# Variant Analysis Report

Eurofins Project ID: **NG-12345**

Date of Processing: **29 June, 2023**

Pipeline: **Variant Analysis Pipeline**

Version: **v2.6.8**

*This report is not a diagnostic / clinical report and is intended for Research Use Only.*

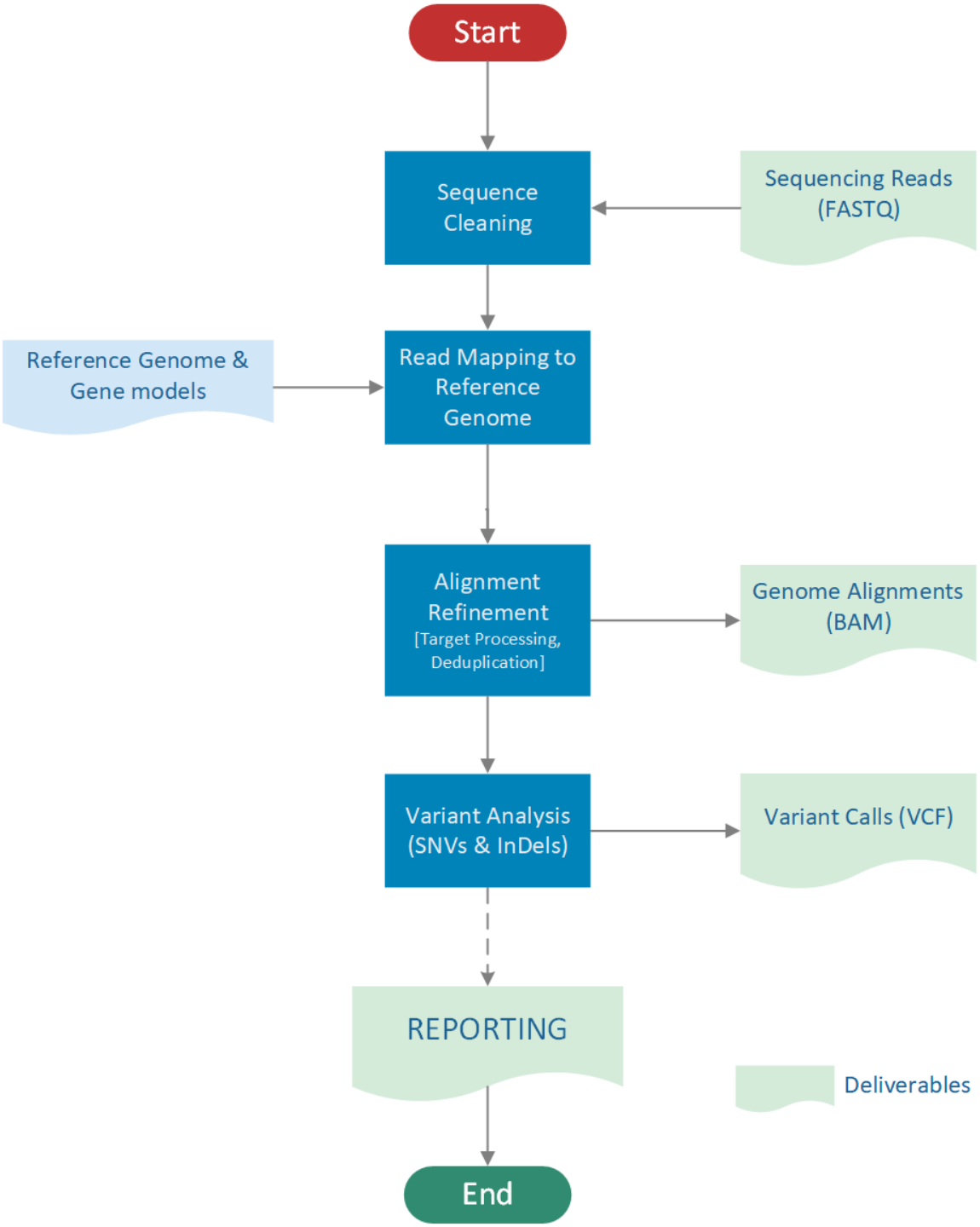
Eurofins Genomics

29 June, 2023

# Contents

# 1 Analysis Workflow

Schematic diagram showing the main steps of the analysis method followed to perform the data analysis.



# 2 Sequence Quality Control

## 2.1 Quality Control of raw sequencing data

Raw sequencing data are preprocessed to generate clean data for downstream analysis.

Quality of raw sequencing data is checked and filtered to retain only high quality bases by performing adapter trimming, quality filtering and per-read quality pruning. Quality is interpreted as the probability of an incorrect base call or, equivalently, the base call accuracy. The quality score is logarithmically based, so a quality score of 10 reflects a base call accuracy of 90%, but a quality score of 20 reflects a base call accuracy of 99% and a quality score of 30 reflects a base call accuracy of 99.9%. These probability values are the results from the base calling algorithm and depend on how much signal was captured for the base incorporation.

Sequencing reads representing reads with quality score at least Q30 is above 90% is of very good quality. For a reasonably good sample source material, according to Illumina specifications, one could expect >75% reads with at least Q30 Phred quality.

Raw sequencing data is processed using fastp[1] software to remove poor quality bases (below Phred Quality 20) using the sliding window approach where in if the average quality of the bases drops below Q20, those bases are removed from the reads. After quality trimming, program checks for presence of any adapters in the reads and removes from the reads. Further, shorter reads which are <30bp length are also removed to retain only high quality sequencing reads for each sample in the analysis. In case of paired-end reads, both the sequencing reads which pass the QC criteria are considered for downstream analysis.

After QC processing, QC metrics such as Q30 reads and GC content can be used to assess the sequencing and sample quality across the samples.

## 2.2 Read Statistics

- Table 1: Sequence Quality Metrics overview. For each sample, the following QC metrics are provided:
  - Sample Name: name of the sample.
  - Total Raw Reads: the total number of raw sequencing reads generated for the sample.
  - Total HQ Reads: the total number of high quality reads after sequence cleaning and filtering.
  - HQ Bases (Q30): Percentage of high quality bases having at least phred quality 30.
  - GC Content: GC content in percentile of high quality sequencing reads.
  - Mean Read Length (bp): Average read length in bp of high quality sequencing reads.
  - HQ Reads %: High Quality Reads percentage.

| ID | Sample Name | Total Raw Reads | Total HQ Reads | HQ Bases (Q30) | GC Content | Mean Read Length (bp) | HQ Reads % |
|---|---|---|---|---|---|---|---|
| 1 | sample1 | 65.09 M | 63.97 M | 90.71% | 49.04% | 148 | 98.28% |
| 2 | sample2 | 59.66 M | 58.7 M | 90.26% | 50.12% | 147 | 98.39% |

# 3 Mapping & Alignment

Mapping to the reference sequence / database is done using BWA[2] run through Sentieon[4] framework.

## 3.1 Alignment Statistics

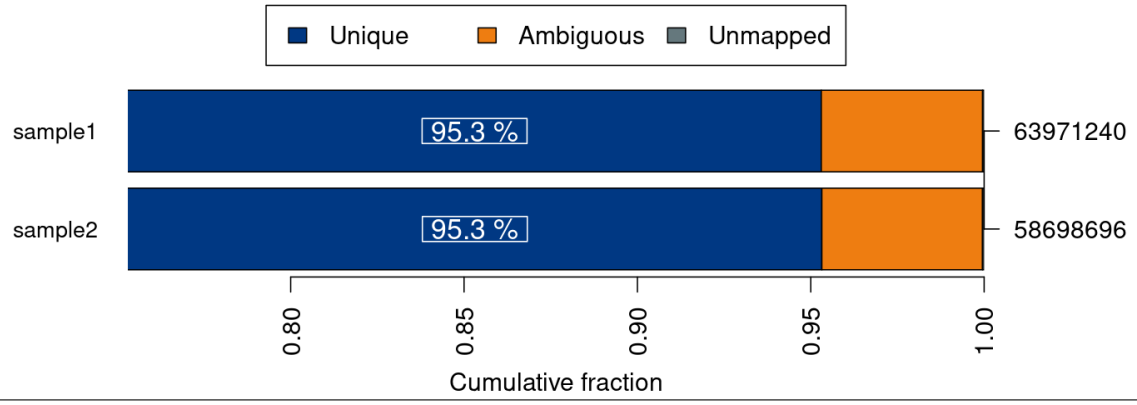
Please note that the mapping efficiency depends on the accuracy of the reference and the quality of sequence reads.

- Mapping statistics overview. For each sample, the following statistics are provided:
  - Total HQ Reads: The total number of reads after filtering the samples.
  - Mapped Reads: Reads mapped to reference.
  - Unique Reads: Reads mapped to exactly one site on the reference.
  - On-target Reads: Uniquely mapped reads that mapped to a target region with +/- 100 bp tolerance.
  - Deduplicated Reads: Reads with PCR duplicates removed.
  - Mean Coverage: Average read coverage of the reference sequence.
  - Mean Coverage (w/o duplicates): Average read coverage of the reference sequence after duplicate removal.

Percentage of reads in category Unique is calculated based on the number of reads mapping to entire reference. Percentage of reads in category On-target is calculated based on the number of reads mapped uniquely. Percentage of reads in category Deduplicated (reads without duplicates) is calculated based on the number of on-target reads.

- Table 2: Mapping metrics per sample.

| ID | Sample | Total HQ Reads | Mapped Reads | Unique Reads | On-target Reads | Deduplicated Reads | Mean Coverage | Mean Coverage (w/o duplicates) |
|---|---|---|---|---|---|---|---|---|
| 1 | sample1 | 63.97M | 63.94M (99.96%) | 60.97M (95.30%) | 51.78M (84.94%) | 38.32M (73.99%) | 116.12x | 85.70x |
| 2 | sample2 | 58.70M | 58.67M (99.95%) | 55.95M (95.32%) | 47.77M (85.38%) | 35.43M (74.16%) | 111.30x | 82.42x |



- Figure 1 : Summary of alignment results. For each sample, the fraction of uniquely mapped, non-uniquely mapped (ambiguous) and unmapped reads relative to the total number of reads per sample (right y-axis) is shown.

## 3.2 Output Files and Descriptions

The following files have been generated for each sample:

- **SAMPLE.REFERENCE.alignment.bam**: These files contain the results of the read mapping in BAM format. The *.bam files are sorted by alignment positions and contain mapped as well as unmapped reads. Use “samtools view -F 4 in.bam > out.sam” to extract mapped reads. Use “samtools view -f 4 in.bam > out.sam” to extract unmapped reads.
- **SAMPLE.REFERENCE.alignment.bam.bai**: The index files of *.bam files are needed e.g. for the visualization with IGV.

# 4 Coverage Report

Depth of coverage has been calculated for each sample.

- Coverage metrics overview. For each sample the following metrics are provided:
  - Mean Coverage: Average read coverage of the reference sequence.
  - Uniformity (0.2x avg.): Sample(s) with high uniformity usually have >90% covered at 0.2x average coverage (e. g. 20x for 100x average coverage)
  - Yx: Percentage of reference covered with minum Y reads.
- Table 3: Depth of coverage summary.

| ID | Sample | Mean Coverage | Uniformity (0.2x avg.) | 2x | 5x | 10x | 20x | 50x | 100x | 120x |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | sample1 | 116.12 | 23x (99.0%) | 99.8 | 99.7 | 99.5 | 99.3 | 96.4 | 58.5 | 37.8 |
| 2 | sample2 | 111.30 | 22x (99.0%) | 99.6 | 99.4 | 99.2 | 99.0 | 94.8 | 53.6 | 34.1 |

The coverage plot shows the base coverage distribution of the aligned data. Depth of coverage is plotted on X-axis and the percentage of the respective reference covered is plotted on Y-axis. The shape of the curve indicates the uniformity of the reference coverage in the samples analysed.

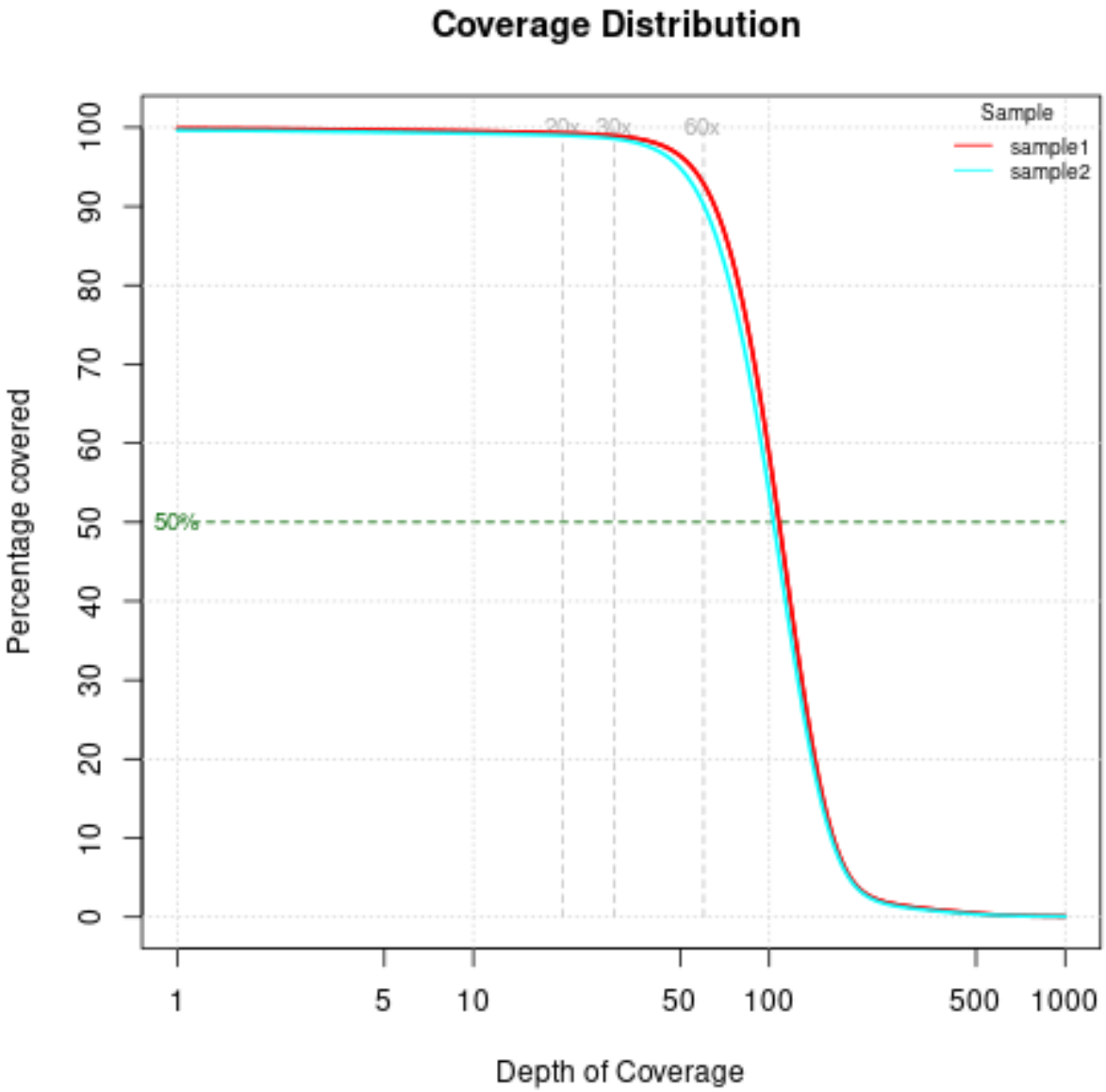



Figure 2 : Coverage plot.

- Table 4: Depth of coverage summary after duplicate removal.

| ID | Sample | Mean Coverage | Uniformity (0.2x avg.) | 2x | 5x | 10x | 20x | 50x | 100x | 120x |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | sample1 | 85.70 | 17x (99.0%) | 99.8 | 99.6 | 99.5 | 99.1 | 89.3 | 23.5 | 9.7 |
| 2 | sample2 | 82.42 | 16x (99.0%) | 99.5 | 99.3 | 99.2 | 98.8 | 85.7 | 21.7 | 9.3 |

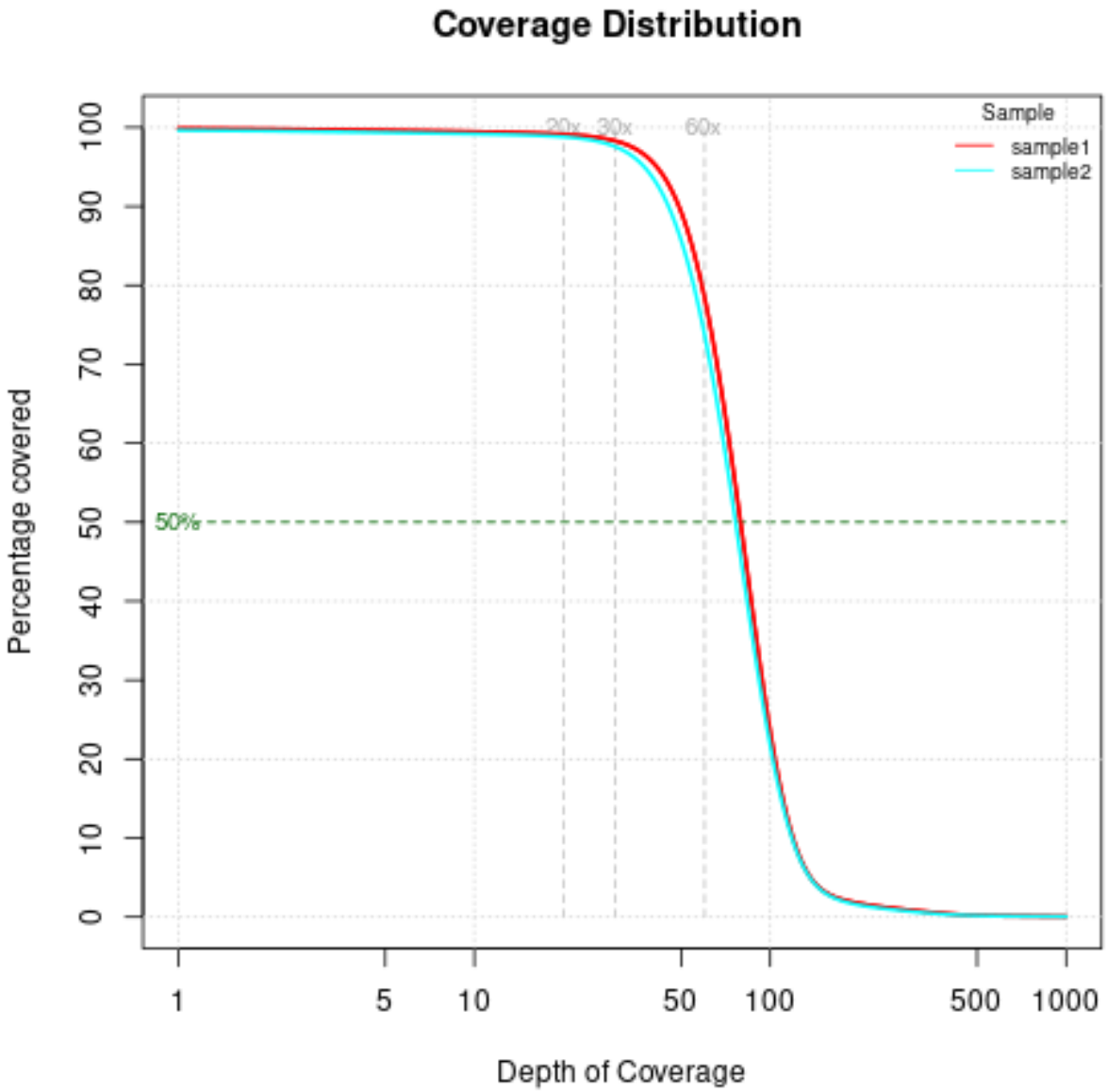



Figure 3 : Coverage plot (excluding duplicated fragments).

# 5 Variant Calling Report

The SNP and InDel calling is done using Sentieon's DNAscope[4].

Variants detected are annotated based on their gene context using VEP[3]. More information about annotations produced by VEP can be found here: https://www.ensembl.org/info/docs/tools/vep/index.html

Please note that the variants are NOT VALIDATED and are provided as they are reported from the programs mentioned above. Therefore it is highly recommended to inspect the variants thoroughly and validate using alternative methods. It is recommended to use filter PASS variants for further downstream analysis. **Note: depending on the input DNA quality and genomic region limit of detection might be > 1%.**

## 5.1 Variant Calling Metrics

- Table 5: Variant metrics per sample.

| ID | Sample | Total | SNP | InDel | Known | Unknown |
|---|---|---|---|---|---|---|
| 1 | sample1 | 140,218 | 122,461 | 17,757 | 139,068 | 1,150 |
| 2 | sample2 | 130,948 | 115,127 | 15,821 | 129,767 | 1,181 |

## 5.2 Variant Consequence Metrics

- Table 6: Count of most severe Variant Consequences per sample.

| Consequence | sample1 | sample2 |
|---|---|---|
| splice_donor_variant | 77 | 73 |
| splice_acceptor_variant | 77 | 69 |
| stop_gained | 159 | 168 |
| frameshift_variant | 418 | 416 |
| stop_lost | 25 | 25 |
| start_lost | 45 | 47 |
| inframe_insertion | 209 | 212 |
| inframe_deletion | 237 | 241 |
| protein_altering_variant | 4 | 5 |
| missense_variant | 14095 | 13931 |
| splice_region_variant | 2716 | 2669 |
| synonymous_variant | 13029 | 12905 |
| stop_retained_variant | 18 | 16 |
| coding_sequence_variant | 4 | 3 |
| 5_prime_UTR_variant | 3778 | 3702 |
| 3_prime_UTR_variant | 6041 | 5542 |
| non_coding_transcript_exon_variant | 2637 | 2404 |
| intron_variant | 90410 | 82364 |
| upstream_gene_variant | 11 | 8 |
| downstream_gene_variant | 10 | 9 |
| TFBS_ablation | 1 | 0 |
| TF_binding_site_variant | 790 | 767 |
| regulatory_region_variant | 2119 | 2095 |
| intergenic_variant | 3308 | 3277 |

## 5.3 VCF Fields

| Field | Description |
|---|---|
| AC | Allele count in genotypes, for each ALT allele, in the same order as listed |
| AD | Allelic depths for the ref and alt alleles in the order listed |
| AF | Allele frequency, for each ALT allele, in the same order as listed |
| AN | Total number of alleles in called genotypes |
| BaseQRankSum | Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities |
| CSQ | Consequence annotations from Ensembl VEP. Format: Allele\| Consequence\| IMPACT\| SYMBOL\| Gene\| Feature_type\| Feature\| BIOTYPE\| EXON\| INTRON\| HGVSc\| HGVSp\| cDNA_position\| CDS_position\| Protein_position\| Amino_acids\| Codons\| Existing_variation\| DISTANCE\| STRAND\| FLAGS\| VARIANT_CLASS\| SYMBOL_SOURCE\| HGNC_ID\| CANONICAL\| MANE_SELECT\| MANE_PLUS_CLINICAL\| TSL\| APPRIS\| CCDS\| ENSP\| SWISSPROT\| TREMBL\| UNIPARC\| UNIPROT_ISOFORM\| REFSEQ_MATCH\| REFSEQ_OFFSET\| GENE_PHENO\| SIFT\| PolyPhen\| DOMAINS\| miRNA\| HGVS_OFFSET\| AF\| AFR_AF\| AMR_AF\| EAS_AF\| EUR_AF\| SAS_AF\| AA_AF\| EA_AF\| gnomAD_AF\| gnomAD_AFR_AF\| gnomAD_AMR_AF\| gnomAD_ASJ_AF\| gnomAD_EAS_AF\| gnomAD_FIN_AF\| gnomAD_NFE_AF\| gnomAD_OTH_AF\| gnomAD_SAS_AF\| MAX_AF\| MAX_AF_POPS\| CLIN_SIG\| SOMATIC\| PHENO\| PUBMED\| MOTIF_NAME\| MOTIF_POS\| HIGH_INF_POS\| MOTIF_SCORE_CHANGE\| TRANSCRIPTION_FACTORS |
| ClippingRankSum | Z-score From Wilcoxon rank sum test of Alt vs. Ref number of hard clipped bases |
| DB | dbSNP membership |
| DP | Combined depth across samples |
| DP | Read depth |
| ExcessHet | Phred-scaled p-value for exact test of excess heterozygosity |
| FS | Phred-scaled p-value using Fisher's exact test to detect strand bias |
| GQ | Genotype quality |
| GT | Genotype |
| InbreedingCoeff | Inbreeding coefficient as estimated from the genotype likelihoods per-sample when compared against the Hardy-Weinberg expectation |
| MLEAC | Maximum likelihood expectation (MLE) for the allele counts, for each ALT allele, in the same order as listed |
| MLEAF | Maximum likelihood expectation (MLE) for the allele frequency, for each ALT allele, in the same order as listed |
| MQ | RMS mapping quality |
| MQRankSum | Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities |
| PGT | Physical phasing haplotype information, describing how the alternate alleles are phased in relation to one another |
| PID | Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group |
| PL | The phred-scaled genotype likelihoods rounded to the closest integer |
| QD | Variant Confidence/Quality by Depth |
| ReadPosRankSum | Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias |
| SOR | Symmetric Odds Ratio of 2x2 contingency table to detect strand bias |

## 5.4 Output Files and Descriptions

The following files have been generated for each sample:

- **SAMPLE.variants.vcf.gz**: Sample-wise variant calls in compressed VCF format.
- **SAMPLE.variants.vcf.gz.tbi**: Index file for sample-wise variant calls in compressed VCF format. This file is needed by different tools when working with vcf.gz files.
- **SAMPLE.variants.passed.vcf.gz**: Filter passed sample-wise variant calls in compressed VCF format.

- **SAMPLE.variants.passed.vcf.gz.tbi**: Index file for filter passed sample-wise variant calls in compressed VCF format. This file is needed by different tools when working with vcf.gz files.
- **SAMPLE.variants.passed.csv**: Filter passed sample-wise variant call in CSV format (can be opened in Excel).
- **SAMPLE.variants.passed.vcf_summary.html**: Detailed information about variant annotation with (VEP / snpEff).

# 6 Relevant Programs

- Table 8: The programs/softwares used in this pipeline.

| Tool | Version | Description |
|---|---|---|
| bamutil | 1.0.14 | BamUtil is a repository that contains several programs that perform operations on SAM/BAM files |
| BCFtools | 1.15 | BCFtools is a set of utilities that manipulate variant calls in the Variant Call Format (VCF) and its binary counterpart BCF |
| BWA | 0.7.17 | BWA is a software package for mapping low-divergent sequences against a large reference genome |
| fastp | 0.20.0 | Fastp is a tool designed to provide fast all-in-one preprocessing for FastQ files. |
| R | 4.1.3 | R is a programming language and environment for statistical computing. |
| sambamba | 0.6.8 | Sambamba is a high performance modern robust and fast tool (and library), for working with SAM and BAM files. |
| Sentieon | 202112.02 | Sentieon® provides complete solutions for secondary DNA/RNA analysis for a variety of sequencing platforms, including short and long reads. |
| snpEff | 4.3 | SnpEff is a genetic variant annotation and effect prediction toolbox. |
| VEP | 104.3 | VEP determines the effect of your variants. |

# 7 Parameter settings

Following parameter have been used for each step:

- Alignment Classification: not unmapped and not supplementary and not secondary_alignment
- Alignment: -k 19 -w 100 -m 50
- Coverage: –count_type 0 –omit_base_output –cov_thresh 2 –cov_thresh 5 –cov_thresh 10 –cov_thresh 20 –cov_thresh 30 –cov_thresh 100 –cov_thresh 120 –histogram_low 1 –histogram_high 1000 –histogram_bin_count 999 –ignore_del_sites
- Multiqc: default
- Sequence Cleaning: –detect_adapter_for_pe –average_qual 20 –length_required 50 –qualified_quality_phred 20
- Variant Calling: default
- Vep: –use_given_ref –everything –sift p –polyphen p –no_intergenic –distance 1 –pick –hgvs –refseq –af_1kg –af_gnomad

# 8 Reference Database

- Table 9: Information about the reference database.

| Tag | Value |
|---|---|
| Genome Name | Homo sapiens |
| Genome Version | hg38 |
| Genome Source | UCSC |
| Genome Size | 3.2 gb |

# 9 Sequence Data Used

- Table 10: The samples used in this pipeline.

| No | Sample | File Name |
|---|---|---|
| 1 | sample1 | NG-12345_sample1_lib123572_1234_2_1.fastq.gz |
| | | NG-12345_sample1_lib123572_1234_2_2.fastq.gz |
| 2 | sample2 | NG-12345_sample2_lib123694_1234_2_1.fastq.gz |
| | | NG-12345_sample2_lib123694_1234_2_2.fastq.gz |

# 10 Deliverables

For details about the files delivered, please refer to each results section.

- **NG-12345.Variant_Analysis_Report.pdf**: This report.
- **SAMPLE.REFERENCE.alignment.bam**: Mapping & Alignment files.
- **NG-12345.alignment_index_files.tar.gz**: Contains all .bam.bai files generated by Mapping & Alignment step.
- **NG-12345.variant_files.tar.gz**: Contains all files generated by Variant Calling step.

# 11 References

[1] Shifu Chen, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, 17 (September 2018), i884–i890. DOI:https://doi.org/10.1093/bioinformatics/bty560

[2] Heng Li and Richard Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* 25, 14 (July 2009), 1754–1760. DOI:https://doi.org/10.1093/bioinformatics/btp324

[3] William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. The ensembl variant effect predictor. *Genome Biology* 17, 1 (June 2016), 122. DOI:https://doi.org/10.1186/s13059-016-0974-4

[4] Sentieon. 2021. *Sentieon provides complete solutions for secondary dna/rna analysis for a variety of sequencing platforms, including short and long reads.* Sentieon. Retrieved from https://www.sentieon.com/