
airr-standards Documentation

Release 1.4

AIRR Community

Feb 14, 2024

CONTENTS

1	Introduction to the AIRR Standards	3
2	Table of Contents	5
2.1	Getting Started	5
2.2	Release Notes	15
2.3	AIRR Standards	24
2.4	Data Submission and Query	163
2.5	Software	168
2.6	Community Resources	211
2.7	Appendix A: Key Terms	212
2.8	References	214
	Bibliography	215
	Index	217

The Adaptive Immune Receptor Repertoire (AIRR) Community of The Antibody Society is a research-driven group that is organizing and coordinating stakeholders in the use of next-generation sequencing (NGS) technologies to study antibody/B-cell and T-cell receptor repertoires. Recent advances in sequencing technology have made it possible to sample the immune repertoire in exquisite detail. AIRR sequencing has enormous promise for understanding the dynamics of the immune repertoire in vaccinology, infectious disease, autoimmunity, and cancer biology, but also poses substantial challenges. The AIRR Community was established to meet these challenges.

INTRODUCTION TO THE AIRR STANDARDS

The AIRR Community is developing a set of standards for describing, reporting, storing, and sharing adaptive immune receptor repertoire (AIRR) data, such as sequences of antibodies and T cell receptors (TCRs). Some specific efforts include:

- The MiAIRR standard for describing minimal information about AIRR datasets, including sample collection and data processing information.
- Data submission guidelines and workflows.
- Data representations (file format) specifications for storing large amounts of annotated AIRR data.
- API to query and download AIRR data from repositories/databases as part of the AIRR Data Commons.
- A community standard for software tools which will allow conforming tools to gain community recognition.
- Set of reference software tools for reading, writing and validating data in the AIRR standards.
- A database and web submission frontend for inferred germline genes

TABLE OF CONTENTS

2.1 Getting Started

This website provides information and resources regarding the AIRR Community Standards for the diverse community of immunology researchers, bioinformaticians, and software developers.

2.1.1 MiAIRR standard for study data submission

- Gather experimental and analysis information about your study to conform to the *MiAIRR* standard (minimal information about adaptive immune receptor repertoires).
- *Submission* of your study data to a public repository.

2.1.2 AIRR Data Commons for query and download of AIRR-seq data

- *Query* publicly available AIRR-seq studies in the *AIRR Data Commons*.

2.1.3 Resources related to data representations and software development

- Schema, definitions and file formats for the *AIRR Data Model*. The AIRR Data Model defines the structure and relationship for the MiAIRR data elements.
- *Software guidelines* for tools developers to enable rigorous and reproducible immune repertoire research.
- *AIRR Data Commons API* provides programmatic access to query and download AIRR-seq data.

2.1.4 Software tools and libraries

- *Python reference library* for reading/writing/validating AIRR data files.
- *R reference library* for reading/writing/validating AIRR data files.
- *ADC API reference implementation* for a local data repository.
- *Resources and tools* that support the AIRR Standards.

2.1.5 Tutorials, examples and workflows

AIRR Rearrangement TSV Interoperability Example

The example that follows illustrates the interoperability provided by the AIRR Rearrangement schema. The code provided demonstrates how to take AIRR formatted data output by IgBLAST and combine it with data processed by IMGT/HighV-QUEST that has converted to the AIRR format by Change-O. Then, the merged output of these two distinct tools is used to (a) create MiAIRR compliant GenBank/TLS submission files, and (b) perform a simple V gene usage analysis task.

Data

We've hosted a small set of example data from BioProject PRJNA338795 (Vander Heiden et al, 2017. J Immunol.) containing both input and output of the example. It may be downloaded from:

[Example Data](#)

Walkthrough

Environment setup

We'll use the Immcantation docker image for this example, which comes loaded with all the tools used in the steps that follow:

```
# Download the image
docker pull kleinstein/immcantation:devel

# Invoke a shell session inside the Immcantation docker image
# Map example data (~/.data) to the container's /data directory
$> docker run -it -v ~/.data:/data:z kleinstein/immcantation:devel bash
```

Generate AIRR formatted TSV files

TSV files compliant with the AIRR Rearrangement schema may be output directly from IgBLAST v1.9+ or generated from IMGT/HighV-QUEST output (or IgBLAST <=1.8 output) using the MakeDb parser provided by Change-O:

```
# Generate TSV directly with IgBLAST
$> cd /data
$> export IGDATA=/usr/local/share/igblast
$> igblastn -query HD13M.fasta -out HD13M_fmt19.tsv -outfmt 19 \
    -germline_db_V $IGDATA/database/imgt_human_ig_v \
    -germline_db_D $IGDATA/database/imgt_human_ig_d \
    -germline_db_J $IGDATA/database/imgt_human_ig_j \
    -auxiliary_data $IGDATA/optional_file/human_gl.aux \
    -ig_seqtype Ig -organism human \
    -domain_system imgt

# Generate TSV from IMGT/HighV-QUEST results using changeo:MakeDb
$> MakeDb.py imgt -i HD13N_imgt.txz -s HD13N.fasta \
    --scores --partial --format airr
```

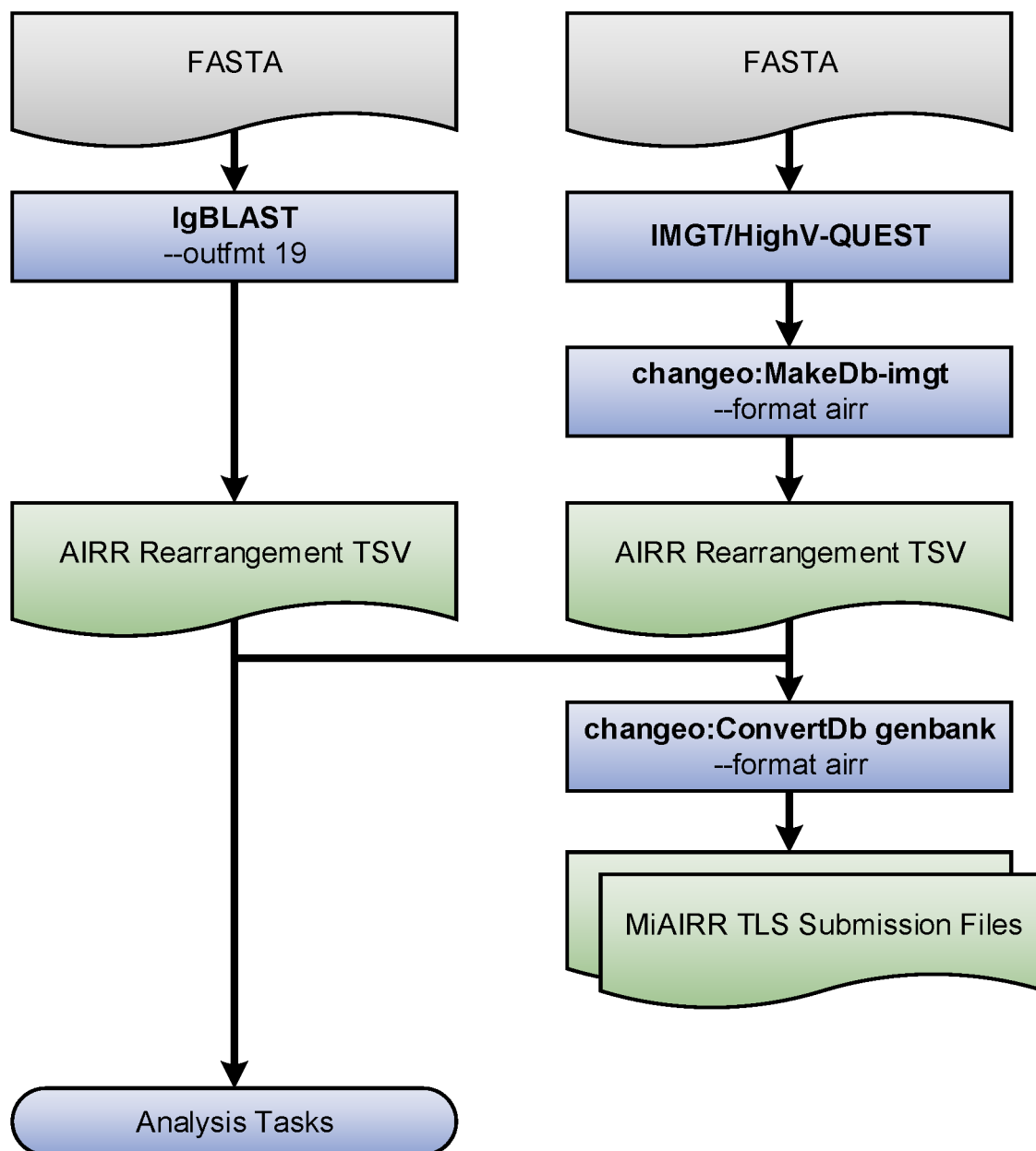


Fig. 1: Flowchart of the example steps.

Generate GenBank/TLS submission files

AIRR TSV files can be input directly in Change-O's ConvertDb-genbank tool to generate MiAIRR compliant files for submission to GenBank/TLS:

```
# Generate ASN files from IgBLAST output
$> ConvertDb.py genbank -d HD13M_fmt7_db-pass.tsv --format airr \
    --inf IgBLAST:1.7.0 --organism "Homo sapiens" \
    --tissue "Peripheral blood" --cell "naive B cell" \
    --id --asn -sbt HD13M.sbt

# Generate ASN files from IMGT/HighV-QUEST output
$> ConvertDb.py genbank -d HD13N_imgt_db-pass.tsv --format airr \
    --inf IMGT/HighV-QUEST:1.5.7.1 --organism "Homo sapiens" \
    --tissue "peripheral blood" --cell "naive B cell" \
    --cregion c_call --id --asn -sbt HD13M.sbt
```

Merge files and count V family usage

AIRR TSV files from different tools and easy combined to perform analysis on data generated using different software. Below is shown a simple V family usage analysis after merging the IgBLAST and IMGT/HighV-QUEST outputs into a single table:

```
# Count V family usage in R
# Imports
$> R
R> library(alakazam)
R> library(dplyr)
R> library(ggplot2)

# Merge IgBLAST and IMGT/HighV-QUEST results
R> db_m <- read.delim("HD13M_fmt7_db-pass.tsv")
R> db_n <- read.delim("HD13N_imgt_db-pass.tsv")
R> db_m$cell_type <- "memory"
R> db_n$cell_type <- "naive"
R> db <- bind_rows(db_m, db_n)

# Subset to heavy chain
R> db <- subset(db, grepl("IGH", v_call))

# Count combined V gene usage
R> v_usage <- countGenes(db, "v_call", groups="cell_type",
    mode="family")

# Plot V family usage
R> ggplot(v_usage, aes(x=GENE, y=SEQ_FREQ, fill=cell_type)) +
    geom_col(position="dodge") +
    scale_fill_brewer(name="Cell type", palette="Set1") +
    xlab("") +
    ylab("Fraction of repertoire")
```

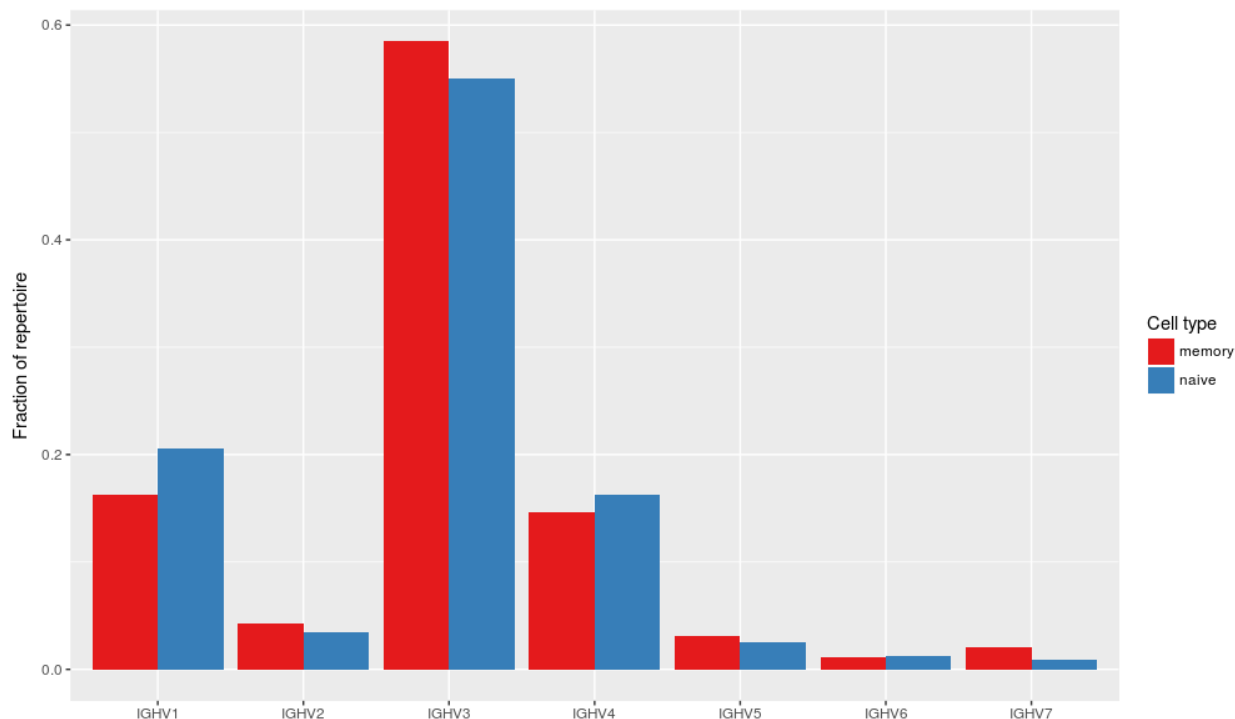


Fig. 2: V family usage for the combined data set.

ADC API Query and Analysis Example

This example shows how repertoires and associated rearrangments may be queried from a data repository using the ADC API and then a simple analysis is performed. The example is split between two python scripts; one that performs the query and saves the data into files, and another that reads the data from the files and generates a grouped CDR3 amino acid length distribution plot. The two scripts could be combined into one, but this example illustrates how the data can be saved into files for later use. The example uses the AIRR standards python library.

Data

This example retrieves data for the following study, which is identified by NCBI BioProject PRJNA300878. In this example, we are only going to query and retrieve the T cell repertoires.

Rubelt, F. et al., 2016. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naive and antigen-experienced cells. Nature communications, 7, p.11112.

Basic study description:

- 5 pairs of human twins
- B-cells and T-cells sequenced
- B-cells sorted into naive and memory
- T-cells sorted into naive CD4, naive CD8, memory CD4 and memory CD8
- Total of 60 repertoires: 20 B-cell repertoires and 40 T-cell repertoires

Walkthrough

We'll use the airr-standards docker image for this example, which comes loaded with all the python packages needed. You will want to map a local directory inside the docker container so you can access the data and analysis results afterwards:

```
# Download the image
docker pull airrc/airr-standards:latest

# Make local temporary directory to hold the data
mkdir adc_example
cd adc_example

# Invoke a shell session inside the docker image
docker run -it -v $PWD:/data airrc/airr-standards:latest bash
```

The first python script queries the data from the VDJServer data repository and saves them into files:

```
# Query the data
cd /data
python3 /airr-standards/docs/examples/api/retrieve_data.py
```

Only a subset of the data is downloaded for illustration purposes, but review the code to see how all data can be downloaded. A total of 40 repertoires and 300,178 rearrangements should be downloaded. The repertoire metadata is saved in the `repertoires.airr.json` file, and the rearrangements are saved in the `rearrangements.tsv` file. The script should take a few minutes to run and produce the following display messages:

```
Info: VDJServer Community Data Portal
version: 1.3
description: VDJServer ADC API response for repertoire query
Received 40 repertoires.
Retrieving rearrangements for repertoire: 5168912186246295065-242ac11c-0001-012
Retrieved 9768 rearrangements for repertoire: 5168912186246295065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 5338391595746455065-242ac11c-0001-012
Retrieved 5521 rearrangements for repertoire: 5338391595746455065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 4858300151399575065-242ac11c-0001-012
Retrieved 2885 rearrangements for repertoire: 4858300151399575065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 5039977268020375065-242ac11c-0001-012
Retrieved 4053 rearrangements for repertoire: 5039977268020375065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 6240077029868695065-242ac11c-0001-012
Retrieved 3506 rearrangements for repertoire: 6240077029868695065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 6389112395039895065-242ac11c-0001-012
Retrieved 2289 rearrangements for repertoire: 6389112395039895065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 5939858815878295065-242ac11c-0001-012
Retrieved 3637 rearrangements for repertoire: 5939858815878295065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 6088937130722455065-242ac11c-0001-012
Retrieved 9028 rearrangements for repertoire: 6088937130722455065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 7446748091679895065-242ac11c-0001-012
Retrieved 1540 rearrangements for repertoire: 7446748091679895065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 7591789137265815065-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 7591789137265815065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 7066128089908375065-242ac11c-0001-012
Retrieved 5662 rearrangements for repertoire: 7066128089908375065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 5624006920930455065-242ac11c-0001-012
```

(continues on next page)

(continued from previous page)

Retrieved 10000 rearrangements for repertoire: 5624006920930455065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 8961797805343895065-242ac11c-0001-012
Retrieved 1179 rearrangements for repertoire: 8961797805343895065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 9084118473933975065-242ac11c-0001-012
Retrieved 4464 rearrangements for repertoire: 9084118473933975065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 8485700680582295065-242ac11c-0001-012
Retrieved 3908 rearrangements for repertoire: 8485700680582295065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 7309695685264535065-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 7309695685264535065-242ac11c-0001-012
Retrieving rearrangements for repertoire: 8425807333172056551-242ac11c-0001-012
Retrieved 6863 rearrangements for repertoire: 8425807333172056551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 8263242821018456551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 8263242821018456551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 8733756488295256551-242ac11c-0001-012
Retrieved 5298 rearrangements for repertoire: 8733756488295256551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 8602072790999896551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 8602072790999896551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 7313153105470296551-242ac11c-0001-012
Retrieved 9121 rearrangements for repertoire: 7313153105470296551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 6964444710708056551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 6964444710708056551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 7640859110155096551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 7640859110155096551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 7461458326201176551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 7461458326201176551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 5953881855632216551-242ac11c-0001-012
Retrieved 5916 rearrangements for repertoire: 5953881855632216551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 7158276584776536551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 7158276584776536551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 6393557657723736551-242ac11c-0001-012
Retrieved 7257 rearrangements for repertoire: 6393557657723736551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 6205695788196696551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 6205695788196696551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 4476756703191896551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 4476756703191896551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 4357957907784536551-242ac11c-0001-012
Retrieved 7033 rearrangements for repertoire: 4357957907784536551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 4931851437876056551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 4931851437876056551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 4744762662462296551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 4744762662462296551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 3252733973504856551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 3252733973504856551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 2989624276951896551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 2989624276951896551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 3628844259615576551-242ac11c-0001-012
Retrieved 5208 rearrangements for repertoire: 3628844259615576551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 3438706057421656551-242ac11c-0001-012
Retrieved 9530 rearrangements for repertoire: 3438706057421656551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 2197374609531736551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 2197374609531736551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 1993707260355416551-242ac11c-0001-012

(continues on next page)

(continued from previous page)

```

Retrieved 10000 rearrangements for repertoire: 1993707260355416551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 2541616238306136551-242ac11c-0001-012
Retrieved 6512 rearrangements for repertoire: 2541616238306136551-242ac11c-0001-012
Retrieving rearrangements for repertoire: 2366080924918616551-242ac11c-0001-012
Retrieved 10000 rearrangements for repertoire: 2366080924918616551-242ac11c-0001-012

```

The second python script loads the data from the files and generates a CDR3 amino acid length distribution that is grouped by the T cell subset. This study performs flow sorting to generate four T cell subsets: naive CD4+, naive CD8+, memory CD4+, memory CD8+. The script uses the repertoire metadata to determine the T cell subset for the rearrangement, tabulates the counts, normalizes them, and generates a grouped bar chart with the results:

```

# Run the analysis
python3 /airr-standards/docs/examples/api/analyze_data.py

```

The figure is placed in the `plot.png` file and should look like this:

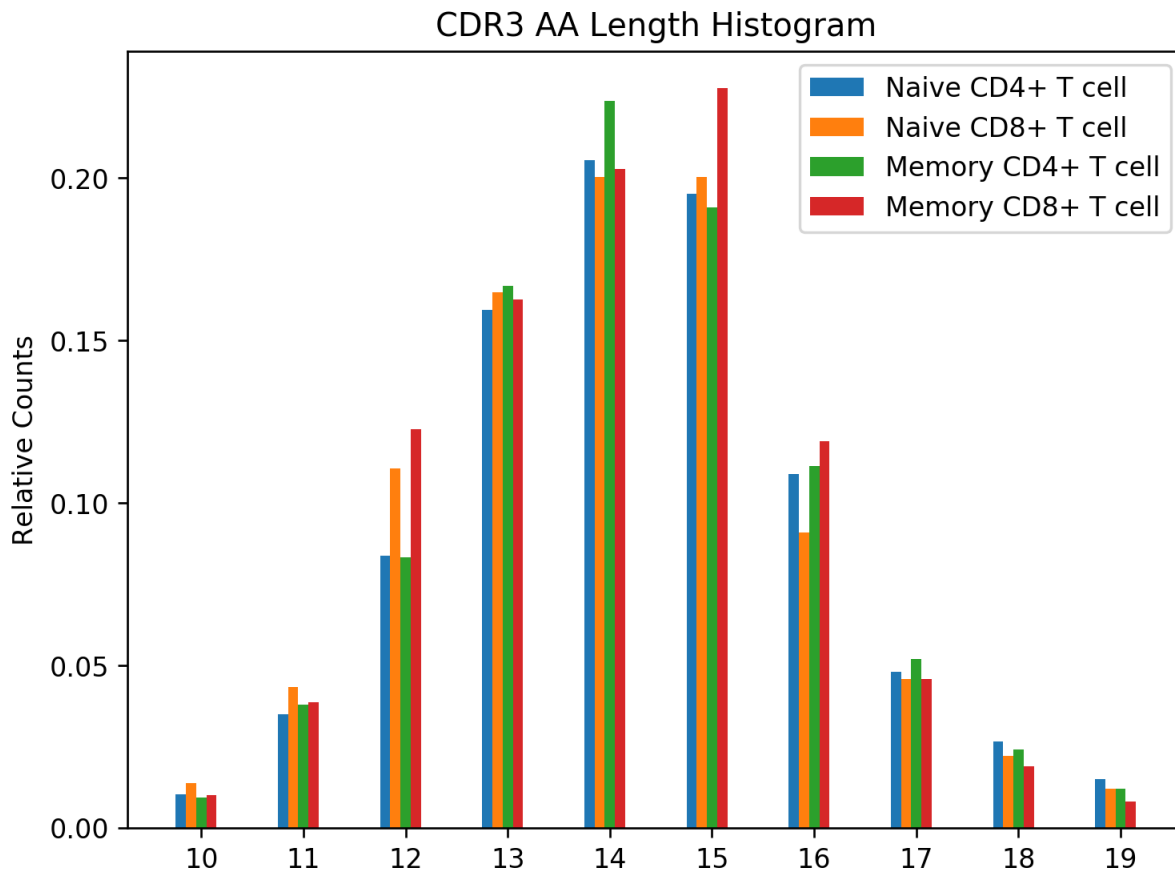


Fig. 3: CDR3 AA Length Histogram grouped by T cell subsets.

Scientific Query Scenarios for AIRR Data Commons API

The AIRR Common Repository Working Group (CRWG) has defined a number of sample scientific query scenarios to guide the design of the ADC API. The [Design Decisions](#) document lists the major design choices for the API, and the [API](#) is currently defined using the [OpenAPI V2.0 Specification](#). This document describes the query examples with associated JSON definitions that can be submitted to an AIRR repository.

There are two main query endpoints in the API: /repertoire for querying MiAIRR-compliant study metadata and /rearrangement for querying rearrangement annotations. Most scientific queries will involve both endpoints. The basic workflow involves first querying /repertoire to get the list of repertoires that meet the search criteria on study, subject, and sample metadata. Secondly, the identifiers from the repertoires in the first query are passed to the /rearrangement endpoint along with any search criteria on the rearrangement annotations. The resultant rearrangements can be downloaded as JSON or in the [AIRR TSV format](#).

Query Example

What human full length TCR-beta sequences have junction amino acid sequence: “CASSYIKLN”?

- The JSON query definition for /repertoire endpoint. The ontology identifier 9606 requests human and TRB is the locus of interest.

```
{
  "filters": {
    "op": "and",
    "content": [
      {
        "op": "=",
        "content": {
          "field": "subject.organism.id",
          "value": "9606"
        }
      },
      {
        "op": "=",
        "content": {
          "field": "sample.pcr_target.pcr_target_locus",
          "value": "TRB"
        }
      }
    ]
  }
}
```

- That query does not request full length sequences. We can enhance the query by adding a clause for the sample.complete_sequences field.

```
{
  "filters": {
    "op": "and",
    "content": [
      {
        "op": "=",
        "content": {
          "field": "subject.species.id",
```

(continues on next page)

(continued from previous page)

```

        "value": "NCBITAXON:9606"
    },
    {
        "op": "=",
        "content": {
            "field": "sample.pcr_target.pcr_target_locus",
            "value": "TRB"
        }
    },
    {
        "op": "or",
        "content": [{
            "op": "=",
            "content": {
                "field": "sample.complete_sequences",
                "value": "complete"
            }
        },
        {
            "op": "=",
            "content": {
                "field": "sample.complete_sequences",
                "value": "complete+untemplated"
            }
        }
    ]
}

```

- The JSON query definition for /rearrangement endpoint. The repertoire identifiers (repertoire_id) in the query are just examples, you would replace them with the actual identifiers returned from the above repertoire query. The query performs an exact match of the junction amino acid sequence.

```

{
  "filters": {
    "op": "and",
    "content": [
      {
        "op": "in",
        "content": {
          "field": "repertoire_id",
          "value": [
            "2603354229190496746-242ac113-0001-012",
            "2618085967015776746-242ac113-0001-012",
            "2633633748627296746-242ac113-0001-012",
            "2564613624180576746-242ac113-0001-012"
          ]
        }
      },
      {
        "op": "=",

```

(continues on next page)

(continued from previous page)

```

        "content": {
            "field": "junction_aa",
            "value": "CARDPRSYHAFDIW"
        }
    ]
},
"fields": ["repertoire_id", "sequence_id", "v_call", "productive"],
"format": "tsv"
}

```

2.2 Release Notes

2.2.1 Schema Release Notes

Version 1.5.0: August 29, 2023

Version 1.5 schema release.

General Schema Changes:

1. Fixed synchronization errors between the OpenAPI v2 and v3 versions of the AIRR Schema (airr-schema.yaml and airr-schema-openapi3.yaml).
2. Set the default value of `x-airr.miarr` attributes to `defined`.
3. Converted all `x-airr.format` attribute values to `snake_case`, which specifically impacts any instance of controlled vocabulary or physical quantity.
4. Corrected numerous instances of missing `x-airr.miarr` and `x-airr.identifier` attributes.
5. Replaced `x-airr.adc-api-optional` attribute with `x-airr.adc-query-support` in multiple fields.
6. Added “IGI” as a valid value to the locus enum fields in multiple schema.
7. Added `null` as a valid value to all nullable enum fields.
8. Removed `discriminator: AIRR` from all object definitions.

Germline and Genotype Schema:

1. Clarified the descriptions of multiple fields in the Germline and Genotype schema.
2. Modified `x-airr: nullable` and `x-airr: identifier` values on multiple fields in the Germline and Genotype schema.
3. Removed the `alignment` field and added the `unaligned_sequence`, `aligned_sequences`, and `alignment_labels` fields to the `SequenceDelineationV` object.
4. Converted the enum values in the `inference_type` field of `AlleleDescription` to `snake_case`.
5. Added the `allele_similarity_cluster_designation` and `allele_similarity_cluster_member_id` fields to `AlleleDescription`.
6. Moved the nested objects `DocumentedAllele`, `UndocumentedAllele`, and `DeletedGenes` out of `Genotype` and defined them as top-level objects references by the `documented_alleles`, `undocumented_alleles`, and `deleted_genes` fields, respectively.

7. Moved the nested object `MHCAllele` out of `MHCGenotype` and defined it as a top-level object referenced by the `mhc_alleles` field.

Single-cell Schema:

1. Added the `property_type` field to the `CellExpression` object.
2. Moved the nested `ReceptorReactivity` object out of `Receptor` and defined it as a top-level object referenced by the `reactivity_measurements` field.

Subject Schema:

1. Removed the nested references to `GenotypeSet` and `MHCGenotypeSet` in the `genotype` field and modified the definition to point to a top-level `SubjectGenotype` object defining these references.

DataProcessing Schema:

1. Clarified the description of `quality_thresholds` to indicate that quality filtering is not mandatory.

Version 1.4.1: August 27, 2022

Version 1.4 schema release.

New General Purpose Schema:

1. Introduced the experimental `DataFile` object, which defines a JSON file holding Repertoire metadata, data processing analysis objects, or any object in the AIRR Data Model.
2. Introduced the experimental `RepertoireGroup` Schema for describing collections of repertoires to be analyzed together.
3. Introduced the experimental `InfoObject` Schema, which provides information about data and ADC API responses.
4. Introduced the experimental `TimePoint` Schema for defining the time point at which an observation or other action was performed.

New Germline and Genotype Schema:

The following experimental schema were introduced to support storage of VDJ germline reference sequences, VDJ genotypes, and MHC genotypes:

1. **GermlineSet**: Defines a collection of `AlleleDescriptions` from the same strain or species.
2. **AlleleDescription**: Details of a putative or confirmed Ig receptor gene/allele inferred from one or more observations.
3. **RearrangedSequence**: Details of a directly observed rearranged sequence or an inference from rearranged sequences contributing support for a gene or allele.
4. **UnrearrangedSequence**: Details of an unrearranged sequence contributing support for a gene or allele.
5. **SequenceDelineationV**: Delineation of a V-gene in a particular system.
6. **GenotypeSet**: Defines a collection a VDJ genotypes for a given subject.
7. **Genotype**: Enumerates the alleles and gene deletions inferred in a single subject for a single locus.
8. **MHCGenotypeSet**: Defines a collection of MHC genotypes for a given subject.
9. **MHCGenotype**: Details the genotype of major histocompatibility complex (MHC) class I, class II and non-classical loci.
10. **Acknowledgement**: Defines contributors to the germline or genotype description.

New Single-cell Schema:

The following experimental schema were introduced to improve support for single-cell data and extend the Cell schema.

1. **CellExpression**: Defines a container to store single-cell expression level measurements.
2. **Receptor**: Describes a complete receptor protein sequence and its reactivity.

Rearrangement Schema:

1. Added the optional fields `v_frameshift`, `j_frameshift`, `d_frame` and `d2_frame` defining annotations related to alignment reading frames.
2. Added the optional field `umi_count` to represent the count of distinct UMIs for a sequence.
3. Modified the definition of `duplicate_count` to remove ambiguity with the new `umi_count` field in a single-cell context. There is now a distinction between duplicate observed sequences (`duplicate_count`) and UMIs (`umi_count`).
4. The optional `quality` and `quality_alignment` alignment fields were added to store Phred quality scores for base calls in the `sequence` and `sequence_alignment` fields, respectively.
5. The following optional fields were added to denote constant region (`c_call`) alignment positions: `c_sequence_start`, `c_sequence_end`, `c_germline_start`, `c_germline_end`, `c_alignment_start`, `c_alignment_end`.

Study Schema:

1. Added the optional fields `study_contact` to store contact information for the primary study contact.
2. Modified the enumerated values supported by `keywords_study` to the following set: `contains_ig`, `contains_tr`, `contains_paired_chain`, `contains_schema_rearrangement`, `contains_schema_clone`, `contains_schema_cell`, `contains_schema_receptor`
3. Added the optional fields `adc_publish_date` and `adc_update_data` that timestamp AIRR Data Commons initial publication and last update, respectively.

Subject Schema:

1. Added the optional `genotype` field linking to the new `GenotypeSet` and `MHCGenotypeSet` objects.

Sample Schema:

1. Added the required field `collection_time_point_relative_unit` defining the units for the sample collection timestamp.
2. Modified the type of the field `collection_time_point_relative` from a string to a number defined in combination with the new unit ontology field `collection_time_point_relative_unit`.

NucleicAcidProcessing Schema:

1. Added the required field `template_amount_unit` defining the units for the input template quantification.
2. Modified the type of the `template_amount` field from a string to a number defined in the combination with the new unit ontology field `template_amount_unit`.

Clone Schema:

1. Added the optional `clone_count` field to specify absolute count of clonal members.
2. Added the optional `umi_count` field to specify the total UMI count of all clonal members.

Cell Schema:

1. Removed the field `expression_tabular` whose functionality has been replaced by the new `CellExpression` schema.

Version 1.3.1: October 13, 2020

Version 1.3 documentation patch release.

Alignment Schema:

1. Added the deprecation tags for `rearrangement_id`, which were accidentally left out of the v1.3.0 release.

Version 1.3.0: May 28, 2020

Version 1.3 schema release.

New Schema:

1. Introduced the Repertoire Schema for describing study meta data.
2. Introduced the PCRTarget Schema for describing primer target locations.
3. Introduced the SampleProcessing Schema for describing experimental processing steps for a sample.
4. Replaced the SoftwareProcessing schema with the DataProcessing schema.
5. Introduced experimental schema for clonal clusters, lineage trees, tree nodes, and cells as `Clone`, `Tree`, `Node`, and `Cell` objects, respectively.

General Updates:

1. Added multiple additional attributes to a large number of schema properties as AIRR extension attributes in the `x-airr` field. The new `Attributes` object contains definitions for these `x-airr` field attributes.
2. Added the top level `required` property to all relevant schema objects.
3. Added the `title` attribute containing the short, descriptive name to all relevant schema object fields.
4. Added an `example` attribute containing an example data value to multiple schema object fields.

AIRR Data Commons API:

1. Added OpenAPI V2 specification (`specs/adc-api.yaml`) for AIRR Data Commons API major version 1.

Ontology Support:

1. Added `Ontology` and `CURIEResolution` objects to support ontologies.
2. Added vocabularies/ontologies as JSON string for: Cell subset, Target substrate, Library generation method, Complete sequences, Physical linkage of different loci.

Rearrangement Schema:

1. Added the `complete_vdj` field to annotate whether a V(D)J alignment was full length.
2. Added the `junction_length_aa` field defining the length of the junction amino acid sequence.
3. Added the `repertoire_id`, `sample_processing_id`, and `data_processing_id` fields to serve as linkers to the appropriate metadata objects.
4. Added a controlled vocabulary to the `locus` field: IGH, IGI, IGK, IGL, TRA, TRB, TRD, TRG.
5. Deprecated the `rearrangement_set_id` and `germline_database` fields.
6. Deprecated `rearrangement_id` field and made the `sequence_id` field be the primary unique identifier for a rearrangement record, both in files and data repositories.
7. Added support secondary D gene rearrangement through the additional fields: `d2_call`, `d2_score`, `d2_identity`, `d2_support`, `d2_cigar`, `np3`, `np3_aa`, `np3_length`, `n3_length`, `p5d2_length`, `p3d2_length`, `d2_sequence_start`, `d2_sequence_end`, `d2_germline_start`, `d2_germline_end`.

d2_alignment_start, d2_alignment_end, d2_sequence_alignment, d2_sequence_alignment_aa, d2_germline_alignment, d2_germline_alignment_aa.

8. Updated field definitions with more concise V(D)J call descriptions.

Alignment Schema:

1. Deprecated the rearrangement_set_id and germline_database fields.
2. Added the data_processing_id field.

Study Schema:

1. Added the study_type field containing an ontology defined term for the study design.

Subject Schema:

1. Deprecated the organism field in favor of the new species field.
2. Deprecated the age field.
3. Introduced age ranges: age_min, age_max, and age_unit.

Diagnosis Schema:

1. Changed the type of the disease_diagnosis field from string to Ontology.

Sample Schema:

1. Changed the type of the tissue field from string to Ontology.

CellProcessing Schema:

1. Changed the type of the cell_subset field from string to Ontology.
2. Introduced the cell_species field which denotes the species from which the analyzed cells originate.

NucleicAcidProcessing Schema:

1. Defined the template_class field as type string.
2. Added a controlled vocabulary the library_generation_method field.
3. Changed the controlled vocabulary terms of complete_sequences. Replacing complete & untemplated with complete+untemplated and adding mixed.
4. Added the pcr_target field referencing the new PCRTarget schema object.

SequencingRun Schema:

1. Added the sequencing_run_id field which serves as the object identifier field.
2. Added the sequencing_files field which links to the RawSequenceData schema objects defining the raw read data.

RawSequenceData Schema:

1. Added the file_type field defining the sequence file type. This field is a controlled vocabulary restricted to: fasta, fastq.
2. Added the paired_read_length field defining mate-pair read lengths.
3. Defined the read_direction and paired_read_direction fields as type string.

DataProcessing Schema:

1. Replaces the SoftwareProcessing object.
2. Added data_processing_id, primary_annotation, data_processing_files, germline_database and analysis_provenance_id fields.

Version 1.2.1: Oct 5, 2018

Minor patch release.

1. Schema gene vs segment terminology corrections
2. Added Info object
3. Updated cell_subset URL in AIRR schema

Version 1.2.0: Aug 18, 2018

Peer reviewed released of the Rearrangement schema.

1. Definition change for the coordinate fields of the Rearrangement and Alignment schema. Coordinates are now defined as 1-based closed intervals, instead of 0-based half-open intervals (as previously defined in v1.1 of the schema).
2. Removed foreign study_id fields
3. Introduced keywords_study field

Version 1.1.0: May 3, 2018

Initial public released of the Rearrangement and Alignment schemas.

1. Added required and nullable constraints to AIRR schema.
2. Schema definitions for MiAIRR attributes and ontology.
3. Introduction of an x-airr object indicating if field is required by MiAIRR.
4. Rename rearrangement_set_id to data_processing_id.
5. Rename study_description to study_type.
6. Added physical_quantity format.
7. Raw sequencing files into separate schema object.
8. Rename Attributes object.
9. Added primary_annotation and repertoire_id.
10. Added diagnosis to repertoire object.
11. Added ontology for organism.
12. Added more detailed specification of sequencing_run, repertoire and rearrangement.
13. Added repertoire schema.
14. Rename definitions.yaml to airr-schema.yaml.
15. Removed c_call, c_score and c_cigar from required as this is not typical reference aligner output.
16. Renamed vdj_score, vdj_identity, vdj_evalue, and vdj_cigar to score, identity, evalue, and cigar.
17. Added missing c_identity and c_evalue fields to Rearrangement spec.
18. Swapped order of N and S operators in CIGAR string.
19. Some description clean up for consistency in Rearrangement spec.

20. Remove repeated objects in `definitions.yaml`.
21. Added `Alignment` object to `definitions.yaml`.
22. Updated MiARR format consistency check TSV with junction change.
23. Changed definition from functional to productive.

Version 1.0.1: Jan 9, 2018

MiAIRR v1 official release and initial draft of Rearrangement and Alignment schemas.

2.2.2 Python Library Release Notes

Version 1.5.0: August 29, 2023

1. Updated schema set and examples to v1.5.
2. Officially dropped support for Python 2.
3. Added check for valid enum values to schema validation routines.
4. Set enum values to first defined value during template generation routines.
5. Removed mock dependency installation in ReadTheDocs environments from setup.
6. Improved package import time.

Version 1.4.1: August 27, 2022

General:

1. Updated pandas requirement to 0.24.0 or higher.
2. Added support for missing integer values (NaN) in `load_rearrangement` by casting to the pandas `Int64` data type.
3. Added gzip support to `read_rearrangement`.
4. Significant internal refactoring to improve schema generalizability, harmonize behavior between the python and R libraries, and prepare for AIRR Standards v2.0.
5. Fixed a bug in the `validate` subcommand of `airr-tools` causing validation errors to only be reporting for the first invalid file when multiple files were specified on the command line.

Data Model and Schema:

1. Added support for arrays of objects in a single JSON or YAML file.
2. Added support for the AIRR Data File and associated schema (`DataFile`, `Info`). The Data File data format holds AIRR object of multiple types and is backwards compatible with Repertoire metadata.
3. Added support for the new germline and genotyping schema (`GermlineSet`, `GenotypeSet`) and associated schema.
4. Renamed `schema.CachedSchema` to `schema.AIRRSchema`.
5. Removed `specs/blank.airr.yaml`.

Deprecations:

1. Deprecated `load_repertoire`. Use `read_airr` instead.

2. Deprecated `write_repertoire`. Use `write_airr` instead.
3. Deprecated `validate_repertoire`. Use `validate_airr` instead.
4. Deprecated `repertoire_template`. Use `schema.RepertoireSchema.template` instead.
5. Deprecated the commandline tool `airr-tools validate repertoire`. Use `airr-tools validate airr` instead.

Version 1.3.1: October 13, 2020

1. Refactored `merge_rearrangement` to allow for larger number of files.
2. Improved error handling in format validation operations.

Version 1.3.0: May 30, 2020

1. Updated schema set to v1.3.
2. Added `load_repertoire`, `write_repertoire`, and `validate_repertoire` to `airr.interface` to read, write and validate Repertoire metadata, respectively.
3. Added `repertoire_template` to `airr.interface` which will return a complete repertoire object where all fields have null values.
4. Added `validate_object` to `airr.schema` that will validate a single repertoire object against the schema.
5. Extended the `airr-tools` commandline program to validate both rearrangement and repertoire files.

Version 1.2.1: October 5, 2018

1. Fixed a bug in the python reference library causing start coordinate values to be empty in some cases when writing data.

Version 1.2.0: August 17, 2018

1. Updated schema set to v1.2.
2. Several improvements to the `validate_rearrangement` function.
3. Changed behavior of all `airr.interface` functions to accept a file path (string) to a single Rearrangement TSV, instead of requiring a file handle as input.
4. Added `base` argument to `RearrangementReader` and `RearrangementWriter` to support optional conversion of 1-based closed intervals in the TSV to python-style 0-based half-open intervals. Defaults to conversion.
5. Added the custom exception `ValidationError` for handling validation checks.
6. Added the `validate` argument to `RearrangementReader` which will raise a `ValidationError` exception when reading files with missing required fields or invalid values for known field types.
7. Added `validate` argument to all type conversion methods in `Schema`, which will now raise a `ValidationError` exception for value that cannot be converted when set to `True`. When set `False` (default), the previous behavior of assigning `None` as the converted value is retained.
8. Added `validate_header` and `validate_row` methods to `Schema` and removed `validations` methods from `RearrangementReader`.
9. Removed automatic closure of file handle upon reaching the iterator end in `RearrangementReader`.

Version 1.1.0: May 1, 2018

Initial release.

2.2.3 Release Notes

Version 1.5.0: August 29, 2023

- Updated schema set and examples to v1.5.

Version 1.4.1: August 27, 2022

Significant internal refactoring to improve schema generalizability, harmonize behavior between the python and R libraries, and prepare for AIRR Standards v2.0.

Rearrangement:

- Added the `aux_types` argument to `read_tabular`, `read_rearrangement`, and `read_alignment` to allow explicit declaration of the type for fields that are not defined in the schema.
- Renamed `read_airr`, `write_airr`, and `validate_airr` to `read_tabular`, `validate_tabular`, and `validate_tabular`, respectively.

Data Model and Schema:

- Defined new `read_airr`, `write_airr`, and `validate_airr` functions that support AIRR Data Model files that store arrays of objects in JSON or YAML.
- Added support for the AIRR Model Data File and associated schema (DataFile, Info). The Data File data format holds AIRR object of multiple types and is backwards compatible with Repertoire metadata.
- Added support for the new germline and genotyping schema (GermlineSet, GenotypeSet) and associated schema.

Version 1.3.0: May 26, 2020

- Updated schema set to v1.3.
- Added `info` slot to Schema object containing general schema information.

Version 1.2.0: August 17, 2018

- Updated schema set to v1.2.
- Changed defaults to `base="1"` for read and write functions.
- Updated example TSV file with coordinate changes, addition of `germline_alignment` data and simplification of `sequence_id` values.

Version 1.1.0: May 1, 2018

Initial release.

2.3 AIRR Standards

Information about all of the AIRR Community standards.

2.3.1 Introduction to MiAIRR

Summary

One of the primary initiatives of the Adaptive Immune Receptor Repertoire (AIRR) Community has been to develop a set of metadata standards for the submission of AIRR sequencing datasets. This work has been carried out by the [AIRR Community Minimal Standards Working Group](#). In order to support reproducibility, standard quality control, and data deposition in a common repository, the AIRR Community has agreed to six high-level data sets that will guide the publication, curation and sharing of AIRR-Seq data and metadata: Study and subject, sample collection, sample processing and sequencing, raw sequences, processing of sequence data, and processed AIRR sequences. The detailed data elements within these sets are defined [here](#) (Download as TSV).

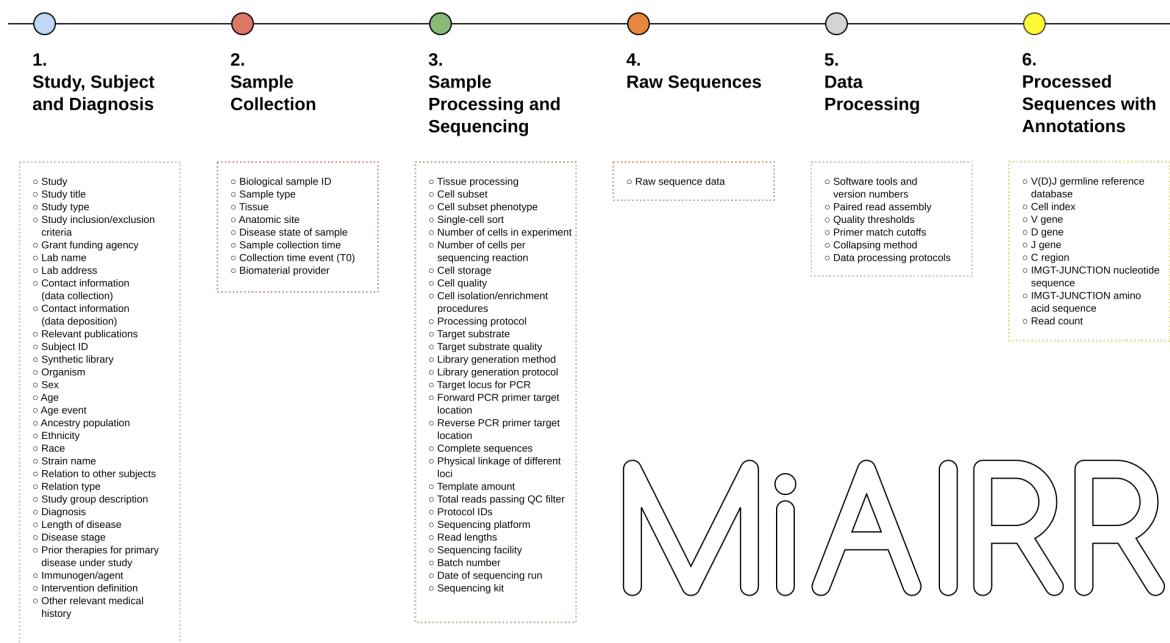


Fig. 4: Schema of MiAIRR data sets and the individual data elements of each set.

Topics

MiAIRR Data Elements

The AIRR Community has agreed to six high-level data sets that will guide the publication, curation and sharing of AIRR-Seq data and metadata: Study and subject, sample collection, sample processing and sequencing, raw sequences, processing of sequence data, and processed AIRR sequences.

Download as TSV.

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
1 / stud	Study ID study_id	string free text	important	Unique ID assigned by study registry such as one of the International Nucleotide Sequence Database Collaboration (INSDC) repositories.	PRJNA001
1 / stud	Study title study_title	string free text	important	Descriptive study title	Effects of sun light exposure of the Treg repertoire
1 / stud	Study type study_type	<i>Ontology</i> <i>Ontology: { top_node: { id: NCIT:C63536, label: Study}}</i>	important	Type of study design	id: NCIT:C15197, label: Case-Control Study
1 / stud	Study inclusion/exclusion criteria inclusion_ex	string free text	important	List of criteria for inclusion/exclusion for the study	Include: Clinical P. falciparum infection; Exclude: Seropositive for HIV
1 / stud	Grant funding agency grants	string free text	important	Funding agencies and grant numbers	NIH, award number R01GM987654
1 / stud	Contact information (data collection) collected_by	string free text	important	Full contact information of the data collector, i.e. the person who is legally responsible for data collection and release. This should include an e-mail address and a persistent identifier such as an ORCID ID.	Dr. P. Stibbons, p.stibbons@unseenu.edu , https://orcid.org/0000-0002-1825-0097
1 / stud	Lab name lab_name	string free text	important	Department of data collector	Department for Planar Immunology
1 / stud	Lab address lab_address	string free text	important	Institution and institutional address of data collector	School of Medicine, Unseen University, Ankh-Morpork, Disk World

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
1 / stud	Contact information (data deposition) submitted_by	string <i>free text</i>	important	Full contact information of the data depositor, i.e., the person submitting the data to a repository. This should include an e-mail address and a persistent identifier such as an ORCID ID. This is supposed to be a short-lived and technical role until the submission is released.	Adrian Turnipseed, a.turnipseed@unseenu.edu , https://orcid.org/0000-0002-1825-0097
1 / stud	Relevant publications pub_ids	string <i>free text</i>	important	Publications describing the rationale and/or outcome of the study. Where ever possible, a persistent identifier should be used such as a DOI or a Pubmed ID	PMID:85642
1 / stud	Keywords for study keywords_stu	array of string **	important	Keywords describing properties of one or more data sets in a study. “contains_schema” keywords indicate that the study contains data objects from the AIRR Schema of that type (Rearrangement, Clone, Cell, Receptor) while the other keywords indicate that the study design considers the type of data indicated (e.g. it is possible to have a study that “contains_paired_chain” but does not “contains_schema_cell”).	['contains_ig', 'contains_schema_rearrangement', 'contains_schema_clone', 'contains_schema_cell']
1 / sub-subject	Subject ID subject_id	string <i>free text</i>	important	Subject ID assigned by submitter, unique within study. If possible, a persistent subject ID linked to an INSDC or similar repository study should be used.	SUB856413
1 / sub-subject	Synthetic library synthetic	boolean <i>true</i> <i>false</i>	essential	TRUE for libraries in which the diversity has been synthetically generated (e.g. phage display)	
1 / sub-subject	Organism species	<i>Ontology</i> <i>Ontology:</i> { <i>top_node:</i> { <i>id:</i> NCBITAXON:7; <i>label:</i> Gnathos-tomata}}	essential	Binomial designation of subject's species	id: NCBITAXON:9606, label: Homo sapiens

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
1 / sub-subject	Sex sex	string <i>free text</i>	important	Biological sex of subject	female
1 / sub-subject	Age minimum age_min	number <i>positive number</i>	important	Specific age or lower boundary of age range.	60
1 / sub-subject	Age maximum age_max	number <i>positive number</i>	important	Upper boundary of age range or equal to age_min for specific age. This field should only be null if age_min is null.	80
1 / sub-subject	Age unit age_unit	<i>Ontology</i> <i>Ontology:</i> { <i>top_node:</i> { <i>id:</i> UO:0000003, <i>label:</i> time unit}}	important	Unit of age range	id: UO:0000036, label: year
1 / sub-subject	Age event age_event	string <i>free text</i>	important	Event in the study schedule to which <i>Age</i> refers. For NCBI BioSample this MUST be <i>sampling</i> . For other implementations submitters need to be aware that there is currently no mechanism to encode to potential delta between <i>Age event</i> and <i>Sample collection time</i> , hence the chosen events should be in temporal proximity.	enrollment
1 / sub-subject	Ancestry population ancestry_population	<i>Ontology</i> <i>Ontology:</i> { <i>top_node:</i> { <i>id:</i> GAZ:00000448, <i>label:</i> geographic location}}	important	Broad geographic origin of ancestry (continent)	id: GAZ:00000459, label: South America
1 / sub-subject	location_birth location_birth	<i>Ontology</i> <i>Ontology:</i> { <i>top_node:</i> { <i>id:</i> GAZ:00000448, <i>label:</i> geographic location}}	important	Self-reported location of birth of the subject, preferred granularity is country-level	id: GAZ:00002939, label: Poland

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
1 / sub- subject	Ethnicity ethnicity	string free text	important	Ethnic group of subject (defined as cultural/language-based membership)	English, Kurds, Manchu, Yakuts (and other fields from Wikipedia)
1 / sub- subject	Race race	string free text	important	Racial group of subject (as defined by NIH)	White, American Indian or Alaska Native, Black, Asian, Native Hawaiian or Other Pacific Islander, Other
1 / sub- subject	Strain name strain_name	string free text	important	Non-human designation of the strain or breed of animal used	C57BL/6J
1 / sub- subject	Relation to other subjects linked_subject	string free text	important	Subject ID to which <i>Relation type</i> refers	SUB1355648
1 / sub- subject	Relation type link_type	string free text	important	Relation between subject and <i>linked_subjects</i> , can be genetic or environmental (e.g.exposure)	father, daughter, household
1 / diagnosis and intervention	Study group description study_group_id	string free text	important	Designation of study arm to which the subject is assigned to	control
1 / diagnosis and intervention	Diagnosis disease_diagnosis	<i>Ontology</i> <i>Ontology: {</i> <i>top_node: {</i> <i>id: DOID:4,</i> <i>label: disease}}</i>	important	Diagnosis of subject	id: DOID:9538, label: multiple myeloma

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
1 / diagnosis and intervention	Length of disease disease_length	string <i>free text</i>	important	Time duration between initial diagnosis and current intervention	23 months
1 / diagnosis and intervention	Disease stage disease_stage	string <i>free text</i>	important	Stage of disease at current intervention	Stage II
1 / diagnosis and intervention	Prior therapies for primary disease under study prior_therapies	string <i>free text</i>	important	List of all relevant previous therapies applied to subject for treatment of <i>Diagnosis</i>	melphalan/prednisone
1 / diagnosis and intervention	Immunogen/agent immunogen	string <i>free text</i>	important	Antigen, vaccine or drug applied to subject at this intervention	bortezomib

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
1 / diagnosis and intervention	Intervention definition intervention	string <i>free text</i>	important	Description of intervention	systemic chemotherapy, 6 cycles, 1.25 mg/m2
1 / diagnosis and intervention	Other relevant medical history medical_history	string <i>free text</i>	important	Medical history of subject that is relevant to assess the course of disease and/or treatment	MGUS, first diagnosed 5 years prior
2 / sample	Biological sample ID sample_id	string <i>free text</i>	important	Sample ID assigned by submitter, unique within study. If possible, a persistent sample ID linked to INSDC or similar repository study should be used.	SUP52415
2 / sample	Sample type sample_type	string <i>free text</i>	important	The way the sample was obtained, e.g. fine-needle aspirate, organ harvest, peripheral venous puncture	Biopsy
2 / sample	Tissue tissue	<i>Ontology</i> <i>Ontology: {</i> <i>top_node: {</i> <i>id: UBERON:00100</i> <i>label: multicellular anatomical structure}}</i>	important	The actual tissue sampled, e.g. lymph node, liver, peripheral blood	id: UBERON:0002371, label: bone marrow
2 / sample	Anatomic site anatomic_site	string <i>free text</i>	important	The anatomic location of the tissue, e.g. Inguinal, femur	Iliac crest
2 / sample	Disease state of sample disease_state	string <i>free text</i>	important	Histopathologic evaluation of the sample	Tumor infiltration
2 / sample	Sample collection time collection_time	number <i>positive number</i>	important	Time point at which sample was taken, relative to <i>Collection time event</i>	14

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
2 / sample collection time unit sample collection_time	Sample collection time unit collection_time	<i>Ontology</i> <i>Ontology:</i> { <i>top_node:</i> { <i>id:</i> UO:0000003, <i>label:</i> time unit}}	important	Unit of Sample collection time	id: UO:0000033, label: day
2 / sample time event sample collection_time	Collection time event collection_time	string <i>free text</i>	important	Event in the study schedule to which <i>Sample collection time</i> relates to	Primary vaccination
2 / sample collection location sample collection_location	Sample collection location collection_location	<i>Ontology</i> <i>Ontology:</i> { <i>top_node:</i> { <i>id:</i> GAZ:00000448, <i>label:</i> geographic location}}	important	Location where the sample was taken, preferred granularity is country-level	id: GAZ:00002939, label: Poland
2 / sample material provider sample biomaterial_provider	Biomaterial provider biomaterial_provider	string <i>free text</i>	important	Name and address of the entity providing the sample	Tissues-R-Us, Tampa, FL, USA
3 / Tissue processing (cell) tissue_processing (cell)	Tissue processing tissue_processing	string <i>free text</i>	important	Enzymatic digestion and/or physical methods used to isolate cells from sample	Collagenase A/Dnase I digested, followed by Percoll gradient
3 / Cell subset process (cell) cell_subset (cell)	Cell subset cell_subset	<i>Ontology</i> <i>Ontology:</i> { <i>top_node:</i> { <i>id:</i> CL:0000542, <i>label:</i> lymphocyte}}	important	Commonly-used designation of isolated cell population	id: CL:0000972, label: class switched memory B cell
3 / Cell subset phenotype process (cell) cell_subset_phenotype (cell)	Cell subset phenotype cell_subset_phenotype	string <i>free text</i>	important	List of cellular markers and their expression levels used to isolate the cell population	CD19+ CD38+ CD27+ IgM- IgD-

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
3 / process (cell)	Cell species cell_species	<i>Ontology</i> <i>Ontology:</i> { <i>top_node:</i> { <i>id:</i> NCBITAXON:7; <i>label:</i> Gnathos- tomata}}	defined	Binomial designation of the species from which the analyzed cells originate. Typically, this value should be identical to <i>species</i> , in which case it SHOULD NOT be set explicitly. However, there are valid experimental setups in which the two might differ, e.g., chimeric animal models. If set, this key will overwrite the <i>species</i> information for all lower layers of the schema.	id: NCBITAXON:9606, label: Homo sapiens
3 / process (cell)	Single-cell sort single_cell	boolean <i>true</i> <i>false</i>	important	TRUE if single cells were isolated into separate compartments	
3 / process (cell)	Number of cells in experiment cell_number	integer <i>positive integer</i>	important	Total number of cells that went into the experiment	1000000
3 / process (cell)	Number of cells per sequencing reaction cells_per_reaction	integer <i>positive integer</i>	important	Number of cells for each biological replicate	50000
3 / process (cell)	Cell storage cell_storage	boolean <i>true</i> <i>false</i>	important	TRUE if cells were cryo-preserved between isolation and further processing	True
3 / process (cell)	Cell quality cell_quality	string <i>free text</i>	important	Relative amount of viable cells after preparation and (if applicable) thawing	90% viability as determined by 7-AAD
3 / process (cell)	Cell isolation / enrichment procedure cell_isolation	string <i>free text</i>	important	Description of the procedure used for marker-based isolation or enrich cells	Cells were stained with fluorochrome labeled antibodies and then sorted on a FlowMerlin (CE) cytometer.

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
3 / process (cell)	Process-ing protocol cell_process:	string <i>free text</i>	important	Description of the methods applied to the sample including cell preparation/ isolation/enrichment and nucleic acid extraction. This should closely mirror the Materials and methods section in the manuscript.	Stimulated with anti-CD3/anti-CD28
3 / process (nucleic acid)	Target substrate template_cla:	string <i>free text</i>	essential	The class of nucleic acid that was used as primary starting material for the following procedures	RNA
3 / process (nucleic acid)	Target substrate quality template_qua:	string <i>free text</i>	important	Description and results of the quality control performed on the template material	RIN 9.2
3 / process (nucleic acid)	Template amount template_amo:	number <i>positive number</i>	important	Amount of template that went into the process	1000
3 / process (nucleic acid)	Template amount unit template_amo:	<i>Ontology</i> <i>Ontology:</i> { <i>top_node:</i> { <i>id:</i> UO:0000002, <i>label:</i> <i>physical quantity}}</i>	important	Unit of template amount	id: UO:0000024, label: nanogram
3 / process (nucleic acid)	Library generation method library_gene:	string <i>free text</i>	essential	Generic type of library generation	RT(oligo-dT)+TS(UMI)+PCR
3 / process (nucleic acid)	Library generation protocol library_gene:	string <i>free text</i>	important	Description of processes applied to substrate to obtain a library that is ready for sequencing	cDNA was generated using

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
3 / process (nucleic acid)	Protocol IDs library_gene:	string <i>free text</i>	important	When using a library generation protocol from a commercial provider, provide the protocol version number	v2.1 (2016-09-15)
3 / process (nucleic acid)	Complete sequences complete_seq	string <i>free text</i>	essential	To be considered <i>complete</i> , the procedure used for library construction MUST generate sequences that 1) include the first V gene codon that encodes the mature polypeptide chain (i.e. after the leader sequence) and 2) include the last complete codon of the J gene (i.e. 1 bp 5' of the J->C splice site) and 3) provide sequence information for all positions between 1) and 2). To be considered <i>complete & untemplated</i> , the sections of the sequences defined in points 1) to 3) of the previous sentence MUST be untemplated, i.e. MUST NOT overlap with the primers used in library preparation. <i>mixed</i> should only be used if the procedure used for library construction will likely produce multiple categories of sequences in the given experiment. It SHOULD NOT be used as a replacement of a NULL value.	partial

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
3 / process (nucleic acid [pcr]	Physical linkage of different re-arrangements <code>physical_link</code>	string <i>free text</i>	essential	In case an experimental setup is used that physically links nucleic acids derived from distinct <i>Rearrangements</i> before library preparation, this field describes the mode of that linkage. All <i>hetero_*</i> terms indicate that in case of paired-read sequencing, the two reads should be expected to map to distinct IG/TR loci. <i>*_head-head</i> refers to techniques that link the 5' ends of transcripts in a single-cell context. <i>*_tail-head</i> refers to techniques that link the 3' end of one transcript to the 5' end of another one in a single-cell context. This term does not provide any information whether a continuous reading-frame between the two is generated. <i>*_prelinked</i> refers to constructs in which the linkage was already present on the DNA level (e.g. scFv).	hetero_head-head
3 / process (nucleic acid [pcr]	Target locus for PCR <code>pcr_target_locus</code>	string <i>free text</i>	important	Designation of the target locus. Note that this field uses a controlled vocabulary that is meant to provide a generic classification of the locus, not necessarily the correct designation according to a specific nomenclature.	IGK
3 / process (nucleic acid [pcr]	Forward primer target location <code>forward_pcr_loc</code>	string <i>free text</i>	important	Position of the most distal nucleotide templated by the forward primer or primer mix	IGHV, +23
3 / process (nucleic acid [pcr]	Reverse primer target location <code>reverse_pcr_loc</code>	string <i>free text</i>	important	Position of the most proximal nucleotide templated by the reverse primer or primer mix	IGHG, +57

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
3 / process (sequencing)	Batch number sequencing_run	string <i>free text</i>	important	ID of sequencing run assigned by the sequencing facility	160101_M01234
3 / process (sequencing)	Total reads passing QC filter total_reads_]	integer <i>positive integer</i>	important	Number of usable reads for analysis	10365118
3 / process (sequencing)	Sequencing platform sequencing_p	string <i>free text</i>	important	Designation of sequencing instrument used	Alumina 1000 LoSeq
3 / process (sequencing)	Sequencing facility sequencing_f	string <i>free text</i>	important	Name and address of sequencing facility	Seqs-R-Us, Vancouver, BC, Canada
3 / process (sequencing)	Date of sequencing run sequencing_r	string <i>free text</i>	important	Date of sequencing run	2016-12-16
3 / process (sequencing)	Sequencing kit sequencing_k	string <i>free text</i>	important	Name, manufacturer, order and lot numbers of sequencing kit	FullSeq 600, Alumina, #M123456C0, 789G1HK
4 / data read (raw data persistent identifier)	Raw sequencing persistent identifier sequencing_d	string <i>free text</i>	important	Persistent identifier of raw data stored in an archive (e.g. INSDC run ID). Data archive should be identified in the CURIE prefix.	SRA:SRR11610494
4 / data read (raw data file type)	Raw sequencing file type read file_type	string <i>free text</i>	important	File format for the raw reads or sequences	

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
4 / Raw sequencing data (raw read filename)	se- quencing data file name	string <i>free text</i>	important	File name for the raw reads or sequences. The first file in paired-read sequencing.	MS10R-NMonson-C7JR9_S1_R1_001.fastq
4 / Read direction (raw read)	Read direction read_directi	string <i>free text</i>	important	Read direction for the raw reads or sequences. The first file in paired-read sequencing.	forward
4 / Forward read length (raw read)	Forward read length read_length	integer <i>posi- tive integer</i>	important	Read length in bases for the first file in paired-read sequencing	300
4 / Paired raw sequencing data (raw read paired_filen	Paired raw sequencing data file name	string <i>free text</i>	important	File name for the second file in paired-read sequencing	MS10R-NMonson-C7JR9_S1_R2_001.fastq
4 / Paired read direction (raw read paired_read_	Paired read direction paired_read_d	string <i>free text</i>	important	Read direction for the second file in paired-read sequencing	reverse
4 / Paired read length (raw read paired_read_	Paired read length paired_read_l	integer <i>posi- tive integer</i>	important	Read length in bases for the second file in paired-read sequencing	300
5 / Software process (con putation: software_ver	Software tools and version numbers	string <i>free text</i>	important	Version number and / or date, include company pipelines	IgBLAST 1.6
5 / Paired assembly process (con putation: paired_reads	Paired assembly paired_reads	string <i>free text</i>	important	How paired end reads were assembled into a single receptor sequence	PandaSeq (minimal overlap 50, threshold 0.8)
5 / Quality process (con putation: quality_thre	Quality thresholds quality_thre	string <i>free text</i>	important	How/if sequences were removed from (4) based on base quality scores	Average Phred score >=20

continues on next page

Table 1 – continued from previous page

Set / Sub set	Designation / Field	Type / Format	Level	Definition	Example
5 / process (con putation:	Primer match cutoffs primer_match	string free text	important	How primers were identified in the sequences, were they removed/masked/etc?	Hamming distance <= 2
5 / process (con putation:	Collapsing method collapsing_m	string free text	important	The method used for combining multiple sequences from (4) into a single sequence in (5)	MUSCLE 3.8.31
5 / process (con putation:	Data processing protocols data_process	string free text	important	General description of how QC is performed	Data was processed using [...]
5 / data (process: se-quer	V(D)J germline reference database germline_data	string free text	important	Source of germline V(D)J genes with version number or date accessed.	ENSEMBL, Homo sapiens build 90, 2017-10-01

MiAIRR-to-NCBI Implementation

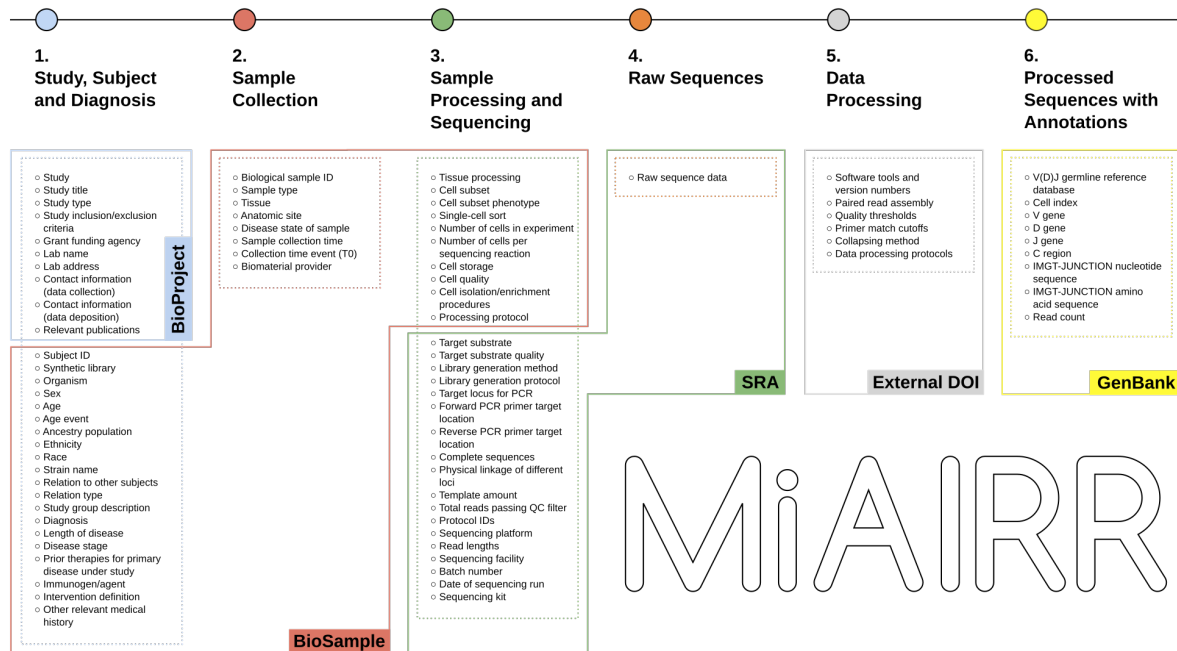
Authors

Christian E. Busse, Florian Rubelt and Syed Ahmad Chan Bukhari

Guide for submission of AIRR-seq data to NCBI

This site provides a detailed “how-to” guide for submission of AIRR-seq data to **NCBI repositories** (BioProject, BioSample, SRA and GenBank). For other implementations of the MiAIRR standard see [here](#).

One of the primary initiatives of the AIRR (Adaptive Immune Receptor Repertoire) Community has been to develop a set of metadata standards for the submission of immune receptor repertoire sequencing datasets. This work has been carried out by the AIRR Community Standards Working Group. In order to support reproducibility, standard quality control, and data deposition in a common repository, the AIRR Community has agreed to six high-level data sets that will guide the publication, curation and sharing of AIRR-Seq data and metadata: Study and subject, sample collection, sample processing and sequencing, raw sequences, processing of sequence data, and processed AIRR sequences. The detailed data elements within these sets are defined [here](#) (Download as TSV). The association between these AIRR sets, the associated data elements, and each of the NCBI repositories is shown below:



Submission of AIRR sequencing data and metadata to NCBI’s public data repositories consists of five sequential steps:

1. Submit study information to [NCBI BioProject](#) using the NCBI web interface.
2. Submit sample-level information to the [NCBI BioSample](#) repository using the [AIRR-BioSample templates](#).
3. Submit raw sequencing data to [NCBI SRA](#) using the [AIRR-SRA data templates](#).
4. Generate a DOI for the protocol describing how raw sequencing data were processed using [Zenodo](#).
5. Submit processed sequencing data with sequence-level annotations to [GenBank](#) using AIRR feature tags.

The [submission manual](#) provides step-by-step instructions on carrying out these steps for an AIRR study submission.

MiAIRR-to-NCBI Submission Manual

Scope of this document

Provide a user manual describing the submission of AIRR data using the NCBI reference implementation described in [\[Rubelt_2017\]](#). This implementation uses NCBI’s BioProject, BioSample, Sequence Read Archive (SRA) and GenBank repositories and metadata standards to report AIRR data.

Step 1. MiAIRR data submission to BioProject, BioSample and SRA

Since we propose to include a combination of raw and processed sequence data, the AIRR standard will sometimes need to be distributed and linked across multiple repositories (e.g., data in SRA linked to related data in GenBank). Besides, the data elements that comprise the standard will be mapped to ontologies in BioPortal through NIH CDE (Common Data Element) terms. These linkages will support more sophisticated validation and logical inference.

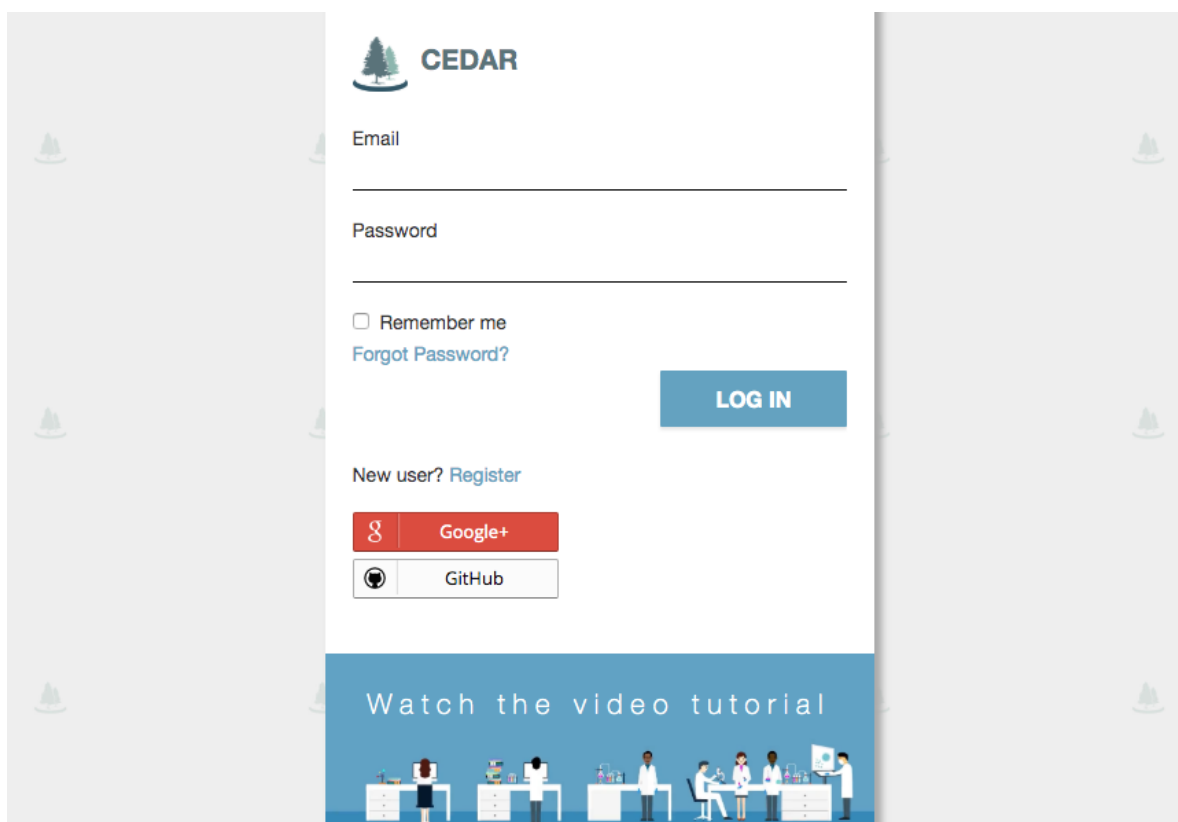
There are three main alternatives to submit raw AIRR data/metadata to NCBI repositories: (1) CEDAR's CAIRR pipeline; (2) NCBI's Web interface; and (3) NCBI's FTP server. These alternatives are described below:

Option 1. Submission via the CEDAR system (CAIRR submission pipeline)

CEDAR's CAIRR submission pipeline helps investigators and curators to edit and validate ontology-controlled meta-data. This pipeline provides a seamless interface to transmit SRA datasets to the NCBI SRA and BioSample repositories from the CEDAR Workbench. The pipeline can be directly be accessed at <http://cairr.airr-community.org>. Note that the CEDAR template and template elements used by this pipeline are publicly available in the following CEDAR folder: [All/Shared/Shared by CEDAR/MiAIRR](#).

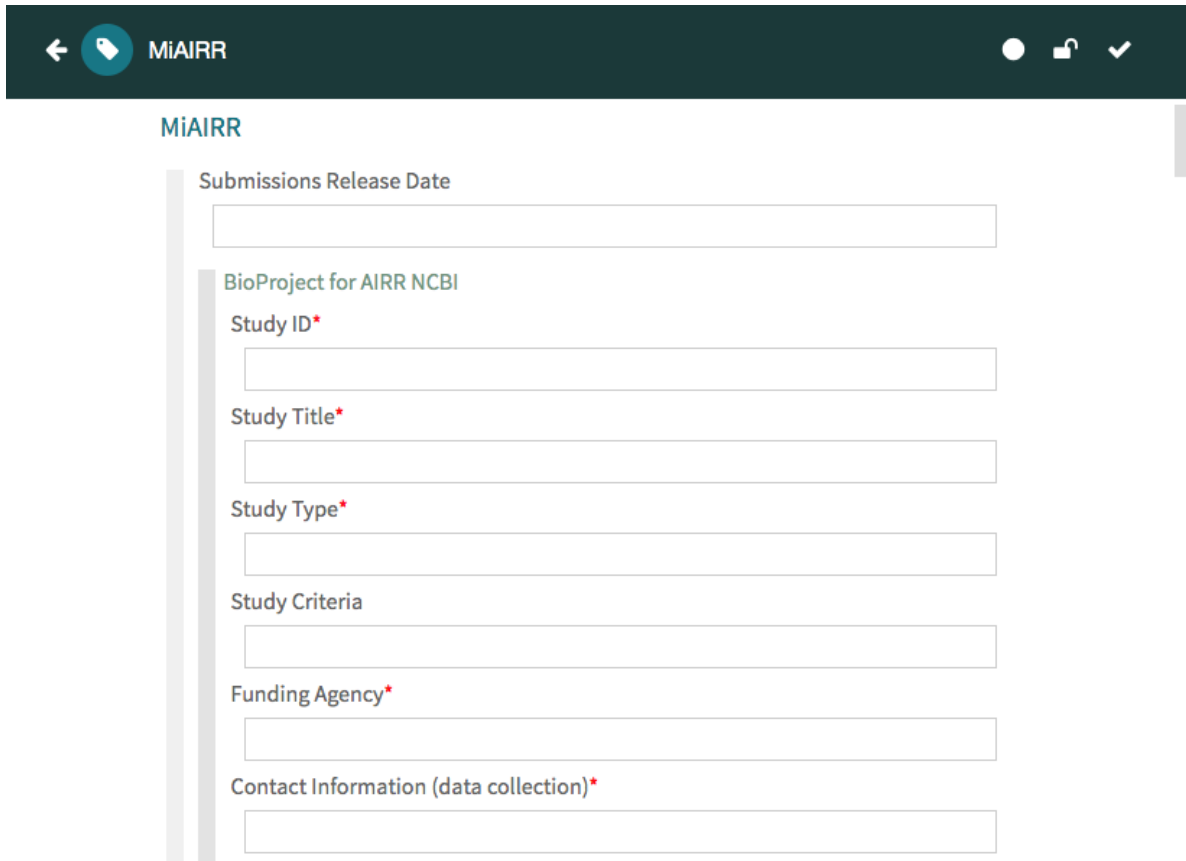
Submission steps:

1. Open CEDAR's MiAIRR template by clicking on <http://cairr.airr-community.org>. If you are not already logged in, this will take you to the CEDAR login panel. If you are a new user, you will have to create an account on the CEDAR Workbench by [clicking here](#).



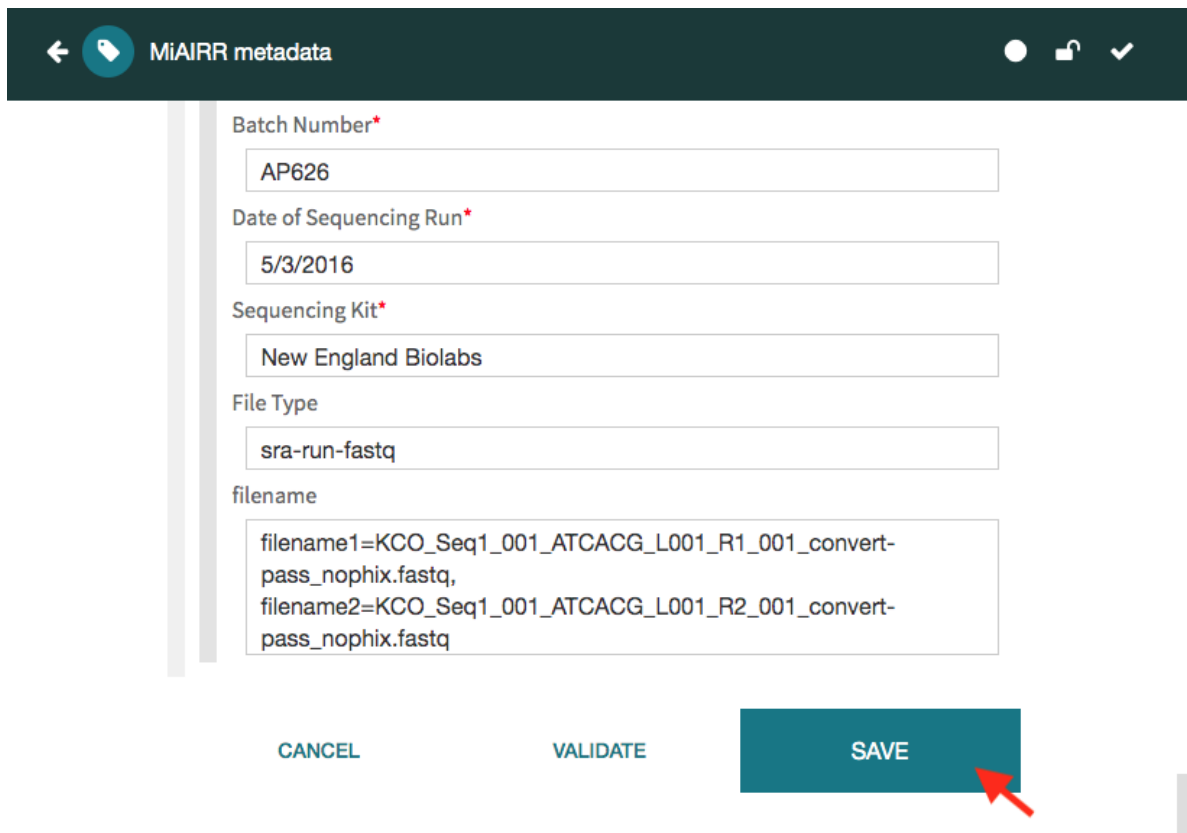
2. After logging in into the system, you will see the 'MiAIRR' template. Fill out the template fields with your metadata. Fields with an asterisk (*) are mandatory. Your submission will fail if any mandatory fields are not completed. If information is unavailable for any mandatory field, please enter 'not collected', 'not applicable' or

‘missing’ as appropriate. Note that you will need to enter a BioProject ID into the field ‘Study ID’. If you do not have a BioProject yet, you can create one at <https://submit.ncbi.nlm.nih.gov/subs/bioproject/>



The screenshot shows the MiAIRR submission form. At the top is a dark teal header bar with a back arrow, a circular icon, the text 'MiAIRR', and three status icons (a circle, a lock, and a checkmark). Below the header, the title 'MiAIRR' is displayed in blue. The form itself is a light gray vertical panel containing several input fields with labels to their left. The labels are: 'Submissions Release Date', 'BioProject for AIRR NCBI', 'Study ID*', 'Study Title*', 'Study Type*', 'Study Criteria', 'Funding Agency*', and 'Contact Information (data collection)*'. Each label is followed by a white rectangular input field with a thin gray border. The 'Study ID*' field is the first of these input fields.

3. Once your metadata is complete, click on the ‘Save’ button to save your metadata into your workspace. You will see a message in a green box confirming that your metadata have been successfully saved, as well as a message in a yellow box letting you know that your metadata have been saved to your personal workspace.



← MiAIRR metadata

Batch Number*

AP626

Date of Sequencing Run*

5/3/2016

Sequencing Kit*

New England Biolabs

File Type

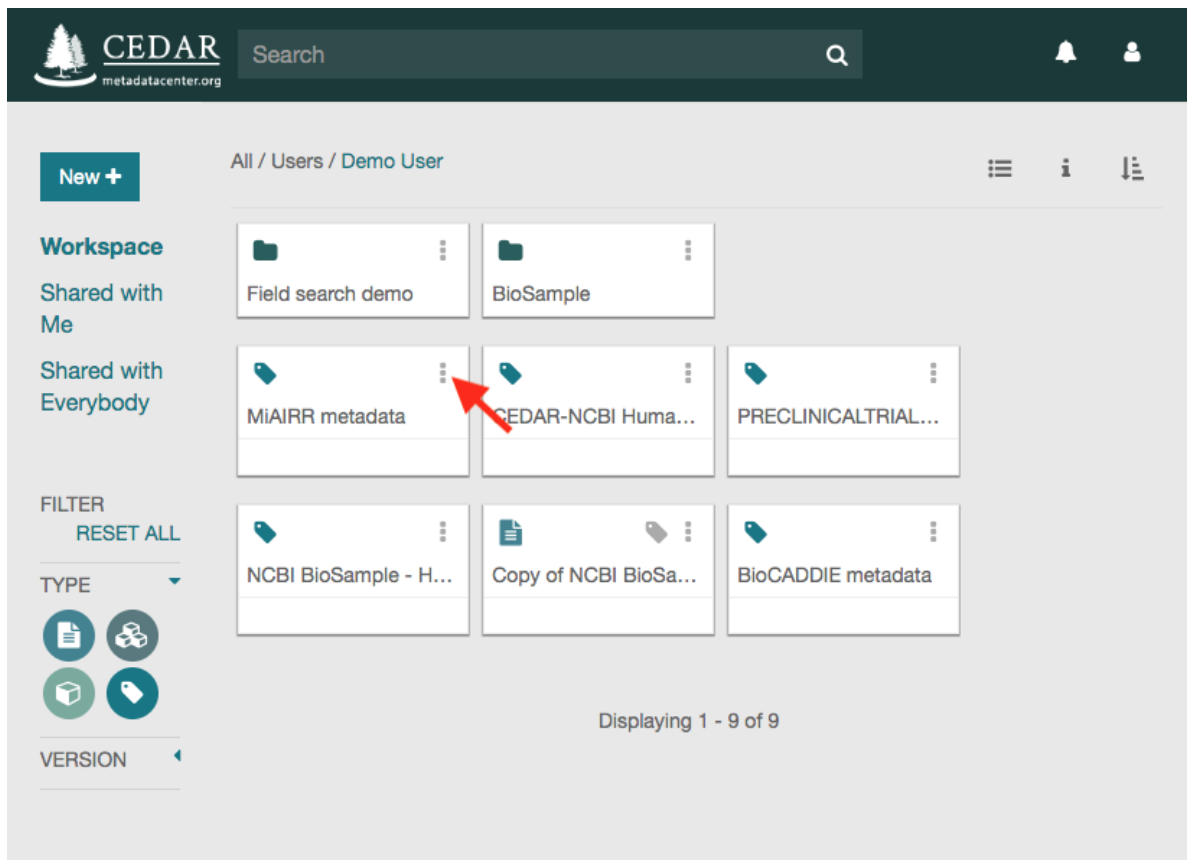
sra-run-fastq

filename

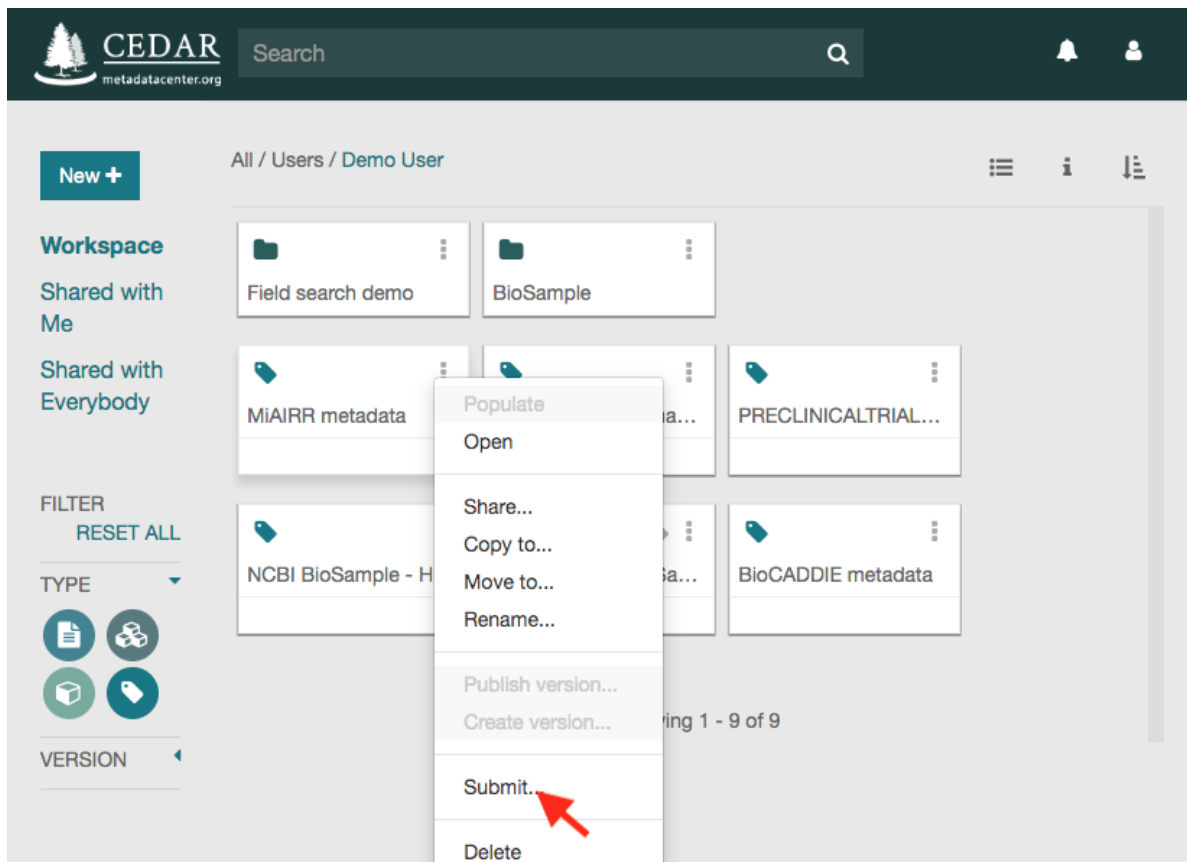
filename1=KCO_Seq1_001_ATCACG_L001_R1_001_convert-pass_nophix.fastq,
filename2=KCO_Seq1_001_ATCACG_L001_R2_001_convert-pass_nophix.fastq

CANCEL VALIDATE SAVE

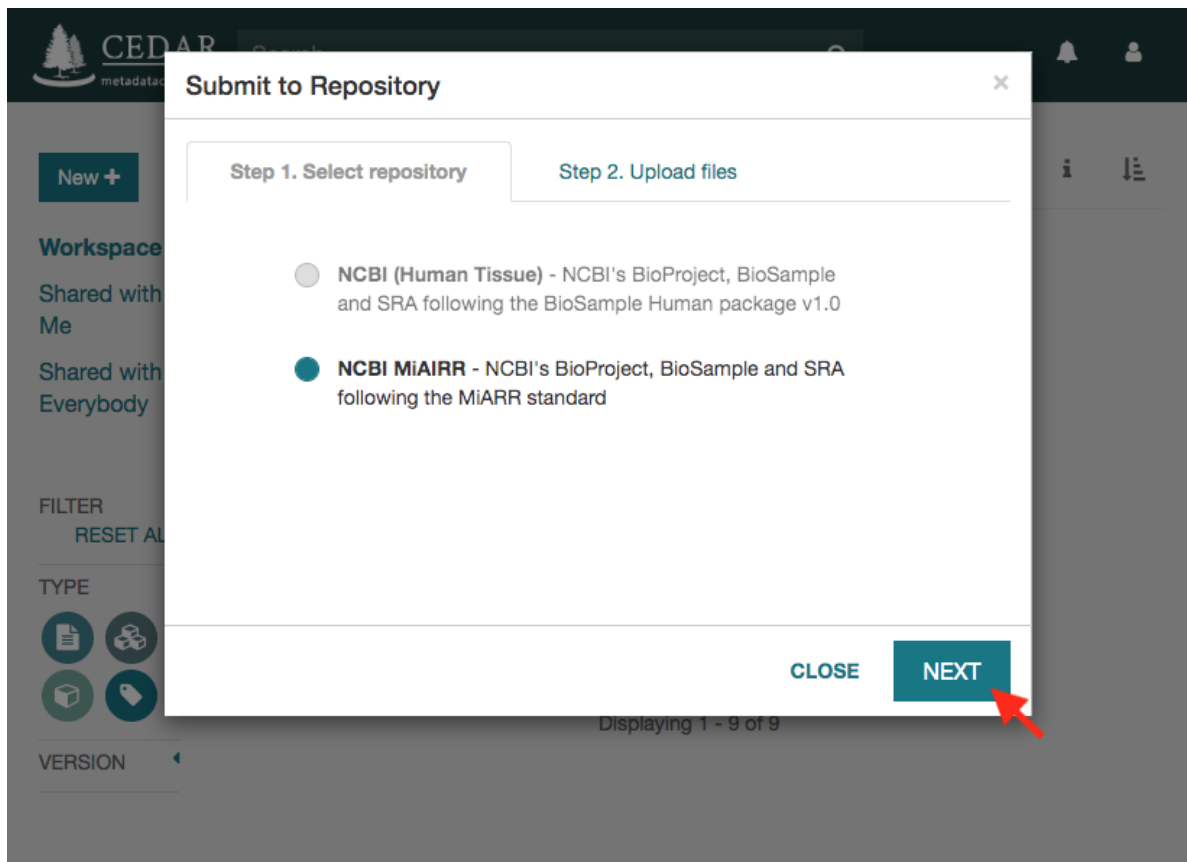
4. Go to your personal workspace by clicking on the left arrow (top left corner) and then on the 'Workspace' link, or by just clicking on: <https://cedar.metadatacenter.org>
5. Once in your workspace, you will see a metadata file called 'MiAIRR metadata'. That file contains the metadata that you have just created and that you want to submit to the NCBI. Click on the three vertical dots on the top-right corner of the file icon to see the available file options.



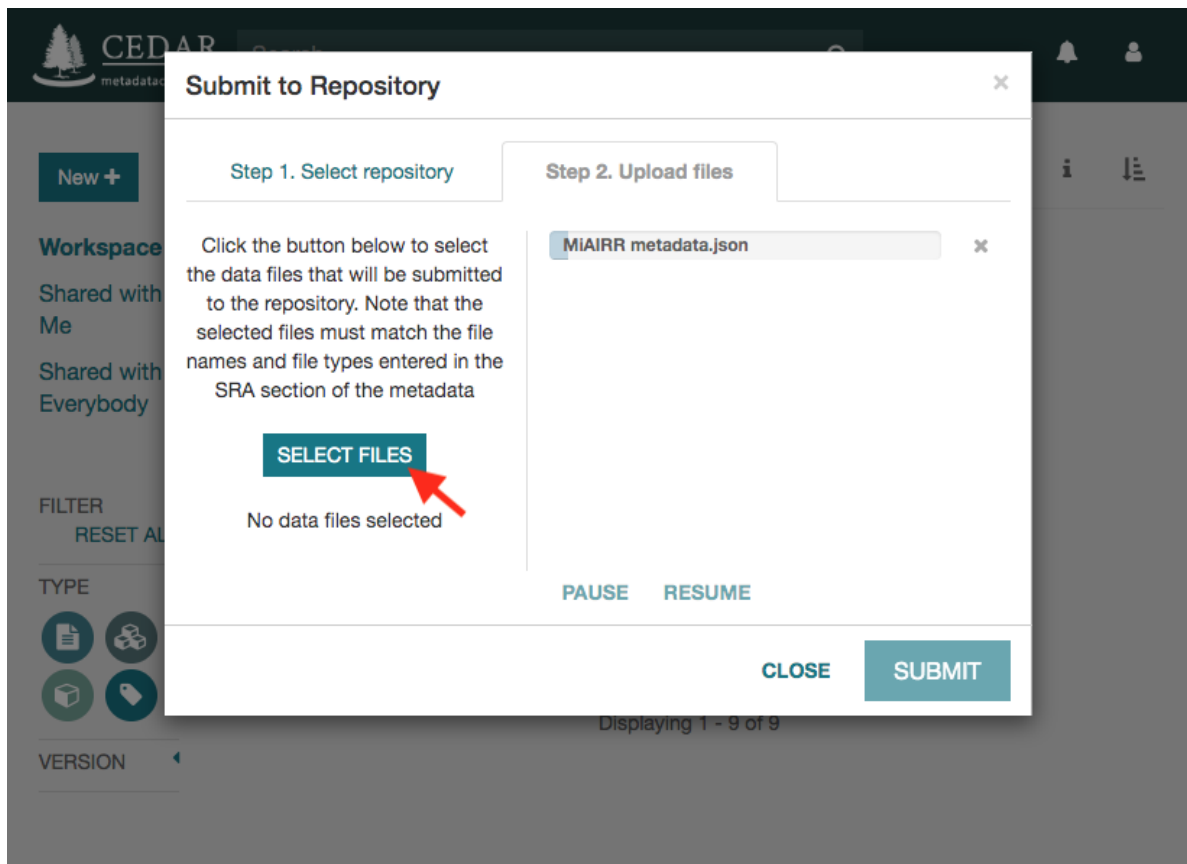
6. Click on the 'Submit' option to open the submission dialog.



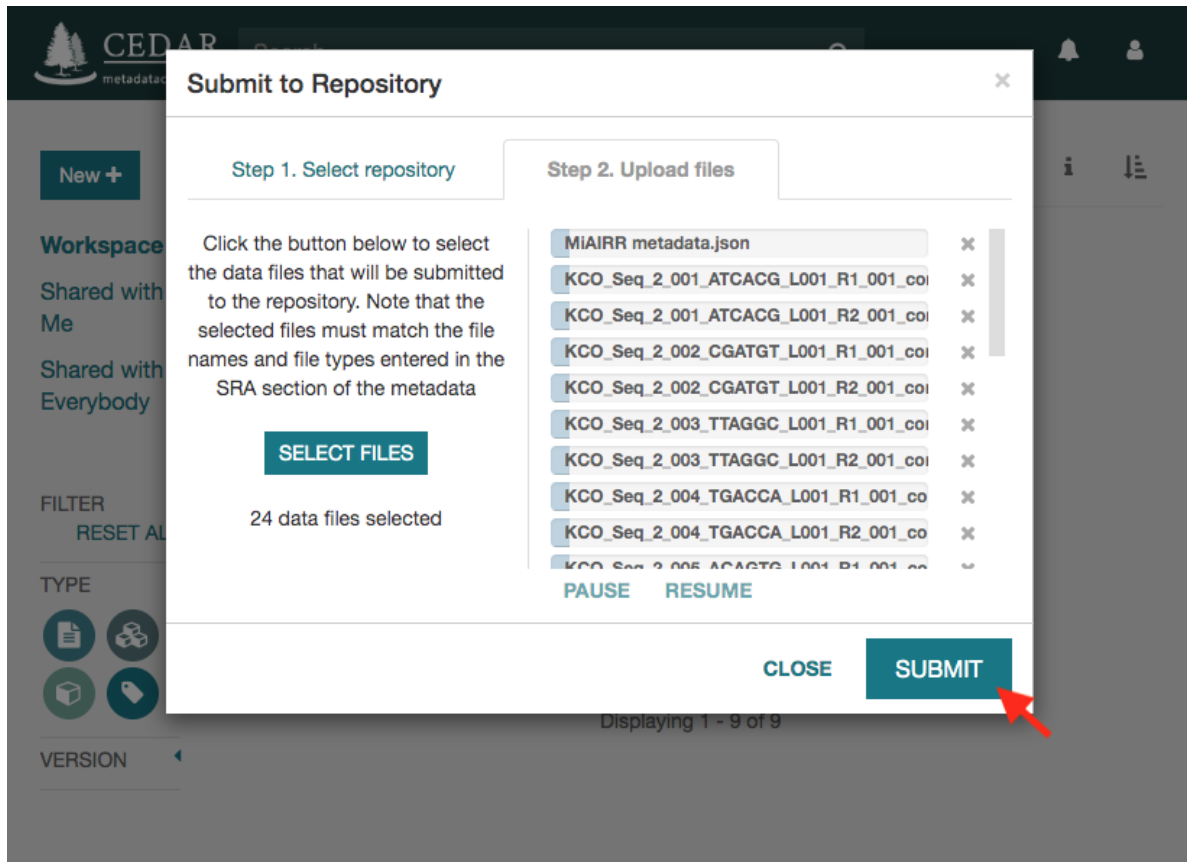
7. The 'NCBI MiAIRR' option will be automatically selected. Click on 'Next' to go to the next step.



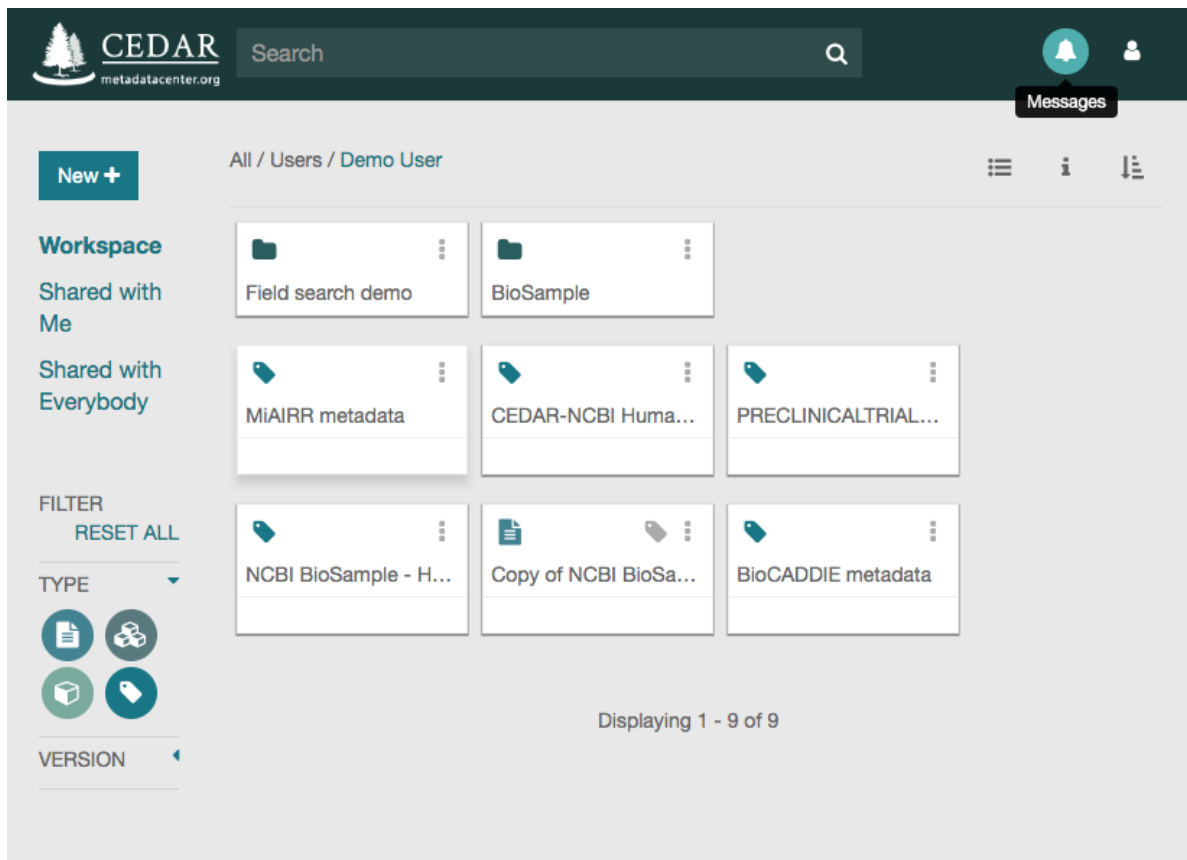
8. Click on the 'Select Files' button to upload the data files. Note that the names of the selected files must match the names in the metadata file. Otherwise, you will receive an error message when trying to start the submission.



9. Click on the 'Submit' button to start the submission. If there are not validation errors, the selected data files and the corresponding metadata will be uploaded to the NCBI servers.



10. Note that the submission may take several hours or even days to be processed by the NCBI. Meanwhile, you will receive status messages about your submission in your workspace (messages icon).



11. Proceed with deposit of processed data, below.

Citing the CAIRR pipeline

Bukhari, Syed Ahmad Chan, Martin J. O'Connor, Marcos Martínez-Romero, Attila L. Egyedi, Debra Debra Willrett, John Graybeal, Mark A. Musen, Florian Rubelt, Kei H. Cheung, and Steven H. Kleinstein. [The CAIRR pipeline for submitting standards-compliant B and T cell receptor repertoire sequencing studies to the NCBI](#). *Frontiers in Immunology* 9 (2018): 1877. DOI: 10.3389/fimmu.2018.01877

Tell Us About It

Please let us know how it went! If you are willing, we would love to have your comments in a [short survey](#), it should just take 5 minutes or so. We also welcome entry of issues and requests in [our GitHub repository](#), and emails can be sent to cedar-users@lists.stanford.edu. Both of these resources are publicly visible.

Support or Contact

Having trouble with NCBI submission process through our pipeline? Please email to [Syed Ahmad Chan Bukhari](#) or to [Marcos Martínez-Romero](#) and we will help you sort it out.

Option 2. Submission via NCBI's web interface

To facilitate AIRR data submissions to NCBI repositories, we have developed the NCBI-compliant metadata submission templates both for single and bulk AIRR data submissions. NCBI provides a web-based interface to create a BioProject and allows to BioSample, Sequence Read Archive (SRA) and GenBank metadata via tab-delimited files for single BioProject related data files submission.

Submitting AIRR data and associated metadata to the Bioproject, BioSample and SRA repositories via NCBI's web interface follows in general the submission procedure described in [\[NCBI_NBK47528\]](#), but uses AIRR-specific template for metadata submission:

1. Go to <https://submit.ncbi.nlm.nih.gov/subs/sra/> and login with your NCBI account (create an account if necessary).
2. Click on “create new submission”. You will see a form as below. Fill the form with required information and click on “continue”.

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information ahmadchan@gmail.com

Submission Portal HOME MY SUBMISSIONS GROUPS TEN

Sequence Read Archive (SRA) submission: SUB2154468
test project, Dec 05 '16

1 SUBMITTER 2 GENERAL INFO 3 PROJECT INFO 4 PUBLICATIONS 5 BIOSAMPLE TYPE 6 BIOSAMPLE ATTRIBUTES 7 SRA METADATA 8 FILES 9 OVERVIEW

Submitter Required fields are r

* First (given) name Middle name * Last (family) name

* E-mail (primary) E-mail (secondary) At least one e-mail should be from the organization's domain.

Select group for this submission

☐ None (affiliation from my personal profile)

☐ Center for Expanded Data Annotation and Retrieval

* Submitting organization Submitting organization URL * Department

Phone Fax

* Street * City * State/Province * Postal code * Country United States of America

Continue ☒ Update my contact information in profile

3. If you are submitting for the first time, check “Yes” on the “new BioProject” or “new BioSample” options to create a new project or sample, respectively.

Submission Portal

HOME MY SUBMISSIONS GROUPS TEMPLATES MY PR

Sequence Read Archive (SRA) submission: SUB2154468
test project, Dec 05 '16

1 SUBMITTER 2 GENERAL INFO 3 PROJECT INFO 4 PUBLICATIONS 5 BIOSAMPLE TYPE 6 BIOSAMPLE ATTRIBUTES 7 SRA METADATA 8 FILES 9 OVERVIEW

General Information Required fields are marked with aster

Do you want to create new BioProject?
* Yes No

Do you want to create new BioSamples for this submission?
* Yes No

Release date

* When should this submission be released to the public:
* Release immediately following processing (recommended)
* Release on specified date or upon publication, whichever is first
Note: Release of BioProject or BioSample is also triggered by the release of linked data.

Continue

- Fill in the project information. Add as much relevant information you can add in description. It will help later in searching the particular submission.

Project Info

* Project title

* Public description

Relevance

Fill project info

* Is your project part of a larger initiative which is already registered with NCBI?
* No Yes (not very common)

External Links

Link description	URL	Delete

Add another link

Select your grants

Use this tool to look up grants from many subscribed governmental funding agencies (eg NIH, CDC, FDA and VA) and some non-governmental funding sources (eg HHMI and Autism Speaks). You can search by grant number, title or grantee name.

Add grants

Consortium name Consortium URL

Data provider Data provider URL Delete

- The AIRR BioSample template is not yet listed on the NCBI website. The template sheet AIRR_BioSample_V1.0.xls can be downloaded from https://github.com/airr-community/airr-standards/tree/master/NCBI_implementation/templates_XLS. Fill in the required field and save the file as *tab-delimited* text file (.TSV format), then upload it.
- To submit the SRA metadata use the AIRR_SRA_v1.0.xls file. Make sure that the column `sample_name` uses sample names that match the record in the BioSample template (if new BioSamples are being submitted) or a previously entered record. Also this file must be saved as *tab-delimited* text file for upload.
- Submit the raw sequence file.
- Complete the submission.
- Proceed with deposit of processed data, below.

Option 3. Submission via NCBI's FTP server, using a predefined XML template

In addition to the web interface, NCBI provides an FTP-based solution to submit bulk metadata. The corresponding AIRR XML templates can be found under https://github.com/airr-community/airr-standards/tree/master/NCBI_implementation/templates_XLS. Otherwise users should refer to the current SRA file upload manual <https://www.ncbi.nlm.nih.gov/sra/docs/submitfiles/>. Users planning to frequently submit AIRR-seq data to SRA using scripts to generate the XML files MUST ensure that the templates are identical to the current upstream version on Github.

Step 2. Processed MiAIRR data submission to GenBank/TLS

Processed sequence data will be submitted to the “Targeted Locus Study” (TLS) section of GenBank. The details of this submission process are currently still being finalized. Basically the procedure is identical to a conventional GenBank submission with the exception of additional keywords marking it as TLS submission.

Non-productive records should be removed before the data submission or use an alternative annotation as described in the specification document.

- Generating MiAIRR compliant GenBank/TLS submissions: <https://changeo.readthedocs.io/en/stable/examples/genbank.html>

GenBank provides multiple tools (GUI and command-line) to submit data:

- BankIt, a web-based submission tool with wizards to guide the submission process
- Sequin, NCBI's stand-alone submission tool with wizards to guide the submission process is available by FTP for use on for Windows, macOS and Unix platforms.
- Tbl2asn is the recommended tool for the bulk data submission. It is a command-line program that automates the creation of sequence records files (.sqn) for submission to GenBank, driven by multiple tabular input data files. Documentation and download options can be found under <https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/>.

MiAIRR-to-NCBI Specification

Outline of INSDC reporting procedure

TODO: Outline the reporting procedure for data sets 1-4

In terms of standard compliance it is currently REQUIRED¹ to deposit information for MiAIRR data sets 5 and 6 in general-purpose sequence repositories for which an AIRR-accepted specification on information mapping MUST exist. However, users should note that in the future additional AIRR-sanctioned mechanisms for data deposition will become available as specified by the AIRR Common Repository Working Group. The mapping of data items in MiAIRR data sets 5 and 6 differs substantially in size and structure and therefore requires distinct reporting procedures:

- Set 5: This is free text information describing the work flow, tools and parameters of the sequence read processing. It is REQUIRED that this information is deposited as a freely available document, permanently linked via a DOI. Note that is currently neither a specific format for this document nor a recommended service provider for obtaining the DOI.
- Set 6: This is specified to contain the consensus sequence and the following information obtained from the initial analysis: V, D and J segment, C region and IMGT-JUNCTION² [LIGMDB_V12]. These will be deposited in a general-purpose INSDC repository, using the record structure described below.

¹ See the “Glossary” section on how to interpret term written in all-caps.

² Note that according to IMGT definition this is a superset of the CDR3.

INSDC records were originally designed to hold individual Sanger sequences. Therefore each record will contain a header with information largely identical between all records in an AIRR sequencing study. Records can be concatenated for uploading.

The INSDC feature table (FT) [INSDC_FT] is a sequence annotation standard used within the INSDC records and assigns information to specified positions on the reported sequence string. In regard to the correct location of the provided annotation, it should especially be noted that some V(D)J inference tools will return coordinates referring to the reference instead of the query sequence. As the sequence submitted in a record **MUST** be identical to the query sequence, the positions provided by the V(D)J inference tool **MUST**, if necessary, be translated back onto the query sequence. In case the start and/or end of a feature cannot be reliably determined or is not present in the reported sequence³, open intervals **CAN** be used for reporting. However, open intervals **MUST NOT** be used to deliberately obfuscate known positions.

In addition to the required information specified in [Table_1](#), users **CAN** use all valid FT keys/qualifiers to provide further annotation for the reported sequences. However, a record **MUST** still be compliant with this specification, if such **OPTIONAL** information would be removed, meaning that it is **FORBIDDEN** to move **REQUIRED** information into **OPTIONAL** keys/qualifiers. In addition, users **MUST NOT** use keys/qualifiers that could create ambiguity with the keys/qualifiers specified here.

element	FT key	FT qualifier	FT value	REQUIRED (if used by original study)
V segment	V_segment	/gene	see [Feature table]	yes
D segment	D_segment	/gene	see [Feature table]	yes; if <i>IGH</i> , <i>TRB</i> or <i>TRD</i> sequence
J segment	J_segment	/gene	see [Feature table]	yes
C region	C_region	/gene	see [Feature table]	yes
JUNCTION	CDS	/function	“JUNCTION”	yes

Table 1: Summary of the mapping of mandatory AIRR MiniStd data set 6 elements to the INSDC feature table (FT). Note that the overall record will contain additional information, such as cross-references linking the deposited sequence reads and metadata.

Element mapping

The broad strategy of element mapping to the various repositories is depicted in [Table_2](#).

MiAIRR data set / subset	target repository
1 / study	BioProject
1 / subject	
1 / diagnosis & treatment	
2 / sample	BioSample
3 / processing (cells)	SRA
3 / processing (nucleic acids)	
4 / raw sequences	
5 / processing (data)	user-defined DOI
6 / Processed sequences & annotations	Genbank

Table 2: Summary of the mapping of MiAIRR data sets to the various repositories

³ This can occur e.g. in paired-end sequencing of head-to-head concatenated transcripts, where the 5' end of the V segment is present in the amplicon, but cannot be precisely determined.

Mapping of data sets 1-4 to BioProject/BioSample/SRA

TODO: Include item-by-item mapping [NCBI_NBK47528]

Mapping of data set 5 to a user-defined repository

While several mandatory item have been defined in this data set, there is currently no mapping as the reporting procedure is implemented as a free text document. AIRR RECOMMENDS to use [Zenodo](#) for deposition of these documents, as it is hosted by CERN and supports versioned DOIs (termed “concept” DOI). Users SHOULD use the existing AIRR tag when submitting documents to increase the visibility of their study.

Mapping of data set 6 to INSDC

Users should note that while the FT is standardized, the overall sequence record structure diverges between the three INSDC repositories. The following section refers to items at or above the hierarchy level of the FT using the GenBank specification [GENBANK_FF], the corresponding designations of ENA [ENA_MANUAL] are provided in parenthesis¹¹.

Record header

The header MUST contain all of the following elements:

- **REQUIRED:** header structure as specified by the respective INSDC repository [ENA_MANUAL] [GENBANK_FF] [GENBANK_SR].
- **FORBIDDEN:** The DEFINITION entry will be autopopulated by information provided in the FT part (misc_feature, /note).
- **REQUIRED:** identifier of the associated SRA record (MiAIRR data set 4) as DBLINK (ENA: DR line). Note that it is **not** possible to refer to individual raw reads, only the full SRA collections can be linked.
- **REQUIRED:** in the KEYWORDS field (ENA: KW line):
 - the term “TLS”
 - the term “Targeted Locus Study”
 - the term “AIRR”
 - the term “MiAIRR:<x>.<y>” with <x> and <y> indicating the used version and subversion of the MiAIRR standard.
- **REQUIRED:** DOI of the associated free-text record containing the information on data processing (MiAIRR data set 5) as REMARK within a REFERENCE⁴ (ENA: RX line).
- **OPTIONAL:** The use of [structured comments](#) is currently evalutated for use in future versions of the MiAIRR standard.

¹¹ Note that there is currently no submission specification for ENA. This information is provided for reference only and will be moved to a separate document in the future.

⁴ The current GenBank record specification does not include a separate key for DOIs.

Feature table

The feature table, indicated by FEATURES (ENA: RX line), MUST or SHOULD contain the following keys/qualifiers:

General sequence information

- REQUIRED: key source containing the following qualifiers:
 - REQUIRED: qualifier /organism (required by [INSDC_FT]).
 - REQUIRED: qualifier /mol_type (required by [INSDC_FT]).
 - REQUIRED: qualifier /citation pointing to the reference in the header (REFERENCE, ENA: RN line) that links to the data set 5 document.
 - REQUIRED: qualifier /rearranged⁵.
 - REQUIRED: qualifier /note containing the AIRR_READ_COUNT keyword to indicate the read number used for the consensus. The criteria for selecting these reads and the procedure used to build the consensus SHOULD be reported as part of data set 5.
 - OPTIONAL: qualifier /note containing the AIRR_INDEX_CELL keyword for single-cell experiments. The value of the keyword SHOULD only contain alpha-numeric characters and MUST be identical for sequences derived from the same cell of origin.
 - RECOMMENDED: qualifiers /assembly_gap and /linkage_evidence to annotate non-overlapping paired-end sequences.
 - RECOMMENDED: qualifier /strain, if /organism is “Mus musculus”.

Note that additional qualifiers might be REQUIRED by GenBank to harmonize the GenBank record with the BioSample referenced by it in the header. A list of known BioSample keyword and GenBank qualifiers that MUST contain the same information can be found below. Whether (and in which direction) the existence of a keyword/qualifiers triggers a requirement in the corresponding record is currently unknown. Please report any undocumented requirements surfacing during submission to the MiAIRR team.

BioSample keyword	GenBank FT qualifier
cell type	/cell_type
isolate	/isolate
sex	/sex
tissue	/tissue_type

Segment and region annotation

The following keys MUST be used for annotation according to their FT definition, if the respective item has been reported by the original study:

- REQUIRED: key V_region. Note that this key MUST NOT be used to annotate V segment leader sequence⁶⁷.

⁵ Although FT does specify a /germline qualifier for non-rearranged sequences it has not been included in this specification as there is no obvious use case for it. In addition, non-rearranged transcripts would lack a number of other features that are assumed to be present, first of all the JUNCTION.

⁶ The FT explicitly states that V_segment does **not** cover the leader sequence. The definition of V_region is slightly more ambiguous, however in combination with the V_segment definition, it becomes clear that the leader is also not considered to be a part of V_region. Therefore the leader sequence should be implicitly annotated as the region between the start of CDS and the start of V_region.

⁷ Previously the leader was implicitly annotated as the region between CDS start and V_region start. As it was decided to drop the “global” CDS to make it easier to accommodate for INDELs, this is currently not an option anymore.

- REQUIRED: key `misc_feature` with coordinates identical to those given in `V_region`. This key MUST contain a `/note` qualifier that contains a string as value, which describes the general type of variable region described by the record. The string MUST match the regular expression

```
/^(immunoglobulin (heavy|light)|T cell receptor (alpha|beta|gamma|delta)) chain_
↪variable region$/
```

This string will be used as record heading upon import into Genbank. Note that while this behavior of Genbank is undocumented, the procedure has been approved by NCBI.

- REQUIRED: key `V_segment`, both coordinates MUST be within `V_region`. Note that this key MUST NOT be used to annotate V segment leader sequence^{Page 54, 6Page 54, 7}.
- REQUIRED: key `D_segment`, both coordinates MUST be within `V_region`. This key is only REQUIRED for sequences of applicable loci (*IGH*, *TRB*, *TRD*⁸). In the rare case of rearrangements using two D segments, this key SHOULD occur twice, but the coordinates of both keys MUST NOT overlap.
- REQUIRED: key `J_segment`, both coordinates MUST be within `V_region`.
- REQUIRED: key `C_region`, both coordinates MUST NOT overlap with `V_region`. If the region can be unambiguously identified, the respective official gene symbol MUST be reported using the `/gene` qualifier. If only the isotype (e.g. IgG) but not the subclass (e.g. IgG1) can be identified, a truncated gene symbol (e.g. IGHG instead of IGHG1) SHOULD be reported instead⁹.

Each `[VDJ]_segment` key MUST or SHOULD contain the following qualifiers:

- REQUIRED: qualifier `/gene`, containing the designation of the inferred segment, according to the database in the first `/db_xref` entry. This qualifier MUST NOT contain any allele information.
- RECOMMENDED: qualifier `/allele`, containing the designation of the inferred allele, according to the database in the first `/db_xref` entry. Note that while INSDC does not specify any format for this qualifier, AIRR compliance REQUIRES that this field only contains the allele string, i.e. without the gene name or separator characters.
- REQUIRED: qualifier `/db_xref`, linking to the reference record of the inferred segment in a germline database [\[INSDC_XREF\]](#). This qualifier can be present multiple times, however only the first entry is mandatory and MUST link to the database used for the segment designation given with `/gene` and (if present) `/allele`.

Note on referencing IMGT databases: There are two IMGT database available in the controlled vocabulary [\[INSDC_XREF\]](#):

- IMGT/GENE-DB: This is the genome database, which requires that a reference sequence has been mapped to genomic DNA. When using this database as reference, note that you can only refer to the gene symbol **not** the allele. In the case of ambiguous allele calls (see below) this means that you MUST NOT annotate any `/allele` at all. Nevertheless, this SHOULD be the default database for applications using IMGT as reference, as the sequence for each gene/allele is unique.
- IMGT/LIGM: This database collects sequences described in INSDC databases (GenBank/ENA/DDBJ). As it might contain multiple entries representing a given gene/allele, it is NOT RECOMMENDED to use it unless that inference gene/allele is only present in IMGT/LIGM and not in IMGT/GENE-DB.
- RECOMMENDED: `/inference` to indicate the tool used for segment inference. The description string SHOULD use COORDINATES as category and alignment as type [\[INSDC_FT\]](#).

Annotation of sequences producing multiple hits with identical scores is problematic and is ultimately at the discretion of the depositing researcher. However, the algorithms used for tie-breaking SHOULD be documented in data set 5. In addition, the following procedures MUST be followed:

⁸ For simplicity, this document only uses human gene symbols. For non-human species the specification pertains to the respective orthologs.

⁹ This approach has been approved by NCBI.

- Certain gene, ambiguous allele: If multiple alleles of the same gene match to the sequence, the `/allele` qualifier MUST NOT be used. As the REQUIRED `/db_xref` qualifier will often refer to a specific allele, all equal hits SHOULD be annotated via this qualifier (which can be used multiple times). Also see the note on the limitations of the IMGT/GENE-DB reference database above.
- Ambiguous gene: Pick one, annotate using the qualifiers as noted for ambiguous allele.

JUNCTION annotation

INSDC does currently not define a key to annotate JUNCTION¹⁰. Therefore the following procedure MUST be used:

- REQUIRED: key CDS, indicating the positions of
 1. the first bp of the first AA of JUNCTION
 2. the last bp of the last AA of JUNCTION as determined by the utilized V(D)J inference tool.

Open coordinates MUST be used for both coordinates to allow for automated creation of the `/translated` qualifier providing the peptide sequence. Further note that a non-productive JUNCTION can have a length not divisible by three. This key contains the following qualifiers:

- REQUIRED: qualifier `/codon_start` with the assigned value “1”.
- REQUIRED: qualifier `/function` with the assigned value “JUNCTION”.
- REQUIRED: qualifier `/product` with an assigned value matching the regular expression

```
/^(immunoglobulin (heavy|light)|T cell receptor (alpha|beta|gamma|delta)) chain_  
↪junction region$/
```

The variable region referred to in the string MUST be the same as the one given in the `misc_feature` key.

- RECOMMENDED: qualifier `/inference`, indicating the tool used for positional inference. The description string SHOULD use `COORDINATES` as category and `protein motif` as type [[INSDC_FT](#)].
- FORBIDDEN: qualifier `/translated`, which will be automatically added by Genbank.

Note that the complete CDS key will be removed by Genbank if the translation contains stop codons or too many “N” (exact number unknown). As such a record will lack a central piece of REQUIRED information it is RECOMMENDED that submitters either

- remove the complete record or
- replace the CDS with a `misc_feature` key while at the same time removing the `/codon_start` and `/product` qualifiers

upfront, as described in the submission manual. If the submitter chooses the replacement option, it has to be ensured that the annotated coordinates are actually valid and not affected by the frame-shift.

¹⁰ NCBI confirmed that once there would be enough datasets using the *JUNCTION* tag as specified here, a motion for an INSDC-sanctioned key could be initiated.

Record body

The record body starts after ORIGIN (ENA: SQ line) and MUST contain:

- the consensus sequence

References

Footnotes

Appendix

Example record (GenBank format)

```

LOCUS      AB123456                420 bp    mRNA    linear    EST 01-JAN-2015
DEFINITION TLS: Mus musculus immunoglobulin heavy chain variable region,
sequence.
ACCESSION  AB123456
VERSION    AB123456.7
KEYWORDS   TLS; Targeted Locus Study; AIRR; MiAIRR:1.0.
SOURCE     Mus musculus
  ORGANISM Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires;
            Rodentia; Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE  1 (bases 1 to 420)
  AUTHORS  Stibbons,P.
  TITLE    Section 5 information for experiment F001
  JOURNAL   published (01-JAN-2000) on Zenodo
  REMARK    DOI:10.1000/0000-12345678
REFERENCE  2 (bases 1 to 420)
  AUTHORS  Stibbons,P.
  TITLE    Direct Submission
  JOURNAL   Submitted (01-JAN-2000) Center for Transcendental Immunology,
            Unseen University, Ankh-Morpork, 12345, DISCWORLD
DBLINK     BioProject: PRJNA000001
            BioSample: SAMN000001
            Sequence Read Archive: SRR0000001
FEATURES   Location/Qualifiers
    source   1..420
             /organism="Mus musculus"
             /mol_type="mRNA"
             /strain="C57BL/6J"
             /citation=[1]
             /rearranged
             /note="AIRR_READ_COUNT:123"
    V_region 1..324
    misc_feature 1..324
             /note="immunoglobulin heavy chain variable region"
    V_segment 1..257
             /gene="IGHV1-34"

```

(continues on next page)

(continued from previous page)

```

        /allele="01"
        /db_xref="IMGT/LIGM:AC073565"
        /inference="COORDINATES:alignment:IgBLAST:1.6"
D_segment 266..272
        /gene="IGHD2-2"
        /allele="01"
        /db_xref="IMGT/LIGM:AJ851868"
        /inference="COORDINATES:alignment:IgBLAST:1.6"
J_segment 291..324
        /gene="IGHJ4"
        /allele="01"
        /db_xref="IMGT/LIGM:V00770"
        /inference="COORDINATES:alignment:IgBLAST:1.6"
CDS <258..>290
        /codon_start=1
        /function="JUNCTION"
        /product="immunoglobulin heavy chain junction region"
        /inference="COORDINATES:protein motif:IgBLAST:1.6"
        /translated="CARAGVYDGYTMDYW"
C_region 325..420
        /gene="Ighg2c"
ORIGIN
    1 agcctggggc ttcagtgaag atgtcctgca aggccttctgg ctacacattc actgactata
   61 acatacactg ggtgaagcag agccatggaa agagccttga gtggattgca tatattaatc
  121 ctaacaatgg tggttatggc tataacgaca agttcaggga caaggccaca ttgactgtcg
  181 acaggtcatc caacacagcc tacatggggc tccgcagcct gacctctgag gactctgcag
  241 tctattactg tgcaagagcg ggagtttacg acggatatac tatggactac tggggtaag
  301 gaacctcagt caccgtctcc tcagccaaaa caacagcccc atcggtctat cactggccc
  361 ctgtgtgtgg aggtacaact ggctcctcgg tgactctagg atgcctggtc aagggaact
//

```

Glossary

- **MUST / REQUIRED:** Indicates that an element or action is necessary to conform to the standard.
- **SHOULD / RECOMMENDED:** Indicates that an element or action is considered to be best practice by AIRR, but not necessary to conform to the standard.
- **CAN / OPTIONAL:** Indicates that it is at the discretion of the user to use an element or perform an action.
- **MUST NOT / FORBIDDEN:** Indicates that an element or action will be in conflict with the standard.

Abbreviations

- **AA:** amino acid
- **bp:** base pair
- **DOI:** digital object identifier
- **FT:** INSDC Feature Table
- **INSDC:** International Nucleotide Sequence Database Collaboration

- SRA: sequence read archive

Introduction

The MiAIRR standard

The MiAIRR standard (minimal information about adaptive immune receptor repertoires) is a minimal reporting standard for experiments using sequencing-based technologies to study adaptive immune receptors (e.g. T cell receptors or immunoglobulins). It is developed and maintained by the Minimal Standards Working Group of the [Adaptive Immune Receptors Repertoire \(AIRR\) Community](#) [Breden_2017]. The current version (1.0) of the standard has been recently published [Rubelt_2017] and was passed by the general assembly at the annual AIRR Community meeting in December 2017. MiAIRR requires researchers to report six sets of information:

1. study, subject, diagnosis & intervention
2. sample collection
3. sample processing and sequencing
4. raw sequencing data
5. data processing
6. processed sequences with a basic analysis results

However, MiAIRR only describes the mandatory data items that have to be reported, but neither provides details how and where to deposit data nor specifies data types and formats. Therefore this document aims to provide both a submission manual for users as well as a detailed data specification for developers.

Requirement Levels of AIRR Schema Fields

Clarification of Terms

- The terms “MUST”, “MUST NOT”, “REQUIRED”, “SHOULD”, “SHOULD NOT”, “RECOMMENDED”, “MAY” and “OPTIONAL” are to be interpreted as described in [RFC2119].
- The terms “IF” and “ONLY IF” are to be interpreted as sufficient and necessary requirement, respectively.
- The term “NULL-LIKE” is an extension of the NULL term in SQL and its equivalents in other programming languages, referring to the absence of data in a field (i.e., the field is empty). NULL-LIKE **additionally** includes the following terms, which also define the reason why the information is missing. As these terms are expected to be provided as text, the field would not be NULL but nevertheless NULL-LIKE (i.e., it lacks biologically interpretable information).
 - `not_applicable`: There is no meaningful value for this field due to study design (e.g., sex for a phage library).
 - `not_collected`: Data for this field was not collected during the study.
 - `missing`: Data for field was collected, but is not available now.

Categories of AIRR Schema Fields

- Fields **MUST** be indicated by the `x-airr:miairr` property IF and ONLY IF the field or its content is governed by the MiAIRR data standard [Rubelt_2017].
- The `x-airr:miairr` property **MUST** be assigned to one of the following three requirement levels:
 - **essential**: Information on this field **MUST** be provided and is considered critical for the meaningful interpretation of the data. Therefore the value of such a field **MUST NOT** be NULL-LIKE. Due to this strict requirement, this level is only assigned to a small set of fields. Importantly, fields are **not** elevated to this level based on the fact that the respective information should typically be available to the data generator. This was decided to simplify MiAIRR-compliant data annotation by third parties, who might perform this task based on publicly available information only.
 - **important**: Information for this field **MUST** be provided. However, the field **MAY** be assigned a NULL-LIKE value if the respective information is not available. The majority of fields governed by the MiAIRR data standard are assigned to this level.
 - **defined**: Information for this field **MAY** be provided. However, IF information matching the semantic definition of the field is provided, this field **MUST** be used for reporting.

Compliance with the MiAIRR Data Standard

- Compliance to the MiAIRR Data Standard is currently a binary state, i.e., a data either is or is not compliant, there are not “grades” of compliance. However, additional requirements for specific use cases might be defined in the future.
- Data sets are considered MiAIRR-compliant **ONLY IF** all **essential** and **important** fields are reported.
- Note that **important** fields with NULL-LIKE values **MUST NOT** be dropped from a data set.
- Implementors of data entry interfaces **SHOULD NOT** set the default value of **important** fields to NULL-LIKE values, i.e., users should be required to actively select the values.

Metadata Annotation Guidelines

Purpose of this Document

This document describes the **RECOMMENDED** ways to provide metadata annotation for various experimental setups.

Clarification of Terms

- The key words “**MUST**”, “**MUST NOT**”, “**REQUIRED**”, “**SHOULD**”, “**SHOULD NOT**”, “**RECOMMENDED**”, “**MAY**”, and “**OPTIONAL**” in this document are to be interpreted as described in [RFC2119].

Individual fields

library_generation_method

The `library_generation_method` describes how the nucleic acid annotated in `template_class` that encodes the V(D)J-rearrangement is reverse-transcribed, amplified and/or otherwise prepared for further processing. Typically this procedure will precede further NGS platform- specific steps, however these procedures MAY be combined. The field uses a controlled vocabulary, the individual values are described below:

template_	library_generat	Methodology
DNA	PCR	Conventional PCR on genomic DNA of a vertebrate host (requires: <code>synthetic == false</code>)
		Conventional PCR on DNA of a synthetic library (requires: <code>synthetic == true</code>)
RNA	RT(RHP)+PCR	RT-PCR using random hexamer primers
	RT(oligo-dT)+PC	RT-PCR using oligo-dT primers
	RT(oligo-dT)+TS	5'-RACE PCR (i.e. RT is followed by a template switch (TS) step) using oligo-dT primers
	RT(oligo-dT)+TS	5'-RACE PCR using oligo-dT primers and template switch primers containing unique molecular identifiers (UMI), i.e., the 5' end is UMI-coded
	RT(specific)+PC	RT-PCR using transcript-specific primers
	RT(specific)+TS	5'-RACE PCR using transcript- specific primers
	RT(specific)+TS	5'-RACE PCR using transcript- specific primers and template switch primers containing UMIs
	RT(specific+UMI	RT-PCR using transcript-specific primers containing UMIs (i.e., the 3' end is UMI-coded)
	RT(specific+UMI	5'-RACE PCR using transcript- specific primers containing UMIs (i.e., the 3' end is UMI-coded)
	RT(specific)+TS	RT-based generation of dsDNA without subsequent PCR. This is used by RNA-seq kits.
any	other	Any methodology not covered above

*_pcr_primer_target_location

The fields `forward_pcr_primer_target_location` and `reverse_pcr_primer_target_location` describe the location of the innermost primer used for a given locus. This information is critical to determine which positions of a sequence reflect a biological process and which are an artifact of the experiment.

The terms “proximal” and “distal” in the field description refer to the start of the respective V or C gene. Using human *TRAC* and a hypothetical primer as an example:

1	5	10	15	20	25	30	35	40	45	50	55	60	position (+1 = start of
↪ exon 1)													
...													
TGCCGTGTACCAGCTGAGAG													TRAC primer (reverse
↪ complement)													
ATATCCAGAACCCTGACCCTGCCGTGTACCAGCTGAGAGACTCTAAATCCAGTGACAAGT													Chr. 14:22,547,506-22,547,
↪ 565 (GRCh38)													
^ ^ ^													
[a] [b] [c]													markers

In this case:

- +19 (marker [a]) is the most distal untemplated nucleotide
- +20 (marker [b]) is the most proximal templated nucleotide
- +39 (marker [c]) is the most distal templated nucleotide

cell_subset

The `cell_subset` field is ontology-controlled, i.e., if present, it MUST refer to a Cell Ontology (CL) term via its `id` field. The field SHOULD NOT be used for values other than `lymphocyte` (CL:0000542) and its descendents. The reasoning behind this is that rearrangements of IG and TR loci are typically confined to this population, so that other nodes, do not provide further information. In addition, the field SHOULD only be used if the subset has been purified to a level that is comparable to flow cytometric cell sorting.

- In general, the provided annotation MUST NOT contradict the experimentally determined phenotype. E.g., if the experiment shows that the population is CD27+ a term that is explicitly defined in CL as CD27- MUST NOT be used.
- However, this does not mean that all markers listed in the description of an ontology term need to be confirmed, as long as the existing information is considered sufficient for classification and not contradictory (see above).
- In case the experimentally isolated cells do not match any “leaf” term, e.g., due to the isolation of multiple populations that contradict the definitions, the general advice is to move up the CL hierarchy to the most distal term that is no-contradictory. In this case, `cell_phenotype` should be used to define the markers that were used experimentally.
- Note that ontology-controlled fields allow exactly one term. Therefore, mixtures of defined cell populations either need to be demultiplexed, or – if this is not possible – use the last (i.e., most distal) common term of all cell populations involved. Again, `cell_phenotype` can be used to provide the markers used in the experiment.

Specific Use Cases and Experimental Setups

Peripheral Blood Mononuclear Cells (PBMCs)

PBMCs are frequently used starting material for AIRR-seq studies in humans and are prepared by a density-gradient centrifugation using Ficoll. As they constitute a mixture of myeloid and lymphoid cells, the following points should be taken into consideration when annotating experiments using PBMCs:

- The `cell_population` and `cell_phenotype` fields should be NULL as PBMCs are neither sufficiently pure nor do they exclusively contain cells of the lymphocytic lineage (see [cell_subset](#)).
- Note that while Cell Ontology does provide a term `peripheral blood mononuclear cell` (CL:2000001), this is a sister node of `lymphocyte` (CL:0000542) and therefore outside of the current specification.
- The typical annotation for PBMC is therefore as follows:

```
sample_type:"peripheral venous puncture"
tissue:
  label:"venous blood"
  id:"UBERON:0013756"
tissue_processing:"Ficoll gradient"
cell_subset:NULL
cell_phenotype:NULL
```

Synthetic libraries

In synthetic libraries (e.g., phage or yeast display), particles present genetically engineered constructs (e.g., scFv fusion receptors) on their surface. As this deviates substantially from other workflows, the following annotation SHOULD/MUST be used:

- In general, Subject should be interpreted as the initial library that undergoes a mutation/selection procedure.
- `synthetic`: MUST be set to `true`
- `species`: It is assumed that every synthetic library is derived from V and J genes that exist in some vertebrate species. This field SHOULD encode that species. Importantly, it MUST NOT encode the phage vector, the bacterial host or a comparable biological component of the library system that constitutes the presenting particle.
- `sample_type`: SHOULD be `NULL`.
- `single_cell`: Only `true` if individual particles are isolated and sequenced. Note that colonies or plaques, even if containing genetically identical particles, *per se* do not match this definition and therefore MUST be annotated as `false`.
- `cell_storage`: SHOULD be used for non-cellular particles analogously.
- `physical_linkage`: For scFv constructs the `hetero_prelinked` term MUST be used. VHH (i.e., camelid) libraries SHOULD annotate `none` as there is only a single rearrangement involved.

10X Chromium

The current 10X V(D)J Kits (07/2020, Rev. G) perform a fully nested PCR, in which only the reverse primers (i.e., complementary to the constant region) are Ig/TCR specific, while the forward primers anneal to the sequence of the template switch primer. For the purpose of annotation, this is considered a gene-specific amplification, therefore such experiments SHOULD be annotated as follows:

- `single_cell`: MUST be `true`
- `library_generation_method`: SHOULD be `RT(specific)+TS(UMI)+PCR`
- `pcr_target` MAY contain multiple entries, one for each locus that is potentially amplified. Within each entry (i.e., each `PCRTarget` object) the following annotations SHOULD be provided:
 - `pcr_target_locus`: The locus described by this object, using the controlled vocabulary defined in the AIRR schema. Note that each object can only describe one locus, multiple loci require multiple `PCRTarget` objects.
 - `forward_pcr_primer_target_location`: `NULL` (as it cannot be reliably determined).
 - `reverse_pcr_primer_target_location`: Locus and position according to the respective set of reverse primers.

Geolocation information

For questions regarding population genetics, the information about the membership of a given individual in a (potentially) distinct population group is of high interest. However, no reasonably complete ontology exists for the concept of “ethnicity”, which in addition is limited to humans (not to mention the challenge of assigning consistent values to this property). Therefore, the AIRR Schema uses annotation of the location of birth of an individual and the location where a given sample was taken as imperfect but at least operationalizable substitute.

The respective properties `Subject.location_birth` and `Sample.collection_location` are ontology controlled and expect values that are geographic locations. The provided value SHOULD allow to resolve at least the country-level

information of the annotated location. Higher granularity SHOULD only be provided, if this does not exposed the data subject to the risk of being re-identified via based on or assisted by this information.

2.3.2 AIRR Data Representations

AIRR Data Representations are versioned specifications that consist of a file format and a well-defined schema. The schema is provided in a machine-readable YAML document that follows the OpenAPI v2.0 specification. The schema defines the data model, field names, data types, and encodings for AIRR standard objects. Strict typing enables interoperability and data sharing between different AIRR-seq analysis tools and repositories, and some fields use a controlled vocabulary or an ontology for value restriction. Specification extensions are utilized to define AIRR-specific attributes.

FAIR Principles

We desire AIRR standard objects to be FAIR (findable, accessible, interoperable and reusable) [[Wilkinson_2016](#)]:

- findable: by giving AIRR standard objects a globally unique identifier
- accessible: by providing an API where AIRR standard objects can be queried and downloaded
- interoperable: by defining a OpenAPI schema for the AIRR standard objects
- reusable: by linking the AIRR standard objects together into a standard formats

AIRR Data Model

The MiAIRR standard defines the minimal information for submission and publication of AIRR-seq datasets. The standard defines a set of data elements for this information and organizes them into six high-level sets.

- Study, Subject and Diagnosis
- Sample Collection
- Sample Processing and Sequencing
- Raw Sequences
- Data Processing
- Processed Sequences with Annotations

However beyond these sets, MiAIRR does not define any structure, data model or relationship between the data elements. This provides flexibility for the information to be stored in various database repositories but is problematic for interoperability and reusability of that information by computer programs. The AIRR Data Model overcomes these issues by defining a schema for the MiAIRR data elements, structuring them within schema objects, defining the relationship between those objects, and defining a file format.

Here are the primary schema objects of the AIRR Data Model:

Schema Object	Description
Study	Information about the experimental study design, including the title of the study, laboratory contact information, funding, and linked publications.
Subject	Information about the study cohorts and individual subjects, including species, sex, age, and ancestry.
Diagnosis	Information about disease state(s), therapies, and study group membership (e.g., control versus disease).
Sample	Information about the origin and expected composition of the biological sample(s). This set aims to capture essential information about the collection of a sample, including its source (e.g., anatomical site), its provenance (provider), and the experimental condition (e.g., the time point during the course of a disease or treatment).
CellProc	Information about the cell subset being profiled, as defined by the investigator, and the flow cytometry or other markers used to select the subset. Additional information includes the number of cells per sample and whether cells were prepared in bulk or captured as single cells.
NucleicAcid	Information about nucleic acid sample type (e.g., RNA versus DNA) and how immune-receptor gene rearrangements were amplified and sequenced (for example, RACE-PCR versus multiplex PCR, paired PCR, and/or varying read length and sequencing chemistries).
SequencingRun	Information about the sequencing run, such as the number of reads, read lengths, quality control parameters, the sequencing kit and instrument(s) used, and run batch number. Also includes information about the raw data for the sequencing run (e.g., FASTQ files).
DataProcessing	Information about the data processing to transform the raw sequencing data into Rearrangements.
Repertoire	Composite object that combines the schema objects Study, Subject, Diagnosis, Sample, CellProcessing, NucleicAcidProcessing, SequencingRun, and DataProcessing. Each Repertoire has a unique identifier repertoire_id for linking with other data files, e.g. Rearrangements. Repertoires have their own schema and file format described here .
Rearrangement	Annotated sequences describing adaptive immune receptor chains. Rearrangements have their own schema and file format described here .
Clones	Information about inferred clones from a study. Clones have their own schema and file format described here .
Cells	Information about an observed Cell in a study. Cells have their own schema and file format described here .
CellExpressionProperties	Information about expression properties observed for a specific cell. CellExpression properties have their own schema and file format described here .
Receptor	Information about adaptive immune receptors (i.e., Ig and TCR) that are linked to observed Cells in a study. Receptors have their own schema and file format described here .
Germline	Lists the receptor germline sequences that have been identified for a single locus within a particular species or sub-species, together with supporting evidence and additional metadata to assist with sequence annotation. Brings together the subsidiary objects AlleleDescription, SequenceDelineationV, RearrangedSequence, UnrearrangedSequence, Acknowledgement.
Genotype	Lists the receptor germline sequences that have been identified within a single subject, including both those that are listed within GermlineSets and those that have not been so listed. References the subsidiary object Genotype, which covers a single locus.

Relationship between Schema Objects

The MiAIRR categories are hierarchical, and includes information about the study, the subjects, the collected samples and how they are processed, details of the sequencing protocol, and information about the data analysis. The top-down relationships are either 1-to-n indicating the top level object can be related to any number of sub-level objects, or n-to-n indicating any number of top level object can be related to any number of sub-level objects. Lastly, 1-to-1 indicates the top level object is related to a single sub-level object.

- **Study** 1-to-n with **Subject**. A study may contain any number of subjects.
- **Subject** 1-to-n with **Diagnosis**. Each subject may contain any number of diagnoses.
- **Subject** 1-to-n with **Sample**. Each subject may contain any number of samples.
- **Sample** 1-to-n with **CellProcessing**. A sample may have any number of cell processing records.
- **CellProcessing** 1-to-n with **NucleicAcidProcessing**. A cell processing record may have any number of nucleic acid processing records.
- **NucleicAcidProcessing** 1-to-n with **SequencingRun**. A nucleic acid processing records may have any number of sequencing runs.
- **SequencingRun** n-to-n with **DataProcessing**. Multiple sequencing runs can be combined in a data processing, and multiple data processing can be done on a sequencing run.

However, this hierarchy is deep and complicated. Therefore to simplify the processing of this information, we denormalized the hierarchy around the conceptual **Repertoire** object. This denormalization represents many relationships as 1-to-1 which simplifies the structure. A single **Repertoire** has these relationships with the primary schema objects.

- **Repertoire** 1-to-1 with **Study**. A repertoire is for a single study, though a study may have multiple repertoires.
- **Repertoire** 1-to-1 with **Subject**. A repertoire is for a single subject, though a subject may have other repertoires defined.
- **SampleProcessing** 1-to-1 with **Sample**, **CellProcessing**, **NucleicAcidProcessing**, and **SequencingRun**. A sample processing is a single chain from initial collection, through cell and nucleic acid processing, to sequencing.
- **Repertoire** 1-to-n with **SampleProcessing**. Generally a repertoire has a single sample processing, but sometimes studies perform technical replicates or re-sequencing to generate additional data, and these studies will have multiple sample processings, which are to be combined and analyzed together as part of the same repertoire.
- **Repertoire** 1-to-n with **DataProcessing**. A repertoire can be analyzed multiple times. More details about multiple data processing is provided below.

The trade-off with denormalization of the hierarchy is that it causes duplication of data. For example, two repertoires for the same study will have the **Study** information duplicated within each of the two repertoire records; likewise multiple repertoires for the same subject will have the **Subject** information duplicated.

While the denormalized **Repertoire** simplifies read-only access to the MiAIRR information, it complicates data entry and write access to the information because updates need to be propagated to all of the duplicate records. Therefore, **Repertoire** was designed to be easily transformed into a normalized form, representing the full hierarchy of the objects, by utilizing the `study_id`, `subject_id`, `sample_id`, and `sample_processing_id` fields to uniquely identify the **Study**, **Subject**, **Sample**, and **SampleProcessing** objects across multiple repertoires. The exception is that **CellProcessing** and **NucleicAcidProcessing** do not have their own unique identifiers, so they are included within **SampleProcessing**.

AIRR extension properties

The OpenAPI V2 and V3 specification provides the ability to define extension properties on schema objects. These are additional properties on the schema definition directly, not to be confused with additional properties on the data. These extension properties allow those schema definitions to be annotated with MiAIRR and AIRR specific information. Instead of creating separate extensions for each property, a single extension `x-airr` property is defined, which is an object that contains any number of properties. Within the AIRR schema, `AIRR_Extension` defines the schema for the `x-airr` object and the properties within it. Here is a list of the currently supported AIRR extension properties:

Extension	Description
<code>miairr</code>	Present if the annotated property is a MiAIRR data standard element. Always has a <i>requirement level</i> assigned to it.
<code>nullable</code>	Assumes <code>miairr</code> . False if the annotated property must not be NULL by the MiAIRR standard, otherwise True or null. This extension is not valid for OpenAPI V3 as the <code>nullable</code> builtin property should be used.
<code>set</code>	Assumes <code>miairr</code> . The MiAIRR set for the annotated property.
<code>subset</code>	Assumes <code>miairr</code> . The MiAIRR subset for the annotated property.
<code>name</code>	Assumes <code>miairr</code> . The MiAIRR field name.
<code>format</code>	Describes the format for the annotated property. Value is either <code>free text</code> , <code>controlled vocabulary</code> or <code>ontology</code> .
<code>ontology</code>	If <code>format=ontology</code> then this provides additional information about the ontology including draft status, name, URL and top node term.
<code>identifier</code>	True if the field is an identifier required to link metadata and/or individual sequence records across objects in the complete AIRR Data Model and ADC API.
<code>adc-query-s</code>	True if an ADC API implementation must support queries on the field. If false, query support for the field in ADC API implementations is optional.
<code>adc-api-opt</code>	True if the field is specific to the ADC API and is not part of the AIRR specification proper. These are typically “convenience” fields that make finding data easy or efficient (can be optimized by a repository).
<code>deprecated</code>	True if the field has been deprecated from the schema.
<code>deprecated-</code>	Information regarding the deprecation of the field.
<code>deprecated-</code>	The deprecated field is replaced by this list of fields.

Schema Definitions

Repertoire Schema

A **Repertoire** is an abstract organizational unit of analysis that is defined by the researcher and consists of study metadata, subject metadata, sample metadata, cell processing metadata, nucleic acid processing metadata, sequencing run metadata, a set of raw sequence files, data processing metadata, and a set of **Rearrangements**. A **Repertoire** gathers all of this information together into a composite object, which can be easily accessed by computer programs for data entry, analysis and visualization.

A **Repertoire** is specific to a single subject otherwise it can consist of any number of samples (which can be processed in different ways), any number of raw sequence files, and any number of rearrangements. It can also consist of any number of data processing metadata objects that describe the processing of raw sequence files into **Rearrangements**.

Typically, a **Repertoire** corresponds to the biological concept of the immune repertoire for that single subject which the researcher experimentally measures and computationally analyzes. However, researchers can have different interpretations about what constitutes the biological immune repertoire; therefore, the **Repertoire** schema attempts to be flexible and broadly useful for all AIRR-seq studies.

Another researcher can take the same raw sequencing data and associated metadata and create their own **Repertoire** that is different from the original researcher's. A common example is to define a repertoire that is a subset such as "productive rearrangements for IGHV4" whereas the original researcher defined a more generic "B cell repertoire". This new **Repertoire** would have much of the same metadata as the original **Repertoire**, except associated with a different study, and with additional information in the data processing metadata that describes how the rearrangements were filtered down to just the "productive rearrangements for IGHV4". Likewise, another researcher may get access to the original biosample material and perform their own sample processing and sequencing, which also would be a new **Repertoire**. That new **Repertoire** could combine samples from the original researcher's **Repertoire** with the new sample data as a large dataset for the subject.

Multiple Data Processing on a Repertoire

Data processing can be a complicated multi-stage process. Documenting the process in a formal way is challenging because of the diversity of actions that may be performed. The MiAIRR standard requires documentation of the process but in an informal way with free text descriptions. A **Repertoire** might undergo multiple different data processing for any number of reasons, e.g. to compare the results from different toolchains, or to compare different settings for the same toolchain.

It is expected that all of the **Samples** of a **Repertoire** will be processed together within a **DataProcessing**. That is, a **DataProcessing** that only uses some but not all samples in a **Repertoire** could be confusing to users and appear as though data is missing. Likewise, processing some samples within a **Repertoire** with one **DataProcessing** and the remaining samples with a different **DataProcessing** could also confuse users. Because **DataProcessing** is unstructured information, it is not possible to validate that all **Samples** in a **Repertoire** are being processed together, so this expectation cannot be strictly enforced.

Having multiple **DataProcessing** for a **Repertoire** will create multiple sets of **Rearrangements** that are distinct and separate from each other. Analysis tools need to be careful not to mix these sets of **Rearrangements** from different **DataProcessing** because it can generate incorrect results. The identifier `data_processing_id` was added so **Rearrangements** can identify their specific **DataProcessing**.

Linking Data

Each **Repertoire** has a unique `repertoire_id` identifier. This identifier should be globally unique so that repertoires from multiple studies can be combined together without conflict. The `repertoire_id` is used to link other AIRR data to a **Repertoire**. Specifically, the *Rearrangements Schema* includes `repertoire_id` for referencing the specific **Repertoire** for that **Rearrangement**.

If a **Repertoire** has multiple **DataProcessing** then `data_processing_id` should be used to distinguish the appropriate **DataProcessing** within the **Repertoire**. The **Rearrangements** contains `data_processing_id` for this purpose. The `data_processing_id` is only unique within a **Repertoire** so `repertoire_id` should first be used to get the appropriate **Repertoire** object and then `data_processing_id` used to acquire the appropriate **DataProcessing**.

It is expected that typical **Repertoires** might only have a single **DataProcessing**, in which case `repertoire_id` and `data_processing_id` will be semantically equivalent and only the former should be used.

If a **Repertoire** has multiple sample processing objects in the sample array then `sample_processing_id` should be used to distinguish the the appropriate sample processing object within the **Repertoire**. The **Rearrangement** object can contain a `sample_processing_id` to uniquely identify a sample processing object within a **Repertoire**. Like `data_processing_id`, the `sample_processing_id` is only unique within the **Repertoire** so `repertoire_id` should first be used to get the appropriate **Repertoire** object and then `sample_processing_id` should be used to determine the appropriate sample processing object that is associated with the **Rearrangement**. If the **Rearrangement** object does not have a `sample_processing_id` then it can be assumed that the rearrangement is associated with all of the samples in the **Repertoire** (e.g. the rearrangement is a collapsed rearrangement across multiple samples).

It is expected that `Repertoires` might often have a single sample processing object, in which case `repertoire_id` and `sample_processing_id` will be semantically equivalent and only the former should be used.

Finally, if it is necessary to link a `Rearrangement` object with a unique pairing of sample processing and `DataProcessing`, the `repertoire_id` of the `Rearrangement` object should be used to identify the correct `Repertoire` object and then the `data_processing_id` should be used to identify the correct `DataProcessing` metadata and the `sample_processing_id` should be used to identify the correct sample processing metadata within that `Repertoire`.

Duality between Repertoires and Rearrangements

There is an important duality relationship between `Repertoires` and `Rearrangements`, specifically with the experimental protocols described in the `Repertoire` versus the annotations on `Rearrangements`. A `Repertoire` defines an experimental design for what a researcher intends to measure or observe, while the `Rearrangements` are what was actually measured and observed. Technically, the border between the two occurs at sequencing, that is when the biological physical entity (prepared DNA) is measured and recorded as information (nucleotide sequence).

This duality is important when considering how to answer certain questions. For example, `locus` for `Rearrangements` may have the value “IGH” which indicates that B cell heavy chain receptors were measured, yet the `Repertoire` might have “T cell” in `cell_subset` which indicates the researcher intended to measure T cells. This conflict between the two indicates something is wrong. Differences can occur in many ways, as with errors in the experimental protocol, or data processing might have incorrectly processed the raw sequencing data leading to invalid annotations.

File Format Specification

Files are YAML/JSON with a structure defined below. Files should be encoded as UTF-8. Identifiers are case-sensitive. Files should have the extension `.yaml`, `.yml`, or `.json`.

File Structure

- The file as a whole is considered a dictionary (key/value pair) structure with the keys `Info` and `Repertoire`.
- The file can (optionally) contain an `Info` object, at the beginning of the file, based upon the `Info` schema in the OpenAPI V2 specification. If provided, `version` in `Info` should reference the version of the AIRR schema for the file.
- The file should correspond to a list of `Repertoire` objects, using `Repertoire` as the key to the list.
- Each `Repertoire` object should contain a top-level key/value pair for `repertoire_id` that uniquely identifies the repertoire.
- Some fields require the use of a particular ontology or controlled vocabulary.
- The structure is the same regardless of whether the data is stored in a file or a data repository. For example, The *ADC API* will return a properly structured JSON object that can be saved to a file and used directly without modification.

Repertoire Fields

Download as TSV

Name	Type	Attributes	Definition
repertoire_id	string	optional, identifier, nullable	Identifier for the repertoire object. This identifier should be globally unique so that repertoires from multiple studies can be combined together without conflict. The repertoire_id is used to link other AIRR data to a Repertoire. Specifically, the Rearrangements Schema includes repertoire_id for referencing the specific Repertoire for that Rearrangement.
repertoire_name	string	optional, nullable	Short generic display name for the repertoire
repertoire_description	string	optional, nullable	Generic repertoire description
study	<i>Study</i>	required	Study object
subject	<i>Subject</i>	required	Subject object
sample	array of SampleProcessing	required	List of Sample Processing objects
data_processing	array of <i>DataProcessing</i>	required	List of Data Processing objects

Study Fields

Download as TSV

Name	Type	Attributes	Definition
study_id	string	required, identifier, nullable	Unique ID assigned by study registry such as one of the International Nucleotide Sequence Database Collaboration (INSDC) repositories.
study_title	string	required, nullable	Descriptive study title
study_type	<i>Ontology</i>	required, nullable	Type of study design
study_description	string	optional, nullable	Generic study description
inclusion_exclusion	string	required, nullable	List of criteria for inclusion/exclusion for the study
grants	string	required, nullable	Funding agencies and grant numbers
study_contact	string	optional, nullable	Full contact information of the contact persons for this study This should include an e-mail address and a persistent identifier such as an ORCID ID.
collected_by	string	required, nullable	Full contact information of the data collector, i.e. the person who is legally responsible for data collection and release. This should include an e-mail address and a persistent identifier such as an ORCID ID.
lab_name	string	required, nullable	Department of data collector
lab_address	string	required, nullable	Institution and institutional address of data collector
submitted_by	string	required, nullable	Full contact information of the data depositor, i.e., the person submitting the data to a repository. This should include an e-mail address and a persistent identifier such as an ORCID ID. This is supposed to be a short-lived and technical role until the submission is released.
pub_ids	string	required, nullable	Publications describing the rationale and/or outcome of the study. Where ever possible, a persistent identifier should be used such as a DOI or a Pubmed ID
keywords_study	array of string	required, nullable	Keywords describing properties of one or more data sets in a study. “contains_schema” keywords indicate that the study contains data objects from the AIRR Schema of that type (Rearrangement, Clone, Cell, Receptor) while the other keywords indicate that the study design considers the type of data indicated (e.g. it is possible to have a study that “contains_paired_chain” but does not “contains_schema_cell”).
adc_publish_date	string	optional, nullable	Date the study was first published in the AIRR Data Commons.
adc_update_date	string	optional, nullable	Date the study data was updated in the AIRR Data Commons.

Subject Fields

Download as TSV

Name	Type	Attributes	Definition
subject_id	string	required, identifier, nullable	Subject ID assigned by submitter, unique within study. If possible, a persistent subject ID linked to an INSDC or similar repository study should be used.
synthetic	boolean	required	TRUE for libraries in which the diversity has been synthetically generated (e.g. phage display)
species	<i>Ontology</i>	required	Binomial designation of subject's species
organism	<i>Ontology</i>	DEPRECATED	Binomial designation of subject's species
sex	string	required, nullable	Biological sex of subject
age_min	number	required, nullable	Specific age or lower boundary of age range.
age_max	number	required, nullable	Upper boundary of age range or equal to age_min for specific age. This field should only be null if age_min is null.
age_unit	<i>Ontology</i>	required, nullable	Unit of age range
age_event	string	required, nullable	Event in the study schedule to which <i>Age</i> refers. For NCBI BioSample this MUST be <i>sampling</i> . For other implementations submitters need to be aware that there is currently no mechanism to encode to potential delta between <i>Age event</i> and <i>Sample collection time</i> , hence the chosen events should be in temporal proximity.
age	string	DEPRECATED	
ancestry_population	<i>Ontology</i>	required, nullable	Broad geographic origin of ancestry (continent)
location_birth	<i>Ontology</i>	optional, nullable	Self-reported location of birth of the subject, preferred granularity is country-level
ethnicity	string	required, nullable	Ethnic group of subject (defined as cultural/language-based membership)
race	string	required, nullable	Racial group of subject (as defined by NIH)
strain_name	string	required, nullable	Non-human designation of the strain or breed of animal used
linked_subjects	string	required, nullable	Subject ID to which <i>Relation type</i> refers
link_type	string	required, nullable	Relation between subject and <i>linked_subjects</i> , can be genetic or environmental (e.g. exposure)
diagnosis	array of <i>Diagnosis</i>	optional	Diagnosis information for subject
genotype	SubjectGenotype	optional, nullable	

Diagnosis Fields

Download as TSV

Name	Type	Attributes	Definition
study_group_descr	string	required, nullable	Designation of study arm to which the subject is assigned to
disease_diagnosis	<i>Ontology</i>	required, nullable	Diagnosis of subject
disease_length	string	required, nullable	Time duration between initial diagnosis and current intervention
disease_stage	string	required, nullable	Stage of disease at current intervention
prior_therapies	string	required, nullable	List of all relevant previous therapies applied to subject for treatment of <i>Diagnosis</i>
immunogen	string	required, nullable	Antigen, vaccine or drug applied to subject at this intervention
intervention	string	required, nullable	Description of intervention
medical_history	string	required, nullable	Medical history of subject that is relevant to assess the course of disease and/or treatment

Sample Fields

Download as TSV

Name	Type	Attributes	Definition
sample_id	string	required, identifier, nullable	Sample ID assigned by submitter, unique within study. If possible, a persistent sample ID linked to INSDC or similar repository study should be used.
sample_type	string	required, nullable	The way the sample was obtained, e.g. fine-needle aspirate, organ harvest, peripheral venous puncture
tissue	<i>Ontology</i>	required, nullable	The actual tissue sampled, e.g. lymph node, liver, peripheral blood
anatomic_site	string	required, nullable	The anatomic location of the tissue, e.g. Inguinal, femur
disease_state_sar	string	required, nullable	Histopathologic evaluation of the sample
collection_time_r	number	required, nullable	Time point at which sample was taken, relative to <i>Collection time event</i>
collection_time_r	<i>Ontology</i>	required, nullable	Unit of Sample collection time
collection_time_r	string	required, nullable	Event in the study schedule to which <i>Sample collection time</i> relates to
collection_locati	<i>Ontology</i>	optional, nullable	Location where the sample was taken, preferred granularity is country-level
biomaterial_provi	string	required, nullable	Name and address of the entity providing the sample

Tissue and Cell Processing Fields

[Download as TSV](#)

Name	Type	Attributes	Definition
tissue_processing	string	required, nullable	Enzymatic digestion and/or physical methods used to isolate cells from sample
cell_subset	<i>Ontology</i>	required, nullable	Commonly-used designation of isolated cell population
cell_phenotype	string	required, nullable	List of cellular markers and their expression levels used to isolate the cell population
cell_species	<i>Ontology</i>	optional, nullable	Binomial designation of the species from which the analyzed cells originate. Typically, this value should be identical to <i>species</i> , in which case it SHOULD NOT be set explicitly. However, there are valid experimental setups in which the two might differ, e.g., chimeric animal models. If set, this key will overwrite the <i>species</i> information for all lower layers of the schema.
single_cell	boolean	required, nullable	TRUE if single cells were isolated into separate compartments
cell_number	integer	required, nullable	Total number of cells that went into the experiment
cells_per_reaction	integer	required, nullable	Number of cells for each biological replicate
cell_storage	boolean	required, nullable	TRUE if cells were cryo-preserved between isolation and further processing
cell_quality	string	required, nullable	Relative amount of viable cells after preparation and (if applicable) thawing
cell_isolation	string	required, nullable	Description of the procedure used for marker-based isolation or enrich cells
cell_processing_protocol	string	required, nullable	Description of the methods applied to the sample including cell preparation/ isolation/enrichment and nucleic acid extraction. This should closely mirror the Materials and methods section in the manuscript.

Nucleic Acid Processing Fields

[Download as TSV](#)

Name	Type	Attributes	Definition
template_class	string	required	The class of nucleic acid that was used as primary starting material for the following procedures
template_quality	string	required, nullable	Description and results of the quality control performed on the template material
template_amount	number	required, nullable	Amount of template that went into the process
template_amount_u	<i>Ontology</i>	required, nullable	Unit of template amount
library_generatic	string	required	Generic type of library generation
library_generatic	string	required, nullable	Description of processes applied to substrate to obtain a library that is ready for sequencing
library_generatic	string	required, nullable	When using a library generation protocol from a commercial provider, provide the protocol version number
pcr_target	array of <i>PCR-Target</i>	optional	If a PCR step was performed that specifically targets the IG/TR loci, the target and primer locations need to be provided here. This field holds an array of PCRTarget objects, so that multiplex PCR setups amplifying multiple loci at the same time can be annotated using one record per locus. PCR setups not targeting any specific locus must not annotate this field but select the appropriate library_generation_method instead.
complete_sequence	string	required	To be considered <i>complete</i> , the procedure used for library construction MUST generate sequences that 1) include the first V gene codon that encodes the mature polypeptide chain (i.e. after the leader sequence) and 2) include the last complete codon of the J gene (i.e. 1 bp 5' of the J->C splice site) and 3) provide sequence information for all positions between 1) and 2). To be considered <i>complete & untemplated</i> , the sections of the sequences defined in points 1) to 3) of the previous sentence MUST be untemplated, i.e. MUST NOT overlap with the primers used in library preparation. <i>mixed</i> should only be used if the procedure used for library construction will likely produce multiple categories of sequences in the given experiment. It SHOULD NOT be used as a replacement of a NULL value.
physical_linkage	string	required	In case an experimental setup is used that physically links nucleic acids derived from distinct <i>Rearrangements</i> before library preparation, this field describes the mode of that linkage. All <i>hetero_*</i> terms indicate that in case of paired-read sequencing, the two reads should be expected to map to distinct IG/TR loci. <i>_head-head</i> refers to techniques that link the 5' ends of transcripts in a single-cell context. <i>_tail-head</i> refers to techniques that link the 3' end of one transcript to the 5' end of another one in a single-cell context. This term does not provide any information whether a continuous reading-frame between the two is generated. <i>_prelinked</i> refers to constructs in which the linkage was already present on the DNA level (e.g. scFv).

PCR Target Locus Fields

Download as TSV

Name	Type	Attributes	Definition
pcr_target_locus	string	required, nullable	Designation of the target locus. Note that this field uses a controlled vocabulary that is meant to provide a generic classification of the locus, not necessarily the correct designation according to a specific nomenclature.
forward_pcr_prime	string	required, nullable	Position of the most distal nucleotide templated by the forward primer or primer mix
reverse_pcr_prime	string	required, nullable	Position of the most proximal nucleotide templated by the reverse primer or primer mix

Raw Sequence Data Fields

Download as TSV

Sequencing Run Fields

Download as TSV

Name	Type	Attributes	Definition
sequencing_run_id	string	required, identifier, nullable	ID of sequencing run assigned by the sequencing facility
total_reads_passi	integer	required, nullable	Number of usable reads for analysis
sequencing_platfc	string	required, nullable	Designation of sequencing instrument used
sequencing_facili	string	required, nullable	Name and address of sequencing facility
sequencing_run_da	string	required, nullable	Date of sequencing run
sequencing_kit	string	required, nullable	Name, manufacturer, order and lot numbers of sequencing kit
sequencing_files	Sequencing-Data	optional	Set of sequencing files produced by the sequencing run

Data Processing Fields

Download as TSV

Name	Type	Attributes	Definition
data_processing_i	string	optional, identifier, nullable	Identifier for the data processing object.
primary_annotatic	boolean	optional, identifier	If true, indicates this is the primary or default data processing for the repertoire and its rearrangements. If false, indicates this is a secondary or additional data processing.
software_versions	string	required, nullable	Version number and / or date, include company pipelines
paired_reads_asse	string	required, nullable	How paired end reads were assembled into a single receptor sequence
quality_threshold	string	required, nullable	How/if sequences were removed from (4) based on base quality scores
primer_match_cut	string	required, nullable	How primers were identified in the sequences, were they removed/masked/etc?
collapsing_method	string	required, nullable	The method used for combining multiple sequences from (4) into a single sequence in (5)
data_processing_r	string	required, nullable	General description of how QC is performed
data_processing_f	array of string	optional, nullable	Array of file names for data produced by this data processing.
germline_database	string	required, nullable	Source of germline V(D)J genes with version number or date accessed.
germline_set_ref	string	optional, nullable	Unique identifier of the germline set and version, in standardized form (Repo:Label:Version)
analysis_provenan	string	optional, nullable	Identifier for machine-readable PROV model of analysis provenance

Rearrangement Schema

A Rearrangement is a sequence which describes a rearranged adaptive immune receptor chain (e.g., antibody heavy chain or TCR beta chain) along with a host of annotations. These annotations are defined by the AIRR Rearrangement schema and comprises eight categories.

Category	Description
Input	The input sequence to the V(D)J assignment process.
Identifiers	Primary and foreign key identifiers for linking AIRR data across files and databases.
Primary Annotations	The primary outputs of the V(D)J assignment process, which includes the gene locus, V, D, J, and C gene calls, various flags, V(D)J junction sequence, copy number (<code>duplicate_count</code>), and the number of reads contributing to a consensus input sequence (<code>consensus_count</code>).
Alignment Annotations	Detailed alignment annotations including the input and germline sequences used in the alignment; score, identity, statistical support (E-value, likelihood, etc); and the alignment itself through CIGAR strings for each aligned gene.
Alignment Positions	The start/end positions for genes in both the input and germline sequences.
Region Sequence	Sequence annotations for the framework regions (FWRs) and complementarity-determining regions (CDRs).
Region Positions	Positional annotations for the framework regions (FWRs) and complementarity-determining regions (CDRs).
Junction Lengths	Lengths for junction sub-regions associated with aspects of the V(D)J recombination process.

File Format Specification

Data for `Rearrangement` or `Alignment` objects are stored as rows in a *tab-delimited* file and should be compatible with any TSV reader.

Encoding

- The file should be encoded as ASCII or UTF-8.
- Everything is case-sensitive.

Dialect

- The record separator is a newline `\n` and the field separator is a tab `\t`.
- Fields or data should not be quoted.
- A header line with the AIRR-specified column names is always required.
- Values must not contain tab or newline characters.
- Values should avoid `@`, `#`, and quote (`"` or `'`) characters, as the result may be implementation dependent.
- Nested delimiters are not supported by the schema explicitly and should be avoided. However, if multiple values must be reported in a single column for an application specific reason, then the use of a comma as the delimiter is recommended.

File names

AIRR formatted TSV files should end with `.tsv`.

File Structure

The data file has two sections in this order:

1. Header. A single line with column names.
2. Data values. One record per line.

A comment section preceding the header (e.g., `#` or `@` blocks) is not part of the specification, but such a section is reserved for potential inclusion in a future release. As such, a comment section should not be included in the file as it *may* be incompatible with a future specification.

Header

A single line containing the column names and specifying the field order. Any field that corresponds to one of the defined fields should use the specified field name.

Required columns

Some of the fields are defined as **required** and therefore must always be present in the header. Note, however, that all columns allow for null values. Therefore, required columns exist to define a core set of fields that are always present in the table structure, but do not mandate that a value be reported.

Custom columns

There are no restrictions on inclusion of additional custom columns in the Rearrangements file, provided such columns do not use the same name as an existing required or optional field. It is recommended that custom fields follow the same naming scheme as existing fields. Meaning, `snake_case` with narrowing scope when read from left to right. For example, `sequence_id` is the “*identifier of the query sequence*”.

Consider submitting a pull request for a field name reservation to the [airr-standards repository](#) if the field may be broadly useful.

Ordering

There are no requirements that fields or records be sorted or ordered in any specific way. However, the field ordering provided by the schema is a recommended default, with top-to-bottom equating to left-to-right.

Data Values

The possible data types are `string`, `boolean`, `number` (floating point), `integer`, and `null` (empty string).

Boolean values

Boolean values must be encoded as T for true and F for false.

Null values

All fields may contain null values. This includes columns that are described as `required`. A null value should be encoded as an empty string.

Coordinate numbering

All alignment sequence coordinates use the same scheme as IMGT and INSDC (DDBJ, ENA, GenBank), with the exception that partial coordinate information should not be used in favor of simply assigning the start/end of the alignment. Meaning, coordinates should be provided as 1-based values with closed intervals, without the use of `>` or `<` annotations that denoted a partial region.

CIGAR specification

Alignments details are specified using the CIGAR format as defined in the [SAM specifications](#), with some vocabulary restrictions on the use of clipping, skipping, and padding operators.

The CIGAR string defines the reference sequence as the germline sequence of the given gene or region; e.g., for `v_cigar` the reference is the V gene germline sequence. The query sequence is what was input into the alignment tool, which must correspond to what is contained in the `sequence` field of the Rearrangement data. For the majority of use cases, this will necessarily exclude alignment spacers from the CIGAR string, such as IMGT numbering gaps. However, any gaps appearing in the query sequence should be accounted for in the CIGAR string so that the alignment between the query and reference is correctly represented.

The valid operator sets and definitions are as follows:

Operator	Description
<code>=</code>	An identical non-gap character.
<code>X</code>	A differing non-gap character.
<code>M</code>	A positional match in the alignment. This can be either an identical (<code>=</code>) or differing (<code>x</code>) non-gap character.
<code>D</code>	Deletion in the query (gap in the query).
<code>I</code>	Insertion in the query (gap in the reference).
<code>S</code>	Positions that appear in the query, but not the reference. Used exclusively to denote the start position of the alignment in the query. Should precede any <code>N</code> operators.
<code>N</code>	A space in the alignment. Used exclusively to denote the start position of the alignment in the reference. Should follow any <code>S</code> operators.

Note, the use of either the `=/X` or `M` syntax is valid, but should be used consistently. While leading `S` and `N` operators are required, trailing `S` and `N` operators are optional.

For example, an D gene alignment that starts at position 419 in the query sequence (leading 418S), that is 16 nucleotides long with no indels (middle 16M), has an 10 nucleotide 5' deletion (leading 10N), a 5 nucleotide 3' deletion (trailing 5N), and ends 72 nucleotides from the end of the query sequence (trailing 71S) would have the following D gene CIGAR string (`d_cigar`) and positional information:

Field	Value
<code>d_cigar</code>	418S10N16M71S5N
<code>d_sequence_start</code>	419
<code>d_sequence_end</code>	434
<code>d_germline_start</code>	11
<code>d_germline_end</code>	26

Definition Clarifications

Junction versus CDR3

We work with the IMGT definitions of the junction and CDR3 regions. Specifically, the IMGT JUNCTION includes the conserved cysteine and tryptophan/phenylalanine residues, while CDR3 excludes those two residues. Therefore, our `junction` and `junction_aa` fields which represent the extracted sequence include the two conserved residues, while the coordinate fields (`cdr3_start` and `cdr3_end`) exclude them.

Productive

The schema does not define a strict definition of a productive rearrangement. However, the IMGT definition is recommended:

1. Coding region has an open reading frame
2. No defect in the start codon, splicing sites or regulatory elements.
3. No internal stop codons.
4. An in-frame junction region.

Locus names

A naming convention for locus names is not strictly enforced, but the IMGT locus names are recommended. For example, in the case of human data, this would be the set: IGH, IGK, IGL, TRA, TRB, TRD, or TRG.

Gene and allele names

Gene call examples use the IMGT nomenclature, but no specific gene or allele nomenclature is strictly mandated. Species denotations may or may not be included in the gene name, as appropriate. For example, “Homo sapiens IGHV4-59*01”, “IGHV4-59*01” and “AB019438” are all valid entries for the same allele.

However, when using an established reference database to assign gene calls adherence to the exact nomenclature used by the reference database is strongly recommended, as this will facilitate mapping to the database entries, cross-study comparison, and upload to public repositories.

Alignments

There is no required alignment scheme for the nucleotide and amino acid alignment fields. These fields may, or may not, include numbering spacers (e.g., IMGT-numbering gaps), variations in case to denote mismatches, deletions, or other features appropriate to the tool that performed the alignment. The only strict requirement is that the query (sequence) and reference (germline) **must** be properly aligned.

Frameshifts

For purposes of annotating alignments, a frameshift is defined as a frameshift that is maintained until the end of the aligned gene, where frames are designated numerically as 1 (in-frame), 2, or 3. For example, an V gene alignment that starts in frame 1 and ends in frame 2, disrupting the conserved cystine, would be defined as a frameshift. Whereas, a V gene alignment with an internal frameshift that corrects with a second frameshift, back to the original frame 1 prior to the conserved cystine, would not need to be annotated as a frameshift.

Fields

The specification includes two classes of fields. Those that are required and those that are optional. Required is defined as a column that must be present in the header of the TSV. Optional is defined as column that may, or may not, appear in the TSV. All fields, including required fields, are nullable by assigning an empty string as the value. There are no requirements for column ordering in the schema, although the Python and R reference APIs enforce ordering for the sake of generating predictable output. The set of optional fields that provide alignment and region coordinates (“_start” and “_end” fields) are defined as 1-based closed intervals, similar to the SAM, VCF, GFF, IMGT, and INDSC formats (GenBank, ENA, and DDJB; <http://www.insdc.org>).

Most fields have strict definitions for the values that they contain. However, some commonly provided information cannot be standardized across diverse toolchains, so a small selection of fields have context-dependent definitions. In particular, these context-dependent fields include the optional “_score,” “_identity,” and “_support” fields used for assessing the quality of alignments which vary considerably in definition based on the methodology used. Similarly, the “_alignment” fields require strict alignment between the corresponding observed and germline sequences, but the manner in which that alignment is conveyed is somewhat flexible in that it allows for any numbering scheme (e.g., IMGT or KABAT) or lack thereof.

By default, data elements representing sequences in the schema contain nucleotide sequences except for data elements ending in “_aa,” which are amino acid translations of the associated nucleotide sequence.

While the format contains an extensive list of reserved field names, there are no restrictions on inclusion of custom fields in the TSV file, provided such custom fields have a unique name. Furthermore, suggestions for extending the format with additional reserved names are welcomed through the issue tracker on the GitHub repository (<https://github.com/airr-community/airr-standards>).

Download as TSV

Name	Type	Attributes	Definition
sequence_id	string	required, identifier, nullable	Unique query sequence identifier for the Rearrangement. Most often this will be the input sequence header or a substring thereof, but may also be a custom identifier defined by the tool in cases where query sequences have been combined in some fashion prior to alignment. When downloaded from an AIRR Data Commons repository, this will usually be a universally unique record locator for linking with other objects in the AIRR Data Model.
sequence	string	required, nullable	The query nucleotide sequence. Usually, this is the unmodified input sequence, which may be reverse complemented if necessary. In some cases, this field may contain consensus sequences or other types of collapsed input sequences if these steps are performed prior to alignment.
quality	string	optional, nullable	The Sanger/Phred quality scores for assessment of sequence quality. Phred quality scores from 0 to 93 are encoded using ASCII 33 to 126 (Used by Illumina from v1.8.)
sequence_aa	string	optional, nullable	Amino acid translation of the query nucleotide sequence.
rev_comp	boolean	required, nullable	True if the alignment is on the opposite strand (reverse complemented) with respect to the query sequence. If True then all output data, such as alignment coordinates and sequences, are based on the reverse complement of 'sequence'.
productive	boolean	required, nullable	True if the V(D)J sequence is predicted to be productive.
vj_in_frame	boolean	optional, nullable	True if the V and J gene alignments are in-frame.
stop_codon	boolean	optional, nullable	True if the aligned sequence contains a stop codon.
complete_vdj	boolean	optional, nullable	True if the sequence alignment spans the entire V(D)J region. Meaning, sequence_alignment includes both the first V gene codon that encodes the mature polypeptide chain (i.e., after the leader sequence) and the last complete codon of the J gene (i.e., before the J-C splice site). This does not require an absence of deletions within the internal FWR and CDR regions of the alignment.
locus	string	optional, nullable	Gene locus (chain type). Note that this field uses a controlled vocabulary that is meant to provide a generic classification of the locus, not necessarily the correct designation according to a specific nomenclature.
locus_species	<i>Ontology</i>	optional, nullable	Binomial designation of the species from which the locus originates. Typically, this value should be identical to <i>organism</i> , if which case it SHOULD NOT be set explicitly. However, there are valid experimental setups in which the two might differ, e.g. transgenic animal models. If set, this key will overwrite the <i>organism</i> information for all lower layers of the schema.

continues on next page

Table 2 – continued from previous page

Name	Type	Attributes	Definition
v_call	string	required, nullable	V gene with allele. If referring to a known reference sequence in a database the relevant gene/allele nomenclature should be followed (e.g., IGHV4-59*01 if using IMGT/GENE-DB).
d_call	string	required, nullable	First or only D gene with allele. If referring to a known reference sequence in a database the relevant gene/allele nomenclature should be followed (e.g., IGHD3-10*01 if using IMGT/GENE-DB).
d2_call	string	optional, nullable	Second D gene with allele. If referring to a known reference sequence in a database the relevant gene/allele nomenclature should be followed (e.g., IGHD3-10*01 if using IMGT/GENE-DB).
j_call	string	required, nullable	J gene with allele. If referring to a known reference sequence in a database the relevant gene/allele nomenclature should be followed (e.g., IGHJ4*02 if using IMGT/GENE-DB).
c_call	string	optional, nullable	Constant region gene with allele. If referring to a known reference sequence in a database the relevant gene/allele nomenclature should be followed (e.g., IGHG1*01 if using IMGT/GENE-DB).
sequence_alignmer	string	required, nullable	Aligned portion of query sequence, including any indel corrections or numbering spacers, such as IMGT-gaps. Typically, this will include only the V(D)J region, but that is not a requirement.
quality_alignment	string	optional, nullable	Sanger/Phred quality scores for assessment of sequence_alignment quality. Phred quality scores from 0 to 93 are encoded using ASCII 33 to 126 (Used by Illumina from v1.8.)
sequence_alignmer	string	optional, nullable	Amino acid translation of the aligned query sequence.
germline_alignmer	string	required, nullable	Assembled, aligned, full-length inferred germline sequence spanning the same region as the sequence_alignment field (typically the V(D)J region) and including the same set of corrections and spacers (if any).
germline_alignmer	string	optional, nullable	Amino acid translation of the assembled germline sequence.
junction	string	required, nullable	Junction region nucleotide sequence, where the junction is defined as the CDR3 plus the two flanking conserved codons.
junction_aa	string	required, nullable	Amino acid translation of the junction.
np1	string	optional, nullable	Nucleotide sequence of the combined N/P region between the V gene and first D gene alignment or between the V gene and J gene alignments.
np1_aa	string	optional, nullable	Amino acid translation of the np1 field.
np2	string	optional, nullable	Nucleotide sequence of the combined N/P region between either the first D gene and J gene alignments or the first D gene and second D gene alignments.

continues on next page

Table 2 – continued from previous page

Name	Type	Attributes	Definition
np2_aa	string	optional, nullable	Amino acid translation of the np2 field.
np3	string	optional, nullable	Nucleotide sequence of the combined N/P region between the second D gene and J gene alignments.
np3_aa	string	optional, nullable	Amino acid translation of the np3 field.
cdr1	string	optional, nullable	Nucleotide sequence of the aligned CDR1 region.
cdr1_aa	string	optional, nullable	Amino acid translation of the cdr1 field.
cdr2	string	optional, nullable	Nucleotide sequence of the aligned CDR2 region.
cdr2_aa	string	optional, nullable	Amino acid translation of the cdr2 field.
cdr3	string	optional, nullable	Nucleotide sequence of the aligned CDR3 region.
cdr3_aa	string	optional, nullable	Amino acid translation of the cdr3 field.
fwr1	string	optional, nullable	Nucleotide sequence of the aligned FWR1 region.
fwr1_aa	string	optional, nullable	Amino acid translation of the fwr1 field.
fwr2	string	optional, nullable	Nucleotide sequence of the aligned FWR2 region.
fwr2_aa	string	optional, nullable	Amino acid translation of the fwr2 field.
fwr3	string	optional, nullable	Nucleotide sequence of the aligned FWR3 region.
fwr3_aa	string	optional, nullable	Amino acid translation of the fwr3 field.
fwr4	string	optional, nullable	Nucleotide sequence of the aligned FWR4 region.
fwr4_aa	string	optional, nullable	Amino acid translation of the fwr4 field.
v_score	number	optional, nullable	Alignment score for the V gene.
v_identity	number	optional, nullable	Fractional identity for the V gene alignment.
v_support	number	optional, nullable	V gene alignment E-value, p-value, likelihood, probability or other similar measure of support for the V gene assignment as defined by the alignment tool.
v_cigar	string	required, nullable	CIGAR string for the V gene alignment.
d_score	number	optional, nullable	Alignment score for the first or only D gene alignment.
d_identity	number	optional, nullable	Fractional identity for the first or only D gene alignment.
d_support	number	optional, nullable	D gene alignment E-value, p-value, likelihood, probability or other similar measure of support for the first or only D gene as defined by the alignment tool.

continues on next page

Table 2 – continued from previous page

Name	Type	Attributes	Definition
d_cigar	string	required, nullable	CIGAR string for the first or only D gene alignment.
d2_score	number	optional, nullable	Alignment score for the second D gene alignment.
d2_identity	number	optional, nullable	Fractional identity for the second D gene alignment.
d2_support	number	optional, nullable	D gene alignment E-value, p-value, likelihood, probability or other similar measure of support for the second D gene as defined by the alignment tool.
d2_cigar	string	optional, nullable	CIGAR string for the second D gene alignment.
j_score	number	optional, nullable	Alignment score for the J gene alignment.
j_identity	number	optional, nullable	Fractional identity for the J gene alignment.
j_support	number	optional, nullable	J gene alignment E-value, p-value, likelihood, probability or other similar measure of support for the J gene assignment as defined by the alignment tool.
j_cigar	string	required, nullable	CIGAR string for the J gene alignment.
c_score	number	optional, nullable	Alignment score for the C gene alignment.
c_identity	number	optional, nullable	Fractional identity for the C gene alignment.
c_support	number	optional, nullable	C gene alignment E-value, p-value, likelihood, probability or other similar measure of support for the C gene assignment as defined by the alignment tool.
c_cigar	string	optional, nullable	CIGAR string for the C gene alignment.
v_sequence_start	integer	optional, nullable	Start position of the V gene in the query sequence (1-based closed interval).
v_sequence_end	integer	optional, nullable	End position of the V gene in the query sequence (1-based closed interval).
v_germline_start	integer	optional, nullable	Alignment start position in the V gene reference sequence (1-based closed interval).
v_germline_end	integer	optional, nullable	Alignment end position in the V gene reference sequence (1-based closed interval).
v_alignment_start	integer	optional, nullable	Start position of the V gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
v_alignment_end	integer	optional, nullable	End position of the V gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
d_sequence_start	integer	optional, nullable	Start position of the first or only D gene in the query sequence. (1-based closed interval).
d_sequence_end	integer	optional, nullable	End position of the first or only D gene in the query sequence. (1-based closed interval).
d_germline_start	integer	optional, nullable	Alignment start position in the D gene reference sequence for the first or only D gene (1-based closed interval).

continues on next page

Table 2 – continued from previous page

Name	Type	Attributes	Definition
d_germline_end	integer	optional, nullable	Alignment end position in the D gene reference sequence for the first or only D gene (1-based closed interval).
d_alignment_start	integer	optional, nullable	Start position of the first or only D gene in both the sequence_alignment and germline_alignment fields (1-based closed interval).
d_alignment_end	integer	optional, nullable	End position of the first or only D gene in both the sequence_alignment and germline_alignment fields (1-based closed interval).
d2_sequence_start	integer	optional, nullable	Start position of the second D gene in the query sequence (1-based closed interval).
d2_sequence_end	integer	optional, nullable	End position of the second D gene in the query sequence (1-based closed interval).
d2_germline_start	integer	optional, nullable	Alignment start position in the second D gene reference sequence (1-based closed interval).
d2_germline_end	integer	optional, nullable	Alignment end position in the second D gene reference sequence (1-based closed interval).
d2_alignment_start	integer	optional, nullable	Start position of the second D gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
d2_alignment_end	integer	optional, nullable	End position of the second D gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
j_sequence_start	integer	optional, nullable	Start position of the J gene in the query sequence (1-based closed interval).
j_sequence_end	integer	optional, nullable	End position of the J gene in the query sequence (1-based closed interval).
j_germline_start	integer	optional, nullable	Alignment start position in the J gene reference sequence (1-based closed interval).
j_germline_end	integer	optional, nullable	Alignment end position in the J gene reference sequence (1-based closed interval).
j_alignment_start	integer	optional, nullable	Start position of the J gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
j_alignment_end	integer	optional, nullable	End position of the J gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
c_sequence_start	integer	optional, nullable	Start position of the C gene in the query sequence (1-based closed interval).
c_sequence_end	integer	optional, nullable	End position of the C gene in the query sequence (1-based closed interval).
c_germline_start	integer	optional, nullable	Alignment start position in the C gene reference sequence (1-based closed interval).
c_germline_end	integer	optional, nullable	Alignment end position in the C gene reference sequence (1-based closed interval).
c_alignment_start	integer	optional, nullable	Start position of the C gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
c_alignment_end	integer	optional, nullable	End position of the C gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).

continues on next page

Table 2 – continued from previous page

Name	Type	Attributes	Definition
cdr1_start	integer	optional, nullable	CDR1 start position in the query sequence (1-based closed interval).
cdr1_end	integer	optional, nullable	CDR1 end position in the query sequence (1-based closed interval).
cdr2_start	integer	optional, nullable	CDR2 start position in the query sequence (1-based closed interval).
cdr2_end	integer	optional, nullable	CDR2 end position in the query sequence (1-based closed interval).
cdr3_start	integer	optional, nullable	CDR3 start position in the query sequence (1-based closed interval).
cdr3_end	integer	optional, nullable	CDR3 end position in the query sequence (1-based closed interval).
fwr1_start	integer	optional, nullable	FWR1 start position in the query sequence (1-based closed interval).
fwr1_end	integer	optional, nullable	FWR1 end position in the query sequence (1-based closed interval).
fwr2_start	integer	optional, nullable	FWR2 start position in the query sequence (1-based closed interval).
fwr2_end	integer	optional, nullable	FWR2 end position in the query sequence (1-based closed interval).
fwr3_start	integer	optional, nullable	FWR3 start position in the query sequence (1-based closed interval).
fwr3_end	integer	optional, nullable	FWR3 end position in the query sequence (1-based closed interval).
fwr4_start	integer	optional, nullable	FWR4 start position in the query sequence (1-based closed interval).
fwr4_end	integer	optional, nullable	FWR4 end position in the query sequence (1-based closed interval).
v_sequence_alignm	string	optional, nullable	Aligned portion of query sequence assigned to the V gene, including any indel corrections or numbering spacers.
v_sequence_alignm	string	optional, nullable	Amino acid translation of the v_sequence_alignment field.
d_sequence_alignm	string	optional, nullable	Aligned portion of query sequence assigned to the first or only D gene, including any indel corrections or numbering spacers.
d_sequence_alignm	string	optional, nullable	Amino acid translation of the d_sequence_alignment field.
d2_sequence_align	string	optional, nullable	Aligned portion of query sequence assigned to the second D gene, including any indel corrections or numbering spacers.
d2_sequence_align	string	optional, nullable	Amino acid translation of the d2_sequence_alignment field.
j_sequence_alignm	string	optional, nullable	Aligned portion of query sequence assigned to the J gene, including any indel corrections or numbering spacers.
j_sequence_alignm	string	optional, nullable	Amino acid translation of the j_sequence_alignment field.

continues on next page

Table 2 – continued from previous page

Name	Type	Attributes	Definition
c_sequence_alignm	string	optional, nullable	Aligned portion of query sequence assigned to the constant region, including any indel corrections or numbering spacers.
c_sequence_alignm	string	optional, nullable	Amino acid translation of the c_sequence_alignment field.
v_germline_alignm	string	optional, nullable	Aligned V gene germline sequence spanning the same region as the v_sequence_alignment field and including the same set of corrections and spacers (if any).
v_germline_alignm	string	optional, nullable	Amino acid translation of the v_germline_alignment field.
d_germline_alignm	string	optional, nullable	Aligned D gene germline sequence spanning the same region as the d_sequence_alignment field and including the same set of corrections and spacers (if any).
d_germline_alignm	string	optional, nullable	Amino acid translation of the d_germline_alignment field.
d2_germline_aligr	string	optional, nullable	Aligned D gene germline sequence spanning the same region as the d2_sequence_alignment field and including the same set of corrections and spacers (if any).
d2_germline_aligr	string	optional, nullable	Amino acid translation of the d2_germline_alignment field.
j_germline_alignm	string	optional, nullable	Aligned J gene germline sequence spanning the same region as the j_sequence_alignment field and including the same set of corrections and spacers (if any).
j_germline_alignm	string	optional, nullable	Amino acid translation of the j_germline_alignment field.
c_germline_alignm	string	optional, nullable	Aligned constant region germline sequence spanning the same region as the c_sequence_alignment field and including the same set of corrections and spacers (if any).
c_germline_alignm	string	optional, nullable	Amino acid translation of the c_germline_aligment field.
junction_length	integer	optional, nullable	Number of nucleotides in the junction sequence.
junction_aa_lengt	integer	optional, nullable	Number of amino acids in the junction sequence.
np1_length	integer	optional, nullable	Number of nucleotides between the V gene and first D gene alignments or between the V gene and J gene alignments.
np2_length	integer	optional, nullable	Number of nucleotides between either the first D gene and J gene alignments or the first D gene and second D gene alignments.
np3_length	integer	optional, nullable	Number of nucleotides between the second D gene and J gene alignments.
n1_length	integer	optional, nullable	Number of untemplated nucleotides 5' of the first or only D gene alignment.
n2_length	integer	optional, nullable	Number of untemplated nucleotides 3' of the first or only D gene alignment.
n3_length	integer	optional, nullable	Number of untemplated nucleotides 3' of the second D gene alignment.
p3v_length	integer	optional, nullable	Number of palindromic nucleotides 3' of the V gene alignment.

continues on next page

Table 2 – continued from previous page

Name	Type	Attributes	Definition
p5d_length	integer	optional, nullable	Number of palindromic nucleotides 5' of the first or only D gene alignment.
p3d_length	integer	optional, nullable	Number of palindromic nucleotides 3' of the first or only D gene alignment.
p5d2_length	integer	optional, nullable	Number of palindromic nucleotides 5' of the second D gene alignment.
p3d2_length	integer	optional, nullable	Number of palindromic nucleotides 3' of the second D gene alignment.
p5j_length	integer	optional, nullable	Number of palindromic nucleotides 5' of the J gene alignment.
v_frameshift	boolean	optional, nullable	True if the V gene in the query nucleotide sequence contains a translational frameshift relative to the frame of the V gene reference sequence.
j_frameshift	boolean	optional, nullable	True if the J gene in the query nucleotide sequence contains a translational frameshift relative to the frame of the J gene reference sequence.
d_frame	integer	optional, nullable	Numerical reading frame (1, 2, 3) of the first or only D gene in the query nucleotide sequence, where frame 1 is relative to the first codon of D gene reference sequence.
d2_frame	integer	optional, nullable	Numerical reading frame (1, 2, 3) of the second D gene in the query nucleotide sequence, where frame 1 is relative to the first codon of D gene reference sequence.
consensus_count	integer	optional, nullable	Number of reads contributing to the UMI consensus or contig assembly for this sequence. For example, the sum of the number of reads for all UMIs that contribute to the query sequence.
duplicate_count	integer	optional, nullable	Copy number or number of duplicate observations for the query sequence. For example, the number of identical reads observed for this sequence.
umi_count	integer	optional, nullable	Number of distinct UMIs represented by this sequence. For example, the total number of UMIs that contribute to the contig assembly for the query sequence.
cell_id	string	optional, identifier, nullable	Identifier defining the cell of origin for the query sequence.
clone_id	string	optional, identifier, nullable	Clonal cluster assignment for the query sequence.
repertoire_id	string	optional, identifier, nullable	Identifier to the associated repertoire in study metadata.
sample_processing	string	optional, identifier, nullable	Identifier to the sample processing object in the repertoire metadata for this rearrangement. If the repertoire has a single sample then this field may be empty or missing. If the repertoire has multiple samples then this field may be empty or missing if the sample cannot be differentiated or the relationship is not maintained by the data processing.
data_processing_id	string	optional, identifier, nullable	Identifier to the data processing object in the repertoire metadata for this rearrangement. If this field is empty then the primary data processing object is assumed.

continues on next page

Table 2 – continued from previous page

Name	Type	Attributes	Definition
rearrangement_id	string	DEPRE-CATED	Identifier for the Rearrangement object. May be identical to sequence_id, but will usually be a universally unique record locator for database applications.
rearrangement_set	string	DEPRE-CATED	Identifier for grouping Rearrangement objects.
germline_database	string	DEPRE-CATED	Source of germline V(D)J genes with version number or date accessed.

Alignment Schema (Experimental)

An Alignment is the output from a V(D)J assignment process for a single V, D, J, or C gene for a sequence. It is not necessary that the V(D)J assignment process performs a sequence alignment algorithm, as the schema can support any algorithmic process. Multiple Alignment records are supported and expected for a single sequence with context-dependent fields (*score*, *identity*, *support*, *rank*) for assessing the quality of assignments that can vary considerably in definition based on the methodology used.

Note, this schema definition is still experimental and should not be considered final.

File Format Specification

The *format specification* describes the file format and details on how to structure this data.

Fields

Download as TSV

Name	Type	Attributes	Definition
sequence_id	string	required, identifier, nullable	Unique query sequence identifier within the file. Most often this will be the input sequence header or a substring thereof, but may also be a custom identifier defined by the tool in cases where query sequences have been combined in some fashion prior to alignment.
segment	string	required, nullable	The segment for this alignment. One of V, D, J or C.
rev_comp	boolean	optional, nullable	Alignment result is from the reverse complement of the query sequence.
call	string	required, nullable	Gene assignment with allele.
score	number	required, nullable	Alignment score.
identity	number	optional, nullable	Alignment fractional identity.
support	number	optional, nullable	Alignment E-value, p-value, likelihood, probability or other similar measure of support for the gene assignment as defined by the alignment tool.
cigar	string	required, nullable	Alignment CIGAR string.
sequence_start	integer	optional, nullable	Start position of the segment in the query sequence (1-based closed interval).
sequence_end	integer	optional, nullable	End position of the segment in the query sequence (1-based closed interval).
germline_start	integer	optional, nullable	Alignment start position in the reference sequence (1-based closed interval).
germline_end	integer	optional, nullable	Alignment end position in the reference sequence (1-based closed interval).
rank	integer	optional, nullable	Alignment rank.
rearrangement_id	string	DEPRECATED	Identifier for the Rearrangement object. May be identical to sequence_id, but will usually be a universally unique record locator for database applications.
data_processing_i	string	optional, nullable	Identifier to the data processing object in the repertoire metadata for this rearrangement. If this field is empty than the primary data processing object is assumed.
germline_database	string	DEPRECATED	Source of germline V(D)J genes with version number or date accessed.

Clone and Lineage Tree Schema (Experimental)

A unique inferred clone object that has been constructed within a single data processing for a single repertoire and a subset of its sequences and/or rearrangements.

A clone object may have one or more inferred lineage trees. Each tree is represented by a Newick string for its edges and a dictionary of node objects.

File Format Specification

The file format has not been specified yet.

Clone Fields

Download as TSV

Name	Type	Attributes	Definition
clone_id	string	required, identifier, nullable	Identifier for the clone.
repertoire_id	string	optional, nullable	Identifier to the associated repertoire in study metadata.
data_processing_i	string	optional, nullable	Identifier of the data processing object in the repertoire metadata for this clone.
sequences	array of string	optional, nullable	List sequence_id strings that act as keys to the Rearrangement records for members of the clone.
v_call	string	optional, nullable	V gene with allele of the inferred ancestral of the clone. For example, IGHV4-59*01.
d_call	string	optional, nullable	D gene with allele of the inferred ancestor of the clone. For example, IGHD3-10*01.
j_call	string	optional, nullable	J gene with allele of the inferred ancestor of the clone. For example, IGHJ4*02.
junction	string	optional, nullable	Nucleotide sequence for the junction region of the inferred ancestor of the clone, where the junction is defined as the CDR3 plus the two flanking conserved codons.
junction_aa	string	optional, nullable	Amino acid translation of the junction.
junction_length	integer	optional, nullable	Number of nucleotides in the junction.
junction_aa_lengt	integer	optional, nullable	Number of amino acids in junction_aa.
germline_alignmer	string	required, nullable	Assembled, aligned, full-length inferred ancestor of the clone spanning the same region as the sequence_alignment field of nodes (typically the V(D)J region) and including the same set of corrections and spacers (if any).
germline_alignmer	string	optional, nullable	Amino acid translation of germline_alignment.
v_alignment_start	integer	optional, nullable	Start position in the V gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
v_alignment_end	integer	optional, nullable	End position in the V gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
d_alignment_start	integer	optional, nullable	Start position of the D gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
d_alignment_end	integer	optional, nullable	End position of the D gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
j_alignment_start	integer	optional, nullable	Start position of the J gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
j_alignment_end	integer	optional, nullable	End position of the J gene alignment in both the sequence_alignment and germline_alignment fields (1-based closed interval).
junction_start	integer	optional, nullable	Junction region start position in the alignment (1-based closed interval).
junction_end	integer	optional, nullable	Junction region end position in the alignment (1-based closed interval).
umi_count	integer	optional, nullable	Number of distinct UMIs observed across all sequences (Rearrangement records) in this clone.
clone_count	integer	optional, nullable	Absolute count of the size (number of members) of this clone in the repertoire. This could simply be the number of sequences (Rearrangement records) observed in this clone, the number of distinct cell barcodes (unique

Tree Fields

Download as TSV

Name	Type	Attributes	Definition
tree_id	string	required, identifier, nullable	Identifier for the tree.
clone_id	string	required, nullable	Identifier for the clone.
newick	string	required, nullable	Newick string of the tree edges.
nodes	object	optional, nullable	Dictionary of nodes in the tree, keyed by sequence_id string

Node Fields

Download as TSV

Name	Type	Attributes	Definition
sequence_id	string	required, identifier, nullable	Identifier for this node that matches the identifier in the newick string and, where possible, the sequence_id in the source repertoire.
sequence_alignmer	string	optional, nullable	Nucleotide sequence of the node, aligned to the germline_alignment for this clone, including including any indel corrections or spacers.
junction	string	optional, nullable	Junction region nucleotide sequence for the node, where the junction is defined as the CDR3 plus the two flanking conserved codons.
junction_aa	string	optional, nullable	Amino acid translation of the junction.

Cell Schema (Experimental)

The cell object acts as point of reference for all data that can be related to an individual cell, either by direct observation or inference.

File Format Specification

The file format has not been specified yet.

Cell Fields

[Download as TSV](#)

Name	Type	Attributes	Definition
cell_id	string	required, identifier	Identifier defining the cell of origin for the query sequence.
rearrangements	array of string	required, nullable	Array of sequence identifiers defined for the Rearrangement object
receptors	array of string	optional, nullable	Array of receptor identifiers defined for the Receptor object
repertoire_id	string	required, nullable	Identifier to the associated repertoire in study metadata.
data_processing_i	string	optional, nullable	Identifier of the data processing object in the repertoire metadata for this clone.
expression_study_	string	optional, nullable	Keyword describing the methodology used to assess expression. This values for this field MUST come from a controlled vocabulary.
expression_raw_dc	string	optional, nullable	DOI of raw data set containing the current event
expression_index	string	optional, nullable	Index addressing the current event within the raw data set.
virtual_pairing	boolean	required, nullable	boolean to indicate if pairing was inferred.

Cell Expression Schema (Experimental)

File Format Specification

The file format has not been specified yet.

Cell Expression Fields

[Download as TSV](#)

Name	Type	Attributes	Definition
expression_id	string	required, identifier	Identifier of this expression property measurement.
cell_id	string	required	Identifier of the cell to which this expression data is related.
repertoire_id	string	required, nullable	Identifier for the associated repertoire in study metadata.
data_processing_id	string	required, nullable	Identifier of the data processing object in the repertoire metadata for this clone.
property_type	string	required	Keyword describing the property type and detection method used to measure the property value. The following keywords are recommended, but custom property types are also valid: “mrna_expression_by_read_count”, “protein_expression_by_fluorescence_intensity”, “antigen_bait_binding_by_fluorescence_intensity”, “protein_expression_by_dna_barcode_count” and “antigen_bait_binding_by_dna_barcode_count”.
property	<i>Ontology</i>	required, nullable	Name of the property observed, typically a gene or antibody identifier (and label) from a canonical resource such as Ensembl (e.g. ENSG00000275747, IGHV3-79) or Antibody Registry (ABREG:1236456, Purified anti-mouse/rat/human CD27 antibody).
value	number	required, nullable	Level at which the property was observed in the experiment (non-normalized).

Germline Schema (Experimental)

Motivation

Understanding and cataloguing receptor germline genes and allele sequences is critical to the analysis of AIRR data. While the human set is relatively well understood in outline, although probably still far from complete, those of other species, even those that are relatively closely studied, is at a much earlier stage. There is an urgent need to define a standardised format for listing such genes, so that they can be shared between researchers and easily consumed by software tools.

Receptor Germline Schema

The receptor germline schema defines the data elements necessary to describe one or more receptor germline genes, together with supporting evidence. The fundamental object is the `AlleleDescription`, which describes a single gene or allele, containing the necessary details for the annotation of a rearranged sequence such as the location of CDRs (in the case of a V-gene) and framing information (in the case of a J-gene). `AlleleDescription` also contains fields to delineate RSS, and the leader regions of V-genes, should those be covered by the sequence provided.

Evidence supporting the gene or allele can be provided in linked `UnrearrangedSequence` and `RearrangedSequence` objects. Information represented in these objects will typically be stored in a repository: either an INSDC repository such as Genbank or SRA, or a lower-tier repository such as OGRDB. Please note that the key distinction between these object types is whether the V(D)J genes have rearranged, rather than the origin of the material, as mature B and T cells carry rearranged sequences in chromosomal DNA. It is most likely that supporting sequences will be `UnrearrangedSequences`, i.e. prior to rearrangement. In the case of a germline inference from a repertoire, the

inferred germline sequence should be provided as a `RearrangedSequence`, if the evidence has been deposited in a repository.

For V-genes, an IMGT-gapped sequence (i.e., a sequence delineated in accordance with the [IMGT numbering scheme](#)) is provided in `AlleleDescription`. Other delineations, such as [Chothia](#) and [Kabat](#), can be provided via linked `SequenceDelineationV` objects. A `GermlineSet` brings together multiple `AlleleDescriptions` from the same locus to form a curated set. The schema assumes that germline sets will be published by multiple repositories. A germline set may be uniquely referenced by means of the `germline_set_ref`, which is a composite field containing the repository id, germline set label, and version.

Gene and Allele Naming

`AlleleDescription` contains a `label` field, which should contain the accepted name for the field, as determined by the authors/curators of the record. The [Nomenclature Committee](#) of the International Union of Immunological Societies (IUIS) allocates gene symbols for receptor genes, and, if a gene symbol has been allocated, this should be used as the label. Where a gene symbol has not been allocated (for example, because the gene or allele has only recently been discovered, or because the available evidence does not meet IUIS standards, a ‘temporary label’ should be used. It is anticipated that publishers of gene sets will provide mechanisms to issue these temporary labels, and to allow researchers to review change history of `AlleleDescriptions` and `GermlineSets`. To provide consistency across research groups, the [Germline Database Working Group of the AIRR Community](#) is developing a [community-wide approach](#) to the allocation of temporary labels.

Genotypes

A `GenotypeSet` describes the specific receptor alleles found in a subject, and also identifies genes that are not found (this could be either because they are not present in the chromosomal locus, or because they are not expressed or expressed only at low levels). Depending on the data available and the inference method used, genotypes may contain haplotyping information, which may be full, or partial. As an example of partial haplotyping, the genotype may have been determined from genomic sequencing in which the sequence of the locus was assembled into contigs, but could not be fully assembled. In this case the co-location of alleles in each contig has been established, but the co-location across the entire locus can not be. Co-location is therefore indicated by means of the `phasing` parameter, which in this case would be assigned a different value for alleles on each contig.

MHC Genotypes

Similar to the IG/TR genotypes, the `MHCGenotype` and `MHCGenotypeSet` objects describe the MHC alleles found in a subject. `MHCGenotype` objects assemble alleles from one class: `MHC-I`, `MHC-II` or `MHC-nonclassical`. The method used to determine the genotype can be provided in the `mhc_genotyping_method` field. As different methods might be used for the various classes, this field is located in the `MHCGenotype` object, not the `MHCGenotypeSet`.

The `mhc_genotyping_method` allows free-text descriptions, however data curators are asked to keep close to the following terms if applicable:

- **PCR-based typing:** Methods whose read-out is the amplification of specific sequences, but which do not provide sequence data by themselves. This includes SSP and SSOP.
- **Sequencing-based typing:** Clinical-grade NGS-based assays, providing high quality and resolution.
- **Inference-based typing:** Allele inference based on genome-wide DNA or RNA sequencing.

File Format Specification

Files are YAML/JSON with a structure defined below. Files should be encoded as UTF-8. Identifiers are case-sensitive. Files should have the extension `.yaml`, `.yml`, or `.json`.

Germline Set File Structure

The Germline Set file has a standardised structure that is utilized by all top-level AIRR Schema Objects and defined by the `DataFile` schema. It is intended to contain all information necessary to annotate receptor sequences derived from a single germline locus, and to be directly usable by annotation tools and other processing software.

The file must contain YAML or JSON representation of one or more `GermlineSet` objects, including the associated `AlleleDescription` objects. It may optionally include other associated objects: `SequenceDelineationV`, `RearrangedSequence`, `UnrearrangedSequence`, `Acknowledgement`. These should all be embedded into the overall `GermlineSet` as specified in the schema.

- The file as a whole is considered a dictionary (key/value pair) structure with the keys `Info`, `GermlineSet`, and `AlleleDescription`.
- The `GermlineSet` contains fields `release_version`, `release_description` and `release_date`, which are intended to be used for version identification, under the control of the authors of the `GermlineSet` as identified by the fields `author`, `lab_name` and `lab_address`. If the set is modified by a party other than these authors, that these 6 fields should be modified to reflect the authors of the modification, and their own version identification. These modifications **MUST** be made if the `GermlineSet` is, or is likely to become, public, in order to avoid confusion with the original set prior to modification. Repositories are encouraged to manage version fields automatically.
- The file can (optionally) contain an `Info` object, at the beginning of the file, based upon the `Info` schema in the OpenAPI specification. If provided, `version` in `Info` should reference the version of the AIRR schema for the file.
- The file should correspond to a list of `GermlineSet` objects, using `GermlineSet` as the key to the list.
- The file should correspond to a list of `AlleleDescription` objects, using `AlleleDescription` as the key to the list.
- There should be only one `AlleleDescription` for each allele in the list.
- Each `AlleleDescription` object should contain a top-level key/value pair for `allele_description_id` that uniquely identifies the allele description object in the file.
- Each `GermlineSet` object should contain a top-level key/value pair for `germline_set_id` that uniquely identifies the germline set object in the file.
- Some fields require the use of a particular ontology or controlled vocabulary.
- `GermlineSet` and `AlleleDescription` contain reference fields `germline_set_ref` and `allele_description_ref`. These are intended to be globally unique references (containing identifiers of the repository, object and version) that can be used in a query API.
- The structure is the same regardless of whether the data is stored in a file or retrieved from a data repository. For example, The [ADC API](#) will return a properly structured JSON object that can be saved to a file and used directly without modification.

GermlineSet Fields

Download as TSV

Name	Type	Attributes	Definition
germline_set_id	string	required, identifier, nullable	Unique identifier of the GermlineSet within this file. Typically, generated by the repository hosting the record.
author	string	required, nullable	Corresponding author
lab_name	string	required, nullable	Department of corresponding author
lab_address	string	required, nullable	Institutional address of corresponding author
acknowledgements	array of Acknowledgement	optional, nullable	List of individuals whose contribution to the germline set should be acknowledged
release_version	number	required, nullable	Version number of this record, allocated automatically
release_descripti	string	required, nullable	Brief descriptive notes of the reason for this release and the changes embodied
release_date	string	required, nullable	Date of this release
germline_set_name	string	required, nullable	descriptive name of this germline set
germline_set_ref	string	required, nullable	Unique identifier of the germline set and version, in standardized form (Repo:Label:Version)
pub_ids	string	optional, nullable	Publications describing the germline set
species	<i>Ontology</i>	required	Binomial designation of subject's species
species_subgroup	string	optional, nullable	Race, strain or other species subgroup to which this subject belongs
species_subgroup_	string	optional, nullable	
locus	string	required	Gene locus
allele_descripti	array of <i>AlleleDescription</i>	required, nullable	list of allele_descriptions in the germline set
curation	string	optional, nullable	Curational notes on the GermlineSet. This can be used to give more extensive notes on the decisions taken than are provided in the release_description.

AlleleDescription Fields

Download as TSV

Name	Type	Attributes	Definition
allele_descripti	string	required, identifier, nullable	Unique identifier of this AlleleDescription within the file. Typically, generated by the repository hosting the record.

continues on next page

Table 3 – continued from previous page

Name	Type	Attributes	Definition
allele_descriptic	string	optional, nullable	Unique reference to the allele description, in standardized form (Repo:Label:Version)
maintainer	string	required, nullable	Maintainer of this sequence record
acknowledgements	array of Acknowledgement	optional, nullable	List of individuals whose contribution to the gene description should be acknowledged
lab_address	string	required, nullable	Institution and full address of corresponding author
release_version	integer	required, nullable	Version number of this record, updated whenever a revised version is published or released
release_date	string	required, nullable	Date of this release
release_descripti	string	required, nullable	Brief descriptive notes of the reason for this release and the changes embodied
label	string	optional, nullable	The accepted name for this gene or allele following the relevant nomenclature. The value in this field should correspond to values in acceptable name fields of other schemas, such as v_call, d_call, and j_call fields.
sequence	string	required	Nucleotide sequence of the gene. This should cover the full length that is available, including where possible RSS, and 5' UTR and lead-in for V-gene sequences.
coding_sequence	string	required, nullable	Nucleotide sequence of the core coding region, such as the coding region of a D-, J- or C- gene or the coding region of a V-gene excluding the leader.
aliases	array of string	optional, nullable	Alternative names for this sequence
locus	string	required	Gene locus
chromosome	integer	optional, nullable	chromosome on which the gene is located
sequence_type	string	required	Sequence type (V, D, J, C)
functional	boolean	required, nullable	True if the gene is functional, false if it is a pseudogene
inference_type	string	required, nullable	Type of inference(s) from which this gene sequence was inferred
species	<i>Ontology</i>	required	Binomial designation of subject's species
species_subgroup	string	optional, nullable	Race, strain or other species subgroup to which this subject belongs
species_subgroup_	string	optional, nullable	
status	string	optional, nullable	Status of record, assumed active if the field is not present
subgroup_designat	string	optional, nullable	Identifier of the gene subgroup or clade, as (and if) defined
gene_designation	string	optional, nullable	Gene number or other identifier, as (and if) defined
allele_designatic	string	optional, nullable	Allele number or other identifier, as (and if) defined
allele_similarity	string	optional, nullable	ID of the similarity cluster used in this germline set, if designated

continues on next page

Table 3 – continued from previous page

Name	Type	Attributes	Definition
allele_similarity	string	optional, nullable	Membership ID of the allele within the similarity cluster, if a cluster is designated
j_codon_frame	integer	optional, nullable	Codon position of the first nucleotide in the ‘coding_sequence’ field. Mandatory for J genes. Not used for V or D genes. ‘1’ means the sequence is in-frame, ‘2’ means that the first bp is missing from the first codon, and ‘3’ means that the first 2 bp are missing.
gene_start	integer	optional, nullable	Co-ordinate in the sequence field of the first nucleotide in the coding_sequence field.
gene_end	integer	optional, nullable	Co-ordinate in the sequence field of the last gene-coding nucleotide in the coding_sequence field.
utr_5_prime_start	integer	optional, nullable	Start co-ordinate in the sequence field of the 5 prime UTR (V-genes only).
utr_5_prime_end	integer	optional, nullable	End co-ordinate in the sequence field of the 5 prime UTR (V-genes only).
leader_1_start	integer	optional, nullable	Start co-ordinate in the sequence field of L-PART1 (V-genes only).
leader_1_end	integer	optional, nullable	End co-ordinate in the sequence field of L-PART1 (V-genes only).
leader_2_start	integer	optional, nullable	Start co-ordinate in the sequence field of L-PART2 (V-genes only).
leader_2_end	integer	optional, nullable	End co-ordinate in the sequence field of L-PART2 (V-genes only).
v_rs_start	integer	optional, nullable	Start co-ordinate in the sequence field of the V recombination site (V-genes only).
v_rs_end	integer	optional, nullable	End co-ordinate in the sequence field of the V recombination site (V-genes only).
d_rs_3_prime_star	integer	optional, nullable	Start co-ordinate in the sequence field of the 3 prime D recombination site (D-genes only).
d_rs_3_prime_end	integer	optional, nullable	End co-ordinate in the sequence field of the 3 prime D recombination site (D-genes only).
d_rs_5_prime_star	integer	optional, nullable	Start co-ordinate in the sequence field of the 5 prime D recombination site (D-genes only).
d_rs_5_prime_end	integer	optional, nullable	End co-ordinate in the sequence field of 5 the prime D recombination site (D-genes only).
j_cdr3_end	integer	optional, nullable	In the case of a J-gene, the co-ordinate in the sequence field of the first nucleotide of the conserved PHE or TRP (IMGT codon position 118).
j_rs_start	integer	optional, nullable	Start co-ordinate in the sequence field of J recombination site (J-genes only).
j_rs_end	integer	optional, nullable	End co-ordinate in the sequence field of J recombination site (J-genes only).
j_donor_splice	integer	optional, nullable	Co-ordinate in the sequence field of the final 3’ nucleotide of the J-REGION (J-genes only).
v_gene_delineatic	array of <i>SequenceDelineationV</i>	optional, nullable	
unrearranged_suppr	array of <i>UnrearrangedSequence</i>	optional, nullable	

continues on next page

Table 3 – continued from previous page

Name	Type	Attributes	Definition
rearranged_support	array of <i>Rearranged-Sequence</i>	optional, nullable	
paralogs	array of string	optional, nullable	Gene symbols of any paralogs
curation	string	optional, nullable	Curational notes on the AlleleDescription. This can be used to give more extensive notes on the decisions taken than are provided in the release_description.
curational_tags	array of string	optional, nullable	Controlled-vocabulary tags applied to this description

RearrangedSequence Fields

Download as TSV

Name	Type	Attributes	Definition
sequence_id	string	required, identifier, nullable	Unique identifier of this RearrangedSequence within the file, typically generated by the repository hosting the schema, for example from the underlying ID of the database record.
sequence	string	required	nucleotide sequence
derivation	string	required, nullable	The class of nucleic acid that was used as primary starting material
observation_type	string	required	The type of observation from which this sequence was drawn, such as direct sequencing or inference from repertoire sequencing data.
curation	string	optional, nullable	Curational notes on the sequence
repository_name	string	required, nullable	Name of the repository in which the sequence has been deposited
repository_ref	string	optional, nullable	Queryable id or accession number of the sequence published by the repository
deposited_version	string	required, nullable	Version number of the sequence within the repository
sequence_start	integer	optional	Start co-ordinate of the sequence detailed in this record, within the sequence deposited
sequence_end	integer	optional	End co-ordinate of the sequence detailed in this record, within the sequence deposited

UnrearrangedSequence Fields

Download as TSV

Name	Type	Attributes	Definition
sequence_id	string	required, identifier, nullable	unique identifier of this UnrearrangedSequence within the file
sequence	string	required	Sequence of interest described in this record. Typically, this will include gene and promoter region.
curation	string	optional, nullable	Curational notes on the sequence
repository_name	string	required, nullable	Name of the repository in which the assembly or contig is deposited
repository_ref	string	optional, nullable	Queryable id or accession number of the sequence published by the repository
patch_no	string	optional, nullable	Genome assembly patch number in which this gene was determined
gff_seqid	string	required, nullable	Sequence (from the assembly) of a window including the gene and preferably also the promoter region.
gff_start	integer	required, nullable	Genomic co-ordinates of the start of the sequence of interest described in this record in Ensemble GFF version 3.
gff_end	integer	required, nullable	Genomic co-ordinates of the end of the sequence of interest described in this record in Ensemble GFF version 3.
strand	string	required, nullable	sense (+ or -)

SequenceDelineationV Fields

Download as TSV

Name	Type	Attributes	Definition
sequence_delineat	string	required, identifier, nullable	Unique identifier of this SequenceDelineationV within the file. Typically, generated by the repository hosting the record.
delineation_schem	string	required, nullable	Name of the delineation scheme
unaligned_sequenc	string	optional, nullable	entire V-sequence covered by this delineation
aligned_sequence	string	optional, nullable	Aligned sequence if this delineation provides an alignment. An aligned sequence should always be provided for IMGT delineations.
fwr1_start	integer	required, nullable	FWR1 start co-ordinate in the ‘unaligned sequence’ field
fwr1_end	integer	required, nullable	FWR1 end co-ordinate in the ‘unaligned sequence’ field
cdr1_start	integer	required, nullable	CDR1 start co-ordinate in the ‘unaligned sequence’ field
cdr1_end	integer	required, nullable	CDR1 end co-ordinate in the ‘unaligned sequence’ field
fwr2_start	integer	required, nullable	FWR2 start co-ordinate in the ‘unaligned sequence’ field
fwr2_end	integer	required, nullable	FWR2 end co-ordinate in the ‘unaligned sequence’ field
cdr2_start	integer	required, nullable	CDR2 start co-ordinate in the ‘unaligned sequence’ field
cdr2_end	integer	required, nullable	CDR2 end co-ordinate in the ‘unaligned sequence’ field
fwr3_start	integer	required, nullable	FWR3 start co-ordinate in the ‘unaligned sequence’ field
fwr3_end	integer	required, nullable	FWR3 end co-ordinate in the ‘unaligned sequence’ field
cdr3_start	integer	required, nullable	CDR3 start co-ordinate in the ‘unaligned sequence’ field
alignment_labels	array of string	optional, nullable	One string for each codon in the aligned_sequence indicating the label of that codon according to the numbering of the delineation scheme if it provides one.

GenotypeSet Fields

Download as TSV

Name	Type	Attributes	Definition
receptor_genotype	string	required, identifier, nullable	A unique identifier for this Receptor Genotype Set, typically generated by the repository hosting the schema, for example from the underlying ID of the database record.
genotype_class_li	array of <i>Genotype</i>	optional, nullable	List of Genotypes included in this Receptor Genotype Set.

Genotype Fields

Download as TSV

Name	Type	Attributes	Definition
receptor_genotype	string	required, identifier, nullable	A unique identifier within the file for this Receptor Genotype, typically generated by the repository hosting the schema, for example from the underlying ID of the database record.
locus	string	required	Gene locus
documented_allele	array of DocumentedAllele	optional, nullable	List of alleles documented in reference set(s)
undocumented_allele	array of UndocumentedAllele	optional, nullable	List of alleles inferred to be present and not documented in an identified GermlineSet
deleted_genes	array of DeletedGene	optional, nullable	Array of genes identified as being deleted in this genotype
inference_process	string	optional, nullable	Information on how the genotype was acquired. Controlled vocabulary.

MHCGenotypeSet Fields

Download as TSV

Name	Type	Attributes	Definition
mhc_genotype_set_	string	required, identifier, nullable	A unique identifier for this MHCGenotypeSet
mhc_genotype_list	array of <i>MHCGenotype</i>	required, nullable	List of MHCGenotypes included in this set

MHCGenotype Fields

Download as TSV

Name	Type	Attributes	Definition
mhc_genotype_id	string	required, identifier, nullable	A unique identifier for this MHCGenotype, assumed to be unique in the context of the study
mhc_class	string	required	Class of MHC alleles described by the MHCGenotype
mhc_alleles	array of MH-CAllele	required, nullable	List of MHC alleles of the indicated mhc_class identified in an individual
mhc_genotyping_method	string	optional, nullable	Information on how the genotype was determined. The content of this field should come from a list of recommended terms provided in the AIRR Schema documentation.

Receptor Schema (Experimental)

General

The purpose of the `Receptor` object is to represent reactivity data of actual *Receptors*, i.e., Ig or TCR, either directly or indirectly. To this end, the `Receptor` object has two main sections:

1. The `receptor_*` properties: These properties describe the receptor as an abstract and global concept, i.e., the actual Ig/TCR protein complex MAY or MAY NOT have been observed in the current study. However, the Rearrangements encoding the respective chains MUST be present in the studies as well as the information linking them (see below). This allows data curators to provide potential reactivities of a potential receptor by referencing to its entry in an external database (e.g., IEDB).
2. The `reactivity_measurements` array: This structure contains one or more entries, describing reactivities of the receptor that have been observed in the current study.

The `Receptor` object explicitly requires full sequence information of the two associated variable domains. This is considered to be an acceptable restriction from an AIRR-seq perspective, where sequencing typically precedes or takes place in combination with the determination of receptor reactivity.

Identifiers

The `Receptor` objects has two properties that serve as identifiers:

- `receptor_id` is a **local** identifier and its uniqueness MUST NOT be assumed beyond the scope of the study the receptor was reported in. This property can be used, e.g., to represent designations for Ig/TCR used in a manuscript.
- `receptor_hash` is the SHA256 hash of the receptors variable domain amino acid sequences, which serves as a **globally unique** identifier that can be independently calculated by repositories without requiring prior communication. It is calculated as follows, where `base16` designates the function described in [RFC4648 Section 6](#):

```
lower_case(
  base16(
    sha256(
      concatenate(
        upper_case(receptor_variable_domain_1_aa),
        upper_case(receptor_variable_domain_2_aa)
      )
    )
  )
)
```

Relations to other AIRR Schema objects

The `Receptor` object is only directly linked to the `Cell` object, which then in turn contains the references to the records in the `Rearrangements` that encode the respective chains of the receptor. Therefore a given rearrangement cannot directly reference to a receptor, which is also not a meaningful thing to do, as the paired chain would be unclear, but is necessary to determine a receptors reactivity.

Annotation guidelines

Reactivity information **SHOULD** be provided if it was either directly recorded in the current study or is indirectly available in a public database, where it can be linked. If no information about the receptor, besides the translated amino acid sequence of the associated chains is available, the respective object **MAY** still be created if the curator considers this information to be beneficial, e.g., if there is some evidence (surface expression, reaction to superantigens) that the receptor is functional and present on the surface. In the absence of any such data, **Receptor** objects **SHOULD NOT** be created.

Note on cells expressing more than a single receptor

Cells that express more than a single IGH/TRB/TRD or a single IGK/IGL/TRA/TRG chain are regularly observed as allelic exclusion is never complete and its efficiency is rather low for loci like TRA. Such dual-expressing cells can technically be accommodated in the current AIRR Schema as an individual **Cell** object can link to more than two rearrangements and to more than a single **Receptor**. In the case of two potential receptors, both **MAY** be created as objects, if the general annotation rules are met (i.e., the direct or indirect presence of reactivity information) for each of them. Note that the annotation of **cell**-based reactivity information is currently not supported for dual-expressing cells, in this case additional information confirming the reactivity of one of the receptors would be required.

Representation of bi-specific antibodies

The goal of the AIRR Standards is primarily to represent naturally occurring receptors. While bi-specific antibodies may arise in dual IGK/IGL expressing B cells their individual reactivity is not measured on a regular basis. Therefore they are currently not supported in the **Receptor** schema.

Receptor Fields

Download as TSV

Name	Type	Attributes	Definition
receptor_id	string	required, identifier	ID of the current Receptor object, unique within the local repository.
receptor_hash	string	required	The SHA256 hash of the receptor amino acid sequence, calculated on the concatenated <code>receptor_variable_domain*_aa</code> sequences and represented as base16-encoded string.
receptor_type	string	required	The top-level receptor type, either Immunoglobulin (Ig) or T Cell Receptor (TCR).
receptor_variable	string	required	Complete amino acid sequence of the mature variable domain of the Ig heavy, TCR beta or TCR delta chain. The mature variable domain is defined as encompassing all AA from and including first AA after the the signal peptide to and including the last AA that is completely encoded by the J gene.
receptor_variable	string	required	Locus from which the variable domain in <code>receptor_variable_domain_1_aa</code> originates
receptor_variable	string	required	Complete amino acid sequence of the mature variable domain of the Ig light, TCR alpha or TCR gamma chain. The mature variable domain is defined as encompassing all AA from and including first AA after the the signal peptide to and including the last AA that is completely encoded by the J gene.
receptor_variable	string	required	Locus from which the variable domain in <code>receptor_variable_domain_2_aa</code> originates
receptor_ref	array of string	optional, nullable	Array of receptor identifiers defined for the Receptor object
reactivity_measur	array of ReceptorReactivity	optional, nullable	Records of reactivity measurement

2.3.3 AIRR Software WG - Guidance for AIRR Software Tools

Version 1.0

AIRR Software WG - Compliance Checklist for AIRR Software Tools

Version 1.0 (when finalised)

This questionnaire should be read in conjunction with the *AIRR Software WG - Guidance for AIRR Software Tools*.

To submit your tool for ratification against the standard, please send the completed questionnaire to software@airrc.antibodysociety.org.

Please provide comments in italics in each response box where these would be helpful to facilitate understanding. We kindly ask for a brief explanatory comment if your answer to a question is *no* or *not applicable*.

Name of Tool:

Brief description of Tool:

Contact Name/Institution:

Contact email:

Re-requirement Ref.	Question	Response
1	Where is the source code published (please provide a link)?	
2	Does the tool support <i>AIRR Data Representations</i> standards? Please list any other standard data formats that are supported	yes/no
3	Does the distribution include example data? Is the example data in MiAIRR format, where applicable? Does the tool provide automated checks for expected output from example data?	yes/no yes/no/not applicable yes/no
4	Does the output of the tool include a summary of the run parameters?	yes/no
5	Is a container build file provided? Container technology used? Is the container automatically built as new versions are released? Does the automated build run the tool against the example data and test the output?	yes/no Docker/Singularity/Other (please specify) yes/no yes/no
6	Where can users see what level of support is available? (Please provide a link)	
7	Under what software licence is the tool published? (please provide the name of the licence (e.g. GPL, MIT) or a link	

AIRR Software WG - List of Tools Certified as Compliant

The following tools have been certified as compliant with v1.0 of the guidelines:

Software	Version	Support	Reference
SONAR	3	Output	Schramm et al. Front Immunol, 2016.
ImmuneDB	0.29.10	Input	Rosenfeld et al. Front Immunol, 2018,
Scirpy	0.4.2	Input	Sturm et al. Bioinformatics, 2020,
Immcantation Framework	4	Input, Output	Vander Heiden et al. Bioinformatics, 2014,; Gupta et al., Bioinformatics, 2015,
CompAIRR	1.3.1	Input	Rognes T, Scheffer L, Greiff V, Sandve GK, BioRxiv, 2021 (preprint),
ImmuneML	2.0.4	Input, Output	Pavlovic et al., BioRxiv, 2021 (preprint),
Dandelion	0.1.10	Output	Suo, C., Polanski, K., Dann, E. et al., Nature Biotechnology, 2023,
TRUST4	1.0.8	Output	Song, L., Cohen, D., Ouyang, Z. et al., Nature Methods, 2021,
partis	0.16.0	Input, Output	Ralph D and Matsen F IV, PLOS Computational Biology, 2022,

Evaluation Data Sets

The Software WG is working on the development and evaluation of simulated data sets.

Lists of published real-world datasets are maintained in the [AIRR Forum Wiki](#).

Introduction

The [Adaptive Immune Receptor Repertoire \(AIRR\) Community](#) will benefit greatly from cooperation among groups developing software tools and resources for AIRR research. The goal of the [AIRR Software Working Group](#) is to promote standards for AIRR software tools and resources in order to enable rigorous and reproducible immune repertoire research at the largest scale possible. As one contribution to this goal, we have established the following standards for software tools. Authors whose tools comply with this standard will, subject to ratification from the AIRR Software WG, be permitted to advertise their tools as being AIRR-compliant.

Requirements

Tools must:

1. Be published in source code form, and hosted on a publicly available repository with a clear versioning system.
2. Support community-curated standard file formats and strive for modularity and interoperability with other tools. In particular, tools must read and write [AIRR Data Representations](#) standards corresponding to their tool.
3. Include example data (in AIRR standard formats where applicable) and an automated check for expected output from that data, in order to provide a minimal example of functionality allowing users to check that the software is performing as described.
4. Provide information about run parameters as part of the output.
5. Provide a container build file that can be used to create an image which encapsulates the software tool, its dependencies, and required run environment. This needs to be remotely and automatically built. The build should conclude by running the example data through the tool (see point 3) and confirming that the expected output is obtained. We currently recognize two software solutions, although we will adapt as software evolves:
 - a. A [Dockerfile](#) that automatically builds a [container image](#) on [Docker Hub](#).
 - b. A [Singularity recipe file](#) that automatically builds a [container image](#) on [Singularity Hub](#).
6. Provide user support, clearly stating which level of support users can expect, and how and from whom to obtain it.

Recommendations

We suggest software tools be published under a license that permits free access, use, modification, and sharing, such as GPL, Apache 2.0, or MIT. However, we understand that this depends on institutional intellectual property restrictions, thus it is a recommendation rather than a requirement.

Explanatory Notes

Open Source Software and Versioned Repositories

Software tools in the AIRR field are evolving rapidly. In the interests of reproducibility and transparency, published work should be based on tools (and versions of tools) that can be obtained easily by other researchers in the future. To that end, AIRR compliant tools must be published in open repositories such as [GitHub](#) or [Bitbucket](#), and we encourage publishing users to provide specifics on the version and configuration of tools that have been employed.

Community-Curated File Formats

The AIRR Data Representation Working Group has defined standards for immune receptor repertoire sequencing datasets. Software tool authors are requested to support these standards as much as possible, for applicable data sets. The currently implemented standard covers submission of reads to NCBI repositories (BioProject, BioSample, SRA and Genbank) and annotated immune receptor rearrangements. Tool authors can assist by easing/guiding the process of submission as much as possible.

Example Data and Checks

Because the installation and operation of the tools in this field may be complex, we require example data and details of expected output, so that users can confirm that their installation is functioning as expected. Furthermore, metadata (for example, germline gene libraries) and other software dependencies should be checked when the tool runs, and informative error messages issued if necessary. A means should be provided to check the expected output automatically.

Dependencies and Containers

Containers encapsulate everything needed to run a piece of software into a single convenient executable that is largely independent of the user's software environment. For the following purposes, providers of AIRR-compliant tools must provide a containerized implementation (based on a published build script as described above) as one download option that users can choose:

- Containers allow users to use and evaluate a tool easily and reproduce results, without the need to resolve dependencies or configure the environment.
- Having these containers be automatically built also provides a self-validated way to understand the fine details of installation from a known starting point.

To ensure that containers are up to date, they must be built automatically when the current release version of the tool is updated. We will use automated builds on Docker Hub and Singularity Hub for this purpose. The corresponding build files document dependencies clearly, and make it easy for the maintainer to keep the container's dependencies up to date in subsequent releases.

An example Docker container is provided on the Software WG [Github Repository](#). This example encapsulates Ig-BLAST, and implements the [bioboxes](#) command-line standard.

Support Statements

Tool authors must provide support for the tool. They must state explicitly what level of support is provided, and explain how support can be obtained. We recommend a method such as the issues tracker on Github, that publishes support requests transparently and links resolutions to specific versions or releases. Users are advised to check that the level of support and the frequency of software updates matches their expectations before committing to a tool.

Analysis Workflows

- At the moment, we do not endorse a specific workflow technology standard:
 - Technology is evolving too rapidly for us to commit to a particular workflow.
 - Typically, AIRR analysis tools have many options and modes, which would make it difficult to support a “plug and play” environment without unduly restricting functionality.
- As tools and workflows evolve, we will keep the position under review and may make stronger technology recommendations in the future.
- We strongly encourage authors of tools to provide concrete, documented, examples of workflows that employ their tools, together with sample input and output data.
- **Likewise we encourage authors of research publications to provide** documented workflows that will enable interested readers to reproduce the results.

Ratification

Authors may submit tools to the AIRR Software WG requesting ratification against the standard. The submitter should provide a completed copy of the *AIRR Software WG - Compliance Checklist for AIRR Software Tools* to evidence reviewable and itemised evidence of compliance with each Requirement listed above.

The Software WG will, where appropriate, issue a Certificate of Compliance, stating the version of the tool reviewed and the version of the Standard with which compliance was ratified. After receiving a Certificate, authors will be entitled to claim compliance with the Standard, and may incorporate any artwork provided by AIRR for that purpose.

The Software WG will maintain and publish a list of compliant software.

If a tool does not achieve ratification, the Software WG will provide an explanation. The Software WG encourages resubmission once issues have been resolved.

Authors must re-submit tools for ratification following major upgrades or substantial modifications. The Software WG may, at its discretion, request resubmission at any time. If a certified tool subsequently fails ratification, or is not re-submitted in response to a request from the Software WG, AIRR compliance may no longer be claimed and the associated artwork may no longer be used.

The Software WG may, at its discretion, issue a new version of this standard at any time. Tools certified against previous version(s) of the standard may continue to claim compliance with those versions and to use the associated artwork. Authors wishing to claim compliance with the new version must submit a new request for certification and may not claim compliance with the new version, or use associated artwork, until and unless certification is obtained.

2.3.4 AIRR Data Commons API V1

Note: This section is about the API, if you are looking for information on the AIRR Data Commons, please see [here](#).

Overview

The AIRR Data Commons (ADC) API provides programmatic access to query and download AIRR-seq data. The ADC API uses JSON as its communication format, and standard HTTP methods like GET and POST. The ADC API is read-only and the mechanism of inclusion of AIRR-seq studies into a data repository is left up to the repository.

This documentation explains how to construct and execute API requests and interpret API responses.

Search and Retrieval

The AIRR Data Commons API specifies endpoints for searching and retrieving AIRR-seq data sets stored in an AIRR-compliant Data Repository according to the AIRR Data Model. This documentation describes Version 1 of the API. The general format of requests and associated parameters are described below.

The design of the AIRR Data Commons API was greatly inspired by National Cancer Institute's [Genomic Data Commons \(GDC\) API](#).

Repository implementation principles

Implementers of the ADC API should follow the following high level principles. Users of the ADC API can expect the following principles to be followed.

- All API endpoints should return JSON encodings as an API response.
- For some API endpoints it is possible to request TSV files, and those endpoints that support TSV files are documented in the <DataCommonsAPIEndpoints> section.
- Endpoints that are not documented as supporting TSV can reject TSV requests.
- If an API endpoint returns a field, then the content of that field in the JSON and TSV response must be equivalent.

Authentication

The ADC API currently does not define an authentication method. Future versions of the API will provide an authentication method so data repositories can support query and download of controlled-access data.

Table of Content

API Endpoints

The ADC API is versioned with the version number (v1) as part of the base path for all endpoints. The ADC API provides a query endpoint for each of the primary, high-level objects in the AIRR Standard to retrieve AIRR-seq related data about that type of object. The *repertoire* endpoint allows querying upon any field in the [Repertoire schema](#) including study, subject, sample, cell processing, nucleic acid processing, sequencing run, raw sequencing files, and data processing information. Queries on the content of raw sequencing files is not support but is supported on file attributes such as name, type and read information. Queries on other objects such as *Rearrangement*, *Clone*, *Cell*, *CellExpression*, *CellReactivity*, and *Receptor*, are provide respectively by the *rearrangement*, *clone*, *cell*, *expression*, *reactivity*, and *receptor* endpoints.

The standard workflow to retrieve all of the data for an AIRR-seq study involves performing a query on the `repertoire` endpoint to retrieve the repertoires in the study, and one or more queries on the other object endpoints (e.g. `rearrangement`) to download the data of that type for each repertoire. The endpoints are designed so the API response can be saved directly into a file and be used by AIRR analysis tools, including the AIRR python and R reference libraries, without requiring modifications or transformation of the data.

Each ADC API endpoint provides specific functionality as summarized in the following table:

Endpoint	Type	HTTP Method	Description
<code>/v1</code>	Service status	GET	Returns success if API service is running.
<code>/v1/info</code>	Service information	GET	Upon success, returns service information such as name, version, etc.
<code>/v1/repertoire/{repertoire_id}</code>	Retrieve a repertoire given its <code>repertoire_id</code>	GET	Upon success, returns the <code>Repertoire</code> information in JSON according to the Repertoire schema .
<code>/v1/repertoire</code>	Query repertoires	POST	Upon success, returns a list of <code>Repertoires</code> in JSON according to the Repertoire schema .
<code>/v1/rearrangement/{sequence_id}</code>	Retrieve a rearrangement given its <code>sequence_id</code>	GET	Upon success, returns the <code>Rearrangement</code> information in JSON format according to the Rearrangement schema .
<code>/v1/rearrangement</code>	Query rearrangements	POST	Upon success, returns a list of <code>Rearrangements</code> in JSON or AIRR TSV format according to the Rearrangement schema .
<code>/v1/clone/{clone_id}</code>	Retrieve a Clone given its <code>clone_id</code>	GET	Upon success, returns the <code>Clone</code> information in JSON format according to the Clone schema .
<code>/v1/clone</code>	Query clones	POST	Upon success, returns a list of <code>Clones</code> in JSON format according to the Clone schema .
<code>/v1/cell/{cell_id}</code>	Retrieve a Cell given its <code>cell_id</code>	GET	Upon success, returns the <code>Cell</code> information in JSON format according to the Cell schema .
<code>/v1/cell</code>	Query cells	POST	Upon success, returns a list of <code>Cells</code> in JSON format according to the Cell schema .
<code>/v1/expression/{expression_id}</code>	Retrieve a Expression Property given its <code>expression_id</code>	GET	Upon success, returns the <code>Expression</code> information in JSON format according to the CellExpression schema .
<code>/v1/expression</code>	Query Cell Expression properties	POST	Upon success, returns a list of <code>Expression Properties</code> in JSON format according to the CellExpression schema .
<code>/v1/receptor/{receptor_id}</code>	Retrieve a Receptor given its <code>receptor_id</code>	GET	Upon success, returns the <code>Receptor</code> information in JSON format according to the Receptor schema .
<code>/v1/receptor</code>	Query Receptor properties	POST	Upon success, returns a list of <code>Receptors</code> in JSON format according to the Receptor schema .
<code>/v1/reactivity/{cell_reactivity_id}</code>	Retrieve a CellReactivity given its <code>cell_reactivity_id</code>	GET	Upon success, returns the <code>CellReactivity</code> information in JSON format according to the <code>CellReactivity</code> schema.
<code>/v1/reactivity</code>	Query CellReactivity properties	POST	Upon success, returns a list of <code>CellReactivity</code> objects in JSON format according to the <code>CellReactivity</code> schema.

Repertoire Endpoint

The `repertoire` endpoint provides access to all fields in the [Repertoire schema](#). There are two type of endpoints; one for retrieving a single repertoire given its identifier, and another for performing a query across all repertoires in the data repository.

It is expected that the number of repertoires in a data repository will never become so large such that queries become

computationally expensive. A data repository might have thousands of repertoires across hundreds of studies, yet such numbers are easily handled by modern databases. Based upon this, the ADC API does not place limits on the repertoire endpoint for the fields that can be queried, the operators that can be used, or the number of results that can be returned.

Retrieve a Single Repertoire

Given a repertoire_id, a single Repertoire object will be returned.

```
curl https://vdjserver.org/airr/v1/repertoire/5993695857891348971-242ac118-0001-012
```

The response will provide the Repertoire data in JSON format.

```
{
  "Info": {
    "title": "AIRR Data Commons API for VDJServer Community Data Portal",
    "description": "VDJServer ADC API response for repertoire query",
    "version": "1.3",
    "contact": {
      "name": "VDJServer",
      "url": "http://vdjserver.org/",
      "email": "vdjserver@utsouthwestern.edu"
    }
  },
  "Repertoire": [
    {
      "repertoire_id": "5993695857891348971-242ac118-0001-012",
      "repertoire_name": null,
      "repertoire_description": null,
      "study": {
        "study_id": "4995411523885404651-242ac118-0001-012",
        "study_title": "T cell Receptor Repertoires Acquired via Routine Pap Testing May_
→Help Refine Cervical Cancer and Precancer Risk Estimates",
        "study_type": {
          "id": "NCIT:C16084",
          "label": "Observational Study"
        },
        "study_description": "Cervical cancer is the fourth most common cancer and
→fourth leading cause of cancer death among women worldwide. In low Human Development
→Index settings, it ranks second. Screening and surveillance involve the cytology-based
→Papanicolaou (Pap) test and testing for high-risk human papillomavirus (hrHPV). The
→Pap test has low sensitivity to detect precursor lesions, while a single hrHPV test
→cannot distinguish a persistent infection from one that the immune system will
→naturally clear. Furthermore, among women who are hrHPV-positive and progress to high-
→grade cervical lesions, testing cannot identify the ~20% who would progress to cancer
→if not treated. Thus, reliable detection and treatment of cancers and precancers
→requires routine screening followed by frequent surveillance among those with past
→abnormal or positive results. The consequence is overtreatment, with its associated
→risks and complications, in screened populations and an increased risk of cancer in
→under-screened populations. Methods to improve cervical cancer risk assessment,
→particularly assays to predict regression of precursor lesions or clearance of hrHPV
→infection, would benefit both populations. Here we show that women who have lower risk
→results on follow-up testing relative to index testing have evidence of enhanced T
→cell clonal expansion in the index cervical cytology sample compared to women who
→persist with higher risk results from index to follow-up. We further show that a
```

(continues on next page)

(continued from previous page)

```

↪machine learning classifier based on the index sample T cells predicts this transition,
↪to lower risk with 95% accuracy (19/20) by leave-one-out cross-validation. Using T_
↪cell receptor deep sequencing and machine learning, we identified a biophysicochemical_
↪motif in the complementarity-determining region 3 of T cell receptor chains whose_
↪presence predicts this transition. While these results must still be tested on an_
↪independent cohort in a prospective study, they suggest that this approach could_
↪improve cervical cancer screening by helping distinguish women likely to spontaneously_
↪regress from those at elevated risk of progression to cancer. The advancement of such_
↪a strategy could reduce surveillance frequency and overtreatment in screened_
↪populations and improve the delivery of screening to under-screened populations.",
    "inclusion_exclusion_criteria": "We included samples from White Hispanic women_
↪age 18 years or older. We excluded women who were HIV+, pregnant, had an intrauterine_
↪device, or had a sexually transmitted disease at the time of sample collection. We_
↪obtained samples across all cytology result categories: Negative for Intraepithelial_
↪Lesion or Malignancy (NILM, Normal), Abnormal Squamous Cells of Undetermined_
↪Significance (ASCUS), Low-grade Squamous Intraepithelial Lesion (LSIL), and High-grade_
↪Squamous Intraepithelial Lesion (HSIL). At Parkland Health and Hospital System (PHHS),_
↪the primary screening strategy is cytology alone with a reflex hrHPV test for women_
↪with an ASCUS cytology result. The test assays for positivity across 14 HPV types, and_
↪the ASCUS result category is divided into ASCUS/HPV- (negative for all 14 types) and_
↪ASCUS/HPV+ (positive for at least one type). An additional exclusion criterion was_
↪applied to women with a result of Normal, ASCUS/HPV-, and ASCUS/HPV+, and that is they_
↪were excluded if they had previously had cervical cancer or previous treatment of_
↪cervical pre-cancerous lesions.\n\nWe applied these inclusion and exclusion criteria_
↪in a quota sampling scheme to ensure adequate representation of women across all five_
↪result categories. We targeted a minimum of 100 samples total with a minimum of 15_
↪samples in each category, and then rescued all samples meeting our criteria each week_
↪until all minimums were reached.",
    "lab_name": "Lindsay G. Cowell",
    "lab_address": "UT Southwestern Medical Center",
    "submitted_by": "Scott Christley, scott.christley@utsouthwestern.edu, UT_
↪Southwestern Medical Center",
    "grants": "This research was supported by Simmons Comprehensive Cancer Center_
↪Development Funds and by a charitable donation from Young Texans Against Cancer, both_
↪to LC and JT.",
    "pub_ids": "PMID: 33868241",
    "keywords_study": [
        "contains_tcr"
    ],
    "adc_publish_date": "2021-08-05T03:50:02.295Z",
    "adc_update_date": "2021-08-05T05:43:14.260Z",
    "collected_by": null
},
"subject": {
    "subject_id": "5_20",
    "synthetic": false,
    "species": {
        "id": "NCBITAXON:9606",
        "label": "Homo sapiens"
    },
    "sex": "female",
    "age_min": 49.1,

```

(continues on next page)

(continued from previous page)

```

    "age_max": 49.1,
    "age_unit": {
      "id": "UO:0000036",
      "label": "year"
    },
    "ethnicity": "Hispanic",
    "race": "White",
    "diagnosis": [
      {
        "disease_diagnosis": {
          "id": null,
          "label": null
        },
        "study_group_description": null,
        "disease_length": null,
        "disease_stage": null,
        "prior_therapies": null,
        "immunogen": null,
        "intervention": null,
        "medical_history": null
      }
    ],
    "age_event": null,
    "ancestry_population": null,
    "strain_name": null,
    "linked_subjects": null,
    "link_type": null
  },
  "sample": [
    {
      "sample_id": "5_20_DNA",
      "sample_type": "cytology",
      "tissue": {
        "id": "UBERON:0004801",
        "label": "cervix epithelium"
      },
      "anatomic_site": "cervix",
      "disease_state_sample": "hsil",
      "cell_species": {
        "id": null,
        "label": null
      },
      "single_cell": false,
      "cell_storage": false,
      "template_class": "DNA",
      "template_amount": "2ug",
      "library_generation_method": "PCR",
      "library_generation_protocol": "Adaptive Biotechnologies",
      "library_generation_kit_version": "v3",
      "pcr_target": [
        {
          "pcr_target_locus": "TRB",

```

(continues on next page)

(continued from previous page)

```

        "forward_pcr_primer_target_location": null,
        "reverse_pcr_primer_target_location": null
    }
],
"complete_sequences": "partial",
"physical_linkage": "none",
"sequencing_run_id": "UTSW-Monson-P02-04",
"sequencing_run_date": "11/16/17",
"sequencing_files": {
    "file_type": "fasta",
    "filename": "5-20_DNA.fasta",
    "read_direction": "forward",
    "read_length": null,
    "paired_filename": null,
    "paired_read_direction": null,
    "paired_read_length": null
},
"sample_processing_id": null,
"collection_time_point_relative": null,
"collection_time_point_reference": null,
"biomaterial_provider": null,
"tissue_processing": null,
"cell_subset": {
    "id": null,
    "label": null
},
"cell_phenotype": null,
"cell_number": null,
"cells_per_reaction": null,
"cell_quality": null,
"cell_isolation": null,
"cell_processing_protocol": null,
"template_quality": null,
"total_reads_passing_qc_filter": null,
"sequencing_platform": null,
"sequencing_facility": null,
"sequencing_kit": null
}
],
"data_processing": [
    {
        "data_processing_id": "bf0617e7-b4a4-480f-99e3-b53eef9ca6d4-007",
        "primary_annotation": true,
        "software_versions": "igblast-ls5-1.14u2",
        "data_processing_files": [
            "5-20_DNA.igblast.airr.tsv.gz"
        ],
        "germline_database": "VDJServer IMGT 2019.01.23",
        "paired_reads_assembly": null,
        "quality_thresholds": null,
        "primer_match_cutoffs": null,
        "collapsing_method": null,
    }
]

```

(continues on next page)

(continued from previous page)

```
        "data_processing_protocols": null,
        "analysis_provenance_id": null
      }
    ]
  }
}
```

Query against all Repertoires

A query in JSON format is passed in a POST request. This example queries for repertoires of human IG heavy chain receptors for all studies in the data repository.

```
curl --data @query2_repertoire.json -H 'content-type: application/json' https://
  ↪ vldjserver.org/airr/v1/repertoire
```

The content of the JSON payload.

```
{
  "filters": {
    "op": "and",
    "content": [
      {
        "op": "=",
        "content": {
          "field": "subject.species.id",
          "value": "NCBITAXON:9606"
        }
      },
      {
        "op": "=",
        "content": {
          "field": "sample.pcr_target.pcr_target_locus",
          "value": "IGH"
        }
      }
    ]
  }
}
```

The response will provide a list of Repertoires in JSON format. The example output is not provided here due to its size.

Rearrangement Endpoint

The `rearrangement` endpoint provides access to all fields in the [Rearrangement schema](#). There are two type of endpoints; one for retrieving a single rearrangement given its identifier, and another for performing a query across all rearrangements in the data repository.

Unlike repertoire data, data repositories are expected to store millions or billions of rearrangement records, where performing “simple” queries can quickly become computationally expensive. Data repositories will need to optimize their databases for performance. Therefore, the ADC API does not require that all fields be queryable and only a limited set of query capabilities must be supported. The queryable fields are described in the Fields section below.

Retrieve a Single Rearrangement

Given a `sequence_id`, a single `Rearrangement` object will be returned.

```
curl https://vdjserver.org/airr/v1/rearrangement/610b77f6d5812c007f79bba3
```

The response will provide the `Rearrangement` data in JSON format.

```
{
  "Info":
  {
    "title": "AIRR Data Commons API reference implementation",
    "description": "API response for rearrangement query",
    "version": 1.3,
    "contact":
    {
      "name": "AIRR Community",
      "url": "https://github.com/airr-community"
    }
  },
  "Rearrangement":
  [
    {
      "sequence_id": "610b77f6d5812c007f79bba3",
      "repertoire_id": "5993695857891348971-242ac118-0001-012",
      "data_processing_id": "bf0617e7-b4a4-480f-99e3-b53eef9ca6d4-007",

      "... remaining fields": "snipped for space"
    }
  ]
}
```

Query against all Rearrangements

Supplying a `repertoire_id`, when it is known, should greatly speed up the query as it can significantly reduce the amount of data to be searched, though it isn't necessary.

This example queries for rearrangements with a specific junction amino acid sequence among a set of repertoires. A limited set of fields is requested to be returned. The resultant data can be requested in JSON or [AIRR TSV](#) format.

```
curl --data @query1_rearrangement.json -H 'content-type: application/json' https://
vdjserver.org/airr/v1/rearrangement
```

The content of the JSON payload.

```
{
  "filters":{
    "op": "and",
    "content": [
      {
        "op": "in",
        "content": {
          "field": "repertoire_id",
          "value": [
            "2603354229190496746-242ac113-0001-012",
            "2618085967015776746-242ac113-0001-012",
            "2633633748627296746-242ac113-0001-012",
```

(continues on next page)

(continued from previous page)

```

        "2564613624180576746-242ac113-0001-012"
      ]
    },
    {
      "op": "=",
      "content": {
        "field": "junction_aa",
        "value": "CARDPRSYHAFDIW"
      }
    }
  ]
},
"fields": ["repertoire_id", "sequence_id", "v_call", "productive"],
"format": "tsv"
}

```

Here is the response in AIRR TSV format.

sequence_id	productive	v_call	repertoire_id
5f70b421e10383007e3038ad	true	IGHV1-69*04	2564613624180576746-242ac113-0001-012
5f70b421e10383007e3038c2	true	IGHV1-69*04	2564613624180576746-242ac113-0001-012
5f70b421e10383007e3038f0	true	IGHV1-69*10	2564613624180576746-242ac113-0001-012
5f70b421e10383007e3039ec	true	IGHV1-69*04	2564613624180576746-242ac113-0001-012
5f70b421e10383007e303a1b	true	IGHV1-69*04	2564613624180576746-242ac113-0001-012
5f70b421e10383007e303a22	true	IGHV1-69*04	2564613624180576746-242ac113-0001-012
5f70b421e10383007e303a23	true	IGHV1-69*04	2564613624180576746-242ac113-0001-012
5f70b421e10383007e303a47	true	IGHV1-24*01	2564613624180576746-242ac113-0001-012
5f70b421e10383007e303b00	true	IGHV1-69*04	2564613624180576746-242ac113-0001-012
5f70b421e10383007e303baf	true	IGHV1-69*04	2564613624180576746-242ac113-0001-012

Clone Endpoint

The `clone` endpoint provides access to all fields in the [Clone schema](#). There are two type of endpoints; one for retrieving a single clone given its identifier, and another for performing a query across all clones in the data repository.

Unlike repertoire data, data repositories are expected to store millions or billions of clone records, where performing “simple” queries can quickly become computationally expensive. Data repositories will need to optimize their databases for performance. Therefore, the ADC API does not require that all fields be queryable and only a limited set of query capabilities must be supported. The queryable fields are described in the Fields section below.

Retrieve a Single Clone

Given a `clone_id`, a single Clone object will be returned.

```
curl https://covid19-1.ireceptor.org/airr/v1/clone/{clone_id}
```

Where `clone_id` is the ID of a clone object in the repository. The response will provide the Clone data in JSON format.

```

{
  "Info":
  {
    "title": "airr-api-ireceptor",
    "description": "AIRR Data Commons API for iReceptor",
    "version": "3.0",

```

(continues on next page)

(continued from previous page)

```

    "last_update": null,
    "contact": {
      "name": "iReceptor",
      "url": "http://www.ireceptor.org",
      "email": "support@ireceptor.org"
    }
  },
  "Clone":
  [
    {
      "clone_id": "clonotype1",
      "repertoire_id": "PRJCA002413-Healthy_Control_1-IG",
      "data_processing_id": "PRJCA002413-Healthy_Control_1",
      "sequences": null,
      "v_call": "IGHV2-70",
      "d_call": "",
      "j_call": "IGHJ3",
      "junction": "TGC GCACGGGCTCATTGTTCTGTTGGGCAGCAGCAGGTTCTGCTTTTGATATGTGG",
      "junction_aa": "CARAHCSWGSSRFGAFDMW",
      "junction_length": 57,
      "junction_aa_length": 19,
      "FIELDS REMOVED" : "FOR SPACE"
    }
  ]
}

```

Query against all Clones

Supplying a `repertoire_id`, when it is known, should greatly speed up the query as it can significantly reduce the amount of data to be searched, though it isn't necessary.

This example queries for clones with a specific junction amino acid sequence among a set of repertoires. A limited set of fields is requested to be returned. The resultant data is provided in JSON format.

```

curl -d '{"filters":{"op":"=","content":{"field":"junction_aa","value":
↪ "CARAHCSWGSSRFGAFDMW"}},"size":1}' -H 'content-type: application/json' http://covid19-
↪ 1.ireceptor.org/airr/v1/clone

```

This query searches the repository for clones that have a specific `junction_aa` field with a value of `CARAHCSWGSSRFGAFDMW` and requests only a single object in the response (`"size":1`). The response would be similar to that provided by the single clone query given above.

Cell Endpoint

The cell endpoint provides access to all fields in the *Cell schema*. There are two type of endpoints; one for retrieving a single cell given its identifier, and another for performing a query across all cells in the data repository.

Unlike repertoire data, data repositories are expected to store millions of cell records, where performing “simple” queries can quickly become computationally expensive. Data repositories will need to optimize their databases for performance. Therefore, the ADC API does not require that all fields be queryable and only a limited set of query capabilities must be supported. The queryable fields are described in the Fields section below.

Retrieve a Single Cell

Given a `cell_id`, a single Cell object will be returned.

```
curl https://covid19-1.ireceptor.org/airr/v1/cell/{cell_id}
```

Where `cell_id` is the ID of a cell object in the repository. The response will provide the Cell data in JSON format.

```
{
  "Info": {
    "title": "airr-api-ireceptor",
    "description": "AIRR Data Commons API for iReceptor",
    "version": "3.0",
    "last_update": null,
    "contact": {
      "name": "iReceptor",
      "url": "http://www.ireceptor.org",
      "email": "support@ireceptor.org"
    }
  },
  "Cell": [
    {
      "cell_id": "AAACCTGCACCGATAT-1",
      "rearrangements": null,
      "receptors": null,
      "reactivity_measurements": [ "61fc6c454f24ed3af5456a54" ],
      "repertoire_id": "PRJCA002413-ERS1-CELL",
      "data_processing_id": "PRJCA002413-ERS1",
      "expression_study_method": "single-cell transcriptome",
      "expression_raw_doi": null,
      "expression_index": null,
      "virtual_pairing": false
    }
  ]
}
```

Query against all Cells

Supplying a `repertoire_id`, when it is known, should greatly speed up the query as it can significantly reduce the amount of data to be searched, though it isn't necessary.

This example queries for clones with a specific junction amino acid sequence among a set of repertoires. A limited set of fields is requested to be returned. The resultant data is provided in JSON format.

```
curl -d '{"filters":{"op":"=", "content":{"field":"repertoire_id", "value":"PRJCA002413-ERS1-CELL"}},"size":1}' -H 'content-type: application/json' http://covid19-1.ireceptor.org/airr/v1/cell
```

This query searches the repository for cells that have a specific `repertoire_id` field with a value of `PRJCA002413-ERS1-CELL` and requests only a single object in the response (`"size":1`). The response would be similar to that provided by the single cell query given above.

Expression Endpoint

The expression endpoint provides access to all fields in the *CellExpression schema*. There are two type of endpoints; one for retrieving a single expression property given its identifier, and another for performing a query across all expression properties in the data repository.

Unlike repertoire data, data repositories are expected to store millions or billions of cell expression records, where performing “simple” queries can quickly become computationally expensive. Data repositories will need to optimize their databases for performance. Therefore, the ADC API does not require that all fields be queryable and only a limited set of query capabilities must be supported. The queryable fields are described in the Fields section below.

Retrieve a Cell Expression Property

Given a `expression_id`, a single Expression object will be returned.

```
curl https://covid19-1.ireceptor.org/airr/v1/expression/{expression_id}
```

Where `expression_id` is the ID of an expression object in the repository. The response will provide the CellExpression data in JSON format.

```
{
  "Info": {
    "title": "airr-api-ireceptor",
    "description": "AIRR Data Commons API for iReceptor",
    "version": "3.0",
    "last_update": null,
    "contact": {
      "name": "iReceptor",
      "url": "http://www.ireceptor.org",
      "email": "support@ireceptor.org"
    }
  },
  "CellExpression": [
    {
      "expression_id": "61fc6c454f24ed3af5456a54",
      "cell_id": "AAACGTGCAGCTTAAC-1",
      "repertoire_id": "PRJCA002413-Healthy_Control_1-CELL",
      "data_processing_id": "PRJCA002413-Healthy_Control_1",
      "property": {
        "label": "ISG15",
        "id": "ENSG:ENSG00000187608"
      },
      "value": 1
    }
  ]
}
```

Query against all Cell Expression data

Supplying a `repertoire_id` or `cell_id`, when it is known, should greatly speed up the query as it can significantly reduce the amount of data to be searched, though it isn't necessary.

This example queries for cell expression data with an ENSEMBL gene ID with the value `ENSG:ENSG0000017575` and requests only a single object response (`"size":1`). The resultant data is provided in JSON format and would be similar to that provided by the single expression property query given above.

```
curl -d '{"filters":{"op":"=","content":{"field":"property.id","value":
↪ "ENSG:ENSG0000017575"}},"size":1}' -H 'content-type: application/json' http://covid19-
↪ 1.ireceptor.org/airr/v1/expression
```

CellReactivity Endpoint

The reactivity endpoint provides access to all fields in the CellReactivitySchema. There are two type of endpoints: One for retrieving a single reactivity object given its identifier, and another for performing a query across all cell reactivity in the data repository.

To allow data repositories to optimize their databases for performance, the ADC API does not require that all fields in the CellReactivity object to be queryable and only a limited set of query capabilities must be supported. The queryable fields are described in the Fields section below.

Retrieve CellReactivity data

Given a `cell_reactivity_id`, a single CellReactivity object will be returned.

```
curl https://covid19-1.ireceptor.org/airr/v1/reactivity/{cell_reactivity_id}
```

Where `cell_reactivity_id` is the ID of a `CellReactivity` object in the repository. The response will provide the object in JSON format.

```
{
  "Info": {
    "title": "airr-api-ireceptor",
    "description": "AIRR Data Commons API for iReceptor",
    "version": "4.0",
    "last_update": null,
    "contact": {
      "name": "iReceptor",
      "url": "http://www.ireceptor.org",
      "email": "support@ireceptor.org"
    }
  },
  "CellReactivity": [
    {
      "cell_reactivity_id": "61fc6c454f24ed3af5456a54",
      "ligand_type": "MHC:peptide",
      "antigen_type": "peptide",
      "antigen": {
        "id": "NCBI:YP_009725307.1",
        "label": "RNA-dependent RNA polymerase_
↪(nsp12)"
      },
      "antigen_source_species": {
        "id": "NCBITAXON:2697049",
        "label": "Severe acute_
↪respiratory syndrome coronavirus 2"
      },
      "peptide_start": 738,
      "peptide_end": 746,
      "peptide_sequence_aa": "DTDFVNEFY",
      "mhc_class": "MHC-I",
      "mhc_gene_1": {
        "id": "MRO:0000046",
        "label": "HLA-A"
      },
      "mhc_allele_1": "HLA-A*01:01",
      "reactivity_method": "dextramer barcoding",
      "reactivity_readout": "barcode count",
      "reactivity_value": 23,
      "reactivity_unit": "absolute count"
    }
  ]
}
```

Query against all CellReactivity data

This example queries for receptor reactivity data that has a `peptide_sequence_aa` of `DTDFVNEFY` and requests only a single object of that type (`"size":1`). The resultant data is provided in JSON format and would be similar to that provided by the cell reactivity query response given above.

```
curl -d '{"filters":{"op":"=","content":{"field":"peptide_sequence_aa","value":"DTDFVNEFY
↪"},"size":1}}' -H 'content-type: application/json' http://covid19-1.ireceptor.org/airr/
↪v1/reactivity
```

Receptor Endpoint

The `receptor` endpoint provides access to all fields in the *Receptor Schema (Experimental)*. There are two type of endpoints: One for retrieving a single receptor given its identifier, and another for performing a query across all receptors in the data repository.

To allow data repositories to optimize their databases for performance, the ADC API does not require that all fields in the `Receptor` object to be queryable and only a limited set of query capabilities must be supported. The queryable

fields are described in the Fields section below.

Retrieve a Receptor

Given a `receptor_id`, a single Receptor object will be returned.

```
curl https://covid19-1.ireceptor.org/airr/v1/receptor/{receptor_id}
```

Where `receptor_id` is the ID of a Receptor object in the repository. The response will provide the object in JSON format.

```
{
  "Info": {
    "title": "airr-api-ireceptor",
    "description": "AIRR Data Commons API for iReceptor",
    "version": "3.0",
    "last_update": null,
    "contact": {
      "name": "iReceptor",
      "url": "http://www.ireceptor.org",
      "email": "support@ireceptor.org"
    }
  },
  "Receptor": [
    {
      "receptor_id": "IG-MM-BALB-123456",
      "receptor_hash": "aa1c4b77a6f4927611ab39f5267415beaa0ba07a952c233d803b07e52261f026",
      "receptor_type": "Ig",
      "receptor_variable_domain_1_aa":
        "QVQLQQPGAELVKPGASVKLSCKASGYTFTSYWMHWVKRPGRGLEWIGRIDPNSGGTKYNEKFKSKATLTVDKPSSTAYMQLSSLTSEDSAVYYCARYD",
      "receptor_variable_domain_1_locus": "IGH",
      "receptor_variable_domain_2_aa":
        "QAVVTQESALTTSPGETVTLTCRSSTGAVTTSNYANWVQEKPDHLFTGLIGGTNNRAPGVPARFSGSLIGDKAALTITGAQTEDEATYFCALWYSNHWVF",
      "receptor_variable_domain_2_locus": "IGL",
      "receptor_ref": [ "IEDB_RECEPTOR:29263" ],
      "reactivity_measurements": [ "61fc6c454f24ed3af5456a54" ]
    }
  ]
}
```

Query against all Receptor data

This example queries for receptor data that has a TCR receptor type and requests only a single object response (`"size":1`). The resultant data is provided in JSON format and would be similar to that provided by the single expression property query given above.

```
curl -d '{"filters":{"op":"=","content":{"field":"receptor_type","value":"TCR"}},"size":1}' -H 'content-type: application/json' http://covid19-1.ireceptor.org/airr/v1/receptor
```

Components of a Request

The ADC API has two classes of endpoints. The endpoints that respond to GET requests are simple services that require few or no parameters. While, the endpoints that response to POST requests are the main query services and provide many parameters for specifying the query as well as the data in the API response.

A typical POST query request specifies the following parameters:

- The `filters` parameter specifies the query.
- The `from` and `size` parameters specify the number of results to skip and the maximum number of results to be returned in the response.
- The `fields` parameter specifies which data elements to be returned in the response. By default all fields (AIRR and non-AIRR) stored in the data repository are returned. This can vary between data repositories based upon how the repository decides to store blank or null fields, so the `fields` and/or `include_fields` parameter should be used to guarantee the existence of data elements in the response.
- The `include_fields` parameter specifies the set of AIRR fields to be included in the response. This parameter can be used in conjunction with the `fields` parameter, in which case the list of fields is merged. This is a mechanism to ensure that specific, well-defined sets of AIRR data elements are returned without requiring all of those fields to be individually provided in the `fields` parameter.

The sets that can be requested are summarized in the table below.

include_fields	MiAIRR	AIRR required	AIRR identifiers	other AIRR/ADC fields
miairr	Y	some	N	N
airr-core	Y	Y	Y	N
airr-schema	Y	Y	Y	Y

Service Status Example

The following is an example GET request to check that the service API is available for VDJServer's data repository.

```
curl https://vdjserver.org/airr/v1
```

The response should indicate success.

```
{"result": "success"}
```

Service Info Example

The following is an example GET request to get information about the service.

```
curl https://vdjserver.org/airr/v1/info
```

The response provides various information about the repository itself, the API that is implemented by the repository, the schema version of the data that is returned from repository queries, as well as repository specific details such as the maximum number of records that are allowed to be requested in a single query as well as the maximum size (in bytes) of the queries sent to the repository.

```
{
  "title": "api-js-tapis",
  "description": "AIRR Data Commons API for VDJServer Community Data Portal",
  "version": "2.0.0",
  "contact": {
```

(continues on next page)

(continued from previous page)

```

    "name": "VDJServer",
    "url": "http://vdjserver.org/",
    "email": "vdjserver@utsouthwestern.edu"
  },
  "license": {
    "name": "GNU AGPL V3"
  },
  "api": {
    "title": "AIRR Data Commons API",
    "version": "1.0.0",
    "contact": {
      "name": "AIRR Community",
      "url": "http://www.airr-community.org/",
      "email": "join@airr-community.org"
    },
    "description": "Major Version 1 of the Adaptive Immune Receptor Repertoire (AIRR)_
↪data repository web service application programming interface (API).\n",
    "license": {
      "name": "Creative Commons Attribution 4.0 International",
      "url": "https://creativecommons.org/licenses/by/4.0/"
    }
  },
  "schema": {
    "title": "AIRR Schema",
    "description": "Schema definitions for AIRR standards objects",
    "version": "1.3",
    "contact": {
      "name": "AIRR Community",
      "url": "https://github.com/airr-community"
    },
    "license": {
      "name": "Creative Commons Attribution 4.0 International",
      "url": "https://creativecommons.org/licenses/by/4.0/"
    }
  },
  "max_size": 1000,
  "max_query_size": 2097152
}

```

Query Repertoire Example

The following is an example POST request to the repertoire endpoint of the ADC API. It queries for repertoires of human TCR beta receptors (filters), skips the first 10 results (from), requests 5 results (size), and requests only the repertoire_id field (fields).

```

curl --data @query1-2_repertoire.json -H 'content-type: application/json' https://
↪vdjserver.org/airr/v1/repertoire

```

The content of the JSON payload.

```

{
  "filters":{
    "op":"and",

```

(continues on next page)

(continued from previous page)

```

    "content": [
      {
        "op": "=",
        "content": {
          "field": "subject.species.id",
          "value": "NCBITAXON:9606"
        }
      },
      {
        "op": "=",
        "content": {
          "field": "sample.pcr_target.pcr_target_locus",
          "value": "TRB"
        }
      }
    ]
  },
  "from": 10,
  "size": 5,
  "fields": ["repertoire_id"]
}

```

The response contains two JSON objects, an Info object that provides information about the API response and a Repertoire object that contains the list of Repertoires that met the query search criteria. In this case, the query returns a list of five repertoire identifiers. Note the Info object is based on the info block as specified in the OpenAPI v2.0 specification.

```

{
  "Info":
  {
    "title": "AIRR Data Commons API reference implementation",
    "description": "API response for repertoire query",
    "version": 1.3,
    "contact":
    {
      "name": "AIRR Community",
      "url": "https://github.com/airr-community"
    }
  },
  "Repertoire":
  [
    {"repertoire_id": "5993695857891348971-242ac118-0001-012"},
    {"repertoire_id": "5981154557681996267-242ac118-0001-012"},
    {"repertoire_id": "6018649617881108971-242ac118-0001-012"},
    {"repertoire_id": "5959121371158548971-242ac118-0001-012"},
    {"repertoire_id": "5939278622251028971-242ac118-0001-012"}
  ]
}

```

Request Parameters

The ADC API supports the follow query parameters. These are only applicable to the query endpoints, i.e. the HTTP POST endpoints.

Parameter	Default	Description
<code>filters</code>	null	Specifies logical expression for query critieria
<code>format</code>	JSON	Specifies the API response format: JSON, AIRR TSV
<code>include_fields</code>	null	Specifies the set of AIRR fields to be included in the response
<code>fields</code>	null	Specifies which fields to include in the response
<code>from</code>	0	Specifies the first record to return from a set of search results
<code>size</code>	repository dependent	Specifies the number of results to return
<code>facets</code>	null	Provide aggregate count information for the specified fields

Filters Query Parameter

The `filters` parameter enables passing complex query criteria to the ADC API. The parameter represents the query in a JSON object.

A `filters` query consists of an operator (or a nested set of operators) with a set of `field` and `value` operands. The query criteria as represented in a JSON object can be considered an expression tree data structure where internal nodes are operators and child nodes are operands. The expression tree can be of any depth, and recursive algorithms are typically used for tree traversal.

The following operators are support by the ADC API.

Op-er-a-tor	Operan	Value Data Types	Description	Example
=	field and value	string, number, integer, or boolean	equals	{“op”：“=”,“content”:{“field”：“junction_aa”,“value”：“CASSYIKLN”}}
!=	field and value	string, number, integer, or boolean	does not equal	{“op”：“!=”,“content”:{“field”：“subject.organism.id”,“value”：“9606”}}
<	field and value	number, integer	less than	{“op”：“<”,“content”:{“field”：“sample.cell_number”,“value”：1000}}
<=	field and value	number, integer	less than or equal	{“op”：“<=”,“content”:{“field”：“sample.cell_number”,“value”：1000}}
>	field and value	number, integer	greater than	{“op”：“>”,“content”:{“field”：“sample.cells_per_reaction”,“value”：10000}}
>=	field and value	number, integer	greater than or equal	{“op”：“>=”,“content”:{“field”：“sample.cells_per_reaction”,“value”：10000}}
is missing	field	n/a	field is missing or is null	{“op”：“is missing”,“content”:{“field”：“sample.tissue”}}
is	field	n/a	identical to “is missing” operator, provided for GDC compatibility	{“op”：“is”,“content”:{“field”：“sample.tissue”}}
is not missing	field	n/a	field is not missing and is not null	{“op”：“is not missing”,“content”:{“field”：“sample.tissue”}}
not	field	n/a	identical to “is not missing” operator, provided for GDC compatibility	{“op”：“not”,“content”:{“field”：“sample.tissue”}}
in	field, multiple values in a list	array of string, number, or integer	matches a string or number in a list	{“op”：“in”,“content”:{“field”：“subject.strain_name”,“value”：[“C57BL/6”,“BALB/c”,“
ex-clude	field, multiple values in a list	array of string, number, or integer	does not match any string or number in a list	{“op”：“exclude”,“content”:{“field”：“subject.strain_name”,“value”：[“SCID”,“NOD”]}}
con-tains	field, value	string	contains the sub-string	{“op”：“contains”,“content”:{“field”：“study.study_title”,“value”：“cancer”}}
132 and	multiple operators	n/a	logical AND	Chapter 2. Table of Contents {“op”：“and”,“content”:[{“op”：“!=”,“content”:{“field”：“subject.organism.id”,“value”：“9606”}}, {“op”：“>=”,“content”:{“field”：“sample.cells_per_reaction”,“value”：10000}}, {“op”：“exclude”,“content”:{“field”：“subject.strain_name”,“value”：[“SCID”,“NOD”]}}]}

Note that the `not` operator is different from a logical NOT operator, and the logical NOT is not needed as the other operators provide negation.

The field operand specifies a fully qualified property name in the AIRR Data Model. Fully qualified AIRR properties are either a JSON/YAML base type (string, number, integer, or boolean) or an array of one of these base types (some AIRR fields are arrays e.g. `study.keywords_study`). The Fields section below describes the available queryable fields.

The value operand specifies one or more values when evaluating the operator for the field operand.

Queries Against Arrays

A number of fields in the AIRR Data Model are arrays, such as `study.keywords_study` which is an array of strings or `subject.diagnosis` which is an array of `Diagnosis` objects. A query operator on an array field will apply that operator to each entry in the array to decide if the query filter is satisfied. The behavior is different for various operators. For operators such as `=` and `in`, the filter behaves like the Boolean OR over the array entries, that is if **any** array entry evaluates to true then the query filter is satisfied. For operators such as `!=` and `exclude`, the filter behaves like the Boolean AND over the array entries, that is **all** array entries must evaluate to true for the query filter to be satisfied.

For complex queries over arrays, it is necessary to compose complex queries into more than one query. For example consider the following subject:

```
* Subject
* diagnosis
  (Diagnosis record 1)
    * disease_diagnosis: "rheumatoid arthritis"
    * disease_length: "20 years"
  (Diagnosis record 2)
    * disease_diagnosis: "pancreatic ductal adenocarcinoma"
    * disease_length: "6 months"
```

If the end result that is required it to find all disease diagnoses of “pancreatic ductal adenocarcinoma” that have a disease length of over 10 years, searching for `disease_diagnosis = pancreatic ductal adenocarcinoma` and `disease_length > 10` will result in the above Subject being returned, even though the subject has not had pancreatic ductal adenocarcinoma for more than 10 years. This is because there is a diagnosis of pancreatic ductal adenocarcinoma and a disease length of more than 10 years but from a different diagnoses. This is a correct response to the query, but does not return the desired outcome.

In order to achieve the desired outcome, it is necessary to search for one of the conditions (e.g. `disease_diagnosis = pancreatic ductal adenocarcinoma`), compile a list of `repertoire_ids` that meet that condition, and then search for the second condition (e.g. `disease_length > 10`) across those `repertoire_ids`.

Examples

A simple query with a single operator looks like this:

```
{
  "filters": {
    "op": "=",
    "content": {
      "field": "junction_aa",
      "value": "CASSYIKLN"
    }
  }
}
```

A more complex query with multiple operators looks like this:

```
{
  "filters": {
    "op": "and",
    "content": [
      {
        "op": "!=",
        "content": {
          "field": "subject.organism.id",
          "value": "9606"
        }
      },
      {
        "op": ">=",
        "content": {
          "field": "sample.cells_per_reaction",
          "value": "10000"
        }
      },
      {
        "op": "exclude",
        "content": {
          "field": "subject.organism.id",
          "value": ["9606", "10090"]
        }
      }
    ]
  }
}
```

Format Query Parameter

Specifies the format of the API response. `json` is the default format and is available for all endpoints. The rearrangement POST endpoint also accepts `tsv` which will provide the data in the *AIRR TSV* format. A specific ordering of fields in the TSV format should not be assumed from one API request to another. Take care to properly merge AIRR TSV data from multiple API requests, e.g. such as with the `airr-tools merge` program.

Fields Query Parameter

The `fields` parameter specifies which fields are to be included in the API response. By default all fields (AIRR and non-AIRR) stored in the data repository are returned. However, this can vary between data repositories based upon how the repository decides to store blank or null fields, so the `fields` and/or `include_fields` parameter should be used to guarantee the existence of data elements in the response.

Include Fields Query Parameter

The `include_fields` parameter specifies that the API response should include a well-defined set of AIRR Standard fields. These sets include:

- `miairr`, for only the MiAIRR fields.
- `airr-core`, for the AIRR required and identifier fields. This is expected to be the most common option as it provides all MiAIRR fields, additional required fields useful for analysis, and all identifier fields for linking objects in the AIRR Data Model.
- `airr-schema`, for all AIRR fields in the AIRR Schema.

The `include_fields` parameter is a mechanism to ensure that specific AIRR data elements are returned without requiring those fields to be individually provided with the `fields` parameter. Any data elements that lack a value will

be assigned null in the response. Any empty array of objects, for example `subject.diagnosis`, will be populated with a single object with all of the object's properties given a null value. Any empty array of primitive data types, like string or number, will be assigned null. Note that if both the `include_fields` and the `fields` parameter are provided, the API response will include the set of AIRR fields and in addition will include any additional fields that are specified in the `fields` parameter.

Size and From Query Parameters

The ADC API provides a pagination feature that limits the number of results returned by the API.

The `from` query parameter specifies which record to start from when returning results. This allows records to be skipped. The default value is 0 indicating that the first record in the set of results will be returned.

The `size` query parameters specifies the maximum number of results to return. The default value is specific to the data repository, and a maximum value may be imposed by the data repository. This is to prevent queries from “accidentally” returning millions of records. The `info` endpoint provides the data repository default and maximum values for the `repertoire` and `rearrangement` endpoints, which may have different values. A value of 0 indicates there is no limit on the number of results to return, but if the data repository does not support this then the default value will be used.

The combination of `from` and `size` can be used to implement pagination in a graphical user interface, or to split a very large download into smaller batches. For example, if an interface displays 10 records at a time, the request would assign `size=10` and `from=0` to get the ten results to display on the first page. When the user traverses to the “next page”, the request would assign `from=10` to skip the first ten results and return the next ten results, and `from=20` for the next page after that, and so on.

Facets Query Parameter

The `facets` parameter provides aggregate count information for the specified field. Only a single field can be specified. The `facets` parameter can be used in conjunction with the `filters` parameter to get aggregate counts for a set of search results. It returns the set of values for the field, and the number of records (repertoires or rearrangement) that have this value. For field values that have no counts, the API service can either return the field value with a 0 count or exclude the field value in the aggregation. The typical use of this parameter is for displaying aggregate information in a graphical user interface.

Here is a simple query with only the `facets` parameter to return the set of values for `sample.pcr_target.pcr_target_locus` and the count of repertoires repertoires that have each value. The content of the JSON payload.

```
{
  "facets": "sample.pcr_target.pcr_target_locus"
}
```

Sending this query in an API request.

```
curl --data @facets1_repertoire.json -H 'content-type: application/json' https://
vdjserver.org/airr/v1/repertoire
```

The output from the request is similar to normal queries except the data is provided with the *Facet* key.

```
{
  "Info": {
    "title": "AIRR Data Commons API for VDJServer Community Data Portal",
    "description": "VDJServer ADC API response for repertoire query",
    "version": "1.3",
    "contact": {
      "name": "VDJServer",
      "url": "http://vdjserver.org/",
      "email": "vdjserver@utsouthwestern.edu"
    }
  }
```

(continues on next page)

(continued from previous page)

```

},
"Facet": [
  {
    "sample.pcr_target.pcr_target_locus": "TRB",
    "count": 2786
  },
  {
    "sample.pcr_target.pcr_target_locus": "TRA",
    "count": 242
  },
  {
    "sample.pcr_target.pcr_target_locus": "IGK",
    "count": 122
  },
  {
    "sample.pcr_target.pcr_target_locus": "IGH",
    "count": 547
  },
  {
    "sample.pcr_target.pcr_target_locus": "IGL",
    "count": 121
  }
]
}

```

Here is a query with both `filters` and `facets` parameters, which restricts the data records used for the facets count. The content of the JSON payload.

```

{
  "filters":{
    "op":"and",
    "content": [
      {
        "op": "=",
        "content": {
          "field": "study.study_id",
          "value": "PRJNA300878"
        }
      },
      {
        "op": "=",
        "content": {
          "field": "sample.pcr_target.pcr_target_locus",
          "value": "IGH"
        }
      }
    ]
  },
  "facets": "subject.subject_id"
}

```

Sending this query in an API request.

```
curl --data @facets2_repertoire.json -H 'content-type: application/json' https://
↳ vdjserver.org/airr/v1/repertoire
```

Example output from the request. This result indicates there are ten subjects each with two IGH repertoires.

```
{
  "Info": {
    "title": "AIRR Data Commons API reference implementation",
    "description": "API response for repertoire query",
    "version": 1.3,
    "contact": {
      "name": "AIRR Community",
      "url": "https://github.com/airr-community"
    }
  },
  "Facet": [
    {"subject.subject_id": "TW05B", "count": 2},
    {"subject.subject_id": "TW05A", "count": 2},
    {"subject.subject_id": "TW03A", "count": 2},
    {"subject.subject_id": "TW04A", "count": 2},
    {"subject.subject_id": "TW01A", "count": 2},
    {"subject.subject_id": "TW04B", "count": 2},
    {"subject.subject_id": "TW02A", "count": 2},
    {"subject.subject_id": "TW03B", "count": 2},
    {"subject.subject_id": "TW01B", "count": 2},
    {"subject.subject_id": "TW02B", "count": 2}
  ]
}
```

Queries on Nested Information

As stated above, in general API response data will have been flattened by the query handler. However, there are several instances in which properties within the top-level entities are arrays of objects, which cannot be flattened because all the information will be expected to present in the response. Therefore, in these cases, the data that is queried and potentially returned will be nested. In addition, while the array of object is obvious from the AIRR Schema, the array component does **not** appear in the hierarchical property names used by the API. Note that this does not create any collisions as the schema allows the existence of multiple properties with the same designation.

However, it results in two possible ways how an AND operator can behave when using such nested properties as input. These are defined as follows:

Given that two or more instances of an object class exists within an array that is a property of a higher-level entity, and given a query that contains an AND operation that uses tests on two or more properties of said object class as an input, the code handling the query will exhibit

- *local* behavior, if for the AND operation to evaluate to TRUE it requires that all tests must succeed within an instance of the object and within at least one object of the array, i.e., the code is aware of the nesting and is able to parse the hierarchy of properties from the provided string, or
- *global* behavior, if for the AND operation to evaluate to TRUE it requires all tests to succeed, but independent of the instances in which the matching properties are located, i.e., the code is agnostic to the nesting and treats all properties within the array as a single set.

While both behaviors have their use cases, ADC API handlers are expected to exhibit “local” behavior, as is easier to implement on the client-side, where it would require joining the the result sets of the queries for each of the properties

individually.

ADC API Limits and Thresholds

Repertoire endpoint query fields

It is expected that the number of repertoires in a data repository will never become so large such that queries become computationally expensive. A data repository might have thousands of repertoires across hundreds of studies, yet such numbers are easily handled by databases. Based upon this, the ADC API does not place limits on the repertoire endpoint for the fields that can be queried or the operators that can be used.

Other endpoint query fields

Unlike repertoire data, data repositories are expected to store billions of other records (e.g. Rearrangement, CellExpression, Clone, Cell), where performing “simple” queries can quickly become computationally expensive. Data repositories are encouraged to optimize their databases for performance. Therefore, based upon a set of query use cases provided by immunology experts, a minimal set of required fields was defined that can be queried. These required fields are described in the following Table. The fields also have the AIRR extension property `adc-query-support: true` in the AIRR Schema.

Minimal Rearrangement Query Fields

Field(s)	Description
sequence_id, repertoire_id, sample_processing_id, data_processing_id, clone_id, cell_id	Identifiers; sequence_id allows for query of that specific rearrangement object in the repository, while repertoire_id, sample_processing_id, and data_processing_id are links to the repertoire metadata for the rearrangement. The clone_id and cell_id are identifiers that group rearrangements based on clone assignment and single cell assignment.
locus, v_call, d_call, j_call, c_call, productive, junction_aa, junction_aa_length	Commonly used rearrangement annotations.

Minimal Clone Query Fields

Field(s)	Description
clone_id, repertoire_id, data_processing_id	Identifiers; clone_id allows for query of that specific clone object in the repository, while repertoire_id and data_processing_id are links to the repertoire metadata for the clone.
v_call, d_call, j_call, junction_aa, junction_aa_length, clone_count	Commonly used clone annotations.

Minimal Cell Query Fields

Field(s)	Description
cell_id, repertoire_id, data_processing_id	Identifiers; cell_id allows for query of that specific cell object in the repository, while repertoire_id and data_processing_id are links to the repertoire metadata for the cell.
rearrangements, receptors, re-activity_measurements, expression_study_method, virtual_pairing	Commonly used cell attributes.

Minimal CellExpression Query Fields

Field(s)	Description
expression_id, cell_id, repertoire_id, data_processing_id	Identifiers; expression_id allows for query of that specific expression object in the repository, while repertoire_id and data_processing_id are links to the repertoire metadata for the cell. cell_id allows for searching for expression data for a specific cell.
property, value	Commonly used cell expression attributes.

Minimal Receptor Query Fields

Field(s)	Description
receptor_id, receptor_hash	Identifiers; receptor_id allows for query of that specific receptor object in the repository. receptor_hash allows for a fast/efficient look up of the globally unique receptor hash.
receptor_type, receptor_variable_domain_1_aa, receptor_variable_domain_1_locus, receptor_variable_domain_2_aa, receptor_variable_domain_2_locus, receptor_ref, reactivity_measurements	Commonly used receptor attributes.

Minimal ReceptorReactivity Query Fields

Field(s)	Description
receptor_reactivity_id	Identifiers; receptor_reactivity_id allows for query of that specific receptor reactivity object in the repository.
ligand_type, antigen_type, antigen, antigen_source_species, peptide_sequence_aa, mhc_class, mhc_gene_1, mhc_allele_1, mhc_gene_2, mhc_allele_2, reactivity_method, reactivity_readout, reactivity_value, reactivity_unit	Commonly used receptor reactivity attributes.

Data repository specific limits

A data repository may impose limits on the size of the data sent to and returned by the repository. This might be because of limitations imposed by the back-end database being used or because of the need to manage the load placed on the server. For example, MongoDB databases have document size limits (16 megabytes) which limit the size of a query that can be sent to a repository and the size of a single repertoire or rearrangement object that is returned. As a result a repository might choose to set a maximum query size.

Size limits can be retrieved from the `info` endpoint. These limits are repository wide (the same for all endpoints). If the data repository does not provide a limit, then no limit is assumed.

Field	Description
<code>max_size</code>	The maximum value for the <code>size</code> query parameter. Attempting to retrieve data from a repository beyond this maximum should trigger an error response. The error response should include information about why the query failed and what the maximum size limit is.
<code>max_query_size</code>	The maximum size (in bytes) of the JSON query object. Attempting to send a query to a repository larger than this size should trigger an error.

ADC API Reference Implementation

The AIRR Community provides a reference implementation for an ADC API service. The reference implementation can be utilized for any number of tasks. For example, a data repository might use the source code as a starting point for their own implementation and can compare the behaviour of their service against the reference. Another example is a tool developer, who wishes to use the API, can setup a local data repository so they can develop and test their tool before sending API requests across the internet to remote data repositories. While the reference implementation is functionally complete, it has minimal security and no optimizations for large data so it should not be used directly for production systems.

The reference implementation consists of three GitHub repositories: [adc-api](#), [adc-api-js-mongodb](#), and [adc-api-mongodb-repository](#). The three repositories correspond to the top-level service composition (`adc-api`), a JavaScript web service that responds to API requests and queries a MongoDB database (`adc-api-js-mongodb`), and a MongoDB database for holding AIRR-seq data (`adc-api-mongodb-repository`). Docker and docker-compose are used to provide a consistent deployment environment and compose the multiple components together into a single service. Complete documentation for configuring and deploying the reference implementation is available in the `adc-api` repository.

2.3.5 AIRR Ontologies and Vocabularies Sub-WG

Summary

The “Ontologies and Vocabularies Team” was initially formed as a joint interest group of the Common Repository (Com-Repo) and the Minimal Standards (MiniStd) working groups (WG) of the AIRR Community. When the two WG merged into the current Standards WG in December 2020, OntoVoc became a Sub-WG of it. The long-term aim of the Sub-WG is to define standard vocabularies and ontologies to be used by AIRR-compliant repositories.

Ontology Data Representation

The nodes in an ontology are typically either concepts (e.g., capital) or instances thereof (e.g., Paris). These nodes have *local IDs* (often numbers), which are unique within an ontology. They also typically have *labels*, which is the human-readable name of the node. Ontology entities in the AIRR Data Standard reflect this model, with each AIRR field that is represented as an ontology recorded with a global *ontology ID* (*id*) and the corresponding *label* (*label*).

Within the AIRR Standards, Compact URIs (**CURIEs**) are used to represent *ontology IDs* or *persistent IDs*. CURIEs are a standardized way to abbreviate International Resource Identifiers (IRI, [RFC3987]), which include URIs and URLs as subsets. They were originally conceived to simplify the handling of attributes, e.g., in XML or SPARQL, by making them more compact and readable. CURIEs are also used by IEDB databases to reduce redundancies (mainly in the leading part of IRIs).

For example, a typical CURIE would look like NCBITAXON:9258. In this case, NCBITAXON is the *prefix*, a custom string that will be replaced by a repository-defined IRI component (e.g., http://purl.obolibrary.org/obo/NCBITaxon_). Note that there is no connection between NCBITAXON in the CURIE and NCBITaxon in the IRI, the former one is just a placeholder. Although common, it is not always the case that a resolved CURIE (the IRI *prefix* plus the *local ID*) can be used as a URL directly to look up the CURIE using a web browser.

The AIRR Schema provides a CURIEMap, a list of AIRR approved CURIE *prefixes* along with a map of at least one *iri_prefix* (i.e., a replacement string to construct the complete IRI) for each *prefix*. As the *iri_prefix* might differ between *provider*-specific implementations of an ontology (e.g., NCBI Taxonomy), the CURIEMap supports multiple *iri_prefix* entries for a given *prefix*. Finally, the CURIEMap should also provide a default map and *provider* for each *prefix*. Complementary to this, the InformationProvider list describes the mechanism to computationally look up a resolved IRI (e.g., the *iri_prefix* and the *local ID*) by specifying how to make a request to the *provider* as well as describing the format in which the request response will be provided.

The CURIEMap serves several purposes:

1. It provides a controlled namespace for CURIE *prefixes* used in the AIRR Schema. For now, custom additions to or replacements of these *prefixes* in the schema are prohibited. This does not affect the ability of repositories to use such custom prefixes internally.
2. It simplifies resolution of CURIEs. The *iri_prefix* lists for each *prefix* should not be considered to be exhaustive. However, when using a custom *iri_prefix*, it must be ensured that the expanded IRI still refers to the same concept/instance as when using the default *iri_prefix*.
3. It simplifies computation using CURIEs. It is possible to use the *provider* for a *prefix* as a mechanism to look up a CURIE from a *provider* with a defined response (See below)

It should be explicitly noted that the CURIEMap should not be interpreted as any kind of recommendation for certain *providers*. It is left up to users to decide how to resolve the resulting IRIs, e.g., via DNS/HTTP (if possible) or by using a *provider* of their choice.

General Policies

Criteria

Ontologies used within AIRR standards

1. MUST **[1]** cover the majority of the required terms, but complete coverage is OPTIONAL
2. MUST have a structure that is scientifically correct and logically coherent
3. MUST NOT feature complexity that makes it hard to use for queries and data representation
4. SHOULD already be widely adopted
5. MUST be actively maintained

6. MUST be available under a free license
7. SHOULD comply to the [OBO Foundry Principles](#). This does not imply a preference.

Comments on criteria:

- ad (1): For most fields it will be difficult to find complete and accurate ontologies. Therefore picking the best available ontology and working with its maintainers to include missing terms is expected to be the most sustainable approach.
- ad (5): This requirement follows from (1), as there needs to be a way for term requests.
- ad (6): A number of ontologies need to be licensed from their respective copyright holders. This results in potential barriers for implementation and distribution of such ontologies. Therefore only ontologies available under a free license are considered suitable for AIRR-compliant databases. The list of suitable licenses is not final, but includes: [CC0](#) and [CC BY](#).
- ad (7): This is an endorsement of the OBO Foundry *Principles*, not of the OBO Foundry *Ontologies* in general. Hence, also non-OBO have an equal standing if they comply to the Principles.

Approved Ontologies

- Cell ontology ([CL](#))
 - used in:
 - * Cell subset (`cell_subset`, *Tissue and Cell Processing*)
 - CURIE summary
 - * CURIE Prefix: CL
 - * CURIE IRI Prefix: http://purl.obolibrary.org/obo/CL_
 - example AIRR use
 - * “`cell_subset.id`” : “CL:0000542”
 - * “`cell_subset.label`” : “lymphocyte”
 - default root node
 - * label: lymphocyte
 - * local id: CL_0000542
 - * path: ``
 - license: [CC BY](#)
 - latest release (as of 2020-05-20): 2020-03-02
 - repo: <https://github.com/obophenotype/cell-ontology>
 - maintainer: Alexander Diehl, Buffalo, NY, US (addiehl@buffalo.edu)
- Human disease ontology ([DOID](#))
 - used in:
 - * Diagnosis (`disease_diagnosis`, *Diagnosis*)
 - CURIE summary
 - * CURIE Prefix: DOID
 - * CURIE IRI Prefix: http://purl.obolibrary.org/obo/DOID_

- example AIRR use
 - * “disease_diagnosis.id” : “DOID:9538”
 - * “disease_diagnosis.label” : “multiple myeloma”
- default root node
 - * label: disease
 - * local ID: DOID:4
 - * path: disease
- license: CC0
- latest release (as of 2020-05-20): 2020-04-20
- repo: <https://github.com/DiseaseOntology/HumanDiseaseOntology>
- maintainer: Lynn Schriml, U Maryland, MD, US (lynn.schriml@gmail.com)
- notes: Features ICD cross-reference
- NCBI organismal taxonomy (NCBITAXON)
 - used in:
 - * Species (species, *Subject*)
 - * Cell species (cell_species, *Tissue and Cell Processing*)
 - CURIE summary
 - * CURIE Prefix: NCBITAXON
 - * CURIE IRI Prefixes: http://purl.obolibrary.org/obo/NCBITaxon_, <http://purl.bioontology.org/ontology/NCBITAXON/>
 - example AIRR use
 - * “species.id” : “NCBITAXON:9606”
 - * “species.label” : “Homo sapiens”
 - default root node
 - * label: Gnathostomata
 - * local ID: 7776
 - * path: cellular organisms/Eukaryota/Opisthokonta/Metazoa/Eumetazoa/Bilateria/Deuterostomia/Chordata/Craniata/Vertebrata/Gnathostomata
 - license: UMLS
 - latest release (as of 2020-05-20): 2020-04-18
 - repo: <https://github.com/obophenotype/ncbitaxon>
 - maintainer: NCBI (info@ncbi.nlm.nih.gov)
- NCI thesaurus (NCIT)
 - used in:
 - * Study type (study_type, *Study*)
 - CURIE summary
 - * CURIE Prefix: NCIT

- * CURIE IRI Prefixes: http://purl.obolibrary.org/obo/NCIT_, <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
- example AIRR use
 - * “study_type.id” : “NCIT:C15197”
 - * “study_type.label” : “Case-Control Study”
- default root node
 - * label: Study
 - * local ID: C63536
 - * path: Activity/Clinical or Research Activity/ Research Activity/Study
- license: Public domain, credit of NCI is requested
- repo: <https://github.com/NCI-Thesaurus/thesaurus-obo-edition>
- latest release (as of 2020-05-20): 2020-05-04
- maintainer: NCI (ncicbiitappssupport@mail.nih.gov)
- Units of measurement ontology (UO)
 - used in:
 - * Age unit (age_unit, *Subject*)
 - CURIE summary
 - * CURIE Prefix: UO
 - * CURIE IRI Prefix: http://purl.obolibrary.org/obo/UO_
 - example AIRR use
 - * “age_unit.id” : “UO:0000036”
 - * “age_unit.label” : “year”
 - default root node
 - * label: time unit
 - * local ID: UO_0000003
 - * path: unit/time unit
 - license: CC BY (per Github repo)
 - repo: <https://github.com/bio-ontology-research-group/unit-ontology>
 - latest release (as of 2020-05-20): 2020-05-18
 - maintainer: unknown
- Uber-anatomy ontology (Uberon)
 - used in:
 - * Tissue (tissue, *Sample*)
 - CURIE summary
 - * CURIE Prefix: UBERON
 - * CURIE IRI Prefix: http://purl.obolibrary.org/obo/UBERON_

- example AIRR use
 - * “tissue.id” : “UBERON:0002371”
 - * “tissue.label” : “bone marrow”
- default root node
 - * label: multicellular anatomical structure
 - * local ID: UBERON:0010000
 - * path: /BFO_0000002/BFO_0000004/anatomical entity/material anatomical entity/anatomical structure/multicellular anatomical structure
- license: CC BY
- repo: <https://github.com/obophenotype/uberon>
- latest release (as of 2020-05-20): 2019-11-22
- maintainer: Chris Mungall, LBL, CA, US (cjmungall@lbl.gov)

Computing with Ontologies

One of the key goals of using ontologies is to enable analysis tools to perform computation using the information in those ontologies. The AIRR Schema’s CURIEMap lists one or more *providers* for each CURIE *prefix* that can be used programmatically by analysis tools. Although the AIRR Schema lists multiple providers for each ontology, this section focuses on the use of the [EBI OLS provider’s OLS Web API](#) interface for querying ontologies.

If we consider the *DOID prefix* from the CURIEMap, the section below defines the use of the Human Disease Ontology ([DOID](#)) within the AIRR Standard:

```
DOID:
  type: ontology
  default:
    map: OBO
    provider: OLS
  map:
    OBO:
      iri_prefix: "http://purl.obolibrary.org/obo/DOID_"
```

We see that the default map for DOID is OBO map, and the OBO map’s *iri_prefix* is http://purl.obolibrary.org/obo/DOID_. Thus the mapping of the CURIE *DOID:9538* (the CURIE for disease “multiple myeloma”) will yield the resolved string http://purl.obolibrary.org/obo/DOID_9538. By the strictest of definitions, this is a valid IRI and should only be considered an identifier, but in this case this IRI is also a URL and can be used to look up the CURIE.

If we consider the default DOID provider in the CURIEMap, we see that it is OLS. Then, in the *InformationProvider* object of the AIRR Schema, under *provider* we see:

```
InformationProvider:
  provider:
    OLS:
      request:
        url: "https://www.ebi.ac.uk/ols/api/ontologies/{ontology_id}/terms?iri={iri}"
        response: application/json
```

And later we see that the parameters for OLS are:

```
parameter:
  CL:
    Ontobee:
      ontology_id: CL
    OLS:
      ontology_id: cl
  DOID:
    Ontobee:
      ontology_id: DOID
    OLS:
      ontology_id: doid
```

The above tells us that we can use the OLS provider to look up ontology terms. The {iri} component of the url string tells us that we need to use the resolved IRI and the {ontology_id} component tells us that we need to replace the ontology_id parameter in the URL with the DOID OLS parameter in the specification, which is the string doid. Thus the fully resolved URL to query for the CURIE DOID:9538 would be:

```
https://www.ebi.ac.uk/ols/api/ontologies/doid/terms?iri=http://purl.obolibrary.org/obo/
↪DOID_9538
```

Again, referring to the OLS provider we see that we can expect an application/json response to the above query, and indeed the response we receive from the above starts with a JSON object as follows.

```
{
  "_embedded" : {
    "terms" : [ {
      "iri" : "http://purl.obolibrary.org/obo/DOID_9538",
      "label" : "multiple myeloma",
      "description" : [ "A myeloid neoplasm that is located in the plasma cells in bone_
↪marrow." ],
      "annotation" : {
        "comment" : [ "OMIM mapping confirmed by DO. [SN]." ],
        "database_cross_reference" : [ "ICD10CM:C90.0", "MESH:D009101", "ICD9CM:203.0",
↪"GARD:7108", "NCI:C3242", "OMIM:254500", "ORDO:29073", "EFO:0001378", "SNOMEDCT_US_
↪2020_09_01:94705007", "UMLS_CUI:C0026764" ],
        "has_obo_namespace" : [ "disease_ontology" ],
        "id" : [ "DOID:9538" ]
      },
      "synonyms" : [ "plasma cell myeloma" ],
      "ontology_name" : "doid",
      "ontology_prefix" : "DOID",
      "ontology_iri" : "http://purl.obolibrary.org/obo/doid.owl",
      "is_obsolete" : false,
      "term_replaced_by" : null,
      "is_defining_ontology" : true,
      "has_children" : true,
      "is_root" : false,
      "short_form" : "DOID_9538",
      "obo_id" : "DOID:9538",
      [Content edited because of length]
```

In this response, you can see that the Ontology object that we requested has a label field that contains the value multiple myeloma and that the id field has a value of DOID:9538.

It is beyond the scope of this document to describe in detail the JSON structure of each of the providers, but this information can be discovered through the `provider` web sites. It should be noted that all Ontology objects in the AIRR specification have the OLS as a `provider` and therefore the method above can be used for any of the ontologies in the AIRR specification. Please see the [OLS Web API](#) documentation for details of the JSON response for the OLS provider.

Sprint Reports

OntoVoc Report - Sprint 11/2018

Objectives

The objectives of this first sprint in November 2018 were to:

1. define criteria for suitable ontologies
2. identify ontologies for five fields/keywords of the MiAIRR data standard and
3. assess technical aspects of ontology integration into databases

General Considerations

The Team initially discussed an approach where only vocabularies (i.e. lists of terms) and not ontologies (i.e. many terms connected by predicates) would have been defined. These vocabularies would have been derived from ontologies, but this process would not necessarily have been reversible. The notion at this time point was, that such an approach would allow to solve a number of problems like combining multiple sources and removing duplicated leaves. However, after some discussions this approach was effectively abandoned for a number of reasons:

- It would discard the UID for an entity. As the UID (in contrast to the name string) is guaranteed to be stable and unique, it facilitates updates, linking and information representation, all of which would otherwise be lost.
- In general, it will be more sustainable to work with the maintainers of an existing ontology to include entities/terms, than just dumping their terms into a list and adding new ones.
- Well-designed ontologies will not contain duplicated entities, although they might appear to do so in a simple browsers (i.e. this is an artifact of representation). Ontologies that actually do contain duplicates are excluded by *criterion 2*.

Criteria for Ontologies

Criteria

Ontologies used within AIRR standards

1. MUST¹ cover the majority of the required terms, but complete coverage is OPTIONAL
2. MUST have a structure that is scientifically correct and logically coherent
3. MUST NOT feature complexity that makes it hard to use for queries and data representation
4. SHOULD already be widely adopted
5. MUST be actively maintained
6. MUST be available under a free license

¹ See the “Glossary” section on how to interpret term written in all-caps.

Comments on criteria:

- ad (1): For most fields it will be difficult to find complete and accurate ontologies. Therefore picking the best available ontology and working with its maintainers to include missing terms is expected to be the most sustainable approach.
- ad (5): This requirement follows from (1), as there needs to be a way for term requests.
- ad (6): A number of ontologies need to be licensed from their respective copyright holders. This results in potential barriers for implementation and distribution of such ontologies. Therefore only ontologies available under a free license are considered suitable for AIRR-compliant databases. The list of suitable licenses is not final, but includes: [CC0](#) and [CC BY](#).

Selected Ontologies

(designations are MiAIRR field names and DataRep keywords)

Completed

- Species (organism)
 - [NCBITAXON](#)
 - license: [UMLS](#)²
 - latest release: 2018-07-06
 - maintainer: NCBI (info@ncbi.nlm.nih.gov)
- Diagnosis (disease_diagnosis)
 - [DOID](#)
 - root node
 - * name: disease
 - * ID: DOID:4
 - * path: /disease
 - License: [CC BY](#)
 - latest release: 2018-03-02
 - maintainer: Lynn Schriml, U Maryland, MD, US (lynn.schriml@gmail.com)
 - notes: Features ICD cross-reference
- Cell subset (cell_subset)
 - [CellOntology](#)
 - license: [CC BY](#)
 - latest release: 2018-07-11
 - maintainer: Alexander Diehl, Buffalo, NY, US (addiehl@buffalo.edu)
- Tissue (tissue)
 - [Uberon](#)

² Will require further review the [UMLS Metathesaurus License](#) is not a free license, however it needs to be clarified how much of it relates to the work (i.e. the taxonomy itself) and how much to the service.

- root node
 - * name: multicellular anatomical structure
 - * ID: UBERON:0010000
 - * path: /BFO_00000002/BFO_00000004/anatomical entity/material anatomical entity/anatomical structure/multicellular anatomical structure
- License: [CC BY](#)
- latest release: 2018-10-15
- Maintainer: Chris Mungall, LBL, CA, US (cjmungall@lbl.gov)

Under evaluation

- Strain name (strain_name)
 - Suggested ontologies:
 - * JAX
 - * IEDB
 - Issues:
 - * Nomenclature
 - * one ontology is not enough

Technical aspects

- Repositories:
 - UID assigned by ontologies are guaranteed to be unique and permanent³.
 - A repository MAY use internal identifiers that are distinct from UIDs. However, to be AIRR-compliant it MUST be able to map UIDs to its identifiers.
 - Points of “AIRR compliance” would typically be:
 - * When data is extracted from the repository through a Query API (CRWG)
 - * When data is extracted from the repository into a file format (DataRep)
- Integration of ontologies into repositories:
 - There are two main ontology providers offering a REST API and all the ontologies listed above:
 - * [NCBO Bioportal](#)
 - * [OLS ontology](#)
 - NCBO can apparently be slow and sometimes not that stable, while OLS seems to be more stable and potentially has a better long-term support.
 - Remote ontology services tend to be slow and create external dependencies. On the other hand, while local hosting of an ontology is possible (and partially supported by NCBO and OLS), it requires non-negligible resources. The Team’s current assumption is that queries to remote ontology services can be substantially

³ This has more recently (early 2020) been called in question and will be revisited during the next sprint. Note that the uncertainty revolves around the question what exactly constitutes a UID, rather than the question whether a UID is permanent and unique.

accelerated if only the relevant section of a respective ontology is queried. Therefore a local service would not be necessary.

- Repositories should store both the IDs and the values in their database. This way, they do not have to query the ontology in a scenario where human-readable output is required. In the case of changes, most ontologies try to follow the practice of not changing a term value but instead create a new term with the new value and a new ID, and deprecating the old term. Therefore term deprecation needs to be handled by the repository.
 - Like for the databases, also the API should be able to handle both IDs and values as query input and return both during a query.
 - The user interface (UI) should offer an ontology-backed autocomplete. NCBO provides some JavaScript code to use. The UI must not offer deprecated terms. To allow entry of terms not present in the ontology, data can be prefixed with some text that will allow the data validation to proceed (e.g., if an entry starts with “other -” the UI will not autocomplete/validate). Later, i.e. when the term has been created, the data will be updated.
- Note that the complete IEDB can be [downloaded as SQL dump](#), it is licensed under [CC BY](#). At a first glance, the main overlap seems to be with `organism`, `strain_name` and to a smaller extent `disease_diagnosis`. However, sample information like `cell_subset` and `tissue` seems to be largely absent from IEDB, so it could currently not be the one-stop solution for AIRR.

Footnotes

Appendix

Glossary

- **MUST / REQUIRED:** Indicates that an element or action is necessary to conform to the standard.
- **SHOULD / RECOMMENDED:** Indicates that an element or action is considered to be best practice by AIRR, but not necessary to conform to the standard.
- **MAY / OPTIONAL:** Indicates that it is at the discretion of the user to use an element or perform an action.
- **MUST NOT / FORBIDDEN:** Indicates that an element or action will be in conflict with the standard.

OntoVoc Report - Sprint 04/2020

Objectives

The objectives of this second sprint in April 2020 were to:

1. revisit general policies around ontologies used in the AIRR schema
2. identify two new ontologies for several fields of the AIRR schema
3. solve technical questions regarding IDs and providers

General Policies

The OntoVoc team revisited the criteria for ontologies used in the AIRR schema that it *defined in the 11/2018 sprint*. While they are still considered to be valid, the team felt that a more detailed guidance could be useful in the process of selecting ontologies for new fields. It therefore evaluated the [OBO Foundry Principles](#), which partially re-iterate some of the existing criteria (e.g., *Openness* and *Maintenance*), but also provide additional recommendations, e.g., the presence of textual definitions, clear scope and a common format, which were considered to be valuable additions to the existing guidelines. The team therefore decided to endorse the OBO Foundry Principles, as RECOMMENDED (but NOT REQUIRED) criteria. It should be noted, that this does not make any statement regarding the use of OBO vs. non-OBO ontologies.

Decisions on Pending Items of Sprint 11/2018

A number of decisions on draft and legacy ontologies as well as root nodes was not officially passed during the last sprint. The team thus revisited and confirmed the following decisions:

- Use of NCIT for `study_type`, top node `Study` (NCIT:C63536).
- Use of UO for `age_unit`, top node `time unit` (UO:0000003).
- Use of `Gnathostomata` (NCBITAXON:7776) as top node for NCBITAXON when used for fields encoding a host species.
- Use of `lymphocyte` (CL:0000542) as top node for CL when used for `cell_subset`.

New Ontologies

Mouse strain

Background

Mouse strain names follow a very [elaborate nomenclature](#) that is capable of describing the genetic background, breeding history and introduced mutation in a detailed manner. However, this nomenclature is rarely used correctly (if at all), which creates uncertainty about the identity of strains used in experimental studies. Therefore an ontology or vocabulary compliant to this nomenclature would be of tremendous help for consistent annotation.

An ontology for the `strain_name` field was already on the list for the last sprint, however it was not possible to identify a single ontology that would contain comprehensive information about strains from multiple species. This situation created a problem that could not be resolved then. In the meantime, the concept of “extensions” has been introduced to the AIRR schema, which create an additional layer of fields (and associated ontologies) on top of a core schema. As these extensions can be made conditional on the value of fields within the core schema, it has now become possible to have multiple extensions defining the `strain_name` field, but for different species and therefore with distinct species-specific ontologies.

Having addressed this issue, the other key problem that remains is the absence of an actual ontology for mouse strains, while a [rat strain ontology](#) exists. Therefore in a first step it is necessary to identify resources that you at least serve as a provider for vocabularies. The two potential candidates that were identified are:

- MGI: The Mouse Genome Informatics database hosted at JAX aims to be comprehensive in regard to all mouse strains that have been published in the literature.
- IEDB: The Immune Epitope Database already ran into the problem of a missing mouse ontology and therefore decided to build up their own reference focused on immunologically relevant strains, as part of their Ontie database.

Once it is clear which of the resources could be used, it will be necessary to approach the current maintainers regarding their willingness to convert the data into an actual ontology (the RS could serve as a template for this). As this will take longer than just a couple of weeks, the second step is out-of-scope for this sprint.

Evaluation

- MGI: The database can be downloaded as a dump, however the licensing conditions are unclear. It contains a total of 60k entries of which 3.2k inbred and 13.8k are congenic strains. The majority of the remaining entries are coisogenic strains, most of them from large- scale gene KO projects.
- IEDB: Database dumps can also be downloaded and are freely available under CC-BY 4.0. It covers over a thousand mouse strains and contains additional information on the genetic background of a strain.

Next steps

- Get in touch with JAX (pending)

Geolocation

There are several (planned) extensions to the AIRR metadata standard that will provide geospatial metadata. Country-level information is typically assumed to be privacy-preserving and easy to operationalize. Therefore, while clearly only capturing some aspects of genetic ancestry, it might serve as a proxy for concepts of “race” and “ethnicity” that are rather ill-defined.

Potential candidate vocabularies/ontologies:

- [ISO3166-1 alpha-2](#): Two-letter code, some ambiguity but well known from ccTLDs.
- [ISO3166-1 alpha-3](#): Three-letter code, less ambiguity than alpha-2.
- UN Stats Division code (currently [M49](#)): Numerical code, not human-readable, maps to ISO3166-1 alpha-3.
- [Gazetter \(GAZ\)](#)
 - Contains 2nd (state) and 3rd (county) level information.
 - Not linked to any actual coordinates
 - ISO3166-1 annotation is incomplete and lacks e.g. for Germany and Switzerland.
 - Does not support German Umlauts. *Äbsölütely* unacceptable, as these are not just diacritical marks (i.e. “Münster” and “Munster” are two different cities).
- [HANCESTRO](#):
 - Seems to be complete, but does not provide ISO3166 codes.
 - Ontology could also be used for other fields relating to genetic ancestry.
 - Links to DBpedia, currently unclear whether it is also populated from there
 - *country* node has pan-240 leaves (surplus seems due to oversea territories), cross-referencing to GAZ (s/a)
- Various pathogen-related repositories:
 - VectorBase ([VBGEO](#)): see link and choose “GADM/VBGEO PlaceNames”
 - Viral Pathogen Resource ([ViPR](#)):
 - * Uses v1.3 of the [GSCID/BRC Project and Sample Application Standard](#).

- * **GSCID/BRC Core Sample** defines four fields for “Collection Location”:
 - “Latitude” (CS11) and “Longitude” (CS12) in **ISO 6709** format
 - “Location” (CS13), using GAZ as controlled vocabulary
 - “Country” (CS14) as by ISO3166-1 (alpha-2).
- Influenza Research Database (**IRD**): Flu-focused version of ViPR, also uses GSCID/BRC Project and Sample Application Standard v1.3.
- Pathosystems Resource Integration Center (**Patric**): Focused on bacterial infectious diseases. Uses an “Isolation Country” field in their “Genome” table, format seems to be full text.

Rejected candidates:

- HL7: own ontology deprecated, now recommends ISO 3166-1 alpha-3 set.
- NCIT: Incomplete, only contains pan-90 entities
- SNOMED: Licensing issues
- GADM data: Good quality and resolution, but not an ontology in itself. Also not under a free license, does not allow redistribution or commercial use.

Evaluation

Given the number of options, there is no obvious candidate to pick. Therefore the team decided to define clear use cases and then evaluate each options against them. However, due to time limitation, we did not really get into this, will have to follow up in the next sprint. The use cases so far were:

- Annotate country of birth / of sampling [REQUIRED]
- Encode higher resolution than country level if legally permitted and scientifically meaningful [RECOMMENDED].
- Linking to geo-spatial coordinates [OPTIONAL]

Technical Questions

Background and Problem

Some nomenclature first: The nodes in an ontology graph are typically either *concepts* (e.g., capital) or *instances* thereof (e.g., Paris). These nodes have *local IDs* (often numbers), which are unique within an ontology. They also typically have *labels*, which is the human- readable name of the node. Nodes can have additional *attributes* (e.g., “population count”) and are connected to other nodes by *relations* (e.g. “is-a”, “superset-of”), which create the edges of the graph.

The complete ontology is usually represented in an XML or OWL file. However, we are looking for a *provider*, i.e. a service that facilitates queries of an ontology via web and/or an API-based interface. Upon querying with a unique ID, it is expected that a *provider* will be able to return the record of a node, which should contain all attributes and relations. Furthermore a *provider* might allow set- and graph-based queries (e.g., is A a complete subset of B; what is the last common ancestor of A and B). Finally a *provider* can offer lookup services, i.e., identify the corresponding *concept* or *instance* in another ontology. Until now we have mainly looked at three providers: **Ontobee**, **OLS** and **BioPortal**. While they all provide similar basic services, it should be noted that some biomedical databases and repositories are, by convention, restricted to use certain *providers*.

As stated above, each node has a *local ID*. To avoid conflicts between the *local IDs* of multiple ontologies, *providers* and ontology collections (e.g., OBO Foundry) use a **namespace**, i.e., some abbreviation for the ontology that is prefixed to the *local ID*. However, as there no common standard how to create these prefixes, this system is only unambiguous

and collision-safe within a single *provider*. To resolve this issue, ontologies often use International Resource Identifiers (IRI, [RFC3987]). While IRIs look like HTTP URLs, they should primarily be considered as permanent and globally unique identifiers, which might resolve to the node's record via DNS/HTTP, but this is optional. In addition, potential intermediate URLs generated in the DNS/HTTP resolving process must be considered internal and therefore should not be used by third parties. Finally, it needs be noted that IRIs should to be considered case-sensitive, especially when used as identifiers (per [RFC3987], Section 5.3.2.1, which only excludes the schema and host (authority) component for case-sensitivity).

While many ontologies already define an entities IRI on the level of the ontology, there are some that do not. For such ontologies, IRIs are then assigned by the provider. The most notable example for this are the UMLS ontologies like the NCBI Taxonomy. This leads to the situation that a single node in an ontology, stored by two providers can have different IRIs. Therefore, a concept from NCBI Taxonomy, e.g., the duck-billed platypus (label: *Ornithorhynchus anatinus*, local ID: 9258) has the IRI `http://purl.obolibrary.org/obo/NCBITaxon_9258` in Ontobee and the IRI `http://purl.bioontology.org/ontology/NCBITAXON/9258` in BioPortal. In addition, other providers might choose to use one of these IRIs too, although it will never resolve to their system via DNS/HTTP (e.g., OLS uses the Ontobee IRIs).

For the AIRR Community, this creates the challenge that we want to be able to have unambiguous identifiers, without requiring any specific provider.

Proposed solution

Compact URIs (**CURIEs**) are a standardized way to abbreviate IRIs, which includes URIs as a subset. They were originally conceived to simplify the handling of attributes, e.g. in XML or SPARQL, by making them more compact and readable. CURIEs are e.g. used by IEDB databases to reduce redundancies (mainly in the leading part of IRIs).

A typical CURIE would, e.g., look like `NCBITAXON:9258`. In this case, `NCBITAXON` is the *prefix*, a custom string that will be replaced by a repository-defined IRI component (e.g., `http://purl.obolibrary.org/obo/NCBITaxon_`). Note that there is no connection between `NCBITAXON` in the CURIE and `NCBITaxon` in the IRI, the former one is just a placeholder.

This resolves the issue of different *providers* usings different IRIs with distinct formatting rules (as described above). As the choice of the *provider* is independent for each ontology, it allows greater flexibility for the repositories, as they do not need a single *provider* that needs be able to resolve all terms. Similarly, different repositories can use the same ontology, but use different *providers*. Note that this would not require changes to the data, as the data would only contains CURIEs, not the (provider-specific) IRIs.

The AIRR schema will provide a list of AIRR approved CURIE *prefixes* along with a list of at least one IRI *prefix* (i.e., replacement string) for each them. This list serves two purposes:

1. It provides a controlled namespace for CURIE *prefixes* used in the AIRR schema. For now, custom additions to or replacements of these *prefixes* in the schema are prohibited. This does not affect the ability of repositories to use such custom prefixes internally.
2. It simplifies resolution of CURIEs by non-repositories. The lists of IRI *prefixes* for each CURIE *prefix* should not be considered to be exhaustive. However, when using custom IRI *prefixes*, it must be ensured that they refer to the same ontology as the provider *prefixes*.

It should be explicitly noted that the IRI *prefix* list should not be interpreted as any kind of recommendation for certain *providers*. It is left up to users to decide how to resolve the resulting IRIs, e.g., via DNS/HTTP (if possible) or by using a *provider* of their choice.

Modifications to the AIRR schema

All changes to the AIRR schema that would be based on the sprint can currently be reviewed on Github in Pull Request #385. These changes are intended to be included into the next major release.

Clarifications

- Root nodes are specific to individual fields, not to an ontology. Therefore, NCBITAXON will use a root node of “Gnathostomata” for the annotations of the host species, but this would not be useful, e.g., if it would be used to annotate pathogenic organisms, which will require a top node at the apex of the hierarchy.
- The `labels` (previous: `values`) that are provided in the schema for ontology-based fields, should be considered an addition for convenience and not as being authoritative. Repositories or applications can choose to link synonyms to given concepts (e.g., “human” for “*Homo sapiens*”) to simply search queries. Repositories further can provide such a synonym in the `label` field upon exporting data. However, repositories importing data should verify the correctness of `labels` that do not match the one provided by the ontology. Importing repositories must not be expected to allow for queries of `labels` other than those present in the ontology.

Annotation guidance

Note that this section is only a parking lot, the respective text will be moved into the AIRR Docs in the final version.

- Cells that come from Ficoll gradients should not be annotated as PBMCs as this is a sister node of `lymphocyte`. For the other sampling related fields, in nearly all cases venous blood (UBERON:0013756) will be the correct `tissue` and it should be used in the case of `sample_type:peripheral venous puncture`. However, if the mode of sampling is not specified, `blood` (UBERON:0000178) should be used instead. Also see <https://github.com/airr-community/airr-standards/issues/242>

OntoVoc Report - Sprint 04/2021

Objectives

The objectives of this third sprint in April 2021 were to:

- Revisit General Policies
- CURIE for PIDs
- Assess whether the use of multiple ontologies for single fields should be allowed
- Use of IDs and labels
- Provide guidance on `cell_population` annotation

General Policies

The OntoVoc team revisited the criteria for ontologies used in the AIRR schema that it *defined in the 11/2018 sprint* and that were expanded in the 04/2020 sprint to include the [OBO Foundry Principles](#).

CURIEs for PIDs

This is not directly related to ontologies, but as it does touch upon the use of CURIEs in the AIRR Schema, it does fall into the scope of OntoVoc. Main question is whether and how we can use, e.g., ORCID in Repos (see [#476](#)). The top-level conclusion on this topic is that CURIE can be used by repositories for non-ontology PIDs, but that the exact specification should be addressed by ComRepo WG. However, it should be noted that there are two functions a CURIE can serve:

- A CURIE can be used to create a unique ID. In this case the prefix serves as/provides a namespace.
- An expanded CURIE (i.e. URI) can be used to retrieve data, i.e., using it as a PID in the strict sense.

As there is no standard what type of information is provided by an URI (in the second case) there should be hints in the schema, e.g., on human vs. machine-readable data.

Use of multiple ontologies for a single field

While it sometime would be preferable to have multiple ontologies available for a given field, the group came to the concluding that this is would more likely than not reduce the interoperability of the data:

- Overlaps between two accepted ontologies would be problematic as this creates an ambiguity about which term to use. This ambiguity jeopardizes the linkage possibilities that ontologies should provide.
- Although OLS has a term-mapping service, this does not ensure that the same term in two different ontologies represents the same concept. Also the logical structure between ontologies might differ, complicating things even further.
- There might however be very special and restricted use cases in which ontologies are orthogonal to each other and therefore the union of both could be used. This would in addition require a common structure that should exist for OBO ontologies. However, until there is a clear use case for this that cannot be solved in any other way, we will not allow multiple ontologies per term.

Use of IDs and labels

It should be reiterated that:

- The `id` is the authoritative and required part of the information, the `label` is at the discretion of the repository and is provided to the user as complementary information (so that the user does not need to resolve the `id`).
- Repositories **MUST** support searches for the official label term, they **SHOULD** support ontology synonyms and they **MAY** support their own synonyms (termed “alias”).
- Data returned by a repository **MUST** provide the official term as `label` and **MAY** provide or not provide a set of synonyms/aliases in the novel `synonyms` array.
- The behavior of a `synonyms` field can be altered by the user query.
- Some PID types might not have a meaningful label. Therefore `label` **MAY** be `NULL`.

Provide guidance on cell_population annotation

Added to documentation

2.3.6 Schema Release Notes

Version 1.5.0: August 29, 2023

Version 1.5 schema release.

General Schema Changes:

1. Fixed synchronization errors between the OpenAPI v2 and v3 versions of the AIRR Schema (airr-schema.yaml and airr-schema-openapi3.yaml).
2. Set the default value of `x-airr.miarr` attributes to `defined`.
3. Converted all `x-airr.format` attribute values to `snake_case`, which specifically impacts any instance of `controlled vocabulary` or `physical quantity`.
4. Corrected numerous instances of missing `x-airr.miarr` and `x-airr.identifier` attributes.
5. Replaced `x-airr.adc-api-optional` attribute with `x-airr.adc-query-support` in multiple fields.
6. Added “IGI” as a valid value to the locus enum fields in multiple schema.
7. Added `null` as a valid value to all nullable enum fields.
8. Removed discriminator: AIRR from all object definitions.

Germline and Genotype Schema:

1. Clarified the descriptions of multiple fields in the Germline and Genotype schema.
2. Modified `x-airr: nullable` and `x-airr: identifier` values on multiple fields in the Germline and Genotype schema.
3. Removed the `alignment` field and added the `unaligned_sequence`, `aligned_sequences`, and `alignment_labels` fields to the `SequenceDelineationV` object.
4. Converted the enum values in the `inference_type` field of `AlleleDescription` to `snake_case`.
5. Added the `allele_similarity_cluster_designation` and `allele_similarity_cluster_member_id` fields to `AlleleDescription`.
6. Moved the nested objects `DocumentedAllele`, `UndocumentedAllele`, and `DeletedGenes` out of `Genotype` and defined them as top-level objects referenced by the `documented_alleles`, `undocumented_alleles`, and `deleted_genes` fields, respectively.
7. Moved the nested object `MHCAllele` out of `MHCGenotype` and defined it as a top-level object referenced by the `mhc_alleles` field.

Single-cell Schema:

1. Added the `property_type` field to the `CellExpression` object.
2. Moved the nested `ReceptorReactivity` object out of `Receptor` and defined it as a top-level object referenced by the `reactivity_measurements` field.

Subject Schema:

1. Removed the nested references to `GenotypeSet` and `MHCGenotypeSet` in the `genotype` field and modified the definition to point to a top-level `SubjectGenotype` object defining these references.

DataProcessing Schema:

1. Clarified the description of `quality_thresholds` to indicate that quality filtering is not mandatory.

Version 1.4.1: August 27, 2022

Version 1.4 schema release.

New General Purpose Schema:

1. Introduced the experimental `DataFile` object, which defines a JSON file holding Repertoire metadata, data processing analysis objects, or any object in the AIRR Data Model.
2. Introduced the experimental `RepertoireGroup` Schema for describing collections of repertoires to be analyzed together.
3. Introduced the experimental `InfoObject` Schema, which provides information about data and ADC API responses.
4. Introduced the experimental `TimePoint` Schema for defining the time point at which an observation or other action was performed.

New Germline and Genotype Schema:

The following experimental schema were introduced to support storage of VDJ germline reference sequences, VDJ genotypes, and MHC genotypes:

1. `GermlineSet`: Defines a collection of `AlleleDescriptions` from the same strain or species.
2. `AlleleDescription`: Details of a putative or confirmed Ig receptor gene/allele inferred from one or more observations.
3. `RearrangedSequence`: Details of a directly observed rearranged sequence or an inference from rearranged sequences contributing support for a gene or allele.
4. `UnrearrangedSequence`: Details of an unrearranged sequence contributing support for a gene or allele.
5. `SequenceDelineationV`: Delineation of a V-gene in a particular system.
6. `GenotypeSet`: Defines a collection a VDJ genotypes for a given subject.
7. `Genotype`: Enumerates the alleles and gene deletions inferred in a single subject for a single locus.
8. `MHCGenotypeSet`: Defines a collection of MHC genotypes for a given subject.
9. `MHCGenotype`: Details the genotype of major histocompatibility complex (MHC) class I, class II and non-classical loci.
10. `Acknowledgement`: Defines contributors to the germline or genotype description.

New Single-cell Schema:

The following experimental schema were introduced to improve support for single-cell data and extend the `Cell` schema.

1. `CellExpression`: Defines a container to store single-cell expression level measurements.
2. `Receptor`: Describes a complete receptor protein sequence and its reactivity.

Rearrangement Schema:

1. Added the optional fields `v_frameshift`, `j_frameshift`, `d_frame` and `d2_frame` defining annotations related to alignment reading frames.
2. Added the optional field `umi_count` to represent the count of distinct UMIs for a sequence.

3. Modified the definition of `duplicate_count` to remove ambiguity with the new `umi_count` field in a single-cell context. There is now a distinction between duplicate observed sequences (`duplicate_count`) and UMIs (`umi_count`).
4. The optional `quality` and `quality_alignment` alignment fields were added to store Phred quality scores for base calls in the `sequence` and `sequence_alignment` fields, respectively.
5. The following optional fields were added to denote constant region (`c_call`) alignment positions: `c_sequence_start`, `c_sequence_end`, `c_germline_start`, `c_germline_end`, `c_alignment_start`, `c_alignment_end`.

Study Schema:

1. Added the optional fields `study_contact` to store contact information for the primary study contact.
2. Modified the enumerated values supported by `keywords_study` to the following set: `contains_ig`, `contains_tr`, `contains_paired_chain`, `contains_schema_rearrangement`, `contains_schema_clone`, `contains_schema_cell`, `contains_schema_receptor`
3. Added the optional fields `adc_publish_date` and `adc_update_data` that timestamp AIRR Data Commons initial publication and last update, respectively.

Subject Schema:

1. Added the optional `genotype` field linking to the new `GenotypeSet` and `MHCGenotypeSet` objects.

Sample Schema:

1. Added the required field `collection_time_point_relative_unit` defining the units for the sample collection timestamp.
2. Modified the type of the field `collection_time_point_relative` from a string to a number defined in combination with the new unit ontology field `collection_time_point_relative_unit`.

NucleicAcidProcessing Schema:

1. Added the required field `template_amount_unit` defining the units for the input template quantification.
2. Modified the type of the `template_amount` field from a string to a number defined in the combination with the new unit ontology field `template_amount_unit`.

Clone Schema:

1. Added the optional `clone_count` field to specify absolute count of clonal members.
2. Added the optional `umi_count` field to specify the total UMI count of all clonal members.

Cell Schema:

1. Removed the field `expression_tabular` whose functionality has been replaced by the new `CellExpression` schema.

Version 1.3.1: October 13, 2020

Version 1.3 documentation patch release.

Alignment Schema:

1. Added the deprecation tags for `rearrangement_id`, which were accidentally left out of the v1.3.0 release.

Version 1.3.0: May 28, 2020

Version 1.3 schema release.

New Schema:

1. Introduced the `Repertoire` Schema for describing study meta data.
2. Introduced the `PCRTarget` Schema for describing primer target locations.
3. Introduced the `SampleProcessing` Schema for describing experimental processing steps for a sample.
4. Replaced the `SoftwareProcessing` schema with the `DataProcessing` schema.
5. Introduced experimental schema for clonal clusters, lineage trees, tree nodes, and cells as `Clone`, `Tree`, `Node`, and `Cell` objects, respectively.

General Updates:

1. Added multiple additional attributes to a large number of schema properties as AIRR extension attributes in the `x-airr` field. The new `Attributes` object contains definitions for these `x-airr` field attributes.
2. Added the top level `required` property to all relevant schema objects.
3. Added the `title` attribute containing the short, descriptive name to all relevant schema object fields.
4. Added an `example` attribute containing an example data value to multiple schema object fields.

AIRR Data Commons API:

1. Added OpenAPI V2 specification (`specs/adc-api.yaml`) for AIRR Data Commons API major version 1.

Ontology Support:

1. Added `Ontology` and `CURIResolution` objects to support ontologies.
2. Added vocabularies/ontologies as JSON string for: Cell subset, Target substrate, Library generation method, Complete sequences, Physical linkage of different loci.

Rearrangement Schema:

1. Added the `complete_vdj` field to annotate whether a V(D)J alignment was full length.
2. Added the `junction_length_aa` field defining the length of the junction amino acid sequence.
3. Added the `repertoire_id`, `sample_processing_id`, and `data_processing_id` fields to serve as linkers to the appropriate metadata objects.
4. Added a controlled vocabulary to the `locus` field: IGH, IGI, IGK, IGL, TRA, TRB, TRD, TRG.
5. Deprecated the `rearrangement_set_id` and `germline_database` fields.
6. Deprecated `rearrangement_id` field and made the `sequence_id` field be the primary unique identifier for a rearrangement record, both in files and data repositories.
7. Added support secondary D gene rearrangement through the additional fields: `d2_call`, `d2_score`, `d2_identity`, `d2_support`, `d2_cigar`, `np3`, `np3_aa`, `np3_length`, `n3_length`, `p5d2_length`, `p3d2_length`, `d2_sequence_start`, `d2_sequence_end`, `d2_germline_start`, `d2_germline_end`, `d2_alignment_start`, `d2_alignment_end`, `d2_sequence_alignment`, `d2_sequence_alignment_aa`, `d2_germline_alignment`, `d2_germline_alignment_aa`.
8. Updated field definitions with more concise V(D)J call descriptions.

Alignment Schema:

1. Deprecated the `rearrangement_set_id` and `germline_database` fields.
2. Added the `data_processing_id` field.

Study Schema:

1. Added the `study_type` field containing an ontology defined term for the study design.

Subject Schema:

1. Deprecated the `organism` field in favor of the new `species` field.
2. Deprecated the `age` field.
3. Introduced age ranges: `age_min`, `age_max`, and `age_unit`.

Diagnosis Schema:

1. Changed the type of the `disease_diagnosis` field from `string` to `Ontology`.

Sample Schema:

1. Changed the type of the `tissue` field from `string` to `Ontology`.

CellProcessing Schema:

1. Changed the type of the `cell_subset` field from `string` to `Ontology`.
2. Introduced the `cell_species` field which denotes the species from which the analyzed cells originate.

NucleicAcidProcessing Schema:

1. Defined the `template_class` field as type `string`.
2. Added a controlled vocabulary the `library_generation_method` field.
3. Changed the controlled vocabulary terms of `complete_sequences`. Replacing `complete` & `untemplated` with `complete+untemplated` and adding `mixed`.
4. Added the `pcr_target` field referencing the new `PCRTarget` schema object.

SequencingRun Schema:

1. Added the `sequencing_run_id` field which serves as the object identifier field.
2. Added the `sequencing_files` field which links to the `RawSequenceData` schema objects defining the raw read data.

RawSequenceData Schema:

1. Added the `file_type` field defining the sequence file type. This field is a controlled vocabulary restricted to: `fasta`, `fastq`.
2. Added the `paired_read_length` field defining mate-pair read lengths.
3. Defined the `read_direction` and `paired_read_direction` fields as type `string`.

DataProcessing Schema:

1. Replaces the `SoftwareProcessing` object.
2. Added `data_processing_id`, `primary_annotation`, `data_processing_files`, `germline_database` and `analysis_provenance_id` fields.

Version 1.2.1: Oct 5, 2018

Minor patch release.

1. Schema gene vs segment terminology corrections
2. Added Info object
3. Updated cell_subset URL in AIRR schema

Version 1.2.0: Aug 18, 2018

Peer reviewed released of the Rearrangement schema.

1. Definition change for the coordinate fields of the Rearrangement and Alignment schema. Coordinates are now defined as 1-based closed intervals, instead of 0-based half-open intervals (as previously defined in v1.1 of the schema).
2. Removed foreign study_id fields
3. Introduced keywords_study field

Version 1.1.0: May 3, 2018

Initial public released of the Rearrangement and Alignment schemas.

1. Added required and nullable constrains to AIRR schema.
2. Schema definitions for MiAIRR attributes and ontology.
3. Introduction of an x-airr object indicating if field is required by MiAIRR.
4. Rename rearrangement_set_id to data_processing_id.
5. Rename study_description to study_type.
6. Added physical_quantity format.
7. Raw sequencing files into separate schema object.
8. Rename Attributes object.
9. Added primary_annotation and repertoire_id.
10. Added diagnosis to repertoire object.
11. Added ontology for organism.
12. Added more detailed specification of sequencing_run, repertoire and rearrangement.
13. Added repertoire schema.
14. Rename definitions.yaml to airr-schema.yaml.
15. Removed c_call, c_score and c_cigar from required as this is not typical reference aligner output.
16. Renamed vdj_score, vdj_identity, vdj_evalue, and vdj_cigar to score, identity, evalue, and cigar.
17. Added missing c_identity and c_evalue fields to Rearrangement spec.
18. Swapped order of N and S operators in CIGAR string.
19. Some description clean up for consistency in Rearrangement spec.

20. Remove repeated objects in `definitions.yaml`.
21. Added `Alignment` object to `definitions.yaml`.
22. Updated MiARR format consistency check TSV with junction change.
23. Changed definition from functional to productive.

Version 1.0.1: Jan 9, 2018

MiAIRR v1 official release and initial draft of Rearrangement and Alignment schemas.

2.4 Data Submission and Query

2.4.1 Data Submission Guides for AIRR-seq studies

There are multiple data repositories that accept submission of AIRR-seq datasets. Each provides different capabilities but all comply with the MiAIRR standard.

CAIRR Pipeline

Introduction: The CAIRR pipeline for submitting standards-compliant B and T cell receptor repertoire sequencing studies to the NCBI

AIRR sequencing (AIRR-seq) has tremendous potential to understand the dynamics of the immune repertoire in vaccinology, infectious disease, autoimmunity, and cancer biology. The adaptation of high-throughput sequencing (HTS) for AIRR (Adaptive Immune Receptor Repertoire) studies has made possible to characterize the AIRR at unprecedented depth and the outcome of such sequencing produces big data. Effective sharing of AIRR-seq big data could potentially reveal amazing scientific insights. The AIRR Community has proposed MiAIRR (Minimum information about an Adaptive Immune Receptor Repertoire Sequencing Experiment), a standard for reporting AIRR-seq studies. The MiAIRR standard has been implemented using the National Center for Biotechnology Information (NCBI) repositories. Submissions of AIRR-seq data to the NCBI repositories typically use a combination of web-based and flat-file templates and include only a minimal amount of terminology validation. As a result, AIRR-seq studies at the NCBI are often described using inconsistent terminologies, limiting scientists' ability to access, find, interoperate, and reuse the data sets and to understand how the experiments were performed. CEDAR (Center for Expanded Data Annotation and Retrieval) develops technologies involving the use of data standards and ontologies to improve metadata quality. In order to improve metadata quality and ease AIRR-seq study submission process, we have developed an AIRR-seq data submission pipeline named CEDAR-AIRR (CAIRR). CAIRR leverages CEDAR's technologies to: i) create web-based templates whose entries are controlled by ontology terms, ii) generate and validate metadata and iii) submit the ontology-linked metadata and sequence files (FASTQ) to the NCBI BioProject, BioSample, and Sequence Read Archive (SRA) databases. Thus, CAIRR provides a web-based metadata submission interface that supports compliance with MiAIRR standards. The interface enables ontology-based validation for several data elements, including: organism, disease, cell type and subtype, and tissue. This pipeline will facilitate the NCBI submission process and improve the metadata quality of AIRR-seq studies.

Submission Steps

The submission steps are described in the MiAIRR-to-NCBI Submission Manual: *Option 1. Submission via the CEDAR system (CAIRR submission pipeline)*. You will need a CEDAR system account; you can self-register at <https://cedar.metadatascenter.org>. You will also need the identifier of a BioProject already entered in the NCBI BioProject database.

Citing the MiAIRR Pipeline

Bukhari, Syed Ahmad Chan, Martin J. O'Connor, Marcos Martínez-Romero, Attila L. Egyedi, Debra Debra Willrett, John Graybeal, Mark A. Musen, Florian Rubelt, Kei H. Cheung, and Steven H. Kleinstein. *The CAIRR pipeline for submitting standards-compliant B and T cell receptor repertoire sequencing studies to the NCBI*. *Frontiers in Immunology* 9 (2018): 1877. DOI: 10.3389/fimmu.2018.01877

Tell Us About It

Please let us know how it went! If you are willing, we would love to have your comments in a [short survey](#), it should just take 5 minutes or so.

We also welcome entry of issues and requests in our [github repository issues](#), and emails can be sent to cedar-users@lists.stanford.edu. Both of these resources are publicly visible.

Support or Contact

Having trouble with NCBI submission process through our pipeline? Please email to [Syed Ahmad Chan Bukhari](#) or to [Marcos Martínez-Romero](#) and we will help you sort it out.

Introduction to VDJServer

VDJServer is a cloud-based analysis portal for immune repertoire sequence data that provides access to a suite of tools for a complete analysis workflow, including modules for preprocessing and quality control of sequence reads, V(D)J gene assignment, repertoire characterization, and repertoire comparison. VDJServer also provides sophisticated visualizations for exploratory analysis. It is accessible through a standard web browser via a graphical user interface designed for use by immunologists, clinicians, and bioinformatics researchers. VDJServer provides a data commons for public sharing of repertoire sequencing data, as well as private sharing of data between users.

- [VDJServer website](#)
- [VDJServer Community Data Portal](#)
- [Email VDJServer](#) for information about submission of your study data.

References

Introduction to iReceptor

iReceptor is a platform for storing, sharing, and exploring AIRR-seq data according to the AIRR Community standards.

- [iReceptor Website](#) (General information)
- [iReceptor Gateway](#) (AIRR Data Commons data query and federation)
- [iReceptor Repositories](#) (AIRR Data Commons repositories)

- [iReceptor Turnkey GitHub](#) (Software)
- [Email iReceptor](#) (Contact).

References

2.4.2 Data Submission for Inferred Genes and Alleles

In 2017, The AIRR Community established the Inferred Allele Review Committee (IARC) to evaluate inferred alleles for inclusion in relevant germline databases. IARC has worked, together with colleagues at IMGT and the US National Institutes of Health, to establish a systematic submission and review process. OGRDB was created and designed to support that process, and provide a real-time record of affirmed sequences.

OGRDB - reference database of inferred immune receptor genes

In recent years it has become possible to sequence immune receptor repertoires (immunoglobulins and T cell receptors) at great depth. The accurate analysis of these repertoires requires a comprehensive understanding of the germline genes that give rise to the repertoire through V(D)J gene recombination.

Even for well-studied species such as humans and mice, our knowledge of allelic variation is incomplete. Identifying new immunoglobulin and T cell receptor polymorphisms from the genome using traditional methods is technically challenging, because of the complex sequence architecture and repetitive nature of these loci. More recently, methods have been developed to infer novel sequences and alleles from sequenced repertoires.

The Adaptive Immune Receptor Repertoire (AIRR) Community was formed to promote and share good practice in adaptive immune repertoire sequencing. In 2017, it established the Inferred Allele Review Committee (IARC) to evaluate inferred alleles for inclusion in relevant germline databases. IARC's work is outlined in more detail in a poster, which was presented at a Systems Immunology Workshop at the University of Surrey, England, in March 2018, and in a recent paper. IARC has worked, together with colleagues at IMGT and the US National Institute of Health, to establish a systematic submission and review process. OGRDB was created and designed to support that process, and provide a real-time record of affirmed sequences. Affirmed sequences will be listed under the Sequences tab above, and the submissions that underpin them will be found under the Submissions tab. You can make your own submissions by following the steps below.

How to submit your sequences

As a first step, IARC is now ready to review submissions of inferred human IGHV genes and alleles. These sequences may be novel, or may extend incomplete sequences currently in the IMGT reference directories. Researchers interested in submitting sequences should:

Submission of IARC gene inference data to NCBI

General outline

IARC submission currently follows a “INSDC first” approach, means that all sequence data related to the reported inference is **REQUIRED** to be properly deposited in a general purpose sequence repository before it is reviewed by IARC. The submitter needs to complete the initial steps of submission to one of the INSDC repositories. Upon submission to IARC, some of this data will be pulled in from NCBI (TODO: What kind of data can we actually pull down from INSDC?)

The aim of this procedure is to ensure that inferences reviewed by IARC are public and will remain available in the long run. It is however explicitly *not* the aim to provide data that deterministically will yield the same inference results.

Deposition of inferred gene data at NCBI

At the end of the deposition process there should be three types of records present at NCBI:

1. A single record containing the final and full-length inferred sequence. The record is deposited in one of the following:
 - Genbank: All inferences that have been performed on the submitters own data CAN be submitted as [??] to Genbank. Note that Genbank typically only holds data that has a physical correlate which is not necessarily true for inferred sequences. Nevertheless NCBI currently accepts this as a kind of consensus building if it is performed on your own data. The Genbank record MUST link to the `select set` record (see 3.) via the DBLINK/DR field. Genbank records will be publicly available independent of other publications. Note that for Genbank, the DBLINK field does not appear to be available through the [BankIt](#) submission interface. You can use `Tb12asn` and `Sequin`, and edit the DBLINK field manually (as “Sequence Read Archive” is not one of the options on the [template creation page](#). A sample Genbank deposit can be found under accession [MK321694](#).
 - TPA (Third-party annotation): A segment of Genbank dedicated inferences. Also the TPA record MUST link to the `select set` record (see 3.) via the DBLINK/DR field. Note that in contrast to Genbank, TPA does REQUIRE a peer-reviewed publication describing the details of the inference process before the record will be made publicly available. A sample TPA deposit can be found under accession [BK01573](#).

The format for both record types the Genbank format (link) with a standardized feature table (FT). Note that your initial submission MUST NOT contain any potential name for the gene as this will be assigned by IARC later on.

TODO: Is there any metadata that should be provided into the GB record?

2. One or multiple SRA records containing all raw reads of the the respective sequencing run. Note that if you are performing inference using third-party data, these records MUST be submitted by the original owner of the data. These record type will typically be present before the other. The metadata annotation of the records SHOULD be MiAIRR compliant [Rubelt et al.].
3. One or multiple SRA records containing the `select set` of reads from (2). The aim of these records is to document the number, quality, coverage and diversity of the reads in a dataset that *potentially* support the inference. This means that the `select set` SHOULD be a superset of the reads that support the inference. It is NOT REQUIRED that inference tools deterministically return the inferred allele upon being fed with the `select set`. Generation of the `select set` from the complete set is described below. When submitting the `select set` to SRA the metadata context, i.e. the original links to project, sample and (if possible) experiment) SHOULD be maintained. Reads originating from multiple subjects or samples MUST NOT be pooled into a single new entry. The new record SHOULD be titled “Reads from <original_run_accession> supporting inference of Homo sapiens immunoglobulin heavy chain variable gene” and contain a design description, e.g., “Experimental workflow as described in original SRA/ENA record [<run_accession>]. Gene inference was performed using <software+version+parameters>. The reported reads were selected based on <selection_criteria>.”

NOTE: It is reasonably likely, in the short term, that you will encounter questions from the SRA/ENA/Genbank staff about the nature of these deposits. If so, you can respond that they are made as part of a community effort to document novel alleles with an emphasis on transparency in data provenance. You can link to the [IARC page](#) and note that we worked together with IMGT and Genbank/TPA staff in designing this procedure.

Generating the `select set`

Below is the current procedure describing how to generate a `select set` using general purpose tools. This procedure was designed in a rather generic fashion so that it is easy to implement and does NOT REQUIRE inference tools to provide their own mechanisms. Note that it is currently assumed that the procedure is not fully deterministic, i.e. the `select set` cannot simply be generated using the complete read data and the inferred sequence, as there are additional filter criteria that apply. In addition the `select set` SHOULD not be subject to any modifications that are not listed below. This includes UMI-based consensus building or other aggregation steps that are not fully transparent to a third-party.

1. Assemble paired-end reads. The two reads MUST overlap. Recommended tool: PandaSeq
2. Perform PHRED filtering that is equivalent to the one performed by inference pipeline. Recommended tool: Immcantation suite
3. Perform a *blastn* search using the data from (2.) as query and bp 1-312 of the inferred gene as reference library. Require matches to be full-length and >99.6% ID. Record all matching read ID. Recommended tool: NCBI BLAST
4. Select the reads with the read ID found in (3.) from the original unmerged FASTQs. Note that each `select set` MUST be derived from a single donor and sample. Recommended tool: Christian's cryptic extractor script
5. Submit the `select set` to SRA

Submit the inferred sequences to IARC via OGRDB, following the

- [OGRDB Submission Guide](#)

Additional information is available at the

- [OGRDB Website](#)

References

2.4.3 Data Query and Download from the AIRR Data Commons

Submission of AIRR-seq datasets to public data repositories means that other researchers can query, download and reuse that data for novel analyses.

AIRR Data Commons

The AIRR Data Commons is a network of distributed repositories that store AIRR-seq data and adhere to the AIRR Community standards. We define the AIRR Data Commons as consisting of the set of repositories that both:

- Adhere to the [AIRR Common Repositories Working Group recommendations](#) for promoting, sharing, and use of AIRR-seq data.
- Implement the [ADC API](#) as a programmatic mechanism to access that data.

More information on repositories in the AIRR Data Commons and how to query these repositories can be found on the AIRR Data Commons page:

Other Public AIRR-Seq Repositories

There are additional data repositories that provide access to AIRR-seq data but which did not implement the ADC API for programmatic access. Information about some of these repositories are provided in a [B-T.CR forum post](#).

Germline Gene Inference and Usage

- [OGRDB](#) provides a list of alleles affirmed by the AIRR Community’s Inferred Allele Review Committee, together with supporting information.
- [VDJbase](#) provides gene usage information derived from a growing base of AIRR-seq repertoires, including inferred genotypes and haplotypes.

2.5 Software

2.5.1 AIRR Standards Reference Implementations

AIRR Python Reference Library

The `airr` reference library provides basic functions and classes for interacting with AIRR Community Data Representation Standards, including tools for read, write and validation.

API Reference

Rearrangement Interface

`airr.read_rearrangement(filename, validate=False, debug=False)`

Open an iterator to read an AIRR rearrangements file

Parameters

- **file** (*str*) – path to the input file.
- **validate** (*bool*) – whether to validate data as it is read, raising a `ValidationError` exception in the event of an error.
- **debug** (*bool*) – debug flag. If `True` print debugging information to standard error.

Returns

iterable reader class.

Return type

`airr.io.RearrangementReader`

`airr.create_rearrangement(filename, fields=None, debug=False)`

Create an empty AIRR rearrangements file writer

Parameters

- **filename** (*str*) – output file path.
- **fields** (*list*) – additional non-required fields to add to the output.
- **debug** (*bool*) – debug flag. If `True` print debugging information to standard error.

Returns

open writer class.

Return type

airr.io.RearrangementWriter

airr.derive_rearrangement(*out_filename*, *in_filename*, *fields=None*, *debug=False*)

Create an empty AIRR rearrangements file with fields derived from an existing file

Parameters

- **out_filename** (*str*) – output file path.
- **in_filename** (*str*) – existing file to derive fields from.
- **fields** (*list*) – additional non-required fields to add to the output.
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

open writer class.

Return type

airr.io.RearrangementWriter

airr.load_rearrangement(*filename*, *validate=False*, *debug=False*)

Load the contents of an AIRR rearrangements file into a data frame

Parameters

- **filename** (*str*) – input file path.
- **validate** (*bool*) – whether to validate data as it is read, raising a `ValidationError` exception in the event of an error.
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

Rearrangement records as rows of a data frame.

Return type

`pandas.DataFrame`

airr.dump_rearrangement(*dataframe*, *filename*, *debug=False*)

Write the contents of a data frame to an AIRR rearrangements file

Parameters

- **dataframe** (*pandas.DataFrame*) – data frame of rearrangement data.
- **filename** (*str*) – output file path.
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

True if the file is written without error.

Return type

`bool`

airr.merge_rearrangement(*out_filename*, *in_filenames*, *drop=False*, *debug=False*)

Merge one or more AIRR rearrangements files

Parameters

- **out_filename** (*str*) – output file path.

- **in_filenames** (*list*) – list of input files to merge.
- **drop** (*bool*) – drop flag. If True then drop fields that do not exist in all input files, otherwise combine fields from all input files.
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

True if files were successfully merged, otherwise False.

Return type

bool

`airr.validate_rearrangement(filename, debug=False)`

Validates an AIRR rearrangements file

Parameters

- **filename** (*str*) – path of the file to validate.
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

True if files passed validation, otherwise False.

Return type

bool

AIRR Data Model Interface

`airr.read_airr(filename, format=None, validate=False, model=True, debug=False)`

Load an AIRR Data file

Parameters

- **filename** (*str*) – path to the input file.
- **format** (*str*) – input file format valid strings are “yaml” or “json”. If set to None, the file format will be automatically detected from the file extension.
- **validate** (*bool*) – whether to validate data as it is read, raising a `ValidationError` exception in the event of a validation failure.
- **model** (*bool*) – If True only validate objects defined in the AIRR DataFile schema. If False, attempt validation of all top-level objects. Ignored if `validate=False`.
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

dictionary of AIRR Data objects.

Return type

dict

`airr.write_airr(filename, data, format=None, info=None, validate=False, model=True, debug=False)`

Write an AIRR Data file

Parameters

- **filename** (*str*) – path to the output file.
- **data** (*dict*) – dictionary of AIRR Data Model objects.

- **format** (*str*) – output file format valid strings are “yaml” or “json”. If set to None, the file format will be automatically detected from the file extension.
- **info** (*object*) – info object to write. Will write current AIRR Schema info if not specified.
- **validate** (*bool*) – whether to validate data before it is written, raising a `ValidationError` exception in the event of a validation failure.
- **model** (*bool*) – If True only validate and write objects defined in the AIRR DataFile schema. If False, attempt validation and write of all top-level objects
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

True if the file is written without error.

Return type

bool

`airr.validate_airr(data, model=True, debug=False)`

Validates an AIRR Data file

Parameters

- **data** (*dict*) – dictionary containing AIRR Data Model objects
- **model** (*bool*) – If True only validate objects defined in the AIRR DataFile schema. If False, attempt validation of all top-level objects
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

True if files passed validation, otherwise False.

Return type

bool

Classes

`class airr.io.RearrangementReader(handle, base=1, validate=False, debug=False)`

Iterator for reading Rearrangement objects in TSV format

fields

field names in the input Rearrangement file.

Type

list

external_fields

list of fields in the input file that are not part of the Rearrangement definition.

Type

list

`__init__(handle, base=1, validate=False, debug=False)`

Initialization

Parameters

- **handle** (*file*) – file handle of the open Rearrangement file.

- **base** (*int*) – one of 0 or 1 specifying the coordinate schema in the input file. If 1, then the file is assumed to contain 1-based closed intervals that will be converted to python style 0-based half-open intervals for known fields. If 0, then values will be unchanged.
- **validate** (*bool*) – perform validation. If True then basic validation will be performed will reading the data. A `ValidationError` exception will be raised if an error is found.
- **debug** (*bool*) – debug state. If True prints debug information.

Returns

reader object.

Return type

airr.io.RearrangementReader

__iter__()

Iterator initializer

Returns

airr.io.RearrangementReader

__next__()

Next method

Returns

parsed Rearrangement data.

Return type

dict

close()

Closes the Rearrangement file

next()

Next method

class *airr.io.RearrangementWriter*(*handle, fields=None, base=1, debug=False*)

Writer class for Rearrangement objects in TSV format

fields

field names in the output Rearrangement file.

Type

list

external_fields

list of fields in the output file that are not part of the Rearrangement definition.

Type

list

__init__(*handle, fields=None, base=1, debug=False*)

Initialization

Parameters

- **handle** (*file*) – file handle of the open Rearrangements file.
- **fields** (*list*) – list of non-required fields to add. May include fields undefined by the schema.

- **base** (*int*) – one of 0 or 1 specifying the coordinate schema in the output file. Data provided to the write is assumed to be in python style 0-based half-open intervals. If 1, then data will be converted to 1-based closed intervals for known fields before writing. If 0, then values will be unchanged.
- **debug** (*bool*) – debug state. If True prints debug information.

Returns

writer object.

Return type

airr.io.RearrangementWriter

close()

Closes the Rearrangement file

write(row)

Write a row to the Rearrangement file

Parameters

row (*dict*) – row to write.

class `airr.schema.Schema`(*definition*)

AIRR schema definitions

definition

name of the schema definition.

info

schema info.

Type

`collections.OrderedDict`

properties

field definitions.

Type

`collections.OrderedDict`

required

list of mandatory fields.

Type

list

optional

list of non-required fields.

Type

list

false_values

accepted string values for False.

Type

list

true_values

accepted values for True.

Type

list

from_bool(*value*, *validate=False*)

Converts a boolean to a string

Parameters

- **value** (*bool*) – logical value.
- **validate** (*bool*) – when True raise a `ValidationError` for an invalid value. Otherwise, set invalid values to None.

Returns

conversion of True or False or 'T' or 'F'.

Return type

str

Raises**airr.ValidationError** – raised if value is invalid when validate is set True.**pandas_types**()

Map of schema types to pandas types

Returns

mapping dictionary for pandas types

Return type

dict

spec(*field*)

Get the properties for a field

Parameters**name** (*str*) – field name.**Returns**

definition for the field.

Return type

collections.OrderedDict

template()

Create an empty template object

Returns

dictionary with all schema properties set as None or an empty list.

Return type

collections.OrderedDict

to_bool(*value*, *validate=False*)

Convert a string to a boolean

Parameters

- **value** (*str*) – logical value as a string.
- **validate** (*bool*) – when True raise a `ValidationError` for an invalid value. Otherwise, set invalid values to None.

Returns

conversion of the string to True or False.

Return type

bool

Raises**airr.ValidationError** – raised if value is invalid when validate is set True.**to_float**(*value*, *validate=False*)

Converts a string to a float

Parameters

- **value** (*str*) – float value as a string.
- **validate** (*bool*) – when True raise a ValidationError for an invalid value. Otherwise, set invalid values to None.

Returns

conversion of the string to a float.

Return type

float

Raises**airr.ValidationError** – raised if value is invalid when validate is set True.**to_int**(*value*, *validate=False*)

Converts a string to an integer

Parameters

- **value** (*str*) – integer value as a string.
- **validate** (*bool*) – when True raise a ValidationError for an invalid value. Otherwise, set invalid values to None.

Returns

conversion of the string to an integer.

Return type

int

Raises**airr.ValidationError** – raised if value is invalid when validate is set True.**type**(*field*)

Get the type for a field

Parameters**name** (*str*) – field name.**Returns**

the type definition for the field

Return type

str

validate_header(*header*)

Validate header against the schema

Parameters**header** (*list*) – list of header fields.**Returns**

True if a ValidationError exception is not raised.

Return type

bool

Raises

airr.ValidationError – raised if header fails validation.

validate_object(*obj*, *missing=True*, *nonairr=True*, *context=None*)

Validate Repertoire object data against schema

Parameters

- **obj** (*dict*) – dictionary containing a single repertoire object.
- **missing** (*bool*) – provides warnings for missing optional fields.
- **(bool (nonairr))** – provides warning for non-AIRR fields that cannot be validated.
- **context** (*string*) – used by recursion to indicate place in object hierarchy

Returns

True if a ValidationError exception is not raised.

Return type

bool

Raises

airr.ValidationError – raised if object fails validation.

validate_row(*row*)

Validate Rearrangements row data against schema

Parameters

row (*dict*) – dictionary containing a single record.

Returns

True if a ValidationError exception is not raised.

Return type

bool

Raises

airr.ValidationError – raised if row fails validation.

Schema

airr.schema.InfoSchema Schema object for the Info definition

AIRR schema definitions

airr.schema.definition

name of the schema definition.

airr.schema.info

schema info.

Type

collections.OrderedDict

airr.schema.properties

field definitions.

Type

collections.OrderedDict

`airr.schema.required`

list of mandatory fields.

Type

list

`airr.schema.optional`

list of non-required fields.

Type

list

`airr.schema.false_values`

accepted string values for False.

Type

list

`airr.schema.true_values`

accepted values for True.

Type

list

`airr.schema.DataFileSchema` Schema object for the DataFile definition

AIRR schema definitions

`airr.schema.definition`

name of the schema definition.

`airr.schema.info`

schema info.

Type

`collections.OrderedDict`

`airr.schema.properties`

field definitions.

Type

`collections.OrderedDict`

`airr.schema.required`

list of mandatory fields.

Type

list

`airr.schema.optional`

list of non-required fields.

Type

list

`airr.schema.false_values`

accepted string values for False.

Type

list

`airr.schema.true_values`

accepted values for True.

Type

list

`airr.schema.AlignmentSchema` Schema object for the Alignment definition

AIRR schema definitions

`airr.schema.definition`

name of the schema definition.

`airr.schema.info`

schema info.

Type

collections.OrderedDict

`airr.schema.properties`

field definitions.

Type

collections.OrderedDict

`airr.schema.required`

list of mandatory fields.

Type

list

`airr.schema.optional`

list of non-required fields.

Type

list

`airr.schema.false_values`

accepted string values for False.

Type

list

`airr.schema.true_values`

accepted values for True.

Type

list

`airr.schema.RearrangementSchema` Schema object for the Rearrangement definition

AIRR schema definitions

`airr.schema.definition`

name of the schema definition.

`airr.schema.info`

schema info.

Type

collections.OrderedDict

`airr.schema.properties`

field definitions.

Type

`collections.OrderedDict`

`airr.schema.required`

list of mandatory fields.

Type

list

`airr.schema.optional`

list of non-required fields.

Type

list

`airr.schema.false_values`

accepted string values for False.

Type

list

`airr.schema.true_values`

accepted values for True.

Type

list

`airr.schema.RepertoireSchema` Schema object for the Repertoire definition

AIRR schema definitions

`airr.schema.definition`

name of the schema definition.

`airr.schema.info`

schema info.

Type

`collections.OrderedDict`

`airr.schema.properties`

field definitions.

Type

`collections.OrderedDict`

`airr.schema.required`

list of mandatory fields.

Type

list

`airr.schema.optional`

list of non-required fields.

Type

list

`airr.schema.false_values`

accepted string values for False.

Type

list

`airr.schema.true_values`

accepted values for True.

Type

list

`airr.schema.GermlineSetSchema` Schema object for the Repertoire definition

AIRR schema definitions

`airr.schema.definition`

name of the schema definition.

`airr.schema.info`

schema info.

Type

collections.OrderedDict

`airr.schema.properties`

field definitions.

Type

collections.OrderedDict

`airr.schema.required`

list of mandatory fields.

Type

list

`airr.schema.optional`

list of non-required fields.

Type

list

`airr.schema.false_values`

accepted string values for False.

Type

list

`airr.schema.true_values`

accepted values for True.

Type

list

`airr.schema.GenotypeSetSchema` Schema object for the Repertoire definition

AIRR schema definitions

`airr.schema.definition`

name of the schema definition.

`airr.schema.info`

schema info.

Type

`collections.OrderedDict`

`airr.schema.properties`

field definitions.

Type

`collections.OrderedDict`

`airr.schema.required`

list of mandatory fields.

Type

list

`airr.schema.optional`

list of non-required fields.

Type

list

`airr.schema.false_values`

accepted string values for False.

Type

list

`airr.schema.true_values`

accepted values for True.

Type

list

Deprecated

`airr.load_repertoire(filename, validate=False, debug=False)`

Load an AIRR repertoire metadata file

Parameters

- **filename** (*str*) – path to the input file.
- **validate** (*bool*) – whether to validate data as it is read, raising a `ValidationError` exception in the event of an error.
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

dictionary of AIRR Data objects.

Return type

dict

Deprecated since version 1.4: Use `read_airr()` instead.

`airr.write_repertoire(filename, repertoires, info=None, debug=False)`

Write an AIRR repertoire metadata file

Parameters

- **file** (*str*) – path to the output file.
- **repertoires** (*list*) – array of repertoire objects.
- **info** (*object*) – info object to write. Will write current AIRR Schema info if not specified.
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

True if the file is written without error.

Return type

bool

Deprecated since version 1.4: Use `write_airr()` instead.

`airr.validate_repertoire(filename, debug=False)`

Validates an AIRR repertoire metadata file

Parameters

- **filename** (*str*) – path of the file to validate.
- **debug** (*bool*) – debug flag. If True print debugging information to standard error.

Returns

True if files passed validation, otherwise False.

Return type

bool

Deprecated since version 1.4: Use `validate_airr()` instead.

`airr.repertoire_template()`

Return a blank repertoire object from the template. This object has the complete structure with all of the fields and all values set to None or empty string.

Returns

empty repertoire object.

Return type

object

Deprecated since version 1.4: Use `schema.Schema.template()` instead.

Commandline Tools

Deprecated since version 1.4: The `validate repertoire` subcommand is deprecated. Use `validate airr` instead.

airr-tools

AIRR Community Standards utility commands.

```
usage: airr-tools [-h] [--version] ...
```

-h, --help

show this help message and exit

--version

show program's version number and exit

airr-tools merge

Merge AIRR rearrangement files.

```
usage: airr-tools merge [--version] [-h] -o OUT_FILE [--drop] -a AIRR_FILES
                        [AIRR_FILES ...]
```

--version

show program's version number and exit

-h, --help

show this help message and exit

-o <out_file>

Output file name.

--drop

If specified, drop fields that do not exist in all input files. Otherwise, include all columns in all files and fill missing data with empty strings.

-a <airr_files>

A list of AIRR rearrangement files.

airr-tools validate

Validate files for AIRR Standards compliance.

```
usage: airr-tools validate [--version] [-h] ...
```

--version

show program's version number and exit

-h, --help

show this help message and exit

airr-tools validate airr

Validate AIRR Data Model files.

```
usage: airr-tools validate airr [--version] [-h] -a AIRR_FILES
                                [AIRR_FILES ...]
```

--version

show program's version number and exit

-h, --help

show this help message and exit

-a <airr_files>

A list of AIRR Data Model files.

airr-tools validate rearrangement

Validate AIRR rearrangement files.

```
usage: airr-tools validate rearrangement [--version] [-h] -a AIRR_FILES
                                         [AIRR_FILES ...]
```

--version

show program's version number and exit

-h, --help

show this help message and exit

-a <airr_files>

A list of AIRR rearrangement files.

airr-tools validate repertoire

Validate AIRR repertoire metadata files.

```
usage: airr-tools validate repertoire [--version] [-h] -a AIRR_FILES
                                       [AIRR_FILES ...]
```

--version

show program's version number and exit

-h, --help

show this help message and exit

-a <airr_files>

A list of AIRR repertoire metadata files.

Python Library Release Notes

Version 1.5.0: August 29, 2023

1. Updated schema set and examples to v1.5.
2. Officially dropped support for Python 2.
3. Added check for valid enum values to schema validation routines.
4. Set enum values to first defined value during template generation routines.
5. Removed mock dependency installation in ReadTheDocs environments from setup.
6. Improved package import time.

Version 1.4.1: August 27, 2022

General:

1. Updated pandas requirement to 0.24.0 or higher.
2. Added support for missing integer values (NaN) in `load_rearrangement` by casting to the pandas `Int64` data type.
3. Added `gzip` support to `read_rearrangement`.
4. Significant internal refactoring to improve schema generalizability, harmonize behavior between the python and R libraries, and prepare for AIRR Standards v2.0.
5. Fixed a bug in the `validate` subcommand of `airr-tools` causing validation errors to only be reporting for the first invalid file when multiple files were specified on the command line.

Data Model and Schema:

1. Added support for arrays of objects in a single JSON or YAML file.
2. Added support for the AIRR Data File and associated schema (`DataFile`, `Info`). The Data File data format holds AIRR object of multiple types and is backwards compatible with Repertoire metadata.
3. Added support for the new germline and genotyping schema (`GermlineSet`, `GenotypeSet`) and associated schema.
4. Renamed `schema.CachedSchema` to `schema.AIRRSchema`.
5. Removed `specs/blank.airr.yaml`.

Deprecations:

1. Deprecated `load_repertoire`. Use `read_airr` instead.
2. Deprecated `write_repertoire`. Use `write_airr` instead.
3. Deprecated `validate_repertoire`. Use `validate_airr` instead.
4. Deprecated `repertoire_template`. Use `schema.RepertoireSchema.template` instead.
5. Deprecated the commandline tool `airr-tools validate repertoire`. Use `airr-tools validate airr` instead.

Version 1.3.1: October 13, 2020

1. Refactored `merge_rearrangement` to allow for larger number of files.
2. Improved error handling in format validation operations.

Version 1.3.0: May 30, 2020

1. Updated schema set to v1.3.
2. Added `load_repertoire`, `write_repertoire`, and `validate_repertoire` to `airr.interface` to read, write and validate Repertoire metadata, respectively.
3. Added `repertoire_template` to `airr.interface` which will return a complete repertoire object where all fields have null values.
4. Added `validate_object` to `airr.schema` that will validate a single repertoire object against the schema.
5. Extended the `airr-tools` commandline program to validate both rearrangement and repertoire files.

Version 1.2.1: October 5, 2018

1. Fixed a bug in the python reference library causing start coordinate values to be empty in some cases when writing data.

Version 1.2.0: August 17, 2018

1. Updated schema set to v1.2.
2. Several improvements to the `validate_rearrangement` function.
3. Changed behavior of all `airr.interface` functions to accept a file path (string) to a single Rearrangement TSV, instead of requiring a file handle as input.
4. Added `base` argument to `RearrangementReader` and `RearrangementWriter` to support optional conversion of 1-based closed intervals in the TSV to python-style 0-based half-open intervals. Defaults to conversion.
5. Added the custom exception `ValidationError` for handling validation checks.
6. Added the `validate` argument to `RearrangementReader` which will raise a `ValidationError` exception when reading files with missing required fields or invalid values for known field types.
7. Added `validate` argument to all type conversion methods in `Schema`, which will now raise a `ValidationError` exception for value that cannot be converted when set to `True`. When set `False` (default), the previous behavior of assigning `None` as the converted value is retained.
8. Added `validate_header` and `validate_row` methods to `Schema` and removed `validations` methods from `RearrangementReader`.
9. Removed automatic closure of file handle upon reaching the iterator end in `RearrangementReader`.

Version 1.1.0: May 1, 2018

Initial release.

Installation

Install in the usual manner from PyPI:

```
> pip3 install airr --user
```

Or from the [downloaded](#) source code directory:

```
> python3 setup.py install --user
```

Quick Start

Deprecation Notice

The `load_repertoire`, `write_repertoire`, and `validate_repertoire` functions have been deprecated for the new generic `load_airr_data`, `write_airr_data`, and `validate_airr_data` functions. These new functions are backwards compatible with the Repertoire metadata format but also support the new AIRR objects such as `GermlineSet`, `RepertoireGroup`, `GenotypeSet`, `Cell` and `Clone`. This new format is defined by the `DataFile` Schema, which describes a standard set of objects included in a file containing AIRR Data Model presentations. Currently, the AIRR `DataFile` does not completely support Rearrangement, so users should continue using AIRR TSV files and its specific functions. Also, the `repertoire_template` function has been deprecated for the `Schema.template` method, which can now be called on any AIRR Schema to create a blank object.

Reading AIRR Data Files

The `airr` package contains functions to read and write AIRR Data Model files. The file format is either YAML or JSON, and the package provides a light wrapper over the standard parsers. The file needs a `json`, `yaml`, or `yaml` file extension so that the proper parser is utilized. All of the AIRR objects are loaded into memory at once and no streaming interface is provided:

```
import airr

# Load the AIRR data
data = airr.read_airr('input.airr.json')
# loop through the repertoires
for rep in data['Repertoire']:
    print(rep)
```

Why are the AIRR objects, such as `Repertoire`, `GermlineSet`, and etc., in a list versus in a dictionary keyed by their identifier (e.g., `repertoire_id`)? There are two primary reasons for this. First, the identifier might not have been assigned yet. Some systems might allow MiAIRR metadata to be entered but the identifier is assigned to that data later by another process. Without the identifier, the data could not be stored in a dictionary. Secondly, the list allows the data to have a default ordering. If you know that the data has a unique identifier then you can quickly create a dictionary object using a comprehension. For example, with repertoires:

```
rep_dict = { obj['repertoire_id'] : obj for obj in data['Repertoire'] }
```

another example with germline sets:

```
germline_dict = { obj['germline_set_id'] : obj for obj in data['GermlineSet'] }
```

Writing AIRR Data Files

Writing an AIRR Data File is also a light wrapper over standard YAML or JSON parsers. Multiple AIRR objects, such as Repertoire, GermlineSet, and etc., can be written together into the same file. In this example, we use the `airr` library `template` method to create some blank Repertoire objects, and write them to a file. As with the `read` function, the complete list of repertoires are written at once, there is no streaming interface:

```
import airr

# Create some blank repertoire objects in a list
data = { 'Repertoire': [] }
for i in range(5):
    data['Repertoire'].append(airr.schema.RepertoireSchema.template())

# Write the AIRR Data
airr.write_airr('output.airr.json', data)
```

Reading AIRR Rearrangement TSV files

The `airr` package contains functions to read and write AIRR Rearrangement TSV files as either iterables or pandas data frames. The usage is straightforward, as the file format is a typical tab delimited file, but the package performs some additional validation and type conversion beyond using a standard CSV reader:

```
import airr

# Create an iterable that returns a dictionary for each row
reader = airr.read_rearrangement('input.tsv')
for row in reader: print(row)

# Load the entire file into a pandas data frame
df = airr.load_rearrangement('input.tsv')
```

Writing AIRR Rearrangement TSV files

Similar to the read operations, write functions are provided for either creating a writer class to perform row-wise output or writing the entire contents of a pandas data frame to a file. Again, usage is straightforward with the `airr` output functions simply performing some type conversion and field ordering operations:

```
import airr

# Create a writer class for iterative row output
writer = airr.create_rearrangement('output.tsv')
for row in reader: writer.write(row)

# Write an entire pandas data frame to a file
airr.dump_rearrangement(df, 'file.tsv')
```


By default, `create_rearrangement` will only write the required fields in the output file. Additional fields can be included in the output file by providing the `fields` parameter with an array of additional field names:

```
# Specify additional fields in the output
fields = ['new_calc', 'another_field']
writer = airr.create_rearrangement('output.tsv', fields=fields)
```

A common operation is to read an AIRR rearrangement file, and then write an AIRR rearrangement file with additional fields in it while keeping all of the existing fields from the original file. The `derive_rearrangement` function provides this capability:

```
import airr

# Read rearrangement data and write new file with additional fields
reader = airr.read_rearrangement('input.tsv')
fields = ['new_calc']
writer = airr.derive_rearrangement('output.tsv', 'input.tsv', fields=fields)
for row in reader:
    row['new_calc'] = 'a value'
    writer.write(row)
```

Validating AIRR data files

The `airr` package can validate AIRR Data Model JSON/YAML files and Rearrangement TSV files to ensure that they contain all required fields and that the fields types match the AIRR Schema. This can be done using the `airr-tools` command line program or the `validate` functions in the library can be called:

```
# Validate a rearrangement TSV file
airr-tools validate rearrangement -a input.tsv

# Validate an AIRR DataFile
airr-tools validate airr -a input.airr.json
```

Combining Repertoire metadata and Rearrangement files

The `airr` package does not currently keep track of which AIRR Data Model files are associated with which Rearrangement TSV files, though there is ongoing work to define a standardized manifest, so users will need to handle those associations themselves. However, in the data, AIRR identifier fields, such as `repertoire_id`, form the link between objects in the AIRR Data Model. The typical usage is that a program is going to perform some computation on the Rearrangements, and it needs access to the Repertoire metadata as part of the computation logic. This example code shows the basic framework for doing that, in this case doing gender specific computation:

```
import airr

# Load AIRR data containing repertoires
data = airr.read_airr('input.airr.json')

# Put repertoires in dictionary keyed by repertoire_id
rep_dict = { obj['repertoire_id'] : obj for obj in data['Repertoire'] }

# Create an iterable for rearrangement data
```

(continues on next page)

(continued from previous page)

```
reader = airr.read_rearrangement('input.tsv')
for row in reader:
    # get repertoire metadata with this rearrangement
    rep = rep_dict[row['repertoire_id']]

    # check the gender
    if rep['subject']['sex'] == 'male':
        # do male specific computation
    elif rep['subject']['sex'] == 'female':
        # do female specific computation
    else:
        # do other specific computation
```

AIRR R Reference Library

airr is an R package for working with data formatted according to the AIRR Standards schemas. It includes the full set of schema definitions along with simple functions for read, write and validation.

Usage Vignette

Introduction

Since the use of High-throughput sequencing (HTS) was first introduced to analyze immunoglobulin (B-cell receptor, antibody) and T-cell receptor repertoires (Freeman et al, 2009; Robins et al, 2009; Weinstein et al, 2009), the increasing number of studies making use of this technique has produced enormous amounts of data and there exists a pressing need to develop and adopt common standards, protocols, and policies for generating and sharing data sets. The [Adaptive Immune Receptor Repertoire \(AIRR\) Community](#) formed in 2015 to address this challenge (Breden et al, 2017) and has established the set of minimal metadata elements (MiAIRR) required for describing published AIRR datasets (Rubelt et al, 2017) as well as file formats to represent this data in a machine-readable form. The airr R package provide read, write and validation of data following the AIRR Data Representation schemas. This vignette provides a set of simple use examples.

AIRR Data Standards

The AIRR Community’s recommendations for a minimal set of metadata that should be used to describe an AIRR-seq data set when published or deposited in a AIRR-compliant public repository are described in Rubelt et al, 2017. The primary aim of this effort is to make published AIRR datasets FAIR (findable, accessible, interoperable, reusable); with sufficient detail such that a person skilled in the art of AIRR sequencing and data analysis will be able to reproduce the experiment and data analyses that were performed.

Following this principles, V(D)J reference alignment annotations are saved in standard tab-delimited files (TSV) with associated metadata provided in accompanying YAML formatted files. The column names and field names in these files have been defined by the AIRR Data Representation Working Group using a controlled vocabulary of standardized terms and types to refer to each piece of information.

Reading AIRR formatted files

The `airr` package contains the function `read_rearrangement` to read and validate files containing AIRR Rearrangement records, where a Rearrangement record describes the collection of optimal annotations on a single sequence that has undergone V(D)J reference alignment. The usage is straightforward, as the file format is a typical tabulated file. The argument that needs attention is `base`, with possible values "0" and "1". `base` denotes the starting index for positional fields in the input file. Positional fields are those that contain alignment coordinates and names ending in `"_start"` and `"_end"`. If the input file is using 1-based closed intervals (R style), as defined by the standard, then positional fields will not be modified under the default setting of `base="1"`. If the input file is using 0-based coordinates with half-open intervals (python style), then positional fields may be converted to 1-based closed intervals using the argument `base="0"`.

Reading Rearrangements

```
# Imports
library(airr)
library(tibble)

# Read Rearrangement example file
f1 <- system.file("extdata", "rearrangement-example.tsv.gz", package="airr")
rearrangement <- read_rearrangement(f1)
glimpse(rearrangement)
```

```
## Rows: 101
## Columns: 33
## $ sequence_id      <chr> "SRR765688.7787", "SRR765688.35420", "SRR765688.36681",
  ↳ "SRR765688.33811", "SRR765688.44149", "SRR765688.15636", "SRR765688.20304", "SRR765688.
  ↳ 13860", "SRR7...
## $ sequence         <chr>
  ↳ "NNNNNNNNNNNNNNNNNNNNNGCTGACCTGCACCTTCTCTGGATTCTCACTCAGTACTAGTGCAGTGGGTGTACACTGGATCCGTCAGCCCCCAGGAAAGG
## $ rev_comp         <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,
  ↳ FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE,
  ↳ FALSE, FALSE, FA...
## $ productive       <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, FALSE, TRUE,
  ↳ FALSE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE,
  ↳ TRUE, TRUE, T...
## $ vj_in_frame      <lgl> TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE,
  ↳ FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, TRUE, TRUE, TRUE, TRUE,
  ↳ TRUE, TRUE, TRUE...
## $ stop_codon       <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, TRUE,
  ↳ FALSE, TRUE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE,
  ↳ FALSE, FALSE, ...
## $ v_call           <chr> "IGHV2-5*02", "IGHV5-51*01", "IGHV7-4-1*02", "IGHV7-4-1*02
  ↳ ", "IGHV7-4-1*02", "IGHV2-5*02", "IGHV7-4-1*02", "IGHV6-1*01,IGHV6-1*02", "IGHV7-4-1*02
  ↳ ", "IGHV4-3...
## $ d_call           <chr> "IGHD5-24*01", "IGHD3-16*02,IGHD3-3*01,IGHD3-3*02", "IGHD3-
  ↳ 22*01", "IGHD3-9*01", "IGHD1-26*01", "IGHD2-21*02", "IGHD1-26*01,IGHD2-21*02,IGHD3/
  ↳ OR15-3a*01", "...
## $ j_call           <chr> "IGHJ4*02", "IGHJ6*02,IGHJ6*04", "IGHJ4*02", "IGHJ6*02",
  ↳ "IGHJ6*01", "IGHJ4*02", "IGHJ5*02", "IGHJ4*02", "IGHJ4*02", "IGHJ4*02", "IGHJ5*02",
  ↳ "IGHJ6*02", "IG...
```

(continues on next page)

(continued from previous page)

```

## $ c_call          <chr> "IGHG", "IGHG", "IGHG", "IGHG", "IGHG", "IGHA", "IGHA",
↳ "IGHG", "IGHG", "IGHA", "IGHA", "IGHG", "IGHA", "IGHA", "IGHG", "IGHA", "IGHG", "IGHA",
↳ "IGHG", "IGH...
## $ sequence_alignment <chr> ".....
↳ .GCTGACCTGCACCTTCTCTGGATTCTCACTCACTG.....
↳ ACTAGTGCAGTGGGTGTACACTGGATCCGTCAGCCCCCAGGAAAGGCCCTGGAGTG...
## $ germline_alignment <chr> "CAGATCACCTTGAAGGAGTCTGGTCTCT...
↳ ACGCTGGTGAACCCACACAGACCTCACGCTGACCTGCACCTTCTCTGGGTCTCACTCAGC.....
↳ ACTAGTGGAGTGGGTGTGGGCTGGATCCGTCAGCCCCCAGGAAAGGCCCTGGAGTG...
## $ junction        <chr> "TGTGCACACAGTGCGGGATGGCTGCCTGATTACTGG",
↳ "TGTGCGAGGCATGGATTATACGGTTGTGATCATACCGGCTGTTATACAAGCTTCTACTACTACGGGATGGACGTCTGG",
↳ "TGTGCGAGAGAAGAACGTCGAAGTAGTGGTTAT...
## $ junction_aa      <chr> "CAHSAGWLPDYW", "CARHGLYGC DHTGCTSFYYYYGMDVW",
↳ "CAREERRSSGYFDHW", "CAREGYFDTTGSPRSHGLDVW", "CARDSGMDVW", "CVLSRRLGDSGVQKYFDYW",
↳ "CAREGLWDGRVVTDLW", "CARTR...
## $ v_cigar          <chr>
↳ "20S56N21=1X11=1X7=1X9=3X62=6D2=1X1=2X2=2X50=1X7=1X4=1X22=1X30=",
↳ "20S40N15=1X15=1X11=1X2=1X1=1X1=2X3=1X7=1X41=2X2=1X10=1X3=1X1=1X5=2X5=1X4=1X9=1X19=1X24=2X...
## $ d_cigar          <chr> "274S5N7=", "305S29N7=", "293S13N12=", "290S9N8=",
↳ "283S4N7=", "273S12N8=", "289S6N6=", "267S9N9=", "281S7N5=", "278S7N5=1X7=", "277S8N7=
↳ ", "297S9N7=", "265...
## $ j_cigar          <chr> "288S11N32=1X4=", "318S7N12=1X15=", "305S5N6=1X14=1X21=",
↳ "321S15N5=1X23=1X17=", "290S17N19=", "296S26=1X21=", "311S11N4=1X33=",
↳ "280S2N17=1X6=1X21=", "299S...
## $ v_sequence_start <int> 21, 21, 21, 21, 21, 21, 21, 21, 20, 22, 21, 21, 20, 21, 21, 21,
↳ 21, 19, 21, 21, 21, 20, 21, 21, 21, 21, 21, 20, 23, 19, 21, 20, 21, 21, 20, 20, 21,
↳ 20, 22, 21, ...
## $ v_sequence_end   <int> 269, 276, 283, 283, 283, 264, 283, 259, 281, 266, 264, 294,
↳ 258, 283, 273, 279, 274, 259, 278, 280, 262, 271, 281, 262, 264, 283, 259, 279, 278,
↳ 280, 261, 2...
## $ v_germline_start <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
↳ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
↳ 1, 1, 1, ...
## $ v_germline_end   <int> 320, 320, 320, 320, 320, 320, 320, 320, 318, 318, 320, 320, 320,
↳ 319, 319, 317, 316, 316, 320, 318, 320, 320, 315, 318, 320, 321, 320, 318, 316, 320,
↳ 317, 319, 3...
## $ d_sequence_start <int> 275, 306, 294, 291, 284, 274, 290, 268, 282, 279, 278, 298,
↳ 266, 292, 301, 285, 276, NA, 291, 284, 273, 279, 289, 277, 272, 299, 270, 292, 282,
↳ 295, 267, 26...
## $ d_sequence_end   <int> 281, 312, 305, 298, 290, 281, 295, 276, 286, 291, 284, 304,
↳ 273, 296, 307, 289, 282, NA, 297, 295, 293, 290, 294, 283, 287, 307, 280, 302, 290,
↳ 301, 280, 27...
## $ d_germline_start <int> 6, 30, 14, 10, 5, 13, 7, 10, 8, 8, 9, 10, 10, 7, 22, 14, 4,
↳ NA, 24, 4, 2, 5, 8, 9, 6, 3, 1, 3, 7, 7, 4, 6, 11, 14, 12, 8, 2, 11, 8, 10, 5, 24, 4,
↳ 17, 5, 5, ...
## $ d_germline_end   <int> 12, 36, 25, 17, 11, 20, 12, 18, 12, 20, 15, 16, 17, 11, 28,
↳ 18, 10, NA, 30, 15, 22, 16, 13, 15, 21, 11, 11, 13, 15, 13, 17, 11, 17, 18, 19, 13, 8,
↳ 16, 20, 1...
## $ j_sequence_start <int> 289, 319, 306, 322, 291, 297, 312, 281, 300, 301, 289, 319,
↳ 276, 300, 317, 299, 296, 271, 336, 321, 303, 304, 300, 297, 293, 322, 289, 311, 315,
↳ 320, 283, 2...
## $ j_sequence_end   <int> 325, 346, 348, 368, 309, 344, 349, 326, 339, 347, 335, 361,

```

(continues on next page)

(continued from previous page)

```

↪ 326, 334, 350, 333, 346, 320, 370, 338, 332, 338, 339, 333, 340, 368, 332, 345, 342, ↪
↪ 362, 327, 3...
## $ j_germline_start    <int> 12, 8, 6, 16, 18, 1, 12, 3, 9, 2, 5, 20, 9, 14, 15, 14, 2, ↪
↪ 13, 28, 18, 6, 14, 9, 15, 1, 16, 5, 14, 8, 5, 4, 12, 9, 1, 20, 15, 6, 6, 5, 6, 5, 9, 1,
↪ 6, 5, 5, ...
## $ j_germline_end      <int> 48, 35, 48, 62, 36, 48, 49, 48, 48, 48, 51, 62, 59, 48, 48,
↪ 48, 52, 62, 62, 35, 35, 48, 48, 51, 48, 62, 48, 48, 35, 47, 48, 46, 48, 44, 62, 48, ↪
↪ 48, 51, 50, ...
## $ junction_length    <int> 36, 78, 45, 66, 33, 60, 48, 45, 36, 61, 51, 48, 51, 30, 54,
↪ 30, 48, 42, 71, 66, 78, 42, 36, 51, 57, 66, 51, 42, 72, 60, 45, 45, 45, 42, 36, 36, ↪
↪ 57, 48, 51, ...
## $ np1_length         <int> 5, 29, 10, 7, 0, 9, 6, 8, 0, 12, 13, 3, 7, 8, 27, 5, 1, 11,
↪ 12, 3, 10, 7, 7, 14, 7, 15, 10, 12, 3, 14, 5, 4, 4, 6, 1, 7, 1, 5, 4, 5, 26, 5, 0, 15,
↪ 26, 26, 8...
## $ np2_length         <int> 7, 6, 0, 23, 0, 15, 16, 4, 13, 9, 4, 14, 2, 3, 9, 9, 13, ↪
↪ NA, 38, 25, 9, 13, 5, 13, 5, 14, 8, 8, 24, 18, 2, 22, 15, 3, 3, 11, 29, 11, 9, 5, 1, 5,
↪ 8, 0, 1, 1,...
## $ duplicate_count    <int> 3, 3, 13, 3, 2, 2, 4, 2, 2, 2, 4, 2, 2, 2, 2, 2, 3, 3, 2, ↪
↪ 2, 3, 2, 2, 2, 2, 2, 2, 5, 2, 2, 3, 2, 4, 2, 3, 4, 8, 2, 2, 2, 2, 3, 2, 4, 3, 4, 2, 5, ↪
↪ 2, 2, 7, 3,...

```

Reading AIRR Data Models

AIRR Data Model records, such as Repertoire and GermlineSet, can be read from either a YAML or JSON formatted file into a nested list.

```

# Read Repertoire example file
f2 <- system.file("extdata", "repertoire-example.yaml", package="airr")
repertoire <- read_airr(f2)
glimpse(repertoire)

```

```

## List of 1
## $ Repertoire:List of 3
## ..$ :List of 5
## .. ..$ repertoire_id : chr "1841923116114776551-242ac11c-0001-012"
## .. ..$ study         :List of 13
## .. ..$ subject       :List of 15
## .. ..$ sample        :List of 1
## .. ..$ data_processing:List of 1
## ..$ :List of 5
## .. ..$ repertoire_id : chr "1602908186092376551-242ac11c-0001-012"
## .. ..$ study         :List of 13
## .. ..$ subject       :List of 15
## .. ..$ sample        :List of 1
## .. ..$ data_processing:List of 1
## ..$ :List of 5
## .. ..$ repertoire_id : chr "2366080924918616551-242ac11c-0001-012"
## .. ..$ study         :List of 13
## .. ..$ subject       :List of 15
## .. ..$ sample        :List of 1
## .. ..$ data_processing:List of 1

```

```
# Read GermlineSet example file
f3 <- system.file("extdata", "germline-example.json", package="airr")
germline <- read_airr(f3)
glimpse(germline)
```

```
## List of 2
## $ GermlineSet:List of 1
## ..$ :List of 17
## .. ..$ germline_set_id      : chr "OGRDB:G000007"
## .. ..$ author              : chr "William Lees"
## .. ..$ lab_name             : chr ""
## .. ..$ lab_address          : chr "Birkbeck College, University of London, Malet_
↳ Street, London"
## .. ..$ acknowledgements     : list()
## .. ..$ release_version      : int 1
## .. ..$ release_description   : chr ""
## .. ..$ release_date         : chr "2021-11-24"
## .. ..$ germline_set_name     : chr "CAST IGH"
## .. ..$ germline_set_ref      : chr "OGRDB:G000007.1"
## .. ..$ pub_ids              : chr ""
## .. ..$ species              :List of 2
## .. ..$ species_subgroup      : chr "CAST_EiJ"
## .. ..$ species_subgroup_type: chr "strain"
## .. ..$ locus                : chr "IGH"
## .. ..$ allele_descriptions   :List of 2
## .. ..$ curation              : NULL
## $ GenotypeSet:List of 1
## ..$ :List of 2
## .. ..$ receptor_genotype_set_id: chr "1"
## .. ..$ genotype_class_list    :List of 1
```

Writing AIRR formatted files

The `airr` package contains the function `write_rearrangement` to write Rearrangement records to the AIRR TSV format.

Writing Rearrangements

```
x1 <- file.path(tempdir(), "airr_out.tsv")
write_rearrangement(rearrangement, x1)
```

Writing AIRR Data Models

AIRR Data Model records can be written to either YAML or JSON using the `write_airr` function.

```
x2 <- file.path(tempdir(), "airr_repertoire_out.yaml")
write_airr(repertoire, x2, format="yaml")

x3 <- file.path(tempdir(), "airr_germline_out.json")
write_airr(germline, x3, format="json")
```

Validating AIRR data structures

The `airr` package contains the function `validate_rearrangement` to validate tabular (`data.frame`) Rearrangement records and AIRR Data Model objects, respectively.

```
# Validate Rearrangement data.frame
validate_rearrangement(rearrangement)
```

```
## [1] TRUE
```

```
# Validate an AIRR Data Model
validate_airr(repertoire)
```

```
## [1] TRUE
```

```
# Validate AIRR Data Model records individual
validate_airr(germline, each=TRUE)
```

```
## GenotypeSet GermlineSet
##      TRUE      TRUE
```

References

1. Breden, F., E. T. Luning Prak, B. Peters, F. Rubelt, C. A. Schramm, C. E. Busse, J. A. Vander Heiden, et al. 2017. Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. *Front Immunol* 8: 1418.
2. Freeman, J. D., R. L. Warren, J. R. Webb, B. H. Nelson, and R. A. Holt. 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Res* 19 (10): 1817-24.
3. Robins, H. S., P. V. Campregher, S. K. Srivastava, A. Wacher, C. J. Turtle, O. Kahsai, S. R. Riddell, E. H. Warren, and C. S. Carlson. 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in alphabeta T cells. *Blood* 114 (19): 4099-4107.
4. Rubelt, F., C. E. Busse, S. A. C. Bukhari, J. P. Burckert, E. Mariotti-Ferrandiz, L. G. Cowell, C. T. Watson, et al. 2017. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat Immunol* 18 (12): 1274-8.
5. Weinstein, J. A., N. Jiang, R. A. White, D. S. Fisher, and S. R. Quake. 2009. High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324 (5928): 807-10.

Reference Topics

read_tabular

Read AIRR tabular data

Description

read_tabular reads a tab-delimited (TSV) file containing tabular AIRR records.

Usage

```
read_tabular(file, schema, base = c("1", "0"), aux_types = NULL, ...)
```

```
read_rearrangement(file, base = c("1", "0"), ...)
```

```
read_alignment(file, base = c("1", "0"), ...)
```

Arguments

file

input file path.

schema

Schema object defining the output format.

base

starting index for positional fields in the input file. If "1", then these fields will not be modified. If "0", then fields ending in "_start" and "_end" are 0-based half-open intervals (python style) in the input file and will be converted to 1-based closed-intervals (R style).

aux_types

named vector or list giving the type for fields that are not defined in schema. The field name is the name, the value the type, denoted by one of "c" (character), "l" (logical), "i" (integer), "d" (double), or "n" (numeric).

...

additional arguments to pass to [read_delim](#).

Value

A `data.frame` of the TSV file with appropriate type and position conversion for fields defined in the specification.

Details

`read_rearrangement` reads an AIRR TSV containing Rearrangement data.

`read_alignment` reads an AIRR TSV containing Alignment data.

Examples

```
# Get path to the rearrangement-example file
file <- system.file("extdata", "rearrangement-example.tsv.gz", package="airr")

# Load data file
df <- read_rearrangement(file)
```

See also

See [Schema](#) for the AIRR schema object definition. See [write_tabular](#) for writing AIRR data.

read_airr

Read an AIRR Data Model file in YAML or JSON format

Description

`read_airr` loads a YAML or JSON file containing AIRR Data Model records.

Usage

```
read_airr(
  file,
  format = c("auto", "yaml", "json"),
  validate = TRUE,
  model = TRUE
)
```

Arguments

file

path to the input file.

format

format of the input file. Must be one of "auto", "yaml", or "json". If "auto" (default), the format will be detected from the file extension.

validate

run schema validation if TRUE.

model

if TRUE validate only AIRR DataFile defined objects. If FALSE attempt validation of all objects in data. Ignored if validate=FALSE

Value

A named nested list contained in the AIRR Data Model with the top-level names reflecting the individual AIRR objects.

Examples

```
# Get path to the Repertoire and GermlineSet example files
f1 <- system.file("extdata", "repertoire-example.yaml", package="airr")
f2 <- system.file("extdata", "germline-example.json", package="airr")

# Load data files
repertoire <- read_airr(f1)
germline <- read_airr(f2)
```

See also

See [Schema](#) for the AIRR schema definition objects. See [write_airr](#) for writing AIRR Data Model records in YAML or JSON format.

write_tabular

Write an AIRR tabular data

Description

write_tabular writes a TSV containing AIRR tabular records.

Usage

```
write_tabular(data, file, schema, base = c("1", "0"), ...)
```

```
write_rearrangement(data, file, base = c("1", "0"), ...)
```

```
write_alignment(data, file, base = c("1", "0"), ...)
```

Arguments

data

`data.frame` of Rearrangement data.

file

output file name.

schema

Schema object defining the output format.

base

starting index for positional fields in the output file. Fields in the input data are assumed to be 1-based closed-intervals (R style). If "1", then these fields will not be modified. If "0", then fields ending in `_start` and `_end` will be converted to 0-based half-open intervals (python style) in the output file.

...

additional arguments to pass to `write_delim`.

Details

`write_rearrangement` writes a `data.frame` containing AIRR Rearrangement data to TSV.

`write_alignment` writes a `data.frame` containing AIRR Alignment data to TSV.

Examples

```
# Get path to the rearrangement-example file
file <- system.file("extdata", "rearrangement-example.tsv.gz", package="airr")

# Load data file
df <- read_rearrangement(file)
```

```
# Write a Rearrangement data file
outfile <- file.path(tempdir(), "output.tsv")
write_tabular(df, outfile, schema=RearrangementSchema)
```

See also

See [Schema](#) for the AIRR schema object definition. See [read_tabular](#) for reading to AIRR files.

`write_airr`

Write AIRR Data Model records to YAML or JSON files

Description

`write_airr` writes a YAML or JSON file containing AIRR Data Model records.

Usage

```
write_airr(  
  data,  
  file,  
  format = c("auto", "yaml", "json"),  
  validate = TRUE,  
  model = TRUE  
)
```

Arguments

data

list containing AIRR Model Records.

file

output file name.

format

format of the output file. Must be one of "auto", "yaml", or "json". If "auto" (default), the format will be detected from the file extension.

validate

run schema validation prior to write if TRUE.

model

if TRUE validate and write only AIRR DataFile defined objects. If FALSE attempt validation and write of all objects in data.

Examples

```
# Get path to the repertoire-example file  
file <- system.file("extdata", "repertoire-example.yaml", package="airr")  
  
# Load data file  
repertoire <- read_airr(file)  
  
# Write a Rearrangement data file  
outfile <- file.path(tempdir(), "output.yaml")  
write_airr(repertoire, outfile)
```

See also

See [Schema](#) for the AIRR schema definition objects. See [read_airr](#) for reading to AIRR Data Model files.

validate_tabular

Validate tabular AIRR data

Description

`validate_tabular` validates compliance of the contents of a `data.frame` to the AIRR standards.

Usage

```
validate_tabular(data, schema)
```

```
validate_rearrangement(data)
```

Arguments

`data`

`data.frame` of tabular data to validate.

`schema`

Schema object defining the data standard of the table.

Value

Returns TRUE if the input data is compliant and FALSE if not.

Details

`validate_rearrangement` validates the standards compliance of AIRR Rearrangement data stored in a `data.frame`

Examples

```
# Get path to the rearrangement-example file
file <- system.file("extdata", "rearrangement-example.tsv.gz", package="airr")

# Load data file
df <- read_rearrangement(file)
```

```
# Validate a data.frame against the Rearrangement schema
validate_rearrangement(df)
```

```
[1] TRUE
```

validate_airr

Validate an AIRR Data Model nested list representation

Description

`validate_airr` validates the fields in a named nested list representation of the AIRR Data Model. Typically, generating by reading of JSON or YAML formatted AIRR files.

Usage

```
validate_airr(data, model = TRUE, each = FALSE)
```

Arguments

data

list containing records of an AIRR Data Model object imported from a YAML or JSON representation.

model

if TRUE validate only AIRR DataFile defined objects. If FALSE attempt validation of all objects in data.

each

if TRUE return a logical vector with results for each object in data instead of a single TRUE or FALSE value.

Value

Returns TRUE if the input data is compliant with AIRR standards and FALSE if not. If `each=TRUE` is set, then a vector with results for each object in data is returned instead.

Examples

```
# Get path to the rearrangement-example file
f1 <- system.file("extdata", "repertoire-example.yaml", package="airr")
f2 <- system.file("extdata", "germline-example.json", package="airr")

# Load data file
repertoire <- read_airr(f1)
germline <- read_airr(f2)

# Validate a single record
validate_airr(repertoire)
```

```
[1] TRUE
```

```
# Return validation for individual objects
validate_airr(germline, each=TRUE)
```

```
GenotypeSet  GermlineSet
      TRUE      TRUE
```

See also

See [Schema](#) for the AIRR schema definitions. See [read_airr](#) for loading AIRR Data Models from a file. See [write_airr](#) for writing AIRR Data Models to a file.

load_schema

Load a schema definition

Description

`load_schema` loads an AIRR object definition from the internal definition set.

Usage

```
load_schema(definition)
```

Arguments

definition

name of the schema definition.

Value

A [Schema](#) object for the definition.

Details

Valid definitions include:

- "Rearrangement"
- "Alignment"
- "Repertoire"
- "Study"
- "Subject"
- "Diagnosis"

- "Sample"
- "SampleProcessing"
- "DataProcessing"
- "GermlineSet"
- "GenotypeSet"

Examples

```
# Load the Rearrangement definition
schema <- load_schema("Rearrangement")

# Load the Repertoire definition
schema <- load_schema("Repertoire")
```

See also

See [Schema](#) for the return object.

Schema-class

S4 class defining an AIRR standard schema

Description

Schema defines a common data structure for AIRR Data Representation standards.

Usage

```
"names"(x)
```

```
"["(x, i)
```

```
"$(x, name)
```

```
InfoSchema
```

```
DataFileSchema
```

```
AlignmentSchema
```

```
RearrangementSchema
```

```
RepertoireSchema
```


GermlineSetSchema

GenotypeSetSchema

AIRRSchema

Arguments

x
Schema object.

i
field name.

name
field name.

Format

A Schema object.

An object of class Schema of length 1.

An object of class Schema of length 1.

An object of class Schema of length 1.

An object of class Schema of length 1.

An object of class Schema of length 1.

An object of class Schema of length 1.

An object of class Schema of length 1.

An object of class list of length 26.

Details

The following predefined Schema objects are defined:

InfoSchema: AIRR Info Schema.

DataFileSchema: AIRR DataFile Schema.

AlignmentSchema: AIRR Alignment Schema.

RearrangementSchema: AIRR Rearrangement Schema.

RepertoireSchema: AIRR Repertoire Schema.

GermlineSetSchema: AIRR GermlineSet Schema.

GenotypeSetSchema: AIRR GenotypeSet Schema.

AIRRSchema: named list containing all non-experimental AIRR Schema objects.

Slots

definition

name of the schema definition.

required

character vector of required fields.

optional

character vector of non-required fields.

properties

list of field definitions.

info

list schema information.

See also

See `load_schema` for loading a Schema from the definition set.

ExampleData

Example AIRR data

Description

Example data files compliant with the the AIRR Data Representation standards.

Format

- `extdata/rearrangement-example.tsv.gz`: Rearrangement TSV file.
- `extdata/repertoire-example.yaml`: Repertoire YAML file.
- `extdata/germline-example.json`: GermlineSet and GenotypeSet JSON file.

Examples

```
# Load Rearrangement example
file <- system.file("extdata", "rearrangement-example.tsv.gz", package="airr")
rearrangement <- read_rearrangement(file)
```

```
# Load Repertoire example
file <- system.file("extdata", "repertoire-example.yaml", package="airr")
repertoire <- read_airr(file)
```

```
# Load GermlineSet and GenotypeSet examples
file <- system.file("extdata", "germline-example.json", package="airr")
germline <- read_airr(file)
```

Release Notes

Version 1.5.0: August 29, 2023

- Updated schema set and examples to v1.5.

Version 1.4.1: August 27, 2022

Significant internal refactoring to improve schema generalizability, harmonize behavior between the python and R libraries, and prepare for AIRR Standards v2.0.

Rearrangement:

- Added the `aux_types` argument to `read_tabular`, `read_rearrangement`, and `read_alignment` to allow explicit declaration of the type for fields that are not defined in the schema.
- Renamed `read_airr`, `write_airr`, and `validate_airr` to `read_tabular`, `validate_tabular`, and `validate_tabular`, respectively.

Data Model and Schema:

- Defined new `read_airr`, `write_airr`, and `validate_airr` functions that support AIRR Data Model files that store arrays of objects in JSON or YAML.
- Added support for the AIRR Model Data File and associated schema (`DataFile`, `Info`). The Data File data format holds AIRR object of multiple types and is backwards compatible with Repertoire metadata.
- Added support for the new germline and genotyping schema (`GermlineSet`, `GenotypeSet`) and associated schema.

Version 1.3.0: May 26, 2020

- Updated schema set to v1.3.
- Added `info` slot to `Schema` object containing general schema information.

Version 1.2.0: August 17, 2018

- Updated schema set to v1.2.
- Changed defaults to `base="1"` for read and write functions.
- Updated example TSV file with coordinate changes, addition of `germline_alignment` data and simplification of `sequence_id` values.

Version 1.1.0: May 1, 2018

Initial release.

Download & Installation

To install the latest release from CRAN:

```
install.packages("airr")
```

To build from the [source code](#), first install the build dependencies:

```
install.packages(c("devtools", "roxygen2"))
```

To install the latest development code via devtools:

```
library(devtools)
install_github("airr-community/airr-standards/lang/R@master")
```

Note, using `install_github` will not build the documentation. To generate the documentation, clone the repository, and then build as normal using the following R commands from the package root `lang/R`:

```
library(devtools)
install_deps(dependencies=T)
document()
install()
```

Dependencies

Depends: FALSE

Imports: jsonlite, methods, readr, stats, stringi, tools, yaml

Suggests: knitr, rmarkdown, tibble, testthat

Authors

Jason Vander Heiden (aut, cre)

Susanna Marquez (aut)

Scott Christley (aut)

Katharina Imkeller (aut)

Ulrik Stervbo (aut)

AIRR Community (cph)

License

CC BY 4.0

2.5.2 Resources and Tools Supporting AIRR Standards

Applications Supporting the Rearrangement Schema

The following list of software tools and databases support the TSV format of v1.2 (or higher) of the *AIRR Rearrangement schema*.

Software	Version	Support	Reference
AIRR Python Library	1.2	Input, output and validation	Vander Heiden et al. Front Immunol, 2018.
AIRR R Library	1.2	Input, output and validation	Vander Heiden et al. Front Immunol, 2018.
Alakazam	1.0.1	Input and output	Gupta & Vander Heiden et al. Bioinformatics, 2015.
Cell Ranger	4.0	Output	10x Genomics, Inc. Pleasanton, CA USA.
Change-O	0.4.2	Input, output and conversion	Gupta & Vander Heiden et al. Bioinformatics, 2015.
compAIRR	0.2.0	Input	Rognes et al. Bioinformatics, 2022.
Decombinator	4.0.1	Output	Oakes et al. Front Immunol, 2017.
dowser	0.1.0	Input	Hoehn et al. PLoS Comput Biol, 2022.
IMGT/HighV-QUEST	1.7.0	Output	Alamyar et al. Methods Mol Biol, 2012.
IMGT/V-QUEST	3.5.16	Output	Giudicelli et al. Cold Spring Harb Protoc, 2011.
IgBLAST	1.11	Output	Ye et al. Nucleic Acids Res, 2013.
immunarch	0.6.5	Input	ImmunoMind Team. 2019
ImmuneDB	0.24.0	Output	Rosenfeld et al. Front Immunol, 2018.
immuneML	2.0.0	Input	Pavlovic et al. Nat Mach Intell, 2021.
immuneSIM	0.8.7	Output	Weber et al. Bioinformatics, 2020.
iReceptor	3.0	Input and output	Corrie et al. Immunol Rev, 2018.
MiXCR	4.0	Output	Bolotin et al. Nat Methods, 2015.
RAbHIT	0.1.5	Input and output	Gidoni et al. Nat Commun, 2019.
scirpy	0.7	Input and output	Sturm et al. Bioinformatics, 2020.
SCOPer	1.0.1	Input and output	Nouri & Kleinstein. Bioinformatics, 2018.
SHazaM	1.0.0	Input and output	Gupta & Vander Heiden et al. Bioinformatics, 2015.
SONAR	3	Output	Schramm et al. Front Immunol, 2016.
sumrep	1.0	Input	Olson et al. Front Immunol, 2019.
TIgGER	1.0.0	Input and output	Gadala-Maria et al. PNAS, 2015.
TRIGS	2	Input	Lees & Shepherd. J Immunol Res, 2015.
VDJServer	1.2.0	Input and output	Christley et al. Front Immunol, 2018
Vidjil-algo	2018.1	Output	Giraud et al. BMC Genomics, 2014.

AIRR Data Commons Repositories

A full list of AIRR Compliant repositories that implement the AIRR Data Commons API can be found at the [AIRR Community ADC Registry github site](#).

2.6 Community Resources

2.6.1 Resources and Tools Supporting AIRR Standards

AIRR Data Commons

The use of high-throughput sequencing for profiling B-cell and T-cell receptors has resulted in a rapid increase in data generation. It is timely, therefore, for the Adaptive Immune Receptor Repertoire (AIRR) community to establish a clear set of community-accepted data and metadata standards; analytical tools; and policies and practices for infrastructure to support data deposit, curation, storage, and use. Such actions are in accordance with international funder and journal policies that promote data deposition and data sharing – at a minimum, data on which scientific publications are based should be made available immediately on publication. Data deposit in publicly accessible databases ensures that published results may be validated. Such deposition also facilitates reuse of data for the generation of new hypotheses and new knowledge.

The AIRR Common Repository Working Group (CRWG) has developed a set of [recommendations](#) that promote the deposit, sharing, and use of AIRR sequence data. These recommendations were refined following community discussions at the AIRR 2016 and 2017 Community Meetings and were approved through a vote by the AIRR Community at the AIRR Community Meeting in December 2017. Updates to these recommendations have continued, with the latest set of Recommendations ratified at the AIRR Community meeting in May 2019.

In May 2020, the AIRR Community released the first version of the AIRR Data Commons Application Programming Interface (ADC API), a specification for programmatic access to query and download AIRR-seq data from repositories that adhere to the AIRR Standards. We define the AIRR Data Commons as consisting of the set of repositories that:

- adhere to the CRWG recommendations for promoting, sharing, and use of AIRR-seq data, and
- that implement the ADC API as a programmatic mechanism to access that data.

This page provides a central location for the community to discover resources that belong to the AIRR Data Commons.

AIRR Data Commons Repositories

The repositories that are part of the ADC are listed on the [AIRR Community ADC Registry github site](#). In order to find data across all of the repositories in the ADC, it is necessary to query all of these repositories.

Querying the AIRR Data Commons

Each of the repositories above can be queried directly using the [ADC API](#). In addition, the following tools and platforms implement web based user interfaces that use the ADC API to query repositories in the AIRR Data Commons:

- *iReceptor Gateway*
- *VDJServer Community Data Portal*

There are *query and analysis use cases* and *a set of example queries* available for the AIRR Data Commons and the ADC API.

2.6.2 Useful Websites for the AIRR Community

- [The AIRR Community](#)
- [B-T.CR Forum](#)
- [The AIRR Community GitHub](#)
- [The AIRR Standards GitHub Repository](#)
- [The AIRR Community Docker Hub](#)

2.7 Appendix A: Key Terms

The following table provides definitions for terms and abbreviations relevant to this documentation.

Table 4: Definitions and abbreviation

Term	Abbreviation	Retired synonyms	Definition
Adaptive Immune Receptor Repertoire	AIRR		A collection of vertebrate immune receptors whose loci typically undergo RAG-mediated rearrangement, i.e., immunoglobulins and T cell receptors.
AIRR Community	AIRR-C		The Adaptive Immune Receptor Repertoire Community is a grass-roots collaboration for developing and promoting best practices for generating, analyzing, curating, and sharing AIRR-seq data, see [DOI:10.3389/fimmu.2017.01418]. As of 12/2017, AIRR-C became a Committee within The Antibody Society (TAbs) and is therefore also known as “The AIRR Community Committee within The Antibody Society”.
AIRR Community Standards	AIRR-C Standards	AIRR Standards	The collection of all norms, including standards, guidelines, best practices, recommendations and definitions, which have been developed and published by the AIRR Community, in their respective current version.
AIRR Data Commons	ADC		The network of data repositories, which make AIRR-seq data FAIR by implementing the ADC API, adhering to the Mi-AIRR metadata standard and provide data in an AIRR Format.
AIRR Data Commons Application Program Interface	ADC API		The standard defining how AIRR-seq data located in a repository can be programmatically searched and retrieved.
AIRR Data Schema			The formal definition of all specified properties of an AIRR-seq study. This includes groupings

2.7. Appendix A: Key Terms

2.8 References

BIBLIOGRAPHY

- [LIGMDB_V12] IMGT-ONTOLOGY definitions. <<http://www.imgt.org/ligmdb/label#JUNCTION>>
- [INSDC_FT] The DDBJ/ENA/GenBank Feature Table Definition. <<http://www.insdc.org/documents/feature-table>>
- [ENA_MANUAL] European Nucleotide Archive Annotated/Assembled Sequences User Manual. <<http://ftp.ebi.ac.uk/pub/databases/ena/sequence/release/doc/usrman.txt>>
- [GENBANK_FF] GenBank Flat File Format. <<https://ftp.ncbi.nih.gov/genbank/gbrel.txt>>
- [GENBANK_SR] GenBank Sample Record. <<https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html>>
- [INSDC_XREF] Controlled vocabulary for /db_xref qualifier. <<http://www.insdc.org/documents/dbxref-qualifier-vocabulary>>
- [NCBI_NBK47528] SRA Handbook. <<https://www.ncbi.nlm.nih.gov/books/NBK47528/>>
- [RFC3987] Internationalized Resource Identifiers (IRIs). DOI:10.17487/RFC3987
- [RFC3987] Internationalized Resource Identifiers (IRIs). **DOI:10.17487/RFC3987**
- [Christley_2018] Christley S *et al.* VDJServer: A Cloud-Based Analysis Portal and Data Commons for Immune Repertoire Sequences and Rearrangements. *Front Immunol* 9:976 (2018) DOI: 10.3389/fimmu.2018.00976
- [Corrie_2018] Corrie *et al.* iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol Rev.* 2018 Jul;284(1):24-41. DOI: 10.1111/imr.12666
- [Ohlin_2019] Ohlin M *et al.* Inferred Allelic Variants of Immunoglobulin Receptor Genes: A System for Their Evaluation, Documentation, and Naming. *Front Immunol* 10:435 (2019) DOI: 10.3389/fimmu.2019.00435
- [Breden_2017] Breden F *et al.* Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. *Front Immunol* 8:1418 (2017). DOI: 10.3389/fimmu.2017.01418
- [Christley_2020] Christley S *et al.* The ADC API: a web API for the programmatic query of the AIRR Data Commons. *Front in Big Data* (2020). DOI: 10.3389/fdata.2020.00022
- [RFC2119] Key words for use in RFCs to Indicate Requirement Levels DOI: 10.17487/RFC2119
- [Rubelt_2017] Rubelt F *et al.* AIRR Community Recommendations for Sharing Immune Repertoire Sequencing Data. *Nat Immunol* 18:1274 (2017). DOI: 10.1038/ni.3873
- [VanderHeiden_2018] Vander Heiden JA *et al.* AIRR Community Standardized Representations for Annotated Immune Repertoires. *Front Immunol* 9:2206 (2018). DOI: 10.3389/fimmu.2018.02206
- [Wilkinson_2016] Wilkinson MD *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018 (2016). DOI: 10.1038/sdata.2016.18
- [Zenodo_1185414] Release archive of the AIRR Standards repository. (2015-2020). DOI: 10.5281/zenodo.1185414

Symbols

`__init__()` (*airr.io.RearrangementReader* method), 171
`__init__()` (*airr.io.RearrangementWriter* method), 172
`__iter__()` (*airr.io.RearrangementReader* method), 172
`__next__()` (*airr.io.RearrangementReader* method), 172
`--drop`
 airr-tools-merge command line option, 183
`--help`
 airr-tools command line option, 183
 airr-tools-merge command line option, 183
 airr-tools-validate command line option, 183
 airr-tools-validate-airr command line option, 184
 airr-tools-validate-rearrangement command line option, 184
 airr-tools-validate-repertoire command line option, 184
`--version`
 airr-tools command line option, 183
 airr-tools-merge command line option, 183
 airr-tools-validate command line option, 183
 airr-tools-validate-airr command line option, 184
 airr-tools-validate-rearrangement command line option, 184
 airr-tools-validate-repertoire command line option, 184
`-a`
 airr-tools-merge command line option, 183
 airr-tools-validate-airr command line option, 184
 airr-tools-validate-rearrangement command line option, 184
 airr-tools-validate-repertoire command line option, 184
`-h`
 airr-tools command line option, 183
 airr-tools-merge command line option, 183
 airr-tools-validate command line option, 183

airr-tools-validate-airr command line option, 184
 airr-tools-validate-rearrangement command line option, 184
 airr-tools-validate-repertoire command line option, 184
`-o`
 airr-tools-merge command line option, 183

A

airr-tools command line option
 `--help`, 183
 `--version`, 183
 `-h`, 183
airr-tools-merge command line option
 `--drop`, 183
 `--help`, 183
 `--version`, 183
 `-a`, 183
 `-h`, 183
 `-o`, 183
airr-tools-validate command line option
 `--help`, 183
 `--version`, 183
 `-h`, 183
airr-tools-validate-airr command line option
 `--help`, 184
 `--version`, 184
 `-a`, 184
 `-h`, 184
airr-tools-validate-rearrangement command line option
 `--help`, 184
 `--version`, 184
 `-a`, 184
 `-h`, 184
airr-tools-validate-repertoire command line option
 `--help`, 184
 `--version`, 184
 `-a`, 184

-h, 184

AlignmentSchema (in module *airr.schema*), 178

C

close() (*airr.io.RearrangementReader* method), 172

close() (*airr.io.RearrangementWriter* method), 173

create_rearrangement() (in module *airr*), 168

D

DataFileSchema (in module *airr.schema*), 177

definition (*airr.schema.Schema* attribute), 173

definition (in module *airr.schema*), 176–180

derive_rearrangement() (in module *airr*), 169

dump_rearrangement() (in module *airr*), 169

E

external_fields (*airr.io.RearrangementReader* attribute), 171

external_fields (*airr.io.RearrangementWriter* attribute), 172

F

false_values (*airr.schema.Schema* attribute), 173

false_values (in module *airr.schema*), 177–181

fields (*airr.io.RearrangementReader* attribute), 171

fields (*airr.io.RearrangementWriter* attribute), 172

from_bool() (*airr.schema.Schema* method), 174

G

GenotypeSetSchema (in module *airr.schema*), 180

GermlineSetSchema (in module *airr.schema*), 180

I

info (*airr.schema.Schema* attribute), 173

info (in module *airr.schema*), 176–180

InfoSchema (in module *airr.schema*), 176

L

load_rearrangement() (in module *airr*), 169

load_repertoire() (in module *airr*), 181

M

merge_rearrangement() (in module *airr*), 169

N

next() (*airr.io.RearrangementReader* method), 172

O

optional (*airr.schema.Schema* attribute), 173

optional (in module *airr.schema*), 177–181

P

pandas_types() (*airr.schema.Schema* method), 174

properties (*airr.schema.Schema* attribute), 173

properties (in module *airr.schema*), 176–181

R

read_airr() (in module *airr*), 170

read_rearrangement() (in module *airr*), 168

RearrangementReader (class in *airr.io*), 171

RearrangementSchema (in module *airr.schema*), 178

RearrangementWriter (class in *airr.io*), 172

repertoire_template() (in module *airr*), 182

RepertoireSchema (in module *airr.schema*), 179

required (*airr.schema.Schema* attribute), 173

required (in module *airr.schema*), 176–181

S

Schema (class in *airr.schema*), 173

spec() (*airr.schema.Schema* method), 174

T

template() (*airr.schema.Schema* method), 174

to_bool() (*airr.schema.Schema* method), 174

to_float() (*airr.schema.Schema* method), 175

to_int() (*airr.schema.Schema* method), 175

true_values (*airr.schema.Schema* attribute), 173

true_values (in module *airr.schema*), 177–181

type() (*airr.schema.Schema* method), 175

V

validate_airr() (in module *airr*), 171

validate_header() (*airr.schema.Schema* method), 175

validate_object() (*airr.schema.Schema* method), 176

validate_rearrangement() (in module *airr*), 170

validate_repertoire() (in module *airr*), 182

validate_row() (*airr.schema.Schema* method), 176

W

write() (*airr.io.RearrangementWriter* method), 173

write_airr() (in module *airr*), 170

write_repertoire() (in module *airr*), 181