

GUI-BIO-002 Genomics England Rare Disease Results Guide

GENOMICS ENGLAND CONFIDENTIAL **UNCONTROLLED IF PRINTED**

Document Key	GUI-BIO-002
Title	Genomics England Rare Disease Results Guide
Document Status	For Review
Confluence Document Version	v6
Date Published	
Policy (only if applicable otherwise N/A)	N/A
Document Author	@Susan Walker
Document Reviewer	@Dalia Kasperaviciute
Document Approver	@Richard Scott
Details of Approval (Completed by the QI team)	<ul style="list-style-type: none"> ⌚ 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	<ul style="list-style-type: none"> ⌚ Default (12 months) ⌚ Other - please specify
Training Format	<ul style="list-style-type: none"> ⌚ Read and understand on Confluence ⌚ Course ⌚ Competency Assessment
Squad/Teams/Roles to be Trained	

GUI-BIO-002 Genomics England Rare Disease Results Guide	1
1 Revision History	3
2 Purpose	4
3 Scope	4
3.1 In Scope	4
3.2 Out of Scope	4
4 Target Audience	4
4.1 Internal Audience	4
4.2 External Audience	4
5 Abbreviations/Definitions	5
6 Introduction/Background	8
6.1 Notes about the current pipeline and future developments	8
6.1.1 Genome build and alignment	8
6.1.2 Secondary Finding	9
6.1.3 Future developments	9
7 The Clinical Reporting Workflow	11

7.1	Pre-interpretation review and virtual gene panel assignment	11
7.2	PanelApp	12
7.3	Genomics England Rare Disease SNV and Indel (Small Variant) Tiering Process	12
7.3.1	Brief overview of Tiers	13
7.3.2	Tiering algorithm.....	14
7.3.1	Tiering Algorithm Criteria.....	14
7.3.1.1	Criterion 1 – Variant FILTER status (configurable).....	14
7.3.1.2	Criterion 2 – Allele Frequency (configurable)	15
7.3.1.3	Criterion 3 - Predicted functional coding impact OR a 'known pathogenic variant'	16
7.3.1.4	Criterion 4 – Segregation with disease.....	16
7.3.1.5	Criterion 5 – (a) Intersection with high evidence gene on specified gene panel AND (b) match with curated mode of inheritance	17
7.3.2	Segregation filters in action	18
7.3.3	Penetrance modes	21
7.3.4	Additional notes regarding Tiering.....	22
7.4	Exomiser Rare Disease SNV and Indel (Small Variant) Prioritisation Process	22
7.4.1	Preliminary validation	22
7.4.2	Genomics England Exomiser Rare Disease Interpretation pipeline.....	23
7.5	Copy Number Variant Reporting and Tiering	24
7.5.1	Additional notes	25
7.5.1.1	Sample quality control	25
7.5.1.2	CNV frequency annotation	26
7.5.1.1	CNVs on sex chromosomes	27
7.5.1.2	Green genes containing common CNVs	27
7.6	Short Tandem Repeats Tiering	27
7.6.1	STR tiering.....	28
7.6.1	STR visualisation.....	28
7.6.2	Internal allele frequencies.....	32
7.7	Short SNV and Indel (small variant) Tiering guide for bioinformaticians	33
7.7.1	Dependencies to run/use the Tiering Pipeline - Bioinformatics	33
7.8	Uniparental Disomy	34
7.9	Clinical Interpretation Partners (CIPs) and the CIP-API.....	35
7.9.1	Clinical Interpretation Partners (CIPs).....	35
7.9.2	Overview.....	35
7.10	Interpretation Request files (for beginners)	35
7.11	Interpretation Request guide for bioinformaticians	44
7.12	Interpretation Request Field Options	44
7.13	Interpreted Genome files	48
7.13.1	Exomiser Interpreted Genome Scores	48
7.14	CIP-API	49
7.15	Quality Assurance Processes	50
7.15.1	Genomics England Quality Assurance processes.....	50
7.15.1	Release of data to NHS GMCs.....	50
7.16	Interpretation Portal.....	50
7.16.1	How to use the Interpretation Portal	50
7.16.2	Reviewing a case	52
7.16.1	Interpretation Flags.....	53
7.16.2	Sex karyotype flag.....	54
7.16.3	Closing a Case	55
7.16.4	Generating a Summary of Findings.....	55
7.16.5	Reporting Outcomes Questionnaire	55
7.16.6	Family level questions	56
7.16.7	Variant and variant pair level questions.....	57
7.16.8	Linking participant to clinical data in LabKey.....	58
7.16.9	Archived cases	58
7.17	Interpretation Browser.....	59
7.17.1	Overview.....	59
7.17.2	Multiple Monogenic Disorders	60
7.17.3	Exomiser Display of Results	61
7.17.4	Display of CNV and STR results	62
7.17.5	CNV Visualisation.....	63
7.17.6	Download variants.....	64
7.17.7	Generating Summary of Findings using the Interpretation Browser.....	64
7.17.8	Search for variants in genes outside of panel used for Tiering	65
7.18	Live Case Interpretation Support	66

8	Process Flow	66
9	Supporting	66
10	Reference Documents	66
11	Appendices.....	67
11.1	Appendix A – PanelApp criteria for diagnostic grade ‘green’ genes	67
11.2	Appendix B – SO terms	69
11.3	Appendix C – Biotypes.....	71

1 Revision History

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

The previous revision history of the document in EQMS is superseded by the revision history in Confluence as per above.

Confluence Version	Date (Day/Month/Year)	Summary of main changes and reasons (section no. + update)
V0.1	05/06/2017	Summary of the workflow to generate the Clinical Report and close/archive a case. Sent to NHSE for consultation with the GMCs.
V 1.0	05/06/2017	First released draft document
V0.2	21/12/2017	Second released draft document
V 2.0	17/05/2018	Pipeline, Interpretation portal and exomiser updates
V 3.0	09/08/2018	Uniparental disomy (UPD), new de novo tiering thresholds, multiple monogenic conditions, summary of findings and case flag updates
V 4.0	14/03/2019	Panelapp and Interpretation Browser updates, Copy Number Variant (CNV) and Short Tandem Repeat tiering, visualisation and reporting
V 5.0	22/11/2019	Updated Interpretation Portal flags for minor sex karyotypes and low coverage genomes
V6.0	06/12/2021	Updated to new ISO template Updated tiering diagrams Correction of minor errors

**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 Genomics Medical Centres (GMCs) with a step-by-step guide to the Genomics England workflow for clinical reporting of primary findings in Rare Disease. This guide walks you through the processes carried out from the receipt of phenotype and genome sequencing data through to finalising your 100,000 Genomes Project Rare Disease result within your clinical interpretation partner's interface at the NHS GMC.

3 Scope

3.1 In Scope

- N/A

3.2 Out of Scope

- N/A

4 Target Audience

4.1 Internal Audience

- N/A

4.2 External Audience

- N/A

Other Third Party Audience

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

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

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

5 Abbreviations/Definitions

Abbreviation	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/phase3/1
Additional Findings	The variants that have been looked for in addition to Main Findings and consented to by the patient. Referred to as 'Secondary Findings' within Genomics England developed systems.
BAM	Binary Alignment Map of a participant's genome.
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/openccb/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
DDG2P	The Developmental Disorders Genotype-Phenotype Database (DDG2P) is a curated list of genes reported to be associated with developmental disorders, compiled by clinicians as part of the DDD study to facilitate clinical feedback of likely causal variants. The list is categorised into the level of certainty that the gene causes developmental disease (confirmed or probable), the consequence of a mutation (loss-of function, activating, etc) and the allelic status associated with disease (monoallelic, biallelic, etc).
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,

Abbreviation	Description
	and to make summary data available for the wider scientific community. http://exac.broadinstitute.org/
Fabric Genomics	Is a CIP system provided by Omicia.
GeCIP	Genomics England Clinical Interpretation Partnership
GEL	Genomics England
GEL_GL_5277	A set of GRCh37 allele frequencies generated using Illumina single-sample small variant calls for 5,277 germline genomes from pilot participants of the 100,000 Genomes Project. Includes indels of length up to 50 bp. 4,879 genomes are from the Rare Disease programme and 398 genomes are non-saliva 'matched normal' samples from Cancer Programme Participants. Contains genomes from close relatives and genomes from participants with rare diseases as well as some unaffected relatives. Aggregation was performed using agg: https://github.com/illumina/agg
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
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 were 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

Abbreviation	Description
<i>Interpretation Portal</i>	<i>Webpage provided by Genomics England to host clinical reports and used to launch cases into a CIP, using the CIPAPI.</i>
<i>JSON</i>	<i>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.</i>
<i>LabKey</i>	<i>Data Server hosting patient clinical and demographic information, excluding VCFs and BAMs.</i>
<i>LDAP</i>	<i>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.</i>
<i>Main Findings</i>	<i>The variants that have been found and associated with the disease/disorder for which the patient has been recruited to the 100,000 Genome Project. Referred to as 'Primary Findings' within Genomics England developed systems.</i>
<i>MDT</i>	<i>Multi-Disciplinary Team</i>
<i>OMIM</i>	<i>Online Mendelian Inheritance in Man</i>
<i>PanelApp</i>	<i>PanelApp (Open Source) was created to enable virtual gene panels to be viewed and commented on by experts</i> https://panelapp.genomicsengland.co.uk/
<i>PID</i>	<i>Patient Identifiable Data</i>
<i>Platypus</i>	<i>Variant Caller -</i> https://github.com/andyrimmer/Platypus
<i>PMID</i>	<i>is the 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</i>
<i>Primary Findings</i>	<i>The variants that have been found and associated with the disease/disorder for which the patient has been recruited to the 100,000 Genome Project.</i>
<i>Sapientia</i>	<i>Is a CIP system provided by Congenica.</i>
<i>Secondary Findings</i>	<i>The variants that have been looked for in addition to Main/Primary Findings and consented to by the patient.</i>
<i>SNV</i>	<i>Single Nucleotide Variant</i>
<i>STR</i>	<i>Short Tandem Repeat</i>
<i>Tier</i>	<i>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</i>
<i>UK10K_ALSPAC</i>	<i>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.</i> https://www.uk10k.org/studies/cohorts.html
<i>UK10K_TWINSUK</i>	<i>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</i>

Abbreviation	Description
	<i>and Molecular Medicine and is the most detailed clinical adult register in the world.</i> https://www.uk10k.org/studies/cohorts.html
UKGTN	UK Genetic Testing Network
UPD	Uniparental Disomy.
VCF	Variant Call Format.

6 Introduction/Background

The primary diagnostic analysis consented to as part of the 100,000 Genomes Project aims to report back to participants variants with sufficient evidence for diagnostic reporting related to their primary condition. The current clinical reporting workflow is summarised in Figure 1.

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 GMC staff. It should be noted that Genomics England are NOT performing a clinical interpretation of the genome sequencing data. It is the responsibility of NHS GMC staff to perform a full clinical review as would be standard in a diagnostic laboratory, validate the presence of selected variants, 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 GMC 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.

6.1 Notes about the current pipeline and future developments

6.1.1 Genome build and alignment

The pipeline is currently reporting using GRCh37 and GRCh38.

The version of the GRCh37 reference genome used for alignment only includes contigs for the main chromosomes – chromosomes 1 to 22, X, Y and the mitochondrial genome. The version of the GRCh38 reference genome used for alignment comprises 2580 contigs including contigs for the main chromosomes, unlocalised and unplaced contigs, and a contig for the Epstein-Barr virus. It does not include ALT loci (alternate representations of highly variable loci such as MHC and LRC/KIR).

Alignment is performed by Illumina’s Isaac aligner. The pipeline currently calls single variant nucleotides and indels for all contigs used for alignment using the Platypus variant caller, but does not tier or report variants for contigs other than 1 to 22, X, and the mitochondrial genome. Short tandem repeat expansions are called at selected loci (Section 7.6) using ExpansionHunter. Copy number variants (CNVs) are called on chromosomes 1 to 22, X and Y using Canvas software. CNV calls are only reported for genomes aligned to GRCh38 reference.

6.1.2 Secondary Finding

The feedback of secondary findings will occur as a separate process, using different approaches not described here. Guidance on secondary findings reporting will be provided before their release.

6.1.3 Future developments

The tools described in this document represent the systems as they stand. Future developments are planned in a number of areas including:

the ability to view previously encountered variants at the case level across all centres

the ability to pass cases or variants to others within another NHS GMC or GeCIP for their expert input

periodic reanalysis of families using the updated versions of the pipeline (for example, to include copy number calls) and incorporating new knowledge accumulated via GeCIP and external research (for example, newly identified genes); the timeframe for this is currently uncertain

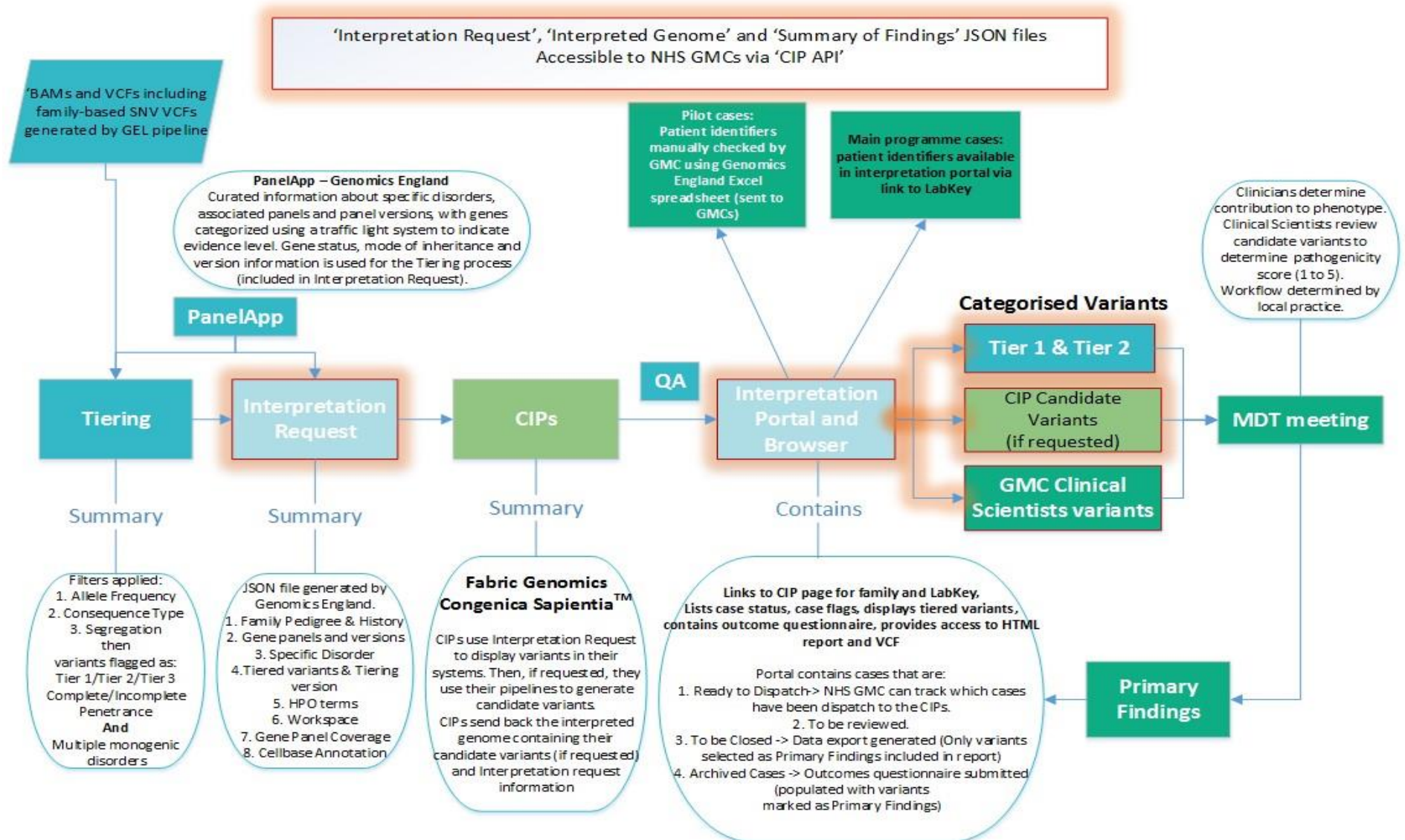


Figure 1: CLINICAL REPORTING WORKFLOW. A GRAPHICAL ILLUSTRATION OF THE WORKFLOW FROM RECEIPT OF CLINICAL DATA FROM THE NHS GMC AND BAM AND VCF FILES FROM ILLUMINA TO PRODUCING THE FINAL CLINICAL REPORT AND ARCHIVING COMPLETED CASES BY THE NHS GMC.

7 The Clinical Reporting Workflow

7.1 Pre-interpretation review and virtual gene panel assignment

Before Tiering takes place, review of the clinical data including HPO terms, pedigree and any additional clinical and phenotypic data submitted by NHS GMCs for each family is performed to:

- I. select the virtual gene panels that will be applied*
- II. select the penetrance settings for Single Nucleotide Variant (SNV) and Indel Tiering (whether the causative genotype is likely to be fully penetrant or incompletely penetrant)
- III. highlight families that need more bespoke analysis, for example where there are likely to be multiple monogenic disorders segregating separately in the family

In the pilot study and main programme, this review has been performed by a clinical geneticist from the Genomics England clinical team. These 'pre-interpretation reviews' are performed in the LabKey clinical data system.

PanelAssigner, an automated tool that suggests panels and penetrance settings for families based on the supplied clinical data has been developed. This is now being used to prepopulate the pre-interpretation review, which can then be manually modified where necessary, for example to add or remove panels or adjust the penetrance setting.

NHS GMC clinical teams can be trained in the use of the system so they can control the Tiering settings for the families that they recruited. Further details regarding this process and PanelAssigner are available via the Genomics England service desk (geservicedesk@genomicsengland.co.uk) or via the portal (www.bit.ly/ge-servicedesk).

*Families will always be analysed using the virtual gene panel relevant to the disorder category that they were recruited under. In addition, reflecting the complexity of many clinical presentations, additional gene panels will be suggested by PanelAssigner based on the HPO terms and other clinical data. For example, a child recruited under the 'Congenital hearing impairment' category who also has intellectual disability, will be analysed with BOTH the 'Congenital hearing impairment' and 'Intellectual disability' gene panels.

7.2 PanelApp

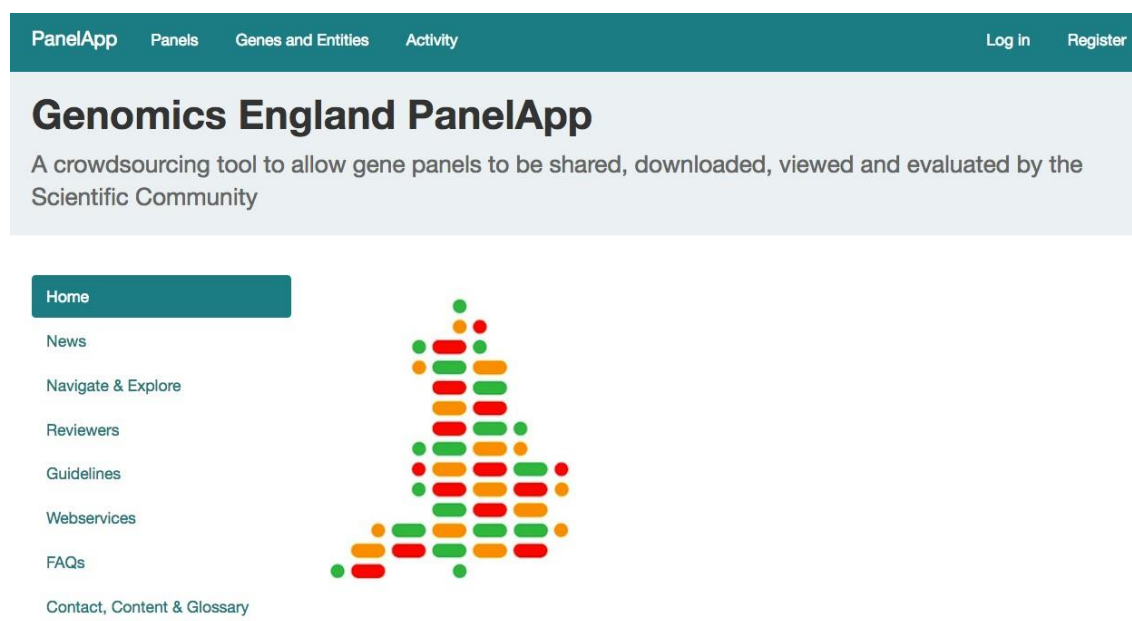


Figure 2 : GENOMICS ENGLAND PANELAPP HOMEPAGE
([HTTPS://PANELAPP.GENOMICSENGLAND.CO.UK/](https://panelapp.genomicsengland.co.uk/))

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 (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 Version 1+ virtual gene panels are used to direct the Tiering process. We encourage NHS GMC teams to continue to contribute their expertise to update existing panels and create new panels. For the most up to date details on how gene panels are defined and how to use PanelApp, refer to the latest version of the PanelApp handbook found on the homepage at <https://panelapp.genomicsengland.co.uk/>.

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

The Genomics England Rare Disease SNV and Indel Tiering Process is to aid NHS GMC 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, effect on protein coding, mode of inheritance and whether they are in a gene in the virtual gene panel(s) applied to the family. The process is summarised in Figure 3. Tiering can be run in two penetrance modes: complete or incomplete.

After Tiering, variants are annotated with a tier (Tier 1, Tier 2, Tier 3 or Empty space (Untiered)) and a penetrance flag (Complete or Incomplete) to indicate the penetrance mode under which they were tiered.

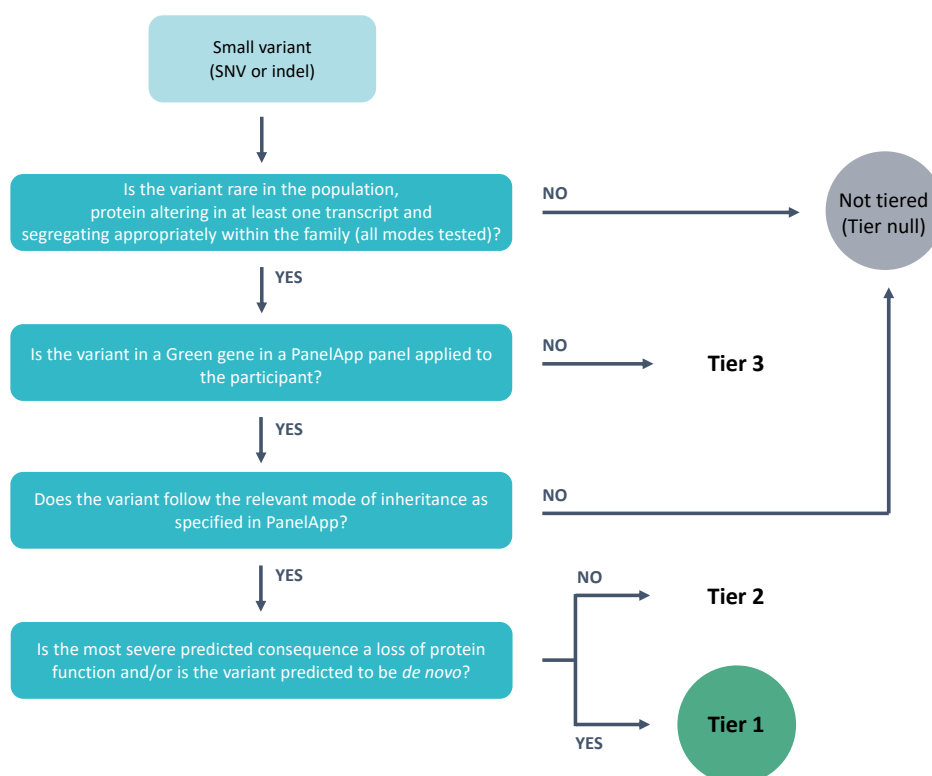


Figure 3: SIMPLIFIED OVERVIEW OF THE TIERING PROCESS

During the Tiering process, variants (which have been previously called, normalised and annotated by the Rare Disease Interpretation Pipeline) pass through multiple filters (allele frequency, consequence type, segregation, quality etc.) in order to classify those that are potentially relevant/causal for a specific case and disease. At the end of the Tiering process, 2 flags will be assigned to the final variants:

- Tier: Tier 1/ Tier 2/ Tier 3 or Empty space (Untiered)
- 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

7.3.1 Brief overview of Tiers

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

- TIER 1: Should be clinically assessed by NHS GMCs.

Includes, high impact variants (e.g. likely loss-of function) and *de novo* moderate impact variants (e.g. missense) within a curated list of Green genes available through PanelApp with sufficient evidence associating them with the patient's phenotype(s). In the future it is anticipated Tier 1 will contain known pathogenic variants once a high confidence curated list has been generated.

- TIER 2: Should be clinically assessed by NHS GMCs.

Includes moderate impact variants (e.g. missense) within a curated list of Green genes available through PanelApp with sufficient evidence associating them with the patient's phenotype(s).

- TIER 3: It is not expected that NHS GMCs will review all of the variants in Tier 3.

Plausible candidate variants identified in genes OUTSIDE of known disease gene panel(s), caution should be used during clinical assessment and interpretation.

Includes high and moderate impact variants outside of the curated list of genes that are associated with the patient's phenotype(s). 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.

7.3.2 Tiering algorithm

5.3.3 Tiering algorithm

The Tiering algorithm considers variants or groups of variants against five criteria:

1. FILTER status
2. Allele frequency
3. Predicted functional coding impact OR (in future) a 'known pathogenic variant'
4. Segregation with disease in the recruited family
5. (a) Intersection with high evidence Green gene on specified gene panel AND
(b) match with the curated mode of inheritance

The algorithm can be summarised as follows:

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 'known pathogenic variant' OR a confident <i>de novo</i> variant	✓			
If a variant (group) passes all of criteria 1-5 and if not predicted high impact consequence type AND not a 'known pathogenic variant' AND not a confident <i>de novo</i> variant		✓		

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 trio of participants cannot be constructed then subsets such as parent-child participant pairs are considered.

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

7.3.1 Tiering Algorithm Criteria

7.3.1.1 Criterion 1 – Variant FILTER status (configurable)

Currently only variants assigned PASS status in the FILTER column of the output VCF are eligible to be classified as Tier 1, 2 or 3.

7.3.1.2 Criterion 2 – Allele Frequency (configurable)

In order for a variant to pass this filter, its allele frequency in the data sets listed below cannot exceed the relevant threshold for that data set, population and the mode of inheritance being considered.

GRCh37 allele frequency annotations and thresholds

Databases	Population Group	Dominant Inherited Disease	Recessive Inherited Disease
ExAC	AFR	0.001	0.01
	AMR	0.001	0.01
	EAS	0.001	0.01
	FIN	0.001	0.01
	NFE	0.001	0.01
	SAS	0.001	0.01
	OTH	0.002	0.01
1000GENOMES_phase_3	AFR	0.002	0.01
	AMR	0.002	0.01
	EAS	0.002	0.01
	EUR	0.002	0.01
	SAS	0.002	0.01
GONL	ALL	0.002	0.01
UK10K_TWINSUK	ALL	0.001	0.01
UK10K_ALSPAC	ALL	0.001	0.01
ESP_6500	EA	0.001	0.01
	AA	0.001	0.01
GEL_GL_5277 (Note: some earlier cases were tiered against earlier sets of frequencies generated using fewer genomes)	Custom Genomics England Frequencies	0.001	0.01

GRCh38 allele frequency annotations and thresholds

Databases	Population Group	Dominant Inherited Disease	Recessive Inherited Disease	Mitochondrial Genome Inherited Disease
GNOMAD_EXOMES	AFR	0.001	0.01	
	AMR	0.001	0.01	
	EAS	0.001	0.01	
	FIN	0.001	0.01	
	NFE	0.001	0.01	
	ASJ	0.001	0.01	
	OTH	0.002	0.01	

Databases	Population Group	Dominant Inherited Disease	Recessive Inherited Disease	Mitochondrial Genome Inherited Disease
1kG_phase3	AFR	0.002	0.01	0.002
	AMR	0.002	0.01	0.002
	EAS	0.002	0.01	0.002
	EUR	0.002	0.01	0.002
	SAS	0.002	0.01	0.002
GEL_GL_ 6628	Custom Genomics England Frequencies	0.001	0.01	

- 7.3.1.3 Criterion 3 - Predicted functional coding impact OR a 'known pathogenic variant'
In order for a variant to pass this filter, it must have a predicted high or moderate impact coding consequence OR be on a list of 'known pathogenic variants'. We DO NOT yet have an approved "white list" of known pathogenic variants, this is currently under development.

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

Consequence Types	
High impact	SO:0001893, SO:0001574, SO:0001575, SO:0001587, SO:0001589, SO:0001578, SO:0001582
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 82 that are associated with certain 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 C.

Biotypes	
<ul style="list-style-type: none"> IG_C_gene IG_D_gene IG_J_gene IG_V_gene IG_V_gene protein_coding 	<ul style="list-style-type: none"> nonsense_mediated_decay non_stop_decay TR_C_gene TR_D_gene TR_J_gene TR_V_gene

- 7.3.1.4 Criterion 4 – Segregation with disease
In order to pass this criterion, a variant or group of variants must pass at least of one of the segregation filters considered. The segregation filters that are considered and their groupings into modes of inheritance are listed below:

Mode of Inheritance	Segregation Filter
Biallelic	SimpleRecessive
	CompoundHeterozygous
	UniparentalIsodisomy
monoallelic_not_imprinted	InheritedAutosomalDominant
	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 7.3.2

7.3.1.5 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/panel being considered has been curated as high evidence ('green' or diagnostic grade) in the PanelApp database in a Version 1+ panel.

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 table below details the PanelApp modes of inheritance that are considered consistent with each mode of inheritance considered by the Tiering process.

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, 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,

Tiering Mode of Inheritance	PanelApp Modes of Inheritance
	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

7.3.2 Segregation filters in action

To illustrate the principals of the segregation filters, we describe below how they work in a simple trio in a full penetrance analysis.

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

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'

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.

CompoundHeterozygous	
Special Filter	None of the NonAffected members of the family are heterozygous for both variants in the pair.

*Each pair of variants in the gene are taken together for the family filter

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

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.

*Each pair of variants in the gene are taken together for the family filter

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'

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	<p>Maternal Imprinted:</p> <ul style="list-style-type: none"> - 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
	<ul style="list-style-type: none"> - Father of unaffected participant (being unaffected 'heterozygous' or 'alternate_homozygous') is not 'alternate_homozygous', 'heterozygous', if father is affected <p>Paternal Imprinted:</p> <ul style="list-style-type: none"> - 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.

XLinkedMonoallelicNotImprinted	
Single sample Filters	Affected samples are not 'reference_homozygous' NonAffected females samples are not 'alternate_homozygous' NonAffected males samples 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'

MitochondrialGenome*	
Single sample Filters	Affected are not 'reference_homozygous'
Single sample Selection	At least one affected is 'heterozygous' or 'alternate_homozygous'

Note: that this Segregation Filter is only considered for variants in the mitochondrial genome.

DeNovo*	
	Maximum fraction of reads supporting the alternate allele in a parent is 3% Minimum fraction of reads supporting the alternate allele in child is 10% At least two reads must support the alternate allele in the child The posterior probability that the variant is <i>de novo</i> exceeds 50%

*The code to apply this filter is a modification of a script which is provided as part of the Platypus project.

7.3.3 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.

7.3.4 Additional notes regarding Tiering

- A single heterozygous variant identified in a gene with a biallelic mode of inheritance will not be assigned a tier.
- 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 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.
- 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 minor contigs (only used for GRCh38 alignment) are not currently considered for tiering.

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

All rare disease cases are now run through the Exomiser automated variant prioritisation framework (see Error! Bookmark not defined.Error! Reference source not found.:

References) 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.

7.4.1 Preliminary validation

Our Exomiser pipeline has been 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 nonpenetrant 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.

7.4.2 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 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.

For a review of how the Exomiser score and rank are displayed in a reported variant, please see section 7.13.1

Exomiser

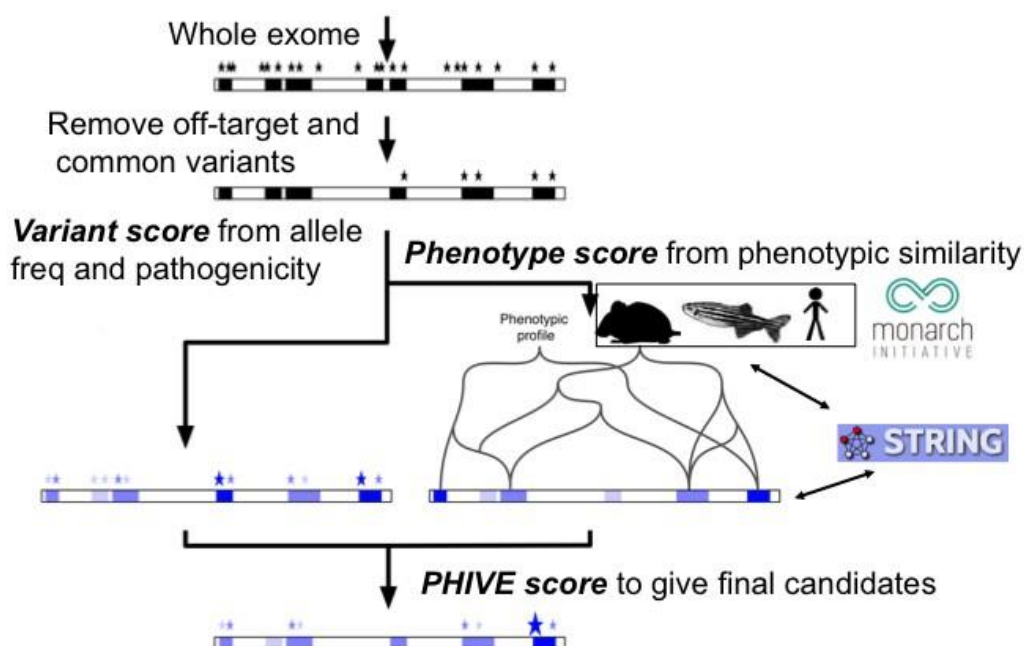


Figure 4 EXOMISER OVERVIEW

7.5 Copy Number Variant Reporting and Tiering

Copy number variant (CNV) calls are produced by Canvas software and are based on sequence coverage and SNV and Indel variant call information. The Genomics England Rare Disease Interpretation Pipeline annotates and reports CNV calls that have PASS filter status assigned by Canvas, i.e. when a call quality score (QS) is ≥ 10 and CNV size is $\geq 10\text{kb}$. Only CNV calls from the proband are annotated and displayed. CNV calls in relatives are NOT currently considered; we hope to improve this in future versions of the pipeline.

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

All PASS CNV variants are categorised into two tiers:

- Tier A. The variant is assigned Tier A if it satisfies one or both of the following criteria:
 - The CNV overlaps the pathogenic region in a panel applied to a participant, the overlap is above the threshold defined in PanelApp for that region, and the variant direction matches (i.e. Gain or Loss) that of the region in PanelApp. Please note, that the mode of inheritance is not considered. In practice in contrast to SNV and Indel (small variant) tiering, a single heterozygous CNV within or encompassing a biallelic gene will be tiered.
 - The CNV overlaps with the green gene in a panel applied to a participant. Any overlap in any gene region is considered. Please note that the mode of inheritance and variant direction is not considered.
- Tier Null. All PASS variants that do not satisfy any of the Tier A conditions.

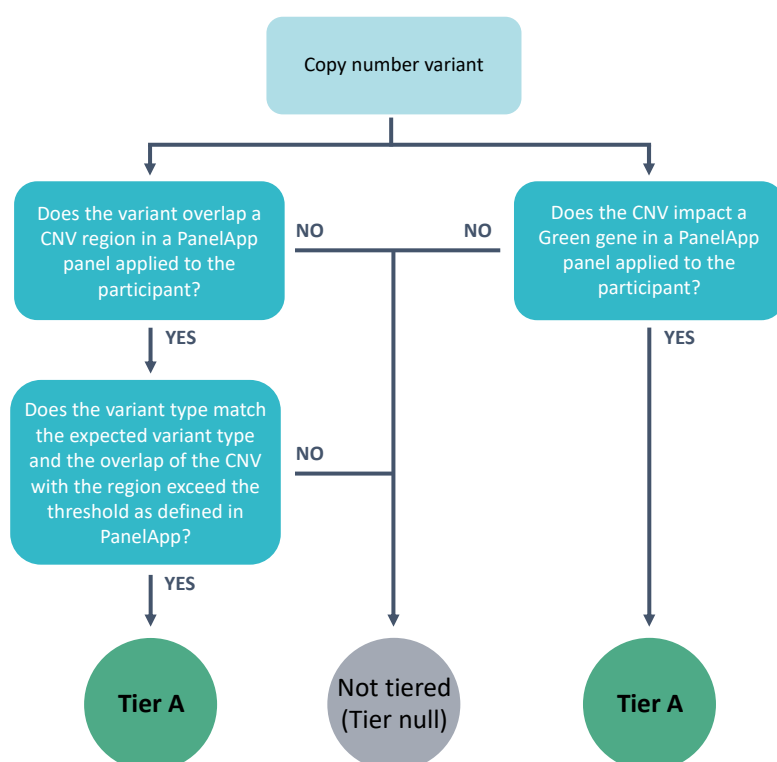


Figure 5: Simplified overview of the copy number variant tiering process

7.5.1 Additional notes

7.5.1.1 Sample quality control

For a small proportion of samples, the sequencing data is of insufficient quality to make reliable CNV calls. We perform sample level quality control based on a number and ratio of different call types and the proportion of common CNVs in a sample. If a sample does not pass this quality control step, then depending upon the value of the QC metric, the CNVs either are not annotated and tiered at all, or are flagged in the Interpretation Portal with one of the following flags:

- “suspected_poor_quality_CNV_calls” – when a sample has an increased number of all types of CNV calls (i.e. both Gain and Loss calls)
- “suspected_increased_number_of_false_positive_heterozygous_LOSS_calls” – when a sample has a suspected increase of false positive heterozygous LOSS calls, but the quality of the other types of calls (Gain and homozygous Loss calls) are not affected.

The current thresholds are described below:

- Do not pass sample through CNV interpretation:
 - if count of autosomal PASS CNVs ≥ 300
- Pass sample through CNV interpretation, but flag in the interpretation portal as “suspected_poor_quality_CNV_calls”:
 - if count of autosomal PASS CNVs ≥ 140 or ≤ 20 , or
 - if $\text{Log2}(\text{Loss}/\text{Gain}) \leq -1.25$ or ≥ 1.35 , or

- if the fraction of common autosomal PASS CNV calls is ≤ 0.3 . 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/>
- Pass sample through CNV interpretation, but flag in the interpretation portal as “suspected_increased_number_of_false_positive_heterozygous_LOSS_calls”:
 - if sample passes thresholds above, but count of autosomal PASS LOH calls ≥ 30 . We do not annotate LOH (loss of heterozygosity) calls that are emitted by Canvas, but use them for quality control purpose only.

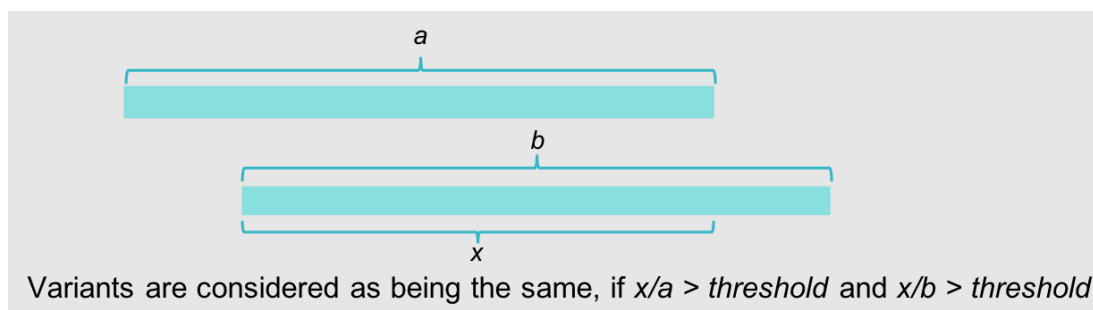
7.5.1.2 CNV frequency annotation

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 would have different breakpoint coordinates in different individuals.
- Large CNVs can be reported as several separate calls (i.e. fragmented calls).
- We cannot distinguish well between different combinations of alleles that can give rise to the same copy number. E.g. a copy number of 3 can be a result of a tandem duplication with 2 copies on one chromosome and a normal other on the other chromosome, or a tandem duplication with 3 copies on one chromosome and a deletion on the other chromosome, or two normal alleles and a copy somewhere 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.

- Reciprocal overlap, defined as shown in a figure below. We use 80% reciprocal overlap threshold. The limitations of this method include its sensitivity to call fragmentation, i.e. fragmented calls can appear as rare, and the possibility of important genes being in the “uncommon” part of the CNV.



- Frequency track – area under the curve method. For that, first CNV calls from multiple samples (22,675 rare disease programme and 6,019 cancer germline samples) are compared. Each base in each of the samples genomes is annotated with the number of chromosomes with a CNV overlapping it, as shown in a Figure 5 below. Then we calculate area under the curve for each CNV and weight it by the maximum possible area (i.e. if allele frequency equals 1).

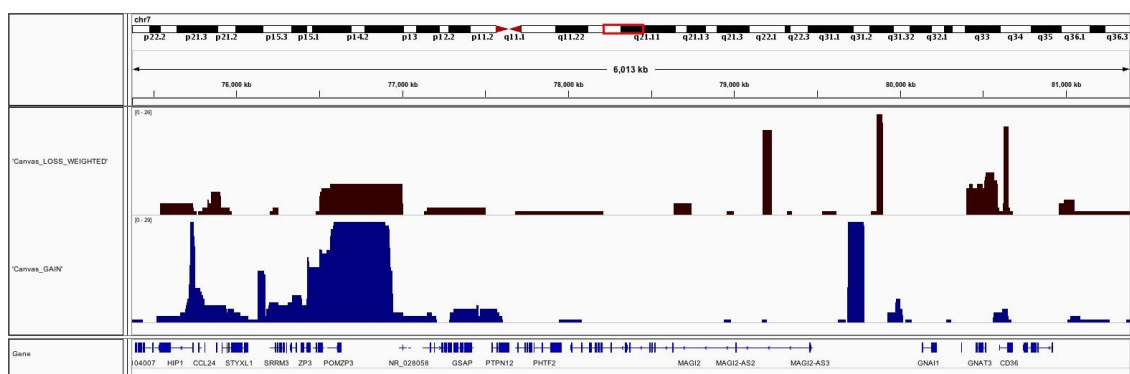


Figure 6: EXAMPLE CNV FREQUENCY TRACK

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

For LOSS variants, we calculate and report allele frequencies. **For GAIN variants, due to difficulties in determining the exact copy number and defining the alleles in all individuals, we calculate and report the proportion of individuals with any GAIN call, not taking copy number into account.**

7.5.1.1 CNVs on sex chromosomes

When calling CNVs on sex chromosomes, the Canvas CNV caller uses inferred sample sex karyotype as inferred by Illumina to define the expected chromosome X and Y ploidies. In a small proportion of samples this sex inference is not correct, which leads to incorrect interpretation of the copy number of the sex chromosome. E.g. if a real sex karyotype is XY, but the CNV calling pipeline assumes that it should be XX, most of the chromosome X regions will be considered to have heterozygous LOSS variants. We check whether the sex karyotype used by the CNV calling pipeline matches the sex karyotype inferred using a more robust method. If a discrepancy is detected, a sample is flagged to note what sex karyotype was used during CNV calling. The relevant flags are "CNV_calls_assumed_XX_karyo" and "CNV_calls_assumed_XY_karyo".

7.5.1.2 Green genes containing common CNVs

CNV tiering does not take allele frequencies into account, therefore some common nonpathogenic CNVs are tiered in a large number of probands, if they affect green genes in the panels applied to participants. Some examples of this include common CNVs in *PRODH* and *KANSL1* genes in the Intellectual Disability panel.

7.6 Short Tandem Repeats Tiering

Short tandem repeats (STRs) or expansions are detected by running Expansion Hunter (<https://github.com/Illumina/ExpansionHunter>). This software has originally been implemented at Illumina, San Diego (Mike Eberle's group). Genomics England has been collaborating with them improving the sensitivity of the tool.

The Interpretation Pipeline runs Expansion Hunter on the loci we have previously defined on PanelApp (see `STRs` at <https://panelapp.genomicsengland.co.uk/panels/entities/>) and reports only expansions detected in affected participants. All VCF files generated in this step are ready to download, containing all loci analysed.

An up to date list and information about specific STR loci and which gene panels each is included within 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 assigned to the case 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, ATXN7, ATXN10, C9orf72, CSTB, DMPK, FMR1, FXN, HTT, JPH3, NOP56, PPP2R2B, and TBP.

Information for each STR includes: (1) the genomic coordinates of the repeat analysed (both GRCh37 and GRCh38 assemblies), (2) the repeat motif or sequence (i.e. `CAG`) and (3) the normal and pathogenic number of repeats associated with each locus. Repeat-lengths larger than the pathogenic threshold will be considered Tier 1, repeat-lengths in between normal and pathogenic thresholds will be considered Tier 2, and repeat-lengths smaller or equal to the normal threshold will not be reported. The internal repeat thresholds have been consulted on and agreed with the NHS GMCs, and are an essential aspect of STR tiering.

Note that although the *FMR1* STR is green in PanelApp, it is NOT reported through the Genomics England interpretation pipeline. More work needs to be done to improve the reliability of the *FMR1* expansion calls before this locus can be included.

7.6.1 STR tiering

The strategy employed by Expansion Hunter when estimating repeat-sizes is to provide confidence intervals and an average for each allele (i.e. $x-y$ and $\text{avg}(x,y)$). The maximum value (i.e. y) of these estimations is taken for each allele and locus. For loci within chromosome X, male genomes are considered as haploids.

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

- Tier 1. The repeat-length for the locus is beyond the pathogenic threshold.
- Tier 2. The repeat-length is between the normal and pathogenic thresholds (defined in PanelApp).

Only affected participants/members in the family are tiered.

For biallelic loci (i.e. *FXN*), affected individuals homozygous for a pathogenic expansion or compound heterozygous (STR and SNV) are also tiered. This is done using the same approach as is used for tiering SNVs and Indels using the incomplete penetrance pipeline for an autosomal recessive mode of inheritance.

It is important to note that internal population repeat size frequencies are not used for STR filtering or tiering.

7.6.1 STR visualisation

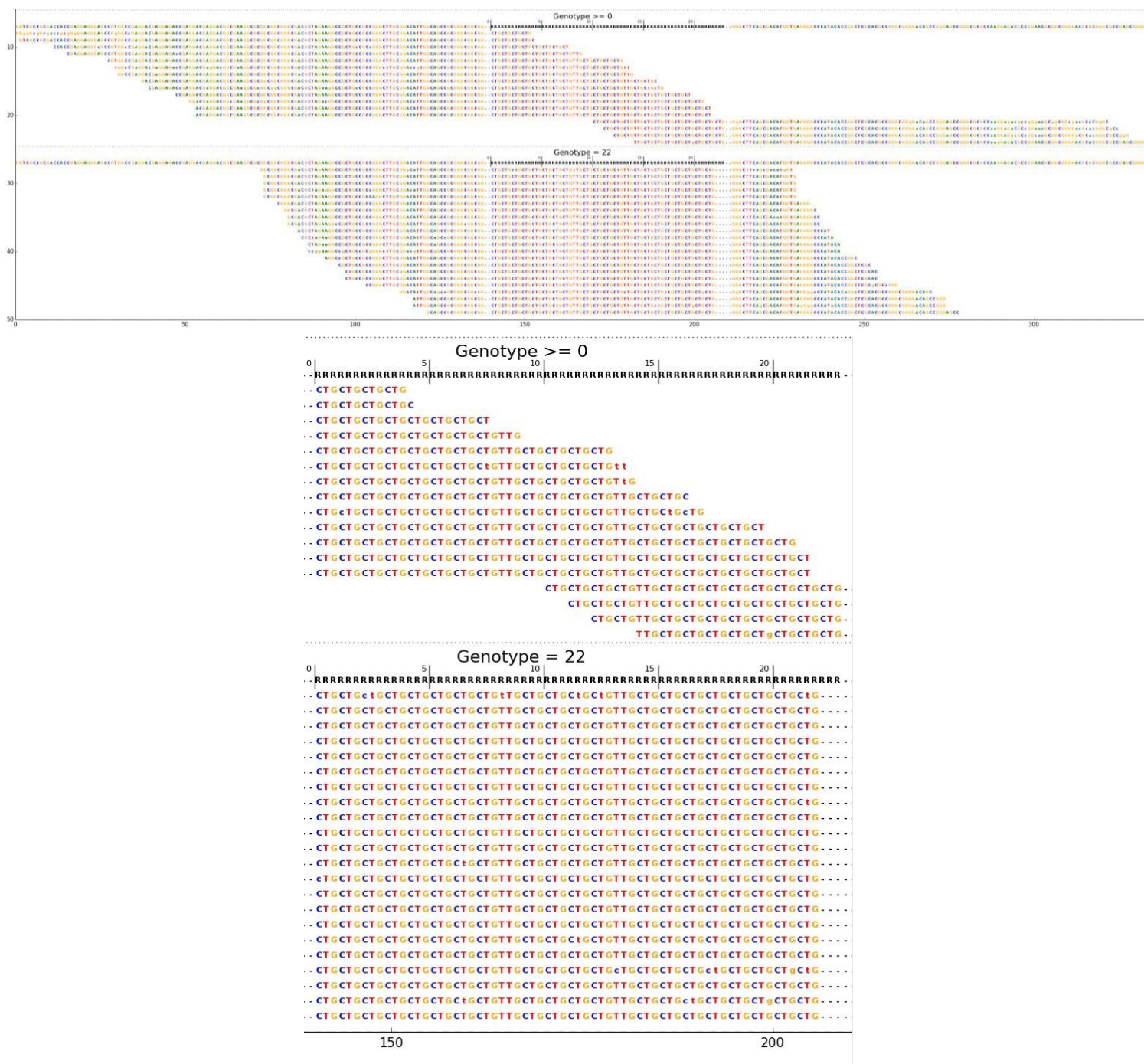
Whenever an STR is reported, a visualization plot of the reads (pile-up) supporting the expansion is provided within the Interpretation Portal (see: STR results). Reviewing this plot is fundamental to the process of assessing the quality of the repeat size estimations computed by Expansion Hunter. Genomics England strongly advises GMCs to use the visualization plot to assess the quality of each call before validation.

Analysing the reads that Expansion Hunter 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) or pathogenic-borderline cases, before orthogonal validation.

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

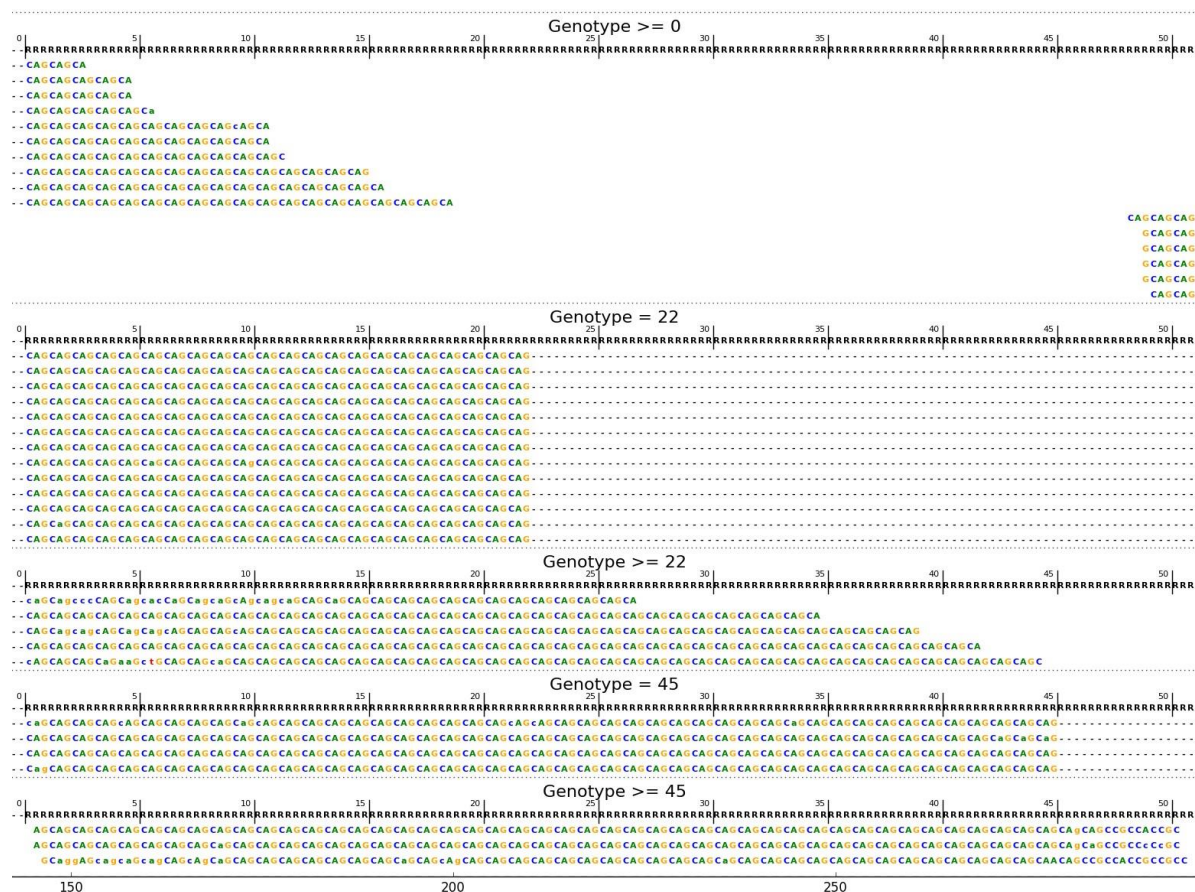
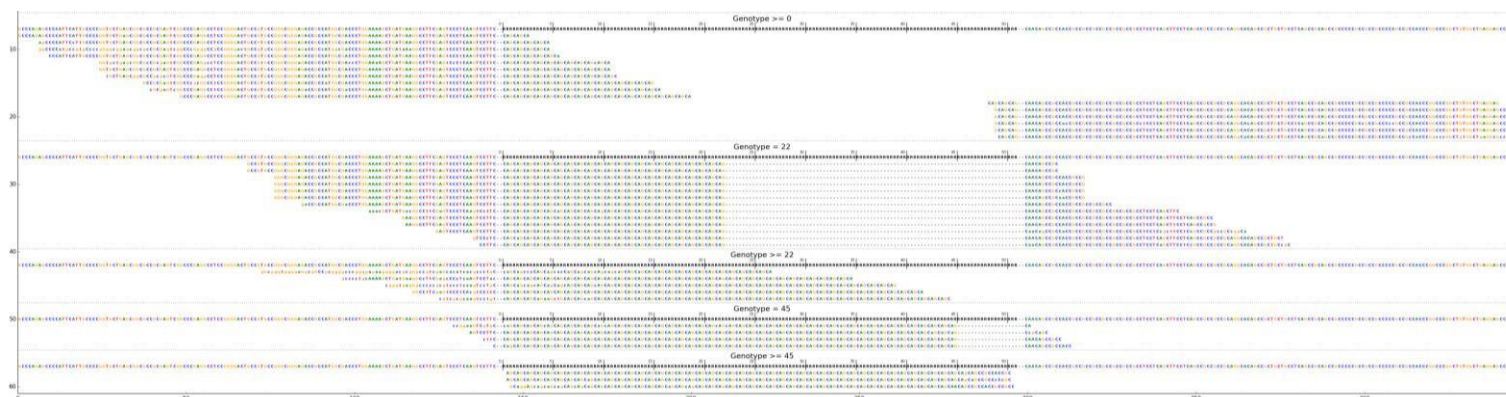
- Examples of a good quality STR call showing alleles within the normal range

Here is an example of a visualisation plot that illustrates the reads that Expansion Hunter uses when estimating expansions in *HTT*. Both alleles have an STR repeat-length of 22. Below you can see 22 reads each containing the `CTG` motif 22 times.



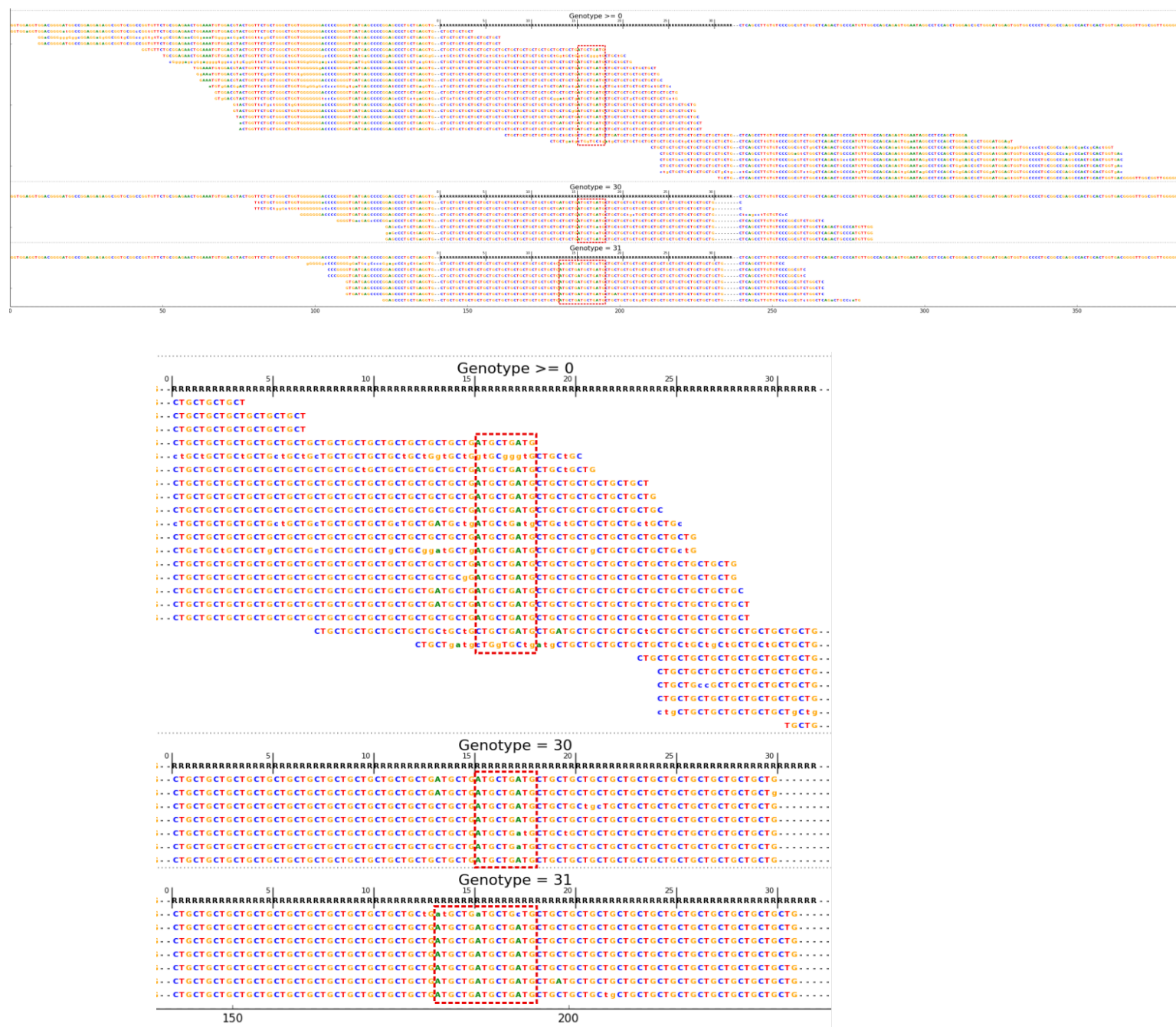
- Example of a good quality STR call showing one allele within the normal range and one expanded allele

Here is a second example of a visualisation plot that illustrates the reads that Expansion Hunter uses when estimating expansions in *HTT*. This time alleles of 22 and 56 repeat lengths are shown. The reads supporting the 22 repeat-length motif are clear.



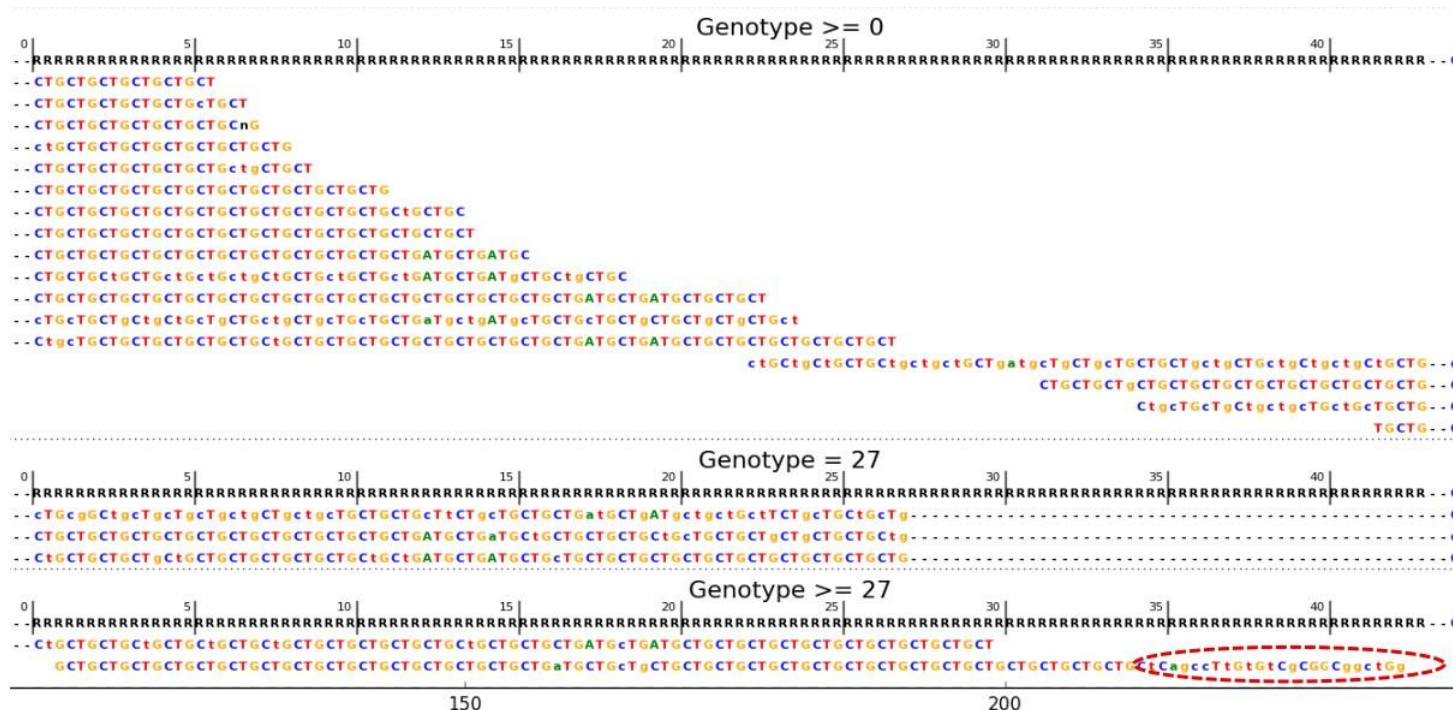
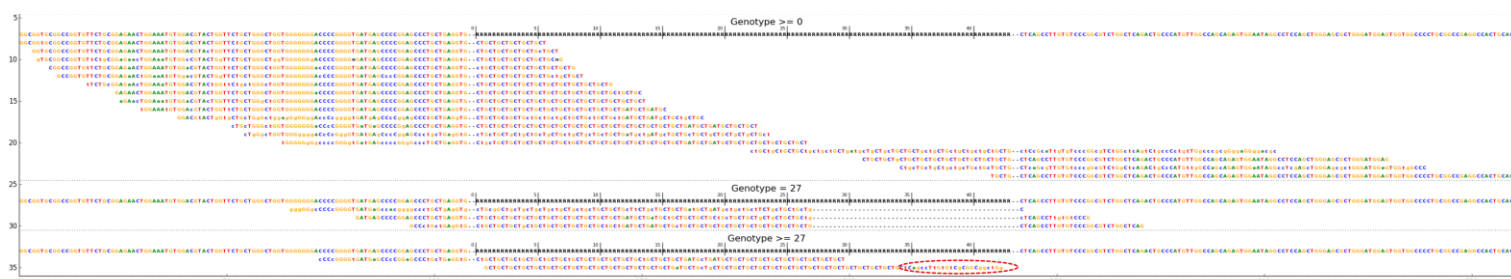
- Example of a good quality STR call with interruptions

Here is an example of the reads that Expansion Hunter uses when estimating expansions in *ATXN1*, assessing a 30 and 53 repeat-length motif for each allele respectively. Below, you can see interruptions (‘ATG’ rather than ‘CTG’) in the reads containing the *ATXN1*



- Example of a poor quality STR call

Here is an example of the reads that Expansion Hunter uses when estimating expansions on *ATXN1*, assessing 27 and 43 repeat-length motifs for each allele respectively. The reads supporting the 27 `CTG` allele are clean good quality reads. Conversely, visual inspection of the reads that Expansion Hunter used for estimating the 43 repeat-length allele, identifies bad quality bases at the end of the read. Bad quality bases within a read are shown in lowercase, while good quality bases are uppercase. In this case the lower case bases are not even following the `CTG` motif. We would suggest that in this case Expansion Hunter should have estimated an allele containing 30-35 repeat-size motif rather than 43.

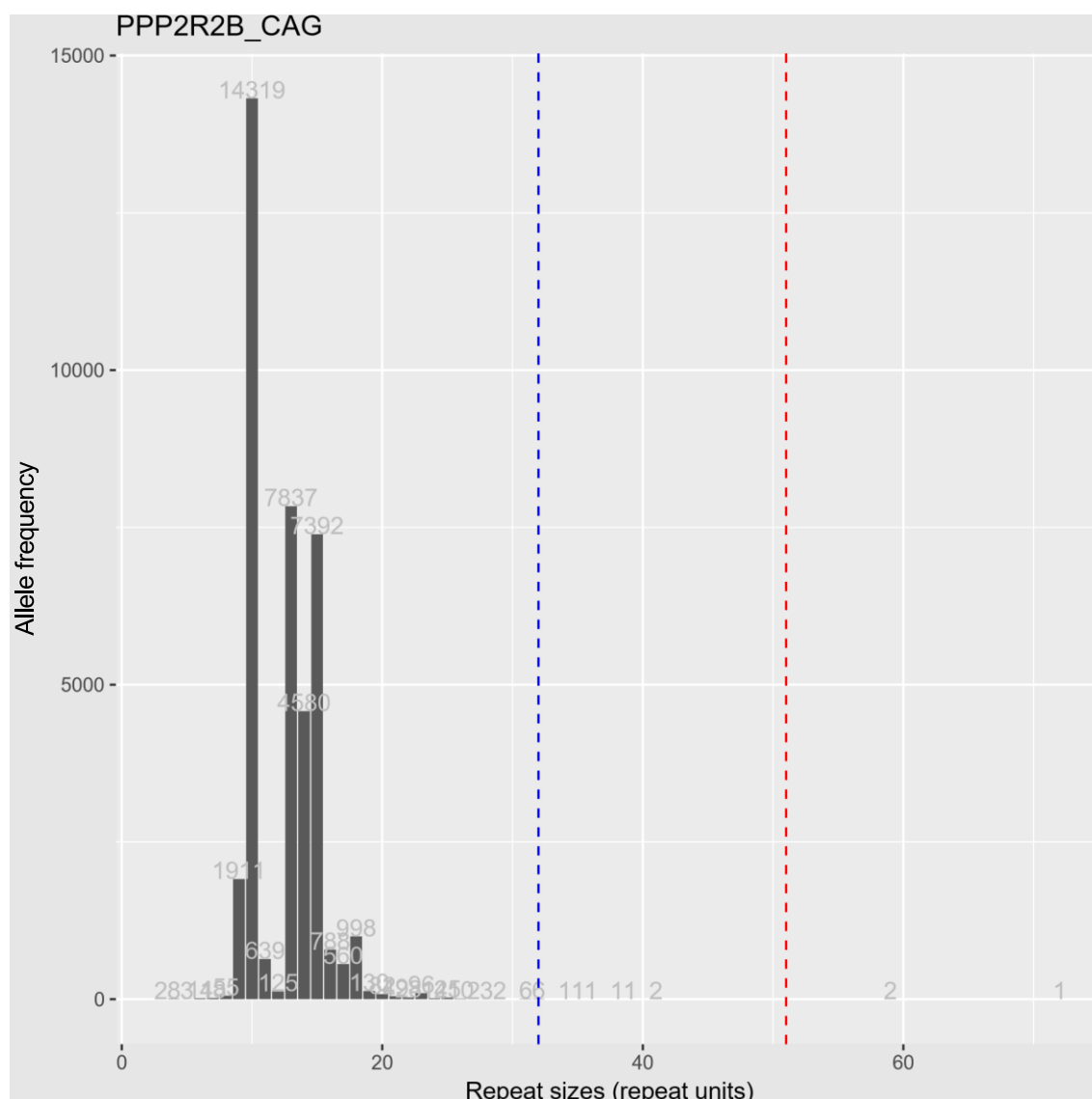


7.6.2 Internal allele frequencies

STR tiering does not take internal allele frequencies into account, but it is useful to have them in order to have some context regarding the frequencies for each locus.

For instance, the example below contains the internal STR allele frequencies for all repeat sizes detected for Spinocerebellar Ataxia 12 (SCA12, *PPP2R2B*) in 30,481 germline genomes in GRCh38. Blue and red vertical lines correspond to normal and pathogenic thresholds

respectively, 32 and 51 respectively for this case. (defined by Genomics England, https://panelapp.genomicsengland.co.uk/panels/20/str/PPP2R2B_CAG/#!details).



From these frequencies we are able to compute the percentiles for each repeat unit (and each locus) and this will be included in the interpretation portal in the future, to give some context for the reporting of the STR variant.

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

7.7.1 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 webservice 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 Platypus 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 and Tier 3 variants

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.

7.8 Uniparental Disomy

The Genomics England WGS pipeline can detect uniparental disomies (UPDs) in rare disease participants of the 100K Genomes Project.

Predicted uniparental disomies (UPDs) are flagged in the CIP-API (and Interpretation Portal – see Figure 8) 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

- dddddddd 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 Uniparental isodisomy segregation filter and that this is independent of flagging.

If you would like to receive approximate coordinates of the approximate regions of isodisomy and/or heterodisomy detected, please contact the Genomics England Service Desk.

7.9 Clinical Interpretation Partners (CIPs) and the CIP-API

7.9.1 Clinical Interpretation Partners (CIPs)

There are currently 2 Clinical Interpretation Partners (CIPs):

- Fabric Genomics
- Congenica

Your GMC will be issued with a CIP specific user guide during your training on their system, if you do not have a copy, or require an updated copy, please contact Genomics England service desk here: ge-servicedesk@genomicsengland.co.uk or via the portal www.bit.ly/geservicedesk



7.9.2 Overview

Following Interpretation, an 'Interpretation Request' is sent from Genomics England to the relevant CIP in JSON format containing all of the information the CIP needs to display and annotate the case (e.g. locations of BAM file(s), VCF file(s), Big Wig coverage file(s), and details of the tiered variants). After the CIP receives the Interpretation Request, the CIP's analyse the data using their annotation and interpretation pipelines (depending on service level applied) and display the associated data (variants, tiers, alignments etc).

Previously there were two service levels that CIPs can provide, silver and bronze. For the majority of the pilot cases and during the early main programme, Silver Service was applied, and therefore the CIP's Clinical Services team undertook a review of each case. For some of these cases the CIP's Clinical Services team may have annotated candidate variants they consider are clinically relevant to the participant's recorded phenotype. This includes reporting variants that the Tiering may not have highlighted due to the filtering thresholds used, or due to the gene not yet being included in a gene panel (e.g. a newly published disease association), or on a panel which was not applied to that participant but could have been considered phenotypically relevant. The approach the CIPs use to perform interpretation is CIP specific. However, they have provided Genomics England with their Standard Operating Procedures and it is expected that they clearly communicate the rationale for any decision that is made to highlight a candidate variant. All main programme cases are now processed on Bronze service, and therefore no such clinical review of case/variants are undertaken, although CIP providers may apply their automated prioritisation algorithms and allow access to these results through the CIP.

Regardless of service level applied to a case, once the CIP's have completed their analysis and ingestion of a case within their system, this is communicated to Genomics England in the form of an 'Interpreted Genome', which is another JSON formatted file.

7.10 Interpretation Request files (for beginners)

The following information can be found within the interpretation request JSON file:

- Family Pedigree and Other Family History

- Analysis Panels & versions
- Specific Disorder
- Tiered Variants and Tiering version
- HPO terms
- Workspace (NHS GMC or LDP site code)
- Gene Panel Coverage
- Disease Penetrance
- Variant Classification

Full descriptions of the terms used above can be found below in the 'Interpretation Request Guide for Bioinformaticians' section

An interpretation request JSON file looks like the example below:

```
"interpretation_request_id": XXX,
  "version": "1",
  "created_at": "2016-12-09T10:43:57.137142Z",
  "cip": "congenica",
  "family_id": "FM50000XZY",
  "sample_type": "raredisease",
  "interpretation_request_data": {
    "json_request": {
      "genomeAssemblyVersion": "GRCh37.p13",
      "virtualPanel": null,
      "pedigreeDiagram": null,
      "analysisVersion": "1",
      "TieringVersion": "0.3.7",
      "additionalInfo": null,
      "analysisReturnURI": "/gel/returns/SAP-XXX-1",
      "pedigree": {
        "gelFamilyId": "FM50000XZY",
        "participants": [
          {
            "personKaryotypicSex": "XX",
            "fatherId": 4,
            "dataModelCatalogueVersion": "v4.2",
            "twinGroup": null,
            "sex": "female",
            "superMotherId": null,
            "superFatherId": null,
            "affectionStatus": "affected",
            "consanguineousParents": "unknown",
```



```
"consentStatus": {
  "primaryFindingConsent": true,
  "carrierStatusConsent": false,
  "programmeConsent": true,
  "secondaryFindingConsent": false
},
```

Ancestries	
mothersEthnicOrigin	Mother's Ethnic Origin
mothersOtherRelevantAncestry	Mother's Ethnic Origin Description
fathersEthnicOrigin	Father's Ethnic Origin
fathersOtherRelevantAncestry	Father's Ethnic Origin Description
chiSquare1KGenomesPhase3Pop	Chi-square test for goodness of fit of this sample to 1000 Genomes Phase 3 populations
CalledGenotype	
gellid	Participant id of the family member
sampleId	LP (sequencing) Number of the family member
genotype	Zygosity
phaseSet	Phase set of variants when variants are phased
depthReference	Depth for Reference Allele
depthAlternate	Depth for Alternate Allele
copyNumber	Copy number genotype for imprecise event
ChiSquare1KGenomesPhase3Pop	
kGSuperPopCategory	1K Super Population
kGPopCategory	1K Population
chiSquare	Chi-square test for goodness of fit of this sample to this 1000 Genomes Phase 3 population

ConsentStatus	
programmeConsent	Is this individual consented to the programme? For example this could include a family member that is not consented but for whom affection status is known
primaryFindingConsent	Consent status for feedback of primary findings
secondaryFindingConsent	Consent status for secondary findings
carrierStatusConsent	Consent status for carrier status
Disorder	
DiseaseGroup	Is the Level 2 Title for this disorder (eg cardiovascular disease)
DiseaseSubGroup	is the Level 3 Title for this disorder (eg cardiomyopathies)
SpecificDisease	is the Level 4 Title for this disorder (eg hypertrophic cardiomyopathy)
Age	Age of onset in months

EthnicCategory This is the reported ethnicity according to the list in ONS16	
D	Mixed: White and Black Caribbean
E	Mixed: White and Black African
F	Mixed: White and Asian
G	Mixed: Any other mixed background
A	White: British
B	White: Irish
C	White: Any other White background
L	Asian or Asian British: Any other Asian background
M	Black or Black British: Caribbean
N	Black or Black British: African
H	Asian or Asian British: Indian
J	Asian or Asian British: Pakistani
K	Asian or Asian British: Bangladeshi
P	Black or Black British: Any other Black background
S	Other Ethnic Groups: Any other ethnic group

EthnicCategory This is the reported ethnicity according to the list in ONS16	
R	Other Ethnic Groups: Chinese
Z	Not stated
File	
SampleId	Unique ID(s) of the Sample, for example in a multisample vcf this would include an array of all the sample ids
URIFile	URI PATH
fileType	This defines a file. This Record is defined by the sampleID and a URI Currently SampleID can be a single String or an array of strings if multiple samples are associated with the same file
md5Sum	This defines a file. This Record is defined by the sampleID and a URI Currently SampleID can be a single String or an array of strings if multiple samples are associated with the same file
Genomic Feature	
FeatureType	Feature Type
ensemblId	Transcript used, this should be a feature ID from Ensembl, (i.e, ENST00000544455)
HGNC	This field is optional, BUT it should be filled when possible
other_ids	Others IDs
HPO term	
Term	Identifier of the HPO term
termPresence	This describes whether the HPO term is present in the participant (default is null=unknown) true=HPO term is present in participant, false=HPO term is not present
modifiers	Modifier associated with the HPO term
ageOfOnset	age of onset in months
OtherFamilyHistory	
MaternalFamilyHistory	Relevant Maternal family history
paternalFamilyHistory	Relevant Maternal family history
Pedigree	
VersionControl	Model version number gelFamilyId which internally translate to a sample set
participants	This is the concept of a family with associated phenotypes as present in the record RDPParticipant
analysisPanels	This is the concept of a family with associated phenotypes as present in the record RDPParticipant
diseasePenetrances	This is the concept of a family with associated phenotypes as present in the record RDPParticipant

RDParticipant	
SuperFatherId	this id is built using the original familyId and the original pedigreeId of the father
superMotherId	this id is built using the original familyId and the original pedigreeId of the mother
twinGroup	Each twin group is numbered, i.e. all members of a group of multiparous births receive the same number
monozygotic	A property of the twinning group but should be entered for all members
adoptedStatus	Adopted Status
lifeStatus	Life Status
consanguineousParents	The parents of this participant has a consanguineous relationship
consanguineousPopulation	Offspring from a consanguineous population
affectionStatus	Disease Status
disorderList	Clinical Data (disorders). If the family member is unaffected as per affectionStatus then this list is empty
hpoTermList	Clinical Data (HPO terms)
ancestries	Participant's ancestries, defined as Mother's/Father's Ethnic Origin and Chi-square test for goodness of fit of this sample to 1000 Genomes Phase 3 populations
consentStatus	What has this participant consented to? A participant that has been consented to the programme should also have sequence data associated with them; however this needs to be programmatically checked
samples	This is an array containing all the samples that belong to this individual, e.g ["LP00002255_GA4"]
inbreedingCoefficient	Inbreeding Coefficient Estimation
additionalInformation	We could add a map here to store additional information for example URIs to images, ECGs, etc Null by default

ReportEvent	
ReportEventId	Unique identifier for each report event, this has to be unique across the whole report, and it will be used by GEL to validate the report
phenotype	This is the phenotype (usually the HPO term or the disorder name) considered to report this variant
panelName	Gene Panel used from panelApp
panelVersion	Gene Panel Version
modeOfInheritance	Mode of inheritance used to analyze the family
genomicFeature	This is the genomicFeature of interest for this reported variant, please note that one variant can overlap more than one gene/transcript. If more than one gene/transcript is considered interesting for this particular variant, should be reported in two different ReportEvents
penetrance	This is the penetrance assumed for scoring or classifying this variant
score	This is the score provided by the company to reflect a variant's likelihood of explaining the phenotype using a specific mode of Inheritance
vendorSpecificScores	Other scores that the interpretation provider may add (for example phenotypically informed or family informed scores)
variantClassification	Classification of the pathogenicity of this variant with respect to the phenotype
fullyExplainsPhenotype	This variant using this mode of inheritance can fully explain the phenotype? true or false
groupOfVariants	This value groups variants that together could explain the phenotype according to the mode of inheritance used. All the variants sharing the same value will be considered in the same group. This value is an integer unique in the whole analysis.
EventJustification	This is the description of why this variant would be reported, for example that it affects the protein in this way and that this gene has been implicated in this disorder in these publications. Publications should be provided as PMIDs using the format [PMID:8075643]. Other sources can be used in the same manner, e.g. [OMIM:163500]. Brackets need to be included.

ReportEvent	
Tier	Tier is a property of the model of inheritance and therefore is subject to change depending on the inheritance assumptions This should be filled with the information provided by GEL

Reported Mode of Inheritance	
monoallelic_not_imprinted	MONOALLELIC, autosomal or pseudoautosomal, Not imprinted
monoallelic_maternally_imprinted	MONOALLELIC, autosomal or pseudoautosomal, Maternally imprinted (paternal allele expressed)
monoallelic_paternally_imprinted	MONOALLELIC, autosomal or pseudoautosomal, Paternally imprinted (maternal allele expressed)
monoallelic	MONOALLELIC, autosomal or pseudoautosomal, Imprinted status unknown
biallelic	BIALLELIC, autosomal or pseudoautosomal, Monoallelic_and_biallelic: BOTH monoallelic and biallelic, autosomal or pseudoautosomal,
monoallelic_and_more_severe_biallelic	BOTH monoallelic and biallelic, autosomal or pseudoautosomal (but BIALLELIC mutations cause a more SEVERE disease form), autosomal or pseudoautosomal
xlinked_biallelic	X-LINKED: hemizygous mutation in males, biallelic mutations in females
xlinked_monoallelic	X-LINKED: hemizygous mutation in males, monoallelic mutations in females may cause disease (may be less severe, later onset than males)
Mitochondrial	MITOCHONDRIAL
Unknown	Unknown

Reported Structural Variant	
Chromosome	Named as: 1-22,X,Y,MT (other contigs name)
Start	Start position 1- based
end	End position 1-based
type	The ID field indicates the type of structural variant, and can be a colon-separated list of types and subtypes (this is VCF Format): DEL, INS, DUP, INV, CNV, DUP:TANDEM, DEL:ME, INS:ME, INS:ME:ALU...
reference	Reference Allele sequence, the same provided in vcf
alternate	alternate
calledGenotypesreportEvents	This is the scores across multiple modes of inheritance, for each model of inheritance analyzed, the variants can have only one Report event.
AdditionalTextualVariantAnnotations	For example HGMD ID
evidencelds	For example HGMD ID, dbSNP ID or Pubmed Id
additionalNumericVariantAnnotations	For Example (Allele Frequency, sift, polyphen, mutationTaster, CADD. ..)
comments	Comments

Reported Variant	
Chromosome	Name as:1-22, C,Y,MT or other contig names as defined in the BAM header
dbSNPid	Variant ID in dbSNP
position	Position 1-based
reference	Reference allele sequence, the same provider vcf
alternate	Alternate allele sequence, the same provider in vcf
calledGenotypes	Array of genotypes for the family
reportEvents	This is the score across multiple mode of inheritance, for each model of inheritance analyzed, the variants can have only one reported event.
AdditionalTextualVariantAnnotations	For example a quote from a paper
Evidenceld	For example HGMD ID, dbSNP ID or Pubmed Id
AdditionalNumericVariantAnnotations	For example (allele frequency, SIFT, Polyphen2, MutationTaster, CADD)
Comments	Comments

Version Control	
GitVersionControl “3.0.0”	This is the version for the entire set of data models as referred to the Git release tag

Version Control	
Zygosity	
Reference_homozygous	0/0,0 0
heterozygous	0/1,1/0, 1 0, 0 1
Alternate_homozygous	1/1, 1 1
Missing	./., . .
Half_missing_reference	./0, 0/., 0 ., . 0
Half_missing_alternate	./1, 1/., 1 ., . 1
Alternate_hemizygous	1
Reference_hemizygous	0
unk	Anything unexpected

7.11 Interpretation Request guide for bioinformaticians

All options for the below fields are explicitly described in the 'Interpretation Request Field Options' section below. Note that not all fields are currently populated. What is shown below is on v3.0.0 of GeLReportModels which is the model being used for all current production Rare Disease cases in the GeL Rare Disease pipeline.

A more detailed description can be found in the model documentation here:

<https://genomicsengland.github.io/GeLReportModels/models.html#org-gel-models-reportavro>

The code repository for GeLReportModels is public. The master branch has the finalised models formatted using the avro schema convention. It is accessible here:

<https://github.com/genomicsengland/GeLReportModels/tree/master>

Examples from Genomics England Workshops using the models can be found here:

https://github.com/genomicsengland/ACGS_GeL_API_workshop

7.12 Interpretation Request Field Options

Interpretation Request Field	Options
AdoptedStatus	<ul style="list-style-type: none"> not_adopted adoptedin adoptedout
AffectionStatus	<ul style="list-style-type: none"> unaffected affected unknown

Interpretation Request Field	Options
AnalysisPanel*	<ul style="list-style-type: none"> • specificDisease • panelName • panelVersion • review_outcome • multiple_genetic_origins
ComplexGeneticPhenomena	<ul style="list-style-type: none"> • Mosaicism • Monosomy • Disomy • uniparental_disomy • Trisomy • Other_aneuploidy
DiseasePenetrance	<ul style="list-style-type: none"> • SpecificDisease • Penetrance
FeatureTypes	<ul style="list-style-type: none"> • RegulatoryRegion • Gene • Transcript
FileType	<ul style="list-style-type: none"> • BAM • gVCF • VCF_small • VCF_somatic_small • VCF_CNV • VCF_somatic_CNV • VCF_SV • VCF_somatic_SV • VCF_SV_CNV • SVG • ANN • BigWig • MD5Sum • ROH • OTHER

Interpretation Request Field	Options
KGPopCategory	<ul style="list-style-type: none"> • ACB • ASW • BEB • CDX • CEU • CHB • CHS • CLM • ESN • FIN • GBR • GIH • GWD • IBS • ITU • JPT • KHV • LWK • MSL • MXL • PEL • PJL • PUR • STU • TSI • YRI
KGSuperPopCategory	<ul style="list-style-type: none"> • AFR • AMR • EAS • EUR • SAS

Interpretation Request Field	Options
LifeStatus	<ul style="list-style-type: none"> • Alive • Aborted • Deceased • Unborn • Stillborn • miscarriage
Penetrance	<ul style="list-style-type: none"> • Complete • Incomplete
PersonKaryotypicSex	<ul style="list-style-type: none"> • Unknown • XX • XY • XO • XXY • XXX • XXYY • XXXY • XXXX • XYY • other
Sex	<ul style="list-style-type: none"> • Male • Female • Unknown • Undetermined
TernaryOption	<ul style="list-style-type: none"> • yes • no • unknown
Tier	<ul style="list-style-type: none"> • NONE • TIER1 • TIER2 • TIER3

Interpretation Request Field	Options
VariantClassification	<ul style="list-style-type: none"> • BENIGN • LIKELY_BENIGN • VUS • LIKELY_PATHOGENIC • PATHOGENIC

* **NOTE:** the specificDisease(s) defined in the Analysis Panel field are the disease(s) for which the family defined in the pedigree were recruited under. There maybe one or more analysis panels associated with that specific disease and these are defined using the Panel App HASH in the panelName field. Details of how to use the panel HASH to get the associated human readable panel name can be found on the PanelApp website here: <https://panelapp.genomicsengland.co.uk/#!Webservices>

7.13 Interpreted Genome files

The same fields as described above in the interpretation requests “possible field options” also apply for interpreted genome format. The CIPs send the Interpreted Genome back to Genomics England for each interpretation request once they have ingested the case and if on silver level service, reviewed the tiered variants and generated any additional candidate variants. The CIP specific processes are described in more detail within the CIP specific user guides. You will be issued with the CIP user guide during your training on the CIP system, if you do not have a copy, or require an updated copy, please contact Genomics England service desk here: ge-servicedesk@genomicsengland.co.uk or via the portal www.bit.ly/geservicedesk

7.13.1 Exomiser Interpreted Genome Scores

Please see section 7.4.2 regarding how Exomiser scores and ranks reported variants and their respective report events. The two fields that are the focus of the Exomiser results are the rank of the report event, and the score of the report event, as shown below (Figure 6).

The CIP-API can be used to further explore the Exomiser results for a case which are displayed as in interpreted_genome with “companyName”: “Exomiser”. HGVS annotations, consequences and genotypes are returned for each of the variants in the reportedVariants list as well as the reportEvents detailing modeOfInheritance, grouped compound-heterozygote reportEvents, gene and detailed phenotype matches in the eventJustification and the genePhenoScore and variantScore that contribute to the overall, combined score.

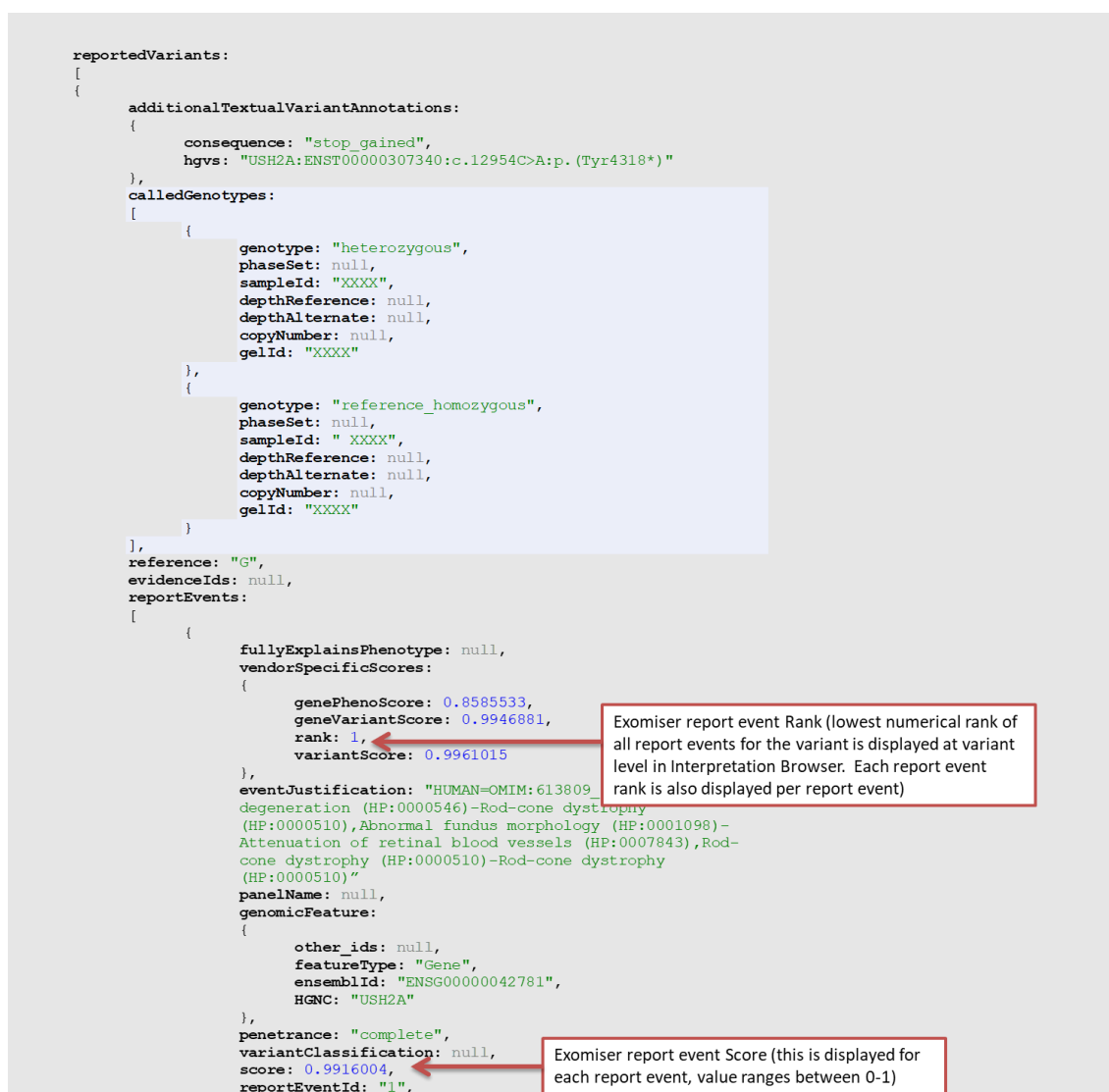


Figure 7: EXOMISER INTERPRETED GENOME: REPORTED VARIANT EXAMPLE

7.14 CIP-API

The CIP-API can be split into four purposes.

1. It communicates with the NHS GMC user and the CIPs regarding which cases are ready for interpretation. It does this by creating an InterpretationRequest.json which is sent to Interpretation Services (e.g. Exomiser) and CIPs and is visible from the CIP-API web services.
2. When an Interpretation Service or CIP generates an InterpretedGenome.json, it pushes this information back to the CIP-API. This json is appended to the InterpretationRequest.json and can be accessed through the CIP-API web services.
3. If an NHS GMC user selects a variant as a Primary Finding in the Interpretation Portal or CIP user interface and decides to produce a clinical report, the Portal and / or CIP pushes this information as a ClinicalReport.json to the CIP-API. The ClinicalReport.json is appended to the InterpretationRequest.json.
4. Via the Interpretation Portal, the CIP-API displays the case status and the ClinicalReport.json as an HTML page visible to the NHS GMC 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/>

7.15 Quality Assurance Processes

The following section describes the Quality Assurance (QA) steps undertaken by Genomics England to ensure the result returned to NHS GMCs via the Interpretation Platform is of high quality before being presented to NHS GMCs. QA is used by Genomics England to ensure mistakes and errors in results produced by the automated pipeline(s) are limited and removed before results are returned to GMCs. QA at Genomics England is performed by the Clinical Bioinformatics Team.

7.15.1 Genomics England Quality Assurance processes

Genomics England do not perform in-depth analysis of each variant, nor assess the pathogenicity of variants. Genomics England carry out genetic checks to ensure clinical and genomic data consistency (sex, relatedness and ethnicity) before any data is sent to the CIPs.

QA can be split into three linked processes:

- User Acceptance Testing (UAT): The process of checking the systems for returning Genomics England results to the NHS GMCs meet our standards.
- Quality Control (QC) spot checks: The process where we regularly review results being returned to the NHS GMCs to ensure their quality
- Verification of internal pipeline updates

7.15.1 Release of data to NHS GMCs

Once QA checks are complete, the cases are made available to the NHS GMC in the Interpretation Portal. The responsibility for individual QC of tiered variants lies with NHS GMC clinical scientists, and it is expected that the NHS GMC clinical scientists will check the read level support for a variant before undertaking any orthogonal (technical) validation.

7.16 Interpretation Portal

The Interpretation Portal enables NHS GMC users to:

- a. See an overview of cases ready for NHS GMC review, and track overall case status.
- b. Review findings from Interpretation Services such as Tiering and Exomiser
- c. Hyperlink to cases in the respective CIP decision support system.
- d. Hyperlink to an overview of summary of the submitted clinical data in LabKey.
- e. Download any available files.
- f. Complete a reporting outcomes questionnaire to close a case.
- g. Save work in progress as draft and return to it later e.g. when completing the outcomes questionnaire

7.16.1 How to use the Interpretation Portal

1. You can access the Interpretation Portal here:
<https://cipapi.genomicsengland.nhs.uk/interpretationportal/ReviewCases.html> on HSCN (previously the N3 network)
2. Use your LDAP credentials (the username and password issued by Genomics England Service desk) to log into the Interpretation Portal.

If you don't have LDAP login details please contact the Genomics England service desk (<http://bit.ly/ge-servicedesk>).

3. Upon logging into the Interpretation Portal with LDAP credentials the user will see a list of their cases which are "To Be Reviewed" (Figure 7).

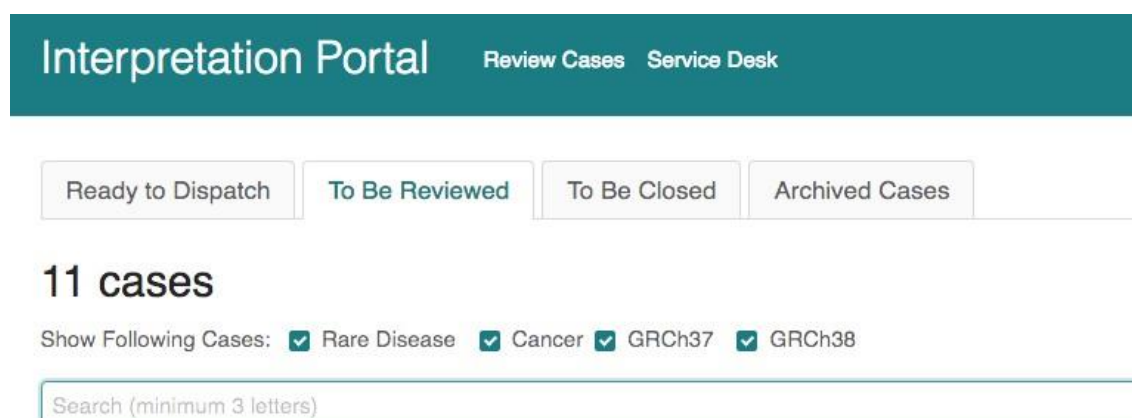


Figure 8: INTERPRETATION PORTAL TABS, FILTERING AND SEARCH

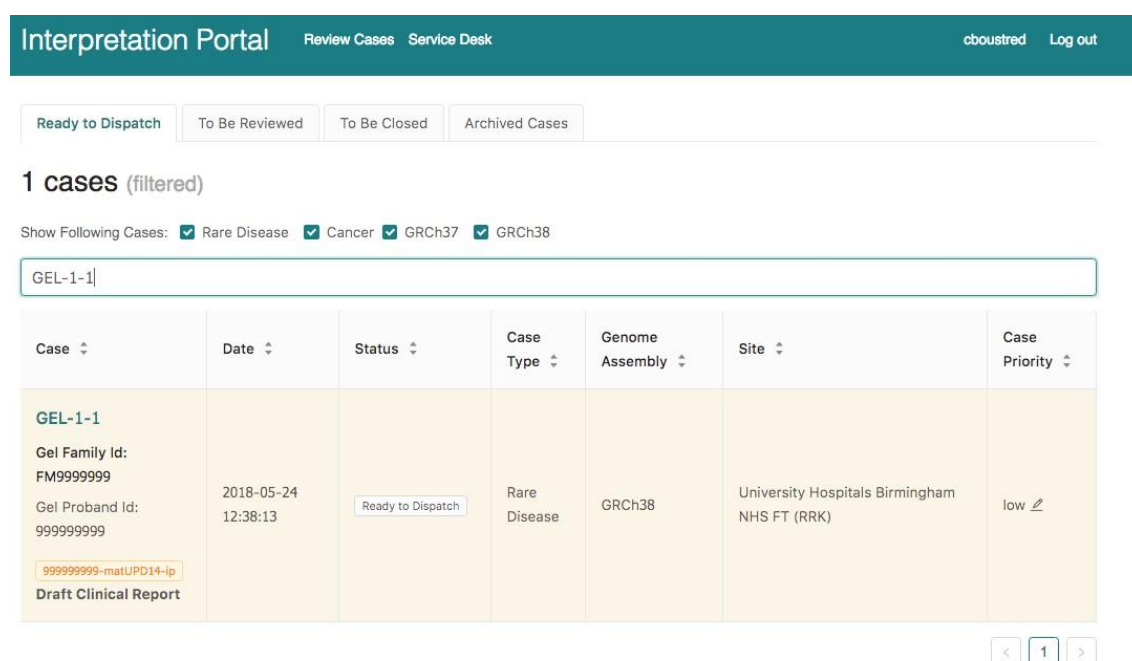


Figure 9: CASE FILTER PAGE WITH EXAMPLE UPD FLAG Cases

will be in one of the four tabs:

1. "Ready to Dispatch" – these are cases that have completed processing (see 7.3) through the Rare Disease Interpretation Pipeline. Interpretation results will be available for review, however, the cases will not yet be accessible in the CIP systems.
2. 'To be Reviewed' – These are cases that are available for NHS GMC users to review and link out to in the CIP decision support system.
3. 'To be Closed' – Cases which have been reviewed and a Summary of Findings HTML generated. The Summary of Findings HTML can be downloaded for any cases in this tab and the Reporting Outcomes questionnaire can be completed.
4. 'Archived Cases' – Cases for which the Reporting Outcomes Questionnaire has been completed.

Note: Cases may be flagged with tags e.g. if a family member is predicted to have Uniparental Disomy (UPD) the case could be tagged with UPD. See 7.8 Uniparental Disomy and Figure 8.

7.16.2 Reviewing a case

NHS GMC's will receive an email each Monday morning alerting them to any new cases that are ready for review from the previous week. The results email alert will be sent to the generic email address for the GMC held by Genomics England. The email will contain a list of Interpretation Request IDs and family IDs. If you wish to update or change the results email address please contact the Genomics England Service desk:

(ge-servicedesk@genomicsengland.co.uk or <http://bit.ly/ge-servicedesk>)

Cases listed in the "Ready to Dispatch" tab enable as early as possible access to rare disease results. The results of tiering and other interpretation services applied by the Genomics England Interpretation Pipeline for a case can be accessed (details below) from this tab. If appropriate a Summary of Findings HTML can be generated based on the Interpretation Service results for cases via the GeL Interpretation Browser. Cases are available in the CIP systems when the case is found in the "Ready to be Reviewed" tab.

1. In the "To Be Reviewed" tab (Figure 7) you will see a list of cases that can be reviewed.

Cases will appear in this list when they pass quality assurance by the Genomics England clinical bioinformatics team and have been dispatched to a CIP and are now visible in the CIP to NHS GMC users for.

2. You can filter the list by sample type and genome build and search for cases using Interpretation Request ID, Family ID, Proband ID, Release Date and Site in the search box (Figure 7).

3. Click on the Interpretation Request ID (e.g. SAP-123-1 or OPA-123-1) for the case you wish to review and you will see summary information about the participant and family members (Figure 9).

4. To review variants highlighted by interpretation services (e.g. Tiering) applied to the case click the "Review Variants" button. This will link to the Interpretation Browser (5.14).

5. To review a case in the CIP decision support system click the "Review Case" button in the top left corner (Figure 9).

6. This will link out to the relevant CIP decision support system and allow you to review the case. Please see the individual company user guides for specific instructions on reviewing a case .

7. Any CSV files generated in the CIP decision support system (e.g. for Sanger confirmations) can be downloaded from the 'Associated files' section at the bottom of the page, by clicking on the download arrow.

8. Once a case is reviewed and the status is changed to 'Approved' in the CIP decision support system, the summary of findings is automatically exported to the CIP-API via Opal or Sapientia. The case will then move to the 'To be Closed' list in the Interpretation Portal.

Please note: in some of the company systems, a report can be generated before the report has been finalised and approved. The case will move to the 'To be Closed' tab when a report has been generated, even if the final approval is still pending in the Decision Support System.

9. To return to the list of cases, click on the 'Review Cases' icon on the green bar at the top of the page

The screenshot shows the 'Interpretation Portal' interface. At the top, there are two red boxes with arrows: 'Return to case overview page' pointing to a link, and 'Link to Service Desk' pointing to a 'Service Desk' button. The main header includes 'Interpretation Portal', 'Review Cases', 'Service Desk', 'cbousted', and 'Log out'.

The main content area is titled 'GEL-1 - Interpretation Browser'. It displays case details: 'Bronze service. Status: ready_to_dispatch Priority: low', 'Proband: [redacted], Year of Birth: [redacted], Sex: male', 'Family: [redacted]', 'Number of Unique Variants: 325', and 'Genome Assembly Version: GRCh38'.

Below this is a table of 'Disorder' and 'Affiliated Panels':

Disorder	Affiliated Panels
Dilated Cardiomyopathy	Dilated Cardiomyopathy and conduction defects (1.33)
Hypertrophic Cardiomyopathy	Familial Genetic Generalised Epilepsies (1.23) Hypertrophic Cardiomyopathy (1.20) Mitochondrial disorders (1.56) RASopathies (1.18) Undiagnosed metabolic disorders (1.72)

Below the table are 'HPO terms' with several tags: 'Hypertrophic cardiomyopathy', 'Concentric hypertrophic cardiomyopathy', 'T-wave inversion', 'Dilated cardiomyopathy', 'Reduced systolic function', 'Generalized tonic seizures', 'Obesity', 'Wolff-Parkinson-White syndrome', and 'Left ventricular noncompaction cardiomyopathy'.

At the bottom is a 'Pedigree Summary' table with columns: 'FamilyID', 'Gel Participant ID', 'Year of Birth', 'Gender', 'Relationship to proband', 'Disorder', and 'Sample ID'. It lists three family members: a mother with Hypertrophic Cardiomyopathy, a maternal aunt with Dilated Cardiomyopathy, and the proband (male) with Hypertrophic Cardiomyopathy.

Annotations on the right side of the screenshot include: 'Disorders and affiliated panels' pointing to the table above, and 'Pedigree Summary Includes disorder associated with each family member' pointing to the pedigree table.

Figure 10 : CASE SUMMARY PAGE

7.16.1 Interpretation Flags

If during processing a case triggers a pre-defined QC check they are automatically flagged.

Table 1 lists the flags, a brief description and where they can be seen in the Interpretation Portal. If you have any concerns or questions relating to case flags please email the geservicedesk@genomicsengland.co.uk and quote the case identifier (e.g. family ID or interpretation request ID) and the flag.

Flag	Description
suspected_poor_quality_CNV_calls	If a sample has an increased number of all types of CNV calls (i.e. both Gain and Loss calls) this flag will be applied. See: Additional notes for more information
suspected_increased_number_of_false_positive_heterozygous_LOSS_calls	If a sample has a suspected increase of false positive heterozygous LOSS calls, but the quality of the other types of calls (Gain and homozygous Loss calls) are not affected this flag will be used. See: Additional notes for more information
CNV_calls_assumed_XX_karyo	When performing CNV calls on the sex chromosomes we check whether the sex karyotype used by the CNV calling pipeline matches the sex karyotype inferred using a more robust method. If a discrepancy is observed cases will be flagged. See: Additional notes for more information
CNV_calls_assumed_XY_karyo	
UPD Flags	See section: Uniparental Disomy

Flag	Description
Sex karyotype flag	At least one family member has a minor sex karyotype and/or there is a discrepancy between the reported and inferred sex. Details pertaining to a specific family can be obtained by raising a Jira Service Desk ticket. See below for further details.
participantID_suboptimal_coverage	At least one family member has lower than expected sequence coverage. The expectation is $\geq 95\%$ of callable autosomal bases are covered by ≥ 15 reads with mapping quality ≥ 11 . Coverage is 90-95% of callable autosomal bases covered by ≥ 15 reads with mapping quality ≥ 11
participantID_low_coverage (If released – decision pending)	At least one family member has lower than expected sequence coverage. The expectation is $\geq 95\%$ of callable autosomal bases covered by ≥ 15 reads with mapping quality ≥ 11 . Coverage is $< 90\%$ of callable autosomal bases covered by ≥ 15 reads with mapping quality of ≥ 11
Other Flags	In some circumstances Genomics England will add additional tags to cases to alert users to issues or features of the case e.g. if a case has been re-issued

7.16.2 Sex karyotype flag

The Genomics England Interpretation Pipeline utilises the expected sex in the logic for tiering of variants on chromosome X, that is, females are considered to have two copies of chrX, and males to have one copy of chrX. In the current pipeline, the expected sex is specified by the reported phenotypic sex, not the reported or inferred karyotypic sex. As such, tiering results for chrX may be suboptimal if there is a difference between the assumed and actual number of copies of chromosome X, as occurs for some families where participants have minor or discordant sex karyotypes.

Most common scenarios that may result in suboptimal tiering for chromosome X:

- A minor sex karyotype is inferred from the genomic data, defined as any karyotype other than XX or XY where there is no discrepancy between the inferred and reported sex (based on the reported karyotype, or the reported phenotypic sex if a karyotype is not provided).
- There is discordance between the reported and inferred sex (with or without a minor sex karyotype) which is expected, as evidenced either by the "Disorders of sexual development" panel being applied or by confirmation by GMC staff that this is known and interpretation should proceed. The reported sex is based on the karyotype, or the reported phenotypic sex if a karyotype is not provided.
- There is a discrepancy between the reported karyotypic and phenotypic sex, where the reported karyotypic sex is in agreement with the karyotype inferred from the genomic data, but a disorder of sexual development is not expected ("Disorders of sexual development" panel is not applied). Results will be flagged in the portal where these errors have not been corrected after contacting NHS GMC staff.

Scenario that is unlikely to affect tiering:

- There is a discrepancy between the reported and inferred karyotypic sex but the reported phenotypic sex is in agreement with the inferred karyotype. Results will be flagged in the portal where these errors have not been corrected after contacting NHS GMC staff.

7.16.3 Closing a Case

1. When a case has been reviewed in the Interpretation Browser or CIP decision support system and a Summary of Findings generated (Please see the individual company user guides for specific instructions) the case will move into the “To Be Closed” tab (Figure 7).
2. From list of cases in the “To Be Closed” tab select the individual case you would like to close.
3. On the case summary page, under “Summary of Findings Table” heading you will see a list of Summary of Findings that have been produced for a case (Figure 9). Please note that it is possible to generate multiple versions of the Summary of Findings.

7.16.4 Generating a Summary of Findings

If you click on the download button on the right hand side you will be able to download and save an HTML copy of the Summary of Findings (Figure 9).




Please note: If any errors are encountered generating the report, a message at the top of the report will appear alerting the user (Figure 10). If this error appears while generating a report, the NHS GMC user should report this to the Genomic England service desk

(ge-servicedesk@genomicsengland.co.uk or via the portal <http://bit.ly/ge-servicedesk>) , with the appropriate Case ID and/or Family ID and a description of what happened, and a screenshot if possible.



Figure 11: HEADER FROM A PRIMARY FINDINGS REPORT ILLUSTRATING AN EXAMPLE OF AN ERROR

7.16.5 Reporting Outcomes Questionnaire

Button	Purpose
	Download Summary of Findings in HTML format
	Complete Reporting Outcomes Questionnaire – close a case
	Edit answers for a previously answered Outcomes Questionnaire

1. To complete the reporting outcomes questionnaire and close the case the user should click on the “+” symbol (Figure 9 and above).
2. The reporting outcomes questionnaire will then be available to populate from a list of drop down menus (Figure 11).

Figure 12: DROP DOWN MENU

3. The variants displayed in the questionnaire are those contained within the associated version of the Summary of Findings. If several versions of a Summary of Findings are generated, it is possible to complete an outcomes questionnaire for each one. It is only necessary to complete one questionnaire per case, usually this would be against the most recent version of the Summary of Findings: if an exit questionnaire is completed for a case but then another summary of findings is generated, the case will move from closed (“Archived”) to “To be Closed”.
4. The exit questionnaire is divided into family level and variant level questions. For ‘negative’ reports containing no variants, the questionnaire will only present the family level questions.

7.16.6 Family level questions

Family level questions
Have the results reported here explained the genetic basis of the family’s presenting phenotype(s)?
Have you done any segregation testing in non-participating family members?

5. The first question asks whether the combined variants between them explain the genetic basis of the family’s presenting phenotype(s). This is asking whether the case can be considered fully or partially solved.
6. The second question asks whether you have done any segregation testing in the family. If you have, please enter details in the free text Comments box below. Please do NOT include any identifying details in this box. For example, ‘Proband’s maternal uncle also affected with ataxia and carries the variant in gene X’ is acceptable; ‘Also tested Robert Smith for the variant in gene X’ is not acceptable.

7.16.7 Variant and variant pair level questions

Variant level questions

Did you carry out confirmation of this variant via an alternative test?

Did the test confirm that the variant is present?

Did you include the variant in your report to the clinician?

What ACMG pathogenicity score (1-5) did you assign to this variant⁴?

Please provide PMIDs for papers which you have used to inform your assessment for this variant

Variant/variant pair level questions

Is evidence for this variant/variant pair sufficient to use it for clinical purposes such as prenatal diagnosis or predictive testing?

Has the clinical team identified any changes to clinical care which could potentially arise as a result of this variant/variant pair?

Did you report this variant/variant pair as being partially or completely causative of the family's presenting phenotype(s)?

7. Five questions are asked about each individual reported variant. It is not mandatory to enter details of publications used to support the reporting outcome, but if you have used any in the clinical report please enter them here.
8. Four questions are asked at the level of a variant or pair of compound heterozygous variants. The first question asks whether you have reported the variant or variant pair as having sufficient supporting evidence to be used for clinical purposes such as prenatal diagnosis or cascade/predictive testing.
9. The second question asks about any potential clinical outcomes which may have been identified during discussion with the clinical team, e.g. at the results MDT meeting. If this information is not available please record it as unknown. Actual clinical outcomes will not be available to collect at the time this questionnaire is completed and do not need to be reported here.
10. The third question asks whether the variant/variant pair is partially or completely causative of the family's presenting phenotype. This is asking whether this variant/variant pair explains all or some of the proband/family's phenotypic features, for example variant A explains the hearing impairment while variant B explains the intellectual disability.
11. The HPO terms entered for the proband are displayed in the final section. If the variant(s) is/are partially responsible for the phenotype, please delete the HPO terms which are NOT explained by those variant(s), leaving only the HPO terms which ARE considered likely to be explained by those variant(s). [Please note: it is not currently possible to add HPO terms to this section; this will be included in a future release.] If this is not clear cut there is no need to change the HPO term display here; it is anticipated that this will only be required in a small proportion of reports.
12. All sections of the form are mandatory apart from the PubMed IDs question. All sections have an 'NA' (not applicable) or 'unknown' option for situations where the question is not relevant.
13. Once completed please click on the 'submit' button at the bottom of the page to close the case and move it into the Archived Cases tab.

14. If you need to revise your report, please open it again as described above, complete it again, and resubmit. This version will over-ride the previous version as only one questionnaire can be stored against each report.

Please note: an exit questionnaire can be saved as a draft and returned to be completed at another time.

7.16.8 Linking participant to clinical data in LabKey

1. For all Genomics England Main Programme participants you can link to the participant's identifiers and clinical summary in LabKey (Figure 9).
2. Select a case from any of the four tabs; if it is a main programme case the Genomics England participant ID will be highlighted green.
3. Click on the Genomics England participant ID to be linked out to LabKey and view clinical and patient identifiable data.

Please note: a link to identifiers is not enabled for participant of the Genomics England Pilot project. Identifiers can be obtained from the spread sheet sent to your site's designated nhs.net account, which will continue to be provided for all pilot cases.

7.16.9 Archived cases

1. Archived cases (those for which the reporting outcomes questionnaire has been completed) can be viewed from the archived cases tab.
2. In the Summary of Findings section there will be a date showing when the case was previously reviewed
3. If you again click on the "paper" icon next to the download button, you are able to review previous answers or re-enter information.
4. To view previous answers, you can click on the Previous Answers link at the top of the page. This information will be provided in a more visually intuitive format in a future release.
5. If any changes are needed to the reporting outcomes questionnaire, it is possible to complete this again by scrolling down the page and completing and submitting as previously. Changes will over-write the previous form, so only the latest version of the form will be kept in the database.
6. As case numbers have increased, an archiving system has been introduced such that 60 days after the final submission of a reporting outcomes questionnaire, the reporting episode will be considered closed, and the latest questionnaire contents will be stored against that case and will not accept any further changes.
7. If at this stage any change to the Summary of Findings contents is required, a new report will need to be generated from the CIP decision support system or Interpretation Browser, and a new reporting outcomes questionnaire should be completed at the close of the new reporting episode.
8. If at this stage any change to the Summary of Findings contents is required, a new report will need to be generated from the CIP decision support system or Interpretation Browser, and a new reporting outcomes questionnaire should be completed at the close of the new reporting episode.

7.17 Interpretation Browser

The Interpretation Browser enables the NHS GMC clinical scientists to review results of Interpretation Services (e.g. Tiering and Exomiser) that have been applied to rare disease cases.

Note: Genomics England strongly advises the NHS GMCs to initial review the case details in the Interpretation Browser as certain new features of the pipeline (eg Multiple Monogenic Disorder Tiering) might not be displayed within the CIP systems.

Results of Interpretation Services can be accessed via the “Review Variants” button on the case overview page (Figure 9)

7.17.1 Overview

The Interpretation Browser enables NHS GMC users to:

- Review variants highlighted by Genomics England Interpretation Services and carry out some basic filtering
- Review results of CNV and STR calling (see sections 7.5 and 7.6)
- Review the zygosity of each variant in all sequenced family members. A filled black circle indicates presence of the variant, with a single black circle indicating heterozygosity and two black circles indicating homozygosity for a variant. A "-" indicates the genotype is undetermined.
- Search for variants in non-panel genes
- Save comments and highlighted variants as draft
- Download selected tiered variants in a TSV file
- Download all variants in VCF format
- Review gene panel coverage
- Review read level support for variants using IGV.js.
- Generate a Summary of Findings.

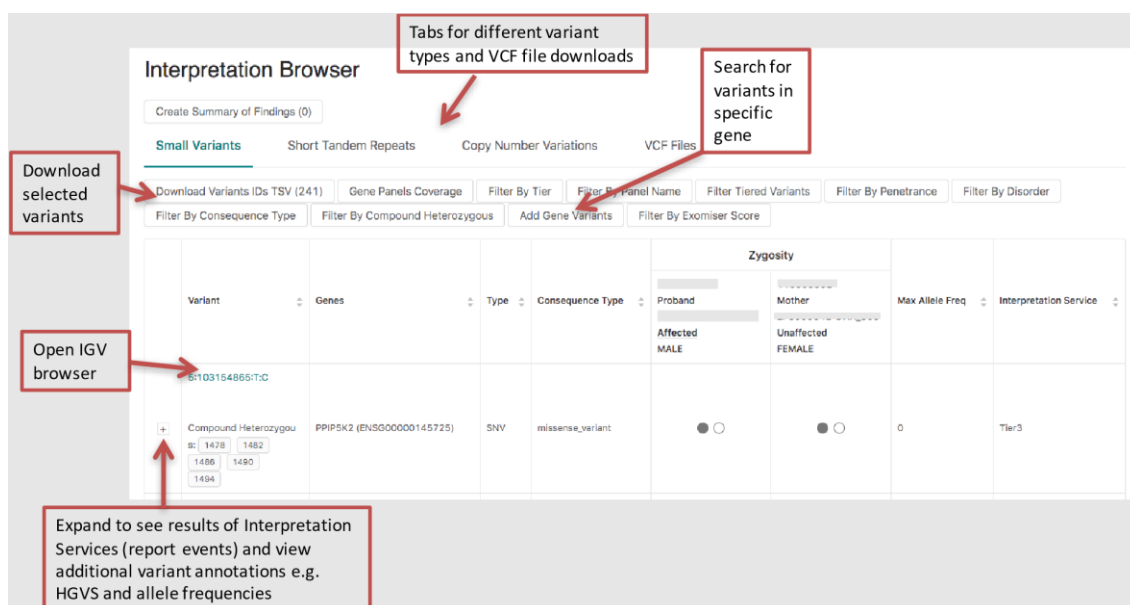


Figure 13: INTERPRETATION BROWSER

7.17.2 Multiple Monogenic Disorders

The rare disease SNV and Indel (small variant) Tiering pipeline can Tier variants in families where there are multiple, non-cosegregating monogenic disorders.

For example, in the case presented in Figure 9 and in Figure 13 there are two disorders in the family, Dilated Cardiomyopathy (DCM) and Hypertrophic cardiomyopathy (HCM). HCM is present in both the Mother and Proband, whereas the DCM is present in only the Aunt. Each disorder is associated with one or more gene panels and these panels are used in the Tiering process. In the Interpretation Browser it is possible to see the disorders used in the pedigree table, the top of the page and by hovering over the word “affected” in the header of the “zygosity” column.

PLEASE NOTE: The current CIP systems are not able to correctly display tiering results of cases with multiple monogenic conditions. The affection status of cases with multiple monogenic conditions should be determined from the information displayed in the Interpretation Portal. If you have any concerns please contact geservicedesk@genomicsengland.co.uk quoting your case identifier.

Multiple monogenic disorders

Disorder	Affiliated Panels
Dilated Cardiomyopathy	Dilated Cardiomyopathy and conduction defects (1.33)
Hypertrophic Cardiomyopathy	Familial Genetic Generalised Epilepsies (1.23) Hypertrophic Cardiomyopathy (1.20) Mitochondrial disorders (1.66) RASopathies (1.18) Undiagnosed metabolic disorders (1.72)

HPO terms: Hypertrophic cardiomyopathy | Concentric hypertrophic cardiomyopathy | T-wave inversion | Dilated cardiomyopathy | Reduced systolic function | Generalized tonic seizures | Obesity | Wolff-Parkinson-White syndrome | Left ventricular noncompaction cardiomyopathy

Multiple monogenic disorders

FamilyID	Gel Participant ID	Year of Birth	Gender	Relationship to proband	Disorder	Sample ID
			female	Mother	Hypertrophic Cardiomyopathy	
			female	MaternalAunt	Dilated Cardiomyopathy	
			male	Proband	Hypertrophic Cardiomyopathy	

Interpretation Browser

Download Variants IDs TSV (330) | Gene Panels Coverage | Create Summary of Findings (0) | Filter By Tier x | Filter By Panel Name | Search

Filter By Penetrance | Filter By Disorder | Filter By Consequence Type | Filter De Novo Variants | Filter Simple Recessive Variants

Filter By Compound Heterozygous | Add Gene Variants | Filter By Exomiser Score

Variant	Genes	Type	Consequence Type	Zygosity		
				Proband	Mother	MaternalAunt
1:11997360:A:C rs756851126	MFN2 (ENSG00000116688)	SNV	missense_variant	affected male	affected female	affected female
10:121						

Hover over "affected" shows the disorder(s)

Figure 14: MULTIPLE MONOGENIC DISORDERS IN THE INTERPRETATION VIEWER

7.17.3 Exomiser Display of Results

The Exomiser results displayed in the Interpretation browser show all rare, coding candidate variants compatible with the patient's pedigree under autosomal or X-linked dominant or recessive modes of inheritance as well as mitochondrial. The variants are ranked according to the Exomiser score (scaled 0 to 1) which is a measure of how rare and pathogenic the variant is predicted to be, as well as how closely the patient's phenotypes match the known phenotypes of diseases and model organisms associated with the gene (Figure 14)

Report Events	Score	HGNC	Ensemblid	Mode of Inheritance	Panel Name	Panel Version	Penetrance	Phenotype	Interpreted Genome Service
<input type="checkbox"/> TIER1		USH2A	ENSG00000042781	biallelic	Posterior segment abnormalities	1.8	complete	Rod-cone dystrophy	Tiering
<input type="checkbox"/>	Rank 1 Score 0.992	USH2A	ENSG00000042781	biallelic			complete	Progressive visual loss, Peripheral retinal degeneration, Abnormality of the retinal vasculature, Rod-cone dystrophy, Visual impairment, Abnormal fundus morphology, Retinal degeneration, Nyctalopia, Bone spicule pigmentation of the retina	Exomiser
<input type="checkbox"/>	Rank 39 Score 0.580	USH2A	ENSG00000042781	monoallelic			complete	Progressive visual loss, Peripheral retinal degeneration, Abnormality of the retinal vasculature, Rod-cone dystrophy, Visual impairment, Abnormal fundus morphology, Retinal	Exomiser

Figure 15: INTERPRETATION BROWSER DISPLAY OF EXOMISER RESULTS

For recessive modes of inheritance, compound heterozygotes with the same rank and score based on an average of the two variants may be highlighted. Such variants may also be identified as contributing under a dominant model as well, in which case two separate report events will be displayed in the browser with different scores and ranks (for the purposes of variant sorting in the browser, the highest score/lowest rank is always used though).

7.17.4 Display of CNV and STR results

For details of the CNV and STR tiering process see sections 7.5 and 7.6. Results of CNV and STR tiering can be accessed via the Interpretation Browser (Figure 12).

NOTE: CNV and STR tiering was added to the tiering application in version 1.0.14.

Cases processed prior to the release 1.0.14 will state:

“Tiering Version v1.0.0: does not include STR analysis therefore no STR results shown”

CNV and STR tiering results will initially only be added to cases prospectively. On request retrospective cases will be re-processed using tiering version 1.0.14 however this will be reviewed on a case by case basis.

STR and CNV report events can be selected and commented on in the same way as SNV / Indel variant results. They can be included in the Summary of Findings HTML that can be downloaded from the portal (see Generating Summary of Findings using the Interpretation Browser).

NOTE: Support CNV and STR tiering results will be implemented in Sapientia during 2019.

CNV results

CNV results are displayed in two tables, the “CNV Call Level Table” and the “CNV Gene Level Table”. The Call Level Table lists all the CNVs identified in the proband and allows users to select and comment on CNV calls they may want to include in Summary of Findings. The Gene Level Table lists all of the genes that overlap CNV calls in the proband.

Note: Currently it is not possible to filter CNV tables. Filters are planned for future releases of the browser.

STR results

The screenshot shows the 'Interpretation Browser' interface with the 'Short Tandem Repeats' tab selected. Annotations highlight key features:

- Create Summary of Findings with selected STR:** Points to the 'Download STRs TSV (1)' button.
- Reference STR ranges from PanelApp:** Points to the 'Normal Number of Repeats: 34' and 'Pathogenic Number of Repeats: 38' fields.
- Download STR pileup graph:** Points to the 'Download Pile-Up Graph' button.
- Number of copies is currently not sex aware:** Points to the '44, 57' value in the 'Number Of Copies' column.
- Select STR to include in Summary of Findings:** Points to the checkbox next to the 'AR' entry in the 'Report Events' table.

TIER1 (Expanded)

Locus: AR
Repeat motif: CAG
Normal Number of Repeats: 34
Pathogenic Number of Repeats: 38
Download Pile-Up Graph
Location: GRCh38 X:67545311-67545383
Panel: Anyotrophic lateral sclerosis/motor neuron disease (1.20)

Zygosity

Participant	Number Of Copies
112008033 Proband LP001079-DNA_B02 Affected MALE	44, 57

Report Events

Score	Genomic Entity	Mode of Inheritance	Panel Name	Panel Version	Penetrance	Clinical Indication	Interpretation Service
<input checked="" type="checkbox"/>	AR ENSG00000169083	xlinked_monosallelic			Incomplete	Charcot-Marie-Tooth disease	Tiering

Note: the current version of Expansion Hunter (the STR caller) is not sex-aware. This means STR calls on the X chromosome in males will show two alleles. This will be improved in the next release of Expansion Hunter.

From the STR tab it is possible to download the pile-up graph produced by Expansion Hunter to get a graphical view of the repeat expansion (see: STR visualisation).

7.17.5 CNV Visualisation

To view CNV results in IGV click the “Genomic Coordinates” hyperlink in either the CNV call or gene level tables. Enter your Genomics England username and password and the IGV browser will appear with a list of files to select. If you deselect all files except the coverage files (coverage.bw) and select “show tracks” the browser will load with the coverage information, labelled for each family member’s LP number. If you zoom out slightly it is possible to get a graphical overview of the CNV for the proband along side any family members (Figure 15).

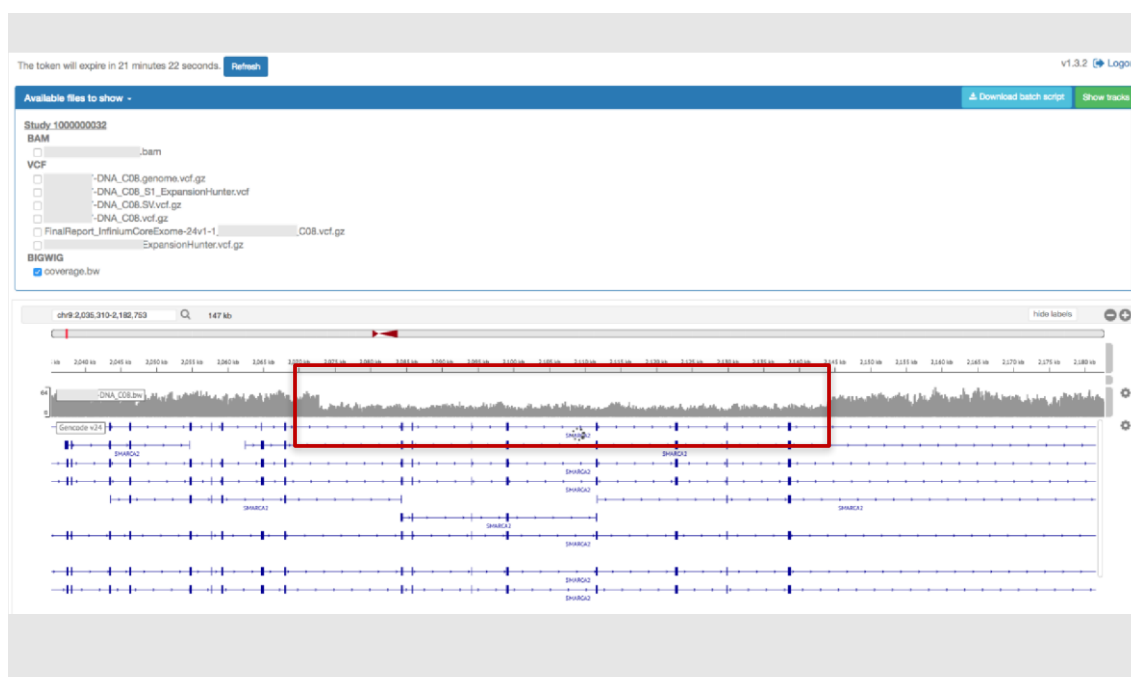


Figure 16: Example CNV view in IGV

If you have any issues loading IGV in your browser (e.g. due to slow internet connection) please try downloading the IGV Desktop application from here:

<http://software.broadinstitute.org/software/igv/download> then use the blue “Download batch script” button (Figure 15) from the web IGV and load this into your Desktop IGV browser.

7.17.6 Download variants

It is possible to download variants shown in the Interpretation Browser in a Tab Separated Values (TSV) format. By default all variants are selected for download, however, it is possible to select one or more individual variants for download (Figure 12).

Variants from the entire genome, in VCF format, can be downloaded for all family members using the VCF files tab of the Interpretation Browser (Figure 12)

7.17.7 Generating Summary of Findings using the Interpretation Browser

To generate a Summary of Findings HTML using the Interpretation Browser first select the variant(s) you wish to be included and click the “Create Summary of Findings” button (Figure 12).

A form will then appear where interpretive comments can be added at the variant and case level (Figure 16).

The screenshot shows a web form titled "Create Summary of Findings (1)". The form contains the following fields and buttons:

- Variant comments:** A text input field with a blue arrow pointing to it from a red box labeled "Variant Level Comments".
- * Your Interpretation:** A text input field with a blue arrow pointing to it from a red box labeled "Family Level Comments". Below this field is a placeholder text: "Summary of the interpretation, this should reflect the positive conclusions of this interpretation".
- Supporting evidence (pubmed ids):** A section with a "+ Add Evidence" button.
- Buttons:** "Cancel", "Save draft", and "Submit". A red box labeled "Save Comments as Draft" has a blue arrow pointing to the "Save draft" button. Another red box labeled "Create Summary of Findings" has a blue arrow pointing to the "Submit" button.

Figure 17: Create Summary of Findings

Upon clicking the submit button the Summary of Findings will be generated, the case status updated and the case moved to the "To Be Closed" tab (as described above).

7.17.8 Search for variants in genes outside of panel used for Tiering

This is a new feature that enables users to search for variants in genes that were not included in the original panel used for variant Tiering. This feature enables GMCs to assess variants in genes that might have only recently been discovered as having an association with the family's disorder.

To add variants in a gene outside of the panel applied click the "Add Gene Variants" button (Figure 12). Enter the HGNC gene symbol for the gene you would like to review variants in and click the "Search" button.

Please note: the search term is currently case sensitive so "KCNJ11" will work whereas "kcnj11" will not.

After clicking the search button variants scroll to the bottom of the variant table to view variants in that gene which can be downloaded and reviewed in the same way as Tiered variants. Please note that a phenotype must be assigned before the variant can be added to the summary of findings (Figure 17).

The screenshot shows the 'Add Gene Variants' dialog box and the 'Report Events' table. Annotations include:

- Un-tick to return all variants within gene including introns:** Points to the 'Exons only' checkbox in the dialog box.
- Search for genes of interest by entering the HGNC symbol (case sensitive):** Points to the 'Gene Symbol' input field in the dialog box.
- After clicking search variants are appended to tiering variant table:** Points to the 'Search' button in the dialog box.
- Variants in non-panel genes are highlighted green:** Points to the green row in the 'Report Events' table.
- Variants included in draft report are highlighted "yellow":** Points to the yellow row in the 'Report Events' table.
- Select report event to download / report:** Points to the 'Score' column in the 'Report Events' table.
- View additional variant annotations e.g. HGVS and allele frequencies:** Points to the 'Show Variant Details' link.
- Link out to Ensembl:** Points to the 'Ensemblid' column in the 'Report Events' table.
- Variants added by "add gene variants" have "Search" as the interpreted genome:** Points to the 'Interpreted Genome Service' column in the 'Report Events' table.
- To include an un-tiered variant in Summary of Findings a phenotype must be added:** Points to the 'Phenotype' column in the 'Report Events' table.

Figure 18: ADD GENE VARIANTS

7.18 Live Case Interpretation Support

If you require live case interpretation support when using any of our interpretation tools, please contact the Help Desk on genomics.results@nhs.net

8 Process Flow

N/A

9 Supporting

- Rare Disease Clinical Data Entry and Review Guide v2.0
- Fabric Genomics User Guide
- Congenica User Guide

10 Reference Documents

1. PanelApp Handbook <https://panelapp.genomicsengland.co.uk/>

2. ClinGen Clinical Validity Classifications Stand et al. Evaluating the clinical validity of genedisease associations: an evidence-based framework developed by the Clinical Genome Resource. 2017
3. The Development Disorder Genotype - Phenotype Database
4. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehm HL; ACMG Laboratory Quality Assurance Committee. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015 May;17(5):405-24. doi: 10.1038/gim.2015.30. Epub 2015 Mar 5. PubMed PMID: 25741868; PubMed Central PMCID: PMC4544753.
5. Smedley D, Jacobsen JO, Jäger M, Köhler S, Holtgrewe M, Schubach M, Siragusa E, Zemojtel T, Buske OJ, Washington NL, Bone WP, Haendel MA, Robinson PN. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*. 2015 10(12):2004-15.
6. Smedley D, Robinson PN. Phenotype-driven strategies for exome prioritization of human Mendelian disease genes. *Genome Med*. 2015 Jul 30;7(1):81.
7. Bone WP, Washington NL, Buske OJ, Adams DR, Davis J, Draper D, Flynn ED, Girdea M, Godfrey R, Golas G, Groden C, Jacobsen J, Köhler S, Lee EM, Links AE, Markello TC, Mungall CJ, Nehrebecky M, Robinson PN, Sincan M, Soldatos AG, Tift CJ, Toro C, Trang H, Valkanas E, Vasilevsky N, Wahl C, Wolfe LA, Boerkoel CF, Brudno M, Haendel MA, Gahl WA, Smedley D. Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficiency. *Genet Med*. 2016 Jun;18(6):608-17. doi: 10.1038/gim.2015.137.
8. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, Graul-Neumann L, Doelken S, Ehmke N, Spielmann M, Oien NC, Schweiger MR, Krüger U, Frommer G, Fischer B, Kornak U, Flöttmann R, Ardeshirdavani A, Moreau Y, Lewis SE, Haendel M, Smedley D, Horn D, Mundlos S, Robinson PN. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med*. 2014 Sep 3;6(252):252ra123.
9. Robinson PN, Köhler S, Oellrich A; Sanger Mouse Genetics Project, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D, Krawitz P, Gilissen C, Haendel M, Smedley D. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*. 2014 Feb;24(2):340-8.

11 Appendices

11.1 Appendix A – PanelApp criteria for diagnostic grade ‘green’ genes

- A. There are plausible disease-causing mutations¹ within, affecting or encompassing an interpretable functional region of this gene² identified in multiple (>3) unrelated cases/families with the phenotype³.

OR

- B. There are plausible disease-causing mutations¹ within, affecting or encompassing cis-regulatory elements convincingly affecting the expression of a single gene identified in multiple (>3) unrelated cases/families with the phenotype³.

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 setting⁴.

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.

11.2 Appendix B – SO terms

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 region at the 3' end of an intron.
Synonyms	Jannovar:splice_acceptor_variant, Seattleseq:spliceacceptor, 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	Stop lost A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript
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, initiator codon variant, initiator codon change, Jannovar:initiator_codon_variant, VAT:startOverlap
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 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: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-nearsplice, 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, 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-nearsplice, 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: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

11.3 Appendix C – Biotypes

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.