Introduction à l'analyse de transcriptome par RNA-seq

Daniel Gautheret

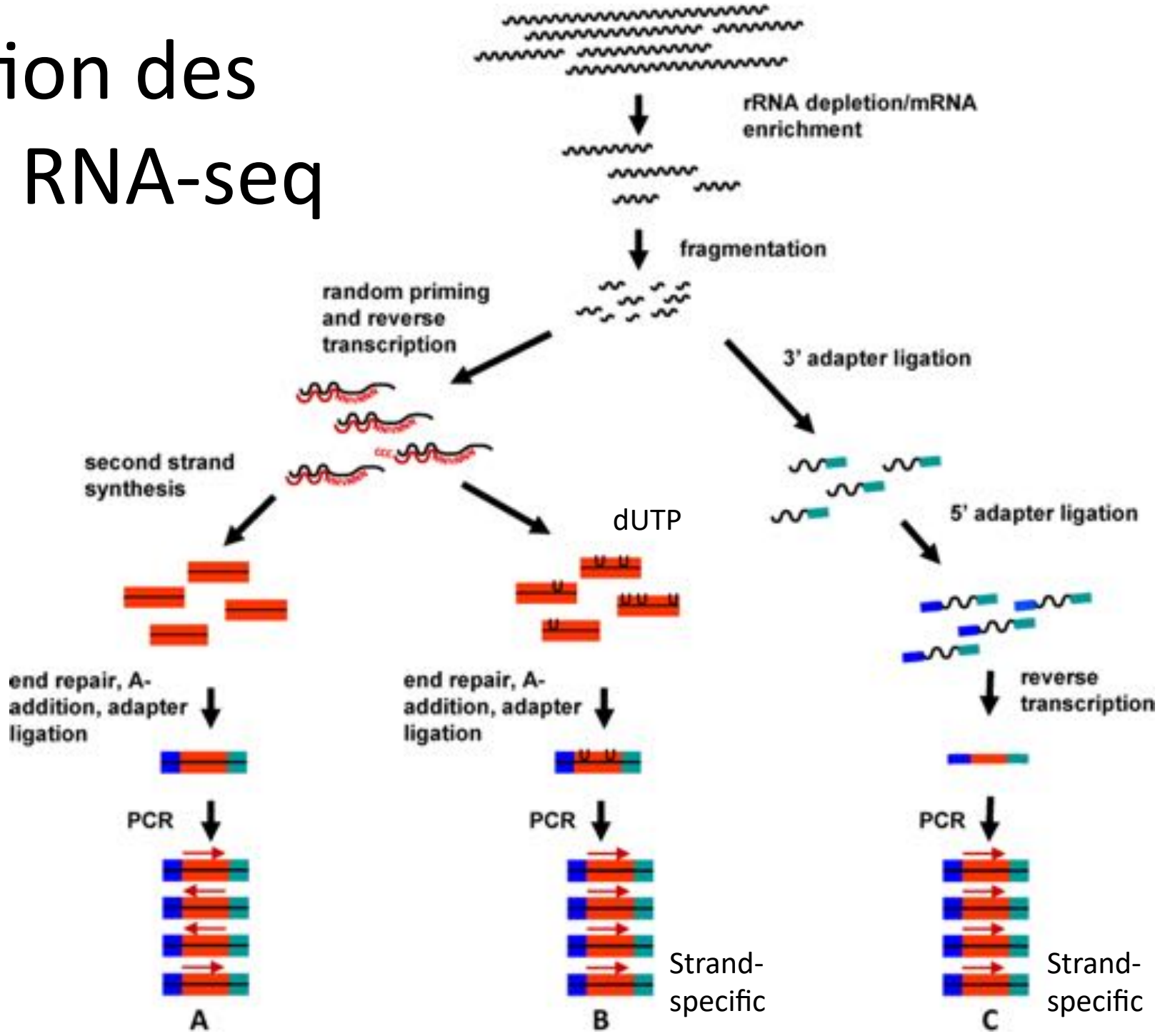
Avec des diapos de:
Yannick Boursin, IGR
Frédéric Lemoine, Institut Pasteur
Sacha Schutz

UNIVERSITÉ PARIS SUD

V. 2018.1

GUSTAVE ROUSSY
CANCER CAMPUS GRAND PARIS

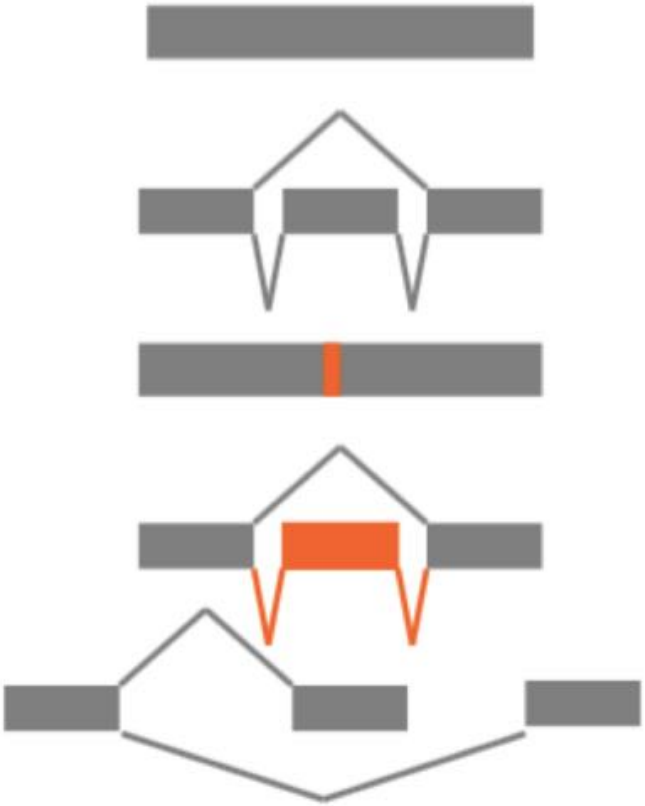
Préparation des banques RNA-seq



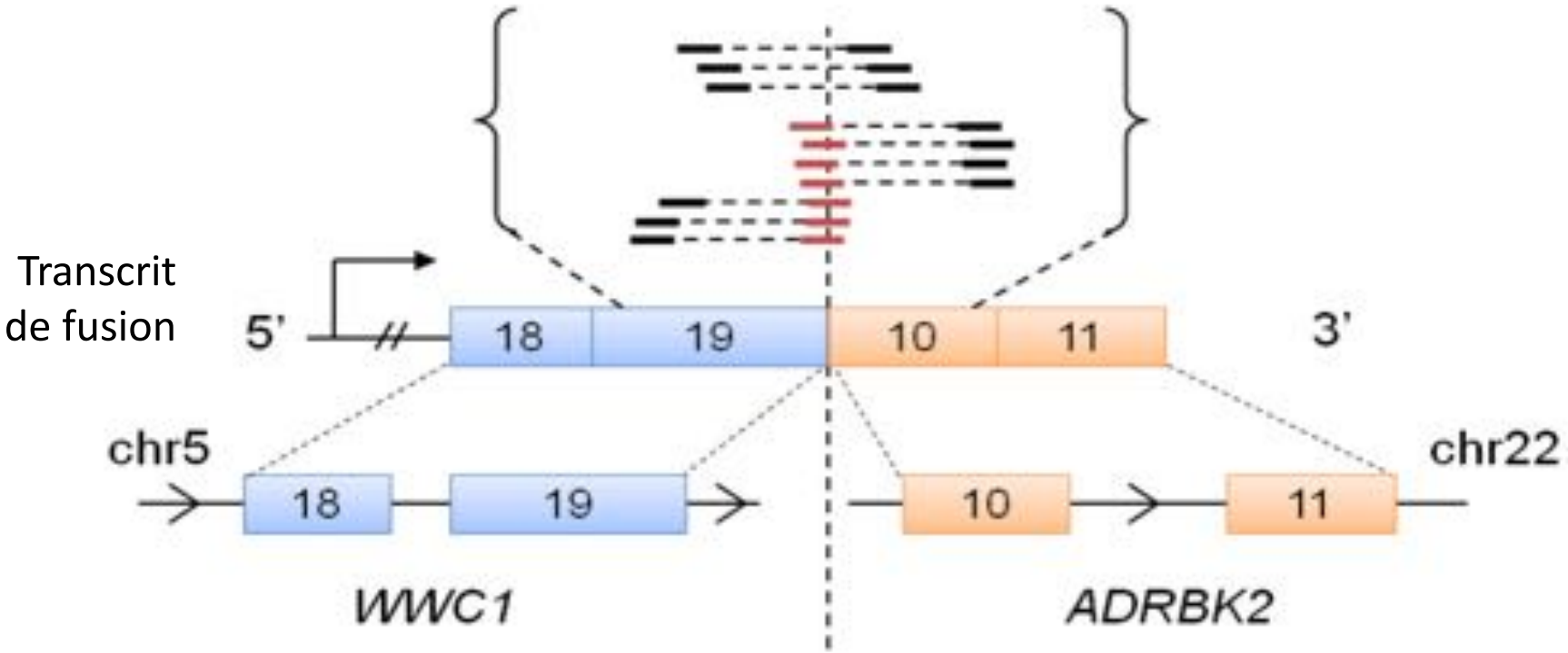
Van Dijk et al.
Exp Cell. Res. 2014

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

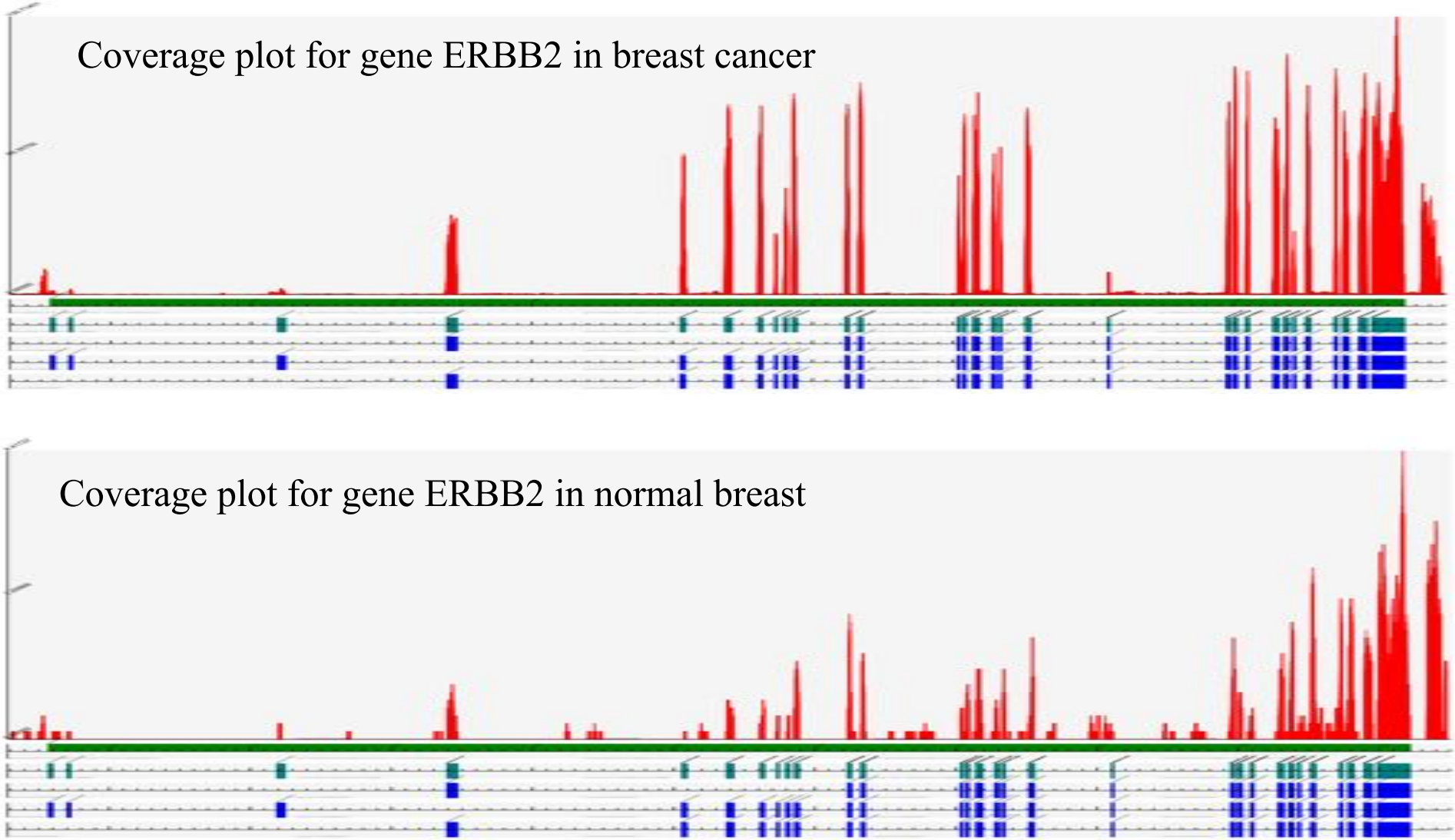


RNA-seq application Cancer #1: Découverte de transcrits de fusion

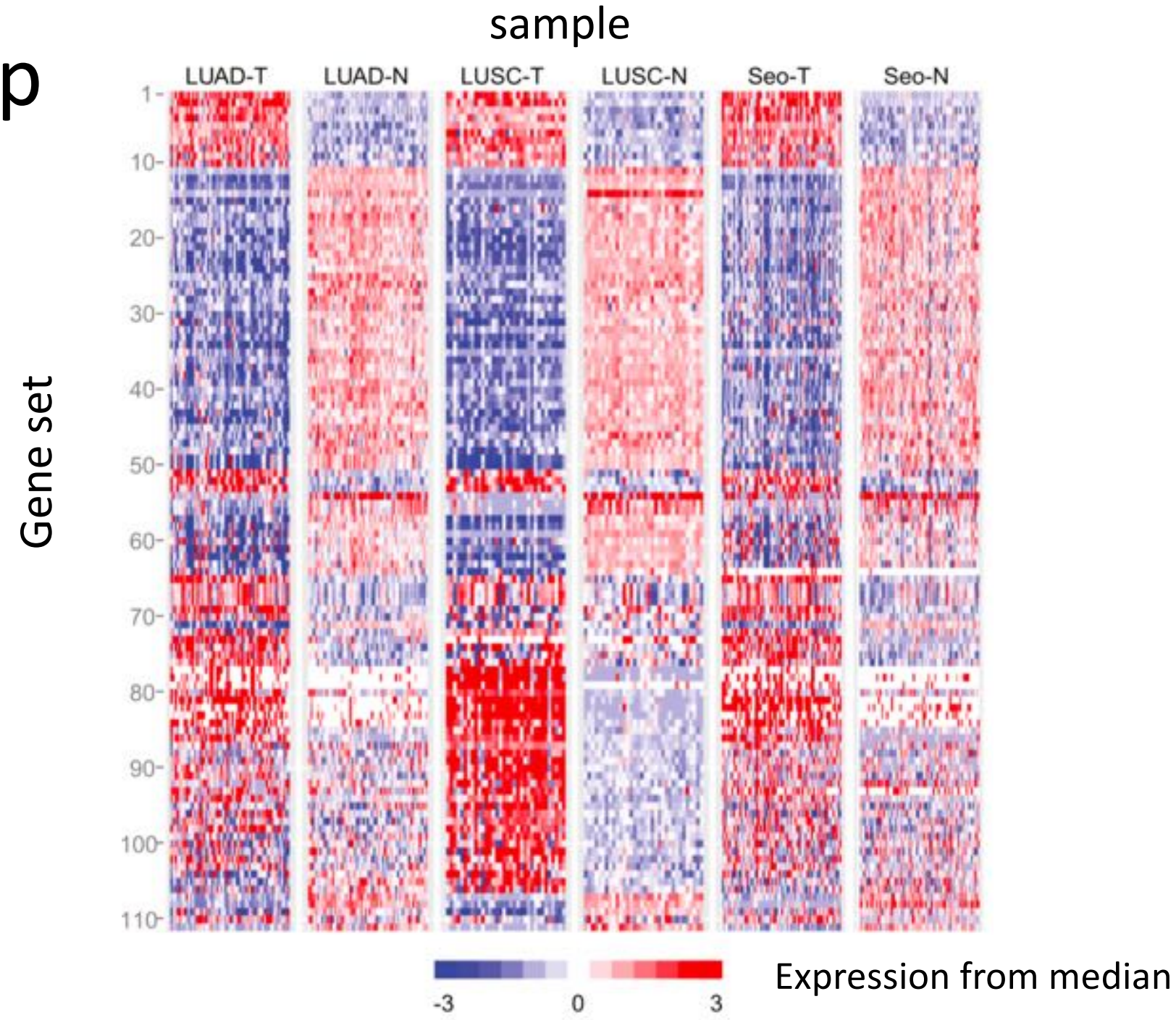


Ha et al. BMC Medical Genomics 2011

RNA-seq application Cancer #2: Expression profiling



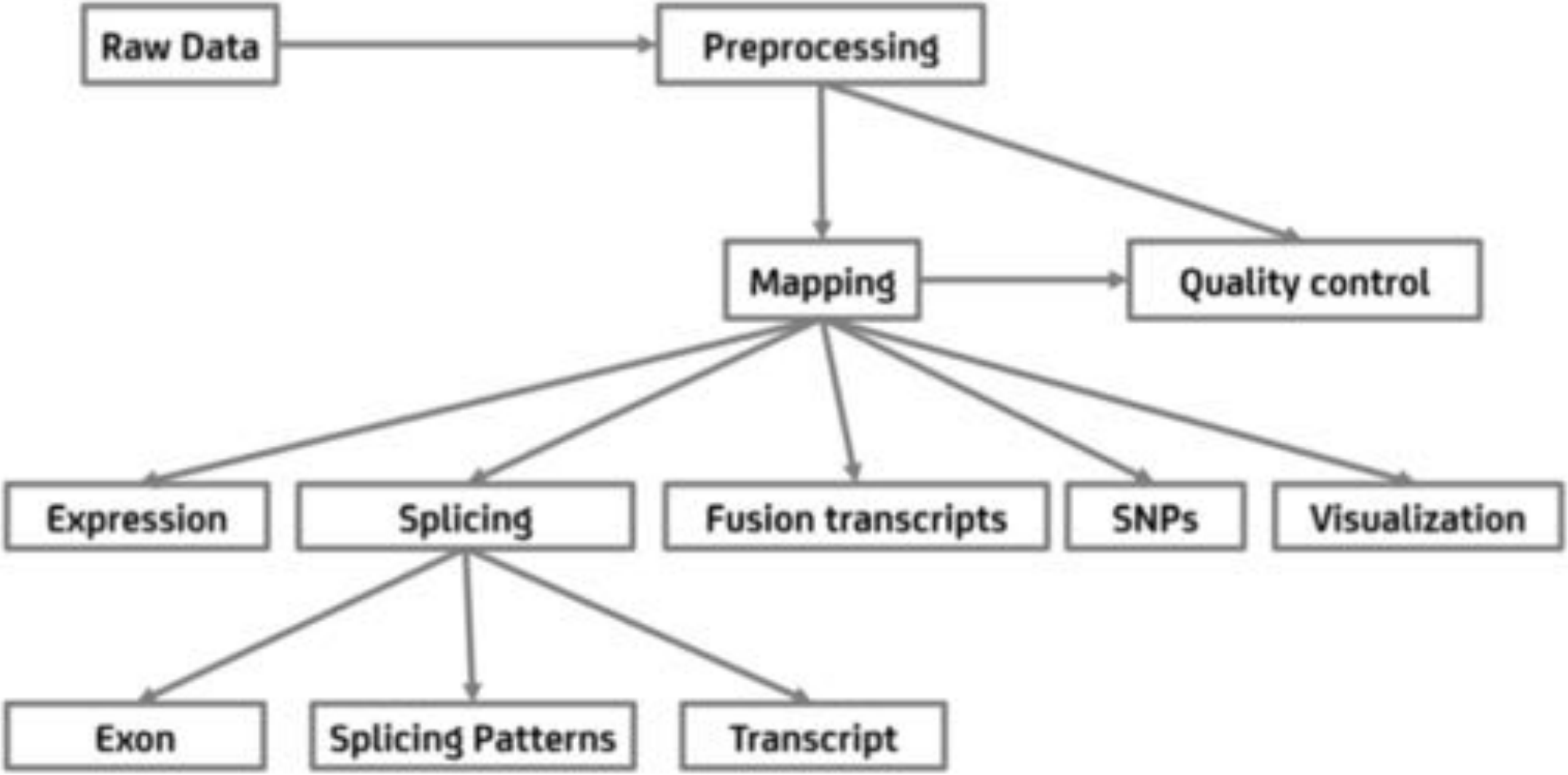
Heatmap

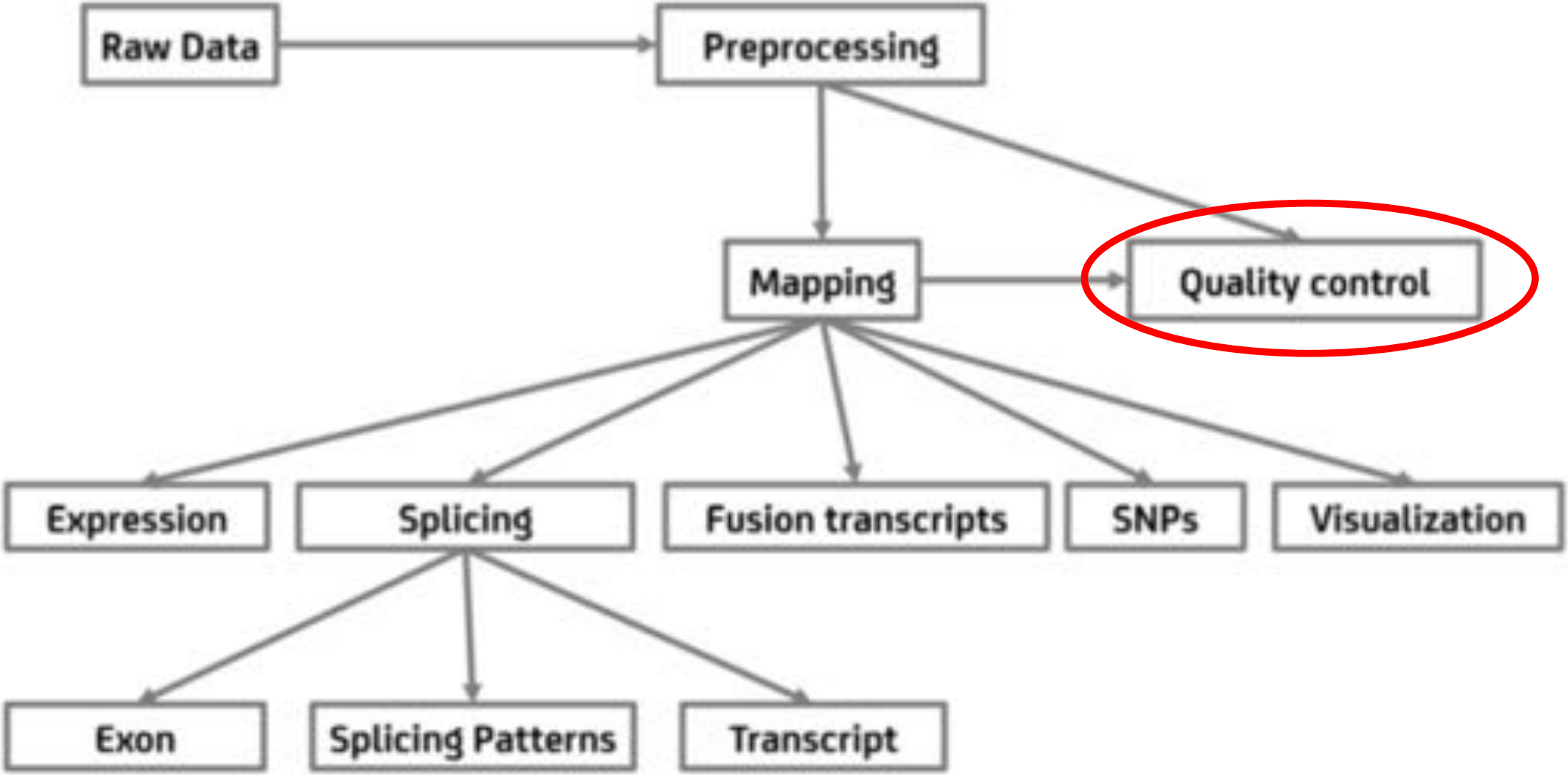

sample
LUAD-T
LUAD-N
LUSC-T
LUSC-N
Seo-T
Seo-N
Gene set
1
10
20
30
40
50
60
70
80
90
100
110
-3
0
3
Expression from median

Variations on RNA-seq

- Small RNA-seq
  - Size filter <30nt (for siRNA, miRNA)
- polyA+ vs. ribozero RNA-seq

Un pipeline d'analyse RNA-seq




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

Quality controls on raw reads: which metrics to check ?

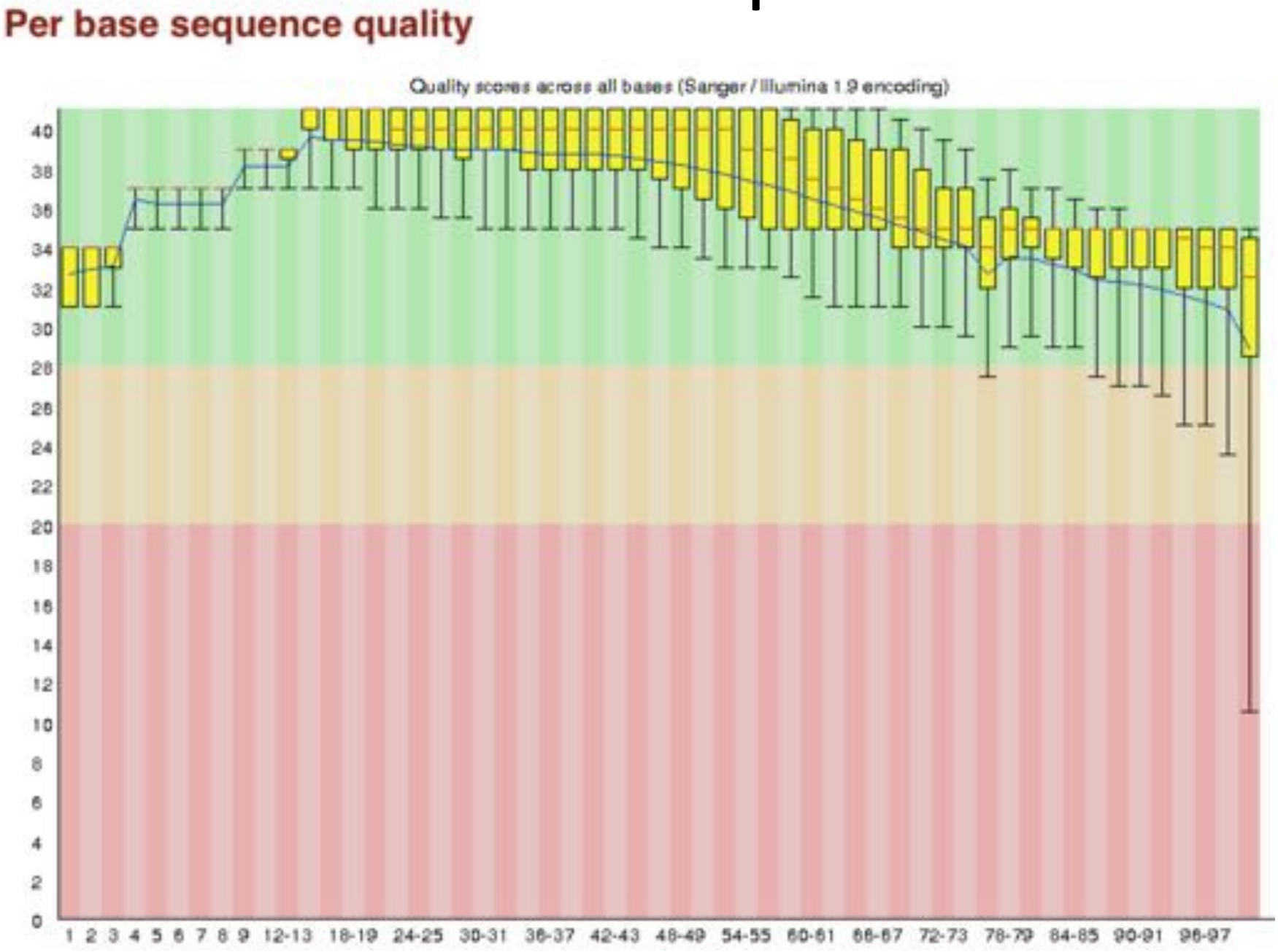
Mainly:

- Quality score per base and over the reads

But also:

- Read length distribution
- Sequence content per base and % of GC
- K-mers content
- Overrepresented sequences
- Duplicated reads

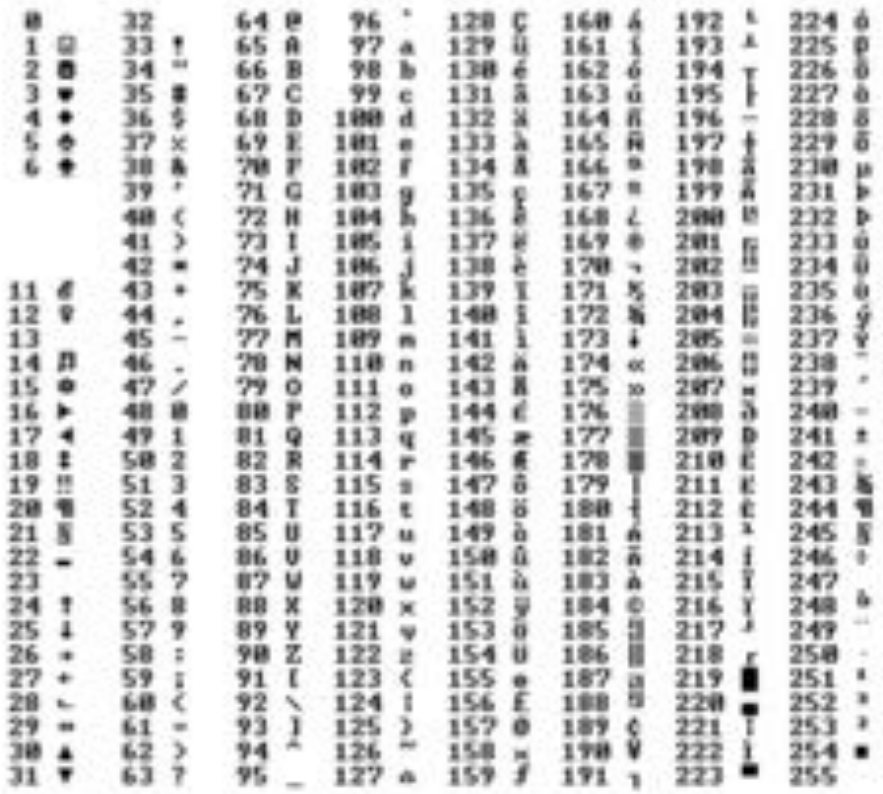
fastqc


Per base sequence quality
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97

Qualité dans le format fastq

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

```
@SEQ_ID1
GATTTGGGGT
+
*(a/$#((((
```



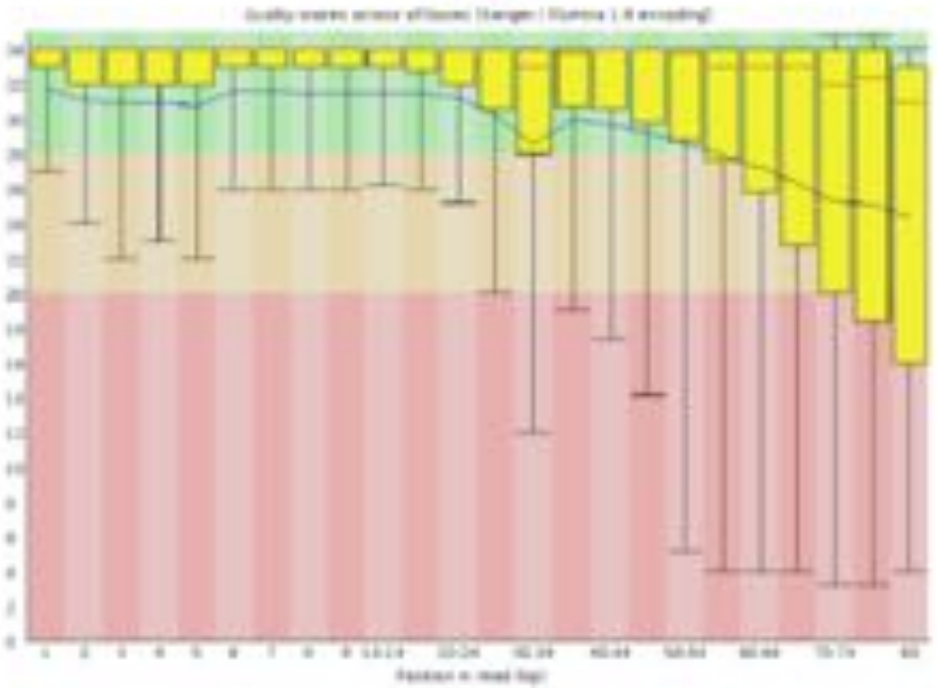
Qualité=-10 $\log_{10}(P_{erreur})$

Score de qualité phred	Probabilité d'une identification incorrecte	Précision de l'identification d'un base
10	1 pour 10	90 %
20	1 pour 100	99 %
30	1 pour 1000	99.9 %
40	1 pour 10000	99.99 %
50	1 pour 100000	99.999 %

Fastqc

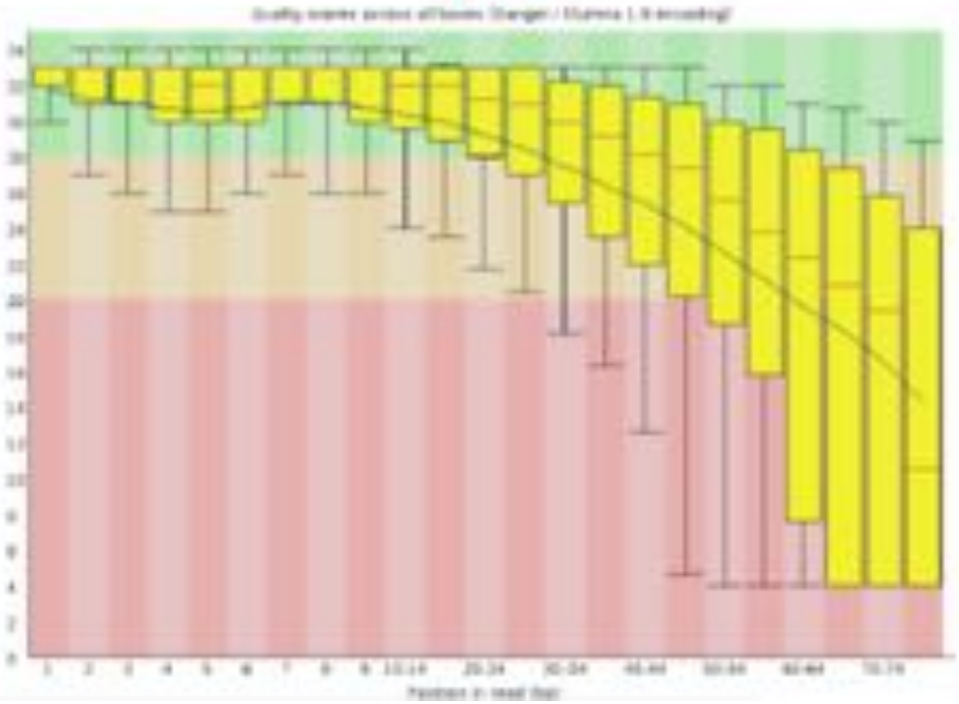
CG10128_RNAi_2

PASS



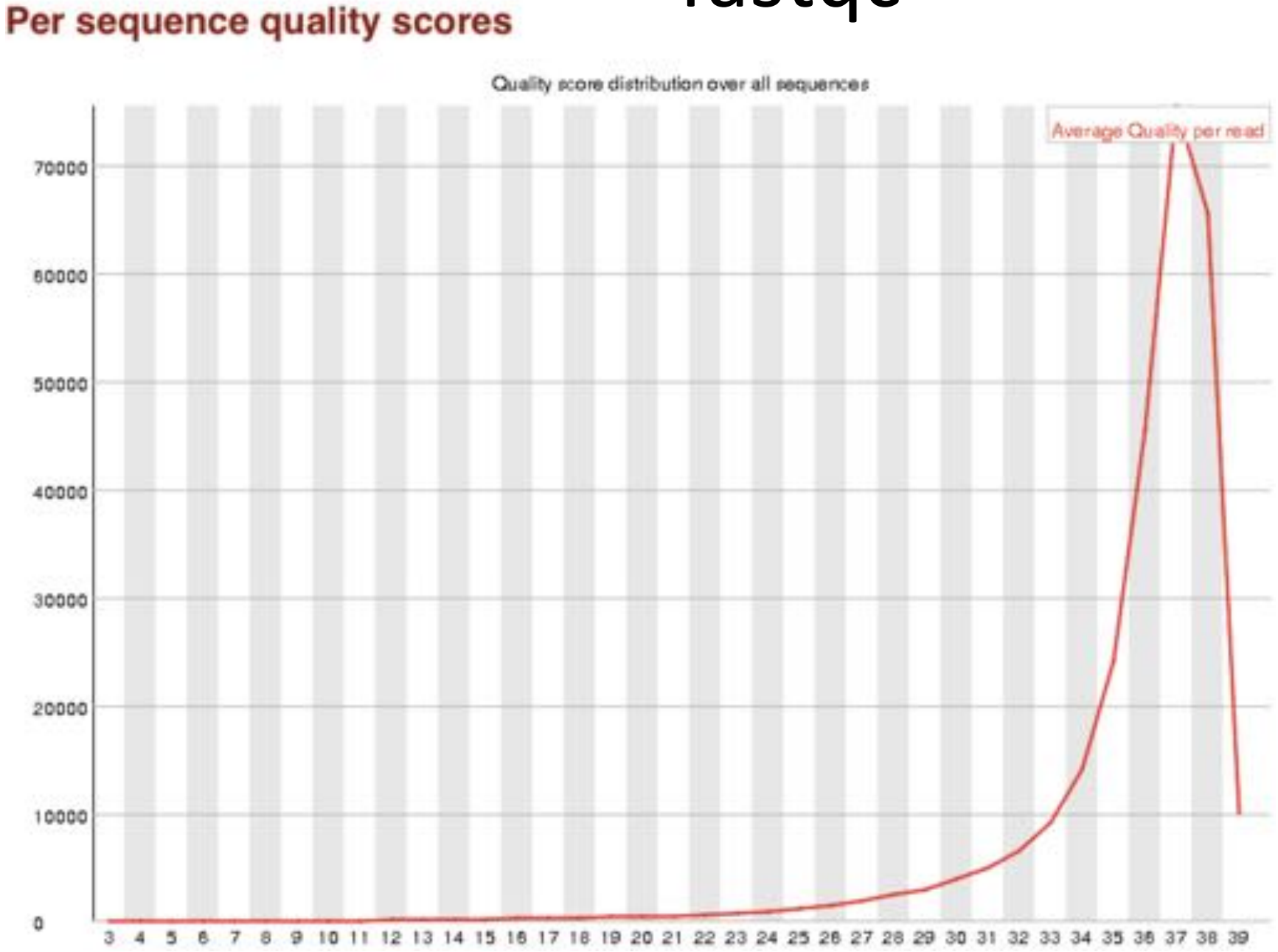
CG10203_RNAi_2

FAIL

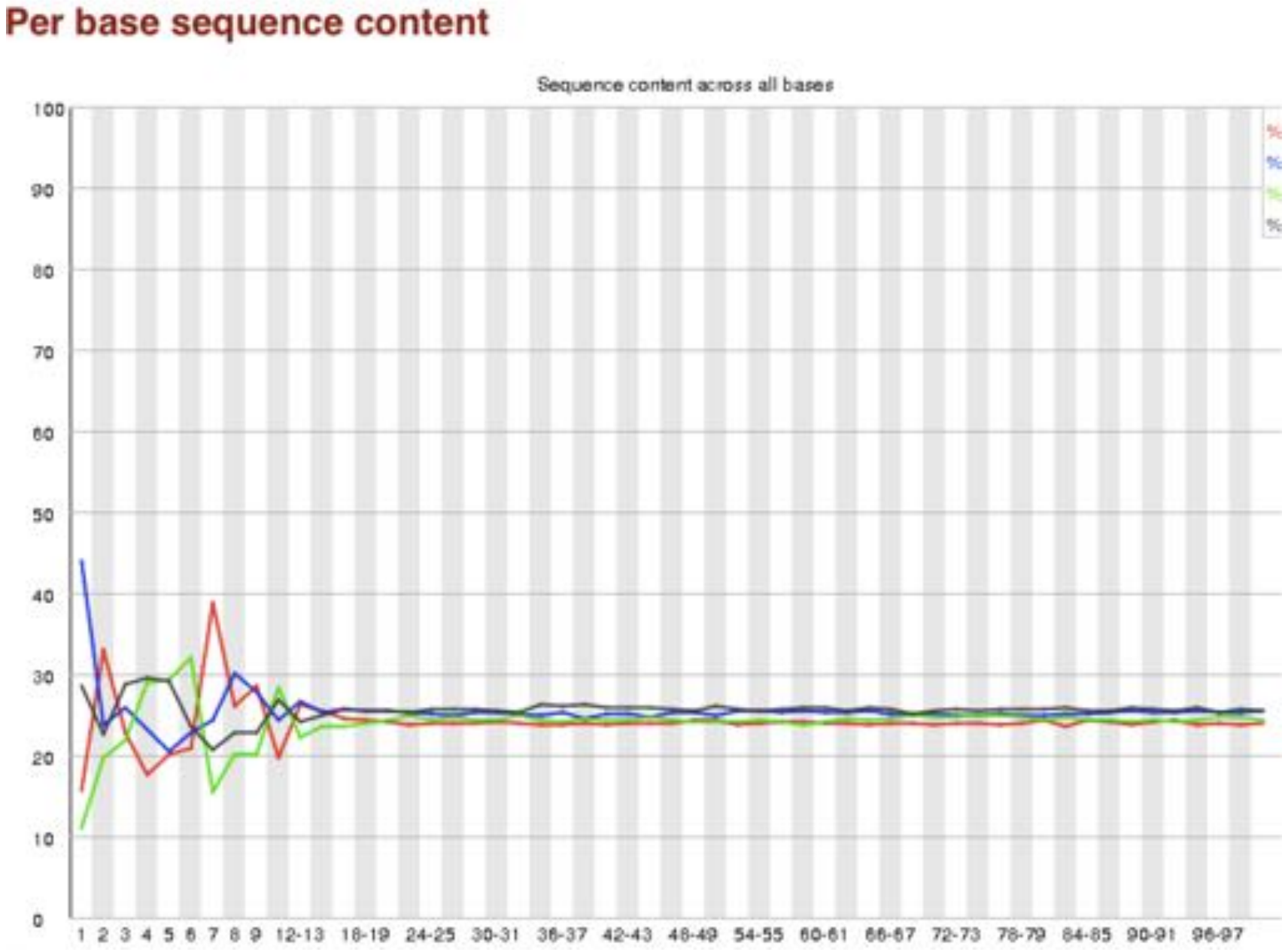


Pour couper les extrémités de basse qualité: FastqTrimmer

fastqc


Per sequence quality scores
Quality score distribution over all sequences
Average Quality per read
70000
60000
50000
40000
30000
20000
10000
0
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39

fastqc

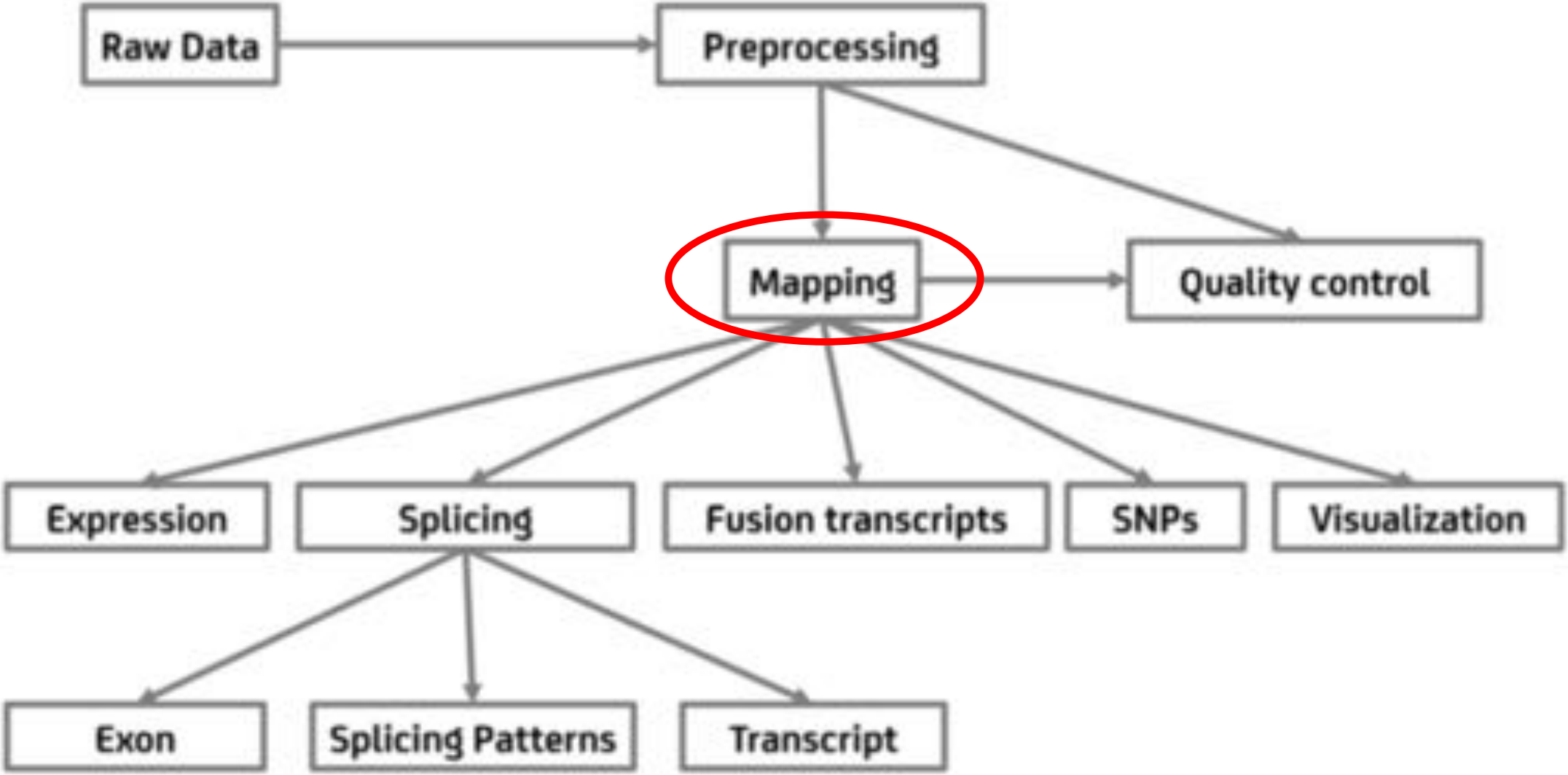

Per base sequence content
Sequence content across all bases
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85 90-91 96-97

fastqc

- Adaptateurs et séquences surreprésentées

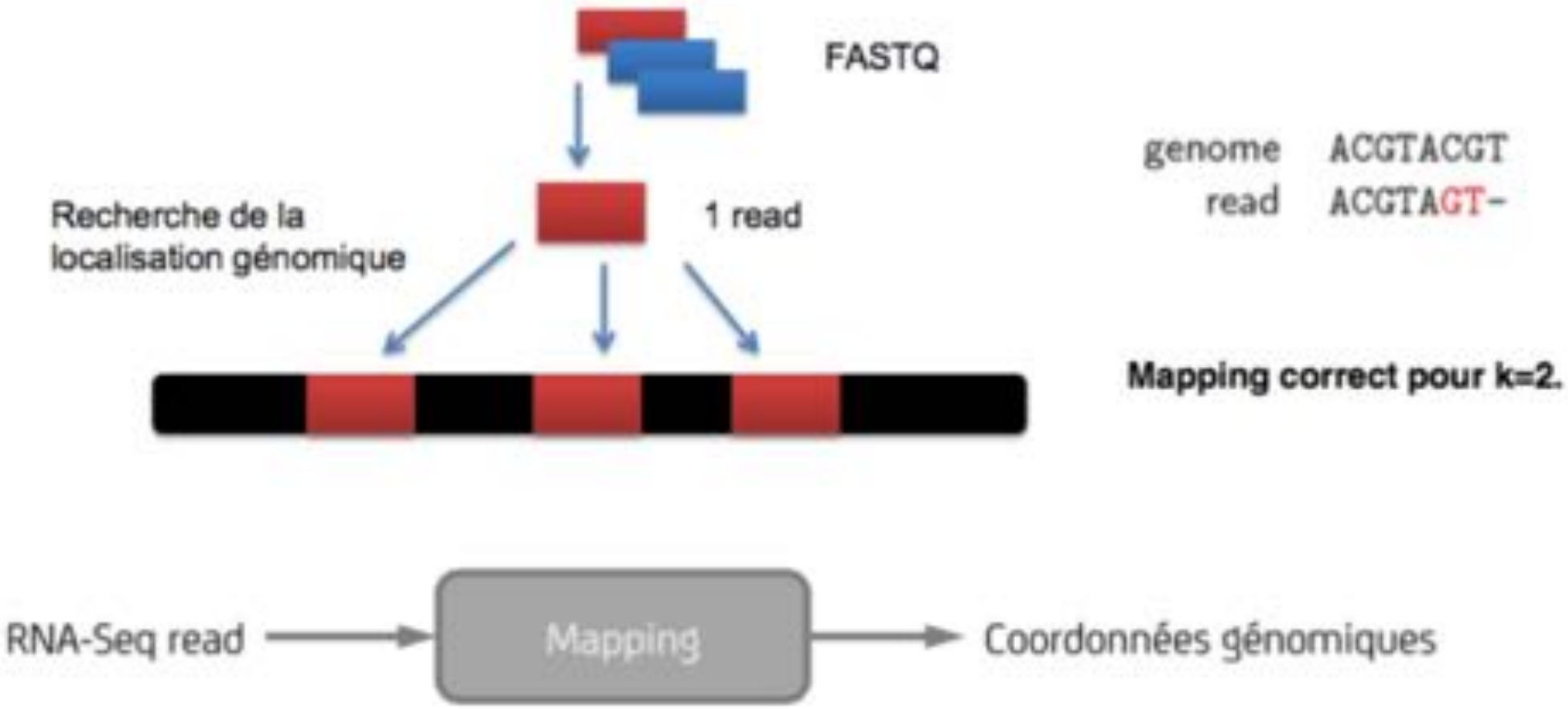
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TTTTTTTGGAAACCTCTGCGCCATGAGAGCCAAGTGGAGGAAGAAGCGAA	608	0.22002120599123534	No Hit
TTTTTGGAAACCTCTGCGCCATGAGAGCCAAGTGGAGGAAGAAGCGAATG	478	0.17297719813126725	No Hit
CTCCAGTCAAAAGTTCTTTGAGACGATGCCATCGGCCTTGGCCAATCGGA	411	0.14873144023420679	No Hit
TTTTTTGGAAACCTCTGCGCCATGAGAGCCAAGTGGAGGAAGAAGCGAAT	356	0.1288282061396049	No Hit
GCAGGCGCAGCCCAGCCTCGAAATGCAGAACGACGCCGGCGAGTTCGTGG	337	0.1219525434523788	No Hit
CAGGCGCAGCCCAGCCTCGAAATGCAGAACGACGCCGGCGAGTTCGTGGA	308	0.11145811092977054	No Hit
CGCAGATAGCATAAGTTTTAAACTGGCCATTAAACCTGCCTGTGACCTTG	288	0.10422057125900062	No Hit


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.



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(3.10e9 x 100)

Mais pour 1e8 reads???

L’algorithme de BLAST

- Dictionnaire de k-mots de la référence
- Recherche des k-mots de la query dans le dictionnaire
- Extension autour des k-mots par alignement

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$		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.

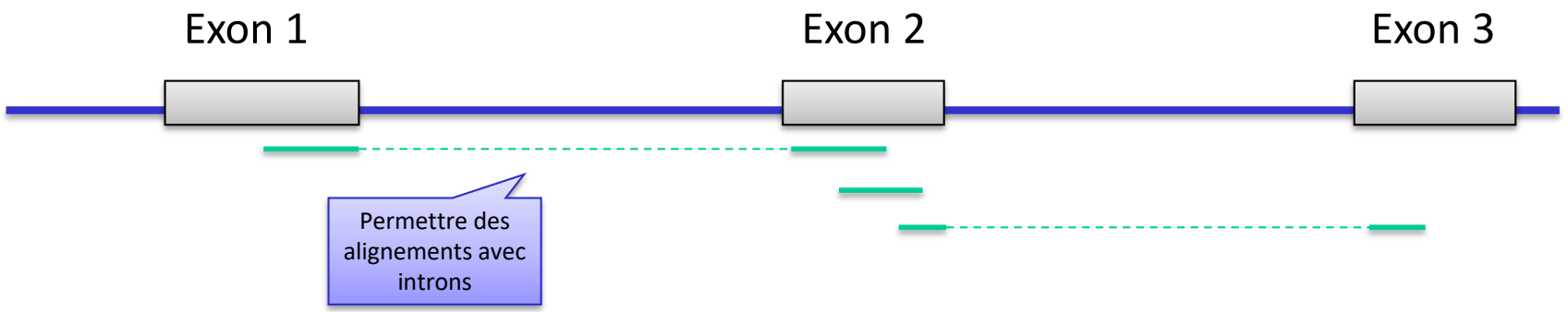
Suffix arrays

Exemple: trouver la chaîne **GO**

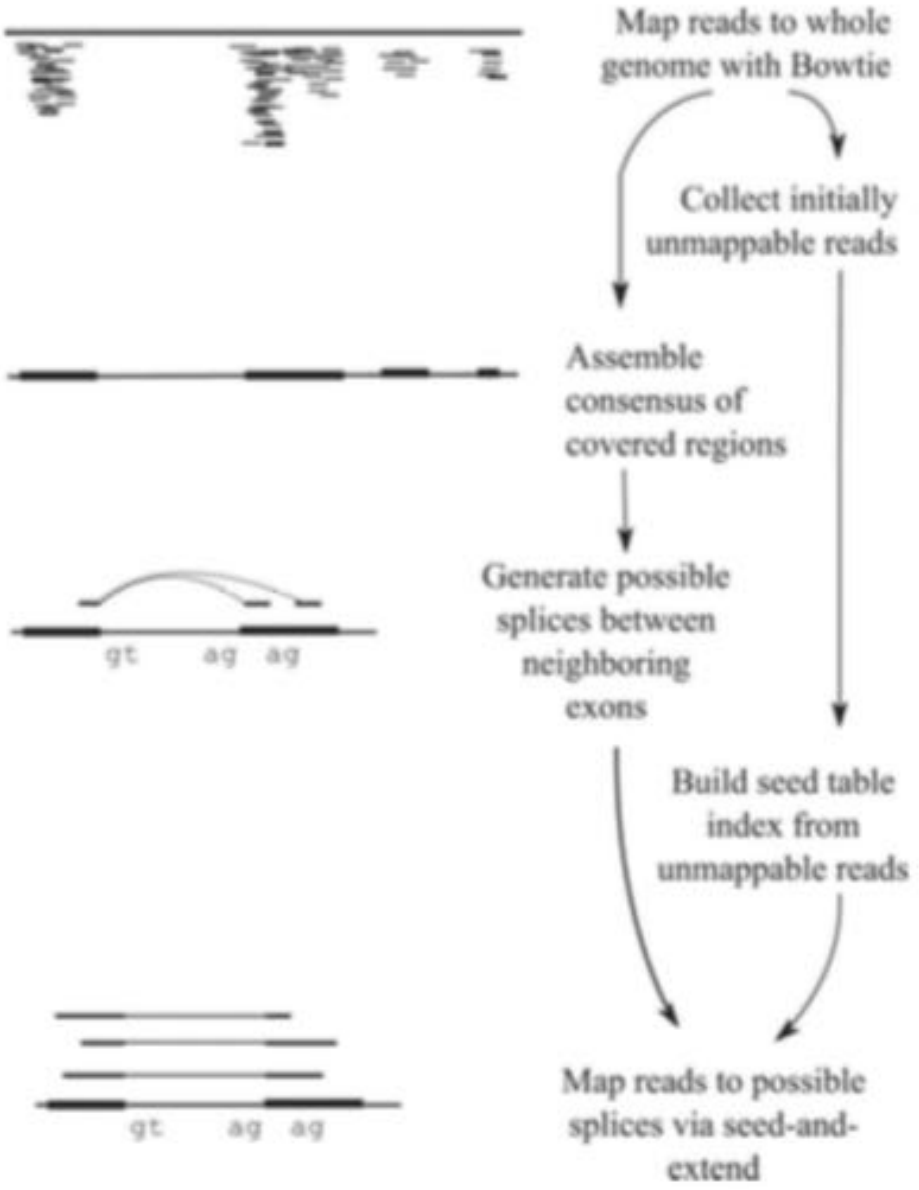
Les algorithmes de mapping

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
Soruman				X		

La spécificité des reads RNA-seq

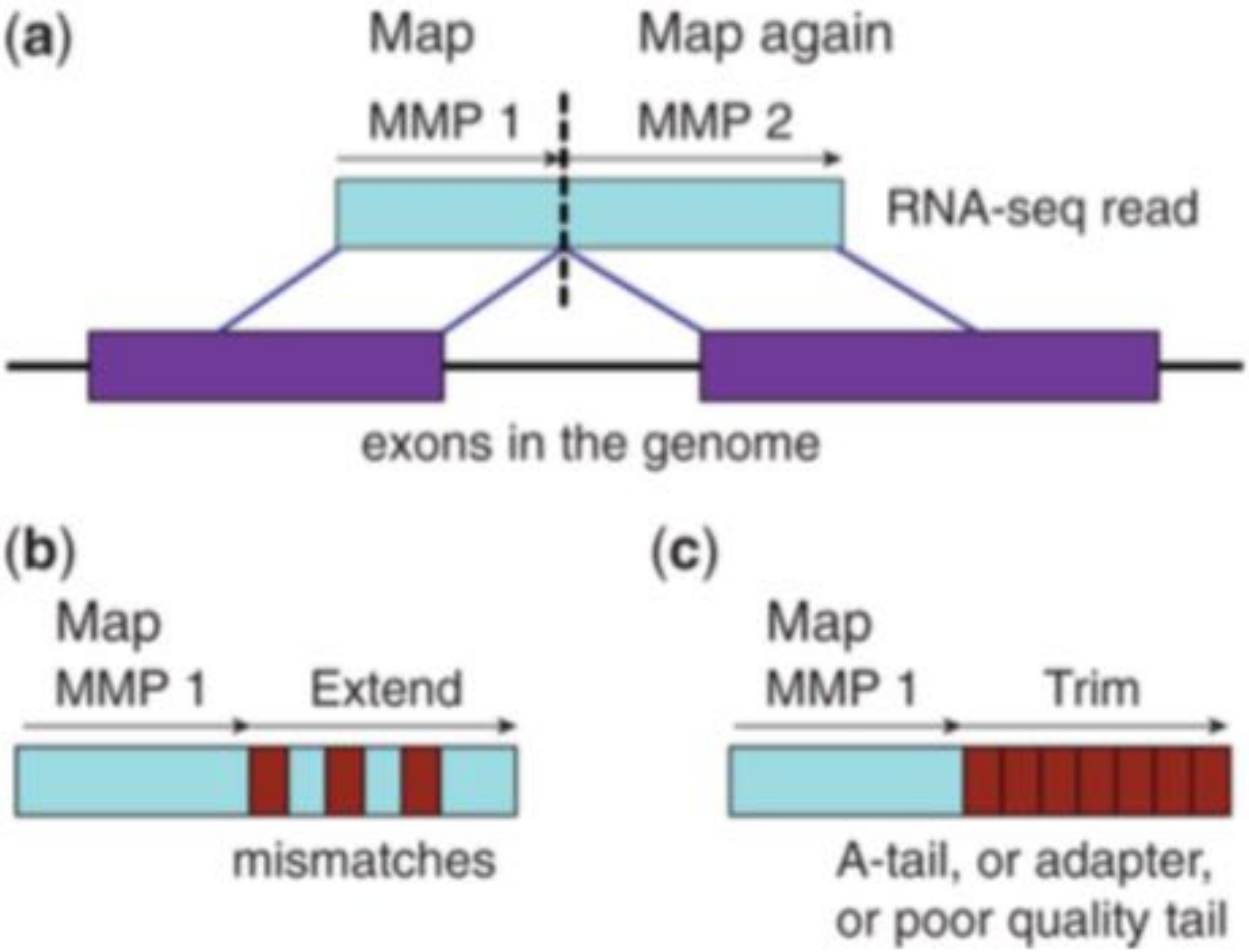


Le programme TopHat (HiSat)



TopHat: Trapnell et. al. Bioinformatics, 2009
HiSat: Kim et al. Nat. Methods, 2015

Le programme STAR



Dobin et. al. Bioinformatics, 2013

La mappabilité: une partie du génome reste invisible

		H.sapiens (hg19)	*M.musculus* (mm9)	*D.melanogaster* (dm3) with het.	*D.melanogaster* (dm3) without het.	*C.elegans* (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%)

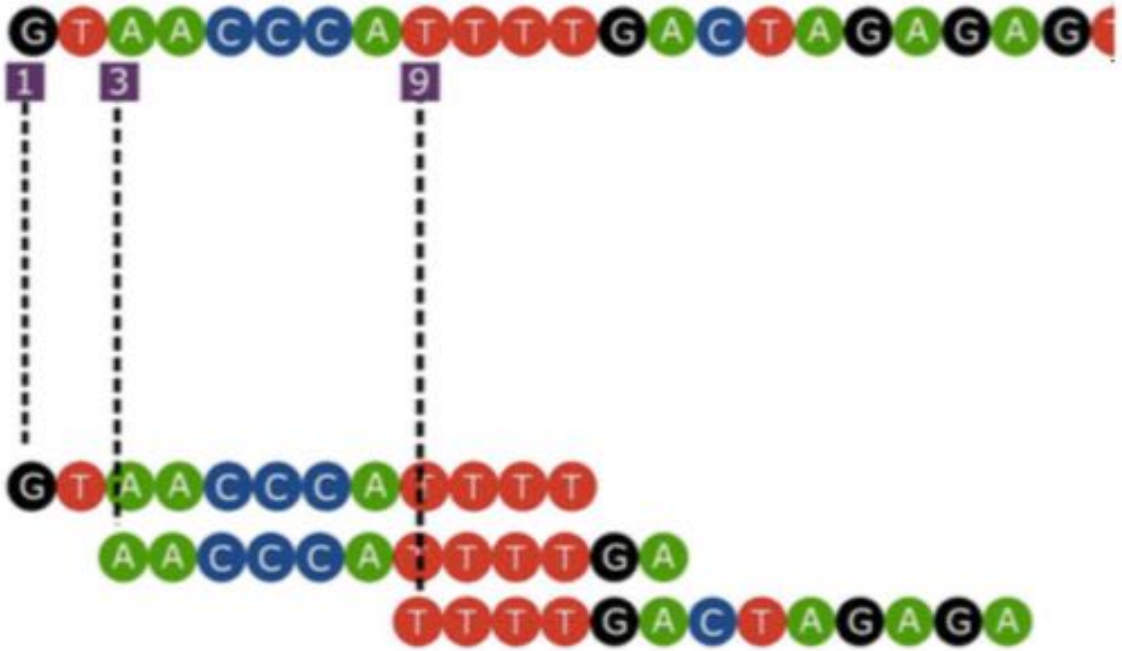
Repeat elements have been identified and classified by the RepeatMasker program [37]. The mappability has been computed for $k = 36, 50$ and 75, with $m = 0$ and 2.
doi:10.1371/journal.pone.0030377.t002

Derrien et al. PloS One 2012

Après le mapping

Format SAM

Contient les reads alignés sur le génome.



Concept:

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

SAM format

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

...

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:

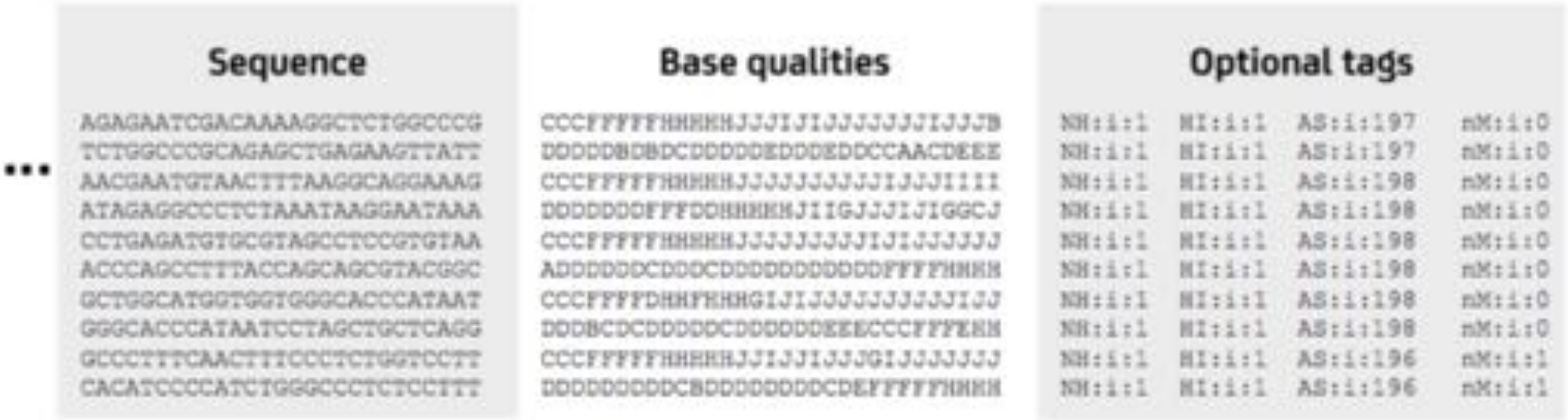
- **Decimal Flag Value** 83
- **Binary Flag Value**

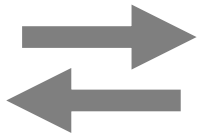
2048	1024	512	256	128	64	32	16	8	4	2	1
0	0	0	0	0	1	0	1	0	0	1	0

- **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

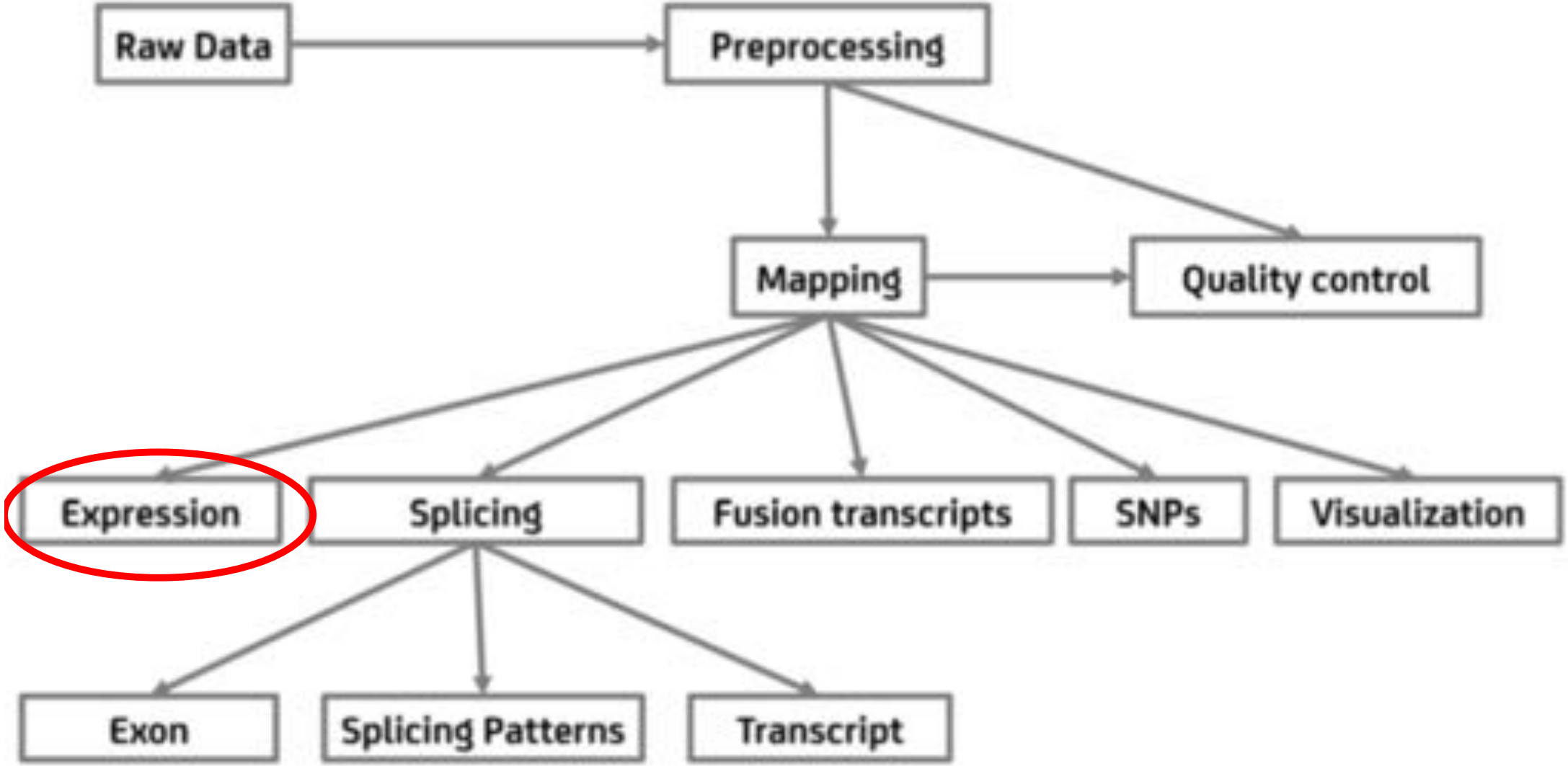

Sequence
Base qualities
Optional tags
...
AGAGAATCGACAAAAGGCTCTGGCCCG
TCTGGCCCGCAGAGCTGAGAAGTTATT
AACGAATGTAACTTTAAGGCAGGAAAG
ATAGAGGCCCTCTAAATAAGGAATAAA
CCTGAGATGTGCGTAGCCTCCGTGTAA
ACCCAGCCTTTACCAGCAGCGTACGGC
GCTGGCATGGTGGTGGGCACCCATAAT
GGGCACCCATAATCCTAGCTGCTCAGG
GCCCTTTCAACTTTCCCTCTGGTCCTT
CACATCCCCATCTGGGCCCTCTCCTTT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

SAM  BAM

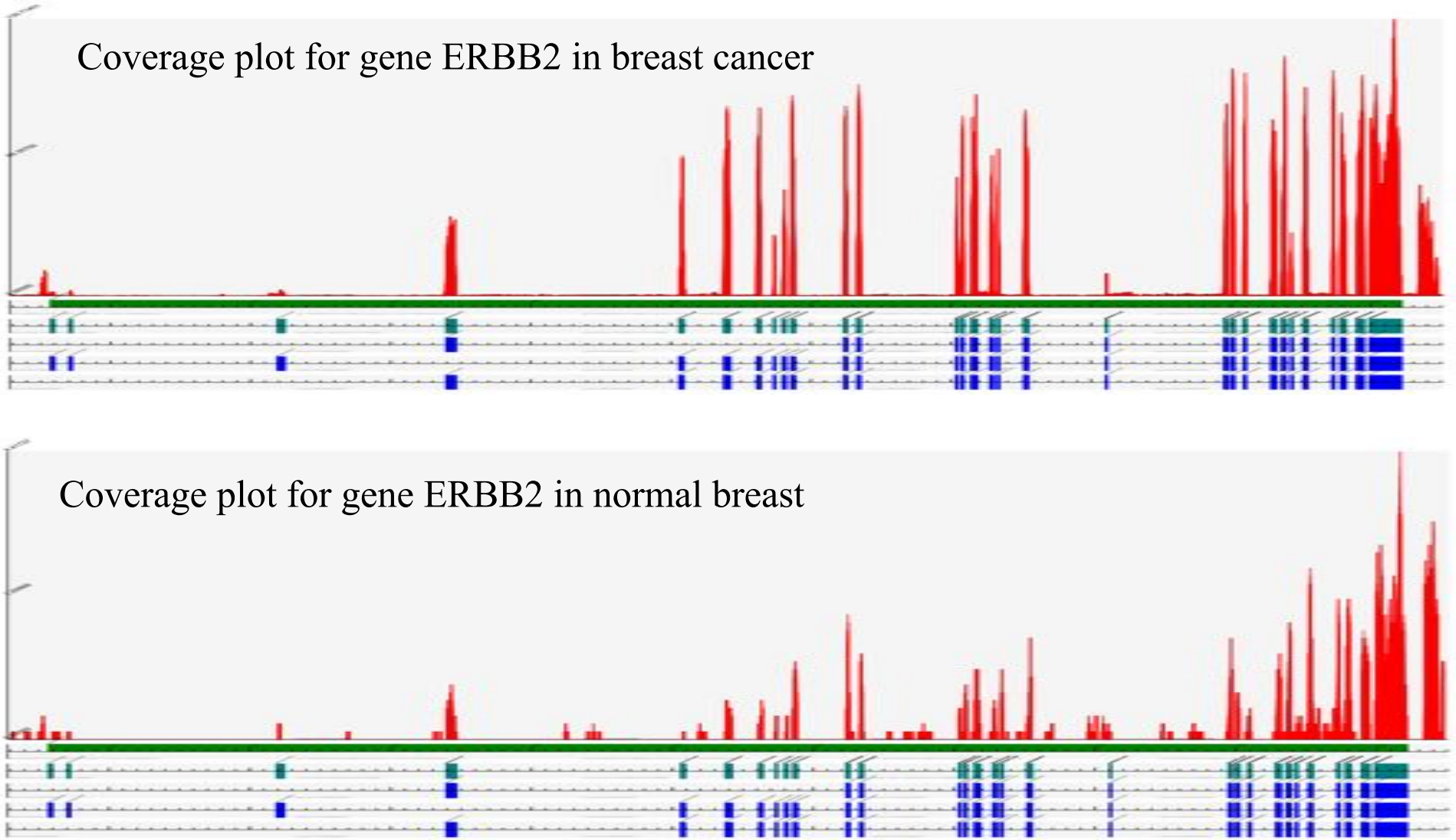
SAMtools

Fichier texte

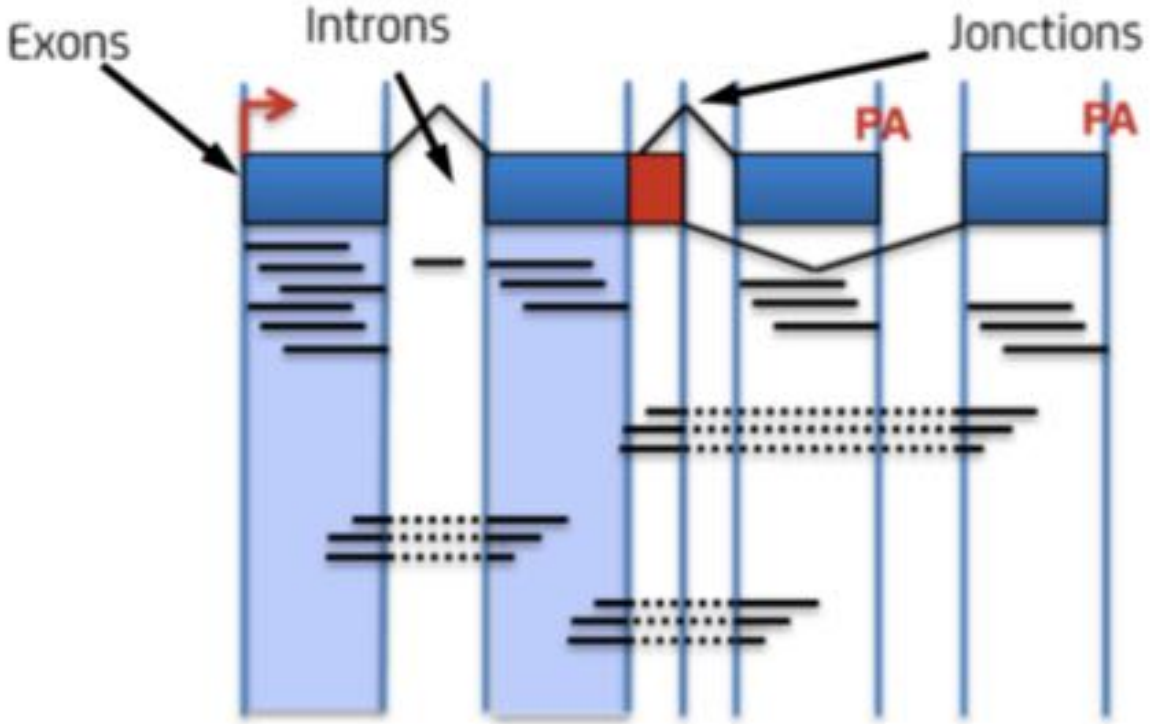
Fichier binaire


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

Differential expression



Mesure de l’expression par RNA-seq



Entrée 1: Fichiers BAM indexés

- 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
```

Entrée 2: fichier de features (format GFF ou GTF)

GFF:

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_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	CDS	133259819	133259866	0.000000	-	1	gene_id	NM_020469;	transcript_id	NM_020469;
chr9	hg38_refGene	exon	133259819	133259866	0.000000	-	.	gene_id	NM_020469;	transcript_id	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;

Récupérer une annotation GTF du génome humain (Gencode)

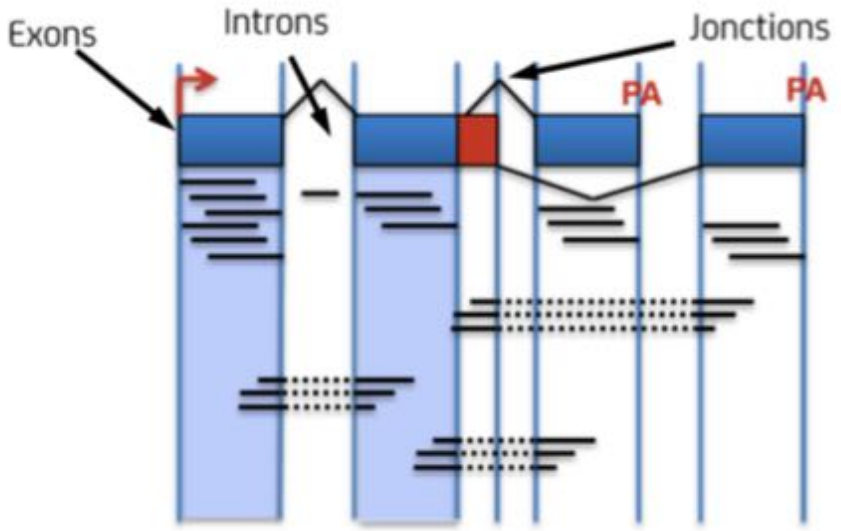
```
wget
ftp://ftp.sanger.ac.uk/pub/gencode/release_3c/genco
de.v3c.annotation.GRCh37.gtf.gz

(curl sur MacOS)
```

Comment estimer l'expression de chaque gène?

L'approche « coverage »

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



Normalisation

Number of mapped reads is related to library size

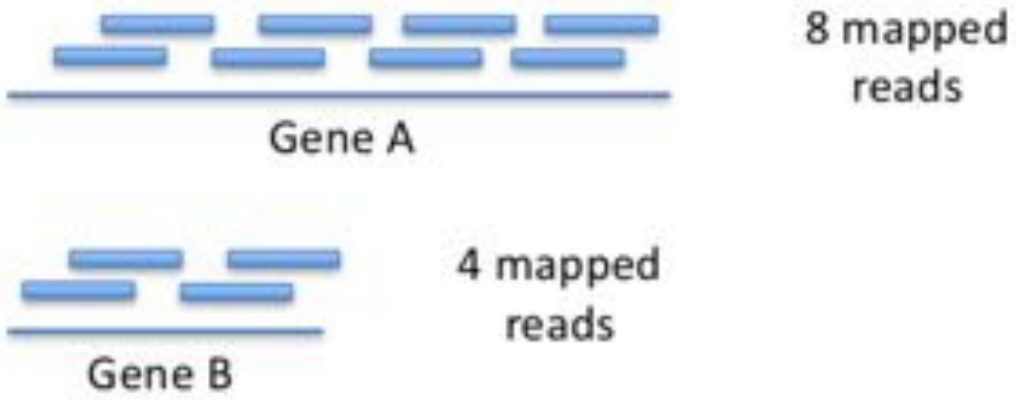
A1BG	4	7
A1CF	41	32
A2M	1	4
A2ML1	3	6
A2MP1	3	1
A3GALT2	1	3
A4GALT	420	327
A4GNT	1	1
AA06	0	0
AAAS	2452	2054
AACS	3234	1678
AACSP1	1544	1926
	Sum 1	Sum 2

Systematic bias ?

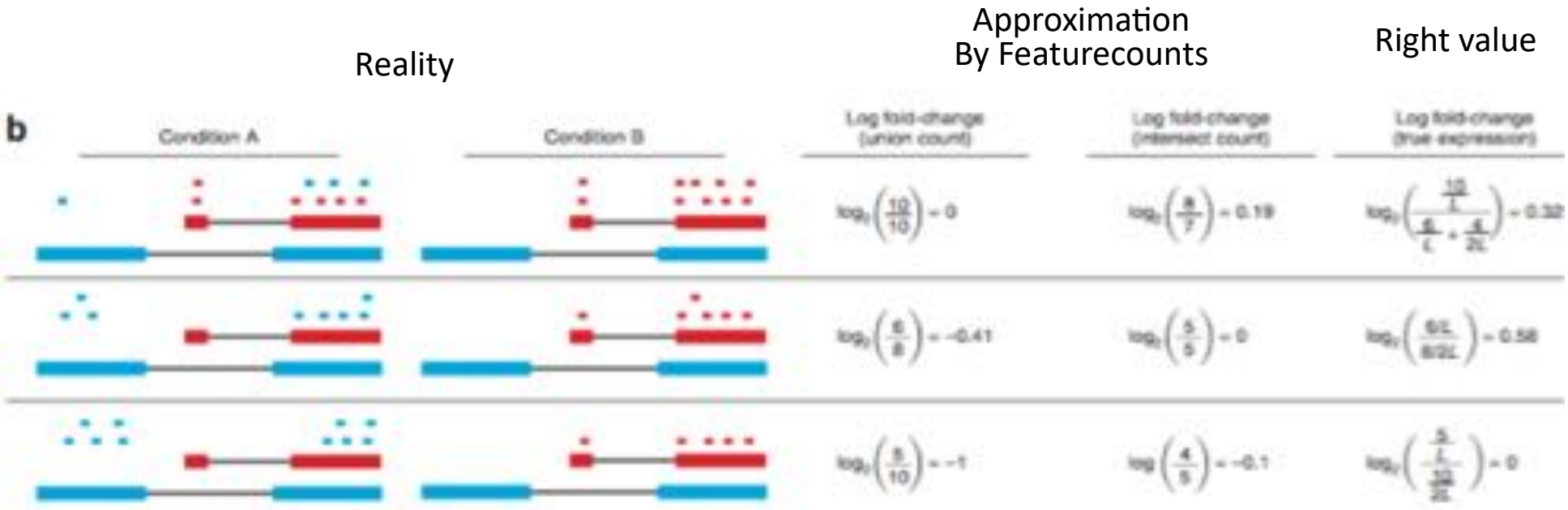
(If Sum 1 = 40.000 and Sum 2 = 30.000)

Normalisation

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



Transcrits alternatifs: pourquoi l'approche « coverage » est problématique

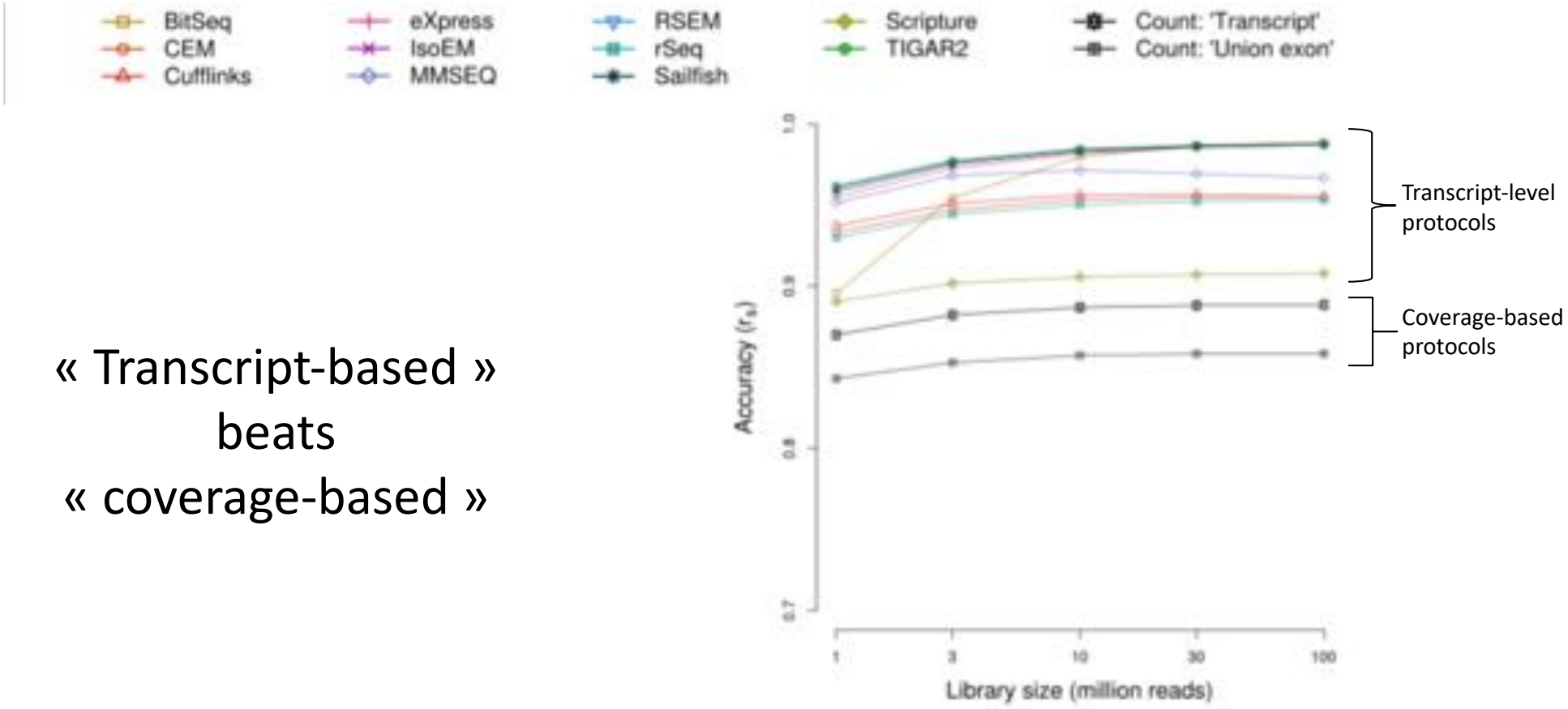


Trapnell et al. Cufflinks 2013

Count transcripts rather than genes?

Transcript-based vs. coverage-based

« Transcript-based »
beats
« coverage-based »



Accuracy of gene expression prediction

Benchmark on simulated data, by *Kanitz et al. Genome Biol. 2015*

Normalized expression units

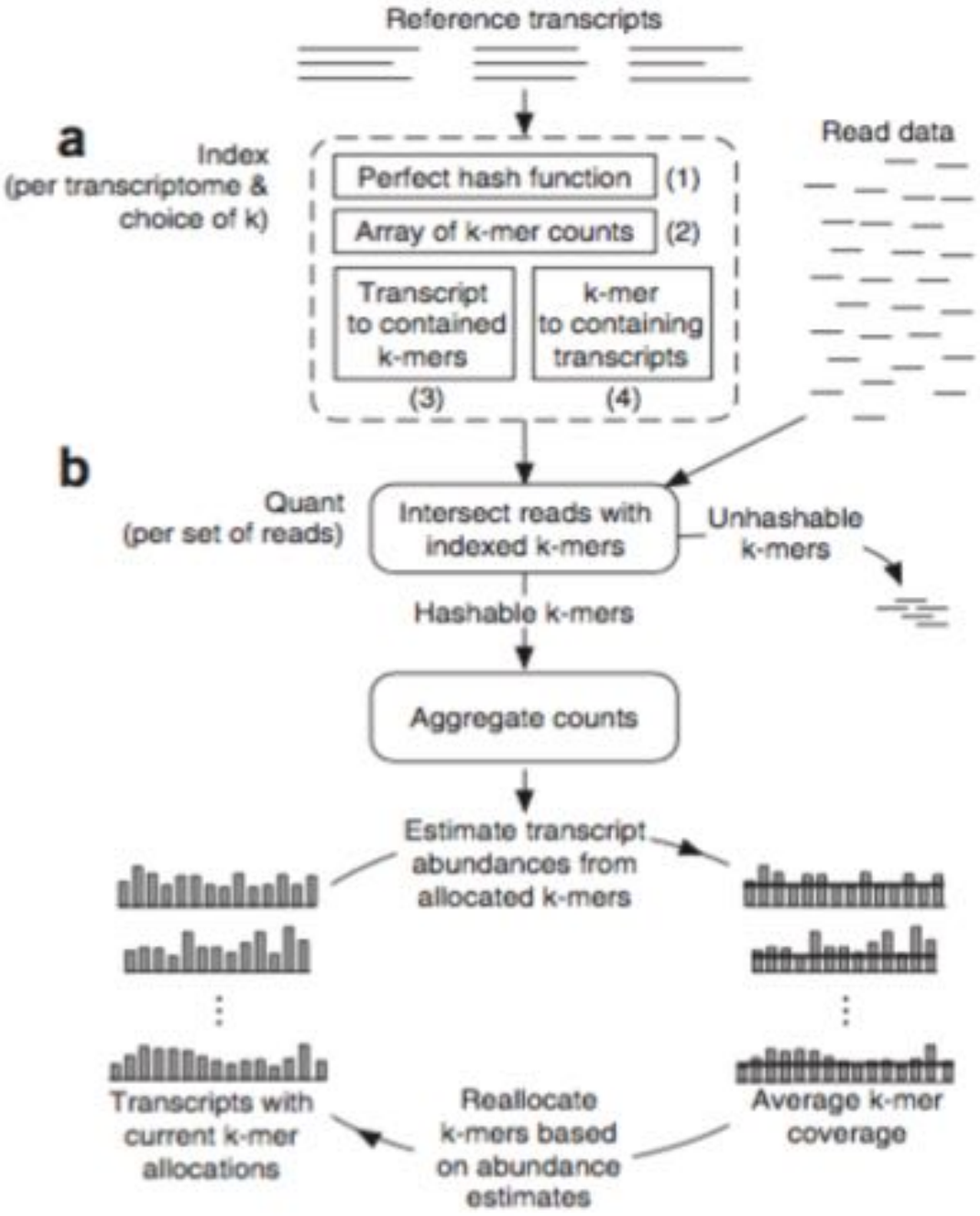
- RPM
  - Normalized by library size
- RPKM
  - Normalized by library size and gene-size
- TPM
  - Transcript-level count, normalized by library size

FeatureCount

Cufflinks, RSEM, Kallisto, Salmon

Towards mapping-free methods

Sailfish
Salmon
Kallisto



Sailfish. Patro et al. 2014

Jeu de données « EMT »

AMERICAN SOCIETY FOR MICROBIOLOGY | Molecular and Cellular Biology

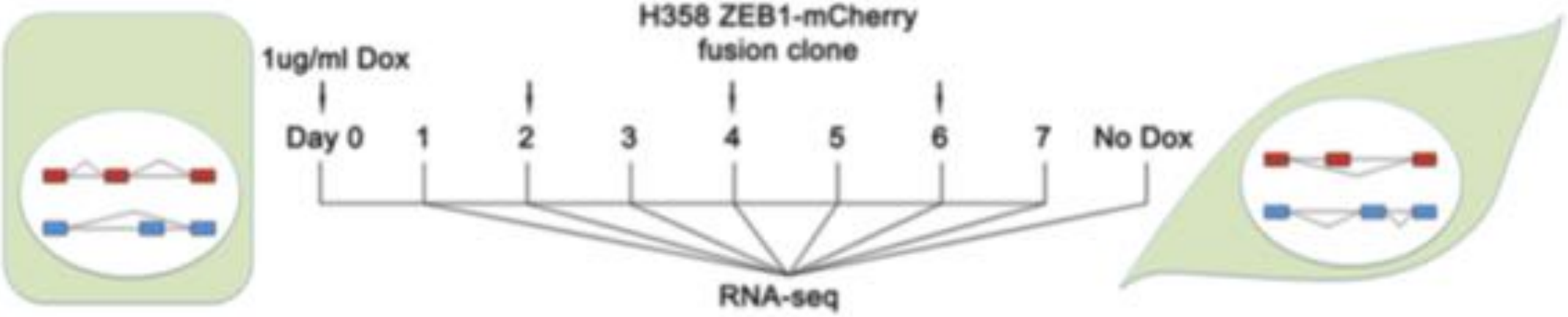
Determination of a Comprehensive Alternative Splicing Regulatory Network and Combinatorial Regulation by Key Factors during the Epithelial-to-Mesenchymal Transition

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.

Exercices IGV

- Lancer IGV
- Vérifiez génome (HG19) et annotation (refseq)
- Chargez fichier BAM indexé « Day0 » (en fait=noDox)
  - Contient uniquement map sur chr 18 + échantillonné (0.5% des reads)
- Naviguez sur le chromosome 18
  - Orientation ? (color by « first of pair »)
  - Introns, exons, annotation étendue (expanded)
- Créez fichier bed décrivant le fragment chr18:47117279-47117482 et visualisez-le
- Trouvez des SNP/indels dans les régions exprimées

Beaux loci à regarder

- Chr18:19449740-19449780 (délétion)
- Chr18:21481000-21481200 (SNP)
- Des cas d'épissage alternatif dans les gènes ZNF397, C18orf21, SLC39A6
- Un lncRNA non annoté: chr18:35328141-35335717
- Un gène différentiel entre Day0 et Day7? (nécessite de charger le BAM pour Day7)

Appendices

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