Version du document
27/01/2025

Pipeline d'analyse RNAseq

Gaëlle Lelandais

gaelle.lelandais@universite-paris-saclay.fr

Les étapes incontournables de l'analyse bioinformatique

Etape 1:

Le contrôle de la qualité des séquences

Etape 2:

L'alignement des séquences sur un (ou des) génome(s) de référence*

* s'ils sont connus

Etape 3:

La visualisation des alignements

Etape 4:

Analyses spécifiques (recherche de gènes DE, de variants, etc.)

Le choix des logiciels (et des valeurs de paramètres) dépend de la technologie de séquençage, de l'organisme étudié et de la question scientifique posée …

Les étapes incontournables de l'analyse bioinformatique

Etape 1:

Le contrôle de la qualité des séquences

Etape 2:

L'alignement des séquences sur un (ou des) génome(s) de référence*

* s'ils sont connus

Etape 3:

La visualisation des alignements

Etape 4:

Analyses spécifiques (recherche de gènes DE, de variants, etc.)

Le choix des logiciels (et des valeurs de paramètres) dépend de la technologie de séquençage, de l'organisme étudié et de la question scientifique posée …

Fichier FASTQ et qualité des *reads*

"FASTQ format is a text-based format for storing both **a biological sequence** (usually nucleotide sequence) and its corresponding **quality scores**. Both the sequence letter and quality score are each encoded with a **single ASCII character** for brevity" (Wikipedia : https://en.wikipedia.org/wiki/FASTQ_format)

Une sequence est nommée "Read" :

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

← Séquence nucléotidique

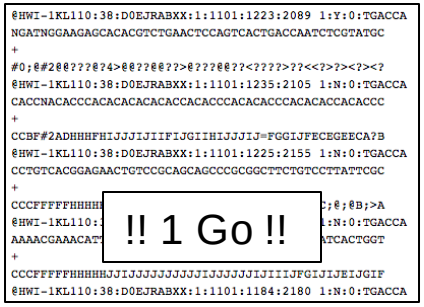
← Score qualité

Caractère ASCII

Phred quality score:

$$Q_{\text{sanger}} = -10 \log_{10} p$$

Plusieurs millions



→ Score qualité entre 0 et 40

Signification : Probabilité d'une erreur de séquençage

Etape 1:

Le contrôle de la qualité des séquences

Exemple de logiciel : FASTQC

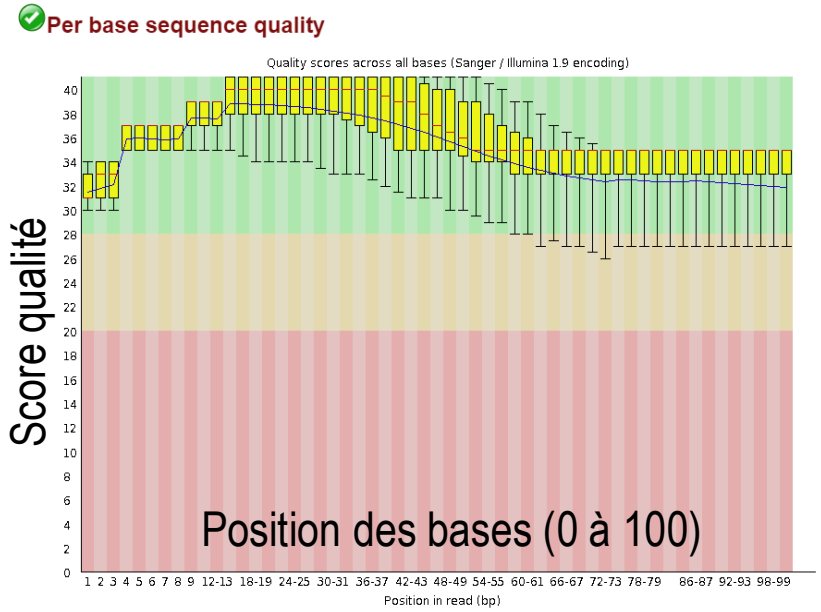
http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

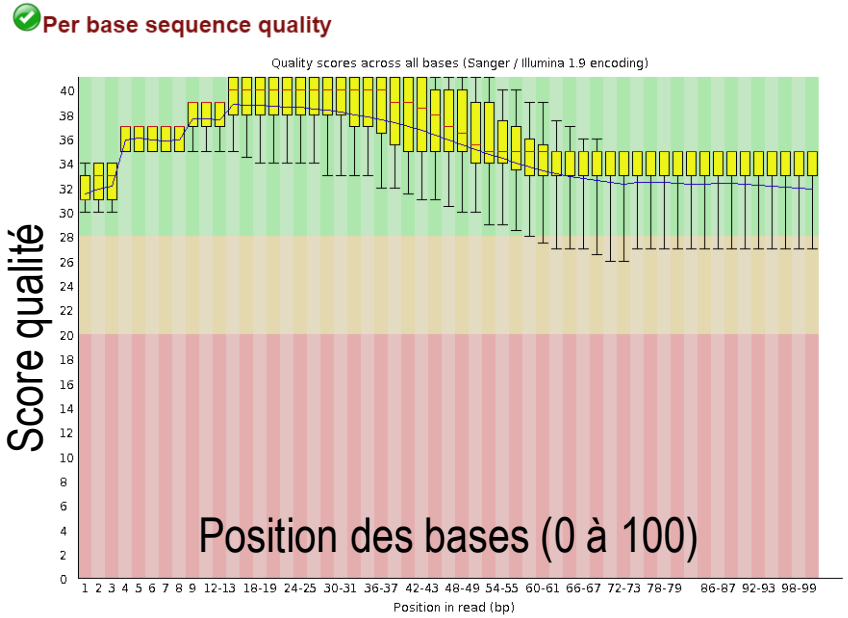
A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.3	STANDARD	LIGHT	3H	SHORT TERM	S2

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.9	DEPLETED	LIGHT	3H	SHORT TERM	S1

5 777 032 séquences (100 bases)

7 007 145 séquences (100 bases)

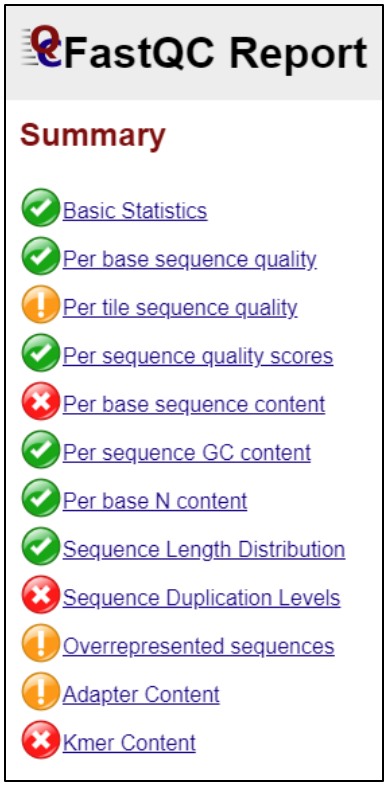


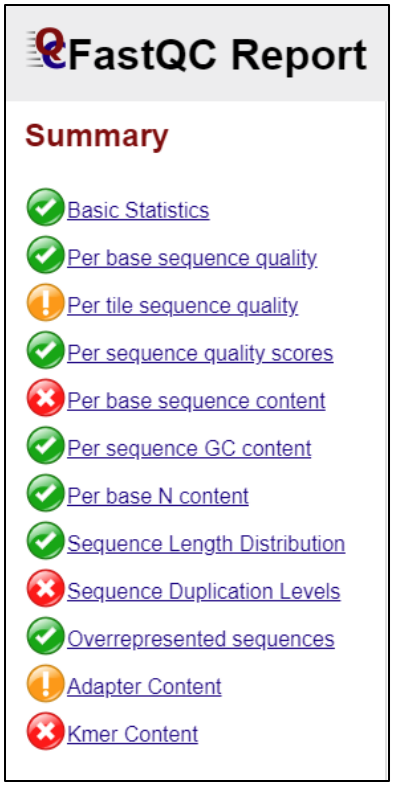


Autres éléments à contrôler

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.3	STANDARD	LIGHT	3H	SHORT TERM	S2

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.9	DEPLETED	LIGHT	3H	SHORT TERM	S1



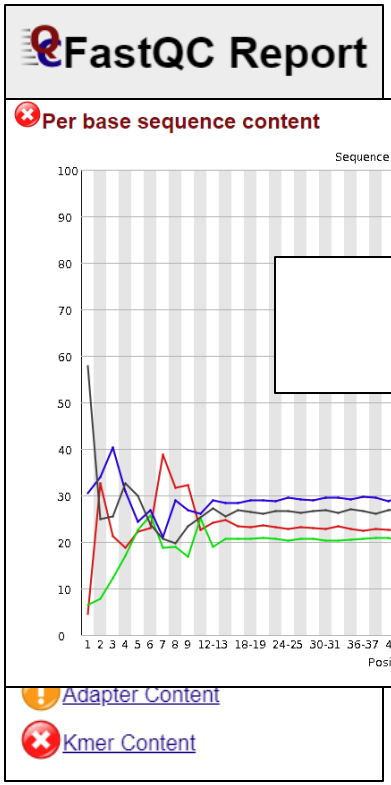


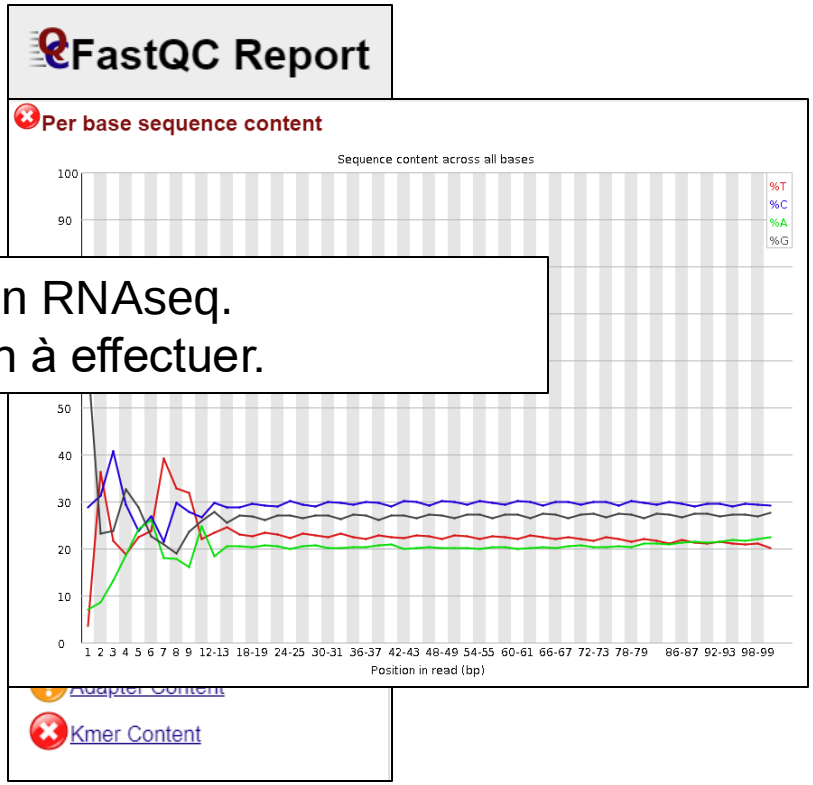
Autres éléments à contrôler

(dépendent du type de données étudiées)

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.3	STANDARD	LIGHT	3H	SHORT TERM	S2

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.9	DEPLETED	LIGHT	3H	SHORT TERM	S1





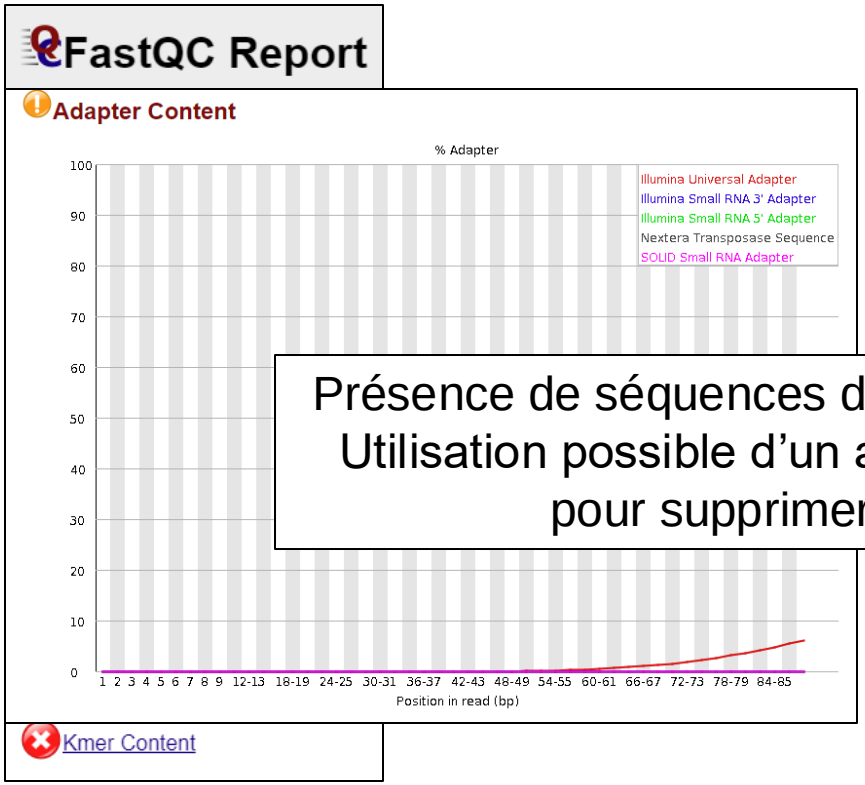
Biais classique en RNAseq.
Il n'y a pas d'action à effectuer.

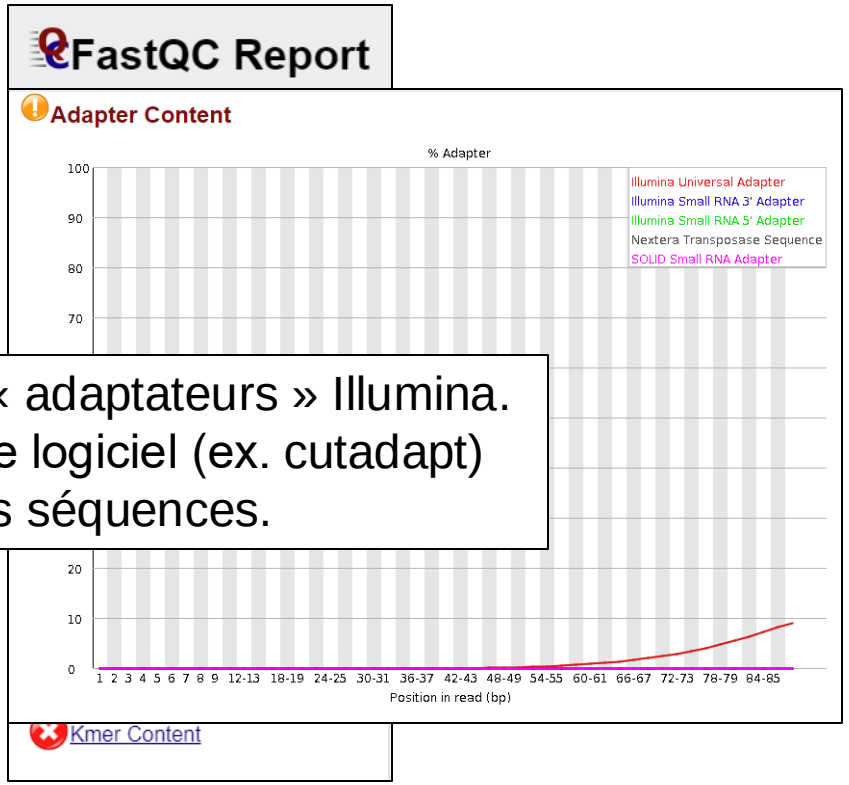
Autres éléments à contrôler

(dépendent du type de données étudiées)

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.3	STANDARD	LIGHT	3H	SHORT TERM	S2

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.9	DEPLETED	LIGHT	3H	SHORT TERM	S1



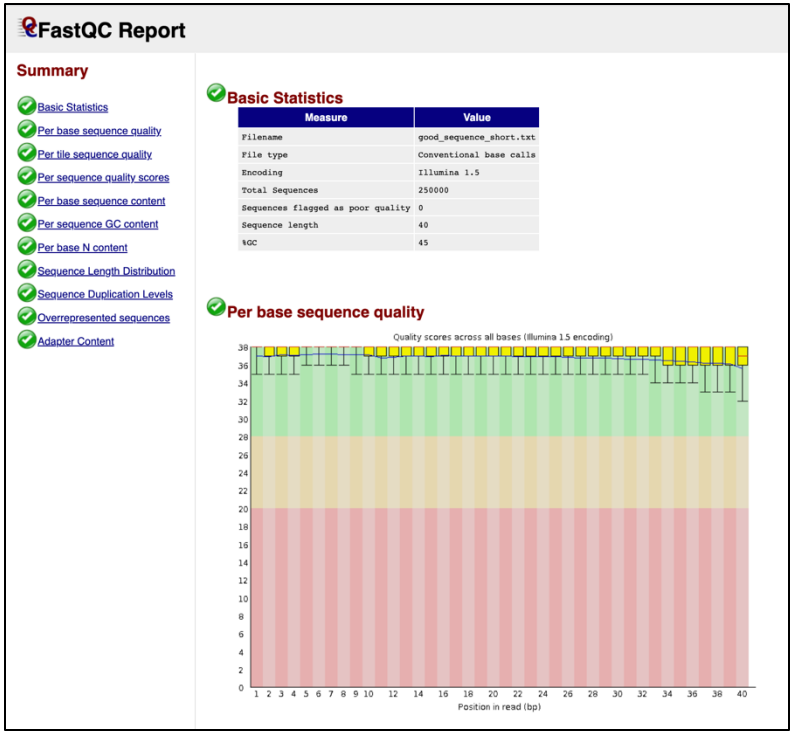


Présence de séquences des « adaptateurs » Illumina.
Utilisation possible d'un autre logiciel (ex. cutadapt) pour supprimer ces séquences.

Pour aller plus loin

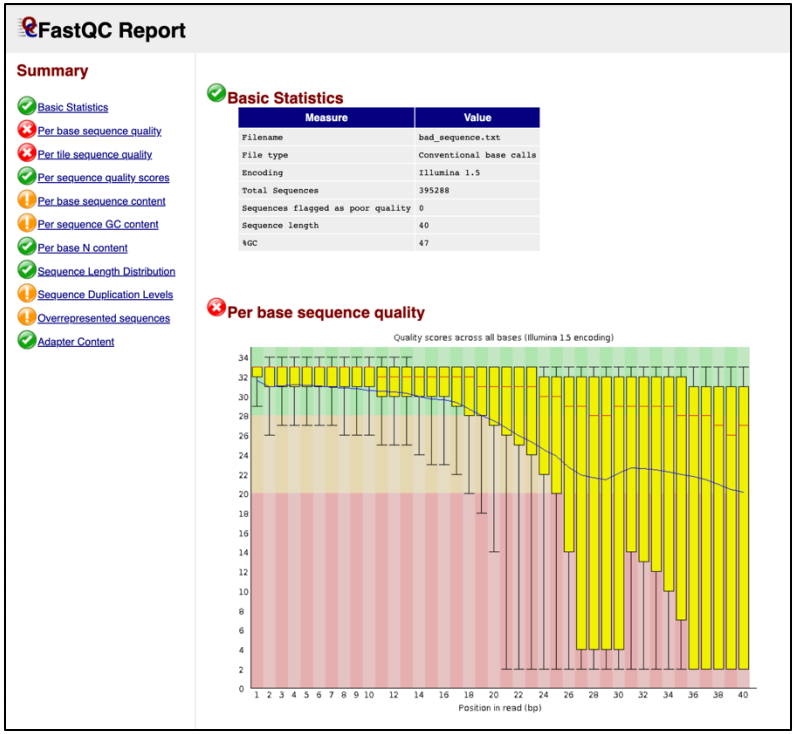
- Tutoriel vidéo (11 minutes) : http://www.youtube.com/watch?v=bz93ReOv87Y

- Exemple de séquences de bonne qualité :



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

- Exemple de séquences de mauvaise qualité :



https://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html

Les étapes incontournables de l'analyse bioinformatique

Etape 1:

Le contrôle de la qualité des séquences

Etape 2:

L'alignement des séquences sur un (ou des) génome(s) de référence*

* s'ils sont connus

Etape 3:

La visualisation des alignements

Etape 4:

Analyses spécifiques (recherche de gènes DE, de variants, etc.)

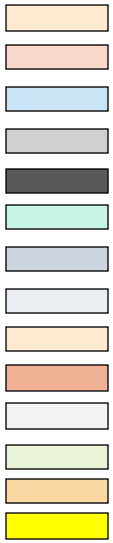
Le choix des logiciels (et des valeurs de paramètres) dépend de la technologie de séquençage, de l'organisme étudié et de la question scientifique posée …

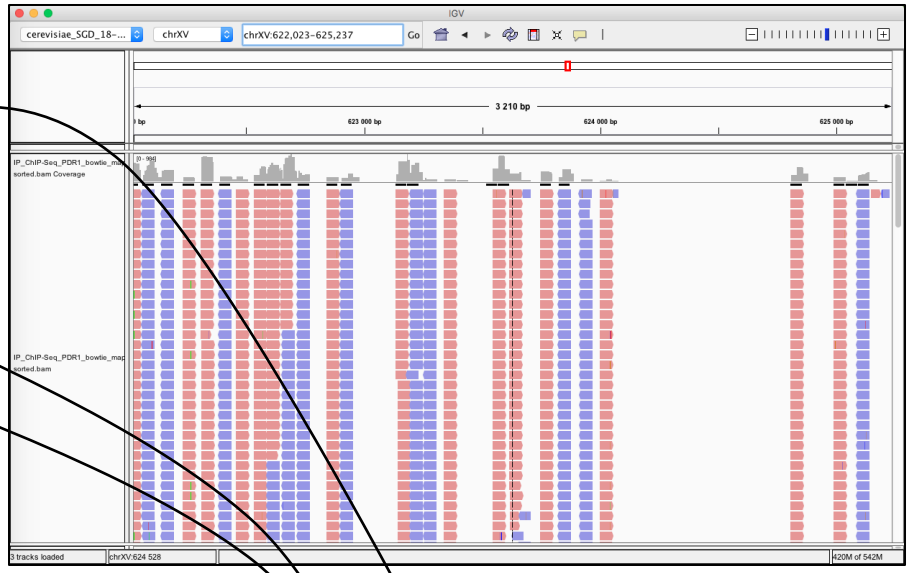
Problématique

D’où viennent ces sequences ?

Courtes séquences, écrites dans un fichier FASTQ

```
@HWI-1KL110:38:D0EJRABXX:1:1101:1223:2089 1:Y:0:TGACCA
NGATNGGAAGAGCACACGTCTGAACTCCAGTCACTGACCAATCTCGTATGC
+
#0;@#2@@???@?4>@@??@@??>@???@@??<????>??<<?>?><?><?
@HWI-1KL110:38:D0EJRABXX:1:1101:1235:2105 1:N:0:TGACCA
CACCNACACCCACACACACACACCACACCCACACACCCACACACCACACCC
+
CCBF#2ADHHHFHIJJJIJIIFIJGIIHIJJJIJ=FGGIJFECEGEECA?B
@HWI-1KL110:38:D0EJRABXX:1:1101:1225:2155 1:N:0:TGACCA
CCTGTCACGGAGAACTGTCCGCAGCAGCCCGCGGCTTCTGTCCTTATTCGC
+
CCCFFFFFHHHHHJJJJIJJJJJJJIGIIJIIJJ>AHECEAAC;@;@B;>A
@HWI-1KL110:38:D0EJRABXX:1:1101:1144:2172 1:N:0:TGACCA
AAAACGAAACATTTGACTTCGCCATCTCAATTGCTGTAGTGCATCACTGGT
+
CCCFFFFFHHHHHJJIJJJJJJJJJJJIJJJJJIIIIJFGIJIJEIJGIF
@HWI-1KL110:38:D0EJRABXX:1:1101:1184:2180 1:N:0:TGACCA
```





Superposition

Pas de résultat !

Solution unique

Régions dupliquées du génomes ?

Génome de référence

Un problème difficile !

Les sequences génomiques sont longues …

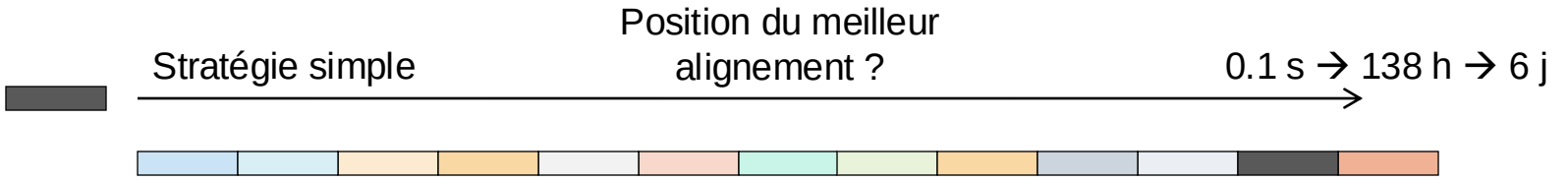
- O. tauri : 12.10^6 bp
- Yeasts : 10 to 20.10^6 bp
- Drosophila : 120.10^6 bp
- Human : 3.10^9 bp

… par rapport à la longueur des *reads*

50 to 250 bases
(en fonction de la technologie de séquençage)

1 / 100 000
à
1 / 1 000 000 000

X 5 à 10 millions



➢ Des algorithmes ont été développés (Bowtie, STAR, etc.) pour trouver une “bonne” solution (qui n’est pas nécessairement “la” meilleure).

Etape 2:

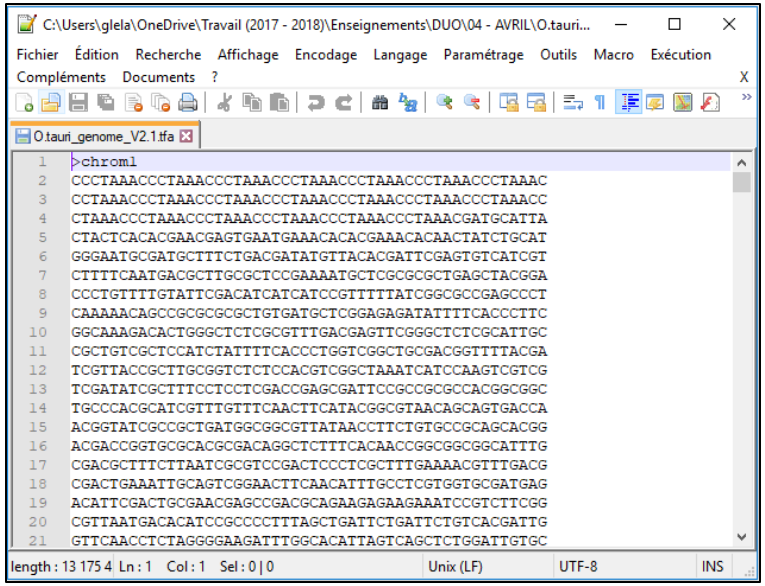
L'alignement des séquences sur un génome de référence

Exemples de logiciels :
BOWTIE, BOWTIE2, STAR, etc.

Logiciel SAMTOOLS (éventuellement)
http://samtools.sourceforge.net/

Fichier FASTQ (après filtrage des séquences si nécessaire)

Fichier FASTA (génome de référence)



BOWTIE, STAR, etc.

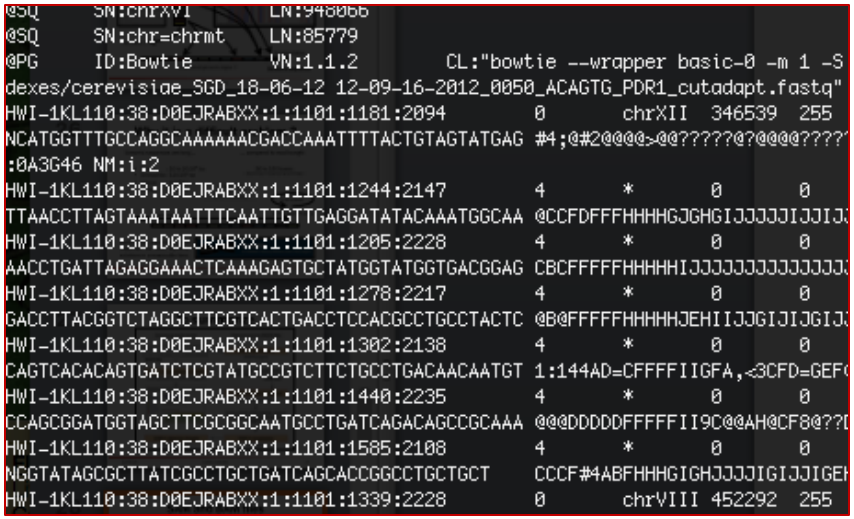
Fichier SAM (résultats des alignements des séquences sur le génome de référence)

SAMTOOLS

Fichier BAM (version binaire et compressée du fichier SAM)

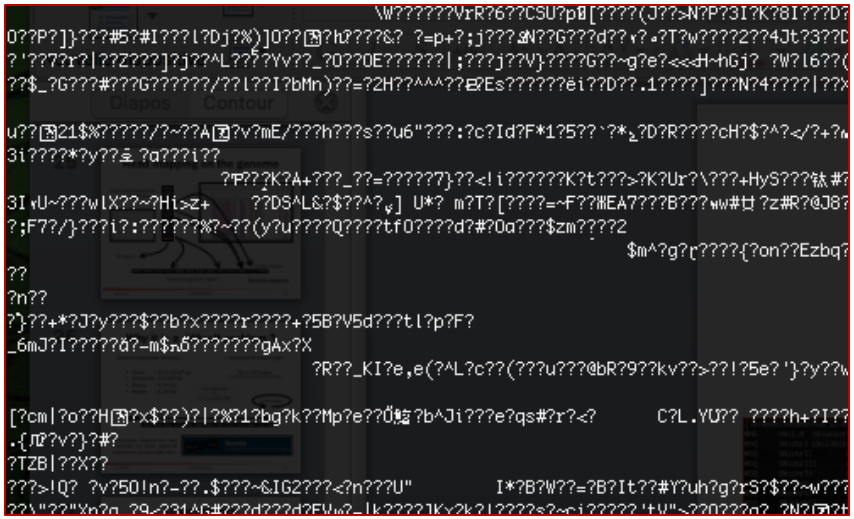
Fichiers SAM et BAM, quelle différence ?

Fichier SAM (5.4 Go)



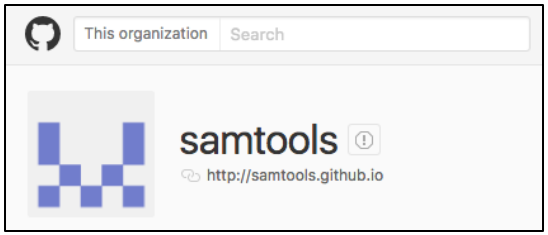
→ Fichier texte (nous pouvons le lire)

Fichier BAM (1.3 Go)



→ Version binaire du SAM (pour les ordinateurs)

La même information est contenue dans les deux fichiers.



Pour aller plus loin

Hatem *et al. BMC Bioinformatics* 2013, **14**:184
http://www.biomedcentral.com/1471-2105/14/184

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Benchmarking short sequence mapping tools

Ayat Hatem[1,2], Doruk Bozdağ[2], Amanda E Toland[3] and Ümit V Çatalyürek[1,2*]

Abstract

Background: The development of next-generation sequencing instruments has led to the generation of millions of short sequences in a single run. The process of aligning these reads to a reference genome is time consuming and demands the development of fast and accurate alignment tools. However, the current proposed tools make different compromises between the accuracy and the speed of mapping. Moreover, many important aspects are overlooked while comparing the performance of a newly developed tool to the state of the art. Therefore, there is a need for an objective evaluation method that covers all the aspects. In this work, we introduce a benchmarking suite to extensively analyze sequencing tools with respect to various aspects and provide an objective comparison.

Results: We applied our benchmarking tests on 9 well known mapping tools, namely, Bowtie, Bowtie2, BWA, SOAP2, MAQ, RMAP, GSNAP, Novoalign, and mrsFAST (mrFAST) using synthetic data and real RNA-Seq data. MAQ and RMAP are based on building hash tables for the reads, whereas the remaining tools are based on indexing the reference genome. The benchmarking tests reveal the strengths and weaknesses of each tool. The results show that no single tool outperforms all others in all metrics. However, Bowtie maintained the best throughput for most of the tests while BWA performed better for longer read lengths. The benchmarking tests are not restricted to the mentioned tools and can be further applied to others.

Conclusion: The mapping process is still a hard problem that is affected by many factors. In this work, we provided a benchmarking suite that reveals and evaluates the different factors affecting the mapping process. Still, there is no tool that outperforms all of the others in all the tests. Therefore, the end user should clearly specify his needs in order to choose the tool that provides the best results.

Keywords: Short sequence mapping, Next-generation sequencing, Benchmark, Sequence analysis

Introduction

Next-generation sequencing (NGS) technology has evolved rapidly in the last five years, leading to the generation of hundreds of millions of sequences (reads) in a single run. The number of generated reads varies between 1 million for long reads generated by Roche/454 sequencer (≈400 base pairs (bps)) and 2.4 billion for short reads generated by Illumina/Solexa and ABI/SOLIDTM sequencers (≈75 bps). The invention of the high-throughput sequencers has led to a significant cost reduction, e.g., a Megabase of DNA sequence costs only $0.1 [1].

Nevertheless, the large amount of generated data tells us almost nothing about the DNA, as stated by Flicek and Birney [2]. This is due to the lack of proper analysis tools and algorithms. Therefore, bioinformatics researchers started to think about new ways to efficiently handle and analyze this large amount of data.

One of the areas that attracted many researchers to work on is the alignment (mapping) of the generated sequences, i.e., the alignment of reads generated by NGS machines to a reference genome. Because, an efficient alignment of this large amount of reads with high accuracy is a crucial part in many applications' workflow, such as genome resequencing [2], DNA methylation [3], RNA-Seq [4], ChIP sequencing, SNPs detection [5], genomic structural variants detection [6], and metagenomics [7]. Therefore, numerous tools have been developed to undertake this challenging task including MAQ [8], RMAP [9], GSNAP [10], Bowtie [11], Bowtie2 [12], BWA [13], SOAP2 [14], Mosaik [15], FANGS [16], SHRIMP [17], BFAST [18],

*Correspondence: umit@bmi.osu.edu
[1]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA
[2]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA
Full list of author information is available at the end of the article

https://pubmed.ncbi.nlm.nih.gov/23758764/

- Les outils disponibles sont nombreux !
- Des études comparent les performances des outils. Stabilité des résultats en fonctions des paramètres choisis ? Temps de calculs ? Etc.

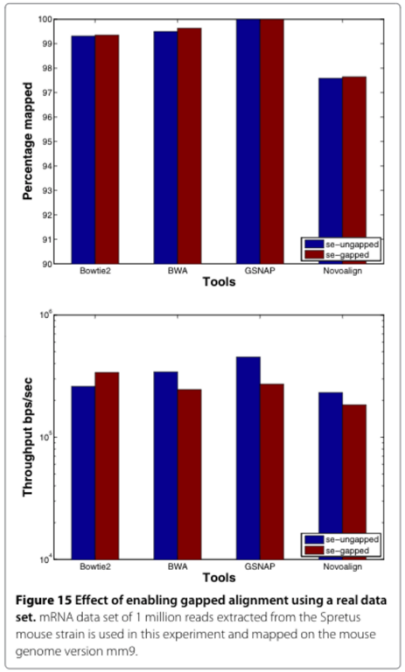


Figure 15 Effect of enabling gapped alignment using a real data set. mRNA data set of 1 million reads extracted from the Spretus mouse strain is used in this experiment and mapped on the mouse genome version mm9.

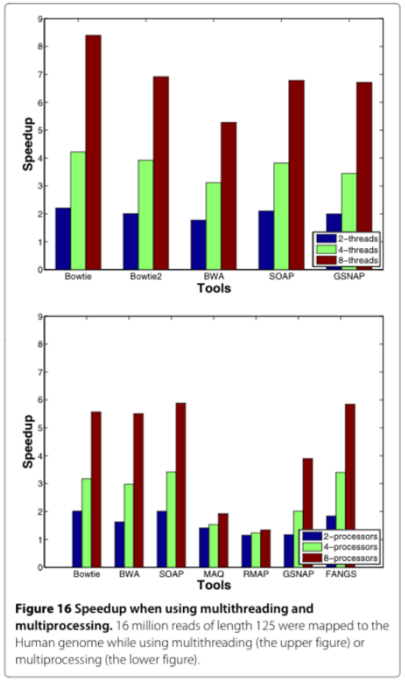


Figure 16 Speedup when using multithreading and multiprocessing. 16 million reads of length 125 were mapped to the Human genome while using multithreading (the upper figure) or multiprocessing (the lower figure).

Les étapes incontournables de l'analyse bioinformatique

Etape 1:

Le contrôle de la qualité des séquences

Etape 2:

L'alignement des séquences sur un (ou des) génome(s) de référence*

* s'ils sont connus

Etape 3:

La visualisation des alignements

Etape 4:

Analyses spécifiques (recherche de gènes DE, de variants, etc.)

Le choix des logiciels (et des valeurs de paramètres) dépend de la technologie de séquençage, de l'organisme étudié et de la question scientifique posée …

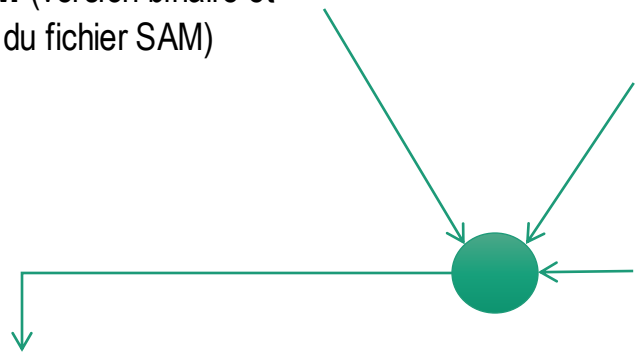
Etape 3:

La visualisation des alignements

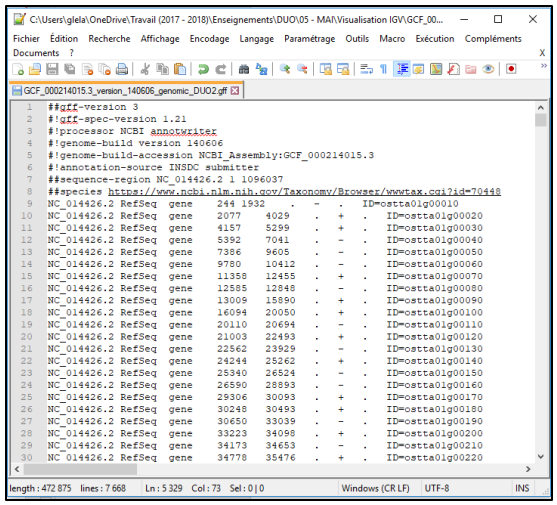
Exemple de logiciel : IGV
https://www.broadinstitute.org/igv/home

Fichiers BAM (version binaire et compressée du fichier SAM)

Fichier FASTA (génome de référence téléchargé depuis NCBI)



Fichier GFF/GTF (positions des éléments génomiques d'intérêts)



La visualisation est une étape très importante dans un projet « omique »

Le format GFF

Fichier composé de 9 colonnes, indiquant les positions génomiques de début et de fin d'éléments génomiques d'intérêts (ici les gènes).

General GFF structure

Position index	Position name	Description
1	sequence	The name of the sequence where the feature is located.
2	source	Keyword identifying the source of the feature, like a program (e.g. Augustus or RepeatMasker) or an organization (like TAIR).
3	feature	The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow their parents in a single block (so all exons of a transcript are put after their parent "transcript" feature line an features and their relationships should Sequence Ontology Project.
4	start	Genomic start of the feature, with a 1- open sequence formats, like BED files
5	end	Genomic end of the feature, with a 1-b offset half-open sequence formats, like
6	score	Numeric value that generally indicates value of "." (a dot) is used to define a r
7	strand	Single character that indicates the Ser assume the values of "+" (positive, or
8	frame (GTF, GFF2) *or* phase (GFF3)	Frame or phase of CDS features; it ca everything else). Frame and Phase ar
9	Attributes.	All the other information pertaining to t is the one which varies the most betwe

GCF_000214015.3_version_140606_genomic.gff

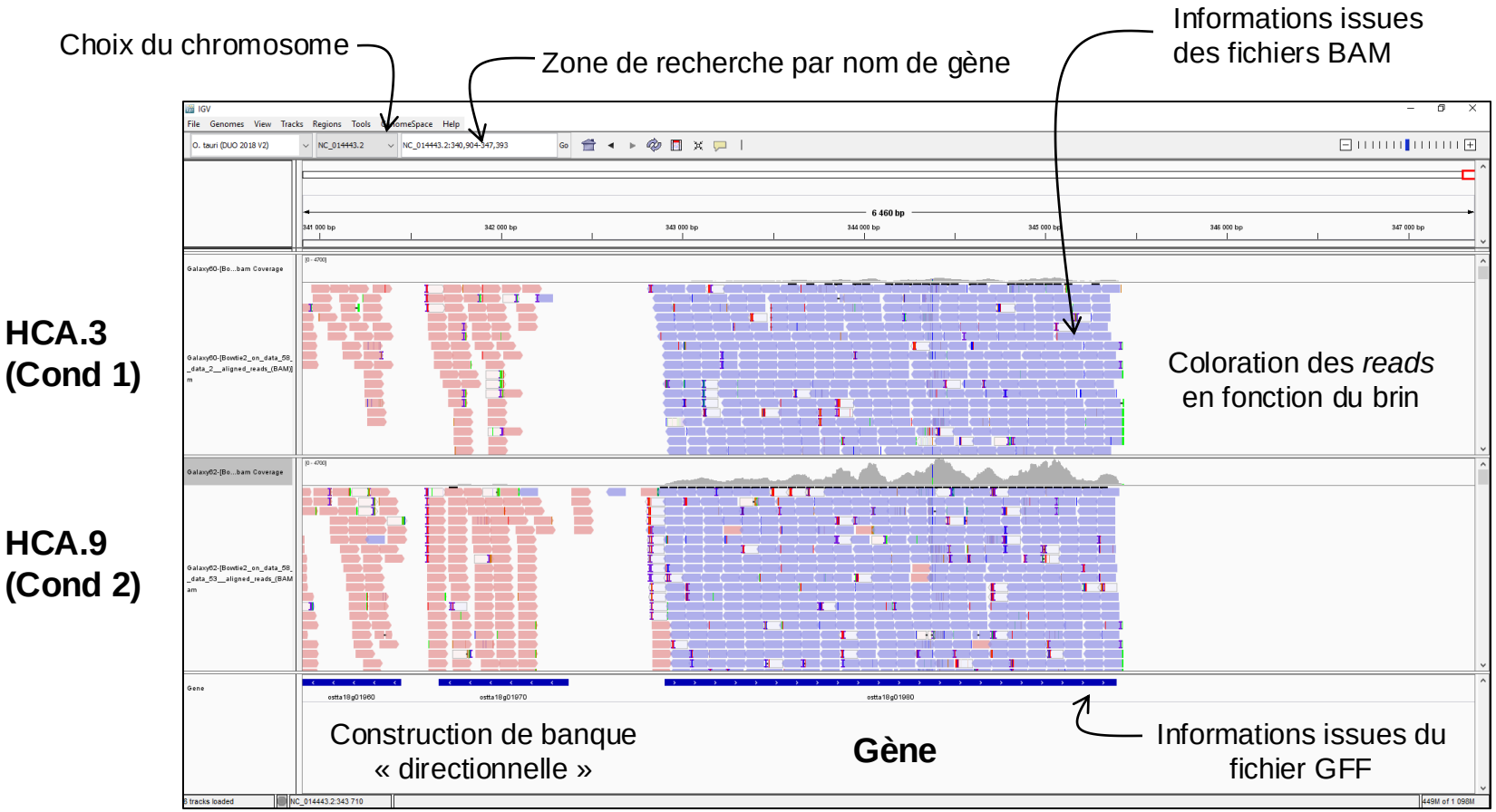
```
1  ##gff-version 3
2  #!gff-spec-version 1.21
3  #!processor NCBI annotwriter
4  #!genome-build version 140606
5  #!genome-build-accession NCBI_Assembly:GCF_000214015.3
6  #!annotation-source INSDC submitter
7  ##sequence-region NC_014426.2 1 1096037
8  ##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=70448
9  NC_014426.2 RefSeq  region  1     1096037 .  +  .  ID=id0;Dbxref=taxon:70448;Name=1;chromosome=1;gbke
10 NC_014426.2 RefSeq  gene    244   1932    .  -  .  ID=gene0;Dbxref=GeneID:9834635;Name=OT_ostta01g000
11 NC_014426.2 RefSeq  mRNA    244   1932    .  -  .  ID=rna0;Parent=gene0;Dbxref=GeneID:9834635,Genbank
12 NC_014426.2 RefSeq  exon    244   1932    .  -  .  ID=id1;Parent=rna0;Dbxref=GeneID:9834635,Genbank:X
13 NC_014426.2 RefSeq  CDS 244 1932    .  -  0  ID=cds0;Parent=rna0;Dbxref=UniProtKB/TrEMBL:Q01H82,Gen
14 NC_014426.2 RefSeq  gene    2077  4029    .  +  .  ID=gene1;Dbxref=GeneID:9832483;Name=OT_ostta01
15 NC_014426.2 RefSeq  mRNA    2077  4029    .  +  .  ID=rna1;Parent=gene1;Dbxref=GeneID:9832483,Genb
16 NC_014426.2 RefSeq  exon    2077  4029    .  +  .  ID=id2;Parent=rna1;Dbxref=GeneID:9832483,Genba
17 NC_014426.2 RefSeq  CDS 2077    4029    .  +  0  ID=cds1;Parent=rna1;Dbxref=GOA:Q01H81,InterPro:IPR
18 NC_014426.2 RefSeq  gene    4157  5299    .  +  .  ID=gene2;Dbxref=GeneID:9832502;Name=OT_ostta01
19 NC_014426.2 RefSeq  mRNA    4157  5299    .  +  .  ID=rna2;Parent=gene2;Dbxref=GeneID:9832502,Genb
20 NC_014426.2 RefSeq  exon    4157  5299    .  +  .  ID=id3;Parent=rna2;Dbxref=GeneID:9832502,Genba
21 NC_014426.2 RefSeq  CDS 4157    5299    .  +  0  ID=cds2;Parent=rna2;Dbxref=UniProtKB/TrEMBL:Q01H80,
22 NC_014426.2 RefSeq  gene    5392  7041    .  -  .  ID=gene3;Dbxref=GeneID:9832501;Name=OT_ostta01
23 NC_014426.2 RefSeq  mRNA    5392  7041    .  -  .  ID=rna3;Parent=gene3;Dbxref=GeneID:9832501,Genb
24 NC_014426.2 RefSeq  exon    5392  7041    .  -  .  ID=id4;Parent=rna3;Dbxref=GeneID:9832501,Genba
25 NC_014426.2 RefSeq  CDS 5392    7041    .  -  0  ID=cds3;Parent=rna3;Dbxref=GOA:Q01H79,InterPro:IPR
```

https://www.ensembl.org/info/website/upload/gff.html ; https://en.wikipedia.org/wiki/General_feature_format

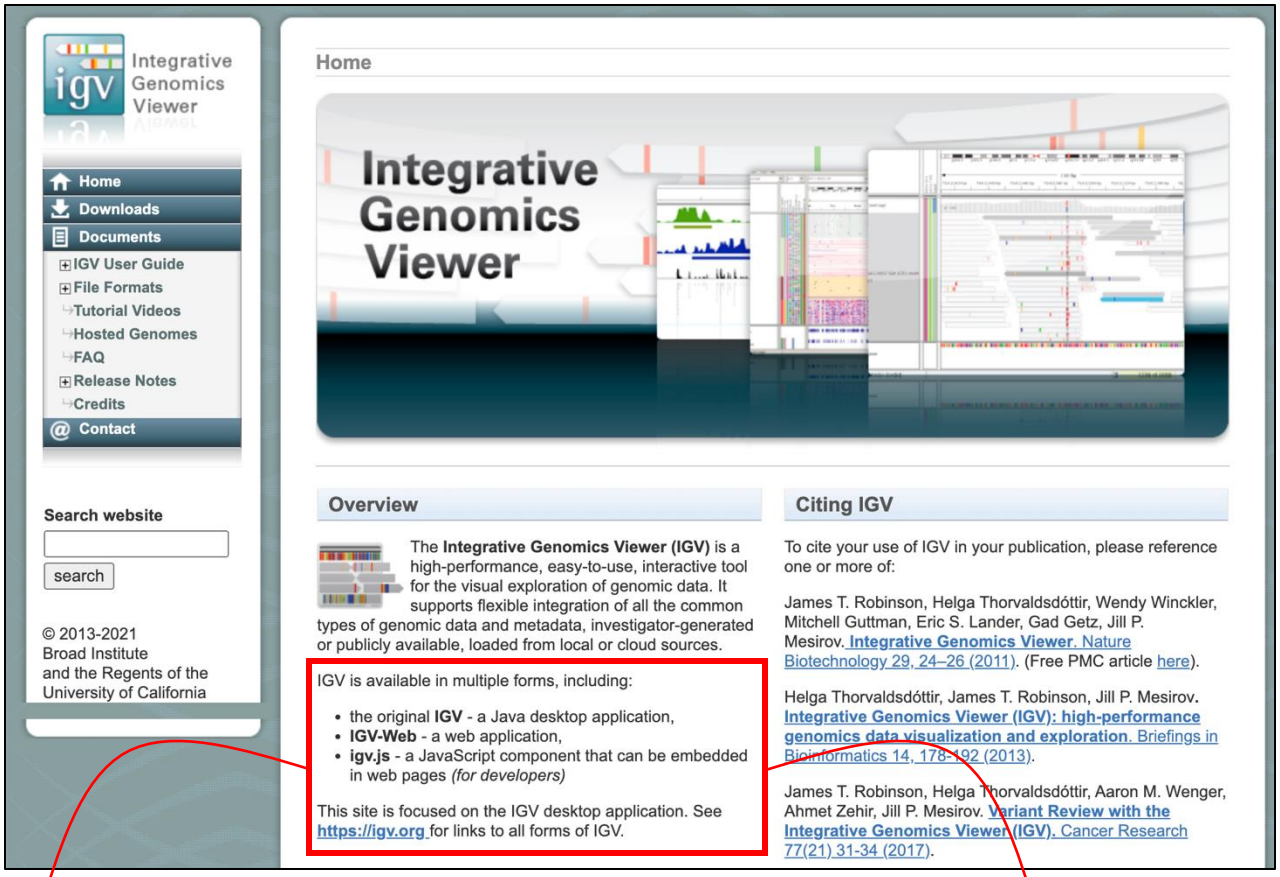
Exemple visualisation

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.3	STANDARD	LIGHT	3H	SHORT TERM	S2

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.9	DEPLETED	LIGHT	3H	SHORT TERM	S1



« IGV Desktop » ou « IGV Web » ?

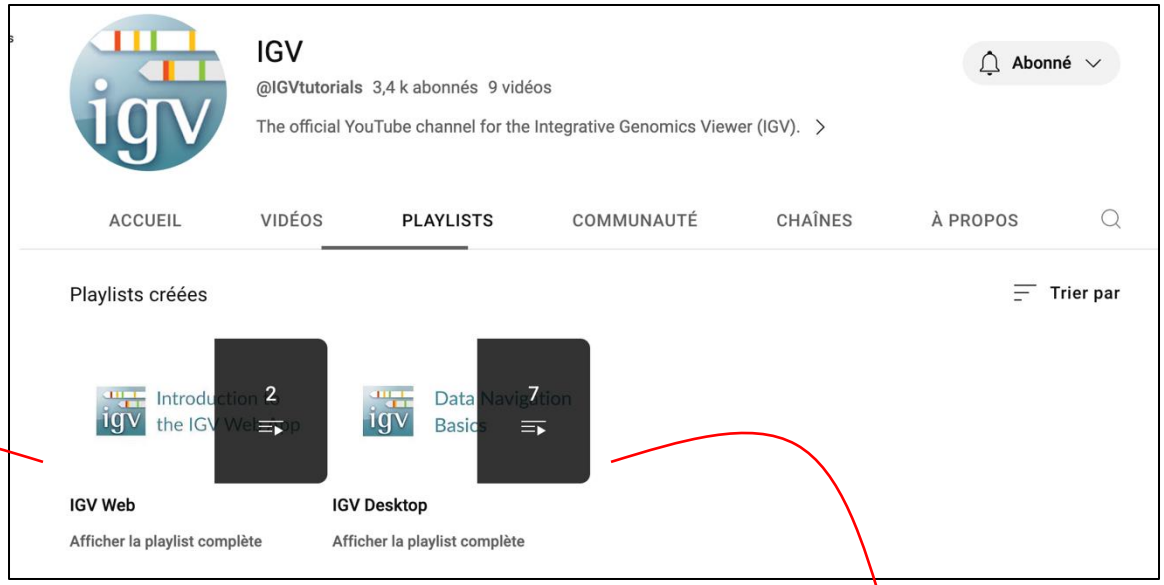


Travail sur son ordinateur « en local »

Travail sur un serveur « en ligne »

Ressources utiles

- Tutoriels vidéos :
  https://www.youtube.com/@IGVtutorials/playlists



Travail sur un serveur « online »

Travail sur son ordinateur « en local »

Les étapes incontournables de l'analyse bioinformatique

Etape 1:

Le contrôle de la qualité des séquences

Etape 2:

L'alignement des séquences sur un (ou des) génome(s) de référence*

* s'ils sont connus

Etape 3:

La visualisation des alignements

Etape 4:

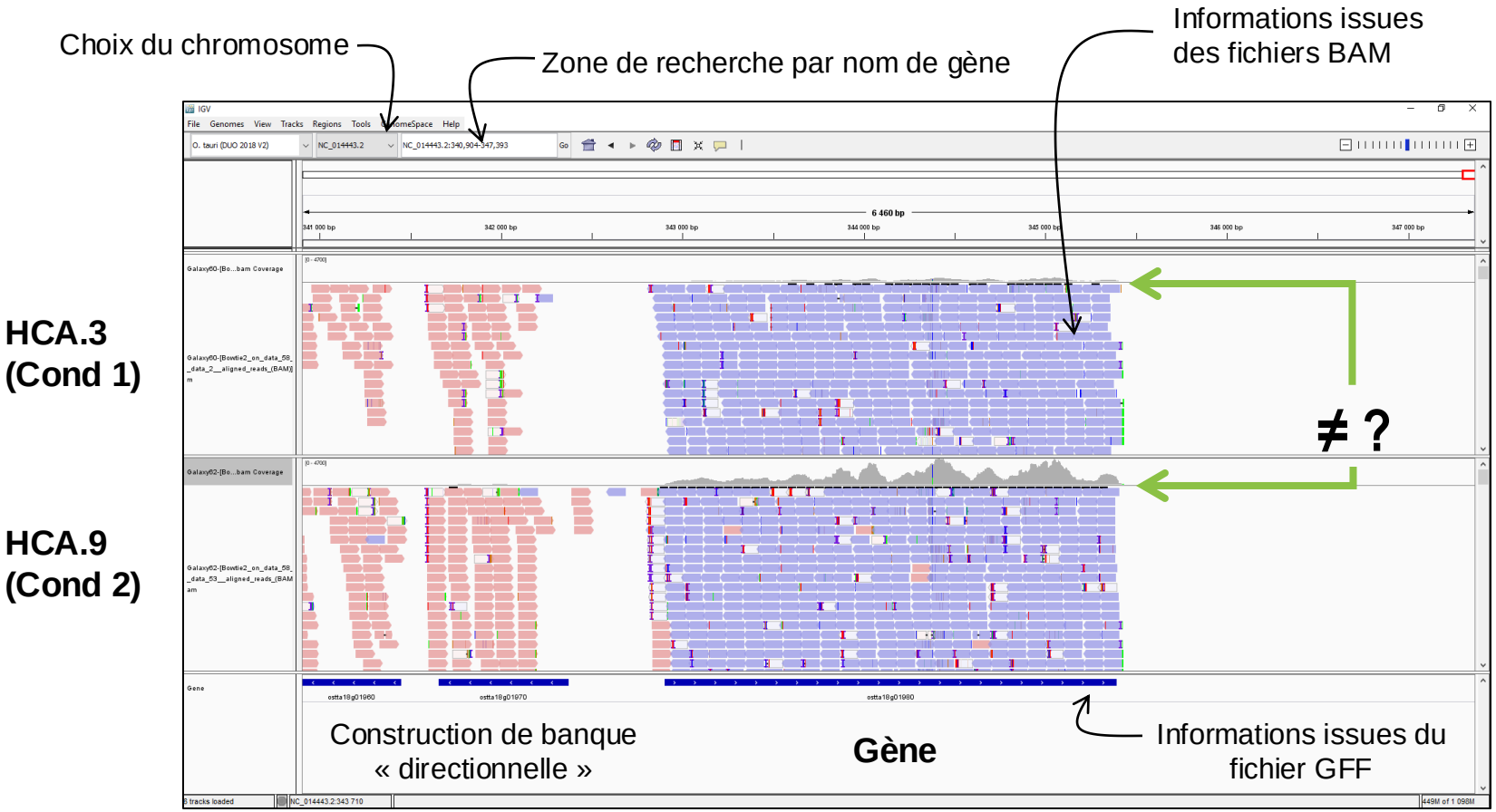
Analyses spécifiques (recherche de gènes DE, de variants, etc.)

Le choix des logiciels (et des valeurs de paramètres) dépend de la technologie de séquençage, de l'organisme étudié et de la question scientifique posée …

Différentiel d'expression ?

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.3	STANDARD	LIGHT	3H	SHORT TERM	S2

A	B	C	D	E	F
SampleID	Iron	Light	Time	Condition	Sample nan
HCA.9	DEPLETED	LIGHT	3H	SHORT TERM	S1

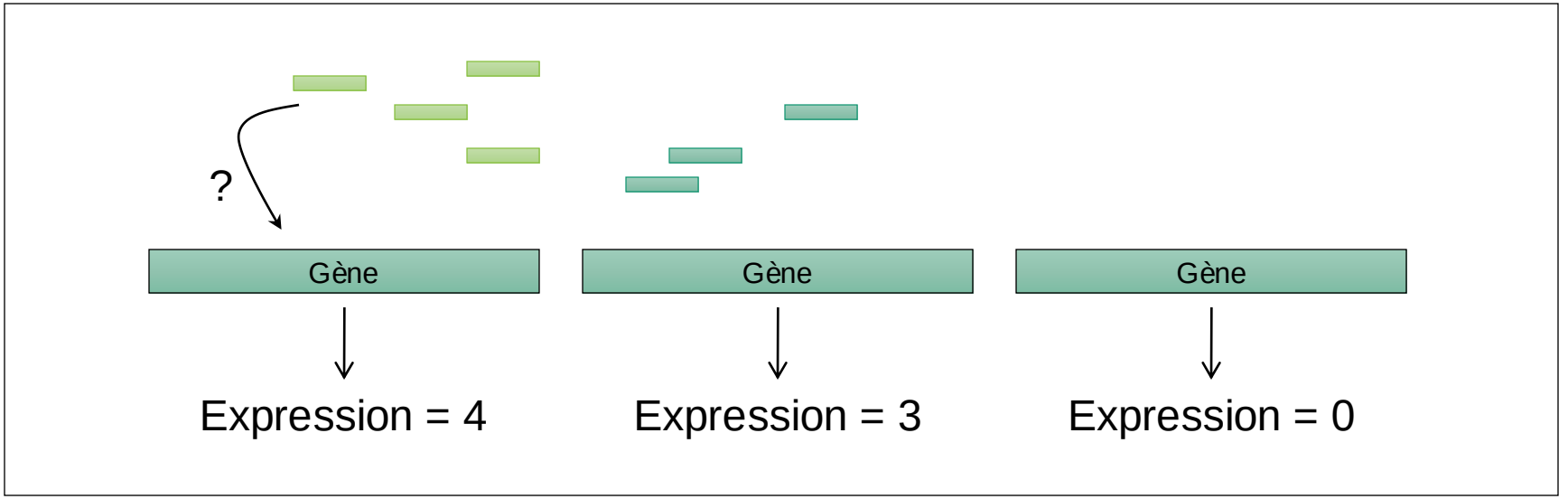


Etape 4 (pour le RNAseq) :

Identification des gènes différentiellement exprimés

Exemple de programme : HTSeq count
https://htseq.readthedocs.io/en/release_0.9.1/counting.html

- Les séquences positionnées au niveau des gènes sont comptées (une annotation du génome est donc nécessaire → Fichier GFF/GTF).



- L'activité transcriptionnelle des gènes est supposée proportionnelle au nombre de séquences alignées.

Etape 4 (pour le RNAseq) :

Identification des gènes différentiellement exprimés

UNIX

- Compilés dans un unique fichier, nommé « table de comptage » :

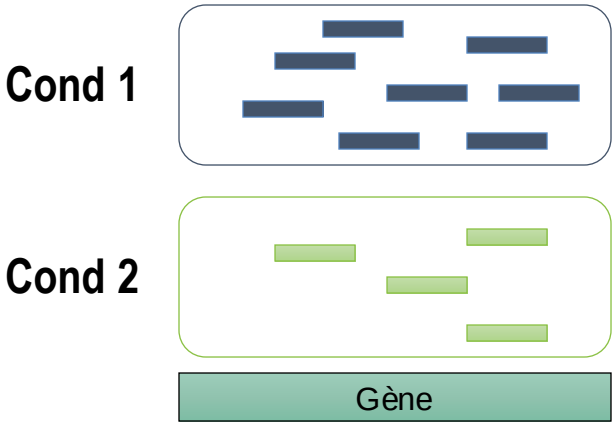
mapping_rawdata_allGenes.txt

gene	length	HCA-3	HCA-4	HCA-5	HCA-6	HCA-7	HCA-8	HCA-9	HCA-10	HCA-11	HCA-12	HCA-13	HCA-14	HCA-15
ostta01g00010	1689	226	176	224	246	236	177	296	194	102	196	225	205	265	285	394	159	207	136	464	225	225	288	126	317
ostta01g00020	1953	256	201	270	299	199	235	391	322	180	190	240	229	212	192	323	225	275	192	399	204	292	389	185	172
ostta01g00030	1143	45	52	32	65	36	53	77	45	18	42	55	62	34	32	46	50	72	41	57	44	29	93	74	17
ostta01g00040	1650	732	595	705	869	723	710	1012		834	442	559	553	571	870	770	1044		695	859	620	1733		891	985
ostta01g00050	2220	311	223	301	481	385	345	386	298	176	248	240	270	384	382	572	277	323	248	524	236	277	407	190	439
ostta01g00060	633 246	202	229	338	284	299	277	270	149	226	304	235	116	128	174	484	592	412	199	80	122	575	297	187	367
ostta01g00070	1098	384	249	359	99	101	100	444	349	196	75	112	103	200	162	222	145	214	122	364	171	187	249	146	125
ostta01g00080	264 1	1	10	13	19	21	14	5	1	25	13	21	17	32	40	10	8	4	36	27	34	22	14	12	28
ostta01g00090	2514	617	571	619	5823		4620		4652		1195		964	643	4491		4774		4505		3187		2677		450
ostta01g00100	3957	808	717	773	756	521	544	1009		963	726	437	301	335	544	469	619	679	831	568	1150		444	533	111
ostta01g00110	585 17	12	25	9	9	8	24	16	5	7	5	12	11	15	24	5	11	7	37	15	7	14	10	4	29
ostta01g00120	1491	375	280	427	156	140	144	504	407	208	107	115	131	203	207	285	206	260	146	330	129	186	340	175	217
ostta01g00130	1368	391	310	371	264	257	250	416	376	182	185	234	249	367	382	453	331	405	256	692	343	431	535	256	271
ostta01g00140	966 816	600	732	64	74	62	812	787	377	74	93	85	771	736	1214		114	123	71	1024		529	546	227	99
ostta01g00150	1185	14	12	21	134	114	109	9	14	8	73	137	97	10	13	12	369	485	345	26	5	11	362	182	0
ostta01g00160	2304	81	92	79	258	199	209	148	102	61	198	208	212	120	106	180	270	292	212	338	168	215	429	226	159
ostta01g00170	747 16	8	15	36	32	22	12	16	5	11	20	16	13	9	29	40	39	32	29	15	16	67	16	28	62
ostta01g00180	246 9	8	12	53	41	55	18	13	8	19	50	53	7	4	9	118	157	123	6	1	4	137	57	6	73
ostta01g00190	2241	589	452	605	2304		1772		1823		915	713	447	1317		1195		1383		275	278	327	4516		532
ostta01g00200	876 99	42	64	134	95	92	145	110	69	89	118	111	60	53	108	119	139	96	186	81	84	192	111	48	173
ostta01g00210	315 41	43	35	43	46	45	69	54	21	49	55	57	42	27	45	47	67	40	87	34	45	83	46	51	95
ostta01g00220	699 121	110	116	146	137	126	189	125	55	128	123	145	127	134	148	161	194	127	230	122	150	242	120	83	201
ostta01g00230	684 82	64	83	53	29	48	84	78	51	27	72	52	84	76	114	48	41	28	123	49	59	69	38	40	67
ostta01g00240	1497	56	37	40	115	91	89	67	51	49	66	95	88	50	59	69	64	84	76	103	57	54	150	84	39
ostta01g00250	876 29	18	35	42	32	42	63	41	15	16	54	43	22	19	23	66	34	34	39	20	27	80	27	19	28
ostta01g00260	2181	0	2	5	19	18	13	4	4	2	27	30	23	3	6	6	18	22	19	13	3	10	70	29	1
ostta01g00270	489 276	198	274	294	240	264	279	200	138	162	224	225	425	383	617	162	228	139	531	285	322	274	112	360	482
ostta01g00280	1350	257	174	270	321	260	272	310	253	161	250	335	292	133	128	175	346	495	341	304	172	164	651	304	89
ostta01g00290	177 5	0	6	6	3	6	3	2	0	5	7	11	9	7	16	8	17	9	11	6	2	15	14	7	10
ostta01g00300	585 35	32	31	36	21	36	69	36	22	28	34	33	52	41	67	42	50	29	83	45	45	51	21	46	68
ostta01g00310	1461	196	165	189	60	51	64	324	271	177	56	47	70	186	182	267	49	64	49	344	202	203	96	78	149
ostta01g00320	2433	311	268	312	252	187	208	430	351	214	171	154	150	208	184	298	299	400	271	423	176	195	444	223	160
ostta01g00330	1152	513	489	648	391	277	303	717	648	352	213	354	293	1160		1121		1941		118	147	73	1641		628
ostta01g00340	1026	15	11	9	59	67	81	15	9	11	59	65	58	32	22	50	44	54	52	74	52	56	109	62	8

Etape 4 (pour le RNAseq) :

Identification des gènes différentiellement exprimés

Exemple du package R : « DESeq2 »
https://bioconductor.org/packages/release/bioc/html/DESeq2.html



seq = 8

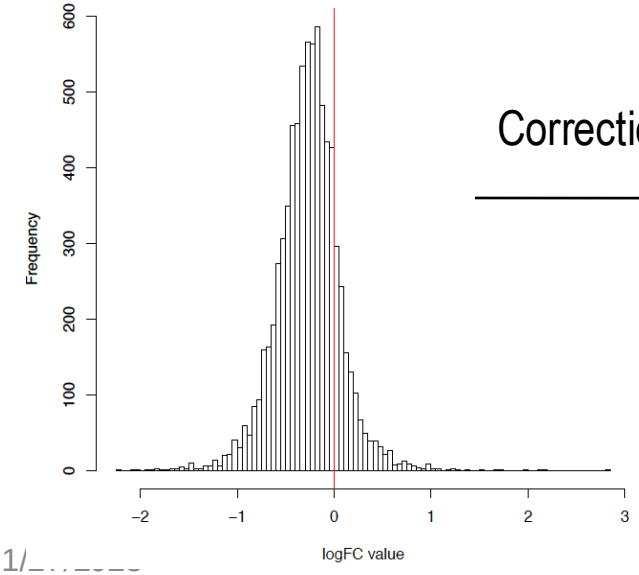
seq = 4

Calcul d'un LogFC

LogFC = Log2(8/4) = 1

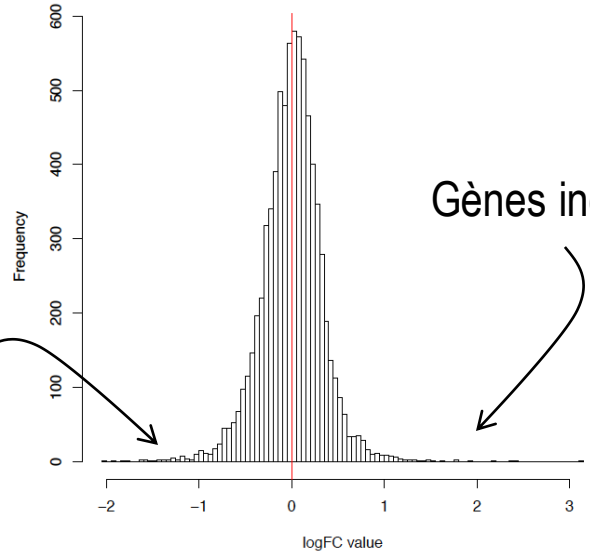
Une normalisation est nécessaire …

Données brutes



Correction statistique

Données normalisées



Gènes induits

Gènes réprimés

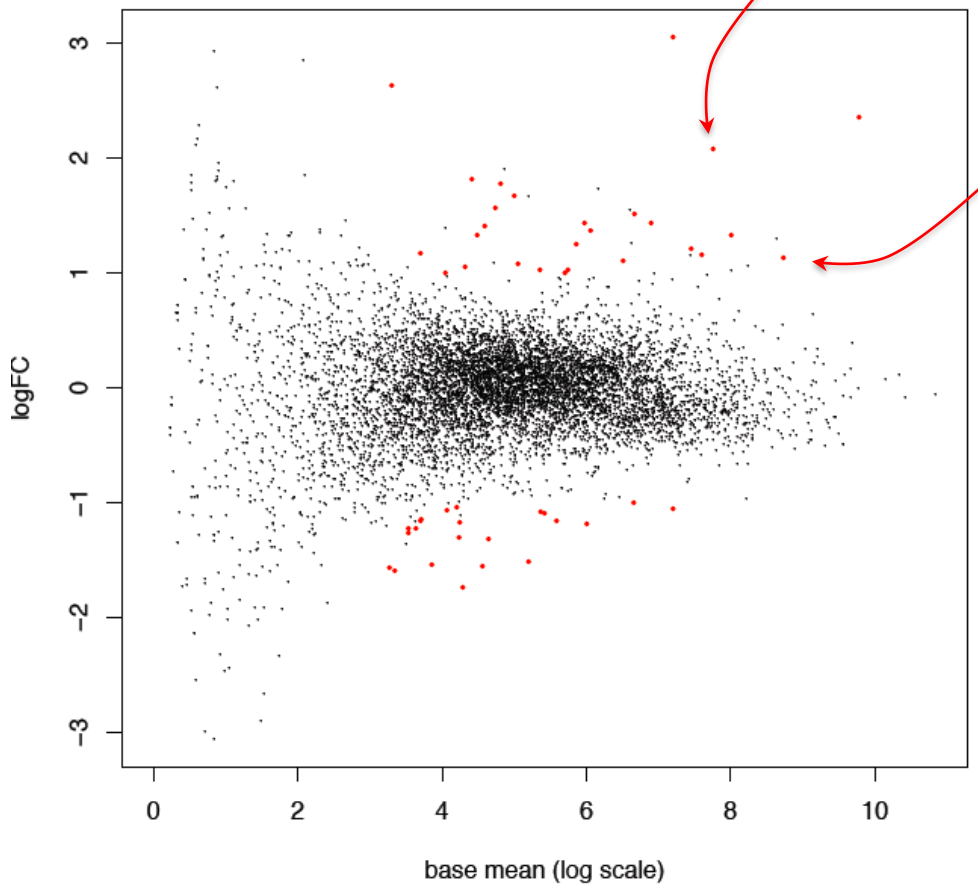
Etape 4 (pour le RNAseq) :

Identification des gènes différentiellement exprimés

Exemple du package R : « DESeq2 »
https://bioconductor.org/packages/release/bioc/html/DESeq2.html

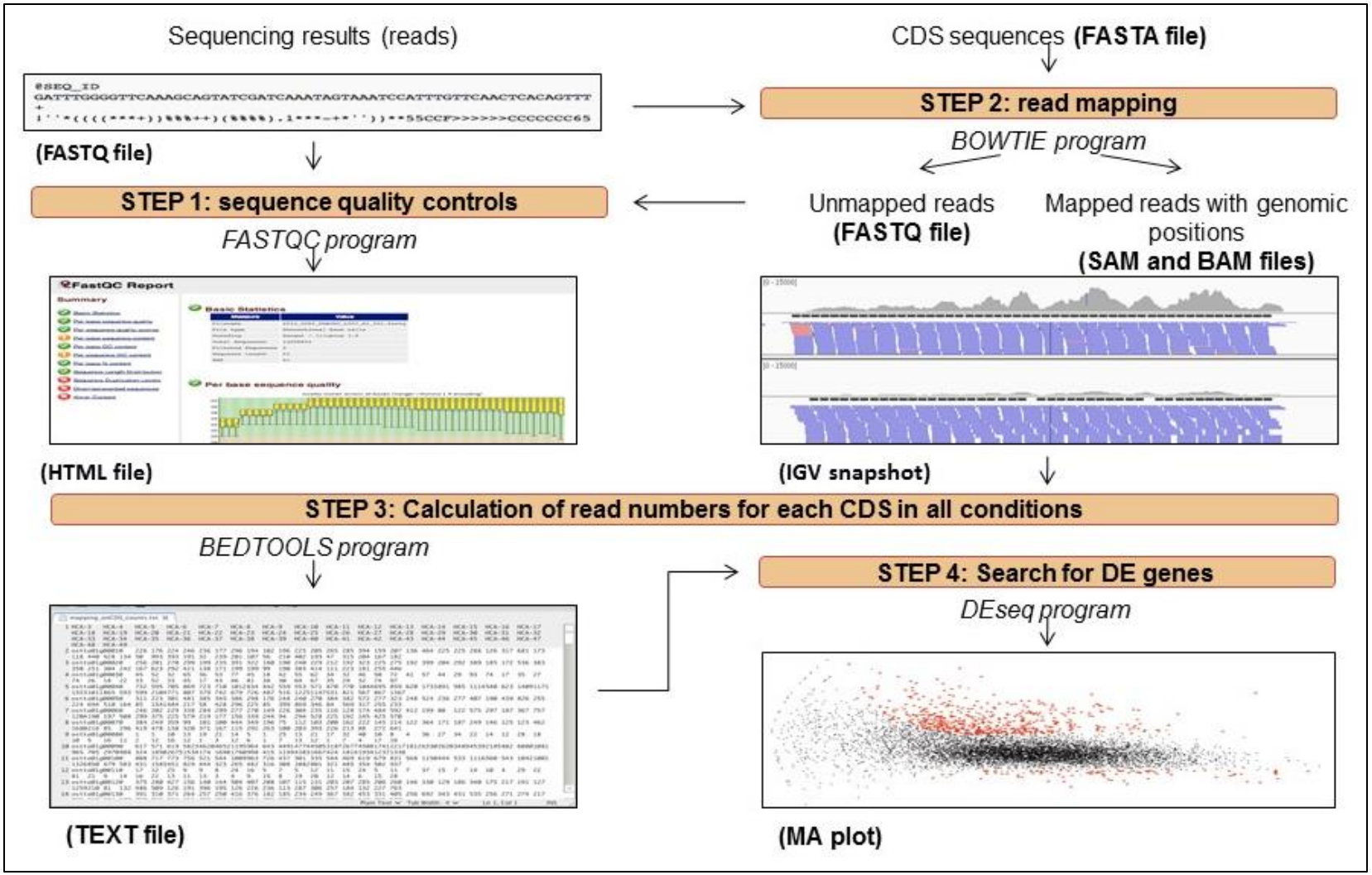
Induction

Repression



Faible expression → Forte expression

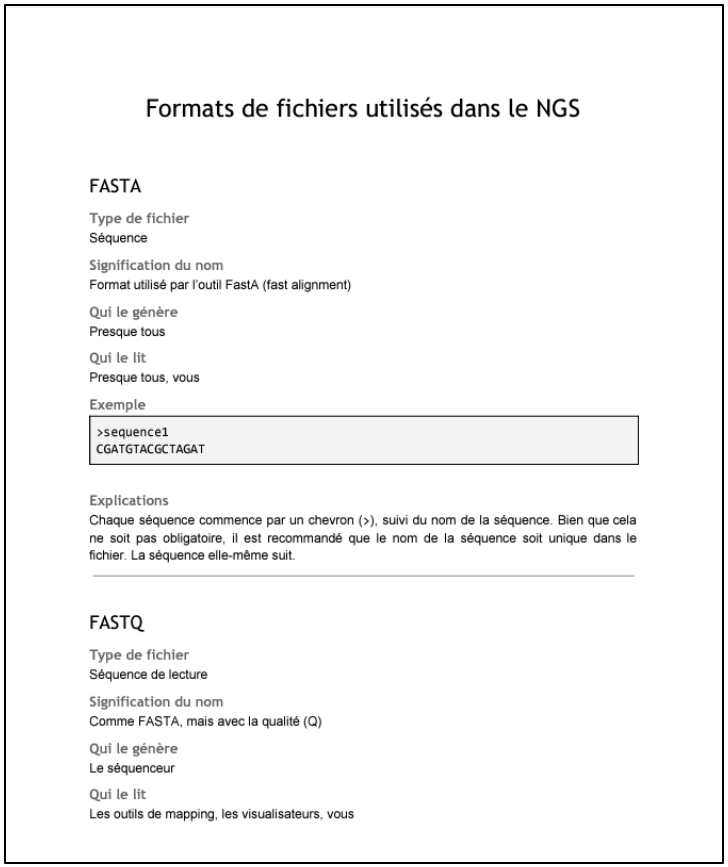
Pour résumer



https://pubmed.ncbi.nlm.nih.gov/27142620/

Ressource utile

https://ressources.france-bioinformatique.fr/sites/default/files/formats.pdf



Formats de fichiers utilisés dans le NGS

FASTA

Type de fichier
Séquence

Signification du nom
Format utilisé par l'outil FastA (fast alignment)

Qui le génère
Presque tous

Qui le lit
Presque tous, vous

Exemple

```
>sequence1
CGATGTACGCTAGAT
```

Explications
Chaque séquence commence par un chevron (>), suivi du nom de la séquence. Bien que cela ne soit pas obligatoire, il est recommandé que le nom de la séquence soit unique dans le fichier. La séquence elle-même suit.

FASTQ

Type de fichier
Séquence de lecture

Signification du nom
Comme FASTA, mais avec la qualité (Q)

Qui le génère
Le séquenceur

Qui le lit
Les outils de mapping, les visualisateurs, vous

FASTA, FASTQ, BED, GTF, GFF, SAM, BAM, BAI, WIG, BedGraph, BigWig, Pileup, VCF.

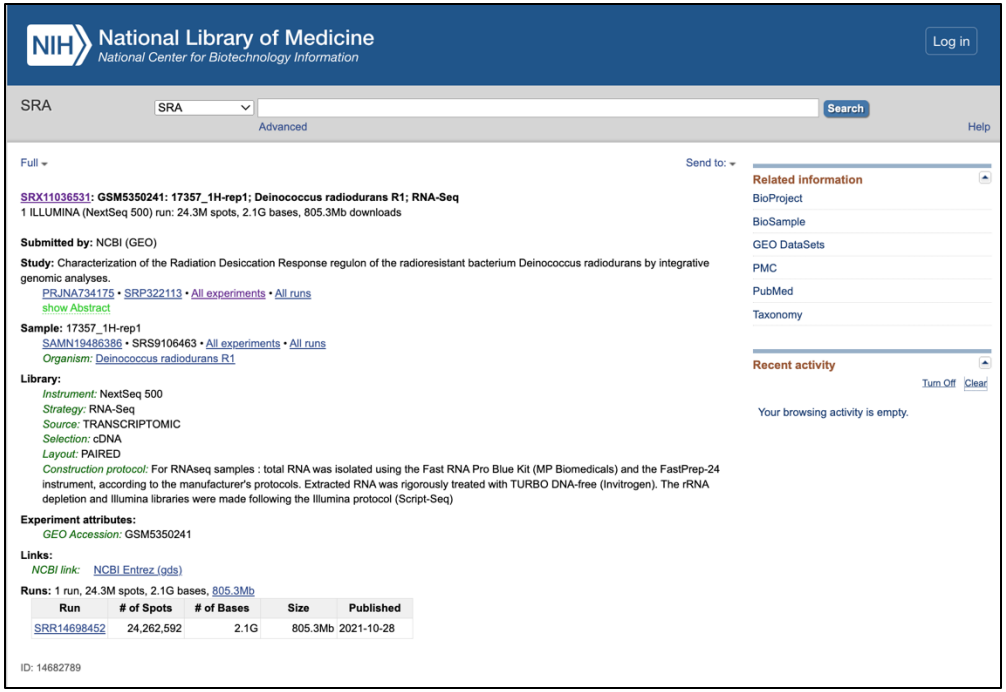
Ou trouver des données ?

- La banque de données SRA est incontournable
  https://www.ncbi.nlm.nih.gov/sra/docs/

SRA Mission

The SRA is a publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis.

- Les données sont trouvées via un moteur de recherche,
- Les données sont décrites via une page Web détaillée,
- Les données sont accessibles via de multiples liens.



https://www.ncbi.nlm.nih.gov/sra/SRX11036531[accn]

Gaëlle LELANDAIS

Version du document : 27/01/2025