AIRR Data Representation Reference Library Usage

The AIRR Community

2018-08-17

Contents

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 stablished 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 Representation 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"`.

```
library(airr)

example_data <- system.file("extdata", "rearrangement-example.tsv.gz", package="airr")
basename(example_data)
```

```
## [1] "rearrangement-example.tsv.gz"
```

```
airr_rearrangement <- read_rearrangement(example_data)
class(airr_rearrangement)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

```
head(airr_rearrangement)
```

```
## # A tibble: 6 x 33
##   sequence_id sequence rev_comp productive vj_in_frame stop_codon v_call
##   <chr>       <chr>    <lgl>    <lgl>      <lgl>       <lgl>      <chr>
## 1 SRR765688.~ NNNNNNN~ FALSE    TRUE       TRUE        FALSE      IGHV2~
## 2 SRR765688.~ NNNNNNN~ FALSE    TRUE       TRUE        FALSE      IGHV5~
## 3 SRR765688.~ NNNNNNN~ FALSE    TRUE       TRUE        FALSE      IGHV7~
## 4 SRR765688.~ NNNNNNN~ FALSE    TRUE       TRUE        FALSE      IGHV7~
## 5 SRR765688.~ NNNNNNN~ FALSE    TRUE       TRUE        FALSE      IGHV7~
## 6 SRR765688.~ NNNNNNN~ FALSE    FALSE      TRUE        TRUE       IGHV2~
## # ... with 26 more variables: d_call <chr>, j_call <chr>, c_call <chr>,
## #   sequence_alignment <chr>, germline_alignment <chr>, junction <chr>,
## #   junction_aa <chr>, v_cigar <chr>, d_cigar <chr>, j_cigar <chr>,
## #   v_sequence_start <int>, v_sequence_end <int>, v_germline_start <int>,
## #   v_germline_end <int>, d_sequence_start <int>, d_sequence_end <int>,
## #   d_germline_start <int>, d_germline_end <int>, j_sequence_start <int>,
## #   j_sequence_end <int>, j_germline_start <int>, j_germline_end <int>,
## #   junction_length <int>, np1_length <int>, np2_length <int>,
## #   duplicate_count <int>
```

Writing AIRR formatted files

The `airr` package contains the function `write_rearrangement` to write Rearrangement records to the AIRR TSV format.

```
out_file <- file.path(tempdir(), "airr_out.tsv")
write_rearrangement(airr_rearrangement, out_file)
```

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.