A Machine-Compiled Database of Genome-Wide Association Studies

Volodymyr Kuleshov
Stanford University
kuleshov@cs.stanford.edu

Braden Hancock
Stanford University
bradenjh@cs.stanford.edu

Alexander Ratner
Stanford University
aratner@stanford.edu

Christopher Re
Stanford University
chrismre@cs.stanford.edu

Serafim Batzoglou
Stanford University
serafim@cs.stanford.edu

Michael P. Snyder
Stanford University
mpsnyder@stanford.edu

Abstract

Tens of thousands of genotype/phenotype associations have been discovered by experimental studies, yet not all of them are available to scientists in a useful, easy to access format. Here, we describe GwasDB, a machine reading system for automatically extracting these associations from the scientific literature in the form of a structured database. Our system reveals that existing manually-curated repositories are highly incomplete, and produces $> 3,000$ previously undocumented associations, which represents about 30% of the size of the largest existing repository of open-access papers. Our results highlight both the importance and the feasibility of using machines to help human curators keep track of important scientific findings obtained using public research funding.

1 Introduction

Genome-wide association (GWAS) studies are one of the most widely-used methodologies for measuring the effects of genomic mutations on human traits. Tens of thousands of genotype/phenotype associations have been discovered by GWAS studies, yet not all of them are available to scientists in a useful, easy to access form.

Multiple efforts are currently trying to catalogue all published GWAS associations. Despite these efforts, it is clear that our grasp on published GWAS results is far from complete. Even the largest hand-curated databases greatly vary in their scope: hundreds to thousands of variants may be present in one repository, but not in any other. Variants that have not been catalogued in a database are effectively lost for most practical applications; this limits the pace of scientific research and represents a wasteful use of public research funding.

Here, we describe GwasDB, a machine reading system that enables researchers to identify thousands of undocumented variant/phenotype associations from the GWAS literature in the form of a structured database like the ones developed by manual curation efforts. When deployed on a set of 589 open-access GWAS publications, GwasDB finds more than 3,000 associations that existing databases do not contain. This represents about 30% of the total number of associations recorded in the largest existing database, GWAS Catalog.

Our results highlight the need and the feasibility to improve curation efforts with machine reading algorithms to avoid losing track of important scientific findings obtained through public funding. More generally, we hope to demonstrate that information extraction has the potential to greatly increase the amount of data available for research in personalized medicine.

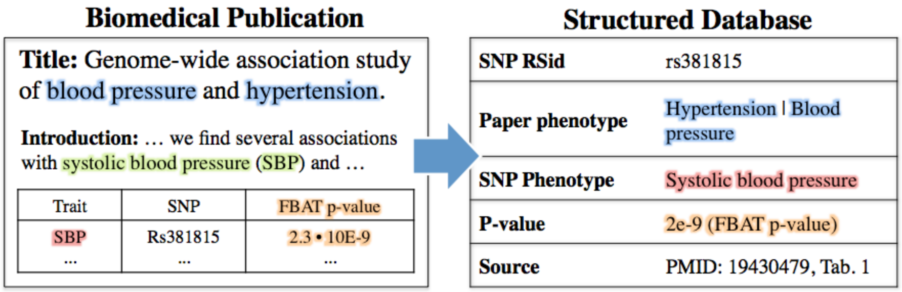

Figure 1: The GwasDB system reads XML papers from PubMed Central (left) and produces a structured database of GWAS associations (right). For each association, it identifies a high-level (paper-level) phenotype (blue), a detailed low-level phenotype (if available; red), and a p-value (orange). Acronyms are also resolved (green).

2 The GwasDB System

Genome-wide association studies. These large case/control studies attempt to find single-nucleotide polymorphisms (SNPs) that confer an increased risk for a given phenotype. Their results are widely used in enrichment analysis, for training classifiers to predict the effects of new mutations [6], and for computing disease risks in personalized genome interpretation [3].

By various estimates, about 2,000-3,000 studies have been performed to date. Their results are manually compiled in databases like GWAS Catalog [1] and GWAS Central [2], which hold approximately 25,000 associations, a third of which are in open-access journals (Table 1).

Inputs and outputs of GwasDB. GwasDB extracts SNP/phenotype relations from biomedical papers in PubMed XML format, which can be downloaded from the PubMed Central (PMC) repository. Due to copyright restrictions, we only deploy our system on open-access papers and we also do not handle supplementary material in proprietary (e.g. MS Word or Excel) formats.

The output of GwasDB is a SQL database of relations extracted from papers. The most important subset can be described as (rsid, phenotype) tuples that represent statistically significant associations identified by the system; we also extract (rsid, p-value) relations, multiple phenotypes, and we support our findings with evidence from the paper (sentences excerpts or locations in tables). GwasDB reports all (rsid, phenotype) associations that are significant at $p < 10^{-5}$ in at least one such cohort or statistical model.

GwasDB reports ta high-level phenotype — such as "heart anomalies" — and, when applicable, a more detailed, low-level phenotype assigned to a particular SNP, e.g. "ventricule size" or "artherial thickness". Low-level phenotypes are needed when a paper studies multiple, related traits; however, extracting low-level phenotypes is more challenging, and we also report the less precise (but often useful) high-level phenotype. We measure accuracy on both tasks.

GwasDB modules. GwasDB is comprised of a set of five modules that each focus on a different relation. The first module parses the paper title and abstract to identify a high-level phenotype. A second module parses the contents of the paper to uncover (rsid, low-level phenotype) relations. Often, the low-level phenotype is given as an abbreviation (e.g. BMI); thus, a third module attempts to resolve these abbreviations (e.g. translate BMI to body mass index). A fourth module parses XML tables to extract (rsid, p-value) relations; we currently look only at tables, as we have found them to contain the vast majority of relations and their relative structure enables us to develop more accurate extraction algorithms. Finally, a fifth module combines and evaluates all our results.

Module structure. Each GwasDB module implements three stages: parsing, candidate generation, and candidate classification. At the parsing stage, we process relevant parts of the paper (e.g. abstract, XML tables) using the Stanford CoreNLP pipeline. CoreNLP performs tokenization, part-of-speech tagging, and a full syntactic parsing. In addition, table contents are parsed into cells. Next, we generate a very large and noisy set of candidates among which are true mentions of SNPs, phenotypes, p-values, or other quantities of interest. Candidates are generated by matching all subspans

Database	Statistics over open-access subset of database			
	Papers	**Variants,** $p < 10^{-5}$	**Unique vars.,** $p < 10^{-5}$	**Precision**
GWAS Catalog	589	$8,384^{\dagger}$	$1,992^{\dagger}$	100%
GWAS Central	$516^{\ddagger}$	5,914	359	100%
GwasDB (ours)	589	6,541	3,030	$\approx 90\%$

†: GWAS Catalog uses more stringent inclusion criteria than the other two databases. ‡: We estimate this by dividing the total number of papers by the fraction that were open-access in GWAS Catalog.

Table 1: Comparison of GwasDB to manually curated databases. "Variants" refer to (pmid, rsid) tuples; unique variants occur only the given database. The same variant can occur in multiple associations.

to regular expressions, or by identifying matches within a known dictionary (e.g. the Snomed ontology of diseases). Finally, we determine which candidates are correct using a generative machine learning classifier trained using the data programming paradigm using a small set of hand-crafted labeling functions.

Module details. To identify phenotypes, we parse paper titles and abstracts and generate candidates via a dictionary search. Our dictionary includes the EFO, Snomed and Mesh ontologies. We label candidates using 11 labeling functions (LFs) and use the data programming approach to learn their weights [5]. We classify mentions using the weighted LFs; we report the high-level phenotype as the three highest scoring mentions exceeding a user-specified score threshold (for example, to capture multiple phenotypes dfined in the abstract) or the highest mention if none exceeds the threshold. Labeling functions used at this stage include: 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. Other GwasDB modules follow a similar approach.

Implementation and availability. GwasDB is implemented in the Python language on top of the Snorkel information extraction framework [4]. A walkthrough of our system is available on GitHub at `github.com/kuleshov/gwasdb`.

3 Evaluation

Datasets. We deploy GwasDB on all open-access papers listed in the GWAS Catalog database (589 in total). The GWAS Catalog is among the most complete databases and its papers can be directly downloaded (unlike GWAS Central). For evaluation, we focus on the GWAS Central database because its reported phenotypes are more detailed than in the GWAS Catalog.

Methodology. The most important output of our system is a table of (rsid, phenotype) associations that are significant at $p < 10^{-5}$ in at least one cohort or study methodology. We evaluate these associations in two ways: we assess recall by looking at the number of GWAS Central relations that our system recovers; we assess precision by manually inspecting the accuracy of 50 random new GwasDB relations not present in GWAS Central.

Recall GWAS Central reports 7583 associations (with $p < 10^{-5}$ in at least one cohort or study) in our input dataset of 589 open-access GWAS papers. Of these, 3087 (41%) were accessible to us (the `rsid`was present somewhere in the XML paper or its plain-text attachments); the remaining associations are found in Word/Excel attachments that we currently do not parse.

Of the 3087 accessible associations, GwasDB recovered 1884 (61%) with maximum precision, and 2574 (83%) with partially-accurate precision (see above); of the 513 relations that could not be recovered, in 42 cases, the `rsid`was found, but the phenotype was incorrect, and for 471 relations the `rsid`was not reported at all.

Precision GwasDB discovered a total of 7838 relations within the 589 input papers, 5222 (67%) of which contained an `rsid`not found in GWAS Central for the corresponding paper. We manually inspected a random subset of 50 novel relations and found that 28 (56%) were to the best of our knowledge fully correct (i.e. satisfying the same criteria as GWAS Central relations from the same paper), 12 (23%) were not significant at 10^{-5} in all cohorts (and thus may have been deemed not

	Source	**General phenotype**	**Precise phenotype**	**min. p-val**
Study	Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs.			
rs17661538	GwasDB	Antipsychotic drugs / Metabolic side effects	Clozapine - Triglycerides	1e-6
	GwasCen	Clozapine-induced change in triglycerides		1e-6
Study	Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists.			
rs12566888	GwasDB	Platelet aggregation	-	5e-19
	GwasCen	Platelet aggregation, epinephrine		5e-19
Study	A genome-wide association study of the Protein C anticoagulant pathway.			
rs13130255	GwasDB	Protein C	funcPS	3e-06
	GwasCen	Anticoagulant levels (funcPS)		3e-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.			
rs4845622	GwasDB	Proteins Involved / Inflammation / Alzheimer's Disease	IL6R	2e-14

Table 2: Examples of a correct (row 1), partially correct (row 2) and incorrect (row 3) relation extracted by GwasDB, and the corresponding entry in GWAS Central. The last entry (row 4) is an undocumented relation that we discovered.

significant by the study, even though they match the specifications of our system), 2 (3%) were correct but were originally identified by a different study (and referenced as background material), and another 5 (10%) were definitely incorrect.

Error analysis A common cause of errors were unusually-formatted tables that our system could not parse. Another issue was phenotype identification: phenotypes are described in the paper using words that our dictionaries don't contain (e.g. "genome-wide association study in bipolar patients"; we only know about "bipolar disorder") or are mentioned only in passing (e.g. "high body fat is a risk for diabetes").

4 Conclusion

In summary, we have introduced in this work a new machine reading system for extracting structured databases from genome-wide association studies, and we have used it to uncover thousands of new relations that were not present in any existing repository.

There variant/phenotype associations were effectively lost for many practical purposes, despite the fact that significant amounts of public funding were spent to obtain them. Our results demonstrate the existence of these "dark associations" and also provide a practical system for accessing a large number of them.

More generally, we show how machine reading algorithms can be used to help harness the large amount of knowledge generated within genomics. This knowledge can be made accessible via new systems that combine the efforts of both human and machines, thus accelerating the pace of discovery in science.

References

[1] Gwas catalog. https://www.ebi.ac.uk/gwas/. Accessed: 2016-09-20.

[2] Gwas central. gwascentral.org. Accessed: 2016-09-20.

[3] Promethease. https://promethease.com/. Accessed: 2016-09-20.

[4] Snorkel. github.com/HazyResearch/snorkel/. Accessed: 2016-09-20.

[5] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *CoRR*, abs/1605.07723, 2016.

[6] J Zhou and OG Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(5):931–4, 2015-10-01 00:00:00.001.