Enhancing Precision Medicine: An Automatic Pipeline Approach for Exploring Genetic Variant-Disease Literature

Lidia Contreras-Ochando$^{1[0000-0001-8213-1765]}$, Pere Marco Garcia$^{1[0009-0003-7026-3543]}$, Ana León$^{1[0000-0003-3516-8893]}$, Lluís-F. Hurtado$^{1[0000-0002-1877-0455]}$, Ferran Pla$^{1[0000-0003-4822-8808]}$, and Encarna Segarra$^{1,2[0000-0002-5890-8957]}$

[1]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camí de Vera s/n, València, 46020, Spain
[2]ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence, Camí de Vera s/n, València, 46020, Spain
{liconoc,anleopa,lhurtado,fpla,esegarra}@upv.es

Abstract. Advancements in genomics have generated vast amounts of data, requiring efficient methods for exploring the relationships between genetic variants and diseases. This paper presents a pipeline approach that automatically integrates diverse biomedical databases, including NCBI Gene, MeSH, LitVar2, PubTator, and SynVar, for retrieving comprehensive information about genes, variants, diseases, and associated literature. The pipeline consists of multiple stages: querying and searching across the different databases, extracting relevant data, and applying filters to refine the results. Its goal is to bridge the gap in information retrieval related to genetic variants and diseases by providing a systematic framework for discovering relevant literature. The pipeline uses open-access sources to uncover additional articles not referenced in expert reports that mention the genetic variants of interest. In this paper, we present the methodology of the pipeline, discuss its limitations and highlight its potential for advancing information systems, data management, and interoperability in the domains of genomics and precision medicine.

Keywords: Genomic variant-disease literature · Precision medicine · Knowledge integration

1 Introduction

The fields of genomics, precision medicine, and genetic variant-disease associations have advanced significantly, reshaping our understanding of human biology and healthcare (8; 11; 7). Precision medicine, driven by genomics, aims to customize medical interventions based on individuals' unique genetic profiles, resulting in more accurate diagnoses and improved treatment outcomes (18). Investigating the associations between genetic variants and diseases is crucial for uncovering disease mechanisms, identifying therapeutic targets, and developing targeted interventions (3).

Research in the field of genetic variant-disease associations has rapidly expanded with the availability of large-scale genomics datasets and advanced computational methods (19). Understanding these associations holds immense implications for healthcare, including improved disease risk assessment, personalized treatment strategies, and advancements in drug development (9). The exploration of genetic variant-disease associations requires the efficient extraction of relevant literature that describes the relationships studied between the variant and the disease. However, the large amount of existing literature and the different nomenclatures for genes, variants and diseases make it extremely difficult to find the relevant information in each case (2).

Our research focuses on developing a pipeline approach to facilitate the discovery of genetic variant-disease associations by finding the relevant literature in each case. By integrating diverse data sources and considering semantic annotations, our pipeline streamlines the identification of relevant literature and the extraction of pertinent information, allowing researchers and clinicians to gain deeper insights into the genetic foundations of diseases. In the era of big data, where the analysis and interpretation of vast genomics and biomedical information are crucial (10; 13), our pipeline offers a valuable solution to navigate and extract meaningful insights from this wealth of data.

The main contributions of this paper are: (1) A novel and automatic system with a pipeline approach for the efficient retrieval of literature related to genetic variants; (2) The use of semantic annotations and disease synonym matching to filter and prioritize articles based on variant-disease co-occurrence; (3) By enabling efficient exploration of genetic variant-disease literature, the pipeline provides valuable insights for personalized treatment strategies, disease risk assessment, targeted interventions and population health management.

The following section contains a summary of relevant works in the field. Section 3 defines the problem we address in this paper. Section 4 gives details of our approach. Section 5 includes the experiments carried out to test the pipeline. Finally, Section 6 closes the paper with the conclusions and future work.

2 Related Work

The field of genomics and precision medicine has witnessed the emergence of several tools and resources aimed at exploring genetic variant-disease literature, supporting researchers and clinicians in uncovering meaningful insights from genomics data. However, many of these tools focus on specific aspects of the exploration process and may not offer a comprehensive framework.

Variomes (14) is a curation-support tool for personalized medicine that enables the triage of publications relevant to support an evidence-based decision. This tool's limitation remains in that it does not consider the supplementary material of the publications. Since 80% of the variants appear only in the supplementary material (2), Variomes is not able to find them.

On the other hand, LitVar2 (1) is a database specifically dedicated to exploring literature mentioning genetic variants. It provides publications related to the variants, even when they appear in the supplementary material. LitVar2 is a valuable resource for researchers interested in the literature aspect of variant-disease associations, but it lacks a disease filter and a more extensive range of variant syntactic variations in its search.

ClinVar (12) is a widely used database that aggregates and curates information about genetic variants and their clinical significance. It serves as a valuable resource for researchers and clinicians to assess the potential impact of variants on human health. While ClinVar provides extensive information on gene-variant-disease associations, its focus is primarily on clinical significance and may not provide comprehensive exploration capabilities. The cites provided in Clinvar are retrieved from Litvar2.

Mastermind (4), a search engine specifically designed for genetic variant exploration, allows users to access a vast collection of scientific literature and identify publications related to specific genetic variants. It provides a valuable resource for researchers to find relevant articles. Even though Mastermind is the most complete tool for the goal of this research, it requires a paid subscription.

Compared to these existing tools, our proposed system encompasses a broader scope by integrating multiple databases, using semantic annotations and facilitating disease-specific filtering in an open-access manner. In addition to the comprehensive integration of databases, semantic annotations and disease-specific filtering, our proposed pipeline stands out for its automation, minimizing the manual effort required from the user. The system takes care of the entire process, from data retrieval to the delivery of filtered articles.

3 Problem Definition

Our research addresses the problem of efficiently retrieving relevant literature where a genetic variant and a disease are mentioned. The system automatically integrates diverse data sources, utilizes semantic annotations, and incorporates disease synonym matching to provide a comprehensive and curated list of articles that mention the gene and variant and are filtered by the specified disease. In this section, we define the problem, outlining the inputs, outputs, and formulation of the problem our research aims to address.

1. The inputs are: A gene, a variant and a disease.
2. Given the inputs, the problem can be formulated as follows:
   (a) Identifying and collecting all the aliases associated with the gene (15).
   (b) Generating all possible syntactic variations of the variant (5; 6; 16; 17).
   (c) Retrieving the synonyms associated with the disease.
   (d) Searching for variant-associated publications that match the gene aliases.
   (e) Obtaining annotations for each identified publication.
   (f) Filtering the articles based on the occurrence of the disease.
3. We produce as output a list of articles of interest.

4 Pipeline Methodology

The pipeline methodology presented in this study provides a systematic and automated approach for efficiently exploring literature related to genetic variants and diseases. The system integrates multiple data sources[1] and uses semantic annotations to automatically guide the retrieval of relevant information which can be further analyzed. The following sections outline the key components and steps involved in the pipeline methodology.

4.1 Data Sources

The system's pipeline utilizes diverse biomedical databases to gather comprehensive information about genes, variants, diseases and associated literature. The main data sources include:

- Gene: The Gene database[2], provided by the National Center for Biotechnology Information (NCBI), is a widely used resource for gene-related information.
- SynVar: The SynVar database[3] focuses on providing syntactic variations of genetic variants associated with specific genes.
- MeSH: The MeSH (Medical Subject Headings)[4] database is a controlled vocabulary resource developed by the National Library of Medicine (NLM).
- LitVar2: The LitVar2 database[5] is a resource that focuses on providing relevant literature for each genetic variant.
- PubTator: The PubTator database[6] provides bioconcepts annotated in biomedical literature.

4.2 Pipeline Workflow

Figure 1 shows the pipeline that works as follows: (1) The pipeline begins by querying the NCBI Gene database using the gene of interest to retrieve gene information; (2) Using the gene aliases obtained in the previous step, the pipeline queries the SynVar database to obtain a list of all unique syntactic variations associated with the specified variant; (3) The pipeline retrieves from the MeSH database synonyms of the specified disease; (4) The pipeline searches the LitVar2 database for variant-associated publications. It retrieves articles that mention the gene and the variant; (5) For all the publications found with LitVar2, the pipeline retrieves the semantic annotations of diseases from Pubtator; (6) Finally,

[1] It is worth noting that the pipeline methodology is flexible and can be adapted to accommodate additional data sources or specific requirements based on the research objectives and available resources.

[2] NCBI website: www.ncbi.nlm.nih.gov/gene

[3] Synvar website: https://goldorak.hesge.ch/synvar/

[4] MeSH website: https://www.ncbi.nlm.nih.gov/mesh/

[5] Litvar2 website: https://www.ncbi.nlm.nih.gov/research/litvar2/

[6] Pubtator website: https://www.ncbi.nlm.nih.gov/research/pubtator/

the pipeline filters the retrieved articles based on disease occurrence. Articles not mentioning the specified disease or its synonymous terms are excluded from further analysis.

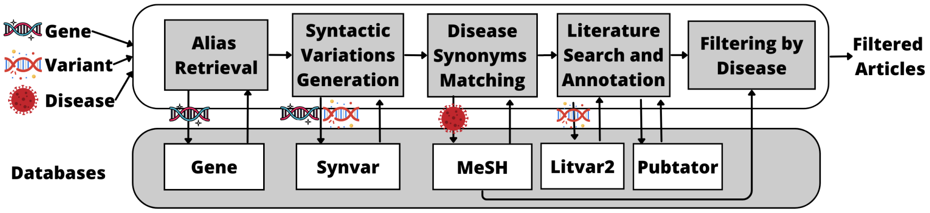

Fig. 1: Pipeline for retrieving filtered publications.

5 Experimental Work

In this section, we present the experiments conducted to evaluate the effectiveness and performance of the proposed system in retrieving variant-associated publications and filtering them by disease occurrence.

Due to the lack of space, Table 1 shows a summary of all the data and results that we will comment in the following sections. We refer the reader to our Github repository for further information, data and code[7] (in Python). All the pipeline connections to databases were performed using Entrez API[8], except for Litvar2 and Synvar, which have their own APIs.

5.1 Test Dataset

For the experiments, we constructed a dataset consisting of 16 examples, each comprising a variant, a gene, a disease, and a list of publications identified by their PubMed IDs (PMIDs). Each example is extracted from one report generated by expert companies in the field of medicine and genetics. Table 1 shows all the examples used.

5.2 Methodology

In the experiments, we applied the pipeline to each of the 16 examples in the dataset. Using the pipeline, we retrieve publications where the genetic variant is mentioned and then filtered the retrieved articles based on the disease occurrence. The results were compared against the expert-generated reports to evaluate the pipeline's accuracy and effectiveness in identifying relevant articles.

[7] Github repository: https://github.com/liconoc/variant_disease_pipeline/
[8] Entrez website: https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html

5.3 Results

We assessed the pipeline's ability to identify articles referenced in the reports, as well as the identification of additional articles that were not mentioned in the reports. This comparison allowed us to evaluate the pipeline's performance and the ability to provide up-to-date information beyond expert reports.

Variant	#names	Gene	#names	Disease	#names	#ref	#sys	#ref_sys
Asn1271Lys	61	MYH7	9	Dilated Cardiomyopathy	60	1	3	0
Asp1272His	61	MYH7	9	Dilated Cardiomyopathy	60	1	0	0
Val411Met	61	SCN5A	15	Long QT Syndrome	48	25	34	7
Asp2Asn	62	PAX5	4	Hepatoblastoma	15	1	0	0
Arg1012Ter	1	STAG2	8	Sarcome Ewing	86	4	0	0
Ile258Val	62	PDGFRA	4	Sarcome Ewing	86	4	2	0
Pro1685Leu	18	PLXNB2	6	Hepatic Sarcoma	1	4	0	0
Arg2072Gln	62	ROS1	4	Hepatic Sarcoma	1	5	4	0
Ala670Val	60	HCN1	10	Ependynoma	1	1	0	0
Ser333Phe	18	ATM	9	Ependynoma	1	1	35	0
Arg365Gln	59	RAD50	4	Ependynoma	1	1	22	0
Ala670Val	60	HCN1	10	Neuroblastome	1	1	0	0
Tyr776Ter	1	TSC2	4	RHABDOMYOSARCOMA	51	1	0	0
Ser273Cys	59	MRE11	5	Nefroblastome	1	1	7	0
Lys1296Arg	61	DOCK3	4	Acute Myeloid leukemia	175	1	2	0
Thr1010Ile	18	MET	17	Malignant tumor of peripheral nerve sheath	1	9	118	1
						61	**227**	**8**

Table 1: List of examples of the test dataset, including a variant, a gene, a disease and a list of publications. #names are the number of alias and synonyms found for the variants, genes and diseases; #ref is the number of articles of reference found in the expert's reports; #sys is the number of articles found using the pipeline system; #exp_found is the number of articles referenced by the experts that have been found with the pipeline.

Aliases and Synonyms To find all the articles that name the variants and to be able to filter by disease, we first need to know all the possible names that the genes, variants and diseases in the examples may have. Table 1 shows the number of names the pipeline could find for them. Note that the number of existing names for a variant depends on which gene it is related to. The same variant can have different names for two different genes.

Identification of articles referenced by the experts In our study, we evaluated the performance of the pipeline in retrieving variant-associated publications and compared the results with expert-generated reports. Out of the 61 publications mentioned in the expert reports, the pipeline successfully retrieved 8 publications (abstracts). However, it is important to note that, upon further investigation, we discovered that some of the articles were not available in the PubMed Central (PMC)[9] database. LitVar2 relies on Pubmed[10] and PMC as

[9] PMC website: https://www.ncbi.nlm.nih.gov/pmc

[10] PubMed website: https://pubmed.ncbi.nlm.nih.gov/

its primary sources of information[11]. However, the articles can be published in other repositories that may not be open access. As a result, it is possible that Litvar2 does not retrieve certain publications referenced in the expert reports. Our results indicate that a significant portion of the expert-generated references were not publicly accessible. In addition, a manual scan of the named articles in the expert reports revealed that only 14 mentioned the variant in the title, abstract, or PMC full-text. One of them would not even name the gene.

Identification of additional articles Contrary to the previous results, the pipeline yielded a significantly higher number of articles than those referenced by the expert reports. In total, we identified 227 publication abstracts and 190 PMCs (full-text). Among these 227 publications, 202 articles mentioned the specific variant of interest accurately [12]. The 25 articles discarded shared the same rsID, but differed in the nucleotide of the mutation. When using the rsID nomenclature[13] on the articles, it may indicate a different mutation. Table 2 shows the different locations where the variants were found within the articles. This comprehensive coverage across different sections of the articles enhances the reliability and robustness of the pipeline.

Title	Abstract	Text	Sup.Mat.	**Total**
6	12	77	96	**202**

Table 2: The number of articles where the variant was mentioned depending on the section of the article where it was found.

Filtering by the disease Upon applying the disease filtering step, we observed that only one disease maintained a substantial number of articles out of the total found. Specifically, this disease retained 28 out of the 34 articles identified. The effectiveness of the disease filtering process highlights the importance of considering the specific disease context when exploring gene-variant literature. By filtering the articles based on disease relevance, we ensure that the final set of articles specifically focuses on the intersection of the gene, variant and disease of interest.

[11] PubMed is a repository of publication abstracts in the field of biomedicine and life sciences, while PMC is a repository of full-text articles.

[12] Variant names indicate the position in which the mutation occurs in the DNA, RNA or protein sequence, as well as the original nucleotide or amino acid (wild type) and the mutated one.

[13] rsID nomenclature indicates the position and wild type but not the specific mutation. Therefore, one rsID can identify all mutations with the same wild type and the same position.

5.4 Discussion

Despite the dependency on PMC, LitVar2 remains a valuable resource for capturing a substantial number of gene-variant-disease open-access literature. Its integration with other databases and resources, along with continuous updates and advancements in literature curation, ensures a comprehensive and up-to-date collection of relevant publications. The pipeline's ability to identify and retrieve a portion of the expert-referenced articles and many up-to-date new ones demonstrates its utility in complementing expert reports and providing an efficient tool for exploring gene-variant-disease literature.

The disparity in the number of articles found and retained for each variant and disease underscores the variability in the availability of literature and its explicit mention of diseases. It is important to note that while some variants and diseases may have fewer associated articles, these articles are expected to provide valuable insights and contribute to understanding the specific gene-variant-disease connections.

6 Conclusions and Future Work

In this paper, we have presented a pipeline to enhance the exploration of gene-variant-disease literature by leveraging advanced search techniques and integrating multiple databases. We identified a significantly large number of articles that contribute valuable scientific evidence to the field of precision medicine and genomics. Our pipeline provides researchers with a comprehensive and efficient tool to investigate gene-variant-disease co-occurrences. However, it is important to consider the limitations imposed by the availability of full-text content. The pipeline's performance relies on the openness and accessibility of literature, which can impact the comprehensive retrieval of all relevant publications.

As future work, ongoing enhancements to the pipeline's algorithms and data integration techniques will contribute to improving its accuracy and efficiency. Furthermore, including additional open data sources and utilizing emerging technologies, such as natural language processing, can further enhance the pipeline's capabilities to extract relevant information from publications and uncover relationships among biomedical entities. A key aspect to address is retrieving information and relationships from supplementary materials, as they often contain valuable data regarding the variants of interest.

Acknowledgements This work is partially supported by MCIN/AEI/10.13039/501100011033, by the "European Union" and "NextGenerationEU/MRR", and by "ERDF A way of making Europe" under grants PDC2021-120846-C44 and PID2021-126061OB-C41. It is also partially supported by the Generalitat Valenciana under project CIPROM/2021/023.

We would like to thank the authors of LitVar2 for their valuable assistance.

Bibliography

[1] Allot, A., Peng, Y., Wei, C.H., Lee, K., Phan, L., Lu, Z.: Litvar: a semantic search engine for linking genomic variant data in pubmed and pmc. Nucleic acids research **46**(W1), W530–W536 (2018)

[2] Allot, A., Wei, C.H., Phan, L., Hefferon, T., Landrum, M., Rehm, H.L., Lu, Z.: Tracking genetic variants in the biomedical literature using litvar 2.0. Nature Genetics pp. 1–3 (2023)

[3] Cano-Gamez, E., Trynka, G.: From gwas to function: using functional genomics to identify the mechanisms underlying complex diseases. Frontiers in genetics **11**, 424 (2020)

[4] Chunn, L.M., Nefcy, D.C., Scouten, R.W., Tarpey, R.P., Chauhan, G., Lim, M.S., Elenitoba-Johnson, K.S., Schwartz, S.A., Kiel, M.J.: Mastermind: a comprehensive genomic association search engine for empirical evidence curation and genetic variant interpretation. Frontiers in genetics **11**, 577152 (2020)

[5] Den Dunnen, J.T., Dalgleish, R., Maglott, D.R., Hart, R.K., Greenblatt, M.S., McGowan-Jordan, J., Roux, A.F., Smith, T., Antonarakis, S.E., Taschner, P.E., et al.: Hgvs recommendations for the description of sequence variants: 2016 update. Human mutation **37**(6), 564–569 (2016)

[6] den Dunnen, J.T.: Sequence variant descriptions: Hgvs nomenclature and mutalyzer. Current Protocols in Human Genetics **90**(1), 7–13 (2016)

[7] England, N.: Accelerating genomic medicine in the nhs. NHS England Website [online] Available at: www. england. nhs. uk/long-read/accelerating-genomic-medicine-in-the-nhs (accessed 23 November 2022) (2022)

[8] Ginsburg, G.S., Phillips, K.A.: Precision medicine: from science to value. Health affairs **37**(5), 694–701 (2018)

[9] Goetz, L.H., Schork, N.J.: Personalized medicine: motivation, challenges, and progress. Fertility and sterility **109**(6), 952–963 (2018)

[10] Hassan, M., Awan, F.M., Naz, A., deAndrés Galiana, E.J., Alvarez, O., Cernea, A., Fernández-Brillet, L., Fernández-Martínez, J.L., Kloczkowski, A.: Innovations in genomics and big data analytics for personalized medicine and health care: A review. International journal of molecular Sciences **23**(9), 4645 (2022)

[11] Krainc, T., Fuentes, A.: Genetic ancestry in precision medicine is reshaping the race debate. Proceedings of the National Academy of Sciences **119**(12), e2203033119 (2022)

[12] Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., Maglott, D.R.: Clinvar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research **42**(D1), D980–D985 (2014)

[13] Luo, J., Wu, M., Gopukumar, D., Zhao, Y.: Big data application in biomedical research and health care: a literature review. Biomedical informatics insights **8**, BII–S31559 (2016)

[14] Pasche, E., Mottaz, A., Caucheteur, D., Gobeill, J., Michel, P.A., Ruch, P.: Variomes: a high recall search engine to support the curation of genomic variants. Bioinformatics **38**(9), 2595–2601 (2022)

[15] Povey, S., Lovering, R., Bruford, E., Wright, M., Lush, M., Wain, H.: The hugo gene nomenclature committee (hgnc). Human genetics **109**, 678–680 (2001)

[16] Saberian, N.: Text Mining of Variant-Genotype-Phenotype Associations from Biomedical Literature. Wayne State University (2020)

[17] Smigielski, E.M., Sirotkin, K., Ward, M., Sherry, S.T.: dbsnp: a database of single nucleotide polymorphisms. Nucleic acids research **28**(1), 352–355 (2000)

[18] Strianese, O., Rizzo, F., Ciccarelli, M., Galasso, G., D'Agostino, Y., Salvati, A., Del Giudice, C., Tesorio, P., Rusciano, M.R.: Precision and personalized medicine: how genomic approach improves the management of cardiovascular and neurodegenerative disease. Genes **11**(7), 747 (2020)

[19] Uffelmann, E., Huang, Q.Q., Munung, N.S., De Vries, J., Okada, Y., Martin, A.R., Martin, H.C., Lappalainen, T., Posthuma, D.: Genome-wide association studies. Nature Reviews Methods Primers **1**(1), 59 (2021)