Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220603

Designing an Optimal Expansion Method to Improve the Recall of a Genomic Variant Curation-Support Service

Anaïs MOTTAZ[a,b,1], Emilie PASCHE[a,b], Pierre-André MICHEL[a,b], Luc MOTTIN[a,b], Douglas TEODORO[a,b,c] and Patrick RUCH[a,b]
[a] *HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland*
[b] *SIB Swiss Institute of Bioinformatics, Geneva, Switzerland*
[c] *Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland*

Abstract. The importance of genomic data for health is rapidly growing but accessing and gathering information about variants from different sources is hindered by highly heterogeneous representations of variants, as outlined by clinical associations (AMP/ASCO/CAP) in their recommendations. To enable a smooth and effective retrieval of variant-containing documents from different resources, we developed a tool (https://goldorak.hesge.ch/synvar/) that generates for any given SNP – including variant not present in existing databases – its corresponding description at the genome, transcript and protein levels. It provides variant descriptions in the HGVS format as well as in many non-standard formats found in the literature along with database identifiers. We present the SynVar service and evaluate its impact on the recall of a genomic variant curation-support service. Using SynVar to search variants in the literature enables to increase the recall by +133.8% without a strong impact on precision (i.e. 93%).

Keywords. genomic variant, biomedical literature, precision medicine

1. Introduction

In the search for accelerating medical discoveries and improving health, genomic data is gaining increasing importance. Cohorts of human genomes are getting larger and cancer sequencing has become routine. Such data facilitate the study of the underlying causes of diseases, help refining care planning and treatment efficacy. But accessing and gathering genomic data and variants information from different sources are highly challenging.

Genetic variants correspond to a change from a template sequence and can be described using various reference sequences and at different levels (genome, transcript, protein). A given variant can thus correspond to several descriptions, even when using nomenclature standards as described by the HGVS society [1]. The combinatorial nature of the different description levels (many-to-many relationships), due to gene overlapping, isoforms and genetic code redundancy, hinder a linear mapping between

[1] Corresponding Author, Anaïs Mottaz, Campus Battelle, Bâtiment B, Rte de la Tambourine 17, 1227 Carouge, Switzerland; E-mail: anais.mottaz@hesge.ch.

them. While many databases of polymorphisms and variants exist, such as ClinVar, COSMIC or dbSNP, using those resources as terminologies is fairly challenging since they do not identify variants at the same level. Besides, although guidelines from clinical associations (AMP/ASCO/CAP) recommend unambiguous variants naming for interpretation and reporting [2], resources such as the biomedical literature tend to use heterogenous and non-standard ways of representing variants like missing reference sequence leading to ambiguous position, or fancy syntaxes.

To enable a smooth and effective retrieval of documents containing variants, we developed a service to expand variant queries, SynVar. This service provides for a given SNP its corresponding description at the genome, transcript and protein levels, in the HGVS format as well as in many non-standard descriptions found in the literature (e.g. BRAFV600E). The SynVar service is being used for query expansion by Variomes (https://candy.hesge.ch/Variomes/), a high recall search engine to support the curation of genomic variants [3]. It is also used by CINECA, which develops a federated infrastructure for genetic data sharing, implementing GA4GH standards, and their Beacon models to query cohorts for genetic variants over many data nodes (https://www.cineca-project.eu). We present, here, the architecture of SynVar, along with an evaluation of its impact on the recall of variant searches in the literature and the frequency distribution of the different expansion patterns encountered in the literature.

2. Methods

2.1. SynVar architecture

The main steps of the SynVar service are presented below. Depending on the description level of the queried variant, the order and the steps may slightly differ.

Query processing. Variants can be queried at the level of the protein, transcript or genome, dbSNP or COSMIC identifier. Variant is extracted from the query using regular expressions. The variant does not need to be in a standard HGVS format. The reference gene, chromosome or sequence is also extracted using regular expressions from the *ref* parameter, if given, or the *variant* parameter. If it is provided within the *variant* parameter, no specific format or order is required (e.g. "Val600Glu, BRAF"). Gene name is validated through the neXtProt API [4]. The description level of the variant is optional and guessed from its format if not provided with the query.

Variant validation. For protein and transcript variants, reference sequence identifiers corresponding to the provided gene are retrieved using the neXtProt API, including isoforms if *iso* parameter is set to true. Protein or transcript fasta sequences are retrieved from neXtProt and NCBI E-utilities [5] respectively. Reference amino acid or base of the variant is then checked against the reference sequences at the variant position. For genomic variants, validation is done against assemblies GRCh37 and GRCh38 using *variantValidator* [6]. If a gene is provided instead of a chromosome, the validation is done on the corresponding chromosome only if the position lies within the given gene.

Translation/Backtranslation. Valid protein and transcript variants in HGVS format, as generated by the validation or conversion steps, are, respectively, backtranslated using *Mutalyzer Back Translator* and translated using *runMutalyzerLight* tools [7]. Several variants may be generated by backtranslation due to amino acid code redundancy.

Mapping/Conversion. Transcript variants are mapped to GRCh37 and GRCh38 assemblies and genomic variants are converted to transcript variants using *variantValidator* or, when not available, *Mutalyzer numberConversion* service.

Variant identifiers. DbSNP identifiers are retrieved based on the chromosome reference sequence and position using NCBI e-Utilities, after the mapping to the genomic build. If a dbSNP identifier is provided as a query, the different genomic variants corresponding to the identifier are collected using NCBI e-Utilities. For COSMIC, the mapping between transcript variants and COSMIC identifiers are retrieved from the downloaded *COSMIC Mutation data* [8] and used to provide COSMIC identifiers in the output and to enable query expansion from a COSMIC identifier.

Syntactic variations. For each level of variant description, a set of syntactic variations is provided. It represents the most common variant description formats as encountered in the literature, based on previously described patterns [9] and on a preliminary evaluation of variant recognition in literature.

Output. Results are returned as a list of genomic variants, along with their corresponding transcript and protein variants in HGVS and non-standard formats. The output can be in XML or JSON Beacon format.

2.2. Evaluation

To evaluate SynVar, we performed two experiments. Both experiments are based on a set of 766 variants in BRCA1 and BRCA2. This set was built using BRCAExchange [10] and corresponds to all missense SNPs for BRCA1 and BRCA2 from LOVD [11].

First, we evaluated the effect of variant expansion generated by SynVar on the recall of a set of variant searches. Literature in PubMed Central was searched for the 766 variants using the Variomes APIs [3]. Two searches were performed for each variant: the first used only the term mentioned in the list (e.g. M18T) while the second expanded the query with all variant descriptions suggested by SynVar (e.g. 53T>C).

Second, we investigated how the different description levels and syntactic synonyms suggested by SynVar were represented in the literature. To this extent, a set of 61 expansion patterns (Table 1) was defined to represent the SynVar synonyms: 18 patterns at the genome level, 18 patterns at the transcript level and 25 patterns at the protein level. The Variomes APIs were used to retrieve all variant occurrences in PubMed Central for the 766 variants. Each occurrence was then mapped to one of the 61 patterns.

Table 1. Example of expansion patterns generated for representing SynVar variant synonyms

Pattern	Level	Example of variant
\d+[ACGT]\s*>\s*[ACGT]	Genome	43124044A > G
\d+[ACGT]/[ACGT]	Transcript	53T/C
p\.[A-Z][a-z]{2}\d+[A-Z][a-z]{2}	Protein	p.Met18Thr

3. Results

3.1. Availability

The SynVar service is available as a SaaS via an Open API (https://goldorak.hesge.ch/synvar/). The GUI is mainly for demonstration and debugging purposes. Output is available in different formats, including Beacon format.

3.2. Impact of SynVar on variant search

Results of the impact of SynVar on the recall for variant search are presented in Table 2. Searching variants in literature using variant expansion resulted in an improvement of the recall by 133.8%. Indeed, while a query with a single term for representing the variant returned on average 3 documents, using all SynVar descriptions enabled retrieving on average 7 documents. It also significantly reduced the number of queries without any results: without SynVar, 255 queries returned no documents, while using SynVar enabled finding documents for 118 of these 255 queries. Among documents retrieved using SynVar, 93% were relevant as estimated by a manual analysis of a subset of 27 documents retrieved from five random queries.

Table 2. Results of the comparison of Variomes with and without SynVar

	Without SynVar	**With SynVar**
Mean number of documents retrieved per query	3	7
Total number of documents retrieved for all queries	2304	5387
Number of queries with no results	255	137

3.3. Evaluation of the SynVar variants' expansion patterns distribution

Out of the 766 queries, 137 queries resulted in no document. In addition, for 6 queries, our system failed to map the variant occurrence to one of the expansion patterns. We thus present the results using the remaining 623 queries. 7037 variant occurrences were identified and mapped to expansion patterns. On average, a variant was represented in the literature under 3.3 different patterns (min: 1; max: 8). 60.6% of variant occurrences corresponded to a variant at the protein level, 28.2% at the transcript level and 11.2% at the genome level (Figure 1). Among the protein patterns proposed by SynVar, three patterns dominated and represented respectively 60.9%, 28.5% and 10.1% of all protein patterns. One pattern was largely represented in the transcript patterns with 94.2% of the occurrences. At the genome level, it is the dbSNP identifier which was mostly impactful with 95.1% of the occurrences.

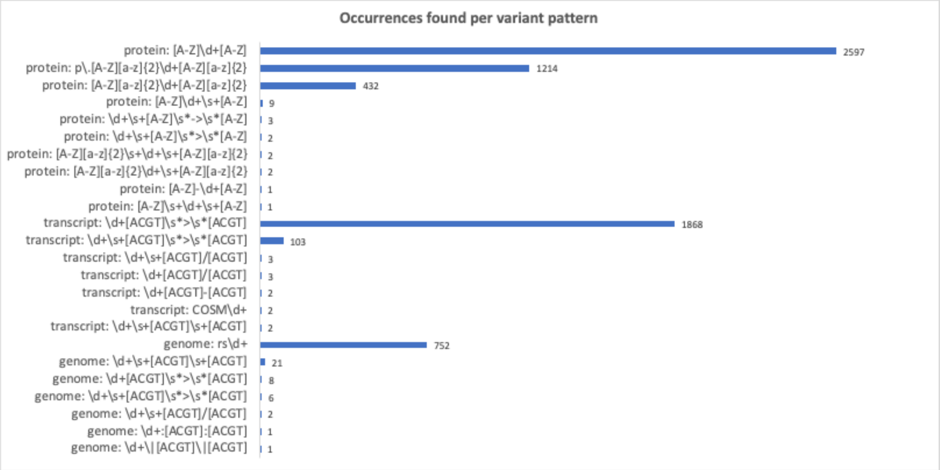

Figure 1. Number of occurrences found in the literature for each variant pattern returning matches

4. Discussion

We propose a service to facilitate the retrieval of variant-containing documents from heterogeneous resources. Its main advantage compared to existing tools, such as tmVar [12,13], is that it processes variants independently of a database, resulting in a much broader recall [3]. Indeed, using a database limits the coverage due to its specific purpose (polymorphism *vs* somatic variants database) and decreases the specificity when position-only based (like dbSNP). The impact of SynVar on literature search recall is important, as it more than doubles the number of retrieved documents without strongly altering precision (i.e. 93%). Searching for the classical pattern would allow to catch only 50% of the occurrences. While six patterns represent more than 95% of occurrences, even a few matches may be of importance for rare variants, which accounts for the vast majority of variants. In a future work, additional patterns as well as non-SNP variants will be considered along with the pre-indexing of documents to improve efficiency, using methods such as Named Entity Recognition [14] and requiring variant normalization.

Acknowledgements. This service has been supported by CINECA - UE H2020 Grant #825775 and the Canadian Institute of Health Research (CIHR) Grant #404896.

References

[1] den Dunnen JT. Describing Sequence Variants Using HGVS Nomenclature. Methods Mol Biol. 2017;1492:243-251.

[2] Li MM, Datto M, Duncavage EJ, Kulkarni S, Lindeman NI, Roy S *et al*. Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer: A Joint Consensus Recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists. J Mol Diagn. 2017 Jan;19(1):4-23.

[3] Pasche E, Mottaz A, Caucheteur D, Michel PA, Gobeill J and Ruch P. Variomes: a high recall search engine to support the curation of genomic variants. Bioinformatics.. 2022;btac146

[4] Zahn-Zabal M, Michel PA, Gateau A, Nikitin F, Schaeffer M, Audot E *et al*. The neXtProt knowledgebase in 2020: data, tools and usability improvements. Nucleic Acids Res. 2020 Jan 8;48(D1):D328-D334.

[5] Sayers E. The E-utilities In-Depth: Parameters, Syntax and More. 2009 May 29 [Updated 2021 Apr 15]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2010-.

[6] Freeman PJ, Hart RK, Gretton LJ, Brookes AJ, Dalgleish R. VariantValidator: Accurate validation, mapping, and formatting of sequence variation descriptions. Hum Mutat. 2018 Jan;39(1):61-68.

[7] Lefter M, Vis JK, Vermaat M, den Dunnen JT, Taschner PEM, Laros JFJ. Next Generation HGVS Nomenclature Checker. Bioinformatics. 2021 Feb 4;37(18):2811–7.

[8] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N *et al*. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019 Jan 8;47(D1):D941-D947.

[9] Yip YL, Lachenal N, Pillet V, Veuthey AL. Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase. J Bioinform Comput Biol. 2007 Dec;5(6):1215-31.

[10] Cline MS, Liao RG, Parsons MT, Paten B, Alquaddoomi F, Antoniou A *et al*. BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. PLoS Genet. 2018 Dec 26;14(12):e1007752.

[11] Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat. 2011 May;32(5):557-63.

[12] Lee K, Wei CH, Lu Z. Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. Brief Bioinform. 2021 May 20;22(3):bbaa142.

[13] Wei CH, 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. 2018 Jan 1;34(1):80-87.

[14] Copara J, Naderi N, Knafou J, Ruch P, Teodoro D. Named entity recognition in chemical patents using ensemble of contextual language models. In Proceedings of the CLEF 2020 conference. 2020 Sep 22-25.