PLOS COMPUTATIONAL BIOLOGY

SOFTWARE

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 G. Jensen[5], David Ladd[6], Mark Polster[2,7,8], Friederike Hanssen[2,7,8], Simon Heumos[2,7,8], nf-core community[¶], Gur Yaari[9], Markus C. Kowarik[10,11], Sven Nahnsen[2,7,8,12‡], Steven H. Kleinstein[1,5,13‡]

1 Department of Pathology, Yale School of Medicine, New Haven, Connecticut, United States of America, **2** Quantitative Biology Center, Eberhard-Karls University of Tübingen, Tübingen, Germany, **3** Yale Center for Research Computing, New Haven, Connecticut, United States of America, **4** Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach, Germany, **5** Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **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, Connecticut, United States of America

☯ These authors contributed equally to this work.
‡ These authors are joint senior authors on this work.
¶ Membership of 'nf-core community' is provided in the Acknowledgements.
* gisela.gabernet@yale.edu

OPEN ACCESS

Citation: Gabernet G, Marquez S, Bjornson R, Peltzer A, Meng H, Aron E, et al. (2024) nf-core/airrflow: An adaptive immune receptor repertoire analysis workflow employing the Immcantation framework. PLoS Comput Biol 20(7): e1012265. https://doi.org/10.1371/journal.pcbi.1012265

Editor: Anna Niarakis, University of Toulouse III Paul Sabatier, Center of Integrative Biology & INRIA Saclay, FRANCE

Received: January 28, 2024

Accepted: June 20, 2024

Published: July 26, 2024

Data Availability Statement: nf-core/airrflow is available free of charge under the MIT license on 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 (fastq.gz) were uploaded to

Abstract

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. 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 with sequencing errors 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, including a mixture of bulk and single-cell sequencing datasets.

Zenodo (https://doi.org/10.5281/zenodo.10989592).

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. SN and MP were funded under the German National Research Infrastructure for Immunology (NFDI4Immuno) [NFDI 49/1 - 501875662] by the DFG. This work was additionally funded by the NIH grant R01AI104739 to SHK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: SHK receives consulting fees from Peraton. AP is an employee of Boehringer Ingelheim Pharma GmbH & Co KG and declares no conflict of interest. DL is an employee of oNKo-innate Pty Ltd 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.

Using this dataset, we extended the convergence findings to 20 additional subjects, highlighting the applicability of nf-core/airrflow to validate findings in small in-house cohorts with reanalysis of large publicly available AIRR datasets.

Author summary

We have created nf-core/airrflow, a workflow to help researchers study the immune system in healthy and disease states, such as infections, autoimmunity, and cancer. The adaptive immune system is responsible for the third line of defense responses, specific to each particular threat, after physical barriers have been compromised and the nonspecific innate immune response has failed to clear the danger. Two types of white blood cells are central players in the adaptive response, namely B cells and T cells. These cells have surface receptors that recognize suspicious elements (antigens). Learning what receptors bind to which antigens is of utmost interest to understand immune responses. Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) is a technique that allows to determine the genetic sequence of these receptors. The amount of data generated in these experiments is large, and the analysis complex. nf-core/airrflow simplifies running a comprehensive analysis connecting tools from Immcantation, a specialized software project to analyze AIRR-seq data. The workflow can efficiently process large datasets on multiple computing platforms. We analyzed immune responses to COVID-19 in 97 infected individuals and 99 healthy people, and confirmed previous findings and provided new insights, demonstrating the workflow's applicability to reanalyzing large publicly available datasets.

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 [1,2]. Additional nucleotide deletion/insertion at the gene boundaries generate practically unique receptors for each maturing cell [3,4]. 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 [5]. 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 [6,7], infectious diseases [8,9], (auto) immune diseases [10–13], and cancer [14].

Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) enables the recovery of the collection of AIRRs sampled from an individual [15]. 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 [16]. The processing of AIRR-seq data requires steps that are distinct from other bioinformatics pipelines [17–20], which has led to the development of a multitude of tools and frameworks [11,21–35]. 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 [36] or Snakemake [37] 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, while 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 [36]. Nextflow has been used in a pipeline for reconstructing T-cell receptor repertoires from RNA sequencing data [38].

To address the need for a high-throughput, portable analysis workflow that incorporates the necessary steps for AIRR-seq data analysis we developed nf-core/airrflow, a workflow written in the Nextflow language. The workflow utilizes the Immcantation framework, a comprehensive collection of open-source software to process AIRR-seq data which has been applied for the last 10 years to analyze diverse datasets investigating the immune responses in autoimmunity [11,12,39,40], allergy [41], viral infections [42,43] and vaccinations [44–46]. Immcantation covers AIRR-seq analysis from start to finish, including sequence assembly and error correction [21], alignment to the international ImMunoGeneTics information system (IMGT) [47,48] BCR and TCR reference data with IgBLAST [49], clonal relationship inference [22,23,27,50], reconstruction of clonal lineages [25,26], and the identification of predictive repertoire properties and sequence motifs [11,51]. A comparison of nf-core/airrflow to other available BCR and TCR AIRR-seq data processing pipelines is detailed in Table A in S1 Text. nf-core/airrflow is a flexible workflow that supports the analysis of both bulk and single-cell sequencing AIRR-seq 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 [52], which defines Nextflow implementation best practices including containerization of all software tools with biocontainers [53] whenever possible, reporting of all software versions in a MultiQC [54] report, 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.*[55] in a larger dataset comprising BCR repertoire samples from 97 individuals diagnosed with COVID-19 and 99 healthy controls retrieved from the from the AIRR Data Commons [56] by querying the iReceptor Gateway [57]. Thus, we showcased the applicability of nf-core/airrflow to validate findings in small patient cohorts with data from publicly available sources.

Design and implementation

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 [11,21–27]. 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 (Fig 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 [58]. The different options can be specified through command line parameters when executing a pipeline or a Nextflow parameters file that is provided at runtime.

Input data. The AIRR-seq data processing with nf-core/airrflow can depart from fastq files or assembled sequences in AIRR rearrangement [58] or fasta format. The path to the input files is specified within a metadata file following the AIRR Study Reporting standard (MiAIRR) [58], 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. For common commercially available AIRR-seq sequencing kits such as New England Biolabs [59], or TAKARA Bio BCR and TCR profiling [60]. However, it is possible to specify in a flexible manner the start position of the primer(s) and the UMI configuration, 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, 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 [61]. 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 [47] 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.

Read QC and sequence assembly. Sequencing read quality control is performed using fastp [62] for general read quality statistics and adapter trimming, if indicated. For bulk sequencing data, the pRESTO [21] 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

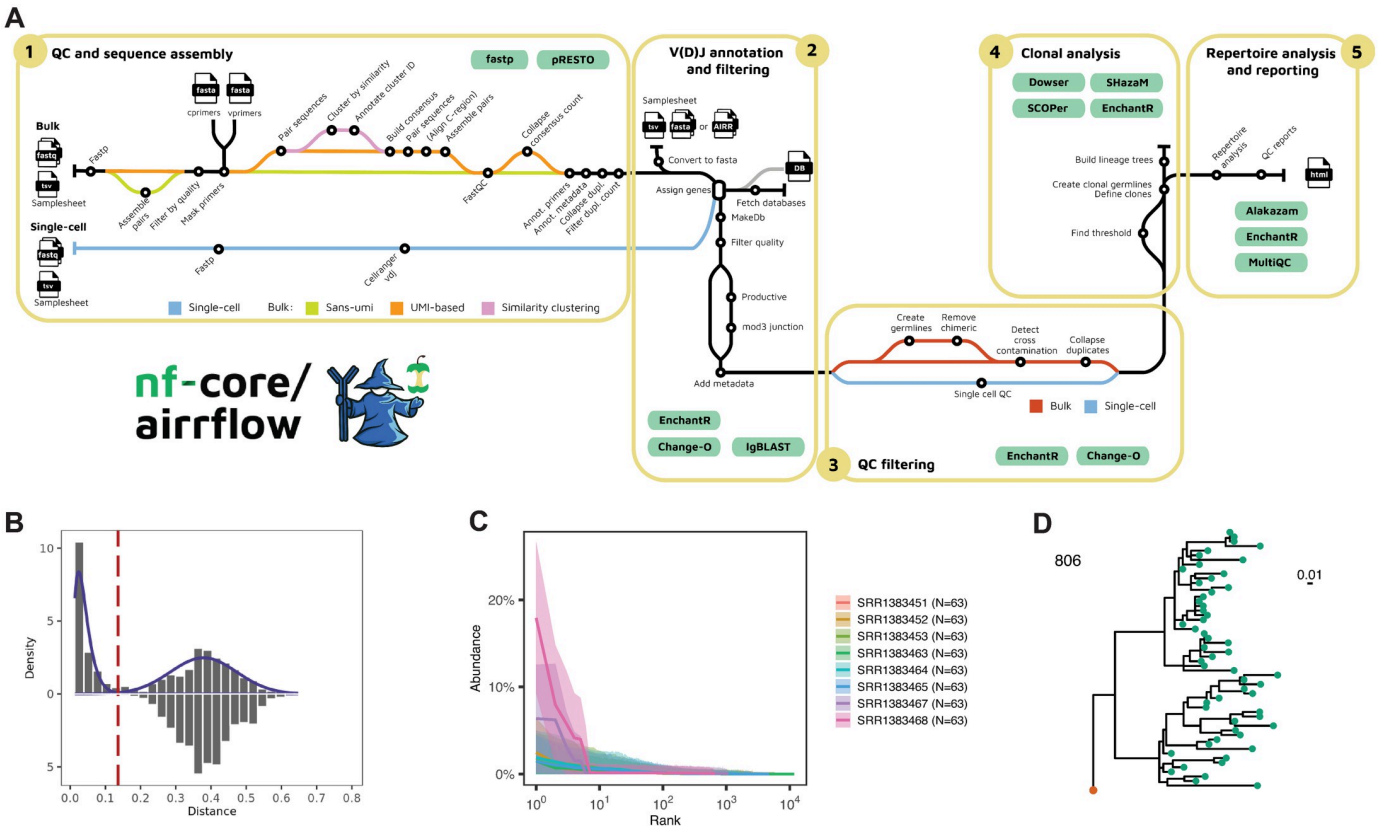

Fig 1. Schematic representation of the nf-core/airrflow workflow processes and detailed analysis steps. A. nf-core/airrflow v4.0 workflow overview. QC and sequence assembly steps (1) are performed on the bulk raw sequencing reads. The workflow supports several sequencing protocols including multiplex 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 the publication of Stern et al. [11] Examples of (B) a hamming distance threshold plot, (C) a clonal abundance plot, and (D) a reconstructed lineage tree from this dataset.

https://doi.org/10.1371/journal.pcbi.1012265.g001

dealing with insufficient UMI diversity, which might occur when UMIs are too short for the library diversity, and certain types of sequencing errors in the UMI region by clustering the sequences within a UMI group by similarity and dividing them into sub-clusters if they belong to dissimilar sequences of origin. Read mates are paired and assembled, and the primer information along with any metadata relevant to the downstream analysis are annotated in the fastq header with a format standardized by pRESTO. Duplicate sequences are collapsed, annotating the number of representative reads where this sequence was observed. Finally, sequences with insufficient support, less than two representatives by default, are filtered out. For single-cell sequencing data, the 10X Genomics *cellranger vdj* tool is used for sequence assembly.

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 [48,49], and processing the alignment results with the Change-O [22] 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.

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 [23,24] 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 [63,64] Immcantation package. As T cells do not undergo somatic hypermutation, T cell clones are defined as having identical TCR nucleotide 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 (Fig 1B). Clonal repertoire properties such as clonal abundance and diversity are calculated with the Alakazam Immcantation package (Fig 1C). Clonal lineage trees can be optionally reconstructed with Dowser [65], also part of Immcantation. Dowser supports several methods for lineage tree reconstruction including IgPhyML [25], a maximum likelihood method specifically designed for B-cell lineage tree reconstruction, other maximum likelihood methods such as RAxML [66] (Fig 1D), and maximum parsimony methods such as the pratchet method implemented within phangorn [67].

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 in Rmarkdown format is generated comprising the number of sequences obtained after each workflow step, which can be downloaded by the user and modified accordingly to meet the specific project analysis needs. Additionally, a MultiQC [54] 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.

Parallelization and portability to alternative computing infrastructures. nf-core/airrflow achieves parallelization across samples, scalability and portability through using Nextflow

[36], a Domain Specific Language for parallel and scalable computational workflows, as well as through providing containers for each analysis step to ensure portability and reproducibility across computing infrastructures. We have tested nf-core/airrflow portability to various computing infrastructure by running the pipeline on full-size AIRR-seq datasets from the publication of Stern et al. [11] (Fig 1C and 1D). The commands to launch nf-core/airrflow on a desktop computer, a high-performance computing cluster and on AWS batch are provided on S1 Text.

Results

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 [68]. 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 [69] (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 S1 Text and Fig A in S1 Text for further details on the repertoire simulation). We then simulated paired-end raw sequencing data for each repertoire with Grinder [70]. 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 [71,72]. 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 (Fig 2 and B in S1 Text). 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 (Fig 2A and 2B) 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 (Fig 2C). 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

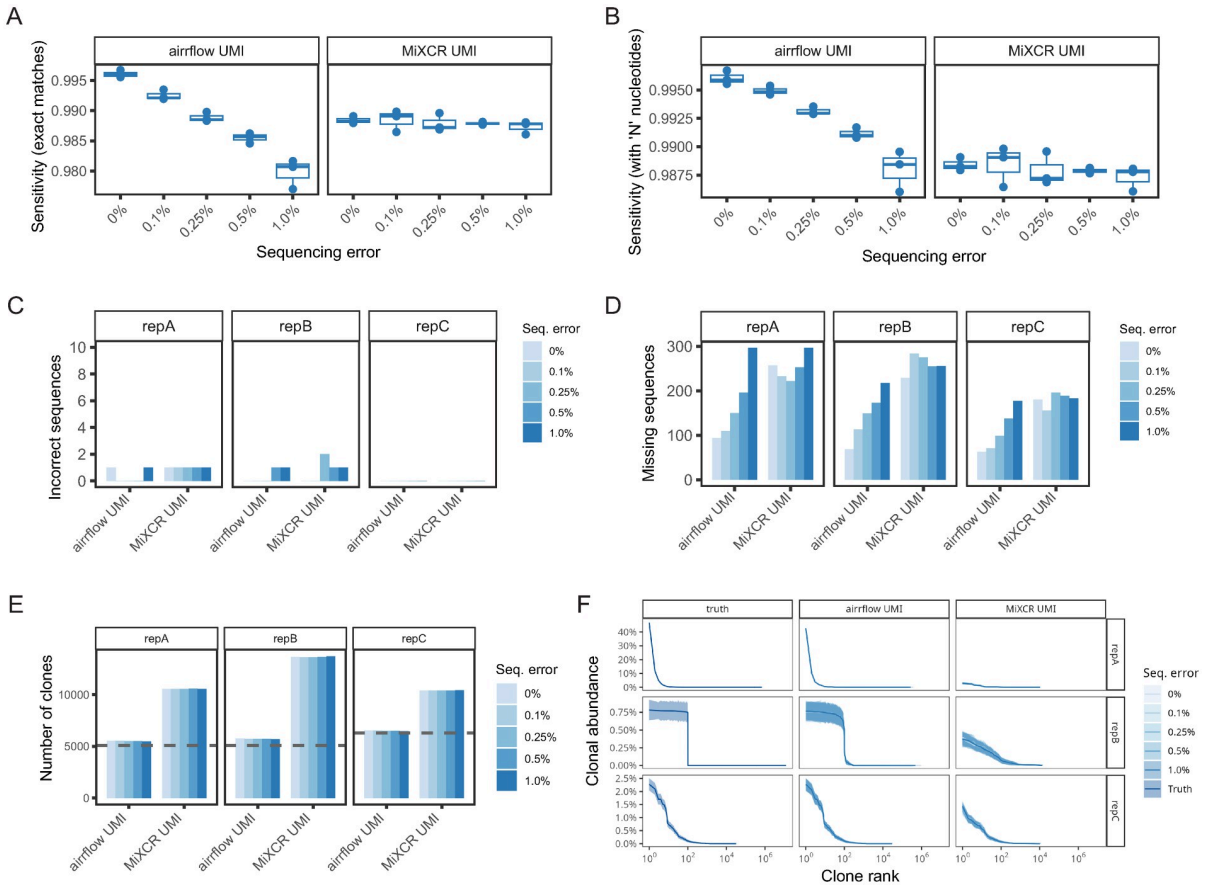

Fig 2. Performance assessment of the nf-core/airrflow pipeline on three simulated BCR repertoires compared to MiXCR. Sensitivity of the nf-core/airrflow and MiXCR pipelines on the data simulated with UMIs. Sensitivity was calculated for exact VDJ sequence matches to the truth repertoires (**A**), and 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. The x axis shows the clone rank number when ordering the clones by size from bigger to smaller.

https://doi.org/10.1371/journal.pcbi.1012265.g002

number of missing sequences was comparable to MiXCR [28], an alternative pipeline for bulk and single-cell AIRR-seq data analysis (Fig 2A–2D and Table A in S1 Text). Differences in the number of correctly recovered sequences can be explained by the differential sequencing error correction and UMI error correction methods used by nf-core/airrflow and MiXCR. nf-core/airrflow performs UMI error correction with pRESTO *B*uildConsensus and *C*lusterSets, which allows the identification of UMI groups with a dissimilar sequence of origin [21]. On the other hand, MiXCR performs UMI error correction by clustering UMI sequences and assigning small UMI groups to the closest larger UMI group [28]. 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 (Fig C in S1 Text), reinforcing the importance of utilizing protocols that include UMIs for sequencing 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 (Fig C in S1 Text).

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 to assess whether a sequence comes from an expanded clone of the same progenitor cell. This step is particularly relevant 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 (Fig 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 (Fig 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 (Fig 2E and 2F). While we used the default parameters for performing clonal inference with the MiXCR pipeline, adjusting these parameters could potentially lead to improved clonal inference results. 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 (Fig B in S1 Text). Regarding runtime, nf-core/airrflow took 2h 2min to process the UMI benchmarking data and 1h 47min to process the sans-umi benchmarking data, with a sample average time of 8min and 7 min, respectively (Table C in S1 Text). When including the lineage tree reconstruction step, which is skipped by default, users should be aware that this is often the most time consuming step and that the method chosen (maximum likelihood vs maximum parsimony), and the presence of large clones in the repertoire, as it is the case in repA, can greatly influence the runtime (Table C in S1 Text).

Case study: Validating convergent antibody responses from public repositories

Previous studies have identified convergent antibody responses to the SARS-CoV-2 virus [8,55,73–75]. 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.* [55] 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 AIRR Data Commons [56] by querying the iReceptor Gateway [57], 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.* [55] (Table D in S1 Text). We processed the repertoires with nf-core/airrflow on an HPC cluster and show that parallelization over several cluster nodes decreases the wall-clock runtime (Fig 3B). nf-core/airrflow was used to perform VDJ gene reassignment, quality control and filtering, clonal inference, and repertoire comparison (steps 2–5 in Fig 1A). This ensures that all sequences retrieved from the various projects in the AIRR Data Commons are aligned to the same germline reference sets with the same software and versions, applied through the same QC criteria, and that the B cell clones are inferred with the same methods and clonal

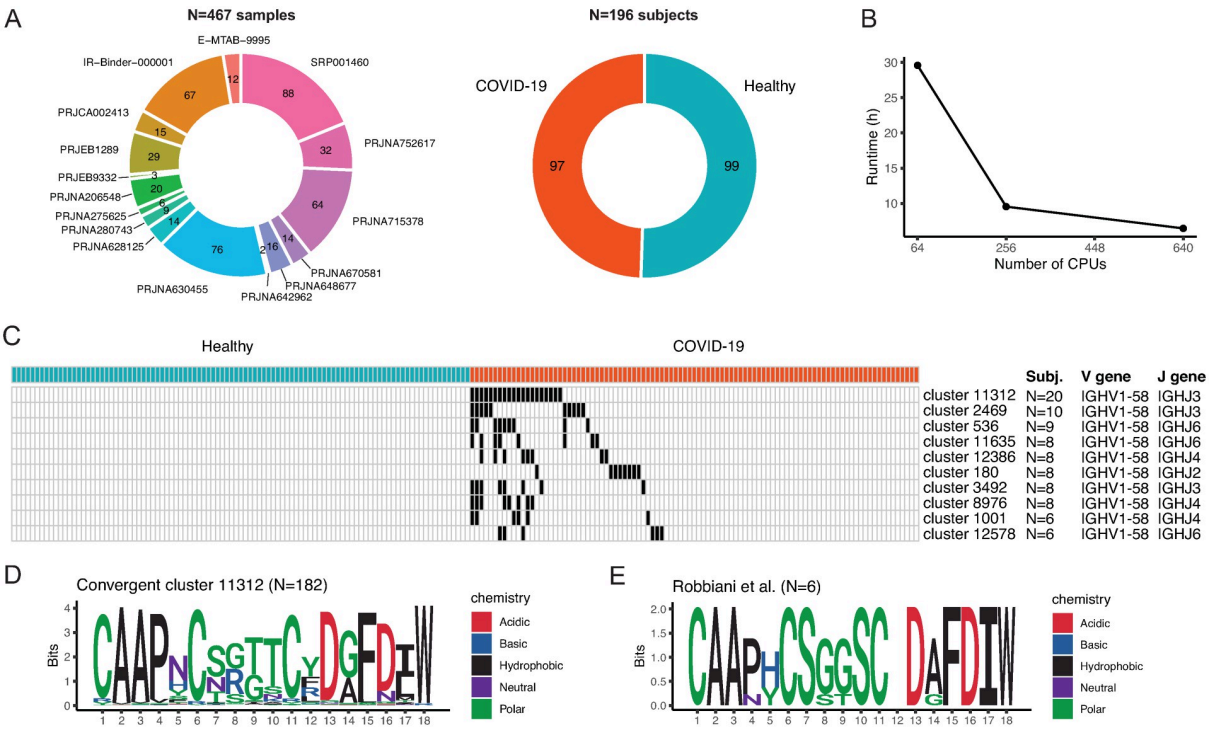

Fig 3. Application of nf-core/airrflow to process repertoires of COVID-19 infected or vaccinated participants together with healthy controls. A. Number of samples that passed the quality control after processing the data with nf-core/airrflow v4.0 and their distribution across projects. Number of subjects in each group (healthy or COVID-19 diagnosed). **B.** Total nf-core/airrflow wall-clock time when processing the COVID datasets on one cluster node (64CPUs), 4 cluster nodes (256 CPUs) and 10 cluster nodes (640 CPUs). **C**. Top 10 antibody convergent clusters (rows) using the IGHV1-58 gene 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 subjects are ordered from higher to lower frequency of sequences in the top ranking cluster, successively until the lowest ranking cluster. 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.* [55], positively tested against the SARS-CoV-2 spike protein receptor binding domain. Amino acids at position 12 are not conserved and thus have 0 information bits.

https://doi.org/10.1371/journal.pcbi.1012265.g003

thresholds. 467 antibody repertoires from 196 subjects (99 healthy controls and 97 COVID-19 diagnosed) passed the QC criteria (Fig C in S1 Text).

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 (Fig D and Table E in S1 Text). 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.* [55]. 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 (Fig 3D and 3E), 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*. [55], 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. nf-core/airrflow facilitates this reanalysis by enabling a highly parallelized, reproducible computation with controlled software versions that is triggered by a single command and can be effortlessly ported to high-performance computing and cloud infrastructure.

Availability and future directions

We implemented nf-core/airrflow, a high-throughput, easy installation and portable workflow for analyzing bulk and single-cell AIRR sequencing data. nf-core/airrflow outputs the results as an AIRR rearrangement table, and was certified as AIRR compliant by the AIRR community (http://airr-community.org), ensuring its interoperability with other AIRR-seq analysis tools following the community standards [58]. These include machine learning frameworks such as ImmuneML [76], repertoire comparison tools such as CompAIRR [77], or tools for downstream annotation of sequences according to their possible targets. The workflow 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 at (https://bitbucket.org/kleinstein/projects).

We plan to further develop the workflow to incorporate additional analysis needs of the users. nf-core/airrflow current and future users are encouraged to join the nf-core community dedicated slack channel (https://nf-co.re/join) for questions or feature requests. 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.

Supporting information

S1 Text. Contains the Supplementary methods, Figs A-D and Tables A-E.
(DOCX)

S1 Data. Contains the raw data for all figures.
(ZIP)

Acknowledgments

The authors thank the nf-core community for the contributions to the nf-core framework that facilitated developing the nf-core/airrflow workflow. A full list of contributors to the nf-core community can be found at https://nf-co.re/contributors.

Author Contributions

Conceptualization: Gisela Gabernet, Susanna Marquez.

Formal analysis: Gisela Gabernet, Susanna Marquez.

Funding acquisition: Gisela Gabernet, Sven Nahnsen, Steven H. Kleinstein.

Methodology: Gisela Gabernet, Susanna Marquez, Robert Bjornson, Steven H. Kleinstein.

Software: Gisela Gabernet, Susanna Marquez, Alexander Peltzer, Hailong Meng, Edel Aron, Noah Y. Lee, Cole G. Jensen, David Ladd, Mark Polster, Friederike Hanssen, Simon Heumos.

Supervision: Gur Yaari, Markus C. Kowarik, Sven Nahnsen, Steven H. Kleinstein.

Writing – original draft: Gisela Gabernet, Susanna Marquez, Sven Nahnsen, Steven H. Kleinstein.

Writing – review & editing: Gisela Gabernet, Susanna Marquez, Robert Bjornson, Alexander Peltzer, Hailong Meng, Edel Aron, Noah Y. Lee, Cole G. Jensen, David Ladd, Mark Polster, Friederike Hanssen, Simon Heumos, Gur Yaari, Markus C. Kowarik, Sven Nahnsen, Steven H. Kleinstein.

References

1. Roth DB. V(D)J Recombination: Mechanism, Errors, and Fidelity. Microbiol Spectr. 2014 Nov 21; 2(6): https://doi.org/10.1128/microbiolspec.MDNA3-0041-2014 PMID: 26104458
2. Tonegawa S. Somatic generation of antibody diversity. Nature. 1983 Apr; 302(5909):575–81. https://doi.org/10.1038/302575a0 PMID: 6300689
3. 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 Jul; 79(13):4118–22. https://doi.org/10.1073/pnas.79.13.4118 PMID: 6287467
4. Lafaille JJ, DeCloux A, Bonneville M, Takagaki Y, Tonegawa S. 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 Dec 1; 59(5):859–70. https://doi.org/10.1016/0092-8674(89)90609-0 PMID: 2590942
5. Papavasiliou FN, Schatz DG. Somatic Hypermutation of Immunoglobulin Genes: Merging Mechanisms for Genetic Diversity. Cell. 2002 Apr 19; 109(2):S35–44. https://doi.org/10.1016/s0092-8674(02)00706-7 PMID: 11983151
6. Safonova Y, Shin SB, Kramer L, Reecy J, Watson CT, Smith TPL, et al. Variations in antibody repertoires correlate with vaccine responses. Genome Res. 2022 Apr; 32(4):791–804. https://doi.org/10.1101/gr.276027.121 PMID: 35361626
7. Kotagiri P, Mescia F, Rae WM, Bergamaschi L, Tuong ZK, Turner L, et al. B cell receptor repertoire kinetics after SARS-CoV-2 infection and vaccination. Cell Rep. 2022 Feb 15; 38(7):110393. https://doi.org/10.1016/j.celrep.2022.110393 PMID: 35143756
8. Chen EC, Gilchuk P, Zost SJ, Suryadevara N, Winkler ES, Cabel CR, et al. Convergent antibody responses to the SARS-CoV-2 spike protein in convalescent and vaccinated individuals. Cell Rep. 2021 Aug 24; 36(8):109604. https://doi.org/10.1016/j.celrep.2021.109604 PMID: 34411541
9. Parameswaran P, Liu Y, Roskin KM, Jackson KKL, Dixit VP, Lee JY, et al. Convergent Antibody Signatures in Human Dengue. Cell Host Microbe. 2013 Jun 12; 13(6):691–700. https://doi.org/10.1016/j.chom.2013.05.008 PMID: 23768493
10. Ramadoss NS, Robinson WH. Characterizing the BCR repertoire in immune-mediated diseases. Nat Rev Rheumatol. 2020 Jan; 16(1):7–8. https://doi.org/10.1038/s41584-019-0339-y PMID: 31780792
11. Stern JNH, Yaari G, Vander Heiden JA, Church G, Donahue WF, Hintzen RQ, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. Sci Transl Med. 2014 Aug 6; 6(248):248ra107. https://doi.org/10.1126/scitranslmed.3008879 PMID: 25100741
12. Vander Heiden JA, Stathopoulos P, Zhou JQ, Chen L, Gilbert TJ, Bolen CR, et al. Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing. J Immunol Baltim Md 1950. 2017 Feb 15; 198(4):1460–73. https://doi.org/10.4049/jimmunol.1601415 PMID: 28087666
13. Bashford-Rogers RJM, Palser AL, Huntly BJ, Rance R, Vassiliou GS, Follows GA, et al. Network properties derived from deep sequencing of human b-cell receptor repertoires delineate b-cell populations. Genome Res. 2013 Nov; 23(11):1874–84. https://doi.org/10.1101/gr.154815.113 PMID: 23742949

14. Frank ML, Lu K, Erdogan C, Han Y, Hu J, Wang T, et al. T-Cell Receptor Repertoire Sequencing in the Era of Cancer Immunotherapy. Clin Cancer Res. 2023 Mar 14; 29(6):994–1008. https://doi.org/10.1158/1078-0432.CCR-22-2469 PMID: 36413126
15. Boyd SD, Joshi SA. High-Throughput DNA Sequencing Analysis of Antibody Repertoires. Microbiol Spectr. 2014 Oct; 2(5). https://doi.org/10.1128/microbiolspec.AID-0017-2014 PMID: 26104353
16. Papalexi E, Satija R. Single-cell RNA sequencing to explore immune cell heterogeneity. Nat Rev Immunol. 2018 Jan; 18(1):35–45. https://doi.org/10.1038/nri.2017.76 PMID: 28787399
17. Georgiou G, Ippolito GC, Beausang J, Busse CE, Wardemann H, Quake SR. The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol. 2014 Feb 19; 32 (2):158–68. https://doi.org/10.1038/nbt.2782 PMID: 24441474
18. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. Genome Med. 2015; 7(121):1–14. https://doi.org/10.1186/s13073-015-0243-2 PMID: 26589402
19. Marquez S, Babrak L, Greiff V, Hoehn KB, Lees WD, Luning Prak ET, et al. Adaptive Immune Receptor Repertoire (AIRR) Community Guide to Repertoire Analysis. Methods Mol Biol Clifton NJ. 2022; 2453:297–316. https://doi.org/10.1007/978-1-0716-2115-8_17 PMID: 35622333
20. Mhanna V, Bashour H, Lê Quý K, Barennes P, Rawat P, Greiff V, et al. Adaptive immune receptor repertoire analysis. Nat Rev Methods Primer. 2024 Jan 25; 4(1):1–25.
21. Vander Heiden JA, Yaari G, Uduman M, Stern JNH, O'Connor KC, Hafler DA, et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics. 2014 Jul 1; 30(13):1930–2. https://doi.org/10.1093/bioinformatics/btu138 PMID: 24618469
22. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. Bioinformatics. 2015 Oct 15; 31(20):3356–8. https://doi.org/10.1093/bioinformatics/btv359 PMID: 26069265
23. Nouri N, Kleinstein SH. A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. Bioinformatics. 2018 Jul 1; 34(13):i341–9. https://doi.org/10.1093/bioinformatics/bty235 PMID: 29949968
24. Nouri N, Kleinstein SH. Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data. PLoS Comput Biol. 2020 Jun; 16(6):e1007977. https://doi.org/10.1371/journal.pcbi.1007977 PMID: 32574157
25. Hoehn KB, Pybus OG, Kleinstein SH. Phylogenetic analysis of migration, differentiation, and class switching in B cells. PLOS Comput Biol. 2022 Apr 25; 18(4):e1009885. https://doi.org/10.1371/journal.pcbi.1009885 PMID: 35468128
26. Hoehn KB, Vander Heiden JA, Zhou JQ, Lunter G, Pybus OG, Kleinstein SH. Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. Proc Natl Acad Sci. 2019 Nov 5; 116(45):22664–72. https://doi.org/10.1073/pnas.1906020116 PMID: 31636219
27. Gadala-maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. Proc Natl Acad Sci. 2015 Feb 24; 112(8):1–9. https://doi.org/10.1073/pnas.1417683112 PMID: 25675496
28. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015 May; 12(5):380–1. https://doi.org/10.1038/nmeth.3364 PMID: 25924071
29. Shugay M, Britanova OV, Merzlyak EM, Turchaninova MA, Mamedov IZ, Tuganbaev TR, et al. Towards error-free profiling of immune repertoires. Nat Methods. 2014 Jun; 11(6):653–5. https://doi.org/10.1038/nmeth.2960 PMID: 24793455
30. Shlemov A, Bankevich S, Bzikadze A, Turchaninova MA, Safonova Y, Pevzner PA. Reconstructing Antibody Repertoires from Error-Prone Immunosequencing Reads. J Immunol Baltim Md 1950. 2017 Nov 1; 199(9):3369–80. https://doi.org/10.4049/jimmunol.1700485 PMID: 28978691
31. Cortina-Ceballos B, Godoy-Lozano EE, Sámano-Sánchez H, Aguilar-Salgado A, Velasco-Herrera MDC, Vargas-Chávez C, et al. Reconstructing and mining the B cell repertoire with ImmunediveRsity. mAbs. 2015; 7(3):516–24. https://doi.org/10.1080/19420862.2015.1026502 PMID: 25875140
32. Sturm G, Szabo T, Fotakis G, Haider M, Rieder D, Trajanoski Z, et al. Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. Bioinformatics. 2020 Sep 15; 36(18):4817–8. https://doi.org/10.1093/bioinformatics/btaa611 PMID: 32614448
33. Suo C, Polanski K, Dann E, Lindeboom RGH, Vilarrasa-Blasi R, Vento-Tormo R, et al. Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins. Nat Biotechnol. 2023 Apr 13;1–12.

34. Lindeman I, Emerton G, Mamanova L, Snir O, Polanski K, Qiao SW, et al. BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. Nat Methods. 2018 Aug; 15(8):563–5. https://doi.org/10.1038/s41592-018-0082-3 PMID: 30065371

35. Kuchenbecker L, Nienen M, Hecht J, Neumann AU, Babel N, Reinert K, et al. IMSEQ—a fast and error aware approach to immunogenetic sequence analysis. Bioinforma Oxf Engl. 2015 Sep 15; 31 (18):2963–71. https://doi.org/10.1093/bioinformatics/btv309 PMID: 25987567

36. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017 Apr 11; 35(4):316–9. https://doi.org/10.1038/nbt.3820 PMID: 28398311

37. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research; 2021, 10:33. https://doi.org/10.12688/f1000research.29032.2 PMID: 34035898

38. Rubio T, Chernigovskaya M, Marquez S, Marti C, Izquierdo-Altarejos P, Urios A, et al. A Nextflow pipeline for T-cell receptor repertoire reconstruction and analysis from RNA sequencing data. ImmunoInformatics. 2022 Jun 1; 6. https://doi.org/10.1016/j.immuno.2022.100012

39. Jiang R, Fichtner ML, Hoehn KB, Pham MC, Stathopoulos P, Nowak RJ, et al. Single-cell repertoire tracing identifies rituximab-resistant B cells during myasthenia gravis relapses. JCI Insight. 2020 Jul 23; 5(14):e136471. https://doi.org/10.1172/jci.insight.136471 PMID: 32573488

40. Brioschi S, Wang WL, Peng V, Wang M, Shchukina I, Greenberg ZJ, et al. Heterogeneity of meningeal B cells reveals a lymphopoietic niche at the CNS borders. Science. 2021 Jul 23; 373(6553):eabf9277. https://doi.org/10.1126/science.abf9277 PMID: 34083450

41. Ota M, Hoehn KB, Fernandes-Braga W, Ota T, Aranda CJ, Friedman S, et al. CD23+IgG1+ memory B cells are poised to switch to pathogenic IgE production in food allergy. Science Translational Medicine. 2024 Feb 7; 16(733):eadi0673. https://doi.org/10.1126/scitranslmed.adi0673 PMID: 38324641

42. Zurbuchen Y, Michler J, Taeschler P, Adamo S, Cervia C, Raeber ME, et al. Human memory B cells show plasticity and adopt multiple fates upon recall response to SARS-CoV-2. Nat Immunol. 2023 Jun; 24(6):955–65. https://doi.org/10.1038/s41590-023-01497-y PMID: 37106039

43. Safra M, Tamari Z, Polak P, Shiber S, Matan M, Karameh H, et al. Altered somatic hypermutation patterns in COVID-19 patients classifies disease severity. Front Immunol. 2023; 14:1031914. https://doi.org/10.3389/fimmu.2023.1031914 PMID: 37153628

44. Turner JS, Zhou JQ, Han J, Schmitz AJ, Rizk AA, Alsoussi WB, et al. Human germinal centres engage memory and naive B cells after influenza vaccination. Nature. 2020 Oct; 586(7827):127–32. https://doi.org/10.1038/s41586-020-2711-0 PMID: 32866963

45. Wang Z, Muecksch F, Raspe R, Johannsen F, Turroja M, Canis M, et al. Memory B cell development elicited by mRNA booster vaccinations in the elderly. Journal of Experimental Medicine. 2023 Jun 27; 220(9):e20230668. https://doi.org/10.1084/jem.20230668 PMID: 37368240

46. Wang M, Jiang R, Mohanty S, Meng H, Shaw AC, Kleinstein SH. High-throughput single-cell profiling of B cell responses following inactivated influenza vaccination in young and older adults. Aging. 2023 Jun 26; 15. https://doi.org/10.18632/aging.204778 PMID: 37367734

47. Lefranc MP, Giudicelli V, Duroux P, Jabado-Michaloud J, Folch G, Aouinti S, et al. IMGT, the international ImMunoGeneTics information system 25 years on. Nucleic Acids Res. 2015 Jan; 43(Database issue):D413–422.

48. Collins AM, Ohlin M, Corcoran M, Heather JM, Ralph D, Law M, et al. AIRR-C IG Reference Sets: curated sets of immunoglobulin heavy and light chain germline genes. Front Immunol. 2024 Feb 9; 14. https://doi.org/10.3389/fimmu.2023.1330153 PMID: 38406579

49. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013 Jul; 41(Web Server issue):W34–40. https://doi.org/10.1093/nar/gkt382 PMID: 23671333

50. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH. Hierarchical Clustering Can Identify B Cell Clones with High Confidence in Ig Repertoire Sequencing Data. J Immunol. 2017 Mar 15; 198(6):2489–99. https://doi.org/10.4049/jimmunol.1601850 PMID: 28179494

51. Olson BJ, Moghimi P, Schramm CA, Obraztsova A, Ralph D, Vander Heiden JA, et al. sumrep: A Summary Statistic Framework for Immune Receptor Repertoire Comparison and Model Validation. Front Immunol. 2019 Nov 1; 10:2533. https://doi.org/10.3389/fimmu.2019.02533 PMID: 31736960

52. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020 Mar 1; 38(3):276–8. https://doi.org/10.1038/s41587-020-0439-x PMID: 32055031

53. da Veiga Leprevost F, Grüning BA, Alves Aflitos S, Röst HL, Uszkoreit J, Barsnes H, et al. BioContainers: an open-source and community-driven framework for software standardization. Bioinformatics. 2017 Aug 15; 33(16):2580–2. https://doi.org/10.1093/bioinformatics/btx192 PMID: 28379341

54. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinforma Oxf Engl. 2016 Oct 1; 32(19):3047–8.

55. Robbiani DF, Gaebler C, Muecksch F, Lorenzi JCC, Wang Z, Cho A, et al. Convergent antibody responses to SARS-CoV-2 in convalescent individuals. Nature. 2020 Aug; 584(7821):437–42. https://doi.org/10.1038/s41586-020-2456-9 PMID: 32555388

56. Christley S, Aguiar A, Blanck G, Breden F, Bukhari SAC, Busse CE, et al. The ADC API: A Web API for the Programmatic Query of the AIRR Data Commons. Front Big Data. 2020; 3:22. https://doi.org/10.3389/fdata.2020.00022 PMID: 33693395

57. Corrie BD, Marthandan N, Zimonja B, Jaglale J, Zhou Y, Barr E, et al. iReceptor: a platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. Immunol Rev. 2018 Jul; 284(1):24–41. https://doi.org/10.1111/imr.12666 PMID: 29944754

58. Vander Heiden JA, Marquez S, Marthandan N, Bukhari SAC, Busse CE, Corrie B, et al. AIRR Community Standardized Representations for Annotated Immune Repertoires. Front Immunol. 2018; 9. https://doi.org/10.3389/fimmu.2018.02206 PMID: 30323809

59. New England Biolabs. NEBNext Immune Sequencing Kit. https://www.neb.com/en-us/products/e6320-nebnext-immune-sequencing-kit-human#Product%20Information

60. 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

61. Song L, Cohen D, Ouyang Z, Cao Y, Hu X, Liu XS. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. Nat Methods. 2021 Jun; 18(6):627–30. https://doi.org/10.1038/s41592-021-01142-2 PMID: 33986545

62. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1; 34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560 PMID: 30423086

63. Yaari G, Vander Heiden JA, Uduman M, Gadala-Maria D, Gupta N, Stern JNH, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. Front Immunol. 2013; 4:358. https://doi.org/10.3389/fimmu.2013.00358 PMID: 24298272

64. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. Nucleic Acids Res. 2012 Sep 1; 40(17):e134. https://doi.org/10.1093/nar/gks457 PMID: 22641856

65. Hoehn KB, Turner JS, Miller FI, Jiang R, Pybus OG, Ellebedy AH, et al. Human B cell lineages associated with germinal centers following influenza vaccination are measurably evolving. eLife. 2021 Nov 17; 10:e70873. https://doi.org/10.7554/eLife.70873 PMID: 34787567

66. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1; 30(9):1312–3. https://doi.org/10.1093/bioinformatics/btu033 PMID: 24451623

67. Schliep KP. phangorn: phylogenetic analysis in R. Bioinformatics. 2011 Feb 15; 27(4):592–3.

68. Weber CR, Akbar R, Yermanos A, Pavlović M, Snapkov I, Sandve GK, et al. immuneSIM: tunable multi-feature simulation of B- and T-cell receptor repertoires for immunoinformatics benchmarking. Bioinformatics. 2020 Jun 1; 36(11):3594–6. https://doi.org/10.1093/bioinformatics/btaa158 PMID: 32154832

69. Ruschil C, Gabernet G, Kemmerer CL, Jarboui MA, Klose F, Poli S, et al. Cladribine treatment specifically affects peripheral blood memory B cell clones and clonal expansion in multiple sclerosis patients. Front Immunol. 2023; 14:1133967. https://doi.org/10.3389/fimmu.2023.1133967 PMID: 36960053

70. Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic Acids Res. 2012 Jul; 40(12):e94. https://doi.org/10.1093/nar/gks251 PMID: 22434876

71. Schirmer M, D'Amore R, Ijaz UZ, Hall N, Quince C. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. BMC Bioinformatics. 2016 Mar 11; 17(1):125. https://doi.org/10.1186/s12859-016-0976-y PMID: 26968756

72. Stoler N, Nekrutenko A. Sequencing error profiles of Illumina sequencing instruments. NAR Genomics Bioinforma. 2021 Mar 1; 3(1):lqab019. https://doi.org/10.1093/nargab/lqab019 PMID: 33817639

73. Schultheiß C, Paschold L, Simnica D, Mohme M, Willscher E, von Wenserski L, 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 Aug; 53(2):442–455.e4. https://doi.org/10.1016/j.immuni.2020.06.024 PMID: 32668194

74. Nielsen SCA, Yang F, Jackson KJL, Hoh RA, Röltgen K, Jean GH, et al. Human B Cell Clonal Expansion and Convergent Antibody Responses to SARS-CoV-2. Cell Host Microbe. 2020 Oct 7; 28(4):516–525.e5. https://doi.org/10.1016/j.chom.2020.09.002 PMID: 32941787

75. Wen W, Su W, Tang H, Le W, Zhang X, Zheng Y, et al. Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. Cell Discov. 2020 May 4; 6:31. https://doi.org/10.1038/s41421-020-0168-9 PMID: 32377375

76. Pavlović M, Scheffer L, Motwani K, Kanduri C, Kompova R, Vazov N, et al. The immuneML ecosystem for machine learning analysis of adaptive immune receptor repertoires. Nat Mach Intell. 2021 Nov; 3 (11):936–44. https://doi.org/10.1038/s42256-021-00413-z PMID: 37396030

77. Rognes T, Scheffer L, Greiff V, Sandve GK. CompAIRR: ultra-fast comparison of adaptive immune receptor repertoires by exact and approximate sequence matching. Bioinforma Oxf Engl. 2022 Sep 2; 38(17):4230–2. https://doi.org/10.1093/bioinformatics/btac505 PMID: 35852318