nature reviews methods primers

https://doi.org/10.1038/s43586-023-00284-1

Primer

Check for updates

Adaptive immune receptor repertoire analysis

Vanessa Mhanna[1,2,6], Habib Bashour[3,4,6], Khang Lê Quý[3], Pierre Barennes[1,2], Puneet Rawat[3], Victor Greiff[3] & Encarnita Mariotti-Ferrandiz[1,2,5]

Abstract

B cell and T cell receptor repertoires compose the adaptive immune receptor repertoire (AIRR) of an individual. The AIRR is a unique collection of antigen-specific receptors that drives adaptive immune responses, which in turn is imprinted in each individual AIRR. This supports the concept that the AIRR could determine disease outcomes, for example in autoimmunity, infectious disease and cancer. AIRR analysis could therefore assist the diagnosis, prognosis and treatment of human diseases towards personalized medicine. High-throughput sequencing, high-dimensional statistical analysis, computational structural biology and machine learning are currently employed to study the shaping and dynamics of the AIRR as a function of time and antigenic challenges. This Primer provides an overview of concepts and state-of-the-art methods that underlie experimental and computational AIRR analysis and illustrates the diversity of relevant applications. The Primer also addresses some of the outstanding challenges in AIRR analysis, such as sampling, sequencing depth, experimental variations and computational biases, while discussing prospects of future AIRR analysis applications for understanding and predicting adaptive immune responses.

Sections

[1]Immunology–Immunopathology–Immunotherapy (i3), UMRS959, Sorbonne Université and INSERM, Paris, France. [2]Clinical Investigation Center for Biotherapies (CIC-BTi), AP-HP and INSERM, Hôpital Pitié-Salpêtrière, Paris, France. [3]Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway. [4]School of Life Sciences, University of Warwick, Coventry, UK. [5]Institut Universitaire de France (IUF), Paris, France. [6]These authors contributed equally: Vanessa Mhanna, Habib Bashour. e-mail: victor.greiff@medisin.uio.no; encarnita.mariotti@sorbonne-universite.fr

Introduction

Adaptive immune responses are driven upon antigen recognition by an array of adaptive immune receptors (AIRs) that include B cell receptors (BCRs) and T cell receptors (TCRs), expressed by B lymphocytes and T lymphocytes, respectively[1]. BCRs and TCRs are composed of combinations of two chains belonging to the immunoglobulin superfamily (IgSF). All AIR chains have two distinct regions: the variable (V) region at the 5′ end of the protein that contains the antigen binding moiety, and the constant (C) region at the 3′ end of the protein. Each BCR is composed of two identical immunoglobulin heavy chains (IgHs) anchored to the cell surface through the constant IgH region (IgHC), and two identical immunoglobulin light chains (IgLs), each of which are bound to the heavy chains through disulfide bonds[2] (Fig. 1). The TCR contains two distinct chains that are both anchored to the cell surface through their respective C regions[3] (Fig. 1). There are two types of TCR heterodimers: the α/β-chain TCR (TCRαβ) is expressed on the surface of the majority of T cells, whereas the γ/δ-chain TCR (TCRγδ) is expressed on γδ T cells[4–6]. The antigen binding site of each BCR and TCR is embedded in the three-dimensional structure that is formed by the V regions of their chain pairs. BCRs can recognize native antigens of proteic, nucleic and lipid nature (Fig. 1), whereas TCRαβ can exclusively bind peptides presented by major histocompatibility (MHC) molecules that are expressed on the cell surface of antigen-presenting cells (Fig. 1). TCRγδ recognizes peptides or lipids presented by various non-classical MHC molecules (as reviewed elsewhere[7,8]).

The generation of a diverse set of BCRs and TCRs within an individual is ensured by a complex somatic recombination machinery that is unique to jawed vertebrates[9,10]. The extraordinary diversity of this adaptive immune receptor repertoire (AIRR) endows the immune system with the potential to recognize a plethora of antigens, including antigens stemming from pathogenic and commensal microorganisms, host-derived molecules and allergens.

AIRR generation

AIRRs have been extensively studied since the discovery of the origin of B cells and T cells[11–13]. In mammals, BCR and TCR repertoires are primarily generated in the primary lymphoid organs (PLOs), the bone marrow and the thymus for B cells and T cells, respectively[14]. The somatic recombination machinery assembles a set of functional immunoglobulin (*IG*) or TCR (*TR*) genes from a larger pool of gene segments – the variable (V), diversity (D) and joining (J) segments – all combined to form the V regions of the BCRs or the TCRs[15,16] (Fig. 1). The collections of V and J gene segments of the *IGL*, *TRA* or *TRG* genes and of the V, D and J gene segments of the *IGH*, *TRB* and *TRD* genes are all encoded on different genomic loci and chromosomes[17]. The variable BCR and TCR regions that are generated through somatic recombination are subdivided into four highly conserved framework regions (FRs) and three complementary determining regions (CDRs), which show increased sequence diversity. CDR1 and CDR2 are germline-encoded by V gene segments, whereas the CDR3 results from stochastic insertions and deletions of nucleotides between the V, (D) and J genes. During the recombination process, the CDR3 becomes variable in length and sequence[18], and greatly contributes to the high diversity of the AIRR[15,19,20] and to antigen recognition[21,22]. By the end of this process, each B lymphocyte and T lymphocyte expresses on its surface multiple copies of a unique BCR or TCR, respectively, that features a specific combination of V, D and J alleles and a unique CDR3. These lymphocytes clonally expand upon encounter of their specific antigen in the secondary lymphoid organs. During the later stages of B cell differentiation, somatic mutations may occur within the recombined variable region of the BCR, namely somatic hypermutations (SHMs). SHM leads to the formation of a clonal lineage that expresses modified nucleotidic versions of the parental BCR, often with improved binding towards a specific antigen[15,23,24] and therefore variations at the amino acid level as well (Fig. 1). Moreover, immunoglobulin class switching can occur in B cells upon antigen stimulation, through an intrachromosomal deletional recombination within the heavy chain constant (C_H) region[25–27]. This allows daughter cells from the same parent clone to produce antibodies of different isotypes, meaning different C regions, that have different effector functions, without altering their antigen specificity.

AIRR diversity and specificity

The potential number of distinct AIRs generated in the PLOs has been estimated at around 10^{19} TCRs[28,29] and 10^{13} BCRs[30] in humans, and this diversity can be further increased by SHM for BCRs. Recently, an AIRR diversity of 10^{61} has been predicted by statistical modelling[31]. However, only a small fraction of the potential BCR and TCR repertoires are present in a given individual, because of the limited number of lymphocytes an organism can harbour (~8×10^{11} in humans and ~10^8–10^9 in mice[32–34]). Furthermore, each AIRR is shaped by various selection events, in the PLOs and in the secondary lymphoid organs, following TCR and BCR interactions with self antigens or non-self antigens (Fig. 2). In the PLOs, such selection processes lead to a highly diverse AIRR, through the selection of B cells and T cells that can recognize non-self antigens and the deletion of high-affinity self-reactive lymphocytes[30,35–38].

The high diversity of AIRRs is essential for developing immunity against pathogenic organisms and for maintaining host homeostasis[1,39]. An individual's AIRR reveals information about ongoing immune responses, but also about previous antigenic encounters[39,40] as those are reflected within the repertoire of adaptive immune memory cells[41]. Furthermore, the study of AIR specificity has led to the development of invaluable tools for experimental research (for example, antibody-based detection methods)[42–46], diagnostics (for example, serum-based diagnostics)[47–50], prevention of disease (vaccine design) and therapeutics (for example, therapies based on TCR, chimeric antigen receptor T cells (CAR T cells) and antibodies)[51–53].

In addition to humans and mice, AIR loci description and rearrangement mechanisms have been studied in other vertebrates, such as marsupials[54], Galliformes[55] and sharks[56], and these studies revealed differences among species and enabled the standardized description and annotation of AIR genes in several species[17]. Furthermore, AIRR analyses have been applied to explore the kinetics of B cell and T cell immune responses and memory formation following a viral infection or during a prime–boost vaccination, in rainbow trout, showing differing observations compared with humans[57,58]. Thus, studying the AIRR in a diversity of species could open new research avenues such as evolutionary and comparative immunology.

Nevertheless, fundamental questions about the diversity, specificity and function of the AIRR have remained unresolved for more than half a century since the clonal selection theory was proposed[59]. The AIRR diversity in an individual at a given time point and its fluctuations over time, the number of distinct lymphocyte clones comprised within a certain AIRR, the clonal size of each clone, the number of clones that are specific to a given antigen and the extent of cross-reactivity are all points that must be resolved. In addition, it remains unclear how the AIRR is shaped throughout lymphocyte ontogeny, selection events, and external and internal perturbations. Finally, functional perspectives on

how individuals can establish an efficient immune response against foreign pathogenic antigens for which no specific B cells and T cells have been positively selected, all while avoiding excessive tissue damage, or on the extent to which the AIRR contributes to the development of pathological autoimmunity are needed. Answering these questions requires a deep quantitative deciphering of AIRRs. As such, thanks to technological advances, it is now possible to study the AIRR via sequencing (AIRR-seq) in bulk and single cells, sometimes at spatial

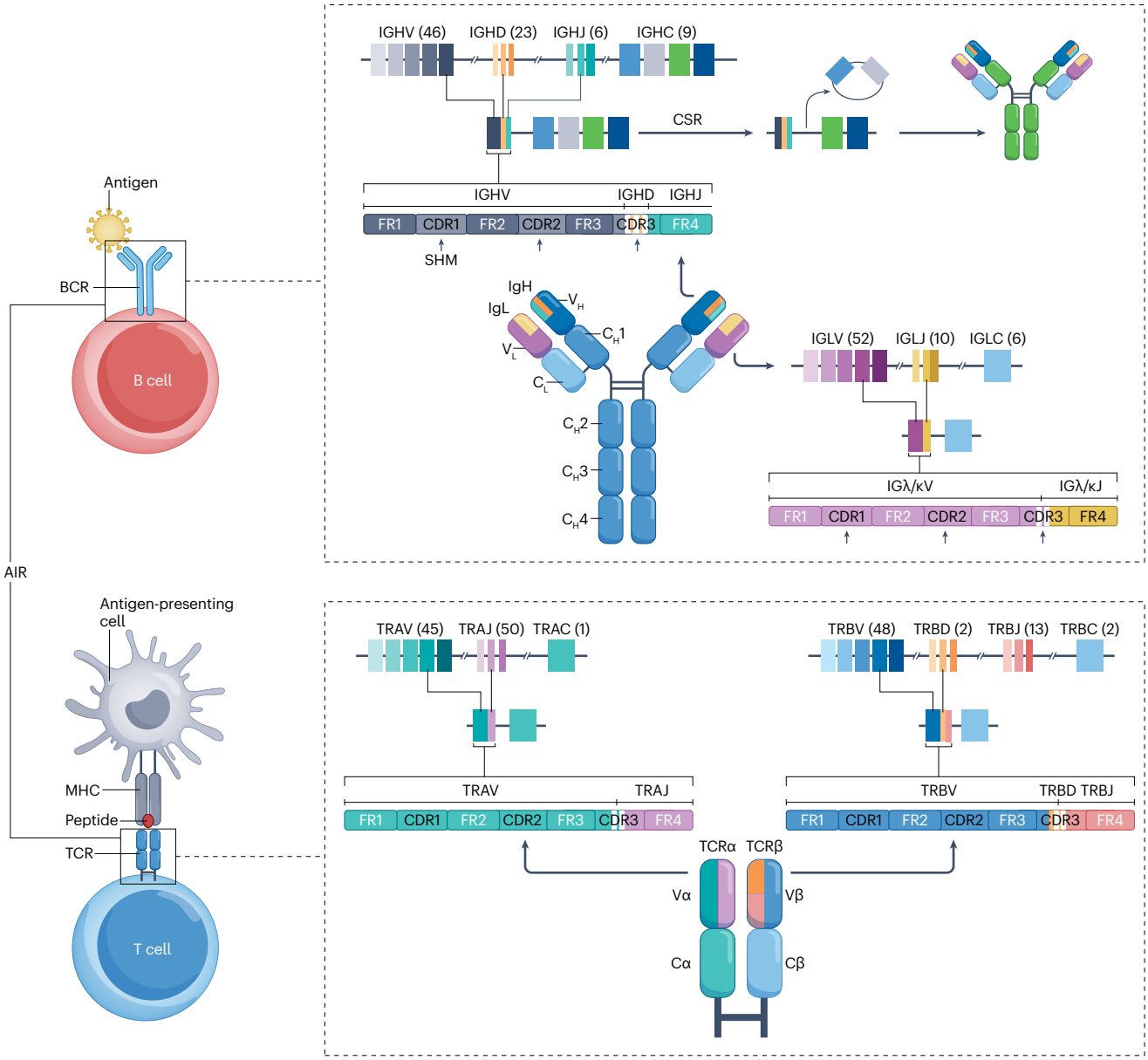

Fig. 1 | AIR generation and structure. B cells express receptors (B cell receptors (BCRs)) that can recognize native antigens of proteic, nucleic and lipid nature. A BCR is composed of two identical immunoglobulin heavy chains (IgHs) and two identical immunoglobulin light chains (IgLs). T cells express receptors (T cell receptors (TCRs)) that bind exclusively to peptides embedded in the major histocompatibility complex (MHC) expressed at the cell surface of antigen-presenting cells. A TCR is composed of two chains, an α-chain (TCRα) and a β-chain (TCRβ) expressed at the surface of the majority of T cells, or a γ-chain (TCRγ) and a δ-chain (TCRδ) expressed on γδ T cells. BCRs and TCRs are generated through a random somatic recombination of immunoglobulin or TCR genes from a set of genes, called variable (V), diversity (D) and joining (J), encoded on different genomic loci and chromosomes. The number of functional genes in humans, provided by IMGT[449] as of 14 November 2023, are shown. The variable region (V_H/V_L for immunoglobulin and Vα/Vβ for TCR) generated through this process can be further subdivided into four framework regions (FRs) and three complementary determining regions (CDRs). Additional insertions and deletions, represented by white bars within CDR3, occur during the recombination process. Upon antigen encounter in the secondary lymphoid organs, activated B cells may undergo point mutations within the CDRs (represented by arrows below the regions), namely somatic hypermutations (SHMs), or isotype switching through a class-switch recombination (CSR) process within the heavy chain constant (C_H) region. AIR, adaptive immune receptor.

resolution[60,61], combined with the application of machine learning approaches to AIRR-seq data for the construction of data-driven predictive models. These progresses are expected to help answer most of the above questions based on insights about AIRR diversity within and across individuals and AIRR clonal architecture.

The field of AIRR-seq has witnessed breakthroughs in the past decade. In this Primer, we focus on key concepts of experimental design for AIRR studies and of computational analyses, including machine learning. For each of the discussed concepts, we provide examples of applications. Moreover, we emphasize the needs for reproducibility as well as experimental and computational optimizations and highlight current limitations encountered in the field.

Experimentation

Originally, cellular biology-based approaches, particularly flow cytometry, were widely used to quantify the relative abundance of B cells and T cells expressing certain V gene segments[62]. The complexity of the recombination machinery prompted the development of molecular-based methods, such as CDR3 spectratyping or immunoscope analysis, which enabled a descriptive, qualitative AIRR analysis[63,64]. Quantitative AIRR analyses were only introduced in 2009 with the advent of high-throughput sequencing (HTS) approaches, which were designed to sequence up to millions of DNA and RNA molecules simultaneously[65–67]. After that, major efforts have been made to improve the experimental methods, to reduce technical biases and ensure reproducibility[68–71], as well as to adopt new technologies, such as single-cell TCR and BCR sequencing[72,73] and spatial transcriptomics[60,74]. To ensure robust and faithful experimental assessment of the AIRR, multiple factors must be considered, including the type of biological sample, the choice of the nucleic acid starting template, the library preparation method and the HTS protocols.

Type of biological sample

The type of sample used in an experiment depends both on its accessibility and on the biological question to be addressed. B cells and T cells may be collected from biological fluids, or fresh and preserved tissues

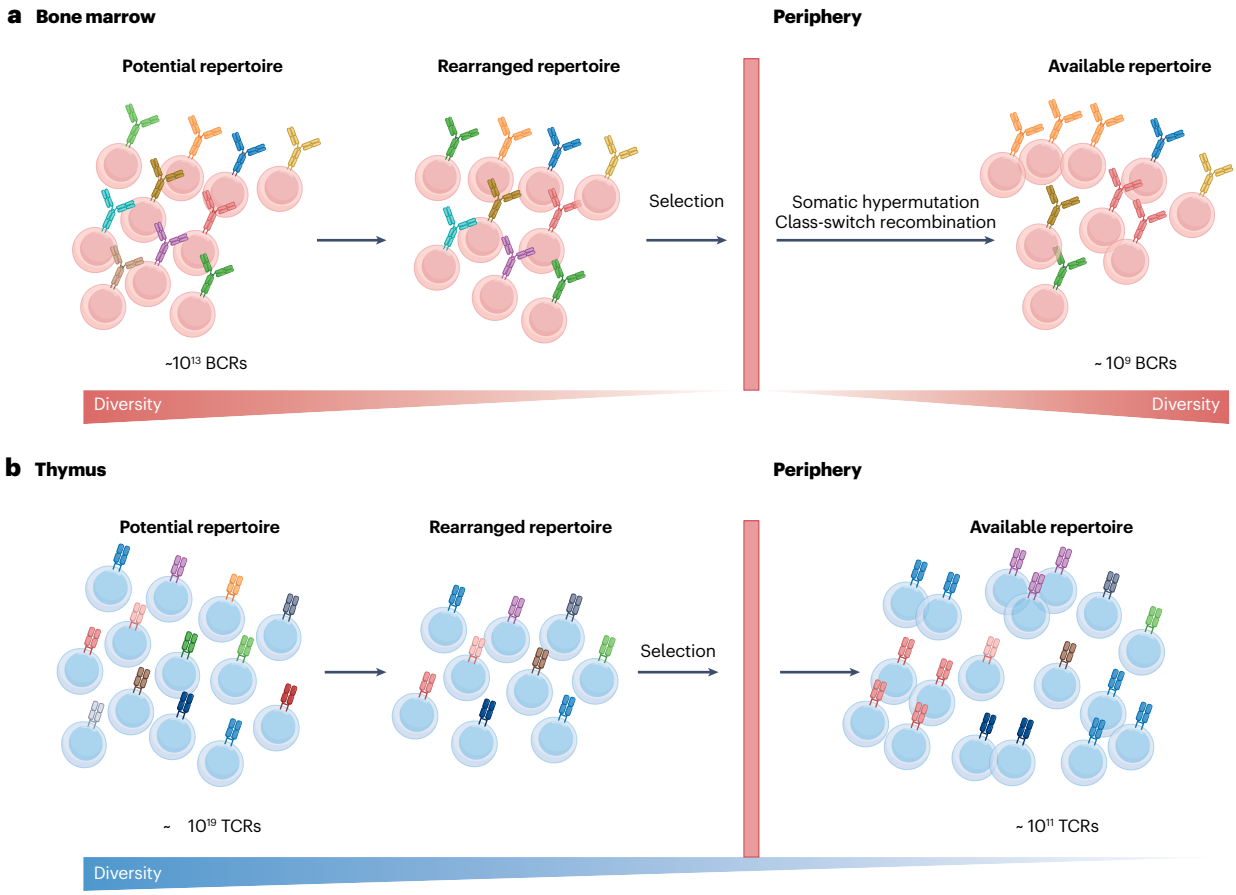

Fig. 2 | Shaping of AIR diversity. a,**b**, Variable, (diversity) and joining (V(D)J) recombination occurs during B cell and T cell development in the bone marrow (part **a**) and the thymus (part **b**), respectively. This process is the first step in shaping of the B cell receptor (BCR) and T cell receptor (TCR) repertoires, where diversity decreases drastically, going from a potential estimation based on the stochastic recombination process, around 10^{19} TCRs and 10^{13} BCRs in humans, to the rearranged repertoire. However, only a fraction of the realized rearrangements expressed by positively selected lymphocytes pass the central selection process and migrate to the periphery where they constitute the available peripheral repertoire. In turn, the latter is reshaped following antigen encounter, further reducing its diversity. The B cell repertoire is exclusively subjected to somatic hypermutations (SHMs) and class-switch recombinations, two phenomena that participate in immunoglobulin diversification in antigen-specific cells. Although no accurate estimate has yet been made to quantify the peripheral diversity, upper bounds can be fixed to 10^{11} for TCRs and 10^{9} for BCRS, which represent the number of circulating B lymphocytes and T lymphocytes in the periphery. AIR, adaptive immune receptor.

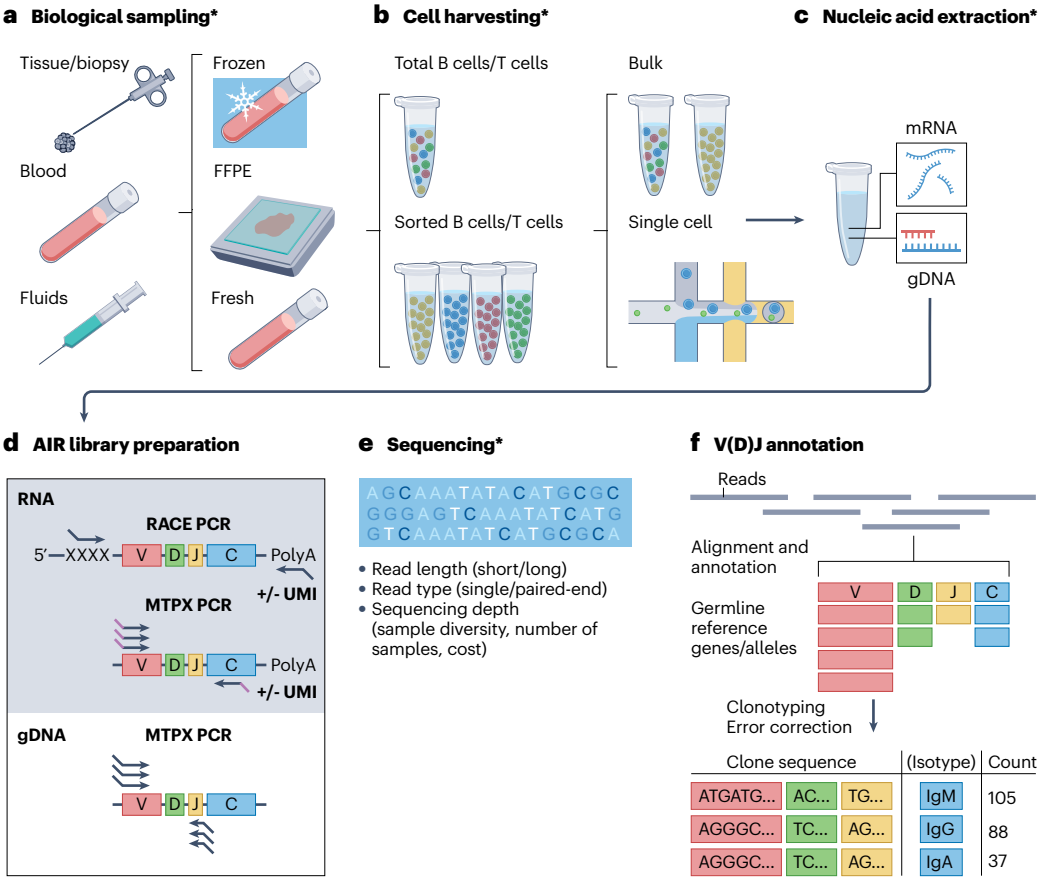

Fig. 3 | AIRR collection, preparation and sequencing. a, Adaptive immune receptor repertoire (AIRR) composition can be studied from peripheral blood, tissues, solid tumour biopsies and body fluids. Collected samples can be handled fresh, frozen or fixed (frozen and formalin-fixed paraffin-embedded (FFPE)) in the case of tissues or biopsies. **b**, B cells or T cells can be recovered and studied as one whole population or sorted into subsets of interest based on their phenotype and functional characteristics. The latter strategy, only from fresh or frozen samples, is advantageous when interested in rare cell subsets that are present in limited numbers. Both bulk and single-cell sequencing techniques can be used considering the technical limitations of each of these methods as detailed in the text. **c**, Genomic DNA (gDNA) or mRNA can be extracted from the samples and used for library preparation, a choice that depends on the quantity of extracted nucleic acids and experimental goals. **d**, Commercial and in-house library preparation methods are now available and based on either multiplex (MTPX) PCR, compatible with both RNA and gDNA molecules, or rapid amplification of complementary DNA ends (RACE) PCR, only applicable on RNA, each with its own advantages and disadvantages. **e**, Multiple factors should be taken into consideration when selecting the appropriate sequencing platform, with the aim of ensuring accurate repertoire coverage and reducing sequencing error rate and experimental costs. **f**, Raw AIRR sequencing data (FASTQ format) are firstly aligned to the species-specific germline genes to extract variable, (diversity) and joining (V(D)J) annotations, as well as isotype classification for BCRs. Subsequently, AIR sequences are assembled depending on a predefined clonal sequence and exported in a human-readable format for downstream analysis. Quality control is necessary following steps in parts **c**, **d** and **e** to ensure that the obtained material can be used in the onward steps. Biological and technical replicates can also be included in the experiment at different steps (indicated by asterisk). C, constant region; UMI, unique molecular identifier.

and organs. Frozen and formalin-fixed paraffin-embedded (FFPE) samples require downstream processing (Fig. 3a).

Whereas organs and tissues can easily be collected for AIRR analyses in experimental animal models, in humans it is only possible to recover entire tissues or organs from organ donors[75–77] and biopsies under routine clinical care. This allows the study of tissue-resident AIRs in the contexts of autoimmune, inflammatory or infectious diseases, as well as in liquid or solid cancers, the latter capturing the tumour microenvironment where adaptive immune cells localize.

Blood remains the preferred sample type in most translational and clinical studies due to its accessibility, non-invasive collection and relative abundance[40,78–80]. It is strongly recommended, when possible, to start from peripheral blood mononuclear cells instead of whole blood, particularly when using RNA as a starting material for library preparation as whole blood contains around 80% of β-globin RNAs, which could reduce the detection of white lymphocytic transcripts[81]. Moreover, to preserve RNA quality in whole blood, stabilization and freezing are required, which are not compatible with cell sorting. Plasma can also be used considering its increased concentration in cell-free DNA, which has been gaining interest in cancer studies[82,83]. Alternatively, lymphocytes can be collected from other sources, such as synovial and cerebrospinal fluid or bronchoalveolar lavage[84–87].

Starting from fresh samples is highly recommended to conserve the original AIRR composition. Alternatively, FFPE or frozen tissues can be analysed, as long as they are preserved in conditions that maintain cell integrity. When tissues sit in paraffin for a long time, quality is diminished, and frozen tissues are susceptible to loss of fragile cells during the thawing–unthawing cycle[88,89]. When handling such samples, nucleic quality and quantity must be checked as well as the lymphocyte infiltration, to ensure having the required material to perform the analysis (details below and elsewhere[90]).

Finally, it is important to consider whether to study bulk lymphocytes or sorted cell subsets (Fig. 3b). Most AIRR data are collected via bulk-sequencing techniques from millions of total B cells or T cells. Although studying total cells can reduce the manipulation time and experimental cost, functionally relevant AIRR modifications and insights about cell subset diversity can be overlooked. For instance, mRNA levels in plasma cells are 10-fold to 100-fold higher than in naive B cells[91,92], which can bias the biological interpretation of the obtained results. A prior cell sorting step is thus preferred to sort out the subsets of interest based on their phenotype and functional characteristics[35,79,93–96]. Cell sorting can also be advantageous when investigating the contribution of a rare cell subset to a given pathology, and this calls for the setting up of an efficient sorting strategy that combines speed, efficiency and cell purity. Finally, whereas bulk sequencing does not allow the study of AIR chain pairing, single-cell technologies now offer this possibility for samples with no more than 10^4 cells, such as small cell subsets[97], cells from tumour biopsies[98,99] or sorted cells with a given antigen specificity[100,101].

Nucleic acid starting template

The number of lymphocytes and the origin of the sample available for analysis are the determining factors for selecting the most suitable starting material for library preparation. This material can be either genomic DNA (gDNA) or mRNA (Fig. 3c) for bulk/sorted cell samples. On the one hand, gDNA quantity is proportional to the number of cells with the corresponding AIR, as a single copy of each rearrangement is found in a cell, which provides perfect linearity between gDNA molecules and cell counts[102–104]. Although gDNA libraries enable the quantification of BCR or TCR clonotypes, they require higher concentration input, and this represents a hurdle when the studied sample is limited in size[105]. On the other hand, using mRNA as a starting material precludes quantification of the absolute abundance of clonotypes, but offers 10–100 times higher sensitivity as compared with gDNA given that each transcript is found in multiple copies in a given cell[69]. Therefore, mRNA offers greater accessibility, particularly for small samples with limited nucleic acid quantities[103,106], provided that it is carefully manipulated, as RNA is sensitive to degradation. Moreover, using mRNA transcripts, which are more likely to be efficiently translated and give rise to functional BCRs or TCRs, reduces the background noise of AIR-seq data, whereas using gDNA introduces an uncertainty on whether a molecule will contribute or not to a productive rearrangement as non-productive rearrangements of the opposite parental allele are also detected. Finally, mRNA allows for the identification of immunoglobulin isotypes, unlike gDNA, where the variable and constant regions are separated by introns[70,107]. Single-cell AIRR-seq is so far exclusively RNA-based, and therefore cell viability is critical to ensure proper encapsulation of RNA from each single cell[90]. Importantly, the quantity and quality of the sample's genetic material must be checked, as these two factors can determine whether the chosen template can be used in onward steps[108].

Amplification and library preparation

Benchmarking studies of DNA-based and RNA-based amplification methods have examined their specific advantages and disadvantages for bulk sequencing, highlighting the importance of methodology choice when developing an experimental design to study the AIRR composition[68–71,103,109,110] (Fig. 3d). Multiplex (MTPX) PCR, which is suitable for both gDNA and mRNA templates, uses a combination of J gene primers of the variable region or a C gene primer, along with a mixture of primers for known V genes. A downside to this method is the potential competition between the vast array of primers that are used in the reaction, causing some genes to be preferentially represented at the expense of others[102,105]. Rapid amplification of 5′ complementary DNA ends (5′ RACE) can be used to overcome this issue. 5′ RACE is only appropriate for mRNA templates. This approach relies on the reverse transcriptase activity to incorporate an adaptor sequence at the 5′ end of the cDNA. This adaptor is then used as a target region for subsequent nested PCR amplifications, which allows, in combination with use of primers that are complementary to C regions, the targeted amplification of transcripts in a V gene-independent manner[111]. Nonetheless, depending on the protocol using either ligation or incorporation, this method can come with a low adaptor incorporation efficiency[112]. Moreover, it is more prone to error as reverse transcriptase enzymes have higher error rates than the DNA polymerases used in gDNA-based MTPX methods[113]. Yet 5′ RACE strategies are becoming popular as they do not require gene-specific primers, thus allowing new allele variants to be captured.

Several commercial and in-house methods are now available for bulk AIRR-seq. MTPX-based kits are increasingly being offered by commercial providers, among which iRepertoire[114] and Adaptive Biotechnologies[66] pioneered the field, whereas Illumina, Archer, Cellecta and probably others are emerging in the field. 5′ RACE-based protocols are also provided by several commercial companies, although less so than MTPX. 5′ RACE kits are currently available, for example, from Takara Bio, NEB and BGI. Milaboratories now offers various MTPX and 5′ RACE products, for application in human and mouse samples. Validated in-house protocols have also been designed, based on either RACE[108,115–117] or MTPX[80,102,118,119].

Protocols for single-cell immune repertoire sequencing use the same library preparation techniques as those applied for bulk sequencing, but differ in their cell isolation strategy, which dictates the number of cells that can be studied per run[120]. For instance, plate-based methods, which sort single cells into 96-well or 384-well plates, enable reliable single-cell profiling, albeit at high costs for a limited number of cells per run[121]. In comparison, the widely used droplet-based approach proposed by 10x Genomics Chromium[122] and inDrop by Illumina[123] owes its popularity to the fact that it permits the encapsulation of up to 10^4–10^5 single cells in individual droplets, each containing reagents for cell lysis, reverse transcription and molecular tagging. Although the cost per cell is relatively lower for the droplet-based method as compared with other approaches, this comes at the expense of reduced sensitivity and increased probability of generating cell doublets[124,125]. Deeper throughput is now achievable through recently launched TCR repertoire profiling protocols[126,127] that allow the analysis of up to one million cells in a single experiment. These technologies, provided by Parse Biosciences and Omniscope, add barcodes to each transcript using four split-pool combinatorial barcoding steps, thus detaching from microfluidic approaches.

Although PCR reagents and protocols have been highly improved in recent years, allowing mostly unbiased amplification, biases can

still occur during library preparation, including a biased amplification of sequences with a particular composition[128]; a stochastic amplification of the DNA molecules; and technical artefacts generating erroneous sequences, also known as jackpot mutations[129]. Jackpot mutations may occur during the early cycles of PCR, leading to exponential amplification of the erroneous sequence. Artefacts can also stem from PCR template switching between similar sequences, producing hybrid sequences[130]. Experimental and computational error correction strategies[102,105,108,131–133], performed independently or combined, can help detect and eliminate these distortions. For example, the use of spike-in controls, which refer to biological or synthetic AIR sequences with defined characteristics such as clonal sequences and frequencies[102,109,134], can aid the detection of enzymatic efficiencies and amplification biases. Other strategies include the introduction of unique molecular identifiers (UMIs)[18,135] in RNA-based amplification methods. UMIs are short sequence barcodes that can be attached to each genetic template during cDNA synthesis, so that each molecule is tagged with a unique barcode. This makes it possible to identify duplicate sequences initially derived from the same template molecule, allowing an accurate quantification of each clonotype abundance and the correction of amplification errors[109,119,136,137].

Characterizing the obtained libraries to ensure successful amplification, purification and size selection of the desired product is the final checkpoint before sequencing. Libraries with a single clear peak at the appropriate fragment length, which depends on the amplification method and the type of the analysed AIR chain, are expected[138]. This can be analysed using gel-based systems along with the provider protocols, such as the ones proposed by Agilent (Bioanalyzer or TapeStation).

High-throughput sequencing

The selection of an appropriate HTS platform depends on the desired read length (short read versus long read), read type (single-end read for a partial amplicon coverage or paired-end read for full amplicon coverage) and sequencing depth, namely the number of transcripts that should be detected per sample, which depends on the diversity of the studied cell populations (Fig. 3e). The most appropriate sequencing platform would ensure accurate and sufficient repertoire coverage of the studied cell populations, while effectively managing error rate and experimental costs.

To perform reliable haplotyping of variable and constant regions, dedicated long-read (>1 kb) sequencing platforms are available[139–143]. Short-read sequencing, defined here as sequencing producing reads 150–300 bp long, can cover the full length of the CDR3 region and variable V(D)J region. For short-read platforms, base call quality is usually poor at the sequence ends, which may lead to sequencing errors. Reduction of sequencing error may be achieved by performing paired-end sequencing, which allows alignment of overlapping regions, at the expense of the clonotype depth when one of the reads is of poor quality. For example, paired-end short-reading sequencing with 2 × 300 bp permits a reliable V gene and V allele assignment as it covers the complete rearrangement by detecting the full CDR1 and CDR2 sequences (see Fig. 1) with SHMs. The position of the sequencing primers determines, in large part, the proportion of the V gene and C region that can be covered (C region coverage is only critical for immunoglobulin isotype determination[144,145]).

Alternatively, UMIs can be used to enable the correction of sequencing errors[109,119,136,137], which is particularly important when studying immunoglobulin intraclonal diversity and antibody evolution[109]. However, choosing the right cutoff for reads per UMI is a critical step, as a stringent threshold could result in a drastic decrease in repertoire coverage and the filtering out of potentially informative low-frequency reads[35,69,135,136]. Combining UMIs with deep sequencing can reduce such loss, a strategy that could be challenging when the studied cell population is quantitatively rare[69].

Sequencing depth diminishes as the number and concentration of libraries increase, owing to a finite read capacity per sequencing run. Moreover, over-sequencing has been shown to alter the clonal distribution of small samples and to generate noise[146]. Hence, the sequencing depth should be adjusted depending on the sample size and their diversity, as well as the number of sequenced samples in a single run. Conversely, although deeper sequencing is more appropriate when analysing large samples, sequencing replicates can ensure a higher coverage of the true repertoire richness and exploration of clonal overlap[35,46,147,148]. MTPX sequencing platforms that can sequence a higher number of samples with deep coverage in a single run reduce the cost but increase the risk of cross-sample contamination. However, this can be addressed via the use of unique dual indexes incorporated into library adaptors. These ensure accurate demultiplexing by filtering out reads resulting from index-hopping, a switch of unique dual indexes between libraries that is common in currently used HTS techniques[149–151].

Altogether, the choice of the biological sample type, starting template material, library preparation method and sequencing platform are all important considerations when planning an AIRR-seq experiment. Although no gold standards are yet established for any of these steps, it is recommended to process samples within the same project as uniformly as possible for a minimally biased experiment[105,131]. First, it is recommended that the same sample type is used across all samples of the same experiment, as the sample type affects the quality and quantity of the collected cells and nucleic material, possibly resulting in diversity variations between identical samples processed differently[152]. Second, abiding by a single sequencing protocol will help reduce variations due to sequencing errors and depth, and therefore ensure accurate comparison and interpretation of the AIRR-seq data, as both are often platform-dependent and technology-dependent[69,103,133]. Strategies such as the implementation of mixed cell populations as in-parallel biological controls have been proposed to detect and correct batch effects in AIRR-seq experiments[105,131]. For instance, by using a lymphoid cell (B cell or T cell) line mixture with predefined V(D)J rearrangements, the relative abundance can be compared with their predefined ones across different sample batches and sequencing runs[153].

Results

AIRR data preprocessing and analysis, starting from raw files provided as sequencing outputs to biological interpretation of the computational results obtained, are discussed in this section. The analysis part will be described by incremental level of granularity and complexity, and the extraordinary number of published tools for AIRR-seq data analysis are summarized in Supplementary Tables 1 and 2.

Data preprocessing

The main sequencing platforms usually produce FASTA or FASTQ output files that contain the unprocessed AIRR-seq data[154].

The main objectives of AIRR data preprocessing are to control sequencing data quality and correct PCR and sequencing errors, annotate germline alleles, assemble clonotypes based on a predefined sequence feature (for example, a specific V(D)J sequence) and export

the output in a human-readable tabular format where columns are sequence features (germline genes or alleles, FRs and CDR sequences) and rows are usually unique clonotypes (Fig. 3f). An elaborate listing of available tools for performing each of these steps is presented in Supplementary Table 1, with a mention of whether the tool supports bulk and/or single-cell AIRR data.

First, the error correction step consists of either filtering reads of low quality based on Phred scores (a measure of base call quality), clustering reads based on sequencing similarity or aggregating reads by UMI[109,136,155]. UMI-based aggregation can be performed by many processing software suites (Supplementary Table 1) and is used particularly to separate true mutations introduced by SHM from PCR and sequencing errors in BCR data. Gupta et al. provide a detailed UMI-based correction workflow for BCR sequencing data using pRESTO[138]. Second, after (or at times along with) error correction, germline gene or germline allele annotation is performed. Here, error-corrected reads are aligned to a species-specific germline database in order to identify, for each read, germline genes, FRs and CDRs, as well as SHMs (SHM count and SHM type) for BCR data. Whereas a single germline gene reference database is generally used, for all AIRR data from across donors, it is now becoming common practice to build reference databases for each donor in order to most accurately represent BCR and TCR germline alleles. Such germline allele-specific annotation can be of importance for downstream comparisons across individuals and for accurate representation of SHMs in BCR data, as individual-specific polymorphisms could be otherwise incorrectly identified as SHMs[156–162]. Although germline polymorphisms have been extensively studied for BCR genes[163–169], the allele analysis of TCRs is just starting[140,159]. Third, sequencing reads that share the predefined assembly feature are aggregated into a single clonotype, for which the abundance is extracted. Fourth, error-corrected and standardized data are output for downstream analysis. The standard output format is the MiAIRR format as developed by members of the AIRR Community (AIRR-C)[170,171]. Of note, AIRR data may also be reconstructed from bulk and scRNA-seq data, albeit with lower efficiency[172–174]. However, such workflows are not a focus of this Primer.

AIRR data exploration and analysis

AIRR data analysis encompasses different levels of granularity, from descriptive analysis to predictive modelling and inference of AIR specificities. The first step of AIRR data analysis involves the calculation of AIRR summary statistics, which mostly describe germline and clonal count information. Subsequently, more detailed analyses are performed focusing on AIRR diversity, AIRR composition similarity, clonal architecture and machine learning-assisted AIRR inference or predictions. A general overview of AIRR data analysis is provided in Fig. 4a and the different approaches and a non-exhaustive list of analytical tools are found in Supplementary Table 2. All tools listed in Supplementary Table 2 (except those listed in the single-cell analysis category) are applicable to bulk sequence data. Some tools outside the single-cell analysis category may also be used with single cells, although it remains an open question as to how to treat paired-chain data diversity, phylogenetics, clustering and machine learning method analyses.

AIRR summary statistics. Germline V, D and J gene usage (the frequency with which a given germline gene is used in a given AIRR) and CDR3 count information within a given AIRR represent fundamental AIRR descriptors. They can be studied with or without sequencing-read based weighting, which can add frequency-based information (if properly corrected for by UMI or other controls[69]). Except for minor variations, germline gene usage is usually stable in the naive compartment across individuals[35,175]. Germline gene usage has also been shown to be similar across different immune states at the peripheral blood mononuclear cell level[176]. However, differences have been observed across some B cell[35] and T cell subpopulations and cell development stages[177,178]. Stark differences in germline gene usage across individuals usually point to technical problems in the library preparation process[179,180]. Whereas germline gene usage is usually similar across individuals in a comparable state, CDR3 counts may vary extensively across samples due to technical biases, or, for example, when comparing cell populations of differing sizes (as, for example, naive versus antigen-experienced cells). Strong variation of CDR3 counts among samples where similar counts are expected are worth investigating and being adjusted prior to downstream data analysis as they might impact the biological conclusions drawn.

AIRR diversity. AIRR diversity is typically calculated using diversity measures that were first developed in ecology (to count and compare animal and plant abundances). These diversity measures both take species (for example, clonotype) richness (unique number of different species) and species abundance distribution into account. Briefly, the diversity of an AIRR of n clonotypes is calculated using the Hill diversity formula (the exponential of the Rényi entropy), which includes many of the commonly used diversity measures as special cases, defined as:

$$ {}^{\alpha}D(f) = \left(\sum_{i=1}^{n} f_i^{\alpha} \right)^{\frac{1}{1-\alpha}} $$

where f is the clonal frequency distribution and f_i is the frequency of each single clonotype, α is any real number greater than or equal to zero and n is the total number of clonotypes in the repertoire[181]. The higher the value of α, the higher the influence of the higher abundance clonotypes on diversity. Special cases of the Hill function relate to diversity indices in the AIRR field: the species richness index, the exponential Shannon–Wiener index, the inverse of the Simpson index, the Gini index, the Pielou index and the Berger–Parker index[182–189]. Two AIRRs may yield qualitatively different αD values depending on the diversity index used, due to the mathematical properties of the Hill diversity function (Schur concavity) (see ref. 181). Diversity profiles, which contain several diversity indices, are suggested to be more accurate compared with single diversity indices[181]. Estimating total AIRR diversity given an experimental sample remains an outstanding challenge to which no satisfying solutions have been proposed so far[19,155,190–192]. Diversity indices may also be used to measure the state of clonal expansion of an AIRR. For this, the Hill diversity values are divided by the sample's species richness, which results in a measure called 'evenness'. Evenness ranges between near zero and one, and quantifies to what extent the clonal frequency distribution (vector of clonal frequencies of an AIRR) is away from a uniform distribution.

Cross-sample diversity estimation can be performed using entropy-based diversity indices[181,190,193–195]. Such approaches revealed, for instance, that naive cell populations have high evenness, whereas antigen-experienced AIRRs have comparatively lower evenness[35,181].

AIRR clonal architecture. AIRR architecture defines the many-to-many sequence similarity landscape between all AIRs within an AIRR. Given the large sequence diversity of AIRRs, AIRR architecture analysis

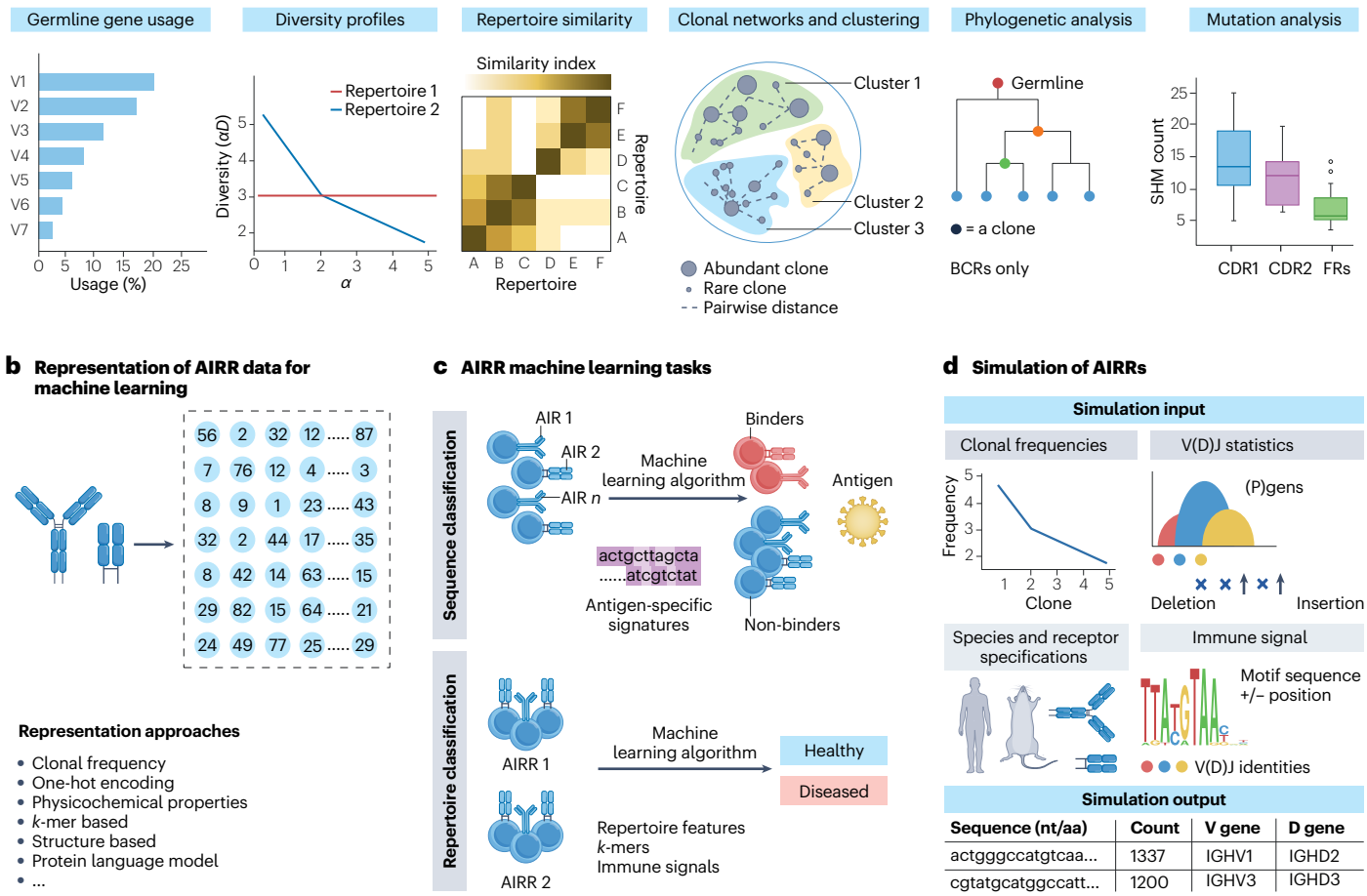

Fig. 4 | Bioinformatic downstream analyses of AIRR-seq data. a, Preprocessed adaptive immune receptor repertoire (AIRR) data harbour valuable information that can be exploited to conduct downstream comparative studies. This can include descriptive analyses (for example, V gene usage), repertoire similarity analyses (for example, clonal overlap) and diversity analyses (for example, diversity profiles and clonal networks). Mutational analysis and somatic hypermutation (SHM)-based phylogenetic studies can also be conducted on B cell receptor (BCR) sequencing data. The results of these analyses are highly dependent on the germline reference used. **b**, AIRR data representation in machine learning. Prior to machine learning analysis, AIRR sequence data require encoding in machine learning-readable format. Several encodings exist, which may take clonal frequency, sequence composition and physicochemical properties into account. **c**, Machine learning algorithms may be implemented on AIRR data to perform classification tasks including sequence classification (for example, antigen binding prediction) and repertoire classification (for example, immune state prediction; health/disease). **d**, Tunable simulation parameters can be implemented to generate synthetic adaptive immune receptor (AIR) sequences in silico. CDR, complementary determining region; FR, framework region; IgH, immunoglobulin heavy chain; V(D)J, variable, (diversity) and joining.

enables identifying low and high sequence similarity regions of the AIRR and relate those to, for instance, antigen binding. Architecture is mathematically based on the all clonotypes versus all clonotypes distance matrix. This distance matrix may be transformed into a network in which nodes are clones and edges represent a chosen sequence similarity cutoff. This cutoff can be calculated using global similarity approaches across the whole sequence. These include the Levenshtein and the Hamming distances, with a default score usually set to one nucleotide or one amino acid difference, although larger distances have been explored[196,197]. It becomes computationally expensive to calculate distance matrices that exceed 10^5 clonotypes, which is the order of magnitude of state-of-the-art experimental AIRR data sets[35,49,148]. For the tasks of large-scale clonal network calculation, the imNet pipeline was developed[196]. Networks of a few thousand nodes may be visualized using software suites such as igraph[198], NetworkX[199], Gephi[200] and Cytoscape[201]. Graph properties and network analysis quantify AIRR architecture properties[196,202,203]. Architecture properties may be analysed at the AIRR level (generally one coefficient per AIRR network) or describe the AIRR at the AIR (clonotype) level (one coefficient per AIR per AIRR)[196]. AIRR-level coefficients include degree distribution, clustering coefficient, diameter and assortativity[196]. The degree of a node is the number of its edges (that is, the number of similar clonotypes to a given clonotype), and an AIR's degree distribution quantifies the abundance of node degrees (that is, clonal similarities) across clonotypes of a repertoire. For example, power law networks have a few highly connected clonotypes and many clonotypes with few connections,

Glossary

Adaptive immune receptor repertoire
(AIRR). The collection of adaptive immune receptors in a single individual at a single point in time.

Adaptive immune receptors
(AIRs). B cell receptors, antibodies and T cell receptors.

Class-switch recombinations
Processes by which proliferating B cells change their antibody production by rearranging the constant region genes in the immunoglobulin heavy chain (IgH) locus to switch from expressing one class of immunoglobulin to another. The produced isotype retains the same antigen specificity but has different effector properties.

Clonotypes
Definitions range from the exact amino acid complementary determining region 3 (CDR3) to clusters of sequences to the sequence of entire variable chain regions. The debate on what constitutes a clonotype is ongoing and beyond the scope of this Primer.

Epitope
The specific part of an antigen that is contacted and recognized by an adaptive immune receptor (AIR).

Generation probability
Probability for observing a given recombined adaptive immune receptor sequence.

Germline alleles
Variants of variable (V), diversity (D) and joining (J) genes, representing the building blocks of recombined variable regions of a B cell receptor/T cell receptor.

Ground truth
An environment where any parameter (and the value thereof) that contributed to training data generation is known and controlled.

Paratope
The set of amino acids in an adaptive immune receptor (AIR) that contribute to antigen/epitope binding and are in direct contact with the epitope during binding.

Peptide–MHC complex
(pMHC). The major histocompatibility complex (MHC) is a highly polymorphic region of the genome that encodes MHC cell surface proteins that present antigenic peptides. T cell receptors recognize and bind peptides that are presented by the MHC. We denote a peptide when presented by the MHC as a pMHC.

Private clonotypes
Adaptive immune receptor sequences that occur exclusively in the adaptive immune receptor repertoire of a single individual.

Public clonotypes
Adaptive immune receptor (AIR) sequences that occur more than n times ($n>1$) across a set of adaptive immune receptor repertoires (AIRRs) collected from different individuals.

Sequencing depth
The number of sequencing reads for a given sample.

Somatic hypermutations
(SHMs). Processes that lead to mutation(s) in the variable, (diversity) and joining (V(D)J) recombined B cell receptor (BCR) sequences, taking place predominantly in anatomical locales called germinal centres, and may be associated with the selection for improved BCR binding of a specific antigen.

Unique dual indexes
Unique pairs of i5 and i7 index primers used for filtering out index-hopped or misassigned reads post sequencing.

Unique molecular identifiers
(UMIs). Short sequences added to DNA/RNA fragments in some high-throughput sequencing library preparation protocols to identify the input DNA/RNA molecule, used to reduce errors and quantitative bias introduced by PCR amplification.

and this network architecture may relate to antigen-driven clonal expansion, whereas exponential networks have a rather even degree distribution across clonotypes more reflective of naive repertoires[196].

Complementarily, AIR-level parameters such as PageRank[196] quantify the importance of the similarity between two CDR3 clonotypes within a clonal network. Extensive mathematical descriptions of network parameters may be found elsewhere[196,204,205]. Of note, AIR-level similarity measures may be generalized to identifying short amino acid motifs. The rationale behind this approach is that short stretches of amino acids, also known as *k*-mers, contribute to the epitope binding affinity[206,207]. Clustering of similar AIRs, for instance, by shared *k*-mers, has thus become a popular method to attempt to identify antigen or epitope-specific receptors[208,209], recently also by including transcriptome information[210–213].

B cell phylogenetics. When exposed to antigens, B cells undergo expansion and hypermutation in their BCR variable regions. This process leads to the development of a B cell lineage, ranging from naive unmutated B cells to memory B cells and plasma cells that have undergone SHM. Studying the evolution of antibody repertoires provides valuable insights into how vaccines and pathogens influence the body's humoral immune response[214,215].

To deduce the ancestral evolutionary connections among individual B cells, lineage trees are created using sequences from a clonal lineage. A clonal lineage is determined by the number of receptor sequences originating from the same recombination event, indicating shared ancestry. When constructing a lineage tree, a typical preprocessing step involves grouping sequences with identical V and J genes and CDR3 length. However, the specifics of this process may vary depending on lineage and clone definitions[24]. Lineage trees may also be identified in a data-driven fashion[216,217]. Standard algorithms for inferring phylogenetic trees that use maximum parsimony and maximum likelihood are often employed in B cell phylogenetic analyses, but it remains challenging to ascertain that a given method has inferred the biologically accurate tree[192,218–220] To account for the unique biology of B cells, more context-aware (for example, favouring hot spots, disfavouring cold spots[221]) phylogenetic methods such as IgPhyML[222] have been developed. Furthermore, BCR repertoires often contain hundreds of independent clones, and standard phylogenetic models consider clonal lineages individually, which can compromise efficiency. The use of repertoire-wide models, which allow some parameters to be shared among the multiple clonal lineages, can improve model precision[222]. Recently, a statistical framework was developed to characterize migration, differentiation and isotype switching along B cell phylogenetic trees, and this framework is implemented in the R package entitled Dowser[223], which now enables inference of B cell phylogenies from paired heavy and light chain BCR sequences, along with other tools[224–226]. Third, B cell

population data have variable clonal abundances, and incorporating clone abundance may be important for accurate tree inference. A few tools use sequence abundance information for phylogenetic tree analysis[227,228]. Finally, visualization of immunoglobulin trees may be performed by various existing tools[229–231]. For more information on immunoglobulin phylogenetic tree analysis, please refer to the following articles[191,192,215,223,232].

Similarity of AIRR composition. Comparison of AIRR composition, at the level of germline gene or CDR3 sequences, is of major interest for the identification of clonotypes that are shared across cell populations or tissues of a given individual or across individuals The presence of such clonotypes, commonly termed public clonotypes, in naive repertoires can be, in part, causally linked to V(D)J recombination statistics[46,233] or convergent recombination[234,235]. For instance, shorter CDR3s tend to have higher generation probability and are thus more likely to be generated and observed[46,236]. Public clonotypes may also reflect preferential central selection[37], or antigen-driven selection, hence the observation of shared clonotypes between individuals in the context of the same immunological encounter[49,237,238] or disease[80,239]. Approaches for repertoire comparison include the measurement of clonal overlap with indices that exclusively consider the presence/absence of the compared repertoire level (such as the Jaccard index[240])[241,242], or additionally consider the frequency information (such as the Morisita–Horn index[243] or the Jensen–Shannon divergence index)[244,245]. Fast identification of public clonotypes, especially across large data sets, can be performed with the tool CompAIRR, which also enables fast identification of similar sequences, that is, clonotypes that differ in a few amino acids across samples[242]. Of interest, structure-based analyses suggest that structure-based similarity may be higher than sequence-based similarity across AIRRs[246].

Recently, basic repertoire statistics and diversity measures have been augmented with sequence-based similarity information[195,247] to account for the highly similar sequences when measuring inter-AIRR and intra-AIRR similarity. Even more generally, immuneREF has been introduced as a tool to measure inter-AIRR similarity by integrating multiple AIRR and sequence and frequency features including gene usage, clonal expansion and clonal overlap. These features allow researchers to interpret differences between immune repertoires using in silico and experimental immunologically interpretable ground truth[176]. However, given that small differences in sequence similarity may lead to differences in antigen binding, such measures may not accurately represent AIRR diversity if considered from the antigen binding perspective.

With the advent of deep learning approaches over the past few years, antibody structure predictions based on the sequence alone have become more commonplace[248,249]. Specifically, there now exist antibody-specific (and to lesser extent TCR-specific) structure prediction tools that enable large-scale prediction of hundreds of thousands or even millions of antibody structures enabling the repertoire-scale structure-based comparison of AIRRs[250–253]. Structure-based AIRR comparison is of heightened interest as the three-dimensional structure of an AIR determines the interaction with an antigen, governing its binding properties[254–256]. AIRs with similar sequences can adopt different conformations and vice versa[257,258]. Of note, although AIR structure prediction methods are steadily improving in performance[249], prediction performance decreases with CDR3 length[248,259] or may suffer from structural inaccuracies such as incorrect *cis*-amide bonds, wrong stereochemistry or clashes[260].

AIRR data-based predictive analysis

AIRRs are both determinants and sensors of health and disease, but their complex architecture hinders straightforward access to features that are associated with antigen binding or the resulting immune response and thereby determine immunity-related outcomes. These AIRR features are collectively referred to, here, as AIRR motifs[261–263]. These immune signals are usually situated in the CDR3 region. Machine learning tools employ pattern recognition and function approximation techniques to identify patterns within groups in (large amounts of) data and were proposed for predictive AIRR analysis more than a decade ago[264,265]. Machine learning can discover statistical associations, for example between AIRR data and immune status or epitope binding, and these associations ideally enable generalizable predictions, aiming not only at developing a model with high predictive performance but also at obtaining biological insights into AIR biology. Therefore, there is a desire for machine learning models to be interpretable. There has been a surge in machine and deep learning methods that can be applied to investigate how immune signal information is encoded in the AIRR[261,266]. Figure 4b–d illustrate these approaches.

Sequence-based and repertoire-based machine learning applications. AIRR-based machine learning may be roughly divided into repertoire-based and sequence-based machine learning tasks. A few of these tools are mentioned non-exhaustively in Supplementary Table 2. Machine learning techniques based on sequence analysis concentrate on classifying AIR sequences using sequence-level labels, such as antigen (epitope) specificity or shared occurrence at the population level. Sequence-based machine learning predictions may be applied for drug discovery, for the in silico design of antigen therapeutics, antibody therapeutics and TCR therapeutics[261,267,268] or, potentially, also for repertoire-wide antigen-specific sequence annotation[269,270]. Repertoire-based machine learning methods and applications emphasize AIRR-based classification and predicting donor immune status. This includes identifying factors such as disease presence, recent vaccinations or prior exposure to specific pathogens. These techniques find significant utility in the field of immunodiagnostics. Nevertheless, repertoire-based machine learning may also be used to infer disease status-associated AIRs or AIR sequence motifs[49,271,272].

AIRR data encoding and embedding for AIRR machine learning analysis. AIR sequences are chains of amino acids of different lengths. Data encoding is the process of assigning a numerical value to each amino acid of a protein sequence, to convert the sequence into a format that can be used by a machine learning algorithm. There are several ways to perform encoding, including one-hot based[273], *k*-mer based[207,233], amino acid-scale based[274] or even whole-sequence based[49]. More recently, neural networks were applied to produce data encodings that are called embeddings and represent sequential data in a high-dimensional vector space. The process of creating an embedding involves mapping each sequence to a point in this vector space, such that similar sequences are close to each other in the space, and dissimilar sequences are far apart (similarity here may not be defined by sequence similarity such as edit distance but, for example, a function, such as binding similarity). Embeddings are commonly used in natural language processing to represent words or phrases in a continuous vector space. In this context, the embedding represents the meaning or context of the word or phrase. One of the most popular methods for creating embeddings is training a neural network to predict a certain variable based on the categorical or discrete input.

Recently, protein language models, which are trained on millions of protein or AIR sequences, have shown great promise for embedding, clustering, predicting and generating protein/AIR function[275–282] as they seem to capture long-range dependencies well beyond conventional sequence similarity. Of interest, joint encoding of sequence and structure has been shown to improve the prediction of paratope and epitope interaction both for antibodies[206,273] and TCRs[283]. Analogously, joint embedding of the AIR sequence and transcriptome profile was suggested to reveal interdependencies between the TCR sequence and transcriptome, allowing for the identification of T cell clusters with previously unidentified disease specificity[284].

AIRR machine learning basic workflow. After having opted between repertoire-based and sequence-based machine learning methods, the basic workflow for AIRR machine learning (AIRR-ML) consistently involves a data preparation step, where the data are gathered and cleaned, and split into a training set, a validation set and a test set. The data preparation step is followed by feature engineering and selection. Feature engineering is the process of creating representations of data that increase the effectiveness of a model[285]. This may include selecting important variables (feature selection), scaling, normalization and encoding of the data. Then, the machine learning model is trained on the training data, with the aim of minimizing the difference between the predicted output and the actual output. Model evaluation is performed on the validation data, using metrics such as accuracy, precision, recall and F1 score[286]. Based on validation results, the model may be fine-tuned to improve performance. Model optimization could involve adjusting the hyperparameters of the model, changing the learning rate or using a different algorithm. Once the model is optimized, it is evaluated on the test data to check its generalization performance. This step helps ensure that the model is not overfitting to the training data and can perform well on new, unseen data, and may involve cross-validation. The AIRR-ML basic workflow is iterative and may involve going back to previous steps to make adjustments based on the evaluation results (for example, nested cross-validation). To streamline the AIRR-ML workflow, immuneML was developed. This tool is an open-source software ecosystem comprising fully specified and shareable workflows. immuneML is available as a command-line tool, is provided through an intuitive Galaxy web interface and contains extensive documentation of workflows, all to promote its widespread use[287]. Specifically, it allows large-scale benchmarking of AIRR-ML methods, which can uncover current blank spots in AIRR-ML development that warrant further investigation[288,289].

Applications

AIRR analyses are currently applied to address various basic and biomedical questions. In this section, the applications are illustrated following the type of analysis methods detailed in the Results section. A synthetic illustration is depicted in Fig. 5.

AIRR diversity

The mechanisms that underlie generation of the large AIRR diversity remain incompletely understood. Nevertheless, the probability with which a given AIR sequence can be generated by V(D)J recombination, also called the generation probability (P_{gen}), can be quantified[290]. A probabilistic model that learns on non-productive rearrangements was developed to estimate the generation probability of each rearrangement event, encompassing segment choice, gene trimming, nucleotide insertions and chain pairing[29,290]. This model was implemented in OLGA (Optimized Likelihood estimate of immunoGlobulin Amino acid sequences), a tool that allows the attribution of a P_{gen} value to any given TCR or BCR CDR3 sequence[29,236,290,291]. Use of OLGA showed that all rearrangements are not generated with equal probabilities, with some highly probable rearrangements that are specific to viral epitopes. Consistently, the thymus was shown to preferentially generate TCRs that are able to interact with multiple and unrelated human viruses[76]. These observations suggest that the AIRR is not stochastically diverse but, rather, skewed towards a highly protective and balanced entity. More recently, the immunoglobulin V(D)J recombination rules and sequence generation probabilities were shown to differ in monozygotic twins or in inbred mice, and this suggested that non-genetic factors, such as epigenetics, influence the recombination process[292]. These observations are additional indicators of the complexity of the AIRRs. Of note, current models of repertoire generation do not account for the recently observed impact of germline AIR gene polymorphisms on V(D)J gene usage and the functional B cell and T cell repertoire[167,293].

AIRR diversity is shaped by selection events in the thymus for T cells and in the bone marrow for B cells, and further accentuated by antigen exposure in the periphery. In mice, repertoire diversity was shown to largely differ between antigen-experienced plasma cells and antigen-inexperienced pre-B cells and splenic naive B cells, highlighting the impact of antigenic encounters on the clonal distribution[35]. Such differences in the BCR repertoires can stem from B cell clonal expansions, but also from class-switch recombinations or SHMs. Indeed, although most of the top expanded clonotypes were shown to be antigen-specific, antigen binding and non-binding clonotypes are evenly distributed among the rest of the repertoire[294]. Furthermore, antibody–antigen binding affinity did not correlate with clonal expansion or SMH, when analysed at a polyclonal and a clonal lineage level. Although SHMs do not occur in the TCR repertoire, the study of naive and activated regulatory T cell repertoires in mice in a physiological context revealed different levels of clonal diversity, mainly caused by increased clonal expansions upon activation[93].

In a pathological context, investigating the AIRR diversity can shed light on the dynamics of the various adaptive immune cell subsets and their implication in particular diseases, and help monitor patient responses to treatment. For instance, studies reported increased clonal expansions within the $CD8^+$TRB rearrangement repertoire in cerebrospinal fluid and peripheral blood of patients with multiple sclerosis[295]. Similar observations were described in the context of type 1 diabetes, particularly showing decreased TCRβ chain repertoire diversity in the pancreatic islets and lymph nodes of patients with type 1 diabetes[296,297]. These results could reconcile over the idea that the disease pathogenesis is T cell-dependent and driven by potential tissue-specific antigens. Increased clonal expansion has also been reported for the BCR repertoire in patients with Crohn's disease and systemic lupus, but not in patients with ANCA-associated vasculitis or IgA vasculitis when compared with healthy individuals[80].

Clonal architecture

Studying the AIR sequence similarity can reveal information about the repertoire clonal architecture in health, as well as its dynamics under pathological conditions. AIRs can be clustered based on sequence similarity, shared amino acid motifs and/or physicochemical properties. Naive B cell and T cell repertoires were found to form highly connected networks around conserved public sequences[196,202], which have been linked to skewed repertoire generation and selection[76,95,177,196]. Conversely, repertoires of antigen-experienced cells were shown to exhibit a lower level of sequence similarities across individuals, and this reflects

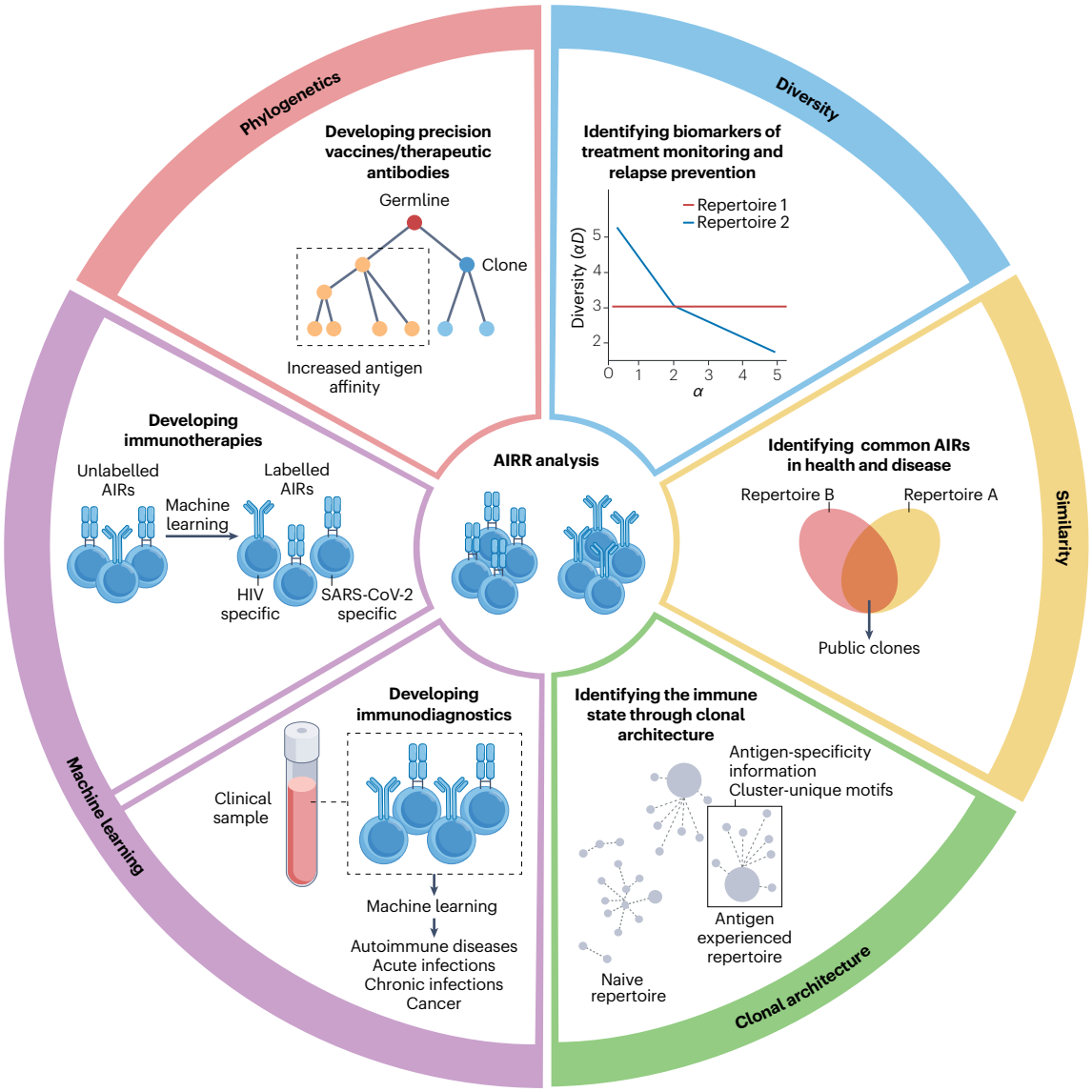

Fig. 5 | Applications of AIRR analysis towards disease diagnostics and immunotherapy development. Different applications of adaptive immune receptor repertoire via sequencing (AIRR-seq) analysis are shown for the development and set-up of novel diagnostics tools as well for the immunotherapy development. Most of the applications can indistinctly arise from B cell receptor (BCR) sequencing and T cell receptor (TCR) sequencing, except phylogenetics given the absence of somatic hypermutations (SHMs) in TCRs. Importantly, all the applications can benefit from the sequencing of functionally distinct cell subsets, which increases the precision/targeting of the treatment approach, for instance. AIR, adaptive immune receptor.

an individual's history of antigenic stimulations and ongoing immune responses[178,298]. Changes in clonal architecture have been observed following antigenic challenge[196], infection with human immunodeficiency virus (HIV) or SARS-CoV-2 (refs. 299,300) and therapeutic antibody treatment[203,241], as well as during tumour development[301]. Moreover, evaluating sequence connectivity relative to the baseline expectation from V(D)J recombination statistics enabled the identification of responding or contracting clonotypes in the peripheral blood of patients receiving cancer immunotherapy[302] and of individuals with SARS-CoV-2 infection[303]. Furthermore, studies have demonstrated that similar sequences are highly likely to recognize the same peptide–MHC complex (pMHC) ligand[202,258,304,305] and may, thus, form specificity groups. Therefore, matching the sequences of specificity groups identified based on sequence similarity with antigen-annotated public data sets might help infer antigen specificity for these groups. This strategy has enabled the identification of groups of sequences enriched in individuals with viral infections[209,306], autoimmune disease[307,308] and cancer[309,310] compared with healthy donors. To summarize, the study of the AIRR sequence architecture in a pathological context, alone or combined with functionally tested antigen-specific sequences from public databases, helps identify motifs across immune responses or AIR specificities and affinities that are implicated in a particular disease.

B cell phylogenetics

The construction of phylogenetic lineage trees can help address questions about B cell clonal expansion, SHM and antigen-driven selection. For example, B cell phylogenetics has been applied to the analysis of differentiation of B cell subtypes during infection with HIV or on sequential isotype switching in the context of food allergy[223]. Moreover, a phylogenetic test of measurable immunoglobulin sequence evolution has identified measurable evolution specifically in germinal centre B cells following influenza vaccination and invalidated the assumption that the inability to induce influenza-specific B cell evolution underlies the generally poor efficacy of seasonal influenza vaccination[311].

In addition, phylogenetic methods that identify sequences that share a common unmutated ancestor have been used to predict antibody affinities and key affinity-increasing mutations[165]. A phylogenetic approach that quantifies tree dissimilarity (Unifrac) has been applied to the repertoires of young and older individuals at baseline and during influenza vaccination, and immunosenescence — the ageing of B cell repertoires — has been demonstrated to be associated with a contracted naive repertoire and diminished intra-lineage diversification[312].

In another example, the parental clones of B cells that reside in the central nervous system of patients with multiple sclerosis were found to arise in the periphery, a finding that raises the question of whether lymphocytes activated in the periphery could be the main drivers of the disease[313]. Understanding such a process could help elucidate how therapeutics that target peripheral B cells can impact B cells that populate the tissues that are affected in a particular disease[313].

Similarity of AIRR composition

As described above, public clonotypes are commonly found across repertoires within and across individuals, and their presence can be due to preferential recombination and/or central selection, and a peripheral antigen-specific selection following an infection or in the context of a chronic disease. Thus, identifying and characterizing public clonotypes could reveal common receptor selection patterns in health and disease[35,38,314]. For instance, CDR3β overlap within the most abundant TCR sequences of thymocytes was found to increase after thymic selection[38]. This observation highlighted the role of thymic selection in preferentially selecting certain sequences, regardless of the cell subset in which they are found. Public clonotypes are also found in unrelated individuals, although twins tend to share higher proportions of their repertoires. Importantly, public clonotypes are a core component of immune responses to vaccination[237,315,316] and infection[49,238,317,318] or in the context of autoimmune diseases[80,239,297,319,320] and malignancies[321,322]. For example, humans exposed to the same antigen showed convergent BCR evolution[315], resulting in the establishment of public clonotypes that harboured protective antibodies, potentially specific against the challenging antigen. These findings could eventually be used for the development of therapeutic antibodies.

Although it has been recognized that both BCR and TCR diversity decreases with age[312,323,324], there are only a few studies exploring the dynamics of repertoires over shorter or longer periods of time[325]. A high degree of clonal persistence has been demonstrated in individual memory B cell subsets across a time span of several months[326]. Furthermore, the identification of some SARS-CoV-2-reactive T cell clones in the memory compartment at a pre-infection time point has indicated the participation of pre-existing cross-reactive memory T cells in the immune response to SARS-CoV-2 (ref. 303). More generally, analysis of BCR repertoires in healthy individuals over the course of 1 month has revealed considerable variation within and across individuals[327].

AIRR-ML applications

The high AIRR diversity and the non-negligible AIR sequence similarity across individuals led immunologists to suggest that convergent AIRR features may contribute to the overall maintenance of health status and that eventual enrichment or loss of such features might contribute to loss of homeostasis. Machine learning approaches can be used to identify such features On the repertoire level, one of the first convincing proofs of principle that AIRRs may be used for disease classification was provided upon the identification of a public TCRβ chain signature from peripheral blood of a cohort of approximately 600 individuals who were CMV^+ and CMV^-[49]. A similar pattern was identified for memory $CD4^+$ T cells in an independent cohort, providing evidence that public TCRs are closely involved in the pathogen-specific T cell response[328]. Interestingly, the CMV-specific TCRβ chain signature was only composed of 164 sequences, and classification accuracy dropped from >90% to nearly random (≈50%) when only one third of the original data were used, demonstrating that large-scale data sets are necessary for detecting immune status-associated immune signals[49,287]. Indeed, a machine learning analysis of ~1,000 synthetic AIRR data sets comprising ≈250,000 AIRRs across different parameters, such as signal occurrences and repertoire size, showed that comparatively simple machine learning algorithms such as L1-penalized logistic regression are able to achieve high prediction accuracy even when a public clonotype occurs only in 1 out of 50,000 AIR sequences. So far, there exist only a few large-scale AIRR data sets. Such large-scale AIRR data sets from of 877 patients with systemic lupus erythematosus and 206 patients with rheumatoid arthritis could be used to differentiate between these autoimmune conditions based on TCRβ chain repertoires[329], whereas AIRR data from patients with COVID-19 (n = 1,815) and healthy individuals (n = 3,500)[330] revealed patterns specific to COVID-19 both early after diagnosis and after recovery. Whereas these approaches only relied on detection of immune status (in other words, detection of disease) based on sharing of public clonotypes, a more recent method leveraged three different machine learning representations, namely overall AIRR composition, convergent clustering of antigen-specific sequences by edit distance and language model feature extraction from BCR and TCR sequences, to classify individuals with SARS-CoV-2 (n = 63), HIV (n = 95) and systemic lupus erythematosus (n = 86) and healthy controls (n = 217)[331].

Sequence-based prediction of AIR–antigen binding may be performed at the sequence level or at the structure level (or with a hybrid approach). Most sequence-based approaches have been applied to the problem of predicting AIR–antigen binding (such as antibody–antigen prediction or TCR–pMHC prediction). Extensive reviews have been published for both antibody–antigen binding[261,267,332,333] and TCR–pMHC predictions[28,334,335]. As T cell epitopes are mostly linear, TCR–pMHC binding prediction approaches have mainly involved sequence-based prediction[28,334–339], with only a few more recent clustering and machine learning approaches also exploring the incorporation of structural data[283,340,341]. In addition to sequence-based AIR–antigen binding prediction, simulation tools based on AIRR sequences, such as IGoR[290], OLGA[342] and immuneSIM[180], enable the generation of large numbers of AIRR sequences with moderate computational resources. These tools offer the advantage of generating native-like sequence data that are nearly identical to experimental data. ImmuneSIM, simAIRR and LIgO, in particular, allow the incorporation of sequence motifs into the generated sequences, enabling the modelling of motifs associated with antigen binding. Consequently, these simulated data can be employed for tasks related to predicting AIR specificity, either in

a binary or multi-class fashion. These predictions involve classifying sequences based on their antigen binding behaviour (see Use Cases 1 and 2 in ref. 273) or for AIRR-based machine learning with applications to immunodiagnostics[287,288,343].

However, although sequence-based data sets are comparatively easier to generate than structural data sets[344–348], sequence-based machine learning often lacks granularity on paratope and conformational epitope binding, and this makes it challenging to resolve conformational antibody–antigen binding[254]. Structural information is used either implicitly in the construction of the features by facilitating epitope identification, using for example gapped *k*-mer encoding, or explicitly by direct incorporation into the machine learning task. More fine-grained information about residue-wise influence on binding may also be gained via deep mutational scanning, which unravels an incredibly complex AIR–antigen binding landscape[51,349–352]. A key future application for AIR–antigen binding prediction is the in silico annotation of AIRR data sets with antigen binding information. This enables quantitative diagnostic profiling for antigen or epitope specificity, and comparison across individuals, antigens and immune states[270,353–356]. In addition to AIR–antigen binding predictions, sequence-based AIRR-ML approaches can be applied to design new AIR sequences. This may find application to simulations[357], design of improved immunotherapy agents[358,359] or antibody drug development[360].

Reproducibility and data deposition

The accumulation and promising potential of AIRR-seq data spurred scientists and industrialists to define common experimental and computational standards and controls for conducting AIRR studies[170,361], as well as for harmonizing data storage and sharing[361–363]. The AIRR-C was established in 2015 as a research-driven group that organizes and coordinates the use of HTS technologies to streamline AIRR-seq study design. Its primary mission is to develop guidelines and standards for the generation, annotation and storage of AIRR-seq data to facilitate its use by the larger research community.

Experimental reproducibility

As any emerging field of research, the experimental procedures involved in AIRR studies were developed at a fast pace without initial standardization, which later hindered performing comparative analysis across studies and data sets[105]. Additionally, the high complexity of the experimental work involved — including biological sample preservation, cellular cytometry, (targeted) nucleic acid isolation, primer design and concentration, PCR reaction, sequencing technology and others — highlighted the several possible biases and errors that can arise during AIRR-seq library preparation and data analysis[131]. Thus, standards and controls are needed for AIRR-seq data generation to provide a key level of reproducibility and minimize experimental errors. This pressing need led to the establishment of the Biological Resources Working Group within the AIRR-C, which aims to develop controls and strategies to streamline AIRR-seq research[105,170,361].

These strategies (further detailed in ref. 105), although not relevant to all experimental platforms or scenarios, include the use of sample-specific barcodes to detect sample crosstalk, the use of standardized sample preparation kits, be they commercial or custom-made, when available as they offer standardized analytical materials and optimized experimental procedures, and the integration of spike-in control sequences within the AIRR-seq library. Although promising, the latter strategy is most effective only when the control sequences model the natural diversity and complexity of AIRRs, while still being distinguishable from the AIRRs of study samples[105]. The implementation of in-parallel biological controls, such as a human lymphoid cell mixture which better captures the diversity of AIRs, provides step-by-step quality monitoring for AIRR-seq library generation and sequencing[105,131,364]. Nevertheless, the identity of genetic rearrangements in this control are not predefined, which can be problematic in the case of PCR and/or sample contamination.

Data sharing and computational reproducibility

As the quantity of AIRR-seq data is growing, providing the community with the raw data sets and their corresponding metadata can facilitate their reuse for secondary analysis or their integration into comparable data sets for greater statistical power. This can support advances in computational strategies, particularly machine learning-based methods, and drive novel scientific discoveries. AIRR-seq data sharing under the FAIR principle (Findability, Accessibility, Interoperability, and Reusability) is a way to ensure reliable and accurate data quality[365]. Although more peer-reviewed journals require raw data to be made publicly available, there is still a long way to go as, for example, until 2022 only 38.1% of TCR sequencing studies have made the raw data available[366]. Standardization of metadata formats could be the key for encouraging researchers to share their raw data, by creating a straightforward ecosystem of databases that can be interchangeably used for data input and output reading. In this context, the AIRR-C has developed data standards (MiAIRR; AIRR file format)[170,171], that uphold reproducibility, standard quality control and data deposition in a shared repository. These standards guide the publication, curation and sharing of AIRR-seq data and metadata. Metadata columns include study and subject information, details about sample collection, processing, sequencing, raw sequences, sequence data processing and processed AIRR sequences[170]. Notably, it is possible to submit AIRR-seq data in the AIRR file format standard to the National Center for Biotechnology Information (NCBI) (see Guide for submission of AIRR-seq data to NCBI)[367]. Additionally, to facilitate data sharing, the AIRR-C has established the AIRR Data Commons (ADC)[368], comprising geographically dispersed AIRR-compliant repositories adhering to AIRR Standards[363]. The ADC interface operates as a web-based query API, making AIRR-seq studies and their associated annotated sequence data in the ADC easily discoverable and accessible. By employing MiAIRR Standards[170] and AIRR file formats[171], the ADC enhances interoperability and data reuse, promoting reproducibility and enabling meta-analysis. The ADC can be explored interactively using the iReceptor gateway web user interface[362]. Apart from large-scale databases, which mostly contain antigen non-annotated data, there are smaller databases with antigen annotation for TCRs[347,348,369] or BCRs[344,346,347,370] (see Supplementary Table 3).

The AIRR-C has also implemented standards for AIRR software tools to ensure that AIRR data standards can be used seamlessly. Tools that comply with the established standards, detailed on the AIRR-C website, can be labelled as AIRR compliant (guidance for AIRR software tools). Currently, there exist nine AIRR-compliant software tools.

Limitations and optimizations

The field of AIRR-seq has been evolving rapidly in the past few years, increasing our understanding of the effect of diseases on our adaptive immune responses. Nevertheless, several technological limitations currently remain (Fig. 6). Limitations are defined as broad-scope shortcomings in data generation and interpretation. The Outlook section below outlines how specific limitations can be addressed in the AIRR field.

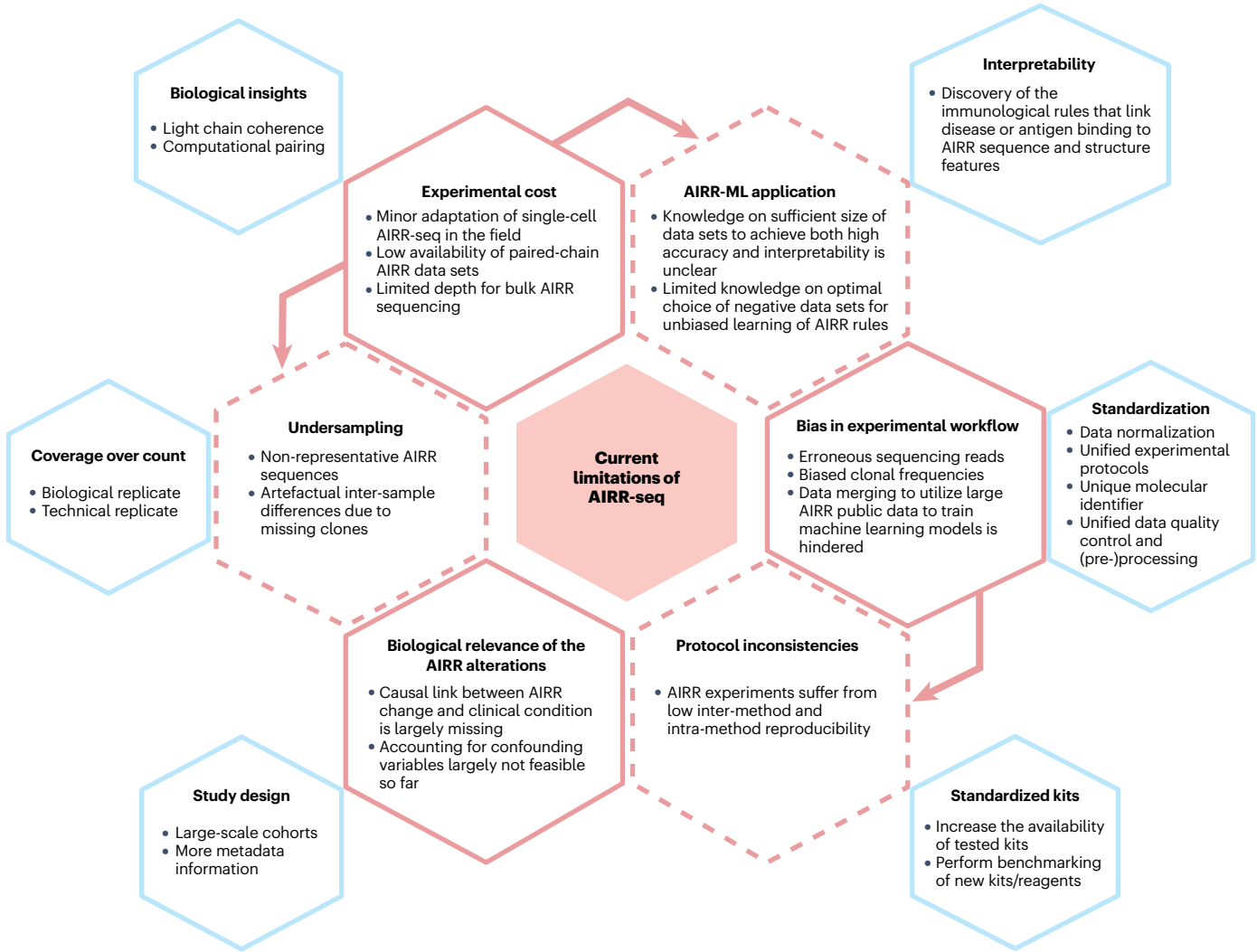

Fig. 6 | Current limitations and their workarounds in the field of AIRR-seq. The major limitations that the field of adaptive immune receptor repertoire sequencing (AIRR-seq) is currently facing are represented (solid red hexagons). Major limitations can exacerbate others (dashed red hexagons), as indicated by a red arrow. Potential/possible workarounds are shown (blue hexagon) for the neighbouring limitation (red hexagon). AIRR-ML, AIRR machine learning.

Experimental biases and data normalization

Technical and biological biases render repertoire cross-comparisons challenging, and inevitable biological heterogeneity exists between samples depending on their source (different individuals, biological compartments, cell subsets)[69,103,105,155,371–373]. As mentioned in 'Experimentation', experimental biases can be introduced during the sequencing workflow (Fig. 3), for example, during the RNA extraction, reverse transcription and PCR amplification steps. Errors introduced during reverse transcription and PCR amplification steps, such as nucleotide misincorporations and amplification biases, may affect the quantification and AIR diversity estimation[105,108]. To identify and mitigate these biases, several strategies can be applied, such as incorporating spike-in controls and UMIs during library preparation[105,135,136]. The use of bioinformatic algorithms such as NoisET[374] can also help reduce amplification biases in AIRR-seq experiments[192,372].

Different protocols may introduce different types of experimental biases, and protocol inconsistencies lead to low reproducibility within and across data sets[69,105], thereby hindering AIRR data comparisons and integration. The preferential use of commercial kits over homebrew methods could help minimize experimental inconsistencies. Biases may also be corrected computationally. For example, the iROAR[372] tool was developed with the vision of generating evergreen data sets that can be merged and compared across. AIRR-ML approaches have been shown to rely on large sample numbers for optimal prediction accuracy[287] as well as well-annotated metadata[343]. Therefore, being able to correct batch effects to combine and utilize existing public data sets would be invaluable for training these machine learning models. Mathematical and statistical methods to correct experimental biases are currently lacking in AIRR-seq studies. A method that accurately discards uninformative sequenced TCR reads based on the Shannon index has been proposed[146], but it does not perform read count corrections.

Normalization strategies developed for other HTS data, such as transcriptome or microbiome sequencing[375], could be considered. Indeed, AIRR data exhibit high species diversity, data sparsity and little overlap between samples, compared with microbiome data.

Experimental cost

The high costs of single-cell AIRR-seq might explain the lack of adaptation of single-cell AIRR-seq studies in academic research and the scant paired-chain AIRR with assigned specificities available in public databases[376]. To this end, mathematical approaches are being developed to acquire TCR paired-chain information based on clonal frequency estimations and observational likelihoods of given pairs from semi-bulk sequencing[377,378]. To further improve this approach, scientists have been investigating the underlying rules of chain pairing[29,376], although no general rules of AIR pairing have been revealed so far[379–382].

High experimental costs in bulk sequencing, particularly in RNA-based approaches, which require higher sequencing depths, could lead to undersampling, namely insufficient read coverage per sample. Increasing the sequencing depth is not the best solution as this can increase sequencing errors, resulting in skewed clone size or, in the case of BCR, skewed SHM profiles, and/or alter the clonal distribution of small samples by generating noise. The best alternative proposed so far is to incorporate technical and/or biological replicates to assist rare clone detection and noise correction[46,292,373,374].

Predicting the antigenic epitopes that are recognized by a given AIR is of interest for the development of immunotherapies. As most of the current prediction approaches are sequence based, the limited number of antigen-annotated sequences in the curated public databases compared with the high diversity of all possible receptor–epitope pairs represents a hurdle for the training of accurate machine learning models. Possible workarounds to this challenge include the use of generic AIR–epitope recognition models to reveal binding patterns and rules[258,338,383]. Nevertheless, a single mutation in the epitope sequence can impact the binding range and receptor affinity, which requires the training of a specific model for each available epitope[269,270,287]. Yet the imbalanced epitope distribution in these databases makes it a challenging approach. Recently, deep mutational scanning experiments were introduced to model the paratope–epitope interactions and gain insight into the complex AIR–antigen binding landscape[51,349–352].

Validation of biological causality

AIRR-seq data analysis has been providing correlations with clinical conditions[49,207,329,384] or underlying genetic background[385–387], but has so far failed to establish causal links[343]. Approaches for biology-based encoding of AIRR data[277] and causal modelling of AIRR data are probably needed to overcome this limitation. A fundamental impediment to larger-scale causality-driven studies is the lack of large-scale cohorts with complete metadata information[388], which would enable controlling for sample selection and confounder variables such as age and sex[343].

AIRR-ML optimization

Both sequence-based and repertoire-based machine learning benchmarking efforts have indicated that further optimization is needed. The lowest bound of sample size required remains to be evaluated[287,288], AIRR-biology adapted machine learning methods must be developed to identify more complex AIRR patterns[271,288], optimal negative data sets still need to be defined[337,389,390] and unbiased approaches to estimate prediction accuracy will be needed[383,391].

Furthermore, interpretation of both sequence-based and repertoire-based AIRR-ML results remains challenging. Strong prediction accuracy indicates that there is an immune signal in the AIRR data set that differs between two labels (such as health and disease, or binders and non-binders). The next step is to understand what underlying AIRR features contribute to the high prediction accuracy (such as the binding rules[273,392]). Interpretability is tied to the encoding of the data and architecture of the AIRR-ML model. For example, by decomposing AIRR data into *k*-mers, one may remove existing amino acid interactions within the sequence, thus potentially removing biological information. Moreover, it becomes increasingly difficult to establish a link between data features and data set labels when complex AIRR-ML models are used. Therefore, simpler machine learning architectures may be preferable for interpretability purposes. Indicatively, models based on amino acid 3-mers with distinct biophysicochemical characteristics and enriched V and J genes have been sufficient to distinguish between patients with coeliac disease and healthy individuals[384]. Models using only specific sub-regions in the CDRH3 have been able to drive classification of public clonotypes and private clonotypes[233], and models based on specific germline V genes could demonstrate how CDR3 varies across immune status and rank sequences based on their likelihood of being associated with a given immune state[331].

It has been shown that the analysis of interpretability results may be complicated by confirmation bias[393], which describes a phenomenon where the researcher unconsciously injects pre-existing beliefs and hypotheses into the analysis. To address this, it is crucial to verify with ground truth (which in most cases would involve synthetic data[394]) that hypotheses drawn from the candidate explanation reflect the intended logic[395].

Outlook

Integrating antibody proteomics into AIRR studies

Proteomics methods, such as mass spectrometry, can be applied to analyse the diversity of antibodies in the blood or mucosal tissues[396,397]. Combining bulk and single-cell BCR sequencing with antibody profiling has the potential to capture humoral immunity in its entirety[47,145,396–399]. De novo protein sequencing, fuelled by deep learning advances, has the potential to revolutionize antibody profiling and, unlike mass spectrometry analyses, does not involve profile deconvolution based on potentially biased BCR sequencing databases[400,401].

Understanding pathophysiology

High-throughput antigen annotation and antigen binding prediction. Although current public databases of AIR sequences and structures are expanding rapidly in size, the majority of the stored data have not been annotated for antigen binding. This limits our knowledge about the variation of antigen specificities across individuals and immune states, as well as about the frequency of antigen-specific AIR for different antigens[156,157].

Antigen-specific AIR-sequencing and structural biology technologies[51,250,268,402–407] as well as growing interdisciplinary expertise in systems immunology, statistics and machine learning are now beginning to offer the opportunity to resolve these questions. Particularly, single-cell sequencing has helped characterize T cells that express multiple TCR chains[408–410], estimated to represent up to 20% of all T cells[411], potentially unravelling the mechanisms underlying multiple TCR chain expression, and predicting epitope specificities[412].

Large-scale antigen-annotated AIRR data may soon enable the development of computational and machine learning methods that

predict antigen specificity from the AIR sequence or structure, unlocking the currently inaccessible antigen-specific information in publicly available AIRR data[268,332]. Importantly, these approaches will further our understanding of AIRR cross-reactivity[76,349,413–416].

In the past decade, numerous competitions have arisen with the aim of addressing unresolved research questions about protein structure, interaction and function prediction, such as Critical Assessment of Structure Prediction (CASP)[417] or Critical Assessment of PRediction of Interactions (CAPRI)[418], or for artificial intelligence-based image recognition, such as ImageNet challenge[419], among others. These competitions have facilitated groundbreaking discoveries such as AlphaFold[420] and may help sharpen our currently insufficient tools for predicting antigen-specific adaptive immunity[391,421].

Integration of AIRR data with transcriptomic data. Integrating AIRR with transcriptomics has recently been employed to understand better how the transcriptional profile of a cell is correlated to AIR–antigen binding. To that end, several groups have reported methods that integrate AIRR and transcriptomics[211–213,284]. Preliminary results suggest that antigen binding specificity and the transcriptional profile may be linked[213,284], and that transcriptomic information may increase pMHC–TCR epitope prediction accuracy. In the future, it remains to be understood to what extent the MHC background influences the interplay of transcriptome and TCR sequence specificity.

The genotype–phenotype link in adaptive immunity. Recently, the link between the germline gene repertoire and humoral immune response has been investigated in depth. For example, immunogens that activate specific germline precursors that have a high likelihood of affinity maturing into broadly neutralizing antibodies have shown promise for development of precision vaccines against major human pathogens[422,423]. The magnitude of the response to germline gene targeting vaccines could be explained to a large degree by the frequencies of the various immunoglobulin genotypes and corresponding B cells rather than by the immunogen dose[157,424,425]. Thus, immunoglobulin allelic variations must be considered when designing and testing germ line-targeting immunogens in clinical trials. Immunologically, these results suggest that genetic variation of the host can modulate the strength of vaccine-induced broadly neutralizing antibody responses. It will be important to understand the evolution and selection of germline gene variants in order to design more targeted vaccines.

The HLA genotype has been associated with various disease susceptibilities[426]. Moreover, the HLA genotype shapes the TCR repertoire of a given individual[427–429]. As recently observed, HLA alleles can influence the composition of the TCRβ chain[387], and HLA type can be predicted based on the presence of some unique TCRs[49,385]. Yet we recently found that TCRs previously known to be restricted to a given HLA type could recognize antigen presented by unmatched HLAs[76]. Although progress in transgenic mice is being made, increasing the number of deep sequencing data sets from paired or unpaired TCRα and TCRβ chain repertoires combined with HLA genotype is needed to provide more accurate knowledge on the association of HLA genotype and TCR repertoires[430].

Translational perspective

AIRR-seq combining advanced statistical and mathematical modelling, including machine learning, can now provide toolkits for the identification of AIRR signatures associated with disease, serological status or response to treatment[49,78,329,331,384,431,432]. This has been shown in various disease indications (cancer, autoimmune disease, transplantation and infection) as well as in response to different therapeutic approaches[40]. Now, the future challenge in the field would be to turn these research-oriented approaches towards clinical application. In the field of B cell and T cell malignancies, the EuroClonality Consortium has already made major progress towards diagnosis and prognosis evaluation through AIRR-seq in clinics[118,153,364]. Both stakeholder support in patient care and collaborative interdisciplinary efforts are essential to achieve success.

Bulk and single-cell approaches are now being employed to mine AIRRs for antigen-specific antibodies[250,294,433] that are as close to human antibody repertoires as possible[434]. Recently, generative machine learning approaches, which involve learning antibody language[277] and antigen-specific binding patterns, are used to generate novel[358,435] or improved[280,436] antibodies, thus replacing experimental with computational antibody discovery. Furthermore, over the past decade adoptive T cell therapy, such as adoptive T cell therapy with CAR T cells or engineered T cells, has gained insight. In non-solid tumour cancers, adoptive T cell therapies targeting well-known cell surface antigens expressed by lymphoma cells, such as CD20 or CD19, are already in clinics[437–439]. Such approaches could be expanded to solid cancers, when AIRR-seq combined with antigen prediction identifies tumour targets that can specifically drive the engineered T cell or the CAR T cell to the tumour[440,441]. Alternatively, AIRR-seq, especially at the single-cell level, could help track CAR T cells in patients and better understand treatment efficacy or failure[442]. Similarly, regulatory CAR T cells are being developed for the treatment of autoimmune diseases[443]. Regulatory T cell-based therapies would also benefit from AIRR-seq analyses aiming at increasing targeting efficacy[444].

To summarize, in our view, there is a need for the AIRR field to go beyond the current predominantly antigen-agnostic analysis of AIRR sequence data. The next frontier is fully antigen-annotated AIRR data analysis. Only when this goal is reached can we really begin to understand the specificity and function of adaptive immunity in health and disease. To achieve this goal, novel breakthroughs in high-throughput AIRR data generation and computational analysis are necessary. Furthermore, whereas the AIRR field has been successful in integrating long-established concepts from other fields, such as diversity analysis from ecology[187,445], there is a need for richer perspectives. For example, given the current findings on extensive and immunity-relevant germline gene diversity[156,158,161], understanding how evolution has shaped the human immune system may lead to evolutionary medicine-driven AIRR-based therapeutics and diagnostics design[156,446–448].

Published online: 25 January 2024

References

1. Rappazzo, C. G. et al. Defining and studying B cell receptor and TCR interactions. *J. Immunol.* **211**, 311–322 (2023).
2. Schroeder, H. W. Jr & Cavacini, L. Structure and function of immunoglobulins. *J. Allergy Clin. Immunol.* **125**, S41–S52 (2010).
3. Kuhns, M. S., Davis, M. M. & Garcia, K. C. Deconstructing the form and function of the TCR/CD3 complex. *Immunity* **24**, 133–139 (2006).
4. Ribot, J. C., Lopes, N. & Silva-Santos, B. γδ T cells in tissue physiology and surveillance. *Nat. Rev. Immunol.* **21**, 221–232 (2021).
5. Boehme, L., Roels, J. & Taghon, T. Development of γδ T cells in the thymus — a human perspective. *Semin. Immunol.* **61–64**, 101662 (2022).
6. Bosselut, R. A beginner's guide to T cell development. *Methods Mol. Biol.* **2580**, 3–24 (2023).
7. Deseke, M. & Prinz, I. Ligand recognition by the γδ TCR and discrimination between homeostasis and stress conditions. *Cell. Mol. Immunol.* **17**, 914–924 (2020).
8. Willcox, B. E. & Willcox, C. R. γδ TCR ligands: the quest to solve a 500-million-year-old mystery. *Nat. Immunol.* **20**, 121–128 (2019).
9. Cooper, M. D. & Alder, M. N. The evolution of adaptive immune systems. *Cell* **124**, 815–822 (2006).

10. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* **11**, 47–59 (2010).
11. Miller, J. F. Immunological function of the thymus. *Lancet* **2**, 748–749 (1961).
12. Miller, J. F. A. P. Effect of neonatal thymectomy on the immunological responsiveness of the mouse. *Proc. R. Soc. Lond.* **156**, 415–428 (1962).
13. Cooper, M. D., Peterson, R. D. & Good, R. A. Delineation of the thymic and bursal lymphoid systems in the chicken. *Nature* **205**, 143–146 (1965).
14. Miller, J. F., Mitchell, G. F. & Weiss, N. S. Cellular basis of the immunological defects in thymectomized mice. *Nature* **214**, 992–997 (1967).
15. Tonegawa, S. Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).
16. Chien, Y. et al. A third type of murine T-cell receptor gene. *Nature* **312**, 31–35 (1984).
17. Alamyar, E., Giudicelli, V., Li, S., Duroux, P. & Lefranc, M.-P. 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 Res.* **8**, 26 (2012).
18. Chaudhary, N. & Wesemann, D. R. Analyzing immunoglobulin repertoires. *Front. Immunol.* **9**, 462 (2018).
19. Chiffelle, J. et al. T-cell repertoire analysis and metrics of diversity and clonality. *Curr. Opin. Biotechnol.* **65**, 284–295 (2020).
20. Marks, C. & Deane, C. M. Antibody H3 structure prediction. *Comput. Struct. Biotechnol. J.* **15**, 222–231 (2017).
21. Davis, M. M. & Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature* **334**, 395–402 (1988).
22. Xu, J. L. & Davis, M. M. Diversity in the CDR3 region of V_H is sufficient for most antibody specificities. *Immunity* **13**, 37–45 (2000).
23. McKean, D. et al. Generation of antibody diversity in the immune response of BALB/c mice to influenza virus hemagglutinin. *Proc. Natl Acad. Sci. USA* **81**, 3180–3184 (1984).
24. Hershberg, U. & Prak, E. T. L. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140239 (2015).
25. Muramatsu, M. et al. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563 (2000).
26. Stavnezer, J., Guikema, J. E. J. & Schrader, C. E. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* **26**, 261–292 (2008).
27. Ng, J. C. F. et al. sciCSR infers B cell state transition and predicts class-switch recombination dynamics using single-cell transcriptomic data. *Nat. Methods* https://doi.org/10.1038/s41592-023-02060-1 (2023).
28. Bradley, P. & Thomas, P. G. Using T cell receptor repertoires to understand the principles of adaptive immune recognition. *Annu. Rev. Immunol.* **37**, 547–570 (2019).
29. Dupic, T., Marcou, Q., Walczak, A. M. & Mora, T. Genesis of the αβ T-cell receptor. *PLoS Comput. Biol.* **15**, e1006874 (2019).
30. Elhanati, Y. et al. Inferring processes underlying B-cell repertoire diversity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20140243 (2015).
31. Mora, T. & Walczak, A. M. Quantifying lymphocyte receptor diversity. Preprint at *bioRxiv* https://doi.org/10.1101/046870 (2016).
32. Trepel, F. Number and distribution of lymphocytes in man. A critical analysis. *Klin. Wochenschr.* **52**, 511–515 (1974).
33. Cosgrove, J., Hustin, L. S. P., de Boer, R. J. & Perié, L. Hematopoiesis in numbers. *Trends Immunol.* **42**, 1100–1112 (2021).
34. Sender, R. et al. The total mass, number, and distribution of immune cells in the human body. *Proc. Natl Acad. Sci. USA* **120**, e2308511120 (2023).
35. Greiff, V. et al. Systems analysis reveals high genetic and antigen-driven predetermination of antibody repertoires throughout B cell development. *Cell Rep.* **19**, 1467–1478 (2017).
36. Klein, L., Kyewski, B., Allen, P. M. & Hogquist, K. A. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).
37. Qi, Q. et al. Diversity and clonal selection in the human T-cell repertoire. *Proc. Natl Acad. Sci. USA* **111**, 13139–13144 (2014).
38. Khosravi-Maharlooei, M. et al. Crossreactive public TCR sequences undergo positive selection in the human thymic repertoire. *J. Clin. Invest.* **129**, 2446–2462 (2019).
39. Brown, A. J. et al. Augmenting adaptive immunity: progress and challenges in the quantitative engineering and analysis of adaptive immune receptor repertoires. *Mol. Syst. Des. Eng.* **4**, 701–736 (2019).
40. Arnaout, R. A., Prak, E. T. L., Schwab, N., Rubelt, F. & Adaptive Immune Receptor Repertoire Community. The future of blood testing is the immunome. *Front. Immunol.* **12**, 626793 (2021).
41. Zinkernagel, R. M. On differences between immunity and immunological memory. *Curr. Opin. Immunol.* **14**, 523–536 (2002).
42. Galson, J. D. et al. Analysis of B cell repertoire dynamics following hepatitis B vaccination in humans, and enrichment of vaccine-specific antibody sequences. *EBioMedicine* **2**, 2070–2079 (2015).
43. Setliff, I. et al. Multi-donor longitudinal antibody repertoire sequencing reveals the existence of public antibody clonotypes in HIV-1 infection. *Cell Host Microbe* **23**, 845–854.e6 (2018).
44. Sui, W. et al. Composition and variation analysis of the TCR β-chain CDR3 repertoire in systemic lupus erythematosus using high-throughput sequencing. *Mol. Immunol.* **67**, 455–464 (2015).
45. Madi, A. et al. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Res.* **24**, 1603–1612 (2014).
46. Briney, B., Inderbitzin, A., Joyce, C. & Burton, D. R. Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* **566**, 393–397 (2019).
47. Lee, J. et al. Persistent antibody clonotypes dominate the serum response to influenza over multiple years and repeated vaccinations. *Cell Host Microbe* **25**, 367–376.e5 (2019).
48. Bournazos, S. et al. Antibody fucosylation predicts disease severity in secondary dengue infection. *Science* **372**, 1102–1105 (2021).
49. Emerson, R. O. et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat. Genet.* **49**, 659–665 (2017).
50. Cowell, L. G. The diagnostic, prognostic, and therapeutic potential of adaptive immune receptor repertoire profiling in cancer. *Cancer Res* **80**, 643–654 (2020).
51. Mason, D. M. et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nat. Biomed. Eng.* **5**, 600–612 (2021).
52. Rapoport, A. P. et al. NY-ESO-1-specific TCR-engineered T cells mediate sustained antigen-specific antitumor effects in myeloma. *Nat. Med.* **21**, 914–921 (2015).
53. Schuster, S. J. et al. Tisagenlecleucel in adult relapsed or refractory diffuse large B-cell lymphoma. *N. Engl. J. Med.* **380**, 45–56 (2019).
54. Morrissey, K. A., Stammnitz, M. R., Murchison, E. & Miller, R. D. Comparative genomics of the T cell receptor μ locus in marsupials and monotremes. *Immunogenetics* **75**, 507–515 (2023).
55. Parra, Z. E., Mitchell, K., Dalloul, R. A. & Miller, R. D. A second TCRδ locus in Galliformes uses antibody-like V domains: insight into the evolution of TCRδ and TCRμ genes in tetrapods. *J. Immunol.* **188**, 3912–3919 (2012).
56. Ott, J. A., Harrison, J., Flajnik, M. F. & Criscitiello, M. F. Nurse shark T-cell receptors employ somatic hypermutation preferentially to alter α/δ variable segments associated with α constant region. *Eur. J. Immunol.* **50**, 1307–1320 (2020).
57. Castro, R. et al. Clonotypic IgH response against systemic viral infection in pronephros and spleen of a teleost fish. *J. Immunol.* **208**, 2573–2582 (2022).
58. Castro, R. et al. Contrasted TCRβ diversity of $CD8^+$ and $CD8^-$ T cells in rainbow trout. *PLoS ONE* **8**, e60175 (2013).
59. Burnet, S. F. M. *The Clonal Selection Theory of Acquired Immunity; The Abraham Flexner Lectures of Vanderbilt University* (Cambridge Univ. Press, 1959).
60. Liu, S. et al. Spatial maps of T cell receptors and transcriptomes reveal distinct immune niches and interactions in the adaptive immune response. *Immunity* **55**, 1940–1952.e5 (2022).
61. Heumos, L. et al. Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.* **24**, 550–572 (2023).
62. Faint, J. M. et al. Quantitative flow cytometry for the analysis of T cell receptor Vβ chain expression. *J. Immunol. Methods* **225**, 53–60 (1999).
63. Pannetier, C. et al. The sizes of the CDR3 hypervariable regions of the murine T-cell receptor β chains vary as a function of the recombined germ-line segments. *Proc. Natl Acad. Sci. USA* **90**, 4319–4323 (1993).
64. Pannetier, C., Even, J. & Kourilsky, P. T-cell repertoire diversity and clonal expansions in normal and clinical samples. *Immunol. Today* **16**, 176–181 (1995).
65. Weinstein, J. A., Jiang, N., White, R. A. III, Fisher, D. S. & Quake, S. R. High-throughput sequencing of the zebrafish antibody repertoire. *Science* **324**, 807–810 (2009).
66. Robins, H. S. et al. Comprehensive assessment of T-cell receptor β-chain diversity in αβ T cells. *Blood* **114**, 4099–4107 (2009).
67. Boyd, S. D. et al. Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Sci. Transl. Med.* **1**, 12ra23 (2009).
68. Rosati, E. et al. Overview of methodologies for T-cell receptor repertoire analysis. *BMC Biotechnol.* **17**, 61 (2017).
69. Barennes, P. et al. Benchmarking of T cell receptor repertoire profiling methods reveals large systematic biases. *Nat. Biotechnol.* **39**, 236–245 (2021).
70. Vázquez Bernat, N. et al. High-quality library preparation for NGS-based immunoglobulin germline gene inference and repertoire expression analysis. *Front. Immunol.* **10**, 660 (2019).
71. Genolet, R. et al. TCR sequencing and cloning methods for repertoire analysis and isolation of tumor-reactive TCRs. *Cell Rep. Methods* **3**, 100459 (2023).
72. Pai, J. A. & Satpathy, A. T. High-throughput and single-cell T cell receptor sequencing technologies. *Nat. Methods* **18**, 881–892 (2021).
73. Friedensohn, S., Khan, T. A. & Reddy, S. T. Advanced methodologies in high-throughput sequencing of immune repertoires. *Trends Biotechnol.* **35**, 203–214 (2017).
74. Wang, Y. et al. Multimodal single-cell and whole-genome sequencing of small, frozen clinical specimens. *Nat. Genet.* **55**, 19–25 (2023).
75. Miron, M. et al. Maintenance of the human memory T cell repertoire by subset and tissue site. *Genome Med.* **13**, 100 (2021).
76. Quiniou, V. et al. Human thymopoiesis produces polyspecific $CD8^+$ α/β T cells responding to multiple viral antigens. *eLife* **12**, e81274 (2023).
77. Meng, W. et al. An atlas of B-cell clonal distribution in the human body. *Nat. Biotechnol.* **35**, 879–884 (2017).
78. Liu, Y.-Y. et al. Characteristics and prognostic significance of profiling the peripheral blood T-cell receptor repertoire in patients with advanced lung cancer. *Int. J. Cancer* **145**, 1423–1431 (2019).
79. Rossetti, M. et al. TCR repertoire sequencing identifies synovial T_{reg} cell clonotypes in the bloodstream during active inflammation in human arthritis. *Ann. Rheum. Dis.* **76**, 435–441 (2017).
80. Bashford-Rogers, R. J. M. et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* **574**, 122–126 (2019).
81. Mastrokolias, A., den Dunnen, J. T., van Ommen, G. B., 't Hoen, P. A. C. & van Roon-Mom, W. M. C. Increased sensitivity of next generation sequencing-based

expression profiling after globin reduction in human blood RNA. *BMC Genomics* **13**, 28 (2012).
82. Valpione, S. et al. Immune-awakening revealed by peripheral T cell dynamics after one cycle of immunotherapy. *Nat. Cancer* **1**, 210–221 (2020).
83. Miljkovic, M. D. et al. Next-generation sequencing-based monitoring of circulating tumor DNA reveals clonotypic heterogeneity in untreated PTCL. *Blood Adv.* **5**, 4198–4210 (2021).
84. Komech, E. A. et al. TCR repertoire profiling revealed antigen-driven $CD8^+$ T cell clonal groups shared in synovial fluid of patients with spondyloarthritis. *Front. Immunol.* **13**, 973243 (2022).
85. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
86. Greenfield, A. L. et al. Longitudinally persistent cerebrospinal fluid B cells can resist treatment in multiple sclerosis. *JCI Insight* **4**, e126599 (2019).
87. Meednu, N. et al. Dynamic spectrum of ectopic lymphoid B cell activation and hypermutation in the RA synovium characterized by NR4A nuclear receptor expression. *Cell Rep.* **39**, 110766 (2022).
88. Gros, A. et al. PD-1 identifies the patient-specific $CD8^+$ tumor-reactive repertoire infiltrating human tumors. *J. Clin. Invest.* **124**, 2246–2259 (2014).
89. Bai, X. et al. Characteristics of tumor infiltrating lymphocyte and circulating lymphocyte repertoires in pancreatic cancer by the sequencing of T cell receptors. *Sci. Rep.* **5**, 13664 (2015).
90. Langerak, A. W. *Immunogenetics* (Springer US, 2022).
91. Klein, U., Küppers, R. & Rajewsky, K. Evidence for a large compartment of IgM-expressing memory B cells in humans. *Blood* **89**, 1288–1298 (1997).
92. Shi, W. et al. Transcriptional profiling of mouse B cell terminal differentiation defines a signature for antibody-secreting plasma cells. *Nat. Immunol.* **16**, 663–673 (2015).
93. Bergot, A.-S. et al. TCR sequences and tissue distribution discriminate the subsets of naïve and activated/memory T_{reg} cells in mice. *Eur. J. Immunol.* **45**, 1524–1534 (2015).
94. Muraro, P. A. et al. T cell repertoire following autologous stem cell transplantation for multiple sclerosis. *J. Clin. Invest.* **124**, 1168–1172 (2014).
95. Bashford-Rogers, R. J. M. et al. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res.* **23**, 1874–1884 (2013).
96. Ghraichy, M. et al. Different B cell subpopulations show distinct patterns in their IgH repertoire metrics. *eLife* **10**, e73111 (2021).
97. King, H. W. et al. Single-cell analysis of human B cell maturation predicts how antibody class switching shapes selection dynamics. *Sci Immunol* **6**, eabe6291 (2021).
98. Han, A., Glanville, J., Hansmann, L. & Davis, M. M. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nat. Biotechnol.* **32**, 684–692 (2014).
99. Munson, D. J. et al. Identification of shared TCR sequences from T cells in human breast cancer using emulsion RT-PCR. *Proc. Natl Acad. Sci. USA* **113**, 8272–8277 (2016).
100. He, B. et al. Rapid isolation and immune profiling of SARS-CoV-2 specific memory B cell in convalescent COVID-19 patients via LIBRA-seq. *Signal. Transduct. Target. Ther.* **6**, 195 (2021).
101. Zhang, S.-Q. et al. High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nat. Biotechnol.* https://doi.org/10.1038/nbt.4282 (2018).
102. Carlson, C. S. et al. Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).
103. Bashford-Rogers, R. J. M. et al. Capturing needles in haystacks: a comparison of B-cell receptor sequencing methods. *BMC Immunol.* **15**, 29 (2014).
104. Chovanec, P. et al. Unbiased quantification of immunoglobulin diversity at the DNA level with VDJ-seq. *Nat. Protoc.* **13**, 1232–1252 (2018).
105. Trück, J. et al. Biological controls for standardization and interpretation of adaptive immune receptor repertoire profiling. *eLife* **10**, e66274 (2021).
106. Menzel, U. et al. Comprehensive evaluation and optimization of amplicon library preparation methods for high-throughput antibody sequencing. *PLoS ONE* **9**, e96727 (2014).
107. Ford, E. E. et al. FLAIRR-seq: a method for single-molecule resolution of near full-length antibody H chain repertoires. *J. Immunol.* **210**, 1607–1619 (2023).
108. Mamedov, I. Z. et al. Preparing unbiased T-cell receptor and antibody cDNA libraries for the deep next generation sequencing profiling. *Front. Immunol.* **4**, 456 (2013).
109. Khan, T. A. et al. Accurate and predictive antibody repertoire profiling by molecular amplification fingerprinting. *Sci. Adv.* **2**, e1501371 (2016).
110. Liu, X. et al. Systematic comparative evaluation of methods for investigating the TCRβ repertoire. *PLoS ONE* **11**, e0152464 (2016).
111. Schaefer, B. C. Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. *Anal. Biochem.* **227**, 255–273 (1995).
112. Lin, Y.-H. et al. Dissecting efficiency of a 5′ rapid amplification of cDNA ends (5′-RACE) approach for profiling T-cell receptor β repertoire. *PLoS ONE* **15**, e0236366 (2020).
113. Ellefson, J. W. et al. Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science* **352**, 1590–1593 (2016).
114. Wang, C. et al. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proc. Natl Acad. Sci. USA* **107**, 1518–1523 (2010).
115. Heather, J. M. et al. Dynamic perturbations of the T-cell receptor repertoire in chronic HIV infection and following antiretroviral therapy. *Front. Immunol.* **6**, 644 (2015).
116. Douek, D. C. et al. A novel approach to the analysis of specificity, clonality, and frequency of HIV-specific T cell responses reveals a potential mechanism for control of viral escape. *J. Immunol.* **168**, 3099–3104 (2002).
117. Turchaninova, M. A. et al. High-quality full-length immunoglobulin profiling with unique molecular barcoding. *Nat. Protoc.* **11**, 1599–1616 (2016).
118. Scheijen, B. et al. Next-generation sequencing of immunoglobulin gene rearrangements for clonality assessment: a technical feasibility study by EuroClonality-NGS. *Leukemia* **33**, 2227–2240 (2019).
119. Baker, A.-M. et al. FUME-TCRseq: sensitive and accurate sequencing of the T-cell receptor from limited input of degraded RNA. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.24.538037 (2023).
120. Gupta, N. et al. Single-cell analysis and tracking of antigen-specific T cells: integrating paired chain AIRR-seq and transcriptome sequencing: a method by the AIRR community. *Methods Mol. Biol.* **2453**, 379–421 (2022).
121. Fan, H. C., Fu, G. K. & Fodor, S. P. A. Combinatorial labeling of single cells for gene expression cytometry. *Science* **347**, 1258367 (2015).
122. Zheng, G. X. Y. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049–14049 (2017).
123. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
124. Mereu, E. et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nat. Biotechnol.* **38**, 747–755 (2020).
125. Yamawaki, T. M. et al. Systematic comparison of high-throughput single-cell RNA-seq methods for immune cell profiling. *BMC Genomics* **22**, 66 (2021).
126. Nadeu, F. et al. Detection of early seeding of Richter transformation in chronic lymphocytic leukemia. *Nat. Med.* **28**, 1662–1671 (2022).
127. Ostendorf, B. N. et al. Common human genetic variants of *APOE* impact murine COVID-19 mortality. *Nature* **611**, 346–351 (2022).
128. Aird, D. et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
129. Kebschull, J. M. & Zador, A. M. Sources of PCR-induced distortions in high-throughput sequencing data sets. *Nucleic Acids Res.* **43**, e143 (2015).
130. Pääbo, S., Irwin, D. M. & Wilson, A. C. DNA damage promotes jumping between templates during enzymatic amplification. *J. Biol. Chem.* **265**, 4718–4721 (1990).
131. Eugster, A. et al. AIRR community guide to planning and performing AIRR-seq experiments. *Methods Mol. Biol.* **2453**, 261–278 (2022).
132. Koraichi, M. B., Touzel, M. P., Mazzolini, A., Mora, T. & Walczak, A. M. NoisET: noise learning and expansion detection of T-cell receptors. *J. Phys. Chem.* **126**, 7407–7414 (2022).
133. Rosenfeld, A. M. et al. Computational evaluation of B-cell clone sizes in bulk populations. *Front. Immunol.* **9**, 1472 (2018).
134. Friedensohn, S. et al. 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. *Front. Immunol.* **9**, 1401 (2018).
135. Vollmers, C., Sit, R. V., Weinstein, J. A., Dekker, C. L. & Quake, S. R. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc. Natl Acad. Sci. USA* **110**, 13463–13468 (2013).
136. Shugay, M. et al. Towards error-free profiling of immune repertoires. *Nat. Methods* **11**, 653–655 (2014).
137. Egorov, E. S. et al. Quantitative profiling of immune repertoires for minor lymphocyte counts using unique molecular identifiers. *J. Immunol.* **194**, 6155–6163 (2015).
138. Gupta, N. et al. Bulk sequencing from mRNA with UMI for evaluation of B-cell isotype and clonal evolution: a method by the AIRR community. *Methods Mol. Biol.* **2453**, 345–377 (2022).
139. Subas Satish, H. P. et al. NAb-seq: an accurate, rapid, and cost-effective method for antibody long-read sequencing in hybridoma cell lines and single B cells. *mAbs* **14**, 2106621 (2022).
140. Rodriguez, O. L., Silver, C. A., Shields, K., Smith, M. L. & Watson, C. T. Targeted long-read sequencing facilitates phased diploid assembly and genotyping of the human T cell receptor α, δ, and β loci. *Cell Genom.* **2**, 100228 (2022).
141. Brochu, H. N. et al. Systematic profiling of full-length Ig and TCR repertoire diversity in rhesus macaque through long read transcriptome sequencing. *J. Immunol.* **204**, 3434–3444 (2020).
142. Singh, M. et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat. Commun.* **10**, 3120 (2019).
143. Amarasinghe, S. L. et al. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol.* **21**, 30 (2020).
144. Schanz, M. et al. High-throughput sequencing of human immunoglobulin variable regions with subtype identification. *PLoS ONE* **9**, e111726 (2014).
145. Quý, K. L. et al. Benchmarking and integrating human B-cell receptor genomic and antibody proteomic profiling.Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.01.565093 (2023).
146. Chaara, W. et al. RepSeq data representativeness and robustness assessment by Shannon entropy. *Front. Immunol.* **9**, 1038 (2018).
147. Shugay, M. et al. Huge overlap of individual TCR β repertoires. *Front. Immunol.* **4**, 466 (2013).
148. Soto, C. et al. High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398–402 (2019).
149. Kircher, M., Sawyer, S. & Meyer, M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic Acids Res.* **40**, e3 (2012).
150. MacConaill, L. E. et al. Unique, dual-indexed sequencing adapters with UMIs effectively eliminate index cross-talk and significantly improve sensitivity of massively parallel sequencing. *BMC Genomics* **19**, 30 (2018).

151. Costello, M. et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* **19**, 332 (2018).
152. Goods, B. A. et al. Blood handling and leukocyte isolation methods impact the global transcriptome of immune cells. *BMC Immunol.* **19**, 30 (2018).
153. Knecht, H. et al. Quality control and quantification in IG/TR next-generation sequencing marker identification: protocols and bioinformatic functionalities by EuroClonality-NGS. *Leukemia* **33**, 2254–2265 (2019).
154. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* **38**, 1767–1771 (2010).
155. Greiff, V., Miho, E., Menzel, U. & Reddy, S. T. Bioinformatic and statistical analysis of adaptive immune repertoires. *Trends Immunol.* **36**, 738–749 (2015).
156. Pennell, M., Rodriguez, O. L., Watson, C. T. & Greiff, V. The evolutionary and functional significance of germline immunoglobulin gene variation. *Trends Immunol.* **44**, 7–21 (2023).
157. deCamp, A. C. et al. Human immunoglobulin gene allelic variation impacts germline-targeting vaccine priming. Preprint at *medRxiv* https://doi.org/10.1101/2023.03.10.23287126 (2023).
158. Mikocziova, I., Greiff, V. & Sollid, L. M. Immunoglobulin germline gene variation and its impact on human disease. *Genes. Immun.* **22**, 205–217 (2021).
159. Omer, A. et al. T cell receptor β germline variability is revealed by inference from repertoire data. *Genome Med.* **14**, 2 (2022).
160. Ohlin, M. et al. Inferred allelic variants of immunoglobulin receptor genes: a system for their evaluation, documentation, and naming. *Front. Immunol.* **10**, 435 (2019).
161. Watson, C. T., Glanville, J. & Marasco, W. A. The individual and population genetics of antibody immunity. *Trends Immunol.* **38**, 459–470 (2017).
162. Mikelov, A. et al. Ultrasensitive allele inference from immune repertoire sequencing data with MiXCR. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.10.561703 (2023).
163. Corcoran, M. M. et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat. Commun.* **7**, 13642 (2016).
164. Gadala-Maria, D., Yaari, G., Uduman, M. & Kleinstein, S. H. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proc. Natl Acad. Sci. USA* **112**, E862–E870 (2015).
165. Ralph, D. K. & Matsen, F. A. IV Using B cell receptor lineage structures to predict affinity. *PLoS Comput. Biol.* **16**, e1008391 (2020).
166. Vázquez Bernat, N. et al. Rhesus and cynomolgus macaque immunoglobulin heavy-chain genotyping yields comprehensive databases of germline VDJ alleles. *Immunity* **54**, 355–366.e4 (2021).
167. Rodriguez, O. L. et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat. Commun.* **14**, 4419 (2023).
168. Gidoni, M. et al. Mosaic deletion patterns of the human antibody heavy chain gene locus shown by Bayesian haplotyping. *Nat. Commun.* **10**, 628 (2019).
169. Mikocziova, I. et al. Germline polymorphisms and alternative splicing of human immunoglobulin light chain genes. *iScience* **24**, 103192 (2021).
170. Rubelt, F. et al. Adaptive immune receptor repertoire community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.* **18**, 1274–1278 (2017).
171. Vander Heiden, J. A. et al. AIRR community standardized representations for annotated immune repertoires. *Front. Immunol.* **9**, 2206 (2018).
172. Song, L. et al. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods* **18**, 627–630 (2021).
173. Rubio, T. et al. A Nextflow pipeline for T-cell receptor repertoire reconstruction and analysis from RNA sequencing data. *ImmunoInformatics* **6**, 100012 (2022).
174. Bolotin, D. A. et al. Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.* **35**, 908–911 (2017).
175. Glanville, J. et al. Naive antibody gene-segment frequencies are heritable and unaltered by chronic lymphocyte ablation. *Proc. Natl Acad. Sci. USA* **108**, 20066–20071 (2011).
176. Weber, C. R. et al. Reference-based comparison of adaptive immune receptor repertoires. *Cell Rep. Methods* **2**, 100269 (2022).
177. Ritvo, P.-G. et al. High-resolution repertoire analysis reveals a major bystander activation of T_{FH} and T_{FR} cells. *Proc. Natl Acad. Sci. USA* **115**, 9604–9609 (2018).
178. Mhanna, V. et al. Impaired activated/memory regulatory T cell clonal expansion instigates diabetes in NOD mice. *Diabetes* **70**, 976–985 (2021).
179. Olson, B. J. et al. sumrep: a summary statistic framework for immune receptor repertoire comparison and model validation. *Front. Immunol.* **10**, 2533 (2019).
180. Weber, C. R. et al. immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics* **36**, 3594–3596 (2020).
181. Greiff, V. et al. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Med.* **7**, 49 (2015).
182. Rényi, A. in *Proc. Fourth Berkeley Symp. Mathematical Statistics and Probability Vol. 1: Contributions to the Theory of Statistics* Vol. 4.1 (ed. Neyman, J.) 547–562 (Univ. of California Press, 1961).
183. Pielou, E. C. The measurement of diversity in different types of biological collections. *J. Theor. Biol.* **13**, 131–144 (1966).
184. Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**, 427–432 (1973).
185. Dorfman, R. A formula for the Gini coefficient. *Rev. Econ. Stat.* **61**, 146–149 (1979).
186. Magurran, A. E. in *Ecological Diversity and Its Measurement* (ed. Magurran, A. E.) 61–80 (Springer Netherlands, 1988).
187. Jost, L. Entropy and diversity. *Oikos* **113**, 363–375 (2006).
188. Nolan, K. A. & Callahan, J. E. Beachcomber biology: The Shannon-Weiner Species Diversity Index. in *Tested Studies for Laboratory Teaching. Proceedings of the 27th Workshop/Conference of the Association for Biology Laboratory Education* Vol. 27 (ed. O'Donnell, M. A.) 334–338 (Association for Biology Laboratory Education, 2006).
189. Somerfield, P. J., Clarke, K. R. & Warwick, R. M. in *Encyclopedia of Ecology* (eds Jorgensen, S. V. & Fath, B.) 3252–3255 (Elsevier, 2008).
190. Kaplinsky, J. & Arnaout, R. Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples. *Nat. Commun.* **7**, 11881 (2016).
191. Miho, E. et al. Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Front. Immunol.* **9**, 224 (2018).
192. Marquez, S. et al. Adaptive immune receptor repertoire (AIRR) community guide to repertoire analysis. *Methods Mol. Biol.* **2453**, 297–316 (2022).
193. Alon, U., Mokryn, O. & Hershberg, U. Using domain based latent personal analysis of B cell clone diversity patterns to identify novel relationships between the B cell clone populations in different tissues. *Front. Immunol.* **12**, 642673 (2021).
194. Strauli, N. B. & Hernandez, R. D. Statistical inference of a convergent antibody repertoire response to influenza vaccine. *Genome Med.* **8**, 60 (2016).
195. Vujović, M., Marcatili, P., Chain, B., Kaplinsky, J. & Andresen, T. L. Signatures of T cell immunity revealed using sequence similarity with TCRDivER algorithm. *Commun. Biol.* **6**, 357 (2023).
196. Miho, E., Roškar, R., Greiff, V. & Reddy, S. T. Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.* **10**, 1321 (2019).
197. Arora, R. & Arnaout, R. Repertoire-scale measures of antigen binding. *Proc. Natl Acad. Sci. USA* **119**, e2203505119 (2022).
198. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *Interj. Complex Syst.* **1695**, 1–9 (2006).
199. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring Network Structure, Dynamics, and Function using NetworkX. in *Proc. 7th Python in Science Conf. (SciPy2008)* (eds Varoquaux, G., Vaught, T. & Millman, J.) 11–15 (Proceedings of the Python in Science Conference, 2008).
200. Bastian, M., Heymann, S. & Jacomy, M. Gephi: an open source software for exploring and manipulating. *Networks. ICWSM* **3**, 361–362 (2009).
201. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
202. Madi, A. et al. T cell receptor repertoires of mice and humans are clustered in similarity networks around conserved public CDR3 sequences. *eLife Sci.* **6**, e22057 (2017).
203. Amoriello, R. et al. TCR repertoire diversity in multiple sclerosis: high-dimensional bioinformatics analysis of sequences from brain, cerebrospinal fluid and peripheral blood. *EBioMedicine* **68**, 103429 (2021).
204. Albert, R., Jeong, H. & Barabasi, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
205. Barabási, A.-L. & Márton, P. *Network Science* (Cambridge Univ. Press, 2016).
206. Akbar, R. et al. A compact vocabulary of paratope–epitope interactions enables predictability of antibody–antigen binding. *Cell Rep.* **34**, 108856 (2021).
207. Ostmeyer, J., Christley, S., Toby, I. T. & Cowell, L. G. Biophysicochemical motifs in T-cell receptor sequences distinguish repertoires from tumor-infiltrating lymphocyte and adjacent healthy tissue. *Cancer Res.* **79**, 1671–1680 (2019).
208. Valkiers, S., Van Houcke, M., Laukens, K. & Meysman, P. ClusTCR: a Python interface for rapid clustering of large sets of CDR3 sequences with unknown antigen specificity. *Bioinformatics* **37**, 4865–4867 (2021).
209. Mayer-Blackwell, K. et al. TCR meta-clonotypes for biomarker discovery with *tcrdist3* enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *eLife* **10**, e68605 (2021).
210. Zhang, H., Zhan, X. & Li, B. GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. *Nat. Commun.* **12**, 4699 (2021).
211. Zhang, Z. et al. Interpreting the B-cell receptor repertoire with single-cell gene expression using Benisse. *Nat. Mach. Intell.* **4**, 596–604 (2022).
212. Zhang, Z., Xiong, D., Wang, X., Liu, H. & Wang, T. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat. Methods* **18**, 92–99 (2021).
213. Schattgen, S. A. et al. Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nat. Biotechnol.* **40**, 54–63 (2022).
214. Hoehn, K. B., Fowler, A., Lunter, G. & Pybus, O. G. The diversity and molecular evolution of B-cell receptors during infection. *Mol. Biol. Evol.* **33**, 1147–1157 (2016).
215. Yermanos, A. D., Dounas, A. K., Stadler, T., Oxenius, A. & Reddy, S. T. Tracing antibody repertoire evolution by systems phylogeny. *Front. Immunol.* **9**, 2149 (2018).
216. Abdollahi, N. et al. A multi-objective based clustering for inferring BCR clonal lineages from high-throughput B cell repertoire data. *PLoS Comput. Biol.* **18**, e1010411 (2022).
217. Nouri, N. & Kleinstein, S. H. A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. *Bioinformatics* **34**, i341–i349 (2018).
218. Yermanos, A. et al. Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim). *Bioinformatics* **33**, 3938–3946 (2017).
219. Davidsen, K. & Matsen, F. A. 4th Benchmarking tree and ancestral sequence inference for B cell receptor sequences. *Front. Immunol.* **9**, 2451 (2018).
220. Zhang, C., Bzikadze, A. V., Safonova, Y. & Mirarab, S. A scalable model for simulating multi-round antibody evolution and benchmarking of clonal tree reconstruction methods. *Front. Immunol.* **13**, 1014439 (2022).

221. Yaari, G. et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front. Immunol.* **4**, 358 (2013).
222. Hoehn, K. B. et al. Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proc. Natl Acad. Sci. USA* **116**, 22664–22672 (2019).
223. Hoehn, K. B., Pybus, O. G. & Kleinstein, S. H. Phylogenetic analysis of migration, differentiation, and class switching in B cells. *PLoS Comput. Biol.* **18**, e1009885 (2022).
224. Jensen, C. G., Sumner, J. A., Kleinstein, S. H. & Hoehn, K. B. Inferring B cell phylogenies from paired heavy and light chain BCR sequences with Dowser. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.29.560187 (2023).
225. Ralph, D. K. & Matsen, F. A. IV Inference of B cell clonal families using heavy/light chain pairing information. *PLoS Comput. Biol.* **18**, e1010723 (2022).
226. Zhou, J. Q. & Kleinstein, S. H. Cutting edge: Ig H chains are sufficient to determine most B cell clonal relationships. *J. Immunol.* **203**, 1687–1692 (2019).
227. DeWitt, W. S. III, Mesin, L., Victora, G. D., Minin, V. N. & Matsen, F. A. IV Using genotype abundance to improve phylogenetic inference. *Mol. Biol. Evol.* **35**, 1253–1265 (2018).
228. Abdollahi, N., Jeusset, L., de Septenville, A., Davi, F. & Bernardes, J. S. Reconstructing B cell lineage trees with minimum spanning tree and genotype abundances. *BMC Bioinformatics* **24**, 70 (2023).
229. Zaragoza-Infante, L. et al. IgIDivA: immunoglobulin intraclonal diversification analysis. *Brief. Bioinform.* **23**, bbac349 (2022).
230. Foglierini, M., Pappas, L., Lanzavecchia, A., Corti, D. & Perez, L. AncesTree: an interactive immunoglobulin lineage tree visualizer. *PLoS Comput. Biol.* **16**, e1007731 (2020).
231. Jeusset, L. et al. ViCloD, an interactive web tool for visualizing B cell repertoires and analyzing intraclonal diversities: application to human B-cell tumors. *NAR Genom. Bioinform.* https://doi.org/10.1093/nargab/lqad064 (2023).
232. Lees, W. D. Tools for adaptive immune receptor repertoire sequencing. *Curr. Opin. Syst. Biol.* **24**, 86–92 (2020).
233. Greiff, V. et al. Learning the high-dimensional immunogenomic features that predict public and private antibody repertoires. *J. Immunol.* **199**, 2985–2997 (2017).
234. Venturi, V. et al. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proc. Natl Acad. Sci. USA* **103**, 18691–18696 (2006).
235. Quigley, M. F. et al. Convergent recombination shapes the clonotypic landscape of the naive T-cell repertoire. *Proc. Natl Acad. Sci. USA* **107**, 19414–19419 (2010).
236. Elhanati, Y., Murugan, A., Callan, C. G. Jr, Mora, T. & Walczak, A. M. Quantifying selection in immune receptor repertoires. *Proc. Natl Acad. Sci. USA* **111**, 9875–9880 (2014).
237. Pogorelyy, M. V. et al. Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. *Proc. Natl Acad. Sci. USA* **115**, 12704–12709 (2018).
238. Galson, J. D. et al. Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. *Front. Immunol.* **11**, 605170 (2020).
239. Gomez-Tourino, I., Kamra, Y., Baptista, R., Lorenc, A. & Peakman, M. T cell receptor β-chains display abnormal shortening and repertoire sharing in type 1 diabetes. *Nat. Commun.* **8**, 1792 (2017).
240. Jaccard, P. The distribution of the flora in the alpine zone.1. *N. Phytol.* **11**, 37–50 (1912).
241. Amoriello, R. et al. The TCR repertoire reconstitution in multiple sclerosis: comparing one-shot and continuous immunosuppressive therapies. *Front. Immunol.* **11**, 559 (2020).
242. Rognes, T., Scheffer, L., Greiff, V. & Sandve, G. K. CompAIRR: ultra-fast comparison of adaptive immune receptor repertoires by exact and approximate sequence matching. *Bioinformatics* **38**, 4230–4232 (2022).
243. Morisita, M. Measuring of the dispersion and analysis of distribution patterns, Memoires of the Faculty of Science, Kyushu University. Series E. Biology. *Sci. Rep.* **2**, 215–235 (1995).
244. Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).
245. Bolen, C. R., Rubelt, F., Vander Heiden, J. A. & Davis, M. M. The repertoire dissimilarity index as a method to compare lymphocyte receptor repertoires. *BMC Bioinformatics* **18**, 155 (2017).
246. Raybould, M. I. J. et al. Public baseline and shared response structures support the theory of antibody repertoire functional commonality. *PLoS Comput. Biol.* **17**, e1008781 (2021).
247. Arora, R., Burke, H. M. & Arnaout, R. Immunological diversity with similarity. Preprint at *bioRxiv* https://doi.org/10.1101/483131 (2018).
248. Ruffolo, J. A., Sulam, J. & Gray, J. J. Antibody structure prediction using interpretable deep learning. *Patterns* **3**, 100406 (2022).
249. Abanades, B. et al. ImmuneBuilder: deep-learning models for predicting the structures of immune proteins. *Commun. Biol.* **6**, 575 (2023).
250. Marks, C. & Deane, C. M. How repertoire data are changing antibody science. *J. Biol. Chem.* **295**, 9823–9837 (2020).
251. Krawczyk, K. et al. Structurally mapping antibody repertoires. *Front. Immunol.* **9**, 1698 (2018).
252. Kovaltsuk, A. et al. How B-cell receptor repertoire sequencing can be enriched with structural antibody data. *Front. Immunol.* **8**, 1753 (2017).
253. Shcherbinin, D. S., Karnaukhov, V. K., Zvyagin, I. V., Chudakov, D. M. & Shugay, M. Large-scale template-based structural modeling of T-cell receptors with known antigen specificity reveals complementarity features. *Front. Immunol.* **14**, 1224969 (2023).
254. Sela-Culang, I., Kunik, V. & Ofran, Y. The structural basis of antibody–antigen recognition. *Front. Immunol.* **4**, 302 (2013).
255. Richardson, E. et al. A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. *mAbs* **13**, 1869406 (2021).
256. Imkeller, K. & Wardemann, H. Assessing human B cell repertoire diversity and convergence. *Immunol. Rev.* **284**, 51–66 (2018).
257. Wong, W. K. et al. Ab-Ligity: identifying sequence-dissimilar antibodies that bind to the same epitope. *MAbs* **13**, 1873478 (2021).
258. Dash, P. et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
259. Ruffolo, J. A., Guerra, C., Mahajan, S. P., Sulam, J. & Gray, J. J. Geometric potentials from deep learning improve prediction of CDR H3 loop structures. *Bioinformatics* **36**, i268–i275 (2020).
260. Fernández-Quintero, M. L. et al. Challenges in antibody structure prediction. *MAbs* **15**, 2175319 (2023).
261. Greiff, V., Yaari, G. & Cowell, L. G. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. *Curr. Opin. Syst. Biol.* **24**, 109–119 (2020).
262. Chakravarthi Kanduri et al. simAIRR: simulation of adaptive immune repertoires with realistic receptor sequence sharing for benchmarking of immune state prediction methods. *GigaScience* **12**, giag074 (2022).
263. Chernigovskaya, M. et al. Simulation of adaptive immune receptors and repertoires with complex immune information to guide the development and benchmarking of AIRR machine learning. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.20.562936 (2023).
264. Mariotti-Ferrandiz, E. et al. A TCRβ repertoire signature can predict experimental cerebral malaria. *PLoS ONE* **11**, e0147871 (2016).
265. Six, A. et al. The past, present, and future of immune repertoire biology — the rise of next-generation repertoire analysis. *Front. Immunol.* **4**, 413 (2013).
266. Pertseva, M., Gao, B., Neumeier, D., Yermanos, A. & Reddy, S. T. Applications of machine and deep learning in adaptive immunity. *Annu. Rev. Chem. Biomol. Eng.* **12**, 39–62 (2021).
267. Hummer, A. M., Abanades, B. & Deane, C. M. Advances in computational structure-based antibody design. *Curr. Opin. Struct. Biol.* **74**, 102379 (2022).
268. Hudson, D., Fernandes, R. A., Basham, M., Ogg, G. & Koohy, H. Can we predict T cell specificity with digital biology and machine learning? *Nat. Rev. Immunol.* **23**, 511–521 (2023).
269. Gielis, S. et al. Detection of enriched T cell epitope specificity in full T cell receptor sequence repertoires. *Front. Immunol.* **10**, 2820 (2019).
270. Luo, J. et al. Quantitative annotations of T-cell repertoire specificity. *Brief. Bioinform.* **24**, bbad175 (2023).
271. Widrich, M. et al. Modern hopfield networks and attention for immune repertoire classification. *Adv. Neural Inf. Process. Syst.* **33**, 18832–18845 (2020).
272. Sidhom, J.-W., Larman, H. B., Pardoll, D. M. & Baras, A. S. DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nat. Commun.* **12**, 1605 (2021).
273. Robert, P. A. et al. Unconstrained generation of synthetic antibody–antigen structures to guide machine learning methodology for antibody specificity prediction. *Nat. Comput. Sci.* **2**, 845–865 (2022).
274. Thomas, N. et al. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* **30**, 3181–3188 (2014).
275. Leem, J., Mitchell, L. S., Farmery, J. H. R., Barton, J. & Galson, J. D. Deciphering the language of antibodies using self-supervised learning. *Patterns* **3**, 100513 (2022).
276. Vu, M. H. et al. Linguistically inspired roadmap for building biologically reliable protein language models. *Nat. Mach. Intell.* **5**, 485–496 (2023).
277. Vu, M. H. et al. ImmunoLingo: linguistics-based formalization of the antibody language. Preprint at https://doi.org/10.48550/arXiv.2209.12635 (2022).
278. Ruffolo, J. A., Chu, L.-S., Mahajan, S. P. & Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nat. Commun.* **14**, 2389 (2023).
279. Wang, M., Patsenker, J., Li, H., Kluger, Y. & Kleinstein, S. Language model-based B cell receptor sequence embeddings can effectively encode receptor specificity. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.21.545145 (2023).
280. Hie, B. L. et al. Efficient evolution of human antibodies from general protein language models. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-023-01763-2 (2023).
281. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
282. Zhang, P, Bang, S., Cai, M & Lee, H. Context-aware amino acid embedding advances analysis of TCR–epitope interactions. *eLife* **12**, RP88837 (2023).
283. Peng, X. et al. Characterizing the interaction conformation between T-cell receptors and epitopes with deep learning. *Nat. Mach. Intell.* **5**, 395–407 (2023).
284. Drost, F. et al. Integrating T-cell receptor and transcriptome for large-scale single-cell immune profiling analysis. Preprint at *bioRxiv* https://doi.org/10.1101/2021.06.24.449733 (2022).
285. Kuhn, M. & Johnson, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models* (CRC, 2019).
286. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **45**, 427–437 (2009).
287. Pavlović, M. et al. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. *Nat. Mach. Intell.* **3**, 936–944 (2021).
288. Kanduri, C. et al. Profiling the baseline performance and limits of machine learning models for adaptive immune receptor repertoire classification. *Gigascience* **11**, giac046 (2022).
289. Katayama, Y. & Kobayashi, T. J. Comparative study of repertoire classification methods reveals data efficiency of *k*-mer feature extraction. *Front. Immunol.* **13**, 797640 (2022).

290. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* **9**, 561 (2018).
291. Murugan, A., Mora, T., Walczak, A. M. & Callan, C. G. Jr Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl Acad. Sci. USA* **109**, 16161–16166 (2012).
292. Slabodkin, A. et al. Individualized VDJ recombination predisposes the available Ig sequence space. *Genome Res.* **31**, 2209–2224 (2021).
293. Russell, M. L. et al. Combining genotypes and T cell receptor distributions to infer genetic loci determining V(D)J recombination probabilities. *eLife* **11**, e73475 (2022).
294. Neumeier, D. et al. Phenotypic determinism and stochasticity in antibody repertoires of clonally expanded plasma cells. *Proc. Natl Acad. Sci. USA* **119**, e2113766119 (2022).
295. Salou, M., Nicol, B., Garcia, A. & Laplaud, D.-A. Involvement of $CD8^+$ T cells in multiple sclerosis. *Front. Immunol.* **6**, 604 (2015).
296. Codina-Busqueta, E. et al. TCR bias of in vivo expanded T cells in pancreatic islets and spleen at the onset in human type 1 diabetes. *J. Immunol.* **186**, 3787–3797 (2011).
297. Seay, H. R. et al. Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight* **1**, e88242 (2016).
298. Kovaltsuk, A. et al. Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. *PLoS Comput. Biol.* **16**, e1007636 (2020).
299. Hoehn, K. B. et al. Dynamics of immunoglobulin sequence diversity in HIV-1 infected individuals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 2014 (2015).
300. Chang, C.-M. et al. Profiling of T cell repertoire in SARS-CoV-2-infected COVID-19 patients between mild disease and pneumonia. *J. Clin. Immunol.* **41**, 1131–1145 (2021).
301. Priel, A., Gordin, M., Philip, H., Zilberberg, A. & Efroni, S. Network representation of T-cell repertoire — a novel tool to analyze immune response to cancer formation. *Front. Immunol.* **9**, 2913 (2018).
302. Pogorelyy, M. V. et al. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol.* **17**, e3000314 (2019).
303. Minervina, A. A. et al. Longitudinal high-throughput TCR repertoire profiling reveals the dynamics of T-cell memory formation after mild COVID-19 infection. *eLife* **10**, e63502 (2021).
304. Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
305. Pogorelyy, M. V. et al. Resolving SARS-CoV-2 $CD4^+$ T cell specificity via reverse epitope discovery. *Cell Rep. Med.* **3**, 100697 (2022).
306. Simnica, D. et al. Landscape of T-cell repertoires with public COVID-19-associated T-cell receptors in pre-pandemic risk cohorts. *Clin. Transl. Immunol.* **10**, e1340 (2021).
307. Servaas, N. H. et al. Longitudinal analysis of T-cell receptor repertoires reveals persistence of antigen-driven $CD4^+$ and $CD8^+$ T-cell clusters in systemic sclerosis. *J. Autoimmun.* **117**, 102574 (2021).
308. Komech, E. A. et al. $CD8^+$ T cells with characteristic T cell receptor β motif are detected in blood and expanded in synovial fluid of ankylosing spondylitis patients. *Rheumatology* **57**, 1097–1104 (2018).
309. Chiou, S.-H. et al. Global analysis of shared T cell specificities in human non-small cell lung cancer enables HLA inference and antigen discovery. *Immunity* **54**, 586–602.e8 (2021).
310. Joshi, K. et al. Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer. *Nat. Med.* **25**, 1549–1559 (2019).
311. Hoehn, K. B. et al. Human B cell lineages associated with germinal centers following influenza vaccination are measurably evolving. *eLife* **10**, e70873 (2021).
312. de Bourcy, C. F. A. et al. Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *Proc. Natl Acad. Sci. USA* **114**, 1105–1110 (2017).
313. Stern, J. N. H. et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl Med.* **6**, 248ra107 (2014).
314. Park, J.-E. et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay33224 (2020).
315. Trück, J. et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J.Immunol.* **194**, 252–261 (2015).
316. Jackson, K. J. L. et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe* **16**, 105–114 (2014).
317. Dong, L., Li, P., Oenema, T., McClurkan, C. L. & Koelle, D. M. Public TCR use by herpes simplex virus-2-specific human CD8 CTLs. *J. Immunol.* **184**, 3063–3071 (2010).
318. Quigley, M. et al. Transcriptional analysis of HIV-specific $CD8^+$ T cells shows that PD-1 inhibits T cell function by upregulating BATF. *Nat. Med.* **16**, 1147–1151 (2010).
319. Eugster, A. et al. High diversity in the TCR repertoire of GAD65 autoantigen-specific human $CD4^+$ T cells. *J. Immunol.* **194**, 2531–2538 (2015).
320. Musters, A. et al. In rheumatoid arthritis, synovitis at different inflammatory sites is dominated by shared but patient-specific T cell clones. *J. Immunol.* **201**, 417–422 (2018).
321. Krasik, S. V. et al. Systematic evaluation of intratumoral and peripheral BCR repertoires in three cancers. Preprint at *bioRxiv* https://doi.org/10.1101/2023.04.16.537028 (2023).
322. Aran, A. et al. Analysis of tumor infiltrating $CD4^+$ and $CD8^+$ CDR3 sequences reveals shared features putatively associated to the anti-tumor immune response. *Front. Immunol.* **14**, 1227766 (2023).
323. Dunn-Walters, D. K. The ageing human B cell repertoire: a failure of selection? *Clin. Exp. Immunol.* **183**, 50–56 (2016).
324. Britanova, O. V. et al. Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling. *J. Immunol.* **192**, 2689–2698 (2014).
325. Minervina, A. A. et al. Primary and secondary anti-viral response captured by the dynamics and phenotype of individual T cell clones. *eLife* **9**, e53704 (2020).
326. Mikelov, A. et al. Memory persistence and differentiation into antibody-secreting cells accompanied by positive selection in longitudinal BCR repertoires. *eLife* **11**, e79254 (2022).
327. Galson, J. D. et al. In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. *Front. Immunol.* **6**, 531 (2015).
328. De Neuter, N. et al. Memory $CD4^+$ T cell receptor repertoire data mining as a tool for identifying cytomegalovirus serostatus. *Genes. Immun.* **20**, 255–260 (2019).
329. Liu, X. et al. T cell receptor β repertoires as novel diagnostic markers for systemic lupus erythematosus and rheumatoid arthritis. *Ann. Rheum. Dis.* **78**, 1070–1078 (2019).
330. Snyder, T. M. et al. Magnitude and dynamics of the T-cell response to SARS-CoV-2 infection at both individual and population levels. Preprint at *medRxiv* https://doi.org/10.1101/2020.07.31.20165647 (2020).
331. Zaslavsky, M. E. et al. Disease diagnostics using machine learning of immune receptors. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.26.489314 (2023).
332. Akbar, R. et al. Progress and challenges for the machine learning-based design of fit-for-purpose monoclonal antibodies. *Abs* **14**, 2008790 (2022).
333. Wilman, W. et al. Machine-designed biotherapeutics: opportunities, feasibility and advantages of deep learning in computational antibody discovery. *Brief. Bioinform.* **23**, bbac267(2022).
334. Valkiers, S. et al. Recent advances in T-cell receptor repertoire analysis: bridging the gap with multimodal single-cell RNA sequencing. *ImmunoInformatics* **5**, 100009 (2022).
335. Katayama, Y., Yokota, R., Akiyama, T. & Kobayashi, T. J. Machine learning approaches to TCR repertoire analysis. *Front. Immunol.* **13**, 858057 (2022).
336. Weber, A., Born, J. & Rodriguez Martínez, M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* **37**, i237–i244 (2021).
337. Montemurro, A., Jessen, L. E. & Nielsen, M. NetTCR-2.1: lessons and guidance on how to develop models for TCR specificity predictions. *Front. Immunol.* **13**, 1055151 (2022).
338. Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S. & Louzoun, Y. Prediction of specific TCR-peptide binding from large dictionaries of TCR–peptide pairs. *Front. Immunol.* **11**, 1803 (2020).
339. Dens, C., Bittremieux, W., Affaticati, F., Laukens, K. & Meysman, P. Interpretable deep learning to uncover the molecular binding patterns determining TCR–epitope interactions. *ImmunoInformatics* **11**, 100027 (2023).
340. Bradley, P. Structure-based prediction of T cell receptor:peptide–MHC interactions. *eLife* **12**, e82813 (2023).
341. Perez, M. A. S. et al. TCRpcDist: estimating TCR physico-chemical similarity to analyze repertoires and predict specificities. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.15.545077 (2023).
342. Sethna, Z., Elhanati, Y., Callan, C. G., Walczak, A. M. & Mora, T. OLGA: fast computation of generation probabilities of B- and T-cell receptor amino acid sequences and motifs. *Bioinformatics* **35**, 2974–2981 (2019).
343. Pavlović, M. et al. Improving generalization of machine learning-identified biomarkers with causal modeling: an investigation into immune receptor diagnostics. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.09291 (2022).
344. Swindells, M. B. et al. abYsis: integrated antibody sequence and structure-management, analysis, and prediction. *J. Mol. Biol.* **429**, 356–364 (2017).
345. Ferdous, S. & Martin, A. C. R. AbDb: antibody structure database — a database of PDB-derived antibody structures. *Database* **2018**, bay040 (2018).
346. Dunbar, J. et al. SAbDab: the structural antibody database. *Nucleic Acids Res.* **42**, D1140–D1146 (2014).
347. Mahajan, S. et al. Epitope specific antibodies and T cell receptors in the immune epitope database. *Front. Immunol.* **9**, 2688 (2018).
348. Shugay, M. et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res* **46**, D419–D427 (2018).
349. Dorigatti, E.et al. Predicting T cell receptor functionality against mutant epitopes. Preprine at *bioRxiv* https://doi.org/10.1101/2023.05.10.540189 (2023).
350. Taft, J. M. et al. Deep mutational learning predicts ACE2 binding and antibody escape to combinatorial mutations in the SARS-CoV-2 receptor-binding domain. *Cell* **185**, 4008–4022.e14 (2022).
351. Straub, A. et al. Recruitment of epitope-specific T cell clones with a low-avidity threshold supports efficacy against mutational escape upon re-infection. *Immunity* **56**, 1269–1284.e6 (2023).
352. Mayer, A. & Callan, C. G. Jr Measures of epitope binding degeneracy from T cell receptor repertoires. *Proc. Natl Acad. Sci. USA* **120**, e2213264120 (2023).
353. Chronister, W. D. et al. TCRMatch: predicting T-cell receptor specificity based on sequence similarity to previously characterized receptors. *Front. Immunol.* **12**, 640725 (2021).
354. De Neuter, N. et al. On the feasibility of mining $CD8^+$ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics* **70**, 159–168 (2018).
355. Elias, G. et al. Preexisting memory CD4 T cells in naïve individuals confer robust immunity upon hepatitis B vaccination. *eLife* **11**, e68388 (2022).
356. Vujkovic, A. et al. Diagnosing viral infections through T cell receptor sequencing of activated $CD8^+$ T cells. *J. Infect. Dis.* **3**, jiad430 (2023).
357. Davidsen, K. et al. Deep generative models for T cell receptor protein sequences. *eLife* **8**, e46935 (2019).
358. Akbar, R. et al. In silico proof of principle of machine learning-based antibody design at unconstrained scale. *mAbs* **14**, 2031482 (2022).

359. Saka, K. et al. Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Sci. Rep.* **11**, 5852 (2021).
360. Amimeur, T. et al. Designing feature-controlled humanoid antibody discovery libraries using generative adversarial networks. Preprint at *bioRxiv* https://doi.org/10.1101/2020.04.12.024844 (2020).
361. Breden, F. et al. Reproducibility and reuse of adaptive immune receptor repertoire data. *Front. Immunol.* **8**, 1418 (2017).
362. Corrie, B. D. et al. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol. Rev.* **284**, 24–41 (2018).
363. Christley, S. et al. VDJServer: a cloud-based analysis portal and data commons for immune repertoire sequences and rearrangements. *Front. Immunol.* **9**, 976 (2018).
364. Brüggemann, M. et al. 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 (2019).
365. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
366. Huang, Y.-N. et al. Data availability of open T-cell receptor repertoire data, a systematic assessment. *Front. Syst. Biol.* **2**, 918792 (2022).
367. Bukhari, S. A. C. et al. The CAIRR pipeline for submitting standards-compliant B and T cell receptor repertoire sequencing studies to the National Center for Biotechnology Information repositories. *Front. Immunol.* **9**, 1877 (2018).
368. Christley, S. et al. The ADC API: a web API for the programmatic query of the AIRR Data Commons. *Front. Big Data* **3**, 22 (2020).
369. Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E. & Friedman, N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**, 2924–2929 (2017).
370. Raybould, M. I. J., Kovaltsuk, A., Marks, C. & Deane, C. M. CoV-AbDab: the coronavirus antibody database. *Bioinformatics* **37**, 734–735 (2021).
371. Li, R. et al. A novel statistical method for decontaminating T-cell receptor sequencing data. *Brief. Bioinform.* **24**, bbad230 (2023).
372. Smirnova, A. O. et al. The use of non-functional clonotypes as a natural calibrator for quantitative bias correction in adaptive immune receptor repertoire profiling. *Elife* **12**, e69157 (2023).
373. Greiff, V. et al. Quantitative assessment of the robustness of next-generation sequencing of antibody variable gene repertoires from immunized mice. *BMC Immunol.* **15**, 40 (2014).
374. Koraichi, M. B., Touzel, M. P., Mazzolini, A., Mora, T. & Walczak, A. M. NoisET: noise learning and expansion detection of T-cell receptors. *J. Phys. Chem. A* **126**, 7407–7414 (2022).
375. Chen, L. et al. GMPR: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* **6**, e4600 (2018).
376. Jaffe, D. B. et al. Functional antibodies exhibit light chain coherence. *Nature* **611**, 352–357 (2022).
377. Holec, P. V., Berleant, J., Bathe, M. & Birnbaum, M. E. A Bayesian framework for high-throughput T cell receptor pairing. *Bioinformatics* **35**, 1318–1325 (2019).
378. Howie, B. et al. High-throughput pairing of T cell receptor α and β sequences. *Sci. Transl. Med.* **7**, 301ra131 (2015).
379. DeKosky, B. J. et al. In-depth determination and analysis of the human paired heavy- and light-chain antibody repertoire. *Nat. Med.* **21**, 86–91 (2015).
380. Tanno, H. et al. Determinants governing T cell receptor α/β-chain pairing in repertoire formation of identical twins. *Proc. Natl Acad. Sci. USA* **117**, 532–540 (2020).
381. Grigaityte, K. et al. Single-cell sequencing reveals αβ chain pairing shapes the T cell repertoire. Preprint at *bioRxiv* https://doi.org/10.1101/213462 (2017).
382. Shcherbinin, D. S., Belousov, V. A. & Shugay, M. Comprehensive analysis of structural and sequencing data reveals almost unconstrained chain pairing in TCRαβ complex. *PLoS Comput. Biol.* **16**, e1007714 (2020).
383. Moris, P. et al. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Brief. Bioinform.* **22**, bbaa318 (2021).
384. Shemesh, O., Polak, P., Lundin, K. E. A., Sollid, L. M. & Yaari, G. Machine learning analysis of naïve B-cell receptor repertoires stratifies celiac disease patients and controls. *Front. Immunol.* **12**, 627813 (2021).
385. Liu, S., Bradley, P. & Sun, W. Neural network models for sequence-based TCR and HLA association prediction. *PLoS Comput. Biol* **19**, e1011664 (2023).
386. DeWitt, W. S. III et al. Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *eLife* **7**, e38358 (2018).
387. Ishigaki, K. et al. HLA autoimmune risk alleles restrict the hypervariable region of T cell receptors. *Nat. Genet.* **54**, 393–402 (2022).
388. Peng, K. et al. Diversity in immunogenomics: the value and the challenge. *Nat. Methods* **18**, 588–591 (2021).
389. Deng, L. et al. Performance comparison of TCR–pMHC prediction tools reveals a strong data dependency. *Front. Immunol.* **14**, 1128326 (2023).
390. Dens, C., Laukens, K., Bittremieux, W. & Meysman, P. The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nat. Mach. Intell.* **5**, 1063–1065 (2023).
391. Meysman, P. et al. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. *ImmunoInformatics* **9**, 100024 (2023).
392. Papadopoulou, I., Nguyen, A.-P., Weber, A. & Martínez, M. R. DECODE: a computational pipeline to discover T cell receptor binding rules. *Bioinformatics* **38**, i246–i254 (2022).
393. Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P. & Preece, A. Sanity checks for saliency metrics. *AAAI* **34**, 6021–6029 (2020).
394. Sandve, G. K. & Greiff, V. Access to ground truth at unconstrained size makes simulated data as indispensable as experimental data for bioinformatics methods development and benchmarking. *Bioinformatics* **38**, 4994–4996 (2022).
395. Chen, V. et al. Best practices for interpretable machine learning in computational biology. Preprint at *bioRxiv* https://doi.org/10.1101/2022.10.28.513978 (2022).
396. Snapkov, I. et al. Progress and challenges in mass spectrometry-based analysis of antibody repertoires. *Trends Biotechnol.* **40**, 463–481 (2021).
397. Ionov, S. & Lee, J. An immunoproteomic survey of the antibody landscape: insights and opportunities revealed by serological repertoire profiling. *Front. Immunol.* **13**, 832533 (2022).
398. Curtis, N. C. et al. Characterization of SARS-CoV-2 convalescent patients' serological repertoire reveals high prevalence of Iso-RBD antibodies. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.08.556349 (2023).
399. Lee, J. et al. Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat. Med.* **22**, 1456–1464 (2016).
400. de Graaf, S. C., Hoek, M., Tamara, S. & Heck, A. J. R. A perspective toward mass spectrometry-based de novo sequencing of endogenous antibodies. *mAbs* **14**, 2079449 (2022).
401. Yilmaz, M., Fondrie, W. E., Bittremieux, W. & Oh, S. De novo mass spectrometry peptide sequencing with a transformer model. Preprint at *bioRxiv* https://doi.org/10.1101/2022.02.07.479481 (2022).
402. Setliff, I. et al. High-throughput mapping of B cell receptor sequences to antigen specificity. *Cell* **179**, 1636–1646.e15 (2019).
403. Eyer, K. et al. Single-cell deep phenotyping of IgG-secreting cells for high-resolution immune monitoring. *Nat. Biotechnol.* **35**, 977–982 (2017).
404. Ma, K.-Y. et al. High-throughput and high-dimensional single-cell analysis of antigen-specific CD8$^+$ T cells. *Nat. Immunol.* **22**, 1590–1598 (2021).
405. Bentzen, A. K. et al. Large-scale detection of antigen-specific T cells using peptide–MHC-I multimers labeled with DNA barcodes. *Nat. Biotechnol.* **34**, 1037–1045 (2016).
406. Minervina, A. A. et al. SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8$^+$ T cells. *Nat. Immunol.* **23**, 781–790 (2022).
407. Gérard, A. et al. High-throughput single-cell activity-based screening and sequencing of antibodies using droplet microfluidics. *Nat. Biotechnol.* **38**, 715–721 (2020).
408. Malissen, M. et al. Regulation of TCR α and β gene allelic exclusion during T-cell development. *Immunol. Today* **13**, 315–322 (1992).
409. Padovan, E. et al. Expression of two T cell receptor α chains: dual receptor T cells. *Science* **262**, 422–424 (1993).
410. Schuldt, N. J. & Binstadt, B. A. Dual TCR T cells: identity crisis or multitaskers? *J. Immunol.* **202**, 637–644 (2019).
411. Zhu, L. et al. scRNA-Seq revealed the special TCR β & α V(D)J allelic inclusion rearrangement and the high proportion dual (or more) TCR-expressing cells. *Cell Death Dis.* **14**, 487 (2023).
412. Croce, G. et al. Deep learning predictions of TCR–epitope interactions reveal epitope-specific chains in dual α T cells. Preprint at *bioRxiv* https://doi.org/10.1101/2023.09.13.557561 (2023).
413. Robert, P. A., Marschall, A. L. & Meyer-Hermann, M. Induction of broadly neutralizing antibodies in germinal centre simulations. *Curr. Opin. Biotechnol.* **51**, 137–145 (2018).
414. Mashiko, S. et al. Broad responses to chemical adducts shape the natural antibody repertoire in early infancy. *Sci. Adv.* **9**, eade8872 (2023).
415. Harvey, E. P. et al. An in silico method to assess antibody fragment polyreactivity. *Nat. Commun.* **13**, 7554 (2022).
416. Sakhnini, L., Lorenzen, N., Sormanni, P., Vendruscolo, M. & Granata, D. Development of machine learning models for prediction of antibody non-specificity. *Biophys. J.* **122**, 463a (2023).
417. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP) — round XIV. *Proteins* **89**, 1607–1617 (2021).
418. Lensink, M. F., Velankar, S. & Wodak, S. J. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins* **85**, 359–377 (2017).
419. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
420. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
421. Armer, C. et al. The protein engineering tournament: an open science benchmark for protein modeling and design. Preprint at *arXiv* arxiv-2309.09955 (2023).
422. Lingwood, D. et al. Structural and genetic basis for development of broadly neutralizing influenza antibodies. *Nature* **489**, 566–570 (2012).
423. Jardine, J. et al. Rational HIV immunogen design to target specific germline B cell receptors. *Science* **340**, 711–716 (2013).
424. Sangesland, M. et al. Allelic polymorphism controls autoreactivity and vaccine elicitation of human broadly neutralizing antibodies against influenza virus. *Immunity* **55**, 1693–1709.e8 (2022).
425. Lee, J. H. et al. Vaccine genetics of IGHV1-2 VRC01-class broadly neutralizing antibody precursor naïve human B cells. *NPJ Vaccines* **6**, 113 (2021).
426. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
427. Zinkernagel, R. M. & Doherty, P. C. Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature* **248**, 701–702 (1974).

428. Zinkernagel, R. M. & Doherty, P. C. Immunological surveillance against altered self components by sensitised T lymphocytes in lymphocytic choriomeningitis. *Nature* **251**, 547–548 (1974).
429. Ashby, K. M. & Hogquist, K. A. A guide to thymic selection of T cells. *Nat. Rev. Immunol.* https://doi.org/10.1038/s41577-023-00911-8 (2023).
430. Brown, A. et al. MHC heterozygosity reduces the T cell receptor repertoire. *Cell* https://doi.org/10.2139/ssrn.4555926 (2023).
431. Xu, J. et al. T cell receptor β repertoires in patients with COVID-19 reveal disease severity signatures. *Front. Immunol.* **14**, 1190844 (2023).
432. Park, J. J. et al. Machine learning identifies T cell receptor repertoire signatures associated with COVID-19 severity. *Commun. Biol.* **6**, 76 (2023).
433. Pedrioli, A. & Oxenius, A. Single B cell technologies for monoclonal antibody discovery. *Trends Immunol.* **42**, 1143–1158 (2021).
434. Raybould, M. I. J. et al. Five computational developability guidelines for therapeutic antibody profiling. *Proc. Natl Acad. Sci. USA* **116**, 4025–4030 (2019).
435. Shuai, R. W., Ruffolo, J. A. & Gray, J. J. IgLM: infilling language modeling for antibody sequence design. *Cell Syst.* **14**, 979–989.e4 (2023).
436. Bachas, S. et al. Antibody optimization enabled by artificial intelligence predictions of binding affinity and naturalness. Preprint at *bioRxiv* https://doi.org/10.1101/2022.08.16.504181 (2022).
437. Rosenberg, S. A. & Restifo, N. P. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* **348**, 62–68 (2015).
438. Sadelain, M., Rivière, I. & Riddell, S. Therapeutic T cell engineering. *Nature* **545**, 423–431 (2017).
439. Shafer, P., Kelly, L. M. & Hoyos, V. Cancer therapy with TCR-engineered T cells: current strategies, challenges, and prospects. *Front. Immunol.* **13**, 835762 (2022).
440. Yang, J., Chen, Y., Jing, Y., Green, M. R. & Han, L. Advancing CAR T cell therapy through the use of multidimensional omics data. *Nat. Rev. Clin. Oncol.* **20**, 211–228 (2023).
441. Joshi, K., Milighetti, M. & Chain, B. M. Application of T cell receptor (TCR) repertoire analysis for the advancement of cancer immunotherapy. *Curr. Opin. Immunol.* **74**, 1–8 (2022).
442. Castellanos-Rueda, R., Di Roberto, R. B., Schlatter, F. S. & Reddy, S. T. Leveraging single-cell sequencing for chimeric antigen receptor T cell therapies. *Trends Biotechnol.* **39**, 1308–1320 (2021).
443. Stucchi, A., Maspes, F., Montee-Rodrigues, E. & Fousteri, G. Engineered T_{reg} cells: the heir to the throne of immunotherapy. *J. Autoimmun*. https://doi.org/10.1016/j.jaut.2022.102986 (2023).
444. Raffin, C. et al. Development of citrullinated-vimentin-specific CAR for targeting T_{regs} to treat autoimmune rheumatoid arthritis. *J. Immunol.* **200**, 176.17 (2018).
445. Venturi, V., Kedzierska, K., Turner, S. J., Doherty, P. C. & Davenport, M. P. Methods for comparing the diversity of samples of the T cell receptor repertoire. *J. Immunol. Methods* **321**, 182–195 (2007).
446. Mayer, A., Balasubramanian, V., Mora, T. & Walczak, A. M. How a well-adapted immune system is organized. *Proc. Natl Acad. Sci. USA* **112**, 5950–5955 (2015).
447. Schnaack, O. H. & Nourmohammad, A. Optimal evolutionary decision-making to store immune memory. *eLife* **10**, e61346 (2021).
448. Vieira, M. C. et al. Germline-encoded specificities and the predictability of the B cell response. *PLoS Pathog.* **19**, e1011603 (2023).
449. Giudicelli, V., Chaume, D. & Lefranc, M.-P. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Res.* **33** (Suppl. 1), D256–D261 (2004).

Author contributions

Introduction (E.M.-F., V.G. and V.M.); Experimentation (K.L.Q., P.B. and V.M.); Results (H.B., K.L.Q., P.R., V.G. and V.M.); Applications (E.M.-F., H.B., V.G. and V.M.); Reproducibility and data deposition (E.M.-F., H.B., V.G. and V.M.); Limitations and optimizations (H.B., V.G. and V.M.); Outlook (E.M.-F. and V.G.); Overview of the Primer (E.M.-F. and V.G.).

Competing interests

V.G. declares advisory board positions in aiNET GmbH, Enpicom B.V, Specifica Inc., Adaptyv Biosystems, EVQLV, Omniscope, Diagonal Therapeutics and Absci; and is a consultant for Roche/Genentech, immunai, Proteinea and LabGenius. P.B. is now an employee of Parean Biotechnologies. E.M.-F., H.B., K.L.Q., P.R. and V.M. declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43586-023-00284-1.

Peer review information *Nature Reviews Methods Primers* thanks B. Chain, S. Efroni, J. Harris, K. Rodgers and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Related links

AIRR-compliant software tools: https://docs.airr-community.org/en/stable/swtools/airr_swtools_compliant.html
Guidance for AIRR software tools: https://docs.airr-community.org/en/stable/swtools/airr_swtools_standard.html
Guide for submission of AIRR-seq data to NCBI: https://docs.airr-community.org/en/stable/miairr/guide_miairr_ncbi.html
iReceptor gateway: https://gateway.ireceptor.org/login
MiAIRR: https://docs.airr-community.org/en/stable/miairr/introduction_miairr.html