

Germline WGS SNV and small-INDEL pipeline

The NYGC automated analysis pipeline for whole genome sequencing matches the CCDG and TOPMed [recommended](#) best practices. Sequencing reads are aligned to the human reference, hs38DH, using [bwa-mem2](#) (v2.2.1). We then fix mate pair information, perform duplicate marking and base quality score recalibration as outlined in the GATK [best-practices](#) to generate a coordinate sorted BAM file. We run several QC tools to gather and analyze a comprehensive range of alignment metrics, including coverage depth, mapping quality scores, GC bias, insert size distributions, and potential contamination. Single nucleotide variants and small indels are called using [GATK](#) HaplotypeCaller (v4.6.0.0), which generates a single-sample gVCF (for genomic VCF) followed by single sample genotype refinement using GATK's GenotypeGVCF which outputs a VCF. Next variant filtration is performed using [Variant Quality Score Recalibration](#) (VQSR at tranche 99.6% for both SNV and INDEL) which identifies annotation profiles of variants that are likely to be real, and assigns a score (VQSLOD) to each variant. SNVs and indels are extensively annotated using publicly available databases and annotation programs. Variant annotation is performed using [VEP](#) (v104) which includes: variant allele frequencies from gnomAD (v3.1.1), prediction of the effect of nucleotide changes on protein sequence, and variant disease associations. Other annotations include cross-species conservation scores from PhyloP, and PhastCons; functional prediction scores from Polyphen2, and SIFT; and clinically relevant variants from Clinvar. Variants and annotations are exported to tabular formats for the ease of downstream analysis. Additional filtering based on functional annotation is applied to report variants with predicted effects on protein coding.

Default Deliverables

Alignments in CRAM format:

- *.final.cram
- *.final.cram.crai

Single sample variant calling with VEP annotation:

- *.recalibrated.haplotypeCalls.vcf.gz
- *.recalibrated.haplotypeCalls.vcf.gz.tbi
- *.recalibrated.haplotypeCalls.annotated.vcf.gz
- *.recalibrated.haplotypeCalls.annotated.vcf.gz.tbi
- *.recalibrated.haplotypeCalls.annotated.txt.gz
- *.recalibrated.haplotypeCalls.annotated.txt.gz.tbi
- *.recalibrated.haplotypeCalls.annotated.high_mod.txt

Note: The variants in *.high_mod.txt are a subset of the original annotation that only includes high and moderate impact variants with frequency less than 1% in the gnomAD database.

Optional Deliverables

Genomic VCF (gVCF):

- *.haplotypeCalls.g.vcf.gz
- *.haplotypeCalls.g.vcf.gz.tbi