/chr18.fa.gz


wget
ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_24/GRCh37_ma
pping/gencode.v24lift37.basic.annotation.gtf.gz


(curl sur MacOS)
```