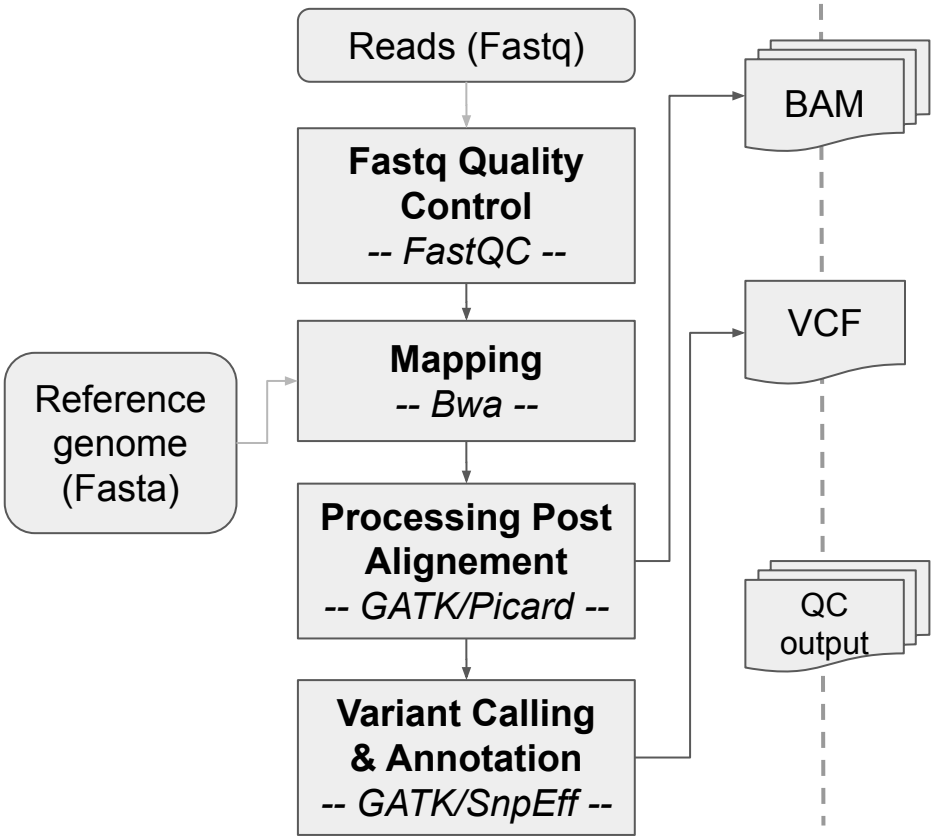
Analyse des variants post-VCF

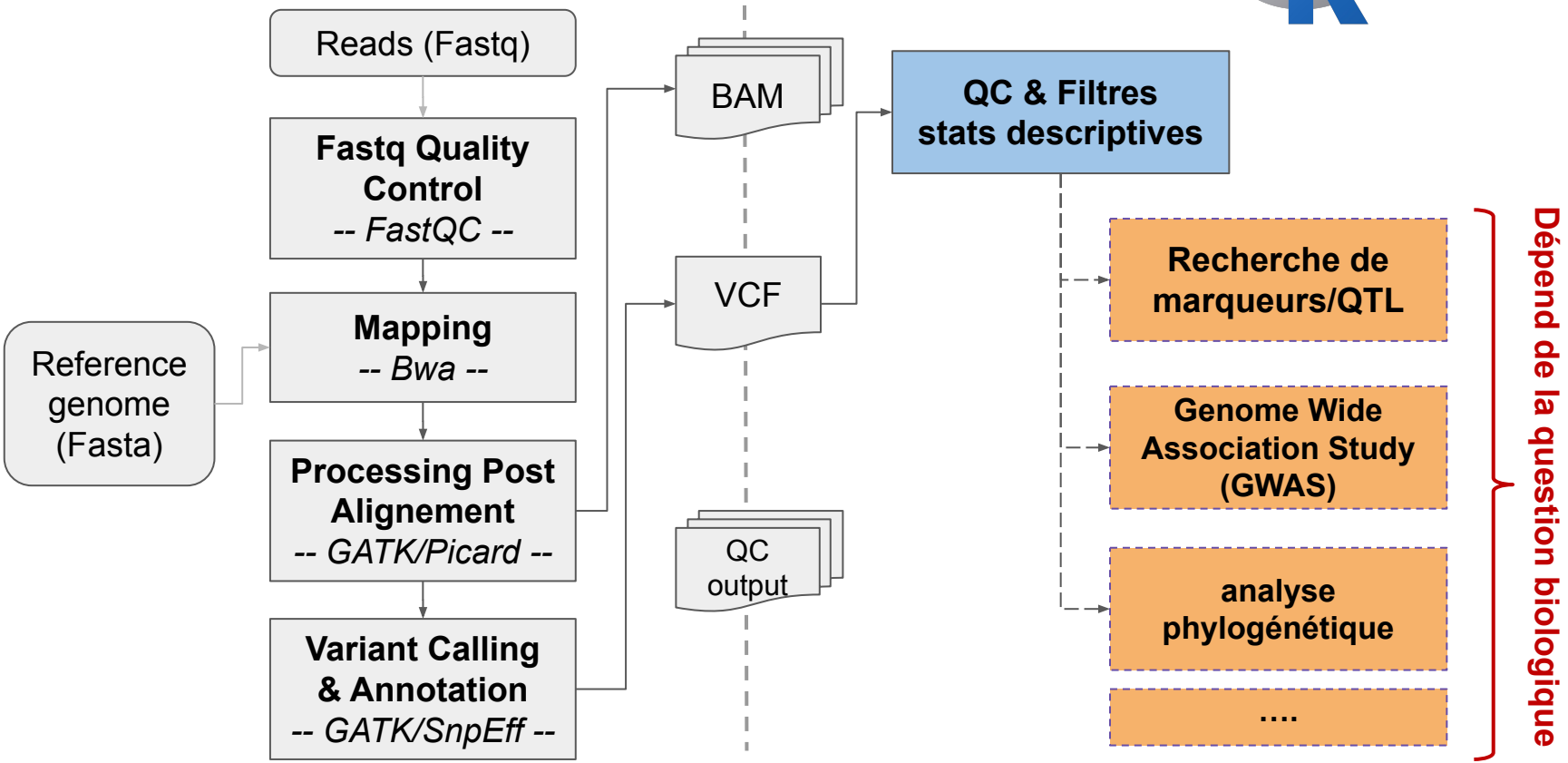
Nadia Bessoltane - INRAE
Vivien Deshaies - AP-HP

École de bioinformatique AVIESAN-IFB-INSERM 2022

Workflow


Reads (Fastq)
Fastq Quality Control
-- FastQC --
Mapping
-- Bwa --
Reference genome (Fasta)
Processing Post Alignement
-- GATK/Picard --
Variant Calling & Annotation
-- GATK/SnpEff --
BAM
VCF
QC output

Workflow : Post-VCF



Rappel : VCF

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

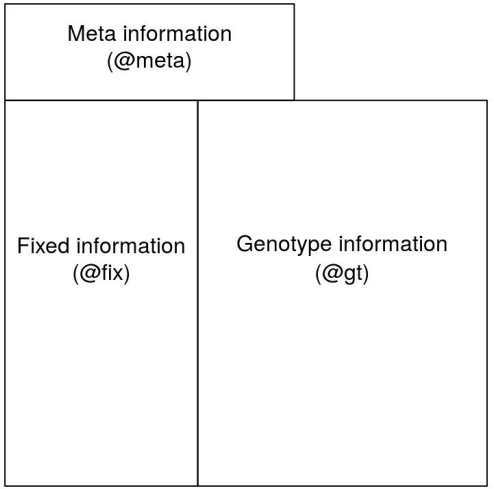
vcfR package

Manipulate and Visualize VCF Data

- vcfR documentation : https://knausb.github.io/vcfR_documentation/index.html

```
> # lire le fichier vcf
> my.vcf <- read.vcfR("pool.vcf")
> # l'objet vcf appartient à quelle class
> is(my.vcf)
[1] "vcfR"
> # la liste des slots (sections)
> slotNames(my.vcf)
[1] "meta" "fix"  "gt"
>
```

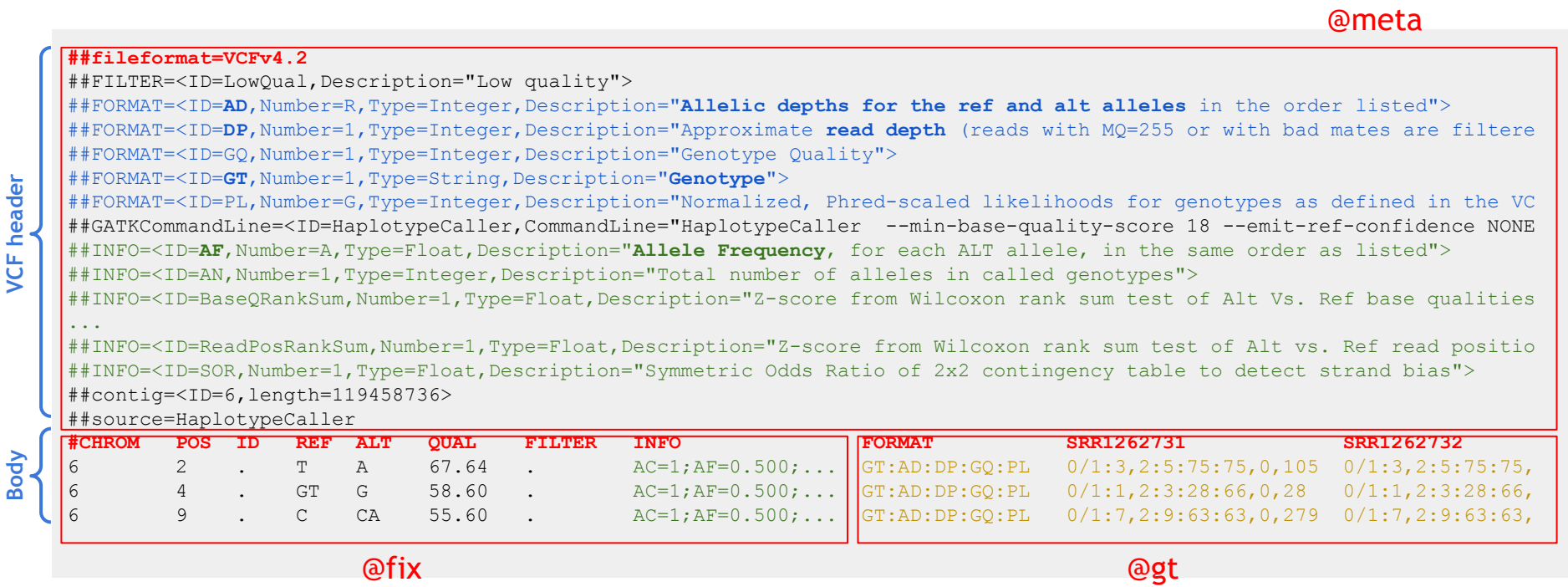
objet de la classe vcfR



Trois sections :

- meta-information : entête du vcf
- Fixed information : information par variant mais commune à tous les échantillons (position, allèles, qualité…)
- Genotype information : information de génotypage par échantillon

objet de la classe vcfR



vcfR package

Manipulate and Visualize VCF Data

- vcfR documentation : https://knausb.github.io/vcfR_documentation/index.html

```
> # lire le fichier vcf
> my.vcf <- read.vcfR("pool.vcf")
> # l'objet vcf appartient à quelle class
> is(my.vcf)
[1] "vcfR"
> # la liste des slots (sections)
> slotNames(my.vcf)
[1] "meta" "fix"  "gt"
> # convertir l'objet vcfR à une liste de
tibbles
> vcf.list <- vcfR2tidy(my.vcf)
> is(vcf.list)
[1] "list"
> names(vcf.list)
[1] "meta" "fix"  "gt"
>
```

TP 1 : recherche de mutation dans un QTL

Jeux de données #1:

Chez le bovin, il existe un locus de caractères quantitatifs (**QTL**) lié à la production de lait, situé sur le chromosome 6, et plus exactement sur une région de 700 kb, composée de 7 gènes.

Les échantillons QTL+ sont caractérisés par une diminution de la production en lait et une augmentation des concentrations en protéine et lipide.

Quelle mutation est responsable de ce QTL ?

Pour le TP nous disposons des résultats du variant calling de 3 échantillons (en Multi-VCF annoté).

Echantillons	Phénotype	Source
SRR1262731	QTL-	projet 1000 génomes bovins
SRR1205992	QTL+	
SRR1205973	QTL+	

Miri Cohen-Zinder et al, Genome Research 2005, DOI:10.1101/gr.3806705

Préparation des données

1- Se connecter sur Rstudio via JupyterLab

2- Copier le matériel de TP dans le dir TP_variants

```
> file.copy(from = “/shared/projects/2325_ebaii/atelier_variant/TP_R”,
            to   = “~/”, recursive=TRUE)
```

2- Positionner l’espace du travail

```
> setwd(“~/TP_R”)
```

5- Changer le workspace sur l’interface RStudio.

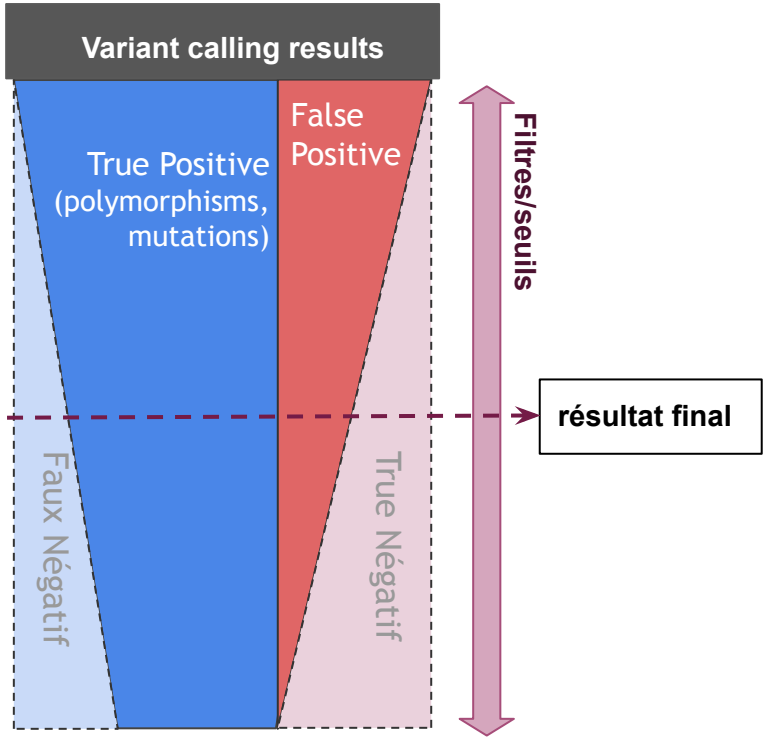
RMD

Rappel : Filtre des variants

RMD

- 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…)

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

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…)

  → GATK Bests Practices : recommendations selon des métriques spécifiques à GATK, différentes pour les SNVs des Indels

  - **QD** - QualByDepth : Score QUAL / AD [profondeur allélique] (>2)
  - **FS** - FisherStrand (<60) :
  - **SOR** - StrandOddsRatio (<3):

  } Score estimant un éventuel biais de brin

  - **MQ** - MappingQuality : Qualité de mapping moyenne sur l'ensemble du read (>40)
  - **MQRankSum** : Teste un biais de différence de qualité de mapping entre allèles (>-12,5)
  - **ReadPosRankSum** : Teste un biais de position des allèles le long du read (>-8.0)

Variant d'intérêt

- Quelle type de mutation est impliquée dans notre phénotype d'intérêt pour l'individu SRR1262731 ?
- Quel est son génotype ? Sur quel gène se situe-elle ?
- Qu'en est-il pour les autres individus ?

RMD

Zinder *et al.*, 2005

Variant d'intérêt

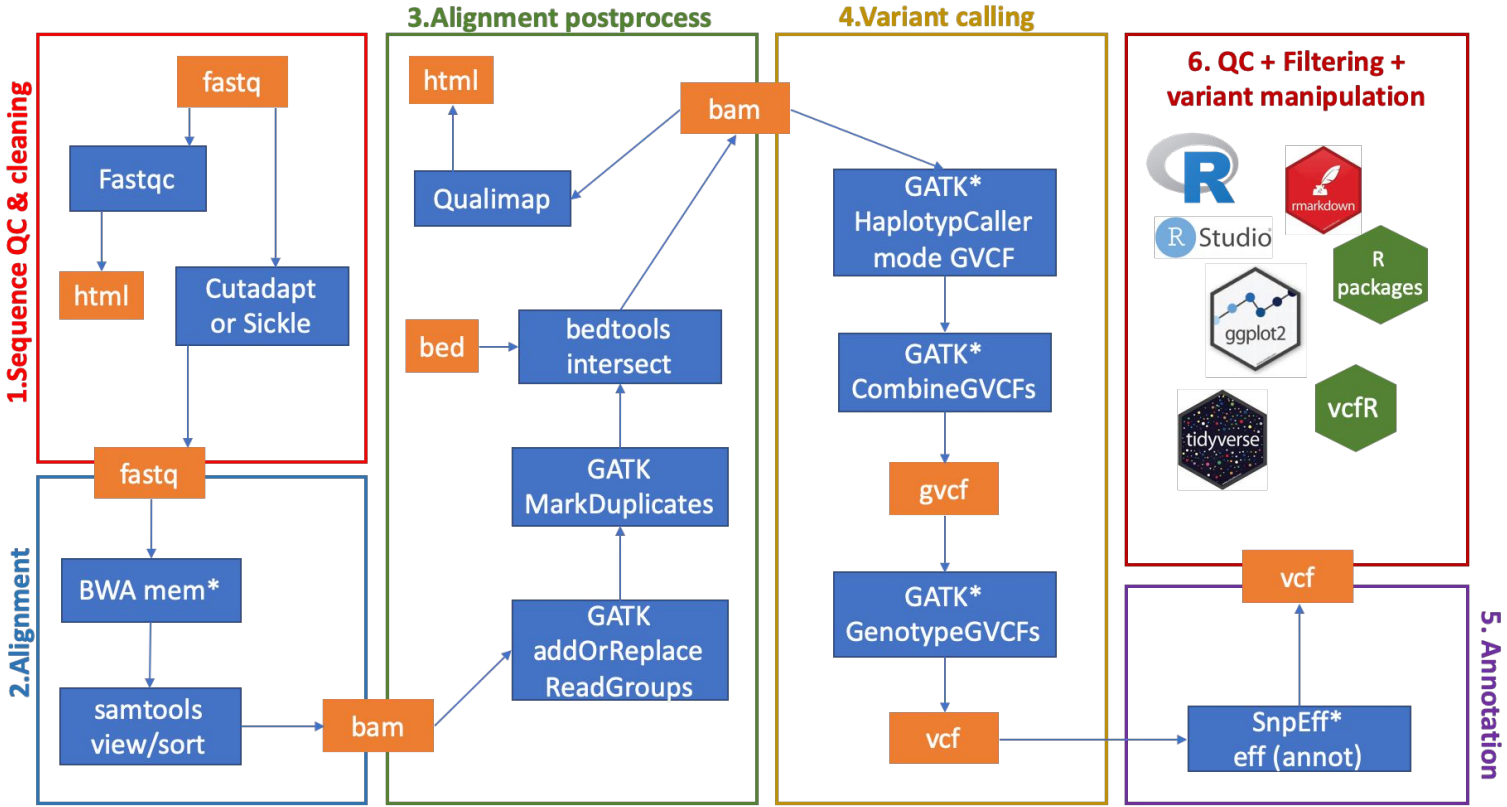
- Quelle type de mutation est impliquée dans notre phénotype d'intérêt pour l'individu SRR1262731 ?
- Quel est son génotype ? Sur quel gène se situe-elle ?
- Qu'en est-il pour les autres individus ?

→ Le variant est **hétérozygote ALT (0/1)** pour l'individu SRR1262731, il comporte une mutation de type SNP (A → C) située sur le gène **ABCG2**, en position **38027010 du chromosome 6.**

→ Pour les deux autres individus, ils ne comportent pas cette mutation : il sont homozygote référence (GT: 0/0).

Zinder *et al.*, 2005

Rappel : du fastq au VCF



* need specific index