GUI-BIO-010 Cancer genome analysis guide

GENOMICS ENGLAND CONFIDENTIAL UNCONTROLLED IF PRINTED

Document Key	GUI-BIO-010
Title	Cancer genome analysis guide
Document Status	Published
Pipeline Version	V5.2 – Petra Release, June 2025
Published Date	11/06/2025
Policy (only if applicable otherwise N/A)	N/A
Document Author(s)	Olena Yavorska, Nadezda Volkova
Document Reviewer	Alex Younger
Document Approver	Alona Sosinsky
Details of Approval (Completed by the QI team)	☒Approved in Confluence ☐ Pre-Approved in EQMS (Evidence in EQMS) ☐ Pre-approved by email (Needs prior authorisation from the Quality Improvement Team) ☐ Reference document - approval not required
Next Review Date	☒ Default (12 months) ☐ Other - please specify
Training Format	☒ Read and understand on Confluence ☐ Course ☐ Competency Assessment
Squad/Teams/Roles to be Trained	

Revision History

The revision history of each document is available in the Confluence Page History. To view details of what was changed, click on the versions to compare and select "Compare Versions".

Pipeline Version	Date (Day/Month/Year)	Summary of main changes and reasons (section no. + Update)
5.2	11/06/2025	Correction to Pipeline Version history – MRD assay design targets were added to version 5.1 of the pipeline. Annotation Data and Software Versions – Updated versions for National Genomic Test Directory and Cancer Census Genes Section 4.2 - SVIG canonical variants: labelling and whitelisting Section 4.3 – Updated rules for interpretation of SV/CNV overlapping regions of interest and thresholds for FLT3 ITD support Section 6.1 - Updated panel name for childhood cancers Section 11.3 - Added disclaimers for MRD Ig/TCR targets

Pipeline Version	Date (Day/Month/Year)	Summary of main changes and reasons (section no. + Update)
		Appendix D – Updated description of CNV plots
5.1 (Corrected)	19/03/2025	Section 9 - Minimal Residual Disease assay targets
5.1	19/02/2025	Section 2.5 - Tumour sample degradation QC Section 3.2 - Dragen v4.0.5b somatic CNV calling and calling DUX4 fusion with Pelops Section 4.3 – Somatic variant interpretation: CNVs & SVs Section 10.2 - Unevenness of Local Genome Coverage Section 10.3 - Potentially/Likely degraded tumour sample disclaimers Section 10.4 – Variant descriptions Section 11 – Additional files available on Interpretation Portal Appendix A – Validation sets and metrics, VAF imprecision, assessment of false positive rates Appendix C – Pipeline limitations Appendix D – CNV visualisation
5	1/10/2024	Section on Annotation Data and Software Versions Section 4.1 – Canonical transcriptions and regions of interest Section 4.3.2 – Domain 1 somatic SV/CNVs Section 5.1 – Germline small variants Supporting or Reference documents
4	22/05/2023	Section 4.2.1 - a note about splice site fusions which currently will have a reported frame but should be considered ambiguous
3.7	01/05/2023	Section 2.8 – explanation of the QC checks applied to tumour only samples Section 3.4 - description of the tumour only variant calling and results Section 6.2.3 - description of germline prioritisation procedure for tumour only samples
3.6.2	15/10/2023	Section 3.3 - explanation of N/A in variant origin field on TINC reports Appendix D – update to linear CNV plot examples, illustrating changes to the visualisations that enable regions of homozygosity (ROH) and coverage drops to be easily distinguished

Pipeline Version	Date (Day/Month/Year)	Summary of main changes and reasons (section no. + Update)
3.5	20/09/2023	Section 2.8 – handling of unaccredited sample types. Section 3.3 – updating the description of the logic behind TINC workflow selection. Lists of actionable genes for Domains 1 and 2 have been updated to align with National Genomic Test Directory v7.1 – please refer to "Cancer analysis additional information v3.5.xlsx" on FutureNHS for details.
2.47.2	09/08/2023	Section 3.1 - additional clarification regarding measurement of uncertainty of variant allele frequencies (VAF).
2.47.1	08/06/2023	Section 3.1 - addition of statement regarding measurement of uncertainty of variant allele frequencies (VAF). Section 3.2 - addition of statement regarding the uncertainty around detection of structural variant (SV) breakpoints. Section 5.1 - removed statement indicating that germline CNVs are not currently reported – germline CNVs have been reported as of version 2.47. Section 8 – statement clarifying that tumour mutational burden are not currently covered in our accredited scope under ISO 15189. Section 9 – statement clarifying that mutational signatures are not currently in our accredited scope under ISO 15189.
2.47	22/03/2023	Section 4: Updated prioritisation of somatic variants according to National Genomic Test Directory Sections 2.4, 5.2, 6.3 and 10.3: New sections for germline CNV findings, quality checks and disclaimers for germline CNVs Appendix C: New disclaimer for accuracy of variant calling in DUX4 and other genes overlapping regions of segmental duplications
2.28	30/11/2022	General: some sections may have moved when new sections have been added, typos corrected, font updated Section 4.2.1: New sub-section describing the fusion outcome prediction feature

Pipeline Version	Date (Day/Month/Year)	Summary of main changes and reasons (section no. + Update)
		Section 8: New section describing the new linear CNV plots – depth of coverage, B-allele, and absolute allele count Section 10: New section describing the additional files that are available for download in the interpretation portal Appendix D: New section providing additional information on the new linear CNV plots
2.27	20/08/2022	Version number updated to reflect pipeline version.
2.23	13/07/2022	Updated list of canonical transcripts to ensure that known sarcoma and haematological fusions are correctly reported: added MECOM downstream, GATA2 enhancer and BCOR downstream regions, alternative transcripts for ABL1 and STAT6, replaced the canonical transcript for CBFB Section 3.3: Explanation of the variant recovery workflow for tumour in normal contaminated samples Section 4.2.1: Updated the information about tiering of certain non-coding regions Section 8.4: Description of the population germline allele frequency column for structural variants
2.21	16/09/2021	Section 3.2: Explanation of somatic score assignment for SS18-SSX2/SS18-SSX4 fusions Section 5: Addition of germline variant prioritisation details for new clinical indications Section 8.4: Updated allele frequency annotation for small variants
2.15.2	14/07/2021	Section 3.2.1 added Description of defect in loss of heterozygosity (LOH) calling Clarification of number of genes used in prioritisation of domain 1 variants Section 4.1 and 5 Update to versions for annotation databases
2.15.2	15/03/2021	Formatted to fit the ISO standard template
2.15.1	16/02/2021	Clarification of sequencing quality metrics
2.15	03/02/2021	Addition of guidance for SS18-SSX2/SS18-SSX4 fusion detection
2.10	24/12/2020	This is the first published version of this document

**Please note latest confluence version cannot be added before document is published and should be amended at the next document review*

Purpose

The purpose of this document is to provide NHS Clinical Scientists, Clinicians, Bioinformaticians and others within the NHS Genomic Laboratory Hubs (GLHs) with a guide to the Genomics England workflow for data analysis, annotation, and interpretation in Cancer. This guide includes the processes carried out from the receipt of clinical and genome sequencing data through to presentation of data in the Interpretation Portal.

Scope

In Scope

- Description of the whole genome sequence analysis performed in the cancer bioinformatics pipeline for GMS, including variant calling and interpretation.

Out of Scope

- Description of the Interpretation Portal or Decision Support tools.

Internal Audience

- Genomic Data Scientists and Bioinformatics Engineers in Cancer Squad

External Audience

- NHS Clinical Scientists, Clinicians, Bioinformaticians
- NHS Genomic Laboratory Hubs (GLH) members

Other Third Party Audience

The external audience for this document may include medical device regulators and associated agencies in the pursuit of medical device regulatory and standards certification including:

- UK Competent Authority: (CAs) the Medicines and Healthcare Products Regulatory Agency (MHRA);
- Notified Bodies (NBs) such as BSI Group;
- NHS Digital; the NHS IT regulator in England and Wales

This document may also be requested by existing and prospective Genomics England customers as part of their procurement process. All external distribution of this document must be approved by a member of the Quality Improvements and Regulatory Affairs team prior to circulation.

Abbreviations/Definitions

Abbreviation	Description
GLH	Genomic Laboratory Hub
CNV	Copy Number Variant
GMS	Genomic Medicine Service
IGV	Integrative Genomics Viewer
ITD	Internal Tandem Duplication
LOH	Loss of Heterozygosity
NGTD	National Genomic Test Directory
MNV	Multi Nucleotide Variant
MRD	Minimal Residual Disease
QC	Quality Control
SNV	Single Nucleotide Variant
SV	Structural Variant
TINC	Tumour in Normal Contamination
VAF	Variant Allele Frequency
WGA	Whole Genome Analysis
WGS	Whole Genome Sequencing

Annotation Data and Software Versions

Annotation Data Versions

Name	Version
Genome	GRCh38+Decoy+HLA
CellBase	v5.4.25
Ensembl	v107
MANE	v1.0

COSMIC	v96
ClinVar	2023-04
gnomAD exomes	2.1.1
gnomAD genomes	3.1.2
National Genomic Test Directory	v11
Cancer Gene Census	V101

Software Versions

Name	Version	Reference(s) and Documentation
Strelka	2.9.9	https://github.com/Illumina/strelka?tab=readme-ov-file Strelka2: fast and accurate calling of germline and somatic variants. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, Saunders CT. Nat Methods. 2018 15(8):591-594
Dragen	v4.0.5b	https://support.illumina.com/content/dam/illumina-support/documents/downloads/software/Dragen/release-notes/200037976_00_DRAGEN-4.0.5-Customer-Release-Notes.pdf
Manta	1.5.0	https://github.com/Illumina/manta Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Bioinformatics. 2016 15;32(8):1220-2.
Juli	v0.1.6	Junction Location Identifier (JuLI): Accurate Detection of DNA Fusions in Clinical Sequencing for Precision Oncology. Shin HT, Kim NKD, Yun JW, Lee B, Kyung S, Lee KW, Ryu D, Kim J, Bae JS, Park D, Choi YL, Lee SH, Ahn MJ, Park K, Park WY. J Mol Diagn. 2020; 22(3):304-318
Pindel	0.2.5b9	Split-Read Indel and Structural Variant Calling Using PINDEL. Ye K, Guo L, Yang X, Lamijer EW, Raine K, Ning Z. Methods Mol Biol. 2018;1833:95-105.
Pelops	v0.8.048b	Whole genome sequencing provides comprehensive genetic testing in childhood B-cell acute lymphoblastic leukaemia. Sarra L. Ryan, John F. Peden, Zoya Kingsbury, Claire J. Schwab, Terena James, Petri Polonen, Martina Mijuskovic, Jenn Becq, Richard Yim, Ruth E. Cranston, Dale J. Hedges, Kathryn G. Roberts, Charles G. Mullighan, Ajay Vora, Lisa J. Russell, Robert Bain, Anthony V. Moorman, David R. Bentley, Christine J. Harrison & Mark T. Ross. Leukemia. 2023; 37, 518-528,

VerifyBamID	2.0.1	Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. Am J Hum Genet. 2012 91(5):839-48
ConPair	0.1.0-1	Conpair: concordance and contamination estimator for matched tumor-normal pairs. Bergmann EA, Chen BJ, Arora K, Vacic V, Zody MC. Bioinformatics. 2016 32(20):3196-3198
SAMtools	1.9	Twelve years of SAMtools and BCFtools Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li *GigaScience*, Volume 10, Issue 2, February 2021, giab008, https://doi.org/10.1093/gigascience/giab008
CCube	1.0	Ccube: A fast and robust method for estimating cancer cell fractions. Ke Yuan, Geoff Macintyre, Wei Liu, PCAWG-11 working group, Florian Markowetz. bioRxiv

Introduction/Background

The Genomics England Cancer pipeline aims to facilitate the identification of genomic variants that may be of actionable benefit for the patient. Genomics England is not performing a clinical interpretation of the genome sequencing data. It is the responsibility of NHS GLH staff to perform a full clinical review, confirm the presence of selected variants where required, and report and authorise any results.

Authorities and Responsibilities

N/A

Procedure Details

N/A

1 Sequencing data and alignment

Cancer Pipeline for Genomic Medicine Services (GMS) is using GRCh38+Decoy+HLA reference human genome. Alignment for both the tumour and germline samples is performed using the Dragen aligner, including alternate haplotypes (ALT contigs) with ALT-aware mapping to improve the specificity of mapping and variant calling. Genome alignments are stored in CRAM files which contain both mapped and unmapped reads.

2 Sample and sequencing quality checks

All genomic data are subject to a series of quality control checks performed in the Genomics England automated pipeline to ensure they are of sufficient quality and are suitable for processing.

2.1 Sequencing data

The following quality checks are performed to assess the data and coverage for each genome sequenced:

- Intake QC (all samples): This includes MD5 check to ensure integrity of the files transferred.
- Germline samples: 95% of the autosomal genome covered at ≥15x calculated from reads with mapping quality >10 AND >85x10^9 bases with Q≥30, after removing duplicate reads and overlapping bases (for forward and reverse read) after adaptor and quality trimming.
- Tumour samples: >2.1x10^{11} bases with Q≥30, after removing duplicate reads and overlapping bases after adaptor and quality trimming.

No genome coverage threshold is required for saliva derived germline samples. A warning will be displayed in the whole genome analysis (WGA) results HTML if <95% of the autosomal genome is covered at 15x OR the mean genome coverage is <15x for a saliva derived germline sample.

2.2 Sample cross-contamination checks

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

Cross-patient contamination in germline samples

Germline samples are assessed with the VerifyBamID algorithm[1] to check for cross-patient contamination. Samples with less than 3% contamination are considered as passing. All samples with germline contamination >3% are reported to NHS GLHs in the Sample Failures report.

If germline contamination is between 3% and 8%, WGA data are returned for somatic variants only with a warning indicating that germline contamination was detected, with the option to replace the germline sample if germline variants are required. If germline contamination is >8% no data are returned.

See Appendix A – Validation data for further information on the derivation of contamination thresholds.

Cross-patient contamination in tumour samples and sample swaps

Tumour samples are assessed using the ConPair algorithm[2]. Samples with ≤1% contamination are considered as passing contamination quality control and if contamination ≥2.5%, the sample is considered as failing. For contamination between 1% and 2.5%, the percentage of contamination is highlighted in the WGA results HTML (Tumour Sample section). Low levels of contamination may result in erroneous reporting of contaminating germline variants as somatic variants. Consequently, specificity of somatic variant detection is significantly reduced and tumour mutation burden can be overestimated. ConPair also detects instances in which tumour and germline samples analysed as a pair belong to different patients (sample swap). These are reported to NHS GLHs in the Sample Failures report and replacement samples are requested.

See Appendix A – Validation data for further information on the derivation of contamination thresholds.

2.3 Discrepancy between the reported and inferred sex

As part of the bioinformatics pipeline, the karyotypic sex is inferred from the genomic data (using the germline sample). This is compared with the sex reported in the test order, which may be taken from the NHS Spine. If there is a discrepancy between the reported and inferred sex, queries are raised with NHS GLH staff. If NHS GLH staff confirm that the discrepancy is expected, genomic data can pass through analysis and will be displayed in the Interpretation Portal with a flag (Inferred_genetic_and_reported_sex_discordant).

2.4 Germline samples with poor quality copy number calls

For a small proportion of germline samples, the sequencing data are not of sufficient high quality to make reliable calls for Copy Number Variants (CNVs). Sample level quality control is performed on the number and ratio of different copy number call types and the proportion of common CNVs detected. If CNV data for the germline sample do not pass this quality control step, the germline sample section in the WGA results HTML will contain a disclaimer. The current thresholds for CNV quality control are applied on the original Dragen CNV VCF and the following:

'Poor quality CNV calls'

i) count of autosomal PASS CNVs >= 600 OR

ii) $\log_2$(Loss/Gain counts) < -0.5

'Suspected poor quality calls'

i) count of autosomal PASS CNVs >= 200 or <= 50 OR

ii) $\log_2$(Loss/Gain counts) <=-0.3 or >= 1.2 OR

iii) the fraction of common autosomal PASS CNV calls is <= 0.4. For this purpose, a CNV is defined as common if it has 50% reciprocal overlap with a CNV from https://www.ncbi.nlm.nih.gov/dbvar/studies/estd20/

2.5 Tumour sequencing and coverage quality metrics: 'likely degraded', 'potentially degraded', and 'sufficient' quality samples

All coverage metrics are calculated by including non-overlapping bases with minimal base quality of 30, where the read has a minimum mapping quality >10, after duplicates are removed. Mapped reads, chimeric DNA fragments and average insert size metrics are calculated with SAMtools. AT/CG dropout and evenness of local genome coverage are calculated with in-house developed tools (see further details for sequencing and coverage quality metrics and typical values for good quality samples in section 11.2). In order to assess sample degradation, median coverage depth for consecutive genomic regions is calculated and compared for each pair of neighboring bins. Evenness of local genome coverage metric is calculated as the proportion of bin-pairs in the genome where the difference in median coverage between the bins is less than 5%.

Tumour sample quality is assessed as "Sufficient", "Potentially Degraded", or "Likely Degraded" by an in-house classification tool developed using quality assessments of 700 tumour samples from the 100,000 Genomes Project and the GMS. The tool uses 7 QC metrics to determine sample quality: evenness of coverage metrics for three bin sizes (100Kb, 1Mb and 10Mb), AT dropout, average and standard deviation of insert size, and the number of structural variants (to avoid false degradation labeling for samples with chromosomal instability). These metrics are combined by a support vector machine to produce a quality score, which is used to determine the sample's quality label. Samples classified as "Potentially Degraded" or "Likely Degraded" have additional disclaimers in their WGS analysis report.

2.6 Samples with low tumour purity or high incidence of somatic variants with low variant allele frequency

The percentage of cancer cells in a tumour sample is calculated using the Ccube algorithm[3] and presented in the WGA results HTML (Tumour Sample section). For samples with low tumour content (<30%), the sensitivity of somatic variant detection is significantly reduced, and tumour mutation burden and mutational signatures are not reliably calculated.

The distribution of variant allele frequencies (VAFs) from the somatic variants detected is also examined as tumour content cannot be reliably estimated from genomic data if the percentage of somatic variants with low VAF (<6%) is high (>40%). High levels of low VAF variants usually indicate that the sample either has very low tumour content or high heterogeneity, which impact the sensitivity of variant detection.

2.7 Tumour contamination in germline samples

If the germline DNA sample is contaminated with DNA originating from the tumour, there is a risk of an increased number of false negative somatic variants as true somatic variants may be inappropriately subtracted in the analysis. This is most commonly observed in haematological cancers. To identify normal samples with tumour in normal contamination (TINC), a quality control component has been designed (TINC test[4]), which identifies clonal mutations in the tumour sample and subsequently estimates the fraction of TINC by assessing the allele fraction of these variants in the germline sample. Warnings are displayed in the WGA results HTML for high TINC when the level of contamination is >5% and low TINC when the level is between 1-5%. In the event of low tumour content (<25%) or small size of the subset of clonal mutations (<40 mutation), TINC cannot be estimated reliably, and warnings are displayed.

2.8 Unaccredited sample types

Any sample types that are not currently within the scope of Genomics England's accreditation must undergo a pilot study to ensure that quality is satisfactory. During this pilot study, cases utilising new sample types will be returned through the Interpretation portal, but with a disclaimer and headers highlighted in green. This disclaimer should be included in any downstream reports issued to clinicians.

When the pilot study for each new sample type is complete, the sample type will automatically be brought into Genomics England's schedule of accreditation. Reports issued before completion of pilot study will not be retroactively modified.

In the absence of a germline sample, submissions that contain only tumour samples (and which are a part of Tumour First – Germline Later workflow) will not undergo cross-patient contamination, reported vs inferred sex discrepancy and tumour content checks as well as any check for the germline sample.

3 Somatic variant detection

3.1 Small variants

Somatic small variant detection (single nucleotide variants (SNVs) and indels < 50bp), with germline subtraction, is being performed using Strelka[5]. Normalisation of variant calls is performed, including left alignment, trimming, decomposition of multi-allelic variants and decomposition of multi nucleotide variants (MNVs).

Strelka2 filters out somatic variant calls based on the following:

- Somatic Empirical Variant Score (SomaticEVS) below 7 for SNVs and below 6 for indels.
- Read depth for the tumour or normal sample below 2
- Read depth for the locus is greater than 3x the mean chromosome depth in the normal sample

Variant Allele Frequency (VAF) is a surrogate measure of the proportion of DNA molecules in the original specimen of tumour tissue (contaminated with normal) carrying the variant. Those performing interpretation should therefore be aware that there is uncertainty around the measurement of VAF due to the near random sampling of the tumour. Quantification of this uncertainty is presented in Appendix A (Measurement of uncertainty (confidence intervals) for Variant Allele Frequency for somatic small variants).

Variants are not currently removed on the basis of low read count/VAF or germline allele frequency in the general population (except in the tumour only or TINC workflows). This is to allow for the detection of low-level variants but may be reviewed in subsequent versions of the pipeline. However, the following flags are added to highlight variants with a higher likelihood of being false positive calls or unsubtracted germline variants.

Variant flag	Indication	Implication
(H)	Small indels intersecting reference homopolymers ≥8bp (where a single non-homopolymer base is permitted)	Commonly arising variants, especially in the context of base-excision repair deficits, but with an overall high incidence of false positive variant calls.
(N)	Small indels in regions with high levels of sequencing noise (>10% of base-calls in a 100bp window around the variant are of poor quality)	Variants with high likelihood of being false calls due to misalignment

Variant flag	Indication	Implication
(GG)	Variants with germline allele frequency >1% in the gnomAD exomes database "-" suggests that this variant was not seen in the gnomAD exomes database	Potential unsubtracted germline variants
(GE)	Variants with germline allele frequency >1% in an internal Genomics England dataset of >6,000 unrelated individuals "-" suggests that this variant was not seen in the Genomics England dataset	Potential unsubtracted germline variants
(R)	Variants with somatic allele frequency >5% in an internal Genomics England dataset	Potential artefact of sequencing or variant calling
(SR)	Variants overlapping simple repeats be Tandem Repeats Finder	Commonly occurring variants with high incidence of false positive variant calls
(PON)	SNVs with a Somatic Fisher Phred score <70 based on comparison of alternative allele depths with those at the equivalent variant site in a panel of normals (genomes for 10,000 GMS unrelated rare disease patients or their relative)	Potential false positive variants due to alignment or sequencing errors
(NP)	Variant that was called by the variant calling tool but did not meet the quality thresholds to be labelled as PASS. Only high impact variants from SVIG-UK canonical variant list are currently being whitelisted and reported as non-PASS variants.	Low confidence variant

3.2 Copy number and structural variants

Copy number variant calling

Copy number variants (CNVs) are being detected by Dragen which utilises read depth and minor allele frequencies to assign copy number states. Dragen integrates germline small variants in its somatic CNV detection model but does not strictly perform subtraction of germline CNVs. CNV start and end coordinates have 1 kb resolution due to the limitations of CNV calling algorithms (coverage and minor allele frequencies are calculated in kb bins).

To avoid fragmentation of CNVs, we implemented a blacklist of recurrent somatic CNVs from 353 tumor genomes. Somatic CNVs from liquid (n=25,000) and solid tumors (n=30,000) were aggregated (separately for GAINs, LOSSes, and LOHs) and genomic regions that appeared in at least 30% of samples of each tumor type were included into this blacklist. Please be aware that blacklisted regions overlap with *DUSP22* and *PDE4DIP* genes. The blacklisted regions are excluded from somatic CNV calling. If flanking CNV segments have matching copy number states as well as other parameters (such as allele frequency, sub-clonal status, and copy number), they are merged across the excluded blacklisted region. The use of this blacklist reduces the occurrence of somatic CNVs under 10kb.

Dragen includes improvements in model selection based on coverage, BAF, and VAF values. When these values provide insufficient information, Dragen defaults to a diploid model and purity estimate is not provided. In such cases, somatic CNV calls are generated, but all CNVs are flagged as non-PASS. We will continue to include all CNVs, including non-PASS calls in the WGA results.

Dragen somatic CNV caller labels CNVs as sub-clonal if the copy number of a segment is heterogenous among different subclones. These segments are annotated with 'LSC' (i.e., likely sub-clonal) flag in the WGA results.

Uncertainty in calling high-ploidy genomes

Dragen has difficulty in accurately distinguishing between genomes with numerous sub-clonal variants and those that are high ploidy. This limitation may lead to false negative calls for tetraploid genomes. We are actively collaborating with the Dragen team to develop a solution for detecting these genomes more reliably. In the meantime, we recommend cross-validating with orthogonal tests when high ploidy has clinical significance.

Structural variant calling

Detection of somatic structural variants (SVs) and indels >50bp with germline subtraction is being performed with Manta[6] which combines paired and split-read evidence for SV discovery and scoring. The following variant calls are filtered out by these tools:

- Manta-called SVs when the depth in the normal sample near one or both breakpoint(s) is three times higher than the chromosomal median
- Manta-called SVs with somatic quality score <30
- Manta-called somatic small variant (<1 kb) where the fraction of reads with MAPQ0 in the normal sample around either breakpoint is >0.4

Users need to be aware of uncertainty for detection of SV breakpoint especially in scenarios where breakpoint overlaps with repetitive sequence and therefore reported location may be imprecise. A manual review of breakpoint location in IGV is always recommended.

Detection of internal tandem duplications (ITDs) at the FLT3 locus is also performed using PINDEL[7] to ensure sensitivity of variant detection. Occasionally, Manta can also call FLT3 ITD variants as duplications or insertions; if the genomic location of Manta FLT3 ITD calls coincides with the ones called by PINDEL, these Manta calls will not be reported.

Due to SSX2 and SSX4 genes residing in the regions of segmental duplication that covers two orthologous genes, SSX2/SSX2B and SSX4/SSX4B, read mapping to these regions is ambiguous. As a result, there is a risk of false negative variants for fusions involving SSX2 or SSX4 using Manta. Consequently, for detection of SSX2/SSX4 fusions with *SS18* gene, the Junction Location Identifier[8] (JuLI) algorithm is being used with adjusted parameters for reads with low mapping quality, with variant calling limited to the *SS18*, *SSX2/SSX2B* and *SSX4/SSX4B* genes. Variants are only called when there is a high level of confidence that the variant is not present in the germline sample. JuLI does not produce somatic quality scores for variant calls, therefore, when detected, SSX fusion variants are artificially assigned a high somatic score to be consistent with the format of other variant calls.

DUX4 rearrangements are detected by Pelops[9], which identifies fusion reads mapping to DUX4 or its paralogs with high sensitivity using mapping results from Dragen. Pelops detects the fact of a rearrangement, but not the precise breakpoint coordinate: since DUX4 has many paralogs in different locations, it is not possible to unambiguously locate the break-ends, so we use a mocked location called "DUX4 region" which is a blanket term for 26 DUX4 paralogues located on chromosomes 4 and 10 that are used for rearrangement detection. Consequently, the partner break-end location also cannot be precisely determined, so Pelops is returning a 1-2 kb region instead. Affected region for the rearrangement partner is annotated based on the center of that region. All variants detected by Pelops have (ABP) flag which highlights the ambiguous nature of reported breakpoints. The presence of such fusions can be assessed by manual review of read alignments (see Appendix B – Data presentation in the Interpretation Portal, HTML files and IGV for more details).

3.3 Variant detection with tumour in normal contamination

Contamination of germline samples with DNA derived from a matched tumour sample (tumour in normal contamination, TINC) introduces additional complexity for somatic variant detection. Small variant detection with Strelka2 provides modest resistance to TINC, with estimated sensitivity of ≥95% for small variants with VAF ≥10% with up to 6% TINC, when tumour purity is high (≥60%). Sensitivity in the presence of TINC decreases with both increased TINC and/or decreased tumour purity. SV calling with Manta is very sensitive to TINC and there is a risk of false negative variants with even a low level of contamination. Somatic CNV detection with Dragen is not affected by TINC.

In order to recover sensitivity which may be lost due to TINC, small variants and structural variants in haematological samples (those with clinical indication in the haematological sheet of the National Genomic Test Directory for Cancer, with an exclusion of Histiocytosis – M117) with TINC >1% are also analysed in the parallel pipeline run without subtracting variants from the patient's germline with subsequent filtering of variants with population frequency >0.01. The results of two pipeline runs are subsequently merged and analysed together in the annotation and interpretation workflow. The germline samples that were submitted as FIBROBLAST where TINC could not be estimated are analysed through the routine tumour-normal workflow.

Haematological cases that were interpreted through TINC workflow are presented on the Interpretation Portal with an interpretation flag of TINC HIGH, TINC LOW or TINC ERROR and corresponding disclaimer is added to WGS results HTML. The TINC workflow is not executed for solid cancers though disclaimers are show on HTML.

To aid interpretation, variants are presented in the WGA results HTML with their origin stated as:

- SOMATIC indicates that the variant has been confidently detected in the standard paired analysis of tumour and normal samples with germline subtraction.

- UNCERTAIN indicates that the variant has been detected only in the analysis that didn't include patient's germline and may represent variants that were subtracted in the paired analysis due to either TINC or presence in the germline. UNCERTAIN variants with population frequency >0.01 are not shown.

- UNCERTAIN (AF>0.001) indicates that the variant has been detected only in analysis that didn't include patient's germline and may represent variants that were subtracted in the paired analysis due to either TINC or presence in the germline, with a population allele frequency between 0.01 and 0.001. These variants are more likely to represent rare germline variants.

- N/A indicates that the variant analysis involved patient's germline sample, but no complete subtraction with the germline has been conducted for this variant; this is the case for all CNVs.

If a sample pair has low level of TINC (between 1% and 5%), both Preliminary and Supplementary WGA result HTML files will be generated, with the circos plot, signature decomposition, tumour mutational burden and somatic variant VAF distribution calculated based on variants detected in the matched germline analysis only.

If a sample pair was reported as having high TINC (over 5%), or if TINC level could not be reliably estimated, only a Preliminary WGA results HTML will be generated. Pan-genomic analysis (circos plot, signature decomposition, tumour mutational burden and somatic variant VAF distribution) and Domain 3 variants are not presented in HTML due to uncertainty of origin for large fraction of variants (can be found in CSV file).

3.4 Tumour only variant detection

For submissions where the germline sample is not submitted with sample for haematological tumour (which are a part of the Tumour First – Germline Later workflow), small variants and structural variants are analysed without subtracting patient's germline variants. The resulting variant lists are filtered to remove small and structural variants with population frequency >0.01. Due to the large number of detected variants, the sensitivity of identifying variants with VAF under 10% may be reduced.

Copy number variants are discovered with Dragen using a down-sampled normal variant aggregate from the gnomAD genomes database instead of patient's own germline variant calls.

To aid interpretation, variants are presented in the WGA results HTML with their origin stated as **UNCERTAIN** (or **UNCERTAIN (AF>0.001)** if a variant has population germline allele frequency between 0.001 and 0.01).

Due to the absence of a germline sample any detected variants may be somatic or germline. It is advised to consider the VAF of a variant and its consequence type to identify variants which are more likely to be germline. If a sample has been taken post haematopoietic stem cell transplant, please be aware that any variant detected could be somatic OR germline, host OR donor.

For tumour only samples, only a Preliminary WGA results HTML will be generated. Pan-genomic analysis (circos plot, signature decomposition, tumour mutational burden and somatic variant VAF distribution) and Domain 3 variants are not presented due to uncertainty of origin of the variants.

Please note that tumour only analysis does not currently fall within the scope of Genomics England's ISO 15189 accreditation.

4 Somatic variant interpretation

4.1 Canonical Transcripts and Regions of Interest

The pipeline reports variants in the pre-selected lists of:

- Canonical transcripts
- Regions of interest
    - specified in the National Genomic Test Directory e.g. IGH
    - included to enhance fusion detection, e.g. GATA2 enhancer region

The list of canonical transcripts and regions of interest for each release is published on NHS Futures as the Cancer Analysis Additional Information document.

4.1.1 Canonical Transcripts

We use Matched Annotation from NCBI and EMBL-EBI (MANE) as the source of canonical transcripts for the majority of protein-coding genes. The "LIST_OF_CANONICAL_TRANSCRIPTS" sheet in Cancer Analysis Additional Information document on NHS Futures lists the transcripts used, with a "source" column indicating the origin of each transcript.

The canonical transcript set is comprised of the following:

- MANE v1.0 transcripts: **19,062** Select and **58** Plus Clinical (Ensembl and RefSeq IDs)
  - Source: MANE Select and MANE Plus Clinical
- Ensembl v107 transcripts: **150** Supplementary transcripts (Ensembl IDs only)
  - **129 Protein-coding genes without a MANE transcript**:
    - Some protein-coding genes do not have a MANE transcript, either because of insufficient review, or because they are out-of-scope for MANE (more details can be found here). This includes the Domain 1/2 genes RUNX1T1, MUC1, and PBRM1. To ensure that we continue to report variants in these genes, canonical Ensembl v107 transcripts are included in the canonical transcript list (more details can be found here).
    - Source: Ensembl v107 Canonical
  - **5 Non-protein-coding genes***:*
    - As MANE is a resource for protein-coding genes only, any non-coding genes are not included, thus we also include canonical Ensembl v107 transcripts for the National Genomic Test Directory targets DLEU2, MALAT1, and PVT1 as well as RMRP and TERC. Please note that these are the only non-protein-coding transcripts included in the canonical transcripts list.
    - Source: Genomics England non-protein coding transcripts
  - **8 Transcripts to patch MANE Clinical v1.0***:*
    - We include an additional 8 transcripts to ensure that the clinically relevant transcripts are up to date with the most current version of MANE Clinical (v1.3)
    - Source: Genomics England additional transcripts
  - **8 Transcripts to enhance fusion and hotspot detection***:*
    - We include an additional 8 transcripts to ensure that known COSMIC hotpots and fusions can be annotated. These transcripts are only considered during structural variant tiering.
    - Source: Genomics England additional SV transcripts

4.1.2 Regions of Interest

The additional loci considered come from 2 sources:

- the National Genomic Test Directory (N=8), and

- a list of loci curated by Genomics England to ensure that clinically relevant fusion breakpoints are detected (N=8)

4.2 Small variants

SNVs and small indels are annotated using Cellbase with the Ensembl and COSMIC databases. Cellbase takes advantage of the data integrated in its database to implement a rich and high-performance variant annotator. Variants annotated with the following consequence types in canonical transcripts (see List of canonical transcripts in Cancer Analysis Additional Information document at NHS Futures) 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
SO:0001792	non_coding_transcript_exon_variant (for RNA coding genes only)

Two non-coding variants in the promoter region of the TERT gene are also reported[10].

Variants resulting in a protein change matching the Somatic Variant Interpretation Guidelines (SVIG-UK) canonical variant list are highlighted by a red lightning sign (in HTML) or [SVIG] flag (in CSV table) in the "CDS change and protein change" column. The current SVIG canonical variant list can be found in the Cancer Analysis Additional Information document at NHS Futures. Variants for which the protein change does not match the SVIG canonical variant list but affects the same amino acid (or exon in case of FLT3 ITDs) are highlighted with a grey lightning sign (in HTML) or [SVIG_overlap] flag (in CSV). To improve the detection rate of oncogenic / likely oncogenic variants, the variants resulting in a protein change matching the SVIG canonical variant list are reported even if they were called as non-PASS variants with an additional (NP) flag in the "Predicted consequences" column.

4.2.1 Domain 1 somatic variants

Variants in a virtual panel of potentially actionable genes are reported in Domain 1 (gene list is available in the Cancer Analysis Additional Information document at NHS Futures). Genes for which small variant testing has been indicated in the National Genomic Test Directory for Cancer for patient's clinical indication or corresponding clinical indication group (Adult Solid, Paediatric, Neurological, Haematological, or Sarcoma) are selected as potentially actionable. Small variants found in these potentially actionable genes are prioritised into Domain 1 and clinical-indication-specific genes are annotated with "*" next to the gene name.

4.2.2 Domain 2 somatic variants

Variants in a virtual panel of cancer-related genes are reported in Domain 2 (gene list is available in the Cancer Analysis Additional Information document at NHS Futures). Cancer-related genes are defined as i) genes in which any variants have been causally implicated in cancer, as defined by the Cancer Gene Census or ii) genes for which small variant testing has been indicated in the National Genomic Test Directory for Cancer that aren't already in Domain 1.

4.2.3 Domain 3 somatic variants

Small variants in genes not included in Domains 1 & 2 are reported in Domain 3.

4.3 Copy number and Structural variants

Prioritisation of SVs only considers variants with breakpoints within introns or exons of consensus transcripts, as well as those falling within Regions of Interest (RoI) (see 4.1.2).

For CNVs detected by Dragen, all genes overlapping the CNV region are reported. For deletions and duplications >50 kb detected by Manta, only genes overlapping the predicted breakpoints are reported due to uncertainty of copy number state between breakpoints.

To capture all fusions of interest we consider breakpoint padding for a subset of RoIs, in particular:

- for fusions involving IGH, IGK, IGL, TRA, TRB or TRD we report partner genes 20 kb up or downstream of the second breakpoint
- for fusions involving DUX4_region we report partner genes 1 kb up or downstream of the second breakpoint

To minimise the number of uninformative SVs and CNVs reported, we limit the types of prioritised RoI variants such that:

- SVs or CNVs where both breakpoints fall into the same RoI are not reported for a subset of RoI
    - For full details, see 'filter_if_both_bp_in_locus' column in the REGIONS_OF_INTEREST sheet in the Cancer Analysis Additional Information document at NHS Futures
- CNVs in upstream / downstream / enhancer RoI are not reported
    - For full details, see 'svtypes_considered' column in the REGIONS_OF_INTEREST sheet in the Cancer Analysis Additional Information document at NHS Futures

CNVs and SVs are presented in the WGA results HTML in three domains as described below. In each domain, variants are listed in two tables: one according to chromosome coordinate (non-redundant list) and one according to gene (with multiple entries for CNVs impacting more than one gene).

For each variant in the chromosome-based list, a confidence score or level of support is displayed.

For Dragen calls, "HC" and "LC" indicate high and low confidence variants, corresponding to PASS variants and all other variants, respectively. For Manta, the number of paired reads (PR) and split reads (SR) supporting the variant for both reference and alternate alleles is provided.

Since the algorithms used for copy number and structural variant detection utilise different methodologies, in some cases a given variant can be detected independently by both Dragen and Manta and may be reported more than once in the WGA results HTML. Support from both methods indicates a higher confidence that a given variant is true, and breakpoints predicted by Manta can be used to refine the coordinates and structure of Dragen copy number variants. Lack of support from the second algorithm does not necessarily indicate that a variant is false.

4.3.1 Fusion frame prediction annotation

For structural variants which can potentially produce a gene fusion (i.e., where the breakpoints fall into different transcripts), a prediction of outcome is performed based on the following algorithm.

First, the breakpoint orientations extracted from Manta calls are compared to the transcription direction of the relevant transcripts, and if they do not align, the outcome is defined as "no fusion". If the transcription directions align with the breakpoint orientations, the types of gene regions involved are checked. If a fusion is only possible with a 3'UTR region of the downstream partner, it is considered non-productive and is reported as "no fusion".

If a fusion is only possible with the 5'UTR region of the upstream partner GENE1, it is reported as "GENE1:GENE2 - 5'UTR", as it is unlikely to produce a fusion transcript, but may lead to a change in transcription level of the downstream partner GENE2.

If at least one of the breakpoints fall anywhere but an intron between coding exons – exon, splice region (first or last 8 bases of an intron), 5'UTR of the downstream partner or 3'UTR of the upstream partner – it is reported as "GENE1:GENE2 – ambiguous", as it is very hard to predict exon skipping or cryptic splice donor/acceptor site utilisation.

If both breakpoints fall into coding introns, we perform frame comparison. If the reading frames of the last present exon of the upstream partner aligns with the reading frame of the first present exon of the downstream partner, the fusion is reported as "GENE1:GENE2 – in-frame" and is likely to be productive. Otherwise, it is reported as "GENE1:GENE2 – out-of-frame" and is unlikely to yield a productive fusion transcript. Please note that the frame prediction is not reliable if any of the breakpoints falls into a splice site (defined as the first 8 or the last 8 bases of an intron between coding exons), therefore the fusions where at least one of the impacted regions is a splice site are considered as having an ambiguous outcome.

Each of the possible outcomes is summarised in the table below.

Outcome	Explanation
No Fusion	Transcription directions do not align with breakpoint orientations, or the fusion involves 3'UTR of the downstream partner; no productive fusion transcript
5'UTR	Fusion involves 5' UTR of the upstream partner; can result in the change of expression, but no fusion transcript
Ambiguous	Cannot predict the frame in the following cases: • fusion involves 5' UTR of the downstream partner • fusion involves 3' UTR of the upstream partner (see e.g. BCOR-CCNB3, PMID 22387997) • fusion involves coding exons (can't predict exon skipping or cryptic splice sites with confidence) • fusion involves splice sites or splice regions (first 8 or last 8 bases of an intron separating two coding exons)

	These cases can potentially result in productive fusion.
In-frame	Fusion between coding introns where the reading frames are aligned
Out-of-frame	Fusion between coding introns where the reading frames are misaligned

The frame prediction annotation appears in the "variant type" column of the structural variants table in the WGA results HTML and should be subject to critical review.

4.3.2 Domain 1 somatic SV/CNVs

CNVs or SVs with breakpoints that overlap genes currently ascribed potential actionability are presented in Domain 1 (gene lists are available in the Cancer Analysis Additional Information document at NHS Futures). Genes for which CNV or SV testing has been indicated in the National Genomic Test Directory for Cancer for patient's clinical indication or corresponding clinical indication group (Adult Solid, Paediatric, Neurological, Haematological, or Sarcoma) are selected as potentially actionable. CNVs and SVs with breakpoints that overlap potentially actionable genes are prioritised into Domain 1 and clinical-indication-specific genes are annotated with "*" next to the gene name.

To improve annotation for regions in which fusion breakpoints are known to occur in non-coding regions, additional regions have been added to Domain 1, see Region of Interest sheet in the Cancer Analysis Additional Information document at NHS Futures.

Alternative transcripts in addition to the canonical ones are also included for multiple genes (see "Source" column in List of Canonical Transcripts in the Cancer Analysis Additional Information document at NHS Futures). For example, additional *ABL1* and *STAT6* transcript are included to ensure the correct detection and reporting of BCR-ABL1 fusions in haematological malignancies and NAB2-STAT6 fusions in sarcomas.

FLT3 ITDs are only reported if they are supported by 3 or more reads, and if they are annotated as in-frame insertions or splice donor/acceptor/region variants. To reduce the number of irrelevant findings reported by *FLT3* ITD caller, the variants with low (3-4 reads) support will be omitted if there are other FLT3 ITDs called with support of at least 5 reads. Occasionally, *FLT3* ITD variant caller may have difficulties identifying the precise location of an ITD and would report multiple overlapping insertions of the same size; in this case, it is recommended to study the variant appearance in IGV to assess its location and confidence.

4.3.3 Domain 2 somatic SV/CNVs

CNVs or SVs with breakpoints that overlap genes in a virtual panel of cancer-related genes (available in the Cancer Analysis Additional Information document at NHS Futures) are reported in Domain 2. Cancer-related genes are defined as i) genes in which any variants have been causally implicated in cancer, as defined by the Cancer Gene Census or ii) genes for which SV or CNV testing has been indicated in the National Genomic Test Directory for Cancer that aren't already in Domain 1.

4.3.4 Domain 3 somatic SV/CNVs

CNVs or SVs with breakpoints that overlap genes not included in Domains 1 & 2 are shown in Domain 3.

5 Germline variant detection

5.1 Germline small variants

Detection of germline small variants is being performed with the Dragen small variant caller. The Dragen software incorporates inferred sex into variant calling such that the overall ploidy of the X chromosome is considered (with possible values of 1 or 2 copies), and haploid calls are produced where appropriate. Annotation of detected small variants is performed with Cellbase with the Ensembl and ClinVar databases.

5.2 Germline copy number variants

Detection of CNVs is performed using Dragen CNV caller with self-normalisation and the Shifting Levels Models (SLM) segmentation mode. High quality CNVs > 10 kb in size are defined as those detected by Dragen CNV with filter status PASS. CNVs between 2 and 10 kb in size are identified by combining the results of Dragen CNV and Manta SV callers which use read depth and anomalously mapped/split reads, respectively, for detection of copy number variants. CNVs in this range detected by both callers with a minimum reciprocal overlap of 50% and with matching CNV type (deletion or duplication) are deemed to be of high quality. The Genomics England Cancer Interpretation Pipeline currently annotates all high quality CNV >= 2 kb in size. Annotation of detected CNVs is performed with Cellbase with the Ensembl database. To account for imprecise breakpoints, Cellbase annotation of CNVs is performed using the extra padding option set to 1000 bp. Please note that annotation of other types of structural variants is currently not performed in the germline analysis.

6 Germline variant interpretation

6.1 Curated gene panels in PanelApp

Interpretation of small variants is performed to prioritise variants of potential clinical relevance, using genes included in curated gene panels, available in PanelApp.

Genomics England PanelApp is a publicly available database created to enable diagnostic grade virtual gene panels to be reviewed and evaluated by experts in the scientific community. All panels are available to view and download on the user interface, or query via webservices and the API. The diagnostic-grade 'Green' genes (and the associated modes of inheritance for pathogenic variants) in virtual gene panels are used to direct the interpretation of germline variants. For details on how gene panels are defined and how to use PanelApp, refer to the latest version of the PanelApp handbook. Signed-off versions of the virtual gene panels used for analysis are available directly at https://nhsgms-panelapp.genomicsengland.co.uk.

Consensus gene panels are finalised through a review process with a disease specialist test group and only signed-off panels are used for analysis, with the most recent signed-off version at the time of interpretation applied. Signed-off panels and associated versions are available in PanelApp.

There are seven applicable germline gene panels in the GMS:

Panel Name	Panel Link
Sarcoma susceptibility	https://panelapp.genomicsengland.co.uk/panels/734/
Childhood solid tumours	https://panelapp.genomicsengland.co.uk/panels/243/
Adult solid tumours cancer susceptibility	https://panelapp.genomicsengland.co.uk/panels/245/
Haematological malignancies cancer susceptibility	https://panelapp.genomicsengland.co.uk/panels/59/
Ovarian cancer pertinent cancer susceptibility	https://panelapp.genomicsengland.co.uk/panels/117/
Breast cancer pertinent cancer susceptibility	https://panelapp.genomicsengland.co.uk/panels/55/
Brain cancer pertinent cancer susceptibility	https://panelapp.genomicsengland.co.uk/panels/166/

Gene panels are applied according to the clinical indication and age of diagnosis using the following rules:

Clinical indication	Age group	Panel(s) for Tier 1 variants
Sarcoma	Childhood	Sarcoma susceptibility Childhood solid tumours
	Adult	Sarcoma susceptibility
Haematological Tumours	Childhood	Haematological malignancies cancer susceptibility Childhood solid tumours
	Adult	Haematological malignancies cancer susceptibility
Paediatric Tumours	Childhood	Childhood solid tumours
	Adult	Childhood solid tumours
Solid Tumours – not high-grade serous ovarian cancer or triple negative breast cancer	Childhood	Childhood solid tumours Adult solid tumours cancer susceptibility
	Adult	Adult solid tumours cancer susceptibility
High-grade Serous Ovarian Cancer	Childhood	Ovarian cancer pertinent cancer susceptibility Childhood solid tumours
	Adult	Ovarian cancer pertinent cancer susceptibility
Triple Negative Breast Cancer	Childhood	Breast cancer pertinent cancer susceptibility Childhood solid tumours
	Adult	Breast cancer pertinent cancer susceptibility
	Childhood	Brain cancer pertinent cancer susceptibility

Clinical indication	Age group	Panel(s) for Tier 1 variants
Neurological Tumours		Childhood solid tumours
	Adult	Brain cancer pertinent cancer susceptibility

Clinical indication	Age group	Panels for Tier 3 variants
Sarcoma	Childhood	Sarcoma susceptibility Childhood solid tumours Adult solid tumours cancer susceptibility
	Adult	Sarcoma susceptibility Adult solid tumours cancer susceptibility
Haematological Tumours	Childhood	Haematological malignancies cancer susceptibility Childhood solid tumours Adult solid tumours cancer susceptibility
	Adult	Haematological malignancies cancer susceptibility Adult solid tumours cancer susceptibility Childhood solid tumours
Paediatric Tumours	Childhood	Childhood solid tumours Adult solid tumours cancer susceptibility
	Adult	Childhood solid tumours Adult solid tumours cancer susceptibility
Solid Tumours – not high-grade serous ovarian cancer or triple negative breast cancer	Childhood	Childhood solid tumours Adult solid tumours cancer susceptibility
	Adult	Adult solid tumours cancer susceptibility
High-grade Serous Ovarian Cancer	Childhood	Ovarian cancer pertinent cancer susceptibility Childhood solid tumours Adult solid tumours cancer susceptibility
	Adult	Ovarian cancer pertinent cancer susceptibility Adult solid tumours cancer susceptibility
Triple Negative Breast Cancer	Childhood	Breast cancer pertinent cancer susceptibility Childhood solid tumours Adult solid tumours cancer susceptibility
	Adult	Breast cancer pertinent cancer susceptibility Adult solid tumours cancer susceptibility
Neurological Tumours	Childhood	Brain cancer pertinent cancer susceptibility Childhood solid tumours Adult solid tumours cancer susceptibility
	Adult	Brain cancer pertinent cancer susceptibility Adult solid tumours cancer susceptibility

Childhood panels are applied when the year of birth and year of diagnosis provided in the test order indicate that the patient was up to and including 25 years of age in the year of diagnosis. If the patient was 26 years or older at the beginning of the year of diagnosis, adult panels are applied.

The panels applied to prioritise Tier 1 and Tier 3 variants are indicated in the WGA results HTML. Variants detected in these genes categorised as being in Tier 1 or Tier 3 (as described below) are presented. Only genes with a high level of evidence for an association with the relevant cancer type are used in variant interpretation (Green Genes in PanelApp panels).

6.2 Germline small variants

ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. However, ClinVar neither curates content nor modifies interpretations independent of an explicit submission. ClinVar reports the level of review supporting the assertion of clinical significance for an individual variant as a review status, and a number of gold stars in assigned accordingly. Further details about the review status provided in the ClinVar database.

ClinVar review status, where available, is displayed in the WGA results HTML for germline variants. However, differences may be observed between the status displayed in the WGA results HTML and the most recent ClinVar page where the review status has been updated since the fixed ClinVar release used for interpretation.

In this analysis multi-nucleotide variants (MNVs) can be reported as multiple consecutive SNVs and/or indels and therefore the potential protein change may require correction. For some complex germline MNVs, annotation with variants described in the ClinVar database may not be correct.

6.2.1 Tier 1

Analysis for pertinent germline findings is performed to detect pathogenic or likely pathogenic variants conferring susceptibility to the relevant clinical indication using a tumour-type specific panel.

Variants are reported in Tier 1 according to the following criteria:

i. predicted protein truncating variants 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)

Variants near the 3’ end of the gene should be carefully evaluated as truncation near the C-terminal end of the protein may or may not impair function. Such variants are flagged with (T) in the WGA results HTML.

ii. variants listed in ClinVar as pathogenic or likely pathogenic (with a rating of at least two stars).

For genes with a biallelic mode of inheritance (as documented in PanelApp), only homozygous or potential compound heterozygous variants are reported in Tier 1. A single heterozygous variant in a gene with a biallelic mode of inheritance that satisfies the criteria for Tier 1 inclusion would be presented in Tier 3.

Please note, for some clinical indications where multiple panels are considered during prioritisation, it’s possible to prioritise the same variant into Tier1 and Tier3 simultaneously, for example if a biallelic mode of inheritance is only required for one of the panels.

Clinical evaluation of variant pathogenicity should be performed locally. If a variant is deemed relevant, it is recommended that the variant is reviewed using the Integrative Genome Viewer (IGV) and assessed via ACMG criteria.

6.2.2 Tier 3

Variants are prioritised to Tier 3 using a broad gene panel(s) spanning cancer susceptibility genes in addition to the tumour-type specific panel. Variants of the consequence types listed in section 4.1 above are included, where the frequency of the variant in an internal Genomics England dataset of >6,000 unrelated individuals is <0.5% (for dominantly-acting genes) and <2% (for recessively acting genes), unless the variant is listed in ClinVar as benign or likely benign with a rating of at least two stars. In the case of susceptibility genes or variants less well reported in ClinVar, bone fide pathogenic missense/splicing variants may not have achieved two-star review status and will be included in Tier 3. Variants in genes on the germline panel for the relevant tumour type are placed at the top of the list and marked with asterisk.

6.2.3 Prioritising potential germline variants in tumour only analyses

As the variant calling results for tumour only analyses are likely to contain rare germline variants private to the patient as well as their somatic variants, we apply germline prioritisation algorithm to all variants detected in tumour only analyses with internal Genomics England population germline allele frequency under 2%. Currently, we consider variants with VAF under 70% to appear as heterozygous, and variants with VAF > 70% as homozygous to enable germline variant prioritisation. Should a variant be classified as Tier 1 or Tier 3, it will have this information recorded in the "Interpretation as germline" field in the WGA results.

Please note that tumour only analysis does not currently fall within the scope of Genomics England's ISO 15189 accreditation.

6.3 Germline copy number variants

Germline copy number variants are prioritised using genes and regions in curated gene panels available in PanelApp (as described in section 6.1). To account for imprecise breakpoints, Cellbase annotation of CNVs is performed with the extra padding option set to 1000 bp. All PASS CNVs that fall within a 'green' pathogenic region or gene on one of the applied PanelApp panels are considered for tiering. Please note that in contrast to small germline tiering, the mode of inheritance is currently not considered in CNV tiering.

6.3.1 Tier 1

Germline CNV prioritisation is performed to detect pathogenic or likely pathogenic variants conferring susceptibility to the relevant clinical indication using a tumour-type specific panel. A CNV is assigned Tier 1 if it satisfies the following criteria:

- The CNV overlaps a pathogenic region in a Tier 1 panel applied in the analysis, the overlap is above the threshold defined in PanelApp for that region, and the variant type matches (i.e. LOSS or GAIN) that of the region in PanelApp;

OR

- The CNV overlaps with a 'green' gene in a Tier 1 panel applied in the analysis and impacts a coding sequence (as defined in the table below). There is no minimal overlap thresholds and variant type (i.e. LOSS or GAIN) is not considered.

SO accession	SO term	Impacted transcript region
SO:0001893	transcript_ablation	full transcript
SO:0001889	transcript_amplification	full transcript
SO:0001580	coding_sequence_variant	partial coding sequence
SO:0001792	non_coding_transcript_exon_variant	partial non-coding exon sequence (RNA genes only)

Variants overlapping the 3' end of the gene should be carefully evaluated as truncation near the C-terminal end of the protein may or may not impair function. Such variants are flagged with (T) in the WGA results HTML.

Clinical evaluation of variant pathogenicity should be performed locally. If a variant is deemed relevant, it is recommended that the variant is reviewed using the Integrative Genome Viewer (IGV) and assessed via ACMG criteria.

6.3.2 Tier 3

Germline CNVs are prioritised to Tier 3 using broader gene panel(s) spanning cancer susceptibility genes. A CNV is assigned Tier 3 if it has not already been tiered as Tier 1 and it satisfies the following criteria:

- The CNV overlaps a pathogenic region in a Tier 3 panel applied in the analysis, the overlap is above the threshold defined in PanelApp for that region, and the variant type matches (i.e. LOSS or GAIN) that of the region in PanelApp;

OR

- The CNV overlaps with a 'green' gene in a Tier 3 panel applied in the analysis and impacts a coding sequence (as defined in the table below). There is no minimal overlap thresholds and variant type (i.e. LOSS or GAIN) is not considered.

6.3.3 CNV frequency annotation

Several factors complicate the assessment of allele frequencies for copy number variants:

- The breakpoints of CNV calls based on sequence coverage are imprecise and therefore the same variant can have different breakpoint coordinates in different individuals.
- Large CNVs can be reported as several separate calls (i.e. fragmented calls). This is often due to a copy number change within the region of a large CNV, for example, due to a smaller nested CNV or a complex structural rearrangement.
- Distinguishing between different combinations of alleles that can give rise to the same diploid copy number is challenging. For example, a copy number of 3 could be the result of a tandem duplication with 2 copies on one allele and a single copy on the other allele, or two single copy alleles with an additional copy elsewhere in the genome.
- It is difficult to make an accurate copy number inference for gain variants with more than 3 copies.

Due to the above issues there is no single perfect method to calculate allele frequencies for CNVs. Therefore, we present two different calculations. CNV frequencies were calculated using 5,757 germline samples from unrelated individuals (participants in the Cancer program of the 100,000 Genomes Project and the COVID-19 research project).

Reciprocal overlap method

CNV population frequency defined using an 80% reciprocal overlap threshold. A limitation of this method is that the frequency may be inaccurate in the event of CNV fragmentation, i.e. fragmented calls can inappropriately appear to be rare.

Area under the curve method

In this method, CNV calls from the 5,757 reference samples are combined. Each base in each of the sampled genomes is annotated with the number of chromosomes for which there is an overlapping CNV. Then the area under the curve for each CNV detected in any patient is calculated, considering both the number of bases and the number of chromosomes in which a CNV is found in the reference dataset. The frequency is then weighted by the maximum possible area. (i.e. an allele frequency of 1 is equivalent to all reference samples having a CNV covering all bases of the patient CNV).

An advantage of this method is that it is robust to CNV fragmentation. A limitation is that we do not know whether the underlying frequency track frequency distribution results from calls of similar size to that detected in the patient, or smaller overlapping CNVs detected in different individuals. If a CNV overlaps two high frequency regions (e.g. at each end) separated by a low frequency region, the overall area under the curve for the region may not be representative of the individual regions, and in particular the contribution of high frequency regions could mask the existence of the low frequency region.

For LOSS variants, allele frequencies are calculated and reported. For GAIN variants, due to difficulties in determining the exact copy number and defining the alleles in all individuals, the proportion of individuals with any GAIN call is calculated and reported, not taking copy number into account.

7 Somatic mutation prevalence (global mutation burden)

We display the tumour mutational burden (TMB) for the patient plotted against the range of TMB values for the respective tumour type and alongside different tumour types for which samples have been sequenced previously. TMB is calculated as total number of small somatic variants (SNVs and indels) in Domains 1-3 per Mb of coding sequence (total 33.2 Mb). Small variants in Domains 1-3 flagged with NP, N, PON, GG, GE, R, SR are removed; indels in homopolymer runs (H) are retained. In the case of very low tumour mutation burden such that no such variants are present, the TMB of the patient is not displayed on the TMB plot provided in the WGA results HTML. Based on analysis of technical replicates for five tumour samples the 95% confidence interval for the TMB calculation is estimated as +/- 5.7%. 95% normal confidence interval is calculated using mean and variance with Bessel correction of absolute deviations from pair-wise TMB averages.

8 Mutational signature analysis

Analysis of large sequencing datasets (10,952 exomes and 1,048 whole-genomes from 40 distinct disease types) has allowed patterns of relative contextual frequencies of different SNVs to be grouped into specific mutational signatures. Using mathematical methods (decomposition by non-negative least squares) the contribution of each of these signatures to the overall mutation burden observed in a tumour can be derived. Further details of the 30 different mutational signatures used for this analysis, their prevalence in different disease types and proposed aetiology can be found at Mutational Signatures (v2 - March 2015). Signatures that contribute < 5% of the overall mutation burden are not reported.

Please be aware that the non-negative least squares fitting tends to over-fit samples by adding many signatures into a single sample. Further, the method tends to favour flatter signatures (i.e., HRD signature 3) and add them incorrectly to samples. The above is especially misleading for samples with low TMB. The quality of fit can be assessed by checking the residual sum of squares (RSS) reported alongside with the signature decomposition. A low RSS value indicates a good fit of the reconstructed profile to the original data, while a larger RSS value suggests a poor fit.

PLEASE NOTE: this feature is not currently UKAS accredited to ISO 15189.

9 Minimal Residual Disease assay targets

Assessment of minimal residual disease (MRD) is essential for risk stratification for patients with haematological cancers. To identify targets suitable for MRD monitoring, reads aligning to Ig/TCR loci are extracted from WGS and analysed using Vidjil algorithm[11] for high-throughput analysis of V(D)J junctions. Vidjil is optimised for the analysis of amplicon-based multiplex PCR assay where reads for the same target are expected to start and end at the same genomic positions. Therefore, we adopted reads re-alignment to account for relative shift of reads that support the same target. The following targets were removed to reduce false positive rate:

- target sequence contains homopolymers of at least 17 nucleotides
- target sequence contains ambiguous basecalls (N nucleotides)
- target sequence contains low complexity sequences that utilise only three out of four nucleotide types
- target sequence contains non-recombined germline sequence as assigned by Vidjil
- Vidjil targets that are initially supported by a single read and then clustered with more than 50 additional reads during the re-alignment step. Investigation showed these were due to V(D)J recombinations inserted into the germline.

Primer selection is performed by random forest model that was trained on a dataset of experimentally validated primers and included the following features: clone size, target locus, length of insertions and deletions, primer length and GC content. We use 95% percentiles to set maximum overlaps of V genes and forward primers (17 nucleotides), V genes and reverse primers (7 nucleotides), J genes and forward primers (7 nucleotides) and J genes and reverse primers (16 nucleotides). Results are presented in an excel spreadsheet containing targets with alignments to V(D)J segments, candidate primers, and supporting reads for each target. Target / primer combinations are prioritised based on primer score and number of reads supporting the target. Targets with fewer than 6 supporting reads are not reported. Targets with 6-11 supporting reads are less likely to result in validated assays and are therefore deprioritised. Primer scores (0-1) are provided as preliminary performance indicators only. Primer sequences frequently require optimization despite favourable computational scores. Validation of all targets and primers via qPCR or dPCR is required before use in an MRD assay.

Information on recurrent ALL and AML fusions/rearrangements that were identified in WGS analysis is also shown in the same excel spreadsheet to facilitate development of MRD assay. The full list of examined fusions/rearrangements is included in the spreadsheet.

PLEASE NOTE: this feature is not currently UKAS accredited to ISO 15189.

10 Depth of coverage, B-allele frequency, and absolute allele counts

Depth of coverage, B-allele frequency, and absolute allele counts plots are included in the Supplementary WGA results HTML (and in Preliminarily HTML for cases with high TINC, cases where TINC can't be estimated, and Tumour-only cases) to aid evaluation of predicted CNV profiles against the observed genomic data. As the observed data is summarised across 500kb windows, these plots are only suitable for evaluation of large CNVs, e.g. loss of a chromosomal arm.

The depth of coverage and B-allele frequency plots show the observed data (dots are coloured according to predicted copy number state) as well as the predicted CNVs calculated by Dragen (orange lines). Overlap between observed and calculated data supports accuracy of CNV calling.

The absolute allele counts plots differentiate between clonal and sub-clonal calls as follows:

- For clonal calls, we show the total CN (solid line) and minor copy number (MCN) (dotted line) for each segment (coloured according to CNV group, e.g. pink for LOSS).
- For sub-clonal calls, we show the total CN (solid line) in black, and CNF (solid line) and minor copy number fraction (MCNF) (dotted line) for each segment (coloured according to CNV group). Differences between CN and CNF are highlighted in yellow.

Please see Appendix D for some examples.

Please note, due to the lack of minor allele count for germline, we cannot identify germline LOHs from the CNV data alone. However, the user can use the B-allele Frequency plot to identify large ROHs from the observed data directly. We also introduce an additional category for recurrent or non-PASS germline CNVs to highlight changes with lower actionability or confidence, respectively. The germline plot for absolute allele counts only shows total copy number as Dragen doesn't provide estimates of minor copy number. Please refer to Appendix D for examples of these plots and their interpretation.

11 Explanation of fields in WGA results HTML

Two WGA results HTML files as well as a machine-readable lists of identified variants are provided in the Interpretation Portal:

The Preliminarily WGA results HTML includes:

- Somatic small variants in Domains 1 and 2
- Somatic fusions/rearrangements and copy number aberrations in Domains 1 and 2
- Pertinent germline findings (small variants and CNVs) in Tiers 1 and 3
- CNV plots (for cases where Supplementary WGA is not provided)

- Additional tables for small/large aberrations (for cases where Supplementary WGA is not provided)

The Supplementary WGA results HTML contains additional information for:

- Somatic small variants in Domain 3
- Somatic fusions/rearrangements and copy number aberrations in Domain 3
- Pan-genomic analysis (circos plot, signature decomposition, tumour mutational burden and somatic variant VAF distribution)

11.1 Sample attributes

The following characteristics are reported in the WGA results HTML:

Attribute	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 by Conpair and samples with contamination <1% are considered as PASS.
Calculated Overall Ploidy	Mean copy number across all bases in autosomal chromosomes, estimated by Dragen. This would be expected to be 2.0 for a diploid genome.
Calculated Chromosome Count	Total number of chromosomes weighted by their copy number (estimated by Dragen)
Calculated Tumour Content	Fraction of cancer cells in a tumour sample calculated by Ccube[3]
Reported Tumour Content	Reported tumour content as estimated in host GLH Pathology lab

11.2 Sequencing and coverage quality metrics

The following metrics are calculated for each sample and used in the assessment of data quality:

Metric	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 value for good quality tumour samples is 98.01% with standard deviation of 0.34%. Median value for germline sample is 98.07% with standard deviation of 0.11%.
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. The median percentage of chimeric DNA fragments in good quality tumour samples is 1.43% with standard deviation of 0.24%.

Metric	Explanation
	The median value for germline samples is 1.28% with standard deviation of 0.33%.
Insert size median, bp	Insert size represents the length of the DNA fragments sequenced. Short fragments could result from DNA fragmentation due to poor sample handling. The median fragment size for good quality tumour samples is 525bp with standard deviation of 22bp. The median value for germline samples is 528bp with standard deviation of 28bp.
Mean genome-wide coverage	Coverage represents the mean number of reads (depth) per base in the reference genome. Coverage is calculated for autosomes only. The median value for good quality tumour samples is 97x with standard deviation of 14x. The median value for germline samples is 42x with standard deviation of 7x.
Evenness of Local Genome Coverage	This metric represents read depth uniformity across the genome. Evenness is calculated as the proportion of 1Mb bp regions ("bins") which differ from their neighboring bin in median coverage depth by less than 5%, and ranges from 0 to 1.
COSMIC content with low coverage	This metric represents the "discoverability" of known somatic mutations. It is calculated as the percentage of hypothetical somatic mutation sites (obtained from COSMIC) with coverage of <30x. Median value for this metric for good quality fresh frozen 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. Expected number of false positive SNVs and indels is discussed in Appendix A.
AT dropout **CG dropout**	These 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 for good quality tumour samples are 1.37% (standard deviation 0.75%) for AT dropout and 2.83% (standard deviation 1.08%) for CG dropout. Median values for germline samples are 2.24% (standard deviation 0.45%) for AT dropout and 1.87% (standard deviation 0.49%) for GC dropout.

NOTE: typical values may be revised as additional data become available.

11.3 Sample and variant quality disclaimers

For samples that do not pass one or more of the quality control checks performed on the genomic data (described in section 2), disclaimers indicating the following maybe displayed in WGA results HTML:

Disclaimer	Description
Low level cross-patient tumour contamination	Cross-patient contamination in a tumour sample between 1% and 2.5%. Sample may have higher incidence of contaminating germline variants reported in the somatic variant calls
Low tumour purity	Calculated tumour content <30% and/or >40% of somatic variants with <6% VAF. Sample may have reduced sensitivity for somatic variants, which may also impact calculation of tumour mutation burden and mutational signatures
Potentially degraded tumour sample / Likely degraded tumour sample	Tumour sample has been assessed by multivariate tumour sample quality QC to be potentially degraded/likely degraded. There is an increased risk of false positive and negative results.
Low germline coverage	Mean germline coverage <15x OR less than 95% of the reference genome is covered with a minimum of 15x. Low germline coverage affects the efficiency of somatic variant detection such that sensitivity is reduced, potentially below 95%. This results in an increased risk of false negative results and tumour mutation burden and mutational signatures are not reliably calculated. Sensitivity and precision of germline variant detection may also be reduced.
Low level cross-patient germline contamination	Cross-patient contamination in a germline sample between 3% and 8%. Germline variants are not reported due to the risk of false positives. Validation experiments show that somatic variant detection is not affected by germline contamination levels between 3-8% therefore somatic variants are reported.
Suspected poor quality germline CNV calls	This sample was reported as having suspected poor quality CNV calls. This means that the sequencing data are not of sufficiently high quality to make reliable CNV calls in this germline sample.
Poor quality germline CNV calls	This sample was reported as having poor quality CNV calls. This means that the sequencing data are not of sufficiently high quality to make reliable CNV calls in this germline sample.
Degenerate diploid	Dragen defaulted to a diploid solution as it could not find a confident model for this sample and was unable to estimate its tumour purity. For these cases, all CNVs will be labelled as low confidence ('LC') and somatic CNV plots will not have an expected (orange) line.
Subclonal calling disabled	A large fraction (>80%) of this genome was labelled as likely subclonal by Dragen (subclonal calling is enabled by default), which is biologically implausible. Therefore somatic CNV calling was performed with subclonal calling disabled for this sample.

Disclaimer	Description
Low tumour in normal contamination	The germline sample for this patient is likely to be contaminated with DNA derived from the tumour. Consequently, the sensitivity of somatic variant detection may be reduced, potentially resulting in an increased risk of false negative findings. To mitigate the potential loss in sensitivity, the results of somatic variant calling with an unmatched germline sample are included in this analysis alongside subtraction with the patient's germline sample.
High tumour in normal contamination	The germline sample for this patient is likely to be contaminated with DNA derived from the tumour. Consequently, the sensitivity of somatic variant detection may be reduced, potentially resulting in an increased risk of false negative findings. To mitigate the potential loss in sensitivity, the results of somatic variant calling with an unmatched germline sample are included in this analysis alongside subtraction with the patient's germline sample.
Tumour in normal contamination estimation not available (haematological malignancies only)	The results of the computational estimation of TINC are not reliable for this patient (which may be due to low tumour content in the tumour sample, or very high tumour contamination in the germline sample). Consequently, TINC cannot be excluded, and the sensitivity of somatic variant detection may be reduced, potentially resulting in an increased risk of false negative findings. To mitigate the potential loss in sensitivity, the results of somatic variant calling with an unmatched germline sample are included in this analysis alongside subtraction with the patient's germline sample.
MRD Ig/TCR targets: single target with high support	Only a single Ig/TCR target with 12 or more supporting reads has been detected. Targets with fewer than 12 reads are less likely to result in validated assays. Any additional targets with at least 6 supporting reads are reported.
MRD Ig/TCR targets: no targets with high support	No target Ig/TCR with 12 or more supporting reads has been detected. Targets with fewer than 12 reads are less likely to result in validated assays. Targets with at least 6 supporting reads are reported.
MRD Ig/TCR targets failure	MRD targets for this sample could not be processed due to an error. A ticket has been raised with the GEL service desk for further investigation.
MRD Ig/TCR targets: no targets	No target Ig/TCR with 6 or more supporting reads has been detected. No Ig/TCR targets have been reported.
Unaccredited sample types	This analysis was performed on a samples type that is not in the scope of accreditation for the Genomics England pipeline. Any finding should be validated before use

11.4 Variant descriptions

Variants presented in WGA results HTML are annotated with the following features:

Variant annotation	Explanation
Small somatic variants	

Variant annotation	Explanation
CDS change	Coding DNA change
Population germline allele frequency	Population germline allele frequencies from two independent datasets are reported: internal Genomics England dataset of >6,000 unrelated individuals and gnomAD v2. '-' denotes absence of the variant in the corresponding database.
VAF (variant allele frequency)	Calculated as alt/(alt + ref) where alt and ref are the number of reads supporting the reference and alternate alleles. Reads with mapping quality <40 and read-pairs with only a single end mapped or with an anomalous insert size are excluded.
Gene mode of action	Classification for the mode of action (oncogene, tumour suppressor or both) associated with the genes. Data extracted from the manually curated list of Cancer Census Genes (v99).
Recruiting Clinical Trials **30 Jan 2023**	Links to potentially recruiting clinical trials at GenomOncology . Genes without a link did not have an associated potentially recruiting trial on 30.01.2023. Note that each gene in a fusion is considered independently.
Structural and Copy Number Variants	
Confidence/support	PR – support for variant from anomalously mapped paired reads for variants called by Manta or JuLI SR – support for variant from split-reads (reads spanning breakpoint) for variants called by Manta or JuLI AD - support for variant from anomalously mapped reads for Pindel calls HC – high confidence Dragen call (CNVs with FILTER=PASS) LC – low confidence Dragen call (CNVs with non-PASS FILTER values) LSC - likely subclonal copy number variant identified by Dragen's subclonal CNV model Please note that to maximise sensitivity for detecting fusions involving *SSX2/4*, different mapping quality thresholds are used with JuLI, and so the number of supporting reads for SS18-SSX2/4 fusions may be higher than for other variants.

Variant type	BND = breakend (translocation) (Manta, JuLI and pelops)
	DEL = deletion (Manta)
	DUP = tandem duplication (Manta)
	GAIN = copy number gain (Dragen)
	INS = insertion (Manta)
	INV = inversion (Manta)
	ITD = internal tandem duplication (Pindel)
	LOH = copy number-neutral loss of heterozygosity (Dragen)
	LOSS = copy number loss (Dragen)

Impacted transcript region	For Manta calls - Location breakpoints within the affected gene (e.g. intron, exon, intergenic region) For Dragen calls - part of the transcript that overlaps with the CNV (e.g. partial coding sequence, full transcript)

Variant annotation	Explanation
Population germline allele frequency	Population germline allele frequency for the breakpoints of a given structural variant based on a panel of normals, which consists of germline structural variants observed in an internal Genomics England dataset of about 6,000 samples from unrelated individuals. If a variant has two breakpoints, maximal value of allele frequency among the two is reported. "-" suggests that this variant was not seen in the Genomics England dataset.
Recruiting Clinical Trials **30 Jan 2023**	Links to potentially recruiting clinical trials at GenomOncology. Genes without a link did not have an associated potentially recruiting trial on 30.01.2023. Note that each gene in a fusion is considered independently.

12 Additional files available on Interpretation Portal

In addition to the Preliminarily and Supplementary WGA results HTML, several other files are available for download in the Interpretation Portal.

These files are summarized in the table below –

File name pattern	Description
<germine_sample_id>.cnv.vcf.gz	Germline CNV VCF
< somatic_dispatched_sample_LSID>_<patient_id>-<somatic_sample_id>-<germline sample_id>- **reported_structural_variants_**xxxxxxx**.csv**	CNV/SV Machine Readable CSV
<somatic_dispatched_sample_LSID>_<patient_id>-<somatic_sample_id>-<germline sample_id>- **reported_variants.csv**	Small Variants Machine Readable CSV
*<somatic_sample_id>-<germline sample_id>.**somatic.CNV.baf.bedgraph.gz***	B-allele frequency in tumour
<somatic_sample_id>.vcf.gz	Annotated somatic small variants VCF gene name, protein change, VAF, etc
*<germline sample_id>-<somatic_sample_id>.**somatic.vcf.gz***	Somatic small variants VCF
*<germline sample_id>-<somatic_sample_id>.**somatic.merged.SV.CNV.vcf.gz***	Somatic SV/CNV VCF
*<somatic_sample_id>.**mrd_with_fusions_excel_report.xlsx***	Minimal Residual Disease Ig/TCR targets

Please note that some of the VCF files in the table above currently reside within the Associated files table and not the VCF files table – this issue will be rectified in an upcoming release, along with a simplification of the file names.

Supporting or Reference Documents

Related documents

1. Cancer analysis additional information (available at NHS Futures)
   - List of canonical transcripts
     - Includes: gene_name, gene_ID, transcript_ID, refseq_transcript_ID, source
     - The "source" column specifies the origin of the transcript. Note that only MANE transcripts will have a RefSeq ID.
   - Regions of Interest:
     - Includes: locus_name, chromosome_38, start_38, end_38
   - Cancer census genes:
     - Includes: gene_name, gene_ID, transcript_ID, role_in_cancer
   - Actionable small variant genes in the National Test Directory
     - Includes: gene_name, gene_ID, transcript_ID, clinical_indication_group, clinical_indication
     - The final columns specify for which clinical indications the transcript would be prioritised into Domain 1.
   - Actionable SV genes in the National Test Directory
     - Includes: gene_name, gene_ID, transcript_ID, clinical_indication_group, clinical_indication
     - The final columns specify for which clinical indications the transcript or locus would be prioritised into Domain 1.
2. Genomics England Interpretation Portal for the NHS Genomic Medicine Service

References

1. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, Boehnke M, Kang HM. Am J Hum Genet. 2012 91(5):839-48

2. Conpair: concordance and contamination estimator for matched tumor-normal pairs. Bergmann EA, Chen BJ, Arora K, Vacic V, Zody MC. Bioinformatics. 2016 32(20):3196-3198

3. Ccube: A fast and robust method for estimating cancer cell fractions. Ke Yuan, Geoff Macintyre, Wei Liu, PCAWG-11 working group, Florian Markowetz. bioRxiv

4. Clinical application of tumour in normal contamination assessment from whole genome sequencing. Jonathan Mitchell, Jack Bartram, Susan Walker, Jane Chalker, Magdalena Zarowiecki, Salvatore Milite, Genomics England Research Consortium, Alona Sosinsky, Giulio Caravagna. bioRxiv

5. Strelka2: fast and accurate calling of germline and somatic variants. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, Saunders CT. Nat Methods. 2018 15(8):591-594

6. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Bioinformatics. 2016 15;32(8):1220-2.

7. Split-Read Indel and Structural Variant Calling Using PINDEL. Ye K, Guo L, Yang X, Lamijer EW, Raine K, Ning Z. Methods Mol Biol. 2018;1833:95-105.

8. Junction Location Identifier (JuLI): Accurate Detection of DNA Fusions in Clinical Sequencing for Precision Oncology. Shin HT, Kim NKD, Yun JW, Lee B, Kyung S, Lee KW, Ryu D, Kim J, Bae JS, Park D, Choi YL, Lee SH, Ahn MJ, Park K, Park WY. J Mol Diagn. 2020; 22(3):304-318

9. Whole genome sequencing provides comprehensive genetic testing in childhood B-cell acute lymphoblastic leukaemia. Sarra L. Ryan, John F. Peden, Zoya Kingsbury, Claire J. Schwab, Terena James, Petri Polonen, Martina Mijuskovic, Jenn Becq, Richard Yim, Ruth E. Cranston, Dale J. Hedges, Kathryn G. Roberts, Charles G. Mullighan, Ajay Vora, Lisa J. Russell, Robert Bain, Anthony V. Moorman, David R. Bentley, Christine J. Harrison & Mark T. Ross. Leukemia 2023;37;518-28

10. Highly recurrent TERT promoter mutations in human melanoma. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Science. 2013;339:957-9.

11. Vidjil: A Web Platform for Analysis of High-Throughput Repertoire Sequencing. Duez M, Giraud M, Herbert R, Rocher T, Salson M, Thonier F. PLoS One. 2016;11.

Appendices

Appendix A – Validation data

Validation of the Cancer Pipeline in accordance with ISO15189:2012 is described in the Pipelines 2.0 Cancer validation report (BIO-VAL-0010 Pipelines 2.0 Cancer validation report October 2019 and VAL-BIO-011 - Alignment and Somatic CNV Calling using DRAGEN v4.0.5). A summary of supplementary validation data is shown below.

Somatic small variant detection

The estimated sensitivity for somatic small variant detection is shown in the Additional Information section at the bottom of the WGA results HTML for the typical (100x) and minimum (70x) levels of genome coverage for tumour genomes. These data were derived by comparison of WGS data with high coverage exome sequencing data, considering variants predicted to be of functional consequence to the protein (that is, those that would be prioritised as being in Domains 1-3 as described in section 3.1) . These estimates represent the minimum expected sensitivity as true variants detected but not passing WGS quality assessments were discounted from calculations and the exome sequencing data were subject to stringent quality filtering.

Additional estimates of the sensitivity and precision of small variant detection across various allelic frequency ranges at both 100x and 70x tumour genome coverage are shown below.

Metric		Mean metric (95% CI)		
	Tumour sample coverage	5-10% VAF	10-15% VAF	15-100% VAF
SNVs				

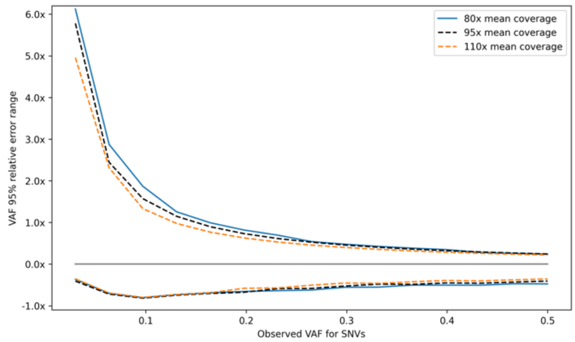
Precision	100x	87% (83-88%)	91% (89-93%)	93% (92-94%)
	70x	82% (79-85%)	91% (89-93%)	93% (92-94%)
Sensitivity	100x	91% (89-93%)	97% (96-99%)	100% (99-100%)
	70x	81% (79-84%)	92% (90-94%)	99% (99-100%)
		INDELs		
Precision	100x	59% (42-75%)	60% (43-77%)	89% (84-93%)
	70x	43% (27-61%)	78% (64-91%)	88% (83-92%)
Sensitivity	100x	86% (67-100%)	92% (84-99%)	100% (99-100%)
	70x	77% (54-98%)	87% (77-96%)	100% (99-100%)

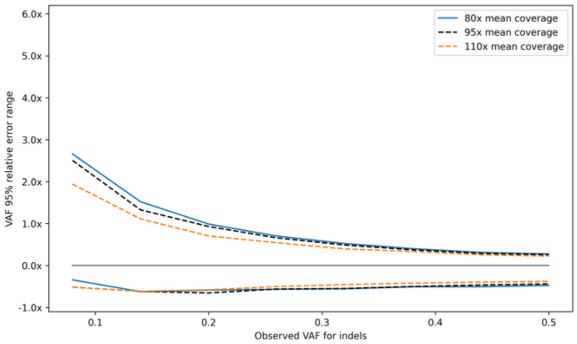
As a further check on sensitivity, two replicates of whole genome sequencing were obtained for three samples. The samples were then pooled and stratified by depth of coverage for high quality reads at the level of individual autosomal variant. Concordance was calculated as the fraction of shared / total variants in each bin. The results are shown below for SNVs and indels:

	Coverage			
	40-80	80-120	120-160	160-300
		SNVs		
VAF 0-10%	0.01	0.90	1.00	0.99
VAF 10-20%	0.81	1.00	1.00	1.00
VAF 20-30%	1.00	1.00	1.00	1.00
VAF 30-40%	0.99	1.00	1.00	1.00
		INDELs		
VAF 0-10%	0.02	0.84	1.00	0.42
VAF 10-20%	0.17	1.00	1.00	1.00
VAF 20-30%	0.80	1.00	1.00	1.00
VAF 30-40%	1.00	1.00	1.00	1.00

Measurement of uncertainty (confidence intervals) for Variant Allele Frequency for somatic small variants

To assess 95% confidence intervals for reported VAF values, three test samples sequenced at high depth in duplicate were selected. Analyses were carried out separately for SNVs and indels. Autosomal chromosomes only were considered. Variants reported in only one replicate were excluded. Figure below shows the 95% confidence intervals normalised to observed VAF for a sequencing technical replicate. The effect of down sampling from 110x to 80x is shown.





The samples were then pooled together, and variants were stratified by depth of coverage with high confidence reads at the level of individual autosomal variant. Confidence intervals at the 95% level for selected VAF levels are reported in the table below for SNVs and indels.

	Coverage			
	40-80	80-120	120-160	160-300
SNVs				
VAF 5%	4-20	3-14	3-12	2-11
VAF 10%	4-26	5-21	5-18	4-17
VAF 20%	9-35	10-30	11-28	11-26
VAF 30%	15-42	17-40	17-38	18-39
VAF 40%	35-66	27-58	34-52	37-52
INDELs				
VAF 5%	6-17	4-15	3-14	6-11
VAF 10%	6-21	5-21	5-21	7-15
VAF 20%	10-31	11-32	12-30	14-25
VAF 30%	20-47	19-41	19-39	24-39
VAF 40%	32-64	23-61	34-52	39-52

Assessment of false positive rate for somatic small variants

To assess the rate of false positive findings in WGS tumour-normal analyses, a normal-normal subtraction experiment where WGS data for "tumour" and "normal" samples are generated from the same germline sample was performed. This experiment is mimicking somatic variant calling with germline subtraction in the Cancer pipeline and is expected to produce no true somatic variants as "tumour" sample does not contain any tumour. Four high-depth Platinum genome sequencing datasets were used, where 40 pairs of 100x and 30x sets were subsampled, mocking matched tumour-normal sequencing experiments. The expectation here, is that variants called from this experiment would represent sequencing artifacts.

There was a median of 60 unflagged small variants observed, with a median of 30 SNVs and 30 INDELs. These findings support the efficiency of filtering steps for false positive small variants.

Variant	Median number of variants (95% CI)	Inter quantile range
PASS indels	30 (29-32)	7.25

PASS SNVs	30.5 (24-34)	18.5

Germline cross-patient contamination

The impact of germline cross-patient contamination of somatic variant detection was assessed using three tumour-normal pairs, with tumour samples spanning a range of levels of genome coverage. Sequencing data for the three tumour samples were artificially contaminated with sequencing reads originating from a fourth germline sample to simulate a range of different levels of contamination. Further assessments were made using additional artificially contaminated tumour-normal pairs for which the tumour genome had high overall ploidy or a low CNV burden and from an individual of non-European ancestry, the results of which supported the conclusions from the initial three pairs.

Somatic small variant detection

For the three test samples selected, a truth set of high confidence variants detected by high-coverage exome sequencing data is available and was used to assess the impact of contamination on somatic variant detection sensitivity and precision. There was no impact of contamination on the sensitivity of somatic variant detection. This is to be expected as it is unlikely that a germline variant from the contaminating sample would overlap with a true somatic variant. Sensitivity (recall) and precision are shown for a range of germline contamination values for all somatic small variants passing basic variant filters on FIGURE 1.

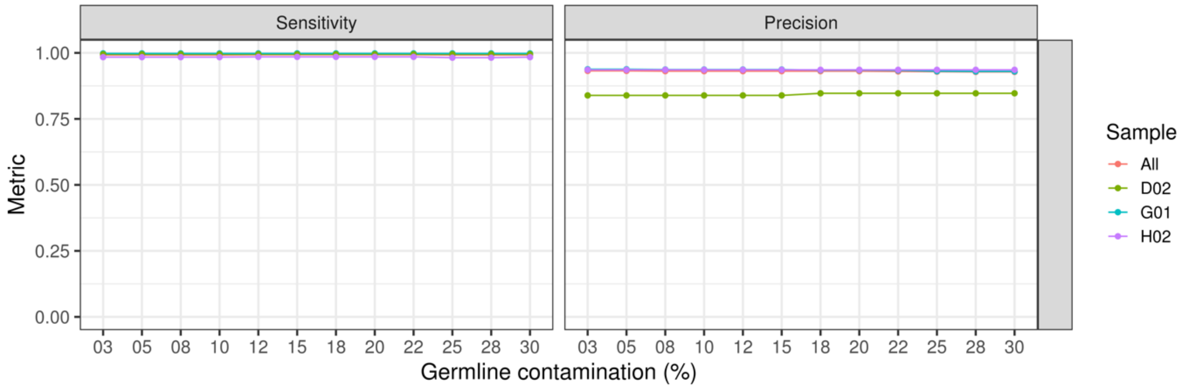


FIGURE 1 SENSITIVITY AND PRECISION OF SOMATIC SMALL VARIANT CALLING IN THE PRESENCE OF GERMLINE SAMPLE CROSS-PATIENT CONTAMINATION FOR THREE SAMPLES WITH DIFFERENT TUMOUR PURITY

Copy number and structural variant detection

To assess the impact of germline contamination on somatic CNV detection, CNVs detected in the tumour-normal pairs with the artificially contaminated samples were compared with the CNVs detected using the corresponding pairs without artificial contamination. Dragen tumour ploidy estimation during somatic CNV calling is not impacted by the presence of contamination in the germline sample. The overall tumour ploidy predicted by Dragen at a range of levels of germline contamination for three tumour-normal pairs is shown in FIGURE 2.

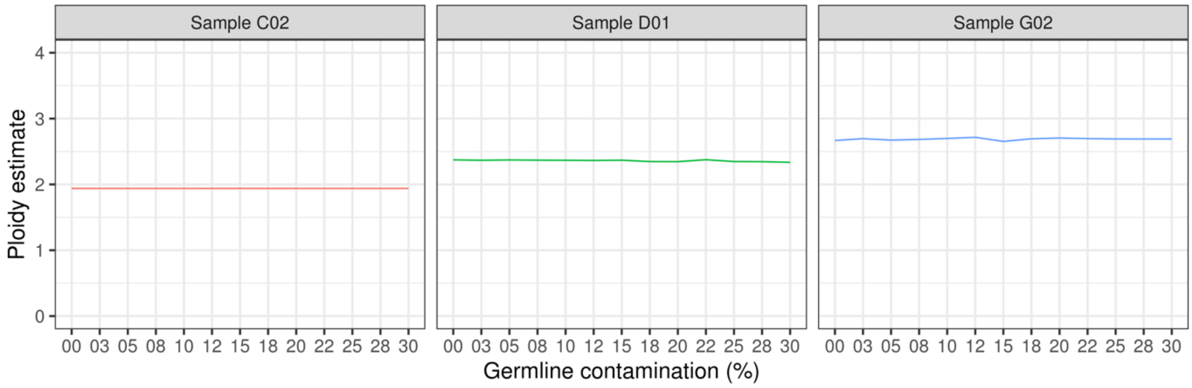


FIGURE 2 - IMPACT OF GERMLINE CROSS-PATIENT CONTAMINATION ON DRAGEN'S TUMOUR PLOIDY ESTIMATION

No impact of germline contamination was observed for the number of structural variants detected by Manta.

Tumour cross-patient contamination

In the event of contamination of a tumour sample with DNA originating from a different patient, germline variants found in the contaminating DNA could be reported as somatic variants. Consequently, in the event of contamination, an increased number of common germline variants may be observed. The impact of cross-patient contamination in tumour samples was evaluated by assessing the fraction of common germline variants detected in the somatic variant set at a range of different contamination levels. An increased number of common germline variants were observed for contamination levels above 2.5%.

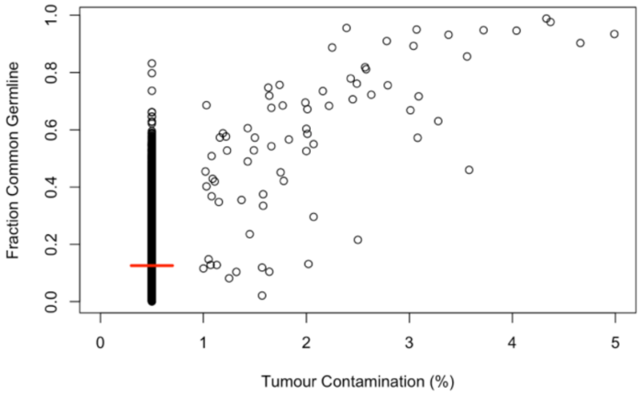


FIGURE 3 - FRACTION OF COMMON GERMLINE VARIANTS DETECTED AS SOMATIC VARIANT AT A RANGE OF LEVELS OF TUMOUR CROSS=PATIENT CONTAMINATION. ALL SAMPLES WITH CONTAMINATION <1% ARE PLOTTED TOGETHER AT 0.5% AND THE MEAN VALUE INDICATED WITH THE RED LINE.

Germline coverage

The impact of genome coverage of the germline sample on the sensitivity of somatic small variant detection was assessed by comparison with a set of high confidence variants detected from high coverage exome sequencing data.

The sensitivity of variant detection was estimated at a range of levels of germline coverage for three samples with varying levels of tumour purity. In each case, the mean coverage for the tumour sample was fixed at 100x. The sensitivity of small somatic variant detection (considering variants predicted to be of functional consequence to the protein with VAF >10% passing all variant quality flags) is <=95% when mean germline coverage is reduced below 15x (data shown in FIGURE 4).

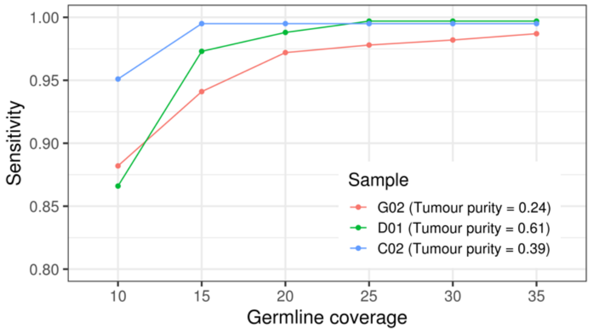


FIGURE 4 - IMPACT OF GERMLINE COVERAGE ON THE SENSITIVITY OF SOMATIC VARIANT DETECTION

The reduction in sensitivity for somatic variant detection at low germline coverage can be attributed to a lower confidence in a somatic variant being a true somatic variant in regions where the germline coverage is low. Consequently, a somatic variant in a region for which the coverage in the germline sample is low is assigned a low-quality score since the probability of the variant being a germline variant that was not detected is higher.

For any tumour-normal pair for which the mean coverage of the germline sample is <15x, a warning will be displayed in the WGA results HTML.

Tumour coverage

The impact of coverage of the tumour sample on the sensitivity of somatic small variant detection was assessed by comparison with a set of high confidence variants detected from high coverage exome sequencing data.

The sensitivity of variant detection was estimated at a range of levels of tumour coverage for three samples with varying levels of tumour purity. In each case, the mean coverage for the germline sample was fixed at 33x. The sensitivity of small somatic variant detection (considering variants predicted to be of functional consequence to the protein with VAF >10% passing all variant quality flags) is above 95% for all values for tumour sample coverage between 70 and 120x.

Tumour purity of the sample impacts sensitivity of calling somatic small variants, where samples with purity >30% have sensitivity >99% and samples with purity <30% have sensitivity >95% (data shown in FIGURE 5).

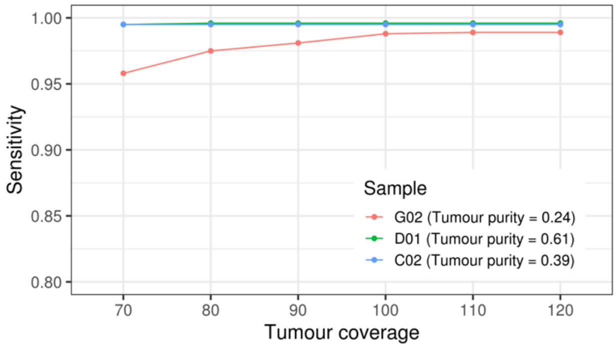


FIGURE 5 - IMPACT OF TUMOUR COVERAGE ON SOMATIC SMALL VARIANT CALLING

Total chromosome count

Dragen does not provide the total chromosome count in the somatic CNV VCF, and so we estimate it based on the Dragen CNV calls. Total chromosome count is calculated as the sum of the mode copy number across all chromosomes. We benchmarked this approach against 24 GMS samples with reported karyotypes and found a strong correlation (n=24; Pearson's correlation = 0.99).

Appendix B – Data presentation in the Interpretation Portal, HTML files and IGV

Interpretation Portal, HTML file and IGV functionality.

The data presented in the WGA result HTML can be used in conjunction with the Integrative Genomics Viewer (IGV) to support data analysis and interpretation. Both the HTML files and IGV viewer are available through the Interpretation Portal.

HTML files can be downloaded from the Interpretation Portal. The coordinates for all variants presented in the Supplementary HTML file are hyperlinks to access the appropriate genomic region using the IGV viewer. After following a link, a login screen for OpenCGA is presented, and after logging in, a list of files available to view in IGV is shown.

A variety of alignment and variant call files are available to view in IGV, of which a summary is shown below. The most relevant files for review are shown in bold.

File	Description
<Germline-sample>.vcf.gz	**Germline small variants (after normalisation)**
<Germline-sample>.repeats.vcf	Germline short tandem repeat (STR) genotypes for select loci detected by ExpansionHunter as part of Dragen
<Germline-sample>.CNV.vcf.gz	Germline copy number variants (CNVs) detected by Dragen CNV
<Germline-sample>_<Tumour-sample>.somatic.CNV.vcf.gz	**Somatic copy number variants detected by Dragen**
<Germline-sample>_<Tumour-sample>.somatic.SV.vcf.gz	**Somatic structural variants detected by Manta**
<Germline-sample>_<Tumour-sample>.somatic.vcf.gz	**Somatic small variants after normalisation**
<Germline-sample>_<Tumour-sample>.somatic.merged.SV.CNV.vcf.gz	Somatic structural and copy number variants in a merged file
<Tumour-sample>.ITD.vcf.gz	Internal tandem duplication genotype at *FLT3* locus
<Tumour-sample>.snv.vcf	Intermediate file from somatic small variant filtering with Panel of Normals
<Tumour-sample>.fisher.snv.vcf.gz	Intermediate file from somatic small variant filtering with Panel of Normals
<Tumour-sample>.fisher.snv.vcf	Intermediate file from somatic small variant filtering with Panel of Normals
<Tumour-sample>.vcf.gz	Somatic small variants detected by Strelka after annotation with quality filters

File	Description
<Germline-sample>.GRCh38DecoyAltHLA_NonN_Regions_autosomes_sex_mt.CHR_full_res.bw	**Germline sample coverage file**
<Germline-sample>.target.counts.bw	Intermediate file from Dragen CNV (germline CNV detection)
<Germline-sample>.cram	**Germline alignment file**
<Tumour-sample>.GRCh38DecoyAltHLA_NonN_Regions_autosomes_sex_mt.CHR_full_res.bw	**Tumour sample coverage file**
<Germline-sample>_<Tumour-sample>.somatic.SV.evidence.normal.bam	**Intermediate file from Manta (germline). Contains reads supporting structural variants.**
<Germline-sample>_<Tumour-sample>.somatic.SV.evidence.tumour.bam	**Intermediate file from Manta (somatic). Contains reads supporting structural variants.**
<Germline-sample>_<Tumour-sample>.somatic.realignment.normal.bam	Intermediate file from Strelka (germline)
<Germline-sample>_<Tumour-sample>.somatic.realignment.tumour.bam	Intermediate file from Strelka (somatic)
<Tumour-sample>.cram	**Tumour sample alignment**

After selecting the appropriate files, data can be viewed using IGV directly in the web browser by clicking “show tracks” or via IGV desktop after downloading a batch script.

Further guidance for using the Interpretation Portal and accessing IGV can be found in the Genomics England Interpretation Portal for the NHS Genomic Medicine.

Interpreting variants using IGV

Variant quality and other characteristics can be visually assessed by viewing genome alignments in IGV.

Typical characteristics of good quality small variants and example sequence data are shown in FIGURE 5 and useful IGV settings for small variant assessment are shown in FIGURE 6.

- Support in tumour
- No support in normal unless TINC
- Mapping quality - bright shade for read
- Basecall quality - bright letter for variant
- No sequencing noise - no ‘Smarties’
- No strand bias - red vs blue reads

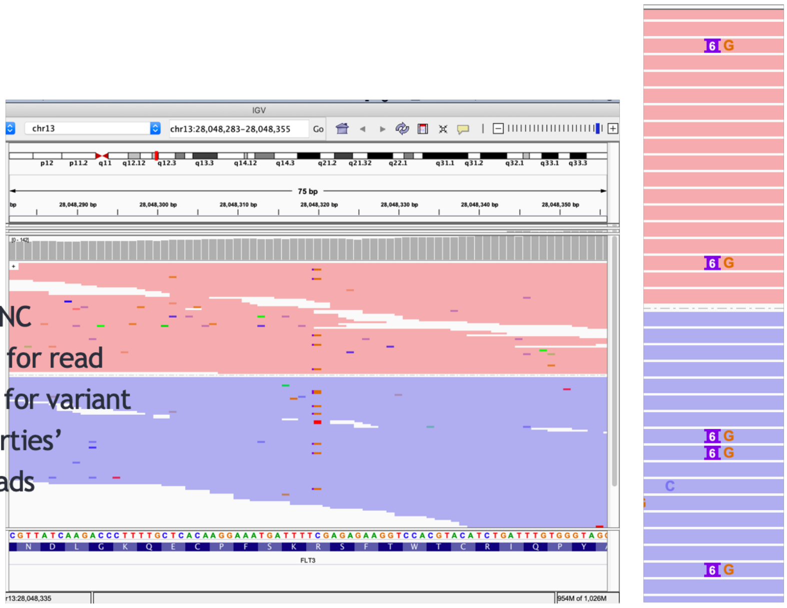


FIGURE 5 - CHARACTERISTICS OF HIGH QUALITY SOMATIC SMALL VARIANTS

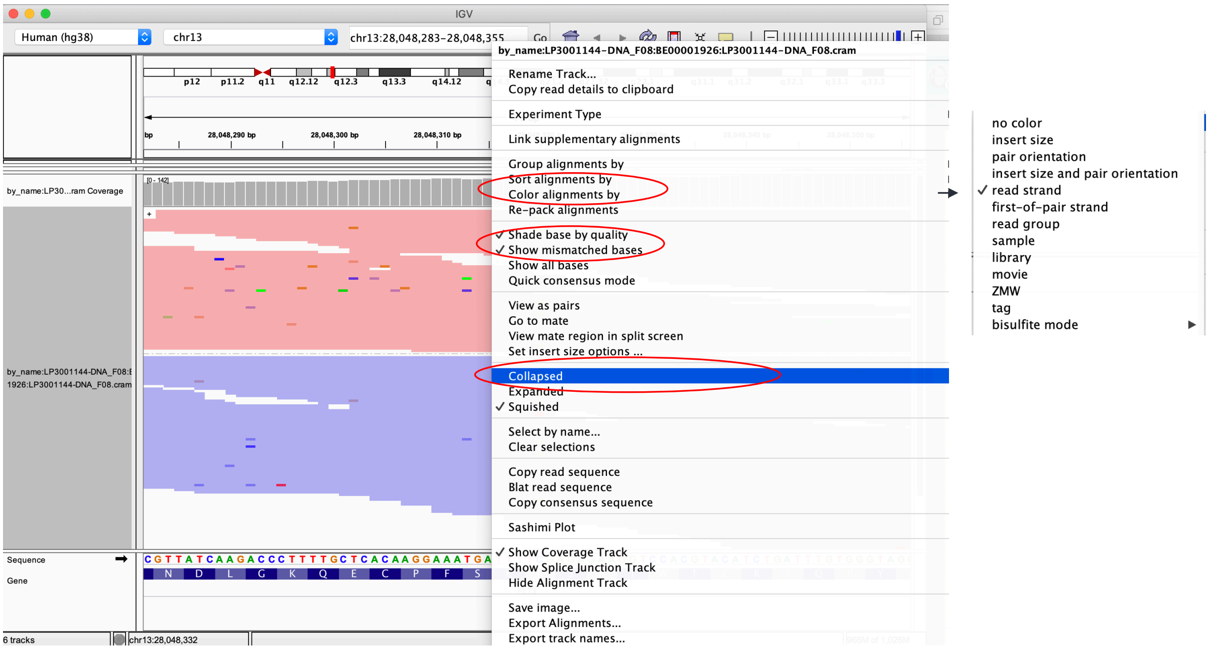
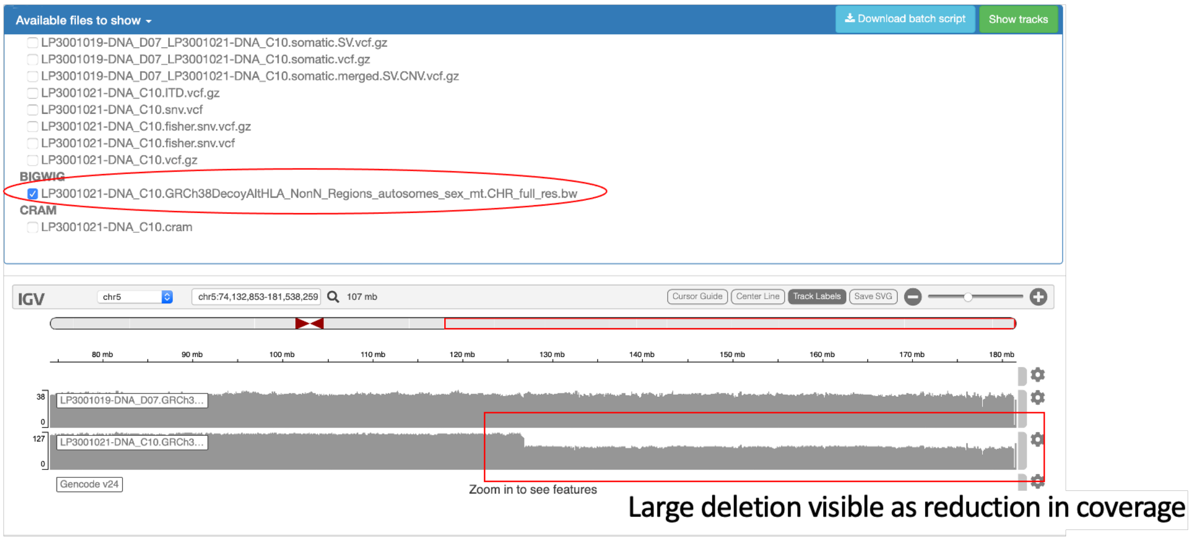


FIGURE 6 - IGV SETTINGS HELP FOR SMALL VARIANT VISUALISATION

Copy number variants can be assessed using coverage profiles (BigWig file), with deletions seen as a reduction in coverage and amplifications as an increase in coverage, as shown in FIGURE 7.



□ LP3001019-DNA_D07_LP3001021-DNA_C10.somatic.SV.vcf.gz
□ LP3001019-DNA_D07_LP3001021-DNA_C10.somatic.vcf.gz
□ LP3001019-DNA_D07_LP3001021-DNA_C10.somatic.merged.SV.CNV.vcf.gz
□ LP3001021-DNA_C10.ITD.vcf.gz
□ LP3001021-DNA_C10.snv.vcf
□ LP3001021-DNA_C10.fisher.snv.vcf.gz
□ LP3001021-DNA_C10.fisher.snv.vcf
□ LP3001021-DNA_C10.vcf.gz

BIGWIG

☑ LP3001021-DNA_C10.GRCh38DecoyAltHLA_NonN_Regions_autosomes_sex_mt.CHR_full_res.bw

CRAM

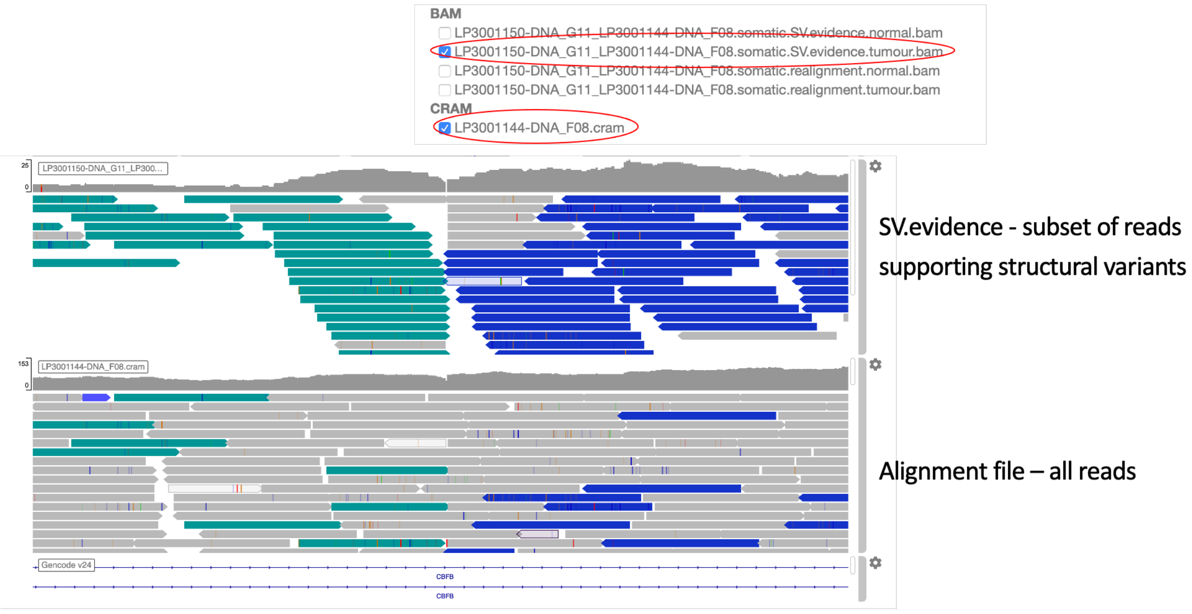
□ LP3001021-DNA_C10.cram

Large deletion visible as reduction in coverage

FIGURE 7 - ASSESSMENT OF LARGE CNVS USING COVERAGE PROFILES

Structural variants, including inversions and translocations, can be assessed by visualising the support from anomalously mapped read pairs. Read pairs for which the distance between reads, orientation of reads or chromosome on which the two reads are aligned are not as expected can indicate the presence of a structural variants. Such read pairs can becoloured coded in IGV. Large CNVs may also be supported by anomalously mapped read pairs at the breakpoints.

Reads supporting structural variants can be viewed in either the genome alignment (CRAM) files or the SV.evidence (BAM) files. The alignment files contain all reads whereas the SV.evidence files contain only reads supporting structural variants and are therefore easier to load and view, as shown in FIGURE 8.



BAM

☐ ~~LP3001150-DNA_G11_LP3001144-DNA_F08.somatic.SV.evidence.normal.bam~~
☑ LP3001150-DNA_G11_LP3001144-DNA_F08.somatic.SV.evidence.tumour.bam
☐ ~~LP3001150-DNA_G11_LP3001144-DNA_F08.somatic.realignment.normal.bam~~
☐ LP3001150-DNA_G11_LP3001144-DNA_F08.somatic.realignment.tumour.bam

CRAM

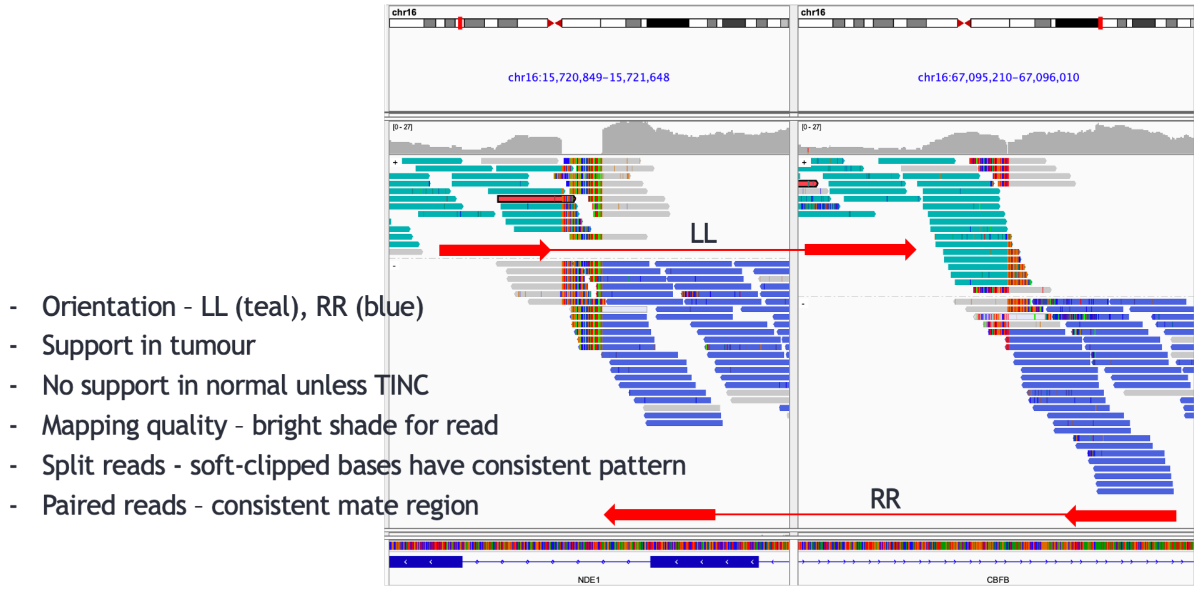
☑ LP3001144-DNA_F08.cram

SV.evidence - subset of reads supporting structural variants

Alignment file – all reads

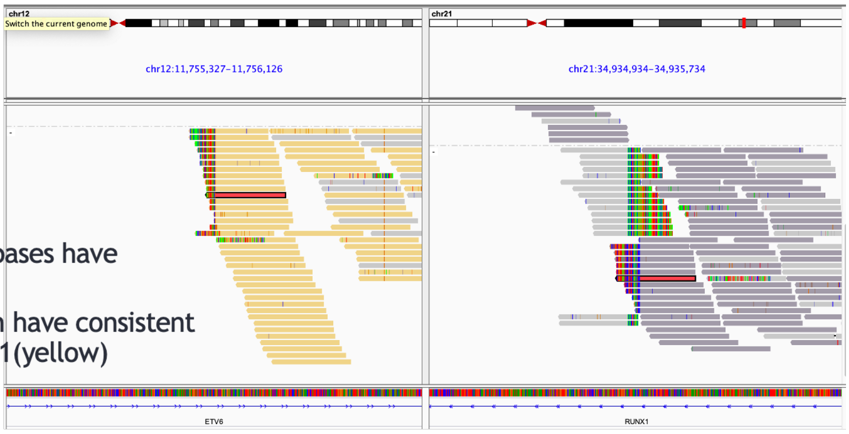
FIGURE 8 - VIEWING EVIDENCE FOR STRUCTURAL VARIANTS USING EVIDENCE AND ALIGNMENT FILES

Characteristics of good quality inversion and translocation (where the two reads in a pair map to different chromosomes) variants are shown in FIGURE 9 and FIGURE 10 respectively.



- Orientation - LL (teal), RR (blue)
- Support in tumour
- No support in normal unless TINC
- Mapping quality - bright shade for read
- Split reads - soft-clipped bases have consistent pattern
- Paired reads - consistent mate region

FIGURE 9 - CHARACTERISTICS OF A HIGH-QUALITY INVERSION


chr12
Switch the current genome
chr12:11,755,327-11,756,126
chr21
chr21:34,934,934-34,935,734
ETV6
RUNX1

- Split reads - soft-clipped bases have consistent pattern
- Paired reads - mate region have consistent pattern: chr12(grey)/chr21(yellow)

FIGURE 10 - CHARACTERISTICS OF HIGH-QUALITY TRANSLOCATION

For reviewing structural variants in IGV, changing the read display is necessary. Alignments should be coloured by "Insert size and pair orientation" (using a similar approach to that shown in FIGURE 6). Read pairings can be shown by selecting "View as pairs" or "view mate region in split screen" depending on the proximity of the reads in a pair. Changing the display to show soft clipped bases is available in the IGV preferences in the alignment tab.
Further information for using the IGV viewer can be found in the IGV user guide

Appendix C – Limitations of the Cancer Bioinformatics Pipeline

A summary of the limitations of the Cancer Bioinformatics pipeline is displayed in the Additional Information section of the WGA results HTML. The following text is shown:

- At the typical genome-wide mean depth of coverage used in WGS analysis (100x), the estimated sensitivity for somatic variants of functional consequence to the protein with allelic frequency >0.1 is 99.3% (95% CI: 99.1-99.6%) for SNVs and 98.1% (95% CI: 96.2-99.7%) for indels (<50bp). At the minimum depth of coverage used for WGS analysis (70x mean coverage), the estimated sensitivity for somatic variants of functional consequence to the protein with allelic frequency >0.1 is 98% (95% CI: 97.5-98.4) for SNVs and 97.1% (95% CI: 94.7% - 99.2%) for indels (<50bp). Estimates are based on comparison of WGS variants passing all quality assessments with high confidence variants detected from high coverage exome sequencing data. These estimates represent the minimum expected sensitivity as true variants detected but not passing WGS quality assessments were discounted from calculations. Variants detected in the WGS analysis which do not meet stringent quality thresholds are shown with a flag. Somatic variants with allelic frequencies <0.1, or in areas of low coverage will be at significantly higher risk of not being detected. The likelihood of failing to detect a variant will increase with progressively lower coverage depth and/or lower allelic frequency. The expected concordance for somatic CNV calling is 96.4% (95% CI: 94-98%) for samples without quality disclaimers (e.g. tumour purity > 30%) compared against equivalent orthogonal tests such as SNP array, FISH and NGS panels. Similarly, the expected concordance for somatic SV calling is 99.2% (95% CI: 98-100%).

- Due to poor mapping in genomic regions within segmental duplications, both somatic and germline variant calling performance may be affected (potentially causing false positive and false negative variants) in genes overlapping such regions (including FCGR2B, HIST2H3C (H3C14), NOTCH2, NUTM2B, NUTM2D, PDE4DIP, and PMS2).

- The expected specificity and precision for all somatic variant types and allele frequencies have not yet been determined. Therefore, false positive results cannot be excluded.
- Variant calls are filtered according to the quality and quantity of reads. Full details of the filters used in this analysis can be found in the Cancer Genome Analysis Guide.
- In this analysis multi-nucleotide variants (MNVs) can be reported as multiple consecutive SNVs and/or indels and therefore the potential protein change may require correction. For some complex germline MNVs, annotation with variants described in the ClinVar database may not be correct.
- A somatic variant may have multiple entries in COSMIC database due to the use of different reference sequences. In these cases, links to all COSMIC entries are provided.
- Links to clinical trials at mycancergenome.org are provided for information purposes only. Status and eligibility criteria may not be up to date.
- For the germline analysis undertaken, it is possible that disease-causing variant(s) are located outside of the list of prioritised variants, for example because they fall outside the gene panels applied, they were located in regions of low coverage, the variant is of a type that could not be detected or the predicted consequence is of a type that is not prioritised. Some complex MNVs involving insertions equivalent to known pathogenic variants in the ClinVar database may not be prioritised. Please note germline structural variants are not currently reported for cancer patients. If the patient has been evaluated as clinically eligible for germline genetic testing on account of their personal and/or family history of cancer, this testing should be performed as per standard local practice.
- If a pathogenic or likely pathogenic germline susceptibility variant is detected, it is recommended that the variant is reviewed by a local clinical laboratory service with expertise in germline cancer genetics. Referral to a clinical cancer genetics unit and technical confirmation of the variant in a new blood sample may be recommended following local variant review.
- For a full description of the methods used to produce these results and for further information regarding QC metrics, please refer to the Cancer Genome Analysis Guide. All related documentation is available at NHS Futures.
- 'N/A' indicates that information is not available or not applicable.

Appendix D – Plots for Depth of Coverage, B-allele Frequency, and Absolute Allele Counts and their interpretation

Depth of Coverage

This plot evaluates how well the model predicts total copy number (CN) and copy number fraction (CNF) by showing the observed coverage along the genome as well as the predicted autosome CN.

Somatic

An intermediate file *.target.counts.gz is used to extract values for the observed coverage. The coverage values in this file are summarised across 500kb bins, with higher opacity indicative of more regions at a given coverage level. The coverage values provided by Dragen are not representative of true coverage of the sample and therefor labels for coverage are not shown to avoid confusion. Instead, predicted CN for autosomes is shown on the y-axis.

To assess model fit against the observed data, we use the purity and diploid coverage estimated by Dragen. We derive the relationship between coverage and CN as follows:

- The subset of tumour coverage that's germline (at this level, tumour CN = 0):
  - $Germline\ Coverage\ =\ (1\ -\ purity_{tumour})\ \times\ diploidcov$

- Then, to calculate the coverage change associated with a single copy:
  - $Tumour\ CN1\ Coverage\ =\ 0.5 \times purity_{tumour} \times\ diploidcov$

- Thus, to calculate the expected tumour coverage at $CN = \alpha$:
  - $Tumour\ CN\alpha\ Coverage\ =\ Germline\ Coverage\ +\ \alpha \times Tumour\ CN1\ Coverage$

- Note: For clonal calls, we use CN, whereas for sub-clonal calls we use CNF in the above calculation.

As expected, coverage depends on the calculated purity of the sample, we expect coverage=0 for regions with CN=0 only for samples with tumour purity of 100%.

Please be aware that the CN scale on the Depth of Coverage plot is calculated based on purity and diploid coverage of the sample. Therefore, this scale should only be used to evaluate regions that are diploid in the germline and coverage for chrX can fall below the CN0 autosome line for male patients.

Please be aware, due to the lack of purity estimate for degenerate diploid samples the above estimates cannot be calculated.

Germline

An intermediate *target.counts.gz file is used to extract values for the observed coverage summarised over 500kb bins. Similarly to the somatic sample, this does not provide an estimate representative of the true coverage of the sample.

To define the set of recurrent CNVs, we calculated CNV frequency in 5,420 germline samples using the AUC frequency to assess recurrency.

The germline CNV analysis does not provide estimates of purity or diploid coverage, so we make the following assumptions when mapping coverage to CN:

- We assume a purity of 100% i.e. tumour in normal contamination (TINC) is 0% when calculating the expected values, thus predictions from samples with high TINC may not match the observed data.
- The diploid coverage is calculated as the median of average autosomal coverage for each chromosome to avoid bias from trisomies.

This simplifies the somatic formula to:

$$Germline\ CN\alpha\ Coverage\ =\ 0.5 \times \alpha \times diploidcov$$

B-allele Frequency

This plot evaluates how well the model predicts haplotypes by showing the observed and predicted B-allele frequency based on the CN calls.

Somatic

For somatic samples, we use the *baf.bedgraph.gz file produced by Dragen to represent the observed B-allele frequencies. These are summarised across 500kb bins on the x-axis and 0.01 bins on the y-axis, with opacity representing the number of variants in each bin.

To evaluate model fit, we first calculate read fraction for each CNV as:

$$CNV_{read\,fraction} = \frac{CNV_{CN} \times purity_{tumour}}{CNV_{CN} \times purity_{tumour} + 2 - 2 \times purity_{tumour}}$$

Then the predicted B-allele frequency is calculated as:

$$CNV_{BAF} = 0.5 \times \left(1 - CNV_{read\,fraction}\right) + CNV_{read\,fraction} \times \frac{CNV_{MCN}}{CNV_{CN}}$$

Where, given a CNV, we define:

- CNV_{CN} as the estimated total CN for clonal calls, and CNF for sub-clonal calls
- CNV_{MCN} as the estimated minor CN for clonal calls, and MCNF for sub-clonal calls
- $purity_{tumour}$ as the estimated tumour purity

For the purposes of the plot, we show both CNV_{BAF} and $(1 - CNV_{BAF})$.

Please be aware, due to the lack of purity estimate for degenerate diploid samples the above estimates cannot be calculated. BAF estimates are also not calculated for complete losses with CN=0.

Germline

Since germline CNV calling does not produce estimates of B-allele frequency, CNV_{MCC}, we cannot use the results from the caller to show observed or estimated B-allele frequency.

Instead, we extract variant allele frequencies (VAFs) from the germline small variant VCF file. We show observed germline B-allele frequency by summarising these VAFs across 500kb bins on the x-axis and 0.01 bins on the y-axis.

Interpretation

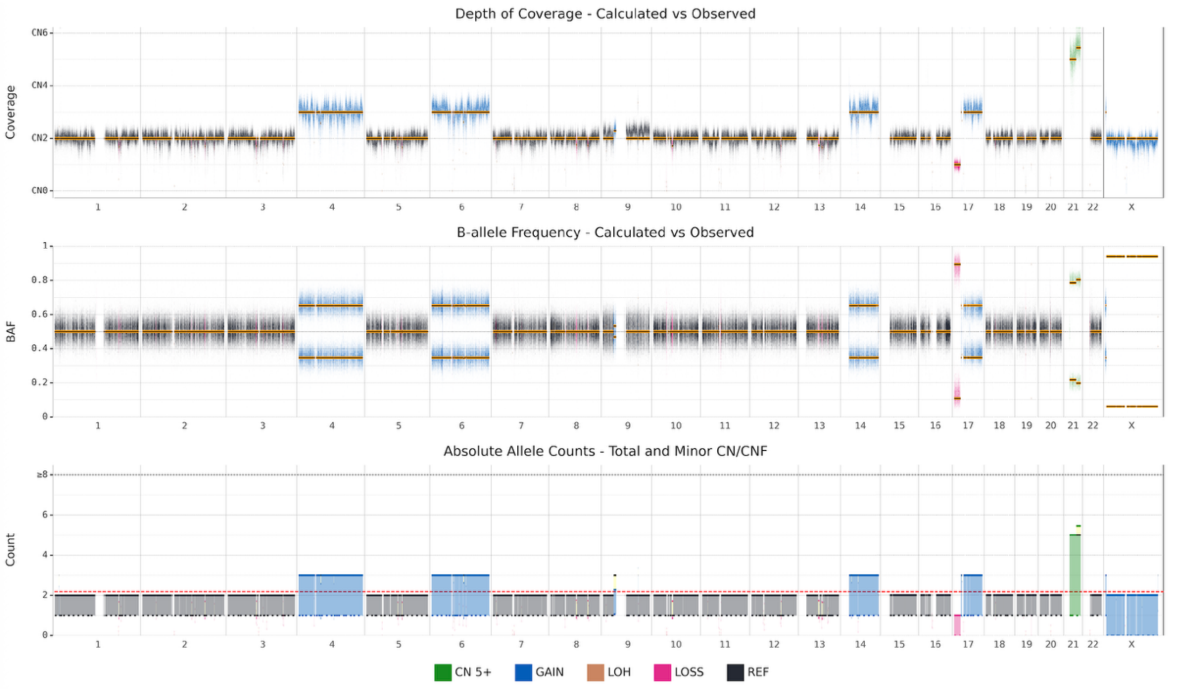
CNVs are represented with the following colour scheme: REF in black (heterozygous diploid), LOSS in pink (CN<2), GAIN in blue (CN=3,4), and CN 5+ (high gain) regions in green. We also include copy neutral LOH in orange (CN=2, Minor=0) for somatic samples and recurrent or non-PASS CNVs in germline in purple. To ensure consistency, we apply the same observed coverage limits to all somatic and germline coverage plots. Any gains above the maximum are shown at the level of that maximum. Regions with subclonal somatic CNVs are highlighted in yellow. The yellow colour highlights the difference between the integer copy number state (black line) and the copy number fraction returned by Dragen, where larger yellow spaces correspond to lower subclone fractions. On Absolute Allele Counts plot, the total allele counts are indicated with solid lines and minor allele counts are indicated with dotted lines.

Genome ploidy is calculated as the mean ploidy across the genome (copied from the header of the Dragen VCF output file) and is shown as a red dashed line.

Example 1: Somatic and Germline plots for patient with germline trisomy 21

- Somatic Plots:

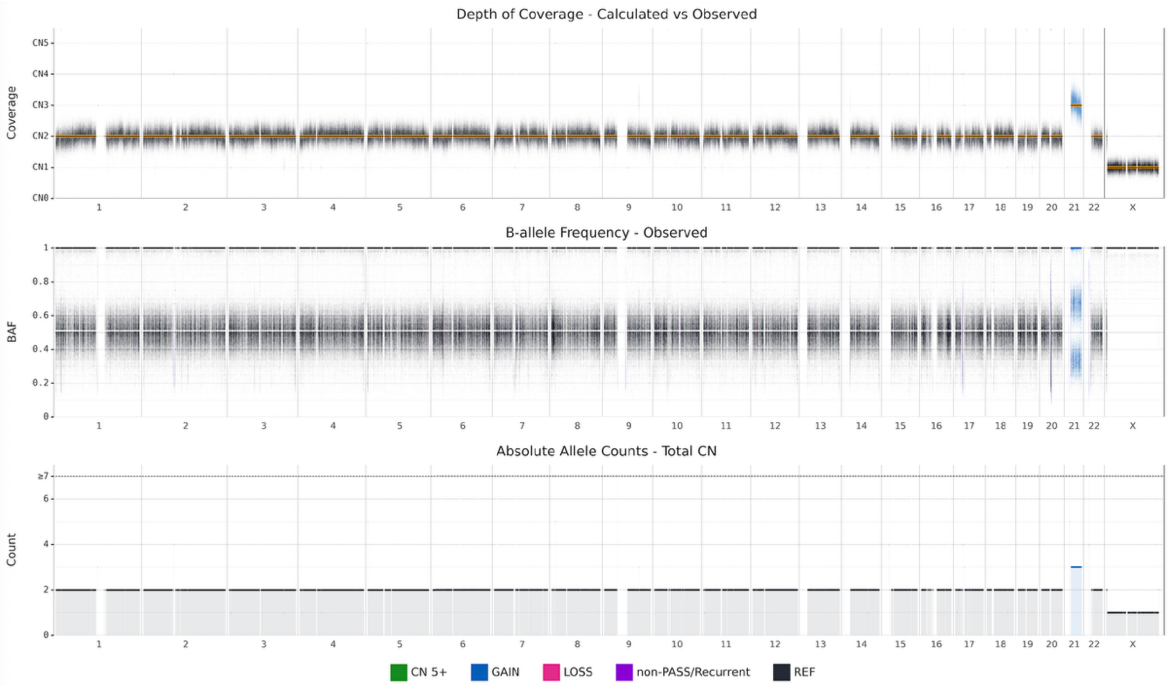
  - The fit (overlap between orange lines for expected values and observed data) is good, but a closer inspection shows that there's likely a missed sub-clonal GAIN on chromosome 9 as the observed coverage sits above the estimated value for diploid region. BAF is also wider spread around this region.
  - There's a small sub-clonal GAIN on chr21, which fits well with coverage and BAF.
  - This is a male patient, as CN2 is classified as a GAIN (in blue) for chrX. Note, you could equally check the germline but this "hack" works for tumour-only plots as well.


Depth of Coverage - Calculated vs Observed

B-allele Frequency - Calculated vs Observed

Absolute Allele Counts - Total and Minor CN/CNF

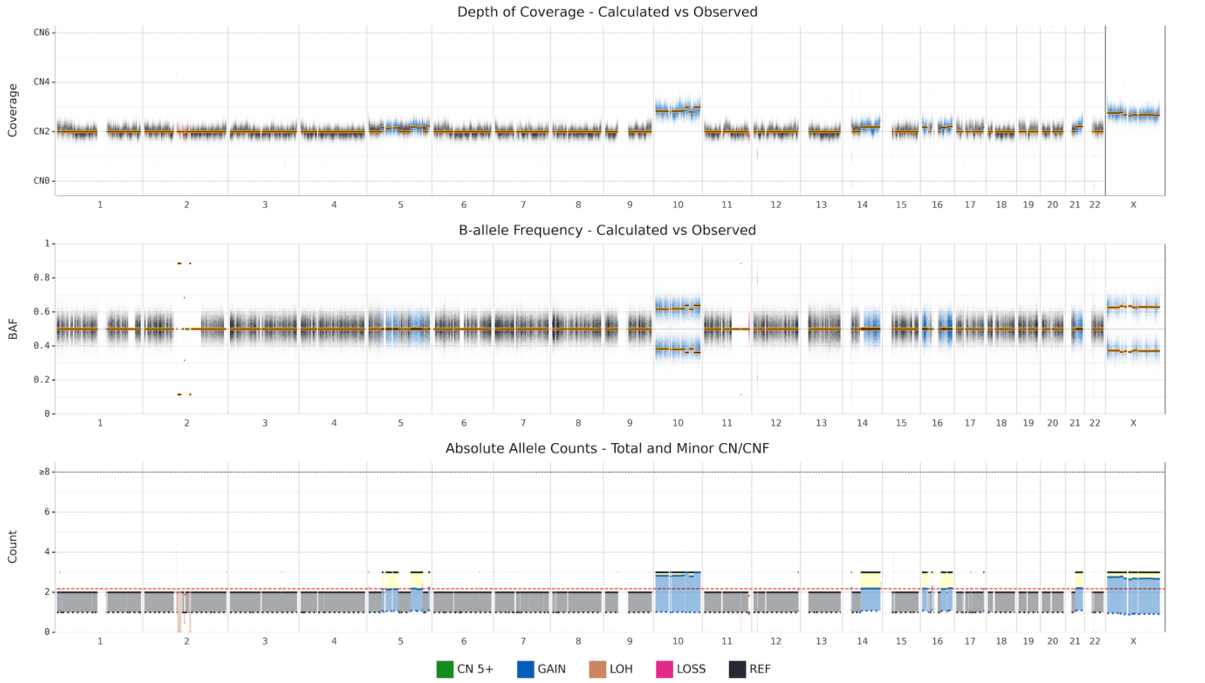
- Germline Plots:
  - The fit is good on the coverage plot.
  - There are no estimated values for the BAF plot, only the observed data is shown.
  - The Absolute Allele Counts plot only shows total CN, as we cannot estimate minor copy number for germline.
  - There's a GAIN on chr21, with CN=3. In this example we can infer that the copy number state of chr21 is likely to be 2:1 based on the split in the BAF plot.



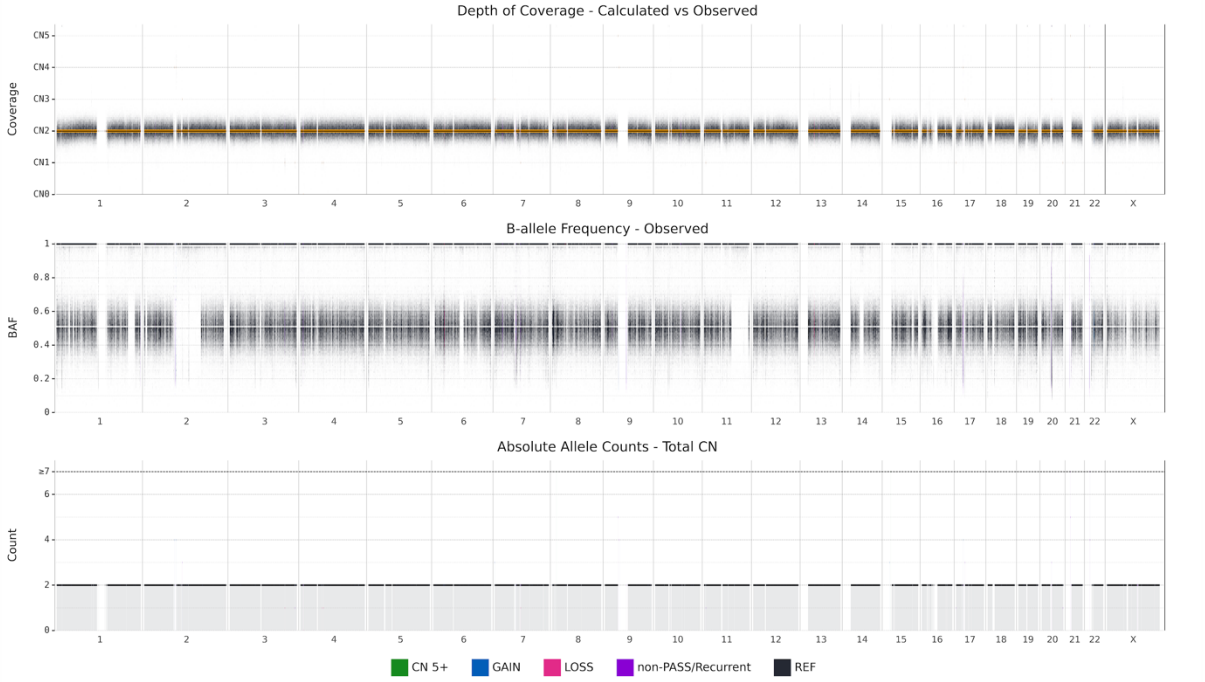
- Combining the information:
    - Using both sets of plots we can infer that chr21 gained additional copies in the tumour compared to the germline.

Example 2: Regions of Homozygosity in the germline

- Somatic Plots:
    - There are some big gaps in the BAF plot (see chr2 and chr11), this is a consequence of the caller relying on heterozygous germline variants to calculate somatic BAF. Checking the germline in these regions we see a lack of observed data at BAF~0.5, suggesting a germline region of homozygosity.
    - There are multiple sub-clonal GAINs (with yellow highlighting the difference between CN and CNF). Multiple GAINs show CN3 (black line) but a range of CNF:
        - On chr10, we note that CNF~3, suggesting that most of the cells have gained an additional copy of chr10, with only a few still diploid.
        - On chr14q, CNF~2, suggesting that only a small proportion of cells have gained an additional copy of chr14q.



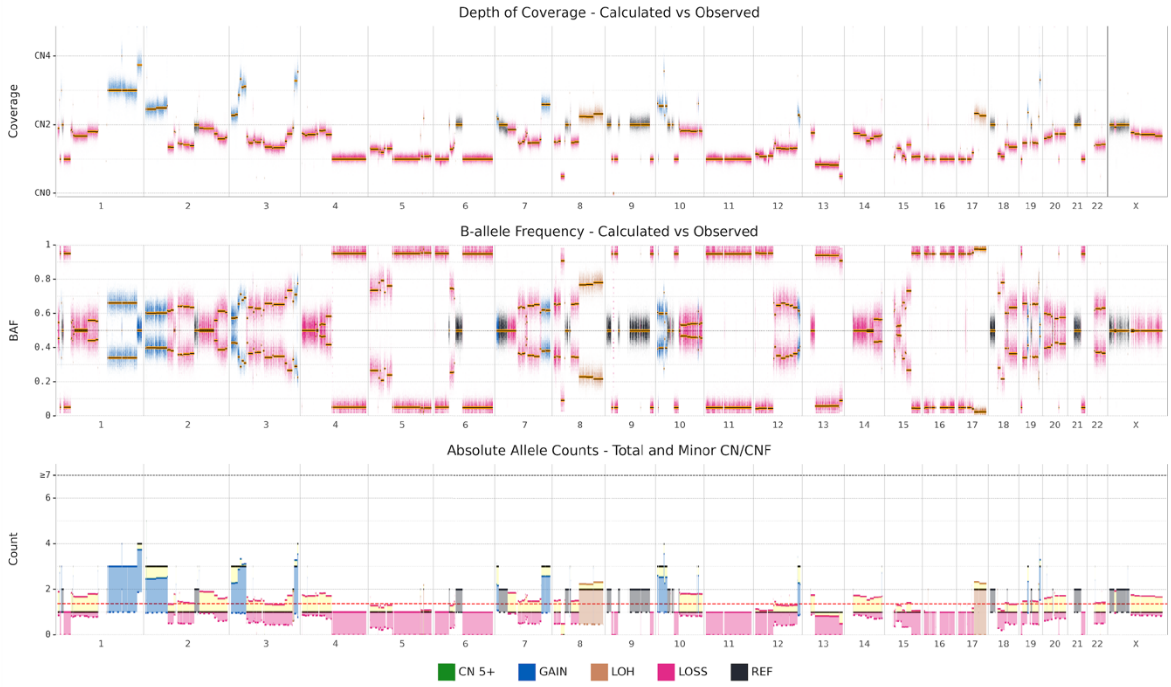
- Germline Plots:
  - The somatic plots already highlighted regions of homozygosity in the germline. This observation can be supported by the abondance of data points with BAF~1 instead of expected BAF~0.5. Note that this is different from small gaps at centromeric regions.



Example 3: Sub-clonal LOSSes

- Based on coverage and BAF plots the CNV model fit is good.
- This sample has multiple sub-clonal calls.
- There are many CN1 sub-clonal LOSSes with a range of CNF and MCNF values:

- On chr13, we see that CNF ~ 0.9, and MCNF=0. This suggests that most cells have lost one copy, and some have lost both copies (as CNF < 1).
- On chr10q, CNF~1.8 and MCNF~0.9. In contrast to the above, here most cells are expected to be diploid with a small fraction that have a LOSS.
- Contrasting chr4p and chr4q, we can see a sub-clonal LOSS affecting some of the cells on 4p but a seemingly clonal loss on 4q. It's important to note again that both of these events will be reported as LOSS(1) in the variant grid, but the former will be flagged as sub-clonal.

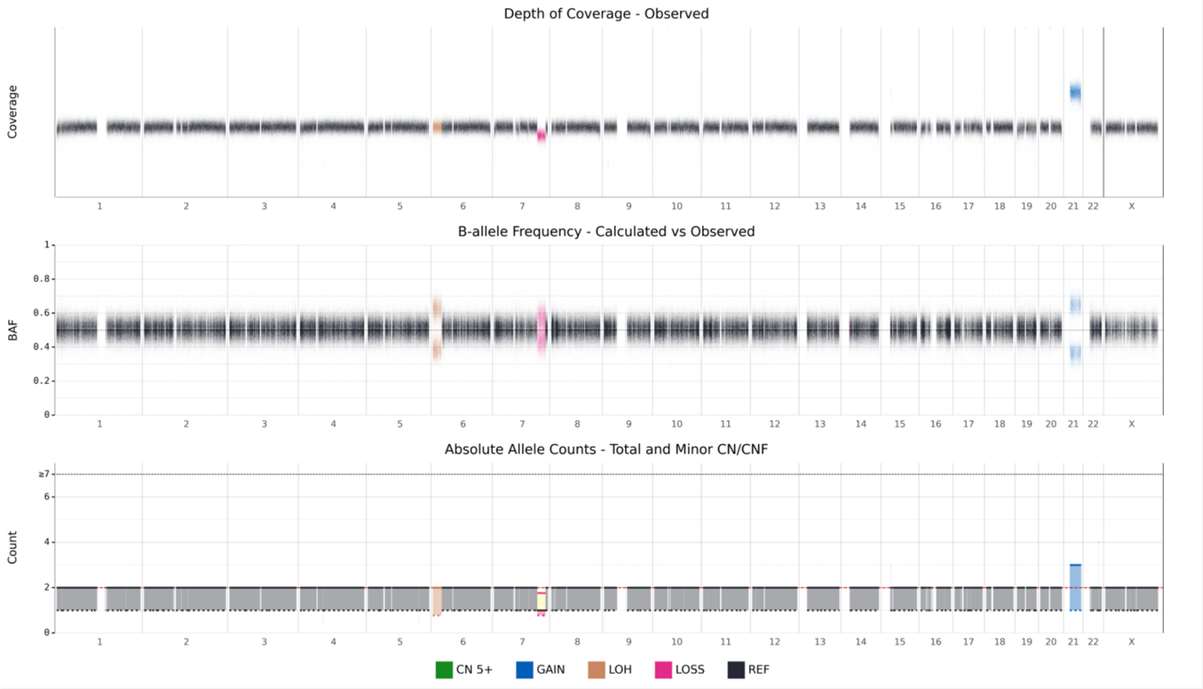

Depth of Coverage - Calculated vs Observed

B-allele Frequency - Calculated vs Observed

Absolute Allele Counts - Total and Minor CN/CNF

Example 4: Degenerate Diploid Tumour-Only

- Expected values for coverage and BAF can't be calculated due to the lack of purity estimate.
- Positions for BAF calculation are being selected from the generic list of common SNPs (instead of the patient's private SNPs) due to the lack of matching germline sample. Only positions with 0<VAF<1 are shown on BAF plot to increase the opacity of interpretable data.


Depth of Coverage - Observed
Coverage
B-allele Frequency - Calculated vs Observed
BAF
1
0.8
0.6
0.4
0.2
0
Absolute Allele Counts - Total and Minor CN/CNF
Count
≥7
6
4
2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
CN 5+
GAIN
LOH
LOSS
REF