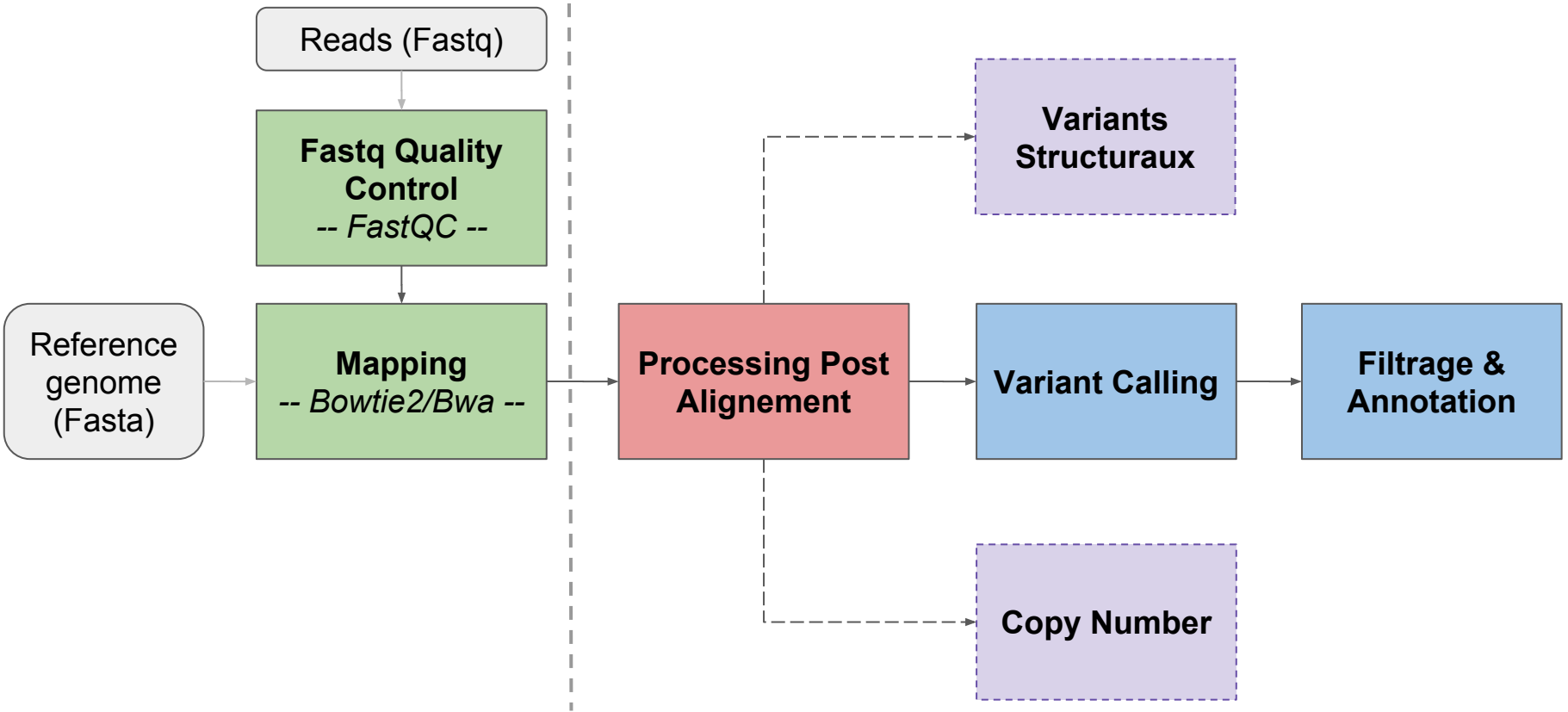
Processing Post-Alignement

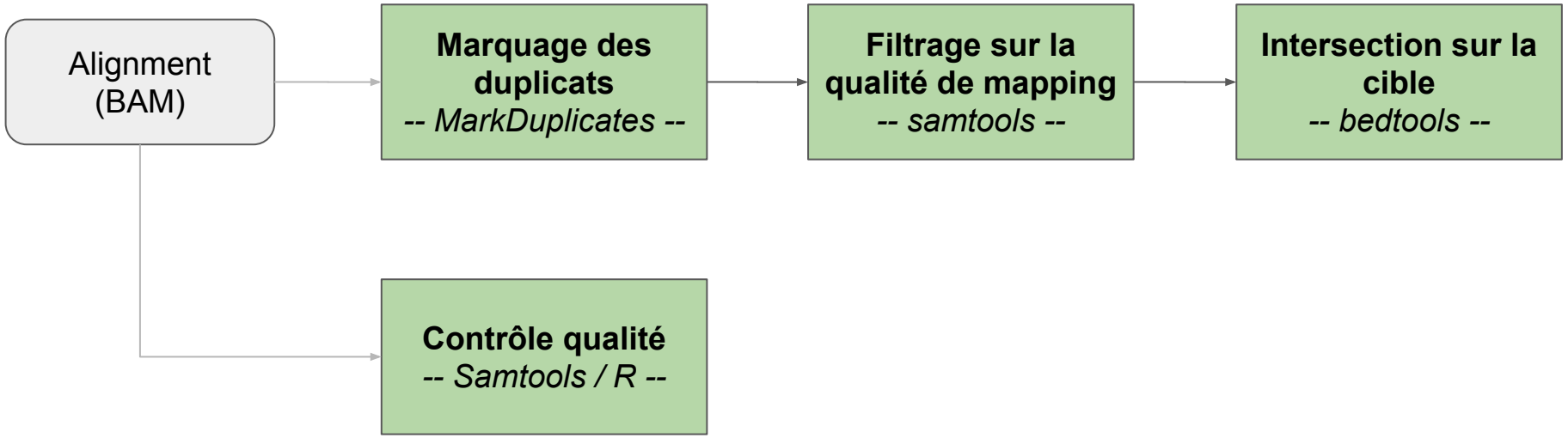
Olivier Rué - INRA

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 - Processing Post Alignement



Copie du jeu de données #1

```
#Listing des fichiers FASTQ, Genome et BAM
$ ls -lh /shared/home/mbernard/atelier_variant/tp_variant/fastq
$ ls -lh /shared/home/mbernard/atelier_variant/tp_variant/genome
$ ls -lh /shared/home/mbernard/atelier_variant/tp_variant/alignment_bwa
```

```
#Copie des fichiers dans notre home
$ cp -r /shared/home/mbernard/atelier_variant/tp_variant/ .

#Se déplacer dans le dossier alignment_bwa
$ cd  ~/tp_variant/alignment_bwa
```

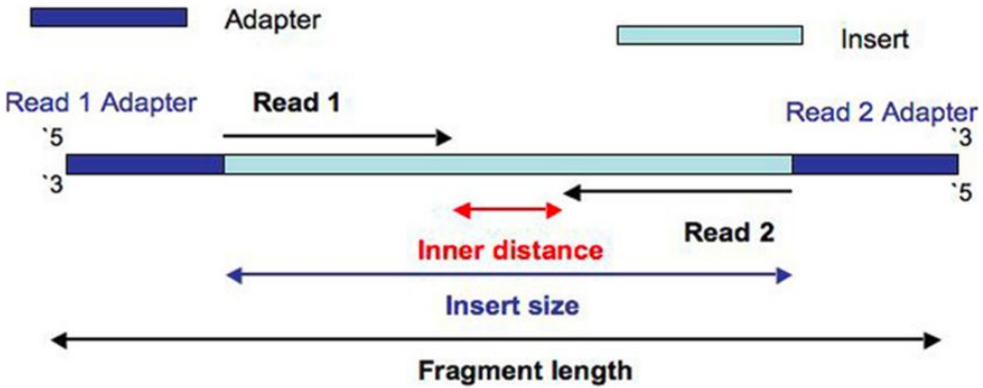
Contrôle qualité des données alignées

- Quelles informations regarder une fois le mapping effectué ?

  → Pourcentage total de reads alignés

  → Pourcentage de reads pairés “**proprement**”

- Quels outils ?
  - Samtools flagstat
  - Qualimap [optionnel]



Contrôle qualité des données alignées

#Lancement de samtools

```
$ samtools --version # affiche la version (v.1.9)
$ samtools flagstat # affiche l'aide

$ samtools flagstat SRR1262731_extract.sort.bam >  SRR1262731.flagstat.txt

$ cat SRR1262731.flagstat.txt # visualisation du résultat
```

#Lancement de Qualimap

```
$ qualimap --version # affiche la version (v2.2.2)
$ qualimap bamqc # affiche l'aide

$ qualimap bamqc -nt 4 -outdir SRR1262731_extract_qualimap_report \
--java-mem-size=4G -bam SRR1262731_extract.sort.bam
```

ReadGroups (RG)

- Associe des **informations sur la provenance des reads**
  - → Identité : run/échantillon
  - → Séquençage, librairie…
- Nécessaire à la recherche de variants

```
Mom's data:
@RG     ID:FLOWCELL1.LANE5      PL:ILLUMINA     LB:LIB-MOM-1 SM:MOM
@RG     ID:FLOWCELL1.LANE6      PL:ILLUMINA     LB:LIB-MOM-1 SM:MOM
@RG     ID:FLOWCELL1.LANE7      PL:ILLUMINA     LB:LIB-MOM-2 SM:MOM
@RG     ID:FLOWCELL1.LANE8      PL:ILLUMINA     LB:LIB-MOM-2 SM:MOM

Kid's data:
@RG     ID:FLOWCELL2.LANE1      PL:ILLUMINA     LB:LIB-KID-1 SM:KID
@RG     ID:FLOWCELL2.LANE2      PL:ILLUMINA     LB:LIB-KID-1 SM:KID
@RG     ID:FLOWCELL2.LANE3      PL:ILLUMINA     LB:LIB-KID-2 SM:KID
@RG     ID:FLOWCELL2.LANE4      PL:ILLUMINA     LB:LIB-KID-2 SM:KID
```

- Comment vérifier la présence de ReadGroups dans un fichier BAM?

```
$ samtools view # affiche l'aide

$ samtools view -H SRR1262731_extract.sort.bam  | grep “^@RG”
```

Comment ajouter des ReadGroups ?

- Au niveau des **paramètres du mapper** :

Bwa : “ -R @RG\tID:ID\tSM:SAMPLE_NAME\tPL:Illumina\tPU:PU\tLB:LB”

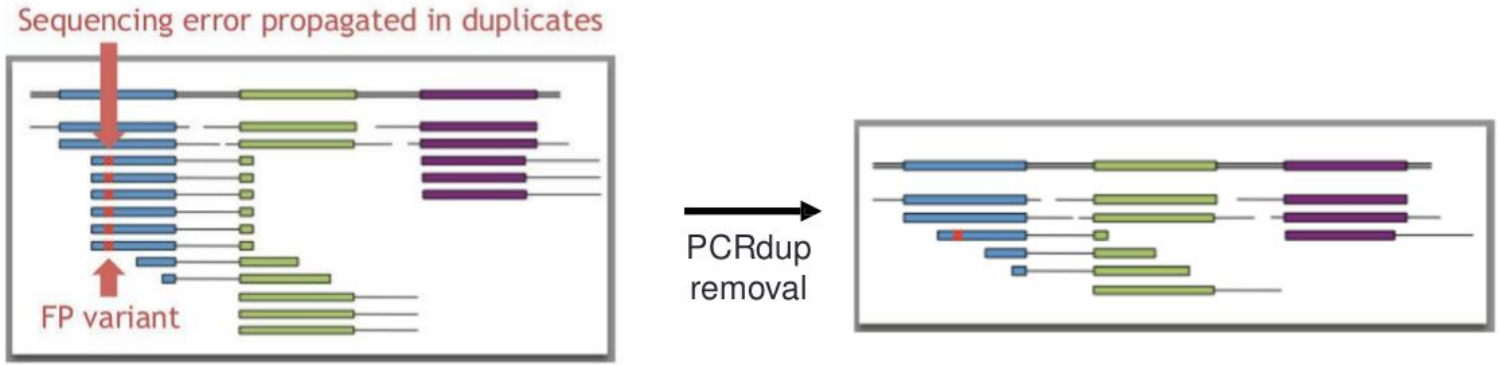
Bowtie2 : “--rg-id ID --rg SM:SAMPLE_NAME --rg PL:Illumina --rg PU:PU --rg LB:LB”

- Avec l’outil **AddOrReplaceReadGroups** de la suite PicardTools

```
$ picard AddOrReplaceReadGroups --version # affiche la version (v2.18.9)
$ picard AddOrReplaceReadGroups --help # affiche l'aide

$ picard AddOrReplaceReadGroups I=SRR1262731_extract.sort.bam \
O=SRR1262731_extract.sort.rg.bam RGID=1 RGPL=Illumina RGPU=PU \
RGSM=SRR1262731 RGLB=LB
```

Marquage des duplicats de PCR

- Identifier les **reads provenant d'une même molécule** issus de :
  → PCR duplicates : amplification PCR durant la préparation de la librairie
  → Optical duplicates : cluster illumina identifié comme deux clusters



Marquage des duplicats de PCR

→ Garder les duplicats de PCR : probabilité importante de confondre les duplicats avec des fragments biologiques issus du même locus

→ Marquer les duplicats mais les conserver dans le fichier BAM : certains outils les supprimeront par défaut (samtools, GATK…)

→ Supprimer les duplicats du fichier BAM

```
$ picard MarkDuplicates --help # affiche l'aide
$ picard -Xmx8G MarkDuplicates I=SRR1262731_extract.sort.rg.bam \
O=SRR1262731_extract.sort.rg.md.bam M=SRR1262731_extract_metrics_md.txt \
VALIDATION_STRINGENCY=SILENT
$ samtools flagstat SRR1262731_extract.sort.rg.md.bam \
> SRR1262731_extract.md.flagstat.txt
$ cat SRR1262731_extract.md.flagstat.txt # nombre de duplicats

$ grep -A1 "LIBRARY" SRR1262731_extract_metrics_md.txt | less -S # % de pcrDup
```

Filtres sur les alignements

Restreindre le fichier BAM en fonction de métriques d'alignements :

- qualité de mapping (MAPQ) suffisante
- retrait des reads non mappés

```
# Suppression des reads non mappés et filtre sur les reads avec MAPQ < 30
$ samtools view -bh -F 4 -q 30 SRR1262731_extract.sort.rg.md.bam \
> SRR1262731_extract.sort.rg.md.filt.bam

$ samtools flagstat SRR1262731_extract.sort.rg.md.filt.bam \
> SRR1262731_extract.filt.flagstat.txt

$ cat SRR1262731_extract.filt.flagstat.txt
```

Filtres sur les alignements

Restreindre le fichier BAM en fonction de métriques d'alignements :

- alignements intersectant les régions d'intérêt
- en fonction du nombre de mismatchs, de la taille d'insert, de paires mappées sur des chromosomes différents…

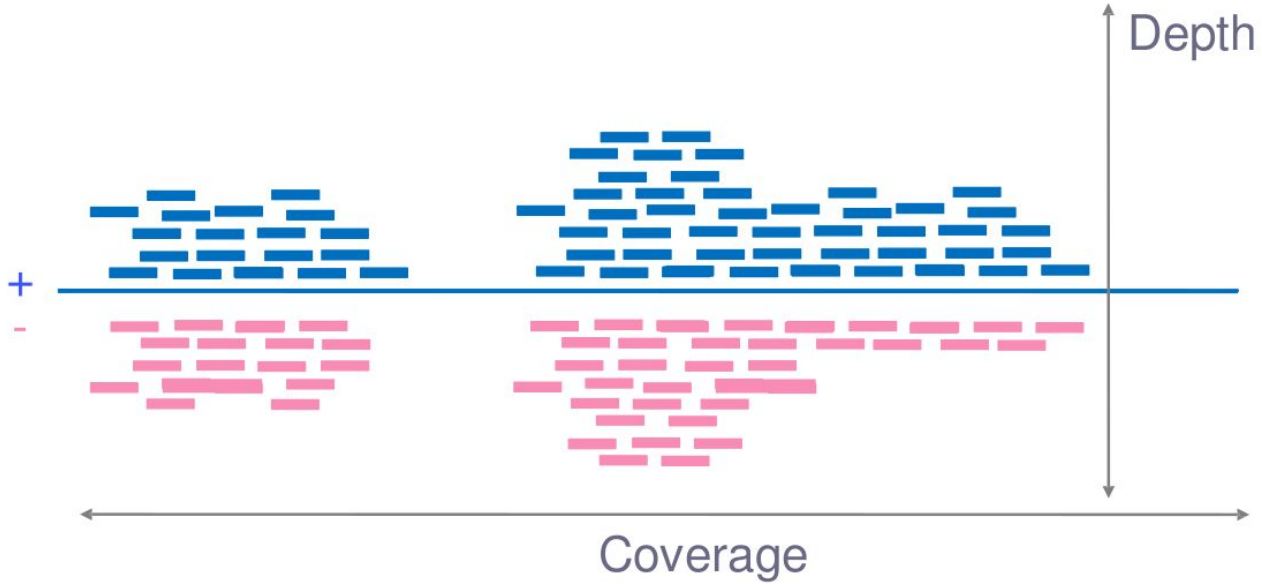
```
# Conservation des alignements dans les régions ciblées
$ bedtools --version # affiche la version (v2.27.1)
$ bedtools intersect  --help # affiche l'aide

$ bedtools intersect -a SRR1262731_extract.sort.rg.md.filt.bam \
-b ../additionnal_data/QTL_BT6.bed \
> SRR1262731_extract.sort.rg.md.filt.onTarget.bam

$ samtools index  SRR1262731_extract.sort.rg.md.filt.onTarget.bam
```

Analyse de la couverture

Contrôle qualité de l'enrichissement de ma capture :

→ Est-ce que ma région est **couverte par suffisamment de reads** ?

→ Cette couverture est-elle homogène sur toute la région ?



Analyse de la couverture

Contrôle qualité de l'enrichissement de ma capture :

→ Est-ce que ma région est **couverte par suffisamment de reads** ?

→ Cette couverture est-elle homogène sur toute la région ?

```
# Calcul de la couverture avec samtools
$ samtools depth --help # affiche l'aide

$ samtools depth -b ../additionnal_data/QTL_BT6.bed \
SRR1262731_extract.sort.rg.md.filt.onTarget.bam \
> SRR1262731_extract.onTarget.depth.txt

$ head SRR1262731_extract.onTarget.depth.txt
```