LocalVar: a local variant collection manager to asynchronously detect synonyms, HGVS expression changes, and variant interpretation changes from ClinVar

Michael Watkins[1], Wendy Kohlmann[2], Therese Berry[2], Neetha Sama[2], Cathryn Koptiuch[2], Shawn Rynearson[1], and Karen Eilbeck[1]

[1]University of Utah Health
[2]Huntsman Cancer Institute Cancer Hospital

June 2, 2021

Abstract

While there are several public repositories of biological sequence variation data and associated annotations, there is little open-source tooling designed specifically for the upkeep of local collections of variant data. Many clinics curate and maintain such local collections and are burdened by frequent changes in the representation of those variants and evolving interpretations of clinical significance. A dictionary of genetic variants from the Huntsman Cancer Institute was analyzed over a period of two years and used to inform the development of LocalVar. This tool is institution-agnostic and uses publicly available ClinVar files to provide the following functionality: auto-complete search bar to pre-empt duplicate entries; single or bulk new variant record entry; auto-detection and merge suggestions for duplicate variant records; auto-detection and merge suggestions for variant records with HGVS expressions that are marked as synonyms in ClinVar; asynchronous suggestion of HGVS expression or variant interpretation updates; history tracking of additions, merges, updates, or other manual edits made to variant records; and the easy export of the collection (.csv), edit history (.json), or HGVS synonym bins (.json).

LocalVar: a local variant collection manager to asynchronously detect synonyms, HGVS expression changes, and variant interpretation changes from ClinVar

Michael Watkins[1], Wendy Kohlmann[2], Therese Berry[2], Neetha Sama[2], Cathryn Koptiuch[2], Shawn Rynearson[3], and Karen Eilbeck[1]

Author Summary

*Corresponding author: michael.watkins.8@gmail.com

[1]The Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, Utah 84108;

[2]Huntsman Cancer Institute, University of Utah, 2000 Circle of Hope Drive, Salt Lake City, Utah 84112;

[3]The Department of Human Genetics, University of Utah, 15 N 2030 E, Salt Lake City, UT 84112

Abstract

While there are several public repositories of biological sequence variation data and associated annotations, there is little open-source tooling designed specifically for the upkeep of local collections of variant data. Many clinics curate and maintain such local collections and are burdened by frequent changes in the representation of those variants and evolving interpretations of clinical significance. A dictionary of genetic variants from

the Huntsman Cancer Institute was analyzed over a period of two years and used to inform the development of LocalVar. This tool is institution-agnostic and uses publicly available ClinVar files to provide the following functionality: auto-complete search bar to pre-empt duplicate entries; single or bulk new variant record entry; auto-detection and merge suggestions for duplicate variant records; auto-detection and merge suggestions for variant records with HGVS expressions that are marked as synonyms in ClinVar; asynchronous suggestion of HGVS expression or variant interpretation updates; history tracking of additions, merges, updates, or other manual edits made to variant records; and the easy export of the collection (.csv), edit history (.json), or HGVS synonym bins (.json).

Keywords

suggestions, update, conflicts, bins, software

Introduction

Huntsman Cancer Institute (HCI) of the University of Utah is the official Cancer Center of Utah. HCI serves as the only and most proximal National Cancer Institute-designated Cancer Center for much of the vast Mountain West region (Utah, Idaho, Montana, Nevada, and Wyoming), which encompasses 17% of the landmass of the continental United States. HCI provides patient and prevention education from three community clinics in the surrounding area, and six affiliate hospitals in neighboring states[1]. With a long history of germline genetics research, HCI maintains a large amount of variant data.

While storing variant data can be a simple matter, keeping that stored collection up-to-date is not. Mutation nomenclature versioning, reference sequences, and scientific discovery all contribute to the ever-changing nature of variant data. One fairly representative example can be seen in the evolving nature of the *ARSA* variants. A number of these variants are disease-causing and can lead to Metachromatic Leukodystrophy (MLD). A study by Cesani, et al. in 2015 focused on an updated recommendation in mutation nomenclature guidelines to ascribe the A of the first ATG translational initiation codon as nucleotide +1. They found that most *ARSA* variants had been reported before this recommendation was made and were described based on the processed mature protein, which differs from the translated protein in six nucleotides at the 5'-terminus of the cDNA sequence and two amino acids at the N-terminus. One example they cited was that the common autosomal dominant-causing variant that had been historically referred to as 1277C>T (Pro426Leu), should be named c.1283C>T (p.Pro428Leu) [2]. This is one of the thousands of instances that highlight the arduous burden of keeping variant data current.

One popular software tool to support research on variant data is the Leiden Open Variation Database (LOVD) [3]. The database was initially created in 2004, with LOVD 2.0 being released in 2007 and the current 3.0 version being released in 2012. This service provides local access to an immense amount of gene/disease annotations with the data all linked to a centralized online LOVD database. While local records can be added using submission templates, there is limited flexibility for custom columns, no support for merging records, no synonymous HGVS record detection, and limited history tracking of edits made to the data.

There are also a growing number of public genetic variant databases, which include insightful annotations.*ClinVar* [4], *ClinGen* [5], *dbSNP* [6], *dbVar* [7], *HGMD* [8], *gnomAD* [9], *CIViC* [10],*OMIM* [11], and *COSMIC* [12]all fill particular niches in the field of medical genetics and have independent funding and partnerships. There have also been REST-based tools created to provide mapping services across the different identifiers used by these databases. The *ClinGen Allele Registry* [13] and*MyVariant.info* [14] are two predominant tools that offer this service. These tools allow interested parties to query the "current" knowledge about a particular variant and benefit from synonym detection and a rich result drawn from across the several databases. These public services are widely used by the research community but are single-query based and not designed for the longitudinal maintenance of local variant collections.

The Variation Representation Specification (VRS) is being developed by the Global Alliance for Genomics and Health (GA4GH)[15]. Currently in its second major version, 1.1, VRS makes several contributions: a terminology and information model that ensures the precise computational definitions for biological concepts

in fields, semantics, objects, and object relationships; a machine-readable schema to enable language-agnostic tests for ensuring compliance to the information models; various conventions that promote reliable data sharing, such as fully justified allele normalization; globally unique computed identifiers that allow data providers and consumers to computationally generate consistent, globally unique identifiers for variation without a central authority; and a python implementation that demonstrates the proper implementation of the specification and facilitates the translation of existing variant representations into VRS. The addition of VRS identifiers to any variant collection will prove to be critical as the variant community moves toward a more computationally stringent system of linking variant knowledge and exchanging variant data across institutions.

This study had two main objectives. The first was to analyze the HCI variant dictionary and discover trends that might indicate needs not currently filled. The second was to create an open-source and institution-agnostic tool to address these needs and otherwise facilitate the management of variant collections.

Methods

Objective 1 - HCI variant dictionary analysis

For over two decades, HCI has maintained a variant dictionary for tracking variants detected through research or clinical genetic analysis. Each clinical variant is assigned one classification. Generally, this is the original classification assigned by the clinical lab that performs the testing. Detected variants whose classifications are in conflict with ClinVar are reviewed by a team of genetic counselors, physicians, and variant specialists who decide upon the final classification to be stored with the variant in the dictionary. These classifications may be revisited by the variant review team if a clinical lab sends an update indicating a variant has been reclassified based on their classification criteria, or if a variant already in the dictionary is identified in a new patient and the clinical lab has assigned a different classification.

Snapshots of the variant dictionary were pulled at three time points that span two years (2019-03-19, 2020-01-29, and 2021-03-02) and used to gauge how the dictionary changed over the two years. While there were several fields included for each entry in the variant dictionary, the coding DNA HGVS expression and the variant interpretation fields were the only ones used for this study. The following metrics were the focus of the analysis:

- The number of entries added to the variant dictionary between each time point.
- The number of added entries that have identical HGVS expressions to existing entries (duplicates).
- The number of entries whose "interpretation" was changed between each time point.

The variant dictionary includes unique identifiers for each record and these identifiers were used to compare HGVS expressions and interpretations across the three snapshots. ClinVar was used to establish a point of reference for the rate of HGVS expression and interpretation changes for variants in the variant dictionary. As part of the ClinVar tab_delimited archive, there are monthly releases of a*variant_summary_-YYYY-MM.txt.gz* file[16]. This file contains several fields. The "AlleleID" (identifier assigned by ClinVar to each simple allele), "Name" (contains the coding DNA HGVS expression), and "ClinicalSignificance" (clinical interpretation of the variant) fields were the only ones used for this study. Three ClinVar variant summary files were downloaded (variant_summary_2019-03.txt, variant_summary_2020-01.txt, variant_summary_2021-03.txt) that corresponded to the dates of the three annual snapshots. These files were parsed and the coding DNA HGVS expressions were compared to those in the variant dictionary. If a match was found, the AlleleID was mapped to the variant's unique identifier in the variant dictionary. These associations were then tracked across the three variant dictionary snapshots and three ClinVar variant summary files (spanning two years) to determine the rate by which ClinVar updated the same HGVS expressions as those stored by HCI in the variant dictionary. To determine changes in clinical interpretation, the "ClinicalSignificance" field of the variant summary file was compared to the interpretation assigned by the HCI variant review team. Generally, the ClinVar variant summary file assigns a single classification for each alleleID. This means that although multiple conflicting interpretations are common in ClinVar, the variant summary file usually presents a single authoritative interpretation for each alleleID. The exception to this is in the

most recent variant summary file used (variant_summary_2021-03.txt) where 32 alleleIDs with conflicting interpretations were found. However, this had no effect on our analysis since the HGVS expressions for those 32 alleleIDs are not present in the HCI variant dictionary.

Objective 2 - LocalVar tool creation

The justification for the functionality of LocalVar is included in the results of the first study objective. This section details the development of this functionality and other design choices for the LocalVar tool. A Flask web application architecture was chosen to allow integration of the various GA4GH Python modules created to provide VRS identifier generation functionality. Because this is typically less versatile than a JavaScript web application, an optional Dockerfile was also included to assist in environment setup.

The tool was designed to be initialized with the upload of a .csv file representing a given institution's variant collection. This format was chosen because it is a common export type of SQL databases, Excel, and other storage services that may be currently used by institutions that maintain variant collections. The tool was created to be institution-agnostic, so a prompt is provided for users to select the names of the column containing the HGVS expressions and the column containing the variant interpretations. This allows the tool to then automatically create VRS identifiers for each entry in the file and place them in a newly added "VRS" column. The merits of VRS identifiers and a justification for their inclusion are provided in the discussion section of this study. The VRS identifiers are generated using HGVS to VRS Allele identifier python code that is provided by the GA4GH vrs-python repository on GitHub[17].

An integral part of the LocalVar functionality is the creation of "HGVS bins" that are subsequently used to detect synonyms and interpretation conflicts/updates. These bins are stored as a JSON object with the HGVS expression as the key and the unique collection ID and variant interpretation as values, as shown in Figure 1. If that HGVS expression is present in ClinVar, additional values are added to the bin as shown in Figure 2. These include the variationID associated with that HGVS expression, synonymous HGVS expressions stored in ClinVar (each associated with the same variationID), and the ClinVar interpretation for that variant. These bins are asynchronously updated by LocalVar with each monthly release of the ClinVar variant summary file (*variant_summary_YYYY-MM.txt.gz* , part of the ClinVar tab_delimited archive) which is where these ClinVar added data are taken from.

Edits can come from the acceptance of any of the suggestions mentioned above, from the addition (single or bulk) or deletion of variant entries, or be made manually to specific variant record fields. All of these edits made to variant records in the collection are time-stamped and stored by LocalVar using a JSON object with the unique collection identifier as key and edit events stored as values.

Results

Objective 1 - HCI variant dictionary analysis

The HCI variant dictionary was analyzed in order to inform the design process of LocalVar. Figure 3 shows that a small percentage of the total variants (1.2% in 2019, 1.1% in 2020, and 1.1% in 2021) were duplicate entries. These findings show that even with high-quality data, there can be a need for tooling to detect the small percentage of duplicates in variant collections. Of the variants in each snapshot of the variant dictionary, 37.8% in 2019, 35.4% in 2020, and 35.4% in 2021 were also found in the ClinVar variant summary files. These lower percentages are due to the fact that affiliate labs of HCI often do not publicly release new variants to ClinVar. Of those that are also found in ClinVar, a few had interpretation conflicts (6.5% in 2019, 5.7% in 2020, 4.6% in 2021). These conflicts were unchanged across the three snapshots. A majority of these conflicts (94%) were not clinically significant ("Benign/Likely benign" vs "Uncertain significance"). A small percentage (5.3%) could be clinically significant ("Pathogenic/Likely pathogenic" vs "Uncertain significance"). Only one (0.2%) of these conflicts was clinically significant ("Pathogenic" vs "Benign"). The severity of each conflict type (clinically significant, could be clinically significant, or clinically significant) is drawn from the ClinVar Miner study where all conflicts in ClinVar are categorized and analyzed[18]. While ClinVar is a widely-used tool containing informative variant interpretations, HCI does not consider such

public knowledge as authoritative. However, the ability to detect and track these conflicts can assist variant review teams (such as the one at HCI), by providing a synthesis of published data via ClinVar that can help to inform their decision.

Figure 4 shows that there were very few changes to the HGVS expressions for the variants in the variant dictionary over the two-year recording period. From 2019–2020, there were 11 total HGVS expression changes in the variant dictionary. This is compared to 700 ClinVar changes to the HGVS expressions of variants found in the variant dictionary. Upon closer inspection, it was found that 695 of those ClinVar changes (99.3%) were transcript updates. From 2020–2021, the number of HGVS expression changes within the variant dictionary rose to 190, but, as was the case with ClinVar, 185 of those changes (97.4%) were transcript updates. ClinVar reported 505 HGVS expression changes over that same period and 100% of them were transcript updates. These findings highlighted the fact that transcript changes are common and may place a burden on individuals tasked with keeping variant collections up-to-date. They also showed that asynchronous updates from external sources, such as ClinVar, can provide useful synonym detection and automated upkeep of variant records.

Figure 4 also suggests that there were clinical interpretation changes in ClinVar (192 from 2019–2020, 244 from 2020–2021) that were not reflected in the HCI variant dictionary (five from 2019–2020, 40 from 2020–2021). There is wisdom in being prudent with updating changes to clinical interpretations based solely on ClinVar. A 2020 study by Xiang, et al. tracked variants interpreted as "Pathogenic" and "Likely pathogenic" by ClinVar. They found that after manual interpretation of 326 qualifying variants, 40% were downgraded to benign, likely benign, or variant of uncertain significance while only 2% were found more likely to be risk factors[19]. It would therefore be alarming to *not* find a high rate of interpretation conflicts when comparing a variant dictionary to ClinVar. However, letting users know that a change occurred, giving them access to evidence and supporting material, and giving them the option to easily update their local variant interpretation can be a useful feature in a variant collection managing tool. A summary of the tooling needs discussed above that were drawn from the analysis of the HCI variant dictionary is included in Table 1.

Objective 2 - LocalVar

LocalVar was created to address several needs associated with the longitudinal maintenance of a variant collection. A demo is available that will allow readers to explore the functionality described in this section (*http://www.watkinscv.com/app-demos/LocalVar*). Once the collection is loaded, the main page (Figure 5) shows an interactive table of the entire collection. This is enhanced with an autocomplete search bar that can pre-empt duplicate entries and a drop-down text area where new single or bulk entries can be added to the collection.

When a variant record from this collection table is clicked or searched, the user is navigated to a record details page (Figure 6). If the HGVS expression of the record is also found in ClinVar, the clinical significance from ClinVar and all associated synonymous HGVS expressions from ClinVar will also be displayed. Additionally, a custom link is provided to view these extra data on the ClinVar online portal using the variationID (ClinVar identifier stored in the HGVS bins) for that variant. In this record details page, any field of the record, except for the system-generated VRS identifier, can be directly edited. Manual changes to the HGVS expression will initiate the auto-creation of a new corresponding VRS identifier and an update of corresponding ClinVar data. Any changes made are tracked by the LocalVar edit history. This history is prominently displayed on the record details page. There is a two-step process for removing variant records. Deleted records will initially be moved into a "trash" collection, shown in Figure 7. Within this trash collection, variant records can still be restored to the main collection without any loss of data and with such an event being recorded in the edit history for that record. These records can also be permanently removed from the trash collection but only after another prompt warning the user that the removed record will not be recoverable.

The HGVS bins are also used to generate a collection of suggested data updates for the user. An example of suggested updates is shown in Figure 8 and can be navigated to using the fixed navigation sidebar. The first suggestion type, "Update Interpretation," is created for each record in the collection with an HGVS

expression that matches a ClinVar entry and has a different interpretation than what is in ClinVar. The user can choose whether to accept this suggestion (which will update the value of the interpretation column for that record) or to reject the suggestion. If rejected, the user is prompted by the tool with the option of marking the "conflicting" interpretations as synonymous. This would be suitable for instances where, for example, the variant collection used the term "Indeterminate" while ClinVar uses the term "Uncertain Significance" to refer to a variant of uncertain significance. If the user chooses to use this option, all conflicts of those two terms in the collection would be removed and the tool would store that preference for any subsequent suggestions. This allows LocalVar to "learn" how to map institution interpretations to those found in ClinVar while remaining institution-agnostic. It also reduces the number of erroneous suggestions for the user and increases their likelihood of finding significant interpretation conflicts. The second suggestion type, "Merge Duplicate," allows the user to merge records in the collection that have separate unique collection identifiers but the same HGVS expression. As shown in Figure 9, the user is given the opportunity to select which fields to carry into the newly merged record. The third suggestion type, "Merge Synonym," utilizes the HGVS bins to allow users to merge records in the collection that have HGVS expressions that are recorded as synonyms by ClinVar. The user is given the opportunity to select which HGVS expression to carry into the merged entry. Any row in the main "View Collection" table can also be selected for the option of merging two or more records into one. In order to merge multiple records into one, the user must select at least one value for each column. For all merge events, the resulting record is saved under the collection ID selected by the user and the records whose collection IDs were not selected are moved to the trash. All such events are tracked in the history of the records involved.

Another feature that was added to LocalVar is the ability to easily download three different "reports." This option is prominently displayed on the fixed sidebar and can be selected from anywhere in the tool. The first report is a .csv snapshot of the entire collection. This updates every time the collection is modified and allows users to easily capture the collection in its current state and either move to another collection managing software or share the collection with interested partners. The second report is a .json file of the entire history of the collection. This allows users to have a detailed record of every edit, accepted suggestion, variant addition or deletion, etc. The JSON format of this file allows the history to be easily searched as each edit event is tied to a specific record. The third report is a .json file of the HGVS bins used to associate the variants in the collection with those from ClinVar. The JSON format of this file also makes it easily searchable and straightforward. This is an effort to make LocalVar suggestions transparent and relatively simple to validate or otherwise audit.

Discussion

Reporting Gaps in Interpretation Updates

One limitation to our analysis of the HCI variant dictionary in relation to ClinVar is that we do not account for reporting gaps. For example, consider the gap in laboratory updates to ClinVar records. Updates occur at different frequencies (quarterly, annually, etc.) depending on the laboratory. A laboratory may have sent an updated interpretation to HCI that has not yet been propagated to ClinVar and this may have caused an interpretation conflict. Additionally, there are possible gaps from HCI between when a variant reclassification notification is received and when the variant review team can meet to discuss the update. In summary, these cross-sectional snapshots may be reflecting discrepancies that are already known and being processed.

Genes Evaluated

The HCI variant dictionary contains many cancer predisposition genes (CPGs). These are genes that can cause a moderate to high increase in risk for cancer when mutated in the germline. This made ClinVar a natural choice for external knowledge about records in the collection since ClinVar contains mostly germline variants (more than half a million). However, the number of somatic entries to ClinVar is rising (>4000 as of January 2020) and this trend will make ClinVar a more flexible knowledge source[20]. The nature of CPGs also leaves them more likely to have multiple classifications and subsequent conflicts as they are more

often the subject of expert review panels (more so than other gene types)[21]. There are also a number of nonclinical variants in the dictionary from research studies conducted by HCI. These variants are not likely to have annotations in ClinVar. Future work to mature this tool could include separate suggestion types and annotation gathering for nonclinical variants. A full list of genes (clinical and non-clinical) in the HCI variant dictionary is included as Supplementary Table 1.

Adding features

Because LocalVar is open-source and python-based, it can be modified by users and have its functionality greatly expanded. One such feature that was out of scope for this study would be to use other knowledge sources in addition to ClinVar. With the HGVS bin structure already in use, implementers would simply need to associate additional synonyms, interpretations, or other annotations via HGVS expressions. The source code is classed, heavily commented, and clearly separated into component files by task. This will assist with expanding the functionality of the tool. Other features could include support for HGVS expressions that use protein references, automated detection and exclusion of somatic variants, the inclusion of additional data points from ClinVar (date submitted, review status, number of submitters, etc.), support for different user types and edit permissions, or the option to submit new variant records to ClinVar directly from LocalVar.

Implementation considerations

The initialization time for new collections varies with the size of the .csv file being loaded (about 1.3 minutes per 1,000 lines). The main factor in increased initialization time is the addition of VRS identifiers to the collection. LocalVar was developed to use local memory rather than a database to store the variant data. Along with making the collection easily downloadable, this design decision also made it much simpler to track the collection edit history and simpler for users to add functionality to the tool. However, it does require more memory (250MB for a benchmark file of 10 columns and 1,000 lines; 600MB for a benchmark file of 10 columns and 10,000 lines).

Conclusion

With the volume of biological sequence variation data ever-rising, there is a need for lightweight and customizable tooling to facilitate the management of local collections of this variant data. An analysis of a variant collection maintained by HCI revealed a need for tooling that can manage duplicate detection and asynchronously generate suggestions for HGVS expression updates and updates to clinical significance interpretations. LocalVar was created as an institution-agnostic prototype. This proof-of-concept application can be installed locally and initialized with any comma-separated file as long as that file contains unique row identifiers and an HGVS expression and some kind of interpretation field for each record. It uses asynchronous monthly updates from ClinVar to provide update suggestions that can be accepted or declined. This tool is intended to replace the use of Excel or SQL to manage local collections of biological sequence variation.

Acknowledgments

NLM T15-LM007124 training predoctoral slot to MW. Research reported in this publication utilized the Genetic Counseling Shared Resource at Huntsman Cancer Institute at the University of Utah and was supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA042014. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Footnotes

Conflicts of Interest: nothing to report.

Data Availability Statement

The source code for the LocalVar tool is openly available in a GitHub repository at https://github.com/mwatkin8/LocalVar. No other new data were created in this study. The HCI

variant dictionary was analyzed in this study but is not publicly available due to privacy or ethical restrictions.

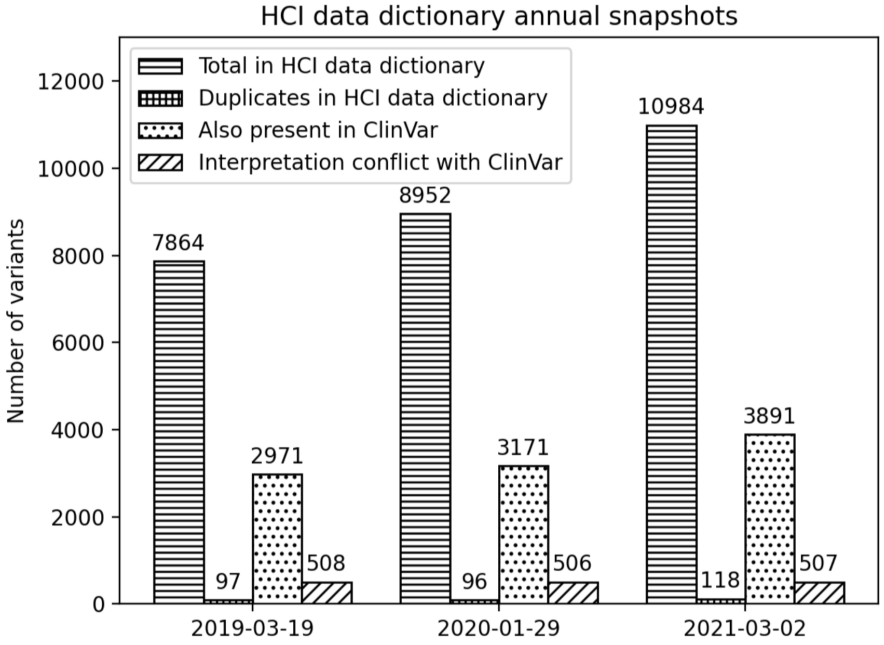
References

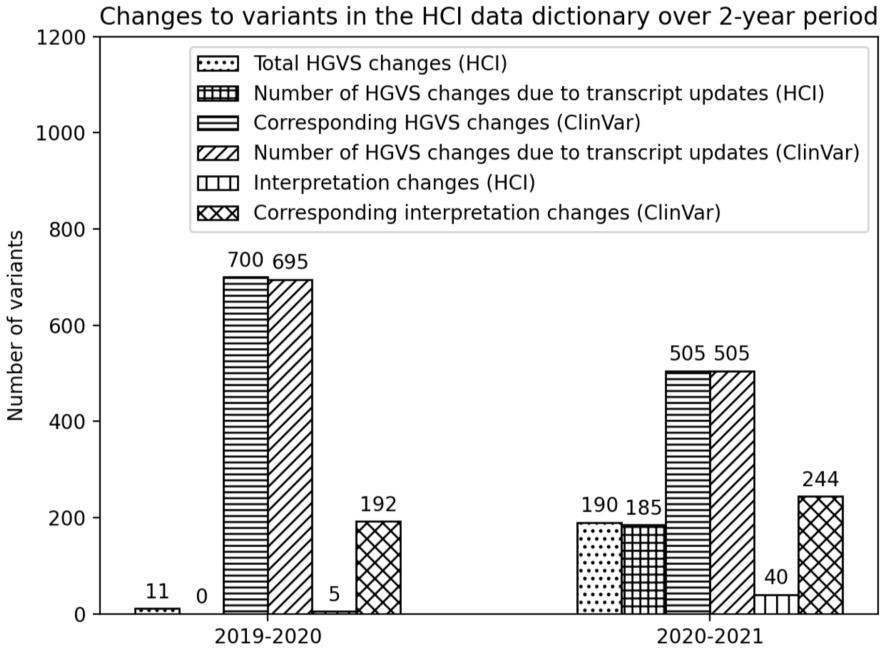
1. Huntsman Cancer Institute. (2018). *Quick Facts* . https://healthcare.utah.edu/huntsmancancerinstitute/news/press-kit.php.
2. Cesani, M., Lorioli, L., Grossi, S., Amico, G., Fumagalli, F., Spiga, I., Filocamo, M. and Biffi, A. (2016), Mutation Update of ARSA and PSAP Genes Causing Metachromatic Leukodystrophy. *Human Mutation* , 37: 16-27.*https://doi-org.ezproxy.lib.utah.edu/10.1002/humu.22919*
3. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT (2011). LOVD v.2.0: the next generation in gene variant databases.*Human Mutation* , 32(5):557-63.
4. U.S. National Library of Medicine. (2013). *ClinVar.* National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/clinvar/.
5. Clinical Genome Resource. (2013). *Explore the clinical relevance of genes & variants* . ClinGen. https://clinicalgenome.org/.
6. U.S. National Library of Medicine. (1999). *Home - SNP - NCBI* . National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/snp/.
7. Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P., & Church, D. M. (2013). DbVar and DGVa: public archives for genomic structural variation.*Nucleic acids research, 41(Database issue)* , D936–D941.*https://doi.org/10.1093/nar/gks1213*
8. Institute of Medical Genetics in Cardiff. (2007). *HGMD® home page* . HGMD. http://www.hgmd.cf.ac.uk/ac/index.php.
9. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* , 581(7809), 434–443.*https://doi.org/10.1038/s41586-020-2308-7*
10. Griffith, M., Spies, N. C., Krysiak, K., McMichael, J. F., Coffman, A. C., Danos, A. M., Ainscough, B. J., Ramirez, C. A., Rieke, D. T., Kujan, L., Barnell, E. K., Wagner, A. H., Skidmore, Z. L., Wollam, A., Liu, C. J., Jones, M. R., Bilski, R. L., Lesurf, R., Feng, Y. Y., Shah, N. M., ... Griffith, O. L. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nature genetics* , 49(2), 170–174.*https://doi.org/10.1038/ng.3774*
11. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). (1998). *Online Mendelian Inheritance in Man (OMIM)* . OMIM. https://omim.org/.
12. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., ... Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic acids research* , 47(D1), D941–D947.*https://doi.org/10.1093/nar/gky1015*
13. Pawliczek, P., Patel, R. Y., Ashmore, L. R., Jackson, A. R., Bizon, C., Nelson, T., Powell, B., Freimuth, R. R., Strande, N., Shah, N., Paithankar, S., Wright, M. W., Dwight, S., Zhen, J., Landrum, M., McGarvey, P., Babb, L., Plon, S. E., Milosavljevic, A., & Clinical Genome (ClinGen) Resource (2018). ClinGen Allele Registry links information about genetic variants. *Human Mutation* , 39(11), 1690–1701.*https://doi.org/10.1002/humu.23637*
14. Department of Integrative, Structural and Computational Biology @ Scripps Research. (2020). *Variant Annotation as a Service* . MyVariant.info. https://myvariant.info/.
15. Global Alliance for Genomics & Health. (2019). *GA4GH Variation Representation Specification* . GA4GH Variation Representation Specification - GA4GH Variation Representation Specification 1.1.2 documentation. https://vrs.ga4gh.org/en/stable/.
16. U.S. National Library of Medicine. (2021). *Index of /pub/clinvar/tab_delimited* . National Center for

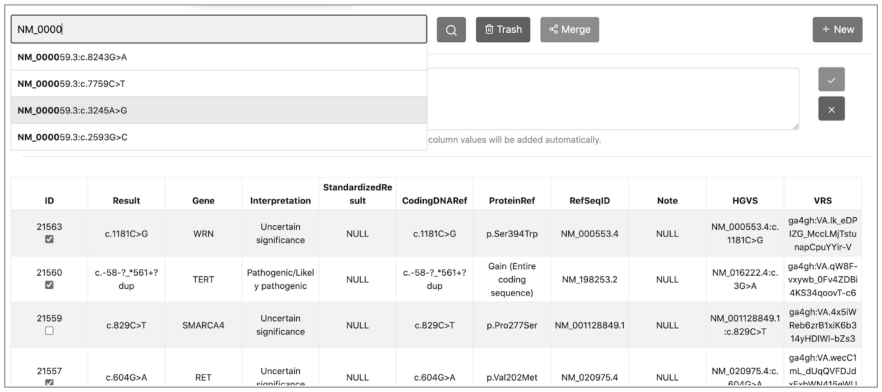
Biotechnology Information.*https://ftp.ncbi.nlm.nih.gov/pub/clinvar/tab_delimited/*.
17. GA4GH, *vrs-python* , (2017), GitHub repository,*https://github.com/ga4gh/vrs-python/*
18. Henrie, A., Hemphill, S. E., Ruiz-Schultz, N., Cushman, B., DiStefano, M. T., Azzariti, D., Harrison, S. M., Rehm, H. L., & Eilbeck, K. (2018). ClinVar Miner: Demonstrating utility of a Web-based tool for viewing and filtering ClinVar data. *Human Mutation* , 39(8), 1051–1060.*https://doi.org/10.1002/humu.23555*
19. Xiang, J., Yang, J., Chen, L. et al. (2020). Reinterpretation of common pathogenic variants in ClinVar revealed a high proportion of downgrades. *Sci Rep* , 10, 331.*https://doi.org/10.1038/s41598-019-57335-5*
20. Melissa J Landrum, Shanmuga Chitipiralla, Garth R Brown, Chao Chen, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Kuljeet Kaur, Chunlei Liu, Vitaly Lyoshin, Zenith Maddipatla, Rama Maiti, Joseph Mitchell, Nuala O'Leary, George R Riley, Wenyao Shi, George Zhou, Valerie Schneider, Donna Maglott, J Bradley Holmes, Brandi L Kattman. (2020). ClinVar: improvements to accessing data, *Nucleic Acids Research,* 48, D1. D835–D844,*https://doi.org/10.1093/nar/gkz972*
21. Rehm, H.L., Fowler, D.M. (2020). Keeping up with the genomes: scaling genomic variant interpretation. *Genome Med* 12, 5 . https://doi.org/10.1186/s13073-019-0700-4
22. Mwatkin8, *LocalVar* , (2021), GitHub repository, https://github.com/mwatkin8/LocalVar

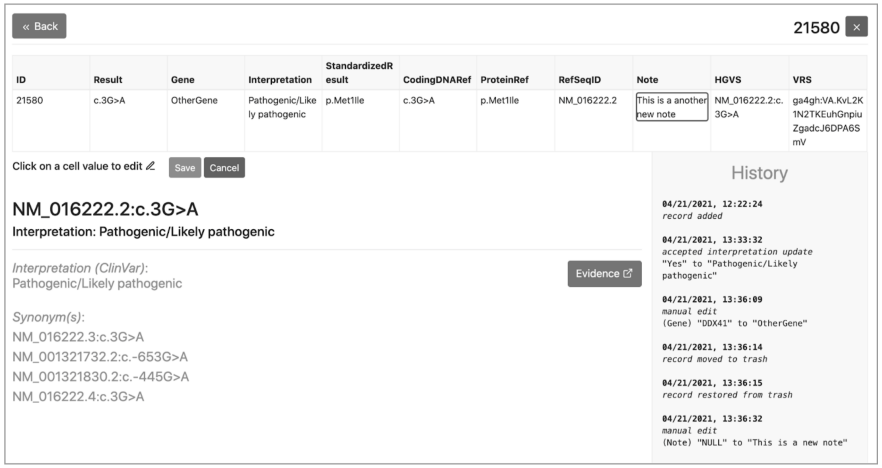

```
"NM_002691.3:c.3221G>A": {
    "ID": "21549",
    "Interpretation": "Indeterminate",
},...
```

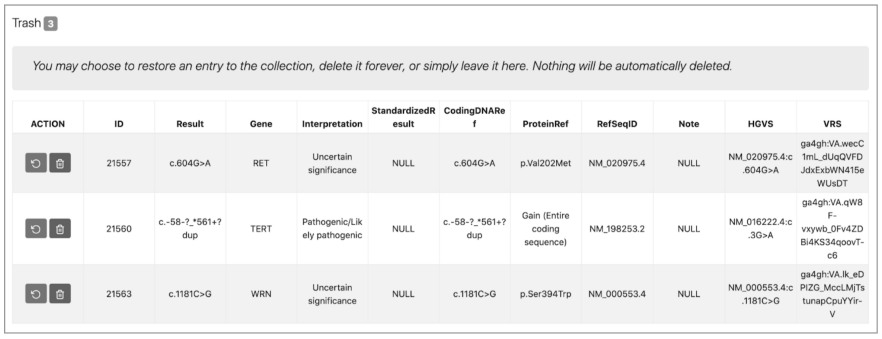
```
"NM_002691.3:c.3221G>A": {
    "ID": "21549",
    "Interpretation": "Indeterminate",
    "ClinVar VariationID": "422817",
    "ClinVar Interpretation": "Uncertain significance",
    "ClinVar Synonyms": [
        {
            "HGVS": "NM_001256849.1:c.3221G>A",
            "VRS": "ga4gh:VA.lBNO0Fr2jHTArc4UOeeKfjqG3bFyhZlS"
        },
        {
            "HGVS": "NM_001308632.1:c.3299G>A",
            "VRS": "ga4gh:VA.A1mWD9XUkELtTApMQoWrcPJrj8fB4dG7"
        },
        {
            "HGVS": "NM_002691.4:c.3221G>A",
            "VRS": "ga4gh:VA.OJuKUg1zfGDba-2d4K7Gv_bgmSemkt-h"
        }
    ]
},...
```

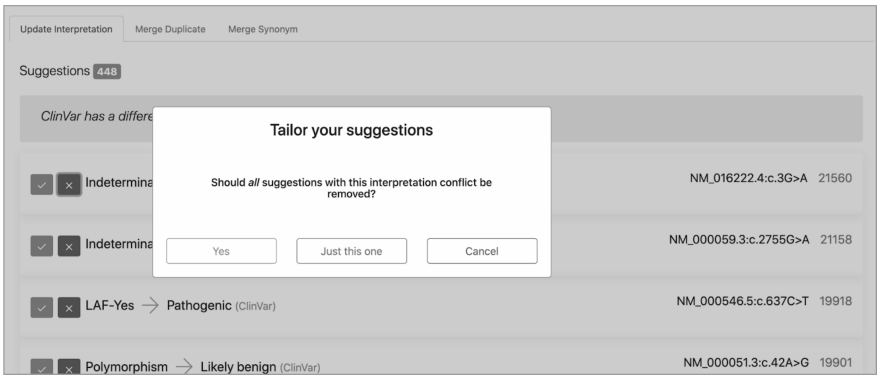


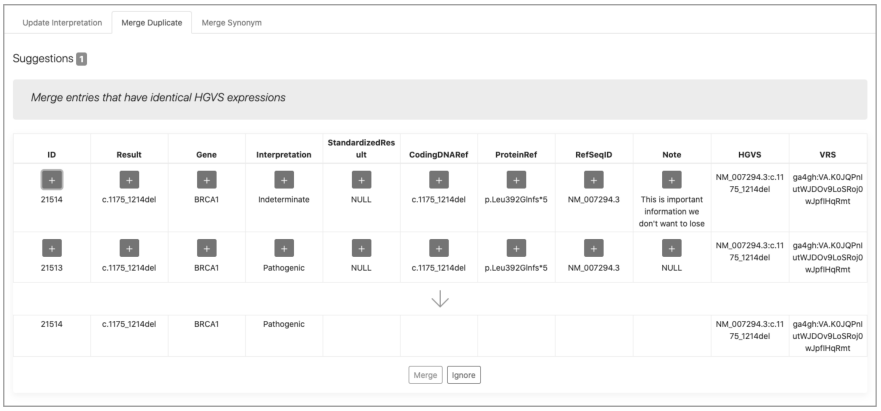












Hosted file

Table 1.docx available at https://authorea.com/users/417469/articles/524572-localvar-a-local-variant-collection-manager-to-asynchronously-detect-synonyms-hgvs-expression-changes-and-variant-interpretation-changes-from-clinvar