Introducing KOIOS: removing impediments in genomic variant identification and mapping

Asieh Golozar[1,2], Laurence Lawrence-Archer[1], Phani Kishore Davineni[2] ,Vlad Korsik[1], Varvara Savitskaya[1], Alexander Davydov[1], Nadia kadakova[1], John Methot[2], Christian Reich[2,4]
[1]Odysseus Data Services, MA, USA [2] OHDSI Center at the Roux Institute, Northeastern University, Boston, MA, USA, [3] Dana-Farber Cancer Institute, Boston, MA, USA [4]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Introduction

Information on somatic genomic variants in cancer patients come from a myriad of different sources:

- Next-generation sequencing, full genome and transcriptome
- Circulating tumor DNA analysis
- Target gene panel sequencing
- Target microarrays
- Polymerase chain reaction (PCR)-based methods
- Fluorescence in situ hybridization (FISH)
- Immunohistochemistry (IHC)
- Mass spectrometry
- Companion diagnostic tests

For one and the same variant, all these methods produce very different notations and formats. But for the OMOP CDM and for standardized OHDSI methods to work, each variant must be represented through one unique concept. In addition, the total number of concepts in a vocabulary should not exceed a few 100 thousand before statistics and performance of OHDSI tools and methods deteriorate.

The Oncology Working Group of OHDSI has developed a canonical, comprehensive, and non-redundant representation of genomic variants that are clinically relevant for cancer. This vocabulary, provided as part of the OHDSI Standardized Vocabularies, is called 'OMOP Genomic'. OMOP Genomic was constructed by selecting only cancer relevant mutations out of the almost infinite possible number of variants. This was done by consolidating genomic variants from public knowledgebases, resulting in over 95,000 variations from 575 cancer genes.

When transforming patient data containing variant information, these must be mapped to the appropriate OMOP concepts. This is a non-trivial bioinformatics task requiring to deal with biological variation (e.g. splice variants), ambiguous identifier space from different public databases, various notations for the same mutation, imprecise naming conventions (e.g. referring to estrogen receptor 1 and 2 as estrogen receptor) and ongoing scientific progress.

Here, we introduce Koios, a new open-source online tool that takes as input a reference to a biological entity (gene, transcript, protein), the nature of the mutation and its location, and matches these to the corresponding OMOP Genomic variant concept.

Methods

Koios is a Python tool generating the best of various possibly fuzzy matches of the input information to the OMOP Genomic concepts. As input, it takes lists of variants in HGVS syntax, variant call format (VCF) files or vendor- or lab-specific custom formats such as xml, json, txt or csv containing the four dimensions of a variant: modality (gene, genome, transcript, protein), identifier (called accession number in molecular biological databases), base pair or amino acid location and nature of mutation (substitution, deletion or insertion of base pairs or amino acids). Koios is built to be robust against the heterogeneity of VCF file formats across vendors and institutions. As output, it returns the OMOP concept IDs and all synonymous HGVS variant notations it computed.

To solve this task, Koios performs a number of transformations of the input information:

- A conversion of identifiers from Genbank, EMBL databases, Clinvar and specialized cancer variant databases
- A conversion of identifiers of different versions of the same biological entity
- A conversion (lift) between human reference genome assemblies or chromosomes and their location coordinate combination
- A conversion between biological heterogeneity, such as splice variants and their respective location coordinate combination
- A matching of gene, modality, mutation and coordinates in the absence of a precise identifier

Koios uses ClinGen as its main source of identifier heterogeneity. In future versions, it might also use the sequence alignment tool blast to increase coverage of target identifiers for matching. In case of ambiguous results (matching to more than one OMOP Genomic concept, it uses a heuristic to pick the most likely hit or outputs an error if this cannot be achieved.

Results

Koios reliably recognizes the input and carries out successful mapping to OMOP Genomic. It is available to ETL analysts converting variant information of patients. It is also used to continuously enhance the comprehensiveness and relevance of the OMOP Genomic vocabulary in capturing clinically significant mutations by identifying the set of clinically relevant mutations not available, yet. KOIOS is currently being tested in the context of a series of use cases and the results will be presented at the symposium.

Conclusion

Koios will enable the use of variant information in patient databases, and therefore support the progress of precision medicines by allowing observational studies using variants for target cohorts or covariates.