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

Emma Tudini,[1,2] James Andrews,[1,3] David M. Lawrence,[3] Sarah L. King-Smith,[1,4] Naomi Baker,[5,6] Leanne Baxter,[7] John Beilby,[8,9] Bruce Bennetts,[10,11] Victoria Beshay,[12] Michael Black,[13] Tiffany F. Boughtwood,[1,14] Kristian Brion,[7] Pak Leng Cheong,[15,16] Michael Christie,[17] John Christodoulou,[1,11,14,18] Belinda Chong,[5] Kathy Cox,[4] Mark R. Davis,[13,19] Lucas Dejong,[4] Marcel E. Dinger,[20] Kenneth D. Doig,[12,21] Evelyn Douglas,[4] Andrew Dubowsky,[4] Melissa Ellul,[4] Andrew Fellowes,[12] Katrina Fisk,[10] Cristina Fortuno,[2] Kathryn Friend,[4] Renee L. Gallagher,[7] Song Gao,[4] Emma Hackett,[10] Johanna Hadler,[4] Michael Hipwell,[22] Gladys Ho,[10,11] Georgina Hollway,[23,24] Amanda J. Hooper,[25,26] Karin S. Kassahn,[4,27] Rahul Krishnaraj,[10] Chiyan Lau,[7,28] Huong Le,[15] Huei San Leong,[12] Ben Lundie,[7] Sebastian Lunke,[5,6] Anthony Marty,[29] Mary McPhillips,[22] Lan T. Nguyen,[25] Katia Nones,[24] Kristen Palmer,[30] John V. Pearson,[31] Michael C.J. Quinn,[1,32] Lesley H. Rawlings,[4] Simon Sadedin,[5,6,14] Louisa Sanchez,[4] Andreas W. Schreiber,[3,33] Emanouil Sigalas,[17] Aygul Simsek,[4] Julien Soubrier,[4,33] Zornitza Stark,[1,5,6] Bryony A. Thompson,[17] James U,[29] Cassandra G. Vakulin,[4] Amanda V. Wells,[4] Cheryl A. Wise,[13] Rick Woods,[7] Andrew Ziolkowski,[22] Marie-Jo Brion,[1,2] Hamish S. Scott,[1,3,4,27,35] Natalie P. Thorne,[1,6,14,29,34,35] Amanda B. Spurdle,[1,2,*] and on behalf of the Shariant Consortium

Summary

Sharing genomic variant interpretations across laboratories promotes consistency in variant assertions. A landscape analysis of Australian clinical genetic-testing laboratories in 2017 identified that, despite the national-accreditation-body recommendations encouraging laboratories to submit genotypic data to clinical databases, fewer than 300 variants had been shared to the ClinVar public database. Consultations with Australian laboratories identified resource constraints limiting routine application of manual processes, consent issues, and differences in interpretation systems as barriers to sharing. This information was used to define key needs and solutions required to enable national sharing of variant interpretations. The Shariant platform, using both the GRCh37 and GRCh38 genome builds, was developed to enable ongoing sharing of variant interpretations and associated evidence between Australian clinical genetic-testing laboratories. Where possible, two-way automated sharing was implemented so that disruption to laboratory workflows would be minimized. Terms of use were developed through consultation and currently restrict access to Australian clinical genetic-testing laboratories. Shariant was designed to store and compare structured evidence, to promote and record resolution of inter-laboratory classification discrepancies, and to streamline the submission of variant assertions to ClinVar. As of December 2021, more than 14,000 largely prospectively curated variant records from 11 participating laboratories have been shared. Discrepant classifications have been identified for 11% (28/260) of variants submitted by more than one laboratory. We have demonstrated that co-design with clinical laboratories is vital to developing and implementing a national variant-interpretation sharing effort. This approach has improved inter-laboratory concordance and enabled opportunities to standardize interpretation practices.

[1]Australian Genomics, Melbourne, VIC 3052, Australia; [2]Population Health, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia; [3]Australian Cancer Research Foundation Cancer Genomics Facility, Centre for Cancer Biology, SA Pathology and University of South Australia, Adelaide, SA 5000, Australia; [4]Genetics and Molecular Pathology, SA Pathology, Adelaide, SA 5000, Australia; [5]Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Melbourne, VIC 3052, Australia; [6]University of Melbourne, Melbourne, VIC 3052, Australia; [7]Pathology Queensland, Brisbane, QLD 4006, Australia; [8]PathWest Laboratory Medicine Western Australia, Perth, WA 6009, Australia; [9]School of Biomedical Sciences, The University of Western Australia, Perth, WA 6009, Australia; [10]Sydney Genome Diagnostics, Western Sydney Genetics Program, The Children's Hospital at Westmead, Sydney, NSW 2145, Australia; [11]Disciplines of Child and Adolescent Health and Genomic Medicine, University of Sydney, Sydney, NSW 2145, Australia; [12]Department of Pathology, Peter MacCallum Cancer Centre, Melbourne, VIC 3052, Australia; [13]Department of Diagnostic Genomics, PathWest Laboratory Medicine Western Australia, Perth, WA 6009, Australia; [14]Murdoch Children's Research Institute, Melbourne, VIC 3052, Australia; [15]Department of Medical Genomics, Royal Prince Alfred Hospital, NSW Health Pathology, Sydney, NSW 2050, Australia; [16]University of Sydney, Sydney, NSW 2006, Australia; [17]Department of Pathology, Royal Melbourne Hospital, Melbourne, VIC 3050, Australia; [18]Department of Paediatrics, University of Melbourne, Melbourne, VIC 3052, Australia; [19]Centre for Medical Research, The University of Western Australia, Perth, WA 6009, Australia; [20]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia; [21]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, VIC 3052, Australia; [22]Division of Molecular Medicine, NSW Health Pathology North, Newcastle, NSW 2305, Australia; [23]Garvan Institute of Medical Research, Sydney, NSW 2010, Australia; [24]Cancer Research, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia; [25]Department of Clinical Biochemistry, PathWest Laboratory Medicine Western Australia, Fiona Stanley Hospital Network, Perth, WA 6150, Australia; [26]School of Medicine, The University of Western Australia, Perth, WA 6009, Australia; [27]Adelaide Medical School, The University of Adelaide, Adelaide, SA 5000, Australia; [28]The University of Queensland, Brisbane, QLD 4072, Australia; [29]Melbourne Genomics Health Alliance, Melbourne, VIC 3052, Australia; [30]Genomics Statewide Services, New South Wales Health Pathology, Newcastle, NSW 2300, Australia; [31]Genome Informatics, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia; [32]Genetic Health Queensland, Royal Brisbane and Women's Hospital, Brisbane, QLD 4006, Australia; [33]School of Biological Sciences, The University of Adelaide, Adelaide, SA 5005, Australia; [34]Walter and Eliza Hall Institute, Melbourne, VIC 3052, Australia
[35]These authors contributed equally
*Correspondence: amanda.spurdle@qimrberghofer.edu.au
https://doi.org/10.1016/j.ajhg.2022.10.006.

The benefits of sharing genomic variant interpretations and associated evidence across laboratories is widely recognized; it allows for improved diagnostic accuracy and patient management.[1,2] One key benefit is the promotion of consistency through the identification and resolution of variant-classification discrepancies: sharing variant interpretations resolved 62% of discrepancies across 41 laboratories submitting to ClinVar[3] and 31% of discrepancies across 12 laboratories participating in the Canadian Open Genetics Repository.[4] As a result, several professional bodies, including the American College of Medical Genetics and Genomics, have incorporated the concept of variant sharing into best-practice guidelines.[5]

Within Australia, clinical genetic-testing laboratories undertaking germline variant curation encompass both the public (usually state-based and centered around major hospitals and genetics clinics) and private sector. It is mandatory for Australian laboratories to have a policy for submitting variants to relevant clinical databases,[6] and laboratories are encouraged to submit genotypic data to clinical databases as part of their accreditation.[6,7] Despite this recommendation, in July 2017 fewer than 300 variants had been submitted to ClinVar[8] by Australian laboratories, a finding corroborated by a national survey reporting that 75% of laboratories did not submit to international databases.[9] These observations indicated that curation knowledge was siloed in individual laboratories, and the Royal College of Pathologists of Australasia identified lack of genetic-data sharing as an area of concern.[9] As a result of the state and territory-based nature of Australian healthcare, compounded by the operation of private laboratories, between-laboratory classifications for the same variant can differ. This can lead to inequities in patient counseling and management, which are particularly problematic where results differ for individuals from the same family.

Australian Genomics is a national research initiative aimed at the implementation of genomic medicine into routine clinical practice.[10] As part of a specific project tasked to standardize variant interpretation, representatives of Australian clinical genetic-testing laboratories were consulted iteratively so that barriers to variant sharing in general could be identified and so that key requirements for sharing variant interpretations nationally could be defined. This consultation drove the development of Shariant: a controlled-access platform designed to simplify sharing of variant interpretations and associated evidence for Australian laboratories while minimizing disruption to their current workflows. Here we describe the evolution and implementation of Shariant and present initial data demonstrating its benefit for diagnostic accuracy.

Consultation

Landscape analysis

Australian clinical testing laboratories were surveyed from November 2016 to February 2017 so that current genetic-testing practices could be mapped (see supplemental material and methods). Responses (full or partial) were received and collated for 22/30 laboratories conducting genetic testing at the time of the survey. Twenty reported conducting germline testing in some capacity, and only five performed exome and/or genome sequencing. Most responding laboratories (81%) used the American College of Medical Genetics and Genomics and Association for Molecular Pathology (ACMG/AMP) guidelines,[11] either as published or with modifications, for classification of germline variants. Somatic testing was conducted by 11 responding laboratories at the time of the survey, and there was no clear consensus on interpretation guidelines for somatic variants.

This evaluation of variant-interpretation practices across Australian clinical genetic-testing laboratories identified areas that had already achieved consensus (e.g., use of ACMG/AMP guidelines for germline variant classification) or would benefit from consultation and standardization (e.g., lack of standards to interpret somatic variants), and it thus informed design of a national variant-sharing platform.

Workshop: Variant-interpretation sharing

A workshop was then conducted at the Human Genetics Society of Australasia conference in August 2017. Representatives from 15 Australian laboratories discussed their willingness to share variant interpretations with the ClinVar database, identified software being used to store variant interpretations, and discussed the potential for variant interpretation sharing.

Consultation with laboratories identified that most respondents were willing to share variant interpretations but wished to seek advice on ethico-legal considerations from their organization before providing commitment to share. The key barriers to sharing between laboratories, and additional incentives that would increase the willingness to share, are shown in Tables 1 and 2, respectively. Solutions to these barriers, and inclusion of incentives, formed the basis for requirements for a national variant-interpretation sharing platform.

Evaluation of tools and platform selection

Nine existing tools (names not disclosed), including commercial and non-commercial as well as national and international, underwent preliminary evaluation so that the potential to share variant interpretations and associated evidence could be assessed (see supplemental material and methods). The preferred tool was selected on the basis of pre-existing functionality against the evaluation framework (Table S1), costs required for software development, technical advantages and limitations, and licensing models. VariantGrid, a tool developed at the Centre for Cancer Biology (an SA Pathology and University of South Australia Alliance), was chosen as the basis for development of the national variant-interpretation sharing solution now known as Shariant.

Table 1. Barriers to sharing and solutions implemented by Shariant

Barrier to sharing	Solution
Resources	
There is no time to manually prepare data for upload to existing databases.	automated connection to laboratory interpretation system
There is limited bioinformatic expertise.	Shariant developer to support integration with laboratory interpretation systems; automated mapping and data transformation of exportable evidence
Consent	
What information can be shared and with whom?	controlled-access platform; laboratories decide on extent of (clinical) data to be shared
Other	
Interpretation tools differ between laboratories and can change over time.	sharing agnostic to interpretation system/s; flexibility in connection solutions; work with interpretation system vendors to improve connection
"Just another (static) database to check."	"real-time" connection from laboratory interpretation system allows viewing variants submitted by other laboratories nationally

Table 2. Incentives to share and solutions implemented by Shariant

Incentives to share	Solution
storage of sufficient evidence to allow review and re-use of existing curations	submission of structured evidence against ACMG/AMP guidelines
identification and resolution of classification discrepancies prior to international sharing	discrepancy-resolution tooling
streamlined submission to ClinVar	automated formatting to ClinVar specifications and submission via the ClinVar submission API

Shariant implementation and features

Shariant operations were approved by the Melbourne Health Human Research Ethics Committee (HREC/16/MH/251) and QIMR Berghofer Medical Research Institute Human Research Ethics Committee (P3447).

Formal evaluation of VariantGrid identified key functionalities that required development or improvement to meet the requirements of Shariant as informed by laboratory consultations: controlled access; extension of the application programming interface (API) to accept structured evidence; ability to accept and liftover variants between GRCh37 and GRCh38; and identification and resolution of classification discrepancies. These functionalities were developed over nine months. In parallel with this initial development, a detailed "terms of use" and security overview were created to address questions and concerns (see supplemental material and methods).

An overview of the main features of the Shariant platform is provided in Figure 1.

Controlled access

Shariant access is currently restricted to Australian laboratories conducting clinically accredited genetic testing (i.e., compliant with the National Association of Testing Authorities) who have signed the Shariant terms of use. Access is controlled via integration with standard international tooling for access management (Keycloak), and hosting is on the Amazon Web Services Sydney node. Amendments are in progress to extend access to New Zealand laboratories, given that they also follow the National Pathology Accreditation Advisory Council accreditation guidelines.[6]

Automated two-way sharing

So that concerns around laboratory resourcing requirements to share variant interpretations would be addressed, Shariant was designed to integrate with the interpretation system already in use by each laboratory via an API where feasible (details provided in Table S2). Further, design allows for a two-way interaction: (1) automated submission of germline variant interpretations and associated evidence to Shariant; (2) importation of data stored in Shariant for viewing in the laboratory interpretation system. Resources were dedicated to work with commercial vendors when required and to maintain connection as laboratories change interpretation systems over time. If a laboratory's interpretation system is unable to be connected via an API, a system-formatted export may be uploaded to the Shariant web portal, and transformation of data can be performed automatically on the Shariant end (supplemental material and methods). Additionally, a configurable system-formatted export (e.g., VCF, JSON) containing interpretations from other participating laboratories can be downloaded from the Shariant webpage for import into a laboratory's interpretation system. That is, laboratories that incorporate an export of other laboratories' variant interpretations from Shariant into their laboratory workflow will be able to see existing classifications along with interpretation evidence.

Sharing of structured variant-interpretation evidence

As indicated by the initial survey, the ACMG/AMP guidelines[11] are the most common criteria used for germline variant interpretation in Australia. Therefore, interpretation evidence was structured around these guidelines, and the flexibility to accept additional evidence fields was incorporated (version accessed 12th May 2022 included in Table S3). A minimal set of fields, including the genome build, a variant representation (usually a coding DNA Human Genome Variation Society (HGVS) expression), condition for which a variant was interpreted, zygosity, and classification/clinical significance, was deemed mandatory. Code was designed to scan for PubMed identifiers in free text, and this information is included as structured evidence with author, title, and abstract.

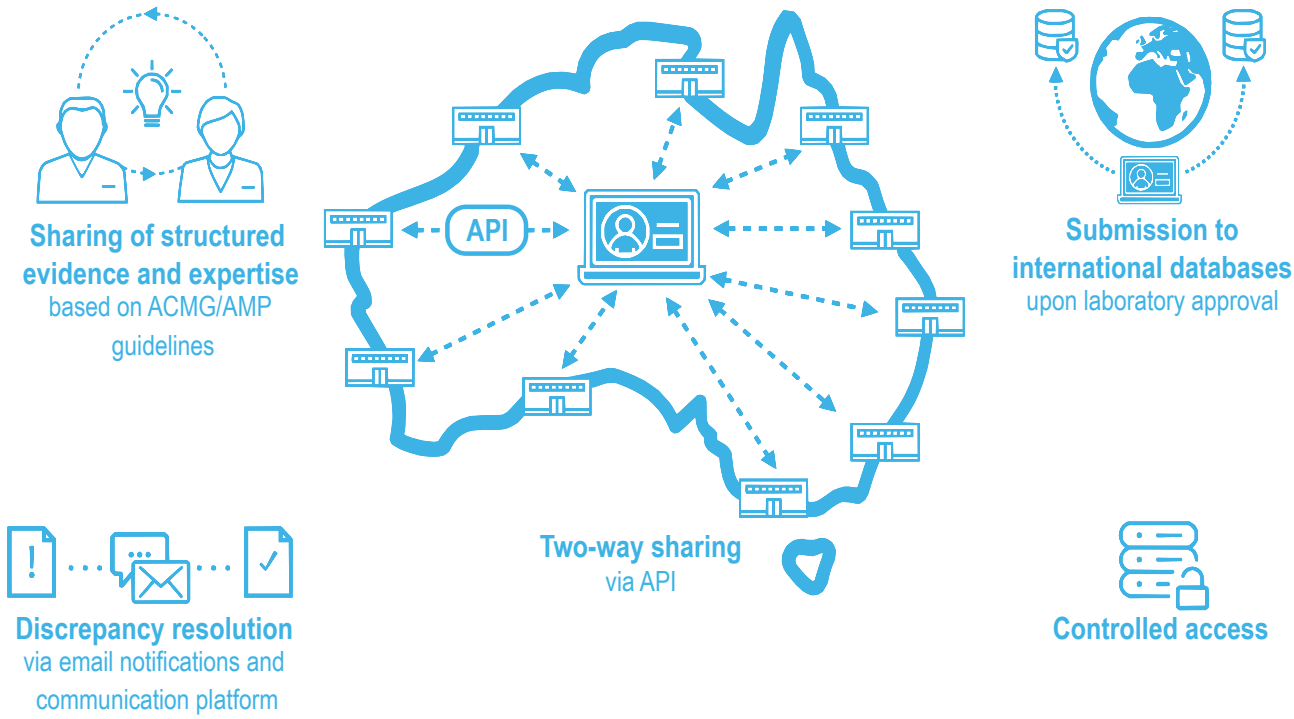


Figure 1. Overview of Shariant features
Main features of Shariant include two-way sharing via an application programming interface (API); sharing of structured evidence and expertise against guidelines from the American College of Medical Genetics and Genomics and Association for Molecular Pathology (ACMG/AMP);[11] discrepancy resolution; submission to international databases, including ClinVar;[8] and controlled access.

Variant normalization and liftover between genome builds

Accurate aggregation and connection of variants across differing variant representations and genome builds is vital for comparison of variant interpretations between laboratories. The submitted variant is matched to a genomic coordinate in the submitted genome build (both RefSeq[12] and Ensembl[13] transcripts are supported) and lifted over to the alternative genome build via the ClinGen Allele Registry[14] and then the National Center for Biotechnology Information (NCBI) Genome Remapping Service if the former does not work. All variants are then mapped to alleles that span both GRCh37 and GRCh38. If there are changes in the variant representation, including transcript version change, right alignment, and reference-base difference, a flag is raised, and the record is not exported until the Shariant team or submitting laboratory has validated the change (see supplemental material and methods; Figure S1).

Discrepancy identification and resolution

As part of another Australian Genomics project, clinical genetic-testing laboratories provided feedback on a process for resolving between-laboratory medically significant classification discrepancies, previously defined[15] as a difference between three classification tiers: pathogenic or likely pathogenic; variant of uncertain significance (VUS); and likely benign or benign. In brief, the project was tasked with analyzing the existing variant reclassification processes across Australian laboratories to inform development of consensus recommendations (H.S.S., unpublished data). Review and discussion of several historical examples of reclassification helped identify areas for improvement. These included the need to overcome logistical barriers to data sharing and communication and the implementation of an agreed-upon process for coordinating variant reinterpretation and notification. The Shariant discrepancy-identification and -resolution functionality was designed to address technical barriers related to these aspects.

Identification of a medically significant classification discrepancy within Shariant triggers an email notification to user/s from each laboratory involved. The notification provides a link allowing comparison of structured evidence between the variant records (Figure 2), and this comparison page points to a dedicated platform that facilitates and records communication within Shariant to resolve the discrepancy. Structuring of evidence has been vital for comparison of discrepant (and concordant) variant interpretations across laboratories. Communication regarding a discrepancy can be seen by any laboratory contributing to Shariant. The implementation of the discrepancy-resolution process has evolved over time, in consultation with the laboratories, highlighting the need for flexibility to address user needs.

Automated submission to ClinVar and other international databases

Streamlined submission to international databases, predominantly ClinVar,[8] is encouraged but remains as an "opt-in" for laboratories contributing to Shariant. The submission

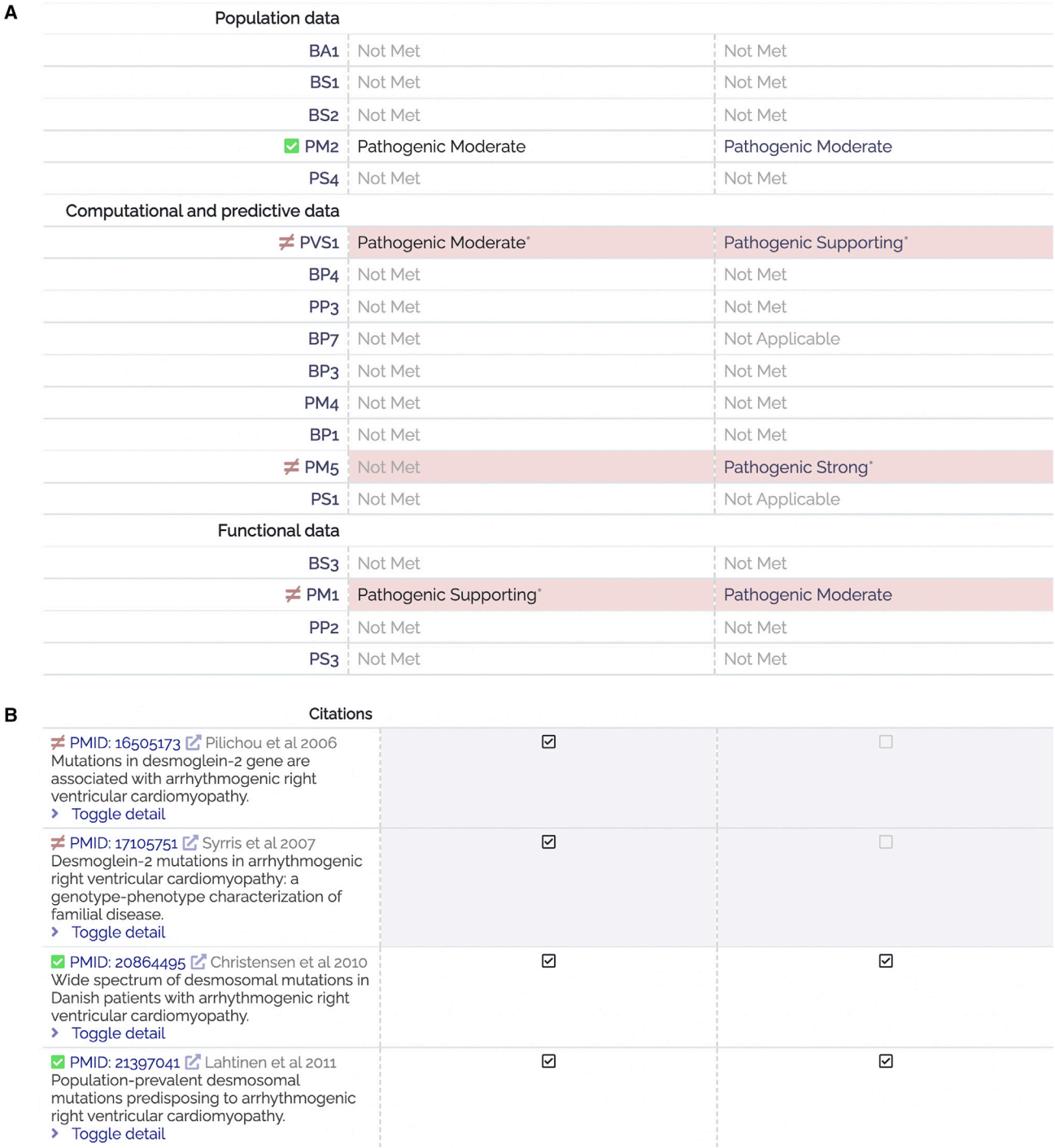


Figure 2. Comparison of structured evidence between variant records

(A) Example comparison of ACMG/AMP[11] criteria applied by two laboratories. Red highlighting indicates a difference in the application of ACMG/AMP codes.

(B) Example comparison of citations referenced for the same variant. Ticks indicate the referencing of a citation by a particular laboratory, and gray highlighting indicates a difference between laboratories.

process uses the recently released ClinVar submission API and recognizes each laboratory individually (i.e., not as a consensus Shariant classification). Shariant mandatory fields overlap with most fields required by ClinVar, with the exception of a standard-condition term (e.g., Mondo Disease Ontology [Mondo],[16] Online Mendelian Inheritance in Man [OMIM],[17] and Human Phenotype Ontology [HPO][18]). Approximately 70% of records submitted to Shariant did not meet this requirement at the time of development; thus, a new functionality was added to match free text conditions to a Mondo ontology identifier. Resources including PanelApp Australia,[19,20] the Gene Curation Coalition[21] and

Mondo[16] were used to assist in matching of a Mondo identifier relevant at the gene level (where possible), as requested by the participating laboratories (see supplemental material and methods; Figures S3 and S4).

Shariant national rollout

In the onboarding of Australian clinical genetic-testing laboratories, initial prioritization was given to public laboratories because of the nature of Australian Genomics funding. The interest of the laboratory, ease of connection to the laboratory's interpretation system, and genetic-testing output of the laboratory were also considered.

Onboarding of laboratories required three stages.

(1) Initial engagement of the laboratory, including discussion of Shariant purpose and features, laboratory willingness to participate, and distribution of Shariant documents, such as the terms of use and security overview.
(2) Organizational sign-off on terms of use.

Legal sign-off by each organization (sometimes pertinent to multiple laboratories) was a significant hurdle for onboarding. Time to sign-off varied greatly, from one week to more than 1.5 years. Where possible, the connection to Shariant was developed and tested prior to the sign-off on the terms of use to minimize delays to data sharing.

(3) Integration of laboratory interpretation system with Shariant.

The complexity and time involved in integration (via API connection or web portal upload) was highly dependent on the laboratory and interpretation system in use. For example, resource needs were greatly reduced during integration of a laboratory with the same commercial interpretation system as an already-connected laboratory.

Automated data transformation from a laboratory's system-formatted export to the structured Shariant format was deemed necessary to permit regular uploads of variant interpretations and associated evidence without additional manual input (supplemental material and methods). Initial automation of this process greatly influenced time to connection, requiring laboratory-specific customization even if the same interpretation system was in use by multiple laboratories. These issues involved parsing of free text fields so that structured information not otherwise provided as a set field in the export (e.g., condition under curation) could be extracted, parsing free text so that laboratory-sensitive information (e.g., internal communication) was excluded, and mapping fields back to standard ACMG/AMP guidelines to resolve differences in interpretation schema. Automation also needed to be flexible to account for changes in the export format and interpretation schema over time.

Shariant statistics

As of December 2021, eleven laboratories across four states in Australia had shared their interpretation of largely prospectively curated variants. A total of 14,045 variant records had been shared across 2,070 genes; seven laboratories routinely submitted multiple records per variant to capture the number of patients. After multiple variant records for relevant laboratories were collapsed, there were 11,655 variants (denoted hereafter as "unique variants per laboratory"). Approximately half of these variants (53%; 6,137 variants) were classified as a VUS, 13% were designated as likely pathogenic, and 27% as pathogenic.

Variant submission schedules varied, depending on the laboratory interpretation system and resourcing, including regular (real-time, weekly, monthly) or more sporadic larger submissions (Figure 3A). It is anticipated that the frequency of variant submission will increase over time as facilitated by implementation of automated data transformation. Ingestion of Shariant data by laboratory interpretation tools generally occurs in parallel to submissions.

It was evident that the number of variants with submissions from multiple laboratories has increased over time, consistent with more per-laboratory submissions and more new laboratories contributing to Shariant (Figures 3B and 3C). Of the 11,377 unique variants across laboratories, only 2% (260) were submitted by multiple laboratories during the time period assessed for this analysis. Of these, 232 (89%) were in complete agreement or concordant within a confidence level.

Discrepancy identification and resolution

The Shariant platform is designed to identify any medically significant classification discrepancies within or between laboratories at the time of any new upload of laboratory information. Thus, discrepancies might arise upon the first submission of a variant from a laboratory or upon resubmission of an updated classification from a laboratory. To date, these automated classification checks have identified 28 unique variants as discrepant between laboratories; this discrepancy detection rate amounts to 11% among unique variants submitted by multiple laboratories (28/260). Shariant processes have assisted resolution of 12/28 (43%) identified discrepancies (Table 3).

Eight discrepancies were resolved without inter-laboratory discussion, mainly utilizing the comparison of structured evidence to review ACMG/AMP criteria, in an average of 17 days (range: 0 to 106 days, median 6 days). Four discrepancies required inter-laboratory discussion in consideration of additional evidence from one laboratory, increasing the average number of days to resolution to 147 (range: 57 to 289, median 121 days). Most (8/10) of the resolved discrepancies resulted in reclassification to a category of more certainty.

One variant has been discussed extensively but could not be resolved via the resolution process (continued

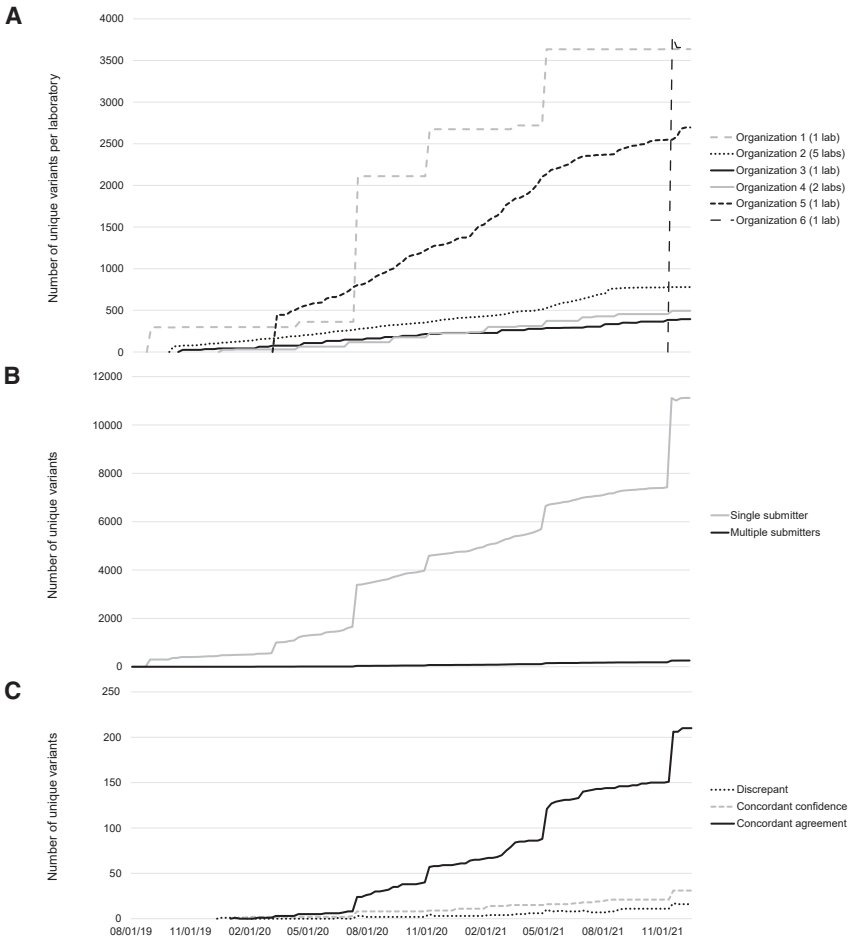


Figure 3. Variants shared via Shariant over time

(A) Submission and sharing of unique variants per laboratory, presented as totals per organization over time (mm-dd-yy). Frequency of upload is dependent on the interpretation system in use by a laboratory (lab). The spike in submissions in November 2021 was due to the recent onboarding of organization 6. Organization 6 has only submitted one large batch of records, and this batch included historical data.

(B) Overall submissions of unique variants to Shariant over time and breakdown of variants submitted by one laboratory compared to variants submitted by multiple laboratories.

(C) Unique variants contributed by more than one laboratory with breakdown of comparison category: concordant-agreement variants, concordant-confidence variants, and discrepant variants. As of December 2021, concordant-agreement variants (n = 211) included those that were pathogenic, n = 145 (69%); likely pathogenic, n = 19 (9%); VUS, n = 37 (18%); likely benign, n = 9 (4%); and benign, n = 1 (0.5%). Concordant-confidence variants (n = 31) included those that were pathogenic or likely pathogenic, n = 30 (97%); and benign or likely benign, n = 1 (3%). Discrepant variants are detailed in Table 3.

discrepancy). Of the remaining 15 discrepancies awaiting resolution, five were identified from recent submissions (<30 days in discrepancy), whereas the others are long-standing (range: 103–311 days). These observations have recently led to the establishment of a formal process utilizing existing clinical-laboratory-directed multidisciplinary team meetings to address unresolved Shariant-detected discrepancies in variant classification. It should be noted that multiple discrepancies involved variants that could be considered reduced-penetrance or risk alleles, an observation reported previously.[3]

The overall initial discrepancy rate observed here is somewhat lower than what has been previously reported (17%–22%),[3,4,22] perhaps reflecting that most variants shared have been identified prospectively, for which curations should use the most up-to-date information from the public domain (e.g., ACMG/AMP guidelines). It might also reflect differences between studies in the relative proportions of variants assigned to classification tiers at baseline. For example, the majority of medically significant discrepancies reported by Mighton et al.[4] involved classification tiers likely benign or benign versus VUS, and 60% of the variants in this dataset were likely benign or benign at baseline; in comparison, only 7% of Shariant submissions were likely benign or benign at baseline. We would also hypothesize that, particularly for laboratories that routinely ingest data from Shariant, discrepancies might be minimized because evidence from other laboratories will now be visible at the time of curation.

Variant reclassification

In addition to discrepancy resolution, laboratories internally reclassified variants over the course of their involvement in Shariant. If we consider only unique variants per laboratory, 102 variants were internally reclassified over two years (Figure 4). This equates to 1.3% of total variants from the 10 laboratories with repeated submissions and is comparable to the proportion reported for ClinVar over four years (2.1%)[23] and lower than that observed for studies where re-evaluation spanned a longer timeframe.[24,25] Lower rates of internal reclassification than in previous studies[24–26] could be attributed to a combination of factors, including that most variants were prospectively evaluated in the past two years according to the ACMG/AMP guidelines and with access to publicly accessible population databases (e.g., gnomAD).

In agreement with Harrison et al.,[23] most reclassifications (91; 89%) moved to a category of more certainty (Figure 4). These included 72 variant upgrades: VUS to likely pathogenic (51, 50%); likely pathogenic to

Table 3. Medically significant between-laboratory classification discrepancies in order of date identified

Gene symbol	Variant (coding HGVS)	Variant (protein HGVS)	Disease area	Classifications[a]	Number of days discrepant (as of 15th December, 2021)	Resolved classification	Upgrade/ downgrade	Inter-laboratory discussion required	Reason for resolution
PROC	NM_000312.3: c.565C>T	NP_000303.1: p.(Arg189Trp)	MONDO: 0005570 hematologic disorder	P vs. VUS	6	LP	upgrade	no	review of ACMG/ AMP criteria
CASR	NM_001178065.1: c.1212C>T	NP_001171536.1: p.(Val404=)	MONDO: 0005066 metabolic disease; MONDO: 0005151 endocrine system disorder	VUS vs. LB	0	LB	downgrade	no	review of ACMG/ AMP criteria
COL4A3	NM_000091.4: c.4421T>C	NP_000082.2: p.(Leu1474Pro)	MONDO: 0005240 kidney disorder	VUS vs. LB	7	VUS	upgrade	no	review of ACMG/ AMP criteria
ABCA4	NM_000350.2: c. 5693G>A	NP_000341.2: p.(Arg1898His)	MONDO: 0005283 retinal disorder	VUS vs. LB	289	LB	downgrade	yes	new functional evidence
FECH	NM_001012515.2: c. 333-48T>C; NM_000140.4: c.315-48T>C	NP_001012533.1: p.?; NP_000 131.2: p.?	MONDO: 0005066 metabolic disease	P vs. VUS	106	P	upgrade	no	reduced penetrance variant (only pathogenic if found with another loss-of-function variant)
TGFBR1	NM_004612.3: c.1468A>G	NP_004603.1: p.(Lys490Glu)	MONDO: 0004995 cardiovascular disorder	LP vs. VUSA	183	LP	upgrade	yes	additional segregation evidence provided by one laboratory
POLG	NM_002693.2: c.2890C>T	NP_002684.1: p.(Arg964Cys)	MONDO: 0004069 inborn mitochondrial metabolism disorder	LP vs. VUSA	385	continued discrepancy	N/A	yes	no resolution—reviewed extensively by mitochondrial experts; awaiting ClinGen expert panel review
POLG	NM_002693.2: c.2209G>C	NP_002684.1: p.(Gly737Arg)	MONDO: 0004069 inborn mitochondrial metabolism disorder	P vs. VUS	0	one classification withdrawn	N/A	no	out-of-date classification uploaded
MYH7	NM_000257.2: c.532G>A	NP_000248.2: p.(Gly178Arg)	MONDO: 0004995 cardiovascular disorder	LP vs. VUS	311				
ABCA4	NM_000350.2: c.71G>A	NP_000341.2: p.(Arg24His)	MONDO: 0005283 retinal disorder	LP vs. VUS	270				
NOD2	NM_022162.2: c.566C>T	NP_071445.1: p.(Thr189Met)	MONDO: 0005046 immune system disorder	VUS vs. LB	249				
DSG2	NM_001943.3: c.3036_3037insG	NP_001934.2: p.(Tyr1013Va lfsTer25)	MONDO: 0004995 cardiovascular disorder	LP vs. VUSA	58	VUS	downgrade	yes	additional evidence from one laboratory and ClinVar
F8	NM_000132.3: c.1094A>G	NP_000123.1: p.(Tyr365Cys)	MONDO: 0005570 hematologic disorder	LP vs. VUS	222				

(*Continued on next page*)

Table 3. ***Continued***

Gene symbol	Variant (coding HGVS)	Variant (protein HGVS)	Disease area	Classifications[a]	Number of days discrepant (as of 15th December, 2021)	Resolved classification	Upgrade/ downgrade	Inter-laboratory discussion required	Reason for resolution
HFE	NM_000410.3: c.187C>G	NP_000401.1: p.(His63Asp)	MONDO: 0005066 metabolic disease	P vs. VUS	222				
NIPBL	NM_133433.3: c.1178A>G	NP_597677.2: p.(Asn393Ser)	MONDO: 0019042 multiple congenital anomalies/dysmorphic syndrome	VUS vs. LB	57	LB	downgrade	yes	additional evidence from one laboratory and external public source
CASR	NM_000388.3: c.190A>G	NP_000379.2: p.(Asn64Asp)	MONDO: 0005066 metabolic disease; MONDO: 0005151 endocrine system disorder	LP vs. VUS	0	LP	upgrade	no	review of ACMG/ AMP criteria
CFTR	NM_000492.3: c.2657+2_ 2657+3insA	NP_000483.3: p.?	MONDO: 0005087 respiratory system disorder	P vs. VUSA	166				
USH2A	NM_206933.2: c.4106C>T	NP_996816.2: p.(Ser1369Leu)	MONDO: 0005283 retinal disorder	P vs. VUS	120				
AHI1	NM_001134831.1: c.2988delA	NP_001128303.1: p.(Val997SerfsTer20)	MONDO: 0019042 multiple congenital anomalies/dysmorphic syndrome	LP vs. VUS	109				
TNFRSF13B	NM_012452.2: c.310T>C	NP_036584.1: p.(Cys104Arg)	MONDO: 0005046 immune system disorder	P vs. VUS	109				
KMT2D	NM_003482.3: c.12862C>T	NP_003473.3: p.(Arg4288Trp)	MONDO: 0019042 multiple congenital anomalies/dysmorphic syndrome	VUS vs. LB	103				
CHEK2	NM_007194.4: c.470T>C	NP_009125.1: p.(Ile157Thr)	MONDO: 0015356 hereditary neoplastic syndrome	LP vs. VUS	6	one classification withdrawn	N/A	no	variant reviewed as a risk factor and therefore doesn't align with ACMG/AMP classification criteria
CDH1	NM_004360.4: c.387+5G>A	NP_004351.1: p.?	MONDO: 0015356 hereditary neoplastic syndrome	VUS vs. LB	26				
POLE	NM_006231.4: c.2090C>G	NP_006222.2: p.(Pro697Arg)	MONDO: 0015356 hereditary neoplastic syndrome	VUS vs. LB	26				

(*Continued on next page*)

Table 3. *Continued*

Gene symbol	Variant (coding HGVS)	Variant (protein HGVS)	Disease area	Classifications[a]	Number of days discrepant (as of 15th December, 2021)	Resolved classification	Upgrade/ downgrade	Inter-laboratory discussion required	Reason for resolution
MSH6	NM_000179.2: c.3226C>T	NP_000170.1: p.(Arg1076Cys)	MONDO: 0015356 hereditary neoplastic syndrome	LP vs. VUS	9	LP	upgrade	no	out-of-date classification uploaded
POLE	NM_006231.3: c.4523G>A	NP_006222.2: p.(Arg1508His)	MONDO: 0015356 hereditary neoplastic syndrome	VUS vs. LB	23				
BRCA2	NM_000059.3: c.10076A>G	NP_000050.2: p.(Glu3359Gly)	MONDO: 0015356 hereditary neoplastic syndrome	VUS vs. LB	23				
MEFV	NM_000243.2: c.910G>A	NP_000234.1: p.(Gly304Arg)	MONDO: 0005046 immune system disorder	VUS vs. LB	11				

Medically significant classification discrepancies have been defined by Harrison et al. 2017.[15]
[a]Pathogenic (P), likely pathogenic (LP), variant of uncertain significance (VUS), likely benign (LB), benign (B). VUSA is a variant of uncertain significance with suspected high clinical significance.

pathogenic (18, 18%); and VUS to pathogenic (3, 3%). There were 19 downgrades from VUS, all to likely benign (19%). Reclassifications to VUS included four variants that were originally likely benign (4%) and five that were originally likely pathogenic (5%). This contrasts with previous studies, where most reclassifications over time were downgrades from VUS to likely benign or benign.[23–26]

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

All laboratories submitting to Shariant include structured evidence that is mapped against the original ACMG/AMP guidelines.[11] These interpretations represent a valuable resource allowing for a nationwide approach to examining the impact of new recommendations for using ACMG/AMP guidelines and prioritizing additional data collation.

Using a points-based approach[27] and modifying PM2 to PM2_Supporting as per recent ClinGen Sequence Variant Interpretation Working Group recommendations[28] (see supplemental material and methods), the classification of 579 unique variants per laboratory applying PM2 (15%) would be changed: 219 (38%) would undergo a medically significant change (likely pathogenic to VUS), 52 (9%) variants would change from VUS to likely benign, and 308 (53%) variants would change confidence from pathogenic to likely pathogenic. This analysis demonstrates the extent to which application of the recommended revised weight to historical records might impact re-reporting, counseling of patients, and ultimately, changes in clinical management.

Functional assay evidence is only assigned a weight for approximately 10% of unique variants per laboratory. If these records are excluded, genes that had the most VUSs in the hereditary cancer context included *BRCA2*, *PALB2*, and *MSH6* (Figure 5A). Across all other diseases, genes with the most VUSs included *PKD1*, *TTN*, and *USH2A* (Figure 5B). Although multiple factors (e.g., length of gene, type of gene) might be contributing to these high VUS numbers, the data provide a snapshot of the minimum number of families for which variant reclassification could be assisted by the incorporation of functional assay data. Such data could be used as a means of prioritizing future functional assays and/or adapting Shariant to include look-ups to large-scale functional datasets such as MAVEdb.[29]

Contribution internationally

The new ClinVar submission API is now being used for Shariant-assisted automated submission. The initial submission of 385 variants from one laboratory more than doubled the number of contributions to ClinVar from Australian clinical genetic-testing laboratories, as compared to contributions prior to project initiation. On the basis of all Shariant data as of December 2021, we estimated that future Australian laboratory submissions to ClinVar would add variant interpretation evidence for

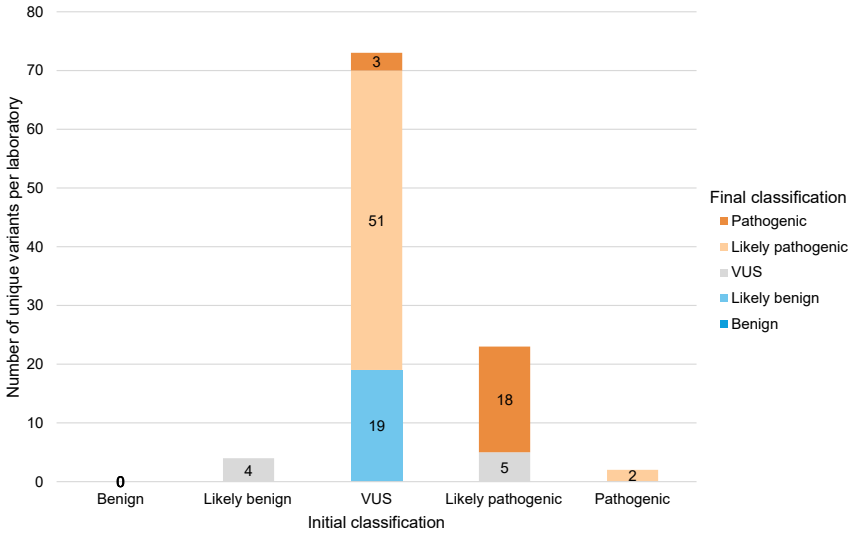


Figure 4. Reclassification of variants in Shariant

Number of unique variants that were reclassified from August 2019 to December 2021. Initial classification is represented on the x axis, and the number of each pathogenicity for reclassified variants is displayed via the colored bars. Most variants that changed classification were initially VUS (n = 73; 72%), and the majority of these resolved to likely pathogenic or pathogenic (n = 54, 74% of this subgroup).

over 11,000 unique variants, of which 4,000 would be novel. Moreover, this will enable routine deposition of variant data to ClinVar, as evidenced by the increased number of variant-classification submissions to ClinVar since manuscript submission; 2,146 submissions from nine laboratories as of August 2022.

Comparison to previous studies

Compared to previous national sharing efforts,[4,30] Shariant has implemented several additional functionalities. These include the ability to automate transformation of data from the laboratories (as opposed to manual manipulation by the laboratory or a database coordinator prior to each upload); system integration between the laboratory interpretation system and a central database (i.e. Shariant) via an API; ability to accept multiple genome builds and facilitate liftover; and a process for identifying classification discrepancies and documenting their resolution.

Future directions

Shariant development priorities are driven collaboratively by the Shariant User Group, which meets monthly and includes all participating laboratories and the Shariant team. Resolution of variants classified as a VUS by multiple laboratories has been flagged as a priority by the User Group and has the potential to be expanded to the resolution of confidence differences. The current implementation of Shariant has focused on sharing germline variant interpretations. However, once we hold further consultations aimed at extending its design and capabilities and implement resultant modifications, we anticipate that Shariant will additionally capture and share somatic variant interpretations in the future.

Conclusion

Implementation of an automated sharing process has facilitated repeated uploads of variant interpretations and has the potential to enable widespread national real-time sharing. Sharing of structured evidence has allowed for the national comparison of interpretation processes and can be used to identify areas for future standardization. By offering a custom solution, Shariant maintains flexibility and agility. Issues around timelines to uptake have been largely governance related.

We have demonstrated that co-design with clinical genetic-testing laboratories is vital to the development and implementation of a national effort to share variant interpretations, and we provide insight into approaches that might be adapted by other national initiatives. However, it should be expected that resources are required for maintaining operations, adapting to new or updated laboratory interpretation systems, and incorporating changes in laboratory-specific or general classification guidelines. Shariant, like any other national sharing project, requires long-term and eventually sustainable funding.

Consortia

Shariant Consortium members: Lauren Akesson, Richard Allcock, Katie Ashton, Damon A. Bell, Anna Brown, Michael Buckley, John R. Burnett, Linda Burrows, Alicia Byrne, Eva Chan, Corrina Cliffe, Roderick Clifton-Bligh, Susan Dooley, Miriam Fanjul Fernandez, Elizabeth Farnsworth, Thuong Ha, Denae Henry, Duncan Holds, Katherine Holman, Matilda Jackson, Sinlay Kang, Catherine Luxford, Sam McManus, Rachael Mehrtens, Cliff Meldrum, David Mossman, Sarah-Jane Pantaleo, Dean Phelan, Electra Pontikinas, Anja Ravine, Tony Roscioli, Rodney Scott, Keryn Simons, and Oliver Vanwageningen.

Data and code availability

The Shariant (VariantGrid) code generated during this study (J.A. and D.M.L.) is freely available for research use under business source license 1.1 on GitHub (https://github.com/SACGF/variantgrid). The datasets supporting the current study have not been deposited in a public repository as a result of restrictions on the sharing of non-aggregate data outlined in the Shariant

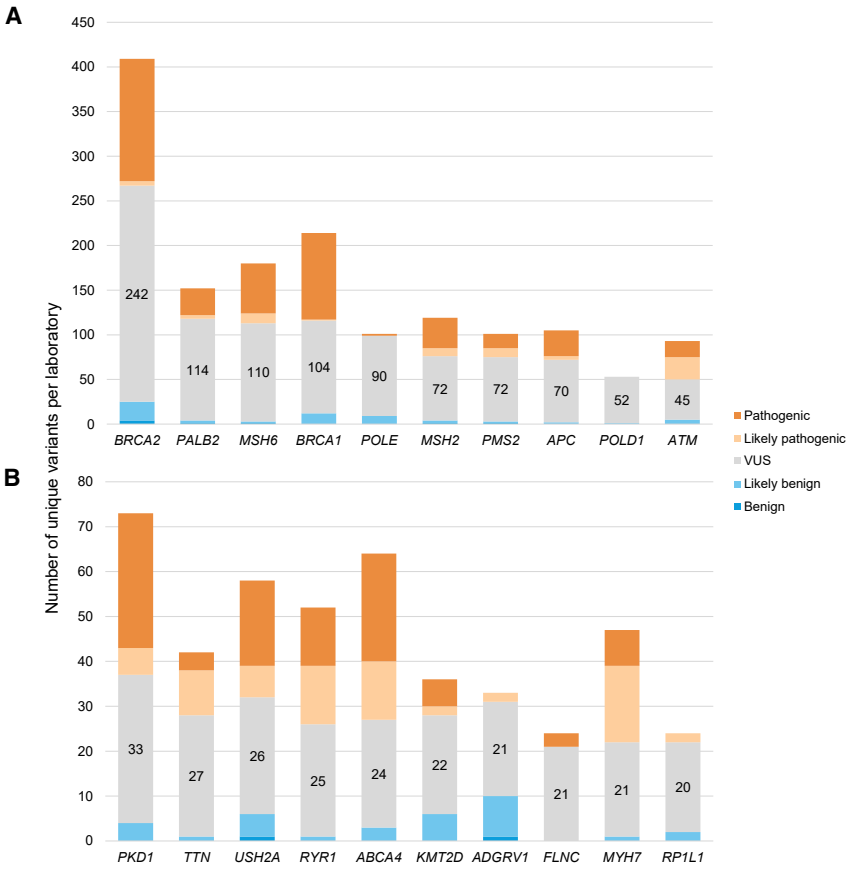


Figure 5. Genes with the most variants of uncertain significance across the context of hereditary cancer versus other diseases

Results are shown for hereditary cancer (A) and other diseases (B). Variant records that were not matched to a variant or had no ACMG/AMP[11] criteria assigned were removed, and only the most recently curated record was included if more than one variant record for a variant had been submitted by the same laboratory. All records that had a strength assigned for BS3 or PS3 were also excluded. Results for the hereditary cancer genes were driven by one laboratory.

terms of use. Variant interpretations and associated evidence will be deposited in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) at submitter request.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.10.006.

Acknowledgments

The authors would like to thank Penny Glenn, who led the development of the Shariant terms of use and multiple rounds of minor amendments in consultation with laboratory legal teams. We also thank the laboratories who participated in the landscape analysis in 2017 and all individuals who attended the Human Genetics Society of Australasia conference workshop. We thank John Rowell and Valentine Hyland for their prior contributions to engagement of Pathology Queensland and the GenoVic team for efforts relating to initial integration of GenoVic laboratories. We wish to acknowledge Agilent Technologies, for its contribution of technical resources and ongoing support to achieve interoperability of Shariant with Agilent's variant tertiary analysis system. Shariant Australia (including funding for E.T. and J.A.) is supported by Australian Genomics (NHMRC grants GNT1113531 and GNT2000001). A.B.S was supported by NHMRC fellowship funding (APP1061779, APP177524). The research conducted at the Murdoch Children's Research Institute was supported by the Victorian Government's Operational Infrastructure Support Program. The Chair in Genomic Medicine awarded to J.C. is generously supported by The Royal Children's Hospital Foundation. Hosting of Shariant on Amazon Web Services between October 2018 and September 2019 was supported by an Amazon Web Services Cloud Credit for Research grant.

Declaration of interests

The authors declare no competing interests.

Web resources

Amazon Web Services, https://aws.amazon.com/
ClinVar Submission API, https://www.ncbi.nlm.nih.gov/clinvar/docs/api_http/
Keycloak, https://www.keycloak.org/
NCBI Genome Remapping Service, www.ncbi.nlm.nih.gov/genome/tools/remap
Shariant evidence keys, https://shariant.org.au/classification/evidence_keys
VariantGrid, https://github.com/SACGF/variantgrid

References

1. Wright, C.F., Ware, J.S., Lucassen, A.M., Hall, A., Middleton, A., Rahman, N., Ellard, S., and Firth, H.V. (2019). Genomic variant sharing: a position statement. Wellcome Open Res. *4*, 22. https://doi.org/10.12688/wellcomeopenres.15090.2.

2. Raza, S., and Hall, A. (2017). Genomic medicine and data sharing. Br. Med. Bull. *123*, 35–45. https://doi.org/10.1093/bmb/ldx024.
3. Harrison, S.M., Dolinksy, J.S., Chen, W., Collins, C.D., Das, S., Deignan, J.L., Garber, K.B., Garcia, J., Jarinova, O., Knight Johnson, A.E., et al. (2018). Scaling resolution of variant classification differences in ClinVar between 41 clinical laboratories through an outlier approach. Hum. Mutat. *39*, 1641–1649. https://doi.org/10.1002/humu.23643.
4. Mighton, C., Smith, A.C., Mayers, J., Tomaszewski, R., Taylor, S., Hume, S., Agatep, R., Spriggs, E., Feilotter, H.E., Semenuk, L., et al. (2021). Data sharing to improve concordance in variant interpretation across laboratories: results from the Canadian Open Genetics Repository. J. Med. Genet. *59*, 571–578. https://doi.org/10.1136/jmedgenet-2021-107738.
5. ACMG Board Of Directors (2017). Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. Genet. Med. *19*, 721–722. https://doi.org/10.1038/gim.2016.196.
6. National Pathology Accreditation Advisory Council (2017). Requirements for Human Medical Genome Testing Utilising Massively Parallel Sequencing Technologies, First edition.
7. The Royal College of Pathologists of Australasia (2014). Massively Parallel Sequencing Implementation Guidelines.
8. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. *46*, D1062–D1067. https://doi.org/10.1093/nar/gkx1153.
9. The Royal College of Pathologists of Australasia (2019). Australian Health Genetics/Genomics Survey 2017. Report of Key Findings to: Department of Health.
10. Stark, Z., Boughtwood, T., Phillips, P., Christodoulou, J., Hansen, D.P., Braithwaite, J., Newson, A.J., Gaff, C.L., Sinclair, A.H., and North, K.N. (2019). Australian Genomics: a federated model for integrating genomics into healthcare. Am. J. Hum. Genet. *105*, 7–14. https://doi.org/10.1016/j.ajhg.2019.06.003.
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. https://doi.org/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–D745. https://doi.org/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. https://doi.org/10.1093/nar/gkaa942.
14. 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. https://doi.org/10.1002/humu.23637.
15. Harrison, S.M., Dolinsky, J.S., Knight Johnson, A.E., Pesaran, T., Azzariti, D.R., Bale, S., Chao, E.C., Das, S., Vincent, L., and Rehm, H.L. (2017). Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. Genet. Med. *19*, 1096–1104. https://doi.org/10.1038/gim.2017.14.
16. Mungall, C.J., McMurry, J.A., Köhler, 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. https://doi.org/10.1093/nar/gkw1128.
17. McKusick, V.A. (1998). Mendelian Inheritance in Man: A Catalog of Human Genes and Genetic Disorders (JHU Press).
18. Köhler, 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. https://doi.org/10.1093/nar/gkaa1043.
19. 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. https://doi.org/10.1016/j.ajhg.2021.06.020.
20. 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. https://doi.org/10.1038/s41588-019-0528-2.
21. 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. *24*, 1732–1742. https://doi.org/10.1016/j.gim.2022.04.017.
22. Furqan, A., Arscott, P., Girolami, F., Cirino, A.L., Michels, M., Day, S.M., Olivotto, I., Ho, C.Y., Ashley, E., Green, E.M., et al. (2017). Care in specialized centers and data sharing increase agreement in hypertrophic cardiomyopathy genetic test interpretation. Circ. Cardiovasc. Genet. *10*, e001700. https://doi.org/10.1161/CIRCGENETICS.116.001700.
23. Harrison, S.M., and Rehm, H.L. (2019). Is 'likely pathogenic' really 90% likely? Reclassification data in ClinVar. Genome Med. *11*, 72. https://doi.org/10.1186/s13073-019-0688-9.
24. Mighton, C., Charames, G.S., Wang, M., Zakoor, K.R., Wong, A., Shickh, S., Watkins, N., Lebo, M.S., Bombard, Y., and Lerner-Ellis, J. (2019). Variant classification changes over time in BRCA1 and BRCA2. Genet. Med. *21*, 2248–2254. https://doi.org/10.1038/s41436-019-0493-2.
25. Slavin, T.P., Van Tongeren, L.R., Behrendt, C.E., Solomon, I., Rybak, C., Nehoray, B., Kuzmich, L., Niell-Swiller, M., Blazer, K.R., Tao, S., et al. (2018). Prospective study of cancer genetic variants: variation in rate of reclassification by ancestry. J. Natl. Cancer Inst. *110*, 1059–1066. https://doi.org/10.1093/jnci/djy027.
26. Kast, K., Wimberger, P., and Arnold, N. (2018). Changes in classification of genetic variants in BRCA1 and BRCA2. Arch. Gynecol. Obstet. *297*, 279–280. https://doi.org/10.1007/s00404-017-4631-2.

27. 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. https://doi.org/10.1002/humu.24088.
28. ClinGen Sequence Variant Interpretation Working Group (2020). ClinGen Sequence Variant Interpretation Recommendation for PM2 - Version 1.0.
29. Esposito, D., Weile, J., Shendure, J., Starita, L.M., Papenfuss, A.T., Roth, F.P., Fowler, D.M., and Rubin, A.F. (2019). MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biol. *20*, 223. https://doi.org/10.1186/s13059-019-1845-6.
30. Fokkema, I.F.A.C., van der Velde, K.J., Slofstra, M.K., Ruivenkamp, C.A.L., Vogel, M.J., Pfundt, R., Blok, M.J., Lekanne Deprez, R.H., Waisfisz, Q., Abbott, K.M., et al. (2019). Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data. Hum. Mutat. *40*, 2230–2238. https://doi.org/10.1002/humu.23896.

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

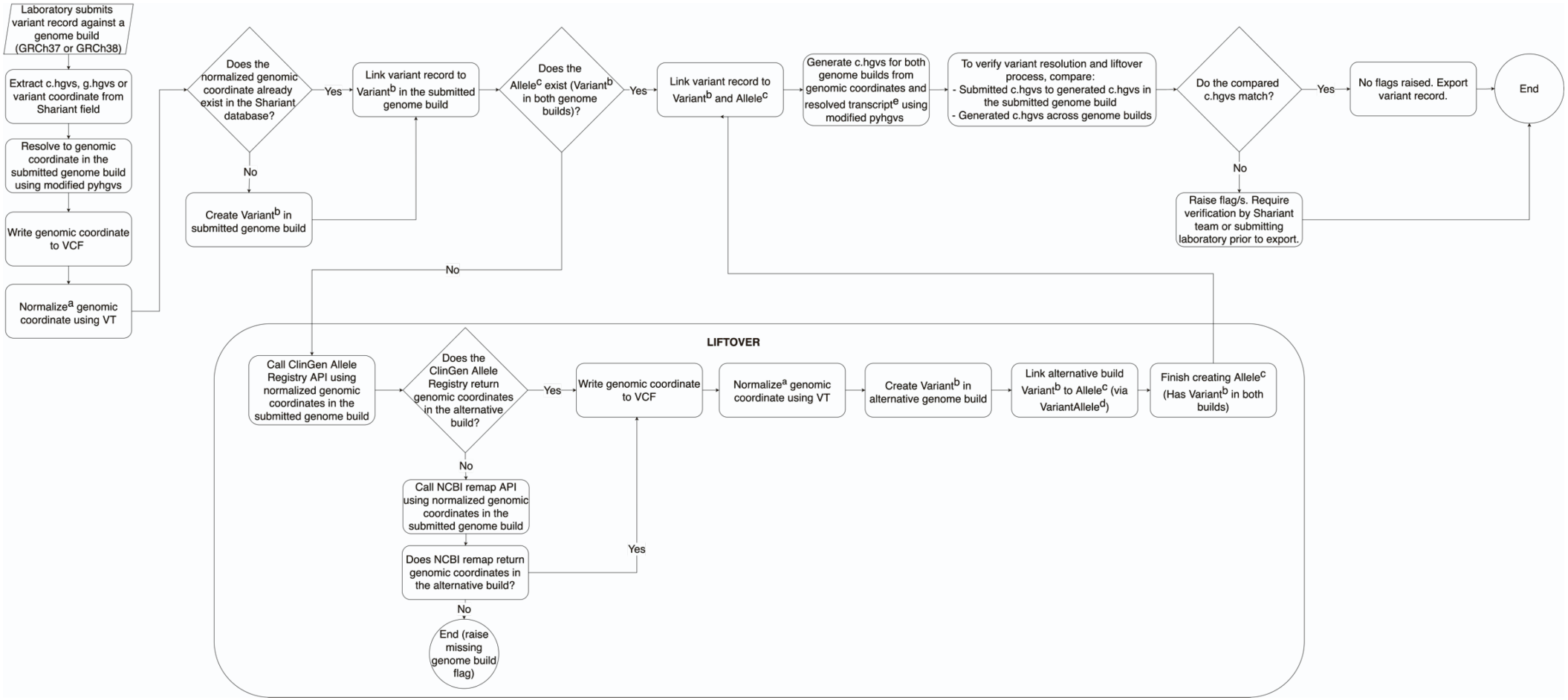


Figure S1. Variant resolution and liftover process.
Representation of process to allow for accurate aggregation and connection of variants across differing variant representations (e.g., coding DNA HGVS nomenclature (c.hgvs), genomic HGVS nomenclature (g.hgvs)) and genome builds GRCh37 and GRCh38. Liftover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

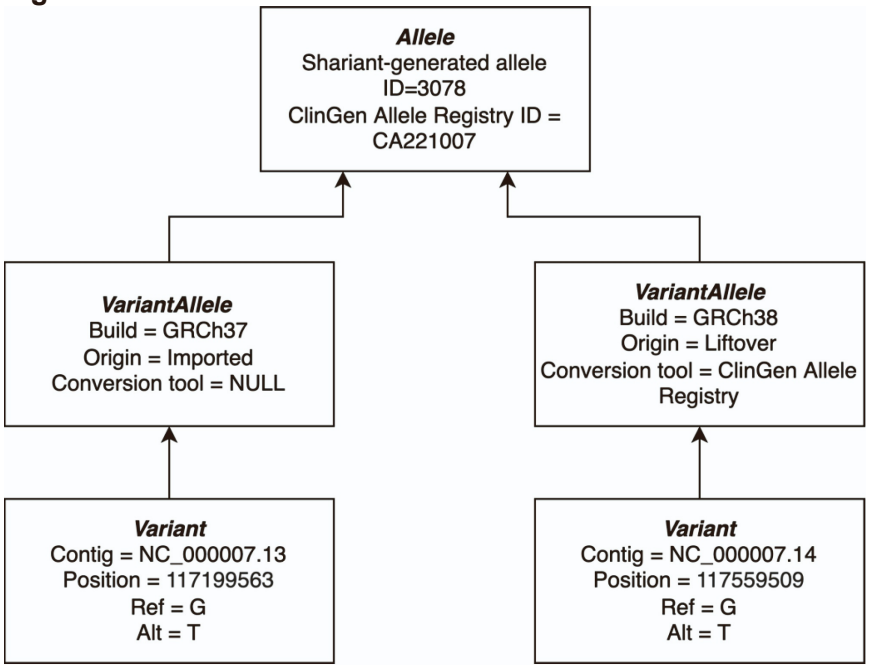


Figure S2. Database representation of example variant record submitted against genome build GRCh37.
Variant is defined as a normalized genomic coordinate against a specific genome build. VariantAllele denotes a database model used to link together Variant and Allele. It also stores the liftover method 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

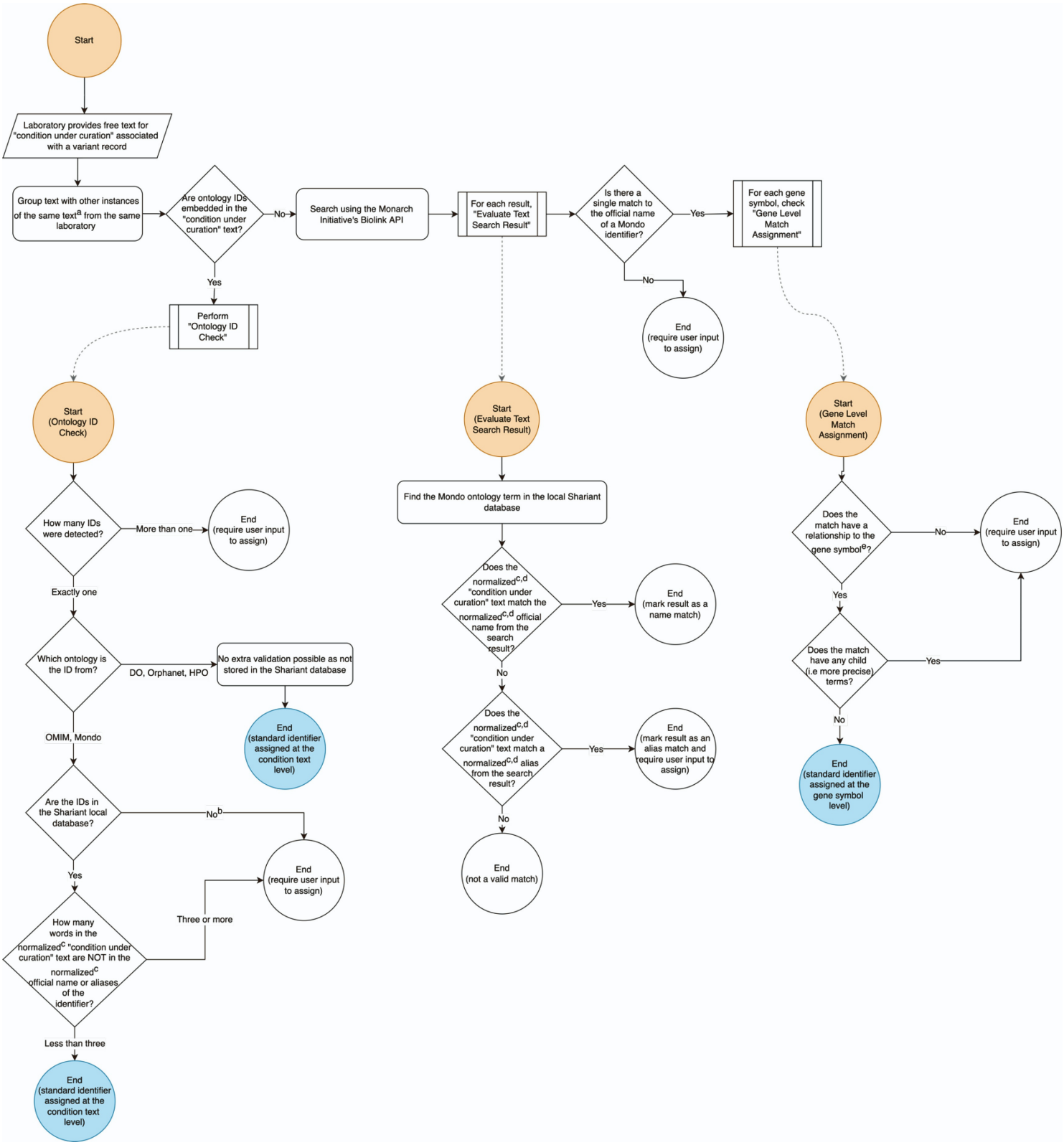


Figure S3. Condition text matching process for the automatic assignment of ontology identifiers (IDs) to "condition under curation" text. Note: This flowchart does not describe the process for suggestion of ontology identifiers that require user input. Ontologies supported include Monarch Disease Ontology (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.

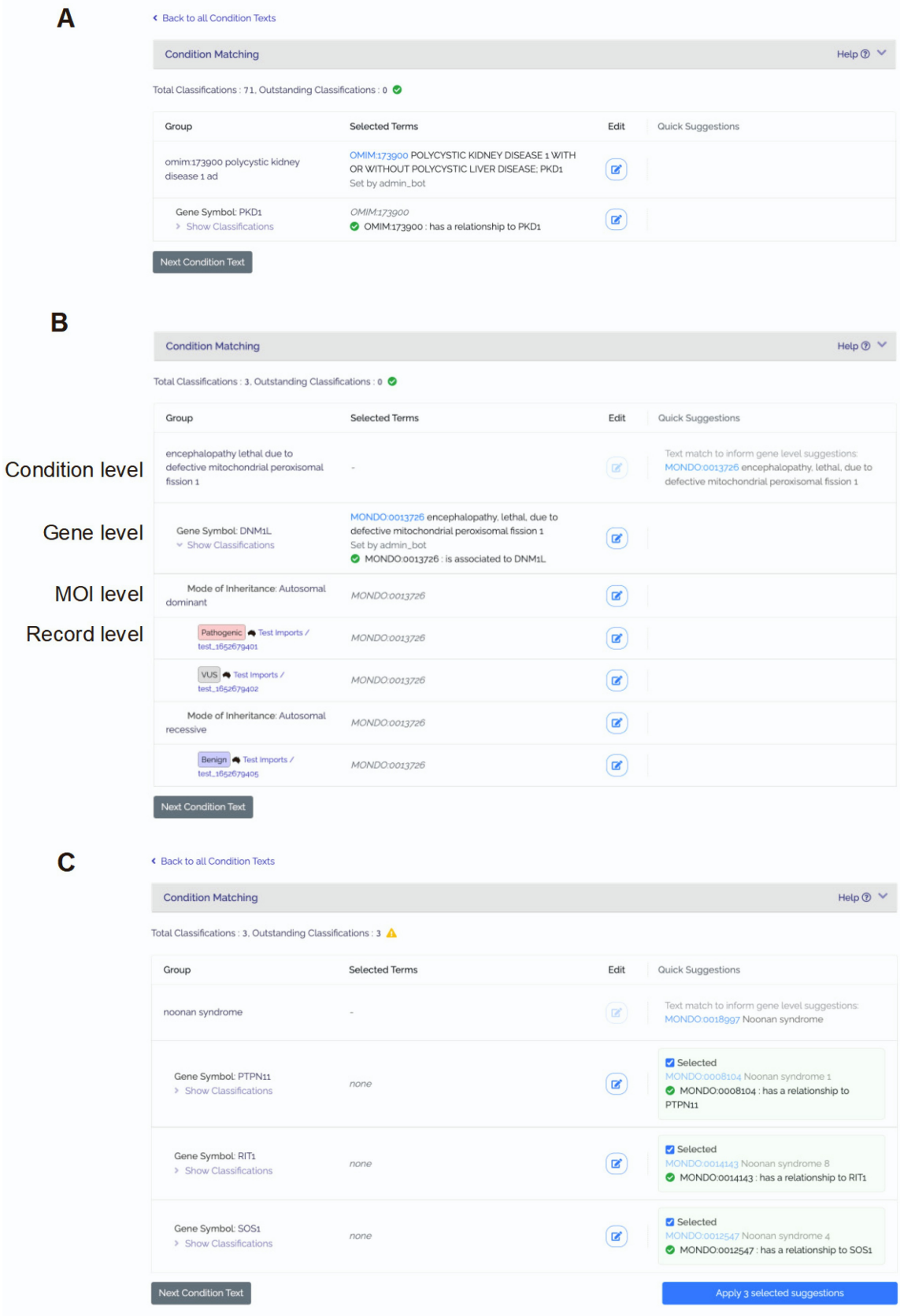
[b]Examples of ontology identifiers that do not exist in the Shariant database include OMIM gene/locus numbers. This function also acts as a means of identifying typographical errors.

[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

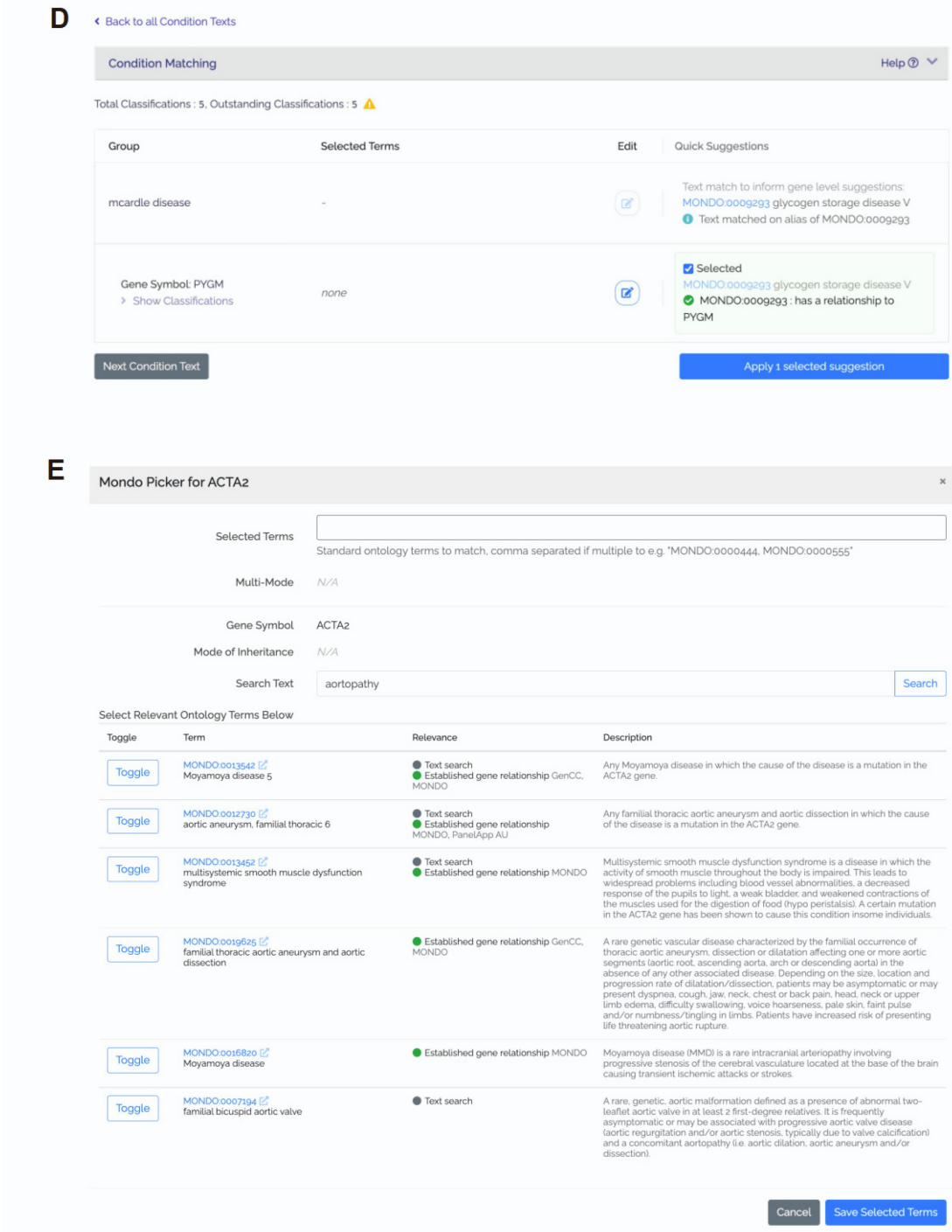


Figure S4. Shariant condition text matching interface.
(A) Automated matching due to submission of a standard ontology. (B) Automated matching of free text condition to a Mondo Disease Ontology (Mondo)[3] identifier with designation of assignment hierarchy at the condition, gene, mode of inheritance (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.