Atelier Variant niveau 2

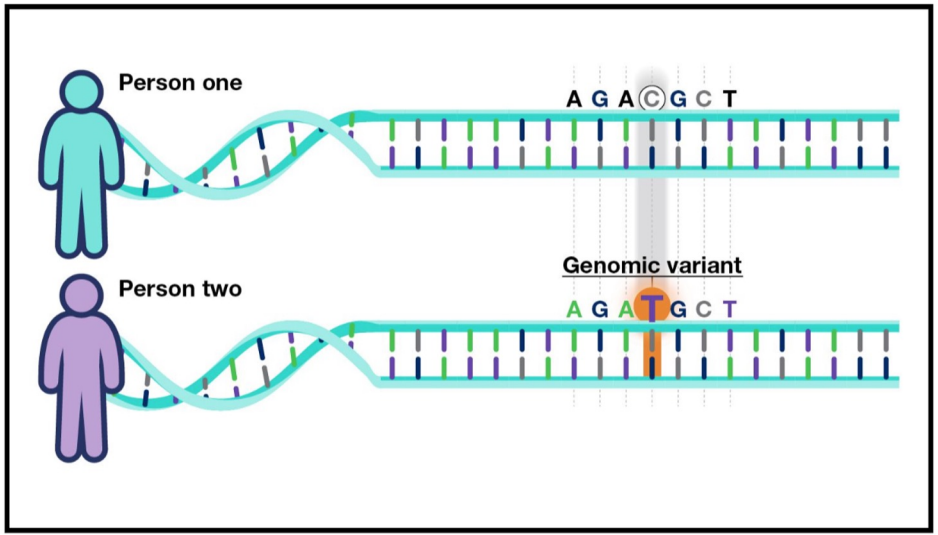
Cathy PHILLIPE - CEA
Nadia BESSOLTANE - INRAe
Pauline FRANCOIS - ANSES

École de bioinformatique niveau 2 IFB-INSERM 2025

Définition

1. Qu'est ce que c'est une variation génomique ?

Une variation génomique est un changement, d'une ou plusieurs bases nucléotides, dans une séquence d'ADN particulière en comparaison avec une séquence d'ADN (un génome) de référence (1). Les variations génomiques se distinguent en deux catégories : polymorphismes et mutations.



Il existe différents types de variations :

- **SNV** : **S**ingle **N**ucleotide **V**ariant
- **INDEL** : **IN**sertion ou **DEL**etion
- **SV** (**S**tructural **V**ariant)
- **CNV** (**C**opy **N**umber **V**ariation)

https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores#one

du Fastq aux VCFs (ebaiin1)

UNIX

Reads (Fastq)

Fastq Quality Control
-- FastQC --

Mapping
-- Bwa --

Reference genome (Fasta)

Processing Post Alignement
-- GATK/Picard --

Structural variants
-- Delly --

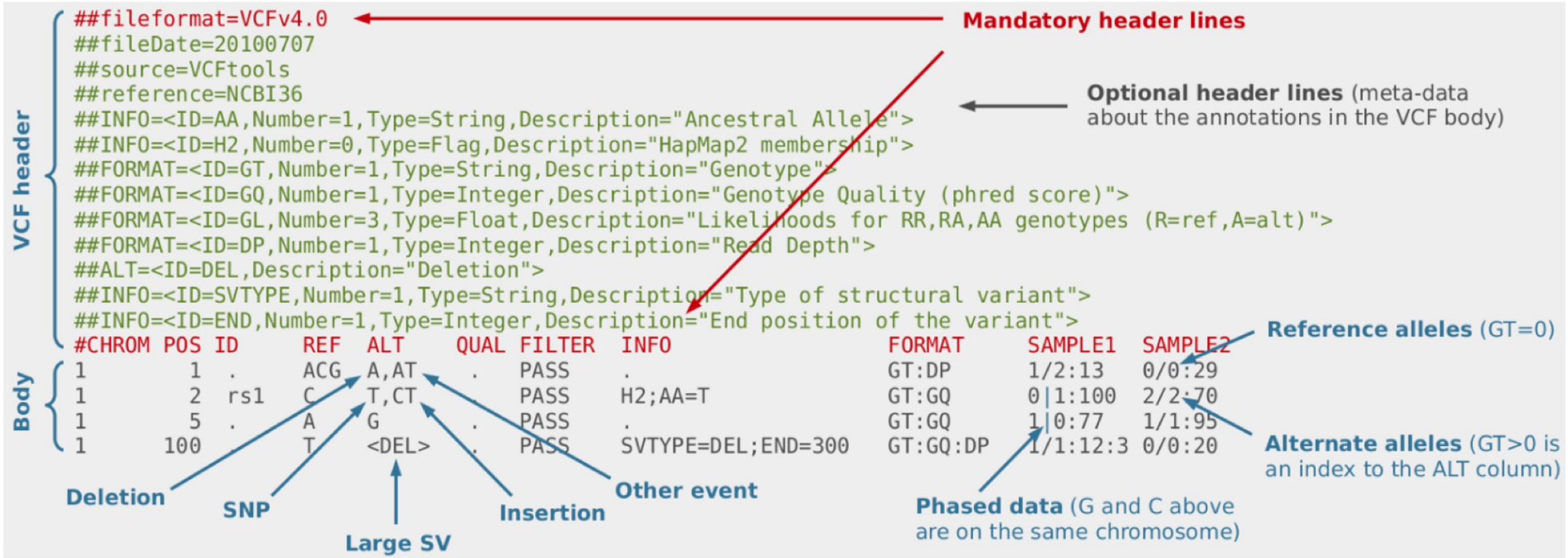
small variants
-- GATK --

Annotation
-- SnpEff --

Variant Calling

VCF

VCF (variant call format)



VCF → 3 parties principales

METADATA

VCF header

```
##fileformat=VCFv4.2
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtere
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VC
##GATKCommandLine=<ID=HaplotypeCaller,CommandLine="HaplotypeCaller  --min-base-quality-score 18 --emit-ref-confidence NONE
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities
...
##INFO=<ID=ReadPosRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt vs. Ref read positio
##INFO=<ID=SOR,Number=1,Type=Float,Description="Symmetric Odds Ratio of 2x2 contingency table to detect strand bias">
##contig=<ID=6,length=119458736>
##source=HaplotypeCaller
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SRR1262731	SRR1262732
6	2	.	T	A	67.64	.	AC=1;AF=0.500;...	GT:AD:DP:GQ:PL	0/1:3,2:5:75:75,0,105	0/1:3,2:5:75:75,
6	4	.	GT	G	58.60	.	AC=1;AF=0.500;...	GT:AD:DP:GQ:PL	0/1:1,2:3:28:66,0,28	0/1:1,2:3:28:66,
6	9	.	C	CA	55.60	.	AC=1;AF=0.500;...	GT:AD:DP:GQ:PL	0/1:7,2:9:63:63,0,279	0/1:7,2:9:63:63,

INFO

GENOTYPE

du Fastq aux VCFs (ebaiin1)

UNIX

Reads (Fastq)

Fastq Quality Control
-- FastQC --

Mapping
-- Bwa --

Reference genome (Fasta)

Processing Post Alignement
-- GATK/Picard --

Structural variants
-- Delly --

small variants
-- GATK --

Annotation
-- SnpEff --

Variant Calling

VCF

du VCFs aux marqueurs (ebaiin2)

QC & Filtres stats descriptives

Rappel : Filtre des variants

- De **nombreux filtres** peuvent être appliqués sur le VCF
  → type de variants à garder (SNVs seulement, Indels...)
  → région d'intérêt
  → filtres sur la qualité (seuils arbitraires : profondeur, génotype (0/1, 1/1), ratio allélique…)

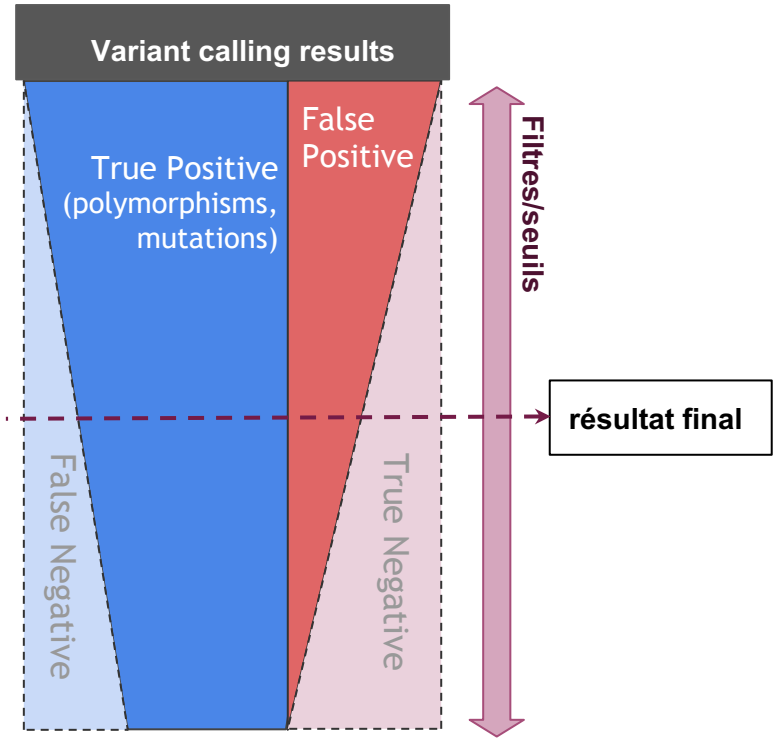
- Filtres difficilement transposables entre analyse :
  → dépendent de la **question biologique**
  → dépendent des outils utilisés

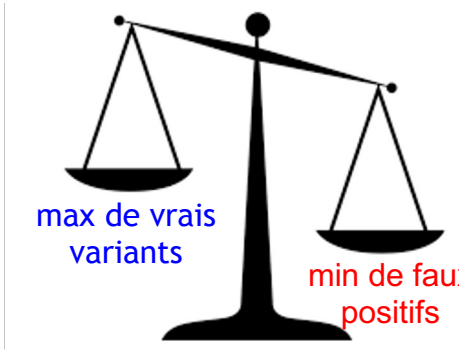
→ d'où l'intérêt de faire ses propres scripts

Rappel : Faux positif

- De nombreux variants Faux Positifs peuvent survenir des étapes précédentes :
  - Artéfacts issus des **cycle PCR** pendant la préparation des échantillons
  - Artéfacts liés à **la technologie de séquençage** (PacBio, HiSeq, NextSeq, …)
  - Difficultées d'**alignement** (régions d'ADN répétées)
  - **Erreurs de lecture** lors du "BaseCalling"

Rappel :
Faux positifs vs Filtres qualité





Plus on est stringent plus on va éliminer les faux positifs mais avec le risque de perdre de vrais variants

du Fastq aux VCFs (ebaiin1)

Reads (Fastq)

Fastq Quality Control
-- FastQC --

Reference genome (Fasta)

Mapping
-- Bwa --

Processing Post Alignement
-- GATK/Picard --

Structural variants
-- Delly --

small variants
-- GATK --

Annotation
-- SnpEff --

Variant Calling

VCF

du VCFs aux marqueurs (ebaiin2)

QC & Filtres stats descriptives

Recherche de marqueurs/QTL

Genome Wide Association Study (GWAS)

analyse phylogénétique

….

Dépend de la question biologique

Objectifs de l’atelier variants

I. Analyse post-VCF des petits variants (SNV & InDel) → 2h (Nadia)

II. Etudes d'association génétique à grande échelle (GWAS) → 4h (Cathy)

Objectifs de l'atelier variants

I. Analyse post-VCF des petits variants (SNV & InDel) → 2h (Nadia)

1- Se connecter sur Rstudio via OnDemand
(1 cpu; 8G ram)

2- Aller sur votre espace projet

```
> setwd("/shared/projects/olala_olala")
```

3- Copier le matériel de TP dans le dir TP_variants

```
> file.copy(from = "/shared/projects/2514_ebiii_n2/dnaseq/TP_small_variants",
            to   = "./", recursive=TRUE)
```

4- Positionner l'espace du travail

```
> setwd("TP_small_variants/")
```

Objectifs de l'atelier variants

I. Analyse post-VCF des petits variants (SNV & InDel) → 2h (Nadia)

II. Etudes d'association génétique à grande échelle (GWAS) → 4h (Cathy)

1- Se connecter sur Rstudio via OnDemand
(4 cpu; 16G ram)

2- Aller sur votre espace projet

```
> setwd("/shared/projects/olala_olala")
```

3- Copier le matériel de TP dans le dir TP_variants

```
> file.copy(from = "/shared/projects/2514_ebiii_n2/dnaseq/GWAS",
            to   = "./", recursive=TRUE)
```

4- Positionner l'espace du travail

```
> setwd("GWAS")
```