AMELIE speeds Mendelian diagnosis by matching patient phenotype and genotype to primary literature

Authors:

Johannes Birgmeier[1], Maximilian Haeussler[2], Cole A. Deisseroth[1], Ethan H. Steinberg[1], Karthik A. Jagadeesh[1], Alexander J. Ratner[1], Harendra Guturu[3], Aaron M. Wenger[3], Mark E. Diekhans[2], Peter D. Stenson[4], David N. Cooper[4], Christopher Ré[1], The Manton Center[5], Alan H. Beggs[5], Jonathan A. Bernstein[3], and Gill Bejerano[1,3,6,7*]

Affiliations:

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

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

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

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

[5]The Manton Center for Orphan Disease Research, Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA 02115, USA

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

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

[*]Corresponding author. Email: bejerano@stanford.edu

Overline: GENETIC DIAGNOSIS

One Sentence Summary:

AMELIE parses primary literature about Mendelian diseases to rank patient candidate causative genes, thereby accelerating diagnosis.

Abstract

The diagnosis of Mendelian disorders requires labor-intensive literature research. Trained clinicians can spend hours looking for the right publication(s) supporting a single gene that best explains a patient's disease. AMELIE (Automatic Mendelian Literature Evaluation) greatly accelerates this process. AMELIE parses all 29 million PubMed abstracts, and downloads and further parses hundreds of thousands of full text articles in search of information supporting the causality and associated phenotypes of most published genetic variants. AMELIE then prioritizes patient candidate variants for their likelihood of explaining any patient's given set of phenotypes. Diagnosis of singleton patients (without relatives' exomes) is the most time-consuming scenario, and AMELIE ranked the causative gene at the very top for 66% of 215 diagnosed singleton Mendelian patients from the Deciphering Developmental Disorders project. Evaluating only the top 11 AMELIE scored genes of 127 (median) candidate genes per patient resulted in a rapid diagnosis in 90+% of cases. AMELIE-based evaluation of all cases was 3-19x more efficient than hand-curated database-based approaches. We replicated these results on a retrospective cohort of clinical cases from Stanford Children's Health and the Manton Center for Orphan Disease Research. An analysis web portal with our most recent update, programmatic interface and code is available at AMELIE.stanford.edu.

Introduction

Millions of babies born worldwide each year are affected by severe genetic, often Mendelian disorders *(1)*. Patients with Mendelian diseases have one or two genetic variants in a single gene primarily responsible for their disease phenotypes *(2)*. Roughly 5,000 Mendelian diseases, each with a characteristic set of phenotypes, have been mapped to about 3,500 genes to date *(3)*. Exome sequencing is often performed to identify candidate causative genes, resulting in a relatively high (currently 30%) diagnostic yield *(4)*. A genetic diagnosis provides a sense of closure to the patient family, aids in patient trajectory prediction and management, allows for better family counseling, and in the age of gene editing even provides first hope for a cure. However, identifying the causative mutation(s) in a patient's exome to arrive at a diagnosis can be very time-consuming, with a typical exome requiring hours of expert analysis *(5)*.

Definitive diagnosis of a known Mendelian disorder is accomplished by matching the patient's genotype and phenotype to previously described cases from the literature. Manually curated databases *(6–10)* are utilized to more efficiently access extracts of the unstructured knowledge in the primary literature. Automatic gene ranking tools *(11–18)* use these databases to prioritize candidate genes in patients' genomes for their ability to explain patient phenotypes. An important feature of many gene ranking tools is the use of phenotype match functions on patient phenotypes and gene- or disease-associated phenotypes. Phenotype match functions exploit the structure of a phenotype ontology *(9)* and known gene-disease-phenotype associations to quantify the inexact match between two sets of phenotypes *(11, 12)*, with recent approaches developed to computationally extract phenotype data from electronic medical notes *(19, 20)*. The goal of all gene ranking tools is to aid a busy clinician in arriving at a definitive diagnosis of any case presented to them in the shortest amount of time, by reading up on genes in the order the algorithm has ranked them.

Given the rapidly growing number of rare diseases with a known molecular basis *(21)* and the difficulty of manually finding a diagnosis for some rare diseases with variable phenotypes, many patients experience long diagnostic odysseys *(22)*. Expert clinician time is expensive and scarce, but machine time is cheap and plentiful. We aimed to accelerate the diagnosis of patients with Mendelian diseases by using information from primary literature to construct gene rankings, thus allowing clinicians to discover the causative gene along with supporting literature in a minimum amount of time.

Here, we introduce AMELIE (Automatic Mendelian Literature Evaluation). AMELIE uses natural language processing (NLP) to automatically construct a homogeneous knowledgebase about Mendelian diseases directly from primary literature. To perform this operation, AMELIE was trained on data from manually curated databases such as Online Mendelian Inheritance in Man (OMIM) *(6)*, Human Gene Mutation Database (HGMD) *(8)* and ClinVar *(10)*. AMELIE then used a machine learning classifier that integrated knowledge about a patient's phenotype and genotype with its knowledgebase to rank candidate genes in the patient's genome for their likelihood of being causative, and simultaneously supported its ranking results with annotated citations to the primary literature. We compare this end-to-end machine learning approach to gene ranking methods that rely on manually curated databases using a total of 271 singleton patients from 3 different sources, including 2 clinical centers and a research cohort.

Results

Overview of AMELIE

Given a patient's genome sequencing data and a phenotypic description of the patient, AMELIE aims to both identify the gene causing the patient's disease (when possible) and supply the clinician with literature supporting the gene's causal role. To this end, AMELIE creates a ranking of candidate causative genes in the patient's genome with the aim of ranking the true causative gene at the top. AMELIE constructs its candidate causative gene ranking by comparing information from the primary literature to information about the patient's genotype and phenotype.

To process information from the full text of primary literature, AMELIE constructs a knowledgebase directly from the primary literature up-front using natural language processing techniques trained on manually curated databases. After knowledgebase construction, AMELIE ranks any patient's candidate causative genes using a classifier, which compares knowledge from the AMELIE knowledgebase with phenotypic and genotypic information about the patient. AMELIE explains each gene's ranking to the clinician by citing articles about this gene in the knowledgebase.

Identification and download of relevant Mendelian disease articles based on all of PubMed

The first step towards building the AMELIE knowledgebase was discovering relevant primary literature. Of 29 million peer-reviewed articles deposited in PubMed, only a fraction is relevant for Mendelian disease diagnosis. We constructed a machine learning classifier that, given titles and abstracts of articles from PubMed, identified potentially relevant articles for the AMELIE knowledgebase.

Machine learning classifiers take as input a numerical vector describing the input, called the “feature vector”. Here, we used a so-called TF-IDF transformation to convert input text into a feature vector. We implemented the title/abstract document classifier as a logistic regression classifier. Logistic regression transforms its output using the logistic sigmoid function to return a probability value which is then mapped into binary (positive/negative) decision making *(23)*.

Machine learning classifiers learn to classify an input as positive (relevant) or negative (irrelevant) by being exposed to a large number of labeled positive and negative examples (the training set). OMIM *(6)* is an online database of Mendelian diseases, genes, and associated phenotypes. HGMD *(8)* is a database of disease-causing mutations in the human genome. The training set for the title/abstract relevance classifier consisted of titles and abstracts of 56,479 Mendelian-disease-related articles cited in OMIM and HGMD as positive training examples, and 67,774 random titles and abstracts of PubMed articles (largely unrelated to Mendelian disease) as negative training examples.

Precision and recall are two standard measures of evaluating classifier performance. Precision measures the fraction of all inputs classified positive that are truly relevant. Recall measures the fraction of truly positive inputs that are classified positive. Five-fold cross-validation (splitting all available labeled training data to include 80% in a training set and evaluating on the remaining 20%, 5 times in round-robin fashion) returned an average precision of 98% and an average recall of 96%.

All 28,925,544 titles and abstracts available in PubMed on September 30, 2018, were downloaded and processed by the document classifier. The classifier identified 578,944 articles as possibly relevant based on their PubMed title and abstract, of which we downloaded 515,659 (89%) full text articles directly from dozens of different publishers.

Building a structured database of information about Mendelian diseases from full text

From the full text of an article, multiple types of information were extracted. Gene mentions in full text were identified using lists of gene and protein names and synonyms from the HUGO Gene Nomenclature Committee (HGNC) *(24)*, UniProt *(25)* and the automatically curated PubTator *(26)*, an NCBI service combining gene mentions discovered by multiple previously published automatic gene recognition methods. AMELIE recognized approximately 93% of disease-causing gene names. However, through a combination of unfortunate gene synonyms (such as "FOR", "TYPE", "ANOVA", or "CO2"), as well as genes mentioned only in titles of cited references, or interaction partners of causative genes, a median of 12 distinct gene candidates were discovered in each article (table S1).

To discover which gene(s) were the subject of the PubMed article, each distinct gene candidate extracted from an article received a "relevant gene score" between 0 and 1 indicating the likelihood of the gene being important in the context of the article. Training data for the relevant gene classifier was obtained from OMIM and HGMD. A total of 304,471 downloaded full-text articles contained at least one gene with a relevance score of 0.1 or higher. These articles, along with their above-threshold scoring genes, formed the AMELIE knowledgebase. Articles in the AMELIE knowledgebase contained a median of 1 gene with a relevant gene score between 0.1 and 1 (table S1). Further, genetic variants (for example, "p.Met88Ile" or "c.251A>G") were identified in the full text of each article and converted to genomic coordinates (chromosome, position, reference and alternative allele) using the AVADA variant extraction method *(27)*. A median of 3 distinct genetic variants were extracted from 123,073 full-text articles in the AMELIE knowledgebase.

Phenotype mentions were recognized in full text articles using a list of phenotype names compiled from Human Phenotype Ontology (HPO) *(9)*. By linking all genes with a relevant gene score of at least 0.5 in an article with all phenotypes mentioned in the same article, we arrived at a total of 872,080 gene-phenotype relationships covering 11,537 genes (fig. S1).

Five scores between 0 and 1 were assigned to the full text of each article. A "full-text document relevance" score assessed the likely relevance of the article for the diagnosis of Mendelian diseases. A "protein-truncating" and a "non-truncating" score each gave an assessment of whether the article was about a disease caused by protein-truncating (splice-site, frameshift, stopgain) or non-truncating (other) variants. A “dominant” and a “recessive” score each gave an assessment of the discussed inheritance mode(s) in the article.

Precision and recall of full-text article information (relevant genes, extracted phenotypes, full-text article scores) varied between 74% and 96%. All the data described in this section were entered into the AMELIE knowledgebase, keyed on the article they were extracted from (Fig. 1A). The top journals from which the most gene-phenotype relationships were extracted are shown in Fig. 1B and table S2. We estimated that the number of newly described gene-phenotype relationships has increased by an average of 10.5% every two years since the year 2000 (fig. S2).

The AMELIE classifier assigns patient genes a likelihood of being causative

Given a patient with a suspected Mendelian disease, AMELIE aims to speed up discovery of the causative gene by ranking patient genes for their ability to describe a patient set of phenotypes. AMELIE performs standard filtering of the patient variant list *(21, 28)* to keep only “candidate causative variants” that are rare in the unaffected population and are predicted to change a protein-coding region (missense, frameshift, nonframeshift indel, core splice-site, stoploss, and stopgain variants). Core splice sites were defined to consist of the 2 basepairs at either end of each intron. Genes containing candidate causative variants were called candidate causative genes (or “candidate genes”). AMELIE ranked approximately 97% of known disease-causing mutations, excluding only those in deeper intronic and non-protein-coding intergenic regions.

We defined an article in the AMELIE knowledgebase to be about a candidate causative gene if the candidate causative gene had a relevant gene score of at least 0.1 in the article to maximize recall while maintaining a median of 1 relevant gene per article. We constructed a machine learning classifier called the “AMELIE classifier” that assigns a score between 0 and 100 to triples *(P, G, A)* consisting of a set of patient phenotypes *P*, a candidate causative gene *G*, and an article *A* about the candidate gene. Given a patient with phenotypes *P* and a candidate gene *G*, the AMELIE score indicates whether the article *A* is likely helpful for diagnosing the patient because it links mutations in *G* to the patient’s phenotypes *P*. Higher AMELIE scores indicate articles more likely relevant to diagnosis. The AMELIE classifier was implemented as a logistic regression classifier and returns a score between 0 and 100 called the “AMELIE score”. The AMELIE score is used to both rank patient candidate genes and explain rankings by citing primary literature, as described below.

The AMELIE classifier uses a set of 27 real-valued features, falling into 6 feature groups (Fig. 1C). The 6 feature groups comprise: **(1)** 5 features containing information about disease inheritance mode extracted from the article and patient variant zygosity. **(2)** 5 features containing information about AVADA-extracted variants from the article and overlap of these variants with patient variants. **(3)** 2 features containing information about patient phenotypes based on the Phrank *(11)* phenotypic match score of phenotypes in article A with the patient phenotypes P. **(4)** 5 features containing information about article and patient variant types. **(5)** 3 features containing information about article relevance and relevance of the candidate gene in the article, and **(6)** 7 features containing a priori information about the patient’s candidate causative variants in G such as in-silico pathogenicity scores *(29)* and gene-level mutation intolerance scores *(30, 31)*.

To train the AMELIE classifier, we constructed a set of 681 simulated patients using data from OMIM *(6)*, ClinVar *(10)*, and the 1000 Genomes Project *(32)*. Each simulated patient s was assigned a disease from OMIM, with phenotypes noisily sampled from the phenotypes associated with the disease. The genome of each simulated patient was based on genome sequencing data from the 1000 Genomes Project. An appropriate disease-causing variant from ClinVar was added to each simulated patient’s genome. Each simulated patient was assigned a diagnostic article A_s describing the genetic cause of the patient’s disease. In total, the simulated patients covered a total of 681 OMIM diseases (1 per patient) and a total of 1,090 distinct phenotypic abnormalities (table S3). The sampled phenotypes for each disease covered an average 21% of the phenotypes manually associated with the disease by HPO.

The AMELIE classifier was trained to recognize the diagnostic article A_s out of all articles about genes with candidate causative variants in a patient s. Of a total of 681 training “patients” constructed using data in OMIM and ClinVar, the single positively labeled article was recognized and downloaded during AMELIE knowledgebase construction in 664 cases (98%), creating 664 positive training examples. The negative training set for the AMELIE classifier consisted of triples *(P_s, G, A)* for each simulated patient *s* where *G* was a non-causative candidate gene in patient *s*, and *A* was an article about *G*. For training efficiency, we used only 664,000 random negative training examples out of all available negative training examples.

The AMELIE classifier assigns each candidate gene *G* an AMELIE score, defined as the best AMELIE classifier score for any paper *A* about gene *G*, as it relates to patient *P* (Fig. 1C). Candidate causative genes were ranked in descending order of their associated score.

Evaluating AMELIE on a retrospective patient test set

We evaluated AMELIE on a set of 215 real singleton patients with an established diagnosis from the Deciphering Developmental Disorders (DDD) project *(33)*. The DDD dataset included HPO phenotypes (a median of 7 per patient), exome data in variant call format (VCF), and the causative gene for each patient (1 per patient). AMELIE's goal was to rank the established causative gene at or near the top of its ranked list of candidate genes for each patient. Filtering for candidate causative variants resulted in a median of 163 variants in 127 candidate genes per patient Fig. 1C). We used the set of 215 patients obtained from the DDD study to evaluate AMELIE against Exomiser *(14)*, Phenolyzer *(15)*, Phen-Gen *(16)*, eXtasy *(17)*, and PubCaseFinder *(18)*. The output of all methods, consisting of a list of ranked genes, was subset to the (median) 127 candidate genes that AMELIE used for each patient based on the filtering criteria previously described (Fig. 2A). This ensured the fair evaluation of all gene ranking methods against the same set of genes.

AMELIE analyzed a median of 4,173 articles per patient and ranked the causative gene at the very top in 142 (66%) out of 215 cases, and in the top 10 in 193 cases (89.7%). Other methods ranked the causative gene at the top between 38% of cases (Exomiser) and 8% of cases (Phen-Gen) (Fig. 2B). AMELIE performed significantly better than all compared methods (all p-values $\leq 1.68*10^{-9}$; one-sided Wilcoxon signed rank test; table S4). Of 117 distinct top-ranked articles supporting the DDD patients where AMELIE ranked the test set causative gene at number 1, only 36 (31%) were cited in OMIM as determined by a systematic Google search of omim.org (table S5).

Due to the large number of patients expected to be sequenced for Mendelian diagnosis *(34)*, one may want to set guidelines for rapid vs. in-depth exome or genome analysis. In our test set of

215 patients, AMELIE offered a diagnosis for 90% of diagnosable cases when evaluating only up to the top 11 AMELIE-ranked genes per case, or 9% of a median of 127 candidate causative genes. If using any of the other methods, the clinician would have to investigate between a median of 30 genes (when using Exomiser to rank patient candidate causative genes) and 108 genes per patient to arrive at the diagnosis in 90% of diagnosable cases (Fig. 2C).

If the clinician used AMELIE to determine the order in which they evaluate their entire candidate gene list, one gene after the other, on the DDD set of 215 patients, they would evaluate a total of 735 gene-patient matches to arrive at the causative gene for all 215 patients. If the clinician went through the list of candidate genes in random order, they would evaluate an expected total sum of 14,383 gene-patient matches to arrive at the causative gene for all patients. By this metric, AMELIE improved diagnosis time by a factor of 19.6x over a random baseline. The next best tool, Exomiser, would require the clinician to read about 2,085 genes until arriving at the causative gene for all patients, an improvement of 6.9x faster over a random baseline. The performance of other methods ranged from a speedup of 3.13x to 1.04x (Fig. 2D). The speedup provided by AMELIE was thus more than twice that provided by the next-best tool, Exomiser.

Replication of AMELIE performance on 56 clinical cases from two sites

To test for result replication across data sources, we evaluated AMELIE using 56 singleton clinical cases seen by the Medical Genetics Service at Stanford Children's Health and the Manton Center for Orphan Disease Research at Boston Children's Hospital. Patient genotype and phenotype data were obtained from Stanford and the Manton Center Gene Discovery Core.

We performed a comparison of gene ranking performance using AMELIE against other methods as above for the DDD patients. AMELIE ranked the causative gene at the very top in 33 (59%) out of 56 cases, and in the top 10 in 50 cases (89%). Again, AMELIE significantly outperformed all other methods (all p-values $\leq 6.65*10^{-3}$; one-sided Wilcoxon signed rank test; fig. S3A and table S6). AMELIE offered a diagnosis for 90% of patients in the test set of 56 Stanford and Manton patients if evaluating the top 15 candidate genes per patient (9% of a median of 172.5), replicating its performance on the DDD set (fig. S3B).

To arrive at the causative gene for each patient in the clinical test set from Stanford and Manton when using AMELIE, a clinician would need to evaluate 300 genes, compared to a baseline of 6,106 genes if evaluating genes in random order. Similar to the DDD patient test set, AMELIE resulted in a speedup of 20x compared to the baseline, 2-20x faster than other methods (fig. S3C). Since the other methods do not use simulated patients for training, gene ranking results using other methods were obtained by running each respective method once on the simulated patients set. 5-fold cross-validation on the 681 simulated patients showed that AMELIE generated significantly better causative gene rankings compared to the other methods (all p-values $\leq 5.24*10^{-10}$; fig. S4, and table S7).

We ran multiple tests with modified AMELIE knowledgebases and AMELIE classifiers to dissect the relative contribution of different AMELIE components to its causative gene ranking performance. For all 175 test cohort patients with the causative gene ranked at the top, we investigated which machine learning features of the AMELIE classifier contributed most to the high score of the causative gene. Overwhelmingly, for 149 (85%) of 175 real test patients, the feature contributing most to the high score was a high phenotypic match between the patient and the article. However, 14 out of a total 27 AMELIE classifier features (52%) occurred at least once within the 3 features contributing most to the top rank of a patient's causative gene (Fig. 3A and table S8).

To measure how much AMELIE relied on certain feature groups, we re-trained the AMELIE classifier 6 times, each time dropping one of its 6 feature groups. With dropped-out features, the number of causative genes ranked at the top across the test set of 271 real patients shrank between 4% and 39% (Fig. 3B and table S9). AMELIE did not better rank causative genes when phenotype recognition was augmented by data from UMLS (35), MeSH (36), and SNOMED-CT (37), three databases containing additional phenotype names and synonyms. However, AMELIE ranked 32% more causative genes at the top when using full-text data rather than data gathered only from titles and abstracts.

AMELIE's performance is not correlated with number of articles about a causative gene

We investigated whether the number of articles about the causative gene in the AMELIE knowledgebase correlated with the causative gene rank by performing linear regression between the causative gene rank and number of articles analyzed for the causative gene. The regression revealed no significant relationship (p=0.85 that the slope of regression equals 0 according to a Wald Test with t-distribution of the test statistic; Fig. 3C), suggesting that AMELIE performs well independent of the number of papers it has analyzed about a causative gene. For the 22 patients (8% of a total of 271 real test patients) with less than 10 papers analyzed for the causative gene, AMELIE ranked causative genes at the top for 10 (45%) cases. In contrast, Exomiser ranked the causative gene at the top in 6 (27%) of these cases.

The AMELIE knowledgebase and AMELIE classifier work together to arrive at high causative gene ranks

We investigated the relative contribution of the AMELIE classifier and the AMELIE knowledgebase to AMELIE's overall gene ranking performance. We re-trained the AMELIE classifier using data from DisGeNET *(38)*, a text mining-based database containing gene-phenotype relationships, disease-causing variants, and links to primary literature from PubMed. Using DisGeNET data resulted in significantly worse causative gene rankings compared to the AMELIE knowledgebase ($p \leq 4.76*10^{-23}$; table S10). We then replaced the AMELIE classifier (Fig. 1C) with the Phrank *(11)* phenotypic match score to estimate the impact of the AMELIE classifier on overall AMELIE performance. Gene ranking by the Phrank phenotypic match score resulted in ranking 94 (35%) of 271 real patients' causative genes at the top, significantly worse compared to the AMELIE classifier, which ranked 175 causative genes at the top ($p=1.33*10^{-11}$, one-sided Wilcoxon signed rank test). We conclude that the AMELIE knowledgebase and the AMELIE classifier work together to achieve AMELIE's high causative gene-ranking performance.

Interactive and programmatic access to AMELIE-based literature analysis

AMELIE can be used through its web portal at https://AMELIE.stanford.edu to utilize AMELIE for patient analysis. The portal offers both an interactive interface (fig. S5) and an application programming interface (API) that enables integrating AMELIE into any computer-assisted clinical workflow. The AMELIE knowledgebase will be updated every year. A pilot of AMELIE has been running at this web address since August 2017, as a service to the community, using an AMELIE knowledgebase automatically curated from articles published until June 2016, and has since served many thousands of queries from more than 40 countries.

Discussion

We present AMELIE, a method for ranking candidate causative genes and supporting articles from the primary literature in patients with suspected Mendelian disorders. We show that AMELIE ranks the causative gene first (among a median of 127 genes) in 2 out of 3 of patients, and within the top 11 genes in over 90% of 215 real patient cases. These results were closely replicated on a cohort of 56 clinical patients from Stanford Children's Health and the Manton Center for Orphan Disease Research.

Mendelian disease diagnosis is a complex problem and clinicians or researchers can spend many hours evaluating a single case. With 5,000 diagnosable Mendelian diseases caused by roughly 3,500 different genes that manifest in different subsets of over 13,000 documented phenotypes, manual patient diagnosis from the primary literature is highly labor intensive. Manually curated databases like OMIM, OrphaNet, and HGMD take a step towards alleviating clinician burden by attempting to summarize the current literature. However, manual curation is growing ever more challenging as the literature about Mendelian diseases is increasing at an accelerating rate. Based on AMELIE analysis, the number of gene-phenotype relationships in Mendelian literature has been increasing by an average of 10.5% every two years since the year 2000. Because AMELIE is an automatic curation approach requiring only an initial critical mass of human curated data to train on, it is not constrained by the bottleneck of on-going human curation. For example, of 117 top-ranked articles supporting the DDD patients where AMELIE ranked the test set causative gene at number 1, only 36 (31%) were cited in OMIM. OMIM, a manually curated database, does not, of course, promise to capture all papers pertaining to any given disease gene, but an automated effort like AMELIE can.

Compared to existing gene-ranking approaches, AMELIE replaces the notion of a fixed disease description (that is, a single set of phenotypes) with the notion of an article and the phenotypes described in it. This approach has multiple advantages. First, it is often fastest to convince a clinician about a diagnosis given an article directly describing the disease, which often includes disease information such as patient images and related literature. Additionally, with considerable phenotypic variability in Mendelian diseases *(39)*, matching patients to specific reports in the literature is conceptually more helpful for definitive diagnosis than matching to a disease, which is effectively a compendium of previously described patient phenotypes.

Due to its dependence on literature and exome sequencing data, AMELIE is subject to a number of limitations. Biomedical literature is not guaranteed to contain the full set of phenotypes known to be associated with a disease, and AMELIE makes no claim about capturing this full set. Rather, AMELIE focuses on causal gene ranking using its knowledgebase, and as we show, it already does it to great practical utility. Certain articles about Mendelian diseases may mention a very small number of phenotypes (or none at all) and just mention disease and causative genes. Although this situation does not appear to be very common in practice (as seen by the good performance of AMELIE), the problem could be alleviated by automatically parsing disease names from such articles and associating diseases with manually curated phenotype information from resources such as HPO. Natural language processing approaches could also be used to read additional texts as well, such as electronic medical notes *(19, 20)*. Further, AMELIE requires as input a list of HPO terms to describe patient phenotypes, although these may be provided by tools such as ClinPhen *(19)* that automatically extract HPO phenotypes directly from free-text clinical notes. Last, AMELIE is hampered by access to literature. Although AMELIE successfully obtained 80% of full-text articles that it deemed relevant based on title and abstract, better publisher programmatic access to full-text literature for the purposes of text mining may lead to even better gene ranking results.

Understanding the impact of hundreds of thousands of variants in thousands of different genes against a body of knowledge of millions of peer reviewed papers that is ever expanding is a challenging task. Because a diagnosis shapes the future management of a patient, there must always be a human expert approving every diagnosis. But the sheer number of patients that can benefit from a molecular diagnosis and our intention to sequence millions of them in the next few years absolutely necessitate automating as much as possible of the diagnostic process, to potentiate rapid, affordable, reproducible and accessible clinical genome-wide diagnosis. As such, along with complementary medical record parsing tools *(19, 20)*, AMELIE provides an important step towards integrating personal genomics into standard clinical practice.

Materials and Methods

Study design

We implemented a natural language processing and machine learning system dubbed "AMELIE" (for Automatic Mendelian Literature Evaluation) to automatically identify candidate causative genes in patients with Mendelian (monogenic) diseases based on information in primary literature. The system consists of two components: a knowledgebase constructed directly from primary literature, and a classifier that ranks candidate causative genes for a patient with a Mendelian disease.

To construct the AMELIE knowledgebase, we trained logistic regression classifiers *(23)* largely on OMIM *(6)* and HGMD *(8)* data to identify potentially relevant PubMed abstracts. Similar classifiers were used to determine full text relevance and identify disease-causing genes, phenotypes, disease inheritance modes, disease-causing variants, and disease-causing variant types from abstract and article text.

The AMELIE classifier was implemented as a logistic regression classifier *(23)*. We constructed a set of 681 simulated patients with a single disease-causing variant using data from the 1000 Genomes Project *(32)*, OMIM *(6)*, Human Phenotype Ontology (HPO) *(9)*, and ClinVar. The AMELIE classifier was trained to recognize the simulated patients' disease-causing genes (positive training examples) against a background of non-disease-causing genes (negative training examples).

We evaluated AMELIE against other knowledgebases and gene ranking tools using a set of 215 previously diagnosed patients from the Deciphering Developmental Disorders (DDD) project *(33)*. The DDD 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). Each patient was associated with a candidate gene list generated using variant frequency filtering techniques, notably by restricting variant frequency to <= 0.5% minor allele frequency in a large control cohort *(30)*. Using the DDD patient data, we compared AMELIE against 5 other gene ranking tools (Exomiser *(14)*, Phenolyzer *(15)*, Phen-Gen *(16)*, eXtasy *(17)*, and PubCaseFinder *(18)*. We replicated the results on the DDD cohort by combining 35 patients from Stanford Children's Health and 21 patients from the Manton Center for Orphan Disease

Research into a further set of 56 test patients. Informed consent was obtained from all participants. Further details about the AMELIE algorithm are provided in the Supplementary Materials and Methods.

Statistical analysis

To test performance differences between any two different gene ranking methods, we used the one-sided Wilcoxon signed-rank test throughout the manuscript. P< 0.05 was considered significant. No adjustments to alpha level or multiple testing correction methods were applied. The Wilcoxon signed rank test is a nonparametric test and does not assume any particular distribution of the data. We used this test to compare two matched samples: in our case, two lists of causative gene ranks on the same set of patients generated by two different methods. To test for significance of the slope of the regression line in Fig. 3C, we used the Wald Test with t-distribution of the test statistic.

Supplementary Materials

Materials and Methods

Table S13. Assignment of features to feature groups.

References and Notes:

1. J. E. Posey, A. H. O’Donnell-Luria, J. X. Chong, T. Harel, S. N. Jhangiani, Z. H. Coban Akdemir, S. Buyske, D. Pehlivan, C. M. B. Carvalho, S. Baxter, N. Sobreira, P. Liu, N. Wu, J. A. Rosenfeld, S. Kumar, D. Avramopoulos, J. J. White, K. F. Doheny, P. D. Witmer, C. Boehm, V. R. Sutton, D. M. Muzny, E. Boerwinkle, M. Günel, D. A. Nickerson, S. Mane, D. G. MacArthur, R. A. Gibbs, A. Hamosh, R. P. Lifton, T. C. Matise, H. L. Rehm, M. Gerstein, M. J. Bamshad, D. Valle, J. R. Lupski, Centers for Mendelian Genomics, Insights into genetics, human biology and disease gleaned from family based genomic studies, *Genet. Med.* **21**, 798–812 (2019).

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

3. OMIM Gene Map Statistics (available at https://omim.org/statistics/geneMap).

4. A. Iglesias, K. Anyane-Yeboa, J. Wynn, A. Wilson, M. Truitt Cho, E. Guzman, R. Sisson, C. Egan, W. K. Chung, The usefulness of whole-exome sequencing in routine clinical practice, *Genet. Med.* **16**, 922–931 (2014).

5. F. E. Dewey, M. E. Grove, C. Pan, B. A. Goldstein, J. A. Bernstein, H. Chaib, J. D. Merker, R. L. Goldfeder, G. M. Enns, S. P. David, N. Pakdaman, K. E. Ormond, C. Caleshu, K. Kingham, T. E. Klein, M. Whirl-Carrillo, K. Sakamoto, M. T. Wheeler, A. J. Butte, J. M. Ford, L. Boxer, J. P. A. Ioannidis, A. C. Yeung, R. B. Altman, T. L. Assimes, M. Snyder, E. A. Ashley, T. Quertermous, Clinical interpretation and implications of whole-genome sequencing, *JAMA* **311**, 1035–1045 (2014).

6. J. S. Amberger, C. A. Bocchini, A. F. Scott, A. Hamosh, OMIM.org: leveraging knowledge across phenotype-gene relationships, *Nucleic Acids Res.* **47**, D1038–D1043 (2019).

7. S. Pavan, K. Rommel, M. E. Mateo Marquina, S. Höhn, V. Lanneau, A. Rath, Clinical practice guidelines for rare diseases: The Orphanet Database, *PLoS ONE* **12**, e0170365 (2017).

8. P. D. Stenson, M. Mort, E. V. Ball, K. Evans, M. Hayden, S. Heywood, M. Hussain, A. D. Phillips, D. N. Cooper, 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.* **136**, 665–677 (2017).

9. S. Köhler, L. Carmody, N. Vasilevsky, J. O. B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N. L. Harris, N. Matentzoglu, J. A. McMurry, D. Osumi-Sutherland, V. Cipriani, J. P. Balhoff, T. Conlin, H. Blau, G. Baynam, R. Palmer, D. Gratian, H. Dawkins, M. Segal, A. C. Jansen, A. Muaz, W. H. Chang, J. Bergerson, S. J. F. Laulederkind, Z. Yüksel, S. Beltran, A. F. Freeman, P. I. Sergouniotis, D. Durkin, A. L. Storm, M. Hanauer, M. Brudno, S. M. Bello, M. Sincan, K. Rageth, M. T. Wheeler, R. Oegema, H. Lourghi, M. G. Della Rocca, R. Thompson, F. Castellanos, J. Priest, C. Cunningham-Rundles, A. Hegde, R. C. Lovering, C. Hajek, A. Olry, L. Notarangelo, M. Similuk, X. A. Zhang, D. Gómez-Andrés, H. Lochmüller, H. Dollfus, S. Rosenzweig, S. Marwaha, A. Rath, K. Sullivan, C. Smith, J. D. Milner, D. Leroux, C. F. Boerkoel, A. Klion, M. C. Carter, T. Groza, D. Smedley, M. A. Haendel, C. Mungall, P. N. Robinson, Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources, *Nucleic Acids Res* **47**, D1018–D1027 (2018).

10. M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, D. R. Maglott, ClinVar: improving access to variant interpretations and supporting evidence, *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

11. K. A. Jagadeesh, J. Birgmeier, H. Guturu, C. A. Deisseroth, A. M. Wenger, J. A. Bernstein, G. Bejerano, Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization, *Genetics in Medicine* **21**, 464–470 (2019).

12. S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, P. N. Robinson, Clinical Diagnostics in Human Genetics with Semantic Similarity Searches in Ontologies, *Am J Hum Genet* **85**, 457–464 (2009).

13. M. V. Singleton, S. L. Guthery, K. V. Voelkerding, K. Chen, B. Kennedy, R. L. Margraf, J. Durtschi, K. Eilbeck, M. G. Reese, L. B. Jorde, C. D. Huff, M. Yandell, Phevor Combines Multiple Biomedical Ontologies for Accurate Identification of Disease-Causing Alleles in Single Individuals and Small Nuclear Families, *The American Journal of Human Genetics* **94**, 599–610 (2014).

14. D. Smedley, J. O. B. Jacobsen, M. Jäger, S. Köhler, M. Holtgrewe, M. Schubach, E. Siragusa, T. Zemojtel, O. J. Buske, N. L. Washington, W. P. Bone, M. A. Haendel, P. N. Robinson, Next-generation diagnostics and disease-gene discovery with the Exomiser, *Nat. Protocols* **10**, 2004–2015 (2015).

15. H. Yang, P. N. Robinson, K. Wang, Phenolyzer: phenotype-based prioritization of candidate genes for human diseases, *Nat Methods* **12**, 841–843 (2015).

16. A. Javed, S. Agrawal, P. C. Ng, Phen-Gen: combining phenotype and genotype to analyze rare disorders, *Nat. Methods* **11**, 935–937 (2014).

17. A. Sifrim, D. Popovic, L.-C. Tranchevent, A. Ardeshirdavani, R. Sakai, P. Konings, J. R. Vermeesch, J. Aerts, B. De Moor, Y. Moreau, eXtasy: variant prioritization by genomic data fusion, *Nat Meth* **10**, 1083–1084 (2013).

18. T. Fujiwara, Y. Yamamoto, J.-D. Kim, O. Buske, T. Takagi, PubCaseFinder: A Case-Report-Based, Phenotype-Driven Differential-Diagnosis System for Rare Diseases, *The American Journal of Human Genetics* **103**, 389–399 (2018).

19. C. A. Deisseroth, J. Birgmeier, E. E. Bodle, J. N. Kohler, D. R. Matalon, Y. Nazarenko, C. A. Genetti, C. A. Brownstein, K. Schmitz-Abe, K. Schoch, H. Cope, R. Signer, J. A. Martinez-Agosto, V. Shashi, A. H. Beggs, M. T. Wheeler, J. A. Bernstein, G. Bejerano, ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis, *Genetics in Medicine* **ePub**, doi:10.1038/s41436-018-0381-1 (2018).

20. J. H. Son, G. Xie, C. Yuan, L. Ena, Z. Li, A. Goldstein, L. Huang, L. Wang, F. Shen, H. Liu, K. Mehl, E. E. Groopman, M. Marasa, K. Kiryluk, A. G. Gharavi, W. K. Chung, G.

Hripcsak, C. Friedman, C. Weng, K. Wang, Deep Phenotyping on Electronic Health Records Facilitates Genetic Diagnosis by Clinical Exomes, *The American Journal of Human Genetics* **103**, 58–73 (2018).

21. A. M. Wenger, H. Guturu, J. A. Bernstein, G. Bejerano, Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers, *Genet Med* **19**, 209–214 (2016).

22. N. Carmichael, J. Tsipis, G. Windmueller, L. Mandel, E. Estrella, “Is it going to hurt?”: the impact of the diagnostic odyssey on children and their families, *J Genet Couns* **24**, 325–335 (2015).

23. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning* (Springer, 2009; http://www.springer.com/us/book/9780387848570).

24. B. Yates, B. Braschi, K. A. Gray, R. L. Seal, S. Tweedie, E. A. Bruford, Genenames.org: the HGNC and VGNC resources in 2017, *Nucleic Acids Res.* **45**, D619–D625 (2017).

25. A. Bateman, M. J. Martin, C. O’Donovan, M. Magrane, E. Alpi, R. Antunes, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, H. Bye-A-Jee, A. Cowley, A. D. Silva, M. D. Giorgi, T. Dogan, F. Fazzini, L. G. Castro, L. Figueira, P. Garmiri, G. Georghiou, D. Gonzalez, E. Hatton-Ellis, W. Li, W. Liu, R. Lopez, J. Luo, Y. Lussi, A. MacDougall, A. Nightingale, B. Palka, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, A. Renaux, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Speretta, E. Turner, N. Tyagi, V. Volynkin, T. Wardell, K. Warner, X. Watkins, R. Zaru, H. Zellner, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.-C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. de Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli,

L. Verbregue, A.-L. Veuthey, C. H. Wu, C. N. Arighi, L. Arminski, C. Chen, Y. Chen, J. S. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Q. Wang, Y. Wang, L.-S. Yeh, J. Zhang, UniProt: the universal protein knowledgebase, *Nucleic Acids Res* **45**, D158–D169 (2017).

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

27. J. Birgmeier, C. A. Deisseroth, L. E. Hayward, L. M. T. Galhardo, A. P. Tierno, K. A. Jagadeesh, P. D. Stenson, D. N. Cooper, J. A. Bernstein, M. Haeussler, G. Bejerano, AVADA: toward automated pathogenic variant evidence retrieval directly from the full-text literature, *Genet. Med.* **ePub** (2019), doi:10.1038/s41436-019-0643-6.

28. K. A. Jagadeesh, D. J. Wu, J. A. Birgmeier, D. Boneh, G. Bejerano, Deriving genomic diagnoses without revealing patient genomes, *Science* **357**, 692–695 (2017).

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

30. M. Lek, K. J. Karczewski, E. V. Minikel, K. E. Samocha, E. Banks, T. Fennell, A. H. O’Donnell-Luria, J. S. Ware, A. J. Hill, B. B. Cummings, T. Tukiainen, D. P. Birnbaum, J. A. Kosmicki, L. E. Duncan, K. Estrada, F. Zhao, J. Zou, E. Pierce-Hoffman, J. Berghout, D. N. Cooper, N. Deflaux, M. DePristo, R. Do, J. Flannick, M. Fromer, L. Gauthier, J. Goldstein, N. Gupta, D. Howrigan, A. Kiezun, M. I. Kurki, A. L. Moonshine, P. Natarajan, L. Orozco, G. M. Peloso, R. Poplin, M. A. Rivas, V. Ruano-Rubio, S. A. Rose, D. M. Ruderfer, K. Shakir, P. D. Stenson, C. Stevens, B. P. Thomas, G. Tiao, M. T. Tusie-Luna, B. Weisburd, H.-H. Won, D. Yu, D. M. Altshuler, D. Ardissino, M. Boehnke, J. Danesh, S. Donnelly, R. Elosua, J. C. Florez, S. B. Gabriel, G. Getz, S. J. Glatt, C. M. Hultman, S. Kathiresan, M. Laakso, S. McCarroll, M. I. McCarthy, D. McGovern, R. McPherson, B. M. Neale, A. Palotie, S. M. Purcell, D. Saleheen, J. M. Scharf, P. Sklar, P. F. Sullivan, J. Tuomilehto, M. T. Tsuang, H. C. Watkins, J. G. Wilson, M.

J. Daly, D. G. MacArthur, Exome Aggregation Consortium, Analysis of protein-coding genetic variation in 60,706 humans, *Nature* **536**, 285–291 (2016).

31. S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, D. B. Goldstein, Genic Intolerance to Functional Variation and the Interpretation of Personal Genomes, *PLoS Genet* **9**, e1003709 (2013).

32. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation, *Nature* **526**, 68–74 (2015).

33. Deciphering Developmental Disorders Study, Large-scale discovery of novel genetic causes of developmental disorders, *Nature* **519**, 223–228 (2015).

34. E. Birney, J. Vamathevan, P. Goodhand, Genomics in healthcare: GA4GH looks to 2022, *bioRxiv* , 203554 (2017).

35. O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Res.* **32**, D267-270 (2004).

36. F. B. Rogers, Medical subject headings, *Bull Med Libr Assoc* **51**, 114–116 (1963).

37. SNOMED CT (available at https://www.nlm.nih.gov/healthit/snomedct/).

38. J. Piñero, N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, L. I. Furlong, DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes, *Database (Oxford)* **2015** (2015), doi:10.1093/database/bav028.

39. K. M. D. Cornett, M. P. Menezes, P. Bray, M. Halaki, R. R. Shy, S. W. Yum, T. Estilow, I. Moroni, M. Foscan, E. Pagliano, D. Pareyson, M. Laurá, T. Bhandari, F. Muntoni, M. M. Reilly, R. S. Finkel, J. Sowden, K. J. Eichinger, D. N. Herrmann, M. E. Shy, J. Burns, Inherited Neuropathies Consortium, Phenotypic Variability of Childhood Charcot-Marie-Tooth Disease, *JAMA Neurol* **73**, 645–651 (2016).

40. I. Lappalainen, J. Almeida-King, V. Kumanduri, A. Senf, J. D. Spalding, S. ur-Rehman, G. Saunders, J. Kandasamy, M. Caccamo, R. Leinonen, B. Vaughan, T. Laurent, F. Rowland, P. Marin-Garcia, J. Barker, P. Jokinen, A. C. Torres, J. R. de Argila, O. M. Llobet, I. Medina, M. S. Puy, M. Alberich, S. de la Torre, A. Navarro, J. Paschall, P. Flicek, The European Genome-phenome Archive of human data consented for biomedical research, *Nat Genet* **47**, 692–695 (2015).

41. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: machine learning in python, *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

42. M. Haeussler, *Download, convert and process the full text of scientific articles: maximilianh/pubMunch3* (2018; https://github.com/maximilianh/pubMunch3).

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

44. O. Tange, Gnu parallel - the command-line power tool, *The USENIX Magazine* **36**, 42–47 (2011).

45. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv:1303.3997 [q-bio]* (2013) (available at http://arxiv.org/abs/1303.3997).

46. M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, M. J. Daly, A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nat. Genet.* **43**, 491–498 (2011).

47. T. Zemojtel, S. Köhler, L. Mackenroth, M. Jäger, J. Hecht, P. Krawitz, L. Graul-Neumann, S. Doelken, N. Ehmke, M. Spielmann, N. C. Øien, M. R. Schweiger, U. Krüger, G. Frommer, B. Fischer, U. Kornak, R. Flöttmann, A. Ardeshirdavani, Y. Moreau, S. E. Lewis, M.

Haendel, D. Smedley, D. Horn, S. Mundlos, P. N. Robinson, Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome, *Science Translational Medicine* **6**, 123 (2014).

48. P. N. Robinson, S. Köhler, A. Oellrich, Sanger Mouse Genetics Project, K. Wang, C. J. Mungall, S. E. Lewis, N. Washington, S. Bauer, D. Seelow, P. Krawitz, C. Gilissen, M. Haendel, D. Smedley, Improved exome prioritization of disease genes through cross-species phenotype comparison, *Genome Res.* **24**, 340–348 (2014).

49. A. Singhal, M. Simmons, Z. Lu, Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine, *PLoS Comput. Biol.* **12**, e1005017 (2016).

50. E. Doughty, A. Kertesz-Farkas, O. Bodenreider, G. Thompson, A. Adadey, T. Peterson, M. G. Kann, Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature, *Bioinformatics* **27**, 408–415 (2011).

51. W. Xing, J. Qi, X. Yuan, L. Li, X. Zhang, Y. Fu, S. Xiong, L. Hu, J. Peng, A gene–phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach, *Bioinformatics* **34**, i386–i394 (2018).

52. A. Coulet, N. Shah, Y. Garten, M. Musen, R. B. Altman, Using text to build semantic networks for pharmacogenomics, *J Biomed Inform* **43**, 1009–1019 (2010).

53. C.-H. Wei, H.-Y. Kao, Z. Lu, GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains, *Biomed Res Int* **2015** (2015), doi:10.1155/2015/918710.

54. D. Campos, S. Matos, I. Lewin, J. L. Oliveira, D. Rebholz-Schuhmann, Harmonization of gene/protein annotations: towards a gold standard MEDLINE, *Bioinformatics* **28**, 1253–1261 (2012).

55. H.-W. Chun, Y. Tsuruoka, J.-D. Kim, R. Shiba, N. Nagata, T. Hishiki, J. Tsujii, Extraction of gene-disease relations from Medline using domain dictionaries and machine learning, *Pac Symp Biocomput* **2006**, 4–15 (2006).

56. A. Özgür, T. Vu, G. Erkan, D. R. Radev, Identifying gene-disease associations using centrality on a literature mined gene-interaction network, *Bioinformatics* **24**, i277–i285 (2008).

57. T. C. Rindflesch, L. Tanabe, J. N. Weinstein, L. Hunter, EDGAR: Extraction of Drugs, Genes And Relations from the Biomedical Literature, *Pac Symp Biocomput* **2000**, 517–528 (2000).

58. D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, D. S. Wishart, PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites, *Nucleic Acids Res.* **36**, W399-405 (2008).

59. N. Collier, T. Groza, D. Smedley, P. N. Robinson, A. Oellrich, D. Rebholz-Schuhmann, PhenoMiner: from text to a database of phenotypes associated with OMIM diseases, *Database (Oxford)* **2015** (2015), doi:10.1093/database/bav104.

60. J. Kim, J. Kim, H. Lee, An analysis of disease-gene relationship from Medline abstracts by DigSee, *Scientific Reports* **7**, 40154 (2017).

Acknowledgments: We thank Dr. Kai Wang for his assistance with ANNOVAR and Erich Weiler for support and guidance. We thank Paul McDonagh and Thomas Defay, as well as the members of the Bejerano lab, for technical advice and helpful discussions. We thank E. Kravets, J. Buckingham and K. MacMillen for assistance with obtaining patient data. We thank P.B. Agrawal, C.A. Brownstein, M. Danowski, C.A. Genetti, J.A. Madden, N. Nori, H. Paterson, K. Schmitz-Abe and T. Yu from the Manton Center for their work. We thank all data sources used in AMELIE, including HPO, Ensembl, HGNC, UniProt, OMIM, ClinVar, HGMD, PubTator, ExAC, pLI, RVIS, M-CAP, and the GWAS catalog. We thank the European Genome-Phenome Archive ***(40)*** (EGA) and the Deciphering Developmental Disorders ***(33)*** (DDD) project. We thank Pankaj B. Agrawal, Catherine A. Brownstein, Morgan Danowski, Casie A. Genetti, Jill A. Madden, Neeharika Nori, Heather Paterson, Klaus Schmitz-Abe, Timothy Yu from The Manton Center for Orphan Disease Research for providing patient data and advice. The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. C.R. is affiliated with SambaNova Systems and Apple**. Funding:** All computational work was funded in part by a Bio-X SIGF fellowship to J.B., DARPA (C.R., G.B.), the Stanford Pediatrics Department (J.A.B., G.B.), a Packard Foundation Fellowship (G.B.), a Microsoft Faculty Fellowship (G.B.), NHGRI U41HG002371-15 (M.H.) and the Stanford Data Science Initiative (G.B., C.R.). Manton Center sequence analysis and diagnosis was supported by NIH 1U54HD090255 IDDRC Molecular Genetics Core grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. 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. **Author contributions:** J.B. and G.B. designed the study and analyzed results. J.B. implemented the text mining software, website and associated databases. M.H. authored PubMunch. J.B., E.H.S., H.G., A.M.W., M.E.D., K.A.J., and

C.A.D. wrote and improved software tools that were used for genotype and phenotype analysis. C.A.D. analyzed EGA data. A.J.R. wrote parts of the gene and phenotype identification. P.D.S. and D.N.C. curated the HGMD data. C.R. provided text mining guidance. The Manton Center and A.H.B. provided Manton patient data from the Gene Discovery Core. J.A.B. provided guidance on clinical aspects of study design, testing set construction and interpretation of results. J.B. and G.B. wrote the manuscript. G.B. supervised the project. All authors commented on and approved the manuscript.

Competing interests: D.N.C. and P.D.S. acknowledge the receipt of financial support from Qiagen Inc through a License Agreement with Cardiff University. C.R. is affiliated with SambaNova Systems and Apple Inc. The other authors declare they have no competing interests. **Data and materials availability:** All data associated with this study is present in the paper or supplementary materials. Data from public databases can be downloaded using the provided links and materials. Deidentified DDD data was obtained through EGA study number EGAS00001000775. Data for Stanford and Manton Center patients require patient consent and are available upon reasonable request, in agreement with the Stanford Children's Health and Manton Center for Orphan Disease Research clinical coordinators and Institutional Review Boards. The AMELIE portal and code are available at https://AMELIE.stanford.edu and https://zenodo.org/deposit/3707012.

Figures:

Figure 1. AMELIE knowledgebase creation and subsequent patient causal gene ranking classifier.

(A) AMELIE knowledgebase creation. AMELIE applies multiple machine learning classifiers to all (current) 29 million PubMed abstracts to parse, predict relevance, download full text, and finally extract Mendelian gene-phenotype relationships and related attributes automatically. **(B)** Number of gene-phenotype relationships extracted from the 10 journals that AMELIE extracted most gene-phenotype relationships from. **(C)** The AMELIE classifier combines 27 features to rank all articles in the AMELIE knowledgebase for their ability to explain any input patient.

Figure 2. AMELIE patient causative gene ranking outperforms methods based on manually curated databases.

(A) Evaluation scheme. The output gene ranking of all algorithms was subset to the same list of candidate genes AMELIE uses for its gene ranking to ensure a fair comparison. **(B)** Fraction of (n=215) DDD cases ranked as 1, 1-2 or 1-3 by six different tools. **(C)** The number of top-ranked genes needed to achieve a 90% diagnosis rate across (n=215) DDD cases by various gene ranking tools. By evaluating up to AMELIE's 11th top-ranked gene, a 90% diagnosis yield on the DDD cases was achieved. The next best tool, Exomiser, achieved a 90% diagnosis yield by evaluating up to Exomiser's 30th gene. **(D)** The speedup in terms of number of genes to investigate when perusing the ranked gene lists provided by each tool from top to bottom until the causative gene was found, compared to the expected value of a random baseline gene ordering for (n=215) DDD cases.

Figure 3. Investigating AMELIE's gene ranking performance.

(A) For each of the 175 patients with AMELIE causative gene rank 1 among all (n=271) real DDD, Stanford, and Manton patients, the 27 features to the AMELIE classifier were ranked by their contribution to the top-ranked article's high score. The panels, left-to-right, show the fraction of patients for which certain features were ranked most-, 2nd most-, or 3rd most-contributing. PTV: protein-truncating variant; NTV: Non-protein-truncating variant, MCAP: Mendelian clinically applicable pathogenicity score, an in-silico pathogenicity score; PV: patient variant; het: heterozygous; EV: full text article-extracted variant. **(B)** Re-training the AMELIE classifier with 5-fold cross-validation, each time omitting one of AMELIE’s 6 feature groups, shows the degree to which feature groups aided performance across all (n=271) DDD, Stanford and Manton patients. **(C)** Each blue dot represents one of (n=271) real DDD, Stanford, or Manton patients in this log-log plot. The red line is a linear regression line between number of articles about causative gene (x-axis) and causative gene rank (y-axis), with red denoting the 95% confidence interval.

Supplementary Materials

Materials and Methods

Parsing all of PubMed to construct a knowledgebase about Mendelian diseases

Downloading titles and abstracts from PubMed

PubMed is an online database that contains titles and abstracts of peer-reviewed biomedical articles. We downloaded titles and abstracts directly from PubMed (ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline/ and ftp://ftp.ncbi.nlm.nih.gov/pubmed/updatefiles/).

Identifying phenotype mentions in titles and abstracts

Human Phenotype Ontology *(9)* (HPO) provides a standardized vocabulary of phenotypic abnormalities encountered in human genetic disease. Phenotypic abnormalities are stored in HPO with a unique identifier, a canonical name and an optional list of synonyms. To identify phenotype mentions in English text, an AMELIE database of 45,435 phenotype phrases (names and synonyms of phenotypes and simplified versions thereof) corresponding to 13,439 HPO phenotypic abnormalities was created using phenotype names and synonyms from the HPO ontology version releases/2018-07-25, which we use throughout this manuscript. HPO names and synonyms are stored in the database in both exact and lemmatized forms. Lemmatization is an algorithmic process that reduces inflection forms of words to a common base form of the word. For example, the word “ovaries” is reduced to the singular form “ovary” after lemmatization. Lemmatized versions of words are created by the Python 3.7.0 NLTK version 3.2.5 WordNetLemmatizer. Stopwords are short words such as “or”, “of”, “a”, “the”, etc. 126 common stopwords were removed from all lemmatized phenotype names in the AMELIE database of phenotype names.

Before any further processing for both phenotype and gene identification (described below), all text is passed through a filter replacing Greek letters and non-alphanumeric characters with canonical symbols, and sentence and word tokenization using Python 3.7.0 NLTK version 3.2.5. Sentence and word tokenization splits a single stream of text into individual sentences and words. To extract phenotypes from English text, all permutations of consecutive word groups of length 1 to 8 are systematically matched against the AMELIE database of phenotype names in both lemmatized forms (after removal of stopwords) and exact forms. If a word group matches, it is associated with the appropriate HPO identifier.

Identifying gene mentions in titles and abstracts

Genes and their protein products are identified by various names in English text. Both the HUGO Gene Nomenclature Committee (HGNC) *(24)* and the UniProt *(25)* database maintain a list of gene and protein names. To identify mentions of human genes or their protein products in text, an AMELIE database of gene names, consisting of 188,975 gene and protein names for 21,346 distinct protein-coding Ensembl version 84 genes and 18,149 non-protein-coding genes (which are largely used as negative training examples for the classifiers described below) was compiled from HGNC and UniProt. Data for mapping Ensembl genes to HGNC and UniProt was downloaded from BioMart (http://uswest.ensembl.org/biomart/) on October 28, 2016. AMELIE identifies gene mentions in articles by matching words groups in the article against the AMELIE database of gene names. Mentions of both gene names and names of their protein products are treated equally, and referred to as “gene mentions”. Further, PubTator *(26)* is an online tool that uses an ensemble of automatic text processing tools to identify gene mentions in PubMed titles and abstracts. Gene mentions identified by PubTator are used to supplement gene identification using the list of gene and protein names described above.

To identify gene mentions in English text, all word groups of length 1-8 in the text are examined. Given an article identified by a PubMed ID, the database of gene and protein names is supplemented by the gene and protein names stored in PubTator for this PubMed ID. To identify gene mentions, word groups from the article are matched against the AMELIE database of gene names (case-insensitive). If a word group matches an entry in the AMELIE gene names database, it is associated with the matching Ensembl ID(s) and corresponding HGNC gene name(s) stored in the database of gene and protein names.

Constructing a title/abstract document classifier

PubMed contains titles and abstracts, but not the full text of articles. AMELIE uses a document classifier to discover articles that may be relevant to Mendelian diseases using the text in the article’s title and abstract, as well as gene and phenotype names discovered in the title and abstract. It then downloads the full text of potentially relevant articles.

To classify titles and abstracts as potentially relevant, gene and phenotype mentions were identified in the text as described above. Subsequently, all gene mentions were replaced by the token “XGENE” and all phenotype mention were replaced by the token “XPHENO” except for phenotypes descending from the following subtrees of HPO: neoplasm, prostate cancer, schizophrenia, abnormality of DNA repair (HP:0002664, HP:0012125, HP:0100753, HP:0003254, respectively). These phenotypes are often not transmitted in a Mendelian fashion, and occur in a large body of literature that is not of interest for Mendelian diagnosis. All words in the title were prefixed with “TITLE_” and all words in the abstract were prefixed with “ABSTRACT_”.

To transform a document into a feature vector, we used the so-called TF-IDF transformation. A TF-IDF transformation treats each document as an unordered bag of words. The document is transformed into a feature vector by assigning each word the scalar product of two statistics: the term frequency (TF) of the word and the inverse document frequency (IDF) of the word. The term frequency *tf(w, d)* of a word *w* in a document *d* is defined to be the number of occurrences of *w* in *d*. The inverse document frequency of a word *w* in a document *d* is defined as

$$idf(w,d) = log\frac{1+n_d}{1+df(w)}+1,$$

where n_d is the total number of documents and *df(w)* is the number of documents that contain the word *w*. (See also http://scikit-learn.org/stable/modules/feature_extraction.html - text-feature-extraction). Then

$$tfidf(w,d) = tf(w,d) \times idf(w,d).$$

Here and below, to transform text data to a TF-IDF vector, we used a scikit-learn *(41)* version 0.20.0 TfidfVectorizer with default parameters.

To construct a training set for the title/abstract document classifier, we used data from OMIM gene entries and entries in HGMD. A “gene entry” in OMIM is an OMIM entry containing information on a particular gene, such as https://omim.org/entry/602635 (the OMIM entry for the gene *DEAF1*). “Allelic Variants” sections are sections in gene entries in the database OMIM (such as https://omim.org/entry/602635#allelicVariants). “Allelic Variants” sections in OMIM gene entries, amongst others, cite literature describing pathogenic mutations in the respective gene. HGMD is a database of pathogenic and disease-associated variants in the human genome that is curated from the literature. We used HGMD PRO version 2018.01. Allelic variants entries from OMIM were downloaded on September 29, 2018. Entries in HGMD reference PubMed IDs of articles from which the pathogenic variants were curated. The training set for the title/abstract document classifier was based on all articles cited in OMIM's "Allelic Variants" sections and in HGMD as positives, and random negative articles from PubMed. Titles and abstracts in the training set were TF-IDF transformed, labeled with 1 ("relevant") or 0 ("irrelevant") and used to train a scikit-learn *(41)* version 0.20.0 LogisticRegression classifier with a maximum of 1000 iterations and default parameters otherwise.

The document classifier was subsequently applied to TF-IDF-transformed titles and abstracts downloaded from PubMed. The classifier returns values between 0 and 1. Here and subsequently, classifiers are evaluated at cutoff value 0.5 to determine precision and recall (with samples with a classifier score >0.5 being classified as positives and samples with a classifier score <= 0.5 being classified as negatives).

Regularization of title/abstract document classifier

"Regularization" in machine learning refers to a process in which the machine learning classifier is penalized for learning complicated parameters that may fit the training data better than real-world data. For training logistic regression classifiers, two regularization schemes named "L2 regularization" and "L1 regularization" are commonly used. We chose the default L2 regularization for all classifiers, including the title/abstract document classifier, as L1 regularization of the title/abstract document classifier led to a slight drop in precision.

Downloading full text of relevant articles

Relevant documents identified using the title/abstract document classifier were downloaded using the pubCrawl2 software from the PubMunch3 version 1.0.3 package *(42)*. Given a PubMed ID, the PubMunch software attempts to retrieve the full text PDF of a scientific article directly from its publisher. Downloaded articles in PDF format were converted to text using pdftotext version 0.26.5 (https://poppler.freedesktop.org/).

Identifying genes and phenotypes in relevant articles' full text

Genes and phenotypes were identified in full text in the same way they were identified in titles and abstracts (see above). PubTator-identified gene names from titles and abstracts were also used to identify gene mentions in full text.

Determining sensitivity of gene name recognition in full text articles

To quantify the sensitivity of gene recognition using HGNC and UniProt, we selected all 39,431 articles cited in HGMD entries on disease-causing mutations that were classified as relevant by the title/abstract classifier for which we had downloaded the full text. Each HGMD entry contains the gene in which the disease-causing mutation occurs. Gene recognition with HGNC and UniProt alone recognized the name of the gene cited in HGMD in 89% of all articles. Augmenting the list of gene names using gene names deposited in PubTator, which relies on multiple previously published specialized gene recognition tools, increased sensitivity to 93%. In recent years, publishers have increasingly pushed to used HGNC gene names in publications; consequently, sensitivity of gene name recognition with UniProt and HGNC gene names alone increases to 92% in articles published since 2013, and to 94% if gene name lists were augmented with PubTator-extracted gene names (as done by AMELIE).

Constructing a relevant gene classifier

Many gene names recognized by AMELIE in an article are either false positives (e.g., "ANOVA") or not relevant for diagnosis of Mendelian diseases (e.g., a gene mentioned in passing). To identify the most relevant (disease-causing) gene(s) in each article, a "relevant gene" classifier was trained to recognize genes that are mentioned in an article as causing a phenotype when mutated. For example, the gene *NOTCH3* is the "relevant gene" of the article "Truncating mutations in the last exon of *NOTCH3* cause lateral meningocele syndrome" (PubMed ID 25394726) and thus has a high relevant gene score in this article. We use articles cited in OMIM "Allelic Variants" sections and HGMD to construct a training set for the "relevant gene" classifier.

Articles cited in OMIM "Allelic Variants" sections were associated with the corresponding OMIM gene in whose entry they were cited. For example, the article "Mutations affecting the SAND domain of DEAF1 cause intellectual disability with severe speech impairment and behavioral problems”, which was cited in the “Allelic Variants” sections of the OMIM entry on *DEAF1* (https://omim.org/entry/602635), was associated with *DEAF1* as its “relevant gene” in the positive training set. This process was repeated across all OMIM “Allelic Variants” sections to construct the positive training set. To further enlarge the positive training set, we used articles associated with variants cited in HGMD. Articles *A* with relevant genes *G* were added to the positive training set if there was a variant in HGMD associated with gene *G* and article *A.* Articles associated with more than 3 relevant genes were omitted from the positive training set. The negative training set consisted of all article-gene mappings where the article is in the positive training set, and the gene is mentioned in the article, but the article-gene mapping is not part of the positive training set. This resulted in 47,357 positive training examples and 919,255 negative training examples.

Given a tuple consisting of an article *A* and a gene *G*, the following features were constructed for the relevant gene classifier: number of mentions of the gene *G* in the title of the article *A,* number of mentions of the gene *G* in the abstract of the article *A*, number of mentions of the gene *G* in the full text of the article *A*, and TF-IDF-transformed word counts (defined above) in 5-word-windows around all mentions of the gene *G* in the full text of the article *A.*

A scikit-learn *(41)* 0.20.0 LogisticRegression classifier with default parameters was subsequently trained using this training set. The classifier was applied to all gene objects in all downloaded articles.

Full-text document classification assigns an article relevance based on the article’s full text

Even if the title and abstract of an article seem potentially relevant enough to download the article, the full text of an article may turn out not to be relevant for AMELIE. We trained a full-text classifier that assigns a likelihood of the article describing a link between genetic mutations and a Mendelian disease based on the article’s full text.

The training set for the full-text document classifier consisted of a large number of (positive) full-text articles that are relevant, and a large number of (negative) full-text articles that are presumably irrelevant. To create the negative training set, random PubMed IDs were selected from all available PubMed IDs that were not part of the positive training set of the title/abstract document classifier and downloaded using PubMunch. 54,023 positive examples were obtained using OMIM and HGMD and 159,665 random negative examples were obtained from all of PubMed. To convert each article's full text into a feature vector, recognized gene and phenotype names were replaced by tokens "XGENE" and "XPHENO" in the full text of each article. A scikit-learn *(41)* 0.20.0 LogisticRegression classifier with default parameters was subsequently trained on TF-IDF transformed documents to serve as the full-text document classifier. This full-text document classifier was applied to all downloaded articles.

Variant type classifier identifies mentions of loss-of-function and gain-of-function variants in articles

Two large classes of variants cause most known Mendelian diseases: variants that completely disrupt the wild-type transcript or translation into a chain of amino acids downstream of the mutation (including frameshift, stopgain and splicing mutations), and, variants that merely change a small portion of the resulting protein (including nonsynonymous and nonframeshift indel mutations). We call the former class "protein-truncating variants" (PTV) and the latter "non-truncating variants" (NTV).

"Allelic Variants" sections in OMIM gene entries list disease-causing genetic mutations and give a textual synopsis of original literature describing the mutation, including a reference to the original article. We used OMIM "Allelic Variants" sections to create a training set for the variant type classifier. To this end, the "Allelic Variants" sections of all available OMIM entries on genes were parsed to construct a training set consisting of article-to-variant-type mappings. "Allelic Variants" sections that described more than one causative mutation or were longer than one paragraph were discarded for simplicity. Parts of sentences delimited by commas and periods in "Allelic Variants" sections that contained the words "originally described" or "originally reported" were ignored because articles referenced in such sentences most often describe patient phenotypes without describing causative mutations.

To identify the variant type described in an "Allelic Variants" section, the mutation given in the "Allelic Variants" section was parsed using regular expression patterns covering missense, stoploss, splicing, deletion, duplication, and insertion variants (table S11). For all mutations fitting a pattern, from the paragraph describing the mutation, all mentioned PubMed IDs were extracted. If a single PubMed ID was extracted, it was labeled as either about NTVs (missense variants, deletion and duplication variants of length 3, 6, or 9) or PTVs (stopgain, stoploss, splicing, and all remaining deletion and duplication variants). This resulted in a training set of 5,131 articles about protein-truncating variants and 11,389 articles about non-truncating variants using data obtained from OMIM. Further, all available 3,178 PubMed IDs from the GWAS catalog v1.0 studies, release 2018-04-10 were added to the training set and labeled as neither about PTVs nor NTVs. A scikit-learn *(41)* 0.20.0 multi-class LogisticRegression classifier with default parameters were subsequently trained on TF-IDF transformed documents using the labeled article set. The trained classifier was applied to all downloaded, TF-IDF transformed articles.

Inheritance mode classifier identifies articles about dominant and recessive diseases

Mendelian diseases are inherited in a number of inheritance modes, notably autosomal dominant and recessive, and X-linked dominant and recessive. For simplicity, AMELIE uses only the notion of "dominant" and "recessive" to distinguish inheritance modes. For the purposes of AMELIE, a dominant disease can manifest itself if only one copy of the causative gene is mutated. A recessive disease manifests itself if all copies of the causative gene are mutated.

The training set for the inheritance mode classifiers consisted of a large number of article-to-inheritance-mode mappings. To create such a training set, the "Allelic Variants" sections of all available OMIM entries on genes were parsed. Allelic Variants sections were downloaded from OMIM on September 29, 2018. Each paragraph describing a mutation was first parsed sentence by sentence to detect keywords indicating a particular inheritance mode. Sentence tokenization was performed by Python 3.7.0 NLTK version 3.2.5. If a sentence contained any of the words "parent", "brother", "sister", "sibling", "family", "mother", or "father", the sentence was omitted. Sentences were subsequently split into words by whitespace. If any sentence in the paragraph contained the words "homozygous", "homozygote", "homozygosity", "recessive", or "compound het", the paragraph was marked as "recessive". If any sentence in the paragraph contained the words "heterozygous", "heterozygote", "heterozygosity", but not "compound heterozygous", the paragraph was marked as "dominant". Unmarked paragraphs or paragraphs marked as both "dominant" and "recessive" were omitted from all further analysis.

Subsequently, PubMed IDs of articles were parsed from each paragraph in the “Allelic Variants” sections. If a paragraph mentioned a single article, the article was associated with the inheritance mode extracted from the paragraph. This resulted in a training set of 6,157 “recessive” articles and 4,276 “dominant” articles using data obtained from OMIM. Similar to the variant type classifier, all available 3,178 PubMed IDs from the GWAS catalog v1.0 studies, release 2018-04-10, were added to the training set as being neither about dominant nor recessive diseases. A scikit-learn *(41)* 0.20.0 multi-class LogisticRegression classifier with default parameters were subsequently trained on TF-IDF transformed documents and applied to all downloaded articles.

Determining precision and recall of information extracted from full text

To estimate the precision of the phenotype identifier on full text, 50 automatically identified phenotype mentions were randomly selected from all downloaded full-text articles and the number of correctly identified phenotypes was counted, resulting in a precision of 74% (table S12). A mention was defined as correct if the word group occurred referred to a phenotype and the associated HPO ID referred to the mentioned phenotype.

5-fold cross-validation was performed using the Python 3.7.0 scikit-learn 0.20.0 function sklearn.model_selection.cross_validate. 5-fold cross-validation of the relevant gene classifier resulted in an average precision of 87% and an average recall of 76%. 5-fold cross-validation of the full-text document relevance classifier returned an average precision of 96% and an average recall of 91%.

For 5-fold cross-validation of the protein-truncating/non-truncating classifier and the dominant/recessive classifier, precision and recall were computed using micro-averaging, in which all outcomes (here, “PTV”, “NTV”, or “neither”) have an equal contribution to the final precision and recall scores. 5-fold cross-validation of the protein-truncating/non-truncating variant type classifier returned an average precision and recall of 79% each. Similarly, 5-fold cross-validation of the dominant/recessive classifier returned an average precision and recall of 83%.

Variant identification in full text

We used AVADA *(27)* to extract variant mentions from full text and convert them to genomic coordinates.

Gene-phenotype statistics

When reporting number of genes and number of gene-phenotype statistics in the AMELIE knowledgebase, we counted HGNC gene symbols and relationships between HGNC gene symbols and HPO phenotypes in the AMELIE knowledgebase except where otherwise noted.

Training a classifier to assign genes a likelihood of being causative

Variant filtering to arrive at candidate causative variants and genes

ANNOVAR *(43)* (Version: $Date: 2017-06-01 23:07:59 -0400 (Thu, 1 Jun 2017) $) was used to annotate patient variants with predicted effect and frequency information. ANNOVAR supporting data for genome assembly GRCh37/hg19 was downloaded from the ANNOVAR website on August 28, 2017 using the command "./annovar/annotate_variation.pl -downdb" for ANNOVAR databases refGene, knownGene, and ensGene. Patient variants are annotated with frequency information from ExAC version 0.3 *(30)* and the 1000 Genomes Project (KGP) phase 3 data *(32)*, as previously described *(21)*. A variant was considered rare if the allele frequency was less than 0.5% in the ExAC and KGP control databases and if it occurred in at most 1 homozygous person in ExAC and KGP. Single heterozygous variants were only considered rare if the allele frequency was at most 0.1% in ExAC and KGP and if the allele count was at most 3. Rare missense, core splice-site (defined as the 2 basepairs at either end of each intron), frameshift, nonframeshift indel, stop-gain and stop-loss variants in Ensembl protein-coding genes were considered to be candidate causative variants. Genes containing candidate causative variants are considered candidate causative genes (or "candidate genes" for short).

Estimating the fraction of known disease-causing mutations that AMELIE can rank

To estimate the fraction of clinically relevant variants that AMELIE can rank using this variant filtering scheme, we annotated all 164,618 disease-causing ("DM") variants available in HGMD version 2018.01 with predicted effect on genomic regions using the ANNOVAR predicted variant effects column "Gene.gene". Of these, 146,423 (89%) fell into exonic regions of protein-coding genes, and 12,590 (8%) fell into core splice-site regions of protein-coding genes. Only 4,469 (2.7%) of disease-causing variants fell into deeper intronic regions and 1,136 (0.69%) fell into other genomic regions. Thus, AMELIE retains approximately 97% of known disease-causing variants.

Discovery of articles about each of the candidate causative genes

Given a patient's list of candidate genes, for each candidate gene *G*, AMELIE analyzes all articles *A* if the relevant gene score of *G* in *A* is at least 0.1.

AMELIE classifier feature set construction

Each patient is associated with a set of phenotypes *P*, a list of candidate causative genes *G*, and a list of all AMELIE knowledgebase articles *A* about each candidate gene *G*. Each triple *(P, G, A)* was transformed into a feature vector that enabled the AMELIE classifier to calculate a likelihood that the article *A* explains the patient's phenotypes *P* in light of the patient's mutation(s) in *G*. Each such triple *(P, G, A)* is associated with a set of 27 real-valued features that enable the AMELIE classifier to calculate a likelihood that the article *A* explains how the patient's variants in *G* cause the patient's phenotypes *P* (Figure 1c). Two of these features use the Phrank *(11)* phenotypic match score. Phrank quantifies the overlap between any two sets of HPO phenotypic abnormalities and needs to be initialized with a set of known gene-phenotype associations. We initialized Phrank by using Ensembl ID-phenotype relationships derived from all Ensembl ID-article-phenotype relationships in the AMELIE knowledgebase if the Ensembl ID had a relevant gene score of at least 0.5 in the article and the article contained at most 100 distinct HPO phenotypes. Given two sets of HPO phenotypic abnormalities P_1 and P_2, Phrank assigns a scalar phenotype match score to the two sets. We denote the Phrank score of two sets of phenotypes with $match(P_1, P_2)$. We use the following abbreviations: "PV" = "patient variant", "PTV" = "protein-truncating variant", "NTV" = "non-truncating variant", "EV" = "extracted variant" (variant mentioned in paper and extracted by AVADA). The 27 features of the AMELIE classifier are as follows:

- **Feature 1:** M-CAP *(29)* ("MCAP score"): the average M-CAP score of all candidate causative variants in *G*. Since M-CAP provides scores only for rare missense variants, all variants for which no M-CAP score exist are assigned the M-CAP$_{100}$ score described below. We used M-CAP version 1.0, downloadable from http://bejerano.stanford.edu/mcap/downloads/dat/mcap_v1_0.txt.gz.

- **Feature 2:** M-CAP$_{100}$ ("100bp-average MCAP"): the arithmetic average of all available M-CAP scores in a window of -50, +50 basepairs adjacent to all candidate causative variants in *G*. If the M-CAP$_{100}$ score does not exist, it is replaced by the M-CAP$_{gene}$ score described below.
- **Feature 3:** M-CAP$_{gene}$ ("MCAP averaged over gene"): the arithmetic average of all available M-CAP scores for *G*. If the M-CAP$_{gene}$ score does not exist, it is replaced by the value 0.0.
- **Feature 4:** ("RVIS score") the RVIS *(31)* score of *G*, or 100.0 if the RVIS score of *G* is not available. RVIS scores, version May 2015, were obtained from http://genic-intolerance.org/data/RVIS_Unpublished_ExAC_May2015.txt .
- **Feature 5:** ("pLI score") the pLI *(30)* score of *G*, or 0.0 if the pLI score of *G* is not available. pLI scores version March 2016, for ExAC version 0.3, were obtained from ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/functional_gene_constraint/forweb_cleaned_exac_r03_march16_z_data_pLI.txt .
- **Feature 6:** ("Patient PTV score") the average protein-truncating score of all candidate causative variants in *G*. The protein-truncating score of a candidate causative variant is defined to be 1.0 if the variant is a splicing, stopgain, stoploss or frameshift variant, and 0.0 otherwise.
- **Feature 7:** ("Average PV allele count") the average ExAC *(30)* allele count of all variants in *G*.
- **Feature 8:** ("Average PV allele count if PV het") the ExAC allele count of the variant in *G* if it is a single heterozygous variant, and 0.0 otherwise.
- **Feature 9:** ("PV is single heterozygous") 1.0 if there is a single heterozygous variant in *G*, and 0.0 otherwise.
- **Feature 10:** ("Number of EV in gene") the number of variants in *G* that are mentioned in *A*.
- **Feature 11:** ("Max relevant gene in paper") the highest relevant gene score extracted from *A*.
- **Feature 12:** ("#Genes with EV in paper") the number of genes with at least one extracted genetic variant in *A*.

- **Feature 13:** ("Number of total EV in paper") the total number of genetic variants extracted from *A*.
- **Feature 14:** ("Paper PTV score") the PTV score of *A*.
- **Feature 15:** ("Paper NTV score") the NTV score of *A*.
- **Feature 16:** ("Paper dominant score") the dominant score of *A*.
- **Feature 17:** ("Paper recessive score") the recessive score of *A*.
- **Feature 18:** ("Full-text document relevance score") the full document relevance score of *A*.
- **Feature 19:** ("Phrank score (paper)") the Phrank score of the patient's phenotypes *P* with the phenotypes mentioned in the article (denoted by *phenotypes(A)*), divided by the maximum Phrank score of the patient:

$$\frac{match(P, phenotypes(A))}{match(P, P)}$$

- **Feature 20:** ("Phrank score (patient)") the Phrank score of the patient's phenotypes *P* with the phenotypes mentioned in the article (denoted by *phenotypes(A)*), divided by the maximum Phrank score of the article:

$$\frac{match(P, phenotypes(A))}{match(phenotypes(A), phenotypes(A))}$$

- **Feature 21:** ("PV equals EV") the average same-type overlap score of each candidate causative variant in *G*. The same-type overlap score of a candidate causative variant is defined to be 1.0 if the patient's variant overlaps an extracted variant in *A* of the same semantic effect, and 0.0 otherwise. Two variants have the same semantic effect if they are both missense, nonframeshift indel, frameshift, stopgain or stoploss variants.
- **Feature 22:** ("PV overlaps EV") the average overlap score of each candidate causative variant in *G*. The overlap score of a candidate causative variant is defined to be 1.0 if the patient's variant overlaps any extracted variant in *A*, and 0.0 otherwise.
- **Feature 23:** ("Paper PTV score*Patient PTV score") the PTV score of *A* (Feature 14) multiplied by the average protein-truncating score of all candidate causative variant in *G* (Feature 6).

- **Feature 24:** ("Paper NTV score*Patient NTV score") the NTV score of *A* (Feature 15) multiplied by the average non-truncating score of all candidate causative variant in *G* (defined as 1-Feature 6).
- **Feature 25:** ("Dominant score if PV het") the dominant score of *A* (Feature 16) if there is a single heterozygous candidate causative variant in *G*, otherwise 0.0.
- **Feature 26:** ("Recessive score if PV non-het") the recessive score of *A* (Feature 17) if there is more than a single heterozygous candidate causative variant in *G*, otherwise 0.0.
- **Feature 27:** ("Relevant gene score") the relevant gene score of *G* in *A*.

Most of these features were normalized by subtracting the mean feature value in the training set and dividing by the standard deviation of feature values in the training set, and subsequently clipped to be between -3.0 and +3.0. Features 9, 21, 22, which encode simple 0 or 1 flags indicating the presence or absence of a particular type of information were not normalized.

Creating simulated patients

To train AMELIE, we generated 681 simulated patients with Mendelian diseases. Each patient *s* was associated with phenotypes P_s and a list of candidate causative variants (including a known causative variant). Further, each patient *s* was associated with a unique article A_s linking the causative variant to the patient's phenotypes P_s. To generate simulated patients, we pursued the following high-level strategy (explained in detail below): first, we selected a set of OMIM diseases caused by a gene *GC*. Each combination of gene *GC* and disease *D* was associated with one article *AC* describing that mutations in *GC* cause *D.* Subsequently, we randomly selected a list of patient phenotypes *P* that are similar to phenotypes associated with *D.* Finally, we constructed a list of candidate causative variants by taking variants from a healthy individual and adding a single mutation in the gene *GC* that causes *D.* The simulated patient *s* was set to have disease $D_s=D$, causative gene $GC_s=GC$, phenotypes $P_s=P$ and an article $A_s=AC$ describing that mutations in GC_s cause D_s.

Selecting a set of OMIM diseases D, each caused by one gene GC, along with an article describing that mutations GC cause D

OMIM "Molecular Genetics" sections are sections in OMIM disease entries that textually describe molecular causes for a disease along with citations to original articles that first described the molecular causes for the disease. (E.g., the "Molecular Genetics section of the

OMIM entry on “Wilson Disease”: https://omim.org/entry/277900#molecularGenetics.) Molecular Genetics sections were downloaded from OMIM on February 15, 2017. To select a set of OMIM diseases, we first parsed the “Molecular Genetics” sections of the set of 4,948 OMIM diseases with a known molecular basis available up to July 2016 to find the first PubMed ID that did not occur in a subclause with the phrase “originally described” or “originally reported”. These “originally reported” subclauses usually refer to articles in which a cohort was originally described without identifying causative mutations. The first PubMed ID in the OMIM “Molecular Genetics” section that does not appear in a subclause with “originally described” or “originally reported” often links mutations in a gene to the disease. Thus, of those OMIM “Molecular Genetics” sections where we could parse a suitable PubMed ID, we took diseases where the first PubMed ID was published after 2011, resulting in a set of 1,363 OMIM diseases with causative genes and a PubMed ID that was published in 2011 or later.

Selecting a random set of phenotypes for a simulated patient with an OMIM disease D

Each of the OMIM diseases associated with an article was examined in order to create phenotypes for a simulated patient affected with the disease. To select a realistic set of simulated patient phenotypes for each disease, it is necessary to select some of the most important phenotypes of the disease with high probability while not selecting such a large set of phenotypes that diagnosing the patient using an automated method is trivial (e.g., by associating the patient with all phenotypes associated with the disease). OMIM “Clinical Features” sections give a textual description of clinical symptoms associated with a disease, often with the most striking and most-often occurring phenotypes mentioned first (e.g., the clinical features of Wilson Disease: https://omim.org/entry/277900#clinicalFeatures). OMIM “Clinical Features” sections were downloaded on February 15, 2017. To determine which the most striking and most-often occurring phenotypes were for a particular disease *D*, we first parsed the “Clinical Features” section of the corresponding OMIM entry for disease *D* using the AMELIE phenotype recognizer. The phenotypes recognized in the “Clinical Features” section were output in the order of first recognition. If no phenotypes were parsed from the “Clinical Features” section, the disease *D* was not used to construct a simulated patient. OMIM “Clinical Synopsis” sections list clinical features using HPO terms (e.g., the clinical synopsis of Wilson Disease: https://omim.org/clinicalSynopsis/277900). OMIM “Clinical Synopsis” sections were downloaded on February 17, 2017. We saved a separate set of HPO phenotypes associated with the disease *D* by OMIM in the “Clinical Synopsis” section that were not already recognized in the “Clinical Features” section. Intuitively, the features described in the free-text “Clinical Features” sections are often the most important phenotypes associated with the disease, while many of the phenotypes mentioned in the “Clinical Synopsis” section are less important and occur less frequently. Subsequently, we refer to the phenotypes parsed from the “Clinical Features” section as “key phenotypes”, and phenotypes listed in “Clinical Synopsis” as “other phenotypes”. To construct a random set of HPO phenotypes for a patient with a given OMIM disease *D*, we selected a subset of the set of key phenotypes parsed from the “Clinical Features” section and the other phenotypes from the “Clinical Synopsis” section of the OMIM entry for *D* using the following strategy: the first mentioned key phenotype in the “Clinical Features” section received a weight of 1.0. Subsequently mentioned key phenotypes received a weight of 0.6 times the weight of the previous key phenotype. The other phenotypes all received a weight of 0.25 times the lowest weighted key phenotype. The weights were converted to a phenotype probability mass function (hence called PMF_P) by dividing each weight by the sum of all weights assigned to phenotypes of the disease. Let $IPMF_P$ be defined as the probability mass function that is achieved by taking the reciprocal value of each entry in PMF_P and dividing by the total sum of weights in $IPMF_P$. Subsequently, a target fraction match score *t* was picked from a normal distribution with mean 20% and standard deviation 10%. If the target fraction match score was less than 5%, it was set to 5%. The goal of the random phenotype generator was now to select a set of simulated phenotypes *F* from the above probability mass function over HPO phenotypes for the given disease *D* such that the fraction match score over the set of all key and non-key phenotypes *O*, defined as

$$\frac{match(F,O)}{match(O,O)}$$

was as close to the target fraction match score as possible. 100 random draws of phenotypes were generated in the following manner: the number of phenotypes *n* to draw was determined by rounding a draw from a Gaussian distribution with mean 7 and standard deviation 4 to the nearest integer. A set *F* of *n* phenotypes were selected using PMF_P as probability mass function. While $\frac{match(F,O)}{match(O,O)}$ was smaller than the target fraction match score, previously unselected phenotypes were uniformly drawn and added to *F* until $\frac{match(F,O)}{match(O,O)}$ was greater than the target fraction match score. Subsequently, while $\frac{match(F,O)}{match(O,O)}$ was greater than the target fraction match score, a phenotype was selected using the probability mass function $IPMF_P$ and replaced by one of its parents in the HPO DAG until $\frac{match(F,O)}{match(O,O)}$ was less than the target fraction match score. Of the 100 sets of simulated phenotypes for disease *D*, the set *F* was selected which minimizes the function

$$J(F) = \left| t - \frac{match(F,O)}{match(O,O)} \right| + 0.001 * \min\left(3, \#abnormalities(F)\right),$$

where *#abnormalities(F)* is defined as the number of HPO phenotypes whose name starts with "abnormality of …" or "abnormal …". These phenotypes are far less likely to be selected by a human phenotype annotator than by an automated algorithm, and were therefore penalized by the automatic phenotype selection algorithm. If the set *F* out of the 100 picked sets with the lowest *J(F)* has $J(F) > 0.05$, the disease *D* was not used to construct a simulated patient.

Creating a set of candidate causative variants for a simulated patient with an OMIM disease D

To construct a genotype of a simulated patient that is causative of disease *D*, all candidate causative variants from a random 1000 genomes project participant were taken. Subsequently, a pathogenic (CLNSIG=5) ClinVar variant (from ClinVar file date 20160705, downloaded from ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/archive_1.0/2016/clinvar_20160705.vcf.gz) in the causative gene *GC* that was associated with disease *D* was randomly selected and added to the patient's genotype. If no such variant existed, the disease *D* was not used to construct a simulated patient. The zygosity of the variant was determined as follows: first, the inheritance mode of the disease *D* was parsed from the OMIM field "phenotypeInheritance" in the "geneMap" section. These data sections were downloaded from OMIM on May 17, 2017. If the disease was annotated as dominant, the variant was set to be heterozygous; if the disease was recessive, the variant was set to be homozygous. Diseases without a unique OMIM-assigned inheritance mode were skipped. If the inheritance mode could not be parsed from OMIM, the disease *D* was not used to construct a simulated patient.

Creating a training set for the AMELIE classifier using simulated patients

The simulated patients, each associated with a disease D_s, a causative gene GC_s, an article A_s describing that mutations in GC_s cause D_s, a list of phenotypes and a list of candidate causative variants, were used to construct a training set for the AMELIE classifier. The training set consists of a set of triples *(P, G, A)*, where *P* is a list of phenotypes, *G* is a candidate causative gene (i.e., a gene containing at least one of the patient's candidate causative variants), and *A* is an article about *G*. Each of these triples either receives a label "true" or "false" (to be used in the training set), or are unlabeled (and discarded from the training set):

- The triple (P_s, GC_s, A_s) containing both the causative gene GC_s and the article A_s describing that mutations in GC_s cause D_s was supervised true.
- All other triples including a patient's causative gene GC_s were discarded from the training set (i.e., neither supervised true nor false).
- All triples (P_s, G, A) containing a gene *G* not equal to the simulated patient *s*'s causative gene GC_s were supervised "false".

Due to the abundance of negative training examples, the total set of negatives was randomly subsampled to 1000x the size of the positive training set. Supervised triples *(P, G, A)* were converted into feature vectors of 27 features described above. A scikit-learn *(41)* 0.20.0 LogisticRegression classifier with options class weight="balanced" and default parameters otherwise was trained on the data created from the simulated patients to act as the AMELIE classifier. GNU parallel *(44)* was used to speed up parts of the computation.

Ranking candidate causative genes using the AMELIE classifier

AMELIE assigns each candidate gene *G* the score of its highest-scoring triple *(P, G, A)* using the following formula:

$$score(G) = max_{A\in\{articles\ about\ G\}}\ \text{AMELIE\_classifier}(P, G, A)$$

Evaluating AMELIE on a patient set from DDD

VCF files of patients submitted to the Deciphering Developmental Disorders *(33)* (DDD) project were downloaded from the European Genome-Phenome Archive *(40)* (EGA) study EGAS00001000775. Variant filtering to candidate causative variants was performed as described above in Methods section “Variant filtering to arrive at candidate causative variants and genes”. All patients with a single-gene diagnosis, with the causative variant present in their associated Variant Call Format (VCF) file and in the candidate causative gene list, that was not due to a structural variant and for which the causative gene was not a novel discovery of the DDD project were selected. From any diagnosed set of siblings, a single diagnosed sibling was selected at random, resulting in a final patient test set of 215 patients.

Ranking candidate causative genes using the AMELIE classifier

The final AMELIE score of a triple *(P, G, A)* is defined to be the output of the AMELIE classifier on *(P, G, A)* if the classifier output is less than 0.95. Otherwise, it is defined to be (0.95 + 0.05 * *match(patient_phenotypes, article_phenotypes)* / *match(patient_phenotypes, patient_phenotypes)*). This is because for genes with a very high AMELIE classifier score (>= 0.95), the raw phenotype match score (calculated as *match(patient_phenotypes, article_phenotypes)* / *match(patient_phenotypes, patient_phenotypes)*) is more indicative of a good match of article to patient than the raw output of the AMELIE classifier. Each candidate causative gene is assigned the AMELIE score of its highest-ranking article.

Comparisons of AMELIE to other methods

All comparisons of AMELIE gene ranking results to other automatic gene ranking algorithms’ ranking results were performed by running the automatic gene ranking algorithms described below, and **subsetting the output list of ranked genes to the same list of candidate genes that were also used by AMELIE** (Figure 2a). This ensures a fair comparison between AMELIE and all other automatic gene ranking methods, because all methods are ultimately measured for their ability to rank the same set of candidate genes. In case an automatic gene ranking method did not output the causative gene in its output list of ranked genes for a patient, the causative gene was given a rank equal to the number of ranked genes plus half the number of all unranked candidate causative genes *(11)*. Since causative genes for DDD, Stanford, and Manton patients are given as HGNC gene symbols, and all compared tools (including AMELIE) output ranked lists of HGNC gene symbols, we computed causative gene ranks using ranked lists of HGNC gene symbols. Tied scores in the output of any methods were broken by alphanumerically sorting HGNC gene symbols.

One-sided Wilcoxon signed-rank tests were performed using the R function

```
wilcox.test(ranks1, ranks2, paired=TRUE, alternative=”less”)
```

When training each of the classifiers that build the AMELIE knowledgebase, we made sure not to train on any article about the causative genes in any of the 215 DDD patients. Nor was any simulated patient assigned a causative gene from the list of DDD patients’ causative genes.

Comparison to Exomiser

The output of Exomiser *(14)* version 11.0.0 was obtained by running the Exomiser Command Line Interface (exomiser-cli) version 11.0.0, obtained from https://github.com/exomiser/Exomiser/releases/download/11.0.0/exomiser-cli-11.0.0-distribution.zip, on all patients. The following command line was used:

```
java -Xms2g -Xmx4g -jar /cluster/u/jbirgmei/Downloads/exomiser/exomiser-cli-11.0.0/exomiser-cli-11.0.0.jar --analysis patient.yml
```

The file patient.yml contained a link to the patient’s VCF file and the patient’s candidate gene list and was based on the file test-analysis-exome.yml in the Exomiser V11 distribution zip file. For each patient, each gene was associated with the “combinedScore” output for the gene by Exomiser. The Exomiser output contained the causative gene in 271 (100%) real patient test cases and 676 (99%) simulated patient cases. Genes were sorted by the Exomiser combinedScore (high-to-low) and subset to the **subset to the list of candidate genes used by AMELIE** in order to arrive at causative gene ranks.

Comparison to Phenolyzer

The output of Phenolyzer *(15)* was obtained by running the Phenolyzer executable from https://github.com/WGLab/phenolyzer, commit number 80596ac3affc565e178dcff3a308e408be0ab94f from September 11, 2018, using the command line

```
perl phenolyzer/disease_annotation.pl "<patient HPO IDs semicolon-separated>" -p -ph -logistic -out <output_file> -addon DB_DISGENET_GENE_DISEASE_SCORE,DB_GAD_GENE_DISEASE_SCORE -addon_weight 0.25
```

The Phenolyzer output contained the causative gene in 265 (98%) real patient cases and 655 (96%) simulated patient cases. **Genes in the output of Phenolyzer were subset to the same candidate gene list used by AMELIE** and all other methods and sorted by associated score (high-to-low) to arrive at causative gene rankings.

Comparison to eXtasy

The output of eXtasy *(17)* was obtained by running the eXtasy Command Line Interface cloned from GitHub (commit number 6fe9d418f05d198e0504f3d0327a4f4ebb7e3fbd from February 19, 2014) on all patients.

```
./extasy.rb -i <patient_vcf_filename> -g <encoded patient HPO IDs> -c
```

The eXtasy output contained the causative gene in 188 (69%) real patient cases and 368 (54%) simulated patient cases. A sample of cases where eXtasy would not rank the true causal gene were spot checked manually, to no better results. The output genes were **subset using the same list of candidate causative genes that was used to evaluated AMELIE** and all other methods and sorted by the eXtasy "extasy_combined_order_statistics" score (low-to-high) to arrive at causative gene ranks.

Comparison to Phen-Gen

The output of Phen-Gen *(16)* was obtained by running the Phen-Gen V1 executable downloaded from http://54.173.20.191/downloadexe.php?file=Phen-GenV1.tar.gz using the command line

```
perl phen-gen.pl input_phenotype=<patient phenotypes file, containing patient HPO IDs newline-separated> input_vcf=<patient VCF file> input_ped=<patient pedigree file>
```

The Phen-Gen output contained the causative gene in 67 (25%) real patient cases and 363 (53%) simulated patient cases. A sample of cases where Phen-Gen would not rank the true causal gene were spot checked manually, to no better results. Genes associated with scores output by Phen-Gen ("PROBABILITY_DAMAGING" column) were **subset to the same list of candidate causative genes used by AMELIE and all other compared methods here** and sorted by associated score (high-to-low) to arrive at causative gene ranks.

Comparison to PubCaseFinder

The output of PubCaseFinder *(18)* was obtained by querying the PubCaseFinder API (https://pubcasefinder.dbcls.jp/mme/match) on November 14, 2018 by querying the URL https://pubcasefinder.dbcls.jp/mme/match with content type application/vnd.ga4gh.matchmaker.v1.0+json based on the instructions given in https://github.com/ga4gh/mme-apis/blob/master/search-api.md. A list of HPO IDs and the list of candidate genes was provided to PubCaseFinder for each patient. Each gene in the output was associated with its maximum associated score in the output of PubCaseFinder. The PubCaseFinder API output contained the causative gene for 189 (70%) real test patients and 392 (58%) simulated patients. **PubCaseFinder output genes were subset to the same list of candidate causative genes used by AMELIE and all other methods compared here** to arrive at the causative gene rank.

Determining the number of genes to examine for 90% diagnosis rate across all gene ranking methods

For each automatic gene ranking method, we determined the minimum number *n* such that the top-ranked *n* genes would include the causative gene in at least 90% of the test patients.

Calculating the speedup achieved by each gene ranking method over a random baseline

For each automatic gene ranking method, the ranks of the causative genes were summed to arrive at the number of genes to investigate until arriving at the causative gene for all patients, assuming a clinician went through the prioritized gene lists in the returned order. The random baseline was estimated by assuming arrival at the causal gene after inspecting half the candidate gene list of every patient. The speedup was calculated by dividing the total size of the set of genes in the random baseline (summed across all patients) by the summed ranks of the causative genes.

Collection of patient data for Stanford and Manton Center cases

VCF files for 21 Manton patients were obtained from the Manton Center Gene Discovery Core. ClinPhen, a tool that automatically extracts phenotypes from medical records, was used to extract 3 top-prioritized HPO phenotypes each from anonymized patient medical records obtained from the Manton Center, as in *(19)*. Raw sequencing data in FASTQ format for 35

Stanford patients were obtained from the Medical Genetics Service at Stanford Children's Health. FASTQ files of Stanford patients were aligned to the GRCh37/hg19 human genome assembly using bwa-mem version 0.7.0-r789, with default parameters *(45)*, and variants were called using Genome Analysis Toolkit (GATK) version 3.4-46-gbc02625 following the HaplotypeCaller workflow in the GATK best practices *(46)* as previously described *(21)*. HPO phenotypes associated with these patients were manually created from the medical record and subsequently reviewed by a clinician. Variant filtering for rare, functional variants was performed in the same manner as for the DDD patients. For both subsets, all cases with available phenotypic and genotypic data for which the causative gene was in the candidate genes list were selected. Ranking of candidate causative genes after processing by the AMELIE classifier, calculating the speedup over a random baseline, determining the number of genes to investigate for a 90% diagnosis rate, and comparison against other gene ranking methods was performed as described for the DDD patients.

Cross-validation of AMELIE classifier on simulated patients

To test gene ranking performance on simulated patients, we split the set of simulated patients into 5 evenly sized chunks. In five round-robin iterations, we re-trained the AMELIE classifier using 4 of the 5 chunks of simulated patients, and evaluated the re-trained classifier's causative gene ranking performance on the remaining fifth chunk of simulated patients (5-fold cross-validation).

Genotypes and phenotypes for simulated patients were obtained as described for the training set of the AMELIE classifier. Calculating the speedup compared to a random baseline, causative gene rank cutoff for a 90% diagnosis rate, and comparison against other methods were performed as above for the DDD patients.

AMELIE ranked the causative gene at the top in 621 cases (91%) and in the top 10 in 672 cases (99%), replicating previous results indicating far higher gene ranking performance on simulated patient data compared to real-world data *(12, 13, 15, 16, 47, 48)*. Compared to a random baseline, AMELIE speeds up causative gene discovery by 34.8x on simulated patients. The next best method, Exomiser, speeds up causative gene discovery by 17.9x and other methods perform worse (figure S4, table S7).

Determining contribution of AMELIE classifier features to performance

Determining the most, 2nd-most, and 3rd-most highly weighted features of the AMELIE classifier across all real patients

To determine the top most influential features for each patient, we analyzed all patients with causative gene rank 1. Because the AMELIE classifier is implemented as a logistic regression classifier, each feature is associated with a global weight. The numerical feature value of each feature associated with the top-ranked article about the causative gene was multiplied by the AMELIE classifier weight of the feature, yielding the 3 highest weighted features for each patient with causative gene rank 1.

Feature ablation analysis

6 AMELIE classifiers with reduced feature sets were re-trained on the same training set used to train the fully featured AMELIE classifier. The features removed from each of the 6 feature-ablated classifiers are listed in table S13.

Augmenting phenotype recognition in articles does not lead to increased performance

HPO cross-links some of its phenotype entries to other databases containing phenotype names. We utilized 19,949 cross-links available in HPO by augmenting HPO phenotype phrases with synonyms from UMLS *(35)*, MeSH *(36)*, and SNOMED-CT *(37)*, three databases containing phenotype names. This augmentation of recognizable phenotype names increased the number of distinct recognized phenotypic abnormalities per article in the AMELIE knowledgebase from 9 to 22. However, on the main task of causative gene ranking, the augmented AMELIE knowledgebase did not perform better (169 instead of 175 patients had the causative gene ranked at the top).

Medical Subject Headings (MeSH) *(36)* data was downloaded from ftp://nlmpubs.nlm.nih.gov/online/mesh/MESH_FILES/asciimesh/c2018.bin and ftp://nlmpubs.nlm.nih.gov/online/mesh/MESH_FILES/asciimesh/d2018.bin on October 15, 2018. SNOMED-CT *(37)* data was downloaded from https://download.nlm.nih.gov/mlb/utsauth/USExt/SnomedCT_USEditionRF2_PRODUCTION_20180901T120000Z.zip on October 15, 2018. Unified Medical Language System (UMLS) *(35)* data was downloaded from https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html on October 15, 2018. Cross-links (XREF) in HPO phenotype entries were used to associated HPO phenotype IDs with IDs of phenotypes in MeSH, SNOMED-CT, and UMLS. UMLS IDs and phenotype phrases were obtained from files MRCONSO.RRF.a*.gz (columns 1 and 15). SNOMED IDs and phenotype phrases were obtained from file Full/Terminology/sct2_Description_Full-en_US1000124_20180901.txt (columns 5 and 8). MeSH IDs and phenotype phrases were parsed from files c2018.bin and d2018.bin, fields "UI" (ID), "NM" (name) and "SY" (synonyms). All UMLS, MeSH and SNOMED CT phenotype phrases were processed in the same fashion HPO phenotype names and synonyms were processed. After augmenting HPO synonyms with synonyms stored in UMLS, MeSH and SNOMED CT, AMELIE could recognize the 13,439 HPO phenotypic abnormalities using 343,705 phenotype names and synonyms. AMELIE phenotype recognition was subsequently run on all downloaded full-text articles, and the AMELIE knowledgebase augmented with newly identified phenotype mentions. The AMELIE classifier was then re-trained on the augmented database, and the augmented database and modified classifier were used to rank candidate causative genes for all real test patients.

AMELIE knowledgebase constructed from full text is up to 43% superior to title/abstract-only knowledgebase on clinical patients

To quantify the information gained from full text for the purpose of causative gene ranking compared to title/abstract-only data, we re-trained all AMELIE knowledgebase classifiers only on title/abstract data from PubMed. Subsequently, we constructed the AMELIE knowledgebase using title/abstract data from PubMed only and re-trained the AMELIE classifier using the title/abstract-only knowledgebase.

To construct the AMELIE knowledgebase using only title/abstract data from PubMed, we fed the title and abstract of each article in place of the full text to all components involved in constructing the AMELIE knowledgebase. Classifiers involved in construction of the AMELIE knowledgebase (relevance classifiers, relevant gene classifier, dominant/recessive classifier, variant type classifier) were re-trained on this data. PubMunch download of full-text articles was disabled. AVADA full text variant data was omitted. The AMELIE classifier was re-trained on the title/abstract-only knowledgebase using data derived from the 681 simulated patients described above.

Title/abstract-based gene ranking performs significantly worse than ranking based on full text information across all 271 real test patients ($p=1.87*10^{-3}$; one-sided Wilcoxon signed rank test). On the set of clinical patients from Stanford and Manton, 43% more patients had their causative genes ranked at the top using full-text based ranking compared to title/abstract-based ranking. Overall, title/abstract-based ranking put the causative genes of 133 of 271 real patients at the top, compared with 175 (+32%) of 271 real patients for full-text based AMELIE.

Comparing causative gene rank vs. number of articles analyzed by AMELIE

Linear regression and the Wald test for the slope of linear regression between AMELIE causative gene ranks and number of articles analyzed by AMELIE for the causative gene for each patient were performed using the Python 3.7.0 package scipy version 1.1.0 using the method scipy.stats.linregress, which returns the Wald test p-value.

Comparisons of AMELIE knowledgebase to DisGeNET and AMELIE classifier to Phrank

Constructing a DisGeNET version of the AMELIE knowledgebase

Multiple previous methods for text mining gene, disease, variant, and phenotype information from literature have been developed *(49–58)*, most of which curate limited information or information from comparatively small sets of articles. Of 3 additional efforts *(38, 59, 60)* attempting to curate gene-phenotype information from a broad set of articles in automatic or semi-automatic fashion, we focused on DisGeNET (the most recently updated such database), containing gene-phenotype relationships, disease-causing variants, and links to primary literature from PubMed. DisGeNET contains both automatically curated data and hand-curated data *(38)*.

DisGeNET data were downloaded from http://www.disgenet.org/ds/DisGeNET/results/ on October 5, 2018. Files containing all gene-disease-PubMed ID associations, all variant-disease-PubMed ID associations, BeFree gene-disease-PubMed ID associations, BeFree variant-disease-PubMed ID associations, mappings from DisGeNET gene identifiers to Ensembl gene identifiers, and mappings from DisGeNET disease identifiers to HPO IDs were downloaded. For BeFree, manually curated data, and both data sources in DisGeNET, we used the following procedure to fill the AMELIE knowledgebase: we converted all DisGeNET disease identifiers to HPO version releases/2018-07-25 IDs where possible according to DisGeNET-provided mappings, all DisGeNET gene identifiers to Ensembl version 84 gene identifiers where possible according to DisGeNET-provided mappings, and then added all articles associated with at least one Ensembl ID and one mentioned HPO ID to the DisGeNET version of the AMELIE knowledgebase. Subsequently, we converted all variants (in form of dbSNP rsIDs) to genomic coordinates, and added those variants that were associated with an article to the variants table in the DisGeNET version of the AMELIE knowledgebase.

Three different versions of DisGeNET information were used: **(1)** containing only data curated by DisGeNET's BeFree automatic curation software, **(2)** containing only manually curated data in DisGeNET, **(3)** containing all suitable data in DisGeNET, both manually and automatically curated. Subsetting DisGeNET to only data curated by DisGeNET's BeFree automatic curation software resulted in a total of 256,902 articles associated with a median of 1 Ensembl gene per article, and a median of 1 genetic variant in 16,292 articles. Subsetting DisGeNET to only manually curated data resulted in a total of 59,472 articles, associated with a median of 1 Ensembl gene per article, and no genetic variants associated with articles. Taking all suitable data in DisGeNET resulted in 287,428 articles.

3 AMELIE classifiers were subsequently trained on data derived from the 681 simulated patients using data in the 3 versions of the DisGeNET knowledgebase analogous to the strategy described above for the full-text AMELIE knowledgebase. Causative gene ranks for the 271 real test patients were obtained using the 3 versions of the DisGeNET knowledgebase analogous to how they were obtained using data in the AMELIE knowledgebase. All 3 versions of DisGeNET performed worse than AMELIE (all p-values $\leq 4.76*10^{-23}$, one-sided Wilcoxon signed rank test). Best-performing was the DisGeNET database with automatically curated data (68 causative genes ranked at the top, compared to 175 for original AMELIE; table S10).

Using Phrank to rank genes using information in the AMELIE knowledgebase

We replaced the AMELIE classifier by the Phrank *(11)* phenotypic match score to rank patient candidate genes. The Phrank phenotypic match score *(11)* was initialized with a database of gene-phenotype associations derived from the AMELIE knowledgebase as above for the

Phrank phenotypic match score features in the AMELIE classifier. Given a patient associated with a list of HPO phenotypes and a list of candidate causative genes, we calculated the Phrank phenotypic match score between the patient and each article about one of the patient's candidate causative genes. Each candidate gene was then associated with: (1) the highest Phrank phenotypic match score of any article about the candidate gene, and (2) the AMELIE full-text document relevance score of the highest-ranked article about the candidate gene. The candidate genes were sorted (high-to-low) first by their associated phenotypic match score, and (to break ties), by their associated full-text relevance score (also high-to-low). This procedure was repeated for each of the 271 real test patients.

Supplementary Figures

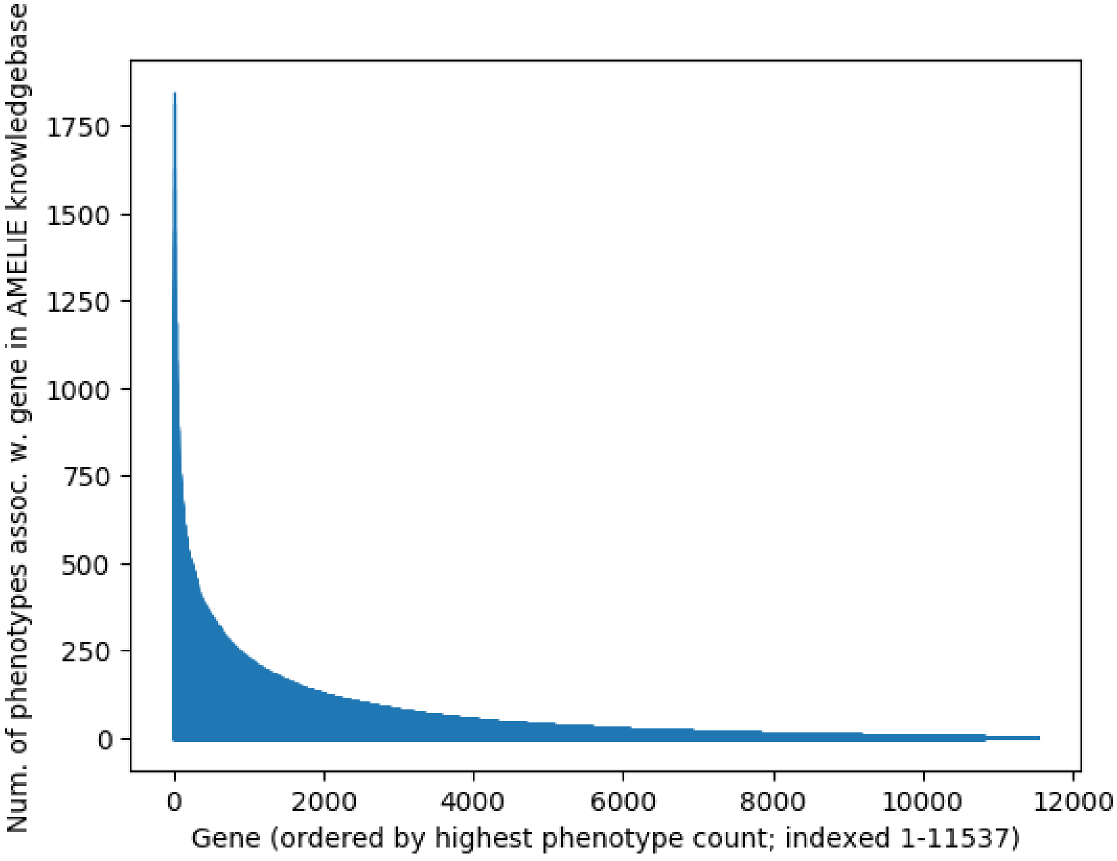


Figure S1. Number of phenotypes associated with genes through articles in the AMELIE knowledgebase. Genes are ordered by number of associated phenotypes (descending left-to-right). The number of phenotypes associated with each gene is plotted.

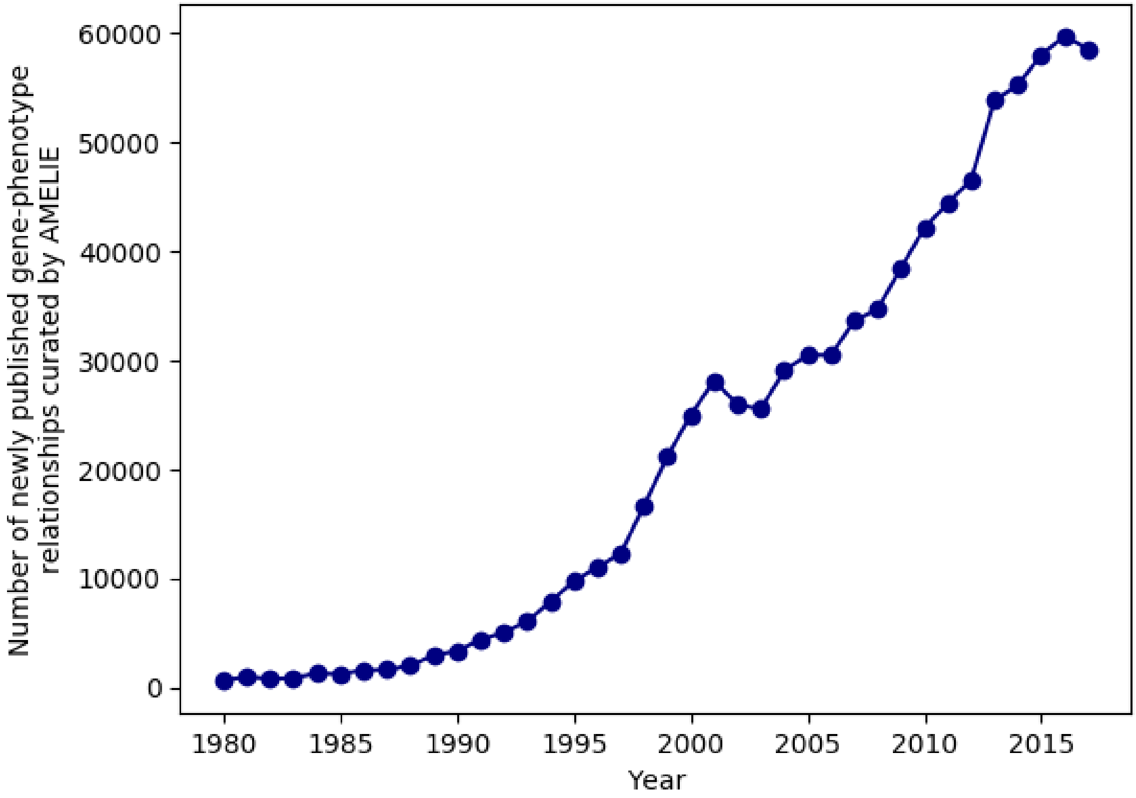


Figure S2. The accelerated accumulation of curated facts in Mendelian genomics. The number of newly published gene-phenotype relationships discovered by AMELIE every year.

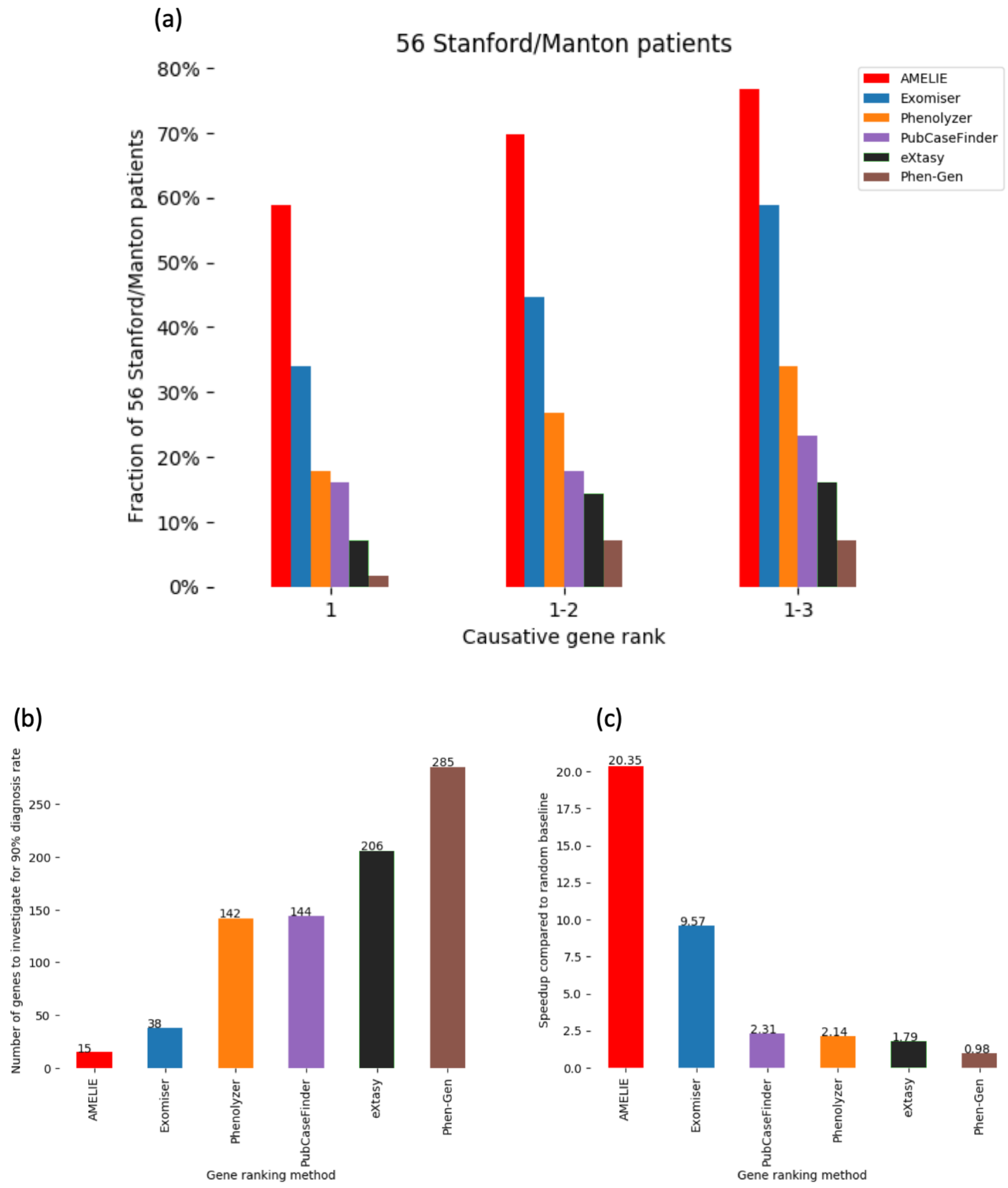


Figure S3. Replication of AMELIE's causative gene ranking performance on 56 clinical patients from Stanford and Manton. We evaluated AMELIE against the other gene ranking tools on a set of 56 patients from Stanford Children's Health and the Manton Center for Orphan

Disease Research. **(a)** Number of genes ranked at the top, ranked 1-2, and ranked 1-3, by different candidate causative gene ranking methods. **(b)** Number of top-ranked genes to investigate for a 90% diagnosis rate by different methods. **(c)** Speedup of diagnosis when evaluating ranked genes top to bottom over a random candidate causative gene search strategy.

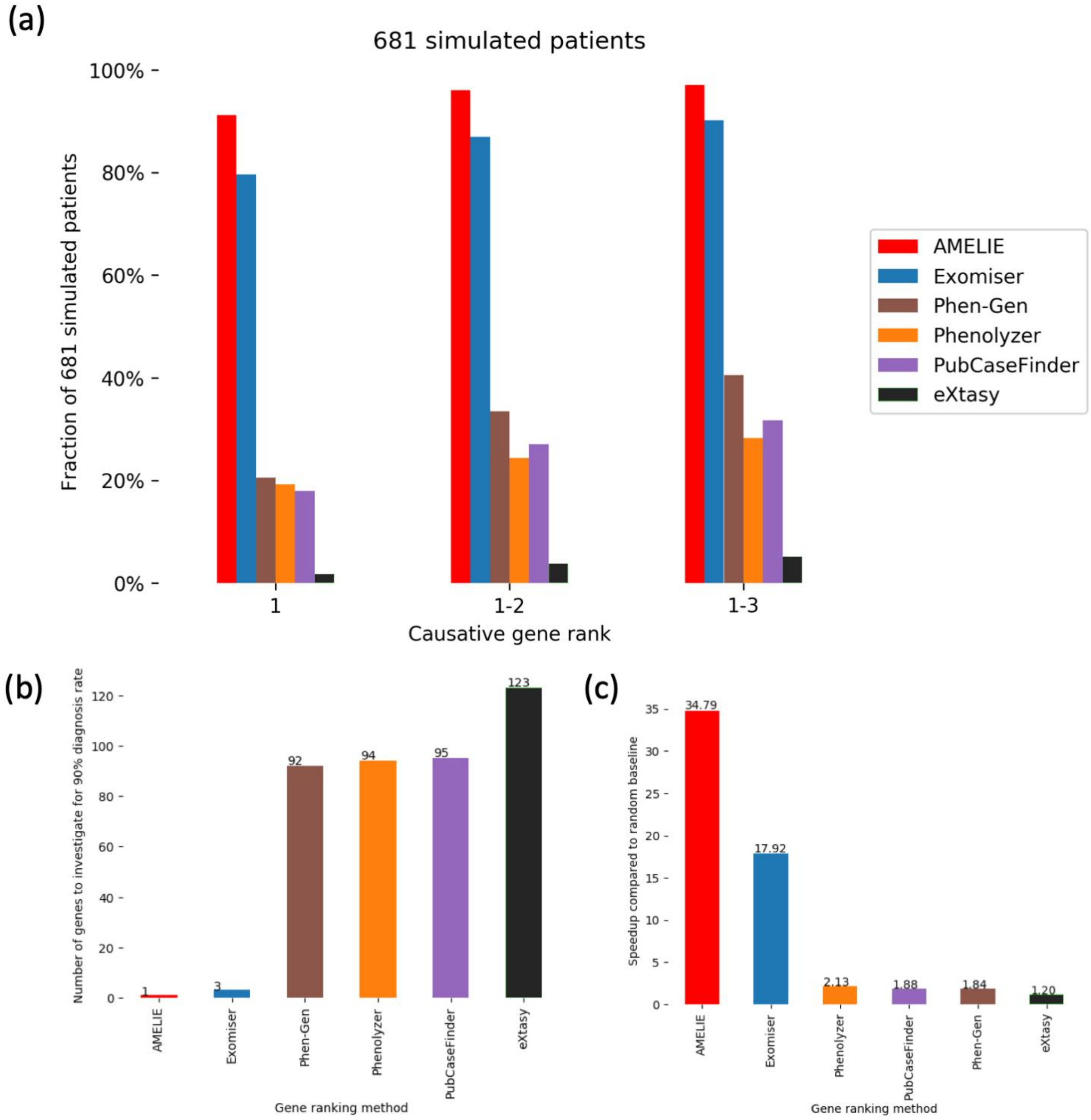
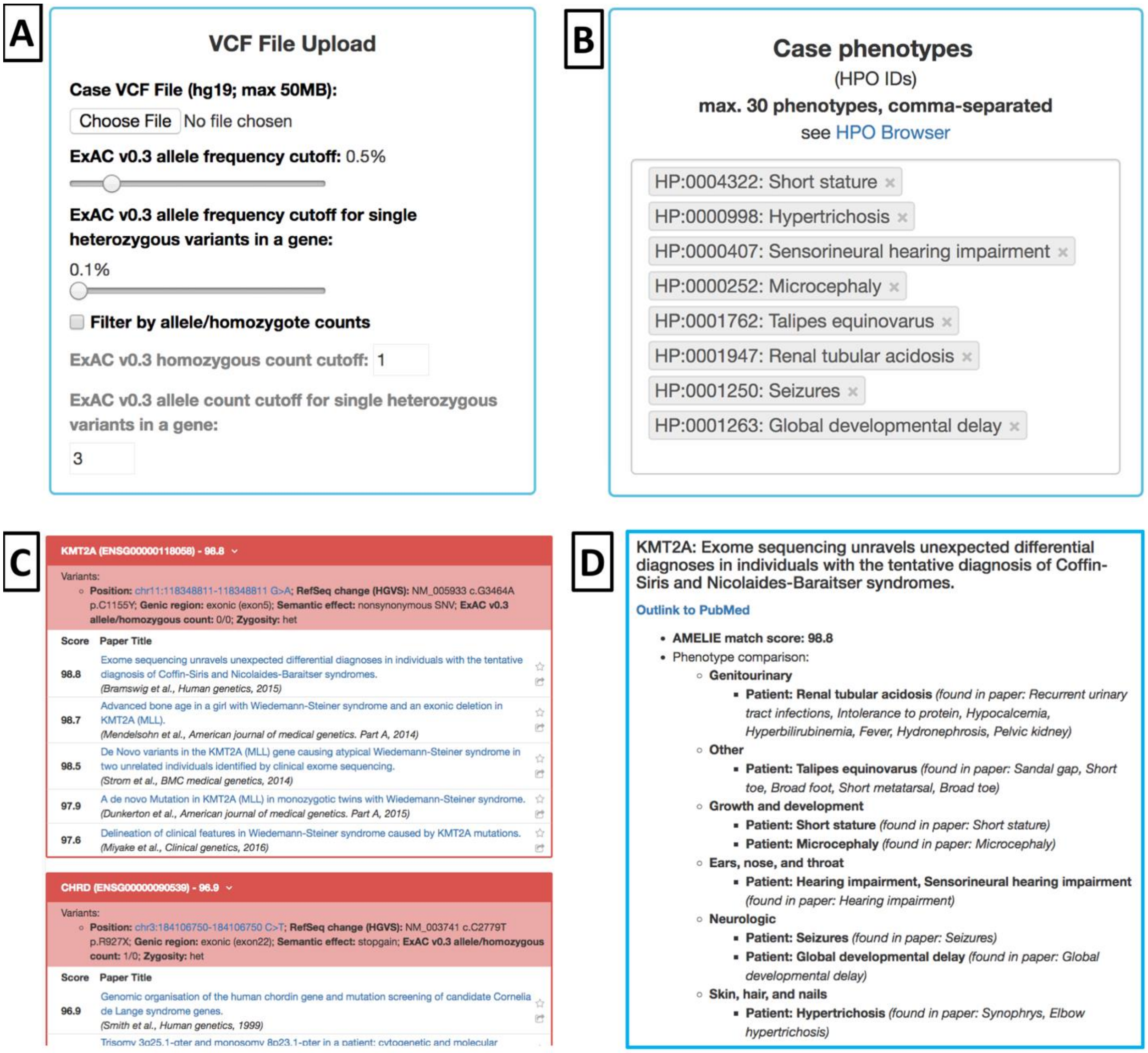


Figure S4. Cross-validation of AMELIE's causative gene ranking performance on 681 simulated patients. We evaluated the AMELIE causative gene ranking performance on 681 simulated patients using 5-fold cross-validation (where 80% of the simulated patients were used for training and 20% were evaluated, in 5 rounds). As in Figure S3, the panels show **(a)** raw causative gene ranking performance, **(b)** number of genes to investigate for a 90% diagnosis rate, and **(c)** speedup of causative gene discovery.



VCF File Upload

Case VCF File (hg19; max 50MB):

Choose File No file chosen

ExAC v0.3 allele frequency cutoff: 0.5%

ExAC v0.3 allele frequency cutoff for single heterozygous variants in a gene:

0.1%

☐ **Filter by allele/homozygote counts**

ExAC v0.3 homozygous count cutoff: 1

ExAC v0.3 allele count cutoff for single heterozygous variants in a gene:

3

Case phenotypes

(HPO IDs)

max. 30 phenotypes, comma-separated

see HPO Browser

- HP:0004322: Short stature ×
- HP:0000998: Hypertrichosis ×
- HP:0000407: Sensorineural hearing impairment ×
- HP:0000252: Microcephaly ×
- HP:0001762: Talipes equinovarus ×
- HP:0001947: Renal tubular acidosis ×
- HP:0001250: Seizures ×
- HP:0001263: Global developmental delay ×

KMT2A (ENSG00000118058) - 98.8

Variants:

- **Position:** chr11:118348811-118348811 G>A; **RefSeq change (HGVS):** NM_005933 c.G3464A p.C1155Y; **Genic region:** exonic (exon5); **Semantic effect:** nonsynonymous SNV; **ExAC v0.3 allele/homozygous count:** 0/0; **Zygosity:** het

Score	Paper Title
98.8	Exome sequencing unravels unexpected differential diagnoses in individuals with the tentative diagnosis of Coffin-Siris and Nicolaides-Baraitser syndromes. *(Bramswig et al., Human genetics, 2015)*
98.7	Advanced bone age in a girl with Wiedemann-Steiner syndrome and an exonic deletion in KMT2A (MLL). *(Mendelsohn et al., American journal of medical genetics. Part A, 2014)*
98.5	De Novo variants in the KMT2A (MLL) gene causing atypical Wiedemann-Steiner syndrome in two unrelated individuals identified by clinical exome sequencing. *(Strom et al., BMC medical genetics, 2014)*
97.9	A de novo Mutation in KMT2A (MLL) in monozygotic twins with Wiedemann-Steiner syndrome. *(Dunkerton et al., American journal of medical genetics. Part A, 2015)*
97.6	Delineation of clinical features in Wiedemann-Steiner syndrome caused by KMT2A mutations. *(Miyake et al., Clinical genetics, 2016)*

CHRD (ENSG00000090539) - 96.9

Variants:

- **Position:** chr3:184106750-184106750 C>T; **RefSeq change (HGVS):** NM_003741 c.C2779T p.R927X; **Genic region:** exonic (exon22); **Semantic effect:** stopgain; **ExAC v0.3 allele/homozygous count:** 1/0; **Zygosity:** het

Score	Paper Title
96.9	Genomic organisation of the human chordin gene and mutation screening of candidate Cornelia de Lange syndrome genes. *(Smith et al., Human genetics, 1999)*
	Trisomy 3q25.1-qter and monosomy 8p23.1-pter in a patient: cytogenetic and molecular

KMT2A: Exome sequencing unravels unexpected differential diagnoses in individuals with the tentative diagnosis of Coffin-Siris and Nicolaides-Baraitser syndromes.

Outlink to PubMed

- **AMELIE match score: 98.8**
- Phenotype comparison:
  - **Genitourinary**
    - **Patient: Renal tubular acidosis** *(found in paper: Recurrent urinary tract infections, Intolerance to protein, Hypocalcemia, Hyperbilirubinemia, Fever, Hydronephrosis, Pelvic kidney)*
  - **Other**
    - **Patient: Talipes equinovarus** *(found in paper: Sandal gap, Short toe, Broad foot, Short metatarsal, Broad toe)*
  - **Growth and development**
    - **Patient: Short stature** *(found in paper: Short stature)*
    - **Patient: Microcephaly** *(found in paper: Microcephaly)*
  - **Ears, nose, and throat**
    - **Patient: Hearing impairment, Sensorineural hearing impairment** *(found in paper: Hearing impairment)*
  - **Neurologic**
    - **Patient: Seizures** *(found in paper: Seizures)*
    - **Patient: Global developmental delay** *(found in paper: Global developmental delay)*
  - **Skin, hair, and nails**
    - **Patient: Hypertrichosis** *(found in paper: Synophrys, Elbow hypertrichosis)*

Figure S5. Essence of the interface at https://AMELIE.stanford.edu**.** Inputs: **(A)** Clinicians upload a VCF file (variant list) of an affected patient. Using a number of adjustable parameters, variant frequency thresholds are set to define a subset of candidate causative variant. Optionally, one can upload just a list of candidate genes. **(B)** Clinicians describe the patient using Human Phenotype Ontology terms. Outputs: **(C)** AMELIE outputs a list of ranked candidate genes along with literature that support this gene as being causative of the patient’s disease. **(D)** Upon selecting an article from Panel C, AMELIE displays a comparison of patient phenotypes with phenotypes detected in the full text of the article.

Supplementary Tables
(See attached Excel file for table contents.)

Table S1. Full-text gene extraction statistics. Mean, median, standard deviation, and interquartile range of number of candidate genes (including many false positives) discovered in all downloaded full text articles, vs. the number of genes with relevance score >= 0.1 per article that ultimately made their way to the AMELIE knowledgebase.

Table S2. Extraction statistics from the 100 most-used journals. Number of gene-phenotype relationships and number of downloaded articles from top 100 journals with most extracted gene-phenotype relationships.

Table S3. Simulated patient details. Containing ID of 1000 genomes person used as basis for genotype, causative gene, assigned OMIM disease, list of synthetically generated phenotypic abnormalities for the disease (canonicalized up to "phenotypic abnormality", i.e., adding all ancestors of the assigned phenotype in the HPO directed acyclic graph of phenotypes up to the node "phenotypic abnormality"), number of phenotypic abnormalities assigned to the patient (canonicalized), number of phenotypic abnormalities associated with the OMIM disease provided by HPO build 149 (downloaded on September 27th, 2018, from http://compbio.charite.de/jenkins/job/hpo.annotations.monthly/149/artifact/annotation/ALL_SOURCES_ALL_FREQUENCIES_diseases_to_genes_to_phenotypes.txt) (canonicalized), fraction of overlapping patient and HPO phenotypes over HPO phenotypes.

Table S4. DDD patient details. DDD patients, with number of candidate causative genes, causative gene, causative gene ranks across all compared gene ranking methods, and list of patient phenotypes.

Table S5. Searching for top-ranked AMELIE articles in OMIM. Top-ranked articles about the causative gene by AMELIE, indicating whether the article is cited in OMIM or not.

Table S6. Stanford and Manton clinical patient details. Stanford and Manton patients with number of candidate causative genes, causative gene, and gene ranks across all compared gene ranking methods.

Table S7. Simulated patient gene ranking results. Simulated patients, number of candidate causative genes, assigned causative gene, and gene ranks across all compared gene ranking methods.

Table S8. Most important features for patients with top-ranked causative genes. Real patients where AMELIE ranked the causative gene at the top, along with the top 3 highest most contributing features to the top-ranked article about the causative gene.

Table S9. AMELIE classifier feature ablation results. Real patients with AMELIE causative gene ranks after removing features from the AMELIE classifier.

Table S10. DisGeNET gene ranking results. AMELIE gene ranks of real patients when using the DisGeNET knowledgebase instead of the AMELIE knowledgebase.

Table S11. Regular expression patterns used to parse variant type from OMIM Allelic Variant entries.

Table S12. Phenotypes extracted from full-text articles by AMELIE, indicating whether the phenotype was extracted correctly or not. Phenotypes are counted as correctly extracted if they are mentioned as applying to humans, not as parts of word groups. E.g., the words "liver failure" are correctly extracted to HP:0001399 (Hepatic failure) in PubMed ID 15703195, but the word "shock" is incorrectly extracted to HP:0031273 (Shock) in PubMed ID 2702668 because it only occurs as part of "heat shock protein" in the article.

Table S13. Assignment of features to feature groups.