# Bioinformatique pour le traitement des données de séquencage

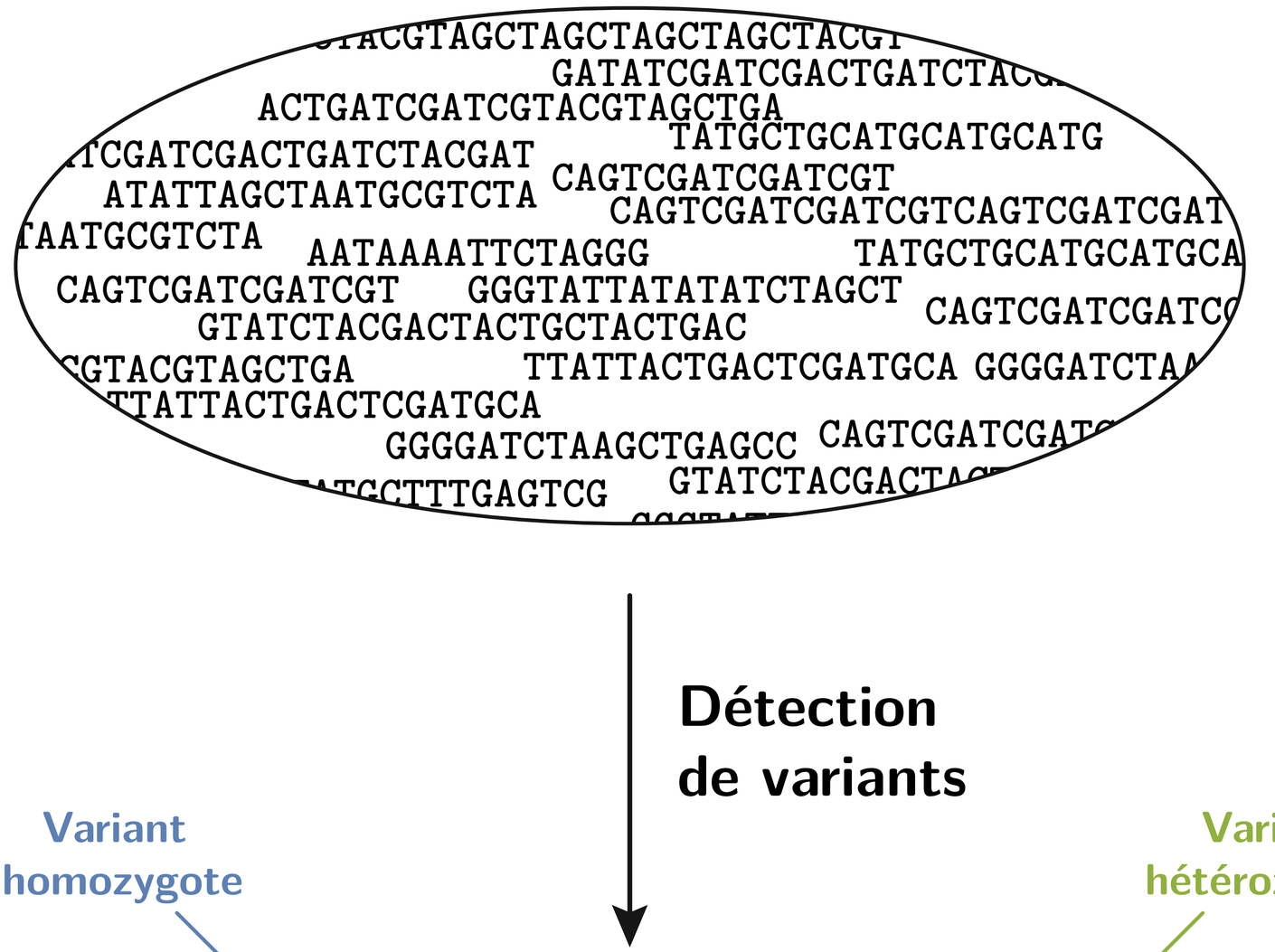
# Détection de variants

**Maud Gautier, Annabelle Haudry, Thibault Latrille**

**17 Septembre 2019**



# Qu'est ce que la détection de variants ?

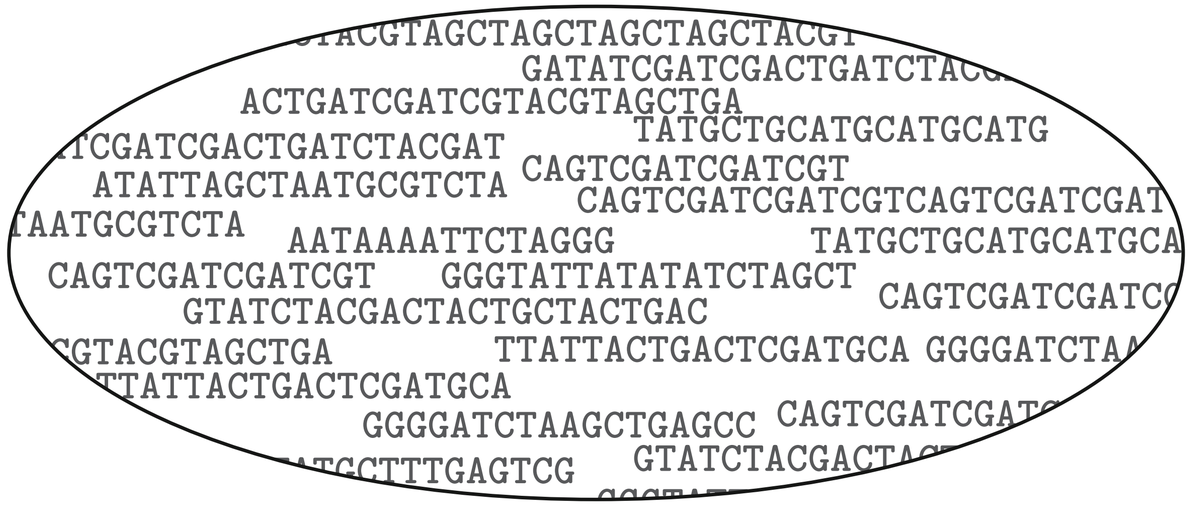



**Données de séquençage d'un individu diploïde**

**Individu diploïde**

**Genome de référence**

# Qu'est ce que la détection de variants ?



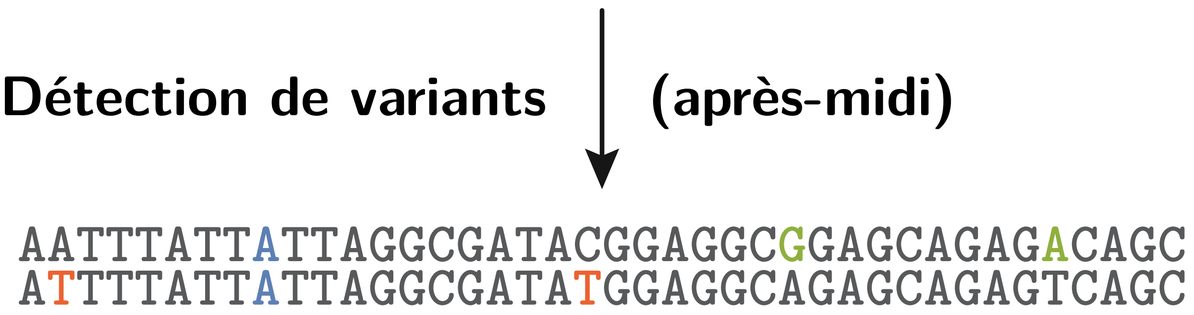

**Données de séquençage d'un individu diploïde (.fastq)**

**Alignement** (matinée)


**Données de séquençage alignées (.bam et .sam)**

**Genome de référence (.fasta)**

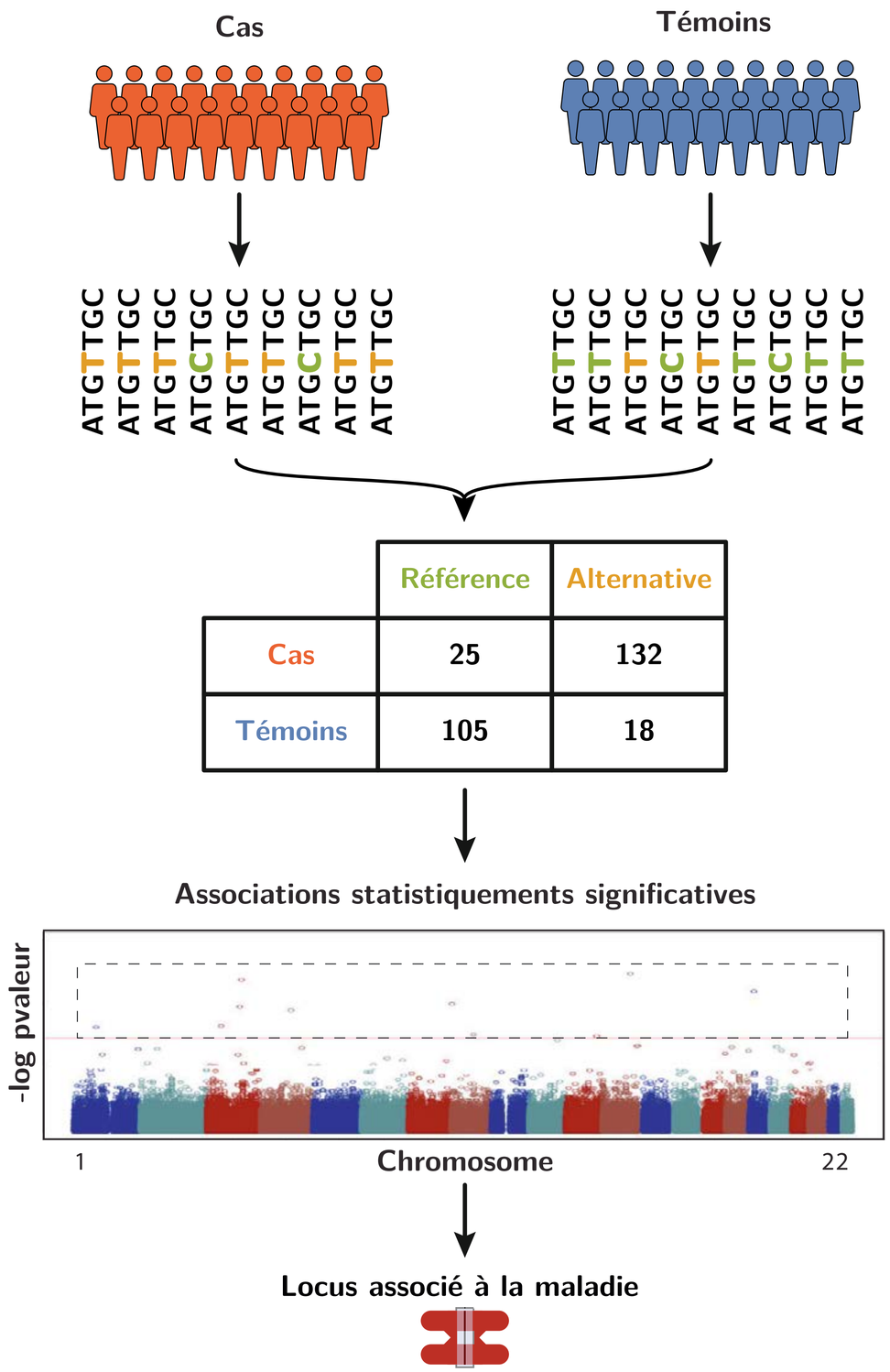
**Détection de variants** (après-midi)



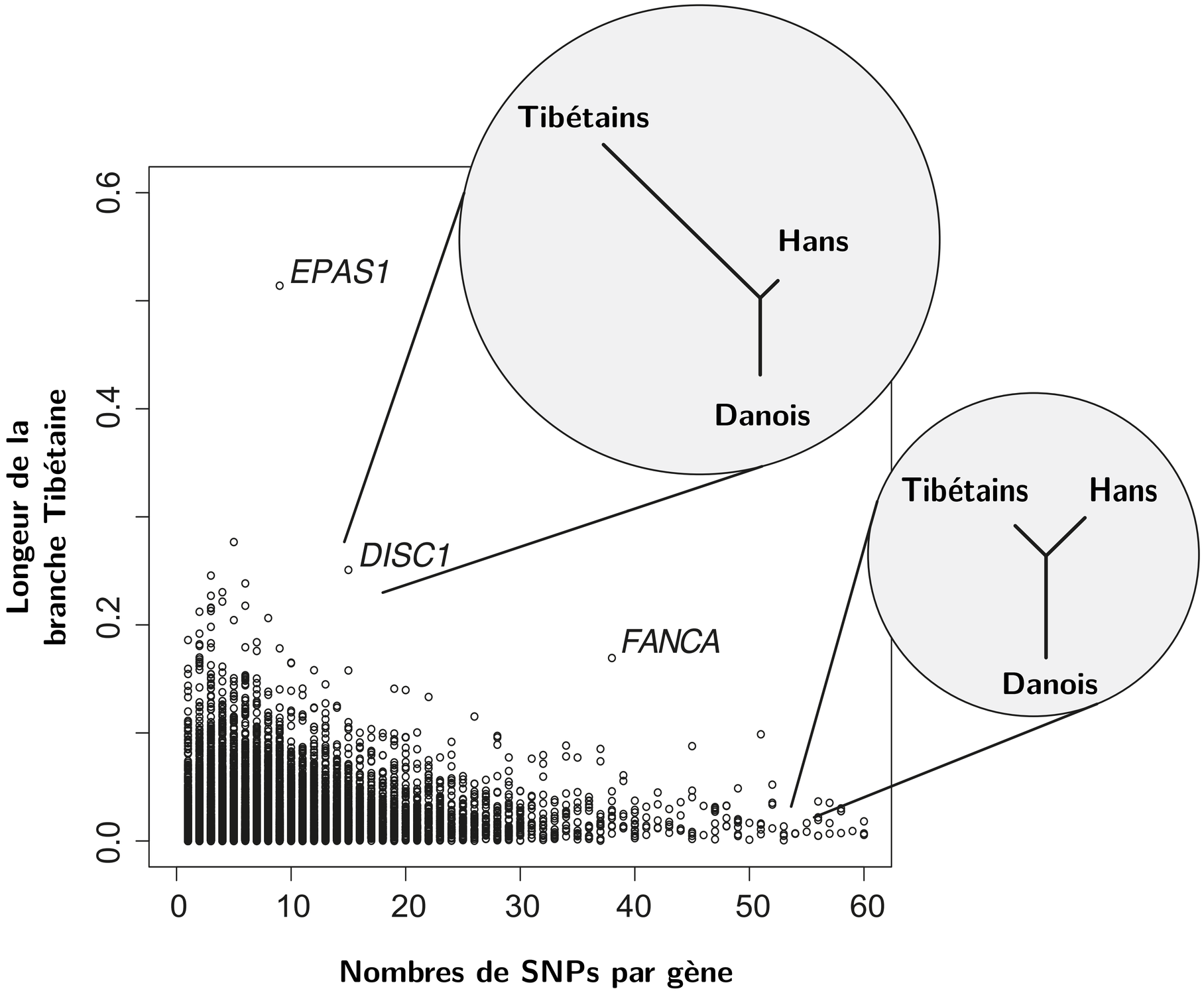

**Individu diploïde (.vcf)**

**Genome de référence (.fasta)**

# Pourquoi cherche-t-on à détecter des variants ?



# Quels gènes sont selectionnés ?



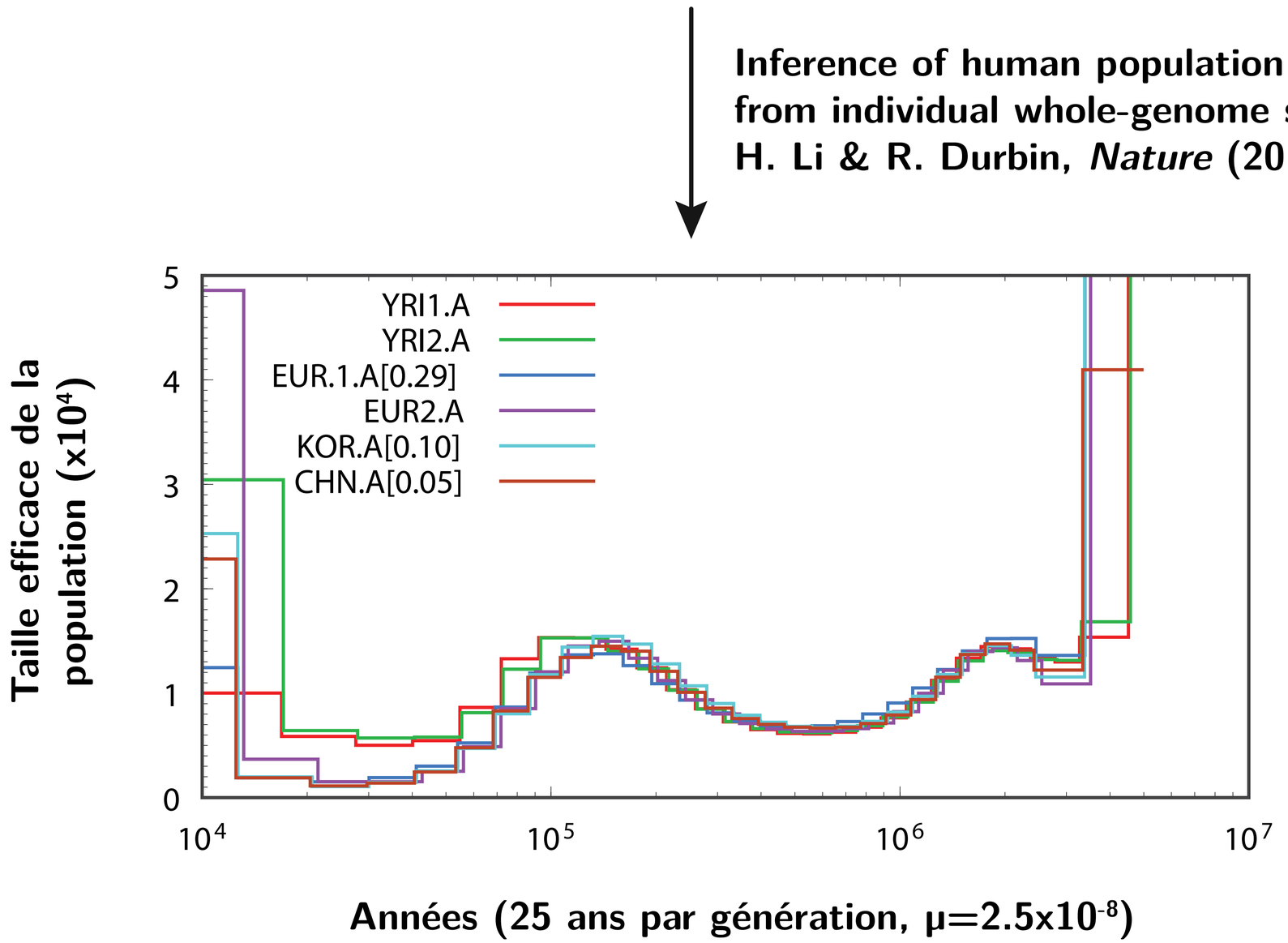

**Yi *et al*, Science (2009)**

# Nos ancêtres étaient-ils nombreux ?

```
-----AATTTATTATTAGGCGATACGGAGGCGGAGCAGAGACAGC-----
-----ATTTTATTATTAGGCGATATGGAGGCAGAGCAGAGTCAGC-----
```

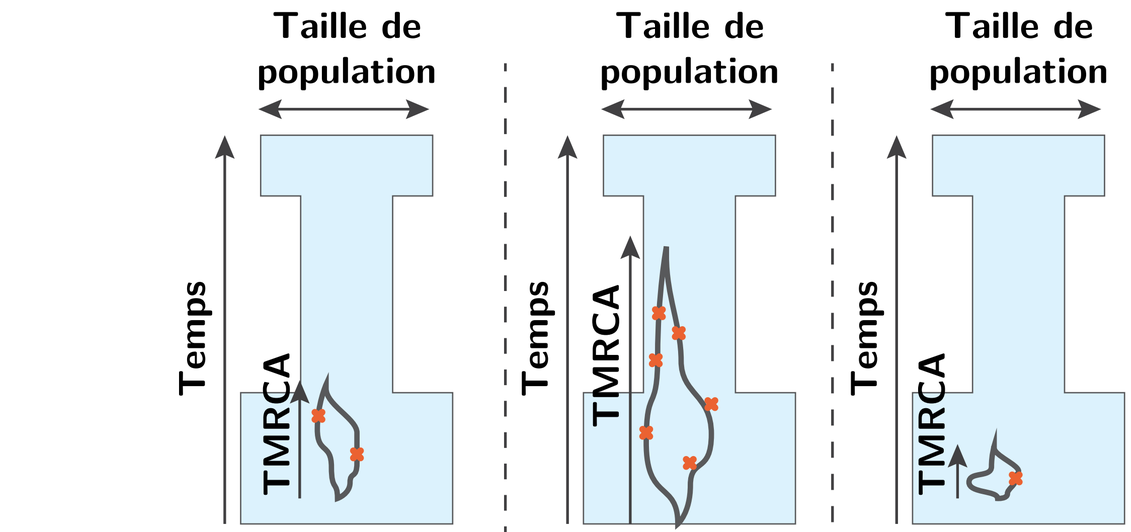
**Individu diploïde**

**Inference of human population history from individual whole-genome sequences, H. Li & R. Durbin, *Nature* (2011)**

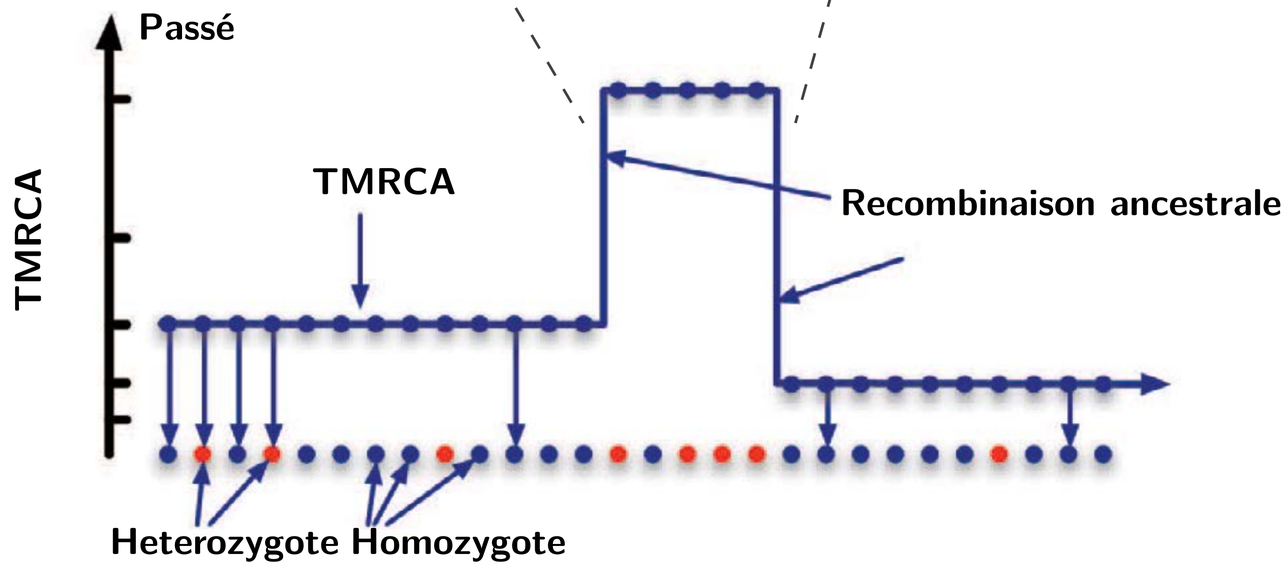


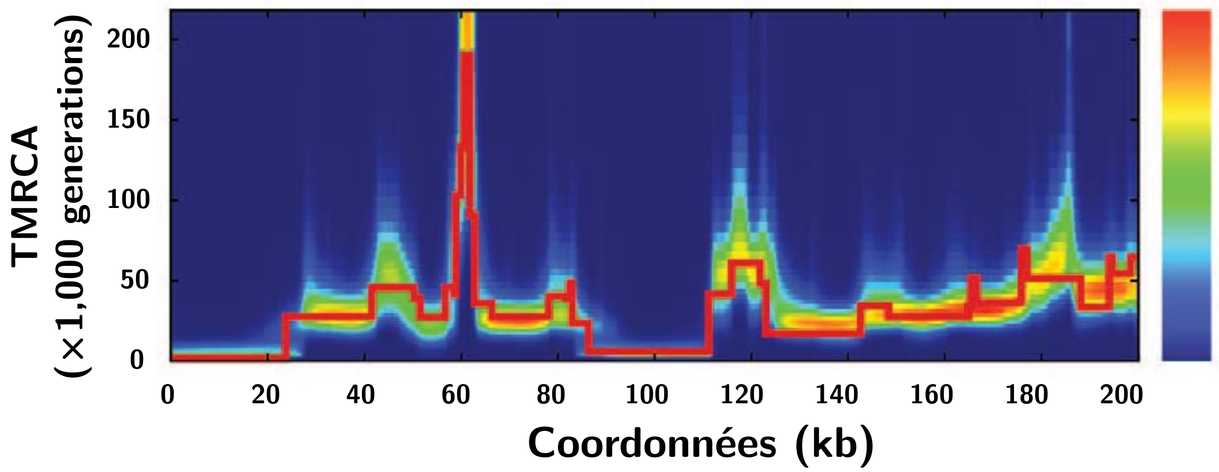

**Li & Durbin, Nature (2011)**

# Nos ancêtres étaient-ils nombreux ?

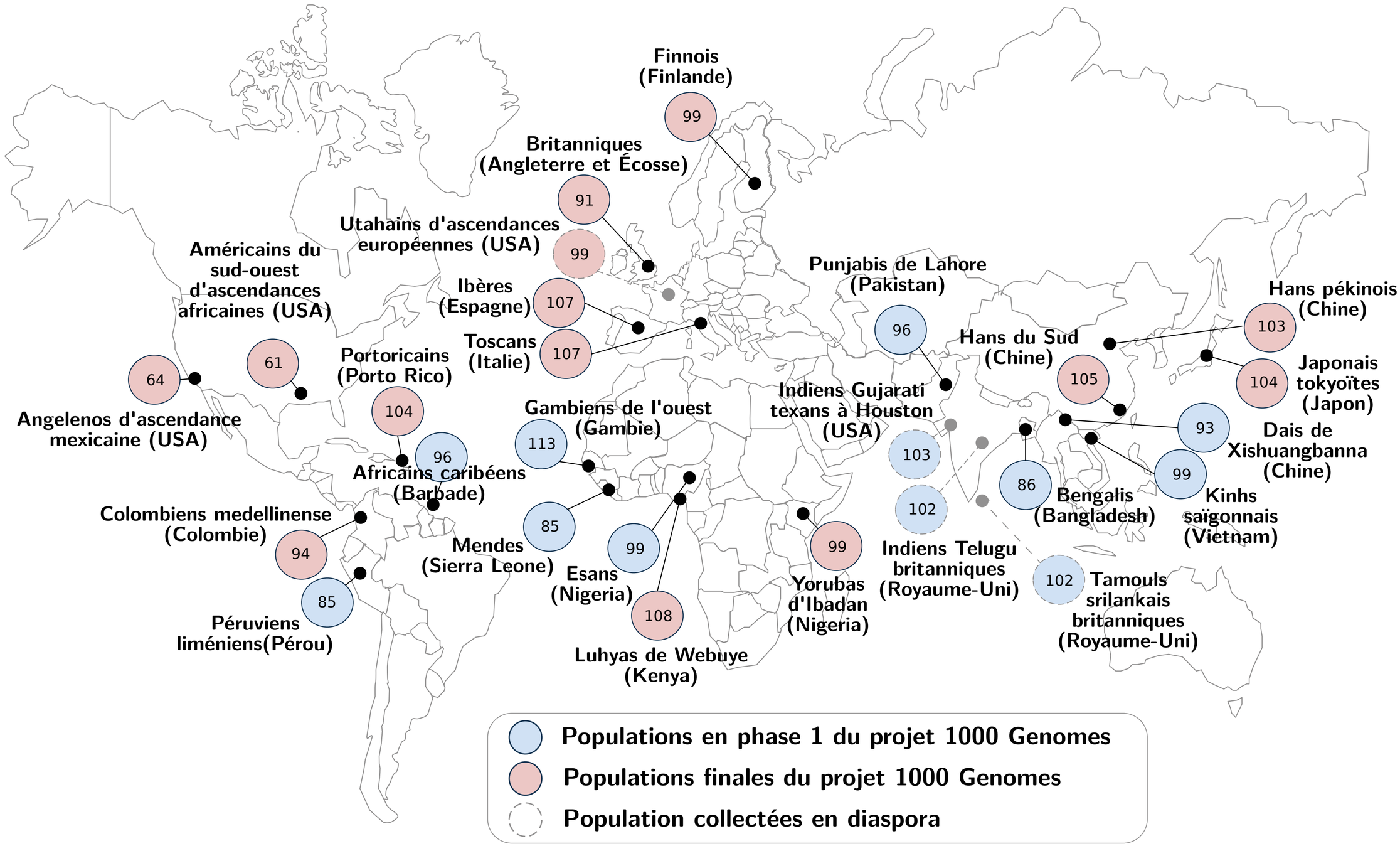



**TMRCA:**
**Temps jusq'à l'ancêtre commun le plus récent**



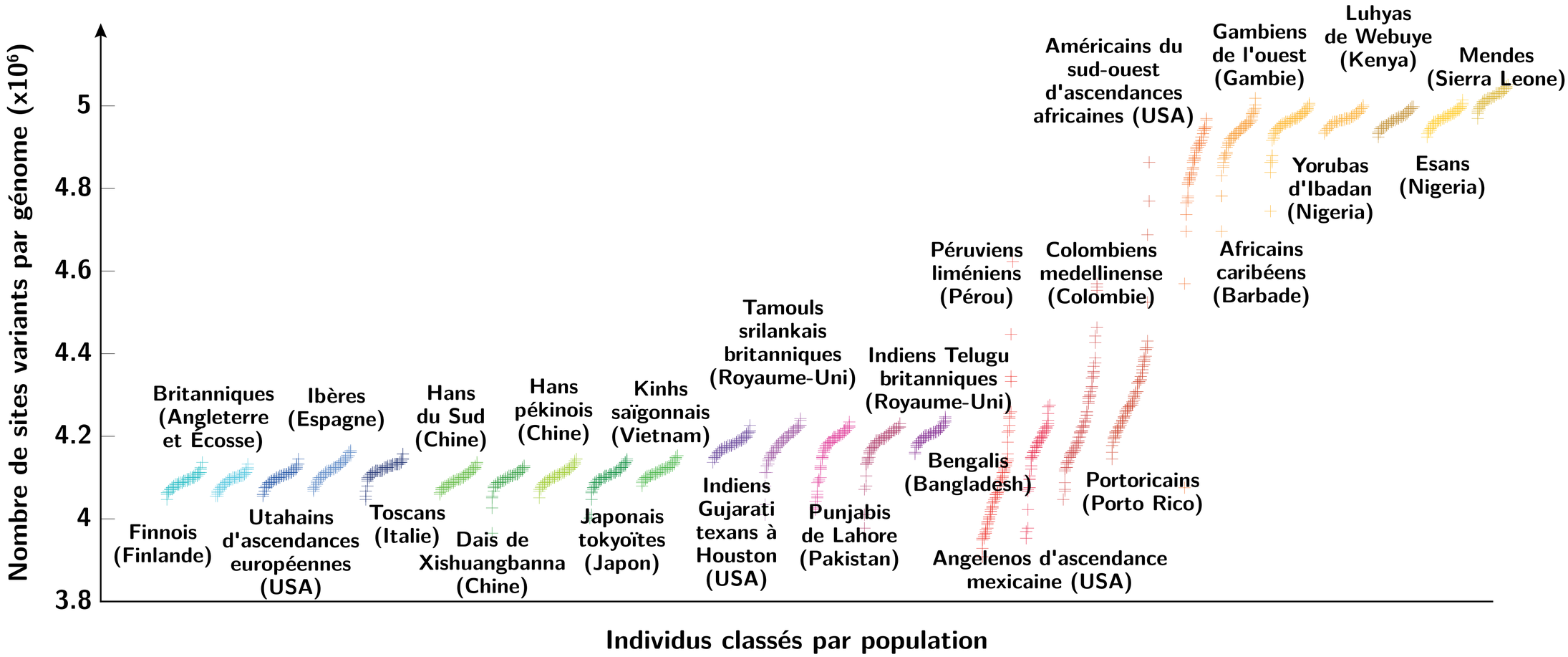



# Quelles données va-t-on utiliser ?



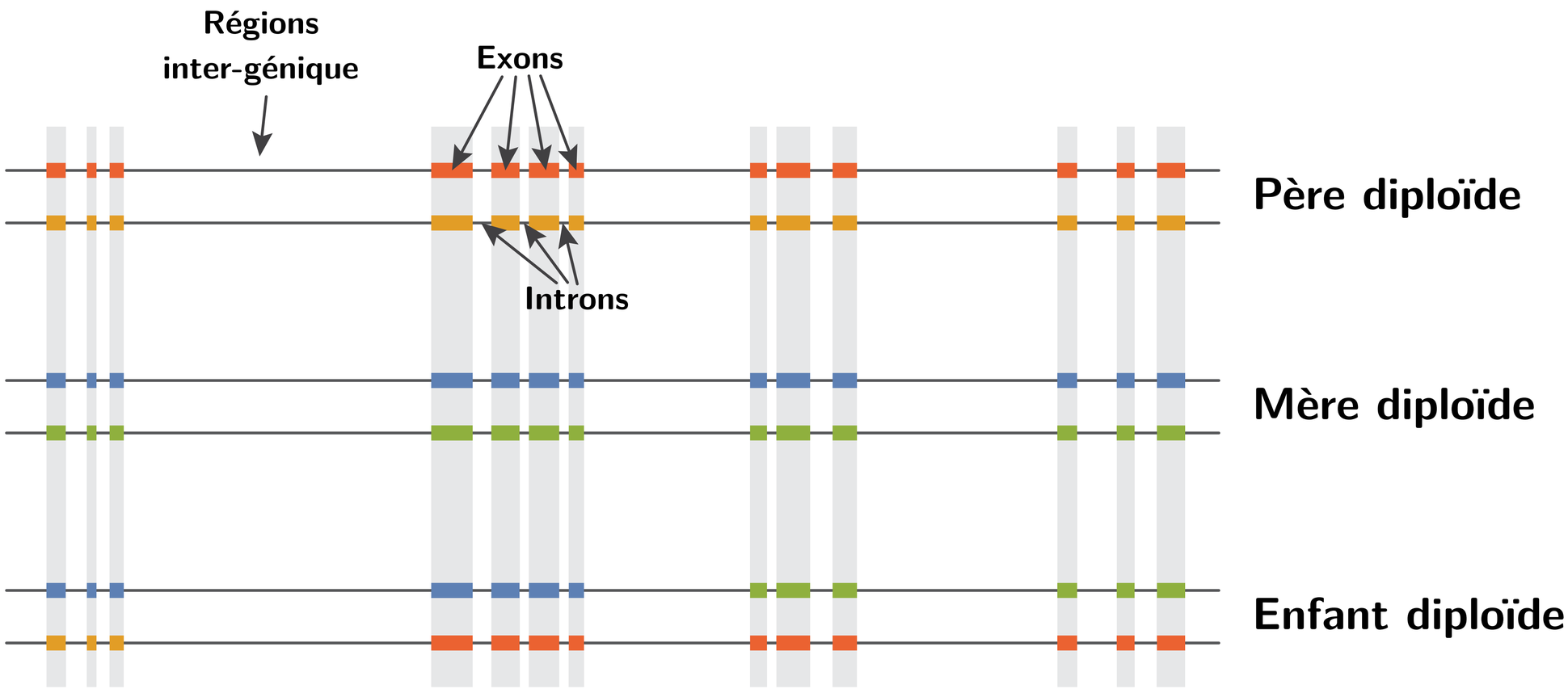

www.internationalegenome.org

# Quels sont les résultats du projet 1000 Génomes ?



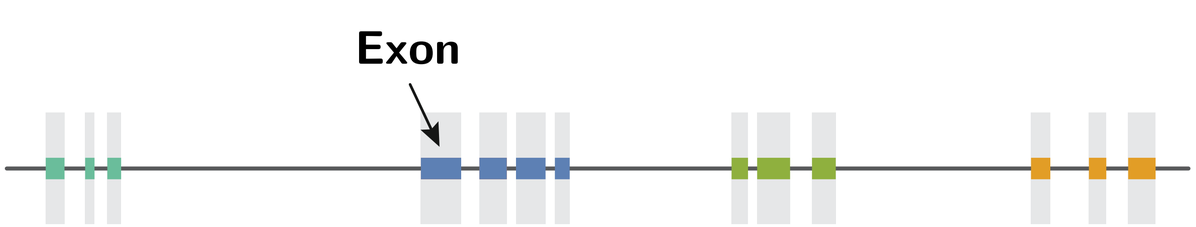

**The 1000 Genomes Project Consortium, Nature (2015)**

# Va-t-on utiliser tout le jeu de données ?

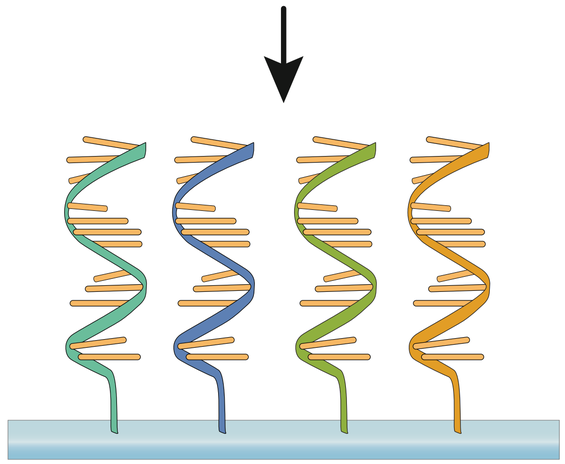



**Séquençage de l'exome pour un trio père-mère-enfant**

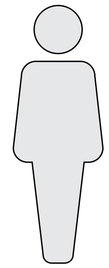
# Comment est obtenu l'exome ?



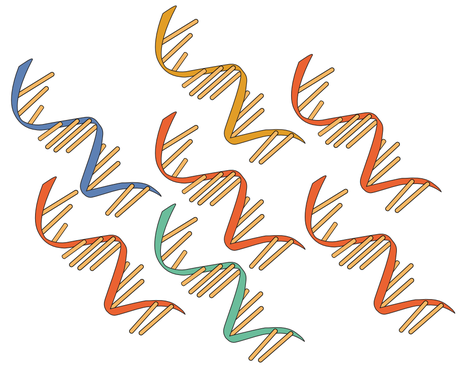

Génome de référence



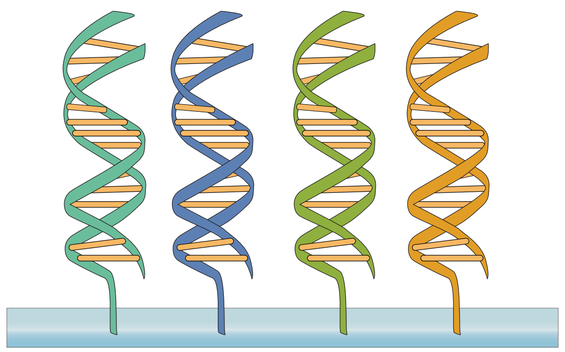
Puce à ADN de l'exome



Individu

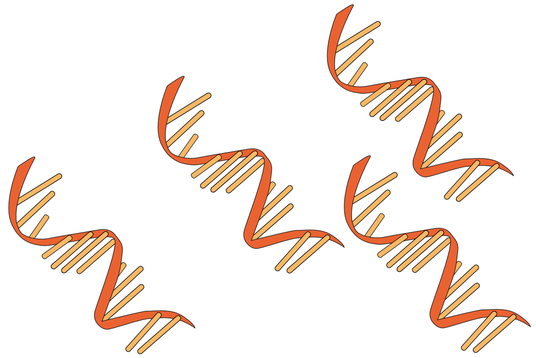


extraction de l'ADN



Hybridisation de l'exome

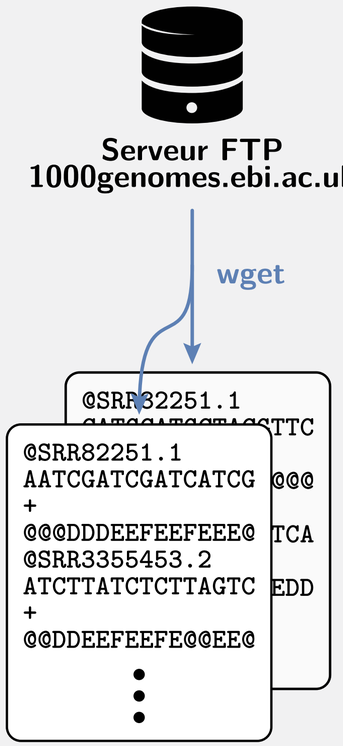
Séquençage de l'exome



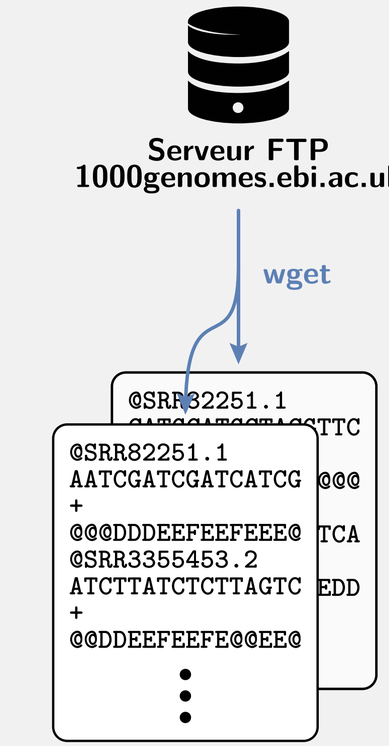
Génome non-exomique

# Donc on fait trois fois la même manipulation ?

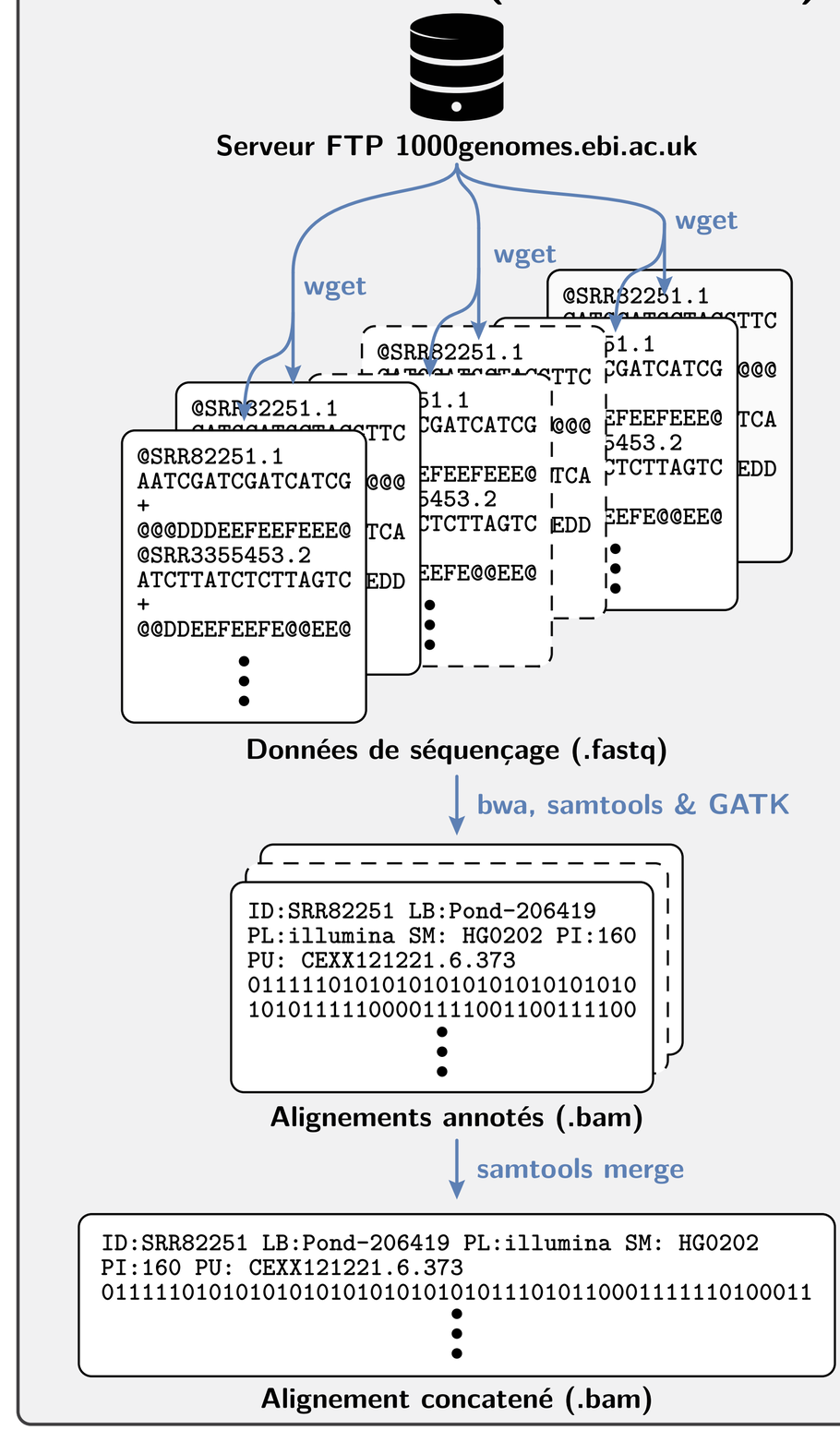
## Niveau 1 - Enfant



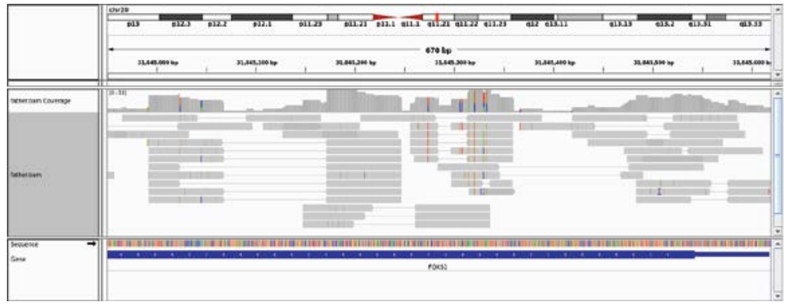

## Niveau 2 - Mère (avec variables)




## Niveau 3 - Père (avec boucle)



# Quels outils va-t-on utiliser ?



Visualisation de l'alignement

Serveur FTP
1000genomes.ebi.ac.uk

wget

Paire 1

```
@SRR82251.1
AATCGATCGATCATCG
+
@@@DDDEEFEEFEEE@
@SRR3355453.2
ATCTTATCTCTTAGTC
+
@@DDEEFEEFE@@EE@
```

Paire 2

```
@SRR82251.1
CATGCATGCTACGTTC
+
DEFEFEEEEFD@@@@@
@SRR3355453.2
ATCGATCGATTTTTCA
+
DDEFFEF####@@EDD
```

Données de séquençage (.fastq)

IGV

```
00110101000011111010000111000 •••
```

Index de l'alignememt (.bam.bai)

samtools index

```
FFQ 0 1 201 12
30  2 2  45  5
```

Statistiques (.txt)

samtools stats

```
ID:SRR82251 LB:Pond-206419 PL:illumina SM: HG0202
PU: CE1E0PACXX121221.6.tagged_373 PI:160
01111101010101010101010101010101011000000111011101010
11101010101000001010010111100101010101010100001001001
```

Alignement annoté (.bam)

GATK AddOrReplaceReadGroups

```
01111101010101010101010101010101011000000111011101010
11101010101000001010010111100101010101010100001001001
```

Alignement ordonné et compressé (.bam)

samtools view & sort

```
2024002 read1
2024002 read2
124714  mapped
```

Statistiques (.txt)

```
@SRR82251   770    CATGCATGCTACGTTC   DEFEFEEEEFD@@@@@
@SRR82251   2147   ATCGATCGATTTTTCA   DDEFFEF####@@EDD
```

Alignement (.sam)

samtools flagstats

Serveur FTP
ensembl.org

wget

```
>homo_sapiens_GRCh38_Chr20
ATATATTTTACTCATGCTAGCATGCTACTTATTGGCGAGG •••
```

Génome de référence (.fasta)

bwa index

```
0101010000111100010101000111000001111010 •••
```

Index du génome de référence (.fasta.*)

bwa mem