

# Outils génomiques appliqués au diagnostic

Détection de variants post-séquençage

AGAGGTATCGCCCCTGTAAATGTAGA  
AGAGGTATCGCCCCTCTAATGTAGA



# 1. Intro - séquençage du génome humain

**Il y a 30 ans ...**

« I expect that within a few years, our technology will be able to sequence one megabase/technician-year. At that rate 100 technicians could sequence the genome in 30 years. »

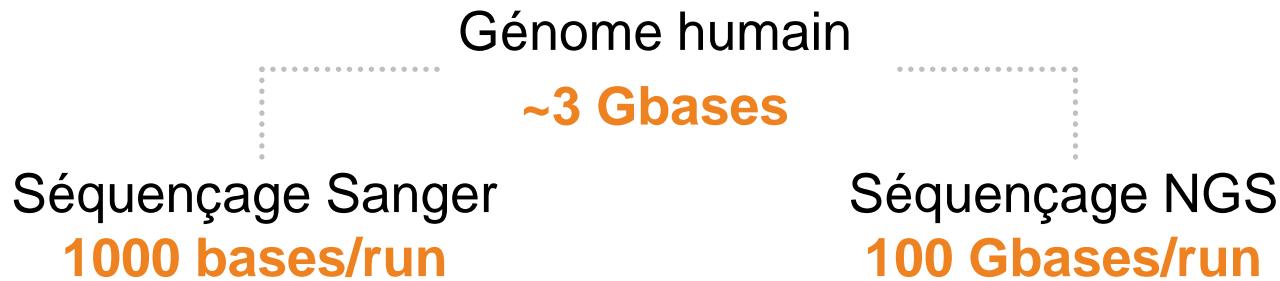
Harvard Nobel Laureate Walter Gilbert 1980  
The Scientist. October 20. 1986

# 1. Intro - séquençage du génome humain

## Apparition du Next Generation Sequencing

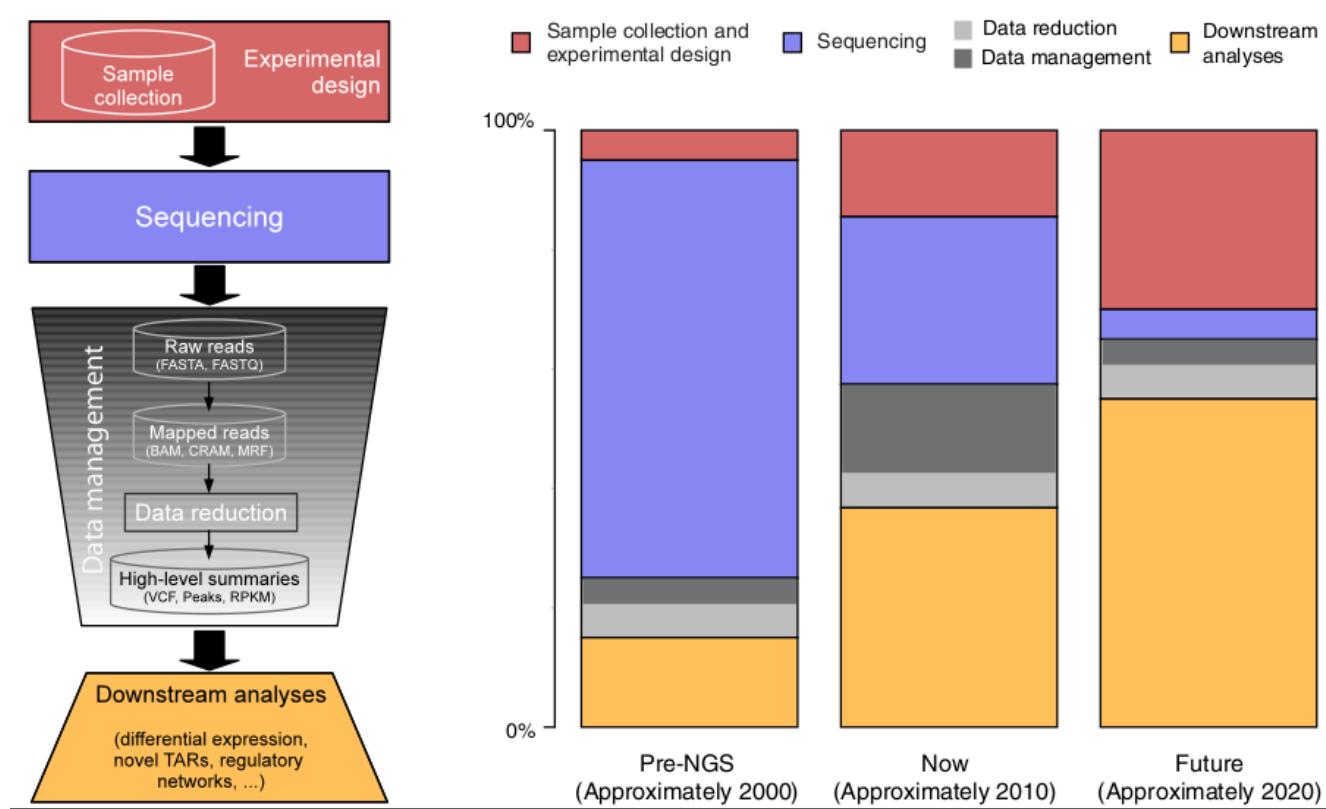
plus rapide, plus de matériel séquencé, moins cher !

Permet l'analyse des séquences nucléiques à une autre échelle en générant des millions de séquences, d'échantillons et de génomes !



# 1. Intro - séquençage du génome humain

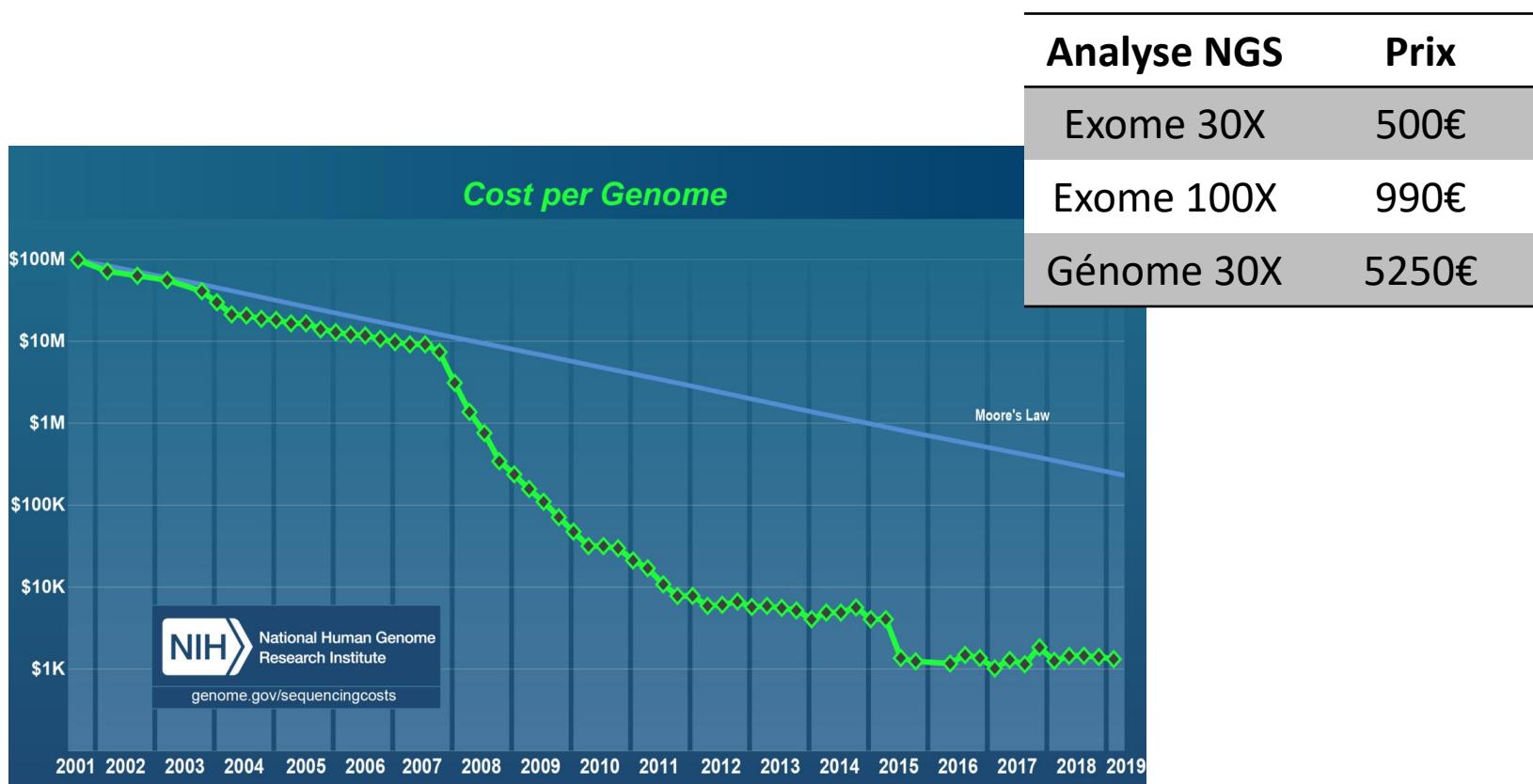
## Evolution des coûts de séquençage



Sboner *et al.* 2011. Genome Biology, 12:125

# 1. Intro - séquençage du génome humain

Coûts de séquençage + faibles que prévus !



# 1. Intro - séquençage du génome humain

## Challenge actuel

Faible coût de séquençage

+

Facilité d'accès à ces nouvelles technologies  
pour les laboratoires



Avalanche de données à analyser

# 1. Intro - séquençage du génome humain



# 1. Intro - séquençage du génome humain

## Analyse NGS et bioinformatique

“The rule of thumb in the genomics community is that every dollar spent on sequencing hardware must be matched by a comparable investment in informatics”

<http://www.the-scientist.com/?articles.view/articleNo/30731/title/Sequence-Analysis-101/>

## 2. Méthodologie pour la détection des variants

### 1. Echantillonnage

Détection de mutations constitutionnelles

- Sang sur tube EDTA
- Salive

Détection de mutations somatiques

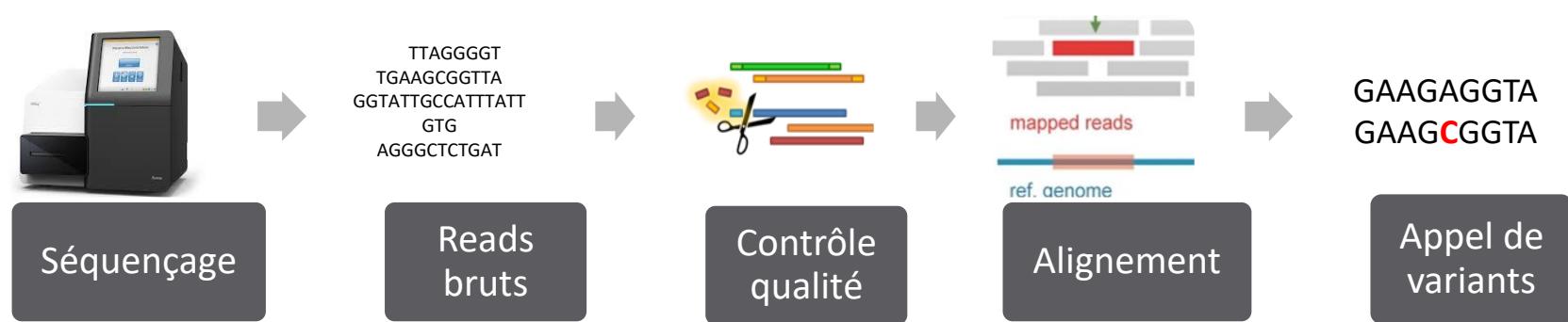
- Tissu tumoral : congelé ou paraffiné
- Sang, ADN tumoral circulant

### 2. Extraction ADN et séquençage

- Fragmentation de l'ADN et construction des librairies (~200b)
- Différentes tailles de cibles: panel, exome (10Kb à 30Mb)
- Différentes technologies : amplicon / capture

## 2. Méthodologie pour la détection des variants

### Pipeline d'analyse bioinformatique



## 2. Méthodologie pour la détection des variants

### Qu'appelle-t-on variant ?

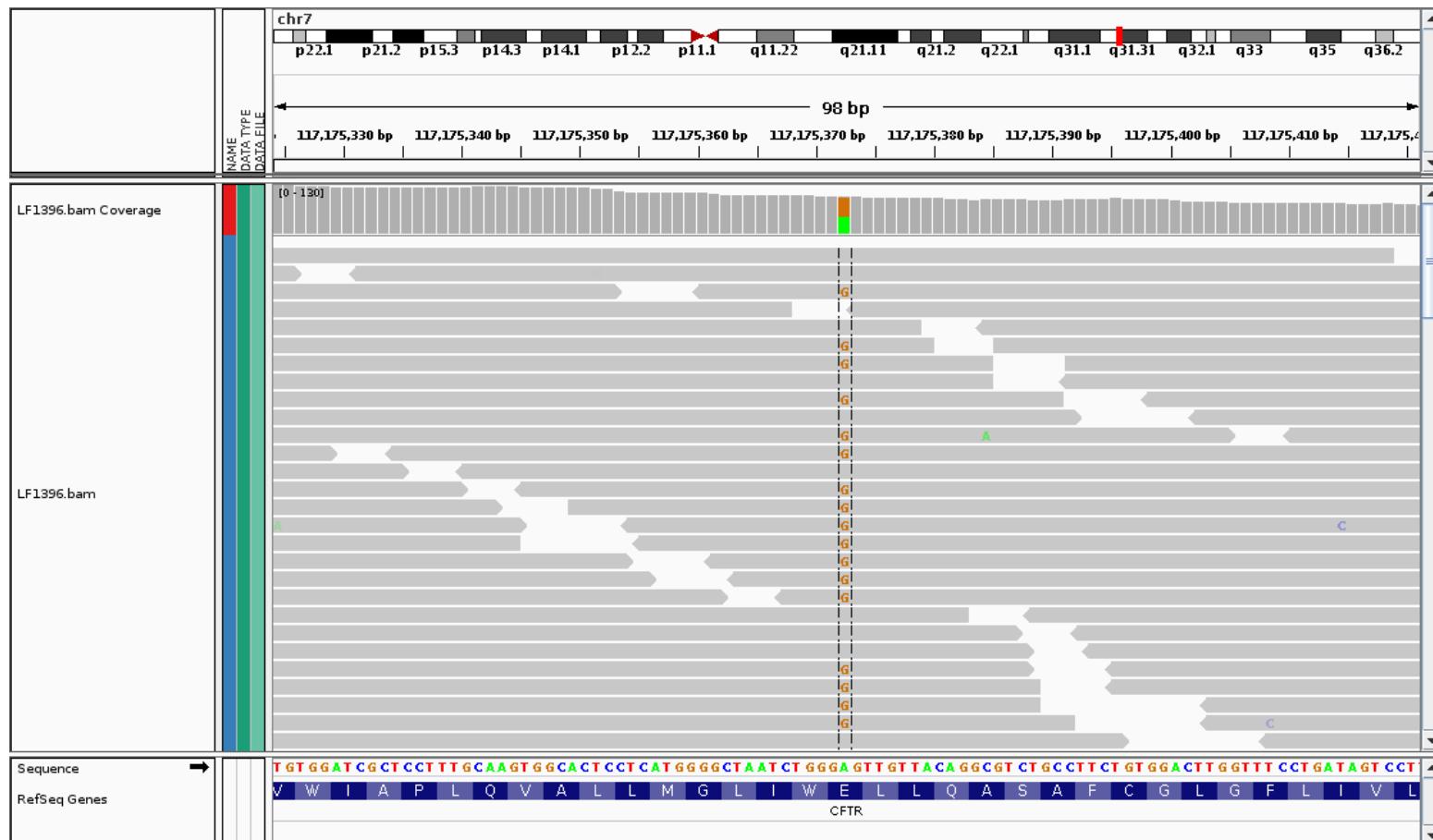
→ Variation génomique dans une séquence nucléique, par rapport à une séquence de référence

Différentes catégories de variants :

- **SNV** : Single Nucleotide Variant
- **INDEL** : INsertion ou DEletion d'une ou plusieurs bases
- **MNV** : Multi Nucleotide Variant, bloc de SNVs et/ou INDELS
- **SV** : Structural Variant, réarrangement génomique > 50bp

## 2. Méthodologie pour la détection des variants

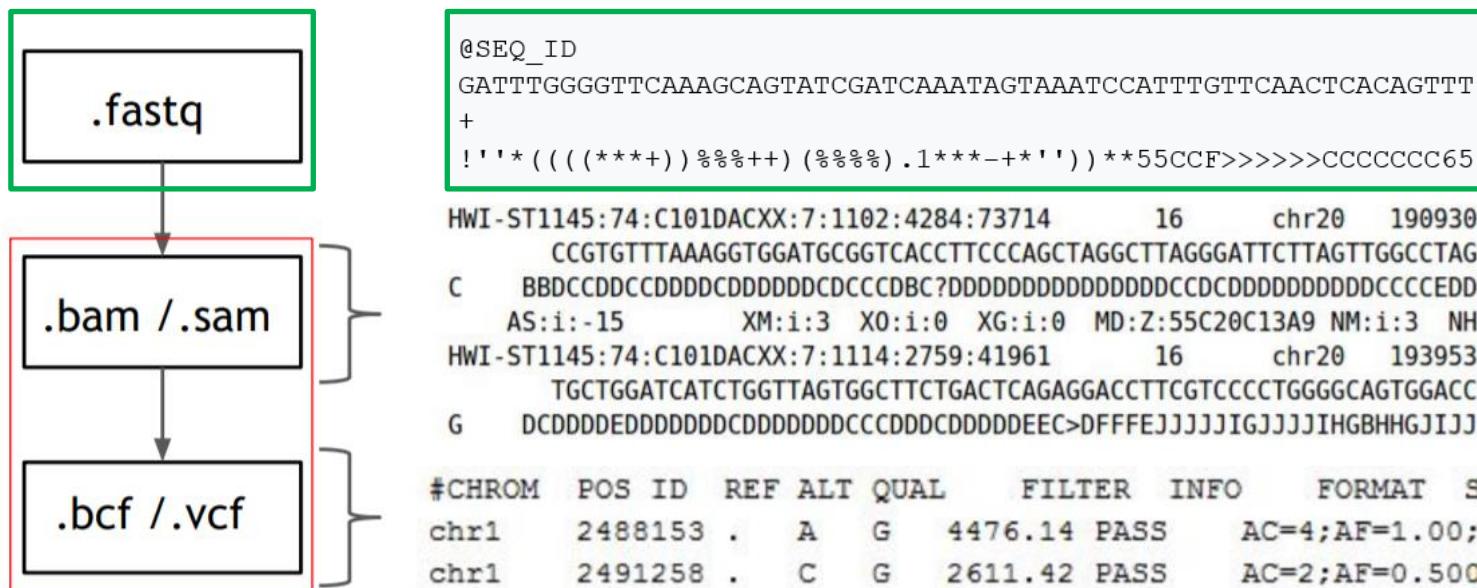
### Visualisation d'un variant (IGV)



## 2. Méthodologie pour la détection des variants

### Qu'est ce qu'un « variant calling » ?

→ Détection automatisée des variants (SNV, INDEL de petites tailles) à partir d'un fichier contenant des données de séquençage alignées.



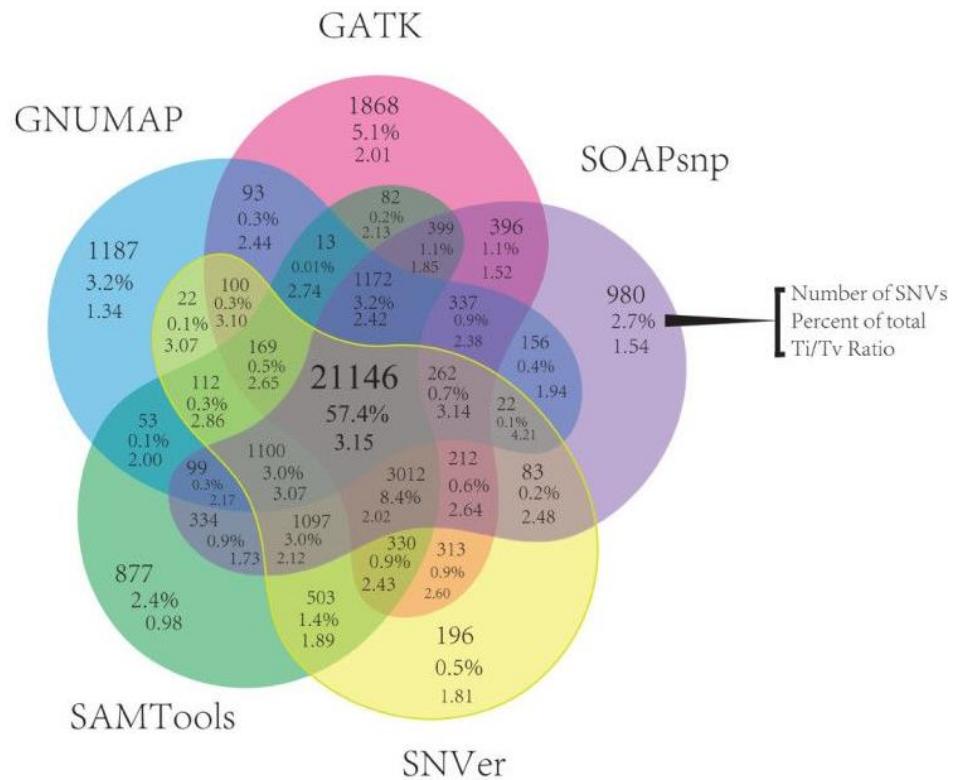
## 2. Méthodologie pour la détection des variants

Chaque outil a ses spécificités

+ de 150 outils de variant calling répertoriés

Faible concordance entre les outils de variant calling

Existence de variants spécifiques à chaque outil



O'Rawe J et al. 2013. Genome Med., 5:28

## 2. Méthodologie pour la détection des variants

### Comment choisir le meilleur outil ?

		Reference variant set	
		Positive	Negative
Variants Called by the Algorithm	Positive	True Positive (TP) Correct variant allele or position call	False Positive (FP) Incorrect variant allele or position call.
	Negative	False Negative (FN) Incorrect reference genotype or no call.	True Negative (TN) Correct reference genotype or no call.

→ Performance de détection

#### Sensibilité

- Mesure la capacité de l'outil à détecter le maximum de véritables variants
- $TP / (TP + FN)$

#### Précision

- Mesure la capacité de l'outil à ne pas détecter de faux variants
- $TN / (TN + FP)$

Olson *et al.* 2015. Front. Genet., 6:235

## 2. Méthodologie pour la détection des variants

### Amélioration de l'alignement

L'appel des indels sur les alignements initiaux peuvent présenter des taux de faux-positifs et faux-négatifs élevés

Pour améliorer la détection des indels : utilisation d'algorithmes plus sophistiqués de réalignement local (ex: GATK )

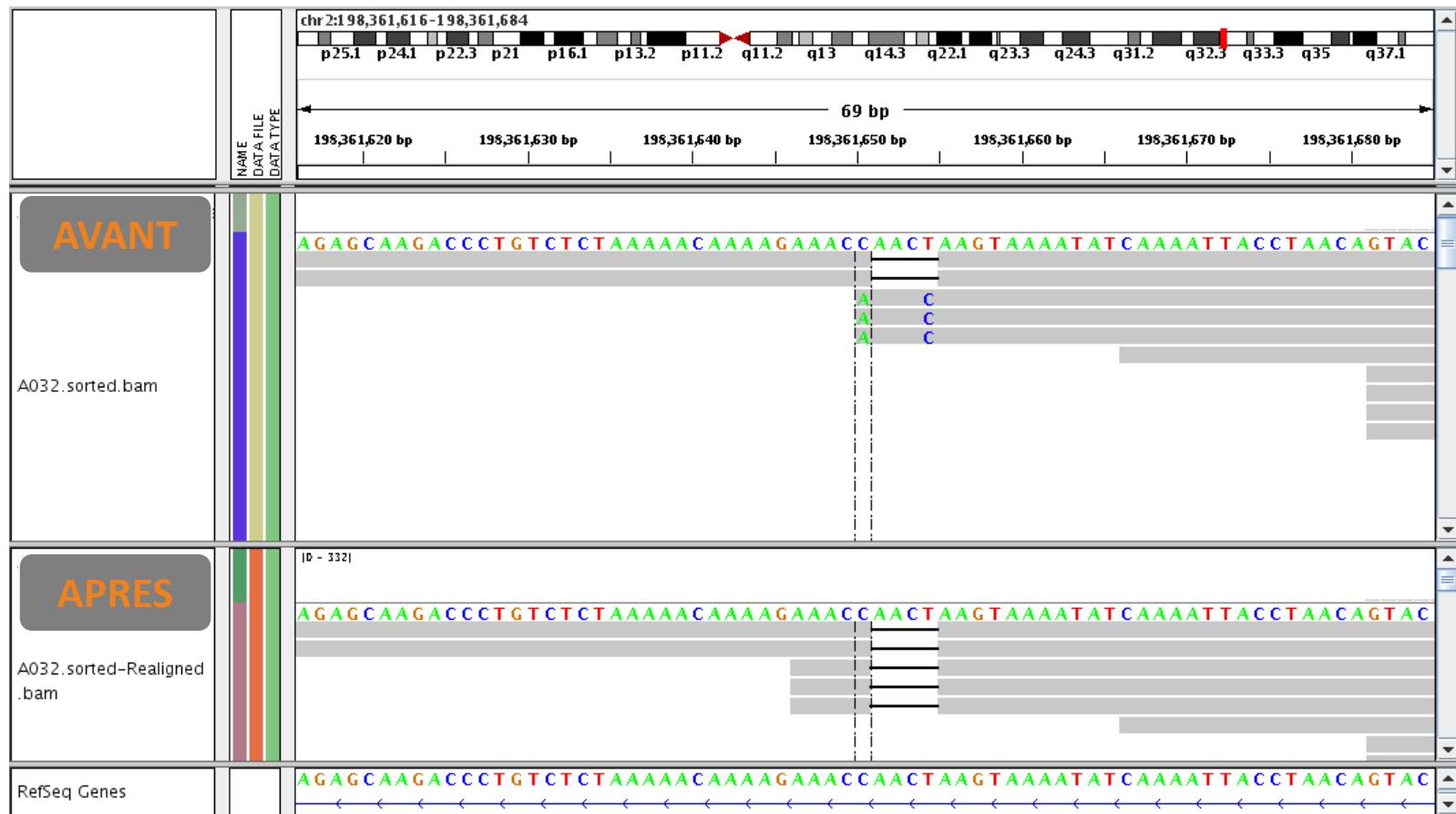
#### Limitation :

Informatiquement intensif si chaque indel possible est testée

#### Solution :

Test uniquement sur un jeu d'indels connues

## 2. Méthodologie pour la détection des variants



## 2. Méthodologie pour la détection des variants

### Validation des variants ?

Liste colossale de variants (surtout pour l'analyse d'exomes)

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Z4R1	
chr1	880933	.	G	C	22	.	DP=29;VDB=0.0363;AF1=0.5;AC1=1;DP4=5,2,21,1;MQ=42;F0=25;PV4=0.14,8.8e-14,0.013,1	GT		
chr1	881593	.	C	T	38	.	DP=31;VDB=0.0138;AF1=0.5;AC1=1;DP4=3,17,2,8;MQ=27;F0=41;PV4=1,1.3e-06,0.27,1	GT:PL:GQ		
chr1	881603	.	G	A	19.1	.	DP=27;VDB=0.0264;AF1=0.5;AC1=1;DP4=4,10,1,7;MQ=28;F0=22;PV4=0.61,2.4e-11,1,1	GT:PL:GQ		
chr1	883530	.	A	G	22	.	DP=4;VDB=0.0427;AF1=1;AC1=2;DP4=0,0,1,2;MQ=33;F0=-36	GT:PL:GQ	1/1:54,9,0:15	
chr1	883961	.	T	C	27	.	DP=10;VDB=0.0218;AF1=0.5001;AC1=1;DP4=0,1,4,0;MQ=39;F0=8.63;PV4=0.2,0.00024,0.081,1	GT		
chr1	886591	.	C	T	72.1	.	DP=6;VDB=0.0454;AF1=1;AC1=2;DP4=0,0,3,3;MQ=37;F0=-45	GT:PL:GQ	1/1:105,18,0:33	
chr1	887446	.	G	A	83	.	DP=17;VDB=0.0410;AF1=0.5;AC1=1;DP4=3,4,4,4;MQ=42;F0=86;PV4=1,3.3e-05,0.026,1	GT:PL:GQ		
chr1	887487	.	A	G	9.52	.	DP=19;VDB=0.0313;AF1=0.5;AC1=1;DP4=5,6,0,6;MQ=41;F0=12,3;PV4=0.1,0.00016,0.0014,0.39	GT		
chr1	887817	.	C	T	40	.	DP=25;VDB=0.0363;AF1=0.5;AC1=1;DP4=4,0,2,6;MQ=37;F0=43;PV4=0.061,1.9e-16,0.37,0.1	GT		
chr1	887881	.	G	A	3.01	.	DP=21;VDB=0.0490;AF1=0.4997;AC1=1;DP4=5,3,3,1;MQ=40;F0=4.77;PV4=1,0.0011,1,0.062	GT		
chr1	887905	.	A	G	90	.	DP=21;VDB=0.0454;AF1=0.5;AC1=1;DP4=5,3,4,7;MQ=41;F0=93;PV4=0.37,3.6e-05,0.03,1	GT:PL:GQ		
chr1	887967	.	G	A	78	.	DP=30;VDB=0.0218;AF1=0.5;AC1=1;DP4=1,14,9,5;MQ=41;F0=81;PV4=0.0017,1.3e-10,0.00052,0.21	GT		
chr1	887982	.	TGCG	TTGAGGACGGCG	43.5	.	INDEL;D=0.0363;AF1=0.5;AC1=1;DP4=0,1,2,3,4;MQ=42;F0=46;PV4=0.037,1,2.4e-0			
chr1	888645	.	A	T	39	.	DP=9;VDB=0.0029;AF1=0.5;AC1=1;DP4=0,2,3,1;MQ=41;F0=37,9;PV4=0.4,0.0039,0.019,0.3	GT		
chr1	889174	.	G	A	30	.	DP=9;VDB=0.0106;AF1=1;AC1=2;DP4=0,0,1,2;MQ=31;F0=-36	GT:PL:GQ	1/1:62,9,0:16	
chr1	889215	.	C	T	7.8	.	DP=14;VDB=0.0026;AF1=0.5;AC1=1;DP4=2,1,3;MQ=30;F0=10.4;PV4=1,3.4e-06,0.023,1	GT:PL:GQ		
chr1	889227	.	G	A	21	.	DP=12;VDB=0.0313;AF1=0.5002;AC1=1;DP4=1,0,0,4;MQ=38;F0=4,75;PV4=0.2,0.0056,0.14,1	GT		
chr1	892325	.	ATTCTCTTTC	ATTCTTC	55.5	.	INDEL;D=26;VDB=0.0321;AF1=0.5;AC1=1;DP4=16,6,4,2;MQ=40;F0=58.5;PV4=1,0.0006,0.0032,1	GT:PL:GQ		
chr1	892552	.	A	T	186	.	DP=47;VDB=0.0535;AF1=0.5;AC1=1;DP4=9,23,3,9;MQ=46;F0=189;PV4=1,0.0006,0.0032,1	GT:PL:GQ		
chr1	909768	.	A	G	5.29	.	DP=9;VDB=0.0539;AF1=1;AC1=2;DP4=0,0,2,0;MQ=46;F0=-33	GT:PL:GQ	1/1:35,6,0:6	
chr1	982941	.	T	C	8.44	.	DP=10;VDB=0.0264;AF1=1;AC1=2;DP4=0,0,1,1;MQ=31;F0=-33	GT:PL:GQ	1/1:39,6,0:8	
chr1	1158631	.	A	G	222	.	DP=11;VDB=0.0518;AF1=1;AC1=2;DP4=0,0,7,3;MQ=45;F0=-57	GT:PL:GQ	1/1:255,30,0:57	
chr1	1246004	.	A	G	174	.	DP=23;VDB=0.0313;AF1=1;AC1=2;DP4=0,0,11,9;MQ=43;F0=-87	GT:PL:GQ	1/1:207,60,0:99	
chr1	1254841	.	C	G	3.98	.	DP=18;VDB=0.0518;AF1=1;AC1=2;DP4=0,0,0,2;MQ=43;F0=-33	GT:PL:GQ	1/1:33,6,0:5	
chr1	1262639	.	C	G	6.02	.	DP=2;VDB=0.0133;AF1=1;AC1=2;DP4=0,0,0,2;MQ=25;F0=-33	GT:PL:GQ	1/1:36,6,0:6	
chr1	1289366	.	ACT	A	41.4	.	INDEL;D=3;VDB=0.0108;AF1=1;AC1=2;DP4=0,0,3,0;MQ=36;F0=-43.5	GT:PL:GQ	1/1:81,9,0:0	
chr1	1323143	.	CCT	C	4.91	.	INDEL;D=4;VDB=0.0274;AF1=1;AC1=2;DP4=0,0,1,1;MQ=34;F0=-40.5	GT:PL:GQ	1/1:42,6,0	
chr1	1480227	.	G	C	13.2	.	DP=18;VDB=0.0014;AF1=0.5;AC1=1;DP4=7,1,6,0;MQ=39;F0=16.1;PV4=1,0.0002,0.00014,1	GT:PL:GQ		
chr1	1647893	.	CTTTCTT	CTTTCTTTCTT	159	.	INDEL;D=47;VDB=0.0454;AF1=0.5;AC1=1;DP4=2,22,2,11;MQ=45;F0=162;PV4=0.6,1,			
chr1	1650787	.	T	C	155	.	DP=79;VDB=0.0410;AF1=0.5;AC1=1;DP4=15,24,25,10;MQ=40;F0=158;PV4=0.0055,1.5e-25,0.061,1	GT		
chr1	1650920	.	G	A	59	.	DP=39;VDB=0.0264;AF1=0.5;AC1=1;DP4=5,19,1,12;MQ=47;F0=62;PV4=0.39,2.5e-25,0.00013,0.33	GT		
chr1	1653004	.	T	C	40.5	.	DP=23;VDB=0.0080;AF1=1;AC1=2;DP4=0,0,0,4;MQ=41;F0=-39	GT:PL:GQ	1/1:73,12,0:21	
chr1	1653069	.	Agttccattatgtatcttcatgcagtatct				AGTTCCATTATGTATCTCCATGCAGTGATCTgttccattatgtatcttcatgcagtatct	19.8	.	IN
chr1	1654004	.	ATT	ATTTGAATGTTTTGTTAATTTT,ATTTTT,ATTTTGTAAATTTT	22.5	.	INDEL;D=27;VDB=0.0175;AF1=1;AC1=2;DP4=0,0			
chr1	1654063	.	C	G	135	.	DP=34;VDB=0.0454;AF1=0.5;AC1=1;DP4=4,1,15,11;MQ=29;F0=10.4;PV4=0.62,1.2e-08,0.21,1	GT		
chr1	1654065	.	AGCG	A	171	.	INDEL;D=31;VDB=0.0106;AF1=0.5;AC1=1;DP4=6,11,3,11;MQ=42;F0=174;PV4=0.46,1.8e-13,2.7e-07,0			
chr1	1686943	.	T	C	44.1	.	DP=7;VDB=0.0059;AF1=1;AC1=2;DP4=0,0,0,6;MQ=45;F0=-45	GT:PL:GQ	1/1:77,18,0:33	
chr1	1887019	.	A	G	3.41	.	DP=8;VDB=0.0029;AF1=1;AC1=2;DP4=0,0,2,0;MQ=32;F0=-33	GT:PL:GQ	1/1:32,6,0:4	
chr1	1890580	.	T	A	4.13	.	DP=12;VDB=0.0304;AF1=0.5001;AC1=1;DP4=2,0,2,0;MQ=36;F0=3.48;PV4=1,1.0,35,1	GT:PL:GQ		
chr1	1900024	.	TG	T	65.5	.	INDEL;D=7;VDB=0.0014;AF1=0.5;AC1=1;DP4=2,0,5,0;MQ=47;F0=9.94;PV4=1,0.38,0.15,1	GT:PL:GQ		
chr1	1900106	.	TCT	TCTCCT	94.5	.	INDEL;D=21;VDB=0.0539;AF1=0.5;AC1=1;DP4=10,3,5,1;MQ=47;F0=97.5;PV4=1,1.0,0.0064,1	GT		
chr1	1900232	.	T	C	90	.	DP=17;VDB=0.0410;AF1=0.5;AC1=1;DP4=2,7,0,7;MQ=40;F0=93;PV4=0.47,6.3e-05,0.0062,1	GT:PL:GQ		
chr1	1916890	.	C	T	48	.	DP=13;VDB=0.0521;AF1=0.5;AC1=1;DP4=6,1,6;MQ=40;F0=51;PV4=1,2e-08,1,0,26	GT:PL:GQ		



## 2. Méthodologie pour la détection des variants

### Validation des variants ?

Liste colossale de variants (surtout pour l'analyse d'exomes)

De nombreux faux positifs parmi les variants trouvés :

- Artéfacts issus des cycles PCR
- Artéfacts spécifiques à la technologie de séquençage
- Erreurs de lecture lors du “Base Calling”
- Mauvaise qualité de l’alignement (régions répétées)

Ordonner les variants trouvés (ex: si cible thérapeutique)



Annotation et tri des variants

### 3. Nomenclature de la description de variants

## Les bonnes pratiques de l'annotation

→ convention de nommage pour échanger/comparer les informations sur les variants de séquences nucléiques.



<https://varnomen.hgvs.org/>

The screenshot shows the homepage of the Sequence Variant Nomenclature website. At the top, there is a navigation bar with links: "Sequence Variant Nomenclature", "Recommendations", "Background Materials", "Recent Additions", "Contact Us", and "Version 19.01". The main title "Sequence Variant Nomenclature" is displayed prominently. Below it, a question "What is the sequence variant nomenclature?" is asked, followed by a detailed explanation: "These pages summarise HGVS-nomenclature: the recommendations for the description of sequence variants. HGVS-nomenclature is used to report and exchange information regarding variants found in DNA, RNA and protein sequences and serves as an international standard. When using the recommendations please cite: *HGVS recommendations for the description of sequence variants - 2016 update, Den Dunnen et al. 2016, Hum.Mutat. 37:564-569*. HGVS-nomenclature is authorised by the Human Genome Variation Society (HGVS), the Human Variome Project (HVP) and the HUman Genome Organization (HUGO).".

#### Current Recommendations

General	DNA	RNA
Protein	Uncertain	Checklist

### 3. Nomenclature de la description de variants

#### Un exemple d'annotation simple ...

En sortie de variant calling :

Chr	Position	Ref	Alt
chr3	178 952 000	A	G

Transcription de la position génomique du variant en position sur la séquence codante d'un gène :

- Classe de variant : **SNV**
- Gène concerné : **PIK3CA**
- Choix d'un transcrit parmi les transcrits alternatifs : **NM\_006218**
- Position dans l'ADN codant: **c.3055** ou la séquence protéique : **p.1019**

NM\_006218.2:c.3055A>G

NM\_006218.2:p.(Ile1019Val)

### 3. Nomenclature de la description de variants

#### Un exemple d'annotation plus compliqué ...

En sortie de variant calling :

Chr	Position	Ref	Alt
chr13	32 968 809	T	-

Transcription de la position génomique du variant en position sur la séquence codante d'un gène :

- Classe de variant : **délétion**
- Gène concerné : **BRCA2**
- Choix d'un transcrit parmi les transcrits alternatifs : **NM\_000059**
- Position dans l'ADN codant : **à 17 bp de l'exon 25**, pas d'équivalence en p.

NM\_000059.3:c.9257-17del

NM\_000059.3:p.?

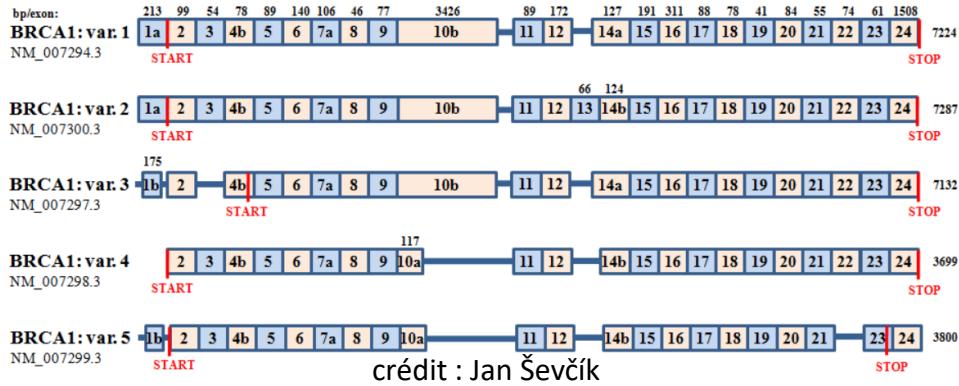
### 3. Nomenclature de la description de variants

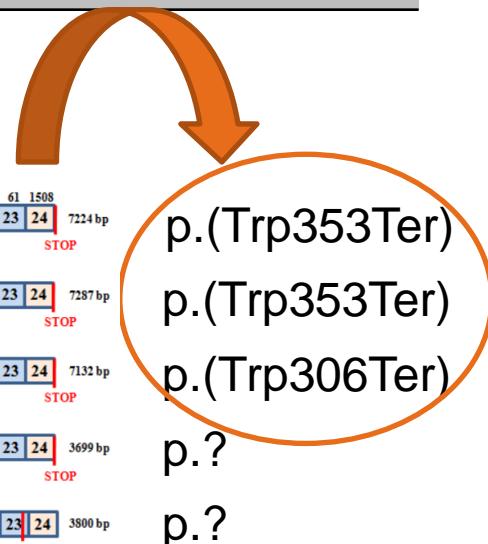
Le choix du bon transcrit est primordial

Chr	Position	Ref	Alt
chr17	41 246 489	C	T

Le gène BRCA1 a 5 transcrits :

- NM\_007294
- NM\_007300
- NM\_007297
- NM\_007299
- NM\_007298



  
p.(Trp353Ter)  
p.(Trp353Ter)  
p.(Trp306Ter)  
p.?  
p.?

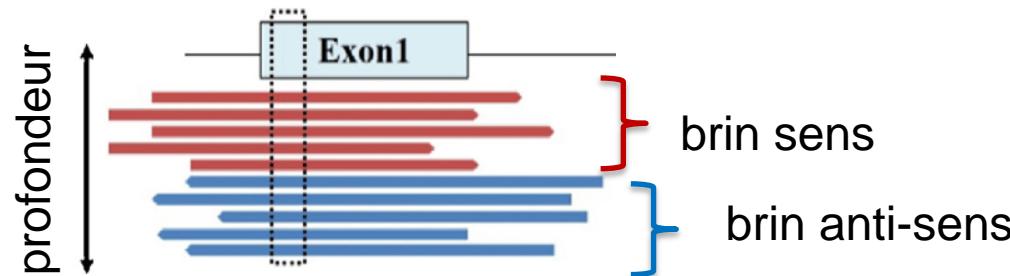
→ seules 3 dénominations sur 5 reflètent l'apparition d'un codon stop

## 4. Tri des variants

### Avec les informations du séquençage

Qualité des variants estimée grâce :

- Au score de qualité phred
  - ✓ un score de 20 → précision de la base estimée à 99%
  - ✓ un score de 30 → précision de la base estimée à 99.9%
- A la couverture à la position étudiée
  - ✓ nombre de reads supportant les bases de référence et alternative
  - ✓ nombre de reads porteurs du variant sur le brin sens et anti-sens



## 4. Tri des variants

### Avec la fréquence dans la population

Variant = polymorphisme si cet allèle a une fréquence égale ou supérieure à 1 % dans la population

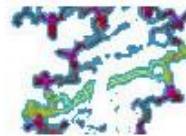
**SNP** : Single-Nucleotide Polymorphism

Base de données qui les répertorie : dbSNP

<http://www.ncbi.nlm.nih.gov/SNP/>



**dbSNP**  
Short Genetic Variations



## 4. Tri des variants

### Avec la fréquence dans la population

#### dbSNP :

- regroupe les polymorphismes de SNVs et de petites indels
- chaque polymorphisme a un rsID

dbSNP v.130 (avril 2009)	dbSNPv.151 (oct. 2017)
3 533 618 rsID <i>(Homo Sapiens)</i>	335 215 764 rsID <i>(Homo Sapiens)</i>

- 100 fois plus de polymorphismes référencés en 8 ans
- soumissions d'un grand nombre de SNPs avec les projets NGS
- croissance exponentielle, mais prudence; contient de faux SNPs

## 4. Tri des variants

# Avec la fréquence dans la population

Have a question about dbSNP? Try searching the SNP FAQ Archive!

Go

**GENERAL**

RSS Feed 

Contact Us

Organism Data

dbSNP Homepage

NCBI Variation Resources

Announcements

dbSNP Summary

FTP Download

**SNP SUBMISSION DOCUMENTATION**

SEARCH

**RELATED SITES**

**dbSNP Summary**

**RELEASE: NCBI dbSNP Build 151**

**dbSNP Component Availability Dates:**

Component	Date available
dbSNP web query for build 151:	Oct 06, 2017
ftp data for build 151:	Oct 06, 2017
Entrez Indexing for build 151:	Oct 06, 2017

- The complete data for build 151 are available at <ftp://ftp.ncbi.nlm.nih.gov/snp/> in multiple formats.  
- All formats and conventions are described in <ftp://ftp.ncbi.nlm.nih.gov/snp/00readme.txt>.  
- Please address any questions or comments regarding the data to [snp-admin@ncbi.nlm.nih.gov](mailto:snp-admin@ncbi.nlm.nih.gov).

**New Submission since previous build:**

Organism	Current Build	New Submissions (ss#'s)	New RefSNP Clusters (rs#'s) (# validated)	New ss# with Genotype	New ss# with Frequency
<a href="#">Homo sapiens</a>	151	<a href="#">893,590,618</a>	335,215,764 (61,743)		
<a href="#">Glycine max</a>	151	<a href="#">4,074,388</a>	846,335 (0)		
<a href="#">Equus caballus</a>	151		(1)		1
<a href="#">Populus trichocarpa</a>	151		2,993,523 (0)		
<a href="#">Oryza sativa</a>	151	<a href="#">232,777</a>	14,108 (0)		
<a href="#">Canis lupus familiaris</a>	151	<a href="#">168</a>	102 (0)		
Total: 6 Organisms		897,897,951	339,069,832 (61,875)		1

## 4. Tri des variants

### Avec la fréquence dans la population

Les données de dbSNP sont à vérifier :

- Aucun critère de fréquence requis pour soumettre un nouveau SNP
- Présence des SNPs de cette base dans le projet 1000 génomes :
  - ✓ sur 39 692 293 positions couvertes
  - ✓ fréquence entre 0.0001 et 100%
  - ✓ problème de curation des données de la base

rs181754315		Current Build 153 Released July 9, 2019
Organism	<i>Homo sapiens</i>	Clinical Significance Not Reported in ClinVar
Position	chr1:51935 (GRCh38.p12)	Gene : Consequence None
Alleles	C>T	Publications 0 citations
Variation Type	SNV Single Nucleotide Variation	Genomic View <a href="#">See rs on genome</a>
Frequency	T=0.000 (1/5008, 1000G)	

## 4. Tri des variants

### Avec la fréquence dans la population

Limites de la base dbSNP :

- la référence dans dbSNP diffère parfois de celle du génome utilisé comme référence
- cas de redondance d'information
- problème de multi-alignement

Pour pallier ces problèmes :

- utilisation de la fréquence des SNPs dans des grands projets du type 1000 génomes, Exome Sequencing Project ...
- utilisation en complément d'autres outils ou bases de données

## 4. Tri des variants

### Avec l'annotation structurale

Variation dans une région codante

- Variation synonyme ?
- Décalage du cadre de lecture (si l'indel n'est pas un multiple de 3) ?
- Implication d'un codon stop (gain ou perte) ?

Variation en dehors des régions codantes

- Dans un intron ?
- Dans un UTR ?
- Dans un promoteur ?
- Sur un site d'épissage - accepteur ou donneur ?
- Dans des régions intergéniques ?

## 4. Tri des variants

### Limites de l'annotation structurale

#### Variations synonymes

- souvent filtrées
- peut supprimer des variants d'intérêt

*A synonymous variation in protease-activated receptor-2 is associated with atopy in Korean children.*

**Lee JH et. al.** J Allergy Clin Immunol. 2011

#### Variants hors des régions codantes

- plus rarement étudiés
- peut avoir d'importantes implications phénotypiques

*A novel mutation in the nerve-specific 5'UTR of the GJB1 gene causes X-linked Charcot-Marie-Tooth disease*

**Murphy SM et al.** J Peripher Nerv Syst 2011

*FA2H-related disorders: a novel c.270+3A>T splice-site mutation leads to a complex neurodegenerative phenotype.*

**Garone C et al.** Dev Med Child Neurol 2011



## 4. Tri des variants

# Avec des bases de données de mutations

Stockage et mise à disposition de mutations somatiques

Variations associées aux informations des projets cancer et de la littérature

The screenshot shows the COSMIC website interface. At the top, there is a navigation bar with links for Projects, Data, Tools, News, Help, About, Genome Version, a search bar, and a login link. The main content area features a banner for 'COSMIC v90, released 05-SEP-19'. Below the banner, there is descriptive text about the project and a search bar. To the right, a box titled 'COSMIC News' contains an article about the release of version 90, featuring a circular logo and a 'Follow @cosmic\_sanger' link. Another section below it discusses the preview of the v90 release.

**COSMIC**  
Catalogue Of Somatic Mutations In Cancer

Projects ▾ Data ▾ Tools ▾ News ▾ Help ▾ About ▾ Genome Version ▾ Search COSMIC... **SEARCH** Login ▾

**COSMIC v90, released 05-SEP-19**

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.

eg Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Camp **SEARCH**

**Projects**

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:

**COSMIC News**

**COSMIC Release v90**

COSMIC version 90 is here. This is a special release for COSMIC because we have focused on updating and upgrading the architecture and systems behind COSMIC, substantially updating and standardising our variant content across current sets of genes, [More...](#)

**Preview the COSMIC v90 release**

The v90 release is almost here. We are currently completing final preparations and tests. We have now prepared files that contain a sample of the complete dataset for each of the COSMIC download files. [More...](#)

## 4. Tri des variants

# Avec des bases de données de mutations

### References

This section shows publications associated with PIK3CA. You can see more information in our [help pages](#).

Show 10 ▾ entries

Export: [CSV](#) [TSV](#) Search:

Reference Title	Author	Year	Journal	Status	COSMIC	Pubmed
Genetic alterations of phosphoinositide 3-kinase subunit genes in human glioblastomas	Mizoguchi M et al	2004	Brain pathology (Zurich, Switzerland);14(4):372-7	Curated	<a href="#">COSP8873</a>	<a href="#">15605984</a>
High frequency of mutations of the PIK3CA gene in human cancers	Samuels Y et al	2004	Science (New York, N.Y.);304(5670):554	Curated	<a href="#">COSP8029</a>	<a href="#">15016963</a>
Mutation of the PIK3CA gene in ovarian and breast cancer	Campbell IG et al	2004	Cancer research;64(21):7678-81	Curated	<a href="#">COSP8848</a>	<a href="#">15520168</a>
Mutations of PIK3CA in anaplastic oligodendroglomas, high-grade astrocytomas, and medulloblastomas	Broderick DK et al	2004	Cancer research;64(15):5048-50	Curated	<a href="#">COSP6149</a>	<a href="#">15289301</a>
PIK3CA gene is frequently mutated in breast carcinomas and hepatocellular carcinomas	Lee JW et al	2005	Oncogene;24(8):1477-80	Curated	<a href="#">COSP8872</a>	<a href="#">15608678</a>
The PIK3CA gene is mutated with high frequency in human breast cancers	Bachman KE et al	2004	Cancer biology & therapy;3(8):772-5	Curated	<a href="#">COSP6155</a>	<a href="#">15254419</a>
Oncogenic mutations of PIK3CA in human cancers	Samuels Y and Velculescu VE	2004	Cell cycle (Georgetown, Tex.);3(10):1221-4	Review	<a href="#">COSP8849</a>	<a href="#">15467468</a>
PI3K: Missense Mutation Motivates Malignancy	Hawthorne VS and Yu D	2004	Cancer biology & therapy;3(8):776-7	Review	<a href="#">COSP15318</a>	<a href="#">15280668</a>
Somatic alterations in the human cancer genome	Weir B et al	2004	Cancer cell;6(5):433-8	Review	<a href="#">COSP8897</a>	<a href="#">15542426</a>
Colorectal cancer: mutations in a signalling pathway	Parsons DW et al	2005	Nature;436(7052):792	Curated	<a href="#">COSP21344</a>	<a href="#">16094359</a>

Showing 1 to 10 of 2,685 entries

First Previous 1 2 3 4 5 ... 269 Next Last



Quelques variations PIK3CA, appuyées par des papiers

## 4. Tri des variants

### Avec des bases de données de mutations

Précision de la base de données COSMIC ?

- Présence de polymorphismes
  - ne devrait pas être dans une base de mutations somatiques
- Base régulièrement mis à jour (par des curateurs)
  - pour faire des ajouts de nouveaux variants
  - mais aussi pour retirer les faux positifs

## 4. Tri des variants

### Avec des outils de prédition d'effet délétère

Algorithmes de prédition:

- prédition de l'impact d'un changement d'acide aminé sur la structure et la fonction de la protéine (ex : SIFT, CADD, ...)
- estimation de la conservation des acides aminés (ex : PhyloP)
- effets sur les transcrits et l'épissage (ex : MaxEntScan)

Des algorithmes différents = des résultats différents

## 4. Tri des variants

### Avec des outils de prédition d'effet délétère

Variations décrites par COSMIC :

- PhyloP : 84% sont sur des sites conservés
- SIFT : 73% sont prédictes délétères
- Polyphen2 : 81% sont prédictes délétères

→ jusqu'à 27% de mauvaises prédictions par les algorithmes individuels.  
→ mais la combinaison des 3 approches prédit un effet délétère dans 95% des variations.



Obtention de bonnes prédictions en combinant les outils

## 4. Tri des variants

# Génération automatique de ces annotations

**ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data**

Kai Wang<sup>1,\*</sup>, Mingyao Li<sup>2</sup> and Hakon Hakonarson<sup>1,3</sup>

<sup>1</sup>Center for Applied Genomics, Children's Hospital of Philadelphia, <sup>2</sup>Department of Biostatistics and Epidemiology and <sup>3</sup>Department of Pediatrics, University of Pennsylvania, Philadelphia, PA 19104, USA

<http://annovar.openbioinformatics.org/en/latest/>



## ANNOVAR Documentation

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, hg38, as well as mouse, worm, fly, yeast and many others). Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, ANNOVAR can perform:

- **Gene-based annotation:** identify whether SNPs or CNVs cause protein coding changes and the amino acids that are affected. Users can flexibly use RefSeq genes, UCSC genes, ENSEMBL genes, GENCODE genes, AceView genes, or many other gene definition systems.
- **Region-based annotation:** identify variants in specific genomic regions, for example, conserved regions among 44 species, predicted transcription factor binding sites, segmental duplication regions, GWAS hits, database of genomic variants, DNase I hypersensitivity sites, ENCODE H3K4Me1/H3K4Me3/H3K27Ac/CTCF sites, ChIP-Seq peaks, RNA-Seq peaks, or many other annotations on genomic intervals.
- **Filter-based annotation:** identify variants that are documented in specific databases, for example, whether a variant is reported in dbSNP, what is the allele frequency in the 1000 Genome Project, NHLBI-ESP 6500 exomes or Exome Aggregation Consortium (ExAC) or Genome Aggregation Database (gnomAD), calculate the SIFT/PolyPhen/LRT/MutationTaster/MutationAssessor/FATHMM/MetaSVM/MetaLR scores, find intergenic variants with GERP++ score<2 or CADD>10, or many other annotations on specific mutations.

## 4. Tri des variants

Avec des bases de données gène-spécifique

Par exemple pour BRCA1



<http://www.umd.be/BRCA1/>

**BRCA Share™ (formerly UMD-BRCA1 mutations database)**

**Home**

Last update 16/06/17

This database includes **11793** records  
(**2495** different variations)

 **Variants Reclassification**  
Last 6 months reclassified variants list

 **Variants Classification**  
Last 6 months annotated variants list

BRCA Share was launched on April 21st 2015

BRCA Share is a novel gene datashare initiative that provides scientists and commercial laboratory organizations around the world with open access to BRCA1 and BRCA2 genetic data. The program's goal is to accelerate research on BRCA gene mutations, particularly variants of uncertain significance, to improve the ability of clinical laboratory diagnostics to predict which individuals are at risk of developing these cancers.

## 4. Tri des variants

### Avec des bases de données locales

Construction et utilisation de bases de données locales pour filtrer des artéfacts dus :

- aux constructions de librairies
- à la technologie de séquençage

→ utilisation des occurrences de variants inter- et intra-runs

# Conclusion

## Bilan sur la recherche de variants

1. Alignement des données de séquençage sur un génome de référence
2. Variant calling, détection des variants par rapport à la référence
3. Annotations structurale et fonctionnelle des variants
  - ✓ Nomenclature HGVS
  - ✓ Vérification du statut de polymorphisme
  - ✓ Recherche dans les bases de données de mutations
  - ✓ Prédiction d'effet délétère
4. Filtration et tri des variants

# Conclusion

## NGS et médecine personnalisée

Petit ensemble de gènes séquencés (**panel**) :

Détection d'altérations génétiques cibles des cancers les plus courants  
(ex : mutations *EGFR*, *KRAS*, *PIK3CA*, *BRCA1/2* ...)

Dans le cas d'analyses exploratoires sur **exome** :

Comparaison avec l'ADN constitutionnel nécessaire pour différentier les variants somatiques des polymorphismes et variants privés

Ces variants permettent de :

- prédire **l'efficacité** d'un traitement (ex : chimiothérapie, thérapie ciblée)
- permettre la prescription d'un traitement **ciblé** quand celui-ci est disponible
- réaliser un suivi dans les cas de détection de **prédisposition**

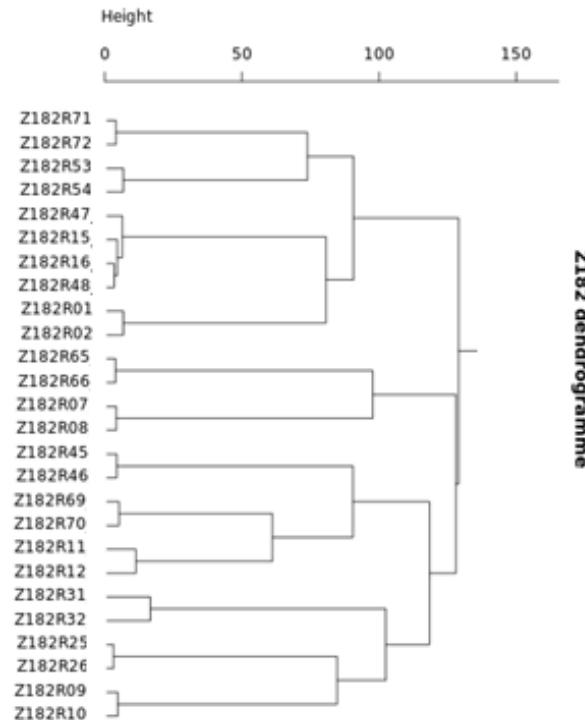
# Autres applications de la détection de variants

## Identitovigilance

Vérification de **l'identité** d'un patient passé en duplicit au sein d'une même analyse ou dans une autre analyse génomique

Clustering basé sur les **ratios alléliques**\*

- d'une liste de variants « hotspot »
- d'une liste de polymorphismes



\* **NB:** un ratio allélique

- est la fréquence à laquelle un variant est observé dans le génome
- est à 0%, 50% ou 100% chez un individu sain (sans mosaïcisme)

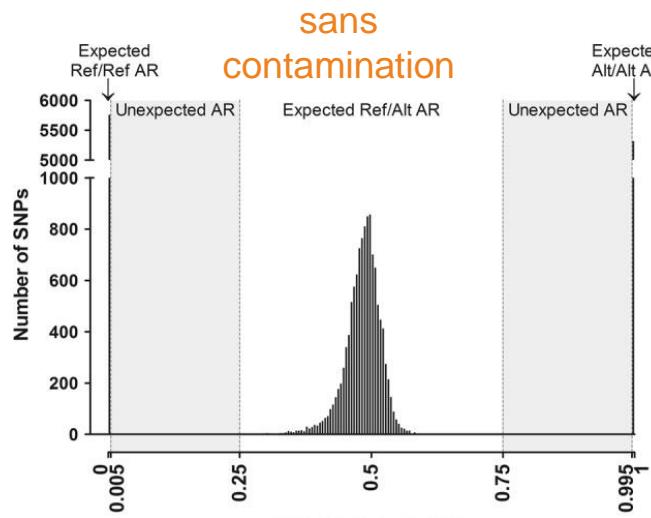
# Autres applications de la détection de variants

## Détection de contaminations

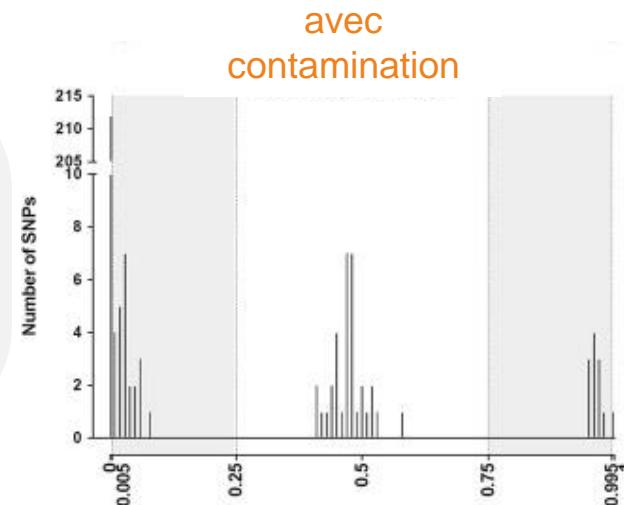
Séquençage haut débit d'échantillons constitutionnels :

- plusieurs patients en parallèle
- risque de contamination intra-analyse

→ Outil pour identifier et quantifier le contaminant



*ART-DeCo: easy tool for detection and characterization of cross-contamination of DNA samples in diagnostic next-generation sequencing analysis*  
Fiévet A et al. Eur J Hum Genet. 2019



# Merci pour votre attention !

Des questions ?

