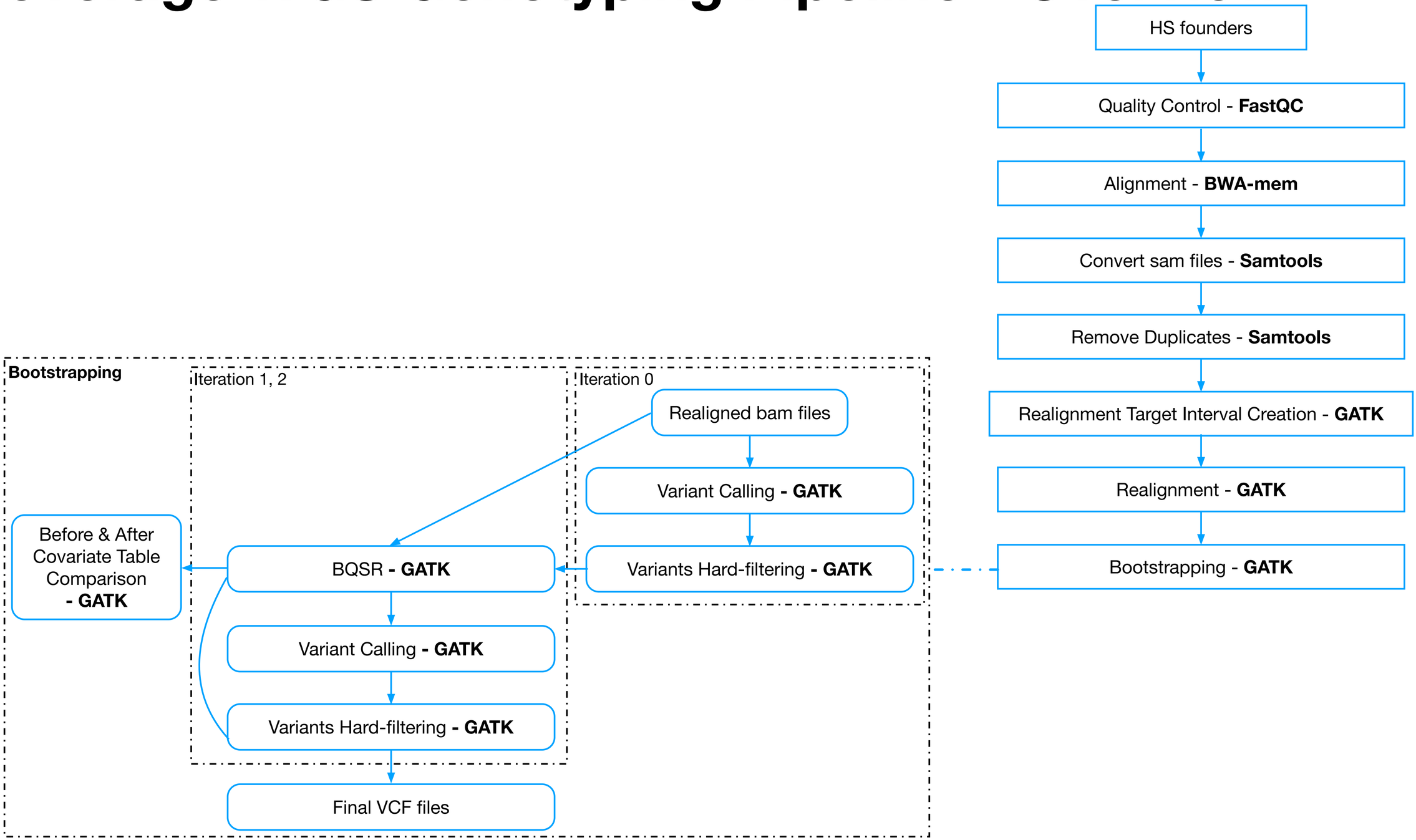
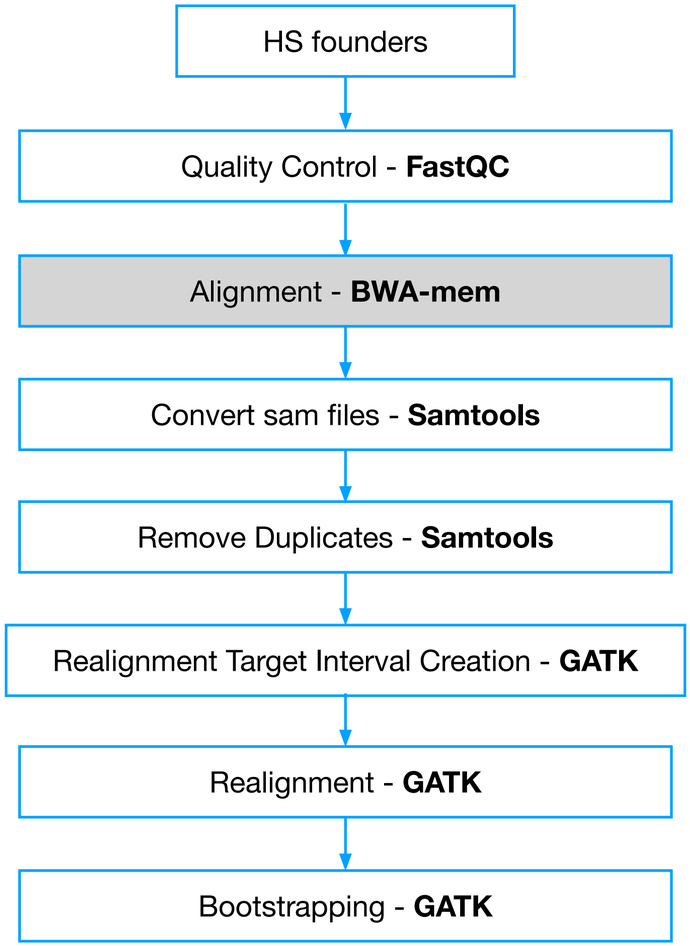
High Coverage WGS Genotyping Pipeline

Denghui Chen, Palmer Lab
May. 26 2022

High Coverage WGS Genotyping Pipeline - Overview



High Coverage WGS Genotyping Pipeline - Alignment



Alignment

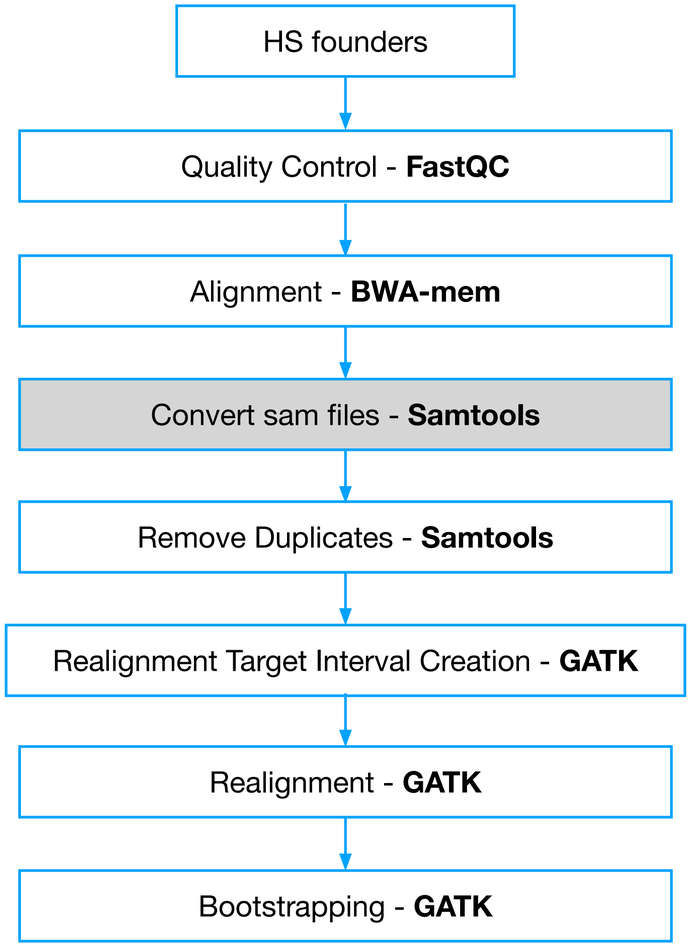
Command:

bwa mem -Y -K 100000000

- -Y: use soft clipping for supplementary alignments (structural variant calling)
- -K: process 100,000,000 input bases in each batch regardless of nThreads (for reproducibility)
- -R: read group header line such as
  “@RG\tID:${flowcell_name}.${lane}\tLB:${library_id}\tPL:${platform}\tSM:${sample}”

High Coverage WGS Genotyping Pipeline - Convert SAM Files

Convert SAM to BAM file (samtools)
samtools view -h -b -o ${sam_prefix}.bam ${sam_prefix}.sam



High Coverage WGS Genotyping Pipeline - Mark Duplicates

Collate BAM (samtools)
samtools collate -o ${bam_prefix}_collated.bam ${bam_file}

- shuffles and groups reads together by their names

Fixmate BAM (samtools)
samtools fixmate -m ${bam_prefix}_collated.bam ${bam_prefix}_fixmate.bam

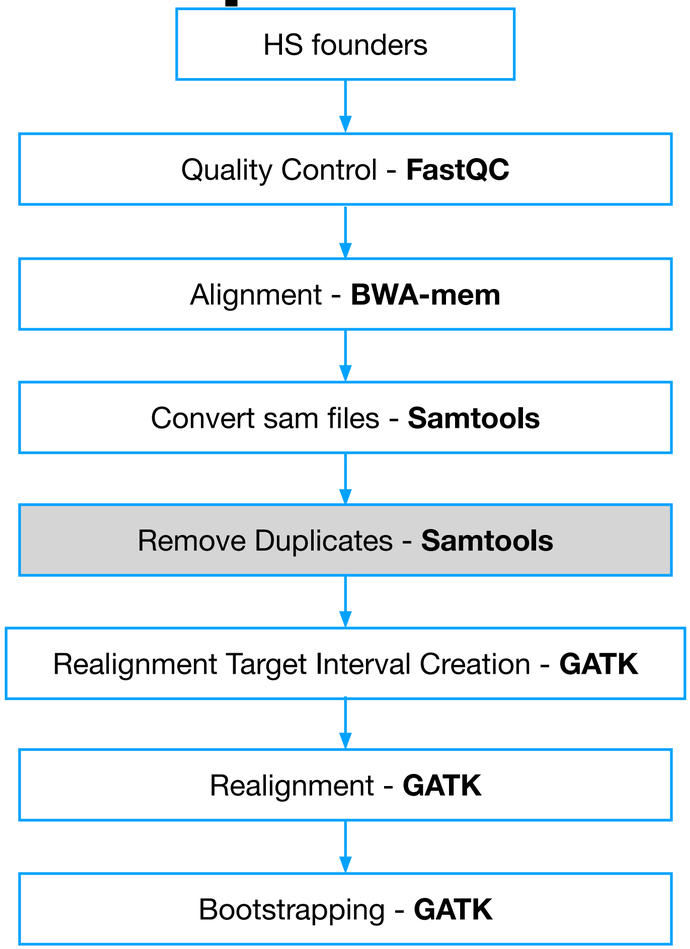
- -m: add ms (mate score) tags. These are used by markdup to select the best reads to keep.
- fills in mate coordinates and insert size fields

Sort BAM by Coordinates (samtools)
samtools sort -o ${bam_prefix}_sorted.bam ${bam_prefix}_fixmate.bam

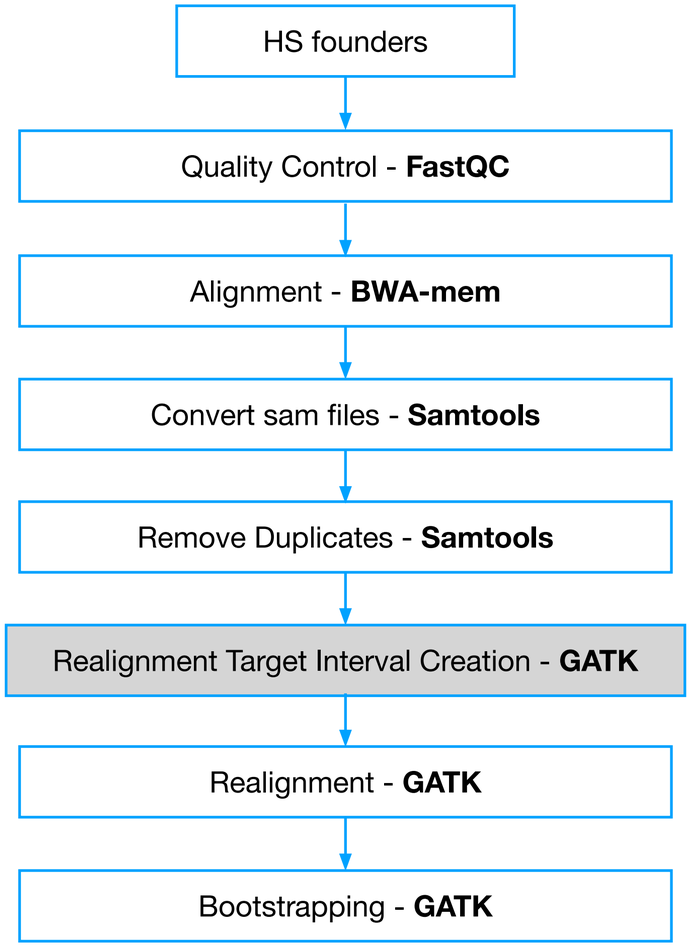
Mark Duplicates (samtools)
samtools markdup -o ${bam_prefix}_rmDup.bam $${bam_prefix}_sorted.bam

- -r: remove duplicate reads

Index BAM (samtools)
samtools index ${bam_prefix}_rmDup.bam ${bam_prefix}_rmDup.bai



High Coverage WGS Genotyping Pipeline - Realignment Target Interval Creation



Realignment Target Interval Creation
Software version

- GATK 3.8.1.0 is used since Indel Realignment is no longer part of GATK4. See more details here.

Commands:

```
java -jar GenomeAnalysisTK.jar RealignerTargetCreator \
    -R ${reference} \
    -I ${bam_file} \
    -o indel.intervals
```

Notes

If there are no known indel or SNP used as a reference in this step, but it can run without it.

High Coverage WGS Genotyping Pipeline - Realignment

Realignment
Filters used

- Default filters on GATK: BadCigarFilter, BadMateFilter, DuplicateReadFilter, FailsVendorQualityCheckFilter, MalformedReadFilter, MappingQualityUnavailableFilter, MappingQualityZeroFilter, NotPrimaryAlignmentFilter, Platform454Filter, UnmappedReadFilter

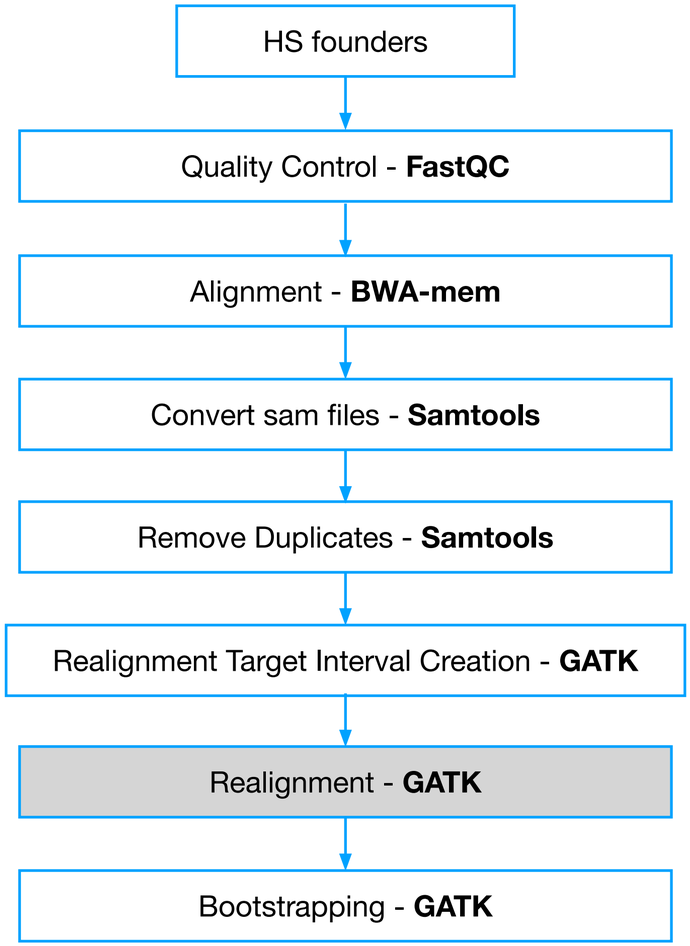
Software version

- GATK 3.8.1.0 is used since Indel Realignment is no longer part of GATK4. See more details here.

Commands:

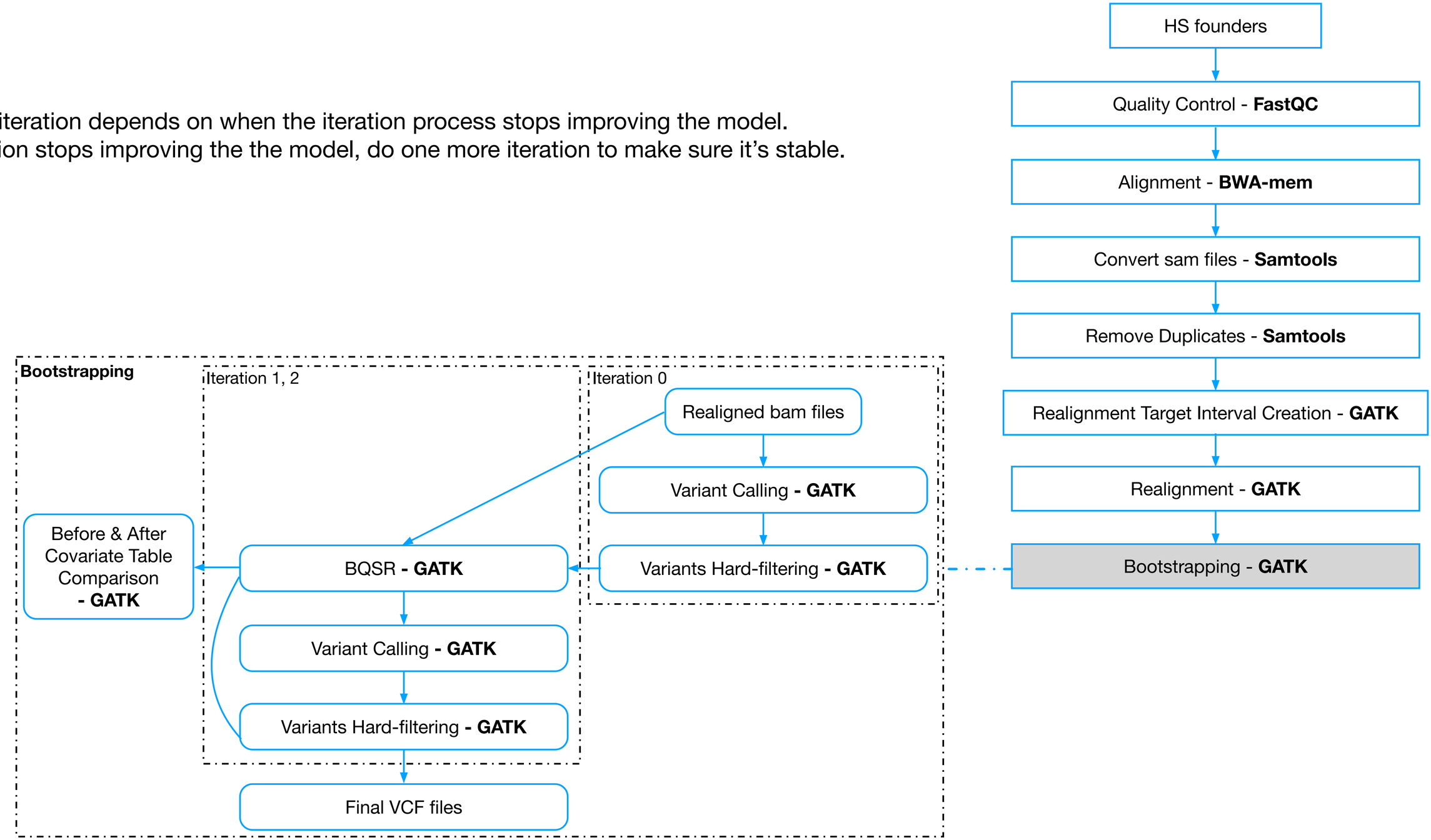
```
java -jar GenomeAnalysisTK.jar IndelRealigner \
    -R ${reference} \
    -targetIntervals indel.intervals \
    --filter_bases_not_stored \
    -I ${bam_file} \
    -o ${bam_prefix}_indelrealigned.bam
```

Notes
If there are no known indel or SNP used as a reference in this step, but it can run without it.



High Coverage WGS Genotyping Pipeline - Bootstrapping Overview

Bootstrapping
The number of iteration depends on when the iteration process stops improving the model.
When the iteration stops improving the the model, do one more iteration to make sure it's stable.



High Coverage WGS Genotyping Pipeline - Bootstrapping Variant Calling

Variant Calling
Filters used

- GATK HyplotypeCaller default filters, see here.
  - NotSecondaryAlignmentReadFilter
  - GoodCigarReadFilter
  - NonZeroReferenceLengthAlignmentReadFilter
  - PassesVendorQualityCheckReadFilter
  - MappedReadFilter
  - MappingQualityAvailableReadFilter
  - NotDuplicateReadFilter
  - MappingQualityReadFilter
  - WellformedReadFilter

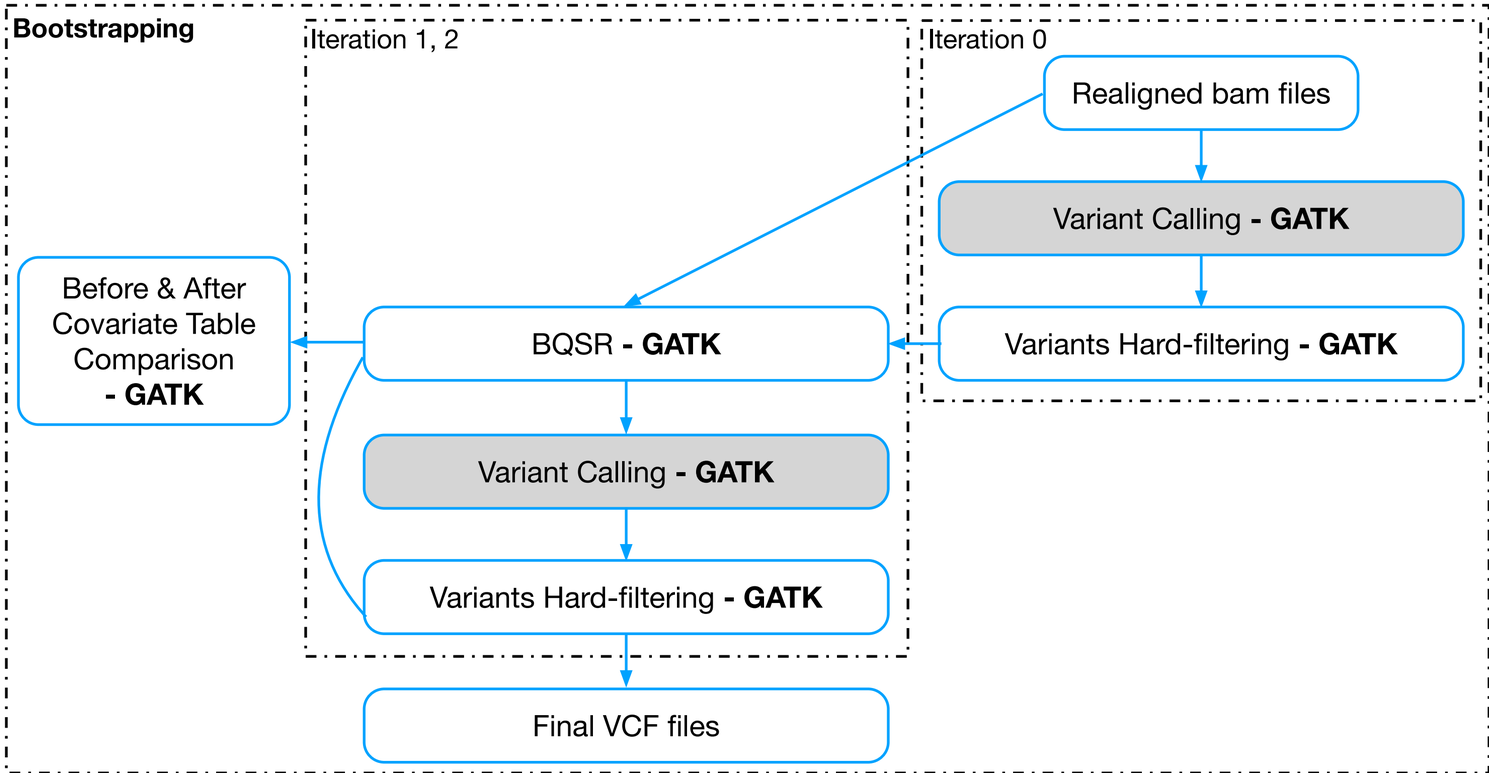
Commands:

Call SNPs and indels via local re-assembly of haplotypes

```
java -jar GenomeAnalysisTK.jar HaplotypeCaller  \
        -R ${reference} \
        -I ${bam_file} \
        -L ${interval} \
        -O ${out_path}/${interval}_it0.vcf.gz

java -jar GenomeAnalysisTK.jar GenomicsDBImport  \
        --genomicsdb-workspace-path ${database_dir}/${chr} \
        -L ${chr} \
        --sample-name-map ${it0_dir}/${chr}_sample_map
```

```
java -jar GenomeAnalysisTK.jar GenotypeGVCFs  \
        -R ${reference} \
        -V gendb://${database_dir}/${chr} \
        -L ${chr} \
        --genomicsdb-shared-posixfs-optimizations true \
        -O ${vcf_dir}/${chr}.vcf.gz
```



High Coverage WGS Genotyping Pipeline - Bootstrapping Variants Hard-filtering (SNPs)

Variants Hard-filtering
Filters used (GATK guidance)

Commands:

Subset to SNPs-only

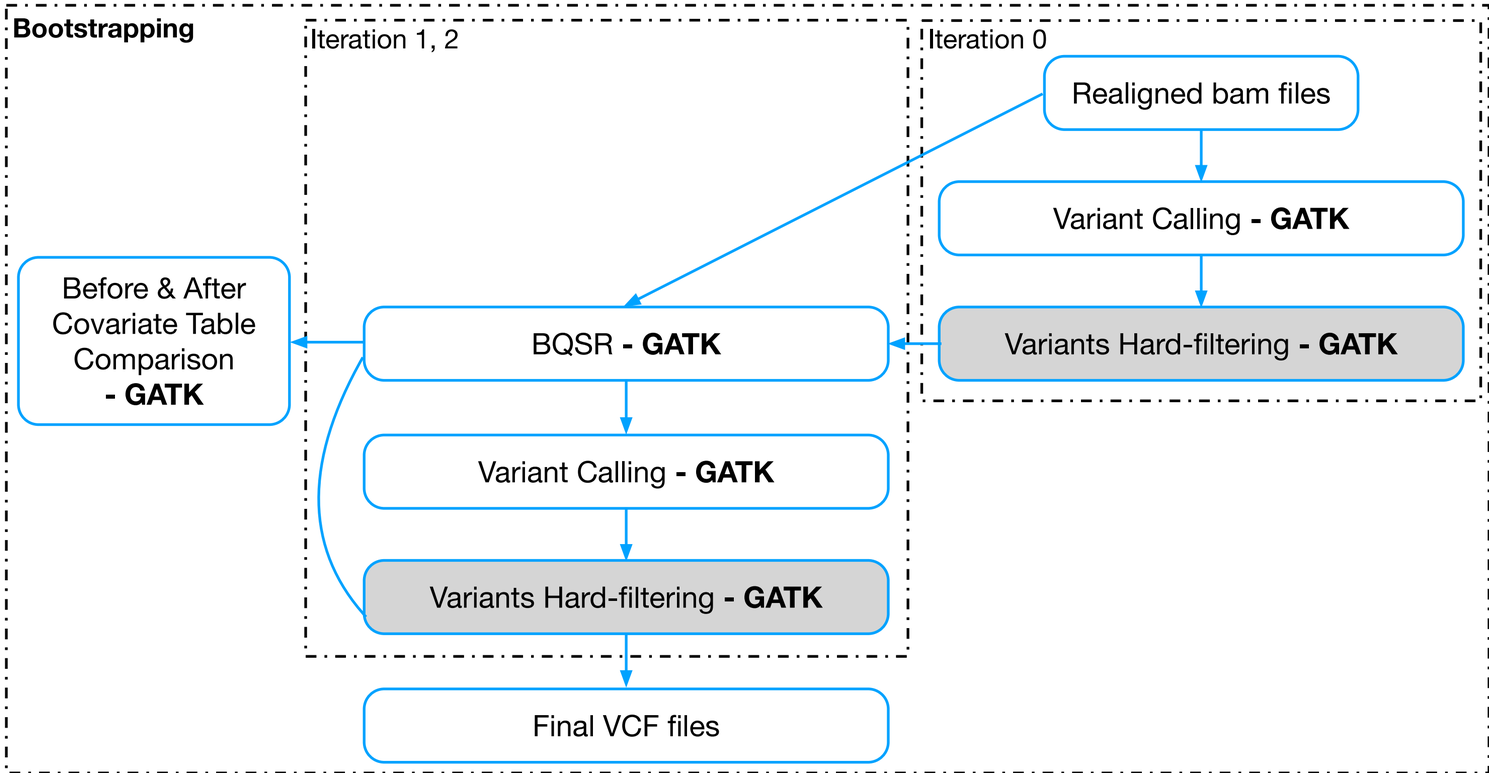
```
java -jar GenomeAnalysisTK.jar SelectVariants \
	-V ${out_path}/${interval}_it0.vcf.gz \
	-select-type SNP \
	-O ${out_path}/${interval}_SNPs_it0.vcf.gz
```

Variants filtering

```
java -jar GenomeAnalysisTK.jar VariantFiltration \
	-V ${out_path}/${interval}_SNPs_it0.vcf.gz \
	-filter "QD < 5.0" --filter-name "QD2" \
	-filter "QUAL < 30.0" --filter-name "QUAL30" \
	-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" \
	-O ${out_path}/${interval}_SNPs_temp_it0.vcf.gz
```

Exclude filtered variants

```
java -jar GenomeAnalysisTK.jar SelectVariants \
	-V ${out_path}/${interval}_SNPs_temp_it0.vcf.gz \
	--exclude-filtered true \
	-O ${out_path}/${interval}_SNPs_filtered_it0.vcf.gz
```



High Coverage WGS Genotyping Pipeline - Bootstrapping Variants Hard-filtering (Indels)

Variants Hard-filtering
Filters used (GATK guidance)

Commands:

Subset to Indels-only

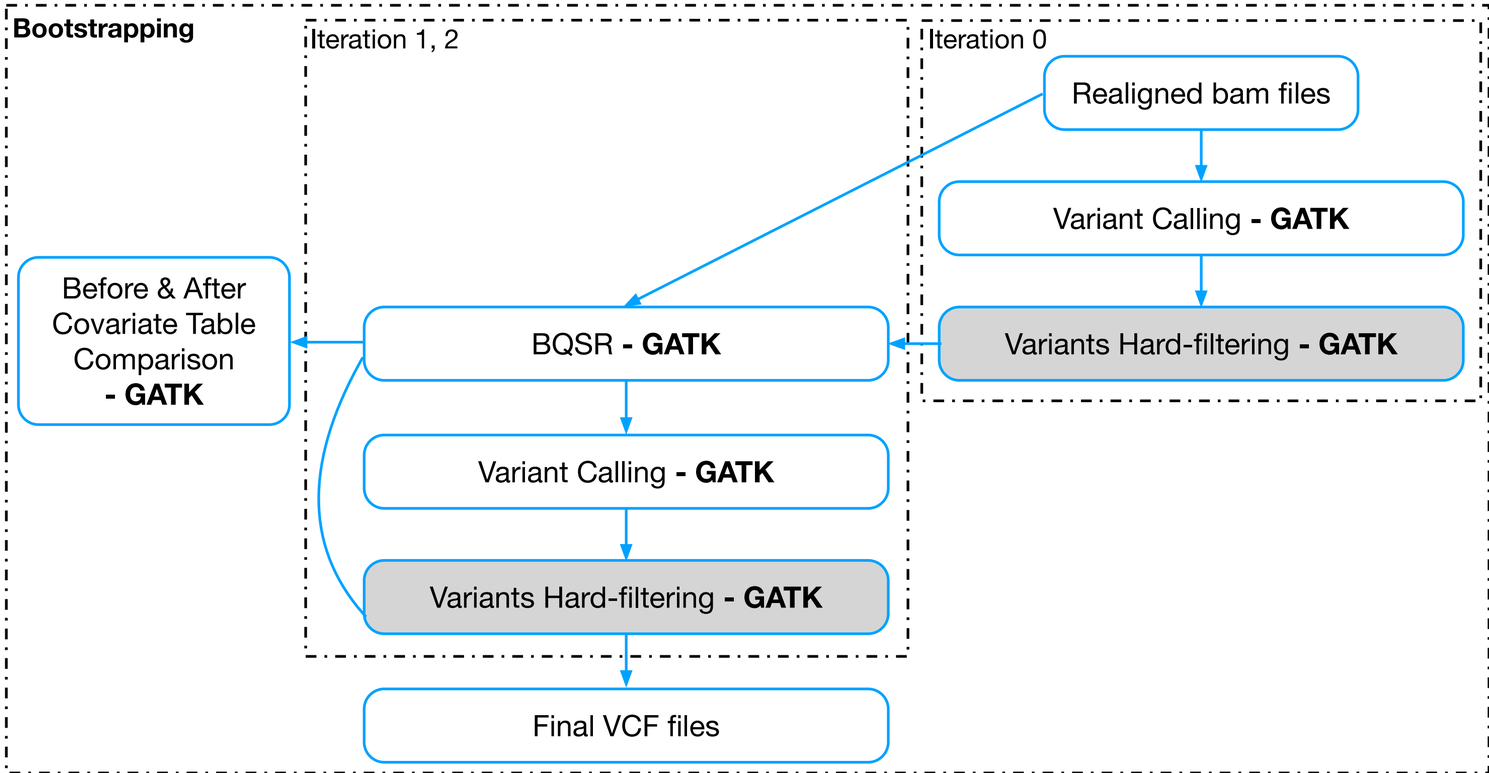
```
java -jar GenomeAnalysisTK.jar SelectVariants \
    -V ${out_path}/${interval}_it0.vcf.gz \
    -select-type INDEL \
    -O ${out_path}/${interval}_indels_it0.vcf.gz
```

Variants filtering

```
java -jar GenomeAnalysisTK.jar VariantFiltration \
    -V ${out_path}/${interval}_indels_it0.vcf.gz \
    -filter "QD < 5.0" --filter-name "QD2" \
    -filter "QUAL < 30.0" --filter-name "QUAL30" \
    -filter "FS > 200.0" --filter-name "FS200" \
    -filter "ReadPosRankSum < -20.0" --filter-name "ReadPosRankSum-20" \
    -O ${out_path}/${interval}_indels_temp_it0.vcf.gz
```

Exclude filtered variants

```
java -jar GenomeAnalysisTK.jar SelectVariants \
    -V ${out_path}/${interval}_indels_temp_it0.vcf.gz \
    --exclude-filtered true \
    -O ${out_path}/${interval}_indels_filtered_it0.vcf.gz
```



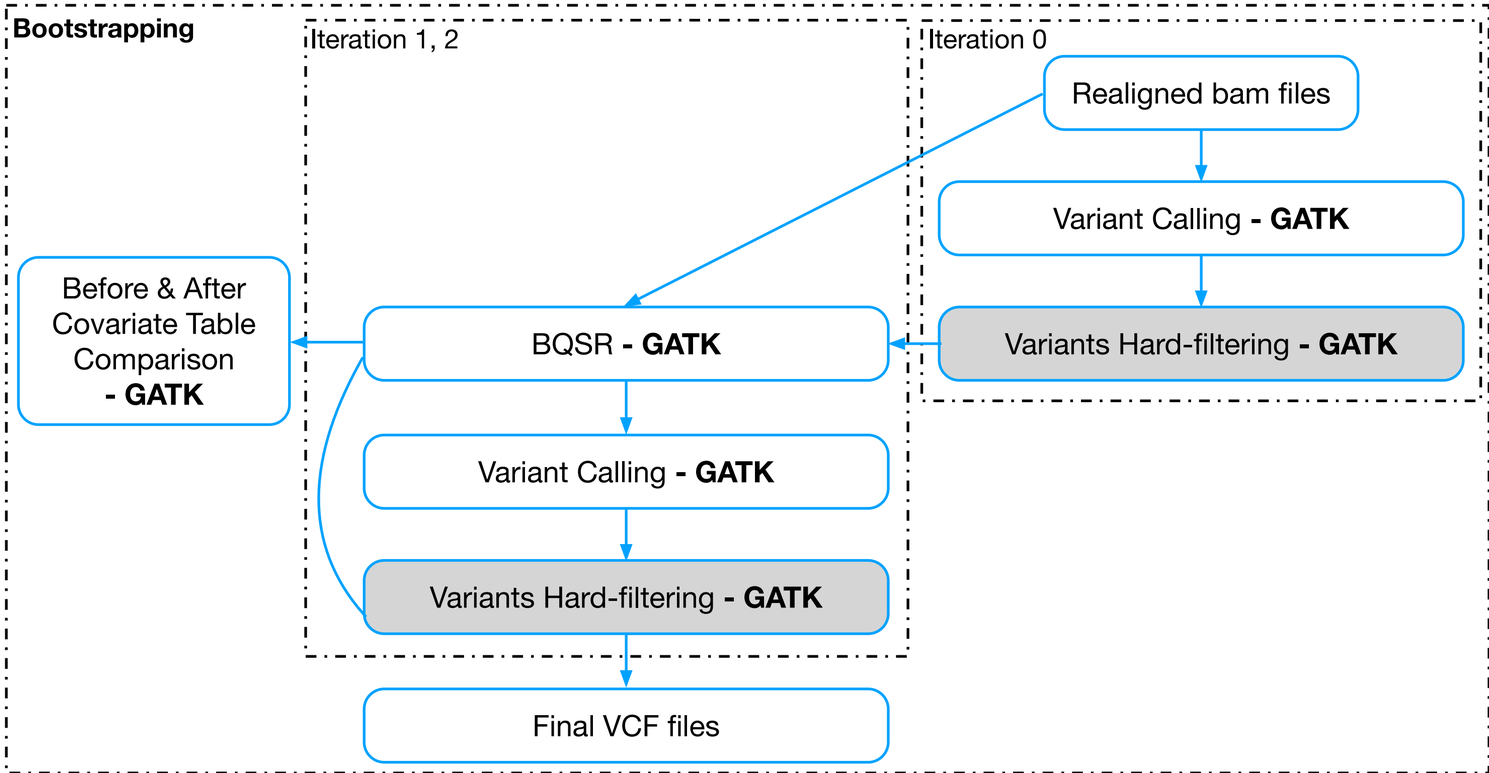
High Coverage WGS Genotyping Pipeline - Bootstrapping Variants Hard-filtering (Concatenate)

Variants Hard-filtering (Concatenate, picard)

Commands:

Concatenate SNPs with Indels

```
java -jar picard.jar MergeVcfs  \
        -INPUT it0_filtered.list \
        -OUTPUT ${prefix}.vcf.gz
```



High Coverage WGS Genotyping Pipeline - Bootstrapping BQSR

Base Quality Score Recalibration

Commands:

Commands:

```
java -jar GenomeAnalysisTK.jar BaseRecalibrator -R ${reference} \
    --known-sites ${prefix}.vcf.gz \
   -I ${bam_prefix}_indelrealigned.bam \
    -o recal1.table

java -jar GenomeAnalysisTK.jar ApplyBQSR -R ${reference} \
    --bqsr-recal-file recal1.table \
    -I ${bam_prefix}_indelrealigned.bam \
    -o ${bam_prefix}_recal.bam

java -jar GenomeAnalysisTK.jar BaseRecalibrator -R ${reference} \
    --known-sites ${know_sites}.vcf.gz \
    -bqsr recal1.table \
    -I ${bam_prefix}_indelrealigned.bam \
    -o recal2.table

java -jar GenomeAnalysisTK.jar AnalyzeCovariates \
    -before recal1.table \
    -after recal2.table \
    -plots AnalyzeCovariates.pdf
```