# WGS Germline Analysis

Whole Genome data are processed on NYGC automated pipeline. Paired-end 150bp reads were aligned to the GRCh38 human reference using the Burrows-Wheeler Aligner (BWA-MEM v0.7.15) and processed using the GATK best-practices workflow that includes marking of duplicate reads by the use of Picard tools (v2.4.1, http://picard.sourceforge.net), local realignment around indels, and base quality score recalibration (BQSR) via Genome Analysis Toolkit (GATK v3.5) (PMID: 20644199, 21478889).

## Single Nucleotide Variant Analysis

Variant discovery is a two-step process. HaplotypeCaller is run on each sample separately in GVCF mode (GATK v3.5). This produces an intermediate file format called gVCF (for genomic VCF). For projects with large number of samples, GVCFs are combined by batches into merged GVCFs. GVCFs are then run through a joint genotyping step (GATK v3.5) to produce a multi-sample VCF. Variant filtration is performed using Variant Quality Score Recalibration (VQSR) which identifies annotation profiles of variants that are likely to be real, and assigns a score (VQSOD) to each variant. Variant effects annotation is performed using SnpEff (PMID: 22728672), bcftools (http://github.com/samtools/bcftools) and in-house software. Other functional annotations include variant frequencies in different populations from 1000 Genomes project (PMID: 20981092), Exome Aggregation Consortium – ExAC (http://biorxiv.org/content/early/2015/10/30/030338), dbSNP 138 (PMID: 11125122); cross-species conservation scores from PhyloP (PMID: 15965027), Genomic Evolutionary Rate Profiling (GERP; PMID: 21152010), PhastCons (PMID: 21278375); functional prediction scores from Polyphen2 (PMID: 20354512) and SIFT (PMID: 19561590); variant disease associations from Clinvar (http://www.ncbi.nlm.nih.gov/clinvar/); regulatory annotations from ENCODE (PMID: 15499007), Regulome (PMID: 22955989). Variants and annotations are exported to tabular formats for the ease of downstream analysis. Additional filtration based on functional annotation is applied to extract variants with predicted effects on protein coding.

## Structural Variant Analysis

Structural Variants are called on an individual sample basis using Manta v1.5.0 (PMID: 26647377) and Canvas version 1.40.0.1613 (PMID: 27153601). Manta is a structural variant caller which uses a combination of discordant and split-read support to call deletions, duplications, inversions, and translocations with sizes greater than 50bp, while Canvas is a read-depth based CNV detection tool which provides copy number gain/loss variants for large events (10kbp+). The observed structural variants are annotated with gene and exon overlap, intersected with known SVs (1kg project, DGV), and overlap with regions of poor mappability, simple repeat content, and segmental duplications. By combining these two platforms, we are able to provide a broad view of the structural variation landscape.

## GenomeSTRiP (archived pipeline)

SV are called per batches of approximately 100 samples using GenomeSTRiP (PMID: 21317889). GenomeSTRiP combines discordant read pair analysis with read depth analysis in order to identify large deletions (>100bp,<1Mb). Joint calling of multiple samples and the use of population scale filters help to distinguish real SVs from artifacts. All deletions annotated as PASS in the GenomeSTRiP results were then processed with in-house scripts. First, deletions were merged within the same sample to remove redundant calls. Next, they were merged across samples and annotated with gene overlap, overlap with known deletions (1kg project, DGV) and overlap of predicted breakpoints with repeat regions. Using this annotation with repeat regions we filtered out SVs whose breakpoints fall into sequence with extensive mapping ambiguity or repeat content and that are likely artifact calls due to mismappings.