Breakout Session 1: Track B

Improving FAIRness and AI / ML Readiness of Bioconductor Data Resources

Dr. Sehyun Oh
Assistant Professor, City University of New York

Improving FAIRness and AI/ML readiness of Bioconductor data resources

Cancer Genomics: Integrative and Scalable Solutions in R/Bioconductor

2024.03.27

Sehyun Oh, PhD (Speaker)
Levi Waldron, PhD (PI)

CUNY SPH GRADUATE SCHOOL OF PUBLIC HEALTH & HEALTH POLICY

ver.2

Challenge

CHALLENGES

OMICS and NON-OMICS data relationship

- Correlation between both data types
- Interaction between both data types
- Presence of sub-phenptypes
- Fairness

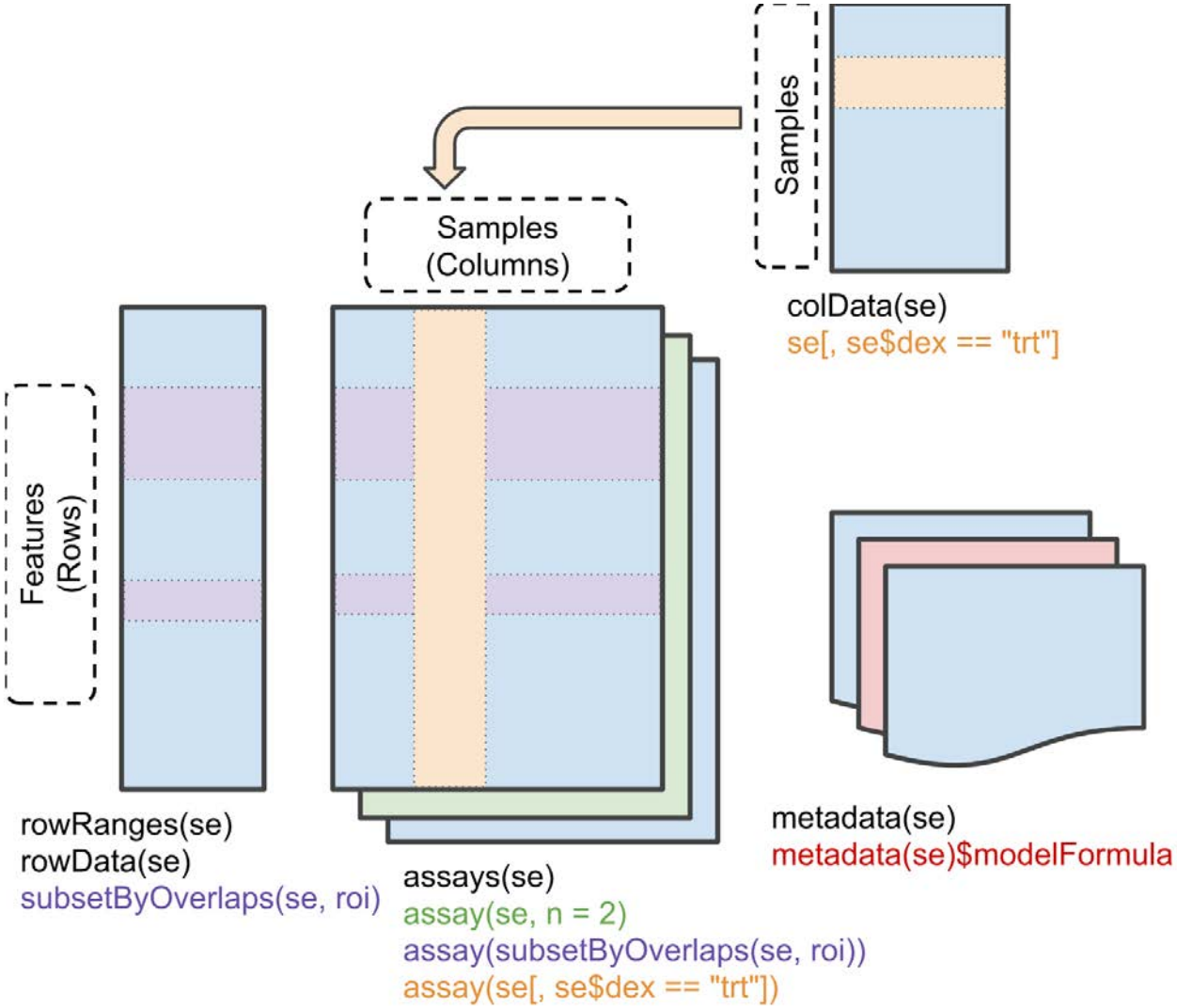
NON-OMICS data

- Lack of standardization
- High heterogeneity
- Large scale (ex. Medical Records, digital data)
- High complexity

Maturana *et al.*, Genes (2019)

Data class from *Bioconductor*

Summarized Experiment



CUNYSPH

Goal

Our project aims to build **the first large-scale, platform-independent, curated, ML-ready data repository for diverse omics and associated non-omics data**.

⬇

Two target Bioconductor data packages:

1) *curatedMetagenomicData* package (human microbiome)
   → 142 attributes collected on 22,599 samples from 93 studies

2) *cBioPortalData* packages (cancer genomics):
   → 3,733 attributes collected on 189,439 samples from 375 studies

Methods

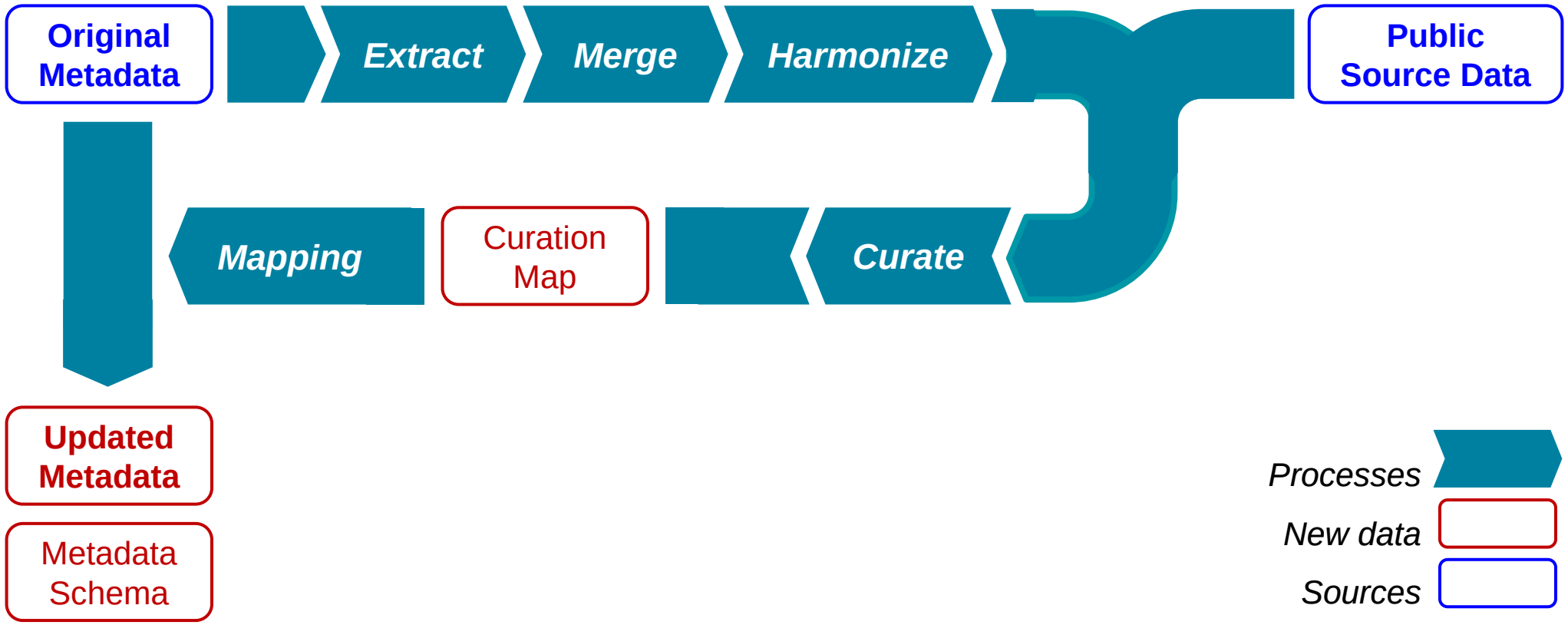
1. [Data export] datasets in R objects into a language-agnostic format
   - *hdf5* for assays, *csv* for metadata, *json* for manifest
   - New Aim for the U24 renewal

2. [Metadata harmonization] sample-level metadata curation/harmonization
   - Major variables: disease/condition, treatment, age, race/ethnicity, sex
   - Compress, consolidate, complete, and correct

3. [Resource distribution] data/harmonized metadata in language-agnostic format
   - Public (e.g., Zenodo) and commercial (e.g., GCP, Azure) for storage
   - User-facing website and API

4. [Use case] use cases of AI/ML-application
   - Cross-studies analysis
   - Analyses In different languages

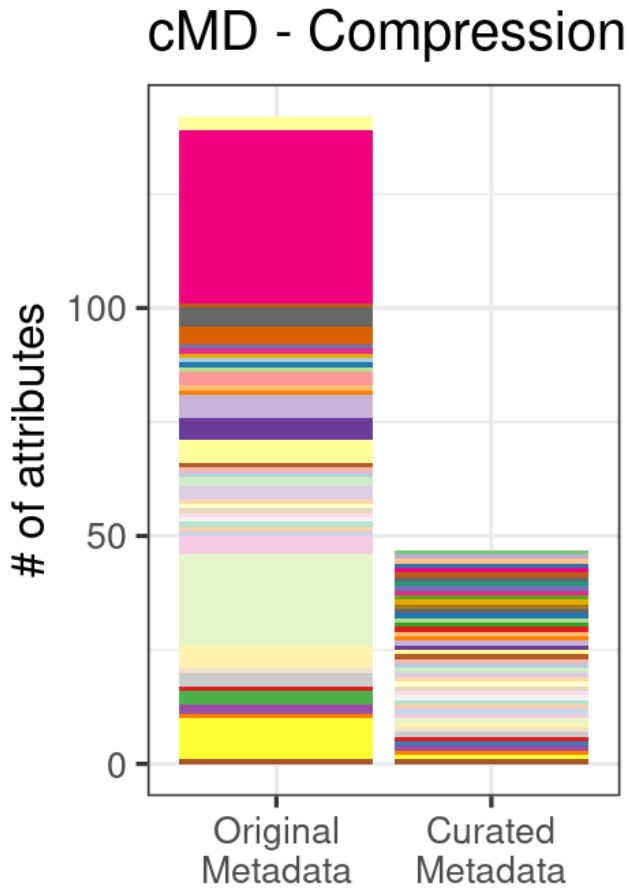
Metadata Harmonization

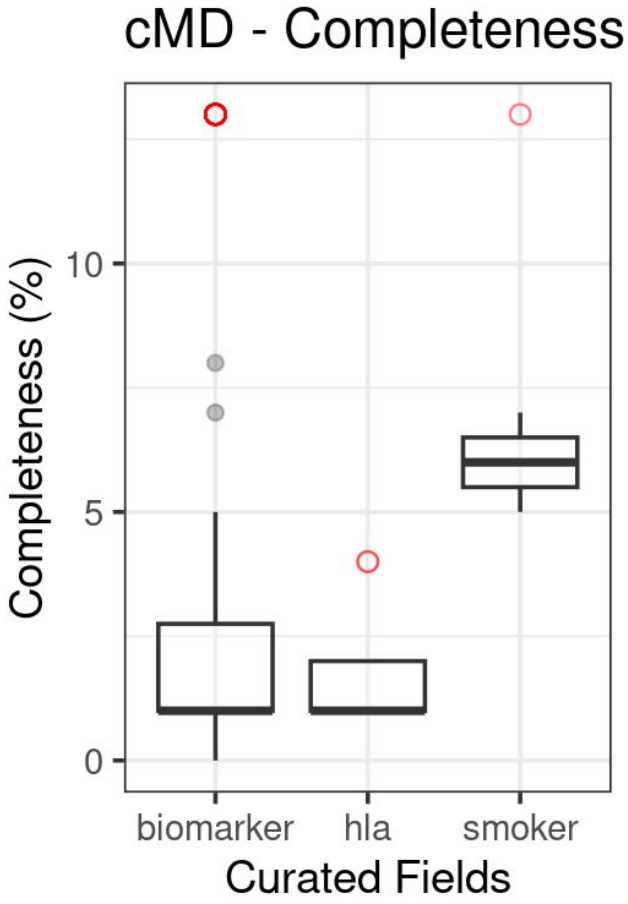


Improvement in data harmonization

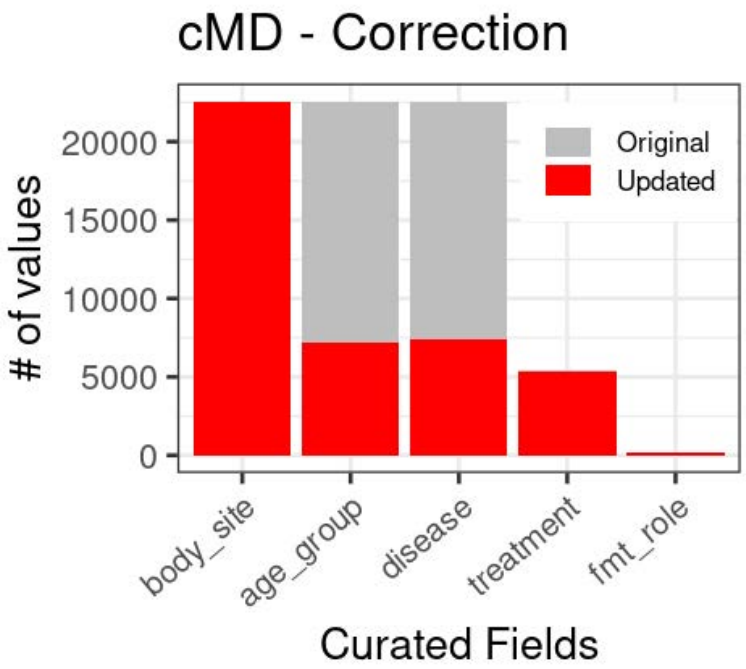
1. ***Compression***: the number of original attributes merged/combined into a new curated attribute
2. ***Consolidation***: the number of unique values for a given attribute
3. ***Completeness***: the proportion of non-missing values
4. ***Correction rate***: the proportion of the values updated during the curation

curatedMetagenomicData





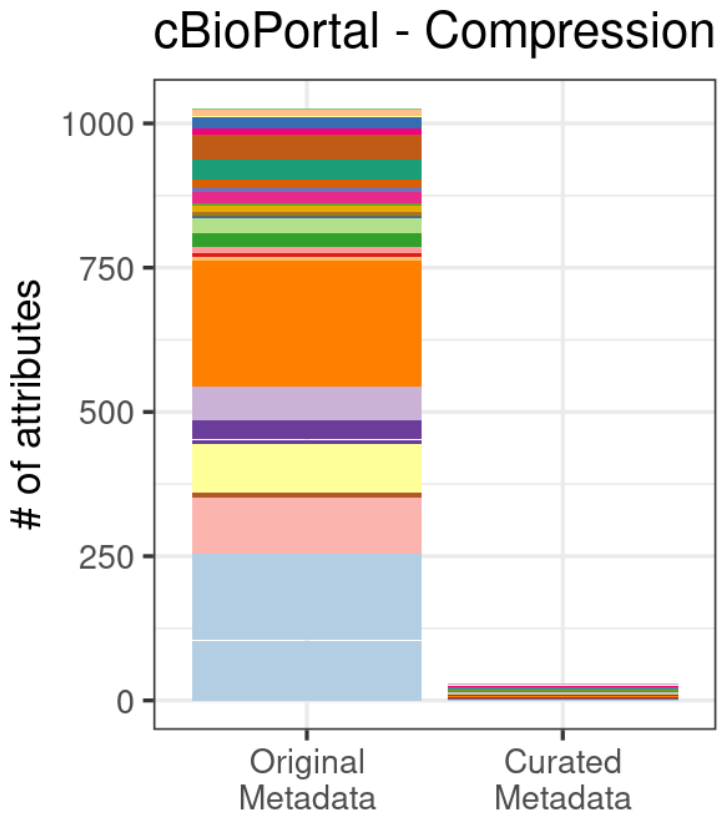
• *original* **o** *curated*

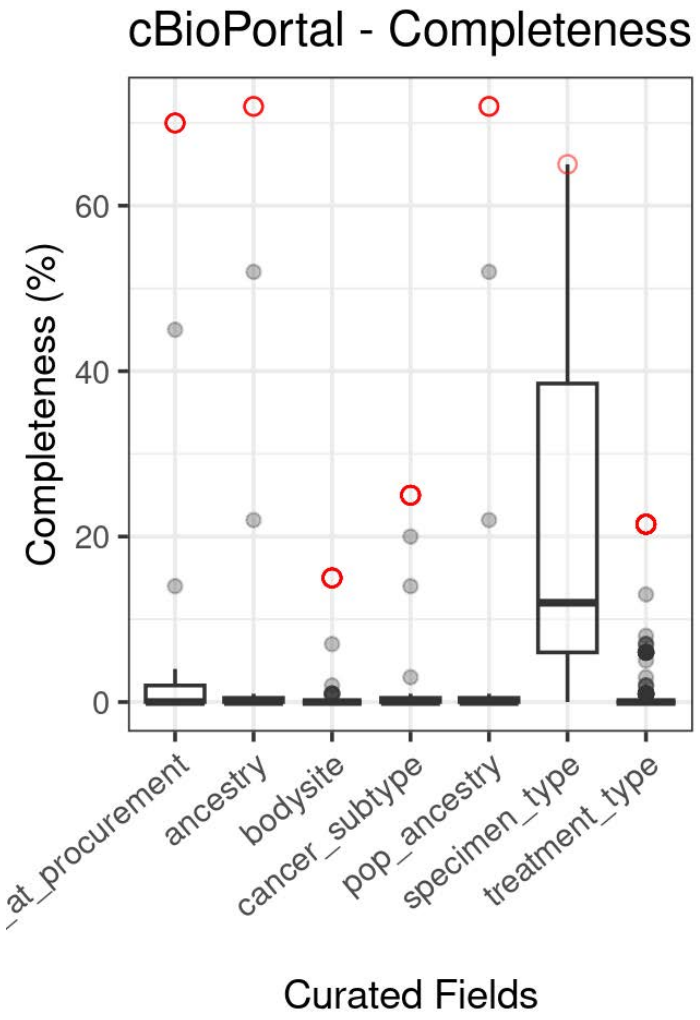


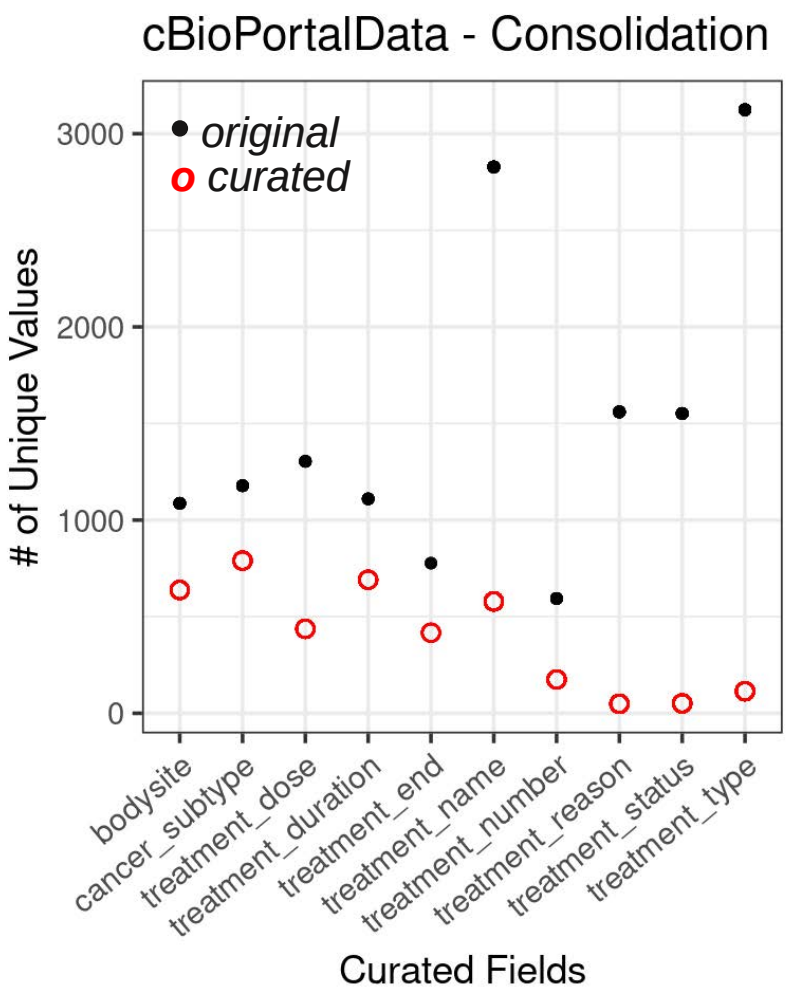
- 12 curated attributes contain completely new values
- control;disease;target_condition

CUNYSPH

cBioPortalData



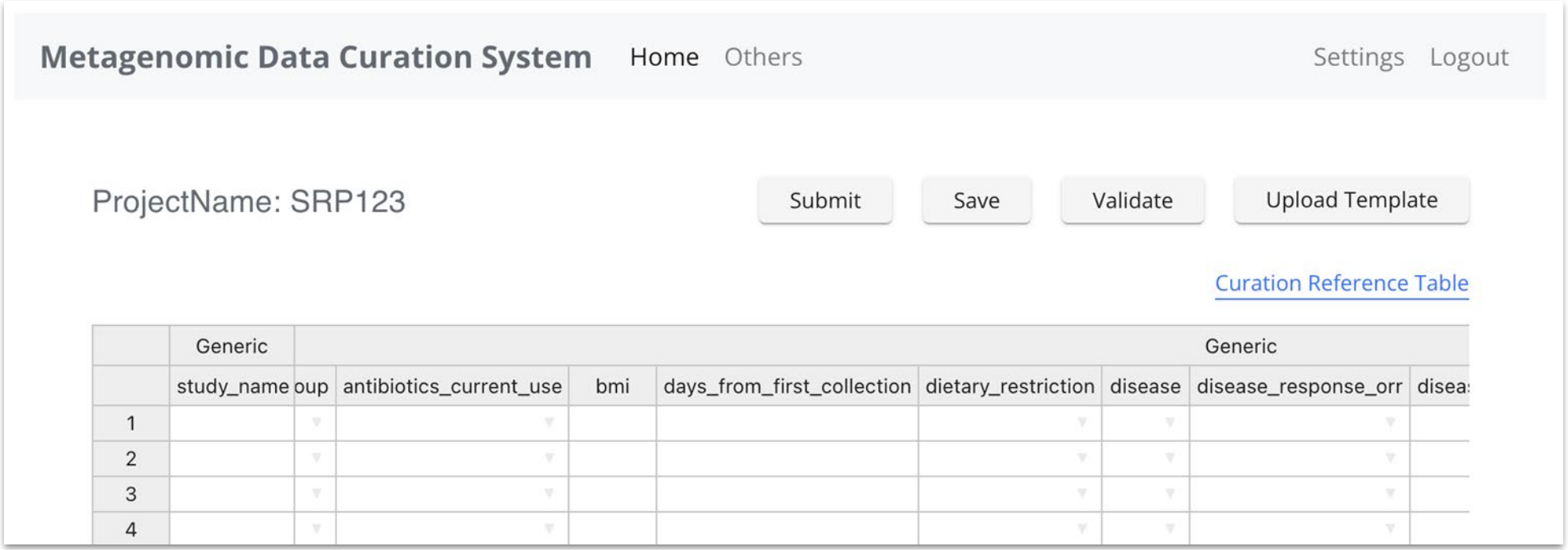




CUNYSPH

Curation Tool

- Current features: automatically fetch info from SRA, validation to the schema, use ontologies
- Ongoing features: version control, automated updates, *etc.*



CUNYSPH

Quan Wan

OmicsMLRepoR

- An R software package for users to access and manipulate metadata
- Inputs are all in non-R-specific formats, such as csv and tsv
- Major functionalities:
  1. Manipulate metadata tables: get long/short/wide/narrow forms of metadata table
  2. Metadata search through ontology tree: *searchMetadata* and *ontoTraverse*

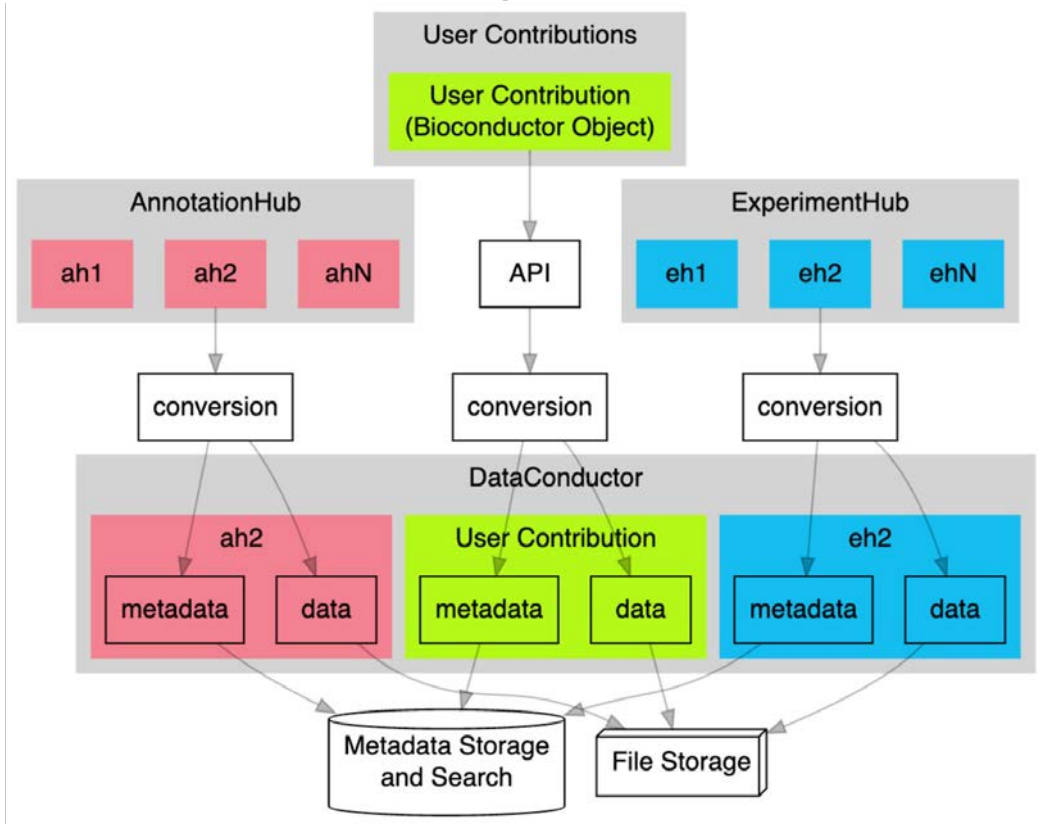
CUNYSPH

Challenges

- Low quality information
  - not covered by the existing ontologies
  - arbitrary abbreviations
  - not self-sufficient information
- Low level or metadata harmonization even some large, widely-used, curated public databases
- Maintain the quality of incoming datasets

Future Plans

1. Release the harmonized version of metadata for *curatedMetagenomicData* *cBioPortalData*
2. Release *OmicsMLRepoR* package
3. Automated metadata harmonization tool
4. Implement *DataConductor*, a unified resource for enhanced data accessibility and integration, which will be featured by:
    - Data in language-agnostic formats
    - Metadata JSON schema
    - User-friendly API and faceted data portal



CUNYSPH

Summary and Conclusions

1. We identified very heterogeneous, sparse, and inaccurate metadata from large, widely used public omics data resources and even from ‘curated’ data resource.
2. We did successful metadata harmonization and curation of two major Bioconductor resources on metagenomics and cancer genomics data.
3. We developed the user-friendly metadata schema based on the already collected information.
4. We developed the manual metadata curation framework and supporting tools.
5. Preliminary work under this Supplementary contributes to develop a new Aim for our renewal application of the parent U24.
6. It is critical to continuously improve and maintain the metadata quality to support diverse use cases of omics data, including AI/ML application.

Acknowledgements

Team

Levi Waldron, PhD

Sean Davis, MD, PhD

Sehyun Oh, PhD

Kaelyn Long, MS

Kai Gravel-Pucillo

Funding

NIH 5U24CA180996

CUNY SPH GRADUATE SCHOOL OF PUBLIC HEALTH & HEALTH POLICY

CUNY THE GREATEST URBAN UNIVERSITY IN THE WORLD