A Machine-Curated Database of Genome-Wide Association Studies

Volodymyr Kuleshov[1,2,*], Jialin Ding[1], Christopher Vo[1], Braden Hancock[1], Alexander Ratner[1], Yang Li[3], Christopher Ré[1], Serafim Batzoglou[1], Michael Snyder[2]

Affiliations

[1]Department of Computer Science, Stanford University, Stanford, CA

[2]Department of Genetics, Stanford University School of Medicine, Stanford, CA

[3]Department of Medicine, University of Chicago, Chicago, IL

Correspondence should be addressed to Volodymyr Kuleshov (kuleshov@cs.stanford.edu)

Abstract

Tens of thousands of genotype-phenotype associations have been discovered to date, yet not all of them are easily accessible to scientists. Here, we describe GwasKB, a novel machine reading system that automatically collects and synthesizes genetic associations from the scientific literature into a structured database. GwasKB helps curators by automatically collecting >3,000 previously documented open-access relations (with an estimated recall of 60-80%) as well as >2,000 associations not present in existing human-curated repositories (with an estimated precision of 82-89%). Our system represents the largest fully automated GWAS curation effort, and is made possible by a novel paradigm for constructing machine learning systems called data programming. Our results demonstrate both the importance and the feasibility of automating the curation of scientific literature.

Genome-wide association studies (GWAS) are widely used for measuring the effects of genomic mutations on human traits[1]. Despite revealing tens of thousands of genotype-phenotype associations, not all GWAS results are available to scientists in a structured form amenable to downstream analyses.

Multiple efforts are underway to catalogue published GWAS associations[2,3], but it is as yet unclear how far we are from a complete GWAS catalogue. Currently, even the most exhaustive databases vary in their scope: hundreds to thousands of variants may be present in one repository, but absent in all others[2,3]. Variants that are omitted in a database are effectively lost for downstream analyses, and as more studies are published, the number of these “dark variants" is expected to increase. This limits the pace of scientific research and represents an inefficient use of research funding.

Here, we describe GwasKB, a machine reading system that automatically collects and synthesizes thousands of genotype-phenotype associations into a structured database. Our system represents the largest GWAS machine curation effort, and is made possible by a novel paradigm for constructing machine learning systems called data programming. When deployed on a set of 589 open-access GWAS publications, GwasKB recovers (at an estimated recall of 60-80%, depending on stringency criteria) >3,000 known associations that were validated in existing GWAS databases, and finds >2,000 associations (with an estimated precision of 82-89%) currently absent in existing repositories. The number of these new variants corresponds to about 20% of all open-access associations recorded in the most up-to-date human-curated database, GWAS Catalog.

We make available to curators an open-source implementation of GwasKB and we also provide an online tool for browsing the associations found by our system[1]. We anticipate that these associations will be used by scientists in future work. More generally, we demonstrate that modern machine reading algorithms have matured to the point of

[1] An online interface to our machine-curated database is available at http://gwaskb.stanford.edu/

significantly improving biomedical curation efforts. Finally, our system may form the basis for further efforts to curate Mendelian mutations and other data.

Automating Biomedical Literature Curation with GwasKB

The results of genome-wide association studies are used to estimate disease risks[4,5], to understand the function of specific genomic regions[6,7], and to train predictors for the effects of new mutations[8]. Overall, about 2,500-3,000 studies have been performed to date[2,3]; they have reported tens of thousands of associations that are manually collected in databases like GWAS Catalog[2] and GWAS Central[3].

However, curating the results of GWAS studies is challenging, as it requires time, domain expertise, and can be prone to errors. As a result, independent human curation efforts are often not consistent, and even the largest GWAS databases are incomplete.

The GwasKB Machine Reading System

We propose that the process of collecting and synthesizing the findings of GWAS studies can be made significantly more efficient using automated machine reading technologies. We demonstrate this by introducing GwasKB, an automated system that extracts genotype-phenotype relations from the biomedical literature and places them in a structured SQL database (Figure 1).

Specifically, GwasKB collects three main pieces of information: genetic variants (as defined by their RSID), their associated phenotypes, and their p-values. We support our findings with evidence from publications (identified by their Pubmed ID), which can take the form of a sentence excerpt or a location in a table.

Several challenges arise when curating GWAS studies. For one, there is no universally adopted threshold for the significance of genotype-phenotype associations. GwasKB reports all (rsid, phenotype) associations that are significant at $p < 10^{-5}$ in at least one

experiment in the study (such as in one cohort or one statistical model) and it records all the other p-values relevant to that association. Our threshold of $p < 10^{-5}$ is the same as the one used in the GWAS Catalog.

A second difficulty arises when describing the study phenotype. Phenotypes can be very general (e.g., "heart disease") or highly specific (e.g., "high systolic blood pressure"), and existing databases often differ in their level detail. GwasKB addresses this issue by providing simple and detailed phenotypes, i.e. a high-level description that applies to every variant in the paper (e.g. "effects of proteins on inflammation"), and, when available, a detailed description for specific variants (e.g., the name of a specific protein).

Lastly, a third difficulty is posed by copyright restrictions. With GwasKB, we restrict ourselves to open-access papers, which represent approximately 25% of all the studies that have been published to date. All open-access publications are catalogued by the PubMed Central (PMC) repository and are made publicly available in XML format. GwasKB takes these XML documents as input, although any paper in HTML format may be parsed by our system after minor preprocessing. In the current version, we also discard any associated files that need to be processed through proprietary software. However, the principles of our system extend to all kinds of studies.

On The Design of GwasKB

GwasKB was designed to extract three key pieces of information: genetic variants, their phenotypes, and their p-values. We have structured GwasKB into a set of five components that extract this information.

The first component of GwasKB parses the title and abstract of every paper to identify a simple phenotype that will be associated with all its variants. The second component parses the body of the paper to find tuples of RSIDs and their associated detailed phenotypes. Often, the detailed phenotype is abbreviated (e.g. BMI) and a third component attempts to resolve these abbreviations (e.g. output "body mass index"). A

fourth component extracts p-values in the form of (rsid, p-value) tuples. Finally, the fifth component constructs a single structured database from all these results.

Each GwasKB component has three stages: parsing, candidate generation, and classification (Figure 2). Parsing is performed with Snorkel[9], a knowledge base construction framework for documents with richly formatted data (data expressed via textual, structural, tabular, and/or visual cues), such as XML documents. Content is first parsed for structure---the XML tree is traversed and converted into a hierarchical data model with text assigned to tables, cells, paragraphs, sentences, etc. Then each sentence or cell is parsed for content using the Stanford CoreNLP pipeline[10], which performs sentence tokenization, part-of-speech tagging, and syntactic parsing. In candidate generation, we identify in the text mentions of some target relation (e.g., p-value/rsid pairs). This is done by generating a large set of substrings from the text of the paper, some of which could contain our target relation. Regular expressions or dictionaries are used to identify candidates that may be valid instances of the relation we are looking for (erring on the side of high recall over high precision). Finally, in the classification stage, we determine which of these candidates are actually correct relation mentions using a machine learning classifier. We use a Naive Bayes classifier with a small number of hand-crafted features (between 4 and 12) and we train the model using the recently proposed data-programming paradigm[11]

One of the most significant bottlenecks in developing machine learning-based applications today is the challenge of collecting large sets of hand-labeled training data. Data programming is a newly proposed paradigm for training models using higher-level, less precise supervision to avoid this bottleneck. In this approach, users write a set of *labeling functions*: black-box functions that label data points, and that can subsume a wide variety of heuristic approaches such as *distant supervision*[12]—where an external knowledge base is used to label data points—regular expression patterns, heuristic rules, and more. These labeling functions are assumed to be better than random, but otherwise may have arbitrary accuracies, may overlap, and may conflict. A generative model is used to learn their accuracies and correlations from unlabeled data. The predictions of

this model can then be used for classification, or to generate labels for a second, discriminative model. We refer the reader to the appendix and to the full data programming paper[11] for more details.

Reproducibility

In order to make our results fully reproducible, we have released Jupyter notebooks that can be used to run GwasKB, generate the database of associations and recreate most of our figures and tables. The notebooks and the source code of GwasKB are freely available on GitHub at github.com/kuleshov/gwasdb.

In addition, we have built an interactive website that enables users to browse associations that have been extracted by GwasKB. Users can search the data by study, phenotype or variant rsid. The entire dataset can also easily be exported in text or SQL format.

Machine Reading Helps Automate GWAS Curation

We next demonstrate how our system can significantly help humans synthesize and understand findings from the biomedical literature. We deploy GwasKB on all the open-access papers listed in the GWAS Catalog database (589 in total), which is the most complete set of such papers that we could access. For evaluation, we also use the GWAS Central database. We use $p < 10^{-5}$ in at least one cohort or study methodology as our significance cutoff, and assess both the precision and the recall of our system (see Table 1).

GwasKB Recovers Up To 80% of Curated Open-Access Associations

GWAS Central and GWAS Catalog contain respectively 3008 and 4023 accessible associations in our set of 589 studies. These are variants whose RSID is contained in the open-access XML content made available through PubMed Central. We also define mappings between GwasKB phenotypes and phenotypes from GWAS Central and

GWAS Catalog (see Methods). These databases often use different levels of precision to describe phenotypes (e.g. “smoking behaviors” vs. “cigarette packs per day”); therefore, we also specify whether our reported phenotype is exact or approximate; in the latter case, it is still useful, but lacks some detail. Table 2 contains examples of relations extracted by GwasKB.

Among the set of open-access papers, GwasKB recovered 2487 (82%) relations with approximately correct phenotypes from GWAS Central and 3245 (81%) relations from the GWAS Catalog. It also recovered 1890 (63%) relations with full accuracy from GWAS Central and 2762 (69%) relations from GWAS Catalog. A number of known associations were not correctly recovered because their reported phenotype was incorrect (89 in GWAS Central and 147 in GWAS Catalog). In the remaining cases, we were not able to report the variant itself. Overall, GwasKB recovered 81-82% of accessible associations at a level of quality that will be useful in many applications.

Machine Curation Uncovers Many Associations Not Found by Human Curators

In total, GwasKB discovered 6422 relations within the 589 input papers, 2959 (46%) of which could not be mapped to GWAS Catalog or GWAS Central. Notably, many of these appeared to be valid.

We investigated this further by first manually inspecting a random subset of 100 novel relations (with independent validation from two independent annotators). We found that 82 relations fully met the specifications of our system, 11 were incorrect, and 7 were originally identified by a different study (and referenced as background material). Most of the errors can be attributed to incorrect phenotypes. Of the 82 relations matching system specifications, 60 appeared to satisfy the same criteria as GWAS Central or GWAS Catalog relations from the same paper, while 22 were not significant at 10^{-5} in all cohorts. The latter may have been omitted by human curators for this reason.

Novel Variants Found by GwasKB Are Correlated with Genomic Function

Linkage Disequilibrium Between Variants from GwasKB and from Existing Databases

To validate the novel variants found by our system, we conducted a series of analyses aimed at characterizing the variants' function. First, we reasoned that detected variants may be in linkage disequilibrium (LD) with known variants (because they originate from the same LD block), or among themselves, thereby inflating our number of truly novel associations.

We estimated LD from the Thousand Genomes dataset (Supplementary Methods); Figure 3 shows the histogram of r^2 distances between each novel variant, and its closest variant in the GWAS Catalog. The distribution of r^2 scores is highly multimodal, with large peaks at $r^2=1$, and many more at $r^2=0$.

Using a threshold of $r^2 > 0.5$, we filtered our set of new [pmid, rsid, phen, pvalue] associations from 3170 to 1494 by removing variants in LD with known manually curated variants; of the 1676 variants that we eliminated, 765 were not in the 1000 Genomes database or their closest previously known variant was not in the database; the remaining 911 SNPs were in LD with known variants. We further reduced this set to 1304 associations by eliminating novel variants that were in LD with each other. Thus, although many variants are in LD with known variants, over 40% of our discovered variants do not appear to be linked to variants previously identified in GWAS databases.

We argue that it is preferable to curate both novel and known variants, since we do not know which mutation in an LD block is truly causal and the r^2 cutoff for defining LD blocks is somewhat arbitrary and may vary. We think that filtering should be performed by the user, depending on their goal; this is also the approach taken by the GWAS Central repository. Moreover, if the authors of a GWAS study report multiple variants in LD, we believe that it is better to report their findings as they are, rather than introducing additional bias through our own filtering.

Comparison to Alternative Approaches for Estimating Variant Significance

Our second analysis focuses on the biological function of the novel variants. We focus on two large classes of phenotypes: neurodegenerative diseases (ND; 27 traits, including Autism, Alzheimer's, Parkinson's, etc.) and autoimmune disorders (AI; 23 traits, including Diabetes, Arthritis, Lupus, etc.); for the analyses below, we consider the subset of variants that are not in LD with any variant in the GWAS Catalog or GWAS Central (283 ND SNPs and 155 AI SNPs).

We also collected two sets of genes that were found to be highly expressed in brain cells as well as in blood cells; specifically, we reasoned that SNPs associated to neuropsychiatric and autoimmune diseases should be more highly enriched near genes expressed in brain and immune cells, respectively. Indeed, we found that variants associated with ND diseases (32 ND SNPs in total) occurred significantly more often within 200Kbp of genes with preferential brain expression, while variants associated with AU traits (15 variants in total) were found much more frequently in near genes with preferential blood expression ($p < 0.05$; see Supplementary Material).

We should note however that the vast majority of ND and AU variants were found far from coding regions. To test whether this set of SNPs also make biological sense, we used GREAT[13], a tool which annotates the function of variants in intergenic areas of the genome. In particular, GREAT links intergenic regions with Disease Ontology (DO) terms, and outputs terms that are significantly enriched for a particular set of variants. When we applied GREAT to ND SNPs, we found a strong enrichment in regions known to play a role in ND-related phenotypes, such as cognitive disease ($p < 10^{-32}$), dementia ($p < 10^{-23}$), and neurodegenerative disease ($p < 10^{-23}$). Similarly, AI variants were significantly associated with AI-related terms, the most significant of which were disease by infectious agent ($p < 10^{-27}$), viral infectious disease ($p < 10^{-19}$), and autoimmune disease ($p < 10^{-17}$). In fact, the top 20 DO terms for either set of variants were all exclusively associated with the correct family of phenotypes (Supplementary Tables 1,2). Hence, our predicted variants were highly consistent with these external annotations.

Examining the Effect Sizes of Novel GwasKB Variants

Finally, we analyzed the magnitude with which novel variants affect their predicted phenotypes and other, related traits. Specifically, we used freely available GWAS summary statistics from the LD Hub project[14] to assess the distribution of SNP effect sizes across novel variants and compared them to those of random SNPs. We focused on the 11 most frequent traits in our dataset for which summary statistics were available; for each trait, we identified an LD Hub study that provides effect sizes (in the form of beta coefficients or log odds ratios) for that trait. Figure 4 compares the distribution of effect sizes of the novel variants identified by GwasKB to the distribution of effects sizes for all SNPs, again restricting to variants that show no LD with other variants in GWAS databases. Whereas the distribution of random SNPs is centered around zero, as one would expect, novel SNP effect sizes appear to follow a different distribution (Kolmogorov-Smirnov Test; see Figure 4 and Supplementary Figures 1,2) and tend to have significantly higher magnitudes than expected.

We also examined the effects of GwasKB variants on phenotypes which are known to be related to their primary, predicted trait. For each pair of diseases, we took the set of variants that GwasKB found to be associated with the first disease, and computed their average absolute effect size using summary statistics from the second disease; in several cases, variants that we determined to be associated with one trait (e.g. Obesity) also had large effect sizes on related traits (e.g. BMI).

Specifically, we used a permutation test to compute the probability of observing the absolute average effect size among novel variants within a random set of SNPs; Figure 5 shows the resulting matrix of p-values (we only include traits for which we computed at least one small p-value). In particular, we found that three traits (Obesity, BMI, Type 2 Diabetes) shared variants with high effect sizes. These three phenotypes are known to be highly correlated.

Interestingly, we also observed an unusual correlation between LDL Cholesterol levels and Alzheimer's disease. To investigate this further, we repeated the same analysis using variants that have been confirmed by the GWAS Catalog (Supplementary Figures 3-5). The resulting matrix resembles closely that of novel variants and also shows correlation between Alzheimer's and LDL cholesterol. We also found a novel variant (rs6857) that was previously found to be associated with Alzheimer's[15]; our system also correctly determined that is associated with LDL[16] at $p < 10^{-7}$; this association is notably missing from current manually-curated databases.

Discussion

The importance of curation. If GWAS associations are not recorded in a database, they are effectively missing for many practical purposes, e.g. for training machine learning systems (to predict SNP function). GWAS studies are also costly (often involving genotyping tens of thousands of subjects), and it thus a waste of research funding to not fully record their results.

An alternative to curation to ask authors to directly report their findings online. This is already possible within GWAS Central, although in practice not all authors do this, and hence the database is far from complete. In addition, past studies still need to be curated. An ideal solution appears to involve a combination of authors, machines, and curators.

Hand-curation is a difficult task. Why do manual curation efforts miss certain associations? Curating papers is often a tedious task involving browsing through highly technical material in search of short snippets of text. Humans are generally not well-suited to this kind of work: they may accidentally skip table rows, or become tired and skip a paragraph. Curation also requires understanding advanced technical concepts such as linkage disequilibrium or multiple hypothesis testing. This makes the task unsuitable for crowdsourcing approaches.

Machines may outperform humans. Computers, on the other hand, don't suffer from the aforementioned limitations: they excel at repetitive work and only need to be programmed by experts once. Crucially, even though machines make errors, these errors are systematic, not random: one may follow an iterative process of fixing these errors and redeploying the system, until a sufficient level of accuracy is reached. Redeploying our system takes on the order of hours, while asking humans to return and correct their errors would take at least months.

Of course, humans also have many advantages over machines. Indeed, the sets of GwasKB and human-curated associations were quite distinct, with thousands of relations present in one set, but not the other. The most accurate and complete GWAS database is in fact a combination of both sources. In the future, we see curation as a collaboration between humans and machines.

Biomedical information extraction. Extracting structured relations from unstructured text is subject of the field of information extraction[17] (IE). Information extraction is widely used in diverse domains such as news[18], finance[19], geology[20], and in the biomedical domain. In the biomedical setting, IE systems have been used to parse electronic medical records[21], identify drug-drug interactions[22], and associate genotypes with drug response[23]. A considerable amount of effort has gone into uncovering gene/disease associations from biomedical literature[24]. Our approach, however, takes a different approach, as it attempts to identify the effects of individual mutations. Recently, Jain et al. applied information extraction to the GWAS domain[25]; their work focused on creating extractors for two specific relations: paper phenotypes and subject ethnicities; these extractors achieved an 87% precision-at-2 and a 83% F1-score on the two tasks, respectively. In contrast, our works introduces an end-to-end system that extracts full (phenotype, rsid, pvalue) relations comparable to ones found in hand-curated databases.

Applications beyond GWAS studies. Dozens of literature curation efforts are currently underway in cancer genomics, pharmacogenomics, and many other fields. Our findings hint at the possibility of using machine curation there as well.

The GWAS domain is in many ways easier than others since variants have standardized identifiers and a lot of information is structured in tables. Nonetheless, it allows us to demonstrate the importance of machine curation and to develop a core system that can be generalized to other domains. Within the GWAS setting, our system can be further improved by extracting additional relations (e.g. risk alleles, odds ratios).

Conclusion

In summary, we have introduced in this work a new machine reading system for extracting structured databases from publications describing genome-wide association studies, and we have used it to both recover many known relations, as well as a number of associations that were not present in any existing repository.

Our results demonstrate how machine reading algorithms may help human curators synthesize the large amount of knowledge contained in the biomedical literature. This knowledge can be made widely accessible using new systems that combine the efforts of both human and machines, thus accelerating the pace of discovery in science.

Online Methods

Detailed Description of GwasKB

GwasKB is implemented in Python on top of the Snorkel information extraction framework[11]. Snorkel provides utilities for parsing XML documents and training machine learning classifiers. GwasKB extends the parsers/classifiers in Snorkel and applies them to the GWAS extraction task. Below, we provide additional details on the various components of GwasKB

Identifying simple phenotypes. We parse paper titles and abstracts and generate candidates from the EFO, Snomed and Mesh ontologies. We use 11 labeling functions (LFs), which include the following: is the mention in the title; is the mention less than 5 characters; does the mention contain nouns; is the mention in the first half of the sentence, etc. We include the full list of labeling functions in our open-source GitHub repository. The high-level phenotype is the set of three highest scoring mentions exceeding a user-specified score threshold or the single highest mention if none exceeds the threshold; this enables us to handle multiple valid phenotypes.

Identifying precise phenotypes. We only parse tables and generate candidates from cells whose header contains the words ”phenotype", “trait", or “outcome". Candidate p-values are generated by matching a regular expression; candidate relations consist of horizontally aligned phenotype and p-value candidates. We use three labeling functions: is the candidate mostly a number; is the header of the cell (indicating it's in a phenotype column) very long; does the mention contain words referring to an rsid. The module is described in more detail on GitHub.

Resolving Acronyms. We resolve acronyms by looking at the entire paper, including tables and the main natural language text in the body of the paper. We extract candidates from aligned pairs table cells, where one row is labeled ”phenotype”, “trait”, or “description”, while the other is labeled “abbreviation”, “acronym”, or “phenotype”. We

generate candidates from the main text using a regular expression. Our labeling functions include the following: is the candidate all in caps; does the candidate match to the Snomed dictionary; does the acronym candidate consist of the letters of each word of the phenotype candidate; is one a prefix of the other; etc. The module for resolving abbreviations is also described on GitHub.

Identifying p-values. We again generate candidates from tables; SNP candidates are generated using a regular expression; p-value candidates are ones that match one of three regular expressions (see GitHub); candidate relations consist of horizontally aligned SNP and p-value candidates (with at most one rsid per row). These candidates were accurate and we report them all.

Mapping Phenotypes Across Databases

In order to compare against GWAS Central and GWAS Catalog, we define mappings between GwasKB phenotypes and ones used in these two repositories. These mappings are tables with about 800 entries each that also indicate whether the mapping is fully or partially correct (e.g. “smoking behaviors" vs “packs per day"). We define the latter as conceptually containing the precise label while also being not so broad as to be useless. See also our earlier discussion on high- and low-level phenotypes. To confirm the validity of our mappings, we asked an independent annotator to label 100 random table entries; their concordance with our labels was 95%. These mappings are available in our GitHub repository.

Understanding The Errors Of GwasKB Components

Simple phenotype extraction. Errors at this stage mostly occur when the true phenotypes are not found in our candidate dictionaries (e.g. “genome-wide association study in bipolar patients"; we can only generate the candidate “bipolar disorder"). The second major source of error are phenotypes mentioned only in passing (e.g. “high body fat is a

risk for diabetes" when diabetes is not the phenotype whose association is being reported).

To estimate the precision of this module, we first restrict ourselves to (paper, rsid, phenotype) relations produced by GwasKB that are also confirmed by an existing database, in the sense that the variant specified by the rsid occurs in *some* relation associated with the paper (but not necessarily one with the same phenotype). Then, we look at the fraction of these relations whose phenotype is also correct (at the approximate level). This gives precisions of 97% in the GWAS Catalog and 96% in GWAS Central.

Detailed phenotype extraction. Most errors occur because we do not correctly resolve acronyms or because low-level phenotypes are not in tables (but rather only in text). Acronyms are not resolved most often because the shortened symbol is not clearly related to the full expression (e.g. CYS5 for Cysteine proteinase inhibitor 5 precursor), and they are presented in tables with confusing formatting. We estimate precision in the same way as for simple phenotypes, but this time, we require that phenotype agree fully. Precision was 73% in GWAS Central, the database with the most precise phenotypes. In GWAS Central, it was 82%.

p-values. To evaluate p-value extraction accuracy, we labeled by hand 100 random relations and found that our rule-based extraction procedure had a precision of 98%. Errors occurred when p-values referred to other entities in the row, such as haplotypes. Note also that oftentimes, variants and their p-values are only provided in text but not in tables. This was the primary reason why we failed to report the rsid's of 584 (15%) GWAS Catalog and 432 (14%) GWAS Central associations.

Error Analysis Over 100 New Relations

Of the incorrect relations, 7 were due to incorrect phenotype labels (but the underlying SNP was significant) and 4 were due to table parsing errors (the p-value was extracted incorrectly). Of the 22 variants that were not significant in all cohorts, 18 could be

identified as such via extracted tags and relations. We also determined that 60 relations were correct because they were either described as “significant" in the paper text (in addition to having $p < 10^{-5}$ in all cohorts) or they had essentially the same or higher level of significance as SNPs that were included in GWAS Catalog or GWAS Central.

To confirm the accuracy of our analysis, we asked two independent annotators with expertise in genomics to label a random subset of 50 associations out of the ones analyzed above. For each annotator, respectively 47 and 48 out of 50 labels were consistent with ours. We publish our 100 samples and their annotation on GitHub; for each example, we add a justification for our label.

Estimating the Precision of GwasKB

We estimate our overall precision at 92%: we consider the 3463 relations confirmed by existing databases as correct, and estimate the error rate on the other relations to be 18% (incorrect and repeat relations).

1. Bush, W.S. & Moore, J.H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **8**, 1-11 (2012).
2. Welter, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001-D1006 (2014).
3. Beck, T., Hastings, R.K., Gollapudi, S., Free, R.C. & Brookes, A.J. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *European Journal of Human Genetics* **22**, 949-952 (2013).
4. Cariaso, M. & Lennon, G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.* **40**, D1308-D1312 (2012).
5. *Promethease*.
6. de Leeuw, C.A., Mooij, J.M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* **11**, e1004219+ (2015).
7. Weng, L. et al. SNP-based pathway enrichment analysis for genome-wide association studies. *BMC Bioinformatics* **12**, 99+ (2011).
8. Zhou, J. & Troyanskaya, O. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Meth* **12**, 931-934 (2015).
9. Ratner, A.J., Bach, S.H., Ehrenberg, H.R. & Ré, C. Snorkel: Fast training set generation for information extraction. *Proceedings of the 2017 ACM International Conference on Management of Data* 1683-1686 (2017).
10. Manning, C.D. et al. The Stanford CoreNLP Natural Language Processing Toolkit. *Association for Computational Linguistics (ACL) System Demonstrations* 55-60 (2014).
11. Ratner, A.J., De Sa, C.M., Wu, S., Selsam, D. & Ré, C. Data programming: Creating large training sets, quickly. *Advances in Neural Information Processing Systems* 3567-3575 (2016).
12. Mintz, M., Bills, S., Snow, R. & Jurafsky, D. Distant supervision for relation extraction without labeled data. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2* 1003-1011 (2009).
13. McLean, C.Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).
14. Zheng, J. et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272-279 (2017).
15. Yu, C. et al. Comprehensive analysis of APOE and selected proximate markers for late-onset Alzheimer's disease: patterns of linkage disequilibrium and disease/marker association. *Genomics* **89**, 655-665 (2007).
16. Deshmukh, H.A. et al. Genome-wide association study of genetic determinants of LDL-c response to atorvastatin therapy: importance of Lp (a). *Journal of lipid research* **53**, 1000-1011 (2012).
17. Moens, M. *Information Extraction: Algorithms and Prospects in a Retrieval Context*. (Springer Netherlands: 2009).

18. Tumarkin, R. & Whitelaw, R.F. News or Noise? Internet Postings and Stock Prices. *Financial Analysts Journal* **57**, 41-51 (2001).
19. Das, S. & Chen, M. Yahoo! for Amazon: Extracting Market Sentiment from Stock Message Boards. *Proceedings of the Asia Pacific Finance Association Annual Conference (APFA)* (2001).
20. Zhang, C. et al. GeoDeepDive: Statistical Inference Using Familiar Data-processing Languages. *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data* 993-996 (2013).doi:10.1145/2463676.2463680
21. Zhou, X., Han, H., Chankai, I., Prestrud, A. & Brooks, A. Approaches to Text Mining for Clinical Medical Records. *Proceedings of the 2006 ACM Symposium on Applied Computing* 235-239 (2006).doi:10.1145/1141277.1141330
22. Percha, B., Garten, Y. & Altman, R.B. Discovery and explanation of drug-drug interactions via text mining. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 410-421 (2012).
23. Rinaldi, F., Schneider, G. & Clematide, S. Relation Mining Experiments in the Pharmacogenomics Domain. *J. of Biomedical Informatics* **45**, 851-861 (2012).
24. Pletscher-Frankild, S., Pallej a, A., Tsafou, K., Binder, J.X. & Jensen, L.J. DISEASES: Text mining and data integration of disease,Äìgene associations. *bioRxiv* 008425+ (2014).doi:10.1101/008425
25. Jain, S. et al. Weakly supervised learning of biomedical information extraction from curated data. *BMC Bioinformatics* **17**, S1 (2016).

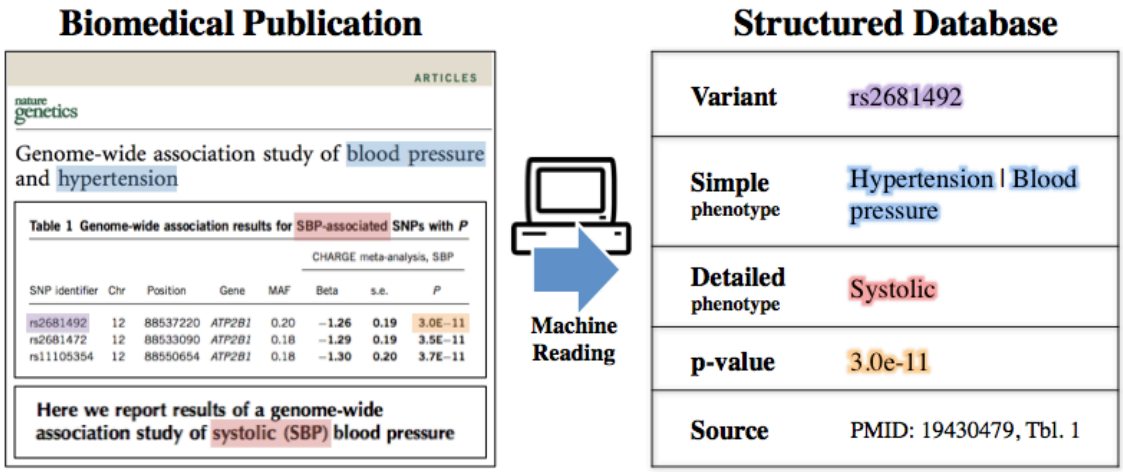


Figure 1: The GwasKB machine reading system. GwasKB takes as input a set of biomedical publications retrieved from PubMed Central (left) and automatically creates a structured database of GWAS associations described in these publications (right). For each association, the system identifies a genetic variant (purple), a high-level phenotype (pertaining to all variants in the publication), a detailed low-level phenotype (specific to individual variants, if available; red), and a p-value (orange). Acronyms are also resolved (red).

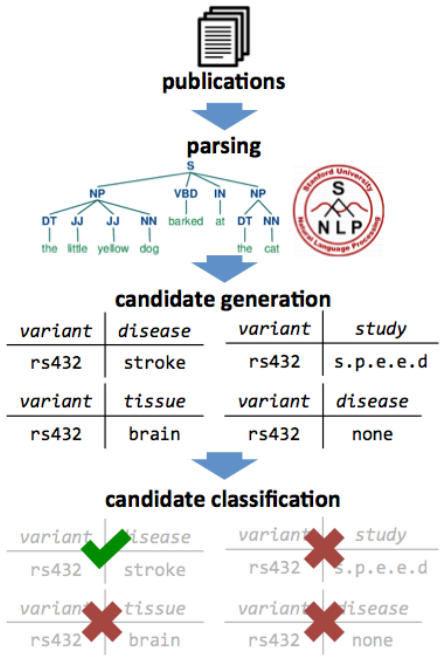


Figure 2: General structure of a GwasKB module. The system contains separate modules for extracting variants, phenotypes, p-values, and for resolving acronyms. Each module consists of three stages. At the parsing stage, we process papers using the Stanford CoreNLP pipeline, performing full syntactic parsing. Next, given a target relation (e.g., variant-phenotype), we generate a large set of candidates, some of which could be correct instances of the target object on relation. Then, at the classification stage, we determine which candidates are correct using a machine learning classifier.

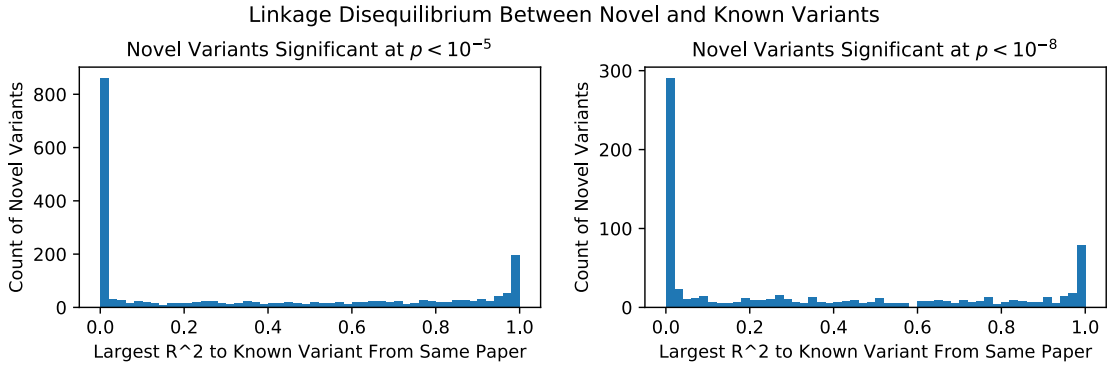
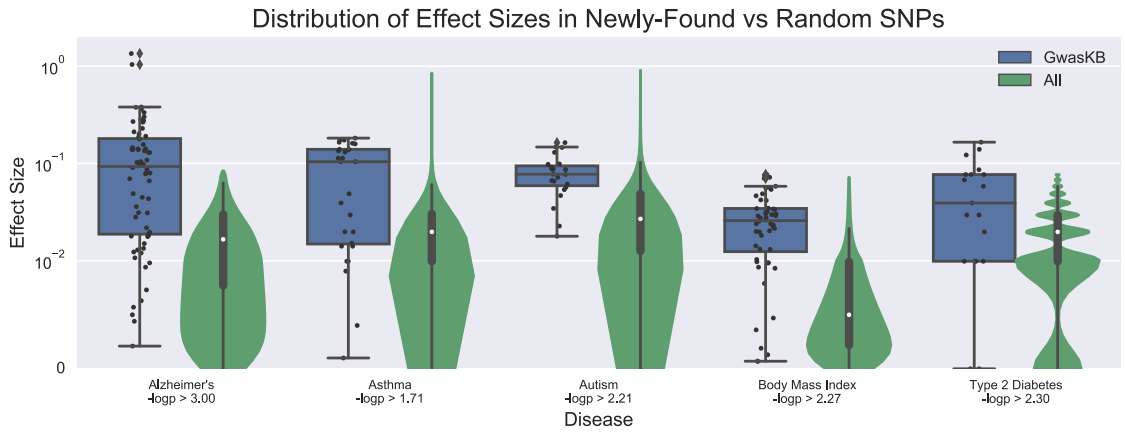
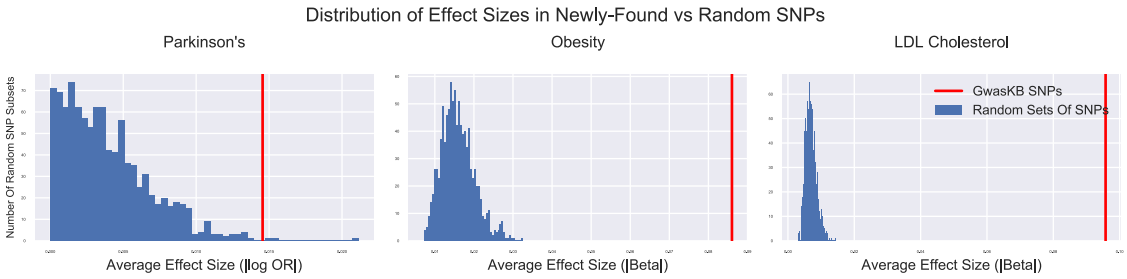


Figure 3: Linkage disequilibrium between GwasKB variants not present in existing human curated databases and variants from the GWAS Catalog. We use the 1000 Genomes dataset to estimate the r^2 metric between pairs of variants, and report distances from each GwasKB variant to the most correlated GWAS Catalog SNP reported in the same paper. The distribution of r^2 scores is highly multimodal; many GwasKB variants are uncorrelated (r^2=0) with GWAS Catalog SNPs.



588



589

590 **Figure 4:** Visualizing the effect sizes of variants identified by GwasKB. *Top:* We
591 compare the distribution of effect sizes (absolute values of beta coefficients or log odds
592 ratios; data from LD Hub) of variants identified by GwasKB (blue) to that of all variants
593 (green) for multiple traits. Blue variant effect sizes cluster away from zero and follow a
594 different distribution (Kolmogorov-Smirnov test). *Bottom:* We subsample 1000 random
595 sets of variants with the same number of elements as the set of GwasKB SNPs for a given
596 disease; the average effect size of GwasKB variants (red) is higher than that of the
597 random subsets (blue). In all settings, we only look at novel GwasKB variants not present
598 in existing human-curated repositories.

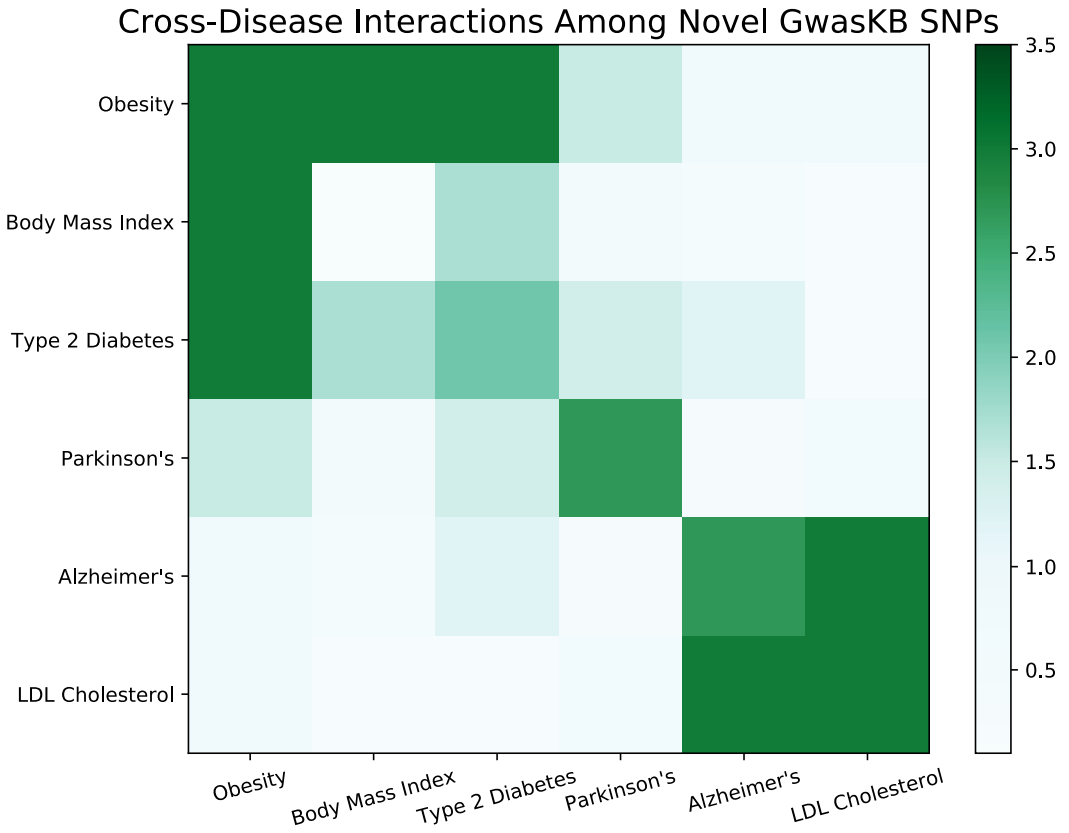


Figure 5: Visualizing the effects of variants identified by GwasKB for pairs of related phenotypes. For each pair of phenotypes, we compute the average absolute effect size of GwasKB SNPs from the first phenotype (left) using summary statistics from the second phenotype (right; summary statistics were obtained from the LD hub). The heat map displays the log-probability of observing an equal or greater effect size by sampling random variants (we thus compute p-values using a one-sided permutation test). Variants predicted by GwasKB to be associated with Obesity, BMI, or Type 2 Diabetes also have significant effects sizes for other, related diseases within this trio. In this analysis, we only look at novel GwasKB variants not present in existing human-curated repositories.

Database	Statistics over open-access papers		
	Papers	**Associations**	**Unique Associations**
GWAS Catalog	589	8,384	>2,026
GWAS Central	516	5,914	>364
GwasKB (ours)	589	6,231	>2,777

Table 1: Numbers of associations contained in different GWAS databases. Unique associations are contained in one database and in none of the others. Human curated databases (GWAS Catalog and GWAS Central) significantly differ in their scope. Our machine-curated repository (GwasKB) automatically recovers a large fraction of known results and also finds a comparable number of unique associations.

	Source	Simple phenotype	Precise phenotype	p-value
Study	*Genome-wide pharmacogenomic study of metabolic side effects to antipsychotic drugs.*			
rs17661538	**GwasKB**	Antipsychotic drugs / Metabolic side effects	Clozapine - Triglycerides	1.00E-06
	GwasCat	Clozapine-induced change in triglycerides		1.00E-06
Study	*Genome-wide meta-analysis identifies seven loci associated with platelet aggregation in response to agonists.*			
rs12566888	**GwasKB**	Platelet aggregation	-	5.00E-19
	GwasCat	Platelet aggregation, epinephrine		5.00E-19
Study	*A genome-wide association study of the Protein C anticoagulant pathway.*			
rs13130255	**GwasKB**	Protein C	funcPS	3.00E-06
	GwasCat	Anticoagulant levels (funcPS)		3.00E-06
Study	*Genome-wide association study of CSF levels of 59 Alzheimer's disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation.*			
rs948399	**GwasKB**	Proteins Involved / Inflammation / Alzheimer's Disease	metalloproteinase-3	1.00E-07

Table 2: Examples of associations identified by GwasKB. Associations can be classified as correct (rs17661538), partially correct (rs12566888; the precise phenotype is missing) and incorrect (rs13130255). We also compare these associations to their corresponding entries in the GWAS Catalog. The last entry (rs948399) is an example of a previously undocumented association discovered by our system.

Contributions

V.K. conceived the study. B.H., V.K., and A.R. developed modules for the Snorkel system. V.K. developed the GwasKB system. V.K., J.D., and C.V. performed computational analysis. J.D. developed the web interface. V.K. and Y.L. wrote the paper. Y.L., C.R., S.B., and M.S. supervised the study.

Competing interests

None declared.