



Analyse de données de séquençage DNA-seq

Recherche et interpretation de variants en génétique somatique et constitutionnelle

Introduction

Qu'est ce que la bioinformatique ?

- Apparition en 1970: B Hesper et P Hogeweg, « Bioinformatica: een werkconcept », *Kameleon*, vol. 1, no 6, 1970, p. 28–29

La bio-informatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation informatique de l'information biologique. Plusieurs champs d'application ou sous-disciplines de la bio-informatique se sont constitués (*Wikipedia*):

- La bio-informatique des séquences
- La bio-informatique structurale
- La bio informatique des réseaux
- La bio-informatique statistique et des populations

Projet du séquençage du Génome Humain

- **Idée** lancée en 1985 par 3 scientifiques, **Renatto Dulbecco**, **Robert Sinsheimer** (directeur de UCSC) et **Charles DeLisi**, qui financera le projet (Directeur dept. de biologie du *Département de l'Energie US*)
- **Séquençage** lancé en 1988 par le *National Research Council*. En suisse est créé HUGO (*Human Genome Organisation*) pour la coordination.
- En 1998, Craig Venter, crée *Celera Genomics* avec pour objectif le **séquençage en 3 ans** par séquençage *Shotgun* et le brevetage du génome (!).
- En 2000, la complétion du séquençage est annoncé pour le Consortium et Celera Genomics (match nul) par le président B. Clinton. Coût: \$3B.
- *Celera Genomics* avouera avoir utilisé les données du Consortium pour son propre assemblage, mais reproduira un séquençage *de Novo* 3 ans plus tard...
- Publication des séquences brutes en 2001 et des séquences finales en 2004.

Différentes technologies de séquençage

<https://planet-vie.ens.fr/thematiques/manipulations-en-laboratoire/la-revolution-de-la-genomique-les-nouvelles-methodes-de>

- **Séquençage Sanger:** faible débit, utilisée pour le séquençage du génome humain
- **Séquençage de nouvelle génération:** adapté au séquençage massif d'un grand nombre de génome pour étudier **les variations génétiques** (GWAS). Le leader du marché aujourd'hui est *Illumina*.

Déroulé d'un séquençage

Le principe d'un séquençage NGS consiste à:

- Créer une *bibliothèque* de fragments d'ADN (par fragmentation enzymatique et mécanique de l'ADN génomique).
- Relier ces fragments à des *adaptateurs*, des petites molécules d'ADN de séquences connues nécessaires au séquenceur.
- Une sélection de la taille minimum et maximum des fragments est effectuée pour des raisons techniques, notamment l'amplification PCR.
- Faire une amplification PCR des fragments.
- Faire le séquençage.

Le séquençage à haut débit et ses applications en oncologie

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 en génétique **constitutionnelles** et **somatiques à faible pourcentages**.
- Grâce au NGS, un grand nombre de patients peuvent être analysés **simultanément** et rapidement.
- L'analyse bioinformatique devient **partie intégrante** du processus de traitement.

Enjeux

- **Explosion de la quantité de données générée**
 - **Volume massif** : Le séquençage à haut débit génère des téraoctets de données brutes, nécessitant des **infrastructures** de stockage et de **gestion adaptées** (Cluster de calcul - Exemple: **IFB**, Cloud de stockage).
 - **Coût du stockage et du calcul** : Le traitement et l'analyse des données nécessitent des ressources informatiques puissantes qui ont un coût.
 - **Archivage et accessibilité** : Les plateformes comme ENA, SRA et EGA doivent assurer une **conservation sécurisée** des données sur le **long terme**.
- **Complexité des approches analytiques**
 - **Prétraitement des données** : Alignement des lectures (BWA, Bowtie2), suppression des erreurs de séquençage, normalisation des données d'expression, etc.
 - **Analyse de variants** : Identification des SNPs, indels et mutations structurales à partir des reads bruts, avec des outils comme GATK ou DeepVariant.
 - **Annotation des variants** : L'association entre variations génétiques et pathologies reste un défi, nécessitant des bases de données fonctionnelles et cliniques (gnomAD, OMIM) (Outil **VEP**: *Variant Effect Predictor*:).

Norme COFRAC et accréditation en laboratoire

- Le **COFRAC** (Comité Français d'Accréditation) est l'unique instance en France chargée de délivrer les accréditations selon les normes internationales.
- En biologie médicale, l'accréditation COFRAC est **obligatoire** pour les laboratoires réalisant des examens de diagnostic.
- Elle garantit la **conformité aux bonnes pratiques**, la **traçabilité**, et le **contrôle qualité** des analyses.
- Elle est souvent fondée sur la norme **NF EN ISO 15189**.
- Le processus d'accréditation comprend des audits réguliers et une déclaration de portée.

NF EN ISO 15189 : exigences pour la qualité des laboratoires

- La norme **NF EN ISO 15189** définit les exigences en termes de **qualité et de compétence technique** pour les laboratoires de biologie médicale.
- Elle couvre :
 - Le **management de la qualité** (documentation, revue de direction, audits internes)
 - Les **exigences techniques** (personnel, équipements, méthodes, validation)
 - La **traçabilité des résultats**, gestion des échantillons et des données.
- Elle place l'**amélioration continue** au cœur du processus qualité.
- L'accréditation selon cette norme est indispensable pour prouver la **fiabilité** et la **reproductibilité** des analyses.

ISO/IEC 27001 : sécurité de l'information

- La norme **ISO/IEC 27001** s'applique à la **gestion de la sécurité de l'information**.
- Elle définit les exigences pour mettre en œuvre un **SMSI** (Système de Management de la Sécurité de l'Information).
- Objectif : garantir la **confidentialité, intégrité et disponibilité** des données (par exemple, données patients, résultats NGS).
- Impliquée dans la protection des systèmes informatiques, des accès, des sauvegardes et des risques.
- De plus en plus importante en **bioinformatique** et pour les infrastructures manipulant des **données de santé sensibles**.

Type d'application du NGS dans le diagnostique

- **Diagnostic des maladies génétiques**

- **Exome sequencing** (WES, *Whole Exome Sequencing*) : Permet d'identifier des mutations dans les régions codantes du génome associées à des maladies rares (ex. : mucoviscidose, dystrophie musculaire, syndrome de Marfan).
- **Génome entier** (WGS, *Whole Genome Sequencing*) : Détecte des variants structurels complexes, tels que des délétions ou duplications impliquées dans des pathologies comme les syndromes de déficience intellectuelle.

- **Oncogénomique et médecine personnalisée**

- **Panel de gènes ciblés** : Analyse de gènes spécifiques liés au cancer (ex. : BRCA1/2 pour le cancer du sein et de l'ovaire, EGFR pour le cancer du poumon).
- **RNA-seq en oncologie** : Détecte des altérations de l'expression génique et identifie des translocations fusionnelles (ex. : BCR-ABL dans la leucémie myéloïde chronique).
- **Suivi des tumeurs et biopsie liquide** : Détection de l'ADN tumoral circulant (ctDNA) pour surveiller l'évolution des cancers et la résistance aux traitements.

- **Immunogénomique et médecine de transplantation**

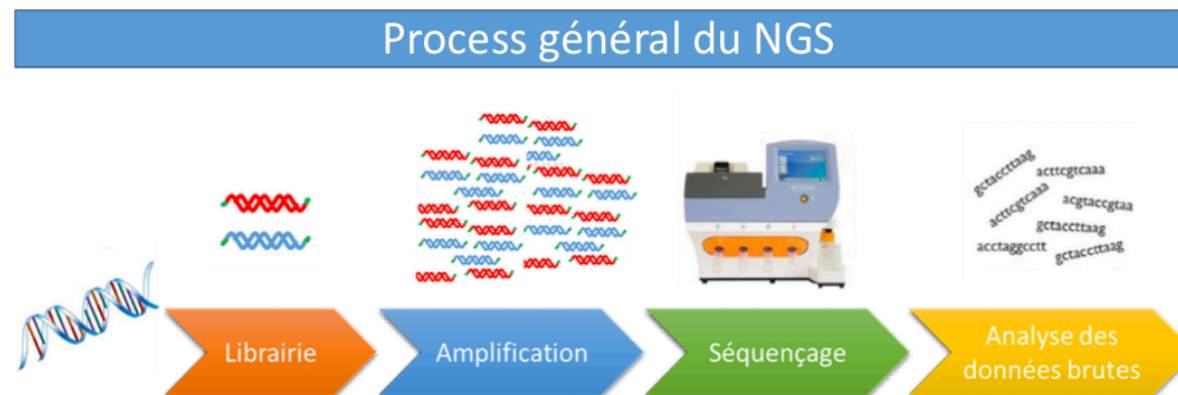
- **Typage HLA par NGS** : Analyse des gènes du complexe majeur d'histocompatibilité (HLA) pour optimiser la compatibilité entre donneur et receveur lors des greffes d'organes et de moelle osseuse.
- **Répertoire des récepteurs T et B (TCR/BCR-seq)** : Analyse de la diversité des lymphocytes pour évaluer la réponse immunitaire et diagnostiquer des maladies auto-immunes ou hématologiques.

Principe général du NGS

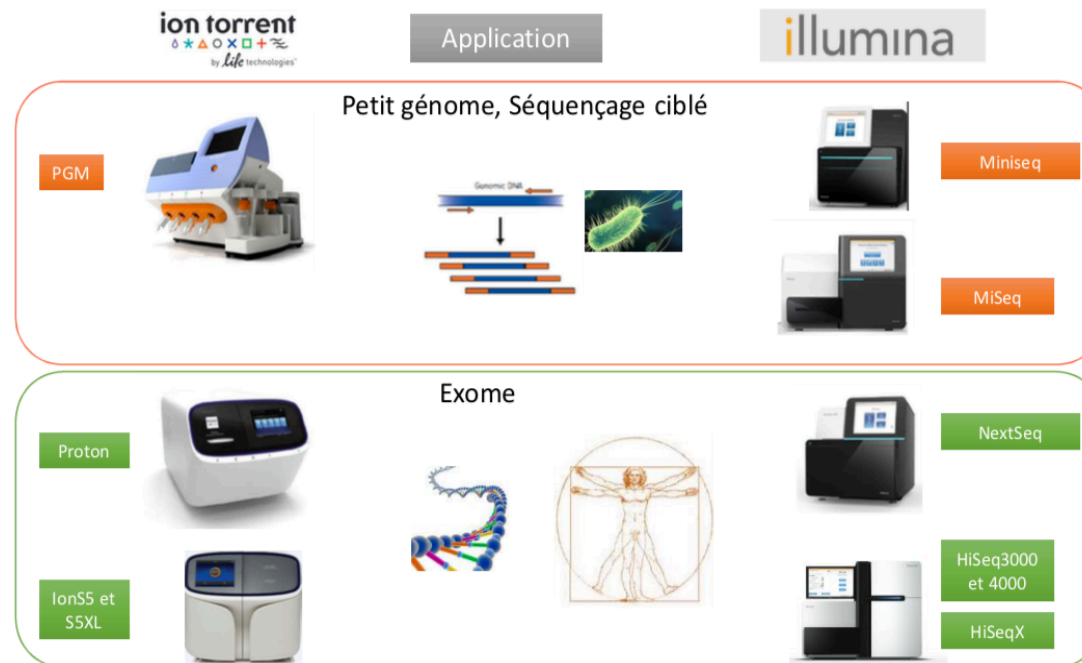
Le NGS ou *séquençage nouvelle génération*

ADN : Whole Genome, Whole Exome ou ciblé

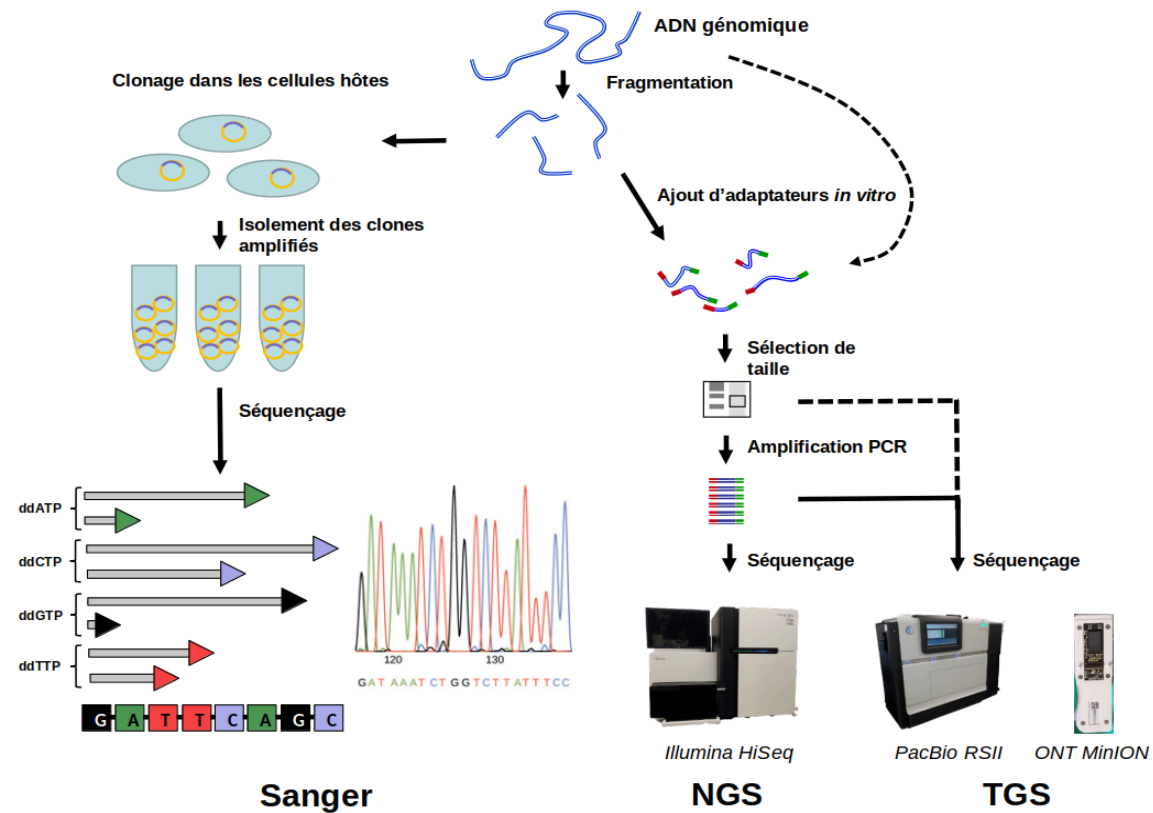
ARN : RNAseq (expression, transcrits de fusion, découverte de nouveaux transcrits...)



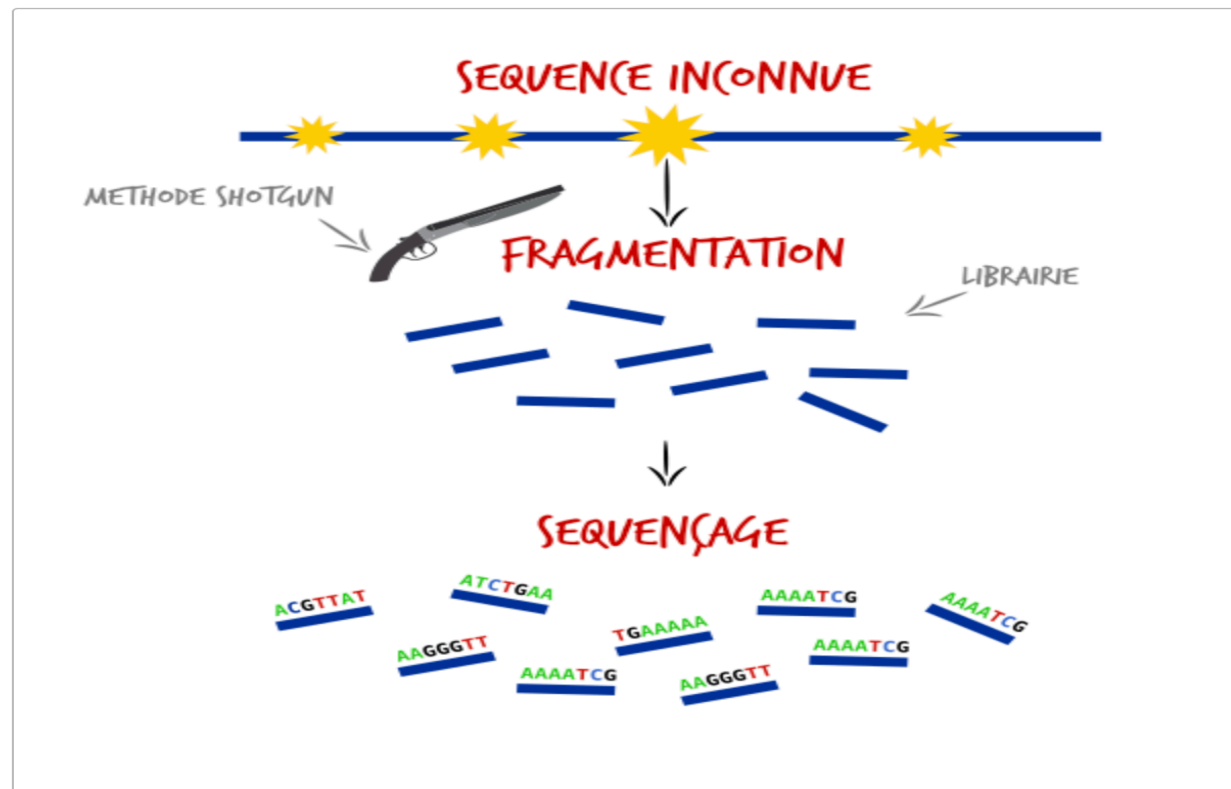
Echelles en fonction de l'application



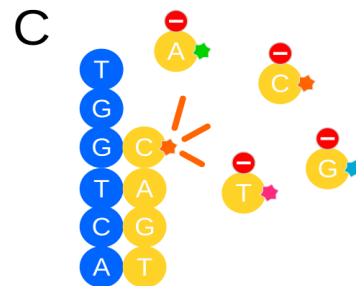
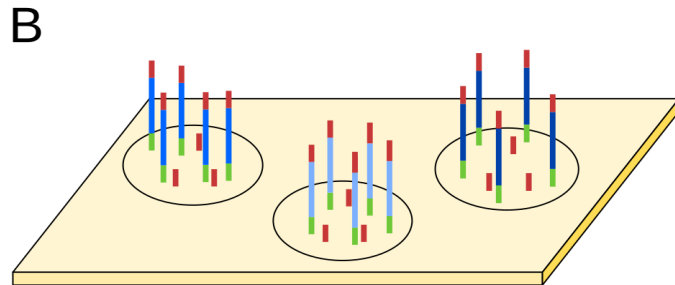
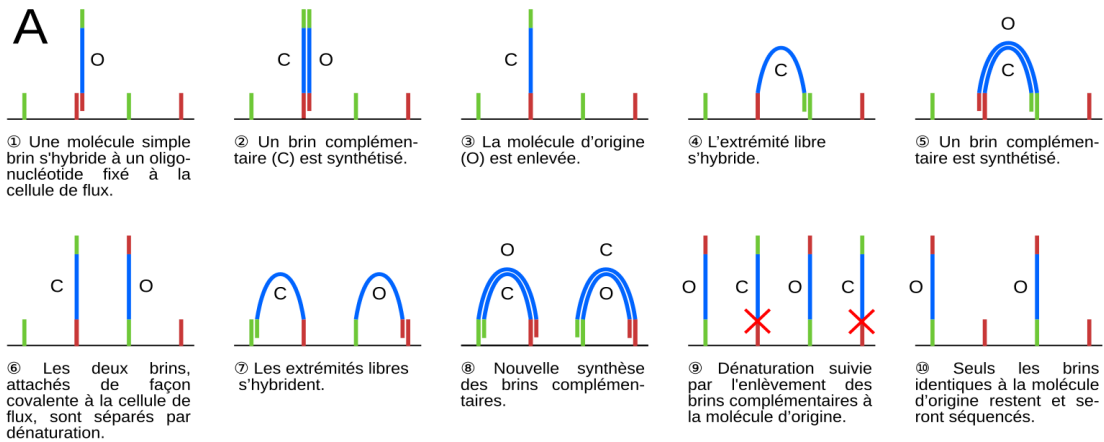
Principe général du séquençage



Principe du séquençage *Shotgun*



Génération des séquences (Illumina)



Séquençage par Illumina - principe

- Hybridation d'un brin sur un oligonucléotide attaché à la *FlowCell*
- Un brin complémentaire est synthétisé.
- La molécule d'origine est enlevée et la molécule libre s'hybride *en pont*.
- Un brin supplémentaire est synthétisé de nouveau.
- Les brins complémentaires à la cellule d'origine sont lavés et il ne reste que plusieurs copies d'une même brin (*clusters*)
- Il reste à séquencer les brins présents: Lors de cette étape, le nucléotide incorporé est identifié grâce à un *groupe fluorescent* identifié par laser, permettant d'enregistrer la séquence de manière informatique.
- <https://www.youtube.com/watch?v=CZeN-IgjYCo>
- <https://www.youtube.com/watch?v=WneZp3fSJlk>

Figure B: Schéma représentant les *clusters* sur une *FlowCell*

Figure C: Réaction de séquençage

Enrichissement

Il est possible de créer une librairie **enrichie en régions d'intérêt**, par exemple pour séquencer uniquement les régions codantes du génome:

- **Par capture:** Il est possible de concevoir des sondes qui vont s'attacher à l'ADN d'intérêt, elle même étant liées à des molécules de *biotines* attachées à des billes magnétiques, qui sont conservées après lavage.
- **par amplicon:** Il est possible de faire une amplification sélectionnée par des amorces PCR choisies.

Importance du choix de la méthode de préparation des librairies

La préparation des librairies est une **étape critique** qui influence directement la qualité et la fiabilité des résultats du séquençage. Plusieurs paramètres doivent être considérés :

1. Type d'échantillon et application

- **ADN génomique (gDNA)** : Utilisé pour le séquençage du génome entier, des exomes ou des panels ciblés.
- **ARN (RNA-seq)** : Essentiel pour l'analyse de l'expression génique et l'identification de fusions oncogéniques.
- **ADN dégradé (FFPE, biopsies liquides)** : Nécessite des protocoles adaptés pour limiter les biais d'amplification.

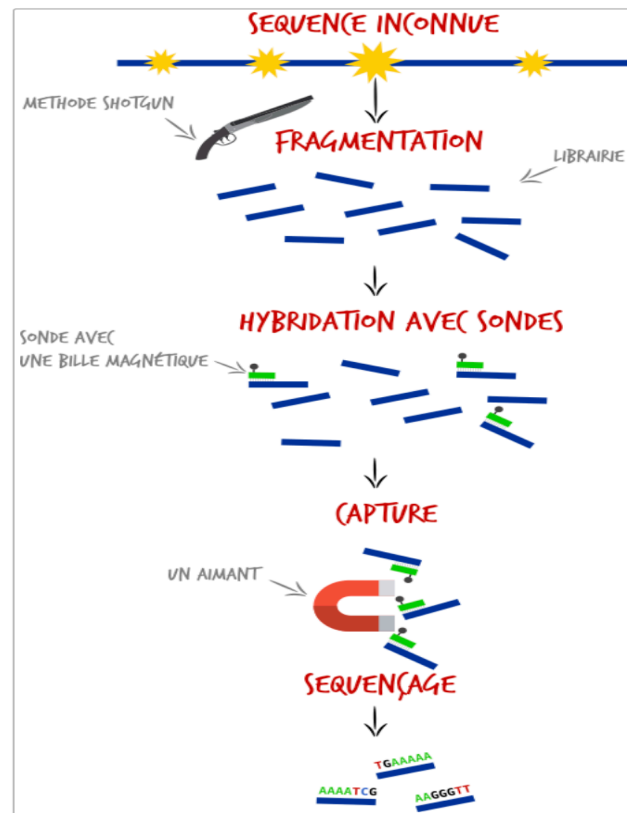
2. Stratégies d'enrichissement

- **Séquençage ciblé** (*capture hybride, amplicons*) : Améliore la profondeur de lecture sur des gènes spécifiques, idéal pour le diagnostic clinique. **-RNA-seq polyA vs. rRNA depletion** : Le choix entre ces méthodes dépend de l'intérêt pour l'ARN codant (mRNA-seq) ou l'ensemble du transcriptome (total RNA-seq).

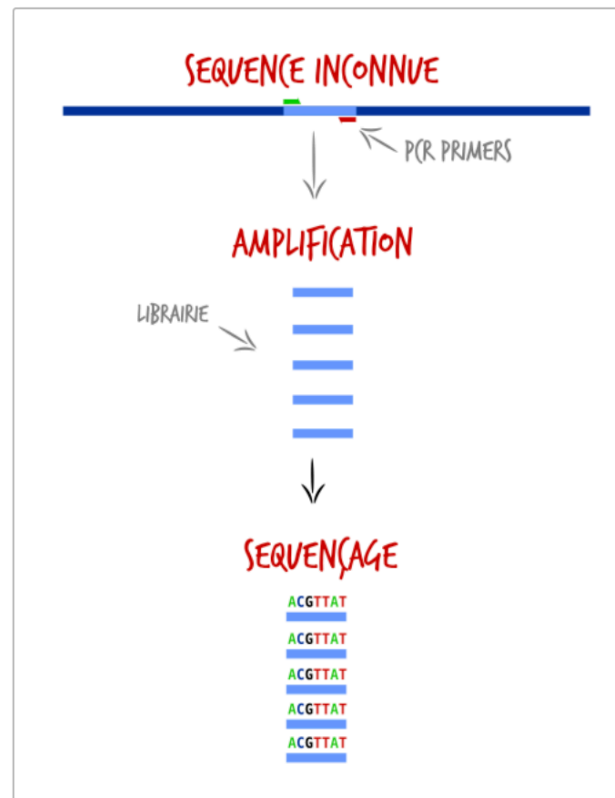
Technologies du NGS et leur spécificités

Technologie	Principe	Longueur des Reads	Avantages	Inconvénients
Illumina	Séquençage par synthèse avec fluorescence	50-300 pb (paired-end)	Haute précision, faible coût par base	Lectures courtes, difficulté sur les répétitions génomiques
Ion Torrent	Détection de pH lors de l'incorporation de nucléotides	200-600 pb	Rapidité, coût modéré	Sensible aux erreurs d'homopolymères
PacBio (SMRT)	Séquençage en temps réel avec polymérase unique	>10 kb	Lectures longues, détection d'épimutations	Taux d'erreur élevé, coût élevé
Oxford Nanopore	Passage d'ADN à travers un pore biologique	Jusqu'à plusieurs Mb	Séquençage ultra-long, faible coût d'infrastructure	Taux d'erreur encore élevé

Séquençage par capture



Séquençage par amplicon



Différence Pair-End (PE) et Single-End (SE)

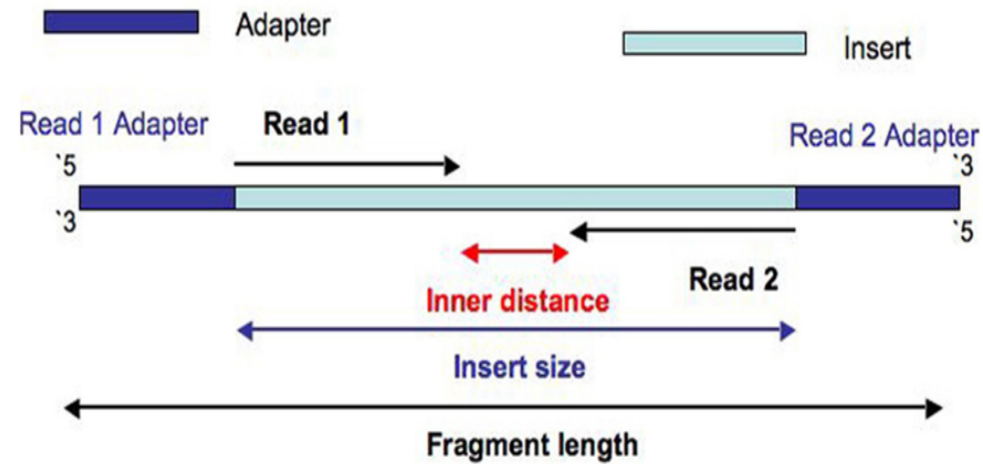
Il est avantageux d'obtenir des fragments de lecture les plus longs possible pour un alignement le plus fiable possible.

- **Lors d'un séquençage simple** (*Single-end*), les brins sont séquencés en partant d'un unique adaptateur. Un *read* correspond donc ensuite à un fragment.
- **Lors d'un séquençage *appairé*** (*Paired-end*), les brins sont séquencés à partir de leur deux extrémités. Les fragments résultats sont appelés R1 et R2 et sont liés, qu'ils soient *recouvrant* ou pas.

Un site expliquant cela de manière intéressante:

<http://thegenomefactory.blogspot.com/2013/08/paired-end-read-confusion-library.html>

Pair-End (PE) et Single-End (SE)



Etapes de l'analyse bioinformatique:

- **Contrôle Qualité** sur les données brutes suivi éventuellement d'un **Trimming**
- **Alignement des reads** sur le génome de référence
- Appel de **variants**
- **Annotation** et production d'un fichier **VCF** et d'un **compte-rendu**

Alignement de séquences

Sortie informatique du séquenceur

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

Mesure et encodage qualité: le Phred

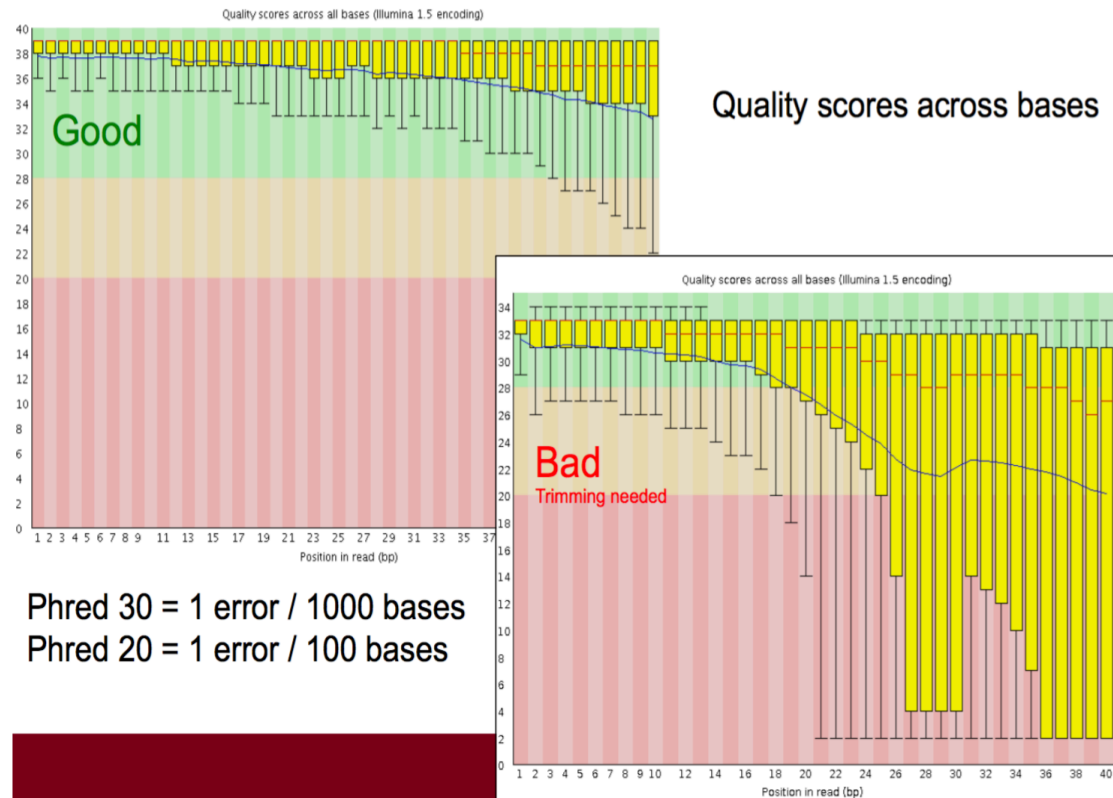
Quelques définitions:

- 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



Filtrage des séquences

Le *trimming* ("rognage") est une étape préliminaire mais cruciale qui consiste à nettoyer les lectures (*reads*) pour améliorer la qualité globale des données en supprimant :

1. Les adaptateurs :

- Séquences artificielles ajoutées pendant la préparation des librairies.
- Peuvent apparaître à l'extrémité des reads si la lecture dépasse l'insert.

2. Les bases de mauvaise qualité :

- La qualité de séquençage chute souvent sur le début ou la fin des reads.
- On retire les bases dont le score de qualité est trop faible (Phred < 20 ou 30).

3. Les séquences trop courtes ou ambiguës :

- Les reads très courts (après trimming) ou contenant trop de N sont parfois éliminés.

Il existe plusieurs programmes tels que **Trimomatic**, **Cutadapt**, **fastp**.

Logiciels d'alignement des séquences

But: Aligner les séquences présentes dans les fichiers **FASTQ** sur le génome de référence (Fichier **fa**).

- **Étapes principales :**

1. Indexation du génome → Construction d'une structure de données pour l'alignement rapide.
2. Alignement des reads → Chaque read est comparé à l'index pour trouver sa position optimale.
3. Gestion des erreurs → Les logiciels tolèrent des mismatches, insertions ou délétions.

Logiciels courants : - **BWA** et sa variant **BWA-mem** (pour l'ADN) - **STAR** (pour l'ARN)

Ces logiciels vont produire un fichier **BAM**: contient les positions, qualités, erreurs, etc.

Logiciels pour l'alignement de séquences

Logiciel	Description	Usage
BWA (Burrows-Wheeler Aligner)	Rapide, précis, adapté à l'ADN	Exome, génome entier
Bowtie2	Très rapide, faible empreinte mémoire	Petits génomes, épigénomique
Minimap2	Alignement long-read + court-read	PacBio, Nanopore, Illumina

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.

Cette transformation demande une **indexation** du génome de référence comme étape préliminaire.

BWA utilise une quantité relativement faible de mémoire (**bwa-mem**) et est parallélisable, pour exploiter les architectures multi-coeurs.

Detection de variants

Détection de variants à partir des fichiers BAM

La détection des variants (*variant calling*) consiste à identifier les différences entre la séquence d'ADN séquencée (*reads* dans le fichier BAM) et une **séquence de référence**. Cette analyse permet de détecter des mutations telles que des **substitutions**, des **insertions**, des **délétions** et des **variations structurelles**.

- **SNP**: Single Nucleotide Polymorphisms: Changement d'un simple nucléotide
- **InDEL**: Insertion-DEletion: Insertion ou délétion d'une séquence jusqu'à 50 nucléotides.
- **CNV**: Copy Number variation: Variant Structurel de **plus de 1kB** (*Autre algorithme de détection*)

SNP (Single Nucleotide Polymorphism)

- **Définition** : Un SNP est une variation de **simple nucléotide** dans une séquence d'ADN. Cela signifie qu'un seul nucléotide (A, T, C ou G) est remplacé par un autre.
- **Fréquence** : Les SNPs sont très courants dans le génome humain et représentent la variation génétique la plus répandue.
- **Exemple** : Si dans une séquence d'ADN on a un "A" à un endroit donné chez une personne, un autre individu peut avoir un "G" à ce même endroit.
- **Effet** : Les SNPs peuvent être **neutres** (n'ayant aucun effet), ou bien influencer l'expression des gènes, la fonction des protéines, ou la susceptibilité aux maladies. Certains SNPs sont également utilisés comme marqueurs génétiques pour étudier l'héritabilité de traits et de maladies.

InDELs (Insertion/Deletion)

- **Définition** : Un InDEL est une variation où une ou plusieurs paires de bases d'ADN sont soit **insérées** (ajoutées) soit **supprimées** (délétion) dans une séquence d'ADN. Les InDELs peuvent être de petite taille (1-50 paires de bases) ou plus grandes.
- **Exemple** : Si une séquence originale est ATGCGT, une insertion pourrait donner ATGCCGT, et une délétion pourrait donner ATGT.
- **Effet** : Les InDELs peuvent provoquer un décalage du cadre de lecture (frameshift), surtout dans les gènes codants, ce qui peut aboutir à des protéines non fonctionnelles. Cela peut être lié à des **maladies génétiques** ou des **traits spécifiques**.

CNV (Copy Number Variation)

- **Définition** : Un CNV est une variation où des **grandes portions de l'ADN** (plus de 1 kb) sont présentes en **copies supplémentaires ou manquantes** par rapport au génome de référence. Cela implique une duplication ou une délétion de segments d'ADN, souvent beaucoup plus grands que les Indels.
- **Exemple** : Une personne peut avoir trois copies d'une région spécifique d'un chromosome, alors que la plupart des gens en ont deux (une copie de chaque parent).
- **Effet** : Les CNVs peuvent affecter plusieurs gènes et ont un **impact majeur sur l'expression des gènes**, car ils augmentent ou diminuent la quantité d'ADN codant disponible. Ils sont associés à **divers troubles génétiques**, comme le syndrome de Down (duplication d'un segment sur le chromosome 21), et à d'autres traits et maladies complexes.
- **Détection**: ils sont détectés par des **algorithmes plus complexes** que les SNPs et InDELS.

Appel de variants ponctuels (SNPs) - Principe général

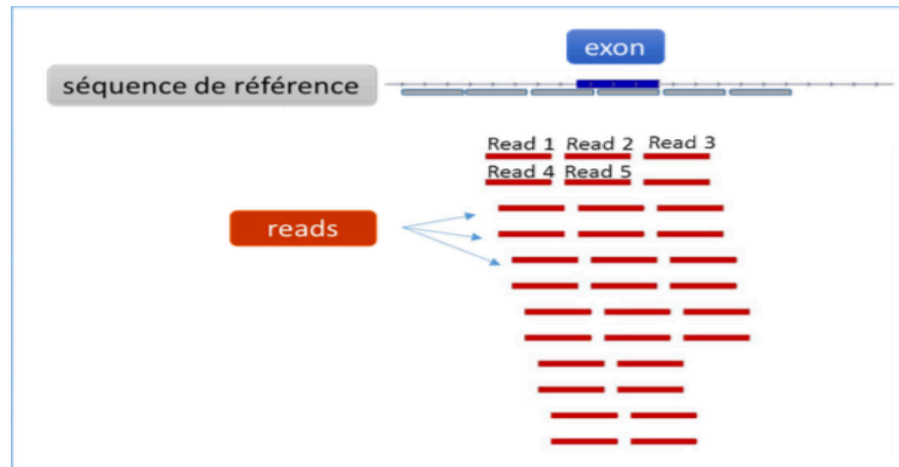
Lors du séquençage, chaque position du génome est couverte par **plusieurs reads**. Pour chaque position, le variant caller compare les bases nucléotidiques observées aux bases attendues dans la séquence de référence et calcule la fréquence des allèles alternatifs.

Quelques définitions: les Reads

Quelques définitions

Reads= lectures= séquences Exemple: ATCGGGTTACCAACCGAAT

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



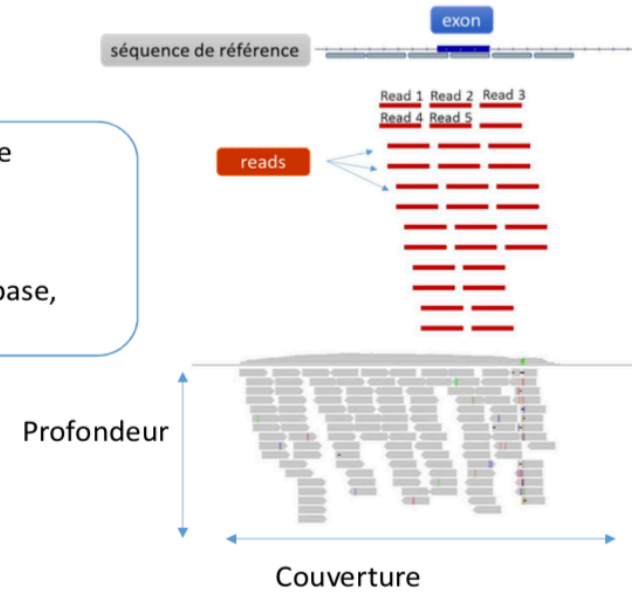
Couverture et profondeur

Quelques définitions

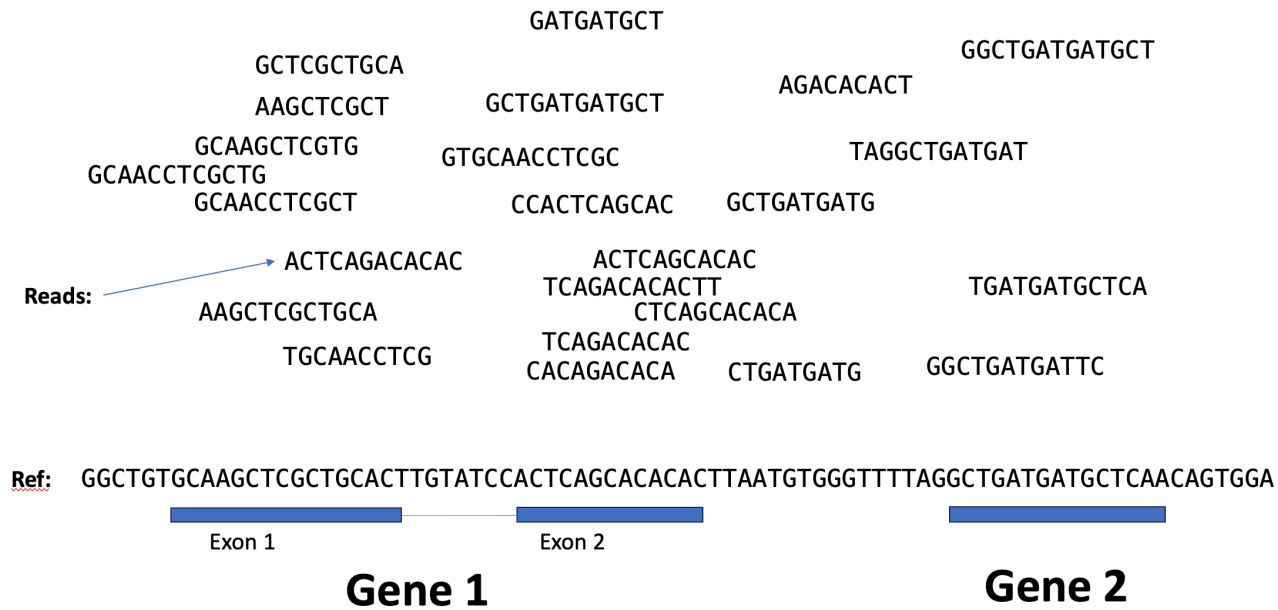
Profondeur-Couverture:

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

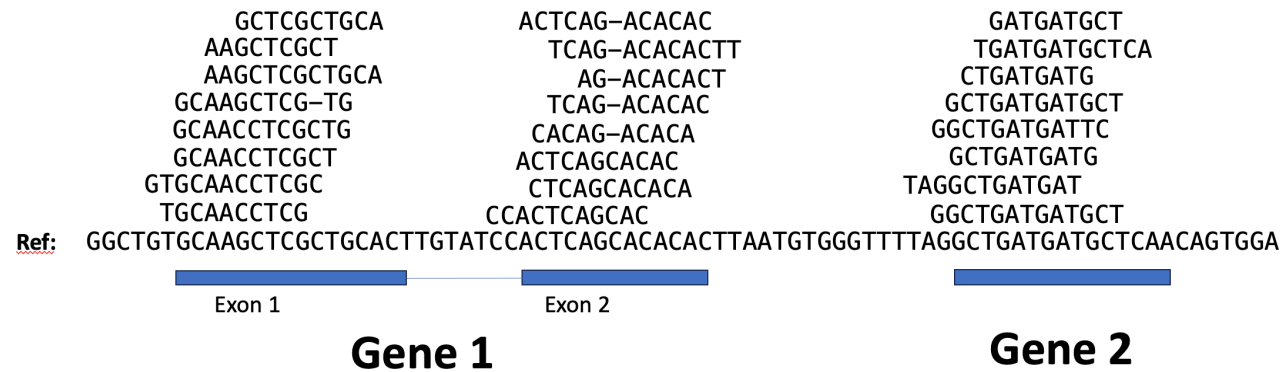
Profondeur: nombre de lecture de chaque base, exprimée en X



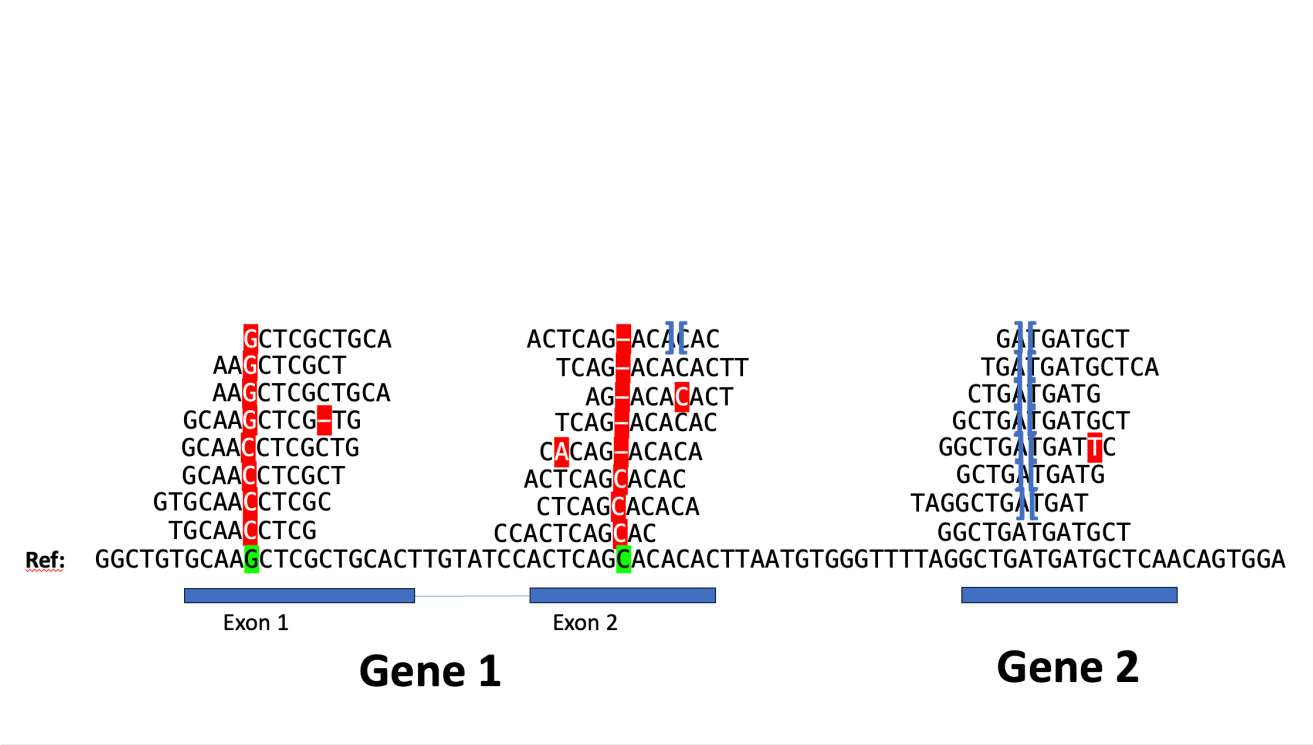
Détection de SNPs par NGS: Sortie du séquenceur



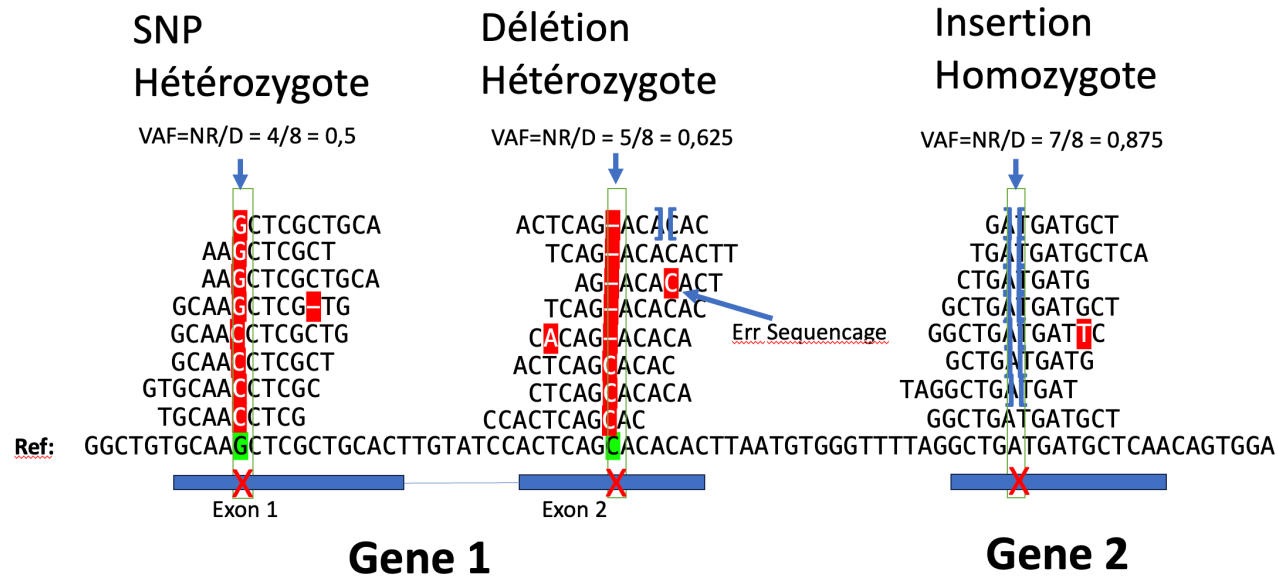
Détection de de SNPs par NGS: Alignement



Détection de de SNPs par NGS: Détection



Détection de de SNPs par NGS: Type/VAF



Détection de SNPs par NGS

But: recherche de **mutations** dans des **gènes d'intérêt** sur un échantillon.

Etapas de l'analyse bioinformatique:

- **Contrôle Qualité** sur les données brutes (**Obligatoire!**)
- **Alignement des reads** sur le génome de référence
- **Appel de variants** (Recherche des SNP et INDELS)
- **Annotation** et production d'un fichier **VCF** et d'un **compte-rendu**
- $VAF = VariantAlleleFrequency = \frac{N(ReadsVariants)}{Profondeur}$

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;OCCURREN
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 GAGA AGC -		ID=COSM6223;OCCURRENCE
chr7	NM_005228	EGFR	20	c.2361G>A	p.(=)	18	338	1913	exonic	synonymous	not found	55249063	G	A	ID=COSM1451600;OCCURREN
chr7	NM_001127500	MET	2	c.534C>T	p.(=)	60	689	1140	exonic	synonymous	not found	116339672	C	T	ID=COSM1579024;OCCURREN
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



Détection de variants par VarScan

Référence: Daniel C. Koboldt *et al*: VarScan: variant detection in massively parallel sequencing of individual and pooled samples: *Bioinformatics*. 2009 Sep 1; 25(17): 2283–2285.

Varscan est un programme de détection de variants utilisable aussi bien en **constitutionnel** qu'en **tumoral**. Il permet également de travailler de manière **individuelle** (un échantillon à la fois) ou sur plusieurs échantillons par le biais d'un **VCF multi-échantillons**.

Site Web: <https://dkoboldt.github.io/varscan/germline-calling.html>

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

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

Référentiels d'annotations: GENCODE

GENCODE est un projet international visant à fournir une annotation complète et précise des éléments génomiques du génome humain (hg19/GRCh37, GRCh38) et de la souris (mm10/GRCm38, GRCm39): <https://www.encodegenes.org/>

- La liste complète des **gènes** (protéiques et non codants)
- Les **transcrits alternatifs**
- Les **pseudogènes**
- Les **exons, introns, UTRs, start/stop codons**
- Les **fichiers GTF/GFF** utilisés dans les pipelines bioinformatiques

A quoi ca sert ?

- **Annotation** des variants (ex: *VEP* ou *ANNOVAR* utilisent GENCODE)
- **Analyse** d'expression (ex: RNA-seq)
- Études des **transcrits non codants**
- Support aux projets de **génomique fonctionnelle**

Référentiels d'annotations: RefSeq

RefSeq (Reference Sequence) est une base de données de séquences de référence maintenue par le NCBI (National Center for Biotechnology Information). Elle fournit des **séquences normalisées et validées** pour :

- ADN génomique
- ARN (messenger, non codant)
- Protéines

Objectif:

- Offrir une référence stable pour la recherche, le diagnostic et les analyses bioinformatiques.
- Uniformiser les annotations et limiter la redondance (une seule séquence de référence par gène).
- Permettre la comparaison entre individus ou espèces via des séquences bien définies.

Comparaison GENCODE/Refseq

Caractéristique	RefSeq (NCBI)	GENCODE / Ensembl
Institution	NCBI (USA)	EMBL-EBI / UCSC / GENCODE (Europe / international)
Objectif principal	Fournir des <i>séquences de référence validées</i>	Offrir une <i>annotation complète et exhaustive</i>
Types de séquences	Sélection filtrée (curation manuelle + automatique)	Annotation large incluant isoformes rares
Nomenclature	NM_, NR_, NP_, NG_, NC_	ENSG_, ENST_, ENSP_
Mise à jour	Moins fréquente, très contrôlée	Rapide et automatisée, cycles fréquents
Utilisation typique	Référence stable pour annotation clinique	Recherche, prédiction d'isoformes, transcriptomique
Genome Browser associé	NCBI Genome Data Viewer	Ensembl / UCSC Genome Browser
Interopérabilité	Compatible avec ClinVar, dbSNP, etc.	Lié à BioMart, VEP

Comparaison GENCODE/Refseq

- Accès RefSeq: <https://www.ncbi.nlm.nih.gov/refseq/>
- Accès GENCODE: <https://www.encodegenes.org/>

Les deux ressources sont **complémentaires** :

- **RefSeq** est souvent utilisé en **diagnostic** pour sa fiabilité.
- **GENCODE** est souvent préféré en **recherche fondamentale** pour sa richesse.

Nomenclature *HGVS*

La nomenclature *HGVS* (*Human Genome Variation Society*) est - un système standardisé - permettant de décrire les variations génétiques de manière **précise et univoque**.

- **Standardisation** : Permet une communication claire entre chercheurs et cliniciens.
- **Interprétabilité** : Utilisée dans les bases de données cliniques (ClinVar, dbSNP, LOVD).
- **Reproductibilité** : Facilite l'intégration des variants dans les outils bioinformatiques.
- *HGVS official guidelines* : <https://varnomen.hgvs.org>
- **Mutalyzer** (outil pour vérifier les annotations) : <https://mutalyzer.nl>

La nomenclature *HGVS* décrit les variations en fonction de la séquence de référence utilisée (ADN, ARN ou protéine)

Principe général

Elle suit la structure générale suivante :

[Type de séquence] [Identifiant de la séquence de référence] : [Type de variation]

Les principaux types de séquences utilisées sont :

- **g.** → **génomique** (ex. : NC_000023.11:g.32389689G>A)
- **c.** → **cDNA** (ex. : NM_004006.2:c.35delG)
- **r.** → **ARN** (ex. : NM_004006.2:r.35_36del)
- **p.** → **protéine** (ex. : NP_003997.1:p.Gly12Val)

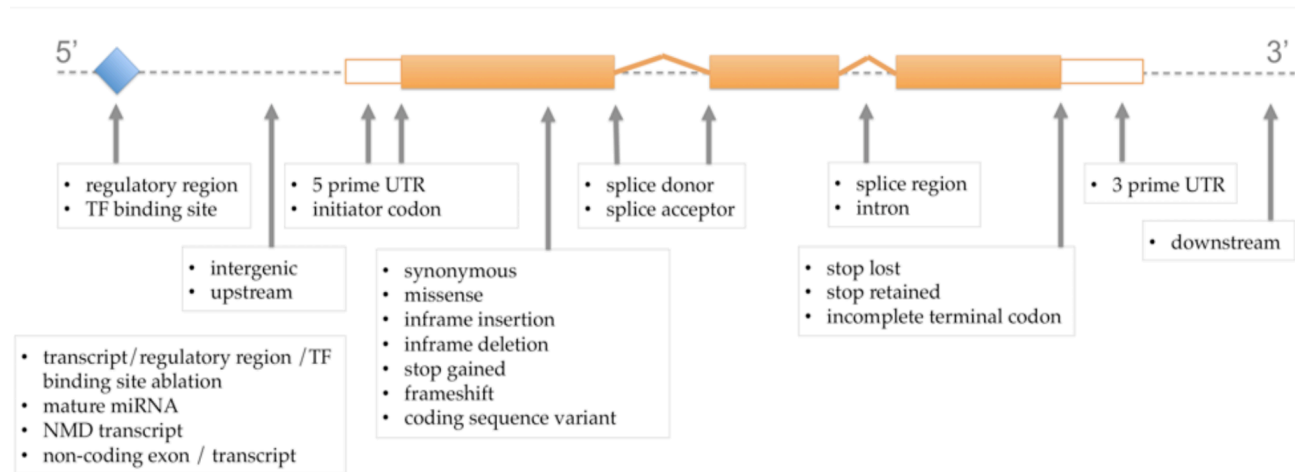
Annotation des variants en pratique

En quoi consiste l'annotation de variants ?

L'**annotation** des variants consiste à collecter de l'**information** sur l'impact du variant:

- Dans quel **gène/exon** se trouve la mutation et quel est son **impact** au regard de la protéine correspondante ?
- Détermination de leur **fréquence dans les populations**, que ce soit parmi la **population générale** ou parmi des patients atteints de **cancers** ou **maladies rares**.
- Conséquences fonctionnelles (Mutation sans effet (**synonyme**), Mutation faux-sens (**missense**), Mutation non-sens (**stop-gain, stop-loss**), Mutation sur **site d'épissage**, variant intronique (**UTR**))
- Interpretation clinique (classification **pathogène, vraisemblablement pathogène, VUS (Variant of Unknown Significance), Bénin**)

Localisation du variant



L'impact peut être plus ou moins important suivant la localisation du variant:

- **Région codante**, **Sites d'épissage** (jonctions exon-intron), **Régions régulatrices** (promoteur, UTRs, enhancers), **Introns** ou **régions intergéniques**

Localisation du variant

Région codante (exon)

- **Synonyme (silent)** : Pas de changement d'acide aminé → souvent neutre.
- **Missense (non synonyme)** : Changement d'un acide aminé → impact variable.
- **Nonsense (stop)** : Création d'un codon stop → protéine tronquée.
- **Frameshift** : Décalage du cadre de lecture → séquence totalement modifiée.

Sites d'épissage (jonctions exon-intron)

- Modification de l'épissage → perte ou ajout d'exons/introns.
- Peut causer des protéines aberrantes.
- ⚠ Détection par des scores comme **SpliceAI**, **MaxEntScan**.

Localisation du variant (Suite)

Régions régulatrices (promoteur, UTRs, enhancers)

- Influence l'expression du gène.
- Peut affecter la stabilité ou la traduction des ARNm.
- Impact plus difficile à prédire, outils : **RegulomeDB**, **FunSeq2**.

Introns ou régions intergéniques

- Souvent neutres.
- Risque : création de nouveaux sites d'épissage ou modification d'éléments régulateurs.
- **Filtrage des variants** : on priorise les exoniques non synonymes ou régulateurs.
- **Diagnostic moléculaire** : prédiction de pathogénicité.
- **Compréhension fonctionnelle** : mécanismes d'action des mutations.

Annotation de variants

- Support bioinformatique
 - Couverture
 - Fraction allélique
 - Qualité
- Scores prédictifs: proportionnels à leur **impact**:
- Scores prédictifs (**SIFT**, **PolyPhen**, **CADD**, **dbNSFP**)
- Impact **VEP** Variant Effect Predictor : *High, Moderate, Low, Modifier*
- Présence dans les bases de données de **pathogenicité** (ClinVar, HGMD...)

VEP (Variant Effect Predictor)

- VEP (Variant Effect Predictor) est un outil développé par [Ensembl](#)
- Il permet d'annoter automatiquement des variants (SNVs, indels, CNVs...)
- Pour chaque variant, il fournit :
 - Le **gène** concerné
 - La **conséquence** biologique (missense, stop gained...)
 - Des effets protéiques, notations **HGVS**, scores **SIFT**, **PolyPhen**
- Supporte les **formats VCF**, texte tabulé, HGVS
- Exemple: Variant : **13:32316461 C>T**
 - Annoté comme missense_variant dans le gène BRCA2
 - Scores : - **SIFT** : deleterious - **PolyPhen** : probably damaging
 - Annotation HGVS : *NM_000059.4:c.7007G>A (p.Arg2336His)*

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 Resource**.

- 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).

COSMIC: 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>

Mesure d'effet d'un variant par SIFT

- **Signification** : *Sorting Intolerant From Tolerant*
- **Objectif** : SIFT prédit si une substitution d'un acide aminé a un effet délétère sur la fonction de la protéine.
- **Principe** : À partir d'une séquence protéique, SIFT sélectionne des protéines homologues et construit un alignement multiple. Pour chaque position de l'alignement, SIFT calcule la **probabilité qu'un acide aminé soit toléré**, en supposant que l'acide aminé le plus fréquent est fonctionnel.

Si cette valeur normalisée est inférieure à un seuil, la substitution est prédite comme **délétère**.

Mesure d'effet d'un variant par

- PolyPhen

- **Signification** : Polymorphism Phenotyping
- **Objectif** : Prédit l'**impact potentiel** d'une substitution d'acide aminé sur la structure et la fonction d'une protéine humaine.
- **Méthode** : Basée sur des critères physiques (structure 3D) et des comparaisons évolutives.


- CADD

- **Signification** : Combined Annotation Dependent Depletion
- **Objectif** : Fournit un score intégré reflétant la probabilité qu'un variant soit délétère.
- **Principe** :
 1. Combine de **multiples annotations fonctionnelles** en un seul score.
 2. Compare les variants **naturellement sélectionnés** avec des **mutations simulées** pour évaluer leur **dangerosité potentielle**.

MobiDetails : Plateforme de visualisation pour l'interprétation clinique des variants génétiques

MobiDetails: an interactive application for clinical interpretation of human genome variations Martelotto L.G., Maussion G., Tournier I. et al.* *Bioinformatics* (2024), btae157

- Application interactive pour **explorer, annoter et interpréter** les variants génétiques.
- Spécialement conçue pour les **laboratoires de diagnostic**.
- Intègre des **annotations fonctionnelles**, des **données cliniques** et des **fréquences populationnelles**.

 <https://mobidetails.iurc.fr>

Analyses en génétique constitutionnelles et
somatiques

Objectifs propres à chaque analyse

Travailler sur des données de séquençage **constitutionnelles** (germinales) et **somatiques** (tumorales) implique des approches et des objectifs très différents, tant sur le plan biologique, **bioinformatique**, que clinique.

Type de variant	Origine	Présence	Exemple
Constitutionnel (germinal)	Hérité ou de novo	Dans toutes les cellules	Mutation BRCA1 germinale
Somatique (Tumoral)	Acquis dans un tissu donné (souvent tumoral)	Dans un sous-ensemble de cellules	Mutation TP53 dans une tumeur

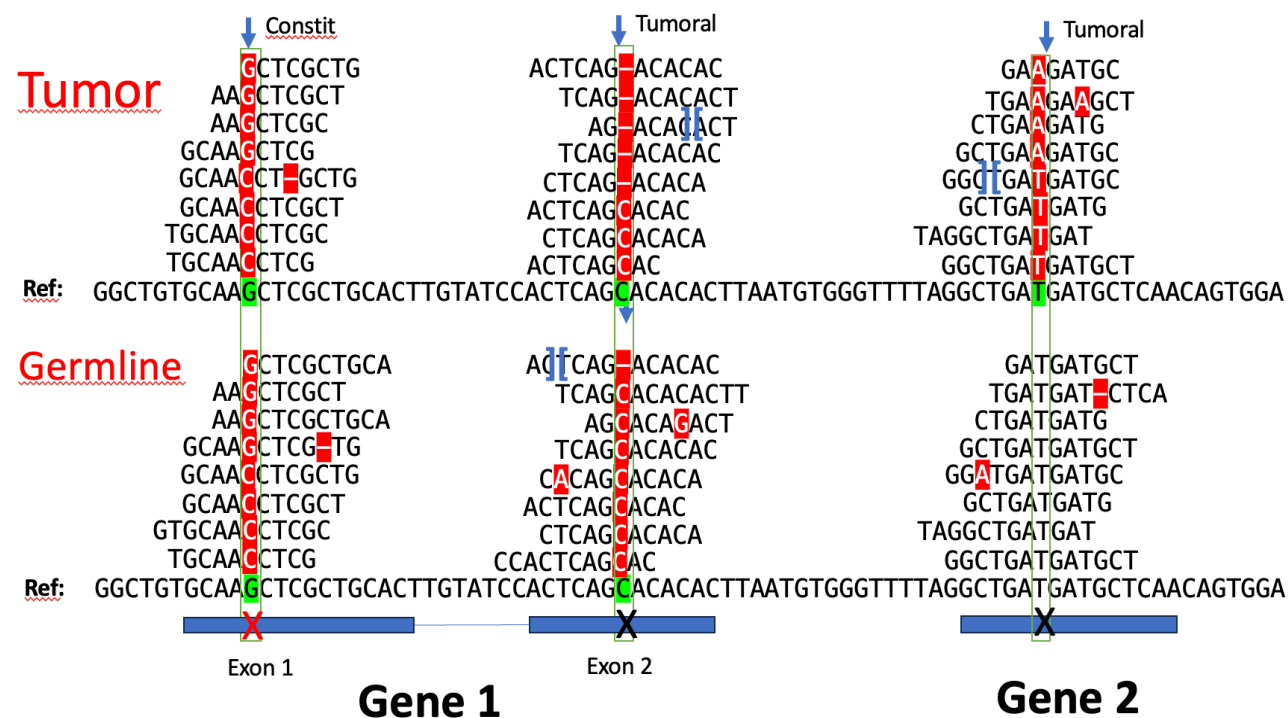
Approches bioinformatiques comparées

Étape	Constitutionnel	Somatique
Type d'échantillon	Sang, salive, peau	Tumeur + sang (paire tumor/normal)
Variant Calling	VarScan2, GATK HaplotypeCaller	Mutect2, VarScan2, Strelka2
Seuils alléliques	50% (hétéro), 100% (homo)	VAF faible possible (<5%)
Annotation	ClinVar, OMIM, VEP	COSMIC
Objectif	Diagnostic / Conseil génétique	Thérapeutique / Recherche cancer

Utilisation de Varscan

- **Varscan** utilise en entrée un fichier **mpileup** qui contient
 - La **profondeur de lecture** par position
 - Les **bases** observées
 - La **qualité** de base (Phred)
- **Analyse par VARSCAN**: Évaluation des positions présentant une variation :
 - Fréquence allélique
 - Qualité des bases
 - Couverture minimale
- **Application de filtres configurables** : fréquence, p-value, strand bias...

Détection de SNP Somatiques



Fonctionnement de Varscan

- **Mode Constitutionnel** (`mpileup2cns`) :
 - Appel de variants indépendamment, sans référence au contexte tumoral
 - **Résultats** : VCF contenant des SNPs et indels **germinaux**
- **Mode Somatique** (`somatic`) :
 - Comparaison normal vs tumeur à chaque **position**
 - **Identification des variants somatiques** via :
 1. Analyse **différentielle** des fréquences (Comparaison des fréquences alléliques sain et tumoral)
 2. Tests statistiques (**Fisher exact test**)

Priorisation des variants en pratique

- Filtrer sur la **fréquence allélique** maximum observable dans la population générale (**pathogénique** = rare dans la population)
- Filtrer sur l'impact mesuré par **SIFT** ou **PolyPhen**

Rappel: Les principales étapes bioinformatiques

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

Les fichiers manipulés en NGS et leurs 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** ou **.fasta**.
- Fichiers listant les mutations/Indels: **.vcf**.

Remerciements

Merci à **Nicole CHARRIERE** (Admin IFB) pour avoir créé l'espace de travail de cette formation sur l'infrastructure IFB



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](https://creativecommons.org/licenses/by-nc-nd/4.0/).