The American Journal of Human Genetics, Volume *109*

Supplemental information

Shariant platform: Enabling evidence sharing across Australian clinical genetic-testing laboratories to support variant interpretation

Emma Tudini, James Andrews, David M. Lawrence, Sarah L. King-Smith, Naomi Baker, Leanne Baxter, John Beilby, Bruce Bennetts, Victoria Beshay, Michael Black, Tiffany F. Boughtwood, Kristian Brion, Pak Leng Cheong, Michael Christie, John Christodoulou, Belinda Chong, Kathy Cox, Mark R. Davis, Lucas Dejong, Marcel E. Dinger, Kenneth D. Doig, Evelyn Douglas, Andrew Dubowsky, Melissa Ellul, Andrew Fellowes, Katrina Fisk, Cristina Fortuno, Kathryn Friend, Renee L. Gallagher, Song Gao, Emma Hackett, Johanna Hadler, Michael Hipwell, Gladys Ho, Georgina Hollway, Amanda J. Hooper, Karin S. Kassahn, Rahul Krishnaraj, Chiyan Lau, Huong Le, Huei San Leong, Ben Lundie, Sebastian Lunke, Anthony Marty, Mary McPhillips, Lan T. Nguyen, Katia Nones, Kristen Palmer, John V. Pearson, Michael C.J. Quinn, Lesley H. Rawlings, Simon Sadedin, Louisa Sanchez, Andreas W. Schreiber, Emanouil Sigalas, Aygul Simsek, Julien Soubrier, Zornitza Stark, Bryony A. Thompson, James U, Cassandra G. Vakulin, Amanda V. Wells, Cheryl A. Wise, Rick Woods, Andrew Ziolkowski, Marie-Jo Brion, Hamish S. Scott, Natalie P. Thorne, Amanda B. Spurdle, and on behalf of the Shariant Consortium

Supplemental Data

Table of Contents

Supplemental Figures

Figure S1

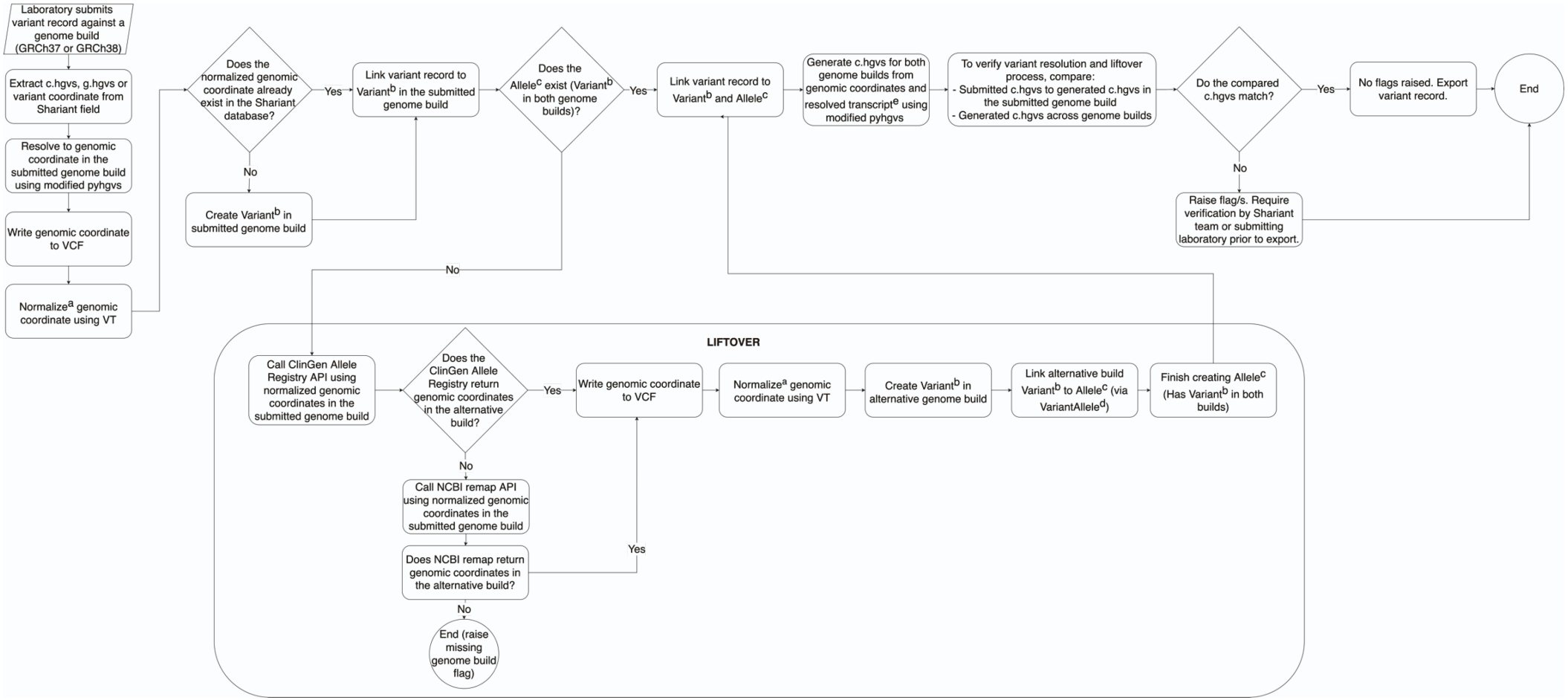


Figure S1. Variant resolution and liftover process.
Representation of process to allow for accurate aggregation and connection of variants across differing variant representations (e.g., coding DNA HGVS nomenclature (c.hgvs), genomic HGVS nomenclature (g.hgvs)) and genome builds GRCh37 and GRCh38. Liftover is performed by

the ClinGen Allele Registry[1] or National Center for Biotechnology Information Genome Remapping Service (NCBI Remap; www.ncbi.nlm.nih.gov/genome/tools/remap) Application Programming Interface (API).
[a]Normalization to left-aligned, parsimonious representation using VT[2].
[b]Variant is defined as a normalized genomic coordinate against a specific genome build.
[c]Allele denotes a Shariant generated, genome build independent identifier to connect "Variants" across genome build GRCh37 and GRCh38.
[d]VariantAllele denotes a database model used to link together Variant and Allele. It also stores the liftover method used to link the Variant to the Allele.
[e]Resolved transcripts refer to either the transcript version submitted or, if that version cannot be used, an alternative transcript version that is used for variant resolution to genomic coordinates.

Figure S2

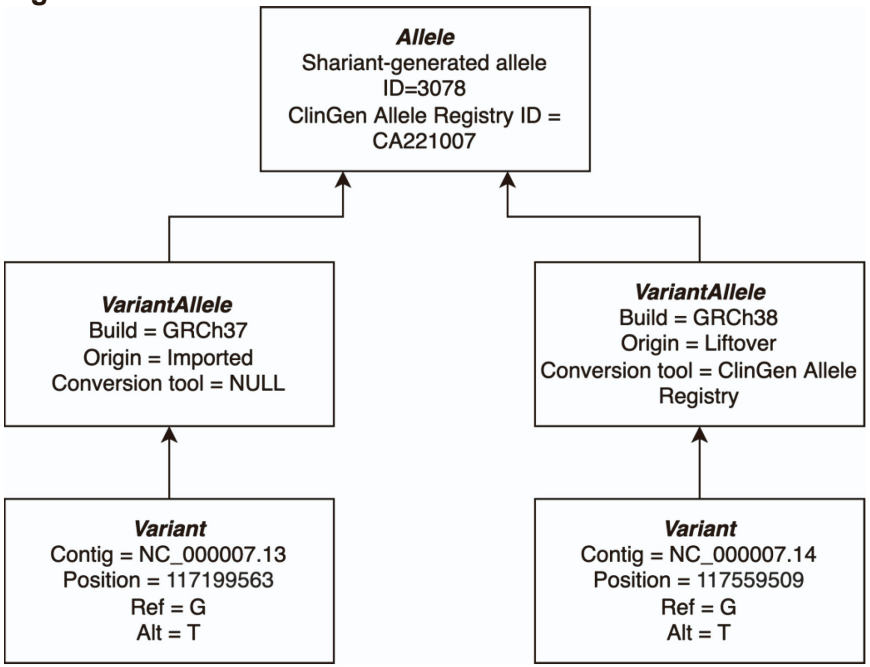


Figure S2. Database representation of example variant record submitted against genome build GRCh37.
Variant is defined as a normalized genomic coordinate against a specific genome build. VariantAllele denotes a database model used to link together Variant and Allele. It also stores the liftover method used to link the Variant to the Allele. Allele denotes a Shariant generated, genome build independent identifier to connect "Variants" across genome build GRCh37 and GRCh38.

Figure S3

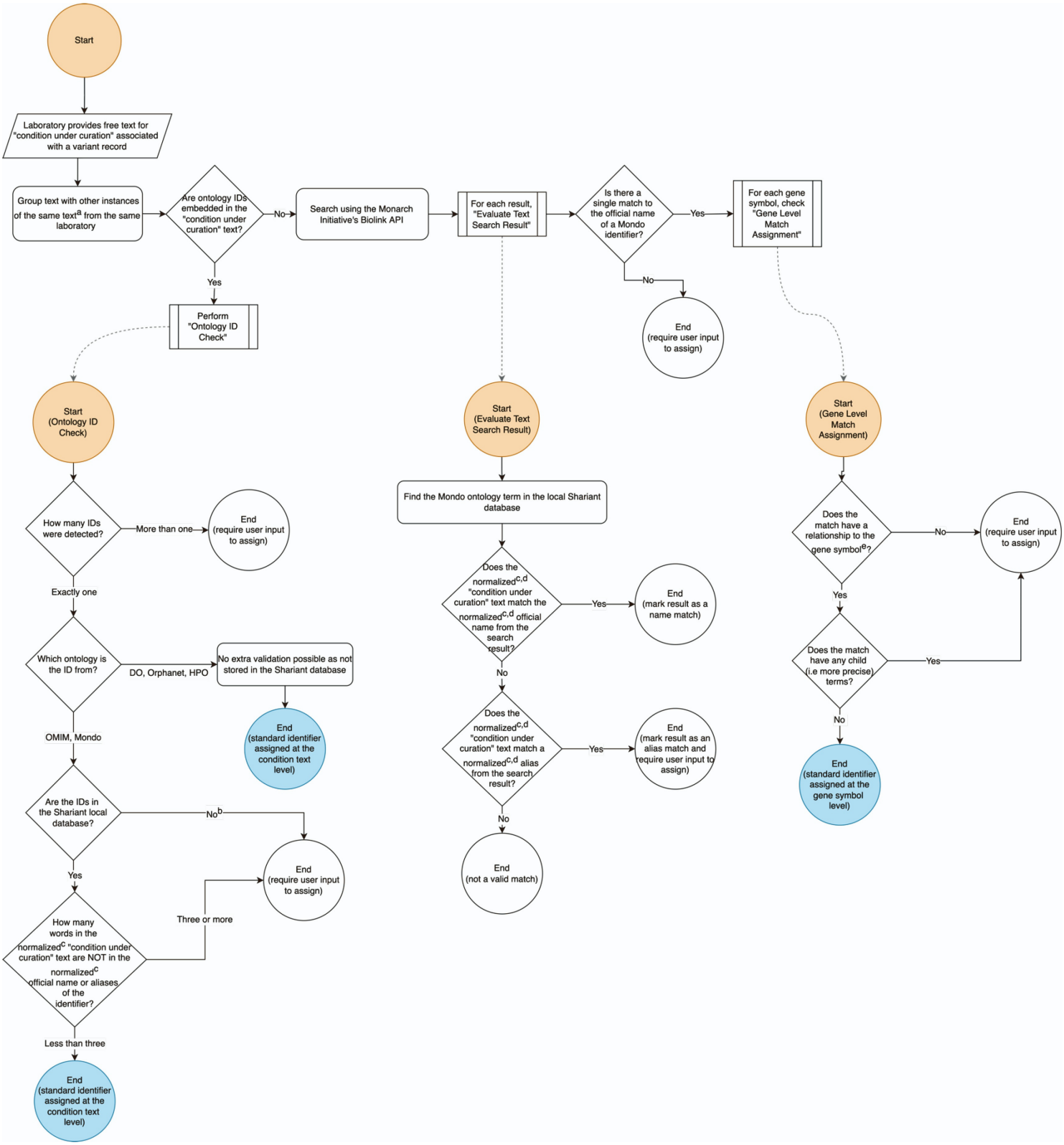


Figure S3. Condition text matching process for the automatic assignment of ontology identifiers (IDs) to "condition under curation" text. Note: This flowchart does not describe the process for suggestion of ontology identifiers that require user input. Ontologies supported include Monarch Disease Ontology (Mondo)[3], Online Mendelian Inheritance in Man (OMIM)[4],

Human Phenotype Ontology (HPO)[5], Orphanet (https://www.orpha.net/) and Disease Ontology (DO)[6].

[a]Condition under curation text grouped by converting all text to lowercase, removing punctuation and removing extra whitespace.

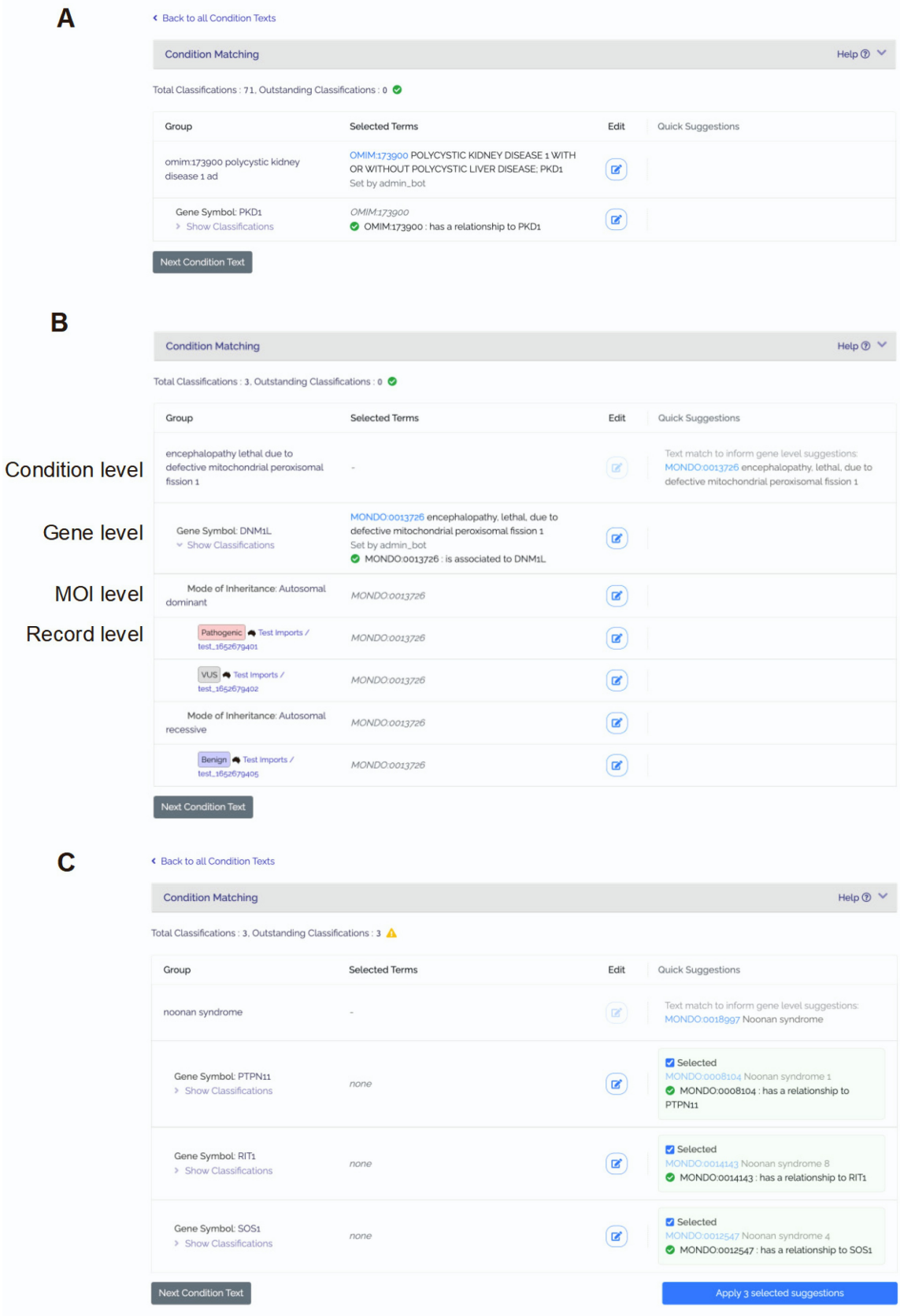
[b]Examples of ontology identifiers that do not exist in the Shariant database include OMIM gene/locus numbers. This function also acts as a means of identifying typographical errors.

[c]Normalization is performed to both the condition under curation text and official ontology name or alias, prior to comparisons for equality. It includes de-pluralizing words, ignoring common words such as “ar”, “ad”, “linked”, “xld”, “xlr”, “disability”, “disorder”, “the”, “an”, “and”, “&”, “or”, “for”, “the”, “type”, “group”, “with” and converting Roman numerals to Arabic Numbers with the exception of “X”.

[d]Normalization is performed to both the condition under curation text and official ontology name or alias, prior to comparing for equality. It involves splitting into "main descriptor" and "subtype". Most ontology terms are divided into a main descriptor and then a subtype, where the subtype can appear before or after the main descriptor. The format for this is rarely consistent e.g., MONDO:0008702 achondrogenesis type II, MONDO:0019257 hemochromatosis type 2, MONDO:0019676 brachydactyly type B. In cases where a subtype is not detected, the entire name will be considered the main descriptor. Additionally, “a” is ignored in the main descriptor but not for subtype e.g., “A brittle bone disorder” versus “Type A”.

[e]As determined by PanelApp Australia[7,8], Gene Curation Coalition[9] and Mondo[3].

Figure S4


A
Back to all Condition Texts
Condition Matching
Help
Total Classifications : 71, Outstanding Classifications : 0
Group
Selected Terms
Edit
Quick Suggestions
omim:173900 polycystic kidney disease 1 ad
OMIM:173900 POLYCYSTIC KIDNEY DISEASE 1 WITH OR WITHOUT POLYCYSTIC LIVER DISEASE; PKD1
Set by admin_bot
Gene Symbol: PKD1
Show Classifications
OMIM:173900
OMIM:173900 : has a relationship to PKD1
Next Condition Text
B
Condition Matching
Help
Total Classifications : 3, Outstanding Classifications : 0
Group
Selected Terms
Edit
Quick Suggestions
Condition level
encephalopathy lethal due to defective mitochondrial peroxisomal fission 1
-
Text match to inform gene level suggestions:
MONDO:0013726 encephalopathy, lethal, due to defective mitochondrial peroxisomal fission 1
Gene level
Gene Symbol: DNM1L
Show Classifications
MONDO:0013726 encephalopathy, lethal, due to defective mitochondrial peroxisomal fission 1
Set by admin_bot
MONDO:0013726 : is associated to DNM1L
MOI level
Mode of Inheritance: Autosomal dominant
MONDO:0013726
Record level
Pathogenic
Test Imports / test_1652679401
MONDO:0013726
VUS
Test Imports / test_1652679402
MONDO:0013726
Mode of Inheritance: Autosomal recessive
MONDO:0013726
Benign
Test Imports / test_1652679405
MONDO:0013726
Next Condition Text
C
Back to all Condition Texts
Condition Matching
Help
Total Classifications : 3, Outstanding Classifications : 3
Group
Selected Terms
Edit
Quick Suggestions
noonan syndrome
-
Text match to inform gene level suggestions:
MONDO:0018997 Noonan syndrome
Gene Symbol: PTPN11
Show Classifications
none
Selected
MONDO:0008104 Noonan syndrome 1
MONDO:0008104 : has a relationship to PTPN11
Gene Symbol: RIT1
Show Classifications
none
Selected
MONDO:0014143 Noonan syndrome 8
MONDO:0014143 : has a relationship to RIT1
Gene Symbol: SOS1
Show Classifications
none
Selected
MONDO:0012547 Noonan syndrome 4
MONDO:0012547 : has a relationship to SOS1
Next Condition Text
Apply 3 selected suggestions

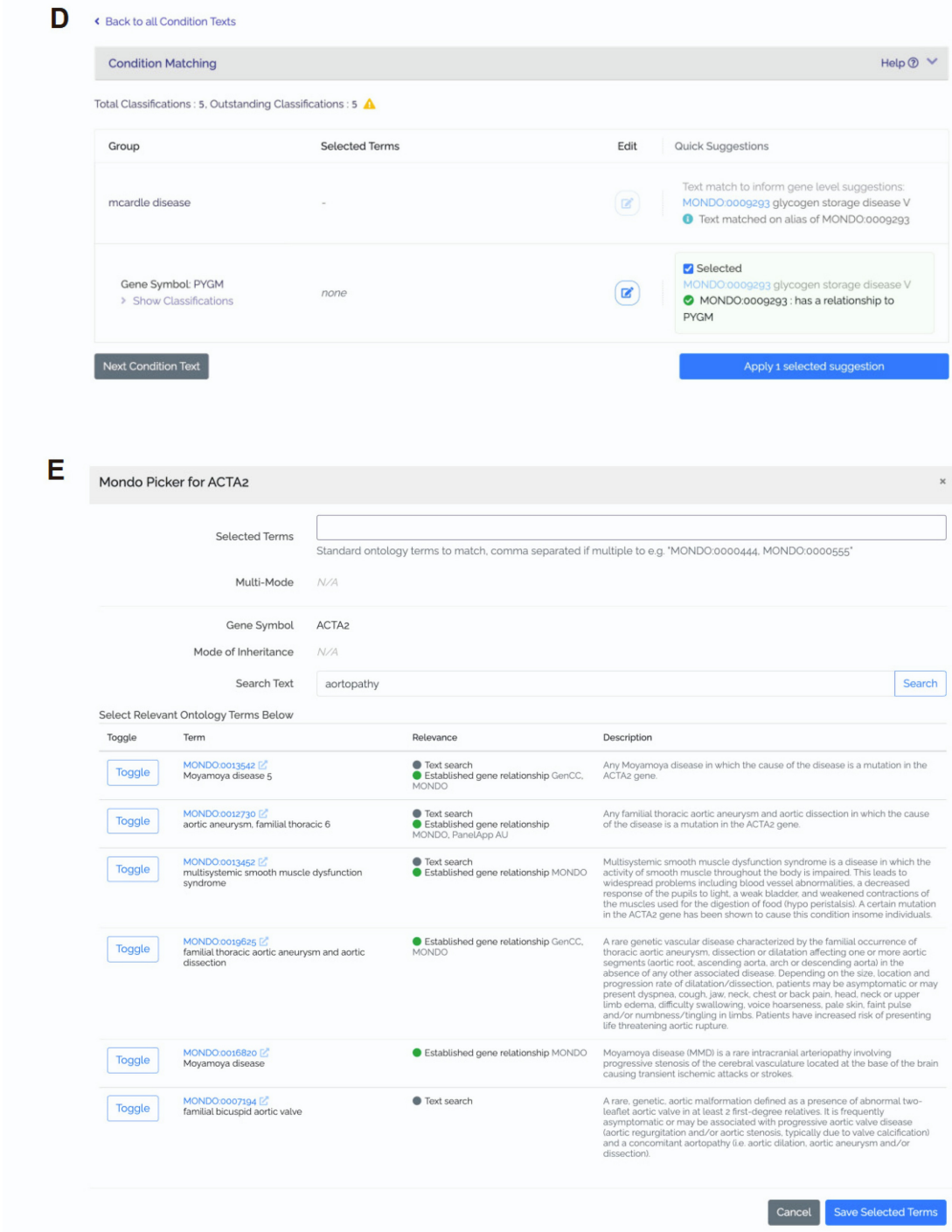


Figure S4. Shariant condition text matching interface.
(A) Automated matching due to submission of a standard ontology. (B) Automated matching of free text condition to a Mondo Disease Ontology (Mondo)[3] identifier with designation of assignment hierarchy at the condition, gene, mode of inheritance (MOI) and record level, respectively. (C) Suggestion of gene-specific Mondo identifiers. (D) Suggestion of Mondo identifier based on the free text condition submission of an alias/synonym. (E) Search functionality against a gene symbol and free text condition showing gene-disease relationships and Mondo identifier description.

Supplemental Tables

Table S1. Evaluation framework for assessment of sharing tools

See separate spreadsheet file

Table S2. Laboratory interpretation software connection to Shariant

Laboratory	Interpretation software	Submission to Shariant	Import into laboratory interpretation system
Organization 1	System 1	Upload of vendor exported file format to Shariant web portal (~ monthly)	Shariant export tailored for import into interpretation software (~ monthly)
Organization 2 (five laboratories)	VariantGrid	API – hourly	API - hourly
Organization 3	System 1	API – weekly	Shariant export tailored for import into interpretation software (monthly)
Organization 4 (two laboratories)	System 1	Upload of vendor exported file format to Shariant web portal (~ every two months)	Shariant export tailored for import into interpretation software (~ quarterly)
Organization 5	In-house tool (submission)/ System 1 (import)	API - weekly	Shariant export tailored for import into interpretation software (~ six monthly)
Organization 6	System 2	Upload of vendor exported file format to Shariant web portal (quarterly)	Not applicable – in progress

Table S3. Shariant evidence fields captured as of May 2022

See separate spreadsheet file

Table S4. Overview of mapping of laboratory variant records to Shariant mandatory/strongly recommended fields

Shariant Field	ClinVar Mandatory Field	Structured Data (number laboratories)	Free Text (number laboratories)	Other Mapping (number laboratories)[a]	Other - Explanation
Genome build	N	11	0	0	
Variant representation (e.g., c.hgvs)	Y	11	0	0	
Clinical significance (classification)	Y	11	0	0	
Date last curated/reviewed	N	7	0	4	Taken from date of last update of the record.
Condition under curation (standard ontology not required)	Y (standard ontology required)	6	4	1	Based on gene symbol e.g., *BRCA2* and breast-ovarian cancer, familial, susceptibility to, 2.
Zygosity	N	6	4	1	Assumed based on variant allele frequency.
Allele origin (germline/somatic)	Y	6	0	5	Auto-populated as laboratories only submit germline interpretations.
Assertion method (e.g., ACMG/AMP guidelines)	Y	10	0	1	ACMG/AMP guidelines auto-populated.
Curation context (e.g., accredited diagnostic testing)	Y	5	0	6	Auto-populated as Shariant laboratories are restricted to those undertaking accredited diagnostic testing.
ACMG/AMP evidence criteria (e.g., BA1)	N	11	0	0	
Interpretation summary	N	8	3	0	
Literature	N	5	2[b]	4	Aggregation of PMIDs under Shariant Citations section[c].
Affected status	Y	5	1	5	Unpopulated at this time.

[a]Field not available in standard export; [b]Free text parsing performed for literature heading, stored as designated literature field; [c]Free text parsing for PubMed identifiers (PMIDs) in all free text submitted, as occurs by default for all laboratories submitting to Shariant.

Table S5. Overview of free text parsing required for population of Shariant fields

Shariant Field	Number of laboratories requiring free text parsing to populate the relevant Shariant fields	Description of parsing, including examples of terms sought (H = heading, T= term)[a]
Condition under curation	4	'Condition' (H), 'Reported Disease Association Name' (H), 'Reported Disease Association ID' (H), 'Phenotype association' (H), Mondo identifier in the forms of 'MONDO:[number]', 'MONDO#[number]', 'MONDO[number], 'MONDO [number] (T), OMIM identifier in the forms described for Mondo, accepting prefixes of OMIM or MIM (T), HPO identifier in the forms described for Mondo, accepting prefixes of HPO or HP (T)
Zygosity	4	'Zygosity' (H), 'homozygous' (T), 'heterozygous' (T), 'compound heterozygous' (T), 'hemizygous' (T)
Interpretation Summary	3	Free text up to a standard delimiter "|||||", free text with removal of 'curated against' and affected status terms, free text with removal of internal communication determined through a block list of 21 keywords such as 'authorised', 'agrees', 'check', 'discussed', 'said', 'to be reviewed', 'remove statement', 'in [sample or patient][6-digit number]'
Literature	2	'References' (H), a combination of 'ACMG justification', 'evidence justification' and 'report description' (H)
Affected Status	1	'Unaffected' (T), 'affected' (T), 'unknown' (T)

[a]Heading refers to a label at the beginning of a section, whereby all free text in that section is included in the Shariant field. Term can refer to a single word or standard prefix, usually followed by a number, that is searched for and included in the appropriate Shariant field.

Supplemental Material and Methods

Landscape analysis

Survey questions were framed to capture activities of genetic testing labs for germline and somatic variation; assess classification methods used and alignment with international standards; assess expertise in variant classification for different diseases; understand protocols for re-evaluation of genes/variants and re-issue of reports and capture views/protocols for report of incidental findings (note: previous surveys by the Royal College of Pathologists of Australasia had predominantly focused on number and types of tests conducted, as well as sources of funding for these tests[10]).

The web-based survey was developed and trialed with representatives from two laboratories and revised for content and clarity. Responses to the survey were obtained in three stages: (1) A link to the survey was emailed to a representative from 46 clinically accredited genetic testing laboratories in November 2016. Contacts for laboratories providing a molecular genetics service under Human Pathology were taken from the National Association of Testing Authorities (NATA; https://nata.com.au/) website in October 2016; (2) Responses were reviewed and incomplete responses flagged; (3) Laboratories that had not completed the survey had their contacts reviewed and were telephoned in January 2017. Laboratories that provided an incomplete response were also followed up by telephone in parallel. Responses to the survey were completed online or by telephone, either by the original contact or a designated replacement contact.

After consultation, 46 laboratories were collapsed to 34 independent organizations (resolving multiple sites for the same laboratory or multiple laboratories of one organization). Of these 34 laboratories, only 30 (16 public and 14 private) were conducting clinical grade genetic testing (i.e. NATA compliant) at the time of the survey.

Evaluation of available variant interpretation sharing tools and selection of a platform

Nine existing tools were identified as candidate sharing tools by ET, ABS and other Australian Genomics' collaborators, including commercial and non-commercial variant interpretation tools and databases in use by Australian laboratories. Preliminary evaluation was undertaken by ET and ABS against an evaluation framework (Table S1). Three tools were prioritized for formal evaluation by representatives from three Australian clinical genetic testing laboratories, including laboratory scientists and bioinformaticians/software developers. The process was as follows: initial demonstration by teleconference, trialing of the tool over the period of one month with fortnightly Q&A calls available to laboratory representatives, assessment of the tool against the evaluation framework (Table S1). Each tool under consideration also submitted a proposal outlining pre-existing functionalities relevant to the purpose of variant interpretation sharing, as well as budget required for further development to meet required functionalities outlined in the evaluation framework. Following formal evaluation, an external clinical genetic testing laboratory was asked to evaluate and rank the prioritized tools. The top two ranked tools then underwent a detailed technical evaluation (JVP).

Shariant Documentation

Terms of Use

Each contributing laboratory is required to undertake legal review and execution of the Terms of Use by an authorized representative. To accommodate modifications introduced at each

separate legal review without the need for laboratories to re-sign, the Terms of Use include a clause allowing for the introduction of minor amendments. Laboratories are notified of and required to acknowledge such amendments at the time of next login to the platform.

At present, conditions include:

- Each laboratory uploading data retains ownership and intellectual property over that data;
- Access to Shariant is limited to Australian clinically accredited genetic testing laboratories and requesting clinicians;
- Upload of patient identifiable information is prohibited;
- Upload of data contributed by an external laboratory to a third-party platform is prohibited.

Additional documentation

Additional documentation was developed to address questions and concerns from the consultation phase and the Terms of Use review. The main document largely focused on issues around security, extent of sequencing and clinical data to be captured, and location of data storage.

Automated transformation of data

To allow for scalability over time, a focus was put on automated transformation of laboratory system-formatted exports, aiming to maximize import of the provided interpretation information. Five laboratories (two interpretation systems) opted for data transformation to occur at the Shariant end, and one laboratory (using a third interpretation system) transformed the data using a co-developed program prior to submitting to Shariant. (Another five laboratories used the VariantGrid interpretation system that shares the same format as Shariant, and thus data transformation was not required).

Code required to transform data at the Shariant end was tailored to each laboratory, even where laboratories used the same interpretation system. Originally, one program was written for each interpretation system, with specific parameters included for different laboratories using the same system. However, with an increasing number of laboratories, it became evident that conforming common functionality based on the interpretation system was not feasible. As a result, a single program was developed to provide simple functionality for generic automated data transformation, while also allowing for implementation of more complex laboratory-specific parameters. This was mainly due to a large reliance on free text parsing to obtain all mandatory/ strongly recommended fields required in Shariant.

All laboratories were able to provide structured data for the mandatory/strongly recommended fields: genome build, variant representation (e.g., c.hgvs), clinical significance and American College of Medical Genetics and Genomics/Association for Molecular Pathology (ACMG/AMP)[11] criteria (Table S4). Standard ACMG/AMP guidelines[11] were used by 9/11 laboratories. Two laboratories used non-standard guidelines that required mapping back to standard ACMG/AMP and inclusion of explanatory text where there was a difference to ACMG/AMP explanations. Mapping data was kept in source code spreadsheets, often versioned, with each variant record providing a version of the guidelines to map to. This allowed for capturing of changes to these non-standard guidelines over time.

Exports from two interpretation systems (relevant to five laboratories) required free text parsing to populate mandatory/strongly recommended Shariant fields (Table S4). Two of these fields (affected status, condition under curation) are considered mandatory for ClinVar submission. Structured text and/or standard terms were searched for and used to populate

the relevant Shariant fields (Table S5). For example, searching for headings such as “zygosity” and “condition”, terms such as “homozygous” and “heterozygous”, or standard ontology identifiers. Although laboratories were asked to use a single standard heading for each Shariant data field where possible, historical records and inconsistent within-laboratory data formats required free text scanning for all possible combinations of structured text and standard terms. Logic was also incorporated to first search for one heading, before falling back to other headings if not available. Free text parsing also required modification over time as laboratories changed their data formats, usually towards that of a more structured or standard format.

Variant resolution and liftover

The following process was developed to allow for accurate aggregation and connection of variants across differing variant representations and genome builds GRCh37 and GRCh38 (Figure S1).

Resolution of submitted variant representation to genomic coordinate in submitted genome build

Submission of variant records to Shariant requires the mandatory field “genome build” as well as at least one of the following variant representations: *variant coordinate* (e.g.,"7:117559509 G>T"), genomic Human Genome Variation Society (HGVS) nomenclature (*g.hgvs)* (e.g.,"NC_000007.14:g.117559509G>T") or coding DNA HGVS nomenclature (*c.hgvs)* (e.g.,"NM_000492.3(CFTR):c.1438G>T").

The variant representation is automatically resolved to the genomic coordinate in the submitted genome build using the Counsyl hgvs (https://github.com/counsyl/hgvs) Python library with modifications (herein referred to as “modified pyhgvs”). Modified pyhgvs code is available at https://github.com/SACGF/hgvs.

Genome coordinate conversion requires alignment information (e.g., exon coordinates) for a large number of transcript versions. To obtain these, transcript data needed for HGVS conversion was extracted from gene annotation files (General Transfer Format (GTF)/General Feature Format (GFF)) available on the RefSeq[12] and Ensembl websites[13] (see *Transcript version GTF/GFF files* below). The transcript data was converted to gzipped JSON format, and code libraries were written to load and convert these transcripts for use in the two most popular Python HGVS libraries: Pip packages pyhgvs (Counsyl; https://github.com/counsyl/hgvs), and hgvs (Biocommons)[14].

Although hgvs (Biocommons) is not currently used for the resolution of variant representations in Shariant, provision of transcripts to this project provides a community resource and reduces future work to adopt that library as a second algorithm to verify conversion.

Additionally, modification of the Counsyl hgvs repository (https://github.com/counsyl/hgvs) was required to match variants that were not previously supported, as well as to correct any coordinate mapping errors found. Support was added for noncoding and LRG transcripts, mitochondrial (m.) HGVS as well as to account for alignment gaps.

Alignment gaps occur when RefSeq transcripts differ from the reference sequence, and align with resulting insertions/deletions, which must be taken into account for accurate coordinate conversion. RefSeq alignment gaps were present in 3% of submitted RefSeq GRCh37 transcripts and 0.17% of GRCh38 transcripts.

RefSeq only reported alignment gaps in GTFs after GRCh37 patch 13 (August 2013), so a small percentage of earlier transcript versions contained unreported gaps. To identify these gaps, the sum of exon lengths is compared with the transcript sequence (accounting for untrimmed poly-A tails). If the length differs, the transcript is marked as unusable for resolution. This additional verification step does not account for unreported gapped alignments with an equal number of insertions and deletions; however, this scenario is captured by a verification at the end of the process (see *Verification of variant resolution and liftover* below).

After accounting for alignment gaps, it was still not possible to obtain all transcript versions for both GRCh37 and GRCh38. As a result, it was necessary to allow for matching to alternative transcript versions than the version submitted by the laboratory (transcripts matched to are denoted as resolved transcripts). Higher transcript versions are first queried in ascending order, followed by lower transcript versions in descending order. If no alternative transcript versions are found or matching to the alternative transcript fails (e.g., coordinate outside transcript boundaries), the c.hgvs variant representation (with the transcript replaced in the same order as above) is sent to the ClinGen Allele Registry[1]. The first successful result is used to retrieve genomic coordinates against GRCh38.

In the event of no version of the transcript being found within Shariant, the RefSeq and Ensembl Application Programming Interfaces (APIs) are queried to identify whether the transcript exists outside of the Shariant platform and/or the transcript is invalid due to submitter error (e.g., typographical or copy/paste error). The information is used by the Shariant team to determine next steps for resolution of the variant (e.g., push back to the laboratory to fix the transcript in their system or for the Shariant team to retrieve new transcript data).

Normalization of genomic coordinate in the submitted genome build

Following resolution to a genomic coordinate in the submitted genome build, the genomic coordinate is written to Variant Call Format (VCF) and normalized (left-aligned, parsimonious) using VT[2]. If the normalized genomic coordinate already exists in Shariant, the variant record is linked to a “Variant” (defined as a normalized genomic coordinate against a specific genome build). A Variant is created if the genomic coordinate does not exist.

Liftover of variant to alternative genome build and creation of allele

Proceeding generation of a Variant, the Shariant database is queried to determine whether an “Allele” (defined as a Shariant generated, genome build independent identifier linking together a Variant in both genome builds) exists. For each new Allele, the Variant in the submitted genome build is used to query the ClinGen Allele Registry[1] using the API, which allows for liftover of the variant by providing genomic coordinates in the alternative build. It also returns a genome build independent unique ClinGen Allele Registry identifier. In the event of an error being returned, the National Center for Biotechnology Information Genome Remapping Service (NCBI Remap) API (www.ncbi.nlm.nih.gov/genome/tools/remap) is queried for genomic coordinates in the alternative build (no unique identifier is returned). The genomic coordinate returned is then written to a VCF and normalized using VT[2], a Variant in the alternative genome build created and the Variant linked to an Allele via a “VariantAllele” (a database model used to link together Variant and Allele. It also stores the liftover method used to link the Variant to the Allele. See Figure S2).

A flag is raised if the ClinGen Allele Registry and NCBI Remap are both unable to provide genomic coordinates in the alternative genome build. This flag is used to inform users but cannot be resolved manually.

Generation of c.hgvs in genome build GRCh37 and GRCh38

For verification of the process, c.hgvs using the resolved transcript is generated from the Variant for both builds (see *Resolution of submitted variant representation to genomic coordinate in submitted genome build* above). The generated c.hgvs is resolved as per HGVS conventions using the modified pyhgvs algorithm which supports a specific subset of HGVS recommendations (e.g., right alignment, insertions to duplications; http://varnomen.hgvs.org/).

Verification of variant resolution and liftover

Comparison includes: (1) submitted c.hgvs and generated c.hgvs in the submitted genome build and (2) generated c.hgvs across GRCh37 and GRCh38. Upon detection of differences, flags are raised and require human intervention (either by the submitting laboratory or the Shariant team) to accept or reject the match. The variant is not exported from Shariant until all flags are resolved. These flags are used to identify a number of differences including submitted c.hgvs that is not normalized (e.g., not right aligned as per HGVS convention, described as an insertion rather than a duplication), reference base not matching the imported reference base (possibly due to genome build patches), transcript version changes between the submitted transcript version and resolved transcript version or the resolved transcript versions across genome builds and change in c.hgvs between genome builds due to undetected alignment gaps. Additionally, if more than one variant representation is provided by the laboratory, all representations are converted to a genomic coordinate and an error raised if the resulting genomic coordinates are not equivalent.

Overview of variant matching issues encountered

As at March 2022, submission of variant records using the variant representation c.hgvs has accounted for 99.9% of the records in Shariant. All variants were lifted over (i.e. all alleles had a genomic coordinate generated in both GRCh37 and GRCh38); however, comparison of generated c.hgvs across GRCh37 and GRCh38 resulted in approximately 3.3% of variants requiring human intervention to accept or reject the match. When examining the submitted c.hgvs and generated c.hgvs in the submitted genome build, 2.7% of total records were flagged for c.hgvs differences (e.g., change of reference base, right alignment, transcript version change).

Additionally, conversion of c.hgvs to genomic coordinates presented a number of difficulties. Laboratories have used transcripts from RefSeq (99.5%) and Ensembl (0.5%), as well as multiple different transcript versions even within a laboratory; over 2700 total transcript versions were identified. Transcripts were outdated (more than one version behind the latest version stored in Shariant) for 88% of variant records. The need to support a large range of transcripts arose due to differences in laboratory interpretation systems, choice of genome build, and in some instances, due to submission of historical data.

Notably, the tooling developed to support the large number of transcripts and versions, increased the number of resolvable transcripts to over 893k, compared to 141k using the previously largest collection Universal Transcript Archive (https://github.com/biocommons/uta). The code to retrieve and convert transcript versions to JSON, and use them with the two Python HGVS libraries has been released as the open source project cdot (http://cdot.cc/).

Transcript version GTF/GFF files

Ensembl GRCh37
ftp://ftp.ensembl.org/pub/grch37/release-82/gff3/homo_sapiens/Homo_sapiens.GRCh37.82.gff3.gz

ftp://ftp.ensembl.org/pub/grch37/release-85/gff3/homo_sapiens/Homo_sapiens.GRCh37.85.gff3.gz
ftp://ftp.ensembl.org/pub/grch37/release-87/gff3/homo_sapiens/Homo_sapiens.GRCh37.87.gff3.gz

Ensembl GRCh38
ftp://ftp.ensembl.org/pub/release-81/gff3/homo_sapiens/Homo_sapiens.GRCh38.81.gff3.gz
ftp://ftp.ensembl.org/pub/release-82/gff3/homo_sapiens/Homo_sapiens.GRCh38.82.gff3.gz
ftp://ftp.ensembl.org/pub/release-83/gff3/homo_sapiens/Homo_sapiens.GRCh38.83.gff3.gz
ftp://ftp.ensembl.org/pub/release-84/gff3/homo_sapiens/Homo_sapiens.GRCh38.84.gff3.gz
ftp://ftp.ensembl.org/pub/release-85/gff3/homo_sapiens/Homo_sapiens.GRCh38.85.gff3.gz
ftp://ftp.ensembl.org/pub/release-86/gff3/homo_sapiens/Homo_sapiens.GRCh38.86.gff3.gz
ftp://ftp.ensembl.org/pub/release-87/gff3/homo_sapiens/Homo_sapiens.GRCh38.87.gff3.gz
ftp://ftp.ensembl.org/pub/release-88/gff3/homo_sapiens/Homo_sapiens.GRCh38.88.gff3.gz
ftp://ftp.ensembl.org/pub/release-89/gff3/homo_sapiens/Homo_sapiens.GRCh38.89.gff3.gz
ftp://ftp.ensembl.org/pub/release-90/gff3/homo_sapiens/Homo_sapiens.GRCh38.90.gff3.gz
ftp://ftp.ensembl.org/pub/release-91/gff3/homo_sapiens/Homo_sapiens.GRCh38.91.gff3.gz
ftp://ftp.ensembl.org/pub/release-92/gff3/homo_sapiens/Homo_sapiens.GRCh38.92.gff3.gz
ftp://ftp.ensembl.org/pub/release-93/gff3/homo_sapiens/Homo_sapiens.GRCh38.93.gff3.gz
ftp://ftp.ensembl.org/pub/release-94/gff3/homo_sapiens/Homo_sapiens.GRCh38.94.gff3.gz
ftp://ftp.ensembl.org/pub/release-95/gff3/homo_sapiens/Homo_sapiens.GRCh38.95.gff3.gz
ftp://ftp.ensembl.org/pub/release-96/gff3/homo_sapiens/Homo_sapiens.GRCh38.96.gff3.gz
ftp://ftp.ensembl.org/pub/release-97/gff3/homo_sapiens/Homo_sapiens.GRCh38.97.gff3.gz
ftp://ftp.ensembl.org/pub/release-98/gff3/homo_sapiens/Homo_sapiens.GRCh38.98.gff3.gz
ftp://ftp.ensembl.org/pub/release-99/gff3/homo_sapiens/Homo_sapiens.GRCh38.99.gff3.gz
ftp://ftp.ensembl.org/pub/release-100/gff3/homo_sapiens/Homo_sapiens.GRCh38.100.gff3.gz
ftp://ftp.ensembl.org/pub/release-101/gff3/homo_sapiens/Homo_sapiens.GRCh38.101.gff3.gz
ftp://ftp.ensembl.org/pub/release-102/gff3/homo_sapiens/Homo_sapiens.GRCh38.102.gff3.gz
ftp://ftp.ensembl.org/pub/release-103/gff3/homo_sapiens/Homo_sapiens.GRCh38.103.gff3.gz
ftp://ftp.ensembl.org/pub/release-104/gff3/homo_sapiens/Homo_sapiens.GRCh38.104.gff3.gz
ftp://ftp.ensembl.org/pub/release-105/gff3/homo_sapiens/Homo_sapiens.GRCh38.105.gff3.gz

RefSeq GRCh37
http://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Homo_sapiens/ARCHIVE/BUILD.37.3/GFF/ref_GRCh37.p5_top_level.gff3.gz
http://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Homo_sapiens/ARCHIVE/ANNOTATION_RELEASE.103/GFF/ref_GRCh37.p9_top_level.gff3.gz
http://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Homo_sapiens/ARCHIVE/ANNOTATION_RELEASE.104/GFF/ref_GRCh37.p10_top_level.gff3.gz
http://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Homo_sapiens/ARCHIVE/ANNOTATION_RELEASE.105/GFF/ref_GRCh37.p13_top_level.gff3.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/105.20190906/GCF_000001405.25_GRCh37.p13/GCF_000001405.25_GRCh37.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/105.20201022/GCF_000001405.25_GRCh37.p13/GCF_000001405.25_GRCh37.p13_genomic.gff.gz

RefSeq GRCh38
http://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Homo_sapiens/ARCHIVE/ANNOTATION_RELEASE.106/GFF/ref_GRCh38_top_level.gff3.gz

http://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Homo_sapiens/ARCHIVE/ANNOTATION_RELEASE.107/GFF/ref_GRCh38.p2_top_level.gff3.gz
http://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Homo_sapiens/ARCHIVE/ANNOTATION_RELEASE.108/GFF/ref_GRCh38.p7_top_level.gff3.gz
http://ftp.ncbi.nlm.nih.gov/genomes/archive/old_refseq/Homo_sapiens/ARCHIVE/ANNOTATION_RELEASE.109/GFF/ref_GRCh38.p12_top_level.gff3.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109/GCF_000001405.38_GRCh38.p12/GCF_000001405.38_GRCh38.p12_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20190607/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20190905/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20191205/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20200228/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20200522/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20200815/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20201120/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20210226/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20210514/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz
http://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/annotation/annotation_releases/109.20211119/GCF_000001405.39_GRCh38.p13/GCF_000001405.39_GRCh38.p13_genomic.gff.gz

Condition Text Matching

Initially, approximately 70% of records did not have a standard ontology identifier assigned. To facilitate submission of variant interpretations from Shariant to ClinVar, as well as to improve the data in Shariant overall, functionality was introduced to identify standard ontology identifiers if provided, and also to match free text conditions to a standard Mondo Disease Ontology (Mondo) identifier[3].

Automated matching

Figure S3 describes the process undertaken to automatically assign ontology identifiers. For Shariant variant records with one standard ontology included in the submitted condition under curation text, the identifier is automatically assigned. Standard ontologies supported include Mondo[3], Online Mendelian Inheritance in Man (OMIM)[4], Human Phenotype Ontology (HPO)[5], Orphanet (https://www.orpha.net/) and Disease Ontology (DO)[6], with additional verification undertaken for OMIM and Mondo identifiers (Figure S4A). Assignment is performed at the condition text level (see *Assignment hierarchy* below).

If no ontology identifiers are included in the condition, free text matching is performed. The Monarch Initiative's[15,16] Biolink API (https://github.com/monarch-initiative/biolink-api) is first queried to find candidates for consideration. This API provides a Solr based search for matching between text and standard terms. Automated matching of free text to a Mondo identifier requires satisfaction of a number of pre-defined criteria such as exact match of the Mondo identifier official name to the free text, presence of a valid gene-disease relationship with the variant (see *Gene-disease relationships* below), and being the most specific match in

the Mondo hierarchy (i.e. a child term) (Figure S3). Assignment is performed at the gene level (see *Assignment hierarchy* below, Figure S4B).

Matches requiring user input

In the event of matches not meeting the pre-specified criteria, Mondo identifiers are provided as suggestions at the gene-level, requiring user confirmation (Figure S4C). Suggestions provided can also include matching of free text submitted to Shariant based on a synonym of a Mondo identifier and/or an acronym (Figure S4D). Human intervention is required to verify, as synonyms are not always exact and the same acronym can match to multiple distinct conditions. Additionally, users are able to search using free text and assign Mondo identifiers manually, with information provided on gene-disease relationships (Figure S4E).

Assignment hierarchy

Assignment can be performed per laboratory within a hierarchy, the top level being the condition text level (i.e. for all records with the same condition text), followed by the gene level (i.e. for all records with the same condition text within a gene), mode of inheritance level (i.e. for all records with the same condition text within a gene and with the same mode of inheritance) and individual record level (i.e. each record can have a specific identifier assigned if needed), respectively (Figure S4B). Records below the level that the ontology identifier has been assigned against, will inherit that identifier. Additionally, assignment of an identifier at a particular level will be applied to all future records that fit at the assigned level or below.

Gene-disease relationships

Matching of free text (automated or manual) was found to be more robust when taking into account the gene symbol of the variant. As a result, gene symbol matching was integrated into the condition text matching process as follows. Gene-disease relationships are deemed valid if present in PanelApp Australia[7,8] (green genes only), Gene Curation Coalition[9] (GenCC; definitive and strong assertions only) or Mondo[3]. PanelApp Australia is queried automatically via the API and GenCC (excluding records from PanelApp Australia) and Mondo loaded periodically via their TSV download (https://search.thegencc.org/download) and JSON file (https://mondo.monarchinitiative.org/pages/download/), respectively.

Analysis of Shariant data to study nationwide impact of new recommendations and evidence

All shared variant records in Shariant were exported on 14th December 2021. Variant records included a combination of laboratories submitting per variant and per patient. Records where the variant submitted was not matched and/or no ACMG/AMP criteria[11] had been assigned a strength, were removed. If duplicate records for the same variant existed for one laboratory, only the most recently curated record was included in the analysis; that is, for each laboratory, only unique variants were considered for analysis. For the PM2 analysis, all variant records from one laboratory were also excluded due to non-conformity with the ACMG/AMP guidelines. Additionally, a point-based approach was used to determine the initial and resulting classification as per Tavtigian et al[17]. For the functional analysis, all variant records that had a strength assigned for BS3/PS3 were also removed.

References

1. Pawliczek, P., Patel, R.Y., Ashmore, L.R., Jackson, A.R., Bizon, C., Nelson, T., Powell, B., Freimuth, R.R., Strande, N., Shah, N., et al. (2018). ClinGen Allele Registry links information about genetic variants. Hum Mutat *39*, 1690-1701. 10.1002/humu.23637.
2. Tan, A., Abecasis, G.R., and Kang, H.M. (2015). Unified representation of genetic variants. Bioinformatics *31*, 2202-2204. 10.1093/bioinformatics/btv112.
3. Mungall, C.J., McMurry, J.A., Kohler, S., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2017). The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res *45*, D712-D722. 10.1093/nar/gkw1128.
4. McKusick, V.A. (1998). Mendelian inheritance in man: a catalog of human genes and genetic disorders (JHU Press).
5. Kohler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., et al. (2021). The Human Phenotype Ontology in 2021. Nucleic Acids Res *49*, D1207-D1217. 10.1093/nar/gkaa1043.
6. Schriml, L.M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., Felix, V., Jeng, L., Bearer, C., Lichenstein, R., et al. (2019). Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res *47*, D955-D962. 10.1093/nar/gky1032.
7. Martin, A.R., Williams, E., Foulger, R.E., Leigh, S., Daugherty, L.C., Niblock, O., Leong, I.U.S., Smith, K.R., Gerasimenko, O., Haraldsdottir, E., et al. (2019). PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet *51*, 1560-1565. 10.1038/s41588-019-0528-2.
8. Stark, Z., Foulger, R.E., Williams, E., Thompson, B.A., Patel, C., Lunke, S., Snow, C., Leong, I.U.S., Puzriakova, A., Daugherty, L.C., et al. (2021). Scaling national and international improvement in virtual gene panel curation via a collaborative approach to discordance resolution. Am J Hum Genet *108*, 1551-1557. 10.1016/j.ajhg.2021.06.020.
9. DiStefano, M.T., Goehringer, S., Babb, L., Alkuraya, F.S., Amberger, J., Amin, M., Austin-Tse, C., Balzotti, M., Berg, J.S., Birney, E., et al. (2022). The Gene Curation Coalition: A global effort to harmonize gene-disease evidence resources. Genet Med. 10.1016/j.gim.2022.04.017.
10. The Royal College of Pathologists of Australasia (2019). Australian Health Genetics/Genomics Survey 2017. Report of Key Findings to: Department of Health.
11. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). 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 *17*, 405-424. 10.1038/gim.2015.30.
12. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res *44*, D733-745. 10.1093/nar/gkv1189.
13. Howe, K.L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., Bhai, J., et al. (2021). Ensembl 2021. Nucleic Acids Res *49*, D884-D891. 10.1093/nar/gkaa942.
14. Wang, M., Callenberg, K.M., Dalgleish, R., Fedtsov, A., Fox, N.K., Freeman, P.J., Jacobs, K.B., Kaleta, P., McMurry, A.J., Prlic, A., et al. (2018). hgvs: A Python package for manipulating sequence variants using HGVS nomenclature: 2018 Update. Hum Mutat *39*, 1803-1813. 10.1002/humu.23615.
15. McMurry, J.A., Kohler, S., Washington, N.L., Balhoff, J.P., Borromeo, C., Brush, M., Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2016). Navigating the

Phenotype Frontier: The Monarch Initiative. Genetics *203*, 1491-1495. 10.1534/genetics.116.188870.
16. Shefchek, K.A., Harris, N.L., Gargano, M., Matentzoglu, N., Unni, D., Brush, M., Keith, D., Conlin, T., Vasilevsky, N., Zhang, X.A., et al. (2020). The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Res *48*, D704-D715. 10.1093/nar/gkz997.
17. Tavtigian, S.V., Harrison, S.M., Boucher, K.M., and Biesecker, L.G. (2020). Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines. Hum Mutat *41*, 1734-1737. 10.1002/humu.24088.