WHOLE EXOME PREPROCESSING PIPELINE

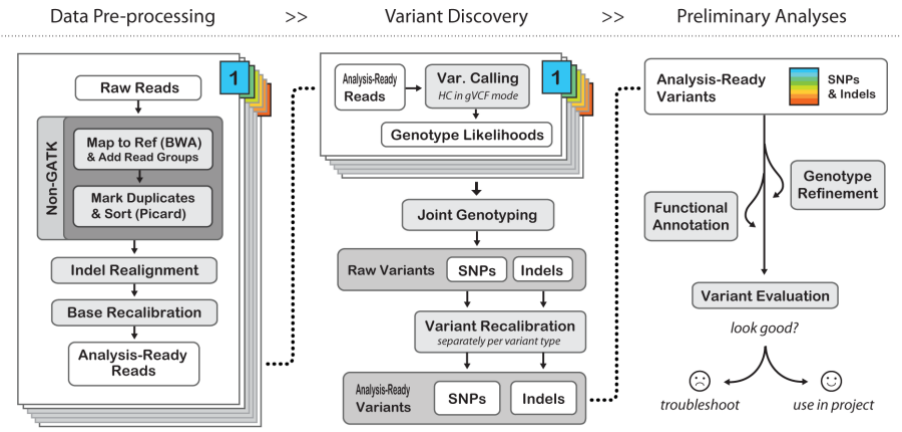

Figure from: GATK Best Practices for SNP and Indel discovery (https://gatk.broadinstitute.org/hc/en-us)

Computation times (TIME) estimated per sample (exome capture sequencing at 30X) on a HPC cluster (using 1 CPU per task).

---

0. INITIAL ASSESSMENT (FastQC/0.11.7, Trimmomatic/0.36-Java-1.8.0_92)

0.1. FASTQ QC TIME ~ 10min

* For exome regions, the GC content is about 49–51% (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4492405/)

0.2. TRIMMING TIME ~ 2h

-PE: Paired End
-phred33
-ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10
-SLIDINGWINDOW:4:25 #Minimum quality 25
-MINLEN:30 #Minimum length 30

Then, redo FastQC: Adapter Content and Overrepre. seq warning disappear, although seq. length distribution warning appears.

---

1. FROM FASTQ TO BAM (BWA/0.7.15; SAMtools/1.6; Picard/2.18.6; GATK/3.7)

First, **index** and **sort** Reference Genome TIME ~ 1h

1.1. MAPPING TIME = max 7-8h

→ Align reads and Add read groups → sam to bam → sort bam → index bam

> bwa mem -M -R @RG\tID:${flowcell}_${lane}\tLB:${library}\tPL:ILLUMINA\tSM:${sample}\tPU:1234 GRCh37.dna.primary_assembly_sorted.fa <(zcat ${sample}_1_paired.fastq.gz) <(zcat ${sample}_2_paired.fastq.gz) | samtools view -Sbh - | java -Xmx8g -jar picard.jar SortSam I=/dev/stdin O=sample}.bam SORT_ORDER=coordinate VALIDATION_STRINGENCY=STRICT; samtools index ${sample}.bam

1.2. REORDER BAM TIME ~ 20 min

> picard.jar ReorderSam I=${sample}.bam O=${sample}.reorder.bam REFERENCE=GRCh37.dna.primary_assembly_sorted.fa VALIDATION_STRINGENCY=STRICT CREATE_INDEX=true

1.3. MARK AND REMOVE DUPLICATES TIME ~ 30min

> picard.jar MarkDuplicates I=${sample}.reorder.bam O=${sample}.rmDup.bam METRICS_FILE=${sample}.rmDup.stats REMOVE_DUPLICATES=true VALIDATION_STRINGENCY=STRICT CREATE_INDEX=true"

1.4. COVERAGE AND MAPPING STATISTICS (for both raw and dedup bams)

Coverage (GATK DepthOfCoverage, DiagnoseTargets) TIME ~ 2h

Mapping statistics

a) % of mapped and unmapped reads → efficiency of mapping to the human
b) GC content (Picard CollectGcBiasMetrics)
c) InsertSize (Picard CollectInsertSizeMetrics)
d) Alignment Summary (Picard CollectAlignmentSummaryMetrics)
e) HS Metrics (Picard CollectHsMetrics): %selected_bases, %off_bait, %target_bases_1x, ... Breadth of coverage

1.5. INDEL REALIGNMENT

RealignerTargetCreator: Create a target list of intervals to be realigned TIME~ 1h

>GenomeAnalysisTK.jar -T RealignerTargetCreator -R GRCh37.dna.primary_assembly_sorted.fa -I $sample.rmDup.bam -known 1000G_phase1.indels.b37.vcf -o $sample.realigner.intervals

IndelRealigner: Perform realignment TIME ~ 40 min

> GenomeAnalysisTK.jar -T IndelRealigner -R GRCh37.dna.primary_assembly_sorted.fa -I $sample.rmDup.bam -known 1000G_phase1.indels.b37.vcf -targetIntervals $sample.realigner.intervals -o $sample.realigned.bam

1.6. BQSR

BaseRecalibrator: Analyze patterns of covariation TIME ~ 2-3h

> GenomeAnalysisTK.jar -T BaseRecalibrator -R GRCh37.dna.primary_assembly_sorted.fa -I sample.realigned.bam -knownSites 1000G_phase1.indels.b37.vcf -knownSites 1000G_omni2.5.b37.vcf -knownSites dbsnp_138.b37.vcf -knownSites hapmap_3.3.b37.vcf -knownSites Mills_and_1000G_gold_standard.indels.b37.vcf -o sample.RecalibrationFile.grp

PrintReads: Apply the recalibration TIME ~ 1-2h

> GenomeAnalysisTK.jar -T PrintReads -R GRCh37.dna.primary_assembly_sorted.fa -I $sample.realigned.bam -BQSR $sample.RecalibrationFile.grp -o $sample.recalibrated.bam

---

2. FROM BAM to VCF (GATK/3.7)

2.1. HAPLOTYPE CALLER TIME ~ 2h

> GenomeAnalysisTK.jar -T HaplotypeCaller -R GRCh37.dna.primary_assembly_sorted.fa - $sample.recalibrated.bam --dbsnp dbsnp_138.b37.vcf --genotyping_mode DISCOVERY -L $PATH_INTERVALS --emitRefConfidence GVCF –o ${sample}.raw_variants.g.vcf

2.2. GENOTYPE GVCFs TIME ~ 4.5h

> GenomeAnalysisTK.jar -T GenotypeGVCFs -R GRCh37.dna.primary_assembly_sorted.fa --dbsnp dbsnp_138.b37.vcf -stand_call_conf 20.0 -L $PATH_INTERVALS –o samples.WES.GC.vcf --variant raw_variants.list

2.3. VQSR for SNPS TIME ~ 20 min

VariantRecalibrator

> GenomeAnalysisTK.jar -T VariantRecalibrator -R GRCh37.dna.primary_assembly_sorted.fa -input all_samples.WES.vcf -recalFile all_samples.WES.joint_variants_raw_SNPs.recal -tranchesFile joint_variants_Raw_SNPs.tranches -nt 4 -mode SNP -resource:hapmap,known=false,training=true,truth=true,prior=15.0 hapmap_3.3.b37.vcf -resource:omni,known=false,training=true,truth=true,prior=12.0 1000G_omni2.5.b37.vcf -resource:1000G,known=false,training=true,truth=false,prior=10.0 1000G_phase1.snps.high_confidence.b37.vcf -resource:dbsnp,known=true,training=false,truth=false,prior=2.0 dbsnp_138.b37.vcf -an QD -an MQ -an MQRankSum -an ReadPosRankSum -an FS -an InbreedingCoeff -rscriptFile recal_snp.plots.R

ApplyRecalibration

> GenomeAnalysisTK.jar -T ApplyRecalibration -R GRCh37.dna.primary_assembly_sorted.fa –input all_samples.WES.vcf –recalFile all_samples.WES.joint_variants_raw_SNPs.recal -tranchesFile joint_variants_Raw_SNPs.tranches -mode SNP --ts_filter_level 99.5 -o all_samples.WES.snp_filtered.vcf

2.4. VQSR for Indels TIME ~ 20 min

- **VariantRecalibrator**

> GenomeAnalysisTK.jar -T VariantRecalibrator -R GRCh37.dna.primary_assembly_sorted.fa –input all_samples.WES.snp_filtered.vcf -recalFile all_samples.WES.joint_variants_raw_INDELS.recal -tranchesFile joint_variants_Raw_INDELS.tranches -nt 4 --maxGaussians 4 -mode INDEL -resource:mills,known=false,training=true,truth=true,prior=12.0 Mills_and_1000G_gold_standard.indels.b37.vcf -resource:dbsnp,known=true,training=false,truth=false,prior=2.0 dbsnp_138.b37.vcf -an QD -an SOR -an MQRankSum -an ReadPosRankSum -an FS -an InbreedingCoeff -rscriptFile recal_indels.plots.R

- **ApplyRecalibration**

> GenomeAnalysisTK.jar -T ApplyRecalibration -R GRCh37.dna.primary_assembly_sorted.fa -input all_samples.WES.snp_filtered.vcf -recalFile all_samples.WES.joint_variants_raw_INDELS.recal -tranchesFile joint_variants_Raw_INDELS.tranches -mode INDEL --ts_filter_level 99.0 -o all_samples.WES.snp_indel_filtered.vcf