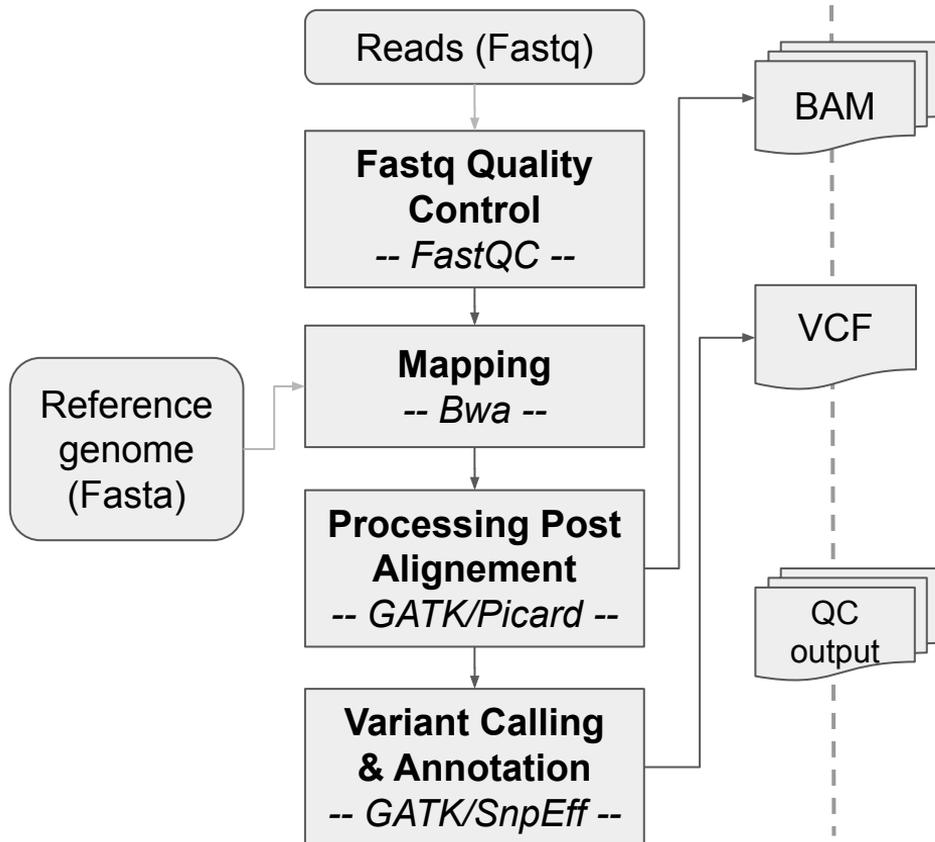




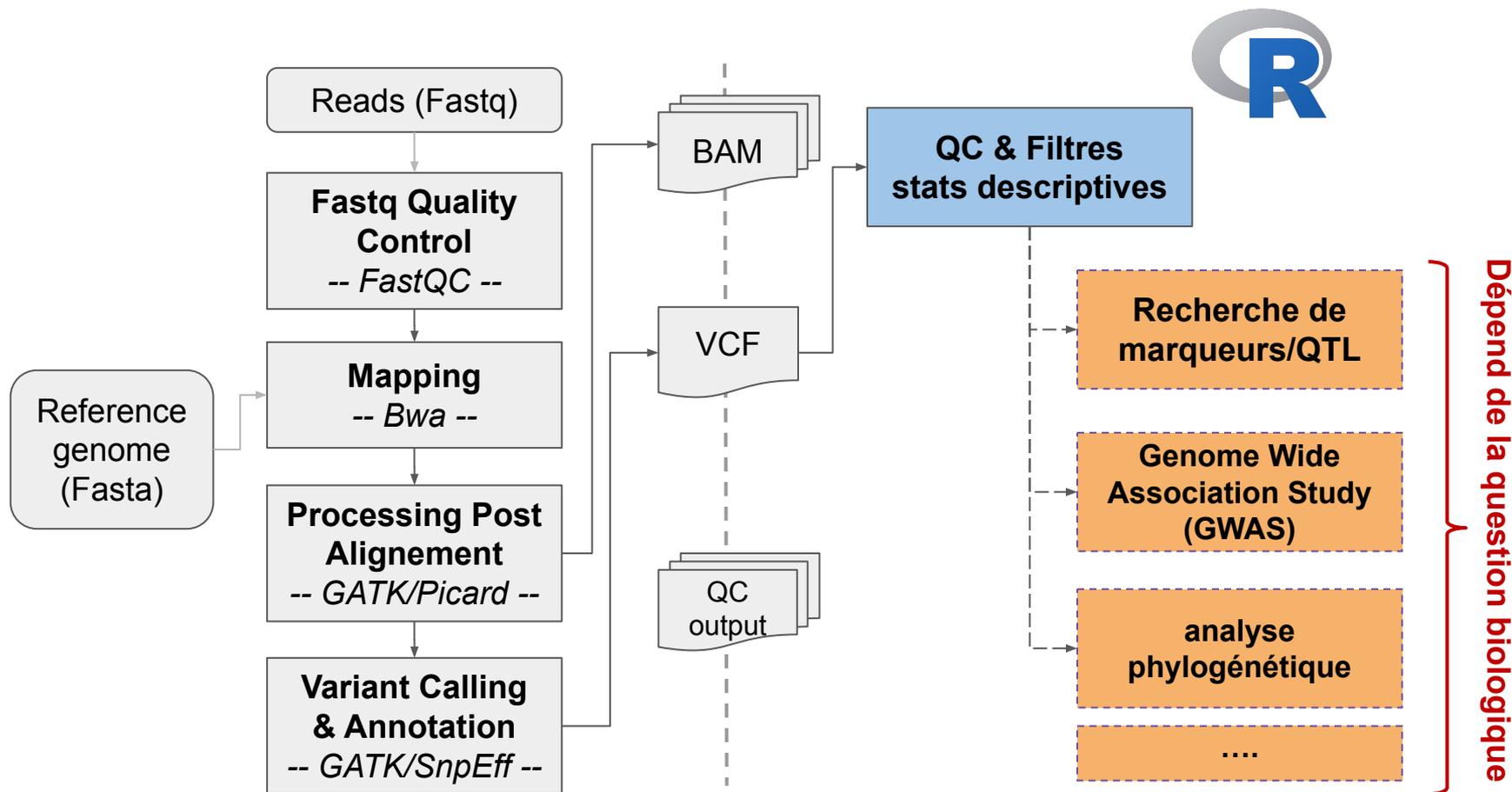
Analyse des variants post-VCF

Nadia Bessoltane - INRAe

Workflow



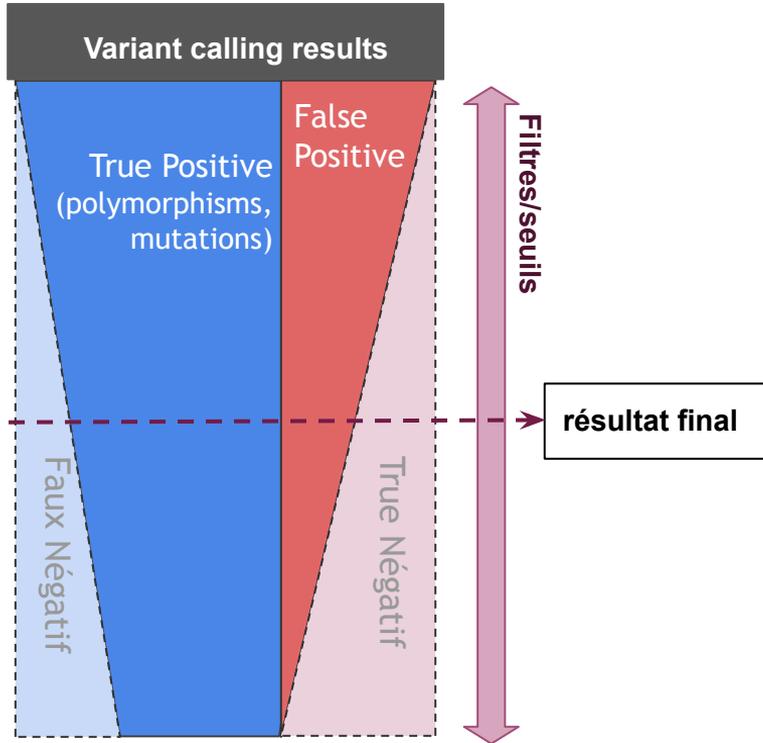
Workflow : Post-VCF



Comment procéder ?

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

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 : Faux positifs

*Rappel : 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, ...)
- Difficultés d'alignement (régions d'ADN répétées)
- Erreurs de lecture lors du “BaseCalling”

Comment procéder ?

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

→ Le reste dépend de la **Question biologique**

- Filtres supplémentaires
- Analyses statistiques
- Représentations graphiques
-

Filtre des variants : GATK Bests Practices recommandation

→ **GATK Bests Practices** : recommandations 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):
 - **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)
- } Score estimant un éventuel biais de brin

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+	

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 ?

Rappel : VCF

METADATA

```
##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 filtered)">
##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 VCF specification">
##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 positions" -->
##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
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
6	2	.	T	A	67.64	.	AC=1;AF=0.500;...
6	4	.	GT	G	58.60	.	AC=1;AF=0.500;...
6	9	.	C	CA	55.60	.	AC=1;AF=0.500;...

FORMAT	SRR1262731	SRR1262732
GT:AD:DP:GQ:PL	0/1:3,2:5:75:75,0,105	0/1:3,2:5:75:75,
GT:AD:DP:GQ:PL	0/1:1,2:3:28:66,0,28	0/1:1,2:3:28:66,
GT:AD:DP:GQ:PL	0/1:7,2:9:63:63,0,279	0/1:7,2:9:63:63,

INFO

GENOTYPE

VCF header

Body

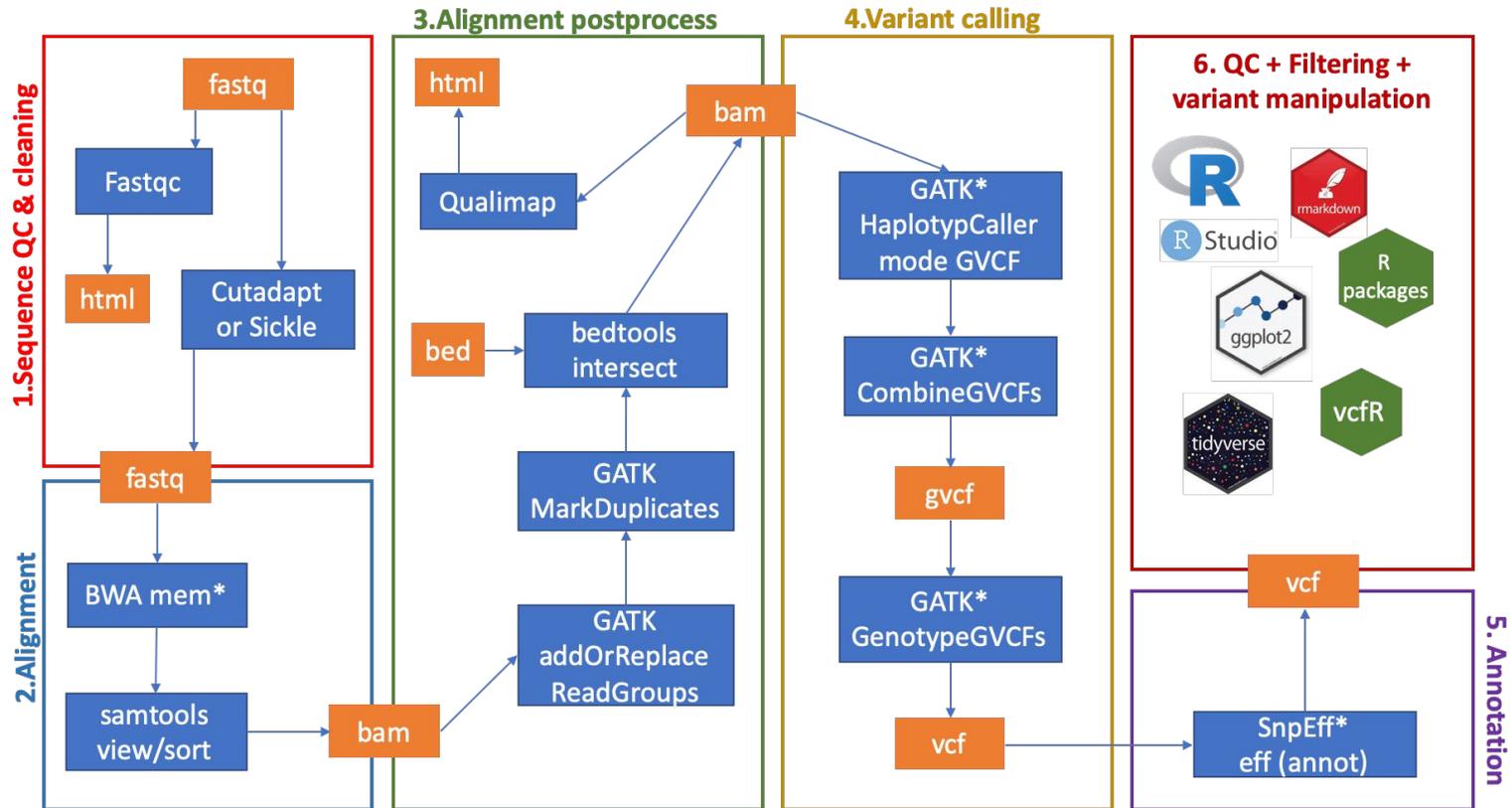
Variant d'intérêt

- 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 : ils sont homozygote référence (GT: 0/0).

Rappel : du fastq au VCF



* need specific index