INRAE

Détection de variants

Thomas Faraut

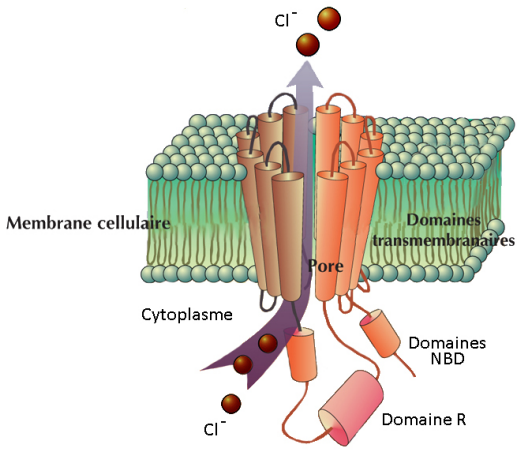
Toulouse, 13 mars 2023

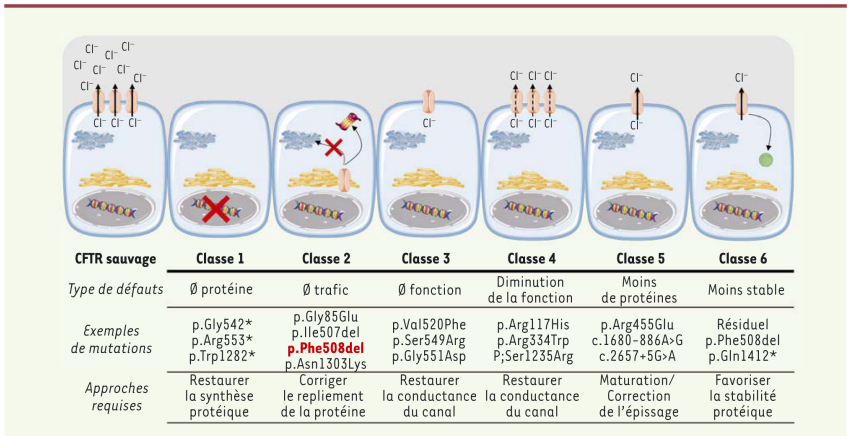
Outline

- What is a genomic variant ?
- Why should we care about them ?
- How can we detect variant ?

Genetic disorders

Cystic fibrosis is caused by mutations in the CFTR gene





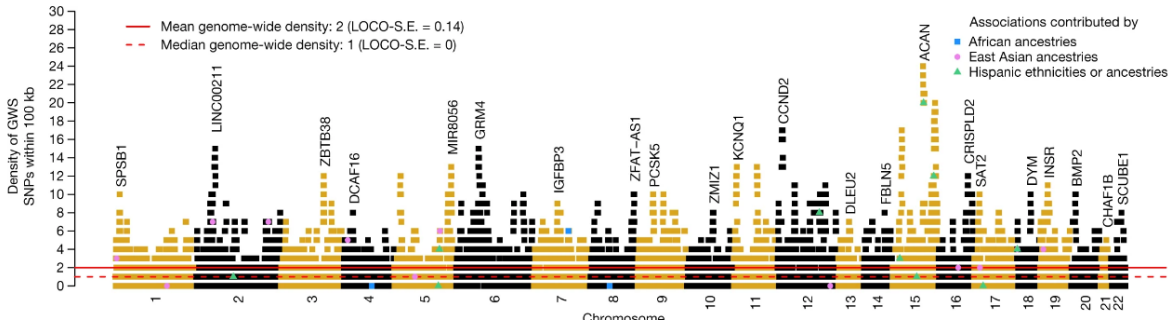
CFTR sauvage	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Classe 6
Type de défauts	Ø protéine	Ø trafic	Ø fonction	Diminution de la fonction	Moins de protéines	Moins stable
Exemples de mutations	p.Gly542* p.Arg553* p.Trp1282*	p.Gly85Glu p.Ile507del **p.Phe508del** p.Asn1303Lys	p.Val520Phe p.Ser549Arg p.Gly551Asp	p.Arg117His p.Arg334Trp P;Ser1235Arg	p.Arg455Glu c.1680-886A>G c.2657+5G>A	Résiduel p.Phe508del p.Gln1412*
Approches requises	Restaurer la synthèse protéique	Corriger le repliement de la protéine	Restaurer la conductance du canal	Restaurer la conductance du canal	Maturation/ Correction de l'épissage	Favoriser la stabilité protéique

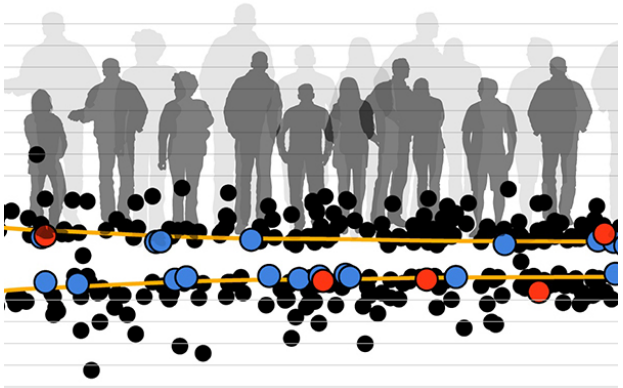
La mucoviscidose. Du gène à la thérapeutique Medecine et Sciciences (2021)

Genomic variants and human height

10% to 40% of all variation in height can be explained by the genome

From: A saturated map of common genetic variants associated with human height

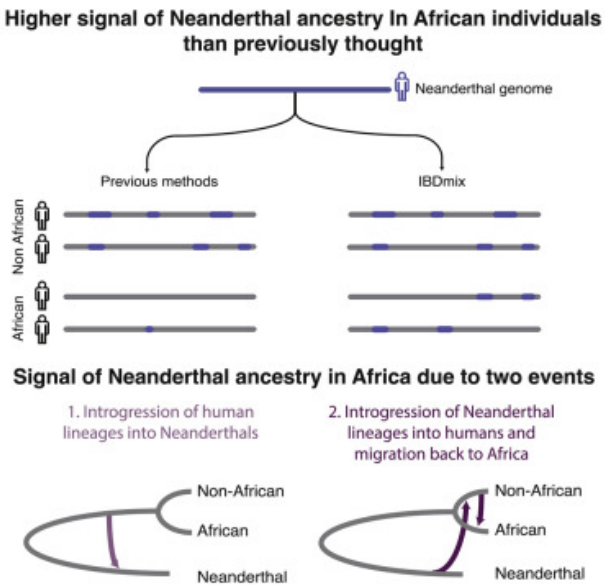


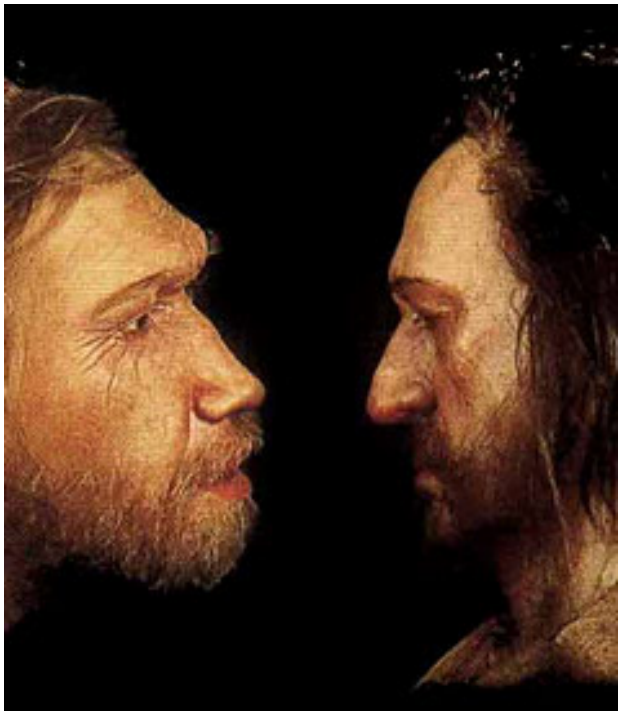


A saturated map of common genetic variants associated with human height. Nature (2022)

Neanderthal our cousin

"East Asian populations have approximately 20% more Neanderthal ancestry compared to Europeans"

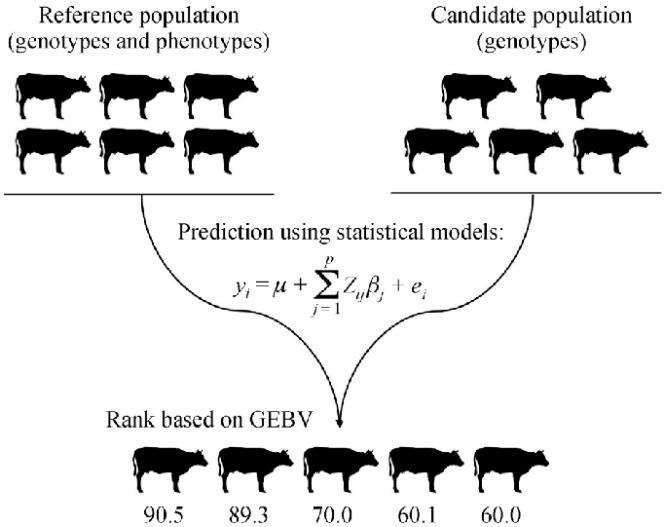




Identifying and Interpreting Apparent Neanderthal Ancestry in African Individuals. Cell (2020)

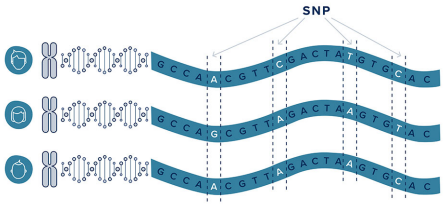
Genomic selection

L'estimation d'une relation entre variations et phénotype permet de prédire le potentiel génétique d'un individu.



Variations

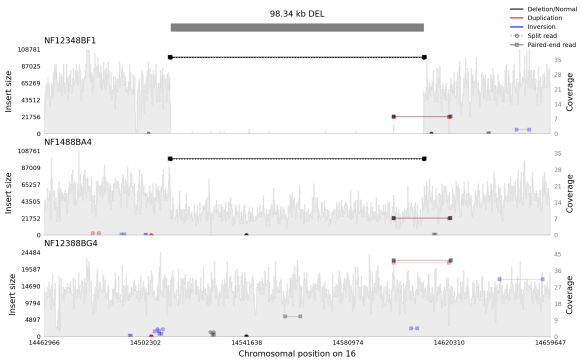
SNP



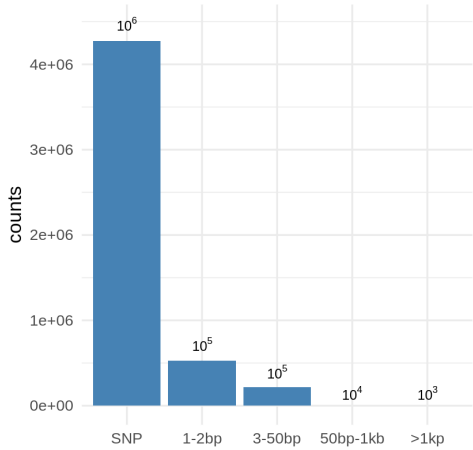
indels

```
AGGCTCCTCCGTCCATGCGATT--GCAGGCC
||  |     |||||||  |||||||||    ||||||
AGTC---TCCGTCCATGGGATTTTGCAGGCC
```

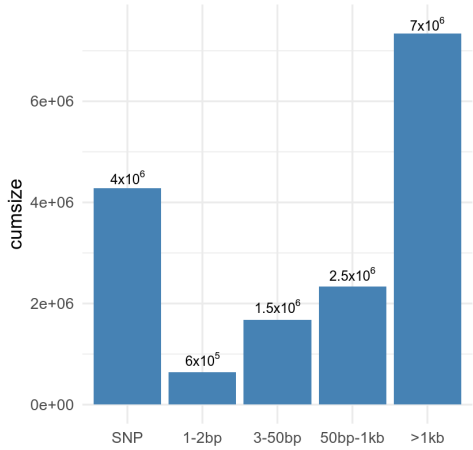
SV



Number of variants

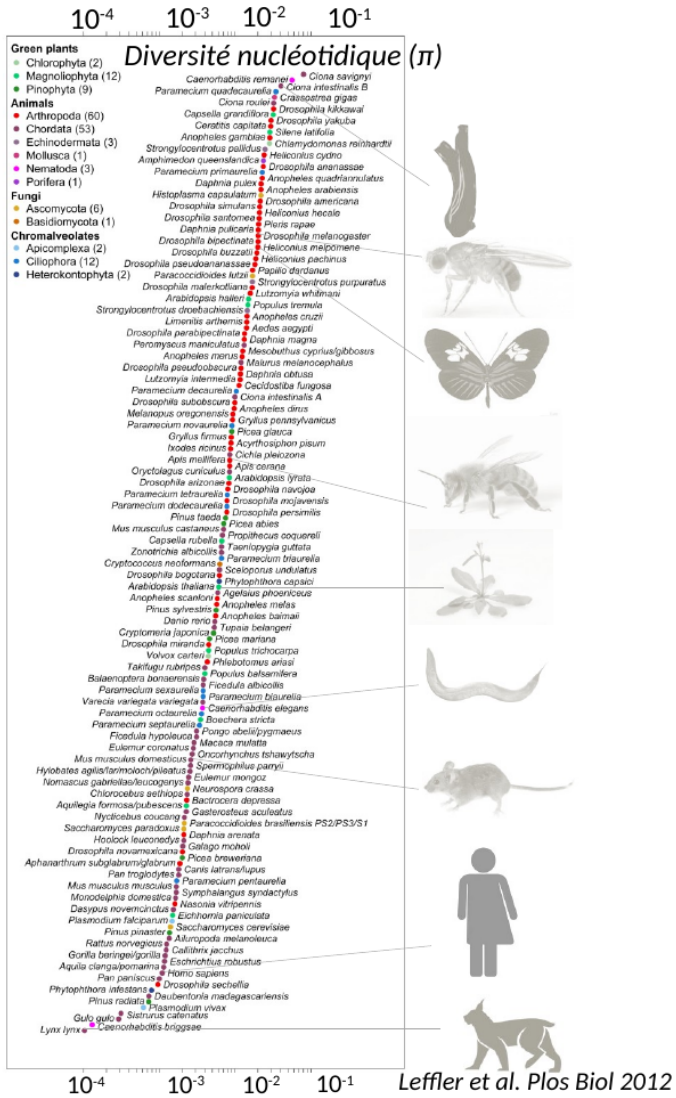


Cumulative size



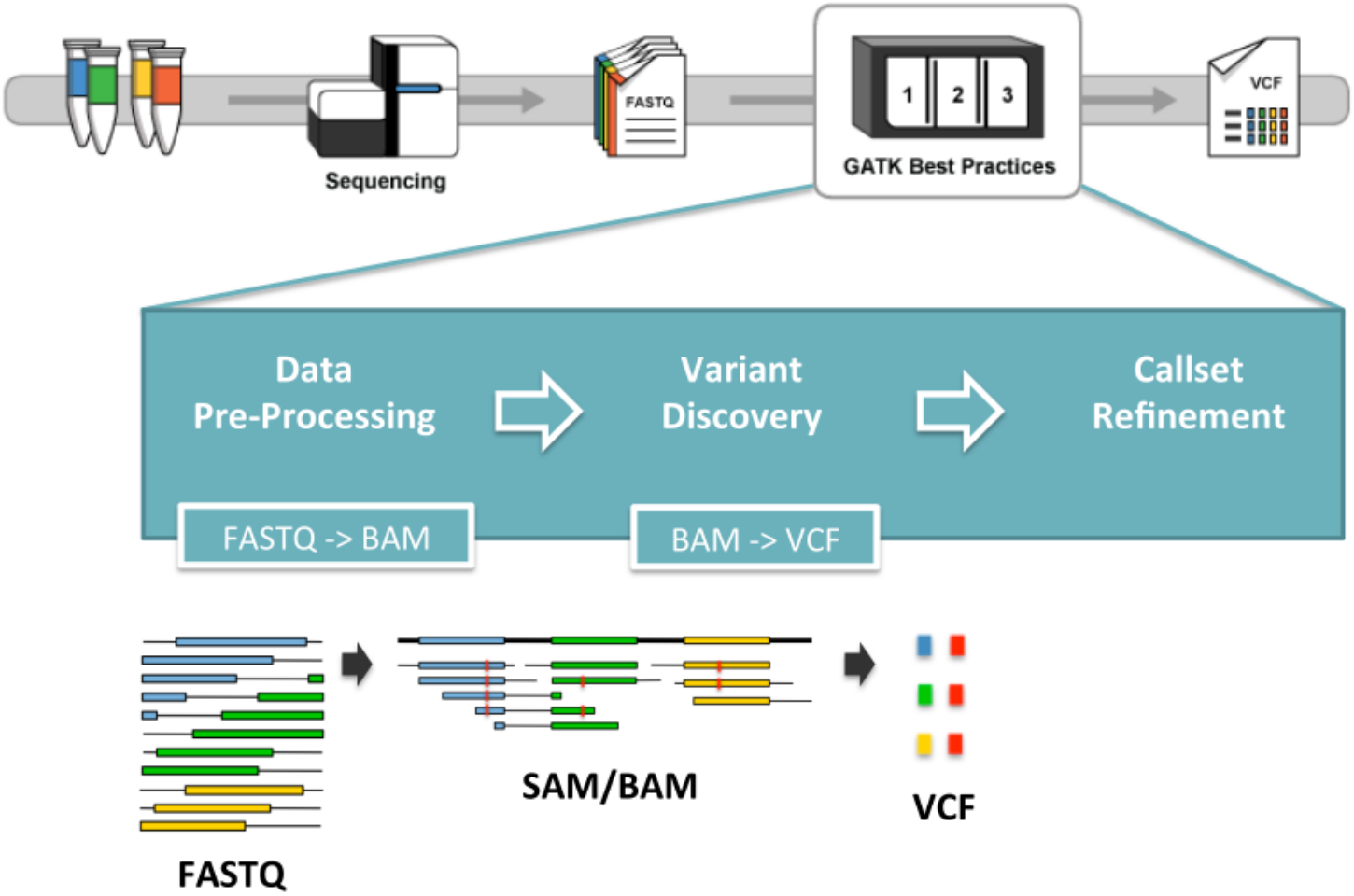
Nucleotide diversity

The level of polymorphism, the nucleotide diversity, differs between species

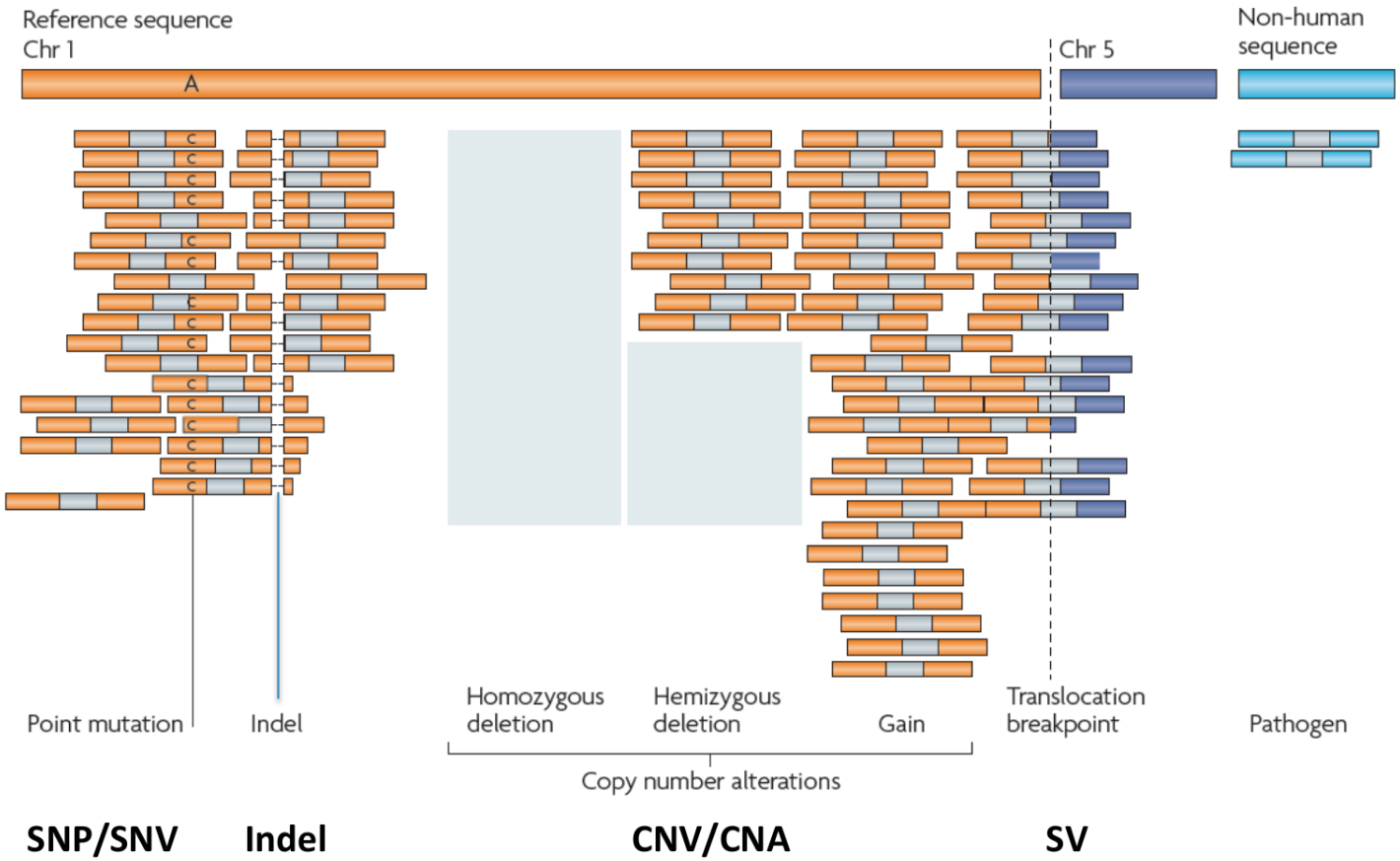


Leffler et al. Plos Biol 2012

NGS worklow

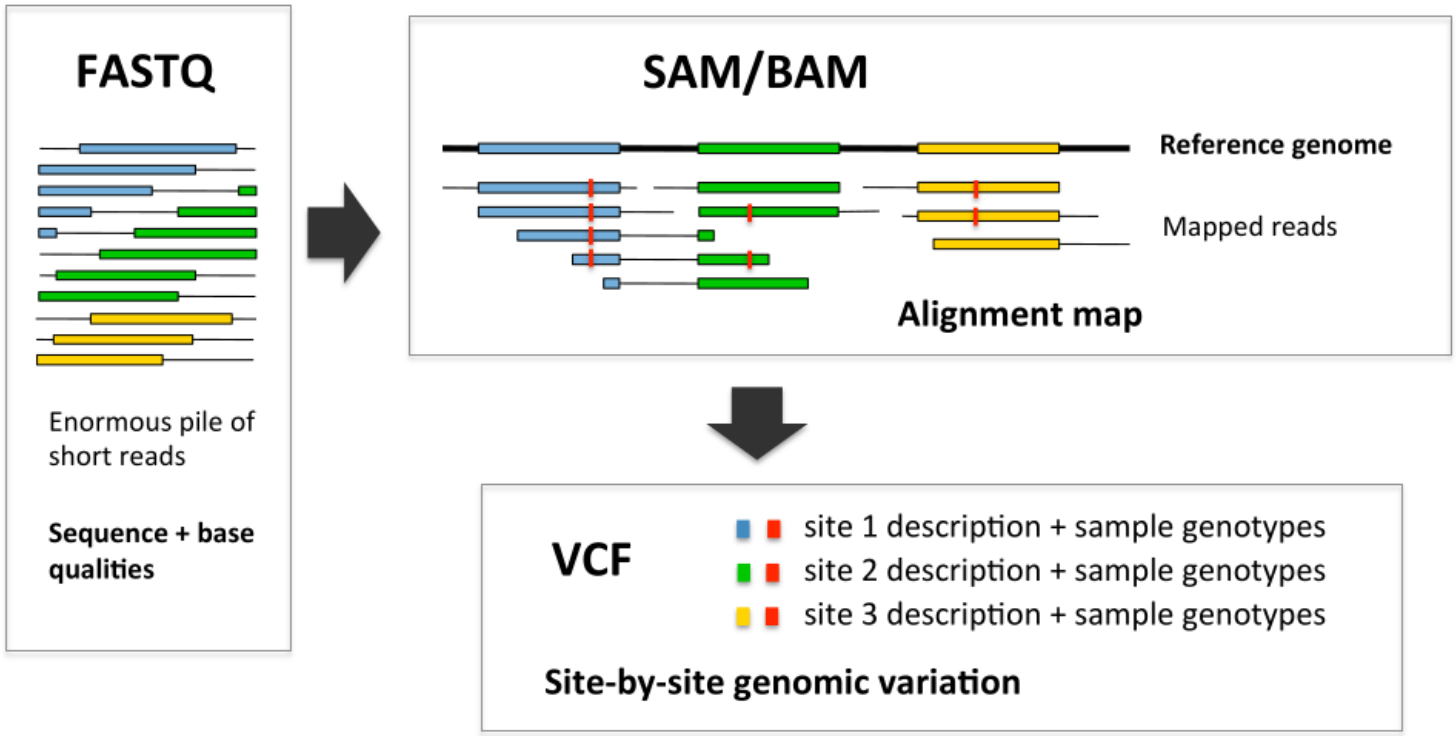

Sequencing
FASTQ
1
2
3
GATK Best Practices
VCF
Data
Pre-Processing
Variant
Discovery
Callset
Refinement
FASTQ -> BAM
BAM -> VCF
FASTQ
SAM/BAM
VCF

Detecting variants with reads

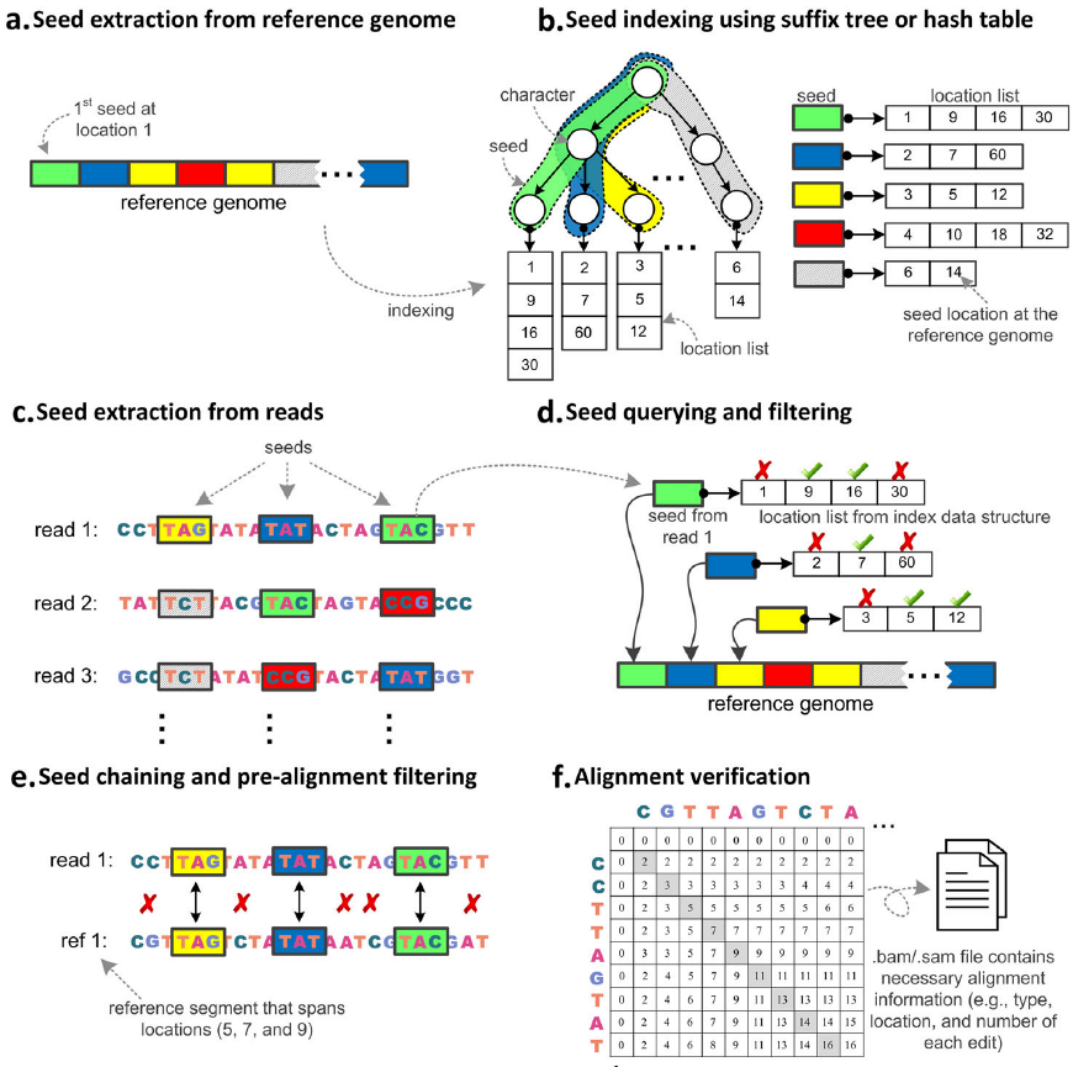


GATK a primer

Variant discovery pipeline

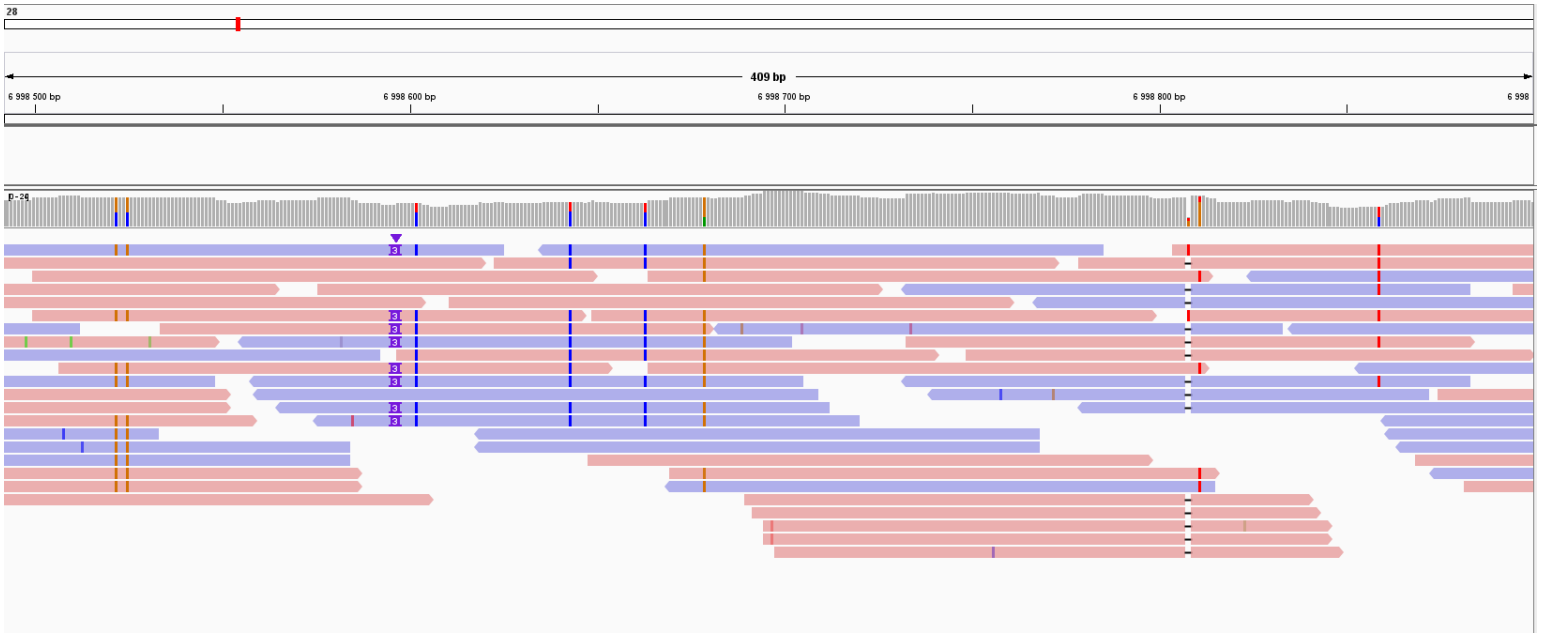


Map to the reference



Technology dictates algorithms: recent developments in read alignment

Starting from Aligned reads



- Reads exhibit descripancies with the reference assembly
- How can we decide if it is an SNP or not ?

Empirical approach

- Count the reads associated with each allele : 4 A and 7 C
- Take a decision according to the proportion of reads for each allele $\Longrightarrow$ heterozygote

Empirical approach

Caveats

- What is a reasonnable choice for the minimum required number of alternate alleles ?
- Cutoff for deciding when an individual is heterozygote ?
- How can we handle the uncertainty inherent of low-coverage data ?
- How to deal with the sequencing errors ?
- No measure of uncertainty in the genotype inference

Probabilistic approach

- What is the probability that a column signs an SNP ?
- What is the most likely genotype for that SNP ?

Probabilistic approach

A single individual

Probabilistic approach

$$P(G_i|D) = \frac{P(G_i)P(D|G_i)}{P(D)}$$

- G_i : one of the possible genotypes
- D : data, i.e. a column of the multiple alignment of reads
- Suppose for the moment a single diploïd individual, hence 3 possible genotypes $\{0/0, 0/1, 1/1\}$

Bayes theorem

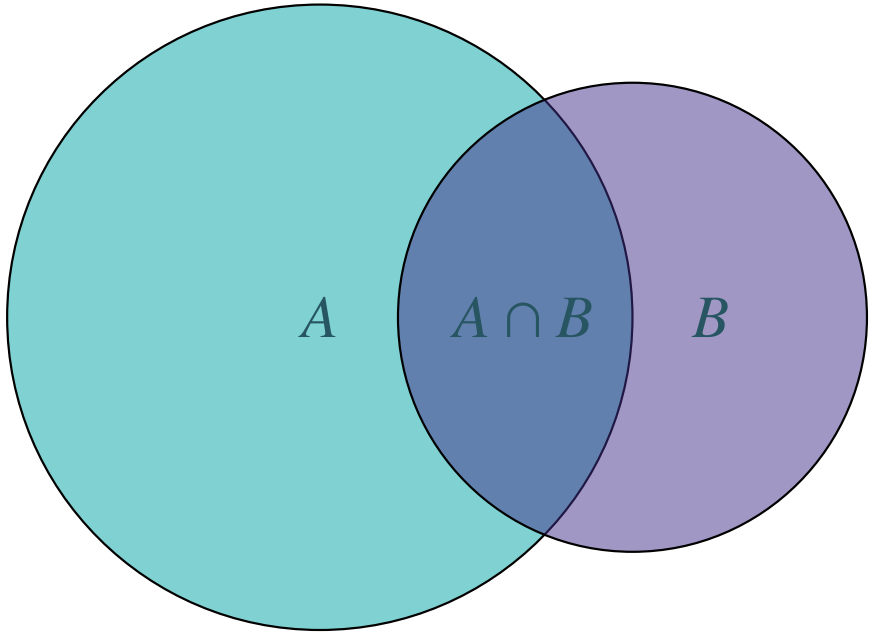
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

and

$$P(A \cap B) = P(A)P(B|A)$$

hence

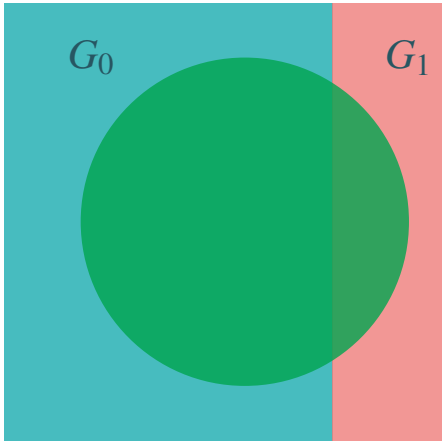
$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$



Bayes theorem

Consider a haploid genome, hence two genotypes 0 and 1

$$P(G_i|D) = \frac{P(G_i)P(D|G_i)}{P(D)}$$



$$P(D) = P(D \cap G_0) + P(D \cap G_1)$$

hence

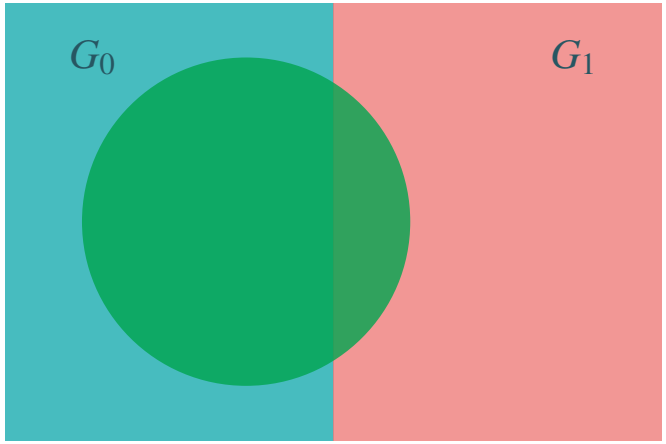
$$P(G_i|D) = \frac{P(G_i)P(D|G_i)}{P(G_0)P(D|G_0) + P(G_1)P(D|G_1)}$$

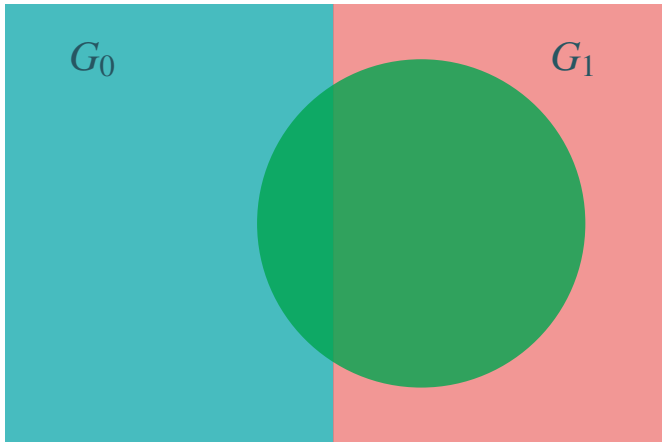
Bayes theorem

Consider a haploid genome, hence two genotypes 0 and 1

$$P(G_i|D) = \frac{P(G_i)P(D|G_i)}{P(D)}$$

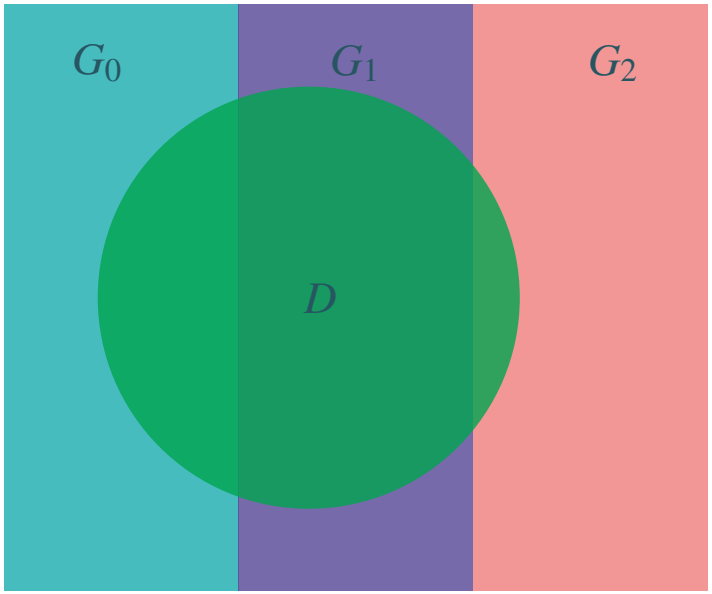
$$\propto P(G_i)P(D|G_i)$$





Bayes theorem

- Consider a diploid genome with 3 genotypes, $G_0 = 0/0$, $G_1 = 0/1$ and $G_2 = 1/1$
- Reads data D gives us some clues about the most likely genotype



Probabilistic approach

Probability of data given the genotype

$$P(G_i|D) \propto P(G_i)P(D|G_i)$$

- $P(D|G_i)$ is the probability to observe the reads for a given genotype
- In the absence of sequencing errors it is straightforward

Probabilitic approach

Figure – A candidate SNP with 10 reads, 6 C and 4 A

- Suppose a candidate SNP with two alelles $(0$ and $1)$

$P(D|G_i)$

- Let ε be the error rate, and
- D be a set of n reads, k reads with 0 and $(n-k)$ reads with 1

$$P(D|00) = \binom{n}{k}(1-\varepsilon)^k \varepsilon^{n-k}$$

$$P(D|11) = \binom{n}{k}\varepsilon^k (1-\varepsilon)^{n-k}$$

$$P(D|01) = \binom{n}{k}p^k (1-p)^{n-k}$$

with

$$p = \frac{1}{2}\varepsilon + \frac{1}{2}(1-\varepsilon) = \frac{1}{2}$$

- The right hand sides are called the genotype likelihoods

$$P(G|D) \propto L(G|D) = P(D|G)$$

$L(G_i) = P(D|G_i)$ the genotype likelihood

We are (only) interested in finding the most likely genotype

$$P(D|00) \propto (1-\varepsilon)^k \varepsilon^{n-k}$$
$$P(D|11) \propto \varepsilon^k (1-\varepsilon)^{n-k}$$
$$P(D|01) \propto (\frac{1}{2})^n$$

- The genotype caller provides the genotype likelihoods coded using a Phred score.

$$PL = -10\lfloor \log_{10}(P(D|G)) \rfloor$$

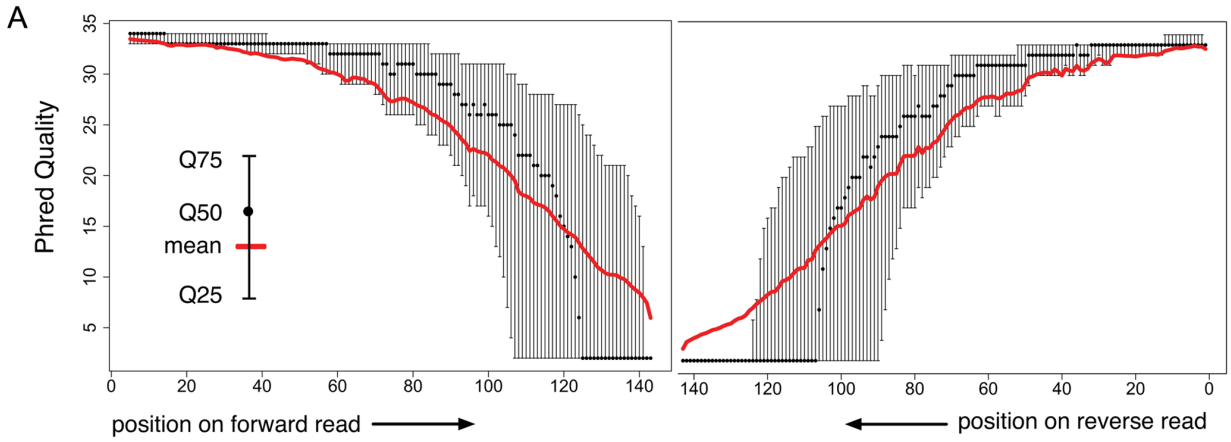
where $\lfloor x \rfloor$ means "the largest integer smaller than x"

- The **called** genotype is the one with the higher likelihood

Probabilistic approach

BaseQuality Score

Sequencing errors



- DNA sequencing is not exempt of errors
- Manufacturers, like Illumina, provide Base Quality Score coded in a phred score

Sequence quality

FASTQ quality values

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
|    |    |    |    |    |    |    |    |
0....5...10...15...20...25...30...35...40
|    |    |    |    |    |    |    |    |
worst.................................best
```

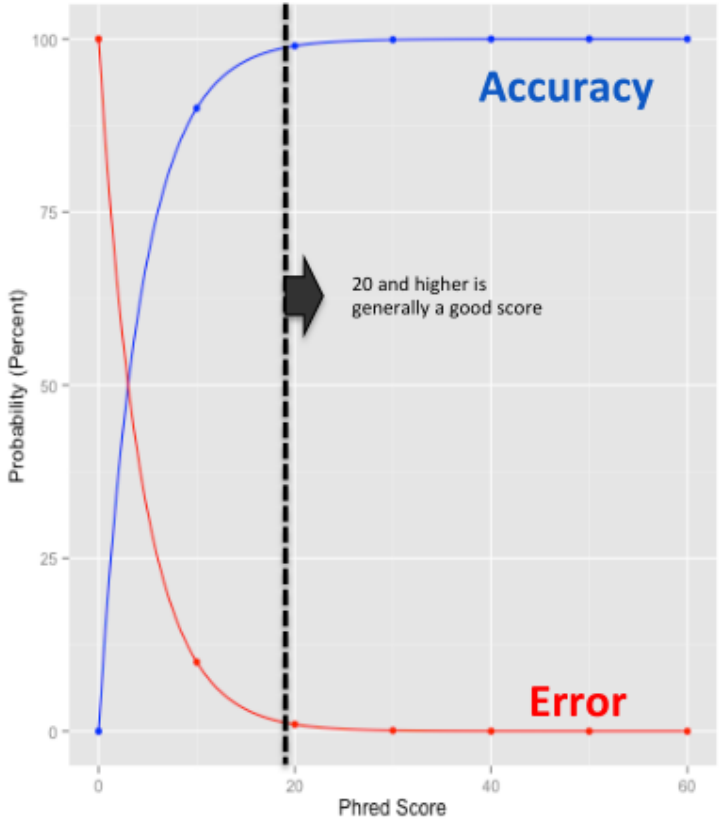
- ASCII coding of the quality (*Phred* quality score)

$$Q = -10\lfloor \log_{10}(p) \rfloor$$

Q	p	erreur
10	10^{-1}	10%
20	10^{-2}	1%
30	10^{-3}	1‰

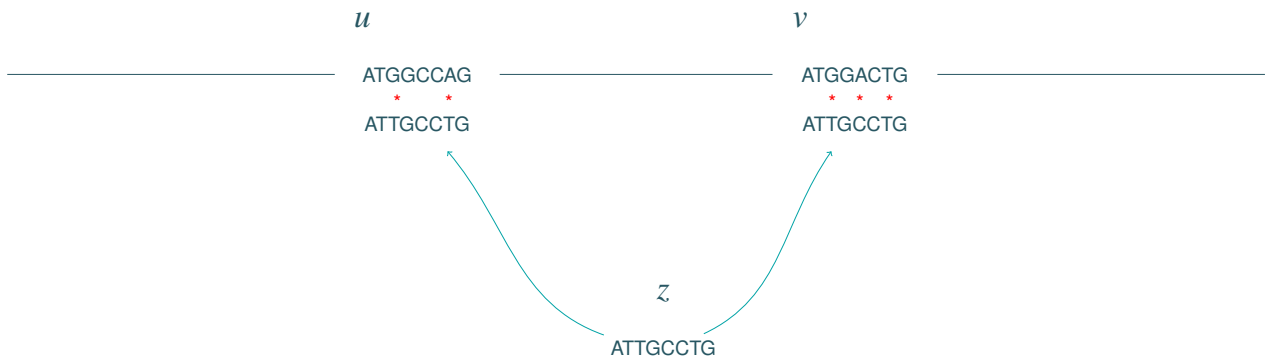
Sequencing errors

- *Phred value* = −10 * *log*10(ε)
- Examples:
  - 90% confidence (10% error rate) = Q10
  - 99% confidence (1% error rate) = Q20
  - 99.9% confidence (.1% error rate) = Q30
- SAM encoding adds 33 to the value
  (because ASCII 33 is the first visible character)



GATK a primer

Mapping quality



Let $p(z|u)$, the probability of z aligned at u, here for example with two errors

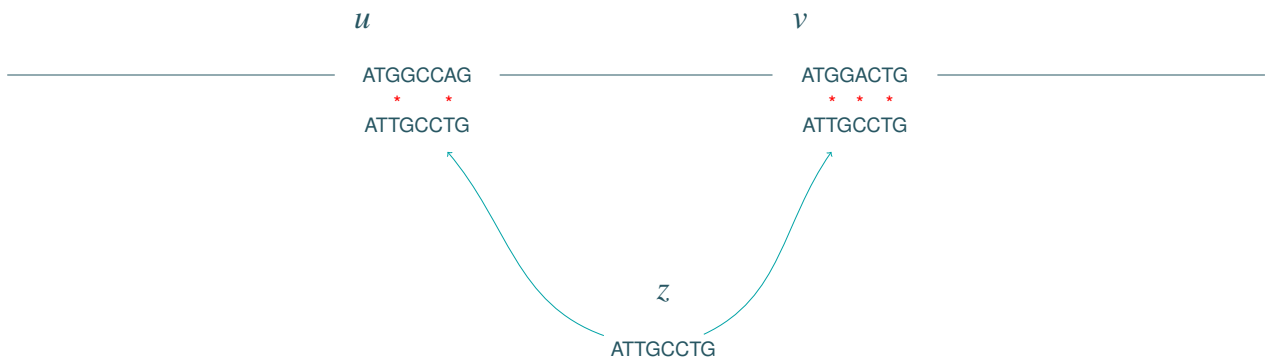
$$p(z|u) = \varepsilon_3 * \varepsilon_7$$

the probability that the read R comes from position u can be computed using

$$p(u|z) = \frac{p(u)p(z|u)}{\sum_v p(v)p(z|v)} = \frac{p(z|u)}{\sum_v p(z|v)}$$

where the sum $\sum_v$ is over all the alternative positions.

Mapping quality



The mapping quality is defined by

$$\begin{aligned} mapQ &= -10\log_{10}[\text{read is wrongly mapped}] \\ &= -10\left\lfloor \log_{10}(1 - p(u|z)) \right\rfloor \end{aligned}$$

$mapqQ = 0 \Longrightarrow$ many alternate positions are equally likely.

$L(G_i)$: genotype likelihood

- The base quality provided in the (fastq) sam files enables to compute a position-specific, read-specific error probability
- If the three genotypes are coded $0, 1, 2$, the likelihood takes the form

$$\mathcal{L}(g) = (\frac{1}{2})^n \prod_{j=1}^{k} [(2-g)\varepsilon_j + g(1-\varepsilon_j)] \prod_{j=k+1}^{n} [(2-g)(1-\varepsilon_j) + g\varepsilon_j]$$

Easy to compute !

Probabilistic approach

Prior on genotype

$$P(G_i|D) = \frac{P(G_i)P(D|G_i)}{P(D)}$$

How do we compute the prior on genotypes ?

$P(G_i)$: prior on genotypes

- Single sample case and diploïd. The current practice is a uniform prior on the 3 genotypes $\{0/0, 0/1, 1/1\}$
- Multiple samples case : all the different genotype configurations are not equally likely, see below

Genotype likelihood : synthèse

- La génotype d'un individu est celui qui maximise la vraisemblance
- C'est à dire $P(D|G)$ la probabilité des données sachant le génotype
- Une mesure d'incertitude est fournie pour chaque génotype sous la forme d'une différence entre les log-vraisemblance des différents génotypes

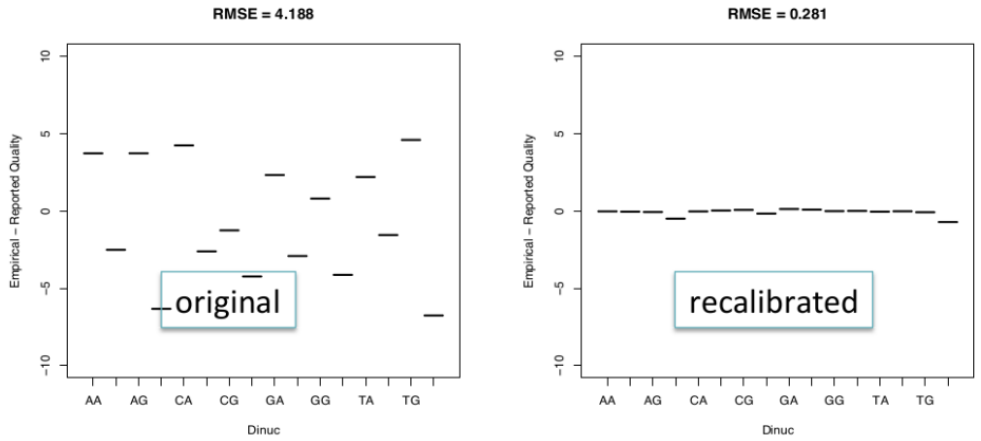
Probabilistic approach

Complications

Complications

- Base quality score are not correctly calibrated $\Longrightarrow$ BQSR
- Alignments around indels are problematic $\Longrightarrow$ Indel Realignment
- Considering a single column is problematic $\Longrightarrow$ use haplotypes

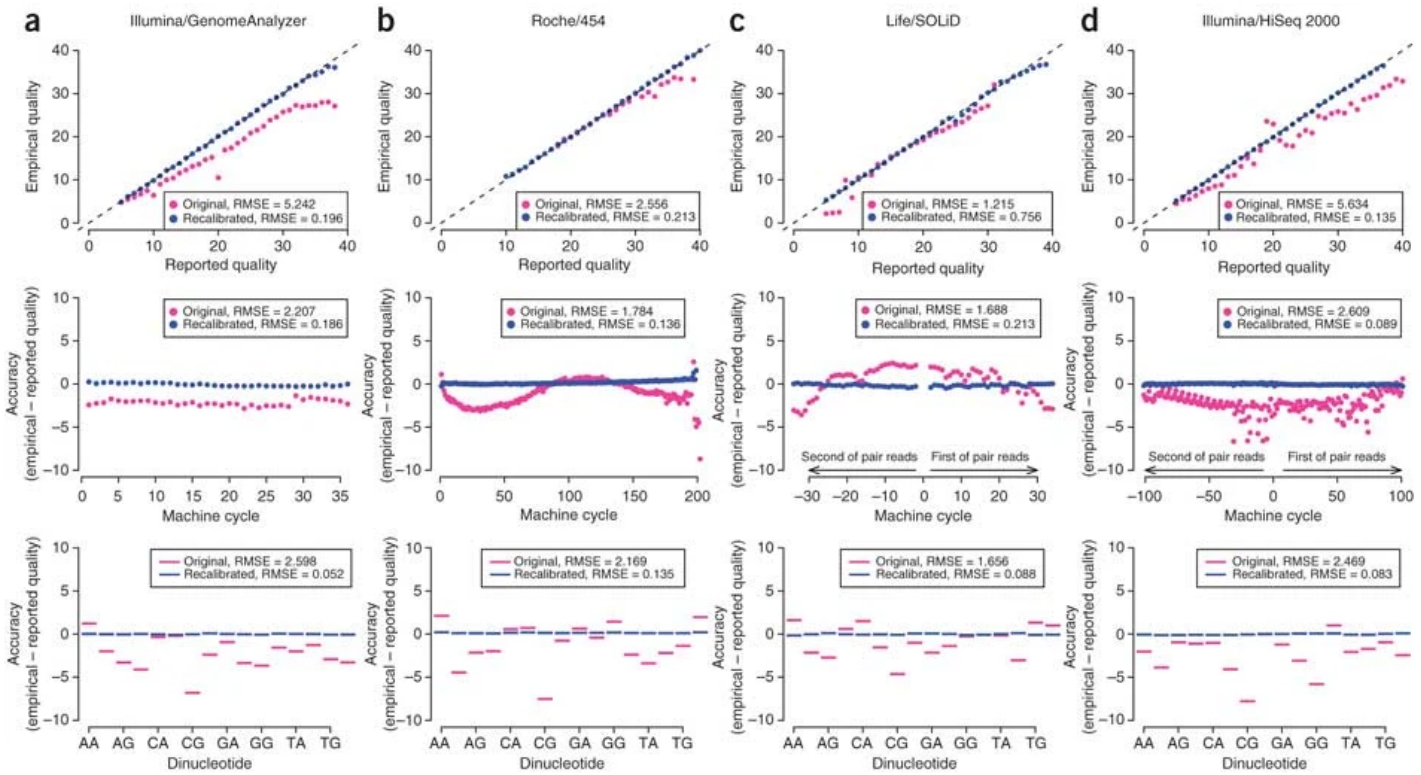
BQSR : Base Quality Score Recalibration



- Error rates can be corrected from the set of reads
- From all the reads, tabulate in bins (original quality score, position in read, dinucleotide context)
- In each bin, count the number of bases and the number of mismatches $\Longrightarrow$ new error estimates

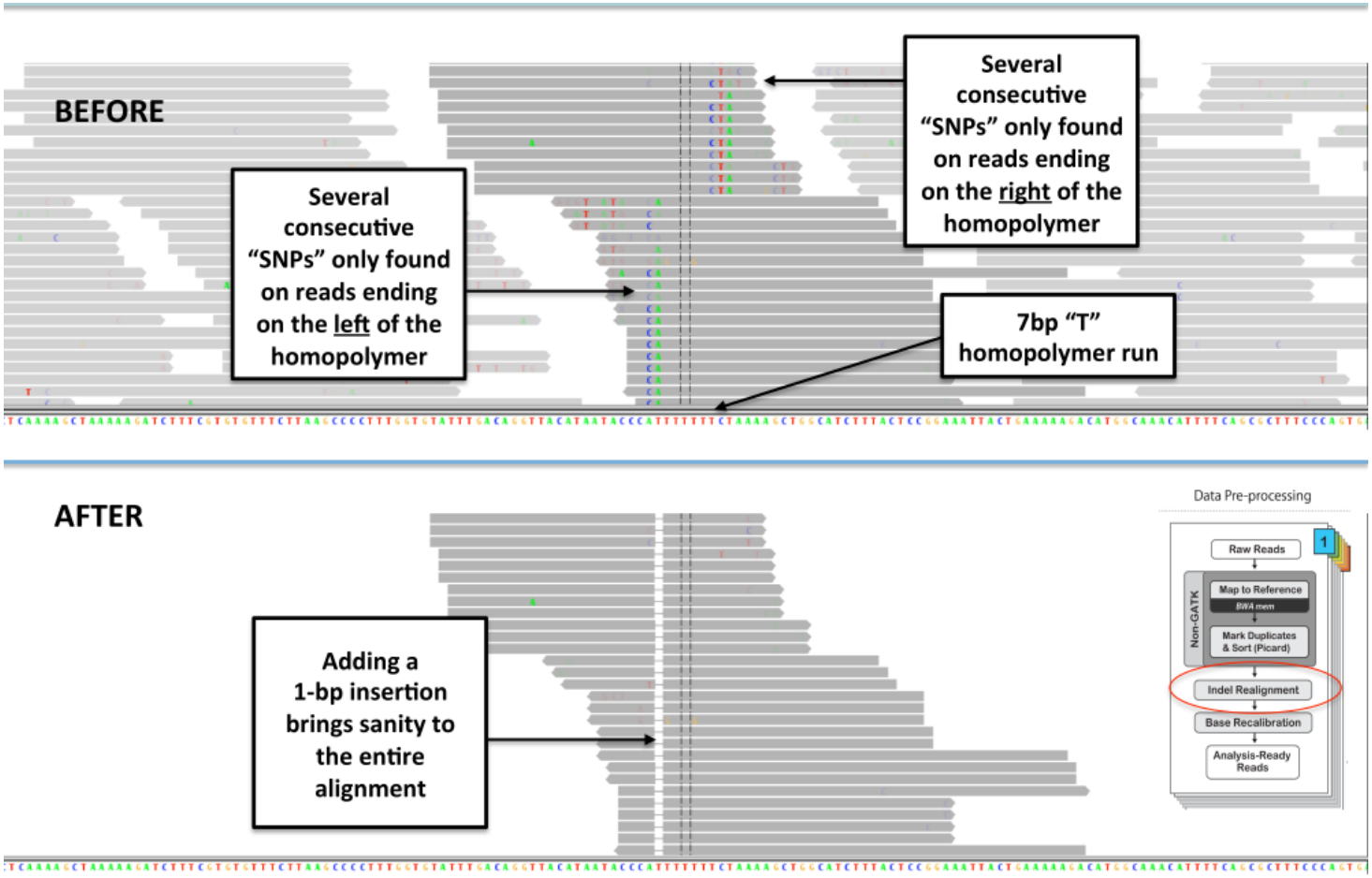
GATK a primer

BQSR : Base Quality Score Recalibration



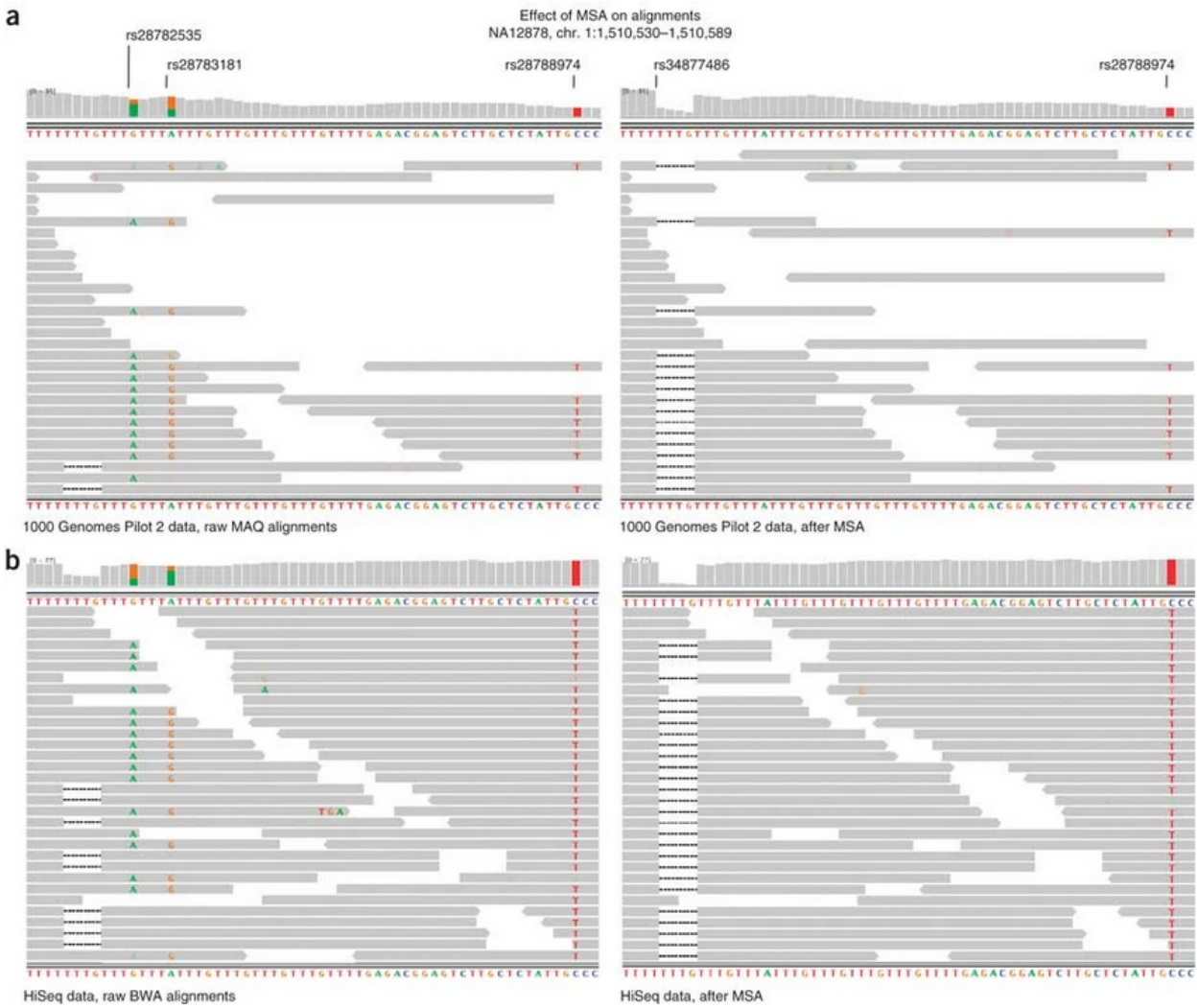
A framework for variation discovery and genotyping using next-generation DNA sequencing data

Indel realignment



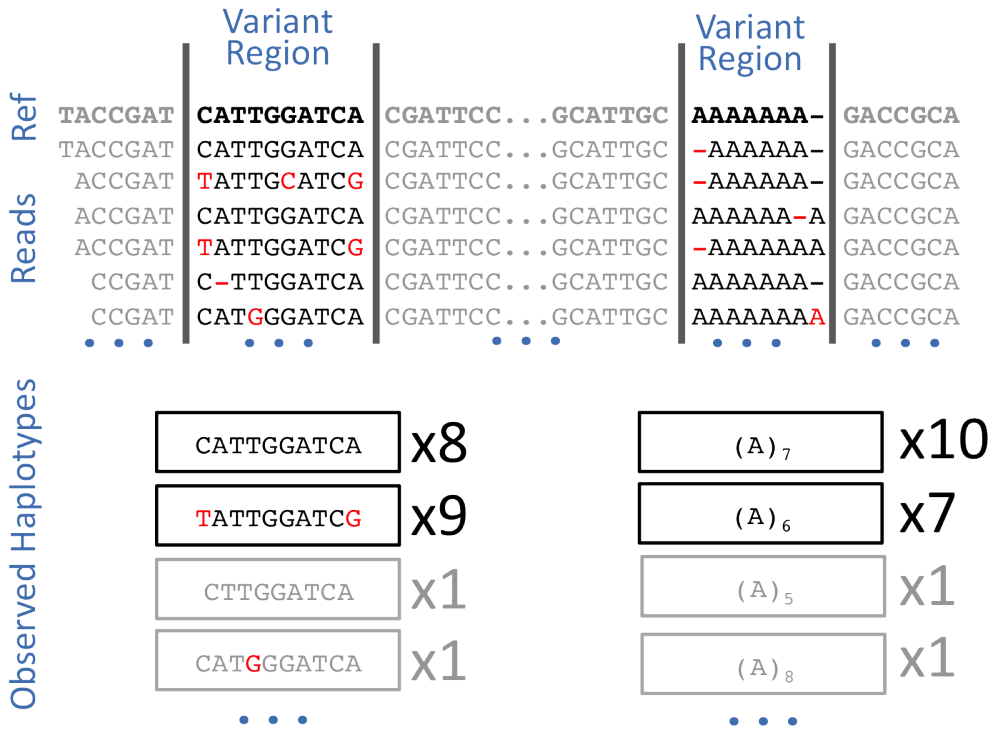
GATK a primer

Indel realignment



Haplotypes with Freebayes

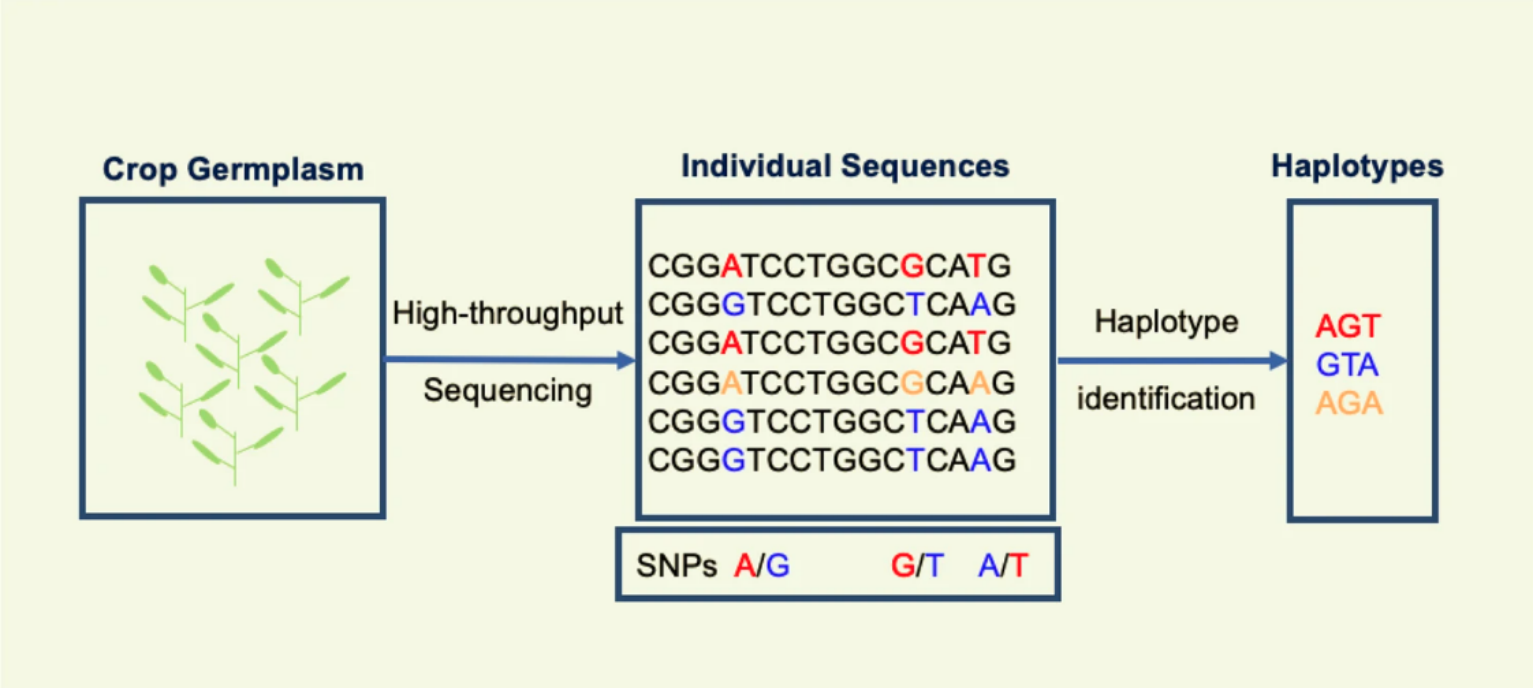
- Indels, and more generally alignments are sometime problematic $\Longrightarrow$ seeking to free itself from alignments



Frebayes

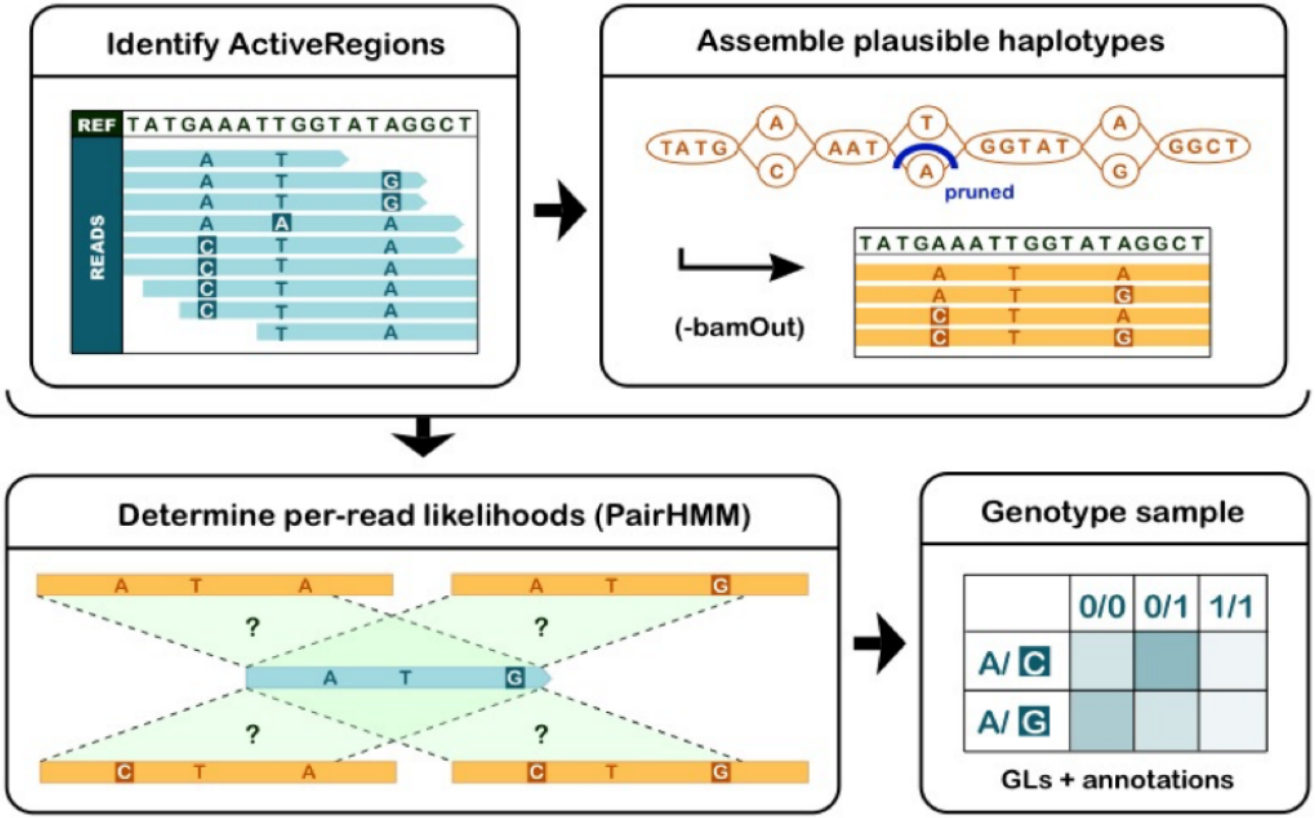
Haplotype : recall

- Haplotypes describe the succession of alleles along a chromosome (segment)



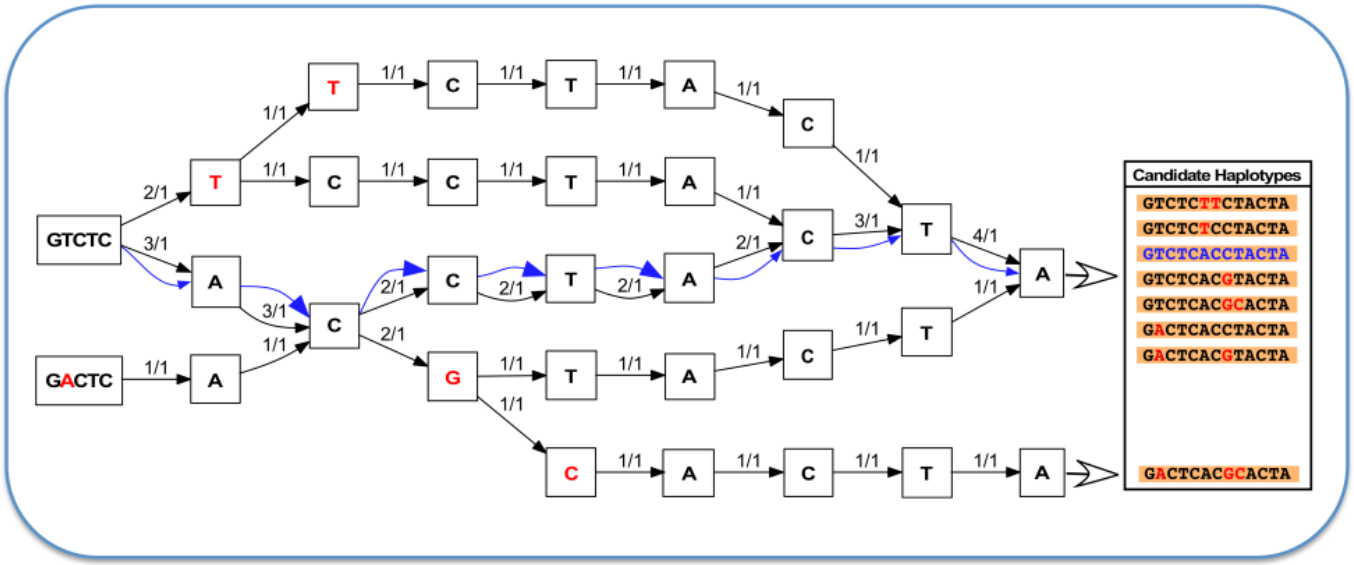
Features and applications of haplotypes in crop breeding

Haplotypes with HaplotypeCaller



A framework for variation discovery and genotyping using next-generation DNA sequencing data

Haplotypes with HaplotypeCaller



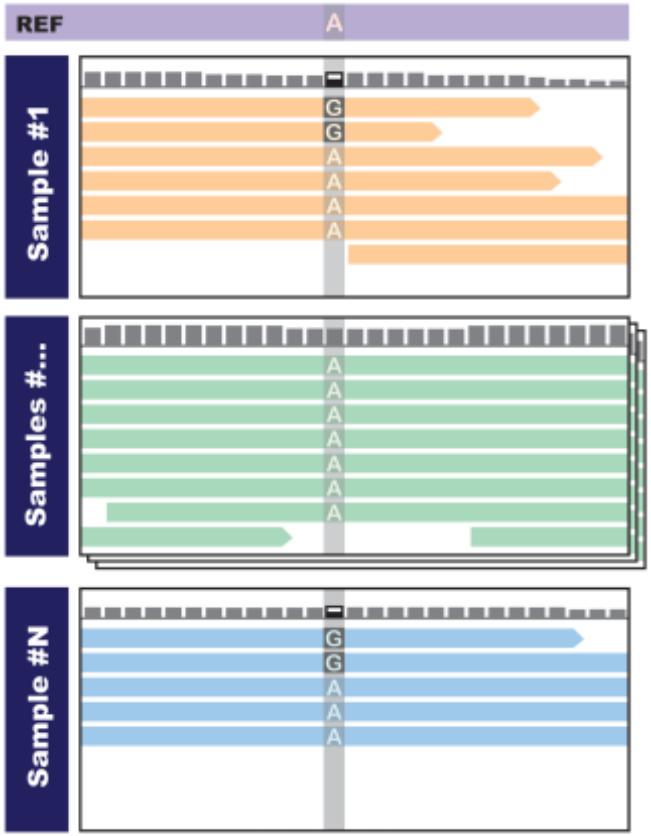
- Previous alignments are ignored
- K-mers consist of every possible sequence combination based on the reads
- Most likely paths through the graph are scored

GATKwr12-5-Variant$_c$*alling*$_j$*oint*$_g$*enotyping*

Probabilistic approach

Multiple individuals

Multiple samples

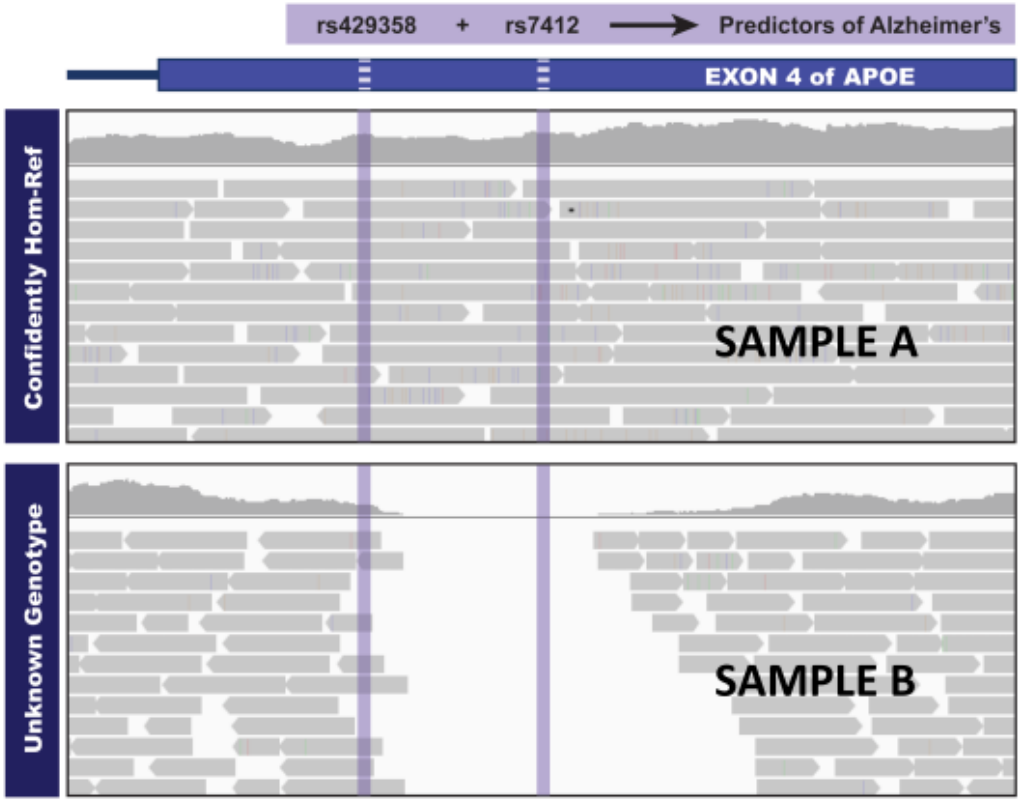


- Sample #1 or Sample #N alone:
  - **weak evidence for variant**
  - **may miss calling the variant**
- Both samples seen together:
  - **unlikely to be artifact**
  - **call the variant more confidently**

GATKwr12-5-Variant$_c$*alling*$_j$*oint*$_g$*enotyping*

Multiple samples

- **Analyzed individually:**
  - No call for either sample
  - Very different reasons!
- **In joint analysis with other samples:**
  - Hom-ref call and no-call genotypes emitted



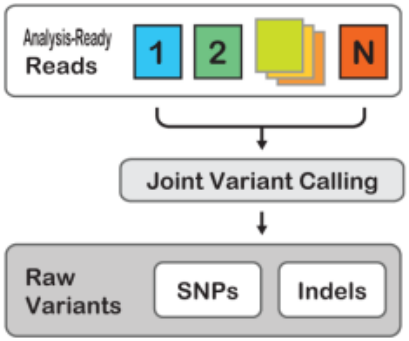
GATKwr12-5-Variant$_{c}alling_{j}oint_{g}enotyping$

Probabilistic approach

$$P(G_1, \ldots, G_n | R_1, \ldots, R_n) = \frac{P(R_1, \ldots, R_n) P(R_1, \ldots, R_n | G_1, \ldots, G_n)}{P(R_1, \ldots, R_n)}$$

- $G_1, \ldots, G_n$: the possible genotypes for the n individuals
- $R_1, \ldots, R_n$: data, i.e. a column of the multiple alignment of reads for all individuals
- The major difference is that the prior probability on genotypes, $P(R_1, \ldots, R_n)$ are generally not taken as uniform
- All the different genotype configurations are not equally likely : for example say 10 $0/0$, 20 $0/1$ and 150 $1/1$) (think of Hardy Weinberg) $\Longrightarrow$ population priors (freebayes, bcftools call, GLnexus)

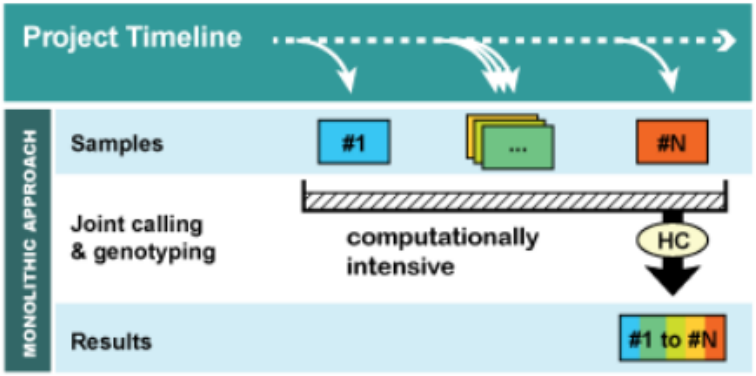
Multiple samples



Compute requirements scale exponentially with number of samples

(combinatorial problem)

It gives us the right answers, but...



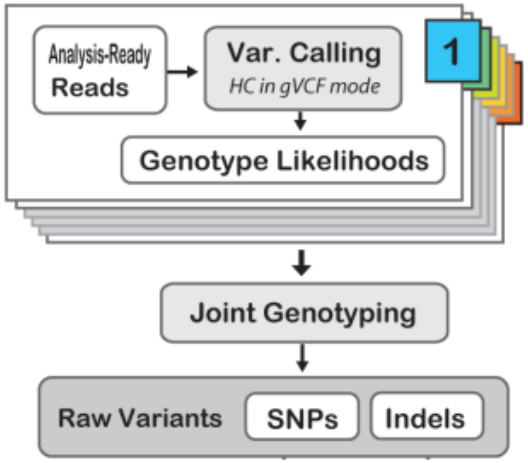
Want to add new samples?

Got to re-run pipeline from scratch!

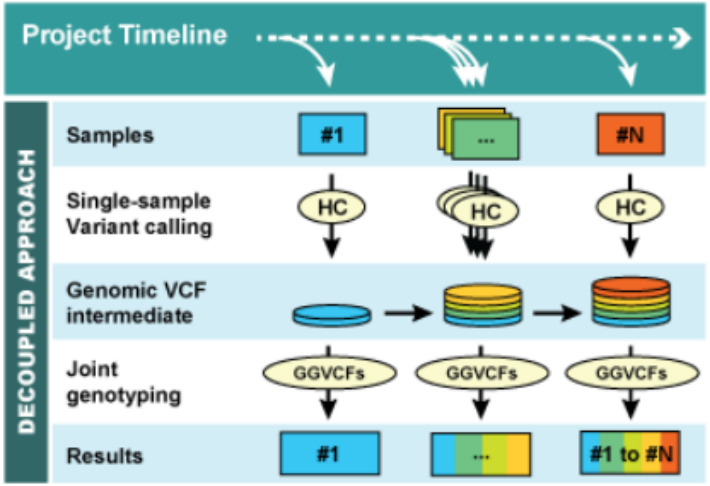
GATKwr12-5-Variant$_{c}alling_{j}oint_{g}enotyping$

Multiple samples : gvcf files

Scalable over sample size



+ Incremental over time



GATKwr12-5-Variant$_c$*alling*$_j$*oint*$_g$*enotyping*

Variant quality score

Each variant is associated with an uncertainty measure (field QUAL in vcf file), coded as a Phred score

$$QUAL = -10\left\lfloor \log_{10}(P[\text{not a variant site}]) \right\rfloor$$

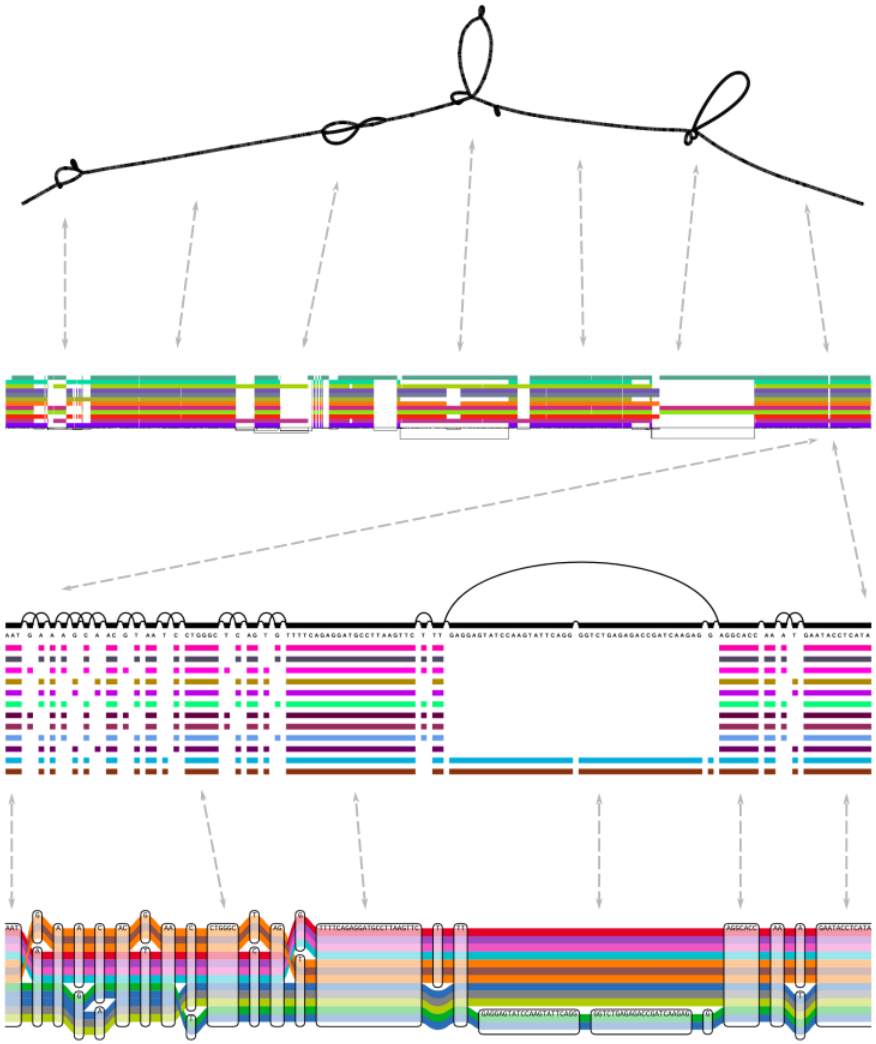
$$= -10\left\lfloor \log_{10}(P(G = 00|D) \right\rfloor$$

In the case of multiple individuals it measures the probability of having at least one anlternate allele.

Variant detection today

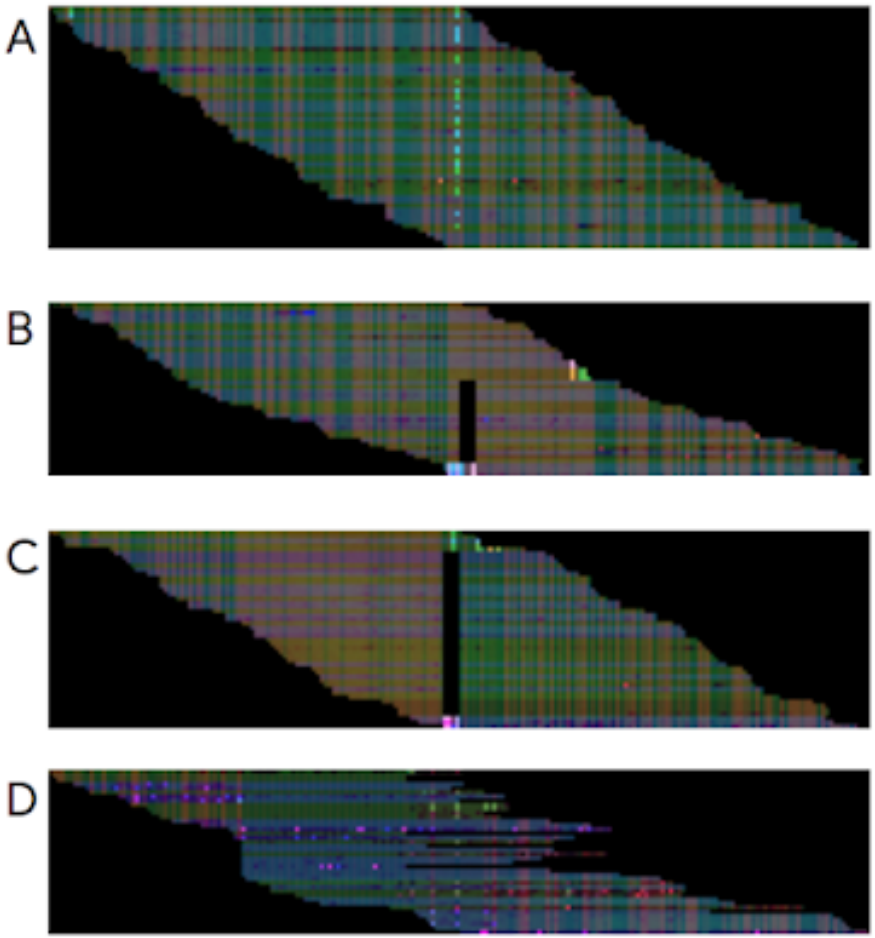
- Structural variant detection and variation graphs
  - In order to bypass the bias induced by the use of a single reference assembly, current approaches aim at combining multiple genomes in a single graph structure and use this structure as a reference on which the reads are mapped
- Variant detection using deep learning
  - Recent developments in variant discovery use deep learning approach to distinguish real variants from sequencing errors or alignment artifacts

Variation graphs



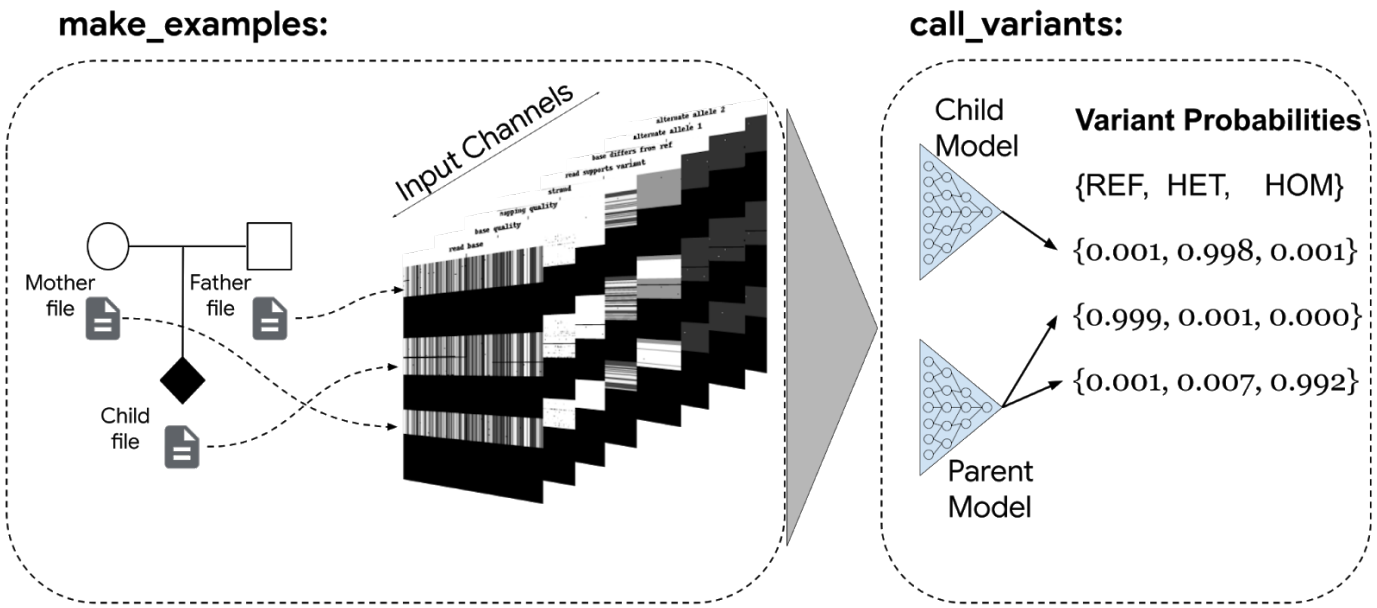
Sequence tube maps: making graph genomes intuitive to commuters

Deep learning for Variant detection



A universal SNP and small-indel variant caller using deep neural networks

Deep learning for Variant detection



A universal SNP and small-indel variant caller using deep neural networks

Deep learning inbioinformatics

Johann Gregor Mendel: the victory of statistics over human imagination