**AIRR-Seq Data Format**

AIRR-Seq data for upload to VDJbase is held in the repo <https://github.com/williamdlees/digby_data/>

The top-level directory AIRR-seq has a sub-directory for each species, and below that a sub-directory for each dataset. The layout for each dataset is standardised, as described below. Please refer to the Human/IGH data set as an example.

The dataset directory must contain two subdirectories: ‘reference’ and ‘samples’.

The ‘reference’ subdirectory must contain:

- gene\_order.py – a file that specifies the order in which genes in this dataset should be displayed, both locus order and alpha order.   
- ambiguous\_allele\_names.csv – a file that lists alleles associated with an ambiguous name. The format is described in a later section below.  
- germline reference files for the genes referred to in the samples. V genes must be IMGT gapped.

There can be any number of germline files with any names, as long as the extension is .fasta: for example V, D and J genes can be stored in separate files, or consolidated into a single file. The reference files must contain full-length sequences of all the non-inferred alleles used by the pipeline to process the samples: import to VDJbase will fail if any alleles used in the pipeline are not defined in the reference set, or if the sequence definitions differ. Note: **any** file in the reference subdirectory with a .fasta extension will be treated as containing reference genes, so only files genuinely containing them should be given this extension. Only full-length sequences should be provided in the reference sets.

Graphical user interface, text

Description automatically generated

The samples subdirectory should contain the samples, using the structure and naming as currently used in the repo for Human/IGH. The naming of directories and files within each sample directory must conform exactly to the standard used for Human/IGH.

The dataset directory *may* contain a projects.yml file, containing metadata for all the samples in the dataset. If this is present, it will be used. If not, the metadata will be taken from individual yml files in each sample directory. The import process will, in this case, construct a ‘consolidated.yml’ file in the samples directory, with the same content that projects.yml would have, if it were present. The content of the .yml file(s) must conform exactly to the standard as used for Human/IGH: any omissions or deviations from the standard are likely to cause the import process to fail.

**Format and use of ambiguous\_allele\_names.csv**

This file is used where some or all of the samples are not full-length. If all the samples are full-length, the file is not required.

For samples that are not full-length, some V-calls will be ambiguous. The pipeline uses a specific notation for these. For example, in Adaptive data sets, the alleles 01,02,03,04 of TRBV10-3 cannot be distinguished. The pipeline calls any matches to these alleles as TRBV10-3\*ap01. Within VDJbase, the corresponding label is TRBV10-3\*01\_02\_03\_04. ambiguous\_allele\_names.csv defines the mapping between these calls.

The file has three columns: GENE, PATTERN and ALLELES. Each ambiguous allele designation used by the pipeline must have a corresponding row in the file. In the case above, GENE would be TRBV10-3, PATTERN would be ap01, and ALLELES would be 01, 02, 03, 04.

Mappings in the file can also be used to est up ‘ambiguous gene’ calls. For example, in the Adaptive sets, the alleles TRBV6-2\*01 and TRBV6-3\*01 cannot be distinguished. The pipeline calls matches to these alleles as TRBV6-23. Within VDJbase, the corresponding label is TRBV6-2\*01\_3.01 and the gene itself is referred to as TRBV6-2/3 in reports. In this case, the row in the file will have GENE TRBV6-23, PATTERN 01, ALLELES 2.01, 3.01

Where a single allele is listed in a cell, if the file is opened in Excel the allele will be treated as an integer and the leading zero will be dropped. Because it’s convenient to manipulate the file in Excel, leading zeroes may be omitted: if VDJbase sees an allele number consisting of a single digit, it will add a leading zero.

Please note that it is not necessary to provide reference sequences for ‘shortened’ ambiguous alleles.

**Some Common Issues**

A short list of issues that can cause problems with VDJbase data import:

* Out-of-date Immcantarion tools. Change-O switched from using upper-case column headings to using mainly lower-case headings. If an out-of-date version is used, the import process won’t parse the files correctly.
* Sample data files that do not follow the naming convention. For example VDJbase extracts the anchor alleles from the name of the haplotype file. If the convention isn’t followed exactly, it won’t be able to work them out.
* Missing metadata: for example subjects or samples not being listed, or fields such as Age or Sex being omitted. The value of many fields can be blank if unknown, but the fields must be included.
* Wrongly-formatted metadata. In particular, values should only be quoted if they contain text: numeric values must not be quoted.