doi:10.3233/SHTI220031

A Multi-Omics Common Data Model for Primary Immunodeficiencies

Mélanie Buy[a], William Digan[a,b], Xiaoyi Chen[b], Julien Husson[a], Mickael Ménager[d], Frédéric Rieux-Laucat[c], Nicolas Garcelon[a], ATRACTion members

[a] Université de Paris, Imagine Institute, Data Science Platform, INSERM UMR 1163, F-75015, Paris, France
[b] Centre de Recherche des Cordeliers, INSERM, Université de Paris, Sorbonne Université, F-75006, Paris, France
[c] Université de Paris, Imagine Institute, Laboratory of Immunogenetics of Pediatric Autoimmune Diseases, INSERM UMR U1163, F-75015 Paris, France
[d] Université de Paris, Imagine Institute, Laboratory of Inflammatory Responses and Transcriptomic Networks in Diseases, Atip-Avenir Team, INSERM UMR 1163, F-75015 Paris, France

Abstract

Primary Immunodeficiencies (PIDs) are associated with more than 400 rare monogenic diseases affecting various biological functions (e.g., development, regulation of the immune response) with a heterogeneous clinical expression (from no symptom to severe manifestations). To better understand PIDs, the ATRACTion project aims to perform a multi-omics analysis of PIDs cases versus a control group patients, including single-cell transcriptomics, epigenetics, proteomics, metabolomics, metagenomics and lipidomics. In this study, our goal is to develop a common data model integrating clinical and omics data, which can be used to obtain standardized information necessary for characterization of PIDs patients and for further systematic analysis. For that purpose, we extend the OMOP Common Data Model (CDM) and propose a multi-omics ATRACTion OMOP-CDM to integrate multi-omics data. This model, available for the community, is customizable for other types of rare diseases (https://framagit.org/imagine-plateforme-bdd/pub-rhu4-atraction).

Keywords:

Cohort Studies, Databases, rare diseases

Introduction

There are about 7000 different types of disorders and rare diseases which have an impact on a large population worldwide [1]. In this context, patient data collection and analysis of rare and undiagnosed diseases have the potential to offer the opportunity to increase knowledge and discover new therapeutic approaches.

In the project ATRACTion (Autoimmunity & inflammation Through RNAseq Analysis at the single Cell level for Therapeutic Innovation), we are interested in rare Primary Immunodeficiencies (PIDs). PIDs gather more than 400 rare monogenic diseases affecting the development, the function or the regulation of the immune response [2]. Moreover, a given monogenic cause could be associated with variable expression ranging from no symptom (no penetrance) to a broad spectrum of severe and debilitating manifestations. As a national reference center of PIDs, Necker hospital registered more than 5000 PIDs. The variable expression observed in PIDs with autoimmunity/inflammation is therefore leading to diagnosis and therapeutics wandering.

In this project, we will include 250 PIDs patients and 250 controls. Their clinical data and multi-omics data will be collected. As some clinical data are mandatory for omics partners for the results analysis and the determination of omic signature, it is important to combine clinical data and omics data into a common data model.

Nevertheless, the combination of clinical and omics data remains challenging [3]. The first challenge comes from the complexity, heterogeneity and scale of non-omics data (i.e., clinical data). They can be structured (e.g., Body Mass Index (BMI), blood pressure) or unstructured (information extracted from clinical narratives), in various data types (qualitative and quantitative), and from different providers. Moreover, some phenotypes can be absent from a patient record, due to the absence of a medical test. The second challenge lies in combining microbiota data with other omics data [4]. In fact, the metabolome state is a back-and-forth process with the immune system. On one hand, metabolites captured the end products of biochemical reactions, which lead to a patient phenotype. On the other hand, metabolites shape the immune response and therefore impact transcriptomics and proteomics. The third challenge comes from the relation between omics and non-omics data. Omics data show an ascertainment bias due to the experiment itself (case *versus* control), which can be specific to a patient condition (e.g., age, gender, drugs).

State of the art

Many efforts have been made for solutions to combine clinical data and omics data. In the medical informatics community, two approaches have been observed. The Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) [5] is dedicated to the integration of heterogeneous hospital data such as administrative data, clinical results, information extraction from electronic health records. Shin SJ, *et al* 2019 [6], [7] proposed the Genome Common Data model as an extension of the OMOP-CDM to include clinical sequencing data (gene name, variant type and actionable mutation). They integrated 114 lung cancer patients [7] from the Ajou University Hospital, Suwon, Republic of Korea and 1060 patients from the Cancer Genome Atlas. To do so, they added four tables linked to the OMOP-CDM, namely Genomic_Test, Target_Gene, Variant_Occurrence and Variant_Annotation. Nevertheless, they did not deliver a general way to integrate other omics data.

Carnival [8] is a graph database model inspired by the Open Biological and Biomedical Ontology (OBO) Foundry ontologies. This database contains over 60,000 patients from the University of Pennsylvania Health System and includes both clinical and whole-exome genomics data (blood and tissue samples). For the clinical data it contains demographics (age, sex), vital signs (BMI, smoking status, blood pressure), ICD9/10 codes and the loss of genes function.

In the bioinformatics community, we had a look at three propositions: Moped 2.5, Aging atlas, LinkedOmics.

Moped 2.5 [9] is an Integrated Multi-Omics Resource which includes approximately 5 million of transcriptomics and proteomics expression records from over 250 experiments coming from four reference organisms: human, mouse, worm, and yeast. Those data contain protein absolute and relative expression and gene relative expression. Relative expression data in Moped are displayed in terms of pairwise comparisons of conditions (case vs control) using the expression ratio for each protein. Moped provides both p-values and false discovery rate (FDR) estimates. Furthermore, to support the goal of reproducible science they provide a metadata checklist information about experimental design, instrument details, sample preparation, data processing, and analysis.

Another step is reached by Aging Atlas [10], which is a curated aging biology database which integrates multi-omics data **(**transcriptomics, single-cell transcriptomics, epigenomics, proteomics, and pharmacogenomics) from heterogeneous sources**.**

The LinkedOmics [11] database contains multi-omics data (genomics, epigenomics, transcriptomics and mass spectrometry) for 32 cancer types associated with 11,158 patients from The Cancer Genome Atlas (TCGA) project.

Goals

In this work, according to the experimental design of the ATRACTion project, we propose an integrative model of clinical data and multi-omics data, which can be used to obtain standardized information necessary for the characterization of PIDs patients and the identification of molecular signatures. Furthermore, it will allow machine-learning based unsupervised approaches such as network inferences and could facilitate the visualization of the complex data via multi-layer networks.

Methods

We here defined a minimal dataset to integrate both clinical data and multi-omics data. Regarding clinical data, the OMOP data model offers a comprehensive patient-centered description for clinical trial purposes. In this work, we have built our data model by selecting only the tables of OMOP relevant for our studies. Indeed, we have included the Person, Observation, Visit_occurrence, Condition_occurrence, Drug_exposure, Measurement, Procedure_exposure, Survey_conduct and all the tables related to the concepts and vocabularies. We have excluded tables like Cost, Care_site, Device_exposure, which provide a specific level of detail but are not relevant for our studies.

ATRACTion multimodal database

As each data type needs a specific storage format, we have looked for ad hoc ontologies and thesaurus.

Samples collection

The cohort includes 250 patients and 250 healthy relatives followed in the Necker hospital. The required clinical data for the project will be collected from the clinical data warehouse, as well as specific interviews driven by medical staff. Plasma, peripheral blood cells, feces and urine samples will be collected from all individuals and be registered in the laboratory information management system, in which the link between the samples and the analyses will be kept.

Clinical data

To compute the analyses and determine omic signatures, some metadata are mandatory for the omics partners, (e.g. breastfeeding information for metagenomic analyses). For that purpose patients will be interviewed by clinicians within the use of RedCap and collected information then transferred in the ATRACTion database. We have defined the minimal dataset based on both administrative information and clinical data needed for further analyses. We also included for each patient, the medical history from the Necker hospital database to complete the clinical data. We mapped the data with the OMOP data model.

Ontologies and knowledge databases

To ensure data interoperability, symptoms will be described using the Medical Subject Headings (MeSH), the Logical Observation Identifiers Names & Codes (LOINC), the Human Phenotype Ontology (HPO), genes with HGNC ontology, and diagnoses with Orphanet Rare Disease Ontology (ORDO). These ontologies and thesaurus will be stored in the OMOP Concept table. Molecules and biological compounds like proteins or lipids will be linked to KEGG databases.

Omics data

Beyond clinical information, six different types of omics will be analyzed from patient samples (Figure 1): transcriptomics, epigenomics, proteomics, lipidomics, metabolomic, and metagenomic. In order to build a well-designed data model for each omic dataset, we have organized meetings with project partners to give insights from omics field specialists and to collect example data files. We were then also able to highlight some data types inherent issues. In our data model we compare cases *versus* controls patients in order to compute fold changes. More details are provided in the following for each omic data.

Transcriptomics

These data will allow us to identify differential gene expression between patients. The data files will be generated from single-cell experiments performed by using 10x Genomics CITE-seq protocol on peripheral blood mononuclear cells (PBMCs) extracted from blood samples.

Epigenomics

Epigenomics data will inform us about chromatin accessibility which is required to allow binding of transcription factors and subsequent gene expression. Data files will be generated through 10X Genomics ATAC-seq protocols performed on nuclei extracted from PBMCs

Proteomics

CYTOF mass cytometry and Olink experiments will generate a large dataset of differential expression for inflammatory-related proteins in blood samples. These experiments will be performed at single-cell level, which could constrain the group comparison, as needed for others single-cell results.

Metabolomics

The data files will be generated from analyses via chromatography and dosage of LPS activity, to investigate the impact of disease on over 300 metabolites belonging to biological pathways. The results data files will give the quantity of a large number of molecules involved in biological pathways linked with the inflammation process.

Metagenomics

Sequencing process will be performed on fecal samples to find which bacterial genes are present in these samples and infer the corresponding bacteria species. The presence of some species could be linked with specific metabolites, allowing the link between metagenomic and metabolomic data.

Lipidomics

Lipidomics gives the concentration of molecules and can be linked with the presence of bacterial species for specific metabolites. This allows lipidomics and metagenomics data to be correlated. Lipidomics data files will be obtained from analyses performed using high-resolution mass spectrometry on blood et urine samples.

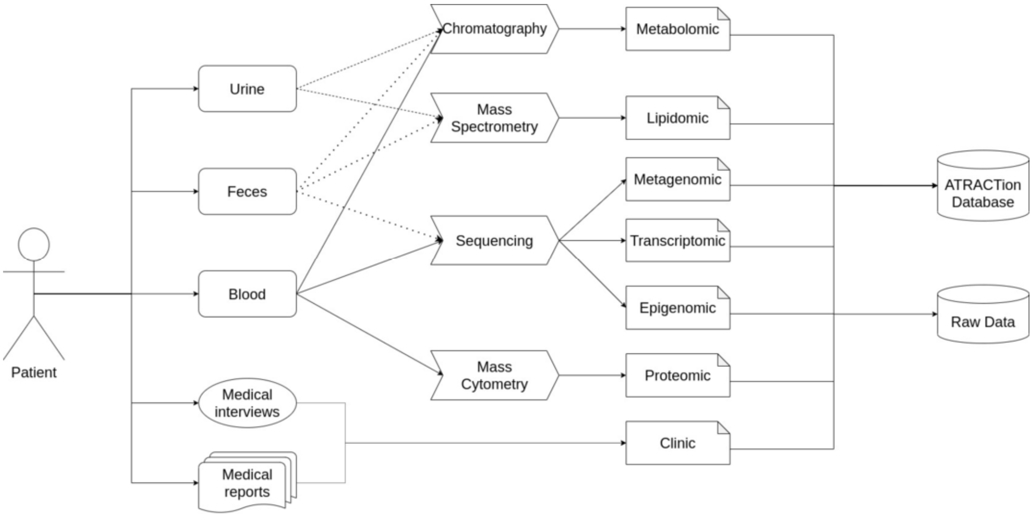

Figure 1: data flow of the ATRACTion data process

Implementation

Regarding the technical aspect, the database is based on an object-relational database implemented in PostgreSQL (available at https://framagit.org/imagine-plateforme-bdd/pub-rhu4-atraction.) and queried via a Django web interface. The relation model of the database stores each type of omic data in specific tables. For each omic, a table suffixed with '_exp' describes metadata about the experience (device, culture medium, statistic model, …) and a table suffixed by '_result' contains numeric results, including p-value, fold change or other types of metrics. For results obtained via group comparisons, the group identifier is stored into the attribute '[omic]_group_id'.

Results

Figure 2 shows the Unified Modeling Language (UML) diagram of the multi-omics ATRACTion OMOP-CDM.

Clinical data set

This dataset consists of 6 parts:

- Patient information and administrative data (e.g. birth date, gender, origin): table PERSON
- Diagnosis and medical condition at the time of the sample (e.g. inflammation state of the patient, BMI, breastfeeding, diet): tables OBSERVATION, CONDITION_OCCURRENCE, SURVEY_CONDUCT, OBSERVATION _PERIOD
- Biological results (creatininemia, autoantibodies etc.): Table MEASUREMENT
- Drugs, one month before the samples: Table DRUG_EXPOSURE
- The type of omic analysis performed can be stored in the PROCEDURE_OCCURRENCE (MeSH Procedure).
- The sample description (type of sample, quantity, etc.): table SPECIMEN, TRANSFORM_SAMPLE.
- In addition, the Vocabulary table stores the different used ontologies, each ontology term being listed in the Concept table, with corresponding entry code.

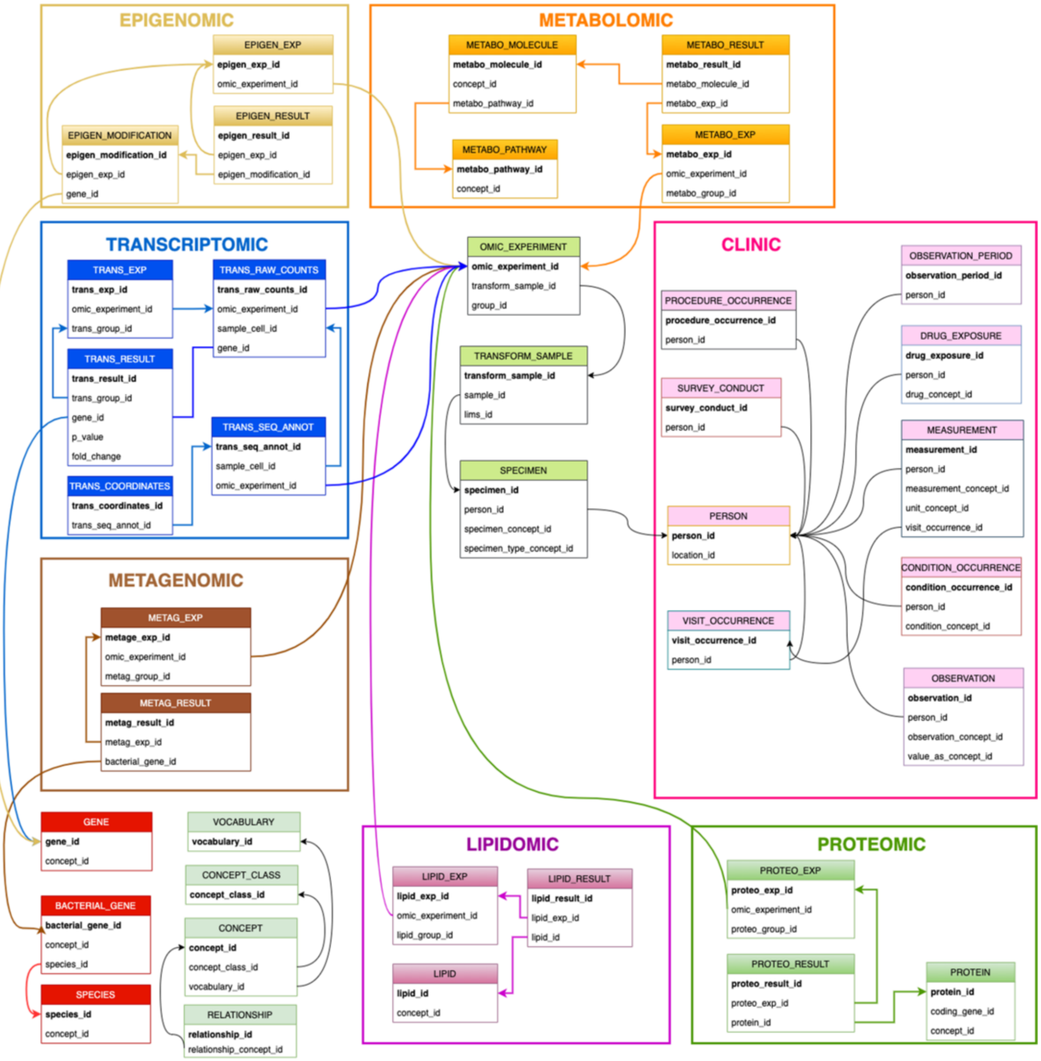

Figure 2: UML diagram of ATRACTion OMOP-CDM, we represent the main key concepts (the complete diagram is in the git project). All the concept_id are linked to the CONCEPT table

Omics

The omic_experiment table is used to group samples sequenced together (batch). Each table with suffix *exp* groups sequenced samples processed together by computational biologist (**_group_id*). Each table with a suffix *_results* contain analysed data and proposed a Fold Change and adjusted pvalue of an experiment.

Transcriptomics

With the help of the raw and normalized files, we build 6 tables: Single_cell_exp, Coordinates, Sequence_annot, Raw_counts, Single_cell_results, and Gene. Single_cell_exp contains metadata about the experiment. Coordinates gathers the sequence reads coordinates for each gene. Sequence_annot contains metadata about sequences, like cellular type, the read sequence in nucleotides. Raw_counts contains raw sequence count associated with an experiment. At last, Gene presents information like gene name and chromosomal location.

Epigenomics

The Modification table relates the genes accessibility by transcription factors associated with peaks of accessible chromatin.

Proteomics

The Protein table details protein information like name, ontology entry, function, and coding gene linked to an experiment.

Metabolomics

3 tables described Metabolomics: Metabolite, Pathway, Metabolomics_Results. The Metabolite table associates a metabolite to a pathway. The Pathway table details pathways within the use of ontology. Metabolomics_Results contains analysed results of an experiment.

Metagenomics

We include 3 tables: Bacterial_gene, Species, and Metagenomic_result. Species table contains relevant information about the bacteria taxonomy. The Bacterial_gene table linked genes to species and biological function of these genes.

Lipidomics

We include 2 tables: Lipidomic_results and Lipid. Lipid contains the lipid name and function.

Discussion

Technical significance

To the best of our knowledge, it is the first description of a clinical and multi-omics (including 6 types of omics data) database implementation, which extended the OMOP-CDM for PIDs diseases. This model can be generalized to the studies of other types of rare diseases, with a focus of case *versus* control comparison.

Significance for secondary use of clinical data

This data representation approach could enhance reproducibility and sharing of models that are learned from this data. Such OMOP-CDM models could be deployed in other reference PIDs centers and researchers will be able to perform distributed queries across them.

Perspectives

In the near future, we will have access to all omics data files to complete and further improve the model. Such a multi-omics database can help identify specific molecular signatures for PIDs, facilitate the visualization of complex data, allow network inferences, and finally build a multi-layer disease network.

Conclusions

In this work, we present a multi-omics ATRACTion OMOP-CDM, which expands the OMOP-CDM in order to include both clinical and multi-omics data. The data definition language is available at https://framagit.org/imagine-plateforme-bdd/pub-rhu4-atraction.

Acknowledgements

This work is supported by a public grant overseen by the French National Research Agency (ANR) as part of the Investments for the Future programme (PIA) under grant agreements No ANR-10-IAHU-01 and ANR-18-RHUS-0010. We thank the ATRACTion members: Zsolt Bocskei, Florence Castelli, Nadine Cerf-Bensussan, Alain Fischer, Nicolas Pons, Anne Remaury, Michel Didier, Francesco Carbone, Mohammad Afshar, Christophe Junot, Dominique Rainteau, Anita Burgun, Josh Waterfall.

References

[1] R. C. Griggs *et al.*, « Clinical research for rare disease: Opportunities, challenges, and solutions », *Mol. Genet. Metab.*, vol. 96, nº 1, p. 20-26, janv. 2009, doi: 10.1016/j.ymgme.2008.10.003.

[2] A. Fischer *et al.*, « Autoimmune and inflammatory manifestations occur frequently in patients with primary immunodeficiencies », *J. Allergy Clin. Immunol.*, vol. 140, nº 5, p. 1388-1393.e8, nov. 2017, doi: 10.1016/j.jaci.2016.12.978.

[3] E. López de Maturana *et al.*, « Challenges in the Integration of Omics and Non-Omics Data », *Genes*, vol. 10, nº 3, mars 2019, doi: 10.3390/genes10030238.

[4] S. H. Chu *et al.*, « Integration of Metabolomic and Other Omics Data in Population-Based Study Designs: An Epidemiological Perspective », *Metabolites*, vol. 9, nº 6, juin 2019, doi: 10.3390/metabo9060117.

[5] R. Makadia et P. B. Ryan, « Transforming the Premier Perspective® Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model », *eGEMs*, vol. 2, nº 1, nov. 2014, doi: 10.13063/2327-9214.1110.

[6] S. J. Shin, S. C. You, J. Roh, Y. R. Park, et R. W. Park, « Genomic Common Data Model for Biomedical Data in Clinical Practice », *MEDINFO 2019 Health Wellbeing E-Netw. All*, p. 1843-1844, 2019, doi: 10.3233/SHTI190676.

[7] S. J. Shin *et al.*, « Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study », *J. Med. Internet Res.*, vol. 21, nº 3, p. e13249, mars 2019, doi: 10.2196/13249.

[8] D. Birtwell, H. Williams, R. Pyeritz, S. Damrauer, et D. L. Mowery, « Carnival: A Graph-Based Data Integration and Query Tool to Support Patient Cohort Generation for Clinical Research », *MEDINFO 2019 Health Wellbeing E-Netw. All*, p. 35-39, 2019, doi: 10.3233/SHTI190178.

[9] E. Montague *et al.*, « MOPED 2.5—An Integrated Multi-Omics Resource: Multi-Omics Profiling Expression Database Now Includes Transcriptomics Data », *OMICS J. Integr. Biol.*, vol. 18, nº 6, p. 335-343, juin 2014, doi: 10.1089/omi.2014.0061.

[10] Aging Atlas Consortium, « Aging Atlas: a multi-omics database for aging biology », *Nucleic Acids Res.*, vol. 49, nº D1, p. D825-D830, janv. 2021, doi: 10.1093/nar/gkaa894.

[11] S. V. Vasaikar, P. Straub, J. Wang, et B. Zhang, « LinkedOmics: analyzing multi-omics data within and across 32 cancer types », *Nucleic Acids Res.*, vol. 46, nº Database issue, p. D956-D963, janv. 2018, doi: 10.1093/nar/gkx1090.

Address for correspondence

Nicolas Garcelon
Institut Imagine, 24 boulevard du Montparnasse,
75015 Paris, France
nicolas.garcelon@institutimagine.org