AIRR Formats WG

Syed Ahmad Chan, Brian Corrie, Chaim Schramm, Corey Watson, Duncan Ralph, Erick Matsen, Felix Breden, Jason Vander Heiden, Nishanth Marthandan, Susanna Marquez

Uri Laserson and Scott Christley, co-chairs
Goal

Define file formats that make it easy to analyze, share, and inter-operate with AIRR data
tl;dr

- Defined file format for VDJ assignment/rearrangement annotations.
- Defined machine-readable specification for MiAIRR schema objects and VDJ rearrangements.
  - Used in iReceptor API to integrate with VDJServer
- Setup GitHub repository with specification files, documentation, and software.
  - [github.com/airr-community/airr-standards](https://github.com/airr-community/airr-standards)
- Python reference library to read/write/validate AIRR format files.
  - Tested with ChangeO and VDJServer.
General file format considerations

- Standard format that provides ease-of-use, non-programmer accessibility, machine-readable specification, and tool interoperability.

- Enables NIH FAIR (findable, accessible, interoperable, reusable) guidelines for a broad spectrum of tools, services and APIs.

- Design for the future to support Big Data and computationally intensive analysis pipelines.
Objects that need modeling

study (MiAIRR metadata)

read (fastq)

germline (IMGT)

alignment (read to germline)

rearrangement (e.g., heavy chain)

clone
Objects that need modeling

study (MiAIRR metadata)

read (fastq)

germline (IMGT)

alignment (read to germline)

rearrangement (e.g., heavy chain)

clone
# AIRR Formats

## Alignments
- `rearrangement_id`
- `sequence`
- `segment (e.g., J)`
- `call (e.g., IGHJ6*01)`
- `score`
- `cigar`
- ...

## Rearrangements
- `rearrangement_id`
- `sequence`
- `v_call`
- `d_call`
- `j_call`
- `junction_nt`
- ...

(approx 70 fields)

Tab-delimited text files
What is a dataset?

1. YAML/JSON metadata
2. Tab-delimited data

```
my_study/my_awesome_airr_study.yaml
my_study/my_awesome_airr_study.tsv
```
<table>
<thead>
<tr>
<th>rearrangement_id</th>
<th>sequence</th>
<th>functional</th>
<th>rev_comp</th>
<th>v_call</th>
<th>d_call</th>
<th>j_call</th>
<th>junction_nt</th>
<th>junction_nt_length</th>
<th>junction_aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>015V14002075_CAF.seq.0</td>
<td>GAGTCGGCTGCTCCCTT</td>
<td>F</td>
<td></td>
<td>TRBV6-2*01,TI</td>
<td>TRBD1*01,TI</td>
<td>TRBJ1-5*01</td>
<td>TGTGCCAGCAGGCGC</td>
<td>51</td>
<td>CASSPGQAYQ</td>
</tr>
<tr>
<td>015V14002075_CAF.seq.1</td>
<td>AATGTTAGGGCTGCTCTT</td>
<td>F</td>
<td></td>
<td>TRBV5-4*01</td>
<td>TRBD2*02</td>
<td>TRBJ2-7*01</td>
<td>CASSSLRYESQ</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.2</td>
<td>CAGCCCTCAAGACCTT</td>
<td>F</td>
<td></td>
<td>TRBV12-3*01,TI</td>
<td>TRBD1*01</td>
<td>TRBJ2-3*01</td>
<td>TGTGCCAGCAGTCCG</td>
<td>51</td>
<td>CASSPQGDBTE</td>
</tr>
<tr>
<td>015V14002075_CAF.seq.3</td>
<td>TCGGCGCTGCTCCCTT</td>
<td>F</td>
<td></td>
<td>TRBV6-1*01</td>
<td>TRBD1*01</td>
<td>TRBJ1-2*01</td>
<td>TGTGCCAGCAGGAGA</td>
<td>54</td>
<td>CASSDSSGGSVS</td>
</tr>
<tr>
<td>015V14002075_CAF.seq.4</td>
<td>GACAGCCGGCTGGGCTT</td>
<td>F</td>
<td></td>
<td>TRBV7-3*01</td>
<td>TRBD1*01,TI</td>
<td>TRBJ2-7*01</td>
<td>TGTGCCAGCAGTTT</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.5</td>
<td>GCTTTGTGGTCTGGGCTT</td>
<td>F</td>
<td></td>
<td>TRBV5-6*01</td>
<td>TRBD2*01</td>
<td>TRBJ2-2*01</td>
<td>TGTGCCAGCAGCTT</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.6</td>
<td>AGGGTCTGCTCCGGCTT</td>
<td>F</td>
<td></td>
<td>TRBV6-5*01</td>
<td>TRBD1*01</td>
<td>TRBJ2-2*01</td>
<td>TGTGCCAGCAGTTT</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.7</td>
<td>ACCCTGGAGCTCGCTGCTT</td>
<td>F</td>
<td></td>
<td>TRBV25-1*01</td>
<td>TRBD2*02</td>
<td>TRBJ2-4*01</td>
<td>TGTGCCAGCAGTTT</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.8</td>
<td>GTGAGCAACATGAGCTT</td>
<td>F</td>
<td></td>
<td>TRBV29-1*01,TI</td>
<td>TRBD1*01</td>
<td>TRBJ2-7*01</td>
<td>TGCAGCGTGTGGAG</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.9</td>
<td>GAGTCGGCTGCTCCCTT</td>
<td>F</td>
<td></td>
<td>TRBV6-1*01</td>
<td>TRBD2*02</td>
<td>TRBJ2-7*01</td>
<td>TGTGCCAGCAGTTT</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.10</td>
<td>CCTCCGACAGCTCGCTGCTT</td>
<td>F</td>
<td></td>
<td>TRBV14*01</td>
<td>TRBD1*01</td>
<td>TRBJ1-6*02</td>
<td>TGTGCCAGCAGCA</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.11</td>
<td>CTGCTCTCCCTCCAAA</td>
<td>F</td>
<td></td>
<td>TRBV6-2*01,TI</td>
<td>TRBD2*01</td>
<td>TRBJ2-3*01</td>
<td>TGTGCCAGCAGTTT</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.12</td>
<td>CTGCTCTCCCTCCAAA</td>
<td>F</td>
<td></td>
<td>TRBV28*01</td>
<td>TRBD2*02</td>
<td>TRBJ2-1*01</td>
<td>TGTGCCAGCAGTTT</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.13</td>
<td>GAGTCGGCTGCTCCCTT</td>
<td>F</td>
<td></td>
<td>TRBV6-2*01,TI</td>
<td>TRBD2*02</td>
<td>TRBJ1-5*01</td>
<td>TGTGCCAGCAGTTT</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.14</td>
<td>ATTCAGCAGCTCGTTTTT</td>
<td>F</td>
<td></td>
<td>TRBV11-2*01,TI</td>
<td>TRBD1*01</td>
<td>TRBJ2-2*01</td>
<td>TGTGCCAGCAGCA0</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.15</td>
<td>CCTGAACTGCTGCTGCTT</td>
<td>F</td>
<td></td>
<td>TRBV11-2*01,TI</td>
<td>TRBD2*02</td>
<td>TRBJ2-7*01</td>
<td>TGTGCCAGCAGTTT</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.16</td>
<td>TCACAGGGTGCTGCTT</td>
<td>F</td>
<td></td>
<td>TRBV15*02</td>
<td>TRBD1*01</td>
<td>TRBJ1-5*01</td>
<td>TGTGCCAGCAGGAGA</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.17</td>
<td>CTACACACCTGCGCTT</td>
<td>F</td>
<td></td>
<td>TRBV4-3*01</td>
<td>TRBD1*01,TI</td>
<td>TRBJ1-2*01</td>
<td>TGCACCAGCAGCCC</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>015V14002075_CAF.seq.18</td>
<td>TGCAGCCAGAAGACTT</td>
<td>F</td>
<td></td>
<td>TRBV4-1*01,TI</td>
<td>TRBD1*01</td>
<td>TRBJ2-5*01</td>
<td>GCTAGCGTGGGAGA</td>
<td>31</td>
<td></td>
</tr>
</tbody>
</table>
### All AIRR rearrangement format mandatory fields

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Mandatory</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>rearrangement_id</td>
<td>string</td>
<td>mandatory</td>
<td>Read/sequence identifier; often identical to a read identifier, but not necessarily (especially where a rearrangement is derived from a multiple read consensus).</td>
</tr>
<tr>
<td>sequence</td>
<td>string</td>
<td>mandatory</td>
<td>Nucleotide sequence (e.g., the &quot;read&quot; sequence; revcomp'd if necessary)</td>
</tr>
<tr>
<td>sample_id</td>
<td>string</td>
<td>mandatory</td>
<td>The biological sample this read derives from (e.g., from BioSample database)</td>
</tr>
<tr>
<td>functional</td>
<td>boolean</td>
<td>mandatory</td>
<td>VDJ sequence is predicted to be functional</td>
</tr>
<tr>
<td>rev_comp</td>
<td>boolean</td>
<td>mandatory</td>
<td>Sequence is reverse complemented</td>
</tr>
<tr>
<td>v_call</td>
<td>string</td>
<td>mandatory</td>
<td>V allele assignment</td>
</tr>
<tr>
<td>d_call</td>
<td>string</td>
<td>mandatory</td>
<td>D allele assignment</td>
</tr>
<tr>
<td>j_call</td>
<td>string</td>
<td>mandatory</td>
<td>J allele assignment</td>
</tr>
<tr>
<td>c_call</td>
<td>string</td>
<td>mandatory</td>
<td>C gene assignment (e.g., IGH4, IGHA2, IGHE, TRBC)</td>
</tr>
<tr>
<td>junction_nt</td>
<td>string</td>
<td>mandatory</td>
<td>Nucleotide sequence of the junction region (CDR3 plus conserved residues; i.e., IMGT's JUNCTION)</td>
</tr>
<tr>
<td>junction_nt_length</td>
<td>integer</td>
<td>mandatory</td>
<td>Number of junction nucleotides in sequence_vdj</td>
</tr>
<tr>
<td>junction_aa</td>
<td>string</td>
<td>mandatory</td>
<td>Amino acid sequence of the junction region (CDR3 plus conserved residues; i.e., IMGT's JUNCTION)</td>
</tr>
<tr>
<td>junction_aa_length</td>
<td>integer</td>
<td>mandatory</td>
<td>Number of junction amino acids in sequence_vdj</td>
</tr>
<tr>
<td>v_score</td>
<td>number</td>
<td>mandatory</td>
<td>V alignment score</td>
</tr>
<tr>
<td>d_score</td>
<td>number</td>
<td>mandatory</td>
<td>D alignment score</td>
</tr>
<tr>
<td>j_score</td>
<td>number</td>
<td>mandatory</td>
<td>J alignment score</td>
</tr>
<tr>
<td>c_score</td>
<td>number</td>
<td>mandatory</td>
<td>C alignment score</td>
</tr>
<tr>
<td>v_cigar</td>
<td>string</td>
<td>mandatory</td>
<td>V alignment CIGAR string</td>
</tr>
<tr>
<td>d_cigar</td>
<td>string</td>
<td>mandatory</td>
<td>D alignment CIGAR string</td>
</tr>
<tr>
<td>j_cigar</td>
<td>string</td>
<td>mandatory</td>
<td>J alignment CIGAR string</td>
</tr>
<tr>
<td>c_cigar</td>
<td>string</td>
<td>mandatory</td>
<td>C alignment CIGAR string</td>
</tr>
</tbody>
</table>

### Study/Sample Metadata
- IMGT annotations
- Change-O clones
- PARTIS clonal lineage

### Study/Sample Metadata
- IgBlast annotations
- MiXCR clones
- IgPhyML clonal lineage

### AIRR compliant
- Repository/Tool 1
- Repository/Tool 2

### AIRR
- YAML/JASON/Tab-delimited via API

### AIRR clones format?
- AIRR lineage format?

---

**IMGT: “AA Junction”**
- MiXCR: “aaSeqCDR3”
- IgBlast: “CDR3”

**AIRR: “junction_aa”**
- AIRR clones format?
- AIRR lineage format?
MiAIRR: AIRR Community Minimal Standards WG  [http://airr-community.org](http://airr-community.org)
Machine-readable specs

- Interoperability
- Continuous-integration
- Automation
- Minimizing ambiguity
AIRR Formats Working Group

Working environment for the AIRR-Formats subgroup.

Projects

- Create a standardized file format for V(D)J rearrangement data.
- Investigate meta data guidelines.
- Investigate standardization of detailed clonal clustering data.
- Investigate standardization of lineage tree data.

Built with MkDocs using a theme provided by Read the Docs.
airr-community / airr-standards

Current Branches Build History Pull Requests

- master  Update README.md  #77 passed  1 min 22 sec
  - Ahmad Syed
- master  Merge remote-tracking branch 'airr-
  - Uri Laserson
- master  Merge pull request #52 from airr-  #75 passed  1 min 3 sec
  - Scott Christley
- master  Merge branch 'master' into 51-consi
  - Scott Christley
- master  Merge pull request #50 from airr-  #72 passed  50 sec
  - Scott Christley
- 51-consi  Refactor consistency check
  - Uri Laserson
- 5-types  Changed data types in table to Open
  - #68 passed  56 sec
Goals for 2018

- Submit manuscript to publicize format
- Develop format for representing clones
- Finish integration of GitHub repository with MiAIRR
- Finish specifying metadata file format
- Public release of reference library to read/write/validate AIRR format files.
  - Initially targeting Python and R
- Releasing documentation incl. example output/use.
"Data Modeling and Representation WG"

- Multiple WGs are designing implementation standards and could use technical input on data representation.
- Coordination with AIRR Working Groups to specify data models, e.g.,
  - Common Repo defining minimal APIs for repositories and REST resources
  - MinStd choosing ontologies for their fields
  - Germline defining new germlines and annotations
- Ensure all AIRR groups are working in mutually compatible ways (in terms of data)
  - Ensure we have liaisons on all other relevant working groups
- Work on representation of provenance of data sets
Thanks to Formats WG!

New contributions and members are welcome!

docs.airr-community.org
General file format considerations

- Ease-of-use
  - Standard container files
    - Commonly available parsers
    - e.g., JSON, CSV, XML
  - Argue over schema, not representation
- Non-programmer accessibility
  - Tabular, non-nested data (Excel-compat)
  - Non-binary data
- Big data
  - Splittability
  - Operate on directories of files
- Metadata
AIRR format overview

- Format
  - Tab-delimited text for data
  - YAML/JSON for metadata
- Coordinates are 0-based with half-open intervals (like Python)
- Data types correspond to OpenAPI spec
- Columns are optional but use AIRR-spec'd names when relevant
  - Please suggest useful columns!
- CIGAR format for alignments