NAR Genomics and Bioinformatics, 2024, **6**, lqae137
https://doi.org/10.1093/nargab/lqae137
Advance access publication date: 15 October 2024
Pipelines and Workflows for Biological Data Analysis

OXFORD

bcRflow: a Nextflow pipeline for characterizing B cell receptor repertoires from non-targeted transcriptomic data

Brent T. Schlegel [1], Michael Morikone [1], Fangping Mu[2], Wan-Yee Tang[3], Gary Kohanbash[4,*] and Dhivyaa Rajasundaram[1,*]

[1]Department of Pediatrics, Division of Health Informatics, University of Pittsburgh School of Medicine, UPMC Children's Hospital of Pittsburgh, John G. Rangos Sr. Research Center, 4401 Penn Avenue, Pittsburgh, PA 15224, USA
[2]Center for Research Computing, University of Pittsburgh, 312 Schenley Place, 4420 Bayard Street, Pittsburgh, PA 15260, USA
[3]Department of Environmental and Occupational Health, University of Pittsburgh, School of Public Health, 130 DeSoto Street, Pittsburgh, PA 15261, USA
[4]Department of Neurological Surgery, University of Pittsburgh School of Medicine, UPMC Children's Hospital of Pittsburgh, John G. Rangos Sr. Research Center, 4401 Penn Avenue, Pittsburgh, PA 15224, USA

[*]To whom correspondence should be addressed. Tel: +1 412 692 8120; Email: dhr11@pitt.edu
Correspondence may also be addressed to Gary Kohanbash. Tel: +1 412 623 1008; Email: gary.kohanbash@pitt.edu

Abstract

B cells play a critical role in the adaptive recognition of foreign antigens through diverse receptor generation. While targeted immune sequencing methods are commonly used to profile B cell receptors (BCRs), they have limitations in cost and tissue availability. Analyzing B cell receptor profiling from non-targeted transcriptomics data is a promising alternative, but a systematic pipeline integrating tools for accurate immune repertoire extraction is lacking. Here, we present bcRflow, a Nextflow pipeline designed to characterize BCR repertoires from non-targeted transcriptomics data, with functional modules for alignment, processing, and visualization. bcRflow is a comprehensive, reproducible, and scalable pipeline that can run on high-performance computing clusters, cloud-based computing resources like Amazon Web Services (AWS), the Open OnDemand framework, or even local desktops. bcRflow utilizes institutional configurations provided by nf-core to ensure maximum portability and accessibility. To demonstrate the functionality of the bcRflow pipeline, we analyzed a public dataset of bulk transcriptomic samples from COVID-19 patients and healthy controls. We have shown that bcRflow streamlines the analysis of BCR repertoires from non-targeted transcriptomics data, providing valuable insights into the B cell immune response for biological and clinical research. bcRflow is available at https://github.com/Bioinformatics-Core-at-Childrens/bcRflow.

Introduction

B cells are instrumental in orchestrating the adaptive immune response, and dysregulation of B cell function contributes to the pathogenesis of many immune-mediated diseases. Each B cell clone expresses a unique antigen receptor known as the B cell receptor (BCR), or immunoglobulins (Ig). These Ig are composed of two heavy chains (IGHs), and two light chains (IGLs) (1). There is a large degree of diversity in BCRs to recognize a wide variety of antigens; this diverse range of BCRs expressed by the total B cell population of an individual is known as the BCR repertoire (2). The diversity of BCR repertoires is largely attributed to recombination in the variable (V), diversity (D) and joining (J) regions of IGH gene segments (1). BCR diversity is also driven by somatic hypermutation (SHM), where antibodies produced by B cells are further diversified to increase antigen binding affinity, as well as IGH class switching, in which deletion and recombination occurs in the constant region to generate new isotypes (1). The third complementarity-determining region of the Ig heavy chain (CDR3) also plays a critical role in antigen recognition and binding affinity (1).

Advances in targeted B cell sequencing at the bulk and single cell level enable the in-depth profiling of the BCR repertoire at an unprecedented level. Interrogation of antibody repertoires at the sequence level includes commonly used techniques which are polymerase chain reaction-based (PCR, or amplicon-based) such as Ig-seq (3), LIBRA-seq (4) and capture-based target enrichment (5). Single cell approaches include single cell immune profiling by 10X Genomics platform, and plate-based approaches such as SMART-seq (6) and SPEC-seq (7). As BCR sequencing is rapidly evolving and producing vast, highly complex datasets, a growing number of bioinformatics tools and algorithms have also been developed.

An alternative approach to BCR profiling is to use immunoglobulin transcripts present in bulk RNA-seq data. The ready availability of RNA-seq datasets has become common in both basic as well as clinical studies and can serve as functionally relevant information on immune receptor repertoires at no additional cost, by allowing further profiling of these repertoires in existing disease studies (8). Separate immune repertoire profiling could often be limited by tissue availability and increases the cost of sequencing. In addition,

Received: May 23, 2024. **Revised:** August 13, 2024. **Editorial Decision:** September 19, 2024. **Accepted:** September 23, 2024

transcriptome sequencing based BCR characterization allows for revealing novel BCR variants or gene fusions that may have not been targeted in a traditional targeted approach, thereby providing new insights into BCR diversity and evolution.

Given the ubiquity of non-targeted sequencing in translational and systems-level research, there is a notable increase in systematic computational workflows to analyze non-targeted data for T cell repertoire characterization (9). An example of a non-targeted immunoinformatic workflow is the TCR_nextflow pipeline (10), which was designed for end-to-end T-cell receptor (TCR) alignment and analysis from bulk transcriptomics data. In recent years, various studies have also focused on characterizing and understanding the BCR repertoire through targeted bulk and single-cell data. One notable example is nf-core/airrflow (11), a Nextflow pipeline developed for immune receptor repertoire analysis (both BCR and TCR). Improved understanding of the BCR repertoire has applications in cancer therapy, vaccine development, autoimmune diseases, SARS-CoV-2 infection, organ transplant, antibody therapy, and risk assessments for environmental exposure (12–14) Several existing algorithms such as MiXCR (15), BASIC (16), BraCeR (17), BALDR (18) and TRUST4 (19) are used for BCR reconstruction from bulk and single cell datasets. In addition, downstream analyses such as diversity analysis, gene usage, and clonal abundance are performed using immcantation (https://immcantation.readthedocs.io/en/stable/about.html), Platypus (20) and immunarch (https://github.com/immunomind/immunarch). Although many studies have investigated the BCR repertoire by targeted high-throughput sequencing, a gold standard in the form of a Nextflow pipeline is still missing for non-targeted data. This is due to the wide variety of library preparation protocols, study designs, and differences in downstream analyses based on the objectives of the study and the available metadata.

To this end, the bcRflow pipeline will provide a comprehensive framework for immunologists to profile BCR repertoires using bulk and single cell transcriptome sequencing datasets to inform future studies. Researchers can utilize this pipeline in complement to T cell receptor workflows or as a standalone analysis to get a more holistic picture of adaptive immune response without incurring the cost of targeted methods.

Materials and methods

Nextflow Implementation

The bcRflow pipeline was developed using base Nextflow (21) (ver. 23.04.2) with DSL 2 enabled. It was constructed using separate modules for each sub-process in the MiXCR pipeline, as well as utilizing custom R scripts for the downstream processing. To profile BCRs with bcRflow, users should provide a comma-separated sheet of file paths containing transcriptomic data and associated metadata; an example of this required input can be found through the bcRflow GitHub repository. bcRflow is then able to process samples and analyze using the user's preferred computing resource. Output from bcRflow includes intermediate files from MiXCR and final figures from downstream analyses. Intermediate outputs from MiXCR include the .tsv report, .vdjca alignments, .clna/.clns clonotypes and alignments, and the .json file for the ImMunoGeneTics (IMGT) reference (22) used. The final output includes a snapshot of the downstream R environment as an .RData file and figures in .pdf and .tiff formats. Currently, bcRflow supports the analysis of paired-end bulk RNA-seq and 10 × 5′ GEX single-cell RNA-seq data, with planned expansion for additional single-cell modalities. Thanks to the containerization utilities of Nextflow and Docker (23), users are not required to download any external packages other than Nextflow itself and either Docker or Singularity (24), depending on the user's system. All other software dependencies are made available as a Docker container, which is automatically downloaded upon execution, and all relevant databases and R scripts for downstream analyses and visualization are provided with the pipeline. Comprehensive documentation outlining pipeline input and templates for sample metadata are provided in the bcRflow GitHub repository.

VDJ segment alignment and assembly

Raw reads in the form of FASTQ files from paired-end bulk or single cell transcriptome sequencing are used as input to the MiXCR algorithm (Ver 4.6.0). Of the benchmarked methods for BCR reconstruction (25), we used the MiXCR algorithm as it allows for the imputation of germline sequences in sparse alignments of variable regions and stringent parameters for CDR3 alignment. TRUST4 was the only other BCR reconstruction option from the benchmarking study by Andreani *et al.* developed for non-targeted bulk sequencing data. However, TRUST4 lacks the capability to impute germline sequences whereas MiXCR provides various presets for sequencing modalities and features a two-stage assembly for short reads. Germline imputation by the MiXCR algorithm helps fill gaps in assembly where reads map only partially, enabling the export of full-length clonal sequences. Additionally, MiXCR provides the option to include or exclude include partial alignments in specific gene regions and, to specifically prevent partial alignments in the CDR3 region, which is critical for BCR sequence reconstruction and accurate clonal identification. To ensure precise clone identification, bcRflow opts not to impute or extend the CDR3 regions. The initial stage of the pipeline is the alignment of raw data against the reference sequences of annotated BCR gene segments from the external ImMunoGeneTics (IMGT) database (22). Two iterations of assemblePartial by MiXCR are used to assemble alignments only partially matching to the regions outside of the CDR3 sequence. Clones are then fully assembled by the VDJ region and exported based on the immune receptor chain of interest (eg: IGH for the B cell heavy chain analysis). Non-productive reads (out of frame and stop codon containing variants) are not considered for further analysis. Each unique combination of CDR3, V and J gene alignments are termed as unique BCR sequences. Germline sequences are critical to assess the degree of somatic mutations and maturity of the repertoire and are inferred for each observed sequence using the germline database from IMGT.

Basic repertoire analysis

The diversity of the BCR repertoire is mainly attributed to V and J gene recombination, which can reveal the unique patterns and quantitative features for the classification of data from different samples (1). bcRflow offers simple visualizations for preliminary comparisons of receptor composition such as isotype frequency, CDR3 length distribution (measured in amino acid residues), and sequence logos highlighting the proportion of hydrophobic amino acids in the CDR3 region.

Gene usage

All species have a collection of germline V, D and J genes to select from when generating antibodies. The detailed landscape of germline gene expression output by bcRflow is determined using the IGHV and IGHJ gene counts. We analyzed V and J gene usage by calculating the proportion of sequences assigned to the V and J gene families for each sample and represented these values as a heatmap of IGHV genes across the samples, and chord diagrams for V and J gene pairs. Additionally, we performed an odds ratio calculation comparing V gene usage against control samples, calculating summary statistics via Fisher's exact test.

Diversity metrics

Immune repertoire diversity is one of the key features to enable broad antigen recognition and can be calculated at various levels based on the diversity of the V, D and J segments, estimation of available repertoire frequency diversity and antibody lineage reconstruction. We utilized several standard metrics to profile the repertoire diversity between sample groups. Sampling depth is a drawback of non-targeted sequencing, especially in the context of diversity estimation (26). To mitigate this, MiXCR was chosen for its proven ability to accurately reconstruct immunoglobulin repertoires and its built-in error correction capabilities (9). Additionally, the bcRflow pipeline includes a 'downsample' parameter, which defaults to 'TRUE', down-sampling data to the size of the smallest sample repertoire prior to any downstream analysis, ensuring improved accuracy in diversity metric calculations.

Hill numbers

Hill numbers estimate both richness and evenness, and calculating Hill diversity at different levels of the order parameter q provides a holistic measure of diversity. As q increases, the Hill number gives increasing weight to the most abundant clones, providing a measure of dominance for hyperexpanded clones. Hill numbers are calculated as follows:

$$D_q = \left(\sum_{i=1}^{S} p_i^q \right)^{\frac{1}{1-q}}$$

where S is the total number of clones, p_i is the proportional abundance of the ith clone, and q is the order parameter.

Repertoire evenness

Evenness is defined as the distribution of the clonotypes or the relative abundance of the clonotypes, and can be calculated using the Pielou index (27):

$$Pielou = \frac{- \sum_{i=1}^{S} \frac{n_i}{N} log_2 \frac{n_i}{N}}{log_2 (S)}$$

where S is the total number of clones, n_i is the number of reads in clone i, and N is the sum of all reads in the BCR repertoire of a given sample.

Shannon diversity is also widely used in antigen receptor diversity analysis to measure uncertainty about the identity of clones in each sample (28,29), as well as characterize and analyze the entropy of information in immune repertoires. Shannon diversity is defined as:

$$H' = - \sum_{i=1}^{S} ln \; p_i^{P_i}$$

where $p_i = n_i/N$ is the proportion of individuals of the ith species, n_i is the number of individuals of the ith species, N is the total number of individuals, and S is the total number of species. The Shannon index considers both clonal richness (the number of different clones) and clonal abundance (proportionality of clones), considering how equally clones are distributed within a repertoire.

As another measure of evenness, the Gini coefficient is used to measure the heterogeneity of different clones, estimating how far the distribution of clones has extended beyond an equal distribution by calculating the normalized area between the Lorenz curve of clonal distribution and a line of perfect evenness (30). Values of the Gini coefficient range from 0 to 1, where 0 is a fully equal distribution and 1 is a fully unequal distribution, and can be calculated as follows:

$$G = \frac{2 \sum_{i=1}^{n} i y_i}{n \sum_{i=1}^{n} y_i} c - \frac{n+1}{n}$$

where y_i is the number of reads for a given clone y, and n is the total number of clones in the repertoire.

Clonal diversity, expansion and tracking

The similarity, or overlap, between the samples of the BCR repertoire is often assessed by estimating the proportion or number of clonotypes in each sample that are common to both samples and is highly sensitive to sample sizes.

Morisita–Horn similarity index

The Morisita–Horn index accounts for both the number of common clonotypes, and the distribution of the clone sizes, and is sensitive to the clone size of the dominant clonotypes. The index is defined as:

$$C_{MH} = \frac{2 \sum_{i=1}^{c} f_i g_i}{\sum_{i=1}^{c} \left(f_i^2 + g_i^2 \right)}$$

where $f_i = n_{1i} / N_1$ and $g_i = n_{2i} / N_2$. Here, n_{1i} and n_{2i} are the clone sizes of the ith clonotype (i.e. number of copies of each distinct CDR3 sequence for the BCR heavy chain) in samples 1 and 2, and N_1 and N_2 are the total number of BCRs in samples 1 and 2, respectively. The summations in the numerator and the denominator are over all c clonotypes in both samples. This index ranges between 0 and 1, with 0 and 1 representing minimal and maximal similarity, respectively.

Jaccard similarity

$$J(X, Y) = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

The Jaccard index measures similarity between sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets. Here, Jaccard similarity is calculated for each pair of samples and visualized as a matrix, with values ranging from 0 to 1, where 0 indicates no correlation and 1 indicates full correlation.

Cosine similarity

$$cos(a, b) = \frac{a * b}{(\|a\| * \|b\|)}$$

Cosine similarity utilizes the normalized dot product between two vectors, in this case BCR clonal abundance, to compare repertoires on a per-sample level (e.g. samples a and b).

The results are visualized as a matrix, showing each pairwise similarity score on a range from −1 to 1, where −1 indicates the samples are entirely distinct, 0 indicates no correlation between the repertoires, and 1 indicates that the samples are identical.

Species richness

Also referred to as alpha diversity, richness represents the total number of unique sequences in a sample but is impacted by the sampling depth. Richness can be used to compare the BCR repertoire differences within and between individuals and can be calculated using Chao1 and Abundance-based Coverage Estimators (ACE). Additional inferences regarding alpha diversity that are less sensitive to sampling depth can be calculated using the inverse of Simpson's index and the Gini–Simpson index. The first of these metrics, Chao1, is calculated as follows:

$$Chao1 = S_{obs} + \frac{n_1^2}{2n_2},$$

where S_{obs} is the observed number of species, n_1 is the number of singletons (samples with count = a), and n_2 is the number of doubletons (samples with count = 2).

Abundance-based coverage estimation (ACE) is a widely used measure of species richness in ecological studies, and is often employed in BCR and TCR studies for estimating the total clonal count based on the observed abundance of clones, particularly applicable when the sample count is less than or equal to 10, calculated by:

$$ACE = S_{abund} + \left(\frac{S_{rare}}{C_{ACE}}\right) + \left(\frac{F_1}{C_{ACE}}\right) \gamma^2\ ACE$$

and

$$C_{ACE} = 1 - \frac{F_1}{N}$$

where S_{abund} is the number of species with count greater than or equal to 10, S_{rare} is the number of species with count less than or equal to 10, F_1 is the number of singletons (species observed only once in the sample) and N is the total number of individuals in the sample. The estimation of the coefficient of variation, ${\gamma^2}_{ACE}$, is given by:

$$\gamma^2_{ACE} = max\left[\frac{S_{rare}\ \sum_{i=1}^{10} i\ (i-1)\ F_i}{C_{ACE}\ (N_{rare})\ (N_{rare}-1)} - 1, 0\right];$$

and the number of rare species, N_{rare} is calculated as:

$$N_{rare} = \sum_{i=1}^{10} iF_i; \text{ where } F_i \text{ is the number of species in with count} = i$$

Simpson's index, while not a direct measurement of species richness, is a more robust representation of the general equation underlying species richness; values of Simpson's index are comparable to those of standard richness measures but are less sensitive to sampling depth (27). Several transformations of the Simpson index are utilized as common measures of diversity, notably the inverse Simpson and Gini-Simpson indices. Simpson's index is calculated as follows, measuring the probability that two randomly selected reads come from the same clone:

$$D = \sum_{i=1}^{S} p_i^2$$

where p_i is the proportional abundance for each clone, and S is the total number of clones. The lower the value of D, the greater the diversity of the population. A more intuitive representation of this is the inverse Simpson index, a commonly presented metric where higher values correlate with higher alpha diversity:

$$Inverse - Simpson = \frac{1}{D}$$

Similarly, the Gini-Simpson index estimates the probability that two random reads stem from *different* clones (31), and is calculated as:

$$Gini - Simpson = 1 - D$$

Diversity 50 (D50)

The D50 metric refers to the number of unique CDR3 sequences that are present in the top 50% of the sequences, with a small D50 index suggestive of large dominant clones and can be used to compare the degree of clonal expansion and clonal dominance during immune response.

Class-switching recombination

One of the critical steps for antibody maturation occurs predominantly in the germinal centers and is termed as class-switch recombination (CSR), where the immunoglobulin class is changed from one isotype to another. We investigated the CSR events using the BrepPhylo package in R and the dnapars utility for constructing maximum parismony trees between clones and the germline sequence. CDR3 amino acid sequences are clustered with at least 70% similarity using Levenshtein distance, then processed by BrepPhylo to construct lineage trees. BrepPhylo then uses these trees compare clonal sequences to IMGT germline sequences and calculates CSR events. Phylogenetic trees and associated calculations are saved in the user-specified output directory under 'CSR_batchAnalysis'. Analysis was limited to clusters with 3 or more members. Phylogenetic trees and associated calculations are saved in the user-specified output directory under 'CSR_batchAnalysis'. BrepPhylo produces a graphical summary of the distribution of CSR events separated by subclasses, as well as CSR event frequency and class-switch distance from germline. IgM and IgD are co-expressed on naïve B-cells, and M-D switches are primarily due to alternative splicing rather than much rarer true CSR events (32). To reflect their co-expression, CSR events involving M and D isotypes are grouped into one category (M/D) in the BrepPhylo graphical output as utilized in previous studies (33).

Somatic hypermutation

B cells respond to infection or immunization through somatic hypermutation (SHM), a process which diversifies the antibodies they produce, and results in an increase in the antigen binding affinity of the antibodies (1). The rate of SHM allows us to estimate the repertoire mutation, and the type of mutation contributing to the emergence of high affinity antibodies, which is an asset for B cell repertoire analysis. To calculate the SHM rate for each clone, sequences were grouped by V

and J genes and clustered based on Levenshtein distance of up to 3 residues between CDR3 amino acid sequences. A maximum edit distance of 3 was used to define a unique cluster of CDR3 sequences. V and J nucleotide sequences from each cluster were then built into lineages and aligned against their respective germline sequences collected from IMGT (22) to calculate mutations following the immunarch SHM pipeline. Calculated lineages were then used to build a phylogenetic tree stemming from the germline sequence to the clonal V and J sequence alignments using the repAlignLineage and repClonalFamily functions. Finally, the number of mutations relative to the germline sequence for each clone was calculated via the repSomaticHypermutation function; the SHM rate for each clone was calculated as the number of total mutations divided by the total nucleotide sequence length, minus the length of the CDR3 region.

Convergent clustering of global CDR3 sequences

Network-based analysis of convergent immune response allows us to model each clone as a node in a global network in the case of bulk data, and a cell in terms of the single cell data. Local and global network properties of the resulting clonal clusters provide insights into the structural organization of the immune network across the sample cohorts (34). The identification of public clones shared across samples or sample groups is a valuable resource for understanding shared immune challenges as well as differential immune response (35). By profiling the sequence-related properties of the shared or divergent clusters in conjunction with results from other analyses, such as diversity estimation and somatic hypermutation load, researchers can gain comprehensive insights into shared gene usage and convergent immune responses in their groups of interest.

To create a set of convergent clusters across sample cohorts, unique CDR3aa sequences from all samples regardless of clinical severity were grouped by their specific V and J gene names and their CDR3 amino acid (CDR3aa) sequence length. Levenshtein distance was calculated between CDR3aa sequences within these groups generating adjacency matrices of sequence similarity. CDR3 amino acid sequences from these groups were then clustered, with clusters defined by at least 70% CDR3aa sequence similarity. For added flexibility, bcRflow includes a 'threshold' parameter that allows users to set a sequence similarity threshold to adjust the stringency of distance-based clustering. Large clusters of 10 or more unique sequences were subset for further analysis. Large clusters were then visualized using a bar plot ranked by cluster size, colored by sample cohort and sized by the number of contributing sequences, for identification of convergent clusters shared between sample groups.

Selected large clusters of clones from multiple sample cohorts were visualized as sequence similarity networks to further profile convergent immune response between groups. Unique sequences are modeled as nodes in the network, and edges are weighted by the Levenshtein distances between CDR3aa sequences. Undirected network diagrams were generated using the igraph R package, with node color indicating group designations and edge thickness signifying relative sequence similarity (ranging from 70–100%). As an additional representation of sequence similarity, sequence logo plots were created to visualize the probability and charge of amino acids within the clustered CDR3 sequences.

Statistical tests and visualization

Statistical analysis was performed in R (v 4.3.0). Fisher's exact test was used to assess the difference in gene usage segments between the samples. To characterize the difference between the samples, group-wise comparisons were performed using the Kruskal–Wallis test. Unless indicated otherwise, multiple hypothesis testing was corrected using the Holm–Bonferroni method, and adjusted P-values <0.05 were considered significant. Data visualization was performed using the following R packages: ggplot2 (36), BrepPhylo (https://github.com/Fraternalilab/BrepPhylo), immunarch, igraph (https://cran.r-project.org/package=igraph) and ComplexHeatmap (37).

Results

Repertoire reconstruction from bulk transcriptome data

The step-by-step pipeline for the processing of BCR repertoire from whole transcriptome data is summarized in Figure 1. To demonstrate the functionality and efficiency of the bcRflow pipeline, we applied our pipeline to bulk transcriptome data downloaded from the Gene Expression Omnibus (GEO) (38). The bulk data consisted of longitudinal samples sequenced from the peripheral blood mononuclear cells of 22 COVID-19 patients exhibiting clinical heterogeneity ranging from seronegative patients exposed to COVID-19 to severe symptoms, together with seven healthy controls collected prior to the COVID-19 pandemic. Disease severity distinguishing the mild and severe cohorts was determined in the original study using a self-reported score screening for 38 different COVID-19 symptoms developed by Duke University Medical Center, with each symptom rated from 0 (none) to 4 (very severe), then summed for a total score. Average scores were 12.8 ± 1.9 for mild patients and 33.6 ± 2.4 for severe patients. The selected sample cohort was composed of both male and female patients and controls, and we utilized only seropositive mild and severe samples for this study. In total there were 48 samples selected (12 from each group), of which 45 samples were successfully processed using the bcRflow pipeline, with 3 samples failing due to low clonal counts. The characterization of B cell repertoire dynamics of immunoglobulin heavy chain (IGH) repertoires between healthy, exposed, mild, and severe disease were tracked through measures of BCR diversity, CDR3 distribution, gene segment usage, SHM rate, isotypes, and CSR between the groups. The resulting analysis and output format allow unified processing and comparison of immune repertoires between the different sample groups.

Here, we implement the bcRflow pipeline based on the MiXCR algorithm (15), which allows for alignment and assembly of clonal sequences from short read data, and export exhaustive information about each clone including nucleotide and amino acid sequences of gene features, gene assignments, read counts, start and end points of key gene regions, and many other statistics. In total, MiXCR successfully aligned, assembled, annotated and exported 17 613 heavy-chain B cell clones across all the groups. The distribution of clonal counts across the sample cohorts was comparable and is as follows: healthy: 2458 (median = 229); exposed: 4045 (median = 335); mild: 4258 (median = 338); severe: 6852 (median = 323.5). Our case study provides detailed insights on BCR repertoire in COVID-19 using bulk transcriptomic data,

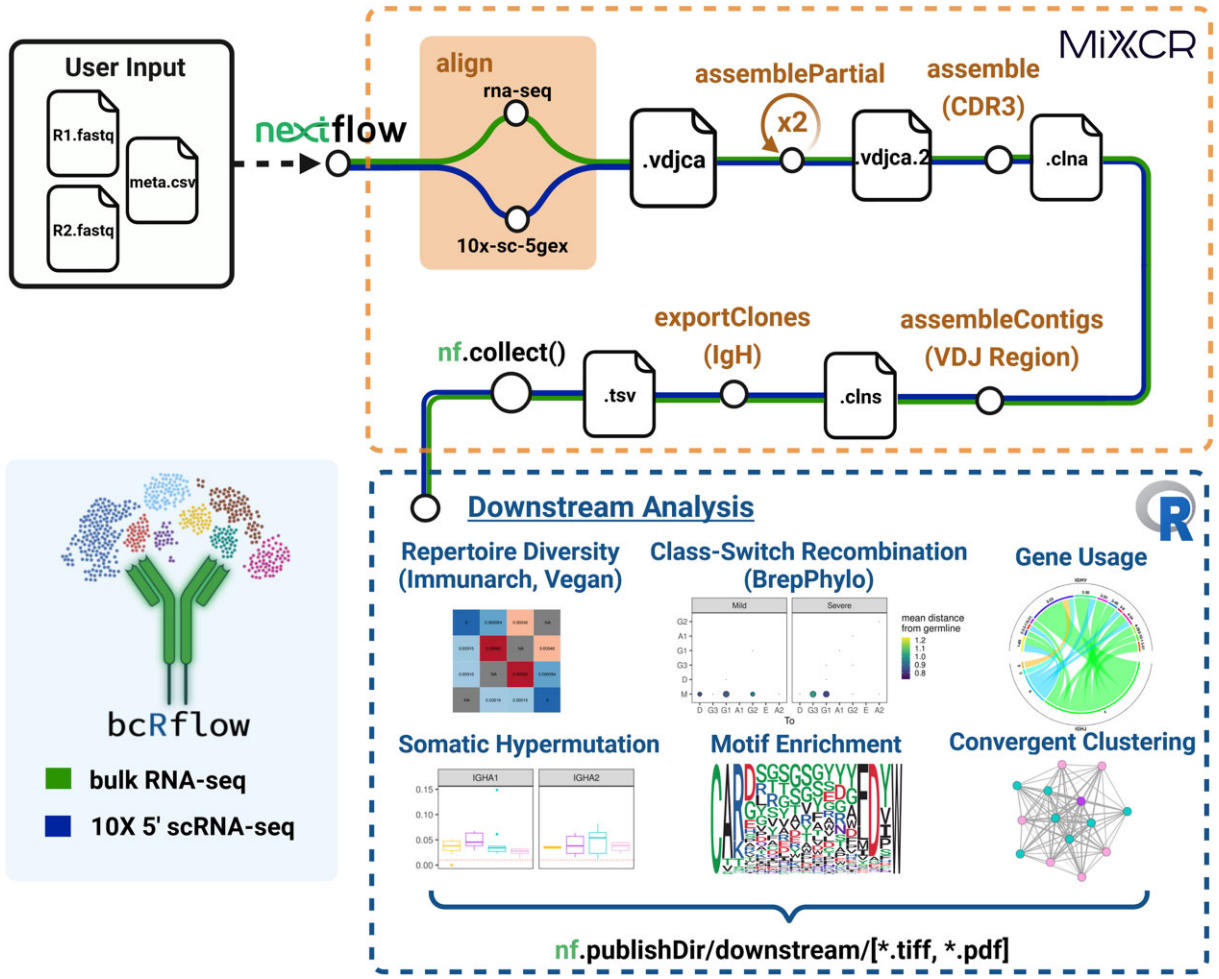

Figure 1. Pipeline overview. A graphical representation of the bcRflow Nextflow pipeline, with colored chords for the separate bulk and single-cell pipelines demonstrating how samples are processed and analyzed, from user input to final plot generation. Nodes along the path represent distinct modules/steps in the pipeline, and file icons highlight the type of file produced by each step which is passed to the subsequent module. The pathway terminates in the Downstream Analysis module, where figure icons highlight the different analyses performed using custom R scripts and standard packages. Created in BioRender. BioRender.com/h85m170.

contributing to a better understanding of the humoral immune response after infection.

For users interested in the application of bcRflow to analyze single-cell RNAseq data, we have included a case study in the GitHub repository that compares BCR repertoires between long COVID and non-long COVID patients (39). This study utilizes 10 × 5′ single-cell RNAseq data retrieved from GEO (accession number GSE235050). We have not included this case study in the results section for brevity, as the primary difference between the single-cell and bulk RNAseq analysis occurs in the alignment stage (Figure 1).

BCR gene usage differs in receptor composition across disease severity

The immense diversity of the BCR repertoire is attributed to the V(D)J recombination of V, D and J gene segments in various combinations, and the heavy chain plays a major role in antigen-binding interactions in most antibodies. Hence, we explored the preferential BCR gene usage bias of V gene (IGHV) segments in disease severity when compared to healthy controls. For V gene segments in the heavy chain, IGHV3, IGHV1 and IGHV2 gene families were frequently used in both COVID-19 samples as well as healthy controls, especially with >70% of all BCRs accounted by IGHV3 and IGHV1 family (Figure 2A). Notably, similar IGHV genes were utilized in SARS-CoV-2 studies using antibody sequencing. In addition, the frequencies of gene segments in each IGHV family were assessed between COVID-19 samples and healthy controls at a P-value <0.05 (Fisher's exact test), and odd ratio >1. IGHV1-18, IGHV1-69, IGHV4-34 and IGHV4-4 were significantly increased in exposed samples when compared to healthy controls (Figure 2B). Previous studies have investigated the role of IGHV4-34 in producing self-reactive antibodies through specific sequence motifs not found in other IGHV gene segments (40,41). Increased use of IGHV1-69, and IGHV3-30 were observed in mild versus healthy controls (Figure 2B) whereas severe versus healthy controls show an increased use of IGHV3-9, IGHV3-30-3, IGHV3-30 and IGHV3-33 (Figure 2B) of the IGHV3 family (42). Increased use of IGHV3-30 and IGHV3-33 were previously reported in the BCR analysis of COVID-19 recovery patients versus healthy controls using V(D)J sequencing data (43–45). Additionally, increased IGHV3-30 usage has been associated with

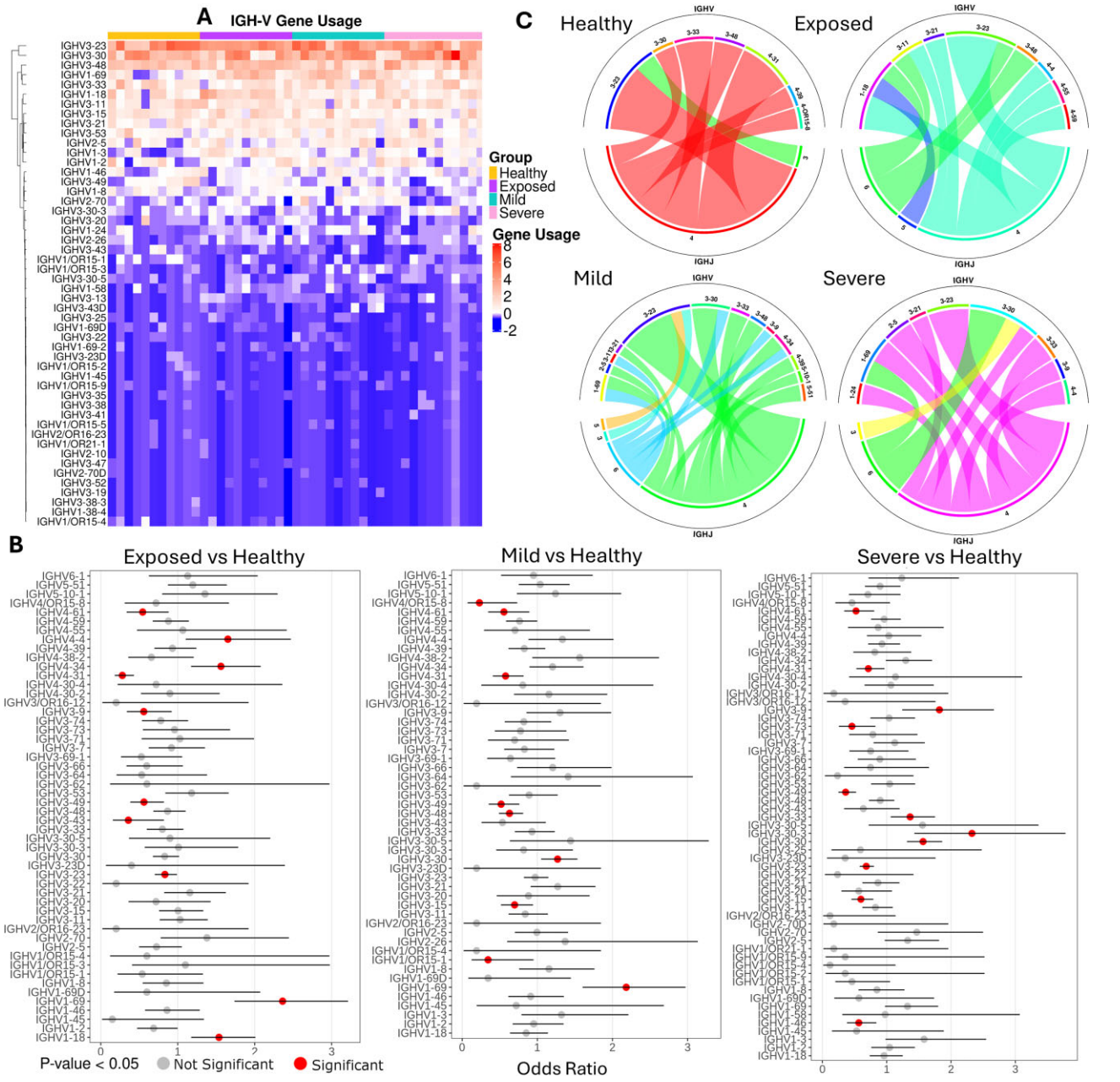

Figure 2. Activated B cell proliferation: IGHV and IGHJ gene usage. (**A**) Heatmap showing frequency of IGHV gene usage across samples with varying disease severity. Each row corresponds to a V gene, each column corresponds to a sample, and the blue-red color spectrum indicates the frequency of the gene in the repertoire. Genes were hierarchically clustered by usage level across samples, highlighted by the dendrogram on the left y-axis. (**B**) Dot plots of the differential analysis of gene usage for multiple pairwise comparisons between sample groups. The red dots represent $P < 0.05$ (by Fisher's exact test), and the gray dots represent $P > 0.05$. Odds Ratio is indicated in the x-axes. (**C**) Circos plots show the V and J segment pairs frequency in the BCR repertoires in representative samples of the different conditions. Arcs represent IGHV and IGHJ gene types with size relative to frequency. Colored links between gene type arcs represent unique IGHV and IGHJ combinations.

higher levels of plasma neutralization of SARS-CoV2 in convalescent patients (43). Taken together, these results highlight the differential usage bias of IGHV gene segments detected through BCR profiling of bulk data and how it corroborates to similar results from V(D)J, and antibody sequencing BCR data analysis.

We then examined the V and J gene recombination as it mainly contributes to the diversity of the BCR repertoire. V-J pairs were visualized using chord plots for representative samples from each sample cohort. The chord diagrams in Figure 2C show gene segments from the healthy controls (IGHV3-23,3–30,3–33,3–48,4–31,4–39), exposed (IGHV1-18,3–11,3–21,3–23,3–48,4–4,4–55,4–59), mild (IGHV1-69,2–5,3–1,13–21,3–23,3–30,3–33,3–48,3–9,4–34,4–39,5–10-1,5–51) and severe (IGHV1-24,1–69,2–5,3–21,3–23,3–30,3–33,3–9,4–4) gene families. In terms of the J segment, IGHJ4 is the most frequently enriched segment consistent with previous studies (42,46). IGHJ3, IGHJ5 and IGHJ6 were other J gene frequencies observed in samples with mild, moderate, and severe symptoms, and consistent with results from single cell immune sequencing results (43).

Quantifying B cell receptor diversity

We assessed the diversity between the repertoires of each sample using the R packages Vegan (47) and immunarch and chose several metrics to cover multiple aspects of diversity. In the context of BCR repertoire profiling, we focus on clonotype abundance, richness (number of unique clones), evenness (degree to which the different clonotypes are equally represented in the samples), and CDR3 sequence similarity. These different measures each place varying levels of importance on specific clonal characteristics.

The complementarity determining region 3 (CDR3) is a highly variable region in the BCR and has a critical role in antigen recognition of B cells (48). We next explored the characteristics of CDR3 length, and Figure 3A depicts the distribution of the CDR3 length in the heavy chain with the curves of the violin plot representing the density of the values. The length of CDR3 was concentrated in two lengths, 15 and 20 amino acid residues. The largest length of 40 amino acid residues was observed in mild samples. The proportion of BCRs with different CDR3 length was highly consistent between diseased and healthy samples.

Hill indices in Figure 3B are a generalization of the species richness, Shannon entropy and the Gini–Simpson index, and define the BCR repertoire diversity as a function of a continuous parameter q. The Hill number gradually decreased in the healthy samples as q increased, but the exposed cohort interestingly surpassed the severe in q 3–6.

We used multiple measures to estimate the dominance of clones, or clonal evenness, in a repertoire as shown in Figure 3C. Pielou evenness in Figure 3C represents Shannon entropy scaled by the maximum number of clones per sample and demonstrates the evenness of clones on a range of 0 to 1, with 1 representing total evenness and 0 representing total unevenness. Similarly, the Gini coefficient quantifies the evenness of the distribution and is used to represent the clonal distribution of the BCR repertoire. The value ranges from 0 (maximal diversity of the repertoire) to 1 (representing extreme inequality). Neither the Gini coefficient nor Pielou evenness metrics showed any statistically significant difference across the sample groups. Lastly, the Shannon diversity index is a common estimator of evenness, with a higher score indicating higher diversity within the group. None of the groups showed statistically significant differences in the Shannon diversity estimation.

As the most common approach to measure similarity, we profiled repertoire overlap by computing the Jaccard index, cosine similarity, and Morisita's overlap index using the CDR3aa sequence. Most of these indices have a value between 0 (no similarity) to 1 (total similarity), and Figure 3D shows very little overlap in repertoires across COVID-19 disease severity.

Most of the diversity metrics are dependent on the number of sampled B cells, and need to be addressed before the metrics are being compared. An alternative to that is to subsample the largest repertoire to match the size of the smallest ones for comparison purposes (26,49). Thus, caution must be exercised while interpreting and comparing immune repertoire diversity metrics within and across samples.

Clonal expansion and differentiation

We investigated clonal expansion and distribution by using clonality indices, which measure clonal volume and the proportions occupied by the most and least abundant clones, both in terms of rank and relative abundance.

To answer the question of how the clonal architecture of the repertoires varies by disease phenotype, we assessed the proportion of the repertoire occupied by the clones of a given size. Ideally, a small clonal index indicates an expanded clonotype whereas a larger index indicates a small clonotype group. Although not significant, the diseased samples had larger clonal proportions in the top 101:1000 (shown in the x-axis of Figure 4A) than healthy controls. Relative abundance of all BCR clones in the repertoire were grouped into four categories ranging from small clones which take up $<1/10\,000$ of clonal space, to hyperexpanded clones which take up $>1/10$th of clonal space.

The clonality spectrum in Figure 4B revealed that most of the COVID-19 sample repertoires were dominated by large clonotypes. Here a 'large' clonotype is defined as a clonotype with an abundance threshold falling within the range of 0.01–1% of the total immune repertoire. In contrast, the BCR repertoire of healthy samples primarily consisted of hyperexpanded clones.

BCR repertoire richness (number of unique BCR sequences) provides an additional measure of expansion and differentiation, which we estimated using the Chao1, ACE and inverse Simpson metrics. The Chao1 metric (Figure 4C) is an additional estimator of richness based on the number of occurrences of rare clones within a repertoire. Chao1 showed a significant difference between severe COVID-19 samples and healthy controls ($P < 0.05$), suggesting lower BCR richness in the healthy samples. Richness measures using Chao1 estimates did not differ significantly between any other groups.

The ACE (abundance-based coverage estimator) index in Figure 4C estimates richness by comparing the number of rare clones and inflating them against the number of highly abundant clones. Here, we see that severe samples have a higher level of species richness compared to both the healthy and mild cohorts, with $P < 0.05$.

The inverse Simpson index was used to assess the probability of two randomly sampled reads belonging to the same clone. High values of the index indicate an even distribution of BCR clones, and lower values indicate enrichment of B cell clones. A significantly higher diversity was observed in severe COVID-19 samples when compared to the healthy control (Figure 4C). The trend was not observed in other sample comparisons. It is of importance to note that the inverse Simpson index should only be used when the data contains high-frequency reads.

The D50 diversity index of each repertoire was calculated (Figure 4D) and corresponds to the percentage of unique CDR3 sequences that account for 50% of the total number of CDR3 in the sample. A smaller D50 value indicates lower diversity with a few dominant clonal expansions, and higher diversity corresponds to more small clonal expansions. Clonotypic expansions, with markedly higher and statistically significant D50 diversity indices, were observed in severe COVID-19 samples compared to healthy controls.

Somatic hypermutation and class-switch recombination

The Ig types in BCR analysis usually have a different degree of antibody affinity, and we compared the frequency of the isotypes between COVID-19 and healthy controls. Distribution

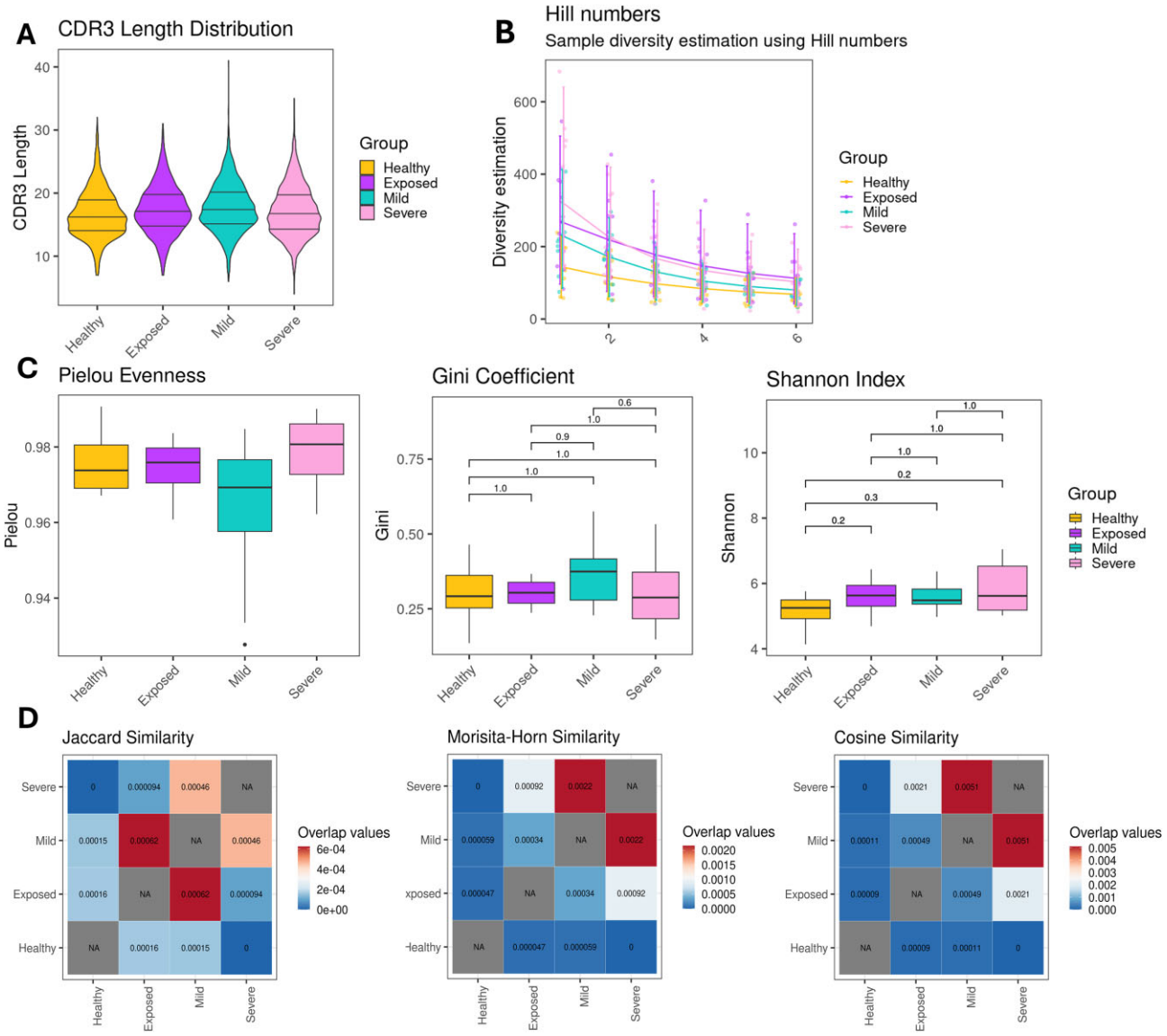

Figure 3. Repertoire diversity metrics. (**A**) Violin plot showing CDR3 length distributions across sample groups, where the y-axis indicates CDR3 length in amino acid residues. (**B**) Hill diversity dot plot showing trends in diversity estimates across increasing values of the order parameter shown on the X axis. Points represent samples, and lines colored by the grouping variable show the mean trend. (**C**) Bar and box plots displaying diversity and evenness metrics, where colors represent distinct groups, and larger y-axis values indicate higher diversity or evenness. *P*-values shown between groups were calculated using the Kruskal–Wallis test and adjusted using the Holm–Bonferroni method to determine if there was a significant difference in diversity or evenness between any two disease states. (**D**) Heatmaps of repertoire overlap scores across disease severity for every pairwise comparison, where overlap values are represented by color. Larger degrees of overlap are indicated by red coloration while smaller degrees of overlap are indicated by blue coloration.

of the 10 Ig isotypes is depicted in Figure 5A, with IGHM having the largest proportion across all groups, highest (>60%) in exposed samples. In contrast, IGHA1 has an increased representation (∼20%) in exposed, mild and severe samples, consistent with single cell sequencing studies of COVID-19 (25).

Many BCRs undergo class switch recombination (CSR) when the B cells respond to an antigen, generating different antibody isotypes and serving as an additional mechanism of affinity maturation. Constant (C) regions of BCRs are typically organized in the following order during immune response via CSR, where the C region is 'switched' while the antigen-binding region is maintained: Cμ (IGHM), Cδ (IGHD), Cγ (IGHG), Cε (IGHE) and Cα (IGHA) (50). To further characterize the CSR profile in the samples, we assessed the progression of CSR between the COVID-19 patients which is visualized in Figure 5B. We can see that very few mutations have been accumulated at the time of the CSR in healthy samples at a very close distance from the germline and mostly corresponding to IGHG1 to IGHE and IGHG2 switches. Exposed samples depict CSR events from IGHM/D to IGHG2, and IGHG1 at a farther distance from the germline. Interestingly, mild, and severe samples appear to have a higher mutational rate with most changes in mild from IGHM/D to IGHG1, IGHG2 and IGHG3. Severe samples show high percentages of switches relative to the number of clusters with three or more members that demonstrated CSR events. These switches are from IGHM/D to IGHG1, and IGHG3. Severe and mild samples also demonstrate broader

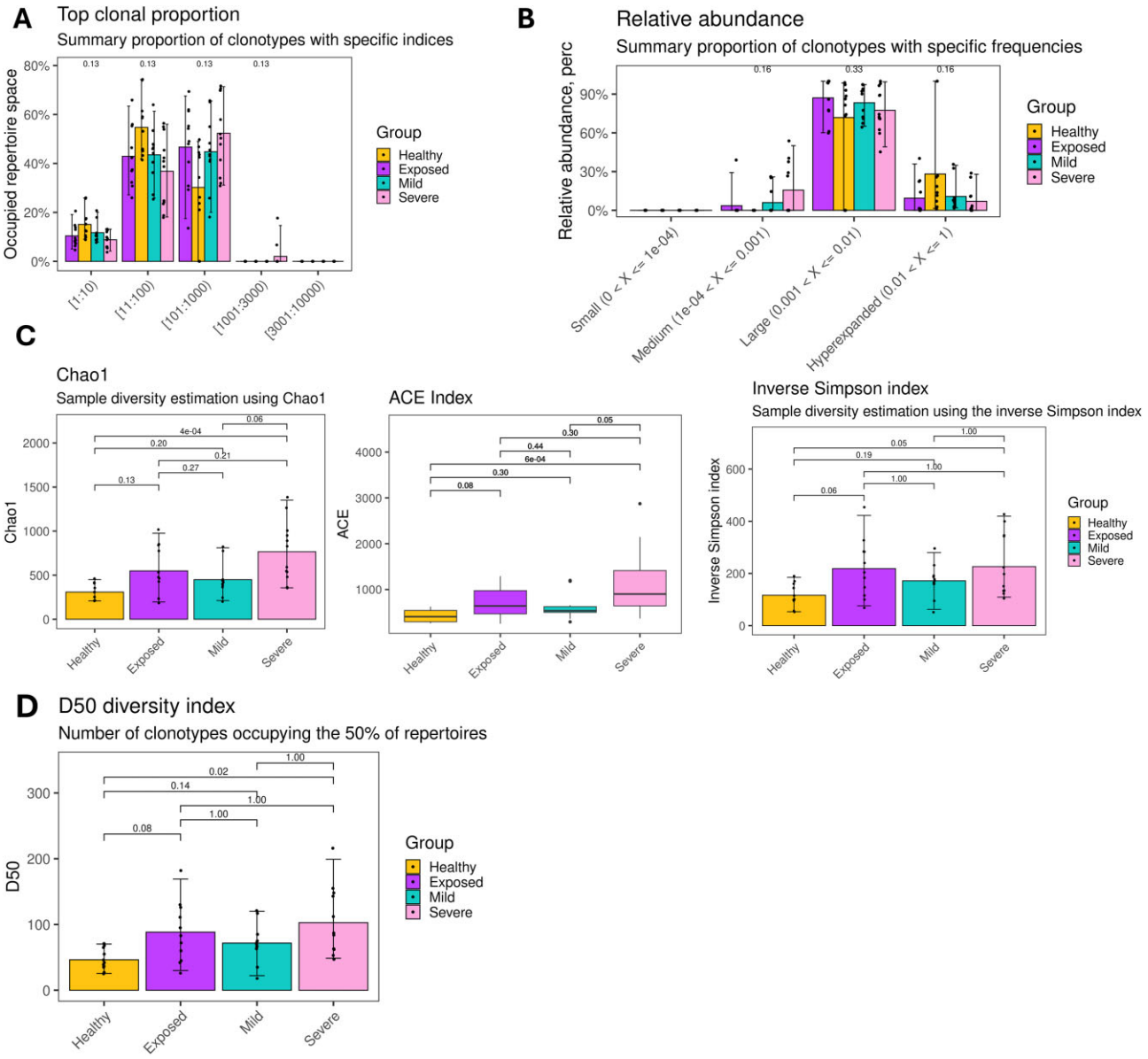

Figure 4. Clonal expansion and differentiation. (**A**) Top clonal proportions across disease states. The x-axis indicates the clonotypes ranked by frequency and grouped in sets of frequencies, while the y-axis shows the proportion of the groups of clonotypes in the entire repertoire. (**B**) Relative abundance of clonotypes across disease states where clonotype frequencies are indicated in the x-axis and relative abundance shown in the y-axis. (**C**) Bar and box plots displaying richness metrics, where colors represent distinct groups, and larger y-axis values indicate higher repertoire richness. P-values calculated using Kruskal-Wallis and Holm–Bonferroni correction. (**D**) Bar plot of the D50 metric, where the diversity index score is shown on the y-axis and disease states are shown on the x-axis. P-values shown above each bar were calculated by Kruskal–Wallis tests to determine significance between D50 values in any two compared disease states.

class switching at lower rates and further from the germline, indicated by the much smaller bubbles. Previous studies have also reported an increase in CSR to IGHG3 in viral infections when compared to healthy individuals (48,51) suggesting a role of IGHG3 as a key factor in virus clearance. Compared with other isotypes, IGHM/D isotypes extensively undergo class switch in the COVID-19 samples in comparison to healthy controls (48,51).

We explored the somatic hypermutation (SHM) rates in IGH repertoires which introduce point mutations in the antibody variable region that encodes the antigen-binding sites, thereby enhancing antibody neutralization, breadth, and potency. The V and J germline, and clonal sequences were used as the input to calculate the SHM rates comparing COVID-19 severity to healthy controls. Reduced SHM is consistent with evidence from other SARS-CoV-2 studies (2,51), and we observed the same across all severity groups (Figure 5C) indicating that activation of SARS-CoV-2-specific antibody response is generated without extensive somatic mutations. When comparing all the samples, mild severe patients have a slight increase in IGHG4, and IGHGP proportions.

Convergent clustering of BCR clones

Convergent clustering of CDR3 sequences among different groups provides insights into shared immune responses during

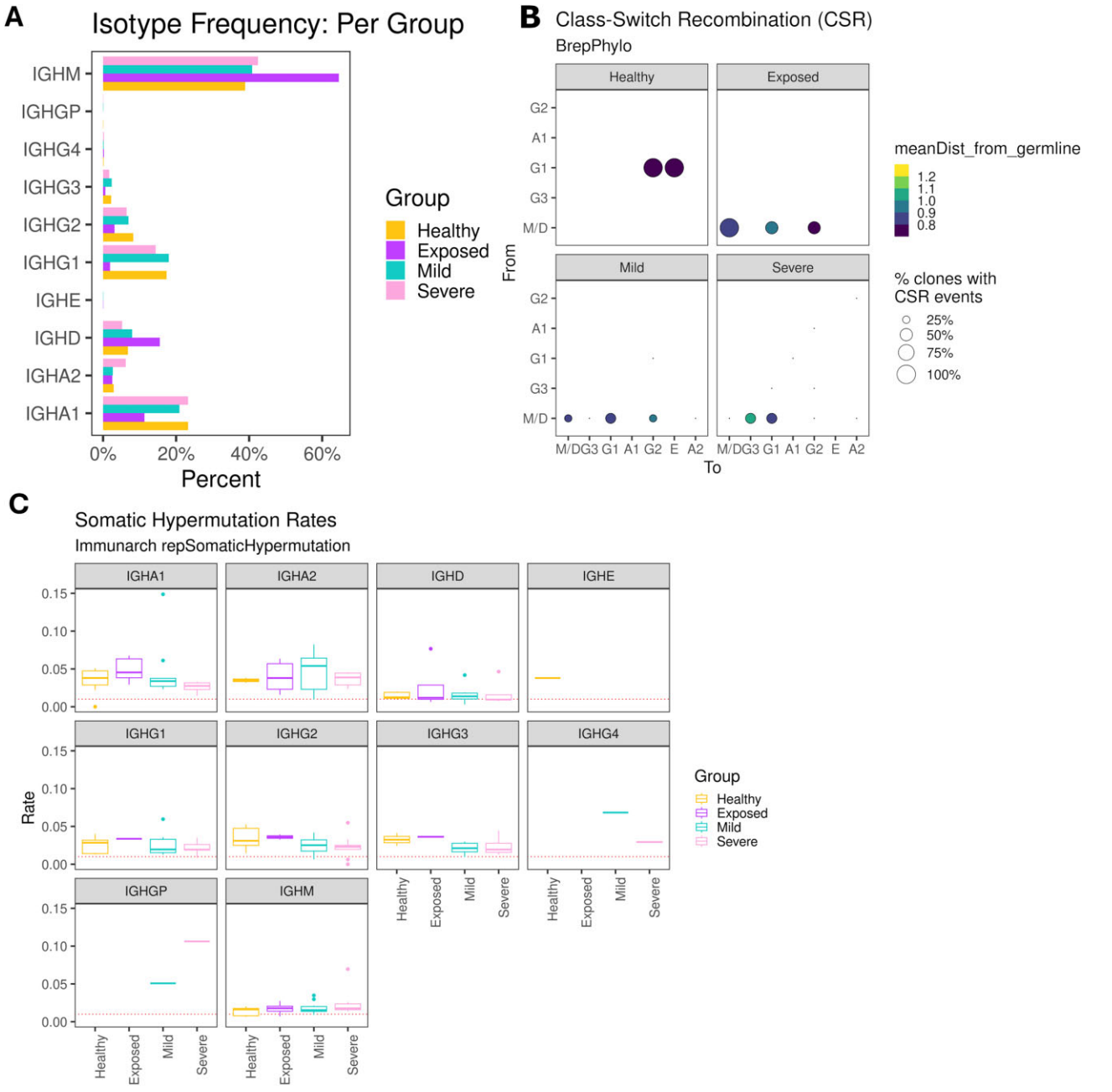

Figure 5. Antigen derived antibody affinity maturation. (**A**) Isotype frequency across IGH sequences per disease state where larger bars represent a higher proportion of a specific IGH isotype within a group. (**B**) Class switch recombination events are shown in a bubble plot. Larger bubbles represent a higher percentage of clusters with three or more unique CDR3 sequences with the respective CSR events. Small bubbles, like in the Severe group highlight broader CSR events, but at much lower rates relative to other events. The y-axis denotes the initial IGH class, the x-axis shows the class switched to, and bubble color reflects the degree of CSR with lighter colors indicating greater distance from the germline sequence. (**C**) Box plots showing somatic hypermutation rates across IGH sequences among disease states. The y-axis indicates the SHM rate with larger values representing higher degrees of mutation.

various stages of infection or disease progression. By examining clusters of shared clones, we can better understand the levels of affinity maturation across different clinical severities. To determine if there was increased clonal sharing in COVID-19 patients compared to healthy individuals, we analyzed the Levenshtein distance between grouped CDR3 amino-acid sequences, considering shared V and J gene usage and equal CDR3 amino-acid sequence length.

We then clustered the CDR3aa sequences within these groups of clones, using a distance threshold of at least 70% similarity, resulting in an annotated list of clustered sequences. Of the 14 586 total clusters across all samples, 1537 (10.5%) were shared between at least two of the samples, and 620 (4.25%) were shared between two or more groups. We focused on the 39 large clusters of 10 or more clones (Figure 6A), a threshold utilized in previous studies to identify convergence (34) and noticed that most of these clusters primarily consisted of clones from either severe or healthy samples. Notably, certain V genes like IGHV1-69 and IGHV3-30 were prominent in these large clusters, indicating their association

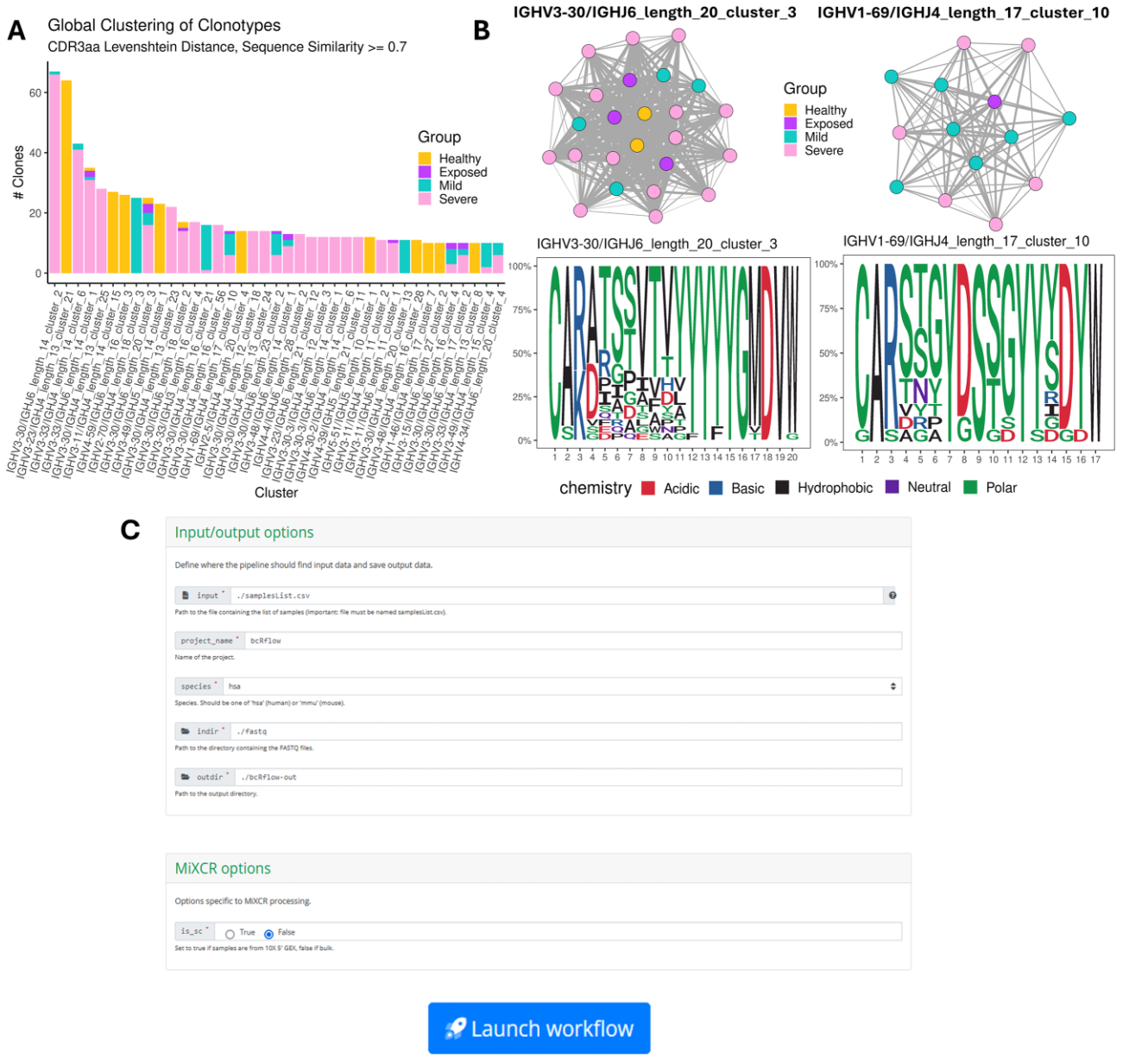

Figure 6. Convergent antibody response and OnDemand user interface. (**A**) Bar plot showing global clustering of IGHV clonotypes where disease state groups were combined based on IGHV genes, IGHJ genes and CDR3 amino acid length. Clusters of 10 or more unique clones were used for visual comparison. Larger bar values represent more unique clones found in any cluster. (**B**) Similarity networks of heterogeneous clusters containing clones across disease states were visualized in undirected network graphs consisting of nodes representing unique clones and edges representing CDR3 amino acid sequence similarity. The color of nodes indicates the disease state of the specific clone. Amino acid composition visualizations were used as an additional representation of CDR3 sequence similarity where colors represent the chemical nature of the amino acid side chains. The y-axis represents the frequency of any amino acid residue encountered at the sequence position indicated by the x-axis. (**C**) Screen capture of the bcRflow Pitt On-Demand instance, the graphical user interface for the implementation of bcRflow using the University of Pittsburgh's Center for Research Computing infrastructure.

with maturation (Figure 6A). These V genes were also found to be significant compared to healthy samples in our gene usage analysis (Figure 2B).

Network diagrams and sequence logo plots for key convergent clusters (Figure 6B) illustrate the similarity of specific antigen sequences, across different sample groups. Nodes in the network represent unique clones, and the weighted edges represent sequence similarity (≥70%) based on Levenshtein distance. The first cluster, IGHV1-69/IGHJ4-length_17, has a V gene that has been associated with affinity maturation in COVID-19 patients, binding to the receptor binding site C (RBS-C) epitope of the SARS CoV-2 virus (41). The identified cluster suggests a more mature immune response in COVID-19 patients, potentially evolving with increased disease severity.

Additionally, the second cluster, IGHV3-30/IGHJ6-length_20 (Figure 6B) consists of clones from all four groups, suggesting the potential use of a versatile antigen for a broad immune response to various challenges, as observed in previous studies (34,52). IGHV3-30 was also enriched in diseased samples compared to healthy controls, demonstrated in our gene usage analysis (Figure 2B).

Computing resources and runtime

It is recommended to use bcRflow with institutional high-performance computing (HPC) clusters or cloud-based systems like AWS for optimal performance. Local execution is possible for small sets of samples, but it is highly dependent on the configuration and available memory of the user's computer. It is suggested to allocate at least 8–16 GB of RAM and 2–4 CPUs per process for efficient processing. MiXCR alignment may take >12 h for large bulk samples, but the average alignment and processing time for the case study finished in 6 h with 8GB of RAM per sub-process. Users can specify memory and CPU allocation in the bcRflow configuration file to customize their setup. Additional configurations for institutional computing clusters and cloud-based systems are provided by default with bcRflow.

Interactive interface using Open OnDemand

The Nextflow community has grown and provides high-quality, scalable bioinformatics pipelines that are reproducible and interoperable. Despite this progress, biologists still encounter difficulties when using high-performance computing environments and need visually engaging and interactive web platforms to execute these pipelines. To address this, we have introduced a graphical user interface for the Open OnDemand framework (53) utilizing the nextflow_schema.json file generated by parsing user input to configure the bcRflow run. Users can specify the input parameters accepted by the bcRflow pipeline (as shown in Figure 6C) through OnDemand's interactive app and initiate the process by clicking 'launch'. The customized bcRflow run will then be sent to the local HPC environment, and once the job is finished users can utilize the web-based file explorer to view the results.

Discussion

Profiling B cell receptor (BCR) repertoires is crucial for understanding the adaptive immune response and immune cell function in health and disease (54,55). Targeted sequencing data has been instrumental in revealing the diversity of BCR repertoires, identifying clonal expansions, and shedding light on B cell responses in various conditions. Analytical tools like Change-O, SCOPer, Partis, MobiLLe, and fastBCR (56,57) are tailored for analyzing targeted BCR sequencing data. The airrflow pipeline (11), based on the immcantation framework, stands out as the sole Nextflow pipeline offering a comprehensive analytical solution for processing targeted BCR data.

Leveraging transcriptomics data for B cell repertoire analysis presents a cost-effective alternative to expensive targeted sequencing methods, enhancing our understanding of the immune response by linking it to gene expression and regulatory mechanisms. However, the lack of a streamlined workflow for BCR profiling from non-targeted data poses a significant challenge in immunology research. To bridge this gap, we introduce bcRflow, a robust computational pipeline designed for immune repertoire analysis from non-targeted RNA-seq reads. Powered by Nextflow, bcRflow incorporates best practices for analysis, ensuring scalability, reproducibility, and user-friendliness. Additionally, bcRflow offers accessible visualization tools, customizable parameters, and publication-ready plots, facilitating seamless integration of immunology and computational research.

The case study we provided showcases the effectiveness of the bcRflow pipeline in analyzing bulk transcriptome data. It successfully processed 17613 heavy-chain B cell clones from both COVID-19 patients and healthy controls, demonstrating its ability to yield results comparable to targeted sequencing methods (58). bcRflow's reliability in capturing B cell repertoire dynamics is highlighted, particularly in revealing preferential V gene usage like IGHV1-69 and IGHV4-34 in COVID-19 patients, corroborated with results from targeted sequencing analysis as important indicators of response and recovery to the COVID-19 virus (40,41). The observed variations in gene segment frequencies across disease severity levels emphasize its utility in capturing nuanced immune responses and maturation. Additionally, the analysis of Ig isotype distribution, class switch recombination, and somatic hypermutation rates offers valuable insights into antibody affinity and maturation processes across disease severity levels. Despite insignificant rates of mutation, greater CSR events were observed in mild and severe cases. Furthermore, the convergent clustering analysis revealed compelling similarities in antibody affinity maturation between different COVID-19 severity levels, as described in studies utilizing targeted sequencing methods (59). These results affirm the utility of the bcRflow pipeline for comprehensive B cell repertoire analysis, providing deeper insights into immune responses and disease pathogenesis in infectious diseases like COVID-19.

Regarding the application of bcRflow, it can be used with bulk and 10 × 5-prime GEX single-cell transcriptomic datasets. Thanks to Nextflow's modular framework and MiXCR's presets for processing various sequencing modalities, we plan to incorporate support for additional single-cell technologies. Currently, the downstream analysis module supports samples from two input species (*Homo sapiens* and *Mus musculus*), but IMGT and MiXCR offer support for many more species that can be integrated into the bcRflow framework.

Moreover, the integration of bcRflow into the OnDemand portal aims to offer researchers a user-friendly interface for seamless access and execution of bcRflow, eliminating the requirement for advanced technical expertise. This enhancement improves accessibility, resource management, collaboration, and overall efficiency, catering to a broader user base. The pipeline is extensively documented and includes a use-case scenario for novice users. We anticipate the release of the workflow and engage with the scientific community to incorporate user input, recommendations, and requests to ensure that bcRflow remains current and vibrant. Collaboration and feature requests are encouraged through the bcRflow GitHub repository (https://github.com/Bioinformatics-Core-at-Childrens/bcRflow).

Data availability

Bulk transcriptomic data used in this study has been downloaded from the NCBI Gene Expression Omnibus (GEO) with the accession number GSE206263). All relevant code and software dependencies have been made open source under the MIT software license, available in the form of a public GitHub repository (https://github.com/Bioinformatics-Core-at-Childrens/bcRflow) and corresponding Docker Hub image (https://hub.docker.com/repository/docker/bioinformaticscoreatchildrens/bcrflow). A

persistent version of bcRflow is available via FigShare (doi: 10.6084/m9.figshare.25881103).

Acknowledgements

Author contributions: **Brent T. Schlegel**: Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Validation, Software, Visualization, Writing - original draft. **Michael Morikone**: Methodology, Validation, Visualization, Writing - original draft. **Fangping Mu**: Software, Resources, Writing - review & editing. **Wan-Yee Tang**: Writing - review & editing. **Gary Kohanbash**: Writing - review & editing. **Dhivyaa Rajasundaram**: Conceptualization, Methodology, Investigation, Validation, Supervision, Project administration, Writing - original draft.

Funding

The computational analysis was implemented using the cluster resources provided by the University of Pittsburgh Center for Research Computing, which is supported by the National Institute of Health [S10OD028483].

Conflict of interest statement

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Chaudhary,N. and Wesemann,D.R. (2018) Analyzing immunoglobulin repertoires. *Front. Immunol.*, **9**, 462.
2. Kotagiri,P., Mescia,F., Rae,W.M., Bergamaschi,L., Tuong,Z.K., Turner,L., Hunter,K., Gerber,P.P., Hosmillo,M., Hess,C., *et al.* (2022) B cell receptor repertoire kinetics after SARS-CoV-2 infection and vaccination. *Cell Rep.*, **38**, 110393.
3. Boyd,S.D., Marshall,E.L., Merker,J.D., Maniar,J.M., Zhang,L.N., Sahaf,B., Jones,C.D., Simen,B.B., Hanczaruk,B., Nguyen,K.D., *et al.* (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci. Transl. Med.*, **1**, 12ra23.
4. Setliff,I., Shiakolas,A.R., Pilewski,K.A., Murji,A.A., Mapengo,R.E., Janowska,K., Richardson,S., Oosthuysen,C., Raju,N., Ronsard,L., *et al.* (2019) High-throughput mapping of B cell receptor sequences to antigen specificity. *Cell*, **179**, 1636–1646.
5. Rodriguez,O.L., Silver,C.A., Shields,K., Smith,M.L. and Watson,C.T. (2022) Targeted long-read sequencing facilitates phased diploid assembly and genotyping of the human T cell receptor alpha, delta, and beta loci. *Cell Genomics*, **2**, https://doi.org/10.1016/j.xgen.2022.100228.
6. Picelli,S., Faridani,O.R., Björklund,Å.K., Winberg,G., Sagasser,S. and Sandberg,R. (2014) Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.*, **9**, 171–181.
7. Stormo,G.D., Zuo,Z. and Chang,Y.K. (2015) Spec-seq: determining protein-DNA-binding specificity by sequencing. *Briefings in Functional Genomics*, **14**, 30–38.
8. Wang,X., Campbell,M.R., Cho,H.Y., Pittman,G.S., Martos,S.N. and Bell,D.A. (2023) Epigenomic profiling of isolated blood cell types reveals highly specific B cell smoking signatures and links to disease risk. *Clin. Epigenet.*, **15**. 90.
9. Bolotin,D.A., Poslavsky,S., Davydov,A.N., Frenkel,F.E., Fanchi,L., Zolotareva,O.I., Hemmers,S., Putintseva,E.v., Obraztsova,A.S., Shugay,M., *et al.* (2017) Antigen receptor repertoire profiling from RNA-seq data. *Nat. Biotechnol.*, **35**, 908–911.
10. Rubio,T., Chernigovskaya,M., Marquez,S., Marti,C., Izquierdo-Altarejos,P., Urios,A., Montoliu,C., Felipo,V., Conesa,A., Greiff,V., *et al.* (2022) A Nextflow pipeline for T-cell receptor repertoire reconstruction and analysis from RNA sequencing data. *ImmunoInformatics*, **6**, 100012.
11. Gabernet,G., Marquez,S., Bjornson,R., Peltzer,A., Meng,H., Aron,E., Lee,N.Y., Jensen,C., Ladd,D., Hanssen,F., *et al.* (2024) nf-core/airrflow: An adaptive immune receptor repertoire analysis workflow employing the Immcantation framework. *PLoS Comput. Biol.*, **20**, e1012265.
12. Bashford-Rogers,R.J.M., Bergamaschi,L., McKinney,E.F., Pombal,D.C., Mescia,F., Lee,J.C., Thomas,D.C., Flint,S.M., Kellam,P., Jayne,D.R.W., *et al.* (2019) Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature*, **574**, 122–126.
13. Hu,X., Zhang,J., Wang,J., Fu,J., Li,T., Zheng,X., Wang,B., Gu,S., Jiang,P., Fan,J., *et al.* (2019) Landscape of B cell immunity and related immune evasion in human cancers. *Nat. Genet.*, **51**, 560–567.
14. Nielsen,S.C.A., Roskin,K.M., Jackson,K.J.L., Joshi,S.A., Nejad,P., Lee,J.-Y., Wagar,L.E., Pham,T.D., Hoh,R.A., Nguyen,K.D., *et al.* (2019) Shaping of infant B cell receptor repertoires by environmental factors and infectious disease. *Sci. Transl. Med.*, **11**, eaat2004.
15. Bolotin,D.A., Poslavsky,S., Mitrophanov,I., Shugay,M., Mamedov,I.Z., Putintseva,E.v. and Chudakov,D.M. (2015) MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods*, **12**, 380–381.
16. Canzar,S., Neu,K.E., Tang,Q., Wilson,P.C. and Khan,A.A. (2017) BASIC: BCR assembly from single cells. *Bioinformatics*, **33** 425–427.
17. Lindeman,I., Emerton,G., Mamanova,L., Snir,O., Polanski,K., Qiao,S.-W., Sollid,L.M., Teichmann,S.A. and Stubbington,M.J.T. (2018) BraCeR: b-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods*, **15**, 563–565.
18. Upadhyay,A.A., Kauffman,R.C., Wolabaugh,A.N., Cho,A., Patel,N.B., Reiss,S.M., Havenar-Daughton,C., Dawoud,R.A., Tharp,G.K., Sanz,I., *et al.* (2018) BALDR: A computational pipeline for paired heavy and light chain immunoglobulin reconstruction in single-cell RNA-seq data. *Genome Med.*, **10**, 20.
19. Song,L., Cohen,D., Ouyang,Z., Cao,Y., Hu,X. and Liu,X.S. (2021) TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods*, **18**, 627–630.
20. Yermanos,A., Agrafiotis,A., Kuhn,R., Robbiani,D., Yates,J., Papadopoulou,C., Han,J., Sandu,I., Weber,C., *et al.* (2021) Platypus: An open-access software for integrating lymphocyte single-cell immune repertoires with transcriptomes. *NAR Genomics Bioinformatics*, **3**, lqab023.
21. Tommaso Paolo,D., Chatzou,M., Floden,E.W., Barja,P.P., Palumbo,E. and Notredame,C. (2017) Nextflow enables reproducible computational workflows. *Nat. Biotechnol.*, **35**, 316–319.
22. Lefranc,M.P. (2011) IMGT, the international imMunoGeneTics information system. *Cold Spring Harb. Protoc.*, **6**, 595–603.
23. Merkel,D. (2014) Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, **2014**, 2.
24. Kurtzer,G.M., Sochat,V. and Bauer,M.W. (2017) Singularity: scientific containers for mobility of compute. *PLoS One*, **12**, e0177459.
25. Andreani,T., Slot,L.M., Gabillard,S., Strübing,C., Reimertz,C., Yaligara,V., Bakker,A.M., Olfati-Saber,R., Toes,R.E.M., Scherer,H.U., *et al.* (2022) Benchmarking computational methods for B-cell receptor reconstruction from single-cell RNA-seq data. *NAR Genomics Bioinformatics*, **4**, lqac049.
26. Yaari,G. and Kleinstein,S.H. (2015) Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.*, **7**, 121.
27. Chiffelle,J., Genolet,R., Perez,M.A., Coukos,G., Zoete,V. and Harari,A. (2020) T-cell repertoire analysis and metrics of diversity

and clonality. In: *Current Opinion in Biotechnology*. Elsevier Ltd. Vol. **65**, pp. 284–295.
28. Roswell,M., Dushoff,J. and Winfree,R. (2021) A conceptual guide to measuring species diversity. *Oikos*, **130**, 321–338.
29. Schwartz,G.W. and Hershberg,U. (2013) Conserved variation: identifying patterns of stability and variability in BCR and TCR v genes with different diversity and richness metrics. *Phys. Biol.*, **10**, 035005.
30. Wittebolle,L., Marzorati,M., Clement,L., Balloi,A., Daffonchio,D., Heylen,K., de Vos,P., Verstraete,W. and Boon,N. (2009) Initial community evenness favours functionality under selective stress. *Nature*, **458**, 623–626.
31. Tran,U.S., Lallai,T., Gyimesi,M., Baliko,J., Ramazanova,D. and Voracek,M. (2021) Harnessing the fifth element of distributional statistics for psychological science: a practical primer and shiny app for measures of statistical inequality and concentration. *Front. Psychol.*, **12**, 716164.
32. Geisberger,R., Lamers,M. and Achatz,G. (2006) The riddle of the dual expression of IgM and IgD. *Immunology*, **118**, 429–437.
33. Wang,P., Luo,M., Zhou,W., Jin,X., Xu,Z., Yan,S., Li,Y., Xu,C., Cheng,R., Huang,Y., *et al.* (2022) Global characterization of peripheral B cells in Parkinson's disease by single-cell RNA and BCR sequencing. *Front. Immunol.*, **13**, 814239.
34. Stewart,A., Sinclair,E., Ng,J.C.F., O'Hare,J.S., Page,A., Serangeli,I., Margreitter,C., Orsenigo,F., Longman,K., Frampas,C., *et al.* (2022) Pandemic, epidemic, endemic: B cell repertoire analysis reveals unique anti-viral responses to SARS-CoV-2, ebola and respiratory syncytial virus. *Front. Immunol.*, **13**, 807104.
35. Miho,E., Roškar,R., Greiff,V. and Reddy,S.T. (2019) Large-scale network analysis reveals the sequence space architecture of antibody repertoires. *Nat. Commun.*, **10**, 1321.
36. Wickham,H. (2016) In: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, NY.
37. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
38. Giroux,N.S., Ding,S., McClain,M.T., Burke,T.W., Petzold,E., Chung,H.A., Rivera,G.O., Wang,E., Xi,R., Bose,S., *et al.* (2022) Differential chromatin accessibility in peripheral blood mononuclear cells underlies COVID-19 disease severity prior to seroconversion. *Sci. Rep.*, **12**, 11714.
39. Yin,K., Peluso,M.J., Luo,X., Thomas,R., Shin,M.G., Neidleman,J., Andrew,A., Young,K.C., Ma,T., Hoh,R., *et al.* (2024) Long COVID manifests with T cell dysregulation, inflammation and an uncoordinated adaptive immune response to SARS-CoV-2. *Nat. Immunol.*, **25**, 218–225.
40. Safra,M., Tamari,Z., Polak,P., Shiber,S., Matan,M., Karameh,H., Helviz,Y., Levy-Barda,A., Yahalom,V., Peretz,A., *et al.* (2023) Altered somatic hypermutation patterns in COVID-19 patients classifies disease severity. *Front. Immunol.*, **14**, 1031914.
41. Zhou,X., Ma,F., Xie,J., Yuan,M., Li,Y., Shaabani,N., Zhao,F., Huang,D., Wu,N.C., Lee,C.C.D., *et al.* (2021) Diverse immunoglobulin gene usage and convergent epitope targeting in neutralizing antibody responses to SARS-CoV-2. *Cell Rep.*, **35**, 109109.
42. Xiang,H., Zhao,Y., Li,X., Liu,P., Wang,L., Wang,M., Tian,L., Sun,H.X., Zhang,W., Xu,Z., *et al.* (2022) Landscapes and dynamic diversifications of B-cell receptor repertoires in COVID-19 patients. *Hum. Immunol.*, **83**, 119–129.
43. Montague,Z., Lv,H., Otwinowski,J., DeWitt,W.S., Isacchini,G., Yip,G.K., Ng,W.W., Tsang,O.T.Y., Yuan,M., Liu,H., *et al.* (2021) Dynamics of B cell repertoires and emergence of cross-reactive responses in patients with different severities of COVID-19. *Cell Rep.*, **35**, 109173.
44. Feldman,J., Bals,J., Altomare,C.G., St Denis,K., Lam,E.C., Hauser,B.M., Ronsard,L., Sangesland,M., Bracamonte Moreno,T., Okonkwo,V., *et al.* (2021) Naive human B cells engage the receptor binding domain of SARS-CoV-2, variants of concern, and related sarbecoviruses. *Sci. Immunol*, **6**, eabl5842.
45. García-Vega,M., Wan,H., Reséndiz-Sandoval,M., Hinojosa-Trujillo,D., Valenzuela,O., Mata-Haro,V., Dehesa-Canseco,F., Solís-Hernández,M., Marcotte,H., Pan-Hammarström,Q., *et al.* (2024) Comparative single-cell transcriptomic profile of hybrid immunity induced by adenovirus vector-based COVID-19 vaccines. *Genes Immun.*, **25**, 158–167.
46. Gao,H., Yu,L., Yan,F., Zheng,Y., Huang,H., Zhuang,X. and Zeng,Y. (2022) Landscape of B cell receptor repertoires in COVID-19 patients revealed through CDR3 sequencing of immunoglobulin heavy and light chains. *Immunol. Invest.*, **51**, 1994–2008.
47. Dixon,P. (2003) VEGAN, a package of R functions for community ecology. *J. Veg. Sci.*, **14**, 927–930.
48. Jin,X., Zhou,W., Luo,M., Wang,P., Xu,Z., Ma,K., Cao,H., Xu,C., Huang,Y., Cheng,R., *et al.* (2021) Global characterization of B cell receptor repertoire in COVID-19 patients by single-cell V(D)J sequencing. *Briefings Bioinf.*, **22**, bbab192.
49. Stern,J.N.H., Yaari,G., vander Heiden,J.A., Church,G., Donahue,W.F., Hintzen,R.Q., Huttner,A.J., Laman,J.D., Nagra,R.M., Nylander,A., *et al.* (2014) B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci. Transl. Med.*, **6**, 248ra107.
50. Xu,Z., Zan,H., Pone,E.J., Mai,T. and Casali,P. (2012) Immunoglobulin class-switch DNA recombination: Induction, targeting and beyond. *Nat. Rev. Immunol.*, **12**, 517–531.
51. Zhang,Y., Yan,Q., Luo,K., He,P., Hou,R., Zhao,X., Wang,Q., Yi,H., Liang,H., Deng,Y., *et al.* (2022) Analysis of B cell receptor repertoires reveals key signatures of the systemic B cell response after SARS-CoV-2 infection. *J. Virol.*, **96**, 1600–1621.
52. Mai,G., Zhang,C., Lan,C., Zhang,J., Wang,Y., Tang,K., Tang,J., Zeng,J., Chen,Y., Cheng,P., *et al.* (2023) Characterizing the dynamics of BCR repertoire from repeated influenza vaccination. *Emerg. Microbes Infect.*, **12**, 2245931.
53. Hudak,D., Johnson,D., Chalker,A., Nicklas,J., Franz,E., Dockendorf,T. and McMichael,B. (2018) Open OnDemand: a web-based client portal for HPC centers. *J. Open Source Softw.*, **3**, 622.
54. Chen,H., Zhang,Y., Ye,A.Y., Du,Z., Xu,M., Lee,C.S., Hwang,J.K., Kyritsis,N., Ba,Z., Neuberg,D., *et al.* (2020) BCR selection and affinity maturation in Peyer's patch germinal centres. *Nature*, **582**, 421–425.
55. Aizik,L., Dror,Y., Taussig,D., Barzel,A., Carmi,Y. and Wine,Y. (2021) Antibody repertoire analysis of tumor-infiltrating B cells reveals distinct signatures and distributions across tissues. *Front. Immunol.*, **12**, 705381.
56. Wang,K., Cai,L., Wang,H., Shan,S., Hu,X. and Zhang,J. (2024) Protocol for fast clonal family inference and analysis from large-scale B cell receptor repertoire sequencing data. *STAR Protocols*, **5**, 102969.
57. Gupta,N.T., vander Heiden,J.A., Uduman,M., Gadala-Maria,D., Yaari,G. and Kleinstein,S.H. (2015) Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics*, **31**, 3356–3358.
58. Ma,J., Bai,H., Gong,T., Mao,W., Nie,Y., Zhang,X., Da,Y., Wang,X., Qin,H., Zeng,Q., *et al.* (2022) Novel skewed usage of B-cell receptors in COVID-19 patients with various clinical presentations. *Immunol. Lett.*, **249**, 23–32.
59. Claireaux,M., Caniels,T.G., de Gast,M., Han,J., Guerra,D., Kerster,G., van Schaik,B.D.C., Jongejan,A., Schriek,A.I., Grobben,M., *et al.* (2022) A public antibody class recognizes an S2 epitope exposed on open conformations of SARS-CoV-2 spike. *Nat. Commun.*, **13**, 4539.

Received: May 23, 2024. **Revised:** August 13, 2024. **Editorial Decision:** September 19, 2024. **Accepted:** September 23, 2024