Ongoing Global and Regional Adaptive Evolution of SARS-CoV-2

Nash D. Rochman[1,*], Yuri I. Wolf[1], Guilhem Faure[2], Feng Zhang[2,3,4,5,6] and Eugene V. Koonin[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894
[2]Broad Institute of MIT and Harvard, Cambridge, MA 02142; [3]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139; [4]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; [5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; and [6]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139

For correspondence: nash.rochman@nih.gov, koonin@ncbi.nlm.nih.gov

Keywords: SARS-Cov-2, phylogeny, ancestral reconstruction, epistasis, globalization

Abstract

Unprecedented sequencing efforts have, as of October 2020, produced nearly 200,000 genomes of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for COVID-19. Understanding the trends in SARS-CoV-2 evolution is paramount to control the pandemic, but analysis of this enormous dataset is a major challenge. We show that the ongoing evolution of SARS-CoV-2 over the course of the pandemic is characterized primarily by purifying selection but a small set of sites, including spike 614 and nucleocapsid 203-204 appear to evolve under positive selection. In addition to the substitutions in the spike protein, multiple substitutions in the nucleocapsid protein appear to be important for SARS-CoV-2 adaptation to the human host. The positively selected mutations form a strongly connected network of apparent epistatic interactions and are signatures of major partitions in the SARS-CoV-2 phylogeny. These partitions show distinct spatial and temporal dynamics, with both globalization and diversification trends being apparent.

Main

High mutation rates of RNA viruses*(1)* enable virus adaptation at a staggering pace. Nevertheless, robust sequence conservation indicates that purifying selection is the principal force shaping the evolution of virus populations*(2,3,4,5)*. The fate of a novel zoonotic virus is in part determined by the race between public health intervention and viral diversification. Even intermittent periods of positive selection can result in lasting immune evasion, leading to oscillations in the size of the susceptible population and ultimately a regular pattern of repeating epidemics, as has been demonstrated for Influenza*(6,7,8)*.

During the current coronavirus pandemic, understanding the degree and dynamics of the diversification of severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) is essential for establishing a practicable, proportionate public health response, from guidelines on isolation and quarantine to vaccination*(9)*. To investigate evolution of SARS-CoV-2, we aggregated all available SARS-Cov-2 genomes as of October 11, 2020, from the three principal repositories: Genbank*(10)*, Gisaid*(11)*, and CNCB*(12)*. From the total of 197,453 submissions in these databases, 82,592 unique SARS-Cov-2 genome sequences were identified, and 39,695 high quality sequences were incorporated into a global multisequence alignment (MSA) consisting of the concatenated open reading frames with stop codons trimmed. The vast majority of the sequences excluded from the MSA were removed due to a preponderance of ambiguous characters (see Methods). The sequences in the final MSA correspond to 73,236 isolates with date and location metadata.

Several methods for coronavirus phylogenetic tree inference have been tested*(13,14)*. The construction of a single high-quality tree from 40,000 30 kilobase (kb) sequences using any of the existing advanced methods is computationally prohibitive. Therefore, building on the available techniques, we assembled a set of maximally diverse subtrees over a reduced alignment. This alignment was constructed mainly through the removal of sites invariant up to the exception of a single sequence (see Methods). The omission

of these sites created additional sequence redundancy which was then eliminated by removing additional sequences. The subtrees generated from the resulting MSA were then used to constrain a single composite tree, both steps requiring a sophisticated phylogeny reconstruction approach (see Methods). This composite tree reflects the correct topology but has incorrect branch lengths and was in turn used to constrain a global tree over the entire MSA (Fig. 1A). A complete reconstruction of ancestral sequences was then performed by leveraging Fitch traceback*(15)* (see Methods), enabling comprehensive identification of nucleotide and amino acid replacements across the tree.

We identified 6 principal partitions within this tree, in general agreement with other work*(16,17,18)*, along with two region-specific clades within partitions 3 and 6 (Fig. 1A) that, as discussed below, are important for the interpretation of the metadata. Given the short evolutionary distances between SARS-CoV-2 isolates, the topology of the tree is a cause of legitimate concern*(14,19,20,21)*. For the analyses presented below, we rely on a single, explicit tree topology which is probably one of many equally likely estimates*(14)*. Therefore, we sought to validate the robustness of the major partitions of the virus genomes using a phylogeny-free approach. To this end, pairwise Hamming distances were computed for all sequences in the MSA and the resulting distance matrix was embedded within a 3-dimensional subspace using classical multidimensional scaling (Fig. 1B). In this embedding, all 6 partitions are nearly completely separated, and the optimal clustering, determined by *k*-means, returned 4 categories (see Methods, Fig. S1), grouping together partitions 1 and 2 as well as 4 and 5. These findings indicate that an alternative tree with a comparable likelihood but a dramatically different coarse-grain topology, most likely, cannot be constructed from this MSA.

Each of the 6 partitions can be characterized by a specific non-synonymous substitution signature (Figs. 1C, S2), generally, corresponding to the most prominent non-synonymous substitutions across the tree (Table S1), some of which are shared by two or more partitions and appear independently many times, consistent with other reports*(22)*. The well known D614G substitution in the spike protein is part of these

signatures, and so are substitutions in two adjacent sites in the nucleocapsid protein (see below). The rest of the signature sites are in the nonstructural proteins 1ab, 3a, and 8 (Fig. 1C). The identification of these prevailing non-synonymous substitutions and an additional set of frequent synonymous substitutions suggested that certain sites in the SARS-CoV-2 genome might be evolving under positive selection. However, uncovering the selective pressures affecting virus evolution was complicated by non-negligible mutational biases. The distributions of the numbers of both synonymous and non-synonymous substitutions across the genome were found to be substantially overdispersed compared to both the Poisson and normal expectations (Fig. S3). Examination of the relative frequencies of all 12 possible nucleotide substitutions indicated a significant genome-wide excess of C to U mutations, approximately 3 fold higher than any other nucleotide substitution, with the exception of G to U, as well as some region-specific trends(Fig. S4).

Motivated by this observation, we compared the trinucleotide contexts of synonymous and non-synonymous substitutions as well as the contexts of low and high frequency substitutions. The context of high-frequency events, both synonymous and non-synonymous, was found to be dramatically different from the background frequencies. The NCN context (that is, all C->D mutations) harbors substantially more events than other contexts (all 16 NCN triplets are within the top 20 most high-frequency-biased ones, see Methods and Fig. S5) and is enriched in mutations uniformly across the genome including both synonymous and non-synonymous sites as well as low and high frequency sites. This pattern suggests a mechanistic bias of the coronavirus RNA-dependent RNA polymerase (RdRP). Evidently, such a bias that increases the likelihood of observing multiple, independent mutations in the NCN context complicates the detection of selection pressures. However, whereas all the sites with an excess of synonymous nucleotide substitutions are NCN and thus can be inferred to originate from the mutational bias, this is not the case for non-synonymous substitutions, suggesting that at least some of these are driven by other mechanisms. Thus, we excluded all synonymous substitutions and the non-synonymous substitutions with the

NCN context from further consideration as candidate sites evolving under positive selection.

Beyond this specific context, the presence of any hypervariable sites complicates the computation of the *dN/dS* ratio, the gauge of protein-level selection*(23*), which requires enumerating the number of synonymous and non-synonymous substitutions within each gene. Hypervariable sites bias this analysis, and therefore, we used two methods to ensure reliable estimation of *dN/dS*. For each protein-coding gene of SARS-CoV-2 (splitting the long orf1ab into 15 constituent non-structural proteins), we obtained both a maximum likelihood estimate of *dN/dS* across 10 sub-alignments and an approximation computed from the global ancestral reconstruction (see Methods). This approach was required due to the size of the alignment, over which a global maximum likelihood estimation would be computationally prohibitive. Despite considerable variability among the genes, we obtained estimates of substantial purifying selection ($0.1<dN/dS<0.5$) across the majority of the genome(Fig. S6), with a reasonable agreement between the two methods. This estimate is compatible with previous demonstrations of purifying selection among diverse RNA viruses*(3)* affecting about 50% of the sites surveyed or more*(2)*.

Thus, the evolution of SARS-CoV-2 appears to be primarily driven by substantial purifying selection. However, a small ensemble of non-synonymous substitutions appeared to have emerged multiple times, independently, and were not subject to an overt mechanistic bias. Due to the existence of many equally likely trees, in principle, in one or more of such trees, any of these mutations could resolve to a single event. However, such a resolution would be at the cost of inducing multiple parallel substitutions for other mutations, and thus, we conclude that more than 100 codons in the genome have undergone multiple parallel mutations in the course of SARS-CoV-2 evolution.

One immediate explanation of this observation is that these sites evolve under positive selection. The possible alternatives could be that these sites are mutational hotspots or

that the appearance of multiple parallel mutations was caused by numerous recombination events (either real or artifacts caused by incorrect genome assembly from mixed infections) in the respective genomic regions. Contrary to what one would expect under the hotspot scenario, we found that codons harboring many synonymous substitutions tend to harbor few non-synonymous substitutions, and vice versa (Fig. S7 A). Although when a moving average with increasing window size was computed, this relationship reversed (Fig. S7 B&C), the correlation between synonymous and non-synonymous substitutions was weak. Most sites in the virus genome are highly conserved, those sites that harbor the highest number of mutations tend to reside in conserved neighborhoods, and the local fraction of sites that harbor at least one mutation strongly correlates with the moving average (Fig. S8). Thus, whereas our observations indicate that SARS-CoV-2 genomes are subject to diverse site-specific and regional selection pressures, we did not detect regions of substantially elevated mutation or recombination.

Given the widespread purifying selection, substantially relaxed selection at any site is expected to permit multiple, parallel non-synonymous mutations to the same degree that any site harbors multiple, parallel synonymous mutations. Thus, seeking to identify sites subject to positive selection, we focused only on those non-synonymous substitutions that independently occurred more frequently than 95% of all synonymous substitutions excluding the mutagenic NCN context (see Methods). Most if not all sites in the SARS-CoV-2 genome that we found to harbor such frequent, parallel non-synonymous substitutions not subject to the restrictions discussed above are likely to evolve under positive selection(Fig. 1D, Table S2).

Having identified the set of potential positively selected residues, we examined the tree for evidence of epistasis*(24)* (see Methods) among these sites and revealed a network of co-occurring substitutions suggestive of epistatic interactions (Fig. 1E, Table S3). Strikingly, both D614G in the spike (S) protein and two adjacent substitutions in the nucleocapsid (N) protein, R203K and G204R, are associated with exceptionally many interactions, forming the two main hubs of the network. Spike D614G appears to boost

the infectivity of the virus, possibly, by increasing the binding affinity between the spike protein and the cell surface receptor of SARS-CoV-2, ACE2 *(25)*. The non-synonymous mutations S|L54H,Q677H, A879S, and V1176F that are strongly linked to D614G in the network are in the spike protein itself although examination of the spike structure does not reveal direct physical interactions among these residues (Fig. S9). Another substitution, S|H49Y, with a weaker statistical association to S|D614G, involves a site that is close to 614 in the structure (Fig. S9). Conceivably, by increasing the receptor affinity, the D614G substitution in the spike protein opens up new adaptive routes for later steps in the viral lifecycle, but the specific mechanisms remain to be investigated experimentally.

Of further interest were four substitutions within the spike protein that have been observed in mink populations: H69del/V70del, Y453F, I692V and M1229I*(26-27)*. I692V never occurred within our tree, and this codon harbors many synonymous substitutions atc(I) to att(I), suggestive of a mutational hotspot. In contrast, 69-, 70-, 453F, and 1229I all appeared multiple times independently throughout the tree. These substitutions were widely spread over the tree although all occurred close to tree tips and none passed our criteria for positive selection. Two sites within the receptor-binding domain (RBD), N331 and N343, have been shown to be important for the maintenance of infectivity*(28)*. As could be expected, these amino acid residues are invariant. Four more substitutions in the RBD, among others, N234Q, L452R, A475V, and V483A, have been demonstrated to confer antibody resistance*(28)*. We found N234 to be invariant, whereas a few L452R and A475V mutations close to tree tips were detected as well as two independent V483A mutations at the base of slightly larger clades within partition 1. None of these sites passed our criteria for positive selection.

The two adjacent amino acid replacements in the N protein, R(agg)203K(aaa) and G(gga)204R(cga), appear together 9 times. Both sites are likely to evolve under positive selection and are adjacent to a third such site, S(agt)202N(aat). The replacements R(agg)203K(aaa) and G(gga)204R(cga) occur via three adjacent nucleotide substitutions which strongly suggests a single mutational event. Evolution of beta-

coronaviruses with high case fatality rates including SARS-CoV-2 was accompanied by accumulation of positive charges in the N protein that are thought to enhance its transport to the nucleus*(29)*. Although positions 202-204 are outside the known nuclear localization signals, the substitutions in these positions are statistically associated with the Q229H substitution in the N protein which occurs in a site known to be responsible for nuclear shuttling *(30)*. Furthermore, the rapid rise of the A220V substitution in the N protein (excluded from considered as a candidate for positive selection in our analysis due to its NCN context) in a European cohort during the summer of 2020 might be related to a transmission advantage of the variant harboring this substitution*(31)*. Conceivably, the substitutions in these two adjacent sites, in particular G(gga)204R(cga), which increases the positive charge, might contribute to the nuclear localization of the N protein as well. This highly unusual cluster of three putative positively selected amino acid substitutions in the N protein is a strong candidate for experimental study that could illuminate the evolution of SARS-CoV-2 pathogenicity.

Orf3a|Q57H is a third hub in the network and, although not considered a candidate for positive selection in our analysis due to its NCN context, ORF8 S84L is a hub in the larger epistatic network including all strongly associated residues (Fig. S10). Also of interest in this larger network is S|N439K that is linked to S|D614G and, despite its NCN context, is potentially subject to positive selection having been demonstrated to enable immune escape*(32)*.

In addition to the recurrent missense mutations that are likely to evolve under positive selection, we identified numerous nonsense mutations (Table S4), the most frequent one, orf8|18(UAA), appearing in 66 unique genome sequences. These nonsense substitutions, apparently, resulting in truncated proteins, occur almost exclusively within the minor SARS-CoV-2 ORFs. ORF8 has been implicated in the modulation of host immunity by SARS-CoV-2, so these truncations might play a role in immune evasion*(33-34)*.

Epistasis in RNA virus evolution, as demonstrated for influenza, can constrain the evolutionary landscape as well as promote compensatory variation in coupled sites, providing an adaptive advantage which would otherwise confer a prohibitive fitness cost*(35)*. Because even sites subject to purifying selection*(36)* can play an adaptive role through interactions with other residues in the epistatic network, the network presented here (Fig. 1E) likely underrepresents the extent of epistatic interactions occurring during SARS-CoV-2 evolution. The early evolutionary events that shaped the epistatic network conceivably laid the foundation for the diversification of the virus relevant to virulence, immune evasion, and transmission. Recent analysis of within-patient genetic diversity has shown that the most common mutations are highly diverse within individuals*(37)*. Such diversity could either result from multiple infections, or otherwise, could present evidence of an even greater role of positive selection affecting a larger number of sites than inferred from our tree (Fig. 1D). Similarly to the case of Influenza, positive selection on these sites could drive virus diversification and might support a regular pattern of repeat epidemics with grave implications for public health. An analysis of the relationships between the sequencing date and location of each isolate and its position within the tree can determine whether diversification is already apparent within the evolutionary history of SARS-CoV-2.

We first demonstrated a strong correlation between the sequencing date of SARS-CoV-2 genomes and the distance to the tree root (Fig. S11), indicating a sufficiently low level of noise in the data for subsequent analyses. Although examination of the global distribution of each of the 6 major SARS-CoV-2 partitions (Figs. S12-13) indicates considerable regional diversification, this variation could partly result from time-dependent fluctuations (Fig. 2). The beginning of the pandemic is primarily characterized by the global extinction of partitions 1 and 2 which were dominant initially, through the beginning of February, in all four regions encompassing the vast majority of the available sequences (United States, Europe, Asia, and Australia/New Zealand). Notably, partitions 1 and 2 lack the substitution S|D614G which is the consensus for all other partitions.

The four “late” partitions began rising to prominence thereafter, with different regional trends. A common global trend is the increasing prevalence of partition 6, the only one in which the adjacent substitutions N|R203K and G204R belong to the consensus. From April to August, partition 6 grew in prominence in all four regions, the only partition to do so (Figs. 2 and 3A, S14). This trend is most dramatic in Australia/New Zealand where all other partitions apparently went extinct by the beginning of June. Notably, most sequences from this region form a clade offset by a long branch within partition 6 (mutational signature shown in Fig. S15 notably including S|S477N). Although the rise of partition 6 in the US has been slow and unsteady, perhaps, the starkest contrast to the global ascension of partition 6 is the resurgence of partition 3 within Europe. Partition 3 seems to be poised to again become dominant in Europe after partition 6 had risen to greater than 80% frequency in July. In this case, as in the case of partition 6 in Australia/New Zealand, clustering of sequences offset by a long branch within partition 3 is observed. The non-synonymous substitution signature of this clade (Fig. S16) includes N|A220V such that nearly all European isolates belong to partition 6 and harbor N|R203K&G204R or belong to partition 3 and harbor N|A220V by the beginning of September. Motivated by this finding, we examined the mutational signature characterizing isolates collected after July 17 from the US, the region with the lowest fraction of partition 6 isolates, compared with those collected after July 17 from the remaining three regions (Fig. S17). Focusing only on substitutions in the N protein (Fig. S18), three additional mutations in the vicinity of motifs implicated in nuclear localization, N|P199L,S194L,T205I, were identified to be prevalent within the US (Fig. S19). Although only few isolates from Asia after late August are available within this dataset, the rapid decrease in partition 6 frequency corresponds to a rapid increase in N|S194L within the region. Taken together, isolates with at least 1 of these 5 mutations compose the majority in all regions by the beginning of August (Fig. 2E), strongly suggesting that this region of the N protein plays a prominent role in the adaptation of SARS-CoV-2 to human hosts.

Despite these global trends, regional dynamics make it difficult to assess whether partition 6 or, perhaps, a distinct clade within partition 3 are indeed more fit than the

other “late” partitions and whether SARS-CoV-2 is beginning to regionally diversify. To better assess the extent of regional divergence, we constructed two diversity measures, one partition-dependent and the other, partition-independent. First, we considered 3 groups of virus genomes: partitions 1&2 (without the consensus substitution S|D614G), partitions 3,4,5, and partition 6 (with consensus substitutions N|R203K&G204R), and computed the Hellinger distance of this three-group frequency distribution between all pairs of regions over a sliding window as a function of time (Fig. 3B). Next, we sampled pairs of isolates both within and between regions, computed the mean tree-distance between pairs over the same sliding window (Figs. 3C, S16), and examined the ratio of inter-regional and intra-regional tree-distances (see Methods).

Both measures reveal three principal stages of the pandemic. 1) The beginning of the year, through early February, when partitions 1 and 2 were dominant, was marked by regional diversification. 2) The extinction of the two “early” partitions signaled globalization and the spread of the mutation S|D614G from February through the end of May. By June, Australia/New Zealand began diverging from the rest of the globe and the remaining regions began to modestly diverge from one another as well. These trends are superimposed over steady, substantial intra-regional diversification (Fig. S16).

Such diversification of the virus could potentially pose problems for both testing and vaccine development. Substitutions in the E protein have already been demonstrated to interfere with a common PCR assay*(38)*. Generally, ORF1ab is more conserved than the spike protein, which itself is more conserved than the remaining ORFs (Figs. S3-4). Using our SARS-CoV-2 MSA, we surveyed 10 regions from ORF1ab(5), N(4), and E(1) genes that are commonly used within PCR assays *(39)* for substitutions relative to the reference sequence. Among the nearly 40,000 genome sequences, there were hundreds to thousands of nucleotide substitutions in each of these regions but those in ORF1ab were markedly less variable than those in N (Supplementary table 5) with one region in N demonstrating variability in nearly half of all isolates. It can be expected that

most targets within the polyprotein will remain subject to the fewest polymorphism-induced false negatives even as the virus continues to diversify.

Of the 9 vaccine candidates currently in phase 3 trials*(40)*, three are inactivated whole-virus (Sinovac, Wuhan Institute of Biological Products/Sinopharm, Beijing Institute of Biological Products/Sinopharm); five utilize the entire spike protein as the antigen (Moderna/NIAID, CanSino Biological Inc./Beijing Institute of Biotechnology, University of Oxford/AstraZeneca, Gamaleya Research Institute, Janssen Pharmaceutical Companies) and one utilizes only the RBD (Pfizer/Fosun Pharma/BioNTech). In addition to the greater sequence conservation of the spike protein relative to all other ORFs outside of the polyprotein, it is the principal host-interacting protein of SARS-CoV-2, making both the whole protein and the RBD obvious antigenic candidates. Most mutations in the RBD were demonstrated to decrease infectivity; however, some conferred resistance to neutralizing antibodies*(27)*. We only identified one mutation in the RBD that was both observed in greater than 1% of the tree branches and passed our criteria for positive selection, S|S477N. Should such mutations become more prominent, the RBD might prove a less effective antigen than the whole spike protein.

To summarize, from these findings, it is clear that, despite dramatically reduced travel*(41)*, the evolution of SARS-Cov-2 is at least partly shaped by globalizing factors, such as the increased virus fitness conferred by S|D614G, N|R203K&G204R, and other positively selected substitutions. There is no strong evidence of “speciation”, that is, formation of stable, diverging variants, a finding that bodes well for a successful vaccination campaign in the midterm. Nevertheless, it is equally clear that SARS-CoV-2 has the capacity to diversify, posing a risk of virus escape from immunity and hence repeat epidemics.

Author contributions

EVK initiated the project; NR and GF collected data; NR, GF, YIW, FZ and EVK analyzed data; NR and EVK wrote the manuscript that was edited and approved by all authors.

Acknowledgements

The authors thank Koonin group members for helpful discussions. NR, YIW and EVK are supported by the Intramural Research Program of the National Institutes of Health (National Library of Medicine).

References

[1] J.W. Drake, J.J. Holland. Mutation rates among RNA viruses. *Proceedings of the National Academy of Sciences* **96.24**, 13910-13913 (1999).

[2] J.O. Wertheim, and S.L. Kosakovsky Pond. Purifying selection can obscure the ancient age of viral lineages. *Molecular biology and evolution* **28.12**, 3355-3365 (2011)

[3] G.M. Jenkins et al. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. *Journal of molecular evolution* **54.2**, 156-165 (2002)

[4] E.C. Holmes. Patterns of intra-and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. Journal of virology **77.20**, 11296-11298 (2003)

[5] G. Jerzak, et al. Genetic variation in West Nile virus from naturally infected mosquitoes and birds suggests quasispecies structure and strong purifying selection. *The Journal of general virology* **86.Pt 8**, 2175 (2005)

[6] Y.I. Wolf, et al. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. *Biology direct* **1.1**, 34 (2006)

[7] R.M. Bush et al. Predicting the Evolution of Human Influenza A. *Science* **286.5446** (1999).

[8] R.M. Bush et al. Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Molecular biology and evolution* **16.11** (1999).

[9] A. Koirala, et al. Vaccines for COVID-19: The current state of play. *Paediatric respiratory reviews* **35**, 43-49 (2020)

[10] D.A. Benson, et al. GenBank. *Nucleic acids research* **41.D1**, D36-D42 (2012)

[11] S. Elbe, G. Buckland-Merrett. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* **1.1**, 33-46 (2017)

[12] W. Zhao et al. The 2019 novel coronavirus resource. *Hereditas* **42.2**, 212-221 (2020)

[13] R. Lanfear. *A global phylogeny of SARS-CoV-2 from GISAID data, including sequences deposited up to 20-August-2020*. *Zenodo* (2020). DOI: 10.5281/zenodo.3958883

[14] B. Morel et al. Phylogenetic analysis of SARS-CoV-2 data is difficult. *bioRxiv* (2020).

[15] W.M. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology* **20.4**, 406-416 (1971)

[16] S. Kumar et al. An evolutionary portrait of the progenitor SARS-CoV-2 and its dominant offshoots in COVID-19 pandemic. *bioRxiv* (2020).

[17] P. Forster et al. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences* **117.17**, 9241-9243 (2020)

[18] N.M. Fountain-Jones et al. Emerging phylogenetic structure of the SARS-CoV-2 pandemic. *bioRxiv* (2020).

[19] C. Mavian et al. Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. *Proceedings of the National Academy of Sciences* **117.23**, 12522-12523 (2020)

[20] S.J. Sánchez-Pacheco et al. Median-joining network analysis of SARS-CoV-2 genomes is neither phylogenetic nor evolutionary. *Proceedings of the National Academy of Sciences* **117.23**, 12518-12519 (2020)

[21] L. Pipes et al. Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. *bioRxiv* (2020).

[22] L. van Dorp et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infection, Genetics and Evolution* **104351** (2020)

[23] Z. Yang and N. Goldman. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution* **11.5** (1994).

[24] N.D. Rochman, Y.I. Wolf, E.V. Koonin. Deep phylogeny of cancer drivers and compensatory mutations. *Communications Biology* **3.1**, 1-11 (2020)

[25] B. Korber et al. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182.4**, 812-827 (2020)

[26] Statens Serum Institut. Mutations in the mink virus. https://www.ssi.dk/aktuelt/nyheder/2020/mutationer-i-minkvirus

[27] B. Munnink et al. Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans. *Science* **eabe5901** (2020)

[28] L. Qianqian et al. The Impact of Mutations in SARS-CoV-2 Spike on Viral Infectivity and Antigenicity. *Cell* **182,** 1284–1294 (2020)

[29] A.B. Gussow et al. Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proceedings of the National Academy of Sciences* (2020).

[30] K.A. Timani et al. Nuclear/nucleolar localization properties of C-terminal nucleocapsid protein of SARS coronavirus. *Virus research* **114.1-2**, 23-34 (2005)

[31] E. Hodcroft et al. Emergence and spread of a SARS-CoV-2 variant through Europe in the Summer of 2020. *medRxiv* (2020).

[32] E. Thomson et al. The circulating SARS-CoV-2 spike variant N439K maintains fitness while evading antibody-mediated immunity. *bioRxiv* (2020).

[33] Y. Zhang et al. The ORF8 Protein of SARS-CoV-2 Mediates Immune Evasion through Potently Downregulating MHC-I. *bioRxiv* (2020).

[34] L. Zinzula Lost in deletion: The enigmatic ORF8 protein of SARS-CoV-2. *Biochemical and Biophysical Research Communications* (2020).

[35] L.I. Gong, M.A. Suchard, J.D. Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* **2**, e00631 (2013)

[36] S. Kryazhimskiy et al. Prevalence of epistasis in the evolution of influenza A surface proteins. *PLoS Genet* **7.2**, e1001301 (2011)

[37] J. Kuipers et al. Within-patient genetic diversity of SARS-CoV-2. *bioRxiv* (2020).

[38] M. Artesi et al. A recurrent mutation at position 26340 of SARS-CoV-2 is associated with failure of the E gene quantitative reverse transcription-PCR utilized in a commercial dual-target diagnostic assay. *Journal of clinical microbiology* **58.10** (2020).

[39] E. Ortiz-Prado et al. Clinical, molecular, and epidemiological characterization of the SARS-CoV-2 virus and the Coronavirus Disease 2019 (COVID-19), a comprehensive literature review. *Diagnostic Microbiology and Infectious Disease*. **115094** (2020).

[40] Y. Dong et al. A systematic review of SARS-CoV-2 vaccine candidates. *Signal Transduction and Targeted Therapy* **5**, 237 (2020).

[41] S. Lai et al. Assessing the effect of global travel and contact reductions to mitigate the COVID-19 pandemic and resurgence. *medRxiv* (2020).

Methods

Multiple alignment of SARS-CoV-2 genomes

All available SARS-CoV-2 genomes as of October 11, 2020 were retrieved from the Genbank*(10)*, Gisaid*(11)*, and CNCB*(12)* datasets. Sequences with apparent anomalies (sequence inversion etc.) were immediately discarded. Sequences were harmonized to DNA (e.g. U was transformed to T to amend software compatibility) and clustered according to 100% identity with no coverage threshold using CD-HIT*(42-43)*, with ambiguous characters masking. All characters excepting ACGT were considered ambiguous. The least ambiguous sequence from each cluster was selected and sequences shorter than 25120 nucleotides were discarded.

Exterior ambiguous characters (preceding/succeeding the first/last defined nucleotide) were removed and sequences with more than 10 remaining, interior, ambiguous characters were discarded. The remaining sequences were aligned using multi-threaded MAFFT*(44)* with 220 cores (--thread 220) and 3.8Tb of RAM to maintain usage of the normal DP algorithm *(44)* (--nomemsave). Sequences sourced from non-human hosts were manually identified from the metadata and those excluded at the previous step were added to the alignment using MAFFT, maintaining the number of columns in the original alignment (specifying --keeplength), again on 220 cores.

Sites corresponding to protein-coding open reading frames were then mapped to the alignment from the reference sequence NC_045512.2 excluding stop codons as follows: 266-13468+13468-21552, orf1ab; 21563-25381, S; 25393-26217, orf3a; 26245-26469, E; 26523-27188, M; 27202-27384, orf6; 27394-27756, orf7a; 27756-27884, orf7b; 27894-28256, orf8; and 28274-29530, N. The remaining sites were discarded.

The resulting alignment contained out-of-frame gaps. Gaps in the reference sequence were found to correspond to gaps in all but fewer than 1% of the remaining sequences. These sites were discarded. The remaining gaps shorter than three nucleotides were replaced with the ambiguous character, N. Longer gaps were shifted into frame and padded with ambiguous characters on either end of the gap, minimizing the number of sites altered.

A fast, approximate tree was then built using FastTree*(45)* (parameters: -nt -gtr -gamma -nosupport -fastest) to unambiguously define two clusters of sequences: an outgroup consisting of 14 sequences sourced from non-human hosts prior to 2020 and the main group. The tree construction requires the resolution of very short branch lengths which makes it necessary to compile FastTree at double precision. Outliers from the remaining sequences were then identified based on the Hamming distance (excluding gaps and ambiguous characters) to the nearest neighbor, the Hamming distance to the consensus, and the degree to which those substitutions relative to consensus were clustered in the genome. At this step, 21 sequences were removed.

The resulting alignment, consisting of 39,695 sequences and 29,119 sites, was maintained for the construction of the global tree and ancestral sequence reconstruction. In an effort to minimize the impact of sequencing error on the tree topology, as well as to decrease computational costs, a reduced alignment was then constructed through the removal of 1) invariant sites, 2) sites invariant with the exception of a single sequence, and 3) sites invariant throughout the main group with the exception of at most one sequence representing each minority nucleotide. Removing these sites created substantial redundancy, so a representative sequence was selected for each cluster of 100% identity to yield an alignment consisting of 34,685 sequences and 10,131 sites.

Tree Construction

We sought to optimize tree topology with IQ-TREE*(46)*; however, we found building the global tree to be computationally prohibitive, and thus, we proceeded to subsample the main group alignment as follows. First, a core set of maximally diverse sequences is selected. The set is initialized with a pair of sequences: a sequence maximizing the number of substitutions relative to consensus and a paired sequence which maximizes the Hamming distance to itself. Sequences are then added to this core set one at a time maximizing the minimum Hamming distance to any representative of the set until *N* sequences are incorporated. Next, $ceil\left(L/(M-N)\right)$ resulting sets are initialized with this core set where *M* is the target number of sequences and *L* is the total number of sequences in the alignment (34,685). Then, sequences that have not yet been incorporated into any resulting set are added to each resulting set, again one at a time, maximizing the minimum distance to any representative of the set until *M* sequences are incorporated. The order of the resulting sets is randomized at each iteration without repeats. Once every (main group) sequence has been incorporated into at least one resulting set, sequences are randomly incorporated into each set until every set contains *M* sequences. Finally, the outgroup is added to each resulting set. We chose *M*=3,000 in an effort to optimize computational efficiency and *N*=300. Note that while increasing *N* increases the number of sets required for alignment coverage, and thus compute time, insufficient overlap between the sequences assigned each sub-alignment

greatly affects the results of subsequent steps. We found *N*=100 to be too low to effectively constrain the global tree for this dataset.

A tree was then built, using IQ-TREE, for each resulting set, with the evolutionary model fixed to GTR+F+G4 and the minimum branch length decreased from the default 10e-6 to 10e-7, according to the results of previous parameter studies*(14)*. These trees were then converted into constraint files and merged to generate a single global constraint file for use within FastTree (parameters: -nt -gtr -gamma -cat 4 -nosupport -constraints).

The remaining sequences excluded from this tree were then reintroduced as unresolved multifurcations and a new constraint file from the multifurcated tree was constructed. A second iteration of FastTree was initiated on the whole alignment including all sites to produce the final tree. This tree was rooted at the outgroup.

Reconstruction of Ancestral Genome Sequences

Ancestral states were estimated by Fitch Traceback*(15)*. Briefly, character sets were constructed from leaf to root where each node was assigned the intersection of the descendant character sets if not empty and the union otherwise. Then, moving from root to leaf, nodes with more than one character in their set were assigned the consensus character if present in their set or a randomly chosen representative character otherwise. Substitutions between states were identified and placed in the middle of the branch bridging the pair of nodes.

Statistical associations between mutations were computed in a manner similar to that previously described*(24)*. Briefly, sequences were leaf-weighted based on the branch lengths of the ultrameterized, tree. Every mutation present across the tree at 50 mean leaf-weight equivalents or more was considered. The probability of independent co-occurrence between any pair was estimated in two ways. An arbitrary member of the pair was selected as the ancestral mutation, and the binomial probability:

$$\sum_{k=N_{pair}}^{N_{total}} \binom{N_{total}}{k} F^k (1-F)^{N_{total}-k}$$

was computed where *N_total* is the number of substitutions to the descendant mutation across the entire ancestral record, *N_pair* is the number of substitutions to the descendant which succeed or appear simultaneously with a substitution to the ancestral mutation, and *F* is the fraction of the tree (fraction of all applicable branch lengths) occupied by the ancestral mutation. The ancestral/descendent designation was then reversed and the “binomial score” was constructed as the negative log of the product of these two terms. Additionally, for each pair, the observed and expected (product of the

tree fractions) tree intersections were calculated and the "Poisson score" (analogous to the log-odds ratio) was calculated:

$$\begin{cases} -\ln\big(1 - PCDF(exp, obs)\big), obs > exp \\ \ln\big(PCDF(exp, obs)\big), obs < exp \end{cases}$$

where PCDF(exp,obs) is the cumulative probability of a Poisson distribution with mean "exp", the expected value of the data, and evaluated at "obs", the observed value of the data. Both scores are reported. Fig. 1E and Table S3 display putative positively selected mutations with both scores above 5 or at least two simultaneous substitutions. Fig. S10 has a relaxed score threshold for 2, an increased weight threshold of 100, and is not restricted to positively selected residues.

Classical Multidimensional Scaling of the MSA

Pairwise Hamming distances were computed for all pairs of rows in the global MSA ignoring gaps and ambiguous characters i.e. the sequences *X*="ATN-A" and *Y*="NTAAT" would be assigned a distance of 1. The resulting distance matrix was embedded in three dimensions with the MATLAB*(47)* routine "cmdscale". 100 rounds of stochastically initiated k-means clustering of the embedding was conducted and the optimum cluster number was determined to be 4 on the basis of the silhouette score distribution (Fig S1).

Validation of Mutagenic Contexts

Mutations were divided into four categories: synonymous vs non-synonymous substitutions and high vs low frequency of independent occurrence. For example, consider codon X with 3 non-synonymous substitutions gat->ggt and 1 non-synonymous substitution gat->cgt. In this context, a non-synonymous nucleotide substitution a->g of frequency 4 would be recorded in nucleotide (X-1)*3+2. The low vs high frequency threshold was determined by the 95th percentile of the synonymous mutation frequency distribution (operationally. 5). For each mutation, the trinucleotide contexts from the ancestral reconstruction at the nodes where the mutation occurred were compared to the background genome-wide frequencies, computed for the inferred common ancestor of SARS-CoV-2.

The expected frequencies of the trinucleotides using the background distribution were tabulated; the Yates correction (+/-0.5 to the original count depending on whether the count is below or above the expectation) was applied to the observed frequencies; the log-odds ratios of the (corrected) observed frequencies to the expectation were computed; and CMDS was applied to the Euclidean distances between the log-odds

vectors to embed the points onto a plane (Fig. S5 A.). This analysis was then repeated, this time, distinguishing only between high and low frequency substitutions but not N and S (Fig. S5 B). Finally, the differences in the contexts of high frequency synonymous vs non-synonymous events were considered in the same manner and the chi-square statistics ((observed-expected)^2/expected) were compared with the critical chi-square value (p=0.05/64, df=1, Fig. S5 C.).

Computation of *dN/dS*

For each of the 24 ORFs (splitting orf1ab into 15 segments corresponding to the 15 mature proteins, nsp11 and nsp12 combined), 10 reduced alignments were constructed as follows. Sequences were ordered based on diversity, in the same order with which they were included in the constraint trees. The first 10 sequences are conserved across every alignment and the remaining 40 are unique to each alignment. The reference sequence, NC_045512.2, was additionally added to each reduced alignment. PAML*(48)* was then used to estimate tN, tS, *dN/dS*, N, S, and N/S for each segment and every reduced alignment.

Given the global ancestral reconstruction from Fitch traceback, the total number of non-synonymous and synonymous substitutions (nN and nS, respectively) as well as these tallies normalized by the respective segment length (tN, and tS, respectively) were retrieved for each segment. . A hybrid *dN/dS* value for each segment was estimated to be (nN/nS)/(N/S)* where (N/S)* is the median value of N/S across all repeats for the segment.

Metadata Assignment

Headers for all isolates belonging to CD-HIT clusters with a representative incorporated into the alignment with fewer than 10 interior ambiguous characters were processed to extract the sequencing date and location. Sequencing location abbreviations were matched to full names and the latitude/longitude of a representative city for each location was retrieved from simplemaps (https://simplemaps.com/data/world-cities)*(49)*.

Regional Divergence Analysis

Two approaches, one partition dependent and one partition independent, were used. First three groups of isolates were constructed belonging to partitions 1&2, partitions 3,4,&5, and partition 6. The Hellinger distance between regions over a sliding time window was then computed between regions for this three group distribution. Next, 400 isolates were randomly selected from each region over a sliding window and 200 pairs within each region as well as 200 pairs between each pair of regions were composed.

683 The tree distance between each pair was computed and the mean for each inter- and
684 intra-regional pair tree-distance distribution was recorded.

685

686 Supplemental References:

687

688 [42] W. Li, A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or
689 nucleotide sequences. Bioinformatics **22.13**, 1658-1659 (2006)

690 [43] L. Fu et al. CD-HIT: accelerated for clustering the next-generation sequencing data.
691 Bioinformatics **28.23**, 3150-3152 (2012)

692 [44] K. Katoh et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast
693 Fourier transform. *Nucleic acids research* **30.14**, 3059-3066 (2002)

694 [45] M.N. Price, P.S. Dehal, A.P. Arkin. FastTree 2–approximately maximum-likelihood trees for
695 large alignments. *PloS one* **5.3**, e9490 (2010)

696 [46] L. Nguyen et al. IQ-TREE: a fast and effective stochastic algorithm for estimating
697 maximum-likelihood phylogenies. *Molecular biology and evolution* **32.1**, 268-274 (2015)

698 [47] MathWorks, Inc, ed. MATLAB, high-performance numeric computation and visualization
699 software: reference guide. MathWorks, (1992)

700 [48] Z. Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and*
701 *evolution* **24.8**, 1586-1591 (2007)

702 [49] World Cities Database. *simplemaps*. https://simplemaps.com/data/world-cities

703

Figure legends

Figure 1. Evolution of SARS-CoV-2.

A. Global tree reconstruction with 6 principal partitions enumerated and color-coded. The gray clades in partition 3/6 are composed of late sequences from Europe/Australia&New Zealand respectively. The majority of sequences from AU/NZ reside in the respective clade as do almost half of all late isolates from Europe. **B.** Projections of the 3D embedding of the pairwise Hamming distance matrix between SARC-CoV-2 genomes. The partitions are color-coded as in A. Wires enclose the convex hulls for each of the four optimal clusters. **C.** Signatures of amino acid replacements for each partition. Sites are ordered by decreasing maximum Kullback-Leibler divergence of the nucleotide distribution (sites are not consecutive in the SARS-CoV-2 proteins; the proteins along with nucleotide and amino acid numbers are indicated underneath each column) of any site in any partition relative to the distribution in that site over all partitions. **D.** Site history trees for spike 614 and nucleocapsid 203 positions. Nodes were included in this reduced tree based on the following criteria: those immediately succeeding a substitution; representing the last common ancestor of at least two substitutions; or terminal nodes representing branches of five sequences or more (approximately, based on tree weight). Edges are colored according to their position in the main partitions and the line type corresponds to the target mutation (solid) or any other state (dashed). These sites are largely binary as are most sites in the genome. The sizes of the terminal node sizes are proportional to the log of the weight descendent from that node beyond which no substitutions in the site occurred. Node color corresponds to target mutation (black) or any other state (gray). **E.** Network of putative epistatic interactions for likely positively selected residues.

Figure 2. Global and regional SARS-CoV-2 partition dynamics during the COVID-19 pandemic. A. US partition distribution over time. **B.** European partition distribution over time **D.** Asian partition distribution over time. **E.** Australian/New Zealand partition distribution over time. **E.** The frequency of mutation S|614G and N|194L, 119L, 203K, 205I, or 220V.

735

736 **Figure 3. Global and regional trends in SARS-CoV-2 evolution. A.** Global
737 distribution of sequences with sequencing locations in the US (brown), Europe (gray),
738 Asia (brown), and Australia/New Zealand (gray) identified. Color scheme is for visual
739 distinction only. Pie charts indicate the partition distributions for each region mid-March
740 through mid-April and mid-July through mid-August. **B.** The Hellinger distance between
741 the six pairs of regions over the four group distribution: partitions 1&2, partitions 3,4&5,
742 and partition 6. **C.** The ratio of the mean tree-distance for pairs of isolates between
743 regions vs. isolates within regions.

Supplemental Figures

Figure S1. 25th, median (solid line), and 75th percentiles of the silhouette score distribution for 100 stochastically initiated rounds of k-means clustering for 2-10 clusters.

Figure S2. The Kullback-Leibler divergence between each clade and the whole for the ten most divergent codons in the genome. The solid line indicates the maximum of any clade and points represent the remaining clades.

Figure S3. A. Distributions of the moving average, respecting segment boundaries, across a 100 codon window for synonymous (blue) and amino acid (orange) substitutions. Solid lines: normal approximations of the distributions (same median and interquartile distance); solid lines: approximation with the same median and theoretical (Poisson) variance. **B.** Moving averages, respecting segment boundaries, across a 100 codon window for synonymous and nonsynonymous substitutions per site, raw (top) and normalized by the median (bottom). There are several regions in the genome with an apparent dramatic excess of synonymous substitutions: 5' end of orf1ab gene; most of the M gene; 3'-half of the N gene, as well as amino acid substitutions: most of the orf3a gene; most of the orf7a gene; most of the orf8 gene; and several regions in of the N gene.

Figure S4. Moving average over a window of 1000 codons, not respecting segment boundaries, of the total number of nucleotide exchanges n1->n2 summed over all substitutions. The ratio to the median over the entire alignment is also displayed as well as the normalized exchange distribution *(i.e.* #c->t/(#c->t+#c->g+#c->a)).

Figure S5 A. Two dimensional embedding of the Euclidean distances between the log-odds vectors of low and high frequency, nonsynonymous and synonymous mutations in the space of trinucleotide contexts relative to background expectation. The context of the high-frequency events (both S and N) is dramatically different from the background frequencies. There is a strong common component in the deviation of both kinds of high-frequency events. The context of the low-frequency events (both S and N) also differs slightly, in the same direction, from the background frequencies. There is a consistent distinction between synonymous and non-synonymous events, suggesting that a single mutagenic context or mechanistic bias does not account for both S and N events. **B.** Log odds ratio of low and high frequency mutations, both synonymous and nonsynonymous, relative to background expectation for each trinucleotide context. The

NCN context (i.e. all mutations C->D) harbors dramatically more mutation events than the other contexts (all 16 NCN events are within the top 20 most-biased high-frequency events). The log-odds ratios for low-frequency events are somewhat correlated with those for high-frequency events (rPearson=0.5, without NCN rPearson=0.64), suggesting the same mechanism may be responsible for the strong bias observed among high frequency events and the weaker bias observed among low frequency events. **C.** Log odds ratio of high frequency nonsynonymous mutations relative to the background expectation from the sum of both high synonymous and high nonsynonymous mutations vs. the sum + 1. There are 12 contexts where synonymous and non-synonymous events differ significantly. All contexts with an excess of synonymous events are NCN, suggesting that high-frequency synonymous events could be driven by mechanistic bias; on the contrary, none of the contexts with an excess of non-synonymous mutations are NCN (in fact all are NGN: aga,agt,agg,ggt,agc,tgt,gct vs. aca,tct,tca,tcg,ccg), suggesting that these non-synonymous events could be driven by other mechanisms. There is no correlation between the frequency of event context and the log-odds ratio for non-synonymous events, further suggesting that the log-odds ratio is not biased by hot-spot mutation context.

Figure S6. Correspondence between the “tree length for dN”, “tree length for dS”, and *dN/dS* between PAML and the results of the ancestral reconstruction utilizing Fitch traceback across 24 ORFs. Two outliers in the PAML tS distribution are identified in each plot.

Figure S7. A. The number of nonsynonymous events vs the number of synonymous events per codon. **B.** The moving average of 100 codons, respecting segment boundaries. **C.** The moving average after removing events with 5 or more independent occurrences. Rho refers to Spearman. Dashed lines are 2/1.3*x reflecting the genome-wide ratio of nonsynonymous to synonymous substitutions, solid lines are linear best fit.

Figure S8. The fraction of sites with at least one substitution vs moving averages, respecting segment boundaries, over windows of 100 codons for synonymous and nonsynonymous substitutions.

Figure S9. Structural analysis for sites epistatically linked to spike D614 within the spike protein. D614 is at the interface between Spike chains. Most regions in the vicinity are not structurally solved potentially indicating that depending on the status of the RBD of the other chains, the regions in close proximity to D614 could become highly flexible. Residue 21 is not structurally solved; however, model inference suggests it is spatially

distant from residue 614. H49 makes a stack cation pi interaction with R44 within the same chain. H49 is spatially distant from D614, however, the domain it belongs to (circled in red) is linked by a linker (dashed red line) that leads to the domain containing D614 (circled in purple). This potentially functions as a holding point to position the purple domain. Note that 614 is very close to the cleavage site, likely requiring accurate positioning of this domain.

Figure S10. Epistatic network for the tree including mutations with binomial/poisson scores above 2 or at least two simultaneous substitutions and weight of at least approximately 100 leaves not restricted to likely positively selected residues.

Figures S11. Correlation between sequencing date and tree distance to the root for all isolates with metadata as well as those which appear explicitly in the tree.

Figures S12-13. Global distribution of sequences. Color represents the number of sequences from that location and size represents the fraction of sequences from the clade displayed. Clade indices are in the top left corner of each map.

Figure S14. Major partition distributions for each region at two fixed timepoints, mid-March to mid-April and mid-July to mid-August, as well as the difference.

Figure S15. The Kullback-Leibler divergence between the gray clade (solid line) and remaining sequences (dashed line) vs. the whole of partition 6. The ten most divergent codons in the genome are shown in the sequence logo.

Figure S16. The Kullback-Leibler divergence between the gray clade (solid line) and remaining sequences (dashed line) vs. the whole of partition 3. The ten most divergent codons in the genome are shown in the sequence logo.

Figure S17. The sequence logo for the 15 most divergent (Kullback-Leibler) codons in the genome considering US sequences sourced after July 17 vs. all sequences sourced after July 17.

Figure S18. The sequence logo for the 15 most divergent (Kullback-Leibler) codons in the nucleocapsid protein considering US sequences sourced after July 17th 2020 vs. all sequences sourced after July 17th 2020.

Figure S19. The frequencies of mutations S|614G, N|194L, N119L, N203K, N205I, and N220V over time.

Figure S20. The mean tree distance between pairs of isolates **A.** from different regions, **B.** within the same region, and **C.** the ratio over time.

Supplemental Tables

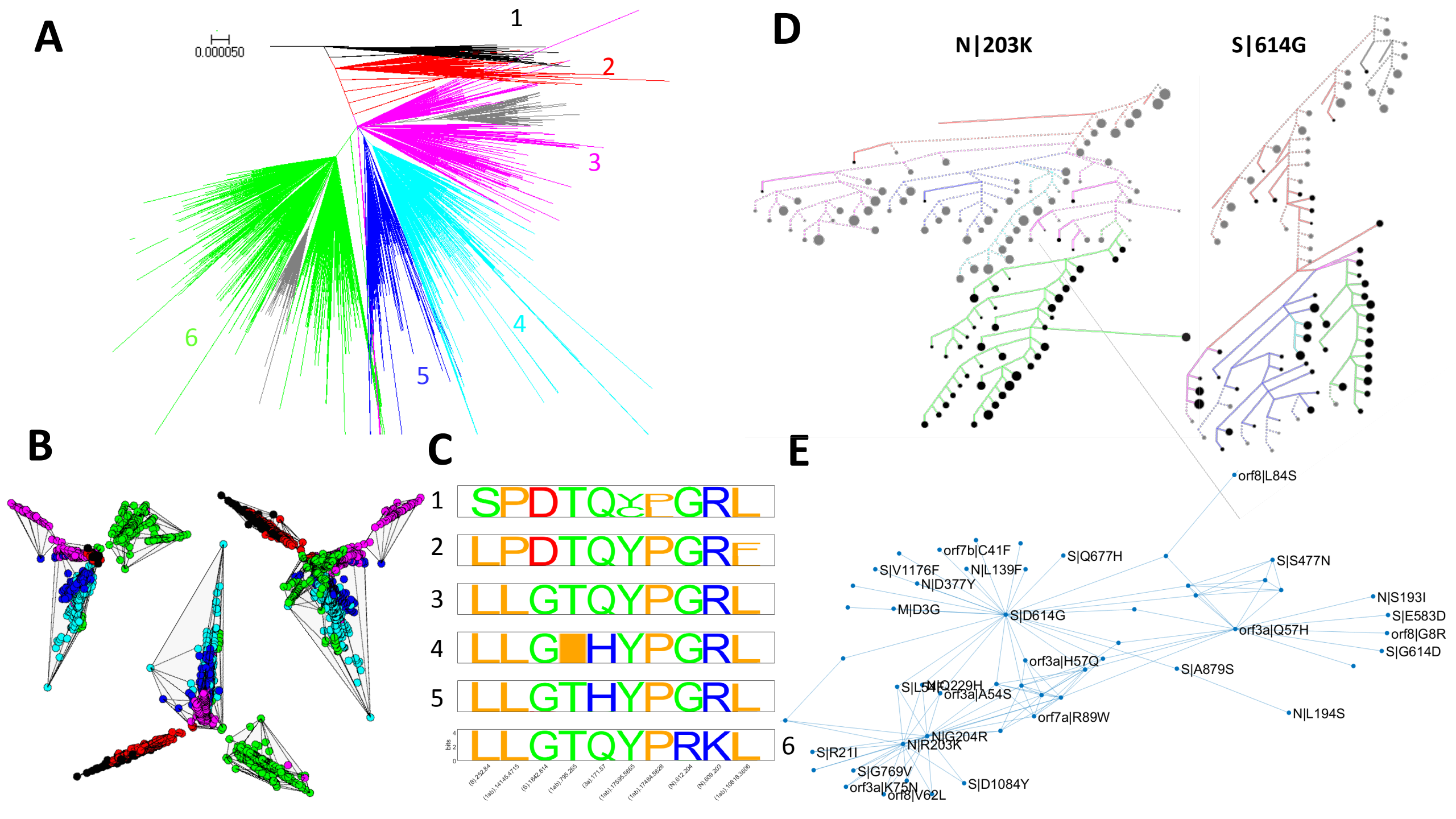
Table S1. The top ten mutations most commonly observed and the top ten with the greatest number of parallel substitutions.

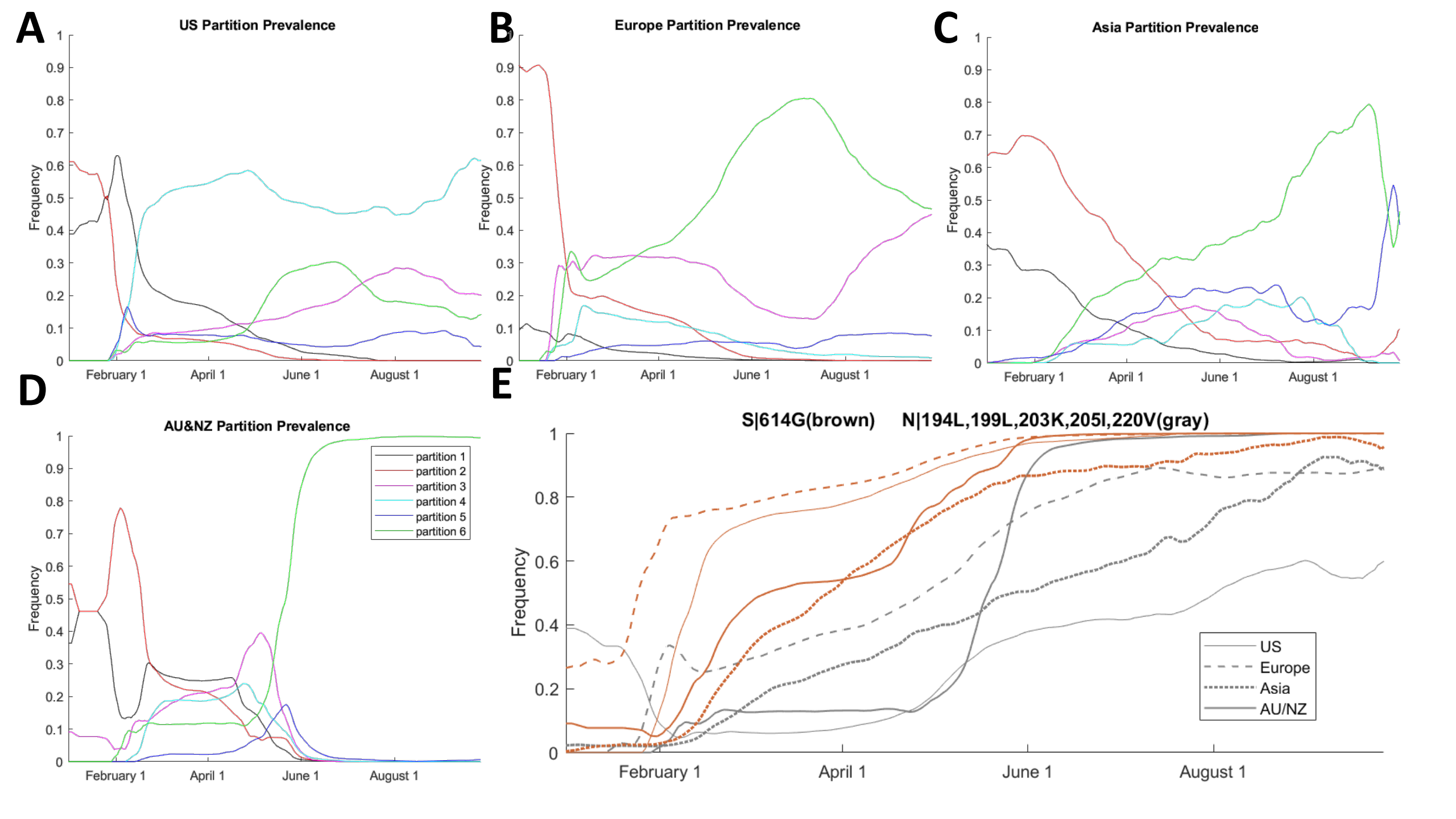
Table S2. List of sites likely to be evolving under positive selection.

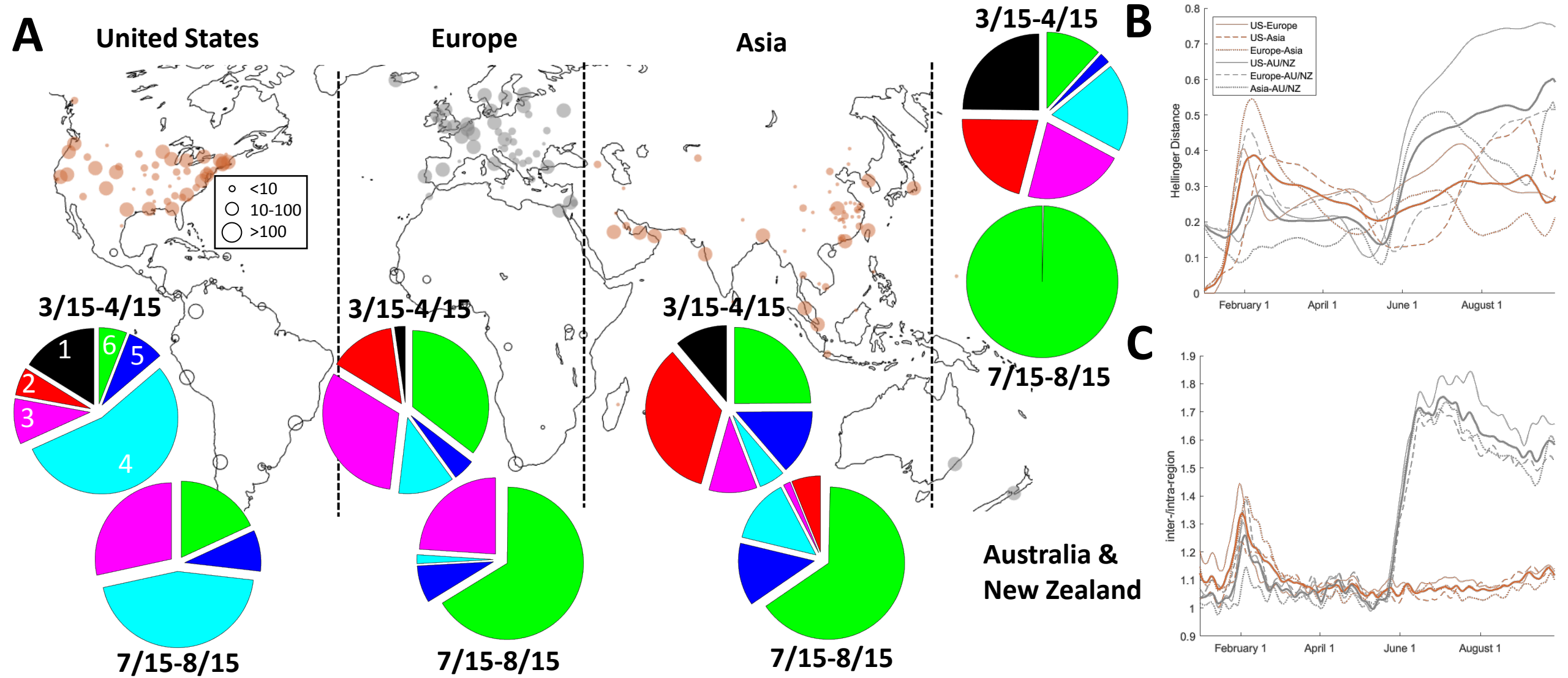
Table S3. All epistatic interactions among states meeting the criteria outlined in the main text for likely positive selection with binomial/Poisson scores greater than 5 or at least 2 simultaneous substitutions. Each mutation must have a minimum weight of approximately 50 leaves and each pair, 30 leaves. Each pair is arbitrarily ordered and the numbers of simultaneous, descendant, and independent substitutions are tabulated.

Table S4. Tabulated three codon neighborhoods for all sites containing at least one stop codon. Sites are ordered in decreasing number of sequences containing the stop. Stops are listed separately before all other neighborhoods.

Table S5. The number of isolates (out of approximately 73k) observed to bear at least one substitution relative to the reference sequence, NC_045512.2, within the regions specified. These regions are commonly used within PCR assays for diagnostic testing.


A
0.000050
1
2
3
4
5
6
B
C
1 SPDTQYPGRL
2 LPDTQYPGRE
3 LLGTQYPGRL
4 LLG HYPGRL
5 LLGTHYPGRL
6 LLGTQYPRKL
bits
(S).252.84
(1ab).14145.4715
(S).1842.614
(1ab).795.265
(3a).171.57
(1ab).17595.5865
(1ab).17484.5828
(N).612.204
(N).609.203
(1ab).10818.3606
D
N|203K
S|614G
E
orf8|L84S
orf7b|C41F
S|Q677H
S|S477N
S|V1176F
N|L139F
N|D377Y
N|S193I
M|D3G
S|D614G
S|E583D
orf3a|Q57H
orf8|G8R
S|G614D
orf3a|H57Q
S|A879S
S|L54F
N|Q229H
orf3a|A54S
N|L194S
orf7a|R89W
N|G204R
S|R21I
N|R203K
S|G769V
orf3a|K75N
S|D1084Y
orf8|V62L


A
US Partition Prevalence
B
Europe Partition Prevalence
C
Asia Partition Prevalence
Frequency
February 1
April 1
June 1
August 1
D
AU&NZ Partition Prevalence
partition 1
partition 2
partition 3
partition 4
partition 5
partition 6
E
S|614G(brown) N|194L,199L,203K,205I,220V(gray)
US
Europe
Asia
AU/NZ


A
United States
Europe
Asia
<10
10-100
>100
3/15-4/15
3/15-4/15
3/15-4/15
3/15-4/15
1
2
3
4
5
6
7/15-8/15
7/15-8/15
7/15-8/15
7/15-8/15
Australia &
New Zealand
B
US-Europe
US-Asia
Europe-Asia
US-AU/NZ
Europe-AU/NZ
Asia-AU/NZ
Hellinger Distance
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
February 1
April 1
June 1
August 1
C
inter-/intra-region
0.9
1
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
February 1
April 1
June 1
August 1