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
Confluence Document Version	V2.27
Published Date	Oct 27, 2022
Policy (only if applicable otherwise N/A)	N/A
Document Author	Susan Walker, Alex Younger
Document Reviewer	Alona Sosinsky
Document Approver	Richard Scott
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

Purpose

Scope
- In Scope
- Out of Scope
- Internal Audience
- External Audience

Abbreviations/Definitions

Introduction/Background

Authorities and Responsibilities

Procedure Details

1 Sequencing data and alignment

2 Sample and sequencing quality checks
- 2.1 Sequencing data

2.2 Sample cross-contamination checks
2.2.1 Germline samples
2.2.2 Tumour samples
2.3 Comparison of reported and inferred genomic data
2.4 Tumour sequencing and coverage quality metrics: 'low' vs 'sufficient' quality tumour samples
2.5 Samples with low tumour content or high incidence of somatic variants with low Variant Allele Frequency (VAF)

3 Somatic variant detection
3.1 Small variants
3.2 Copy number and Structural variants
3.2.1 Defect in loss of heterozygosity calling
3.3 Variant detection with tumour in normal contamination

4 Somatic variant interpretation
4.1 Small variants
4.1.1 Domain 1 somatic variants
4.1.2 Domain 2 somatic variants
4.1.3 Domain 3 somatic variants
4.2 Copy number and Structural variants
4.2.1 Domain 1 somatic SV/CNVs
4.2.2 Domain 2 somatic variants
4.2.3 Domain 3 somatic variants

5 Germline findings
5.1 Tier 1
5.2 Tier 3

6 Somatic mutation prevalence (global mutation burden)

7 Mutational signature analysis

8 Explanation of report fields
8.1 Sample attributes
8.2 Sequencing and coverage quality metrics
8.3 Sample and variant quality disclaimers
8.4 Variant descriptions

Process Flow

Supporting or Reference Documents
Related documents
References

Appendices
Appendix A – Validation data
Somatic small variant detection
Germline cross-patient contamination
Somatic small variant detection
Copy number and structural variant detection
Tumour cross-patient contamination
Germline coverage
Appendix B – Data presentation in the Interpretation Portal, HTML files and IGV
Interpretation Portal, HTML file and IGV functionality.
Interpreting variants using IGV
Appendix C – Limitations of the Cancer bioinformatics pipeline

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".

~~**Confluence Version**~~ **Pipeline Version**	**Date** (Day/Month/Year)	**Summary of main changes and reasons** (section no. + Update)
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 reflected: 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 and reporting 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 2.0, including variant calling and interpretation.

Out of Scope

- Description of the Interpretation Portal or Decision Support tools.

Target Audience

Internal Audience

- N/A

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

Introduction/Background

N/A

Authorities and Responsibilities

N/A

Procedure Details

1 Sequencing data and alignment

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.

Cancer Pipeline 2.0 is reporting using GRCh38+Decoy+HLA. Alignment for both 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. The current version of the DRAGEN software is 3.2.22.

Further details of the quality control, variant calling and variant prioritisation processes are outlined in later sections.

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 >85x109 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.
- Tumour samples: 2.1x1011 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 displays in the whole genome analysis (WGA) 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. Germline samples may also be contaminated with DNA from the tumour of the same patient, which may lead to reduced sensitivity for somatic variants.

2.2.1 Germline samples

Cross-patient contamination

Germline samples are assessed with the VerifyBamID algorithm1 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.

Tumour in Normal (TIN) contamination

In the event that 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. In order to identify normal samples with TIN contamination, a specialised quality control component has been designed (TINC test), which identifies clonal mutations in the tumour sample and subsequently estimates the fraction of TIN contamination by assessing the allele fraction of these variants in the germline sample. Warnings are displayed in the whole genome analysis (WGA) HTML file for high TIN contamination when the level of contamination is >5% and low TIN contamination when the level is between 1-5%. In the event of low tumour content (<25%), TIN contamination cannot be estimated reliably, and warnings are displayed (see section 2.5 for further details).

2.2.2 Tumour samples

Tumour samples are assessed using the ConPair algorithm2. 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%, a report is produced with the percentage of contamination highlighted in the WGA results HTML file (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. 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 Comparison of reported and inferred genomic data

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 Tumour sequencing and coverage quality metrics: ‘low’ vs ‘sufficient’ quality tumour 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 (v1.9). AT/CG Dropout and Unevenness of Local Genome Coverage are calculated with in-house developed tools (see further details for sequencing and coverage quality metrics and typical values for good quality samples in section 8.2).

By assessing data quality from a set of 15,000 fresh frozen samples (from the Cancer Programme of the 100,000 Genomes Project), four of these metrics were established as being critical in determining data quality of tumour samples: evenness of coverage, GC dropout, AT dropout and average fragment size. Thresholds have been established for each of these metrics as five standard deviations from the median value and when a sample fails one or more of these parameters, the whole genome analysis (WGA) is returned to NHS GLH staff with a warning that the sample has failed sequencing QC. Such samples are at risk of an increased number of false positive and/or negative variants.

Thresholds are the following:

i. coverage unevenness > 23 AND { CGdrop <-5.4 OR CGdrop > 8.5 OR ATdrop < -3.3 OR ATdrop > 9.0}

OR

ii. average fragment size < 330 bp OR > 630 bp

The above six metrics have been calculated from a cohort of 2000 samples. Metrics for each new sample analysed are compared with this cohort using PCA analysis. Samples with p<10-4 (p: probability density after multivariate normal fitting) are classified as outliers. Outliers are regularly reviewed to identify systematic issues arising during sample preparation or sequencing processes, and GLHs may be notified accordingly.

2.5 Samples with low tumour content or high incidence of somatic variants with low Variant Allele Frequency (VAF)

The percentage of cancer cells in a tumour sample is calculated using the Ccube algorithm3 and presented in the WGA results HTML file (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 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 indicates that the sample either has very low tumour content or high heterogeneity, which impact the sensitivity of variant detection.

3 Somatic variant detection

3.1 Small variants

Somatic small variant detection (single nucleotide variants (SNVs) and indels < 50bp), with tumour normal subtraction, is being performed using Strelka4 (v2.9.9). 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

Variants are not currently removed on the basis of low read count/VAF or germline allele frequency in the general population. 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 (when 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
(GG)	Variants with germline allele frequency >1% in the gnomAD v2 database	Potential unsubtracted germline variants

Variant flag	Indication	Implication
(GE)	Variants with germline allele frequency >1% in an internal 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	Commonly occurring variants with high incidence of false positive variant calls
(PON)	Variants with a Somatic Fisher Phred score <50 based on comparison of allele depths with those at the equivalent variant site in a panel of normals (genomes for 7,000 unrelated rare disease patients or their relative(s))	Potential false positive variants due to alignment or sequencing errors

3.2 Copy number and Structural variants

Detection of somatic structural variants (SVs) and indels >50bp with tumour normal subtraction is being performed with Manta5 (version 1.5) which combines paired and split-read evidence for SV discovery and scoring. Copy number variants (CNVs) are being detected with Canvas6 (version 1.39) which utilises read depth and minor allele frequencies to assign copy number states. Canvas integrates germline small variants in its somatic CNV detection model but does not strictly perform subtraction of germline CNVs. 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
- Canvas-called somatic CNVs with length <10kb

Detection of internal tandem duplications (ITDs) at the FLT3 locus is performed using PINDEL7 (v0.2.5b9).

Fusions involving the SSX2 and SSX4 genes are recurrent rearrangements in genomes of patients with synovial sarcoma and important diagnostic marker. These genes reside within a region of segmental duplication that covers two orthologous genes, SSX2/SSX2B or SSX4/SSX4B. Due to the high homology of these two genes, with short read sequencing, it is not possible for reads to be mapped unambiguously, which subsequently impacts variant calling. 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, the Junction Location Identifier (JuLI) algorithm is being used with adjusted parameters for reads with low mapping quality, with variant calling limited to the *SS18*, *SSX2* and *SSX4* 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.

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).

3.2.1 Defect in loss of heterozygosity calling

A defect in loss of heterozygosity (LOH) calling has been identified in Cancer pipeline 2.0. The defect is the result of a bug in the version of Canvas implemented in the pipeline. During the CNV calling process, neighboring segments of the genome with the same copy number are merged. The result of this bug is that LOH regions are merged with adjacent segments that have the same total copy number (copy number 2) and the whole merged region is reported with the call type (LOH or REF) that applies to the most 5' segment. Consequently, the bug may result in false positive (LOH regions can be detected as longer than appropriate) and false negative LOH calls. Review of B-allele frequency plots in the decision support tool (BSVI) is strongly recommended. B-allele frequency ratios and detection of other variant types are not impacted by this bug.

3.3 Variant detection with tumour in normal contamination (TINC)

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. CNV detection with Canvas is not affected by TINC.

In order to recover variants which may be lost due to TINC, haematological samples with TINC >1% (as detected by TINC test) are also analysed with an unmatched germline sample for small variant and structural variant detection (copy number variant detection is not changed for samples with TINC). The results of variant calling using matched and unmatched germline samples are subsequently merged and analysed together in the annotation and interpretation workflow.

Haematological cases that have an interpretation flag of **TINC HIGH**, **TINC LOW** or **TINC ERROR** in the interpretation portal have been through the TINC workflow (see explanation for interpretation flags below).

- **TINC HIGH:** 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.

- **TINC LOW:** 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.

- **TINC ERROR:** The results of the computational estimation of tumour in normal contamination (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.

TINC None will not have any disclaimers. The TINC workflow is not executed for solid cancers.

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

- **SOMATIC** indicates that the variant has been detected in the matched analysis
- **UNCERTAIN** indicates that the variant has been detected only in the unmatched analysis and may represent variants that were subtracted in the matched analysis due to either TiNC or presence in the germline
- **UNCERTAIN (AF>0.001)** indicates that the variant has been detected only in the unmatched analysis and may represent variants that were subtracted in the matched 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

Variants detected in only the unmatched analysis may represent variants that were subtracted in the match analysis due to TINC or germline variants. To improve processing time for such samples, Somatic Fisher Phred score is only calculated for small variants with VAF < 40%. The variants which come only from unmatched germline variant calling and have population germline allele frequency higher that 1% are not tiered.

If a sample pair has low level of TIN (between 1% and 5%), both preliminary and supplementary reports will be generated, with the copy number profile plots, 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 TIN contamination (over 5%), or if TIN contamination level could not be reliably estimated, only a preliminary report will be generated, and no Domain 3 variants will be shown.

4 Somatic variant interpretation

4.1 Small variants

SNVs and small indels are annotated using Cellbase (v4.7.1 for analysis prior to May 26th 2021, v4.9.5 from May 26th to July 28th 2021 and v4.9.6 from July 28th 2021) with the ENSEMBL (version 90/GRCh38) and COSMIC (version v90/GRCh38) databases. CellBase takes advantage of the data integrated in its database to implement a rich and high-performance variant annotator (with 99.9991% concordance with Ensembl VEP Consequence Types across 1000 genomes phase 3 variants). Variants annotated with the following consequence types in canonical transcripts (see List of canonical transcripts v2) 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)

Complex indels and frameshift variants are not annotated at the protein level. Two non-coding variants in the promotor region of the TERT gene are also reported (Huang et al. 20138).

4.1.1 Domain 1 somatic variants

Variants in a virtual panel of potentially actionable genes are reported in domain 1 (168 genes listed in Actionable genes in solid tumour v2 and 170 genes listed in Actionable genes in haemonc v2 (152 for haematological malignancies of lymphoid lineages or 53 genes for haematological malignancies of myeloid lineages); available in the Cancer analysis additional information document at NHS Futures). For haematological malignancies, the information displayed relates to either lymphoid tumours (where the referred tumour type is lymphoid), myeloid tumours (where the referred tumour type is myeloid) or all haematological tumours i.e. lymphoid and myeloid (in the unusual case that the referred tumour type has features of both e.g. biphenotypic leukaemia). Actionable genes are defined as genes in which small variants (SNVs and indels <50bp) have reported therapeutic, prognostic or clinical trial (both actively recruiting participants or closed to recruitment UK trials) associations, as defined by the GenomOncology Knowledge Management System. Where known, the 'variant-level actionability' category and applicable tumour type are indicated. For other variants in these genes, their impact on gene function has not yet been characterised and therefore their actionability status is unclear. This means:

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

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

4.1.2 Domain 2 somatic variants

Variants in a virtual panel of cancer-related genes (536 genes, listed in the Cancer census genes v2; available in the Cancer analysis additional information document) are reported in domain 2. Cancer-related genes are defined as genes in which any variants have been causally implicated in cancer, as defined by the Cancer Gene Census (https://cancer.sanger.ac.uk/census).

4.1.3 Domain 3 somatic variants

Small variants in genes not included in domains 1 & 2 are reported in domain 3.

4.2 Copy number and Structural variants

Prioritisation of SVs only considers variants with breakpoints within introns or exons of consensus transcripts, with the exception of the immunoglobulin and T-cell receptor loci. For fusions involving IGH, IGK, IGL, TRA, TRB or TRD partner genes 20 kb up or downstream of the breakpoint are also considered. For CNVs detected by Canvas, all genes within 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.

CNVs and SVs are presented in the WGA HTML report 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 Canvas calls, “HC” and “LC” indicate high and low confidence variants, with the Canvas confidence score shown. Quality scores for CNVs take into account:

i. bin count — longer CNVs will be given higher score

ii. coverage for CNV should fit to predicted coverage — sub-clonal CNVs will have low score

iii. distance between current copy number solution and the next one — this distance will be low for high copy numbers where the relative fold change between neighboring copy number states is small. Therefore, copy number variants with a high number of copies (such as focal amplifications) may be designated as low confidence where the specific copy number is uncertain.

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 Canvas and Manta and may be reported more than once in the HTML report. 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 Canvas copy number variants. Lack of support from the second algorithm does not necessarily indicate that a variant is false.

4.2.1 Domain 1 somatic SV/CNVs

CNVs or SVs with breakpoints that overlap genes currently ascribed potential actionability are presented in domain 1 (179 genes listed in Actionable genes in solid tumour SV v2, 147 genes for haematological malignancies of lymphoid lineages or 41 genes for haematological malignancies of myeloid lineages listed in Actionable genes in haemonc SV v2; available in the Cancer analysis additional information document at NHS Futures). For haematological malignancies, gene lists relate to either myeloid, lymphoid or all haematological tumours (i.e. myeloid and lymphoid if the malignancy has features of both lineages). Actionable genes are defined as those in which SVs/CNVs have reported diagnostic, therapeutic, prognostic or clinical trial associations, as defined by the GenomOncology Knowledge Management System. Due to uncertainties in SV interpretation and imprecise CNV breakpoints, SVs and CNVs are included into Domain 1 both when the type of variant detected is equivalent to the actionable variant type ("Strongly matched actionability") and when only the gene (but not the variant type) is concordant ("Weakly matched actionability").

To improve annotation for regions in which fusion breakpoints are known to occur in non-coding regions, the following additional regions have been added to domain 1:

Region label	Coordinates (GRCh38)
MECOM downstream region	chr3:168830599- 169084761
GATA2 enhancer region	chr3:1284928990-128619969
BCOR downstream region	chrX:40030374-40051251

Alternative transcripts in addition to the canonical ones are included for *ABL1* and *STAT6* genes to ensure the correct detection and reporting of BCR-ABL1 fusions in haematological malignancies and NAB2-STAT6 fusions in sarcomas (see List of transcripts in the Cancer analysis additional information document).

4.2.2 Domain 2 somatic variants

CNVs or SVs with breakpoints that overlap genes in a virtual panel of cancer-related genes (536 genes, listed in the Cancer census genes v2 in the Cancer analysis additional information document) are reported in domain 2. Cancer-related genes are defined as genes in which any variants have been causally implicated in cancer, as defined by the Cancer Gene Census (https://cancer.sanger.ac.uk/census).

4.2.3 Domain 3 somatic variants

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

5 Germline findings

Detection of germline small variants is being performed with the DRAGEN small variant caller v3.2.22. 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 (v4.7.1 for analysis prior to May 26th 2021, v4.9.5 from May 26th to July 28th 2021 and v4.9.6 from July 28th 2021) with the ENSEMBL (version 90/GRCh38) and ClinVar (June 19 release for analysis prior to May 26th 2021 and January 2021 release after May 26th 2021) databases. Annotation of germline copy number and structural variants is currently not performed.

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 found on the homepage at https://panelapp.genomicsengland.co.uk/. 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/
Tumour predisposition - childhood onset	https://panelapp.genomicsengland.co.uk/panels/243/
Adult solid tumours cancer susceptibility	https://panelapp.genomicsengland.co.uk/panels/245/

Panel Name	Panel Link
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 Tumour predisposition - childhood onset
	Adult	Sarcoma susceptibility
Haematological Tumours	Childhood	Haematological malignancies cancer susceptibility Tumour predisposition - childhood onset
	Adult	Haematological malignancies cancer susceptibility
Paediatric Tumours	Childhood	Tumour predisposition - childhood onset
	Adult	Tumour predisposition - childhood onset
Solid Tumours – not high-grade serous ovarian cancer or triple negative breast cancer	Childhood	Tumour predisposition - childhood onset Adult solid tumours cancer susceptibility
	Adult	Adult solid tumours cancer susceptibility
High-grade Serous Ovarian Cancer	Childhood	Ovarian cancer pertinent cancer susceptibility Tumour predisposition - childhood onset
	Adult	Ovarian cancer pertinent cancer susceptibility
Triple Negative Breast Cancer	Childhood	Breast cancer pertinent cancer susceptibility Tumour predisposition - childhood onset
	Adult	Breast cancer pertinent cancer susceptibility
Neurological Tumours	Childhood	Brain cancer pertinent cancer susceptibility Tumour predisposition - childhood onset
	Adult	Brain cancer pertinent cancer susceptibility

Clinical indication	Age group	Panels for tier 3 variants
Sarcoma	Childhood	Sarcoma susceptibility Tumour predisposition - childhood onset Adult solid tumours cancer susceptibility
	Adult	Sarcoma susceptibility Adult solid tumours cancer susceptibility
Haematological Tumours	Childhood	Haematological malignancies cancer susceptibility Tumour predisposition - childhood onset Adult solid tumours cancer susceptibility
	Adult	Haematological malignancies cancer susceptibility

Clinical indication	Age group	Panels for tier 3 variants
		Adult solid tumours cancer susceptibility Tumour predisposition - childhood onset
Paediatric Tumours	Childhood	Tumour predisposition - childhood onset Adult solid tumours cancer susceptibility
	Adult	Tumour predisposition - childhood onset Adult solid tumours cancer susceptibility
Solid Tumours – not high-grade serous ovarian cancer or triple negative breast cancer	Childhood	Tumour predisposition - childhood onset Adult solid tumours cancer susceptibility
	Adult	Adult solid tumours cancer susceptibility
High-grade Serous Ovarian Cancer	Childhood	Ovarian cancer pertinent cancer susceptibility Tumour predisposition - childhood onset 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 Tumour predisposition - childhood onset Adult solid tumours cancer susceptibility
	Adult	Breast cancer pertinent cancer susceptibility Adult solid tumours cancer susceptibility
Neurological Tumours	Childhood	Brain cancer pertinent cancer susceptibility Tumour predisposition - childhood onset 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 in the WGA are indicated in the WGS HTML files. 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).

ClinVar (https://www.ncbi.nlm.nih.gov/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 are available here: https://www.ncbi.nlm.nih.gov/clinvar/docs/review_status/ .

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

5.1 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. For genes with a biallelic mode of inheritance (as documented in PanelApp), only homozygous or potential compound heterozygous variants are reported.

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 HTML report.

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. 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.

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 Intergrative Genome Viewer (IGV) and assessed via ACMG criteria.

5.2 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 2 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.

It is not anticipated or required that Tier 3 will be reviewed for all patients. Please see the NHS England Guidelines for Cancer Whole Genome Sequencing & Next Generation Sequencing Panel Interpretation & Reporting guidance document for further details.

6 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 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 HTML report.

7 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) (https://cancer.sanger.ac.uk/cosmic/signatures_v2). 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.

8 Explanation of report fields

Two whole genome analysis (WGS) HTML results files as well as a machine-readable lists of identified mutations are provided in the Interpretation Portal; a short report and a longer report with supplementary analyses.

The short WGA results HTML file includes:

- Somatic small variants in Domains 1 and 2
- Somatic fusions/rearrangements and copy number aberrations in Domains 1 and 2
- Pertinent germline findings in Tier 1

The supplementary WGA results HTML file contains additional information for:

- Somatic small variants in Domain 3
- Somatic fusions/rearrangements and copy number aberrations in Domain 3
- COSMIC signatures
- Mutation burden
- Pertinent germline findings in Tier 3

8.1 Sample attributes

The following characteristics are reported in the HTML results:

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, estimated by Canvas. 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 Canvas)
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

8.2 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%. The median value for germline samples is 1.28% with standard deviation of 0.33%.
Median Insert size, 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.
Unevenness of Local Genome Coverage	This metric represents read depth uniformity across the genome. Unevenness is calculated as median for the root mean square deviation (RMSD) of coverage calculated in non-overlapping 100 kb windows. This metric would be 0 for a genome with absolutely uniform coverage. The median value for good quality tumour samples is 13.8 with standard deviation of 1.7. The median value for germline samples is 7.8 with standard deviation of 0.8.
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.

Metric	Explanation
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.

8.3 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 HTML reports:

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
Potential low-quality sample	Tumour sample fails one or more of the sequencing quality control thresholds applied for evenness of coverage, GC dropout, AT dropout and average fragment size. As a result, there may be an increased risk of false positive or negative variants.
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.
Low tumour in normal (TiN) contamination	The germline sample has a low level of contamination with DNA derived from the tumour (between 1% and 5%). Consequently, the sensitivity of somatic variant detection may be reduced, potentially resulting in an increased risk of false negative findings
High tumour in normal (TiN) contamination	Germline sample has a high level of contamination with DNA derived from the tumour (above 5%). Consequently, the sensitivity of somatic variant detection is likely to be reduced, resulting in an increased risk of false negative findings
Tumour in normal contamination estimation not available	The computational estimation of tumour in normal contamination is not reliable (likely due to low tumour content in the tumour sample). Consequently, TiN cannot be excluded and the sensitivity of somatic variant

Disclaimer	Description
(haematological malignancies only)	detection may be reduced, potentially resulting in an increased risk of false negative findings.

8.4 Variant descriptions

Variants presented in HTML reports are annotated with the following features:

Variant annotation	Explanation
Small somatic variants	
Gene- or variant– level actionability	Therapies or clinical trials for which the patient may be eligible. Cancer type abbreviations have been used and full annotation is available in the Cancer type abbreviations v2 document
CDS change	Coding DNA change calculated with the Mutalyzer API
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 (downloaded in July 2020 from http://cancer.sanger.ac.uk/census; see the list in Cancer census genes v2)
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 Canvas call (quality score >=10) LC – low confidence Canvas call (quality score < 10) 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 or JuLI) DEL = deletion (Manta) DUP = tandem duplication (Manta) GAIN = copy number gain (Canvas) INS = insertion (Manta) INV = inversion (Manta) ITD = internal tandem duplication (Pindel) LOH = copy number-neutral loss of heterozygosity (Canvas) LOSS = copy number loss (Canvas)
Impacted transcript region	For Manta calls - Location breakpoints within the affected gene (e.g. intron, exon, intergenic region) For Canvas calls - part of the transcript that overlaps with the CNV (e.g. partial coding sequence, full transcript)

Variant annotation	Explanation
Population germline allele frequency (GESG | GECG)	Population germline allele frequency for the breakpoints of a given structural variant based on two internal panels of normals: GESG, which consists of germline variants coming from single germline analysis of about 2,200 samples, and GECG, which consists of the variants detected as germline in paired tumour-normal variant calling for about 2,500 cancer samples.
Weakly or strongly matched actionability	Due to uncertainties in SV interpretation and imprecise CNV breakpoints, SVs and CNVs are included in Domain 1 where the variant type is equivalent to the known actionable variant type (annotated as "Strongly matched actionability") and where the known actionable variant type differs (annotated as "Weakly matched actionability")

Process Flow

(Non-mandatory - state N/A here if this section is not applicable. Simple flowchart providing an overview of the actual process)

Supporting or Reference Documents

Related documents

1. Cancer analysis additional information (available at NHS Futures)
a. List of canonical transcripts v2
b. Actionable genes in solid tumour v2
c. Actionable genes in solid tumour SV v2
d. Actionable genes in haemonc v2
e. Actionable genes in haemonc SV v2
f. Cancer census genes v2
g. Cancer type abbreviations v2
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 doi: https://doi.org/10.1101/484402

4. 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

5. 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.

6. Canvas: versatile and scalable detection of copy number variants. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Bioinformatics. 2016 32(15):2375-7

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. 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.

Appendices

Appendix A – Validation data

Validation of the Cancer Pipeline 2.0 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 available on NHS Futures). 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 of the HTML reports 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 at range of different allelic frequency ranges at both 100x and 70x tumour genome coverage are shown below.

Metric (tumour coverage)	5-10% VAF mean (95% CI)	10-15% VAF mean (95% CI)	15-100% VAF mean (95% CI)
Precision SNVs (100x)	0.87 (0.84-0.90)	0.92 (0.90-0.94)	0.91 (0.90-0.92)
Sensitivity SNVs (100x)	0.91 (0.89-0.93)	0.98 (0.97-0.99)	0.99 (0.99-1.0)
Sensitivity SNVs (70x)	0.85 (0.82-0.87)	0.94 (0.92-0.96)	0.99 (0.99-1.0)
Precision Indels (100x)	0.44 (0.25-0.62)	0.56 (0.36-0.75)	0.92 (0.87-0.97)
Sensitivity indels (100x)	0.90 (0.63-1.0)	0.86 (0.75-0.97)	0.97 (0.94-1.0)
Sensitivity indels (70x)	0.90 (0.63-1.0)	0.81 (0.68-0.93)	0.97 (0.94-1.0)

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. No impact of contamination on the sensitivity of somatic variant detection was observed, which is to be expected as it is unlikely that a germline variant from the contaminating sample corresponds with a true somatic variant. However, a reduction in precision was observed when germline contamination is greater than 12% (both for all variants and those of functional consequence to the protein). Sensitivity (recall) and precision are shown for a range of germline contamination values for all somatic small variants passing basic variant filters (PASS status) are shown in FIGURE 1.

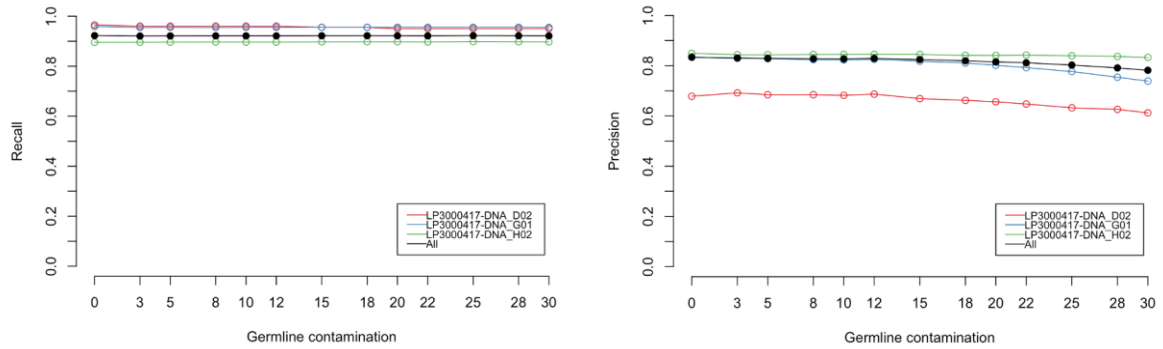


FIGURE 1 SENSITIVITY AND PRECISION OF SOMATIC VARIANT CALLING IN THE PRESENCE OF GERMLINE SAMPLE CROSS-PATIENT CONTAMINATION

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. These data show that from approximately 10% germline contamination, there can be a drastic increase in the number of gain CNVs detected as a result of the overall ploidy of the tumour being incorrectly estimated, depending on the nature of the tumour. The overall tumour ploidy predicted by Canvas at a range of levels of germline contamination for three tumour-normal pairs is shown in FIGURE 2.

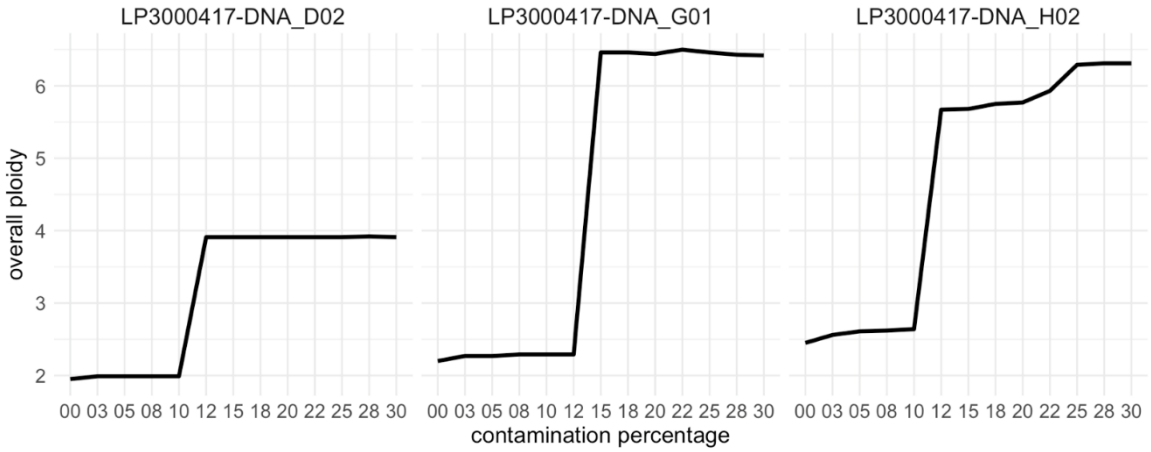


FIGURE 2 IMPACT OF GERMLINE CROSS-PATIENT CONTAMINATION ON TUMOUR PLOIDY ESTIMATION

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%.

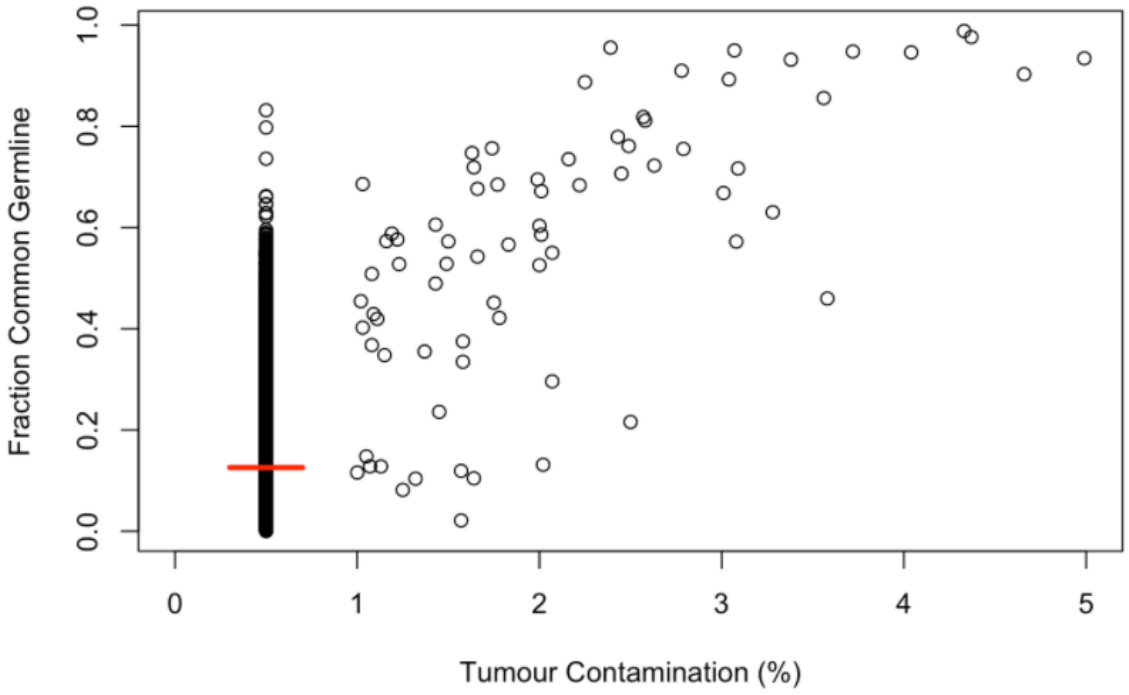


FIGURE 3 FRACTION OF COMMON GERMLINE VARIANTS DETECTED AS SOMATIC VARIANT AT A RANGE OF LEVELS OF TUMOUR CROSS=PATIENT CONTAMINATION. ALL SAMPLES WITH CONTAMINATION <1% ARE PLOTTED TOGETHER AT 0.5% AND THE MEAN VALUE INDICATED WITH THE RED LINE.

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) falls below 95% when mean germline coverage is reduced below 15x (data shown in FIGURE 4).

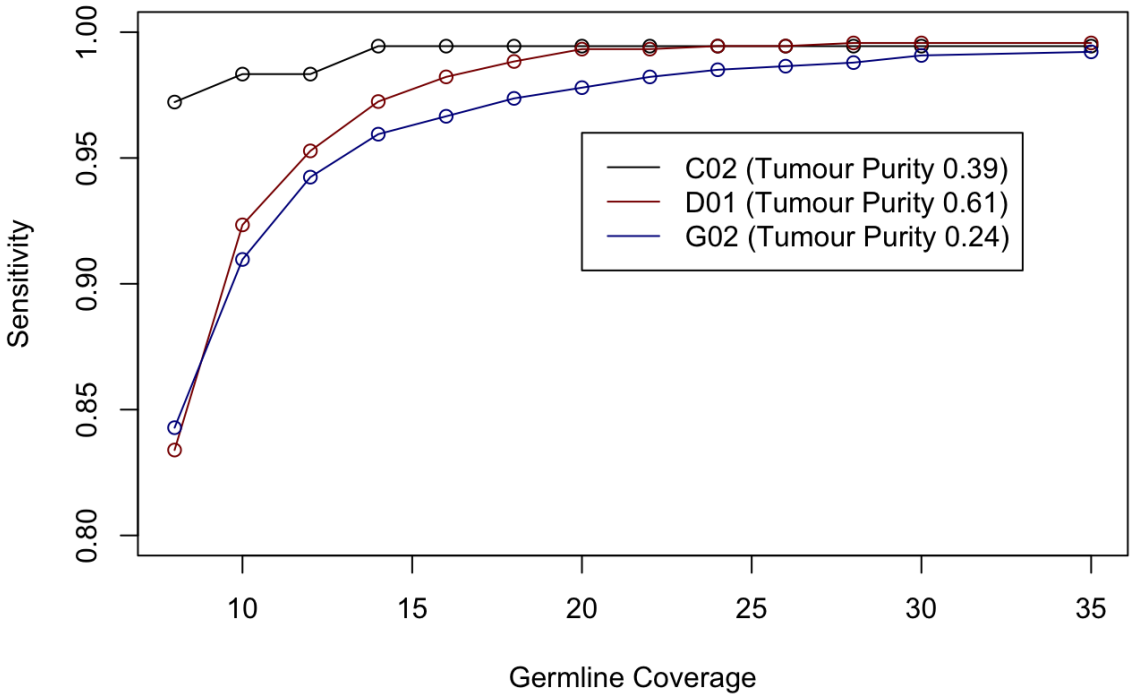


FIGURE 4 IMPACT OF GERMLINE COVERAGE ON THE SENSITIVITY OF SOMATIC VARIANT DETECTION

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 HTML WGA report.

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

Interpretation Portal, HTML file and IGV functionality.

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

Supplementary 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 Canvas**
<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
<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.

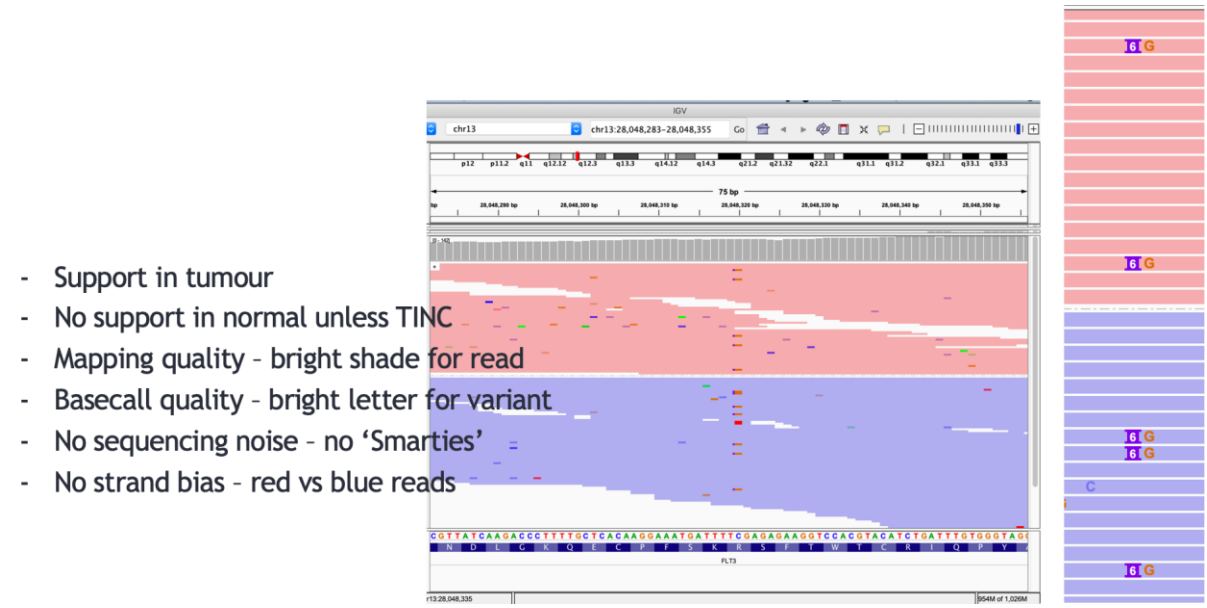


FIGURE 5 CHARACTERISTICS OF HIGH QUALITY SOMATIC SMALL VARIANTS

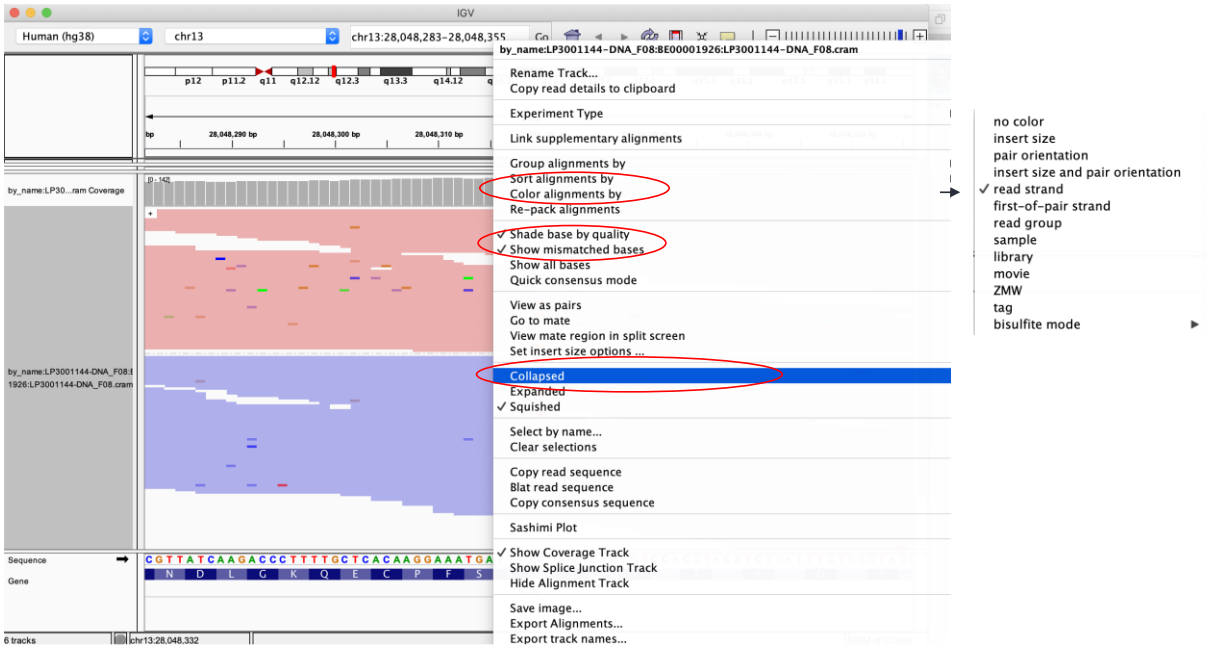


FIGURE 6 IGV SETTINGS HELP FOR SMALL VARIANT VISUALISATION

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.

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 be

coloured 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.

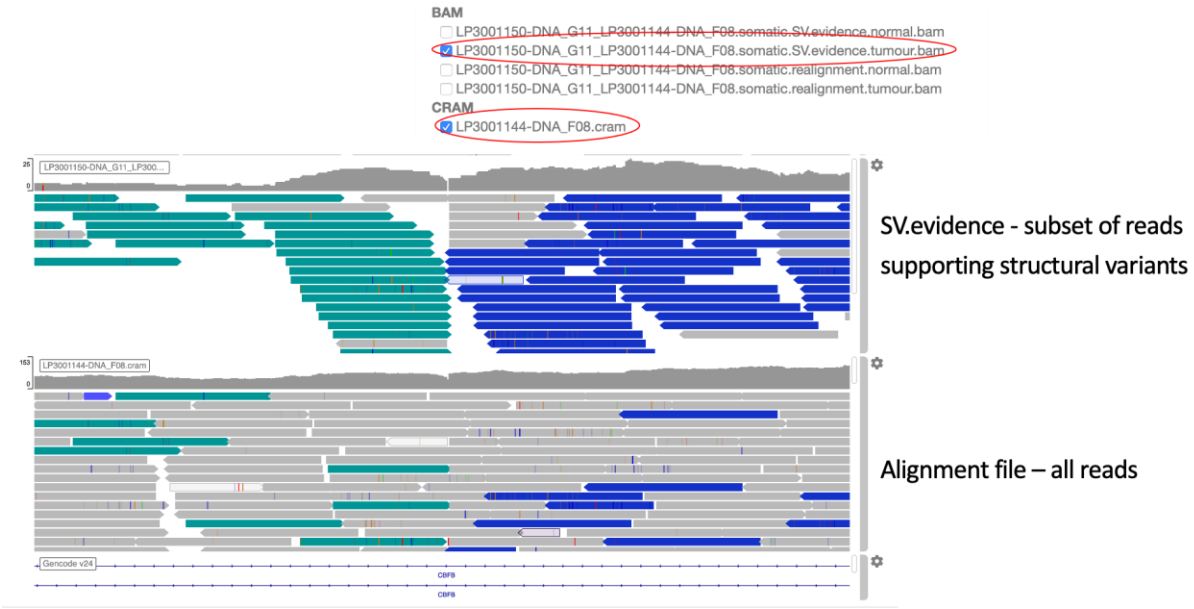
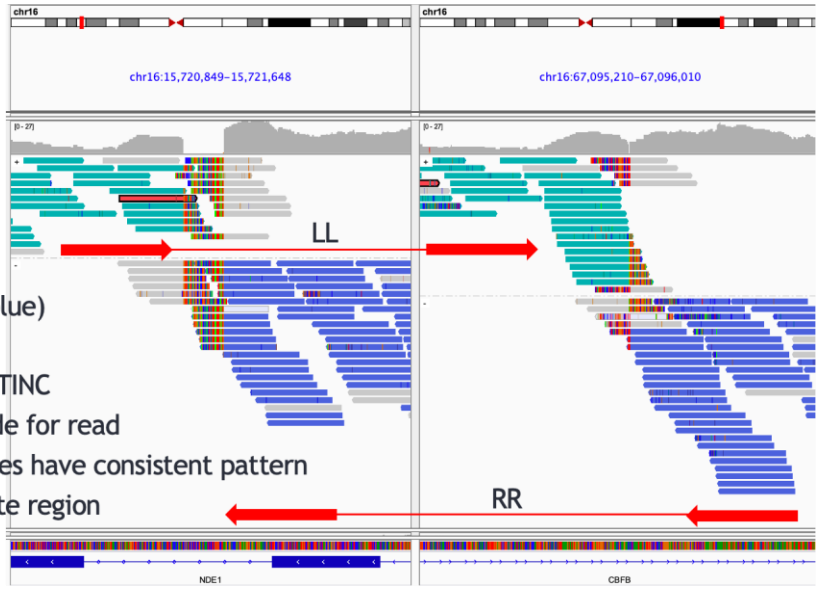


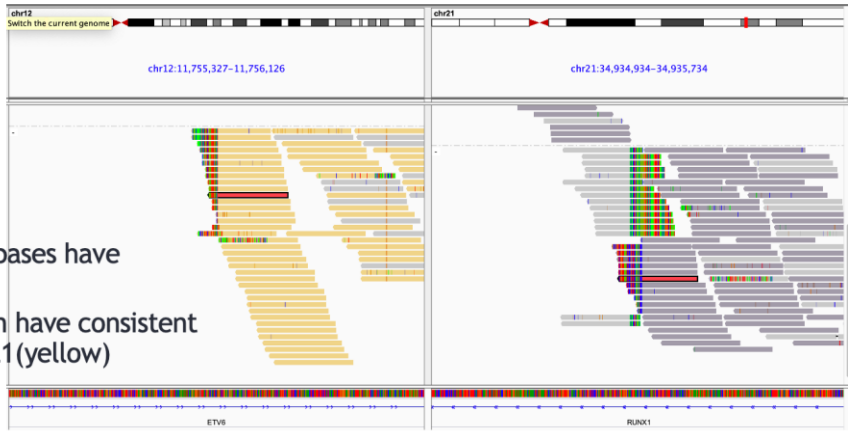
FIGURE 8 VIEWING EVIDENCE FOR STRUCTURAL VARIANTS USING EVIDENCE AND ALIGNMENT FILES

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



- 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: http://software.broadinstitute.org/software/igv/UserGuide

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 HTML WGS reports. 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

99.4% (95% CI: 99.1-99.7) for SNVs and 94.8% (95% CI: 90.9-98.3) 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 98.5% (95% CI: 98.0-98.9) for SNVs and 93.2% (95% CI: 88.7% - 97.3%) 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 sensitivity for detection of SVs and CNVs is yet to be determined. False negative results cannot be excluded.

- 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 clinicaltrials.gov 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 and copy number 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.