variant2literature: full text literature search for genetic variants

Yin-Hung Lin[1,#], Yu-Chen Lu[1,#], Ting-Fu Chen[1], Jacob Shujui Hsu[2,3], Ko-Han Lee[1,4], Yi-Wei Cheng[1,4], Yi-Chieh Chen[1,5], Jhih-Sheng Fan[1], Chien-Ta Tu[5], Chen-Ming Hsu[6], Chih-Chen Chou[1], Pei-Lung Chen[5,7,8,*], Yi-Chin Ethan Tu[1,*], and Chien-Yu Chen[9,*]

[1]Taiwan AI Labs, Taipei, Taiwan; [2]Institute of Molecular and Genomic Medicine, National Health Research Institutes, Taiwan; [3]Center for Genomics Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China; [4]Department of Medicine, National Taiwan University College of Medicine, Taipei, Taiwan; [5]Graduate Institute of Medical Genomics and Proteomics, National Taiwan University College of Medicine, Taipei, Taiwan; [6]Department of Computer Science and Information Engineering, Chien Hsin University of Science and Technology; [7]Graduate Institute of Clinical Medicine, National Taiwan University College of Medicine, Taipei, Taiwan; [8]Department of Medical Genetics, National Taiwan University Hospital, Taipei, Taiwan; [9]Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei, Taiwan.

#The authors contributed equally to this work.

*To whom correspondence should be addressed.

Abstract

Motivation: Whole genome sequencing (WGS) by next-generation sequencing produces millions of variants for an individual. The retrieval of biomedical literature for such a large number of genetic variants remains challenging, because in many cases the variants are only present in tables as images, or in the supplementary documents of which the file formats are diverse.

Results: The proposed tool named variant2literature from the TaiGenomics (Toolkits for AI genomics) resolves the problem by incorporating text recognition with image processing. In addition to the adoption of advanced image-based text retrieval, the recall rate of finding the literature containing the variants of interest is further improved by employing the skill of variant normalization. Different variant presentations are transformed into chromosome coordinates (standard VCF format) such that false negatives can be largely avoided. variant2literature is available in two ways. First, a web-based interface is provided to search all the literature in PMC Open Access Subset. Second, the command-line executable can be downloaded such that the users are free to search all the files in a specified directory locally.

Availability: http://variant2literature.taigenomics.com/

Contact: chienyuchen@ntu.edu.tw

1 Introduction

Whole-exome sequencing (WES) and whole-genome sequencing (WGS) become more and more popular in diagnosis of diseases and discovery of phenotype-related variants

[1-3]. The annotation of variants involves a critical step that finds literature to explain the potential influence of a discovered variant on protein functions, cell behaviors, organ normality, etc. In this regard, many efforts have been made to improve the search performance [4, 5]. Even though, we observed that the sensitivity of variant queries still has space to increase. For example, when a researcher sends a query in the literature database, it is most likely that the related papers are missing in the result page owing to the fact that the search is not comprehensively performed on the full text of the papers as a well as all the supplementary files. The challenge of this task is that most accessible literature are released in PDF format. In this study, we used PubMed Central (PMC) Open Access Subset (PMC OA) to demonstrate that the adoption of image processing followed by named-entity recognition (NER) and variant normalization can largely increase the recall rate of literature search.

In addition to tackle the barrier due to diverse file formats, variant2literature also resolves the format issue of variant queries by translating all types of variant presentations into chromosome-based coordinates. The literature files were preprocessed to record all the variants using chromosome-based coordinates in a local database. Moreover, if a gene has multiple isoforms, a known variant of one particular isoform will be also annotated on the other isoforms with corrected coordinates. Meanwhile, RSID (Reference SNP cluster ID) is also acceptable. All of these designs were shown to improve the sensitivity of variant query on literature search. Currently, variant2literature accepts documents files in XML, PDF, DOCX, DOC, XLSX, and CSV formats. Both web interface and command-line executable are available now.

2 Materials and methods

When handling PDF files, fasterRCNN [6] was used to detect table regions. Figure and table captions in PDF were retrieved by PDFFigures 2.0 [7]. variant2lieterature used tmVar [8] to extract variants from paragraphs and tables, and used GNormplus [9] to extract genes from paragraphs and tables. It is important that variant representations were made consistent, in-cluding the variants in the literature and the user queries. For example, ‘GJB2, c.109G>A’, ‘GJB2, p.Val37Ile’, ‘GJB2, V37I’, and ‘rs72474224’ all represent the same variant. Moreover, ‘TNNT2, R141W’, ‘TNNT2, R134W’, and ‘TNNT2, R143W’ might present the same variant owing to position shifts on different isoforms of TNNT2. After normalization, the variants and the associated genes were recorded in VCF format [10]. Accordingly, the literature files were indexed by the normalized variants.

variant2literature is available in two ways: web interface and command-line executable. In the web server, query can be made in different ways. Some examples can be found on the web site (http://variant2literature.taigenomics.com/). Meanwhile, uploading a VCF file is acceptable in the web interface.

3 Results and discussions

The disease 'deafness' is used as an example to demonstrate the power of variant2literature. The Deafness Variation Database (DVD) [11] contains 876,135 known variants (v8.1.2017-11-08). There are 10,036 variants being linked to 3,095 papers, constituting 18,410 variant-paper pairs. Among the 3,095 papers, 206 papers can be found in the PMC OA database, containing 962 variant-paper pairs, involving 874 variants. It is noticed that not all the variant-paper relationships for these 874 variants were curated in the DVD database. In this regard, three domain experts were recruited in this study to identify the missing variant-paper pairs present in the 206 papers for these 874 variants. Finally, 996 pairs were recorded as true relationships.

After invoking variant2literature on 874 variants against the 206 pa-pers, 1,016 pairs were reported as positives, resulting a recall rate of 98.39% and a precision rate of 96.46%. For comparison, we performed the same query of variant set on LitVar [4]. By uploading the 874 variants to LitVar, 672 pairs were reported as positives, resulting a recall rate of 40.15% and a precision rate of 96.43%.

Table 1. Performance of variant2literature with different settings.

Methods	Variant-paper pairs predicted	Recall	Precision	F1 score
variant2literature (without suppl.)	444	43.47%	97.52%	0.60
variant2literature (with suppl.)	831	80.42%	96.39%	0.88
variant2literature (with suppl. + table text retrieval)	1,016	98.39%	96.46%	0.97
LitVar	672	40.15%	96.43%	0.57

Abbreviation: suppl. stands for supplementary files.

3.1 Limitations

The following limitations of variant2literature should be noticed. First, when a user specifies a query like this: 'Gene_name, A123C', variant2literature cannot tell whether the position 123 is the coordinate of bases on cDNA or residues on proteins. Currently, variant2literautre assumes the number is regarding residue coordinates. Second, variant2literature considers all potential isoform coordinates of a gene when searching variants in literature, this might result in some false positives. Such limitations will be gradually resolved according to user feedback in the future.

Acknowledgements

The authors would like to thank Ministry of Science and Technology (MOST), R. O. C., for the financial support under the contract: MOST 105-2221-E-002-129-MY3. The funder had no role in the design, collection, analysis, or interpretation of the data; writing the manuscript; or the decision to submit the manuscript for publication.

Conflict of Interest: none declared.

References

1. Yang, Y., et al., *Clinical whole-exome sequencing for the diagnosis of mendelian disorders.* The New England Journal of Medicine, 2013. **369**(16): p. 1502-1511.
2. Martin, H.C., et al., *Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis.* Human Molecular Genetics, 2014. **23**(12): p. 3200-3211.
3. Stavropoulos, D.J., et al., *Whole-genome sequencing expands diagnostic utility and improves clinical management in paediatric medicine.* Genomic Medicine, 2016. **1**: p. 15012.
4. Allot, A., et al., *LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC.* Nucleic Acids Research 2018. **46**(W1): p. W530–W536.
5. Cejuela, J.M., et al., *nala: text mining natural language mutation mentions.* Bioinformatics, 2017. **33**(12): p. 1852-1858.
6. Ren, S., et al. *Faster r-cnn: Towards real-time object detection with region proposal networks*. in *Advances in Neural Information Processing Systems (NIPS)*. 2015.
7. Clark, C. and S. Divvala. *Pdffigures 2.0: Mining figures from research papers*. in *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. 2016. IEEE.
8. Wei, C.-H., et al., *tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine.* Bioinformatics, 2017. **34**(1): p. 80-87.
9. Wei, C.-H., H.-Y. Kao, and Z.J.B.r.i. Lu, *GNormPlus: an integrative approach for tagging genes, gene families, and protein domains.* BioMed Research International, 2015. **2015**.
10. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-2158.
11. Azaiez, H., et al., *Genomic Landscape and Mutational Signatures of Deafness-Associated Genes.* The American Journal of Human Genetics, 2018. **103**(4): p. 484-497.