**INTRODUCTION:** A systematic characterization of the humoral response to severe acute respiratory system coronavirus 2 (SARS-CoV-2) epitopes has yet to be performed. This analysis is important for understanding the immunogenicity of the viral proteome and the basis for cross-reactivity with the common-cold coronaviruses.

Coronavirus disease 2019 (COVID-19), caused by SARS-CoV-2, is notable for its variable course, with some individuals remaining asymptomatic whereas others experience fever, respiratory distress, or even death. A comprehensive investigation of the antibody response in individuals with severe versus mild COVID-19—as well as an examination of past viral exposure history—is needed.

**RATIONALE:** An understanding of humoral responses to SARS-CoV-2 is critical for improving diagnostics and vaccines and gaining insight into variable clinical outcomes. To this end, we used VirScan, a high-throughput method to analyze epitopes of antiviral antibodies in human sera. We supplemented the original VirScan library with additional libraries of peptides spanning the proteomes of SARS-CoV-2 and all other human coronaviruses. These libraries enabled us to precisely map epitope locations and investigate cross-reactivity between SARS-CoV-2 and other coronavirus strains. The original VirScan library allowed us to simultaneously investigate antibody responses to prior infections and viral exposure history.

**RESULTS:** We screened sera from 232 COVID-19 patients and 190 pre–COVID-19 era controls against the original VirScan and supplemental coronavirus libraries, assaying more than 10^8 antibody repertoire–peptide interactions. We identified epitopes ranging from "private" (recognized by antibodies in only a small number of individuals) to "public" (recognized by antibodies in many individuals) and detected SARS-CoV-2-specific epitopes as well as those that cross-react with common-cold coronaviruses. Several of these cross-reacting antibodies are present in pre–COVID-19 era samples. We developed a machine learning model that predicted SARS-CoV-2 exposure history with 99% sensitivity and 98% specificity from VirScan data. We used the most discriminatory SARS-CoV-2 peptides to produce a Luminex-based serological assay, which performed similarly to gold-standard enzyme-linked immunosorbent assays. We stratified the COVID-19 patient samples by disease severity and found that patients who had required hospitalization exhibited stronger and broader antibody responses to SARS-CoV-2 but weaker overall responses to past infections compared with those who did not need hospitalization. Further, the hospitalized group had higher seroprevalence rates for cytomegalovirus and herpes simplex virus 1. These findings may be influenced by differences in demographic compositions between the two groups, but they raise hypotheses that may be tested in future studies. Using alanine scanning mutagenesis, we precisely mapped 823 distinct epitopes across the entire SARS-CoV-2 proteome, 10 of which are likely targets of neutralizing antibodies. One cross-reactive antibody epitope in S2 has been previously suggested to be neutralizing and, as it exists in pre–COVID-19 era samples, could affect the severity of COVID-19.

**CONCLUSION:** We present a highly detailed view of the epitope landscape within the SARS-CoV-2 proteome. This knowledge may be used to produce diagnostics with improved specificity and can provide a stepping stone to the isolation and functional dissection of both neutralizing antibodies and antibodies that might exacerbate patient outcomes through antibody-dependent enhancement or immune distraction.

Our study reveals notable correlations between COVID-19 severity and both viral exposure history and overall strength of the antibody response to past infections. These findings are likely influenced by demographic covariates, but they generate hypotheses that may be tested with larger patient cohorts matched for age, gender, race, and other demographic variables.

Author affiliations and MGH COVID-19 Collection & Processing Team participants and affiliations are available in the full article online.

*These authors contributed equally to this work.

†These authors contributed equally to this work.

‡Corresponding author. Email: selledge@genetics.med.harvard.edu

This is an open-access article distributed under the terms of the Creative Commons Attribution license (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Cite this article as E. Shrock et al., Science 370, eabd4250 (2020). DOI: 10.1126/science.abd4250

READ THE FULL ARTICLE AT

https://doi.org/10.1126/science.abd4250
Understanding humoral responses to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is critical for improving diagnostics, therapeutics, and vaccines. Deep serological profiling of 232 coronavirus disease 2019 (COVID-19) patients and 190 pre–COVID-19 era controls using VirScan revealed more than 800 epitopes in the SARS-CoV-2 proteome, including 10 epitopes likely recognized by neutralizing antibodies. Preexisting antibodies in controls recognized SARS-CoV-2 ORF1, whereas only COVID-19 patient antibodies primarily recognized spike protein and nucleoprotein. A machine learning model trained on VirScan data predicted SARS-CoV-2 exposure history with 99% sensitivity and 98% specificity; a rapid Luminox-based diagnostic was developed from the most discriminatory SARS-CoV-2 peptides. Individuals with more severe COVID-19 exhibited stronger and broader SARS-CoV-2 responses, weaker antibody responses to prior infections, and higher incidence of cytokagelovirus and herpes simplex virus 1, possibly influenced by demographic covariates. Among hospitalized patients, males produce stronger SARS-CoV-2 antibody responses than females.

Coronaviruses constitute a large family of enveloped, positive-sense single-stranded RNA viruses that cause diseases in birds and mammals (1). Among the strains that infect humans are the alphacoronaviruses HCoV-229E and HCoV-NL63 and the betacoronaviruses HCoV-OC43 and HCoV-HKU1, which cause common colds (Fig. 1A). Three additional betacoronavirus species result in severe infections in humans: severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a novel coronavirus that emerged in late 2019 in Asia and quickly spread throughout the world (2). As of November 2020, SARS-CoV-2 has caused more than 50 million confirmed infections and nearly 1.3 million deaths (3). The clinical course of coronavirus disease 19 (COVID-19)—the disease resulting from SARS-CoV-2 infection—is notable for its extreme variability: Some individuals remain entirely asymptomatic, whereas others experience fever, anosmia, diarrhea, severe respiratory distress, pneumonia, cardiac arrhythmia, blood clotting disorders, liver and kidney distress, enhanced cytokine release and, in a small percentage of cases, death (4). Therefore, understanding the factors that influence how this spectrum of outcomes is an intense area of research. Disease severity is correlated with advanced age, sex, ethnicity, socioeconomic status, and comorbidities such as diabetes, cardiovascular disease, chronic lung disease, obesity, and reduced immune function (4). Additional relevant factors likely include the inoculum of virus at infection and individual genetic background and viral exposure history. The complex interplay of these elements also determines how individuals respond to therapies aimed at mitigating disease severity. Detailed knowledge of the immune response to SARS-CoV-2 could improve our understanding of diverse outcomes and inform the development of improved diagnostics, vaccines, and antibody-based therapies.

In this study, we used VirScan, a programmable phage-display immunoprecipitation and sequencing (PhIP-Seq) technology that we developed previously, to explore antiviral antibody responses across the human virome (5–8). Here we describe a detailed analysis of the humoral response in COVID-19 patients.

Results

Development of a VirScan library targeting human coronaviruses

Our existing VirScan phage-display platform is based on an oligonucleotide library encoding 56-mer amino acid (56-mer) peptides tiling every 28 amino acids across the proteomes of all known pathogenic human viruses (~400 species and strains) plus many bacterial proteins (8). To investigate the serological response to SARS-CoV-2 and other human coronaviruses (HCoVs), we supplemented this library with three additional sublibraries: Sublibrary 1 encodes a 56-mer peptide library tiling every 28 amino acids through each of the open reading frames (ORFs) expressed by the six HCoVs and three bat coronaviruses closely related to SARS-CoV-2; sublibrary 2 encodes 20-mer peptides tiling every 5 amino acids across the SARS-CoV-2 proteome, enabling more precise localization of epitopes; and sublibrary 3 encodes triple-alanine scanning mutants of the 56-mer peptides tiling across the SARS-CoV-2 proteome, enabling the mapping of epitope boundaries at amino acid resolution (Fig. 1, A to C, and table S1) (9, 10).

We used VirScan (Fig. 1C) to profile the antibody repertoires of nine cohorts of individuals from multiple locations, including Baltimore, MD, Boston, MA, and Seattle, WA (tables S2 to S8). These cohorts comprised longitudinal samples from individuals enrolled in prospective studies of COVID-19 infection, cross-sectional samples from patients with active...
Fig. 1. VirScan detects the humoral response to SARS-CoV-2 in sera from COVID-19 patients. (A) Phylogeny tree of 50 coronavirus sequences (32) constructed using MEGA X (33, 34). The scale bar indicates the estimated number of base substitutions per site (35). Coronaviruses included in the updated VirScan library are indicated in red. (B) Schematic representation of the ORFs encoded by the SARS-CoV-2 genome (10, 36). (C) Overview of the VirScan procedure (5–8). The coronavirus oligonucleotide library includes 56-mer peptides tiling every 28 amino acids (aa) across the proteomes of 10 coronavirus strains and 20-mer peptides tiling every 5 amino acids across the SARS-CoV-2 proteome. Oligonucleotides were cloned into a T7 bacteriophage display vector and packaged into phage particles displaying the encoded peptides on their surface. The phage library was mixed with sera containing antibodies that bind to their cognate epitopes on the phage surface; bound phage were isolated by IP with either anti-IgG- or anti-IgA-coated magnetic beads. Lastly, PCR amplification and Illumina sequencing from the DNA of the bound phage revealed the peptides targeted by the serum antibodies. (D) Detection of antibodies targeting coronavirus epitopes by VirScan. Heatmaps depict the humoral response from COVID-19 patients (n = 232) and pre–COVID-19 era control samples (n = 190). Each column represents a sample from a distinct individual. The color intensity indicates the number of 56-mer peptides from the indicated coronaviruses significantly enriched by IgG antibodies in the serum sample. (E) Box plots illustrate the number of peptide hits from the indicated coronaviruses in COVID-19 patients and pre–COVID-19 era controls. The box indicates the interquartile range, with a line at the median. The whiskers span 1.5 times the interquartile range.
COVID-19 who were receiving treatment in either hospital or outpatient settings, and cross-sectional samples from convalescent individuals with a past history of COVID-19. Our cohorts also included a diverse set of control sera collected before the COVID-19 outbreak. We profiled the targets of IgG and IgA (immunoglobulins G and A) antibodies separately: IgG and IgA are the most abundant isotypes in blood, whereas IgA is the principal isotype secreted on mucosal surfaces, including the respiratory tract. Collectively, we analyzed ~550 samples in duplicate, in total assessing ~100 million potential antibody repertoire–peptide interactions.

Detection of SARS-CoV-2 seropositivity with VirScan

To measure immune responses to SARS-CoV-2, we compared VirScan profiles of serum samples from COVID-19 patients to those of controls obtained before the emergence of SARS-CoV-2 in 2019. These pre–COVID-19 era controls facilitate identification of (i) SARS-CoV-2 peptides encoding epitopes specific to COVID-19 patients and (ii) SARS-CoV-2 peptides encoding epitopes that are cross-reactive with antibodies developed in response to the ubiquitous common-cold HCoVs. Sera from COVID-19 patients exhibited much more SARS-CoV-2 reactivity than did sera from pre–COVID-19 era controls (Fig. 1, D and E). Some cross-reactivity toward SARS-CoV-2 peptides was observed in the pre–COVID-19 era samples, but this was expected because nearly all people have been exposed to one or more HCoVs (II).

COVID-19 patient sera also showed significant levels of cross-reactivity with the other highly pathogenic HCoVs, SARS-CoV and MERS-CoV, although less was observed against the more distantly related MERS-CoV. Extensive cross-reactivity was also observed against peptides derived from the three bat coronaviruses that share the greatest proportion of sequence identity with SARS-CoV-2 (Fig. 1, A, D, and E) (9). We know that these represent cross-reactivities because, given the low prevalence and circumscribed geographical location of SARS-CoV and MERS-CoV, none of the individuals in this study are likely to have encountered these viruses.

COVID-19 patient sera also exhibited a significantly higher level of reactivity to seasonal HCoV peptides than did sera from pre–COVID-19 era controls (Fig. 1, D and E). This could be due to the elicitation of novel antibodies that cross-react or to an anamnestic response that boosts B cell memory against HCoVs. The converse is not always true: Many pre–COVID-19 era samples exhibit strong recognition of seasonal HCoV peptides but little or no recognition of SARS-CoV-2 peptides (Fig. 1D).

In some cases, the concentrations of antibodies against seasonal HCoVs may be below the threshold of detection in the pre–COVID-19 era samples.

Coronavirus proteins targeted by antibodies in COVID-19 patients

Analysis of SARS-CoV-2 proteins targeted by COVID-19 patient antibodies revealed that the primary responses to SARS-CoV-2 are reactive with peptides derived from spike protein (S) and nucleoprotein (N) (Fig. 2, A and B). Compared with sera from pre–COVID-19 era controls, COVID-19 patient sera exhibit significant differential recognition of these two proteins, indicating that this recognition is a result of antibody responses to SARS-CoV-2. Third-most frequently recognized is the replicase polyprotein ORF1, but unlike S and N, ORF1 is recognized to a similar extent by sera from COVID-19 patients and pre–COVID-19 era controls. This suggests that recognition of SARS-CoV-2 ORF1 is a result of cross-reactions from antibodies elicited by exposure to other pathogens, possibly HCoVs. Antibody responses to peptides from membrane glycoprotein (M), ORF3, and ORF9b were occasionally detected in COVID-19 patients.

We also analyzed longitudinal samples from 23 COVID-19 patients. Most patients displayed an antibody response to peptides derived from S or N in the second week after symptom onset, with many displaying an antibody response by the end of the first week (Fig. 2C). The relative strength and onset of the antibody response to S and N differed markedly between individuals, and the initial immune response showed no preference for S or N. The signal intensity of antibodies recognizing SARS-CoV-2 ORF1 epitopes did not increase over time, further suggesting that ORF1 antibodies likely represent a preexisting cross-reactive response.

Identification of immunogenic regions of SARS-CoV-2 proteins

To more precisely define the immunogenic regions of the SARS-CoV-2 proteome, we examined the specific 56-mer and 20-mer peptides detected by VirScan in COVID-19 patients compared with those in pre–COVID-19 era controls. An example IgG response from a single patient to SARS-CoV-2 S and N is shown in Fig. 3A. We observed strong concordance between the viral regions enriched by the 56-mer and 20-mer fragments, demonstrating the robustness of VirScan. In many cases, we observed recognition of overlapping 56-mer peptides, indicating an epitope in the common region.

Next, we compared the protein regions recognized by IgG and IgA across COVID-19 patients (Fig. 3B). We identified four regions each in S and N that are recurrently targeted by antibodies from >15% of COVID-19 patients, with additional regions recognized less frequently. Overall, IgG and IgA recognize the same protein regions with similar frequencies across the population. However, when IgG and IgA responses were compared within individuals, we observed considerable divergence (Fig. 3C): Many epitopes were recognized by only IgG, only IgA, or both IgG and IgA within an individual patient. Together, these data suggest that patients generate distinct IgG and IgA antibody responses to SARS-CoV-2, but the targeted regions are largely shared at the population level.

Machine learning guides the design of a Luminex assay for rapid COVID-19 diagnosis

To predict SARS-CoV-2 exposure history from VirScan data, we developed a gradient-boosting algorithm (XGBoost) that integrates both IgG and IgA data and predicts current or past COVID-19 with 99.1% sensitivity and 98.4% specificity (Fig. 4, A and B). We used Shapley Additive exPlanations (SHAP)—a method to compute the contribution of each feature of the data to the predictive model (12)—to identify peptides from SARS-CoV-2 S and N plus homologous peptides from SARS-CoV and BatCoV-HKU-3 and BatCoV-279 that were highly predictive of SARS-CoV-2 exposure (Fig. 4, C and D). We leveraged these insights to develop a simple, rapid Luminex-based diagnostic for COVID-19. We chose 12 SARS-CoV-2 peptides predicted by VirScan data and the machine learning model to be highly indicative of SARS-CoV-2 exposure history (table S9). These SARS-CoV-2 peptides, two positive control peptides from rhinovirus A and Epstein-Barr virus (EBV) that are recognized in >80% of seropositive individuals by VirScan (7), and a negative control peptide from HIV-1 were coupled to Luminex beads (13). We tested 163 COVID-19 patient samples and 165 pre–COVID-19 era controls for IgG reactivity to the Luminex panel. We detected clear responses to SARS-CoV-2 peptides in COVID-19 patient samples but rarely in the pre–COVID-19 era controls (Fig. 4E). Using the Luminex data, we developed a logistic regression model that predicts COVID-19 infection history with 89.6% sensitivity and 95.2% specificity [area under the curve (AUC) = 0.97] (Fig. 4, F and G). A subset of COVID-19–positive samples (n = 107) was also examined with an in-house enzyme-linked immunosorbent assay (ELISA) using three SARS-CoV-2 antigens: N, S, and the S receptor-binding domain (RBD). Considering a sample to be positive if it scored above the 99% specificity threshold on any one of the three ELISA antigens, we determined that the sensitivity of the Luminex assay for this subset (88.8%) was similar to that of the ELISA (90.7%) (fig. S1). Among samples run on all three assays, VirScan significantly outperformed both the Luminex and ELISAs (fig. S1, A and C). Notably, our optimal model integrated only three SARS-CoV-2 peptides—residues 386 to 406 of N (N 386-406), residues 810 to 830 of S (S 810-830), and residues 1146 to 1166 of S (S 1146-1166)—which were also the
most discriminatory 20-mers in the VirScan data. IgG responses in COVID-19 patients were highly correlated between the Luminex and VirScan assays, providing orthogonal validation of the VirScan data and supporting the prevalence of SARS-CoV-2–induced humoral responses to these regions of S and N (fig. S1D).

Differential antibody responses to common viruses in hospitalized versus nonhospitalized COVID-19 patients

We next considered whether differences in the antibody response to SARS-CoV-2 or to other viruses might be associated with the severity of COVID-19. We grouped the COVID-19 patients into two subsets: those who required hospitalization (n = 101) and those who did not (n = 131). We compared the responses to peptides derived from the SARS-CoV-2 S and N proteins between the hospitalized (H) and nonhospitalized (NH) groups and found that the H group exhibited stronger and broader antibody responses to S and N peptides that might be due to epitope spreading (Fig. 5A).

We then analyzed 32 NH COVID-19 samples, 32 H COVID-19 samples, and 32 pre–COVID-19 era negative controls with the Luminex assay and similarly observed that the H group had stronger and broader antibody responses to SARS-CoV-2–specific peptides than did the NH group (Fig. 5B).

VirScan also offers the opportunity to examine the history of previous viral infections and to determine correlates of COVID-19 outcomes. For example, prior viral exposure could provide some protection if cross-reactive neutralizing antibodies or T cell responses are stimulated upon exposure to SARS-CoV-2 (14, 15). Alternatively, cross-reactive antibodies to viral
Fig. 3. IgG and IgA recognition of immunodominant regions in SARS-CoV-2 spike and nucleoprotein. (A) Example response to S and N proteins from a single COVID-19 patient. The y axis indicates the strength of enrichment (z-score; see Materials and methods) of each 56-mer (blue) or 20-mer (red) peptide recognized by the IgG antibodies present in the serum sample. (B) Common responses to S and N proteins across COVID-19 patients. The y axis indicates the fraction of COVID-19 patient samples (n = 348) enriching each 20-mer peptide with either IgG (top) or IgA (bottom) antibodies. (C) Comparison of the IgA and IgG responses in individual COVID-19 patients. Each set of two rows represents the IgG and IgA antibody specificities of a single patient, with data displayed for 10 representative COVID-19 patients. Numeric values indicate the degree of enrichment (z-score) of each peptide tiling across the S and N proteins.
Fig. 4. Machine learning models trained on VirScan data discriminate COVID-19–positive and –negative individuals with very high sensitivity and specificity. (A) Gradient-boosting machine learning models were trained on IgG and IgA VirScan data from 232 COVID-19 patients and 190 pre–COVID-19 era controls. Separate models were created for the IgG and IgA data, and then a third model (Ensemble) was trained to combine the outputs of the first two. (B) The plot shows the predicted probability that each sample is positive for COVID-19. True COVID-19–positive samples are shown as red dots; true COVID-19–negative samples are shown as gray dots. The corresponding confusion matrix for each model is shown on the right. (C and D) SHAP analysis to identify the most discriminatory peptides informing the models in (B). The chart in (C) summarizes the relative importance of the most discriminatory peptides increased among COVID-19 patients identified by the IgG and IgA gradient-boosting models. The enrichment [log$_2$(fold change)] of the normalized read counts in the sample IP versus in no-serum control reactions] of each of these peptides across all samples is shown in (D). (E) Luminex assay using highly discriminatory SARS-CoV-2 peptides identifies IgG antibody responses in COVID-19 patients but rarely in pre–COVID-19 era controls. Each column represents a COVID-19 patient (n = 163) or pre–COVID-19 era control (n = 165); each row is a SARS-CoV-2–specific peptide. Peptides containing public epitopes from rhinovirus A, EBV, and HIV-1 served as positive and negative controls. The color scale indicates the median fluorescence intensity (MFI) signals after background subtraction. (F) Receiver operating characteristic (ROC) curve for the Luminex assay predicting SARS-CoV-2 infection history, evaluated by 10× cross-validation. The light red lines indicate the ROC curve for each test set, the dark line indicates the average, and the gray region represents ±1 SD. The average area under the curve (AUC) is shown. (G) (Left) Predicted probability that each sample is positive for COVID-19, using the Luminex model, as in (B). The dashed line indicates the model threshold. (Right) Confusion matrix for the Luminex model.
Fig. 5. Correlates of COVID-19 severity. (A) Differential recognition of peptides from SARS-CoV-2 N and S between COVID-19 nonhospitalized patients (n = 131), hospitalized patients (n = 101), and pre–COVID-19 era negative controls (n = 190). Each column represents a specific patient and each row represents a peptide tile; tiles are labeled by amino acid start and end position and may be duplicated for intervals for which amino acid sequence diversity is represented in the library. Color intensity represents the degree of enrichment (z-score) of each peptide in IgG samples. Asterisks indicate peptides that exhibit a significant increase in recognition by sera from hospitalized versus nonhospitalized patients (Kolmogorov-Smirnov test, Bonferroni-corrected P value thresholds of 0.001 for S and 0.0025 for N). (B) SARS-CoV-2 Luminex assay identifies stronger IgG responses in hospitalized COVID-19 patients than in nonhospitalized COVID-19 patients. Each column represents either a nonhospitalized (n = 32) or hospitalized (n = 32) COVID-19+ patient or a pre–COVID-19 era control (n = 32); each row represents a peptide in the Luminex assay. The color scale indicates the MFI signals after background subtraction. (C) All peptides in the VirScan library are plotted by the fraction of nonhospitalized (x axis) and hospitalized COVID-19 patient IgG samples (y axis) in which they are recognized. A z-score threshold of 3.5 was used as an enrichment cutoff to count a peptide as positive. Peptides that exhibit statistically significant associations with hospitalization status are colored by virus of origin (Fisher’s exact test, Bonferroni-corrected P value threshold of $8.52 \times 10^{-7}$). All peptides that do not exhibit significant association with hospitalization status are shown in gray. The significant peptides shown are collapsed for high sequence identity. (D) All peptides derived from CMV that are present in the VirScan library are plotted by median z-score for the nonhospitalized (x axis) and hospitalized COVID-19 patients (y axis). The line $y = x$ is shown as a dashed line. (E) Reduced recognition of mild disease–associated antigens with age. The histogram shows the relative recognition in healthy donors at age 58 compared with age 42 for each distinct antigen that was more strongly recognized by antibodies in nonhospitalized than hospitalized COVID-19 patients.
To examine these possibilities, we analyzed certain viruses could affect the response to amine the relative strength of the antibody versus 45.8% (60/131) of the NH group. To examine 101) of the H group were positive for HSV-1 in infection: 82.2% (83 of 101) of the H group had a higher incidence of both CMV and HSV-1 whereas the H patients displayed more robust responses to peptides from cytomegalovirus (CMV) and herpes simplex virus 1 (HSV-1) (Fig. 5C). These observations may be influenced by demographic differences in the NH and H cohorts, as described below. We sought to understand whether the differential reactivity to CMV and HSV-1 between the H and NH patients was due to differences in the strength of antibody responses or the prevalence of infection (these viruses are common, but not ubiquitous like rhinoviruses, enteroviruses, and influenza viruses). Using VirScan data, we found that the H group had a higher incidence of both CMV and HSV-1 infection: 82.2% (83 of 101) of the H group were positive for CMV versus 37.4% (49 of 131) of the NH group, whereas 92.1% (93 of 101) of the H group were positive for HSV-1 versus 45.8% (60/131) of the NH group. To examine the relative strength of the antibody responses, we considered only CMV- or HSV-1-seropositive individuals from the NH and H groups. The antibody response to both CMV (Fig. 5D) and HSV-1 (fig. S2) was stronger among the NH individuals. Thus, the differing seroprevalence of CMV and HSV-1 in the NH versus H groups likely explains the results shown in Fig. 5C. We conclude that antibody responses to nearly all viruses, except SARS-CoV-2, were weaker in the H patients than in the NH patients. These notable differences led us to examine potential demographic covariates between the NH and H groups. We found that age, sex, and race were all significantly associated with COVID-19 severity (fig. S3), as has been reported (18, 19). Older age, male sex, and non-white racial groups were significantly over-represented in the H group compared with the NH group (fig. S3 and table S3). Furthermore, hospitalized males exhibited stronger responses to N than hospitalized females, whereas nonhospitalized males and females did not exhibit differential responses to any SARS-CoV-2 proteins (fig. S3E). Advanced age is a dominant risk factor for severe COVID-19 and is correlated with reduced immune function (20). In light of the age difference between the H (median age: 58) and NH (median age: 42) patients in our cohort, we reasoned that the antigens recognized more strongly in the NH group might reflect more general age-associated changes in humoral immunity. To test this hypothesis, we examined VirScan data for a cohort of 648 healthy, pre-pandemic donors. We characterized the recognition of each NH-associated peptide in subsets of healthy donors representing different age groups and observed a general decline in recognition with age, including a median 19% reduction in recognition from age 42 to 58 (Fig. 5E). These data suggest that age-related changes to the immune system may partially explain the observation of weaker antibody responses to most viruses in the H group. Although it is correlative and potentially influenced by other demographic differences between the NH and H cohorts, the broad age-related diminution in immune system activity that we observed could be an important aspect of the increased severity in the H group.

**Cross reactivity of SARS-CoV-2 epitopes**

We returned to the question of epitope cross-reactivity, this time examining antibody responses to the triple-alanine scanning library. For each 56-mer peptide spanning the SARS-CoV-2 proteome, this library contained a collection of scanning mutants: The first mutant peptide encoded three alanines instead of the first three residues, the second mutant peptide contained the three alanines moved one residue downstream, and so on (fig. S4). Antibodies that recognize the wild-type 56-mer peptide will not recognize mutant versions of the peptide containing alanine substitutions at critical residues. Thus, the location of the linear epitope can be deduced by looking for “antibody footprints,” indicated by stretches of alanine mutants missing from the pool of immunoprecipitated phage. The first and last triple-alanine mutations to interfere with binding are expected to start two amino acids before the first residue that is essential for antibody binding and end two amino acids after the last.

With respect to cross-reactivity, IgG from COVID-19 patients recognized more 56-mer peptides from the common HCoVs HKU1, OC43, 29E, and NL63 than IgG from pre-COVID-19 era controls. This difference is primarily driven by a pronounced increase in recognition of S peptides from the HCoVs and is likely a result of cross-reactivity of antibodies developed during SARS-CoV-2 infection (Fig. 6A).

We mapped the position of all HCoV S peptides that display increased recognition in COVID-19 patient samples onto the SARS-CoV-2 S protein. This revealed four immunodominant regions recognized by >25% of COVID-19 patients (Fig. 6B). Comparing the frequency of peptide recognition between the COVID-19 patients and pre-COVID-19 era controls showed that two of these immunogenic regions in SARS-CoV-2 S are likely to cross-react strongly with homologous regions of other HCoVs, as the frequency of recognition of the HCoV peptides at these regions rises in COVID-19 patients. For instance, peptides from all four seasonal HCoVs that span the region corresponding to residues 811 to 830 of SARS-CoV-2 S are frequently recognized by COVID-19 patients but much less so by pre-COVID-19 era controls, suggesting that this recognition is a result of antibodies developed or boosted in response to SARS-CoV-2 infection. Using triple-alanine scanning mutagenesis (fig. S4), we mapped the antibody footprints in this region to an 11-amino acid stretch that is highly conserved between SARS-CoV-2 and all four common HCoVs, which explains the cross-reactivity (Fig. 6, C and D). Similarly, both SARS-CoV-2 and HCoV-OC43 peptides corresponding to S 1144-1163 are recognized much more frequently by COVID-19 patients than pre-COVID-19 era controls, and triple-alanine-scanning mutagenesis confirmed that the antibody footprints are located within a 10-amino acid stretch conserved between SARS-CoV-2 and HCoV-OC43 but not the other HCoVs. By contrast, the epitope sequences around S 551-570 and S 766-785 are not conserved between SARS-CoV-2 and the seasonal HCoVs, and indeed these epitopes are not cross-reactive. One HCoV-HKU1 peptide spanning S 551-570 scores in both COVID-19 patients and pre-COVID-19 era control samples; however, its frequency of detection is not further boosted in COVID-19 patients, suggesting that the antibody that recognizes the SARS-CoV-2 S 551-570 peptide is distinct from the antibody recognizing the HCoV-HKU1 peptide, consistent with sequence differences at this location (Fig. 6C).

Notably, we detect antibody responses to SARS-CoV-2 S 811-830 in 79.9% of COVID-19 patients. However, we also see responses to the corresponding peptides from OC43 and 29E in ~20% of the pre-COVID-19 era controls, and these responses seem to cross-react with SARS-CoV-2. It is possible that some patients have preexisting antibodies to this region that cross-react and are expanded during SARS-CoV-2 infection. This might explain the very high prevalence of antibody responses to this epitope and suggests that anamnestic responses to seasonal coronaviruses may influence the antibody response to SARS-CoV-2. Of note, this region is located directly after the predicted S2’ cleavage site for SARS-CoV-2 and overlaps the fusion peptide. A recent study showed that adding an excess of the fusion peptide reduced neutralization, implying that an antibody that binds the fusion peptide might contribute to neutralization by interfering with membrane fusion (21, 22). Given the frequency of seroreactivity toward this epitope in COVID-19 patients, it will be important to determine whether the antibodies that recognize this epitope are neutralizing in future studies. If so, the prior presence of antibodies recognizing this
Fig. 6. Cross-reactive epitopes among human coronaviruses. (A) Bar graphs depict the average number of 56-mer peptides derived from SARS-CoV-2, SARS-CoV, and each of the four common HCoVs that are significantly enriched per sample (IgG IP). Error bars represent 95% confidence intervals. (B) Analysis of cross-reactive epitopes for HCoV S proteins. The upper plot shows the similarity of each region of the SARS-CoV-2 S protein to the corresponding region in the four common HCoVs (see Materials and methods). The frequency of peptide recognition is shown in the bottom two plots. Peptides from each virus are indicated by colored lines: The length of each line along the x axis indicates the corresponding region of the SARS-CoV-2 S protein covered by each peptide according to a pairwise protein alignment; the height of each line corresponds to the fraction of samples in which that peptide scored in either the IgG or IgA IPs. The epitopes mapped in (C) and (D) are highlighted in pink. (C and D) Mapping of recurrently recognized SARS-CoV-2 S IgG (C) and IgA (D) epitopes by triple-alanine scanning mutagenesis. Each plot represents a 20–amino acid region of the SARS-CoV-2 S protein within the regions highlighted in (B). Each column of the heatmap corresponds to an amino acid position, and each row represents a sample. The color intensity indicates the average enrichment of 56-mer peptides containing an alanine mutation at that site relative to the median enrichment of all mutants of that 56-mer in each sample. COVID-19 patients with a minimum relative enrichment below 0.6 in the specified window are shown. The amino acid sequence across each region of SARS-CoV-2 S, as well as an alignment of the corresponding sequences in the common HCoVs, is shown below each heatmap. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
epitope may affect the course of COVID-19 and mitigate severity.

**Epitope mapping reveals hundreds of distinct SARS-CoV-2 epitopes, including likely epitopes of neutralizing antibodies**

We also used the triple-alanine scanning mutagenesis library to map antibody footprints across the entire SARS-CoV-2 proteome (Fig. 7, fig. S5, and tables S10 to S19). We used a hidden Markov model (HMM) to analyze the mutagenesis data and detect antibody footprints. By integrating signals across stretches of consecutive residues, the HMM successfully distinguished antibody footprints from random noise and thus detected regions containing epitopes with improved sensitivity and far greater resolution than was possible with the 56-mer peptide data alone (see Materials and methods) (figs. S6 and S7 and tables S15 to S18). We performed hierarchical clustering on the antibody footprints identified by the HMM to determine the number of distinct epitopes (here defined as distinct antibody footprints) that we detected across the SARS-CoV-2 proteome (fig. S8 and table S10). Overall, we identified 3103 antibody footprints across 169 COVID-19 patient samples and mapped 823 distinct epitopes (table S19). These epitopes are not evenly distributed along the proteins but rather fall into 303 epitope clusters, each of which contains multiple overlapping epitopes (fig. S8). For example, across the 89 IgA samples that recognized the epitope cluster from S 1135-1165, we identified nine epitopes that overlap but have distinct triple-alanine scanning profiles that indicate distinct antibody-epitope interactions (fig. S8C). Individual epitopes are recognized at a wide range of frequencies in the COVID-19 patients. The average COVID-19 patient sample contained antibodies to ~18 distinct linear epitopes (fig. S9), although this is likely an underestimate of the total epitope count per person, as VirScan does not efficiently detect antibodies recognizing discontinuous (conformational) epitopes (although such antibodies may retain some affinity to linear peptides that constitute the epitope).

The SARS-CoV-2 epitope landscape includes regions recognized by antibodies in a large fraction of COVID-19 patients (“public” epitopes) and regions recognized by antibodies in only one or a few individuals (“private” epitopes). For example, we mapped six distinct epitopes in the region spanning N 151-175 (fig. S5C). One of these epitopes was recognized by nearly one-third of the COVID-19 patients, whereas the rest were detected by <2% of the COVID-19 patients. Similarly, the region spanning S 766-835 contained more than 20 distinct epitopes, including the highly public epitope cluster near S815 and the public epitope cluster near S770 that is preferentially recognized by IgA (Fig. 7B). The public epitope cluster near S770 was recognized in 43% of COVID-19 patient IgA samples but only 4% of COVID-19 patient IgG samples. In another example, we detected at least 20 distinct epitopes within a stretch of just 46 residues in N 363-408, 10 of which were specific to IgA and 2 of which were specific to IgG (fig. S5D). The positions of several public epitope clusters are shown mapped onto the structure of SARS-CoV-2 in fig. S10.

We also mapped at least 12 distinct epitopes in the SARS-CoV-2 RBD, including 5 in the receptor binding motif that binds ACE2, the human receptor for SARS-CoV-2, and 6 that overlap ACE2 binding sites (Fig. 7, C and D, and fig. S6A). For example, S 414-427 (labeled E2 in Fig. 7) spans residue K417 in the RBD; K417 makes a direct contact with the human ACE2 protein in structures of ACE2 bound to the RBD. Thus, antibodies that recognize E2 are likely to block ACE2 binding and have neutralizing activity (Fig. 7E). Epitope S 454-463 (labeled E6 in Fig. 7) also overlaps ACE2 contact residues and partially overlaps the binding site of the neutralizing antibody CB6, which suggests that antibodies recognizing this epitope also have neutralizing potential (23–25) (Fig. 7G). Several other epitopes also span or are adjacent to critical residues contacted by ACE2 (Fig. 7, F and H). Thus, our data reveal some of the likely binding sites for neutralizing antibodies.

**Discussion**

In this study, we used VirScan to analyze sera from COVID-19 patients and pre-COVID-19 era controls to provide an in-depth serological description of antibody responses to SARS-CoV-2. We mapped the landscape of linear epitopes in the SARS-CoV-2 proteome, characterized their specificity or cross-reactivity, and investigated serological and viral exposure history correlates of COVID-19 severity.

**Identification of SARS-CoV-2 epitopes recognized by COVID-19 patients**

VirScan detected robust antibody responses to SARS-CoV-2 in COVID-19 patients. These were primarily directed against the S and N proteins, with substantial cross-reactivity to SARS-CoV and milder cross-reactivity with the distantly related MERS-CoV and seasonal HCoVs. Cross-reactive responses to SARS-CoV-2 ORF1 were frequently detected in pre–COVID-19 era controls, suggesting that these result from antibodies induced by other pathogens.

At the population level, most SARS-CoV-2 epitopes were recognized by both IgA and IgG antibodies. We found that individuals often exhibited a “checkerboard” pattern, using either IgG or IgA antibodies against a given epitope. This suggests that a given IgM clone often evolves into either an IgG or IgA antibody, potentially influenced by local signals, and that, within an individual, there may often be a largely monoclonal response to a given epitope.

Examination of the humoral response to SARS-CoV-2 at the epitope level using the triple-alanine scanning mutagenesis library revealed 145 epitopes in S, 116 in N, and 562 across the remainder of the SARS-CoV-2 proteome (table S10). Most S epitopes were located on the surface of the protein or within unstructured regions that often abut, but seldom overlap, glycosylation sites (fig. S11). These epitopes ranged from private to highly public, with one region of S (S 811-830) being recognized by 79.9% of COVID-19 patients. Triple-alanine scanning mutagenesis showed highly conserved antibody footprints for some epitope clusters and diverse antibody footprints for others, indicating varying levels of conservation at the antibody-epitope interface among individuals (fig. S8).

Peptides containing public epitopes could be used to isolate and clone antibodies from B cells bearing antigen-specific receptors. If these antibodies are found to lack protective effects or have deleterious effects, these regions could be mutated in future vaccines to divert the immunological response to other regions of S that might have more protective effects. Epitopes also varied in cross-reactivity, which can be explained by the presence or absence of sequence conservation between seasonal HCoVs and SARS-CoV-2 at these regions. Antibodies against several conserved epitopes in HCoVs seemed to be anamnestically boosted in COVID-19 patients. Antibodies recognizing one of these epitopes in the fusion peptide of S2 have been implicated in neutralization, and their presence prior to SARS-CoV-2 infection could mitigate the severity of COVID-19. Collectively, these data help explain why many serological assays for SARS-CoV-2 produce false positives due to preexisting cross-reactive antibodies, some of which may potentially affect the consequences of future SARS-CoV-2 infections.

**Development of SARS-CoV-2 signature peptides for detecting seroconversion by Luminex**

Using machine learning models trained on VirScan data, we developed a classifier that predicts SARS-CoV-2 exposure history with 99% sensitivity and 98% specificity. We identified peptides frequently and specifically recognized by COVID-19 patients and used these to create a Luminex assay that predicts SARS-CoV-2 exposure with 90% sensitivity and 95% specificity. Notably, the Luminex assay requires only three peptides to perform comparably to full-antigen ELISAs and could be further optimized in the future. This highlights the utility of VirScan-based serological profiling in the development of rapid and efficient diagnostic assays based on public epitopes.

**Correlates of severity in COVID-19 patients**

An important goal is to uncover serological correlates of COVID-19 severity. To this end, we compared cohorts of COVID-19 patients...
Fig. 7. High-resolution mapping of SARS-CoV-2 epitopes.

(A) Mapping of antibody epitopes in the SARS-CoV-2 S protein using triple-alanine scanning mutagenesis. Each column of the heatmap corresponds to an amino acid position, and each row represents a COVID-19 patient. The color intensity indicates the average enrichment of three triple-alanine mutant 56-mer peptides containing an alanine mutation at that site, relative to the median enrichment of all mutants of that 56-mer. The upper panel shows the fraction of samples that recognized each region of S as mapped by the IgA 56-mer (gray) versus the IgA and IgG triple-alanine scanning data (blue and red, respectively).

(B and C) Detailed plot of the triple-alanine scanning mutagenesis in (A) to show the epitope complexity within two regions: S 766-835 (B) and S 406-520 (C). The amino acid sequence at each position is shown on the x-axis. In (B), the fusion peptide and predicted S2’ cleavage site are indicated below the sequence (21, 22). In (C) the distinct IgA epitopes identified by the HMM and clustering algorithms are depicted by colored bars. Black dots correspond to ACE2 contact residues in the crystal structure of the RBD receptor complex (6M0J) (23). Epitopes in regions E9 and E10 were not picked up by the HMM classifier because of their short length; however, these regions score in multiple samples and correspond to accessible regions in the crystal structure, which suggests that they may represent true epitopes. (D) Cryo–electron microscopy (cryo-EM) structure of the partially open SARS-CoV-2 S trimer (6VSB) (24), highlighting the locations of the antibody epitopes mapped by triple-alanine scanning mutagenesis. The three S monomers are depicted in tan, green, and gray for the two closed and single open-conformation monomers, respectively. The RBD of the open monomer is shown in light gray. Three of the RBD epitopes from (C) that overlap ACE2 contact residues and are resolved in the cryo-EM structure (E2, E5, and E6) are highlighted in red, purple, and blue, respectively. The locations of additional public epitopes that were mapped in at least 10 samples across the IgG and IgA experiments are depicted in yellow, pink, and cyan. (E to H) The locations of four of the epitope footprints mapped in (C) are shown in relation to the RBD-ACE2 binding interface. The upper image for each panel shows the structure (6M0J) of SARS-CoV-2 RBD (green) in complex with ACE2 (cyan). The E2, E5, E6, and E8 epitopes are highlighted in red, purple, blue, and orange, respectively. Below each structure image is the sequence alignment of the regions of the SARS-CoV-2 and the SARS-CoV S proteins encompassing each epitope. The colored bars indicate the epitope, the black dots indicate residues that directly interact with ACE2 in the crystal structure, and the shaded residues indicate conservation between SARS-CoV-2 and SARS-CoV.
who did (H) or did not (NH) require hospitalization. Using both VirScan and the COVID-19 Luminex assay, we noticed a pronounced and somewhat counterintuitive increase in recognition of peptides derived from the SARS-CoV-2 S and N proteins among the H group, with more extensive epitope spreading. Whether this is a cause or a consequence of severe disease is not clear. Individuals whose innate and adaptive immune responses are not able to quell the infection early may experience a higher viral antigen load, a prolonged period of antibody evolution, and epitope spreading. Consequently, these patients might develop stronger and broader antibody responses to SARS-CoV-2 and could be more likely to have hyperinflammatory reactions such as cytokine storms that increase the probability of hospitalization. We noticed that hospitalized males had more robust antibody responses to SARS-CoV-2 than hospitalized females. This finding may indicate that males in this group are less able to control the virus soon after infection, and it is consistent with reported differences in disease outcomes for males and females (18, 19).

VirScan allowed us to examine viral exposure history, which revealed two notable correlations. First, the seroprevalence of CMV and HSV-1 was much greater in the H group than the NH group. The demographic differences in our relatively small cohort of H versus NH COVID-19 patients make it impossible for us to conclusively determine whether CMV or HSV-1 infection affects disease outcome or is simply associated with other covariates such as age, race, and socioeconomic status. Although CMV prevalence slightly increases with age after 40, it also differs greatly among ethnic and socioeconomic groups (26, 27). CMV is a chronic herpesvirus that is known to have a profound impact on the immune system: it can skew the naïve T-cell repertoire (28) and decrease T and B cell function (29) and is associated with higher systemic levels of inflammatory mediators (30) and increased mortality of people >65 years of age (31). The effects of CMV on the immune system could potentially influence COVID-19 outcomes.

The second notable correlation we observed was a substantial decrease in the levels of antibodies that target ubiquitous viruses such as rhinoviruses, enteroviruses, and influenza viruses in COVID-19 H patients compared with NH patients. When we examined only the CMV+ or HSV-1+ individuals in the two groups, we found that the strength of the antibody response to CMV and HSV-1 peptides was also reduced in the H group. We examined the effects of age on viral antibody levels in a pre–COVID-19 era cohort and found a diminution with age in the antibody response against viral peptides differentially recognized between the H and NH groups, consistent with previous studies on the effects of aging on the immune system (20). This inferred reduced immunity during aging could affect the severity of COVID-19 outcomes.

In correlative analyses such as these, it is difficult to draw strong conclusions about causality, given the demographic differences in the NH versus H groups. The NH group is younger and has a higher percentage of white and female individuals (average age 42, 66% female) than the H group (average age: 58; 42% female) (fig. S2), consistent with well-documented demographic skews in severely affected COVID-19 patients (18, 19). However, even if age and other demographic factors are covariates, CMV seropositivity and age-related reduction in antibody titers against viral antigens, as described here, could still influence the severity of infection. To test these hypotheses, a much larger cohort of COVID-19 patients with severe and mild disease that could be matched for age, race, and sex is required. Such future studies have the potential to enhance our understanding of the biological mechanisms underlying variable outcomes of COVID-19.

Deep serological profiling can provide a window into the breadth of viral responses, how they differ in patients with diverse outcomes, and how past infections may influence present responses to viral infections. Understanding the epitope landscape of SARS-CoV-2, particularly within S, provides a stepping stone to the isolation and functional dissection of both neutralizing antibodies and antibodies that might exacerbate patient outcomes through ADE and could inform the production of improved diagnostics and vaccines for SARS-CoV-2.

Materials and methods

Sources of serum used in this study

Cohort 1

Plasma samples were from volunteers recruited at Brigham and Women’s Hospital who had recovered from a confirmed case of COVID-19. All volunteers had a polymerase chain reaction (PCR)–confirmed diagnosis of COVID-19 before being admitted to the study. Volunteers were invited to donate specimens after recovering from their illness and were required to be symptom free for a minimum of 7 days. Participants provided verbal and/or written informed consent and provided blood specimens for analysis. Clinical data, including date of initial symptom onset, symptom type, date of diagnosis, date of symptom cessation, and severity of symptoms, were recorded for all participants, as were results of COVID-19 molecular testing. Participation in these studies was voluntary, and the study protocols have been approved by the Partners Institutional Review Board.

Cohort 2

Serum samples from patients with PCR-confirmed COVID-19 cases while admitted to the hospital and from patients who were actively enrolled into a prospective study of COVID-19 infection were provided by collaborators from the University of Washington. Residual clinical blood specimens were used. Clinical data, including symptom duration and comorbidities, were extracted from medical records and participant-completed questionnaires. All study procedures have been approved by the University of Washington Institutional Review Board.

Cohort 3

Plasma samples were provided by collaborators from Ragon Institute of MGH, MIT, and Harvard and Massachusetts General Hospital from study participants in three categories: (i) PCR-confirmed COVID-19 cases while admitted to the hospital; (ii) PCR-confirmed SARS-CoV-2–infected cases seen in an ambulatory setting; and (iii) PCR-confirmed COVID-19 cases in their convalescent stage. All study participants provided verbal and/or written informed consent. Basic data on days since symptom onset were recorded for all participants, as were results of COVID-19 molecular testing. Participation in these studies was voluntary, and the study protocols have been approved by the Partners Institutional Review Board.

Cohort 4

Patients were enrolled in the emergency department (ED) at Massachusetts General Hospital in Boston from 15 March to 15 April 2020 during the peak of the COVID-19 surge, with an IRB-approved waiver of informed consent. These included patients 18 years or older with a clinical concern for COVID-19 upon ED arrival and acute respiratory distress with at least one of the following: (i) tachypnea ≥22 breaths per minute, (ii) oxygen saturation ≤92% on room air, (iii) a requirement for supplemental oxygen, or (iv) positive-pressure ventilation. A blood sample was obtained in a 10-ml EDTA tube concurrent with the initial clinical blood draw in the ED. Blood was also drawn on days 3 and 7 if the patient was still hospitalized on those dates. Clinical course was followed to 28 days post-enrollment or until hospital discharge if that occurred after 28 days.

Enrolled individuals who were positive for SARS-CoV-2 were categorized into four outcome groups: (i) requiring mechanical ventilation, with subsequent death; (ii) requiring mechanical ventilation and subsequently recovered; (iii) requiring hospitalization on supplemental oxygen but not mechanical ventilation; and (iv) discharged from ED and not subsequently readmitted with supplemental oxygen. Demographic, past medical, and clinical data were collected and summarized for each outcome group, using medians with interquartile ranges and proportions with 95% confidence intervals, where appropriate.
Cohorts 5 and 6
Longitudinal Hopkins cohort: Remnant serum specimens were collected longitudinally from PCR-confirmed COVID-19 patients seen at Johns Hopkins Hospital. Samples were de-identified before analysis, with linked time since onset of symptom information. Specimens were obtained and used in accordance with an approved IRB protocol.

Cohorts 7 and 8
Cohorts 7 and 8 were from previous studies (7, 8).

Cohort 9
Plasma samples were collected from consenting participants (37 female and 51 male individuals; 18 to 85 years old) of the Partner’s Biobank program at Brigham and Women’s Hospital during the period from July to August 2016. Plasma was harvested after a 10-min 1200xg ficoll density centrifugation from blood tubes (STEMCELL) containing 16 ml of Ficoll 1:2 ratio to facilitate cell separation for other PBMC separation procedures. Plasma was then centrifuged at 1200 rcf for 20 min at 20°C. After aliquoting plasma (1.5-ml conical tubes) into 1.5-ml cryovials and placed on ice during transport, plasma was frozen at −30°C in 1-ml aliquots. All samples were collected with Partners Institutional Review Board approval.

Blood sample collection methods
For cohorts 1 to 3: Blood samples were collected into EDTA (ethylenediamina tetraacetic acid) tubes and spun for 15 min at 2600 rpm according to standard protocol. Plasma was aliquoted into 1.5-ml cryovials and stored at −80°C until analyzed. Only de-identified plasma aliquots including metadata (e.g., days since symptom onset, severity of illness, hospitalization, ICU status, survival) were shared for this study. When appropriate for nonconvalescent samples, plasma or serum was also heat-inactivated at 56°C for survival (appropriate for nonconvalescent samples) or serum was also heat-inactivated at 56°C for 3 hours post–blood draw in a biosafety level 2+ laboratory on site. Whole blood was diluted with room temperature RPMI medium in a 1:2 ratio to facilitate cell separation for other analyses using the SepMate PBMC isolation tubes (STEMCELL) containing 16 ml of Ficoll (GE Healthcare). Diluted whole blood was centrifuged at 1200 rcf for 20 min at 20°C. After centrifugation, plasma (5 ml) was pipetted into 15-ml conical tubes and placed on ice during PBMC separation procedures. Plasma was then centrifuged at 1000 rcf for 5 min at 4°C, pipetted in 1.5-ml aliquots into three cryovials (4.5 ml total), and stored at −80°C. For the current study, samples (200 μl) were first randomly allocated onto a 96-well plate on the basis of disease outcome grouping.

Design and cloning of the SARS-CoV-2 tiling and triple-alanine scanning library
Multiple VirScan libraries were constructed as described below. We created ~200-nt oligos encoding peptide sequences 56 amino acids in length, tiled with 28–amino acid overlap through the proteomes of all coronaviruses known to infect humans, including HCoV-NL63, HCoV-229E, HCoV-OcCoV, HCoV- HKU1, SARS-CoV, MERS-CoV, and SARS-CoV-2, as well as three closely related bat viruses (BatCoV-Rp3, BatCoV-HKU3, and BatCoV-279). For SARS-CoV-2, we included a number of coding variants available in early sequencing of the viruses. For SARS-CoV-2, we additionally made a 20–amino acid peptide library tiling every five amino acids. Additionally, for SARS-CoV-2 we made triple-alanine mutant sequences scanning through all 56-mer peptides. Non-alanine amino acids were mutated to alanine, and alanines were mutated to glycine. Each peptide in all three libraries was encoded in two distinct ways such that there were duplicate peptides that could be distinguished by DNA sequencing. We reverse-translated the peptide sequences into DNA sequences that were codon-optimized for expression in Escherichia coli, that lacked restriction sites used in downstream cloning steps (EcoRI and Xhol), and that were distinct in the 50 nt at the 5’ end to allow for unambiguous mapping of the sequencing results. Then we added adapter sequences to the 5’ and 3’ ends to form the final oligonucleotide sequences (table S1): These adapter sequences facilitated downstream PCR and cloning steps. Different adapters were added to each sublibrary so that they could be amplified separately. The resulting sequences were synthesized on a releasable DNA microarray (Agilent). We PCR-amplified the DNA oligo library with the primers shown below, digested the product with EcoRI and Xhol, and cloned it into the EcoRI/SalI site of the T7FNS2 vector (3). We packaged the resultant library into T7 bacteriophage using the T7 Select Packaging Kit (EMD Millipore) and amplified the library according to the manufacturer’s protocol.

Primers used for analysis of the different libraries employed:
CoV 56-mer library:
5’ adapter: 5′-GAATTCGAGACCTGTTTCCGTGAG-3′
3’ adapter: 5′-CAGGGAAGACGCTCCTGAAGGCGGTGCC-3′
Forward primer: 5′-AATGATACGGCGGGAAATTCCGCCTGCCGCT-3′
Reverse primer: 5′-TGACTATACCTACGGAGGCCTCG-3′
SARS-CoV-2 triple-alanine scanning library:
5’ adapter: 5′-GAATTCGAGACCTGTTTCCGTGAG-3′
3’ adapter: 5′-CAGGGAAGACGCTCCTGAAGGCGGTGCC-3′
Forward primer: 5′-AATGATACGGCGGGAAATTCCGCCTGCCGCT-3′
Reverse primer: 5′-CAAGCAGAAGACTCGAGCAACCGC-3′
SARS-CoV-2 20-mer library:
5’ adapter: 5′-GAATTCGAGACCTGTTTCCGTGAG-3′
3’ adapter: 5′-GAATTCGAGACCTGTTTCCGTGAG-3′
Forward primer: 5′-AATGATACGGCGGGAAATTCCGCCTGCCGCT-3′
Reverse primer: 5′-CAAGCAGAAGACTCGAGCAACCGC-3′
CoV 56-mer library:
5’ adapter: 5′-GAATTCGAGACCTGTTTCCGTGAG-3′
3’ adapter: 5′-CAGGGAAGACGCTCCTGAAGGCGGTGCC-3′
Forward primer: 5′-AATGATACGGCGGGAAATTCCGCCTGCCGCT-3′
Reverse primer: 5′-TGACTATACCTACGGAGGCCTCG-3′

Phage immunoprecipitation and sequencing
We performed phage IP and sequencing as described previously or with slight modifications (5–8). For the IgA and IgG chain isotype-specific IPs, we substituted magnetic protein A and protein G Dynabeads (Invitrogen) with 6 μg of Mouse Anti-Human IgG Fe-CIOT (Southern Biotech) or 4 μg of Goat Anti-Human IgA-BIOT (Southern Biotech) antibodies. We added these antibodies to the phage and serum mixture and incubated the reactions overnight at 4°C. Next, we added 25 or 20 μl of Pierce Streptavidin Magnetic Beads (Thermo-Fisher) to the IgG or IgA reactions, respectively, and incubated the reactions for 4 hours at room temperature, then continued with the washing steps and the remainder of the protocol, as previously described (5–8).

Machine learning classifiers
Gradient-boosting classifier models for the VirScan data were generated using the XGBoost algorithm (version 1.0.2). Classifier models were trained to discriminate either COVID-19+ and COVID-19− patients (n = 232 and 190, respectively) or severe disease and mild disease (n = 101 hospitalized patients and n = 131 nonhospitalized patients). Two models were generated in each case, one using the z-scores for each VirScan peptide from the IgG IP as input features, and the other using the z-scores for each VirScan peptide from the IgA IP as input features. Additionally, a third logistic regression classifier was trained on the output probabilities from the IgG and IgA models to generate a combined prediction. The performance of each of the three model was assessed using a 20-fold cross-validation procedure, whereby predictions for each 5% of the data points were generated from a model trained on the remaining 95%. The SHAP package was used to identify the top discriminatory peptide features from each of the XGBoost models. The logistic regression models for the Luminex data were generated using the scikit-learn python package. The raw median fluorescence intensity (MFI) values were preprocessed using the RobustScalar function, then a logistic regression model was trained using the three most discriminate SARS-CoV-2 peptides. The model performance was quantified by 10-fold cross-validation.

High-resolution epitope identification and clustering
For each position in the 56-mer, the relative enrichment for each amino acid was calculated as the mean fold change of the three mutant peptides containing an alanine mutation.

Shrock et al., Science 370, eabd4250 (2020) 27 November 2020
at that location relative to the median fold change of all alanine mutants for the 56-mer. Overlapping 56-mers were combined by taking the minimum value at each shared position to account for the possibility that an epitope is interrupted in one of the tiles by the peptide junction. To map the boundaries of antibody footprints from the triple-alanine scanning data for each sample we used the Hmmlearn python package to develop a three-state HMM assuming a Gaussian distribution of relative enrichment emissions for each state. Mapped antibody footprints smaller than five amino acids in length were removed from the subsequent analysis. Next, we performed a two-step hierarchical clustering procedure to identify the number of distinct epitopes. First, for each protein all antibody footprints identified across the 169 COVID-19+ patient samples were clustered based on the start and stop locations predicted by the HMM classifier to generate epitope clusters. Next, to identify distinct epitopes, we performed an additional step of hierarchical clustering on the samples with epitopes within each epitope cluster based on the relative enrichment values of the triple-alanine mutants spanning the epitope (fig. S8).

**Similarity-score calculation**

Pairwise alignments were generated for the S proteins of SARS-CoV-2 and each of the four common HCoVs. Similarity scores were calculated separately for a 21-mer amino acid window at each position of the SARS-CoV-2 S protein. The mean similarity score between SARS-CoV-2 and the corresponding sequence of the other HCoV was calculated for each window using the BLOSUM62 substitution matrix. Sequence alignments of the common HCoVs were then used to map the boundaries to identify the number of distinct epitopes.

**ELISA serology assays**

ELISAs were performed separately using the SARS-CoV-2 N protein, S protein, or the S receptor-binding domain (RBD). 96-well plates were coated with antigen overnight. The plates were then blocked in PBS + 3% BSA. After washing with PBS + 0.05% Tween-20, the sample was incubated at 4°C for 1 hour. Bound antigen was quantified by measuring the OD405, and the reported values were calculated as the fold change over the pre-COVID-19 controls.

**REFERENCES AND NOTES**

3. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU); https://coronavirus.jhu.edu/map.html
ACKNOWLEDGMENTS

We thank J. Bensko, D. Gakpo, G. DeGregorio, and S. Fisch (Brigham and Women’s Hospital) for their expertise and instrument availability that supported this work. This work was also supported by grants from the National Institutes of Health (NIH AI121394 and AI138318, AR070253, and AI101387/Z/16/Z). S.J.E., B.D.W., and H.B.L. are investigators of the Howard Hughes Medical Institute. S.J.E. serves on the scientific advisory board of Homology Medicines, TSCAN Therapeutics, MAZE, MAZE Therapeutics and Mirimus. S.J.E. is a founder of H.Y.C., N.H., M.B.G., B.D.W., and H.B.L. A.S., T.K. are founders of TSCAN Therapeutics. S.J.E. is a founder of Ragon Institute of MGH, MIT and Harvard. S.J.E. and B.D.W. are Investigators with the Howard Hughes Medical Institute. S.J.E. and B.D.W. are authors of the VirScan library to identify pathogen antibodies included in the article that is credited to a third party; obtain authorization from the rights holder before using such material.

Inflammatory Diseases, Boston, MA, USA. 5Massachusetts General Hospital, Boston, MA, USA. 6Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. 7Inventors of VirScan (AI138331). 8Inventors of VirScan (AI070253 and AI070495). 9Inventors of VirScan (AI242244 and AI138331). 10Inventors of VirScan (AI138318, AR070253, and AI101387/Z/16/Z). 11Inventors of VirScan (AI138318, AR070253, and AI101387/Z/16/Z). 12Inventors of VirScan (AI138318, AR070253, and AI101387/Z/16/Z).

Copyright © 2020 the Authors. Published by The American Society for Cell Biology. This is an open access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited and no changes are made to the licensed material. This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original work is properly cited. See https://creativecommons.org/licenses/by/4.0/ for more information.

Supplemental methods

See https://science.sciencemag.org/content/370/6520/eabd4250/suppl/DC1 for supplemental methods, tables and figures. doi: 10.1126/science.abd4250


REFERENCES

1Department of Emergency Medicine, Massachusetts General Hospital, Boston, MA, USA. 2Massachusetts General Hospital Cancer Center, Boston, MA, USA. 3Division of Infectious Diseases, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. 4Massachusetts General Hospital Center for Immunology and Inflammatory Diseases, Boston, MA, USA. 5Massachusetts General Hospital, Boston, MA, USA. 6Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. 7Inventors of VirScan (AI138331). 8Inventors of VirScan (AI070253 and AI070495). 9Inventors of VirScan (AI242244 and AI138331). 10Inventors of VirScan (AI138318, AR070253, and AI101387/Z/16/Z). 11Inventors of VirScan (AI138318, AR070253, and AI101387/Z/16/Z). 12Inventors of VirScan (AI138318, AR070253, and AI101387/Z/16/Z).
Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity


Science 370 (6520), eabd4250.
DOI: 10.1126/science.abd4250 Originally published online September 29, 2020

Profiling coronaviruses

Among the coronaviruses that infect humans, four cause mild common colds, whereas three others, including the currently circulating severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), result in severe infections. Shrock et al. used a technology known as VirScan to probe the antibody repertoires of hundreds of coronavirus disease 2019 (COVID-19) patients and pre-COVID-19 era controls. They identified hundreds of antibody targets, including several antibody epitopes shared by the mild and severe coronaviruses and many specific to SARS-CoV-2. A machine-learning model accurately classified patients infected with SARS-CoV-2 and guided the design of an assay for rapid SARS-CoV-2 antibody detection. The study also looked at how the antibody response and viral exposure history differ in patients with diverging outcomes, which could inform the production of improved vaccine and antibody therapies.

Science, this issue p. eabd4250