AIRR-C Meeting VI - 2022
GLDB WG Report

Current

Date of this report: May 2022
SC/WG Name: Germline Database WG
SC/WG Co-leaders: Andrew Collins, Corey Watson
SC/WG Active Members (list):

Pierre Boudinot, Steve Bosinger, Felix Breden, Christian Busse, Scott Christley, Andrew Collins (Co-lead), Martin Corcoran, Chris Cottrell, Jamie Heather, Gunilla Hedestam, Katherine Jackson, Justin Kos, William Lees, Susana Magadan, Mats Ohlin, Ayelet Peres, Oscar Rodriguez, Cathrine Scheepers, Chaim Schramm, Jamie Scott, Amit Upadhyay, Henk-Jan van den Ham, Corey Watson (Co-lead), Gur Yaari and Jian Ye

Purpose:
This Working Group (WG) was formed to promote the development of complete and accurate sets of reference germline IG and TCR genes, and to promote the accurate analysis and reporting of the germline genes that can be identified in repertoire studies. The WG works to establish processes for documenting novel germline genes and alleles and standards for versioned, inclusive databases. The WG also provides guidance on specific topics relating to data assessment that may be referred to it by the Inferred Allele Review Sub-committee (IARC).

Goals:

- Develop and seek ratification of universal principles for nomenclature systems that meet the needs of IG and TCR researchers
- Work to develop systems to document and report variation in non-coding regions of the IG/TR loci
- Develop appropriate IG Reference Sets for mouse strains and for rhesus macaque
- Establish an IARC focused on the review of TCR gene inferences
- Develop and benchmark tools and systems to improve capacity to fully leverage data types of the future, while maintaining the quality of References Sets
- Continue to work on questions regarding database versioning, programmatic access and licensing of germline reference databases

Products (if any):

- Release of new germline gene sets on OGRDB (see link). Together these represent a significant increase in available V, D, and J germline gene sequences for the mouse:
    - IGH germline genes (based on AIRR-seq inference) for the BALB/cByJ inbred mouse strain. (see publication Jackson et al.)
    - IGH germline genes (based on AIRR-seq inference) for four inbred wild-derived mouse strains representing diverse sub-species origins (CAST/EiJ, LEWES/EiJ, MSM/MsJ and PWD/PhJ), as well as the inbred strain NOD/ShiltJ. (see publication Watson et al.)
    - IGL and IGK germline genes (based on AIRR-seq inference) for 18 inbred mouse strains, including disease models and wild-derived strains representing diverse sub-species origins. (see preprint Kos et al.)
- Contribution of BALB/cByJ IGH germline reference set to IgBLAST (now available: https://www.ncbi.nlm.nih.gov/igblast/)
- Integrated IG germline database for the Rhesus macaque. This database consolidates germline gene/allele sequences from multiple sources, including KIMDB, RhGLDB, and IMGT. Critically, this database includes an extensive set of germlines that more fully represent diversity in this species. Although many germline sequences in this database have not yet formally received formal IUIS names, the database leverages our new temporary label system, and concepts from our germline database schema (see below) to ensure that sequences can receive usable identifiers to improve communication between research groups.
- Tool and registry for the assignment of temporary gene/allele/sequence identifiers for IG and TR sequences that are characterised/published, but do not meet current requirements of review by IARC or the IUIS nomenclature committee. This tool can be found here: https://github.com/williamdlees/IgLabel
- Development of a “new” germline database schema. This schema is being developed to improve the documentation of germline IG and TR sequences through multiple levels of curation. Critically, the schema allows for flexibility/adaptability in the future development of germline databases that draw on more diverse supporting data types (e.g., AIRR-seq, germline DNA sequencing, and other genomic datasets).
- Links between OGRDB and VDJbase to make it easier to track novel alleles discovered in VDJbase and submit them to OGRDB
- Enhancements to VDJbase to support the MiAIRR metadata standard and to hold genomic data alongside AIRR-seq
- Group directed/affiliated manuscripts:
    - A paper describing the future of germline gene databases, outlining the development of a new database schema, and providing short- and

long-term views for leveraging data from multiple sources and building more adaptable systems for IG/TR germline gene/allele curation and nomenclature. This manuscript is currently under consideration for AIRR-C endorsement.
    - A paper describing a BALB/c IGHV Reference Set was published in Frontiers in Immunology (link). Although not an official AIRR-C publication, the paper was the work of the Mouse subgroup of the GLDB WG. The Reference Set is now integrated with IgBLAST.

Resources (if any):

- Expanded features and resources at GLDB-WG affiliated databases:
    - VDJbase: https://vdjbase.org/ (see products list above)
    - OGRDB: https://ogrdb.airr-community.org/ (see products list above)

Progress report on current purpose, goals, products and resources:

- Develop and seek ratification of universal principles for nomenclature systems that meet the needs of IG and TCR researchers
    - Actions are underway to restructure governance of the the IG/TR/MH Nomenclature Sub-committee of IUIS's Nomenclature Committee. Until these issues are fully resolved, the GLDB-WG is reserving specific recommendations.
- Work to develop systems to document and report variation in non-coding regions of the IG/TR loci
    - There are now multiple tools available in the community to capture non-coding variation using either AIRR-seq data (e.g., 5'UTRs) and genomic data. Efforts to develop systems to catalogue and share such genetic variation have advanced considerably over the past interval. The focus has been on the development and implementation of tools/reporting features made available in VDJbase; this work has been driven by William Lees and the group of Gur Yaari (see vdjbase.org for more details). Specifically, VDJbase will soon offer users the ability to explore and analyse variant data available in both coding and non-coding regions of the IG and TCR gene regions.
- Develop appropriate IG Reference Sets for mouse strains and for Rhesus macaque
    - As outlined in the products section above, IG germline reference sets have been created for 18 mouse strains, and from a compilation of datasets for rhesus macaque, representing population level surveys of allelic variants. Mouse germline sets are currently available on OGRDB. Sources for complete germline sets for Rhesus will be shared in the coming weeks.

- Establish an IARC focused on the review of TCR gene inferences
    - This work has been transferred to the IARC; please see IARC 2022 report for updates on this goal.
- Develop and benchmark tools and systems to improve capacity to fully leverage data types of the future, while maintaining the quality of Reference Sets
    - This goal has been tackled at multiple levels. First, we have laid out a plan for a new germline database schema. This schema is designed to better facilitate the documentation of germline genes and alleles described for any species, including data sources, metadata, and identifiers. This schema is outlined in a manuscript currently undergoing AIRR-C endorsement procedures. As means to address this goal, the schema is structured to provide flexibility at multiple levels: 1) it more easily allows for the documenting of germline gene/alleles coming from a variety of data sources (e.g., AIRR-seq, genomic data, future data types.); 2) it allows for more seamless linking of sequence records and identifiers, promoting transparency and provenance.
    - Second, the reliance on genomic data for the application of existing germline gene/allele nomenclatures has limited our ability to efficiently name and share sequence sets curated from data that don't align with the "gold-standard" data types. For example, there is a growing collection of germline genes and alleles that have been discovered using non-traditional approaches (e.g., AIRR-seq), particularly in non-human species. In these cases, these germlines can currently neither be reviewed by IUIS nor IARC. And while, in the short-term these sets represent high-value data for the community, the effective sharing of these data between research groups is stilted by our inability to apply stable naming systems. Thus we developed an approach that will allow research groups to contribute to growing germline sets, and assign temporary gene/allele labels to newly discovered sequences in any species of interest. These identifiers are being managed by a registry system set up by the GLDB, in which interested participants can employ the registry to get stable unique gene/allele identifiers assigned to their germline sequences. This allows us to circumvent IUIS/IARC for the cataloguing of sequences for short-term utility, if and until such sequences can obtain formal names when genomic evidence is made available. Critically, these names can be utilised in conjunction with formal IUIS names within the database schema mentioned above.
    - Third, several members of the GLDB have been working on method development for curating IG/TR genes from high-throughput genomic data. While complete systems have not been developed yet for the review of such data, we expect these efforts to provide foundational resources for developing such capabilities moving forward (e.g., see progress at VDJbase noted above).

- Continue to work on questions regarding database versioning, programmatic access and licensing of germline reference databases

    - All three topics have seen some progress in the past Interval:
        - Database versioning: The new germline database schema discussed above foresees mechanisms that facilitate fine-grained and transparent updates of germline reference databases.
        - Programmatic access: Germline sets on OGRDB are accessible via a REST API. A defined AIRR Data Commons API will be added in the coming period.
        - Database licensing: In line with previous recommendations of Legal & Ethics WG, all data on ORGDB and VDJbase is licensed under a CC0 licence, allowing for completely unrestricted use and reuse. In addition, GLDB Participants have contributed various use cases to the upcoming detailed analysis of the EU Database Directive (96/9/EG), which has been conducted by Legal & Ethics WG.

Proposed plans for the coming interval:

Purpose:

This Working Group (WG) was formed to promote the development of complete and accurate sets of reference germline IG and TCR genes, and to promote the accurate analysis and reporting of the germline genes that can be identified in repertoire studies. The WG works to establish processes for documenting novel germline genes and alleles and standards for versioned, inclusive databases. The WG also provides guidance on specific topics relating to data assessment that may be referred to it by the Inferred Allele Review Sub-committee (IARC).

Goals:

- Continue to develop appropriate IG Reference Sets for human, various mouse strains, and for rhesus macaque, building upon recent efforts by integrating genomic data currently being generated by multiple groups. This integrated effort should result in robust germline sets for these species that could serve as useful models for other non-human organisms.
- Continue the development and implementation of the germline database schema.
- Collaborate with VDJbase to further develop features for cataloguing and sharing non-coding variation.

- Create an outreach subgroup to identify academic and commercial partners involved in the generation and use of germline data, as a means to encourage broader inclusion of key stakeholders. Specifically, we want to:
    - Improve the provision of datasets from additional species
    - Identify stakeholders and integrate them into community efforts
    - Promote the uptake of germline sets and schema in tools
    - Develop a plan for sustainability through ongoing funding

Long-term vision and how WG products integrate with the AIRR-C mission:

Enhance the accuracy and species coverage of AIRR-seq by providing comprehensive and regularly updated germline reference sets for species of interest. Leverage next-generation techniques (e.g. inference from AIRR-seq and long-read genomic sequencing) to make this possible. Provide this on a sustainable basis with respect to resourcing and funding.

Build understanding and awareness of the importance of comprehensive reference sets in AIRR-seq analysis. Promote the application of 'personalised' genotypes and haplotypes within AIRR-seq analysis, demonstrating the value they can add.

Develop and publish best practice in the application of current and future tools and methods for germline gene discovery.

Proposed SC/WG Co-leaders:

TBD