Dynamics of B-cell repertoires and emergence of cross-reactive responses in COVID-19 patients with different disease severity

Zachary Montague[1*], Huibin Lv[2*], Jakub Otwinowski[3], William S. DeWitt[4,5], Giulio Isacchini[3,6], Garrick K. Yip[2], Wilson W. Ng[2], Owen Tak-Yin Tsang[7], Meng Yuan[8], Hejun Liu[8], Ian A. Wilson[8,9], J. S. Malik Peiris[2], Nicholas C. Wu[8], Armita Nourmohammad[1,3,5#], Chris Ka Pun Mok[2#]

1. Department of Physics, University of Washington, 3910 15th Ave Northeast, Seattle, WA 98195, USA
2. HKU-Pasteur Research Pole, School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.
3. Max Planck Institute for Dynamics and Self-organization, Am Faßberg 17, 37077 Göttingen, Germany
4. Department of Genome Sciences, University of Washington, 3720 15th Ave NE, Seattle, WA 98195, USA
5. Fred Hutchinson Cancer Research Center, 1100 Fairview Ave N, Seattle, WA 98109, USA
6. Laboratoire de physique de l'ecole normale supérieure (PSL University), CNRS, Sorbonne Université, and Université de Paris, 75005 Paris, France
7. Infectious Diseases Centre, Princess Margaret Hospital, Hospital Authority of Hong Kong
8. Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA
9. The Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

* equal contribution

\# equal contribution and corresponding authors: Armita Nourmohammad: armita@uw.edu, and Chris Ka Pun Mok: ch02mkp@hku.hk

Keywords: SARS-CoV-2, COVID-19, B-cell repertoires, cross-reactivity

Abstract

COVID-19 patients show varying severity of the disease ranging from asymptomatic to requiring intensive care. Although a number of monoclonal antibodies against SARS-CoV-2 have been identified, we still lack an understanding of the overall landscape of B-cell receptor (BCR) repertoires in COVID-19 patients. Here, we used high-throughput sequencing of BCR repertoires collected over multiple time points during an infection to characterize statistical and dynamical signatures of the B-cell response to SARS-CoV-2 in 19 patients with different disease severities. Based on principled statistical approaches, we determined differential sequence features of BCRs associated with different disease severity. We identified 34 significantly expanded rare clonal lineages shared among patients as candidates for a specific response to SARS-CoV-2. Moreover, we identified natural emergence of a BCR with cross-reactivity to SARS-CoV and SARS-CoV-2 in a number of patients. Overall, our results provide important insights for development of rational therapies and vaccines against COVID-19.

Introduction

The novel coronavirus SARS-CoV-2, which causes the severe respiratory disease COVID-19, has now spread to 216 countries and caused more than ten million infections with a mortality rate around 5% (WHO, 2020). COVID-19 patients show varying disease severity ranging from asymptomatic to requiring intensive care. While epidemiological and clinical data report that many factors such as age, gender, genetic background, and preexisting conditions are associated with disease severity, host immunity against the virus infection is the crucial component of controlling disease progression (Ellinghaus et al., 2020; Guan et al., 2020; McKechnie and Blish, 2020; Vabret et al., 2020; Wu et al., 2020a). Shedding light on signatures of a protective immune response against SARS-CoV-2 infections can help understand the nature of COVID-19 and guide the therapeutic developments as well as vaccine design and assessment.

Adaptive immunity is considered as one of the core protective mechanisms in humans against infectious diseases. A vast diversity of surface receptors on B- and T-cells enables us to recognize and counter new or repeated invasions from a multitude of pathogens (Janeway et al., 2005; Nielsen and Boyd, 2018). In particular, antibodies produced by B-cells can provide long-lasting protection against specific pathogens through neutralization or other antibody-mediated immune mechanisms (Janeway et al., 2005). During the early phase of an infection, antigens of a pathogen are recognized by a group of naïve B-cells, which then undergo affinity maturation in a germinal center through somatic hypermutation and selection. The B-cell receptors (BCRs) of mature B-cells can strongly react to infecting antigens, resulting in B-cell stimulation, clonal expansion, and ultimately secretion of high affinity antibodies in the blood (Burnet, 1959; Burnet, 1960; Cyster and Allen, 2019). The specificity of a BCR is determined by a number of features such as V-, (D-) or J-gene usage, length and sequence composition of the HCDR3 region. It has been found that SARS-CoV-2-specific IgG antibodies can be detected in plasma samples of COVID-19 patients starting from the first week after the onset of symptoms (Perera et al., 2020). These antibodies bind to different antigens including spike protein and nucleoprotein as well as other structural or non-structural proteins (Hachim et al., 2020). In addition, multiple studies have isolated SARS-CoV-2-specific B-cells from COVID-19 patients and determined their germline origin (Brouwer et al., 2020; Cao et al., 2020; Chi et al., 2020; Ju et al., 2020; Liu et al., 2020; Robbiani et al., 2020; Rogers et al., 2020; Seydoux et al., 2020a; Seydoux et al., 2020b; Wu et al., 2020b; Zost et al., 2020). However, we still lack a comprehensive view of the patients' entire BCR repertoires during SARS-CoV-2 infections.

Antibody repertoire sequencing has advanced our understanding of the diversity of adaptive immune repertoires and their response to pathogens (Boyd et al., 2009; Georgiou et al., 2014; Kreer et al., 2020; Robins, 2013). A few studies have performed BCR repertoire sequencing to characterize the statistical signatures of the immune response to SARS-CoV-2 (Galson et al., 2020; Nielsen et al., 2020;

Schultheiß et al., 2020). However, these studies have limited data on the dynamics of BCR repertoires, which could otherwise provide significant insight into specific responses to the infection. In this study, we have established a principled statistical approach to study the statistics and dynamics of BCR repertoires and to characterize the immune responses in 19 COVID-19 patients with different disease severities. Importantly, by combining information from the statistics of sequence features in BCR repertoires, the expanding dynamics of clonal lineages during infection, and sharing of BCRs among COVID-19 patients, we identified 34 clonal lineages that are potential candidates for a response to SARS-CoV-2. Moreover, we identified cross-reactive responses to SARS-CoV in some of the COVID-19 patients and a natural emergence of a previously isolated SARS-reactive antibody (Pinto et al., 2020) in three patients.

Results

B-cell repertoires differ in receptor compositions across cohorts. We obtained total RNA from the PBMC isolated from 19 patients infected with SARS-CoV-2 and three healthy individuals (see Methods and Tables S1, S2 for details). The patients showed different severity of symptoms, forming three categories of infected cohorts: 2 patients with mild symptoms, 12 patients with moderate symptoms, and 5 patients with severe symptoms. Specimens from all patients, except patient 19, were collected over two or more time points during the course of the infection. The sampled time points for all patients are indicated in Fig. 1 and Tables S1. IgG heavy chains of B-cell repertoires were sequenced by next-generation sequencing, and the statistics of the collected BCR read data from each sample are shown in Table S1. Statistical models were applied to analyze the length of the HCDR3 region, IGHV- or IGHJ-gene usage, and insertion/deletion patterns as well as the expansion of specific clonal lineages (Fig. 1).

We first aimed to investigate whether cohorts with different disease severities can be distinguished by molecular features of their B-cell repertoires. Since sequence features of immune receptors (e.g. HCDR3 length, V- or J-gene usage, and insertion/deletion patterns) are often associated with their binding specificity, we used statistical methods to compare these features both at the level of clonal lineages, including the inferred receptor sequence of lineage progenitors (Figs. 2, S2), and unique sequences (Figs. S1, S2) in a repertoire. Lineage progenitors of IgG repertoires are closest to the ensemble of naïve receptors in the periphery. Features of lineage progenitors reflect receptor characteristics that are necessary for activating and forming a clonal lineage in response to an infection. Statistics of unique sequences in a repertoire, on the other hand, contain information about the size of the circulating lineages. Importantly, these two statistical ensembles are relatively robust to PCR amplification biases that directly impact read abundances (see Methods for error correction and processing of sequence reads).

IGHV genes cover a large part of pathogen-engaging regions of BCRs, including the three complementarity-determining regions HCDR1, HCDR2, and a portion of HCDR3. Therefore, we investigated if there are any differences in V-gene usage across cohorts, which may indicate preferences relevant for response to a particular pathogen. We found that the variation in V-gene usage among individuals within each cohort was far larger than differences among cohorts (Fig. 2A). Data from unique sequences also indicated large background amplitudes due to vast differences in the sizes of lineages within a repertoire (Fig. S1). Similarly, IGHJ-gene usage was also comparable across different cohorts (Fig. S2). Our results suggest that the SARS-CoV-2 V-gene specific responses are highly individualized at the repertoire level.

HCDR3 is part of the variable chain of B-cell receptors and is often a crucial region in determining specificity. Importantly, HCDR3 is highly variable in its sequence content and length due to insertion and deletion of sequence fragments at the VD and DJ junctions of the germline receptor. Therefore, differential characteristics of the HCDR3 sequence in BCR repertoires of different cohorts can signal preferences for sequence features specific to a class of antigens. We find that HCDR3s in COVID-19 patients with moderate and severe symptoms are significantly longer than in the healthy control, both at the level of clonal lineages (Fig. 2B) and unique sequences in each repertoire (Fig. S1B). The difference between HCDR3s length in healthy individuals and patients with mild symptoms is insignificant (p-values for significance of these deviations are given in the corresponding Figure captions). Recent studies have also identified propensity for longer HCDR3s in BCR repertoires of COVID-19 patients (Galson et al., 2020; Nielsen et al., 2020; Schultheiß et al., 2020). We find that the longer HCDR3 length in patients with moderate and severe symptoms is primarily due to insertion of larger sequence fragments and also shorter deletions at the VD junction (Figs. 2C-D, S1C-D). Interestingly, the insertion/deletion patterns at the DJ junctions did not show a substantial difference across cohorts (Fig. S2C-F).

Longer sequence insertions in HCDR3 can introduce more sequence diversity at the repertoire level. Quantifying sequence diversity of a B-cell repertoire, however, is very sensitive to the sampling depth in each individual. Despite progress in the quality of high-throughput repertoire sequencing (RepSeq) techniques, sequenced BCRs still present a highly under-sampled view of the entire repertoire. To characterize the diversity of repertoires and the statistics of sequence features that make up this diversity, we inferred principled models of repertoire generation and selection for the entry of receptors into the periphery (Methods) (Elhanati et al., 2014; Marcou et al., 2018). To do so, we first used data from unproductive B-cell receptors to infer the highly non-uniform baseline model that characterizes the probability $P_{\mathrm{gen}}(\sigma)$ to generate a given receptor sequence (Elhanati et al., 2014; Marcou et al., 2018) (Fig. 1 and Methods). This baseline model reflects sequence biases in generating BCRs in the bone marrow.

The functional, yet pathogen-naïve BCRs that enter the periphery experience selection through processes known as central tolerance (Janeway et al., 2005). In addition, the inferred progenitors of clonal lineages in the IgG repertoire have undergone antigen-dependent selection that led to expansion of their clonal lineage in response to an infection. These two levels of selection make sequence features of functional lineage progenitors distinct from the pool of unproductive BCRs that reflects biases of the generation process prior to any selection. To identify these distinguishing sequence features, we inferred a selection model for lineage progenitors (Methods). We characterized the probability to observe a clonal lineage ancestor in the periphery as $P_{\text{post}}(\sigma) \sim P_{\text{gen}}(\sigma) e^{\Sigma_{f:\text{features}} q_f(\sigma)}$, which deviates from the inferred generation probability of the receptor $P_{\text{gen}}(\sigma)$ by selection factors $q_f(\sigma)$ (Isacchini et al., 2020a; Isacchini et al., 2020b; Sethna et al., 2020) . These selection factors $q_f(\sigma)$ depend on sequence features, including IGHV-gene and IGHJ-gene usages, HCDR3 length, and amino acid preferences at different positions in the HCDR3 (Methods) (Elhanati et al., 2014; Isacchini et al., 2020a; Isacchini et al., 2020b; Marcou et al., 2018; Sethna et al., 2020).

The distribution of the log-probability $\log_{10} P_{\text{post}}(\sigma)$ for the inferred progenitors of clonal lineages observed in individuals from different cohorts is shown in Fig. 3A. We find an overabundance of BCR lineages with progenitors that have a low probability of entering the periphery (i.e., a lower $P_{\text{post}}(\sigma)$) in COVID-19 patients compared to healthy individuals (Fig. 3A). Moreover, we estimated the diversity of the repertoires in each cohort by evaluating the entropy of receptor sequences generated by the respective repertoire models (see Methods). In particular, diverse repertoires that contain B-cell lineages with rare receptors (i.e., those with a lower $P_{\text{post}}(\sigma)$), should have larger entropies. Based on this analysis, we find that immune repertoires are more diverse in COVID-19 patients compared to healthy individuals (Fig. 3A and Methods). Specifically, the entropy (i.e., diversity) of BCR repertoires grew with severity of the disease, from 38.7 bits in the healthy cohort to 40.5 bits in the mild cohort, to 41.6 bits in the moderate cohort, and to 41.3 bits in the severe cohort (see Methods).

Selection factors $q_f(\sigma)$ determine the deviation in preferences for different sequence features of BCRs in each cohort, including their HCDR3 length and composition, and IGHV-gene usages. A comparison of selection factors among cohorts can characterize their distinctive sequence features. Consistent with our previous analysis of lineage characteristics in Fig. 2A, we find that selection factors associated with V-gene usage are strongly correlated across cohorts (Fig. 3B). However, we find a weaker correlation for selection factors associated with preference for HCDR3 length between healthy individuals and the cohorts of COVID-19 patients (Fig. 3B), which is consistent with our observation in Fig. 2B. Moreover, there are only weak correlations for HCDR3 amino acid preferences between healthy individuals and the cohorts of COVID-19 patients (Fig. 3B). Interestingly, the HCDR3 composition was most distinct between patients with moderate symptoms and the healthy cohort, with 73% correlation for length and 65% correlation for amino acid preferences (Fig. 3B). The HCDR3 composition was most similar between patients with moderate and severe symptoms, with

92% correlation for length and 89% correlation for amino acid preferences (Fig. 3B). Taken together, HCDR3 represents a molecular feature that is distinguishable among cohorts.

Expansion of BCR clonal lineages over time as a proxy for response to SARS-CoV-2. Next, we examined the dynamics of BCR repertoires in the COVID-19 patients. The binding level (measured by OD_{450} in ELISA assays) of both IgM and IgG antibodies against the receptor-binding domain (RBD) or N-terminal domain (NTD) of SARS-CoV-2 increased in most of the COVID-19 patients in our study over the course of their infection (Figs. 4A, S3). We expected that the increase of OD_{450} binding level is associated with activation of specific B-cells, resulting in an increase in mRNA production of the corresponding BCRs. Detecting expansion of specific clonal lineages is challenging due to subsampling of the repertoires. In fact, only a limited overlap of BCR lineages was found if we simply compared the data between different time points or between replicates of a repertoire sampled at the same time point (Fig. S4). To identify expanding clonal lineages, we examined lineages only in patients whose plasma showed an increase in binding level (OD_{450}) to the RBD of SARS-CoV-2 and compared the sequence abundance of those lineages that appear in two or more time points (Figs. 4A, S3 and Methods). Using a hypothesis test with a false discovery rate of 7.5%, as determined by analyzing replicate data (Methods, Fig. S5), we detected significant expansion of clonal lineages within all investigated patients. The results reflect a dynamic repertoire in all patients, ranging from 5% to 15% of lineages with significant expansion and large changes in sequence abundances over time (Figs. 4, S6). The expanding lineages had comparable HCDR3 length to the rest of the repertoire in COVID-19 patients (Fig. S6). Moreover, we observed expanding lineages to show V-gene preferences comparable to those of previously identified antibodies against SARS-CoV-2 (RBD). This includes the abundance of IGHV4-59, IGHV4-39, IGHV3-23, IGHV3-53, IGH3-66, IGHV2-5, and IGHV2-70 (Brouwer et al., 2020; Ju et al., 2020; Pinto et al., 2020; Rogers et al., 2020). However, it should be noted that these preferences in V-gene usage among expanding lineages are comparable to the overall biases in V-gene usage within patients, and expanded lineages roughly make up 25% of lineages with a given V gene (Fig. 4C). Therefore, Our results suggest that the overall response to SARS-CoV-2 is not driven by only a specific class of IGHV gene.

Sharing of BCRs among individuals. Despite the vast diversity of BCRs, we observe a significant number of identical progenitors of BCR clonal lineages among COVID-19 patients (Fig. 5) and among healthy individuals (Fig. S7). Previous work has also identified sharing of BCRs among COVID-19 patients, which was interpreted by the authors as evidence for large-scale convergence of immune responses (Galson et al., 2020; Nielsen et al., 2020; Schultheiß et al., 2020). Although BCR sharing can be due to convergent response to common antigens, it can also arise from convergent recombination leading to the same receptor sequence (Elhanati et al., 2018; Pogorelyy et al., 2018a) or simply from experimental biases. Therefore, it is imperative to formulate a null statistical model to identify the outliers among shared BCRs as candidates for common responses to antigens. Convergent

recombination defines a null expectation for the amount of sharing within a cohort based on only the underlying biases for receptor generation within a repertoire (Elhanati et al., 2018; Pogorelyy et al., 2018a) (Methods). Intuitively, sharing is more likely among commonly generated receptors (i.e., with a high $P_{\text{post}}(\sigma)$) and within cohorts with larger sampling (Methods). Importantly, rare receptors (i.e., with a low $P_{\text{post}}(\sigma)$) that are shared among individuals in a common disease group can signal commonality in function and a response to a common antigen, as previously observed for TCRs in response to a yellow fever vaccine (Pogorelyy et al., 2018b), CMV and diabetes (Pogorelyy et al., 2018a). We used the receptors' probabilities $P_{\text{post}}(\sigma)$ to assess the significance of sharing by identifying a probabilistic threshold to limit the shared outliers both among the COVID-19 patients (dashed line in Fig. 5) and the healthy individuals (dashed line in Fig. S7). Out of a total of 33,660 (unique) progenitors of clonal lineages (Fig. 5A), we found 5,112 progenitors to be shared among at least two individuals; 152 of the 5,112 were classified as rare, having a probability of occurrence below the indicated threshold (dashed line) in Fig. 5B. Moreover, we found that 439 lineages shared a common sequence ancestor in at least two individuals and have expanded in at least one of the individuals (Fig. 5C-D). 34 of these shared, expanding lineages stemmed from rare naïve progenitors (below the dashed line in Fig. 5B, D). The sharing of these rare expanding BCRs among COVID-19 patients indicates a potentially convergent response to SARS-CoV-2 (Table S3).

Interestingly, we found that 20% of receptors in these rare shared, expanding lineages contain multiple cysteines in their HCDR3s, in contrast to only 10% of the receptors in the whole repertoire. Such sequence patterns with cysteine pairs in the HCDR3 have been associated with stabilization of the HCDR3 loop by forming disulfide bonds with particular patterns and spacings of the cysteines (Lee et al., 2014; Prabakaran and Chowdhury, 2020). Disulfide bonds in the HCDR3 can decrease the conformational flexibility of the loop, thus decreasing the entropic cost of binding to improve the affinity of the receptor (Almagro et al., 2012). The significantly larger fraction of multi-cysteine HCDR3s among the candidate SARS-CoV-2 responsive receptors (p-value = 10^{-9} based on binomial sampling) indicates an underlying molecular mechanism for developing a potent response to SARS-CoV-2.

Presence of SARS-CoV-2 and SARS-specific neutralizing antibodies within repertoires. Previous studies have identified a number of antibodies responsive to the RBD of SARS-CoV-2 (Cao et al., 2020; Pinto et al., 2020; Robbiani et al., 2020). In repertoires of the COVID-19 patients, we found that several HCDR3s matched with SARS-CoV-2-specific monoclonal antibodies (mAbs) that were previously isolated in other studies (Table S3). Specifically, three distinct SARS-CoV-2-specific mAbs were found to be close in sequence to HCDR3s in our data, with only one amino acid difference, and one of these mAbs also had a matching V gene (Table 1). Among these, we found a class of antibodies containing multiple cysteines in their HCDR3s, with the potential to form disulfide bonds

to enhance their affinity to antigens and paratope diversity (Almagro et al., 2012; Lee et al., 2014; Prabakaran and Chowdhury, 2020).

In addition, we found that two patients had exact HCDR3 matches to a previously identified antibody (S304) that has cross-reactivity to SARS-CoV and SARS-CoV-2 (Pinto et al., 2020). We also observed in one patient a HCDR3 with only one amino acid difference to this antibody (Table 1). Importantly, the plasma in these patients showed a substantial binding level (OD_{450}) to SARS-CoV (Fig. S3), which indicates a possibility of cross-reactive antibody responses to SARS-CoV and SARS-CoV-2.

It should be noted that all of the previously verified antibodies that we identify in the patients' repertoires have a relatively high probability $P_{\text{post}}(\sigma)$ of generation and entry to the periphery (indicated by yellow diamonds in Fig. 5B). This is not surprising as it is very unlikely to observe rare BCRs (with small $P_{\text{post}}(\sigma)$) to be shared in multiple individuals. Overall, our results are encouraging for vaccine development since they indicate that even common antibodies can confer specific responses against SARS-CoV-2.

Discussion

A repertoire of immune receptor sequences presents a unique snapshot of the history of immune responses in an individual (Boyd et al., 2009; Georgiou et al., 2014; Kreer et al., 2020; Robins, 2013), and the changes in a repertoire during an infection can signal specific responses to pathogens (Horns et al., 2019; Nourmohammad et al., 2019). Identifying signatures of a functional response to a given pathogen from the pool of mostly unspecific BCRs collected from the blood is challenging—it is a problem of finding a needle in a haystack. Therefore, principled statistical inference approaches are necessary to extract functional signal from such data. Here, we characterized the B-cell repertoire response to SARS-CoV-2 in COVID-19 patients with different disease severity by combining evidence from the overall statistics of repertoires together with dynamics of clonal lineages during infection and the sharing of immune receptors among patients.

At the repertoire level, we showed that HCDR3 of BCRs in COVID-19 patients are significantly longer than HDCR3 in healthy individuals, and the amino acid composition of this receptor region varies among cohorts of patients with mild, moderate, and severe symptoms. Moreover, we observed large-scale sharing of B-cell receptors among COVID-19 patients, consistent with previous findings in COVID-19 patients (Galson et al., 2020; Nielsen et al., 2020; Schultheiß et al., 2020). Sharing of receptors among individuals can signal common immune responses to a pathogen. However, BCR sharing can also be due to convergent recombination leading to the same receptor sequence or other experimental biases that influence statistics of shared sequences. These statistical nuances can substantially sway conclusions drawn from the sharing analysis and, therefore, should be carefully

accounted for. Here, we established a null expectation of BCR sharing due to convergent recombination by inferring a model of receptor generation and migration to the periphery and used this null model to identify sequence outliers. Our analysis identified a subset of rare BCRs shared among COVID-19 patients, which appears to signal convergent responses to SARS-CoV-2.

In addition to the statistics of repertoires, we observed that the activity of many B-cell lineages (i.e., mRNA production) in COVID-19 patients significantly increases during infection, accompanied by an increase in the binding level (OD_{450}) of the patients' sera to the RBD and NTD of SARS-CoV-2. Dynamics of clonal lineages during an infection provide significant insights into the characteristics of responsive antibodies (Horns et al., 2019; Nourmohammad et al., 2019). By taking advantage of data collected at multiple time points in most patients, we identified expanded lineages shared among patients and found 34 clonal lineages that are candidates for a specific response specific to SARS-CoV-2 antigens. Further analysis would be required to determine the reactivity of these antibodies against the RBD or NTD of SARS-CoV-2.

Identifying antibodies with cross-reactive neutralization abilities against viruses in the SARS family is of significant interest. Interestingly, in nine patients, we see a substantial increase in the binding level (OD_{450}) of their sera to SARS-CoV epitopes during the course of COVID-19 infection. Moreover, in three patients, we identify a BCR identical to the heavy chain of antibody S304 (Pinto et al., 2020), which was previously isolated from a patient who recovered from a SARS-CoV infection. This antibody was shown to be moderately cross-reactive to both SARS-CoV and SARS-CoV-2, and our results further indicate a possibility for such cross-reactive antibodies to emerge naturally in response to SARS-CoV-2 (Brouwer et al., 2020; Lv et al., 2020; Rogers et al., 2020). Taken together, our findings provide substantial insight and strong implications for devising vaccines and therapies with a broad applicability against SARS-CoV-2.

Materials and Methods

Experimental procedures

Sample collection and PBMC isolation. Specimens of heparinized blood were collected from the RT-PCR-confirmed COVID-19 patients at the Infectious Disease Centre of the Princess Margaret Hospital, Hong Kong. The study was approved by the institutional review board of the Hong Kong West Cluster of the Hospital Authority of Hong Kong (approval number: UW20-169). All study procedures were performed after informed consent was obtained. Day 1 of clinical onset was defined as the first day of the appearance of clinical symptoms. The severity of the COVID-19 cases was classified based on the adaptation of the Sixth Revised Trial Version of the Novel Coronavirus Pneumonia Diagnosis and Treatment Guidance. The severity of the patients was categorized as follows: Mild - no sign of pneumonia on imaging, mild clinical symptoms; Moderate - fever, respiratory symptoms and radiological evidence of pneumonia; Severe - dyspnea, respiratory frequency >30/min, blood oxygen saturation 93%, partial pressure of arterial oxygen to fraction of inspired oxygen ratio <300, and/or lung infiltrates >50% within 24 to 48 hours; Critical - respiratory failure, septic shock, and/or multiple organ dysfunction or failure or death.

The blood samples were first centrifuged at 3000 xg for 10 minutes at room temperature for plasma collection. The remaining blood was diluted with equal volume of PBS buffer, transferred onto the Ficoll-Paque Plus medium (GE Healthcare), and centrifuged at 400 xg for 20 minutes. Peripheral Blood Mononuclear Cells (PBMC) samples were then collected and washed with cold RPMI-1640 medium for three times. The isolated PBMC samples were finally stored at cell freezing solution (10% DMSO + 90% FBS) and kept in -80°C until used.

RNA extraction and reverse transcription. Total RNA was extracted from 5×10^5 PBMC using the RNeasy Mini isolation kit (Qiagen) according to the manufacturer's protocol. Reverse transcription of the RNA samples was performed using the Proto- Script® II Reverse Transcriptase kit (New England Biolabs, NEB) with random hexamer primers according to the manufacturer's protocol. The thermal cycling conditions were designed as follows: 25°C for 5 minutes, 42°C for 60 minutes, and 80°C for 5 minutes. The resulting cDNA samples were stored in 80°C freezer before PCR was performed.

Amplification of B cell repertoire from the samples by PCR. The cDNA samples were used as a template to amplify the antibody IgG heavy chain gene with six FR1-specific forward primers and one constant region-specific reversed primer using the Phusion® High-Fidelity DNA Polymerase. The primer sequences were the same as previously described (Wu et al., 2015); primer sequences are listed in Table S2. The thermal cycling conditions were set as follows: 98°C for 30 seconds; 30 cycles of

98ºC for 10 seconds, 58ºC for 15 seconds, and 72ºC for 30 seconds; and 72ºC for 10 minutes. Then 10 ng of the PCR product was used as a template for the next round of gene amplification with sample-specific barcode primers. The thermal cycling conditions were set as follow: 98ºC for 3 min; 30 cycles of 98ºC for 10 seconds, 58ºC for 15 seconds, and 72ºC for 15 seconds; and a final extension at 72ºC for 10 min using Phusion® High-Fidelity DNA Polymerase. The PCR product was purified by QIAquick Gel Extraction Kit (Qiagen), and quantified by NanoDrop Spectrophotometers (Thermofisher).

Cell lines. Sf9 cells (*Spodoptera frugiperda* ovarian cells, female, ATCC catalogue no. CRL-1711) and High Five cells (*Trichoplusia ni* ovarian cells, female; Thermo Fisher Scientific, Waltham, United States (US), catalogue number: B85502) were maintained in HyClone (GE Health Care, Chicago, US) insect cell culture medium.

Protein expression and purification. The receptor-binding domain (RBD, residues 319–541) and N-terminal domain (NTD, residues 14 to 305) of the SARS-CoV-2 spike protein (GenBank: QHD43416.1) as well as the RBD (residues 306-527) and NTD (residues 14-292) of SARS-CoV spike protein (GenBank: ABF65836.1) were cloned into a customized pFastBac vector (Lv et al., 2020; Wec et al., 2020). The RBD and NTD constructs were fused with an N-terminal gp67 signal peptide and a C-terminal His_6 tag. Recombinant bacmid DNA was generated using the Bac-to-Bac system (Life Technologies, Thermo Fisher Scientific). Baculovirus was generated by transfecting purified bacmid DNA into Sf9 cells using FuGENE HD (Promega, Madison, US) and subsequently used to infect suspension cultures of High Five cells (Life Technologies) at a multiplicity of infection (moi) of 5 to 10. Infected High Five cells were incubated at 28 °C with shaking at 110 rpm for 72 h for protein expression. The supernatant was then concentrated using a Centramate cassette (10 kDa molecular weight cutoff for RBD, Pall Corporation, New York, USA). RBD and NTD proteins were purified by Ni-NTA Superflow (Qiagen, Hilden, Germany), followed by size exclusion chromatography and buffer exchange to phosphate-buffered saline (PBS).

ELISA. A 96-well enzyme-linked immunosorbent assay (ELISA) plate (Nunc MaxiSorp, Thermo Fisher Scientific) was first coated overnight with 100 ng per well of purified recombinant protein in PBS buffer. The plates were then blocked with 100 µl of Chonblock blocking/sample dilution ELISA buffer (Chondrex Inc, Redmon, US) and incubated at room temperature for 1 h. Each human plasma sample was diluted to 1:100 in Chonblock blocking/sample dilution ELISA buffer. Each sample was then added into the ELISA plates for a two-hour incubation at 37°C. After extensive washing with PBS containing 0.1% Tween 20, each well in the plate was further incubated with the anti-human IgG secondary antibody (1:5000, Thermo Fisher Scientific) for 1 hour at 37°C. The ELISA plates were then washed five times with PBS containing 0.1% Tween 20. Subsequently, 100 µL of HRP substrate (Ncm TMB One; New Cell and Molecular Biotech Co. Ltd, Suzhou, China) was added into each well.

After 15 min of incubation, the reaction was stopped by adding 50 µL of 2 M H_2SO_4 solution and analyzed on a Sunrise (Tecan, Männedorf, Switzerland) absorbance microplate reader at 450 nm wavelength.

BCR processing and computational procedures

BCR preprocessing. For initial processing of the raw reads, we used pRESTO (version 0.5.13) (Vander Heiden et al., 2014) to assemble paired-end reads, remove sequences with a mean quality score less than 30, mask primer subsequences, and collapse duplicate sequences into unique sequences. The small fraction of paired-end reads that overlapped were assumed to be anomalous and were discarded from the analysis. Additionally, after preprocessing with pRESTO, we discarded unique reads that contained ambiguous calls (N's) in their receptor sequence.

BCR error correction. We performed two rounds of error correction on sequences that passed the quality control check. In the first round, we clustered singletons and other low-frequency sequences into larger sequences if they were similar in sequence. The intent of this round was to correct for sequencing errors (e.g. from reverse transcription of mRNA to cDNA) that caused large abundance clones to be split into many similar sequences. We used two parameters: $\Delta_r = 1.0$, the marginal Hamming distance tolerance per decade in log-ratio abundance (each $\log_{10}$ unit allowing Δ_r additional sequence differences), and $\Delta_a = 1.0$, the marginal abundance tolerance of clusterable sequences per decade in log-ratio abundance (each $\log_{10}$ unit allowing abundance Δ_a higher as clusterable). For example, a sequence with abundance a_1 and a Hamming distance d away from a higher abundance sequence with abundance a_2 was absorbed into the latter only if $d \leq \Delta_r \log_{10} \frac{a_2}{a_1}$ and $a_1 \leq \Delta_a \log_{10} \frac{a_2}{a_1}$. We used the output of this first round as input for the second round of error correction, in which we more aggressively target correction of reverse transcriptase errors. In the second round, we used two different parameters to assess sequence similarity: $d_{\text{thresh}} = 2.0$, the Hamming distance between sequences, and $a_{\text{thresh}} = 1.0$, the ratio of sequence abundances. A sequence with abundance a_1 and a Hamming distance d away from a sequence of larger abundance a_2 was absorbed into the latter only if $d \leq d_{\text{thresh}}$ and the ratio of the sequence abundances was greater than a_{thresh}, i.e. $\frac{a_2}{a_1} \geq a_{\text{thresh}}$. This round of error correction allows much larger abundance sequences to potentially be clustered than is possible in the first round. For both of the above steps, we performed clustering greedily and approximately by operating on sequences sorted by descending abundance, assigning the counts of the lower abundance sequence to the higher abundance one iteratively.

After error correction, the sequences still contained a large number of singletons, i.e. sequences with no duplicates (Table S1). We discarded these singletons from all analyses that relied on statistics of unique sequences (i.e., the results presented in Figs. S1, S2C,E,F).

BCR annotation. For each individual, error-corrected sequences from all timepoints and replicates were pooled and annotated by abstar (version 0.3.5) (Briney and Burton, 2018). We processed the output of abstar, which included the estimated V gene/allele, J gene/allele, location of the HCDR3 region, and an inferred naïve sequence (germline before hypermutation). Sequences which had indels outside of the HCDR3 were discarded. We partitioned the sequences into two sets: productive BCRs, which were in-frame and had no stop codons, and unproductive BCRs, which were out-of-frame.

Unproductive BCRs. Due to a larger sequencing depth in healthy individuals, we were able to reconstruct relatively large unproductive BCR lineages. Unproductive sequences are BCRs that are generated but, due to a frameshift or insertion of stop codons, are never expressed. These BCRs reside with productive (functional) BCRs in a nucleus and undergo hypermutation during B-cell replication and, therefore, provide a suitable null expectation for generation of BCRs in immune repertoires.

Clonal lineage reconstruction. To identify BCR clonal lineages, we first grouped sequences by their assigned V gene, J gene, and HCDR3 length and then used single-linkage clustering with a threshold of 85% Hamming distance. A similar threshold has been suggested previously by (Gupta et al., 2017) to identify BCR lineages. Clusters of size smaller than three were discarded from our analysis, size being defined as the sum of the amount of unique sequences per time point within a lineage. For each cluster, there may have been multiple inferred naïve sequences, as this was an uncertain estimate. Therefore, the most common naïve sequence was chosen to be the naïve progenitor of the lineage. When the most common naïve sequence of a productive lineage contained a stop codon, the progenitor of the lineage was chosen iteratively by examining the next most common naïve sequence until it did not contain any stop codons. If all inferred naïve sequences in a productive lineage had a stop codon, that lineage was discarded from the analysis. Table S1 shows the statistics of constructed clonal lineages in each individual.

Inference of generation probability and selection for BCRs. We used IGoR (version 1.4) (Marcou et al., 2018) to obtain a model of receptor generation. This model characterized the probability of generation $P_{\text{gen}}(\sigma)$ of a receptor dependent on the features of the receptor, including the V, D, and J genes and the deletion and insertion profiles at the VD and DJ junctions. To characterize the parameters of this model, we trained IGoR on the progenitors of all unproductive lineages, regardless of size, pooled from all cohorts. For consistency with our receptor annotations based on abstar, we used abstar's genomic templates and the HCDR3 anchors of abstar's reference genome as inputs for IGoR's genomic templates and HCDR3 anchors.

We used SONIA (version 0.27) (Sethna et al., 2020) to infer a selection model for progenitors of productive clonal lineages. The SONIA model evaluated selection factors q to characterize the deviation in the probability $P_{\text{post}}(\sigma)$ to observe a functional sequence in the periphery from the null expectation, based on the generation probability $P_{\text{gen}}(\sigma)$: $P_{\text{post}}(\sigma) = \frac{1}{Z} P_{\text{gen}}(\sigma) e^{\Sigma_{f:\text{features}} q_f(\sigma)}$, where Z is the normalization factor and $q_f(\sigma)$ are selection factors dependent on the sequence features f. These sequence features include V-gene and J-gene usages and HCDR3 length and amino acid composition (Sethna et al., 2020).

In our analysis, we used the SONIA left-right model with independent V- and J-gene usages (Sethna et al., 2020). We used the output from IGoR as the receptor generation model for SONIA. We trained four cohort-specific SONIA models on progenitors of all productive lineages, regardless of size, pooled from all individuals within a cohort. 150 epochs and 500,000 generated sequences were used to train each SONIA model. Fig. 3 shows the distributions for the probabilities of observing productive receptors sampled from each cohort $P_{\text{post}}(\sigma)$ and the correlation of feature-specific selection factors q_f among cohorts.

Characterizing repertoire diversity. We quantified the diversity of each cohort by evaluating the entropy of receptor sequences in each cohort. Estimates of entropy can be influenced by how each distribution is sampled. To avoid such sampling biases, we used the inferred IGoR and SONIA models to generate 500,000 synthetics receptors based on each of the cohort-specific models. We evaluated cohort entropies H as the expected log-probabilities to observe a functional sequence in the respective cohort: $H = -\sum_{\sigma} P_{\text{post}}(\sigma) \log P_{\text{post}}(\sigma)$.

For comparison, we also evaluated the entropy estimated on the repertoire data in each cohort, which showed a similar pattern to the estimates from the generated cohorts (in the main text). Specifically, the entropy of BCR repertoires estimated from the data follows: 39.8 ± 0.3 bits in healthy individuals, 41.9 ± 0.7 bits for patients in the mild cohort, 42.6 ± 0.3 bits for patients in the moderate cohort, and 42.8 ± 0.5 for patients in the severe cohort. The error bars indicate the standard error due to differences among individuals within a cohort.

Clonal Lineage expansion. We studied clonal lineage expansion of BCR repertoires in individuals that showed an increase in the binding level (OD_{450}) of their plasma to SARS-CoV-2 (RBD) during infection (Figs. 4A, S3): patients 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, 14. Other individuals showed no increase in IgG binding to SARS-CoV-2 (RBD), either due to already high levels of binding at early time points or to natural variation and noise (Fig. S3). Our expansion test compared two time points. Therefore, for individuals with three time points, we combined data from different time points such that the

separated times coincided with larger changes in binding levels (OD_{450}). Specifically, we combined the last two time points for patients 2 and 7 and the first two time points for patient 9. In addition, we combined replicates at the same time point and filtered out small lineages with size < 3, where size was defined as the sum of the amount of unique sequences per time-point within a lineage.

To test for expansion, we compared lineage abundances (i.e., total number of reads in a lineage) between early and late time points. Many lineages appeared only in one time point due to the sparse sampling of clonal lineages and the cells that generate them (Fig. S4). Therefore, we tested for expansion only for lineages that had nonzero abundances at both time points.

Our expansion test relied on comparing the relative abundance of a given lineage with other lineages. However, due to primer-specific amplification biases, abundances were not comparable between reads amplified with different primers. Therefore, in our analysis we only compare a lineage with all other lineages that were amplified with the same primer.

We applied a hypergeometric test (Fisher's exact test) to characterize significance of abundance fold change for a focal lineage. A similar method was used to study clonal expansion in TCRs (DeWitt et al., 2015). For each focal clonal lineage i (in a given individual), we defined a 2×2 contingency matrix C,

$$\mathrm{C} = \begin{pmatrix} n_i^{\text{early}} & N_{/i}^{\text{early}} \\ n_i^{\text{late}} & N_{/i}^{\text{late}} \end{pmatrix}$$

where n_i^{early} and n_i^{late} are the abundances of the focal lineage at the early and late time, and $N_{/i}^{\text{early}}$ and $N_{/i}^{\text{late}}$ are the total abundances of all reads (with the same primer) minus those from lineage i at the early and late times. The ratio $(n_i^{\text{late}}/n_i^{\text{early}})/(N_{/i}^{\text{late}}/N_{/i}^{\text{early}})$ describes the fold change, or odds ratio, of lineage i relative to the rest of the reads in the same primer group. Based on the contingency matrix C, one-sided p-values for Fisher's exact test were calculated using the "fisher.test" function in R version 4.0. Fold change and p-values are shown in Fig. S6A.

To determine a significance threshold for the Fisher's exact test, we examined the replicate data from samples collected from the same time point in each individual because we did not expect any significant expansion among replicates. We calculated the expansion test on pairs of replicates (Fig. S5B,C), and we compared the empirical cumulative distributions of the time point and replicate expansion data (Fig. S5D,E) (Storey, 2002; Storey and Tibshirani, 2003). We chose a p-value

threshold of 10^{-300}, where there were 12.3 as many significant expansions as in the replicate data, and therefore the false discovery rate was approximately $1/(1+12.3) = 0.075$.

Significance of BCR sharing among individuals. The probability that receptor σ is shared among a given number of individuals due to convergent recombination can be evaluated based on the probability to observe the receptor in the periphery $P_{\text{post}}(\sigma)$, the size of the cohort M, and the size of the repertoire (sequence sample size) N. First, we evaluated the probability $\rho(\sigma; N)$ that receptor σ with probability $P_{\text{post}}(\sigma)$ appears at least once in a sample of size N,

$$\rho(\sigma; N) = 1 - \left(1 - P_{\text{post}}(\sigma)\right)^N \simeq 1 - e^{-NP_{\text{post}}}$$

The probability that receptor σ is shared among m individuals out of a cohort of M individuals, each with a (comparable) sample size N, follows the binomial distribution,

$$P_{\text{share}}(\sigma; m, M, N) = \binom{M}{m} [\rho(\sigma; N)]^m [1 - \rho(\sigma; N)]^{M-m}$$

We aimed to identify shared receptors that were outliers such that their probability of sharing is too small to be explained by convergent recombination or other biases in the data. To do so, we identified the receptors with the smallest sharing probabilities P_{share} and found a threshold of P_{post} (dashed lines in Fig. 5 and Fig. S7) at the 1% quantile of P_{share} in the data. Specifically, since P_{share} is a function of P_{post} and m (number of individuals sharing), for each m we solved for P_{post} such that $P_{\text{share}} = c$, and tuned the constant c such that only 1% of the data lay below P_{share}. This was a conservative choice to identify the rare shared outliers in the data.

Accessibility and resources.

BCR repertoire data can be accessed through:
https://www.ncbi.nlm.nih.gov/bioproject/PRJNA645245

All codes for data processing and statistical analysis can be found at:
https://github.com/StatPhysBio/covid-BCR

Acknowledgments

This work was supported by DFG grant (SFB1310) on Predictability in Evolution (A.N., Z.M., J.O., G.I.), the Max Planck Society through MPRG funding (A.N., Z.M., J.O., G.I.), Department of Physics at the University of Washington (A.N., Z.M.), NIH NIAID F31AI150163 (WSD), Calmette and Yersin scholarship from the Pasteur International Network Association (H.L.), Bill and Melinda Gates

Foundation OPP1170236 (I.A.W.), NIH K99 AI139445 (N.C.W.), US National Institutes of Health (contract no. HHSN272201400006C) (J.S.M.P), National Natural Science Foundation of China (NSFC)/Research Grants Council (RGC) Joint Research Scheme (N_HKU737/18) (C.K.P.M. and J.S.M.P) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project no. T11-712/19-N) (J.S.M.P). We acknowledge the support of the clinicians who facilitated this study, including Drs Wai Shing Leung, Jacky Man Chun Chan, Thomas Shiu Hong Chik, Chris Yau Chung Choi, John Yu Hong Chan, Daphne Pui-Lin Lau, and Ying Man Ho; the dedicated clinical team at Infectious Diseases Centre, Princess Margaret Hospital, Hospital Authority of Hong Kong; and the patients who kindly consented to participate in this investigation. We also thank the Center for PanorOmic Sciences (CPOS), LKS Faculty of Medicine, and University of Hong Kong for their support on next-generation sequencing and acknowledge the use of the computational infrastructure provided by the Hyak supercomputer system funded by the student technology fund (STF) at the University of Washington.

Competing Interests

The authors declare no competing interests.

References

Almagro, J.C., Raghunathan, G., Beil, E., Janecki, D.J., Chen, Q., Dinh, T., LaCombe, A., Connor, J., Ware, M., Kim, P.H., *et al.* (2012). Characterization of a high-affinity human antibody with a disulfide bridge in the third complementarity-determining region of the heavy chain. J Mol Recognit *25*, 125-135.

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 VDJ pyrosequencing. Sci Transl Med *1*, 12ra23.

Briney, B., and Burton, D.R. (2018). Massively scalable genetic analysis of antibody repertoires. bioRxiv 10.1101/447813, 447813.

Brouwer, P.J.M., Caniels, T.G., van der Straten, K., Snitselaar, J.L., Aldon, Y., Bangaru, S., Torres, J.L., Okba, N.M.A., Claireaux, M., Kerster, G., *et al.* (2020). Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. Science 10.1126/science.abc5902.

Burnet, F.M. (1959). The clonal selection theory of acquired immunity 10.5962/bhl.title.8281 (Vanderbilt University Press).

Burnet, F.M. (1960). Immunity as an aspect of general biology. In Mechanisms of antibody formation,

M. Holub, and J. Jaroskova, eds. (Prague: *Publishing House of Czech. Acad. Sci.*), pp. 15-21.

Cao, Y., Su, B., Guo, X., Sun, W., Deng, Y., Bao, L., Zhu, Q., Zhang, X., Zheng, Y., Geng, C.*, et al.* (2020). Potent neutralizing antibodies against SARS-CoV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. Cell 10.1016/j.cell.2020.05.025.

Chi, X., Yan, R., Zhang, J., Zhang, G., Zhang, Y., Hao, M., Zhang, Z., Fan, P., Dong, Y., Yang, Y.*, et al.* (2020). A neutralizing human antibody binds to the N-terminal domain of the spike protein of SARS-CoV-2. Science 10.1126/science.abc6952.

Cyster, J.G., and Allen, C.D.C. (2019). B cell responses: cell interaction dynamics and decisions. Cell *177*, 524-540.

DeWitt, W.S., Emerson, R.O., Lindau, P., Vignali, M., Snyder, T.M., Desmarais, C., Sanders, C., Utsugi, H., Warren, E.H., McElrath, J.*, et al.* (2015). Dynamics of the cytotoxic T cell response to a model of acute viral infection. J Virol *89*, 4517-4526.

Elhanati, Y., Murugan, A., Callan, C.G., Mora, T., and Walczak, A.M. (2014). Quantifying selection in immune receptor repertoires. Proc Natl Acad Sci U S A *111*, 9875-9880.

Elhanati, Y., Sethna, Z., Callan, C.G., Mora, T., and Walczak, A.M. (2018). Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. Immunol Rev *284*, 167-179.

Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J., Prati, D., Baselli, G., Asselta, R.*, et al.* (2020). Genomewide association study of severe Covid-19 with respiratory failure. N Engl J Med 10.1056/NEJMoa2020283.

Galson, J.D., Schaetzle, S., Bashford-Rogers, R.J.M., Raybould, M.I.J., Kovaltsuk, A., Kilpatrick, G.J., Minter, R., Finch, D.K., Dias, J., James, L.*, et al.* (2020). Deep sequencing of B cell receptor repertoires from COVID-19 patients reveals strong convergent immune signatures. bioRxiv 10.1101/2020.05.20.106294, 2020.2005.2020.106294.

Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H., and Quake, S.R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat Biotechnol *32*, 158-168.

Guan, W.J., Ni, Z.Y., Hu, Y., Liang, W.H., Ou, C.Q., He, J.X., Liu, L., Shan, H., Lei, C.L., Hui, D.S.C.*, et al.* (2020). Clinical characteristics of coronavirus disease 2019 in China. N Engl J Med *382*, 1708-1720.

Gupta, N.T., Adams, K.D., Briggs, A.W., Timberlake, S.C., Vigneault, F., and Kleinstein, S.H. (2017). Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. J Immunol *198*, 2489-2499.

Hachim, A., Kavian, N., Cohen, C.A., Chin, A.W., Chu, D.K., Mok, C.K.P., Tsang, O.T., Yeung, Y.C., Perera, R.A., Poon, L.L.*, et al.* (2020). Beyond the spike: identification of viral targets of the antibody response to SARS-CoV-2 in COVID-19 patients. medRxiv 10.1101/2020.04.30.20085670, 2020.2004.2030.20085670.

Horns, F., Vollmers, C., Dekker, C.L., and Quake, S.R. (2019). Signatures of selection in the human antibody repertoire: selective sweeps, competing subclones, and neutral drift. Proc Natl Acad Sci U S A *116*, 1261-1266.

Isacchini, G., Olivares, C., Nourmohammad, A., Walczak, A.M., and Mora, T. (2020a). SOS: online probability estimation and generation of T and B cell receptors. Bioinformatics 10.1093/bioinformatics/btaa574.

Isacchini, G., Sethna, Z., Elhanati, Y., Nourmohammad, A., Walczak, A.M., and Mora, T. (2020b). Generative models of T-cell receptor sequences. Phys Rev E *101*, 062414.

Janeway, C.A., Travers, P., Walport, M., and Shlomchik, M. (2005). Immunobiology: the immune system in health and disease, 6 edn (New York: Garland Science).

Ju, B., Zhang, Q., Ge, J., Wang, R., Sun, J., Ge, X., Yu, J., Shan, S., Zhou, B., Song, S.*, et al.* (2020). Human neutralizing antibodies elicited by SARS-CoV-2 infection. Nature 10.1038/s41586-020-2380-z.

Kreer, C., Gruell, H., Mora, T., Walczak, A.M., and Klein, F. (2020). Exploiting B cell receptor analyses to inform on HIV-1 vaccination strategies. Vaccines (Basel) *8*.

Lee, P.S., Ohshima, N., Stanfield, R.L., Yu, W., Iba, Y., Okuno, Y., Kurosawa, Y., and Wilson, I.A. (2014). Receptor mimicry by antibody F045-092 facilitates universal binding to the H3 subtype of influenza virus. Nat Commun *5*, 3614.

Liu, L., Wang, P., Nair, M.S., Yu, J., Huang, Y., Rapp, M.A., Wang, Q., Luo, Y., Sahi, V., Figueroa, A.*, et al.* (2020). Potent neutralizing monoclonal antibodies directed to multiple epitopes on the SARS-CoV-2 spike. bioRxiv 10.1101/2020.06.17.153486.

Lv, H., Wu, N.C., Tsang, O.T., Yuan, M., Perera, R., Leung, W.S., So, R.T.Y., Chan, J.M.C., Yip, G.K., Chik, T.S.H.*, et al.* (2020). Cross-reactive antibody response between SARS-CoV-2 and SARS-CoV infections. Cell Rep *31*, 107725.

Marcou, Q., Mora, T., and Walczak, A.M. (2018). High-throughput immune repertoire analysis with IGoR. Nat Commun *9*, 561.

McKechnie, J.L., and Blish, C.A. (2020). The innate immune system: fighting on the front lines or fanning the flames of COVID-19? Cell Host Microbe *27*, 863-869.

Nielsen, S.C.A., and Boyd, S.D. (2018). Human adaptive immune receptor repertoire analysis—past, present, and future. Immunol Rev *284*, 9-23.

Nielsen, S.C.A., Yang, F., Hoh, R.A., Jackson, K.J.L., Roeltgen, K., Lee, J.-Y., Rustagi, A., Rogers, A.J., Powell, A.E., Kim, P.S.*, et al.* (2020). B cell clonal expansion and convergent antibody responses to SARS-CoV-2. Res Sq 10.21203/rs.3.rs-27220/v1.

Nourmohammad, A., Otwinowski, J., Łuksza, M., Mora, T., and Walczak, A.M. (2019). Fierce selection and interference in B-cell repertoire response to chronic HIV-1. Mol Biol Evol *36*, 2184-2194.

Perera, R.A., Mok, C.K., Tsang, O.T., Lv, H., Ko, R.L., Wu, N.C., Yuan, M., Leung, W.S., Chan, J.M., Chik, T.S.*, et al.* (2020). Serological assays for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), March 2020. Euro Surveill *25*.

Pinto, D., Park, Y.-J., Beltramello, M., Walls, A.C., Tortorici, M.A., Bianchi, S., Jaconi, S., Culap, K., Zatta, F., De Marco, A.*, et al.* (2020). Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. Nature 10.1038/s41586-020-2349-y.

Pogorelyy, M.V., Minervina, A.A., Chudakov, D.M., Mamedov, I.Z., Lebedev, Y.B., Mora, T., and Walczak, A.M. (2018a). Method for identification of condition-associated public antigen receptor sequences. eLife *7*, e33050.

Pogorelyy, M.V., Minervina, A.A., Touzel, M.P., Sycheva, A.L., Komech, E.A., Kovalenko, E.I., Karganova, G.G., Egorov, E.S., Komkov, A.Y., Chudakov, D.M.*, et al.* (2018b). Precise tracking of vaccine-responding T cell clones reveals convergent and personalized response in identical twins. Proc Natl Acad Sci U S A *115*, 12704-12709.

Prabakaran, P., and Chowdhury, P.S. (2020). Landscape of non-canonical cysteines in human VH repertoire revealed by immunogenetic analysis. Cell Rep *31*, 107831.

Robbiani, D.F., Gaebler, C., Muecksch, F., Lorenzi, J.C.C., Wang, Z., Cho, A., Agudelo, M., Barnes, C.O., Gazumyan, A., Finkin, S.*, et al.* (2020). Convergent antibody responses to SARS-CoV-2 in convalescent individuals. Nature 10.1038/s41586-020-2456-9.

Robins, H. (2013). Immunosequencing: applications of immune repertoire deep sequencing. Curr Opin Immunol *25*, 646-652.

Rogers, T.F., Zhao, F., Huang, D., Beutler, N., Burns, A., He, W.T., Limbo, O., Smith, C., Song, G., Woehl, J.*, et al.* (2020). Isolation of potent SARS-CoV-2 neutralizing antibodies and protection from disease in a small animal model. Science 10.1126/science.abc7520.

Schultheiß, C., Paschold, L., Simnica, D., Mohme, M., Willscher, E., von Wenserski, L., Scholz, R., Wieters, I., Dahlke, C., Tolosa, E.*, et al.* (2020). Next generation sequencing of T and B cell receptor repertoires from COVID-19 patients showed signatures associated with severity of disease. Immunity 10.1016/j.immuni.2020.06.024.

Sethna, Z., Isacchini, G., Dupic, T., Mora, T., Walczak, A.M., and Elhanati, Y. (2020). Population variability in the generation and thymic selection of T-cell repertoires. bioRxiv 10.1101/2020.01.08.899682, 2020.2001.2008.899682.

Seydoux, E., Homad, L.J., MacCamy, A.J., Parks, K.R., Hurlburt, N.K., Jennewein, M.F., Akins, N.R., Stuart, A.B., Wan, Y.H., Feng, J.*, et al.* (2020a). Characterization of neutralizing antibodies from a SARS-CoV-2 infected individual. bioRxiv 10.1101/2020.05.12.091298.

Seydoux, E., Homad, L.J., MacCamy, A.J., Parks, K.R., Hurlburt, N.K., Jennewein, M.F., Akins, N.R., Stuart, A.B., Wan, Y.H., Feng, J.*, et al.* (2020b). Analysis of a SARS-CoV-2-infected individual reveals development of potent neutralizing antibodies with limited somatic mutation. Immunity

10.1016/j.immuni.2020.06.001.

Storey, J.D. (2002). A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol *64*, 479-498.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci U S A *100*, 9440-9445.

Vabret, N., Britton, G.J., Gruber, C., Hegde, S., Kim, J., Kuksin, M., Levantovsky, R., Malle, L., Moreira, A., Park, M.D.*, et al.* (2020). Immunology of COVID-19: current state of the science. Immunity *52*, 910-941.

Vander Heiden, J.A., Yaari, G., Uduman, M., Stern, J.N., O'Connor, K.C., Hafler, D.A., Vigneault, F., and Kleinstein, S.H. (2014). pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. Bioinformatics *30*, 1930-1932.

Wec, A.Z., Wrapp, D., Herbert, A.S., Maurer, D., Haslwanter, D., Sakharkar, M., Jangra, R.K., Dieterle, M.E., Lilov, A., Huang, D.*, et al.* (2020). Broad sarbecovirus neutralizing antibodies define a key site of vulnerability on the SARS-CoV-2 spike protein. bioRxiv 10.1101/2020.05.15.096511.

WHO (2020). Coronavirus disease (COVID-19) pandemic.

Wu, J.T., Leung, K., Bushman, M., Kishore, N., Niehus, R., de Salazar, P.M., Cowling, B.J., Lipsitch, M., and Leung, G.M. (2020a). Estimating clinical severity of COVID-19 from the transmission dynamics in Wuhan, China. Nat Med *26*, 506-510.

Wu, Y., Wang, F., Shen, C., Peng, W., Li, D., Zhao, C., Li, Z., Li, S., Bi, Y., Yang, Y.*, et al.* (2020b). A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. Science *368*, 1274-1278.

Wu, Y.C., Kipling, D., and Dunn-Walters, D. (2015). Assessment of B cell repertoire in humans. Methods Mol Biol *1343*, 199-218.

Zost, S.J., Gilchuk, P., Chen, R.E., Case, J.B., Reidy, J.X., Trivette, A., Nargi, R.S., Sutton, R.E., Suryadevara, N., Chen, E.C.*, et al.* (2020). Rapid isolation and profiling of a diverse panel of human monoclonal antibodies targeting the SARS-CoV-2 spike protein. bioRxiv 10.1101/2020.05.12.091462.

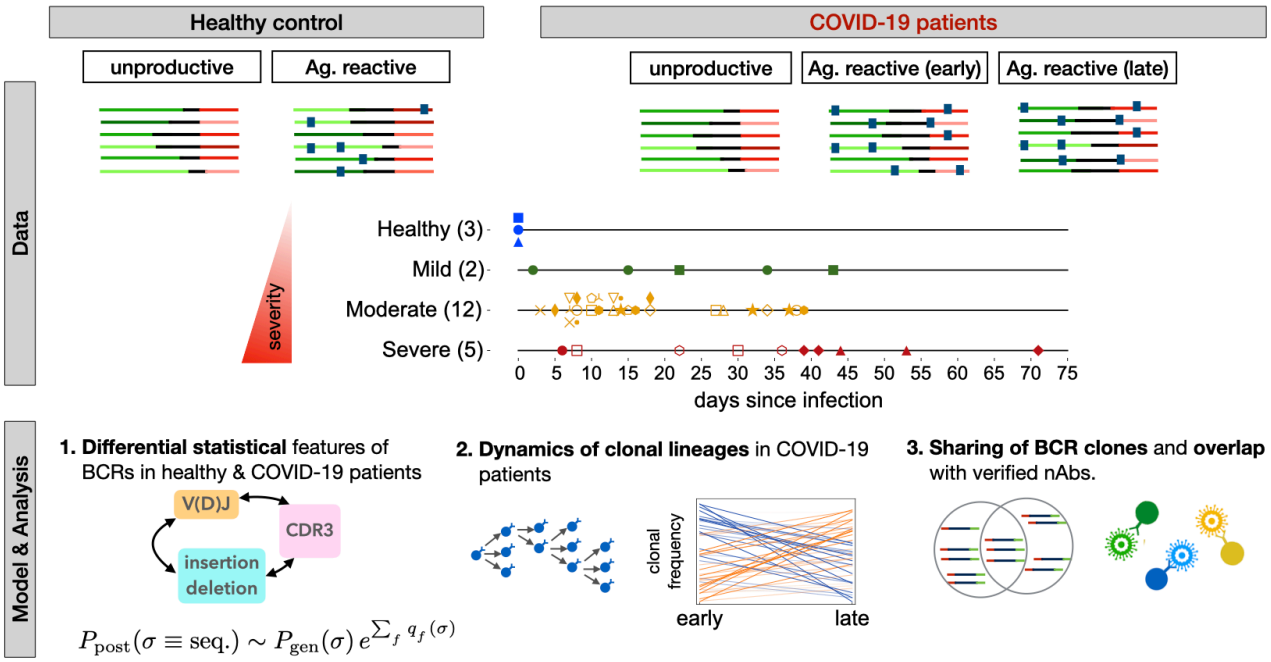

Figure 1. Roadmap for analysis of BCR repertoires. Top: We collected bulk blood IgG BCR samples from three healthy individuals and 19 COVID-19 patients where two patients had mild symptoms, 12 had moderate symptoms, and five had severe symptoms (different markers and colors) (See Methods). Samples were collected at different time points during infection (shown in center). We distinguished between productive receptors and unproductive receptors that had frameshifts due to V(D)J recombination. In each patient, we constructed clonal lineages for productive and unproductive BCRs and inferred the naïve progenitor of the lineage (Methods). **Bottom:** 1. We inferred a model (Marcou et al., 2018) to characterize a null probability for generation of receptors $P_{\text{gen}}(\sigma)$, trained on the set of unproductive inferred naïve BCRs. We inferred a selection model (Isacchini et al., 2020b) to characterize the deviation from the null among inferred naïve productive BCRs using the probability of entry to the periphery $P_{\text{post}}(\sigma)$ and characterize selection factors $q_f(\sigma)$ dependent on sequence features to differentiate between cohort statistics. 2. Based on temporal information of sampled BCRs, we identified clonal lineages that showed significant expansion during infection. 3. We identified progenitors of clonal lineages shared among individuals and assess the significance of these sharing statistics based on the probabilities to find each receptor in the periphery. We also identified previously characterized monoclonal antibodies (mAbs) specific to SARS-CoV-2 and SARS-CoV.

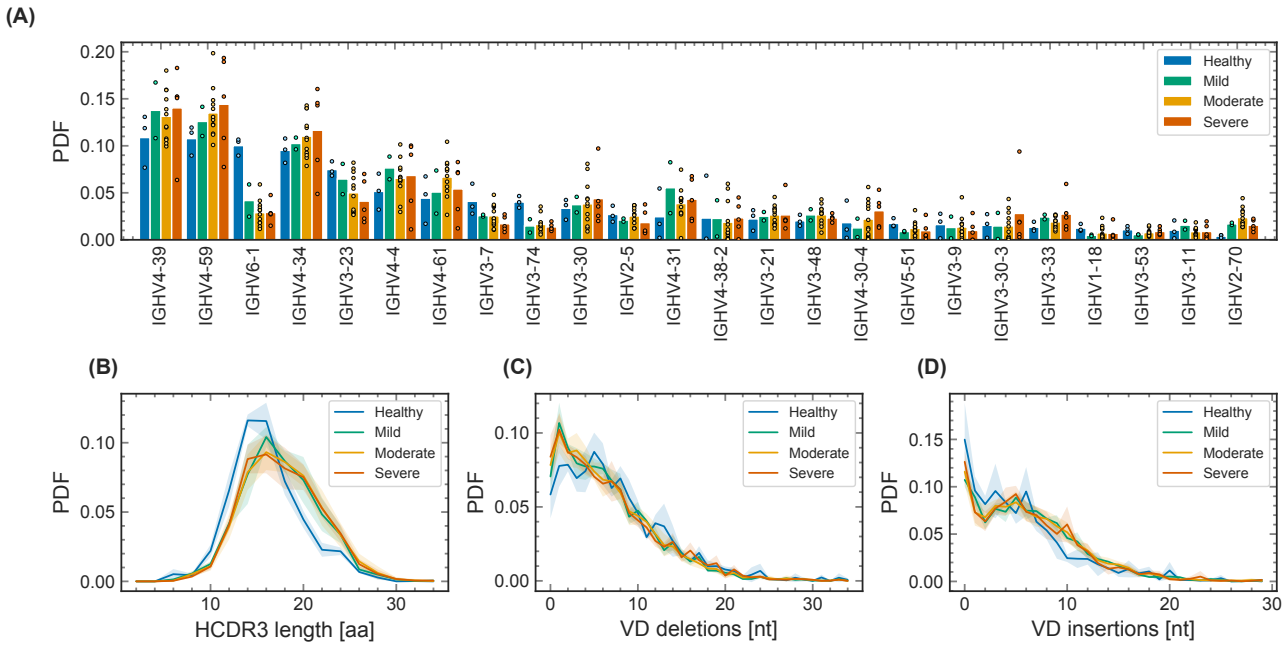

Figure 2. Sequence features of immune receptors across cohorts. (A) The normalized probability density function (PDF) for IGHV-gene usage is shown for inferred naïve progenitors of clonal lineages in cohorts of healthy individuals and COVID-19 cohorts of patients exhibiting mild, moderate, and severe symptoms. The bars indicate the usage frequency averaged over individuals in each cohort, and dots indicate the variation in V-gene frequencies across individuals within each cohort. **(B)** Distribution of length of HCDR3 amino acid sequence, **(C)** length of nucleotide sequence insertions at the VD junction, and **(D)** length of nucleotide sequence deletions at the VD junction are shown for each cohort. Full lines show distributions averaged over individuals in each cohort, and shadings indicate regions containing one standard deviation of variation among individuals within a cohort. The inferred naïve progenitors of clonal lineages in healthy individuals show significantly shorter HCDR3s, shorter VD deletions, and longer VD insertions compared to COVID-19 patients with moderate and sever symptoms. ANOVA statistics for mean HCDR3 length between cohorts: Healthy-Mild: $F_{1,3} = 12.0$, p-value = 0.040; Healthy-Moderate: $F_{1,13} = 15.7$, p-value = 0.002; Healthy-Severe: $F_{1,6} = 37.3$, p-value = 0.001. ANOVA statistics for mean VD deletion length between cohorts: Healthy-Mild: $F_{1,3} = 4.2$, p-value = 0.133; Healthy-Moderate: $F_{1,13} = 12.6$, p-value = 0.004; Healthy-Severe: $F_{1,6} = 19.1$, p-value = 0.005. ANOVA statistics for mean VD insertion length between cohorts: Healthy-Mild: $F_{1,3} = 3.7$, p-value = 0.152; Healthy-Moderate: $F_{1,13} = 10.0$, p-value = 0.008; Healthy-Severe: $F_{1,6} = 8.2$, p-value = 0.03.

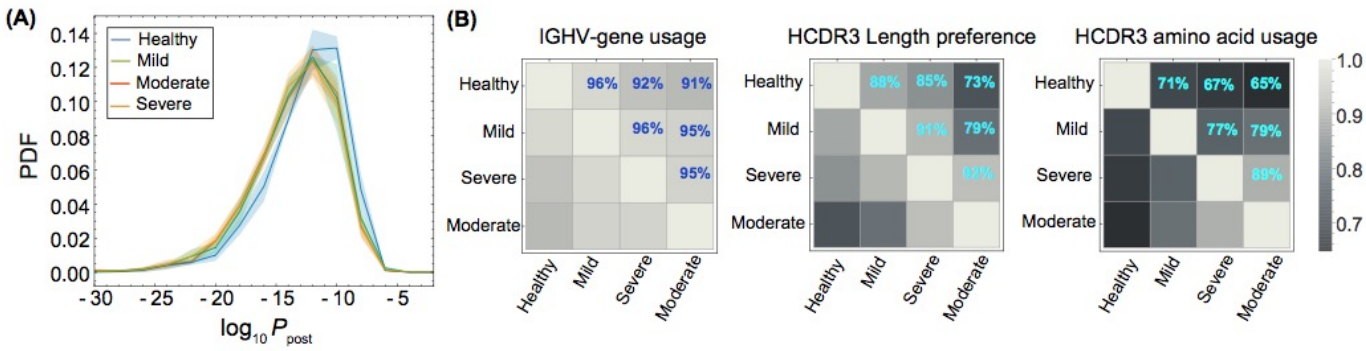

Figure 3. Differential statistics of immune repertoires across cohorts. (A) The distribution of the log-probability to observe a sequence σ in the periphery $\log_{10} P_{\text{post}}(\sigma)$ is shown for inferred naïve progenitors of clonal lineages in cohorts of healthy individuals and the mild, moderate, and severe cohorts of COVID-19 patients. Full lines show distributions averaged over individuals in each cohort, and shadings indicate regions containing one standard deviation of variation among individuals within a cohort. **(B)** Correlation between the inferred feature-specific selection factors q_f for V-gene usage (left), HCDR3 length preference (middle), and HCDR3 amino acid usage (right) is shown among cohorts. Correlation values associated with each color is indicated in the matrices.

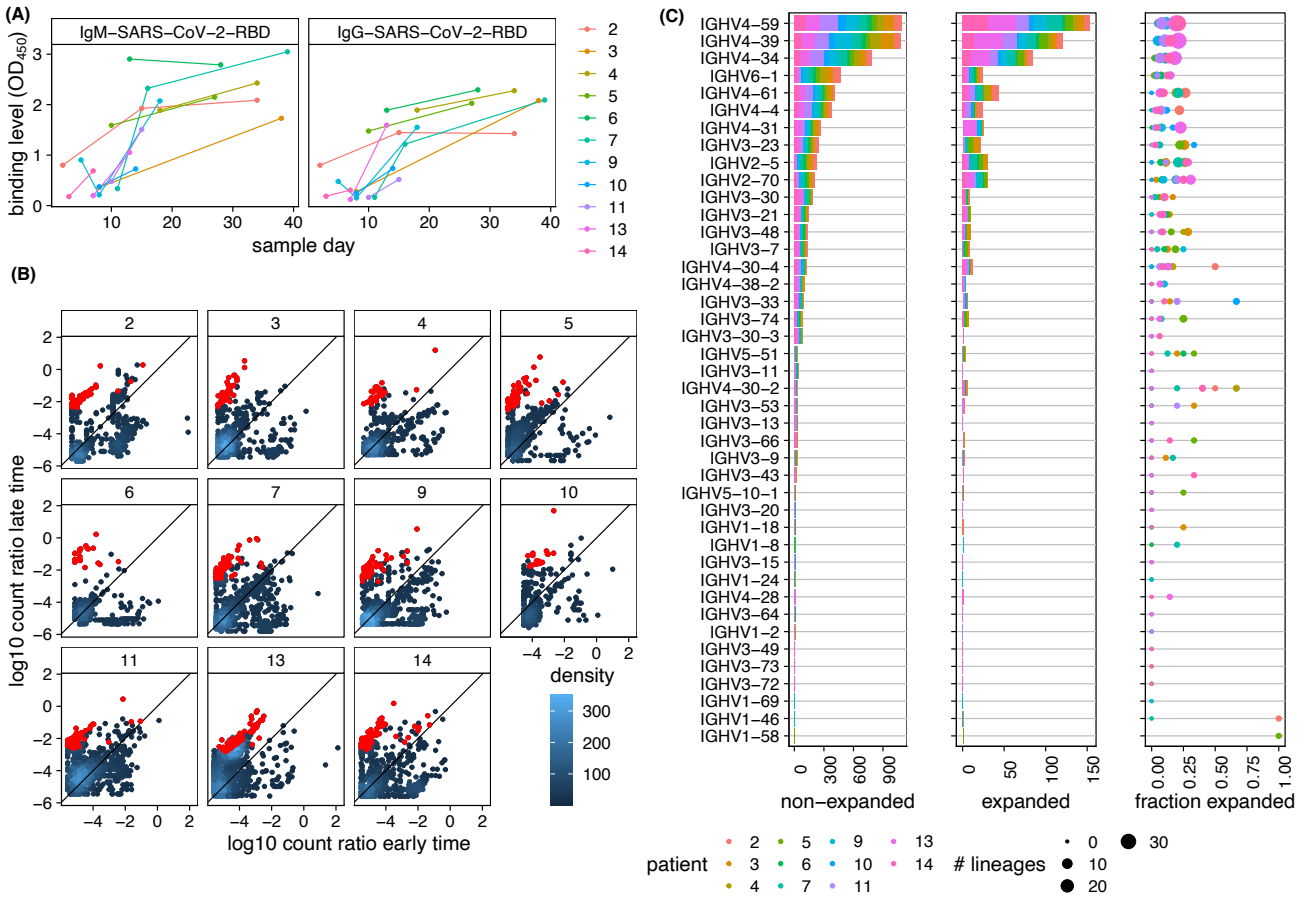

Figure 4: Dynamics of BCR repertoires during infection. (A) The binding level (measured by OD_{450} in ELISA assay) of the IgM (left) and IgG (right) repertoires to SARS-CoV-2 (RBD) epitopes increases over time in most individuals. **(B)** The log-ratio of BCR (mRNA) abundance at late time versus early time is shown for all clonal lineages that are present at least in two time points (see Methods). Each panel shows dynamics of lineages for a given individual, as indicated in the label. The analysis is shown in individuals for whom the binding level (OD_{450}) of the IgG repertoire increases over time (shown in (**A**)). The count density indicates the number of lineages at each point. Lineages that show a significant expansion over time are indicated in red (see Methods for estimation of associated p-values). **(C)** IGHV-gene usage of lineages is shown for non-expanded (left) and expanded (middle) lineages in all individuals (colors). The right panel shows, for each patient (colors), the fraction of expanded lineages with a given IGHV gene as the number of expanded lineages divided by the total number of lineages with that given IGHV gene. The size of the circles indicates the total number of lineages in each category.

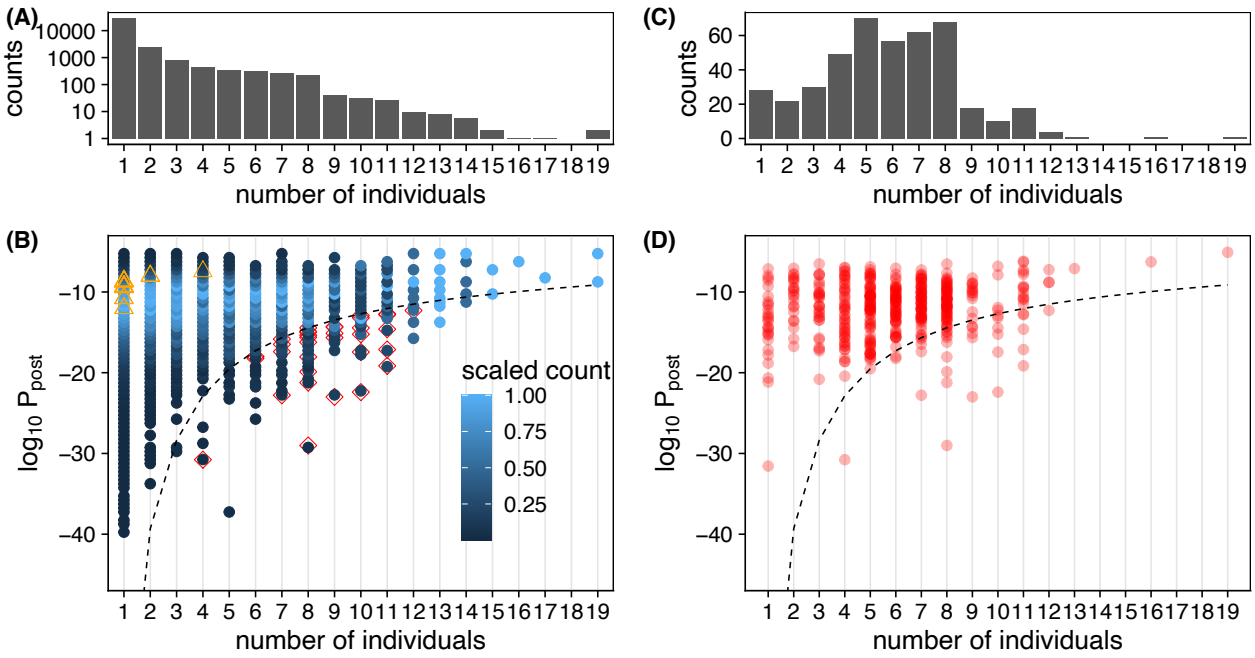

Figure 5. Sharing of BCRs among patients. (A) The histogram shows the number of clonal lineages that share a common progenitor in a given number of individuals, indicated on the horizontal axis. **(B)** The density plot shows the distribution of P_{post} for progenitors of clonal lineages shared in a given number of individuals, indicated on the horizontal axis; Histogram bin size is 0.5. The scaling of sequence counts sets the maximum of density in each column to one. Sharing of rare lineages with P_{post} below the dashed line is statistically significant (see Methods). Red diamonds indicate clonal lineages below the dashed line with significant expansion in at least one of the individuals, identified in Fig. 4. Yellow triangles indicate HCDR3 matches with previously identified SARS-CoV-2 specific mAbs. **(C)** The histogram shows the number of clonal lineages that share a common progenitor in a given number individuals, indicated on the horizontal axis, and which have significantly expanded (Fig. 4) during infection in at least one of the individuals. **(D)** The scatter plot with transparent overlapping markers shows P_{post} for progenitors of clonal lineages shared in a given number individuals that have expanded in at least one individual, shown in (C). The dashed line indicates P_{post}-dependent significance of sharing, similar to (B).

verified antibody				observed antibody		
ID	**IGHV**	**HCDR3**	**target**	**IGHV**	**HCDR3**	**Patient ID**
BD-494 (Cao et al., 2020) BD-498 (Cao et al., 2020) BD-507 (Cao et al., 2020)	IGHV3-53 IGHV3-66 IGHV3-53	ARDLVVYGMDV	SARS-2: RBD	IGHV4-59 IGHV4-39 IGHV4-61 IGHV4-34	ARDLVVGGMDV	7
COV21.2 (Robbiani et al., 2020) COV57.2 (Robbiani et al., 2020) COV107.2 (Robbiani et al., 2020)	IGHV1-58	AAPHCSGGSCYDAFDI AAPYCSGGSCNDAFDI AAPYCSGGSCSDAFDI	SARS-2: RBD	IGHV1-58	AAPYCSGGSCYDAFDI AAPYCSGGSCYDAFDI AAPYCSGGSCFDAFDI	6 6 9
B38 (Wu et al., 2020b)	IGHV3-53	AREAYGMDV	SARS-2: RBD	IGHV3-66	ARGAYGMDV	13 14 15
S304 (Pinto et al., 2020)	IGHV3-13	AKGDSSGYYYYFDY	SARS	IGHV3-13	AKGDSSGYYYYFDY AKGDSSGYYYYFDY ARGYSSGYYYYFDY	5 7 13

Table 1. Verified antibodies detected in BCR repertoires of COVID-19 patients. Verified antibodies responsive to SARS-CoV-2 or SARS-CoV, whose HCDR3 sequences match with a receptor (with up to Hamming distance of one amino acid) in the BCR repertoires of patients in this study are shown. Mutations with respect to the original HCDR3 are in red. Single amino acid mutations among HCDR3 of verified antibodies are in cyan. The complete list of verified antibodies is given in Table S3.

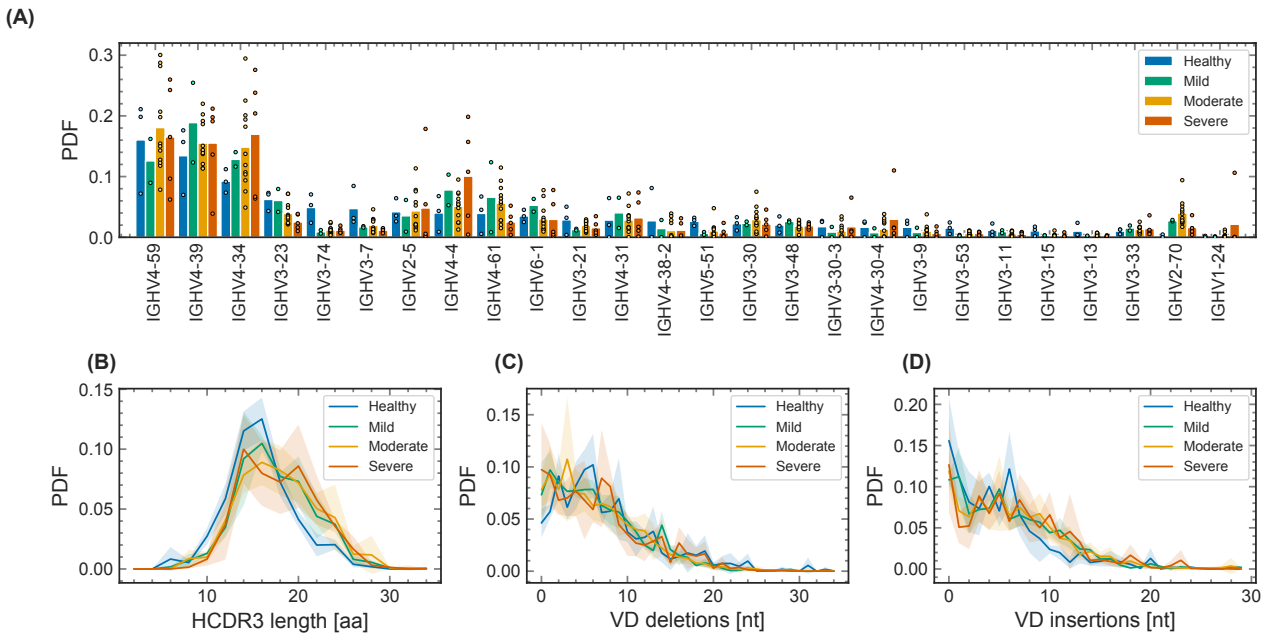

Figure S1. Features of unique immune receptors across cohorts. (A) IGHV-gene usage is shown for unique receptors in cohorts of healthy individuals and COVID-19 cohorts of patients exhibiting mild, moderate, and severe symptoms. The bars indicate the usage frequency averaged over individuals in each cohort, and dots indicate the variation in V-gene frequencies across individuals within each cohort. **(B)** Distribution of the length of HCDR3 amino acid sequence, **(C)** the length of nucleotide sequence insertion at the VD junction, and **(D)** the length of nucleotide sequence deletion at the VD junction are shown for each cohort. Full lines show distributions averaged over individuals in each cohort, and shading indicates regions containing one standard deviation of variation among individuals within a cohort. Unique BCRs in healthy individuals show significantly shorter HCDR3s compared to moderate and sever cohorts, and significantly shorter VD deletions, and longer VD insertions compared to only the moderate cohort. ANOVA statistics for mean HCDR3 length between cohorts: Healthy-Mild: $F_{1,3} = 8.7$, p-value = 0.060; Healthy-Moderate: $F_{1,13} = 17.2$, p-value = 0.001; Healthy-Severe: $F_{1,6} = 10.0$, p-value = 0.020. ANOVA statistics for mean VD deletion length between cohorts: Healthy-Mild: $F_{1,3} = 1.7$, p-value = 0.288; Healthy-Moderate: $F_{1,13} = 9.5$, p-value = 0.009; Healthy-Severe: $F_{1,6} = 4.5$, p-value = 0.08. ANOVA statistics for mean VD insertion length between cohorts: Healthy-Mild: $F_{1,3} = 2.1$, p-value = 0.241; Healthy-Moderate: $F_{1,13} = 10.1$, p-value = 0.007; Healthy-Severe: $F_{1,6} = 4.3$, p-value = 0.08.

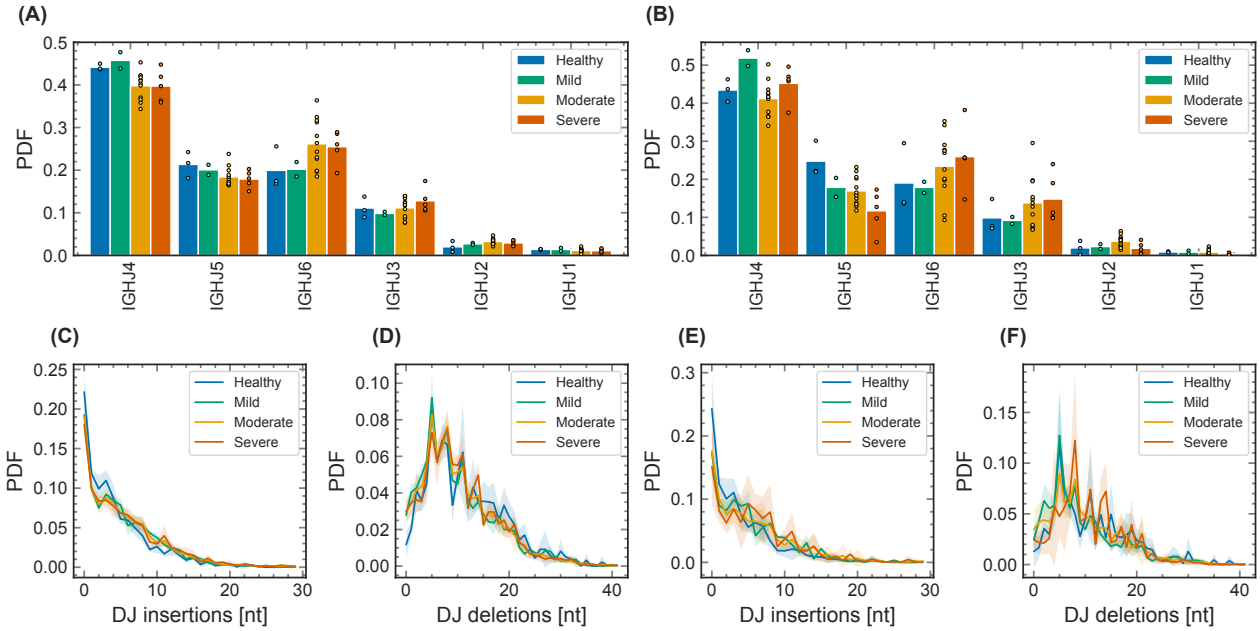

Figure S2. IGHJ-gene usage and insertion/deletion at DJ junctions. IGHJ-gene usage is shown for lineage progenitors **(A)** and for unique receptors **(B)** in cohorts of healthy individuals and COVID-19 cohorts of patients exhibiting mild, moderate, and severe symptoms. The bars indicate the usage frequency averaged over individuals in each cohort, and dots indicate the variation in J-gene frequencies across individuals within each cohort. Distributions of the length of nucleotide sequence insertion **(C)** and deletion **(D)** at the DJ junction are shown for lineage progenitors. **(E)** and **(F)** show similar statistics to (**C**) and (**D**) for unique receptors in each cohort. (**C-F**) Full lines show distributions averaged over individuals in each cohort, and shadings indicates regions containing one standard deviation of variation among individuals within a cohort. The inferred naïve progenitors of clonal lineages in healthy individuals show significantly longer DJ insertions, and shorter DJ deletions compared to COVID-19 patients with moderate symptoms. ANOVA statistics for mean DJ insertion length between Healthy-Moderate: $F_{1,13} = 12.4$, p-value = 0.004, and for mean DJ deletion length between Healthy-Moderate: $F_{1,13} = 10.9$, p-value = 0.006. The differences in the statistics of DJ indels between healthy individuals and other infected cohorts are insignificant (ANOVA p-value > 0.01). The differences in statistics of DJ indels for unique BCRs are insignificant between cohorts.

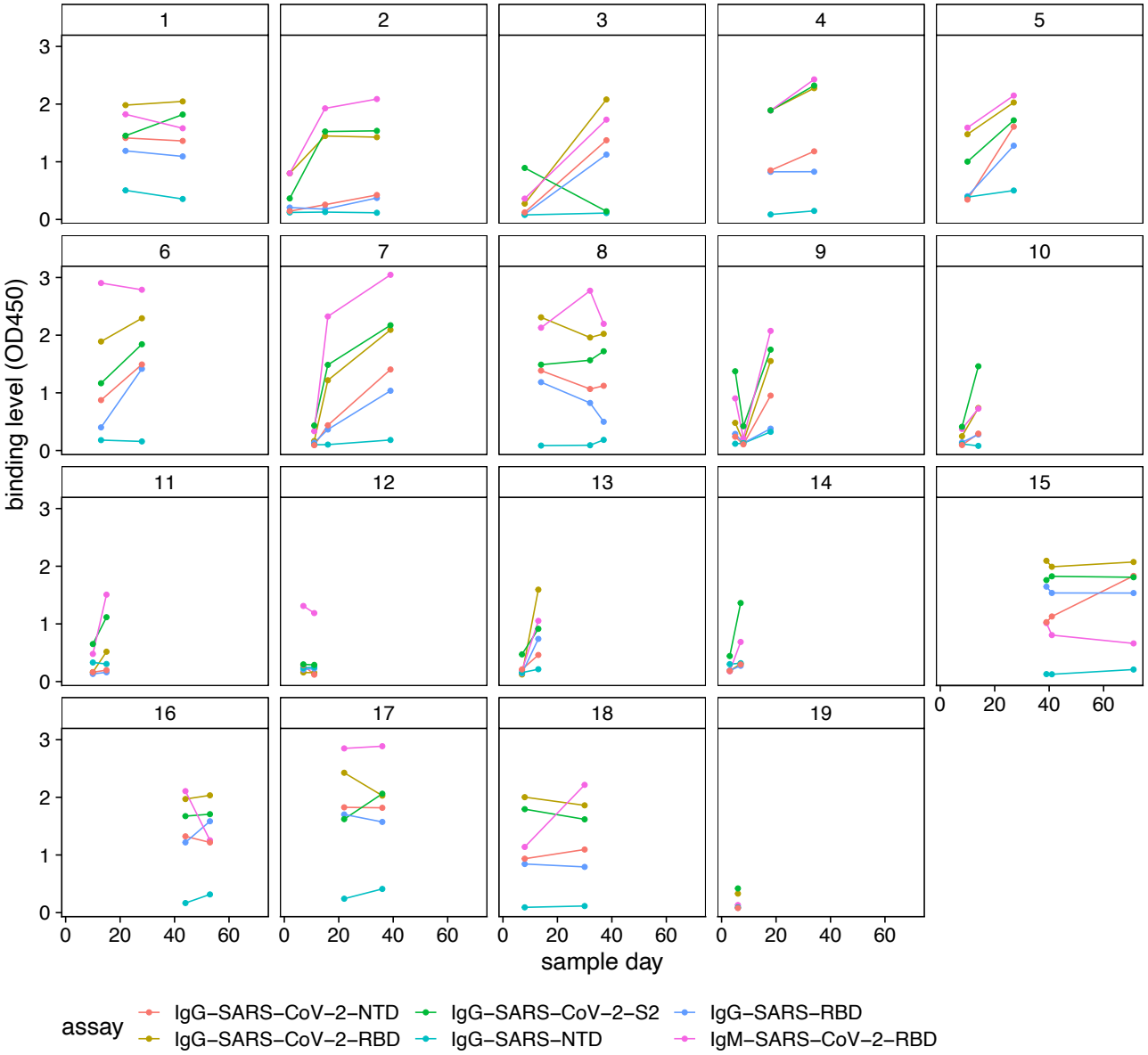

Figure S3. ELISA binding assays for IgG and IgM repertoires against SARS-CoV-2 and SARS-CoV. Plasma binding levels (measured by OD_{450} in ELISA) against RBD, NTD, and S2 epitopes of SARS-CoV-2 and against RBD and NTD epitopes of SARS-CoV are shown. As seen in binding assays, many individuals developed a cross-reactive response to SARS-CoV-2 and SARS-CoV. Some individuals showed no increase in IgG binding to SARS-CoV-2 RBD due to already high levels at sampling time or natural variation. For the expansion analysis (Fig. 4), we analyzed only individuals whose IgG repertoires showed an increase in binding to SARS-CoV-2 (RBD): 2, 3, 4, 5, 6, 7, 9, 10, 11, 13, and 14.

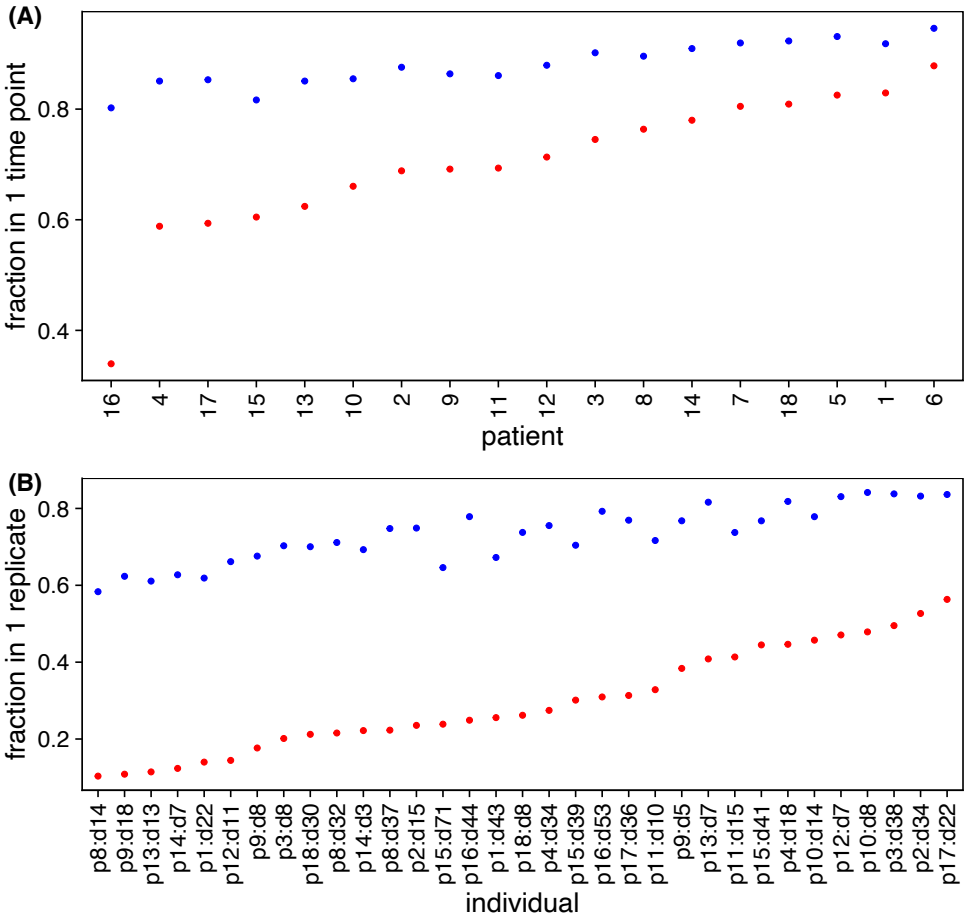

Figure S4. Overlap of BCR lineages between samples from each individual. BCR repertoires are highly under-sampled, and relatively few BCR lineages appear in multiple time points and replicates. **(A)** Fraction of lineages present in only one time point before (blue) and after (red) filtering out small lineages (i.e., those with less than three unique sequences per time point) are shown. **(B)** Fraction of lineages present in only one replicate before (blue) and after (red) filtering out small lineages (i.e., those with less than three unique sequences per time point) are shown.

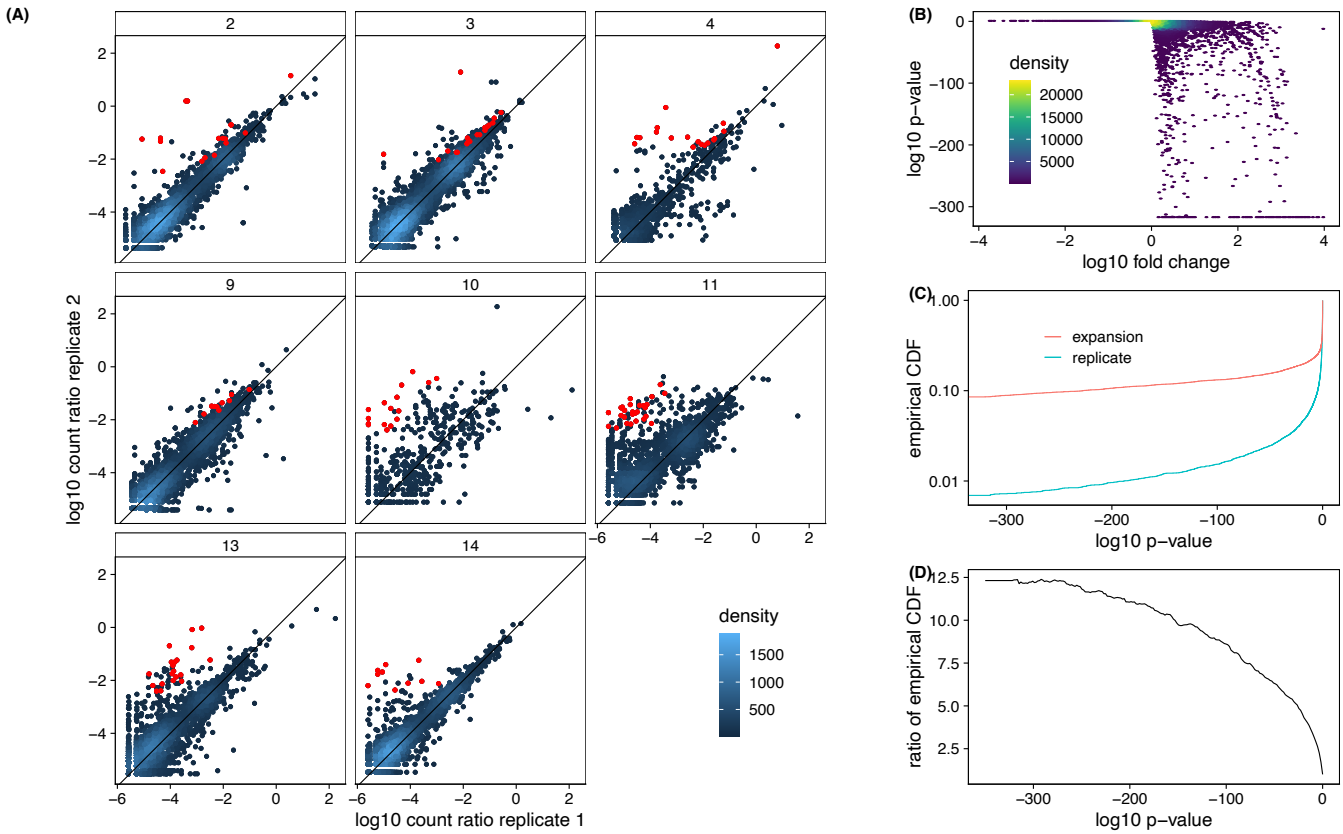

Figure S5. Expansion test with replicate data. (A) The log-ratio of abundance of receptors for all clonal lineages present in two replicates are shown. Each panel shows the test result for a given patient, as indicated in the label. The count density indicates the number of lineages at each point. Lineages that show a significant expansion over time are indicated in red. Since this is replicate data and represents a null model, red points indicate false positives. **(B)** $\log_{10}$ p-values of the expansion test versus $\log_{10}$ fold change (or odds ratio) for replicate data are shown. Color indicates density of points, and p-values of zero are displayed at the minimum nonzero value. See Methods for normalization, data processing, and hypothesis test. **(C)** Empirical cumulative distributions of expansion data from multiple time points (red) and replicate data (blue) show that many more tests in expansion data result in low p-values compared to replicate data. **(D)** Ratio of empirical cumulative distributions indicates that at a significance threshold of 10^{-300} there are roughly 12.3 times more positives than false positives.

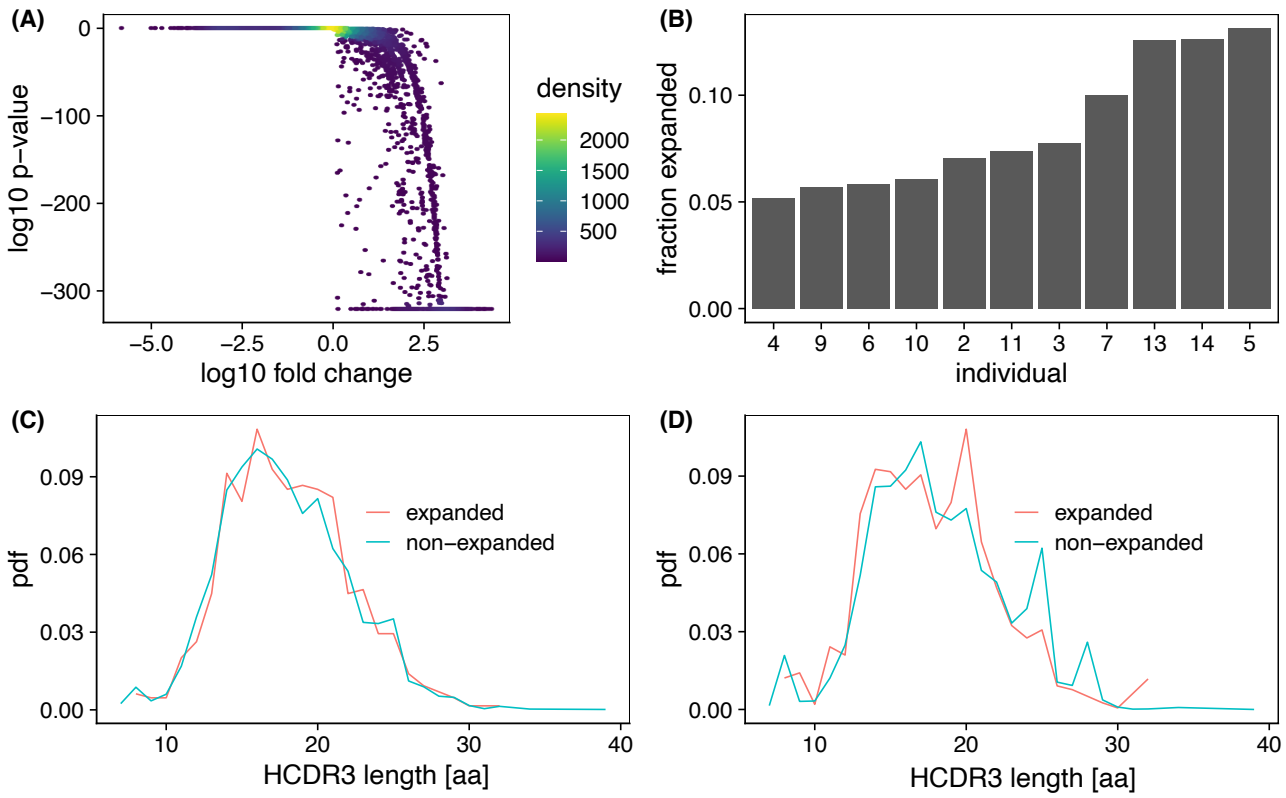

Figure S6. Statistics of expansion tests. (A) $\log_{10}$ p-values of the expansion test versus $\log_{10}$ fold change (or odds ratio) for data corresponding to Fig. 4B is shown. Color indicates density of points, and p-values of zero are displayed at the minimum nonzero value. See Methods for normalization, data processing, and hypothesis test. **(B)** Fraction of lineages expanded for different individuals is shown. HCDR3 length distributions of expanded and non-expanded lineages, **(C)** with each lineage having equal weight, and **(D)** with each lineage weighted by the number of unique sequences per time point (excluding singletons) are shown.

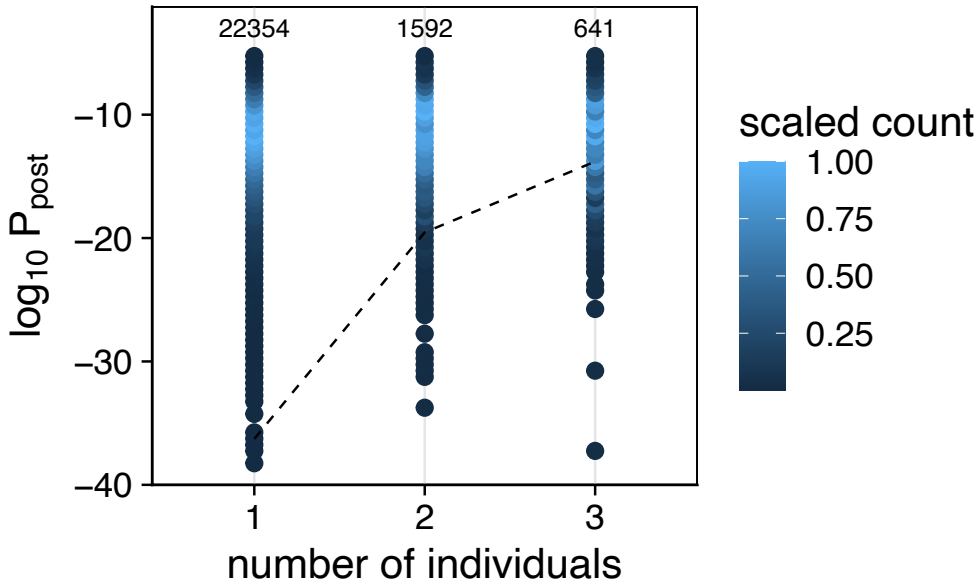

Figure S7. Sharing of BCRs among healthy individuals. The density plot shows the distribution of $\log_{10} P_{post}$ for progenitors of clonal lineages shared in a given number of healthy individuals, indicated on the horizontal axis; Histogram bin size is 0.5. The counts in each bin are scaled such that the maximum is equal to one for each column. The numbers above each column indicate the total number of sequences in the respective column. Sharing of rare lineages with $\log_{10} P_{post}$ below the dashed line is statistically significant (see Methods).

Severity	Patient (Sex, Age)	Time (replicate)	Abundance	Singletons	Unique sequences	Lineages (size > 2)	Lineages (size > 10)
Healthy	H1 (F, 28)	0 (1)	585886	67557	69985	623	282
	H2 (F, 30)	0 (1)	366626	54765	57112	826	369
	H3 (M, 45)	0 (1)	336339	42613	44106	469	221
Mild	1 (F, 62)	22 (1)	340773	62846	106373	2474	1221
		22 (2)	187755	38058			
		43 (1)	313753	55415	101573		
		43 (2)	216832	41867			
	2 (F, 37)	2 (1)	234000	44165	44814	3107	1271
		15 (1)	269510	38632	90655		
		15 (2)	136105	24335			
		15 (3)	122493	23898			
		34 (1)	877806	104879	209879		
		34 (2)	299677	48613			
		34 (3)	285678	47777			
Moderate	3 (M, 47)	8 (1)	513662	73778	164904	2718	1251
		8 (2)	127864	27294			
		8 (3)	325384	56658			
		38 (1)	571769	65093	132895		
		38 (2)	141638	24684			
		38 (3)	241868	37894			
	4 (F, 73)	18 (1)	317855	37128	60831	1222	441
		18 (2)	111427	21487			
		34 (1)	235260	39501	69040		
		34 (2)	159067	26999			
	5 (F, 72)	10 (1)	677259	97781	103198	2696	1216
		27 (1)	829769	105705	110757		
	6 (M, 56)	13 (1)	366489	62644	67365	2825	1241
		28 (1)	575454	71227	74550		
	7 (F, 55)	11 (1)	559836	85695	91002	5846	2677
		16 (1)	256747	45229	47957		
		39 (1)	1385280	183388	192451		
	8 (M, 37)	14 (1)	341447	60690	101153	35456	1627
		14 (2)	177401	35116			
		32 (1)	394388	58632	119555		
		32 (2)	367276	56757			

		37 (1)	330463	48171	88008		
		37 (2)	239495	36775			
	9 (F, 52)	5 (1)	88825	19062	67036	3026	1533
		5 (2)	281929	45587			
		8 (1)	337280	59940	114397		
		8 (2)	260919	49560			
		18 (1)	322636	57263	118522		
		18 (2)	310443	55570			
	10 (M, 27)	8 (1)	56146	7437	10332	917	405
		8 (2)	12101	2483			
		14 (1)	465583	54671	81512		
		14 (2)	138907	22548			
	11 (F, 20)	10 (1)	526891	75605	118907	3010	1421
		10 (2)	177963	36654			
		15 (1)	411802	71675	85572		
		15 (2)	25773	8636			
	12 (F, 48)	7 (1)	167253	16442	20963	1285	541
		7 (2)	9967	3982			
		11 (1)	485491	68900	130095		
		11 (2)	353809	55396			
	13 (F, 44)	7 (1)	502644	44087	79069	2868	1491
		7 (2)	269464	31602			
		13 (1)	556448	85160	172192		
		13 (2)	436959	75728			
	14 (F, 51)	3 (1)	556305	69852	137602	3207	1631
		3 (2)	387828	60574			
		7 (1)	707053	107229	191634		
		7 (2)	424161	73551			
Severe	15 (F, 68)	39 (1)	330027	52159	102887	2467	1232
		39 (2)	252573	44751			
		41 (1)	285227	40809	70733		
		41 (2)	111393	21975			
		41 (3)	22780	3473			
		71 (1)	2722	2550	32925		
		71 (2)	27529	8412			
		71 (3)	114391	12003			
		71 (4)	64427	8053			

	16 (M, 75)	44 (1)	393239	43502	84714	848	328
		44 (2)	249452	37010			
		53 (1)	319869	36988	93905		
		53 (2)	390043	52473			
	17 (M, 60)	22 (1)	217292	25681	88196	972	393
		22 (2)	445077	57879			
		36 (1)	579192	47906	63325		
		36 (2)	77254	10672			
		36 (3)	29979	2990			
	18 (F, 62)	8 (1)	268701	40873	70530	2142	938
		8 (2)	167760	26507			
		30 (1)	339144	48349	159879		
		30 (2)	248845	48291			
		30 (3)	333632	56928			
	19 (M, 56)	6 (1)	428221	69313	75128	1885	964
Unproductive	Pooled	Pooled	270632	79463	88312	4826	1065

Table S1. Statistics of BCR repertoire sequence data from healthy individuals, and COVID-19 patients. The Information about individuals in each cohort is shown. Detailed statistics of the processed data for productive BCRs are shown for read abundance, number of singletons, and number of unique sequences for all replicates and sampled time-points in each individual. For each individual, the number of lineages with more than two and ten unique sequences across all time points are shown in separate columns. Read statistics for unproductive receptors pooled from all individuals is shown separately.

5'-end primer	Sequence (5'-3')
IGHV1	CCTCAGTGAAGGTCTCCTGCAAGG
IGHV2	TCCTGCGCTGGTGAAACCCACACA
IGHV3	GGTCCCTGAGACTCTCCTGTGCA
IGHV4	TCGGAGACCCTGTCCCTCACCTGC
IGHV5	CAGTCTGGAGCAGAGGTGAAA
IGHV6	CCTGTGCCATCTCCGGGGACAGTG
3'-end primer	**Sequence (5'-3')**
CHG-R	GCGCCTGAGTTCCACGACAC

Table S2. List of primers used for PCR amplification of B-cell repertoires samples.

IGHV gene	IGHJ gene	HCDR3	# ind.	$\log_{10} P_{post}$
IGHV4-39	IGHJ3	CARLKRYYDSSGSPSAFDIW	12	-12.29
IGHV6-1	IGHJ5	CARITFAVIPPW	11	-14.61
IGHV4-61	IGHJ1	CARIGYDSSGYYYIFQHW	11	-12.81
IGHV4-39	IGHJ5	CARQKMGSVTIFGVVMHDNWFDPW	11	-17.13
IGHV4-39	IGHJ4	CARHVHMTWFEPSPGFDYW	11	-19.15
IGHV4-30-4	IGHJ4	CARDRGLGVIENFDYW	11	-12.72
IGHV6-1	IGHJ4	CARGGWGAPGSGRLFDYW	10	-14.45
IGHV4-61	IGHJ4	CAREVGSYYDWRPDYW	10	-13.20
IGHV4-59	IGHJ5	CARQDAWAYCGGDCYRGWFDPW	10	-15.24
IGHV4-59	IGHJ3	CARDHPYGDPRGAFDIW	10	-12.92
IGHV4-4	IGHJ5	CARGVAGEYWNDGKIRDFNWFDPW	10	-22.40
IGHV2-5	IGHJ4	CAHSIKPRRSGYCSGGSCYHFDYW	10	-17.44
IGHV4-59	IGHJ4	CARSRSGGSCYSFPSWYFDYW	9	-14.35
IGHV4-4	IGHJ6	CATLSRRAAAGTRDYYGMDVW	9	-14.30
IGHV4-34	IGHJ6	CARGQRSGACSGGSCVYYYGMDVW	9	-15.12
IGHV4-34	IGHJ6	CARGDDNYCTYSYGYGRLNRYYYYGMDVW	9	-22.99
IGHV2-70	IGHJ4	CARSYGTTVVTRAFDYW	9	-15.66
IGHV4-59	IGHJ4	CARQRKATVVTAILFDYW	8	-14.92
IGHV4-59	IGHJ4	CARHALGYFDWLNEGYFDYW	8	-15.36
IGHV4-39	IGHJ6	CARQPVLRYFDWLPWFGGMDVW	8	-14.45
IGHV4-39	IGHJ4	CVRSRAERYFDWTKLDAHVKPYYFDYW	8	-29.01
IGHV4-34	IGHJ5	CARESYAYTLTRGSTWFDPW	8	-16.22
IGHV4-34	IGHJ5	CAMGFHCSGGSCYDGVLGSW	8	-21.22
IGHV4-34	IGHJ3	CARGRRPVVVPAAGPAGRAFDIW	8	-15.62
IGHV4-34	IGHJ2	CARLNGWGIVVVIAIPRKNWYFDLW	8	-18.05
IGHV3-48	IGHJ4	CARDKYCSSTSCLHGRHYFDYW	8	-14.97
IGHV2-5	IGHJ4	CAQSNDVLRYFDWLYKMRGFDYW	8	-19.88
IGHV6-1	IGHJ6	CARGRMVPYSSSWTTLYYYYGMDVW	7	-16.47
IGHV4-59	IGHJ4	CARHRRLGLISDSSVDYW	7	-17.39
IGHV4-59	IGHJ4	CARGDVEMATIKSGGPFYHFDYW	7	-22.78
IGHV4-39	IGHJ3	CARRGRKLLPVAGTIDAFDIW	7	-15.83
IGHV4-34	IGHJ6	CARARTVTTFFLGVRVGMDVW	6	-17.99
IGHV4-34	IGHJ4	CARAIQSQHIVVVIAIPRPFYYFDYW	6	-18.23
IGHV4-28	IGHJ4	CATGFSGYSSSWYDKPLLEVEHAGPYFDYW	4	-30.77

Table S3. Rare expanding BCRs shared among individuals. The list of rare progenitors of clonal lineages (i.e., with P_{post} below the dashed line in Fig. 5) that exhibit lineage expansion in at least one individual is shown. These receptors are indicated by red diamonds in Fig. 5. The 34 rare expanding lineage progenitors shown here are shared among four to 12 COVID-19 patients.