An Initiative to Develop *a More Complete and Accurate Inventory of Allelic Variants of BCR and TCR Genes*

Corey T. Watson[1], Andrew M. Collins[2], Christian E. Busse[3], William Lees[4], Cathrine Scheepers[5], Martin M. Corcoran[6], Mats Ohlin[7]

[1]University of Louisville School of Medicine, Louisville, KY, USA; [2]University of New South Wales, Sydney AUSTRALIA; German Cancer Research Center, Heidelberg, GERMANY; [4]University of London, Birkbeck, UNITED KINGDOM; [5]National Institute for Communicable Diseases, Johannesburg, SOUTH AFRICA; [6]Karolinska Institute, Stockholm, SWEDEN; [7]Lund University, Lund, SWEDEN

Background

Comprehensive germline immunoglobulin (IG) and T cell receptor (TCR) gene/allele reference collections are critical for the accurate analysis of adaptive immune receptor repertoire (AIRR) datasets. It is now clear, however, that existing germline gene databases (GLDBs) for humans and critical model species are neither complete nor fully accurate. Historically, the documentation of IG and TCR germline sets at the genomic level has been tedious and technically challenging. Recently, methods for profiling the expressed IG and TCR repertoires at great depth using high-throughput sequencing have opened the door to alternative approaches for the discovery of novel IG/TCR variable (V), diversity (D), and joining (J) genes and alleles by inference.

In 2015, the Adaptive Immune Receptor Repertoire (AIRR) Community was formed. This community-driven organization aims to organize and to coordinate stakeholders in the use of Next Generation Sequencing technologies to study IG and TCR repertoires. In 2016, the AIRR Community Germline Gene Database Working Group (GLDB WG) was formed to promote the complete and accurate description of germline IG and TCR genes, and their allelic variants, across species, strains, and populations.

At the 2017 meeting of the AIRR Community in Washington DC, the recommendations of the GLDB WG were adopted. These recommendations included the formation of an Inferred Allele Review Committee. This committee has now been formed, and has commenced the evaluation of inferred human IGHV alleles.

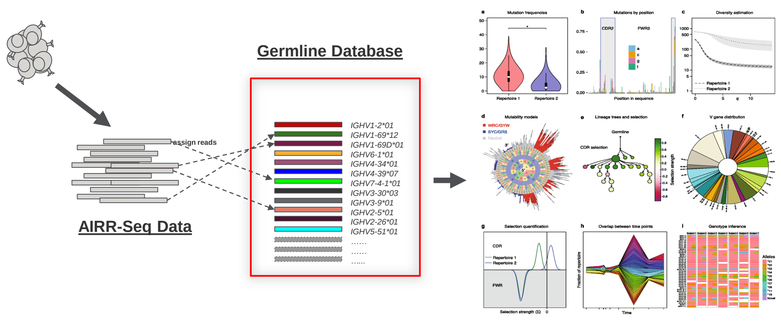

Figure 1. Almost every analysis of AIRR data requires alignment of V(D)J sequences against a database of Germline Genes. For analysis to be accurate, the Germline Database must be as complete and accurate as possible (right panel adapted from Yaari and Kleinstein, 2015).

Methods for inferring novel germline alleles from AIRR data

Methods have been developed in recent years which seek to infer germline genes and alleles present in AIRR sequencing data; these take various approaches that leverage sequence clustering, phylogenetic, and probabilistic methods. In all cases, the use of such tools can lead to the identification of alleles that are not currently found in existing GLDBs.

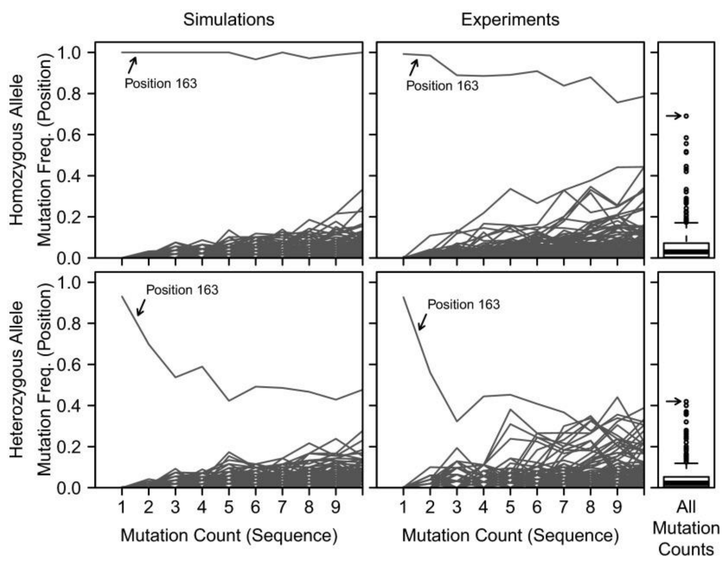

Figure 2. Figure from Gadala-Maria et al. (2015) demonstrating how TIgGER models and identifies putative IGHV polymorphisms observed in AIRR sequencing data.

Germline Gene Database Working Group

The MISSION of the Germline Gene Database Working Group of the AIRR Community:

To promote the comprehensive and accurate ***identification***, ***description***, ***classification***, ***annotation***, ***curation***, and ***consistent use*** of germline IG and TCR genes/alleles across species, strains, and populations.

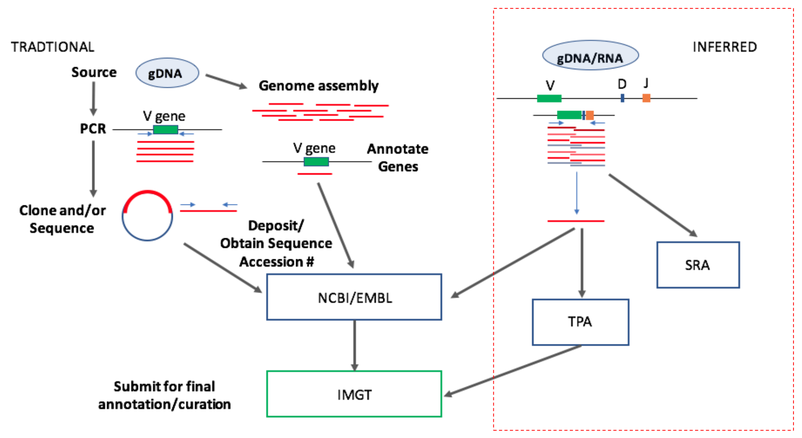

Figure 3: Databases like those of IMGT and NCBI are not presently set up to allow the deposition of inferred sequences.

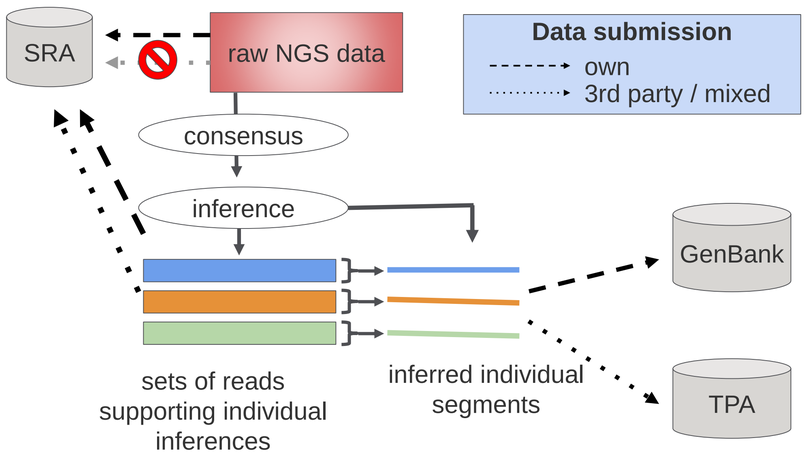

Figure 4. The GLDB WG is working to define processes by which inferred gene sequences can be submitted to SRA, Genbank and other databases.

Detecting and quantifying chimerism in AIRR data

A new initiative of the GLDB WG in 2018 has been the formation of a subgroup to raise awareness of the problem of chimeric sequences in AIRR data. Such sequences can compromise our ability to infer unreported polymorphisms in AIRR data, as well as degrading the quality of almost any analysis of such data. The subgroup aims to develop utilities for the identification of chimeric sequences and for their quantification within datasets.

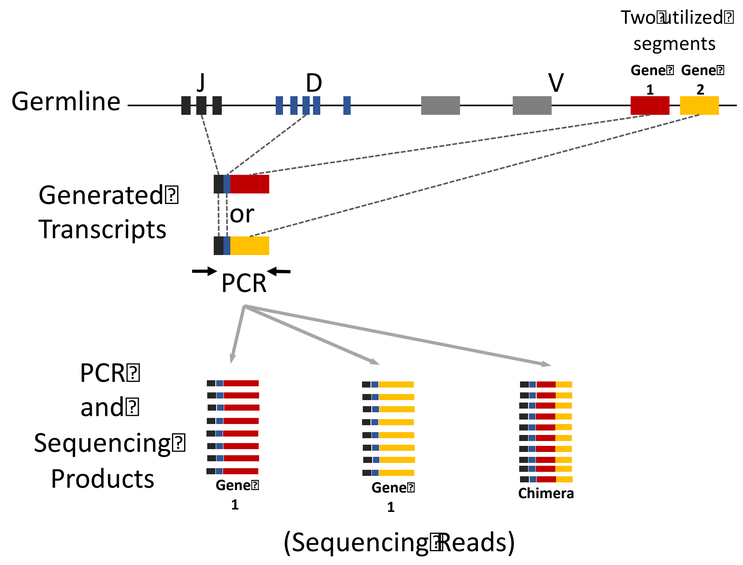

Figure 5. Illustration demonstrating the generation of PCR/Sequencing reads that include genuine representations of germline genes/alleles, as well as chimeric products that can masquerade as potential "novel" alleles.

The Inferred Allele Review Committee

The MISSION of the Inferred Allele Review Committee:

To review data in support of inferred alleles of IG and TCR genes, submitted by the research community, and to advise IMGT and the IMGT/IUIS/WHO BCR/TCR nomenclature committee of allele sequences that should be named and included in databases of germline BCR and TCR gene sequences.

In evaluating submitted IGHV inferences, the IARC considers a range of issues.

Inferences are increasingly identified using utilities such as IgDiscover, TIgGER, partis and ImPre. Identification of an inference using more than one utility gives added confidence in an inference.

Inferred sequences, without accumulated somatic point mutations, should be present at a relatively high frequency within the database.

Where an inferred IGHV sequence is identified in an individual who appears to be heterozygous at the locus of interest, the inferred sequence should be seen in at least 10% of all rearrangements of that gene.

Sets of rearranged genes utilizing an inferred IGHV sequence should include diverse IGHD and IGHJ genes.

In some situations, inferences will need to be confirmed using haplotype analysis.

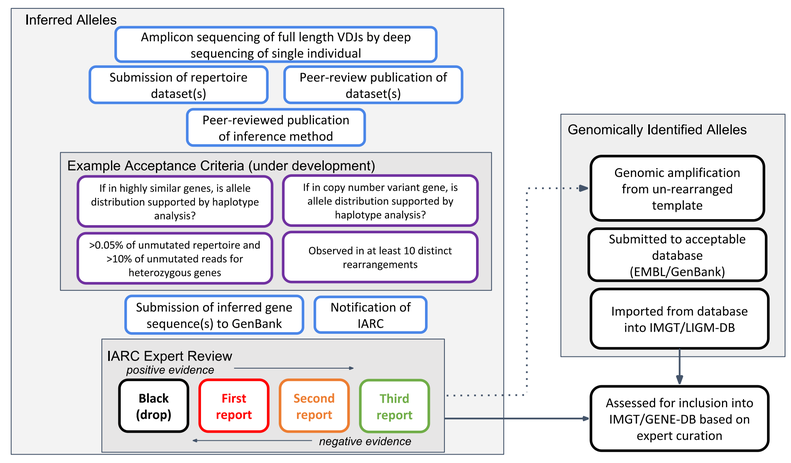

Figure 6. Flowchart depicting allele submission, review, and curation.

A Three Level System

When a submitted sequence is considered by the IARC, and its existence is affirmed, it will be designated as a Level 1 sequence.

If a Level 1 sequences is inferred in a second independent study, and it is again submitted for consideration by the IARC, and if its existence is again affirmed, it will be designated as a Level 2 sequence.

If a Level 2 sequences is inferred in a third independent study, and it is again submitted for consideration by the IARC, and if its existence is again affirmed, it will be designated as a Level 3 sequence.

The IARC will use the same evaluation process and the same standards in reaching each decision.

Submitting Sequences

Although the GLDB WG and IARC are still developing processes and procedures for sequence submission, they welcome approaches from researchers who may be interested in submitting sequences at a later date. Contact corey.watson@louisville.edu or a.collins@unsw.edu.au for further information