eLife

REVIEW ARTICLE

Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling

Johannes Trück[1†], Anne Eugster[2†], Pierre Barennes[3,4†], Christopher M Tipton[5], Eline T Luning Prak[6], Davide Bagnara[7], Cinque Soto[8,9], Jacob S Sherkow[10,11,12], Aimee S Payne[6], Marie-Paule Lefranc[13,14,15], Andrew Farmer[16], The AIRR Community, Magnolia Bostick[16‡§*], Encarnita Mariotti-Ferrandiz[3‡*]

[1]University Children's Hospital and the Children's Research Center, University of Zurich, Zurich, Switzerland; [2]CRTD Center for Regenerative Therapies Dresden, Faculty of Medicine, Technische Universität Dresden, Dresden, Germany; [3]Sorbonne Université U959, Immunology-Immunopathology-Immunotherapy (i3), Paris, France; [4]AP-HP Hôpital Pitié-Salpêtrière, Biotherapy (CIC-BTi), Paris, France; [5]Lowance Center for Human Immunology, Emory University School of Medicine, Atlanta, United States; [6]Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States; [7]University of Genoa, Department of Experimental Medicine, Genoa, Italy; [8]The Vanderbilt Vaccine Center, Vanderbilt University Medical Center, Nashville, United States; [9]Department of Pediatrics, Vanderbilt University Medical Center, Nashville, United States; [10]College of Law, University of Illinois, Champaign, United States; [11]Center for Advanced Studies in Biomedical Innovation Law, University of Copenhagen Faculty of Law, Copenhagen, Denmark; [12]Carl R. Woese Institute for Genomic Biology, University of Illinois, Urbana, Illinois, United States; [13]IMGT, The International ImMunoGeneTics Information System (IMGT), Laboratoire d'ImmunoGénétique Moléculaire (LIGM), Institut de Génétique Humaine (IGH), CNRS, University of Montpellier, Montpellier, France; [14] Laboratoire d'ImmunoGénétique Moléculaire (LIGM) CNRS, University of Montpellier, Montpellier, France; [15]Institut de Génétique Humaine (IGH), CNRS, University of Montpellier, Montpellier, France; [16]Takara Bio USA, Inc., Mountain View, United States

***For correspondence:** magnolia.bostick@gmail.com (MB); encarnita.mariotti@sorbonne-universite.fr (EM-F)

[†]These authors contributed equally to this work
[‡]These authors also contributed equally to this work

Present address: [§]Process Development, PACTPharma, Inc, South San Francisco, United States

Competing interest: See page 13

Funding: See page 13

Received: 13 January 2021
Accepted: 15 May 2021
Published: 26 May 2021

Reviewing editor: Lindsay Cowell, , United States

© Copyright Trück et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Abstract Use of adaptive immune receptor repertoire sequencing (AIRR-seq) has become widespread, providing new insights into the immune system with potential broad clinical and diagnostic applications. However, like many high-throughput technologies, it comes with several problems, and the AIRR Community was established to understand and help solve them. We, the AIRR Community's Biological Resources Working Group, have surveyed scientists about the need for standards and controls in generating and annotating AIRR-seq data. Here, we review the current status of AIRR-seq, provide the results of our survey, and based on them, offer recommendations for developing AIRR-seq standards and controls, including future work.

Introduction

Immunoglobulin chains (IG) and T-cell receptor chains (TR) are generated by DNA recombination, a process of somatic rearrangement of variable (V), diversity (D), and joining (J) genes

(**Tonegawa, 1983**). The diversity of the resulting rearranged genes (referred as V-J and V-D-J) is very high, due to not only the combination of different germline V, D, and J genes, but also to the deletion and addition of templated (P) nucleotides and the addition of non-templated (N) nucleotides at the junctions between the rearranged genes, and somatic hypermutation of expressed IG (**Papavasiliou and Schatz, 2002**; **Lefranc and Lefranc, 2020**). The total number of potential expressed rearranged IG and TR sequences in an individual is referred to as the adaptive immune receptor repertoire (AIRR). The adaptive immune repertoire is very diverse in a healthy individual, with the theoretically possible number of clonotypes reaching more than 10^{19} different TR (**Bradley and Thomas, 2019**) and 10^{11} IG (**Glanville et al., 2009**), far exceeds the number of B and T cells in a given individual (**Davis and Bjorkman, 1988**; **Freeman et al., 2009**; **Elhanati et al., 2015**). Thanks to next-generation sequencing (NGS), the AIRR can be sampled with sufficient depth for some of its complexity to be studied (**Weinstein et al., 2009**; **Six et al., 2013**). AIRR sequencing (AIRR-seq) provides insights into the immune status of an individual at steady-state or in altered conditions such as malignancy, autoimmune disease, immunodeficiency, infectious disease, or vaccination, and allows comparison of B- and T-cell populations between individuals and time points (**Benichou et al., 2012**; **Kirsch et al., 2015**; **Dziubianau et al., 2013**; **Hou et al., 2016**; **Ghraichy et al., 2018**). AIRR-seq permits the description and quantification of global diversity and characteristics of AIRR, the identification of clonal expansions, the tracking of particular clonotypes, and the prediction of their specificities (**Miho et al., 2018**; **Zvyagin et al., 2020**; **Sidhom et al., 2018**; **Glanville et al., 2017**; **Huang et al., 2020**; **Jokinen et al., 2021**; **Akbar et al., 2021**; **Hayashi et al., 2021**) as well as the antibody selection through phage display (**Rouet et al., 2018**; **Ravn et al., 2013**), thereby providing opportunities for new biomarker identification (**Gittelman, 2021**; **Dines, 2020**), therapeutic antibody discovery (**Akbar et al., 2021**; **Richardson et al., 2021**), CAR-T cell bioengineering (**Sheih et al., 2020**), vaccine development, cancer diagnostics and treatment (**Linette et al., 2019**; **Zhang et al., 2018**; **Lu et al., 2018**), including neoantigen discovery (**Chiou et al., 2021**; **Richters et al., 2019**) and immune intervention monitoring in diverse pathologies, such as stem cell transplantation (**Robinson, 2015**; **Fink, 2019**; **Jiang et al., 2019**; **Jacobsen et al., 2017**; **Theil, 2017**; **Link-Rachner et al., 2019**; **Rubelt et al., 2017**; **Parola et al., 2018**; **Georgiou et al., 2014**; **Arnaout et al., 2021**; **Anand et al., 2021**).

NGS-based approaches and methods have multiplied, now including high-throughput bulk sequencing of IG or TR starting from genomic DNA (gDNA) or mRNA (as cDNA), which typically provides information on one receptor chain only, and more recently to the sequencing of the two IG or TR chains expressed in a single cell, which provides information on the antigen-specific receptor. These approaches are increasingly applied, mostly to human AIRRs, but also to study AIRRs from other organisms (**Chaudhary and Wesemann, 2018**; **Minervina et al., 2019**). Molecular protocols to amplify IG or TR chains typically rely on Polymerase Chain Reaction (PCR), such as multiplex-PCR or RACE-PCR (**Robins et al., 2009**; **Wang et al., 2010**; **DeKosky et al., 2013**; **Eugster et al., 2013**; **Heather et al., 2015**; **Mamedov et al., 2013**).

To obtain reliable and comparable AIRR-seq data, the methods for performing AIRR-seq need to fulfill a number of requirements. First, the data generated by AIRR-seq must reflect the composition and diversity of the 'real cellular repertoire.' The cell subset(s), the cell sample size, and the sequencing depth used in an AIRR-seq experiment all influence the downstream data and should therefore be carefully adapted to the experimental question (**Rosenfeld et al., 2018**). PCR amplification of AIRR-seq libraries can introduce bias by preferentially targeting certain genes, by missing certain alleles (in the case of multiplex PCR with primers anchored in the V or J genes), and by overamplifying targets of certain genes and length (in multiplex PCR and RACE-PCR) (**Calis and Rosenberg, 2014**; **Alamyar et al., 2012**; **Gadala-Maria et al., 2015**; **Primi et al., 1986**; **Barennes et al., 2021**). All these parameters will influence how accurately AIRR profiling reflects the true abundance and diversity of clones in the immune repertoire. Second, AIRR V and J genes (and constant (C) genes for IG) must be identified in an unambiguous and unbiased manner as knowledge of the complementarity determining region 3 (CDR3; the site of V, (D), and J recombination in an IG or TR, and hence its greatest sequence diversity), but also of the CDR1 and CDR2 (and that of the C gene in IG) is critical to assess the physicochemical constraints that define specificity, affinity, and function of the

TR or the IG (***Li et al., 2013***; ***Rossjohn et al., 2015***). Third, AIRR-seq data should be as free from sequencing error as possible. The CDR3, key for assigning a sequence to a clonotype, is by definition unknown, and in the case of IG sequences it is important to be able to distinguish *bona fide* somatic hypermutation from artifactual mutations introduced by PCR or by sequencing errors all along the rearranged molecule, as the latter can generate falsely elevated inter- and intra-clonal variation. Further complicating the matter is the fact that the germline V genes from the same subgroup (e.g., IGHV4-31 vs. IGHV4-30-4 for IG) and V alleles or polymorphic variants of a given gene (e.g., TRAV14/DV4 for TR) may have very few nucleotide differences (***Lefranc, 2014***; ***Lefranc and Lefranc, 2001***). Therefore, distinguishing errors from true biological variants can be a major challenge. Finally, samples for AIRR-seq should be free from cross-sample contamination, to which AIRR-seq experiments are prone as multiple samples are often processed in parallel and sequenced on a single lane of a sequencing run.

In the past decade, multiple molecular biology protocols and approaches have been developed by academic and industrial investigators, rendering comparisons among studies difficult. Moreover, experimental and analytical protocols are highly complex and therefore prone to intra- and inter-experimental variability (***Barennes et al., 2021***; ***Liu et al., 2016***; ***Rosati et al., 2017***; ***Bashford-Rogers et al., 2014***). AIRR-seq can be performed either on gDNA or cDNA from multiple T/B cell populations (bulk sequencing) or on individual cells (single-cell sequencing). The starting material and sequencing method used depend on the application, as each has advantages and disadvantages (***Table 1***). The availability of commercial kits can be helpful, since they are produced following standards and rigorous quality control, thus offering standardized reagents and protocols across laboratories. Currently, available commercial kits include gDNA-based methods (e.g., Adaptive Biotechnologies, iRepertoire) as well as mRNA-based methods that use cDNA (e.g., Illumina, Takara Bio, iRepertoire) for bulk sequencing. Mainly mRNA-based commercial methods (e.g., 10X Genomics, Takara Bio, HiFiBio) are used for single-cell analysis, which can provide sequences for full receptors or antibodies. This is an important consideration as the determination of the isotype for IG requires a full or partial sequence of the constant region. Single-cell approaches are further helping in the detection of clonotypes because they provide both paired chains and potentially allow full-length cDNA sequencing (***Stubbington et al., 2017***). The fidelity of sequence is an additional factor to consider. Unique molecular identifiers (UMI) consisting of random stretches of 8–12 nucleotides are incorporated into oligonucleotides that are used to generate cDNA from mRNA, such that statistically each cDNA molecule contains a unique sequence. Analysis of sequences that share the same UMI is used to generate a consensus sequence, greatly reducing sequencing errors (***Shugay et al., 2014***). In contrast, multiplex PCR approaches can be associated with artifacts arising from primer competition or off-target primer binding. Although this favors RACE-PCR when considering mRNA-based methods, gDNA-based multiplex PCR may offer higher fidelity since it does not rely on reverse transcription (reverse transcriptase enzymes have higher error rates than DNA polymerases [***Ellefson et al., 2016***; ***Holland et al., 1982***]). Finally, cost may influence the choice of a particular protocol. There are many factors that contribute to the cost of AIRR-seq data generation. For example, the cost of sequencing, the sequencing depth, and the number of cells analyzed per sample are variable; also, the choice between commercial kits and ‘homebrew’ methods will influence costs. In general, gDNA analysis is the most cost-effective, because it requires the lowest sequencing depth with the largest representation of cells per sample, whereas single-cell analysis is on the opposite end of the scale, with bulk cDNA sequencing in the middle.

Several considerations should be taken into account when designing and planning an AIRR-seq experiment. In addition to the large number of different methods and protocols, other factors including budgetary constraints, timelines, sample types, and processing are also important. Given the diversity of AIRR-seq workflows, comparisons between different data sets are challenging or even impossible. Standards and controls are needed for optimal AIRR-seq data harmonization, interpretation, and sharing (***Rubelt et al., 2017***; ***Breden et al., 2017***; ***Vander Heiden et al., 2018***). This need led to the formation in 2015 of a grassroots community of scientists and other interested parties, known as the AIRR Community (https://www.antibodysociety.org/the-airr-community/). The objective of the Biological Resources Working Group (WG) within the AIRR Community is to coordinate the assessment and development of AIRR-seq controls, ultimately providing the scientific community with controls and standards for the generation, harmonization, and rigorous comparison and interpretation of AIRR-seq data. In order to recommend biological standards that are needed and to

Table 1. Current AIRR-seq methods and their typical use(s).

Bulk gDNA, bulk cDNA, and single-cell cDNA-based sequencing methods are compared with respect to their general features, uses, methods, and potential issues. Each is ranked using a semi-quantitative scale (from '+++" for best to '-" for worst or non-existent).

		Bulk gDNA sequencing	Bulk cDNA sequencing	Single-cell cDNA sequencing
General Features	PCR method	Multiplex	Multiplex and 5′ RACE	Multiplex and 5′ RACE
	Cell number	10^2–10^6	10^2–10^6	10^2–10^3
	Sample throughput	Low-high	Low-moderate	Low
	Length of receptor sequences	100–600 bp	150–600 bp	700–800 bp
	Availability of commercial kits and service providers	++	+++	+
Uses	Gene usage	++	++	+
	CDR3 length and properties	++	++	+
	Somatic hypermutation (for IG)	++	++	+
	Repertoire diversity	++	++	+/-
	Clonal expansion	+++	++	+
	Clonal evolution	++	+++	++
	Tracking of clonotypes	+++	++	+
	Clinical use (e.g., MRD detection)	++	+/-	-
	Unbiased detection of unproductive rearrangements	++	-	-
	Inference of germline	++	+	+/-
	Determination of constant gene	-	++	+
	Structural annotation	+/-	++	+
	Linkage of both antigen receptor chains	+/-	+/-	++
	Direct combination of AIRR-seq with single-cell immunophenotype (e.g., transcriptome or cell surface protein expression)	-	-	++
	Characterization of clonotype full antigen receptor/Functional testing	-	+/-	++
	Rare clonotype detection	++	++	+/-
Methods	Simplicity of workflow (library preparation)	+++	++	+
	Cost for library preparation commercial kits (per sample)	Low	Moderate	High
	Fidelity in sequences	Moderate	High	High
	Molecular barcoding (correcting PCR/sequencing error)	+/-	++	++
Potential Issues	V-gene amplification bias	++	+	+/-
	V-gene annotation issues	++	+	+
	PCR and sequencing error	++	+	+/-
	Difficulty with translation of copy number to cells	+/-	++	+/-
	Degradation of template	+	++	++

bp = base pairs; CDR3 = complementarity determining region 3; MRD = minimal residual disease; RACE = rapid amplification of cDNA ends; V = variable.

prioritize their development, a written survey was developed asking participants about the use of and need for controls for AIRR-seq experiments. The survey gathered information on participants' research interests, sample types, sequencing methodologies, currently available controls, and desired controls. In addition to the survey results, different AIRR-seq methods and controls were also gleaned from the literature, and finally, the WG also invited scientists with unpublished research on controls for their input. These three sources of data were then used to provide a comprehensive overview of sequencing methods, technical issues, current standards, and potential priorities for the development of future standards. Here we describe the progress of the Biological Resources WG of the AIRR Community, its information collection and proposed strategies to define, develop, and use AIRR-seq standards.

Biological controls in AIRR-seq experiments

AIRR Community survey: Overview and respondent demographics

To address the use, needs, and requirements for AIRR-seq controls and standards, we designed and disseminated a questionnaire to researchers in the AIRR Community, as well as to users of IMGT, the international ImMunoGeneTics information system (http://www.imgt.org) (***Lefranc, 2014***). The questionnaire was composed of 4 sections and included 28 questions, with tick-box predefined answers and free-text options allowing for participants' personalized answers (**Supplementary material**). After 6 months, 105 responses were recorded, including one incomplete response from a participant who neither produces nor analyzes AIRR-seq data. Three respondents participated twice, with consistent answers and same name and contact information, therefore only one completed form was considered for each. Answers from 101 remaining participants originated mainly from North America and Europe (***Figure 1A***) and were further analyzed. At the time of the survey, 96% of respondents were involved in AIRR-seq studies and 4% had plans to perform AIRR-seq studies in the future. Of the respondents, 92% were engaged in human studies, 48% in mouse studies, and 38% in the use of AIRR-seq to study other species or synthetic molecules (e.g., from phage-displayed antibody libraries; ***Figure 1B***). Approximately half of the respondents focused exclusively on IG while a quarter each studied TR and IG or TR alone (***Figure 1C***). Several respondents were interested in many different topics (***Figure 1D*** and ***Figure 1—figure supplement 1***), with their fields of interest dominated by 'immune system diseases' including infection, autoimmune disease, and cancer. Furthermore, the survey results clearly indicate an interest among the majority of respondents in developing bioinformatic tools for the analysis of AIRR-seq data, followed by major interests in other research areas such as vaccinology, immune repertoire homeostasis, immunotherapy, antibody engineering, hematology, and aging (***Figure 1D***). In addition, respondents with bioinformatic skills tended to be those who had broader research interests (***Figure 1—figure supplement 1***). Finally, 89/101 of survey respondents indicated an interest in using AIRR-seq to either track clonotypes over time or across samples, whereas 88/101 were interested in identifying highly expanded clonotypes, 87/101 on analyzing diversity, and 83/101 on studying clonal selection. In conclusion, participants in this survey came from diverse backgrounds, had wet and dry bench expertise, and had a breadth of research interests covering different aspects of AIRR-seq studies.

Survey results on sequencing methodologies used

Whereas 91% (92/101) questionnaire respondents commonly perform bulk sequencing experiments, 67% (68/101) combine bulk sequencing with the use of single-cell technologies. Only 8% of respondents focused exclusively on single-cell sequencing or phage display technologies only. With respect to the input biological material used in bulk sequencing, the majority of participants (83%) preferred to sequence long amplicons that covered the entire (or almost the entire) V-(D)-J region and part of the C region despite the associated higher sequencing cost per read and restrictions on the type of compatible sequencer. AIRR-seq researchers mainly used mRNA for cDNA sequencing (69%), while both mRNA and gDNA were used by 25%, and gDNA alone by 6% of respondents. ***Figure 2A*** shows that the majority of survey respondents performing bulk sequencing used multiplex PCR or the template-switching approach with a considerable number of AIRR-seq researchers using both methods. For those using either approach, mRNA was still the preferred starting material. In addition, UMIs were more commonly used with template switching than multiplex PCR approaches (***Figure 2B***). The association of UMIs with template switching methods is likely related to the

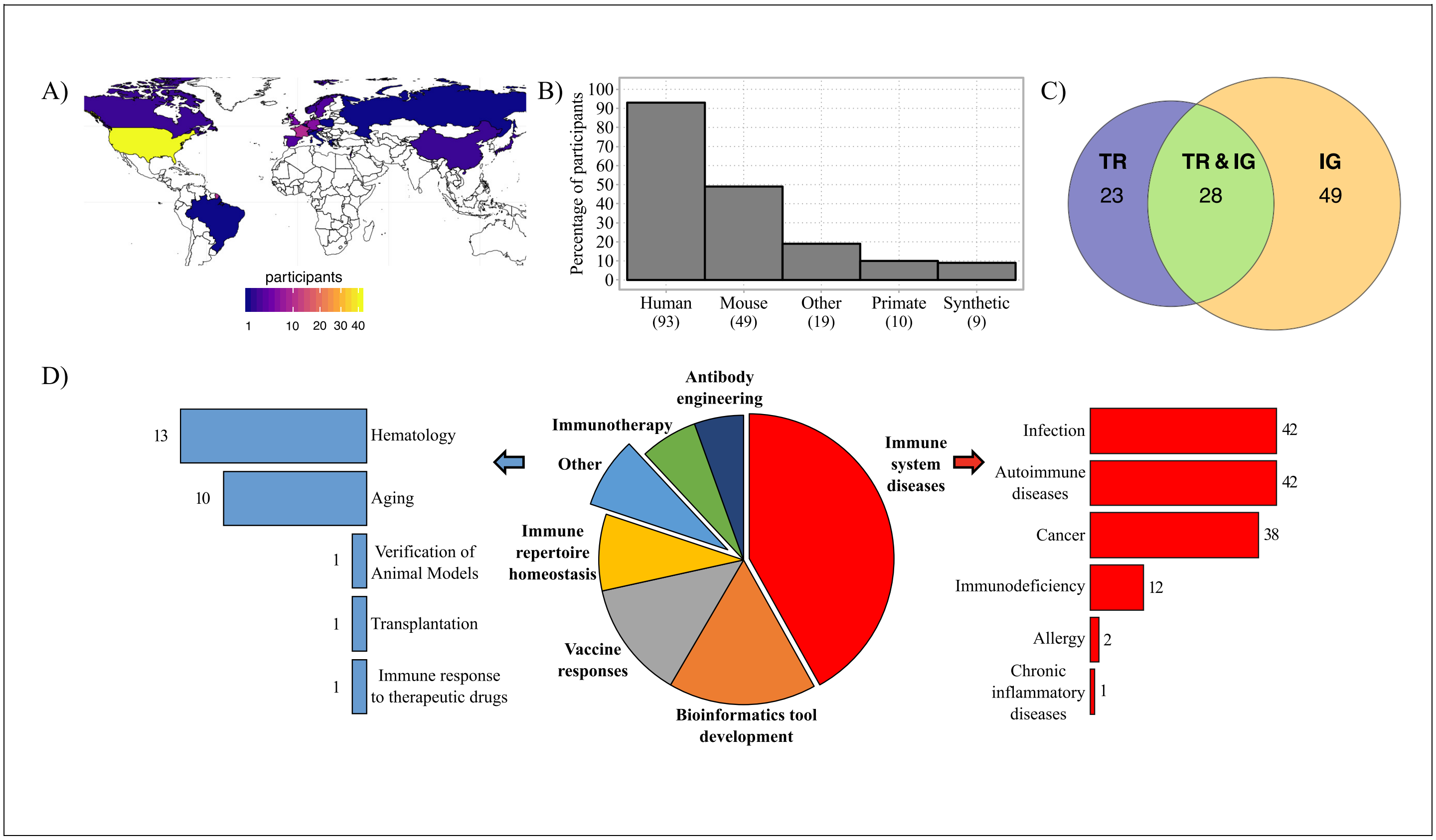

Figure 1. Geographic distribution of survey participants and their AIRR-seq research interests. (A) Map with geographic distribution of survey participants. (B) Histogram showing the principal studied organisms among the participants. The 'Other' category includes rat, ferret, rabbit, goat, pig, canine, bovis, cattle, chicken, fish, teleost, salmon, zebrafish, other fish species, transgenic animals. (C) Venn diagram representing the percentage of participants according to their interest in AIRR template type. (D) Pie-chart representing the distribution of survey participants according to their research interest(s). Immune system diseases and other categories are described in more detail in the bar plots (right and left). Numbers of respondents for each category are shown next to the bars.

The online version of this article includes the following figure supplement(s) for figure 1:

Figure supplement 1. Heatmaps of the areas of study depending on the interest.

template choice, since UMIs are not easy to use with gDNA-based templates, due to the incorporation of the UMI into the amplified products; nevertheless, two participants in the survey reported using UMIs with gDNA (*Figure 2B*).

Altogether, these initial results suggest that among survey respondents bulk sequencing on mRNA and gDNA are the most frequently used methods. However, the literature reflects increasing use of single-cell approaches, including the computational construction of IG or TR from RNA-sequencing (RNA-seq) libraries (*Figure 2—figure supplement 1*) and the combined use of target capture plus single-cell RNA-seq (e.g., 10X Genomics' 5′ end kits, etc.).

To gain insight into which standards should be prioritized for development and sharing with the scientific community, we asked survey participants about the standards they currently use and about the types of standards they would like to use. Based on the results, we concentrate below on potential controls for bulk analysis, as this approach is being used by the majority of the respondents (91%) and because the development of standards for single-cell applications is somewhat distinct.

Survey results on controls used and desired controls

Most respondents (88%) were interested in using standards or controls in AIRR-seq experiments. The 47 respondents who already use controls (n = 47) did so for protocol development (12/47 = 26%), everyday use (7/47 = 15%), or for both (15/47 = 32%), with the remaining (13/47 = 28%) not indicating their specific application. Commercial controls were used by 11, and homebrew

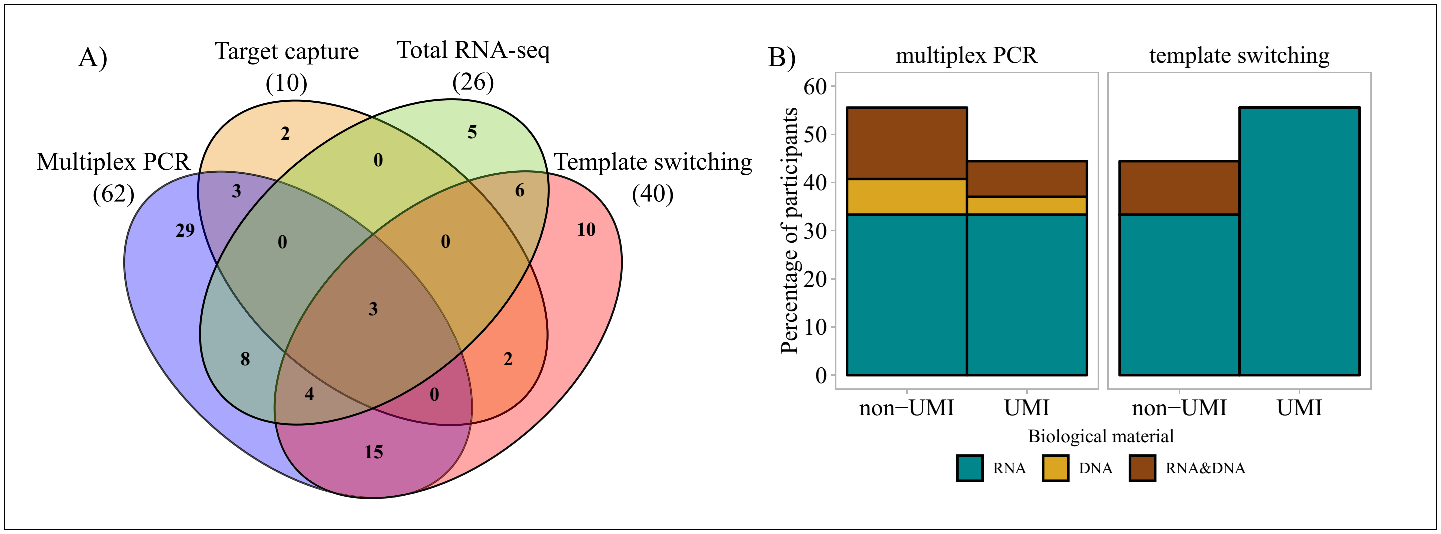

Figure 2. Molecular approaches used in bulk sequencing. (**A**) Venn diagram representing the most important molecular approaches used and their usage and sharing among participants. Numbers of respondents in each of the four main categories are shown in parentheses. (**B**) Bar plots representing biological material used and molecular barcoding proportion for the two major molecular biology approaches (multiplex PCR and template switching). Only the answers of respondents who used one technology exclusively are shown (Multiplex PCR: n = 29; Template switching: n = 10). UMI = unique molecular identifier.

The online version of this article includes the following figure supplement(s) for figure 2:

Figure supplement 1. Yearly number of PubMed entries referring to single-cell AIRR sequencing (left panel) and bioinformatic AIRR reconstruction from RNA-seq studies (right panel) 1980–2020 (via https://www.ncbi.nlm.nih.gov/pubmed/, accessed on 16 January 2020).

controls were used by 23 participants, with 3 people using both and 16 not specifying their source. ***Figure 3A*** indicates that the most commonly used homebrew controls were cell lines or pooled-cell preparations. Respondents who did not use controls (black bars) were also asked how they might want to use controls. ***Figure 3B*** depicts the community survey responses to these questions.

In summary, the survey, as well as further discussions within the AIRR Biological Resources WG, identified major concerns arising from different steps in the AIRR-seq workflow (***Table 2***, left and middle columns). Additionally, based on these identified issues, the right column of ***Table 2*** describes potential controls to address them. These controls are described in more detail in the subsequent sections of the manuscript.

Current concepts in the use of biological standards for RNA-seq experiments

To prioritize the development of community-wide standards for AIRR data, we turned to examples of community-wide standards in the NGS space. Several such standards have been developed and are actively used, for example, the External RNA Control Consortium (ERCC) spike-in controls and the Microarray Quality Control (MAQC) RNA standards. Both of these standards were generated through broad collaborations of stakeholders including the United States government, industry, and academic researchers (***Reid and External RNA Controls Consortium, 2005***; ***Su et al., 2014***).

The ERCC spike-in controls were designed specifically to be used with RNA-seq experiments, for normalization of expression values during analysis. Two tubes with different compositions of the RNA sequences are commercially available from Thermo Fisher Scientific as Invitrogen ERCC RNA Spike-In Mix (https://perma.cc/WW9Z-D2NY) and ready for use with eukaryotic samples. Each tube contains 92 RNA species with each containing a predefined polyA tail with a different sequence, with their relative abundance covering a ~10^6 fold range, and a limited range of GC content and lengths. The external RNA mixture can be used to normalize relative quantities of transcripts across samples within a single experiment or project, and to optimize protocols for reproducibility and accuracy. The ERCC standards are widely accepted by the transcriptome community, but do suffer from a few limitations: (i) the GC content and range of lengths are not broad - representing the average, but not all possible mRNA moieties; (ii) the polyA region is shorter than endogenous mRNA,

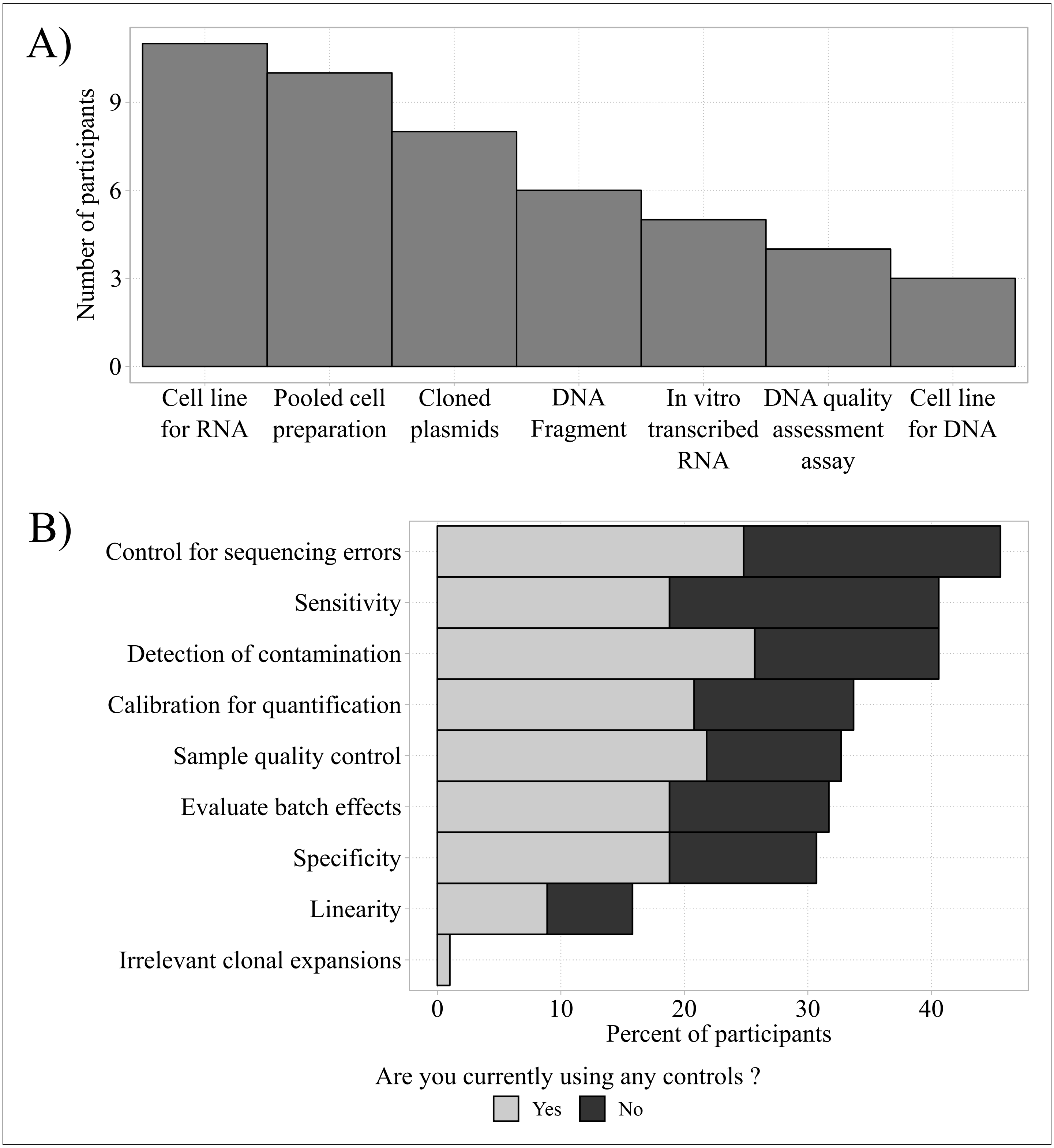

Figure 3. Homebrew controls and their desired applications. (**A**) Most frequently used homebrew controls (total n = 47). (**B**) Total frequencies of desired applications of homebrew controls for respondents currently using (gray bars; total n = 47) and currently not using controls (black bars; total n = 42).

Table 2. Concerns and expected errors introduced during AIRR-seq workflows and possible controls to detect them.
A typical workflow consists of 5 steps: Sample collection > Extraction > Amplification > Sequencing > Analysis.

Concern	Mechanism(s)	Example of potential controls
Sequence errors	Enzyme errors (RT, DNA polymerase); Sequencing errors	UMIs for bioinformatic error correction; Spike-in controls with defined sequences to evaluate error rates
Sensitivity	Enzymatic inefficiencies (RT or PCR conditions/polymerase); Sample collection size (e.g., cell input number, purity); Sequencing depth	Spike-in controls (synthetic or cellular) at known concentrations
Specificity	Enzyme bias (RT, DNA polymerase); Analysis pipelines (annotation, error correction)	Spike-in controls with defined sequences to identify overall V/D/J gene amplification bias
Detection of contamination	Bench-level cross contamination (sample mixing or PCR contamination) or barcode jumping during sequencing	Unique spike-in (synthetic) for each sample; UDIs for sequencing barcode crosstalk
Sample quality control	Sample collection or nucleic acid purification	Identified by spectroscopy or agarose electrophoresis
Evaluate batch effects	Subtle differences introduced at all stages of the workflow	Spike-in controls (synthetic or cellular); Parallel biological (clonal or complex) sample
Linearity/accuracy of clonotype quantification	Enzymatic inefficiencies (RT or PCR conditions); Analytical error correction	Spike-in controls (synthetic or cellular) at known concentrations
Reproducibility/ Batch effects	All stages	Spike-in controls (synthetic or cellular); Parallel biological (clonal or complex) sample; Comparison of replicate amplifications of the same sample; Comparison of sequences generated on the same sample in different sequencing runs
Data processing	Database/annotation limitations; filtering; error correction; collapsing/consensus algorithms	Spike-in controls (synthetic or cellular); Parallel biological (clonal or complex) sample

RT, reverse transcriptase; UMIs, unique molecular identifiers; UDIs, unique dual indices.

and (iii) no splicing variants are included (***Jiang et al., 2011***). These limitations result from the typical compromise accompanying any process for generating useful (and well validated) controls in a timely fashion.

The MAQC RNA standards were not generated as NGS controls initially, but instead were developed by the U.S. Food and Drug Administration (FDA)-led, community-wide consortium for the purpose of validating microarrays, instruments, and analysis methods. The MAQC project consists of four phases, with the first two focusing on microarray methods and the second two on NGS methods (SEQC; Sequencing QC). Initially, differential gene expression levels of nearly 1000 genes between two human reference RNA samples (Human Brain RNA and Human Universal RNA) were assessed by qRT-PCR and microarrays; these highly characterized RNA samples were subsequently used to validate different microarray or NGS library preparation methods, instruments, and data analysis methods (***Shi et al., 2006***). Phase 3 of the MAQC project applied similar concepts to NGS platforms and the comparison of results obtained by microarray or RNA sequencing (***Su et al., 2014***). Phase 4 of this project is ongoing, with the goal of developing robust analysis protocols and providing quality control metrics.

In addition, UMIs and unique dual indices (UDIs) have been proposed in the RNA-seq field (https://perma.cc/AMB4-WC86) (***Kircher et al., 2012***) to control and correct for sequencing errors, as well as sequencing index crosstalk.

Current practices for controls in AIRR-seq experiments

While the standards described above cannot be directly applied to AIRR-seq experimentation, they can serve as a blueprint for the development of standards. In ***Table 2*** (right column), we highlight different possible approaches to address the concerns of the AIRR Community. These approaches closely resemble the general standards described above, using spike-in controls or well-characterized biological samples, including UMIs or UDIs. As described below, AIRR-seq researchers have already initiated some studies to address the use of such controls.

Several groups have recently developed synthetic standards for use with AIRR-seq samples of mouse and human origin, all based on common principles, and generally available for academic institutions via a material transfer agreement. These synthetic templates are either generated from plasmids (via in vitro transcription followed by RT-PCR) or directly produced as synthetic dsDNAs, with numbers of unique sequences produced ranging from dozens to over 1 million (**Khan et al., 2016**; **Friedensohn et al., 2018**; **Carlson et al., 2013**) (and unpublished work of J. Trück, University Children's Hospital Zurich and C. Tipton, Emory University).

The first use of a synthetic repertoire was reported by **Carlson et al., 2013**. They developed synthetic DNA templates combining 14 and 4 different V and J genes of the TR gamma (TRG) locus, resulting in a total of 56 templates of equal length. All sequences contained three barcodes to unambiguously identify individual V-J combinations. In addition, universal primer sites that were identical in all synthetic templates were added on both ends. This approach allowed identification of amplification bias and optimization of primer concentrations as well as informing the computational correction of residual bias. Work on both murine and human IG repertoire was performed by **Khan et al., 2016** and **Friedensohn et al., 2018** using a similar strategy in the same research laboratory. There, a total of 16 murine and 85 human synthetic sequences were used to assess primer bias from multiplex PCR library generation. In combination with incorporation of UMIs into amplicons and an error correction analysis pipeline, this approach increased workflow fidelity and produced more accurate data. A very important element in the strategy used in this process was the integration of UMI during initial reverse transcription, resulting in labeling of each cDNA on a single molecule level. In contrast to the study by Carlson et al., synthetic sequences used in both studies from Khan et al. and Friedensohn et al. contained different CDR3 sequences and were used in different relative concentrations within the spike-in pool. This approach allowed to not only assess primer-dependent amplification bias but also the impact of variable input concentrations of synthetic sequences on their relative abundance following sequencing.

In principle, synthetic templates are designed such that each mimics an individual recombined V-(D)-J region and a partial C region while also containing universal priming sites, barcodes for unique template identification. The universal priming sites allow for unbiased quantification. Furthermore, comparison of amplification efficiency with the universal vs. targeted primers (the latter usually binding to the leader or V and J or C regions) may be used to correct for target-specific differences in amplification efficiency. Through amplification of synthetic templates alone, multiplex primer sets can be tuned to individual concentrations that will more accurately amplify known targets within repertoires or they can be used to eliminate primers that perform poorly altogether. Through the use of known templated sequences of known abundance (e.g., cell lines spiked into other cells), quantification and amplification efficiency can be calculated. Experience from early testing has identified certain limitations of this approach (see below). Some standards additionally harbor mutations deviating from the germline (unmutated) IGHV sequences; these can be used to model the efficiency of amplification of somatically mutated templates (**Friedensohn et al., 2018**). In practical terms, synthetic standards can be used during method development (primer optimization, alteration of methods to account for amplification bias, etc.), as spike-ins, or can be run as separate positive controls alongside the samples of interest. Synthetic templates can also serve as spike-in controls for concurrent quantification measurements of run-to-run variability, amplification and sequencing efficiency, as a positive control, or as a measurement of sample-to-sample contamination and/or index of misidentification (using different synthetic library spike-ins for individual amplifications in a pooled sequencing run).

Although theoretically promising, the molecular design and bioinformatic analyses of synthetic sequences are challenging. Controls should mimic biological repertoires as closely as possible, and therefore are most effective when they contain a representative level of the biological diversity, which is tedious and expensive. In addition, they should also be distinguishable from other biological sequences so that even following nucleotide changes introduced through PCR or sequencing errors, such synthetic sequences can be unambiguously resolved from their biological counterparts. These challenges may explain the rare usage of spike-in controls except for initial method optimization or very specific applications.

Mixed-cell populations and cell lines can also be used as workflow controls, as has been documented recently by the Euroclonality Consortium (**Knecht et al., 2019**; **Brüggemann et al., 2019**). The first type of Euroclonality (**Knecht et al., 2019**) control monitors general primer and sequencing performance of a sequencing run (batch of samples) and consists of a poly-target control, comprised of gDNA isolated from healthy human thymus, tonsil, and peripheral blood mononuclear cells

(PBMCs), the latter derived from apheresis donors (in a 1:1:1 ratio). This control is included in the workflow alongside experimental samples in a separate tube. By bioinformatic identification of the primer sequences and comparison to stored reference sequencing profiles from the sample control mix, unusual amplification patterns or batch effects can be identified. One advantage of such a cell mixture control is that it fully models the immune repertoire complexity of a bulk cell population and provides quality monitoring for every step of the process including amplification of the template and its sequencing. A second advantage is that this type of control is very easy to generate and is therefore accessible to many laboratories. A disadvantage of using this type of complex control is that one does not fully know the identity of all of the rearrangements in the sample, which can be problematic if there is sample or PCR contamination. A second disadvantage is that it is generated in a finite amount - and once used up, the process of validating the control must be repeated.

The second type of control used by the Euroclonality Consortium (***Brüggemann et al., 2019***) is designed to evaluate assay sensitivity and linearity within each library. This in-parallel control consists of a gDNA sample obtained from 59 human B/T lymphoid cell lines with a total of 46 well-defined rearrangements mixed together in different ratios and added to each processed sample. An advantage of using cell lines as in-parallel controls is that their gene rearrangements are defined and thus easily identified; theoretically allowing for the conversion of reads into cell numbers and permitting relative quantification of template abundance. In practice, however, the use of in-parallel amplification controls can be very challenging, and requires careful interpretation. For example, in samples with poor gDNA quality or low template abundance, the control templates may outcompete the test sample. The depth of sequencing and relative amounts of sample input can affect the measured abundance (***Barennes et al., 2021***; ***Chaara et al., 2018***). An additional disadvantage is that cell lines do not model a fully diverse repertoire, only a fraction of V and J gene combinations are represented by the cell lines, and thus primer performance and bias, especially between samples, are not fully controlled.

Discussion and future work

AIRR-seq experiments are becoming increasingly commonplace, in both the research and clinical settings. In contrast, the development of controls and best practices for assay validation, interpretation and standardization have lagged behind. Here, the members of the AIRR Community Biological Resources WG have summarized the current practice regarding the use of standards and controls among its members as well as among other international AIRR-seq experts and in the literature. We also have identified differences in the types of standards and controls that are used among users. Some of these differences depend on the sample type (fresh vs. fixed cells), the starting material (single cell vs. bulk), the template (mRNA vs. gDNA), as well as the quality and quantity of the relevant cells and templates. In addition, the selection of controls is influenced by the amplification method, with single cell and mRNA-based methods relying more heavily on cellular and molecular barcoding approaches, for example. Last, but certainly not least, the downstream application of the assay can profoundly influence the choice and prioritization of controls. In some cases, assays need to be sensitive and specific (e.g., a clinical grade assay that detects minimal residual disease) whereas in others quantitation (e.g., clonal size analysis for monitoring clonal expansions) or unbiased amplification (e.g., assessment of repertoire skewing during an immune response) may be more important.

All AIRR-seq assays can clearly benefit from rigorous controls. There is broad agreement that controls and standards are desirable, with over half of AIRR-seq survey respondents currently already using controls (mostly of a homebrew variety) in their experiments. Furthermore, whether individuals used controls or did not, they appeared to agree on the types of issues in analyzing and interpreting AIRR-seq data that would benefit from the use of controls - most importantly measuring and controlling for sample quality, assay sensitivity and specificity, and calibration for the quantification of clonal size. Also, with the progress regarding antigen-specificity inference using computational tools (***Glanville et al., 2017***; ***Huang et al., 2020***; ***Jokinen et al., 2021***; ***Akbar et al., 2021***; ***Hayashi et al., 2021***; ***Richardson et al., 2021***; ***Galson et al., 2015***; ***Shomuradova et al., 2020***; ***Chronister et al., 2021***; ***Sidhom et al., 2021***; ***Pogorelyy et al., 2019***; ***Dash et al., 2017***) or more conventionally through technologically challenging using antigen-binding approaches, including single-cell (***Johnson et al., 2020***; ***Fuchs et al., 2019***), having controls would be of major interest to ensure the accuracy of the TR or IG identified. It is unlikely that a single control can fulfill all of these needs across all methods and applications. There is at the moment also no obvious front-runner for a 'gold standard' that can be used to judge the adequacy of different types of controls.

Having identified a need, a diversity of methodologic approaches and the lack of a 'gold standard' for AIRR-seq controls, the AIRR Community Biological Resources WG is now coordinating the development of such controls. Although AIRR-seq researchers are aware of potential methodological problems, current solutions have not been systematically evaluated or compared. Based upon the current use of controls and needs identified by our survey respondents, the WG plans to focus on three forms of controls: one in-parallel (synthetic standards), one in-parallel and computational (UMIs), and one that is external (a complex cell mixture that is run in parallel to monitor amplification and sequencing run batch effects). To optimize these three types of standards, we first must determine how well they work. We therefore propose to carry out a multi-center analysis of three types of controls: (1) synthetic calibrators for bulk gDNA sequencing to measure clonal size and amplification bias; (2) UMIs for samples studied in parallel by bulk mRNA (with UMIs) and single cell sequencing (with cellular barcodes) for the analysis of amplification bias and sequencing error; and (3) a mixed cell population (either a human apheresis and/or pooled tissue product or murine spleen samples) for the evaluation of batch effects that compare between sequencing runs performed at the same site or between runs performed at different sites. These three types of controls can be run in parallel or in separate, dedicated experiments, allowing for greater participation of AIRR-seq investigators.

In order to perform these studies, the Biological Resources WG will first establish a framework suited to the analysis and quantitation of potential issues, depending on the type and amount of input material, the assay(s) used, and the analysis method(s). Since the method used for the production of AIRR-seq data can impact the results, as shown by the benchmarking of TR library preparation methods study (***Barennes et al., 2021***), we plan to evaluate different standards in the framework of a molecular biology method benchmarking study as well. For the TR repertoire, we will take advantage of the already evaluated methods to include more gDNA-based methods. For the IG repertoire analysis, we will launch a systematic study, leveraging high-volume pooled-cell collections and synthetic standards that can be shared by investigators at multiple sites. We plan four major experiments: (1) analysis of TR and IG rearrangements in PBMCs using bulk gDNA and RNA approaches; (2) analysis of TR on sorted naïve polyclonal T cells using bulk gDNA; (3) analysis of IG rearrangements on sorted naïve polyclonal B cells using bulk gDNA and RNA; and (4) analysis of IG rearrangements on spleen cells from organ donors using bulk gDNA and RNA approaches. Spleen cells are enriched for memory B-cell clones and are useful for modeling clonal expansion and somatic hypermutation (***Meng et al., 2017***). Synthetic controls and UMIs (in the case of RNA-based sequencing) will be added to triplicate samples of PBMCs and polyclonal splenocytes. Samples will be run with and without spike-in controls that will be included at different ratios. Ideally, the selected methods will all be handle by 2 to 3 labs, a compromise between feasibility and inter-lab validation. Protocols and workflows will be standardized and shared. To avoid sequencing batch effect, we will sequence all the replicates through the same facility. Based on the results and to determine the lowest possible cell input levels, we will then evaluate the impact of decreasing the quantity of cells and repeat the same schema, focusing on the most reproducible methods. Finally, we will work closely with other AIRR Community working groups and additional experts in the field to harmonize standardization efforts. Together with the AIRR Community Software WG, we will select a series of tools for data quality control, alignment, and annotation and identify the analysis pipelines required for the detection of contamination, amplification bias, and batch effects. By leveraging the diverse skills of AIRR Community investigators, the development, optimization, and dissemination of biological standards for AIRR-seq data should progress quickly. Such ambitious project will require financial support in order to help volunteer labs to handle the experiments, already under discussion at the level of the AIRR-community.

Using, testing, and comparing these standards is but the first step. Beyond that is their adoption by the wider scientific community. For widespread adoption, commercial and other partnerships are essential for high-quality production and dissemination of the standards. In addition, to ensure broad distribution, controls should either be free of significant intellectual property restrictions or, if proprietary, be well-documented. For this, the Biological Resources WG will reach out to academic research groups (e.g., IMGT), non-profit organizations (e.g., the Global Alliance for Genomics and Health), governmental organizations (e.g., the US Department of Commerce's National Institute of Standards and Technology), as well as commercial entities (e.g., the ATCC). With these efforts, we will be able to improve the rigor, robustness and interchangeability of AIRR-seq studies, and to increase their utility for downstream applications, including clinical diagnostics.

Additional information

Competing interests

Eline T Luning Prak: is consulting or is an advisor for Roche Diagnostics Corporation, Enpicom, The Antibody Society, The American Autoimmune Related Diseases Association and IEDB. Andrew Farmer: works for Takara Bio USA, but has no ownership or stock in the company. Magnolia Bostick: During the writing of the manuscript, Magnolia Bostick was employed by Takara Bio USA, but has no ownership or stock in the company. The other authors declare that no competing interests exist.

Funding

Funder	Grant reference number	Author
Swiss National Science Foundation	PZ00P3_161147	Johannes Trück
Swiss National Science Foundation	PZ00P3_183777	Johannes Trück
Agence Nationale de la Recherche	ANR-16-rHUS-0001	Pierre Barennes
Agence Nationale de la Recherche	ANR-11-IDEX-0004-02	Encarnita Mariotti-Ferrandiz
FP7 Ideas: European Research Council	Fp7-IdEAS-ErC-322856	Encarnita Mariotti-Ferrandiz
Agence Nationale de la Recherche	ANR-18-ECVD-0001	Encarnita Mariotti-Ferrandiz
H2020 Fast Track to Innovation	825821	Encarnita Mariotti-Ferrandiz
DFG	FZ 111	Anne Eugster
National Institutes of Health	UM1 AI144288	Eline T Luning Prak
National Institutes of Health	P01 AI106697	Eline T Luning Prak
National Institutes of Health	P30-AI0450080	Eline T Luning Prak
National Institutes of Health	UC4 DK112217	Eline T Luning Prak
National Institutes of Health	P30-CA016520	Eline T Luning Prak

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Johannes Trück, Conceptualization, Formal analysis, Writing - original draft, Writing - review and editing; Anne Eugster, Conceptualization, Writing - original draft, Writing - review and editing; Pierre Barennes, Data curation, Formal analysis, Writing - original draft; Christopher M Tipton, Writing - original draft; Eline T Luning Prak, Conceptualization, Methodology, Writing - original draft, Writing - review and editing; Davide Bagnara, Investigation, Writing - review and editing; Cinque Soto, Writing - original draft, Writing - review and editing; Jacob S Sherkow, Investigation; Aimee S Payne, Marie-Paule Lefranc, Andrew Farmer, Writing - review and editing; The AIRR Community, Validation; Magnolia Bostick, Conceptualization, Supervision, Investigation, Writing - original draft, Project administration, Writing - review and editing; Encarnita Mariotti-Ferrandiz, Conceptualization, Supervision, Methodology, Writing - original draft, Project administration, Writing - review and editing

Author ORCIDs

Johannes Trück https://orcid.org/0000-0002-0418-7381
Anne Eugster http://orcid.org/0000-0001-8009-5959
Eline T Luning Prak https://orcid.org/0000-0002-9478-9211
Cinque Soto https://orcid.org/0000-0002-3997-6217
Jacob S Sherkow http://orcid.org/0000-0002-9724-9261
Marie-Paule Lefranc https://orcid.org/0000-0003-0116-9353
Encarnita Mariotti-Ferrandiz https://orcid.org/0000-0002-8770-0186

References

Akbar R, Robert PA, Pavlović M, Jeliazkov JR, Snapkov I, Slabodkin A, Weber CR, Scheffer L, Miho E, Haff IH, Haug DTT, Lund-Johansen F, Safonova Y, Sandve GK, Greiff V. 2021. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports* **34**:108856. DOI: https://doi.org/10.1016/j.celrep.2021.108856, PMID: 33730590

Alamyar E, Giudicelli V, Li S, Duroux P, Lefranc M-P. 2012. IMGT/HIGHV-QUEST: the IMGT web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Research* **882**:569–604. DOI: https://doi.org/10.1007/978-1-61779-842-9_32

Anand T, Virmani N, Bera BC, Vaid RK, Vashisth M, Bardajatya P, Kumar A, Tripathi BN. 2021. Phage display technique as a tool for diagnosis and antibody selection for coronaviruses. *Current Microbiology* **78**:1124–1134. DOI: https://doi.org/10.1007/s00284-021-02398-9, PMID: 33687511

Arnaout RA, Prak ETL, Schwab N, Rubelt F, Adaptive Immune Receptor Repertoire Community. 2021. The future of blood testing is the immunome. *Frontiers in Immunology* **12**:626793. DOI: https://doi.org/10.3389/fimmu.2021.626793, PMID: 33790897

Barennes P, Quiniou V, Shugay M, Egorov ES, Davydov AN, Chudakov DM, Uddin I, Ismail M, Oakes T, Chain B, Eugster A, Kashofer K, Rainer PP, Darko S, Ransier A, Douek DC, Klatzmann D, Mariotti-Ferrandiz E. 2021. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nature Biotechnology* **39**:236–245. DOI: https://doi.org/10.1038/s41587-020-0656-3, PMID: 32895550

Bashford-Rogers RJ, Palser AL, Idris SF, Carter L, Epstein M, Callard RE, Douek DC, Vassiliou GS, Follows GA, Hubank M, Kellam P. 2014. Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunology* **15**:29. DOI: https://doi.org/10.1186/s12865-014-0029-0, PMID: 25189176

Benichou J, Ben-Hamo R, Louzoun Y, Efroni S. 2012. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing: rep-seq: ngs for the immunological repertoire. *Immunology* **135**:183–191. DOI: https://doi.org/10.1111/j.1365-2567.2011.03527.x

Bradley P, Thomas PG. 2019. Using T cell receptor repertoires to understand the principles of adaptive immune recognition. *Annual Review of Immunology* **37**:547–570. DOI: https://doi.org/10.1146/annurev-immunol-042718-041757, PMID: 30699000

Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE, Vander Heiden JA, Christley S, Bukhari SAC, Thorogood A, Matsen Iv FA, Wine Y, Laserson U, Klatzmann D, Douek DC, Lefranc MP, Collins AM, Bubela T, Kleinstein SH, Watson CT, et al. 2017. Reproducibility and reuse of adaptive immune receptor repertoire data. *Frontiers in Immunology* **8**:1418. DOI: https://doi.org/10.3389/fimmu.2017.01418, PMID: 29163494

Brüggemann M, Kotrová M, Knecht H, Bartram J, Boudjogrha M, Bystry V, Fazio G, Froňková E, Giraud M, Grioni A, Hancock J, Herrmann D, Jiménez C, Krejci A, Moppett J, Reigl T, Salson M, Scheijen B, Schwarz M, Songia S, et al. 2019. Standardized next-generation sequencing of immunoglobulin and T-cell receptor gene recombinations for MRD marker identification in acute lymphoblastic leukaemia; a EuroClonality-NGS validation study. *Leukemia* **33**:2241–2253. DOI: https://doi.org/10.1038/s41375-019-0496-7, PMID: 31243313

Calis JJ, Rosenberg BR. 2014. Characterizing immune repertoires by high throughput sequencing: strategies and applications. *Trends in Immunology* **35**:581–590. DOI: https://doi.org/10.1016/j.it.2014.09.004, PMID: 25306219

Carlson CS, Emerson RO, Sherwood AM, Desmarais C, Chung MW, Parsons JM, Steen MS, LaMadrid-Herrmannsfeldt MA, Williamson DW, Livingston RJ, Wu D, Wood BL, Rieder MJ, Robins H. 2013. Using synthetic templates to design an unbiased multiplex PCR assay. *Nature Communications* **4**:2680. DOI: https://doi.org/10.1038/ncomms3680, PMID: 24157944

Chaara W, Gonzalez-Tort A, Florez LM, Klatzmann D, Mariotti-Ferrandiz E, Six A. 2018. RepSeq data representativeness and robustness assessment by shannon entropy. *Frontiers in Immunology* **9**:1038. DOI: https://doi.org/10.3389/fimmu.2018.01038, PMID: 29868003

Chaudhary N, Wesemann DR. 2018. Analyzing immunoglobulin repertoires. *Frontiers in Immunology* **9**:462. DOI: https://doi.org/10.3389/fimmu.2018.00462, PMID: 29593723

Chiou SH, Tseng D, Reuben A, Mallajosyula V, Molina IS, Conley S, Wilhelmy J, McSween AM, Yang X, Nishimiya D, Sinha R, Nabet BY, Wang C, Shrager JB, Berry MF, Backhus L, Lui NS, Wakelee HA, Neal JW, Padda SK, et al. 2021. Global analysis of shared T cell specificities in human non-small cell lung cancer enables HLA inference and antigen discovery. *Immunity* **54**:586–602. DOI: https://doi.org/10.1016/j.immuni.2021.02.014, PMID: 33691136

Chronister WD, Crinklaw A, Mahajan S, Vita R, Koşaloğlu-Yalçın Z, Yan Z, Greenbaum JA, Jessen LE, Nielsen M, Christley S, Cowell LG, Sette A, Peters B. 2021. TCRMatch: predicting T-Cell receptor specificity based on sequence similarity to previously characterized receptors. *Frontiers in Immunology* **12**:640725. DOI: https://doi.org/10.3389/fimmu.2021.640725, PMID: 33777034

Dash P, Fiore-Gartland AJ, Hertz T, Wang GC, Sharma S, Souquette A, Crawford JC, Clemens EB, Nguyen THO, Kedzierska K, La Gruta NL, Bradley P, Thomas PG. 2017. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**:89–93. DOI: https://doi.org/10.1038/nature22383, PMID: 28636592

Davis MM, Bjorkman PJ. 1988. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**:395–402. DOI: https://doi.org/10.1038/334395a0, PMID: 3043226

DeKosky BJ, Ippolito GC, Deschner RP, Lavinder JJ, Wine Y, Rawlings BM, Varadarajan N, Giesecke C, Dörner T, Andrews SF, Wilson PC, Hunicke-Smith SP, Willson CG, Ellington AD, Georgiou G. 2013. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nature Biotechnology* **31**:166–169. DOI: https://doi.org/10.1038/nbt.2492, PMID: 23334449

Dines JN. 2020. The ImmuneRACE study: a prospective multicohort study of immune response action to COVID-19 events with the ImmuneCODETM open access database. *medRxiv*. DOI: https://doi.org/10.1101/2020.08.17.20175158

Dziubianau M, Hecht J, Kuchenbecker L, Sattler A, Stervbo U, Rödelsperger C, Nickel P, Neumann AU, Robinson PN, Mundlos S, Volk H-D, Thiel A, Reinke P, Babel N. 2013. TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T Cell-Related pathology. *American Journal of Transplantation* **13**:2842–2854. DOI: https://doi.org/10.1111/ajt.12431

Elhanati Y, Sethna Z, Marcou Q, Callan CG, Mora T, Walczak AM. 2015. Inferring processes underlying B-cell repertoire diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**:20140243. DOI: https://doi.org/10.1098/rstb.2014.0243

Ellefson JW, Gollihar J, Shroff R, Shivram H, Iyer VR, Ellington AD. 2016. Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science* **352**:1590–1593. DOI: https://doi.org/10.1126/science.aaf5409, PMID: 27339990

Eugster A, Lindner A, Heninger A-K, Wilhelm C, Dietz S, Catani M, Ziegler A-G, Bonifacio E. 2013. Measuring T cell receptor and T cell gene expression diversity in antigen-responsive human CD4+ T cells. *Journal of Immunological Methods* **400-401**:13–22. DOI: https://doi.org/10.1016/j.jim.2013.11.003

Fink K. 2019. Can we improve vaccine efficacy by targeting T and B cell repertoire convergence? *Frontiers in Immunology* **10**:110. DOI: https://doi.org/10.3389/fimmu.2019.00110, PMID: 30814993

Freeman JD, Warren RL, Webb JR, Nelson BH, Holt RA. 2009. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome Research* **19**:1817–1824. DOI: https://doi.org/10.1101/gr.092924.109, PMID: 19541912

Friedensohn S, Lindner JM, Cornacchione V, Iazeolla M, Miho E, Zingg A, Meng S, Traggiai E, Reddy ST. 2018. Synthetic standards combined with error and Bias correction improve the accuracy and quantitative resolution of antibody repertoire sequencing in human naïve and memory B cells. *Frontiers in Immunology* **9**:1401. DOI: https://doi.org/10.3389/fimmu.2018.01401, PMID: 29973938

Fuchs YF, Sharma V, Eugster A, Kraus G, Morgenstern R, Dahl A, Reinhardt S, Petzold A, Lindner A, Löbel D, Bonifacio E. 2019. Gene Expression-Based identification of Antigen-Responsive CD8$^+$ T cells on a Single-Cell level. *Frontiers in Immunology* **10**:2568. DOI: https://doi.org/10.3389/fimmu.2019.02568, PMID: 31781096

Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. 2015. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *PNAS* **112**:E862–E870. DOI: https://doi.org/10.1073/pnas.1417683112, PMID: 25675496

Galson JD, Kelly DF, Truck J. 2015. Identification of Antigen-Specific B-Cell receptor sequences from the total B-Cell repertoire. *Critical Reviews in Immunology* **35**:463–478. DOI: https://doi.org/10.1615/CritRevImmunol.2016016462, PMID: 27279044

Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. 2014. The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology* **32**:158–168. DOI: https://doi.org/10.1038/nbt.2782, PMID: 24441474

Ghraichy M, Galson JD, Kelly DF, Trück J. 2018. B-cell receptor repertoire sequencing in patients with primary immunodeficiency: a review. *Immunology* **153**:145–160. DOI: https://doi.org/10.1111/imm.12865, PMID: 29140551

Gittelman RM. 2021. Diagnosis and tracking of SARS-CoV-2 infection by T-Cell receptor sequencing. *medRxiv*. DOI: https://doi.org/10.1101/2020.11.09.20228023

Glanville J, Zhai W, Berka J, Telman D, Huerta G, Mehta GR, Ni I, Mei L, Sundar PD, Day GM, Cox D, Rajpal A, Pons J. 2009. Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *PNAS* **106**:20216–20221. DOI: https://doi.org/10.1073/pnas.0909775106, PMID: 19875695

Glanville J, Huang H, Nau A, Hatton O, Wagar LE, Rubelt F, Ji X, Han A, Krams SM, Pettus C, Haas N, Arlehamn CSL, Sette A, Boyd SD, Scriba TJ, Martinez OM, Davis MM. 2017. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**:94–98. DOI: https://doi.org/10.1038/nature22976, PMID: 28636589

Hayashi F, Isobe N, Glanville J, Matsushita T, Maimaitijiang G, Fukumoto S, Watanabe M, Masaki K, Kira JI. 2021. A new clustering method identifies multiple sclerosis-specific T-cell receptors. *Annals of Clinical and Translational Neurology* **8**:163–176. DOI: https://doi.org/10.1002/acn3.51264, PMID: 33400858

Heather JM, Best K, Oakes T, Gray ER, Roe JK, Thomas N, Friedman N, Noursadeghi M, Chain B. 2015. Dynamic perturbations of the T-Cell receptor repertoire in chronic HIV infection and following antiretroviral therapy. *Frontiers in Immunology* **6**:644. DOI: https://doi.org/10.3389/fimmu.2015.00644, PMID: 26793190

Holland J, Spindler K, Horodyski F, Grabau E, Nichol S, VandePol S. 1982. Rapid evolution of RNA genomes. *Science* **215**:1577–1585. DOI: https://doi.org/10.1126/science.7041255, PMID: 7041255

Hou D, Chen C, Seely EJ, Chen S, Song Y. 2016. High-Throughput Sequencing-Based immune repertoire study during infectious disease. *Frontiers in Immunology* **7**:336. DOI: https://doi.org/10.3389/fimmu.2016.00336, PMID: 27630639

Huang H, Wang C, Rubelt F, Scriba TJ, Davis MM. 2020. Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nature Biotechnology* **38**:1194–1202. DOI: https://doi.org/10.1038/s41587-020-0505-4, PMID: 32341563

Jacobsen LM, Posgai A, Seay HR, Haller MJ, Brusko TM. 2017. T cell receptor profiling in type 1 diabetes. *Current Diabetes Reports* **17**:118. DOI: https://doi.org/10.1007/s11892-017-0946-4, PMID: 29022222

Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Research* **21**:1543–1551. DOI: https://doi.org/10.1101/gr.121095.111, PMID: 21816910

Jiang N, Schonnesen AA, Ma KY. 2019. Ushering in integrated T cell repertoire profiling in Cancer. *Trends in Cancer* **5**:85–94. DOI: https://doi.org/10.1016/j.trecan.2018.11.005, PMID: 30755308

Johnson JL, Rosenthal RL, Knox JJ, Myles A, Naradikian MS, Madej J, Kostiv M, Rosenfeld AM, Meng W, Christensen SR, Hensley SE, Yewdell J, Canaday DH, Zhu J, McDermott AB, Dori Y, Itkin M, Wherry EJ, Pardi N, Weissman D, et al. 2020. The transcription factor T-bet resolves memory B cell subsets with distinct tissue distributions and antibody specificities in mice and humans. *Immunity* **52**:842–855. DOI: https://doi.org/10.1016/j.immuni.2020.03.020, PMID: 32353250

Jokinen E, Huuhtanen J, Mustjoki S, Heinonen M, Lähdesmäki H. 2021. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLOS Computational Biology* **17**:e1008814. DOI: https://doi.org/10.1371/journal.pcbi.1008814, PMID: 33764977

Khan TA, Friedensohn S, Gorter de Vries AR, Straszewski J, Ruscheweyh HJ, Reddy ST. 2016. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Science Advances* **2**: e1501371. DOI: https://doi.org/10.1126/sciadv.1501371, PMID: 26998518

Kircher M, Sawyer S, Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Research* **40**:e3. DOI: https://doi.org/10.1093/nar/gkr771, PMID: 22021376

Kirsch IR, Watanabe R, O'Malley JT, Williamson DW, Scott L-L, Elco CP, Teague JE, Gehad A, Lowry EL, LeBoeuf NR, Krueger JG, Robins HS, Kupper TS, Clark RA. 2015. TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. *Science Translational Medicine* **7**:308ra158. DOI: https://doi.org/10.1126/scitranslmed.aaa9122

Knecht H, Reigl T, Kotrová M, Appelt F, Stewart P, Bystry V, Krejci A, Grioni A, Pal K, Stranska K, Plevova K, Rijntjes J, Songia S, Svatoň M, Froňková E, Bartram J, Scheijen B, Herrmann D, García-Sanz R, Hancock J, et al. 2019. Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. *Leukemia* **33**:2254–2265. DOI: https://doi.org/10.1038/s41375-019-0499-4, PMID: 31227779

Lefranc MP. 2014. Immunoglobulin and T cell receptor genes: IMGT and the birth and rise of immunoinformatics. *Frontiers in Immunology* **5**:22. DOI: https://doi.org/10.3389/fimmu.2014.00022, PMID: 24600447

Lefranc M-P, Lefranc G. 2001. *The Immunoglobulin FactsBook*. Elsevier.

Lefranc M-P, Lefranc G. 2020. Immunoglobulins or antibodies: IMGT bridging genes, structures and functions. *Biomedicines* **8**:319. DOI: https://doi.org/10.3390/biomedicines8090319

Li S, Lefranc MP, Miles JJ, Alamyar E, Giudicelli V, Duroux P, Freeman JD, Corbin VD, Scheerlinck JP, Frohman MA, Cameron PU, Plebanski M, Loveland B, Burrows SR, Papenfuss AT, Gowans EJ. 2013. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nature Communications* **4**:2333. DOI: https://doi.org/10.1038/ncomms3333, PMID: 23995877

Linette GP, Becker-Hapak M, Skidmore ZL, Baroja ML, Xu C, Hundal J, Spencer DH, Fu W, Cummins C, Robnett M, Kaabinejadian S, Hildebrand WH, Magrini V, Demeter R, Krupnick AS, Griffith OL, Griffith M, Mardis ER, Carreno BM. 2019. Immunological ignorance is an enabling feature of the oligo-clonal T cell response to melanoma neoantigens. *PNAS* **116**:23662–23670. DOI: https://doi.org/10.1073/pnas.1906026116, PMID: 31685621

Link-Rachner CS, Eugster A, Rücker-Braun E, Heidenreich F, Oelschlägel U, Dahl A, Klesse C, Kuhn M, Middeke JM, Bornhäuser M, Bonifacio E, Schetelig J. 2019. T-cell receptor-α repertoire of CD8+ T cells following allogeneic stem cell transplantation using next-generation sequencing. *Haematologica* **104**:622–631. DOI: https://doi.org/10.3324/haematol.2018.199802, PMID: 30262565

Liu X, Zhang W, Zeng X, Zhang R, Du Y, Hong X, Cao H, Su Z, Wang C, Wu J, Nie C, Xu X, Kristiansen K. 2016. Systematic comparative evaluation of methods for investigating the tcrβ repertoire. *PLOS ONE* **11**:e0152464. DOI: https://doi.org/10.1371/journal.pone.0152464, PMID: 27019362

Lu YC, Zheng Z, Robbins PF, Tran E, Prickett TD, Gartner JJ, Li YF, Ray S, Franco Z, Bliskovsky V, Fitzgerald PC, Rosenberg SA. 2018. An efficient Single-Cell RNA-Seq approach to identify Neoantigen-Specific T cell receptors. *Molecular Therapy : The Journal of the American Society of Gene Therapy* **26**:379–389. DOI: https://doi.org/10.1016/j.ymthe.2017.10.018, PMID: 29174843

Mamedov IZ, Britanova OV, Zvyagin IV, Turchaninova MA, Bolotin DA, Putintseva EV, Lebedev YB, Chudakov DM. 2013. Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Frontiers in Immunology* **4**:456. DOI: https://doi.org/10.3389/fimmu.2013.00456, PMID: 24391640

Meng W, Zhang B, Schwartz GW, Rosenfeld AM, Ren D, Thome JJC, Carpenter DJ, Matsuoka N, Lerner H, Friedman AL, Granot T, Farber DL, Shlomchik MJ, Hershberg U, Luning Prak ET. 2017. An atlas of B-cell clonal distribution in the human body. *Nature Biotechnology* **35**:879–884. DOI: https://doi.org/10.1038/nbt.3942, PMID: 28829438

Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. 2018. Computational strategies for dissecting the High-Dimensional complexity of adaptive immune repertoires. *Frontiers in Immunology* **9**:224. DOI: https://doi.org/10.3389/fimmu.2018.00224, PMID: 29515569

Minervina A, Pogorelyy M, Mamedov I. 2019. T-cell receptor and B-cell receptor repertoire profiling in adaptive immunity. *Transplant International* **32**:1111–1123. DOI: https://doi.org/10.1111/tri.13475, PMID: 31250479

Papavasiliou FN, Schatz DG. 2002. Somatic hypermutation of immunoglobulin genes: merging mechanisms for genetic diversity. *Cell* **109 Suppl**:S35–S44. DOI: https://doi.org/10.1016/s0092-8674(02)00706-7, PMID: 11983151

Parola C, Neumeier D, Reddy ST. 2018. Integrating high-throughput screening and sequencing for monoclonal antibody discovery and engineering. *Immunology* **153**:31–41. DOI: https://doi.org/10.1111/imm.12838, PMID: 28898398

Pogorelyy MV, Minervina AA, Shugay M, Chudakov DM, Lebedev YB, Mora T, Walczak AM. 2019. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLOS Biology* **17**:e3000314. DOI: https://doi.org/10.1371/journal.pbio.3000314, PMID: 31194732

Primi D, Barbier E, Cazenave PA. 1986. Structural polymorphism of V kappa 21 E and V kappa 21 D gene products in laboratory mice. *European Journal of Immunology* **16**:292–296. DOI: https://doi.org/10.1002/eji.1830160315, PMID: 3082651

Ravn U, Didelot G, Venet S, Ng KT, Gueneau F, Rousseau F, Calloud S, Kosco-Vilbois M, Fischer N. 2013. Deep sequencing of phage display libraries to support antibody discovery. *Methods* **60**:99–110. DOI: https://doi.org/10.1016/j.ymeth.2013.03.001, PMID: 23500657

Reid L, External RNA Controls Consortium. 2005. Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics* **6**:150. DOI: https://doi.org/10.1186/1471-2164-6-150, PMID: 16266432

Richardson E, Galson JD, Kellam P, Kelly DF, Smith SE, Palser A, Watson S, Deane CM. 2021. A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. *mAbs* **13**:1869406. DOI: https://doi.org/10.1080/19420862.2020.1869406, PMID: 33427589

Richters MM, Xia H, Campbell KM, Gillanders WE, Griffith OL, Griffith M. 2019. Best practices for bioinformatic characterization of neoantigens for clinical utility. *Genome Medicine* **11**:56. DOI: https://doi.org/10.1186/s13073-019-0666-2, PMID: 31462330

Robins HS, Campregher PV, Srivastava SK, Wacher A, Turtle CJ, Kahsai O, Riddell SR, Warren EH, Carlson CS. 2009. Comprehensive assessment of T-cell receptor beta-chain diversity in Alphabeta T cells. *Blood* **114**:4099–4107. DOI: https://doi.org/10.1182/blood-2009-04-217604, PMID: 19706884

Robinson WH. 2015. Sequencing the functional antibody repertoire–diagnostic and therapeutic discovery. *Nature Reviews Rheumatology* **11**:171–182. DOI: https://doi.org/10.1038/nrrheum.2014.220, PMID: 25536486

Rosati E, Dowds CM, Liaskou E, Henriksen EKK, Karlsen TH, Franke A. 2017. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnology* **17**:61. DOI: https://doi.org/10.1186/s12896-017-0379-9, PMID: 28693542

Rosenfeld AM, Meng W, Chen DY, Zhang B, Granot T, Farber DL, Hershberg U, Luning Prak ET. 2018. Computational evaluation of B-Cell clone sizes in bulk populations. *Frontiers in Immunology* **9**:1472. DOI: https://doi.org/10.3389/fimmu.2018.01472, PMID: 30008715

Rossjohn J, Gras S, Miles JJ, Turner SJ, Godfrey DI, McCluskey J. 2015. T cell antigen receptor recognition of antigen-presenting molecules. *Annual Review of Immunology* **33**:169–200. DOI: https://doi.org/10.1146/annurev-immunol-032414-112334, PMID: 25493333

Rouet R, Jackson KJL, Langley DB, Christ D. 2018. Next-Generation sequencing of antibody display repertoires. *Frontiers in Immunology* **9**:118. DOI: https://doi.org/10.3389/fimmu.2018.00118, PMID: 29472918

Rubelt F, Busse CE, Bukhari SAC, Bürckert JP, Mariotti-Ferrandiz E, Cowell LG, Watson CT, Marthandan N, Faison WJ, Hershberg U, Laserson U, Corrie BD, Davis MM, Peters B, Lefranc MP, Scott JK, Breden F, Luning Prak ET, Kleinstein SH, AIRR Community. 2017. Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nature Immunology* **18**:1274–1278. DOI: https://doi.org/10.1038/ni.3873, PMID: 29144493

Sheih A, Voillet V, Hanafi LA, DeBerg HA, Yajima M, Hawkins R, Gersuk V, Riddell SR, Maloney DG, Wohlfahrt ME, Pande D, Enstrom MR, Kiem HP, Adair JE, Gottardo R, Linsley PS, Turtle CJ. 2020. Clonal kinetics and single-cell transcriptional profiling of CAR-T cells in patients undergoing CD19 CAR-T immunotherapy. *Nature Communications* **11**:219. DOI: https://doi.org/10.1038/s41467-019-13880-1, PMID: 31924795

Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, et al. 2006. The MicroArray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**:1151–1161. DOI: https://doi.org/10.1038/nbt1239, PMID: 16964229

Shomuradova AS, Vagida MS, Sheetikov SA, Zornikova KV, Kiryukhin D, Titov A, Peshkova IO, Khmelevskaya A, Dianov DV, Malasheva M, Shmelev A, Serdyuk Y, Bagaev DV, Pivnyuk A, Shcherbinin DS, Maleeva AV, Shakirova NT, Pilunov A, Malko DB, Khamaganova EG, et al. 2020. SARS-CoV-2 epitopes are recognized by a public and diverse repertoire of human T cell receptors. *Immunity* **53**:1245–1257. DOI: https://doi.org/10.1016/j.immuni.2020.11.004, PMID: 33326767

Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, Bolotin DA, Staroverov DB, Putintseva EV, Plevova K, Linnemann C, Shagin D, Pospisilova S, Lukyanov S, Schumacher TN, Chudakov DM. 2014. Towards error-free profiling of immune repertoires. *Nature Methods* **11**:653–655. DOI: https://doi.org/10.1038/nmeth.2960, PMID: 24793455

Sidhom J-W, Larman HB, Pardoll DM, Baras AS. 2018. DeepTCR: a deep learning framework for revealing structural concepts within TCR repertoire. *bioRxiv*. DOI: https://doi.org/10.1101/464107

Sidhom JW, Larman HB, Pardoll DM, Baras AS. 2021. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications* **12**:1605. DOI: https://doi.org/10.1038/s41467-021-21879-w, PMID: 33707415

Six A, Mariotti-Ferrandiz ME, Chaara W, Magadan S, Pham HP, Lefranc MP, Mora T, Thomas-Vaslin V, Walczak AM, Boudinot P. 2013. The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Frontiers in Immunology* **4**:413. DOI: https://doi.org/10.3389/fimmu.2013.00413, PMID: 24348479

Stubbington MJT, Rozenblatt-Rosen O, Regev A, Teichmann SA. 2017. Single-cell transcriptomics to explore the immune system in health and disease. *Science* **358**:58–63. DOI: https://doi.org/10.1126/science.aan6828, PMID: 28983043

Su Z, Labaj P, Li S, SEQC/MAQC-III Consortium. 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nature Biotechnology* **32**:903–914. DOI: https://doi.org/10.1038/nbt.2957, PMID: 25150838

Theil A. 2017. T cell receptor repertoires after adoptive transfer of expanded allogeneic regulatory T cells: t cell receptor repertoires post-T $_{reg}$ cell therapy. *Clinical and Experimental Immunology* **187**:316–324. DOI: https://doi.org/10.1111/cei.12887

Tonegawa S. 1983. Somatic generation of antibody diversity. *Nature* **302**:575–581. DOI: https://doi.org/10.1038/302575a0, PMID: 6300689

Vander Heiden JA, Marquez S, Marthandan N, Bukhari SAC, Busse CE, Corrie B, Hershberg U, Kleinstein SH, Matsen Iv FA, Ralph DK, Rosenfeld AM, Schramm CA, Christley S, Laserson U, AIRR Community. 2018. AIRR community standardized representations for annotated immune repertoires. *Frontiers in Immunology* **9**:2206. DOI: https://doi.org/10.3389/fimmu.2018.02206, PMID: 30323809

Wang C, Sanders CM, Yang Q, Schroeder HW, Wang E, Babrzadeh F, Gharizadeh B, Myers RM, Hudson JR, Davis RW, Han J. 2010. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *PNAS* **107**:1518–1523. DOI: https://doi.org/10.1073/pnas.0913939107, PMID: 20080641

Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR. 2009. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**:807–810. DOI: https://doi.org/10.1126/science.1170020, PMID: 19423829

Zhang S-Q, Ma K-Y, Schonnesen AA, Zhang M, He C, Sun E, Williams CM, Jia W, Jiang N. 2018. High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nature Biotechnology* **36**:1156–1159. DOI: https://doi.org/10.1038/nbt.4282

Zvyagin IV, Tsvetkov VO, Chudakov DM, Shugay M. 2020. An overview of immunoinformatics approaches and databases linking T cell receptor repertoires to their antigen specificity. *Immunogenetics* **72**:77–84. DOI: https://doi.org/10.1007/s00251-019-01139-4, PMID: 31741011