Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data

Florian Rubelt[1,21], Christian E Busse[2,21], Syed Ahmad Chan Bukhari[3,21], Jean-Philippe Bürckert[4], Encarnita Mariotti-Ferrandiz[5], Lindsay G Cowell[6], Corey T Watson[7], Nishanth Marthandan[8], William J Faison[9], Uri Hershberg[10], Uri Laserson[11], Brian D Corrie[12,13], Mark M Davis[1,14], Bjoern Peters[15], Marie-Paule Lefranc[16], Jamie K Scott[8,12,17], Felix Breden[12,13], The AIRR Community[18], Eline T Luning Prak[19,22] & Steven H Kleinstein[3,20,22]

High-throughput sequencing of B and T cell receptors is routinely being applied in studies of adaptive immunity. The Adaptive Immune Receptor Repertoire (AIRR) Community was formed in 2015 to address issues in AIRR sequencing studies, including the development of reporting standards for the sharing of data sets.

Antigen specificity is a cardinal feature of adaptive immunity that underlies immune homeostasis and control of pathogenic attack in higher vertebrates[1]. B and T cells are the two pillars of the adaptive immune system, and both express antigen-specific receptors at their surface, namely, B cell receptors (BCRs) and T cell receptors (TCRs), respectively. These receptors are produced by somatic gene-segment rearrangement that generates unique antigen-specific variable regions[2,3]. The collection of BCRs and TCRs in an individual forms the adaptive immune receptor repertoire (AIRR), which is capable of recognizing a vast array of antigens, including pathogens, autoantigens, allergens, toxins, and tumors[4,5]. For decades, characterization of the AIRR relied on low-resolution approaches such as flow cytometry, spectratyping, and Sanger sequencing. With the advent of high-throughput sequencing (HTS), it became possible to characterize the AIRR at unprecedented depth, with typical runs generating tens to hundreds of millions of receptor sequences[6]. Profiling of the AIRR with HTS (AIRR-seq) has since become an important part of basic and clinical immunology research, including vaccine design, therapeutic antibody discovery, minimal-residual disease detection, and monitoring of responses to therapy[4,6–8]. With high-throughput technologies generating large, complex data sets, AIRR-seq has led to the development of a diverse set of sample-processing strategies[9] and bioinformatics data analysis tools[10,11]. However, the ability to generate these data has outpaced the infrastructure available to manage it. Hundreds of studies are being published without common rules or standard procedures for the acquisition, storage, annotation, or sharing of the associated AIRR-seq data sets.

A full list of affiliations can be found at the end of the paper.

The AIRR Community

Efforts to organize the AIRR Community[12] formally began in September 2014 with a planning meeting held at the Interdisciplinary Research in the Mathematical and Computational Sciences Centre of Simon Fraser University (Burnaby, British Columbia, Canada). The meeting was initiated by Felix Breden, Jamie Scott, and Thomas Kepler to plan for a larger community meeting that would bring together a wide range of stakeholders interested in the use of HTS technologies to study BCR and TCR repertoires. The result was the first AIRR Community meeting, held in May/June 2015 in Vancouver, British Columbia, Canada, on the topic of “analysis, management, and sharing of antigen receptor repertoire sequence data.” This 2015 meeting successfully initiated a grassroots movement to develop guidelines for best practices for the generation, use, and sharing of AIRR-seq data, and to address challenges involved in the management and analysis of these large data sets. To this end, three working groups were formed: (1) Common Repository, (2) Minimal Standards, and (3) Tools & Resources (**Fig. 1**). These working groups spent the next year prioritizing goals and defining recommendations that were discussed at the second AIRR Community meeting, in June 2016, which was held at the US National Institutes of Health in Rockville, Maryland. The resulting set of ratified recommendations, along with video recordings of the discussions, is available on the AIRR Community website (http://airr-community.org). The third AIRR Community meeting will take place in December 2017, also at the US National Institutes of Health.

The three AIRR Community working groups meet regularly via teleconference, with many discussions continued online through the B-T.CR forum (https://b-t.cr/). A major goal of the Common Repository Working Group is to develop mechanisms to enable data sharing between multiple repositories that store AIRR-seq data, such as iReceptor (http://ireceptor.irmacs.sfu.ca/) and VDJServer (https://vdjserver.org/). This infrastructure will allow queries (such as a search for a particular CDR3 sequence) to span participating repositories. Along with technical hurdles, major issues involve legal constraints such as donor

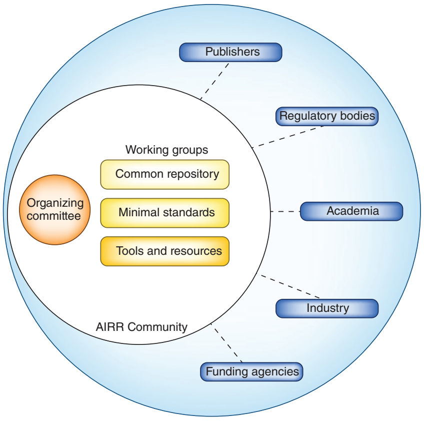

Figure 1 The organization of and interactions among the AIRR Community.

consent and privacy, as well as intellectual property concerns. The Tools & Resources Working Group has several ongoing initiatives, including (1) the establishment of a common file format that builds on previous efforts[13,14] to allow interoperability of AIRR-seq analysis tools; (2) improvements to existing germline gene/allele databases and nomenclature, including the development of a framework for reporting novel variable (V), diversity (D) and joining (J) alleles computationally inferred from AIRR-seq data[15,16] (such alleles do not meet current criteria for inclusion in the widely used International Immunogenetics Information system IMGT database); (3) the construction of a foundation for software tool validation, based on benchmarking of repertoire simulation tools with a variety of summary statistics to ensure that they accurately reflect the characteristics of biological data sets; and (4) the development of biological reference samples that can be used as controls for amplification and sequencing protocols in collaboration with industry and the US National Institute of Standards and Technology. Finally, the Minimal Standards Working Group is seeking to improve the reproducibility of AIRR-seq experiments and promote data sharing and reuse through the establishment of standards for depositing AIRR-seq data in the public domain. The result of this effort is the first release of the Minimal Information about Adaptive Immune Receptor Repertoire (MiAIRR) (pronounced “my air”) data standard described herein, and its implementation at the US National Center for Biotechnology Information (NCBI).

The MiAIRR data standard

The definition of a community-based standard for the reporting of experimental results has become vital for complex data-driven research as a way to enhance reproducibility and allow efficient data sharing and comparison[17]. Minimum Information for Biological and Biomedical Investigations (MIBBI) provides a portal (https://fairsharing.org/collection/MIBBI) with links to nearly 40 minimum-information checklists for several biological disciplines, including the well-known Minimum Information about a Microarray Experiment standard (MIAME). The MIAME guidelines were proposed in 2001[18] and, within eight years, were widely accepted by journals, with >10,000 studies deposited into MIAME-compliant databases[19]. Ample experience with repositories such as GEO (https://www.ncbi.nlm.nih.gov/geo), TCGA (https://cancergenome.nih.gov/), and ImmPort[20] shows the benefits of a consistent strategy for annotation, storage, and availability.

We hereby propose the MiAIRR standard, which we consider necessary for the interpretation and comparison of AIRR-seq experiments conducted by different groups. A draft MiAIRR proposal was approved at the second AIRR Community meeting, and the MiAIRR standard described here presents a refinement of this initial proposal that is the result of a year-long collaboration by database curators, informaticians, and biologists from the AIRR Community Minimal Standards Working Group. MiAIRR has been endorsed for this publication by the authors and the listed members of the AIRR Community.

The standard consists of six high-level sets (**Fig. 2**), with each set providing information on a distinct aspect of the study. By design, the combined information in these six sets should allow a researcher skilled in the art of AIRR-seq data generation and analysis to reproduce the results of the study. The MiAIRR standard does not stipulate how samples should be obtained, selected, processed, sequenced, or analyzed. The field of repertoire analysis is evolving rapidly, and the MiAIRR standard is intended to facilitate data sharing, rather than constrain experimental or analytical methods. Nevertheless, following the MiAIRR standards during the design phase of an experiment could help scientists determine whether they have captured all of the essential information about their experiment. Furthermore, incorporation of these standards might enhance the robustness of the experimental design and data collection, thereby improving the quality and significance of the planned study.

The six high-level MiAIRR sets are (1) Study, Subject and Diagnosis; (2) Sample Collection; (3) Sample Processing and Sequencing; (4) Raw Sequences; (5) Data Processing; and (6) Processed Sequences with Annotations (**Fig. 2**). The MiAIRR sets are arranged in chronological order from study design to data generation. Their focus is primarily on experimental setup and data generation, rather than data analysis. As with gene expression microarrays, it is presumed that subsequent analyses will most often use the raw or minimally processed data from several different studies. Because the immune repertoire is influenced by genetic factors, environmental conditions, and disease history, the value of AIRR-seq data increases dramatically if sufficient information is included. In addition, the inclusion of basic sequence annotations, such as the VDJ gene segment and junction sequence assignments, will enable rapid querying of the data. To ensure that MiAIRR is discoverable by third parties, and especially those outside our domain, we have registered the standard in FAIRsharing (bsg-s000689).

We briefly discuss each of the six MiAIRR sets below; the complete list of data fields associated with each set is shown in **Figure 2**. A tabular and more detailed version of the element list is available through the AIRR Community website. The site also provides information about ongoing efforts by the Minimal Standards and Tools & Resources Working Groups to map these data fields to controlled vocabularies and specify the data types and formats for each of the fields.

Set 1 describes the experimental study design, including the title of the study, laboratory contact information, funding, linked publications, and information about the study cohorts and individual subjects. For individual subjects, the species, sex, age, and ancestry are included, along with information about disease state(s), therapies, and study group membership (e.g., control versus disease). Although “subject” typically refers to an individual from which a sample is derived, the standard also allows for synthetic (e.g., phage-based) libraries. For collection and reporting of this information, the use of controlled vocabularies is in many cases possible and recommended.

Set 2 includes information about the origin and expected composition of the biological sample(s). This set aims to capture essential information about the collection of a sample, including its source (e.g., anatomical site), its provenance (provider), and the experimental condition (e.g., the time point during the course of a disease or treatment).

Set 3 includes information on sample processing and sequencing. Repertoire properties can vary dramatically between cell types (e.g., naive versus memory). This set includes the cell subset being profiled, as defined by the investigator, and the flow cytometry or other markers used to select the subset. The latter field can provide objective information on the identity of a cell population (for example, when designations change over time, are inconsis-

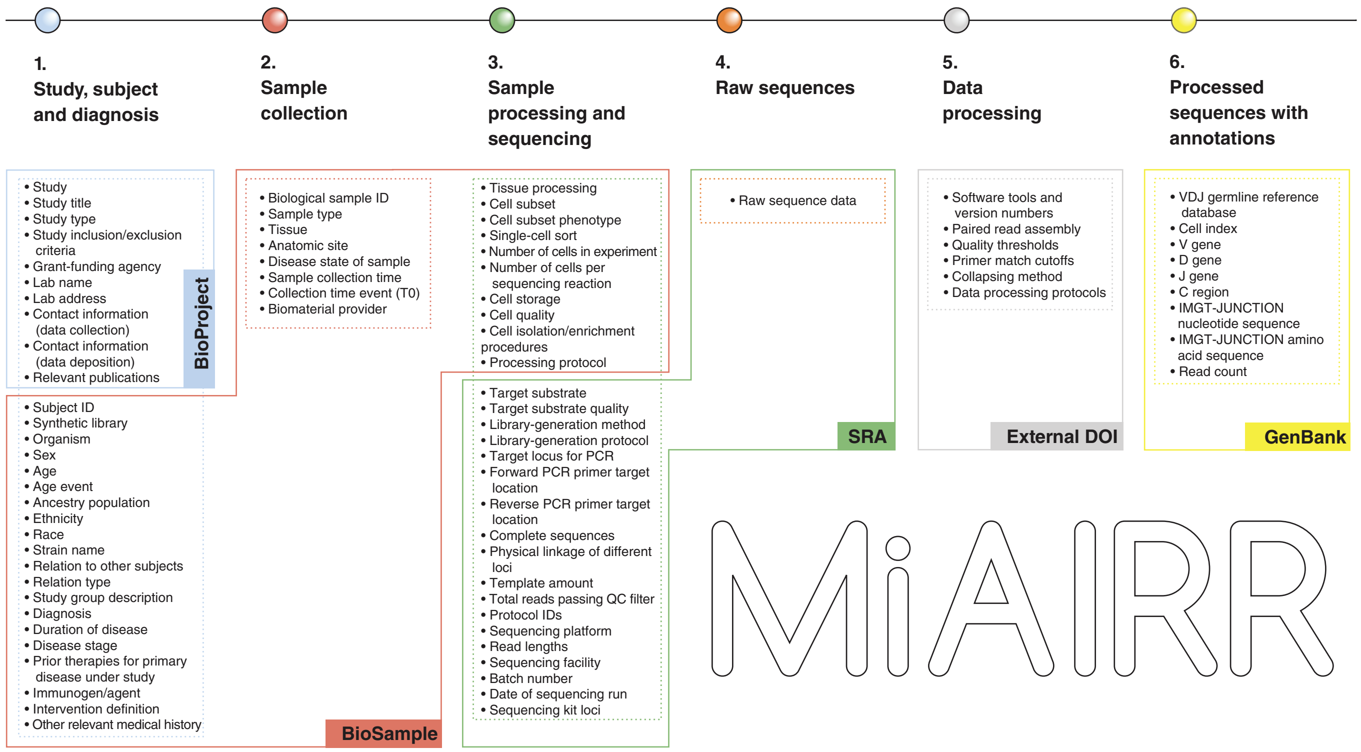

Figure 2 An overview of the six MiAIRR sets and associated data fields, along with their target submission repositories at NCBI.

tently used, or are controversial). Additional information includes the number of cells per sample and whether cells were prepared in bulk or captured as single cells, as well as the nucleic acid sample type (e.g., RNA versus DNA) and how immune-receptor gene rearrangements were amplified and sequenced (for example, RACE-PCR versus multiplex PCR, paired PCR, and/or varying read length and sequencing chemistries). Annotation of the amplicons being sequenced should be described in sufficient detail to allow the raw sequencing data (set 4) to be transformed into processed sequences (set 6) suitable for downstream analyses, such as VDJ annotation and lineage construction. This includes any sample-specific sequences or barcodes, primer sequences and locations, unique molecular identifiers, and so on. Information about the sequencing run, such as the number of reads, read lengths, quality control parameters, the sequencing kit and instrument(s) used, and run batch number, is also included.

Set 4 provides the raw HTS data for each sequencing run (e.g., FASTQ files) and permits the reanalysis, secondary analysis, and combination of multiple data sets from different studies via meta-analysis techniques. The inclusion of raw data allows the most up-to-date data processing to be carried out, as analysis tools for AIRR-seq data are undergoing rapid evolution[7,10,11]. The raw sequence data linked to the information mentioned in the previous sections form the basis for the analysis of the repertoire data.

Set 5 includes information on data processing. The MiAIRR standard requires that sufficient information be provided to allow the raw sequencing data in set 4 to be transformed into the processed sequencing data provided in set 6. Because of the wide variety of sequencing and processing methods in use, as well as rapid innovation in this area, we have not attempted to define detailed data elements to capture all of this information. Rather, MiAIRR defines broad categories that cover the essential data-processing steps (for example, the software tools with version numbers, quality thresholds, and primer match and length cutoffs). For example, if V, D, and J gene segments and/or isotypes are reported, the software or method used to assign these annotations should also be described.

Finally, set 6 consists of the processed sequences with annotations. Rapid querying of AIRR-seq data to screen for recurring sequences, or junction motifs, linked to a given stimulus or medical condition is an important aspect of AIRR research. This MiAIRR set comprises the list of processed sequences along with sequence-level annotations. This includes the VDJ gene segment and constant region (isotype) annotation if used in the associated publication, along with the junction sequence (i.e., CDR3 plus the conserved flanking amino acid residues).

Leveraging NCBI resources

NCBI hosts a collection of databases, tools, and services to address the increasing need for data archival and analysis in biomedicine. To leverage their robust infrastructure and long-term support, we have developed a specification for how the MiAIRR standard can be implemented with NCBI's existing data repositories. The six MiAIRR sets include metadata and data that span four NCBI repositories: BioProject, BioSample, Sequence Read Archive (SRA), and GenBank (**Fig. 2**). Study, subject, and sample information (sets 1–3) is submitted to BioProject and BioSample, and the sequencing information and linked raw sequencing data (sets 3 and 4) are submitted to SRA. Because processed sequencing data (set 6) are beyond the scope of SRA, they are submitted to GenBank with a link to the associated BioProject ID. The information describing how raw sequencing data were processed to generate the GenBank submission (set 5) does not have a natural home in the current NCBI framework. To capture this information, metadata associated with set 5 are submitted to a digital object identifier (DOI)-granting service, and the resulting DOI is stored with the NCBI submission. Thus, all of the information required by

MiAIRR is linked together and made available to the wider scientific community for search and download.

In practice, the process for submission to NCBI's public data repositories consists of five sequential steps. First, study information is submitted to BioProject via the NCBI web interface, which currently captures all of the information required by MiAIRR. Second, sample-level information is submitted to the NCBI BioSample repository in AIRR-specific templates. These data templates are made available in Excel and XML formats to support individual and bulk submissions, respectively. Third, raw sequencing data are uploaded to SRA, again in AIRR-specific data templates. Fourth, a DOI is generated for the protocol that describes how raw sequencing data were processed. This DOI can be generated by Zenodo (https://zenodo.org) or an equivalent DOI-granting service. Finally, the processed sequencing data and protocol DOI are submitted to GenBank. Sequence-level annotations in MiAIRR are captured in GenBank via the International Nucleotide Sequence Database Collaboration (INSDC)[21] feature table. INSDC currently provides feature tags that cover most of the data elements required by MiARR, including VDJ gene annotations. AIRR-specific information, such as the V(D)J junction region, is specified using custom keywords. The AIRR data templates to support these NCBI submissions, along with detailed step-by-step instructions, are available through the AIRR Community website (http://airr-community.org/working_groups/minimal_standards).

Conclusion

High-throughput AIRR sequencing is revolutionizing the investigation of adaptive immunity. Through annual meetings and standing working groups, the AIRR Community[12] is bringing together stakeholders to address a broad range of questions inherent to the generation, analysis, comparison, and sharing of data on adaptive immune repertoires. The AIRR Community consists of researchers who produce and use BCR and TCR data; industrial partners; statisticians; bioinformaticians; data security experts; and scholars in the relevant disciplines of ethics, law, and policy. One of the primary initiatives of the AIRR Community has been to develop a set of standards for the reporting and sharing of AIRR-seq data. Journals have different (or no) requirements for providing AIRR-seq data to readers. Although funding agencies are beginning to require the sharing of genomic data[22,23], no such requirements currently apply to AIRR-seq data. Differences in data quality, analysis, curation, and storage impose difficulties for data sharing, and this in turn makes it impossible to fully exploit the potential of AIRR-seq data. The MiAIRR data standard described here is a first step that will promote reproducibility, as well as secondary and meta-analysis. We do not consider MiAIRR to be a static definition. It is likely that as the field of AIRR-seq continues to develop, new data elements will gain importance, whereas others will become irrelevant. The AIRR Community is committed to the long-term maintenance of MiAIRR and will continue to update and revise the standard according to the needs of the larger community.

The MiAIRR standard provides a checklist of data elements to include with AIRR study data submissions. Along with this standard, we have described a specification for implementing this standard by combining four NCBI repositories (BioProject, BioSample, SRA, and GenBank). We envision that other repositories will also provide mechanisms for AIRR-seq data submissions in the future. For example, an implementation that used the NCBI Database of Genotypes and Phenotypes or the European Genome-phenome Archive (https://ega-archive.org/) would allow for sharing of data that require the protection of human subjects, and an implementation that involved the European Nucleotide Archive (https://www.ebi.ac.uk/ena) could address EU-specific legal concerns. Additional work is also required to make the data-submission process more user-friendly. To this end, we are working with the Center for Expanded Data Annotation and Retrieval (https://metadatacenter.org/) to develop a unified web interface for the submission of ontology-constrained AIRR-seq metadata and data to NCBI. The adoption of a common standard for AIRR-seq studies will lay the foundation for a unified environment of data sets and analysis tools, thereby creating economies of scale for individual researchers.

We invite all interested parties to join the AIRR Community, to attend the annual meetings, and to participate in one or more of the AIRR Community working groups. We hope that readers of this Comment will use the MiAIRR standard, and encourage their publishers to require authors to use it, for AIRR-seq data submission and sharing. These efforts will benefit the entire community as a step toward making AIRR studies findable, accessible, interoperable, and reusable (FAIR)[24].

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Litman, G.W., Cannon, J.P. & Dishaw, L.J. *Nat. Rev. Immunol.* **5**, 866–879 (2005).
2. Tonegawa, S. *Nature* **302**, 575–581 (1983).
3. Davis, M.M. & Bjorkman, P.J. *Nature* **334**, 395–402 (1988).
4. Liu, X.S. & Mardis, E.R. *Cell* **168**, 600–612 (2017).
5. Hou, D., Chen, C., Seely, E.J., Chen, S. & Song, Y. *Front. Immunol.* **7**, 336 (2016).
6. Georgiou, G. *et al. Nat. Biotechnol.* **32**, 158–168 (2014).
7. Wardemann, H. & Busse, C.E. *Trends Immunol.* **38**, 471–482 (2017).
8. Burel, J.G., Apte, S.H. & Doolan, D.L. *Trends Immunol.* **37**, 53–67 (2016).
9. Friedensohn, S., Khan, T.A. & Reddy, S.T. *Trends Biotechnol.* **35**, 203–214 (2017).
10. Yaari, G. & Kleinstein, S.H. *Genome Med.* **7**, 121 (2015).
11. Greiff, V., Miho, E., Menzel, U. & Reddy, S.T. *Trends Immunol.* **36**, 738–749 (2015).
12. Breden, F. *et al. Front. Immunol.* http://dx.doi.org/10.3389/fimmu.2017.01418 (2017).
13. Toby, I.T. *et al. BMC Bioinformatics* **17**, 333 (2016).
14. Gupta, N.T. *et al. Bioinformatics* **31**, 3356–3358 (2015).
15. Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S.H. *Proc. Natl. Acad. Sci. USA* **112**, E862–E870 (2015).
16. Corcoran, M.M. *et al. Nat. Commun.* **7**, 13642 (2016).
17. Taylor, C.F. *et al. Nat. Biotechnol.* **26**, 889–896 (2008).
18. Brazma, A. *et al. Nat. Genet.* **29**, 365–371 (2001).
19. Brazma, A. *ScientificWorldJournal* **9**, 420–423 (2009).
20. Bhattacharya, S. *et al. Immunol. Res.* **58**, 234–239 (2014).
21. Nakamura, Y., Cochrane, G. & Karsch-Mizrachi, I. *Nucleic Acids Res.* **41**, D21–D24 (2013).
22. Contreras, J.L. *Trends Genet.* **31**, 55–57 (2015).
23. European Commission. *H2020 Program: Guidelines on FAIR Data Management in Horizon 2020, Version 3.0.* (European Commission, 2016).
24. Wilkinson, M.D. *et al. Sci. Data* **3**, 160018 (2016).

[1]Department of Microbiology and Immunology and Institute for Immunity, Transplantation and Infection, Stanford University School of Medicine, Stanford, California, USA. [2]Division of B Cell Immunology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [3]Department of Pathology, Yale School of Medicine, New Haven, Connecticut, USA. [4]Department of Infection and Immunity, Luxembourg Institute of Health, Luxembourg, Luxembourg. [5]Sorbonne Universités, UPMC Univ Paris 06, INSERM, UMR_S 959, Immunology-Immunopathology-Immunotherapy (i3), Paris, France. [6]Department of Clinical Sciences, UT Southwestern Medical Center, Dallas, Texas, USA. [7]Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, Kentucky, USA. [8]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada. [9]Duke Human Vaccine Institute, Duke University School of Medicine, Durham, North Carolina, USA. [10]School of Biomedical Engineering, Science & Health Systems, and Department of Microbiology and Immunology, College of Medicine, Drexel University, Philadelphia, Pennsylvania, USA. [11]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA. [12]iReceptor, Simon Fraser University, Burnaby, British Columbia, Canada. [13]Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia, Canada. [14]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA. [15]La Jolla Institute for Allergy and Immunology, La Jolla, California, USA. [16]IMGT, the international ImMunoGeneTics information system, LIGM, Institut de Génétique Humaine IGH, CNRS, University of Montpellier, Montpellier, France. [17]Faculty of Health Sciences, Simon Fraser University, Burnaby, British Columbia, Canada. [18]A list of members endorsing this manuscript and their affiliations is available in **Supplementary Note 1**. [19]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [20]Department of Immunobiology, Yale School of Medicine, and Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, USA. [21]These authors contributed equally to this work [22]These authors jointly supervised this work.
e-mail: luning@pennmedicine.upenn.edu or steven.kleinstein@yale.edu