Article

Selection for immune evasion in SARS-CoV-2 revealed by high-resolution epitope mapping and sequence analysis

Arnaud
N'Guessan,
Senthilkumar
Kailasam, Fatima
Mostefai, ..., B.
Jesse Shapiro,
Jörg H. Fritz,
Ciriaco A. Piccirillo

jorg.fritz@mcgill.ca (J.H.F.)
ciro.piccirillo@mcgill.ca
(C.A.P.)

Highlights

High-density peptide
arrays provide high
resolution antibody
epitope maps

Seasonal coronavirus
cross-reactivity drives
antibody responses to
SARS-CoV-2 epitopes

Point mutations and
natural selection in
epitopes occur at higher
rates upon transmission

iScience

Article

Selection for immune evasion in SARS-CoV-2 revealed by high-resolution epitope mapping and sequence analysis

Arnaud N'Guessan,[1,2,13] Senthilkumar Kailasam,[3,4,12,13] Fatima Mostefai,[5,6] Raphaël Poujol,[5] Jean-Christophe Grenier,[5] Nailya Ismailova,[1,10,12] Paola Contini,[7] Raffaele De Palma,[7] Carsten Haber,[8] Volker Stadler,[8] Guillaume Bourque,[3,4,12] Julie G. Hussin,[5,9] B. Jesse Shapiro,[1,2,12] Jörg H. Fritz,[1,10,12,14,15,*] and Ciriaco A. Piccirillo[1,10,11,12,14,*]

SUMMARY

Here, we exploit a deep serological profiling strategy coupled with an integrated, computational framework for the analysis of SARS-CoV-2 humoral immune responses. Applying a high-density peptide array (HDPA) spanning the entire proteomes of SARS-CoV-2 and endemic human coronaviruses allowed identification of B cell epitopes and relate them to their evolutionary and structural properties. We identify hotspots of pre-existing immunity and identify cross-reactive epitopes that contribute to increasing the overall humoral immune response to SARS-CoV-2. Using a public dataset of over 38,000 viral genomes from the early phase of the pandemic, capturing both inter- and within-host genetic viral diversity, we determined the evolutionary profile of epitopes and the differences across proteins, waves, and SARS-CoV-2 variants. Lastly, we show that mutations in spike and nucleocapsid epitopes are under stronger selection between than within patients, suggesting that most of the selective pressure for immune evasion occurs upon transmission between hosts.

INTRODUCTION

Coronaviruses constitute a large family of enveloped, positive-sense single-stranded RNA viruses that cause frequent diseases in birds and mammals. The *Coronaviridae* family includes four species that are endemic in the human population (hCoVs): the alpha-coronaviruses that include hCoV-229E and hCoV-NL63 and beta-coronavirus species that include hCoV-HKU1 and hCoV-OC43, and are usually associated with mild, self-limiting upper respiratory tract infections, although they can cause severe illness in immuno-compromised patients.[1] Three other beta-coronavirus species have recently emerged: Middle East respiratory syndrome-CoV (MERS-CoV), SARS-CoV-1, and SARS-CoV-2, all causing severe disease in humans.[2,3] Severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2) is a novel virus belonging to the *Coronaviridae* family that emerged in late 2019 and quickly spread throughout the world, causing a pandemic with morbidity, mortality, and economic disruption on a global scale with few precedents.[3] The clinical course of COVID-19 is highly variable: some infected individuals are completely asymptomatic,[4] while others experience a spectrum of clinical manifestations including fever, severe respiratory distress, pneumonia, diarrhea, blood clotting disorders, increased systemic cytokine release and, in <5% of cases, prolonged hospitalization and death.[5] In addition to factors like viral exposure history, viral inoculum at infection, and the genetic background of the individual, the severity of COVID-19 and the response to treatment is also heavily influenced by other factors like sex, advanced age, ethnicity, and comorbidities such as cardiovascular disease, chronic lung disease, obesity, diabetes, and compromised immune function.[6–8] An in-depth understanding of the immune response to SARS-CoV-2, particularly humoral, could improve our understanding of the diverse courses of disease and better guide the development of improved diagnostics and vaccines.

SARS-CoV-2 infection can elicit robust antibody responses in humans, and this response represents the primary focus of global efforts to develop accurate serology-based diagnostics and vaccination strategies against infection.[9,10] Pre-existing cross-reactive immune responses to SARS-CoV-2 proteins in

[1]Department of Microbiology and Immunology, McGill University, Montréal, QC, Canada

[2]McGill Genome Centre, McGill University, Montréal, QC, Canada

[3]Canadian Center for Computational Genomics, Montréal, QC, Canada

[4]Department of Human Genetics, McGill University, Montréal, QC, Canada

[5]Research Centre, Montreal Heart Institute, Montreal, QC, Canada

[6]Département de Biochimie et Médecine Moléculaire, Université de Montréal, Montréal, QC, Canada

[7]Department of Internal Medicine, University of Genoa and IRCCS IST-Ospedale San Martino, Genoa, Italy

[8]PEPperPRINT GmbH, Heidelberg, Germany

[9]Département de Médecine, Université de Montréal, Montréal, QC, Canada

[10]McGill University Research Center on Complex Traits (MRCCT), McGill University, Montréal, QC, Canada

[11]Infectious Diseases and Immunity in Global Health Program of the Research Institute of McGill Health Center, Montréal, QC, Canada

[12]Dahdaleh Institute of Genomic Medicine (DIgM), McGill University, Montréal, QC, Canada

[13]These authors contributed equally

[14]These authors contributed equally

[15]Lead contact

Continued

infection-naive subjects have been described by several studies.[11] Notably, regions of high homology between SARS-CoV-2 and endemic hCoVs have been highlighted as a likely source of this cross-reactivity, although the role of such cross-reactive responses in the outcome of SARS-CoV-2 infection and vaccination is currently not fully understood (reviewed in[11]).

Cross-reactive B cells generated to epitopes of low homology may interfere with responses to a secondary heterologous infection by skewing the response to antibodies with higher avidity for the initial infectious agent, reducing the generation of *de novo* antibody responses. Evidence for such cross-reactive mechanisms was reported in common cold coronaviruses as well as SARS-CoV-2.[12–14] Beneficial cross-reactive memory B cells and antibodies can also be generated to epitopes of high similarity between a primary and a heterologous secondary infection. Cross-reactive B cells specific to highly conserved epitopes might then go on to produce an enhanced memory-like response to heterologous infection, including the production of cross-neutralizing antibodies that can prevent viral entry to cells. Evidence for cross-reactive B cell responses was reported in coronavirus infections, including beta-coronaviruses and SARS-CoV-2.[15,16] Age and disease severity have been suggested to be a contributing factor for these observations.[11] However, none of these studies directly explored effects of pre-existing cross-reactive antibodies on COVID-19 severity. As such, the causal relationship between cross-reactivity of hCoVs and SARS-CoV-2 and fatal outcomes remains unclear.

Cross-reactivity can also impact the T cell compartment. Cross-reactive T cells generated to epitopes of relatively low homology may detrimentally impact a secondary infection with heterologous agents by dominating the response to the secondary infection, disrupting the development of high-avidity *de novo* T cell responses, thereby leading to immunopathology and reduced viral clearance. Evidence for these detrimental cross-reactive mechanisms has been reported in studies of flavivirus infection,[17,18] but to our knowledge not in coronavirus infections. Beneficial cross-reactive T cell responses can be generated by the stimulation of T cells through high-homology epitopes. These T cells then cross-react with high avidity during a heterologous secondary infection and might block pathogenicity by preventing invasive infection or expedite the rate of viral clearance by forming a 'secondary-like' memory immune response with an increased magnitude of B and T cell responses. Evidence for such cross-reactive mechanisms was reported in studies of coronavirus infections, including SARS-CoV-2.[19–24]

Overall, there is no evidence from longitudinal cohort studies or population-level studies to suggest that pre-existing SARS-CoV-2 cross-reactive antibodies exacerbate COVID-19.[11,25] Rather, current evidence suggests that alongside the *de novo* immune response, cross-reactive T cells and antibodies form part of the protective immune response to SARS-CoV-2 infection. Considering the high propensity for SARS-CoV-2 to mutate viral proteins, notably in S protein, variants of concern (VOCs) and variants under investigation (VUIs) can acquire properties for increased transmissibility, disease severity, and/or immune evasion.[26] Thus, promoting this cross-reactive, pre-existing memory immune response to common hCoVs may be an effective strategy against SARS-CoV-2 and future VOCs.[27,28]

To better understand the molecular determinants underlying protective immunity to pathogens, including viruses, one must define the epitopes in various viral proteins, the minimal unit of an antigen that can be recognized by T and B cells and can elicit potent cellular and humoral immune responses, respectively. A recent study used VirScan technology, a high-throughput, programmable phage-display immunoprecipitation and sequencing (PhIP-Seq) method,[29] to analyze epitopes of antiviral antibodies in sera of COVID-19 patients relative to pre–COVID-19 sera controls.[30] However, the nature and dynamics of the peptide pools of VirScan/PHIP-seq may limit the resolution, sensitivity, and breadth of specific epitope detection in infected individuals, in turn, providing a fragmented view of the complete footprint of epitope recognition by antibodies.[29,31]

In the current study, we provide a comprehensive analysis of SARS-CoV-2 humoral immune responses in a dataset of symptomatic or recovered COVID-19-positive and COVID-19-negative patients. We exploited a high-density peptide array (HDPA) by spotting overlapping 15-mer peptides derived from the entire SARS-CoV-2 and hCoVs proteomes to rapidly identify B cell epitopes recognized by distinct antibody isotypes in patients' blood sera of individual patient groups. We then subjected our data to an integrated computational pipeline to evaluate the fine immunological properties of detected SARS-CoV-2 epitopes and relate them to their evolutionary and structural characteristics. We show that while some epitopes are common

*Correspondence:
jorg.fritz@mcgill.ca (J.H.F.),
ciro.piccirillo@mcgill.ca
(C.A.P.)
https://doi.org/10.1016/j.isci.
2023.107394
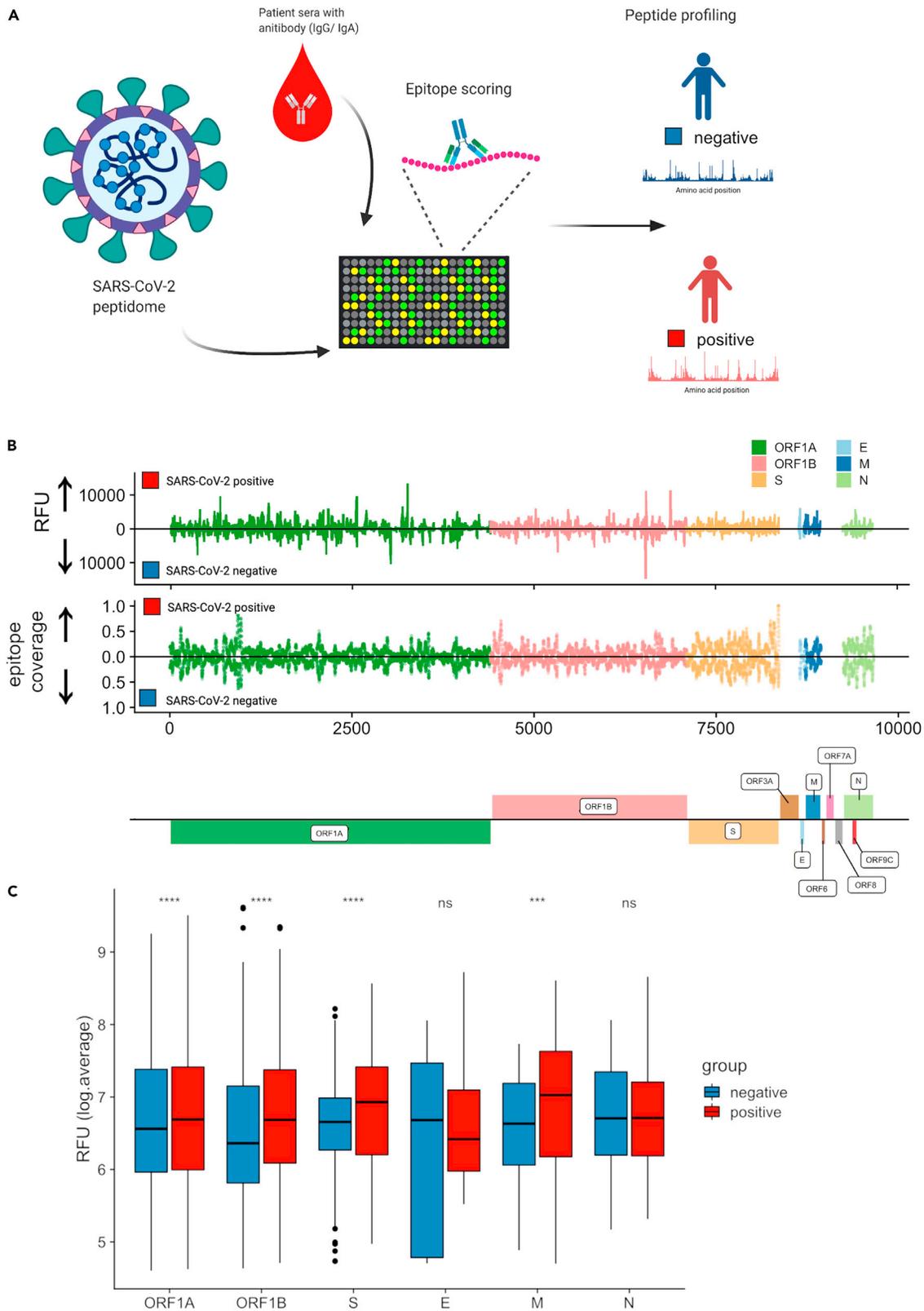
Figure 1. High-density peptide arrays (HDPA) provide high-resolution antibody epitope maps across the SARS-CoV-2 proteome

(A) Overview of analytical pipeline. The proteome of SARS-CoV-2 was translated into 15-mer overlapping peptides with a peptide-to-peptide overlap of 13 amino acids. The resulting individual peptides were printed in duplicates on the microarray. Sera from confirmed SARS-CoV-2-positive and -negative individuals were incubated on PEPperCHIP the HDPA. Serum antibody binding was visualized using respective fluorescently labeled secondary antibodies (anti-human IgG and anti-human IgA). Image acquisition and data quantification resulted in epitope-specific antibody profiles for SARS-CoV-2.

(B) Average relative fluorescent units (RFU) profiles and peptide coverages are plotted across the SARS-CoV-2 proteome (ORF1A, ORF1B, Spike (S) protein, Envelope (E) protein, Membrane (M) glycoprotein, Nucleocapsid (N) phosphoprotein). Antibody responses to each linear 15-mer peptide were mapped across the SARS-CoV-2 proteome and average RFU calculated for each amino acid residue. The normalized positional 'epitope coverage' at each protein residue location is defined as the ratio of total peptides mapped to each position by the total expected peptides (see STAR methods section). 'Hotspots' can be seen as spiked in the RFU or coverage distributions.

(C) Comparison of mean RFU (log-scale) between SARS-CoV-2-positive and -negative sample groups for each viral protein. (unpaired t-test, ns: p > 0.05; *: p ≤ 0.05; **: p ≤ 0.01; ***: p ≤ 0.001; ****: p ≤ 0.0001).

(*public epitopes*) across all studied hCoVs (including SARS-CoV-2), others are unique (*private epitopes*) to a specific hCoV. Then, to highlight epitopes that have an important role for protecting against SARS-CoV-2 when an individual gets infected, we defined differential epitopes as epitope for which the response is at least two-times higher in COVID-19-positive than COVID-19-negative individuals. We also highlight hot-spots of pre-existing immunity and a subset of cross-reactive epitopes that contributes to increasing the average humoral immune response to SARS-CoV-2. Finally, using a dataset of over 38,000 publicly available genome sequences, collected during the first two waves of the pandemic, we tracked single nucleotide variants (SNVs) within and between COVID-19 patients and found evidence for positive selection on nonsynonymous mutations in epitopes. Selection is stronger between than within patients, indicating that selection for immune evasion occurs mostly upon transmission between hosts. Overall, our results have implications for future genomic surveillance and vaccine design.

RESULTS

Antibody fingerprinting with high-density peptide arrays provides a high-resolution antibody epitope map across the SARS-CoV-2 proteome

Most previously reported high-resolution SARS-CoV-2 B cell epitope mapping strategies relied on VirScan/PHIP-seq methodology.[30,32] However, the nature and dynamics of the peptide pools of VirScan/PHIP-seq limit the resolution, sensitivity, and breadth of specific epitope detection in infected individuals, providing a fragmented view of the complete footprint of epitope recognition by antibodies.[29,31] To assess the humoral immune response against SARS-CoV-2 at the epitope level, we used a HDPA technology to define virus protein-specific B cell epitopes and potential antigenic hotspots for antibody reactivity. A high-resolution linear epitope map across the entire SARS-CoV-2 proteome was achieved using the PEPperCHIP SARS-CoV-2 proteome microarray technology (Figure 1A).[33] We performed this assay on sera obtained from ten SARS-CoV-2-positive individuals (asymptomatic and recovered) and five SARS-CoV-2-negative, control subjects (SARS-CoV-2-negative) (Table S1). The degree of immune reactivity to spike protein (S), envelope protein (E), membrane glycoprotein (M), nucleocapsid phosphoprotein (N) and ORF1AB was measured in relative fluorescence units (RFUs). Linear overlapping peptides of 15 amino acid length were used for each protein and a dual isotype analysis, determining IgG- and IgA-specific antibody responses, was performed (Figure 1A). This was followed by a comprehensive analysis workflow to characterize the differential epitopes, their structural properties and utilize genome sequence analysis of arising SARS-CoV-2 variants to assess immune evasion potential.

Sera from SARS-CoV-2-positive individuals yielded strong immune reactivity (measured in RFU) in S and N proteins, as well as in select regions of ORF1AB (Figure 1B). Antibody responses were also identified in samples from SARS-CoV-2-negative individuals (Figure 1B), suggesting that HDPA technology is well suited to detect epitopes of pre-existing immune responses conferred through prior infections with hCoVs. Although sera from SARS-CoV-2-positive and -negative individuals had very similar epitope coverage per amino acid, some regions in the SARS-CoV-2 proteome were more immuno-dominant than others, exerting higher RFU values (Figure 1B). Except for M protein, all proteins analyzed had stronger antibody responses to more unique peptides in the SARS-CoV-2-positive patient group (Figure S1, Tables 1, and S2). In addition, the mean RFU values of SARS-CoV-2-positive sera were higher toward most regions of the SARS-CoV-2 proteome than in the SARS-CoV-2-negative group, demonstrating the elicitation of robust antibody responses to immuno-dominant epitopes upon SARS-CoV-2 infection (Figure 1C). For further analysis of epitopes with greatest immuno-dominance, only peptide RFU values greater than or equal to

Table 1. Numbers of SARS-CoV-2-specific epitope-defining peptides identified by HDPA

SARS-CoV2-specific Peptides	Spike	Nucleocapsid	Envelope	Membrane	ORF1ab	TOTAL
SARS-CoV-2 negative	119	41	1	47	294	502
SARS-CoV-2 positive	195	69	6	29	549	848
Overlap	90	35	6	17	353	501
Total	404	145	13	93	1196	1851

Number of SARS-CoV-2-specific epitope-defining peptides identified with high density peptide arrays (HDPA) in spike (S) protein, envelope (E) protein, membrane (M) glycoprotein, nucleocapsid (N) phosphoprotein, and ORF1ab. Number of unique peptides that showed a significant antibody response (RFU ≥ 1000) in SARS-CoV-2-negative and SARS-CoV-2-positive groups are depicted. Some peptides are present in both groups, referred to as overlap.

1000 were used in our further analysis. Taken together, our results demonstrate that the applied HDPA approach allows highly sensitive detection of a large pool of epitopes across the SARS-CoV-2 proteome.

Structural features of identified epitopes and comparison with computationally predicted epitopes

An epitope is the minimal unit of an antigen that can be recognized by T and B cells and can elicit potent cellular and humoral immune responses, respectively. B cell epitopes can be divided into two major categories, namely linear and conformational epitopes. In a linear epitope, a stretch of continuous amino acids forms the antibody binding site, while amino acid residues that are brought together by protein folding form conformational epitopes. In our current study, antibody responses are detected using linear peptide arrays, and thus these epitopes are primarily linear in nature, although some linear epitopes contributing to conformational components of a protein may also be detected. Though there is a significant interest for short linear epitopes in vaccine design, most of the current SARS-CoV-2 vaccine immunogens are structural S proteins.[34] Thus, we asked whether the short peptide-based approach in the applied HDPA approach can also reveal conformational epitope sites, as has been recently suggested.[30] Using 3D structures and biophysical properties of the SARS-CoV-2 proteome, we applied the DiscoTope algorithm[35] to computationally predict conformational B cell epitopes as well as the BepiPred algorithm[36] to obtain linear B cell epitopes. We then compared these to the epitope sites identified in our HDPA experiment. Apart from E and M proteins, we observed significant overlap of experimentally identified epitopes with predicted epitopes (Figure 2A), with approximately 38% of the total proteome being part of the amino acid residues contributing to the B cell epitome of SARS-CoV-2 recognized in infected individuals. We observed overlap of mapped epitope sites with conformational epitopes predicted by DiscoTope, suggesting that the applied HDPA approach also identifies a considerable number of conformational epitopes.

It is important for epitopes to be solvent exposed to allow their amino acid side chains to interact with the antibody. In our study, there was no significant difference in the average normalized solvent accessibility (SASA) of epitopes compared to non-epitope regions (mean SASA = 5.3 Å^2). However, residues in highly conserved epitope sites have lower solvent accessibility (Figures S2A–S2C), suggesting that many identified epitopes might only get exposed through proteolysis or conformational changes throughout the infectious cycle or stage of anti-viral immune response. This implies that our linear peptide-based HDPA approach can capture more epitopes compared to those that use full-length antigens or protein domains to study immune profiling of antibody responses.[37,38] Using structural models we mapped the epitopes of S protein of SARS-CoV-2 identified by the HDPA approach, which revealed that many of the epitope sites identified in the S protein are in the N terminal domain (NTD) and the receptor binding domain (RBD) (Figures 2B and 2C). Interestingly, most epitope sites identified in the NTD and RBD have low conservation scores, while many other epitope sites identified in the S protein had high conservation scores (Figures 3A and 3B). In addition, HDPA analysis revealed strong antibody immunoreactivity in a few epitope sites of E (Figure 3C), M (Figure 3D), N proteins (Figure 3E), as well as ORF1A (Figure S3) and ORF1B (Figure S4).

Cryo-EM studies have indicated that the SARS-CoV-2 spike protein is highly flexible and exhibits several prefusion conformations where three RBDs adopt distinct orientations: "up" (receptor-accessible state) and "down" (receptor-inaccessible). Protomers with the up-conformation can facilitate the binding between the spike protein and the angiotensin-converting enzyme 2 (ACE2) receptor, thereby allowing
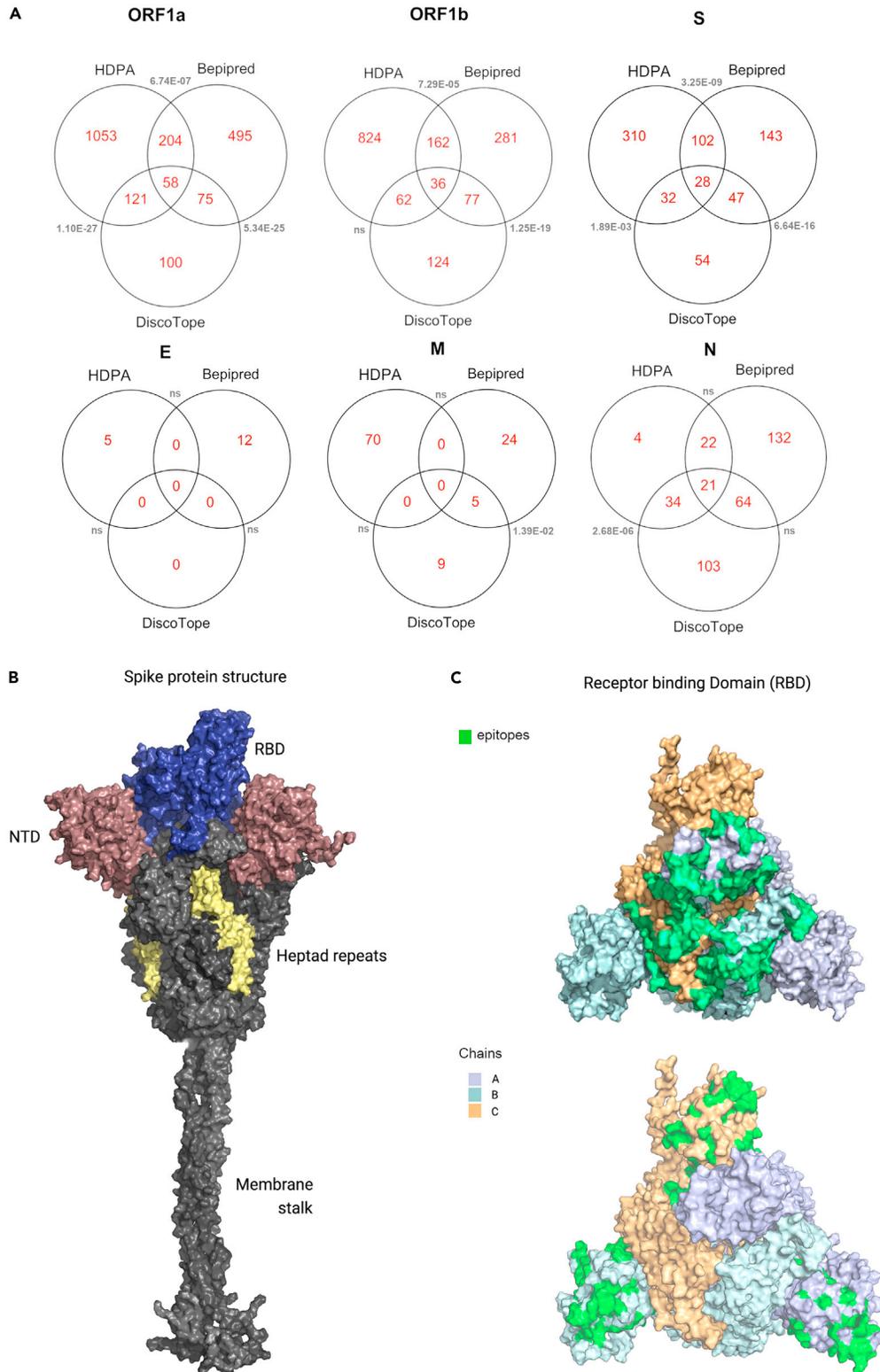
Figure 2. Structural features of identified epitopes and comparison with computationally predicted epitopes

(A) Venn diagrams comparing epitope sites identified by HDPA with computational approaches including BepiPred and DiscoTope. Overlap hypergeometric test (R package *phyper*) for significance between epitope sites from HDPA and computational prediction methods; p value (<0.10) shown in gray.

Figure 2. *Continued*

(B) Three-dimensional structural model of the full-length spike protein trimer in an open conformation with domains labeled; receptor binding domain (RBD); N-terminal domain (NTD).

(C) Three-dimensional model of RBD and NTD highlighting epitope sites (green) identified by HDPA analysis on the surface of the RBD and NTD domains.

and facilitating host cell infection.[39] Analysis of structures of distinct SARS-CoV-2 human neutralizing antibodies (NAbs) in complex with the SARS-CoV-2 spike trimer or RBD revealed structural correlates of SARS-CoV-2 neutralization and allowed classification of antibodies into categories.[40] Focusing on RBD-binding a distinction into four distinct classes has been proposed: Class 1 - NAbs that block ACE2 and bind only to "up" RBDs; Class 2 - ACE2-blocking NAbs that bind both "up" and "down" RBDs; Class 3 - NAbs that bind outside the ACE2 site and recognize both "up" and "down" RBDs; and Class 4 – NAbs that do not block ACE2 and bind only to "up" RBDs.[40] To analyze the extent with which HDPA epitope sites overlap with known NAb-binding residues in SARS-CoV-2 RBD we aligned the epitopes sites identified by HDPA with NAb-binding residue sites mapped by Cryo-EM studies 40. Epitopes found by HDPA generally tend to agree with cryo-EM, particularly in more conserved regions of the protein sequence (Figure 4A). More than 60% of RBD residues in NAb-binding sites are identified by HDPA analysis, mapping epitope sites in all four distinct structural correlate classes of SARS-CoV-2 RBD-binding NAbs identified by cryo-EM (Figure 4B). Taken together, our results demonstrate that the applied HDPA profiling strategy can identify a set of linear and conformational B cell epitopes unique in sera of SARS-CoV-2-infected individuals. In addition, a large portion of the identified epitope sites are residues in previously reported NAb-binding sites, demonstrating that epitope sites mapped by HDPA analysis are of functional relevance in antibody-mediated immunity to SARS-CoV-2.

Cross-reactivity to endemic seasonal human coronaviruses is a significant driver of antibody responses to SARS-CoV-2 epitopes

In addition to the zoonotic pathogens SARS-CoV, MERS-CoV, and SARS-CoV-2, four other low-pathogenicity hCoVs are endemic and co-circulating in the human population[41]: strains OC43 and HKU1 (beta-CoVs like SARS-CoV-2), and NL63 & 229E (alpha-CoVs), of which OC43 and 229E are the most common, accounting for 5–30% of common colds.[42] Notably, structural proteins of SARS-CoV-2 show some degree of amino acid sequence identity with hCoVs.[43,44] One prevailing view in our understanding of COVID-19 immunopathogenesis is that an underlying immune response toward endemic hCoVs is a hallmark feature of SARS-CoV-2-infected asymptomatic individuals.[11] This pre-existing immunity is hypothesized to partially control viral replication and eliminate infected cells resulting in less severe pathology and inflammation.[11,12,45–48]

Having established the link between structure accessibility and protein conservation (Figures 3B and S2), we next asked how conservation is related to the humoral immune response. One way protein conservation could affect adaptive immunity is through cross-reactivity to related viruses. To this end, we first analyzed the humoral immune response against the four hCoVs OC43, HKU1, NL63 and 229E at the epitope level using HDPA on sera from the same ten SARS-CoV-2-positive (asymptomatic or recovered) patients and five control subjects (SARS-CoV-2-negative; Table S1). HDPA yielded strong antibody reactivities to many distinct sites across the proteomes of all hCoVs (Tables 2, S3, S4, S5, and S6). We then defined cross-reactive epitopes based on the conservation of peptide residues across hCoVs and the presence of an immune response to epitopes in SARS-CoV-2 and at least of one of the endemic hCoVs (HKU1, NL63, OC43 or 229E; Figure 5A). To evaluate the conservation of peptide sequences across hCoVs, we aligned the protein sequences of these viruses and calculated a conservation score, reflecting the conservation of physicochemical properties in the alignment where identical residues score the highest.[49] We also defined cross-reactivity at the level of epitope sites (single amino acids) to account for the possibility that a particular amino acid site within a 15-mers epitopes is associated with cross-reactive immunity. Epitope sites with a conservation score $\geq$ 6 and for which we obtained antibody reactivity for both SARS-CoV-2 and at least one of the hCoVs were considered as cross-reactive epitope sites. These cross-reactive epitope sites represent 27.2% of the pool of detected epitope sites by the applied HDPA assay (Table S7). We also carried out local alignment of the peptides from the HDPA (with RFU $\geq$ 1000) of all five viral strains to the SARS-CoV-2 proteome to evaluate the cross-reactivity profile of SARS-CoV-2 epitopes and identified hotspots of conserved epitopes (example for S protein in Figure 5B).

A

Spike protein

NTD
RBD

RFU Ratio
2-5
5-10
10-15
16-20
>30
NA

B

Spike protein

Conservation Score
9
8
7
6
5
4
3
2
1
NA

C

Envelope protein

RFU Ratio
2-5
5-10
10-15
>30
NA

D

Membrane protein

RFU Ratio
2-5
5-10
10-15
>30
NA

E

Nucleocapsid protein

RFU Ratio
2-5
5-10
10-15
>30
NA
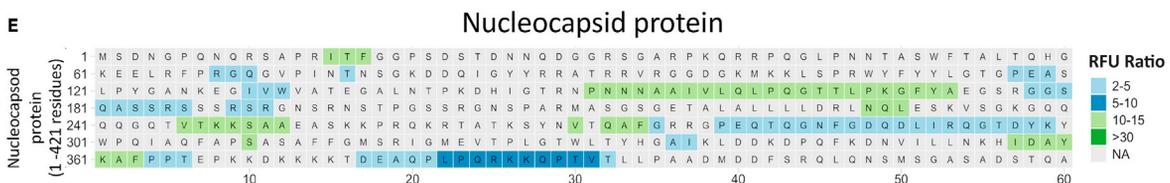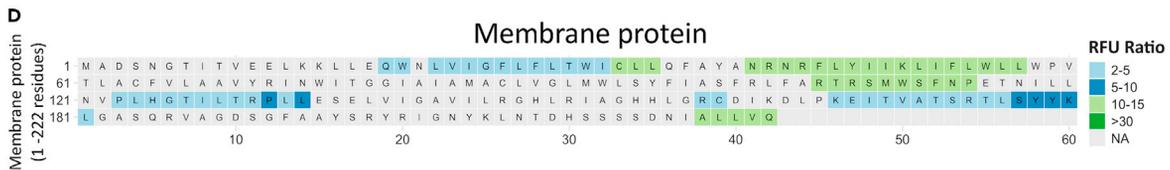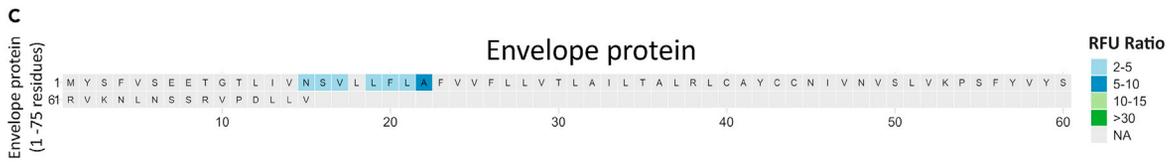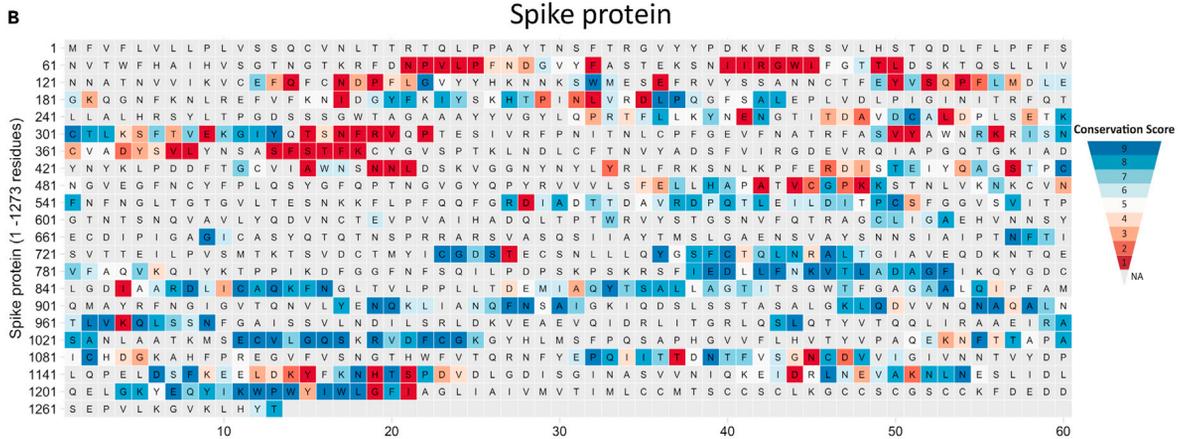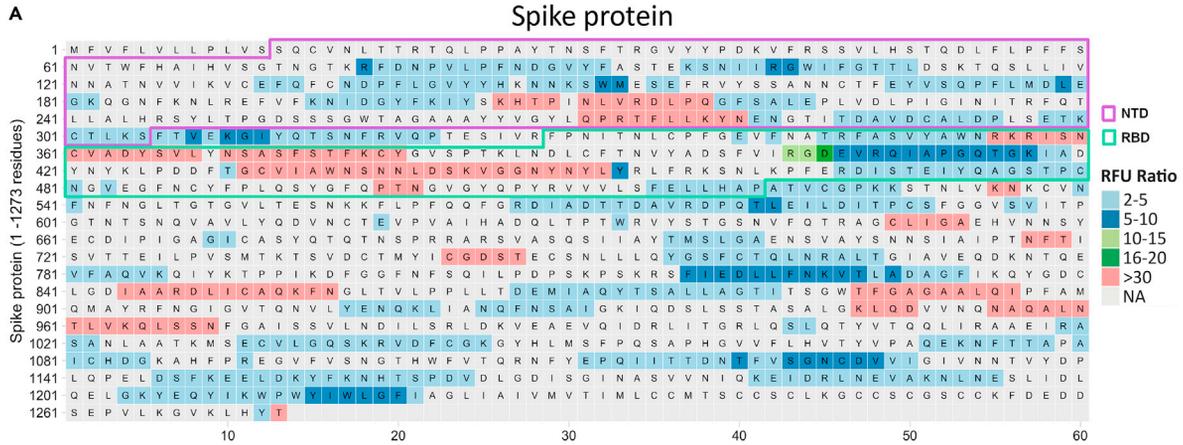
Figure 3. Identified differential epitope sites in structural proteins of SARS-CoV-2

Relative Fluorescence Unit (RFU) values of HDPA analysis were used to calculate ratio values to define differential epitope sites and are color coded on the distinct SARS-CoV-2 proteins. Residues that are not part of epitopes are marked in gray (NA, not applicable).

(A) RFU values of differential epitope sites identified in SARS-CoV-2 Spike protein. N-terminal domain (NTD) and receptor-binding domain (RBD) are highlighted.

(B) Conservation scores of physicochemical properties of the SARS-CoV-2 Spike protein.

(C) RFU values of differential epitope sites identified in SARS-CoV-2 Envelope protein.

(D) RFU values of differential epitope sites identified in SARS-CoV-2 Membrane protein.

(E) RFU values of differential epitope sites identified in SARS-CoV-2 Nucleocapsid protein.

Next, to highlight cross-reactive epitope sites that are particularly important for the humoral immune response after exposure to SARS-CoV-2, we focused on differential cross-reactive epitope sites that give an antibody response signal in sera of SARS-CoV-2-positive over SARS-CoV-2-negative individuals. Although we analyzed sera from a smaller sized cohort compared to a recent study using PhIP-Seq,[30] we were nonetheless able to sensitively detect more differential cross-reactive epitope sites discriminating virus exposed from non-exposed individuals (Figure 5C). The increased sensitivity of HDPA over PhIP-Seq analysis for identifying cross-reactive epitope sites was further highlighted when performing a sensitivity analysis on the number of cross-reactive epitope sites that define a cross-reactive epitope (Figure S5A).

We next analyzed if the humoral immune response to SARS-CoV-2 epitopes correlated with the number of cross-reactive epitopes identified. In other words, to what extent is the response to SARS-CoV-2 predict-able based on cross-reactivity to other endemic hCoVs? We defined cross-reactive epitopes as peptide sequences with at least five cross-reactive epitope sites. We found a positive correlation between the average humoral immune response to SARS-CoV-2 epitopes and the number of cross-reactive epitopes per patient in a recently published PhIP-Seq dataset[30] (Figure 5D) and in our HDPA dataset (Figure S5). We detected a stronger positive correlation between the average antibody response to SARS-CoV-2 epitopes and the number of cross-reactive epitopes in SARS-CoV-2-positive compared to negative patients (Figure 5D; correlation coefficients: 4.42e-3 vs. 1.54e-3; ANOVA p for *Covid*19 *status* = 7.43e-10). This positive correlation is robust to the threshold number of cross-reactive epitope sites defining a cross-reactive epitope and is also replicated in our HDPA dataset (Figure S5).

Leveraging the sample size in this same PhIP-Seq dataset, we aimed to identify the subset of cross-reactive epitopes with the greatest contribution in humoral immunity in SARS-CoV-2-positive patients using the IndVal test, which is a non-parametric test to identify significant associations from presence-absence data.[50] This test allowed us to identify 75 epitopes that are significantly associated with SARS-CoV-2-positive over SARS-CoV-2-negative samples, with 16 out of these 75 epitopes (21.3%) being cross-reactive (Table S8.). These results again highlight the contribution of a subset of hCoV-cross-reactive epitopes for the humoral immune response to SARS-CoV-2.

Point mutations and natural selection in epitopes occur at higher rates upon transmission than within patients

There is mounting evidence that mutations in SARS-CoV-2 enhance viral fitness, replication rate and trans-missibility, and/or partially evade adaptive immunity that has been induced by prior infection or vaccina-tion.[26] Thus, it is essential to shed light on the interplay between SARS-CoV-2 mutations and the acquired immune response in infection. To this end, we tracked the evolution of SARS-CoV-2 B cell epitopes using SNVs identified in 38,685 SARS-CoV-2 genome sequences from the NCBI sequence read archive (Table S9). We selected SARS-CoV-2 samples from the first pandemic wave (defined as January 1 to July 31, 2020) and the second wave (defined as August 1 to December 31, 2020) sequenced using Illumina paired-end ampli-cons with a minimum average depth of coverage of 200× and fewer than 10,000 sites with a depth of coverage lower than 100×. Combined with additional filters to remove sequencing errors (see STAR methods for details), such deep coverage allowed us to identify SNVs that are polymorphic within patients, reflecting within-patient evolution,[51,52] as well as those that are shared between the consensus sequences of different patients. We refer to these within-patient SNVs as 'mutations' and to between-patient SNVs (those present at >75% frequency within a sample and observed in three or more samples) as 'substitutions' that have likely been transmitted across multiple patients. Our definitions of mutations and substitutions are not mutually exclusive: an SNV can be a mutation in one sample and a substitution in another. We counted the absolute number of substitutions relative to the Wuhan-1 reference genome, so the count does not reflect the unique number of substitution events along a phylogeny, but rather the prevalence

Figure 4. Epitope sites identified by high-density peptide arrays (HDPA) are prominently found in structural correlate classes of SARS-CoV-2 neutralization antibodies

(A) Fluorescence Unit (RFU) values of HDPA analysis were used to calculate ratio values to define differential epitope sites (color coded) identified in the SARS-CoV-2 receptor binding domain (RBD) of the Spike protein. Residues that are not part of epitopes are marked in gray (NA, not applicable). Structural correlate classes 1–4 of SARS-CoV-2 neutralization antibodies (NAbs) identified by cryo-EM are shown as asterisks below the RBD sequence; epitopes identified by HDPA are also shown with asterisks. Conservation scores of residues ranging from low (1) to highest (9) are indicated as numbers with colored background.

(B) Pie chart quantifying frequencies of RBD epitope sites recognized by HDPA stratified according to the unique or shared structural correlate classes 1–4 of SARS-CoV-2 NAbs they fall in.

Table 2. Numbers of seasonal endemic hCoV-specific epitope-defining peptides identified by HDPA

	Spike	Nucleocapsid	Envelope	Membrane	ORF1ab	TOTAL
OC43-specific peptides						
SARS-CoV-2 negative	126	37	6	4	293	466
SARS-CoV-2 positive	209	70	10	27	508	824
Overlap	104	35	2	8	280	429
Total	439	142	18	39	1081	1719
HKU1-specific peptides						
SARS-CoV-2 negative	104	43	7	17	293	464
SARS-CoV-2 positive	220	105	7	23	503	858
Overlap	90	35	2	10	254	391
Total	414	183	16	50	1050	1713
NL63-specific peptides						
SARS-CoV-2 negative	139	54	12	17	296	518
SARS-CoV-2 positive	183	77	8	24	571	863
Overlap	70	56	3	14	269	412
Total	392	187	23	55	1136	1793
229E-specific peptides						
SARS-CoV-2 negative	116	43	3	15	306	483
SARS-CoV-2 positive	158	99	7	38	592	894
Overlap	72	46	2	12	325	457
Total	346	188	12	65	1223	1834

Number of seasonal endemic hCoV-specific (229E, HKU1, OC43 and NL63) epitope-defining peptides identified with high density peptide arrays (HDPA) in spike (S) protein, envelope (E) protein, membrane (M) glycoprotein, nucleocapsid (N) phosphoprotein, and ORF1ab. Number of unique peptides that showed a significant antibody response (RFU ≥ 1000) in SARS-CoV-2-negative and SARS-CoV-2-positive groups are depicted. Some peptides are present in both groups, referred to as overlap.

of the substitutions in the database. As such, substitution counts are weighted to reflect their 'success' in transmitting widely.

Using this dataset of mutations and substitutions, we first asked whether cross-reactive (public) epitopes evolved differently than epitopes private to SARS-CoV-2. We found that cross-reactive epitopes tend to evolve more slowly than SARS-CoV-2 private epitopes, accumulating fewer substitutions and having lower ratios of nonsynonymous to synonymous substitutions (Figures S6A–S6D). The same trend of slower evolution in cross-reactive epitopes is also observed at the level of within-patient mutations, but the effect is much stronger at the level of substitutions between patients (Figures S6A–S6D). This is consistent with the fact that these epitopes are conserved across multiple distinct hCoV strains and could be evolving under strong and long-standing functional constraints. However, purifying selection has less time to purge deleterious mutations within hosts, and is therefore more detectable over longer time scales spanning multiple transmission events.

Mutations in epitopes have the potential to evade or lessen the effectiveness of adaptive immunity conferred by infection or vaccination. A recent topic of debate has been the extent to which natural selection for immune evasion acts on SARS-CoV-2 during infection, or upon transmission.[26] During influenza virus infection, most of the selective pressure for immune evasion occurs upon transmission, not within a patient.[53] This is because viral loads often peak before the priming of adaptive immune responses. As such, peak viral transmission occurs before there is time for selection to act within a patient, and for immune evasion to occur. A similar 'asynchrony' transmission model has been proposed for SARS-CoV-2,[54] although data supporting such model has been lacking. To test the asynchrony model in SARS-CoV-2 we tracked SNVs within as well as between patients, within and outside epitope sites, and across the first two pandemic waves (Figures 6 and S7). Throughout both waves, we found
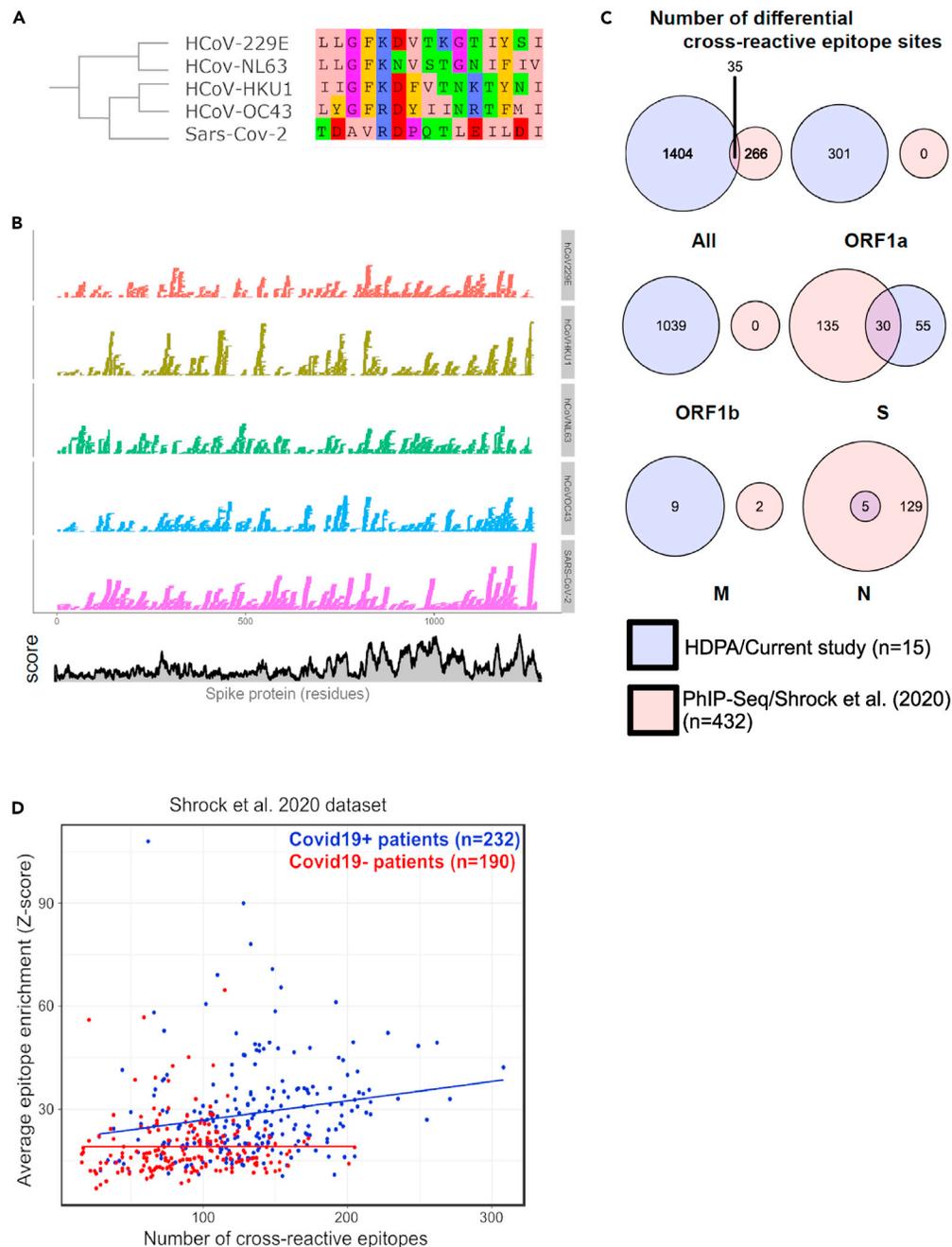
Figure 5. Cross-reactivity to endemic seasonal human coronaviruses is a significant driver of antibody responses to SARS-CoV-2 epitopes

(A) Schematic multiple sequence alignment of proteome sequences between hCoVs and SARS-CoV-2. We defined cross-reactive epitope sites based on peptide sequence conservation between human coronavirus strains (SARS-CoV-2, HKU1, OC43 and NL63) and the presence of an antibody response to the corresponding peptide in SARS-CoV-2 and at least one of the endemic human coronaviruses (229E, HKU1, OC43 and NL63).

(B) Mapping of hCoVs epitope-defining peptides within the Spike (S) protein. The colors represent the 15-mer peptides to which an antibody response to the human coronavirus strains (SARS-CoV-2, 229E, HKU1, OC43 and NL63) has been detected. Conservation score (Cscore) was calculated based on this alignment.

(C) Numbers of differential cross-reactive epitope sites across distinct viral proteins (ORF1A, ORF1B, Spike (S) protein, Envelope (E) protein, Membrane (M) glycoprotein, Nucleocapsid (N) phosphoprotein) detected in the current study using HDPA (blue) compared with a recently published PhIP-Seq study (red, n = 432.[30]

Figure 5. *Continued*

(D) The average immune response to SARS-CoV-2 positively correlates with the number of cross-reactive epitopes. A linear regression between the average epitope Z score per patient and the number of cross-reactive epitopes for both SARS-CoV-2-positive (blue; adjusted R^2 = 0.033; slope = 0.059; p = 0.003) and both SARS-CoV-2 negative (red; not significant; p > 0.5) patients was performed.

consistently lower within-host mutation rates in epitopes sites when compared to non-epitope sites across most SARS-CoV-2 proteins (Figures 6A and S7A). In contrast, the structural proteins S, M, and N had significantly higher rates of between-host substitution in epitope sites compared to non-epitope sites (Figures 6B and S7B), suggesting stronger positive selection for epitope changes upon transmission than within hosts. To further assess the evidence for selection on epitopes, we used the ratio of nonsynonymous to synonymous SNVs both between patients (dN/dS) and within patients (pN/pS) calculated separately within and outside epitopes. Higher ratios indicate positive or relaxed purifying selection, whereas lower ratios indicate stronger purifying selection. We found that the structural proteins S and N have consistently higher nonsynonymous SNV rates in epitope sites, both within and between patients, and across both pandemic waves (Figures 6C, 6D, S7C, and S7D). While this result is consistent with positive selection of altered epitopes (immune evasion) occurring both within and between patients, dN/dS ratios (between patients) are consistently higher than pN/pS ratios (within patients). These observations indicate that nonsynonymous substitutions in S and N epitope sites accumulate most rapidly upon transmission, rather than within patients. Taken together these results support the notion that most of the selective pressure for immune evasion of SARS-CoV-2 occurs upon transmission between hosts, consistent with the asynchrony model.[53]

Assessing the immune evasion potential of SARS-CoV-2 variants

The observation of similar pattern of mutations and selective pressures in epitopes across pandemic waves 1 and 2 (Figures 6 and S7) was surprising, given the expectation that increasing levels of immunity in the population would lead to increased selection for immune evasion over time. The second wave is characterized by the rise of VOCs and VUIs with higher transmissibility and, in some cases, increased disease severity and acquired immune evasion phenotypes.[26] The rise of VOCs has been suggested to be due to a shift in the SARS-CoV-2 fitness landscape.[55] If part of this shift were due to rising population immunity from the first to second wave, one would expect increasing selection for immune evasion variants, resulting in higher frequencies of SNVs in epitopes in wave 2. Although we found a higher total number of nonsynonymous SNVs (including both mutations and substitutions) in epitope sites unique to wave 1 than unique to wave 2 (Figure 7A), wave 2-specific SNVs reached higher frequencies across samples compared to wave 1-specific SNVs, consistent with increased selection for immune evasion over time (Figure 7B). However, mutations common to both waves achieved the highest frequencies, indicating their early appearance and persistence over time (Figure 7B).

A likely driver of VOC evolution is selection for increased transmissibility. For example, the Delta VOC is estimated to be 76–117% more transmissible than non-VOCs and non-VUIs, while Gamma is 29–48% more transmissible and Alpha is 24–33% more transmissible than the original Wuhan SARS-CoV-2 strain.[56] However, selection for immune evasion could also play a significant role for increased spread of VOCs and VUIs. To test this hypothesis, we defined signature mutations of each variant (see STAR methods; Table S10) as substitutions that are present in ≥90% of sequences assigned to that lineage. We calculated the prevalence of substitutions in thousands of publicly available consensus sequences collected from NCBI during 2020 and added data from CoV-Spectrum about under-represented lineage in the database or lineages that emerged during 2021.[57] We focused on nonsynonymous signature mutations located in epitope sites and found that VOCs and VUIs contain significantly more signature mutations in epitopes compared to non-VOCs and non-VUIs (Figure 7C) suggesting that evasion of the humoral immune response could be a significant driver of VOC/VUI evolution. We then ranked VOCs and VUIs based on their number of signature mutations in epitopes (Figure 7D). We observed that Delta has an intermediate number of mutations in mapped epitopes (Figure 7D). However, the most nonsynonymous epitope mutations were observed in Omicronand its sublineages and recombinants such as XBB (Figure 7D), which other studies have shown to be highly immune evasive.[58–60] Most epitope mutations in VOC/VUI occur in the S protein (Figure 7D). Normalizing by gene length revealed a relatively high density of epitope mutations in the N protein, especially at sites of the N protein that overlap with ORF9c (Figure S8), a
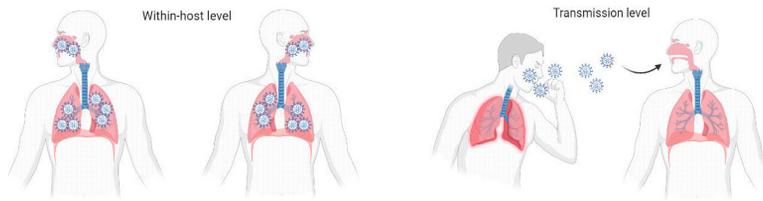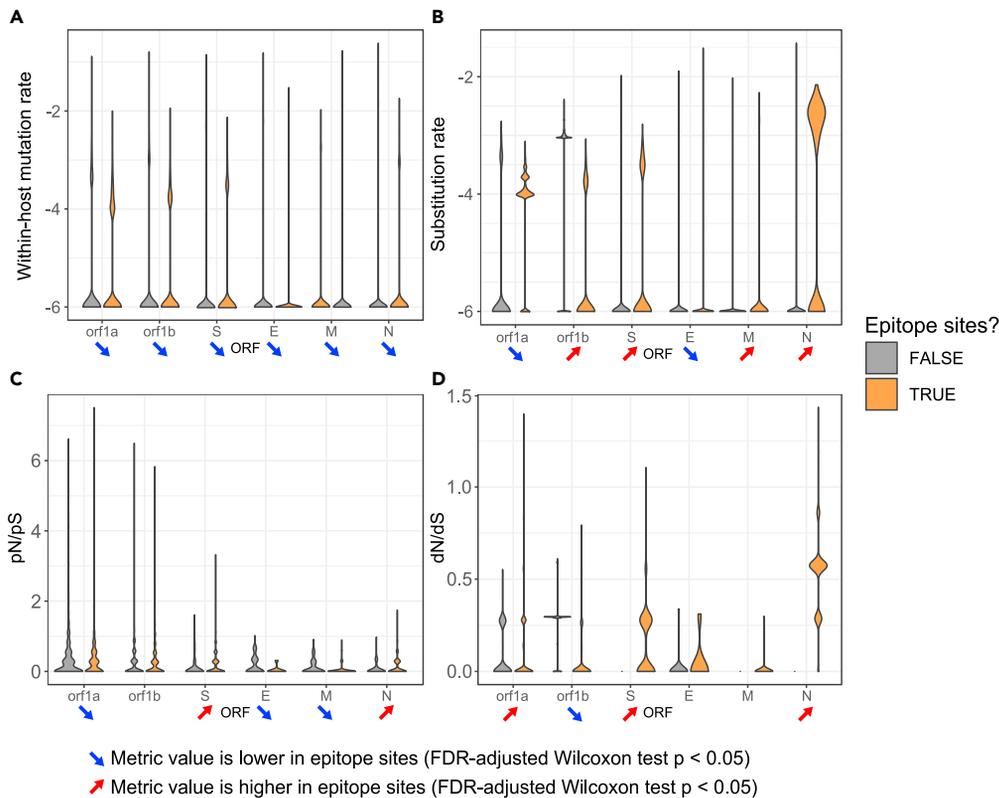
Figure 6. Evolutionary profiles of SARS-CoV-2 epitopes during the first pandemic wave

Mutations (within hosts) and substitutions (between hosts) in epitope sites (orange) vs. non-epitope sites (gray) during the first pandemic SARS-CoV-2 wave (defined as January 1 to July 31, 2020) is depicted. For each metric, significantly lower values in epitope sites of a certain gene are represented by a blue arrow pointing down while significantly higher values in epitope sites of a certain gene are represented by a red arrow pointing up (FDR-adjusted Wilcoxon test p < 0.05).

(A) Distributions of sample mutation rates across proteins.

(B) Distributions of sample substitution rates across proteins.

(C) Distributions of pN/pS across proteins.

(D) Distributions of dN/dS across proteins. Mutation and substitution rates are normalized by gene length and plotted on a log10 scale.

membrane-anchored protein of SARS-CoV-2 that can hinder interferon signaling, viral protein degradation and other stress response pathways when expressed in human lung epithelial cell lines.[61] Finally, having established these general evolutionary patterns of mutation and selection on epitopes, we attempted to pinpoint specific epitope mutations that could hinder the immune response. For each epitope site, we extracted both the measured patient immune responses and the prevalence of nonsynonymous (missense) mutations from the NCBI dataset (Table S11). Among the most prevalent mutations identified are two mutations occurring at consecutive sites in the N protein (N:R203K and N:G204R) that overlap with ORF9c (encoded within the N gene). Taken together our observations show that high resolution epitope mapping combined with genome sequence analysis provides a powerful strategy to rapidly assess the immune evasion potential of emerging SARS-CoV-2 variants.
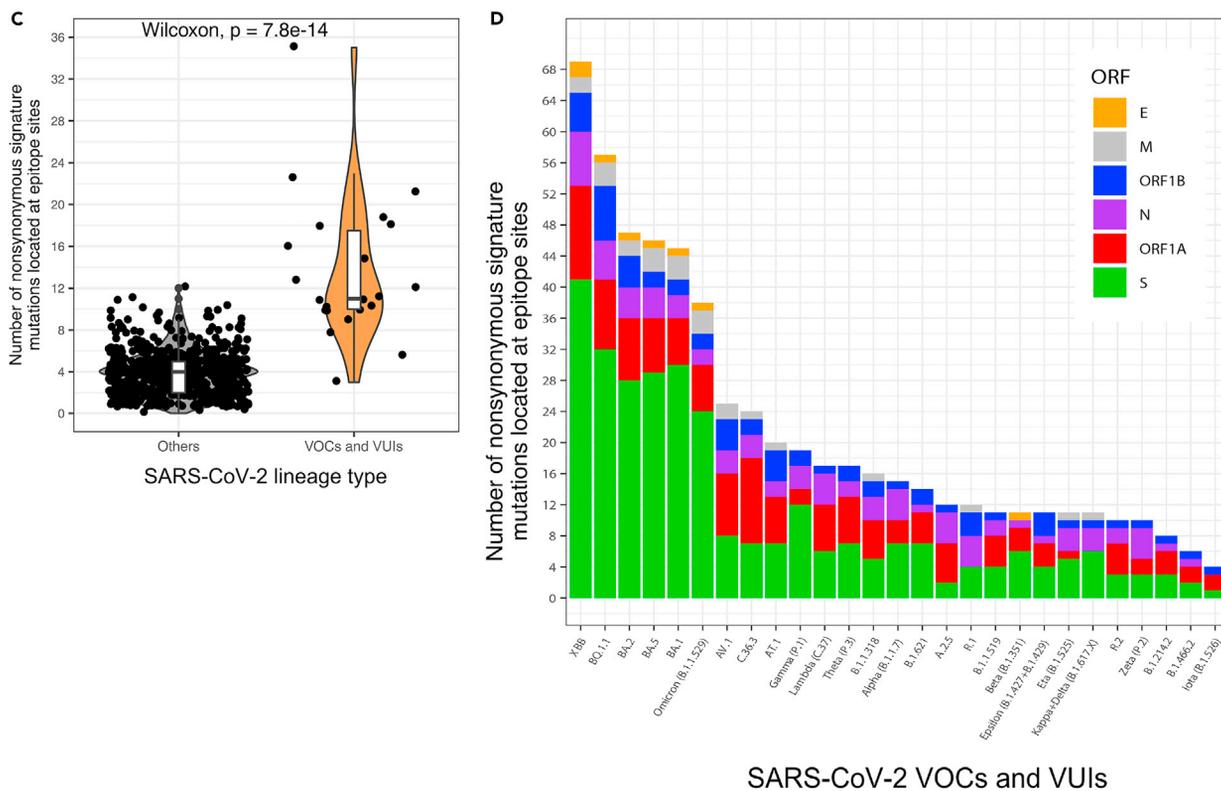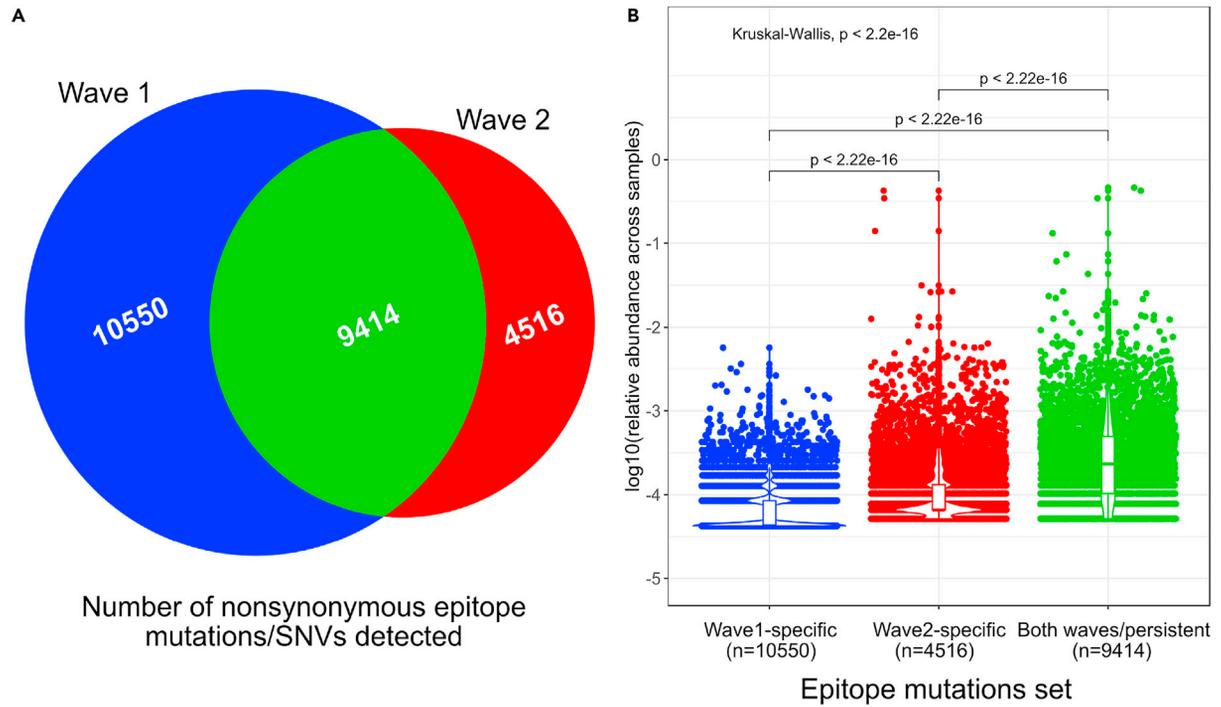
Figure 7. Assessing the immune evasion potential of SARS-CoV-2 variants

(A) Venn diagram showing the numbers of non-synonymous epitope mutations specific to SARS-CoV-2 pandemic wave 1 (blue; defined as January 1 to July 31, 2020), specific to wave 2 (red; defined as August 1 to December 31, 2020) and shared between both waves (green).

(B) Distribution of the relative abundance of non-synonymous mutations in epitopes across pandemic waves. For better visualization, we plotted the Y axis on a log10 scale and represented the distributions with a jitter plot, a violin plot, and a boxplot. The Wilcoxon test p values are indicated above each pair of distributions. The Kruskal-Wallis test p values are indicated at the top left to indicate the significance of the differences across all distributions.

(C) Distribution of the numbers of signature mutations located at epitope sites across SARS-CoV-2 groups (gray for non-VOCs and non-VUIs, orange for VOCs and VUIs). The Wilcoxon test p value is indicated at the top of the panel to show the significance of the differences between the two groups.

(D) Distribution of the numbers of nonsynonymous signature mutations in epitopes of selected VOCs and VUIs. For each VOC or VUI we indicate the number of signature mutations in epitopes identified with HDPA across all analyzed ORFs: envelope (E) protein (orange), membrane (M) glycoprotein (gray), N sites overlapping ORF9c (black), ORF1b (blue), nucleocapsid (N) phosphoprotein (purple), ORF1A (red), and spike (S) protein (green).

DISCUSSION

An in-depth map of the breadth of the antigenic determinants of the immune response following infection with SARS-CoV-2 is key for a better understanding of the diagnostic markers, the identification of the correlates of protection and the monitoring of vaccine efficacy. We therefore set out to define the antigenic hotspots and epitope signatures of SARS-CoV-2-specific humoral immune responses in patients with COVID-19 and uninfected healthy controls using high-density peptide microarrays (HDPA) covering the entire proteomes of SARS-CoV-2 as well as of the four seasonal hCoVs (OC43, NL63, HKU1, and 229E). Our results demonstrate that the HDPA approach provides a sensitive, high-throughput antibody profiling strategy to identify linear and conformational B cell epitopes. Using structural models, we found that many of the epitope sites identified in the S protein are in the NTD and RBD region of the S protein. Interestingly, most epitope sites identified in the NTD and RBD are poorly conserved across coronaviruses, while epitope elsewhere in the S protein were more highly conserved. In addition, HDPA analysis revealed strong and specific antibody immunoreactivity in select epitope sites of structural SARS-CoV-2 proteins (E, M, N proteins), as well as ORF1AB.

Antibody cross-reactivity with similar viral antigens affects the accuracy of serological tests, but also has the potential to elicit beneficial immunological memory responses that could affect the course of SARS-CoV-2 infections. Our results highlight a significant cross-reactivity between SARS-CoV-2 and hCoV B cell epitope sites in many viral proteins, demonstrating that HDPA allows to uncover another dimension of cross-reactive immunity relative to PhIP-Seq. The fact that more differential epitope sites in the S and N proteins were detected by a recent study using PhIP-Seq[30] probably reflects the lower sample size of our dataset. However, HDPA detected more cross-reactive epitope sites than PhIP-Seq with fewer patient samples analyzed, reflecting one of the benefits of our applied methodology. Sensitivity limitations of PhIP-seq to broadly detected polio epitopes have been previously reported[29,31] and might contribute to the observed differences, similarly affecting detectability of CoV antigens. Such limitations are not observed in our HDPA approach[62] which typically yielded strong polio responsiveness in over 90% of sampled individuals.[63]

While HDPA assays show significant benefits over PhIP-seq analyses, a few limitations still apply in our study. It needs to be taken into consideration that our analysis only targeted ten patient and five control sera, and the antibody-binding epitope profile reflects the breadth of anti-viral immunity in these specific patients. Hence, the non-epitope sites may not necessarily be sites that cannot be recognized by antibodies and increasing the number of analyzed sera could also potentially increase the breadth of the antibody response detected in our assay. Moreover, the signal readout of the HDPA microarray combines several parameters which could potentially affect epitope availability including antibody titer, antibody affinity and antibody off-rates since assays are stopped at a specific timepoint followed by washing and addition of a secondary antibody. Folding effects of the immobilized peptides on the chip can also play a role in masking available epitopes, as peptides tend to form structures (e.g., helical in nature) with increasing distance from the glass slide coating. Such folding may be beneficial or counterproductive for antibody binding. Based on our experience running HDPA, in general, RFU of equal to or higher than 100 reveal a positive antibody response. However, to focus on the most prominent antibody responses, our analysis was performed with an RFU cut-off of 1000 or higher, representing the most significant epitope sites of the cohort analyzed.

Although HDPA is primarily considered to provide linear epitope information, folding effects of the immobilized peptides may provide a large structural interface diversity for conformational antibody-binding sites. In addition, studies have indicated that the SARS-CoV-2 spike protein is very flexible and exhibits

several prefusion conformations.[39] The HDPA approach might therefore provide a larger and more diverse structural interface for antibody-binding than full-length antigens, in turn, reflecting enhanced diversities of structural interfaces to viral antigen sites during viral infection. Using RFU values of HDPA analysis to calculate ratio values of the analyzed sera (uninfected vs. infected) allowed for the definition of differential epitope sites per amino acid residue within the distinct SARS-CoV-2 proteins. Using structural modeling, we also mapped the epitope sites identified by HDPA within the S protein of SARS-CoV-2 and showed that many of the epitope sites are in the NTD and the receptor binding domain (RBD) (Figure 3).

To analyze the extent to which HDPA epitope sites overlap with known neutralizing antibody (NAb)-binding residues in SARS-CoV-2 RBD, we aligned the epitopes sites identified by HDPA with NAb-binding residue sites mapped by cryo-EM studies 39,40. We demonstrate that more than 60% of RBD residues in NAb-binding sites are identified by HDPA analysis, mapping epitope sites in all four distinct structural correlate classes of SARS-CoV-2 RBD-binding NAbs identified by cryo-EM 39,40. Hence, our data clearly shows that our HDPA profiling strategy can identify sets of linear and conformational B cell epitopes unique in sera of SARS-CoV-2-infected individuals. Importantly, a large portion of the identified epitope sites are residues located in previously reported NAb-binding sites, demonstrating that epitope sites mapped by HDPA analysis are functionally relevant in Ab-mediated immunity to SARS-CoV-2.

Importantly, the cross-reactivity in identified B cell epitope sites positively relates to previous infections with seasonal common cold hCoVs. This suggests that immune memory conferred by previous seasonal hCoV infections positively influences SARS-CoV-2-specific antibody responses and may explain the large portion of SARS-CoV-2-infected individuals with mild and asymptomatic disease symptoms.[4] Notably, there was little to no correlation between cross-reactivity and immune response in COVID-19 negative patients, suggesting that resistance to infection is not easily explained by cross-reactivity. However, this molecular cross-reactivity can pose important complications in serological tests, particularly when studying asymptomatic patients. Cross-reactivity in immunodominant epitopes can be molecular determinants of strong immunity in individuals and therefore may serve as the basis for future pan-coronavirus vaccine design strategies. In turn, mutations in these cross-reactive epitopes can potentially breach pre-existing immune protection conferred by previous viral exposures, contributing to viral evolution, immune selection, and immune evasion.

By combining our epitope dataset with publicly available SARS-CoV-2 genome sequences, we were able to study mutations that occur in epitopes and compare their rates of evolution and selective pressures to non-epitope sites. Ideally, we would have matched epitopes and viral mutations arising from the same patients to infer selection more directly for immune evasion. Although such matched data is currently rare, we were still able to make inferences about evolution within epitopes on a population-wide scale. First, we found that mutations in SARS-CoV-2 epitopes are under evolutionary constraints. SARS-CoV-2-specific epitopes that are cross-reactive with other endemic seasonal hCoVs tend to accumulate fewer substitutions and are under purifying selection against nonsynonymous changes. Second, epitopes in structural proteins S, M, and N accumulate more substitutions and are under stronger positive selection for nonsynonymous changes than non-epitopes. Natural selection favoring changes in epitope sites was therefore detectable during the first two pandemic waves. As population immunity accumulates over time, we would expect increasing selection for immune evasion. Consistent with this expectation, we observed that mutations in epitopes increased in frequency from the first to the second pandemic wave, and we expect this trend to continue.

Notably, we found much slower rates of evolution and weaker evidence for positive selection on epitopes within patients, indicating that most selection for immune evasion occurs upon transmission rather than within patients. This is consistent with asynchrony between peak viral loads (when selection is most efficient) and the adaptive immune response, as is the case for influenza.[53,64] Notable exceptions are chronic infections, in which significant adaptive evolution occurs within patients, likely including antibody evasion.[65,66] However, such infections likely represent a small minority of the sequences included in our dataset. While they may be important – particularly if a chronic infection is transmitted – they do not represent most COVID-19 cases. Another non-exclusive explanation for the higher between-host mutation rate in epitope sites is the small transmission bottleneck.[67,68]

Consistent with a general trend of immune evasion, we observed that VOCs and VUIs contain significantly more signature mutations in epitopes than non-VOCs and non-VUIs demonstrating that evasion of the

humoral immune response is a significant driver of VOC/VUI evolution. The most mutations in epitopes were found in the VOCs Delta, C.36.3, and especially Omicron (B.1.1.529) and its sublineages. Most of these epitope mutations in VOCs are localized to the S protein, highlighting that polymorphism in the S protein critically impacts antigenicity in highly transmissible variants.

Much research rightly focuses on the S protein, but we also find mutation in the N protein epitopes that could be selected for immune evasion. The N protein had the highest dN/dS values during both pandemic waves analyzed, suggesting the presence of a subset of epitope substitutions under positive selection. After normalizing by gene length, we found the highest density of epitope mutations in the N protein, especially in regions of overlap with ORF9c. Orf9c is one of the four conserved overlapping genes (OLGs) of SARS-CoV-2,[69] wherein a single stretch of nucleotides encodes two distinct proteins in different reading frames. OLGs are 'genes within genes' that compress genomic information, thereby allowing genetic innovation via *overprinting*.[70,71] However, a single mutation in an OLG may alter two proteins at the same time, constraining evolution of the pre-existing open reading frame (ORF). Although, OLGs are known entities that contribute to the emergence and pathogenicity of new viruses,[72] unfortunately, genome annotation methods typically miss OLGs, instead favoring one ORF per genomic region.[72] Similarly, they remain inconsistently reported in viruses of the SARS *coronavirus* species.[73] Importantly, annotations of *ORF9b* and *ORF9c* are conflicting or absent in the SARS-CoV-2 reference genome Wuhan-Hu-1 (NCBI: NC_045512.2) and genomic studies.[74,75] In addition, OLGs are often not displayed in genome browsers[76] and therefore such inconsistencies complicate research to decipher their role in infection and immunity.

The small protein encoded by the ORF9c OLG has recently been shown to constitute a membrane-associated protein to suppresses antiviral interferon and antigen-presentation responses and modify innate immune responses.[61,77,78] Here, we found that N protein epitopes in the region overlap with ORF9c constitute an antigenic target of the humoral immune response and accumulate a high density of mutations in VOCs. It remains to be investigated if and to what extent ORF9c-specific immune responses contribute to host protection and if mutations could also affect these responses. Other OLG-derived proteins, including Orf3d, ORF8 and Orf9b, have been shown to elicit strong antibody responses in sera from COVID-19 patients,[79–81] although their contribution to host protection remains unknown. Concerns have arisen that S-specific vaccine immunity conferred solely to S protein may fail to neutralize emerging variants of SARS-CoV-2 and contribute to selection of immune escape variants.[82–84] Vaccination studies in rodent models using N protein as antigenic target have recently shown the establishment of protective immunity.[85] Hence, expansion of viral antigenic targets in SARS-CoV-2 vaccines, including OLG proteins, to broaden epitope coverage and immune effector mechanisms should be a goal in the development of new COVID-19 vaccines.

Limitation of the study

While HDPA assays show significant benefits over other epitope mapping strategies, a few limitations still apply in our study. It needs to be taken into consideration that we were only able to analyze 10 patients and 5 control sera. Hence, the non-epitope sites may not necessarily be sites that cannot be recognized by antibodies, and increasing the number of analyzed sera could potentially increase the breadth of the antibody response detected by the HDPA assay. Moreover, the signal readout of the HDPA microarray comprises several parameters which could potentially affect epitope availability including antibody titer, antibody affinity and antibody off-rates since assays are stopped at a specific timepoint followed by washing and addition of a secondary antibody. Folding effects of the immobilized peptides on the chip can also play a role in masking available epitopes, as peptides tend to form structures (e.g., helical in nature) with increasing distance from the glass slide coating. Such folding may be beneficial or detrimental for antibody binding.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  ○ Lead contact
  ○ Materials availability
  ○ Data and code availability
- EXPERIMENTAL ANALYSIS AND STUDY PARTICIPANTS DETAILS

- ○ Human participants
- ● METHOD DETAILS
  - ○ High-density peptide array (HDPA)
  - ○ Defining cross-reactivity using protein conservation and immune response to endemic human coronaviruses
  - ○ Detecting epitopes that are significantly more prevalent in SARS-CoV-2 positive patients
  - ○ Structural properties of B cell epitopes and B cell epitope prediction
  - ○ Epitope evolution profiling
  - ○ Selection for immune escape in VOCs and VUIs genomes
- ● QUANTIFICATION AND STATISTICAL ANALYSIS
- ● ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107394.

AUTHOR CONTRIBUTIONS

J.H.F., C.A.P., and B.J.S. conceived the study and designed the experiments. P.C. and R.D. provided serum samples. C.H. and C.S. performed HDPA analysis. A.N.G., S.K., F.M., R.P., J.C.G., G.B., J.G.H., and N.I. analyzed data and designed Figures and Tables. All authors contributed to the writing and editing of the manuscript.

DECLARATION OF INTERESTS

Volker Stadler is the Chief Executive Officer of PEPperPRINT. All other authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

REFERENCES

1. Corman, V.M., Muth, D., Niemeyer, D., and Drosten, C. (2018). Hosts and Sources of Endemic Human Coronaviruses. Adv. Virus Res. *100*, 163–188. https://doi.org/10.1016/bs.aivir.2018.01.001.

2. Cui, J., Li, F., and Shi, Z.L. (2019). Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. *17*, 181–192. https://doi.org/10.1038/s41579-018-0118-9.

3. Hu, B., Guo, H., Zhou, P., and Shi, Z.L. (2021). Characteristics of SARS-CoV-2 and COVID-19. Nat. Rev. Microbiol. *19*, 141–154. https://doi.org/10.1038/s41579-020-00459-7.

4. Sah, P., Fitzpatrick, M.C., Zimmer, C.F., Abdollahi, E., Juden-Kelly, L., Moghadas, S.M., Singer, B.H., and Galvani, A.P. (2021). Asymptomatic SARS-CoV-2 infection: A systematic review and meta-analysis. Proc. Natl. Acad. Sci. USA *118*, e2109229118. https://doi.org/10.1073/pnas.2109229118.

5. Ramos-Casals, M., Brito-Zerón, P., and Mariette, X. (2021). Systemic and organ-specific immune-related manifestations of COVID-19. Nat. Rev. Rheumatol. *17*, 315–332. https://doi.org/10.1038/s41584-021-00608-z.

6. Velavan, T.P., Pallerla, S.R., Rüter, J., Augustin, Y., Kremsner, P.G., Krishna, S., and Meyer, C.G. (2021). Host genetic factors determining COVID-19 susceptibility and severity. EBioMedicine *72*, 103629. https://doi.org/10.1016/j.ebiom.2021.103629.

7. Ng, W.H., Tipih, T., Makoah, N.A., Vermeulen, J.G., Goedhals, D., Sempa, J.B., Burt, F.J., Taylor, A., and Mahalingam, S. (2021). Comorbidities in SARS-CoV-2 Patients: a Systematic Review and Meta-Analysis. mBio *12*, e03647-20. https://doi.org/10.1128/mBio.03647-20.

8. Fernández Villalobos, N.V., Ott, J.J., Klett-Tammen, C.J., Bockey, A., Vanella, P., Krause, G., and Lange, B. (2021). Effect modification of the association between comorbidities and severe course of COVID-19 disease by age of study participants: a systematic review and meta-analysis. Syst. Rev. *10*, 194. https://doi.org/10.1186/s13643-021-01732-3.

9. Cromer, D., Juno, J.A., Khoury, D., Reynaldi, A., Wheatley, A.K., Kent, S.J., and Davenport, M.P. (2021). Prospects for durable immune control of SARS-CoV-2 and prevention of reinfection. Nat. Rev. Immunol. *21*, 395–404. https://doi.org/10.1038/s41577-021-00550-x.

10. Lipsitch, M., Krammer, F., Regev-Yochay, G., Lustig, Y., and Balicer, R.D. (2022). SARS-CoV-2 breakthrough infections in vaccinated individuals: measurement, causes and impact. Nat. Rev. Immunol. *22*, 57–65. https://doi.org/10.1038/s41577-021-00662-4.

11. Murray, S.M., Ansari, A.M., Frater, J., Klenerman, P., Dunachie, S., Barnes, E., and Ogbe, A. (2023). The impact of pre-existing cross-reactive immunity on SARS-CoV-2 infection and vaccine responses. Nat. Rev. Immunol. *23*, 304–316. https://doi.org/10.1038/s41577-022-00809-x.

12. Selva, K.J., van de Sandt, C.E., Lemke, M.M., Lee, C.Y., Shoffner, S.K., Chua, B.Y., Davis, S.K., Nguyen, T.H.O., Rowntree, L.C., Hensen, L., et al. (2021). Systems serology detects functionally distinct coronavirus antibody features in children and elderly. Nat. Commun. *12*, 2037. https://doi.org/10.1038/s41467-021-22236-7.

13. McNaughton, A.L., Paton, R.S., Edmans, M., Youngs, J., Wellens, J., Phalora, P., Fyfe, A., Belij-Rammerstorfer, S., Bolton, J.S., Ball, J., et al. (2022). Fatal COVID-19 outcomes are associated with an antibody response targeting epitopes shared with endemic coronaviruses. JCI Insight *7*, e156372. https://doi.org/10.1172/jci.insight.156372.

14. Wratil, P.R., Schmacke, N.A., Karakoc, B., Dulovic, A., Junker, D., Becker, M., Rothbauer, U., Osterman, A., Spaeth, P.M., Ruhle, A., et al. (2021). Evidence for increased SARS-CoV-2 susceptibility and COVID-19 severity related to pre-existing immunity to seasonal coronaviruses. Cell Rep. *37*, 110169. https://doi.org/10.1016/j.celrep.2021.110169.

15. Ng, K.W., Faulkner, N., Cornish, G.H., Rosa, A., Harvey, R., Hussain, S., Ulferts, R., Earl, C., Wrobel, A.G., Benton, D.J., et al. (2020). Preexisting and de novo humoral immunity to SARS-CoV-2 in humans. Science *370*, 1339–1343. https://doi.org/10.1126/science.abe1107.

16. Morgenlander, W.R., Henson, S.N., Monaco, D.R., Chen, A., Littlefield, K., Bloch, E.M., Fujimura, E., Ruczinski, I., Crowley, A.R., Natarajan, H., et al. (2021). Antibody responses to endemic coronaviruses modulate COVID-19 convalescent plasma functionality. J. Clin. Invest. *131*, e146927. https://doi.org/10.1172/JCI146927.

17. Welsh, R.M., Che, J.W., Brehm, M.A., and Selin, L.K. (2010). Heterologous immunity between viruses. Immunol. Rev. *235*, 244–266. https://doi.org/10.1111/j.0105-2896.2010.00897.x.

18. Screaton, G., Mongkolsapaya, J., Yacoub, S., and Roberts, C. (2015). New insights into the immunopathology and control of dengue virus infection. Nat. Rev. Immunol. *15*, 745–759. https://doi.org/10.1038/nri3916.

19. Nelde, A., Bilich, T., Heitmann, J.S., Maringer, Y., Salih, H.R., Roerden, M., Lübke, M., Bauer, J., Rieth, J., Wacker, M., et al. (2021). SARS-CoV-2-derived peptides define heterologous and COVID-19-induced T cell recognition. Nat. Immunol. *22*, 74–85. https://doi.org/10.1038/s41590-020-00808-x.

20. Swadling, L., Diniz, M.O., Schmidt, N.M., Amin, O.E., Chandran, A., Shaw, E., Pade, C., Gibbons, J.M., Le Bert, N., Tan, A.T., et al. (2022). Pre-existing polymerase-specific T cells expand in abortive seronegative SARS-CoV-2. Nature *601*, 110–117. https://doi.org/10.1038/s41586-021-04186-8.

21. Loyal, L., Braun, J., Henze, L., Kruse, B., Dingeldey, M., Reimer, U., Kern, F., Schwarz, T., Mangold, M., Unger, C., et al. (2021). Cross-reactive CD4(+) T cells enhance SARS-CoV-2 immune responses upon infection and vaccination. Science *374*, eabh1823. https://doi.org/10.1126/science.abh1823.

22. da Silva Antunes, R., Pallikkuth, S., Williams, E., Dawen Yu, E., Mateus, J., Quiambao, L., Wang, E., Rawlings, S.A., Stadlbauer, D., Jiang, K., et al. (2021). Differential T-Cell Reactivity to Endemic Coronaviruses and SARS-CoV-2 in Community and Health Care Workers. J. Infect. Dis. *224*, 70–80. https://doi.org/10.1093/infdis/jiab176.

23. Kundu, R., Narean, J.S., Wang, L., Fenn, J., Pillay, T., Fernandez, N.D., Conibear, E., Koycheva, A., Davies, M., Tolosa-Wright, M., et al. (2022). Cross-reactive memory T cells associate with protection against SARS-CoV-2 infection in COVID-19 contacts. Nat. Commun. *13*, 80. https://doi.org/10.1038/s41467-021-27674-x.

24. Ogbe, A., Pace, M., Bittaye, M., Tipoe, T., Adele, S., Alagaratnam, J., Aley, P.K., Ansari, M.A., Bara, A., Broadhead, S., et al. (2022). Durability of ChAdOx1 nCoV-19 vaccination in people living with HIV. JCI Insight *7*, e157031. https://doi.org/10.1172/jci.insight.157031.

25. Gartlan, C., Tipton, T., Salguero, F.J., Sattentau, Q., Gorringe, A., and Carroll, M.W. (2022). Vaccine-Associated Enhanced Disease and Pathogenic Human Coronaviruses. Front. Immunol. *13*, 882972. https://doi.org/10.3389/fimmu.2022.882972.

26. Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., et al.; COVID-19 Genomics UK COG-UK Consortium (2021). SARS-CoV-2 variants, spike mutations and immune escape. Nat. Rev. Microbiol. *19*, 409–424. https://doi.org/10.1038/s41579-021-00573-0.

27. Ng, K.W., Faulkner, N., Finsterbusch, K., Wu, M., Harvey, R., Hussain, S., Greco, M., Liu, Y., Kjaer, S., Swanton, C., et al. (2022). SARS-CoV-2 S2-targeted vaccination elicits broadly neutralizing antibodies. Sci. Transl. Med. *14*, eabn3715. https://doi.org/10.1126/scitranslmed.abn3715.

28. Zhou, P., Yuan, M., Song, G., Beutler, N., Shaabani, N., Huang, D., He, W.T., Zhu, X., Callaghan, S., Yong, P., et al. (2022). A human antibody reveals a conserved site on beta-coronavirus spike proteins and confers protection against SARS-CoV-2 infection. Sci. Transl. Med. *14*, eabi9215. https://doi.org/10.1126/scitranslmed.abi9215.

29. Xu, G.J., Kula, T., Xu, Q., Li, M.Z., Vernon, S.D., Ndung'u, T., Ruxrungtham, K., Sanchez, J., Brander, C., Chung, R.T., et al. (2015). Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome. Science *348*, aaa0698. https://doi.org/10.1126/science.aaa0698.

30. Shrock, E., Fujimura, E., Kula, T., Timms, R.T., Lee, I.H., Leng, Y., Robinson, M.L., Sie, B.M., Li, M.Z., Chen, Y., et al. (2020). Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. Science *370*, eabd4250. https://doi.org/10.1126/science.abd4250.

31. Weber, L.K., Palermo, A., Kügler, J., Armant, O., Isse, A., Rentschler, S., Jaenisch, T., Hubbuch, J., Dübel, S., Nesterov-Mueller, A., et al. (2017). Single amino acid fingerprinting of the human antibody repertoire with high density peptide arrays. J. Immunol. Methods *443*, 45–54. https://doi.org/10.1016/j.jim.2017.01.012.

32. Klompus, S., Leviatan, S., Vogl, T., Mazor, R.D., Kalka, I.N., Stoler-Barak, L., Nathan, N., Peres, A., Moss, L., Godneva, A., et al. (2021). Cross-reactive antibodies against human coronaviruses and the animal coronavirome suggest diagnostics for future zoonotic spillovers. Sci. Immunol. *6*, eabe9950. https://doi.org/10.1126/sciimmunol.abe9950.

33. Heiss, K., Heidepriem, J., Fischer, N., Weber, L.K., Dahlke, C., Jaenisch, T., and Loeffler, F.F. (2020). Rapid Response to Pandemic Threats: Immunogenic Epitope Detection of Pandemic Pathogens for Diagnostics and Vaccine Development Using Peptide Microarrays. J. Proteome Res. *19*, 4339–4354. https://doi.org/10.1021/acs.jproteome.0c00484.

34. Dai, L., and Gao, G.F. (2021). Viral targets for vaccines against COVID-19. Nat. Rev. Immunol. *21*, 73–82. https://doi.org/10.1038/s41577-020-00480-0.

35. Haste Andersen, P., Nielsen, M., and Lund, O. (2006). Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. *15*, 2558–2567. https://doi.org/10.1110/ps.062405906.

36. Jespersen, M.C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: improving

sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. *45*, W24–W29. https://doi.org/10.1093/nar/gkx346.

37. Atanackovic, D., Avila, S.V., Lutfi, F., de Miguel-Perez, D., Fan, X., Sanchez-Petitto, G., Vander Mause, E., Siglin, J., Baddley, J., Mannuel, H.D., et al. (2021). Deep dissection of the antiviral immune profile of patients with COVID-19. Commun. Biol. *4*, 1389. https://doi.org/10.1038/s42003-021-02852-1.

38. Premkumar, L., Segovia-Chumbez, B., Jadi, R., Martinez, D.R., Raut, R., Markmann, A., Cornaby, C., Bartelt, L., Weiss, S., Park, Y., et al. (2020). The receptor binding domain of the viral spike protein is an immunodominant and highly specific target of antibodies in SARS-CoV-2 patients. Sci. Immunol. *5*, eabc8413. https://doi.org/10.1126/sciimmunol.abc8413.

39. Mittal, A., Khattri, A., and Verma, V. (2022). Structural and antigenic variations in the spike protein of emerging SARS-CoV-2 variants. PLoS Pathog. *18*, e1010260. https://doi.org/10.1371/journal.ppat.1010260.

40. Barnes, C.O., Jette, C.A., Abernathy, M.E., Dam, K.M.A., Esswein, S.R., Gristick, H.B., Malyutin, A.G., Sharaf, N.G., Huey-Tubman, K.E., Lee, Y.E., et al. (2020). SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. Nature *588*, 682–687. https://doi.org/10.1038/s41586-020-2852-1.

41. Fung, T.S., and Liu, D.X. (2019). Human Coronavirus: Host-Pathogen Interaction. Annu. Rev. Microbiol. *73*, 529–557. https://doi.org/10.1146/annurev-micro-020518-115759.

42. Gorse, G.J., Patel, G.B., Vitale, J.N., and O'Connor, T.Z. (2010). Prevalence of antibodies to four human coronaviruses is lower in nasal secretions than in serum. Clin. Vaccine Immunol. *17*, 1875–1880. https://doi.org/10.1128/CVI.00278-10.

43. Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R.H., Peters, B., and Sette, A. (2020). A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. Cell Host Microbe *27*, 671–680.e2. https://doi.org/10.1016/j.chom.2020.03.002.

44. Galipeau, Y., Greig, M., Liu, G., Driedger, M., and Langlois, M.A. (2020). Humoral Responses and Serological Assays in SARS-CoV-2 Infections. Front. Immunol. *11*, 610688. https://doi.org/10.3389/fimmu.2020.610688.

45. van der Heide, V. (2020). SARS-CoV-2 cross-reactivity in healthy donors. Nat. Rev. Immunol. *20*, 408. https://doi.org/10.1038/s41577-020-0362-x.

46. Lipsitch, M., Grad, Y.H., Sette, A., and Crotty, S. (2020). Cross-reactive memory T cells and herd immunity to SARS-CoV-2. Nat. Rev. Immunol. *20*, 709–713. https://doi.org/10.1038/s41577-020-00460-4.

47. Sette, A., and Crotty, S. (2020). Pre-existing immunity to SARS-CoV-2: the knowns and unknowns. Nat. Rev. Immunol. *20*, 457–458. https://doi.org/10.1038/s41577-020-0389-z.

48. Yang, F., Nielsen, S.C.A., Hoh, R.A., Röltgen, K., Wirz, O.F., Haraguchi, E., Jean, G.H., Lee, J.Y., Pham, T.D., Jackson, K.J.L., et al. (2021). Shared B cell memory to coronaviruses and other pathogens varies in human age groups and tissues. Science *372*, 738–741. https://doi.org/10.1126/science.abf6648.

49. Livingstone, C.D., and Barton, G.J. (1993). Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput. Appl. Biosci. *9*, 745–756. https://doi.org/10.1093/bioinformatics/9.6.745.

50. De Cáceres, M., and Legendre, P. (2009). Associations between species and groups of sites: indices and statistical inference. Ecology *90*, 3566–3574. https://doi.org/10.1890/08-1823.1.

51. Wang, Y., Wang, D., Zhang, L., Sun, W., Zhang, Z., Chen, W., Zhu, A., Huang, Y., Xiao, F., Yao, J., et al. (2021). Intra-host variation and evolutionary dynamics of SARS-CoV-2 populations in COVID-19 patients. Genome Med. *13*, 30. https://doi.org/10.1186/s13073-021-00847-5.

52. Lythgoe, K.A., Hall, M., Ferretti, L., de Cesare, M., MacIntyre-Cockett, G., Trebes, A., Andersson, M., Otecko, N., Wise, E.L., Moore, N., et al. (2021). SARS-CoV-2 within-host diversity and transmission. Science *372*, eabg0821. https://doi.org/10.1126/science.abg0821.

53. Petrova, V.N., and Russell, C.A. (2018). The evolution of seasonal influenza viruses. Nat. Rev. Microbiol. *16*, 47–60. https://doi.org/10.1038/nrmicro.2017.118.

54. Sette, A., and Crotty, S. (2021). Adaptive immunity to SARS-CoV-2 and COVID-19. Cell *184*, 861–880. https://doi.org/10.1016/j.cell.2021.01.007.

55. Martin, D.P., Weaver, S., Tegally, H., San, J.E., Shank, S.D., Wilkinson, E., Lucaci, A.G., Giandhari, J., Naidoo, S., Pillay, Y., et al. (2021). The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. Cell *184*, 5189–5200.e7. https://doi.org/10.1016/j.cell.2021.09.003.

56. Campbell, F., Archer, B., Laurenson-Schafer, H., Jinnai, Y., Konings, F., Batra, N., Pavlin, B., Vandemaele, K., Van Kerkhove, M.D., Jombart, T., et al. (2021). Increased transmissibility and global spread of SARS-CoV-2 variants of concern as at June 2021. Euro Surveill. *26*, 2100509. https://doi.org/10.2807/1560-7917.ES.2021.26.24.2100509.

57. Chen, C., Nadeau, S., Yared, M., Voinov, P., Xie, N., Roemer, C., and Stadler, T. (2022). CoV-Spectrum: Analysis of Globally Shared SARS-CoV-2 Data to Identify and Characterize New Variants. Bioinformatics *38*, 1735–1737. https://doi.org/10.1093/bioinformatics/btab856.

58. Cele, S., Jackson, L., Khoury, D.S., Khan, K., Moyo-Gwete, T., Tegally, H., San, J.E., Cromer, D., Scheepers, C., Amoako, D.G., et al. (2022). Omicron extensively but incompletely escapes Pfizer BNT162b2 neutralization. Nature *602*, 654–656. https://doi.org/10.1038/s41586-021-04387-1.

59. Mykytyn, A.Z., Rosu, M.E., Kok, A., Rissmann, M., van Amerongen, G., Geurtsvankessel, C., de Vries, R.D., Munnink, B.B.O., Smith, D.J., Koopmans, M.P.G., et al. (2023). Antigenic mapping of emerging SARS-CoV-2 omicron variants BM.1.1.1, BQ.1.1, and XBB.1. Lancet. Microbe *4*, e294–e295. https://doi.org/10.1016/S2666-5247(22)00384-6.

60. Cao, Y., Jian, F., Wang, J., Yu, Y., Song, W., Yisimayi, A., Wang, J., An, R., Chen, X., Zhang, N., et al. (2023). Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. Nature *614*, 521–529. https://doi.org/10.1038/s41586-022-05644-7.

61. Dominguez Andres, A., Feng, Y., Campos, A.R., Yin, J., Yang, C.C., James, B., Murad, R., Kim, H., Deshpande, A.J., Gordon, D.E., et al. (2020). SARS-CoV-2 ORF9c Is a Membrane-Associated Protein that Suppresses Antiviral Responses in Cells. Preprint at bioRxiv. https://doi.org/10.1101/2020.08.18.256776.

62. Stadler, V., Felgenhauer, T., Beyer, M., Fernandez, S., Leibe, K., Güttler, S., Gröning, M., König, K., Torralba, G., Hausmann, M., et al. (2008). Combinatorial synthesis of peptide arrays with a laser printer. Angew Chem. Int. Ed. Engl. *47*, 7132–7135. https://doi.org/10.1002/anie.200801616.

63. Blank, A., Fürle, K., Jäschke, A., Mikus, G., Lehmann, M., Hüsing, J., Heiss, K., Giese, T., Carter, D., Böhnlein, E., et al. (2020). Immunization with full-length Plasmodium falciparum merozoite surface protein 1 is safe and elicits functional cytophilic antibodies in a randomized first-in-human trial. NPJ Vaccines *5*, 10. https://doi.org/10.1038/s41541-020-0160-2.

64. Xue, K.S., and Bloom, J.D. (2020). Linking influenza virus evolution within and between human hosts. Virus Evol. *6*, veaa010. https://doi.org/10.1093/ve/veaa010.

65. Kemp, S.A., Collier, D.A., Datir, R.P., Ferreira, I.A.T.M., Gayed, S., Jahun, A., Hosmillo, M., Rees-Spear, C., Mlcochova, P., Lumb, I.U., et al. (2021). SARS-CoV-2 evolution during treatment of chronic infection. Nature *592*, 277–282. https://doi.org/10.1038/s41586-021-03291-y.

66. Clark, S.A., Clark, L.E., Pan, J., Coscia, A., McKay, L.G.A., Shankar, S., Johnson, R.I., Brusic, V., Choudhary, M.C., Regan, J., et al. (2021). SARS-CoV-2 evolution in an immunocompromised host reveals shared neutralization escape mechanisms. Cell *184*, 2605–2617.e18. https://doi.org/10.1016/j.cell.2021.03.027.

67. Braun, K.M., Moreno, G.K., Wagner, C., Accola, M.A., Rehrauer, W.M., Baker, D.A., Koelle, K., O'Connor, D.H., Bedford, T., Friedrich, T.C., and Moncla, L.H. (2021). Acute SARS-CoV-2 infections harbor limited within-host diversity and transmit via tight transmission bottlenecks. PLoS Pathog. *17*, e1009849. https://doi.org/10.1371/journal.ppat.1009849.

68. Martin, M.A., and Koelle, K. (2021). Comment on "Genomic epidemiology of superspreading events in Austria reveals mutational dynamics and transmission properties of SARS-CoV-2. Sci. Transl. Med. *13*, eabh1803. https://doi.org/10.1126/scitranslmed.abh1803.

69. Nelson, C.W., Ardern, Z., Goldberg, T.L., Meng, C., Kuo, C.H., Ludwig, C., Kolokotronis, S.O., and Wei, X. (2020). Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic. Elife *9*, e59633. https://doi.org/10.7554/eLife.59633.

70. Keese, P.K., and Gibbs, A. (1992). Origins of genes: "big bang" or continuous creation? Proc. Natl. Acad. Sci. USA *89*, 9489–9493. https://doi.org/10.1073/pnas.89.20.9489.

71. Bartonek, L., Braun, D., and Zagrovic, B. (2020). Frameshifting preserves key physicochemical properties of proteins. Proc. Natl. Acad. Sci. USA *117*, 5907–5912. https://doi.org/10.1073/pnas.1911203117.

72. Pavesi, A. (2021). Origin, Evolution and Stability of Overlapping Genes in Viruses: A Systematic Review. Genes *12*, 809. https://doi.org/10.3390/genes12060809.

73. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. *5*, 536–544. https://doi.org/10.1038/s41564-020-0695-z.

74. Chan, J.F.W., Kok, K.H., Zhu, Z., Chu, H., To, K.K.W., Yuan, S., and Yuen, K.Y. (2020). Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg. Microbes Infect. *9*, 221–236. https://doi.org/10.1080/22221751.2020.1719902.

75. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature *579*, 265–269. https://doi.org/10.1038/s41586-020-2008-3.

76. Flynn, J.A., Purushotham, D., Choudhary, M.N.K., Zhuo, X., Fan, C., Matt, G., Li, D., and Wang, T. (2020). Exploring the coronavirus pandemic with the WashU Virus Genome Browser. Nat. Genet. *52*, 986–991. https://doi.org/10.1038/s41588-020-0697-z.

77. Redondo, N., Zaldívar-López, S., Garrido, J.J., and Montoya, M. (2021). SARS-CoV-2 Accessory Proteins in Viral Pathogenesis: Knowns and Unknowns. Front. Immunol. *12*, 708264. https://doi.org/10.3389/fimmu.2021.708264.

78. Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O'Meara, M.J., Rezelj, V.V., Guo, J.Z., Swaney, D.L., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature *583*, 459–468. https://doi.org/10.1038/s41586-020-2286-9.

79. Hachim, A., Kavian, N., Cohen, C.A., Chin, A.W.H., Chu, D.K.W., Mok, C.K.P., Tsang, O.T.Y., Yeung, Y.C., Perera, R.A.P.M., Poon, L.L.M., et al. (2020). ORF8 and ORF3b antibodies are accurate serological markers of early and late SARS-CoV-2 infection. Nat. Immunol. *21*, 1293–1301. https://doi.org/10.1038/s41590-020-0773-7.

80. Wang, X., Lam, J.Y., Wong, W.M., Yuen, C.K., Cai, J.P., Au, S.W.N., Chan, J.F.W., To, K.K.W., Kok, K.H., and Yuen, K.Y. (2020). Accurate Diagnosis of COVID-19 by a Novel Immunogenic Secreted SARS-CoV-2 orf8 Protein. mBio *11*, e02431-20. https://doi.org/10.1128/mBio.02431-20.

81. Jiang, H.W., Li, Y., Zhang, H.N., Wang, W., Yang, X., Qi, H., Li, H., Men, D., Zhou, J., and Tao, S.C. (2020). SARS-CoV-2 proteome microarray for global profiling of COVID-19 specific IgG and IgM responses. Nat. Commun. *11*, 3581. https://doi.org/10.1038/s41467-020-17488-8.

82. Hanage, W.P., and Russell, C.A. (2021). Partial immunity and SARS-CoV-2 mutations. Science *372*, 354. https://doi.org/10.1126/science.abi4727.

83. Saad-Roy, C.M., Morris, S.E., Metcalf, C.J.E., Mina, M.J., Baker, R.E., Farrar, J., Holmes, E.C., Pybus, O.G., Graham, A.L., Levin, S.A., et al. (2021). Epidemiological and evolutionary considerations of SARS-CoV-2 vaccine dosing regimes. Science *372*, 363–370. https://doi.org/10.1126/science.abg8663.

84. Cobey, S., Larremore, D.B., Grad, Y.H., and Lipsitch, M. (2021). Concerns about SARS-CoV-2 evolution should not hold back efforts to expand vaccination. Nat. Rev. Immunol. *21*, 330–335. https://doi.org/10.1038/s41577-021-00544-9.

85. Matchett, W.E., Joag, V., Stolley, J.M., Shepherd, F.K., Quarnstrom, C.F., Mickelson, C.K., Wijeyesinghe, S., Soerens, A.G., Becker, S., Thiede, J.M., et al. (2021). Cutting Edge: Nucleocapsid Vaccine Elicits Spike-Independent SARS-CoV-2 Protective Immunity. J. Immunol. *207*, 376–379. https://doi.org/10.4049/jimmunol.2100421.

86. Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. methods *9*, 671–675. https://doi.org/10.1038/nmeth.2089.

87. Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189–1191. https://doi.org/10.1093/bioinformatics/btp033.

88. Haste Andersen, P., Nielsen, M., and Lund, O. (2006). Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. *15*, 2558–2567. https://doi.org/10.1110/ps.062405906.

89. Jespersen, M.C., Peters, B., Nielsen, M., and Marcatili, P. (2017). BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. Nucleic Acids Res. *45*, W24–W29. https://doi.org/10.1093/nar/gkx346.

90. De Cáceres, M., Legendre, P., and Moretti, M. (2010). Improving indicator species analysis by combining groups of sites. Oikos *119*, 1674–1684. https://doi.org/10.1111/j.1600-0706.2010.18334.x.

91. Clamp, M., Cuff, J., Searle, S.M., and Barton, G.J. (2004). The Jalview Java alignment editor. Bioinformatics *20*, 426–427. https://doi.org/10.1093/bioinformatics/btg430.

92. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. *32*, D115–D119. https://doi.org/10.1093/nar/gkh131.

93. Zheng, W., Zhang, C., Li, Y., Pearce, R., Bell, E.W., and Zhang, Y. (2021). Folding non-homologous proteins by coupling deep-learning contact maps with I-TASSER assembly simulations. Cell Rep. Methods *1*, 100014. https://doi.org/10.1016/j.crmeth.2021.100014.

94. Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. F1000Res. *5*, 189. https://doi.org/10.12688/f1000research.7931.1.

95. De Maio, N., Walker, C., Borges, R., Weilguny, L., Slodkowicz, G., and Goldman, N. (2020). Issues with SARS-CoV-2 Sequencing Data. Virological.

96. Andersen, K.G., Shapiro, B.J., Matranga, C.B., Sealfon, R., Lin, A.E., Moses, L.M., Folarin, O.A., Goba, A., Odia, I., Ehiane, P.E., et al. (2015). Clinical Sequencing Uncovers Origins and Evolution of Lassa Virus. Cell *162*, 738–750. https://doi.org/10.1016/j.cell.2015.07.020.

97. Spielman, S.J., and Wilke, C.O. (2015). The relationship between dN/dS and scaled selection coefficients. Mol. Biol. Evol. *32*, 1097–1108. https://doi.org/10.1093/molbev/msv003.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Goat anti-Human IgG Fc Cross-Adsorbed Secondary Antibody, DyLight™ 680	ThermoFisher Scientific	Cat#SA5-10138; RRID: AB_2556718
Goat anti-Human IgA (alpha chain) Antibody DyLight™800 Conjugated	Rockland Immunochemicals	Cat#609-145-006
Chemicals, peptides, and recombinant proteins		
Phosphate Buffered Saline with 0.05% TWEEN®20, pH 7.4	Millipore Sigma	Cat#PPB005-20PAK
Rockland Blocking Buffer	Rockland Immunochemicals	Cat#MB-070
TRIS	ThermoFisher Scientific	Cat#17926
Deposited data		
Single nucleotide variants from whole genome sequences	NCBI sequence read archive	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/4ZXDW0
Proteome of seasonal coronaviruses	NCBI	Table S12
Software and algorithms		
ImageJ	Schneider et al. (2012)[86]	https://imagej.nih.gov/ij/index.html
PepSlide Analyzer	SICASYS Software GmbH	https://www.sicasys.de/pepslide/
Jalview	Waterhouse et al. (2009)[87]	https://www.jalview.org/
Pymol	The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.	https://pymol.org/2/
DiscoTope	Andersen et al. (2006)[88]	http://tools.iedb.org/discotope/
BepiPred-2.0	Jespersen et al. (2017)[89]	https://services.healthtech.dtu.dk/services/BepiPred-2.0
R Software Version 4.2.1		https://www.r-project.org/
R package *indicspecies*	De Caceres et al. (2010)[90]	https://cran.r-project.org/web/packages/indicspecies/index.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jörg Fritz (jorg.fritz@mcgill.ca).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- All data reported in this paper will be shared by the lead contact upon request.

- Code is publicly available at https://github.com/arnaud00013/SARS-CoV-2-HPDA-evolutionary-analysis.

- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

- Additional Supplemental Items are available from Mendeley Data at https://doi.org/10.17632/fbs5k97hkz.1.

EXPERIMENTAL ANALYSIS AND STUDY PARTICIPANTS DETAILS

Human participants

Recruitment of patients at the San Martino University Hospital (Genoa, Italy) was approved by the Institutional Review Board at Genoa University, approved by the Ethics Committee of Liguria Region (Comitato Etico Regione Liguria; N. CER Liguria 114/2020–ID 10420) and carried out in accordance with the principles of the Declaration of Helsinki. Positivity of SARS-CoV-2 infection was assessed both by PCR and measurement of specific antibodies (Cobas-Roche using Elecsys Anti-SARS-CoV-2 S). All patients gave their consent for participation in this study. Negative (3 male, 2 female, mean age 47 [27–66]) and asymptomatic (1 male, 4 female, mean age 41 [28–56]) individuals were health workers who were tested regularly in the hospital and classified according to serological and molecular tests for COVID-19 (Table S1). Recovered individuals (convalescent post-infection) (3 male, 2 female, mean age 55 [42–67]) were all patients previously admitted at the hospital due to lung pneumonia and were found to be positive to COVID-19, having severe (n = 2) and mild (n = 3) disease. Sera were collected according to standard procedures, by centrifugation.

METHOD DETAILS

High-density peptide array (HDPA)

To analyze the antibody responses to SARS-CoV-2 at the epitope level we used a recently developed high-density peptide array (HDPA), the PEPperCHIP Microarray (PEPperPRINT GmbH, Germany), covering the proteome of the SARS-CoV-2 isolate Wuhan-Hu-1 as well as the four seasonal hCoVs OC43, HKU1, NL63 and 229E (see Table S12 for accession numbers used). The protein sequences of ORF1A/B, Spike (S) protein, Envelope (E) protein, Membrane (M) glycoprotein, Nucleocapsid (N) phosphoprotein were translated into 15 amino acid peptides with a peptide overlap of 13 amino acids. This results in 27,540 individual peptides, which were printed in duplicates (55,080 spots). In addition, to ensure sensitivity controls of the PEPperCHIP HDPA, positive controls were included to probe for antibody reactivity for influenza hemagglutinin (HA; YPYDVPDYAG, 360 spots) and polio virus (KEVPALTAVETGAT, 355 spots). These additional HA and polio peptides framing the microarrays were simultaneously stained as internal quality control to confirm assay performance and peptide microarray integrity. With this setup per chip, 15 samples (see Table S1) were analyzed.

At first, the peptide microarrays were incubated for 15 min in phosphate buffered saline supplemented with 0.05% Tween 20 (PBS-T, pH 7.4) and blocked for 30 min with Rockland Blocking Buffer (RL) (Rockland Immunochemicals) at room temperature. Prior to immunoassay, sera of patients were first heat-inactivated at 56°C for 30 min, and then the microarrays were incubated at serum dilutions of 1:500, 1:100 and 1:20 in 10% RL/PBS-T overnight at 4°C with orbital shaking. Microarrays were then washed (three times with PBS-T for 1 min) and peptide binding was detected with isotype-specific secondary goat anti-human IgG (Fc) DyLight680 (ThermoFisher Scientific) and goat anti-human IgA (alpha chain) DyLight800 (Rockland Immunochemicals) antibodies at a final concentration of 0.1–1 μg/mL, respectively (in 10% RL/PBS-T for 45 min at room temperature). Subsequent washing (three times with PBS-T for 1 min) was followed by dipping the microarrays in 1 mM TRIS pH 7.4 followed by drying with pressurized air. Acquisition of images was done using an LI-COR Odyssey CLx Infrared Imaging System (scanning offset 0.65 mm, resolution 21 μm). Data quantification and analysis was based on the assays at dilution 1:20. Using ImageJ software the resulting 32-bit gray-scale TIFF files were converted into 16-bit gray-scale TIFF files and then further analyzed using the PepSlide Analyzer (SICASYS Software GmbH). The in house developed PEPperPRINT software algorithm was used to calculate median foreground intensities (background-corrected intensities) of each spot and spot-to-spot deviations of spot duplicates. A maximum spot-to-spot deviation of 40% was tolerated, otherwise, the corresponding intensity values were zeroed. To complement this analysis, acquired microarray scans were reassessed with respect to artifacts by visual inspection, and erroneous values were corrected manually. Based on averaged median foreground intensities, intensity maps were generated and interactions in the peptide maps highlighted by an intensity color code with red (IgG) or green (IgA) for high and white for low spot intensities. To identify the top IgG and IgA antibody responses of the human serum samples, the averaged intensity values were sorted by decreasing spot intensities. We further plotted averaged spot intensities of the assay against the peptide microarray content from the N-terminus of Spike (SARS-CoV-2) to the C-terminus of ORF1AB (HUK1) to visualize overall spot intensities and signal-to-noise ratios. The intensity plot was correlated with the peptide and intensity map as well as with visual inspection of the microarray scans to identify the main antibody responses of the human sera. In

general, relative fluorescent units (RFU) of equal to or higher than 100 was considered a positive antibody response. However, as mentioned in the results sections, to focus on the most prominent antibody responses, our analysis was performed with an RFU cut-off of 1000 or higher.

Defining cross-reactivity using protein conservation and immune response to endemic human coronaviruses

To find epitope sites associated with cross-reactivity, we first calculated the conservation of peptide sequences across endemic hCoVs (HKU1, NL63, OC43 or 229E). To do so, we aligned the reference sequences of the hCoVs (proteome) in Jalview and extracted the conservation score (Cscore).[91] This conservation score reflects the conservation of physico-chemical properties in the alignment, where identical residues score the highest.[49] Epitope sites with a conservation score ≥6 and for which we detected antibody responses for both SARS-CoV-2 and at least one of the endemic hCoVs were considered as cross-reactive epitope sites. To find epitope sites associated with cross-reactivity, we first calculated the conservation of peptide sequences across endemic hCoVs (HKU1, NL63, OC43 or 229E). To do so, we aligned the reference sequences of the hCoVs (proteome) in Jalview and extracted the conservation score (Cscore) (86).[91] This conservation score reflects the conservation of physico-chemical properties in the alignment, where identical residues score the highest (49).[49] Epitope sites with a conservation score ≥6 and for which we detected antibody responses for both SARS-CoV-2 and at least one of the endemic hCoVs were considered as cross-reactive epitope sites. To find epitope sites associated with cross-reactivity, we first calculated the conservation of peptide sequences across endemic hCoVs (HKU1, NL63, OC43 or 229E). To do so, we aligned the reference sequences of the hCoVs (proteome) in Jalview and extracted the conservation score (Cscore).[91] This conservation score reflects the conservation of physico-chemical properties in the alignment, where identical residues score the highest.[49] Epitope sites with a conservation score ≥6 and for which we detected antibody responses for both SARS-CoV-2 and at least one of the endemic hCoVs were considered as cross-reactive epitope sites.

Detecting epitopes that are significantly more prevalent in SARS-CoV-2 positive patients

First, antibody responses to each linear 15-mer peptide were mapped across the SARS-CoV-2 proteome and average RFU calculated for each amino acid residue.

Second, the normalized positional 'epitope coverage' at each amino acid residue within the proteins was defined as the ratio of total peptides mapped to each position by the total expected peptides, with values ranging between 0 and 1. A value of 1 in the SARS-CoV-2-positive group means that amino acid residues within the proteins were covered by peptides that showed immune response in all 10 SARS-CoV-2-positive patients and 14 peptides that overlap that position. (14 × 10 = 140 is the theoretical expected positional coverage to be 100%). Similarly, a value of 1 in SARS-CoV-2-negative group is 70 peptides with response (14 peptides x 5 SARS-CoV-2-negative patients = 70. i.e., all 70 unique peptides that cover residue locations).

Third, to identify the epitopes that are particularly prevalent in SARS-CoV-2-positive subjects, we performed an indicator value analysis.[50] This type of analysis is frequently used in ecology to determine whether species have significant associations with certain site groups. We applied this method to epitopes presence/absence data by replacing species with epitopes and site groups with patient groups defined by SARS-CoV-2 PCR status (positive or negative). The indicator value analysis measures the IndVal metric, which is the product of the specificity (e.g., the proportion of individuals within the whole dataset that exhibit a response to the epitope and belongs to a certain patient group) and the fidelity (e.g., the proportion of individuals within a certain patient group that exhibits a response) to the epitope. To control for the differences in sample size between patient groups, we used the group-equalized version of IndVal, IndVal-pag.[50] The R function *multipatt* from the R package *indicspecies* allowed us to perform this analysis and evaluate the significance of the associations through permutation tests.[50]

Structural properties of B cell epitopes and B cell epitope prediction

SARS-CoV-2 protein sequences were obtained from Uniprot.[92] Structure models of all 24 proteins in SARS-CoV-2 were obtained from I-TASSER.[93] Solvent accessibility was calculated using freeSASA.[94] Pymol was used for visualization. Using 3D structures and biophysical properties of the SARS-CoV-2 proteome, we applied the DiscoTope algorithm[35] to computationally predict conformational B cell epitopes with a significance threshold of −7.7 (75% specificity, 45% sensitivity). In addition, we used the Bepipred algorithm[36]

to obtain linear B cell epitopes. Epitopes with minimum length of 7 amino acid residues and minimum score of 0.55 (80% specificity, 30% sensitivity) were used for the analysis.

Epitope evolution profiling

To understand the evolution of SARS-CoV-2 epitopes in SARS-CoV2-positive patients, we made use of single nucleotide variants (SNVs) from 38,685 whole genome sequences from the NCBI sequence read archive (Table S9, see https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/4ZXDW0). We selected SARS-CoV-2 samples from the first pandemic wave (defined as January 1 to July 31, 2020) and the second wave (defined as August 1 to December 31, 2020) sequenced using Illumina paired-end amplicons with a minimum average depth of coverage of 200 x and fewer than 10,000 sites with a depth of coverage lower than 100x. We then retained single nucleotide variants present in both minus and plus strands at a minimum frequency of 2%, occurring at sites with a minimum depth of 100x, having a minimum within-sample frequency of 5% and located between sites 101 and 29778 of the genome to exclude sites at the extremities that are prone to sequencing errors and have been frequently masked.[95] These additional filters allowed us to remove sequencing errors and provided deep coverage to identify SNVs that are polymorphic within patients, reflecting within-patient evolution,[51,52] as well as those that are shared between the consensus sequences of different patients.

Next, to compare epitope evolution from the evolution of non-epitope sites of the same protein, we measured the evolution rates at within-host (SNVs with a frequency <75% that are not transmitted for sure) and between-host/transmission level (SNVs with a frequency ≥75% that are observed in at least three samples). Because the number of SNVs observed will vary depending on sample coverage, which varies across samples, we estimated the evolution rates in each sample separately using the number of SNVs observed per site with adequate coverage. Such sites are defined as having a detection power of at least 80%, which is the probability of detecting five reads supporting the presence of a SNVs with a frequency of at least 5% in a site of coverage C, i.e., the minimum adequate coverage, under a binomial distribution. This approach has been used previously for similar purposes with the Lassa virus.[96]

We also inferred selection in the proteins of interest (ORF1A/B, Spike (S) protein, Envelope (E) protein, Membrane (M) glycoprotein, Nucleocapsid (N) phosphoprotein) using dN/dS, the ratio of non-synonymous (dN) and synonymous substitutions rates (dS), which we calculated from the called SNVs in each sample.[97]

$$dN / dS = (Nb_{nsub} / Nb_{nss}) / (Nb_{ssub} / Nb_{ss}) \qquad \text{Equation 1}$$

where Nb_{nsub} is the number of non-synonymous substitutions, Nb_{nss} is the number of non-synonymous sites, Nb_{ssub} is the number of synonymous substitutions, and Nb_{ss} is the number of synonymous sites. dN/dS can detect purifying selection (dN/dS < 1), neutral evolution (dN/dS ≈ 1) and positive selection (dN/dS > 1). In each sample, we calculated dN/dS only if there were more than three SNVs including at least one synonymous SNV.

Finally, we inferred selection at the within-host level, using pN/pS, which we calculated from intrahost SNVs (iSNVs), i.e., SNVs that are not fixed (within-sample frequency <75%):

$$pN / pS = (Nb_{nmut} / Nb_{nss}) / (Nb_{smut} / Nb_{ss}) \qquad \text{Equation 2}$$

where Nb_{nmut} is the number of non-synonymous iSNVs, Nb_{nss} is the number of non-synonymous sites, Nb_{smut} is the number of synonymous iSNVs, and Nb_{ss} is the number of synonymous sites. These analyses have been implemented in R (https://github.com/arnaud00013/SARS-CoV-2-HPDA-evolutionary-analysis).

Selection for immune escape in VOCs and VUIs genomes

To reveal VOCs and VUIs mutations possibly involved in selection for immune escape, we first defined the signature mutations of each variant (Table S10) as substitutions that are present in ≥90% of sequences assigned to that lineage. We calculated the prevalence of substitutions in thousands of publicly available consensus sequences collected during 2020 and added data from CoV-Spectrum about under-represented lineage in the database or lineages that emerged during 2021.[57] Then, we only focused on nonsynonymous signature mutations in our database and asked if these signature mutations are located at epitope sites as these mutations can change the antibodies' ability to recognize the epitopes. The signature mutation prevalence data were collected from our database of NCBI samples for the earlier lineages

(PANGO v.2.1.7) and from GISAID data obtained from CoV-spectrum for more recent lineages like Omicron. The database of lineage signature mutations is available on Github (https://github.com/arnaud00013/SARS-CoV-2-HPDA-evolutionary-analysis).

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical details of experiments can be found in the figure legends. Findings were considered significant with a p value of less than 0.05. All analyses were performed with R Software Version 4.2.1.

ADDITIONAL RESOURCES

R scripts used for analysis are publicly available at https://github.com/arnaud00013/SARS-CoV-2-HPDA-evolutionary-analysis.

Whole genome sequences used for evolution profiling are publicly available. See https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/4ZXDW0.