Table of Contents

III.II Small variant tiering

III.III Copy number variant tiering

III.IV Short tandem repeats

I. Background

1 Pipeline Overview

The primary diagnostic analysis consented to as part of the Genomic Medicine Service aims to provide prioritised variants for patients with sufficient evidence for diagnostic reporting related to their primary condition.

The Genomics England pipeline aims to facilitate this by annotating a shortlist of 'tiered' variants that are likely or plausibly disease causing for assessment by NHS GLH staff. It should be noted that 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 as would be standard in a diagnostic laboratory, confirm the presence of selected variants where required, and report and authorise any results.

A major component of the Tiering process is the application of diagnostic grade virtual panels relevant to each family's phenotype. These reflect the current EuroGentest and ESHG guidelines that state:

> 'For diagnostic purpose, only genes with a known (i.e., published and confirmed) relationship between the aberrant genotype and the pathology, should be included in the analysis.'

Note

Variants that are prioritised by the Genomics England Tiering process are available, along with their associated annotations in the Interpreted Genome output files (available in json format).

The Genomics England Interpretation Portal and Clinical Interpretation Partner's tools also allow NHS GLH staff to explore the genome beyond the tiered variants so that variants outside the virtual gene panels applied or that do not pass default filters can be explored.

2 Feedback

If you have any feedback on the Genomics England Rare Disease bioinformatics pipeline please contact the Genomics England Service Desk at ge-servicedesk@genomicsengland.co.uk.

3 Purpose

The purpose of this documentation 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 clinical reporting of primary findings in Rare Disease. This guide includes the processes carried out from the receipt of phenotype and genome sequencing data through to presentation of data in the Interpretation Portal.

4 Scope

4.1 In scope

This site describes the Whole Genome Sequence (WGS) analysis performed in the Rare Disease bioinformatics pipeline 2.0 including variant calling and interpretation.

4.2 Out of scope

- Description of the Interpretation Portal
- Description of Decision Support tools

5 Target Audience

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

Other Third Party Audiences

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:

- Competent Authorities (CAs) from within the European Union (EU), including the Medicines and Healthcare Products Regulatory Agency (MHRA); the United Kingdom (UK) CA;
- Notified Bodies (NBs) from within the EU, 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.

6 Authorities and Responsibilities

N/A

7 Accreditation

Genomics England Limited is a UKAS accredited medical laboratory No. 10170. Genomics England's Schedule of Accreditation includes Fixed Scope and Flexible Scope permissions;

This schedule includes Fixed Scope and Flexible Scope permissions; the latter allows Genomics England to make changes within defined and agreed UKAS boundaries and report the results as accredited. Any results reported that are outside the Fixed Schedule and Flexible Scope boundaries are noted in the statement of limitations and highlighted as outside the scope of UKAS accreditation.

Further detailed information of changes made to the bioinformatics pipeline are summarised in the release notes and in the relevant Rare Disease Genome Analysis Guide for each software version.

II. Bioinformatics pipeline

II.I Genome alignment and variant detection

8 Overview

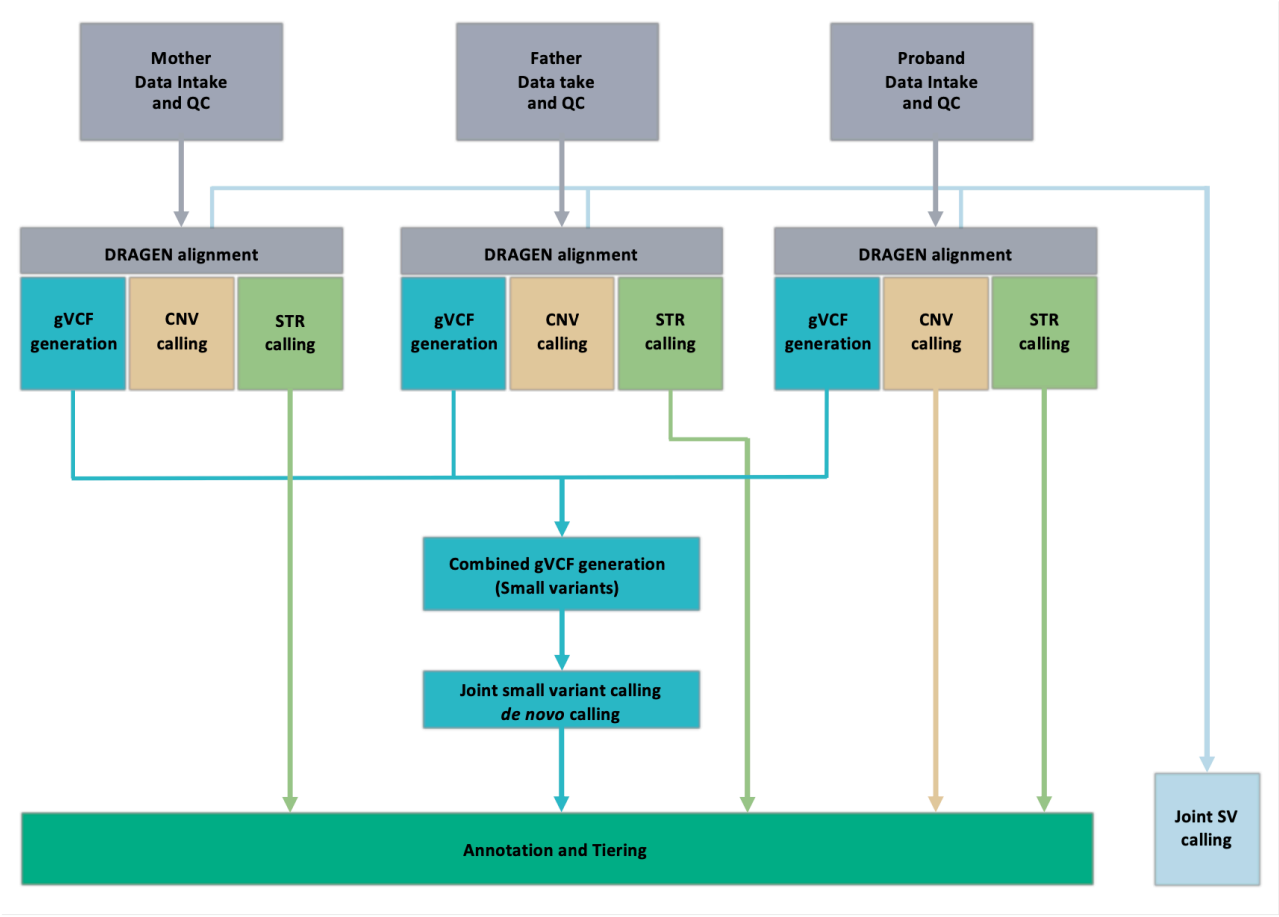
The Rare Disease Pipeline uses genome reference GRCh38. Sequencing read alignment to the genome reference including decoy contigs and alternate haplotypes (ALT contigs) is performed using the DRAGEN aligner, with graph-based ALT-aware mapping and variant calling to improve specificity. The graph-based method enables accuracy improvements in difficult-to-map regions and segmental duplications. For further details, see DRAGEN graph mapper

Alignments are stored in CRAM files which contain both mapped and unmapped reads.

Detection of small variants (single nucleotide variants (SNVs) and indels) and copy number variants (CNVs) are performed using the DRAGEN small variant caller and DRAGEN CNV respectively. Short tandem repeat (STR) expansions are being detected using ExpansionHunter (v5) as part of the DRAGEN software.

DRAGEN software is used for alignment and variant calling. Small variants and CNVs are being tiered and reported for chromosomes 1 – 22 and chrX. Small variants are also tiered and reported for the mitochondrial genome. STR expansions are being detected and tiered at selected loci. Structural variants (SVs) are not tiered but are being detected in the pipeline using a specialised SV caller integrated within DRAGEN software (DRAGEN SV, derived from Manta). Tiering is described further in sections 9.3, 9.5 and 9.6.2.

DRAGEN software incorporates the 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. Variant calling is performed assuming a haploid model for chromosome X for individuals inferred to have a single copy of chromosome X (for example, XY, XO, XYY karyotypes) and assuming a diploid model for individuals inferred to have two or more copies of chromosome X (for example, XX, XXX, XXY karyotypes). A summary of the alignment and variant calling process is shown below.


Mother
Data Intake
and QC
Father
Data take
and QC
Proband
Data Intake
and QC
DRAGEN alignment
DRAGEN alignment
DRAGEN alignment
gVCF generation
CNV calling
STR calling
gVCF generation
CNV calling
STR calling
gVCF generation
CNV calling
STR calling
Combined gVCF generation (Small variants)
Joint small variant calling
de novo calling
Annotation and Tiering
Joint SV calling

9 *de novo* variant detection

The detection of *de novo* small variants using the DRAGEN algorithm is performed directly from gVCF files rather than alignment files (as in some other algorithms such as Platypus). In generating gVCF files, homozygous reference variants with similar quality scores from consecutive genomic positions are collapsed into a group and represented by a single entry in the resulting file. Consequently, metrics (e.g., quality scores, depth of coverage, allelic fractions etc) relating to specific sites with homozygous reference genotypes may not be available. Thus, care should be taken when interpreting the apparent allelic depth in parental samples at sites corresponding to *de novo* variants in their offspring (i.e., homozygous reference positions in the parents) as the metrics presented in VCF files may not correspond to the anticipated position.

This behavior applies to all chromosomes, including sex chromosomes and the mitochondrial genome. Allele counts for homozygous reference positions can be obtained directly from the CRAM file, for example by viewing the CRAM file in IGV or generating a pile-up using bcftools for a specified genomic position.

The DRAGEN *de novo* small variant detection algorithm determines all positions for which the genotypes in a trio are not consistent with a Mendelian inheritance pattern. Detection of *de novo* variants is not restricted to variant positions with homozygous reference genotypes for the parents and a heterozygous genotype for the offspring.

There are three possible groups that a variant can be assigned from small variant *de novo* detection:

DN value	Description
Inherited	Genotype is consistent with Mendelian inheritance in the trio
LowDQ	Genotype is inconsistent with Mendelian inheritance in the trio DQ score is less than quality threshold
DeNovo	Genotype is inconsistent with Mendelian inheritance in the trio DQ score is greater than or equal to quality threshold

Prioritisation of *de novo* variants does not take into account the affection status of family members other than the proband. This does not affect biallelic genes but in genes with a monoallelic mode of inheritance can result in the prioritisation of variants that are absent in affected parents.

Note

de novo quality score (DQ) thresholds are posterior probability scores calculated from the consideration of possible genotypes within the trio - with the probability of error assumed to be independent for each member of the trio. The rare disease bioinformatics pipeline applies default DQ values for SNVs (≥0.0013) and indels (≥0.02).

Note

Only *de novo* variants with quality (DQ) scores that match or exceed the thresholds indicated in the box above will be considered during the Genomics England variant tiering approach and displayed to users in the Interpretation Portal for further consideration. The DQ scores are not included in the displayed information to the user, but are present in the VCF file available for download.

10 Mitochondrial variant detection

Detection of variants in the mitochondrial genome with the DRAGEN small variant detection is outside the scope of ISO 15189 accreditation for the pipeline, but variants are provided to the user for consideration.

The detection algorithm utilises a continuous allele frequency model. Given that there are many copies of the mitochondrial genome per cell, the continuous allele frequency model is more appropriate for mitochondrial variant detection as it assumes that the variant allele fraction can vary between 0 – 100% and facilitates detection of low level heteroplasmy.

Please note that the measurement of uncertainty (i.e. the accuracy of heteroplasmy level value) was not determined against standard of care tests. The manufacturer's (i.e. Illumina) stated limit of detection for mtDNA heteroplasmy is 1%, with recommended threshold of 2% to increase calling specificity (i.e. decrease the number of false positive calls), however these values have not been formally validated by Genomics England against standard of care tests.

The Genomics England Rare Disease bioinformatics pipeline will consider small mitochondrial variants for prioritisation during variant tiering if they are present at ≥5% allelic fractions.

Information relevant to how variants will be presented in the Interpretation Portal, if they are prioritised, are included below:

Variant state	Allele Fraction	Genotype representation in VCF
Heteroplasmic	≥5% and <95%	0/1
Homoplasmic	≥95%	1/1

II.II Quality control and genomic identity checks

11 Genomic and data checks

As part of the data quality processes followed in the genomic data analysis pipeline, comparisons are made between the pedigree and clinical data supplied and the corresponding information inferred from the genomic sequencing data, particularly to confirm that the sex and family relationships are as expected. These checks are performed by calculating the relative coverage of the sex chromosomes, identity by descent genotype sharing between family members and the number of mendelian inconsistencies per chromosome (where appropriate).

In the test order system, there are 3 relevant fields relating to the reported sex:

Field	Mandatory	Description
Gender	✓	Used to infer phenotypic sex if other fields left blank
Phenotypic sex		Should be completed if different from gender
Karyotypic sex		Should be completed if unusual or discordant from phenotypic sex or gender

In some cases where discrepancies are observed between the reported sex(es) and that inferred from the genomic data, data will pass through the Genomics England Interpretation Pipeline and a flag will be displayed in the Interpretation Portal (see Referral Flags). A separate flag will be displayed if a sex chromosome aneuploidy (also known as minor sex karyotype) is predicted from the genomic data.

Since the DRAGEN algorithms utilise the expected ploidy of the X and Y chromosomes (based on the inferred sex karyotype) in variant detection, erroneous genotypes for variants on the sex chromosomes may be observed for individuals with some minor sex karyotypes and tiering of variants on the X chromosome may be compromised (see Small Variant Tiering). The reported and inferred sex for each individual is displayed in the Interpretation Portal, along with the number of X chromosomes used for analysis. If further detail is required for a specific flagged case, a Jira service desk ticket should be raised by NHS GLH staff and a response will be provided by nhs.net email to an approved recipient. GLH staff may be contacted in the rare event that the inferred sex karyotype is ambiguous.

Flag	Description
UNUSUAL SEX KARYOTYPE	Applies when at least one member of the family has a sex karyotype that is not XX or XY
INCORRECT OR DISCORDANT SEX KARYOTYPE	Applies when the reported karyotypic and phenotypic sex or gender are discrepant but the karyotypic sex is supported by the sex inferred from the genomic data

Flag	Description
INFERRED GENETIC AND REPORTED SEX DISCORDANT	Applies when the reported sex is discrepant from the inferred genetic sex and the Disorders of Sexual Development panel has been applied, or the GLH has confirmed that the discordance is not due to a data entry error
UNKNOWN PHENOTYPIC SEX	Applies when at least one member of a family has an unknown phenotypic sex

If there is discrepancy between the reported sex and the inferred genetic sex and the Disorders of Sex Development panel has not been applied, queries will be raised with NHS GLHs to confirm or correct the phenotypic information prior to the sample proceeding through the interpretation pipeline.

Queries may also be raised when the expected relationships between family members are not supported by the genomic data. These queries will be returned in the DQ report (and in future in the MI portal). In the event of a complex inconsistency, such as a sex discrepancy or misattributed relationship, the DQ report will indicate only the query category and specific details will be sent by nhs.net email to a nominated address.

12 SNP identity checks (Sample Matching Service)

The Sample Matching Service has been deprecated in the Rare Disease pipeline.

Please refer to the Sample Matching Service online help page for further guidance.

13 Quality control

Genomic sequencing data are subject to a series of QC checks performed by the Genomics England automated pipeline to ensure they are of sufficient quality and are suitable for processing.

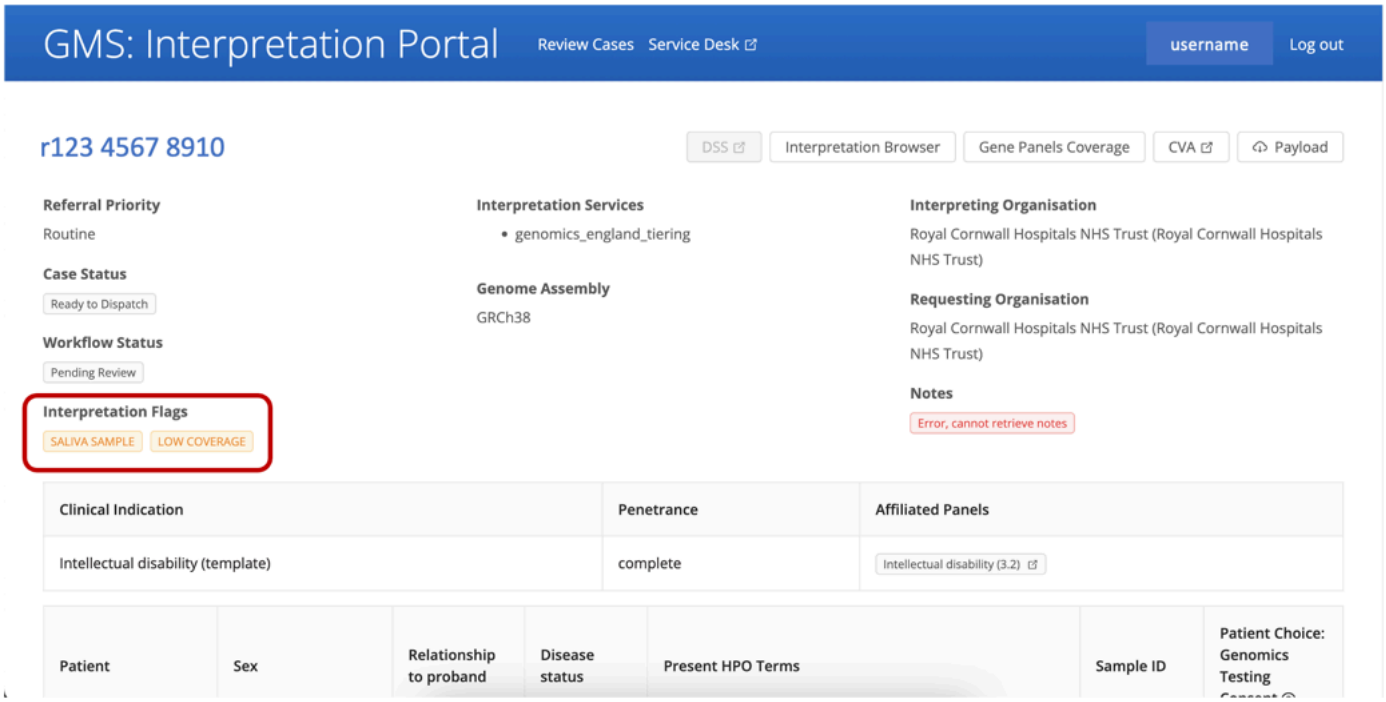
The following QC checks are completed as part of the pipeline:

Quality control	Description
Data integrity	md5sum check to confirm integrity of the genomic data transferred from the sequencing provider
Genome coverage	Alignments must cover at least 95% of the reference genome at 15x or above. Coverage will be calculated using bases from reads with mapping quality >10 and only counting sequences that remain as output by Illumina's current aligner after: - removal of overlapping bases - removal of duplicated reads - trimming adaptors - quality trimming ends of reads - clipping semi-aligned reads *(Note: Saliva samples are exempt from this check)*
Base quality	Sequence data for each sample will contain more than 85x10^9 bases with quality >=30. This threshold will be met by reads that are not duplicated and will not double count overlapping bases in the same fragment, after adaptor trimming and quality trimming.
Contamination	Germline cross-sample contamination performed using VerifyBamID. Samples with >3% contamination are considered as failing.

Note

- *Adaptor trimming:* when adaptor sequences are found at the end of the reads they are clipped. These sequences can be specified on the command line or in the sample sheet.
- *Quality trimming:* when the average base quality at the 3' end of the read is below a given threshold (15) the end of the read is trimmed.
- *Semi-aligned read clipping:* when a large number of mismatches accumulate at the end of a read, possibly indicating an indel or a structural variant, the mismatching end of the read is soft clipped.

Samples not passing these criteria are reported to the NHS GLHs via the Sample Failures Report. Saliva-derived DNA samples are exempt from the minimum coverage requirement and a flag `LOW_COVERAGE` will be displayed in the Interpretation Portal for any sample which does not pass the coverage QC metric (indicated in red box in image below).



14 Case flags in the CIP-API and Interpretation Portal

A case may have one or more of the following flags applied:

Flag	Description
`LOW_COVERAGE`	<95% of the autosomal genome covered at ≥15x calculated from reads with mapping quality >10 after filtering reads as described in Genome coverage. Samples with this quality flag will have >85x10^9 bases with quality ≥30, achieved from reads that are not duplicated and without double counting overlapping bases in the same fragment.
`SALIVA_SAMPLE`	Genome sequencing performed using DNA extracted from saliva. The increased risk of baceterial contamination in saliva samples can reduce genome sequencing quality, resulting in reduced quality associated with real genomic variants and/or an increased number of false positive variants detected.
`UNUSUAL_SEX_KARYOTYPE`	At least one member of the family has a sex karyotype that is not XX or XY.
`INCORRECT_OR_DISCORDANT_SEX_KARYOTYPE`	Reported karyotypic and phenotypic sex are discrepant but the karyotypic sex is supported by the sex inferred from the genomic data in at least one family member.
`INFERRED_GENETIC_AND_REPORTED_SEX_DISCORDANT`	Reported sex is discrepant from the inferred genetic sex and the Disorders of Sex development panel has been applied.
`UNKNOWN_PHENOTYPIC_SEX`	Phenotypic sex is unknown for at least one member of a family.
`POOR_QUALITY_CNV_CALLS`	Majority of CNV calls in the proband are expected to be of poor quality.
`SUSPECTED_POOR_QUALITY_CNV_CALLS`	An increased number of poor quality CNV calls is suspected.
`ddddddddd[mat|pat]UPDnn[i|h|m][c|p]`	Uniparental disomy detected that segregates with disease. Refer to uniparental disomy section for further details on interpretation of the details provided in this case flag.

III. Variant prioritisation approaches

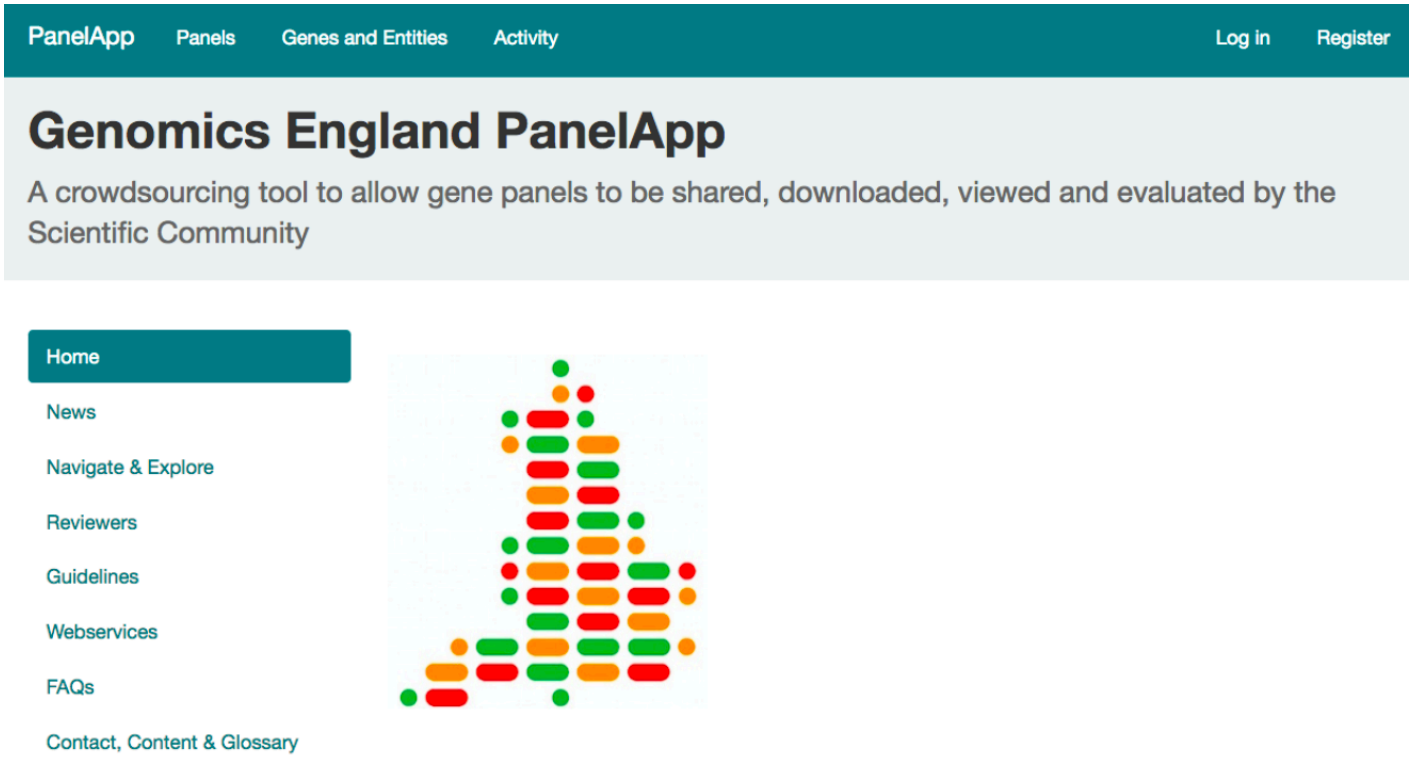
15 Pre-interpretation review and virtual gene panel assignment

Specification of analysis parameters including penetrance settings and panel assignment is carried out by GLH or Ordering Entity staff in the NGIS test order system. For each clinical indication, a default virtual panel and penetrance setting are attributed, but these can be modified according to local SOPs. Please see the Transcribing Whole Genome Sequence Test Requests in the NGIS User Guide for details of how to set these parameters. If the penetrance setting, disease status or panel assignment have been set incorrectly, please consult the Cancelling or Updating a Test Request or Patient Record in the NGIS SOP document for guidance.

III.I PanelApp

16 Overview

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 (see PanelApp API for more details). As described in PanelApp criteria for diagnostic grade 'green' genes, the diagnostic-grade 'Green' genomic entities (genes, STRs and regions e.g., CNVs), and their modes of inheritance in virtual gene panels are used to direct the Tiering process.



17 Use of virtual gene panels

Panels used for whole genome sequencing indications in the GMS will be denoted by the panel type field 'GMS Rare Disease Virtual'. Other GMS test types have the panel type ‘GMS Rare Disease’. For the GMS, consensus gene panels are finalised through a review process with a disease specialist test group and only signed-off panels are used for analysis. Signed-off panels and associated versions are available in PanelApp. We encourage NHS GLH members to continue to contribute their expertise by reviewing genes on panels, adding new genes or evidence over time, which will then be assessed for periodic updates.

18 PanelApp criteria for diagnostic grade 'green' genes

One of the three criteria A, B or C below must be met.

Criterion A

There are plausible disease-causing mutations[1] within, affecting or encompassing an interpretable functional region of this gene[2] or identified in multiple (>3) unrelated cases/families with the phenotype[3].

Criterion B

There are plausible disease-causing mutations[1] within, affecting or encompassing cis-regulatory elements convincingly affecting the expression of a single gene identified in multiple (>3) unrelated cases/families with the phenotype[3].

Criterion C

As Critera A or B but in 2 or 3 unrelated cases/families with the phenotype, with the addition of convincing bioinformatic or functional evidence of causation e.g., known inborn error of metabolism with mutation in orthologous gene which is known to have the relevant deficient enzymatic activity in other species; existence of an animal model which recapitulates the human phenotype.

Both criteria D and E must be met.

Criterion D

Evidence indicates that disease-causing mutations follow a Mendelian pattern of causation appropriate for reporting in a diagnostic setting[4].

Criterion E

No convincing evidence exists or has emerged that contradicts the role of the gene in the specified phenotype.

---

1. Plausible disease-causing mutations: Recurrent de novo mutations convincingly affecting gene function. Rare, fully-penetrant mutations - relevant genotype never, or very rarely, seen in controls. ↩↩
2. Interpretable functional region: ORF in protein coding genes miRNA stem or loop. ↩
3. Phenotype: the rare disease category, as described in the eligibility statement. ↩↩
4. Intermediate penetrance genes should not be included. ↩

III.II Small variant tiering

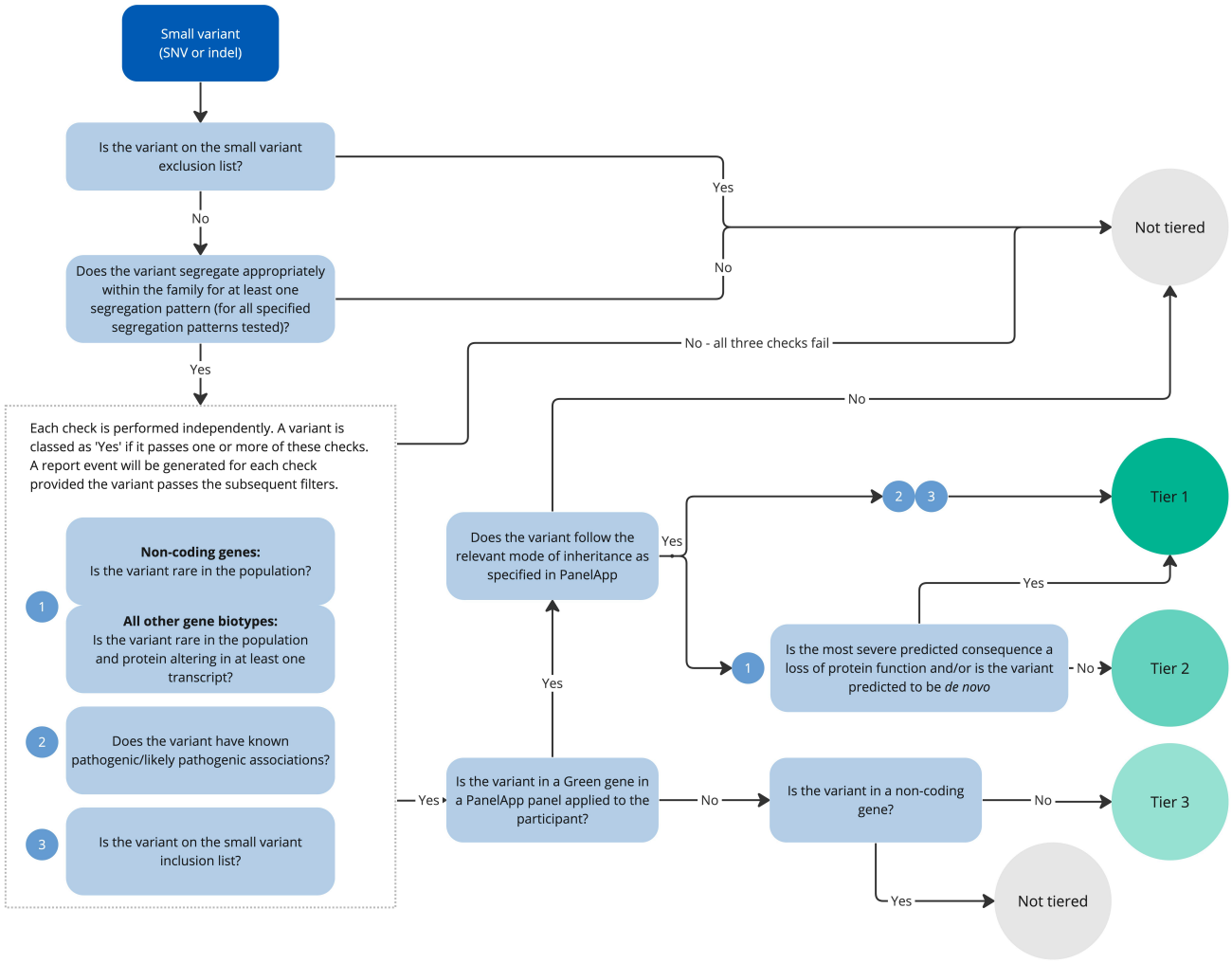
19 Overview

The Genomics England Rare Disease SNV and Indel Tiering Process is designed to aid NHS GLH evaluation of Rare Disease primary finding results by annotating variants that are plausibly pathogenic, based on their segregation in the family, frequency in control populations, predicted impact on the relevant protein(s), whether they have known pathogenic/likely pathogenic associations with disease and whether they are in a gene in a virtual panel(s) applied in the analysis incorporating the associated mode of inheritance. The process is summarised below.

Tiering can be performed in two penetrance modes:
1. Variants to be reported under ***complete penetrance***
2. Variants to be reported under ***incomplete penetrance***

After Tiering, variants are annotated with a tier (*Tier 1, Tier 2, Tier 3*) and a penetrance flag (*Complete* or *Incomplete*) to indicate the penetrance mode under which they were tiered. Incompletely penetrant variants are only reported if requested in the test order.

During the Tiering process, variants (detected and normalised by the Rare Disease Pipeline) are annotated and passed through multiple filters (population allele frequency, consequence type, segregation, quality etc.) in order to prioritise those that are potentially relevant/causal for a specific case and disease. The Genomics England Rare Disease Interpretation Pipeline annotates and reports small variants that have “PASS” filter status assigned - the current DRAGEN software implemented in the Rare Disease pipeline uses machine learning recalibration of variant quality scores, with a threshold for "PASS" filter status (using default quality score thresholds).



Simplified overview of the small variant tiering process. *Note: if a variant satisfies criteria for several tiers (e.g. Tier 1 and Tier 2), multiple records are written to the interpreted genome output file, but the highest tier is assigned to the variant for analysis in the interpretation portal.*

20 Tiers

Variants of potential relevance to the patient's clinical presentation will be automatically categorised into one of three tiers:

20.1 Tier 1

Rare variants that impact 'Green' genes (see PanelApp) in the applied gene panel, that are:

- high impact variants (e.g., likely loss-of function) that impact 'Green' genes in the applied gene panel;
- *de novo* moderate impact variants (e.g., missense);
- *de novo* variants in non-coding genes; or
- variants with known pathogenic or likely pathogenic disease associations

> **Note**
>
> *de novo* moderate impact variants (e.g. missense) are currently assigned to Tier 2 if the variant: (1) is part of a group of variants in a gene associated with a biallelic mode of inheritance (i.e. compound heterozygous segregation logic), and; (2) fulfils other prioritisation criteria (e.g. appropriate allele frequency), and; (3) has not been previously assigned as pathogenic or likely pathogenic status in the data resources currently considered in the rare disease tiering algorithm

20.2 Tier 2

Rare variants that impact 'Green' genes (see PanelApp) in the applied gene panel, that are:

- moderate impact variants (e.g., missense) in a gene biotype considered for tiering; or
- a non coding exon transcript variant in a non-coding gene

20.3 Tier 3

Plausible candidate variants identified with high or moderate impact or with known pathogenic/likely pathogenic disease associations in genes OUTSIDE those included in the analysis panel(s).

Caution should be used during clinical assessment and interpretation. Although most tier 3 variants will NOT be pathogenic, sometimes the causal variant will lie within tier 3. This could occur because there is insufficient evidence to support the inclusion of the gene within the relevant panel(s) at the time of analysis, or because the relevant panel was not applied.

Note

Occasionally a diagnostic variant may not be tiered, for example if the segregation pattern (disease status) provided in the test order tool for sequenced family members is inconsistent with the segregation pattern of variant(s).

Note

Variants in non-coding genes outside of the applied gene panel are not included in Tier 3 variants.

21 Tiering algorithm

A simplified view of the approach taken during tiering is included in the Overview. The Tiering algorithm considers variants or groups of variants in relation to eight criteria, each of which is further elaborated in separate sections:

1. FILTER status
2. Population frequency
3. Predicted functional impact
4. Segregation with disease in the recruited family
5. (a) Intersection with a high-evidence Green gene on the specified gene panel(s) AND (b) match with the curated mode of inheritance, as detailed in PanelApp
6. Known pathogenic/likely pathogenic disease associations
7. Internal and/or stakeholder knowledgebase of variant pathogenicity (small variant inclusion list)
8. Internal and/or stakeholder knowledgebase of variant benignity or artefacts

The table below summarises the relationship between the eight criteria and the assigned Tier groups:

Assumption	Tier 1	Tier 2	Tier 3	Untiered
If a variant or group of variants does not pass all of criteria 1-4				✅
If a variant (group) passes all criteria 1-4 but does not pass 5(a)			✅	
If a variant (group) passes all of criteria 1-4 and 5(a) but not 5(b)				✅
If a variant (group) passes all of criteria 1-5 and if predicted high impact consequence type OR a *de novo* variant with predicted functional impact (high, moderate, or in a non-coding gene)	✅			
If a variant (group) passes all of criteria 1-5 and if not predicted high impact consequence type (excluding de novo variants)		✅		
If a variant (group) passes criteria 1, 4, 5 and 6	✅			

Assumption	Tier 1	Tier 2	Tier 3	Untiered
If a variant (group) passes criteria 1, 4 and 6 but not 5(a)			✅	
If a variant (group) passes criteria 1, 4, 5(a) and 6 but not 5(b)				✅
If a variant (group) passes criteria 1, 4, 5 and 7	✅			
If a variant (group) passes criteria 1, 4 and 7 but not 5(a)			✅	
If a variant (group) passes criteria 1, 4, 5(a) and 7 but not 5(b)				✅
If a variant (group) passes criteria 8				✅

Variants that fail criteria 1-4, or pass critera 1-4 and are in a green gene but fail 5a will be `Untiered`

The Tiering pipeline analyses any family structure (by organising participants in trios of mother, father and offspring), regardless of the complexity of pedigree. All the trios must pass the defined filters. Where a family trio cannot be constructed, subsets are considered, such as parent-child pairs.

Where multiple gene panels have been assigned to a family, Tiering is performed independently using each panel.

22 Filter status

Only variants assigned PASS status in the FILTER column of the VCF file of variant calls are eligible to be classified as Tier 1, 2 or 3.

23 Population frequency

23.1 Nuclear genome

A population frequency filter is applied to prevent common variants being prioritised. For a variant to pass this filter, the population frequency cannot exceed **any** of the thresholds applied for the mode of inheritance being considered. Variants for which there is no allele population provided in the particular data set and population combination are considered zero.

The datasets used for allele frequency filtering are:

- gnomAD genomes v3.1.2
- gnomAD exomes v2.1.1
- Genomics England allele frequencies

The current GRCh38 allele frequency thresholds for the following populations are as follows:

Dataset tag	Population	Dataset size (individuals)	Dominant inherited disease	Recessive inherited disease
GNOMAD_EXOMES	AMR	17,296	0.001	0.01
GNOMAD_EXOMES	ASJ	5,040	0.001	0.01
GNOMAD_EXOMES	EAS	9,157	0.001	0.01
GNOMAD_EXOMES	FIN	10,824	0.001	0.01
GNOMAD_EXOMES	NFE	56,855	0.001	0.01
GNOMAD_EXOMES	SAS	15,308	0.001	0.01
GNOMAD_GENOMES	AFR	20,744	0.001	0.01
GNOMAD_GENOMES	AMI	456	0.100	0.10
GNOMAD_GENOMES	AMR	7,647	0.001	0.01
GNOMAD_GENOMES	ASJ	1,736	0.003	0.01
GNOMAD_GENOMES	EAS	2,604	0.002	0.01

Dataset tag	Population	Dataset size (individuals)	Dominant inherited disease	Recessive inherited disease
GNOMAD_GENOMES	FIN	5,316	0.001	0.01
GNOMAD_GENOMES	MID	158	0.100	0.10
GNOMAD_GENOMES	NFE	34,029	0.001	0.01
GNOMAD_GENOMES	SAS	2,419	0.002	0.01
GEL_aggCOVID_DRAGENv4.0-20230921 (internal ref: *20230921-aggDRAGENv4.0_COVID_v1.1-AFgt0*)	Genomics England custom frequencies	5,415	0.001	0.01

Note

gnomAD frequencies are extracted from gnomAD genomes v3.1.2 and gnomAD exomes v2.1.1. Variants present in gnomAD can receive flags that impact the variant's annotation and/or confidence - further information on possible flags is available through the gnomAD website. Variant frequencies in gnomAD are considered in the Rare Disease tiering pipeline regardless of the flags present.

Note

Several populations considered during variant tiering in earlier versions of the variant tiering pipeline are no longer considered, including: UK10K, 1000 Genomes Phase 3 and DiscovEHR

23.2 Mitochondrial genome

From the Orion NGIS release onwards (see release dates), there is consideration of mitochondrial allele frequencies in gnomAD during variant tiering. This includes consideration of homoplasmic or near-homoplasmic variants (95-100% allele fraction) in gnomAD, and exclusion of some genomic sequencing datasets that were included in the nuclear genome datasets for the same release. More details of the cohort composition and allele frequency data generation are available through the gnomAD webpage.

The datasets used for allele frequency filtering are:

- gnomAD genomes v3.1.2

The current GRCh38 allele frequency thresholds for the following populations are as follows:

Dataset tag	Population	Dataset size (individuals)	Mitochondrial genome inherited disease
GNOMAD_MT	AFR	14,347	0.001
GNOMAD_MT	AMI	392	0.1
GNOMAD_MT	AMR	5,718	0.001
GNOMAD_MT	ASJ	1,415	0.003
GNOMAD_MT	EAS	1,482	0.002
GNOMAD_MT	FIN	4,892	0.001
GNOMAD_MT	NFE	25,849	0.001
GNOMAD_MT	SAS	1,493	0.002
GEL_aggCOVID_DRAGENv4.0-20230921 (internal ref: *20230921-aggDRAGENv4.0_COVID_v1.1-AFgt0*)	Genomics England custom frequencies	5,415	0.001

24 Predicted functional impact

In order for a variant to pass this filter, it must have specific predicted functional impacts: - For protein-coding genes, variants must have high or moderate functional coding impact - For non-coding genes, specific functional impact terms are considered

The tables below lists the Sequence Ontology (SO) terms that are considered during variant tiering. A list of all possible SO terms can be found at the Sequence Ontology homepage.

24.1 High Impact

Variants with high impact sequence ontology terms will be prioritised as Tier 1 variants, providing that they pass the other variant filtering criteria.

Sequence ontology term	Definition	Synonyms
SO:0001893	A feature ablation whereby the deleted region includes a transcript feature.	Jannovar:transcript_ablation, transcript ablation, VEP:transcript_ablation
SO:0001574	A splice variant that changes the 2 base pair region at the 3' end of an intron.	Jannovar:splice_acceptor_variant Seattleseq:splice-acceptor snpEff:SPLICE_SITE_ACCEPTOR splice acceptor variant VAAST:splice_acceptor_variant VEP:splice_acceptor_variant
SO:0001575	A splice variant that changes the 2 base pair region at the 5' end of an intron.	Jannovar:splice_donor_variant Seattleseq:splice-donor snpEff:SPLICE_SITE_DONOR splice donor variant VAAST:splice_donor_variant VEP:splice_donor_variant
SO:0001587	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened polypeptide.	Seattleseq:stop-gained-near-splice stop codon gained ANNOVAR:stopgain Jannovar:stop_gained nonsense nonsense codon Seattleseq:stop-gained

Sequence ontology term	Definition	Synonyms
		snpEff:STOP_GAINED stop gained VAAST:stop_gained VAT:prematureStop VEP:stop_gained
SO:0001589	A sequence variant which causes a disruption of the translational reading fram, because the number of nucleotides inserted or deleted is not a multiple of three.	ANNOVAR:frameshift block substitution ANNOVAR:frameshift substitution Seattleseq:frameshift-near-splice VAT:deletionFS VAT:insertionFS frameshift variant frameshift_ frameshift_coding Jannovar:frameshift_variant Seattleseq:frameshift snpEff:FRAME_SHIFT VAAST:frameshift_variant VEP:frameshift_variant
SO:0001578	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	Seattleseq:stop-lost-near-splice ANNOVAR:stoploss Jannovar:stop_lost Seattleseq:stop-lost snpEff:STOP_LOST stop codon lost stop lost VAAST:stop_lost VAT:removedStop VEP:stop_lost
SO:0001582	A codon variant that changes at least one base of the first codon of a transcript.	snpEff:NON_SYNONYMOUS_START initiatior codon variant initiator codon change Jannovar:initiator_codon_variant VAT:startOverlap
SO:0002012	A codon variant that changes at least one base of the canonical start codon.	Jannovar:start_lost snpEff:START_LOST VEP:start_lost

24.2 Moderate impact

Variants with moderate impact sequence ontology terms will be prioritised as Tier 2 variants, providing that they pass the other variant filtering criteria.

Sequence ontology term	Definition	Synonyms
SO:0001889	A feature amplification of a region containing a transcript	transcript amplification, VEP:transcript_amplification
SO:0001821	An inframe non synonymous variant that inserts bases into in the coding sequence.	inframe codon gain, ANNOVAR:nonframeshift insertion, inframe increase in CDS length, inframe insertion, inframe_codon_gain, Jannovar:inframe_insertion, snpEFF:CODON_INSERTION, VAT:insertionNFS, VEP:inframe_insertion, SO:0001651
SO:0001822	An inframe non synonymous variant that deletes bases into in the coding sequence.	inframe codon loss, inframe deletion, snpEff:CODON_DELETION, ANNOVAR:nonframeshift deletion, inframe decrease in CDS length, inframe_codon_loss, Jannovar:inframe_deletion, VAT:deletionNFS, VEP:inframe_deletion, SO:0001652
SO:0001583	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved.	ANNOVAR:nonsynonymous SNV, Seattleseq:missense-near-splice, VAAST:non_synonymous_codon, Jannovar:missense_variant, missense, missense codon, Seattleseq:missense, snpEff:NON_SYNONYMOUS_CODING, VAAST:missense_variant, VAT:nonsynonymous, VEP:missense_variant, SO:0001584, SO:0001783
SO:0001630	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron.	ANNOVAR:splicing, snpEff:SPLICE_SITE_BRANCH, snpEff:SPLICE_SITE_BRANCH_U12, Jannovar:splice_region_variant, snpEff:SPLICE_SITE_REGION, splice region variant, VAAST:splice_region_variant, VEP:splice_region_variant
SO:0001626	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed.	incomplete terminal codon variant, partial_codon, VEP:incomplete_terminal_codon_variant

24.3 Non-coding variant impacts

Rare variants in non-coding genes with relevant sequence ontology terms will be prioritised as Tier 2 variants, providing that they pass the other variant filtering criteria.

Sequence ontology term	Definition	Synonyms
SO:0001792	A sequence variant that changes non-coding exon sequence in a non-coding transcript.	Seattleseq:non-coding-exon-near-splice, ANNOVAR:ncRNA_exonic, Jannovar:non_coding_transcript_exon_variant, non coding transcript exon variant, non_coding_transcript_exon_variant, Seattleseq:non-coding-exon, snpEff:non_coding_exon_variant, VEP:non_coding_transcript_exon_variant

24.4 Transcript biotypes

Consequence type is considered relative to the set of GENCODE Basic transcripts on Ensembl version 90 (GRCh38) that are associated with certain biological significance (biotype) categories. All GENCODE basic transcripts associated with the gene are evaluated. The biotypes considered are listed below.

Biotype	Description
IG_C_gene IG_D_gene IG_J_gene IG_V_gene TR_C_gene TR_D_gene TR_J_gene TR_V_gene	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
protein_coding	Contains an open reading frame (ORF).
nonsense_mediated_decay	If the coding sequence (following the appropriate reference) of a transcript finishes >50bp from a downstream splice site then it is tagged as NMD. If the variant does not cover the full reference coding sequence then it is annotated as NMD if NMD is unavoidable i.e. no matter what the exon structure of the missing portion is the transcript will be subject to NMD.

Biotype	Description
non_stop_decay	Transcripts that have polyA features (including signal) without a prior stop codon in the CDS, i.e. a non-genomic polyA tail attached directly to the CDS without 3' UTR. These transcripts are subject to degradation.
miRNA	*Non-coding gene biotype*: A small RNA (~22bp) that silences the expression of target mRNA.
lncRNA	*Non-coding gene biotype*: Transcripts that are long intergenic non-coding RNA locus with a length >200bp. Requires lack of coding potential and may not be conserved between species.
snRNA	*Non-coding gene biotype*: Small RNA molecules that are found in the cell nucleus and are involved in the processing of pre messenger RNAs.
snoRNA	*Non-coding gene biotype*: Small RNA molecules that are found in the cell nucleolus and are involved in the post-transcriptional modification of other RNAs.
Mt_tRNA	*Non-coding gene biotype*: A transfer RNA, which acts as an adaptor molecule for translation of mRNA. Gene is present in the mitochondrial genome.

Note

From the NGIS Nembus release, the `lincRNA` biotype has been modified to `lncRNA` as a result of changes to the Ensembl database that is hosted in the CellBase version that is included in this release.

25 Segregation with disease

25.1 Segregation filters

In order to pass this filter, a variant or group of variants must pass at least one of the segregation filters considered. The segregation filters that are considered and their groupings into modes of inheritance are listed below:

Mode of inheritance	Segregation Filter
biallelic	SimpleRecessive CompoundHeterozygous UniparentalIsodisomy
monoallelic_not_imprinted	InheritedAutosomalDominant *de novo*
monoallelic_paternally_imprinted	InheritedAutosomalDominantPaternallyImprinted
monoallelic_maternally_imprinted	InheritedAutosomalDominantMaternallyImprinted
xlinked_biallelic	XLinkedSimpleRecessive XLinkedCompoundHeterozygous
xlinked_monoallelic	XLinkedMonoallelic *de novo*
mitochondrial	MitochondrialGenome

All segregation filters are considered, i.e., there is no attempt to exclude any mode of inheritance based on the pattern of disease that is observed in the family's pedigree.

In practice, for a gene with an autosomal recessive mode of inheritance, this means that where only one variant in a gene passes the tiering filters, the variant will not be tiered as it is not consistent with the mode of inheritance.

Note

From the NGIS Nembus release, copy number variants will also be considered for compound heterozygous variants across variant types, see compound heterozygous variants across variant types for additional details.

The segregation filters are described in greater detail below.

25.2 Intersection with high evidence gene on specified gene panel and matching curated mode of inheritance

In order to be prioritised as Tier 1, a variant must be located in a gene whose association with the disorder being considered has been curated as high evidence (‘green’ or diagnostic grade) in the PanelApp panel applied.

The variant must also pass a segregation filter that is consistent with the curated mode of inheritance in PanelApp for that gene-disease association.

The nomenclature for modes of inheritance differs between PanelApp and tiering. The table below details which Tiering mode of inheritance would be considered appropriate for the different PanelApp modes of inheritances.

Tiering mode of inheritance	**PanelApp modes of inheritance**
biallelic	biallelic monoallelic_and_biallelic monoallelic_and_more_severe_biallelic not_provided unknown
monoallelic_not_imprinted	monoallelic_not_imprinted monoallelic monoallelic_and_biallelic monoallelic_and_more_severe_biallelic xlinked_biallelic xlinked_monoallelic mitochondrial not_provided unknown
monoallelic_paternally_imprinted	monoallelic_paternally_imprinted not_provided unknown
monoallelic_maternally_imprinted	monoallelic_maternally_imprinted not_provided unknown
xlinked_biallelic	xlinked_biallelic not_provided unknown

Tiering mode of inheritance	PanelApp modes of inheritance
xlinked_monoallelic	xlinked_monoallelic not_provided unknown
mitochondrial	mitochondrial not_provided unknown
de novo	monoallelic_not_imprinted monoallelic monoallelic_and_biallelic monoallelic_and_more_severe_biallelic monoallelic_paternally_imprinted monoallelic_maternally_imprinted xlinked_biallelic xlinked_monoallelic mitochondrial not_provided unknown

In PanelApp some panels are “superpanels”. This means that they contain a list of “subpanels” and they inherit all the gene-disease associations from all the panels in this list. In this situation it is possible that different subpanels may list the same gene with different modes of inheritance. In this situation, all applicable modes of inheritance are considered.

25.3 Segregation filters in action

To illustrate the principals of the segregation filters, an illustrative example is described below for a simple trio in a full penetrance analysis.

For each segregation filter, a number of individual filters are applied; variants are only tiered when all of these filters in each family member pass.

25.3.1 SimpleRecessive

Single sample filters	Single sample selection	Family filter
Affected samples are not 'reference_homozygous' or 'heterozygous'	At least one affected sample is 'alternate_homozygous'	Father and mother cannot be 'reference_homozygous'

Single sample filters	Single sample selection	Family filter
NonAffected samples are not 'alternate_homozygous'		

Note

Variants may be tiered under both `SimpleRecessive` and `CompoundHeterozygote` segregation filters if they fulfill both criteria. For example a rare alternate homozygous variant that impacts the same gene as a rare alternate heterozygous variant will be further considered in tiering under both segregation filters.

25.3.2 UniparentalIsodisomy

Single sample filters	Single sample selection	Family filter
Affected samples are not 'reference_homozygous' or 'heterozygous' NonAffected samples are not 'alternate_homozygous'	At least one affected sample is 'alternate_homozygous'	Father or mother (and only one of them) is 'reference_homozygous'

25.3.3 CompoundHeterozygote

Info

This filter is not applied when Tiering is performed with the incomplete penetrance mode. Each pair of variants in the gene are taken together for the family filter

Single sample filters	Single sample selection	Family filter[1]	Special Filter
Affected samples are not 'reference_homozygous' NonAffected samples are not 'alternate_homozygous'	At least one affected is 'heterozygous', 'alternate_hemizygous' or 'alternate_homozygous'	Father and mother are not both reference homozygous for the same variant in the pair, except where one or both variants in a proband are *de novo*	None of the NonAffected members of the family have the same combination of heterozygous or alternate homozygous variants for both variants included in the pair.

Note

Variants may be tiered under both `SimpleRecessive` and `CompoundHeterozygote` segregation filters if they fulfill both criteria. For example a rare alternate homozygous variant that impacts the same gene as a rare alternate heterozygous variant will be further considered in tiering under both segregation filters.

25.3.4 XLinkedSimpleRecessive

Single sample filters	Single sample selection	Family filter
Affected males are not 'reference_homozygous' or 'heterozygous' NonAffected females are not 'alternate_homozygous'	At least one affected sample is 'alternate_homozygous'	Mother must be 'heterozygous' (if mother is present), Father cannot be affected

25.3.5 XLinkedCompoundHeterozygote

Info

This filter is not applied when Tiering is performed with the incomplete penetrance mode. Each pair of variants in the gene are taken together for the family filter.

Single sample filters	Single sample selection	Family filter[1]	Special Filter
Affected samples are not 'reference_homozygous' NonAffected samples are not 'alternate_homozygous'	At least one affected female 'heterozygous' or 'alternate_homozygous'	Father and mother are not both reference homozygous for the same variant in the pair, except where one or both variants in a proband are *de novo*	None of the unaffected members of the family have the same combination of heterozygous or alternate homozygous variants for both variants included in the pair.

25.4 Monoallelic segregation filters

25.4.1 InheritedAutosomalDominant

Single sample filters	Single sample selection	Family filter
Affected samples are not 'reference_homozygous' NonAffected samples are not 'heterozygous' or 'alternate_homozygous'	At least one affected sample is 'alternate_homozygous' or 'heterozygous'	Father and mother cannot be 'reference_homozygous'

25.4.2 InheritedAutosomalDominantPaternallyImprinted

Info

Variants on the X chromosome and the mitochondrial genome are not considered under this mode of inheritance.

Single sample filters	Single sample selection	Family filter
Affected samples are not 'reference_homozygous'	At least one affected is 'alternate_homozygous' or 'heterozygous'	Father is not 'alternate_homozygous', 'heterozygous', if both parents unaffected Father is not 'alternate_homozygous', 'heterozygous', if father affected

Single sample filters	Single sample selection	Family filter
		Mother of unaffected participant (being unaffected 'heterozygous' or 'alternate_homozygous') is not 'alternate_homozygous', 'heterozygous', if both parents unaffected Mother of unaffected participant (being unaffected 'heterozygous' or 'alternate_homozygous') is not 'alternate_homozygous', 'heterozygous', if mother is affected

25.4.3 InheritedAutosomalDominantMaternallyImprinted

Info

Variants on the X chromosome and the mitochondrial genome are not considered under this mode of inheritance.

Single sample filters	Single sample selection	Family filter
Affected samples are not 'reference_homozygous'	At least one affected is 'alternate_homozygous' or 'heterozygous'	Mother is not 'alternate_homozygous', 'heterozygous', if both parents unaffected Mother is not 'alternate_homozygous', 'heterozygous', if mother affected Father of unaffected participant (being unaffected 'heterozygous' or 'alternate_homozygous') is not 'alternate_homozygous', 'heterozygous', if both parents unaffected Father of unaffected participant (being unaffected 'heterozygous' or 'alternate_homozygous') is not 'alternate_homozygous', 'heterozygous', if father is affected

25.4.4 XLinkedMonoallelicNotImprinted

Single sample filters	Single sample selection	Family filter
Affected samples are not 'reference_homozygous' NonAffected females are not 'alternate_homozygous' NonAffected males are not 'alternate_homozygous' or 'heterozygous	At least one affected sample is 'alternate_homozygous' or 'heterozygous'	Both parents are not 'reference_homozygous'

25.4.5 MitochondrialGenome

Info

The MitochondrialGenome Segregation Filter is only considered for variants in the mitochondrial genome.

Single sample filters	Single sample selection
Affected samples are not 'reference_homozygous'	Allele fraction ≥0.05 in affected individuals

25.4.6 deNovo

Info

The DeNovo Segregation Filter is considered independently of other segregation filers. The DeNovo Segregation Filter does not take into account the affected status of other family members. This does not affect biallelic genes but in genes with a monoallelic mode of inheritance can result in the prioritisation of variants that are absent in affected parents.

SNVs	Indels
DQ value ≥0.0013	DQ value ≥0.02

Note

Only *de novo* variants with quality (DQ) scores as indicated in the table above will be considered during the Genomics England variant tiering approach and displayed to users in the Interpretation Portal for further consideration. The DQ scores are not included in the displayed information to the user, but are present in the VCF file available for download.

Note

From the NGIS Nembus release, *de novo* small variants are considered as part of compound heterozygous pairs of variants, if they fulfill the quality criteria declared in the table above.

---

1. each pair of variants in the gene are taken together for the family filter ↩↩

26 Variants with pathogenic/likely pathogenic disease associations

This criterion enables data from external sources to be used to prioritise variants with known pathogenic/likely pathogenic associations that would not otherwise be prioritised. This also enables variants that do not occur in transcripts with a considered biotype (see Criterion 3) to be prioritised.

The population frequency filter (Criterion 2) or predicted functional impact filter (Criterion 3) may result in potentially diagnostic variants with known pathogenic/likely pathogenic associations not being prioritised, e.g. if they are known to be pathogenic but above allele frequency threshold used during variant tiering.

26.1 ClinVar

ClinVar aggregates information about genomic variation and its relationship to human health. An individual submission to ClinVar asserting a relationship between a variant and condition are represented by a SCV record. All submissions about the same variant are aggregated into a VCV record.

A variant passes Criterion 6 if all of the criteria outlined below are fulfilled:

1. The variant has an exact match with the genomic position and alternate allele of variant with a VCV record in ClinVar **OR** the variant has a HGVSp match[1] with a variant with a VCV record in ClinVar
2. The ClinVar VCV record has the following properties:

   > A clinical significance that is one of:
   > 1. Pathogenic
   > 2. Likely pathogenic
   > 3. Conflicting interpretations of pathogenicity
   >
   > A review status that is one of:
   > 1. Practice guideline
   > 2. Reviewed by expert panel
   > 3. Criteria provided, multiple submitters, no conflicts
   > 4. Criteria provided, conflicting interpretations[2]
   > 5. Criteria provided, single submitter
   > 6. No assertion criteria provided

3. The variant has a population frequency in the custom Genomics England frequencies dataset less than 0.05 (i.e. 5%)

> **Note**
>
> The ClinVar version utilised in the pipeline is available in software and database versions. The Rare Disease tiering pipeline does not consider ClinVar variants that are `Established risk alleles` or `drug response`.

26.2 Clinical Variant Ark (CVA)

CVA is a Genomics England database which stores information assigned to genomic variants for samples processed by Genomics England bioinformatics pipelines, including:

- variant classifications
- variant interpretation logs
- tiered variants
- clinical indications upon referral
- data reported in outcome questionnaires and summary of findings

Details of all the information available are further elaborated on the CVA online help page

A variant passes Criterion 6 if:

1. The variant has at least one pathogenic or likely pathogenic classification in CVA at the time of interpretation, or
2. A variant with the same protein change, based on HGVSp annotation for the same transcript, has at least one pathogenic or likely pathogenic classification in CVA at the time of interpretation
3. The variant must have a population frequency in the custom Genomics England frequencies dataset less than 0.05 (i.e. 5%)

> **Note**
>
> *HGVSp annotations are only available for variants classified in cases run following the NGIS Izar release in March 2023*

> **Note**
>
> Queries against the CVA database are done in real-time at the time of analysis; CVA is updated daily, therefore reanalysis at a later date can produce different results.

Note

Variant classifications in CVA are continually updated as GMS referrals are reviewed and findings returned. A variant with no pathogenic or likely pathogenic classifications in CVA at the point of interpretation will not pass criterion 6 for CVA even if the variant is subsequently classified as pathogenic or likely pathogenic before being reviewed. However, the variant will display the current CVA classification in the Interpretation Portal.

---

1. *Variants matched through exact (#1) or protein-matched (#2) logic will be displayed in the same way in the interpretation portal. Users must manually review these events to determine if they were prioritised through the known pathogenic tiering criterion by exact match or through protein-matching.* ↩
2. *Variants with a review status of "criteria provided, conflicting interpretations" must have at least one associated SCV record with a pathogenic/likely pathogenic assertion.* ↩

27 Variant inclusion list

Variants that are in the curated list of variants below will be prioritised if they are present, providing they pass at least the variant filter status filter (Criterion 1) and segregate with the disease (Criterion 4):

Variant	Gene	ClinVar Accession
chr1:94010911:T:A	ABCA4	VCV000099390
chr11:89284793:G:A	TYR	VCV000003779
chr18:57571588:A:G	FECH	VCV000000562

28 Variant exclusion list

Variants that are in the list of curated variants below will not be prioritised under any circumstances.

- No variants currently on the exclusion list

29 Penetrance modes

By default, Tiering is performed assuming complete penetrance and therefore any genotypes that are present in unaffected individuals would be excluded from Tiering.

Where incomplete penetrance analysis is selected, Tiering is performed first using the complete penetrance settings and then again under incomplete penetrance. If a tiered variant is annotated with a tier under the complete penetrance segregation filter, it will not also be tiered under an incomplete penetrance segregation filter.

In the incomplete penetrance analysis, genotypes must be present in all affected individuals but are not excluded if they are also present in unaffected individuals. Genotypes in unaffected individuals may still be used to check that genotype patterns are consistent with inheritance, e.g., for phasing of compound heterozygous variants.

Incomplete penetrance analysis does not currently consider the pattern of disease in the family's pedigree. If a disease skips generations in the pedigree, then it may be possible to deduce that particular unaffected family members should have the disease genotype. The Tiering process does not currently perform this deduction.

30 Additional notes

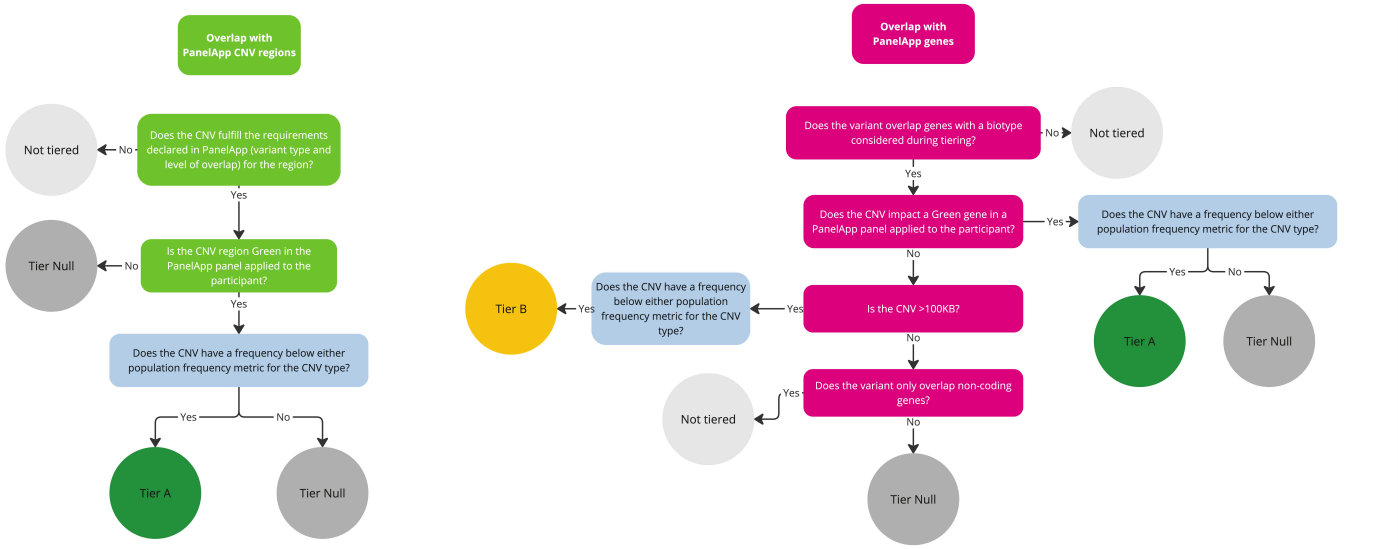
- Tiering is performed using a signed-off version of a panel, with the version most recently approved at the time of interpretation. The gene content of current and previous signed-off versions is available in PanelApp (see Use of virtual gene panels).
- Tiering of variants on chrX is performed according to an individual's inferred number of X chromosomes from coverage levels in the genomic data. The number of X chromosomes used in tiering will be the same as used in variant calling (see Genome build, alignment and variant detection), which is displayed in the Interpretation Portal.
- Tiering supports only haploid and diploid models, thus tiering of variants on chrX may be suboptimal for individuals with minor sex karyotypes with different X chromosome ploidy.
- A single heterozygous SNV or indel variant identified in a gene with a biallelic mode of inheritance will not be assigned a tier, but can be explored using the decision support system (see Decision Support Systems). Mode of inheritance is not used in CNV tiering, therefore all CNVs that impact genes are assigned a tier (see Copy number variant tiering).
- The Segregation Filter for *de novo* variants is considered independently of other segregation filters. The DRAGEN *de novo* variant detection algorithm considers all variants with a non-Mendelian inheritance pattern, and in rare cases, this can result in variants being inappropriately tiered. For example, a **homozygous variant** in a proband will be considered as *de novo* and tiered (with monoallelic mode of pathogenicity and complete penetrance) if it is **heterozygous** in one parent and **homozygous reference** in the other parent.
- The pipeline reports all the classified variants in a structured format and ignores missing values. For example, in a fully penetrant autosomal dominant setting, a variant would be tiered even if it were missing in one of the affected individuals if it passed all of the other necessary criteria.
- Input genotypes (from the normalised VCF file produced by the Rare Disease Pipeline) can be phased or unphased, but phase information is currently ignored. It is possible to view alignment files (CRAM) through igv.js, and manually assess phase for variants in close proximity.
- The Tiering algorithm does not prioritise small variants in pseudoautosomal regions, but CNVs in these regions can be prioritised.
- In a scenario of compound heterozygosity for a large deletion and a small variant, whereby the small variant is detected within the single copy region, tiering of the small variant may indicate prioritisation under the UniparentalIsodisomy mode (with the corresponding CNV likely being in Tier A).
- Variants on chromosomes other than 1 to 22, the X chromosome and the mitochondrial genome are not currently considered in Tiering, i.e., variants on the Y chromosome or on alternate and decoy contigs are not currently considered for tiering.

III.III Copy number variant tiering

31 Overview

Copy number variants (CNVs) are detected using DRAGEN CNV with self-normalisation and the Shifting Level 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 DRAGEN SV callers. CNVs in this range detected by both callers with a minimum reciprocal overlap of 50% are deemed to be high quality. The Genomics England Rare Disease Interpretation Pipeline currently annotates and reports all high quality CNVs that overlap with gene biotypes considered during tiering and are ≥ 2 kb in size. Annotations include internal allele frequencies calculated for 5,415 samples.

An overview of the CNV tiering pipeline can be seen below.



Note

There are two separate streams of analysis applied during tiering of CNVs (both use information available in PanelApp: (1) overlap of CNV regions, and (2) overlap of genes; as indicated by different colour boxes in diagram). The tiering diagram provided aims to assist with understanding of prioritisation approaches applied to CNVs and does not account for all ossible scenarios, please see text below for more information.

A CNV is considered under both prioritisation streams **independently** (PanelApp genes and PanelApp regions). It is possible that a CNV is appropriately prioritised and assigned multiple tiers. In this scenario, the final tier will be the highest assigned tier (see table), and is the final tier displayed in the Interpretation Portal. All assigned tiers are also displayed in variant report events information in the Interpretation Portal.

Please note that CNVs that are classified as “Not tiered” via one prioritisation stream will only be displayed in the Interpretation Portal if they receive a higher tier via the alternative CNV prioritisation stream.

Priority Rank	Tier
1	A
2	B
3	Null
-	Not tiered

Examples of CNVs that receive more than one tier include, but are not limited to:

- CNVs which impact multiple genes with appropriate biotypes, where at least one gene is Green on the applied gene panel and at least one gene is not on the applied gene panel **(Tier A and Tier Null)**
- CNVs which are >100Kb and overlap a PanelApp region and do not fulfil the requirements for that region (variant type, level of overlap), but do overlap a gene with an appropriate biotype that is not included in the applied gene panel **(Not tiered, Tier B and Tier Null)**

Only CNV calls from the proband are annotated and displayed. CNV calls in relatives are **not** currently considered in tiering. However, visual assessment of CNVs for all family members can be performed using coverage profiles displayed in the IGV viewer following the links from the Interpretation Portal.

Note

Mode of inheritance is not considered in CNV tiering. In contrast to small variant tiering, a single heterozygous CNV within or impacting a gene or region (with appropriate variant type and overlap) will be tiered regardless of the expected mode of inheritance recorded in PanelApp.

31.1 Tier A

A CNV is assigned Tier A if it satisfies the following criteria:

- The CNV must not exceed the frequency threshold for **both** of the CNV frequency metrics (CNV_AF, CNV_AUC). The thresholds are unique to the CNV type:
  - LOSS: 0.005
  - GAIN: 0.01

> **Note**
>
> A CNV can be included as Tier A if one of the CNV frequency metrics (e.g. CNV_AUC) is exceeded but the other is below the threshold (e.g. CNV_AF)

AND one or both of the following criteria:

- The CNV overlaps with a green gene in a panel applied in the analysis. All CNVs (i.e. loss and gain) overlapping any gene are considered, without requiring a minimum overlap threshold.
- The CNV overlaps a pathogenic region in a virtual 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.

> **Note**
>
> If a Tier A CNV impacts a gene with a biallelic mode of inheritance then the prioritisation algorithm will include assessment of compound heterozygous variants across variant types.

31.2 Tier B

A CNV is assigned Tier B if it satisfies the following criteria:

- The CNV must not exceed the frequency threshold for **both** of the CNV frequency metrics (CNV_AF, CNV_AUC). The thresholds are unique to the CNV type:
  - LOSS: 0.005
  - GAIN: 0.01
- The CNV overlaps a gene with a biotype included for consideration (see CNV biotypes table below)
- The CNV is larger than 100Kb

> **Note**
>
> A CNV can be included as Tier B if one of the CNV frequency metrics (e.g. CNV_AUC) is exceeded but the other is below the threshold (e.g. CNV_AF)

31.3 Tier Null

All PASS variants that overlap with a gene with a biotype included for consideration (see table below).

Note

- Tier A CNVs may also be classified as Tier B if they are >100kb in size
- Tier A CNVs may also be classified as Tier Null if they also impact genes that do not fulfil the criteria for Tier A but do fulfil the criteria for Tier Null
- All CNVs that are classified as Tier B will also have a Tier Null classification
- The highest reported tier will be considered the final tier for the CNV

31.4 CNV biotypes considered during variant tiering

CNVs will be considered during variant tiering if they impact genes with any of the following biotypes:

Biotype	Description
IG_C_gene IG_D_gene IG_J_gene IG_V_gene TR_C_gene TR_D_gene TR_J_gene TR_V_gene	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
protein_coding	Contains an open reading frame (ORF).
nonsense_mediated_decay	If the coding sequence (following the appropriate reference) of a transcript finishes >50bp from a downstream splice site then it is tagged as NMD. If the variant does not cover the full reference coding sequence then it is annotated as NMD if NMD is unavoidable i.e. no matter what the exon structure of the missing portion is the transcript will be subject to NMD.
non_stop_decay	Transcripts that have polyA features (including signal) without a prior stop codon in the CDS, i.e. a non-genomic polyA tail attached directly to the CDS without 3' UTR. These transcripts are subject to degradation.

32 Sample quality control

For a small proportion of samples, the sequencing data are not of sufficiently high quality to make reliable CNV calls. Sample level quality control is performed based on the number and ratio of different call types and the proportion of common CNVs detected. If CNV data for a proband do not pass this quality control step, the family is flagged in the Interpretation Portal with one of the following flags:

Flag	Summary	Metrics causing flag activation
`poor_quality_CNV_calls`	majority of the CNV calls in the sample are expected to be of poor quality	- count of autosomal PASS CNVs ≥ 980 **or** - Log2(Loss/Gain) < -2.2
`suspected_poor_quality_CNV_calls`	some of the CNV calls in the sample are expected to be of poor quality	- count of autosomal PASS CNVs ≥ 195 and < 980, or ≤ 121 **or** - Log2(Loss/Gain) ≥ -2.2 and ≤ -1.0, or ≥ 0.1 **or** - fraction of common[1] autosomal PASS CNV calls < 0.375

1. a CNV is defined as common if it has 50% reciprocal overlap with a CNV from Conrad et al. 2010. ↩

33 CNV frequency annotation

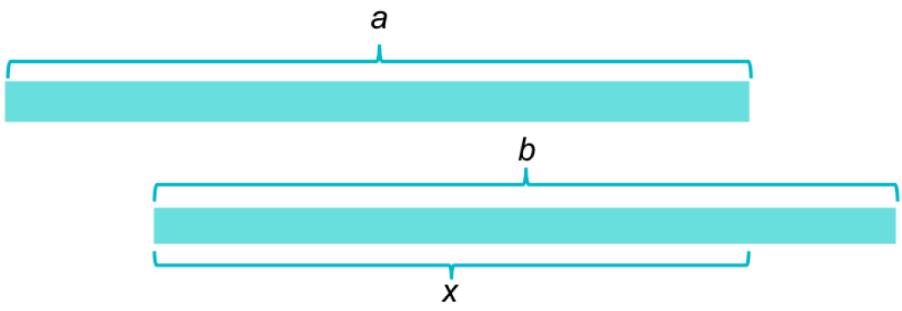
Several factors complicate the assessment of allele frequencies for copy number variants:

- 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 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 a tandem duplication with 3 copies on one allele and a deletion on the other allele, or two single copy alleles with an additional copy elsewhere in the genome.
- The accuracy of exact copy number inference for gains with more than 3 copies is not known.

Due to these issues, there is no single perfect method to calculate allele frequencies for CNVs. Therefore, we present two calculations through alternative strategies. CNV frequencies were calculated using 5,415 germline samples from unrelated individuals (participants in the Cancer program of the 100,000 Genomes Project and the COVID-19 research project)

33.1 Reciprocal overlap

In this method, CNV calls from the 5,415 reference samples are compared individually to CNVs in the sample as shown in a figure below, 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.

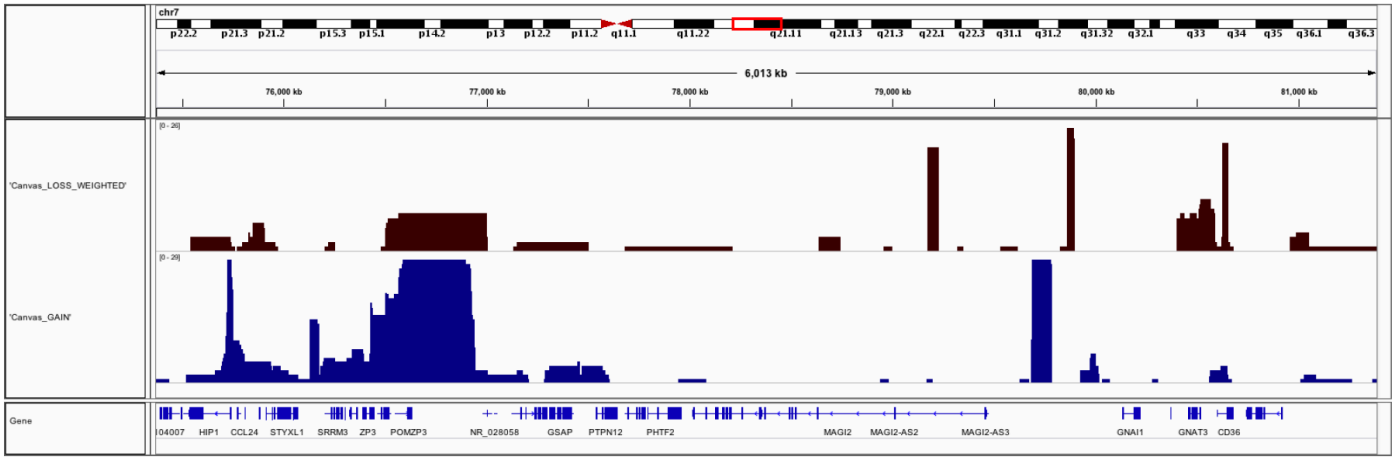


Variants are considered as being the same, if $x/a > threshold$ and $x/b > threshold$

33.2 Frequency track – area under the curve method

In this method, CNV calls from the 5,415 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.



33.3 Differences in LOSS and GAIN calculations

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.

34 Detection of CNVs between 2-10kb

Detection of CNVs between 2 and 10 kb was introduced to the Rare Disease pipeline on July 28th 2021 (NGIS Danny release). All analysis performed after this date includes this pipeline enhancement. For any given referral, the limitations described in the Summary of Findings will indicate if small CNV analysis was performed (see Pipeline limitations for further details).

Note

Detection of CNVs between 2 and 10kb in size is performed for probands only.

CNVs between 2 and 10 kb are identified using a combination of two different variant callers that utilise different signals to detect copy number variation, *DRAGEN CNV* and *DRAGEN SV*, which use read depth and anomalously mapped/split reads respectively.

CNVs <10 kb detected by DRAGEN CNV that are also supported by CNVs detected by DRAGEN SV with a minimum reciprocal overlap of at least 50% and with matching CNV type (deletion or duplication) are considered to be high quality CNVs and are subsequently included for annotation and reporting. For high quality CNVs, the filter status in the VCF file for the proband is set to PASS, and additional annotation is added to the VCF file including the CNV coordinates detected by DRAGEN SV (which are likely more accurate due to the use of split reads in CNV detection) and a flag to indicate a CNV is *de novo* where parental data are available.

The updated version of the CNV VCF file for the proband is renamed to ".enhanced.cnv.vcf.gz" and the following additional annotations are included in the INFO field:

Field	Description
`MANTA_SUPPORT`	(Flag) CNV supported by DRAGEN SV (reciprocal overlap >= 50%)
`MANTA_POS`	Coordinates of the overlapping DRAGEN SV call
`DN`	(Flag) De novo variant, based on DRAGEN SV joint calling
`DN_TYPE`	Type of de novo variant, based on DRAGEN SV joint calling (format: probandGT-fatherGT-motherGT)
`PREV_FILTER`	Original non-PASS filter in DRAGEN CNV VCF file before small CNV enhancement

> **Note**
>
> `MANTA_SUPPORT` and `MANTA_POS` flags are legacy names prior to the conversion of MANTA to DRAGEN SV

35 Measurement of uncertainty of CNV breakpoints

The DRAGEN CNV caller algorithm uses deviation in expected read depth in genomic regions (~1,000bp windows) to identify gains or losses in genomic content.

Due to the nature of this algorithm, the breakpoints reported by the pipeline are not an accurate representation of exact CNV breakpoints.

Our comparison of DRAGEN CNV breakpoints with the breakpoints obtained from more precise approaches (Manta software, which uses information from split reads and unproperly paired sequencing reads) demonstrated:

- For CNVs with the breakpoints in well mappable regions (i.e. where sequencing reads can be unambiguously aligned to genome reference), 86.3% of breakpoints determined by DRAGEN CNV are within 1000bp from the breakpoint determined by split or unproperly paired reads, and 94.3% are within 2000bp.[1]
- For CNVs where the breakpoint falls within a difficult to map region, e.g. a segmental duplication, the breakpoints cannot be unambiguously determined using short read sequencing technology, and the uncertainty depends on the size of the difficult to map region.

---

1. CNV events which are fragmented (i.e. multiple DRAGEN calls overlap with a continuous region that is lost or gained in the sample) will decrease measurement of uncertainty metrics. Fragmentation in DRAGEN calls can occur as a result of CNV events overlapping regions of the genome that are excluded from CNV calling, e.g. regions where read alignment is difficult. ↵

III.IV Short tandem repeats

36 Overview

Expansions of short tandem repeats (STRs) are detected by ExpansionHunter (v5) as part of the DRAGEN software. STRs are detected only at loci defined in PanelApp STRs.

STR expansions are only reported in affected participants. All repeat expansion VCF files are available to download, containing all loci analysed.

An up-to-date list and information about specific STR loci and the associated gene panels can be found in PanelApp. PanelApp information for each STR includes: - the genomic coordinates of the repeat analysed (both GRCh37 and GRCh38 assemblies) - the repeat motif or sequence (e.g. `CAG`) - the normal and pathogenic number of repeats associated with each locus.

The internal repeat thresholds have been reviewed and agreed by NHS clinical experts, and are an essential aspect of STR tiering.

> **Note**
>
> Some STR loci are green on some panels, and red on others.

Only STR loci that are green on a panel applied in the analysis and that follow the relevant mode of inheritance will be reported.

37 STR tiering

STRs are only tiered for affected members of a family.

When estimating repeat sizes, ExpansionHunter provides confidence intervals and an average for each allele (i.e., x-y and avg(x,y)). For increased sensitivity, the maximum value (i.e., y) of these estimations is taken for each allele and locus to assign tier.

> **Note**
>
> The average value is displayed in the Interpretation Portal, as that value is expected to be closer to the real allele size. On rare occasions, a Tier 1 or Tier 2 STR expansion (tiered due to higher confidence interval exceeding the relevant threshold) may have an average allele size below the threshold applied during tiering.

STR loci that are green in PanelApp for the panel(s) applied for analysis and have a `PASS` status after expansion detection will be tiered. Two different ranges of thresholds are used when tiering:

37.1 Tier 1

The repeat-length for the locus is greater than or equal to the pathogenic threshold.

37.2 Tier 2

The repeat-length is greater than or equal to the threshold for normal alleles but less than the pathogenic threshold.

37.3 Tier Null

The repeat-length is less than the threshold for normal alleles.

> **Note**
>
> In some rare instances, an STR locus may not receive a `PASS` status, due to the quality and the amount of informative sequencing reads available. In such instances the STR locus will not be considered for tiering in the Rare Disease Pipeline.

37.4 Unified tiering of STRs and small variants

For biallelic loci (e.g. *FXN*), affected individuals homozygous for a pathogenic expansion or compound heterozygous (STR and SNV) are also tiered. Please refer to compound heterozygous variants across variant types for additional details.

> **Note**
>
> Due to a limitation of the ExpansionHunter algorithm, in some cases, biallelic expansions (e.g. in *FXN* or *FMR1* in females) may be incorrectly detected as monoallelic expansions.

37.5 Special notes regarding FMR1

Full FMR1 expansions (>200 repeats) cannot be distinguished from pre-pathogenic expansions and therefore in majority of cases will be reported as pre-expansions in Tier 2.

Tiering of FMR1 was introduced in the NGIS Grace release. For the cases processed with the earlier version of tiering, that was not considering FMR1 despite it being green in PanelApp, a warning will be displayed in the Interpretation Portal notifying that FMR1 was not considered in tiering.

> **Note**
>
> Reanalysis of cases that were referred before the NGIS Grace release will not include FMR1 STR tiering as these expansions were not detected prior to Grace, and the reanalysis does not include mapping and variant detection approaches.

37.6 Special notes regarding NOP56

There is a known limitation of ExpansionHunter's performance for NOP56, where on some occasions it silently skips genotype calling for this locus when there are no reliably mapped reads due to locus complexity. This leads to missing NOP56 genotypes and therefore the testing for NOP56 should be considered as not performed. Clinical Scientists may want to caveat reports appropriately if there is no results for NOP56 displayed in the Interpretation Portal when they would be expected (i.e. STR locus is green on the panel applied).

38 Measurement of uncertainty for STR allele sizing

38.1 General recommendations

Based on the assessments of measurement of uncertainty described below, it is recommended that all expansion-positive results, as well as normal alleles close to the threshold (+/2 repeat units) and close to the sequencing read length (e.g. 49 repeats of trinucleotide repeat unit), if deemed as possibly clinically relevant, are assessed by reviewing STR pileup plots (see STR visualisation), and, if required, an orthogonal test is performed.

For detailed recommendations for each STR loci please refer to NHS England Guidelines for Rare Disease Whole Genome Sequencing & Next Generation Sequencing Panel Interpretation & Reporting.

The accuracy and measurement of uncertainty of STR repeat allele sizing by ExpansionHunter depends on overall allele length.

38.2 Alleles shorter than sequencing read length

For alleles shorter than a sequencing read length, i.e. where the whole repeat and some flanking regions on both sides can be spanned by a single sequencing read (i.e. <150bp), the estimate of allele size by ExpansionHunter is very accurate.

In comparison with the allele sizes obtained through standard of care testing for the same samples, 98% of the allele sizes from ExpansionHunter were within +/-2 repeat units of the sizes reported by the standard of care tests, and 92% within +/-1 repeat unit.

In many cases, manual inspection of sequencing reads through visualization software (REViewer plots) revealed that ExpansionHunter was likely to estimate the allele size more accurately than standard of care tests. In a small number of cases where the detected difference from the standard of care test was >2 repeat units, ExpansionHunter typically overestimated the allele size, suggesting that it is more likely to produce false positive than false negative calls.

38.3 Alleles longer than sequencing read length

When allele length exceeds a sequencing read length, the allele sizing accuracy of ExpansionHunter is reduced.

Validation against standard of care methods showed that clinically relevant STR expansions can still be detected with high sensitivity (see Appendix G for up to date values of STR detection performance), however estimates of exact repeat length are less accurate.

39 STR visualisation

STR results are displayed in the Interpretation Portal and reviewing this plot is fundamental to the process of assessing the quality of the repeat size estimations computed by ExpansionHunter. Genomics England strongly advises GLHs to use the visualization plot to assess the quality of each call before validation of expansions with an alternative method. It is also advised to review normal alleles of the sizes larger than the read length (e.g. >49 repeats for trinucleotide repeat loci) and normal alleles very close to the threshold (e.g. +/- 2 repeat unit)

Analysing the reads that ExpansionHunter considers when assessing the repeat lengths is essential for determining the quality of the call but also for characterising interruptions (e.g. for Spinocerebellar Ataxias) or pathogenic-borderline cases, before orthogonal confirmation.

Below are visualisation plots and scenarios to illustrate how Expansion Hunter estimates STR genotypes.

Note

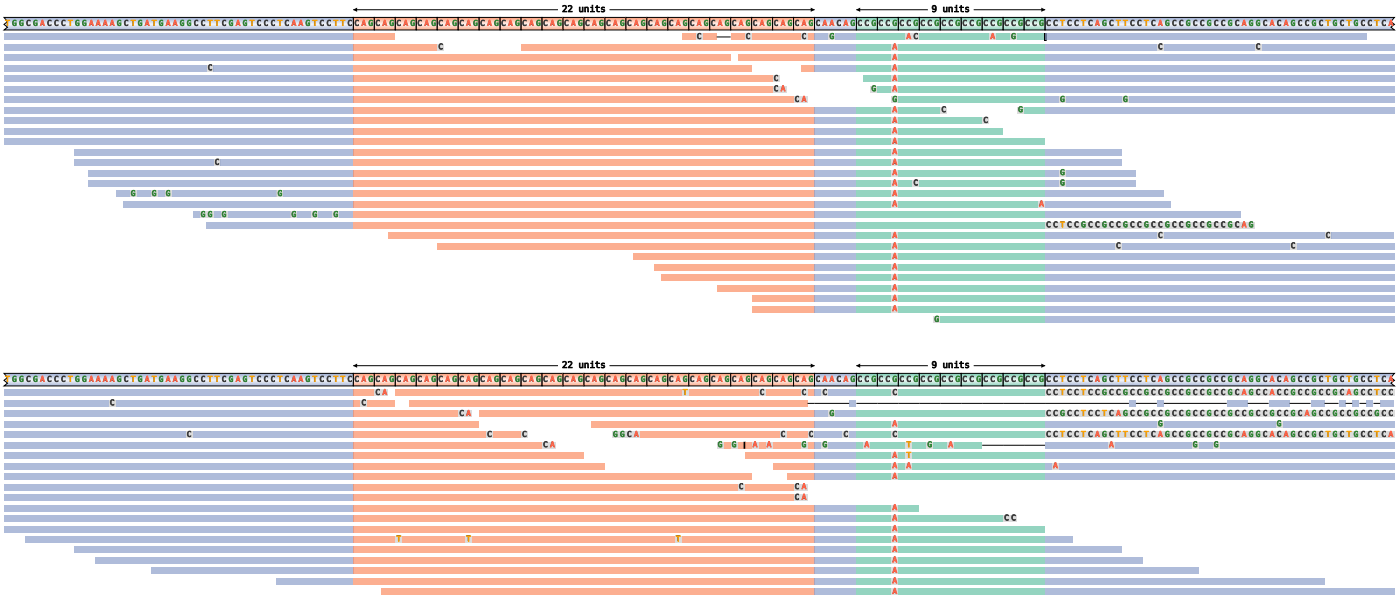
The software utilised to generate STR plots for DRAGEN v4.0.5 (NGIS Mira release) has been altered to REViewer. The output format from REViewer is `.svg` which can be downloaded and viewed in most web browsers.

Note

In some rare instances, a REViewer plot is not created for the loci because there are no reads aligned to the region of interest. In this scenario, the REViewer plot is replaced by text stating `No plot generated due to missing genotype. Please consult the repeats.vcf.gz file`. REViewer plots will be created for all other loci.

39.1 A good quality STR call showing alleles within the normal range

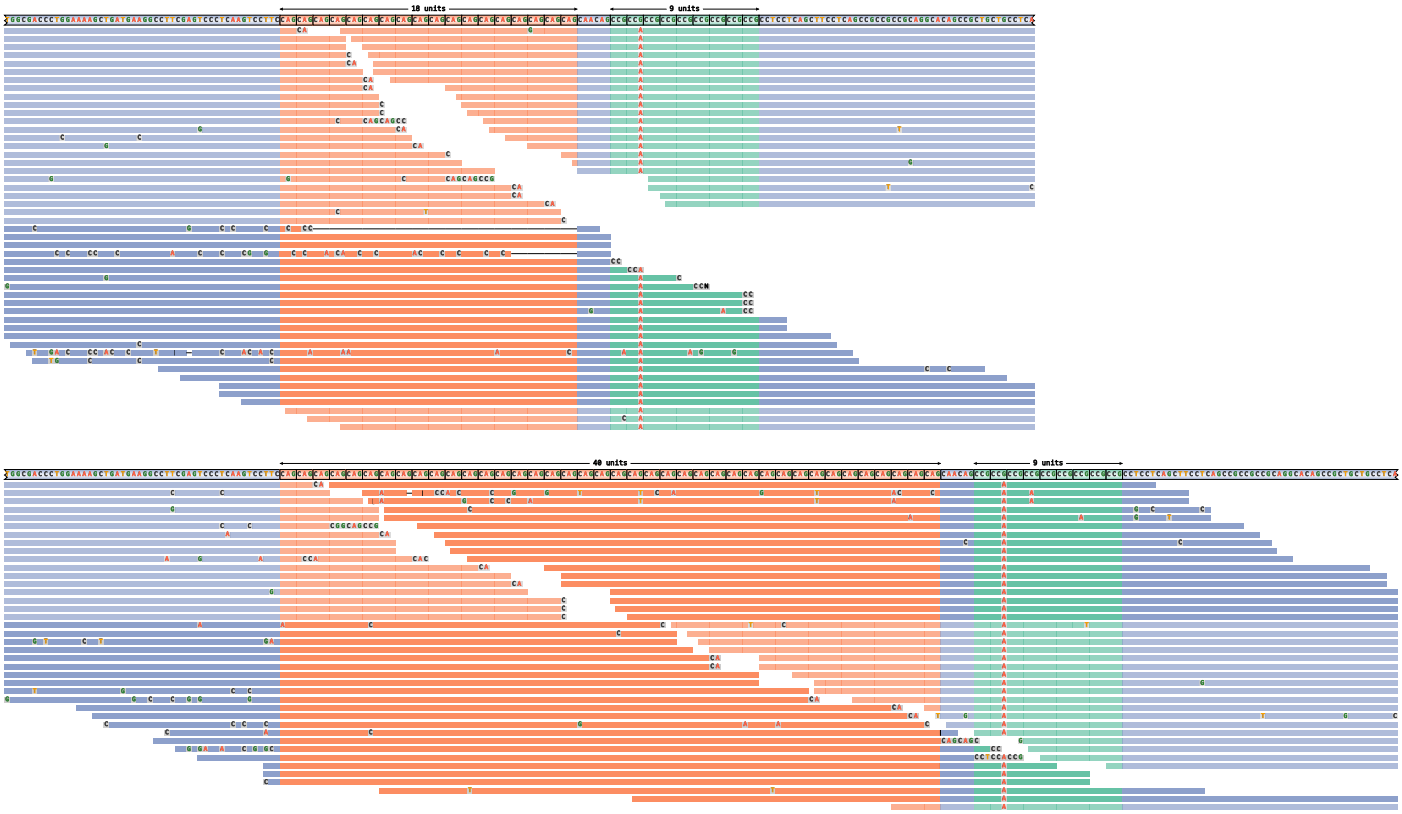
An example of a visualisation plot that illustrates the reads used by ExpansionHunter when estimating expansions in *HTT* is shown in below. Both alleles have an STR repeat-length of 22 and accordingly, reads each containing the `CAG` motif 22 times are visible.



39.2 A good quality STR showing one allele within the normal range and one expanded allele within read length

An example of a visualisation plot illustrating the reads used by ExpansionHunter when estimating expansions in *HTT* is shown below. Alleles of 18 and 40 repeat-lengths are shown in the plot. For the expanded allele, support is provided by:

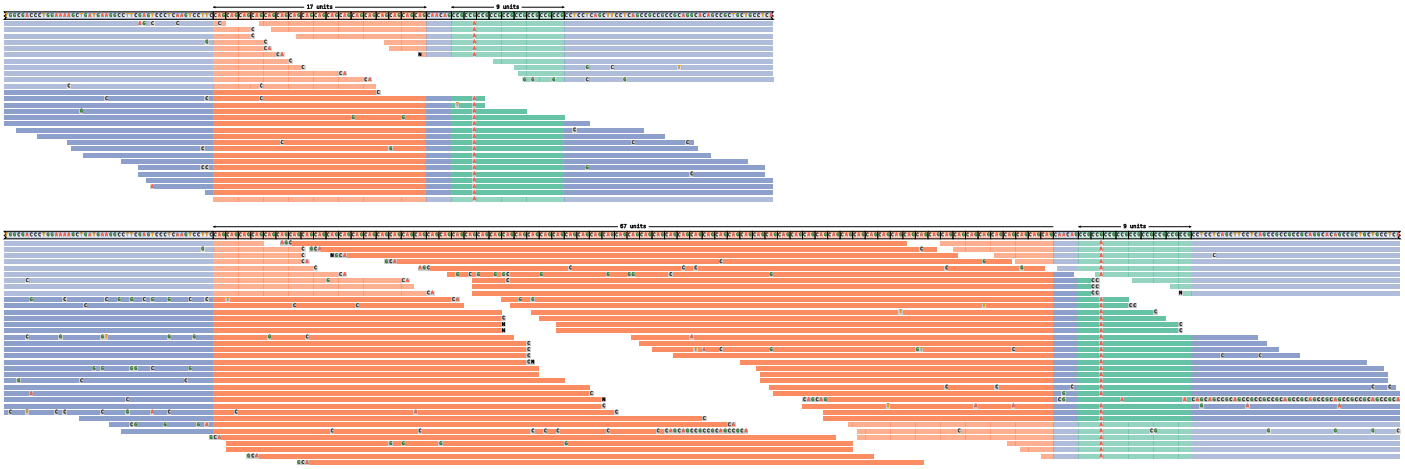
- "**Flanking reads**" that are anchored only on one side of the repeat. Some of these reads support more than 18 repeats but cannot be used to determine the exact number of repeats.
- **"Spanning reads"** that are anchored on both sides of the repeat and support exactly 40 repeats for one of the alleles.
- There are no **"in repeat reads"** that are not anchored at either end of the repeat



39.3 A good quality STR showing one allele within the normal range and one expanded allele with "in repeat reads"

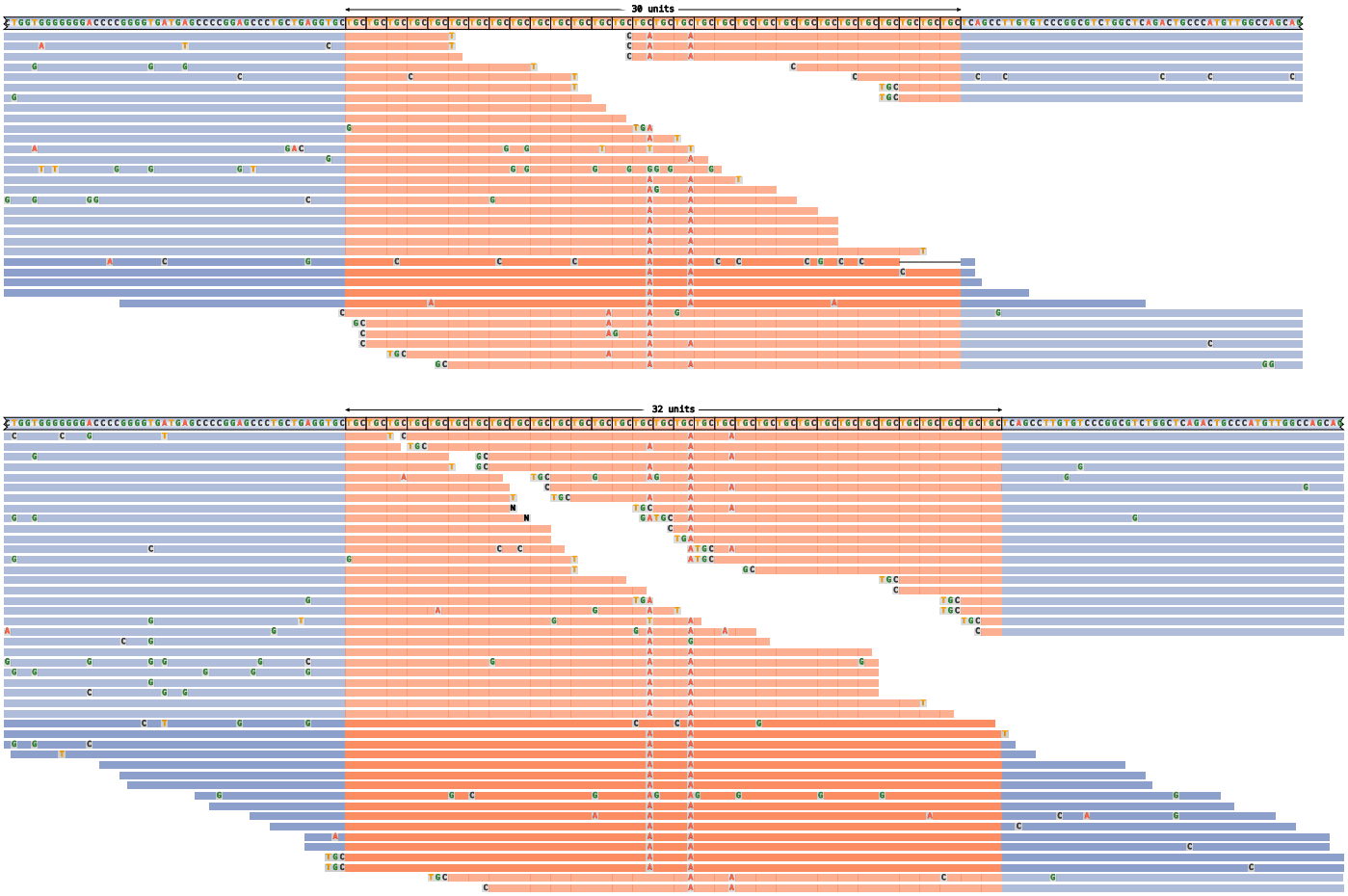
An example of a visualisation plot illustrating the reads used by ExpansionHunter when estimating expansions in *HTT* is shown below. Alleles of 17 and 67 repeat-lengths are shown in the plot. For the expanded allele, support is provided by:

- **"Flanking reads"** that are anchored only on one side of the repeat. Some of these reads support more than 17 repeats but cannot be used to determine the exact number of repeats.
- **"In repeat reads"** that are not anchored at either end of the repeat. These reads support at least 50 repeats.
- There are no **"spanning reads"** that are anchored on both sides of the repeat.



39.4 A good quality STR call with interruptions

An example plot with the reads used by ExpansionHunter when estimating expansions in *ATXN1* is shown below. Interruptions (`ATG` rather than `CTG`) in the reads containing the *ATXN1* repeat motif are visible (*red* `A` *s* in figure). In certain disorders (i.e., ataxias) it is important to use the visualisation plots to check for such interruptions.



40 Compound heterozygous variants across variant types

As discussed in other sections, the Genomics England bioinformatics pipeline will prioritise small variants, copy number variants and short tandem repeat expansions when they fulfil certain criteria.

If a gene is associated with **biallelic modes of inheritance** then variants that are in compound heterozygosity across variant types (e.g. a SNV and a CNV) are also considered together during variant prioritisation.

In order to be considered as a compound heterozygous variant across variant types, the variants **must impact a PanelApp green gene on the applied panel**, with a biallelic mode of inheritance.

In addition:

- CNVs must fulfil all criteria required to be classified as Tier A
- STRs must fulfil other criteria to be considered as a Tier 1 or Tier 2 variant, i.e. have an upper confidence interval beyond the expansion thresholds in PanelApp[1]
- Small variants must fulfil other criteria to be considered as a Tier 1 or Tier 2 variant, including population frequency, consequence type and/or pathogenicity status[1]

Note

[1]In order for STRs and small variants to be considered as Tier1 or Tier2, they must segregate appropriately with the stated mode of inheritance in PanelApp. For biallelic genes, this approach considers variants impacting the gene across the different variant types (Copy Number Variants, Short Tandem Repeat expansions, small insertions, small deletions and single nucleotide variants).

Note

Genes that are not green in the applied gene panel will not include consideration of compound heterozygous variants across variant types. For example, a CNV assigned to Tier Null will not be considered in combination with a small variant assigned to Tier 3.

Tiering of compound heterozygous variants across variant types can be applied for singletons and for probands referred with other family members.

Unlike small variants analysed in trios - which considers variant phase during prioritisation - **the phase of CNVs and STRs are not considered**. Scientists reviewing groups of variants across variant types are recommended to assess segregation of variants in the family to inform whether the variants are *in-cis* or *in-trans*.

Note

de novo small variants are considered during tiering of compound heterozygous variants

III.V Exomiser

41 Overview

For all rare disease referral, interpretation is performed using the Exomiser automated variant prioritisation framework (Next-generation diagnostics and disease-gene discovery with the Exomiser) developed by members of the Monarch initiative: principally Prof. Damian Smedley's team at Queen Mary University London and Professor Peter Robinson's team at Jackson Laboratory, USA, with previous contributions from staff at Charité – Universitätsmedizin, Berlin and the Sanger Institute.

Given a multi-sample VCF file, family pedigree and proband phenotypes encoded by Human Phenotype Ontology (HPO) terms, Exomiser annotates the consequence of variants (based on Ensembl transcripts) and then filters and prioritises them for how likely they are to be causative of the proband's disease based on:

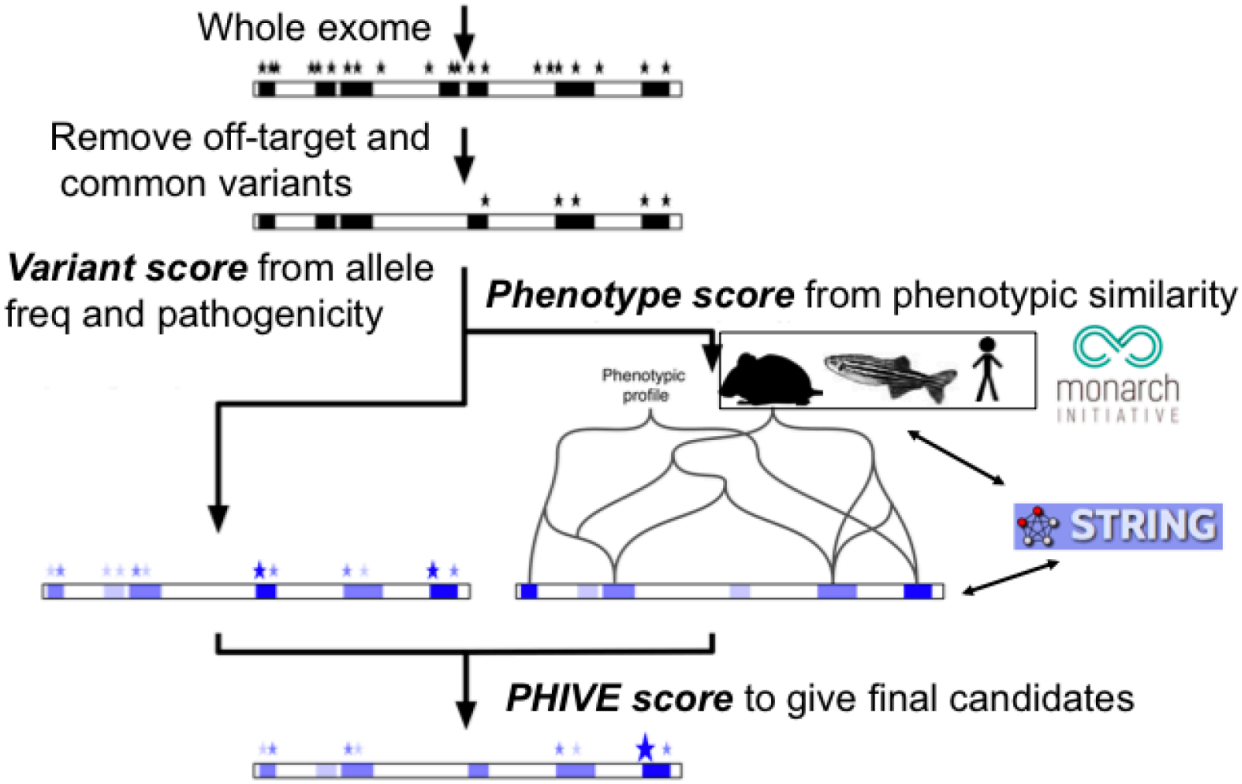
- the predicted pathogenicity and allele frequency of the variant in reference databases
- how closely the patient's phenotypes match the known phenotypes of diseases and model organisms associated with the gene

Note

Both Genomics England and Congenica run Exomiser independently, versions and configuration may differ meaning the two systems may show different Exomiser scores and ranks.

Graphical summary of Exomiser approach:

Exomiser


Whole exome
Remove off-target and common variants
Variant score from allele freq and pathogenicity
Phenotype score from phenotypic similarity
Phenotypic profile
monarch INITIATIVE
STRING
PHIVE score to give final candidates

42 Exomiser implementation in Genomics England Rare Disease pipeline

42.1 Modes of Inheritance

In the Genomics England Rare Disease pipeline, Exomiser is configured to remove low-quality (i.e. non PASS) and non-coding variants (that are not in the variant inclusion list) and then for each of the modes of inheritance (MOI) being considered:

- autosomal dominant
- autosomal recessive
- x-linked dominant
- x-linked recessive
- mitochondrial

42.2 Population frequencies

Variants compatible with the MOI are retained if below a minor allele frequency of 0.1% (or 2% for compound heterozygotes, 0.2% for mitochondrial variants) in all of the following reference databases:

- 100,000 Genomes Project samples
- 1000 Genomes
- ESP
- TOPMed
- UK10K
- ExAC
- gnomAD (excluding the Ashkenazi Jewish population)

For exact Exomiser configurations and database versions see Exomiser configuration.

42.3 Variant score calculation

Exomiser then calculates a score for how rare and deleterious each variant is (on a scale of 0 to 1) using the above frequency sources and predicted deleteriousness scores by Polyphen2, SIFT and MutationTaster from dbNSFP.

For each MOI, the highest scoring compatible variant for each gene, or top two highest for compound-heterozygous candidates, are then selected as the contributing variant(s) for that gene under that MOI and used to assign a gene-level variant score (taking the mean for compound heterozygotes).

Additionally, a variant inclusion list is configured that is based on data from ClinVar, but extended with pathogenic/likely pathogenic variants from GMS data. Exomiser will consider any variant on the inclusion list to be maximally deleterious (i.e. score of 1), regardless of additional annotation (e.g. variant effect, allele frequency, predicted deleteriousness). This means that in some cases Exomiser results will contain variants that would be otherwise excluded from consideration, e.g. non-coding variants.

42.4 Phenotype score calculation

In parallel, Exomiser produces a phenotype score for each gene (on a scale of 0 to 1) based on how phenotypically similar the patient's phenotypes are to:

- OMIM and Orphanet rare diseases known to be associated with the gene,
- mouse and zebrafish models associated with the orthologue of the gene,
- disease, mouse or zebrafish phenotypes associated with neighbouring genes in the StringDB protein-protein association database (scores weighted down based on network distance from the gene under consideration).

This scoring makes use of the OWLSim algorithm to semantically compare phenotypes such that similar but non-exact phenotypes can be identified and weighted according to how distant the two terms are in the ontology as well as how frequently the phenotype in common is observed. The highest score from these comparisons is assigned as the gene-level phenotype score.

42.5 Overall Exomiser score

Finally, a logistic regression model is used to combine the phenotype and variant scores and produce an overall Exomiser score for each gene and its contributing variants for each compatible MOI (scaled from 0 to 1). Variants are ranked based on their overall Exomiser score, with the highest ranked (rank = 1) variant(s) representing the most-likely disease-causing candidate according to Exomiser.

Note

A particular variant can be identified as contributing under a dominant MOI as well as a recessive MOI as a compound heterozygote, and in this scenario will receive two different Exomiser scores. In this scenario, each MOI-specific score is returned as a separate reportEvent for that variant. The maximum Exomiser score out of any of the reportEvents for a variant is used to rank all of the returned variants.

43 Validation of Exomiser performance

43.1 Exomiser versions and validation method

Since the NGIS Jabbah release (September 2023) the Genomics England pipeline uses Exomiser version 13.2.0.

We validated the performance of the latest major release of Exomiser (v13) to identify known genetic diagnoses using 1869 variants in 1659 cases, representing all diagnostic variants reported in the outcome questionnaires in the Genomic Medicine Service by NHS GLHs at the time of analysis (Nov 2022).

The rankings of variants were compared to the previous version of Exomiser used in the pipeline (v12.0.0).

Note

Validations were performed using Exomiser (v13.1.0). There are no expected deviations in performance from the minor updated version of the software integrated into the pipeline (v13.2.0)

43.2 Sensitivity for known diagnostic variants

The 1869 variant(s) reported as diagnostic were returned in the top 3 ranked candidates from Exomiser (v13) for 1709/1869 variants (91%, 95% CI 90%-93%). This favourably compares with 80% (95% CI 77%-81%) in the top 3 using the previous version of Exomiser (v12.0.0).

43.3 Differences in behaviour between versions

Despite the overall improved performance, 68 variants that were previously prioritised rank 1-3 (using Exomiser v12) have dropped below rank 3 using Exomiser v13.

We identified several common reasons for why these variants were below rank 3 in Exomiser 13.1.0:

- Low phenotype match which contributes to a low score (25%)
- The score of a diagnostic variant stayed the same but a different variant improved in score (16%)
- MAF >1%, which while below the threshold of 2% still contributes to a low score (15%)
- The variant was still ranked highly despite being outside of rank 3 (15% in rank 4-5).

Overall, including ranks 1-5 instead of ranks 1-3 increases the recall of variants from 0.91 to 0.94 (95% CI 0.93- 0.95).

43.4 Limitations

One limitation of Exomiser is that heteroplasmy is not taken into account for mitochondrial variants. We see that Exomiser prioritises mitochondrial variants that would usually be filtered by the thresholds used for heteroplasmic variants in the rare disease tiering algorithm (allele fraction ≥ 0.05). As a result, care should be taken interpreting these variants and the heteroplasmy level should be checked.

44 Exomiser configuration

This is a complete description of the configuration used for Exomiser in the Rare Disease GMS. A summary of the general configurations applied is included below:

- Only variants with a "PASS" status are included
- Population frequency cut-offs are applied
- Predicted pathogenicity tools are enabled and this contributes to the score/rank
- Inheritance filters are enabled and this contributes to the score/rank
- The OMIM prioritiser is enabled and this contributes to the score/rank

45 Population allele frequencies

The table below defines allele frequency thresholds by each inheritance pattern considered by Exomiser.

Inheritance pattern	Frequency threshold
AUTOSOMAL_DOMINANT	0.1
AUTOSOMAL_RECESSIVE_COMP_HET	2.0
AUTOSOMAL_RECESSIVE_HOM_ALT	0.1
X_DOMINANT	0.1
X_RECESSIVE_COMP_HET	2.0
X_RECESSIVE_HOM_ALT	0.1
MITOCHONDRIAL	0.2

The allele frequency cutoff is compared against the following populations.

Source	Sub-Population
ESP	ESP_AFRICAN_AMERICAN ESP_EUROPEAN_AMERICAN ESP_ALL

Source	Sub-Population
ExAC	EXAC_AFRICAN_INC_AFRICAN_AMERICAN EXAC_AMERICAN EXAC_EAST_ASIAN EXAC_FINNISH EXAC_NON_FINNISH_EUROPEAN EXAC_SOUTH_ASIAN EXAC_OTHER
gnomAD exomes	GNOMAD_E_AFR GNOMAD_E_AMR GNOMAD_E_EAS GNOMAD_E_FIN GNOMAD_E_NFE GNOMAD_E_OTH GNOMAD_E_SAS
gnomAD genomes	GNOMAD_G_AFR GNOMAD_G_AMR GNOMAD_G_EAS GNOMAD_G_FIN GNOMAD_G_NFE GNOMAD_G_OTH GNOMAD_G_SAS
Others	THOUSAND_GENOMES UK10K TOPMED

Additionally an internal Genomics England allele frequencies dataset is used (see Exomiser database versions)

46 Phenotype scoring algorithms

The HiPhive algorithm is configured to use human, mouse, fish organism data and to include protein-protein interaction proximities in phenotype scores.

47 Variant scoring algorithms

REVEL and MVP are configured as “pathogenicitySources”.

> **Note**
>
> REVEL is an ensemble method that includes data from Polyphen.

48 Variant consequences

A full list of possible variant consequences is available here

The following variant consequences are filtered out, and not considered.

Region	Specific consequence
Untranslated region (UTR)	FIVE_PRIME_UTR_EXON_VARIANT FIVE_PRIME_UTR_INTRON_VARIANT THREE_PRIME_UTR_EXON_VARIANT THREE_PRIME_UTR_INTRON_VARIANT
Transcript	NON_CODING_TRANSCRIPT_EXON_VARIANT NON_CODING_TRANSCRIPT_INTRON_VARIANT CODING_TRANSCRIPT_INTRON_VARIANT
Intergenic	UPSTREAM_GENE_VARIANT INTERGENIC_VARIANT REGULATORY_REGION_VARIANT

49 Short tandem repeat expansion maskings

The following STR loci showed a large number of artifacts caused by the variability between individuals. As these will be better handled by our dedicated STR caller we have excluded these regions from analysis in Exomiser:

- chr1:149390802-149390840 (NOTCH2NLC)
- chr4:3074876-3074963 (HTT)
- chr6:16327633-16327700 (ATXN1)
- chr12:6936727-6936750 (ATN1)
- chr12:111598949-111599000 (ATXN2)
- chr14:92071009-92071011 (ATXN3)
- chr19:45770204-45770252 (DMPK)
- chr20:46022942-46022952 (SLC12A5)

- chrX:147912048-147912058 (FMR1)
- chrX:67545316-67545317 (AR)

50 Exomiser database versions

The Exomiser data release 2209 is used along with Exomiser v13.2.0. This dataset along with a description of its contents can be found here.

The versions datasets inside the bundle are:

Group	Data	Version/Date
Transcripts	Ensembl (GRCh37)	87
	Ensembl (GRCh38)	99
Population frequencies	gnomAD exomes	r2.0.1
	gnomAD genomes	r2.0.1
	ExAC	0.3
	ESP	ESP6500SI-V2-SSA137.GRCh38-liftover
	UK10K_COHORT	20160215
	dbSNP	155
	gnomAD-SV	v2.1
	dbVar	2022-08-03
	DGV	2020-02-25
	GONL	2016-10-13
	DECIPHER	2015
Pathogenicity sources	dbNSFP (Polyphen, MutationTaster, SIFT, MVP, REVEL)	4.0b2a
	ClinVar	2022-08-24

Group	Data	Version/Date
Phenotype data	OMIM	2022-08-26
	Orphanet	2022-06-14
	HPO	2022-06-11
	HPO annotations	2022-06-11
	IMPC	release 17 2022-08-01
	MP	2022-08-04
	Zfin	2022-08-27
	ZP	2022-08-27

51 Genomics England data sources

Internal data sources are also used when running Exomiser v13.2.0, these are described below

Data	Version
Internal allele frequency data[1]	GEL_aggCOVID_DRAGENv4.0-20230921 (internal ref: *20230921-aggDRAGENv4.0_COVID_v1.1-AFgt0*)
Inclusion list[2]	2209_hg38_clinvar_and_20220912_gel_includes_list

1. The same internal allele frequency dataset used by the tiering algorithm. This data was aggregated from 5,415 samples and then converted to an Exomiser specific format (`.pg.gz`) ↩
2. The dataset provided by Exomiser is extended to include variants classified as pathogenic in CVA up to 2022-09-12. ↩

52 Uniparental disomy

The Genomics England WGS pipeline can detect uniparental disomies (UPDs) in individuals for whom both parents have been sequenced.

Predicted uniparental disomies (UPDs) are flagged in the CIP-API and the Interpretation Portal using a label with the format `pddddddddddd [mat|pat]UPDnn [i|h|m][c|p]`.

- `pddddddddddd` is the participant ID of the person in whom the UPD was detected
- `mat|pat` indicates whether the parent who contributed two chromosomes was the mother (mat) or the father (pat)
- `nn` indicates the chromosome where two homologues were inherited from one parent
- `i|h|m` indicates whether the UPD event involves isodisomy (i), heterodisomy (h) or both (m for mixed)
- `c|p` indicates whether the UPD event involves an entire chromosome (c for complete) or part of a chromosome (p for partial)

For example, `p99999999999 matUPD14 ic` denotes that complete maternal isodisomy of chromosome 14 was detected in participant p99999999999.

Note

Uniparental disomies will be flagged irrespective of the penetrance setting used.

Note

Variants showing the appropriate segregation pattern can be tiered under the UniparentalIsodisomy segregation filter and this is independent of flagging.

If approximate coordinates of the predicted regions of isodisomy and/or heterodisomy detected are required, please contact the Genomics England Service Desk.

IV. Additional information

53 Abbreviations and Glossary

Abbreviation / Term	Description
1000GENOMES_phase_3	The 1000 Genomes Project ran between 2008 and 2015, creating the largest public catalogue of human variation and genotype data. As the project ended, the Data Coordination Centre at EMBL-EBI has received continued funding from the Wellcome Trust to maintain and expand the resource.
BAM	Binary Alignment Map of a participant's genome.
CRAM	Compressed Reference-oriented Alignment Map file. A compressed efficient reference-based alternative to the BAM file.
Catalog (OpenCGA)	Catalog is been developed to provide authentication, ACLs and to keep track all of the files and sample annotation.
Cellbase	Annotation Database - https://github.com/opencb/cellbase
CIP	Clinical Interpretation Provider (CIP) is the software company which manages the CIP decision support system used by an NHS GMC user to interpret variants from a case.
CIP-API	Clinical Interpretation Provider Application Programming Interface (CIP-API) is the defined endpoint computer program that communicates between the CIP and the Genomics England bioinformatics pipeline using Genomics England data models.
CNV	Copy Number Variant
ESHG guidelines	The European Society of Human Genetics Guidelines
ESP_6500	NHLBI GO Exome Sequencing Project (ESP) is to discover novel genes and mechanisms contributing to heart, lung and blood disorders by pioneering the application of next-generation sequencing of the protein coding regions of the human genome across diverse, richly-phenotyped populations and to share these datasets and findings with the scientific community to extend and enrich the diagnosis, management and treatment of heart, lung and blood disorders.
EuroGentest	EuroGentest is a project funded by the European Commission to harmonize the process of genetic testing, from sampling to counselling, across Europe. The ultimate goal is to ensure that all aspects of genetic testing are of high quality thereby providing accurate and reliable results for the benefit of the patients.

Abbreviation / Term	Description
EVS	Exome Variant Server
ExAC	The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.
GeCIP	Genomics England Clinical Interpretation Partnership
GEL	Genomics England
GelPedigree	The Model of the pedigree is defined with the following parameters: 1. Model version number, 2. Family id which internally translates to a sample set, 3. Participants, members of a family with associated phenotypes as present in the record RD Participant, 4. Analysis Panels, in a family with associated phenotypes as present in the record Participants 5. Penetrance of a disease, in a family with associated phenotypes as present in the record Participants
NHS GLH	NHS Genomics Laboratory Hub
GnomAD	Genome Aggregation Database. This is a coalition of investigators seeking to aggregate and harmonize exome and genome sequencing data from a variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.
GONL	The Genome of the Netherlands is a consortium funded as part of the Netherlands Biobanking and Biomolecular Research Infrastructure. Samples where contributed by LifeLines, The Leiden Longevity Study, The Netherlands Twin Registry (NTR), The Rotterdam studies, and The Genetic Research in Isolated Populations program.
GRCh37	The human genome assembly GRCh37 (also known as hg19)
GRCh38	The human genome assembly GRCh38
HPO	Human Phenotype Ontology
HPO terms	Human Phenotype Ontology terms

Abbreviation / Term	Description
HTML	HyperText Markup Language – used to provide a human-readable presentation of key information from the JSON data export (a report).
Interpretation Browser	The Interpretation Browser is within the Genomics England Interpretation Portal enables the NHS GMC clinical scientists to review results of Genomics England Interpretation Services (e.g. Tiering and Exomiser) that have been applied to rare disease cases
Interpretation Portal	Webpage provided by Genomics England to host clinical reports and used to launch cases into a CIP, using the CIPAPI.
JSON	JavaScript Object Notation (JSON) is a lightweight data-interchange format used to encapsulate Genomics England's interpreted genome and interpretation request through the CIP-API.
LabKey	Data Server hosting patient clinical and demographic information, excluding VCFs and BAMs.
LDAP	Lightweight Directory Access Protocol (LDAP) is a client/server protocol used to access and manage directory information. It reads and edits directories over IP networks and runs directly over TCP/IP using simple string formats for data transfer.
Main findings	Variants which have been found and potentially associated with the disease/disorder for which the patient has given consent for the genetic test. Referred to as 'Primary Findings' within Genomics England developed systems.
MDT	Multi-Disciplinary Team
OMIM	Online Mendelian Inheritance in Man
PanelApp	PanelApp (Open Source) was created to enable virtual gene panels to be viewed and commented on by experts
PID	Patient Identifiable Data
PMID	Unique identifier number used in PubMed. They are assigned to each article record when it enters the PubMed system, so an in-press publication will not have one unless it is issued as an electronic pre-pub

Abbreviation / Term	Description
Primary findings	The variants that have been found and associated with the disease/disorder for which the patient has given consent for the genetic test.
SNV	Single Nucleotide Variant
STR	Short Tandem Repeat
SV	Structural variant
Tier	Flag used by Genomics England to signify variants of potential relevance to the patient's condition - will be automatically categorised into Tiers to aid evaluation
UK10K_ALSPAC	The Avon Longitudinal Study of Parents and Children (ALSPAC) is a long-term health research project. More than 14,000 mothers enrolled during pregnancy in 1991 and 1992, and the health and development of their children has been followed in great detail ever since. The ALSPAC families have provided a vast amount of genetic and environmental information over the years.
UK10K_TWINSUK	The database used to study the genetic and environmental aetiology of age-related complex traits and diseases. It is one of the major departments of King's College London Division of Genetics and Molecular Medicine and is the most detailed clinical adult register in the world.
UPD	Uniparental Disomy
VCF	Variant Call Format

54 Software and Database Versions

54.1 Software

Task	Software	Version
Genome alignment	DRAGEN	4.0.5
Small variant detection	DRAGEN	4.0.5
Copy number detection	DRAGEN	4.0.5
Structural variant detection	DRAGEN	4.0.5
Short tandem repeat (STR) expansions	ExpansionHunter (as part of DRAGEN)	5
Variant annotation	CellBase	5.4.25

54.2 Databases

Source	Version
Ensembl	107
Genomics England Tiering	5.166.3
ClinVar	2023-04
gnomAD_genomes	3.1.2
gnomAD_exomes	2.1.1

Note

The version of Ensembl indicated in the table above is used for the annotation of variants during the Genomics England Rare Disease variant tiering approach, and uses the version for the GRCh38 reference build. Information that is considered by Exomiser during variant prioritisation is available here.

55 Release dates

The table below indicates the production dates of NGIS releases since 2023.

Release name	Production release date
Quasar	tbc
Petra	11/06/2025
Orion increment (PanelApp update)	30/04/2025
Orion	19/02/2025
Nembus increment (PanelApp update)	30/10/2024
Nembus	09/10/2024
Mira	01/05/2024
Lyra	06/12/2023
Kraz	15/11/2023
Jabbah	20/09/2023
Izar	22/03/2023

56 B-Allele Frequency Plots

56.1 Access to B-Allele Frequency Plots

From the Orion release onwards, B-Allele Frequency plots generated by the Genomics England Rare Disease bioinformatics pipeline are available for download for all samples through the Interpretation Portal. Further instructions on how to access these plots are available here.

56.2 Background

B-Allele frequency plots provided by Genomics England display a combination of metrics that are useful for the detection of features and events that may not be immediately accessible through other approaches for variant detection.

B-Allele frequencies are quantified as *"the proportion of sequencing read support for an alternate allele ('B-Allele') in comparison to the reference genome"*.

Hypothetically, we expect certain B-Allele frequencies in specific scenarios on diploid chromosomes:

Scenario	Genotype	Expected B-Allele frequency
Reference Homozygous	0/0	0.0
Alternate Heterozygous	0/1	0.5
Alternate Homozygous	1/1	1.0

Normal ranges do differ from these hypothetical values, but when considered in combination with other datasets (coverage and copy number count), B-Allele frequencies can be useful to identify and characterise several event types, including:

- uniparental disomy
- regions of homozygosity
- large mosaic events (e.g. copy number variants, uniparental disomy)

56.3 Limitations of B-Allele frequency plots

B-Allele frequencies are calculated from genomic sequencing datasets, and are currently limited to genomic sites that pass quality filters and are covered by at least 10 sequencing reads. At an average genome-wide sequencing coverage of 30-40x, B-Allele frequency plots can be useful for the detection of large CNVs, particularly large mosaic CNVs, and other genomic events that are not detected or prioritised through other approaches in the Genomics England rare disease bioinformatics pipeline.

However, there are known limitations, for example, the level of mosaicism that can be detected, the range of allele fraction distributions that can be observed and the static nature of the plots provided.

It is strongly encouraged that B-Allele frequency plots are utilised alongside other datasets provided by the Genomics England rare disease bioinformatics pipeline to characterise and detect complex genomic events, including alignment files and structural variant vcfs. This may include, but is not limited to, assessment of the orientation of read pairs, the soft clipping of sequencing reads, and the relative read coverage within informative regions of the genome (e.g. pseudoautosomal regions on chrX to infer minor sex karyotypes). Moreover, in cases where the quantitative level of mosaicism is explicitly considered, inferences (whilst limited) can be made from the level of read support for variants within the small variant vcfs.

56.4 Information available in B-Allele frequency plots

As shown below, B-Allele frequency plots produced by the Rare Disease bioinformatics pipeline contain three different data types:

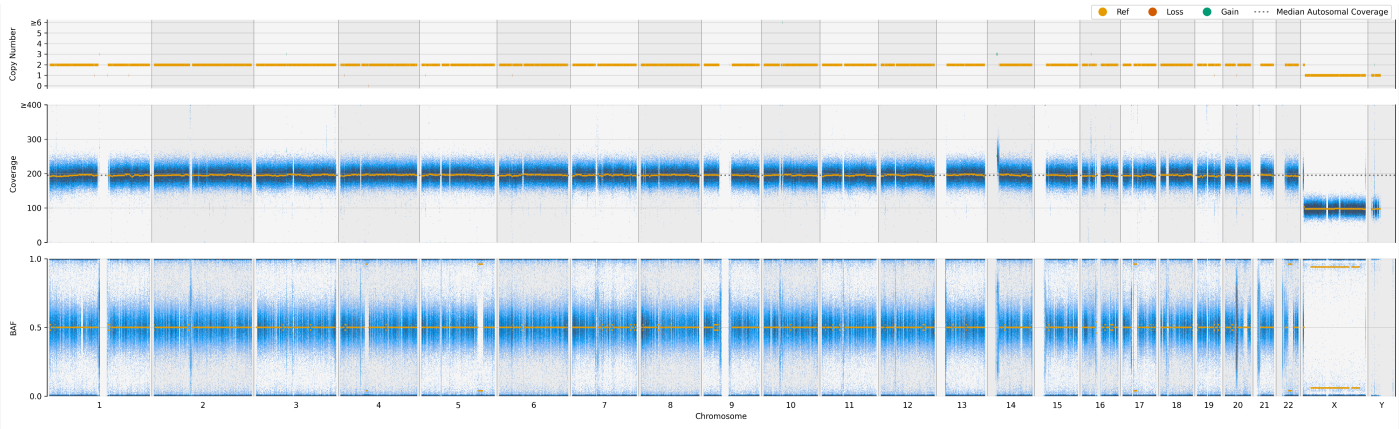
- **top panel:** copy number states
- **middle panel:** sequencing read coverage values
- **bottom panel:** B-Allele frequencies for small variants

Orange lines indicate general trends in different sections of the plots.

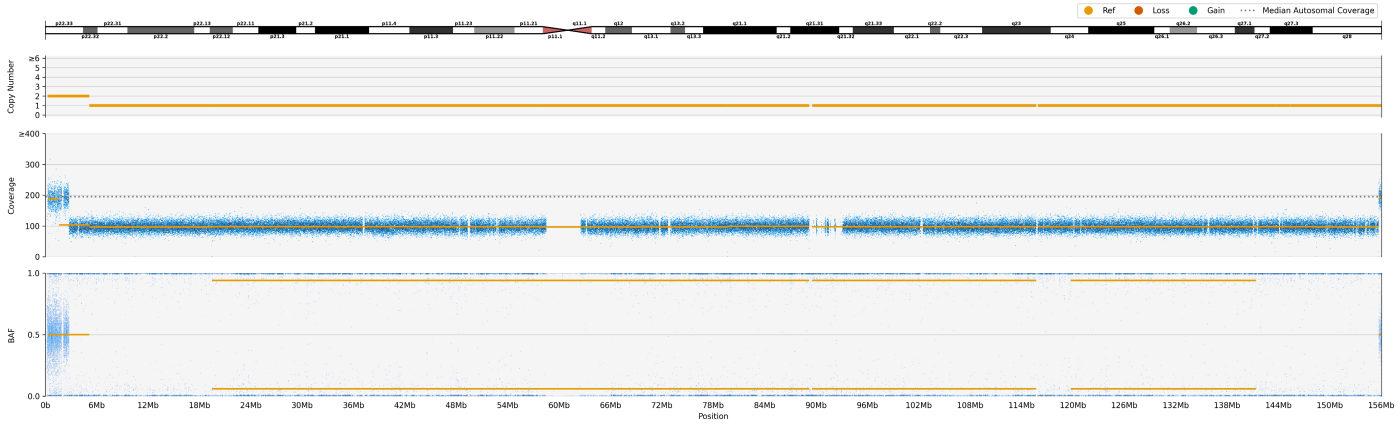
B-Allele frequency plots are provided for all autosomal and sex chromosomes on a single plot, and also individually for each chromosome. The resolution of chromosomal regions is higher in individual chromosome plots.

56.4.1 Examples of typical B-Allele frequency plots

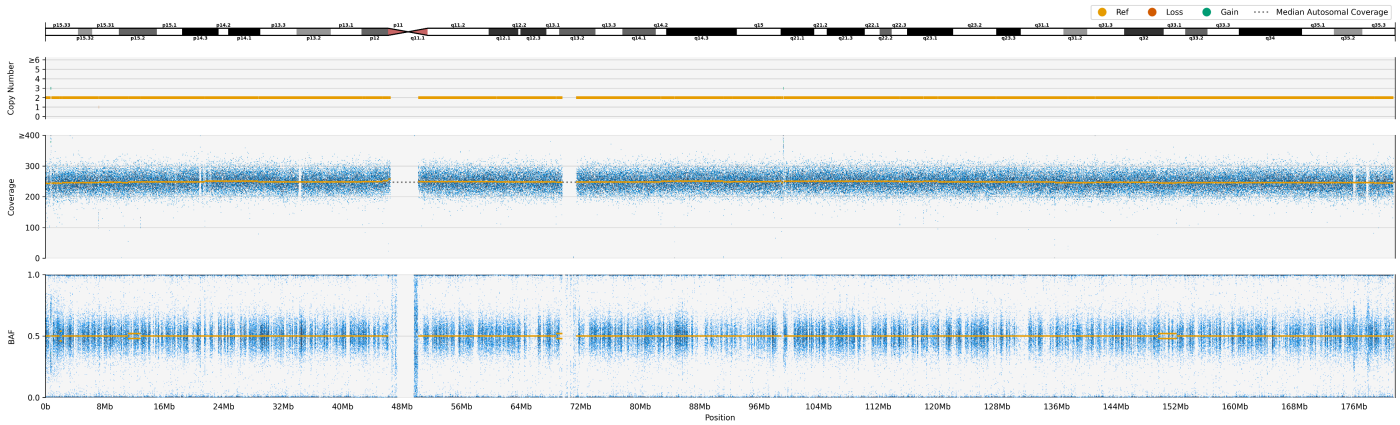
56.4.1.1 Typical whole genome plot



56.4.1.2 Typical chromosome X plot, XY karyotype



56.4.1.3 Typical autosomal chromosome plot



There are several features to note in the typical plots included above:

Feature	Observations
karyotypic sex	The whole genome plot suggests this sample is from an individual with an XY karyotype. Specifically, the coverage values and copy number states indicate presence of 1 copy each for chromosome X and Y. Of note, the pseudoautosomal (PAR) regions of chrX, which are also present on chrY, have features consistent with a diploid state (CN=2). PAR1 (~first 3Mb of chrX) is viewable on the chromosome X single chromosome plot above. PAR2 is not easily viewable.
normal copy number state	There are no clear indications of large copy number gains or losses across any chromosomes, and this can be confirmed for individual chromosomes (if appropriate to do so) on individual chromosome plots, shown above for chr8. This can be interpreted by the presence of: (1) normal copy number state, (2) no obvious deviation in coverage values, and (3) no abnormalities in B-Allele frequencies - please

Feature	Observations
	see additional examples on this page for key features of B-Allele frequency distributions indicative of abnormal copy number state.
distribution of B-Allele frequencies	The plots below show the typical distributions of B-Allele frequencies in a sample from an individual with an XY karyotype. There is natural deviation from the hypothetical values suggested in the table above. This reflects both real biological variation and the protocols put in place to increase interpretability of the plots. There are two trends observable in the whole genome and autosomal chromosome plots that are consistent with normal diploid states: (1) examples of B-Allele frequencies close to 0 or 1, indicative of homozygous sites, and (2) examples of B-Allele frequencies around 0.5 (approx range of 0.3-0.7), indicative of heterozygous sites.
general trends (solid lines) for coverage and B-Allele frequencies	Solid lines are provided to indicate general trends observed for sequencing coverage and B-Allele frequencies. The regions that are plotted as solid lines mirror the regions that are reported in the CNV vcf, and therefore may not perfectly align with the boundaries of other detectable features (e.g. homozygous regions or regions implicated in uniparental disomies). For coverage, the solid line marks the median coverage and the line colour reflects the CNV status. For B-Allele frequencies, the orange line represents the modal (i.e. most common) value for non-homozygous B-allele frequencies, and is included for all regions with >250 non-homozygous B-Allele frequency observations.

Note

Copy number variants detected with high confidence in a proband will be considered through variant tiering approaches.

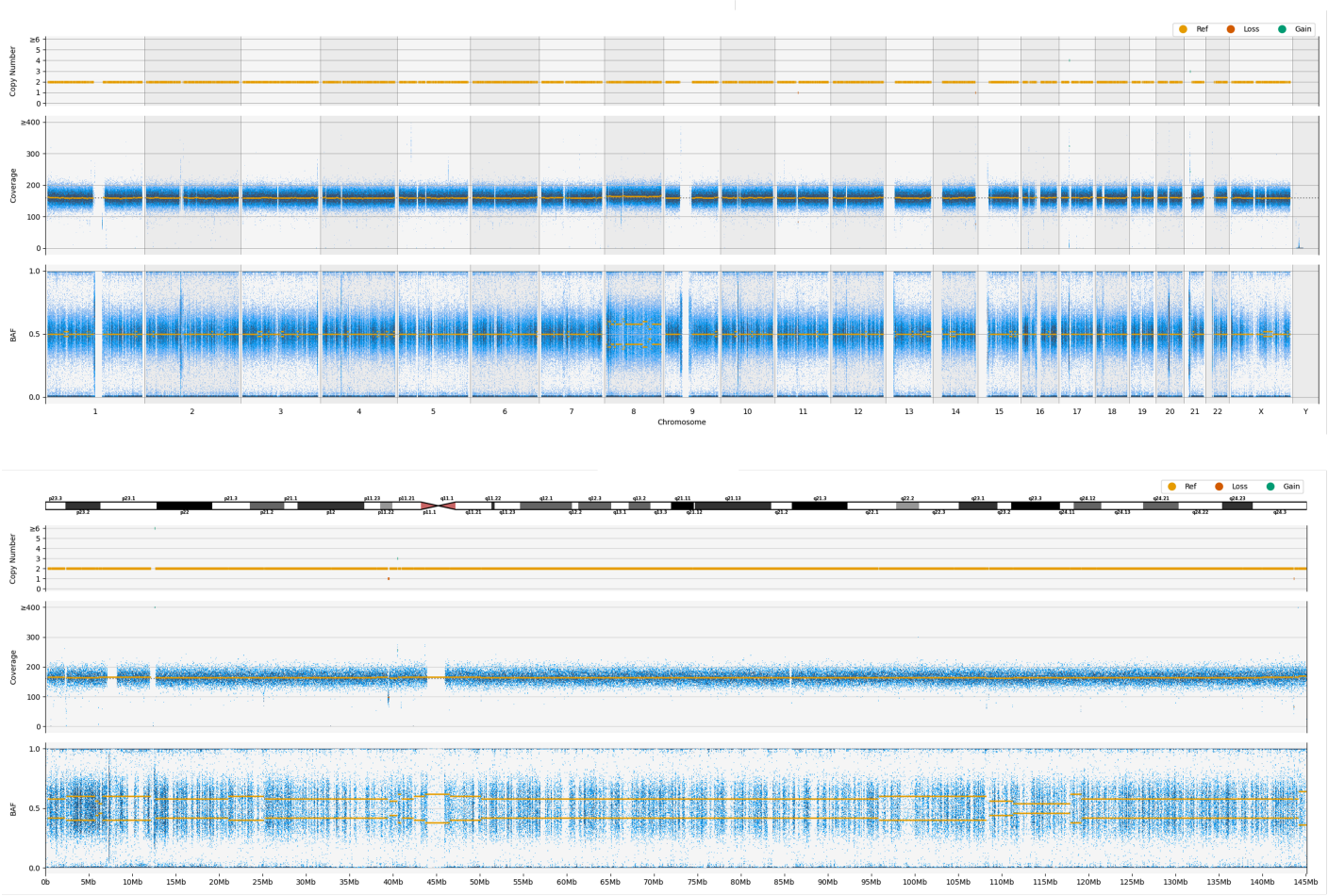
Note

Regions proximal to the centromere and telomeres may appear "messier" (i.e. greater range of B-allele frequencies) or absent of data. This is expected, and is due to the difficulty of alignment and variant detection in these regions with short-read sequencing technologies.

Note

Due to the CNV regions utilised to calculate trends, some regions will be missing from trend calculations as they are not considered for CNV detection, and the solid lines are unlikely to be representative of trends associated with uniparental disomies.

56.4.2 Mosaic CNV viewable in B-Allele frequency plots

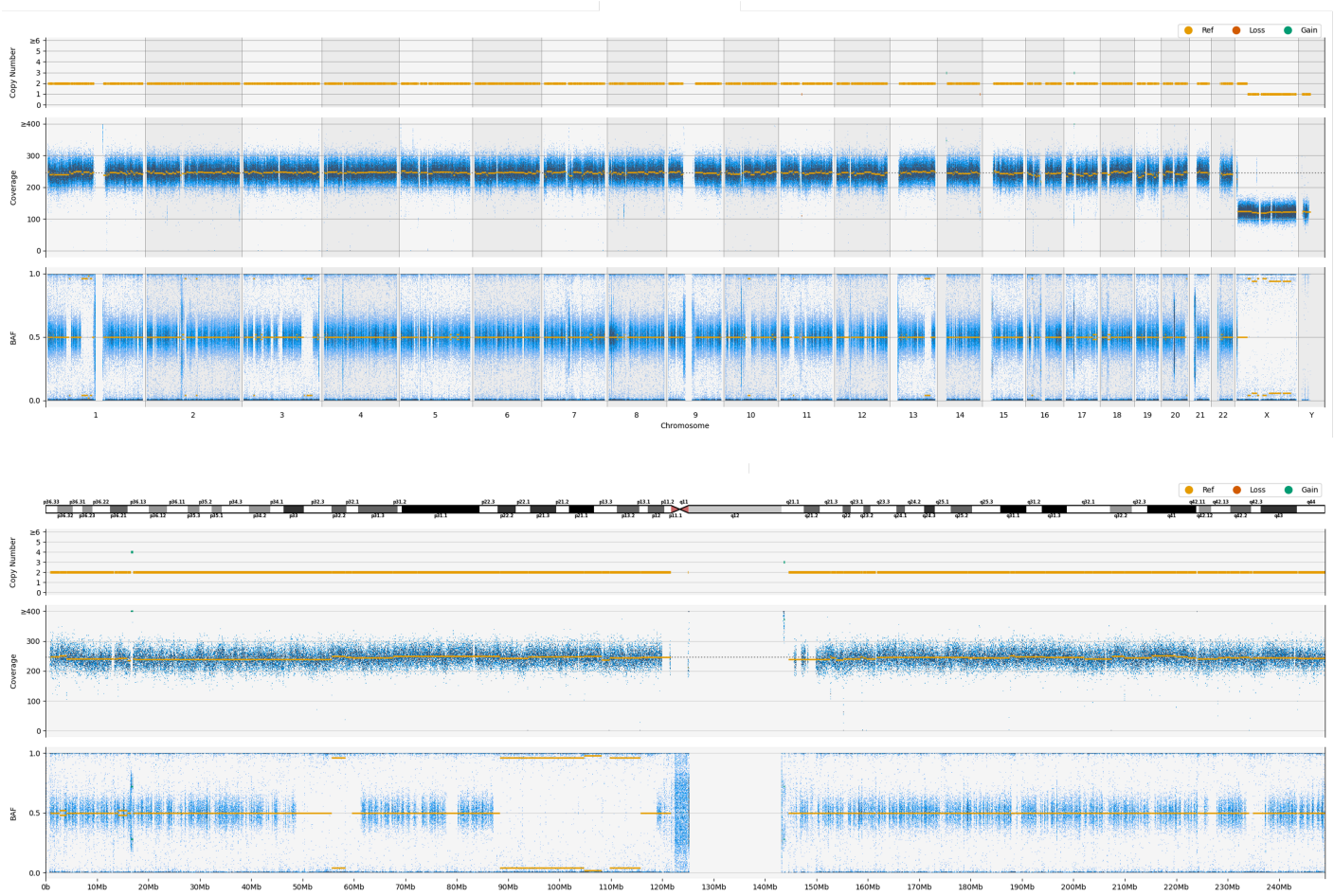


This example illustrates trends indicative of a mosaic copy number gain, in this case impacting the whole of chromosome 8.

In both the whole genome and single chromosome plots, it can be observed that chromosome 8 has B-Allele Frequencies that deviate from the values typical for heterozygous variants. This occurs across the complete length of chromosome 8.

On the whole genome plot, it can also be observed that there is a slight increase in the coverage values across the length of the chromosome, although not significant enough to be detected as change in predicted copy number.

56.4.3 Regions of homozygosity



This example illustrates trends associated with regions of homozygosity (ROH) present on several chromosomes.

The characteristic features of ROHs can be observed, with B-Allele frequencies almost exclusively at 0 or 1, but clear indication in the coverage and copy number panels of the plots that there are two copies of the chromosome.

This scenario may occur due to consanguinity, uniparental isodisomy or a balanced CNV event where the region lost on one copy is replaced with that region from its pair.

These trends can also be seen in the individual chromosome plots, and are shown here for chromosome 1, with ROH impacting approximate regions 50-60 Mb and 90-120 Mb.

57 Pipeline sensitivity and precision

Variant Type	Measure	Sensitivity	Precision	Truth set
Single Nucleotide Variants	Mean	0.9978	0.9995	HG002 (high confidence regions)
Single Nucleotide Variants	95% credible interval	0.9978-0.9979	0.9995-0.9995	HG002 (high confidence regions)
Indels	Mean	0.9979	0.9991	HG002 (high confidence regions)
Indels	95% credible interval	0.9977-0.9980	0.9990-0.9991	HG002 (high confidence regions)
Copy Number Variants (>2Kb)	Mean	0.9779	N/A	Clinically significant CNVs detected by standard of care tests in accredited labratories
Copy Number Variants (>2Kb)	95% credible interval	0.9609-0.9929	N/A	Clinically significant CNVs detected by standard of care tests in accredited laboratories
Short Tandem Repeat Expansions	Mean	0.9828	0.9194	STR expansion tests (positive and negative) at target loci performed in accredited laboratories
Short Tandem Repeat Expansions	95% credible interval	0.9343-1	0.8252-0.9943	STR expansion tests (positive and negative) at target loci performed in accredited laboratories

Note

HG002: (Child, Ashkenazi Jewish Trio) sample with Genome in a Bottle truth set (high confidence regions) - metrics reported here relate to variant detection performed as a singleton

58 Coverage profile data

Genomics England provide an analysis of the coverage profile for all the genes that are green in any PanelApp panel. This has been updated to reflect recent panel changes.

The coverage profile data can be accessed from the NHS Futures website (https://future.nhs.uk/) under “NHS Genomic Medicine Service” > “Guidance” > “Genomics England Documentation”, and can be downloaded from the following link:

Coverage profile dataset

59 Clinical Interpretation Portal-API

The Clinical Interpretation Portal (CIP)-API serves four functions:

1. It communicates with the NHS GLH user and the Decision Support Systems (DSS) regarding which cases are ready for interpretation. It does this by creating an “Interpretation Request” which is sent to Interpretation Services (e.g., Tiering & Exomiser) and DSS and is visible from the CIP-API web services.
2. When an Interpretation Service generates an “Interpreted Genome”, it pushes this information back to the CIP-API. These data are appended to the Interpretation Request for the case and can be accessed through the CIP-API web services.
3. If an NHS GLH user selects a variant as a Primary Finding in the Interpretation Portal (or DSS) user interface and decides to produce a Summary of Findings, the Portal and / or DSS pushes this information as a “ClinicalReport” to the CIP-API. The “ClinicalReport” is appended to the InterpretationRequest for the case.
4. Via the Interpretation Portal, the CIP-API displays the case status and the ClinicalReport.json as an HTML page visible to the NHS GLH user.

Note

Further information about how to access and query the API and all the endpoints are documented here: https://cipapi-documentation.genomicsengland.co.uk/

60 Decision Support Systems (DSS)

Congenica provides Decision Support Services for Rare Disease cases in the GMS.

NHS GLHs will be issued with a DSS user guide during training on the system. If additional copies are required, please contact Genomics England service desk here: ge-servicedesk@genomicsengland.co.uk or via the Genomic England Service Desk

61 GMS interpretation portal

The GMS Interpretation Portal allows users to:

- See an overview of cases ready for NHS GLH review, and track overall case status.
- Review findings from Interpretation Services such as Tiering and Exomiser.
- Download any available files and link out to Decision Support Systems
- Complete a reporting outcomes questionnaire to close a case.
- Save work in progress as draft and return to it later e.g., when completing the outcomes questionnaire.
- Review alignment and variant calls from BAM and VCF files using IGV.js.

The GMS Interpretation Portal is accessible here and online help pages are available here.

62 Genome data available through IGV.js

Genomic data are available for browsing using IGV.js through the Interpretation Portal. A variety of data and files generated by the Rare Disease Pipeline are available to view, a summary of which is shown below. The most relevant files for review are shown in bold.

File	Description
[SampleID].repeats.vcf.gz	Short tandem repeat genotypes estimated by ExpansionHunter
[SampleID].enhanced.cnv.vcf.gz[1]	Copy number variants detected by DRAGEN CNV in the proband. Including annotations for high quality small CNVs (2-10 kb) detected by DRAGEN CNV and DRAGEN SV
[SampleID].cnv.vcf.gz	Copy number variants detected by DRAGEN CNV in other family members
[SampleID].forceGT.vcf.gz[1]	Genotypes of approximately 500,000 SNPs used for Genomic and Data Checks
[SampleID].FGT_SMS.SNP.vcf.gz	Genotypes of SNPs used by the Sample Matching Service
[referralID_XXXXX].duprem.left.split.vcf.gz	Small variants detected by the DRAGEN small variant caller after normalisation
[referralID_XXXXX].sv.vcf.gz	Structural Variants detected by DRAGEN SV, joint called for family

File	Description
	members where available
[SampleID].GRCh38DecoyAltHLA_NonN_Regions_autosomes_sex_mt.CHR_full_res.bw	Genome coverage file
[SampleID].target.counts.bw	Intermediate file from DRAGEN CNV. This file can be used to review dropout regions for which CNV signals are not extracted from the alignments for inclusion in CNV calling. CNV events may span these intervals if there is sufficient signal in flanking regions.
[SampleID].cram	Genome alignment (CRAM format) generated by the DRAGEN aligner

Note

[1]currently there is a known issue causing this file to be inaccessible

63 Limitations of the Rare Disease bioinformatics pipeline

A summary of the limitations of the Rare Disease bioinformatics pipeline is provided in the Summary of Findings available in the Interpretation Portal.

The rare disease bioinformatics pipeline has been through several iterations of innovation. As a result, the features and approaches included in the rare disease bioinformatics pipeline can differ between major NGIS releases. Examples of innovations include:

- detection of CNVs between 2 and 10Kb (released in NGIS Danny release, July 2021);
- prioritisation of variants with known pathogenic/likely pathogenic status (released in NGIS Izar release for ClinVar variants, March 2023);
- upgrade of the DRAGEN software used for read mapping and variant calling to DRAGEN v4.0.5 (released in NGIS Mira release, April 2024).

The specific NGIS release version that the Statement of Limitations text relates to is included in the first paragraph.

> This sample was processed through the Rare Disease bioinformatics pipeline included in the latest NGIS release. Additional details of the Rare Disease analytical pipeline are available in the Rare Disease Genome Analysis Guide (available online, and through the NHS Futures Website).
>
> The variants described below were selected by the NHS Genomic Laboratory Hub following review of prioritised variants from the Genomics England interpretation (tiering) pipeline. It may include single nucleotide variants and small insertions/deletions in the virtual gene panel(s) classified as Tier 1 or 2 and/or other types of prioritised variants that may be of relevance to the patient's phenotype. The variants identified here were detected from whole genome sequencing data with a variant prioritisation process that focused on protein coding genes, selected non-coding genes, and loci in accordance with most currently diagnostic reportable genomic variation.
>
> The single nucleotide variant (SNV), small insertion/deletion (indel) and copy number variant (CNV) tiering has been carried out based on the clinical indication and pedigree data as given in the referral. It is the responsibility of the reporting laboratory to check that this information is correct before issuing a clinical report.
>
> Tiered SNVs and indels are rare variants that segregate with disease under the penetrance mode defined in the referral.
>
> Tier 1 includes rare variants in the applied virtual gene panel that are:
>
> - high impact variants
>   - predicted consequence types: stop-gain, stop-loss, start-loss, splice donor/acceptor, frameshift, transcript ablation
> - de novo variants predicted to be of functional consequence
>   - only applies to genes associated with a phenotype with monoallelic mode of inheritance
> - ClinVar and/or Clinical Variant Ark (CVA) variants with at least one pathogenic or likely pathogenic assertion in genes in the virtual gene panel(s) applied for the patient
>   - variants with the same protein change as a ClinVar variant or a CVA variant are also included for consideration* - this amino acid matching approach only applies to variants that were classified as pathogenic/likely pathogenic in CVA from the NGIS Izar release (March 2023) onwards
>
> Tier 2 includes rare variants in the applied virtual gene panel that are:
>
> - moderate impact variants
>   - predicted consequence types in protein-coding genes: missense, splice region variant (+/- 8bp from the nearest exon), in-frame insertion/deletion, transcript amplification, incomplete terminator codon
>   - predicted consequence types in non-coding genes: non-coding transcript exon variant
>
> Tier 3 includes rare variants outside of the applied virtual gene panel that are:
>
> - high impact variants
>   - predicted consequence types: stop-gain, stop-loss, start-loss, splice donor/acceptor, frameshift, transcript ablation consequence types
> - moderate impact variants in protein-coding genes

  - predicted consequence types in protein-coding genes: missense, splice region variant (+/- 8bp from the nearest exon), in-frame insertion/deletion, transcript amplification, incomplete terminator codon
- ClinVar and/or CVA variants with at least one pathogenic or likely pathogenic assertion
  - variants with the same protein change as a ClinVar variant or a CVA variant are also included for consideration* – this amino acid matching approach only applies to variants that were classified as pathogenic/likely pathogenic in CVA from the NGIS Izar release (March 2023) onwards

*It is recommended that HGVSp. automated predictions are verified during the reporting process, particularly for variants with complex nomenclature.

Tiered CNVs are high quality calls >2 kb derived from the proband only. Tier A includes rare CNVs that overlap with genes or contain regions defined in the virtual gene panel(s) applied to the patient. Tier B includes rare CNV calls >100 kb that overlap with protein coding genes, but do not overlap genes that are classified as 'Green' on the applied gene panel. Tier A and Tier B CNVs may have duplicate Tier Null entries if they also overlap genes with relevant biotypes that fulfil the criteria for Tier Null. Tier Null includes high quality CNV calls >2 kb that neither overlap with genes nor regions defined in the virtual gene panel(s), but does overlap a gene with a relevant biotype.

Short Tandem Repeats (STRs) are only included in the prioritised variants for specific loci defined in the virtual gene panel(s) applied to the patient.

It is possible that disease-causing variant(s) were not detected, for example because they are in a region of low coverage, low mappability, or poor sequence quality, they are of a type that could not be detected, they have a lower than expected allelic balance due to mosaicism or they are mitochondrial variants with very low heteroplasmy levels. Variants may also not be included in this list of prioritised variants if the variant falls outside of the virtual gene panels applied, has a consequence type that is not prioritised, a population allele frequency above the threshold applied, a segregation pattern not considered or not in accordance with the mode of inheritance for pathogenic variants attributed to the relevant gene or entity, or the segregation pattern in the family is not as expected (for example, incomplete penetrance was not anticipated). In some cases, biallelic STR expansions may be detected as monoallelic expansions and may not be included in the list of prioritised variants where the genotype is not in accordance with the anticipated mode of inheritance attributed to the STR. In these cases, the expansion will be reported as Tier null. For longer expansions where the length of the repeat is longer than the WGS read length (150bp or 50 repeats for a triplet repeat), pre-expansions may not be reliably distinguished from full expansions. In such cases, the full expansion length may be underestimated and reported in either Tier 1 or Tier 2. Please note that CNVs <2 kb and structural variants are not currently reported. All GENCODE Basic transcripts (Ensembl version 90, GRCh38) associated with specified biological significance categories are considered in the tiering algorithm. Further diagnostic or research analysis may lead to updated prioritised variants being issued in the future.

The estimated sensitivity and precision of the Rare Disease pipeline 2.0 for variant detection (not including tiering) of small variants, CNVs and STR expansions are summarised below. Estimates may be revised as availability of appropriate data for validation improves.

> Genomics England Limited is a UKAS accredited medical laboratory No. 10170. Genomics England's Schedule of Accreditation includes Fixed Scope and Flexible Scope permissions; the latter allows Genomics England to make changes within defined and agreed UKAS boundaries and report the results as accredited. Any results reported that are outside the Fixed Schedule and Flexible Scope boundaries are noted below and highlighted as outside the scope of UKAS accreditation. Further detailed information can be found in the release notes for each software release, and in the Rare Disease Genome Analysis Guide.
>
> Mitochondrial DNA variants were not included in calculation of sensitivity and precision and are outside the scope of ISO 15189 accreditation for the pipeline. The measurement of uncertainty and the limit of detection of heteroplasmic mitochondrial DNA variants was not determined.
>
> Pipeline sensitivity and precision metrics

64 Links to supporting documentation

- GMS Interpretation Portal
- 100K Interpretation Portal
- Interpretation Platform Documentation
- Clinical Variant Ark

65 Feedback

If you have any feedback on the Genomics England Rare Disease bioinformatics pipeline please contact the Genomics England Service Desk at ge-servicedesk@genomicsengland.co.uk.

66 Release notes

Releases for the Rare Disease pipeline are bundled with NGIS releases, details of which can be found here.

The dates of NGIS releases can be found here.

66.1 Quasar

Release information

There are no major alterations to the rare disease bioinformatics pipeline in the Quasar release

66.2 Petra

Release information

The new features that are released in Petra are:

- Extend the list of regions excluded from the Exomiser configuration.

66.3 Orion increment

Release information

The new features that are released in the Orion increment are:

- Update of gene sequencing coverage report information to include updated version of PanelApp
- Update of gene sequencing coverage report to include samples prepared through the Illumina DNA PCR-Free library preparation method

66.4 Orion

Release information

The new features that are released in Orion are:

- Addition of mitochondrial variant allele frequencies from gnomAD v3.1.2 for consideration during variant tiering
- Inclusion of appropriate large CNVs (>100Kb) as Tier B report events
- Availability of B-Allele Frequency plots for download and assessment

66.5 Nembus

Release information

The new features that are released in Nembus are:

- Upgrade of the database (CellBase) and datasets utilised during variant annotation, and considered during variant tiering, including:
  - upgrade from Ensembl v90 to Ensembl v107
  - upgrade from gnomAD v2.0.1 to gnomAD v3.1.2 (genomes) and gnomAD v2.1.1 (exomes)
  - removal of some populations from population frequency consideration (e.g. 1000 genomes datasets)
  - updated ClinVar version (v2023-04) considered during tiering of known pathogenic variants
- Inclusion of non-coding mitochondrial transfer RNA (tRNA) genes in variant tiering
- Updates to the variant inclusion list
- Updates to STR tiering and STR visualisation behaviour
- Inclusion of copy number variants and *de novo* small variants in consideration of compound heterozygous variants

66.6 Mira

Release information

The major new features that are released in Mira are:

- Upgrade of the DRAGEN software utilised for mapping and variant calling from DRAGEN v3.2.22 to DRAGEN v4.0.5 *(updates throughout user guide)*, including:
  - updated internal allele frequency cohort used during assessment of population frequencies
  - updated approach and thresholds for small variant 'PASS' filters and *de novo* variants
  - updated approach for CNV detection (2-10Kb)
- Upgrade to ExpansionHunter v5 software (within DRAGEN), with introduction of REViewer software for short tandem repeat expansion visualisation
- Tiering of rare variants impacting non-coding green genes in the applied panel (altered biotypes and variant consequences included for consideration of Tier2 variants)
- Variants-with-pathogenic-associations in the Genomics England Clinical Variant Ark are only included as tiered variants if the internal allele frequency is less than 5%, or if variants are on the variant inclusion list
- Heterozygous variants which fulfill relevant tiering criteria are tiered in combination with alternate homozygous variants that impact the same gene and also fulfill relevant tiering criteria under compound heterozygote segregation filters

Minor updates subsequently added to original release of the Mira genome analysis user guide:

- modifications to variant inclusion list
- increased clarity about handling of *de novo* variants
- added details of Ensembl database version used in the pipeline

66.7 Lyra

Release information

The Lyra release of the online user guide is an early access release and replicates the v2.4.1 rare disease user guide available through the NHS Futures website under “NHS Genomic Medicine Service” > “Guidance” > “Genomics England Documentation”).

There are some minor changes of the content and presentation after migration of the PDF to the online user guide:

- "The Clinical Reporting Workflow" section renamed to "Variant Prioritisation Approaches"
- Reordering and joining of Sections 1-7 into "Background"
- Update of population allele frequency information used during small variant tiering
- Removal of Section 9.7 "Short SNV and Indel (small variant) Tiering guide for bioinformaticians"
- Changes in presentation format across all sections, including splitting up of larger sections of the guide and reorganisation of content where appropriate

66.8 Prior to Lyra

Release information prior to Lyra

- information about specific changes to releases prior to Lyra are available in previous versions of the Genomics England rare disease user guide (available as PDFs through the NHS Futures website under “NHS Genomic Medicine Service” > “Guidance” > “Genomics England Documentation”)