ODharmonizer: Enabling Omics Data Harmonization in OMOP CDM

Erwin Tantoso[1], Cindy Ho[1,2], Ismail Mohd[1,2], Sebastian Maurer-Stroh[1,3,4], Johan G Eriksson[2,5,6,7], Ngiam Kee Yuan[8,9], Mukkesh Kumar[1,2]

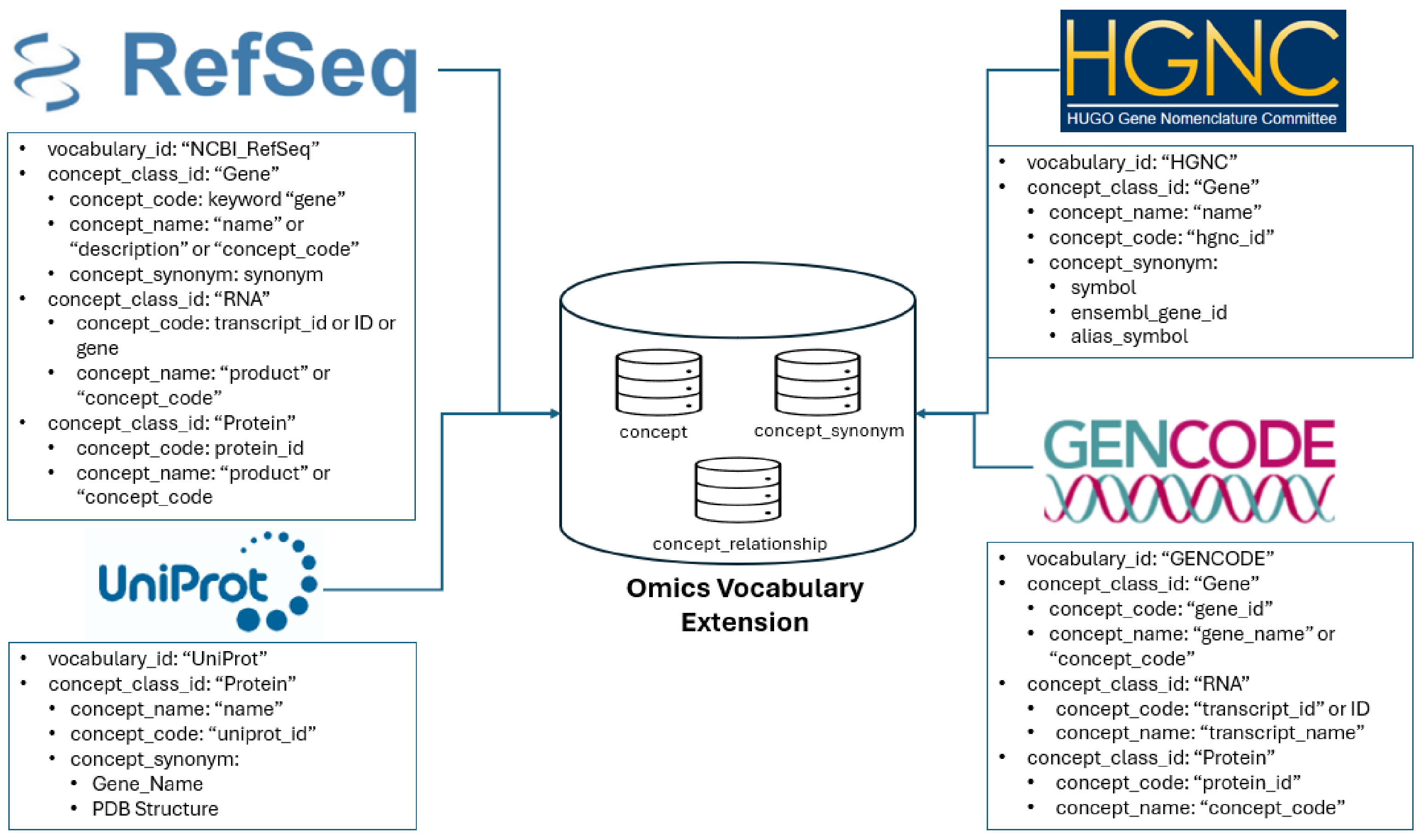
[1] Bioinformatics Institute, Agency for Science Technology and Research, Singapore
[2] Institute for Human Development and Potential, Agency for Science Technology and Research, Singapore
[3] Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[4] Department of Biological Sciences, National University of Singapore, Singapore
[5] Department of Obstetrics and Gynaecology and Human Potential Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
[6] Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland
[7] Folkhälsan Research Center, Helsinki, Finland
[8] Division of General Surgery (Thyroid & Endocrine Surgery), National University Hospital Singapore, Singapore
[9] Department of Biomedical Informatics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Background

The Observational Health Data Sciences and Informatics (OHDSI) initiative established a global research network using the OMOP Common Data Model (CDM) to standardize diverse clinical data[1]. While OMOP CDM has become the global standard for harmonizing electronic health records, it lacks support for high-throughput omics data. The existing OMOP Genomic vocabulary focuses mainly on oncology[2]. At the ISO TC 215 Health Informatics meeting in London (May 2025), genomic and multi-omics data standardization was identified as a strategic priority. Building on our previous work presented at the OHDSI APAC Symposium 2024[3], we expanded genomic vocabularies beyond oncology, but integration of transcriptomic and proteomic data remains a challenge. To address this, we developed ODharmonizer, a modular suite that transforms genomic, transcriptomic, and proteomic data into OMOP CDM-compliant structures. We demonstrate its application by harmonizing Olink proteomics data from the Singapore S-PRESTO cohort, enabling integrated clinical–molecular analyses within OHDSI.

Methods

1. Omics Vocabulary Extension



- vocabulary_id: "NCBI_RefSeq"
- concept_class_id: "Gene"
  - concept_code: keyword "gene"
  - concept_name: "name" or "description" or "concept_code"
  - concept_synonym: synonym
- concept_class_id: "RNA"
  - concept_code: transcript_id or ID or gene
  - concept_name: "product" or "concept_code"
- concept_class_id: "Protein"
  - concept_code: protein_id
  - concept_name: "product" or "concept_code

- vocabulary_id: "UniProt"
- concept_class_id: "Protein"
  - concept_name: "name"
  - concept_code: "uniprot_id"
  - concept_synonym:
    - Gene_Name
    - PDB Structure

- vocabulary_id: "HGNC"
- concept_class_id: "Gene"
  - concept_name: "name"
  - concept_code: "hgnc_id"
  - concept_synonym:
    - symbol
    - ensembl_gene_id
    - alias_symbol

- vocabulary_id: "GENCODE"
- concept_class_id: "Gene"
  - concept_code: "gene_id"
  - concept_name: "gene_name" or "concept_code"
- concept_class_id: "RNA"
  - concept_code: "transcript_id" or ID
  - concept_name: "transcript_name"
- concept_class_id: "Protein"
  - concept_code: "protein_id"
  - concept_name: "concept_code"

Figure 1: Omics vocabulary extension to capture gene or transcript or protein level expression data. HGNC/NCBI_RefSeq/UniProt are recommended as the standard for gene/transcript/protein id respectively. The mapping relationship from NCBI_RefSeq/GENCODE gene to HGNC gene, GENCODE transcript to NCBI_RefSeq transcript and NCBI_RefSeq/GENCODE protein accession to UniProt is defined in the concept_relationship table.

2. ODharmonizer Suite (Software Component Architecture)

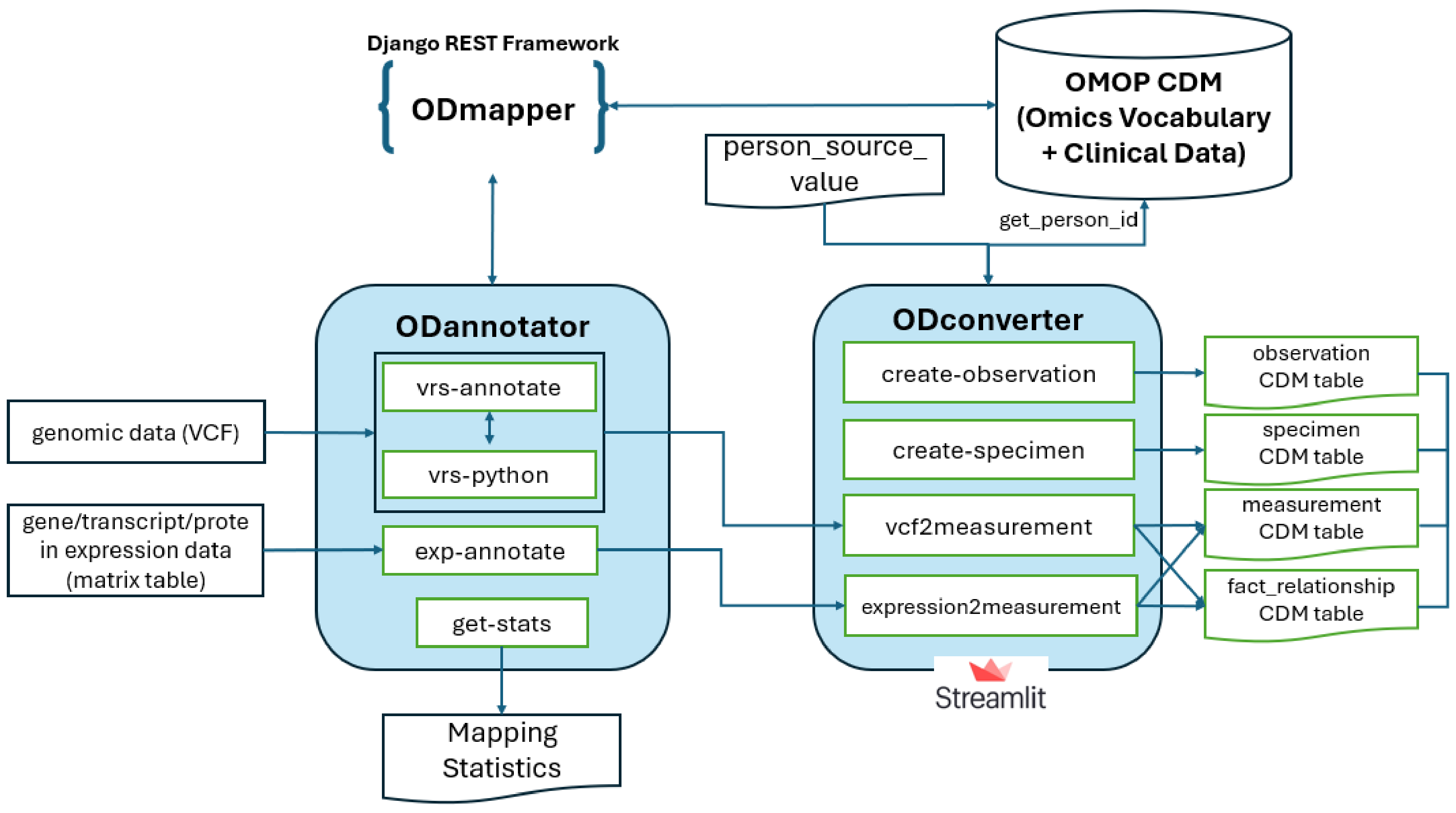


Figure 2: ODharmonizer (https://github.com/biierwint/ODharmonizer) workflow overview. ODannotator is used to annotate the input data leveraging on ODmapper. Genomic data (in VCF format) will be annotated using vrs-annotate to get the ga4gh_id, whereas the expression data will be annotated using exp-annotate. ODannotator leverages on ODmapper to map to concept_id. The annotated data is then converted to OMOP CDM tables using ODconverter to generate observation, specimen, measurement and fact_relationship CDM tables.

Results (1)

1. Omics Vocabulary Extension

Table 1: The statistics of the Omics Vocabulary Extension

Vocabulary	Version	#concept	#concept_synonym
HGNC	2025-08-01	44,433	888,77
NCBI_RefSeq	GCF_00000140540 -RS_2024_08	381,760	63,789
GENCODE	Release 48	576,701	-
UniProt	Release 2025_03	20,420	131,832

Conclusion

We have developed ODharmonizer to harmonize omics data into OMOP CDM schema, enabling the integration of omics data into EHR and clinical phenotype data. This enables researchers to leverage OHDSI resources to analyze multi-omics data as well as studies that require combined clinical and molecular data, such as pharmacogenomics. We have also used ODharmonizer to transform S-PRESTO Olink proteomics data into OMOP CDM. Future development includes the expansion into other omics data types, such as epigenetics, microbiomes, metabolomics and lipidomics.

Results (2)

2. Application on Singapore S-PRESTO Olink Dataset

Table 2: S-PRESTO dataset. There are 8 timepoints measured, which include preconception, pregnancy visit (5 times), delivery and 3 months postnatal. Not all subjects have all the 8 visits.

Info	Counts	Remarks
Total unique subjects	1015	Preconception (908), Pregnancy V1 (326), Pregnancy V2 (325), Pregnancy V3 (325), Pregnancy V4 (325), Pregnancy V5 (312), Delivery (241), Postnatal (270)
Total accumulated visits	3,032	
Olink markers	92	

Table 3: Total number of records in the OMOP CDM tables following the visits by each subject

Tables	#records	Remarks
specimen	3,032	**Linearity preserved**: specimen → observation → measurement **Fact_relationship**: measurement ↔ specimen
observation	3,032	
measurement	278,944	
fact_relationship	557,888	

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. Tantoso E, Ngiam KY, Kumar M. Enabling Genomic Data Harmonization in OMOP CDM. In Singapore; 2024. Available from: https://www.ohdsi.org/wp-content/uploads/2025/01/04_Erwin-Tantoso_Genomic-Data-Harmonization-in-OMOP-CDM.pdf

Acknowledgement

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