A large-scale systematic survey of SARS-CoV-2 antibodies reveals recurring molecular features

Yiquan Wang[1,*], Meng Yuan[2,*], Jian Peng[3], Ian A. Wilson[2,4], Nicholas C. Wu[1,5,6,7,§]

[1] Department of Biochemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[2] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

[3] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[4] The Skaggs Institute for Chemical Biology, The Scripps Research Institute, La Jolla, CA 92037, USA

[5] Center for Biophysics and Quantitative Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[6] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

[7] Carle Illinois College of Medicine, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

* These authors contributed equally to this work.

§Correspondence: nicwu@illinois.edu (N.C.W.)

ABSTRACT

In the past two years, the global research in combating COVID-19 pandemic has led to isolation and characterization of numerous human antibodies to the SARS-CoV-2 spike. This enormous collection of antibodies provides an unprecedented opportunity to study the antibody response to a single antigen. From mining information derived from 88 research publications and 13 patents, we have assembled a dataset of ~8,000 human antibodies to the SARS-CoV-2 spike from >200 donors. Analysis of antibody targeting of different domains of the spike protein reveals a number of common (public) responses to SARS-CoV-2, exemplified via recurring IGHV/IGK(L)V pairs, CDR H3 sequences, IGHD usage, and somatic hypermutation. We further present a proof-of-concept for prediction of antigen specificity using deep learning to differentiate sequences of antibodies to SARS-CoV-2 spike and to influenza hemagglutinin. Overall, this study not only provides an informative resource for antibody and vaccine research, but fundamentally advances our molecular understanding of public antibody responses to a viral pathogen.

INTRODUCTION

From the beginning of COVID-19 pandemic, many research groups worldwide turned their attention to SARS-CoV-2 and, in particular, to the immune response to infection and vaccination. Over the past two years, thousands of human monoclonal antibodies to SARS-CoV-2 have been isolated and characterized [1, 2]. The major surface antigen to which antibodies are elicited is the SARS-CoV-2 spike (S) protein, which is a homotrimeric glycoprotein that facilitates virus entry by first engaging the host receptor ACE2 and then mediating membrane fusion [3, 4]. The S protein has three major domains, namely the N-terminal domain (NTD), receptor-binding domain (RBD), and S2 domain [5, 6]. Most studies on SARS-CoV-2 antibodies have focused on the immunodominant RBD [7], because neutralizing antibodies can be elicited to it with very high potency [8, 9]. Antibodies to the NTD and the highly conserved S2 domain have also been discovered, but usually exhibit lower neutralizing potency [10-16].

A common or public antibody response describes antibodies to the same antigen in different donors that share genetic elements that usually result in similar modes of antigen recognition. Deciphering public responses to particular antigens is not only critical for uncovering the molecular features of recurring antibodies within the diverse antibody repertoire at the population level, but also important for development of effective vaccines [17, 18]. A conventional approach to study public antibody responses is to identify public clonotypes, which are antibodies from different donors that share the same immunoglobulin heavy variable (IGHV) gene and with similar complementarity-determining region (CDR) H3 sequences [19-23]. While this definition of public clonotypes has improved our understanding of public antibody response, it generally ignores the contribution of the light chain. Moreover, our recent study has shown that a public antibody response to influenza hemagglutinin is driven by an IGHD gene with minimal dependence on the IGHV gene [24]. Therefore, the true extent and molecular characterization of public antibody responses remain to be explored.

Although information of many human clonal antibodies to SARS-CoV-2 is now publicly available, it has been difficult to leverage all available information to investigate public antibody responses to SARS-CoV-2. One major challenge is that the data from different studies are rarely in the same format. This inconsistency imposes a huge barrier to data mining. The establishment of the coronavirus antibody database (CoV-AbDab) has enabled researchers to deposit their antibody data in a standardized format and has partially resolved the data formatting issue [2]. However, not every SARS-CoV-2 antibody study has deposited their data to CoV-AbDab. Furthermore, IGHD gene identities, nucleotide sequences, and donor IDs are not available in CoV-AbDab, which makes it challenging to study public antibody responses using CoV-AbDab. Thus, additional efforts must be made to fully synergize the information across many different SARS-CoV-2 antibody studies to investigate and decipher public antibody responses.

In this study, we performed a systematic literature survey and assembled a large dataset of human SARS-CoV-2 monoclonal antibodies with donor information. We then analyzed this dataset and uncovered many previously unknown antibody sequence features that contribute to public antibody responses to SARS-CoV-2 S. For example, we identified a public antibody response to RBD that is largely independent of the IGHV gene, as well as involvement of a particular IGHD gene in a public antibody response to S2. Our analysis also revealed a number of recurring somatic hypermutations (SHMs) in different public clonotypes.

RESULTS

Collection of SARS-CoV-2 antibody information

Information for 8,048 human antibodies was collected from 88 research publications and 13 patents that described the discovery and characterization of antibodies to SARS-CoV-2 (**Figure S1, Data S1**). Among these antibodies, which were isolated from 215 different donors, 7,997

(99.4%) react with SARS-CoV-2, and the remaining 51 react with SARS-CoV or seasonal coronaviruses. While 99.1% (7,923/7,997) SARS-CoV-2 antibodies in our dataset bind to S protein, 49 bind to N and 25 to ORF8. Epitope information was available for most SARS-CoV-2 S antibodies, with 5,002 to RBD, 513 to NTD, and 890 to S2. In addition, information on neutralization activity, germline gene usage, sequence, structure, bait for isolation (e.g. RBD, S), and donor status (e.g. infected patient, vaccinee, etc.), if available, was collected for individual antibodies.

Epitope-dependent V gene usage bias in SARS-CoV-2 S antibodies

To identify the sequence features in RBD, NTD, and S2 antibodies, we first performed an analysis on V gene usage. Our analysis identified several commonly used IGHV/IGK(L)V pairs among RBD antibodies (**Figure 1A**), such as IGHV3-53/IGKV1-9 and IGHV3-53/IGKV3-20, which represent two known public clonotypes [25-30]. We also observed substantial enrichment of IGHV1-24 among NTD antibodies over the naïve baseline (**Figure 1B**), which was established by published datasets of antibody repertoire sequencing from 26 healthy donors [31-33]. IGHV1-24 is in fact a known public antibody response that targets an antigenic supersite on NTD [10-13]. These observations illustrate that the gene usage pattern in our dataset is consistent with previous findings. Importantly, our dataset also enabled us to discover previously unknown patterns in gene usage. For example, IGHV3-30 and IGHV3-30-3 were highly enriched among S2 antibodies over baseline (**Figure 1B**). For our subsequent analyses, IGHV3-30-3 was also labeled as IGHV3-30, since IGHV3-30 and IGHV3-30-3 have an identical amino acid sequence in the framework regions, CDR H1 and CDR H2. V gene usage bias was also observed in the light chain. For example, IGKV3-20 and IGKV3-11 were most used among S2 antibodies, whereas IGKV1-33 and IGKV1-39 were most used among RBD antibodies (**Figure 1C**). Overall, these results demonstrated that RBD, NTD, and S2 antibodies have distinct patterns of V gene usage.

CDR H3 analysis reveals public antibody response

Although heavy and light chain V genes together encode four of the six CDRs, most of the antibody sequence diversity comes from the CDR H3 region due to V(D)J recombination. Since CDR H3 is typically an important determinant for binding and may even dominate the paratope [24, 34-37], characterization of CDR H3 sequences in S antibodies is essential for understanding the antibody response to SARS-CoV-2. Here, we aimed to examine the convergence of CDR H3 sequences among S antibodies. Briefly, CDR H3 sequences with the same length were clustered by an 80% sequence identity cutoff. Only those clusters that contained antibodies from at least two different donors were subjected to further analysis. A total of 170 clusters were identified (**Figure 2A and Data S1**). Interestingly, antibodies within the same cluster often share the same binding region on the S protein (RBD, NTD, or S2), consistent with the notion that the CDR H3 sequence has a critical role in determining the epitope that is recognized.

The largest cluster (cluster 1) consisted of 139 antibodies from 57 donors (**Figure 2B**). Most of the antibodies in cluster 1 belonged to a well-characterized public clonotype to RBD that is encoded by IGHV3-53/3-66 and IGKV1-9 [25-27, 29, 30]. IGHV3-53/3-66, which is frequently used in RBD antibodies [28], was also enriched among antibodies in several other major CDR H3 clusters (e.g. clusters 2, 4, 8, and 14). Antibodies that bind to quaternary epitopes by bridging two RBDs on the same spike are found in clusters 14 and 17 [38] (**Figure S2**). Notably, both clusters 3 and 5, which target the RBD, contained a conserved disulfide bond (**Figure 2B**). Cluster 3 represents another well-characterized public clonotype that is encoded by IGHV1-58/IGKV3-20 [8, 9, 39, 40]. On the other hand, antibodies in cluster 5, which are largely encoded by IGHV3-30/IGKV1-33, have not been extensively studied. Most antibodies within cluster 5 had relatively weak neutralizing activity, if any, despite having reasonable binding affinity (**Table S1**). This result suggests the existence of an RBD-targeting public clonotype that had minimal neutralizing activity.

Similar observation was made with RBD antibodies encoded by IGHV3-13/IGKV1-39, although most of these antibodies did not share a similar CDR H3 (**Figure S3 and Table S2**).

Furthermore, we also discovered several S2-specific CDR H3 clusters (clusters 6, 9, and 11) that were predominantly encoded by IGHV3-30 with diverse IGK(L)V genes, suggesting a public heavy chain response to S2 (**Figure 2B**). Clusters 10 and 15 were also of interest to us. Cluster 10 was featured by a very short CDR H3 (6 amino acids, IMGT numbering) and was encoded by IGHV4-59/IGKV3-20, which was a frequent V gene pair among the S2 antibodies. Cluster 15 was encoded by IGHV1-69/IGKV3-11, which was the most used V gene pair among the S2 antibodies. Therefore, clusters 10 and 15 represented two major S2 public clonotypes, despite their minimal neutralizing activity (**Table S1**). In contrast to RBD- and S2-specific clusters, all NTD-specific CDR H3 clusters had a relatively small size (**Figure 2A**), suggesting that the paratopes for most NTD antibodies are not dominated by CDR H3.

A public antibody response dominated by the light chain and CDR H3

While most clusters have a dominant IGHV gene, diverse IGHV genes were observed in cluster 7 (**Figure 2B-C**). Most antibodies (42 out of 45) in cluster 7 used IGLV6-57, suggesting their paratopes are mainly composed of CDR H3 and light chain. S2A4, which is encoded by IGHV3-7/IGLV6-57 [41], is an antibody in cluster 7. A previously determined structure of S2A4 in complex with RBD indeed demonstrates that its CDR H3 contributes 38% of the buried surface area (BSA) of the epitope, whereas the light chain contributes 53% (**Figure 2D-E**). Specifically, IGLV6-57 forms an extensive H-bond network with the RBD (**Figure 2F**), whereas a ^{97}WLRG^{100} motif at the tip of CDR H3 interacts with the RBD through H-bonds, π-π stacking, and hydrophobic interactions (**Figure 2G**). Although G100 does not participate in binding, it exhibits backbone torsion angles (Φ = −94°, Ψ = −160°) that are in the preferred region of Ramachandran plot for glycine, but in the allowed region for non-glycine (**Figure S4**). Consistently, this ^{97}WLRG^{100} motif

is highly conserved in cluster 7 (**Figure 2B**). These results illustrate that our CDR H3 clustering analysis not only captured existing knowledge about public SARS-CoV-2 antibody responses, but was able to uncover recurring sequence features among SARS-CoV-2 antibodies that were previously unknown.

IGHV3-30/IGHD1-26 is a recurring feature in S2 antibodies

As a major contributor to CDR H3, the IGHD gene can also drive a public antibody response [24]. Consequently, we aimed to understand if there are any signature IGHD genes in SARS-CoV-2 S antibodies. While the frequency of most IGHD genes were within the baseline level, IGHD1-26 was highly enriched among S2 antibodies (**Figure 3A**). These IGHD1-26 S2 antibodies were predominantly encoded by IGHV3-30 (**Figure 3B**), which is one of the most used IGHV genes among S2 antibodies (**Figure 1B**). In contrast, the IGK(L)V gene usage was more diverse among these IGHD1-26 S2 antibodies, although several were more frequently used than others (**Figure 3C**), implying that this public antibody response to S2 is mainly driven by the heavy chain. Interestingly, 70% of these IGHD1-26 S2 antibodies had a CDR H3 of 14 amino acids, whereas only <20% of other S antibodies had a CDRH3 of 14 amino acids (**Figure 3D**). In fact, most members of clusters 6, 9, and 11 in our CDR H3 analysis above (**Figure 2B**) represented this public antibody response to S2. While CDR H3 is also encoded by the IGHJ gene, the distribution of IGHJ gene usage in these IGHD1-26 S2 antibodies did not show a strong deviation from that of other S antibodies in our dataset (**Figure 3E**).

In our dataset, there were 110 IGHD1-26 S2 antibodies from 17 donors with a CDR H3 length of 14 amino acids. Sequence logo analysis of these 110 antibodies revealed a conserved ^{97}[S/G]G[S/N]Y^{100} motif in the middle of their CDR H3 sequences (**Figure 3F**). In-depth analysis of the CDR H3 sequences from three representative IGHD1-26 S2 antibodies, namely P008_088, G32M4, and ADI-56059, further indicated that the conserved ^{97}[S/G]G[S/N]Y^{100} motif was within

the IGHD1-26-encoded region (**Figure 3G**). Of note, P008_088, G32M4, and ADI-56059 were isolated from three different donors by three independent research groups [42-44]. While P008_088 and G32M4 were from SARS-CoV-2 infected individuals, ADI-56059 was from a SARS-CoV survivor. Although 87 out of these 110 IGHD1-26 S2 antibodies can cross-react with SARS-CoV, they generally have minimal neutralization activity (**Table S3**). Together, these results show that IGHV3-30/IGHD1-26 represents a public antibody response to a highly conserved epitope in S2.

Recurring somatic hypermutations in public antibody responses

Our recent study has shown that V_H Y58F is a recurring somatic hypermutation (SHM) among IGHV3-53 antibodies to SARS-CoV-2 RBD [25]. Here, we aimed to identify additional recurring SHMs in other public clonotypes to SARS-CoV-2 S. In this analysis, antibodies from at least two donors that had the same IGHV/IGK(L)V genes and CDR H3s from the same CDR H3 cluster were classified as a public clonotype (**Figure 4A**). SHM that occurred in at least two donors within a public clonotypes was defined as a recurring SHM. Our analysis here only focused on major public clonotypes with antibodies from at least nine donors. This analysis led to the identification of several recurring SHMs in IGHV3-53/3-66-encoded public clonotypes that were previously characterized, including V_H F27V, T28I, and Y58F [25, 45, 46] (**Figure S5**). We also identified many other previously unknown recurring SHMs in both heavy and light chains (**Figure 4A-B**), including V_L S29R in a IGHV1-58/IGKV3-20 public clonotype that belongs to cluster 3 of our CDR H3 clustering analysis (**Figure 2A-B**). V_L S29R emerged in 8 out of 26 (31%) donors that carried this IGHV1-58/IGKV3-20 public clonotype.

Antibodies of this IGHV1-58/IGKV3-20 public clonotype bind to the ridge region of SARS-CoV-2 RBD (**Figure 5A**), and can be robustly elicited by infection with antigenically distinct variants of SARS-CoV-2 [39, 47] and by vaccination [48, 49]. These antibodies are also able to potently

neutralize multiple variants of concern (VOC) [9, 48, 50]. We compared two previously determined structures of IGHV1-58/IGKV3-20 antibodies in complex with RBD [40, 51], where one has the germline-encoded V_L S29 (**Figure 5B**) and the other carries a somatically mutated V_L R29 (**Figure 5C**). While neither V_L S29 nor V_L R29 directly interact with RBD, V_L R29 is able to form a cation-π interaction with V_L Y32, which in turn forms a T-shaped π-π stacking with RBD-F486 and H-bonds with RBD-C480 (**Figure 5C**). In the absence of SHM V_L S29R, the rotamer adopted by V_L Y32 does not permit these interactions to be formed. During our structural analysis, we discovered that V_L S29R forms a salt bridge with another SHM V_L G92D (**Figure 5C**), which can further stabilize the interactions between V_L Y32 and with RBD. In fact, it is likely that V_L S29R promoted the emergence of V_L G92D, since V_L G92D was found in four out of the 67 antibodies and all four that carried V_L S29R (**Figure 5D-E**). This analysis substantiates the notion that recurring SHM can be found among antibodies within a public clonotype and further suggests the existence of common affinity maturation pathways that involve emergence of multiple SHMs in a defined order.

Antigen identification by deep learning

Since many sequence features of public antibody responses to the S protein can be observed in our dataset, we postulated that the dataset is sufficiently large to train a deep learning model to identify S antibodies. To provide a proof-of-concept, we aimed to train a deep learning model to distinguish between antibodies to S and to influenza hemagglutinin (HA). Among different antigens, HA was chosen here because there are a large number of HA antibodies with published sequences, albeit still lower than the published SARS-CoV-2 S antibodies. Here, 4,736 unique SARS-CoV-2 S antibodies and 2,204 unique influenza HA antibodies with complete information for all six CDR sequences were used (**Data S2**). Sequences for HA antibodies were retrieved from GenBank [52]. None of these antibodies have identical sequences in all six CDRs. These antibodies to S and HA were divided into a training set (64%), a validation set (16%), and a test set (20%), with no overlap between the three sets. The training set was used to train the deep

learning model. The validation set was used to evaluate the model performance during training. The test set was used to evaluate the performance of the final model.

Our deep learning model has a simple architecture, which consisted of one encoder per CDR followed by three fully connected layers (**Figure 6A**). To evaluate the model performance on the test set, the area under the curves of receiver operating characteristic (ROC AUC) and precision-recall (PR AUC) were used to measure the model's ability to avoid misclassification. While ROC AUC is popular evaluation metric [53], PR AUC is shown to be more informative for evaluating models that are trained with imbalanced datasets [54]. Model performance was the best when all six CDRs (i.e. H1, H2, H3, L1, L2, and L3) were used to train the model, which resulted in an ROC AUC and an PR AUC of 0.87 and 0.92, respectively (**Figure 6B and Table S4**). Interesting, a similar performance was observed when the model was trained by the three heavy-chain CDRs (i.e. H1, H2, and H3) (ROC AUC = 0.86, PR AUC = 0.91), indicating that the heavy chain sequence captures most of the information to distinguish between HA antibodies and S antibodies. A reasonable performance was also observed when the model was trained by the three light-chain CDRs (i.e. L1, L2, and L3) (ROC AUC = 0.77, PR AUC = 0.86). For other types of inputs that we have tested, including CDR H3 only, CDR L3 only, CDR H3+L3, CDR H1+H2, and CDR L1+L2, the ROC AUCs were between 0.72 and 0.83 and the PR AUCs were between 0.82 and 0.90. These results imply that IGHV-encoded region (H1+H2), IGK(L)V-encoded region (L1+L2), and the V(D)J junctions (CDR H3 and CDR L3) are all informative for predicting antigen specificity. Overall, while our deep learning model had a relatively simple architecture, it was able to discriminate between antibodies to two different antigens based on primary sequences.

A recent study reported 81 antibodies to SARS-CoV-2 RBD that were elicited by Beta variant infection [47]. While these 81 antibodies were not included in the dataset that we assembled (**Data S1**), they provided an opportunity to further evaluate the performance of our deep learning model.

Our deep learning model that was trained by all six CDRs (see above) successfully predicted that 72 of the 81 (89%) antibodies as SARS-CoV-2 S antibodies (**Figure 6C and Table S5**). This result further demonstrates the possibility of predicting antibody specificity solely based on the primary sequence.

DISCUSSION

Through a systematic survey of published information on SARS-CoV-2 antibodies, we identified many molecular features of public antibody responses to SARS-CoV-2. The large amount of published information has allowed us to explore distinct patterns of germline gene usages in antibodies that target different domains on the S protein (i.e. RBD, NTD, and S2). Notably, the types and nature of public antibody responses to different domains appear to be quite different. For example, convergence of CDR H3 sequences can be readily identified in the public antibody responses to RBD and S2. In contrast, the public antibody response to NTD seems to be largely independent of the CDR H3 sequence. Furthermore, an IGHD-dependent public antibody response was enriched against S2, but not RBD or NTD. Together, our study demonstrates the diversity of sequence features that can constitute a public antibody response against a single antigen.

The public antibody response to SARS-CoV-2 has also been examined by a recent data mining study that focused on identifying public clonotypes [55]. This previous study defined public clonotypes as antibodies with the same IGHV/IGHJ/IGK(L)V/IGK(L)V genes and high similarity of CDR H3 [55]. While multiple public clonotypes were identified using this stringent definition [55], the characterization of public antibody response is likely far from comprehensive. A public antibody response may not always involve a defined pair of IGHV/IGK(L)V genes, especially when either IGHV or IGK(L)V gene-encoded residues only make a minimal contribution to the paratope. In fact, a well-characterized public antibody response to the highly conserved stem region of

296 influenza hemagglutinin has a paratope that is entirely attributed to the IGHV1-69 heavy chain
297 [56-59]. IGHV3-30/IGHD1-26 antibodies to S2 in our study may represent a similar type of
298 IGK(L)V-independent public antibody response, although it still needs to be confirmed by
299 structural analysis. On the other extreme, RBD antibodies that are encoded by IGLV6-57 with a
300 ^{97}WLRG^{100} motif in the CDR H3 represent a public response that is largely independent of IGHV
301 gene usage. Given the diverse types of public antibody responses to SARS-CoV-2 S, we need to
302 acknowledge the limitation of using the conventional strict definition of public clonotype to study
303 public antibody responses.
304
305 Public antibody response to different antigens can have very different sequence features. For
306 example, IGHV6-1 and IGHD3-9 are signatures of public antibody response to influenza virus [24,
307 60-62], whereas IGHV3-23 is frequently used in antibodies to Dengue and Zika viruses [63]. In
308 contrast, these germline genes are seldom used in the antibody response to SARS-CoV-2 as
309 compared to the naïve baseline (**Figure 1B-C and Figure 3A**). Since the binding specificity of an
310 antibody is determined by its structure, which in turn is determined by its amino acid sequence,
311 the antigen specificity of an antibody can theoretically be identified based on its sequence. This
312 study provides a proof-of-concept by training a deep learning model to distinguish between SARS-
313 CoV-2 S antibodies and influenza HA antibodies, solely based on primary sequence information.
314 Technological advancements, such as the development of single-cell high-throughput screen
315 using the Berkeley Lights Beacon optofluidics device [64] and advances in paired B-cell receptor
316 sequencing [65], have been accelerating the speed of antibody discovery and characterization.
317 As more sequence information on antibodies to different antigens is accumulated, we may be
318 able in the future to construct a generalized sequence-based model to accurately predict the
319 antigen specificity of any antibody.
320

In summary, the amount of publicly available information on SARS-CoV-2 antibodies has provided invaluable biological insights that have not been readily obtained for other pathogens. One reason is that the COVID-19 pandemic has gathered scientists from many fields and around the globe to work intensively on SARS-CoV-2. The parallel efforts by many different research groups have enabled SARS-CoV-2 antibodies to be discovered in unprecedented speed and scale that have not been possible for other pathogens. We anticipate that knowledge of the molecular features of the antibody response to SARS-CoV-2 will keep accumulating as more antibodies are isolated and characterized. Ultimately, the extensive characterization of antibodies to the SARS-CoV-2 S protein may allow us to address some of the most fundamental questions about antigenicity and immunogenicity, as well as how the human immune repertoire has evolved to respond to specific classes of viral pathogens that have coexisted with humans for hundreds to thousands of years.

MATERIALS AND METHODS

Collection of antibody information

Information on the monoclonal antibodies is derived from the original papers (Supplementary Table 1). Sequences of each monoclonal antibody are from the original papers and/or NCBI GenBank database (www.ncbi.nlm.nih.gov/genbank) [52]. Putative germline genes were identified by IgBLAST [66]. Some studies isolated antibodies from multiple donors, but the donor identity for each antibody was not always clear. For example, some studies mixed B cells from multiple donors before isolating individual B cell clones. Since the donor identity cannot be distinguished among those antibodies, we considered them from the same donor with "_mix" as the suffix of the donor ID. In addition, the PBMCs of SARS-CoV survivors in three separate studies were all from NIH/VRC [12, 44, 67]. Since it is unclear If they are the same SARS-CoV survivor, the same donor ID "VRC_SARS1" was assigned to them to avoid overestimation of public antibody response. the neutralization activity of a given antibody was only measured at a single concentration, 50% neutralization activity or below was classified as non-neutralizing. We also

downloaded the CoV-AbDab [2] in September 2021 to fill in any additional information. As of September 2021, there were 2,582 human SARS-CoV-2 antibodies in CoV-AbDab. Information in the finalized dataset was manually inspected by three different individuals. For antibodies that were shown to bind to S1 but not RBD, they were classified as NTD antibodies. Due to having identical nucleotide sequences, IGKV1D-39*01 was classified as IGKV1-39*01, IGHV1-68D*02 as IGHV1-68*02, IGHV1-69D*01 as IGHV1-69*19, IGHV3-23D*01 as IGHV3-23*01, and IGHV3-29*01 as IGHV3-30-42*01.

Analysis of germline gene usages

Non-functional germline genes were ignored in our germline gene usage analysis. Except for the analysis presented in **Figure 1**, IGHV3-30-3 was classified as IGHV3-30 since they have identical amino-acid sequence in the framework regions, CDR H1, and CDR H2. To establish the baseline germline usage frequency, published antibody repertoire sequencing datasets from 26 healthy donors [31, 32] were downloaded from cAb-Rep [33]. Putative germline genes for each antibody sequence in these repertoire sequencing datasets from healthy donors were identified by were identified by IgBLAST [66].

CDR H3 clustering analysis

Using a deterministic clustering approach, antibodies with CDR H3 sequences that had the same length and at least 80% amino-acid sequence identity were assigned to the same cluster. As a result, CDR H3 of every antibody in a cluster would have >20% difference in amino-acid sequence identity with that of every antibody in another cluster. A cluster would be discarded if all of its antibody members were from the same donor. The number of antibodies within a cluster was defined as the cluster size. Sequence logos were generated by Logomaker in Python [68]. For each cluster, epitope assignment was performed using the following scoring scheme. Briefly, there were three scoring categories, namely “RBD”, “NTD”, and “S2”.

- 1 point was added to category "RBD" for each antibody with an epitope label equals to "S:RBD" or "S:S1".
- 1 point was added to category "NTD" for each antibody with an epitope label equals to "S:NTD", "S:S1", "S:non-RBD", or "S:S1 non-RBD".
- 1 point was added to category "S2" for each antibody with an epitope label equals to "S:S2", " S:S2 Stem Helix", "S:non-RBD".

The category with >50% of the total points would be classified as the epitope for a given cluster. If no category had >50% of the total points, the epitope for the cluster would be classified as "unknown".

Identification of recurring somatic hypermutation (SHM)

In this study, a public clonotype was classified as antibodies from at least two donors that had the same IGHV/IGK(L)V genes and CDR H3s from the same CDR H3 cluster (see "CDR H3 clustering analysis" above). For each antibody, ANARCI was used to number the position of each residue according to Kabat numbering [69]. The amino-acid identity at each residue position of an antibody was then compared to that of the putative germline gene. CDR H3, CDR L3, and framework region 4 in both heavy and light chains were not included in this analysis. Insertions and deletions were also ignored in this analysis. SHM that occurred in at least two donors within a public clonotype was defined as a recurring SHM.

Deep learning model for antigen identification

Model construction

The deep learning model consisted of two networks, namely multi-encoder (ME) and a stack of multi-layered perceptrons (MLP). The CDR amino-acid sequences were taken as input and passed to ME. Specifically, each CDR amino-acid sequence was described by a 21-letter alphabet vector $\vec{\boldsymbol{x}} = (x_1, x_2, \dots, x_{L-1}, x_L), x \in \mathbb{R}^L$, where L represented the length of sequence, and

399 x represented the amino acid category. Each of the 20 canonical amino acids was one category,
400 whereas all the ambiguous amino acids were grouped as the 21st category. Before passing to
401 ME, inputs were tokenized at the amino-acid level and processed by zero padding, so that the
402 embedding layers represented the character-level tokens (i.e. amino acids) and the size of each
403 input was the same. Subsequently, the inputs were mapped to the embedding vectors with
404 additional dimension d. The sinusoidal positional encoding vectors were added to the embedding
405 vectors to encode the relative position of tokens (i.e. amino acids) in the sequence. Each
406 embedding vector, $\vec{\boldsymbol{x}} \in \mathbb{R}^{L\times d}$, with size of $L \times d$, was passed into transformer encoder layer by
407 self-attention mechanism to learn the sequence feature [70]. All learned sequence features were
408 then concatenated together and passed to multi-layered perceptron (MLP). Each MLP layer
409 contained leaky rectified linear unit (ReLU) activations to avoid the vanishing gradient. Dropout
410 layers were placed after each MLP block to avoid model overfitting [71]. The final output layer
411 was followed by a sigmoid activation function to predict the probability of different classes. The
412 prediction losses were calculated by binary cross-entropy loss.
413
414 Training detail
415 SARS-CoV-2 S antibodies and influenza HA antibodies with complete information for all six CDR
416 sequences were identified. Sequences of each antibody were from the original papers (**Data S2**)
417 or NCBI GenBank database (www.ncbi.nlm.nih.gov/genbank) [52]. If all six CDR sequences were
418 the same between two or more antibodies, only one of these antibodies would be retained. After
419 filtering duplicates, there were 4,736 antibodies to SARS-CoV-2 and 2,204 to influenza HA. The
420 CDR sequences were identified by IgBLAST and PyIR [66, 72]. This dataset was randomly split
421 into a training set (64%), a validation set (16%), and a test set (20%). The training set was used
422 to train the deep learning model. The validation set was used to evaluate the model performance
423 during training. The test set was used to evaluate the performance of the final model. There was
424 no overlap of antibody sequences among the training set, validation set, and test set. The Adam

algorithm was used to optimize the model. The following hyper-parameters were used for model training:

- CDR embedding size: 256
- The number of attention heads for self-attention on CDR feature learning: 4
- The number of encoder layer for CDR encoder: 4
- Size of stacking MLP layers: 512, 128, and 64
- Learning rate: 0.0001
- Batch size: 256

Using the same training set, validation set and test set, the model performance of using the following inputs was compared:

1. CDR H1 + H2
2. CDR L1 + L2
3. CDR H3
4. CDR L3
5. CDR H3 + L3
6. CDR H1 + H2 + H3
7. CDR L1 + L2 + L3
8. CDR H1 + H2 + H3 + L1 + L2 + L3

Performance Metrics

For evaluating model performance, S antibodies and HA antibodies were considered “positive” and “negative”, respectively. False positives (FP) and false negatives (FN) were samples that were misclassified by the model while true negatives (TN) and true positives (TP) were correctly classified one. The following metrics were computed to evaluate model performance:

$$accuarcy = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$precision = \frac{TP}{TP + FP} \quad (2)$$

$$recall = \frac{TP}{TP + FN} \quad (3)$$

In addition, we also used the receiver operating characteristic (ROC) curve and precision-recall (PR) curve to measure the model's ability to avoid misclassification [53, 54]. Area under the curves of ROC (i.e. ROC AUC) and PR (i.e. PR AUC) were computed using the "keras.metrics" module in TensorFlow [73].

DATA AVAILABILITY

The assembled SARS-CoV-2 antibody dataset is in **Data S1**. The dataset for constructing and testing the deep learning model is in **Data S2**.

CODE AVAILABILITY

Custom python scripts for all analyses have been deposited to https://github.com/nicwulab/SARS-CoV-2_Abs.

ACKNOWLEDGEMENT

This work was supported by National Institutes of Health (NIH) R00 AI139445 (N.C.W.), DP2 AT011966 (N.C.W.), and Bill and Melinda Gates Foundation INV-004923 (I.A.W.). We thank Seth Zost and Huibin Lv for helpful discussion.

AUTHOR CONTRIBUTIONS

All authors conceived and designed the study. Y.W, M.Y. and N.C.W. assembled the dataset and performed data analysis. J.P. provided technical expertise in deep learning, Y.W., M.Y. I.A.W, and N.C.W. wrote the paper and all authors reviewed and/or edited the paper.

REFERENCES

1. Li D, Sempowski GD, Saunders KO, Acharya P, Haynes BF. SARS-CoV-2 neutralizing antibodies for COVID-19 prevention and treatment. Annu Rev Med. 2021. Epub 2021/08/25. doi: 10.1146/annurev-med-042420-113838. PubMed PMID: 34428080.

2. Raybould MIJ, Kovaltsuk A, Marks C, Deane CM. CoV-AbDab: the coronavirus antibody database. Bioinformatics. 2021;37(5):734-5. Epub 2020/08/18. doi: 10.1093/bioinformatics/btaa739. PubMed PMID: 32805021; PubMed Central PMCID: PMCPMC7558925.

3. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270-3. Epub 2020/02/06. doi: 10.1038/s41586-020-2012-7. PubMed PMID: 32015507.

4. Shang J, Wan Y, Luo C, Ye G, Geng Q, Auerbach A, et al. Cell entry mechanisms of SARS-CoV-2. Proc Natl Acad Sci U S A. 2020;117(21):11727-34. Epub 2020/05/08. doi: 10.1073/pnas.2003138117. PubMed PMID: 32376634; PubMed Central PMCID: PMCPMC7260975.

5. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell. 2020;181(2):281-92.e6. Epub 2020/03/11. doi: 10.1016/j.cell.2020.02.058. PubMed PMID: 32155444.

6. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020;367(6483):1260-3. Epub 2020/02/23. doi: 10.1126/science.abb2507. PubMed PMID: 32075877.

7. Yuan M, Liu H, Wu NC, Wilson IA. Recognition of the SARS-CoV-2 receptor binding domain by neutralizing antibodies. Biochem Biophys Res Commun. 2021;538:192-203. Epub 2020/10/19. doi: 10.1016/j.bbrc.2020.10.012. PubMed PMID: 33069360; PubMed Central PMCID: PMCPMC7547570.

8. Tortorici MA, Beltramello M, Lempp FA, Pinto D, Dang HV, Rosen LE, et al. Ultrapotent human antibodies protect against SARS-CoV-2 challenge via multiple mechanisms. Science. 2020;370(6519):950-7. Epub 2020/09/26. doi: 10.1126/science.abe3354. PubMed PMID: 32972994; PubMed Central PMCID: PMCPMC7857395.

9. Wang L, Zhou T, Zhang Y, Yang ES, Schramm CA, Shi W, et al. Ultrapotent antibodies against diverse and highly transmissible SARS-CoV-2 variants. Science. 2021;373(6556):eabh1766. Epub 2021/07/03. doi: 10.1126/science.abh1766. PubMed PMID: 34210892.

10. Voss WN, Hou YJ, Johnson NV, Delidakis G, Kim JE, Javanmardi K, et al. Prevalent, protective, and convergent IgG recognition of SARS-CoV-2 non-RBD spike epitopes. Science. 2021;372(6546):1108-12. Epub 2021/05/06. doi: 10.1126/science.abg5268. PubMed PMID: 33947773; PubMed Central PMCID: PMCPMC8224265.

11. Cerutti G, Guo Y, Zhou T, Gorman J, Lee M, Rapp M, et al. Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite. Cell Host Microbe. 2021;29(5):819-33.e7. Epub 2021/04/01. doi: 10.1016/j.chom.2021.03.005. PubMed PMID: 33789084; PubMed Central PMCID: PMCPMC7953435.

12. Li D, Edwards RJ, Manne K, Martinez DR, Schafer A, Alam SM, et al. In vitro and in vivo functions of SARS-CoV-2 infection-enhancing and neutralizing antibodies. Cell. 2021;184(16):4203-19.e32. Epub 2021/07/10. doi: 10.1016/j.cell.2021.06.021. PubMed PMID: 34242577; PubMed Central PMCID: PMCPMC8232969.

13. Chi X, Yan R, Zhang J, Zhang G, Zhang Y, Hao M, et al. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. Science. 2020;369(6504):650-5. Epub 2020/06/24. doi: 10.1126/science.abc6952. PubMed PMID: 32571838; PubMed Central PMCID: PMCPMC7319273.

14. Zhou P, Yuan M, Song G, Beutler N, Shaabani N, Huang D, et al. A protective broadly cross-reactive human antibody defines a conserved site of vulnerability on beta-coronavirus spikes. bioRxiv. 2021. Epub 2021/04/07. doi: 10.1101/2021.03.30.437769. PubMed PMID: 33821273; PubMed Central PMCID: PMCPMC8020973.

15. Pinto D, Sauer MM, Czudnochowski N, Low JS, Tortorici MA, Housley MP, et al. Broad betacoronavirus neutralization by a stem helix-specific human antibody. Science. 2021;373(6559):1109-16. Epub 2021/08/05. doi: 10.1126/science.abj3321. PubMed PMID: 34344823.

16. Li W, Chen Y, Prevost J, Ullah I, Lu M, Gong SY, et al. Structural basis and mode of action for two broadly neutralizing antibodies against SARS-CoV-2 emerging variants of concern. bioRxiv. 2021. Epub 2021/08/11. doi: 10.1101/2021.08.02.454546. PubMed PMID: 34373853; PubMed Central PMCID: PMCPMC8351775.

17. Lanzavecchia A, Fruhwirth A, Perez L, Corti D. Antibody-guided vaccine design: identification of protective epitopes. Curr Opin Immunol. 2016;41:62-7. Epub 2016/06/28. doi: 10.1016/j.coi.2016.06.001. PubMed PMID: 27343848.

18. Andrews SF, McDermott AB. Shaping a universally broad antibody response to influenza amidst a variable immunoglobulin landscape. Curr Opin Immunol. 2018;53:96-101. doi: 10.1016/j.coi.2018.04.009. PubMed PMID: 29730560.

19. Setliff I, McDonnell WJ, Raju N, Bombardi RG, Murji AA, Scheepers C, et al. Multi-donor longitudinal antibody repertoire sequencing reveals the existence of public antibody clonotypes in HIV-1 infection. Cell Host Microbe. 2018;23(6):845-54.e6. Epub 2018/06/05. doi: 10.1016/j.chom.2018.05.001. PubMed PMID: 29861170; PubMed Central PMCID: PMCPMC6002606.

20. Jackson KJ, Liu Y, Roskin KM, Glanville J, Hoh RA, Seo K, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. Cell Host Microbe. 2014;16(1):105-14. Epub 2014/07/02. doi: 10.1016/j.chom.2014.05.013. PubMed PMID: 24981332; PubMed Central PMCID: PMCPMC4158033.

21. Truck J, Ramasamy MN, Galson JD, Rance R, Parkhill J, Lunter G, et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. J Immunol. 2015;194(1):252-61. Epub 2014/11/14. doi: 10.4049/jimmunol.1401405. PubMed PMID: 25392534; PubMed Central PMCID: PMCPMC4272858.

22. Henry Dunand CJ, Wilson PC. Restricted, canonical, stereotyped and convergent immunoglobulin responses. Philos Trans R Soc Lond B Biol Sci. 2015;370(1676):20140238. Epub 2015/07/22. doi: 10.1098/rstb.2014.0238. PubMed PMID: 26194752; PubMed Central PMCID: PMCPMC4528415.

23. Pieper K, Tan J, Piccoli L, Foglierini M, Barbieri S, Chen Y, et al. Public antibodies to malaria antigens generated by two LAIR1 insertion modalities. Nature. 2017;548(7669):597-601. Epub 2017/08/29. doi: 10.1038/nature23670. PubMed PMID: 28847005; PubMed Central PMCID: PMCPMC5635981.

24. Wu NC, Yamayoshi S, Ito M, Uraki R, Kawaoka Y, Wilson IA. Recurring and adaptable binding motifs in broadly neutralizing antibodies to influenza virus are encoded on the D3-9 segment of the Ig gene. Cell Host Microbe. 2018;24(4):569-78.e4. doi: 10.1016/j.chom.2018.09.010. PubMed PMID: 30308159.

25. Tan TJC, Yuan M, Kuzelka K, Padron GC, Beal JR, Chen X, et al. Sequence signatures of two public antibody clonotypes that bind SARS-CoV-2 receptor binding domain. Nat Commun. 2021;12(1):3815. Epub 2021/06/23. doi: 10.1038/s41467-021-24123-7. PubMed PMID: 34155209.

26. Cao Y, Su B, Guo X, Sun W, Deng Y, Bao L, et al. Potent neutralizing antibodies against SARS-CoV-2 identified by high-throughput single-cell sequencing of convalescent patients' B cells. Cell. 2020;182(1):73-84.e16. Epub 2020/05/20. doi: 10.1016/j.cell.2020.05.025. PubMed PMID: 32425270; PubMed Central PMCID: PMCPMC7231725.

27. Kim SI, Noh J, Kim S, Choi Y, Yoo DK, Lee Y, et al. Stereotypic neutralizing VH antibodies against SARS-CoV-2 spike protein receptor binding domain in patients with COVID-19 and healthy individuals. Sci Transl Med. 2021;13(578):eabd6990. Epub 2021/01/06. doi: 10.1126/scitranslmed.abd6990. PubMed PMID: 33397677; PubMed Central PMCID: PMCPMC7875332.

28. Yuan M, Liu H, Wu NC, Lee CD, Zhu X, Zhao F, et al. Structural basis of a shared antibody response to SARS-CoV-2. Science. 2020;369(6507):1119-23. Epub 2020/07/15. doi: 10.1126/science.abd2321. PubMed PMID: 32661058; PubMed Central PMCID: PMCPMC7402627.

29. Clark SA, Clark LE, Pan J, Coscia A, McKay LGA, Shankar S, et al. SARS-CoV-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms. Cell. 2021;184(10):2605-17.e18. Epub 2021/04/09. doi:

10.1016/j.cell.2021.03.027. PubMed PMID: 33831372; PubMed Central PMCID: PMCPMC7962548.

30. Zhang Q, Ju B, Ge J, Chan JF, Cheng L, Wang R, et al. Potent and protective IGHV3-53/3-66 public antibodies and their shared escape mutant on the spike of SARS-CoV-2. Nat Commun. 2021;12(1):4210. Epub 2021/07/11. doi: 10.1038/s41467-021-24514-w. PubMed PMID: 34244522; PubMed Central PMCID: PMCPMC8270942.

31. Soto C, Bombardi RG, Branchizio A, Kose N, Matta P, Sevy AM, et al. High frequency of shared clonotypes in human B cell receptor repertoires. Nature. 2019;566(7744):398-402. Epub 2019/02/15. doi: 10.1038/s41586-019-0934-8. PubMed PMID: 30760926; PubMed Central PMCID: PMCPMC6949180.

32. Briney B, Inderbitzin A, Joyce C, Burton DR. Commonality despite exceptional diversity in the baseline human antibody repertoire. Nature. 2019;566(7744):393-7. Epub 2019/01/22. doi: 10.1038/s41586-019-0879-y. PubMed PMID: 30664748; PubMed Central PMCID: PMCPMC6411386.

33. Guo Y, Chen K, Kwong PD, Shapiro L, Sheng Z. cAb-Rep: a database of curated antibody repertoires for exploring antibody diversity and predicting antibody prevalence. Front Immunol. 2019;10:2365. Epub 2019/10/28. doi: 10.3389/fimmu.2019.02365. PubMed PMID: 31649674; PubMed Central PMCID: PMCPMC6794461.

34. Liu H, Wu NC, Yuan M, Bangaru S, Torres JL, Caniels TG, et al. Cross-neutralization of a SARS-CoV-2 antibody to a functionally conserved site Is mediated by avidity. Immunity. 2020;53(6):1272-80.e5. Epub 2020/11/27. doi: 10.1016/j.immuni.2020.10.023. PubMed PMID: 33242394; PubMed Central PMCID: PMCPMC7687367 COVA1-16 and other antibodies first disclosed by Brouwer et al. (2020) has been filed by Amsterdam UMC under application number 2020-039EP-PR. I.A.W. is a member of the Immunity Editorial Board.

35. Ekiert DC, Kashyap AK, Steel J, Rubrum A, Bhabha G, Khayat R, et al. Cross-neutralization of influenza A viruses mediated by a single antibody loop. Nature. 2012;489(7417):526-32. doi: 10.1038/nature11414. PubMed PMID: 22982990; PubMed Central PMCID: PMCPMC3538848.

36. Jette CA, Cohen AA, Gnanapragasam PNP, Muecksch F, Lee YE, Huey-Tubman KE, et al. Broad cross-reactivity across sarbecoviruses exhibited by a subset of COVID-19 donor-derived neutralizing antibodies. Cell Rep. 2021;36(13):109760. Epub 2021/09/18. doi: 10.1016/j.celrep.2021.109760. PubMed PMID: 34534459; PubMed Central PMCID: PMCPMC8423902.

37. Pancera M, Changela A, Kwong PD. How HIV-1 entry mechanism and broadly neutralizing antibodies guide structure-based vaccine design. Curr Opin HIV AIDS. 2017;12(3):229-40. Epub 2017/04/20. doi: 10.1097/COH.0000000000000360. PubMed PMID: 28422787; PubMed Central PMCID: PMCPMC5557343.

38. Barnes CO, Jette CA, Abernathy ME, Dam KA, Esswein SR, Gristick HB, et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. Nature. 2020;588(7839):682-7. Epub 2020/10/13. doi: 10.1038/s41586-020-2852-1. PubMed PMID: 33045718.

39. Robbiani DF, Gaebler C, Muecksch F, Lorenzi JCC, Wang Z, Cho A, et al. Convergent antibody responses to SARS-CoV-2 in convalescent individuals. Nature. 2020;584:437-42. doi: 10.1038/s41586-020-2456-9.

40. Dejnirattisai W, Zhou D, Ginn HM, Duyvesteyn HME, Supasa P, Case JB, et al. The antigenic anatomy of SARS-CoV-2 receptor binding domain. Cell. 2021;184(8):2183-200.e22. Epub 2021/03/24. doi: 10.1016/j.cell.2021.02.032. PubMed PMID: 33756110; PubMed Central PMCID: PMCPMC7891125.

41. Piccoli L, Park YJ, Tortorici MA, Czudnochowski N, Walls AC, Beltramello M, et al. Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. Cell. 2020;183(4):1024-42.e21. Epub 2020/09/30. doi: 10.1016/j.cell.2020.09.037. PubMed PMID: 32991844; PubMed Central PMCID: PMCPMC7494283.

42. Graham C, Seow J, Huettner I, Khan H, Kouphou N, Acors S, et al. Neutralization potency of monoclonal antibodies recognizing dominant and subdominant epitopes on SARS-CoV-2 Spike is impacted by the B.1.1.7 variant. Immunity. 2021;54(6):1276-89.e6. Epub 2021/04/10. doi: 10.1016/j.immuni.2021.03.023. PubMed PMID: 33836142; PubMed Central PMCID: PMCPMC8015430.

43. Tong P, Gautam A, Windsor IW, Travers M, Chen Y, Garcia N, et al. Memory B cell repertoire for recognition of evolving SARS-CoV-2 spike. Cell. 2021;184(19):4969-80.e15. Epub 2021/08/02. doi: 10.1016/j.cell.2021.07.025. PubMed PMID: 34332650; PubMed Central PMCID: PMCPMC8299219.

44. Wec AZ, Wrapp D, Herbert AS, Maurer DP, Haslwanter D, Sakharkar M, et al. Broad neutralization of SARS-related viruses by human monoclonal antibodies. Science. 2020;369(6504):731-6. Epub 2020/06/17. doi: 10.1126/science.abc7424. PubMed PMID: 32540900; PubMed Central PMCID: PMCPMC7299279.

45. Scheid JF, Barnes CO, Eraslan B, Hudak A, Keeffe JR, Cosimi LA, et al. B cell genomics behind cross-neutralization of SARS-CoV-2 variants and SARS-CoV. Cell. 2021;184(12):3205-21.e24. Epub 2021/05/21. doi: 10.1016/j.cell.2021.04.032. PubMed PMID: 34015271; PubMed Central PMCID: PMCPMC8064835.

46. Hurlburt NK, Seydoux E, Wan YH, Edara VV, Stuart AB, Feng J, et al. Structural basis for potent neutralization of SARS-CoV-2 and role of antibody affinity maturation. Nat Commun. 2020;11(1):5413. Epub 2020/10/29. doi: 10.1038/s41467-020-19231-9. PubMed PMID: 33110068; PubMed Central PMCID: PMCPMC7591918.

47. Reincke SM, Yuan M, Kornau H-C, Corman VM, van Hoof S, Sánchez-Sendin E, et al. SARS-CoV-2 Beta variant infection elicits potent lineage-specific and cross-reactive antibodies. bioRxiv. 2021. doi: 10.1101/2021.09.30.462420.

48. Schmitz AJ, Turner JS, Liu Z, Zhou JQ, Aziati ID, Chen RE, et al. A vaccine-induced public antibody protects against SARS-CoV-2 and emerging variants. Immunity. 2021;54(9):2159-66.e6. Epub 2021/09/01. doi: 10.1016/j.immuni.2021.08.013. PubMed PMID: 34464596; PubMed Central PMCID: PMCPMC8367776.

49. Andreano E, Paciello I, Piccini G, Manganaro N, Pileri P, Hyseni I, et al. Hybrid immunity improves B cells and antibodies against SARS-CoV-2 variants. Nature. 2021. Epub 2021/10/21. doi: 10.1038/s41586-021-04117-7. PubMed PMID: 34670266.

50. Li T, Han X, Gu C, Guo H, Zhang H, Wang Y, et al. Potent SARS-CoV-2 neutralizing antibodies with protective efficacy against newly emerged mutational variants. Nat Commun. 2021;12(1):6304. Epub 2021/11/04. doi: 10.1038/s41467-021-26539-7. PubMed PMID: 34728625.

51. Wheatley AK, Pymm P, Esterbauer R, Dietrich MH, Lee WS, Drew D, et al. Landscape of human antibody recognition of the SARS-CoV-2 receptor binding domain. Cell Rep. 2021;37(2):109822. Epub 2021/10/06. doi: 10.1016/j.celrep.2021.109822. PubMed PMID: 34610292; PubMed Central PMCID: PMCPMC8463300.

52. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013;41(Database issue):D36-42. Epub 2012/11/30. doi: 10.1093/nar/gks1195. PubMed PMID: 23193287; PubMed Central PMCID: PMCPMC3531190.

53. Flach P, Hernández-Orallo J, Ferri C. A coherent interpretation of AUC as a measure of aggregated classification performance. Proceedings of the 28th International Conference on International Conference on Machine Learning; Bellevue, WA, USA2011. p. 657-64.

54. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432. Epub 2015/03/05. doi: 10.1371/journal.pone.0118432. PubMed PMID: 25738806; PubMed Central PMCID: PMCPMC4349800.

55. Chen EC, Gilchuk P, Zost SJ, Suryadevara N, Winkler ES, Cabel CR, et al. Convergent antibody responses to the SARS-CoV-2 spike protein in convalescent and vaccinated individuals. Cell Rep. 2021;36(8):109604. Epub 2021/08/20. doi: 10.1016/j.celrep.2021.109604. PubMed PMID: 34411541; PubMed Central PMCID: PMCPMC8352653.

56. Lang S, Xie J, Zhu X, Wu NC, Lerner RA, Wilson IA. Antibody 27F3 broadly targets influenza A group 1 and 2 hemagglutinins through a further variation in VH1-69 antibody orientation on the HA stem. Cell Rep. 2017;20(12):2935-43. doi: 10.1016/j.celrep.2017.08.084. PubMed PMID: 28930686.

57. Dreyfus C, Laursen NS, Kwaks T, Zuijdgeest D, Khayat R, Ekiert DC, et al. Highly conserved protective epitopes on influenza B viruses. Science. 2012;337(6100):1343-8. doi: 10.1126/science.1222908. PubMed PMID: 22878502; PubMed Central PMCID: PMCPMC3538841.

58. Sui J, Hwang WC, Perez S, Wei G, Aird D, Chen LM, et al. Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses. Nat Struct Mol Biol. 2009;16(3):265-73. doi: 10.1038/nsmb.1566. PubMed PMID: 19234466; PubMed Central PMCID: PMCPMC2692245.

59. Ekiert DC, Bhabha G, Elsliger MA, Friesen RH, Jongeneelen M, Throsby M, et al. Antibody recognition of a highly conserved influenza virus epitope. Science. 2009;324(5924):246-51. doi: 10.1126/science.1171491. PubMed PMID: 19251591; PubMed Central PMCID: PMCPMC2758658.

60. Wu NC, Andrews SF, Raab JE, O'Connell S, Schramm CA, Ding X, et al. Convergent evolution in breadth of two VH6-1-encoded influenza antibody clonotypes from a single donor. Cell Host Microbe. 2020;28:434-44. Epub 2020/07/04. doi: 10.1016/j.chom.2020.06.003. PubMed PMID: 32619441.

61. Joyce MG, Wheatley AK, Thomas PV, Chuang GY, Soto C, Bailer RT, et al. Vaccine-induced antibodies that neutralize group 1 and group 2 influenza A viruses. Cell. 2016;166(3):609-23. doi: 10.1016/j.cell.2016.06.043. PubMed PMID: 27453470; PubMed Central PMCID: PMCPMC4978566.

62. Kallewaard NL, Corti D, Collins PJ, Neu U, McAuliffe JM, Benjamin E, et al. Structure and function analysis of an antibody recognizing all influenza A subtypes. Cell. 2016;166(3):596-608. doi: 10.1016/j.cell.2016.05.073. PubMed PMID: 27453466; PubMed Central PMCID: PMCPMC4967455.

63. Robbiani DF, Bozzacco L, Keeffe JR, Khouri R, Olsen PC, Gazumyan A, et al. Recurrent potent human neutralizing antibodies to Zika virus in Brazil and Mexico. Cell. 2017;169(4):597-609.e11. Epub 2017/05/06. doi: 10.1016/j.cell.2017.04.024. PubMed PMID: 28475892; PubMed Central PMCID: PMCPMC5492969.

64. Winters A, McFadden K, Bergen J, Landas J, Berry KA, Gonzalez A, et al. Rapid single B cell antibody discovery using nanopens and structured light. mAbs. 2019;11(6):1025-35. Epub 2019/06/13. doi: 10.1080/19420862.2019.1624126. PubMed PMID: 31185801; PubMed Central PMCID: PMCPMC6748590.

65. Curtis NC, Lee J. Beyond bulk single-chain sequencing: Getting at the whole receptor. Curr Opin Syst Biol. 2020;24:93-9. Epub 2020/10/27. doi: 10.1016/j.coisb.2020.10.008. PubMed PMID: 33102951; PubMed Central PMCID: PMCPMC7568503.

66. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013;41(Web Server issue):W34-40. doi: 10.1093/nar/gkt382. PubMed PMID: 23671333; PubMed Central PMCID: PMCPMC3692102.

67. Shiakolas AR, Kramer KJ, Wrapp D, Richardson SI, Schafer A, Wall S, et al. Cross-reactive coronavirus antibodies with diverse epitope specificities and Fc effector functions. Cell Rep Med. 2021;2(6):100313. Epub 2021/06/01. doi: 10.1016/j.xcrm.2021.100313. PubMed PMID: 34056628; PubMed Central PMCID: PMCPMC8139315.

68. Tareen A, Kinney JB. Logomaker: beautiful sequence logos in Python. Bioinformatics. 2020;36(7):2272-4. Epub 2019/12/11. doi: 10.1093/bioinformatics/btz921. PubMed PMID: 31821414; PubMed Central PMCID: PMCPMC7141850.

69. Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. Bioinformatics. 2016;32(2):298-300. Epub 2015/10/02. doi:

10.1093/bioinformatics/btv552. PubMed PMID: 26424857; PubMed Central PMCID: PMCPMC4708101.

70. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017); Long Beach, CA, USA2017.

71. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929-58.

72. Soto C, Finn JA, Willis JR, Day SB, Sinkovits RS, Jones T, et al. PyIR: a scalable wrapper for processing billions of immunoglobulin and T cell receptor sequences using IgBLAST. BMC Bioinformatics. 2020;21(1):314. Epub 2020/07/18. doi: 10.1186/s12859-020-03649-5. PubMed PMID: 32677886; PubMed Central PMCID: PMCPMC7364545.

73. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al., editors. TensorFlow: A system for large-scale machine learning. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation; 2016; Savannah, GA, USA.

74. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol. 2007;372(3):774-97. Epub 2007/08/08. doi: 10.1016/j.jmb.2007.05.022. PubMed PMID: 17681537.

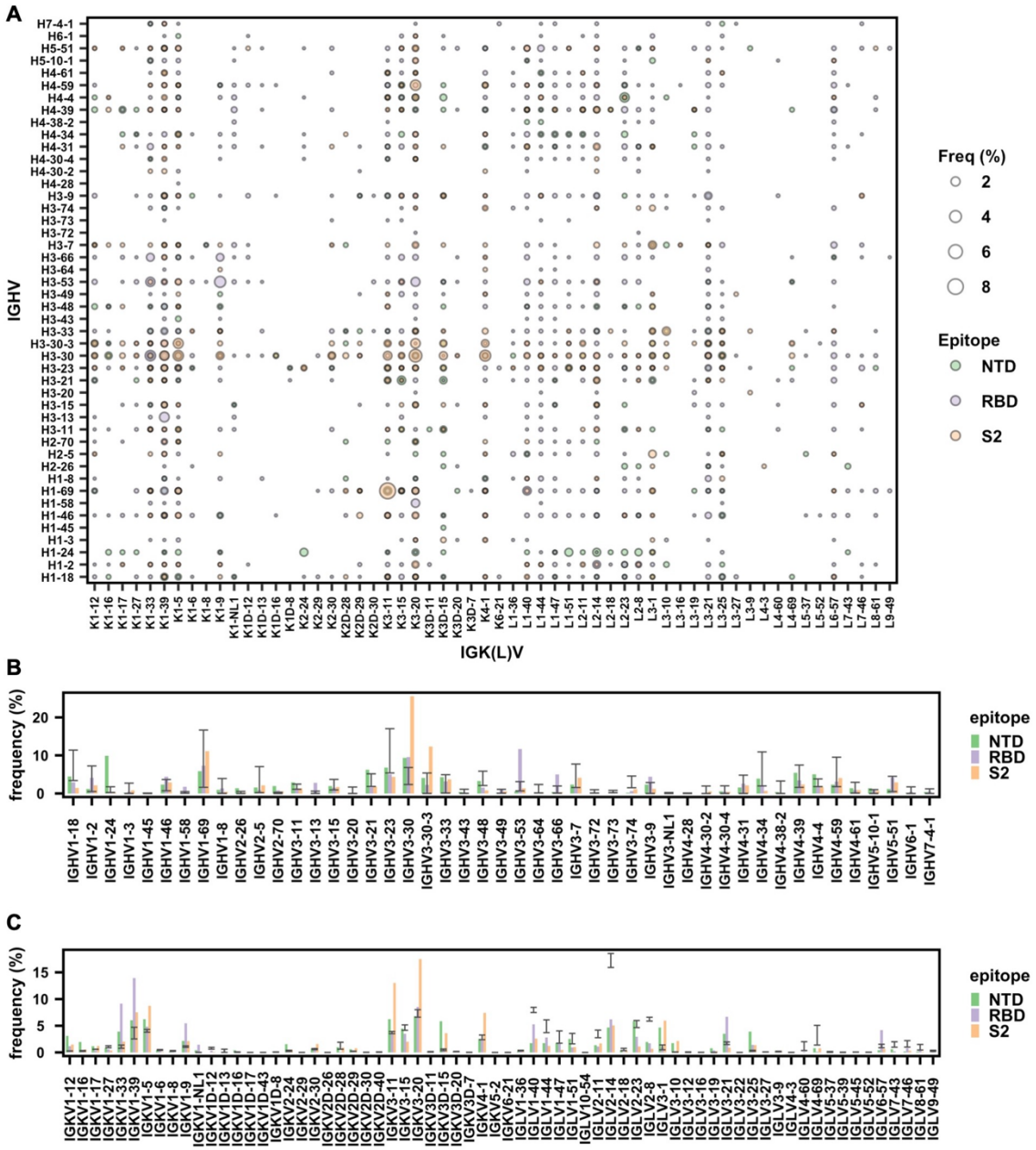


Figure 1. Analysis of V gene usage in SARS-CoV-2 S antibodies. (A) The frequency of different V gene pairings between heavy and light chains are shown for SARS-CoV-2 S antibodies to RBD, NTD, and S2. The size of each datapoint represents the frequency of the corresponding IGHV/IGK(L)V pair within its epitope category. Only those antibodies where both IGHV and IGK(L)V information is available for both heavy and light chains was included in this analysis. **(B)** The IGHV gene usage in antibodies to NTD, RBD, and S2 are shown. Only those antibodies with

IGHV information available were included in this analysis. **(C)** The IGK(L)V gene usage in antibodies to NTD, RBD, and S2 are shown. Only those antibodies with IGK(L)V information available were included in this analysis. **(B-C)** Error bars represent the frequency range among 26 healthy donors [31-33].

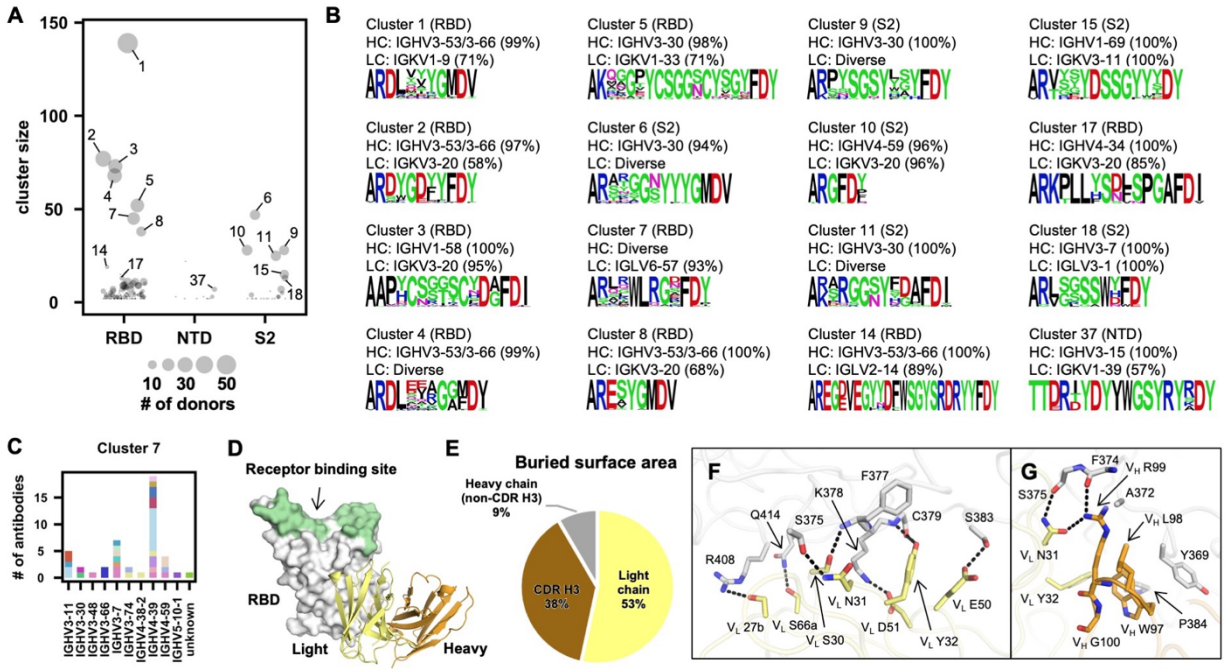


Figure 2. Convergent CDR H3 sequences among SARS-CoV-2 S antibodies. (A) CDR H3 sequences from individual antibodies were clustered using a 20% cutoff (see Materials and Methods). The epitope of each CDR H3 cluster is classified based on that of its antibody members. Cluster size represents the number of antibodies within the cluster. **(B)** The V gene usage and CDR H3 sequence are shown for each of the 16 CDR H3 clusters of interest. For each of the CDR H3 cluster of interest, the CDR H3 sequences are shown as a sequence logo, where the height of each letter represents the frequency of the corresponding amino-acid variant (single-letter amino-acid code) at the indicated position. The dominant germline V genes (>50% usage among all antibodies within a given CDR H3 cluster) are listed. Diverse: no germline V genes had >50% frequency among all antibodies within a given CDR H3 cluster. HC: heavy chain. LC: light chain. **(C)** IGHV usage in cluster 7 is shown. Different colors represent different donors. Unknown: IGHV information is not available. **(D)** An overall view of SARS-CoV-2 RBD in complex with IGLV6-57 antibody S2A4 (PDB 7JVA) [41], which belongs to cluster 7, is shown. The RBD is in white with the receptor binding site highlighted in green. The heavy and light chains of S2A4 are in orange and yellow, respectively. **(E)** Percentages of the S2A4 epitope that are buried by the

light chain, heavy chain (without CDR H3), and CDR H3 are shown as a pie chart. Buried surface area (BSA) was calculated by PISA (Proteins, Interfaces, Structures and Assemblies) at the European Bioinformatics Institute (https://www.ebi.ac.uk/pdbe/prot_int/pistart.html) [74]. **(F-G)** Detailed interactions between the **(F)** light and **(G)** heavy chains of S2A4 and SARS-CoV-2 RBD. Hydrogen bonds and salt bridges are represented by black dashed lines. The color coding is the same as panel D.

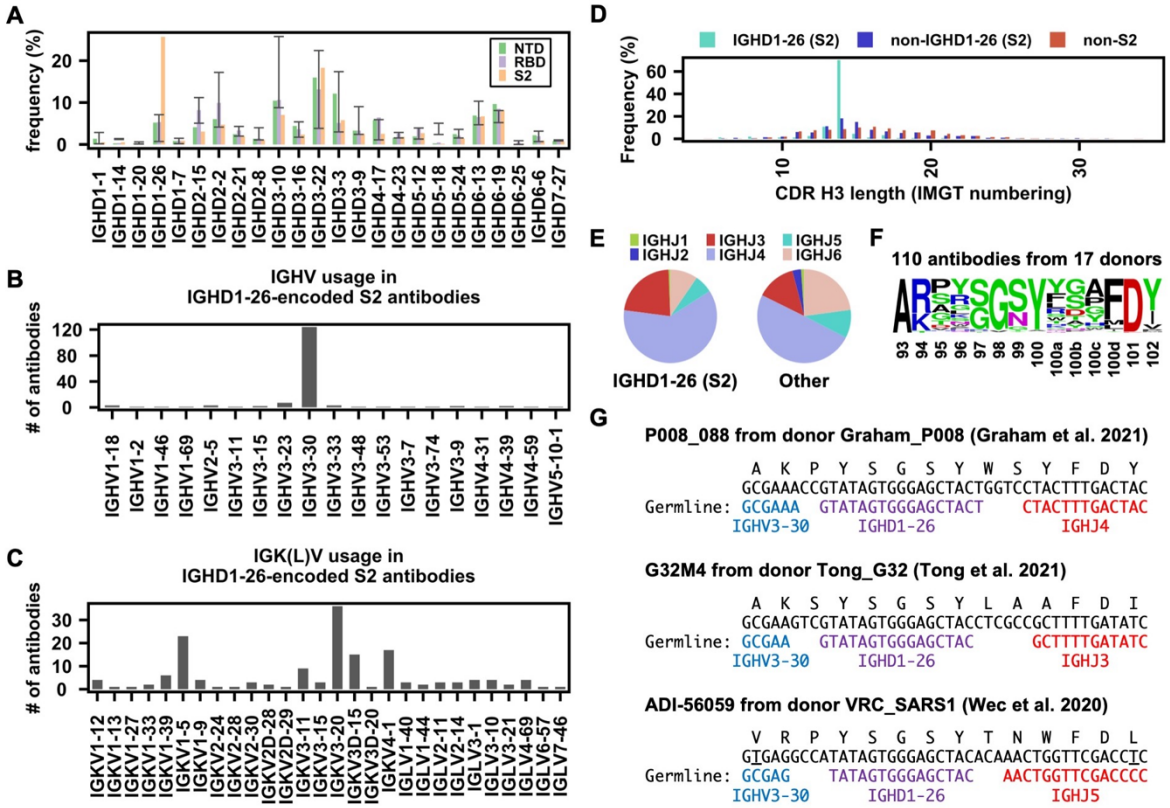


Figure 3. Enrichment of IGHD1-26 in SARS-CoV-2 S2 antibodies. (A) The IGHD gene usage in NTD, RBD, S2 antibodies is shown. Error bars represent the frequency range among 26 healthy donors. **(B)** IGHV gene usage and **(C)** IGK(L)V gene usage among IGHD1-26 S2 antibodies is shown (n = 157). **(D)** The distribution of CDR H3 length (IMGT numbering) in IGHD1-26 S2 antibodies (n = 157), non-IGHD1-26 S2 antibodies (n = 533), and other non-S2 S antibodies that do not target S2 (n = 5,090), are shown. **(E)** The IGHJ gene usage among IGHD1-26 S2 antibodies (n = 157) and other S antibodies with well-defined epitopes (n = 5,623) is shown. **(F)** The CDR H3 sequences for IGHD1-26 S2 antibodies (n = 110) are shown as a sequence logo. **(G)** Amino acid and nucleotide sequences of the V-D-J junction are shown for three IGHD1-26 S2 antibodies [42-44]. Putative germline sequences and segments were identified by IgBlast [66] and are indicated. Somatically mutated nucleotides are underlined. Intervening spaces at the V-D and D-J junctions are N-nucleotide additions.

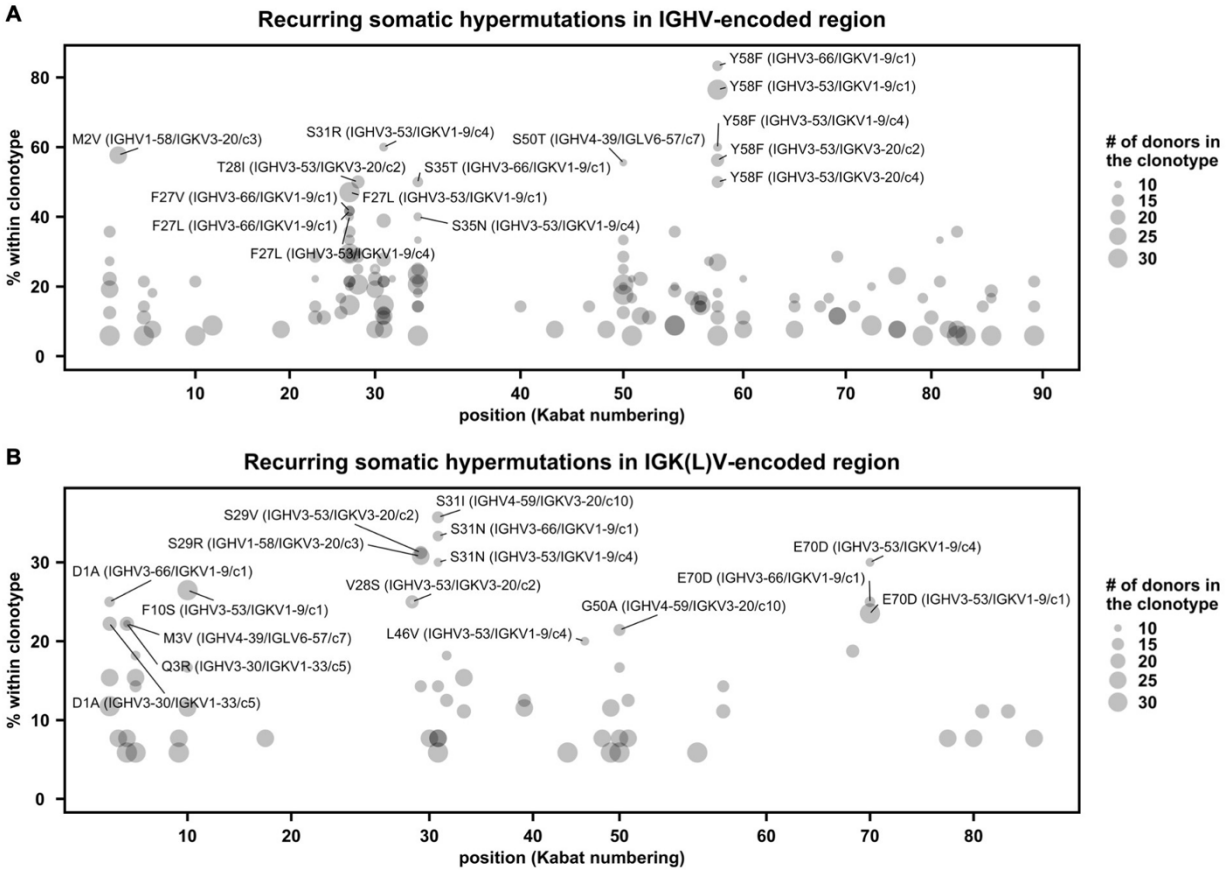
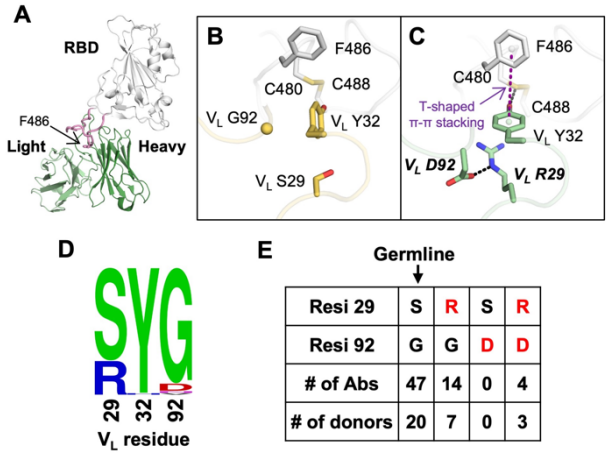


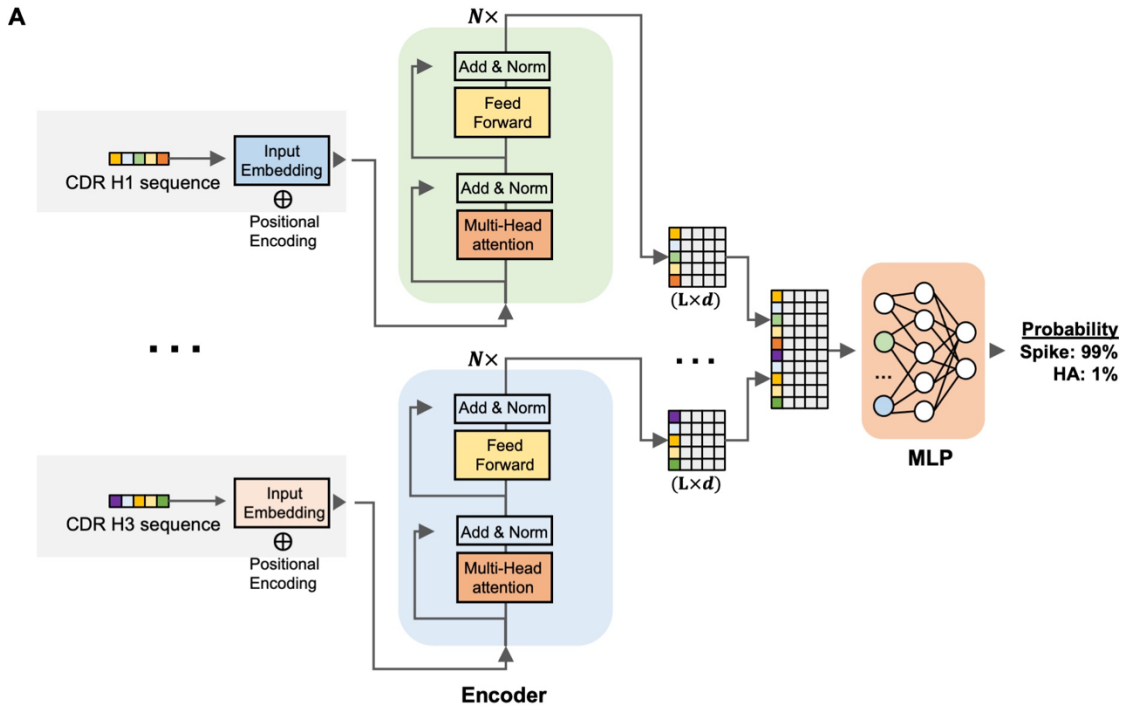
Figure 4. Recurring somatic hypermutations (SHMs) in SARS-CoV-2 S antibodies. (A-B) For each public clonotype, if the exact same SHM emerged in at least two donors, such SHM is classified as a recurring SHM. Only those public clonotypes that can be observed in at least nine donors are shown. **(A)** Recurring SHMs in heavy chain V genes. **(B)** Recurring SHMs in light chain V genes. X-axis represents the position on the V gene (Kabat numbering). Y-axis represents the percentage of donors who carry a given recurring SHM among those who carry the public clonotype of interest. For example, V_L S29R emerged in 8 donors out of 26 donors that carry an public clonotype that is encoded by IGHV1-58/IGKV3-20. As a result, V_L S29R (IGHV1-58/IGKV3-20) is 31% (8/26) within the corresponding clonotype. Of note, since each public clonotype is also defined by the similarity of CDR H3 (see Materials and Methods), there could be multiple clonotypes with the same heavy and light chain V genes (e.g. IGHV3-53/IGKV1-9). The CDR H3 cluster ID for each clonotype is indicated with a prefix “c”, following the information of the V genes.

832 For heavy chain, SHMs that emerged in at least 40% of the donors of the corresponding clonotype
833 are labeled. For light chain, SHMs that emerged in at least 20% of the donors of the corresponding
834 clonotype are labeled.



Resi 29	S	R	S	R
Resi 92	G	G	D	D
# of Abs	47	14	0	4
# of donors	20	7	0	3

Figure 5. Structural analysis of a recurring SHM in the IGHV1-58/IGKV3-20 public clonotype. (A) An overall view of SARS-CoV-2 RBD in complex with the IGHV1-58/IGKV3-20 antibody PDI 222 (PDB 7RR0) [51]. The RBD is shown in white, while the heavy and light chains of the antibody are in dark and light green, respectively. The ridge region (residues 471-491) is shown in pink, with F486 highlighted as sticks. **(B-C)** Structural comparison between two IGHV1-58/IGKV3-20 antibodies that either **(B)** carry germline residues V_L S29/G92 (COVOX-253, PDB 7BEN) [40] and **(C)** somatically hypermutated residues V_L R29/D92 (PDI 222, PDB 7RR0) [51]. SARS-CoV-2 RBD is in white, while antibodies are in yellow (COVOX-253) and green (PDI 222). Somatically mutated residues are labeled with bold and italic letters. The T-shaped π-π stacking between RBD-F486 and V_L Y32 is indicated by a purple dashed line. Hydrogen bond and salt bridge are represented by black dashed lines. **(D)** Sequence logo of V_L residues 29, 32, and 92 among 67 IGHV1-58/IGKV3-20 RBD antibodies are shown. **(E)** Numbers of antibodies in the IGHV1-58/IGKV3-20 public clonotype carrying the germline-encoded variant at V_L residues 29 and 92 (S29, G92), as well as V_L SHM S29R and G92D (red) are listed. Of note, one antibody in this IGHV1-58/IGKV3-20 public clonotype carries S29/N92 and another carries S29/V92. However, they are not listed in the table here.



B

Input CDR	TP	FP	TN	FN	Accuracy	Precision	Recall	ROC AUC	PR AUC
H3	766	229	237	156	0.72	0.77	0.83	0.74	0.83
L3	796	287	179	126	0.70	0.73	0.86	0.72	0.82
H3+L3	795	212	254	127	0.76	0.79	0.86	0.79	0.87
H1+H2	786	173	293	136	0.78	0.82	0.85	0.83	0.90
H1+H2+H3	804	152	314	118	0.81	0.84	0.87	0.86	0.91
L1+L2	774	222	244	148	0.73	0.78	0.84	0.76	0.84
L1+L2+L3	766	226	240	156	0.72	0.77	0.83	0.77	0.86
H1+H2+H3+L1+L2+L3	778	124	342	144	0.81	0.86	0.84	0.87	0.92

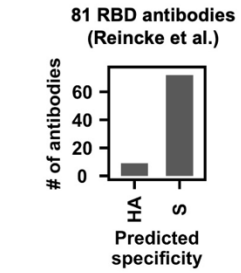


Figure 6. Antigen identification by deep learning. (A) A schematic overview of the deep learning model architecture. **(B)** For evaluating model performance, S antibodies and HA antibodies were considered “positive” and “negative”, respectively. Model performance on the test set was compared when different input types were used. Of note, the test set has no overlap with the training set and the validation set, both of which were used to construct the deep learning model. True positive (TP) represents the number of S antibodies being correctly classified as S antibodies. False positive (FP) represents the number of HA antibodies being misclassified as S antibodies. True negative (TN) represents the number of HA antibodies being correctly classified as HA antibodies. False negative (FN) represents the number of S

antibodies being misclassified as HA antibodies. See Materials and Methods for the calculations of accuracy, precision, recall, ROC AUC, and PR AUC for the training and test sets. **(C)** The antigen specificity of 81 RBD antibodies from Reincke et al. [47] were predicted by a deep learning model that was trained to distinguish between S antibodies and HA antibodies.