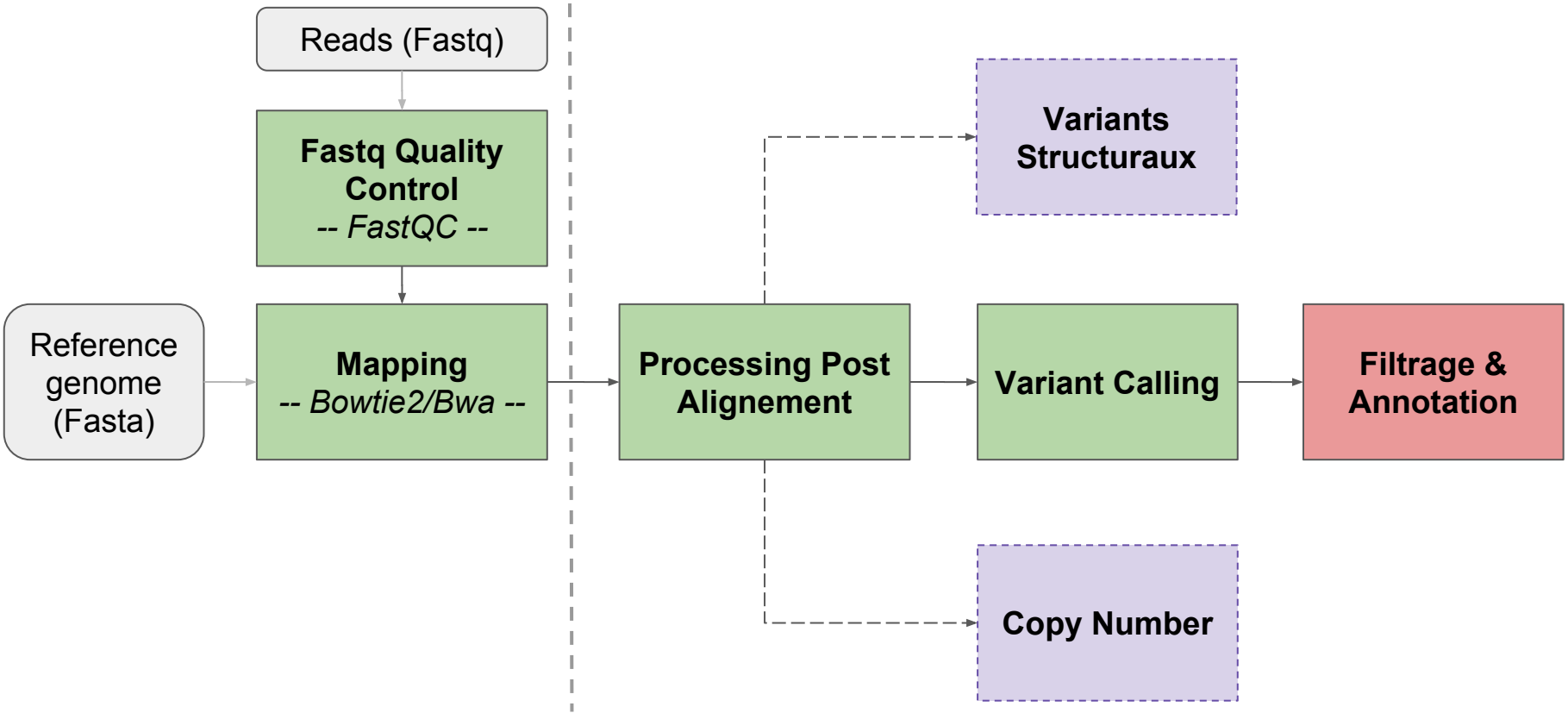
Filtrage & Annotation

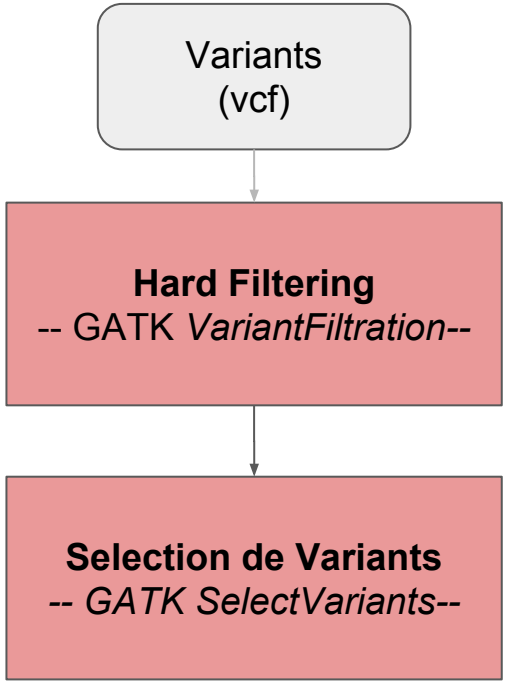
Olivier Rué - INRA

École de bioinformatique AVIESAN-IFB 2018

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



Plan

- Filtrage simple avec vcftools ? + Gatk HardFilter + combine varscan - HaplotypeCaller + snpSift
- Annotation avec snpEFF - createdb
- filtre sur the variant annoté
- diagramme de Venn → interlude R

- Generation d'une screenshot automatique sur IGV du variant d'intérêt

- Bonus : clustering variant

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
$ cd ..
$ mkdir filter_and_annot
$ cd filter_and_annot


# Extraction des SNVs dans un fichier séparé pour GATK
$ gatk SelectVariants -R ../genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa \
-V ../GATK/gvcf/pool_GATK.vcf \
-O pool_GATK.SNP.vcf \
--select-type SNP

# Extraction des SNVs dans un fichier séparé pour Varscan
$ gatk SelectVariants -R ../genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa \
-V ../Varscan/pool_Varscan.vcf \
-O pool_Varscan.SNP.vcf \
--select-type SNP
```

SelectVariants et Hard filtering

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

SelectVariants et Hard filtering

```
# Filtrage des SNVs selon les filtres recommandés par GATK
$ gatk VariantFiltration -R ../genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa \
-V pool_GATK.SNP.vcf \
-O pool_GATK.SNP.prefilt.vcf \
--filter-expression "QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 ||
ReadPosRankSum < -8.0" \
--filter-name “hard_filtering_snv”


# Sélection des variants passant ce filtre
$ gatk SelectVariants -R ../genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa \
-V pool_GATK.SNP.prefilt.vcf  \
-O pool_GATK.SNP.filtered.vcf  \
--exclude-filtered
```

Intersection des résultats des variant callers

```
# Intersection des variants obtenus avec Varscan et avec GATK
$ vcftools # v0.1.16


# 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

$ vcf-isec -f -n +2 pool_GATK.SNP.filtered.vcf.gz pool_Varscan.SNP.vcf.gz >
GATK_varscan_inter.vcf
```

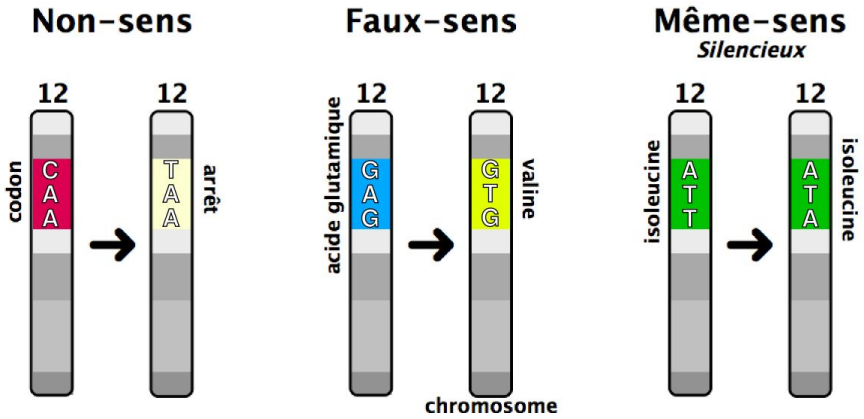
Annotation des variants

- Ajout d'informations biologiques pertinentes aux variants :

  → Est-ce que mes variants sont connus ?

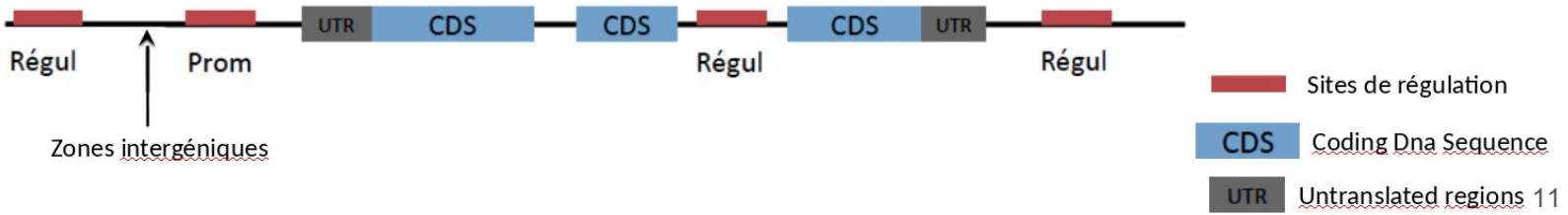
  → Où se positionnent mes variants ?

  → 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
$ snpEff -version # affiche la version (v4.3t)

$ echo BosTaurus.genome >> snpeff.config # <genome_name>.genome
$ mkdir -p BosTaurus
$ cp ../genome/Bos_taurus.UMD3.1.dna.toplevel.6.fa BosTaurus/sequences.fa
$ cp ../genome/Bos_taurus.UMD3.1.93_6.gtf BosTaurus/genes.gtf
$ echo -e "BosTaurus\nSnpEff4.1" > BosTaurus.db

$ snpEff build -c snpeff.config -gtf22 -v BosTaurus -dataDir .
```

```
# Annotation avec notre base de données
$ snpEff eff -c snpeff.config -dataDir . BosTaurus -s snpeff_resultat.html
GATK_varscan_inter.vcf > GATK_varscan_inter.annot.vcf
```

SnpSift

```
$ SnpSift filter -h  # affiche l'aide (v 4.3t)

# Garder les variants codant qui ne sont pas des synonymes :
$ cat GATK_varscan_inter.annot.vcf | SnpSift filter "(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)
$ cat GATK_varscan_inter.annot.coding.nosyn.vcf | SnpSift filter "ANN[*].EFFECT = 'missense_variant' & isHet( GEN[2] ) & isVariant( GEN[2] ) & isRef( GEN[0] ) & isRef( GEN[1] )" > GATK_varscan_inter.annot.coding.nosyn.filtered.vcf
```

SnpSift

```
# Sélectionner notre variant d’intérêt ayant un impact
$ cat
pooled_calling_GATK.SNP.filtered.intersect.annot.vcf.NO_SYN.NO_INT.coding.vcf |
java -jar SnpSift filter "ANN[*].EFFECT = 'missense_variant'" >
pooled_calling_GATK.SNP.filtered.intersect.annot.vcf.NO_SYN.NO_INT.coding.missen
se.vcf

cat variants.vcf | java -jar SnpSift.jar filter "isHom( GEN[0] ) & isVariant(
GEN[0] ) & isRef( GEN[1] )" > 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, il ne comporte pas cette mutation : il est homozygote référence (GT: 0/0).

Zinder *et al.*, 2005

Quitter sans fermer la session

```
# Pour quitter sans terminer sa session
$ screen -d

# Vérification que l'on a toujours un job qui tourne
$ squeue -u <userName>

# Cela doit vous afficher
#  JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
#    772      fast mbernard mbernard  R    1:00:52      1 cpu-node-1

# Vérification plus détaillée
$ scontrol show job 772 | grep JobName

# JobId=772 JobName=mbernard_TPVariant
```