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[10,11,12#], 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
10. Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
11. Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
12. Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.

* equal contribution

† current address: Dyno Therapeutics, 1 Kendall Square, Building 1400E, Suite 202, Cambridge, MA 02139

\# equal contribution and corresponding authors: Armita Nourmohammad: armita@uw.edu, Nicholas C. Wu: nicwu@illinois.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 SARS-CoV-2 specific monoclonal antibodies 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 bulk and plasma B-cells collected over multiple time points during infection to characterize signatures of B-cell response to SARS-CoV-2 in 19 patients. Using principled statistical approaches, we determined differential features of BCRs associated with different disease severity. We identified 38 significantly expanded clonal lineages shared among patients as candidates for specific responses to SARS-CoV-2. Using single-cell sequencing, we verified reactivity of BCRs shared among individuals to SARS-CoV-2 epitopes. Moreover, we identified natural emergence of a BCR with cross-reactivity to SARS-CoV-1 and SARS-CoV-2 in a number of patients. 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 120 million infections with a mortality rate around 2.2% (WHO, 2021). 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 elucidate the nature of COVID-19 and guide 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 react strongly to infecting antigens, resulting in B-cell stimulation, clonal expansion, and ultimately secretion of high-affinity antibodies in the blood (Burnet, 1959, 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 and 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 post-symptom onset (Perera et al., 2020). These antibodies bind to different antigens including the 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 (Barnes et al., 2020; Brouwer et al., 2020; Cao et al., 2020; Chi et al., 2020; Han et al., 2020; Hansen et al., 2020; Hurlburt et al., 2020; Ju et al., 2020; Kreer et al., 2020a; Kreye et al., 2020; Liu et al., 2020b; Noy-

Porat et al., 2020; Robbiani et al., 2020; Rogers et al., 2020; Seydoux et al., 2020a, 2020b; Shi et al., 2020; Wu et al., 2020b; Yuan et al., 2020; Zost et al., 2020). However, we still lack a comprehensive view of 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., 2020b; Robins, 2013). A few studies have performed BCR repertoire bulk sequencing to characterize the statistical signatures of the immune response to SARS-CoV-2 (Galson et al., 2020; Nielsen et al., 2020; Niu 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 responses specific to the infection. Moreover, they do not probe the composition of plasma B-cells during infection, which is the direct indicator of antibody production within an individual.

In this study, we have established a principled statistical approach to study the statistics and dynamics of bulk and plasma B-cell repertoires and to characterize the immune responses in 19 COVID-19 patients with different disease severities. 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 38 clonal lineages that are potential candidates for a response to SARS-CoV-2. Importantly, eight of these lineages contain BCRs from the plasma B-cell repertoire and, hence, are likely to have been secreting antibodies during infection. Moreover, using single-cell sequencing, we have verified the reactivity of BCRs shared among individuals to the epitopes of the receptor-binding domain (RBD) and the N-terminal domain (NTD) of SARS-CoV-2. Lastly, we identified cross-reactive responses to SARS-CoV-1 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

Strong correlation between composition of bulk and plasma B-cell repertoires. 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). To broaden our healthy control pool, we also incorporated into our analyses IgG B-cells from ten individuals in the Great Repertoire Project (GRP) (Briney et al., 2019). Sequence statistics for the first three biological replicates pooled together for each individual from the GRP are shown in Table S3 (see Methods). The patients showed different severities of symptoms, forming three categories of infected cohorts: two patients with mild symptoms, 12 patients with moderate symptoms, and five patients with severe symptoms. Specimens from all but one patient were collected over two or more time points during the course of the infection (Table S1). In addition to the bulk repertoire, we also isolated $CD38^{+}$ plasma B-cells from PBMC samples over at least two time points from seven patients in this cohort (six moderate, one severe) and from seven additional patients (two asymptomatic, three mild, two moderate) and three healthy individuals (Figure S1 and Table S4). The sampled time points for all patients in this study are indicated in Fig. 1 and Tables S1 and S4. 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 Tables S1 and S2. Statistical models were applied to analyze the length of the HCDR3 region, IGHV- or IGHJ-gene usage, and expansion and sharing of specific clonal lineages (Fig. 1).

The bulk repertoire is a collection of all BCRs circulating in the blood, including receptors from naïve, memory, and plasma B-cells. Plasma B-cells are actively producing antibodies and their receptors are more likely to be engaged in responding to an ongoing infection. Interestingly, the abundance of B-cell clonal lineages in the bulk and the plasma are strongly correlated (Fig. S3A), with Pearson correlations ranging from 0.55 – 0.88 across patients and significance $\mathrm{p-values} < 5 \times 10^{-8}$ across patients; correlations and p-values are given for each patient in Fig. S3. The significant correspondence between the bulk and plasma B-cell repertoires in Fig. 2 indicates that samples from the bulk, which cover a larger depth, are representative of functional immune responses, at least in the course of the infection.

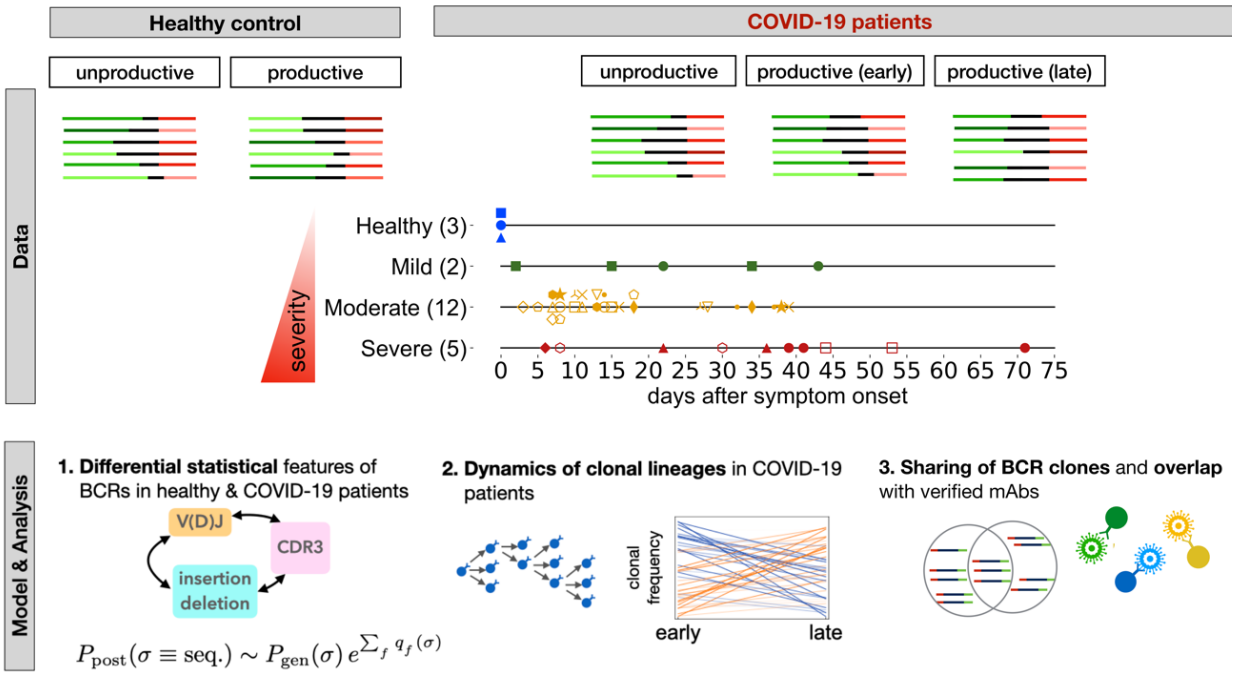

Figure 1. Roadmap for analysis of BCR repertoires. Top: We collected bulk blood IgG BCR samples from three healthy individuals and 19 COVID-19 patients where two patients had mild symptoms, 12 had moderate symptoms, and five had severe symptoms (different markers and colors); see Methods. We also collected $CD38^+$ plasma B-cells from PBMC samples of seven patients in this cohort (six moderate, one severe) and from seven additional patients (two asymptomatic, three mild, two moderate), and three healthy individuals (Fig. S2, Tables S1, S2). Samples were collected at different time points during infection (shown in center for bulk repertoires). We distinguished between productive receptors and unproductive receptors that had frameshifts due to V(D)J recombination. Line segments of varying lengths represent full V(D)J rearrangements (colors). In each patient, we constructed clonal lineages for productive and unproductive BCRs and inferred the naïve progenitor of the lineage (Methods). **Bottom:** 1. Using the set of unproductive inferred naïve BCRs, we inferred a model to characterize the null probability for generation of receptors $P_{\text{gen}}(\sigma)$(Marcou et al., 2018). We inferred a selection model (Sethna et al. 2020) to characterize the deviation from the null among inferred naïve productive BCRs, with the probability of entry to the periphery $P_{\text{post}}(\sigma)$ and selection factors $q_f(\sigma)$, dependent on receptor sequence features. 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 assessed the significance of these sharing statistics based on the probabilities to find each receptor in the periphery. The shared expanding clonal lineages that contain plasma B-cells, are likely candidates for secreting responsive antibodies during infection. We verified reactivity of receptors to SARS-CoV-2 antigenic epitopes using sorted single-cell data. We also identified previously characterized monoclonal antibodies (mAbs) specific to SARS-CoV-2 and SARS-CoV-1.

B-cell repertoires differ in receptor compositions across cohorts. We 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 and V- or J-gene usage) are often associated with their binding specificity, we used statistical methods to

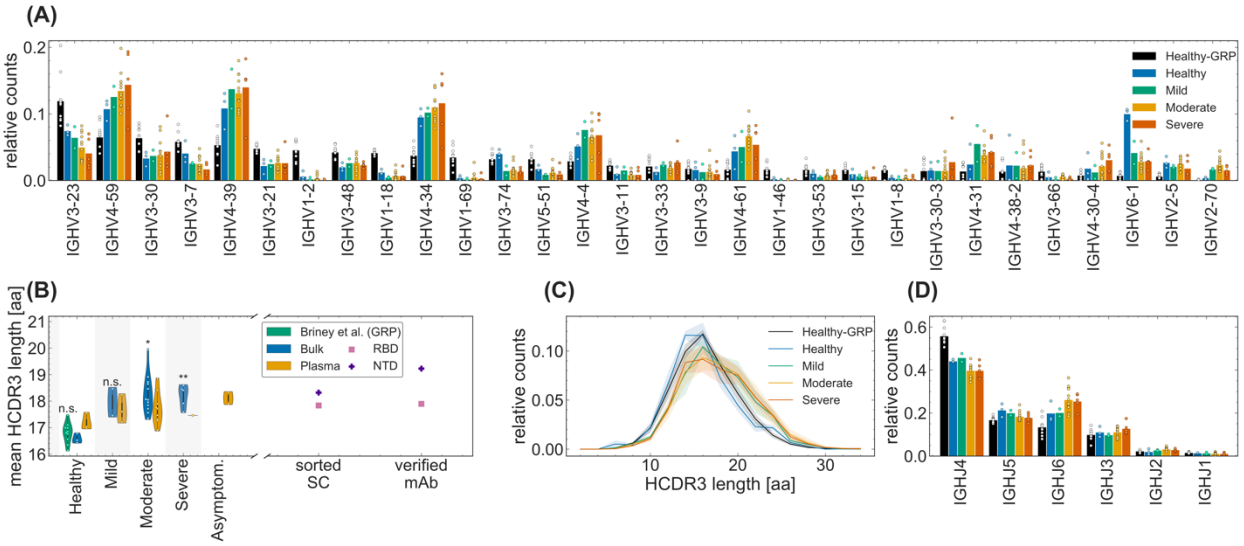

Figure 2. Sequence features of immune receptors in the bulk repertoire across cohorts. (A) The relative counts for IGHV-gene usage is shown for inferred naïve progenitors of clonal lineages in cohorts of healthy individuals and COVID-19 cohorts of patients exhibiting mild, moderate, and severe symptoms. 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, C)** Statistics of length of HCDR3 amino acid sequence is shown for different patients in each cohort. The violin plots in **(B)** show the mean HCDR3 length of each patient (dots) in a given cohort (color), with violin plot cut parameter set to 0.1. The mean HCDR3 lengths of the sorted single cells and verified monoclonal antibodies (axis) for RBD-reactive (pink squares) and NTD-reactive (purple pluses) receptors are shown on the right. Full lines in **(C)** show distributions averaged over individuals in each cohort (color), and shadings indicate regions containing one standard deviation of variation across individuals within a cohort. One-way ANOVA statistical tests were performed comparing the means HCDR3 of all the COVID-19 cohort and the healthy repertoires from Great Repertoire Project (GRP) dataset (Briney et al. 2019), with the healthy control from this study: Healthy-Mild: $F_{1,3} = 12.0$, p-value = 0.04; Healthy-Moderate: $F_{1,13} = 15.7$, p-value = 0.0016; Healthy-Severe: $F_{1,6} = 37.5$, p-value = 0.00087; Healthy-GRP: $F_{1,11} = 0.9$, p-value = 0.359. Significance cutoffs: n.s. $\mathrm{p-value} > 0.01$, * $\mathrm{p-value} \leq 0.01$, ** $\mathrm{p-value} < 0.001$. **(D)** The relative counts for IGHJ-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 J-gene frequencies across individuals within each cohort.

compare these features both at the level of clonal lineages, including the inferred receptor sequence of lineage progenitors in the bulk (Figs. 2, S2) and in the plasma B-cell repertoires (Figs. S3), and also the unique sequences in the bulk (Figs. S2) and in the plasma B-cell repertoires (Figs. S3); see Tables S1, S3, S4 for detailed statistics of clonal lineages in each individual.

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. In particular, the subset of lineages that contain plasma B-cell receptors can signal specific responses for antibody production against the infecting pathogen. Statistics of unique sequences in the bulk and the plasma B-cell repertoires, on the other hand, contain information about the size of the circulating lineages. Importantly, these statistical ensembles are relatively robust to PCR amplification biases that directly impact read abundances (see Methods for error correction and processing of 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 both in the bulk (Fig. 3A) and in the plasma B-cell repertoire (Fig. S3B). Data from unique sequences also indicated large background amplitudes due to vast differences in the sizes of lineages within a repertoire (Figs. S2A, S3E). Similarly, IGHJ-gene usage was also comparable across different cohorts for both bulk and plasma B-cell repertoires (Figs. 2D, S2C, and S3D,G). Moreover, we do not see a significant distinction in statistics of gene usage between the bulk and the plasma B cell repertoires (Figs. 2, S2 for bulk and Fig. S3 for plasma B-cells). 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 found that HCDR3s of lineages in COVID-19 patients with moderate and severe symptoms are significantly longer than in the healthy controls both from this study and from the GRP (Briney et al., 2019) (see Fig. 3B-C; One-way ANOVA statistics for differences in mean HCDR3 length: Healthy-Moderate: $F_{1,13} = 15.7$, p-value $= 1.6 \times 10^{-3}$; Healthy-Severe: $F_{1,6} = 37.5$, p-value $= 8.7 \times 10^{-4}$; GRP-Moderate: $F_{1,20} = 34.0$, p-value $= 1.1 \times 10^{-5}$; GRP-Severe: $F_{1,13} = 41.5$, p-value $= 2.2 \times 10^{-5}$). The difference between HCDR3 length in healthy individuals and patients with mild

symptoms were less significant. These differences are also observed at the level of unique productive BCRs (Fig. S2B). These findings are consistent with previous reports of longer HCDR3 lengths in COVID-19 patients (Galson et al., 2020; Nielsen et al., 2020; Schultheiß et al., 2020). It should be noted that despite differences in experimental protocols, the HCDR3 length of the healthy cohort from this study and from GRP (Briney et al., 2019) are comparable to each other (Figs. 2B-C, S2B). In addition, we found no significant difference between the HCDR3 length of the unproductive BCR repertoires of healthy individuals and COVID-19 patients (Figs. S2E), which should reflect biases in the generation of receptors prior to functional selection. Taken together, these finding indicate that BCRs with a longer HCDR3 tend to be preferentially elicited in repertoires of individuals responding to SARS-CoV-2 infections. This preference seems to have a functional significance as longer HCDR3 is also observed among monoclonal antibodies (mAbs) specific to the receptor binding domain (RBD) and the N-terminal domain (NTD) of SARS-CoV-2 (Fig. 2B) which were identified in previous studies (Brouwer et al., 2020; Han et al., 2020; Hurlburt et al., 2020; Kreye et al., 2020; Pinto et al., 2020; Robbiani et al., 2020; Wu et al., 2020b; Zost et al., 2020)**.**

Differential selection on B-cell repertoires in response to SARS-CoV-2. Longer HCDR3 sequences in COVID-19 patients can introduce more sequence diversity at the repertoire level. Quantifying sequence diversity of a B-cell repertoire can be very sensitive to the sampling depth in each individual. Despite progress in the quality of high-throughput repertoire sequencing 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; Sethna et al., 2020). To do so, we first used data from unproductive lineage progenitors of B-cell receptors in the bulk repertoire to infer the highly non-uniform baseline model that characterizes the probability $P_{\mathrm{gen}}(\sigma)$ to generate a given receptor sequence, dependent on its sequence features including the V-, D-, and J- gene choices and also the inserted and deleted sequences at the VD and DJ junctions (Elhanati et al., 2014; Marcou et al., 2018; Sethna et al., 2020) (Fig. 1 and Methods). The resulting model reflects the biased preferences in generating BCRs in the bone marrow by V(D)J recombination.

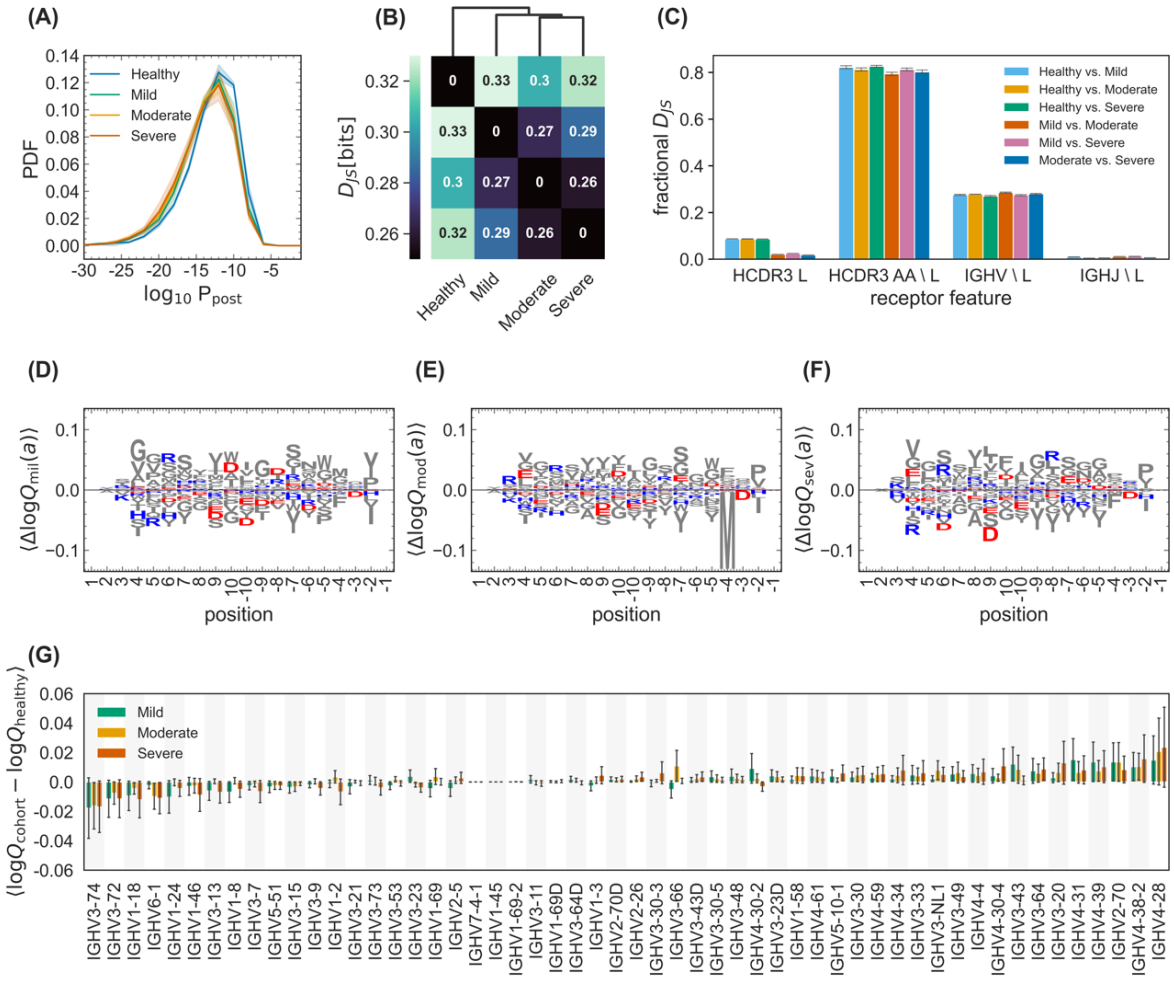

Figure 3. Differential statistics of immune repertoires across cohorts. (A) The distribution of the log-probability to observe a sequence σ in the periphery $\log_{10} P_{\text{post}}(\sigma)$ is shown as a normalized probability density function (PDF) for inferred naïve progenitors of clonal lineages 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)** Clustering of cohorts based on their pairwise Jensen-Shannon divergences D_{JS} as a measure of differential selection on cohorts is shown (Methods). **(C)** The bar graph shows how incorporating different features into a SONIA model contributes to the fractional Jensen-Shannon divergence between models trained on different cohorts. The error bars show the variations of these estimates over five independently inferred models (Methods). Logo plots show the expected differences in the log-selection factors for amino acid usage, $\langle \Delta \log Q_{\text{cohort}}(a) \rangle = \langle \log Q_{\text{cohort}}(a) - \log Q_{\text{healthy}}(a) \rangle$ for the **(D)** mild, **(E)** moderate, and **(F)** severe COVID-19 cohorts. The expectation values $\langle \cdot \rangle$ are evaluated on the mixture distribution $\frac{1}{2}\left(\mathrm{P}_{\text{post}}^{\text{cohort}} + \mathrm{P}_{\text{post}}^{\text{healthy}}\right)$. Positively charged amino acids (lysine, K; arginine, R; and histidine, H) are shown in blue while negatively charged amino acids (aspartate, D, and glutamate, E) are shown in red. All other amino acids are grey. Positions along the HCDR3 are shown up to 10 residues starting from the 3' (positive position values) and the 5' ends (negative position values). **(G)** The bar graph shows the average mean difference between the log-selection factors for IGHV-gene usage for the mild (green), moderate (yellow), and severe (red) COVID-19 cohorts, with the mean computed using the mixture distribution $\frac{1}{2}\left(\mathrm{P}_{\text{post}}^{\text{cohort}} + \mathrm{P}_{\text{post}}^{\text{healthy}}\right)$ and the average taken over the mean differences of 30 independently trained SONIA models for each cohort. Error bars show one standard deviation for the estimated mean, due to variations in the inferred SONIA models.

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 lineages 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. In addition, differential selection on receptor features can be used to quantify a distance between repertoires of different cohorts that reflect their functional differences in responses to immune challenges (Isacchini et al., 2021).

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, 2020b, 2021; 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, 2020b, 2021; Marcou et al., 2018; Sethna et al., 2020). Importantly, the inferred selection models are robust to the differences in the sample size of the repertoires, as long as enough data is available to train the models (Methods and Fig. S4C-F). As a result, selection models offer a robust approach to compare functional differences even between repertoires with widely different sample sizes, as is the case for our cohorts (Methods and Fig. S4C-F).

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). A similar pattern is observed at the level of generation probability $P_{\text{gen}}(\sigma)$ for functional receptors in the healthy versus COVID-19 infected individuals (Fig. S4A). Notably, the inferred selection models from the GRP healthy repertoires are comparable to the healthy cohort in this study (Fig. S4B). Thus, the

overabundance of rare receptors in COVID-19 patients is likely to be linked to functional responses associated with the stimulation of the repertoires against SARS-CoV-2.

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 bulk repertoires grows with severity of the disease, from 39.18 bits in the healthy cohort to 40.81 ± 0.03 bits in the mild cohort, to 41.03 ± 0.25 bits in the moderate cohort, and to 41.32 ± 0.11 bits in the severe cohort (see Methods). The error bars indicate variations over different models inferred in each of the COVID-19 cohorts, from repertoires subsampled to the same size as the healthy control (Methods). As indicated in Fig. S4, the models inferred from subsampled repertoires are highly consistent within each cohort.

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. To quantify the selection differences across cohorts, we evaluated the Jensen-Shannon divergence (D_{JS}) between repertoires of different cohorts, which measures the distance between the features of their receptor repertoire distributions (Isacchini et al., 2021) (Methods). Clustering of the cohorts based on their pairwise Jensen-Shannon divergences indicates that repertoires diverge with growing disease severity, and the COVID-19 cohorts are more similar with each other than with the healthy cohort (Fig. 3B, Methods).

The inferred selection models enabled us to quantify how different receptor features affect the pairwise divergence D_{JS} of BCR repertoires (Methods). In particular, we found that HCDR3 length contributes the most to differences in receptor distributions between the healthy and COVID-19 cohorts (Fig. 3C), consistent with the significant differences in the HCDR3 length distributions shown in Fig. 2C. In addition, we found that the amino acid composition of HCDR3 is the second most distinguishing factor between repertoires (Fig. 3C), indicating that negatively charged amino

acids are slightly suppressed at the center of HCDR3s in COVID-19 cohorts compared to healthy repertoires (Fig. 3D-F). The selection differences of IGHV- and IGHJ-gene usages between the healthy and the COVID-19 patients are insignificant (Figs. 3C,G), consistent with our previous analysis of lineage characteristics in Fig. 2A,D. Taken together, HCDR3 length and composition represents the molecular features that are most distinguishable at the repertoire level across different cohorts. Nonetheless, further work is necessary to understand the molecular underpinnings that may make these receptor features apt in response to a SARS-CoV-2 challenge.

Expansion of BCR clonal lineages over time indicates responses 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, S5). 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. S6). 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 in the bulk repertoire that appeared in two or more time points (Figs. 4A, S6 and Methods). Using a hypothesis test with a false discovery rate of 7.5%, as determined by analyzing replicate data (Methods, Fig. S6), we detected significant expansion of clonal lineages of receptors harvested from the bulk repertoire 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-

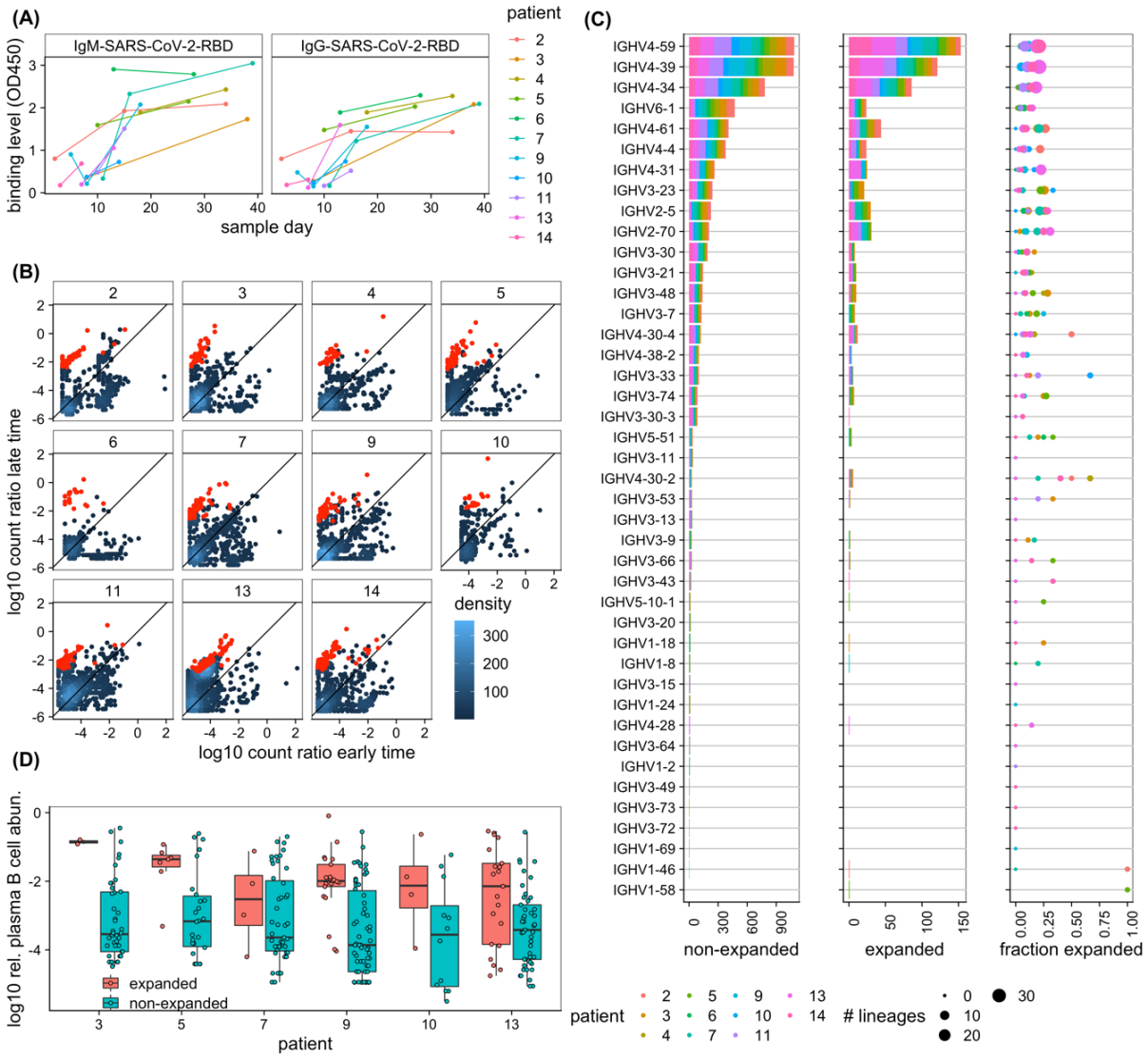

Figure 4. Dynamics of BCR repertoires during infection. (A) The binding level (measured by OD_{450} in ELISA assay) of the IgM (left) and IgG (right) repertoires to SARS-CoV-2 (RBD) epitopes increases over time in most individuals. **(B)** The log-ratio of BCR (mRNA) abundance at late time versus early 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. **(D)** Boxplot of $\log_{10}$ relative read abundance in the plasma B-cell (Methods) are shown for expanding (red) and non-expanding (cyan) lineages that contain reads from the plasma B-cell in different patients. Receptors from the plasma B-cell are significantly more abundant in expanding lineages in a number of patients based on the ANOVA test statistics: patient 3: $F_{1,42}$ = 5.4, p-value = 0.02; patient 5: $F_{1,31}$ = 0.5, p-value = 0.5; patient 7: $F_{1,49}$ = 0.01, p-value = 0.91; patient 9: $F_{1,42}$ = 4.1, p-value = 0.04; patient 10: $F_{1,42}$ =2.9, p-value = 0.1; patient 13: $F_{1,64}$ =7.7, p-value = 0.007.

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. We expect clonal expansions to reflect responses to SARS-CoV-2 during infection. Indeed, we observed that expanding lineages (based on the bulk data) show an over-representation of receptors harvested from plasma B-cells, which are likely to be associated with antibody-secreting B-cells (Fig. 4D and Methods); patient-specific significance p-values are reported in the caption of Fig. 4D.

Sharing of BCRs among individuals. Despite the vast diversity of BCRs, we observe a substantial number of identical progenitors of BCR clonal lineages among COVID-19 patients (Fig. 5) and among healthy individuals from our dataset and from the GRP (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) and 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 lines in Fig. S7). Out of a total of 40,128 (unique)

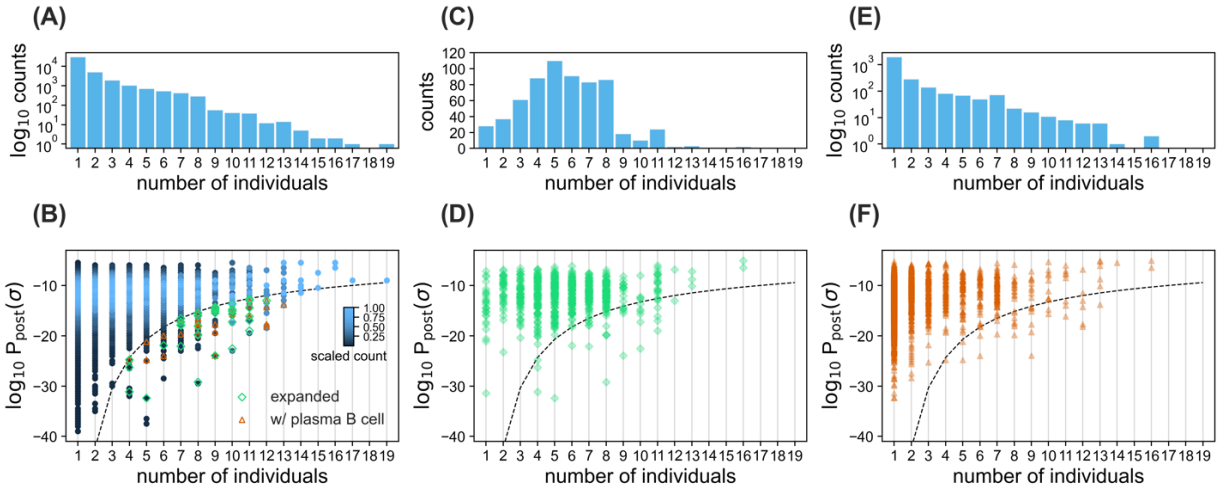

Figure 5. Sharing of BCRs among patients. (A) The histogram shows the number of clonal lineages that share a common progenitor in a given number of individuals, indicated on the horizontal axis. **(B)** The density plot shows the distribution of $\log_{10} P_{\text{post}}$ for progenitors of clonal lineages 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 the density in each column to one. Sharing of rare lineages with $\log_{10} P_{\text{post}}$ below the dashed line is statistically significant (Methods). Green diamonds indicate clonal lineages below the dashed line with significant expansion in at least one of the individuals. Orange triangles indicate clonal lineages below the dashed line that contain reads from the plasma B cell repertoire in at least one of the individuals. **(C, E)** The histograms show the number of clonal lineages that share a common progenitor in a given number individuals, which have significantly expanded during infection in at least one of the individuals (C), or contained reads from the plasma B-cell repertoire in at least one of the individuals (E). **(D, F)** The scatter plots with transparent overlapping markers show $\log_{10} P_{\text{post}}$ for progenitors of clonal lineages shared in a given number individuals that have expanded (D), or contain reads from the plasma B cell repertoire (E), in at least one individual. The dashed line is similar to (B).

progenitors of clonal lineages reconstructed from the pooled bulk+plasma B-cell repertoires (Fig. 5A, Tables S1, S4), we found 10,146 progenitors to be shared among at least two individuals, and 761 of these lineages contained receptors found in the plasma B-cell of at least one individual. 167 of the 10,146 lineage progenitors were classified as rare, having a probability of occurrence below the indicated threshold (dashed line) in Fig. 5B, with 30 of them containing receptors harvested from plasma B-cell, indicating a significant over-abundance of plasma B-cells among the rare, shared receptors (p-value=7.2×10^{-6}). Moreover, we found that 615 lineages shared a common sequence ancestor in at least two individuals and have expanded in at least one of the individuals (Fig. 5C-D). 38 of these shared, expanding lineages stemmed from rare naïve progenitors (below the dashed line in Fig. 5B, D), eight of which contain receptors found in the plasma B-cell of at least one individual. The over-abundance of plasma B-cell receptors in the rare, shared expanding

lineages is significant (p-value=0.04). The sharing of these rare, expanding BCRs among COVID-19 patients, with an over-abundance of receptors associated with antibody production in the plasma B-cell data, indicates a potentially convergent response to SARS-CoV-2; these receptors are listed in Table S5.

Interestingly, we found that 24% of receptors in the 38 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 = 0.013 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-CoV-1 specific neutralizing antibodies within repertoires.

To further investigate the functional response in the repertoire of COVID-19 patients, we performed single-cell sequencing on pooled samples from all patients, sorted for reactivity to RBD or NTD epitopes of SARS-CoV-2 (Methods). This analysis suggests that about 0.2% of these single cells are RBD-reactive as opposed to only 0.02% that are NTD-reactive (Fig. S1). This inferred fraction of reactive antibodies is consistent with previous estimates (Kreer et al., 2020a).

Next, we characterized the sequence features of RBD- and NTD-sorted antibodies. The IGHV-gene usage of these reactive receptors is shown in Fig. 6 and is compared to gene usage in monoclonal antibodies (mAbs) identified in previous studies (Brouwer et al., 2020; Han et al., 2020; Hurlburt et al., 2020; Kreye et al., 2020; Pinto et al., 2020; Robbiani et al., 2020; Wu et al., 2020b; Zost et al., 2020). Despite the broad range of IGHV-gene usages associated with epitope reactivity, sorted single-cell data show common IGHV-gene preferences to that of the previously

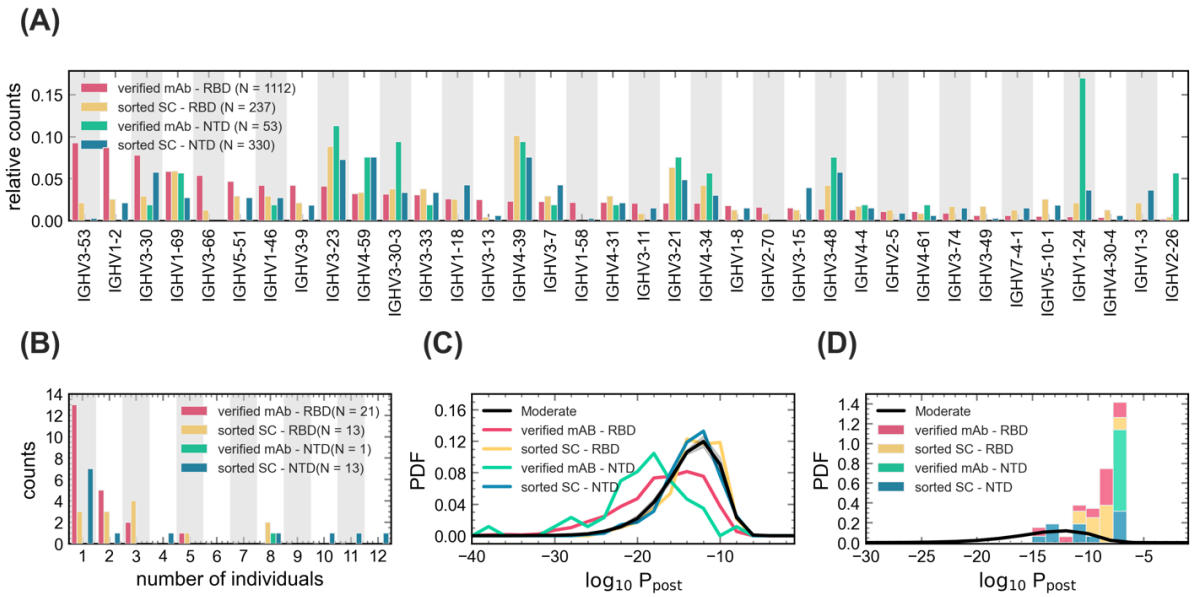

Figure 6: Statistics of BCRs reactive to RBD and NTD epitopes. (A) The relative counts for IGHV-gene usage is shown for known mAbs (Table S8) reactive to RBD (pink) and NTD (green) epitopes of SARS-CoV-2 and for receptors obtained from single cell sequencing of the pooled sample from all patients (Methods), sorted for RBD (yellow) and NTD (blue) epitopes. **(B)** The histogram shows the number of NTD-sorted receptors from single cell sequencing (Table S6) and RBD- and NTD-specific verified mAbs (Table S7) found in the bulk+plasma B-cell repertoires of a given number of individuals (Methods), indicated on the horizontal axis. **(C)** The distribution of the log-probability to observe a sequence σ in the periphery $\log_{10} P_{\text{post}}(\sigma)$ is shown as a normalized probability density function (PDF) for inferred naïve progenitors of known RBD- and NTD-specific mAb and for RBD- and NTD-sorted receptors from single cell sequencing. $P_{\text{post}}(\sigma)$ values were evaluated based on the repertoire model created from patients with moderate symptoms. The corresponding $\log_{10} P_{\text{post}}$ distribution for bulk repertoires of the moderate cohort (similar to Fig. 3A) is shown in black as a reference. **(D)** Similar to (C) but restricted to receptors that are found in the bulk+ plasma B-cell repertoire of at least one patient in the cohort (Tables S6, S7). Colors are consistent between panels and the number of samples used to evaluate the statistics in each panel is indicated in the legend.

identified mAbs against SARS-CoV-2 epitopes. This includes an abundance of IGHV1-69, IGHV4-59, IGHV3-30-3, IGHV3-33, IGHV1-18, IGHV5-51, and IGHV1-46 against RBD, and IGHV3-23, IGHV4-59, IGHV4-39, IGHV3-21, and IGHV3-48 against NTD (Fig. 7A). Similarly, we observe consistent biases in V- and J- gene usages of the κ and λ light chains for the sorted single-cell data and the verified mABs (Fig. S8). Moreover, the HCDR3 length distributions of the sorted single-cell data are comparable to those of the verified mABs (Fig. S8). The average length of the HCDR3 for both the verified mAbs and the sorted single-cell receptors are comparable to that of bulk repertoires from COVID-19 patients, which is significantly longer than that of healthy individuals (Fig. 2B).

To characterize how SARS-CoV-2 reactive receptors make up the patients' repertoires, we mapped the heavy chain receptors from the sorted single-cell data onto BCR lineages constructed from the bulk+plasma B-cell data in the COVID-19 patients (Methods, Table S6). We found that 13 (from 237) RBD-sorted and 13 (from 330) NTD-sorted antibodies from the single-cell data matched receptor lineages in at least one individual (Fig. 6B). Interestingly, we found a broad sharing of these antibodies with 10 RBD- and 6 NTD-sorted single cells present in at least two patients (Fig. 6B).

In repertoires of the COVID-19 patients, we found that several HCDR3s matched with SARS-CoV-2-specific mAbs that were previously isolated in other studies (Brouwer et al., 2020; Han et al., 2020; Hurlburt et al., 2020; Kreye et al., 2020; Pinto et al., 2020; Robbiani et al., 2020; Wu et al., 2020b; Zost et al., 2020). Specifically, a total of 20 mAb families specific to SARS-CoV-2 epitopes were found to be close in sequence to HCDR3s in our data (with up to one amino acid difference), among which are 14 RBD-specific, one NTD-specific, and five S1-specific (reactive to either RBD or NTD) mAbs (Fig. 7B, Table S7). Interestingly, nine of these mAbs are shared among at least two individuals, and the NTD-specific antibody is found in eight individuals (Fig. 7B).

In addition, we found that two patients had exact HCDR3 matches to a previously identified antibody, S304, that has cross-reactivity to SARS-CoV-1 and SARS-CoV-2 (Pinto et al., 2020). We also observed in one patient an HCDR3 with only one amino acid difference to this antibody (Table S7). Importantly, the plasma in these patients showed a substantial binding level (OD_{450}) to SARS-CoV-1 (Fig. S5), which indicates a possibility of cross-reactive antibody responses to SARS-CoV-1 and SARS-CoV-2.

We also investigated the matches between the RBD- and NTD-sorted single-cell receptors with the verified mAbs from previous studies (Brouwer et al., 2020; Han et al., 2020; Hurlburt et al., 2020; Kreye et al., 2020; Pinto et al., 2020; Robbiani et al., 2020; Wu et al., 2020b; Zost et al., 2020). Although we found no matches between the heavy chain CDR3 of sorted single-cell receptors and the verified mAbs, we found a large number of matches between the κ and λ light

chain CDR3s of the sets (Fig. S8). Notably, 59 of 142 IG_{κ} and 47 of 110 IG_{λ} from the RBD-reactive single cells, and 1 of 202 IG_{κ}, and 22 of 155 IG_{λ} from the NTD-reactive single cells matched to light chain CDR3s of mAbs in those respective subsets (Fig. S8). Given the low sequence diversity of light chain receptors, it remains to be seen as to whether these matches between the light chain mAbs and sorted single-cell data are statistically significant—a question that would require modeling the generation and selection of the light chain receptors' repertoire.

Lastly, we observed that the previously verified mAbs have a lower probability $P_{\text{post}}(\sigma)$ of generation and entry to the periphery compared to the overall repertoire (Fig. 6C). This is in part expected since the selection models used to evaluate these probabilities were trained on different repertoires than those from which the mAbs were originally harvested. Consistently, the evaluated probabilities for the sorted single sorted receptors are within the range for the bulk repertoire (Fig. 6C), as the two datasets were derived from the same cohort. It should also be noted that all of the verified mAbs and the sorted receptors from the single-cell data that we can match to the patients' repertoires have a relatively high probability $P_{\text{post}}(\sigma)$ (Fig. 6D). This is not surprising as it is very unlikely to observe rare BCRs (with small $P_{\text{post}}(\sigma)$) to be shared in across different cohorts. Overall, our results are encouraging for vaccine development since they indicate that even common antibodies can confer specific responses against SARS-CoV-2.

Discussion

COVID-19 will remain an ongoing threat to public health until an effective SARS-CoV-2 vaccine is available globally. Understanding the human B-cell immune response to SARS-CoV-2 is critical for vaccine development and assessment (Wec et al., 2020a). A repertoire of immune receptor sequences represents a unique snapshot of the history of immune responses in an individual (Boyd et al., 2009; Georgiou et al., 2014; Kreer et al., 2020b; 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 a 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 systematically characterize 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 the HCDR3 of BCRs in COVID-19 patients are significantly longer than HCDR3 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.

Bulk B-cell repertoires predominantly contain a mixture of naïve, memory, and plasma-B cells. At the early stages of viral infection, antigen-specific plasma B-cells may develop, which act as antibody factories and confer neutralization against the infecting pathogen (Wrammert et al., 2008). Almost all prior work on immune repertoires has focused on bulk repertoires, which are often easier to sample from and to analyze. Moreover, functional studies, using single-cell sequencing of antigen-sorted B-cell receptors, have often been disconnected from the large-scale analysis of receptor repertoires. Our study synergizes data from bulk and plasma B-cell sequencing with antigen-sorted single-cell B-cell receptors to draw a more complete picture of the human immune response to SARS-CoV-2. Importantly, our joint longitudinal analysis of the bulk and the plasma B-cell repertoires in COVID-19 patients brings insight into the dynamics of antigen-specific B-cells as well as the statistics of receptor sequence features associated with 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' plasma 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 38 clonal lineages that are candidates for a response specific to SARS-CoV-2 antigens (Fig. 5, Table S5). Importantly, the over-representation of plasma B-cells among these shared expanding lineages signifies their potential role in mounting protective antibody responses against SARS-CoV-2. It should be noted that none of these 38 clonal lineages matched with the verified mAbs. This is in part expected since the verified mAbs that matched the bulk repertoires have relatively high probabilities P_{post} (Fig. 6C), whereas these 38 lineages are chosen explicitly to be rare.

Our analysis of repertoire dynamics has identified a large-scale expansion of B-cell clonal lineages (5 -15% of lineages) over the course of COVID-19 infections. However, it is hard to imagine that all of these expanding clones that account for a sizeable portion of the repertoire are engaged in responding to SARS-CoV-2 specifically. In contrast, our single-cell analysis identified only about 0.2% of receptors as reactive to RBD and only 0.02% as reactive to NTD epitopes (Fig. S1)—an estimate that is consistent with previous findings (Kreer et al., 2020a). This disparity raises an outstanding question: why do we observe such a large-scale expansion of clonal lineages during an acute immune response?

Identifying antibodies with cross-reactive neutralization abilities against viruses in the SARS family is of significant interest. While cross-neutralization antibodies have been isolated from COVID-19 patients (Brouwer et al., 2020; Liu et al., 2020a; Zhou et al., 2020), it remains unclear how prevalent they are. Interestingly, in nine patients, we see a substantial increase in the binding level (OD_{450}) of their plasma to SARS-CoV-1 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-1

infection. This antibody was shown to be moderately cross-reactive to both SARS-CoV-1 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

▪ Data and code availability

BCR repertoire data and single-cell data can be accessed through:
https://www.ncbi.nlm.nih.gov/bioproject/PRJNA645245
https://www.ncbi.nlm.nih.gov/bioproject/PRJNA679920

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

▪ Experimental Procedures

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.

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: 98ºC 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).

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-1 spike protein (GenBank: ABF65836.1) were cloned into a customized pFastBac vector (Lv et al., 2020; Wec et al., 2020b). 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).

$CD38^+$ plasma B-cell enrichment. $CD38^+$ plasma B-cells were isolated from the PBMC samples by performing two subsequent magnetic separation steps according to the manufacturer's protocol

(Plasma Cell Isolation Kit II, human, Miltenyi Biotec). Briefly, non-plasma B-cells are labeled with magnetic beads combined with cocktail antibodies and separated using the MACS column. Then, $CD38^+$ plasma B-cell are directly labeled with CD38 MicroBeads and isolated from the pre-enriched B cell pool. Purified $CD38^+$ plasma B-cell were eluted and washed in PBS containing 2% (v/v) fetal bovine serum (FBS) and kept for the following RNA isolation step. In order to test the purity of the $CD38^+$ plasma B cells, we also added staining antibodies and 10 µl of Anti-human CD19-BV510 (BioLegend) and CD38-PE-Cy7 (BioLegend) and incubated them for 15 minutes in the dark in the refrigerator (2-8°C). Cells were finally fixed with 4% PFA for 20 minutes on ice. The stained samples were acquired by flow cytometry on a FACS Attune (Invitrogen) and analyzed with FlowJo software (Fig. S1).

RBD and NTD protein specific binding B cell enrichment. B-cells were enriched from the PBMC samples according to the manufacture's protocol (B Cell Isolation Kit II, human, Miltenyi Biotec). Briefly, non-B-cells are labeled with a cocktail of biotin-conjugated antibodies and separated by the MACS column. Purified B-cells were eluted and kept in the PBS buffer with 2% (v/v) FBS. The enriched B cells were then incubated with 2 µg Biotin-RBD or NTD protein for 30 min at 4°C. After incubation, Anti-Biotin MicroBeads were added and incubated for 30 min. RBD and NTD specific bead binding B cells were washed and eluted in PBS and stored on ice until use. In order to test the purity of the RBD- or NTD-specific B cells, we also added staining antibodies, 10 µl of Anti-human CD19-BV510 (BioLegend), and 2 µg of SARS-CoV-2 RBD-PE or NTD-PE and incubated them for one hour in the dark in the refrigerator (2-8°C). Cells were finally fixed with 4% PFA for 20 minutes on ice. The stained samples were acquired by flow cytometry on a FACS Attune (Invitrogen) and analyzed with FlowJo software (Fig. S1).

Single B cell 5' mRNA and VDJ sequencing. After RBD or NTD specific B-cells enrichment, cells were counted by using 0.4% (w/v) trypan blue stain solution in the microscope and directly loaded on the 10X Chromium™ Single Cell A Chip. Then single B cell lysis and RNA first-strand synthesis were carried out following the 10X Chromium™ Single Cell 5′ Library & Gel Bead Kit protocol. The RNA sample were used for the next step B cell VDJ library construction following the Chromium™ Single Cell V(D)J Enrichment Kits protocol. VDJ library sequencing was performed on a NovaSeq PE150 and the sequencing data were processed by Cell Ranger.

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.

▪ Statistical Inference and Methods

BCR preprocessing. We used a similar procedure for processing of the bulk and the plasma B-cell receptor repertoires. 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 (Tables S1, S4). We discarded these singletons from all analyses that relied on statistics of unique sequences (i.e., the results presented in Figs. S2A-C and S3E-G).

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 IGHV gene/allele, IGHJ 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 IGHV gene, IGHJ 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. Defining size as the sum of the number of unique sequences per time point within a lineage, clusters of size smaller than three were discarded from most analyses. They were retained only for training IGoR and SONIA models and were not discarded in the sharing analysis only if the progenitor of that small cluster was also a progenitor of a cluster of size at least three in another patient. 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. Tables S1 and S4 show the statistics of constructed clonal lineages in each individual for the bulk repertoire and combined bulk+plasma B-cell repertoire, respectively.

Mapping of single-cell data onto reconstructed clonal lineages: Like the repertoire datasets, the single-cell sequences were annotated by abstar (Briney and Burton, 2018). For each receptor acquired by single-cell sequencing, we identified a subset of reconstructed clonal lineages from the bulk repertoire which had identical HCDR3 length as the sequence and which also had an IGHV gene which was 90% similar to that of the single-cell receptor. This flexibility in V-gene choice would identify functionally homologous receptors and associate a receptor to a lineage with a sequence divergence in the V-segment, compatible with the expectation under somatic hypermutations (Lee et al., 2017). A single-cell sequence was matched to a reconstructed clonal

lineage from this subset if its HCDR3 could be clustered with other members of the lineages, using single-linkage clustering with a similarity threshold of 85% Hamming distance (similar to the criteria for lineage reconstruction for bulk repertoires).

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 IGHV, IGHD, and IGHJ 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 unproductive lineages, regardless of size, pooled from the bulk repertoire of all individuals, restricted to progenitors whose HCDR3 began with a cysteine and ended with a tryptophan. 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. $P_{\text{gen}}(\sigma)$ distributions of the healthy and COVID-19 cohorts in this study are shown in Fig. S4A.

We used SONIA (version 0.45) (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 IGHV-gene and IGHJ-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 IGHV- and IGHJ-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 productive lineages, regardless of size, pooled from the bulk repertoire of all individuals within a cohort, restricted to progenitors whose HCDR3 began with a cysteine and ended with a tryptophan. 150 epochs, L_2 regularization with strength 0.001, 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. A SONIA model was also trained on all the productive lineage progenitors in the GRP dataset (Briney et al., 2019) and used 5,000,000 generated sequences, keeping the other parameters unchanged. We refrain from comparing directly $P_{\text{post}}(\sigma)$ associated with GRP BCRs to BCRs in this study due to experimental differences.

It should be noted that the (pre-selection) generation model $P_{\text{gen}}(\sigma)$ inferred by IGoR (Marcou et al., 2018) is robust to sequence errors due to experimental errors or hypermutations in the IgG repertoires. However, hypermutations in BCRs could introduce errors in inference of selection models and estimation of receptor probabilities by SONIA (Sethna et al., 2020). Therefore, we have restricted our selection analyses to only the inferred progenitors of clonal lineages. Although the inferred progenitors of lineages can still deviate from the *true* (likely IgM naïve) progenitors, the selection models inferred from *ensembles* of inferred progenitors in IgG repertoires seem to be comparable to the models inferred from the IgM repertoires (Ruiz Ortega et al., 2021). The resulting selection models, trained on either true or inferred progenitors, reflect preferences for sequence features of unmutated receptors, including IGHV- and IGHJ- genes and HCDR3 length and composition, but they do not account for the hypermutation preferences that may distinguish one cohort from another.

Characterizing the robustness of selection inference. To test the sensitivity of the inferred selection models on the size of the training sets, we down-sampled the receptor data of each COVID-19 cohort to a size comparable to the smallest cohort, i.e., the healthy repertoire sequenced in this study. This down-sampling resulted in two independent training datasets for the mild COVID-19 cohort, 13 independent training datasets for the moderate COVID-19 cohort, and three independent training datasets for the severe COVID-19 cohort. Though this down-sampling resulted in over 400 independent training datasets for the GRP, we elected to use only 15. We then inferred a separate selection model with SONIA for each of these training datasets and used each model to evaluate the receptor log-probabilities $\log_{10} P_{\text{post}}(\sigma)$ for a set of 500,000 generated receptors. The evaluated probabilities are strongly correlated between models inferred from the

down-sampled data in each cohort, with a Pearson correlation of $r > 0.99$ and p-value = 0 (Fig. S4C-F).

We used a similar approach to compare the selection model inferred from the healthy repertoires sequenced in this study and the GRP study (Briney et al., 2019). Fig. S4B shows that, using the model inferred with our healthy repertoire and 30 down-sampled independently inferred selection models using the GRP dataset, the evaluated log-probabilities $\log_{10} P_{\text{post}}(\sigma)$ based on these two datasets are strongly correlated, with a Pearson correlation of $r > 0.99$ and p-value = 0 (Fig. S4B).

Characterizing repertoire diversity. We quantified the diversity of each cohort by evaluating the entropy of receptor sequences in each cohort. Entropy can be influenced by the size of the training dataset for the selection models. To produce reliable estimates of repertoires' diversities (and entropies), we used the procedure described above to learn independent selection models for subsampled repertoires in each cohort. We then used the inferred IGoR and SONIA models to generate 500,000 synthetic receptors based on each of the subsampled, cohort-specific models. We evaluated cohort entropies H as the expected log-probabilities to observe a functional sequence in the respective cohort: $H = -\sum_{\sigma} P_{\text{post}}(\sigma) \log P_{\text{post}}(\sigma)$; the estimates based on the generated receptors are reported in the main text. The error bars reported for these entropy estimates are due to variations across the inferred models in each cohort.

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, 42.7 ± 0.3 bits for patients in the moderate cohort, and 42.9 ± 0.5 for patients in the severe cohort. The error bars indicate the standard error due to differences among individuals within a cohort.

Comparing selection between repertoires of cohorts. Selection models enable us to characterize the sequence features of immune repertoires that differ between cohorts. We evaluated the Jensen-Shannon divergence $D_{JS}(r, r')$ between the distribution of repertoires r and r', P_{post}^{r} and $P_{\text{post}}^{r'}$, defined as

$$
\begin{aligned}
D_{JS}(r,r') &= \frac{1}{2}\sum_{\sigma:\,\text{sequences}} P^{r}_{\text{post}}(\sigma)\log\frac{P^{r}_{\text{post}}(\sigma)}{\left(P^{r}_{\text{post}}(\sigma)+P^{r'}_{\text{post}}(\sigma)\right)/2} + \frac{1}{2}\sum_{\sigma:\,\text{sequences}} P^{r'}_{\text{post}}(\sigma)\log\frac{P^{r'}_{\text{post}}(\sigma)}{\left(P^{r}_{\text{post}}(\sigma)+P^{r'}_{\text{post}}(\sigma)\right)/2} \\
&= \frac{1}{2}\sum_{\sigma:\,\text{sequences}} P^{r}_{\text{post}}(\sigma)\log\frac{2\,Q^{r}(\sigma)}{Q^{r}(\sigma)+Q^{r'}(\sigma)} + \frac{1}{2}\sum_{\sigma:\,\text{sequences}} P^{r'}_{\text{post}}(\sigma)\log\frac{2\,Q^{r'}(\sigma)}{Q^{r}(\sigma)+Q^{r'}(\sigma)}
\end{aligned}
$$

where we used the relationship between a receptor's generation probability $P_{\text{gen}}(\sigma)$ and its probability after selection $P^{r}_{\text{post}}(\sigma)$, using the inferred selection factor $Q^{r}(\sigma) = \frac{1}{Z}\, e^{\Sigma_{f:\text{features}} q^{r}_{f}(\sigma)}$ in repertoire r: $P^{r}_{\text{post}}(\sigma) = P_{\text{gen}}(\sigma)\, Q^{r}(\sigma)$. The Jensen-Shannon divergence $D_{JS}(r,r')$ is a symmetric measure of distance between two repertoires, which we can calculate using their relative selection factors (Isacchini et al., 2021). Fig. 3 shows the expected partial Jensen-Shannon divergences evaluated over five independent realizations of 100,000 generated sequences for each partial selection model. The error bars show the variations of these estimates over the five independent realizations in this procedure.

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. 5A, S5): 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. S5). 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 less than three, 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. S8). 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 $\frac{\frac{n_i^{\text{late}}}{n_i^{\text{early}}}}{\frac{N_{/i}^{\text{late}}}{N_{/i}^{\text{early}}}}$ 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. S6G.

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 performed the expansion test on pairs of replicates (Fig. S6C) and compared the empirical cumulative distributions of the time point and replicate expansion data (Fig. S6E,F) (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 a 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. 6 and Fig. S11) at the 2% 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 2% of the data lay below P_{share}. This was a conservative choice to identify the rare shared outliers in the data.

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.), Royalty Research Fund 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.), a startup fund at the University of Illinois at Urbana-Champaign (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.

Author Contributions

Z.M., H.Lv, J. O, I.A.W., J.S.M.P. N.C.W., A.N., and C.K.P.M. conceived and designed the study. O.T.-Y.T. organized patient recruitment, data collection, and sampling. H.Lv, G.K.Y., W.W.N., and C.K.P.M. prepared the next-generation sequencing libraries and performed the ELISA experiments. M.Y., H.Liu, and N.C.W. and expressed and purified the proteins. Z.M., J.O., W.S.D., G.I., and A.N. analyzed the data and performed the modelling work and statistical inference. Z.M., H.Lv, N.C.W., A.N., and C.K.P.M. wrote the paper. All authors reviewed and edited the paper.

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.

Barnes, C.O., West, A.P., Huey-Tubman, K.E., Hoffmann, M.A.G., Sharaf, N.G., Hoffman, P.R., Koranda, N., Gristick, H.B., Gaebler, C., Muecksch, F., et al. (2020). Structures of human antibodies bound to SARS-CoV-2 spike reveal common epitopes and recurrent features of antibodies. Cell *182*, 828-842.e16.

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.

Briney, B., Inderbitzin, A., Joyce, C., and Burton, D.R. (2019). Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature *566*, 393–397.

Brouwer, P.J.M., Caniels, T.G., Straten, K. van der, 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 *369*, 643–650.

Burnet, F.M. (1959). The clonal selection theory of acquired immunity (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.

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.

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 *383*, 1522–1534.

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.

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., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., 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.

Han, X., Wang, Y., Li, S., Hu, C., Li, T., Gu, C., Wang, K., Shen, M., Wang, J., Hu, J., et al. (2020). A rapid and efficient screening system for neutralizing antibodies and its application for the discovery of potent neutralizing antibodies to SARS-CoV-2 S-RBD. BioRxiv 10.1101/2020.08.19.253369.

Hansen, J., Baum, A., Pascal, K.E., Russo, V., Giordano, S., Wloga, E., Fulton, B.O., Yan, Y., Koon, K., Patel, K., et al. (2020). Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. Science *369*, 1010–1014.

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.

Hurlburt, N.K., Seydoux, E., Wan, Y.-H., Edara, V.V., Stuart, A.B., Feng, J., Suthar, M.S., McGuire, A.T., Stamatatos, L., and Pancera, M. (2020). Structural basis for potent neutralization of SARS-CoV-2 and role of antibody affinity maturation. Nat Commun *11*, 5413.

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

Isacchini, G., Olivares, C., Nourmohammad, A., Walczak, A.M., and Mora, T. (2020b). SOS: online probability estimation and generation of T-and B-cell receptors. Bioinformatics *36*, 4510–4512.

Isacchini, G., Walczak, A.M., Mora, T., and Nourmohammad, A. (2021). Deep generative selection models of T and B cell receptor repertoires with soNNia. Proc Natl Acad Sci U S A *118*, e2023141118.

Janeway, C.A., Travers, P., Walport, M., and Shlomchik, M.J. (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 *584*, 115–119.

Kreer, C., Zehner, M., Weber, T., Ercanoglu, M.S., Gieselmann, L., Rohde, C., Halwe, S., Korenkov, M., Schommers, P., Vanshylla, K., et al. (2020a). Longitudinal isolation of potent near-germline SARS-CoV-2-neutralizing antibodies from COVID-19 patients. Cell *182*, 843-854.e12.

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

Kreye, J., Reincke, S.M., Kornau, H.-C., Sánchez-Sendin, E., Max Corman, V., Liu, H., Yuan, M., Wu, N.C., Zhu, X., Lee, C.-C.D., et al. (2020). A SARS-CoV-2 neutralizing antibody protects from lung pathology in a COVID-19 hamster model. BioRxiv 10.1101/2020.08.15.252320.

Lee, D.W., Khavrutskii, I.V., Wallqvist, A., Bavari, S., Cooper, C.L., and Chaudhury, S. (2017). BRILIA: Integrated Tool for High-Throughput Annotation and Lineage Tree Assembly of B-Cell Repertoires. Front Immunol *7*.

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, H., Wu, N.C., Yuan, M., Bangaru, S., Torres, J.L., Caniels, T.G., van Schooten, J., Zhu, X., Lee, C.-C.D., Brouwer, P.J.M., et al. (2020a). Cross-neutralization of a SARS-CoV-2 antibody to a functionally conserved site is mediated by avidity. BioRxiv 10.1101/2020.08.02.233536.

Liu, L., Wang, P., Nair, M.S., Yu, J., Rapp, M., Wang, Q., Luo, Y., Chan, J.F.-W., Sahi, V., Figueroa, A., et al. (2020b). Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. Nature *584*, 450–456.

Lv, H., Wu, N.C., Tsang, O.T.-Y., Yuan, M., Perera, R.A.P.M., 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 Reports *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., Jackson, K.J.L., Hoh, R.A., Röltgen, K., Jean, G.H., Stevens, B.A., Lee, J.-Y., Rustagi, A., Rogers, A.J., et al. (2020). Human B cell clonal expansion and convergent antibody responses to SARS-CoV-2. Cell Host & Microbe *28*, 516-525.e5.

Niu, X., Li, S., Li, P., Pan, W., Wang, Q., Feng, Y., Mo, X., Yan, Q., Ye, X., Luo, J., et al. (2020). Longitudinal Analysis of T and B Cell Receptor Repertoire Transcripts Reveal Dynamic Immune Response in COVID-19 Patients. Front Immunol *11*, 582010.

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.

Noy-Porat, T., Makdasi, E., Alcalay, R., Mechaly, A., Levy, Y., Bercovich-Kinori, A., Zauberman, A., Tamir, H., Yahalom-Ronen, Y., Israeli, M., et al. (2020). A panel of human neutralizing mAbs targeting SARS-CoV-2 spike at multiple epitopes. Nat Commun *11*, 4303.

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). Euro Surveill *25*, 2000421.

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 *583*, 290–295.

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 *584*, 437–442.

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 *369*, 956–963.

Ruiz Ortega et al., M. (2021). Private communication.

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 *53*, 442-455.e4.

Sethna, Z., Isacchini, G., Dupic, T., Mora, T., Walczak, A.M., and Elhanati, Y. (2020). Population variability in the generation and selection of T-cell repertoires. PLoS Comput Biol *16*, e1008394.

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). Analysis of a SARS-CoV-2-infected individual reveals development of potent neutralizing antibodies with limited somatic mutation. Immunity *53*, 98-105.e5.

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). Characterization of neutralizing antibodies from a SARS-CoV-2 infected individual. BioRxiv 10.1101/2020.05.12.091298.

Shi, R., Shan, C., Duan, X., Chen, Z., Liu, P., Song, J., Song, T., Bi, X., Han, C., Wu, L., et al. (2020). A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. Nature *584*, 120–124.

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.H., 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., Haslwanter, D., Abdiche, Y.N., Shehata, L., Pedreño-Lopez, N., Moyer, C.L., Bornholdt, Z.A., Lilov, A., Nett, J.H., Jangra, R.K., et al. (2020a). Longitudinal dynamics of the human B cell response to the yellow fever 17D vaccine. Proc Natl Acad Sci U S A *117*, 6675–6685.

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. (2020b). 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 (2021). Coronavirus disease (COVID-19) pandemic.

Wrammert, J., Smith, K., Miller, J., Langley, W.A., Kokko, K., Larsen, C., Zheng, N.-Y., Mays, I., Garman, L., Helms, C., et al. (2008). Rapid cloning of high-affinity human monoclonal antibodies against influenza virus. Nature *453*, 667–671.

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.

Yuan, M., Wu, N.C., Zhu, X., Lee, C.-C.D., So, R.T.Y., Lv, H., Mok, C.K.P., and Wilson, I.A. (2020). A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science *368*, 630–633.

Zhou, D., Duyvesteyn, H.M.E., Chen, C.-P., Huang, C.-G., Chen, T.-H., Shih, S.-R., Lin, Y.-C., Cheng, C.-Y., Cheng, S.-H., Huang, Y.-C., et al. (2020). Structural basis for the neutralization of SARS-CoV-2 by an antibody from a convalescent patient. Nat Struct Mol Biol *27*, 950–958.

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. Nat Med *26*, 1422–1427.