

Bases de données en bioinformatique

Sacha Schutz
Interne en biologie moléculaire
M2 bioinformatique
10 mai 2016

Sommaire

1. Bases en informatique
2. Formats de données en bioinformatique
3. Bases de données publiques

Bases en informatique

Le numérique

Binaire : Base 2

2 Symboles : 1 0

1 0 0 0 1 1 0 1 1 0 0 0 1 1 0 1 1 1 0 1

1 bit

1 0 0 0 1 1 0 1 1 0 0 0 1 1 0 1 1 1 0 1

8 bits

1 octet ou 1 bytes



Combinaison = bases^{taille} = $2^8 = 256$

clef	valeur
1000001	A
1000010	B
1000000	@
0001101	<end ine>

⋮

⋮

Code ASCII

ADN : Base 4

4 Symboles : A T C G

A T C G C G T A A A A T C G C G T A A A

1 nucléotide

A T C G C G T A A A A T C G C G T A A A

3 nucléotides

1 codon ou 1 triplet



Combinaison = bases^{taille} = $4^3 = 64$

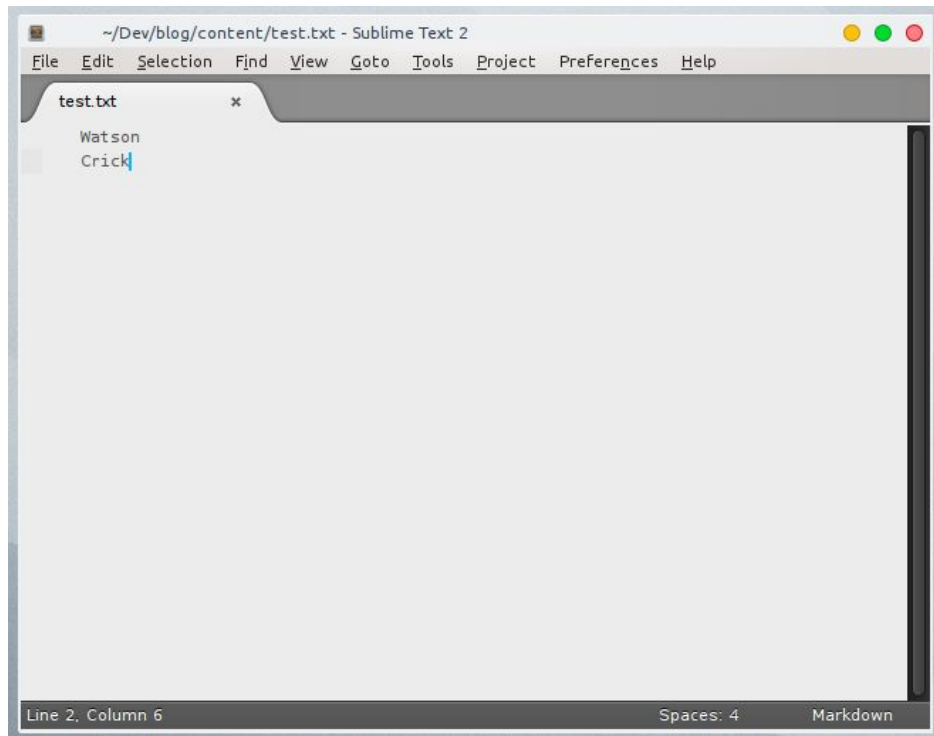
clef	valeur
ACG	Thr
AAG	Lys
GCG	Ala
TAG	<Stop>

⋮

⋮

Code génétique

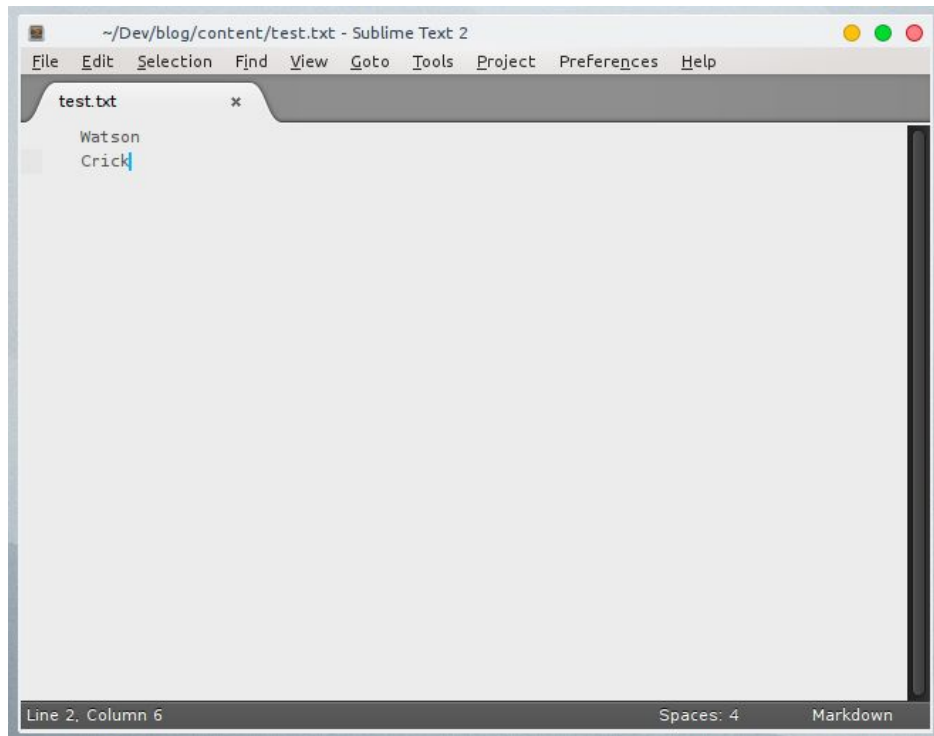
Taille d'un fichier



Quelle est la
taille du fichier ?

- En bytes ?
- En bits ?

Taille d'un fichier



Quelle est la
taille du fichier ?

12 bytes $\Rightarrow 11 + 1$

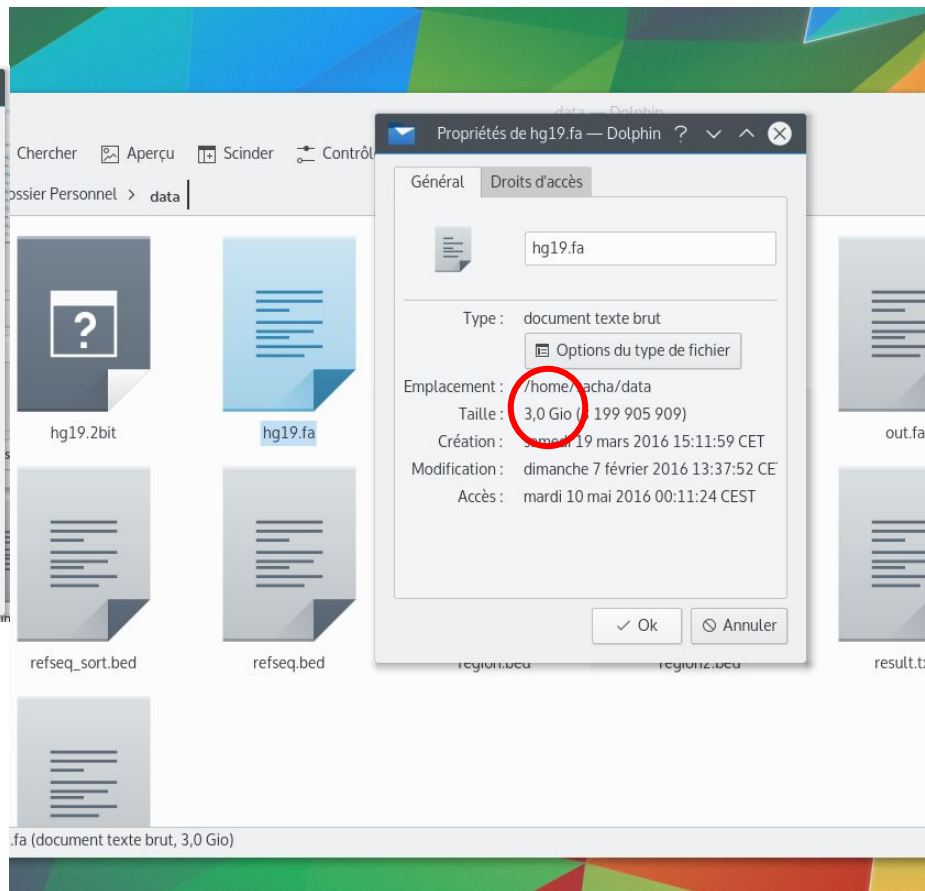
96 bits $\Rightarrow 12 \times 8$

Taille d'un fichier

Quelle est la taille d'un fichier contenant tout le génome ?

Taille d'un fichier

```
data : more
Fichier  Édition  Affichage  Signets  Configuration  Aide
aacgaggggtgaaccaggtccaggaagaaggtgcaaaagacagcattccagg
taaaagaacacagcttgaacaaaaagtgtgtaggggaaCCGCAAGCGGTCT
TGAGTGCTGAGGGTACAATCATCCTTGGGGAAGTACTAGAAGAAAGAATG
ATAAACAGAGGCCAGTTTGTAAAAACACTCAAAATTAAAGCTAGGAGTT
TGGACTTGTGGCAGGAATgaaatccttagacctgtgctgtccaatatggt
agccaccaggcacatgcagccactgagcacttgaatgtggatagtctga
attgagatgtgccataagtgtaaaatatgcaccaaatttcaaggctaga
aaaaaagaatgtaaaatatcttattttttatattgattacgtgctaaaa
taaccatatttgggatatactggattttaaaaatatatcactaatttcat
ctgtttctttttacttttAGAAATCACATATGTGACTTAAATATTTCTTT
TCTTTTCTTTTCTCTCACTCAGCGTCTGTGATTCCAAAGAAATGAGTC
TCTGCTGTTTTTGGGCAGCAGATATCCTAGAATGGACTCTGACCTAAGCA
TCAAAATTAATCATCATAACGTTATCATTTTATGGCCCCTTCTTCCTATA
TCTGGTAGCTTTTAAATGATGACCATGTAGATAATCTTTATTGTCCCTCT
TTCAGCAGACGGTATTTTCTTATGCTACAGTATGACTGCTAATAATACCT
```



Fichier texte et binaire

2 types de fichier

Fichier texte

- Unité : 1 **caractère** (Byte)
- Lisible par un humain
- Dans un éditeur de texte
- Prend beaucoup d'espace
- Exemple :
 - ◆ fasta
 - ◆ sam
 - ◆ csv, xml
 - ◆ vcf, fastq

Fichier binaire

- Unité : Un **bit**
- Non lisible par un humain
- Dans un éditeur hexadécimal
- Prend moins d'espace
- Exemple :
 - ◆ png, jpg, mp3, wav
 - ◆ bam
 - ◆ excel, word
 - ◆ vcf.gz, fastq.gz

Fichier texte et binaire

2 types de fichiers

Fichier texte

Exemple : informations [Vrai ou faux]

Vrai,Vrai,Faux,Faux,Vrai,Faux,Vrai,Faux

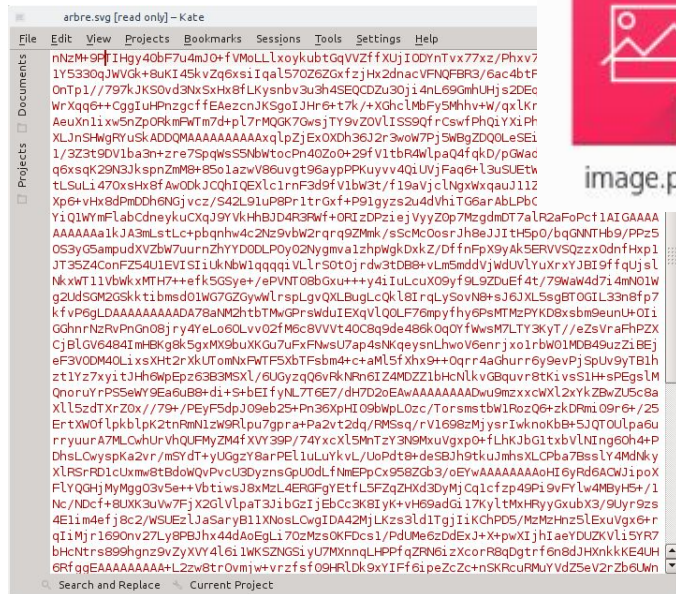
Total : 39 octets

Fichier binaire

Exemple : informations [Vrai ou faux]

11001010

Total : 1 octet





Formats de données en bioinformatique



Les différents formats

- En bioinformatique, la majorité des données sont dans un format textuel
 - FASTA, FASTQ, SFF, SAM, VCF, BED, BEDGRAPH, GFF, GTF3, MAF, TSV, CSV, XML, JSON
- Pour des raisons de performance et de compression, certaines sont dans un format binaire
 - 2BIT, ABIF, BAM, VCF.GZ, FASTQ.GZ

Format et spécification

Le **format** d'un fichier décrit comment les données sont représentées. Cette description est fournie par une documentation qu'on appelle **spécification**.

Format et spécification

Mêmes données, différents formats

```
users : {  
  first_name: "James",  
  last_name: "Watson",  
  birthday: "1928-04-06"  
}
```

Format **JSON**

<https://tools.ietf.org/html/rfc4627>

```
<users>  
  <first_name>James</firstname>  
  <last_name> Watson</last_name>  
  <birthday>19280406</birthday>  
</users>
```

Format **XML**

<https://www.w3.org/TR/REC-xml/>

Séquences et régions

En génomique, on peut catégoriser les formats en deux groupes

- Les formats décrivant des **séquences**
- Les formats décrivant des **régions**

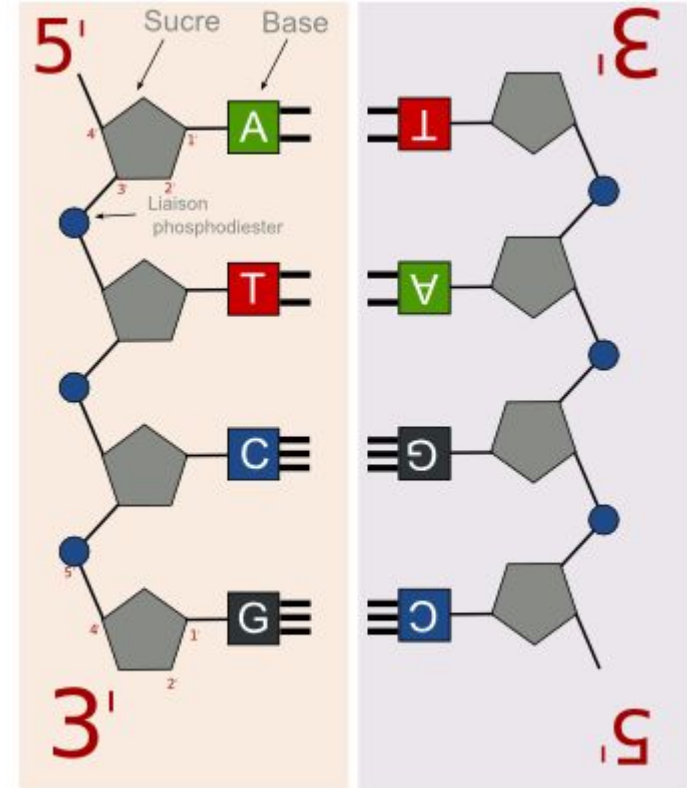
Les séquences

Une séquence d'**ADN**

s'écrit **toujours** dans le sens **5'- 3'**.

Quelle est la séquence sur :

- Le brin de gauche ?
- Le brin de droite ?



Les séquences

Format Fasta [*.fa ; *.fasta]

>Identifiant1 Commentaire

ACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACG
TGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGG
GTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTG
CTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTA
GTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGT

>Identifiant2 Commentaire

ACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACG
TGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGG
GTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTGCTAGTGGACTGTACGTACGTGGGTG

Les séquences

Format FastQ [*.fastq]

@SEQ_ID1

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

! ' * ((((* * * +)) % % % + +) (% % % %) . 1 * * * - + * ' ')) * * 5 5 C C F > > > > > C C C C C C C C 6 5

@SEQ_ID2

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

! ' * ((((* * * +)) % % % + +) (% % % %) . 1 * * * - + * ' ')) * * 5 5 C C F > > > > > C C C C C C C C 6 5

Les séquences

Format FastQ [*.fastq]

@SEQ_ID1

GATTTGGGGT

42,40,134,47,36 35 40 40 40



@SEQ_ID1

GATTTGGGGT

+

*(å/\$#((((



0	32	64	96	128	160	192	224	256
1	33	65	97	129	161	193	225	257
2	34	66	98	130	162	194	226	258
3	35	67	99	131	163	195	227	259
4	36	68	100	132	164	196	228	260
5	37	69	101	133	165	197	229	261
6	38	70	102	134	166	198	230	262
	39	71	103	135	167	199	231	263
	40	72	104	136	168	200	232	264
	41	73	105	137	169	201	233	265
	42	74	106	138	170	202	234	266
11	43	75	107	139	171	203	235	267
12	44	76	108	140	172	204	236	268
13	45	77	109	141	173	205	237	269
14	46	78	110	142	174	206	238	270
15	47	79	111	143	175	207	239	271
16	48	80	112	144	176	208	240	272
17	49	81	113	145	177	209	241	273
18	50	82	114	146	178	210	242	274
19	51	83	115	147	179	211	243	275
20	52	84	116	148	180	212	244	276
21	53	85	117	149	181	213	245	277
22	54	86	118	150	182	214	246	278
23	55	87	119	151	183	215	247	279
24	56	88	120	152	184	216	248	280
25	57	89	121	153	185	217	249	281
26	58	90	122	154	186	218	250	282
27	59	91	123	155	187	219	251	283
28	60	92	124	156	188	220	252	284
29	61	93	125	157	189	221	253	285
30	62	94	126	158	190	222	254	286
31	63	95	127	159	191	223	255	287

Les séquences

Format GeneBank

Cherchez sur **NCBI** le gène *GJB2* dans la section **Nucléotide**

- Sur quel chromosome est le gène ?
- Combien d'exons comporte le gène ?
- Quelle est la séquence des 4 premiers nucléotides du premier intron ?

http://www.ncbi.nlm.nih.gov/nuccore/196115124#feature_196115124_gene_0

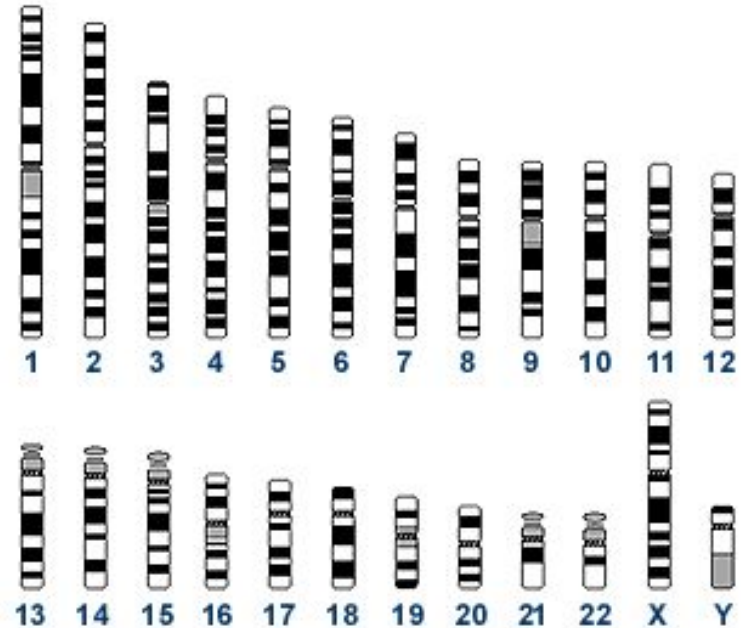


Les régions

Les coordonnées génomiques permettent de localiser avec précision une région du génome.

<chromosome>:<start>-<end>

chr7:117465784-117715971



Les régions

Exemple : Récupérer une
séquence depuis
Ensembl.

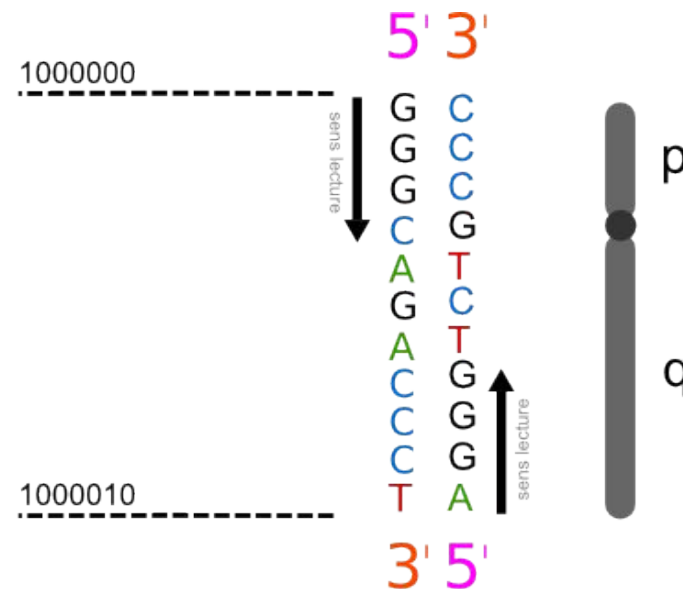
GRCh38 ? hg19 ?



<http://rest.ensembl.org/sequence/region/human/7:117465784..1177159171>



<http://grch37.rest.ensembl.org/sequence/region/human/7:117465784..1177159171>



5:1000000..1000010:1
5' GGGCAGACCCT 3'

5:1000000..1000010:-1
5' AGGGTCTGCCC 3'

Les régions

Attention, suivant le format, la première base est numéroté **0** ou **1**.

A T C G C G T A A A

▲
0

A T C G C G T A A A

▲
1

Table 9-1. Range types of common bioinformatics formats

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

Les régions

Format BED

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

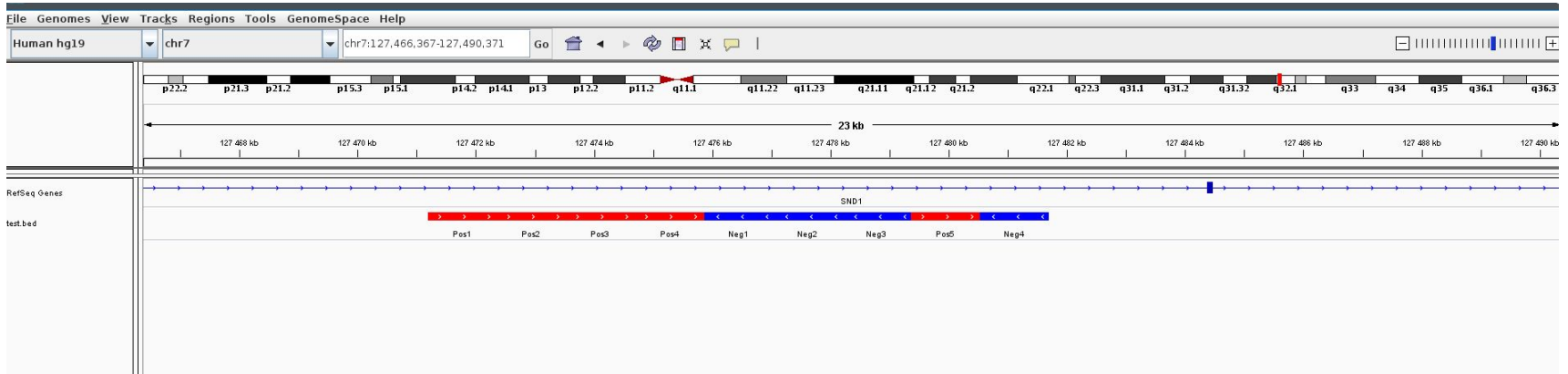
Attention

- Le premier nucléotide est numéroté 0.
- $\text{end} - \text{start} = \text{taille de la séquence}$

0 9
▼ ▼
ACGTGTCATG
chr7:0-10

Les régions

Format BED



Les régions

Format BEDGRAPH

```
#Note, zero-relative, half-open coordinate system in use for bedGraph  
format
```

```
track type=bedGraph name="BedGraph Format" description="BedGraph  
format" visibility=full color=200,100,0 altColor=0,100,200 priority=20
```

```
chr19 49302000 49302300 -1.0
```

```
chr19 49302300 49302600 -0.75
```

```
chr19 49302600 49302900 -0.50
```

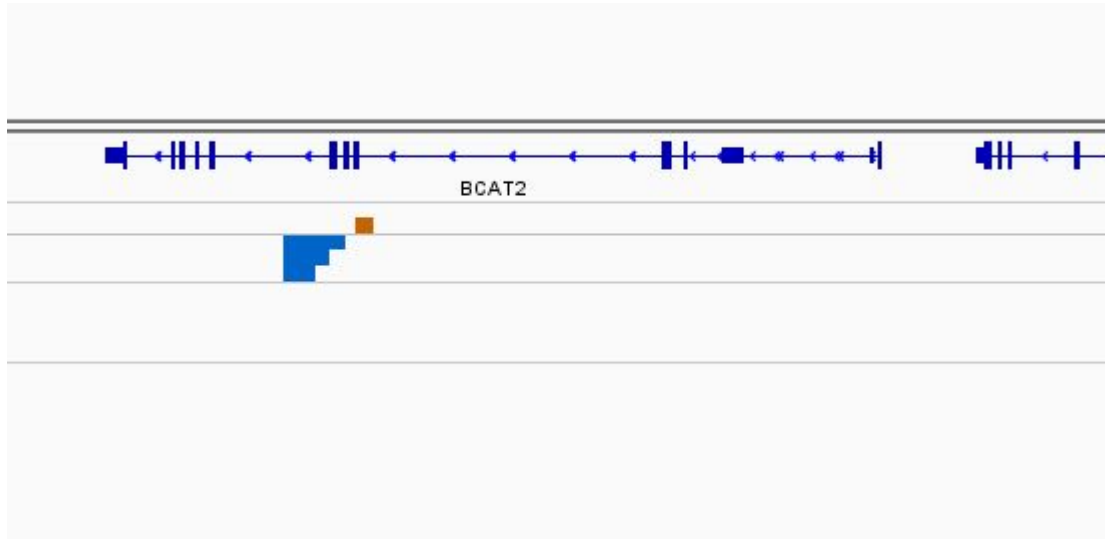
```
chr19 49302900 49303200 -0.25
```

```
chr19 49303200 49303500 0.0
```

```
chr19 49303500 49303800 0.25
```

Les régions

Format BEDGRAPH



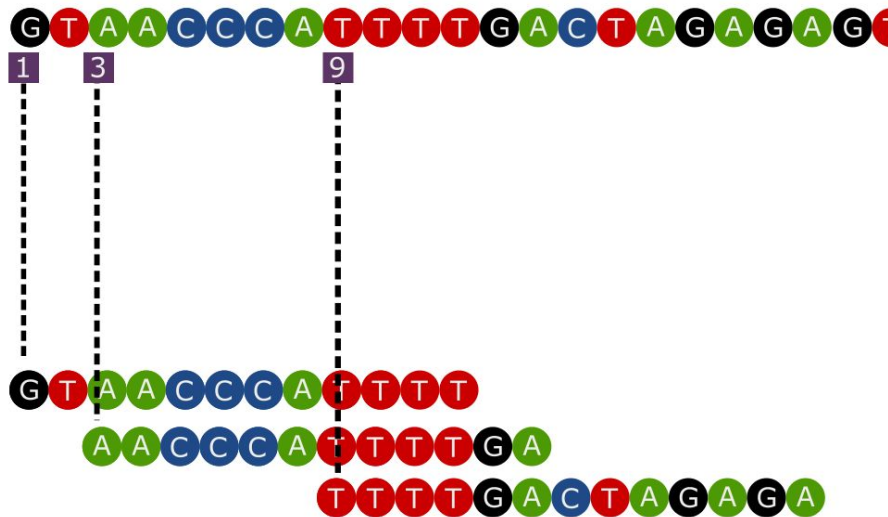
Les régions

Format SAM

Contient des séquences (reads) alignées sur le génome.

Dans l'idéal :

chr7	1324324	ACGTGCGTTTTCGT
chr8	1724354	GCGTGATGCGTAAG
chr8	1424324	GTATGTTATATGTA



Les régions

Format SAM

En vrai...

```
@HD VN:1.5 SO:coordinate
```

```
@SQ SN:ref LN:45
```

```
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

Les régions

Format SAM

Fichier texte

SAM



SAMtools

Fichier binaire

BAM

Profondeur ?
Couverture ?

[illegible]

Les régions

Format GFF / GTF3

```
chr5 illumina CDS      380   401   .   +   0   gene_id "001";
chr5 illumina CDS      501   650   .   +   2   gene_id "001";
chr5 illumina CDS      700   707   .   +   2   gene_id "001";
chr5 illumina start_codon 380   382   .   +   0   gene_id "001";
chr5 illumina stop_codon  708   710   .   +   0   gene_id "001";
```

Les régions

Format VCF

- **L'empilement** (pileup) consiste à comptabiliser la proportion en base pour chaque position.
- Le **variant calling** consiste à détecter les positions différentes de la référence.

Les régions

Format VCF

reference { ACGT **G**TCATG ACGTGTCATG

reads {

ACGT	G	AC		ACGT	GT
ACGT	G	ACAT		TGTCAT	G
	C	ACAT		TGTCAT	G
ACGT	G	ACAT		TCAT	G
ACGT	G	AGAT	GA	ACGT	
ACGT	G	AGAT	GA	ACGT	GTC
ACGT	G	AGAT	GA		

Les régions

Format VCF

reference { ACGTGTTCATGACGTGTCATG

reads {

ACGTGAC	ACGTGT
ACGTGACATG	TGTCATG
CGACATG	TGTCATG
ACGTGACATG	TCATG
ACGTGAGATGACGT	
ACGTGAGATGACGTGTC	
ACGTGAGATGA	

Les régions

Format VCF

reference { ACGTGTCATGACGTGTCATG

reads {

ACGTGAC	ACGTGT
ACGTGACATG	TGTCATG
CGACATG	TGTCATG
ACGTGACATG	TCATG
ACGTGAGATGACGT	
ACGTGAGATGACGTGTC	
ACGTGAGATGA	

Les régions

Format VCF

VCF header

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T		.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Deletion

SNP

Large SV

Insertion

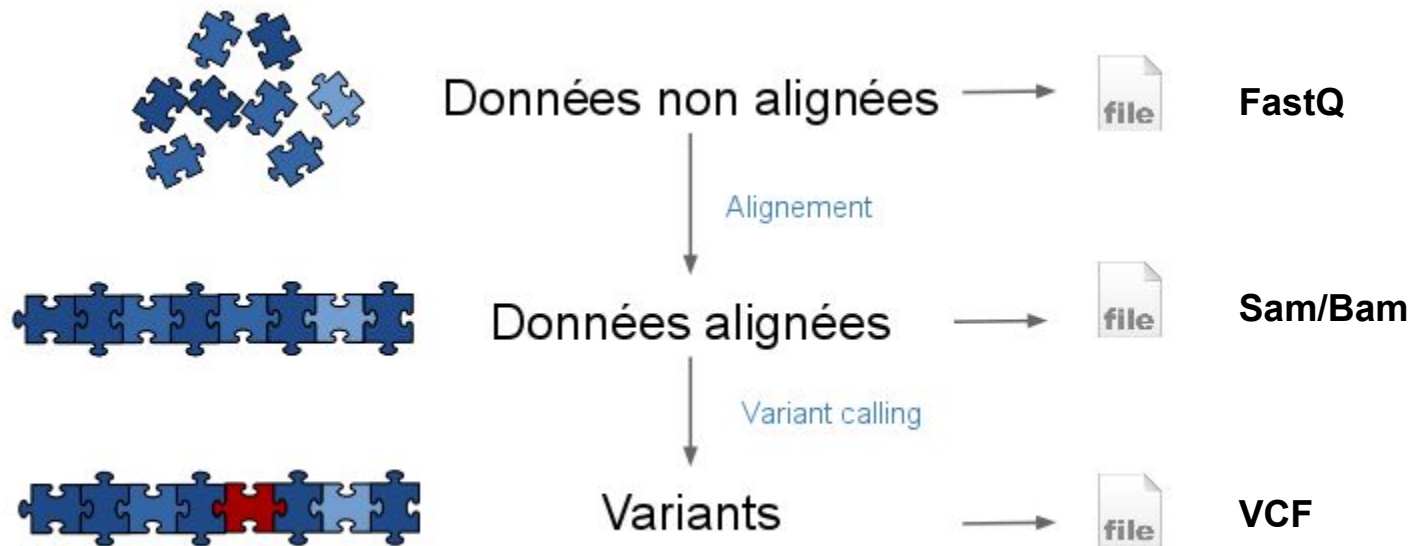
Other event

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Pipeline NGS



Bases de données publiques

Différentes bases de données

Généraliste

- **UCSC**
- Ensembl

Cancer spécifique

- TCGA
- **ICGC**
- Cosmic

UCSC : Le Golden path



<http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/>

UCSC : Table browser

UCSC : Table Browser

Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see [Using the Table Browser](#) for a description of the controls in this form, the [User's Guide](#) for general information and sample queries, and the OpenHelix Table Browser [tutorial](#) for a narrated presentation of the software features and usage. For more complex queries, you may want to use [Galaxy](#) or our [public MySQL server](#). To examine the biological function of your set through annotation enrichments, send the data to [GREAT](#). Send data to [GenomeSpace](#) for use with diverse computational tools. Refer to the [Credits](#) page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the [Sequence and Annotation Downloads](#) page.

clade: Mammal : **genome:** Human : **assembly:** Dec. 2013 (GRCh38/hg38) :

group: Comparative Genomics : **track:** Conservation :

table: Cons 100 Verts (phastCons100way) :

region: ☒ genome ☐ position chr21:30000000-40000000

filter:

subtrack merge:

intersection:

correlation:

output format: data points : Send output to ☐ [Galaxy](#) ☐ [GREAT](#) ☐ [GenomeSpace](#)

output file: (leave blank to keep output in browser)

file type returned: ☒ plain text ☐ gzip compressed

<https://genome.ucsc.edu/cgi-bin/hgTables>

Cosmic

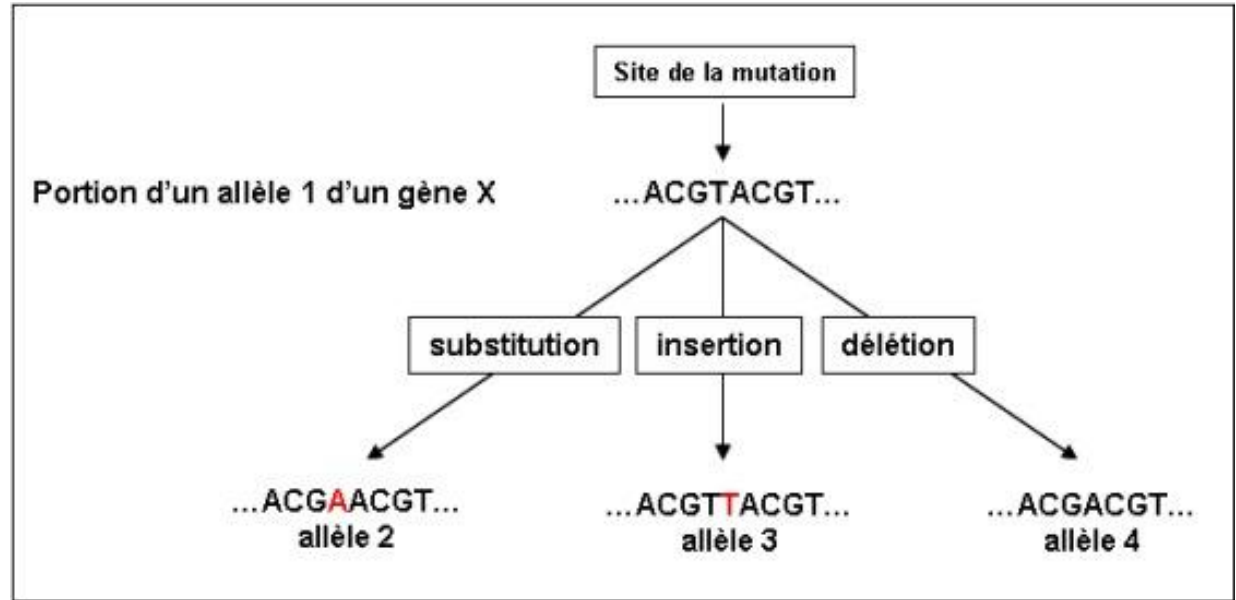
- Mutations trouvées :
 - Par croisement d'article
 - Données de séquençage (TCGA / ICGC)
 - Cancer Gene census : 572 gènes

<http://cancer.sanger.ac.uk>

Nomenclature des mutations

A l'échelle de la base

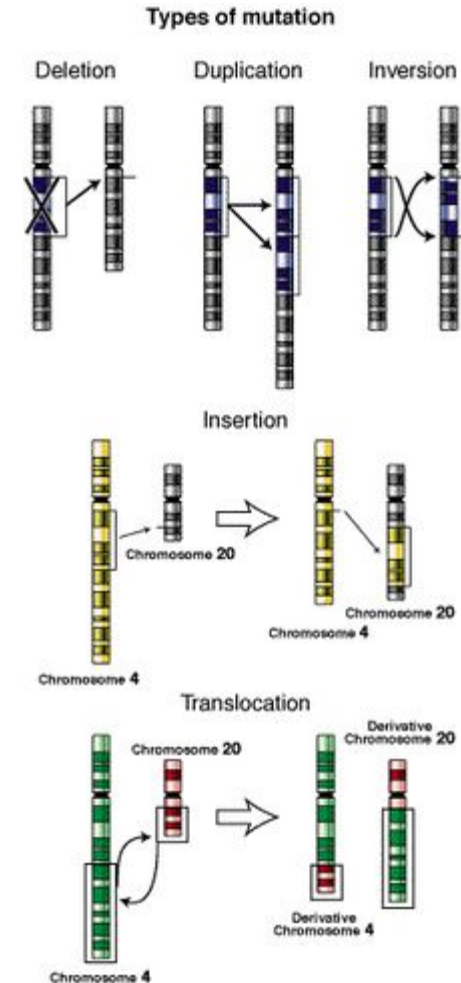
- Substitution
- Délétion
- Insertion



Nomenclature des mutations

A l'échelle du chromosome

- Anomalie du nombre
- Anomalie de structure



Nomenclature des mutations

Norme HGVS

Coordonnée génomique

g.231333423 A>C

g.43331234DelC

g.324234InsA

Coordonnée exonique

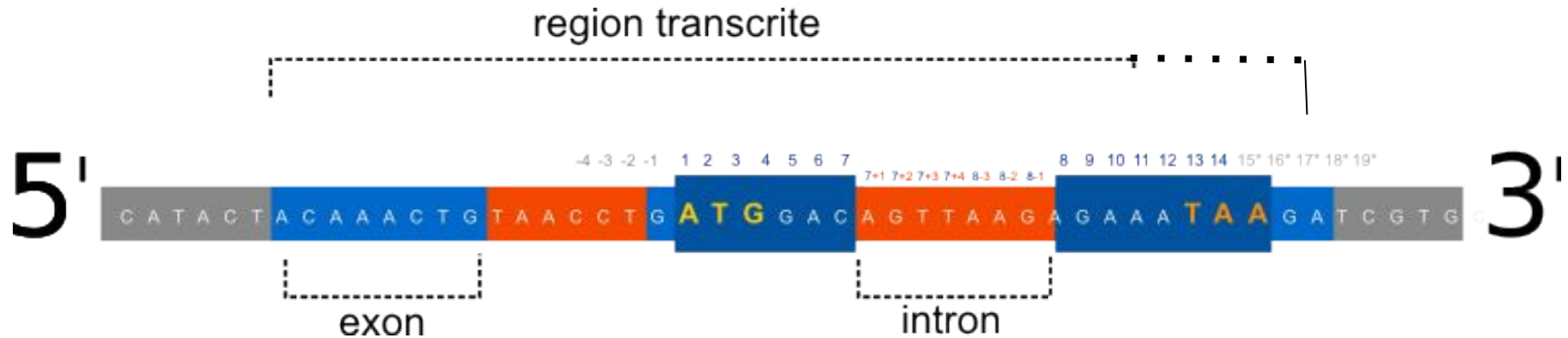
c.504A>C

c.423DelC

c.32InsA

Nomenclature des mutations

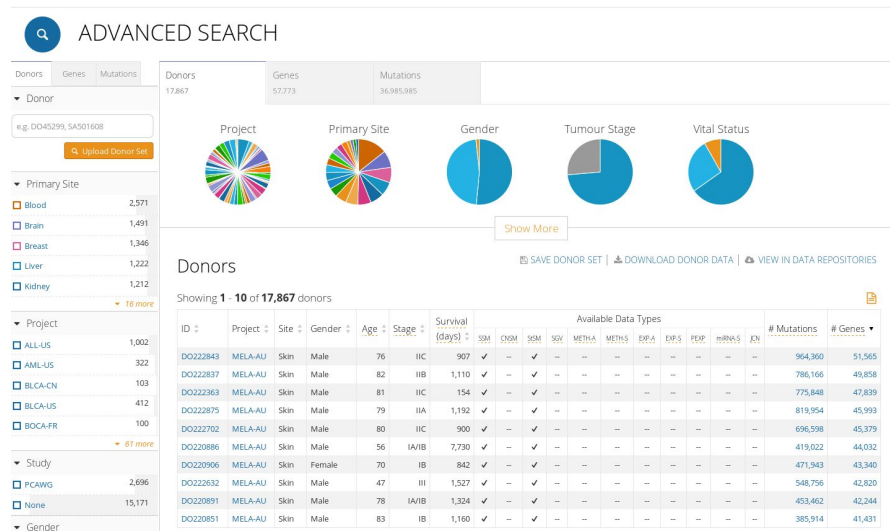
Norme HGVS



<http://www.hgvs.org/mutnomen/recs.html>

ICGC

ICGC : International Cancer Genome Consortium



<http://icgc.org/>

Références

Comprendre le NGS

https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_introduction.php#similarities

<http://www.biorigami.com/>

<http://bioinfo-fr.net/>

Mon blog

<http://dridk.me>

Twitter

[@dridk](https://twitter.com/dridk)