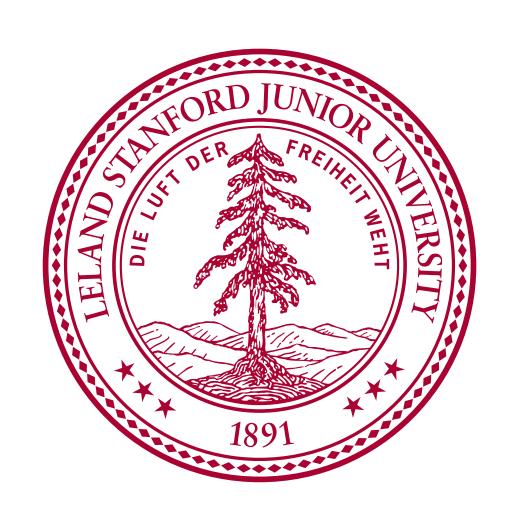
A Machine-Compiled Database of Genetic Disease

Volodymyr Kuleshov, Braden Hancock, Alex Ratner, Christopher Re, Serafim Batzoglou, Michael Snyder Department of Computer Science and Department of Genetics, Stanford University

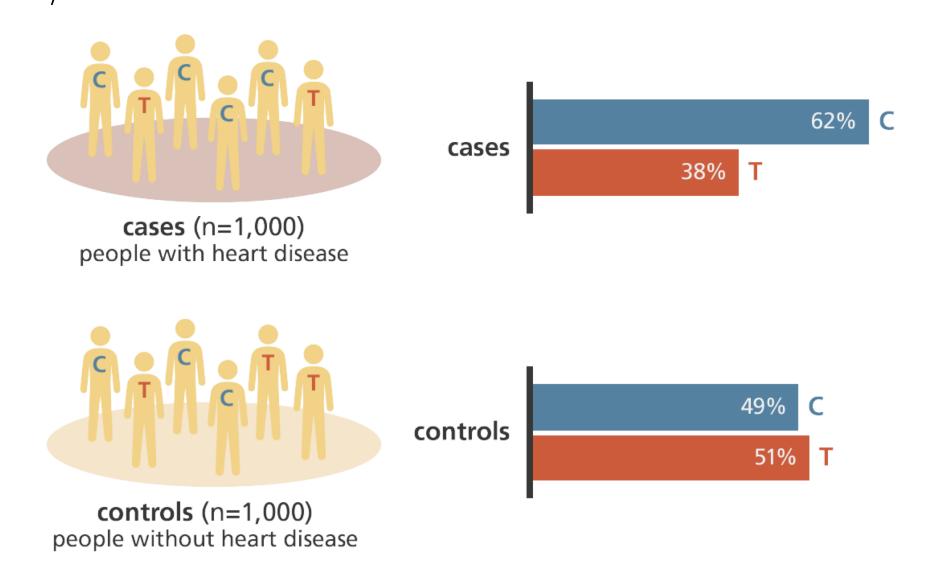


Highlights

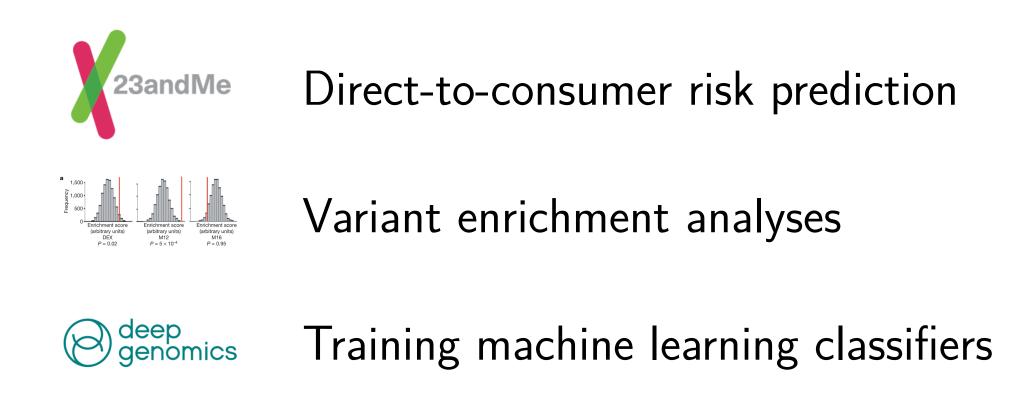
- We built a machine reading system for extracting genotype/phenotype associations from literature.
- Our system uncovers thousands of associations not present in any human-curated database
- \bullet We estimate having 92% precision and 81% recall relative to existing knowledge bases.

Genome-Wide Association Studies

Case/control studies to find disease-related variants.



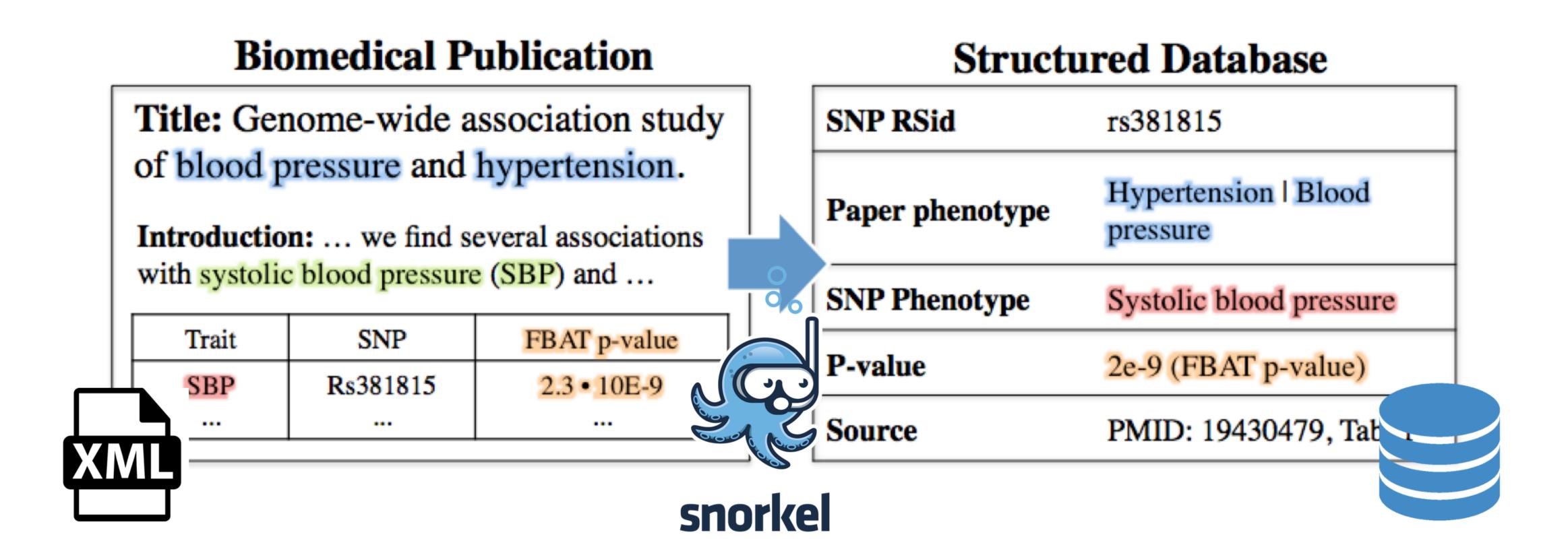
Applications of GWAS



Current Curation Efforts

- **GWAS Catalog** contains > 25k associations from 589 papers.
- **GWAS Central** contains > 60k associations from papers published before 2014; not every variant is significant at 1e-5.
- **Open-access** papers represent about 25% of papers/associations contained in these databases.

GwasDB: A Machine-Compiled Database of Genome-Wide Association Studies



For each association, GwasDB 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).

GWASdb Finds Thousands of Uncatalogued Relations in Open-Access Papers

We find \approx 3k new GWAS associations in open-access papers, i.e. about 30% of largest open-access database.

	Papers (OA)	Associations	Unique Associations
GWAS Catalog	589	8,384	2,026
GWAS Central	516	5,917	364
GwasDB (ours)	589	6,231	2,777 new!

Examples of Associations Extracted by Our System

Correct

Genomewide pharmacogenomic study of metabolic side effects to antipsychotic drugs (rs17661538)

GwasDB Antipsychotic drugs, Metabolic side effects 1e-6
(Clozapine - Triglycerides)
Clozapine-induced change in triglycerides 1e-6

Previously undocumented

Genome-wide association study of CSF levels of 59 alzheimer's disease candidate proteins: significant associations with proteins involved in amyloid processing and inflammation.

GwasDB Proteins Involved, Inflammation, Alzhei- 3e-06 mer's Disease (IL6R)

Imprecise

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

Erroneous

A genome-wide association study of the Protein C anticoagulant pathway (rs13130255)

GwasDB	Protein C (funcPS)	3e-06
GwasCen	Anticoagulant levels (funcPS)	3e-06

Recall Relative to Existing Databases

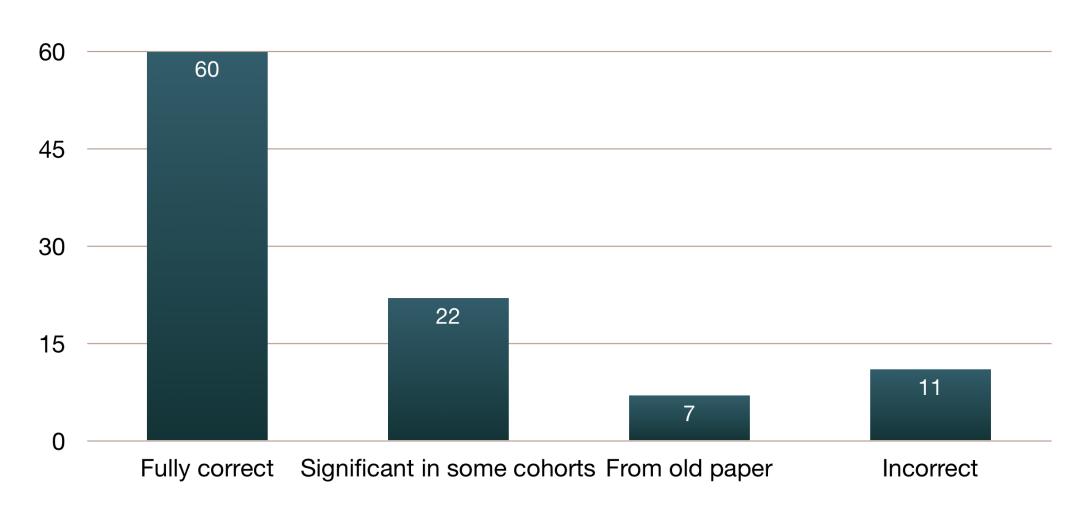
- We only look at open-access (OA) papers
- About 50% of OA associations are in Excel/Word attachments, which we don't parse

We report recall over the accessible associations.

	GWAS Cat.	GWAS Cent.
Accessible associa-	4,023	3,008
tions		
Found with variant	2,762 (69%)	1,890 (63%)
phenotype		
Found with study	3,245 (81%)	2,487 (82%)
phenotype		

Precision of New Associations

We manually inspected 100 new relations and had our results confirmed by two independent annotators



- Half of our assocations were new; half were confirmed by existing sources
- We assume confirmed associations are correct; this gives overall precision of 92%.

Takeaways

- A significant fraction of GWAS results are effectively lost.
- Hand-curation is an error-prone task
- Machine-augmented curation holds promise to improve scientific discovery.