

Andy Wei<sup>1,2</sup>, Li Fang<sup>1</sup>, Kai Wang<sup>1,3</sup>

<sup>1</sup> Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA

<sup>2</sup> Belmont High School, Belmont, MA 02478, USA

<sup>3</sup> Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Question? Please contact PI: Prof. Kai Wang via wangk@email.chop.edu (wglab.org)

## Background

A comprehensive curation of disease-causal genes and variants from published literature can greatly accelerate the identification of disease-causal variants for Mendelian disorders. For example, the HGMD database [1] was built for this purpose, but it is mainly built by manual curation and may be less scalable in an era of explosion of knowledge on genetic findings. Due to the large amounts of papers from scientific literature, text mining tools can be essential for extracting such information before manual review. In order to train a highly accurate text-mining tool to extract disease-causal genes and variants, a corpus with annotations of diseases, gene names, and variants is needed. In this work, we present the DiseaseMutation corpus, a collection of 213 full-text articles (1,732 paragraphs) fully annotated with disease, gene, and variant mentions manually to serve as a research resource for the biomedical natural language processing community.

## METHODS

Pubtator is NCBI's web-based system that provides automatic annotations of biomedical concepts using text-mining tools[2]. DiseaseMutation corpus is based on PubTator's automatic annotations, but with manual validation. We downloaded the curation of disease-causal variants from ClinGen and obtained the PMIDs of corresponding articles. Next, we retrieved the full text (with entity annotations) from PubTator [2] using these PMIDs. We then manually validated and modified PubTator's annotation using our in-house tool named LabelTools (Figure 1).

Our manual validation includes 4 categories: "Genes/proteins," "Diseases/Phenotypes," "SNPs/Indels," and "Structural Variations". If the annotations were correctly categorized, they were marked with a "Y" while annotations that were incorrect were marked with a "N." If an entity's annotation was missing, we manually added it and marked it with a "Y."

We then evaluated PubTator's automatic annotations by directly comparing them with our manual annotations. Of note, PubTator does not have the entity type of "Structural Variations," so all "Structural Variations" terms were manually added.

## RESULTS

We annotated four types of entities: gene/protein names, disease names, SNP/Indels and structural variants (SVs). Our corpus has 7,623 gene/protein annotations, 4,699 disease annotations, 8,877 SNP/Indel annotations and 32 structural variant annotations. We compared our manual annotations with PubTator [2] annotations, which were labeled by text-mining tools. In total, PubTator had 1,682 false annotations and missed 3,852 annotations. PubTator also lacks the annotation of structural variants.

Out of 22,978 total annotations, we found that 17,347 (75.49%) of the annotations were true, including 3,884 manual annotations. Manual annotations comprised of missing annotations in long lists of variants; genes, proteins, diseases, or phenotypes that belonged in a different category; many abbreviations of diseases/phenotypes, including USH, EVA, DCM, and HL; and annotations that included unrelated terms/phrases. Deleted annotations can be attributed to incorrect punctuation/syntax (parentheses, separation of variants, etc.), which accounts for 3.98% of the total annotations. The other 767 false annotations included abbreviations, like "muM" and "SIFT" and reference DNA sequences, which didn't belong in any of the categories.

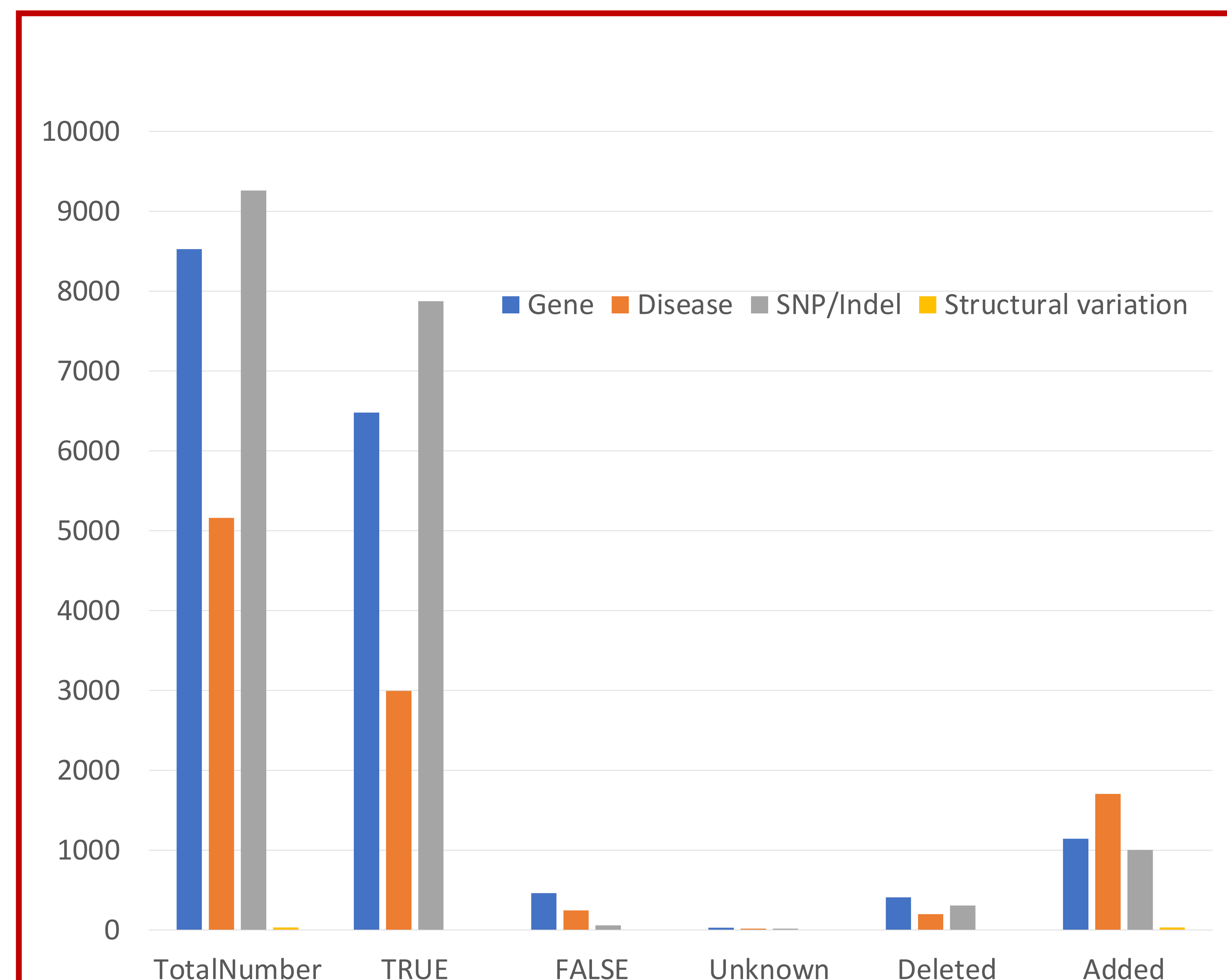


Figure 2. The summary of different annotation categories manually curated by LabelTools.

## CONCLUSIONS

These results show that the DiseaseMutation corpus has the potential to improve the accuracy of biomedical text-mining tools and facilitate the development of tools for automated extraction of disease-causal variants.

## REFERENCES

- Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeyasinghe S, Krawczak M, Cooper DN. Human Gene Mutation Database (HGMD): 2003 update. 21(6):577-81, 2003
- Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 41:W518–W522, 2013

## ACKNOWLEDGEMENTS



**LabelTools: A tool box for text labeling**  
Named Entity Labeling

Please choose your input json file: Choose File pubtator-2021-variants.json

Please check the named entity types that you want to highlight (<=7 types):

- ☒ Gene (count: 7381 )
- ☒ Disease (count: 3458 )
- ☒ SNP/Indel (count: 8255 )
- ☐ Species (count: 3132 )
- ☐ Chemical (count: 2067 )
- ☐ CellLine (count: 260 )
- ☒ Structural variation (count: 0 )
- ☐ SNP/Indel-Text (count: 0 )
- ☐ mutantGene (count: 0 )

Confirm selection Clear

< Previous passage Reset passage Next passage > Save

Title: Diverse spectrum of rare deafness genes underlies early-childhood hearing loss in Japanese patients: a cross-sectional, multi-center next-generation sequencing study

This was a multi-center study of 58 subjects (36 subjects with ? hearing loss ? and 22 subjects with normal hearing) from 15 unrelated Japanese families in which at least two family members had bilateral ? hearing loss ?. All subjects were patients at the National Hospital Organization Tokyo Medical Center or a collaborating hospital. Medical histories were obtained and physical, audiological, and radiological examinations were carried out for the subjects and family members. Subjects with ? hearing loss ? related to environmental factors were excluded. Subjects with ? GJB2 ? mutations or mitochondrial ? m.1555A>G ? or ? 3243A>G ? mutations were excluded. Subjects with ? enlarged vestibular aqueduct ?, which is often associated with ? SLC26A4 ? mutations, and subjects with clinical features that suggested ? syndromic hearing loss ? were excluded. Subjects with ? auditory neuropathy ? were tested for ? OTOF ? mutations, which are associated with ? auditory neuropathy ?, and subjects with ? OTOF ? mutations were excluded. The Ethics Review Committees of the National Hospital Organization Tokyo Medical Center and all collaborating hospitals approved the study procedures. All procedures were conducted after written informed consent had been obtained from each subject or their parents.

To add an annotation, select the text in the above box and then click the corresponding button:

Gene Disease SNP/Indel Structural variation

showing passage 2 of document 1. PMID: 24164807 PMCID: 4231469

< Previous passage Reset passage Next passage > Save

Title: Diverse spectrum of rare deafness genes underlies early-childhood hearing loss in Japanese patients: a cross-sectional, multi-center next-generation sequencing study

This was a multi-center study of 58 subjects (36 subjects with Y hearing loss ? and 22 subjects with normal hearing) from 15 unrelated Japanese families in which at least two family members had bilateral Y hearing loss ?. All subjects were patients at the National Hospital Organization Tokyo Medical Center or a collaborating hospital. Medical histories were obtained and physical, audiological, and radiological examinations were carried out for the subjects and family members. Subjects with Y hearing loss ? related to environmental factors were excluded. Subjects with Y GJB2 ? mutations or mitochondrial Y m.1555A>G ? or Y 3243A>G ? mutations were excluded. Subjects with Y enlarged vestibular aqueduct ?, which is often associated with Y SLC26A4 ? mutations, and subjects with clinical features that suggested Y syndromic hearing loss ? were excluded. Subjects with Y auditory neuropathy ? were tested for Y OTOF ? mutations, which are associated with Y auditory neuropathy ?, and subjects with Y OTOF ? mutations were excluded. The Ethics Review Committees of the National Hospital Organization Tokyo Medical Center and all collaborating hospitals approved the study procedures. All procedures were conducted after written informed consent had been obtained from each subject or their parents.

To add an annotation, select the text in the above box and then click the corresponding button:

Disease Gene SNP/Indel Structural variation

showing passage 2 of document 1. PMID: 24164807 PMCID: 4231469

Figure 1. Examples of manual entity annotations A: Before annotation, B: After annotation.