All Genes Lead to ROMOPomics

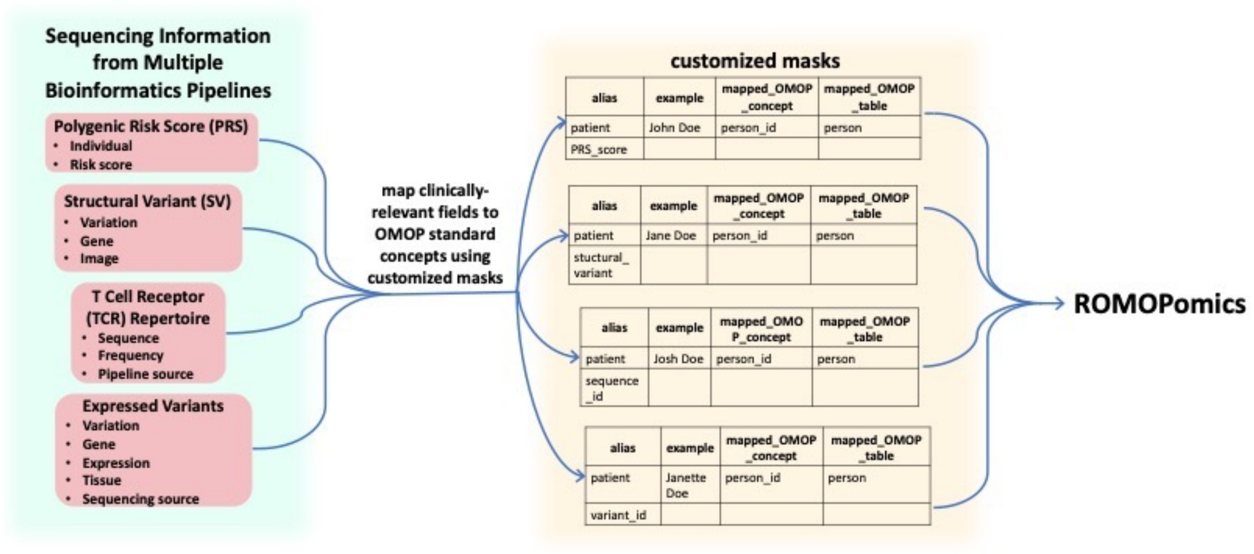
Nicholas Giangreco, MA, MPhil,
Salvatore G Volpe, MA,
Meghana Tandon,
Kamileh Narsinh, MA,
Ben Busby

Background

Clinical applications of sequencing and -omics data remain an underutilized resource for patient care in the push for precision medicine.[1] Clinicians can learn valuable patient-specific information from sequencing data, including PRSs (polygenic risk scores)[2,3], SVs (structural variants),[4,5] SNPs (single nucleotide permutations),[6] and TCR repertoire profiling[7]. Unfortunately, this crucial information is often buried in -omics databases that are impractical for clinical use.[8,9] Furthermore, the multitude of biological silos housing this data do not conform to the same naming conventions, formatting, etc. Standardization is overall lacking. Bioinformaticians, clinical informaticists, computational biologists, and other stakeholders aim to provide clinicians with diagnostically relevant genetic information in a manner that is interpretable and useful at the point of care.[10] To do this, we map select -omics data to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)[11] to enforce standardization, and facilitate clinical decision support.

Methods

We developed a pipeline (Figure 1) for integrating diverse sequencing and molecular datasets into the OMOP CDM by expanding the ROMOPOmics R package (https://github.com/ngiangre/ROMOPOmics).[12] The ROMOPOmics R package facilitates the conversion of sample-centric to observation-centric tabular data for generating a SQLite database of tables in the common data model.



alias	example	mapped_OMOP_concept	mapped_OMOP_table
patient	John Doe	person_id	person
PRS_score			

alias	example	mapped_OMOP_concept	mapped_OMOP_table
patient	Jane Doe	person_id	person
stuctural_variant			

alias	example	mapped_OMOP_concept	mapped_OMOP_table
patient	Josh Doe	person_id	person
sequence_id			

alias	example	mapped_OMOP_concept	mapped_OMOP_table
patient	Janette Doe	person_id	person
variant_id			

Figure 1. Pipeline overview. We first collect and curate sequencing-based datasets. We then correspond the input data fields to fields mapping to tables in the common data model, using a custom OMOP version 6 CDM file. Then the sample-centric data is converted to observation-centric data via ROMOPOmics and into a SQLite database of tables following the common data model.

Conclusion

We have developed a proof-of-concept pipeline and database infrastructure for clinical bioinformatics. This database infrastructure facilitates referencing deriving patient-level information from sample-level data using simple SQL queries, which is rare in bioinformatics and allows future feasibility and comparative research in a clinical context. Future goals include automating the process of corresponding fields and tables in the CDM to fields from input data such that a custom "mask" does not have to be developed for each bioinformatics pipeline.

References/Citations

1. Park JY, Kricka LJ, Clark P, Londin E, Fortina P. Clinical genomics: when whole genome sequencing is like a whole-body CT scan. Clinical chemistry. 2014 Nov 1;60(11):1390-2. https://doi.org/10.1373/clinchem.2014.230276
2. Duncan L, Shen H, Gelaye B, Meijsen J, Ressler K, Feldman M, Peterson R, Domingue B. Analysis of polygenic risk score usage and performance in diverse human populations. Nature communications. 2019 Jul 25;10(1):1-9. https://doi.org/10.1038/s41467-019-11112-0
3. Escott-Price V, Myers AJ, Huentelman M, Hardy J. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. Annals of neurology. 2017 Aug;82(2):311-4. https://doi.org/10.1002/ana.24999
4. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: changing the landscape of chromosomes and design of disease studies. Human molecular genetics. 2006 Apr 15;15(suppl_1):R57-66. https://doi.org/10.1093/hmg/ddl057
5. Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. Frontiers in bioengineering and biotechnology. 2015 Jun 25;3:92.
6. Taylor JG, Choi EH, Foster CB, Chanock SJ. Using genetic variation to study human disease. Trends in molecular medicine. 2001 Nov 1;7(11):507-12.
7. Van der Velden VH, Cazzaniga G, Schrauder A, Hancock J, Bader P, Panzer-Grumayer ER, Flohr T, Sutton R, Cavé H, Madsen HO, Cayuela JM. Analysis of minimal residual disease by Ig/TCR gene rearrangements: guidelines for interpretation of real-time quantitative PCR data. Leukemia. 2007 Apr;21(4):604-11.
8. Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, Ballestar E, Bongcam-Rudloff E, Conesa A, Tegnér J. Data integration in the era of omics: current and future challenges. BMC systems biology. 2014 Mar;8(2):1-0. https://doi.org/10.1186/1752-0509-8-S2-I1
9. López de Maturana E, Alonso L, Alarcón P, Martín-Antoniano IA, Pineda S, Piorno L, Calle ML, Malats N. Challenges in the integration of omics and non-omics data. Genes. 2019 Mar;10(3):238. https://doi.org/10.3390/genes10030238
10. Conesa A, Beck S. Making multi-omics data accessible to researchers. Scientific data. 2019 Oct 31;6(1):1-4. https://doi.org/10.1038/s41597-019-0258-4
11. OMOP Common Data Model – OHDSI. Accessed June 15, 2021. https://www.ohdsi.org/data-standardization/the-common-data-model/
12. Andrew Clugston, & Nick Giangreco. (2021, January 25). ngiangre/ROMOPOmics: First release (Version v.1.0.0). Zenodo. http://doi.org/10.5281/zenodo.4463257.