Forum TERATEC 2017
Atelier 2

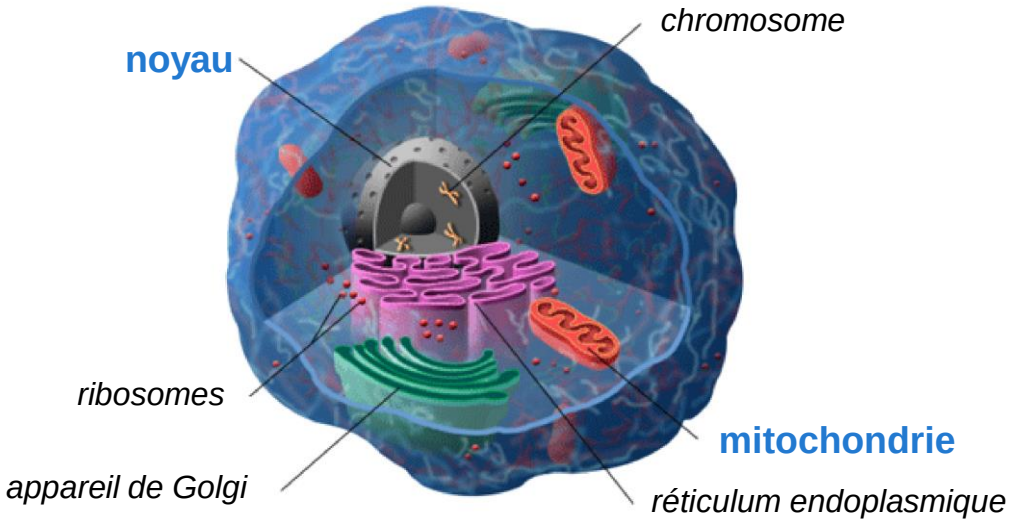
Santé - Perspectives pour une médecine personnalisée en 2025

Les challenges de la Génétique Humaine à l'ère du NGS

Professeur Christophe Béroud
INSERM UMR_S910, équipe "Génétique et Bioinformatique"
Aix-Marseille Université
christophe.beroud@univ-amu.fr

Quelques rappels

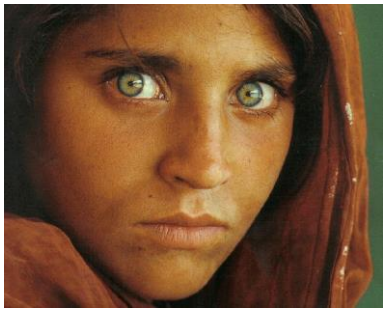
- Le génome humain est contenu dans le noyau et les mitochondries



- Le génome nucléaire est réparti en 23 paires de chromosomes (22 autosomes et X, Y)
- Le génome mitochondrial est en multiples copies

Apports de la génétique humaine

- **Comprendre les bases génétiques de la diversité entre individus**



Apports de la génétique humaine

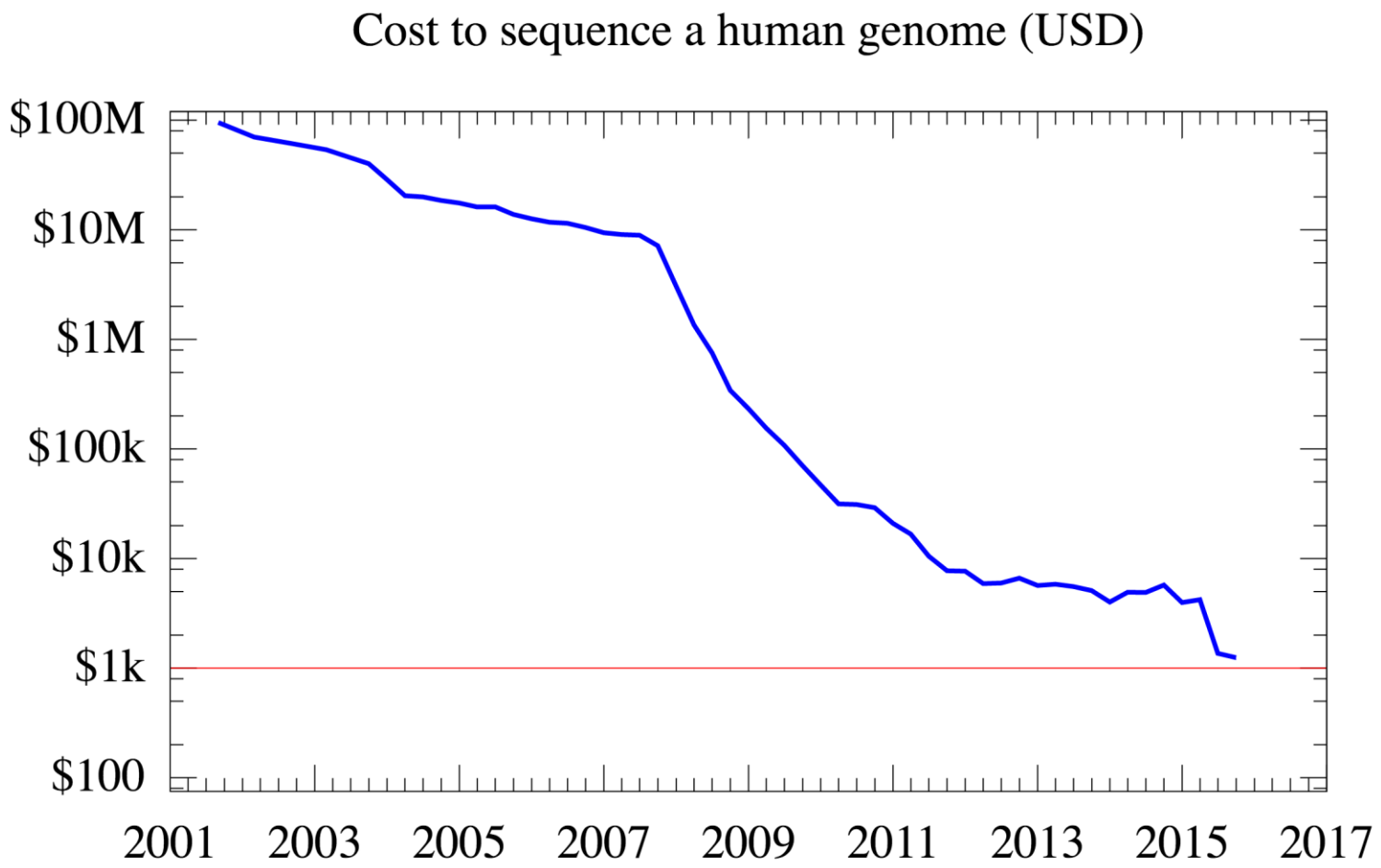
- **Comprendre les bases génétiques de la diversité entre individus**
- **Identifier les mutations responsables de maladies génétiques**
  - Eviter l'errance diagnostique
  - Meilleure prise en charge des patients
  - Conseil génétique dans la famille
    - Diagnostic prénatal
    - Diagnostic préimplantatoire
  - Développements thérapeutiques
    - "Thérapie génique"
- **Médecine personnalisée** "donner le bon médicament au bon patient à la bonne dose"

Le projet de séquençage du Génome Humain

Plus grand projet scientifique mondial lancé en 1988/1989, commencé en 1990 :

- Ampleur de la tâche
  - 1 page = 3000 bases
  - 1 tome de 500 pages = 1 500 000 bases
  - 1 génome diploïde = 2 000 tomes !
  - 1 génome humain (3,2 Gb) = 60 000 tomes de données brutes (30X) !
- Capacité de séquençage
  - En 1975, 1 000 bases/semaine ➛ 15 000 ans pour 100 personnes !
  - En 1986, 10 000 bases/jour ➛ 240 ans pour 100 machines
  - En 1998, 200 000 bases/jour ➛ 12 ans pour 100 machines

Le projet de séquençage du Génome Humain



Applications du séquençage de nouvelle génération (NGS)

- **Séquençage *de novo***
  - Génomes
  - Transcriptomes
- **Reséquençage**
  - Génomes
  - Exomes
  - Panels de gènes
- **Séquençage d'ARN**
  - ARNm
  - miARN
  - Dégradome
- **Chip-Seq** (identification de sites ADN permettant la fixation de protéines)
- **Methyl-Seq** (régulation génique médiée par la méthylation de l'ADN)
- **RIP-Seq** (Immunoprécipitation de l'ARN pour identifier les ARN liant des protéines)

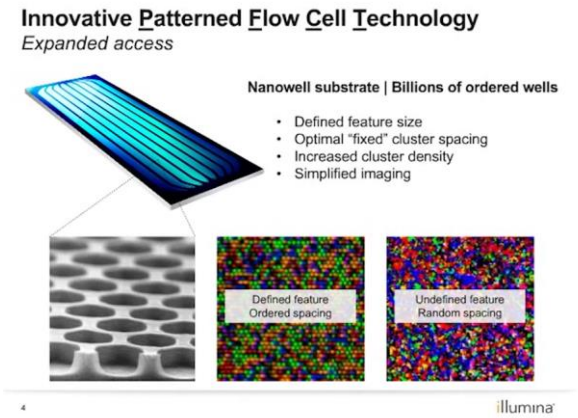
Les différentes étapes de l'analyse NGS

Le "*Wet Laboratory*"

- Prélèvement ADN
- Fragmentation
- Etapes de capture (Exome, séquençage ciblé, ...) *
- Construction de la librairie (amorces+ adaptateurs)
- Amplification clonale
- Séquençage
- Acquisition des signaux (fluorescence, différence de potentiels…)
- *Base calling* (Transformation des signaux en nucléotides de séquence)

Les différentes étapes de l'analyse NGS

Le "*Wet Laboratory*"



	HiSeq 3000 System	HiSeq 4000 System
Number of Flow Cells per Run	1	1 or 2
Output[b]		
1 × 50 bp	105–125 Gb	210–250 Gb
2 × 75 bp	325–375 Gb	650–750 Gb
2 × 150 bp	650–750 Gb	1300–1500 Gb
Clusters Passing Filter (Single Reads)	2.1–2.5 billion	4.3–5 billion
Quality Scores		
2 × 50 bp	≥ 85% of bases above Q30	
2 × 75 bp	≥ 80% of bases above Q30	
2 × 150 bp	≥ 75% of bases above Q30	
Daily Throughput	> 200 Gb	> 400 Gb
Run Time	< 1–3.5 days	< 1–3.5 days
Human Genomes per Run[c]	up to 6	up to 12
Exomes per Run[d]	up to 48	up to 96
Transcriptomes per Run[e]	up to 50	up to 100

Les différentes étapes de l'analyse NGS

Le "*Dry Laboratory*"

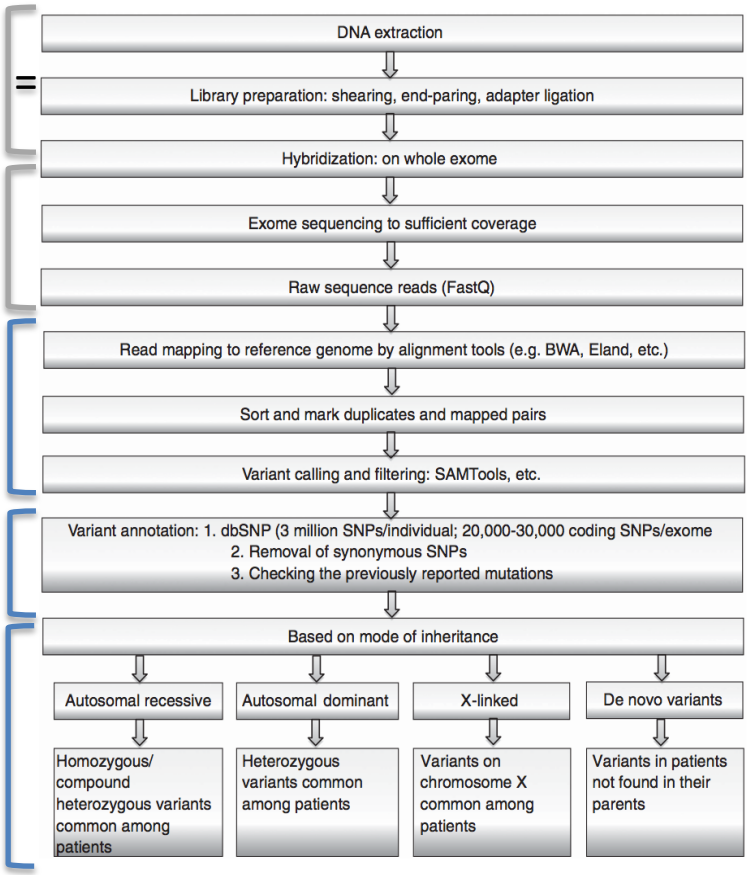
Préparation de l'échantillon

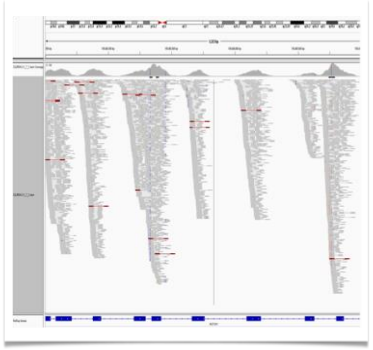
Etape de séquençage

Alignement et variant calling

Annotation de variants

Priorisation de variants





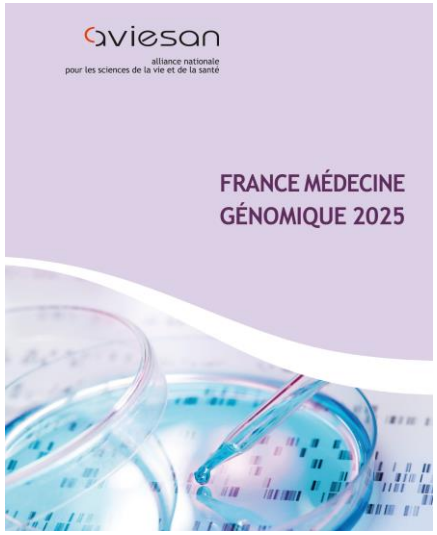
WES	WGS
100 000	4 500 000
.	.
.	.
.	.
.	.
.	.
.	.
1-10	1-10

Les challenges du NGS

Les challenges du NGS

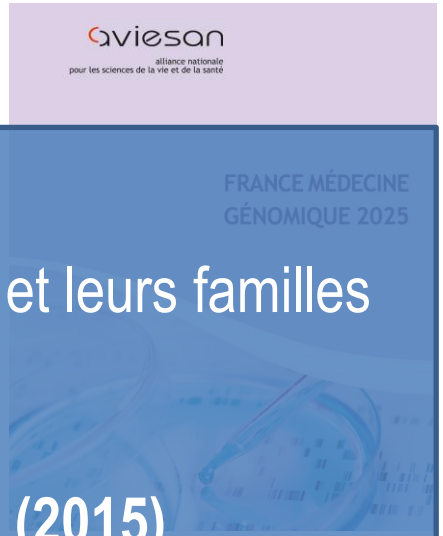
1- La taille des données

- Génome Humain haploïde ≃ 3.2Gb
  - Profondeur de couverture de 50X
    - Fichier FastQ compressé ≃ 120 Go
    - Fichier BAM ≃ 200 Go
  - Petit projet de 100 WGS : ≃ 20 To



- **Projet France Médecine Génomique 2025**
  - > 200 000 équivalents WGS/an (12 centres)
    - ≃ 20 000 équivalents WGS/an par centre : ≃ **4 Po/an**
    - Diagnostic ➛ conservation des données brutes = 30 ans

Le Plan France Médecine Génomique 2025



Maladies rares : 3-4 millions individu en France

Objectif : 20 000 patients atteints de maladies rares et leurs familles (environ 60 000 WGS)

Cancers : 384 500 nouveaux cas ; 149 500 décès (2015)

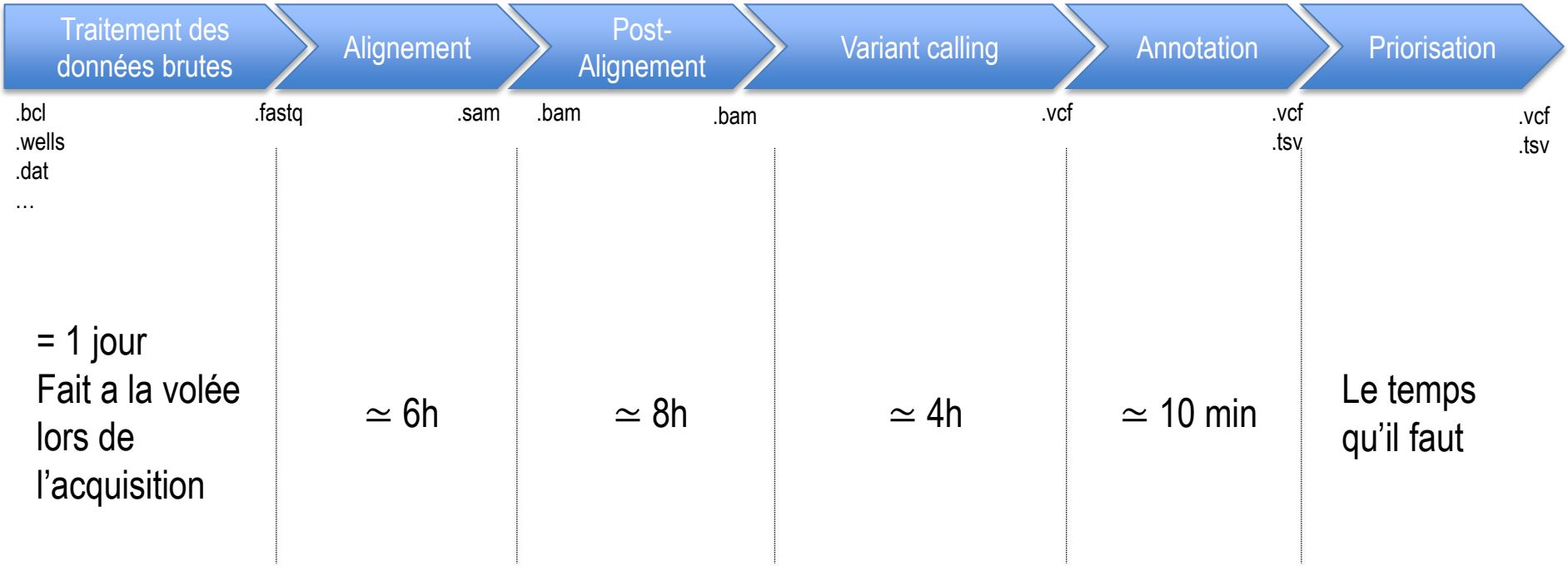
Objectif : 50 000 patients prioritaires car atteints de cancers métastatiques/réfractaires au traitement (environ 175 000 WGG)

- Filière de soins
- Filière industrielle

Les challenges du NGS

2- La puissance de calcul nécessaire

- **Pour un Exome**, sur une station de travail (2014) ➛ 18h



Les challenges du NGS

2- La puissance de calcul nécessaire

- **Pour un Exome**
  - Station de travail (2014) ➛ 18h
  - UV 2000 (1To, 256 cœurs) (2017) ➛ 40 mn (2 WES en parallèle)
  - Amélioration si RAM ↗ (et NAND) et SSD
- **Pour un Génome (30X)**
  - Station de travail (2014) ➛ 5j
  - UV 2000 (1To, 256 cœurs) (2017) ➛ 6 h
  - Clusters (52To, 1440 cœurs) (2017) ➛ 35h / WGS ; 55 / j
  - **UV 300 (96To, 1200 cœurs) (2017) ➛ 19 mn ; 72 / j**

Augmenter la RAM ➛ augmente la rapidité d'analyse
Augmenter les cœurs ➛ augmente le nombre d'analyses /j

Les challenges du NGS

2- Pourquoi faut-il de la puissance de calcul ?

- ≃ 20 000 équivalents WGS/an par centre : ≃ **4 Po/an**
- Diagnostic ➛ conservation des données brutes = 30 ans

- Pour **éviter un goulot d'étranglement**
  - ≃ 20 000 WGS/an ➛ 385/semaine ➛ **55/j**
- assurer un **rendu des résultats rapides**
  - Analyses secondaires complexes
  - Situations d'urgence (choix des traitements)

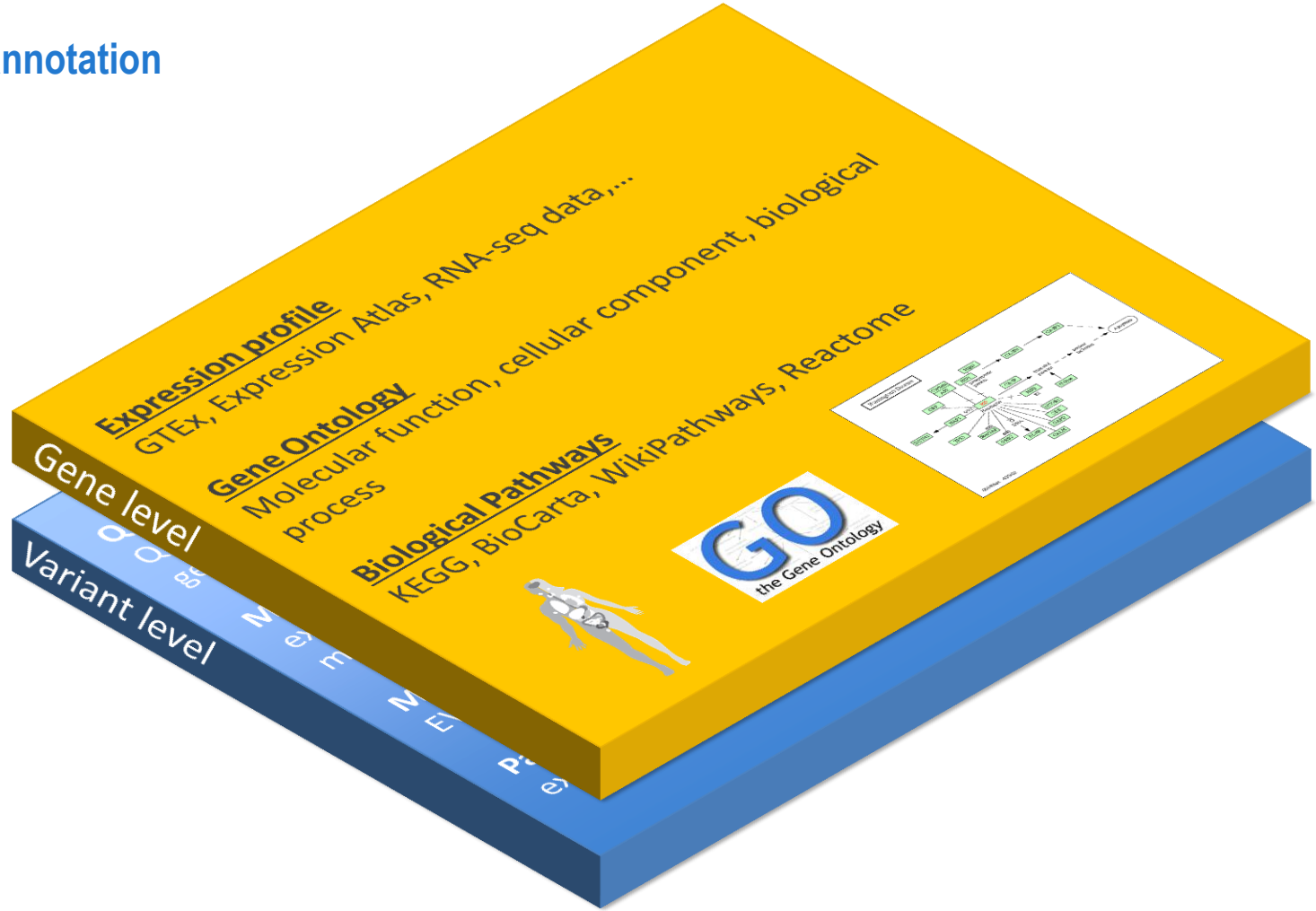
- **Stocker les données** dans des bases de données
  - Partager le savoir

Les challenges du NGS

3- L'annotation

Les challenges du NGS

3- L'annotation



Les challenges du NGS

3- L'annotation

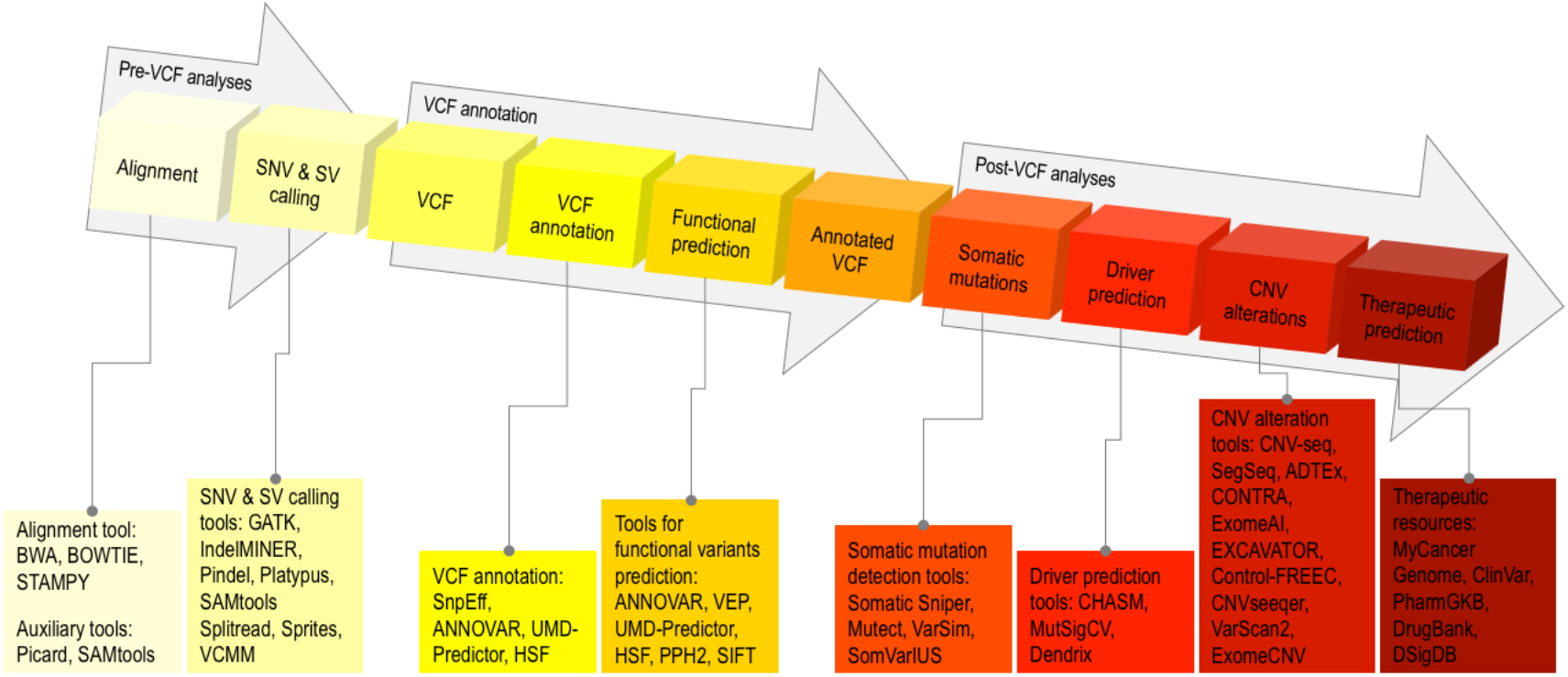
Les challenges du NGS

4- L'interprétation : ou comment identifier la ou les 2 mutations responsables du phénotype ?

Les challenges du NGS

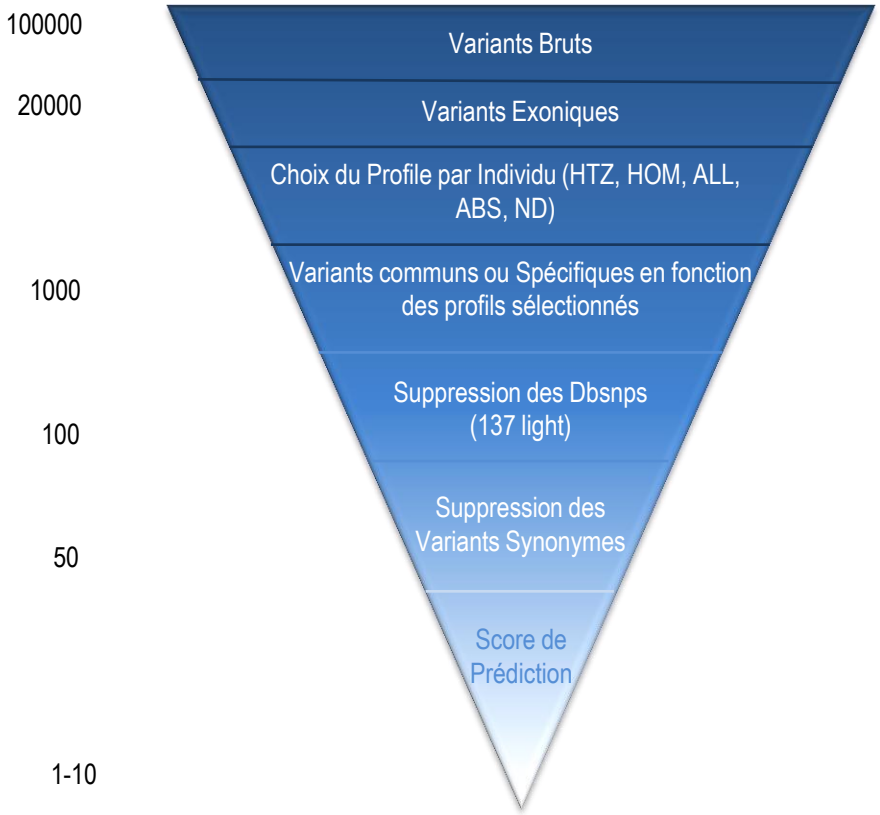
5- Bien construire son pipeline d'analyse

- **Nombreux outils disponibles pour chaque étape**



5- Bien construire son pipeline d'analyse

- **Filtration : réduire le nombre de variants à valider**

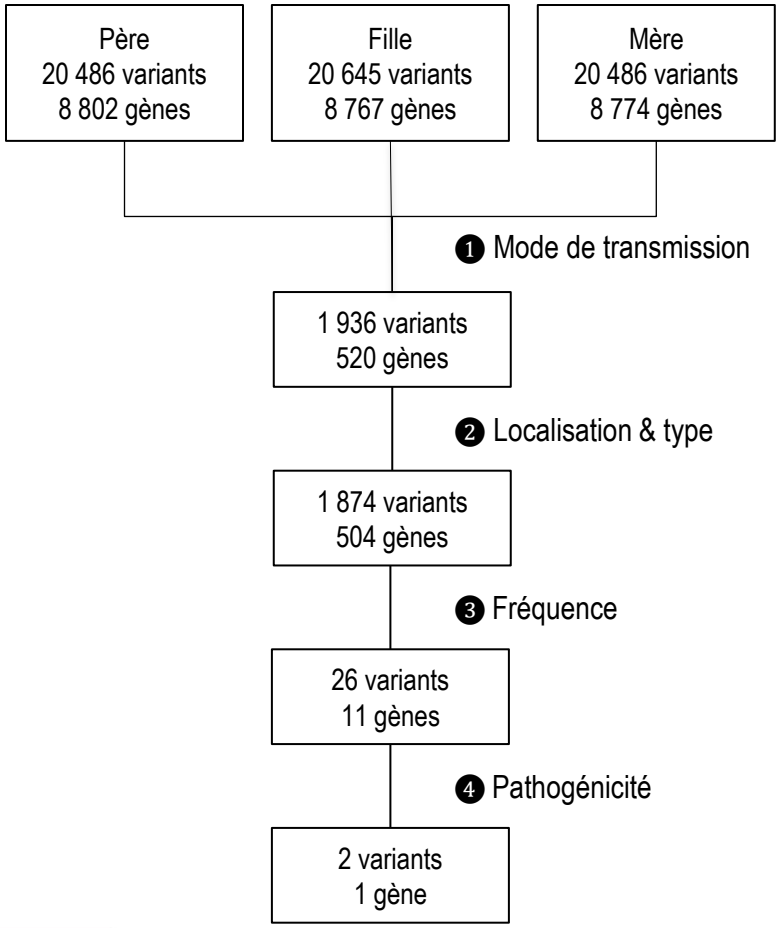


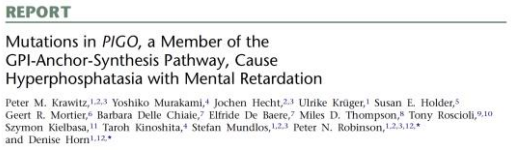
Exemples

Analyse d'une famille (utilisation de l'outil d'annotation/filtration VarAFT)

- Trio avec une fille malade et 2 parents sains
- Syndrome de Mabry : déficience intellectuelle, dysmorphie faciale, hyperphosphatémie …
- Application de 4 filtres :
  - Transmission (AR)
    - ➛ Homozygote ou Hétérozygote composite
  - Localisation & type
  - Fréquence faible dans la population
  - Prédiction de la pathogénicité

  (UMD-Predictor et Human Splicing Finder)





REPORT

Mutations in *PIGO*, a Member of the GPI-Anchor-Synthesis Pathway, Cause Hyperphosphatasia with Mental Retardation

Peter M. Krawitz,[1,2,3] Yoshiko Murakami,[4] Jochen Hecht,[2,3] Ulrike Krüger,[1] Susan E. Holder,[5] Geert R. Mortier,[6] Barbara Delle Chiaie,[7] Elfride De Baere,[7] Miles D. Thompson,[8] Tony Roscioli,[9,10] Szymon Kielbasa,[11] Taroh Kinoshita,[4] Stefan Mundlos,[1,2,3] Peter N. Robinson,[1,2,3,12,*] and Denise Horn[1,12,*]

Analyse d'une autre famille (comparaison des outils de prédiction de la pathogénicité)

ARTICLE

Disruption of POF1B Binding to Nonmuscle Actin Filaments Is Associated with Premature Ovarian Failure

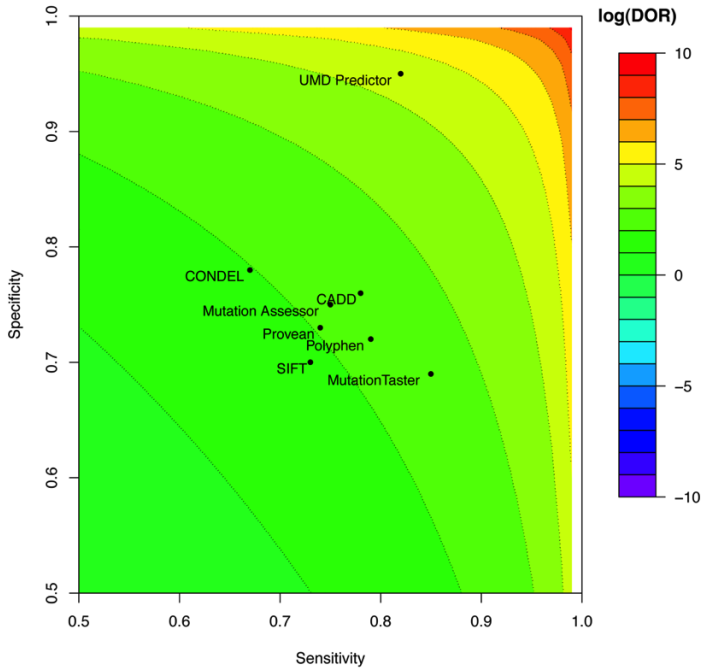
Arnaud Lacombe,* Hane Lee,* Laila Zahed, Mahmoud Choucair, Jean-Marc Muller, Stanley F. Nelson, Wael Salameh, and Eric Vilain

Outil	Valeur	Conclusion
UMD Predictor	66	Pathogène
SIFT	0,05	Pathogène
Polyphen 2 HDIV	0,392	Non pathogène
Polyphen 2 HVAR	0,027	Non pathogène
LRT	0,000	Neutre
MutationTaster	0,176	Pathogène Automatique
Mutation Assessor	1,355	Neutre
Provean	-1,87	Neutre
M-Cap	0,095	Pathogène
CADD	23,5	Pathogène
DANN	0,999	Pathogène

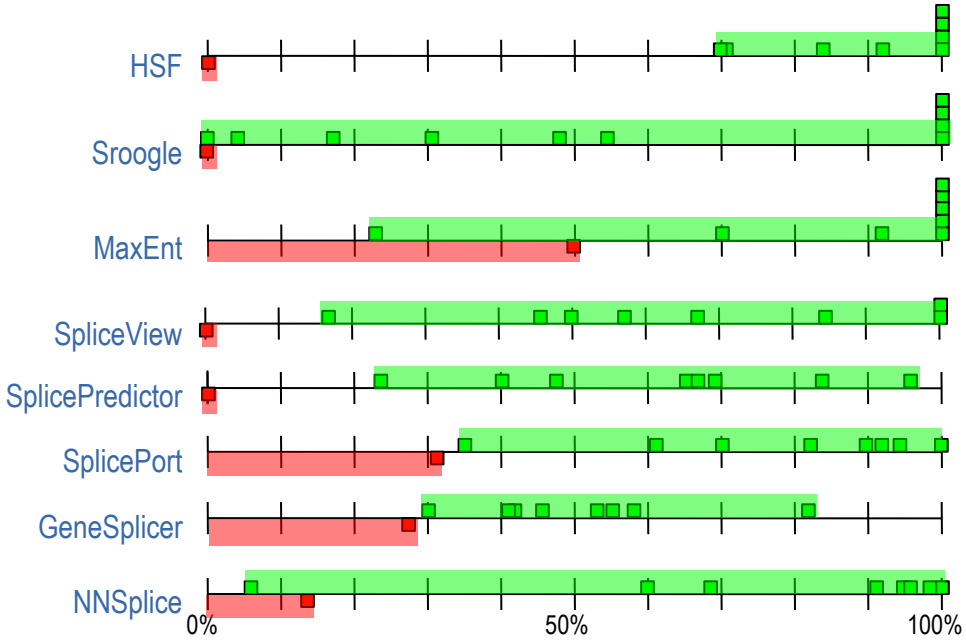
Notre recommandation : limiter le nombre de prédicteurs utilisés
Uniquement UMD-Predictor pour SNP et HSF pour épissage

"le mieux est le mortel ennemi du bien", Montesquieu

UMD-Predictor (umd-predictor.eu) et Human Splicing Finder (umd.be/HSF3/)



UMD-Predictor



Human Splicing Finder (HSF)

Les limites des outils actuels

Les limites

1- Identification des variants

- Choix du génome de référence (HG19, HG19 decoy5, HG38 …)
- Choix du pipeline de traitement
- Différence d'efficacité en fonction
  - Technologie de séquençage (Illumina, Proton, Nanopore, PacBio …)
  - De la région (séquences répétées, régions riches en GC)
  - De la nature des mutations (ins/del, CNV, mutations somatiques)

Les limites

2- Annotation

- Nécessite la manipulation de nombreuses sources de données
- Des annotations incorrectes ou incomplètes peuvent aboutir à des conclusions erronées (faux positifs)
- Il est intéressant de restreindre l'analyse à certains transcrits lorsqu'ils sont spécifiques du tissu d'intérêt
- Fréquence dans la population générale : est-elle la bonne référence ?

Les limites

3- Filtration

- Il est impératif d'avoir une bonne description clinique du patient
- Le mode de transmission doit également être bien évalué
- Il n'existe pas de "Gold Standard", mais :
    - Fréquence dans la population de référence (à bien choisir)
    - Génotype (fonction du mode de transmission)
    - Type de mutation / pathogénicité
    - Appliquer les filtres pas à pas
- Attention à l'utilisation de multiples outils de prédiction
- Attention à la protection des données
    - Données génétiques et cliniques = données sensibles
    - S'assurer de la protection des données (utilisation des systèmes on-line)

Les limites

- Taux de succès du NGS : **25 à 40% des cas** résolus
- Devrait ↗ avec l'utilisation systématique du **WGS**
- Connaissances biologiques

- **Filtration** : formuler les bonnes hypothèses
- **Coût** : séquençage, calcul, stockage, bases de données
- **Ethique** et **partage de données** : données sensibles

Le futur

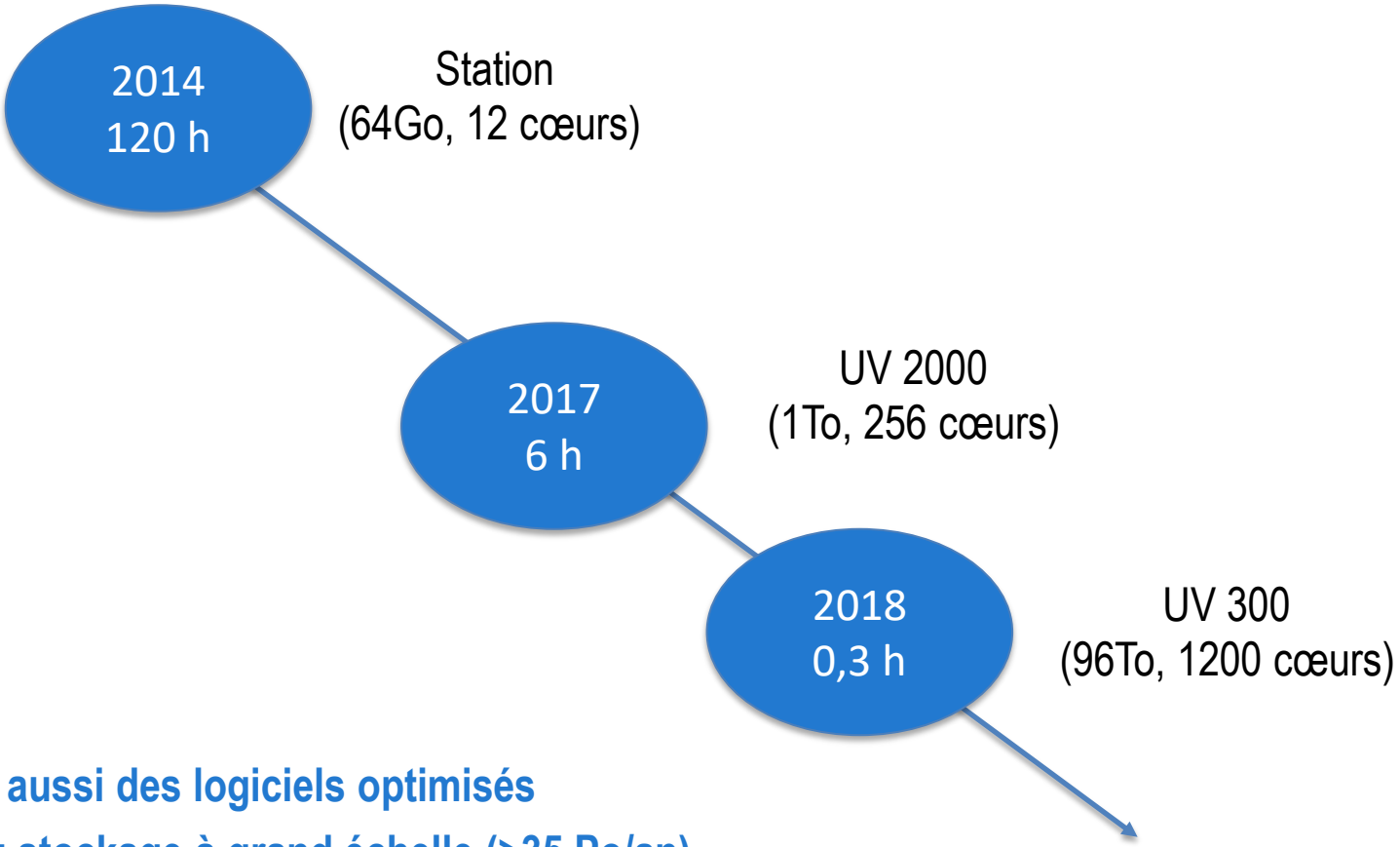
Les challenges du futur

- Augmenter le taux de résolution des analyses
  - Maladies rares monogéniques (AR, AD, XL, *de novo*, mosaiques)
  - Oncologie : détecter le plus tôt possible les mutations (seuil de détection)

- Analyser l'ensemble des données
  - Gènes modificateurs
  - Médecine personnalisée

- Maladies complexes
  - Polygénisme
  - Mutations à effet limité

Des ordinateurs plus puissants



Mais aussi des logiciels optimisés

Et du stockage à grand échelle (>35 Po/an)