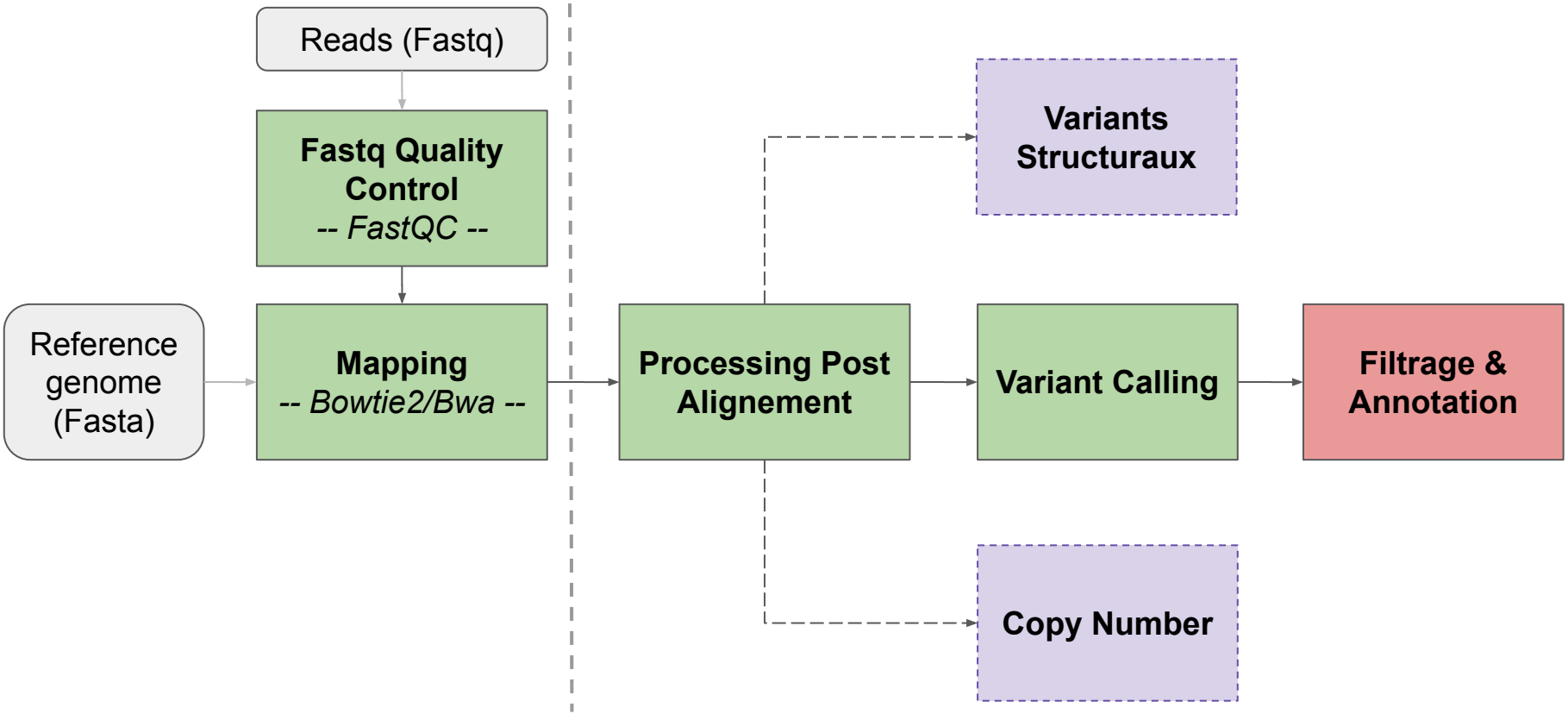
Filtrage & Annotation

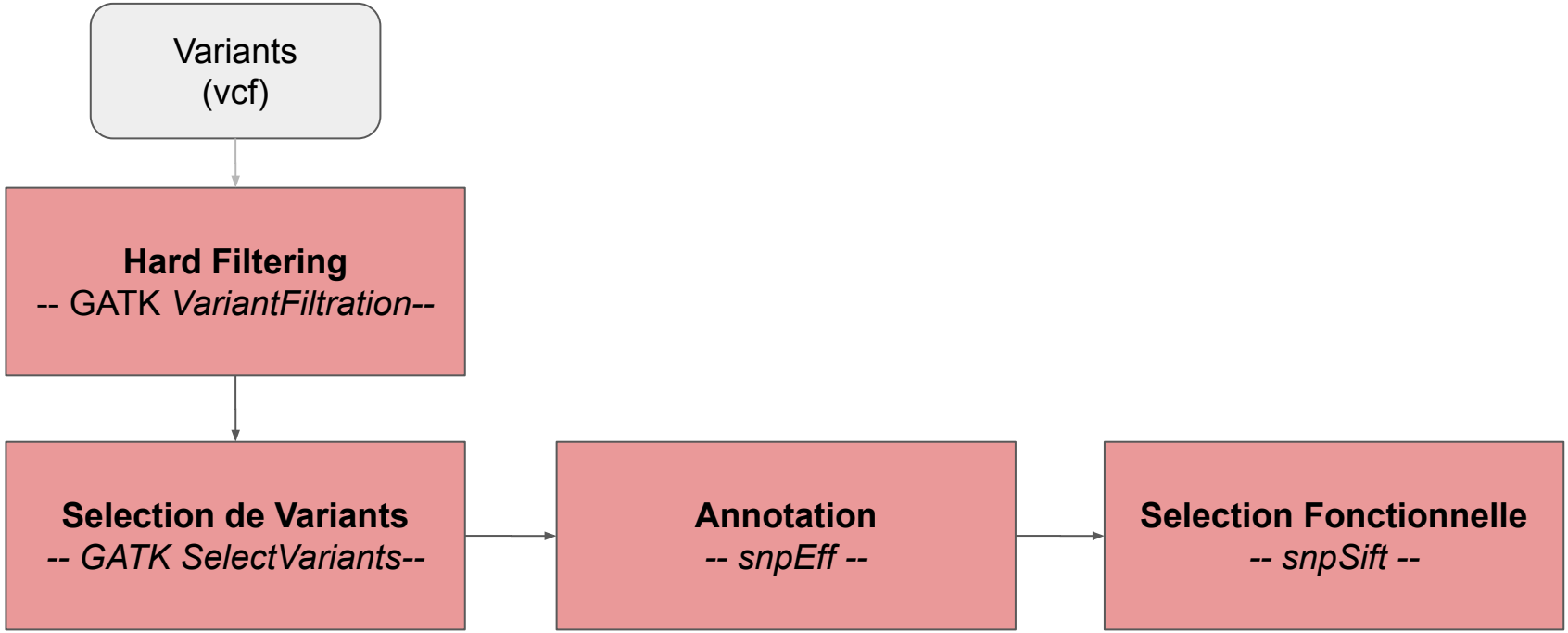
Mathieu Charles - INRAE

École de bioinformatique AVIESAN-IFB-INSERM 2021

Workflow


Reads (Fastq)
Fastq Quality Control
-- FastQC --
Reference genome (Fasta)
Mapping
-- Bowtie2/Bwa --
Processing Post Alignement
Variants Structuraux
Variant Calling
Filtrage & Annotation
Copy Number

Workflow - Filtrage et Annotation



Filtres des variants

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

- Filtres difficilement transposables entre analyse :

  → dépendent de la **question biologique**
  → dépendent des outils utilisés

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

SelectVariants et Hard filtering

```
# Préparation d'un nouveau répertoire de résultats
$ mkdir -p ~/tp_variant/filter_and_annot/logs
$ cd ~/tp_variant/filter_and_annot

# Extraction des SNVs dans un fichier séparé pour GATK
$ sbatch -J GATK_SNP -o logs/GATK_SNP.out -e logs/GATK_SNP.err --mem=8G --wrap=" \
    gatk SelectVariants --java-options '-Xmx8G' \
    -R ~/tp_variant/genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa \
    -V ~/tp_variant/GATK/vcf/pool_GATK.vcf \
    --select-type SNP -O pool_GATK.SNP.vcf"

# Extraction des SNVs dans un fichier séparé pour Varscan
$ sbatch -J Varscan_SNP -o logs/Varscan_SNP.out -e logs/Varscan_SNP.err --mem=8G
--wrap="gatk SelectVariants --java-options '-Xmx8G' \
    -R ~/tp_variant/genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa \
    -V ~/tp_variant/Varscan/pool_Varscan_dict.vcf \
    --select-type SNP -O pool_Varscan.SNP.vcf"
```

SelectVariants et Hard filtering

- **QD** - QualByDepth : Score QUAL / AD [profondeur allélique]
- **FS** - FisherStrand :
- **SOR** - StrandOddsRatio:

} Score estimant un éventuel biais de brin

- **MQ** - MappingQuality : Qualité de mapping moyenne sur l'ensemble du read
- **MQRankSum** : Teste un biais de différence de qualité de mapping entre allèles
- **ReadPosRankSum** : Teste un biais de position des allèles le long du read

doc GATK

HowTo: Apply hard filters to a call set
I am unable to use VQSR (recalibration) to filter variants
how to understand and improve upon the generic hard filtering recommendations.

SelectVariants et Hard filtering

```
# Filtrage des SNVs selon les filtres recommandés par GATK
$ sbatch -J GATK_SNP_filter -o logs/GATK_SNP_filter.out -e logs/GATK_SNP_filter.err
--mem=8G --wrap="gatk VariantFiltration --java-options '-Xmx8G' \
  -R ~/tp_variant/genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa \
  -V pool_GATK.SNP.vcf -O pool_GATK.SNP.prefilt.vcf \
  -filter 'QD < 2.0' --filter-name 'QD2' -filter 'SOR > 3.0' --filter-name 'SOR3' \
  -filter 'FS > 60.0' --filter-name 'FS60' -filter 'MQ < 40.0' --filter-name 'MQ40' \
  -filter 'MQRankSum < -12.5' --filter-name 'MQRankSum-12.5' \
  -filter 'ReadPosRankSum < -8.0' --filter-name 'ReadPosRankSum-8'"

# Sélection des variants passant ce filtre
$ sbatch -J GATK_SNP_PASS -o logs/GATK_SNP_PASS.out -e logs/GATK_SNP_PASS.err
--mem=8G --wrap="gatk SelectVariants --java-options '-Xmx8G' \
    -R ~/tp_variant/genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa \
    -V pool_GATK.SNP.prefilt.vcf \
    --exclude-filtered \
    -O pool_GATK.SNP.filtered.vcf"
```

Intersection des résultats des variant callers

```
# Intersection des variants obtenus avec Varscan et avec GATK post filtering

# Compression et indexation des fichiers vcfs
$ bgzip -c pool_GATK.SNP.filtered.vcf > pool_GATK.SNP.filtered.vcf.gz
$ tabix -p vcf pool_GATK.SNP.filtered.vcf.gz

$ bgzip -c pool_Varscan.SNP.vcf > pool_Varscan.SNP.vcf.gz
$ tabix -p vcf pool_Varscan.SNP.vcf.gz

$ sbatch -J GATK_varscan_isec -o logs/GATK_varscan_isec.out \
    -e logs/GATK_varscan_isec.err --mem=8G --wrap=" \
    bcftools isec -f PASS -n +2 -w 1 -O v \
    pool_GATK.SNP.filtered.vcf.gz pool_Varscan.SNP.vcf.gz \
    > GATK_varscan_inter.vcf "
```

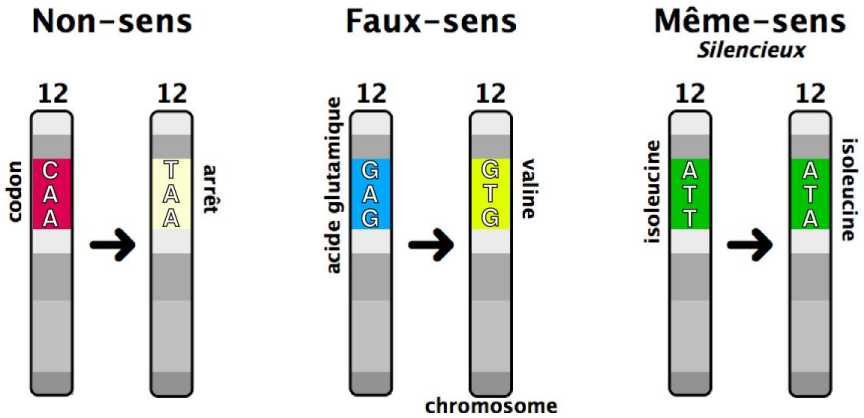
Annotation des variants

- Ajout d'informations biologiques pertinentes aux variants :

→ Est-ce que mes variants sont connus ?

→ Où se positionnent mes variants ?

→ Quel est l'effet d'une mutation sur le CDS qui le contient ?

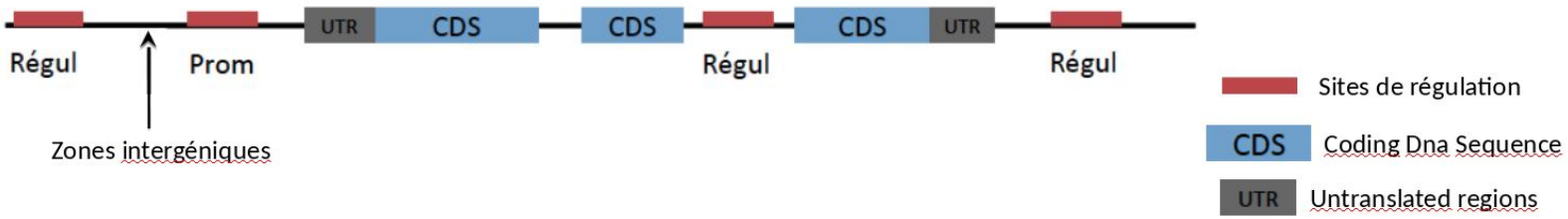


Annotation des variants

- Annotation structurale :
  → Mon variant se trouve-t-il dans un intron, un exon ?

- Annotation fonctionnelle :
  → Informations sur la région ? Exemple : CDS codant pour une protéine

- Impacts potentiels :
  → Dans le cas d'un CDS, protéine produite tronquée, allongée, décalée… ou silencieuse (redondance du code génétique)



Annotation des variants

- Nécessité d'avoir des bases de données associées aux organismes étudiés (Ensembl, Refseq...)

- Exemples d'**outils/algorithmes** :
  → SnpEff
  → VEP
  → Annovar
  → SIFT, POLYPHEN2, CADD…

SnpEff

```
# Création de la base de données SnpEff
$ module load snpeff/4.3.1t
$ snpEff -version                # affiche la version (v4.3t)

$ echo BosTaurus.genome >> snpeff.config        # <genome_name>.genome
$ mkdir -p BosTaurus
$ cp ~/tp_variant/genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa BosTaurus/sequences.fa
$ cp ~/tp_variant/genome/Bos_taurus.UMD3.1.93.chromosome.6.gff3 BosTaurus/genes.gff
$ echo -e "BosTaurus\nSnpEff4.3t" > BosTaurus.db

$ sbatch -J snpeffBuild -o logs/snpeffBuild.out -e logs/snpeffBuild.err --mem=8G \
--wrap="snpEff build -c snpeff.config -gff3 -v BosTaurus -dataDir ."
```

```
# Annotation avec notre base de données
$ sbatch -J snpeffAnnot -o logs/snpeffAnnot.out -e logs/snpeffAnnot.err --mem=8G \
--wrap="snpEff eff -c snpeff.config -dataDir . BosTaurus -s snpeff_res.html \
GATK_varscan_inter.vcf > GATK_varscan_inter.annot.vcf"
```

SnpSift

```
$ module load snpsift/4.3.1t
$ SnpSift filter -h                    # affiche l'aide (v 4.3t)

# Garder les variants codant qui ne sont pas des synonymes :
$ sbatch -J snpsift1 -o logs/snpsift1.out -e logs/snpsift1.err --mem=8G --wrap=" \
 cat GATK_varscan_inter.annot.vcf | SnpSift filter -Xmx8G \
 \"(ANN[*].EFFECT != 'synonymous_variant') && (ANN[*].BIOTYPE = 'protein_coding')\" \
   > GATK_varscan_inter.annot.coding.nosyn.vcf"
```

```
# Sélectionner notre variant d'intérêt parmi les variants hétérozygotes ayant un
impact (missense)
$ sbatch -J snpsift2 -o logs/snpsift2.out -e logs/snpsift2.err --mem=8G --wrap=" \
 cat GATK_varscan_inter.annot.coding.nosyn.vcf | SnpSift filter -Xmx8G \
 \"ANN[*].EFFECT = 'missense_variant' & isHet( GEN[2] ) & isVariant( GEN[2] ) \
 & isRef( GEN[0] ) & isRef( GEN[1] ) \" \
> GATK_varscan_inter.annot.coding.nosyn.filtered.vcf"
```

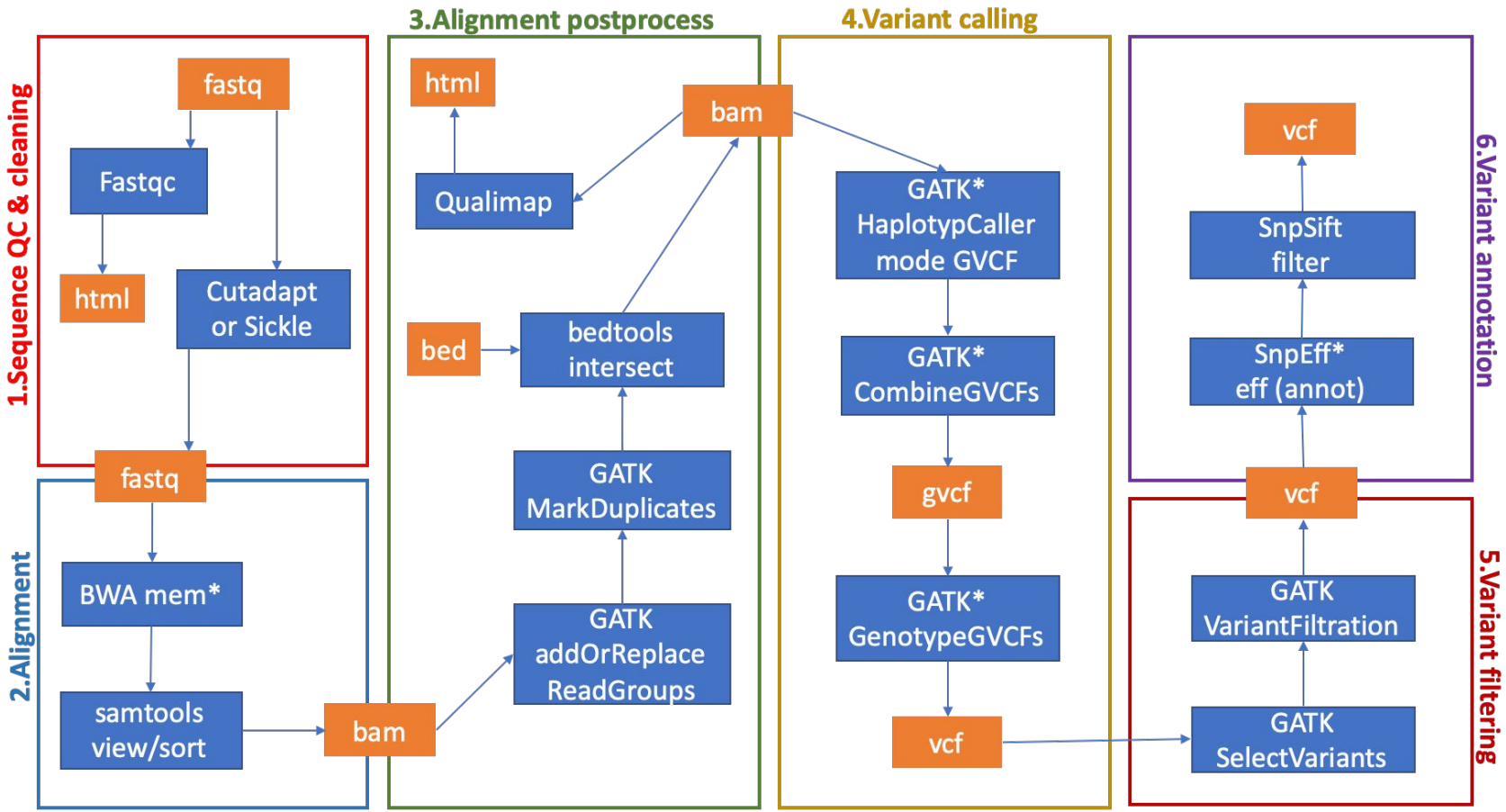
Variant d'intérêt

- 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



* need specific index