Cancer Analysis Technical Information Document

Document Key: 616	**Version:** 2.0	**Approval Date:** 17 May 2019 11:24	**Effective Date:** Two weeks after the approval date
Document Author: Ellen McDonagh	**Document Approver:** Augusto Rendon	**Status:** Live	
Department: Bioinformatics	**Document Type:** User Guideline		
Read & Understood Required: No	**Training Session Required:** Yes		
Next Review Date: 17 May 2020	**Total Pages:** **12**		

Table of Contents

1. Document Version History

Version *(newest on top)*	Summary of main changes and reasons
v.2	Ellen McDonagh added explanation that the DPYD gene is on the -ve chromosomal strand, and the genomic allele will be complemented – information pooled from unreleased interpretation document
v.1	New version of technical document corresponding to v1.11 of the Cancer Pipeline

Note: The document version history detailed above is effective for the last live version of this document only. Details of changes for prior versions are available in EQMS.

1. 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.

2. Sample cross-contamination checks

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

2.1 Germline samples

Germline samples are processed with VerifyBamID algorithm and PASS status is assigned to the samples with less than 3% of contamination.

2.2 Tumour samples

Tumour samples are processed with ConPair algorithm. 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 (Tumour Information section). 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. ConPair also highlights the pairs of tumour and germline samples that belong to different patients. Those are reported back to GMCs and replacement samples are requested.

3. Sequencing QC: 'low' vs 'sufficient' quality FF samples

3.1. 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.

4. Variant detection

4.1. 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)
- All calls that overlap LINE repeat region

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. Variants with germline allele frequency > 1% in an internal Genomic England data set have been highlighted (G) to indicate potential un-subtracted germline variant. Recurrently identified somatic variants with somatic allele frequency > 5% in an internal Genomic England data set have been highlighted (R) to indicate potential technical artefact. Variants overlapping simple repeats are denoted with (SR)

4.2. 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

5. Small variant annotation

SNVs and small indels were normalized (left aligned, trimmed, multi-allelic variants decomposed), annotated using Cellbase with the ENSEMBL (version 90/GRCh38), COSMIC (version v86/GRCh38) and ClinVar (October 2018 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 v1.10) 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 (143 genes, listed at Actionable genes in solid tumour and 118 genes, listed at Actionable genes in haemonc v1.10 document). For haematological malignancies the information displayed relates to either lymphoid tumours (where the referred tumour type is lymphoid), myeloid tumours (where the referred tumour type is myeloid) or all haematological tumours i.e. lymphoid and myeloid (in the unusual case that the referred tumour type has features of both e.g. biphenotypic leukaemia). 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 (535 genes, listed at Cancer census genes v1.10 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.
4) Variants with germline allele frequency > 1% in an internal Genomic England data set (indicates potential un-subtracted germline variant) are denoted with an (G) after the annotation in the predicted consequence column
5) Recurrently identified somatic variants with somatic allele frequency > 5% in an internal Genomic England data set (indicates potential technical artefact) are denoted with an (R) after the annotation in the predicted consequence column
6) Variants overlapping simple repeats are denoted with an (SR) after the annotation in the predicted consequence column

6. Germline findings

6.1. Tier 1

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 corresponding gene panel at https://panelapp.genomicsengland.co.uk/. 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). Only homozygote or compound heterozygote variants are reported for genes with biallelic mode of inheritance (as documented on PanelApp). Variants near the 3' end of the gene should be carefully evaluated as truncation near the C-terminal end of the protein may not impair its function. Two variants in MSH2 gene that arise as technical artefacts with high recurrence are blacklisted from the report (chr2:47414419:GT>G and chr2:47414419:GTA>G)

Clinical evaluation of pathogenicity should be confirmed locally. If a variant is deemed relevant, it is recommended that you review the variant in the IGV provided and assess via ACMG criteria. Variants should be technically validated locally. Variants returned to the patient should be reported in the feedback form

6.2. Tier 3

The criteria for Tier 3 analysis were evolved by the NHSE V&R working subgroup on cancer germline findings. In addition to the tumour-type specific panel, a broad panel spanning cancer susceptibility genes is applied (see below for panel names and https://panelapp.genomicsengland.co.uk/ for the gene lists). Variants of the consequence type listed in section 5 are included, where the frequency of the variant in the full dataset is <0.5% (for dominantly-acting genes) and <2% (for recessively acting genes) unless variant is listed in ClinVar as benign or likely benign with a rating of at least two stars. Variants in genes on the germline panel for the relevant tumour type are placed at the top of the list and marked with asterisk.

Patient	**Panels applied**
Adult with solid tumour	Adult solid tumours
Child with solid tumour (age < 18 years old or submitted with disease type 'childhood')	Adult solid tumours + Childhood solid tumours
Haematological tumour	Adult solid tumours + Childhood solid tumours + Haematological malignancies

A variant in FANCD2 gene that arise as alignment artefact artefacts with high recurrence is blacklisted from the report (chr3:10046720 CTTAGTAAG>TTTAT)

It is not anticipated or required that Tier 3 will be reviewed for all patients, but it may be appropriate to review if (i) the likely susceptibility gene is less well reported on ClinVar such that bone fide pathogenic missense/splicing variants have not achieved 2 stars (ii) there is a high index of suspicion of germline susceptibility in the individual and/or (iii) a strong family history of other cancers.

7. Germline analysis: Pharmacogenomics

The variants listed in Table 7.1 were screened for in the germline genome as part of pharmacogenomics findings. In the csv file of genome analysis results, these variants are denoted as 'PGx' in the Domain field.

These are research results being returned to help clinical teams reduce risk of harm from medications. If the result is intended for use in informing clinical management it should be confirmed using a test accredited for clinical use. It remains the responsibility of the health-care provider to determine the best course of treatment for a patient.

Table 7.1: Pharmacogenetic variants screened

Gene	Nucleotide change[i]	Protein change[i]	rsID	Haplotype name	Genomic coordinates (Grh38)	Allele Functional Status	Activity Score	Evidence level for functional status[iii]
DPYD	c.1905+1G>A	N/A	rs3918290	**2A*	1:97450058C>T	No function	0	Strong Evidence supporting function
DPYD	c.1679T>G	p.I560S	rs55886062	**13*	1: 97515787A>C	No function	0	Strong Evidence supporting function
DPYD	c.2846A>T	p.D949V	rs67376798		1:97082391T>A	Decreased function	0.5	Strong Evidence supporting function
DPYD	c.1129-5923C>G c.1236G>A[ii]	N/A, p.E412E	rs75017182 rs56038477	*Part of the HapB3 haplotype*	1:97579893G>C 1:97573863C>T	Decreased function	0.5	Strong Evidence supporting function

Legend Table 7.1

This table is adapted from the PharmGKB DPYD Allele Functionality Table worksheet[1], as part of the supplemental information for the DPYD Clinical Pharmacogenetics Implementation Consortium[2] (CPIC) guideline[3] (2017).

[i] Nucleotide changes according to reference sequence ENST00000370192, encoded on the negative chromosome strand. Note that nucleotide change with genomic coordinates (Grh38) are complemented to the positive chromosome strand. The PharmGKB database will also represent the genotypes complemented to the positive chromsome strand.

[ii] Genome analysis includes two variants from the HapB3 haplotype, likely to be identified in cis. For interpretation, they should not be considered as two independent variants. According to the CPIC guideline, the c.1129-5923C>G variant (rs75017182) is the likely causative variant underlying the HapB3 haplotype; this variant is in intron 10 and introduces a cryptic splice site and the partial production of a nonfunctional transcript[3]. c.1236G>A (rs56038477) is a synonymous variant in LD with the likely causative variant in Europeans, and thus can act as a proxy/tag Single Nucleotide Polymorphism (SNP) – for example, the *DPYD* prospective testing in the Netherlands screened for the exonic c.1236G>A variant rather than c.1129-5923C>G[4].

[iii] Only variants with strong evidence supporting an effect on function, according to the CPIC guideline[3], were chosen for reporting from the genome analysis. The CPIC guideline[3] includes an additional six variants that have moderate evidence supporting function; these were deemed by the

NHSE-Genomics England Pharmacogenetics Working Group not to have enough evidence at this stage to be included in the analysis.

8. Somatic mutation prevalence (global mutation burden)

We display the tumour mutational burden (TMB) of the patient plotted against the range of TMB values for the respective tumour type and alongside different tumour types that have been sequenced in the 100,000 Genomes Project. TMB is calculated as total number of small somatic variants in domains 1-3 per Mb of coding sequence (32.61 Mb)

9. Explanation of report fields

9.1. 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 1%.
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 v1.10 document
CDS change	Coding DNA change was calculated with the Mutalyzer API
Population germline allele frequency	Population 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 in November 2018 from http://cancer.sanger.ac.uk/census; see the list at Cancer census genes v1.10)

9.2. 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
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%.
Insert size medium, 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.
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.
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

10. References

1. PharmGKB, DYPD reference materials https://www.pharmgkb.org/page/dpydRefMaterials (accessed 6th August 2018).

2. CPIC https://cpicpgx.org/ (accessed 13th February 2019).

3. Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for Dihydropyrimidine Dehydrogenase Genotype and Fluoropyrimidine Dosing: 2017 Update. Amstutz U, Henricks LM, Offer SM, Barbarino J, Schellens JHM, Swen JJ, Klein TE, McLeod HL, Caudle KE, Diasio RB, Schwab M. Clin Pharmacol Ther. 2018 Feb;103(2):210-216. For an interactive version of the guidelines table for fluorouracil therapeutic guidelines see https://www.pharmgkb.org/guideline/PA166122686, and for capecitabine see https://www.pharmgkb.org/guideline/PA166109594.

4. DPYD genotype-guided dose individualisation of fluoropyrimidine therapy in patients with cancer: a prospective safety analysis. Henricks LM et al. Lancet Oncol. 2018 Oct 18 pii: S1470-2045(18)30686-7. doi: 10.1016/S1470-2045(18)30686-7.