Enabling Genomic Data Harmonization in OMOP CDM

Erwin Tantoso[1], Ngiam Kee Yuan[2,3], Mukkesh Kumar[1,4]

[1] Bioinformatics Institute, Agency for Science Technology and Research, Singapore
[2] Division of General Surgery (Thyroid & Endocrine Surgery), National University Hospital Singapore, Singapore
[3] Department of Biomedical Informatics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[4] Institute for Human Development and Potential, Agency for Science Technology and Research, Singapore

Background

The Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) has revolutionized the idea of large-scale analysis of clinical data from diverse sources by enabling the harmonization of these disparate data models into a common data model and common vocabularies. The adoption of OMOP CDM across multiple institutions in multiple countries has enabled cross-institutional collaborations in various disease domains with the intention to generate real-world evidence and ultimately improve patient care[1]. To enable precision medicine, it requires the integration of genomic variants into the CDM. While the OHDSI tools and vocabularies have been developed in multiple fronts, to date, the focus of OMOP vocabulary for genomic variants (OMOP Genomic) has been placed on genomic variants that are clinically relevant to cancer[2]. This limits the effort of precision medicine in other disease domains and healthy populations; therefore, we believe that improvement on 1) genomic vocabulary; and 2) mapping tools are important to minimize this limitation. Incidentally, the US Food and Drug Administration (US FDA) have identified a gap in interoperable genomic data standards and therefore, it is of strategic value to develop an OMOP/GA4GH interoperability framework using OMOP CDM.

Methods

First: Enriched Genomic Vocabulary (Figure 1)

Objective: To enrich the genomic vocabulary with clinically relevant variants from publicly available literature/curated datasets in the local Singapore context

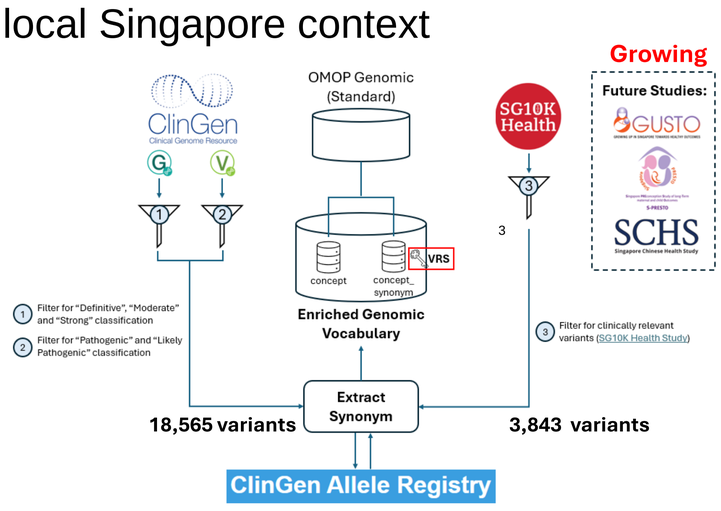

Figure 1. Workflow for Genomic Vocabulary Enrichment

Second: Enabling Query and Mapping of Variants to OMOP Genomic Concept IDs (Figure 2)

Objective: To develop Django REST API for querying and mapping of variants to the OMOP Genomic vocabulary and enriched genomic vocabulary

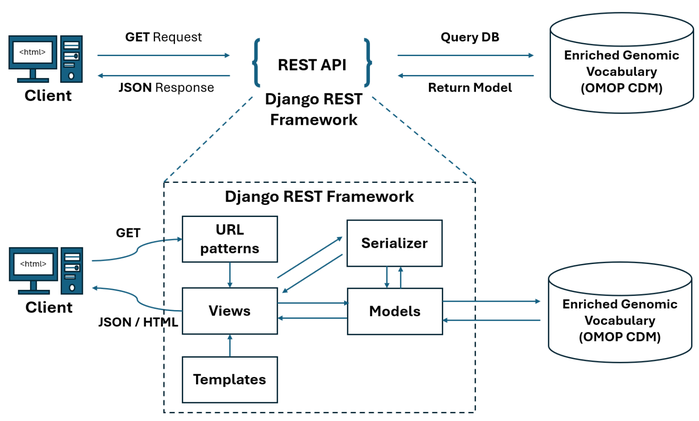

Figure 2. Django REST API Framework for Genomic Vocabulary Query

Results

1) Total Number of Variants in Enriched Genomic Vocabulary

- ClinGen[4] Gene-Disease + Variant Pathogenicity: 18,565 variants
- SG10K Health Study[5]: 3,843 variants

2) Enriched Genomic Vocabulary Contributes Towards a More Comprehensive List Extending the Coverage of Clinically Relevant Variants (Table 1)

Phenotype (ACMG v3.2)	#Genes	OMOP Genomic	Enriched Genomic Vocabulary
Genes related to cancer phenotypes	28	28	28
Genes related to cardiovascular phenotypes	40	10	40
Genes related to inborn errors of metabolism phenotypes	4	1	4
Genes related to miscellaneous phenotypes	9	7	9

Table 1. Coverage of enriched genomic vocabulary on ACMG v3.2 gene list (81 genes)

3) ODmapper – Django REST API framework for genomic data mapping and harmonization (Figure 3)

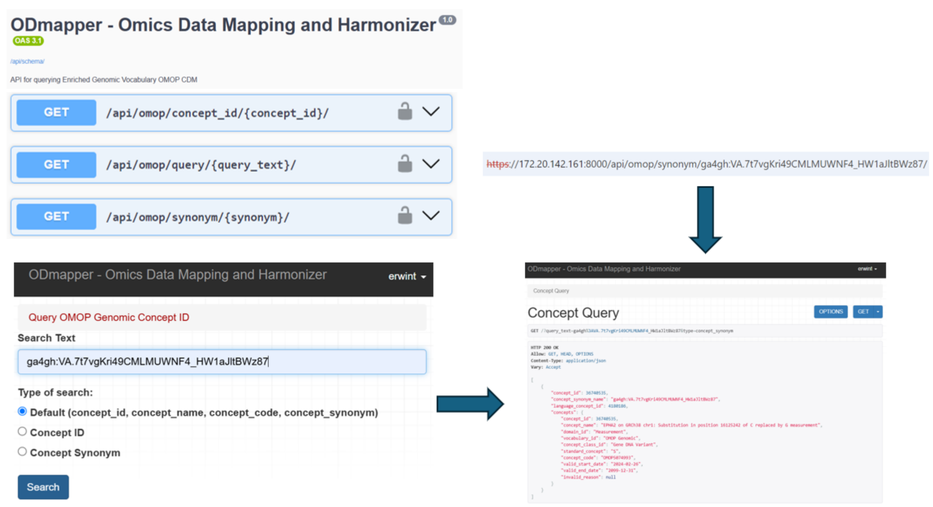

Figure 3. ODmapper GUI for querying OMOP Genomic Concept ID

Conclusion

We have enriched the genomic vocabulary by including clinically relevant variants from public resources to enable harmonization of genomic variants to OMOP CDM space. The development of OMOP Genomic vocabulary REST API facilitates the mapping of variants to their OMOP concept id. This provides the foundation for harmonization of clinically relevant variants in multiple clinical cohorts which will then facilitate precision medicine in diverse medical domains. We envision that the tool will serve as the foundation for development of automatic OMOP CDM conversion for genomic data from diverse cohort studies across the APAC and global OHDSI data network.

References

1. Observational Health Data Sciences and Informatics. Chapter 1. The OHDSI Community. In: The Book of OHDSI.
2. Golozar A, Reich C. 82. Enabling large scale precision oncology research with a new standard for genomic variants: OMOP Genomic. Cancer Genet. 2022 Nov 1;268–269:27.
3. Wagner AH, Babb L, Alterovitz G, Baudis M, Brush M, Cameron DL, et al. The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification. Cell Genomics. 2021 Nov 10;1(2):100027.
4. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen — The Clinical Genome Resource. N Engl J Med. 2015 Jun 4;372(23):2235–42.
5. Chan SH, Bylstra Y, Teo JX, Kuan JL, Bertin N, Gonzalez-Porta M, et al. Analysis of clinically relevant variants from ancestrally diverse Asian genomes. Nat Commun. 2022 Nov 5;13:6694.

Future Work

First: Development of automated OMOP CDM converter for genomic data

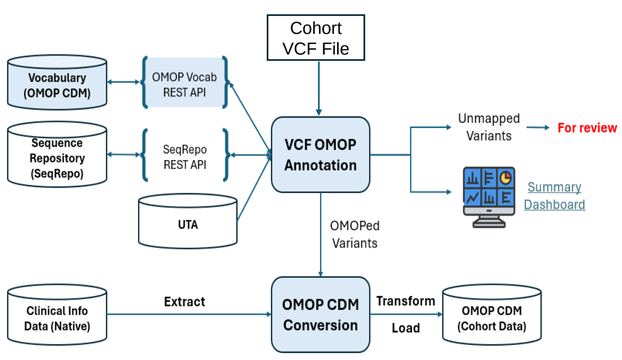

Figure 4. VCF-to-OMOP CDM Converter. The cohort VCF file will be annotated with OMOP Concept IDs based on enriched genomic vocabulary. The OMOPed variants will be converted to OMOP CDM (v5.4).

Second: Deployment of application on the MOH-TRUST TRE (enTRUST)

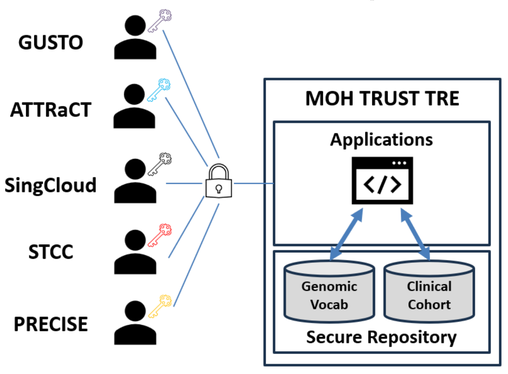

Figure 5. MOH-TRUST Trusted Research Environment (TRE). MOH TRUST TRE as the central TRE to host the OMOP CDM enriched genomic vocabulary and sensitive clinical cohorts which can only be accessible by the trusted users.

Acknowledgement

We are thankful for the continuous support from the OHDSI community, MOH TRUST and A*STAR in advancing the genomic data harmonization initiatives.