White Paper

Manual Curation vs. Artificial Intelligence: Can Automated Variant Evidence Retrieval Replace Human Judgment?

Stanford University compares data quality from their Automatic VAriant evidence DAtabase (AVADA) to the Human Gene Mutation Database (HGMD)

Variant Annotation in Clinical Genomics

Next Generation Sequencing (NGS) methodologies such as Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) give us the ability to search for clinically significant variations across the human genome, providing detailed genetic information that influences patient diagnosis and treatment. NGS applications greatly expand the number of novel genetic variants — changes in DNA sequence not previously reviewed by the medical community. Earlier approaches to genetic testing assayed established variants in specific genes. These targeted variants are considered established because they have been reviewed extensively by the medical community and the clinical relevance is evident prior to testing (1). The vast increase in genetic data from NGS poses new challenges as the process of annotating and understanding the clinical relevance of genetic information is time-consuming and requires the expertise of highly trained individuals (2,3). The clinical advantages of data produced through NGS may not be leveraged to its full extent due to the rate-limiting step of researching and assessing genetic variants.

Variant classification is an essential step in the genetic testing workflow and refers to assigning clinical significance to the observed DNA variations in patient samples. This process is complex, time-consuming, and challenging, requiring expert knowledge and experience. Data sources that provide information on variants are numerous, heterogeneous, quickly evolving, and sometimes conflicting. Because of this, differences in variant classifications can exist between laboratories; these discrepancies have the potential to impact clinical decision making (4).

The American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) published guidelines in 2015 to standardize variant classification across clinical testing labs. These guidelines are considered the gold standard for the interpretation of sequence variants (5). According to ACMG guidelines the retrieval of variant evidence from the literature should be combined with the evaluation of the validity and strength of available evidence to reach a final pathogenicity categorization for the clinical significance of genetic data.

Narrowing Down the Data Retrieved From the Available Sources

The retrieval of variant evidence from various sources can be daunting when thousands of variants in hundreds of genes need review. Considering that a typical singleton patient exome contains 200–500 rare variants, variant classification alone can take up to a month per patient if performed manually (6). This is because each of the detected variants requires independent evaluation, giving scientists the difficult task of determining which variants have a functional impact on the protein and a clinical impact. Filtering and prioritizing genetic variants before considering literature evidence to reduce the variant list to those that need further attention is one of the most challenging tasks in clinical genomics.

Variant scientists evaluate different types of variant evidence to begin prioritizing a list of variants for literature review. They review allele frequencies from population data for healthy individuals, computational and predictive data, and reports from other labs in public databases such as ClinVar to begin narrowing down the list. This job is time-consuming and almost impossible to perform manually, considering a sequenced human exome can contain more than 20,000 variants that must be checked across a diversity of sources and databases.

Scientific literature is one of the most important sources of empirical evidence to help a scientist determine whether a variant is clinically relevant, but it is also the most challenging and the most time-consuming source to evaluate. After initial filtering of the variant data, variant scientists may still be left with a large list of variants that require comprehensive literature review. Simply identifying the relevant literature evidence for a variant can be a tedious process. Scientists must search for literature across multiple platforms, using variant nomenclature across multiple transcripts, and sometimes research non-standard or historical naming schemes for variants. The time and effort put into the literature search can be substantial before taking into account the time required to read and analyze the data within the publications found. Often, scientific publications report conflicting data and interpretations for a particular variant and the variant scientist must review not only the content, but the validity of the study.

Machine Learning to Identify Relevant Evidence

The Automatic VAriant evidence DAtabase (AVADA), is a modern, advanced machine learning tool that was developed to accelerate the process of retrieving variant evidence from the literature. AVADA automates identification of variant-level evidence from PubMed indexed literature and converts the data to genomic coordinates. AVADA was also developed with the aim of improving the quality and precision of the extracted data since the 26% error rate in associating a variant with the correct gene with previously available crowd-sourcing and artificial intelligence (AI) applications was unsatisfactory (7).

To investigate the clinical utility of the automatically retrieved data and to show the amount of valuable evidence, AVADA evidence was compared with the gold-standard curated Human Gene Mutation Database (HGMD) and the ClinVar database of lab observed variants. From the AVADA database that contained 61,116 articles, data for 203,536 variants was automatically retrieved (GRCh37/hg19 chromosome, position, reference allele, and alternative allele) (Figure 1a). Of the variants retrieved by AVADA, 85,888 coincided with disease-causing variants in HGMD, corresponding to 61% of disease-causing variants in HGMD (Figure 1b). AVADA contained 26,033 (55%) of all likely/pathogenic variants listed in the ClinVar (Figure 1c). It outperformed other automated text processing methods such as tmVar (8), which retrieved only 14% disease-causing HGMD variants and only 31% of likely/pathogenic variants in ClinVar.

Comparing the data between the three, AVADA retrieved 62,180 variants known to be disease-causing in HGMD but that have not been reported in ClinVar. Only 2,325 variants retrieved in AVADA are reported in ClinVar as likely/pathogenic but were not listed in the HGMD. Of the variants extracted by AVADA, 115,323 (56%) are neither present in the HGMD nor in the ClinVar (Figure 1d), so are of uncertain quality and clinical relevance.

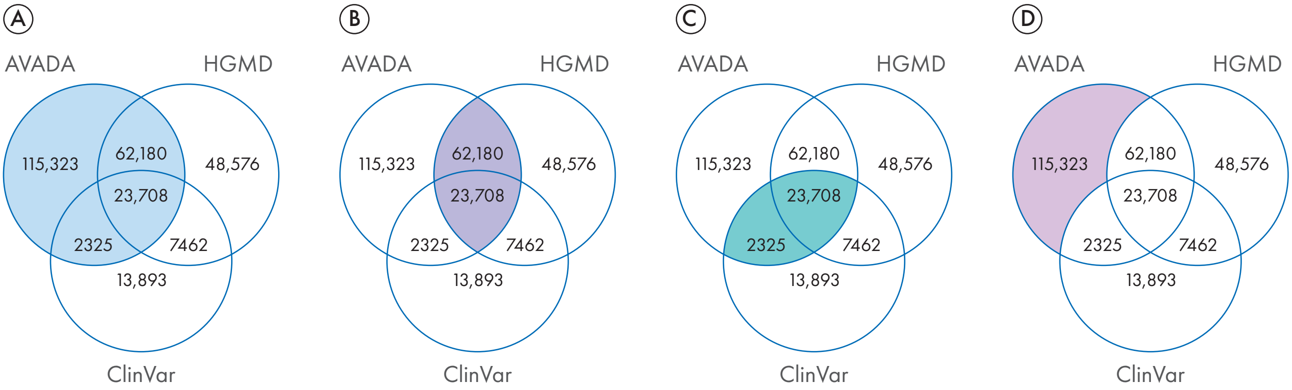

Figure 1. A) 203,536 total variants were automatically extracted by AVADA; B) 61% of variants reported as disease-causing in the expert-curated HGMD were found by AVADA; C) 55% of variants reported as likely/pathogenic in the ClinVar were also found in AVADA; D) 56% of variants retrieved by ADAVA were neither in HGMD nor in ClinVar. These variants are of questionable quality and clinical utility.

Automatization – Quantity vs. Quantity?

There is no doubt that automatic data collection using machine learning and artificial intelligence can aid in variant evidence retrieval. Using these systems, we will be able to automatically retrieve data for hundreds of thousand variants from tens of thousands of downloaded and parsed publications. From a total of 61,116 articles that made it into the final AVADA database, 203,536 distinct variants in 5827 genes were automatically retrieved (9). Comparing the data between the AVADA and HGMD that does not overlap, AVADA had twice as many unique automatically extracted variants as HGMD had unique expert-exacted variants (118,000 vs. 56,000) (Figure 2).

However, we must not forget the presence of false-positives, such as from indirect gene-variant-article references from tables, incompletely described variants or variants lacking Human Genome Variation Society (HGVS) nomenclature, imprecise chromosome positions, alternative variant nomenclature (ex. historical names including for genes), and alternative transcripts resulting in different protein positions corresponding to same DNA variant, among other variability in variant descriptions that often exist in the literature. These are hard for automated text mining approaches to parse and translate accurately, so data collection and variant classification requires additional scrutiny and caution. Assessing the validity

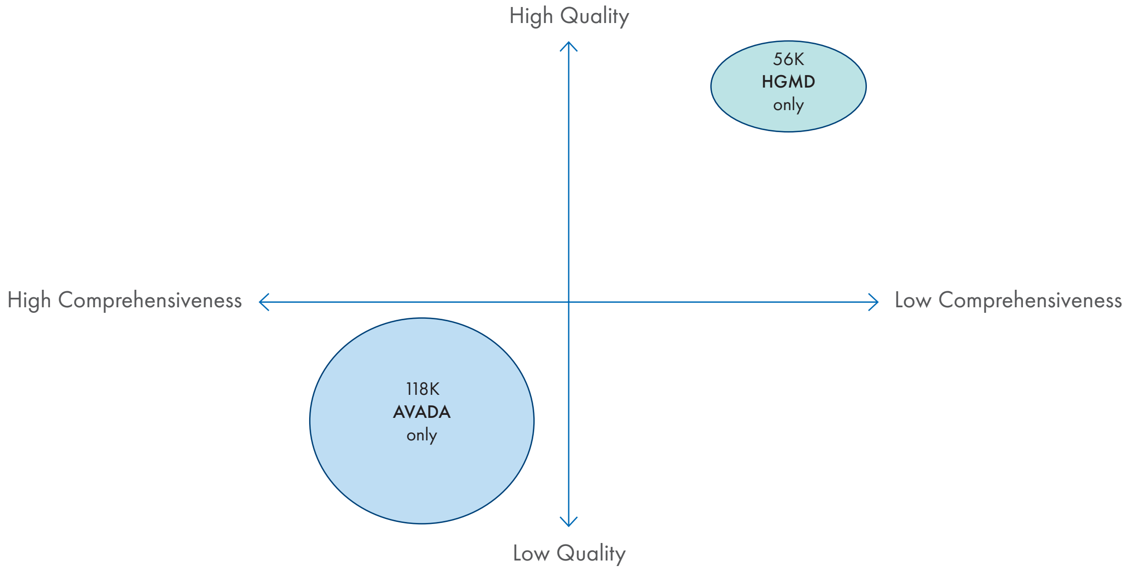

Figure 2. Manual inspection of randomly selected AVADA variants showed lower quality and comprehensiveness compared to expert-curated, high-quality HGMD data even though twice as much data was automatically extracted.

and weight of individual papers and deciding on a final classification on the potentially conflicting sources is a challenging task. Some of the existing discrepancies can only be resolved manually by critically assessing and reviewing supplementary materials in the articles or by direct contact with the authors.

To assess the quality of automatically retrieved data, 200 AVADA variants were randomly selected and manually reviewed to assess quality (9). It was found that the majority were incorrectly extracted. This means that when using automatically text-mined content, the associated literature reference provided for a variant will be irrelevant about 4 in 10 times, yielding false-positive results. Thus, even though large numbers of potential variants can be automatically extracted from the literature, there is a data quality issue beyond their clinical relevance (Figure 2). Because of this, the AVADA publication website warns users of these significant quality and incompleteness concerns.

Due to many existing uncertainties and the inability of known AI systems to address them — including AVADA — , expert curation by variant scientists is irreplaceable in the field of clinical genomics.

In an effort to minimize false-positive variant mentions as much as possible, AVADA scoped its processing to abstracts and full-text papers they first auto-classified as about hereditary disease. Even so, the resulting estimated 73% recall and 49.5% precision of relevant articles indicates the challenges of automatically accurately identifying relevant articles beyond extracting mutation references. AVADA recovered nearly 60% of disease-causing HGMD variants leaving over 40% of disease-associated mutations undetected. This indicates that modern text extraction approaches including AVADA are not suitable as the sole source of literature identification for clinical purposes.

HGMD Data Quality

HGMD combines biomedical informatics and human search procedures for variant data curation to provide high-quality variant references. HGMD is constantly updated by a team of expert variant scientists. They screen peer-reviewed biomedical literature on a daily basis via manual inspection of scientific journals to classify variants as disease-causing or possibly disease-causing (for Mendelian conditions), or as disease-associated (for multifactorial diseases).

To retrieve genetic variants from the literature, it is important to account for differing nomenclature across gene transcripts. HGMD catalogues variants across multiple genome builds and transcripts, providing reliable variant data. To identify which variant description maps to which mentioned gene in the article, AVADA first forms gene–variant candidate mappings between each variant nomenclature description and each mentioned gene if the variant matches at least one RefSeq transcript of the gene (9). However, transcript identifiers are sometimes omitted from the papers and differing nomenclature can exist across transcripts making the task difficult for AI systems.

HGMD is the only database that pursues a policy of continuous curation and reclassification, not relying solely on the original submitter updating their submission. Variant reclassification continually takes place in HGMD, giving high-quality data to the laboratory scientists who use it – making sure that they do not miss pivotal studies that may change the assessment of pathogenicity for a variant.

HGMD offers the most comprehensive database of articles supporting the clinical significance of hereditary disease mutations. The value of HGMD is that its associated literature evidence is essential for the variant classification, and users can trust and rely on the information it provides. AI systems for automatic data retrieval should be considered an aid in variant retrieval but are far from sufficient. HGMD also includes variants identified by text mining approaches, but all such variants are validated for their clinical significance by HGMD expert curators.

Conclusion

Clinical genomics has high competency and quality demands because the final interpretative results will only be as good as the evidence used. AVADA and other AI tools are meant to identify relevant information for variant scientists to evaluate and are helpful. However, if the scientist solely relies on automated text mining approaches such as AVADA to identify relevant literature, critical evidence will often be missed. With a significant percentage of undetected disease-associated mutations and false-positive article associations, modern text mining approaches cannot compete with HGMD's data quality and expert curation in terms of the accuracy and completeness of the clinical literature data.

Notes

References

1 Bewicke-Copley F, Arjun Kumar E, Palladino G, Korfi K, Wang J. (2019). Applications and analysis of targeted genomic sequencing in cancer studies. Comput. Struct. Biotechnol. J., **17**, 1348–1359.

2 Lappalainen T, Scott AJ, Brandt M, Hall IM. (2019). Genomic Analysis in the Age of Human Genome Sequencing. Cell, **177**(1), 70–84.

3 Ormondroyd E, Mackley MP, Blair E, Craft J, Knight JC, Taylor JC, Taylor J, Watkins H. (2018). "Not pathogenic until proven otherwise": perspectives of UK clinical genomics professionals toward secondary findings in context of a Genomic Medicine Multidisciplinary Team and the 100,000 Genomes Project. Genet. Med., **20**(3), 320–328.

4 Rehm HL, et al. (2015). ClinGen – the Clinical Genome Resource. N. Engl. J. Med., **372**(23), 2235–2242.

5 Richards S, 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**(5), 405–424.

6 Dewey FE, et al. (2014). Clinical interpretation and implications of whole-genome sequencing. JAMA, **311**(10), 1035–1045.

7 Bungartz KD, Lalowski K, Elkin SK. (2018). Making the right calls in precision oncology. Nature biotechnology, **36**(8), 692–696.

8 Wei CH, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. (2018). tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. Bioinformatics, **34**(1), 80–87.

9 Birgmeier J, Deisseroth CA, Hayward LE, Galhardo L, Tierno AP, Jagadeesh KA, Stenson PD, Cooper DN, Bernstein JA, Haeussler M, Bejerano G. (2020). AVADA: toward automated pathogenic variant evidence retrieval directly from the full-text literature. Genet. Med., **22**(2), 362–370.

10 http://bejerano.stanford.edu/AVADA/

Learn more at **digitalinsights.qiagen.com/HGMD**

Ordering **www.qiagen.com/shop** | Technical Support **support.qiagen.com** | Website **digitalinsights.qiagen.com**