

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/125422/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Birgmeier, Johannes, Deisseroth, Cole A., Hayward, Laura E., Galhardo, Luisa M. T., Tierno, Andrew P., Jagadeesh, Karthik A., Stenson, Peter D., Cooper, David N. , Bernstein, Jonathan A., Haeussler, Maximilian and Bejerano, Gill 2020. AVADA: toward automated pathogenic variant evidence retrieval directly from the full-text literature. *Genetics in Medicine* 22 (2) , pp. 362-370. 10.1038/s41436-019-0643-6

Publishers page: <http://dx.doi.org/10.1038/s41436-019-0643-6>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.



1 **Title**

2 AVADA Improves Automated Genetic Variant Database
3 Construction Directly from Full Text Literature

4 **Authors**

5 Johannes Birgmeier¹, Andrew P. Tierno¹, Peter D. Stenson², Cole A. Deisseroth¹,
6 Karthik A. Jagadeesh¹, David N. Cooper², Jonathan A. Bernstein³, Maximilian Haeussler⁴
7 and Gill Bejerano^{1,3,5,6*}

8 **Affiliations**

9 ¹ Department of Computer Science, Stanford University, Stanford, California 94305, USA

10 ² Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, UK

11 ³ Department of Pediatrics, Stanford School of Medicine, Stanford, California 94305, USA

12 ⁴ Santa Cruz Genomics Institute, MS CBSE, University of California Santa Cruz, California
13 95064, USA

14 ⁵ Department of Developmental Biology, Stanford University, Stanford, California 94305, USA

15 ⁶ Department of Biomedical Data Science, Stanford University, Stanford, California 94305, USA

16 * To whom correspondence should be addressed, bejerano@stanford.edu

17 **Abstract**

18 The primary literature on human genetic diseases with high penetrance includes descriptions of
19 large numbers of pathogenic variants that can be essential for clinical diagnosis. Variant
20 databases such as ClinVar and HGMD collect pathogenic variants by manual curation of either
21 voluntary submissions or the published literature. AVADA (Automatically curated Variant
22 Database) represents the first automated tool designed to construct a comprehensive database of
23 highly penetrant genetic variants directly from full-text articles about human genetic disease.

24 AVADA was able to automatically curate almost 60% of the pathogenic variants deposited in
25 HGMD, over 4 times more than approaches parsing only PubMed abstracts. AVADA also
26 contains more than 60,000 pathogenic variants that are in HGMD, but not in ClinVar. Despite
27 being fully automated, 9 of AVADA's top 10 yielding journals are shared with HGMD's top 10,
28 and its mutation type distribution strongly resembles that of both HGMD and ClinVar. We
29 demonstrate the utility of AVADA in clinical practice on a cohort of 245 patients with already
30 diagnosed genetic diseases. Out of 260 causative variants originally reported for these patients,
31 AVADA contained 38 variants described in the literature prior to publication of the patient
32 cohort, compared to 43 using HGMD, 20 using ClinVar and only 13 (wholly subsumed by
33 AVADA's) using an automated abstracts-only based approach. The database of automatically
34 curated variants will be made available upon publication at
35 <http://bejerano.stanford.edu/AVADA>.

36 **Introduction**

37 Rare genetic diseases affect 7 million infants born every year worldwide¹. Exome or genome
38 sequencing is now entering clinical practice in relation to the identification of molecular causes
39 of highly penetrant genetic diseases, and in particular Mendelian disorders (genetic diseases
40 caused by mutations in a single gene²⁻⁵). In a Mendelian context, typically one or two of the
41 patient's genetic variants in a single gene are causative of the patient's disease. After following
42 standard variant filtering procedures, a typical singleton patient exome contains 200-500 rare
43 functional variants⁶. Identifying causative variants is therefore very time-consuming, as
44 investigating each variant and deciding whether or not it is causative can take up to an hour
45 [GILL: This sounds like best practice to me! Surely most labs would take much longer]⁷. Various
46 approaches are in development to accelerate this process⁸⁻¹¹. Identifying causative variants can
47 be greatly accelerated if the patient's genome contains a previously reported pathogenic variant
48 that partly or fully explains their phenotype. The American College of Medical Genetics
49 (ACMG) guidelines for the interpretation of sequence variants recommend variant annotation
50 using databases of reported pathogenic variants¹².

51 The rapidly growing literature on human genetic diseases¹³, the costly process of manual variant
52 curation¹⁴, and improved computational access to the full text of primary literature^{15,16} serve to

53 incentivize automatic variant curation. Creating a variant database from the primary literature
54 involves finding variant descriptions (such as “c.123A>G”), linking them to a transcript of the
55 correct gene mention, and converting them to genomic coordinates (chromosome, position,
56 reference and alternative alleles) so they can be readily intersected with any patient variants.
57 Previous work on automatic variant discovery in the literature has largely focused on finding
58 variant descriptions in paper titles and abstracts with high accuracy without converting the
59 discovered variants to genomic coordinates^{17–23}. Previous automatic variant curation tools have
60 focused on mapping variant mentions to dbSNP²⁴ variant identifiers (rsIDs). Mapping textual
61 variant descriptions directly to reference genome coordinates requires significant effort, and has
62 thus far largely been left to manually curated databases such as HGMD²⁵ and ClinVar²⁶, which
63 devote many thousands of wo/man-hours to the task of collecting genetic variants from either the
64 scientific literature or clinical laboratories.

65 We posed the question as to whether manual variant curation to genome coordinates could be
66 accelerated with the help of machine learning approaches by first training an automatic curation
67 system on a sample of manually curated variants (from ClinVar and HGMD), and then applying
68 the trained system to the entire body of PubMed indexed literature for automatic curation of
69 published variants. AVADA (Automatically curated Variant Database), our automated variant
70 extractor, identifies variants in genetic disease literature and converts all detected variants into a
71 database of genomic (hg19) coordinates, reference and alternative alleles. We show that
72 AVADA improves on the state of the art in automated variant extraction, by comparing it to
73 tmVar 2.0²⁷, a best-in-class tool used to harvest variants from PubMed abstracts. Combining the
74 free ClinVar and AVADA variant databases, we find that we can recover a significant fraction of
75 diagnostic disease-causing variants in a cohort of 245 patients with Mendelian diseases.

76 **Materials and Methods**

77 **Identification of relevant literature**

78 PubMed is a database containing titles and abstracts of biomedical articles, only a subset of
79 which contain descriptions of variants that cause human genetic disease. A document classifier is
80 a machine learning classifier that takes as its input arbitrary text and classifies it as “positive”
81 (here, meaning an article about genetic disease) or “negative” (otherwise). We trained a scikit-

82 learn²⁸ LogisticRegression²⁹ classifier to identify relevant documents using positive input texts
83 (titles and abstracts of articles cited in OMIM³⁰ and HGMD²⁵) and negative input texts (random
84 titles and abstracts from PubMed). Machine learning classifiers take as input a real-valued vector
85 (the “feature vector”) describing the input numerically. Input texts were converted into a feature
86 vector by means of a scikit-learn CountVectorizer followed by a TF-IDF³¹ transformer (an
87 operation that converts input text to a feature vector based on the frequency of words in input
88 documents). After training the title/abstract document classifier, we applied it to all 25,793,020
89 titles and abstracts in PubMed to identify articles that might be relevant to the diagnosis of
90 genetic diseases. Full text PDFs of relevant articles were then downloaded and converted to text
91 using pdftotext³² version 0.26.5. Because identifying potentially relevant articles based upon title
92 and abstract alone often yields articles whose full text does not turn out to be relevant for the
93 diagnosis of genetic diseases, we subsequently trained a full-text scikit-learn LogisticRegression
94 classifier to classify downloaded full-text documents as “relevant” or “irrelevant” based upon the
95 article’s full text. As with the title/abstract classifier, full text documents were converted to a
96 feature vector by means of a CountVectorizer followed by a TF-IDF transformer. Filtering full-
97 text articles for relevance resulted in a subset of downloaded articles more relevant to the
98 diagnosis of genetic disease (Supplemental Methods). A total of 133,410 articles were
99 downloaded and subsequently classified as relevant to the diagnosis of human genetic diseases
100 based on the articles’ full text. We refer to this set of articles as the “AVADA full-text articles”
101 (Figure 1).

102 **Variant and gene mention detection**

103 In order to extract genetic variants from the full-text articles about human genetic disease and
104 convert them to genomic coordinates, it is necessary to detect both mentions of genes and variant
105 descriptions in articles about genetic disease. Extracting variant descriptions alone does not
106 suffice, because variant descriptions in HGVS notation, such as “c.123A>G”, can only be
107 converted to genomic coordinates if a transcript of the gene that the variant refers to is identified
108 (Table 1).

109 AVADA extracts gene mentions from articles’ full text using a custom-built database of gene
110 names containing gene name entries from the HUGO Gene Nomenclature Committee (HGNC)
111 and UniProt databases. Gene and protein names from these were matched case-insensitive to

112 word groups of length 1-8 in the document to identify gene mentions. To identify variant
113 mentions, we manually developed a set of 47 regular expressions based on commonly observed
114 HGVS-like variant notations in articles about human genetic disease (Supplemental Methods,
115 Supplemental Table S1 and Figure 2A). At this step, we refer to every string that matches one of
116 the 47 regular expressions as a “variant description”. In the AVADA full-text articles, variant
117 descriptions in 92,436 articles were identified, with a mean of 11.1 variant descriptions per
118 article (Figure 1).

119 **Mentioned genes form gene-variant candidate mappings with all mentioned variants** 120 **that “fit” the gene**

121 Having identified gene mentions and variant descriptions in text, it is now necessary to link
122 variant descriptions with the genes that they refer to. Articles often mention variant descriptions
123 without explicitly stating to which gene each variant description maps. The gene to which each
124 variant description maps can be inferred by expert readers of the article. However, an automatic
125 algorithm cannot easily infer to which gene a variant description maps, because gene mention
126 and variant description do not necessarily occur in the same sentence or even the same paragraph
127 or page.

128 To identify which variant description maps to which mentioned gene in the article, AVADA first
129 forms so-called *gene-variant candidate mappings* between each variant description and each
130 mentioned gene if the variant appears to “fit” at least one RefSeq³³ transcript of the gene. Given
131 an extracted variant description “c.123A>G”, the variant description forms gene-variant
132 candidate mappings with all mentioned genes that have an “A” at coding position 123 of at least
133 one transcript (Supplemental Methods and Figure 2B). A variant description can form gene-
134 variant candidate mappings with multiple genes, which are filtered in the next step. Gene-variant
135 candidate mappings are converted to genomic coordinates in the hg19/GRCh37 reference
136 assembly. In the AVADA full-text articles, an extracted variant description initially mapped to a
137 mean of 4.6 different genomic coordinates (Figure 1).

138 **Machine learning classifier selects the correct gene-variant mapping out of multiple** 139 **gene-variant candidate mappings**

140 AVADA uses a machine learning framework to decide which gene-variant candidate mappings
141 are likely to be correct. The machine learning classifier is a scikit-learn²⁸

142 GradientBoostingClassifier³⁴. The training set for the classifier comprised positive gene-variant
143 mappings curated from the literature in ClinVar, and a set of negative gene-variant mappings
144 created by assigning variants from the positive training set to genes mentioned in the paper to
145 which they did not map. Each gene-variant mapping was converted to a feature vector, based
146 upon which the classifier decided if the gene-variant candidate mapping was true or false. The
147 feature vector included the Euclidean distance between the 2D coordinates (consisting of page
148 number, x and y coordinates of a mention) of the closest mentions of the variant and the gene in
149 the PDF, the number of words between variant and gene mentions, the number of short
150 “stopwords” (like “and”, “or”, “of”, ...) around gene and variant mentions, and a number of
151 other textual features containing information about the relationship between gene and variant
152 mentions (Supplemental Methods and Figure 2C; performance analyzed below).

153 The classifier successfully reduced 4.6 candidate gene-variant mappings per variant description
154 to a mean of 1.2 genomic positions in the AVADA full-text articles (Supplemental Methods and
155 Figures 1, 2D).

156 **Results**

157 **AVADA identified 203,608 variants in 5,827 genes from 61,117 articles**

158 A total of 61,117 articles made it into the final AVADA database, with a mean of 8.8 identified
159 variant descriptions per article. From these articles, 203,608 distinct genetic variants in 5,827
160 genes were automatically curated (Figure 1), comprising a variety of different variant types in a
161 distribution strikingly similar to that of manually curated HGMD and ClinVar: for each of 6
162 categories of variant (stoploss, nonframeshift, splicing, stopgain, frameshift, missense), the
163 fraction of variants AVADA extracted are between the fraction of the respective category in
164 HGMD and ClinVar $\pm 1\%$ (Table 2). The articles used to construct AVADA are from a variety of
165 journals, which are similar to the journals targeted by HGMD to curate its variants (9 out of the
166 top 10 journals being the same between AVADA and HGMD; Figure 3A,B).

167 Each variant, defined by chromosome, position, reference and alternative allele, is annotated
168 with: PubMed ID(s) of publications where this variant was extracted from; HUGO Gene
169 Nomenclature Committee³⁵ (HGNC) gene symbol, Ensembl ID³⁶, and Entrez ID³⁷ of the gene in
170 which the variant is found, the **inferred variant effect** [GILL: what do you mean?] (e.g.,

171 “missense”), the RefSeq ID of the gene’s transcript to which the variant was mapped (e.g.,
172 NM_005101.3), and the exact variant description from the original article (e.g., “c.163C.T”). The
173 latter allows clinicians to later rapidly locate mentions of this variant within the body of the
174 article.

175 **AVADA is 72% precise**

176 To estimate the precision (the fraction of extracted variants that are correctly extracted), 100
177 distinct random variants in AVADA were manually examined. AVADA variants were manually
178 counted as true extractions whenever the scientist reading the paper (using all lines of evidence
179 in the paper such as Sanger sequencing reads, UCSC genome browser shots etc.) independently
180 mapped the paper’s variant mention to the same genomic coordinates as AVADA. Of the 100
181 distinct random variants, 72% were extracted and mapped to the correct genomic position
182 without error by AVADA (Supplemental Table S2).

183 **AVADA recovers nearly 60% of disease-causing HGMD variants directly from the** 184 **primary literature**

185 We compared AVADA to HGMD and ClinVar versions with synchronized time stamps
186 (Supplemental Methods). 85,888 AVADA variants coincided with variants identified in HGMD
187 and marked as disease-causing (“DM”), corresponding to 61% of all disease-causing variants in
188 HGMD. From this set of 85,888 AVADA variants, we selected 100 random variants and
189 manually verified that the genomic coordinates (chromosome, position, reference and alternative
190 alleles) were correctly extracted and the variant was reported as disease-causing. Of the 100
191 variants examined, 97% fulfilled these criteria (Supplemental Table S3). Thus, we infer that
192 AVADA contains 59% of all disease-causing variants identified by HGMD.

193 We compared AVADA’s performance to the best previously published automatic variant
194 curation tool, tmVar 2.0, which attempts to map variant mentions in all PubMed abstracts to
195 dbSNP identifiers (rsIDs). tmVar extracted only 19,424 disease-causing HGMD variants, or 14%
196 of HGMD (Supplemental Figure 1 and Figure 3C).

197 Considering only single nucleotide variants (SNVs), the largest class of known pathogenic
198 variant, AVADA contains 70% of all DM SNVs in HGMD, of which an estimated 97% were
199 extracted correctly. Similarly, AVADA contains 55% of all likely pathogenic or pathogenic
200 variants in ClinVar (clinical significance level 4 or 5) and 62% of pathogenic or likely

201 pathogenic SNVs in ClinVar. tmVar 2.0 extracted only 13,664, or 31%, of pathogenic or likely
202 pathogenic variants in ClinVar.

203 Strikingly, AVADA contains 63,521 variants that are in HGMD (“DM” only) but not in ClinVar
204 (clinical significance level 4 or 5). An analysis of a representative subset of 100 of the remaining
205 115,612 variants that were extracted by AVADA, but not reported as disease-causing in either
206 HGMD or ClinVar, revealed them to be mostly benign or incorrectly extracted variants
207 (Supplemental Table S4).

208 **Diagnosis of patients with Mendelian diseases using AVADA**

209 We analyzed the accuracy of patient variant annotation with AVADA, tmVar, ClinVar and
210 HGMD using a set of 245 patients from the Deciphering Developmental Disorders³⁸ (DDD)
211 study, harboring 260 causative variants reported by the original DDD study. De-identified DDD
212 data were obtained from EGA³⁹ study number EGAS00001000775 (Supplemental Methods).
213 The DDD study is a large-scale sequencing study in which children affected with developmental
214 disorders were sequenced with a view to attempting a diagnosis. Disease-causing variants
215 reported in DDD were obtained from Supplemental Table 4 in reference³⁸.

216 **Sensitivity of variant annotation using AVADA, tmVar, HGMD and ClinVar**

217 The more complete a variant database is, the higher its sensitivity when annotating patient
218 genomes and the higher the likelihood of finding a causative variant in the patient’s genome. We
219 determined how many of the 260 causative DDD variants were found in AVADA, tmVar,
220 HGMD and ClinVar. The more causative variants are found in a database, the more rapidly a
221 given patient can be diagnosed. For the DDD patient variant annotation comparison, we subset
222 AVADA and tmVar 2.0 to reference only articles until 2014 (before the publication of the DDD
223 study), **HGMD use only variants added until 2014** [GILL: not sure if this makes sense. used only
224 HGMD variants....?], and took the latest ClinVar version from 2014 (ClinVar version
225 20141202).

226 Of 260 different causative variants reported by the DDD study, a total of 45 variants were found
227 by AVADA in the scientific literature. For each of these variants, all articles from which the
228 variant was extracted were manually inspected. If at least one article was found in which the
229 variant’s genomic coordinates (chromosome, position, reference and alternative allele) were

230 correctly extracted, the variant was reported as causative and the article did not cite the DDD
231 study, the variant was counted as correct. 38 of the 45 variants found by AVADA fulfilled these
232 criteria (Supplemental Table S5).

233 Only 20 variants reported to be causative by the DDD study were listed in ClinVar and ascribed
234 a pathogenicity level such as “pathogenic” or “likely pathogenic”. 43 variants were in HGMD,
235 reported as “DM” (disease-causing). tmVar 2.0 contained 13 causative variants (Supplemental
236 Table S6). AVADA and ClinVar together contained 41 causative variants. All of tmVar’s
237 variants were either in AVADA or ClinVar. Thus, combining the free variant databases AVADA
238 and ClinVar resulted in our annotating almost as many causative variants as are listed in HGMD.
239 Combining all three databases yielded 51 variants (Figure 3D).

240 **Discussion**

241 We present AVADA, an automated approach to constructing a highly penetrant variant database
242 from full-text articles about human genetic diseases. AVADA automatically curated nearly a
243 hundred thousand disease-causing variants from tens of thousands of downloaded and parsed
244 full-text articles. All AVADA mutations are stored in a Variant Call Format⁴⁰ (VCF) file that
245 includes the chromosome, position, reference and alternative alleles, variant strings as reported
246 in the original article, and PubMed IDs of the original articles mentioning the variants. AVADA
247 recovers nearly 60% of all disease-causing variants deposited in HGMD at a fraction of the cost
248 of constructing a manually curated database⁴¹, over 4 times as many as the tmVar 2.0 database
249 that relies on PubMed abstracts, and maps only to dbSNP rsIDs. From a cohort of 245 previously
250 diagnosed patients from the Deciphering Developmental Disorders (DDD) project, AVADA
251 pinpoints 38 DDD-reported disease-causing variants, fewer than HGMD (43) but almost twice as
252 many as ClinVar (20) and almost three times as many as tmVar 2.0 (13), showing that this new
253 resource will be useful in clinical practice. Combining the free variant databases AVADA and
254 ClinVar recovers 41 diagnostic variants.

255 Multiple lessons were learned from AVADA. First, curating variants from full text articles
256 scattered between dozens of publishers’ web portals is worth the extra effort. However, while
257 gene to variant linking is often relatively simple in the context of an abstract, this task is much
258 more challenging in the context of sprawling full texts that may well discuss many additional

259 genes beyond the causal few. A two-pronged approach is therefore necessary to further improve
260 AVADA's precision. First, our ability to link variants to the correct transcripts and genes can be
261 improved. Second, non-pathogenic mentioned variants need to be better distinguished from
262 pathogenic mentioned variants. Implementing patterns for more exotic variant notations and
263 parsing supplements of articles would improve sensitivity, but would decrease precision.

264 AVADA curates variants without costly human input and can be re-run continually to discover
265 newly reported variants without incurring significant additional cost. While the approach cannot
266 currently replicate manual curation efforts, it is nevertheless well suited to supporting the work
267 of manual curators in improving and extending existing variant databases. Blending the AVADA
268 automatic variant curation approach with manual verification should facilitate rapid variant
269 classification⁴² and the cost-effective annotation of patient variants.

270 Publishers could help to further improve the automatic variant curation process by supplying
271 database curation tools with simpler, stable programmatic access to full text and supplemental
272 data of appropriate articles, a win-win step that would lead to both better variant databases, and
273 increase the circulation of articles among their target audience. Requiring authors to abide by
274 strict HGVS notation would also help. Moreover, the approach presented here can be extended to
275 the automatic curation of genetic variants [GILL: HGVS is not appropriate for animal models or
276 non-model organisms] or related notation from other valuable modalities beyond human patients,
277 such as animal models, cell lines, or non-model organisms with reference genomes and
278 transcripts. The approach described could therefore support the rapid and cost-effective creation
279 and upkeep of multiple different variant databases beyond human genetic diseases⁴³ directly
280 from the primary literature [GILL: You could also mention somatic mutations in cancer genes?].

281 By comprehensively annotating each variant with information from the original articles (such as
282 the originally reported variant string), AVADA enables rapid re-discovery and verification of a
283 large fraction of previously reported variants in the scientific literature. AVADA shows that
284 automatic variant curation from the literature is feasible and useful with regard to accelerating
285 the creation of genetic variant databases that enable rapid diagnostics in a clinical setting.

286 Previously, manual curation efforts such as HGMD²⁵ have demonstrated the power of systematic
287 manual curation of pathogenic variants from the primary literature. Combining automatic
288 curation approaches like AVADA with manual curation will enable the rapid construction of

289 clinically useful variant databases from the primary literature enabling both rapid diagnosis⁴² and
290 reanalysis¹³.

291 **Supplemental Data**

292 Supplemental Methods describe the AVADA variant curation process in detail. Supplemental
293 Tables S1-S8 contain additional data referenced in main text and Supplemental Methods.

294 **Acknowledgments**

295 This work was funded in part by a Bio-X Stanford Interdisciplinary Graduate Fellowship to J.B.;
296 by grants EMBO ALTF292-2011 and NIH/NHGRI 5U41HG002371-15 to M.H.; and by
297 DARPA, the Stanford Pediatrics Department, a Packard Foundation Fellowship, a Microsoft
298 Faculty Fellowship and the Stanford Data Science Initiative to G.B. We would like to thank the
299 European Genome-Phenome Archive³⁹ (EGA) and the Deciphering Developmental Diseases³⁸
300 (DDD) project. The DDD study presents independent research commissioned by the Health
301 Innovation Challenge Fund [grant number HICF-1009-003], a parallel funding partnership
302 between the Wellcome Trust and the Department of Health, and the Wellcome Trust Sanger
303 Institute [grant number WT098051]. The views expressed in this publication are those of the
304 author(s) and not necessarily those of the Wellcome Trust or the Department of Health. The
305 study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge
306 South REC, and GEN/284/12 granted by the Republic of Ireland REC). Deidentified DDD data
307 was obtained through EGA. The research team acknowledges the support of the National
308 Institute for Health Research, through the Comprehensive Clinical Research Network.

309 **Author Contributions**

310 J.B. and M.H. wrote software to map variants to the reference genome using a database of
311 RefSeq transcripts. A.P.T. verified AVADA-extracted variants. J.B. wrote the machine learning
312 classifiers and performed performance evaluations. J.B., M.H., A.P.T., and G.B. wrote the
313 manuscript. C.D. and K.A.J. downloaded and processed DDD data. P.D.S. and D.N.C. created
314 HGMD and helped with manual variant inspection. J.A.B. provided guidance on clinical aspects
315 of study design, testing set construction and interpretation of results. G.B. supervised the project.

316 All authors read and commented on the manuscript.

317 The authors declare no conflicts of interest.

318 **Web resources**

319 All code for automatic variant curation with AVADA, as well as the automatically curated
320 variants database presented here, will be available upon publication for non-commercial use at
321 <http://bejerano.stanford.edu/AVADA>.

322 **References**

323 1. Church, G. (2017). Compelling reasons for repairing human germlines. *N. Engl. J. Med.* *377*,
324 1909–1911.

325 2. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T.,
326 Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel
327 sequencing of 12 human exomes. *Nature* *461*, 272–276.

328 3. Simpson, M.A., Irving, M.D., Asilmaz, E., Gray, M.J., Dafou, D., Elmslie, F.V., Mansour, S.,
329 Holder, S.E., Brain, C.E., Burton, B.K., et al. (2011). Mutations in *NOTCH2* cause Hajdu-
330 Cheney syndrome, a disorder of severe and progressive bone loss. *Nat. Genet.* *43*, 303–305.

331 4. Ng, S.B., Buckingham, K.J., Lee, C., Bigham, A.W., Tabor, H.K., Dent, K.M., Huff, C.D.,
332 Shannon, P.T., Jabs, E.W., Nickerson, D.A., et al. (2010). Exome sequencing identifies the cause
333 of a mendelian disorder. *Nat. Genet.* *42*, 30–35.

334 5. Jones, W.D., Dafou, D., McEntagart, M., Woollard, W.J., Elmslie, F.V., Holder-Espinasse,
335 M., Irving, M., Saggart, A.K., Smithson, S., Trembath, R.C., et al. (2012). *De novo* mutations in
336 *MLL* cause Wiedemann-Steiner syndrome. *Am. J. Hum. Genet.* *91*, 358–364.

337 6. Jagadeesh, K.A., Wenger, A.M., Berger, M.J., Guturu, H., Stenson, P.D., Cooper, D.N.,
338 Bernstein, J.A., and Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain
339 significance in clinical exomes at high sensitivity. *Nat. Genet.* *48*, 1581–1586.

340 7. Dewey, F.E., Grove, M.E., Pan, C., Goldstein, B.A., Bernstein, J.A., Chaib, H., Merker, J.D.,
341 Goldfeder, R.L., Enns, G.M., David, S.P., et al. (2014). Clinical interpretation and implications
342 of whole-genome sequencing. *JAMA* *311*, 1035.

343 8. Smedley, D., Jacobsen, J.O.B., Jäger, M., Köhler, S., Holtgrewe, M., Schubach, M., Siragusa,
344 E., Zemojtel, T., Buske, O.J., Washington, N.L., et al. (2015). Next-generation diagnostics and
345 disease-gene discovery with the Exomiser. *Nat. Protoc.* *10*, 2004–2015.

- 346 9. Jagadeesh, K.A., Birgmeier, J., Guturu, H., Deisseroth, C.A., Wenger, A.M., Bernstein, J.A.,
347 and Bejerano, G. (2018). Phrank measures phenotype sets similarity to greatly improve
348 Mendelian diagnostic disease prioritization. *Genet. Med. Off. J. Am. Coll. Med. Genet. ePub.*
- 349 10. Birgmeier, J., Haeussler, M., Deisseroth, C.A., Jagadeesh, K.A., Ratner, A.J., Guturu, H.,
350 Wenger, A.M., Stenson, P.D., Cooper, D.N., Re, C., et al. (2017). AMELIE accelerates
351 Mendelian patient diagnosis directly from the primary literature. *BioRxiv* 171322.
- 352 11. Deisseroth, C.A., Birgmeier, J., Bodle, E.E., Bernstein, J.A., and Bejerano, G. (2018).
353 ClinPhen extracts and prioritizes patient phenotypes directly from medical records to accelerate
354 genetic disease diagnosis. *BioRxiv* 362111.
- 355 12. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde,
356 M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of
357 sequence variants: a joint consensus recommendation of the American College of Medical
358 Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am.*
359 *Coll. Med. Genet.* *17*, 405–424.
- 360 13. Wenger, A.M., Guturu, H., Bernstein, J.A., and Bejerano, G. (2016). Systematic reanalysis of
361 clinical exome data yields additional diagnoses: implications for providers. *Genet. Med.* *19*, 209–
362 214.
- 363 14. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J.,
364 Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen--the
365 Clinical Genome Resource. *N. Engl. J. Med.* *372*, 2235–2242.
- 366 15. Van Noorden, R. (2013). Text-mining spat heats up. *Nature* *495*, 295.
- 367 16. Westergaard, D., Stærfeldt, H.-H., Tønsberg, C., Jensen, L.J., and Brunak, S. (2018). A
368 comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus
369 their corresponding abstracts. *PLoS Comput. Biol.* *14*, e1005962.
- 370 17. Jimeno Yepes, A., and Verspoor, K. (2014). Mutation extraction tools can be combined for
371 robust recognition of genetic variants in the literature. *F1000Research* *3*, 18.
- 372 18. Baker, C.J.O., and Witte, R. (2006). Mutation Mining—A Prospector’s Tale. *Inf. Syst. Front.*
373 *8*, 47–57.
- 374 19. Xuan, W., Wang, P., Watson, S.J., and Meng, F. (2007). Medline search engine for finding
375 genetic markers with biological significance. *Bioinformatics* *23*, 2477–2484.
- 376 20. Doughty, E., Kertesz-Farkas, A., Bodenreider, O., Thompson, G., Adadey, A., Peterson, T.,
377 and Kann, M.G. (2011). Toward an automatic method for extracting cancer- and other disease-
378 related point mutations from the biomedical literature. *Bioinformatics* *27*, 408–415.
- 379 21. Caporaso, J.G., Baumgartner, W.A., Randolph, D.A., Cohen, K.B., and Hunter, L. (2007).
380 MutationFinder: a high-performance system for extracting point mutation mentions from text.
381 *Bioinforma. Oxf. Engl.* *23*, 1862–1865.

- 382 22. Wei, C.-H., Harris, B.R., Kao, H.-Y., and Lu, Z. (2013). tmVar: a text mining approach for
383 extracting sequence variants in biomedical literature. *Bioinformatics* 29, 1433–1439.
- 384 23. Thomas, P., Rocktäschel, T., Hakenberg, J., Lichtblau, Y., and Leser, U. (2016). SETH
385 detects and normalizes genetic variants in text. *Bioinforma. Oxf. Engl.* 32, 2883–2885.
- 386 24. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin,
387 K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311.
- 388 25. Stenson PD, Mort M, Ball EV, Evans K, Hayden M, Heywood S, Hussain M, Phillips AD,
389 Cooper DN. (2017) The Human Gene Mutation Database: towards a comprehensive repository
390 of inherited mutation data for medical research, genetic diagnosis and next-generation
391 sequencing studies. *Hum. Genet.* 136, 665-677.
- 392 26. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J.,
393 Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically
394 relevant variants. *Nucleic Acids Res.* 44, D862-868.
- 395 27. Wei, C.-H., Phan, L., Feltz, J., Maiti, R., Hefferon, T., and Lu, Z. (2018). tmVar 2.0:
396 integrating genomic variant information from literature with dbSNP and ClinVar for precision
397 medicine. *Bioinformatics* 34, 80–87.
- 398 28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,
399 Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python.
400 *J. Mach. Learn. Res.* 12, 2825–2830.
- 401 29. Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*
402 (Springer).
- 403 30. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015).
404 OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes
405 and genetic disorders. *Nucleic Acids Res.* 43, D789-798.
- 406 31. Jurafsky, D., and Martin, J.H. (2000). *Speech and Language Processing: An Introduction to*
407 *Natural Language Processing, Computational Linguistics, and Speech Recognition* (Upper
408 Saddle River, NJ, USA: Prentice Hall PTR).
- 409 32. Poppler. <https://poppler.freedesktop.org/>.
- 410 33. O’Leary, N.A., Wright, M.W., Brister, J.R., Ciuffo, S., Haddad, D., McVeigh, R., Rajput, B.,
411 Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq)
412 database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids*
413 *Res.* 44, D733–D745.
- 414 34. Friedman, J.H. (2001). Greedy Function Approximation: A Gradient Boosting Machine.
415 *Ann. Stat.* 29, 1189–1232.

- 416 35. Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., and Bruford, E.A. (2015). Genenames.org:
417 the HGNC resources in 2015. *Nucleic Acids Res.* *43*, D1079-1085.
- 418 36. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins,
419 C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. *Nucleic Acids Res.* *44*,
420 D710-716.
- 421 37. Maglott, D., Ostell, J., Pruitt, K.D., and Tatusova, T. (2011). Entrez Gene: gene-centered
422 information at NCBI. *Nucleic Acids Res.* *39*, D52–D57.
- 423 38. Deciphering Developmental Disorders Study (2015). Large-scale discovery of novel genetic
424 causes of developmental disorders. *Nature* *519*, 223–228.
- 425 39. Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J.D., ur-Rehman, S.,
426 Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., et al. (2015). The European Genome-
427 phenome Archive of human data consented for biomedical research. *Nat. Genet.* *47*, 692–695.
- 428 40. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker,
429 R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools.
430 *Bioinformatics* *27*, 2156–2158.
- 431 41. Project Information - NIH RePORTER - NIH Research Portfolio Online Reporting Tools
432 Expenditures and Results.
- 433 42. Patel, R.Y., Shah, N., Jackson, A.R., Ghosh, R., Pawliczek, P., Paithankar, S., Baker, A.,
434 Riehle, K., Chen, H., Milosavljevic, S., et al. (2017). ClinGen Pathogenicity Calculator: a
435 configurable system for assessing pathogenicity of genetic variants. *Genome Med.* *9*, 3.
- 436 43. McMurry, J.A., Köhler, S., Washington, N.L., Balhoff, J.P., Borromeo, C., Brush, M.,
437 Carbon, S., Conlin, T., Dunn, N., Engelstad, M., et al. (2016). Navigating the phenotype frontier:
438 The Monarch Initiative. *Genetics* *203*, 1491–1495.
- 439 44. Tsao, C.Y., and Paulson, G. (2005). Type 1 ataxia with oculomotor apraxia with aprataxin
440 gene mutations in two American children. *J. Child Neurol.* *20*, 619–620.
- 441 45. Le Ber, I., Moreira, M.-C., Rivaud-Péchoux, S., Chamayou, C., Ochsner, F., Kuntzer, T.,
442 Tardieu, M., Saïd, G., Habert, M.-O., Demarquay, G., et al. (2003). Cerebellar ataxia with
443 oculomotor apraxia type 1: clinical and genetic studies. *Brain* *126*, 2761–2772.
- 444 46. Cryns, K., Sivakumaran, T.A., Van den Ouweland, J.M.W., Pennings, R.J.E., Cremers,
445 C.W.R.J., Flothmann, K., Young, T.-L., Smith, R.J.H., Lesperance, M.M., and Van Camp, G.
446 (2003). Mutational spectrum of the *WFS1* gene in Wolfram syndrome, nonsyndromic hearing
447 impairment, diabetes mellitus, and psychiatric disease. *Hum. Mutat.* *22*, 275–287.
- 448 47. Khanim, F., Kirk, J., Latif, F., and Barrett, T.G. (2001). *WFS1*/wolframin mutations,
449 Wolfram syndrome, and associated diseases. *Hum. Mutat.* *17*, 357–367.

450 48. Taylor, A., Tabrah, S., Wang, D., Sozen, M., Duxbury, N., Whittall, R., Humphries, S.E., and
451 Norbury, G. (2007). Multiplex ARMS analysis to detect 13 common mutations in familial
452 hypercholesterolaemia. *Clin. Genet.* *71*, 561–568.

453 49. Hooper, A.J., Nguyen, L.T., Burnett, J.R., Bates, T.R., Bell, D.A., Redgrave, T.G., Watts,
454 G.F., and van Bockxmeer, F.M. (2012). Genetic analysis of familial hypercholesterolaemia in
455 Western Australia. *Atherosclerosis* *224*, 430–434.

456 50. Haeussler, M. (2018). pubMunch. <https://github.com/maximilianh/pubMunch>.

457 51. The UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Res.*
458 *43*, D204–D212.

459 52. Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). PubTator: a web-based text mining tool for
460 assisting biocuration. *Nucleic Acids Res.* *41*, W518–522.

461 53. den Dunnen, J.T., Dalgleish, R., Maglott, D.R., Hart, R.K., Greenblatt, M.S., McGowan-
462 Jordan, J., Roux, A.-F., Smith, T., Antonarakis, S.E., and Taschner, P.E.M. (2016). HGVS
463 recommendations for the description of sequence variants: 2016 Update. *Hum. Mutat.* *37*, 564–
464 569.

465 54. (2018). bcftools. <https://github.com/samtools/bcftools/> (Github: “samtools”).

466 55. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic
467 variants from high-throughput sequencing data. *Nucleic Acids Res.* *38*, e164–e164.

468 56. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O’Donnell-
469 Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding
470 genetic variation in 60,706 humans. *Nature* *536*, 285–291.

471 57. 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D.,
472 Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome
473 variation from population-scale sequencing. *Nature* *467*, 1061–1073.

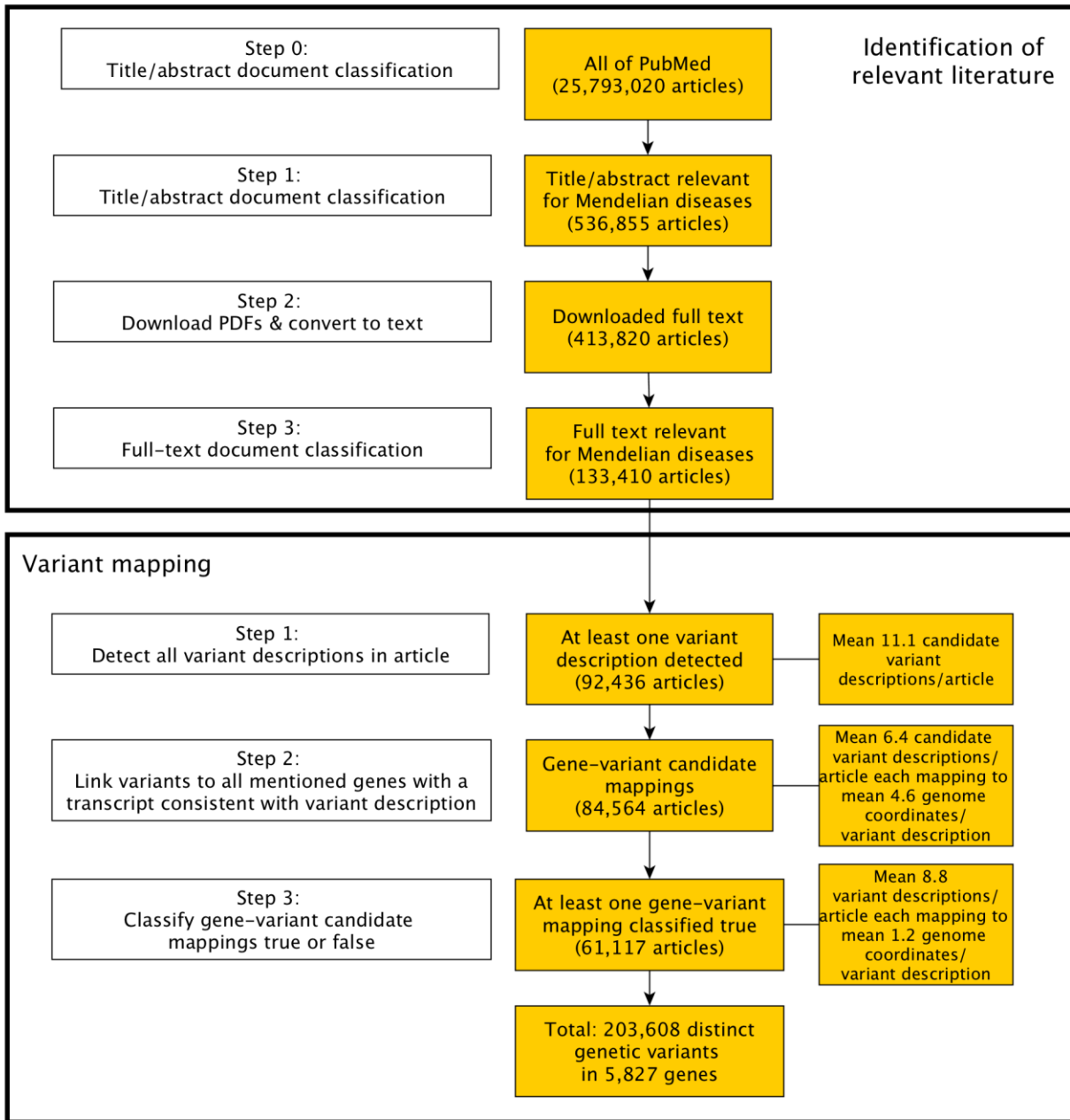
474 58. The UK10K Consortium (2015). The UK10K project identifies rare variants in health and
475 disease. *Nature* *526*, 82.

476

477

478 **Figures**

479 **Figure 1**



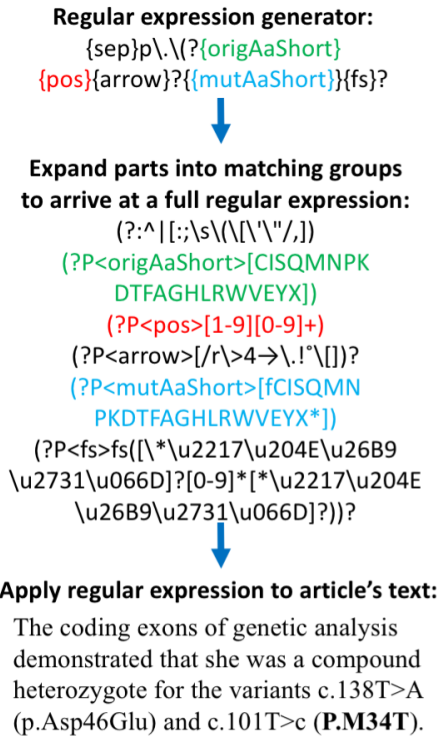
480

481 **Figure 1. Construction of the automated variant database AVADA. Identification of**
 482 **relevant literature:** Step 0: titles and abstracts of articles are downloaded from PubMed. Step
 483 1: a suitable subset of relevant literature is identified by a document classifier that classifies titles
 484 and abstracts deposited in PubMed as possibly relevant or irrelevant to genetic disease. Step 2:

485 full text PDFs of potentially relevant articles are downloaded wherever possible and converted to
486 text. Step 3: the full text of potentially relevant articles is filtered by a separate full-text
487 document classifier that again tests for relevance to genetic diseases. **Variant mapping:** Step 1:
488 gene mentions are detected using a list of gene names, and variant mentions are detected using
489 47 manually built regular expressions (Figure 2A). Step 2: a super-set of possible gene-variant
490 candidate mappings is constructed out of all mentioned variants and genes in a paper where the
491 variant appears to “fit” the gene: e.g., if a variant description is “c.123A>G”, the variant fits all
492 genes mentioned in the paper that have at least one transcript with an “A” at coding position 123
493 (Figure 2B). Step 3: A machine learning classifier using a number of textual features (Figure 2C)
494 describing the relationship between variant and gene mention in the article’s full text decides
495 which of the previously constructed gene-variant candidate mappings are true, i.e., which variant
496 actually refers to which gene (Figure 2D). AVADA extracts 203,608 distinct genetic variants in
497 5,827 genes from 61,117 articles.

498

A

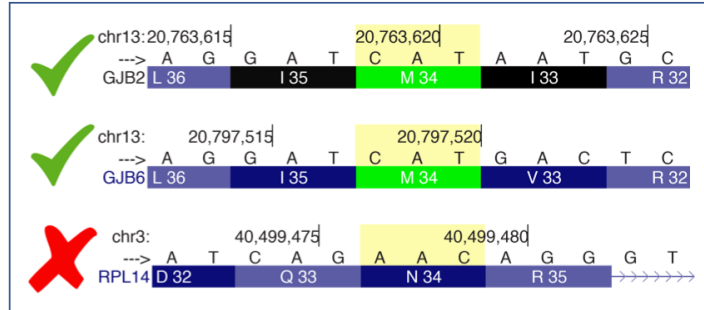


Matched string: P.M34T

Matching group	Value
origAaShort	M
pos	34
mutAaShort	T

B

p.M34T: Check for M at residue position 34 in all mentioned genes in paper (GJB2, GJB6 and RPL14) to arrive at candidate gene-variant mappings:



C

Annotate candidate gene-variant mapping (in this example: GJB2 & p.M34T) with 125 textual features

Euclidean distance, angle, x-distance, y-distance, word distance ... between gene and variant mentions (*GJB6 = connexin 30 here*):

Her blood was screened for pathogenic splice site mutations in exon 1 and for pathogenic mutations in the protein coding exon 2 region of the **GJB2** gene by sequence analysis. She was also tested for the **connexin 30** deletion (**GJB6-D15S1850**) mutation using a PCR assay.

The coding exons of genetic analysis demonstrated that she was a compound heterozygote for the variants c138T>A (p.Asp46Glu) and c.101T>c (**P.M34T**). **Connexin 30** mutation deletion (**GJB6-D15S1850**) was absent.

Words and counts of alphanumeric characters surrounding gene and variant mentions:

Her blood was screened for pathogenic splice site mutations in exon 1 and for pathogenic mutations in the protein coding exon 2 region of the **GJB2** gene by sequence analysis. She was also tested for the **connexin 30** deletion (**GJB6-D15S1850**) mutation using a PCR assay.

The coding exons of genetic analysis demonstrated that she was a compound heterozygote for the variants c138T>A (p.Asp46Glu) and c.101T>c (**P.M34T**). **Connexin 30** mutation deletion (**GJB6-D15S1850**) was absent.

D

Classify candidate gene-variant mapping using GradientBoostingClassifier on 125 features

500

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

502 **literature. (A)** AVADA uses regular expressions to detect variants in articles. Regular

503 expressions are designed in forms of regular expression generators such as

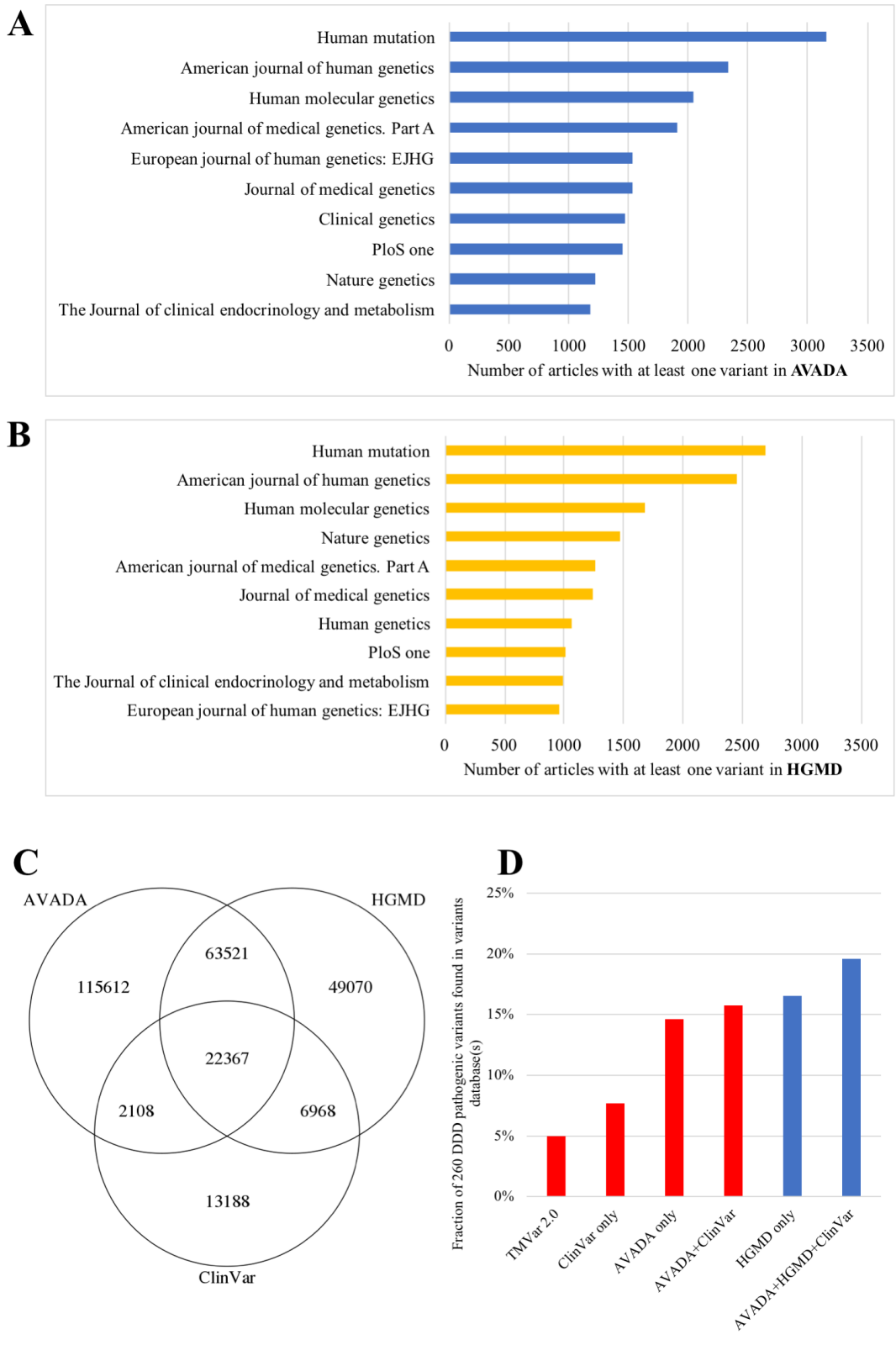
504 “{sep}c.{pos}{space}?{plusMinus}{space}?{offset}[.,]*?{origDna}{space}?{arrow}{space}?{mutDna}”.

505 These regular expression generators contain named matching group generators, such as

506 “{origDna}” (reference nucleotide, such as “A” or “T”) or “{pos}” (numeric position of the

507 mutated nucleotide relative to the start of the transcript). Named matching group generators
508 describe parts of the HGVS description that contain information about the mutation. Regular
509 expression generators are expanded into regular expressions by replacing the matching group
510 generators, such as “{pos}”, into a named matching group, such as “(?P<pos>[1-9][0-9]*)”.
511 Expanding all named matching group generators into named matching groups gives a full regular
512 expression. If a full regular expression matches any string in a given article, the matched string is
513 assumed to be a variant description. **(B)** Given a detected variant description and a set of genes
514 detected in the text of an article, AVADA first checks if the variant matches any of the gene’s
515 transcripts. In the current example, the variant p.M34T matches transcripts of the genes *GJB2*
516 and *GJB6* because both have a methionine residue at position 34, but not the gene *RPL14* (with
517 an asparagine at position 34). The variant p.M34T therefore forms gene-variant candidate
518 mappings (p.M34T, *GJB2*) and (p.M34T, *GJB6*), which are filtered in the next step. **(C)** Given a
519 gene-variant candidate mapping (variant=p.M34T and gene=*GJB2* in this example, highlighted
520 in green), AVADA lets a Gradient Boosting classifier decide if the variant refers to the candidate
521 gene using a set of 125 numerical features that contain information about the textual relationship
522 between the variant mention and the closest mentions of the candidate gene (*GJB2*), as well as
523 the closest mentions [GILL: what does ‘closest mentions’ mean?] of alternative nearby
524 mentioned genes (connexin 30 (encoded by *GJB6*) in the example, in red). The 125 features are
525 based on the relative positions of the closest candidate gene mentions to the variant mention,
526 closest alternative gene mentions to the variant mention, information about the genes’
527 importance in the article, and words and characters surrounding the gene and variant mentions.
528 **(D)** The Gradient Boosting classifier takes these 125 features as input and returns a probability
529 between 0 and 100% indicating the classifier’s assessment of whether the variant actually refers
530 to the given candidate gene. If the classifier returns a likelihood greater than 90%, the gene-
531 variant candidate mapping is transformed to Variant Call Format (chromosome, position,
532 reference and alternative alleles) and entered into the AVADA database. In the present example,
533 AVADA correctly decides that p.M34T only maps to *GJB2* and not connexin 30 (encoded by the
534 gene *GJB6*). Example taken from PubMed ID 23808595.

535



538 **Figure 3. Automatic variant curation results.** (A) Journals with most articles with AVADA
539 curated variants. AVADA extracted variants from 3,159 articles in “Human Mutation”, 2,330
540 articles in “American Journal of Human Genetics”, 2,042 articles in “Human Molecular
541 Genetics” etc. (B) Journals with most articles containing variants curated by HGMD. Similarly
542 to AVADA, the top three journals are “Human Mutation”, the “American Journal of Human
543 Genetics”, and “Human Molecular Genetics”. The two lists share 9 of the top 10 journals even
544 though HGMD is manually curated whereas AVADA is entirely based on automated curation.
545 (C) Extracted variants in AVADA intersected with all disease-causing variants in HGMD and
546 ClinVar. AVADA extracts 85,888 variants in literature-based HGMD (subset to disease-causing
547 variants) and 24,475 variants in submission-based ClinVar (subset to pathogenic and likely
548 pathogenic variants). (D) Comparison of the fraction of Deciphering Developmental Disorders
549 (DDD) causative variants found in various combinations of databases. 260 different variants
550 were reported to be causative of 245 patients’ diseases in the DDD project, a large-scale
551 diagnostic sequencing research project. We subset AVADA, HGMD, ClinVar and the
552 automatically curated variant database tmVar 2.0 to sources pre-dating the publication of the
553 DDD patient set. Of the causative variants, tmVar 2.0 which automatically parses on PubMed
554 abstracts, contained 5%, ClinVar contained 8% reported as (likely) pathogenic, full text-based
555 AVADA contained 15% and HGMD contained 17% reported as disease-causing. All tmVar 2.0
556 variants were either in AVADA or ClinVar. Combining the free (bars in red) AVADA and
557 ClinVar databases recovers 16% of causative variants. Combining all databases facilitates rapid
558 diagnosis for 20% of causative variants.

559

560 **Tables**

561 **Table 1**

HGVS(-like) variant descriptions (alternatives describing same genetic event)	Explanation of HGVS variant description	Disease caused by variant (cited literature 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 ^{44,45}
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 ^{46,47}
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 ^{48,49}

562

563 **Table 1. Examples of HGVS or common HGVS-like variant descriptions.** Each row contains
564 examples of a disease-causing variant description in HGVS or a common HGVS-like notation.
565 Each of these variant descriptions describes a single genetic event causing a disease, usually by
566 giving at least the position of the change in the gene's transcript, an optional reference sequence
567 and a novel alternative (mutated) sequence. All given variants can be described using multiple
568 commonly used notations. Examples of alternatives to the notations are shown in the left hand
569 column that denote the exact same genetic variants. Transcript identifiers for variant
570 descriptions, which enable the mapping of a variant to a reference genome, are usually omitted
571 by article authors. Therefore, an automated method like AVADA must identify the gene's
572 transcript that the variant occurs in [GILL: unclear! The variant may occur within multiple
573 transcripts associated with the gene in question. Do you mean the predominant transcript in a
574 given tissue? Do you mean the longest known transcript associated with that gene?]. The right
575 hand column lists the disease along with two articles using the variant descriptions given in the
576 left hand column. The difficulty of parsing different variant notations that refer to the same
577 genetic event warrants the development of automated approaches for variant curation from the
578 literature.

579

580 **Table 2**

Variant type	AVADA	HGMD	ClinVar
stoploss	0.30%	0.14%	0.10%
nonframeshift	2%	3%	3%
splicing	8%	7%	4%
stopgain	12%	14%	9%
frameshift	14%	22%	11%
missense	65%	53%	74%

581 **Table 2. Variant type percentages in AVADA, HGMD and ClinVar.** Despite being based
582 purely on automatic Natural Language Processing methods, AVADA variant type fractions are
583 always within the range between manually curated HGMD and ClinVar $\pm 1\%$. [GILL: Hmmm!
584 Not sure I like this spin! In three cases, AVADA variant type fractions lie outwith the range
585 between HGMD and ClinVar]

586

587 **Supplemental Methods**

588 **Variant Extraction Directly from Primary Literature**

589 **Download of literature**

590 Articles were identified as potentially relevant based upon title and abstract in PubMed as
591 previously described¹⁰. Briefly, all 25,793,020 available titles and abstracts from PubMed were
592 downloaded. Subsequently, we trained a scikit-learn²⁸ LogisticRegression²⁹ classifier featurized
593 by TF-IDF-transformed words (a common transformation of word frequencies into a feature
594 vector). The training set for the title/abstract document classifier was based on 51,637 positive
595 titles and abstracts cited in OMIM “Allelic Variants” sections or HGMD PRO version 2016.02,
596 and 66,424 random negative titles and abstracts from PubMed. PDFs of articles were
597 downloaded directly from publishers using PubMunch⁵⁰.

598 **Identification of relevant articles based on the full text of articles**

599 We created a full-text classifier that assigns a score between 0 and 1 to each downloaded article,
600 providing an estimate of the article’s likelihood of containing human pathogenic mutation data.
601 To create a TF-IDF feature vector, for use by a machine learning classifier, out of an article’s full
602 text, each article was transformed by means of a scikit-learn²⁸ CountVectorizer with parameters
603 max_df=0.95 and min_df=100 followed by a TfidfTransformer with default parameters. The
604 training set was based on 267,267 random articles in PubMed that were downloaded as a
605 negative training set, and 46,291 full text articles cited in OMIM “Allelic Variants” sections or
606 HGMD PRO version 2016.02. Based on this training set, a scikit-learn²⁸ LogisticRegression²⁹
607 classifier was trained.

608 **Identifying candidate gene mentions in full text**

609 Identification of candidate genes in full text was performed as previously described¹⁰. Briefly, a
610 list of 188,975 gene and protein names was compiled from HGNC³⁵ and UniProt⁵¹. Gene and
611 protein names in this list were matched to word groups in the PDF text. Extractions were
612 supplemented by PubTator⁵² gene extractions where available by matching gene names
613 deposited in PubTator for a particular article to words occurring in that article.

614 **Identifying candidate variant descriptions in full text**

615 Candidate variant descriptions in Human Genome Variation Society (HGVS) or HGVS-like

616 notation⁵³ were identified using 47 regular expressions (Supplemental Table S1 and
617 Supplemental Table S7). We partition mentioned variants into 3 broad categories: cDNA
618 variants (“c.” variants, such as “c.123T>C”), protein variants (“p.” variants such as “p.T34Y”)
619 and splicing variants (“c.” variants with a position and an offset, such as “c.123-2A>G” or “IVS”
620 variants, such as “IVS4-2A>G”). Variant descriptions generally consist of a subset of the
621 following components: variant type (cDNA, protein, splicing), position of the mutation relative
622 to the given transcript, reference nucleotide or amino acid, mutated nucleotide or amino acid, and
623 type of genetic event (deletion, insertion, ...). Using regular expression matching groups,
624 information about all of these components is saved for each identified variant.

625 To create Figure 1, when counting the number of variant descriptions in articles, we removed all
626 non-alphanumeric characters from variant descriptions because inconsistencies throughout the
627 article with respect to spacing and parentheses used can otherwise lead to double-counting
628 variant descriptions.

629 **Mapping variants to candidate genes**

630 A gene-variant candidate mapping of a variant onto a gene is a tuple (g, v) comprising a variant
631 description v and a gene g such that there is at least one transcript t of g that has the variant’s
632 given reference nucleotide/amino acid at the position given in the variant description v . If this is
633 the case, the variant v is supported by the gene g , and (g, v) forms a candidate mapping.

634 To identify all gene-variant candidate mappings in an article with a set of mentioned variant
635 descriptions V and a set of mentioned genes G , AVADA examines each pairwise combination $(g,$
636 $v)$ of a variant v in V and a gene g in G to determine if they form a candidate mapping. Each gene
637 is represented by its set of transcripts deposited in the RefSeq³³ database. All known RefSeq
638 transcripts of g are successively examined to establish if g supports v . Most variants are written
639 in a form that includes the position of the mutation inside the gene’s transcript, the reference
640 sequence, and the mutated sequence (e.g., “c.123A>G”: the position is “123”, the reference
641 sequence is “A” and the mutated sequence is “G”). However, some variants only contain a
642 position and a mutated sequence, not the original reference sequence (e.g., “c.153_154insGG”:
643 the reference sequence is not included, just the novel insertion of “GG” between positions 153
644 and 154 inside the transcript). If the variant description v does not contain a reference sequence,
645 all candidate genes form candidate gene-variant mappings with the variant. These gene-variant

646 candidate mappings are further filtered using a machine learning classifier in the next section.

647 All gene-variant candidate mappings are converted to genomic coordinates (chromosome,
648 position, reference allele and alternative allele). A conversion attempt is unsuccessful if the
649 underlying nucleotide change cannot be identified given the variant description: e.g., this is the
650 case for frameshift variants in “p.” notation such as “p.Val330fsX30”. Here, the precise
651 underlying nucleotide change cannot be inferred from the variant description because the given
652 frameshift may be caused by a very large number of possible nucleotide indel mutations.

653 In the case of a missense protein variant (e.g., NM_000025.2:p.Trp64Arg), the variant was
654 translated to all possible single nucleotide variants that could cause such an amino acid change at
655 the given position in the transcript. Since the Trp at position 64 in NM_000025.2 is encoded by
656 the nucleotides TGG, both changing the T to a C (CGG) and the T to an A (AGG) result in an
657 Arg codon. All further analysis was performed only on variants where conversion to genomic
658 coordinates was successful.

659 **Distinguishing true from false candidate gene-variant mappings**

660 Given a set of candidate gene-variant mappings $\{(g_1, v), (g_2, v), (g_3, v), (g_4, v), \dots\}$, most of the
661 genes g_i associated with v through a candidate mapping are false: the variant v does not map to
662 gene g_i . We constructed a machine learning classifier that distinguishes true gene-variant
663 candidate mappings from false gene-variant candidate mappings. This classifier uses a number of
664 real-valued [GILL: real-value?] numbers, called features, to determine if a gene-variant
665 candidate mapping is true or false. In order to describe these features, some terminology must
666 first be introduced:

- 667 • A “stopword” is a short word such as “by”, “of”, “there”, “if”, “or”, etc. The variant
668 classifier uses a list of 122 stopwords (Supplemental Table S8).
- 669 • An alphanumeric character is a character in the ranges a-z, A-Z, and 0-9.
- 670 • A 2D position of a description in a PDF file consists of a page number and x and y
671 coordinates of the mention on the page.
- 672 • A word position of a description in a PDF file consists of a single integer that gives the
673 index of a word in the PDF document that contains the description.

- 674 • The Euclidean distance of two mentions associated with x and y coordinates (x_1, y_1) and
675 (x_2, y_2) is defined as $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
- 676 • The word distance between two mentions m_1 and m_2 of some genes or variants in an
677 article A is defined as $|w_2 - w_1|$.
- 678 • A mention m_1 occurs “above” a mention m_2 in the document if the page number of the 2D
679 position of mention m_1 is smaller than the page number of the 2D position of m_2 . If the
680 page numbers of the two mentions are the same, m_1 occurs before m_2 if the y coordinate
681 of m_1 in the PDF is smaller than the y coordinate of m_2 in the PDF.

682 Contextual information about a gene or variant mention in a PDF file is defined to consist of the
683 following:

- 684 • the number of stopwords among the 20 words preceding the mention in the article’s text
- 685 • the number of stopwords among the 20 words following the mention in the article’s text
- 686 • the number of alphanumeric characters among the 20 characters preceding the mention in
687 the article’s text
- 688 • the number of alphanumeric characters among the 20 characters following the mention in
689 the article’s text.

690 Each gene g is mentioned 1 to n times in an article. Let $mention(g)_1 \dots mention(g)_n$ be the
691 mentions of the gene g in the article. Similarly, each variant v is mentioned 1 to m times in an
692 article. Let $mention(v)_1 \dots mention(v)_m$ be the mentions of the variant v in the article.

693 The machine learning classifier used by AVADA to distinguish true from false gene-variant
694 candidate mappings is a scikit-learn²⁸ GradientBoostingClassifier³⁴. To decide whether a given
695 gene-variant candidate mapping is true or false, the GradientBoostingClassifier takes a list of 125
696 numerical features containing information about the relationship between mentions of the gene
697 and mentions of the variant in the original article. Based on these features, the classifier returns a
698 number between 0 and 1 that gives the likelihood of the gene-variant mapping being true or not.
699 The 125 features are constructed in 8 different feature groups describing the textual and
700 geometric relationship between the candidate gene and candidate variant mention, and other
701 genes mentioned close to the candidate variant mention. Further information is available in the

702 accompanying code (see “variant_classifier_features.py”, functions “relationship_2d” and
703 “relationship_wordspace”).

704 The variant classifier decides if $mention(v)_j$ maps to gene g for $1 \leq j \leq m$ based on these 125
705 features. The value of these features is determined separately for each variant mention
706 $mention(v)_j$. If the classifier decides that any variant mention in $mention(v)_1 \dots mention(v)_m$ maps
707 to g with classifier score greater or equal to 0.9, the variant v is considered to map to the gene g .

708 To train the classifier, it was presented with a large number of annotated true and false gene-
709 variant candidate mappings, called a training set. The training set for the classifier was created as
710 follows: gene-variant candidate mappings (g, v) discovered by AVADA in a given article A were
711 converted to genomic coordinates in form of chromosome, position, reference and alternative
712 allele. If the genomic coordinates of a gene-variant candidate mapping extracted from A were
713 deposited in ClinVar version 20170228 and annotated as curated from A , the mapping (g, v) was
714 supervised true and all mappings of other genes to the same variant v in the article were
715 supervised false. Otherwise, the variant was discarded. Synonymous variants (e.g.,
716 “p.Trp88Trp”) were also discarded due to the fact that they were largely not disease-causing, or
717 were false extractions. This strategy yielded a training set comprising 25,218 positive training
718 examples and 91,742 negative training examples from 7,823 articles. The importance assigned to
719 each of the 125 features by the GradientBoostingClassifier is listed in Supplemental Table S9.

720 All extracted variants in AVADA were pre-processed using bcftools⁵⁴ to normalize all variants
721 (left-align indels and exclude variants where the RefSeq reference nucleotide did not match the
722 hg19 nucleotide):

```
723 bcftools norm --check-ref x -f human_g1k_v37.fasta -o avada.vcf  
724 avada_non_normalized.vcf
```

725 **Comparison of AVADA to HGMD, ClinVar, and tmVar 2.0**

726 The first version of AVADA was created on articles downloaded until June 2016. To ensure a
727 fair comparison, we compare AVADA with HGMD PRO version 2016.02 and ClinVar version
728 20160705 [GILL: date accessed?]. These were obtained from
729 <http://www.hgmd.cf.ac.uk/ac/index.php> and ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/
730 , respectively. tmVar 2.0 variants were obtained from

731 <ftp://ftp.ncbi.nlm.nih.gov/pub/lu/PubTator/mutation2pubtator.gz> . The tmVar file was subset to
732 contain only tmVar-extracted variants in articles from 2016 and before (same set of articles used
733 as input to AVADA). tmVar-extracted rsIDs were converted to genome coordinates by joining
734 with the official dbSNP database mapping rsIDs to genome coordinates at
735 ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_GRCh37p13/VCF/All_20180423.vcf.gz .
736

737 Variants reported in AVADA, HGMD, ClinVar, and tmVar 2.0 were normalized (as above)
738 using bcftools:

```
739 bcftools norm --check-ref x -f human_g1k_v37.fasta -o  
740 <database_normalized>.vcf <database>.vcf
```

741 Variants were counted to be in two variant databases if the full variant description (chromosome,
742 position, reference and alternative alleles) in both databases matched exactly. HGMD contained
743 165,051 distinct variants, of which 141,926 were marked as disease-causing [GILL: DMs only?
744 DMs plus DM?s ?]. ClinVar contained 142,396 distinct variants, of which 44,631 were marked
745 as “pathogenic” or “likely pathogenic”. tmVar 2.0 contained 80,159 distinct variants.

746 **Variant types contained in AVADA**

747 To count the fractions of variant types contained in AVADA, each variant was assigned one of
748 the types “missense” (single nucleotide variants changing an amino acid in the mapped gene),
749 “nonframeshift” (insertion, deletion and indel variants adding a multiple of 3 nucleotides to a
750 coding exon), “frameshift” (all other insertion, deletion and indel variants in coding exons),
751 “splicing” (splice-site variants), “stopgain” (single nucleotide variants changing an amino acid
752 codon in a coding exon to a stop codon) and “stoploss” (single nucleotide variants changing a
753 stop codon to an amino acid codon) by automatically analyzing the effect of the variant on the
754 mapped transcript. Variants of all types were summed, and fractions of variant types were
755 calculated as the number of variants of a particular type over the total number of variants of all
756 types in AVADA.

757 **Variant types contained in ClinVar and HGMD**

758 To generate fractions of variant types in HGMD and ClinVar, variants in these databases were
759 annotated with semantic effect using ANNOVAR⁵⁵. All HGMD or ClinVar variants that had a
760 missense, stoploss, stopgain, splice-site, frameshift or nonframeshift effect in ENSEMBL³⁶ and

761 RefSeq³³ coding exons, and had a variant frequency of less than 3% in ExAC⁵⁶ v0.3 and the
762 1000 Genomes Project⁵⁷ phase 3 were counted, and percentages of each variant type were
763 calculated as the number of variants of a particular type over the total number of missense,
764 stoploss, stopgain, splice-site, frameshift and nonframeshift variants in HGMD and ClinVar,
765 respectively.

766 **Diagnosis of patients with Mendelian diseases using AVADA**

767 DDD patient Variant Call Format (VCF) files were obtained from the European Genome-
768 Phenome Archive³⁹ (EGA) study number EGAS00001000775. We identified VCF files for
769 affected patients by matching the phenotypes that each VCF file was annotated with the
770 phenotypes that each patient identifier and causative variant were annotated with, and verifying
771 that the causative variant was contained in the patient's associated VCF file. If unique
772 identification of a patient's VCF file was not possible, we omitted the patient. Reported disease-
773 causing variants that were not found in a VCF file were omitted. Bcftools were used to normalize
774 all variants in DDD VCF files using the following command:

```
775 bcftools norm -f human_g1k_v37.fasta -o <normed DDD VCF file>  
776 <original DDD VCF file>
```

777 **Sensitivity of variant annotation using AVADA, HGMD, ClinVar, and tmVar 2.0**

778 ANNOVAR⁵⁵ was used to annotate variants with a predicted effect on protein-coding genes from
779 ENSEMBL³⁶ and RefSeq³³, and allele frequencies from the ExAC⁵⁶ v0.3, the 1000 Genomes
780 Project⁵⁷ phase 3 and the UK10K⁵⁸ ALSPAC and TWINS sub-cohorts. All variants with a
781 frequency of at most 0.5% in all sub-populations of ExAC v0.3, 1000 Genomes Project and the
782 UK10K ALSPAC and TWINS sub-cohorts, that affected a protein-coding gene and were
783 missense, stopgain, stoploss, frameshift indel, nonframeshift indel or splice-site disrupting were
784 retained.

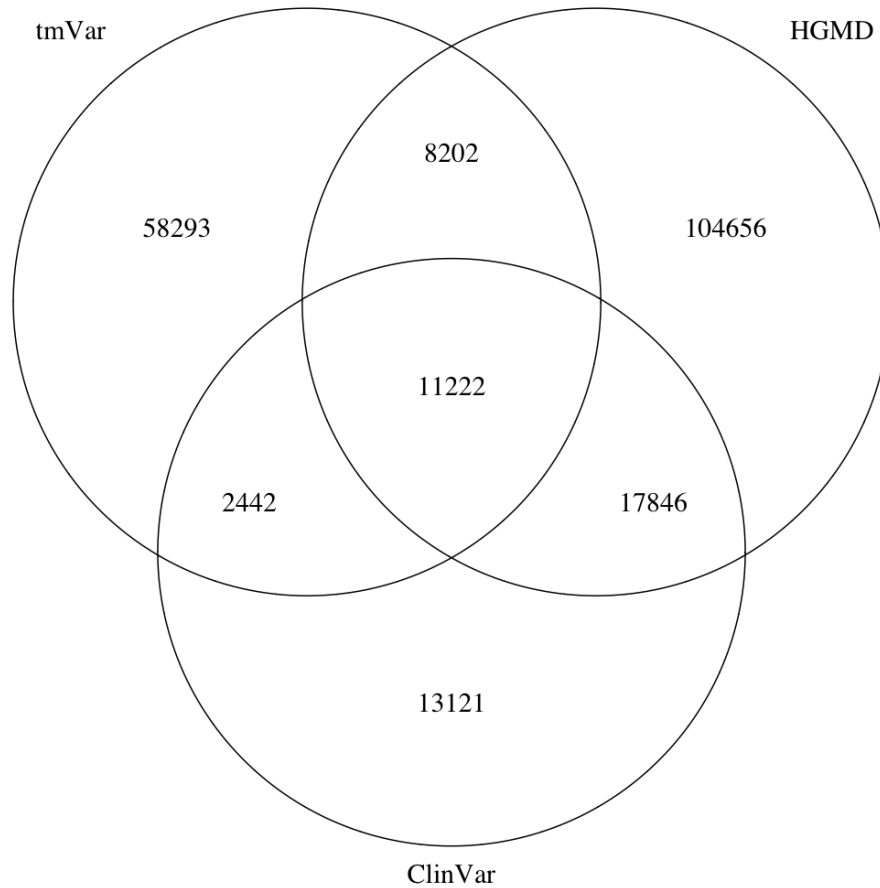
785 AVADA and tmVar 2.0 were subset to variants from articles until 2014 by associating each
786 article with the publication date stored in PubMed and subsetting to articles until 2014. HGMD
787 variants were subset to 2014 by removing all variants with a "new_date" greater than 2014.
788 ClinVar version 20141202 was obtained from
789 ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_1.0/2014/.

790 **Specificity of variant annotation using AVADA, HGMD, ClinVar, and tmVar 2.0**
791 Candidate causative variants in patient’s exomes were defined using the same variant filtering
792 criteria as above (0.5% minor allele frequency thresholds for all mutations affecting protein-
793 coding regions). VCF files of the 245 patients annotated with AVADA, HGMD, ClinVar, or
794 tmVar 2.0 were processed as described above to arrive at a list of rare candidate causative
795 variants per patient. A candidate causative variant was counted as “annotated” by a database if
796 the identical variant (chromosome, position, reference and alternative allele) occurred in the
797 database.
798

799 **Supplemental Figures**

800 **Supplemental Figure 1**

801



802

803 **Supplemental Figure 1. Extracted variants in tmVar intersected with all disease-causing**
804 **variants in HGMD and ClinVar.** tmVar extracts 19,424 variants in HGMD (subset to disease-
805 causing variants), as compared to 85,888 variants for AVADA and 13,664 variants in ClinVar
806 (subset to pathogenic and likely pathogenic variants), as compared to 24,475 for AVADA.