nf-core/airrflow: an adaptive immune receptor repertoire analysis workflow employing the Immcantation framework

Gisela Gabernet[1,2]*, Susanna Marquez[1]*, Robert Bjornson[3], Alexander Peltzer[4], Hailong Meng[1], Edel Aron[5], Noah Y. Lee[5], Cole Jensen[5], David Ladd[6], Friederike Hanssen[2,7,8], Simon Heumos[2,7,8], nf-core community, Gur Yaari[9], Markus C. Kowarik[10,11], Sven Nahnsen[||2,8,12], Steven H. Kleinstein[||1,5,13]

[1]Department of Pathology, Yale School of Medicine, New Haven, CT, USA.
[2]Quantitative Biology Center, Eberhard-Karls University of Tübingen, Tübingen, Germany.
[3]Yale Center for Research Computing, New Haven, CT, USA.
[4]Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany.
[5]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA.
[6]oNKo-Innate Pty Ltd, Melbourne, Victoria, Australia.
[7]Department of Computer Science, Eberhard-Karls University of Tübingen, Tübingen, Germany.
[8]M3 Research Center, University Hospital, Tübingen, Germany.
[9]Faculty of Engineering, Bar Ilan University, Ramat Gan, Israel.
[10]Department of Neurology and Stroke, Center for Neurology, Eberhard-Karls University of Tübingen, Tübingen, Germany.
[11]Hertie Institute for Clinical Brain Research, Eberhard-Karls University of Tübingen, Tübingen, Germany.
[12]Institute for Bioinformatics and Medical Informatics (IBMI), Eberhard-Karls University of Tübingen, Tübingen, Germany.
[13]Department of Immunobiology, Yale School of Medicine, New Haven, CT, USA.
*Equal contribution; [||]co-senior authors.
Corresponding author: gisela.gabernet@yale.edu

ABSTRACT

Motivation

Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) is a valuable experimental tool to study the immune state in health and following immune challenges such as infectious diseases, (auto)immune diseases, and cancer. Several tools have been developed to reconstruct B cell and T cell receptor sequences from AIRR-seq data and infer B and T cell clonal relationships. However, currently available tools offer limited parallelization across samples, scalability or portability to high-performance computing infrastructures.

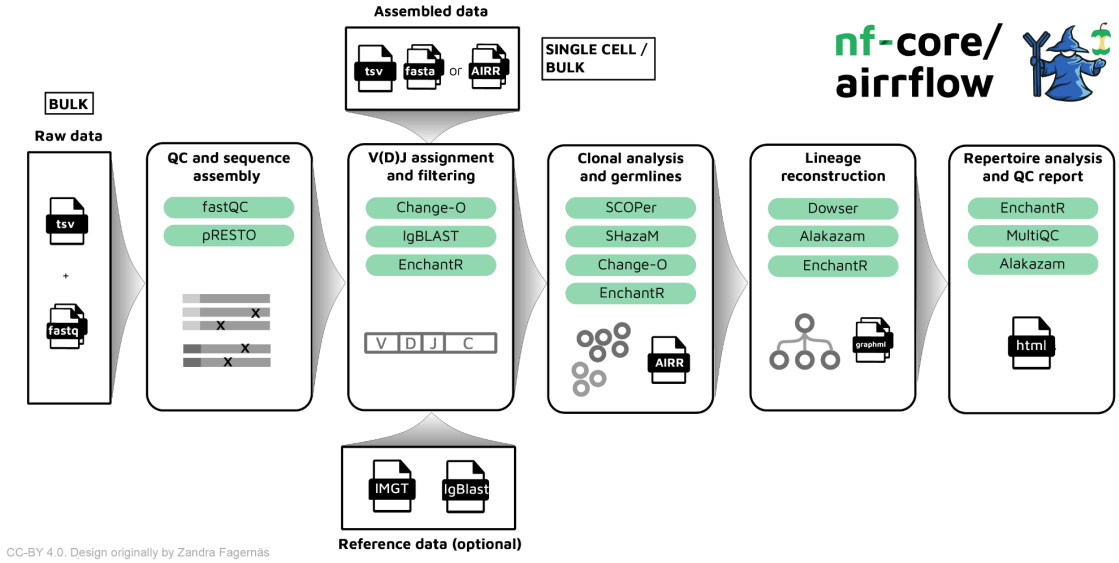
Results

To address this need, we developed nf-core/airrflow, an end-to-end bulk and single-cell AIRR-seq processing workflow which integrates the Immcantation Framework following BCR and TCR sequencing data analysis best practices. The Immcantation Framework is a comprehensive toolset, which allows the processing of bulk and single-cell AIRR-seq data from raw read processing to clonal inference. nf-core/airrflow is written in Nextflow and is part of the nf-core project, which collects community contributed and curated Nextflow workflows for a wide variety of analysis tasks. We assessed the performance of nf-core/airrflow on simulated sequencing data and show example results with real datasets. To demonstrate the applicability of nf-core/airrflow to the high-throughput processing of large AIRR-seq datasets, we validated and extended previously reported findings of convergent antibody responses to SARS-CoV-2 by analyzing 97 COVID-19 infected individuals and 99 healthy controls retrieved from public databases.

Availability and implementation

nf-core/airrflow is available free of charge, under the MIT license on GitHub (https://github.com/nf-core/airrflow). Documentation and example results are available at https://nf-co.re/airrflow.

Visual abstract



1. Introduction

B cells and T cells, key components of adaptive immunity, recognize foreign pathogens and provide long-term protection against them. They are also implicated in auto-immune diseases when eliciting a deleterious response against self-antigens. The antigen recognition is performed through membrane receptors, termed B cell and T cell receptors (BCR and TCR, respectively). BCRs in their secreted form are termed antibodies. BCRs and TCRs are generated during cell maturation by the somatic DNA recombination of a number of variable (V), diversity (D), and joining (J) gene segments in the immunoglobulin (IGH, IGK and IGL) and TCR (TRA and TRB) loci (Tonegawa 1983; Roth 2014). Additional nucleotide deletion/insertion at the gene boundaries generate practically unique receptors for each maturing cell (Alt and Baltimore 1982; Lafaille *et al.* 1989). The constant region is encoded in additional exons and added to the sequence during transcription and splicing. Two identical heavy and light chains constitute the BCR, while an alpha and a beta chain constitute the TCR. In BCRs and antibodies, the heavy chain constant region determines the isotype. The diversity of BCRs is further increased upon antigen encounter and activation by a process termed somatic hypermutation (SHM), which provides the substrate for antigen driven selection leading to increased affinity to their targets (Papavasiliou and Schatz 2002). The collection of BCRs and TCRs in an individual, tissue, or cell subset is referred to as the adaptive immune receptor repertoire (AIRR). Characterizing AIRRs is relevant for the study of the immune state of individuals in health and following alterations such as vaccines (Kotagiri *et al.* 2022; Safonova *et al.* 2022), infectious diseases (Parameswaran *et al.* 2013; Chen *et al.* 2021), (auto)immune diseases (Bashford-Rogers *et al.* 2013; Stern *et al.* 2014; Vander Heiden *et al.* 2017; Ramadoss and Robinson 2020), and cancer (Frank *et al.* 2023).

Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) enables the recovery of the collection of AIRRs sampled from an individual (Boyd and Joshi 2014). AIRR-seq can be performed from the genomic DNA or expressed RNA libraries. Methods based on the targeted amplification of the expressed BCR and TCR transcripts are more widely used due to the possibility of additionally sequencing the constant region, which determines the BCR isotype, and the presence of multiple RNA molecules per cell that can enhance sensitivity and allow for error correction when employing protocols that incorporate Unique Molecular Identifiers (UMIs). Single-cell sequencing technologies have revolutionized the field by allowing the characterization of the paired heavy and light (BCR) or alpha and beta (TCR) chain receptor sequences linked to the individual cell's transcriptomic profiles (Papalexi and Satija 2018). The processing of AIRR-seq data requires steps that are distinct from other bioinformatics pipelines (Georgiou *et al.* 2014; Yaari and Kleinstein 2015), which has led to the development of a multitude of tools and frameworks (Shugay *et al.* 2014; Stern *et al.* 2014; Vander Heiden *et al.* 2014; Bolotin *et al.* 2015; Cortina-Ceballos *et al.* 2015; Gadala-maria *et al.* 2015; Gupta *et al.* 2015; Kuchenbecker *et al.* 2015; Shlemov *et al.* 2017; Lindeman *et al.* 2018; Nouri and Kleinstein 2018, 2020; Hoehn *et al.* 2019; Sturm *et al.* 2020; Hoehn, Pybus and Kleinstein 2022; Suo *et al.* 2023). However, individual tools offer limited parallelization across samples, scalability

or native support for alternative compute infrastructures such as high-performance computing (HPC) clusters or commercial clouds.

Workflow management systems such as Nextflow (Di Tommaso *et al.* 2017) or Snakemake (Mölder *et al.* 2021) allow the incorporation of tools into highly configurable analysis workflows which can be triggered with a single command, and offer implicit massive parallelization across the analyzed samples. Workflow management systems additionally help ensure reproducible analysis results in several ways: by stating the commands used to execute the individual tools, reporting the parameters used to launch the workflow, and easing the use of container engines in the individual analysis steps. As an example, Nextflow supports the download and execution of Docker, Apptainer, and other container engines from public container repositories by solely specifying their registry address, name and version in the desired workflow steps. Volume mounts for the start and stop commands are set by Nextflow in the background. Containers provide a controlled running environment with fixed tool versions and system libraries, which is critical in ensuring reproducible results and portability across compute infrastructures, such as clouds (Di Tommaso *et al.* 2017).

To address the need for a high-throughput, portable analysis workflow which incorporates the necessary steps for AIRR-seq data analysis we developed nf-core/airrflow, a flexible workflow written in the Nextflow language. The workflow utilizes the Immcantation framework, a comprehensive collection of open-source software to process AIRR-seq data from start-to-finish, covering sequence assembly and error correction (Vander Heiden *et al.* 2014), alignment to an international immunogenetics information system (IMGT) (Lefranc *et al.* 2015; Collins *et al.* 2023) BCR and TCR reference data with IgBLAST (Ye *et al.* 2013), clonal relationship inference (Gadala-maria *et al.* 2015; Gupta *et al.* 2015, 2017; Nouri and Kleinstein 2018), reconstruction of clonal lineages (Hoehn *et al.* 2019; Hoehn, Pybus and Kleinstein 2022), and the identification of predictive repertoire properties and sequence motifs (Stern *et al.* 2014; Olson *et al.* 2019). nf-core/airrflow is a flexible workflow which supports the analysis of both bulk and single-cell sequencing datasets generated with a multitude of protocols, allowing as well the combined processing of datasets with mixed bulk and single-cell modalities. The workflow is part of the nf-core project (Ewels *et al.* 2020), which defines Nextflow implementation best practices including containerization of all software tools with biocontainers (da Veiga Leprevost *et al.* 2017) whenever possible, reporting of all software versions in a MultiQC report (Ewels *et al.* 2016), continuous integration testing with example test data, and portability testing to cloud infrastructures. These features aim at making nf-core/airrflow easy to use and install, requiring only Nextflow and a container engine as dependencies, as well as ensuring that the obtained results will be reproducible and portable across compute infrastructures.

We benchmarked the workflow on simulated ground truth BCR repertoire sequencing data and showed its superiority to alternative existing tools. As an application use-case, we validated the findings of convergent antibody sequences across three COVID-19 diagnosed individuals identified by Robbiani *et al.* (Robbiani *et al.* 2020) in a larger dataset comprising BCR repertoire samples from 97 individuals diagnosed with COVID-19 and 99 healthy controls retrieved from

the iReceptor database (Corrie *et al.* 2018). Thus, we showcased the applicability of nf-core/airrflow to validate findings in small patient cohorts with data from publicly available sources.

2. Workflow design and implementation

2.1 Workflow outline

nf-core/airrflow is a high-throughput workflow for the end-to-end analysis of AIRR bulk and single-cell sequencing data utilizing the Immcantation framework (Stern *et al.* 2014; Vander Heiden *et al.* 2014; Gadala-maria *et al.* 2015; Gupta *et al.* 2015; Nouri and Kleinstein 2018, 2020; Hoehn *et al.* 2019; Hoehn, Pybus and Kleinstein 2022). It encompasses sequencing read assembly, V(D)J and constant region allele assignment, clonal inference, and repertoire analysis including quality filtering and quality control reporting in each step (Figure 1A). Multiple protocols for raw bulk sequencing data analysis are supported, including multiplexed PCR and 5' Rapid amplification of cDNA ends (5'-RACE) based protocols, with or without the inclusion of UMIs that allow the correction of sequencing errors and amplification biases. The analysis of assembled bulk and single-cell BCR and TCR sequencing data generated with the 10x genomics platform or other platforms provided in the AIRR rearrangement format is also supported (Cohn *et al.* 2023). The different options can be specified through command line parameters when executing a pipeline or a Nextflow parameters file that is provided at runtime.

2.2 Input data

The bulk raw AIRR-seq data processing starts from the FastQ files. Those are specified within a metadata file following the AIRR Study Reporting standard (MiAIRR) (Vander Heiden *et al.* 2018), which includes the sample identifiers, species, target locus (IG or TR), and other information required for the repertoire analysis. The primer and/or RACE linker sequences used for library preparation are additional required inputs for processing bulk AIRR-seq data. Examples of supported kits are the commercially available New England Biolabs kits (New England Biolabs), or TAKARA Bio BCR and TCR profiling kits (Takara Bio). It is possible to specify in a flexible manner the start position of the primer(s) and UMI position and length, extending the applicability to other custom protocols. For single-cell AIRR-seq data analysis, the pipeline accepts as input sequence tables in the AIRR rearrangement format (Vander Heiden *et al.* 2018), such as the ones provided by the 10X Genomics Cell Ranger, or tools that can extract BCR and TCR sequences from single-cell RNA-seq sequencing data such as TRUST4 (Song *et al.* 2021). The analysis workflow then starts with the optional V(D)J and constant region allele reassignment with IgBLAST, to ensure that the sequences are annotated with the same IMGT (Lefranc *et al.* 2015) BCR and TCR reference data, and IMGT gaps are included in the alignments. It is also possible to process readily assembled bulk sequencing data starting with this step, by providing the fully assembled and error-corrected reads in fasta or AIRR rearrangement format.

2.3 Read QC and sequence assembly

Sequencing read quality control is performed using fastp (Chen *et al.* 2018) for general read quality statistics and adapter trimming, if indicated. The pRESTO (Vander Heiden *et al.* 2014) Immcantation tool is employed for all other sequence assembly and error correction steps. Pre-processing involves the filtering of low-quality reads, extraction of UMIs, and primer masking. For protocols including UMIs, a consensus sequence is built with sequences comprising the same UMI. Optionally, the *cluster set* process allows dealing with insufficient UMI diversity, which might occur when UMIs are too short for the library diversity. Read mates are paired and assembled, and the primer information along with any metadata relevant to the downstream analysis is annotated in the fastq header with a format standardized by pRESTO. Duplicate sequences are collapsed, annotating the individual number of representative reads where this sequence was observed. Finally, sequences with insufficient support, less than two representatives by default, are filtered out.

2.4 V(D)J annotation and alignment QC

V(D)J and constant region allele assignment is performed by aligning the assembled reads to the relevant BCR and TCR reference data with IgBLAST (Ye *et al.* 2013; Collins *et al.* 2023), and processing the alignment results with the Change-O (Gupta *et al.* 2015) Immcantation tool. Currently, human and mouse IMGT reference data are supported and directly downloaded by the workflow. Alternatively, users can provide curated custom reference data for other species in the same format. Several QC and filtering steps are performed after alignment. First, the provided desired locus is checked to match the assigned V gene chain. By default, sequence alignments that contain less than 200 informative positions, more than 10% N nucleotides, or that generate non-productive transcripts are filtered out. For bulk sequencing data, chimeric reads sometimes generated during PCR amplification can be optionally removed. Likely cross-sample sequence contamination is detected by identifying samples with a high percentage of overlapping identical VDJ sequences. As a last step for bulk sequencing QC, duplicated sequences are collapsed and the duplicate number is annotated. Filtering and quality control steps for single-cell sequencing data include removing cells containing multiple heavy chains or only light chains, as well as detecting and removing cross-sample sequence contaminants with identical VDJ region sequences and cell barcodes.

2.5 Clonal inference

Clonal inference involves clustering the set of BCR and TCR sequences into clones, which are defined as the group of cells that descend from a common ancestor. This is a harder problem for B cells than T cells, as SHM introduces targeted mutations in the BCR sequences during B cell clonal expansion, generating a diversity of sequences within the same clone that are relevant for antibody affinity maturation. Clonal relationships are identified with the Immcantation SCOPer (Nouri and Kleinstein 2018, 2020) integrated methods. By default, sequences are partitioned into groups with the same V gene (IGHV or TRBV), J gene (IGHJ or TRBJ) and junction length - the junction sequence corresponds to the highly variable complementarity

determining region 3 (CDR3) region with additional flanking nucleotides that encode two conserved amino acids in the 5' and 3' end -. Hierarchical clustering with single-linkage is applied to the heavy chain junction nucleotide sequence, with the length-normalized hamming distance as a distance metric. For B cell repertoires, the hamming distance threshold used to distinguish sequences belonging to the same clonal group can be estimated using the SHazaM (Yaari, Uduman and Kleinstein 2012; Yaari *et al.* 2013) Immcantation package. As T cells do not undergo somatic hypermutation, T cell clones are defined as having identical TCR sequences. In the current implementation, TCR sequences sharing the same TRBV and TRBJ alleles and displaying identical junction nucleotide sequences (hamming distance threshold of zero) are assigned to the same clone, because in some sequencing protocols the full V(D)J sequence will not be recovered. For single-cell data where paired heavy and light chains are available, the clones can be additionally partitioned according to the light chain (or alpha chain) V genes, J genes and junction length. Summary reports of the identified distance threshold for all the samples and clonal relationships facilitate a visual validation of the results (Figure 1B). Clonal repertoire properties such as clonal abundance and diversity are calculated with the Alakazam Immcantation package (Figure 1C). Clonal lineage trees can be optionally reconstructed with Dowser (Hoehn *et al.* 2021), an Immcantation tool which includes several methods for lineage tree reconstruction such as IgPhyML (Hoehn, Pybus and Kleinstein 2022), a method specifically designed for B-cell lineage tree reconstruction (Figure 1D).

2.6 Reporting

HTML reports on individual workflow steps can be found in the respective process output folders, and are generated through EnchantR, an R helper package developed along with the Nextflow workflow, which is now part of Immcantation. An interactive report is generated comprising the number of sequences obtained after each workflow step. The report in Rmarkdown format can be downloaded by the user and modified accordingly to meet the specific project analysis needs. Additionally, a MultiQC interactive report is generated, which contains a read quality control report over all the processed samples and the software versions employed by each workflow process.

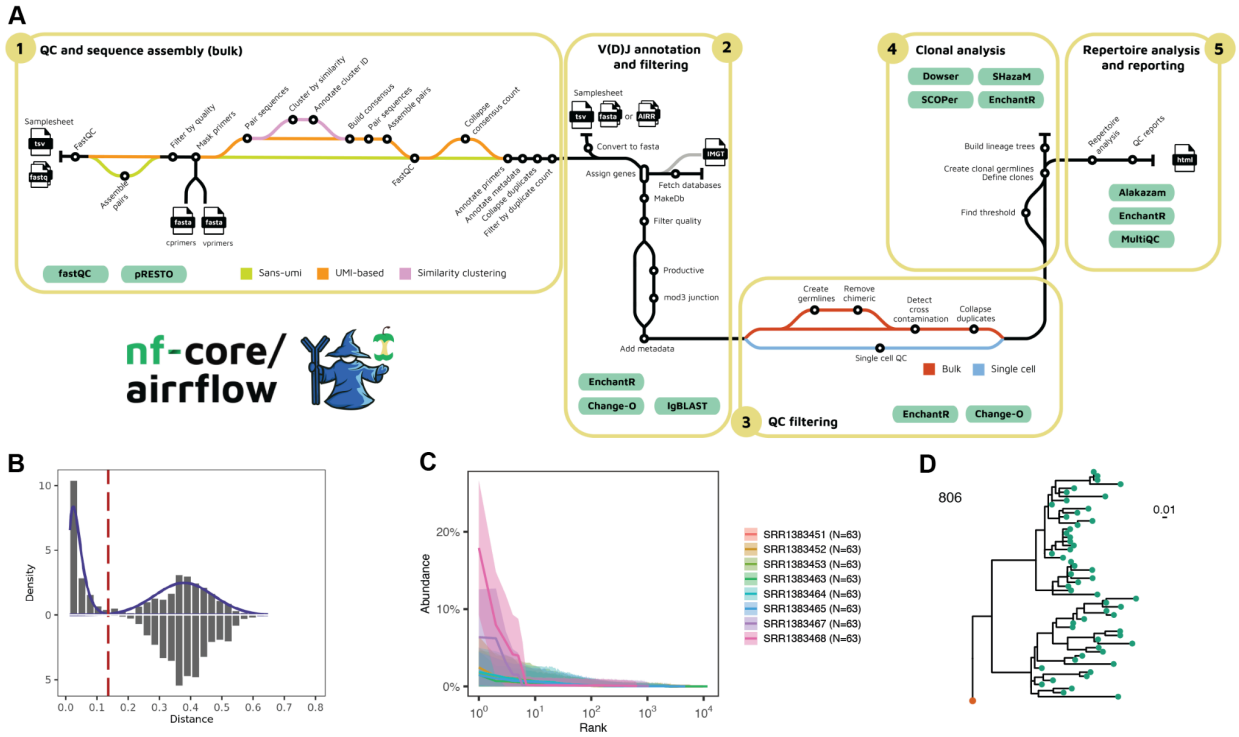


Figure 1. Schematic representation of the nf-core/airrflow workflow processes and detailed analysis steps. A. Workflow overview. QC and sequence assembly steps (1) are performed on the bulk raw sequencing reads. The workflow supports several sequencing protocols including multiplex PCR and 5'-RACE, with or without UMIs. For single-cell sequencing data or readily-assembled bulk sequencing reads, the analysis starts with V(D)J and C gene assignment with IgBLAST (2). Reference BCR and TCR sequences can directly be pulled from IMGT, or provided by the user for enhanced reproducibility. After V(D)J assignment, the sequences are filtered for meaningful alignments (minimum of base alignments, as well as % of accepted N nucleotides) and productive sequences. Additional QC steps can be applied to remove chimeric reads and duplicate sequences, and detect cross-sample contamination (3). Clonal groups are defined by hierarchical clustering using a hamming distance threshold which can be provided or estimated from the data (4). Lineage trees can be optionally reconstructed with the Dowser package. A final report summarizes the repertoire analysis results (5). MultiQC integrates the quality control results of the QC steps and reports the software versions of all executed processes in the workflow. To generate example data, nf-core/airrflow was run on full-size test data on AWS batch extracted from (Stern *et al.* 2014). Example of a hamming distance threshold plot (**B**), clonal abundance plot (**C**), and reconstructed lineage tree (**D**) from this dataset.

3. Results

3.1 nf-core/airrflow benchmarking with simulated data

To evaluate the ability of nf-core/airrflow to recover immune repertoire sequences and infer clonal relationships from ground truth sequencing data, we simulated three BCR receptor repertoires with known V(D)J sequences, varying clonal abundances and increasing frequency of sequencing errors. The germline sequences were generated by simulating V(D)J recombination with ImmuneSIM (Weber *et al.* 2020). Clonal lineage trees and somatic hypermutation were simulated with SHazaM to obtain a power law and a uniform clonal size distribution (repA and repB, respectively) or extracted from a real BCR repertoire sample previously published (Ruschil *et al.* 2023) (repC). 5000 singleton sequences - representing naive B cells that are not clonally expanded - were added to the synthetic repertoires repA and repB to achieve a similar frequency as observed in the real BCR repertoire repC (see Supplementary methods and Supplementary Figure 1 for further details on the repertoire simulation).
We then simulated paired-end raw sequencing data for each repertoire with Grinder (Angly *et al.* 2012). To assess the impact of sequencing errors on the final sequence recovery, the sequencing data simulations were performed with increasing percentages of sequencing errors modeled in a linear fashion along the read length, increasing from zero to a predetermined percentage, in accordance with previous studies on Illumina sequencing error values and their distribution along the read positions (Schirmer *et al.* 2016; Stoler and Nekrutenko 2021). Five simulated libraries were prepared, with 0%, 0.1%, 0.25%, 0.5%, 1.0% sequencing errors in the center of the reads. Additionally, two library preparation protocols were compared: with (*UMI*) and without (*sans-UMI*) UMIs. Adding UMIs to the library preparation procedure offers the potential for sequencing error correction by constructing a consensus sequence of the recovered sequences with identical UMIs, which is a widely used strategy in BCR and TCR sequencing protocols.

The simulated BCR sequencing libraries were processed with nf-core/airrflow, and the ability to recover the original sequences in the simulated repertoires was evaluated (Figure 2, Suppl. Figure 2). We evaluated the proportion of correctly identified sequences for each of the repertoires with exact sequence matches (sensitivity exact matches) and additionally considered the matches of sequences that contain “N” nucleotides (Figure 2A and B) for a protocol including UMIs. The N-nucleotides are introduced when UMIs are used for error correction by building a sequence consensus, and the consensus base is under a certain frequency threshold (default minimum frequency 0.6) or quality threshold (default minimum quality 0) so there is insufficient consensus to call a particular base. nf-core/airrflow correctly recovered over 99% of the sequences in each of the three repertoires when no simulated sequencing errors were present. The data with sequencing errors decreased the proportion of correctly identified sequences, but sensitivity was maintained above 97% for exact sequence

matches and 98% for matches containing “N” nucleotides due to not reaching sufficient consensus. Two incorrect sequences were reported due to a rare occurrence of duplicate UMIs being assigned to two highly similar sequences that are part of the same clone, and two sequences with the same UMI with simulated errors at the same position (Figure 2 C). The number of missing sequences ranged from 100 to 300 from a total of 21,321 (repA), 20,959 (repB), and 15,329 (repC). The sensitivity and number of missing sequences was comparable to MiXCR (Bolotin *et al.* 2015), an alternative tool for bulk and single-cell AIRR-seq data analysis (Figure 2 A-D, Supplementary Table 1). When simulating a protocol without UMI, the sensitivity started at 99% with the absence of sequencing errors, but dropped below 50% and up to 125 incorrect sequences with increasing sequencing errors (Suppl. Figure 3A and B), reinforcing the importance of utilizing protocols that include UMI error correction. The high number of missing sequences is due to a quality control step that eliminates sequences that do not contain at least two representative copies, which are attributed to sequencing errors (Suppl. Figure 3C).

In addition to recovering the original sequences in the sample, AIRR-seq data analysis often involves determining the clonal relationships of the individual sequences. This is important in order to assess whether a sequence comes from an expanded clone of the same progenitor cell. This step is particularly important in BCR data analysis, as mutations are introduced during clonal expansion by targeted somatic hypermutation. Thus, we assessed the ability of the workflow to recover the original number and size distribution of B-cell clones (Figure 2E). Close to 5,500 clones and 5,700 clones were identified for repA and repB (ground truth 5,100 for both), and 6,500 clones for repC (ground truth 6,305). The number of clones was overestimated by 7%, 11%, and 3%, respectively, but was robust with respect to simulated sequencing errors. The clonal abundance distribution (Figure 2F) reflected the true clonal distribution in all three repertoires for the simulated protocol with UMIs. Clonal inference by the Immcantation tools incorporated into nf-core/airrflow was superior to the inference method implemented in MiXCR (Figure 2E-F). The inferred number of clones and clonal abundance by both tools were affected by the increasing sequencing errors in the sans-UMI protocol, highlighting once more the importance of UMI error correction (Suppl. Figure 2 D and E).

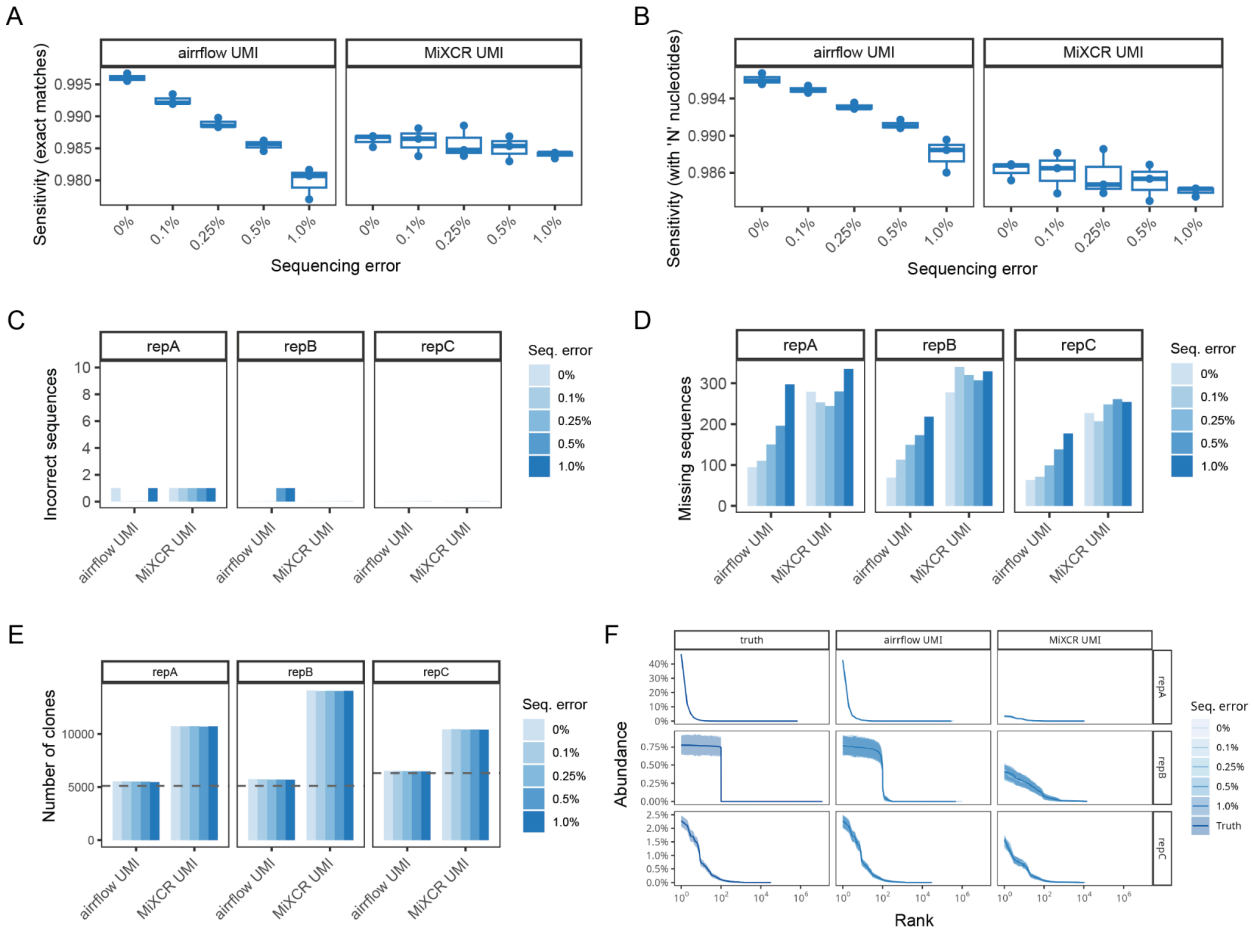


Figure 2. Performance assessment of the nf-core/airrflow pipeline on three simulated BCR repertoires compared to MiXCR. Sensitivity of the nf-core/airrflow and MiXCR pipelines on the data simulated with UMIs. Sensitivity was calculated for exact VDJ sequence matches to the truth repertoires (**A**), and matches that contain N nucleotides due to not reaching sufficient consensus (**B**). **C**. Number of incorrect sequences. **D.** Number of sequences present in the truth repertoires that were not identified by the pipelines. **E**. Number of clones identified by each pipeline. The discontinuous line indicates the true number of clones in the simulated repertoires. **F**. Mean clonal abundance (solid line) with max and min intervals (shaded area) from n=200 bootstrap samples of N=15,250 sequences of the three simulated repertoires.

3.2 Case study: validating convergent antibody responses from public repositories

Previous studies have identified convergent antibody responses to the SARS-CoV-2 virus (Nielsen *et al.* 2020; Robbiani *et al.* 2020; Schultheiß *et al.* 2020; Wen *et al.* 2020; Chen *et al.* 2021). These are sequences with high sequence similarity across patients, that can indicate BCRs targeting epitopes important for the neutralization of the SARS-CoV-2 virus or that are

conserved across different strains. Robbiani *et al.* (Robbiani *et al.* 2020) identified a convergent antibody cluster comprising six BCR sequences across three COVID-19 infected subjects that were experimentally validated to bind the SARS-CoV-2 spike protein receptor binding domain (RBD). The identified convergent antibodies shared the same IGHV and IGHJ genes (IGHV1-58 and IGHJ3) and showed high similarity of the heavy chain CDR3 sequences (1 to 3 amino acid edits). To validate this finding and extend it to a larger number of subjects, we retrieved BCR repertoire datasets from COVID-19 diagnosed subjects from the iReceptor gateway (Corrie *et al.* 2018), together with healthy controls. These comprised datasets from 17 studies, including 496 antibody repertoires from 213 subjects sequenced with single-cell and bulk-based protocols, but did not include those reported by Robbiani *et al*. (Supl. Table 3). We processed the repertoires with nf-core/airrflow on an HPC cluster in just over 2 days (Figure 3B). 467 antibody repertoires from 196 subjects (99 healthy controls and 97 COVID-19 diagnosed) passed the QC criteria (Figure 3A, Suppl. Figure 3). We defined convergent antibody clusters by partitioning the processed sequences according to IGHV, IGHJ genes and junction length. We performed single-linkage hierarchical clustering on the junction amino acid sequences according to their pairwise hamming distances and defined convergent groups setting a length-normalized hamming distance threshold of 0.2, for each IGHV gene. We excluded convergent clusters that contained antibody sequences from a single subject, and clusters with sequences from healthy controls. We obtained 416 convergent antibody clusters of varying sizes with antibodies using the IGHV1-58 gene segment (Supplementary Figure 4, Figure3C). The convergent cluster with most sequences contained 1,189 sequences (182 unique junction sequences) from 20 subjects and 6 different studies (cluster 11312). This convergent cluster comprised BCRs with IGHV1-58 and IGHJ3 genes, and junction sequences at an edit distance of 1 to 3 amino acids from the six antibodies described by Robbiani *et al*. None of the other convergent clusters with IGHV1-58 and same junction length contained junction sequences at an edit distance of 3 amino acids or less. A sequence logo of all of the sequences in the convergent cluster 11312 revealed amino acid residue and properties conservation similar to the sequences reported in the original publication, with two conserved cysteines at positions 6 and 11, as well as two aspartate negative charges at positions 13 and 16 and a Phenylalanine at position 15 (Figure 3D and E), providing an indication of conserved specificity and the likely amino-acids important for binding. Thus, we conclude that this cluster represents the same convergent antibody sequences found by Robbiani *et al.*, and extended this convergence to additionally 20 individuals. This use case highlights the applicability of nf-core/airrflow to validate findings originally made in small in-house cohorts with reanalysis of large publicly available AIRR-seq datasets.

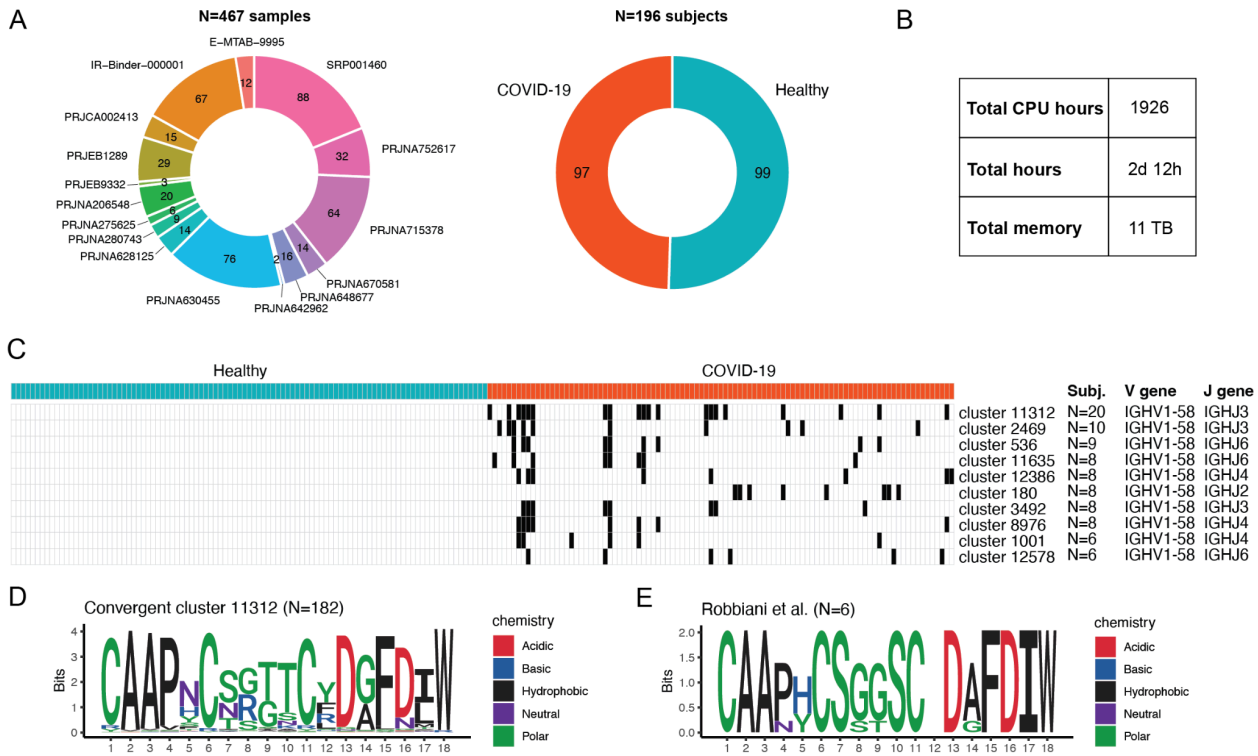


Figure 3. Application of nf-core/airrflow to process repertoires of COVID-19 infected or vaccinated participants together with healthy controls. A. Number of samples that passed the quality control after processing the data with nf-core/airrflow v3.2.0 and their distribution across projects. Number of subjects in each group (healthy or COVID-19 diagnosed). **B.** Resources needed to process the datasets on an HPC cluster. **C**. Top 10 antibody convergent clusters (rows) using the IGHV1-58 gene segment according to the number of unique COVID-19 subjects included in each cluster. Clusters containing sequences from a single individual, or from healthy controls were excluded. Each column represents a subject colored by their status (healthy or COVID-19 diagnosed). Black denotes the presence of sequences belonging to an antibody cluster for each particular subject. The annotations on the right denote the number of subjects with antibody sequences included in each convergent cluster, the heavy chain V gene assignment and the heavy chain J gene assignment of the majority of the sequences in the convergent cluster. **D**. Sequence logo of the N=182 junction sequences contained in the convergent cluster 11312. Coloring according to amino acid side chain properties. **E**. Sequence logo of the sequences published by Robbiani *et al.*, positively tested against the SARS-CoV-2 spike protein receptor binding domain.

4. Conclusions

In this work we present nf-core/airrflow, a high-throughput, easy installation and portable workflow for the analysis of bulk and single-cell AIRR sequencing data. The workflow enables following best practices for the end-to-end analysis of BCR and TCR sequencing data obtained with various experimental protocols, with minimal installation requirements and ensuring

reproducible results across different computing infrastructures, such as cloud and HPC. Implemented using the versatile Nextflow workflow manager, nf-core/airrflow enables the analysis of AIRRseq datasets by users less versed in this data modality and with limited programming backgrounds. We have demonstrated the scalability of nf-core/airrflow to hundreds of samples retrieved from public repositories. As part of the nf-core community project, contributions to the workflow from other developers are welcome and encouraged, and will be reviewed before incorporation into the pipeline. nf-core/airrflow current or future users are encouraged to join the nf-core community dedicated slack channel (https://nf-co.re/join) for questions or feature requests.

Funding

This work was funded in part by the Chan Zuckerberg Initiative EOSS4 [2021-237742] to GG, the Deutsche Forschungsgemeinschaft (DFG) under Germany's excellence Strategy [EXC 2180–390900677] (iFIT) and under [EXC 2124–390838134] (CMFI) to SN, and the National Institutes of Health grant R01AI104739 to SHK.

Conflict of interest

SHK receives consulting fees from Peraton. AP is an employee of Boehringer Ingelheim Pharma GmbH & Co KG and declares no conflict of interest. MCK has served on advisory boards and received speaker fees / travel grants from Merck, Sanofi-Genzyme, Novartis, Biogen, Janssen, Alexion, Celgene / Bristol-Myers Squibb and Roche. He has received research grants from Merck, Roche, Novartis, Sanofi-Genzyme and Celgene / Bristol-Myers Squibb. All other authors declare no conflicts of interest.

Data availability

The pipeline is available free of charge under the MIT license in GitHub (https://github.com/nf-core/airrflow), as part of the nf-core project. Detailed documentation and example results are available at https://nf-co.re/airrflow. The code to reproduce the data simulation, pipeline benchmark, and scripts to reproduce the analysis of the COVID and healthy datasets can be found on https://bitbucket.org/kleinstein/projects. The simulated repertoire sequencing data files were uploaded to Zenodo (https://doi.org/10.5281/zenodo.10070476).

References

Alt FW, Baltimore D. Joining of immunoglobulin heavy chain gene segments: implications from a chromosome with evidence of three D-JH fusions. *Proc Natl Acad Sci U S A* 1982;**79**:4118–22.

Angly FE, Willner D, Rohwer F *et al.* Grinder: a versatile amplicon and shotgun sequence

simulator. *Nucleic Acids Res* 2012;**40**:e94.
Bashford-Rogers RJM, Palser AL, Huntly BJ *et al.* Network properties derived from deep sequencing of human b-cell receptor repertoires delineate b-cell populations. *Genome Research* 2013;**23**:1874–84.
Bolotin DA, Poslavsky S, Mitrophanov I *et al.* MiXCR: software for comprehensive adaptive immunity profiling. *Nat Methods* 2015;**12**:380–1.
Boyd SD, Joshi SA. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. *Microbiol Spectr* 2014;**2**, DOI: 10.1128/microbiolspec.AID-0017-2014.
Chen EC, Gilchuk P, Zost SJ *et al.* Convergent antibody responses to the SARS-CoV-2 spike protein in convalescent and vaccinated individuals. *Cell Rep* 2021;**36**:109604.
Chen S, Zhou Y, Chen Y *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90.
Cohn H, Bloom N, Cai GY *et al.* Mpox vaccine and infection-driven human immune signatures: an immunological analysis of an observational study. *Lancet Infect Dis* 2023;**23**:1302–12.
Collins AM, Ohlin M, Corcoran M *et al.* AIRR-C Human IG Reference Sets: curated sets of immunoglobulin heavy and light chain germline genes. 2023:2023.09.01.555348.
Corrie BD, Marthandan N, Zimonja B *et al.* iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunol Rev* 2018;**284**:24–41.
Cortina-Ceballos B, Godoy-Lozano EE, Sámano-Sánchez H *et al.* Reconstructing and mining the B cell repertoire with ImmunediveRsity. *MAbs* 2015;**7**:516–24.
Di Tommaso P, Chatzou M, Floden EW *et al.* Nextflow enables reproducible computational workflows. *Nature Biotechnology* 2017;**35**:316–9.
Ewels P, Magnusson M, Lundin S *et al.* MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8.
Ewels PA, Peltzer A, Fillinger S *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* 2020;**38**:276–8.
Frank ML, Lu K, Erdogan C *et al.* T-Cell Receptor Repertoire Sequencing in the Era of Cancer Immunotherapy. *Clin Cancer Res* 2023;**29**:994–1008.
Gadala-maria D, Yaari G, Uduman M *et al.* Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proceedings of the National Academy of Sciences* 2015;**112**:1–9.
Georgiou G, Ippolito GC, Beausang J *et al.* The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology* 2014;**32**:158–68.
Gupta NT, Adams KD, Briggs AW *et al.* Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. *The Journal of Immunology* 2017;**198**:2489–99.
Gupta NT, Vander Heiden JA, Uduman M *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* 2015;**31**:3356–8.
Hoehn KB, Pybus OG, Kleinstein SH. Phylogenetic analysis of migration, differentiation, and class switching in B cells. *PLOS Computational Biology* 2022;**18**:e1009885.
Hoehn KB, Turner JS, Miller FI *et al.* Human B cell lineages associated with germinal centers following influenza vaccination are measurably evolving. *eLife* 2021;**10**:e70873.
Hoehn KB, Vander Heiden JA, Zhou JQ *et al.* Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proceedings of the National Academy of Sciences* 2019;**116**:22664–72.
Kotagiri P, Mescia F, Rae WM *et al.* B cell receptor repertoire kinetics after SARS-CoV-2

infection and vaccination. *Cell Reports* 2022;**38**:110393.
Kuchenbecker L, Nienen M, Hecht J *et al.* IMSEQ--a fast and error aware approach to immunogenetic sequence analysis. *Bioinformatics* 2015;**31**:2963–71.
Lafaille JJ, DeCloux A, Bonneville M *et al.* Junctional sequences of T cell receptor gamma delta genes: implications for gamma delta T cell lineages and for a novel intermediate of V-(D)-J joining. *Cell* 1989;**59**:859–70.
Lefranc M-P, Giudicelli V, Duroux P *et al.* IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res* 2015;**43**:D413-422.
Lindeman I, Emerton G, Mamanova L *et al.* BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat Methods* 2018;**15**:563–5.
Mölder F, Jablonski KP, Letcher B *et al.* Sustainable data analysis with Snakemake. 2021, DOI: 10.12688/f1000research.29032.1.
New England Biolabs. NEBNext Immune Sequencing Kit. https://www.neb.com/en-us/products/e6320-nebnext-immune-sequencing-kit-human#Product%20Information. (15 December 2023, date last accessed).
Nielsen SCA, Yang F, Jackson KJL *et al.* Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. *Cell Host & Microbe* 2020;**28**:516-525.e5.
Nouri N, Kleinstein SH. A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. *Bioinformatics* 2018;**34**:i341–9.
Nouri N, Kleinstein SH. Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data. *PLoS Comput Biol* 2020;**16**:e1007977.
Olson BJ, Moghimi P, Schramm CA *et al.* sumrep: A Summary Statistic Framework for Immune Receptor Repertoire Comparison and Model Validation. *Frontiers in Immunology* 2019;**10**:2533.
Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* 2018;**18**:35–45.
Papavasiliou FN, Schatz DG. Somatic Hypermutation of Immunoglobulin Genes: Merging Mechanisms for Genetic Diversity. *Cell* 2002;**109**:S35–44.
Parameswaran P, Liu Y, Roskin KM *et al.* Convergent Antibody Signatures in Human Dengue. *Cell Host & Microbe* 2013;**13**:691–700.
Ramadoss NS, Robinson WH. Characterizing the BCR repertoire in immune-mediated diseases. *Nat Rev Rheumatol* 2020;**16**:7–8.
Robbiani DF, Gaebler C, Muecksch F *et al.* Convergent antibody responses to SARS-CoV-2 in convalescent individuals | Nature. *Nature* 2020;**584**:437–42.
Roth DB. V(D)J Recombination: Mechanism, Errors, and Fidelity. *Microbiology Spectrum* 2014;**2**:10.1128/microbiolspec.mdna3-0041–2014.
Ruschil C, Gabernet G, Kemmerer CL *et al.* Cladribine treatment specifically affects peripheral blood memory B cell clones and clonal expansion in multiple sclerosis patients. *Frontiers in Immunology* 2023;**14**.
Safonova Y, Shin SB, Kramer L *et al.* Variations in antibody repertoires correlate with vaccine responses. *Genome Res* 2022;**32**:791–804.
Schirmer M, D'Amore R, Ijaz UZ *et al.* Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics* 2016;**17**:125.
Schultheiß C, Paschold L, Simnica D *et al.* Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. *Immunity* 2020;**53**:442-455.e4.
Shlemov A, Bankevich S, Bzikadze A *et al.* Reconstructing Antibody Repertoires from

Error-Prone Immunosequencing Reads. *J Immunol* 2017;**199**:3369–80.
Shugay M, Britanova OV, Merzlyak EM *et al.* Towards error-free profiling of immune repertoires. *Nat Methods* 2014;**11**:653–5.
Song L, Cohen D, Ouyang Z *et al.* TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat Methods* 2021;**18**:627–30.
Stern JNH, Yaari G, Vander Heiden JA *et al.* B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci Transl Med* 2014;**6**:248ra107.
Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics* 2021;**3**:lqab019.
Sturm G, Szabo T, Fotakis G *et al.* Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. *Bioinformatics* 2020;**36**:4817–8.
Suo C, Polanski K, Dann E *et al.* Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins. *Nat Biotechnol* 2023:1–12.
Takara Bio. SMARTer Human BCR Profiling Kit. https://www.takarabio.com/products/next-generation-sequencing/immune-profiling/human-repertoire/human-bcr-profiling-kit-for-illumina-sequencing. (15 December 2023, date last accessed).
Tonegawa S. Somatic generation of antibody diversity. *Nature* 1983;**302**:575–81.
Vander Heiden JA, Marquez S, Marthandan N *et al.* AIRR Community Standardized Representations for Annotated Immune Repertoires. *Frontiers in Immunology* 2018;**9**.
Vander Heiden JA, Stathopoulos P, Zhou JQ *et al.* Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing. *J Immunol* 2017;**198**:1460–73.
Vander Heiden JA, Yaari G, Uduman M *et al.* pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* 2014;**30**:1930–2.
da Veiga Leprevost F, Grüning BA, Alves Aflitos S *et al.* BioContainers: an open-source and community-driven framework for software standardization. *Bioinformatics* 2017;**33**:2580–2.
Weber CR, Akbar R, Yermanos A *et al.* immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. *Bioinformatics* 2020;**36**:3594–6.
Wen W, Su W, Tang H *et al.* Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *Cell Discov* 2020;**6**:31.
Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine* 2015;**7**:1–14.
Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. *Nucleic Acids Res* 2012;**40**:e134.
Yaari G, Vander Heiden JA, Uduman M *et al.* Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front Immunol* 2013;**4**:358.
Ye J, Ma N, Madden TL *et al.* IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res* 2013;**41**:W34-40.