|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Document Record ID Key | |
| Work stream |  |  | |
| Programme Director |  | **Status** |  |
| Document Owner |  | **Version** | 1.0.main |
| Document Author | ­­­Alona Sosinsky | **Version Date** | 03/10/2016 |

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 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Reviewers

This document must be reviewed by the following:

|  |  |  |
| --- | --- | --- |
| Name | Title | Version |
| Shirley Henderson |  | 1.0.main |
| Angela Hamblin |  | 1.0.main |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## 

## Approvers

This document must be approved by the following:

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Responsibility | Date | Version |
| Augusto Rendon | Director of Bioinformatics | 05/10/2016 | 1.0.main |
| Clare Turnbull | Clinical Lead for Cancer Data | 05/10/2016 | 1.0.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

Illumina’s North Star pipeline (version 2.5.55.16) was used for primary WGS analysis. Read alignment against human reference genome GRCh37 was performed with ISAAC (version iSAAC-SAAC00776.15.01.27); small variants calling together with tumour-normal subtraction - with Strelka (version 2.0.14.1). 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 SNVs with high mismatch density where the fraction of basecalls filtered out is > 0.4 in either sample
* Somatic SNVs overlapping spanning deletions, i.e. where the fraction of reads crossing the site with spanning deletion in either sample is > 0.75
* Somatic SNVs with quality score < 15 (joint probability of the somatic variant and a homo ref normal genotype)
* Somatic indels where fraction of basecalls filtered in a window extending 50 bases to either side of the indel’s call position is > 0.3
* Somatic indels with a reference repeat count is > 8
* Somatic indels overlapping ‘interrupted homopolymers’ of > 14 bp in the reference sequence

Variants were not filtered out based on their frequency in common population.

# 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/GRCh37) and COSMIC (version v75/GRCh37) 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). Only variants annotated with the following consequence types in canonical transcripts (see [List of canonical transcripts v1.0](https://www.genomicsengland.co.uk/download/list-of-canonical-transcripts/?wpdmdl=10184&masterkey=57f4ba41621f5)) 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:0001582 | initiator codon variant | | |
| SO:0002012 | start\_lost | | |
| SO:0001889 | transcript\_amplificatio | | |
| SO:0001821 | inframe\_insertion | | |
| SO:0001822 | inframe\_deletion | | |
| SO:0001583 | missense\_variant | | |
| SO:0001630 | splice\_region\_variant | | |
| SO:0001626 | incomplete\_terminal\_codon\_variant | | |

**.**

# Explanation of report fields

## Sample and variant description

|  |  |  |  |
| --- | --- | --- | --- |
| Column name | Explanation |  |  |
| Tumour Sample Cross-contamination | Cross-contamination for tumour samples is calculated by ContEst software. 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 v1.0](https://www.genomicsengland.co.uk/download/cancer-type-abbreviations/?wpdmdl=10183&masterkey=57f4bd65037b5) document | | |
| 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 v1.0](https://www.genomicsengland.co.uk/download/cancer-census-genes/?wpdmdl=10182&masterkey=57f4b8d6f2105)) | | |

## 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%. | | |