Metaviromic identification of discriminative genomic features in SARS-CoV-2 using machine learning

Jonathan J. Park, Sidi Chen

PII: S2666-3899(21)00281-6

DOI: https://doi.org/10.1016/j.patter.2021.100407

Reference: PATTER 100407

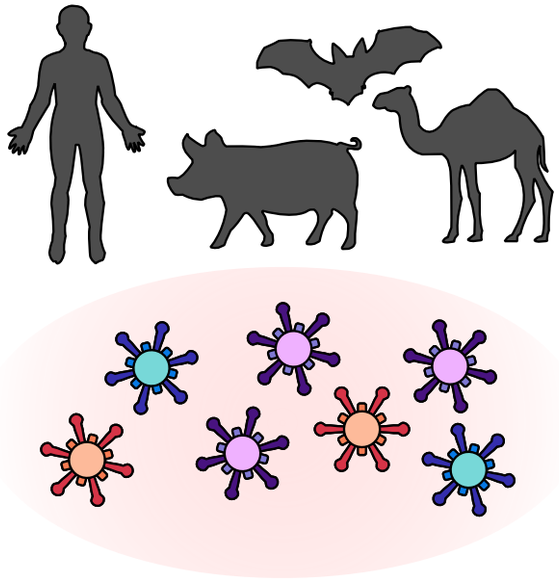
To appear in: *Patterns*

Received Date: 13 July 2021

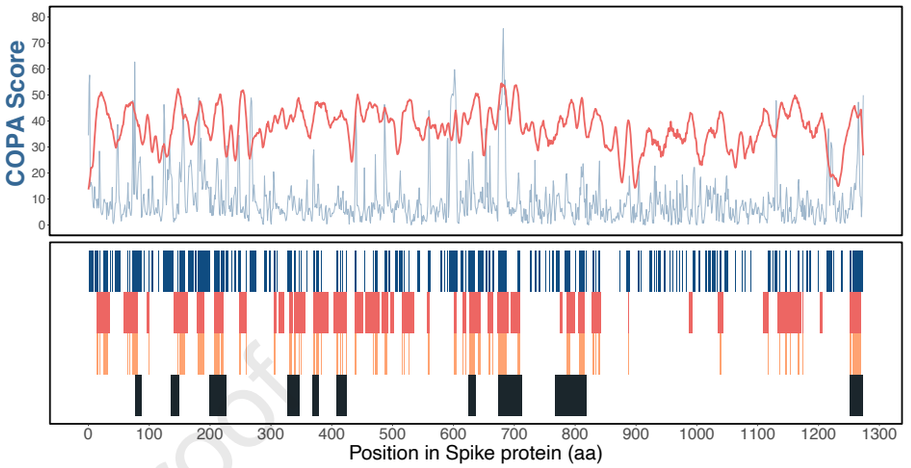
Revised Date: 12 August 2021

Accepted Date: 11 November 2021

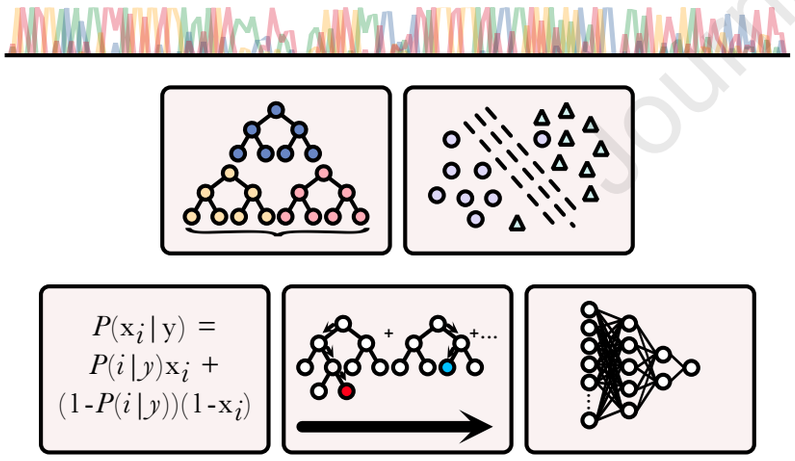
Please cite this article as: Park JJ, Chen S, Metaviromic identification of discriminative genomic features in SARS-CoV-2 using machine learning, *Patterns* (2021), doi: https://doi.org/10.1016/j.patter.2021.100407.



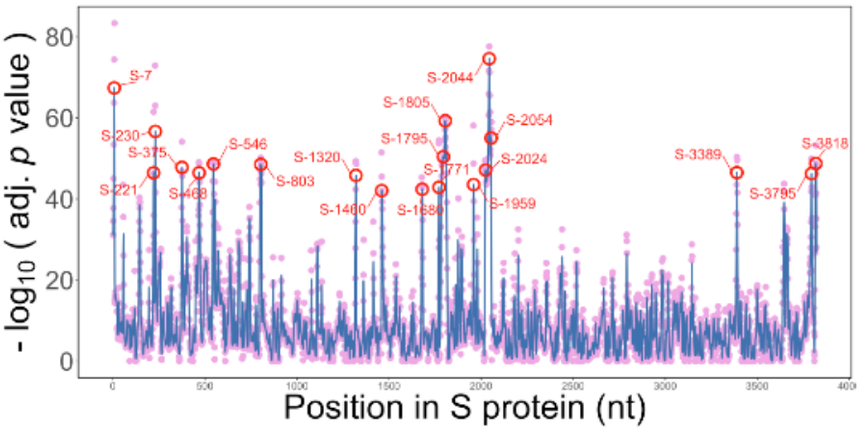
Coronavirus genomes



Integrative analysis



Machine learning models



Discriminative genomic features

Metaviromic identification of discriminative genomic features in SARS-CoV-2 using machine learning

Jonathan J. Park [1,2,3,4] and Sidi Chen [1,2,3,4,5,6,7,8,9,10,11,12,13,14,#]

Affiliations

1. Department of Genetics, Yale University School of Medicine, New Haven, Connecticut, USA
2. System Biology Institute, Yale University, West Haven, Connecticut, USA
3. Center for Cancer Systems Biology, Yale University, West Haven, Connecticut, USA
4. M.D.-Ph.D. Program, Yale University, West Haven, Connecticut, USA
5. Immunobiology Program, Yale University, New Haven, Connecticut, USA
6. Molecular Cell Biology, Genetics, and Development Program, Yale University, New Haven, Connecticut, USA
7. Combined Program in the Biological and Biomedical Sciences, Yale University, New Haven, Connecticut, USA
8. Department of Neurosurgery, Yale University School of Medicine, New Haven, Connecticut, USA
9. Comprehensive Cancer Center, Yale University School of Medicine, New Haven, Connecticut, USA
10. Stem Cell Center, Yale University School of Medicine, New Haven, Connecticut, USA
11. Liver Center, Yale University School of Medicine, New Haven, Connecticut, USA
12. Center for Biomedical Data Science, Yale University School of Medicine, New Haven, Connecticut, USA
13. Center for RNA Science and Medicine, Yale University School of Medicine, New Haven, Connecticut, USA
14. Corresponding author and lead contact

Correspondence:
SC (sidi.chen@yale.edu)
+1-203-737-3825 (office)
+1-203-737-4952 (lab)

Summary:

The COVID-19 pandemic caused by SARS-CoV-2 has become a major threat across the globe. Here, we developed machine learning approaches to identify key pathogenic regions in coronavirus genomes. We trained and evaluated 7,562,625 models on 3,665 genomes including SARS-CoV-2, MERS-CoV, SARS-CoV and other coronaviruses of human and animal origins to return quantitative and biologically interpretable signatures at nucleotide and amino acid resolutions. We identified hotspots across the SARS-CoV-2 genome including previously unappreciated features in spike, RdRp and other proteins. Finally, we integrated pathogenicity genomic profiles with B cell and T cell epitope predictions for enrichment of sequence targets to help guide vaccine development. These results provide a systematic map of predicted pathogenicity in SARS-CoV-2 that incorporates sequence, structural and immunological features, providing an unbiased collection of genetic elements for functional studies. This metavirome-based framework can also be applied for rapid characterization of new coronavirus strains or emerging pathogenic viruses.

Introduction

The coronavirus disease 2019 (COVID-19) pandemic caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has become an unprecedented on-going global public health and economic crisis since its emergence at the end of the year 2019 [1,2]. The SARS-CoV-2 virus has infected more than 180 million people and caused over 3.9 million deaths globally as of July 1st, 2021 [3]. Although pathogenic coronaviruses have repeatedly emerged from the wild to become infectious to human populations, the common genetic and molecular features that drive the disease-causing potential of these viruses are still unclear. Identifying genetic elements and specific regions of the SARS-CoV-2 genome that make it dangerous is critical for public health prevention and disease mitigation, as well as the development of vaccines and therapeutics.

Machine learning (ML) methods have become important for the interpretation of large and complex genomic data sets [4], and have been used in a variety classification tasks including transcription start site recognition [5], gene expression prediction [6], or complex disease phenotype prediction [7]. Given the large scale of viral genome datasets and potential for ML methods to recognize patterns in DNA sequences, such methods are well suited for the classification task of identifying pathogenicity-associated genomic features in coronaviruses. We therefore developed a set of ML approaches focused on unbiased scanning and scoring of key pathogenicity-linked regions in the genomes of SARS-CoV-2 and other high case fatality rate (CFR) coronaviruses [8] that distinguish them from other coronaviruses strains.

There are a number of challenges when setting up ML models for sequence-based classification tasks as performed in this study. First, because we were comparing genomes from different coronavirus species which have different lengths (**Figure S1A**), there must be a way to standardize sequence inputs in a way that conserves information on the evolutionary relationship between species. The comparative genomics approach for doing so is by multiple sequence alignment. Second, since ML methods typically require numerical inputs, we encoded the categorical alignment data into integer representation using one-hot encoding. Third, since we were interested in identifying specific local genomic regions that were predictive for coronavirus pathogenicity, we partitioned the alignment into smaller sliding windows for training and evaluation of the ML models. Fourth, there is the limited experimental data available on characterizing the pathogenicity of genomic sequences for coronaviruses. For example, our group and collaborators have identified nonstructural protein 1 (Nsp1) through an ORF mini-screen as a key protein that causes reduction of host cell viability [9] and Gordon et al. mapped physical interactions between human host proteins and SARS-CoV-2 proteins [10]; however, these studies are limited to the scale of whole open reading frames. To address the challenge of defining labels, we use evolution and species-based annotations comparable to the approach of other groups [11]. Fifth, many ML techniques applied to genomic sequencing data use an arbitrary accuracy threshold for determining significance. We utilized ML-derived accuracy scores as a proxy for "learned, predictive information content" and developed a statistically rigorous meta-model based on the hypothesis that highly gapped alignment regions should not be predictive of coronavirus pathogenicity. Sixth, in order to demonstrate the biological significance and utility of the scores obtained by our pipeline, we performed comprehensive evolutionary, structural, immunological, and emerging variant of concern analyses.

These methods provided us with highly quantitative, and biologically interpretable, coronavirus pathogenicity (COPA) scores for every nucleotide in the SARS-CoV-2 genome. We believe that the ML-based approach developed here can be generally applied for functional genomic characterization of novel viruses across the metavirome, such as new coronavirus strains, new emerging pathogenic viruses, or other pathogenic microbials, where traditional analytic methods are limited.

Results

High-CFR coronavirus strains have shared genomic features that distinguish them from other coronaviruses.

We hypothesized that the increased pathogenicity of high-CFR coronavirus strains is due in part to shared genomic features that distinguish them from other coronaviruses. To test this hypothesis, we performed

principal component analyses (PCA) on encoded representations of the coronavirus genomes used in the study. We aligned 3,665 Coronaviridae family genomes obtained from the Virus Pathogen Database and Analysis Resource (ViPR) database (Pickett et al., 2012) with diverse taxonomic and host features (**Figure 1B**), then performed one-hot encoding of the entire genomes, followed by PCA. Alphacoronaviruses and betacoronaviruses typically cause respiratory illness in humans or gastroenteritis in birds, while gammacoronaviruses and deltacoronaviruses typically infect birds. Although low-CFR coronaviruses that infect humans (e.g. HCoV-NL63, HCoV-229E, HCoV-OC43 and HKU1) span both alpha- and betacoronaviruses, the highly pathogenic, high-CFR strains that infect humans (e.g. MERS-CoV, SARS-CoV, and SARS-CoV-2) are betacoronaviruses [8]. Of note, some low-CFR strains can still cause to severe infections in children, elderly, or immunocompromised patients [12]. Visualizations from our PCA analyses revealed that coronavirus genomes can cluster by genus, host, and species (**Figure S1G**). Specifically, alpha-, beta-, and gamma- coronaviruses were clearly segregated in the first four principal components, and genomes further clustered by host (e.g. human hosts in betacoronaviruses for PC1 and PC2) and species (e.g. avian coronavirus in gammacoronaviruses for PC3 and PC4). In order to see if high-CFR virus genomes (MERS-CoV, SARS-CoV, and SARS-CoV-2) also cluster after dimensionality reduction, we labelled the genomes accordingly (1 representing high-CFR genomes, 0 representing all other genomes, **Figure S1H**) and observed that the high-CFR genomes clustered together along with associated features such as betacoronaviruses or their respective species.

Machine learning and statistical meta-model identifies high-resolution discriminative features in coronavirus genomes.

We then developed a rigorous, integrative ML-based approach to identify regions that contribute to coronavirus pathogenicity, incorporating Random Forests (RF), Support Vector Machines (SVM), Bernoulli Naïve Bayes (BNB), Gradient Boosting Classifiers (GBC) and Multi-layer Perceptron Classifiers (MLPC) (**Experimental Procedures**) (**Figure 1A**). We chose a set of five different supervised learning algorithms that have robust performance and represent methodological diversity including ensembles of decision trees, Bayes' theorem, and neural networks. We then trained and evaluated 7,562,625 advanced ML models (see **Experimental Procedures** for further details) on six bp-wide sliding windows with stratified fivefold cross-validation across the aligned coronavirus genomes for different predictors based on the classification strategy. To set up our predictor classes, we established several classification strategies to capture signatures associated with pathogenicity (**Figure 1C**). We considered that sequence features that enable coronaviruses to jump from animal populations to humans (strategy A) and that distinguish SARS-CoV, MERS-CoV and SARS-CoV-2 from other coronaviruses (strategy B) to likely be important

contributors to pathogenicity. We also considered features that specifically distinguish SARS-CoV, MERS-CoV and SARS-CoV-2 that infect human hosts (strategy C) from all other coronaviruses. To highlight the evolutionary relationship between the samples in relation to our classification strategies, we have run a set of phylogenetic analyses on the genomes used for training our ML models (**Figure 1E** and **Figure S1I**). Consistent with our PCA analyses, we observe that the predictor class genomes across the different strategies have evolutionary proximity, and have overlap with Sarbecoviruses and Merbecoviruses. After training and evaluating our base ML models on windows tiled across the alignment (100,835 windows), we integrated performance accuracy scores into a statistically rigorous meta-model based on minimum entropy windows (**Experimental Procedures, Figures S1B-F**) to obtain biologically interpretable nucleotide-level coronavirus pathogenicity (NT-COPA) scores for every nucleotide in the SARS-CoV-2 genome (**Figure 1D**). To be specific, for a given window, each of the individual classifiers were trained and tested for each fold permutation (using 4 folds for training, 1 fold for testing), yielding 5 accuracy scores for each classifier. These scores were used as a surrogate for how well the high-CFR genomes can be differentiated from other genomes at that particular position. That is, the higher the score, the better the predictive performance of the classifier, the better that particular genomic region can distinguish high-CFR genomes. All of these scores are then tested against scores from the minimal entropy control group using the two-sided Wilcoxon rank-sum test, adjusted for false discovery rate, and then negative log10 transformed to obtain NT-COPA score.

Because the samples in the standard dataset are ordered by alignment, individual models for different cross-validation folds may have dissimilar training compositions and therefore accuracy scores (**Figure S2A-D**); however, this tradeoff may come with a greater diversity of biologically meaningful learned features. Since we pool together all cross-validation scores with training coverage of all samples, no genomic information is lost for our statistical meta-model analyses. We focused our subsequent analyses on results obtained from this pipeline for biological interpretability, and provide the NT-COPA scores as a resource.

Identification of local discriminative hotspots in SARS-CoV-2 proteins.

Next, we looked at NT-COPA score distributions intersected with the annotated SARS-CoV-2 genome to see if they can be used to identify potential discriminative hotspots. We found that the NT-COPA scores reflected quantitative and high-resolution signatures for characterizing individual base pairs and amino acids within SARS-CoV-2 features (**Figure 2A**). In order compare the NT-COPA scores with more naïve methods, we obtained the consensus score for each position in the multiple sequence alignment, with values corresponding the percentage identity to the consensus sequence. A score of 100 would therefore

correspond to 100% identity to the consensus sequence, which in turn means there is no sequencing variation across all viruses at that position. We plotted the consensus scores against the NT-COPA score, performed linear regression analyses, and found there to be a negative linear relationship with a significant model p-value (**Figure S2E**). We expect our NT-COPA scores to be higher with increased diversity at a given position, since this in turn corresponds to increased information for better ML model performance. Therefore, a negative relationship between the consensus and NT-COPA scores are in line with our expectations, as increased consensus scores correspond to decreased diversity at a given position.

To address the challenge of systematically defining hotspot regions from such high-resolution data, we considered that scores for a given base pair should reflect local genomic information capture due to our sliding window based approach for training the base ML models. Therefore, we considered kernel smoothing to be an appropriate nonparametric curve estimation method for a region-based approach to identify hotspots (**Experimental Procedures**). We calculated the kernel regression estimate at each base pair using the NT-COPA scores, and used the estimates to determine local signal maxima (peaks) within SARS-CoV-2 features (**Figure 2A, Figure S3**). This approach yielded 2,473 peaks across the SARS-CoV-2 genome, which mark local discriminative hotspots (**Figure 2B**). Limitations to the kernel smoothing based approach include that identified peaks may have relatively low NT-COPA scores as they only reflect local maxima (which may be addressed by using score thresholds), and that high signal density regions may only return a single peak. The advantage of the approach is that it is unbiased and systematic. Both systematic and customized strategies can be applied to generate biologically meaningful insights from these pathogenicity-associated scores.

Spike protein hotspots reveal a furin cleavage site and contact sites with ACE2.

To biologically validate the significance of our candidate hotspots, we performed a series of in-depth evolutionary and structural analyses. There has been considerable focus on the spike protein as it facilitates coronavirus entry into target cells [13,14]. For SARS-CoV-2, interaction between the trimeric spike glycoprotein and the human host ACE2 receptor triggers a cascade of events that leads to the fusion between cell and viral membranes [15]. We examined the NT-COPA score distributions and peaks for the spike protein and found the strongest signal to be peak S-2044, corresponding to amino acid position 682 (**Figure 2A**). In order to determine the evolutionary significance of this hotspot, we aligned the spike protein amino acid sequences for Coronaviridae family viruses across various species and hosts and compared the alignment with the NT-COPA score density for peak S-2044 and nearby residues (**Figure 3B**).

We find that this peak corresponds to a functional polybasic furin cleavage site (RRAR) at the junction between the S1/S2 subunits, which has been reported to expand SARS-CoV-2 tropism and/or enhance its infectivity [16]. The leading proline that is also inserted at the site for SARS-CoV-2 (for PRRA insertion) has been shown to result in addition of O-linked glycans to S673, T678 and S686 which flank the cleavage site by structural analysis [17]. Nucleotides for the T678 codon and the first nucleotide for the S686 codon (corresponding to position S-2056) all have high NT-COPA scores and are included as part of the peak S-2044 associated hotspot. More generally, this hotspot, which spans nucleotide positions 2,021 to 2,056 (amino acid positions 674 to 686, all with NT-COPA scores > 40), corresponds to amino acid insertions that contribute to distinguishing betacoronaviruses from alpha- and gammacoronaviruses (**Figure 3A-B**). The functional consequence of the polybasic cleavage site and the predicted O-linked glycans in SARS-CoV-2 remains unclear, although possibilities for the latter include creation of mucin-like glycan shields involved in immune evasion [17]. These analyses showed that the ML-based approach independently learned pathogenicity signals that correspond to important features of the SARS-CoV-2 genome, several of which have been previously validated or are under active investigation.

To see if ML-scored discriminative hotspots can offer functionally significant structural insights, we examined the spike protein receptor-binding domain (RBD) interface with ACE2. We calculated the amino acid resolution COPA scores (AA-COPA, or COPA for short) by averaging the NT-COPA scores for codons. We then examined the high AA-COPA regions in the spike protein RBD, and identified two hotspot regions comprising residues NNL at positions 439-441 and residues NCYF at positions 487-490 (**Figure 3C**). We then mapped the COPA scores onto a recently solved crystal structure of the wild-type SARS-CoV-2 RBD bound to human ACE2 [15], and found that the NCYF hotspot included contact site residues at the RBD-ACE2 interface (**Figure 3D**). Of the 13 hydrogen bonds at the SARS-CoV-2 RBD - ACE2 interface identified from the wild-type structure (**Figure S4B**), 3 hydrogen bonds are included in the NCYF hotspot: N487-Q24, N487-Y83, and Y489-Y83. Notably, all three of these SARS-CoV-2 - ACE2 hydrogen bonds are conserved for the SARS-CoV RBD - ACE2 interface, as N473-Q24, N473-Q24, Y475-Y83 [15]. Both of the coronavirus contact site residues in the SARS-CoV-2 NCYF hotspot (N487 and Y489) are relatively conserved amongst proximal strains, but differ in less proximal strains (**Figure 3A, 3E**), suggesting that acquisition of these sites were important evolutionary events in development of high affinity coronavirus binding to the human ACE2 receptor. Interestingly, an alternative, chimeric RBD-engineered structure of the SARS-CoV-2 spike protein-ACE2 complex demonstrated that structural changes in one of the ridge loops that differentiate SARS-CoV-2 from SARS-CoV introduces an additional main-chain hydrogen bond between residues N487 and A475 in the SARS-CoV-2 receptor binding motif (RBM),

causing the ridge to form more contacts with the N-terminal helix of ACE2 [18]. The COPA-structural joint analysis suggested that the ML models automatically learned the SARS-CoV-2 NCYF hotspot as a proximally conserved contributor to coronavirus pathogenicity.

We then examined the other hotspot region identified in the SARS-CoV-2 RBD, comprising residues N439, N440, and L441. Residue N439 was not identified to be involved in contacts between SARS-CoV-2 RBD and ACE2 receptor in the wild-type structure (**Figure S4C**). However, its associated residue in the SARS-CoV RBD, R426, forms a strong salt bridge with E329 on ACE2 and a hydrogen bond with Q325 [15,18,19] (**Figure S4D**). Evolutionary analysis reveals that the NNL (SARS-CoV-2 coordinates) or RNI (SARS-CoV coordinates) hotspot has substantial sequence divergence from other coronaviruses across species and hosts (**Figure S4A**). While the significance for the NNL hotspot for SARS-CoV-2 is unclear, R426 is a functionally important residue for ACE2 receptor binding in SARS-CoV, and scored highly in the classification strategies focused on learning sequence determinants of pathogenicity that are generalizable across respiratory disease-causing coronaviruses.

RdRp hotspots reveal RNA contact sites and codon composition biases.

Another key component of the SARS-CoV-2 virus is the RNA-dependent RNA polymerase (RdRp), also known as nonstructural protein 12 (NSP12). RdRp/NSP12 forms a complex with accessory factors including NSP7 and NSP8, which increase template binding and processivity, to catalyze the synthesis of viral RNA [20,21]. As this complex plays an important role in the viral replication and transcription cycle, RdRp is currently being investigated as a target of nucleotide analog antiviral drugs such as remdesivir for COVID-19 treatment [22,23]. To identify discriminative hotspot regions in RdRp potentially associated with pathogenicity, we intersected its sequence with the ML-generated COPA scores (**Figure 4A**). We then mapped the COPA scores onto a recently solved cryo-EM structure of the SARS-CoV-2 NSP12-NSP7-NSP8 complex bound to template-primer RNA and the triphosphate form of remdesivir (RTP) [21] (**Figure 4B**). We focus on two structural regions of interest in SARS-CoV-2 RdRp with high COPA score signal density. Region (1), which comprises residues ERVRQ (positions 180-184) and DRY (positions 284-286), reflects a previously uncharacterized feature of RdRp with a high density of hydrophobic and hydrophilic amino acid residues. Whether the hotspot residues in region (1) create networks of hydrophilic interactions that contribute to pathogenicity require further experimental study; nevertheless, this region highlights discriminative features that were learned from the ML models in an unbiased manner. Region (2), which comprise residues K500, S501, W509, and I847, includes key residues involved in direct RdRp protein-RNA interactions. We observed that the identified COPA hotspot residues generally exhibit high amino

acid conservation amongst proximal strains and differentiation in less proximal strains (**Figure 4C and Figure S5B**), with a notable exception of residues K500 and S501.

We were surprised at this exception since initially it was unclear why our ML approach would assign residues K500 and S501 high COPA scores if these positions exhibit such strong evolutionary conservation across species and hosts, as these positions should then not be able to distinguish pathogenic coronaviruses. To examine these regions at the nucleotide resolution, we returned to our aligned genome used for training the base ML models, and generated nucleotide composition frequencies (presented as motifs of sequence logos) for codons associated with the hotspot residues (**Figure 4D and Figure S5A).** The ERVRQ motif reveals conservation amongst the pathogenic coronaviruses that differentiate them from the high diversity of non-pathogenic coronaviruses in this region. These results are expected given the goals of our methods. The KS motif, however, reveals codon composition bias that differentiate SARS-CoV-2 and SARS-CoV at the nucleotide level from MERS-CoV and non-pathogenic coronaviruses. This bias is particularly striking for residue S501, where all three nucleotides differentiate the SARS strains from other coronaviruses, despite conserving a serine residue. Whether these codon composition biases reflect selection, recombination, or more generalized codon usage biases require further study. Nevertheless, these results highlight the learned evolution signatures of critical features in the SARS-CoV-2 genome at different levels.

Integration of genomic discriminative profiles with B cell and T cell immunogenic features.

Although a few vaccine candidates have been approved for SARS-CoV-2 (e.g. Moderna, Pfizer/BioNtech), most of the current approaches use the spike glycoprotein as a target and primarily use full length or simple partial ORFs [24]. It is still unclear if this single target will prove to be sufficient for mounting long-term protective immunity for humans, and whether novel variants of concerns will lead to decreased efficacy. More generally, limited information on which parts of the virus are recognized by human immune responses is a major knowledge gap impeding novel vaccine design and surveillance, although efforts are currently underway to study patterns of immunodominance [25] and to identify conserved epitopes for cross-reactive antibody binding [26]. While current vaccine strategies focus on inducing B cell humoral responses, T cell immunity comprises another dominant domain of immune responses essential for viral vaccines [27–29] and may play an important role in eliminating SARS-CoV-2 [25,30]. Therefore, it is important to examine both B cell and T cell epitopes and consider more precise pathogenic and immunogenic regions that could potentially induce a stronger immune response.

We thus set out to identify those regions by intersecting both discriminative NT-COPA score hotspots and immunogenic hotspots. To identify regions in the SARS-CoV-2 proteome that are predicted to be both pathogenic and immunologically relevant, we ran B cell epitope analysis (**Figure 5A**) as well as T cell MHC-I and MHC-II binder predictions, and then integrated them with the ML-generated COPA scores (**Figure 5B, Figure 6B, and Figure S6A**). Surprisingly, we found that for spike and nucleocapsid proteins, high COPA pathogenic regions significantly overlap with potential B cell epitopes (hypergeometric test, $p < 0.008$ for spike, and $p < 0.0012$ for nucleocapsid) (**Figure 5A, Figure 6A**). For T cell epitopes, we prioritize peptides by counts of discriminative hotspot peaks obtained from the kernel smoothing analysis. These convergent regions may help prioritize epitopes that overlap with potentially functionally important regions of SARS-CoV-2 (**Figure 5B, Figure 6B, Figure S6A**). For example, incorporating these discriminative signals may help for developing vaccines that generate immune responses enriched in neutralization of the more dangerous viral elements. We join other efforts for systematic characterization of SARS-CoV-2 features [10] and provide in this study all the regional hotspots as consensus regions for next-generation precision vaccine development.

Integrative analyses of discriminative profiles with mutations associated with SARS-CoV-2 variants of concern.

Due to the urgency of the pandemic, SARS-CoV-2 genomes have been sequenced at an unprecedented rate, with over a million sequences available through the Global Initiative on Sharing All Influenza Data (GISAID) [31,32]. Other resources such as the Nextstrain project provide genomic epidemiology analyses on the number of accumulated mutational events across the SARS-CoV-2 genome (**Figure S7D**) [33]. Although most mutations are expected to be neutral or mildly deleterious, a small proportion of mutations can also be expected to confer some fitness advantage; and indeed, several 'variants of concern' have emerged for SARS-CoV-2 with altered viral characteristics. We performed a set of analyses intersecting high COPA score residues with mutations from UK variant B.1.1.7 [34], SA variant B.1.351 [35], and Brazil variant P.1 [36], and surprisingly found that a number of high score residues and mutations overlapped (including positions 681 in spike protein, position 183 in NSP3, and position 3 in N protein for UK variant B.1.1.7), as shown in **Figure 7 and Figure S7A-C**. For further comparison with emerging mutations, we extracted the variant emergence rankings and cumulative number of locations for Spike mutation combinations from the GISAID database [37,38] and found 452R_478K_681R_1263L to be the top ranked variant combination (**Figure S7E-F**). The COPA scores for the corresponding mutations are 4.7, 6.76, 66.09, and 23.65, suggesting that our study has identified two of the four top Spike mutations that are currently in circulation. We anticipate that the NT-COPA scores from this study can be used together with emerging data on SARS-CoV-2 evolution

and transmission for prioritizing which mutations may potentially contribute to variant fitness advantage and warrant further study through feasible reverse genetics experiments.

Discussion

This study developed a rigorous framework that integrates base ML models and a statistical meta-model to distinguish pathogenic sequence features of coronaviruses down to base pair and amino acid resolutions with quantitative and biologically interpretable COPA scores. By training and evaluating a high number of diverse ML models on a large collection of coronavirus genomes of human and animal origins, we identified discriminative hotspots across the SARS-CoV-2 viral genome with potential significance for viral fitness. Comparative validation with previous work through in-depth, biologically-motivated investigation showed various intersections of common key features, while the ML approach itself is fully unbiased in terms of scoring and generation of a large number of previously unidentified candidate hotspots. For example, the significance of these hotspots was shown with in-depth evolutionary and structural analyses of the spike protein and RdRp, which are important SARS-CoV-2 genetic elements under active investigation. The integrative analysis of pathogenicity-associated genomic profiles with B cell and T cell epitopes converged on regions of the SARS-CoV-2 proteome that are predicted to be both pathogenic and immunologically relevant, which provides a collection of feature-rich elements that potentially serve as candidates for prioritization and enrichment of key sequence targets to guide vaccine development.

While we focused our downstream analysis here on spike protein and RdRp to demonstrate the interpretability and functional significance of the COPA scores learned from our framework, we emphasize that the learned features from this study are genome-wide and may provide insights into less characterized SARS-CoV-2 structural and non-structural proteins. For example, we noticed that our framework identified a high density of pathogenic peaks in ORF8 (**Figure 2B**), a protein whose function remains mysterious [39]. A recent study has identified a 382-nt deletion variant that covers nearly the entirety of ORF8 from strains isolated from hospitalized patients in Singapore [40], which were implied to lead to reduced virulence of SARS-CoV-2 based on experimental data from SARS-CoV OFR8 deletion variants [41]. Whether ORF8 is in fact an important driver of SARS-CoV-2 pathogenicity will require further study, and it should be noted that the low sequence identity across species for ORF8 may lead to increases in NT-COPA scores of unclear significance. However, the discriminative signatures identified here may constitute an unbiased collection for regional dissection through viral experiments.

Though the application of ML methods for identifying pathogenic sequence elements in viral genomes at scale has been limited to date, the ongoing COVID-19 pandemic has highlighted the importance of this field. Recently, another group has used comparative genomics and ML methods to identify determinants of pathogenicity in SARS-CoV-2 [11]. Though there are some similarities in goals, this study has several differentiating factors: (1) we trained on 3,665 genomes including both human and animal coronaviruses compared to 944 human coronavirus genomes only, and should be able to capture host-based or evolutionary signals; (2) we encode our alignments to include both nucleotide type as well as gaps, as opposed to only encoding gaps, and should therefore capture information on indels and substitutions rather than indels only; (3) we developed a statistical meta-model that integrates signals to provide COPA scores that are unbiased, nucleotide resolution, and quantitative, rather than using pre-defined thresholds to identify regions of interest; (4) we use multiple classifiers and classification strategies rather than one; and (5) we have performed both immune epitope and variant of concern analyses. We anticipate that both the integrative analytical methods and results described here will provide substantial value to the COVID-19 research community in conjunction with other studies.

Given the ongoing nature of the COVID-19 pandemic, there is an urgent need to identify functionally important features of SARS-CoV-2. While much effort is currently underway to characterize the spike protein, RdRp, and other proteins suggested to be important from prior studies on coronaviruses, there has been limited information on sequence determinants of pathogenicity at the global, metavirome-wide scale. We demonstrate here how harnessing the predictive power of ML or other artificial intelligence algorithms may be used to identify such features in a systematic manner. While our ML strategies are based on primary sequences, future ML algorithms that incorporate 3D structures may generate additional insights that cannot be obtained from linear sequence analysis alone, and further enhance the prediction of pathogenicity, immunogenicity, or other important elements of viral proteins. This study demonstrates the development and application of ML to coronavirus genomes with integrative analyses, which is not limited to coronaviruses but can be broadly applied to other viral genomes or microbial pathogens to gain insights on pathogenicity and immunogenicity.

Limitations of the Study

The primary limitation to this study is that although the identified features are potentially related to coronavirus pathogenicity by design of the ML-based approach, genomic regions may be substantially different without necessarily contributing to increased pathogenicity. Moreover, the resulting scores do not

add information on the functional nature of the hotspots. Therefore, further experimental studies are necessary to determine the functional significance of these discriminative genomic features.

Experimental Procedures

Resource availability

No new biological materials was generated by this study. New computational methods were developed. The codes are made available via Zenodo or via lead contact Sidi Chen (sidi.chen@yale.edu).

Lead contact

Requests for resources and code used throughout the study should be directed to and will be fulfilled by the lead contact Sidi Chen (sidi.chen@yale.edu).

Materials availability

No new biological materials was generated by this study.

Data and code availability

The authors are committed to freely share all COVID-19 related data, knowledge and resources to the community to facilitate the development of new treatment or prevention approaches against SARS-CoV-2 / COVID-19 as soon as possible. All relevant processed data generated during this study are included in this article and its supplemental information files. Raw data are from various sources as described below. Data and resources related to this study have been deposited at Zenodo under the DOI https://doi.org/10.5281/zenodo.5652344 and are freely available upon request to the corresponding author. Additional Supplemental Items are available from Mendeley Data at http://dx.doi.org/10.17632/tfmzjdkxh6.1.

Sequence data collection:

A total of 3,665 complete nucleotide genomes of the "Coronaviridae" family were downloaded from the Virus Pathogen Database and Analysis Resource (ViPR) database (Pickett et al., 2012) to be used for machine learning algorithm training. Genbank accession MN908947 was used as the reference SARS-CoV-2 sequence for downstream analyses. Coronavirus protein sequences for spike protein (YP_009755834, ACN89696, ABD75577, QIQ54048, QHR63300, QHD43416, QDF43825, ATO98157, AAP13441, ASO66810, ALD51904, AYF53093, AKG92640, ALA50214, AFD98757, AJP67426, AHX26163, AVM80492) and ORF1ab (QIT08254, QJE38280, QJD07686, QHR63299, QIA48640, QDF43824,

AAP13442, QCC20711, AJD81438, AHE78095, ATP66760, ABD75543, YP_009019180, AVM80693, AFU92121, AFD98805, APZ73768, ATP66783, YP_002308496) used for evolutionary analyses were obtained from the NCBI Virus community portal. Amino acid sequences for SARS-CoV-2 were obtained from translations from reference sequence NC_045512 (equivalent to MN908947). FASTA sequences for S protein (YP_009724390), E protein (YP_009724392), M protein (YP_009724393), N protein (YP_009724397), NSP3 (YP_009742610), NSP5 (YP_009742612), NSP8 (YP_009742615), NSP9 (YP_009742616), and NSP12 (YP_009725307) were obtained from the NCBI Protein database and used for downstream evolutionary and immune epitope analyses.

Pre-processing:

Sequences were aligned with MAFFT [42] version 7 with the --auto strategy. Degenerate IUPAC base symbols that represent multiple bases were converted to "N" and ultimately masked prior to training algorithms. Six bp-wide sliding windows with 1bp shifts were generated across every position in the alignment for a total of 100,835 alignment-tiled windows. Genetic features including nucleotides and gaps for a given window were converted to binary vector representations using LabelEncoder and OneHotEncoder from the Python scikit-learn library [43], for integer encoding of labels and one-hot encoding respectively. Additional Python libraries used include BioPython [44], NumPy , and pandas [46].

Principal components analysis:

Dimensionality reduction of encoded whole coronavirus genomes was performed primarily using R scripts. The MSA was converted to cell-based representations in a CSV file, followed by one hot encoding, PCA, and visualization with metadata labelling. One hot encoding with performed with the "mltools" R package and PCA was performed with the "prcomp" R function.

Training and evaluating machine learning base models:

Genome metadata was converted to binary vector classifications with "1" representing predictor class genomes depending on classification strategy and "0" representing all other genomes. Three different classification strategies were used: (1) predictor class comprised coronavirus samples infecting human hosts, (2) predictor class comprised all SARS-CoV-2, SARS-CoV, and MERS-CoV samples, and (3) predictor class comprised SARS-CoV-2, SARS-CoV, and MERS-CoV samples specifically infecting human hosts. Five supervised learning classifiers from scikit-learn were used for training and evaluation, with seeds set at 17 for algorithms that use a random number generator. Support vector classifiers (SVC) were trained with a linear kernel and regularization parameter of 1.0; random forest (RF) classifiers were

trained with 100 estimators; Bernoulli Naïve Bayes (BNB) were trained with alpha of 1.0 with the "fit_prior" parameter set as true to learn class prior probabilities; multi-layer perceptron (MLPC) classifiers were trained with "lbfgs" solver, alpha of 1e-5, 5 neurons in the first hidden layer, and 2 neurons in the second hidden layer; gradient boosting classifiers (GBC) were trained with "deviance" loss function, learning rate of 0.1, and 100 estimators. All estimators were trained and evaluated with stratified 5-fold cross-validation on each window, using 80% of the data for training and 20% of the data for validation. Each of the 5-fold cross-validations were performed once with the cross_val_score function from scikit-learn, with folds created preserving the percentage of samples for each class.

Statistical hypothesis test-based meta-model:

Accuracy scores obtained from machine learning base models were used as a proxy for "learned, predictive information content" to determine coronavirus pathogenicity (COPA) scores using a statistical hypothesis test-based meta-model. First, Shannon entropy values were calculated for each window across the alignment. Windows with minimal entropy values (n = 10,383), typically found in highly gapped regions, were used to define a biologically meaningful control group; i.e., we hypothesized that windows with low information content in highly gapped regions should not be predictive of coronavirus pathogenicity and should have minimal discriminative value. For each position across the alignment (100,840 positions), scores associated with windows that overlap with the position (typically ~six windows) were pooled and tested to see if statistically significantly different from the minimal entropy control group using the nonparametric two-sided Wilcoxon rank-sum test. For the main NT-COPA score calculations and evolution-based analyses, all scores across the three classification strategies were used for testing; in supplemental analyses, scores for individual classification strategies were used separately. This procedure was performed across the alignment, and p-values were adjusted for multiple comparisons using the Benjamini & Hochberg procedure. P-values were transformed to nucleotide resolution coronavirus pathogenicity scores by negative log base 10 (also referred to as NT-COPA scores). Amino acid resolution scores were obtained by averaging the NT-COPA scores for a given residue's codon (referred to simply as COPA scores).

Kernel regression smoothing for hotspot peak identification:

For a systematic strategy to identify pathogenicity hotspots across the SARS-CoV-2 genome using COPA scores, we combined kernel regression smoothing with local maxima identification. For each position across the alignment, we determined the Nadaraya-Watson kernel regression estimate using the ksmooth function in R with a "normal" kernel and various bandwidth sizes. Peaks highlighted in this study are primarily based

on estimates calculated with bandwidth size of 3. Local peaks were determined from kernel regression estimates using the "findpeaks" function with nups parameter set at 2, from the "pracma" R package.

Evolutionary analyses:

Protein sequences used for evolutionary analyses were aligned using MAFFT version 7 with the "L-INS-i" strategy [42]. Alignments were visualized using Jalview 2.11.1.0 [47]. Phylogenic analyses were performed using MEGA10.1.8 software [48]. Phylogeny trees were generated with the Maximum Likelihood statistical method, Jones-Taylor-Thornton (JTT) substitution model, uniform rates among sites, use of all sites, Nearest-Neighbor-Interchange (NNI) heuristic method, and default NJ/BioNJ initial tree. For spike protein analysis, all obtained sequences were used for alignment and phylogeny. For NSP12 analysis, all obtained ORF1ab sequences and reference SARS-CoV-2 NSP12 (YP_009725307) were used for alignment, but only ORF1ab sequences were used for phylogeny.

For large scale phylogenetic analysis, efficient tree inference on the full genome set multiple sequence alignment was performed using IQ-TREE version 2.0.6 [49] with the GTR+F+R10 model, which was selected automatically using ModelFinder [50]. Circular phylogenetic trees were then generated for visualization and labelled using FigTree v1.4.4.

Structural analyses:

The crystal structure of SARS-CoV-2 spike receptor-binding domain bound with ACE2 was obtained from Protein Data Bank (PDB) with accession code 6M0J [15]. The cryo-EM structure of the SARS-CoV-2 NSP12-NSP7-NSP8 complex bound to the template-primer RNA and the triphosphate form of remdesivir (RTP) was obtained from PDB with accession code 7BV2 [21]. The crystal structure of SARS-CoV spike RBD bound with ACE2 was obtained from PDB with accession code 2AJF [19]. Molecular graphics and analyses including mapping of COPA scores onto structures were performed with UCSF ChimeraX version 0.94 [51].

B cell epitope analysis:

FASTA sequences for reference SARS-CoV-2 structural proteins were used to predict B cell epitopes. Linear B-cell epitopes probability scores were obtained using BepiPred-2.0 [52]. "Consensus Regions" were defined as amino acid residues with epitope scores > 0.5 and COPA scores > 8. Hypergeometric test of overlap of high COPA score (> 8) and high epitope score (> 0.5) residues was performed to determine the statistical significance of consensus regions. "Compound Regions" were identified using k-means clustering. Briefly, the R function "kmeans" was run with variable number of clusters and nstart parameter

25 on a dataset containing residue position, epitope score, and COPA score. Residues were marked as compound regions if they belonged to clusters with epitope score centers > 0.5 and COPA score centers > 8. Flagged residues that did not belong to a contiguous run of amino acids ≥ 5 residues were filtered out.

T cell epitope analysis:

FASTA sequences for reference SARS-CoV-2 structural proteins and select nonstructural proteins were used to predict T cell epitopes. Prediction of peptides binding to MHC class I and class II molecules was then performed using TepiTool [53] from the Immune Epitope Database (IEDB) Analysis Resource. MHC-I binder predictions were made for the "Human" host species and the 27 most frequent A & B alleles in the global population. Default settings for low number of peptides (only 9mer peptides), IEDB recommended prediction method, and predicted percentile rank cutoff ≤ 1.0 were used for peptide selection. MHC-II binder predictions were made for the "Human" host species using the "7-allele method" (median of percentile ranks from DRB1*03:01, DRB1*07:01, DRB1*15:01, DRB3*01:01, DRB3*02:02, DRB4*01:01, DRB5*01:01). Median consensus percentile rank ≤ 20.0 was used for peptide selection. Pathogenicity associated peaks within the proteins with NT-COPA scores greater than 8 were then mapped to the predicted peptides for prioritization.

Variant of concern analysis:

Mutation profiles for the United Kingdom variant B.1.1.7 [34], South African variant B.1.351 [35], and Brazilian variant P.1 [36] were obtained for comparison with NT-COPA score profiles. Individual mutations were mapped onto COPA score signal density plots for separate features and mutations overlapping with highly discriminative regions were marked.

Statistical information summary

Comprehensive information on the statistical analyses used are included in various places, including the figures, figure legends and results, where the methods, significance, p-values and/or tails are described. All error bars have been defined in the figure legends or methods. Standard statistical calculations such as Spearman's rho were performed in R with functions such as "cor".

Acknowledgments

We thank Richard Sutton, Yong Xiong, Hongyu Zhao, Albert Ko, Yong Xiong, Craig Wilen, Katie Zhu, Ruth Montgomery, and a number of other colleagues for discussions. We thank Antonio Giraldez, Andre Levchenko, Chris Incarvito, Mike Crair, and Scott Strobel for their support on COVID-19 research. We

thank Chen lab members such as Matthew Dong, Ariel Zhou, Vino Peng, Ryan Chow, Paul Clark, Viola Lee, Stanley Lam, as well as our colleagues in the Genetics Department, the Systems Biology Institute and various Yale entities. We personally thank all the frontline healthcare workers directly fighting this disease.

Author contributions

JJP and SC conceived and designed the study. JJP developed the analysis approach, performed all data analyses, and created the figures. JJP and SC prepared the manuscript. SC supervised the work.

Declaration of interests

The authors declare no competing interests.

Figure Legends

Figure 1. Machine learning and statistical meta-model identifies high-resolution discriminative features in coronavirus genomes.

(A) Schematic detailing ML-based strategy to learn discriminative genomic features of coronaviruses. Complete genome sequences of the Coronaviridae family in the ViPR database (n = 3,665) were obtained, aligned, and encoded into binary vector representations. Base machine learning models with different classification strategies were trained on sliding windows tiled across the alignment. A statistical hypothesis test-based meta-model integrated signals into a coronavirus pathogenicity (COPA) score to identify discriminative hotspot regions in the SARS-CoV-2 genome.

(B) (Left) Donut chart showing distribution of host species for coronavirus genomes used in study. (Middle) Donut chart showing the distribution of virus genus. (Right) Donut chart showing the distribution of virus species.

(C) Pie charts showing class membership proportions for different classification strategies. (Left) Strategy A defines the predictor class as coronavirus samples that infect human hosts. (Middle) Strategy B defines the predictor class as all SARS-CoV-2, SARS-CoV, and MERS-CoV samples, including those that infect human or animal hosts. (Right) Strategy C defines the predictor class as specifically those SARS-CoV-2, SARS-CoV, and MERS-CoV samples that infect human hosts.

(D) NT-COPA scores (negative log base 10 of adjusted p-values obtained from meta-model, see **Experimental Procedures**) for every nucleotide position across the reference SARS-CoV-2 genome. Larger NT-COPA scores represent stronger discriminative signals learned from our models.

(E) Circular phylogenetic trees built from all Coronaviridae genome sequences used for training ML models labelled by genus, host, or species.

See also Figures S1-S2

Figure 2. Identification of local discriminative hotspots in SARS-CoV-2 proteins.

(A) High resolution NT-COPA score distributions shown for spike protein, membrane protein, ORF8, NSP1, NSP5 (3C-like protease), and NSP12 (RNA-dependent RNA polymerase). Scores at each nucleotide position are shown as pink dots. Smoothed kernel regression estimates are shown a blue line graph, with select local peaks circled and labeled in red.

(B) NT-COPA scores for local highly discriminative peaks identified across SARS-CoV-2 genome shown plotted against rank by kernel regression estimate. Smoothing and peak identification was used as an unbiased strategy to identify hotspots.

See also Figure S3

Figure 3. Spike protein hotspots reveal a furin cleavage site and contact sites with ACE2.

(A) Phylogeny tree for spike protein sequences of coronaviruses across species and hosts. Sequences for alphacoronavirus are labelled in green, betacoronaviruses labelled in purple, gammacoronaviruses labelled in brown, and the reference SARS-CoV-2 labelled in red.

(B) NT-COPA score signal density near peak S-2044 (amino acid position 682) in spike protein compared to alignment. Peak S-2044 corresponds to a furin-like cleavage site.

(C) AA-COPA score, i.e. COPA score signal density in receptor-binding domain (RBD) in spike protein reveals two primary discriminative hotspot regions.

(D) COPA scores mapped onto structure of SARS-CoV-2 RBD complexed with ACE2 receptor reveal that the NCYF hotspot contains residues that mediate viral binding to host receptor.

(E) Protein alignment reveals NCYF hotspot for SARS-CoV-2 (NCYW hotspot for SARS-CoV) has high sequence divergence from other coronaviruses across species and hosts. Residues are colored using the Clustal X color scheme. Hotspot residues for SARS-CoV-2 labelled in red, with corresponding residues for SARS-CoV labelled in blue.

See also Figure S4

Figure 4. RdRp hotspots reveal RNA contact sites and codon composition biases.

(A) COPA score signal density across the NSP12/RdRp amino acid sequence. Select hotspot residues marked in red and directed towards their location on structure in **(B)**.

(B) COPA scores mapped onto RdRp in structure of SARS-CoV-2 RdRp-NSP7-NSP8 complex bound to the template-primer RNA and triphosphate form of remdesivir (RTP). (Left) Spatial region in RdRp with high density of hotspots. (Right) Discriminative hotspot residues correspond to contact sites in RdRp that directly participate in the binding of RNA.

(C) Protein alignment reveals discriminative hotspot residues in SARS-CoV-2 RdRp has high sequence divergence from other coronaviruses across species and hosts, with exceptions as noted in **(D)**. For the ERVRQ hotspot region (amino acid positions 180 to 184), residues are colored according to hydrophobicity (where most hydrophobic residues are colored red and the most hydrophilic ones are colored blue). For other regions, residues are colored according to their Blosum62 score (where residues matching the consensus sequence residue at that position are colored dark blue).

(D) Sequence logos for codons from the genome alignment used for ML training associated with the ERVRQ hotspot region and KS hotspot contact sites. Logos were generated separately for MERS-CoV, SARS-CoV, and SARS-CoV-2 infecting human hosts, and non-pathogenic coronaviruses.

See also Figure S5

Figure 5. Integration of genomic discriminative profiles with B cell and T cell immunogenic features.

(A) B cell epitope integrative analyses for spike protein, membrane protein, envelope protein, and nucleocapsid protein. (Upper) COPA scores and B cell epitope prediction scores plotted across the amino acid sequences. Thresholds used for identifying key residues (COPA score > 8 and epitope score > 0.5) marked with horizontal line. Statistical significance of overlap of key residues was determined by hypergeometric test (see Figure 6A). (Lower) Residues marked for COPA score > 8, epitope score > 0.5, consensus regions, and compound regions.

(B) T cell epitope integrative analyses for spike protein and RdRp for MHC-I and MHC-II. Highly discriminative peaks identified from kernel regression estimates were mapped onto predicted MHC-I and MHC-II binders. Peptides with high peak counts are highlighted.

See also Figure S6

Figure 6. Additional integrative analyses of discriminative profiles with B cell and T cell epitopes for SARS-CoV-2 structural proteins.

(A) Venn diagrams showing overlap of key residues in spike protein, membrane protein, envelope protein, and nucleocapsid protein identified with thresholds of COPA score > 8 and epitope score > 0.5. Statistical significance of overlap of key residues was determined by hypergeometric test.

(B) T cell epitope integrative analyses for nucleocapsid protein, membrane protein, and envelope protein for MHC-I and MHC-II. Highly discriminative peaks identified from kernel regression estimates were mapped onto predicted MHC-I and MHC-II binders. Peptides with high peak counts are highlighted.

See also Figure S6

Figure 7. Integrative analyses of discriminative profiles with mutations associated with SARS-CoV-2 variants of concern.

(A) COPA score signal density across the spike protein, NSP3, and N protein amino acid sequences mapped with mutations associated with the United Kingdom variant B.1.1.7. Vertical dashed lines represent locations of specific B.1.1.7 mutations. Mutations overlapping with high COPA score hotspot regions are marked with a red arrow and labelled.

(B) COPA score signal density across the N protein and ORF3a amino acid sequences mapped with mutations associated with the South African variant B.1.351. Vertical dashed lines represent locations of specific B.1.351 mutations. Mutations overlapping with high COPA score hotspot regions are marked with a red arrow and labelled.

(C) COPA score signal density across the N protein and NSP3 amino acid sequences mapped with mutations associated with the Brazilian variant P.1. Vertical dashed lines represent locations of specific P.1 mutations. Mutations overlapping with high COPA score hotspot regions are marked with a red arrow and labelled.

See also Figure S7

Supplemental Datasets

Data S1. Metadata for sequences used in study, NT-COPA scores, and statistics.

Dataset S1. Metadata, classification strategy predictor classes, and principal components (PC1-PC4) for Coronaviridae family genomes obtained from ViPR database.

Dataset S2. Cumulative Shannon entropy scores for six bp windows tile across alignment.

Dataset S3. Meta-model q-values, NT-COPA scores, kernel regression estimate, and consensus scores for each nucleotide position in SARS-CoV-2 genome with standard dataset.

Dataset S4. Amino acid resolution COPA scores for residues in SARS-CoV-2 features.

Dataset S5. Meta-model q-values, NT-COPA scores, and kernel regression estimate for each nucleotide position in SARS-CoV-2 genome for each independent (A, B, C) and evolution-based classification strategies.

Dataset S6. Metadata for coronavirus spike protein sequences used for evolutionary analysis.

Dataset S7. Metadata for coronavirus ORF1ab polypeptide and NSP12 protein sequences used for evolutionary analysis.

Data S2. T and B cell integrative analysis scores.

Dataset S8. B cell epitope scores, COPA scores, and key regions for spike protein.

Dataset S9. B cell epitope scores, COPA scores, and key regions for membrane protein.

Dataset S10. B cell epitope scores, COPA scores, and key regions for envelope protein.

Dataset S11. B cell epitope scores, COPA scores, and key regions for nucleocapsid protein.

Dataset S12. T cell epitopes and peak counts for spike protein and MHC class I.

Dataset S13. T cell epitopes and peak counts for spike protein and MHC class II.

Dataset S14. T cell epitopes and peak counts for NSP12 and MHC class I.

Dataset S15. T cell epitopes and peak counts for NSP12 and MHC class II.

Dataset S16. T cell epitopes and peak counts for nucleocapsid protein and MHC class I.

Dataset S17. T cell epitopes and peak counts for nucleocapsid protein and MHC class II.

Dataset S18. T cell epitopes and peak counts for membrane protein and MHC class I.

Dataset S19. T cell epitopes and peak counts for membrane protein and MHC class II.

Dataset S20. T cell epitopes and peak counts for envelope protein and MHC class I.

Dataset S21. T cell epitopes and peak counts for envelope protein and MHC class II.

Dataset S22. T cell epitopes and peak counts for NSP3 and MHC class I.

Dataset S23. T cell epitopes and peak counts for NSP3 and MHC class II.

Dataset S24. T cell epitopes and peak counts for NSP5 and MHC class I.

Dataset S25. T cell epitopes and peak counts for NSP5 and MHC class II.

Dataset S26. T cell epitopes and peak counts for NSP8 and MHC class I.

Dataset S27. T cell epitopes and peak counts for NSP8 and MHC class II.

Dataset S28. T cell epitopes and peak counts for NSP9 and MHC class I.

Dataset S29. T cell epitopes and peak counts for NSP9 and MHC class II.

Data S3. Classifier performance accuracy scores.

Dataset S30. Classifier performance accuracy scores for permutations of machine learning algorithm, fold, and classification strategy (A, B, or C) with standard dataset.

References:

1. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., Xiao, Y., Gao, H., Guo, L., Xie, J., Wang, G., Jiang, R., Gao, Z., Jin, Q., Wang, J. & Cao, B. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395**, 497–506 (2020).
2. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S. M., Lau, E. H. Y., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T. T. Y., Wu, J. T., Gao, G. F., Cowling, B. J., Yang, B., Leung, G. M. & Feng, Z. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *N. Engl. J. Med.* (2020) doi:10.1056/NEJMoa2001316.
3. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534 (2020).
4. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **16**, 321–332 (2015).
5. Ohler, U., Liao, G., Niemann, H. & Rubin, G. M. Computational analysis of core promoters in the Drosophila genome. *Genome Biol.* **3**, research0087.1 (2002).
6. Beer, M. A. & Tavazoie, S. Predicting Gene Expression from Sequence. *Cell* **117**, 185–198 (2004).
7. Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F. C. P., Clarke, D., Gu, M., Emani, P., Yang, Y. T., Xu, M., Gandal, M. J., Lou, S., Zhang, J., Park, J. J., Yan, C., Rhie, S. K., Manakongtreecheep, K., Zhou, H., Nathan, A., Peters, M., Mattei, E., Fitzgerald, D., Brunetti, T., Moore, J., Jiang, Y., Girdhar, K., Hoffman, G. E., Kalayci, S., Gümüş, Z. H., Crawford, G. E., Consortium, P., Roussos, P., Akbarian, S., Jaffe, A. E., White, K. P., Weng, Z., Sestan, N., Geschwind, D. H., Knowles, J. A. & Gerstein, M. B.

Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, (2018).

8. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **17**, 181–192 (2019).

9. Yuan, S., Peng, L., Park, J. J., Hu, Y., Devarkar, S. C., Dong, M. B., Shen, Q., Wu, S., Chen, S., Lomakin, I. B. & Xiong, Y. Nonstructural Protein 1 of SARS-CoV-2 Is a Potent Pathogenicity Factor Redirecting Host Protein Synthesis Machinery toward Viral RNA. *Mol. Cell* **80**, 1055-1066.e6 (2020).

10. 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., Tummino, T. A., Huettenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., Kim, M., Haas, P., Polacco, B. J., Braberg, H., Fabius, J. M., Eckhardt, M., Soucheray, M., Bennett, M. J., Cakir, M., McGregor, M. J., Li, Q., Meyer, B., Roesch, F., Vallet, T., Mac Kain, A., Miorin, L., Moreno, E., Naing, Z. Z. C., Zhou, Y., Peng, S., Shi, Y., Zhang, Z., Shen, W., Kirby, I. T., Melnyk, J. E., Chorba, J. S., Lou, K., Dai, S. A., Barrio-Hernandez, I., Memon, D., Hernandez-Armenta, C., Lyu, J., Mathy, C. J. P., Perica, T., Pilla, K. B., Ganesan, S. J., Saltzberg, D. J., Rakesh, R., Liu, X., Rosenthal, S. B., Calviello, L., Venkataramanan, S., Liboy-Lugo, J., Lin, Y., Huang, X.-P., Liu, Y., Wankowicz, S. A., Bohn, M., Safari, M., Ugur, F. S., Koh, C., Savar, N. S., Tran, Q. D., Shengjuler, D., Fletcher, S. J., O'Neal, M. C., Cai, Y., Chang, J. C. J., Broadhurst, D. J., Klippsten, S., Sharp, P. P., Wenzell, N. A., Kuzuoglu, D., Wang, H.-Y., Trenker, R., Young, J. M., Cavero, D. A., Hiatt, J., Roth, T. L., Rathore, U., Subramanian, A., Noack, J., Hubert, M., Stroud, R. M., Frankel, A. D., Rosenberg, O. S., Verba, K. A., Agard, D. A., Ott, M., Emerman, M., Jura, N., von Zastrow, M., Verdin, E., Ashworth, A., Schwartz, O., d'Enfert, C., Mukherjee, S., Jacobson, M., Malik, H. S., Fujimori, D. G., Ideker, T., Craik, C. S., Floor, S. N., Fraser, J. S., Gross, J. D., Sali, A., Roth, B. L., Ruggero, D., Taunton, J., Kortemme, T., Beltrao, P., Vignuzzi, M., García-Sastre, A., Shokat, K.

M., Shoichet, B. K. & Krogan, N. J. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* 1–13 (2020) doi:10.1038/s41586-020-2286-9.

11. Gussow, A. B., Auslander, N., Faure, G., Wolf, Y. I., Zhang, F. & Koonin, E. V. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proc. Natl. Acad. Sci.* **117**, 15193–15199 (2020).
12. Su, S., Wong, G., Shi, W., Liu, J., Lai, A. C. K., Zhou, J., Liu, W., Bi, Y. & Gao, G. F. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. *Trends Microbiol.* **24**, 490–502 (2016).
13. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S., Herrler, G., Wu, N.-H., Nitsche, A., Müller, M. A., Drosten, C. & Pöhlmann, S. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271-280.e8 (2020).
14. Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., Graham, B. S. & McLellan, J. S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
15. Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L. & Wang, X. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215–220 (2020).
16. Walls, A. C., Park, Y.-J., Tortorici, M. A., Wall, A., McGuire, A. T. & Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).
17. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).
18. Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A. & Li, F. Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).

19. Li, F., Li, W., Farzan, M. & Harrison, S. C. Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. *Science* **309**, 1864–1868 (2005).

20. Gao, Y., Yan, L., Huang, Y., Liu, F., Zhao, Y., Cao, L., Wang, T., Sun, Q., Ming, Z., Zhang, L., Ge, J., Zheng, L., Zhang, Y., Wang, H., Zhu, Y., Zhu, C., Hu, T., Hua, T., Zhang, B., Yang, X., Li, J., Yang, H., Liu, Z., Xu, W., Guddat, L. W., Wang, Q., Lou, Z. & Rao, Z. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779–782 (2020).

21. Yin, W., Mao, C., Luan, X., Shen, D.-D., Shen, Q., Su, H., Wang, X., Zhou, F., Zhao, W., Gao, M., Chang, S., Xie, Y.-C., Tian, G., Jiang, H.-W., Tao, S.-C., Shen, J., Jiang, Y., Jiang, H., Xu, Y., Zhang, S., Zhang, Y. & Xu, H. E. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science* (2020) doi:10.1126/science.abc1560.

22. Beigel, J. H., Tomashek, K. M., Dodd, L. E., Mehta, A. K., Zingman, B. S., Kalil, A. C., Hohmann, E., Chu, H. Y., Luetkemeyer, A., Kline, S., Lopez de Castilla, D., Finberg, R. W., Dierberg, K., Tapson, V., Hsieh, L., Patterson, T. F., Paredes, R., Sweeney, D. A., Short, W. R., Touloumi, G., Lye, D. C., Ohmagari, N., Oh, M., Ruiz-Palacios, G. M., Benfield, T., Fätkenheuer, G., Kortepeter, M. G., Atmar, R. L., Creech, C. B., Lundgren, J., Babiker, A. G., Pett, S., Neaton, J. D., Burgess, T. H., Bonnett, T., Green, M., Makowski, M., Osinusi, A., Nayak, S. & Lane, H. C. Remdesivir for the Treatment of Covid-19 — Preliminary Report. *N. Engl. J. Med.* **0**, null (2020).

23. Holshue, M. L., DeBolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., Spitters, C., Ericson, K., Wilkerson, S., Tural, A., Diaz, G., Cohn, A., Fox, L., Patel, A., Gerber, S. I., Kim, L., Tong, S., Lu, X., Lindstrom, S., Pallansch, M. A., Weldon, W. C., Biggs, H. M., Uyeki, T. M. & Pillai, S. K. First Case of 2019 Novel Coronavirus in the United States. *N. Engl. J. Med.* **382**, 929–936 (2020).

24. Amanat, F. & Krammer, F. SARS-CoV-2 Vaccines: Status Report. *Immunity* (2020) doi:10.1016/j.immuni.2020.03.007.

25. Grifoni, A., Weiskopf, D., Ramirez, S. I., Mateus, J., Dan, J. M., Moderbacher, C. R., Rawlings, S. A., Sutherland, A., Premkumar, L., Jadi, R. S., Marrama, D., de Silva, A. M., Frazier, A., Carlin, A., Greenbaum, J. A., Peters, B., Krammer, F., Smith, D. M., Crotty, S. & Sette, A. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell* S0092867420306103 (2020) doi:10.1016/j.cell.2020.05.015.

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

27. Arunachalam, P. S., Charles, T. P., Joag, V., Bollimpelli, V. S., Scott, M. K. D., Wimmers, F., Burton, S. L., Labranche, C. C., Petitdemange, C., Gangadhara, S., Styles, T. M., Quarnstrom, C. F., Walter, K. A., Ketas, T. J., Legere, T., Jagadeesh Reddy, P. B., Kasturi, S. P., Tsai, A., Yeung, B. Z., Gupta, S., Tomai, M., Vasilakos, J., Shaw, G. M., Kang, C.-Y., Moore, J. P., Subramaniam, S., Khatri, P., Montefiori, D., Kozlowski, P. A., Derdeyn, C. A., Hunter, E., Masopust, D., Amara, R. R. & Pulendran, B. T cell-inducing vaccine durably prevents mucosal SHIV infection even with lower neutralizing antibody titers. *Nat. Med.* 1–9 (2020) doi:10.1038/s41591-020-0858-8.

28. Miller, J. D., van der Most, R. G., Akondy, R. S., Glidewell, J. T., Albott, S., Masopust, D., Murali-Krishna, K., Mahar, P. L., Edupuganti, S., Lalor, S., Germon, S., Del Rio, C., Mulligan, M. J., Staprans, S. I., Altman, J. D., Feinberg, M. B. & Ahmed, R. Human Effector and Memory CD8+ T Cell Responses to Smallpox and Yellow Fever Vaccines. *Immunity* **28**, 710–722 (2008).

29. Akondy, R. S., Monson, N. D., Miller, J. D., Edupuganti, S., Teuwen, D., Wu, H., Quyyumi, F., Garg, S., Altman, J. D., Del Rio, C., Keyserling, H. L., Ploss, A., Rice, C. M., Orenstein, W. A., Mulligan, M. J. & Ahmed, R. The Yellow Fever Virus Vaccine Induces a Broad and Polyfunctional Human Memory CD8+ T Cell Response. *J. Immunol. Baltim. Md 1950* **183**, 7919–7930 (2009).

30. Braun, J., Loyal, L., Frentsch, M., Wendisch, D., Georg, P., Kurth, F., Hippenstiel, S., Dingeldey, M., Kruse, B., Fauchere, F., Baysal, E., Mangold, M., Henze, L., Lauster, R., Mall, M., Beyer, K., Roehmel, J., Schmitz, J., Miltenyi, S., Mueller, M. A., Witzenrath, M., Suttorp, N., Kern, F., Reimer, U., Wenschuh, H., Drosten, C., Corman, V. M., Giesecke-Thiel, C., Sander, L.-E. & Thiel, A. Presence of SARS-CoV-2 reactive T cells in COVID-19 patients and healthy donors. *medRxiv* 2020.04.17.20061440 (2020) doi:10.1101/2020.04.17.20061440.

31. Meredith, L. W., Hamilton, W. L., Warne, B., Houldcroft, C. J., Hosmillo, M., Jahun, A. S., Curran, M. D., Parmar, S., Caller, L. G., Caddy, S. L., Khokhar, F. A., Yakovleva, A., Hall, G., Feltwell, T., Forrest, S., Sridhar, S., Weekes, M. P., Baker, S., Brown, N., Moore, E., Popay, A., Roddick, I., Reacher, M., Gouliouris, T., Peacock, S. J., Dougan, G., Török, M. E. & Goodfellow, I. Rapid implementation of SARS-CoV-2 sequencing to investigate cases of health-care associated COVID-19: a prospective genomic surveillance study. *Lancet Infect. Dis.* **20**, 1263–1271 (2020).

32. Harvey, W. T., Carabelli, A. M., Jackson, B., Gupta, R. K., Thomson, E. C., Harrison, E. M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S. J. & Robertson, D. L. SARS-CoV-2 variants, spike mutations and immune escape. *Nat. Rev. Microbiol.* 1–16 (2021) doi:10.1038/s41579-021-00573-0.

33. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T. & Neher, R. A. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).

34. Chand, M., Hopkins, S., Dabrera, G., Achison, C., Barclay, W., Ferguson, N., Volz, E., Loman, N., Rambaut, A. & Barrett, J. Investigation of novel SARS-COV-2 variant: Variant of Concern 202012/01. (2020).

35. Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E. J., Msomi, N., Mlisana, K., Gottberg, A. von, Walaza, S., Allam, M., Ismail, A., Mohale, T., Glass, A. J., Engelbrecht, S., Zyl, G. V., Preiser, W., Petruccione, F., Sigal, A., Hardie, D., Marais, G.,

Hsiao, M., Korsman, S., Davies, M.-A., Tyers, L., Mudau, I., York, D., Maslo, C., Goedhals, D., Abrahams, S., Laguda-Akingba, O., Alisoltani-Dehkordi, A., Godzik, A., Wibmer, C. K., Sewell, B. T., Lourenço, J., Alcantara, L. C. J., Pond, S. L. K., Weaver, S., Martin, D., Lessells, R. J., Bhiman, J. N., Williamson, C. & Oliveira, T. de. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv* 2020.12.21.20248640 (2020) doi:10.1101/2020.12.21.20248640.

36. Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. da S., Mishra, S., Crispim, M. A. E., Sales, F. C. S., Hawryluk, I., McCrone, J. T., Hulswit, R. J. G., Franco, L. A. M., Ramundo, M. S., Jesus, J. G. de, Andrade, P. S., Coletti, T. M., Ferreira, G. M., Silva, C. A. M., Manuli, E. R., Pereira, R. H. M., Peixoto, P. S., Kraemer, M. U. G., Gaburo, N., Camilo, C. da C., Hoeltgebaum, H., Souza, W. M., Rocha, E. C., Souza, L. M. de, Pinho, M. C. de, Araujo, L. J. T., Malta, F. S. V., Lima, A. B. de, Silva, J. do P., Zauli, D. A. G., Ferreira, A. C. de S., Schnekenberg, R. P., Laydon, D. J., Walker, P. G. T., Schlüter, H. M., Santos, A. L. P. dos, Vidal, M. S., Caro, V. S. D., Filho, R. M. F., Santos, H. M. dos, Aguiar, R. S., Proença-Modena, J. L., Nelson, B., Hay, J. A., Monod, M., Miscouridou, X., Coupland, H., Sonabend, R., Vollmer, M., Gandy, A., Prete, C. A., Nascimento, V. H., Suchard, M. A., Bowden, T. A., Pond, S. L. K., Wu, C.-H., Ratmann, O., Ferguson, N. M., Dye, C., Loman, N. J., Lemey, P., Rambaut, A., Fraiji, N. A., Carvalho, M. do P. S. S., Pybus, O. G., Flaxman, S., Bhatt, S. & Sabino, E. C. Genomics and epidemiology of the P.1 SARS-CoV-2 lineage in Manaus, Brazil. *Science* **372**, 815–821 (2021).
37. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017).
38. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
39. Scudellari, M. The sprint to solve coronavirus protein structures — and disarm them with drugs. *Nature* **581**, 252–255 (2020).

40. Su, Y. C. F., Anderson, D. E., Young, B. E., Linster, M., Zhu, F., Jayakumar, J., Zhuang, Y., Kalimuddin, S., Low, J. G. H., Tan, C. W., Chia, W. N., Mak, T. M., Octavia, S., Chavatte, J.-M., Lee, R. T. C., Pada, S., Tan, S. Y., Sun, L., Yan, G. Z., Maurer-Stroh, S., Mendenhall, I. H., Leo, Y.-S., Lye, D. C., Wang, L.-F. & Smith, G. J. D. Discovery and Genomic Characterization of a 382-Nucleotide Deletion in ORF7b and ORF8 during the Early Evolution of SARS-CoV-2. *mBio* **11**, e01610-20.
41. Muth, D., Corman, V. M., Roth, H., Binger, T., Dijkman, R., Gottula, L. T., Gloza-Rausch, F., Balboni, A., Battilani, M., Rihtarič, D., Toplak, I., Ameneiros, R. S., Pfeifer, A., Thiel, V., Drexler, J. F., Müller, M. A. & Drosten, C. Attenuation of replication by a 29 nucleotide deletion in SARS-coronavirus acquired during the early stages of human-to-human transmission. *Sci. Rep.* **8**, 15177 (2018).
42. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).
43. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, É. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
44. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B. & de Hoon, M. J. L. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
45. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
46. McKinney, W. Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.* 56–61 (2010) doi:10.25080/Majora-92bf1922-00a.
47. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

48. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

49. Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A. & Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

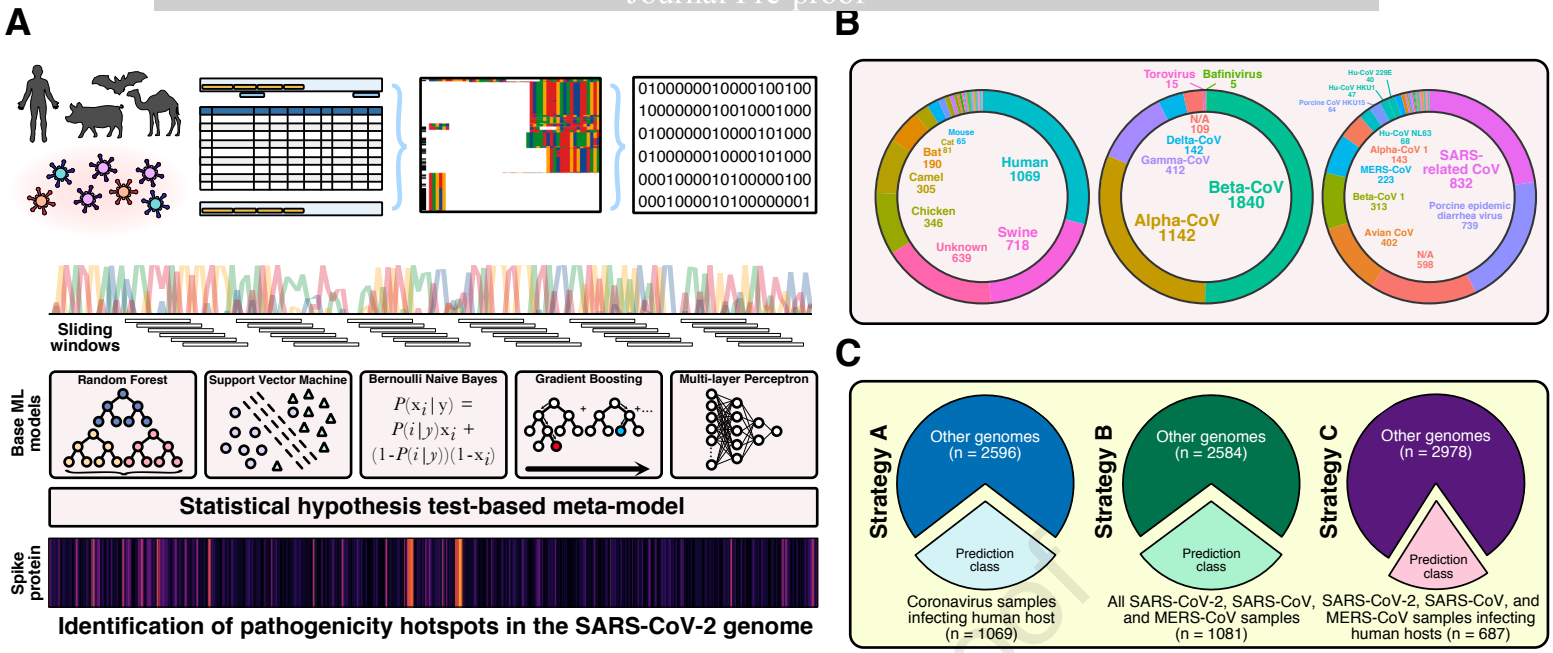
50. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

51. Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H. & Ferrin, T. E. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci. Publ. Protein Soc.* **27**, 14–25 (2018).

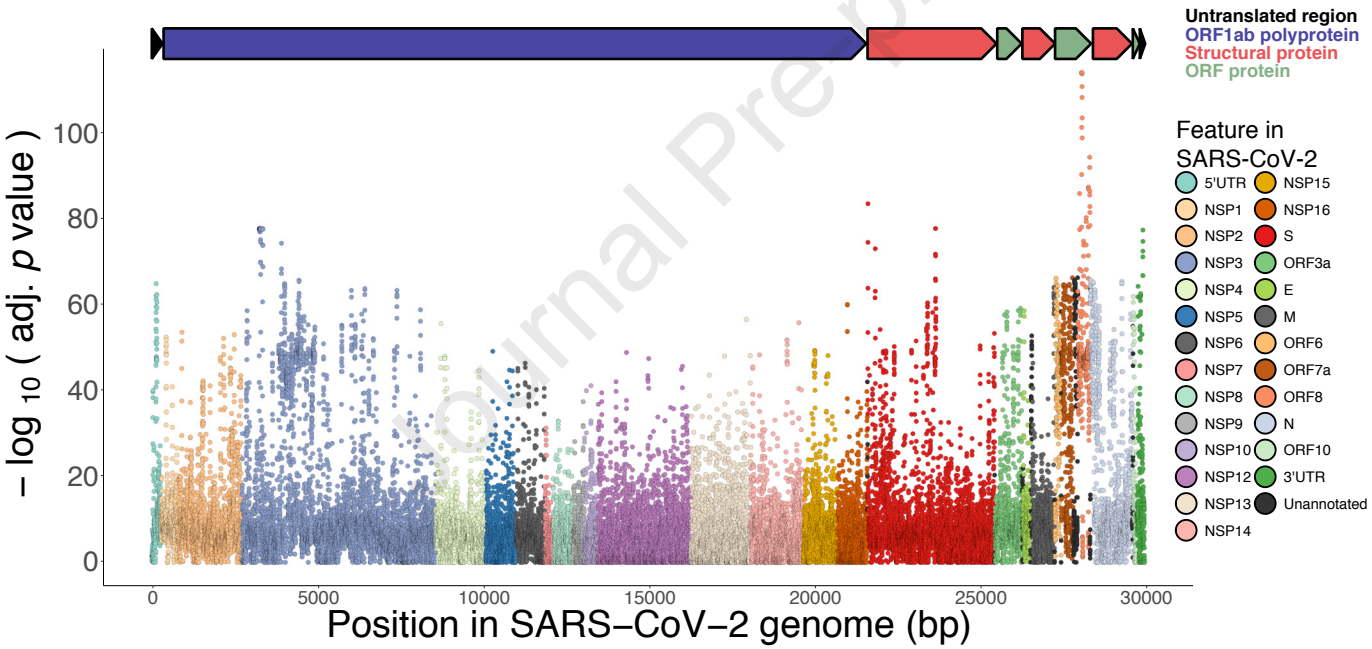
52. Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* **45**, W24–W29 (2017).

53. Paul, S., Sidney, J., Sette, A. & Peters, B. TepiTool: A Pipeline for Computational Prediction of T Cell Epitope Candidates. *Curr. Protoc. Immunol.* **114**, 18.19.1-18.19.24 (2016).

Figure 1



D



E

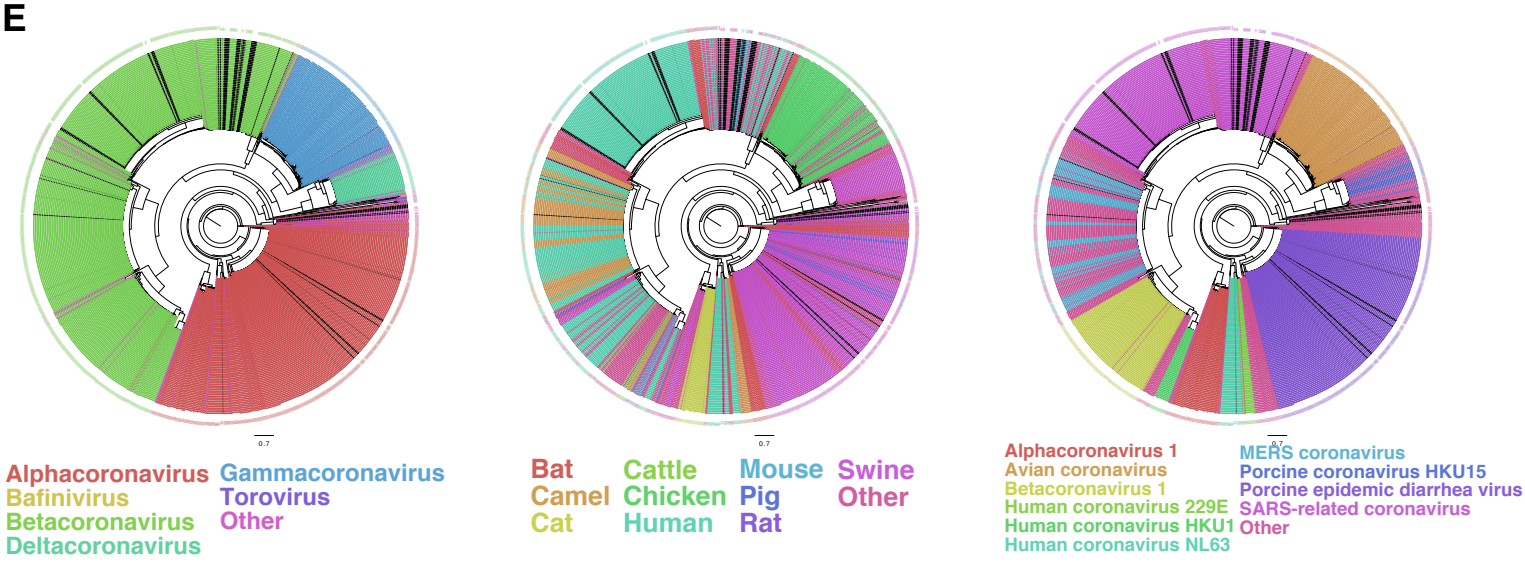
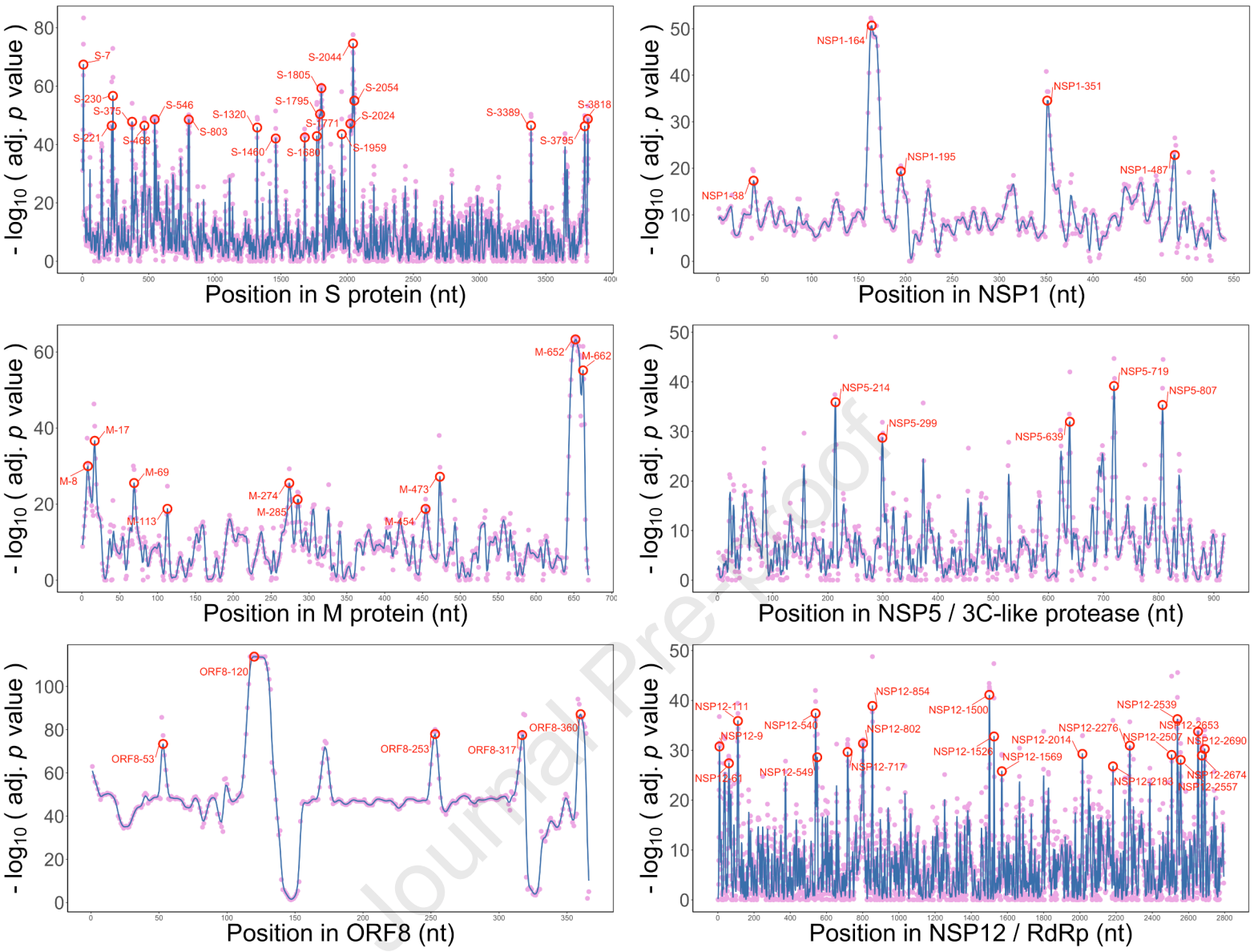


Figure 2

A



B

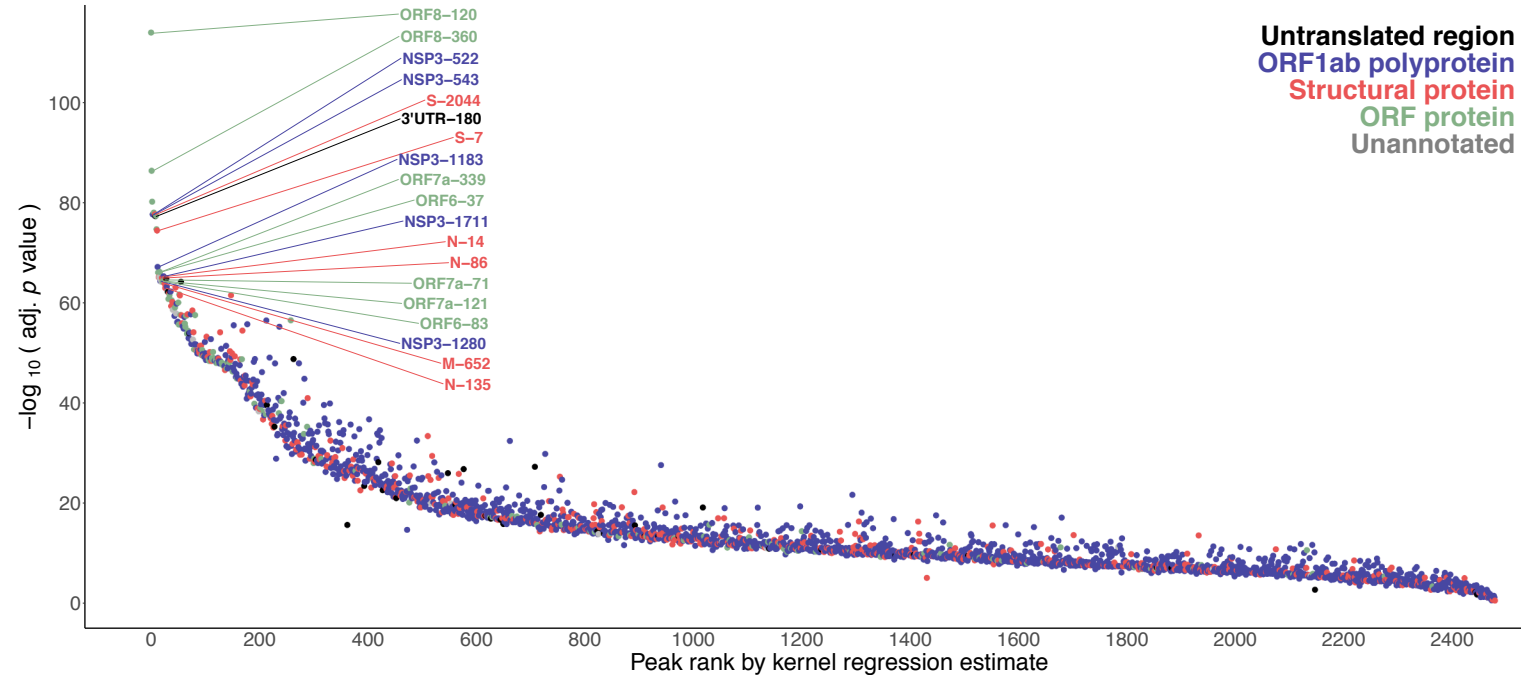


Figure 3

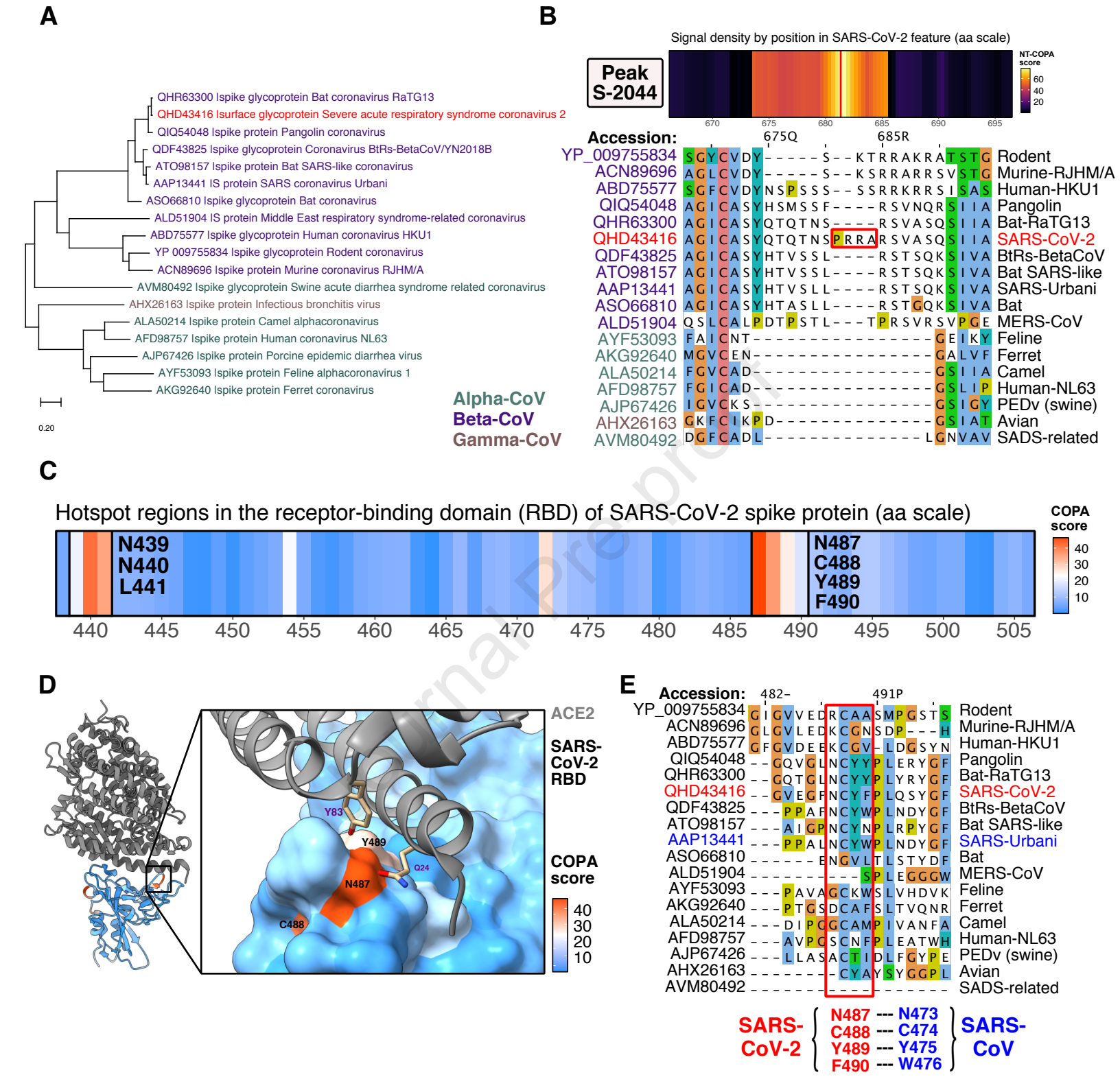


Figure 4

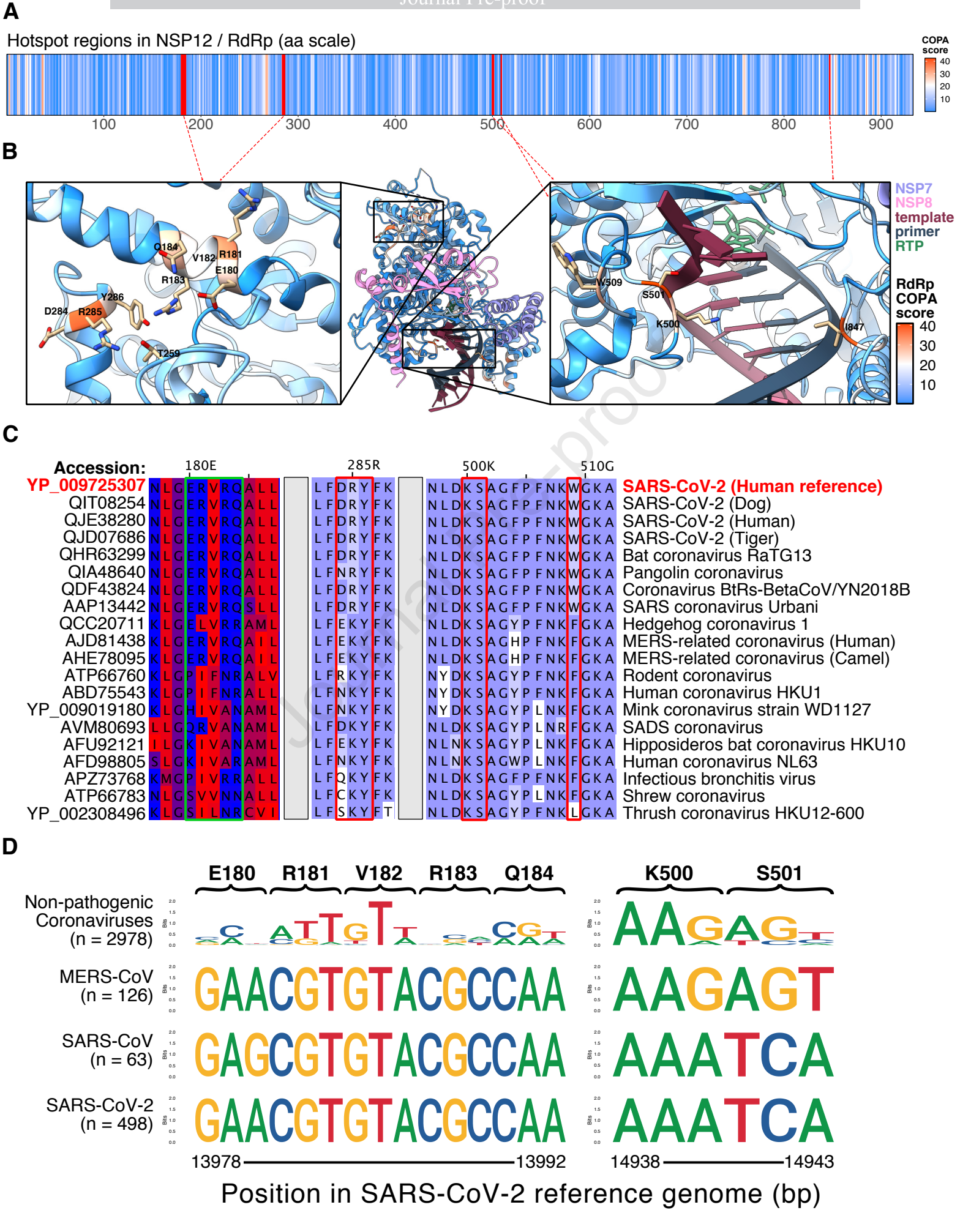
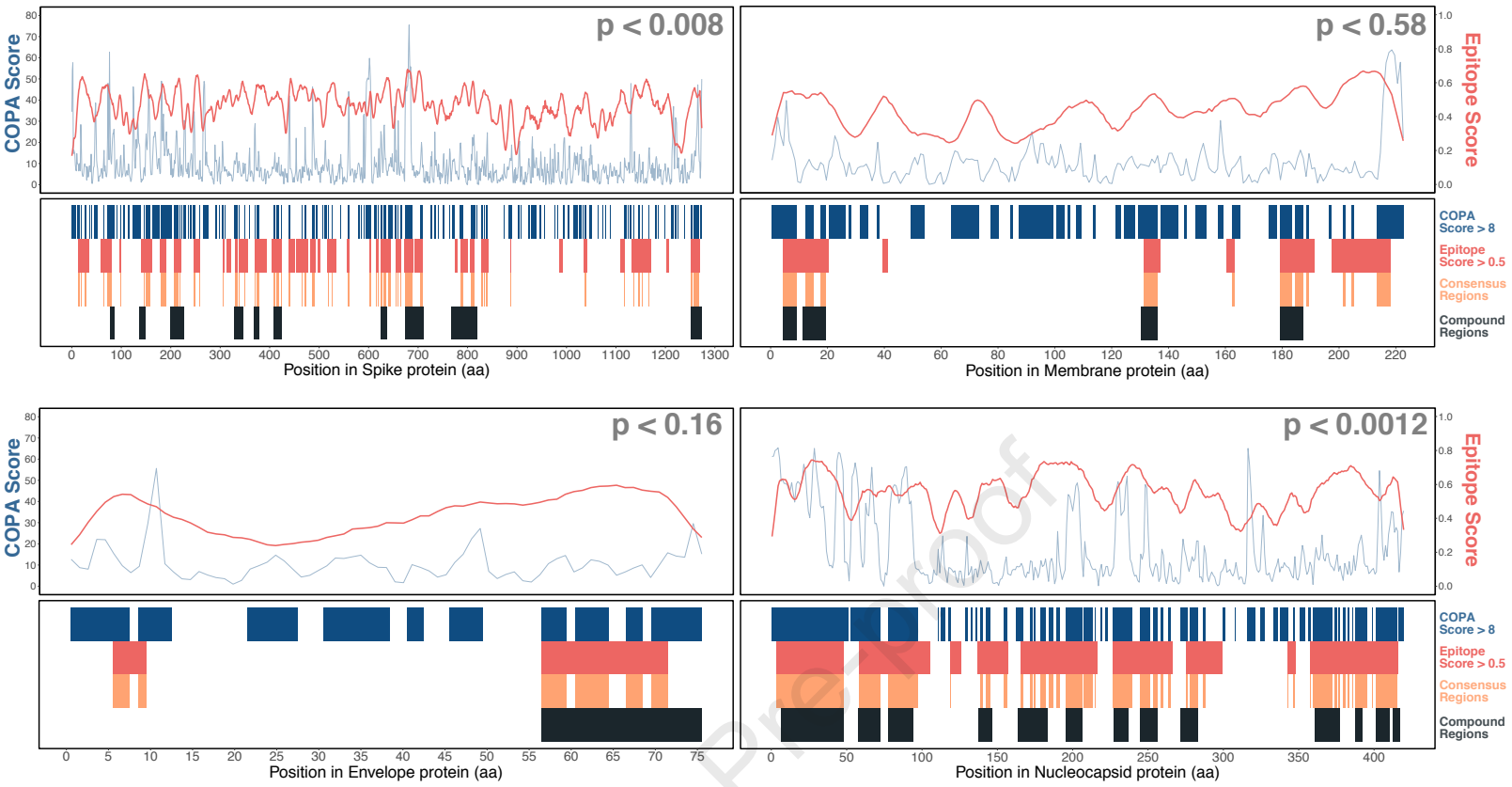


Figure 5

A



B

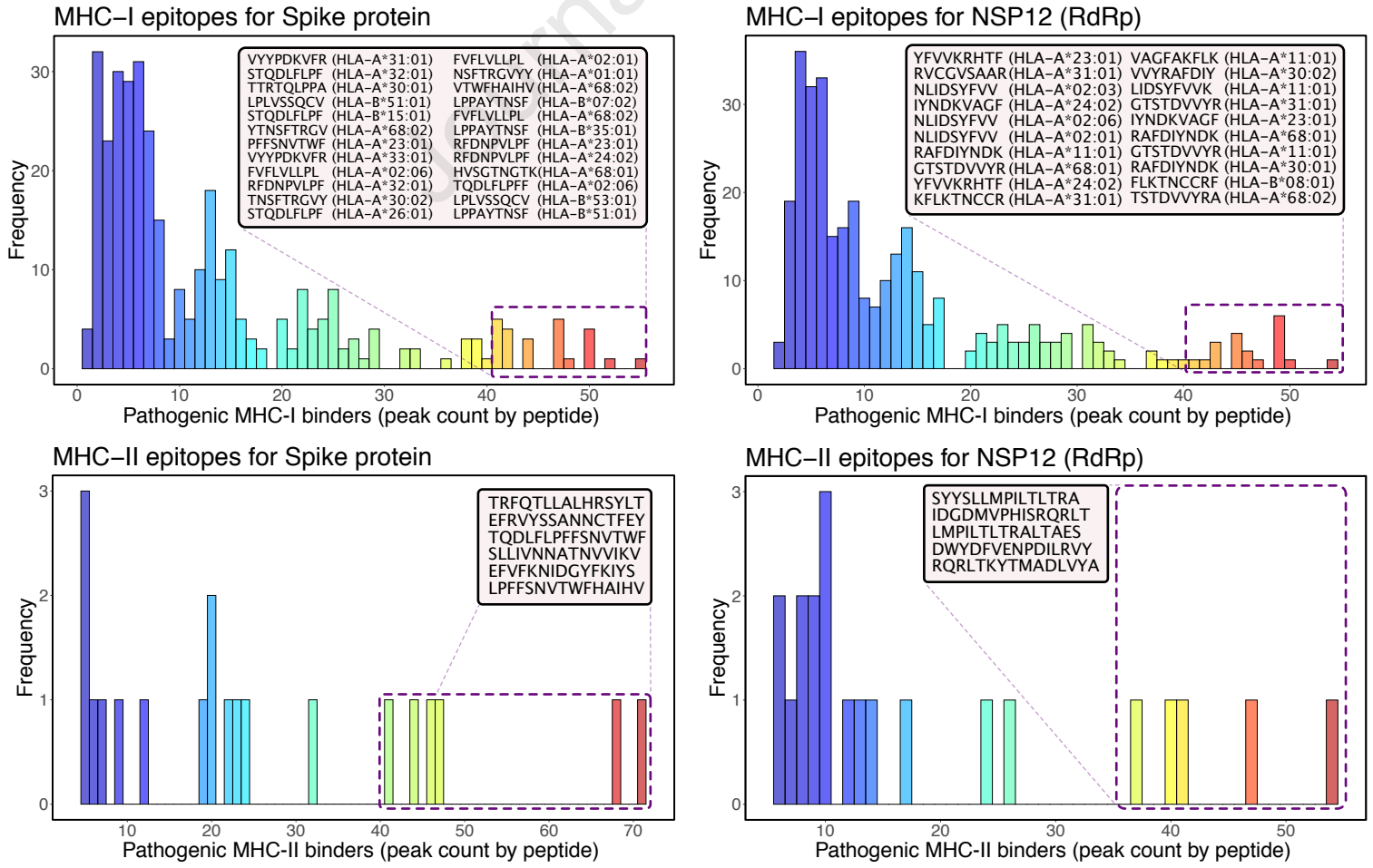
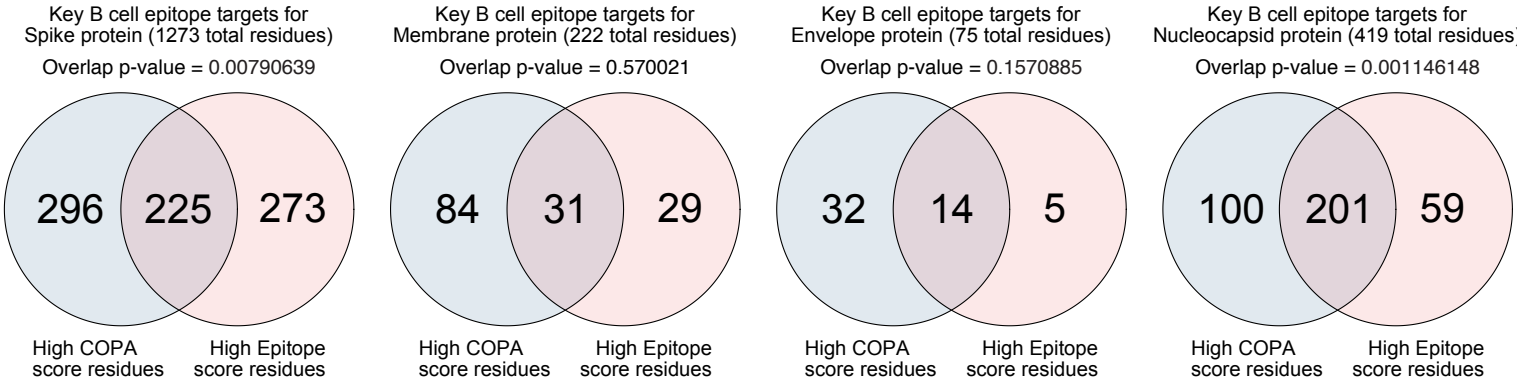
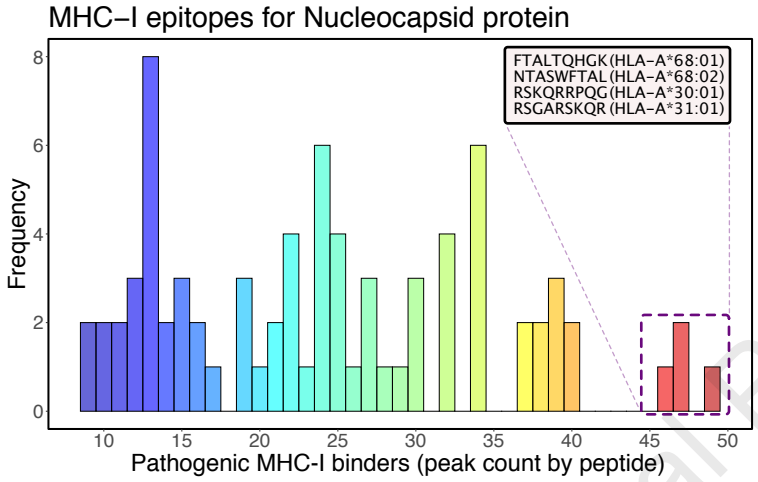


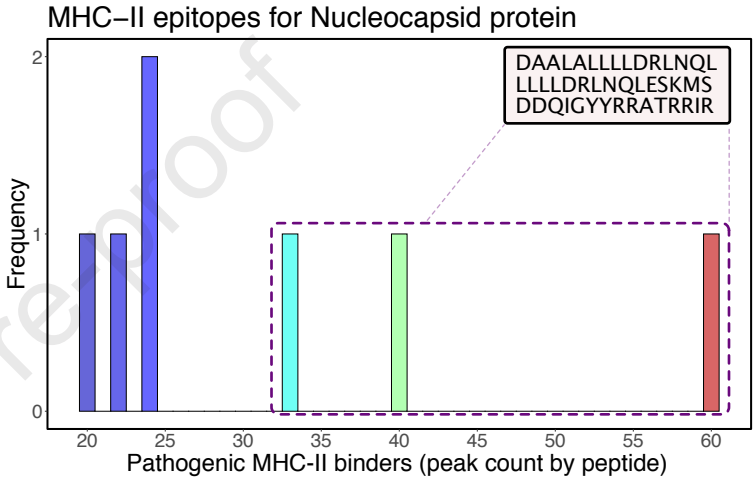
Figure 6

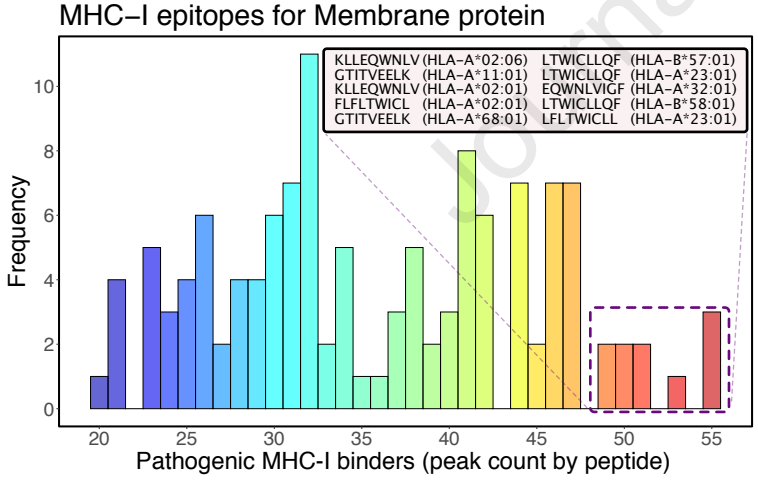
A

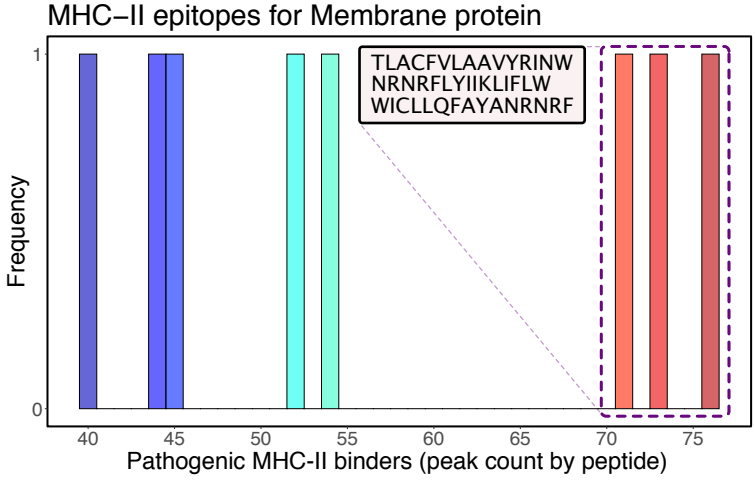


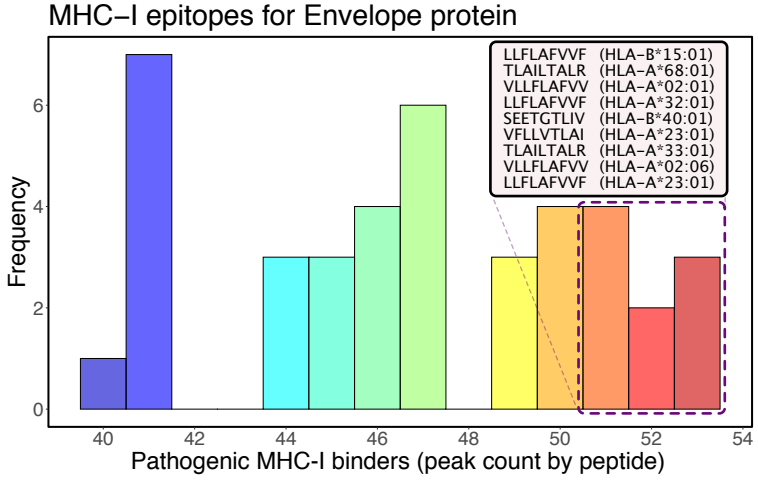
B











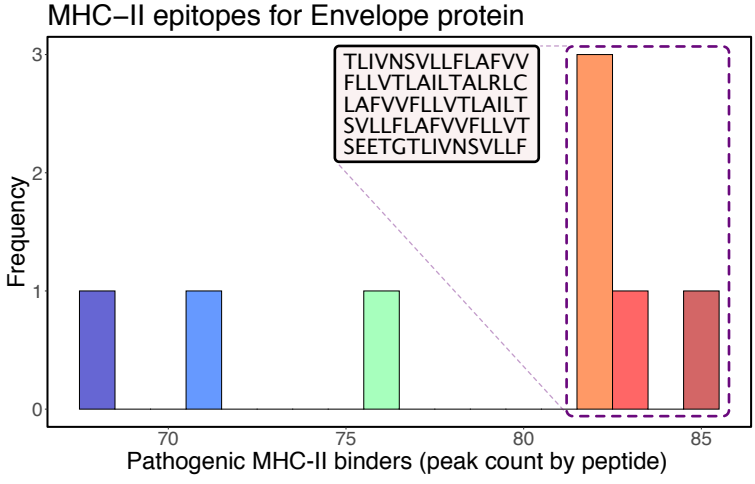
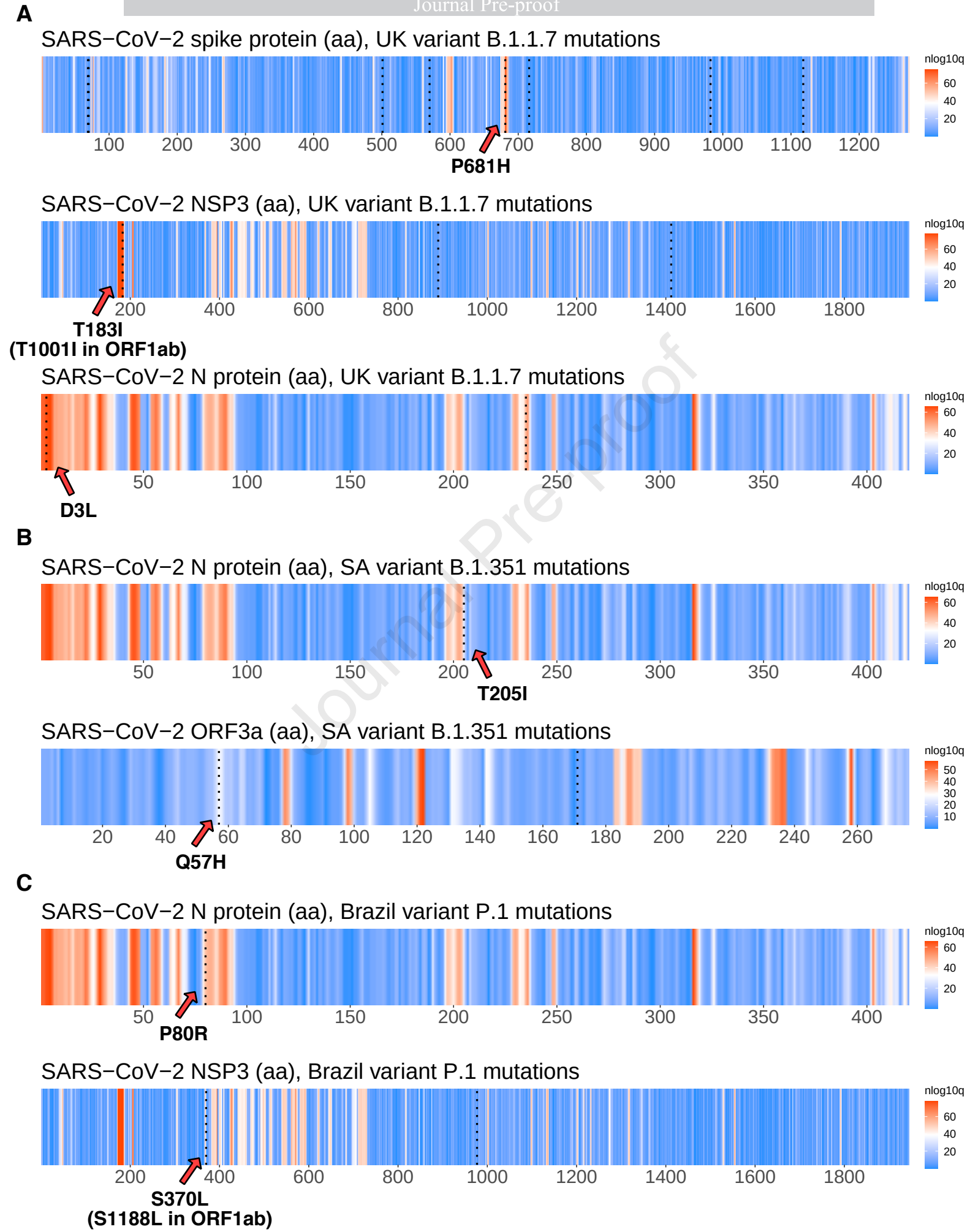


Figure 7



The bigger picture

Identifying which genomic regions of the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virus are pathogenic remains a major challenge in COVID-19 research. However, there is currently a lack of systematic and unbiased methods for such functional characterization. In this study, we set up a machine learning based approach to identify which genomic regions distinguish SARS-CoV-2 and other high case fatality rate (CFR) coronaviruses from other coronaviruses. Discriminative scores were obtained for every nucleotide in the SARS-CoV-2 genome. We then performed a series of evolutionary and structural analyses of candidate hotspots, as well as integrative analyses with predicted B cell and T cell epitopes and emerging variants of concern. Our approach can be extended to other viral genomes or microbial pathogens to gain insights on which sequence features are pathogenic or immunogenic.

Highlights

- Machine learning identifies discriminative signatures in coronavirus genomes.
- Hotspots in key viral proteins have evolutionary and structural significance.
- Integration of hotspots with B cell and T cell epitopes identify joint features.
- Hotspots correlate with emerging variants of concern for mutation prioritization.

eTOC blurb

To identify key pathogenic regions in coronavirus genomes, this study developed machine learning approaches and provide a systematic map of predicted descriptive genomic features in SARS-CoV-2.