AIRR-C Standards Working Group Reporting for 2021-2022

Date of this report: April 20, 2022

SC/WG Name: Standards

SC/WG Co-leaders: Christian Busse, Jason Vander Heiden, Scott Christley

SC/WG Active Members: Felix Breden, Christian Busse, Brian Corrie, Veronique Giudicelli, Eli Harkins, Kenneth Hoehn, Susanna Marquez, Chaim Schramm, Scott Christley, Jason Vander Heiden, Ulrik Stervbo, William Lees, Kira Neller, Aditi Jain, Jingyun Li, Adrien Six, Artur Roca, Edward Lee, Marco Oliveira, Bjorn Peters, Francisco Arcila, Florian Rubelt, Katharina Imkeller, Lindsay Cowell, Nina Luning Prak, Nicole Knoetze, Enkelejda Miho

Purpose:

Develop a set of metadata standards (MiAIRR) for the submission of adaptive immune receptor repertoire sequencing (AIRR-seq) datasets. Develop standardized file formats, schemas and data field names to represent MiAIRR metadata, annotated antibody and T cell receptor sequences, and any downstream data representations. These standards are defined in formal machine-readable specifications, allowing interoperability between software from different developers.

Long-term vision and how WG products integrate with the AIRR-C mission:

The Standards WG aims to facilitate data sharing and interoperability of analysis tools within the AIRR-seq field through common data and metadata standards and documentation.

Products:

- Machine-readable, open source schema for AIRR-seq data.
  - https://github.com/airr-community/airr-standards
- Reference API libraries in R and python providing read, write and validation operations for finalized schema.
  - https://pypi.org/project/airr
  - https://cran.r-project.org/web/packages/airr
- Detailed schema and software documentation for Standards WG products and those of other WGs, along with documentation resources for public data submission and compliant community tools.
  - https://docs.airr-community.org

Progress in 2021-2022:

- Merged the Minimal Standards WG and Data Representations WG into a single Standards WG.
- Experimental germline database schema finalized, in collaboration with GLDB.

Included provisional support in the R and python libraries.

- Experimental single-cell schemas finalized.
- Experimental clonal lineage schemas finalized.
- Experimental receptor schema development is ongoing.
- Review and harmonization of AIRR terminology documents.
- Rough draft schemas for both a file manifest and aggregation of multiple repertoires.
- Various process improvements on GitHub, concerning unit tests, meeting minutes, and project management.
- Release of AIRR Standards v1.3.1 and associated python and R libraries.

Proposed plans for 2022-2023:

- The next cycle will focus primarily on refinement of experimental schemas for release in production ready versions along with a manuscript.
- Release AIRR Standards v1.4, which is scheduled to include:
    - Experimental release of the germline database schemas.
    - Experimental release of the single-cell schemas.
    - Experimental release of the receptor schema.
    - Updates to abundance fields to account for new technologies.
    - Support for additional schemas in the R and python libraries.
    - Abandonment of Python v2 support.
    - Various minor improvements to field definitions and documentation.
- Release AIRR Standards v2.0, which is scheduled to include:
    - Production release of the germline database schemas.
    - Production release of the single-cell schemas.
    - Production release of receptor schemas.
    - Production release of the lineage schemas.
    - Experimental release of a file manifest schema, repertoire grouping schema, and a persistent identifier definition.
    - Several small, but backwards incompatible changes.
- Draft a manuscript to accompany the v2.0 release describing new standards development since the original Minimal Standards (https://doi.org/10.1038/ni.3873) and Data Representations (https://doi.org/10.3389/fimmu.2018.02206) publications in 2017 and 2018, respectively.

Proposed SC/WG Co-leaders for 2022-2023: Christian Busse, Jason Vander Heiden