Leveraging the CEDAR Workbench for Ontology-linked Submission of Adaptive Immune Receptor Repertoire Data to the Sequence Read Archive (SRA)

Syed Ahmad Chan Bukhari[1], Martin J. O'Connor[2], John Graybeal[2], Mark A. Musen[2], Kei-Hoi Cheung[3], Steven H. Kleinstein[1]

[1]Department of Pathology, Yale School of Medicine, New Haven, CT , [2]Center for Expanded Data Annotation and Retrieval, Stanford Center for Biomedical Informatics Research, Stanford University and [3]Department of Emergency Medicine, Yale School of Medicine, New Haven, CT

Introduction

Next-generation sequencing technologies have led to a rapid production of high-throughput sequence data characterizing adaptive immune-receptor repertoires (AIRRs). As part of the AIRR community (http://airr-community.org) data standards working group, we have developed an initial set of metadata recommendations for publishing AIRR sequencing studies. These recommendations will be implemented in several public repositories, including the NCBI sequence read archive (SRA). Submissions to SRA typically use a flat-file template and include only a minimal amount of term validation. In order to ease the metadata authoring and to implement the ontological terms validation of repertoire sequence data, we are developing an interactive template through CEDAR workbench that will allow for ontological validation, and subsequent deposition in SRA. CEDAR workbench also allows the user to populate the template with metadata for data submission to various data repositories. The incorporation of template-element level ontology mapping not only facilitates validation of data submission, but also enables intelligent queries within and across repositories.

High-quality Metadata and Challenges

High-quality metadata are seen as crucial to facilitate knowledge discovery. The biomedical community has a strong history of tackling metadata challenge by driving the development of metadata templates. These templates focus on addressing the reproducibility challenge by providing detailed checklists of the metadata needed to describe particular types of experimental data sources. The key goal is to provide sufficient metadata to enable the source studies to be reproduced. While individual metadata templates can provide a standard format for a particular data source, they rarely share common structure or semantics. There is also a disconnect between the high-level checklist-based template definitions developed by scientific communities and the submission formats required by metadata repositories. Moreover, different repositories provide their locally defined templates for describing metadata. These templates lack the use of common data elements and standard vocabularies. This creates a barrier for sharing and using metadata to enable knowledge discovery. We use CEDAR workbench to create common templates for entering metadata. To enhance machine readability, we use CEDAR's capability to link individual data elements and their values to ontology concepts

The CEDAR Workbench

The Center for Expanded Data Annotation and Retrieval is studying the creation of comprehensive and expressive metadata for biomedical datasets to facilitate data discovery, data interpretation, and data reuse. CEDAR takes advantage of emerging community-based standard templates for describing different kinds of biomedical datasets. CEDAR workbench investigates the use of computational techniques to help investigators to assemble templates and to fill in their values. We are creating a repository of metadata from which we plan to identify metadata patterns that will drive predictive data entry when filling in metadata templates. The metadata repository not only will capture annotations specified when experimental datasets are initially created, but also will incorporate links to the published literature, including secondary analyses and possible refinements or retractions of experimental interpretations. By working initially with the Human Immunology Project Consortium and the developers of the ImmPort data repository, we are developing and evaluating an end-to-end solution to the problems of metadata authoring and management that will generalize to other data-management environments.

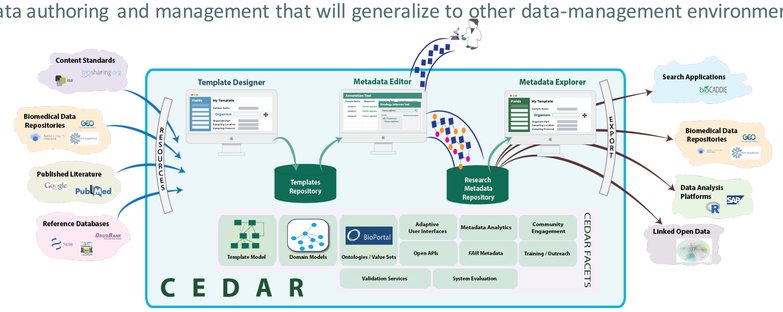

Figure 1. Metadata Life Cycle in CEDAR Workbench

Minimal Standards WG Recommendations for The AIRR Sequencing Data

As high throughput experiments become more prevalent in the field of Immunology and elsewhere, there is an increased need for collective organization of data and standardized methods of data reporting. No current standards exist for adaptive immune receptor repertoire sequencing data. Data and metadata formats need to be harmonized so that data from different experiments can be mined. Once recovered, the mined data need to have sufficient descriptive metadata in order to be useful. To fulfill these unmet needs, we propose a set of minimal standards that we recommend journals adopt and that could form the requirements for submission to a public data repository:

1. *The experimental study design* including sample data relationships (e.g., which raw data file(s) relate to which sample, which samples are technical, which are biological replicates).
2. *The essential sample annotation* including experimental factors and their values (e.g., the set of markers used to sort the cell population being studied).

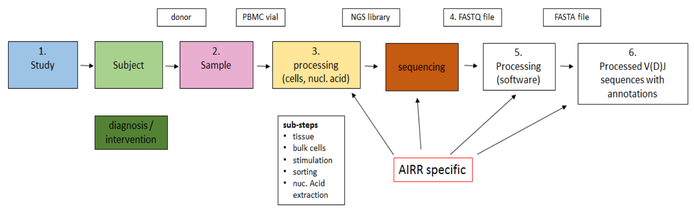

Figure 2. Overview of the six high-level principles and associated data elements that comprise the AIRR standard draft agreed to at the second annual AIRR Community meeting in 2016.

3. *Sufficient annotation of the amplicon being sequenced* that would allow the raw data to be transformed into the processed sequences (e.g., barcodes, primers, unique molecular identifiers).
4. *The raw data* for each sequencing run (e.g., FASTQ files)
5. *The essential laboratory and data processing protocols* (e.g., software tools with version numbers, quality thresholds, primer match cutoffs, etc.) that have been used to obtain the final processed data.
6. *The final processed antigen receptor sequences* for the set of samples in the experiment (e.g., the set of sequences used for V(D)J assignment), along with the V(D)J assignments for each sequence.

AIRR Data Submission to SRA Leveraging CEDAR Workbench

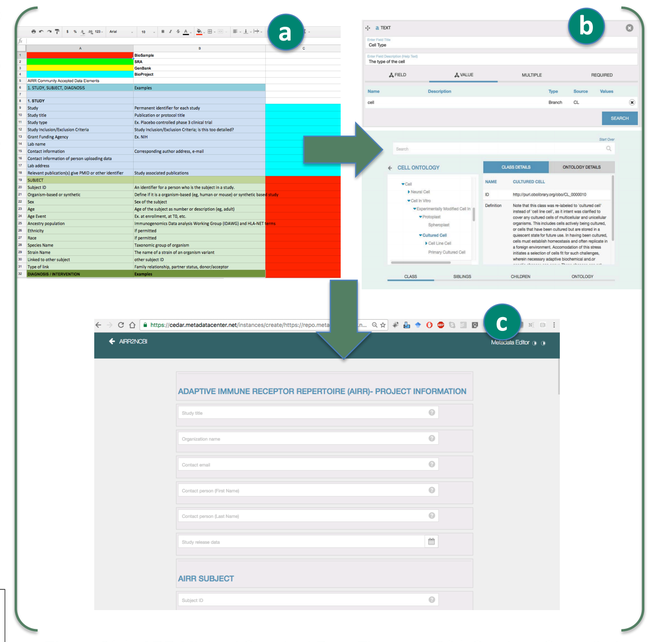

Figure 3(a). ARR Minimal Standard Data Elements, **3(b)** Ontology Controlled Template Authoring Through CEDAR Workbench and **3(c)** AIRR Data Submission Template

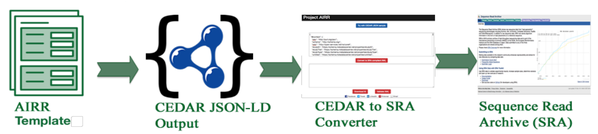

Figure 4. CEDAR Workbench to SRA Conversion Workflow

CEDAR JSON-LD to SRA XML Converter Demo

References

1- Musen, Mark A., et al. "The center for expanded data annotation and retrieval." *Journal of the American Medical Informatics Association* 22.6 (2015): 1148-1152.
2- Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway. "The sequence read archive." *Nucleic acids research* (2010): gkq1019.

Acknowledgement: We acknowledge **Dr. Ben Busby** from NCBI for his valuable suggestions during this research work.

CEDAR is supported by grant U54 AI117925 awarded by the National Institute of Allergy and Infectious Diseases through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov).