GUI-BIO-009 Rare Disease Genome Analysis Guide

GENOMICS ENGLAND CONFIDENTIAL **UNCONTROLLED IF PRINTED**

Document Key	GUI-BIO-009
Title	Rare Disease Genome Analysis Guide
Document Status	Published
Confluence Document Version	V2.2
Published Date	Nov 15, 2022
Policy (only if applicable otherwise N/A)	N/A
Document Author	Jamie Ellingford
Document Reviewer	Dalia Kasperaviciute, Ellen Thomas
Document Approver	Richard Scott
Details of Approval **(Completed by the QI team)**	☒ Approved in Confluence ☐ Pre-Approved in EQMS (Evidence in EQMS) ☐ Pre-approved by email (Needs prior authorisation from the Quality Improvement Team) ☐ Reference document - approval not required
Next Review Date	☒ Default (12 months) ☐ Other - please specify
Training Format	☒ Read and understand on Confluence ☐ Read and understand on EQMS if applicable ☐ Course ☐ Competency Assessment
Squad/Teams/Roles to be Trained	

1 Revision History

ℹ The revision history of each document is available in the Confluence Page History. To view details of what was changed, click on the versions to compare and select "Compare Versions". If the document was previously held in EQMS, the version number and history is superseded by the revision history now held in Confluence as per above.

Confluence Version* (newest on top)	**Date**	**Summary of main changes and reasons**
See Confluence Document Version above	April 2021	Converted to ISO Template
Version 2.0	24th December 2020	This is the first published version of this document
Version 2.1	14th July 2021	Description of analysis for copy number variants 2-10kb Change update in tiering for *de novo* variants in imprinted genes
Version 2.2	10th November 2022	STR prioritization rule changes (9.6) Small variant internal AF data update (9.3.4.29.3.4.2 Clarify super panel mode of inheritance (9.3.4.5) Link to updated coverage profile data (Appendix H)

		Correction of the sample size for internal CNV frequencies (9.5.2)

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

2 Purpose

The purpose of this document is to provide NHS Clinical Scientists, Clinicians, Bioinformaticians and others within the NHS Genomic Laboratory Hubs (GLHs) with a guide to the Genomics England workflow for data analysis and 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.

3 Scope

3.1 In Scope

- Description of the whole genome sequence analysis performed in the Rare Disease bioinformatics pipeline 2.0, including variant calling and interpretation.

3.2 Out of Scope

- Description of the Interpretation Portal or Decision Support tools

4 Target Audience

4.1 Internal Audience

- N/A

4.2 External Audience

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

ℹ **Other Third Party Audience**

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

- 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.

5 Abbreviations/Definitions

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. http://www.internationalgenome.org/category/phase-3/
BAM (or CRAM)	Binary Alignment Map of a participant's genome. A CRAM file is a compressed alternative to a BAM file.
Catalog-OpenCGA	Catalog is been developed to provide authentication, ACLs and to keep track all of the files and sample annotation. OpenCGA is an open-source project that aims to provide a Big Data storage engine and analysis framework for genomic scale data analysis.
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. http://evs.gs.washington.edu/EVS/
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.**
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. http://exac.broadinstitute.org/
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. https://gnomad.broadinstitute.org/
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. http://www.nlgenome.nl/

GRCh37	The human assembly GRCh37 (also known as hg19)
GRCh38	The human assembly GRCh38
HPO	Human Phenotype Ontology.
HPO terms	Human Phenotype Ontology terms
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	JSON (JavaScript Object Notation) 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 https://panelapp.genomicsengland.co.uk/
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
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. https://www.uk10k.org/studies/cohorts.html
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. https://www.uk10k.org/studies/cohorts.html
UPD	Uniparental Disomy.
VCF	Variant Call Format.

6 Introduction/Background

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, reflecting current EuroGentest and ESHG guidelines that "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."

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.

7 Authorities and Responsibilities

N/A

8 The Bioinformatics Pipeline

8.1 Genome build, alignment and variant detection

Rare Disease Pipeline 2.0 is reporting using 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 ALT-aware mapping and variant calling to improve specificity. 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 (v2.5.6) as part of the DRAGEN software. The DRAGEN software v3.2.22 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 being detected using Manta (v1.5) and are not tiered. Tiering is described further in sections 9.3, 9.5 and 9.6.2.The 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 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 in Appendix A – Summary of the analytical pipeline.

8.1.1 *de novo* variant detection

Detection of *de novo* small variants using the DRAGEN algorithm is performed directly from gVCF files, rather than BAM 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 behaviour applies to all chromosomes, including sex chromosomes and the mitochondrial genome. Allele counts for homozygous reference positions can be obtained directly from the BAM file, for example by viewing the BAM 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.

8.1.2 Mitochondrial variant detection

Detection of variants in the mitochondrial genome with the DRAGEN small variant detection algorithm utilises a continuous allele frequency model. This is different from some other variant callers, including Platypus, which use diploid or haploid models. 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.

8.2 Quality control of sequencing data

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:

- md5sum check to confirm integrity of the genomic data transferred from the sequencing provider.
- 95% of the autosomal genome covered at ≥15x calculated from reads with mapping quality >10 and >85x10^9 bases with Q≥30, after removing duplicate reads and overlapping bases after adaptor and quality trimming. (Saliva samples exempt)
- Germline cross-sample contamination performed using VerifyBamID. Samples with >3% contamination are considered as failing.

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.

8.3 Genomic Identity Checks

8.3.1 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:

- Gender: may be pulled from NHS spine, mandatory field, will be used to infer phenotypic sex if other fields left blank
- Phenotypic sex: non-mandatory, should be completed if different from gender
- Karyotypic sex: non-mandatory, 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 section 9.9 for more detail about the Interpretation

Portal). 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 section 9.3). 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.

Table 1

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
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.

8.3.2 SNP identity checks (Sample Matching Service)

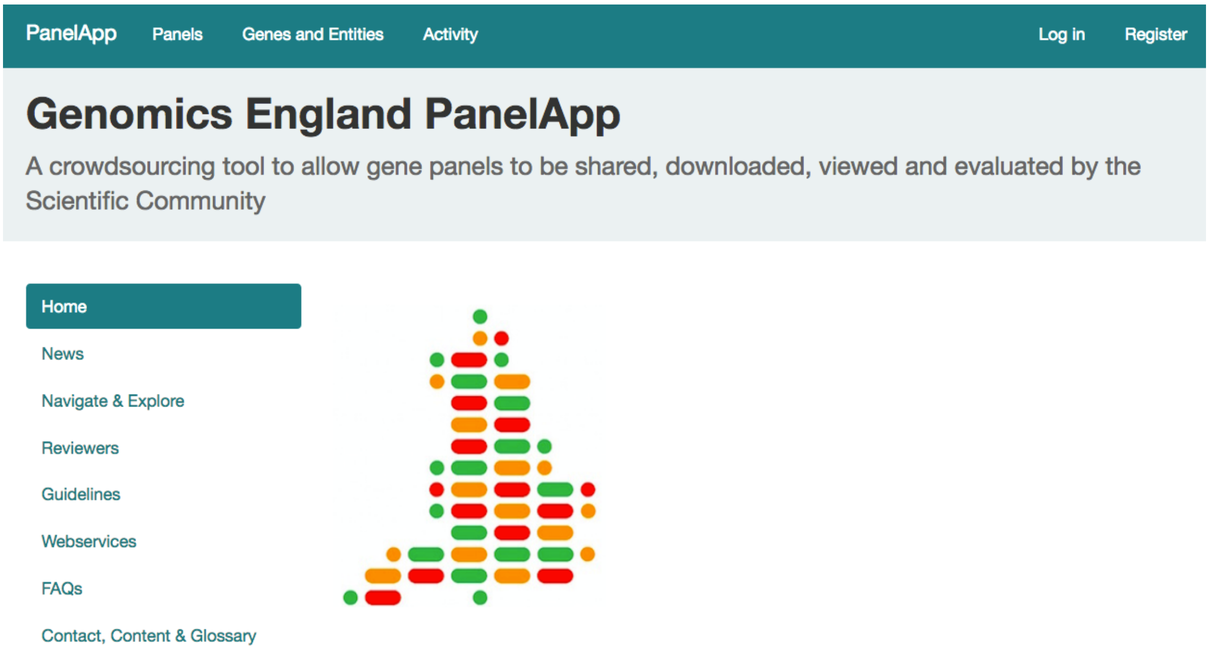
The Sample Matching Service allows positive sample identification by comparing a VCF file produced locally at an NHS GLH (using, for example, a SNP genotyping assay) with a VCF file produced from Rare Disease Pipeline 2.0. For all cases passing genomic identity checks, the results of the sample matching service will be automatically displayed in the Interpretation Portal. When queries are raised by the Genomic Identity checks, the result of the sample matching service will be incorporated into the query investigation and provided to the GLH via the sample failures report. Further details about the Sample Matching Service are available in the Sample Matching Service – Spec for Positive Sample Identification document.

9 The Clinical Reporting Workflow

9.1 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 into 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 NGIS SOP document for guidance.

9.2 PanelApp



Genomics England PanelApp

A crowdsourcing tool to allow gene panels to be shared, downloaded, viewed and evaluated by the Scientific Community

FIGURE 1 Genomics England PanelApp homepage (https://panelapp.genomicsengland.co.uk/)

9.2.1 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 https://panelapp.genomicsengland.co.uk/#!Webservices for more details). As described in greater detail later in this document, 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 (see Appendix B for further details).

9.2.2 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.

9.3 Genomics England Rare Disease SNV and Indel (Small Variant) Tiering Process

9.3.1 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) 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 in FIGURE 2. Tiering can be performed in two penetrance modes: complete or incomplete.

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.

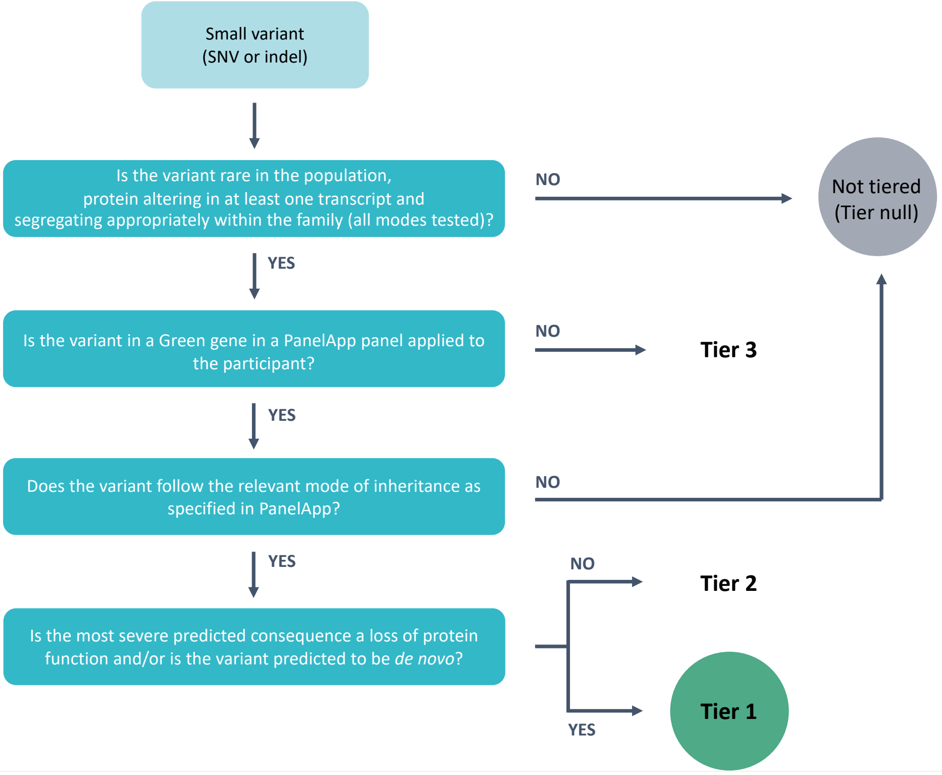

Small variant
(SNV or indel)
Is the variant rare in the population, protein altering in at least one transcript and segregating appropriately within the family (all modes tested)?
NO
Not tiered
(Tier null)
YES
Is the variant in a Green gene in a PanelApp panel applied to the participant?
NO
Tier 3
YES
Does the variant follow the relevant mode of inheritance as specified in PanelApp?
NO
YES
Is the most severe predicted consequence a loss of protein function and/or is the variant predicted to be *de novo*?
NO
Tier 2
YES
Tier 1

FIGURE 2 Simplified overview of the small variant tiering process

During the Tiering process, variants (detected and normalised by the Rare Disease Pipeline 2.0) are annotated and passed through multiple filters (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 (QUAL > 10.4139). At the end of the Tiering process, 2 flags will be assigned to the final variants:

- Tier: Tier 1/ Tier 2/ Tier 3
- Penetrance: Complete/Incomplete

The penetrance analysis can run in two modes:

1. Variants to be reported under complete penetrance
2. Variants to be reported under incomplete penetrance

9.3.2 Brief overview of Tiers

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

- **TIER 1:**
  Includes, high impact variants (e.g., likely loss-of function) and *de novo* moderate impact variants (e.g., missense) within the curated list of Green genes in the panels applied for analysis (available through PanelApp).
- **TIER 2**:
  Includes moderate impact variants (e.g., missense) within the curated list of Green genes in the panels applied for analysis (available through PanelApp).
- **TIER 3:**
  Plausible candidate variants identified with high or moderate impact 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.

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 incorrect.

9.3.3 Tiering algorithm

The Tiering algorithm considers variants or groups of variants in relation to five criteria:

1. FILTER status
2. Allele frequency
3. Predicted functional coding 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

The algorithm can be summarised as follows:

Table 2

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 of 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 high or moderate impact	✓			
If a variant (group) passes all of criteria 1-5 and if not predicted high impact consequence type (excluding *de novo* variants)		✓		

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.

9.3.4 Tiering Algorithm Criteria

9.3.4.1 Criterion 1 – Variant FILTER status (configurable)

Currently 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.

9.3.4.2 Criterion 2 – Allele Frequency (configurable)

In order for a variant to pass this filter, the allele frequency in the data sets cannot exceed the relevant thresholds listed below. All individual population frequencies are considered independently in accordance with the specified thresholds and mode of inheritance.

The current GRCh38 allele frequency annotations and thresholds are as follows:

Table 3

Dataset	Population	Dataset size* (individuals)	Dominant inherited disease	Recessive inherited disease	Mitochondrial genome inherited disease
UK10K	ALSPAC	3,854	0.001	0.01	NA
UK10K	TWINSUK	3,571	0.001	0.01	NA
GNOMAD_GENOMES	AFR	4,359	0.001	0.01	NA
GNOMAD_GENOMES	AMR	424	0.003	0.02	NA
GNOMAD_GENOMES	EAS	780	0.003	0.02	NA
GNOMAD_GENOMES	FIN	1,738	0.002	0.01	NA
GNOMAD_GENOMES	NFE	7,718	0.001	0.01	NA
GNOMAD_GENOMES	OTH	544	0.003	0.02	NA
GNOMAD_EXOMES	AFR	8,128	0.001	0.01	NA
GNOMAD_EXOMES	AMR	17,296	0.001	0.01	NA
GNOMAD_EXOMES	EAS	9,197	0.001	0.01	NA
GNOMAD_EXOMES	FIN	10,824	0.001	0.01	NA
GNOMAD_EXOMES	NFE	56,885	0.001	0.01	NA
GNOMAD_EXOMES	ASJ	5,040	0.001	0.01	NA
GNOMAD_EXOMES	OTH	3,070	0.001	0.01	NA
1kG_phase3	EUR	505	0.002	0.01	0.002
1kG_phase3	SAS	494	0.002	0.01	0.002
GEL_aggCOVID_V4.2-20220117	Custom Genomics England Frequencies	13,699	0.001	0.01	0.001

* maximum number of individuals available for surveillance at each genomic position

NOTE: gnomAD frequencies are extracted from gnomAD v2.0.1.

9.3.4.3 Criterion 3 - Predicted functional coding impact

In order for a variant to pass this filter, it must have a predicted high or moderate impact coding consequence.

The table below lists the Sequence Ontology (SO) terms that are considered to have high and moderate impact consequences. For explanations of the SO terms, please see Appendix C.

Table 4

Consequence types	Sequence Ontology terms
High impact	SO:0001893, SO:0001574, SO:0001575, SO:0001587, SO:0001589, SO:0001578, SO:0001582; SO:0002012
Moderate impact	SO:0001889, SO:0001821, SO:0001822, SO:0001583, SO:0001630, SO:0001626

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 table below lists the biotypes considered. For explanations of biotypes, please see Appendix D.

Table 5

Biotypes	
• IG_C_gene • IG_D_gene • IG_J_gene • IG_V_gene • protein_coding	• nonsense_mediated_decay • non_stop_decay • TR_C_gene • TR_D_gene • TR_J_gene • TR_V_gene

9.3.4.4 Criterion 4 – Segregation with disease

In order to pass this criterion, 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:

Table 6

Mode of Inheritance	Segregation Filter
Biallelic	SimpleRecessive
	CompoundHeterozygous
	UniparentalIsodisomy
monoallelic_not_imprinted	InheritedAutosomalDominant
	deNovo
monoallelic_paternally_imprinted	InheritedAutosomalDominantPaternallyImprinted
monoallelic_maternally_imprinted	InheritedAutosomalDominantMaternallyImprinted
xlinked_biallelic	XLinked SimpleRecessive
	XLinkedCompoundHeterozygous
xlinked_monoallelic	XLinkedMonoallelic
	DeNovo
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.

Variants in each gene that have passed criteria 1-3 above are grouped together in a batch and the segregation filter is applied.

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.

The segregation filters are described in greater detail in section 9.3.5.

9.3.4.5 Criterion 5 – (a) Intersection with high evidence gene on specified gene panel AND (b) match with curated mode of inheritance

In order to pass this criterion, a variant or group of variants must meet two criteria:

- 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.
- Must pass a segregation filter (see Criterion 4) 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*.

Table 7

Tiering Mode of Inheritance	PanelApp Modes of Inheritance
Biallelic	biallelic, monoallelic_and_biallelic, monoallelic_and_more_severe_biallelic, not_provided, unknown
Xlinked_biallelic	xlinked_biallelic, 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
Xlinked_monoallelic	xlinked_monoallelic, 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
Mitochondrial	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.

9.3.5 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.

Table 8

SimpleRecessive	
Single sample Filters	Affected samples are not 'reference_homozygous' or 'heterozygous' NonAffected samples are not 'alternate_homozygous’
Single sample Selection	At least one affected sample is 'alternate_homozygous'
Family Filter	Father and mother cannot be 'reference_homozygous'
UniparentalIsodisomy	
Single sample Filters	Affected samples are not 'reference_homozygous' or 'heterozygous' NonAffected samples are not 'alternate_homozygous’
Single sample Selection	At least one affected sample is 'alternate_homozygous'
Family Filter	Father or mother (and only one of them) is 'reference_homozygous'

Table 9

CompoundHeterozygous	
Single sample Filters	Affected samples are not 'reference_homozygous' or 'alternate_homozygous' NonAffected samples are not 'alternate_homozygous'
Single sample Selection	At least one affected is 'heterozygous' or 'alternate_hemizygous'
Family Filter*	Father and mother are not both reference homozygous for the same variant in the pair.
Special Filter	None of the NonAffected members of the family are heterozygous for both variants in the pair. NOTE: 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

Table 10

XLinkedSimpleRecessive	
Single sample Filters	Affected males are not 'reference_homozygous' or 'heterozygous' NonAffected females are not 'alternate_homozygous’
Single sample Selection	At least one affected is 'alternate_homozygous'
Family Filter	Mother must be 'heterozygous' (if mother is present), Father cannot be affected

Table 11

XLinkedCompoundHeterozygous	
Single sample Filters	Affected females are not 'reference_homozygous' NonAffected are not 'alternate_homozygous'
Single sample Selection	At least one affected female is 'heterozygous'
Family Filter*	Father and mother are not both reference_homozygous for the same variant in the pair. No parent is reference_homozygous for both variants in the pair.
Special Filter	None of the NonAffected females of the family are heterozygous for both variants in the pair. NOTE: 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

Table 12

InheritedAutosomalDominant	
Single sample Filters	Affected samples are not 'reference_homozygous' NonAffected samples are not 'heterozygous' or 'alternate_homozygous'
Single sample Selection	At least one affected is 'alternate_homozygous' or 'heterozygous'
Family Filter	Both Parents are not 'reference_homozygous'

Table 13

InheritedAutosomalDominantMaternallyImprinted/ InheritedAutosomalDominantPaternallyImprinted	
Single sample Filters	Affected samples are not 'reference_homozygous'
Single sample Selection	At least one affected is 'alternate_homozygous' or 'heterozygous'
Family Filter	Maternal Imprinted: - Mother is not ‘alternate_homozygous' or 'heterozygous', if both parents unaffected - Mother is not 'alternate_homozygous' or '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 Paternal Imprinted: - Father is not ‘alternate_homozygous', 'heterozygous', if both parents unaffected - Father is not ‘alternate_homozygous', 'heterozygous’, if father affected - 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

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

Table 14

XLinkedMonoallelicNotImprinted	
Single sample Filters	Affected samples are not 'reference_homozygous' NonAffected females are not 'alternate_homozygous' NonAffected males are not 'alternate_homozygous' or 'heterozygous
Single sample Selection	At least One affected is 'alternate_homozygous' or ‘heterozygous'
Family Filter	Both Parents are not 'reference_homozygous'

Table 15

MitochondrialGenome*	
Single sample Filters	Affected are not 'reference_homozygous'
Single sample Selection	Allele fraction ≥0.05 in affected individuals

*Note that the MitochondrialGenome Segregation Filter is only considered for variants in the mitochondrial genome.

DeNovo*	
	For SNVs, DQ value ≥0.05 AND DN=DeNovo For indels, DQ value ≥0.02 AND DN=DeNovo

*Note that the DeNovo Segregation Filter is considered independently from other segregation filers.

9.3.6 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.

9.3.7 Additional notes regarding Tiering

- 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 section 9.2).
- Tiering of variants on chrX is performed according to an individual’s anticipated number of X chromosomes, inferred from the genomic data. The number of X chromosomes used in tiering will be the same as used in variant calling (see section 8.1), 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 section 9.11). Mode of inheritance is not used in CNV tiering, therefore all CNVs are assigned a tier (see section 9.5 for more details).
- The Segregation Filter for de novo variants is considered independently of other segregation filers. Since the DRAGEN de novo variant detection algorithm considers all variants with a non-Mendelian inheritance pattern and the DeNovo Segregation Filter is independently of other segregation filters, in rare cases, variants may be inappropriately tiered. For example, a variant for which the parental genotypes are heterozygous for a variant allele and homozygous reference and the child is homozygous for the variant allele will be considered as *de novo* and tiered when a monoallelic mode of pathogenicity and complete penetrance is expected.
- 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 2.0) can be phased or unphased, but phase information is currently ignored.
- Information from non-recruited family members may inform likely segregation patterns of variants, but this information is not currently included in the Tiering pipeline.
- The Tiering algorithm does not treat pseudoautosomal regions as autosomal ones.
- 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.

9.4 Exomiser Rare Disease SNV and Indel (Small Variant) Prioritisation Process

9.4.1 Overview

For all rare disease referral, interpretation is performed using the Exomiser automated variant prioritisation framework (Smedley *et al.* 2015 Nature Protocols 10(12):2004-15) developed by members of the Monarch initiative: principally Dr. 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.

9.4.2 Preliminary validation

The Exomiser pipeline was validated on 62, randomly selected, 100,000 Genomes Project cases with a positive diagnosis from the NHS GMCs (50 GrCh37 and 12 GrCh38). The variant(s) reported as diagnostic by the NHS GMCs were correctly returned as the top ranked candidate(s) in 44/62 (71%) of cases (sensitivity = 0.71, precision = 0.71) and in the top 5 for 57/62 (92%) of cases (sensitivity=0.92, precision=0.18). The 5 cases where the diagnosed variant lay outside the top 5 ranked Exomiser candidates included non-coding and non-penetrant diagnoses that Exomiser would not detect with the current pipeline.

Exomiser offers a complementary approach to the panel-based, tiering pipeline as shown below by an analysis of ~200 clinically solved cases. 72% of the diagnoses were identified in the applied gene panels by the tiering pipeline with high precision (1-2 candidates per case). Exomiser identified 81% of the diagnoses in its top 5 ranked results. Combining the tiering and Exomiser results leads to an increased recall of 90% of the diagnoses compared to using either approach alone, with a precision of 0.17, meaning an average of 5-6 variants are presented for consideration.

9.4.3 Genomics England Exomiser Rare Disease Interpretation pipeline

For the Genomics England Rare Disease genome interpretation pipeline, Exomiser was configured to remove all low-quality and non-coding variants and then for each of the modes of inheritance (MOI) being considered (autosomal dominant, autosomal recessive, x-linked dominant, x-linked recessive and mitochondrial), variants compatible with the MOI were retained if below a minor allele frequency of 0.1% (or 2% for compound-heterozygotes) in all of the following reference databases: 100,000 Genomes Project reference samples, 1000 Genomes, ESP, TOPMed, UK10K, ExAC and gnomAD (excluding the Ashkenazi Jewish population).

Exomiser then calculates a score for how rare and pathogenic each variant is (on a scale of 0 to 1) using the above frequency sources and predicted pathogenicity 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).

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 (i) OMIM and Orphanet rare diseases known to be associated with the gene, (ii) mouse and zebrafish models associated with the orthologue of the gene, and (iii) 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 observed is the phenotype in common. The highest score from these comparisons is assigned as the gene-level phenotype 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). Note that 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 with rank = 1 representing the most-likely candidate according to Exomiser and hopefully describing a rare, predicted pathogenic variant that disrupts a gene that has previously been associated with similar phenotypes to the patient.

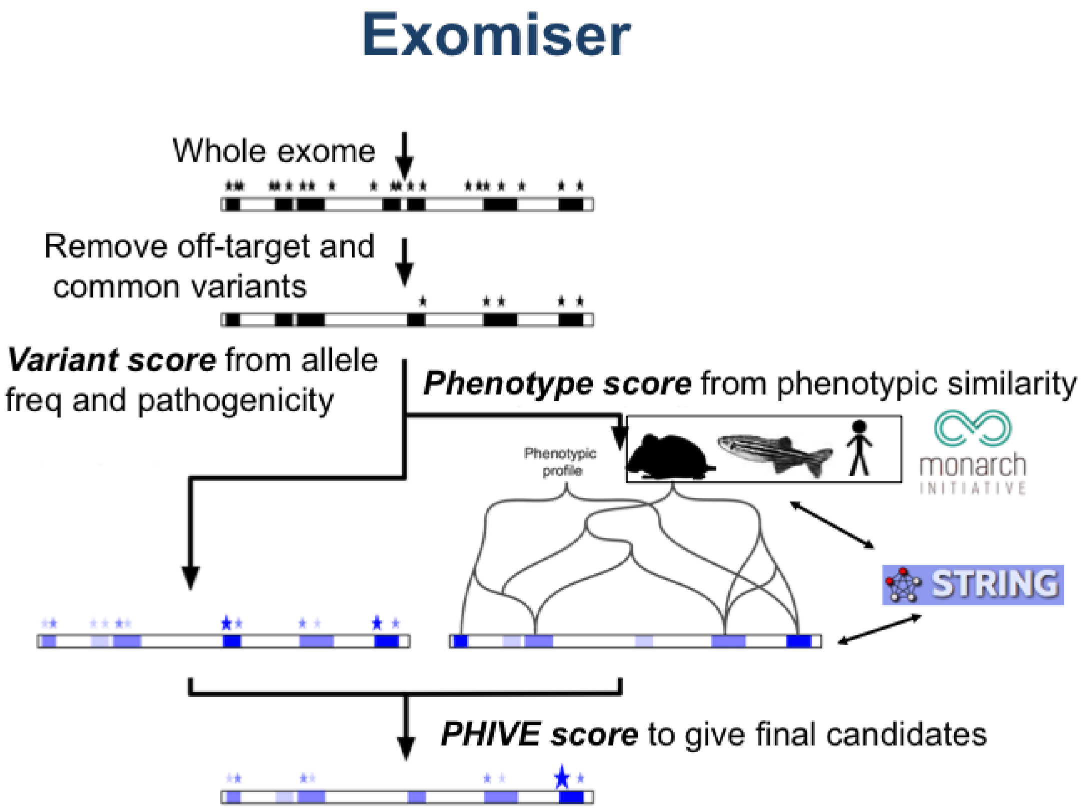

Exomiser
Whole exome
Remove off-target and common variants
Variant score from allele freq and pathogenicity
Phenotype score from phenotypic similarity
Phenotypic profile
monarch INITIATIVE
STRING
PHIVE score to give final candidates

FIGURE 3 Exomiser Overview

9.5 Copy Number Variant Reporting and Tiering

9.5.1 Overview

Copy number variants (CNVs) are detected using DRAGEN CNV (v3.2.22) 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 Manta (v1.5) 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 ≥2 kb in size.

Only CNV calls from the proband are annotated and displayed. CNV calls in relatives are **NOT** currently considered in tiering; this feature is in development for future versions of the pipeline. 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.

The annotations, including internal allele frequencies, for all PASS Gain and Loss calls are displayed in the Interpretation Portal. All PASS calls are assigned a tier, either Tier A or Tier null.

All PASS CNVs are categorised into two tiers:

- **Tier A**. A CNV is assigned Tier A if it satisfies one or both of the following criteria:
  - 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.
  - The CNV overlaps with a green gene in a panel applied in the analysis. All CNVs overlapping any gene are considered, without requiring a minimum overlap threshold. Variant type (i.e., loss or gain) is not considered.
- **Tier Null.** All PASS variants that do not satisfy either of the Tier A conditions.

Please note that the 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.

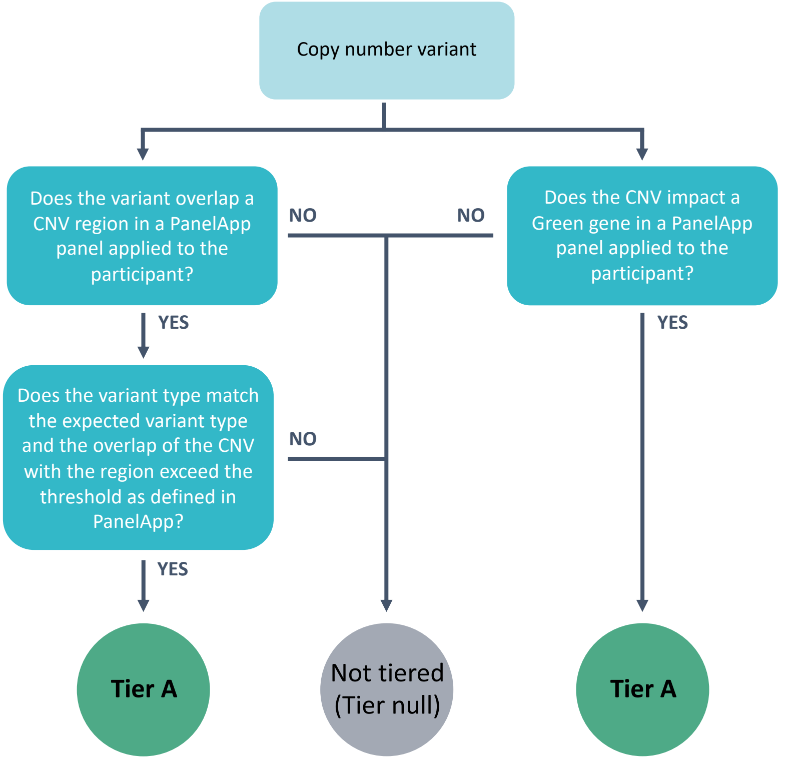

Copy number variant
Does the variant overlap a CNV region in a PanelApp panel applied to the participant?
NO
NO
Does the CNV impact a Green gene in a PanelApp panel applied to the participant?
YES
YES
Does the variant type match the expected variant type and the overlap of the CNV with the region exceed the threshold as defined in PanelApp?
NO
YES
Tier A
Not tiered (Tier null)
Tier A

FIGURE 4 Simplified overview of the copy number variant tiering process

9.5.2 Additional notes regarding CNVs

Detection of CNVs between 2 and 10 kb

Detection of CNVs between 2 and 10 kb was introduced to the Rare Disease pipeline on July 28th 2021 (Danny release). All analysis performed after this date include this pipeline enhancement. For any given referral, the limitations described in the Summary of Findings will indicate if small CNV analysis was performed (see Appendix G for further details). 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 Manta, 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 Manta 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 Manta (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 "<SampleID>.enhanced.cnv.vcf.gz" and the following additional annotations are included in the INFO field:

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

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:

- "Poor_quality_CNV_calls" – most of the CNV calls in the proband are expected to be of poor quality
- "Suspected_poor_quality_CNV_calls" – when a sample is suspected to have an increased number of poor quality CNV calls

The current thresholds are described below:

- "Poor quality CNV calls":
  - if count of autosomal PASS CNVs ≥ 600, or
  - if Log2(Loss/Gain) < -0.5

- "Suspected_poor_quality_CNV_calls":
    - if count of autosomal PASS CNVs ≥ 200 or ≤ 50, or
    - if Log2(Loss/Gain) ≤ -0.3 or ≥ 1.2, or
    - if the fraction of common autosomal PASS CNV calls is <= 0.4. For this purpose, a CNV is defined as common if it has 50% reciprocal overlap with a CNV from Conrad et al. 2010, https://www.ncbi.nlm.nih.gov/dbvar/studies/estd20/.

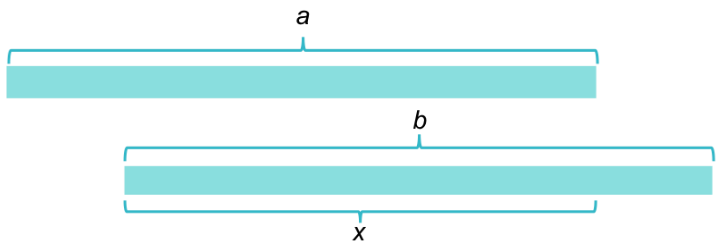
CNV frequency annotation

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

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

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

Reciprocal overlap - defined 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*

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

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

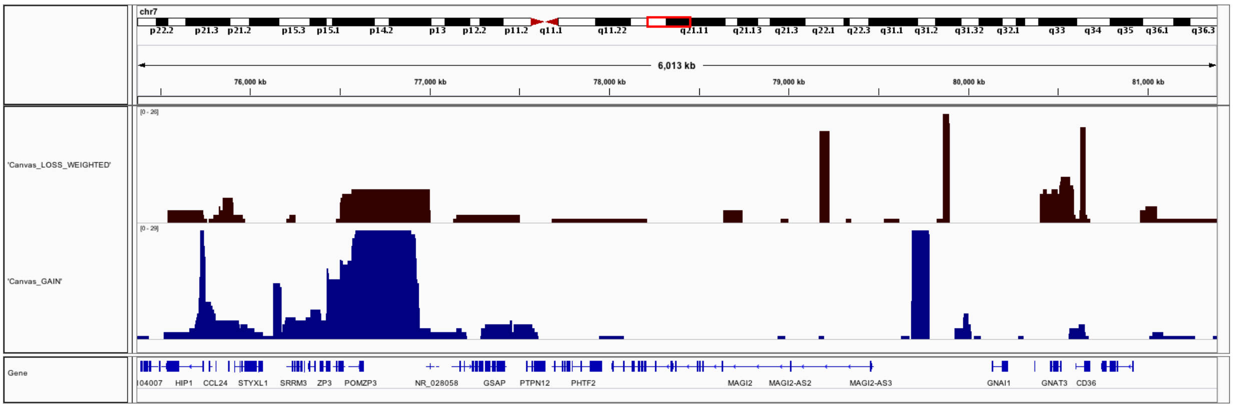


FIGURE 5 Example CNV frequency track

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.**

Green genes containing common CNVs

CNV tiering does not take allele frequencies into account, therefore some common non-pathogenic CNVs are tiered in a large number of probands, if they affect green genes in the panels applied in the analysis. Some examples of this include common CNVs in *PRODH* (Inborn errors of metabolism and Intellectual disability panels), *KANSL1* (DDG2P, Fetal anomalies and Intellectual disability panels) and *OTOA* (Hearing loss panel).

Visualisation and fine-mapping of CNV breakpoints

Additional guidance on visualisation and fine-mapping of CNV breakpoints is available in the Interpretation Portal user guide.

9.6 Short Tandem Repeat Reporting and Tiering

9.6.1 Overview

Expansions of short tandem repeats (STRs) are detected by ExpansionHunter (v2.5.6) as part of the DRAGEN software. STRs are detected only at loci defined in PanelApp (see `STRs` at https://panelapp.genomicsengland.co.uk/panels/entities/). 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. **Note that 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.

Information found in PanelApp (https://panelapp.genomicsengland.co.uk/panels/entities/) relates to the following STR loci:

AR, ATN1, ATXN1, ATXN2, ATXN3, CACNA1A, CNBP, ATXN7, ATXN10, C9orf72, CSTB, DMPK, FMR1, FXN, HTT, JPH3, NOP56, PPP2R2B, and *TBP*.

Information for each STR includes:

- the genomic coordinates of the repeat analysed (both GRCh37 and GRCh38 assemblies)
- the repeat motif or sequence (i.e., `CAG`)
- the normal and pathogenic thresholds as 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.

9.6.2 STR tiering

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

STR loci that are green in PanelApp for the panel(s) applied for analysis will be tiered. Two different ranges of thresholds are used when tiering:

- **Tier 1**. The repeat-length for the locus is greater than or equal to the pathogenic threshold.
- **Tier 2**. The repeat-length is greater than or equal to the threshold for normal alleles but less than the pathogenic threshold
- **Tier Null**. The repeat-length is less that the threshold for normal alleles (defined in PanelApp).

STRs are only tiered for affected members of a family.

For biallelic loci (i.e., *FXN*), affected individuals homozygous for a pathogenic expansion or compound heterozygous (STR and SNV) are also tiered. This is achieved using the same approach as is used for tiering SNVs and Indels using the incomplete penetrance pipeline for an autosomal recessive mode of inheritance. Due to a limitation of the ExpansionHunter algorithm, in some cases, biallelic expansions of the *FXN* expansion may be incorrectly detected as monoallelic expansions, and will be tiered as Tier null rather than Tier 1/2.

It is important to note that population allele frequencies for repeat sizes are not currently being used for STR filtering or tiering.

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 preexpasions in Tier 2.

Tiering of FMR1 was introduced in 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.

9.6.3 STR visualisation

Whenever an STR is green in a panel used in the interpretation, a visualization plot of the reads (pile-up) supporting the detected allele lengths is provided for each family member within the Interpretation Portal. 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 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. +/- 1 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 (i.e., for Spinocerebellar Ataxias, see section 3 below) or pathogenic-borderline cases, before orthogonal confirmation.

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

1. Example of 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 FIGURE 6 Example of a good quality STR call showing alleles within the normal range. Both alleles have an STR repeat-length of 22 and accordingly, reads each containing the `CTG` motif 22 times are visible.

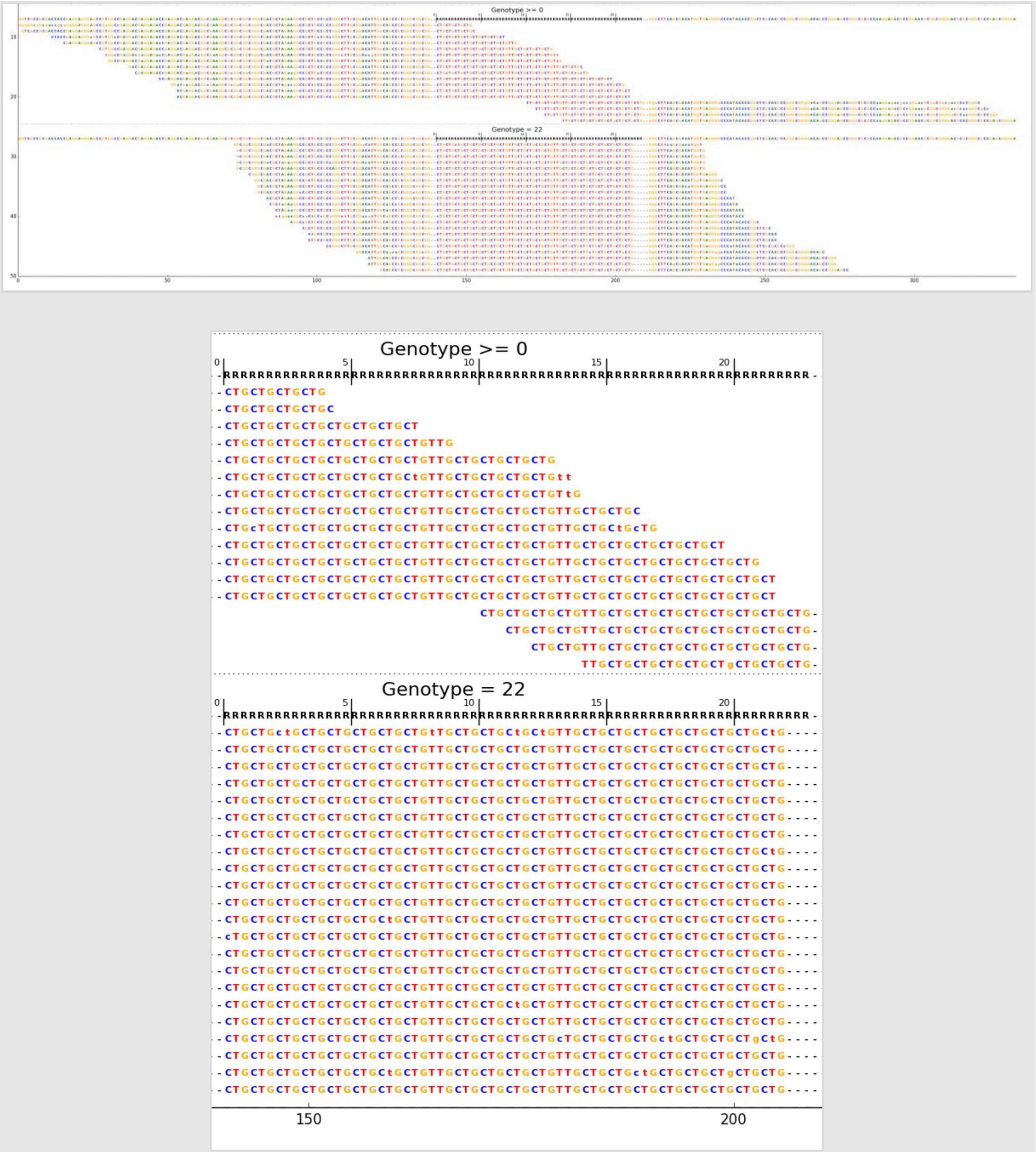


FIGURE 6 Example of a good quality STR call showing alleles within the normal range

2. Example of a good quality STR showing one allele within the normal range and one expanded allele

An example of a visualisation plot illustrating the reads used by ExpansionHunter when estimating expansions in *HTT* is shown in “FIGURE 7 Example of a good quality STR showing one allele within the normal range and one expanded allele”. Alleles of 24 and 47 repeat-lengths are shown in the plot which is divided into 5 subplots:

- *(Genotype>=0)* There are "flanking reads" that are anchored only on one side of the repeat. These support more than 0 repeats but cannot be used to determine the exact number of repeats.
- *(Genotype=24)* There are "spanning reads" that are anchored on both sides of the repeat and support exactly 24 repeats for one of the alleles.
- *(Genotype>=24)* There are additional flanking reads that support more than 24 repeats for the second allele but again cannot be used to determine the exact number of repeats.
- *(Genotype=47)* There are spanning reads that support exactly 47 repeats for the second allele.
- *(Genotype>=47)* There are "inrepeat reads" that are not anchored at either end of the repeat and support at least 47 repeats. Note that these reads do not contradict 47 repeats.

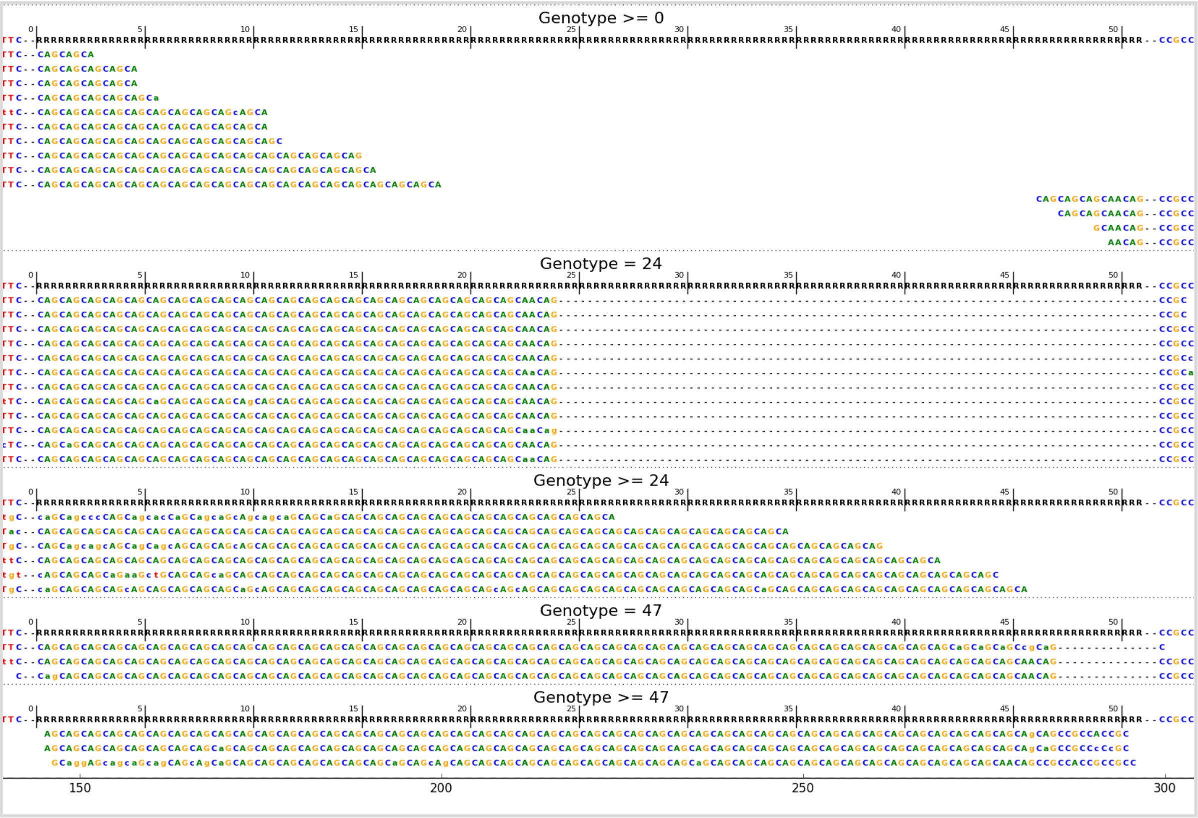


FIGURE 7 Example of a good quality STR showing one allele within the normal range and one expanded allele

3. Example of a good quality STR call with interruptions

An example plot with the reads used by ExpansionHunter when estimating expansions in *ATXN1* is shown in FIGURE 8 Example of a good quality STR expansion with interruptions, with alleles of 30 and 53 repeat lengths. Interruptions (`ATG` rather than `CTG`) in the reads containing the *ATXN1* repeat motif are visible (within the red boxes in FIGURE 8 Example of a good quality STR expansion with interruptions). In certain disorders (i.e., ataxias) it is important to use the visualisation plots to check for such interruptions.

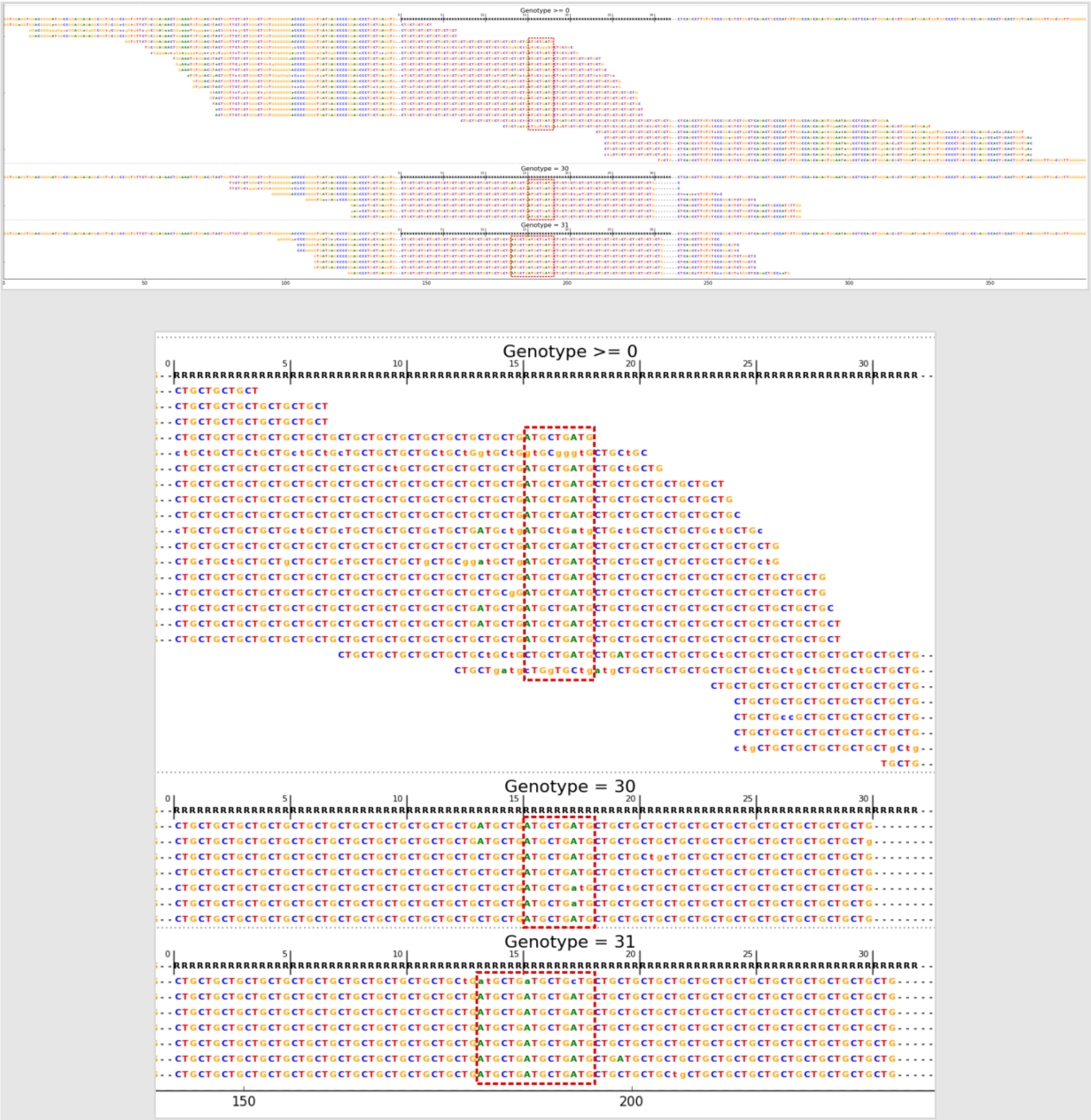


FIGURE 8 Example of a good quality STR expansion with interruptions

4. Example of a poor quality STR call

An example plot showing the reads used by ExpansionHunter uses when estimating expansions at *ATXN1* is shown in FIGURE 9 Example of a poor quality STR call, with alleles of 27 and 43 repeat-lengths. The reads supporting the 27 `CTG` allele are clean good quality reads. However, for the 43 repeat-length allele, lower quality bases are visible at the at the end of the read. Lower quality bases within a read are shown in lowercase, while higher quality bases are uppercase. Furthermore, in this case the lower-quality bases do not follow the `CTG` motif (see red dashed oval in FIGURE 9 Example of a poor quality STR call). In this case, it is likely that

ExpansionHunter should have estimated an allele containing 30-35 repeat-size motif rather than 43.

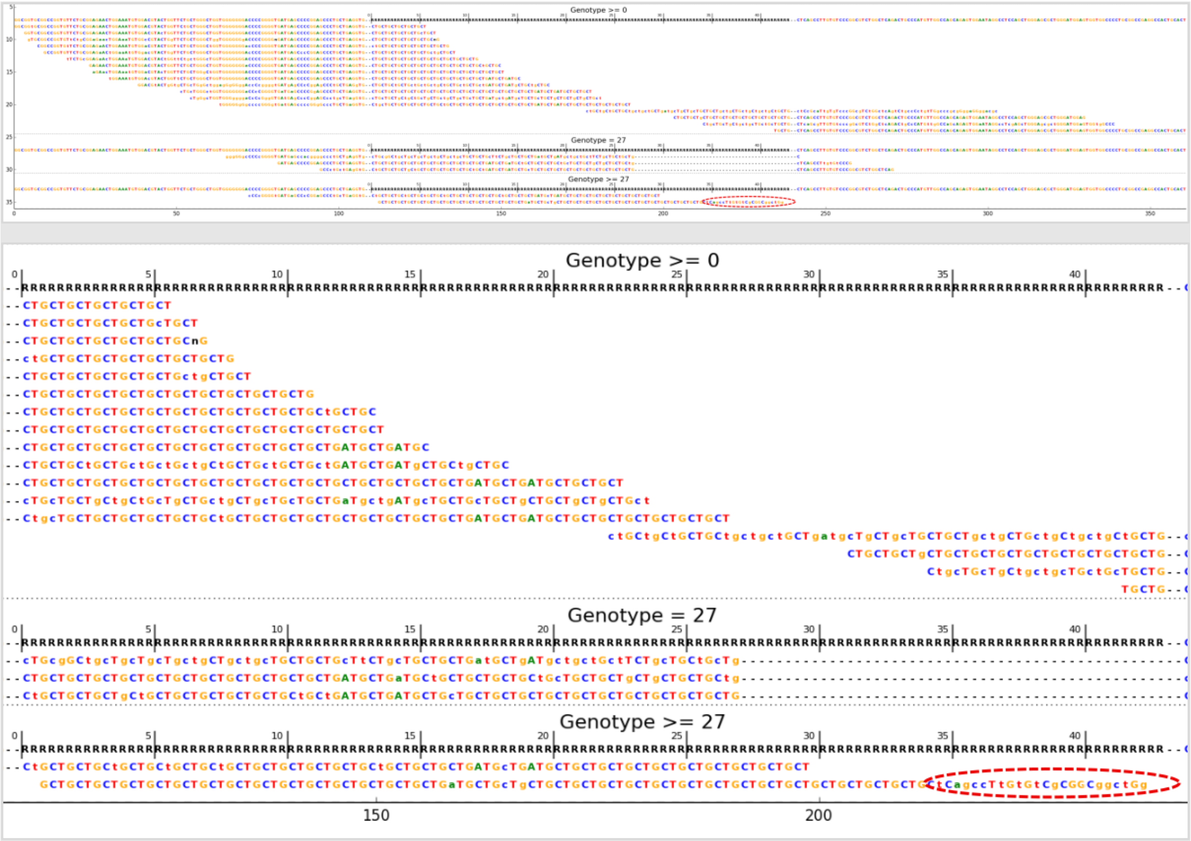


FIGURE 9 Example of a poor quality STR call

9.7 Short SNV and Indel (small variant) Tiering guide for bioinformaticians

Dependencies to run/use the Tiering Pipeline - Bioinformatics

By default, this project relies on Catalog-OpenCGA to get the information needed and store the results (however it could be run independently of OpenCGA)

PythonCommonLibs: This is a library created by Genomics England, where all the common methods are implemented.

PanelApp: WebServices are used in the Rare Disease Tiering (however it can be run without them). See https://panelapp.genomicsengland.co.uk/#!Webservices for webservices on available query options e.g. How to query gene panels that are now retired from PanelApp.

GelReportModels: This library is the result of a big effort in Genomics England to standardise all the information pieces involved in the process of interpretation.

What do I have to know before using it?

GelTiering is a python project made to be consistent and compatible with Genomics England project.

Input/Output data formats are described in GelReportModels.

Rare Disease Tiering takes as input:

A pedigree File as described in GelPedigree

A multisample VCF file produced by DRAGEN v3.2.22 containing genotypes for the samples to be analysed.

A file containing Cellbase annotations for the variants in the VCF file

A configuration file describing the criteria according to which variants are filtered/classified

Rare Disease Tiering output

A JSON file listing Tier 1, Tier 2, Tier 3 variants for small variants and Tier A and Tier null variants for CNVs.

VCF files listing Tier 1, Tier 2 and Tier 3 variants. A separate VCF is produced for each combination of gene panel and mode of inheritance, as well as a combined VCF file

Excel spreadsheet listing Tier 1, Tier 2 and Tier 3 variants. There are a metadata tab and a tab for each combination of gene panel and mode of inheritance.

9.8 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) that segregate with disease status are flagged in the CIP-API and the Interpretation Portal using a label with the format 'ddddddddd [mat|pat]UPDnn [i|h|m][c|p]'.

E.g., '999999999 matUPD14 ic' denotes that complete maternal isodisomy of chromosome 14 was detected in participant 999999999 and segregates with the disease

- ddddddddd 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)

Uniparental disomies are only flagged where:

- the UPD segregates with disease in the family under the assumption of complete penetrance
- the UPD can be specifically identified as a UPD. Regions of homozygosity that could result from either UPD or consanguinity are not flagged

Note that variants showing the appropriate segregation pattern can be tiered under the UniparentalIsodisomy segregation filter and that 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.

9.9 GMS Interpretation Portal

The GMS Interpretation Portal allows users to:

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

The GMS Interpretation Portal is accessible here:

https://cipapi.genomicsengland.nhs.uk/interpretationportal/?category=gms#/

Note: you must be on the HSCN (previously the N3) network

Click “Login with NGIS AD Credentials” – these will be your NHS.net username and password which will have been set up by Genomics England Service Desk for BETA testing.

If encounter login issues please contact the Genomics England service desk (http://bit.ly/ge-servicedesk).

Upon logging into the GMS BETA Interpretation Portal with AD credentials the user will see a list of their cases which are “To Be Reviewed”.

In some cases, flags may be displayed in the Interpretation Portal and in the CIP-API (see section 9.10) with alerts for complex scenarios or potential data quality issues. Please see Appendix E for a summary and description of flags.

Further details of how to use the GMS Interpretation Portal can be found in the accompanying user guide: “Genomics England Interpretation Portal for the NHS Genomic Medicine Service”.

9.10 CIP-API

The 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/

9.11 Decision Support Systems (DSS)

Congenica will be providing 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 portal www.bit.ly/ge-servicedesk

10 Process Flow

N/A

11 Supporting or Reference Documents

Transcribing Whole Genome Sequence Test Requests into NGIS User Guide
Cancelling or Updating a Test Request or Patient Record in NGIS SOP
Genomics England Interpretation Portal for the NHS Genomic Medicine Service
Genomic Laboratory Hubs (GLH) Congenica User Guide Clinical Genomic Interpretation

12 Appendices

Appendix A – Overview of the analytical pipeline

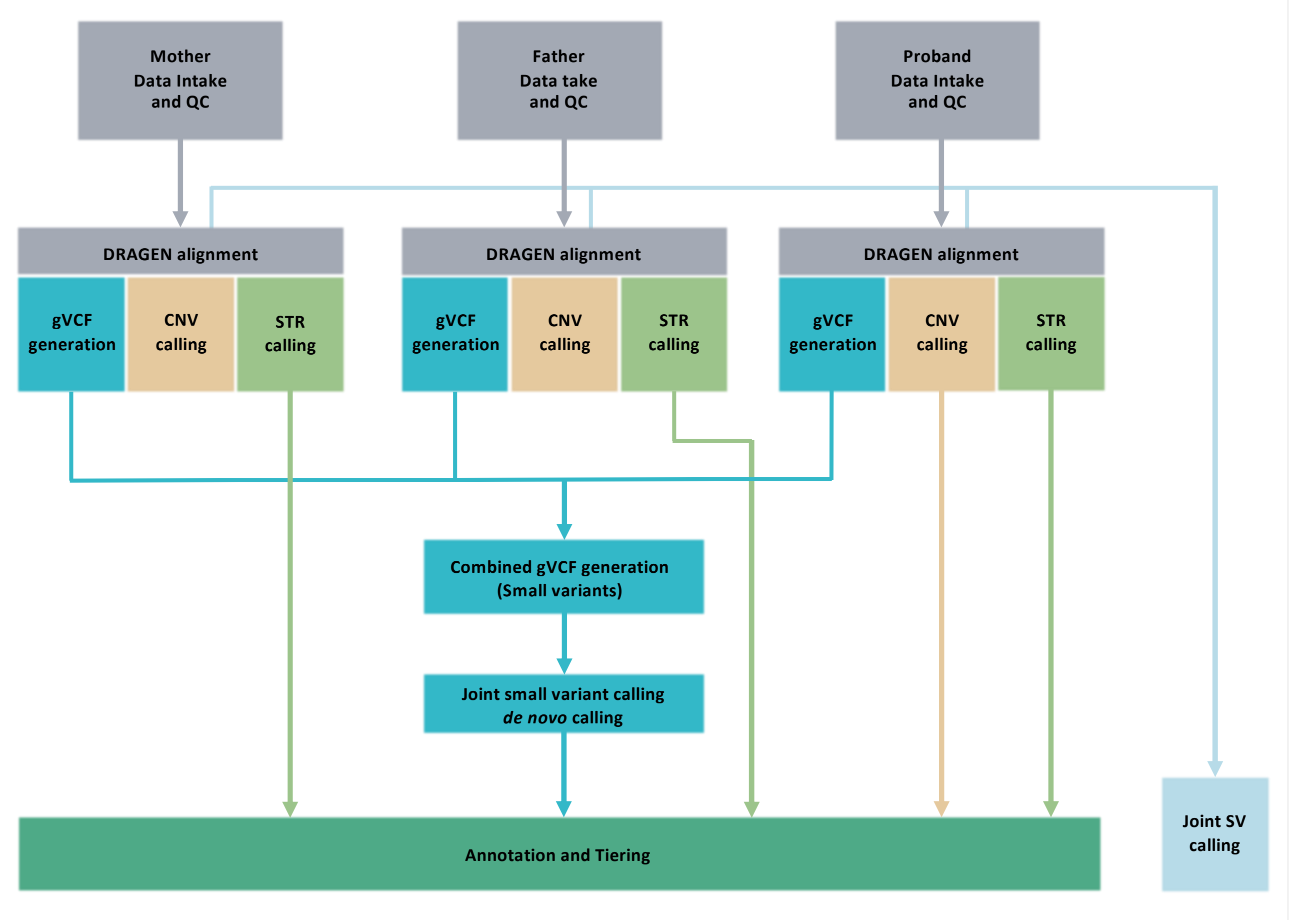


FIGURE 10 A summary of the bioinformatics analytical pipeline for alignment, variant calling and tiering.

Appendix B – PanelApp criteria for diagnostic grade ‘green’ genes

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

OR

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].

OR

C. As definitions 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.

AND

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

AND

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.*

Appendix C – SO terms

Table 16

SO:0001893	
Definition	A feature ablation whereby the deleted region includes a transcript feature.
Synonyms	Jannovar:transcript_ablation, transcript ablation, VEP:transcript_ablation
SO:0001574	
Definition	A splice variant that changes the 2 base pair region at the 3' end of an intron.
Synonyms	Jannovar:splice_acceptor_variant, Seattleseq:splice-acceptor, snpEff:SPLICE_SITE_ACCEPTOR, splice acceptor variant, VAAST:splice_acceptor_variant, VEP:splice_acceptor_variant

SO:0001575	
Definition	A splice variant that changes the 2 base pair region at the 5' end of an intron.
Synonyms	Jannovar:splice_donor_variant, Seattleseq:splice-donor, snpEff:SPLICE_SITE_DONOR, splice donor variant, VAAST:splice_donor_variant, VEP:splice_donor_variant
SO:0001587	
Definition	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened polypeptide.
Synonyms	Seattleseq:stop-gained-near-splice, stop codon gained, ANNOVAR:stopgain, Jannovar:stop_gained, nonsense, nonsense codon, Seattleseq:stop-gained, snpEff:STOP_GAINED, stop gained, VAAST:stop_gained, VAT:prematureStop, VEP:stop_gained
SO:0001589	
Definition	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three.
Synonyms	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	
Definition	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript
Synonyms	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	
Definition	Definition: A codon variant that changes at least one base of the first codon of a transcript.
Synonyms	snpEff:NON_SYNONYMOUS_START, initiatior codon variant, initiator codon change, Jannovar:initiator_codon_variant, VAT:startOverlap
SO:0002012	
Definition	Definition: A codon variant that changes at least one base of the canonical start codon.
Synonyms	Jannovar:start_lost, snpEff:START_LOST, VEP:start_lost
SO:0001889	
Definition	A feature amplification of a region containing a transcript.
Synonyms	transcript amplification, VEP:transcript_amplification

SO:0001821	
Definition	An inframe non synonymous variant that inserts bases into in the coding sequence.
Synonyms	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	
Definition	An inframe non synonymous variant that deletes bases into in the coding sequence.
Synonyms	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	
Definition	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved.
Synonyms	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	
Definition	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.
Synonyms	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	
Definition	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed.
Synonyms	incomplete terminal codon variant, partial_codon, VEP:incomplete_terminal_codon_variant

Appendix D – Biotypes

Table 17

IG_C_gene, IG_D_gene, IG_J_gene, IG_V_gene,TR_C_gene, TR_D_gene, TR_J_gene, TR_V_gene	
Description	Immunoglobulin (Ig) variable chain and T-cell receptor (TcR) genes imported or annotated according to the IMGT.
protein_coding	
Description	Contains an open reading frame (ORF).
nonsense_mediated_decay	
Description	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.

Appendix E – Case flags in the CIPAPI and Interpretation Portal

Table 18

Flag	**Description**
LOW COVERAGE	<95% of the autosomal genome covered at ≥15x calculated from reads with mapping quality >10 and >85x10^9 bases with Q≥30, after removing duplicate reads and overlapping bases after adaptor and quality trimming.
SALIVA SAMPLE	Genome sequencing performed using DNA extracted from saliva. There is a risk of poorer quality sequencing data for saliva samples.
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 on 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.

Appendix F – Genomic 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 2.0 available to view, a summary of which is shown below. The most relevant files for review are shown in **bold**.

Table 19

File	
<SampleID>.repeats.vcf	Short tandem repeat genotypes detected by ExpansionHunter
<SampleID>.enhanced.cnv.vcf.gz	Copy number variants detected by DRAGEN CNV in the proband. Including annotations for high quality small CNVs (2-10 kb) detected by DRAGEN and Manta
<SampleID>.cnv.vcf.gz	Copy number variants detected by DRAGEN CNV in other family members
<SampleID>.forceGT.vcf.gz	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
candidateSmallIndels.vcf.gz	Subset of the candidateSV.vcf.gz file containing only simple insertion and deletion variants of size 50 or less.
<referralID_XXXXX>.diploidSV.vcf.gz	Structural Variants detected by Manta, joint called for family members where available
candidateSV.vcf.gz	Unscored SV and indel candidates detected by Manta. Includes low quality and small (<50bp) variants.
<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

File	
	intervals if there is sufficient signal in flanking regions.
<SampleID>.cram	Genome alignment (CRAM format) generated by the DRAGEN aligner

Appendix G – 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.

Detection of CNVs between 2 and 10 kb was introduced to the Rare Disease pipeline on July 28th 2021. All analysis performed after this date include this pipeline enhancement. For any given referral, the limitations described in the Summary of Findings will indicate if small CNV analysis was performed.

The following text is displayed for cases for which interpretation of CNVs 2-10 kb in size was included. For referrals analysed prior to the small CNV pipeline enhancement, the text displayed will indicate that tiered CNVs are >10 kb in size.

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 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 high impact variants (stop-gain, stop-loss, start-loss, splice donor/acceptor, frameshift, transcript ablation consequence types, or de novo variant in a gene associated with a phenotype with monoallelic mode of inheritance) in genes in the virtual gene panel(s) applied for the patient. Tier 2: Includes moderate impact variants (missense, splice region variant (+/- 8bp from the nearest exon), in-frame insertion/deletion, transcript amplification, incomplete terminator codon) in genes included in the virtual gene panel(s) applied for the patient. Tier 3: Includes high and moderate impact variants in genes not in the virtual gene panel(s) applied for the patient.

Tiered CNVs are high quality calls ≥2 kb in size derived from the proband only. Tier A: Includes CNVs that overlap with genes or contain regions defined in the virtual gene panel(s) applied to the patient.

Tier null: Includes high quality CNV calls ≥2 kb that neither overlap with genes nor contain regions defined in the virtual gene panel(s). Please note that detection of CNVs between 2 kb and 10 kb is not currently included in the ISO15189 schedule of accreditation issued by UKAS. Accreditation is anticipated in Q3 2021. Please refer to the detail of accreditation for further details https://www.ukas.com/wp-content/uploads/schedule_uploads/00007/10170Medical-Single.pdf

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, or they have a lower than expected allelic balance due to mosaicism. 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. 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. Additional details of the Rare Disease analytical pipeline are available in the Rare Disease Genome Analysis Guide.

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.

Table 20

Variant type	Measure	Sensitivity	Precision	Truth set
Single Nucleotide Variants	Mean	0.999	0.996	NA12878 with Genome in a Bottle truth set (high confidence regions)
	95% credible interval	0.999-0.999	0.996-0.996	
Indels	Mean	0.936	0.949	
	95% credible interval	0.935-0.936	0.948-0.950	
Copy Number Variants	Mean	0.971	N/A	Clinically significant CNVs detected by chromosomal microarray in accredited laboratories
	95% credible interval	0.937 – 0.993	N/A	
Short Tandem	Mean	0.959	0.986	STR expansion tests (positive and negative) at targeted loci
	95% credible interval	0.896 - 1	0.947 - 1	

Variant type	Measure	Sensitivity	Precision	Truth set
Repeat expansions				performed in accredited laboratories

Appendix H – 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 and can be found in the same location as this document as REP-BIO-021.