Utilisation de GriCAD pour les analyses de données de séquençage à haut débit pour le diagnostic au CHU de Lyon

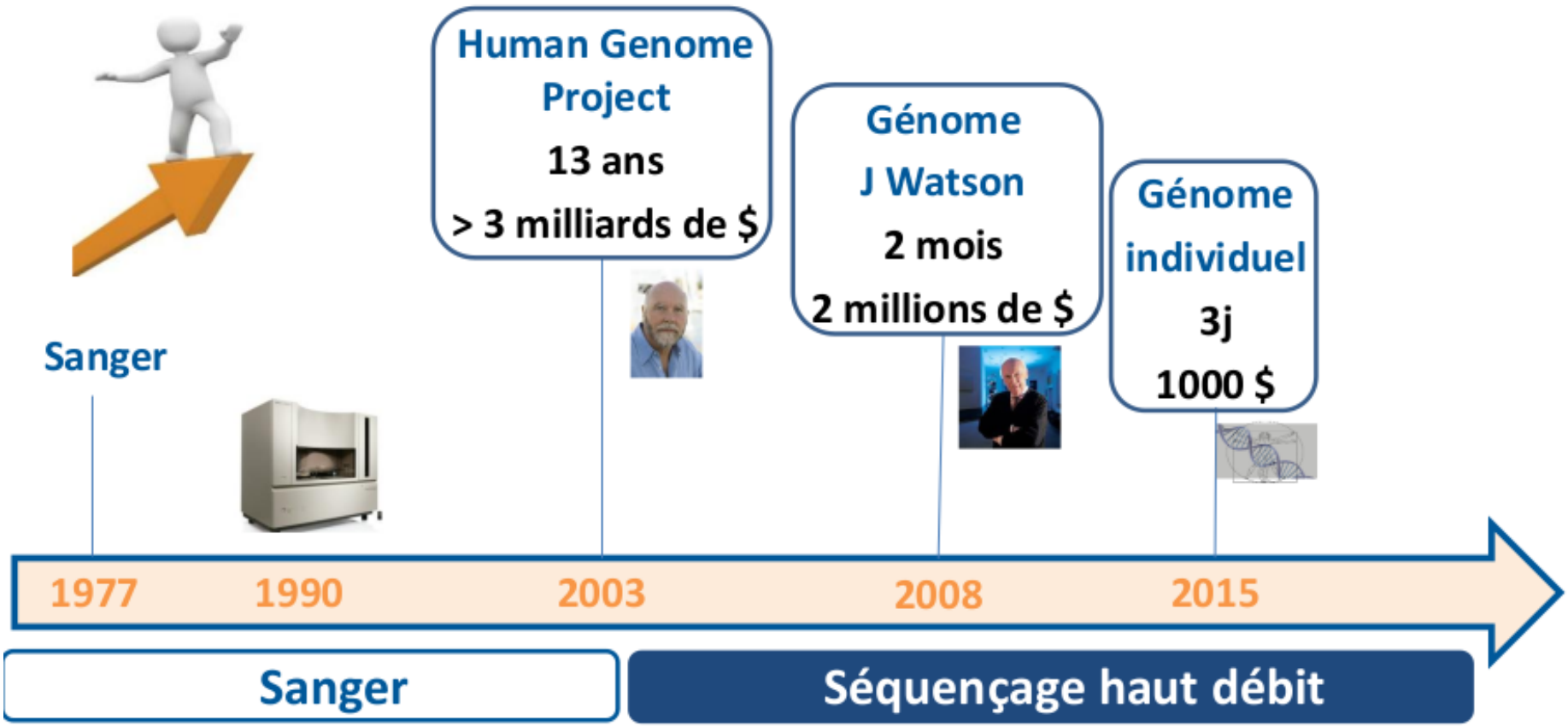
Claire Bardel
Cellule Bioinformatique des HCL

2 novembre 2020

Contexte : séquençage à haut débit

Séquençage à haut débit

- Séquençage : identification de la séquence de l'ADN d'un organisme
- Haut débit : possibilité de séquencer de nombreux fragments d'ADN en parallèle (plusieurs centaines de millions)



Contexte : séquençage à haut débit dans un CHU

Pourquoi séquence-t'on ?

- **Visée diagnostique** : identification de la variation génétique responsable d'une pathologie (épilepsie, cardiopathies, maladies métaboliques *etc.*)
- **Visée pronostique** : identification d'une variation génétique dans de l'ADN tumoral permettant de dire si l'évolution de la maladie sera favorable ou non
- **Visée théranostique** : identification d'un variation génétique tumorale, ou virale permettant d'adapter le traitement
- **Suivi épidémiologique :** identification des variations génétiques virales ou bactériennes et de leur diffusion spatiale et temporelle

Contexte : séquençage à haut débit dans un CHU

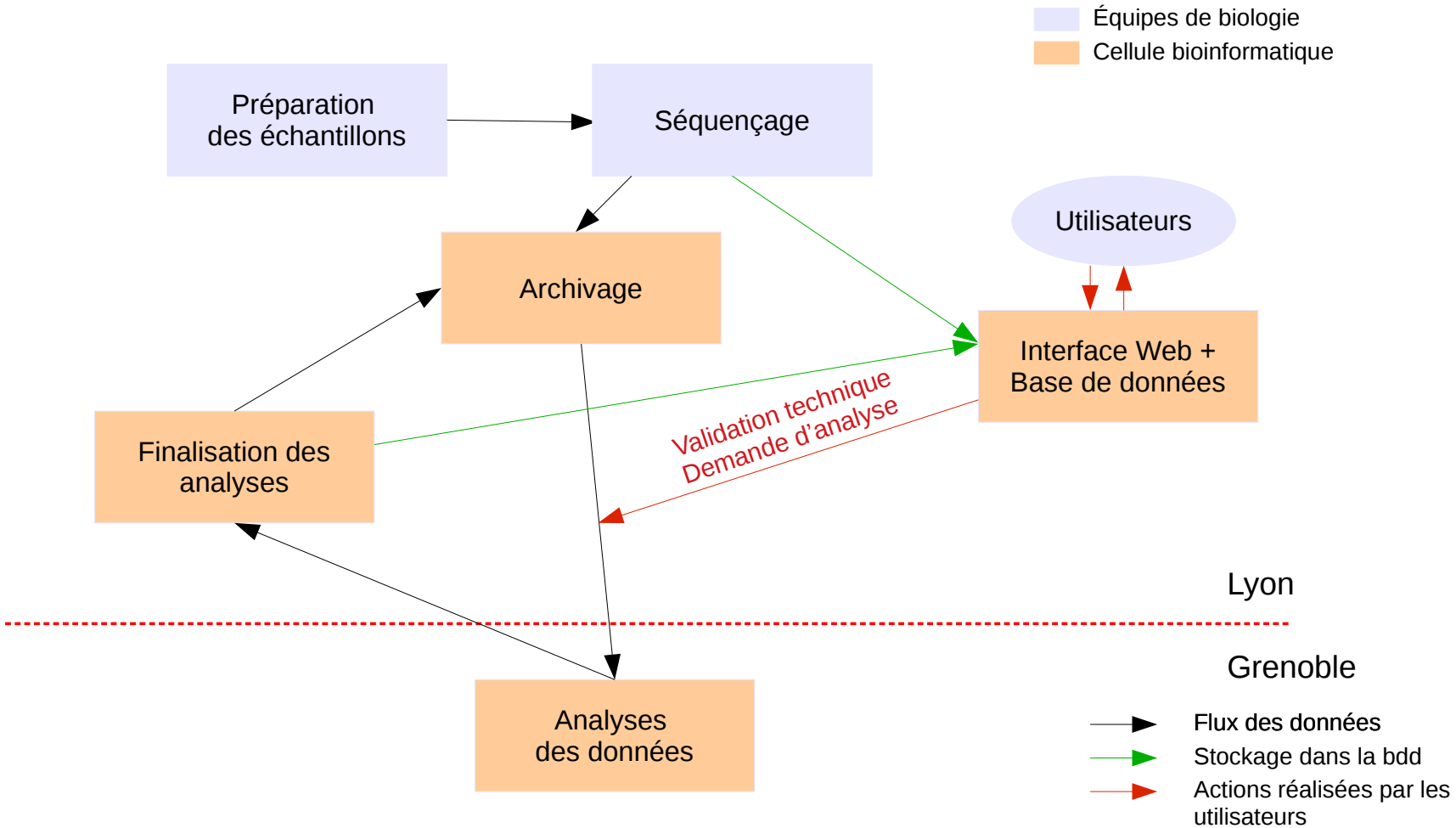
Qu'est-ce qu'on séquence ?

- Majoritairement de l'**ADN ou de l'ARN de patients**
  - séquençage d'un **panel** de gènes en lien avec une pathologie
  - séquençage des régions codantes de l'ADN de tous les gènes (2-3% du génome) = séquençage d'**exome**
  - séquençage du **génome**
- Mais aussi de l'**ADN ou de l'ARN viral ou bactérien**

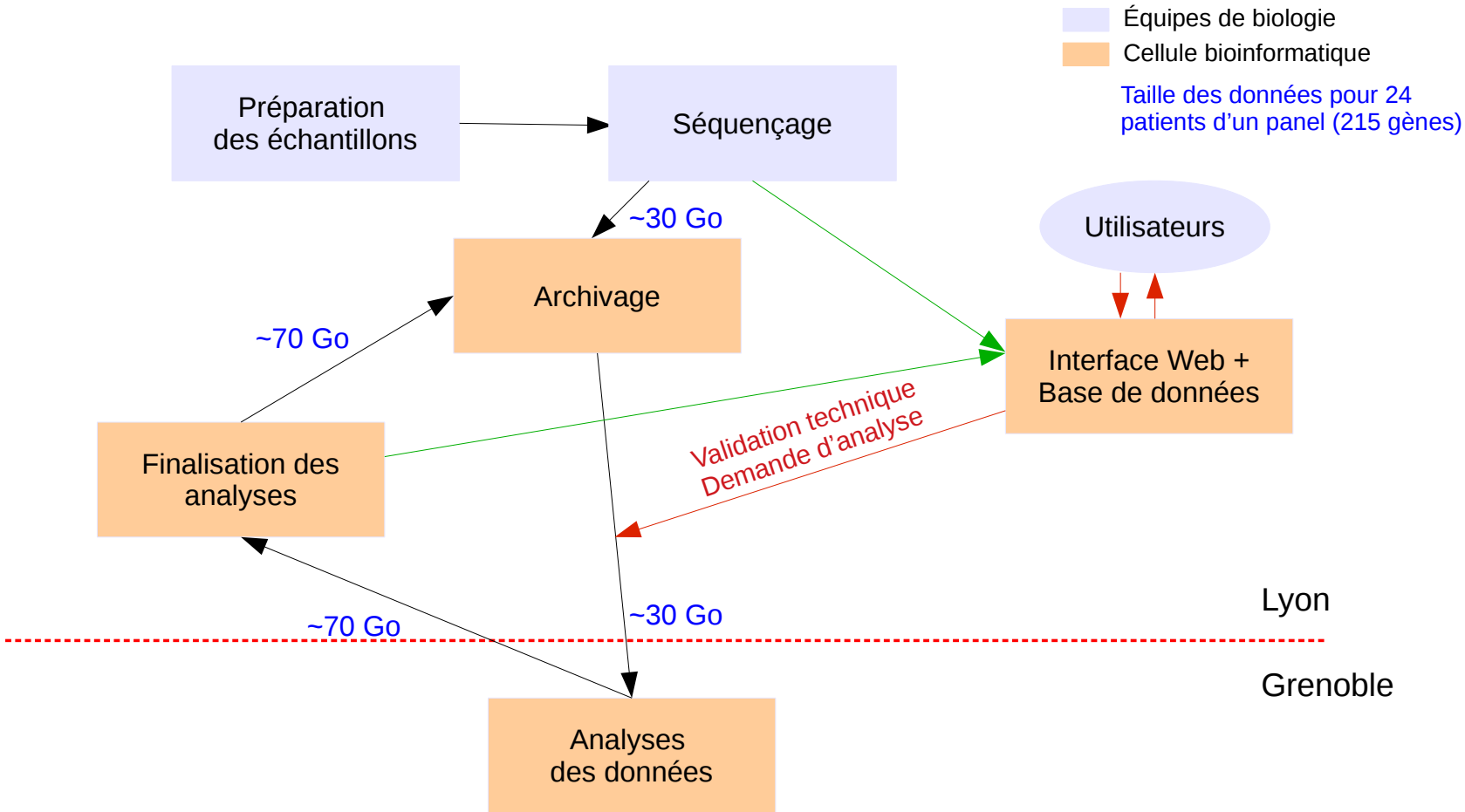
Conséquence sur les tailles des données initiales

- Panel de gène : 100 Mo à 2 ou 3 Go par patient
- Exome : 7 à 10 Go par patient
- Génome : 50 à 80 Go par patient

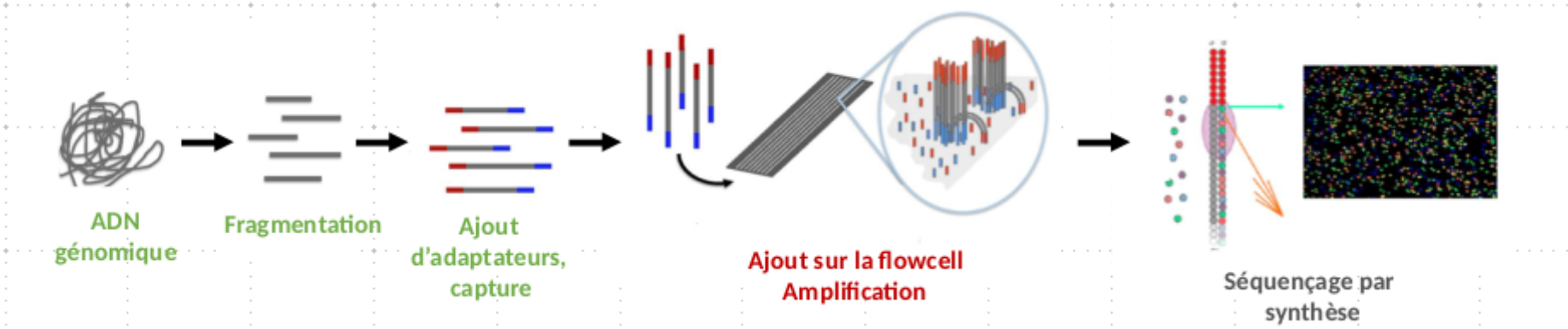
Organisation générale des analyses



Organisation générale des analyses



Détail des analyses : production des données



Technique

- **Préparation des librairies** : fragmentation de l'ADN génomique, ajout d'adaptateurs + codes-barres individuels, capture (sélection des fragments d'ADN d'intérêts)
- **Séquençage par synthèse** : détection de fluorescence émise lors de l'incorporation de nucléotides marqués

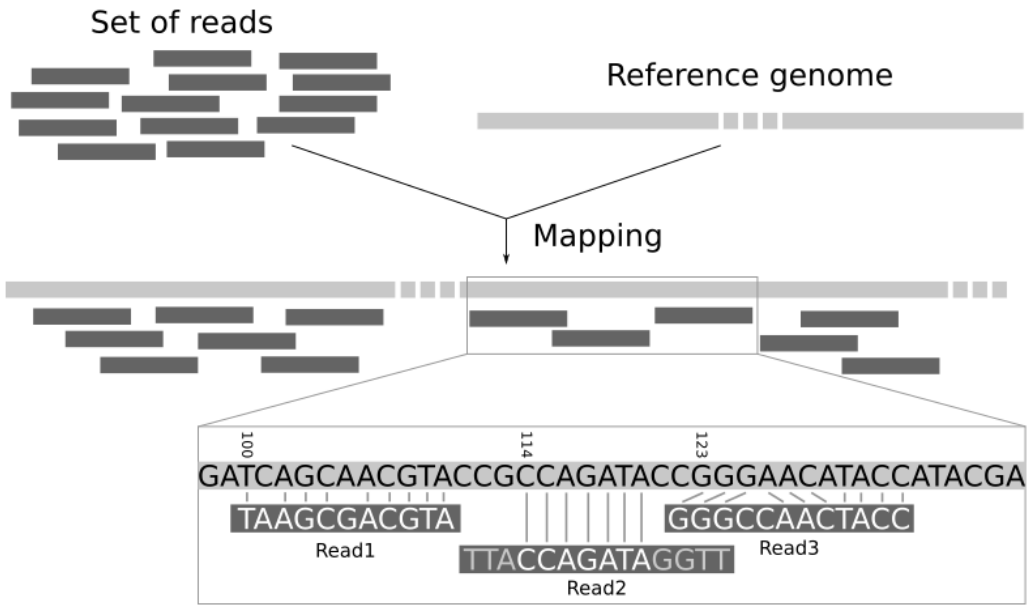
Démultiplexage

Attribution des séquences à chacun des patients grâce au code-barre individuel

⇒ **Obtention de fichiers au format fastq (texte)**

Détail des analyses : étapes réalisées sur GriCAD (1)

Alignement sur le génome de référence



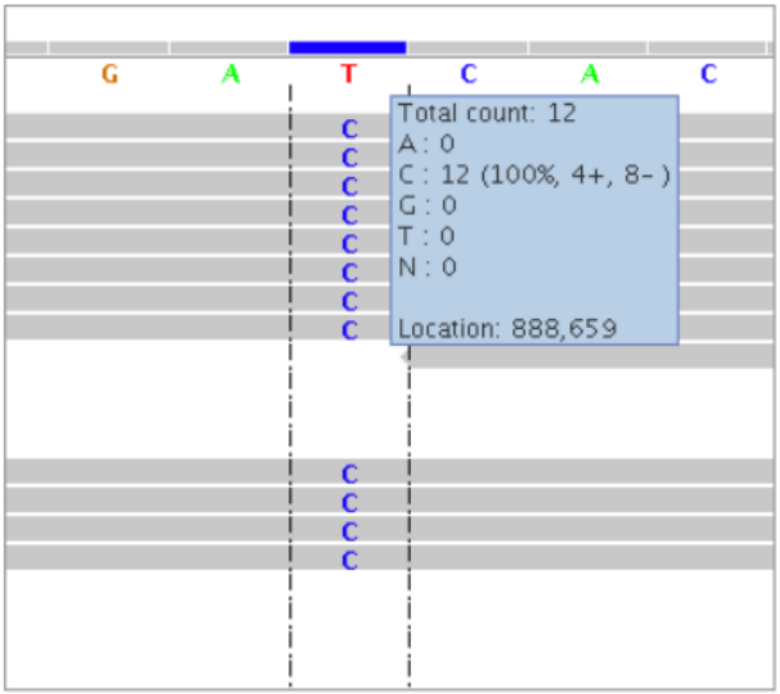
source de l'image :

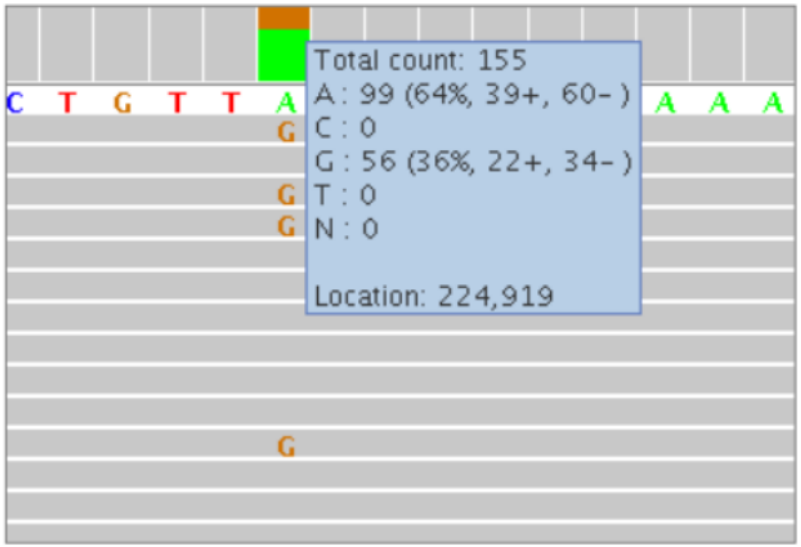
https ://galaxyproject.github.io/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

⇒ **Obtention de fichiers au format BAM (binaire)** + métriques de qualité

Détail des analyses : étapes réalisées sur GriCAD (2)

Appel de variants





Annotation des variants à l'aide de bdd

Prédiction de l'effet, fréquence dans des bases de personnes saines, pathogénicité établie dans d'autres études, *etc.*

⇒ **Obtention de fichiers au format VCF (texte)**

Détail des analyses : étapes réalisées sur GriCAD (3)

Pipeline d'analyse

- Pipeline Nextflow
  - intégration de OAR dans Nextflow (M. Vallée)
  - nombreux logiciels, besoins CPU/RAM variés
- Impératifs de production → image Singularity

Taille des fichiers manipulés

Pour un patient

	Panel	Exome	Genome
FASTQ	100 Mo - 2 à 3 Go	7 - 10 Go	50 - 80 Go
BAM	60 Mo - 1 à 2 Go	10 - 15Go	100 - 200 Go
VCF	qqs 10 Mo (24 p.)	1 à 2 Go (12 p.)	–

Détail des analyses : finalisation et mise à disposition

Fin des analyses à Lyon

- Fin de l'annotation
- Génération d'un rapport de qualité de l'ensemble de l'analyse
- Archivage des données
- Mise à disposition des résultats sur l'interface web

	VARIANT	GENE	GNOMAD	GENOTYPE	OMIM	EFFET	CLINVAR	HGMD	MOSAIQUE
1	Chr16:56374792A>G	*GNAO1*	40	Heterozygote	AD	missense_variant	Benin	Not Found	0.4763
1	Chr16:78312527A>G	*WWOX*	16379	Heterozygote	AR	missense_variant	NA	DM?	0.4227
4	Chr16:9934917C>T	*GRIN2A*	Not Found	Heterozygote	AD	missense_variant	Non Trouvé	Not Found	0.5091
	Chr19:13319691GGAT>G	*CACNA1A*	265	Heterozygote	AD	disruptive_inframe_deletion	Benin_Probable	Not Found	0.1316
2	Chr19:42482916T>C	*ATP1A3*	6	Heterozygote	AD	missense_variant	Non Trouvé	Not Found	0.4888
2	Chr5:125928405T>A	*ALDH7A1*	3	Heterozygote	AR	missense_variant	VOUS	Not Found	0.4690

Autres analyses réalisées sur GriCAD

Analyse de l'ADN humain

- Recherche de variants de nombre de copie
- Recherche de points de cassure de l'ADN

Analyse de l'ARN

- Recherche de transcrits de fusion
- Quantification du niveau d'expression des transcrits

Séquençage nanopore

- Utilisation des GPU pour réaliser l'appel des bases (passage du signal brut à des séquences de A, T, G et C)
- Appel des bases réalisé en 4h30 contre + de 3 semaines
- Réalisé au printemps sur les runs de coronavirus

Bilan

Contraintes spécifiques à notre activité

- Confidentialité des données : anonymisation stricte, HDS
- Rapidité des analyses : délais de rendu courts pour certaines analyses (cancérologie, diagnostic prénatal)
  Exemple en cancérologie : délai de rendu total = 1 semaine
- Nécessité de prévoir une solution dégradée sur des serveurs de calcul internes en cas de maintenance ou de pannes trop longues
- Données parfois très volumineuses

Bilan des analyses réalisées en 1 an

- $\sim$ 200 000 h de calcul réalisées
- $\sim$ 300 runs de séquençage analysés, correspondant à plusieurs milliers de patients

Remerciements

La cellule bioinformatique des HCL

- Claire Bardel
- Sylvain Mareschal
- Pierre-Antoine Rollat-Farnier
- Thomas Simonet
- Maxime Vallée

L'ensemble du personnel de la plateforme NGS des HCL

En particulier :

- Nicolas Chatron et Florence Roucher pour les illustrations des diapos 2 et 7