

# Recherche de variants par analyse DNA-seq

Ghislain Bidaut, Plateforme Cibi, CRCM, Aix-Marseille Université

24/03/2022

# Recherche de variants par Séquencage à haut débit (DNA-seq)

## Applications principales

- **En Recherche:** Recherche de mutations dans des panels *larges* ou des *exomes* complet) à visée de découverte.
- **En Clinique:** Recherche de mutations dans des panels restreints pour le diagnostique.
- Permet l'étude de mutations **constitutionnelles** et **somatiques à faible pourcentages**.
- Un grand nombre de patients peuvent être analysés **simultanément** et rapidement.
- L'analyse bioinformatique devient **partie intégrante** du processus de traitement.

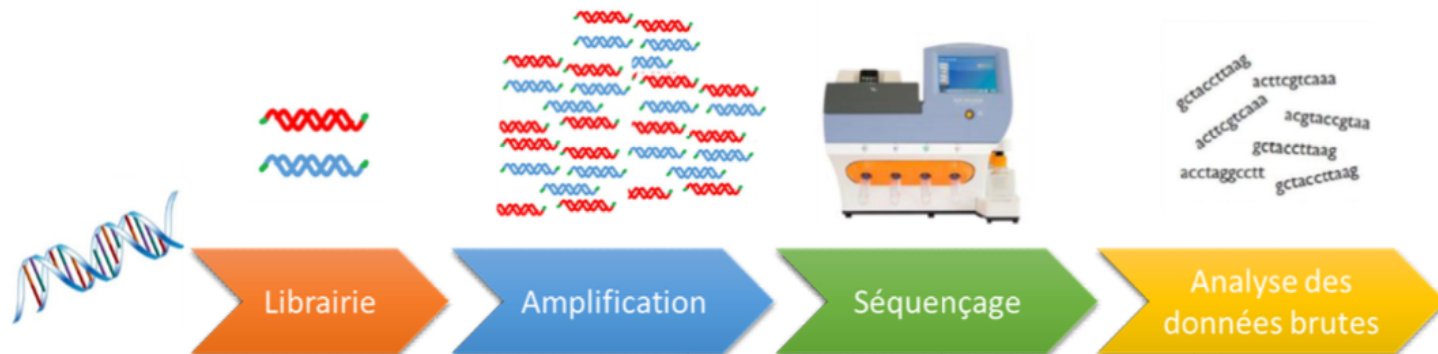
# Principe général du NGS

Le NGS ou *séquençage nouvelle génération*

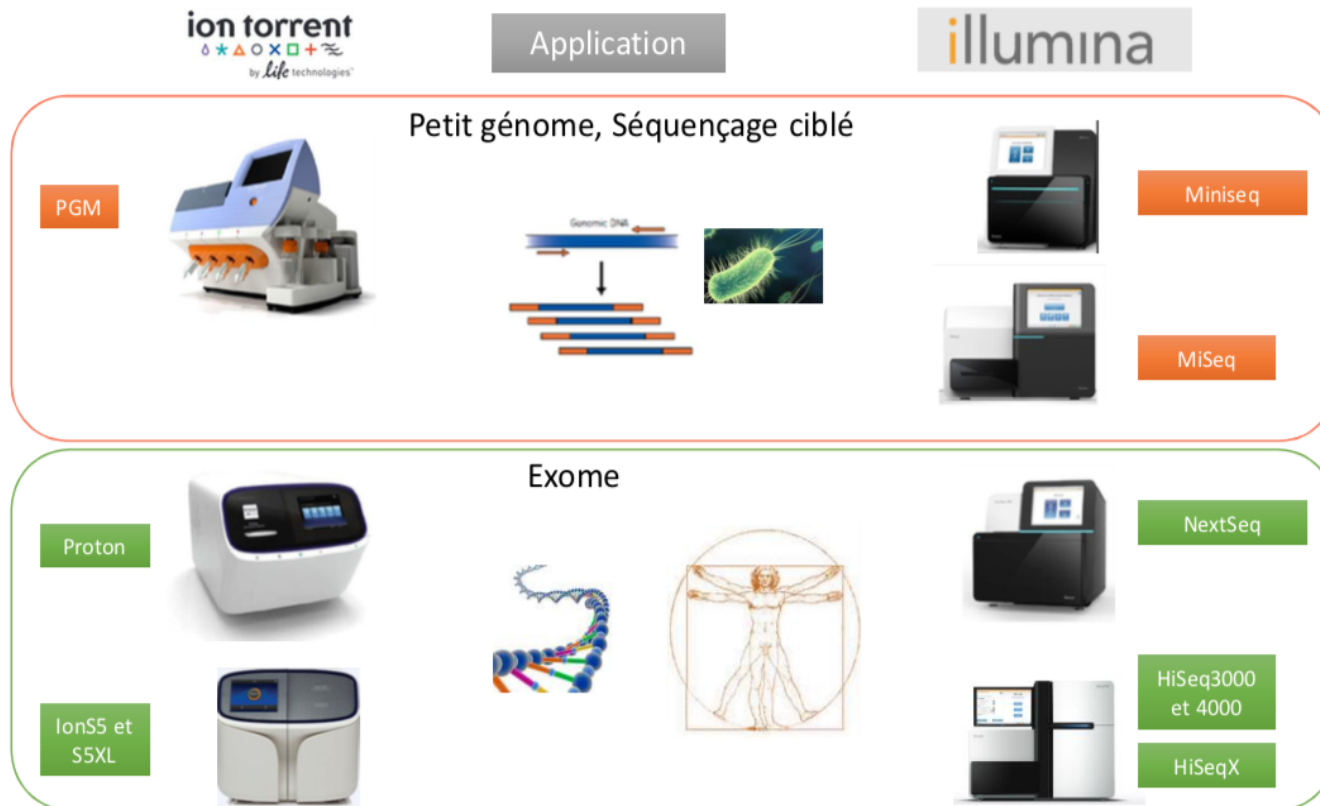
**ADN** : Whole Genome, Whole Exome ou ciblé

**ARN** : RNAseq (expression, transcrits de fusion, découverte de nouveaux transcrits...)

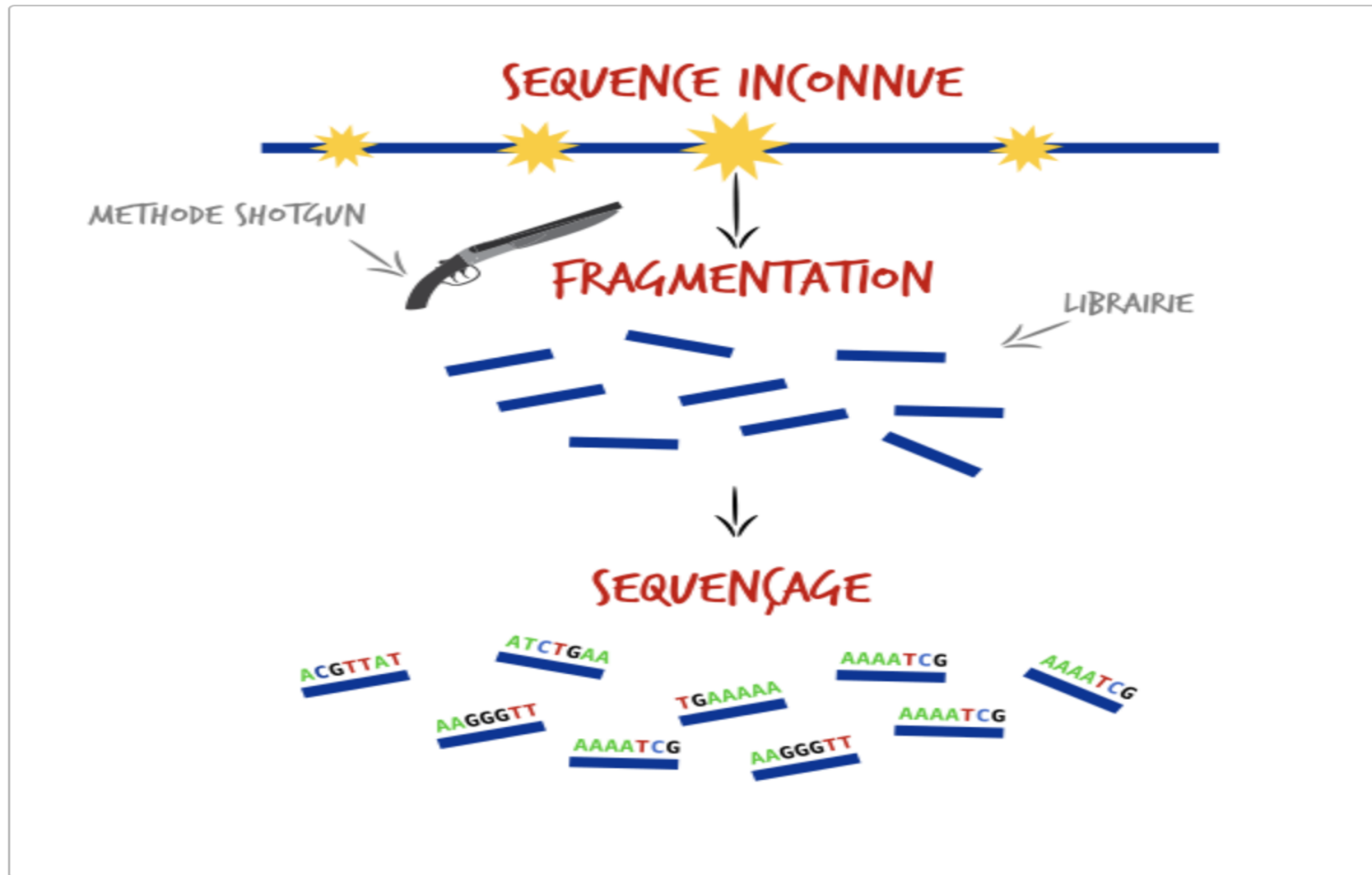
## Process général du NGS



# Echelles en fonction de l'application



# Principe du séquençage *Shotgun*



# Détection de variants par NGS

**But:** recherche de mutations dans des gènes d'intérêt pour poser un diagnostic sur un patient.

Etapes de l'analyse bioinformatique:

- **Contrôle Qualité** sur les données brutes suivi éventuellement d'un **Trimming**
- **Alignement des reads** sur le génome de référence
  - **Alignement principal.**
  - **Ré-alignement local** pour la recherche d'**INDELS**.
- Appel de **variants**
- **Annotation** et production d'un fichier **VCF** et d'un **compte-rendu**

# Formats de fichiers utilisés en génomique

- **FASTA**: Stockage de génomes et séquences de références
- **FASTQ**: Stockage de fragments de lectures issus d'un séquenceur
- **BAM**: Stockage de fragments de lectures alignés (format **binaire**: faire un `samtools view` pour le lire)
- **VCF**: Stockage de variants pour un ou plusieurs échantillons
- **BED**: Stockage de régions génomiques d'intérêt

# Départ: les fichiers issus du séquenceur

## (Fichiers FASTQ)

Ils contiennent les *reads*: petite séquence d'un fragment d'ADN de longueurs plus ou moins fixe.

- **Single-end**
  - Chaque read est indépendant
- **Paired-end**
  - Le séquençage est fait par chaque extrémité de chaque brin. Dans ce cas, les reads sont organisés par paires

```
@HWI-ST865:166:D0C4KACXX:2:1101:1042:1954 1:Y:0:
CNANAAATNAANNNGNNNNNNNNNANNNNAAANNNTNNNNNNNNNTNNTGNNNTTGTTTNNTTGTGGGTTTCTCTGTCCCN
+
#####
@HWI-ST865:166:D0C4KACXX:2:1101:1241:1970 1:N:0:
CCAGCGACACTTGCAGCTTAGGGGCAAGAGGCTCCCACAACACCCTGTGCGATCGGAAGAGCGGTCAGCAGGGATGCCGCGGCC
+
GFFIGIIIFGEHHIJJJIIGGGHIIBD=BFG?EDECC@FGCHC?BCCBB)53(;;B;?8299?#####
```



# Mesure et encodage qualité: le Phred

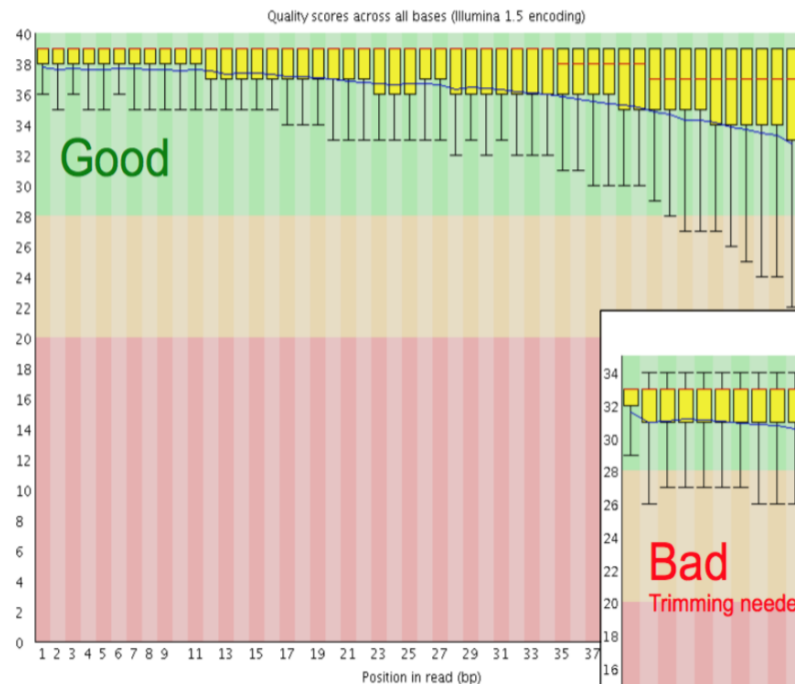
Quelques définitions:

- Valeur de qualité exprimée en *QPhred*
- *QPhred* = probabilité  $p$  d'erreur de mauvaise identification de la base
- $QPhred = -10.\log_{10}(p)$

Exemple:

- Q20 correspond à une probabilité d'erreur de 1%
- Q30 correspond à une probabilité d'erreur de 0,1%

# Contrôle Qualité par FastQC

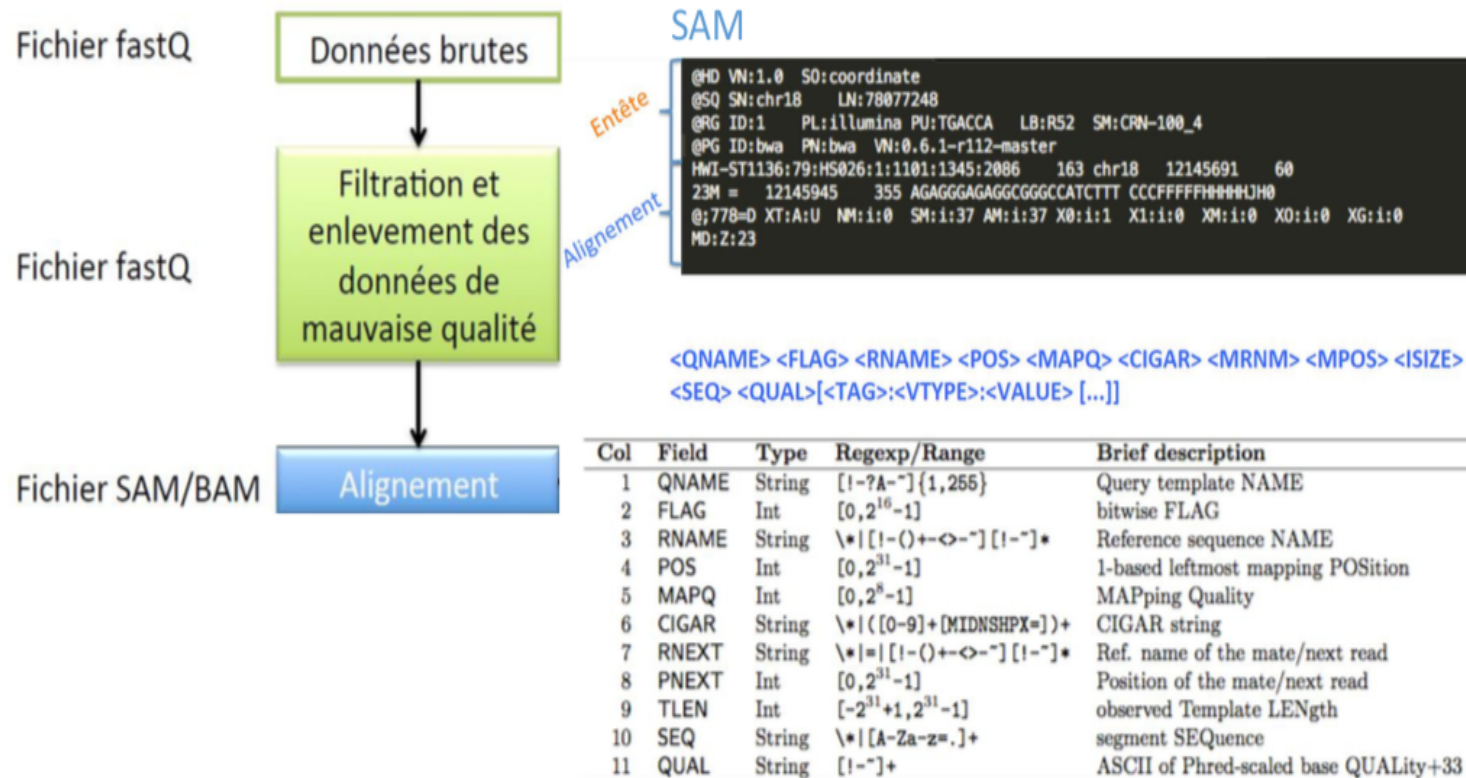


Quality scores across bases

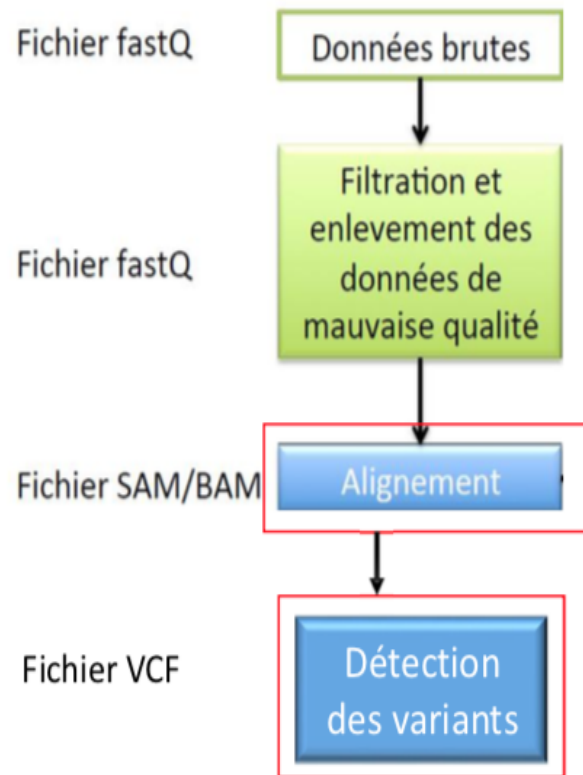
Phred 30 = 1 error / 1000 bases  
Phred 20 = 1 error / 100 bases



# Alignement sur le génome de référence (BWA)

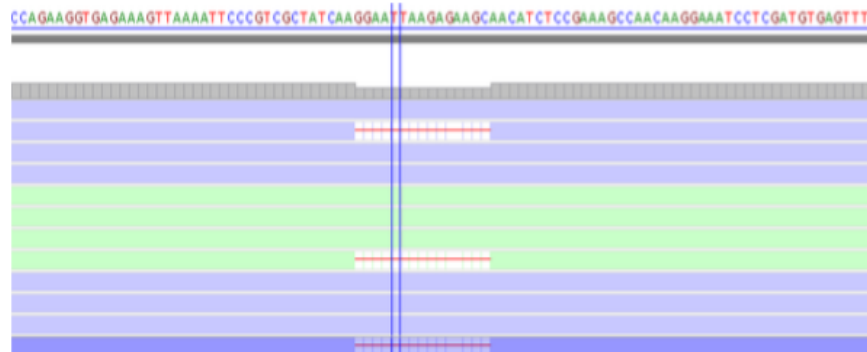


# Détection des variants



BAM

visualisation des alignements via un viewer



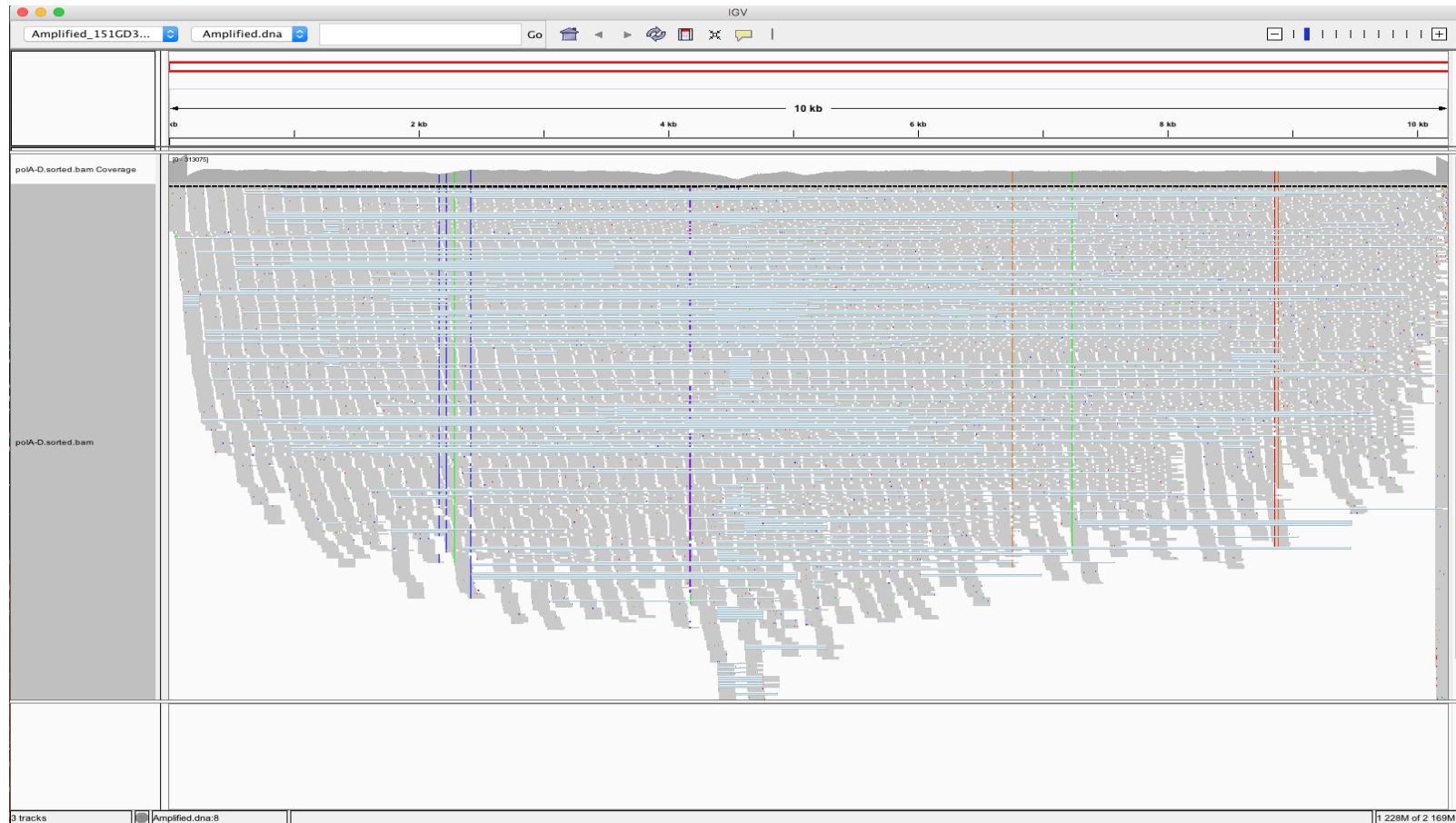
# Production des VCF (Variant Calling Files)

Résultat: exemple d'un fichier VCF annoté

Chr	Ref.NM	Gene	exon	c.(Mutalyzer)	p.(Mutalyzer)	Var.freq	Var.Cov.	Pos.Cov.	Region	Type	Sensitivity	Start_Position	Ref.seq	Var.seq	COSMIC
chr4	NM_000142	FGFR3	9	c.1173G>T	p.(=)	2	2	84	exonic	synonymous	not found	1806154	G	T	NA
chr4	NM_000142	FGFR3	14			99	164	165	exonic	synonymous	not found	1807894	G	A	NA
chr4	NM_006206	PDGFRA	12	c.1701A>G	p.(=)	100	583	583	exonic	synonymous	not found	55141055	A	G	ID=COSM1430082; OCCUREN
chr4	NM_000222	KIT				100	954	954	intronic	NA	not found	55599436	T	C	NA
chr7	NM_005228	EGFR	19	c.2235_2249cp.(Glu746_Ala750del)		76	1115	1466	exonic	nonframeshift	sensible	55242465	GGAATTAA GAGAAGC	-	ID=COSM6223; OCCURENCE
chr7	NM_005228	EGFR	20	c.2361G>A	p.(=)	18	338	1913	exonic	synonymous	not found	55249063	G	A	ID=COSM1451600; OCCUREN
chr7	NM_001127500	MET	2	c.534C>T	p.(=)	60	689	1140	exonic	synonymous	not found	116339672	C	T	ID=COSM1579024; OCCUREN
chr7	NM_001127500	MET				24	133	543	intronic	NA	not found	116421963	TAAAT	ATAAAAC	NA
chr7	NM_001127500	MET				74	401	543	intronic	NA	not found	116421967	T	C	NA
chr10	NM_000141	FGFR2				100	318	318	intronic	NA	not found	123279745	C	T	NA

Reporte les variations nucléotidiques détectées par rapport au génome de référence  
→ mutations et polymorphismes

# Visualisation sous IGV

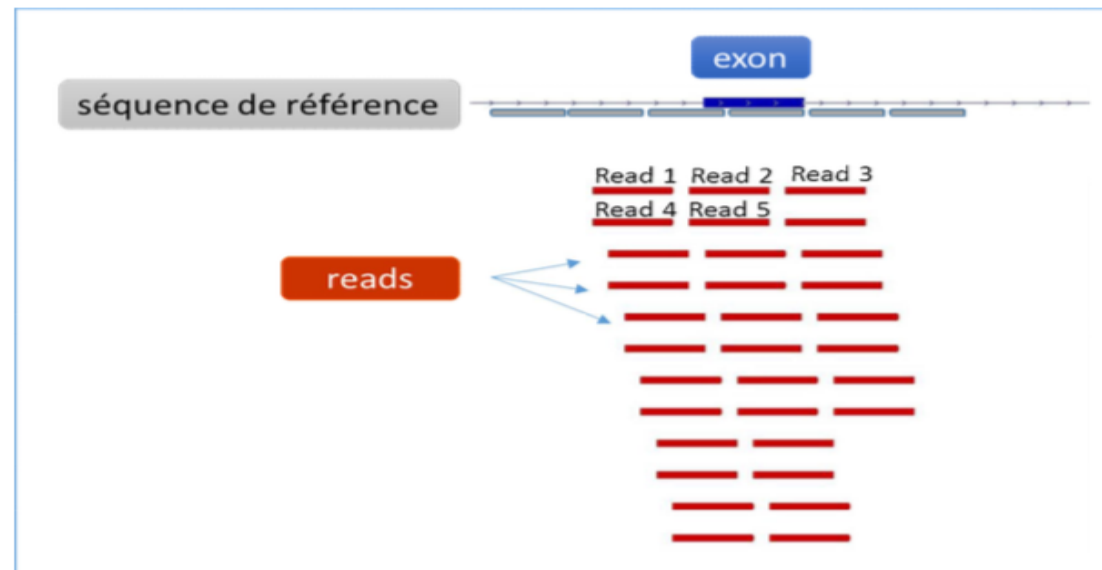


# Quelques définitions: les Reads

## Quelques définitions

**Reads**= lectures= séquences Exemple: ATCGGGTTACCAACCGAAT

**Alignement** des reads (=mots) sur la séquence de référence (=phrase)



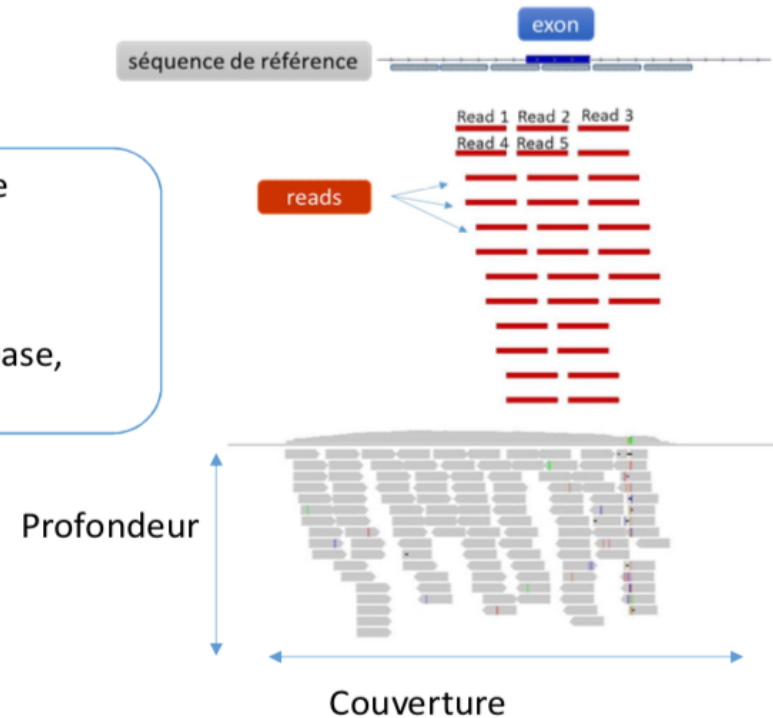
# Couverture et profondeur

## Quelques définitions

### Profondeur-Couverture:

Couverture: zone couverte par au moins une lecture, exprimée en %

Profondeur: nombre de lecture de chaque base, exprimée en X






# Analyse de panels: Exemple du panel INCa






Gène	Exon	Molécule	AMM/essai
<i>ALK</i>	22, 23, 24 et 25	crizotinib et inhibiteur de ALK	AMM
<i>BRAF</i>	11 et 15	vemurafenib et dabrafenib	AMM
<i>EGFR</i>	18, 19, 20 et 21	antiEGFR	AMM
<i>ERBB2</i>	20	trastuzumab et neratinib	Essais cliniques
<i>ERBB4</i>	10 et 12	afatinib	Essais cliniques
<i>FGFR2</i>	8, 14 et 16	Inhibiteurs de FGFR	Essais cliniques
<i>FGFR3</i>	7, 10 et 15	Inhibiteurs de FGFR	Essais cliniques
<i>HRAS</i>	2, 3 et 4	inhibiteurs de MEK	Essais cliniques
<i>KIT</i>	8, 9, 11, 13, 17 et 18	imatinib	AMM
<i>KRAS</i>	2, 3 et 4	panitumumab et cetuximab	AMM
<i>MAP2K1</i>	2	inhibiteurs de MEK	Essais cliniques
<i>MET</i>	2, 14, 15, 16, 17, 18, 19 et 20	crizotinib	Essais cliniques
<i>NRAS</i>	2, 3 et 4	panitumumab, inh MEK, inh BRAF	AMM et essais cliniques
<i>PDGFRA</i>	12, 14 et 18	imatinib	AMM
<i>PIK3CA</i>	10 et 21	Inh PI3K	Essais cliniques

# Du prélèvement au compte-rendu scientifique

Extraction de l'ADN (FFPE, congélation ou plasma)

Analyse moléculaire par NGS: synthèse librairie, amplification clonale et séquençage



Patient 1		KRAS, BRAF, NRAS, KIT, EGFR...
Patient 2		KRAS, BRAF, NRAS, KIT, EGFR...
...		
Patient 24		KRAS, BRAF, NRAS, KIT, EGFR...
Témoin positif		KRAS, BRAF, NRAS, KIT, EGFR...
Témoin négatif		KRAS, BRAF, NRAS, KIT, EGFR...

Gene	Start	Position	Exon	c.	p.	Var. freq.
PIK3CA	178952085	21		c.3140A>G	p.(His1047Arg)	25
EGFR	55241707	18		c.2155G>A	p.(Gly719Ser)	16,65
EGFR	55242465	19		c.2235_2249del	p.(Glu746_Ala750del)	6
EGFR	55249071	20		c.2369C>T	p.(Thr790Met)	6
EGFR	55259515	21		c.2573T>G	p.(Leu858Arg)	6
EGFR	55259524	21		c.2582T>A	p.(Leu861Gln)	6
BRAF	140453136	15		c.1799T>A	p.(Val600Glu)	34
KRAS	25378562	4		c.436G>A	p.(Ala146Thr)	8
KRAS	25398281	2		c.38G>A	p.(Gly13Asp)	8
KRAS	25398284	2		c.35G>A	p.(Gly12Asp)	8
ERBB2	37881082	20		c.2411G>A	p.(Gly804Asp)	8

Interprétation de l'analyse:

- ✓ Vérification des témoins, vérification de l'absence de contamination
- ✓ Analyse des ADN patient (+/- technique complémentaire)

Réalisation du **Compte-Rendu** biologique

# Alignement par BWA

**Référence:** Li *et al*: Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009 Jul 15; 25(14): 1754–1760.

**BWA** (Burrows-Wheeler Alignment tool) a été spécialement conçu pour l'alignement de millions de séquences peu divergentes d'un génome de référence.

Il est basé sur la *Transformée Burrows-Wheeler* associé à un algorithme de tri par arbre. Il permet l'alignement de *reads* relativement longs pour lesquels il existe des seuils (gap) en cas de présence d'INDELS.

Il utilise une quantité relativement faible de mémoire et est parallélisable, pour exploiter les architectures multi-coeurs.

# GATK (Genome Analysis Toolkit)

**Référence:** Van der Auwera GA & O'Connor BD. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition). *O'Reilly Media*.

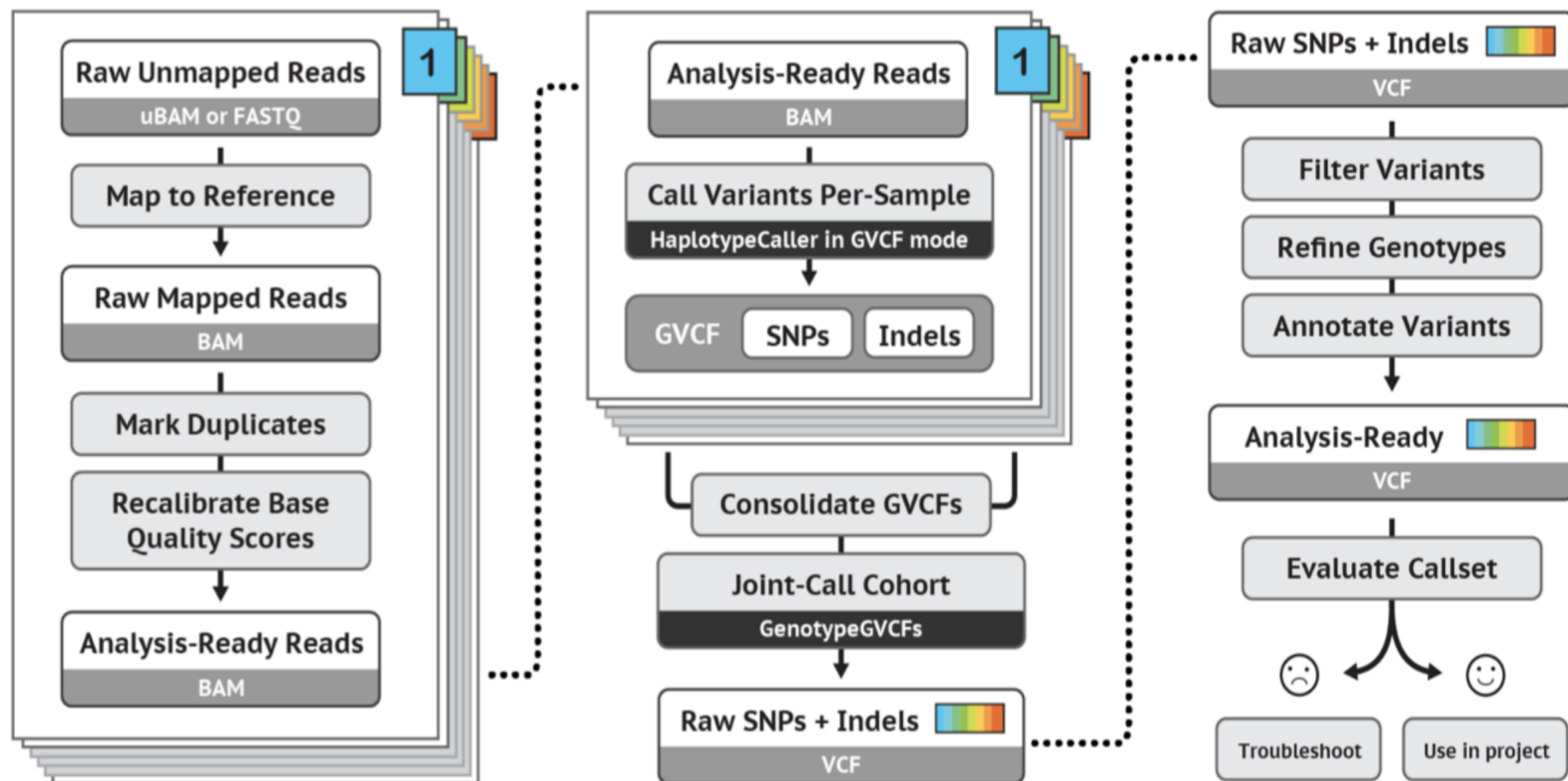
**Tutoriel:** Van der Auwera GA *et al.* (2013). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics*, 43:11.10.1-11.10.33. DOI: 10.1002/0471250953.bi1110s43.

**GATK** est une suite d'outil qui permettent

- Le marquage des réplicats PCR dans les reads (outil *picard*)
- La recalibration des valeurs qualité des bases des reads
- Le réalignement pour l'appel d'INDELS
- L'appel de variants (SNPs+INDELS)
- La recalibration du score des variants et leur filtrage

# Utilisation de GATK pour la détection de variants

Pipeline issu des bonnes pratiques définies par le Broad Institute.



# Annotation de variants

# Format d'échange de données: VCF

Le format de fichier VCF (**Variant Call Format**) est typiquement utilisé pour l'échange de données. (Nous en sommes à la version 4.3) (<https://samtools.github.io/hts-specs/VCFv4.3.pdf>).

Ce format a été développé dans le cadre de grands projets génomiques (1000 Genome Project). Certains sites ont développé leur **propre spécification** du format VCF.

# Structure d'un fichier VCF

Il est divisé en 2 parties:

- **Une en-tête** (marquée avec des **##**) contenant les métadonnées:
  - Génome
  - Logiciel utilisé pour l'appel de variants
  - Définition de plusieurs variables qualité (DP, Génotype)
  - Définitions des entrées **FILTER**, **INFO** et **FORMAT**
- **(Optionnel) La liste des régions** du génome analysé (Format **gVCF**)
- **La liste des variants** contenant:
  - Chromosome
  - Coordonnées chromosomiques
  - ID, REF=allèle de référence, ALT=allèle mutant
  - QUAL= qualité du variant), FILTER (PASS ou FAILED)
  - INFO= Informations définies dans l'en-tête
  - FORMAT= Information génomiques



# En-tête VCF (Metadonnées) Précédées par ##

- Entrées **FILTER**: Descripton du filtre utilisé pour le contrôle qualité
- Entrées **INFO**: Informations sur l'ensemble des échantillons
- Entrées **FORMAT**: Informations spécifique à chaque échantillon

# VCF en pratique

C'est un fichier **texte délimité par des tabulations**. On peut donc l'ouvrir avec M\$ *Excel* mais en général, nous travaillons sur ce type de fichiers avec des **outils dédiés** ou des scripts "maison" du fait de leur volume.

Voici quelques outils très intéressants pour la manipulation de fichiers VCF:

- VCFTools (<https://vcftools.github.io/index.html>)
- BCFTools
- VT...

Il existe également un format étendu **genomic VCF** (gVCF), utilisé avec **GATK**, qui contient des informations sur les blocs qui correspondent à la référence et à leur qualité.

# En quoi consiste l'annotation de variants ?

L'**annotation** des variant consiste à collecter de l'**information biologique** correspondant aux variant que l'on analyse.

Nous allons pouvoir savoir dans quel gène/exon se trouve la mutation et quel est son impact au regard de la protéine correspondante.

Nous allons pouvoir également trouver leur **fréquence allélique** dans les populations, que ce soit parmi la population générale ou parmi des patients atteints de cancers ou maladies génétiques rares.

Il est également possible de contribuer nos découvertes dans des bases de données publiques.

Bien qu'il soit possible de parcourir manuellement les bases de données publiques, il est plus simple d'interroger directement les bases publiques avec **des programmes spécialisés dans l'annotation**.

Par Exemple: **Annovar, snpEff**.

# Annovar

**Annovar** est un utilitaire d'annotations de variants. Il fonctionne à partir de génomes divers (**hg38**, **hg19**, souris, drosophile, levure, etc...)

Il permet de faire des annotations au niveau des **gènes** et donc de retrouver les gènes à partir des bases de données **RefSeq**, **Ensembl**, etc...

Il permet aussi d'annoter des **Régions** ainsi que de construire des **filtres** basés sur le contenu des bases de données.

Les principales bases de données qu'il permet d'utiliser pour l'annotation sont:

(voir <https://annovar.openbioinformatics.org/en/latest/user-guide/filter/#summary-of-databases>)

- 1000 genomes
- GnomAD
- RefSeq...

# SnEff

SnEff permet de **prédire l'effet d'un variant** sur les gènes ou protéines (changement dans les acides aminés par exemple).

**Référence:** Cingolani P *et al.*, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3." *Fly* (Austin). 2012 Apr-Jun;6(2):80-92. PMID: 22728672

- Il supporte 38 000+ génomes
- Il supporte le format standard d'annotation **ANN**
- Notation **HGVS** (<https://varnomen.hgvs.org/bg-material/simple/>)
- Support de Sequence Ontology
- Compatible **GATK** en natif
- 1000 Genomes Project
- GnomAD
- COSMIC...

# Base de données: *GnomAD*

La base de données **Genome Aggregation Database** est une base développée à l'intention de la communauté scientifique et médicale pour l'annotation de séquences humaines.

Elle contient les **fréquences alléliques** de variants structuraux dans différentes populations pour plus de 76000 génomes (pour hg38) et 10000 génomes (pour hg37) ayant été séquencés dans le cadre d'analyses de maladies rares et de cancers.

**Référence:** Karczewski, K.J., Francioli, L.C., Tiao, G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443 (2020).

<https://doi.org/10.1038/s41586-020-2308-7>

# Base de données: *1000 genomes project*

C'est un catalogue de variations génétiques communes (existantes dans au moins 1% de la population) obtenues à partir de **donneurs sains**, constituant une ressource de référence utilisée par la communauté biomédicale.

Ce catalogue est accessible à travers l'**International Genome Sample Ressource**.

- Il est continuellement maintenu et mis à jour avec les dernières versions du génome humain et des données provenant de nouvelles populations.
- A ce jour, il contient des variants pour **2504 individus** obtenus dans 26 populations.
- Il n'y a aucune donnée phénotypique ou médicales associée.

**Référence:** A global reference for human genetic variation, The 1000 Genomes Project Consortium, *Nature* 526, 68-74 (01 October 2015) [doi:10.1038/nature15393](https://doi.org/10.1038/nature15393).

# Base de données: The Catalog of Somatic Mutations in Cancer

URL: <https://cancer.sanger.ac.uk/cosmic> Cette base constitue une ressource pour l'exploration de l'impact des mutations somatiques dans les cancers.

Il contient des données traitées manuellement associées à des panels de gènes ciblés. Elles sont disponibles sur les versions hg37 et hg38 du génome humain.

Les données consistent en un **catalogue de mutations liées à 1.4 millions de tumeurs** obtenues à partir de 26000 publications. Les données sont associées à des meta-données (facteurs environnementaux et historique des patients).

**Référence:** COSMIC: the Catalogue Of Somatic Mutations In Cancer. John G Tate et al. *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D941–D947, <https://doi.org/10.1093/nar/gky1015>



# Rappel des étapes bioinformatiques

- Contrôle Qualité (**FASTQC**)
- Alignement sur le génome de référence (**BWA**)
- Trimming des séquences adaptatrices (**Trimmomatic**)
- Ré-alignement (**GATK**)
- Détection des mutations (**GATK**)
- Annotation des variants (**SnpEff**)
- Visualisation des données (Read, SNPs) (**IGV - Integrative Genomics Viewer**)

# Rappels sur les extensions de fichiers

- Fichiers de séquences brutes: **.fastq** (Compressé: **.fastq.gz**)
- Fichiers de séquences alignées **.BAM**
- Index de fichiers de séquences alignées **.BAI**
- Génome complet au format FASTA: **.fa**
- Fichiers listant les mutations/Indels: **.VCF** ou **.txt**

# Licence



Ce(tte) œuvre est mise à disposition selon les termes de la [Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International](#).