AIRRscape: an interactive tool for exploring B-cell receptor repertoires and antibody responses

Eric Waltari[1,*], Saba Nafees[1], Joan Wong[1], Krista M. McCutcheon[1,#], John E. Pak[1,*]

[1]Chan Zuckerberg Biohub, San Francisco, California, United States of America

[*]Corresponding authors

E-mail: eric.waltari@czbiohub.org (EW) or john.pak@czbiohub.org (JEP)

[#]Current address: Immune-Onc Therapeutics Inc, Palo Alto, California, United States of America

Short title (65 characters): *AIRRscape: an interactive tool for exploring antibody repertoires*

Abstract

The sequencing of antibody repertoires of B-cells at increasing coverage and depth has led to the identification of vast numbers of immunoglobulin heavy and light chains. However, the size and complexity of these Adaptive Immune Receptor Repertoire sequencing (AIRR-seq) datasets makes it difficult to perform exploratory analyses. To aid in data exploration, we have developed AIRRscape, an R Shiny-based interactive web browser application that enables B-cell receptor (BCR) and antibody feature discovery through comparisons among multiple repertoires. Using AIRR-seq data as input, AIRRscape starts by aggregating and sorting repertoires into interactive and explorable bins of germline V-gene, germline J-gene, and CDR3 length, providing a high-level view of the entire repertoire. Interesting subsets of repertoires can be quickly identified and selected, and then network topologies of CDR3 motifs can be generated for further exploration. Here we demonstrate AIRRscape using patient BCR repertoires and sequences of published monoclonal antibodies to investigate patterns of humoral immunity to three viral pathogens: SARS-CoV-2, HIV-1, and DENV (dengue virus). AIRRscape reveals convergent antibody sequences among datasets for all three pathogens, although HIV-1 antibody datasets display limited convergence and idiosyncratic responses. We have made AIRRscape available on GitHub to encourage its open development and use by immuno-informaticians, virologists, immunologists, vaccine developers, and other scientists that are interested in exploring and comparing multiple immune receptor repertoires.

Author Summary

Technological advances in next generation sequencing have allowed for broad experimental sampling of immune repertoires, providing insight into how our immune system responds to infection, vaccination, autoimmunity, and cancer. The scale of these “big data”, however, make it difficult to bioinformatically extract the key sequence features that are shared across multiple repertoires. With AIRRscape, we enable large-scale immune repertoire visualization and analysis that requires no knowledge of the command line or advanced programming. By providing the community with an open-source, interactive, and user-friendly interface, we reduce the barriers to exploring immune repertoires at scale. We demonstrate the use of AIRRscape to characterize features of immune responses to viral infection that are shared across multiple repertoire datasets.

1. Introduction

Individual B-cell receptor (BCR) repertoires have exceptional diversity, estimated to be greater than 10^9 in a single adult [1,2]. Improvements in sequencing capability over the past twenty years have allowed a wider sampling of BCR sequences to be uncovered [e.g. 3–5]. These adaptive immune receptor response sequence (AIRR-seq) datasets can offer a complex and information-rich glimpse into B-cell immune responses to vaccination and infection. For example, with analyses of AIRR-Seq data comes an increasing recognition that convergence among BCR repertoires, sometimes termed 'public' clonotypes [1,2,6,7], is important for understanding the humoral immune response to natural infections and for improving vaccine design.

The size of AIRR-Seq datasets can make it challenging to visualize the BCR sequence features of individual repertoires, let alone of multiple repertoires concurrently. Several groups have developed metrics to assess similarity among entire BCR repertoires [8,9], but few open-source tools enable facile, multi-dimensional visualization and exploration of these repertoires. Notable examples of higher-level BCR repertoire visualizations (e.g. for features such as V-gene and J-gene usage, CDR3 length, and CDR3 amino acid sequence motifs) include Circos plots [10], radial phylogenies [11], and clouds summarizing clonotype networks [12]. While powerful in offering a global overview of AIRR-Seq data, these methods are not amenable to interactive exploration, in particular to uncover and display antibody convergence.

To enable simultaneous interactive visualization of multiple BCR repertoires and in-depth data exploration, we have developed an open-source tool called AIRRscape. Analysis using AIRRscape begins with the generation of sequence feature heatmaps, which can span both individual or combined AIRR-Seq datasets. Visual comparison of these heatmaps in their entirety provides a simple and intuitive global overview of differences in three coupled AIRR-Seq dataset features: 1) V- and J-gene usage; 2) CDR length, and 3) either somatic hypermutation (SHM) or sequence read count. AIRRscape enables extraction of all sequences that fit a particular set of features from the heatmap (individual or combined datasets) into bins for finer scale local analysis. Topologies or phylogenetic networks of CDRH3 or CDRL3 amino acid motifs can be generated, on the fly, to examine clonotypes and study antibody convergence among multiple studies and patients. Recognizing the lack of consensus on how to define clonal lineages [2], we allow users to select from a range of CDR3 identity thresholds to define and visualize antibody convergence. AIRRscape can be deployed using R Shiny [13], allowing the tool to be easily used on any web browser without advanced programming or command-line expertise.

Here we use AIRRscape to examine collections of antibodies and BCR repertoires from three types of viral infections to address several questions. For SARS-CoV-2: how representative are published antibodies relative to healthy and COVID-19 patient repertoires, and how common are convergent antibody responses [14–16]? For HIV-1: how comparable are published anti-HIV-1 antibodies relative to SARS-CoV-2 antibodies as well as to HIV-1 patient repertoires, and how common are public antibody motifs [reported in 7]? For DENV: how

common are the convergent antibodies reported in recent studies [17–19]? We show that AIRRscape provides an intuitive and explorable visualization of the convergent antibody responses that occur in patients infected with SARS-CoV-2 or DENV, as well as the idiosyncratic responses that are typical in HIV-1 due to high SHM.

2. Material and Methods

2.1 Datasets

Recent efforts by the AIRR Community [20,21] have led to the organization and standardization of T-cell and B-cell repertoire analysis, enhancing the accessibility of BCR repertoire datasets. Here we utilize two AIRR data repositories, iReceptor and Observed Antibody Space [22–24], to gather BCR repertoire datasets in response to three viral pathogens: SARS-CoV-2, HIV-1, and DENV, the causative agents of COVID-19, AIDS, and dengue, respectively. Datasets primarily come from studies examining patient bulk BCR repertoires [7,11,14,17,18,25–29]. We also examine repositories of known monoclonal antibodies collated from multiple antibody discovery studies: CoV-AbDab for COVID-19 [30], and for HIV/AIDS the IEDB and CATNAP databases [31,32] and a recent dataset by Yacoob et al. [33]. For dengue, we focus on a set of described plasmablast sequences from two patients [17,19]. All datasets are summarized in Table 1.

98 **Table 1. Datasets used in AIRRscape.**

Dataset	Data type	Sample	Reference	Source	Number of sequences
COVID-19	*SARS-CoV-2 mAbs*		2	CoV-AbDab database (2)	*4,306*
COVID-19	patient bulk repertoire	p11	11	iReceptor (22)	9,385
COVID-19	patient bulk repertoire	7450	26	iReceptor	15,645
COVID-19	patient bulk repertoire	galson-01	14	iReceptor	29,745
COVID-19	patient bulk repertoire	M5	25	iReceptor	18,711
Dengue	*dengue mAbs*		19	(19)	*79*
Dengue	patient bulk repertoire	d13	17	SRA	32,495
Dengue	patient bulk repertoires	45 patients	18	Observed Antibody Space (23)	198,119
Healthy control	bulk repertoire		28	SRA	50,942
HIV-1	*HIV-1 mAbs*		31	IEDB (31)	*98*
HIV-1	*HIV-1 mAbs*		32	CATNAP (32)	*441*
HIV-1	*HIV-1 mAbs*		33	(33)	*83*
HIV-1	patient bulk repertoire	NIH45	29	iReceptor	14,644
HIV-1	patient bulk repertoire	MT1214	27	iReceptor	33,855
HIV-1	patient bulk repertoires	6 CAPRISA patients	7	Observed Antibody Space	184,294

99

100 ## 2.2 Data processing

101 BCR repertoire datasets are first downloaded either as flat files, which follow AIRR
102 community standards, or as sequence files that are partially processed [e.g. the pRESTO

module of the Immcantation suite, 34]. For the repertoires in the latter category, the processing is completed using two Change-O scripts from the Immcantation suite, 'changeo-igblast' and 'changeo-clone' [35]. The collated lists of antibodies against SARS-CoV-2 and HIV-1 are focused on protein sequence and do not include SHM, the % nucleotide divergence from germline, which thus needs to be calculated separately. For those antibody sequences we first run a tblastn search to find perfectly matched nucleotide sequences, and then use the Change-O scripts for processing (after checking to ensure the source is the same antibody without codon optimization). The resulting repertoire datasets now follow the AIRR community standards, with columns indicating the assigned V-gene (*v_call*), assigned J-gene (*j_call*), mutation from germline V-gene (*v_identity*), and assigned V(D)J junction amino acid sequence (*junction_aa*).

A function within a custom R script is used to further modify the BCR repertoire datasets before visualization with Shiny. First we calculate the level of SHM from germline V-gene using the standard column *v_identity*. We also calculate CDR3 amino acid length as defined by the international ImMunoGeneTics information system® (IMGT) by removing the first and last residue in the standard column *junction_aa*, with the new value called *cdr3length_imgt*. All sequences with stop codons and sequences with CDR3 lengths that are missing, not divisible by 3, or fewer than 3 amino acids are then removed. Lastly, we create additional columns used in our tool, namely *gf* for the assigned V-gene family, *gf_jgene* for the V-gene family + J-gene assignment, *ncount* for the count of sequences per unique combination of *gf_jgene* + *cdr3length_imgt*, and *shm_mean* and *shm_max* for mean or maximum SHM, respectively, per unique combination of *gf_jgene* + *cdr3length_imgt*.

To visualize multiple repertoires simultaneously, sets of antibodies and/or repertoires are combined in two ways. We first visualize a set as separate faceted panels for individual repertoires (see 2.3). We then combine these repertoires into a single panel, to enable searches of antibody convergence. Due to their broader importance, only heavy chain antibody sequences are visualized in each of these combined repertoires [36,37]. For all of the repertoire-based visualizations, we compact the individual repertoires by collapsing all sequences with the same *gf_jgene* + *cdr3length_imgt* bin assignments and identical CDR3 amino acid motifs. In datasets containing multiple individuals, we compact only individual repertoires before combining them. A depiction of the workflow of dataset processing is summarized in Fig 1.

Fig 1. Workflow of repertoire data retrieval and processing.

2.3 AIRRscape application

AIRRscape is developed as an interactive web application using R Shiny. We use AIRRscape to visualize the main features of BCR repertoire datasets from processed flat files. To begin, researchers can explore a generated heatmap with either multiple panels of individual BCR repertoires (Fig 2A) or a single panel combining multiple repertoires (Fig 2B). The x-axis of the heatmap shows antibodies binned according to their assigned V-gene family + J-gene germline, while the y-axis shows CDR3 length. For this visualization, the user selects a dataset

along with one of three parameters for coloring the bins: by percent of overall total antibody sequences, by average SHM of the bin, or by maximum SHM of the bin. The user can then hover over the paneled heatmaps to view an interactive popup displaying a bin's attributes (Fig 2B). Clicking on a single bin or selecting multiple bins using a drawn rectangle produces a table of antibodies and their features (Fig 3). Within the table, users can query by or focus on any of the antibody features loaded in the datasets. Selected antibodies can be downloaded or further explored via their topologies (i.e. their phylogenetic relationship, with the caveat that convergent antibodies are not clonally related).

Fig 2. AIRRscape visualization of immune repertoires. Heatmaps comparing characteristics of (A) separate and (B) combined datasets for anti-SARS-CoV-2 and anti-HIV-1 antibodies.

Fig 3. AIRRscape interface showing antibodies and CDR3 amino acid topology of selected antibody bin. Selected bin is highlighted in the red box of Fig 2B.

Below the table of selected antibodies, AIRRscape displays interactive topologies of CDR3 amino acid sequence similarity (Fig 3). Binning antibodies both by germline assignment as well as CDR3 length enables examination of these CDR3 motifs in a phylogenetic framework, since the CDR3 motifs will be 'aligned' when constrained to bins with a given CDR3 length. The labels of the topology tips conveniently display antibody names, assigned V-genes, and CDR3 amino acid sequences. The major advantage of this approach is that thousands of antibody CDR3 sequences can be visualized via a web browser for quick and easy exploration by researchers without command-line expertise.

AIRRscape allows for multiple options to create CDR3 topologies. First, a user may select a set of antibodies from the table for constructing their topology via either the neighbor-joining or parsimony tree building methods [38]. Alternatively, a user can select a single antibody sequence of interest to identify all similar sequences in that bin within a chosen amino acid identity threshold. AIRRscape will then display a parsimony topology of these nearest sequences. The options for identity threshold are 50%, 70% [used in 7], 80% [used in 6], and 100% [used in 1]. Topology tip colors are unique to each dataset. Lastly, the user has the ability to change both the height and width of the topology. Height adjustment allows the user to more easily visualize a small topology or examine a large topology (up to 500 sequences) across more than one visible page; adjusting the width allows the user to see the full tip label when there are long branch lengths.

Two custom R scripts, *airrscape_preprocessing.R* and *airrscape_processing.R*, are included to document the data processing and to convert files that follow the aforementioned AIRR community standards for loading into AIRRscape. AIRRscape can be run locally using RStudio within a browser or be deployed onto a local server or service such as https://shinyapps.io. Both the pre-processing scripts and the R Shiny code for AIRRscape are available on GitHub at https://github.com/czbiohub/AIRRscape.

3. Results

To illustrate the utility of AIRRscape, we used it to explore the adaptive BCR responses to three viral pathogens: SARS-CoV-2, HIV-1 and DENV.

3.1 anti-SARS-CoV-2 antibody datasets share characteristics and convergent motifs

We first examined the CoV-AbDab database of validated monoclonal anti-SARS-CoV-2 antibodies and compared them to published bulk BCR repertoires of four COVID-19 patients, as well as to a healthy control bulk BCR repertoire. We visualized 2,153 paired antibody sequences from the CoV-AbDab dataset and a single representative repertoire from each study of COVID-19 patient bulk repertoires [11,14,25,26; Table 1]. After removing duplicate sequences (as described in 2.2), we compared 73,536 COVID-19 patient antibody CDRH3 motifs and 50,942 healthy control motifs (Fig 4).

Fig 4. AIRRscape heatmaps comparing anti-SARS-CoV-2 antibodies, bulk BCR repertoires of four COVID-19 patients, and a healthy control bulk BCR repertoire.

We first asked how representative the anti-SARS-CoV-2 antibodies are relative to patient repertoires and relative to a healthy control repertoire, based on the heatmap visualizations in AIRRscape. This analysis revealed that the overall pattern of heavy chain V-gene + J-gene usage versus CDRH3 length distribution does not differ among the CoV-AbDab dataset, the four patient repertoires, and the healthy control repertoire (Fig 4). The most common V+J gene family assignments in all datasets were IGHV3+IGHJ4, IGHV3+IGHJ6, IGHV4+IGHJ4, IGHV4+IGHJ6, IGHV1+IGHJ4 and IGHV1+IGHJ6. A notable visual difference between the datasets was the presence or absence of sequences assigned to IGHV7 (Fig 4), which is not unexpected given that this gene ‘family’ consists of a single functional V-gene that does not occur in all individuals [1]. SHM levels among the anti-SARS-CoV-2 antibodies (mean 2.3%) were lower than the four patient repertoires (overall mean 3.7%) and the healthy control repertoire (mean 3.3%), although the patient from the Galson et al. study had noticeably higher SHM levels (mean 5.5%). These data suggest that a majority of neutralizing antibodies against SARS-CoV-2 do not diverge greatly from germline and therefore would be more easily and broadly elicited among patients.

Next we used AIRRscape to visualize convergent clonotypes among anti-SARS-CoV-2 antibodies and COVID-19 patient bulk BCR repertoires. Visualizations were created based on their *gf_jgene* + *cdr3length_imgt* bin assignments, where a bin *x_y_z* represents IGHV*x* germline assignment + IGHJ*y* germline assignment + CDR3 length of *z* aa. As use cases, we focused on three recently reported convergences, visualized as four bins: both *3_4_11* and *3_6_11* for an IGHV3-53-based cluster as reported by Yuan et al. [16] and Galson et al. [14]; *3_6_14* for a large cluster found by Galson et al. [14]; and *1_3_16* for a cluster reported by Robbiani et al. [15]. Within each bin, we selected one published monoclonal antibody and examined the topology of similar CDRH3 motifs among all the COVID-19 datasets, using the

80% aa sequence identity threshold selected by Soto et al. [6]. As expected, convergence across all four bins was observed (Fig 5; S1-3 Figs), with each containing 12–30 CDRH3 motifs from the CoV-AbDab dataset; these motifs were found across seven or more unique studies. Three of the bins show convergent motifs among multiple patient repertoires (Fig 5; S1 and S3 Fig), with two also showing convergence among three of the four patient repertoires. Furthermore, two of the bins show convergence to the healthy control repertoire (Fig 5; S1 Fig). The three convergent motifs seen in multiple COVID-19 patient repertoires were also found in the dengue patient repertoires (see 3.3), although convergence across both CDRH3 and exact V-gene usage was only observed for one motif. While SHM is not included in the CDRH3 topologies, examination of the convergent sequences indicated that they all show low SHM (0-4%).

Fig 5. SARS-CoV-2 convergent clonotypes to mAb DH1149 in the 3_6_14 bin. An 80% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are published anti-COVID-19 antibodies from 7 different studies, dark gray tips are antibody sequences from a healthy donor BCR repertoire, and orange through brown shaded tips are antibody sequences from COVID-19 patient BCR repertoires.

3.2 anti-HIV-1 antibody datasets display idiosyncrasies

We collated anti-HIV-1 monoclonal antibodies from two databases and one study [31–33], and compared this set to published bulk BCR repertoires of eight HIV-1 patients. From the databases we visualized 358 sequences. For the HIV-1 patient bulk BCR repertoires, 33,855 antibody motifs from patient MT1214 [27], 14,644 from patient NIH45 [29], and 184,294 from six CAPRISA patients [7] were visualized after removing duplicate sequences (see 2.2; S4 Fig). Although patient NIH45 was only sequenced for IGHV1 genes, this partial ‘repertoire’ was included because it is the original source of the broadly neutralizing antibody VRC01 [29].

The heatmap visualizations in AIRRscape were used to compare published anti-HIV-1 antibodies to anti-SARS-CoV-2 antibodies and to two HIV-1 patient repertoires. While the number of anti-HIV-1 antibodies was sparse relative to that of anti-SARS-CoV-2 antibodies, they appeared to have similar heavy chain V-gene + J-gene usage (Fig 2A). There were noticeable differences in both SHM and CDRH3 length, consistent with previous anti-HIV-1 antibody repertoire characterizations [39,40]. The anti-HIV-1 antibody dataset showed higher SHM (mean 12.6% vs. 2.3%, Fig 2A) and longer CDRH3 lengths (20.4 vs. 15.8 aa) than the CoV-AbDab dataset. Comparing published anti-HIV-1 antibodies with the HIV-1 patient repertoires also revealed higher SHM (mean 12.6% vs. 5.0%) and longer CDRH3 lengths (mean 20.4 vs. 16.0 aa), which contrasts with the COVID-19 results in 3.1 (S4 Fig) and indicates strong diversifying selection away from unmutated, naive antibodies.

We used AIRRscape to visualize convergent clonotypes among anti-HIV-1 antibodies and HIV-1 patient bulk BCR repertoires, focusing on ten recently reported convergences from Setliff [7], visualized as eight bins, as well as the inferred VRC01 germline CDR3 motif [29]. As per 3.1, we selected one published monoclonal antibody within each bin and examined the topology of similar CDRH3 motifs among all the HIV-1 datasets, this time using a lower 70% aa

sequence identity threshold and requiring identical V-gene assignment, following Setliff et al. [7]. Four of the bins showed limited convergence among HIV-1 datasets (Fig 6; S5 Fig), with motifs shared among two, three, or four patient repertoires. Despite including two antibody datasets from studies on the VRC01 antibody lineage, we notably could not find any convergence to the VRC01 germline CDR3 motif among the HIV-1 datasets. The complex evolution from germline and mutational requirements of neutralizing HIV-1 antibodies are consistent with the rarity of success in naturally achieving humoral neutralization in infected individuals and with challenges encountered in vaccine development.

Fig 6. HIV-1 convergent clonotypes to antibodies from Setliff (2018; Fig 4). (A) Convergent clonotypes to mAb 02-o in the 1_4_13 bin. (B) Convergent clonotypes to mAb 02-s in the 1_4_14 bin. (C) Convergent clonotypes to mAb HK20 in the 1_3_15 bin. A 70% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are published anti-COVID-19 antibodies, and green shaded tips are antibody sequences from HIV-1 patient BCR repertoires.

3.3 Dengue antibody datasets share convergent motifs

AIRRscape was used to visualize both reported and potentially undiscovered convergence among the collated dengue datasets. We searched for convergence between 38 unique antibody sequences from plasmablasts isolated from two Colombian dengue patients and bulk BCR repertoires of dengue patients from Colombian and Nicaraguan cohorts. Patient bulk BCR repertoires consist of one Colombian patient [d13; 17] and a collection of 45 Nicaraguan patient repertoires [18]. After removing duplicate sequences we compared 32,495 antibody motifs from Patient d13 and 198,119 from the Nicaraguan cohort (S6 Fig).

Evidence of convergent antibody motifs was previously found in independent studies of the Colombian (n = 2) and Nicaraguan (n = 45) patients. Plasmablast sequences in the Colombian patients were reported to cluster into 15 clonal families, with one clonal family (CF1) found in both patients [19]. Similarly, analysis of the Nicaraguan cohort found six CDRH3 motifs common to multiple patients [18].

In our analysis, we began with each of the 15 plasmablast clonal families, and searched for similar CDRH3 motifs among the full dengue dataset, using the 80% aa sequence identity threshold as per Soto et al. [6]. Five of the 15 clonal families showed convergence across both Colombian and Nicaraguan cohorts (Table 2; Fig 7; S7-9 Figs). Of these five, we identified CF1 from the Colombian cohort as the most prevalent clonal family in the Nicaraguan dataset, occurring in 16 donors (Table 2; Fig 7). We also found a previously unreported instance of convergence in the Colombian dataset. CDRH3 motifs from the CF6 clonal family identified in patient d20 were also found in the patient d13 bulk BCR repertoire, as well as in two patient repertoires from the Nicaraguan bulk dataset (Table 2; S7 Fig).

319 **Table 2. Dengue convergence to plasmablast clonal families.**

Clonal family	V_J_CDR3 length	Plasmablast donor	Colombian p13 Bulk matches?	Nicaraguan cohort matches?	Nicaraguan sequence matches	Nicaraguan donors with match
CF1	4_5_10	p13, p20	yes	yes	41	**16**
CF2	1_4_15	p13	yes	no		
CF3	1_5_16	p13	yes	no		
CF4	1_4_16	p20	no	no		
CF5	1_3_17	p13	yes	no		
CF6	1_4_13	p20	**yes**	yes	2	**2**
CF7	1_5_16	p13	yes	yes	24	**5**
CF8	1_5_17	p13	yes	yes	1	**1**
CF9	1_5_20	p13	yes	yes	7	**3**
CF10	3_5_14	p13	yes	no		
CF11	3_6_23	p13	yes	no		
CF12	3_4_17	p20	no	no		
CF13	4_5_18	p13	yes	no		
CF14	4_4_15	p20	no	no		
CF15	4_5_23	p13	yes	no		

320
321 **Fig 7. Dengue convergent clonotypes to CF1 (Zanini 2018).** An 80% identity threshold is
322 used to calculate convergence. Tips are colored by dataset source. Purple tips are plasmablast
323 sequences reported by Zanini (2018) isolated from two Colombian patients (d13 and d20), blue
324 tips are antibody sequences from the BCR repertoire of patient d13, and gold tips are antibody
325 sequences from a cohort of Nicaraguan patient BCR repertoires.
326

The six common CDRH3 motifs reported in the Nicaraguan dengue patients were also explored using AIRRscape. These six motifs group into two clusters by amino acid similarity. We found that the first cluster is common among the Nicaraguan dengue cohort, occurring in 17 patients, but not seen in the Colombian dataset. Notably, we found that the second cluster was convergent with the CF1 clonal family [19; S9 Fig], demonstrating its frequent occurrence. These data demonstrate how repertoire convergence in populations can occur at different geographic levels, some more restricted than others. This could be due to regional differences or similarities in factors such as infection or vaccine history, genetics, and environment.

4. Discussion

The importance of understanding BCR repertoires has gained increasing appreciation, especially as a result of the current COVID-19 pandemic. Along with rapidly advancing sequencing capabilities, this has led to the study and publication of dozens of individual BCR repertoires, particularly from patients infected by well-studied infectious pathogens as well as from healthy controls [e.g. 1,6,11,14,15,25,29]. As with other examples of ‘big data’, there are many challenges in managing these datasets. In response, the AIRR community has made significant efforts to standardize and curate BCR repertoires [20–22]. However, visualization and comparative analysis of these repertoires remains a challenge. To begin to address this gap we developed AIRRscape, a tool that leverages AIRR community standards to visualize multiple BCR repertoires quickly and simultaneously. AIRRscape provides a high-level view of large datasets that allows researchers to straightforwardly investigate differences in commonly used repertoire characteristics, such as V-gene usage, J-gene usage, CDR3 length, and somatic hypermutation.

Recent studies have found measurable convergence between individual BCR repertoires, particularly among individuals infected with a common antigen [2]. Such convergence has great implications for understanding immune responses to antigens and informing vaccine design. Therefore, a major aim of AIRRscape is to visualize related antibody sequences, particularly from different repertoires, using the commonly accepted characteristics of convergence, V-gene assignment, J-gene assignment, and CDR3 motif. We use phylogenetic methods to more easily visualize and understand differences among the motifs, keeping in mind that the visualizations are topologies of CDR3 motifs and not necessarily phylogenetic trees indicating common ancestry, e.g. when multiple individual repertoires are compared. The characteristics used to find convergence are the same as those used to define clonotypes, or clonal clusters of antibodies within individuals. Given that no consensus on clonotype definitions has been reached, with debate largely centered around thresholds of sequence identity in the CDR3 motif, AIRRscape creates topologies based on CDR3 amino acid sequences while being agnostic with respect to thresholds. Users may currently select one of four amino acid identity thresholds, three of which are used in recent large-scale studies of BCR repertoires [1,6,7]. AIRRscape also does not require *a priori* clustering of clonotypes in the input datasets. Users with pre-defined clonal families can employ another tool (Olmsted: https://olmstedviz.org) to generate such visualizations.

We validated the utility of AIRRscape by exploring datasets of both antibodies and patient bulk BCR repertoires for three viral pathogens: SARS-CoV-2, HIV-1, and DENV. Among the COVID-19 datasets, we first visually confirmed that the set of known anti-SARS-CoV-2 antibodies is broadly similar both to a healthy BCR repertoire and to a collection of COVID-19 patient repertoires, as measured by heavy chain V-gene + J-gene usage, CDRH3 length distribution, and relatively low levels of SHM from germline. This is consistent with studies made in the first months of the COVID-19 pandemic, which concluded that anti-SARS-CoV-2 antibodies could likely be induced by vaccines [41]. We then used AIRRscape to examine convergence among anti-SARS-CoV-2 antibodies and COVID-19 patient BCR repertoires as reported by multiple studies [14–16]. With an amino acid sequence identity threshold of 80%, we find that the COVID-19 convergent motifs are indeed present in currently known anti-SARS-CoV-2 antibodies as well as in four COVID-19 patient bulk BCR repertoires. COVID-19 convergent motifs were also identified among repertoires of healthy controls and dengue patients, which would not be expected to contain antibodies against SARS-CoV-2, but could be explained by the relative proximity of the motifs to germline sequences and/or to infection with other coronaviruses. Convergence of neutralizing antibody sequences among multiple COVID-19 repertoires is a strong indicator of similarity in SARS-CoV-2 immune responses and suggests that vaccines eliciting these antibodies will be broadly effective [15,42]. The discovery that these antibody sequences are public (i.e. also identified in our dengue and healthy control samples) suggests that there is a general pre-existing foundation in populations towards COVID-19 protection. Such a primed immune repertoire bodes well for robust vaccine responses. Repertoires of patients infected by different SARS-CoV-2 variants could be explored using AIRRscape to find potential convergence among these datasets. Such convergence would indicate antibody motifs with potential to neutralize the range of known SARS-CoV-2 variants, which could be prioritized for reverse vaccinology research or development.

In contrast to the COVID-19 datasets, the HIV-1 datasets are mostly idiosyncratic but do show limited convergence. Using AIRRscape, we first visualized characteristics of anti-HIV-1 antibodies and eight HIV-1 patient bulk BCR repertoires. The heatmaps show high SHM, a common feature of neutralizing antibodies against HIV-1. Notably, this collection of antibodies displays much higher SHM levels than anti-SARS-CoV-2 antibodies. We then searched for convergence in CDRH3 motifs between anti-HIV-1 antibodies and eight HIV-1 patient bulk BCR repertoires, using a more permissive threshold of 70% CDR3 sequence identity. Still, we found limited convergence among the HIV-1 datasets and no convergence to the VRC01 germline CDR3 motif [29]. While the limited convergence appears at odds with the reporting of VRC01-like antibodies in multiple HIV-1 patients [7,43], the VRC01 class varies in CDRH3 similarity and is instead defined by 20-35% SHM in the germline VH1-2 gene and a 5 aa CDRL3 length [44,45]. These idiosyncratic characteristics underscore the difficulty in vaccine development for HIV-1.

Examination of dengue datasets revealed evidence of convergent antibody motifs among two patient cohorts from Colombia and Nicaragua. Visualizing 15 focal plasmablast clonal families in AIRRscape, we found five were common to both patient cohorts, often in multiple patients. The most common convergent lineage CF1, seen in both Parameswaran [18]

and Zanini [19], was found not to be broadly neutralizing against all DENV serotypes [17]. However, the second most commonly found lineage, CF7, includes the mAb J9 that Durham et al. [17] found to be broadly neutralizing. That one of two tested convergent antibody lineages shows broad neutralization suggests further investigations of convergence may be beneficial for focusing antibody discovery and vaccine design. Broad neutralization could be particularly important for treating dengue, where antibody-enhanced disease is a problem.

As AIRRscape is in active development, we note some limitations and also envision future enhancements. AIRRscape is primarily an exploratory tool focused on visualization but various metrics could be used to formally measure overlap between repertoires [RDI; 46,sumrep; 47]. AIRRscape does not visualize all of the sequences in the bulk BCR repertoire datasets; due to size and computing limitations in the Shiny application, large datasets are collapsed when V-gene + J-gene assignments as well as CDR3 amino acid motifs are identical. Similarly, to facilitate construction of topologies of CDR3 amino acid motifs we focus on those with exact same lengths, although it has been shown that functionally similar antibodies may have different CDR3 lengths [48,49]. While the variability of CDR3 is typically crucial for defining antigen specificity [37,50], other CDRs can also be important in antigen recognition; in those cases, AIRRscape could be modified to visualize changes outside of CDR3. With respect to identifying convergence, AIRRscape creates bins based on assigned germline V-gene families; thus, CDR3 topologies can indicate convergence even if antibodies have different germline V-genes within a family. While this definition of convergence is intended to cluster closely related germline genes such as IGHV3-53 & IGHV3-66, it may be too broad in some cases. As an aid, we include the full V-gene assignments in the topology tip names, and we note that restricting the antibody list to one gene will limit searches to only convergences with matching V-genes. Lastly, although AIRRscape can quickly and easily search for known antibody sequences, examine overall repertoire patterns, and find convergence, users currently cannot initiate searches for identical motifs across datasets.

Despite these current limitations, AIRRscape is a useful tool for quick and thorough exploration of large BCR repertoire datasets. We have released the source code on GitHub so that interested research groups may use other datasets or incorporate more features. While we have loaded datasets from one healthy control and three diseases, we have designed the pre-processing R script so that any AIRR-compliant datasets may be visualized. Similarly, the code could be further modified to meet the desires of immuno-informaticians, virologists, and immunologists, for instance with respect to the CDR3 sequence identity thresholds, phylogenetic tree options, or color and visualization schemes. While our focus is on infectious disease, AIRRscape can be similarly used for autoimmunity and cancer biology, for example to compare healthy vs. autoimmune or cancer patient repertoires. We hope that AIRRscape will be useful for the research community and that interested parties will take advantage of its open source nature to contribute improvements.

Acknowledgments

We would like to acknowledge members of the CZ Biohub Data Science platform for thoughtful discussion, and Sandra L. Schmid for comments on the manuscript.

Author Contributions

E.W. and J.E.P. conceived of the project. E.W. designed and wrote software. E.W. performed bioinformatic analysis. E.W. and S.N. documented and implemented code repository. E.W. wrote the first draft of the manuscript. E.W., S.N., J.W., K.M.M., and J.E.P revised the draft and approved the final manuscript.

Data availability statement

All AIRRscape code and data loaded for visualization, as well as code for processing similar datasets are available on a GitHub repository at https://github.com/czbiohub/AIRRscape.

References

1. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature. 2019;566: 393–397. doi:10.1038/s41586-019-0879-y
2. Raybould MIJ, Rees AR, Deane CM. Current strategies for detecting functional convergence across B-cell receptor repertoires. mAbs. 2021;13: 1996732. doi:10.1080/19420862.2021.1996732
3. Chaudhary N, Wesemann DR. Analyzing Immunoglobulin Repertoires. Frontiers in Immunology. 2018;9: 462. doi:10.3389/fimmu.2018.00462
4. Dunn-Walters D, Townsend C, Sinclair E, Stewart A. Immunoglobulin gene analysis as a tool for investigating human immune responses. Immunological Reviews. 2018;284: 132–147. doi:10.1111/imr.12659
5. Wardemann H, Busse CE. Novel Approaches to Analyze Immunoglobulin Repertoires. Trends in Immunology. 2017;38: 471–482. doi:10.1016/j.it.2017.05.003
6. Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, et al. High frequency of shared clonotypes in human B cell receptor repertoires. Nature. 2019;566: 398–402. doi:10.1038/s41586-019-0934-8
7. Setliff I, McDonnell WJ, Raju N, Bombardi RG, Murji AA, Scheepers C, et al. Multi-Donor Longitudinal Antibody Repertoire Sequencing Reveals the Existence of Public Antibody Clonotypes in HIV-1 Infection. Cell Host & Microbe. 2018;23: 845-854.e6. doi:10.1016/j.chom.2018.05.001
8. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. Front Immunol. 2018;9. doi:10.3389/fimmu.2018.00224
9. Rubelt F, Bolen CR, McGuire HM, Heiden JAV, Gadala-Maria D, Levin M, et al. Individual heritable differences result in unique cell lymphocyte receptor repertoires of naïve and antigen-experienced cells. Nature Communications. 2016;7: 11112. doi:10.1038/ncomms11112
10. Muggen AF, de Jong M, Wolvers-Tettero ILM, Kallemeijn MJ, Teodósio C, Darzentas N, et al. The presence of CLL-associated stereotypic B cell receptors in the normal BCR repertoire from healthy individuals increases with age. Immun Ageing. 2019;16: 22. doi:10.1186/s12979-019-0163-x
11. Schultheiß C, Paschold L, Simnica D, Mohme M, Willscher E, Wenserski L von, et al. Next-Generation Sequencing of T and B Cell Receptor Repertoires from COVID-19 Patients Showed Signatures Associated with Severity of Disease. Immunity. 2020;53: 442-455.e4. doi:10.1016/j.immuni.2020.06.024

12. Horns F, Vollmers C, Croote D, Mackey SF, Swan GE, Dekker CL, et al. Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. Neher RA, editor. eLife. 2016;5: e16578. doi:10.7554/eLife.16578
13. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J. shiny: Web Application Framework for R. 2020. Available: https://CRAN.R-project.org/package=shiny
14. Galson JD, Schaetzle S, Bashford-Rogers RJM, Raybould MIJ, Kovaltsuk A, Kilpatrick GJ, et al. Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. bioRxiv. 2020; 2020.05.20.106294. doi:10.1101/2020.05.20.106294
15. 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;584: 437–442. doi:10.1038/s41586-020-2456-9
16. Yuan M, Liu H, Wu NC, Lee C-CD, Zhu X, Zhao F, et al. Structural basis of a shared antibody response to SARS-CoV-2. Science. 2020;369: 1119–1123. doi:10.1126/science.abd2321
17. Durham ND, Agrawal A, Waltari E, Croote D, Zanini F, Fouch M, et al. Broadly neutralizing human antibodies against dengue virus identified by single B cell transcriptomics. eLife. 2019;8: e52384. doi:10.7554/eLife.52384
18. Parameswaran P, Liu Y, Roskin KM, Jackson KKL, Dixit VP, Lee J-Y, et al. Convergent Antibody Signatures in Human Dengue. Cell Host & Microbe. 2013;13: 691–700. doi:10.1016/j.chom.2013.05.008
19. Zanini F, Robinson ML, Croote D, Sahoo MK, Sanz AM, Ortiz-Lasso E, et al. Virus-inclusive single-cell RNA sequencing reveals the molecular signature of progression to severe dengue. PNAS. 2018;115: E12363–E12369. doi:10.1073/pnas.1813819115
20. Breden F, Luning Prak ET, Peters B, Rubelt F, Schramm CA, Busse CE, et al. Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. Frontiers in Immunology. 2017;8: 1418. doi:10.3389/fimmu.2017.01418
21. Rubelt F, Busse CE, Bukhari SAC, Bürckert J-P, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat Immunol. 2017;18: 1274–1278. doi:10.1038/ni.3873
22. 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. Immunological Reviews. 2018;284: 24–41. doi:10.1111/imr.12666
23. Kovaltsuk A, Leem J, Kelm S, Snowden J, Deane CM, Krawczyk K. Observed Antibody Space: A Resource for Data Mining Next-Generation Sequencing of Antibody Repertoires. The Journal of Immunology. 2018;201: 2502–2509. doi:10.4049/jimmunol.1800708
24. Olsen TH, Boyles F, Deane CM. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. Protein Science. 2021;2021: 1–6. doi:10.1002/pro.4205
25. Kuri-Cervantes L, Pampena MB, Meng W, Rosenfeld AM, Ittner CAG, Weisman AR, et al. Comprehensive mapping of immune perturbations associated with severe COVID-19. Science Immunology. 2020;5: eabd7114. doi:10.1126/sciimmunol.abd7114
26. 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;28: 516-525.e5. doi:10.1016/j.chom.2020.09.002
27. Waltari E, Jia M, Jiang CS, Lu H, Huang J, Fernandez C, et al. 5′ Rapid Amplification of cDNA Ends and Illumina MiSeq Reveals B Cell Receptor Features in Healthy Adults, Adults With Chronic HIV-1 Infection, Cord Blood, and Humanized Mice. Frontiers in Immunology. 2018;9: 628. doi:10.3389/fimmu.2018.00628
28. Waltari E, McGeever, Aaron, Friedland, Natalia, Kim PS, McCutcheon KM. Functional

557 enrichment and analysis of Antigen-Specific memory B cell antibody repertoires in PBMCs.
558 Frontiers in immunology. 2019;10: 1452. doi:10.3389/fimmu.2019.01452
559 29. Wu X, Zhang Z, Schramm CA, Joyce MG, Do Kwon Y, Zhou T, et al. Maturation and
560 Diversity of the VRC01-Antibody Lineage over 15 Years of Chronic HIV-1 Infection. Cell.
561 2015;161: 470–485. doi:10.1016/j.cell.2015.03.004
562 30. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. CoV-AbDab: the coronavirus antibody
563 database. Bioinformatics. 2020;37: 734–735. doi:10.1093/bioinformatics/btaa739
564 31. Mahajan S, Vita R, Shackelford D, Lane J, Schulten V, Zarebski L, et al. Epitope Specific
565 Antibodies and T Cell Receptors in the Immune Epitope Database. Front Immunol. 2018;9.
566 doi:10.3389/fimmu.2018.02688
567 32. Yoon H, Macke J, West AP, Foley B, Bjorkman PJ, Korber B, et al. CATNAP: a tool to
568 compile, analyze and tally neutralizing antibody panels. Nucleic Acids Res. 2015;43:
569 W213–W219. doi:10.1093/nar/gkv404
570 33. Yacoob C, Pancera M, Vigdorovich V, Oliver BG, Glenn JA, Feng J, et al. Differences in
571 Allelic Frequency and CDRH3 Region Limit the Engagement of HIV Env Immunogens by
572 Putative VRC01 Neutralizing Antibody Precursors. Cell Reports. 2016;17: 1560–1570.
573 doi:10.1016/j.celrep.2016.10.017
574 34. Vander Heiden JA, Yaari G, Uduman M, Stern JN, O'Connor KC, Hafler DA, et al.
575 pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte
576 receptor repertoires. Bioinformatics. 2014;30: 1930–1932.
577 doi:10.1093/bioinformatics/btu138
578 35. Gupta NT, Vander Heiden JA, Uduman M, Gadala-Maria D, Yaari G, Kleinstein SH.
579 Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing
580 data. Bioinformatics. 2015;31: 3356–3358. doi:10.1093/bioinformatics/btv359
581 36. Zhou JQ, Kleinstein SH. Cutting Edge: Ig H Chains Are Sufficient to Determine Most B Cell
582 Clonal Relationships. The Journal of Immunology. 2019;203: 1687–1692.
583 doi:10.4049/jimmunol.1900666
584 37. Xu JL, Davis MM. Diversity in the CDR3 Region of VH Is Sufficient for Most Antibody
585 Specificities. Immunity. 2000;13: 37–45. doi:10.1016/S1074-7613(00)00006-6
586 38. Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev
587 Genet. 2003;4: 275–284. doi:10.1038/nrg1044
588 39. Scheid JF, Mouquet H, Feldhahn N, Seaman MS, Velinzon K, Pietzsch J, et al. Broad
589 diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals.
590 Nature. 2009;458: 636–640. doi:10.1038/nature07930
591 40. Sok D, Burton DR. Recent progress in broadly neutralizing antibodies to HIV. Nat Immunol.
592 2018;19: 1179–1188. doi:10.1038/s41590-018-0235-7
593 41. Kreer C, Zehner M, Weber T, Ercanoglu MS, Gieselmann L, Rohde C, et al. Longitudinal
594 Isolation of Potent Near-Germline SARS-CoV-2-Neutralizing Antibodies from COVID-19
595 Patients. Cell. 2020;182: 843-854.e12. doi:10.1016/j.cell.2020.06.044
596 42. Kreer C, Gruell H, Mora T, Walczak AM, Klein F. Exploiting B Cell Receptor Analyses to
597 Inform on HIV-1 Vaccination Strategies. Vaccines. 2020;8: 13.
598 doi:10.3390/vaccines8010013
599 43. Zhou T, Zhu J, Wu X, Moquin S, Zhang B, Acharya P, et al. Multidonor Analysis Reveals
600 Structural Elements, Genetic Determinants, and Maturation Pathway for HIV-1
601 Neutralization by VRC01-Class Antibodies. Immunity. 2013;39: 245–258.
602 doi:10.1016/j.immuni.2013.04.012
603 44. West AP, Diskin R, Nussenzweig MC, Bjorkman PJ. Structural basis for germ-line gene
604 usage of a potent class of antibodies targeting the CD4-binding site of HIV-1 gp120.
605 PNAS. 2012;109: E2083–E2090. doi:10.1073/pnas.1208984109
606 45. Wu X, Kong X-P. Antigenic landscape of the HIV-1 envelope and new immunological
607 concepts defined by HIV-1 broadly neutralizing antibodies. Current Opinion in Immunology.

2016;42: 56–64. doi:10.1016/j.coi.2016.05.013
46. Bolen CR, Rubelt F, Vander Heiden JA, Davis MM. The Repertoire Dissimilarity Index as a method to compare lymphocyte receptor repertoires. BMC Bioinformatics. 2017;18: 155. doi:10.1186/s12859-017-1556-5
47. 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;10. doi:10.3389/fimmu.2019.02533
48. Barnes CO, Jette CA, Abernathy ME, Dam K-MA, Esswein SR, Gristick HB, et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. Nature. 2020; 1–6. doi:10.1038/s41586-020-2852-1
49. Barnes CO, West AP, Huey-Tubman KE, Hoffmann MAG, Sharaf NG, Hoffman PR, et al. Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. Cell. 2020;182: 828-842.e16. doi:10.1016/j.cell.2020.06.025
50. Greiff V, Bhat P, Cook SC, Menzel U, Kang W, Reddy ST. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. Genome Medicine. 2015;7: 49. doi:10.1186/s13073-015-0169-8

Figures

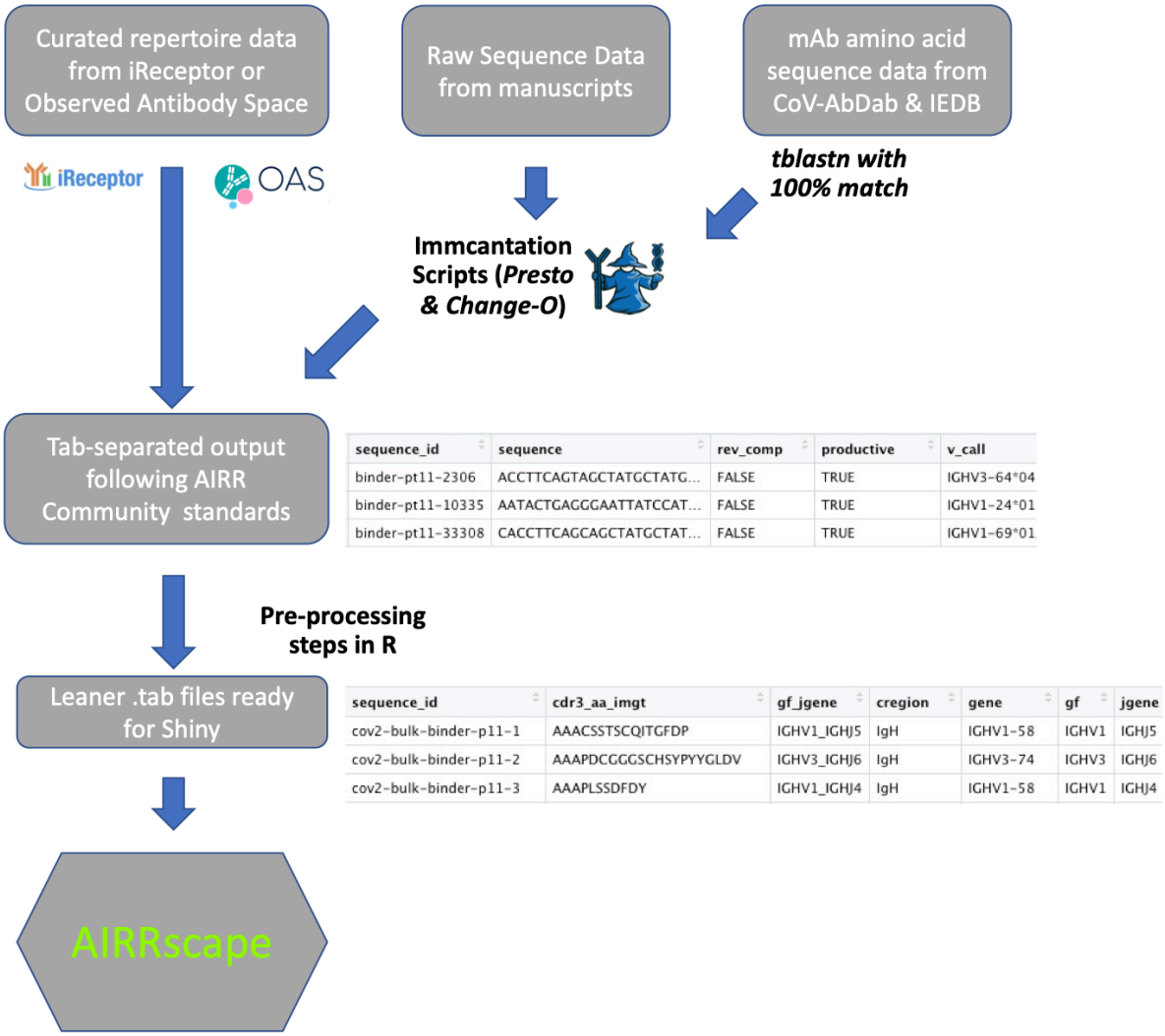

Fig 1. Workflow of repertoire data retrieval and processing.

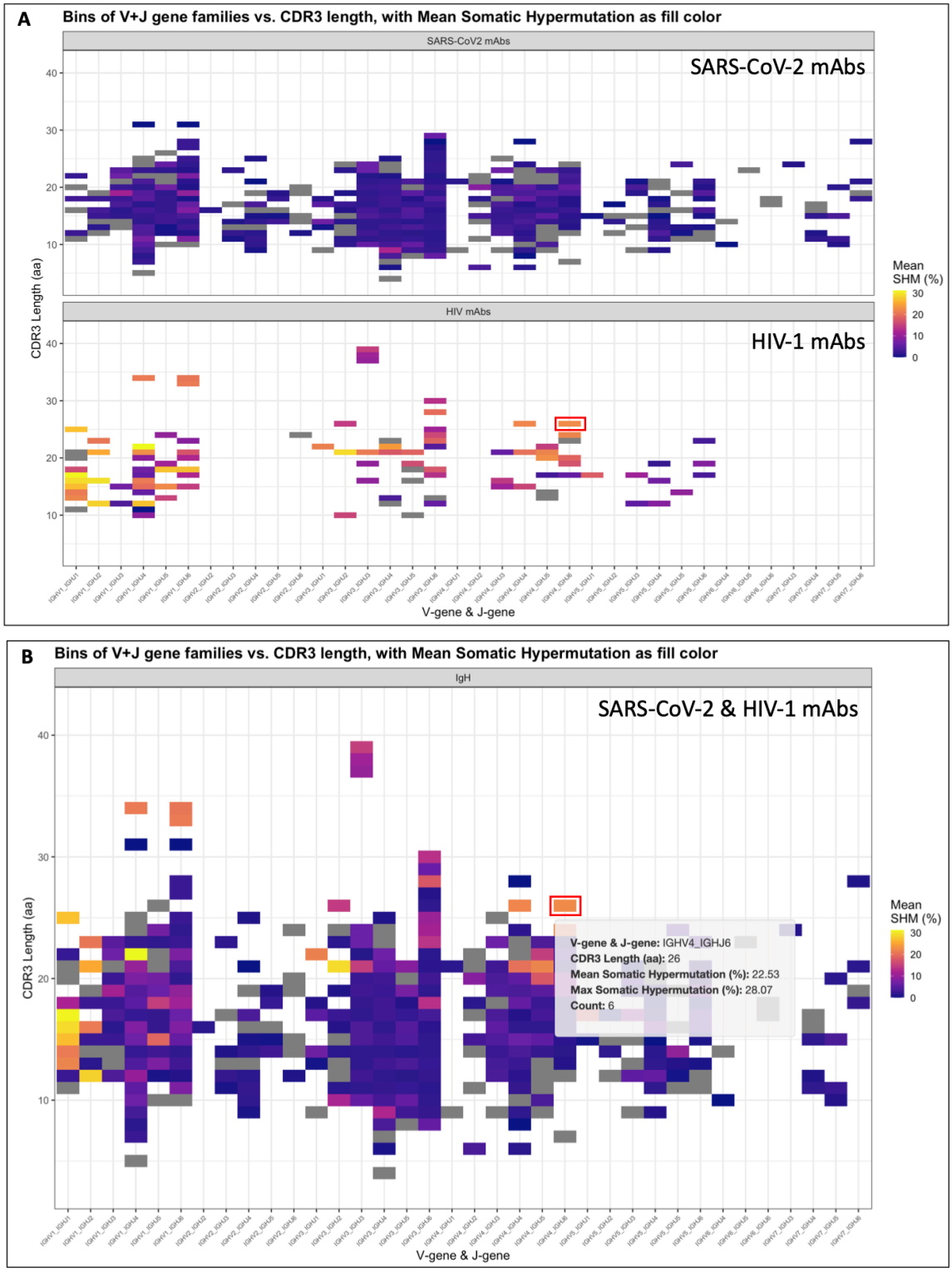

Fig 2. AIRRscape visualization of immune repertoires. Heatmaps comparing characteristics of (A) separate and (B) combined datasets for anti-SARS-CoV-2 and anti-HIV-1 antibodies.

Fig 3. AIRRscape interface showing antibodies and CDR3 amino acid topology of selected antibody bin. Selected bin is highlighted in the red box of Fig 2B.

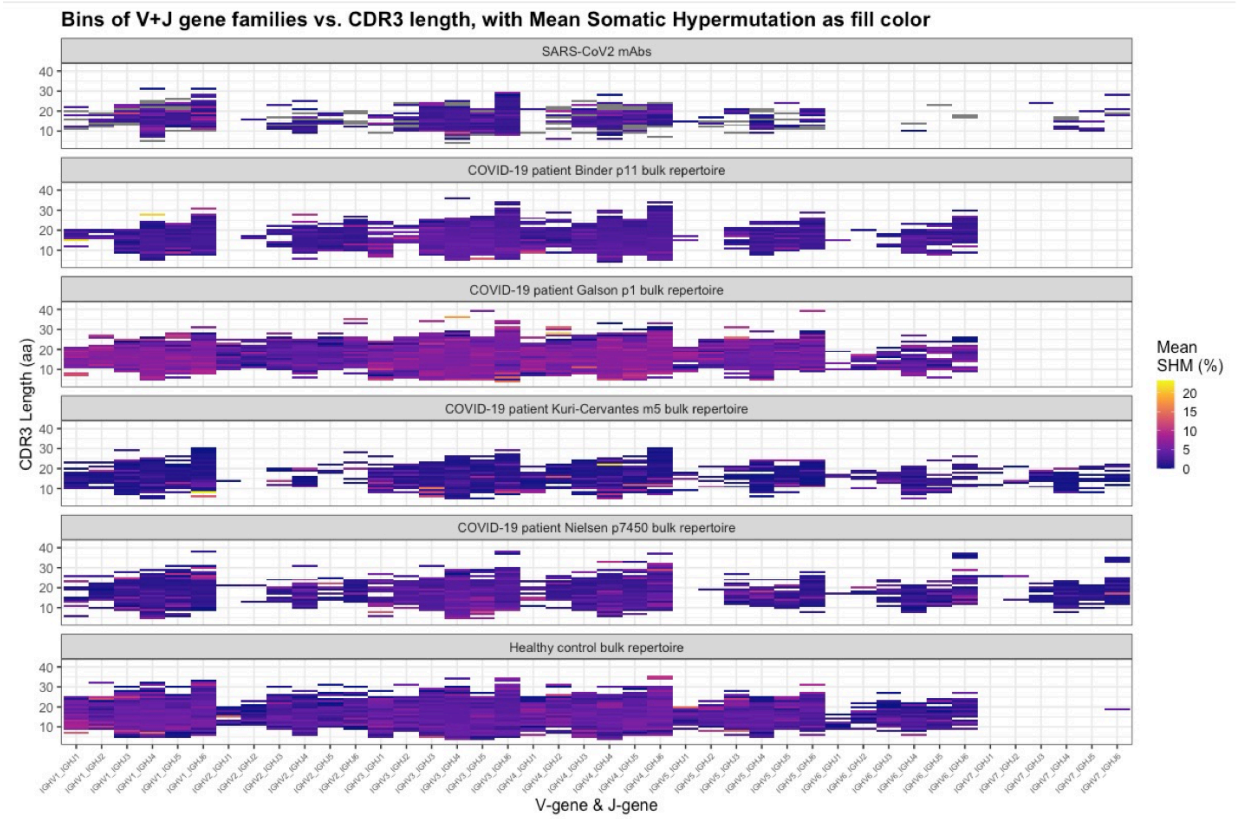

Fig 4. AIRRscape heatmaps comparing anti-SARS-CoV-2 antibodies, bulk BCR repertoires of four COVID-19 patients, and a healthy control bulk BCR repertoire.

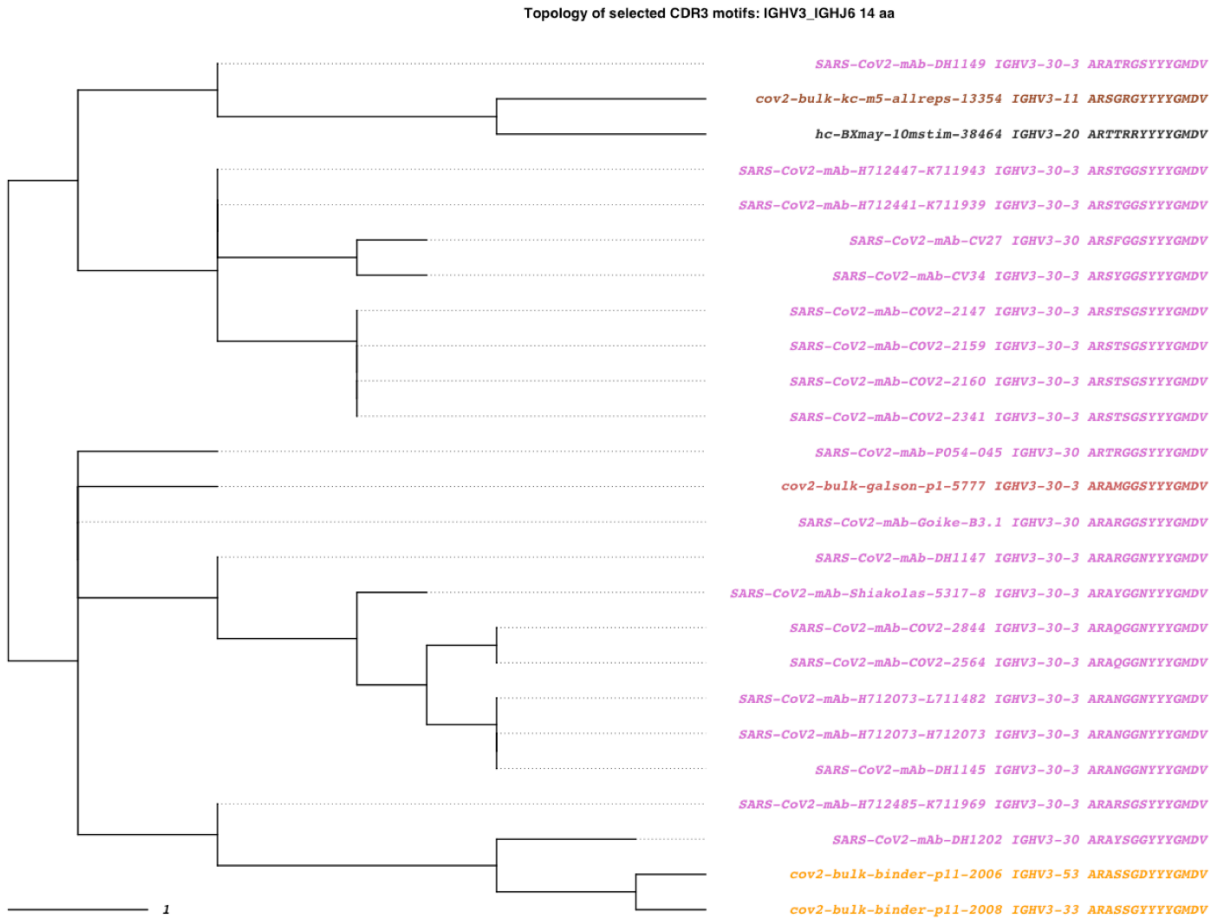

Fig 5. SARS-CoV-2 convergent clonotypes to mAb DH1149 in the 3_6_14 bin. An 80% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are published anti-COVID-19 antibodies from 7 different studies, dark gray tips are antibody sequences from a healthy donor BCR repertoire, and orange through brown shaded tips are antibody sequences from COVID-19 patient BCR repertoires.

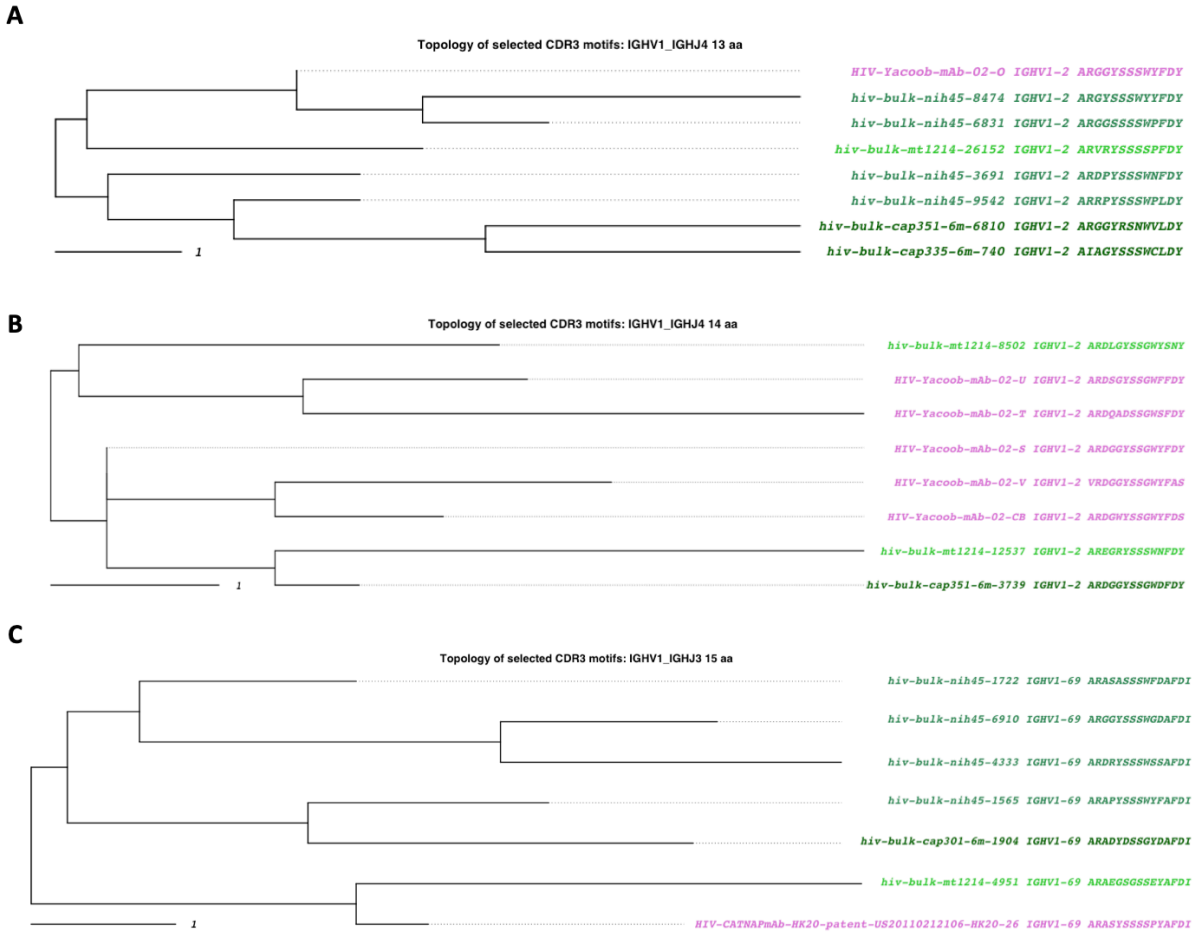

Fig 6. HIV-1 convergent clonotypes to antibodies from Setliff (2018; Fig 4). (A) Convergent clonotypes to mAb 02-o in the 1_4_13 bin. (B) Convergent clonotypes to mAb 02-s in the 1_4_14 bin. (C) Convergent clonotypes to mAb HK20 in the 1_3_15 bin. A 70% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are published anti-HIV-1 antibodies, while green shaded tips are antibody sequences from HIV-1 patient BCR repertoires.

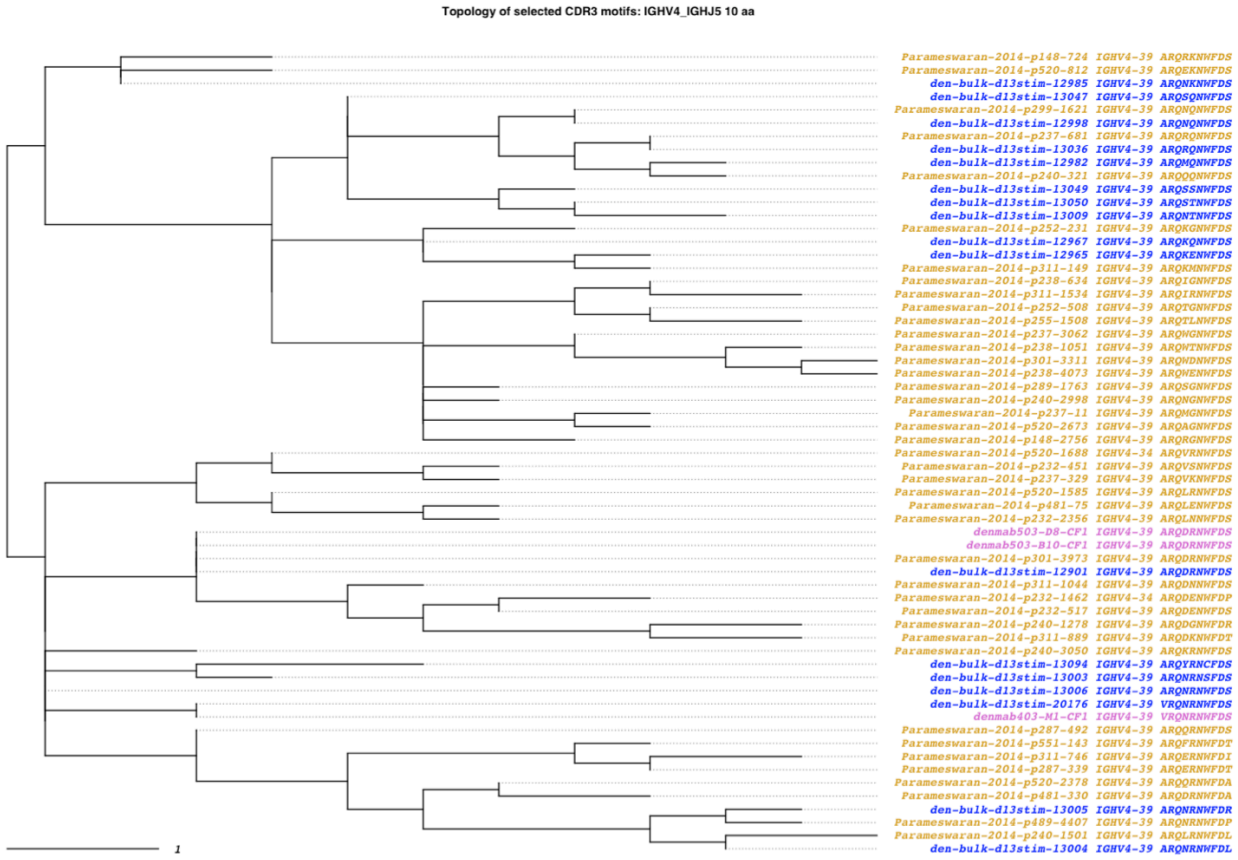

Fig 7. Dengue convergent clonotypes to CF1 (Zanini 2018). An 80% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are plasmablast sequences reported by Zanini (2018) isolated from two Colombian patients (d13 and d20), blue tips are antibody sequences from the BCR repertoire of patient d13, and gold tips are antibody sequences from a cohort of Nicaraguan patient BCR repertoires.

Supporting Information

S1 Fig. SARS-CoV-2 convergent clonotypes to mAb C102 in the 3_4_11 bin. An 80% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are published anti-COVID-19 antibodies from 12 different studies, dark gray tips are antibody sequences from a healthy donor BCR repertoire, and orange through brown shaded tips are antibody sequences from COVID-19 patient BCR repertoires.

S2 Fig. SARS-CoV-2 convergent clonotypes to mAb C125 in the 1_3_16 bin. An 80% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are published anti-COVID-19 antibodies from 7 different studies, dark gray tips are antibody sequences from a healthy donor BCR repertoire, and orange through brown shaded tips are antibody sequences from COVID-19 patient BCR repertoires.

S3 Fig. SARS-CoV-2 convergent clonotypes to mAb BD-498 in the 3_6_11 bin. An 80% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are published anti-COVID-19 antibodies from 15 different studies, dark gray tips are antibody sequences from a healthy donor BCR repertoire, and orange through brown shaded tips are antibody sequences from COVID-19 patient BCR repertoires.

S4 Fig. AIRRscape heatmaps comparing anti-HIV-1 antibodies and bulk BCR repertoires of eight HIV-1 patients.

S5 Fig. HIV-1 convergent clonotypes to mAb CAP351_6m_6041 (Setliff 2018; Fig 3) in the 1_6_14 bin. A 70% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are published anti-COVID-19 antibodies, and green shaded tips are antibody sequences from HIV-1 patient BCR repertoires.

S6 Fig. AIRRscape heatmaps comparing isolated dengue plasmablasts and bulk BCR repertoires of dengue patients from Colombian and Nicaraguan cohorts.

S7 Fig. Dengue convergent clonotypes to CF6 (Zanini 2018). An 80% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are plasmablast sequences reported by Zanini (2018) isolated from two Colombian patients (d13 and d20), blue tips are antibody sequences from the BCR repertoire of patient d13, and gold tips are antibody sequences from a cohort of Nicaraguan patient BCR repertoires.

S8 Fig. Dengue convergent clonotypes to CF7 (Zanini 2018). An 80% identity threshold is used to calculate convergence. Tips are colored by dataset source. Purple tips are plasmablast sequences reported by Zanini (2018) isolated from two Colombian patients (d13 and d20), blue tips are antibody sequences from the BCR repertoire of patient d13, and gold tips are antibody sequences from a cohort of Nicaraguan patient BCR repertoires.

668 **S9 Fig. Dengue convergent clonotypes to Parameswaran (2018) motif ARQIGNWFDP**
669 **similar to CF1 (Zanini 2018).** An 80% identity threshold is used to calculate convergence. Tips
670 are colored by dataset source. Purple tips are plasmablast sequences reported by Zanini (2018)
671 isolated from two Colombian patients (d13 and d20), blue tips are antibody sequences from the
672 BCR repertoire of patient d13, and gold tips are antibody sequences from a cohort of
673 Nicaraguan patient BCR repertoires.
674