Next-generation germline sets for B-cell and T-cell repertoire sequencing

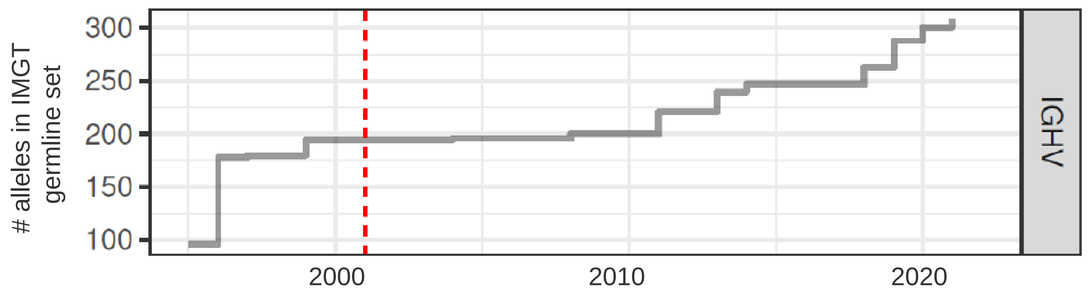
William Lees[1,2], Ayelet Peres[3], Justin Kos[4], Lindsay Cowell[5], Gur Yaari[3], Mats Ohlin[6], Andrew Collins[7], Corey T. Watson[4]

[1]Institute of Structural and Molecular Biology, Birkbeck College, London, UK [2]Institute for Systems and Computer Engineering, Technology and Science, Porto, Portugal,
[3]Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel, [4]Department of Biochemistry and Molecular Genetics, School of Medicine, University of Louisville, KY, United States,
[5]Department of Population and Data, University of Texas Southwestern Medical Center, Dallas, TX, United States, [6]Department of Immunotechnology, Lund University, Lund, Sweden,
[7]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

Correspondence: william@lees.org.uk

The Challenge

High-fidelity analysis of repertoire sequencing requires the use of accurate and comprehensive germline sets. Missing genes or alleles can impact downstream analysis: for example, the determination of mutation levels in B-cell analyses, or the inference of CDR1 and CDR2 sequences in short-read T-cell data. Today's sets are likely to be far from complete. Approx. 100 alleles have been added to the IMGT IGHV set since 2015, and the rate shows no sign of slowing.



Next-generation techniques, such as high-fidelity long-read sequencing, and the inference of alleles from AIRR-seq repertoires, make it possible to survey many 10s or 100s of subjects. Germline sets based on such studies have been published, but each publication tends to use a unique naming system, making it difficult to integrate information from multiple sources, and meaning that the same sequence can be known by multiple names. Official names are issued by the International Union of Immunological Societies, but the process is lengthy and requires information that may not be readily available from next-generation sources: for example, it may not be possible to map a discovered allele conclusively to a specific gene. In such a case, knowledge of the allele can still be useful in AIRR-seq analyses, for example for IG clonal determination.

Fields required in the official IUIS name may not be available at first discovery

IGHV1-69-2*01

- IGHV1 — Family number, determined phylogenetically: *Requires a representative enumeration of genes In the locus, so that the phylogeny can be established.*
- 69 — Identifier of the gene in the reference assembly: *Requires, at minimum, a good understanding of the structure of the locus, so that existence of the gene can be confirmed.*
- 2*01 — Allele number: *Requires confirmed mapping of the allele sequence to a specific genomic location.*

How can researchers collaborate to publish germline sets that are named consistently and evolve quickly as new information becomes available?

An Approach

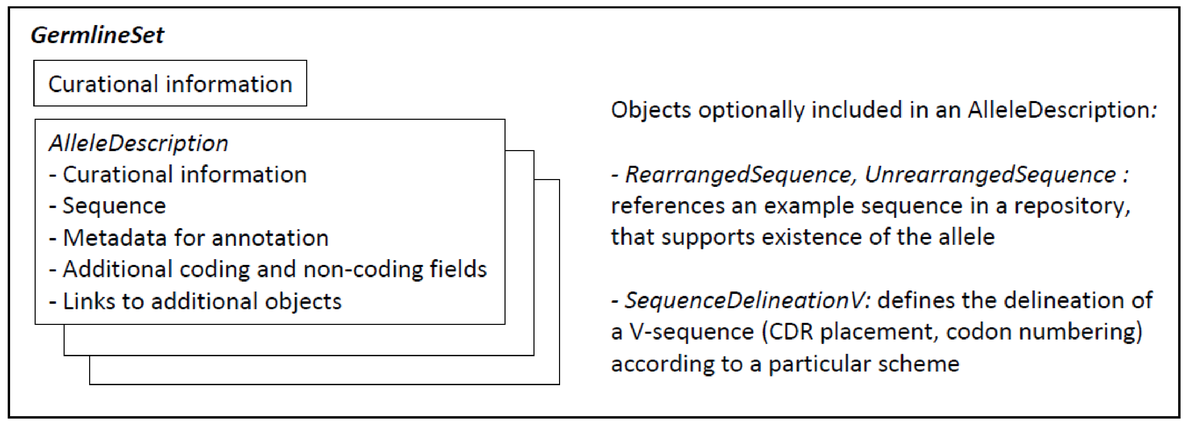
Here we describe an approach that has been developed within the Germline Database Working Group of the AIRR Community, and which is currently being applied to mouse and macaque strains from sequences and sets developed by its members. We intend to extend this work to other species, including humans. The working group is open to all with an interest in this work: please contact us if you are interested in joining.

As a first step, we assign each candidate germline sequence a 'temporary label' containing four letters or numbers assigned at random, for example IGHV-GY4A. The sequences are stored in a database, along with aliases denoting other names that they have been known by. This allows work from multiple researchers to be pooled, identifying duplicates and preserving traceability, while creating a consistent naming scheme. We have developed a simple tool, IgLabel, to allocate labels and maintain the database.

The AIRR-C Germline Set Schema

Once the database is populated, sequences can be selected to form a germline set. The selection is likely to be based on specific criteria, typically related to the degree of supporting evidence and can be fine-tuned for specific purposes. We term these selected sequences a *germline set*, in contrast to the *germline database*, which contains the complete collection of putative sequences, together with indicators of confidence. Germline sets may be selected to target specific sub-species or human populations.

The AIRR-C germline set schema defines a file format which fully describes the sequences, including fields that support annotation, and delineation in various numbering schemes.



The lifecycle of a sequence

As more information becomes available, a sequence can be confirmed as functional, and identified as an allele of a specific gene. The table below shows an illustrative progression to the point that official ratification is achieved.

3 sequences are submitted by researchers. The second is a sub-sequence of the first.		2 alleles are identified from the 3 sequences		The two alleles are established to be from the same gene.		Further confirmation becomes available and the sequences are sent to IUIS. IUIS ratifies the gene and alleles and assigns names	
sequence	**label**	**label**	**alias**	**label**	**alias**	**label**	**alias**
acacgta	IGHV-A5B2	IGHV-A5B2	IGHV-C89D	IGHV-A5B2*01	IGHV-C89D	IGHV3-189*01	IGHV-A5B2*01, IGHV-A5B2, IGHV-C89D
acgta	IGHV-C89D	IGHV-KZO5		IGHV-A5B2*02	IGHV-KZO5	IGHV3-189*02	IGHV-A5B2*02, IGHV-KZO5
acgaga	IGHV-KZO5						

Implications for Software Tools

The germline set schema is intended to contain all the information necessary to load a set into an annotation or analysis tool. We hope that authors of tools will use this information to simplify and streamline the process of updating sets within their tools, so that users can be encouraged to update frequently. We will shortly publish a tool for extracting information such as family number, gene identifier from the schema and adding it to an AIRR-standard alignment. We encourage all tool authors to move in this direction and accept a free-format name, rather than parsing information from the name on-the-fly.

Germline sets created by the AIRR-C Germline Database Working Group are published on OGRDB, and are downloadable via the web, or via a REST API. Currently OGRDB holds ~40 sets covering IG sequences for selected mouse strains. Sets for other species are in preparation.

Please use the QR code to access an online version of this poster with links to references