Covariance predicts conserved protein residue interactions important to the emergence and continued evolution of SARS-CoV-2 as a human pathogen

William P. Robins[2] and John J. Mekalanos[1]

[1,2]Department of Microbiology, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115

[1]Corresponding authors

Abstract

SARS-CoV-2 is one of three recognized coronaviruses (CoVs) that have caused epidemics or pandemics in the 21st century and that likely emerged from animal reservoirs. Differences in nucleotide and protein sequence composition within related β-coronaviruses are often used to better understand CoV evolution, host adaptation, and their emergence as human pathogens. Here we report the comprehensive analysis of amino acid residue changes that have occurred in lineage B β-coronaviruses that show covariance with each other. This analysis revealed patterns of covariance within conserved viral proteins that potentially define conserved interactions within and between core proteins encoded by SARS-CoV-2 related β-coranaviruses. We identified not only individual pairs but also networks of amino acid residues that exhibited statistically high frequencies of covariance with each other using an independent pair model followed by a tandem model approach. Using 149 different CoV genomes that vary in their relatedness, we identified networks of unique combinations of alleles that can be incrementally traced genome by genome within different phylogenic lineages. Remarkably, covariant residues and their respective regions most abundantly represented are implicated in the emergence of SARS-CoV-2 are also enriched in dominant SARS-CoV-2 variants.

Introduction

The prior emergence of SARS-CoV and MERS-CoV as human pathogens is attributed to zoonotic viruses that transferred from bats to civets and camels, respectively, while SARS-CoV-2 is most similar to viruses isolated from both bats and pangolins [1–6]. The ~30kb genome size of all SARS-related CoVs renders sequence alignment and pairwise distance methods effective for phylogenic studies and determining genetic events that correlate with their adaption to the human host. While nucleic acid sequence-based phylogenies are informative, they clearly have limitations as not all single nucleotide polymorphisms are equal. For example, single-strand RNA viruses possess significant base-pairing in regions of their genomes that can result in different fitness costs even for synonymous mutations because higher-ordered RNA structures and non-coding regions can impact replication, transcription, and recognition by the host immune system [7,8].

Nucleotide polymorphisms also distinctly influence encoded amino acid (AA) residues depending on their position in a codon and thus further protein expression is often correlated with codon frequency or cognate tRNA abundance [9]. Similarly, codon usage has been studied in the context of mutational pressure and natural selection [10–13]. For example, the presence of repeating rare codons in the SARS-CoV-2 Spike protein corresponding to the furin cleavage site (FCS) has been explored as circumstantial evidence for genetic manipulation [14]. Mutational pressure rather than translational selection is however reported in other work to be one dominant factor in the observed codon usage in other human RNA viruses [15,16]. In RNA viruses, the host immune system can impose additional selective pressure on genomic nucleotide content; thus the high frequency of A- and U-ending codons and underrepresentation of CpG dinucleotides in RNA viruses including CoVs is attributed primarily to cytosine deamination and the pressure exerted by innate immune mechanisms [17–19]. Comparative analysis of SARS-CoV-2 to closely related CoVs suggests that C-to-U conversion played a significant role in the evolution of the SARS-CoV-2 [20]. In sum, codon usage, nucleotide sequences, and the AA content may affect virus adaptation with the latter being most relevant to protein folding, stability, function, and adaptive immune recognition in the host.

Many previous and ongoing efforts to study human CoV virus-host interactions are focused on AA residues and domains within the quaternary and tertiary structure of the Spike trimer protein. For SARS-CoV, the stepwise adaption from the ancestral bat CoV to variants that infect civet, human, and even laboratory mice is well-understood and can be traced to the domains in the Spike protein that confer specificity to the host, especially in the context of residues within the receptor-binding domain (RBD) [21,22,3]. An ancestor to SARS-CoV-2 is yet to be established with certainty, but both residues within the Spike RBD that interact with the host receptor angiotensin-converting enzyme 2 (ACE-2), and a unique furin cleavage site are believed to have contributed to its adaption to the human host and its enhanced transmission [23–25]. RATG13, one of the closest bat CoV relatives to SARS-CoV-2 that shares ~96% nucleotide identity [26], is measured to have a reduced affinity for human ACE-2 when both are compared and this is in part conferred by residues in Spike [27]. However, molecular evidence for ACE-2 affinity being the primary determinant in host specificity for CoVs is also confounding. SARS-CoV and SARS-CoV-2 viruses that infect human cell lines via the ACE-2 receptor are found to vary in their ability to infect bat cell lines suggesting that the host range of β-coronaviruses is not only specified by Spike RBD-ACE-2 interactions [24,28].

Evidence for the selective adaptability and the plasticity of Spike protein domains has been documented by the existence of single and multiple mutations that have been temporally enriched in newly dominant variant lineages during the ongoing pandemic. For example, the Spike D164G allele, a stand-alone defining mutation of the dominant SARS-CoV-2 A2a clade that emerged early in the pandemic, has been demonstrated to increase viral fitness and infectivity, possibly by influencing proteolytic processing, incorporation of Spike protein in the virion, and conformational states of the Spike protein [29,30]. Subsequent dominant emerging variants within this clade notably possess additional mutations in Spike and other genes. Moreover, a broad and diverse collection of variant mutations are represented independently in other distinct lineages [31,32]. In Spike, some of these mutations are attributed to immunity evasion, host-receptor interactions, and Spike structure and conformational dynamics [33]. Distinct attributes or roles of each of these mutations are yet to be entirely elucidated and the actual extent of the contribution of any allele may be intricate. Importantly, it is not yet known what is the contribution of each of these single mutations in the context of other mutations and in general, the effects of mutations in many proteins other than Spike have been less studied or are completely unknown.

Here we report the results of an investigation that sought to use the evolutionary history of sarbecoviruses to identify the most-conserved interactions between AA residues in key proteins encoded by CoV viruses most related to SARS-CoV and SARS–CoV-2. Specifically, we identified all covariant AA pairs and also larger correlated tandem model-based networks (clusters) of AA residues that exhibited statistically high frequencies of covariance with each other. We examined conserved covariance between protein sequences to uncover new insights into CoV evolution through the identification of apparent inter and intra-protein interactions. We propose that these covariant interactions of residues are important for virus evolution and may drive adaption to other hosts and influence transmission and pathogenicity by emergent variant viruses.

Results

Estimated phylogeny of β-coronaviruses with completed genomes

We first estimated the extent of the evolutionary relatedness of 169 β-coronaviruses using whole-genome nucleotide sequences (**Figure 1B**). We postulated covariant amino acid (AA) residues play diverse roles in viral protein structure, interactions, and functions or instead may be a consequence of mutational accumulation and drift that is not biologically relevant to viral protein function. Phylogeny and relatedness of genomes are recognized to bias observed apparent coupling of AA mutations and influence covariation [34–38]. By generating a phylogenic tree to assist in identifying such phylogenic effects, AA variability at covariant residues can also be traced using tree topology and even branch length and therefore analyzed in the context of evolution [38,39].

We aligned the deposited nucleotide (nt) sequences of 169 unique lineage β-coronaviruses between the initiation codon of *Nsp1* of the 16 polyprotein-encoding gene *Orf1ab* through to the termination codon of the *N* gene that encodes the nucleocapsid protein **(Genome Accession numbers listed in Supplemental**

File S1). The 30,553 nt gapped alignment of these strains spans all core and accessory genes except for the hypothetical gene *ORF10* downstream of *N*. *ORF10* is not predicted to be conserved in all represented CoVs in this group and is not essential for SARS-CoV-2 *in vitro* or *in vivo* [40]. In their entirety, CoVs represented in our analysis were isolated from bat (147), Civet (10), Pangolin (6), and human (6). SARS-CoV is notably represented as both civet and human isolates and SARS-CoV-2 is represented by an initial reference sequence isolated in Wuhan during December 2019.

The topology and structure of the maximum likelihood (ML) tree (Figure 1B) generated by our alignment is largely consistent with published work using aligned variable regions of the genome [26,41–45]. Though certain genomic regions are predicted to be prone to recombination events that can lead to mosaicism during the evolution in CoVs [46–49], we did not alter our phylogenic analysis based on the exclusion or inclusion of any single core gene or gene region, apart from hypothetical gene *ORF10*. In an unrooted tree, distinct lineages are apparent; several CoVs are most related to SARS-CoV, a set of sublineages of CoVs more closely related to SARS-CoV-2, and a third distinct group of β-coronaviruses more closely related to HKU-3. Only a small subset of viruses in this analysis are predicted to be most closely related to SARS-CoV and SARS-CoV-2 relative to all other CoVs represented in this phylogeny. How these viruses differ from those more distantly related in the context of amino acid residue covariance was one aim of the comparative analysis presented here.

The relatedness of SARS-CoV-2 to certain bat and pangolin CoVs supports the emergence of this virus from a zoonotic reservoir. Using our nucleotide alignment-based tree, we identify RATG13 and other more recently identified bat CoVs from Laos to be most closely related to SARS-CoV-2, designated as Group 1. Clusters of other pangolin and bat CoVs, some nearly clonal, comprise the next tier of related assemblages designated as Group 2 and 3, respectively. A small number of other CoVs, designed Groups 4 and 5, are less closely related to those in Groups 1-3 CoVs but also distinct from other SARS-CoV and HKU-3-related viruses in our generated phylogeny. All CoVs in groups 1-5 are more likely to share a common ancestor with SARS-CoV-2 and we propose mutational and/or recombination events and also selective processes have generated the observed diversity within this subclade (**Figure 2A and 2B**).

Alignment of conserved proteins in β-coronaviruses

We selected 149 CoVs represented in our phylogenic reconstruction based on the availability of annotated proteins and aligned the amino acids of core proteins using identified open reading framed (ORFs) common to all genomes. This includes the conserved CoV polyprotein genes called *ORF1a* and *ORF1b* which together encodes at least 16 smaller non-structural proteins (nsps) when processed by a viral protease. Others include genes for *S* (Spike), the two viroporins *ORF3a* and *E* (envelope), *M* (membrane), and *N* (nucleocapsid). Our goal was to align the AA sequences of these proteins to identity covariant pairs and residues within and between core non-structural and structural viral proteins. The alignment resulted in a 9458 AA consensus sequence with only 1.5% of sites being gaps with low coverage and 2% with residue conservation of less than 80%.

When the AA sequences for each protein are aligned and compared, the degree of conservation varies between each protein and also within individual and discrete protein domains. Genes that encode nonstructural proteins (nsp)s with roles in RNA metabolism and genome replication (i.e., nsp12-14) are among the most highly conserved in AA identity. Others that encode nsp2, nsp3, and nsp4 are highly variable. The NTD and RBD of Spike show the most significant variability in both residue identity and length. In contrast, the sequence of much of the CTD of the S1 and the entire S2 subunit of Spike is highly conserved. The channel-forming E (envelope) viroporin protein is one of the most highly conserved proteins in contrast to the viroporin ORF3a that exhibits high variability in its NTD and other CTD subdomains. We propose that this is evidence of evolutionary pressure on residues in CoV proteins and that certain protein subdomains may be under higher selective pressure than others.

In addition to greater AA sequence variability in some genes, individual residues and continuous sections of AA are either uniquely present or absent in a portion of CoVs. One key aim of our study was to identify and evaluate the covariance of all residues in our selected proteins among these 149 CoVs without bias. We predicted that his analysis might reveal the existence of critical amino acid residue conservation or changes that would possibly correlate with changes in the virus-host range or biological properties. In this regard, SARS-CoV-2 is recognized to possess unique sequences not present in other closely related CoVs including those that define and enhance a furin cleavage site (FCS) of the Spike protein [25,50]. After binding to the ACE-2 receptor, cleavage of S by furin and or other proteases is critical to conformational changes that allow viral envelope-host membrane fusion and subsequent viral RNA entry into the host cytosol [25]. Other changes in S are less understood. For example, certain residues in the NTD of SARS-CoV-2 Spike are present in other unrelated CoVs but are notably absent in the corresponding regions of SARS-CoV [51]. Conversely, there are residues in many CoVs with no positional equivalent in SARS-CoV-2. To address gaps in the alignment, we have indicated such residues with a “Z” designation to accommodate possible covariance between residues and such deletion occurrences.

Extracted networks of covariant residues are informative about evolutionary relationships in CoVs

Covariance is a quantitative measurement of how often the identity of one AA is correlated to the identity of another AA or AAs in either the same protein or in a completely different protein [52–54]. Because covariant AA residues change in concert with each other, they can define critical AA-AA residue interactions within a protein or in homologous or heterologous protein-protein pairs or instead indicate phylogenetic relatedness based on their co-existence [37,38]. These putative residue interactions may provide new insights into the evolution and relatedness of the CoV family of viruses [55]. To survey the frequency of covariance among a reference collection CoVs, we identified correlative pairs and also assembled groups of three or more (here designated as ‘clusters’) of covarying amino acid residues using a correlating tandem model [56]. These clusters are not typically generated using other typical pairwise algorithms tailored for determining protein structure or docking interfaces. We chose the FastCoV approach for its distinct quality in identifying larger networks of putative compensatory mutations generated by selection and adaptation which seems well-suited for studying the emergence of viruses similar to SARS-CoV and SAR-CoV-2 [56]. This differs from DCA-based and other covariance approaches used to predict co-evolving residue interactions that may use corrected and weighted correlative data that can also be coupled with other various predictive secondary structure motifs to assist in protein structure and interaction predictions [57]. Our approach simply provides a raw covariance purity and percentage score for pairs and larger networks of covariant residues with no goal for structure-based predictions. We selected a purity threshold (0.7) based on our small sampling size and extracted 973,649 unique pairs and 741 clusters. In this preliminary analysis, we identified a collection of gaps and also unique sequences selectively conserved in some CoVs and we concluded that such deletions or insertions, like residues, may also covary with AA sequences in proteins. Deletions were temporarily substituted as rare alternate amino acids in the alignment and covariance was analyzed to reveal putative covariance between all residues and also deletions. This expanded the total number of unique correlating residue pairs (1,089,836) and clusters (769) (**Supplemental File S2)**. We identified many deletions that correlated with AA residues and also other deletions.

All CoVs genomes, clusters, and residues were graphed using a force mapping algorithm (**Supplementary File S7**, shown in **Figure 1A**). This interactive graph facilitates the extraction of clusters and respective residues and deletions uniquely present to different groups and subsets of CoVs. Remarkably, the spatial organization of graphed CoVs is highly consistent with our phylogenic estimate based on nucleotide alignment (**Figure 1B**). CoVs most closely related to SARS-CoV, SARS-CoV-2, and HKU-3 are spatially positioned close to one another in each group solely based on shared covariant residues. Other CoVs that vary regarding relatedness are distributed in between these three indicated groups. Because some covariant pairs and clusters are entirely inclusive to single groupings or instead shared between certain CoVs, we conclude that covariant residues can be enriched through a common evolutionary history such as ancestry or can be selected by adaption to a specific host(s).

To provide information about the phylogenic distribution of any cluster that may be due to ancestry, an average taxonomic distribution score (ATDS) was calculated for each cluster based on the number of CoVs present in a given cluster and their average distribution based on branch lengths estimated in the ML tree (**Supplemental File S3)**. Though this score is relative and also determined by the relatedness of all CoVs in the phylogenic reconstructions, clusters and their respective alleles with a larger ATDS are more broadly represented in the evolutionary record within the scope of these 149 β-coronaviruses analyzed. A small ATDS value indicates these covariant residues in a given cluster are restricted to CoVs that are very similar or almost identical. We predict this class of clusters may be biologically informative about covariant residues specifically enriched in SARS-CoV and SARS-CoV-2 and their respective relatives. Conversely, clusters with large ATDS values are those clusters with residues that are present in more evolutionary disjunctively distributed single or groups of CoVs. These may be the result of divergent or independent selective events or are instead conserved covariant residues that have persisted during the evolution of CoVs and are possibly ancestral or even essential to the lineage of these viruses.

Of the 1,089,836 unique pairs with varying degrees of residue identity at each two positions, we identified 522,336 correlative AA residue pair positions and also calculated the number of unique amino acid identities that can exist for each position in the pair (tabulated in **Supplemental File S2)**. This degree of residue representation of each pair varied between a minimum of two (481,024) and a maximum of seven (2). Only ~8% (41,312) of all pairs are represented by three or more unique residue identities and this representation drops significantly stepwise for each unique identity between three and seven. We hypothesized that the increased number of independent residue pairs represented at any two correlative positions in the evolutionary record increases the probability that there is a true interacting relationship between such residues.

The position of every covariant residue in the alignment was mapped to the respective residue position in SARS-CoV-2. For an overwhelming majority of residues that show high conservation and are present among the 149 CoVs, this translational numbering assignment based on the sequence alignments is straightforward. For residues in less conserved regions such as those that exist as gaps or insertions in some CoVs including SARS-CoV-2, this created residues positions that are represented by gaps for sites missing in SARS-CoV-2 and duplicate numbering. For example, several CoVs possess between two and eight additional residues between the aligned numbered positions of AA 7 and 8 of SARS-CoV-2 Spike. If any of these residues are covariant with other residues or gaps, the use of SARS-CoV-2 residue numbering necessitates either no assignment or the number of the residue that flanks the missing residues in SARS-CoV-2 to preserve the information position of gapped-residue covariance. We chose the latter to indicate the position, thus some residues may appear to be duplicated or even covariant with themselves when SARS-CoV-2 numbering is employed. We have provided tables that indicates these positions to identify such occurrences (**Supplemental File S2**).

The presence of more variable covariant residues in other proteins varies significantly. For pairs with at least five independent identities (319), nearly half (154) of these are located in the Spike protein. Various structures of Spike trimer are elucidated and well-studied due to the roles in receptor recognition, cell entry, and interactions with monoclonal antibodies [23,58–61]. Using available high-resolution PDB structures, we screened for predicted interacting residues and then referenced our identified covariant pairs to establish a correlation between the number of unique identities in pairs. As observed in the alignment, Spike sequences and high order structures vary between CoVs, and we adjusted scoring both for directly interacting residues and those directly adjacent by one residue position (**Supplemental File S5)**. Residue pairs with increased representation are more likely to interact or be in close proximity to one another in the Spike trimer protein (~24%) than those represented by only two identities (~5%). This provided confidence that residues with increased representation are more likely to have direct interactions with their identified cognate pairs.

AA covariance in the CoVs closely related to SARS-CoV-2 is enriched in Spike

We examined the identity and distribution of covariant residues within the lineages of CoVs most closely related to SARS-CoV-2 identified in this work designated as Groups 1-5 (**Figure 2A**). We reasoned that the selective pressure imposed on the AA identity of truly covariant residues should be different than for all other residues. Thus the collection of putative covariant AAs in SARS-CoV-2 and other CoVs that share a common ancestor provides a new perspective about the evolutionary relationships between these viruses.

Group 5 exhibits the most numerous and distributed covariant residues identified in distinct clusters (Figure 2C). This is not surprising based on the apparent evolutionary divergence within these CoVs when compared to Groups 1-4 (**Figure 2B**). Both Group 2, which is entirely represented by Pangolin CoVs, and Group 3, all bat CoVs, similarly exhibit a greater number of covariant residues when compared to Group 1, also likely due to differences in the overall relatedness of CoVs. The CoVs represented in Group 1 exhibit high conservation and some members are nearly clonal. For these, nearly all covariant residues in this group are restricted to Spike and ORF3a, except for two residues in Nsp3 (AA 149 and 175) and a single residue in Nsp15 (115).

Because CoVs in Groups 1-3 are most closely related to SARS-CoV-2 based on their nucleotide identity, we focused on these and examined covariant residues in Spike (**Figures 3A-C**). Residues in Spike are recognized to be among the most relevant in the emergence and persistence of SARS-CoV-2 and also implicated in host adaption in both SARS-CoV and SARS-CoV-2 [43,62]. We vetted residues in Groups 1-3 because we hypothesized these covariant alleles might be important for selection, adaptation, and viral fitness for the most closely SARS-CoV-2-related viruses in human, pangolin, and bat hosts. Furthermore, we identified covariant residues co-present in two or in all three of these groups. By definition, the AA identity of individual covariant residues is not highly conserved in CoVs, but instead, their conserved identity varies with other residues. Thus any covariant pair or cluster of residues may indicate a direct or indirect conserved interaction between AAs important during the adaption of a CoV. We find a common subset of conserved covariant residues between both bat and pangolin CoVs with those closely related to SARS-CoV-2 in Group 1. These may indicate specific interactions between residues and residue identities especially relevant to the biology of SARS-CoV-2 and related CoVs.

A majority of the Spike-specific covariant residues common to clusters in Groups 1-3 are located within discrete domains primarily in the S1 domain (**Figures 3B and 3C**). These regions are in the NTD (AA 67-112, 137-155, and 239-271), RBD (AA 439-508 & 529-589), and CTD of the S1 subunit (AA 632-640), and also at the FCS within the S1/S2 subunit boundary (AA 675-690). The NTD, RBD, and FCS are also notably enriched in both mutations and deletions identified in dominant variants of SARS-CoV-2. For example, the deletions and flanking mutations at AA positions 69-70, 142-145, 156-157, and 241-253 found in SARS-CoV-2 dominant variants align well with enriched covariant residues, including deletions, in these three groups. In regions of very sparse covariance such as AA 529-590, two variant mutations 547 (Omicron) and 570 (Alpha) also align with covariant residues in Group 1. Conversely, other mutations in current SARS-CoV-2 variants do not align with these enriched covariant residues. True covariant residues require additional compensatory changes at other residue positions and we expect a portion of these residues to be less mutable than other noncovariant residues with low conservation.

Clinical and pan covariant residues are similarly represented

The resolution and extent of our pan-CoV covariance analysis are in part defined by the number of distinct genomes and also their relatedness. Roles for all identified covariant residues in all proteins cannot be readily ascertained, but this generated data may be further validated as more SARS-CoV-2 protein structures become available. Because of the vast scope and magnitude of SARS-CoV-2 infections during the ongoing pandemic and availability of whole genomes sequenced, we supposed covariant residues may also be apparent in the millions of sampled clinical strains over 18 months. Genes enriched residues with covariant relationships that appear to be co-present in both the pan and clinical covariant analysis are more likely to be of special interest to SARS-CoV-2 biology. We extracted and stringently selected whole-genome sequences that were nearly or entirely complete to avoid artifacts that may bias our analysis and then compared the positions of covariant residues of 252,102 randomly selected sequences deposited between December 2019 and August 2021 (**Supplemental File S4)**.

Due to the near clonality of SARS-CoV-2 sequences, we were unsurprised to find only 1.2% of the total covariant residues when compared to those identified in pan-CoV analysis. 13,041 uniquely represented pairs of AA residues can be reduced to 6,137 correlative pairs for all proteins. As observed in the pan-analysis, the distribution is exceedingly skewed toward several genes encoding proteins including Spike. When the distribution of every single residue identified in both the pan and clinical analysis is compared gene-by-gene, regardless of the positions of correlative partner, we discovered the contributed representation for each encoded protein follows a similar trend (**Figure 4A and 4B**). Genes encoding nsp5 (3CL-pro), nsp7-nsp16, Envelope, and Membrane proteins are sparse in coverage of co-identified single residues. In contrast, covariant residues are most abundant in frequency in genes encoding nsp1-3, Spike, and ORF3a. Remarkably, in either category, there are proteins and specific regions of proteins similarly enriched in the distribution of covariant residues for both analyses. When the co-occurrence for each residue is quantified for the entire protein, the observed overlap between clinical and pan covariance is found to be statistically significant for nsp2-4, nsp13, nsp16, Spike, and nucleocapsid. For proteins with overlap measured to be above the threshold of significance, such as nsp1, envelope, and ORF3a, this graphing allows us to observe both similar patterns and frequencies of covariant residues across the protein. Conversely, for nsp6, nsp10, and membrane proteins we see no significant similarities in residue distribution or patterns.

Mapping of identified conserved pairs in both analyses

We accounted for the distribution of intra- and inter-protein residue pairs identified in both analyses (**Figure 5A**). We reasoned these pairs are more informative about conserved residue interactions than the distribution of single residues mapped per protein. As with single residues shown in **Figure 4**, the distribution of linked residue pairs is not uniform and the density varies significantly by protein and within protein regions. The majority of linked pairs are mapped to genes encoding nsp1, nsp2, nsp3, Spike, ORF3a, and nucleocapsid. For genes encoding nsp4 and nsp16, the occurrence is sparse or absent in residue pair representation. Of 538 total pairs (**Supplemental File S6)**, 90% (485) are represented by residues within the same protein (**Figure 5B**), are frequently proximal or adjacent to each other by position, and are not random in distribution. We expect this bias as most interacting residues should be present within the same protein. The remaining 10% (53) intra-protein pairs are similarly clustered and nonrandom in their positional enrichment (**Figure 5C**). Intra-protein residues in nsp3, Spike, and nucleocapsid are linked with the most diverse partner proteins (**Figure 5F**). In contrast, nsp12, nsp13, nsp15, nsp16, and envelope possess only one or two intraprotein pairs.

Evidence for interactions within and between Spike and ORF3a linked to viral emergence and adaption

The enrichment of residue pairs in the subdomains of Spike and ORF3a were of special interest (**Figure 5D**). First, the distribution and abundance of these residues are similar to the covariant residues we identified within the CoVs most related to SARS-CoV-2 (**Figure 2C**). Furthermore, of these 224 residues, 40 are identified in the 88 Spike and ORF3a residues as 31 pairs present in dominant variants circulating including Omicron. When 31 residue pairs are mapped to Spike and ORF3a, most links are enriched between the NTD and RBD of Spike with two notable links between the Spike NTD and AA 26 in the NTD of ORF3a (**Figure 5E**).

We find evidence that subsets of these 31 residue pairs likely interact directly or are positioned proximal to one other within particular regions of the Spike protein (**Figure 6A-E**). In a solved quaternary structure of the Spike trimer PDB (7JJI.PDB), NTD residue 20 is adjacent to its covariant pair residue 138 in Spike Cryo-EM reconstruction (**Figure 6D**). Moreover, residues 17 and 21 interact directly with 138 Residues between 138 through 157 include identified covariant residues in this work that are notably deletions and/or mutated in dominant variants. Similarly, covariant residues 241 through 252 are also frequently deleted and/or mutated and these directly interact with 138-157. Residues 248-250 are identified in our work to covariant with residue 75. With residue 75, deletions and mutations between residues 65 and 82 are also among the most abundant identified in dominant circulating variants. Residues 212 and 215 reside in yet another covariant hotspot and have covariant pairings with residues 142 and 241/242, respectively. Curiously, based on structure, AA residues 212-215 have no apparent direct interactions with 138-157, 241-252, or 67-75. All of these mutation and deletion hotspots in Spike NTD have generated much interest regarding their roles as superantigens and the escape from neutralizing antibodies (more below). The stand-alone identification of these in both pan and clinical covariance analyses and co-occurrence in dominant circulating variants indicate these are often under significant selective pressure to either mutate or become absent by consequence of in-frame deletion, often in the context of distal residues.

Spike protein remarkably accounts for 58% of residues in our identified 538 coincident pairs and also is recognized to possess the majority of mutations in dominant variants. We examined all 538 residue pairs and then cross-listed the occurrence of each residue in dominant variants to identify those in all proteins. 119 (30%) of 394 total represented residues are found in 15 dominant variants including the two current Omicron sublineages [31]. When both residue pairs are present in a dominant variant (49), we find 76% of these are remarkably present together in the same variant lineage (**Figure 7**). This observation is suggestive of covariance pressure operative in the emergence of variants during the ongoing pandemic.

Discussion

For both nucleotide and amino acid identity-based approaches, the sequence conservation in either complete or partial regions of CoV genomes continues to be applied to understand the relatedness between β-coronaviruses and SARS-CoV-2. This extends to the emergence of SARS-CoV-2 as a human pathogen responsible for a global pandemic and its continued adaptation. In this work, we examined the conservation of correlative covariant pairs and even clusters of amino acids that appear to change in concert with one another across the entire genome. We acknowledge apparent covariant mutations can also be a simple consequence of spontaneously emerged mutations and other common concurrent mutations at less conserved sites. Conversely, these could be a result of spontaneous mutations that imposes a selection for one or even more compensatory mutations at other sites to maintain or even increase viral fitness. Both instances are certain to be present in this dataset. An applied hypergeometric probability distribution predicts the overlap of 538 covariant pairs between the pan and clinical datasets is exceedingly significant (**Figure 7**). We acknowledge that the conservation and essentiality of amino acid residues vary significantly in CoVs based on the scope of evolutionary relatedness. This should also influence the probability of true coupling because not all residues are equally conserved or mutable. We propose future efforts should examine such covariance in the context of residue mutability. Recent efforts that applied DCA to predict epistatic interactions in the context of SARS-CoV-2 residue mutability concluded coupling also played a minor role in emerging mutations in variants. The authors surmised that a restricted number of unique genomes and a broad scope of evolutionary divergence among all coronaviruses also limited the analysis performance for epistatis-mutation comparisons in SARS-CoV-2 proteins [63]. We chose to limit our pan-analysis to only lineage B β-coronaviruses for our work based on this very principle.

We find evidence of true covariance in this work when we compare the total number of independent changes present in each pair with a known structure. In Spike, the probability of either a direct or possible indirect interaction by one flanking residue increased from 5% to 24% when the minimum number of unique AA changes in any given residue pair of two identities were compared to those with five. Furthermore, for covariant residues with increased independent residue representation in the pan covariance analysis, these were also more likely to be identified in the clinical covariance analysis. Notably, when residues identified in both analyses are compared, 90% are found within the same gene and enriched in certain genes with important virus-host interactions such as Spike. This observation is consistent with an expected model where intraprotein covariance is predicted to be more abundant than that for interprotein, including proteins that form homotrimers such as Spike.

A benefit of comparing residues from both covariant analyses is demonstrated by their co-presence and enrichment in dominant circulating variants. Mutations and deletion-enriched hotspots in the Spike NTD described in this work have been recently identified and studied as antigenic regions responsible for antibody escape [64,51,65–68], but not yet investigated comprehensively in the context of pan-covariance to our knowledge. Notably, many NTD covariant hotspots are also deletions identified in SARS-CoV Spike protein when aligned to SARS-CoV-2 [51]. Though clinical SARS-CoV-2 covariance data should by definition reveal co-present residues common to variants, the independence occurrence of these in the Pan-CoV covariance is intriguing. We note these clinical sequences were collected through August 2021, over three months prior to the first emergence of Omicron, and yet we identify some Omicron-specific residues and pairs in this work. We propose all covariant residues identified in Spike and other conserved CoV proteins in this work serve as one reference for possible future single and multiple mutations that might arise in dominant variants. These could inform about epitopes and antigenic regions that are possibly vulnerable to enriched mutations also in part due to covariance such as specific regions in the Spike NTD. Furthermore, these may reveal key residues that contribute to yet undiscovered interactions between viral proteins of SARS-CoV-2 including Spike and ORF3a as discussed in an earlier description of our initial results [55].

Methods

Genome and protein sequence acquisition and alignment

Nucleotide and protein sequences for the 169 individual CoVs and 252,102 clinical samples were downloaded from available NCBI and GISAID public databases [32,69,70]. All genomes and accession numbers are provided in **Supplemental Table S1**. The GISAID sequences have been provided from various sources and published work and these source data are acknowledged in **Supplemental Table S4**. Nucleotide sequences of 169 CoVs were aligned using the MAFFT iterative consistency-based setting (G-INS-i) and we used nucleotides that spanned the aligned start codon of SARS-CoV-2 *Orf1a* gene through the stop codon of the *N* gene [71]. Protein sequences for NSP1 through NSP16, Spike, ORF3a, E, M, and N of 149 CoVs were concatenated and then aligned using the MAFFT iterative consistency-based setting (G-INS-i)[71]. For clinical samples, we initially selected a total of 882,364 sequences isolated, sequenced, and deposited in GISAID between December 2019 and August 21, 2021 based on their near completeness (>95%) of sequence coverage for all proteins used in pan-CoV amino acid covariance. This facilitated the inclusion and identification of sequences with small deletions and rare insertions. Due to computational limits, two independent sets of 126,051 randomly selected sequences of the 882,364 were aligned using MAFFT using respective references of each alignment to maintain identical length (List provided in **Supplemental Table S4)** [71]. We expect some deletions and insertions are due to sequencing and assembly errors but only co-varying deletions should become apparent during covariance analysis. The clinical alignment spans both known and predicted genes between nsp1 and Orf9c, but only covariance between the proteins also studied in pan covariance set are analyzed and compared in this work.

Phylogeny and ATDS calculation

We inferred phylogeny by reconstructing a maximum likelihood (ML) tree with IQTree after first testing and comparing 286 DNA models by creating initial parsimony trees scored according to Bayesian information criterion (BIC) using IQTree ModelFinder [72]. We then applied the best fit DNA model which is a general time reversible model using empirical base frequencies allowing for the FreeRate heterogeneity model across sites GTR+F+R6 (invariable site plus discrete Gamma model) with 1000 replicates using bootstrap resampling analysis [72]. This tree file is available in **Supplemental File**. Bootstrap resampling analysis was completed using 1000 replicates. Bootstrap values and branch lengths are indicated in an unrooted tree shown as a circular phylogram. Branch lengths shorter than 0.0368 are shown as having length 0.0368.

For all genomes that belong to each cluster, the sum of branch lengths between every possible pair was extracted from the tree file and averaged to calculate the Average Taxonomic Distribution Score (ATDS). This relative score is provided as additional metadata in the Gephi Force mapping file.

Covariance analysis and force mapping

Pairwise and multiple residue covariance and scores were calculated using FastCov [56]. Alignment files for both pan and clinical CoVs were substituted to provide a “W’ in place of absent/deleted residues.

Using the known position of true "W", residues, all "W" deletions were replaced as "Z" to indicate absence following analysis. We set a purity score (0.7) for stringency cutoffs in both the pan-CoV and clinical-CoV sequence alignment. A raw table of predicted covariant pairs is provided as **Supplemental Table S2.** This allowed the binning of clusters and respective strains for Force Mapping in Gephi using the Multigravity ForceAtlas 2 setting and comparison of covariant residues based on clusters and strains [73]. All clusters and residues and their respective occurrence in CoVs for both analyses are tabulated in **Supplemental Table S3**. Genomes, clusters, residues were mapped in Gephi using the MultiGravity ForceAtlas 2 algorithm. [73]. This data is provided in **Supplemental File S6** for interactive application using Gephi Software.

Prediction of interacting residues and mapping of residues in Spike trimer structure

The Arpeggio program was used to calculate inter and intramolecular interactions between residues in the 7JJI.PDB file (**Supplemental Table S5)** [59,74]. To accommodate minor sequence and structure variability between Spike proteins in the pan-CoV analysis, the position of any two residues identified to interact in SARS-CoV-2 was extended by one flanking position both amino and carboxyl to each residue when calculating possible interactions for all 149 CoVs. Residues in Spike were mapped onto the PDB structure for Spike (7JJI.pdb) using PyMol (v.2.3.4) [58,75,76].

Cross-referencing residues present in dominant variants

Mutations identified in previous and dominant circulating variants of clinical interest were extracted from data compiled by CoVariants.org and enabled by GISAID [31,32,77]. The WHO label for each variant is used for reference.

Statistics

Hypergeometric probability was applied in R using the abundance and distribution of single residues in each analyzed gene in the pan and clinical covariance datasets. Residue identity by position was approximated for the pan covariance and then numbered by position in SARS-CoV-2. For comparative analyses of covariant pairs identified in both analyses across all genes, residue identity by position was approximated for the pan covariance and then numbered by position in SARS-CoV-2. The total number of unique residues identified as covariant in each independent analyses and the total pairs co-present (overlap) was examined as above by applying a hypergeometric probability formula.

Plotting

Circular graphing of key collections of residues was graphically plotted using Circos [78].

Declarations

Ethics approval and consent to participate

This study includes sequence and metadata of 252,102 CoV virus strains from a publically available database (GISAID) and though patient age and sex has been approved to be publically available in this database, only the locations and date of virus isolation are noted in this work. No IRB approval is needed for this data or and the all acknowledged sources and authors for every sequence in this source data tabulated from the public GISAID are provided in **Supplemental Table S4**.

Availability of data and materials

All data was generated using publically deposited and available genome and protein sequences and the identity and accession numbers are provided. All generated and analyzed raw output data for which this study is based is provided in this published article and can be referenced in the tabulated supplemental spreadsheet files.

Competing interests

We declare there are no financial or non-financial competing interests with the published work.

Funding

This work was supported by National Institutes of Health Grant AI-018045-JJM

Authors' contributions

WPR completed all data analysis and contributed to the biological interpretation of data, discussion, and conclusions. JJM contributed to the biological interpretation of data, discussion, and conclusions.

Acknowledgements

We are especially appreciative and gratefully acknowledge the authors and originating laboratories and hospitals responsible for obtaining the virus specimens and the laboratories where genetic sequence data were generated and shared via the GISAID Initiative (sources provided in **Supplemental File S4**).

Abbreviations

AA: Amino acid

ATDS: Average Taxonomic Distribution Score.

CoV: Coronavirus

CTD: Carboxy-terminal domain

FCS: Furin Cleavage Site

FP: Fusion Peptide

GISAID : Global Initiative on Sharing All Influenza Data

NCBI: National Center for Biotechnology Information

nsp: Nonstructural protein

NTD: Amino-terminal domain

PCA: Principal Component Analysis

RBD: Receptor binding domain

WHO: World Health Organization

References

1. Al-Omari, A., Rabaan, A. A., Salih, S., Al-Tawfiq, J. A. & Memish, Z. A. MERS coronavirus outbreak: Implications for emerging viral infections. *Diagnostic Microbiology and Infectious Disease* **93**, 265–285 (2019).

2. Corman, V. M. *et al.* Rooting the Phylogenetic Tree of Middle East Respiratory Syndrome Coronavirus by Characterization of a Conspecific Virus from an African Bat. *J. Virol.* **88**, 11297 (2014).

3. Hu, B. *et al.* Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. *PLoS Pathog* **13**, e1006698 (2017).

4. Li, W. Bats Are Natural Reservoirs of SARS-Like Coronaviruses. *Science* **310**, 676–679 (2005).

5. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).

6. Zhang, T., Wu, Q. & Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Current Biology* **30**, 1346-1351.e2 (2020).

7. Patiño-Galindo, J. Á., González-Candelas, F. & Pybus, O. G. The Effect of RNA Substitution Models on Viroid and RNA Virus Phylogenies. *Genome Biology and Evolution* **10**, 657–666 (2018).

8. Smyth, R. P., Negroni, M., Lever, A. M., Mak, J. & Kenyon, J. C. RNA Structure—A Neglected Puppet Master for the Evolution of Virus and Host Immunity. *Frontiers in Immunology* **9**, 2097 (2018).

9. Akashi, H. & Eyre-Walker, A. Translational selection and molecular evolution. *Current Opinion in Genetics & Development* **8**, 688–693 (1998).

10. Tort, F. L., Castells, M. & Cristina, J. A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses. *Virus Res* **283**, 197976–197976 (2020).

11. Hou, W. Characterization of codon usage pattern in SARS-CoV-2. *Virology Journal* **17**, 138 (2020).

12. Kandeel, M., Ibrahim, A., Fayez, M. & Al-Nazawi, M. From SARS and MERS CoVs to SARS-CoV-2: Moving toward more biased codon usage in viral structural and nonstructural genes. *Journal of Medical Virology* **92**, 660–666 (2020).

13. Roy, A. *et al.* Base Composition and Host Adaptation of the SARS-CoV-2: Insight From the Codon Usage Perspective. *Frontiers in Microbiology* **12**, 747 (2021).

14. Segreto, R. & Deigin, Y. The genetic structure of SARS-CoV-2 does not rule out a laboratory origin: SARS-COV-2 chimeric structure and furin cleavage site might be the result of genetic manipulation. *Bioessays* **43**, e2000240–e2000240 (2021).

15. Jenkins, G. M. & Holmes, E. C. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Research* **92**, 1–7 (2003).

16. Belalov, I. S. & Lukashev, A. N. Causes and Implications of Codon Usage Bias in RNA Viruses. *PLOS ONE* **8**, e56642 (2013).

17. Greenbaum, B. D., Levine, A. J., Bhanot, G. & Rabadan, R. Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses. *PLOS Pathogens* **4**, e1000079 (2008).

18. Greenbaum, B. D., Rabadan, R. & Levine, A. J. Patterns of Oligonucleotide Sequences in Viral and Host Cell RNA Identify Mediators of the Host Innate Immune System. *PLOS ONE* **4**, e5969 (2009).

19. Takata, M. A. *et al.* CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature* **550**, 124–127 (2017).

20. Matyášek, R. & Kovařík, A. Mutation Patterns of Human SARS-CoV-2 and Bat RaTG13 Coronavirus Genomes Are Strongly Biased Towards C>U Transitions, Indicating Rapid Evolution in Their Hosts. *Genes* **11**, (2020).

21. Song, H.-D. *et al.* Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *Proceedings of the National Academy of Sciences* **102**, 2430–2435 (2005).

22. Roberts, A. *et al.* A Mouse-Adapted SARS-Coronavirus Causes Disease and Mortality in BALB/c Mice. *PLOS Pathogens* **3**, e5 (2007).

23. Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* (2020).

24. Yan, H. *et al.* ACE2 receptor usage reveals variation in susceptibility to SARS-CoV and SARS-CoV-2 infection among bat species. *Nature Ecology & Evolution* **5**, 600–608 (2021).

25. Coutard, B. *et al.* The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Research* **176**, 104742 (2020).

26. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).

27. Li, P. *et al.* The Rhinolophus affinis bat ACE2 and multiple animal orthologs are functional receptors for bat coronavirus RaTG13 and SARS-CoV-2. *Science Bulletin* **66**, 1215–1227 (2021).

28. Lau, S. K. P. *et al.* Differential Tropism of SARS-CoV and SARS-CoV-2 in Bat Cells. *Emerg Infect Dis* **26**, 2961–2965 (2020).

29. Zhang, L. *et al.* SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun* **11**, 6013–6013 (2020).

30. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* (2020) doi:10.1016/j.cell.2020.06.043.

31. Hodcroft, E. *CoVariants: SARS-CoV-2 Mutations and Variants of Interest.* https://covariants.org (2021).

32. https://www.gisaid.org. *www.gisaid.org* https://www.gisaid.org (2020).

33. Gobeil Sophie M.-C. *et al.* Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. *Science* **373**, eabi6226.

34. Felsenstein, J. Phylogenies and the Comparative Method. *The American Naturalist* **125**, 1–15 (1985).

35. Rivas, E. Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinformatics* **6**, 63–63 (2005).

36. Dunn, S. D., Wahl, L. M. & Gloor, G. B. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340 (2007).

37. Talavera, D., Lovell, S. C. & Whelan, S. Covariation Is a Poor Measure of Molecular Coevolution. *Molecular Biology and Evolution* **32**, 2456–2468 (2015).

38. Qin, C. & Colwell, L. J. Power law tails in phylogenetic systems. *Proc Natl Acad Sci USA* **115**, 690 (2018).

39. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An Evolutionary Trace Method Defines Binding Surfaces Common to Protein Families. *Journal of Molecular Biology* **257**, 342–358 (1996).

40. Pancer, K. *et al.* The SARS-CoV-2 ORF10 is not essential in vitro or in vivo in humans. *PLOS Pathogens* **16**, e1008959 (2020).

41. Gorbalenya, A. E. *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology* **5**, 536–544 (2020).

42. Liu, P. *et al.* Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? *PLOS Pathogens* **16**, e1008421 (2020).

43. Li, X. *et al.* Emergence of SARS-CoV-2 through recombination and strong purifying selection. *Sci Adv* eabb9153 (2020).

44. Delaune, D. *et al.* A novel SARS-CoV-2 related coronavirus in bats from Cambodia. *Nat Commun* **12**, 6563–6563 (2021).

45. Sarah Temmam *et al.* Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula. *Nature Portfolio* (2021).

46. Graham, R. L. & Baric, R. S. Recombination, Reservoirs, and the Modular Spike: Mechanisms of Coronavirus Cross-Species Transmission. *J. Virol.* **84**, 3134 (2010).

47. Simon-Loriere, E. & Holmes, E. C. Why do RNA viruses recombine? *Nature Reviews Microbiology* **9**, 617–626 (2011).

48. Paraskevis, D. *et al.* Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infection, Genetics and Evolution* **79**, 104212 (2020).

49. Pollett, S. *et al.* A comparative recombination analysis of human coronaviruses and implications for the SARS-CoV-2 pandemic. *Scientific Reports* **11**, 17365 (2021).

50. Wu, Y. & Zhao, S. Furin cleavage sites naturally occur in coronaviruses. *Stem Cell Research* **50**, 102115 (2021).

51. McCarthy Kevin R. *et al.* Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape. *Science* **371**, 1139–1142 (2021).

52. Fitch, W. M. & Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics* **4**, 579–593 (1970).

53. Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics* **18**, 309–317 (1994).

54. Taylor, W. R. & Hatrick, K. Compensating changes in protein multiple sequence alignments. *Protein Engineering, Design and Selection* **7**, 341–348 (1994).

55. Robins, W. P. & Mekalanos, J. J. Protein covariance networks reveal interactions important to the emergence of SARS coronaviruses as human pathogens. *bioRxiv* 2020.06.05.136887 (2020).

56. Shen, W. & Li, Y. A novel algorithm for detecting multiple covariance and clustering of biological sequences. *Scientific Reports* **6**, 30425 (2016).

57. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci USA* **108**, E1293 (2011).

58. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).

59. Bangaru, S. *et al.* Structural analysis of full-length SARS-CoV-2 spike protein from an advanced vaccine candidate. *Science* **370**, 1089–1094 (2020).

60. Huo, J. *et al.* Neutralization of SARS-CoV-2 by Destruction of the Prefusion Spike. *Cell Host Microbe* **28**, 445-454.e6 (2020).

61. Huo, J. *et al.* Neutralizing nanobodies bind SARS-CoV-2 spike RBD and block interaction with ACE2. *Nature Structural & Molecular Biology* **27**, 846–854 (2020).

62. Lu, G., Wang, Q. & Gao, G. F. Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. *Trends in Microbiology* **23**, 468–478 (2015).

63. Rodriguez-Rivas, J., Croce, G., Muscat, M. & Weigt, M. Epistatic models predict mutable sites in SARS-CoV-2 proteins and epitopes. *Proc Natl Acad Sci USA* **119**, e2113118119 (2022).

64. Choi, B. *et al.* Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N Engl J Med* **383**, 2291–2293 (2020).

65. McCallum, M. *et al.* N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2. *Cell* **184**, 2332-2347.e16 (2021).

66. Ribes, M., Chaccour, C. & Moncunill, G. Adapt or perish: SARS-CoV-2 antibody escape variants defined by deletions in the Spike N-terminal Domain. *Signal Transduction and Targeted Therapy* **6**, 164 (2021).

67. Venkatakrishnan, A. J. *et al.* Antigenic minimalism of SARS-CoV-2 is linked to surges in COVID-19 community transmission and vaccine breakthrough infections. *medRxiv* 2021.05.23.21257668 (2021).

68. Meng, B. *et al.* Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7. *Cell Reports* **35**, (2021).

69. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33–46 (2017).

70. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **46**, D8–D13 (2018).

71. Nakamura, T., Yamada, K. D., Tomii, K. & Katoh, K. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics* **34**, 2490–2492 (2018).

72. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* **32**, 268–274 (2015).

73. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE* **9**, e98679 (2014).

74. Jubb, H. C. *et al.* Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology* **429**, 365–371 (2017).

75. *The PyMOL Molecular Graphics System*. (Schrödinger, LLC).

76. Song, W., Gui, M., Wang, X. & Xiang, Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. *PLOS Pathogens* **14**, e1007236 (2018).

77. Khare, S. *et al.* GISAID's Role in Pandemic Response. *China CDC Wkly* **3**, 1049–1051 (2021).

78. Krzywinski, M. I. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Research* (2009).

Main Figure Legend

Figure 1. Force mapping graph cluster based on amino acid residue covariance and nucleotide alignment-based ML tree phylogeny both reveal similar relationships. (A) Gephi force mapping graph (Multiforce ForceAtlas 2) showing 149 CoVs based on all predicted clusters of covariant residues. The respective host for each CoV is indicated by color and the average taxonomic distribution score (ATDS) for each cluster is indicated by cluster circle size. All CoV, clusters, and residues can be accessed and extracted from the interactive Gephi file and supplemental tables. CoVs most closely related to SARS-CoV, SARS-CoV-2, and HKU3 based on phylogeny are circles and labeled. (B) Overview of ML tree of 169 CoVs that spans nts of genes *ORF1a/b* through *N* (Nucleocapsid). CoVs are colored by host. SARS-CoV, SARS-CoV-2, and HKU3-related CoVs are circled to match groups in Figure 1A.

Figure 2. The distribution of covariant residues in CoVs most closely related to SARS-CoV-2. (A) the identification and CoVs in Groups 1-5 in a subset of the maximum likelihood tree showing branch length and bootstrap values. (B) Position of tree subset in entire maximum likelihood tree (from Figure 1B). (C) Distribution of covariant residues in core genes for Groups 1-5 based on clusters specific to each group. Each group is indicated by color and the position of every residue is colored.

Figure 3. The distribution of Spike covariant residues found in CoVs Groups 1-3. (A) the number of unique and shared residues between Groups 1, 2, and 3. (B) Covariant residue-enriched of NTD, RBD, and FCS that are shown in detail in (C). (C) Aligned AA sequences of covariant residue-enriched regions of Spike with representatives from Group 1, Group 2, and Group 3. Covariant sequences that overlap between these are boxed and those common to two groups colored as yellow and those to all three purple. Residues identified in the clinical covariant analysis are indicated (+). Residues present in dominant circulating variants are indicated (*). Residues deleted in these regions when these groups are compared are shown in red.

Figure 4. Comparing individual covariant residues in clinical and pan analyses in conserved genes. (A and B) Outline of comparative analysis. (C) Trace showing the position and frequency of both Clinical (orange) and Pan (blue) identified covariant residues in each gene. The length of the SARS-CoV-2 gene is labeled and the number of covariant residues in each is indicated with the number overlapping in a Venn diagram. The size of each Venn circle is proportional to the percentage of residues identified in each gene., The significance and P-value calculated for each gene based on length of the gene, number of covariant residues in each, and overlap is shown.

Figure 5. Mapping covariant pairs common to both clinical and pan analyses. (A) Diagram showing the concept of conserved residue pair. (B) Distribution of intra-gene covariant pairs. Links are colored by the gene. (C) Distribution of inter-gene covariant pairs. (D) Distribution of Spike and ORF3a inter-and intra-gene covariant pairs. Domains and boundaries of ORF3a (NTD) and Spike (NTD, RBD, and FCS) are shown. (E) Distribution of Spike and ORF3a covariant pairs from (D) that are also present in dominant circulating variants. (F) Network graph showing the number of covariant residues represented in each gene (size of circle) and the occurrence with the number of inter-gene covariant residues between each gene (value indicated for each linkage).

Figure 6. Mapping of Spike NTD covariant residues and deletion identified in clinical and pan analyses and also found in dominant circulating SARS-CoV-2 variants. (A-C) The top, side, and bottom PDB structure (7JJI.PDB) of the Spike homotrimer shows the position of these residues. Residues are colored by position in the NTD. (D) Labeled regions and residue numbers of regions are indicated. Amino acids structures are shown for highlighted residues. (E) The sequences of highlighted residues are shown. Residues predicted to directly interact at the molecular level in the PBD are linked by dotted lines.

Covariant residues found in both Pan and Clinical covariant analyses and dominant SARS-CoV-2 variants are colored red.

Figure 7. Flow chart showing the accounting of identified pan and clinical covariant residues and overlap. 14 dominant variants used in comparative analyses and respective residues found in each are shown in parenthesis. WHO label for each variant is used for reference. The gene by gene distribution of the 538 co-present pairs in genes is graphed. The abundance and overlap of pairs and single residues identified in the 14 dominant lineages are graphed. Hypergeometric probability of common residues is shown and representation value indicates the overlap value divided by the number of expected pairs to overlap if all covariant pairs were equally probable.

Figure 1

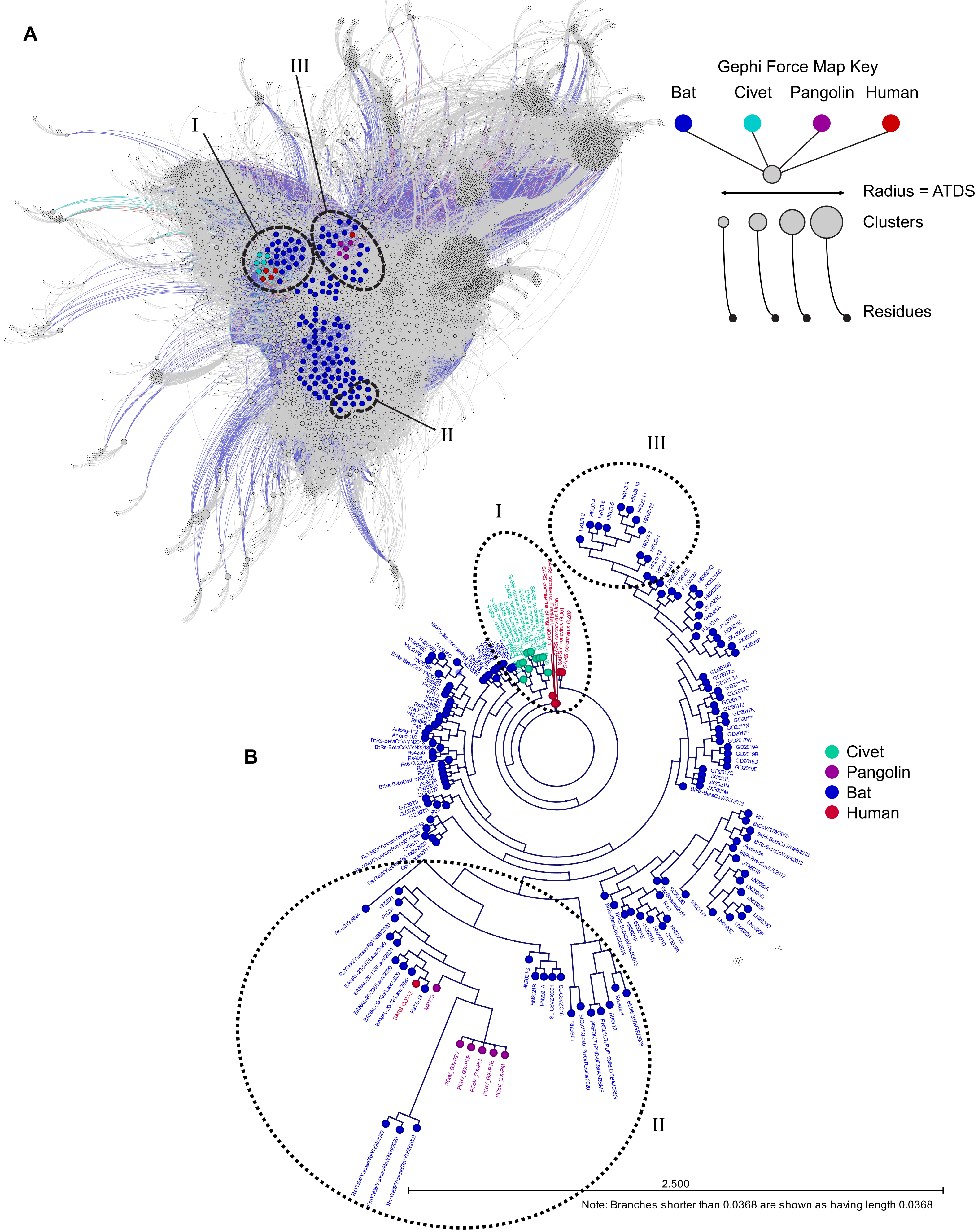

A
I
II
III
Gephi Force Map Key
Bat
Civet
Pangolin
Human
Radius = ATDS
Clusters
Residues
B
I
II
III
Civet
Pangolin
Bat
Human
HKU3-2
HKU3-4
HKU3-6
HKU3-5
HKU3-9
HKU3-10
HKU3-11
HKU3-13
HKU3-3
HKU3-1
HKU3-12
HKU3-7
HKU3-8
SARS coronavirus GZ02
SARS-CoV-2
RaTG13
MP789
PCoV_GX-P2V
PCoV_GX-P3B
PCoV_GX-P5L
PCoV_GX-P1E
PCoV_GX-P4L
RsYN04/Yunnan/RsYN04/2020
RmYN08/Yunnan/RmYN08/2020
RmYN05/Yunnan/RmYN05/2020
BANAL-20-247/Laos/2020
BANAL-20-116/Laos/2020
BANAL-20-236/Laos/2020
BANAL-20-103/Laos/2020
BANAL-20-52/Laos/2020
RpYN06/Yunnan/RpYN06/2020
Rc-o319 RNA
YN2021
PrC31
HN2021G
HN2021B
HN2021A
SL-CoVZXC21
SL-CoVZC45
RhGB01
BtCoV/Khosta-2/Rh/Russia/2020
PREDICT/PRD-0038/ABSMF
PREDICT/PDF-2386/OTBA40RSV
BtKY72
Khosta-1
BM48-31/BGR/2008
Rf1
BtCoV/273/2005
BtRf-BetaCoV/HeB2013
BtRf-BetaCoV/SX2013
Jiyuan-84
BtRf-BetaCoV/JL2012
JTMC15
LN2020A
LN2020G
LN2020B
LN2020C
LN2020F
LN2020H
LN2020E
16BO133
Rm1
SC2018B
BtRs-BetaCoV/GX2013
GD2016B
GD2017G
GD2017M
GD2017H
GD2017O
GD2019A
GD2019B
GD2019D
GD2019E
JX2021L
JX2021N
JX2021M
HB2020D
JX2021AC
HB2020E
JX2021G
JX2021K
JX2021J
JX2021O
JX2021P
Rs3367
WIV1
RsSHC014
Rs4084
YNLF_34C
YNLF_31C
Rf4092
F46
Anlong-112
Anlong-103
BtRs-BetaCoV/YN2013
BtRs-BetaCoV/YN2018
Rs4255
Rs4081
Rs672/2006
Rs4247
Rs4237
BtRs-BetaCoV/YN2018C
As6526
YN2020A
GD2017F
GZ2021I
GZ2021H
GZ2021C
Rp3
RsYN03/Yunnan/RsYN03/2019
RmYN07/Yunnan/RmYN07/2020
LYRa11
YN09/2020
Cp/Yunnan2011
RsYN09/Yunnan
2.500
Note: Branches shorter than 0.0368 are shown as having length 0.0368

Figure 2

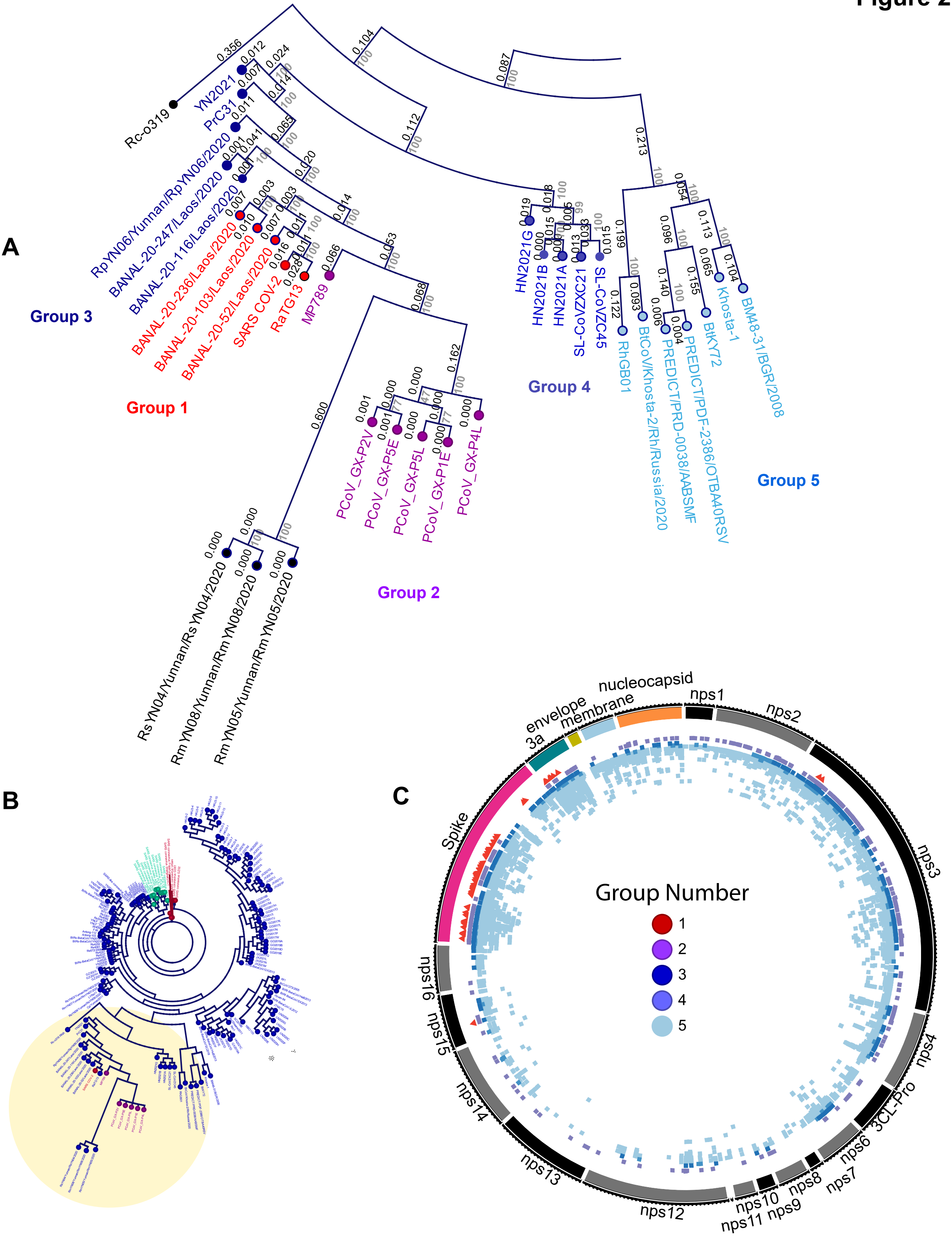

A
Rc-o319
YN2021
PrC31
RpYN06/Yunnan/RpYN06/2020
BANAL-20-247/Laos/2020
BANAL-20-116/Laos/2020
BANAL-20-236/Laos/2020
BANAL-20-103/Laos/2020
BANAL-20-52/Laos/2020
SARS COV-2
RaTG13
MP789
PCoV_GX-P2V
PCoV_GX-P5E
PCoV_GX-P5L
PCoV_GX-P1E
PCoV_GX-P4L
RsYN04/Yunnan/RsYN04/2020
RmYN08/Yunnan/RmYN08/2020
RmYN05/Yunnan/RmYN05/2020
HN2021G
HN2021B
HN2021A
SL-CoVZXC21
SL-CoVZC45
RhGB01
BtCoV/Khosta-2/Rh/Russia/2020
PREDICT/PRD-0038/AABSMF
PREDICT/PDF-2386/OTBA40RSV
BtKY72
Khosta-1
BM48-31/BGR/2008
Group 1
Group 2
Group 3
Group 4
Group 5
B
C
nucleocapsid
nps1
nps2
nps3
nps4
3CL-Pro
nps6
nps7
nps8
nps9
nps10
nps11
nps12
nps13
nps14
nps15
nps16
Spike
3a
envelope
membrane
Group Number
1
2
3
4
5

Figure 3

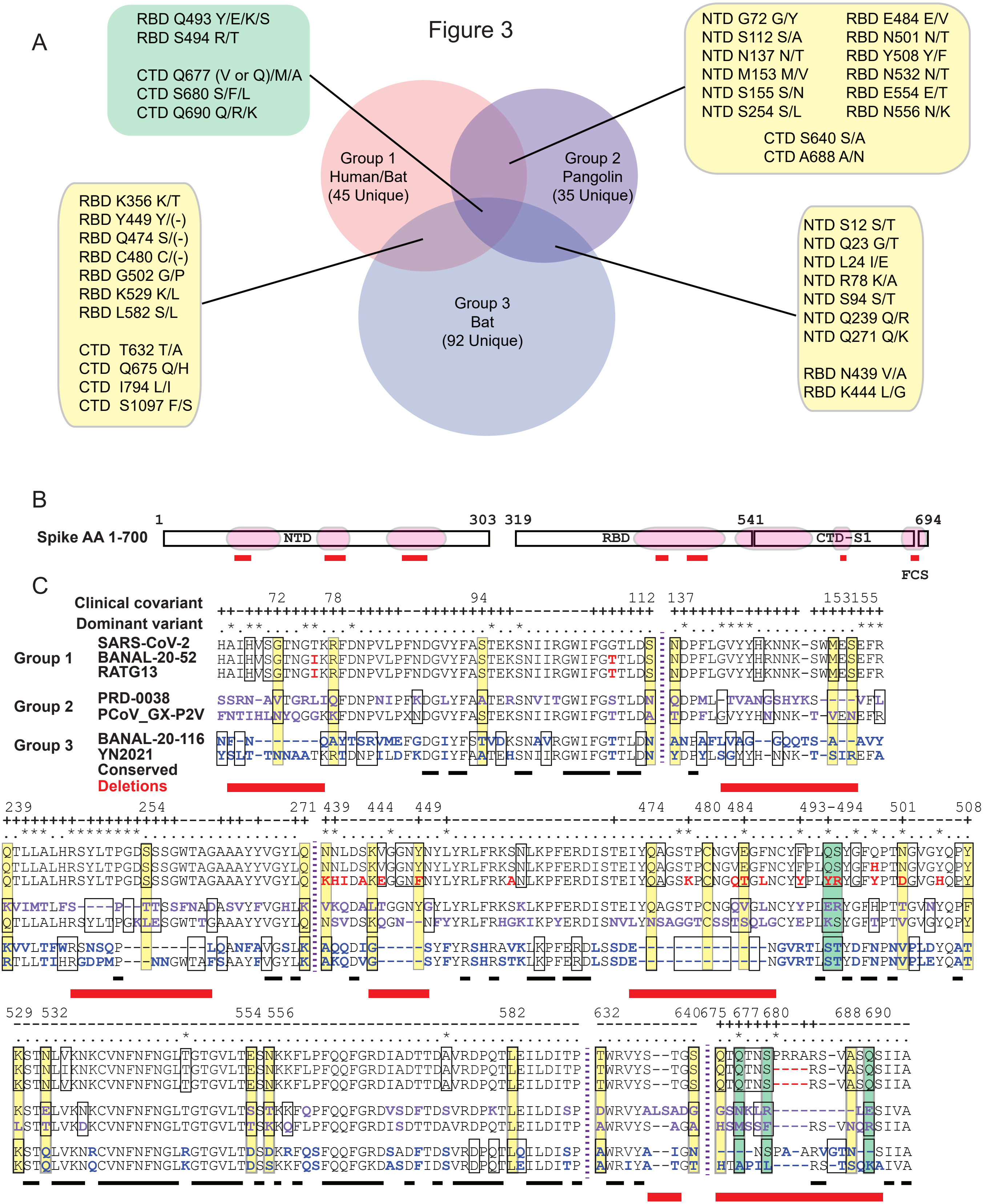

A
RBD Q493 Y/E/K/S
RBD S494 R/T
CTD Q677 (V or Q)/M/A
CTD S680 S/F/L
CTD Q690 Q/R/K
NTD G72 G/Y
NTD S112 S/A
NTD N137 N/T
NTD M153 M/V
NTD S155 S/N
NTD S254 S/L
RBD E484 E/V
RBD N501 N/T
RBD Y508 Y/F
RBD N532 N/T
RBD E554 E/T
RBD N556 N/K
CTD S640 S/A
CTD A688 A/N
Group 1
Human/Bat
(45 Unique)
Group 2
Pangolin
(35 Unique)
Group 3
Bat
(92 Unique)
RBD K356 K/T
RBD Y449 Y/(-)
RBD Q474 S/(-)
RBD C480 C/(-)
RBD G502 G/P
RBD K529 K/L
RBD L582 S/L
CTD T632 T/A
CTD Q675 Q/H
CTD I794 L/I
CTD S1097 F/S
NTD S12 S/T
NTD Q23 G/T
NTD L24 I/E
NTD R78 K/A
NTD S94 S/T
NTD Q239 Q/R
NTD Q271 Q/K
RBD N439 V/A
RBD K444 L/G
B
Spike AA 1-700
1
303
319
541
694
NTD
RBD
CTD-S1
FCS
C
Clinical covariant
Dominant variant
Group 1
SARS-CoV-2
BANAL-20-52
RATG13
Group 2
PRD-0038
PCoV_GX-P2V
Group 3
BANAL-20-116
YN2021
Conserved
Deletions
72 78 94 112 137 153 155
HAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDS NDPFLGVYYHKNNK-SWMESEFR
HAIHVSGTNGIKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDS NDPFLGVYYHRNNK-SWMESEFR
HAIHVSGTNGIKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDS NDPFLGVYYHKNNK-SWMESEFR
SSRN-AVTGRLIQFDNPNIPFKDGLYFAATERSNVITGWIFGSTLDN QDPML-TVANGSHYKS----VFL
FNTIHLNYQGGKKFDNPVLPXNDGVYFASTEKSNIIRGWIFGTTLDA TDPFLGVYYHNNNK-T-VENEFR
NF-N-------QAYTSRVMEFGDGIYFSTVDKSNAVRGWIFGTTLDN ANPAFLVAG--GQQTS-A--AVY
YSLT-TNNAATKRTDNPILDFKDGIYFAATEHSNIIRGWIFGTTLDN YDPYLSGYYH-NNK-T-SIREFA
239 254 271 439 444 449 474 480 484 493-494 501 508
QTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQ NNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPY
QTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQ NNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFHPTNGVGYQPY
QTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQ KHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQAGSKPCNGQTGLNCYYPLYRYGFYPTDGVGHQPY
KVIMTLFS----P--TTSSFNADASVYFVGHLK VKQDALTGGNYGYLYRLFRKSKLKPFERDISTEIYQAGSTPCNGQVGLNCYYPLERYGFHPTTGVNYQPF
QTLLALHRSYLTPGKLESGWTTGAAAYYVGYLQ NSVDSKQGN--NFYYRLFRHGKIKPYERDISNVLYNSAGGTCSSTSQLGCYEPLKSYGFTPTVGVGYQPY
KVVLTFWRSNSQP-------LQANFAVGSLK AQQDIG-----SYFYRSHRAVKLKPFERDLSSDE-----------NGVRTLSTYDFNPNVPLDYQAT
RTLLTIHRGDPMP---NNGWTAFSAAYFVGYLK AKQDVG-----SYFYRSHRSTKLKPFERDLSSDE-----------NGVRTLSTYDFNPNVPLEYQAT
529 532 554 556 582 632 640 675 677 680 688 690
KSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITP TWRVYS--TGS QTQTNSPRRARS-VASQSIIA
KSTNLIKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITP TWRVYS--TGS QTQTNS----RS-VASQSIIA
KSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITP TWRVYS--TGS QTQTNS----RS-VASQSIIA
KSTELVKNKCVNFNFNGLTGTGVLTSSTKKFQPFQQFGRDVSDFTDSVRDPKTLEILDISP DWRVYALSADG GSNKLR---------LESIVA
LSTTLVKDKCVNFNFNGLTGTGVLTTSKKQFLPFQQFGRDISDTTDAVRDPQTLEILDITP AWRVYS--AGA HSMSSF----RS-VNQRSIIA
KSTQLVKNRCVNFNFNGLRGTGVLTDSDKRFQSFQQFGRDSADFTDSVRDPQTLQILDISP AWRVYA--IGN ----NSPA-ARVGTN--SIIA
KSTQLVKNQCVNFNFNGLKGTGVLTDSSKKFQSFQQFGKDASDFIDSVRDPQTLEILDITP AWRIYA--TGT HTAPIL----RS-TSQKAIVA

Figure 4

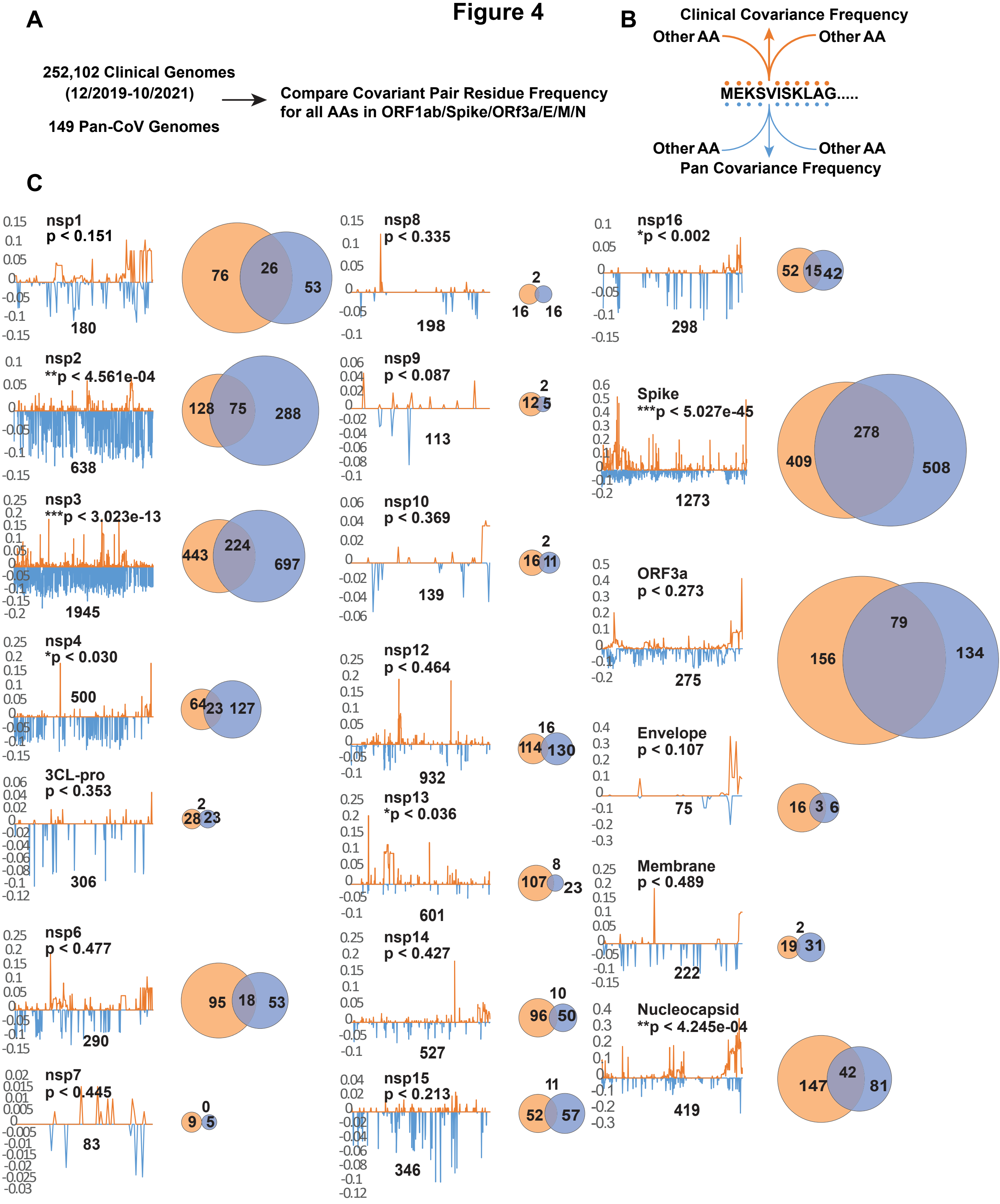

A
252,102 Clinical Genomes
(12/2019-10/2021)
149 Pan-CoV Genomes
Compare Covariant Pair Residue Frequency
for all AAs in ORF1ab/Spike/ORf3a/E/M/N
B
Clinical Covariance Frequency
Other AA
Other AA
MEKSVISKLAG.....
Other AA
Other AA
Pan Covariance Frequency
C
nsp1
p < 0.151
180
76
26
53
nsp2
**p < 4.561e-04
638
128
75
288
nsp3
***p < 3.023e-13
1945
443
224
697
nsp4
*p < 0.030
500
64
23
127
3CL-pro
p < 0.353
306
28
2
23
nsp6
p < 0.477
290
95
18
53
nsp7
p < 0.445
83
9
0
5
nsp8
p < 0.335
198
16
2
16
nsp9
p < 0.087
113
12
2
5
nsp10
p < 0.369
139
16
2
11
nsp12
p < 0.464
932
114
16
130
nsp13
*p < 0.036
601
107
8
23
nsp14
p < 0.427
527
96
10
50
nsp15
p < 0.213
346
52
11
57
nsp16
*p < 0.002
298
52
15
42
Spike
***p < 5.027e-45
1273
409
278
508
ORF3a
p < 0.273
275
156
79
134
Envelope
p < 0.107
75
16
3
6
Membrane
p < 0.489
222
19
2
31
Nucleocapsid
**p < 4.245e-04
419
147
42
81

Figure 5

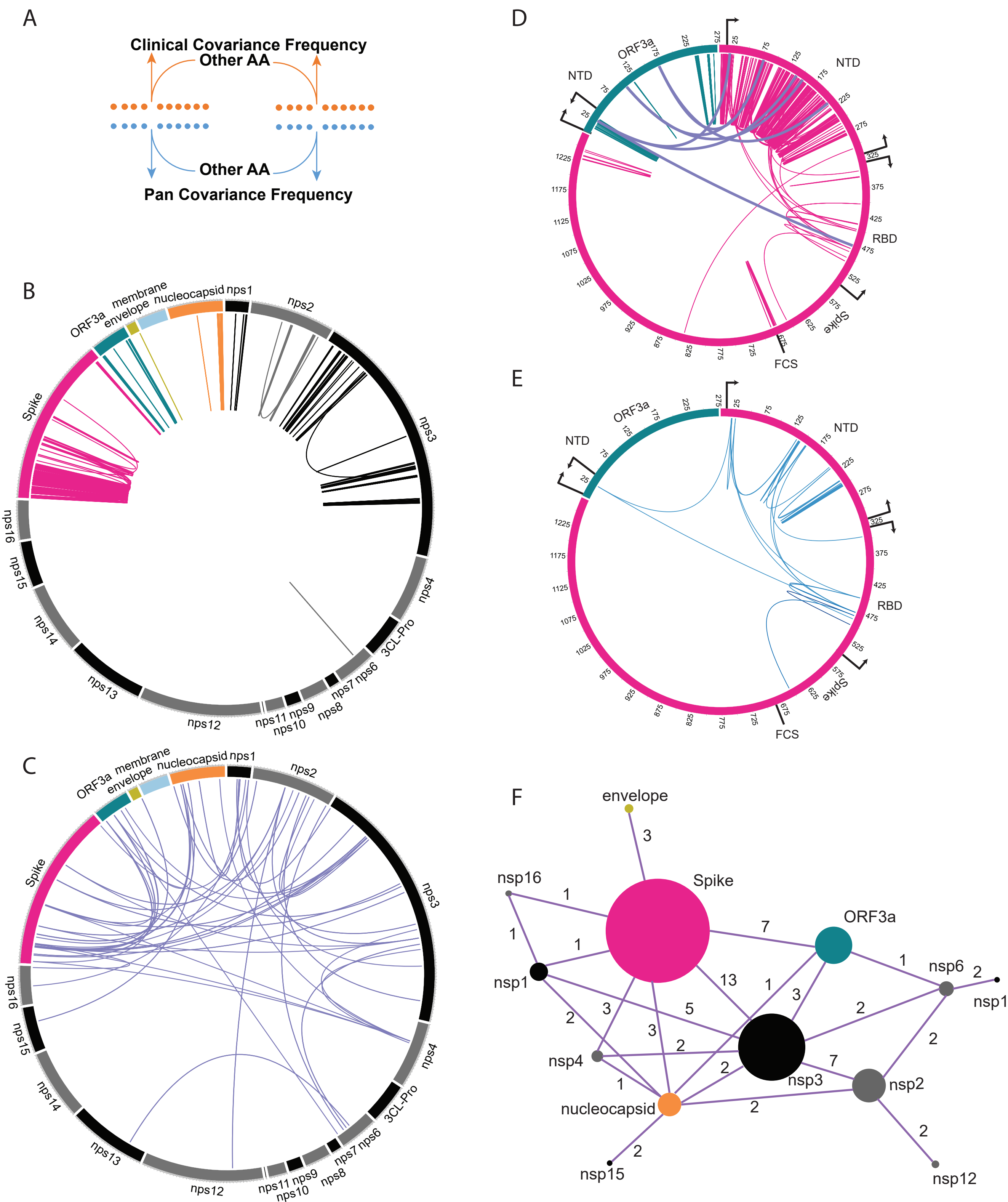

A
Clinical Covariance Frequency
Other AA
Other AA
Pan Covariance Frequency
B
Spike
ORF3a
envelope
membrane
nucleocapsid
nps1
nps2
nps3
nps4
3CL-Pro
nps6
nps7
nps8
nps9
nps10
nps11
nps12
nps13
nps14
nps15
nps16
C
Spike
ORF3a
envelope
membrane
nucleocapsid
nps1
nps2
nps3
nps4
3CL-Pro
nps6
nps7
nps8
nps9
nps10
nps11
nps12
nps13
nps14
nps15
nps16
D
ORF3a
NTD
NTD
RBD
Spike
FCS
25
75
125
175
225
275
25
75
125
175
225
275
325
375
425
475
525
575
625
675
725
775
825
875
925
975
1025
1075
1125
1175
1225
E
ORF3a
NTD
NTD
RBD
Spike
FCS
25
75
125
175
225
275
25
75
125
175
225
275
325
375
425
475
525
575
625
675
725
775
825
875
925
975
1025
1075
1125
1175
1225
F
envelope
3
Spike
nsp16
1
1
1
7
ORF3a
1
nsp6
2
nsp13
nsp1
13
1
3
2
3
5
2
3
2
nsp4
2
7
nsp3
nsp2
1
nucleocapsid
2
2
2
nsp15
nsp12

Figure 6

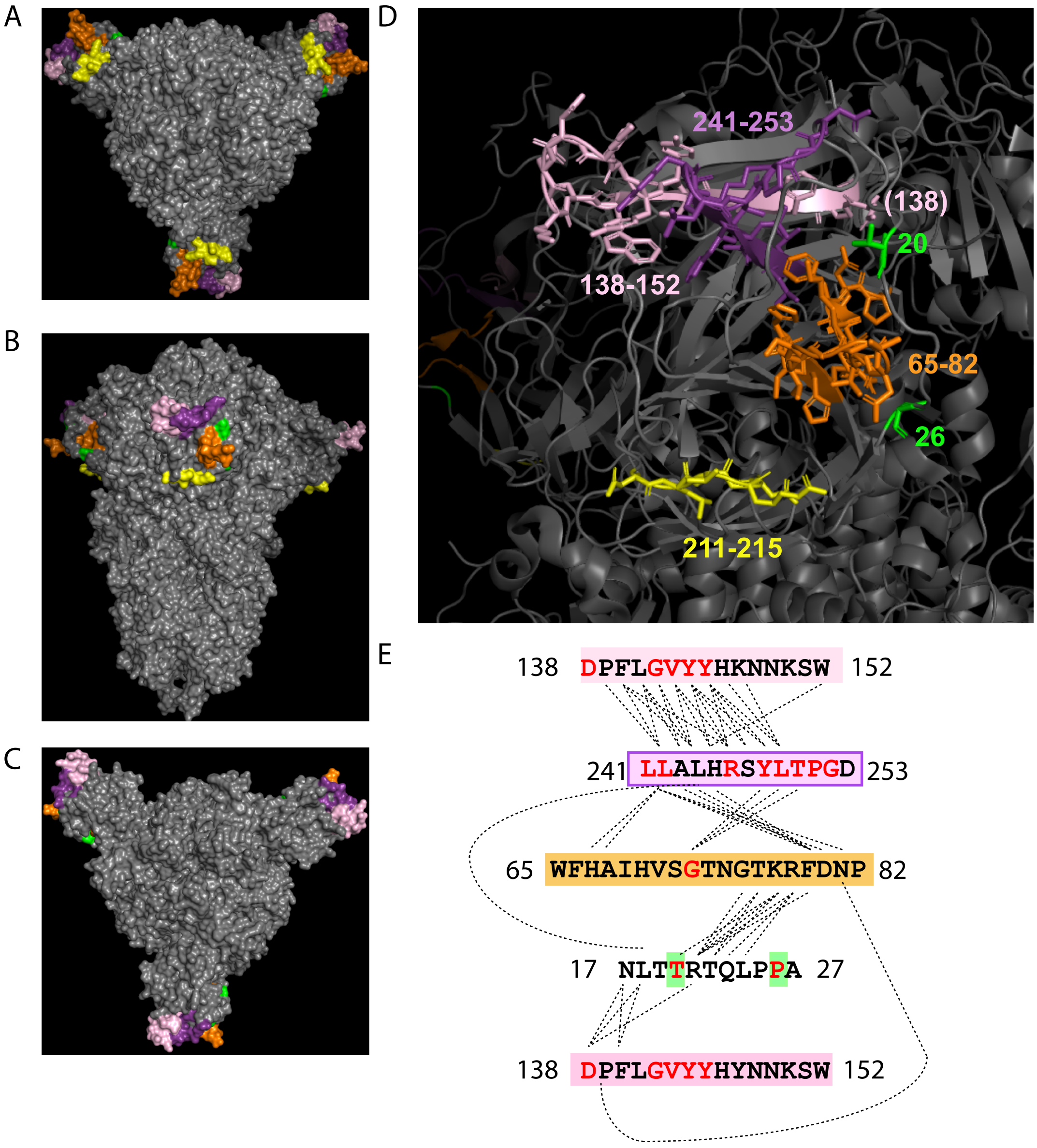
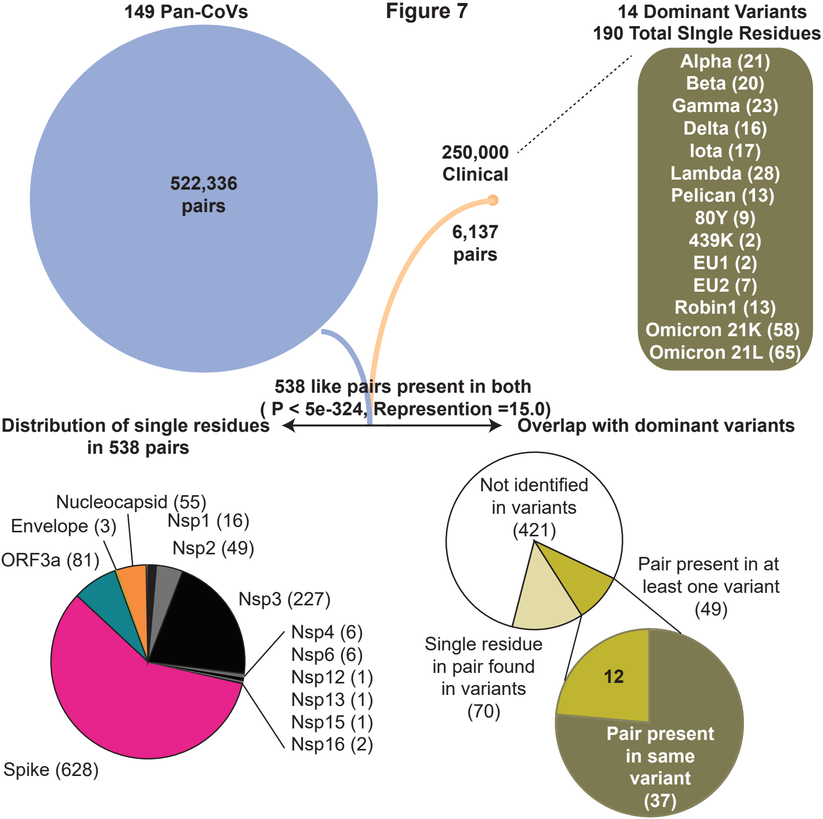

A
B
C
D
241-253
(138)
20
138-152
65-82
26
211-215
E
138 DPFLGVYYHKNNKSW 152
241 LLALHRSYLTPGD 253
65 WFHAIHVSGTNGTKRFDNP 82
17 NLTTRTQLPPA 27
138 DPFLGVYYHYNNKSW 152

Figure 7


149 Pan-CoVs
522,336
pairs
250,000
Clinical
6,137
pairs
14 Dominant Variants
190 Total SIngle Residues
Alpha (21)
Beta (20)
Gamma (23)
Delta (16)
Iota (17)
Lambda (28)
Pelican (13)
80Y (9)
439K (2)
EU1 (2)
EU2 (7)
Robin1 (13)
Omicron 21K (58)
Omicron 21L (65)
538 like pairs present in both
(P < 5e-324, Represention =15.0)
Distribution of single residues
in 538 pairs
Overlap with dominant variants
Nucleocapsid (55)
Envelope (3)
Nsp1 (16)
Nsp2 (49)
ORF3a (81)
Nsp3 (227)
Nsp4 (6)
Nsp6 (6)
Nsp12 (1)
Nsp13 (1)
Nsp15 (1)
Nsp16 (2)
Spike (628)
Not identified
in variants
(421)
Pair present in at
least one variant
(49)
Single residue
in pair found
in variants
(70)
12
Pair present
in same
variant
(37)