GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

Bases de données en bioinformatique

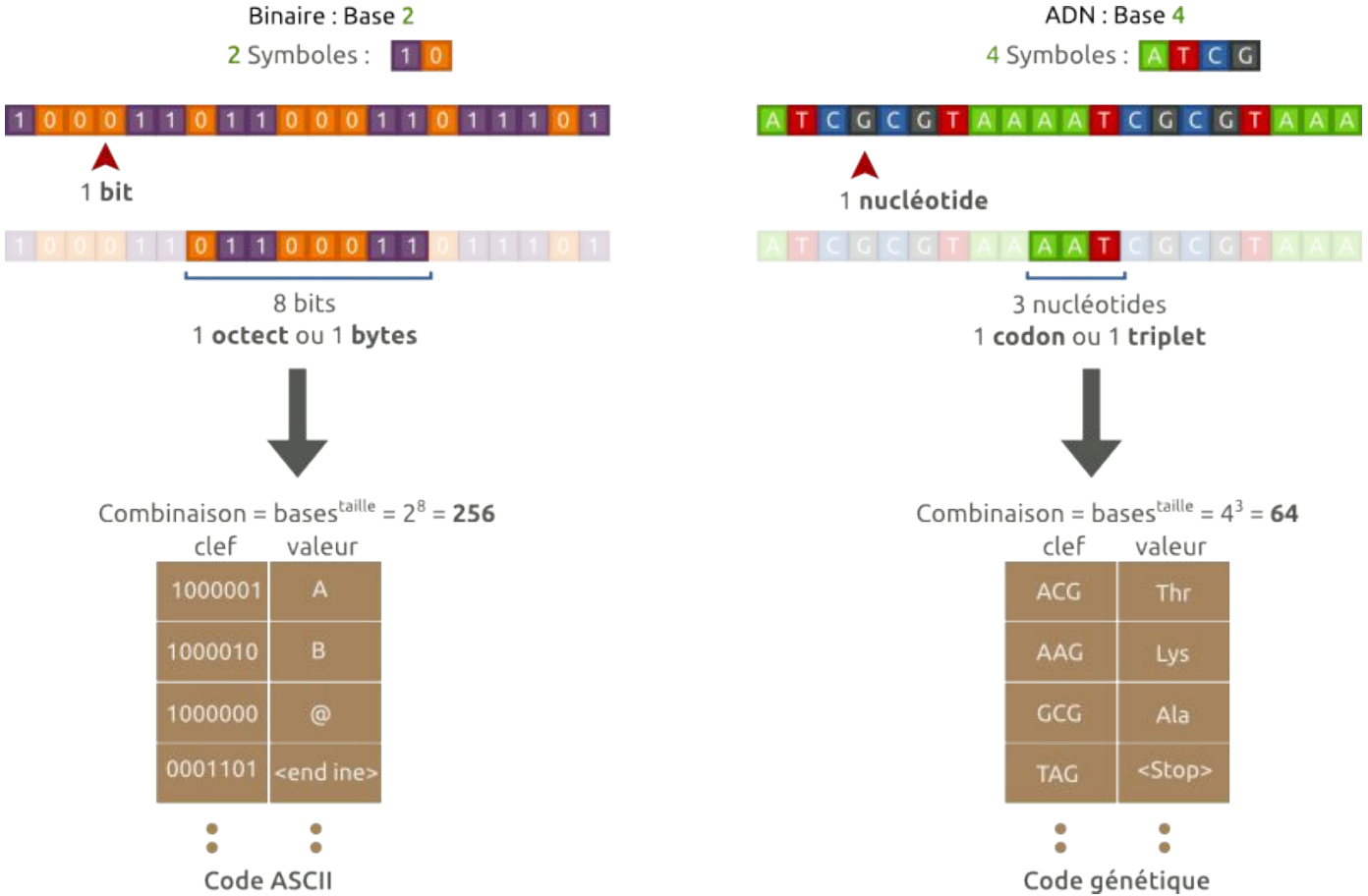
Sacha Schutz
Interne en biologie moléculaire
M2 bioinformatique
10 mai 2016

Sommaire

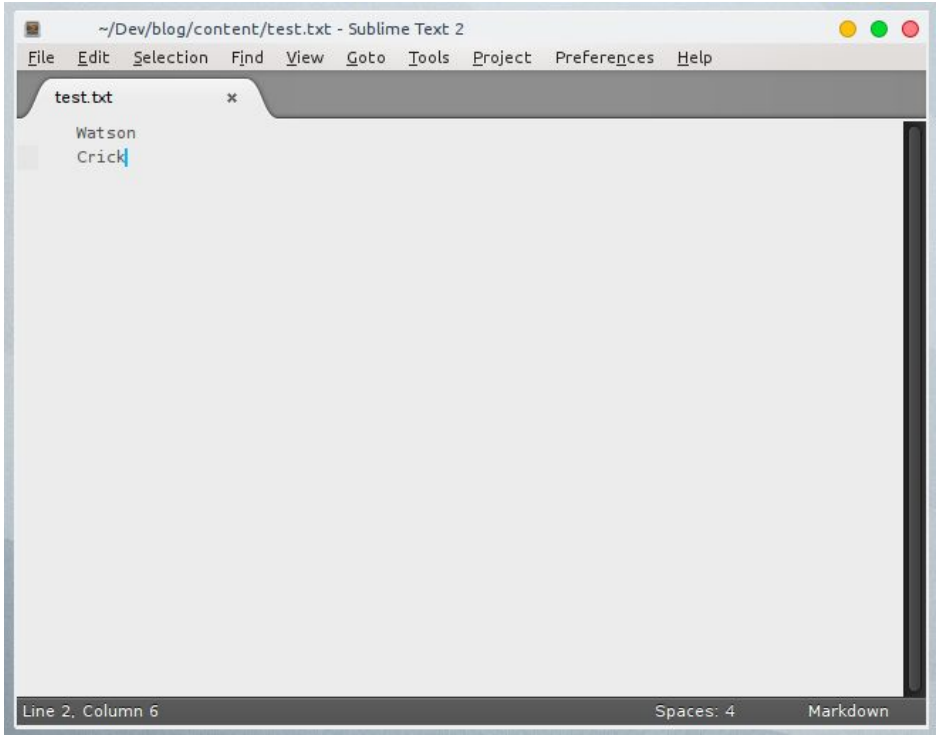
1. Bases en informatique
2. Formats de données en bioinformatique
3. Bases de données publiques

Bases en informatique

Le numérique



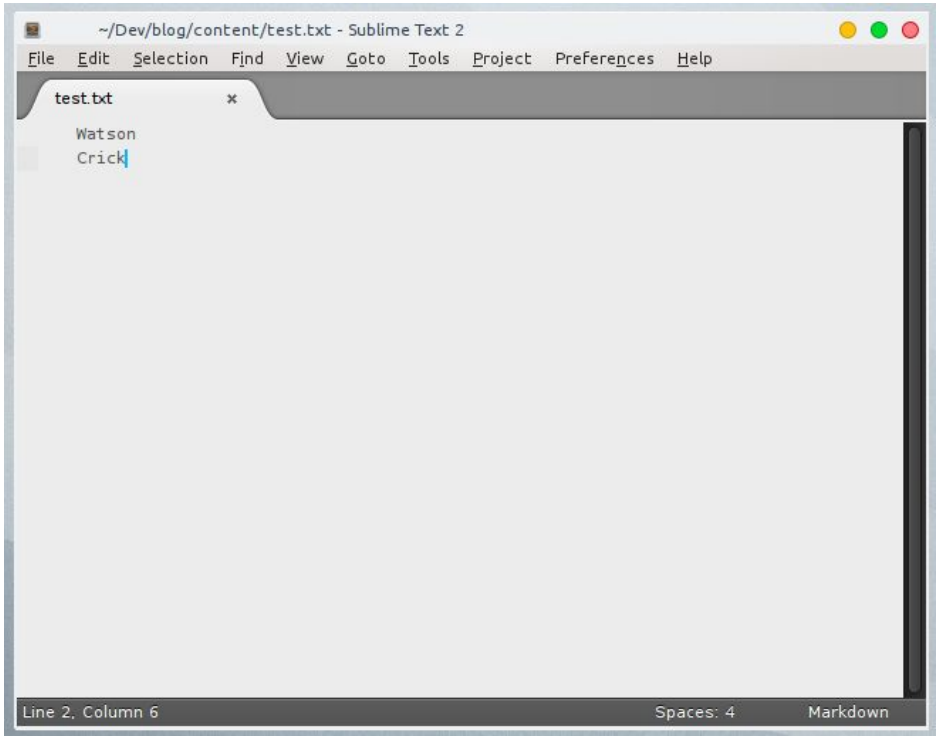
Taille d’un fichier



Quelle est la taille du fichier ?

- En bytes ?
- En bits ?

Taille d’un fichier



Quelle est la taille du fichier ?

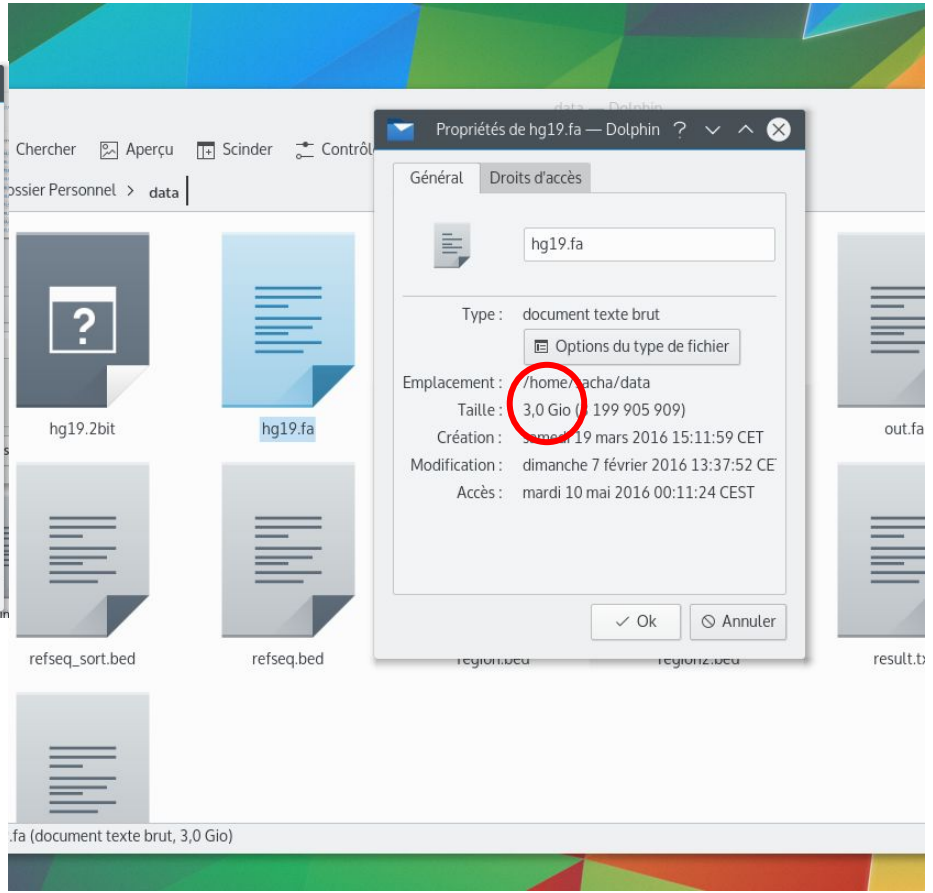
12 bytes ⇒ 11 + 1

96 bits ⇒ 12 × 8

Taille d’un fichier

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

Taille d'un fichier



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]

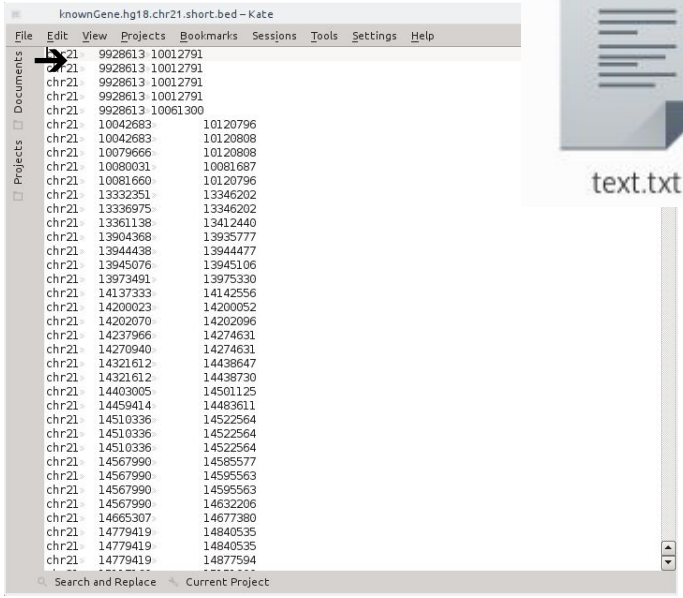
11001010

Total : 1 octet

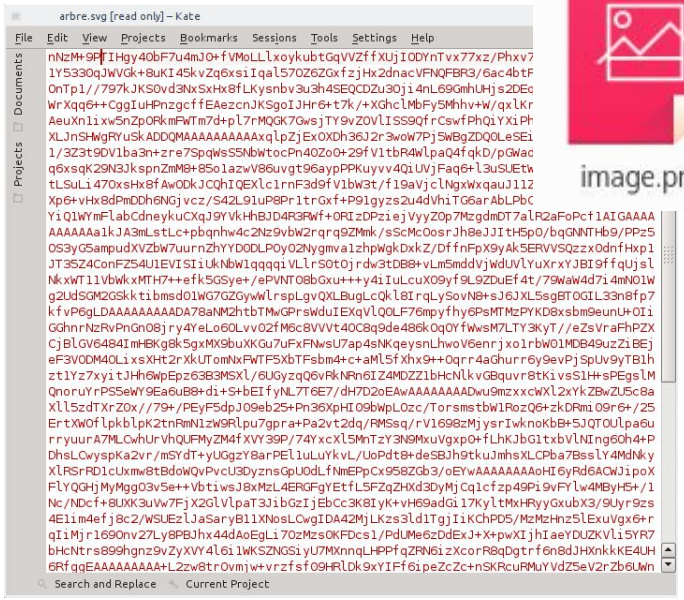
Fichier texte et binaire

2 types de fichiers

Fichier texte



Fichier binaire



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**

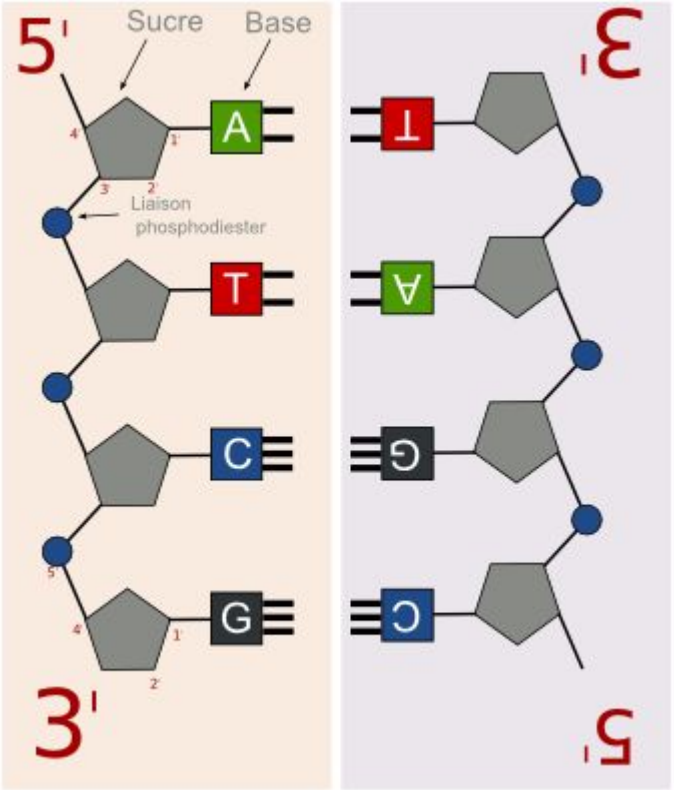
Les séquences

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]

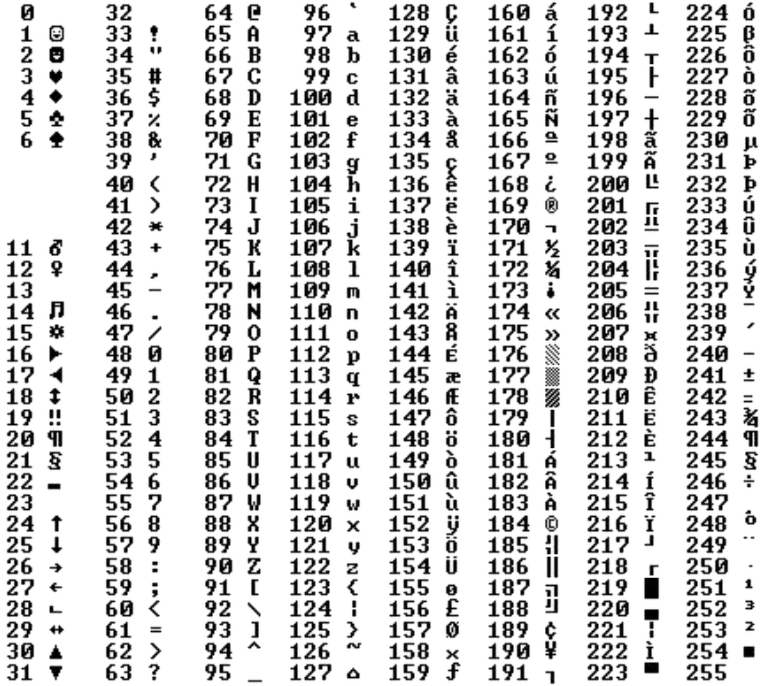
```
@SEQ_ID1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65

@SEQ_ID2
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Les séquences

Format FastQ [*.fastq]

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

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



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 ?

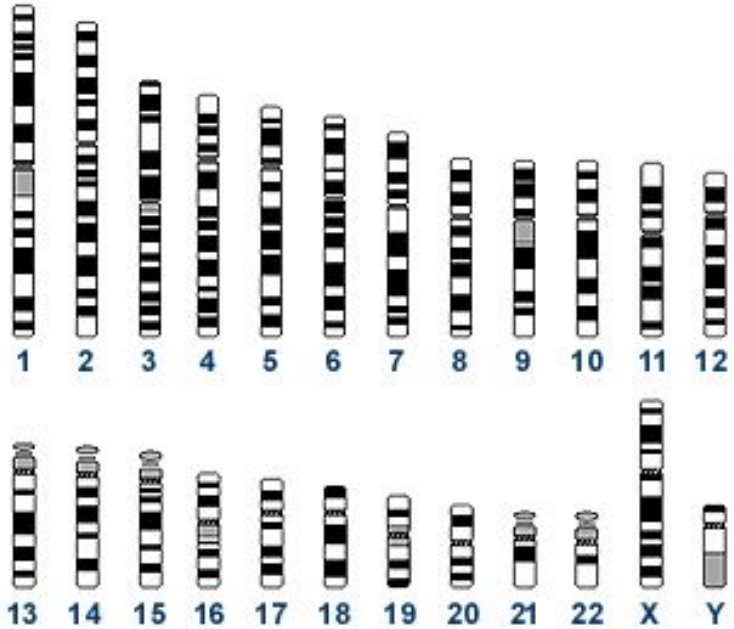
http://www.ncbi.nlm.nih.gov/nuccore/196115124#feature_196115124_gene_0

Les régions

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

<**chromosome**>:<**start**>-<**end**>

chr7:117465784-117715971



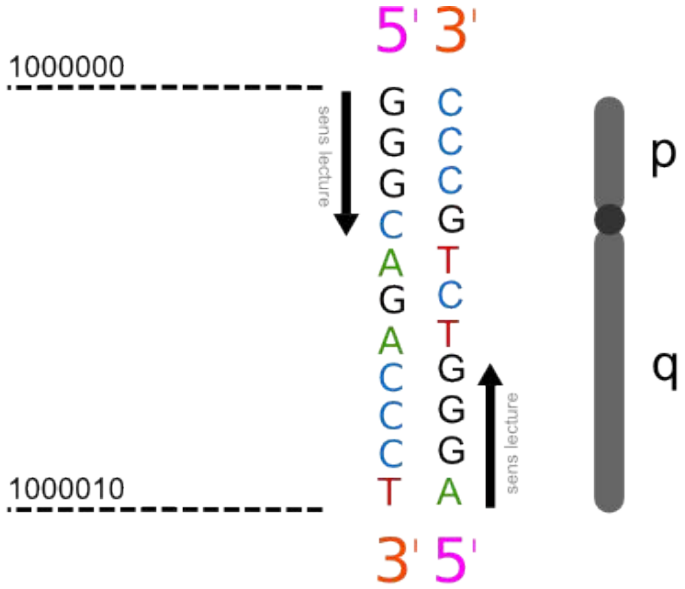
Les régions

Exemple : Récupérer une séquence depuis Ensembl.

GRCh38 ? hg19 ?

http://rest.ensembl.org/sequence/region/human/7:117465784..177159171

http://grch37.rest.ensembl.org/sequence/region/human/7:117465784..117715971



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**.

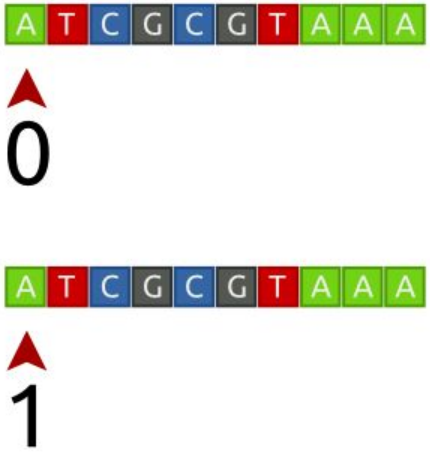


Table 9-1. Range types of common bioinformatics formats

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.
- end - start = taille de la séquence

Les régions

Format BED

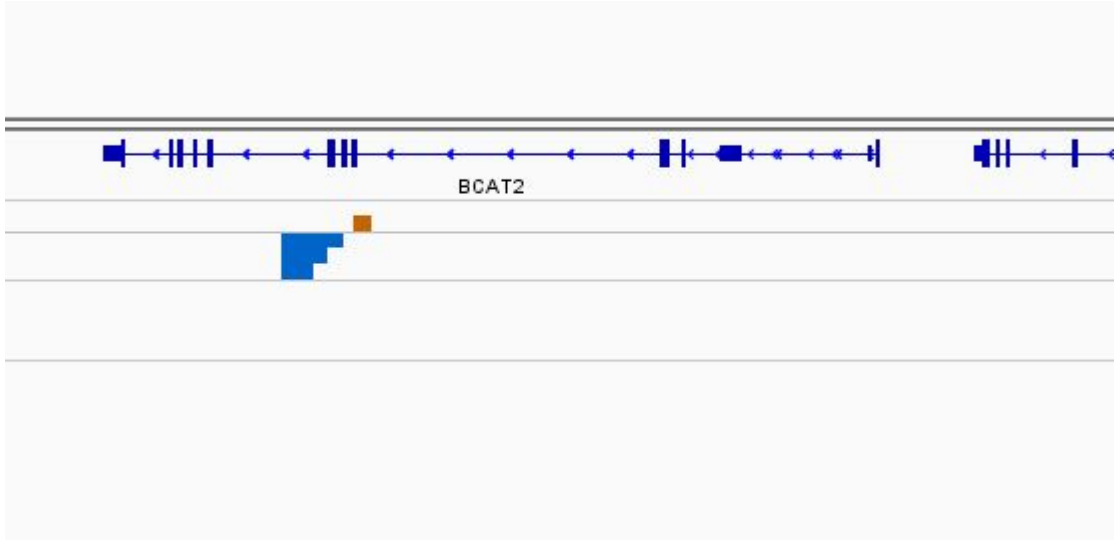
Les régions

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



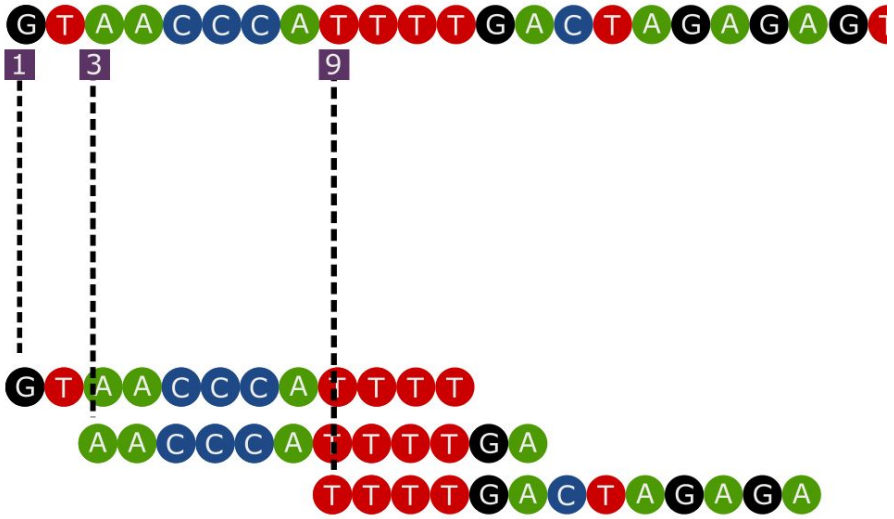
Les régions

Format SAM

Contient des séquences (reads) alignées sur le génome.

Dans l'idéal :

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



Les régions

Format SAM

En vrai...

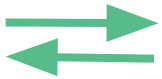
```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
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

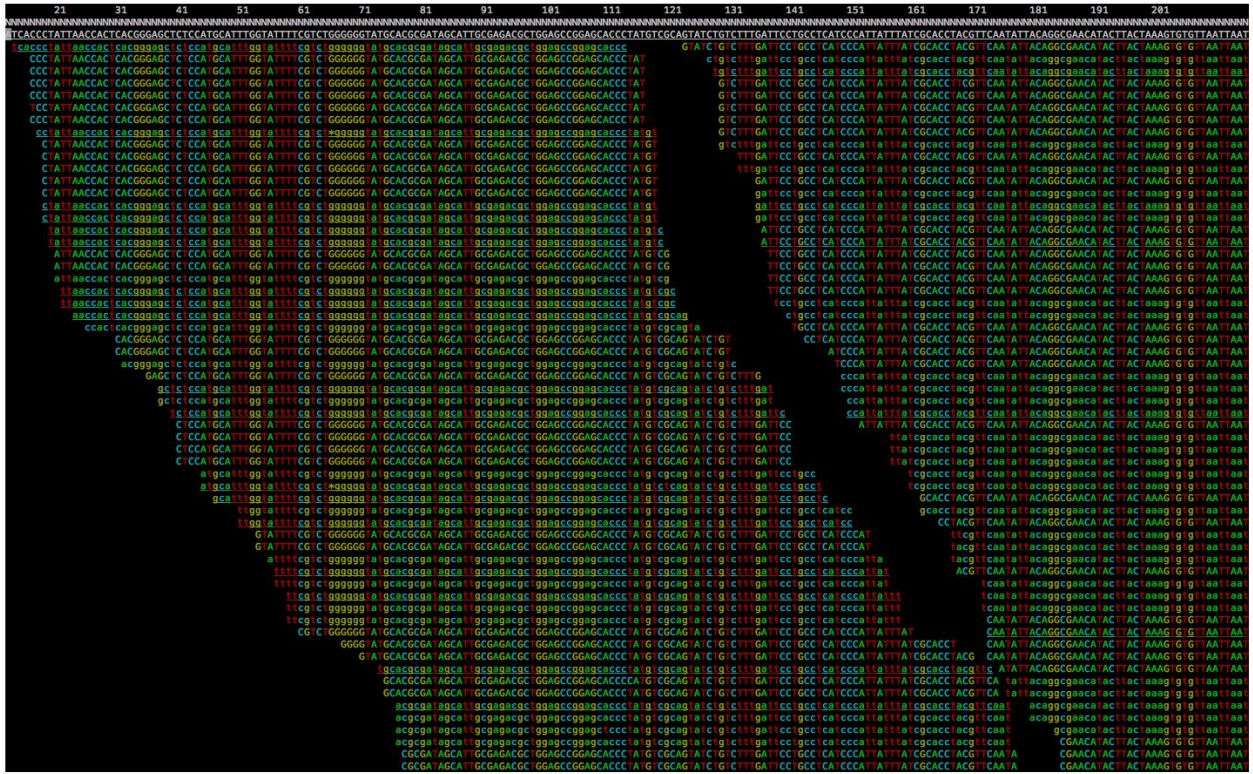
Fichier binaire

BAM

Les régions

Format SAM

Profondeur ?
Couverture ?



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

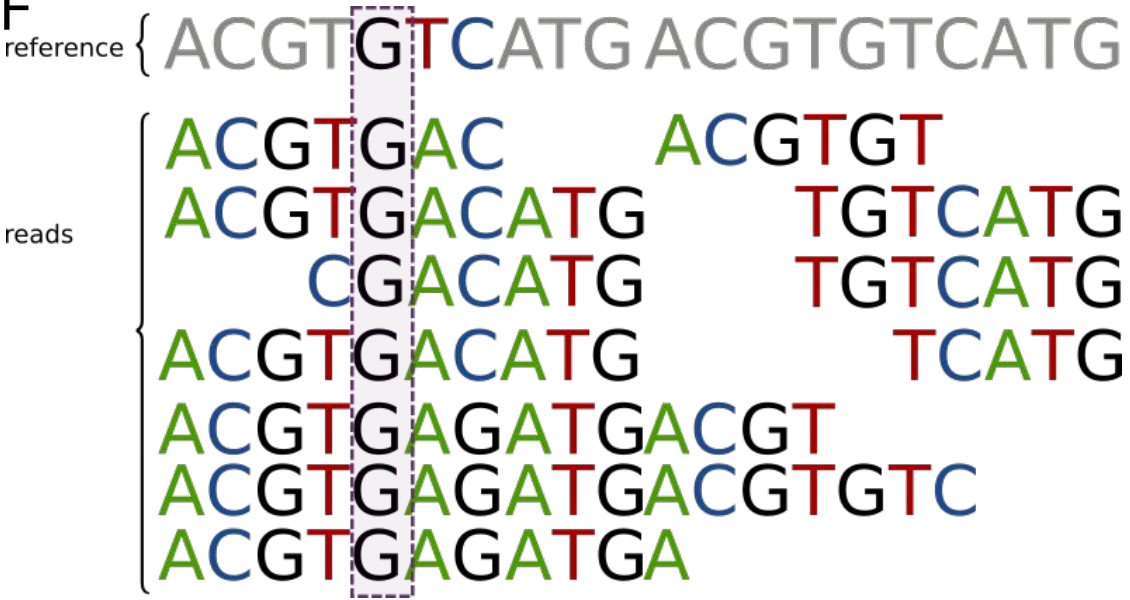
Format VCF

- **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.

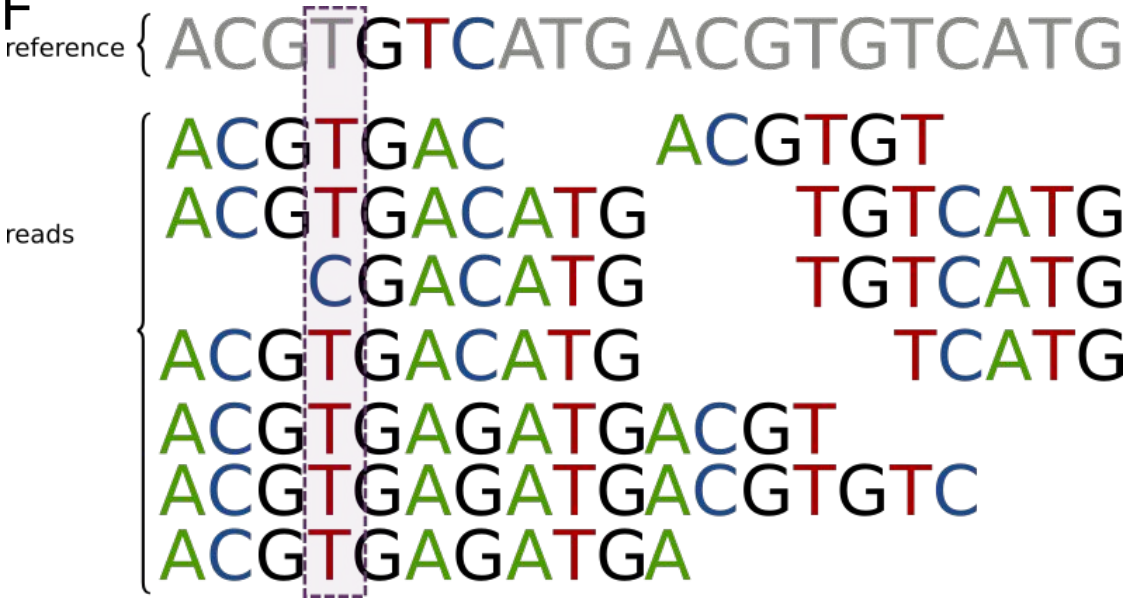
Les régions

Format VCF



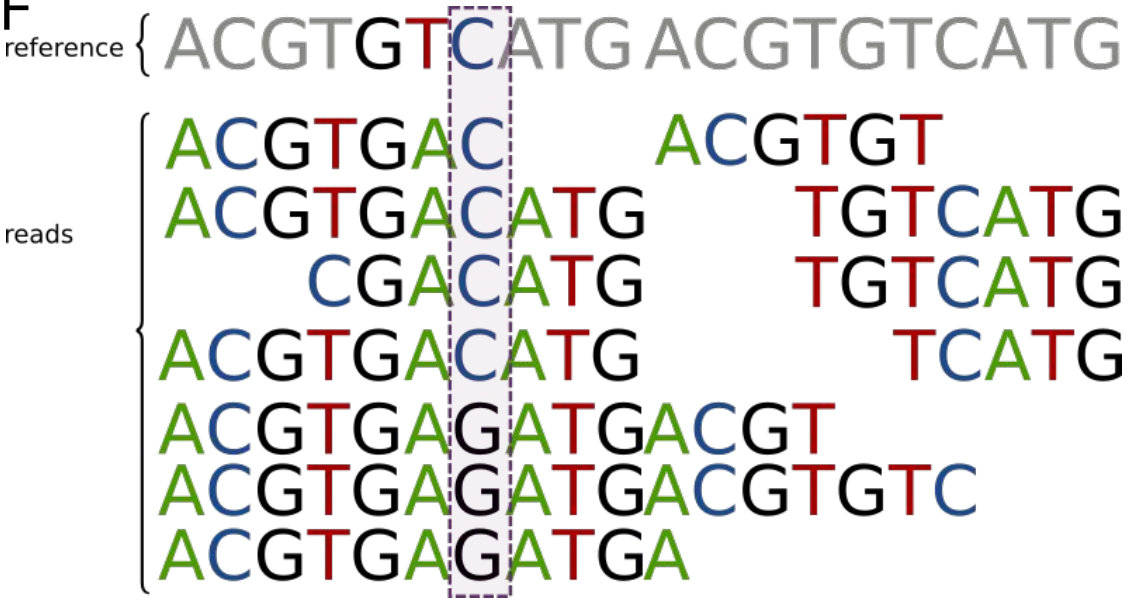
Les régions

Format VCF



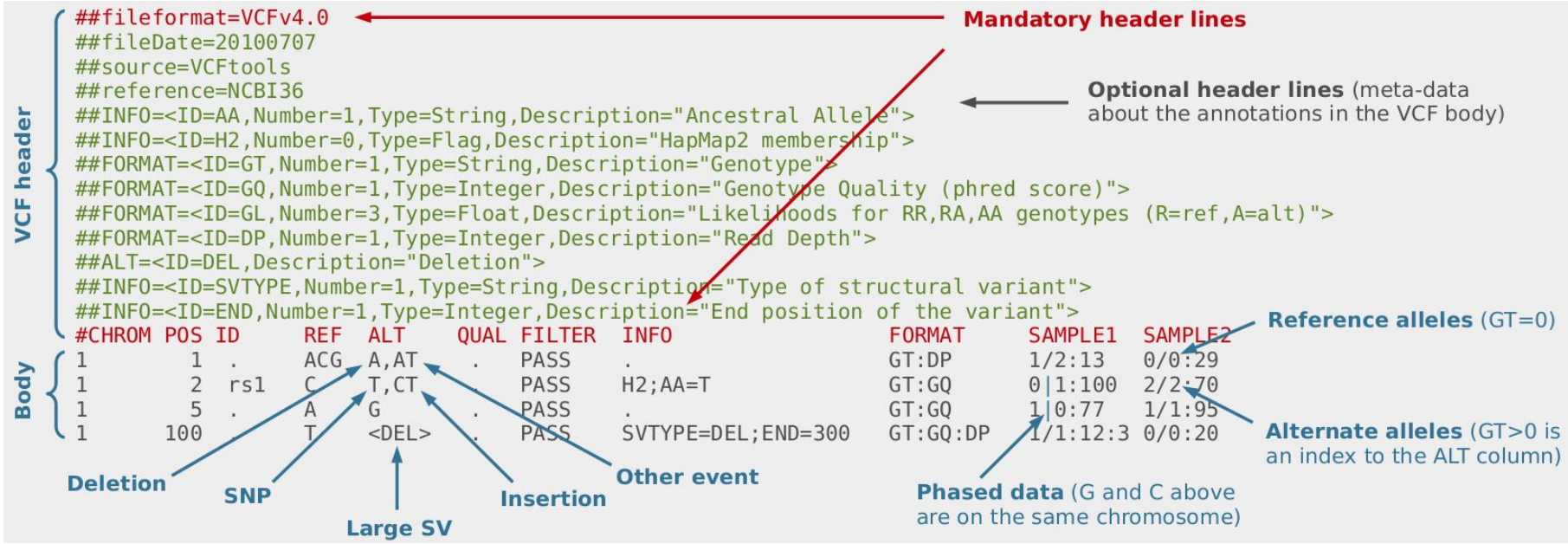
Les régions

Format VCF



Les régions

Format VCF



Pipeline NGS

Données non alignées → file **FastQ**

Alignement

Données alignées → file **Sam/Bam**

Variant calling

Variants → file **VCF**

Bases de données publiques

Différentes bases de données

Généraliste

- **UCSC**
- Ensembl

Cancer spécifique

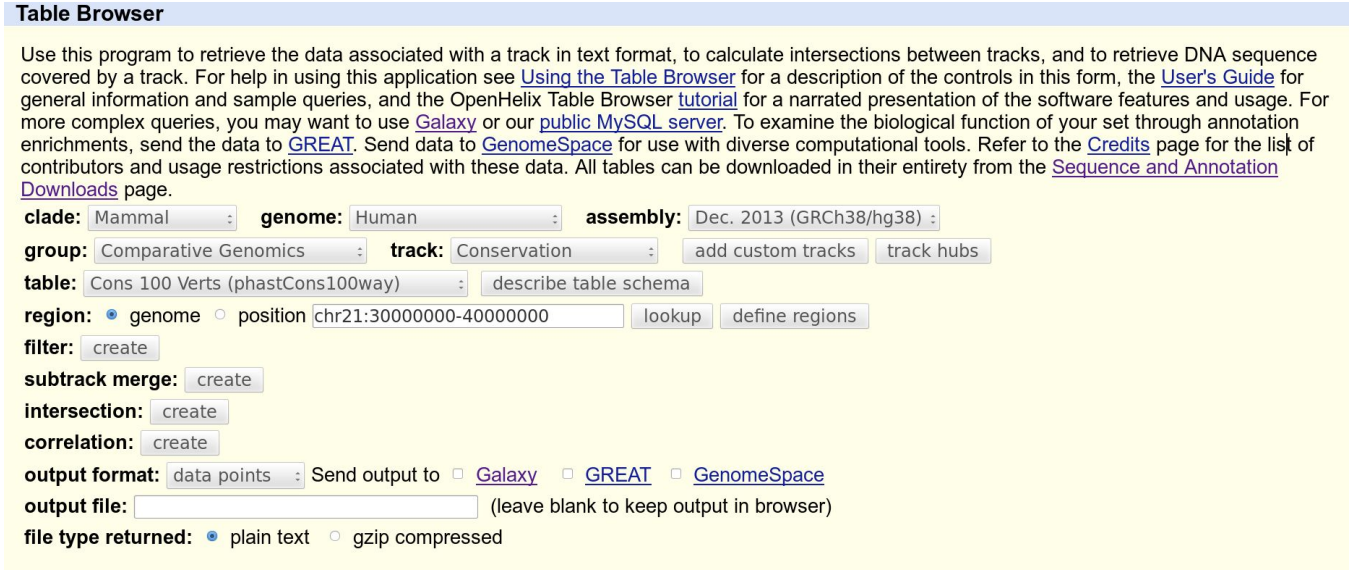
- TCGA
- **ICGC**
- Cosmic

UCSC : Le Golden path

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

UCSC : Table browser

UCSC : Table Browser



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

Cosmic

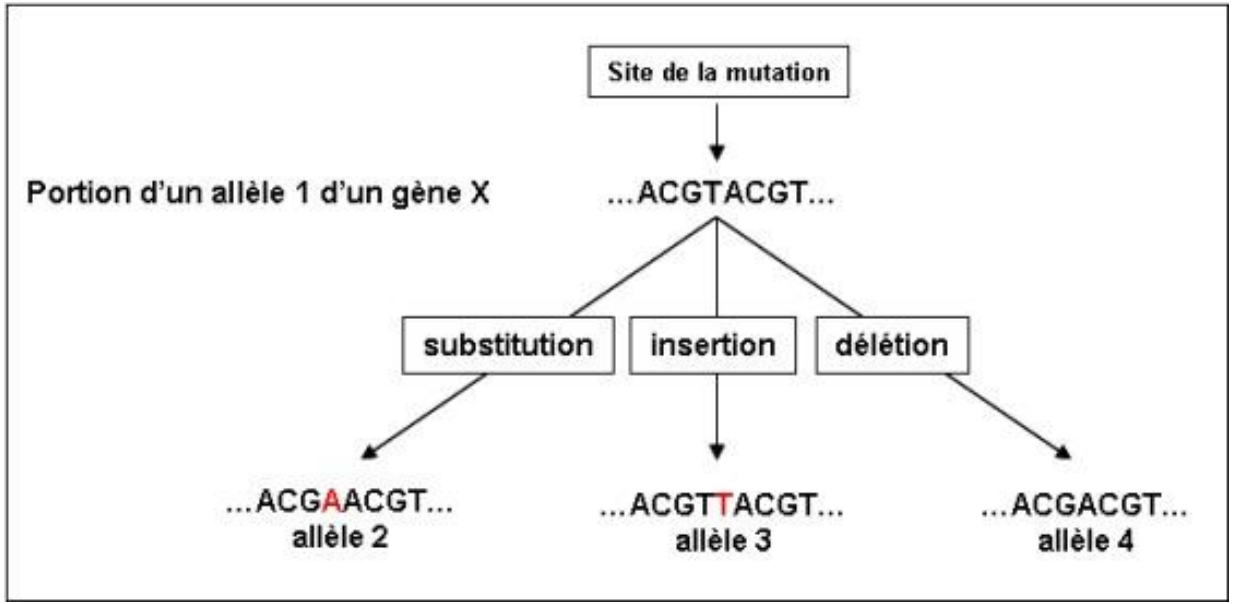
- Mutations trouvées :
  - Par croisement d’article
  - Données de sequençage (TCGA / ICGC)
  - Cancer Gene census : 572 gènes

http://cancer.sanger.ac.uk

Nomenclature des mutations

A l'échelle de la base

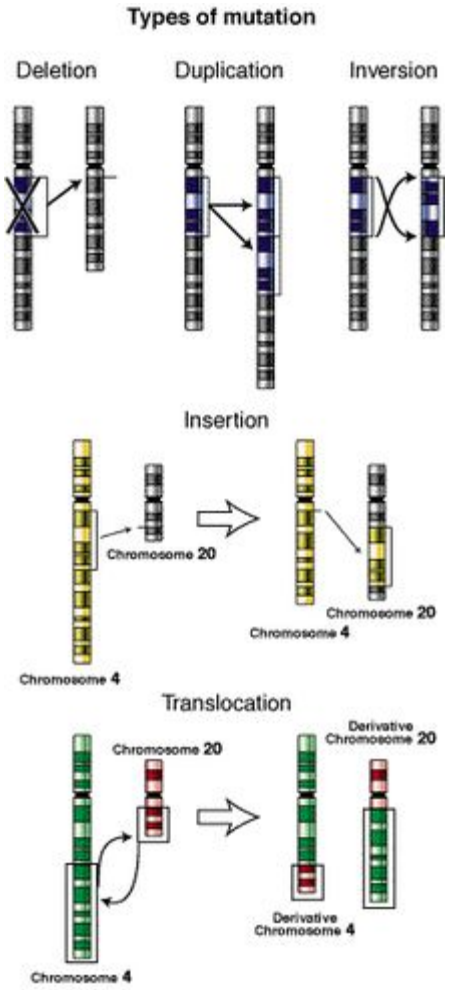
- Substitution
- Délétion
- Insertion



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

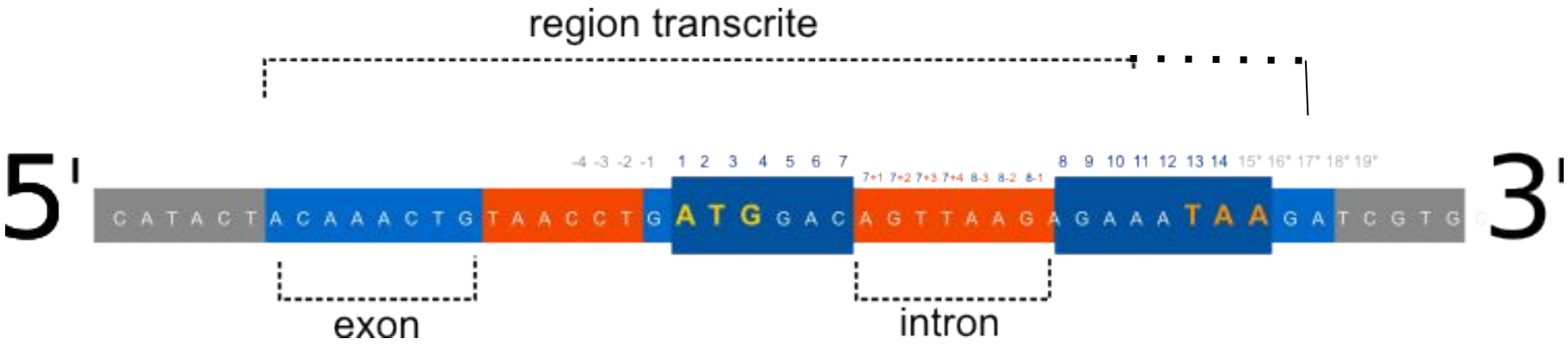
Coordonnée exonique

c.504A>C
c.423DelC
c.32InsA

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

Nomenclature des mutations

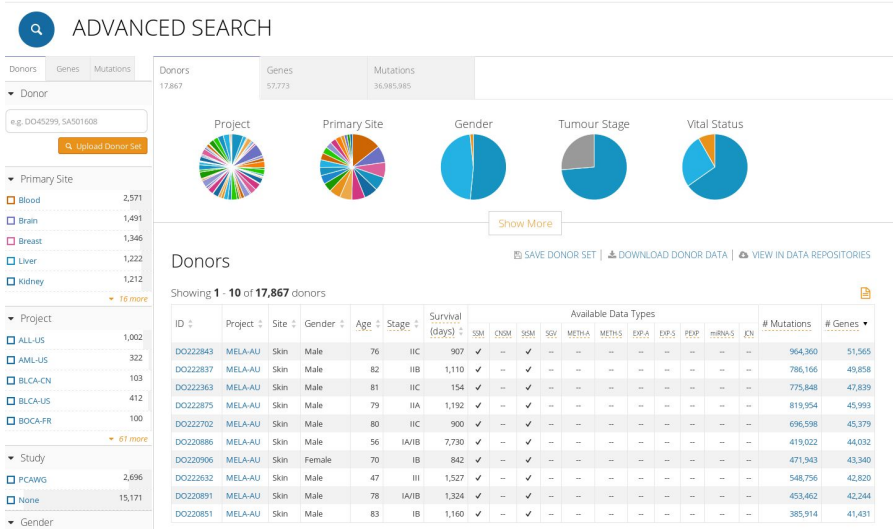
Norme HGVS



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

ICGC

ICGC : International Cancer Genome Consortium



http://icgc.org/

Référénces

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