|  |  |  |
| --- | --- | --- |
|  |  | Document Record ID Key |
| Work stream | Project B |  |
| Programme Director | Augusto Rendon | **Status** |  |
| Document Owner | Alona Sosinsky | **Version** | 1.2.main |
| Document Author | ­­­Alona Sosinsky | **Version Date** | 16/01/2017 |

Technical information Document

# Document History

The controlled copy of this document is maintained in the Genomics England internal document management system. Any copies of this document held outside of that system, in whatever format (for example, paper, email attachment), are considered to have passed out of control and should be checked for currency and validity. This document is uncontrolled when printed.

## Version History

|  |  |  |
| --- | --- | --- |
| Version | Date | Description |
| 1.0.main | 05/10/2016 | This Technical information Document will accompany Whole Genome Analysis: Preliminary Analysis document |
| 1.1.main | 11/11/2016 | This Technical information Document will accompany Whole Genome Analysis: Preliminary and Supplementary Analysis documents |
| 1.2.main | 16/01/2017 | This Technical information Document will accompany Whole Genome Analysis: Preliminary and Supplementary Analysis documents |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Reviewers

This document must be reviewed by the following:

|  |  |  |
| --- | --- | --- |
| Name | Title | Version |
| Shirley Henderson |  | 1.2.main |
| Angela Hamblin |  | 1.2.main |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

##

## Approvers

This document must be approved by the following:

|  |  |  |  |
| --- | --- | --- | --- |
| Name  | Responsibility | Date  | Version |
| Augusto Rendon | Director of Bioinformatics | Dec 2016 | 1.1.main |
| Clare Turnbull | Clinical Lead for Cancer Data | 18/01/2017 | 1.2.main |
| Joanne Mason | Director of Sequencing | 05/10/2016 | 1.0.main |

Technical Information Document

Main program

# Sequencing and alignment

Samples were prepared using an Illumina TruSeq DNA Nano, TruSeq DNA PCR-Free or FFPE library preparation kit and then sequenced on a HiSeq X generating 150 bp paired-end reads. Germline samples were sequenced to produce at least 85 Gb of sequences with sequencing quality of at least 30. For tumour samples at least 212.5 Gb were required. Alignments for the germline sample must cover at least 95% of genome at 15x or above with well mapped reads (mapping quality > 10) after discarding duplicates.

# Variant detection

## Small variants

Illumina’s North Star pipeline (version 2.6.53.23) was used for primary WGS analysis. Read alignment against human reference genome GRCh38-ALT+Decoy+EBV was performed with ISAAC (version iSAAC-03.16.02.19); small variants calling together with tumour-normal subtraction - with Strelka (version 2.4.7). Strelka filters out the following variant calls:

* All calls with a normal sample depth three times higher than the chromosomal mean
* All calls where the site in normal sample is not a homozygous reference
* Somatic SNV calls with empirically fitted VQSR score < 2.75 (recalibrated quality score expressing the phred scaled probability of the somatic call being a false positive observation)
* Somatic indels where fraction of basecalls filtered out in a window extending 50 bases to either side of the indel’s call position is > 0.3
* Somatic indels with quality score < 30 (joint probability of the somatic variant and a homo ref normal genotype)

Variants are not removed on the basis of low read count/frequency in the current version of the analysis pipeline. This is to allow for the detection of low level variants but may be reviewed in subsequent versions of the pipeline.

Variants were not filtered out on the basis of being common in the general population. Small indels intersecting with reference homopolymers of at least 8 nucleotides in length have been highlighted on the analysis with an (H): such variants arise commonly, especially in the context of deficits in base-excision repair, but overall have a higher likelihood of being false positive artefacts of sequencing or calling.

## Structural variants

Structural variants (SVs) and long indel (>50bp) calling was performed with Manta (version 0.28.0) that combines paired and split-read evidence for SV discovery and scoring; copy number variants (CNVs) calling - with Canvas (version 1.3.1) that employs coverage and minor allele frequencies into assigning copy number. These tools filter out the following variant calls:

* Manta-called SVs with a normal sample depth near one or both variant break-ends three times higher than the chromosomal mean
* Manta-called SVs with somatic quality score < 30
* Manta-called somatic deletions and duplications with length > 10kb
* Manta-called somatic small variant (<1kb) where fraction of reads with MAPQ0 around either break-end > 0.4
* Canvas-called somatic CNVs with length < 10kb
* Canvas-called somatic CNVs with quality score < 10

# Variant annotation

SNVs and small indels were normalized (left aligned, trimmed, MNVs decomposed), uploaded to Open-CGA and annotated by Cellbase against ENSEMBL (version 82/GRCh38) and COSMIC (version v78/GRCh38) databases. CellBase takes advantage of the data integrated in its database to implement a rich and high-performance variant annotator (with 99.9991% concordance with Ensembl VEP Consequence Types across 1000 genomes phase 3 variants). Only variants annotated with the following consequence types in canonical transcripts (see List of canonical transcripts v1.2) are reported:

|  |  |  |  |
| --- | --- | --- | --- |
| SO term | Consequence type |  |  |
| SO:0001893 | transcript ablation |
| SO:0001574 | splice\_acceptor\_variant |
| SO:0001575 | splice\_donor\_variant |
| SO:0001587 | stop\_gained |
| SO:0001589 | frameshift\_variant |
| SO:0001578 | stop\_lost |
| SO:0002012 | start\_lost |
| SO:0001889 | transcript\_amplification |
| SO:0001821 | inframe\_insertion |
| SO:0001822 | inframe\_deletion |
| SO:0001583 | missense\_variant |
| SO:0001630 | splice\_region\_variant |

PLEASE NOTE:-

1. Complex indels are not annotated at the protein level and are described as “complex variant” in the predicted consequence column
2. Indels intersecting with reference homopolymers of at least 8 nucleotides in length are denoted with an (H) after the annotation in the predicted consequence column.

# Explanation of report fields

## Sample and variant description

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Explanation |  |  |
| Tumour Sample Cross-contamination | Cross-contamination is a measure, which indicates whether the tumour DNA sample is contaminated with DNA from other individuals. Contamination is calculated at homozygote sites derived from germline genotyping array. PASS status means that contamination is below 2%. |
| Reported Tumour Content | Reported tumour content as estimated in host GMC Pathology lab (Low <40%; Medium 40-60%; High >60%.). |
| Gene- or variant– level actionability | List of cancer types with abbreviations can be seen in the Cancer type abbreviations v1.2 document |
| cDNA change | cDNA change was calculated with the [Mutalyzer API](https://www.mutalyzer.nl) |
| VAF | Calculated as alt/(alt + ref) where alt and ref are the number of reads passing filter (read pairs with a mapping quality < 40; read pairs with only a single end mapped or with an anomalous insert size) |
| Gene mode of action | Classification for gene mode of action (oncogene, tumour suppressor or both) was extracted from the manually curated list of Cancer Census Genes (downloaded on 26/09/2016 from <http://cancer.sanger.ac.uk/census>; see the list at Cancer census genes v1.2) |

## Sequencing and coverage quality metrics

All coverage metrics are calculated by including fragments (rather than reads) with minimal base quality of 30 and minimal mapping quality of 10, with duplicates removed. Quality metrics (mapped reads, chimeric DNA fragments and insert size) were calculated with samtools (version 1.1).

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Explanation |  |  |
| Genome-wide coverage mean | Coverage is calculated for autosomes only |
| Unevenness of Local Genome Coverage | Unevenness is calculated as median for the root mean square deviation (RMSD) of coverage calculated in non-overlapping 100 kb windows. This metric would be 0 for genome with absolutely uniform coverage. Typical value for FF samples is in the range 12-15. |
| COSMIC content with low coverage | Percentage of somatic mutations in coding regions reported in COSMIC in multiple samples for which coverage is <30x. Typical value for this metric is < 2%. |