Title

AVADA: toward automated pathogenic variant evidence retrieval directly from the full-text literature

Journal

Genetics in Medicine, 22(2)

ISSN

1098-3600

Authors

Birgmeier, Johannes
Deisseroth, Cole A
Hayward, Laura E
et al.

Publication Date

2020-02-01

DOI

10.1038/s41436-019-0643-6

Peer reviewed

HHS Public Access

Author manuscript

Genet Med. Author manuscript; available in PMC 2020 June 18.

Published in final edited form as:

Genet Med. 2020 February ; 22(2): 362–370. doi:10.1038/s41436-019-0643-6.

AVADA: Towards Automated Pathogenic Variant Evidence Retrieval Directly from the Full Text Literature

Johannes Birgmeier, MSc[1], **Cole A. Deisseroth, BSc**[1], **Laura E. Hayward, MS**[2], **Luisa M. T. Galhardo, BSc**[1], **Andrew P. Tierno**[1], **Karthik A. Jagadeesh, PhD**[1], **Peter D. Stenson, BSc**[3], **David N. Cooper, BSc, PhD**[3], **Jonathan A. Bernstein, MD, PhD**[4], **Maximilian Haeussler, PhD**[5], **Gill Bejerano, PhD**[1,4,6,7,*]

[1]Department of Computer Science, Stanford University, Stanford, California 94305, USA

[2]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

[3]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, UK

[4]Department of Pediatrics, Stanford School of Medicine, Stanford, California 94305, USA

[5]Santa Cruz Genomics Institute, MS CBSE, University of California Santa Cruz, California 95064, USA

[6]Department of Developmental Biology, Stanford University, Stanford, California 94305, USA

[7]Department of Biomedical Data Science, Stanford University, Stanford, California 94305, USA

Abstract

Purpose: Both monogenic pathogenic variant cataloging, and clinical patient diagnosis start with variant-level evidence retrieval followed by expert evidence integration in search of diagnostic variants and genes. Here, we try to accelerate pathogenic variant evidence retrieval by an automatic approach.

Methods: AVADA (Automatic Variant evidence DAtabase) is a novel machine learning tool that uses natural language processing to automatically identify pathogenic genetic variant evidence in full text primary literature about monogenic disease and convert them to genomic coordinates.

[*]**Corresponding author:** Gill Bejerano, bejerano@stanford.edu, Stanford University, Stanford, CA 94305, 1 (650) 725-6792.

Author Contributions

J.B. and M.H. wrote software to map variants to a reference set of mapped transcripts. J.B., C.A.D., L.E.H., L.M.T.G., and A.P.T. validated AVADA-retrieved variants. J.B. wrote the machine learning classifiers and executed performance evaluations. J.B., M.H., and G.B. wrote the manuscript. C.A.D. and K.A.J. downloaded and processed DDD data. P.D.S. and D.N.C. created HGMD and helped with manual variant inspection. J.A.B. provided guidance on clinical aspects of study design, testing set construction and interpretation of results. G.B. supervised the project. All authors read and commented on the manuscript.

Conflicts of interest

P.D.S. and D.N.C are the creators of HGMD. They receive financial support for it from Qiagen LTD through a License Agreement with Cardiff University. The other authors declare no conflict of interest.

Web resources

AVADA code for automatic variant evidence retrieval, and the automatically retrieved (but not validated) variant evidence database, will be available for non-commercial use at http://bejerano.stanford.edu/AVADA

Results: AVADA automatically retrieved almost 60% of likely disease-causing variants deposited in HGMD, a 4.4x-fold improvement over the current best open source automated variant extractor. AVADA contains over 60,000 likely disease-causing variants that are in HGMD, but not in ClinVar. AVADA also highlights the challenges of automated variant mapping and pathogenicity curation. However, when combined with manual validation, on 245 diagnosed patients, AVADA provides valuable evidence for an additional 18 diagnostic variants, on top of ClinVar's 21, vs. only 2 using the best current automated approach.

Conclusion: AVADA advances automated retrieval of pathogenic monogenic variant evidence from full-text literature. Far from perfect, but much faster than PubMed/Google Scholar search, careful curation of AVADA-retrieved evidence can aid both database curation and patient diagnosis.

Keywords

automatic variant retrieval; machine learning; natural language processing; full-text extraction; variants database

Introduction

Exome or genome sequencing is entering clinical practice in aid of the identification of molecular causes of highly penetrant genetic diseases, particularly Mendelian disorders, where typically one or two of the patient's genetic variants in a single gene are causative (diagnostic) of the patient's disease. After variant filtering, a typical singleton patient exome contains 200–500 rare variants[1,2]. Identifying causative variants is therefore time-consuming, as investigating each variant can take up to an hour[3]. Various approaches strive to accelerate this process[4–6]. Identifying diagnostic variants can be greatly accelerated if the patient's genome contains a previously reported disease-causing variant that partly or fully explains their phenotype. The American College of Medical Genetics (ACMG) guidelines for the interpretation of sequence variants recommend variant annotation using databases of reported pathogenic variants[7].

Variant curation from the literature includes (a) retrieval of variant evidence from the literature and (b) variant evaluation, which assesses the validity and power of every piece of evidence, and combines all (possibly conflicting) evidence together to make variant- and gene-level diagnostic assertions. The rapidly growing literature on human genetic diseases[8], the costly process of manual variant curation[9], and improved computational access to the full text of primary literature[10] serve to incentivize automating parts of the variant curation pipeline. Here, we focus on automating the process of variant evidence retrieval from the primary literature about Mendelian diseases without attempting to automate variant evaluation. Automatic pathogenic variant evidence retrieval from the primary literature involves finding papers about Mendelian diseases that include variant descriptions (such as "c.123A>G"), linking them to a transcript of the correct gene, and converting these to genomic coordinates for ease of downstream use. Previous work on automatic variant evidence retrieval from the literature has largely focused on finding variant descriptions in article titles and abstracts without converting the discovered variants to genomic coordinates[11–14] or only to dbSNP[15] variant identifiers (rsIDs). Mapping textual variant

descriptions directly to reference genome coordinates requires significant effort, and has thus far largely been left to manually curated databases such as HGMD[16] and ClinVar[17], which devote many worker-hours to the task.

The ClinGen[9] project has proposed to "develop machine-learning algorithms to improve the throughput of variant interpretation" and notes that a rate-limiting factor for clinical use of variant information is the lack of openly accessible knowledgebases capturing known genetic variants. AVADA (Automatic VAriant evidence DAtabase) is trained on a sample of manually curated variants (from ClinVar and HGMD), and then applied to the entire body of PubMed indexed literature for automatic retrieval of published variants in papers about Mendelian disease. We show that AVADA improves on the state of the art in automated variant retrieval by comparing it to tmVar 2.0, a best-in-class tool[18] used to harvest variants from PubMed abstracts.

To show the amount of valuable evidence retrieved by AVADA, we also compare variants collected by our approach to the fully curated HGMD and ClinVar databases. We also show, on 245 real patients, that when AVADA is coupled with manual curation, it can aid diagnosis.

We note that the creators of HGMD, Drs. Stenson and Cooper, are co-authors of this study. They provided HGMD data and advised on validating and curating AVADA literature-extracted variant evidence and the comparison of AVADA to HGMD.

Materials and Methods

Identification of relevant literature

PubMed is a database containing titles and abstracts of biomedical articles, only a subset of which contain descriptions of variants that cause human genetic disease. A document classifier is a machine learning classifier that takes as input arbitrary text and classifies it as "positive" (here, an article about genetic disease) or "negative" (otherwise). We trained a scikit-learn[19] LogisticRegression[20] classifier to identify relevant documents using positive input texts (titles and abstracts of articles cited in "Allelic Variants" sections of OMIM[21], and all of HGMD[16] which curates germline disease-causing and disease-associated variants) and negative input texts (random PubMed titles and abstracts). Input texts were converted into a feature vector based on the frequency of words in input documents by means of a scikit-learn CountVectorizer followed by a Term Frequency-Inverse Document Frequency[20] (TF-IDF) transformer. After training the title/abstract document classifier, we applied it to all 25,793,020 titles and abstracts in PubMed to identify articles that might be relevant to the diagnosis of genetic diseases (if the title/abstract classifier returned a score of at least 0.5). Full text PDFs of potentially relevant articles were downloaded, converted to text using pdftotext[22] version 0.26.5, and evaluated for full-text relevance using a TF-IDF transformer, a full-text scikit-learn LogisticRegression classifier, and a threshold of 0.5 on the full text relevance scores.

A total of 133,410 articles were downloaded and subsequently classified as relevant to the diagnosis of human genetic diseases based on the articles' full text (Supplementary Methods). We refer to this set of articles as the "AVADA full-text articles" (Figure 1).

To estimate recall (the fraction of relevant articles that are contained in the AVADA full-text article set) of our pipeline, we took the set of 27,361 articles associated with "likely/pathogenic" variants in ClinVar. Of these, 24,724 (90%) were classified relevant by the title/abstract classifier, 23,978 (88%) were downloaded in full text, and 20,014 (73%) were finally classified relevant by the full-text classifier. To determine the precision (fraction of relevant articles among all AVADA full-text articles) of our pipeline, we randomly selected 200 articles from the set of AVADA full-text articles and manually assessed their relevance. An article was considered relevant if it discussed a Mendelian disease and its causative gene. 99 (49.5%) articles were determined to be relevant (Supplementary Table S1).

Variant and gene mention detection

In order to retrieve genetic variants from full-text articles about human genetic disease and convert them to genomic coordinates, it is necessary to detect both mentions of genes and variant descriptions in articles about genetic disease. This is because variant descriptions in HGVS-like notation, such as "c.123A>G", often do not contain the identity of the specific transcript or gene they reside in (Table 1).

AVADA extracts gene mentions from articles' full text using a custom-built database of gene and protein name entries from the HUGO Gene Nomenclature Committee (HGNC) and UniProt databases. These were matched case-insensitive to word groups of length 1–8 in the document. To identify variant mentions, we manually developed 47 regular expressions based on commonly observed HGVS-like variant notations in articles about human genetic disease (Supplementary Methods, Supplementary Table S2 and Figure 2A). At this step, we refer to every string that matches one of the 47 regular expressions as a "variant description". In the AVADA full-text articles, variant descriptions were identified in 92,436 articles, with a mean of 11.1 variant descriptions per article (Figure 1).

Mentioned genes form gene-variant candidate mappings with all plausible mentioned variant descriptions

Variant descriptions and their host gene mentions do not necessarily occur in the same sentence or even the same paragraph or page. To identify which variant description maps to which mentioned gene in the article, AVADA first forms so-called *gene-variant candidate mappings* between each variant description and each mentioned gene if the variant matches at least one RefSeq[23] transcript of the gene. For example, the variant description "c.123A>G" forms gene-variant candidate mappings with all mentioned genes that have an "A" at coding position 123 in at least one transcript (Supplementary Methods and Figure 2B). A variant description can form gene-variant candidate mappings with multiple genes, which are filtered in the next step. Gene-variant candidate mappings are converted to genomic coordinates in the GRCh37/hg19 reference assembly and initially result in a mean of 4.6 different genomic coordinates per extracted variant description (Figure 1).

Machine learning classifier selects the correct gene-variant mapping candidate

AVADA uses a scikit-learn GradientBoostingClassifier[24] to decide which gene-variant candidate mappings are likely to be correct. The training set comprised positive gene-variant mappings extracted from the literature that were referenced in all ClinVar entries, and a set of negative gene-variant mappings created by assigning variants from the positive training set to other genes mentioned in the article. For classification, each gene-variant mapping was converted to a feature vector, including the Euclidean distance between the 2D coordinates (consisting of page number, x and y coordinates of a mention) of the closest mentions of the variant and the gene in the PDF, the number of words between variant and gene mentions, and a number of other textual features containing information about the relationship between gene and variant mentions (Supplementary Methods and Figure 2C). Using a threshold of 0.9 (justified below), the gene-variant candidate classifier successfully reduced 4.6 candidate gene-variant mappings per variant description to a mean of 1.2 genomic coordinates (chromosome, position, reference, and alternative allele) in the final set of AVADA full-text articles (Supplementary Methods and Figures 1, 2D).

Results

AVADA retrieved 203,536 variants in 5,827 genes from 61,116 articles

A total of 61,116 articles made it into the final AVADA database, with a mean of 8.8 identified variant descriptions per article. From these articles, 203,536 distinct variants (GRCh37/hg19 chromosome, position, reference allele, and alternative allele) in 5,827 genes were automatically retrieved (Figure 1). The distribution of types of rare variants (<= 3% variant frequency in the healthy population[2]) in AVADA is strikingly similar to that of manually curated HGMD and ClinVar: for each of 6 variant categories (stoploss, nonframeshift indel, splicing, stopgain, frameshift, missense), the fraction of rare variants in AVADA are between the fraction of the respective category of rare variants in all of HGMD and ClinVar ±1% (Table 2). The articles used to construct AVADA are from a variety of journals, which are similar to the journals targeted by all of HGMD to curate its variants (9 out of the top 10 journals being the same between AVADA and all of HGMD; Figure 3A,B).

Each variant in AVADA is annotated with the PubMed ID(s) of publications where this variant was retrieved from, a HGNC[25] gene symbol, an Ensembl ID[26] and Entrez ID[27], the transcript RefSeq ID (e.g., NM_005101.3), and the exact variant description from the original article (e.g., “c.163C.T”). The latter allows AVADA users to later rapidly locate mentions of the variant within the body of the article.

AVADA is 61% precise at mapping gene-variant pairs to genomic coordinates—To estimate the precision of AVADA at extracting gene-variant candidate mappings in articles into genomic coordinates, 200 distinct random variants in AVADA were manually examined. For each of these variants, we selected the article associated with the (alphanumerically sorted) first PubMed ID in AVADA, and let 2 reviewers determine if the gene-variant candidate mapping from the article was correctly extracted to genomic coordinates using all lines of evidence in the article such as Sanger sequencing reads, UCSC

genome browser shots etc. 122 (61%) random variants were correctly extracted by AVADA (Supplementary Table S3).

AVADA recovers nearly 60% of disease-causing HGMD variants directly from the primary literature—We compared AVADA to HGMD and ClinVar versions with synchronized time stamps (Supplementary Methods). In this section, we subset HGMD to "DM" (disease-causing) variants and ClinVar to variants marked as "likely/pathogenic". 85,888 variants in AVADA coincided with variants marked as disease-causing ("DM") in HGMD, corresponding to 61% of all disease-causing variants in HGMD. We selected 200 distinct random variants from this set for verification by 2 reviewers. A variant was counted as correct and likely disease-causing if the reviewers came to a consensus that the original gene-variant candidate mapping was converted to the variant's genomic coordinates correctly in at least one of its associated articles, and the variant was described as likely disease-causing in the article. This was the case in 96.5% of the 200 variants (Supplementary Table S4). Thus, we infer that AVADA contains 58% of all disease-causing variants identified by HGMD.

We compared AVADA's performance to the best previously published automatic variant retrieval tool, tmVar 2.0[18], which attempts to map variant mentions in all PubMed abstracts to dbSNP identifiers (rsIDs). We converted rsIDs in tmVar 2.0 to genomic coordinates using mappings provided by dbSNP. tmVar retrieved only 19,481 (14%) disease-causing ("DM") HGMD variants (Supplementary Figure 1 and Figure 3C).

Considering only single nucleotide variants (SNVs), the largest class of known disease-causing variants, AVADA contains 70% of all "DM" SNVs in HGMD. Similarly, AVADA contains 26,033 (55%) of all "likely/pathogenic" variants in ClinVar and 62% of all "likely/pathogenic" SNVs in ClinVar. tmVar 2.0 retrieved only 14,841, or 31%, of pathogenic or likely pathogenic variants in ClinVar. Strikingly, AVADA contains 62,180 variants noted to be disease-causing in HGMD ("DM") but not in ClinVar ("likely/pathogenic").

2 reviewers evaluated a subset containing 200 distinct random variants of the remaining 115,323 variants that were retrieved by AVADA, but not reported as disease-causing in either HGMD ("DM") or ClinVar ("likely/pathogenic"). 68 (34%) of the 200 variants were correctly converted to genomic coordinates. 8 variants (4%) were further reported to be likely disease-causing (Supplementary Table S5). 7 of these 8 are contained in later versions of HGMD ("DM"), suggesting AVADA could both help curators accelerate variant retrieval as well as unearth a modest amount of undocumented likely disease-causing variants still hidden in the literature.

Diagnosis of patients with Mendelian diseases using AVADA

We analyzed the utility of known variant databases using 260 diagnostic (i.e., causative) variants from 245 patients with developmental disorders, diagnosed in Supplementary Table 4 of the Deciphering Developmental Disorders (DDD) study[28], obtained from EGA[29] study number EGAS00001000775 (Supplementary Methods).

Accuracy of variant annotation using AVADA, tmVar, HGMD and ClinVar—The more complete a variant database is, the higher its sensitivity when annotating patient genomes and the higher the likelihood of finding a diagnostic variant in the patient's genome. We determined how many of the 260 reported diagnostic DDD variants were found in AVADA, tmVar, HGMD ("DM" variants only), and ClinVar ("likely/pathogenic" variants only). The more disease-causing variants are contained in a database, the more rapidly some patients can be diagnosed. For this comparison, we subset AVADA and tmVar 2.0 to articles published until 2014 (before DDD publication), used only disease-causing "DM" variants entered until 2014 in HGMD, and used only likely/pathogenic variants from ClinVar version 20141202.

Of 260 different diagnostic variants reported by the DDD study, a total of 45 had evidence in AVADA from the scientific literature. Because AVADA retrieves variant evidence without validating it, all AVADA evidence needs to be manually assessed. Each patient variant found in AVADA was counted as correct if our 2 reviewers agreed that AVADA cited at least one article from which the variant was correctly mapped to genomic coordinates and the variant was reported as likely disease-causing in this article. 35 of the 45 variants found by AVADA fulfilled these criteria (Supplementary Table S6). Only 21 DDD diagnostic variants were listed in ClinVar and ascribed a pathogenicity level of "likely/pathogenic". Combining the free variant databases yielded 39 variants, almost as many as the 43 variants listed in HGMD ("DM"). Combining all three databases yielded 48 variants (Figure 3D). tmVar 2.0 contained only 13 diagnostic variants (Supplementary Table S7), all of which were in AVADA as well (Figure 3D).

We defined patient variants to be "candidate causative variants" if they were non-silent exonic or core splice-site mutations and occurred at an allele frequency of at most 0.5% in large databases of healthy controls[2,30,31] (Supplementary Methods). The 245 patients' data contained a mean of 435 non-diagnostic candidate variants each. To determine the variant annotation precision of AVADA, HGMD, ClinVar, and tmVar 2.0, we divided the number of distinct annotated diagnostic variants by the number of distinct annotated candidate variants across the 245 patients. A mean of 6.7 variants per patient were found in AVADA (2.8% precision), 3.5 in HGMD ("DM") (6.2% precision), 1.6 in ClinVar (clinical significance 4 or 5) (7.2% precision) in accordance with previous observations[2], and 3.4 in tmVar 2.0 (2.2% precision). Therefore, AVADA was less precise than the manually curated databases, but more precise than the previously best-in-class automatic variant retrieval tool tmVar 2.0.

Gene-variant pairing prediction classifier is robust to changes in training data

To examine how robust the gene-variant candidate pairing prediction classifier is to changes in training data, we first trained it on HGMD ("DM") instead of ClinVar data. The resulting variant set is highly similar to the original, having almost (−0.6%) the same size and 97% of it identical to the original (Supplementary Figure 2).

Next, we re-trained the gene-variant candidate classifier on 3 different subsets of the original ClinVar-based training data, each containing a random half of the articles in the original training data (Supplementary Methods). Again, after running variant extraction on the original set of AVADA articles, the variants returned from each re-training was highly

similar to the original AVADA variants, (+2.7%−0.6% bigger in size and containing 98% of the original set; see Supplementary Figure 3).

Picking the gene-variant candidate classifier threshold

To set the gene-variant classifier threshold at 0.9, we evaluated potential thresholds between 0.5 and 0.99 (Supplementary Figure 4). Including all gene-variant candidate mapping with a gene-variant candidate classifier score of at least 0.5 increased the number of distinct genetic variants to 291,281 (+43% compared to the original AVADA database). Since we previously selected 200 distinct variants to estimate AVADA precision (Supplementary Table S3), we now selected a proportional 87 distinct random variants between confidence levels 0.5 and 0.9 that were not already in the original AVADA variant set to manually determine database precision at lower gene-variant candidate classifier score levels. Again, 2 reviewers independently evaluated each variant.

Overall, these 87 variants were less than 21% correctly mapped to genomic coordinates (Supplementary Table S8), compared to 61% correct mapping in the original set. Inversely, if we subset the previously verified 200 AVADA variants (Supplementary Table S3) to variants arising from gene-variant mappings scored only above 0.95 and 0.99, respectively, variant extraction precision would modestly increase to 63.5%−68.5% (from 61%), but the fraction of recovered HGMD variants would decrease by 2.4%−15.1% (Supplementary Figure 4). Based on this search, we chose a gene-variant candidate classifier score threshold of 0.9 that balances precision and recall (Supplementary Figure 4).

Discussion

We studied the potential and challenges of creating an end-to-end machine learning tool for the automatic retrieval of variant evidence directly from full text literature about Mendelian disease. AVADA automatically retrieved nearly a hundred thousand disease-causing variants from tens of thousands of downloaded and parsed full-text articles. All AVADA variants are stored in a Variant Call Format[32] (VCF) file that includes the chromosome, position, reference and alternative alleles, variant descriptions as reported in the original article, and PubMed IDs of the original articles mentioning the variants.

AVADA makes a special effort not to curate just any variant mention, but rather to process only abstracts, and later full text papers that appear to our classifiers to describe pathogenic variants in the context of Mendelian diseases. While this undoubtedly removes a great number of false positive human variant mentions, AVADA's estimated 73% recall and 49.5% precision over relevant papers suggests that more can be done to optimize this process. AVADA also takes full advantage of recent success in allowing computerized access to the scientific literature. Its large gains over abstract-based tmVar 2.0 justify the engineering feat. However, the length and complexity of biomedical texts also result in AVADA's 61% precision in mapping mentioned variants to their correct genomic coordinates (currently not including mitochondrial variants). Clearly computational effort should continue to bring precision up. It is also worthwhile noting that while we wrote 47 regular expressions to capture most common variant mentions, HGVS formalism does request that variant mentions be preceded with an accepted reference sequence (e.g.,

NP_003995.2 for the GJB2 variant in Figure 2C). The more journals enforce and the more papers are written in strict HGVS notation, the easier the computerized coordinate conversion task becomes.

AVADA currently only tackles the first step in variant curation, that of evidence collection. Assessing the validity and power of individual papers, combining evidence from multiple, potentially conflicting papers, to arrive at variant and gene level assertions are even more challenging tasks. In this sense, AVADA can be thought of as a much quicker (and more hit-or-miss) means of evidence hunting than PubMed or Google Scholar. In order to assess the validity of AVADA's evidence we used the HGMD and ClinVar human curated databases of pathogenic variants. Neither of these is entirely accurate. For example, the ExAC paper[2] highlights 68 HGMD "DM" variants (55 of which are also in AVADA) that appear in >1% of at least one of its populations, requiring re-evaluation of HGMD's linked-papers based conclusions. With these caveats in mind, we estimate roughly 44% of AVADA collected variants as pathogenic Mendelian ones.

Despite these limitations, AVADA does recover nearly 60% of all disease-causing ("DM") variants deposited in HGMD at a fraction of the cost of constructing a manually curated database[9], and over 4 times as many as the abstract - rsID based tmVar 2.0. AVADA offers an approximate 64,617 (= 96.5% * 61,180 + 4% * 115,323; Figure 3C) disease-causing variants not present in ClinVar (136% increase over ClinVar alone), at the cost of nearly twice as many (112,886; complement of the above) additional benign or incorrectly extracted variants. In patient context if one is willing to manually validate AVADA evidence (as one is obliged to do with any compelling HGMD or ClinVar evidence as well), over the DDD example, ClinVar had evidence for 21 pathogenic variants. AVADA offers unvalidated evidence for an additional 27 variants, of which 18 were manually validated to be correctly mapped to genomic coordinates and correctly reported as disease-causing, essentially doubling ClinVar's reach.

AVADA shows the potential to (a) improve the state of the art in machine learning based evidence collection of literature-mentioned pathogenic variants and their mapping to reference genome coordinates, (b) enable first attempts to automate aspects of variant curation, and (c) motivate curation of benign variants as well as variants in other domains (such as cancer, mouse models, and other research fields where manually curated data may be scarce[33]). Combining AVADA-based rapid variant retrieval with validation will enable the creation and upkeep of cheaper, better, faster updating variant databases, which will ultimately empower both rapid diagnosis[9] and reanalysis[8].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Mark A. Lemley and Henry T. Greely for advice. This work was funded in part by a Bio-X Stanford Interdisciplinary Graduate Fellowship to J.B.; by grants EMBO ALTF292–2011 and NIH/NHGRI 5U41HG002371-15 to M.H.; and by DARPA, the Stanford Pediatrics Department, a Packard Foundation Fellowship, a Microsoft Faculty Fellowship and the Stanford Data Science Initiative to G.B. We are obliged to

thank the European Genome-Phenome Archive[29] (EGA) and the Deciphering Developmental Diseases[28] (DDD) project. The DDD study presents independent research commissioned by the Health Innovation Challenge Fund [grant number HICF-1009–003], a parallel funding partnership between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger Institute [grant number WT098051]. The views expressed in this publication are those of the author(s) and not necessarily those of the Wellcome Trust or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). Deidentified DDD data was obtained through EGA. This research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network.

References

1. Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet. 2015;47(7):717–726. doi:10.1038/ng.3304 [PubMed: 25985138]
2. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285–291. [PubMed: 27535533]
3. Dewey FE, Grove ME, Pan C, et al. Clinical interpretation and implications of whole-genome sequencing. JAMA. 2014;311(10):1035. doi:10.1001/jama.2014.1717 [PubMed: 24618965]
4. Smedley D, Jacobsen JOB, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015;10(12):2004–2015. doi:10.1038/nprot.2015.124 [PubMed: 26562621]
5. Jagadeesh KA, Birgmeier J, Guturu H, et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. Genet Med Off J Am Coll Med Genet. 7 2018. doi:10.1038/s41436-018-0072-y
6. Deisseroth CA, Birgmeier J, Bodle EE, et al. ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. Genet Med. 12 2018:1. doi:10.1038/s41436-018-0381-1
7. Richards S, Aziz N, Bale S, et al. 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 Off J Am Coll Med Genet. 2015;17(5):405–424. doi:10.1038/gim.2015.30
8. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. Genet Med. 2016;19(2):209–214. [PubMed: 27441994]
9. Clinical Genome (ClinGen) Resource. National Human Genome Research Institute (NHGRI). https://www.genome.gov/27558993/clinical-genome-clingen-resource/. Accessed September 27, 2018.
10. Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol. 2018;14(2):e1005962. doi:10.1371/journal.pcbi.1005962 [PubMed: 29447159]
11. Doughty E, Kertesz-Farkas A, Bodenreider O, et al. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. Bioinformatics. 2011;27(3):408–415. doi:10.1093/bioinformatics/btq667 [PubMed: 21138947]
12. Wei C-H, Harris BR, Kao H-Y, Lu Z. tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics. 2013;29(11):1433–1439. doi:10.1093/bioinformatics/btt156 [PubMed: 23564842]
13. Jimeno Yepes A, Verspoor K. Mutation extraction tools can be combined for robust recognition of genetic variants in the literature. F1000Research. 2014;3:18. doi:10.12688/f1000research.3-18.v2 [PubMed: 25285203]
14. Thomas P, Rocktäschel T, Hakenberg J, Lichtblau Y, Leser U. SETH detects and normalizes genetic variants in text. Bioinforma Oxf Engl. 2016;32(18):2883–2885. doi:10.1093/bioinformatics/btw234
15. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001;29(1):308–311. [PubMed: 11125122]

16. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet. 2017;136(6):665–677. doi:10.1007/s00439-017-1779-6 [PubMed: 28349240]

17. Landrum MJ, Lee JM, Benson M, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. 2016;44(D1):D862–868. doi:10.1093/nar/gkv1222 [PubMed: 26582918]

18. Wei C-H, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. Bioinformatics. 2018;34(1):80–87. doi:10.1093/bioinformatics/btx541 [PubMed: 28968638]

19. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–2830.

20. Jurafsky D, Martin JH. Speech and Language Processing. 2nd edition Upper Saddle River, N.J: Prentice Hall; 2008.

21. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015;43(Database issue):D789–798. [PubMed: 25428349]

22. Poppler. https://poppler.freedesktop.org/. https://poppler.freedesktop.org/. Accessed September 24, 2018.

23. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(Database issue):D733–D745. doi:10.1093/nar/gkv1189 [PubMed: 26553804]

24. Jagadeesh KA, Wenger AM, Berger MJ, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581–1586. doi:10.1038/ng.3703 [PubMed: 27776117]

25. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43(Database issue):D1079–1085. doi:10.1093/nar/gku1071 [PubMed: 25361968]

26. Yates A, Akanni W, Amode MR, et al. Ensembl 2016. Nucleic Acids Res. 2016;44(D1):D710–716. doi:10.1093/nar/gkv1157 [PubMed: 26687719]

27. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 2011;39(Database issue):D52–D57. doi:10.1093/nar/gkq1237 [PubMed: 21115458]

28. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015;519(7542):223–228. doi:10.1038/nature14135 [PubMed: 25533962]

29. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-phenome Archive of human data consented for biomedical research. Nat Genet. 2015;47(7):692–695. doi:10.1038/ng.3312 [PubMed: 26111507]

30. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, et al. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–1073. doi:10.1038/nature09534 [PubMed: 20981092]

31. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. Nature. 2015;526(7571):82. doi:10.1038/nature14962 [PubMed: 26367797]

32. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–2158. doi:10.1093/bioinformatics/btr330 [PubMed: 21653522]

33. McMurry JA, Köhler S, Washington NL, et al. Navigating the phenotype frontier: The Monarch Initiative. Genetics. 2016;203(4):1491–1495. doi:10.1534/genetics.116.188870 [PubMed: 27516611]

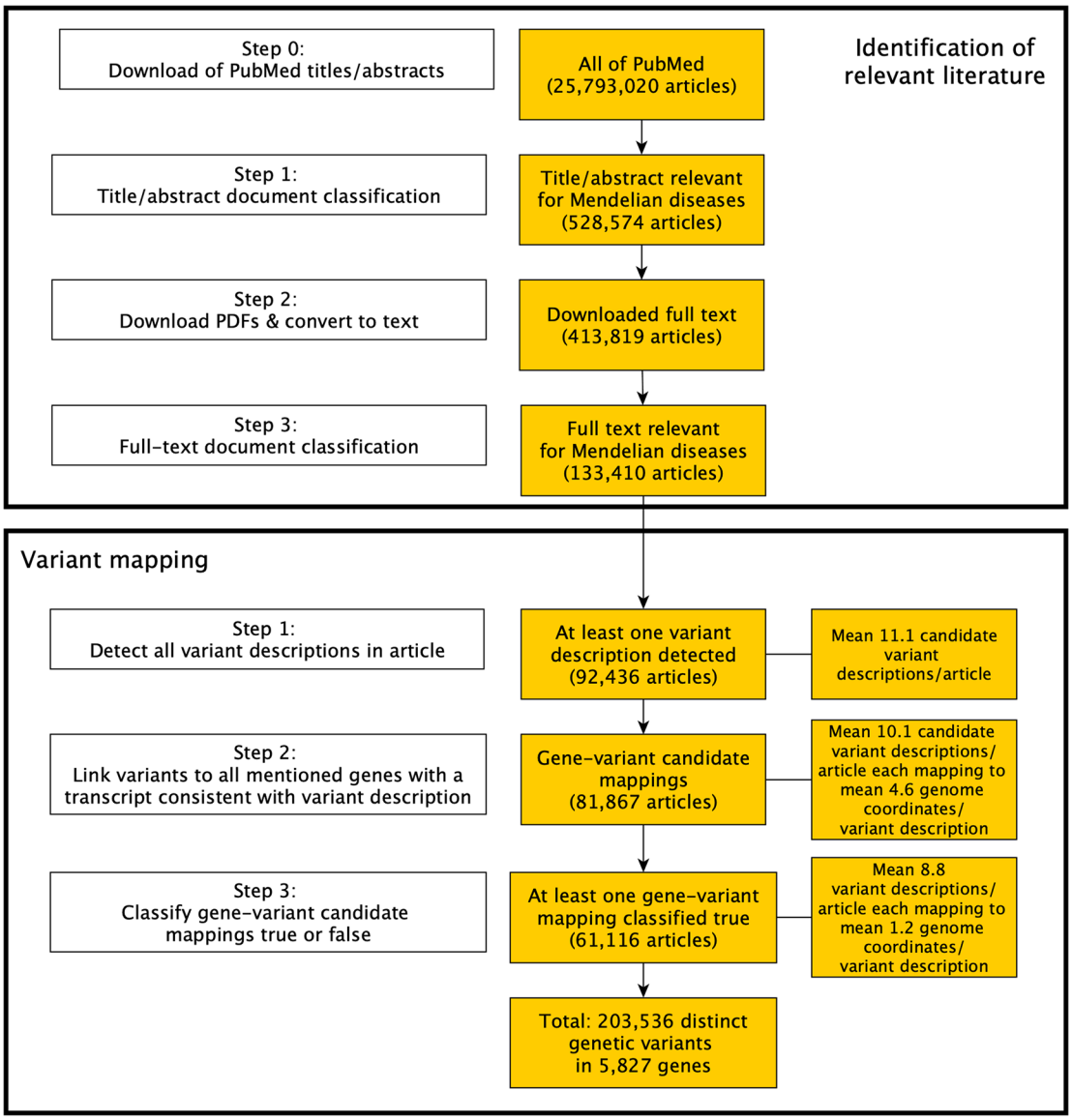

Figure 1. Construction of the automated variant evidence database AVADA. Identification of relevant literature:

AVADA discovers potentially relevant articles (about the genetic causes of Mendelian diseases) from PubMed, downloads their full text, and again filters potentially relevant articles based on the articles' full text. **Variant mapping**: Variant descriptions are detected in articles using 47 manually built regular expressions. Variant descriptions are then linked to mentioned genes to form gene-variant candidate mappings. Gene-variant candidate mappings are filtered using a gene-variant candidate classifier and converted to genomic coordinates. AVADA ultimately retrieves (unvalidated) evidence about 203,536 distinct genetic variants in 5,827 genes from 61,116 articles.

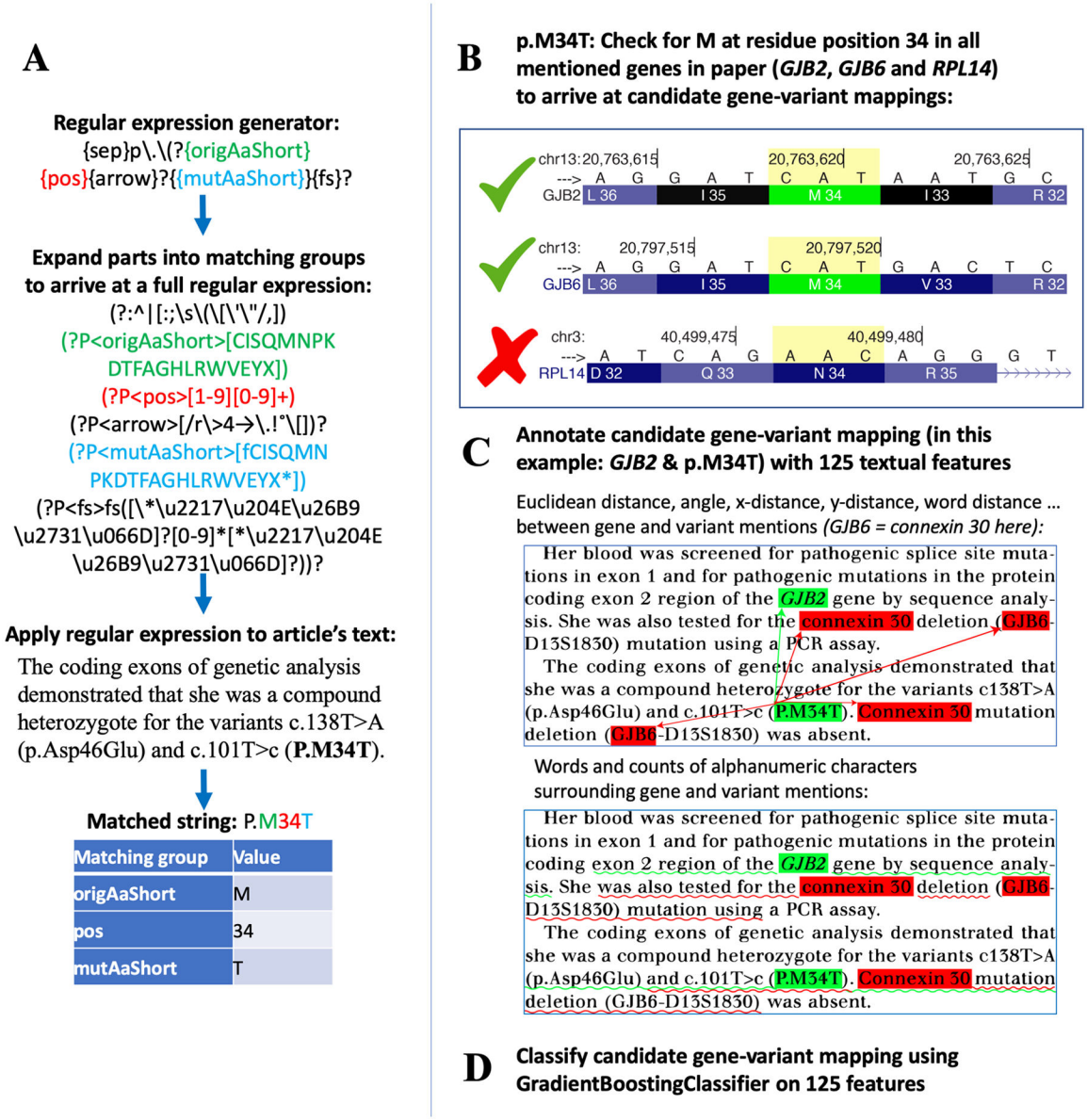

Figure 2. Automatic conversion of variant mentions to genomic coordinates from full-text literature.

(A) AVADA uses a regular expression to detect a variant mention (e.g., p.M34T) in the full text of an article. The position of the variant in the transcript (34), reference (M) and alternative alleles (T) are parsed using the regular expression. **(B)** AVADA detects mentioned genes in the article using a list of gene names and synonyms, and the help of a classifier that decides if recognized words are indeed a gene mention. The variant description detected in step A forms gene-variant candidate mappings with those genes that have the reference "M" at amino acid number 34. **(C)** Gene-variant candidate mappings (variant=p.M34T and gene=*GJB2* in this example, highlighted in green) are associated with 125 numerical features based on the relative positions of the closest mention of the candidate gene to the variant mention, information about the candidate gene's importance in the article, and words and characters surrounding the gene and variant mentions and nearby gene mentions (the

latter highlighted in red; see Supplementary Methods). **(D)** A machine learning classifier (implemented as a Gradient Boosting classifier) takes these 125 features as input and returns a score between 0 and 1 indicating the classifier's assessment of whether the variant actually refers to the given candidate gene. If the classifier returns a score greater than 0.9, the gene-variant candidate mapping is transformed to Variant Call Format (chromosome, position, reference and alternative allele) and entered into the AVADA database. In the present example, AVADA correctly decides that p.M34T only maps to *GJB2* and not connexin 30 (encoded by the gene *GJB6*). Example taken from PubMed ID 23808595.

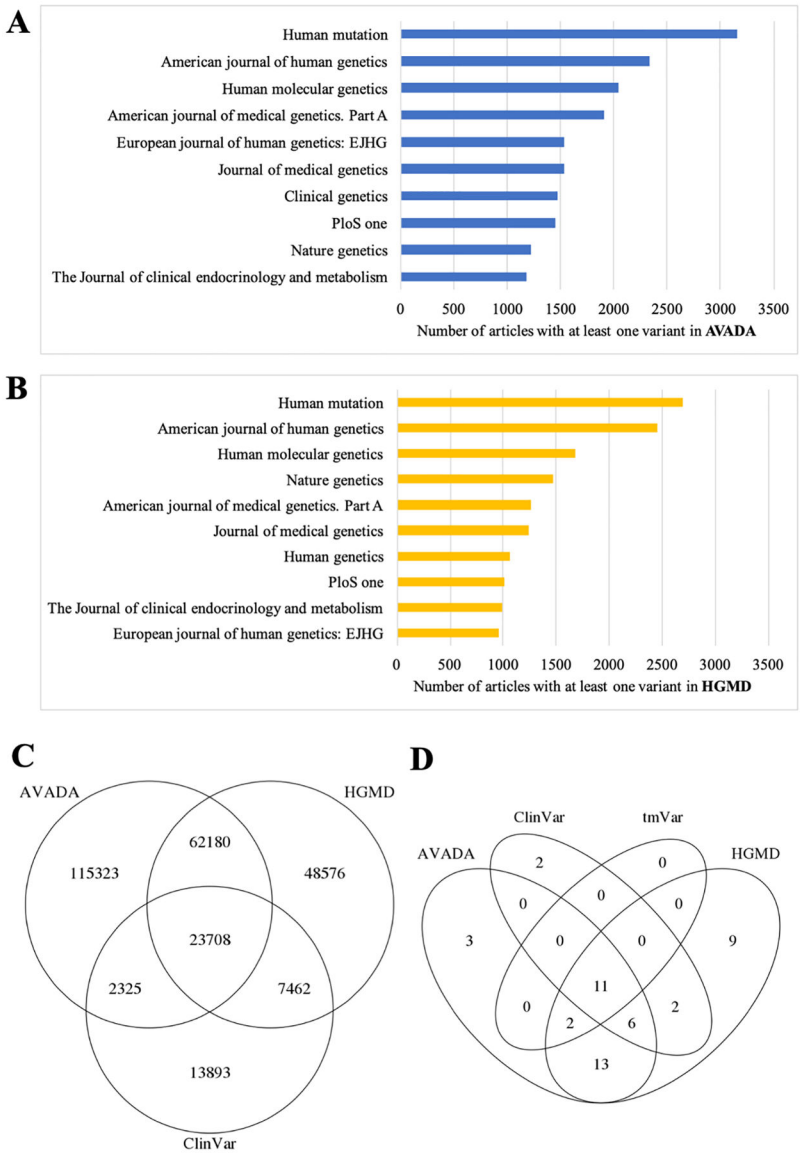

Figure 3. Automatic variant retrieval results.

(A) Top ten journals in AVADA. AVADA retrieved variants from 3,159 articles in “Human Mutation”, 2,330 articles in “American Journal of Human Genetics”, 2,042 articles in “Human Molecular Genetics” etc. **(B)** Top ten journals in all of HGMD. Similar to AVADA, the top three journals are “Human Mutation”, the “American Journal of Human Genetics”, and “Human Molecular Genetics”. Reassuringly, the two lists share 9 of the top 10 journals even though HGMD is manually curated whereas AVADA automatically retrieves variant evidence, but does not validate it. **(C)** (Unvalidated) AVADA variants intersected with all curated disease-causing variants in HGMD (“DM” variants only) and ClinVar (“likely/pathogenic” variants only). AVADA retrieves 85,888 variants also in the HGMD set (subset to disease-causing variants) and 26,033 variants also in the ClinVar set (subset to pathogenic and likely pathogenic variants). **(D)** AVADA’s potential value in patient diagnosis. We enumerate the number of patient diagnostic variants found in each of four databases, for 245

Deciphering Developmental Disorders (DDD) diagnosed patients. Curated HGMD and ClinVar (predating the DDD publication) are subset to disease-causing ("DM"), and "likely/pathogenic", respectively. For tmVar and AVADA, we manually validated all diagnostic evidence shown. AVADA completely subsumes and almost triples abstract-based tmVar. And while ClinVar alone implicates 21 diagnostic variants, AVADA offers unvalidated evidence for an additional 27 variants, of which 18 are valid, virtually doubling ClinVar's reach.

Table 1.

Examples of HGVS or common HGVS-like variant descriptions.

Each row contains examples of a disease-causing variant description in HGVS or a common HGVS-like notation. Each of these variant descriptions describes a single genetic event causing a disease, usually by giving at least the position of the change in the gene's transcript, an optional reference sequence and a novel alternative (mutated) sequence. All given variants can be described using multiple commonly used notations. Examples of alternatives to the notations are shown in the left-hand column. Transcript identifiers for variant descriptions, which greatly ease the mapping of variants to reference genome positions, are often omitted by article authors and must therefore be inferred by automated methods like AVADA. The more papers are written fully adhering to HGVS guidelines, the easier this task will become.

HGVS(-like) variant descriptions (alternatives describing same genetic event)	Explanation of HGVS variant description	Disease caused by variant (cited PMIDs' full text uses all variant notations shown in left column)
NM_175073.2 593C>T (NP_778243.1 p.A198V)	**DNA single nucleotide substitution** reference C replaced by alternative T at position 593 in the transcript NM_175073.2	Cerebellar ataxia with oculomotor apraxia type 1 (14506070, 16159533)
NM_006005.3 460+1G→A (NM_006005.3 IVS4+1G>A)	**Splicing variant** reference G replaced by alternative A at the genomic position 1 basepairs downstream of the 3' end of the exon of transcript NM_006005.3 that ends at position 460	Wolfram syndrome (11317350, 12955714)
NP_000518.1 p.Asp221Thrfs*44 (NM_000527.4 c.660delC; NP_000518.1 p.Pro220Profsx45)	**Protein frameshift variant** reference aspartic acid at residue number 221 in transcript NP_000518.1 impacted by an indel resulting in an alternative threonine, with the rest of the protein being frameshifted, introducing a stop codon 44 amino acid residues downstream of residue number 221	Familial hypercholesterolaemia (17539906, 22883975)

Table 2.

Percentage of rare variant types in AVADA, HGMD and ClinVar.

The table shows fractions of variant types in roughly synchronized time-wise AVADA, HGMD, and ClinVar, each subset to rare variants (<= 3% allele frequency in a large healthy control cohort[2]) of the shown variant types. Despite being based purely on automatic Natural Language Processing methods, AVADA (unvalidated) variant type fractions are always within the range between all variants deposited in manually curated HGMD and ClinVar at roughly synchronized timestamps ± 1%.

Variant type	AVADA	HGMD	ClinVar
stoploss	0.08%	0.14%	0.10%
nonframeshift indel	1.87%	3.12%	2.62%
splicing	4.05%	7.35%	3.82%
stopgain	12.37%	13.87%	8.58%
frameshift	14.60%	22.16%	11.22%
missense	67.03%	53.36%	73.67%