|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Document Record ID Key | |
| Work stream | Project B |  | |
| Programme Director | Augusto Rendon | **Status** |  |
| Document Owner | Alona Sosinsky | **Version** | 1.6.main |
| Document Author | ­­­Alona Sosinsky | **Version Date** | 19/12/2017 |

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 |  |
| 1.2.main | 16/01/2017 |  |
| 1.3.main | 05/03/2017 | Section on germline variant calling was added and explanation of report fields was extended. |
| 1.4.main | 30/03/2017 |  |
| 1.5.main | 18/08/2017 |  |
| 1.6.main | 19/12/2017 | Section on samples QC, description of domains for small variants and refined procedure for analysis of germline findings were added |

## Reviewers

This document must be reviewed by the following:

|  |  |  |
| --- | --- | --- |
| Name | Title | Version |
| Shirley Henderson | Lead for Cancer Molecular Diagnostics | 1.2.main |
| Angela Hamblin |  | 1.3.main |
|  |  |  |
|  |  |  |

## 

## Approvers

This document must be approved by the following:

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

# Sample cross-contamination checks

Cross-contamination is a measure, which indicates whether the tumour DNA sample is contaminated with DNA from other individuals. Cross-contamination could potentially lead to false positive results.

## Germline samples

Germline samples are processed with VerifyBamID algorithm and PASS status is assigned to the samples with less than 2% of contamination. We also use bcftools in order to estimate discordance between germline genotyping array and germline WGS as a number of non-matching genotypes. PASS status is assigned to the samples with less than 1% of non-matching genotypes.

## Tumour samples

Tumour samples are processed with ContEst and ConPair algorithms. Two algorithms are currently used to give a greater degree of assurance with the percentage of contaminating DNA. If both algorithms signify the sample has passed or failed then no further action is necessary. If the figures are discordant (i.e. don’t both indicate PASS or FAIL) then further manual investigation of called variants will be required. However no discordant cases have been identified so far (~700 cases are analysed as of October 2017). For some special deliveries (e.g. FT TAT samples) only ContEst is implemented as these samples are delivered without germline genotyping array information.

The PASS status means that contamination is below 1%. Sample FAILs if contamination is above 5%. For contamination in the range between 1% and 5% the report is produced with the percentage of contamination highlighted in the report. In these highlighted cases there will be a greater chance that variants with low variant allele frequency could be false positives due to the cross contamination.

# Sequencing QC: ‘low’ vs ‘sufficient’ quality FF samples

## Sequencing and coverage quality metrics

All coverage metrics are calculated by including non-overlapping bases with minimal base quality of 30, where the read has a minimum mapping quality of 10, after duplicates are removed. Mapped Reads, Chimeric DNA Fragments and Average Insert Size metrics are calculated with samtools (version 1.1). AT/CG Dropout and Unevenness of Local Genome Coverage are calculated with in-house developed tools (see further details for sequencing and coverage quality metrics in section 7.2)

There is a stored cohort of 672 fresh frozen samples for which the above six metrics have been calculated. Metrics for each new sample are compared with this cohort using PCA analysis. Samples with a p < 10-4 (p: probability density after multivariate normal fitting) are classified as outliers. Outliers are of poorer quality and therefore more likely to give false positive or negative result. Outliers should be manually reviewed and a decision made whether the sample should be classified as a potential low quality sample. We are in the process of developing protocol for issuing whole-genome analysis for potential low quality samples as well as FFPE samples.

# 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-Decoy+EBV was performed with ISAAC (version iSAAC-03.16.02.19); small variant calling together with tumour-normal subtraction was performed using Strelka (version 2.4.7).

Strelka filters out the following germline variant calls:

* All calls with a sample depth three times higher than the chromosomal mean
* Site genotype conflicts with proximal indel call. This is typically a heterozygous SNV call made inside of a heterozygous deletion
* Locus read evidence displays unbalanced phasing patterns
* Genotype call from variant caller not consistent with chromosome ploidy
* The fraction of basecalls filtered out at a site > 0.4
* Locus quality score < 14 for for het or hom SNP
* Locus quality score < 6 for het, hom or het-alt indels
* Locus quality score < 30 for other small variant types or quality score is not calculated

Strelka filters out the following somatic variant calls:

* All calls with a normal sample depth three times higher than the chromosomal mean
* All calls where the site in the 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. Small indels in regions with high levels of sequencing noise have been highlighted (N) if at least 10% of the basecalls in a window extending 50 bases to either side of the variant have been filtered out due to the poor quality. These indels have a higher likelihood of being false positive artefacts of misalignment.

## Structural variants

Structural variants (SVs) and long indel (>50bp) calling was performed with Manta (version 0.28.0) which combines paired and split-read evidence for SV discovery and scoring. Copy number variants (CNVs) were called with Canvas (version 1.3.1) that employs coverage and minor allele frequencies to assign 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

# Small variant annotation

SNVs and small indels were normalized (left aligned, trimmed, MNVs decomposed), uploaded to Open-CGA and annotated by Cellbase against ENSEMBL (version 90/GRCh38), COSMIC (version v79/GRCh38) and ClinVar (December 2016 release) 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 v1.6) 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:0001650 | Inframe\_variant | | |
| SO:0001583 | missense\_variant | | |
| SO:0001630 | splice\_region\_variant | | |

**Domain 1 somatic variants**: variants in a virtual panel of potentially actionable genes (132 genes, listed at Actionable genes in solid tumour v1.6 document). Actionable genes are defined as genes in which small variants (SNVs and indels <50bp) have reported therapeutic, prognostic or clinical trial (both actively recruiting participants or closed to recruitment UK trials) associations, as defined by the GenomOncology Knowledge Management System. Where known, the 'variant-level actionability' category and applicable tumour type are indicated. For other variants in these genes, their impact on gene function has not yet been characterised and therefore their actionability status is unclear. This means:

(i) local evaluation will be required for listed variants which are not yet characterised

(ii) even if well characterised as actionable for some tumour types, the listed variants may not be actionable in the participant's specific tumour type

**Domain 2 somatic variants:** variants in a virtual panel of cancer-related genes (566 genes, listed at Cancer census genes v1.6 document). Cancer-related genes are defined as genes in which any variants have been causally implicated in cancer, as defined by the Cancer Gene Census (http://cancer.sanger.ac.uk/census)

**Domain 3 somatic variants:** small variants in genes not included in domains 1 & 2. These are not included in this document but are accessible via the Supplementary Analysis.

PLEASE NOTE:-

1. Complex indels and frameshift variants are not annotated at the protein level
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.
3. Small indels in regions with high levels of sequencing noise where at least 10% of the basecalls in a window extending 50 bases to either side of the indel’s call have been filtered out due to the poor quality are denoted with an (N) after the annotation in the predicted consequence column.

# Germline findings

Analysis for pertinent germline findings was performed to detect pathogenic or likely pathogenic variants in genes conferring susceptibility to the relevant tumour type. For a list of the genes analysed for each tumour type, see The Gene List for Reporting Germline Findings in Cancer Patients v1.6. Pathogenic or likely pathogenic variants include (i) variants predicted to truncate the protein in genes for which the mechanism of pathogenicity is loss of function (variants listed in ClinVar as benign or likely benign with a rating of at least two stars are excluded) (ii) variants listed in ClinVar as pathogenic or likely pathogenic (with a rating of at least two stars). Clinical review within Genomics England was undertaken for all pathogenic or likely pathogenic variants.

# 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 the 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 | Therapies or clinical trials which the patient may be eligible for. Cancer type abbreviations have been used expansions can be accessed in Cancer type abbreviations v1.6 document | | |
| cDNA change | cDNA change was calculated with the [Mutalyzer API](https://www.mutalyzer.nl) | | |
| Population germline allele  frequency | Popula**ti**on germline allele frequencies from two independent datasets are reported: 1000 Genomes and gnomAD. ‘.’ Denotes a luck of variant in the corresponding database. | | |
| VAF | Calculated as alt/(alt + ref) where alt and ref are the number of reads passing filter (reads excluded are 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 14/07/2017 from <http://cancer.sanger.ac.uk/census>; see the list at Cancer census genes v1.6) | | |

## 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 |  |  |
| Mapped Reads | The percentage of reads which can be mapped to the reference sequence. A low percentage could indicate DNA degradation and/or cross-species (e.g. bacterial) contamination. Median values for good quality FF samples are 95.7% with standard deviation of 0.6%. | | |
| Chimeric DNA fragments, % | This metric indicates the proportion of chimeric DNA fragments. Random Inter-chromosomal DNA cross-linking due to DNA strand breakage can cause high proportions of chimeric DNA fragments. This can reflect problems with tissue processing or DNA extraction. For good quality FF samples the median percentage of chimeric DNA fragments is 0.3 % with standard deviation of 0.1%. | | |
| Average fragment size, bp | Short fragments could result from DNA fragmentation due to poor sample handling. Very long fragments (2 or 3 times longer than expected) could result from artefacts introduced during the PCR amplification step, which is required to obtain libraries from samples containing low levels of DNA. Median fragment size for good quality FF samples is 490 bp with standard deviation of 29 bp. | | |
| Genome-wide coverage mean | Coverage represents the median number of reads (depth) per base in the reference genome. Coverage is calculated for autosomes only. The median value for good quality FF samples is 101x with standard deviation of 8x. | | |
| AT dropout  CG dropout | This metrics calculate the percentage of reads that are missing from AT-rich or GC-reach genomic regions. This metric would be 0 for a genome with absolutely uniform coverage. Median values are: 2.0% for AT dropout and 2.1% for CG dropout | | |
| Unevenness of Local Genome Coverage | This metric represents how evenly the read coverage is distributed across the genome. 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 a genome with absolutely uniform coverage. Median value for good quality FF samples is 16.2 with standard deviation of 1.4. | | |
| COSMIC content with low coverage | This metric represents the “discoverability” of known somatic mutations. It is calculated as the % of hypothetical somatic mutation sites (obtained from COSMIC) with coverage of <30x. Median value for this metric for good quality FF samples is 0.9% with standard deviation of 0.2%. | | |
| Total somatic  SNVs, indels and SVs | High numbers of somatic calls can signal a high rate of false positives. However caution is required when interpreting this metric as different tumour types typically have different levels of mutation burden. Additionally, tumours arising from particular mechanisms (e.g severe loss of function in DNA repair genes) may contain very high numbers of somatic mutations. | | |