Semantifying Genomic Variant Data: VCF to RDF Conversion Framework

Semantifying Genomic Variant Data: VCF to RDF Conversion Framework Elias Crum[1,2,*], Ruben Taelman[1], Bart Buelens[2], Gokhan Ertaylan[2] and Ruben Verborgh[1]

[1]IDLab, Department of Electronics and Information Systems, Ghent University – imec, Belgium
[2]Flemish institute for Technological Research (VITO) Mol, Belgium

Abstract

Variant Call Format (VCF) files, the standard for representing patient genomic variant data, face limitations in interoperability, data linking, querying, and semantic interpretability. We propose a framework for converting VCF data into semantic data using the Resource Description Framework (RDF) to address these limitations. Our approach includes a comprehensive ontology, a storage-efficient RDF representation using Header Dictionary Triples (HDT), and integration with clinical metadata schemas like that proposed by SPHN. Our framework and the representation of VCF data semantically will contribute to greater integration, scalability, and usability of these genomic variant data in both genomic medicine and research.

Keywords

Knowledge Representation, Genomic Data, Semantics, Ontology

1. Problem Identification

Genomic medicine, specifically in the emerging fields of pharmacogenomics and rare genetic disease diagnosis, is increasingly informed by patient genomic variant data. To represent patient genomic variant data, the current state-of-the-art is using Variant Call Format (VCF) files [**?**]. VCF files are flat, tab-delineated, and human-readable and represent the variation between an individual and a reference genome.

The current VCF format does have drawbacks. Specifically, VCF format limits interoperability (VCF files are highly specific to genomics and lack inherent support for integration with other biomedical data formats or systems), data-linking capabilities (VCF files are standalone and lack native mechanisms for linking to external resources (e.g., dbSNP, ClinVar, PubMed), querying capabilities (genomic queries in VCF often require bespoke tools and scripts, which can be limiting and non-scalable), and data interpretability (while VCF has a fixed schema, it lacks semantic context, making it challenging to interpret data consistently without domain expertise).

To increase the usability and scalability of patient variant data for medical and research use, one potential solution is developing a conversion framework for VCF representation semantically. Here, we use previously proposed solutions (i.e. VCF ontologies and semantic technologies) to inform an approach for such a conversion process.

2. Proposed Framework

There are existing methods for converting VCF to semantic representations such as the Resource Description Framework (RDF) [1, 2, 3]. While these approaches each have their own strengths and weaknesses, we propose an ontology that offers a complete representation of single-sample VCF file data

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.
elias.crum@ugent.be (S. G. V. D. V. t. R. C. F. E. Crum); ruben.taelman@ugent.be (R. Taelman); bart.buelens@vito.be (B. Buelens); gokhan.ertaylan@vito.be (G. Ertaylan); ruben.verborgh@ugent.be (R. Verborgh)
0009-0005-3991-754X (S. G. V. D. V. t. R. C. F. E. Crum); 0000-0001-5118-256X (R. Taelman); 0000-0001-7734-3747 (B. Buelens); 0000-0001-5602-6435 (G. Ertaylan); 0000-0002-8596-222X (R. Verborgh)

that utilizes preexisting ontologies and represents concept relationships in a manner that encourages performant analysis in downstream applications such as querying and data linking. Along with the components for knowledge graph construction, we also propose a method for the representation of genomic variant RDF data in a compressed, storage-efficient format using Header Dictionary Triples (HDT) [4]. In addition to these aims, we also propose integrating our approach with clinically-related meta-data representation schema such as that proposed by SPHN [5].

Ultimately, this proposed framework will provide a workflow for the representation of VCF genomic variant data semantically, thus encouraging greater use of these data in research and clinical practice

Acknowledgments

Project funding provided from the Research Foundation – Flanders (FWO) (SB Fellowship 1S27825N).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] E. D. S. Penha, E. Iriabho, A. Dussaq, D. M. de Oliveira, J. S. Almeida, Isomorphic semantic mapping of variant call format (VCF2RDF), Bioinformatics 33 (2017) 547–548.

[2] S. Prasanna, A. Kumar, D. Rao, E. J. Simoes, P. Rao, A scalable tool for analyzing genomic variants of humans using knowledge graphs and graph machine learning, Front. Big Data 7 (2024) 1466391.

[3] J. R., Sparqling-genomics · gitlab, 2021. URL: https://gitlab.com/roelj/sparqling-genomics.

[4] J. D. Fernández, M. A. Martínez-Prieto, C. Gutiérrez, A. Polleres, M. Arias, Binary RDF representation for publication and exchange (HDT), Web Semant. 19 (2013) 22–41.

[5] E. van der Horst, D. Unni, F. C. Kopmels, J. Armida, V. Touré, W. Franke, K. Crameri, E. Cirillo, Österle Sabine, Bridging clinical and genomic knowledge: An extension of the sphn rdf schema for seamless integration and fairification of omics data, 2023.