Bioinformatics, 38(9), 2022, 2595–2601
https://doi.org/10.1093/bioinformatics/btac146
Advance Access Publication Date: 11 March 2022
Original Paper

Data and text mining

Variomes: a high recall search engine to support the curation of genomic variants

Emilie Pasche [1,2,*], Anaïs Mottaz [1,2], Déborah Caucheteur [1,2], Julien Gobeill [1,2], Pierre-André Michel[1,2] and Patrick Ruch [1,2]

[1]SIB Text Mining Group, Swiss Institute of Bioinformatics, 1206 Geneva, Switzerland and [2]BiTeM Group, Information Sciences, HES-SO/HEG, 1227 Carouge, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on September 30, 2021; revised on February 7, 2022; editorial decision on March 6, 2022; accepted on March 10, 2022

Abstract

Motivation: Identification and interpretation of clinically actionable variants is a critical bottleneck. Searching for evidence in the literature is mandatory according to ASCO/AMP/CAP practice guidelines; however, it is both labor-intensive and error-prone. We developed a system to perform triage of publications relevant to support an evidence-based decision. The system is also able to prioritize variants. Our system searches within pre-annotated collections such as MEDLINE and PubMed Central.

Results: We assess the search effectiveness of the system using three different experimental settings: literature triage; variant prioritization and comparison of Variomes with LitVar. Almost two-thirds of the publications returned in the top-5 are relevant for clinical decision-support. Our approach enabled identifying 81.8% of clinically actionable variants in the top-3. Variomes retrieves on average +21.3% more articles than LitVar and returns the same number of results or more results than LitVar for 90% of the queries when tested on a set of 803 queries; thus, establishing a new baseline for searching the literature about variants.

Availability and implementation: Variomes is publicly available at https://candy.hesge.ch/Variomes. Source code is freely available at https://github.com/variomes/sibtm-variomes. SynVar is publicly available at https://goldorak.hesge.ch/synvar.

Contact: emilie.pasche@hesge.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advances in personalized medicine make it now possible to select a treatment targeting specific tumor variants. Indeed, based on the tumor's molecular profile coupled with clinical information such as the diagnosis, it is possible to better determine which treatment resulting in a likely favorable response can be proposed. To this extent, a tissue sample is sequenced, resulting in the identification of hundreds of variants. The clinical experts, assisted by bioinformatics tools, are then in charge of determining which variants are actionable, i.e. are likely to result in a better or worse prognosis and derived treatment response. It results in the generation of a tumor board report which can be used by physicians for the treatment of the patient.

The identification and interpretation of clinically actionable variants is a critical bottleneck. Indeed, it is necessary to look for evidence in genomic variant knowledgebases such as OncoKB, COSMIC, CIViC, as well as in the literature. These high-quality resources rely on manual curation. However, since manual curation is not scalable, information on curated variants is sometimes incomplete or out-of-date. Moreover, their coverage is very diverse. Lee *et al.* (2021) reported that only eight variants overlapped over six well-known cancer genomic variant databases. Scientific literature is thus an indispensable source of content. However, screening out scholarly publications poses a number of challenges. Curators must cope with large and increasing volumes of publications: Lee *et al.* (2021) showed that among the last 5 years 50 000 publications containing genomic variants were published every year. Moreover, the information is 'hidden' in unstructured text in which the genetic variants, but also the other relevant entities (e.g. genes, diseases, demographic information, etc.), are labeled in diverse forms such as synonyms and abbreviations. While a few variants received massive attention, most of them—the so-called Variant or Unknown/Uncertain Significance—are described in a handful of published reports. For instance, only 13 MEDLINE abstracts mention the genomic variant P871L of the Breast Cancer 1 gene and gene product.

However, there are 37 additional abstracts of potential interest in which the variant is named with alternative forms, such as p.Pro871Leu or c.2612C>T. By using PubMed-like search engines, it is therefore necessary to multiply the queries to avoid missing an important publication. In addition, when the number of publications is large (e.g. for highly studied variants), the triage of the literature to select the most relevant documents is a tedious task.

In recent years, search engines dedicated to variant-related tasks have aroused the interest of the Natural Language Processing (NLP) community. In particular, NLP competitions have accelerated the investigation of more sophisticated approaches. Since 2017, TREC Precision Medicine track (Roberts *et al.*, 2018, 2019) has been proposing a task aiming at finding relevant publications given a particular case containing a genetic variant, a disease and demographic information. The best resulting systems were able to retrieve almost two-thirds of relevant publications in their top-10 results.

When developing a search engine to retrieve literature for genomic variants, four aspects should be considered: the literature collections, the normalization of variant names, the type of search algorithms and the literature triage. Regarding the literature collections, MEDLINE proposes a corpus of more than 30 million citations, out of which about 800 000 are related to mutations (Jiang *et al.*, 2019). While abstracts can be considered as sufficient for variant prioritization, full-text access is needed to support the clinical interpretation of variants. Indeed, information related to genomic variants are often mentioned in the body of the scientific reports, including results and tables. According to Jimeno Yepes and Verspoor (2014) a significant subset of the information related to genomic variants is even reported in Supplementary Data. The normalization of variants' names into a single form is an essential step. Indeed, variants can be represented in a multitude of standard formats (e.g. amino acids can be represented using one-letter codes or three-letters codes) or even by using non-standard expressions as described by Yip *et al.* (2007). Lee *et al.* (2021) reported that 76% of the variants mentioned in the literature do not follow the standard Human Genome Variation Society (HGVS) nomenclature. Thus, the use of variant-specific name entity recognition tools is essential.

Various variant-specific search engines have been developed in recent years. We present here a brief overview. LitVar (Allot *et al.*, 2018), one of the most cited ones, uses both abstracts and full-texts. Using information matching, it returns a chronologically ordered set of publications. variant2literature (Lin *et al.*, 2019) not only uses full-text articles but also processes Supplementary Data, in particular tables represented as images in PDF. Like LitVar, it returns literature in a chronological order. Overcoming this chronological ordering limitation, VIST (Ševa *et al.*, 2019) uses a support vector machine model to rank publications by relevance. However, this tool does not cope with full-text articles: only abstracts are used. Nevertheless, it is one of the rare tools including searching for clinical trials. LitVar, variant2literature and VIST all rely on the use of tmVar (Wei *et al.*, 2018) for the recognition and normalization of variant names in both publications and users' queries. Other types of approaches focus on the literature triage, such as LitGen (Nie *et al.*, 2020), which collects publications returned by LitVar and filters them by types of evidence. Further, Lv *et al.* (2020) propose a method to classify papers as relevant or not to support the curation of genomics databases instead of trying to search for variant occurrences. It is based on a Knowledge-enhanced Multi-channel CNN model to identify relevant publications both in abstracts and full-texts.

In this context, we designed Variomes (Caucheteur *et al.*, 2020), an original application to support the search of human variants. The system can be used as a literature triage system in the same way as LitVar. It can also be used to prioritize variants (e.g. from a Variant Calling File) to facilitate the identification of clinically actionable variants. When the system is used to rank variants, it assumes that the clinical interest of a variant is associated with the volume of literature published about the variant. The ranking of variants consists of establishing a score for each variant by summing the scores—more precisely the Retrieval Status Value as returned by the search engine, see Ehrler *et al.* (2005) for more information—of publications retrieved for each variant. On the contrary, a variant without clinical significance (e.g. silent mutation) will result in no or very few citations.

Our tool aims to facilitate the annotation of the variants by curators by suggesting them a set of publications of interest. Variomes enables searching the biomedical literature. The collections are preprocessed with a set of medical terminologies (Gobeill *et al.*, 2020). At query time, user queries are automatically processed to map keywords to the terminologies and expand genetic variants using a dedicated variant expansion system (Caucheteur *et al.*, 2020). Finally, different strategies are investigated to maximize the performance of the literature triage based on the tuning of an optimal ranking function (Caucheteur *et al.*, 2019; Pasche *et al.*, 2018). Variomes is available through a user-friendly interface as well as through a set of application programming interfaces (APIs). Finally, it is also integrated within the SVIP curation platform (Stekhoven *et al.*, 2018), a national Swiss repository for clinically verified variant annotations in oncology.

2 Materials and methods

Our approach is based on the use of three collections of scientific literature: abstracts from MEDLINE, full-text articles from PubMed Central and clinical trials from ClinicalTrials.gov.

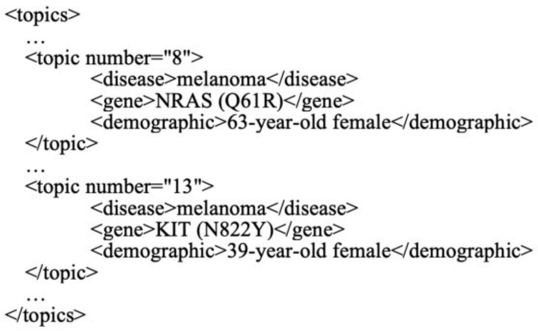
2.1 System's architecture

The collection is first normalized with a set of terminologies to ease the matching of user's information requests. The following terminologies were selected: neXtProt (Gaudet *et al.*, 2017) for genes, NCI Thesaurus (Sioutos *et al.*, 2007) for diseases and DrugBank (Wishart *et al.*, 2018) for drugs. Pre-processing the collections enables first to retrieve results faster thanks to pre-computed indexes and second to increase the recall. Indeed, querying the collections through the annotations permits retrieving not only the exact term, but also its synonyms as well as string variations. For instance, while querying the MEDLINE collection using the keyword *BRAF* returns 15 907 hits, querying the annotations using *NX_P15056* (i.e. the unique neXtProt identifier corresponding to the *BRAF* gene) returns 17 191 documents, thus increasing the recall by almost +8%. Indeed, the annotations recognized not only *BRAF* and its official synonyms *BRAF1* and *RAFB1* but also syntactic variations such as *B-RAF*.

Documents and annotations are loaded into a MongoDB document database and indexed into an ElasticSearch index. Our system uses the ElasticSearch index for querying, while the MongoDB database serves for annotations-based re-ranking as well as for enriching the display of the documents. At the time of writing (November 16, 2021), the MEDLINE collection consists of 33 289 693 citations, the PubMed Central collection consists of 4 416 483 full-texts, and the ClinialTrials.gov collection consists of 395 229 clinical trials. The abstracts and full-text collections are managed by the SIB Literature Services (SIBiLS) (Gobeill *et al.*, 2020) with daily updates, while the clinical trials collection is updated every quarter.

The search engine is based on a two-steps system. The first step focuses on recall as it aims at gathering a large set of documents related to a particular case. The second step focuses on precision as it attempts to properly rank the set of documents.

Each variant query can be represented as a triplet: a variant in a gene for a given diagnosis. To collect the most comprehensive set of abstracts, an Elasticsearch query is generated. This query is composed of three 'must' clauses (i.e. one clause for each entity of the triplet) that must appear in the matching documents. For the disease and the gene clauses, two 'should' sub-clauses are defined and at least one of the clauses must appear in the matching documents: the exact query term is searched in the publications text and the corresponding unique identifier is searched in the annotations. For the variant clause, a set of 'should' sub-clauses are generated: one for the exact query term and one for each of its synonyms generated by the SynVar service. SynVar is a synonym generator for single nucleotide polymorphisms (Caucheteur *et al.*, 2020). It provides



```
<topics>
  ...
  <topic number="8">
          <disease>melanoma</disease>
          <gene>NRAS (Q61R)</gene>
          <demographic>63-year-old female</demographic>
  </topic>
  ...
  <topic number="13">
          <disease>melanoma</disease>
          <gene>KIT (N822Y)</gene>
          <demographic>39-year-old female</demographic>
  </topic>
  ...
</topics>
```

Fig. 1. Example of TREC PM 2018 topics

descriptions of the variant at other levels (e.g. genomic level), as well as syntactic variations encountered in the literature. It proposes up to 50 synonyms for a variant, with descriptors at the protein, transcript and genomic levels.

However, the triplet-based query is sometimes too specific and documents not strictly targeting a given triplet might still be valuable, so that a constraint relaxing strategy is needed. For instance, a document about the given variant in the given gene but for another diagnosis—e.g. melanoma instead of breast cancer—may still be valuable from a clinical point of view. Moreover, while full-text articles reporting on treatments usually mention all the information regarding the disease, gene and variant, it is not always the case with abstracts. Thus, our system also collects documents with decreasing levels of specificity. Three additional queries are generated, each omitting one of the entities of the triplet. The respective outputs of these additional queries are linearly combined (Belkin *et al.*, 1995; Fox and Shaw, 1994), which has proven highly effective when combining results returned by different search methods (Aslam and Montague, 2001; Lee, 1997; Savoy, 2004).

A final score is calculated for each publication based on a linear combination of the score computed on each index. Each index score corresponds to a ranking strategy: (i) the Retrieval Status Value (Ehrler *et al.*, 2005) provided by ElasticSearch, (ii) a score based on the constraint relaxing strategy described above, (iii) a score based on the density of some specific named-entities in the document, (iv) a score based on the demographic concordance if demographic information is available in the query and (v) a score based on the density of some predefined keywords. More details for each strategy are available below. All weights are determined by direct search using TREC benchmarks. The optimal settings are based on a staged strategy: we maximize the R-Prec (R-Precision). However, when two runs share the same R-Prec, we maximize the P5 (Precision at rank 5) and finally infNDCG (inferred Non-Discounted Cumulative Gain). R-Prec returns the number of relevant documents returned in the top-R documents, where R corresponds to the number of relevant documents for the query. P5 represents the proportion of relevant documents retrieved in the top-5 results. Finally, infNDCG reflects the gain brought by a document based on its position in the ranked results. Finally, the scores of documents published in languages other than English are downgraded to be ranked after English documents.

The relevance score provided by ElasticSearch is calculated using BM25 (Robertson and Zaragoza, 2009), which is a strong baseline for retrieval effectiveness (Thakur *et al.*, 2021). The score for the density of some specific named-entities is based on the number of occurrences of gene descriptors, disease descriptors and drug descriptors. To accelerate the scoring of the retrieved documents at query time, the collection is pre-annotated with a large set of named-entity types. The score for the demographic concordance is based on age-groups and genders extracted from the Medical Subject Headings (MeSH) terms associated with each MEDLINE record. Currently, this approach is applied only on the MEDLINE collection because it is available via the MeSH terms. In PMC, such information is not available. When an abstract is targeting the required gender and/or age-group, a positive boost is applied. The score based on predefined keywords aims to classify a document as being related to precision medicine or not. A list of positive stemmed keywords (e.g. *treat*) and negative stemmed keywords (e.g. *marker*) has been manually defined through a manual screening of a subset of documents. The presence of positive keywords in a document will improve the scoring, while the negative keywords occurrences will decrease its relevance.

2.2 Experimental evaluation setting

We perform three types of evaluation of the services: (i) as a literature triage system to support the curation of a given variant using different standard metrics, which balance both recall and precision; (ii) as a variant triage system using a VCF as input with focus on precision; (iii) a comparison between our system and LitVar with focus on recall.

In the absence of benchmarks with VCF queries, the system is initially tuned to perform a literature triage task; it means that most queries contain only a single variant. The literature triage task is evaluated using TREC Precision Medicine benchmarks. TREC Precision Medicine track aimed at retrieving MEDLINE abstracts providing relevant information to clinicians treating patients with cancer. The benchmark consists of semi-structured synthetic cases created by precision oncologists at the University of Texas MD Anderson Cancer Center. Each query mentions a disease, one or several mutated gene(s) and, optionally, some demographic information. About half of the topics contain a gene variant. A sample of TREC queries is shown in Figure 1. The tuning was performed using the TREC PM 2018 benchmark (Roberts *et al.*, 2018), composed of 50 topics, while the evaluation is done using the TREC PM 2019 benchmark (Roberts *et al.*, 2019), composed of 40 topics. In addition, we also perform an experiment to assess the use of synonyms of variants for literature triage. To this extent three runs are performed: querying the system with all the variant synonyms, including gene to protein translation (default settings as described above), querying the system with strict protein synonyms (i.e. *V600E* is expanded to *Val600Glu* but not to *1799 T>A*) and finally querying the system with no synonym.

Further, the variant prioritization is evaluated using a dataset of 756 variants, originating from eight patients, a partial set of the SwissMTB study (Singer *et al.*, 2018). For each patient (i.e. a case), called variants and manually curated tumor board reports are evaluated. A case consists of a diagnosis, a set of variants containing on average 94.5 (16–439) single nucleotide variants (SNV) and optionally a gender and/or an age. For each topic, the VCF has been pre-processed to generate a set of queries. Each query corresponds to a different SNV, as available in the VCF, and has the following content: a gene, a variant, a diagnosis and when available the demographic information. Only single nucleotides at the protein level are selected. The objective is to rank the queries to return on top the variants judged as relevant based on the tumor board reports. Five cases contain one relevant SNV and three cases contain two relevant SNVs. The settings defined for the literature triage task are used. For each variant, a score is calculated based on the sum of the scores of each article retrieved for the variant.

Furthermore, a comparison between LitVar and Variomes is also performed to evaluate the recall of our system. A set of 803 queries (see Supplementary Data) containing variants in BRCA1 and BRCA2, originating from BRCAExchange (Cline *et al.*, 2018) is used. These queries correspond to all coding SNPs for BRCA1 and BRCA2 from the LOVD dataset (Fokkema *et al.* 2011). Queries are sent to both systems and an automatic comparison of the returned documents is performed. For Variomes, the settings defined for the literature triage task are used. However, since the queries contain only gene and variant, no constraint relaxing is performed. Thus, the score attributed to constraint relaxing is set to 0. In addition to the quantitative comparison, we also perform a qualitative analysis on a random subset of queries. The manual analysis of the results aims to identify what features could explain such differences.

3 Results and discussion

3.1 Tuning of the system

The tuning of the system is based on five steps: (i) the scoring of the constraint relaxing strategy, (ii) the scoring of the density of named-

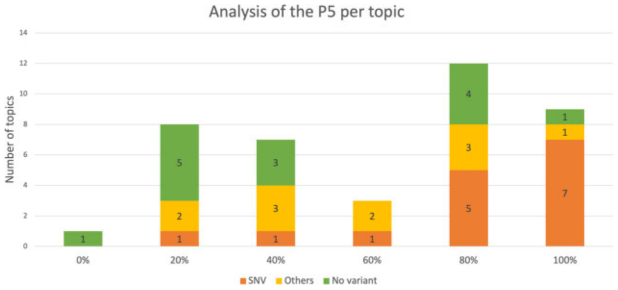


Fig. 2. Distribution of topics (number and type) according to the P5 values

entity types, (iii) the scoring of the demographic concordance, (iv) the scoring of the predefined keywords and (v) the linear combination of all strategies. We present here the best settings selected for each step. The final equation is available in Supplementary Data.

The constraint relaxing strategy was tested by attributing a weight between 0.0 and 1.0 to each of the three relaxed queries. The best results were obtained when using a weight of 0.95 for the query containing the disease and the gene, a weight of 0.07 for the query containing the disease and the variant and a weight of 0.05 for the query containing the gene and the variant.

The scoring of the density of named-entity types was tested by attributing a weight between 0.0 and 1.0 to the following named-entity types: gene, disease and drug. The best results were obtained when using a weight 0.97 for the disease, 0.51 for the gene and 0.57 for the drug.

The scoring of the demographic concordance was tested by attributing a weight to the age score and the gender score, between 0.0 and 1.0. The age and gender scores are calculated by attributing a strong bonus to documents matching the requested age and gender (bonus between 0.7 and 1.0) and a moderate bonus to documents not discussing the age or gender (bonus between 0.1 and 0.4). The best results were obtained when using a weight of 0.7 for the age and 0.5 for the gender, as well as a bonus of 0.7 for documents matching the age, 0.4 for documents not mentioning the age, a bonus of 0.7 for documents matching the gender and 0.4 for documents not mentioning the gender.

The scoring of the predefined keywords consisted first to define a list of positive and negative stemmed keywords. Two lists were tested for each modality: *treat*; *drug*; *therap*; *prognos*; *surviv* and *treat*; *drug*; *therap* for positive keywords; *immuno*; *marker*; *detect*; *sequencing* and *immuno*; *marker*; *detect* for negative keywords. A bonus between 0.0 and 1.0 was attributed to each occurrence of a positive keyword, while a penalty between 0.0 and -1.0 was attributed to each occurrence of a negative keyword. The best results were obtained when using the following keywords lists for positive and negative keywords: *treat; drug*; *therap*; *prognos; surviv* and *immuno*; *marker*; *detect*. The best settings occurred when using a bonus of 0.2 for the positive keywords and a penalty of -0.1 for the negative keywords.

Finally, the weight attributed to each of these strategies was defined by testing weights between 0.0 and 1.0. The ElasticSearch Retrieval Status Value was granted a weight of 1.0. The best results were obtained when giving a weight of 0.65 to the constraint relaxing, 0.1 to the named-entity types density, 0.05 to the demographic concordance and 0.1 to the predefined keywords.

3.2 Experimental setting 1: literature triage

The settings defined in the previous section were used to evaluate the literature triage task. The system resulted in a R-Prec of 32.5%, an infNDCG of 49.8% and a P5 of 62%, which means that almost two thirds of the top-5 returned abstracts are judged relevant. One third of the relevant documents are retrieved in the top-R documents. The official results of TREC 2019 are available in Roberts *et al.* (2019), where our methods ranked in the top 3 out of 14 participants. The top competitors adopted similar methods (Faessler *et al.*, 2019) but did not use powerful variant expansion methods likely to associate a transcript variant (e.g. 182A>G) to a protein variant (e.g. Q61R).

Table 1. Comparison of the effect of variants' synonyms on the results of literature triage

	P5	P10	R-Prec	infNDCG
Run with all variants' synonyms	62%	55.8%	32.5%	49.8%
Run with no variants' synonym	63.5%	57.5%	32.4%	49.6%
Run with basic variants' synonyms	64%	56.5%	32.7%	50.1%
TREC PM best run (Faessler *et al.*, 2019)	—	65.3%	35.7%	57.8%
TREC PM Variomes (Caucheteur *et al.*, 2019)	—	62.8%	31.7%	53.4%

An analysis of the P5 topic per topic is presented in Figure 2. The system performs the best with queries related to SNV. Indeed, 80% of the SNV queries resulted in a P5 equal or greater to 80%. However, queries related to gene fusion (classified as others in Fig. 2) do not perform well, with a P5 between 20% and 60%.

Our system is composed of two steps: collecting a set of abstracts and re-ranking these abstracts. While the second step is important to prioritize the literature, the first step is mandatory: the relevant abstracts must be broadly captured in our set to be properly ranked. We thus investigate which proportions of relevant abstracts have been successfully retrieved by our system, at whatever ranks they were returned. However, for queries resulting in a large amount of documents, only the top-1000 documents are retrieved. For more than half of the topics (22/40), at least 70% of the relevant documents are retrieved in our abstracts set. For such topics, the focus should be put on improving the ranking. However, for twelve topics, less than half of the relevant documents are retrieved. Further investigations are needed to analyze such abstracts and define possible actions to improve their gathering, such as improving the annotations of genes or diseases, expanding the diseases to parent and/or children diseases, defining better synonym lists for genetic variants, etc.

The results of the three runs to assess the effect of variant's synonyms are presented in Table 1. We observed that using all the synonyms proposed by the SynVar service surprisingly decreased the precision by −2.3%. Indeed, we obtained a P5 of 63.5% when no synonym was used to query ElasticSearch. However, using only the basic synonyms (i.e. those obtained without any mapping) resulted in the best P5 (64%). It appeared from discussions with the TREC PM organizers that complex synonyms of variants were not taken into account by the assessors. This is a limitation of the TREC benchmark since publications retrieved using complex synonyms are of equal clinical significance for the curation of variants.

3.3 Experimental setting 2: variants prioritization

The variant prioritization resulted in a P5 of 25%, which means that a quarter of the variants returned in the top-5 are judged as clinically actionable according to the tumor board. Because the relevance judgments only contain one or two relevant variants per topic, the R-Prec is a more appropriate metric: 71.4% of the variants returned in the top-R results are relevant. Out of the eleven variants reported in the tumor board, 81.8% of the variants were returned in the top-3: five were returned at rank #1, three at rank #2 and one at rank #3. For two variants, no literature was found; thus, suggesting that recall remains the main challenge for such a task.

3.4 Experimental setting 3: comparison with LitVar

In this section, we perform a comparison of Variomes and LitVar. LitVar is used as a baseline system to evaluate the performance of

Table 2. Comparison of the effect of variants' synonyms on the results of literature triage

	Union	LitVar	Variomes	Relative gain of Variomes compared to LitVar
1. Average number of documents retrieved per query	8.9	6.1	7.4	+21.3%
2. Average percentage of the content retrieved per query	100%	58.6%	90.8%	+54.9%
3. Total number of documents retrieved	7110	4896	5915	+20.8%
4. Number of queries with more results than the other system	—	80	462	+477.5%
5. Number of queries with no result	139	253	144	−43.1%

Table 3. Random sample of selected queries for the manual comparison of Variomes and LitVar.

	Results by Variomes	Results by LitVar	Total results	Only in Variomes	Only in LitVar
BRCA1 (L246V)	10	8	10	2	0
BRCA1 (I1044V)	2	1	2	1	0
BRCA2 (V3091I)	4	2	4	2	0
BRCA2 (G2813E)	3	1	3	2	0
BRCA1 (S1497A)	5	0	5	5	0
BRCA2 (R3052W)	27	16	27	11	0
BRCA2 (R2842L)	3	5	6	1	3
BRCA1 (D67E)	5	4	7	3	2
BRCA1 (A1830T)	2	3	3	0	1
BRCA2 (Q3066E)	1	0	1	1	0

Variomes. We first provide a direct comparison of the two systems on a reference benchmark. The reference benchmark is automatically generated based on the results retrieved by the two systems. Further, we perform an error analysis based on a subsample of articles.

3.4.1 Comparison of tools

Quantitative results of the comparison of LitVar and Variomes are presented in Table 2: we design measures to tentatively contrast both the recall for queries and the retrieved articles (cf. metrics #1, #2, #3, #4) and the silence (metric #5) of Variomes. Results were obtained in February 2021.

Seven thousand hundred and ten documents were retrieved in total with 3701 documents (52%) in common between the two systems. LitVar retrieves on average 6.1 documents per query, while Variomes returns 7.4 documents, corresponding to a relative recall gain of +21.3% per query. Further, Variomes was able to retrieve on average 90.8% of the published content per query (i.e. results retrieved by LitVar + results retrieved by Variomes), while LitVar retrieved on average 58.6%. For 261 queries (32.5%), both systems returned the same number of documents. For 462 queries (57.5%), Variomes retrieved more documents than LitVar, while for 80 queries (10%), LitVar returned more documents. It means that Variomes returns the same number of results or more results for about 90% of the queries. Symmetrically the silence—as measured by counting the number of queries with no results—of Variomes is 43% lower. Furthermore, 139 queries (17.3%) returned no result whatever systems are used, which again suggests that recall remains a challenge for variant retrieval. We indeed observed that many curated articles did not mention the variant in the textual content of the article but rather in other files and in particular in Supplementary Data, which would require specific and sometimes relatively complex pre-processing steps such as table parsing (MS-Excel, CSV, …) or image analysis (e.g. PNG, TIFF, JPEG). With a silence of 17.9%, our results are clearly better but consistent with Jimeno Yepes and Verspoor (2014).

3.4.2 Error analysis

Ten queries were randomly selected to perform a query-by-query error analysis. These queries were manually analyzed (Table 3) in order to better understand where lies the respective power of Variomes and LitVar. For two queries, only Variomes returned results. For five queries, Variomes returned the total set retrieved by LitVar, as well as one to eleven additional documents. For the other three queries, Variomes missed one to three documents retrieved by LitVar, but for two of these queries, it also proposed one to three documents not retrieved by LitVar. In total, 68 documents were retrieved for the five queries, with 34 documents in common among both systems.

Variomes failed to retrieve six documents, which were retrieved by LitVar. In two of these documents, the variant is observed in the reference section, i.e. in the title of another publication. In such cases, the relevance of the citing article is doubtful. In the four other documents, the variant highlighted by LitVar is not the one mentioned in the query (e.g. *R2842H* instead of *R2842L*). Indeed, LitVar is using reference SNP identifiers (RSIDs), which combine variants occurring at the same locus. This is also the main cause observed for the two queries where Variomes retrieved fewer articles than LitVar (*BRCA1 R1443G* and *BRCA1 A1708V*). Thus, we can consider from this error analysis that no or very few relevant articles were really missed by our system compared to LitVar. The fraction of articles not found by Variomes (9.2%) seems therefore an experimental artifact and the error analysis suggests that Variomes' recall is about 70% higher than LitVar.

To evaluate the quality of documents retrieved solely by Variomes, we merged the 28 pairs of variants/documents from Table 3 together with 50 other randomly selected documents (see Supplementary Data). We reached a precision of 86%. Regarding the relevant documents specific to Variomes, in most cases, the variant was present, but with a different form (e.g. *9154C>T→G* for *R3052W*). We can thus deduct that the variant expansion services of Variomes (so-called SynVar and whose API can be accessed independently from Variomes) proposes a larger set of patterns for variant synonyms. Further, some of the retrieved documents were functional assays of human BRCA genes performed in various cells such as mouse stem cells, bacteria, yeast and human cancer cells. They may not have been considered as 'human' research by LitVar while they could clearly be of interest for some clinical research. Regarding the non-relevant documents, they had two origins. First, some variants were representing the same residue change at the same position but in a different gene. Second, some documents were retrieved through the occurrence of a RSID but were related to a different residue at the same position.

3.5 User interface

The Variomes service is publicly available at https://candy.hesge.ch/Variomes/. Users can either query the service with a single variant (or a combination of several variants) or they can upload a file containing a list of variants. A few parameters enable to personalize the search: specification of the timeline, addition of keywords to re-rank the documents, specification of facultative entities (e.g. disease), etc. The variants are then returned in a ranked table. Users can select a variant to access the retrieved literature (MEDLINE abstracts, EuropePMC full-texts and clinical trials). Each document is displayed with highlighted annotations. Users can mark publications of interest. Thus they can at the end generate and export a report (i.e. JSON or CSV) that summarize all the variants of interest with publications selected as relevant by the user. In addition, public APIs are also available for the major functionalities: e.g. retrieving ranked literature for a given variant, retrieving annotated literature and retrieving variant synonyms. Finally, Variomes services are also integrated with the SVIP prototype.

4 Conclusion

We are proposing an efficient tool for retrieving literature associated with variants. The system defines a new state of the art for retrieving genomic variants. For literature triage, our system was able to retrieve almost two thirds of the relevant publications in the top-5. The Variomes system is particularly efficient with single nucleotide variants, where the P5 was greater than 80% for most of the SNVs queries. Such a result is consistent with the targeted application of the tool: SNV accounts for the vast majority of contents in VCF files, whether they are based on gene panels or on WGS/WES. We are now including other types of variants in the system to better cover the needs of personalized medicine. The expansion of the tool beyond somatic mutations is also a work area. Germline mutations, as well as mutations in non-human genomes, including viruses, are currently being considered.

In comparison to LitVar, one of the most popular variant-specific literature search tools, we obtained competitive results. Indeed, our system retrieved more relevant documents than LitVar. On the one hand, thanks to the various patterns proposed by the SynVar service, more relevant documents were retrieved. On the other hand, our system offers a larger collection as it covers not only human research but also animal studies. In addition, our system is also more focused on the exact variant requested by the curator, while LitVar returns any variation occurring at the same locus, which may result in an overload of publications to curate. Further, the literature content of Variomes, which is powered by the SIB Literature Services (SIBiLS) (Gobeill *et al.*, 2020), is also wider than LitVar as it provides access to complementary data sources, such as the ClinicalTrials.gov, which are directly relevant for variant curation, functional assays of human genes in non-human cells, or collections describing biodiversity-related variants, from mammals (e.g. Bats, Pangolin, etc.) to viruses, but which could have relevance for healthcare. Such additional content was not used in the present study but could potentially broaden the scope of the variant search.

As future work we consider four aspects. First, we would like to investigate the use of Supplementary Data, in particular spreadsheets and images. Second, identifying textual evidence in publication to re-rank publications might have a positive impact on the literature triage (Allot *et al.*, 2021; Mottin *et al.*, 2017). Third, pretrained language and ensemble learning models (Knafou *et al.*, 2020) could be opportunely used to provide the curator with a more focused evidence passage to support the curation work of mutation databases (Stekhoven *et al.*, 2018). Finally, a current constraint of our system is the slow response time (up to 5 min for variants associated with large sets of literature). Parallelization and/or pre-annotation would be investigated to accelerate the system.

To conclude, the system we developed has the potential to significantly propel variant curation. It is however to be noted that such a system is neither intended to replace human curators, nor clinical expertise, but rather to support these professionals by cutting down the cost of the manual triage of the literature.

Acknowledgements

The authors thank the SVIP project team members of the following groups: Clinical Bioinformatics Unit of NEXUS Personalized Health Technologies, SIB Clinical Bioinformatics and Swiss-Prot group, namely Daniel Stekhoven, Valérie Barbié, Anne Estreicher, Livia Famiglietti, Faisal Al Quaddoomi, David Meyer, Linda Grob, Franziska Singer and Nora Toussaint. This work also benefited from discussions with Melissa Cline.

Funding

This work was supported by the Swiss Personalized Health Network (SPHN) and the BioMedIT infrastructure. The work presented in this report is built on top of SIBiLS, the SIB Literature Services, which is supported by the Elixir Data Platform.

Conflict of Interest: none declared.

Data availability

Variomes is publicly available at https://candy.hesge.ch/Variomes. Source code is freely available at https://github.com/variomes/sibtm-variomes. SynVar is publicly available at https://goldorak.hesge.ch/synvar.

References

Allot,A. *et al.* (2018) LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Res.*, **46**, W530–W536.

Allot,A. *et al.* (2021) LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res.*, **49**, W352–W358.

Aslam,J.A. and Montague,M. (2001) Models for metasearch. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, United States, pp. 276–284.

Belkin,N.J. *et al.* (1995) Combining the evidence of multiple query representations for information retrieval. *Inf. Process. Manag.*, **31**, 431–448.

Caucheteur,D. *et al.* (2019) Designing retrieval models to contrast precision-driven ad hoc search vs. recall-driven treatment extraction in precision medicine. In: *TREC Proceedings*, Gaithersburg.

Caucheteur,D. *et al.* (2020) Text-mining services of the Swiss variant interpretation platform for oncology. *Stud. Health Technol. Inf.*, **270**, 884–888.

Cline,M.S. *et al.*; BRCA Challenge Authors. (2018) BRCA challenge: BRCA exchange as a global resource for variants in BRCA1 and BRCA2. *PLoS Genet.*, **14**, e1007752.

Ehrler,F. *et al.* (2005) Data-poor categorization and passage retrieval for gene ontology annotation in Swiss-Prot. *BMC Bioinformatics*, **6**, S23.

Faessler,E. *et al.* (2019) JULIE Lab & Med Uni Graz @ TREC 2019 precision medicine track. In: *TREC Proceedings*, Gaithersburg.

Fokkema,I.F. *et al.* (2011) LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat*, **32**, 557–563.

Fox,E.A. and Shaw,J.A. (1994) Combination of multiple searches. *NIST Special Publ. SP*, **243**, 6.

Gaudet,P. *et al.* (2017) The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.*, **45**, D177–D182.

Gobeill,J. *et al.* (2020) SIB literature services: RESTful customizable search engines in biomedical literature, enriched with automatically mapped biomedical concepts. *Nucleic Acids Res.*, **48**, W12–W16.

Jiang,Y. *et al.* (2019) GTX.Digest.VCF: an online NGS data interpretation system based on intelligent gene ranking and large-scale text mining. *BMC Med. Genomics*, **12**, 193.

Jimeno Yepes,A. and Verspoor,K. (2014) Literature mining of genetic variants for curation: quantifying the importance of supplementary material. *Database*, **2014**, bau003.

Knafou,J. *et al.* (2020) BiTeM at WNUT 2020 shared task-1: named entity recognition over wet lab protocols using an ensemble of contextual language models. In: *Proceedings of the Sixth Workshop on Noisy User-generated Text, 2020*. pp. 305–313.

Lee,J.H. (1997) Analyses of multiple evidence combination. In: *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia. pp. 267–276.

Lee,K. *et al.* (2021) Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. *Brief. Bioinf.*, **22**, bbaa142.

Lin,Y.H. *et al.* (2019) variant2literature: full text literature search for genetic variants. *bioRxiv*, 583450.

Lv,C. *et al.* (2020) Literature triage on genomic variation publication by knowledge-enhanced multi-channel CNN. *arXiv*, arXiv:2005.04044.

Mottin,L. *et al.* (2017) Triage by ranking to support the curation of protein interactions. *Database*, **2017**, bax040.

Nie,A. *et al.* (2020) LitGen: genetic literature recommendation guided by human explanations. *Pac. Symp. Biocomput.*, **25**, 67–78.

Pasche,E. *et al.* (2018) SIB text mining at TREC 2018 precision medicine track. In: *TREC Proceedings*, Gaithersburg.

Roberts,K. *et al.* (2018) Overview of the TREC 2018 precision medicine track. In: *TREC Proceedings*, Gaithersburg.

Roberts,K. *et al.* (2019) Overview of the TREC 2019 precision medicine track. In: *TREC Proceedings,* Gaithersburg.

Robertson,S. and Zaragoza,H. (2009) The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retrieval*, **3**, 333–389.

Savoy,J. (2004) Combining multiple strategies for effective monolingual and cross-language retrieval. *Inf. Retrieval*, **7**, 121–148.

Ševa,J. *et al.* (2019) VIST – a Variant-Information Search Tool for precision oncology. *BMC Bioinformatics*, **20**, 429.

Singer,F. *et al.* (2018) SwissMTB: establishing comprehensive molecular cancer diagnostics in Swiss clinics. *BMC Med. Inf. Decis. Mak.*, **18**, 89.

Sioutos,N. *et al.* (2007) NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inf.*, **40**, 30–43.

Stekhoven,D.J. *et al.* (2018) Swiss Variant Interpretation Platform for Oncology (SVIP-O). *Swiss Med. Inf.*, **34, w00411**.

Thakur,N. *et al.* (2021) BEIR: a heterogeneous benchmark for zero-shot evaluation of information retrieval models. *arXiv*:2104.08663.

Wei,C.H. *et al.* (2018) tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinformatics*, **34**, 80–87.

Wishart,D.S. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.

Yip,Y.L. *et al.* (2007) Retrieving mutation-specific information for human proteins in UniProt/Swiss-Prot Knowledgebase. *J. Bioinf. Comput. Biol.*, **5**, 1215–1231.