Immunity, Volume *55*

Supplemental information

A large-scale systematic survey reveals recurring molecular features of public antibody responses to SARS-CoV-2

Yiquan Wang, Meng Yuan, Huibin Lv, Jian Peng, Ian A. Wilson, and Nicholas C. Wu

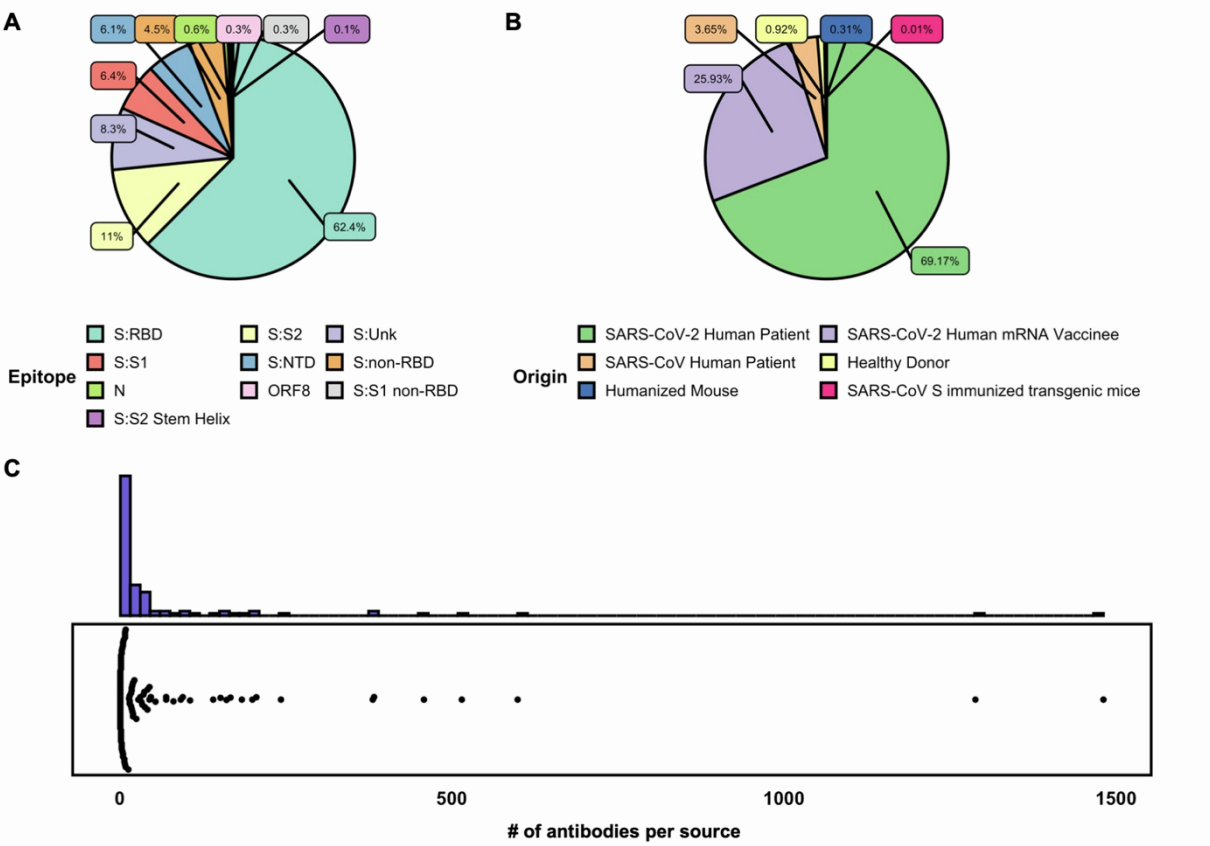

Figure S1. Summary statistics of the antibody dataset, Related to Figure 1. (A) The percentages of antibodies in our antibody dataset that bind to different epitopes are shown. S:Unk represents spike protein with unknown binding domain. **(B)** The percentages of antibodies in our antibody dataset from different origins are shown. **(C)** Each data point in the bottom panel represents one source (i.e. one of the 88 research publications and 13 patents). The number of antibodies from each of the different sources is shown. The distribution of number of antibodies per source is also shown as a histogram in the upper panel.

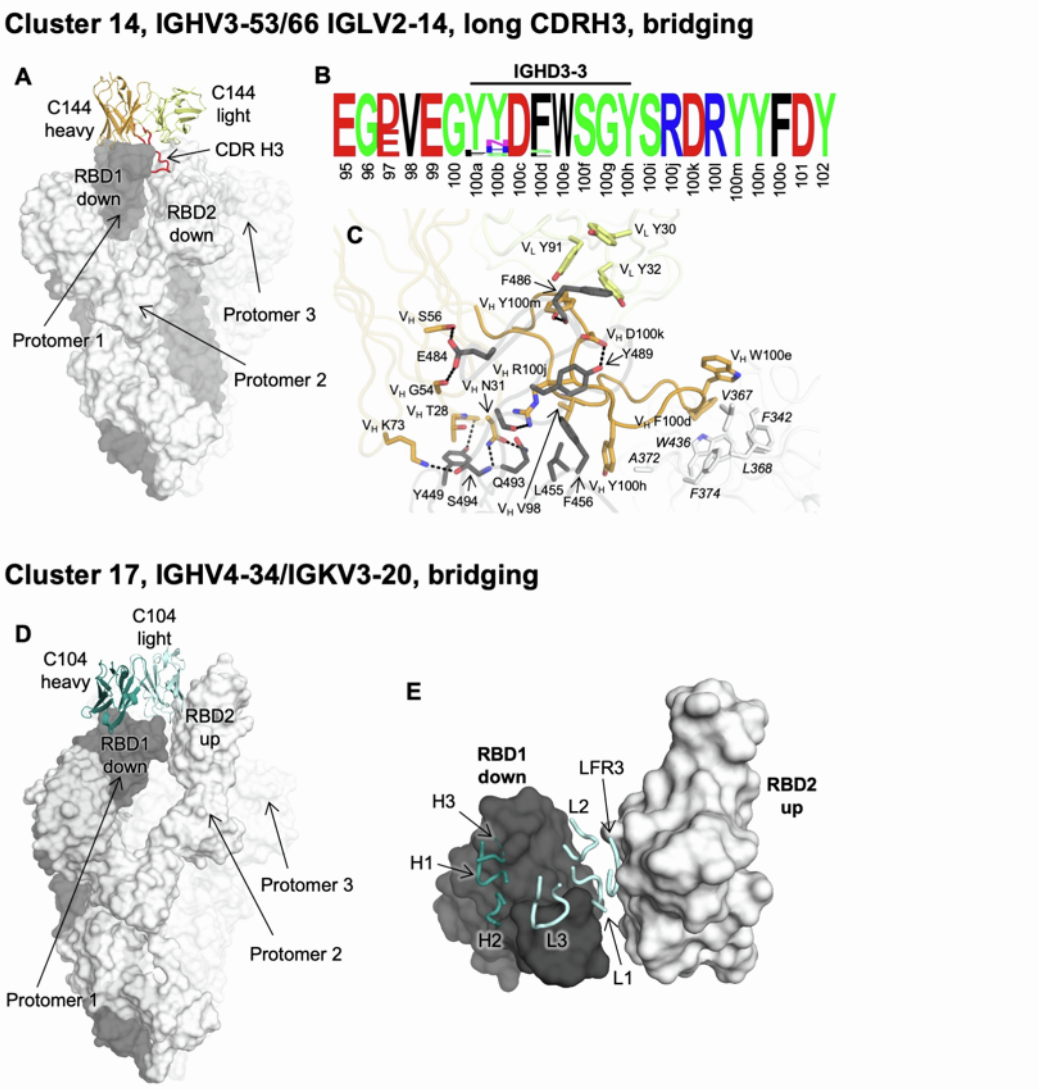

Figure S2. CDR H3 clusters 14 and 17 represent antibodies that bridge two RBDs in the same spike trimer, Related to Figure 2. (A-C) Molecular feature of IGHV3-53/3-66 IGLV2-14 antibodies (cluster 14) that bridges two down RBDs simultaneously in a spike trimer. Cluster 14 contains 19 antibodies from 3 donors. **(A)** An overall view of SARS-CoV-2 spike in complex with an IGHV3-53/3-66 IGLV2-14 antibody. C144 (PDB 7K90) is used as a representative antibody (Barnes et al., 2020). Three protomers of the spike trimer are shown in charcoal, white, and transparent white, respectively. Heavy and light chains of C144 are shown in orange and yellow, respectively, with CDR H3 highlighted in red. The two RBDs bridged by the antibody are indicated in the figure. **(B)** Sequence logo of CDR H3 of IGHV3-53/3-66 IGLV2-14 antibodies, with residue

positions labeled according to Kabat numbering. **(C)** Detailed interactions between C144 and the SARS-CoV-2 spike trimer. The color coding corresponds to that in panel A, where the RBD residues on protomer 1 is shown in charcoal and the RBD residues on protomer 2 are in white. Residue numbers of the RBD on protomer 2 are shown in italic. Hydrogen bonds and salt bridges are represented by black dashed lines. The RBD on protomer 1 is recognized by CDRs H1, H2, and H3 as well as a few light chain residues, where the RBD on protomer 2 is also targeted by CDR H3 of the same antibody molecule. CDR H3, which is encoded by IGHD3-3, contains ^{100d}FW^{100e} at tip that stacks extensively with aromatic residues in the adjacent RBD. **(D-E)** Molecular features of IGHV4-34/IGKV3-20 antibodies (cluster 17) bridging two RBDs (one up and one down) simultaneously in a spike trimer. Cluster 17 contains 13 antibodies from 4 donors. **(D)** An overall view of SARS-CoV-2 spike in complex with an IGHV4-34/IGKV3-20 antibody. C104 (PDB 7K8U) is used as a representative antibody (Barnes et al., 2020). Three protomers of the spike trimer are shown in charcoal, white, and transparent white, respectively. Heavy and light chains of C104 are shown in teal and pale cyan, respectively. The two RBDs bridged by the antibody are indicated in the figure. **(E)** A zoomed-in view of the interactions between the CDR loops of C104 and the two RBDs that display one-up-one-down conformation. Atomic interactions are not shown here as side chains of paratope residues were truncated in the original structure due to the low resolution of the structure. CDR H3 was partially truncated in the original coordinates.

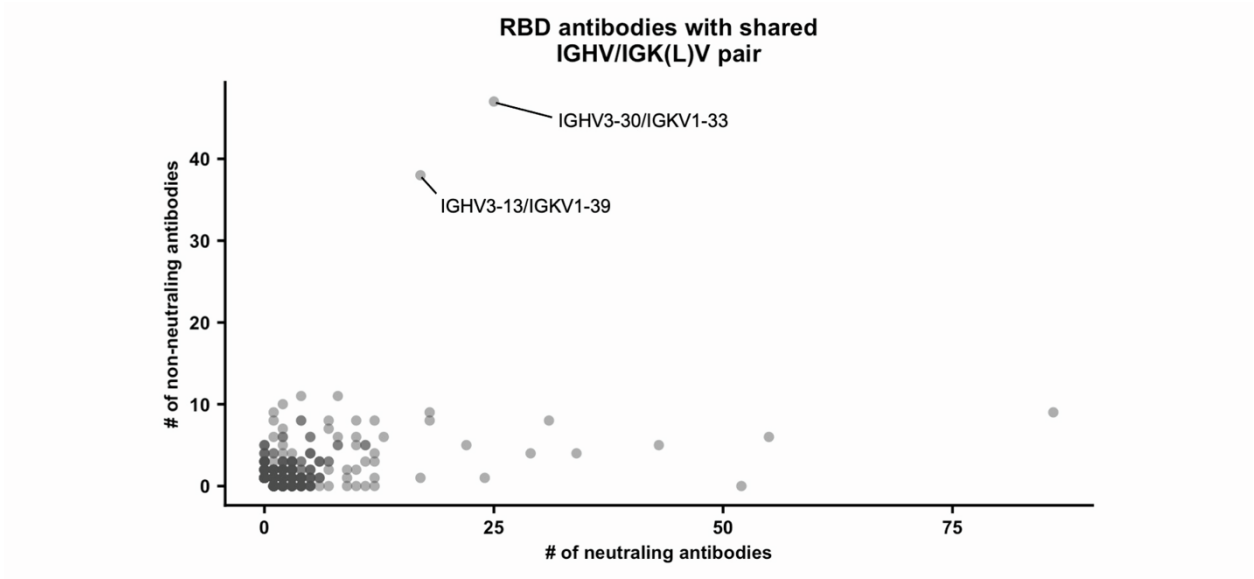

Figure S3. Two IGHV/IGK(L)V pairs are enriched in non-neutralizing RBD antibodies, Related to Figure 2. Each data point represents a defined IGHV/IGK(L)V pair. For each IGHV/IGK(L)V pair, the number of neutralizing RBD antibodies is plotted against the number of non-neutralizing RBD antibodies. Two IGHV/IGK(L)V pairs with a large number of non-neutralizing antibodies are labeled.

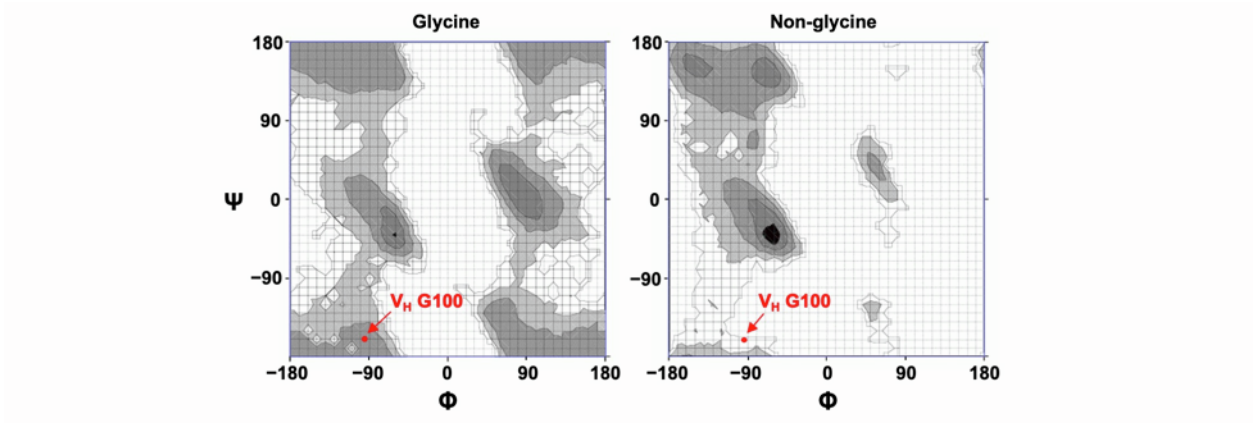

Figure S4. V_H G100 of antibody S2A4 is in the allowed region of Ramachandran plot for non-glycine, Related to Figure 2. V_H G100 of antibody S2A4 (PDB 7JVA) (Piccoli et al., 2020), which is a representative antibody of cluster 7, is shown as red dots on the Ramachandran plots. Black and grey areas represent highly preferred conformations, whereas outlined white area represents allowed conformations.

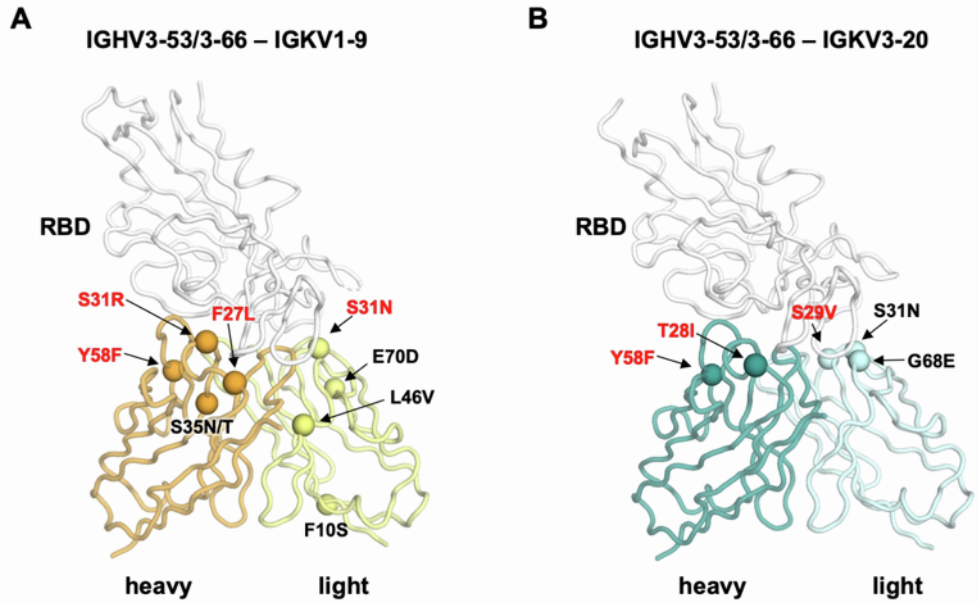

Figure S5. IGHV3-53/3-66-encoded public clonotypes have many recurring somatic mutations, Related to Figure 4. (A) Recurring somatic mutations in the public clonotype encoded by IGHV3-53/3-66 and IGKV1-9 are shown. CC12.1 (PDB 6XC3) is used as an example here (Yuan et al., 2020). **(B)** Recurring somatic mutations in the public clonotype encoded by IGHV3-53/3-66 and IGKV3-20 are shown. CC12.3 (PDB 6XC4) is used as an example here (Yuan et al., 2020). Paratope residues (defined as buried surface area upon binding > 0 $Å^2$ by RBD) are shown in red.

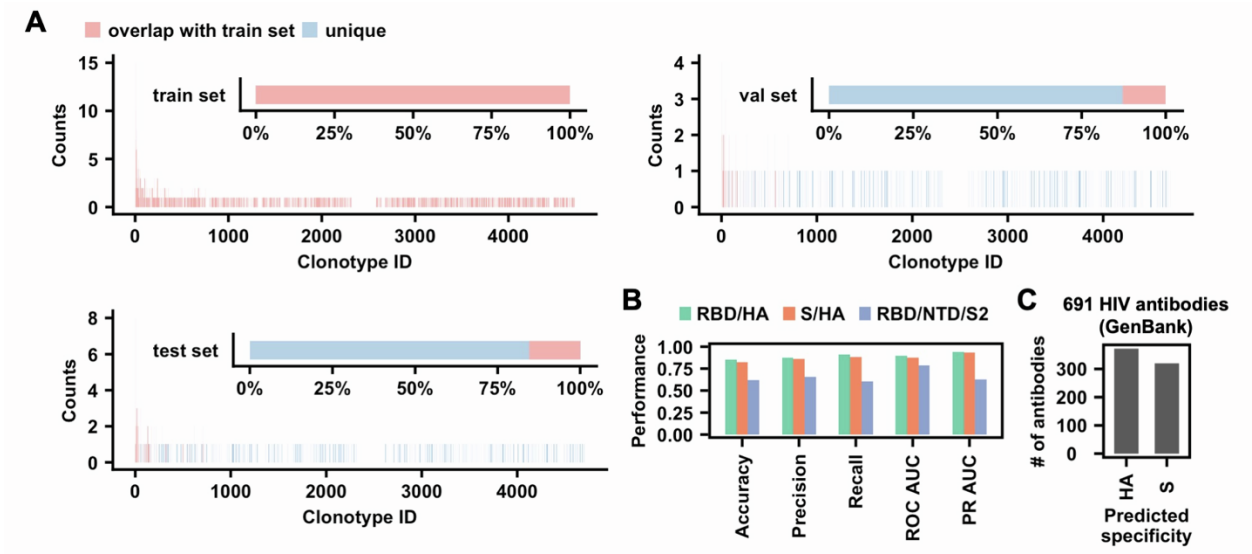

Figure S6. The deep learning model has a robust performance, Related to Figure 6. (A) Antibodies with identical heavy and light chain immunoglobulin variable (V) genes, junction (J) genes, and belonging to the same CDR H3 cluster are defined as a clonotype. This definition was adopted from a recent study on SARS-CoV-2 public clonotypes (Chen et al., 2021). Each clonotype was assigned a unique ID. By comparing the distribution of clonotypes IDs in different sets, we found that only 12.7% and 15.4% clonotypes in the validation (val) and test sets, respectively, overlapped with the training set (train). **(B)** The performances of different models are shown. The RBD/NTD/S2 model was trained by the heavy chain CDRs (H1, H2, and H3), whereas the RBD/HA model was trained by all six CDRs. The dataset for the RBD/NTD/S2 model included 389 NTD antibodies and 674 S2 antibodies with sequence information for all heavy chain CDRs. In addition, the number of RBD antibodies was down-sampled to 800 to avoid data imbalance (see STAR Methods). For the RBD/HA model, the number of RBD antibodies were down-sampled to 3,000, and the same 1,356 HA antibodies as in the S/HA model were used. **(C)** The S/HA model that was trained by six CDRs was applied to a dataset of 691 HIV antibodies with both heavy and light chain sequence information available (**Table S6**).