Panel: Innovation on Digital Curation

Text Mining for Improving the Prioritization, Curation, and Integration of Knowledge for Clinically Relevant Variants

Zhiyong Lu, NCBI, NLM, NIH

Abstract

Understanding the associations of genomic variants with diseases and conditions and assessing their clinical significance is critical for genomic research precision medicine. Despite significant efforts in expert curation, information about most of the 154 million dbSNP Build 149 reference variants (RS) remains unknown. On the contrary, a wealth of human knowledge about the variant biological function and disease impact is buried in unstructured literature data. Previous studies have attempted to harvest and unlock such information with text-mining techniques but are of limited use in practice.

I will first present a new text-mining method (tmVar 2.0) for extracting variant mentions in the literature and subsequently normalizing them to standardized database identifiers, followed by a large-scale analysis of text-mined results vs. curated data from existing databases [1].

Next, through several real-world use cases (e.g. [2, 3]), I will demonstrate that our approach can identify high impact variants from publications and that our results can be combined with existing data to prioritize and rank the variants by various attributes (e.g. functional consequence and allele frequency). I will conclude by summarizing the opportunities and challenges of using text mining for the manual curation and interpretation of variation effects on biological functions and diseases to enrich our current knowledge.

References:

[1]: Wei C-H, Phan L, Feltz J, Maiti R, Hefferon T, Lu Z. tmVar 2.0: Integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. Bioinformatics (Oxford, England). btx541. https://doi.org/10.1093/bioinformatics/btx541

[2]: Poux S, Arighi CN, Magrane M, Bateman A, Wei CH, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, UniProt Consortium. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. Bioinformatics. 2017 Jul 13:btx439.

[3]: Singhal A, Simmons M, Lu Z. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. PLoS computational biology. 2016;12(11):e1005017. doi:10.1371/journal.pcbi.1005017.