Ongoing Adaptive Evolution and Globalization 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 over 100,000 genomes of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that is responsible for the ongoing COVID-19 crisis. Understanding the trends in SARS-CoV-2 evolution is paramount to control the pandemic. Although this extensive data availability quickly facilitated the development of vaccine candidates[1], major challenges in the analysis of this enormous dataset persist, limiting the ability of public health officials to translate science into policy. Having evolved over a short period of time, the SARS-CoV-2 isolates show low diversity, necessitating analysis of trees built from genome-scale data. Here we provide a complete ancestral genome reconstruction for SARS-CoV-2 leveraging Fitch Traceback[2]. We show that the ongoing evolution of SARS-CoV-2 over the course of the pandemic is characterized primarily by purifying selection. However, a small set of sites, including the extensively studied spike 614[3], harbor mutations which recurred on multiple, independent occasions, indicative of positive selection. These mutations form a strongly connected network of apparent epistatic interactions. The phylogenetic tree of SARS-CoV-2 consists of 7 major clades which show distinct global and temporal dynamics. Periods of regional diversification of SARS-CoV-2 are short and, despite dramatically reduced travel[4], globalization of the virus is apparent.

Main

High mutation rates among RNA viruses[5] enable host adaptation at a staggering pace. Nevertheless, robust sequence conservation makes purifying selection the principal evolutionary force shaping virus populations[6,7,8,9]. 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 permit lasting immune evasion leading to oscillations in the size of the susceptible population and ultimately a regular pattern of repeat epidemics, as has been demonstrated for Influenza[10.]

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. To investigate evolution of SARS-CoV-2, we aggregated all available Sars-Cov-2 genomes as of July 28, 2020, from the three principle repositories: Genbank[11], Gisaid[12], and CNCB[13]. Out of 97,000 submissions, 45,000 unique sequences were identified and 20,000 were incorporated into a global multisequence alignment (MSA) consisting of the concatenated open reading frames with stop codons trimmed. The vast majority of sequences excluded from the MSA were removed due to a preponderance of ambiguous characters (see Methods).

A variety of methods for coronavirus phylogenetic tree inference have been tested[14,15]. The construction of a single high-quality tree from 20,000 30 kb sequences using any of the existing advanced methods is computationally prohibitive. Therefore, building on the available techniques, we assembled an ensemble of maximally diverse subtrees over a reduced alignment which contains fewer sites and consequently fewer unique sequences. These subtrees were then used to constrain a single composite tree. This composite tree reflects the correct topology but incorrect branch lengths and was in turn used to constrain a global tree over the entire MSA (Fig. 1A). A comprehensive reconstruction of ancestral sequences was then performed (see Methods), enabling the identification of nucleotide and amino acid replacements across the tree.

We identified 7 principal clades within this tree, in a general agreement with other work[16,17,18]; however, given the short evolutionary distances between SARS-CoV-2 isolates, the topology of the tree is a cause of legitimate concern[15,19,20,21]. For the analyses presented below, we rely on a single, explicit tree topology which is likely one of many equally likely trees[15]. Therefore, we sought to validate the robustness of the major clades using a phylogeny-free approach. Pairwise Hamming distances, ignoring ambiguous characters and gaps, were computed for all rows of the MSA and the resulting distance matrix was embedded within a 3-dimensional subspace using classical multidimensional scaling (Fig. 1B). In this embedding, all 7 clades are nearly completely separated and the optimal clustering, determined by k-means, returned 4 categories (see Methods, Fig. S1), two of which correspond to the major clades 3 and 5. These findings indicate it is unlikely an alternative tree with a comparable likelihood, but a dramatically different coarse-grain topology could be constructed for this MSA.

Each of the 7 clades 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 multiple clades and appear independently many times, consistent with other reports[22]. The well known D614G site in the spike protein is part of these signatures, and so are two adjacent sites in the nucleocapsid protein (see below). The rest of the signature sites are in the nonstructural proteins 1ab and 3a (Figure 1C). The identification of these prevailing non-synonymous substitutions and an additional set of frequent synonymous substitutions raised the possibility that certain sites in the SARS-CoV-2 genome might be evolving under positive selection. However, uncovering the selective pressures acting on this genome was complicated by non-negligible mutational biases. The distribution of the number of events per site is highly non-uniform for both synonymous and non-synonymous substitutions across the genome(Fig. S3). Both distributions are substantially overdispersed compared to both the Poisson and normal expectations, and examination of the relative frequencies of all 12 possible nucleotide substitutions indicates 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(Figs. S4-5).

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, Methods, Table S2) and is enriched 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 events are NCN and thus can be inferred to originate from the mutational bias, this is not the case for non-synonymous mutations, suggesting that at least some of the non-synonymous events could be driven by other mechanisms. We conservatively excluded all synonymous mutations and all non-synonymous mutations with 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 which is the gauge of protein-level selection. Therefore, for each protein-coding gene, splitting the long orf1ab into 15 constituent non-structural proteins, we obtained maximum likelihood estimates of *dN/dS* across 10 sub-alignments as well as approximations 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 the considerable variability between methods and among genes, we obtained estimates of substantial purifying selection ($0.1<dN/dS<0.5$) across the majority of the genome(Fig. S6). This estimate is compatible with previous work demonstrating purifying selection among disparate RNA viruses[7] affecting about 50% of the sites surveyed or more[6]

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 be resolved to a single event. However, such a resolution would be at the cost of inducing multiple parallel substitutions for other mutations, and thus, we can state conclusively that a small ensemble of sites in the genome have undergone multiple parallel mutations in the course of SARS-CoV-2 evolution. The 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 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 correlates well with the moving average (Fig. S8). Thus, overall, although our observations indicate that SARS-CoV-2 genomes are subject to diverse site-specific and regional selection pressures, we did not detect obvious regions of substantially elevated mutation or recombination.

Given the expectation of widespread purifying selection, it is reasonable to suspect that substantially relaxed selection in any given site would permit multiple, parallel non-synonymous mutations to the same degree that any site harbors multiple, parallel synonymous mutations. Accordingly, we focus only on those non-synonymous substitutions that independently occurred more frequently than 95% of all synonymous substitutions excluding the mutagenic context NCN (see Methods). Therefore, we have to conclude that most if not all sites in the SARS-CoV-2 genome that we found to harbor multiple, parallel non-synonymous substitutions not subject to the restrictions discussed above evolve under positive selection(Figs. 1D, Table S3).

Having identified the set of potential positively selected residues, we examined the tree for evidence of epistasis[23] (see Methods) among these sites and revealed a network of putative epistatic interactions (Fig. 1E, Table S4). Strikingly, D614G in the spike protein is associated with exceptionally many interactions and is the main hub of the network. Spike D614G is thought to increase the infectivity of the virus[3], possibly, by increasing the binding affinity between the spike protein and the cell receptor. This high affinity for the receptor might relax selection pressures related to cell entry acting on other regions of the genome and induces positive selection on the sites in this epistatic network. Two non-synonymous mutations linked to spike 614G in this network, S|R21I and S|L54H, are in the spike protein itself though we were unable to validate physical interaction through structural analysis. Another mutation, S|H49Y, less likely to evolve under positive selection but also epistatically linked to S|D614G (Fig. S9) is indirectly supported in the structure(Fig. S10). The majority of the mutations in the epistatic cluster of D614G are located in the non-structural polyprotein (orf1ab) and thus are even less amenable to direct interpretation. Conceivably, 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.

Two adjacent amino acid replacements in the nucleocapsid protein (N): R(agg)203K(aaa) and G(gga)204R(cga) appear simultaneously 7 times. Both sites are likely to evolve under positive selection and are adjacent to yet a third such site, S(agt)202N(aat). 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 that are thought to enhance the transport of the protein to the nucleus[24]. Although positions 202-204 are outside the known nuclear localization signals[25], it appears possible that the substitutions in these sites, in particular G(gga)204R(cga), 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 might illuminate the evolution of SARS-CoV-2 pathogenicity.

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. S9). It is associated with ORF7a Q62*, one of the 6 stop mutations that are observed in at least 10 sequences (Table S5). Stop codon substitutions, apparently, resulting in truncated proteins, occur almost exclusively within the minor SARS-CoV-2 ORFs. The products of ORF8 and ORF7 have been implicated in the modulation of host immunity by SARS-CoV-2, and the strong epistatic connection suggests that the two proteins act in concert. The rest of the connections of S84L are with mutations in orf1ab which, as in the case of D614G, implies uncharacterized functional links between virus-host interactions and virus replication.

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[26]. Because even sites subject to purifying selection[27] 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 diversification relevant to virulence, immune evasion and transmission. Similarly to the case of Influenza, such a diversification process could potentially support a regular pattern of repeat epidemics with grave implications for public health. Strikingly, this is not what we observe.

We first established that sequencing date strongly correlated with tree distance to the root (Fig. S11), indicating a sufficiently low level of noise in the metadata for subsequent analysis. Although examination of the global distribution of each of the 7 major SARS-CoV-2 clades (Figs. S12-13) indicates some regional diversification, this variation is likely to be largely accounted for by time-dependent fluctuations(Fig. 2). Clade 1 is small and was only prevalent early in the year, primarily, within the US, potentially corresponding to sequences descendant from early, limited community spread[28]. Clades 2 and 3, initially dominant, have largely gone extinct, with clade 3 representing only 30% of the sequences from Asia towards the end of June. Clade 6 has been a stable minority throughout the pandemic. Clades 4 and 7 were most prominent in

Europe and the US, respectively, with clade 7 becoming the dominant variant within the US at the height of the April outbreak. Clade 5, growing in prominence throughout the pandemic in Europe, substantially increased in the US as well, and by late June, was poised to become the dominant clade globally.

A comparison of regional clade distributions from the end of April to the beginning of June (Figs. 3A, S14) illustrates the extinction of regionally-dominant early clades and the increasing global prevalence of clade 5. Analysis of the Jenson-Shannon divergence between all pairs of regions (Fig. 3B) shows fluctuations of less than two months in duration and no clear trend towards increased diversity. Normalization by the divergence among triplets of randomized regions, where all sequencing locations are randomly assigned to one of the three regions (Fig. S15), both reduces these fluctuations and demonstrates a clear downward trend (Fig. 3C). Thus, the clade distribution among disparate locations has substantially homogenized relative to expectation over the course of the year. From these observations, it is clear that, despite the dramatically reduced travel[4], Sars-Cov-2 continues to evolve globally. The apparent fitness advantage conferred by the small ensemble of mutations in sites evolving under positive selection, as described here, appears to be sufficient to cause rapid extinction of the less fit variants and to stymie virus diversification. This finding bodes well for a successful vaccination campaign in the midterm.

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).

Figure legends

Figure 1. Evolution of SARS-CoV-2.

A. Global tree reconstruction with 7 principal clades enumerated and color-coded. **B.** Projections of the 3D embedding of the pairwise Hamming distance matrix between SARC-CoV-2 genomes. The clades 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 clade. 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 clade relative to the distribution in that site over all clades. **D.** Site history tree for spike 614. Nodes immediately succeeding a substitution, representing the last common ancestor of at least two substitutions, or terminal nodes are included. Labels correspond to mutations or the tree weight (in mean leaf weight equivalents; see Methods) descendent from that node beyond which no events in the site occur. (Top) Black corresponds to 614D, red to 614G, and green to 614N. **E.** Network of putative epistatic interactions for likely positively selected residues.

Figure 2. Global and regional SARS-CoV-2 clade dynamics during the COVID-19 pandemic. A. Global clade distribution over time. **B.** US clade distribution over time. **C.** European clade distribution over time **D.** Asian clade distribution over time.

Figure 3. Global and regional trends in SARS-CoV-2 evolution. A. Global distribution of sequences with sequencing locations in the US, Europe, and East/Southeast Asia identified. Pie charts indicate the clade distributions for each region mid March through mid April and mid June through mid July. **B.** The Jenson-Shannon divergence between the three pairs of regions. **C.** The mean Jenson-Shannon divergence among the three pairs normalized by the expected divergence between pairs of three randomized regions. Solid line indicates median, shading indicates 25th to 75th percentile.

References

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

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

[3] Korber, Bette, 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)

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

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

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

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

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

[9] Jerzak, Greta, 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)

[10] Wolf, Yuri I., 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)

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

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

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

[14] Lanfear, Rob. *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

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

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

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

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

[19] Mavian, Carla, 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] Sánchez-Pacheco, Santiago J., 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] Pipes, Lenore, et al. Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. *bioRxiv* (2020).

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

[23] Rochman, Nash D., Yuri I. Wolf, and Eugene V. Koonin. Deep phylogeny of cancer drivers and compensatory mutations. *Communications Biology* **3.1**, 1-11 (2020)

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

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

[26] Gong, Lizhi Ian, Marc A. Suchard, and Jesse D. Bloom. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* **2**, e00631 (2013)

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

[28] COVID, CDC, et al. Evidence for Limited Early Spread of COVID-19 Within the United States, January–February 2020. *Morbidity and Mortality Weekly Report* **69.22**, 680 (2020)

Methods

Alignment

All available Sars-Cov-2 genomes as of July 28, 2020 were retrieved from the Genbank[11], Gisaid[12], and CNCB[13] datasets. 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[29,30], masking ambiguous characters. 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 MAFFT[31] with 150 cores. 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 150 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. 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[32] (parameters: -nt -gtr -gamma -nosupport -fastest) to unambiguously define two clusters of sequences: an outgroup consisting of 13 sequences sourced from non-human hosts prior to 2020 as well as sequence GWHABKP00000001 from the CNCB dataset, and the main group. Tree construction requires the resolution of very short branch lengths and it is necessary to compile FastTree at double precision.

The resulting alignment, consisting of 19,327 sequences and 29,119 sites, was maintained for the construction of the global tree and ancestry. 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 significant redundancy and a representative sequence was selected for each cluster of 100% identity to yield an alignment consisting of 15,977 sequences and 6035 sites.

Tree Construction

We sought to optimize tree topology with IQ-TREE[33]; 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, $ceil\big(L/(M-N)\big)$ resulting sets are initialized with this core set where *M* is the desired number of sequences and *L* is the total number of sequences in the alignment (15,977). After this sequences which 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*=1,000 in an effort to optimize computational efficiency and *N*=100. Insufficient overlap greatly affects the results of subsequent steps.

We proceeded to build a tree, using IQ-TREE, for each resulting set fixing the evolutionary model to GTR+F+G4 and decreasing the minimum branch length from the default 10e-6 to 10e-7 following according to the results of previous parameter studies[15]. 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[2]. Briefly, character sets were constructed from leaf to root where each node was assigned the intersection of the descendant character sets if non-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[23.] Briefly, sequences were leaf-weighted based on the branch lengths of the, ultrametrized, tree. Every mutation present across the tree at three mean leaf-weight equivalents of more was considered. The probability of independent co-occurrence between any pair was estimated 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. 1D and Table S3 display putative positively selected mutations with a binomial score above 50 or at least two simultaneous substitutions. Fig. S9 is not restricted to positively selected residues but is restricted to mutations with at least two such pairings.

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[34] 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 substitution events in the codon and high vs low frequency of independent occurrence. For example, consider codon X with 3 nonsynonymous substitution events gat->ggt and 1 nonsynonymous substitution event gat->cgt. In this context, a nonsynonymous nucleotide substitution a->g of frequency 4 would be recorded in nucleotide (X-1)*3+2. The low/high frequency threshold was determined by the 95th percentile of the synonymous mutation frequency distribution (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. Altogether 13,145 mutation events were recorded.

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 (Table S2, sheet 1). This analysis revealed that the context of the high-frequency events (both S and N) is dramatically different from the background frequencies and that 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) differs from the background frequencies in the same direction as that to the high-frequency events, but to a lesser degree. Finally 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.

This analysis was then repeated, this time, distinguishing only between high and low frequency events but not N and S (Table S2, sheet 2) solidifying 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). Furthermore, the log-odds ratios for low-frequency events are strongly correlated with those for high-frequency events (rPearson=0.77), 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.

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, Table S2, sheet 3). This analysis revealed seven contexts where synonymous and non-synonymous events differ significantly. While 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, only 1/4 contexts with an excess of non-synonymous mutations are NCN, suggesting that these non-synonymous events could be driven by other mechanisms. Lastly, 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

Computation of *dN/dS*

For each of the 24 ORFs (nsp11 and nsp12 combined), 10 reduced alignments were constructed as follows. First the core set of maximally diverse sequences selected during constraint tree construction were equally divided (10 sequences for each alignment). Next 10 constraint trees were randomly chosen and the first 40 sequences uniquely incorporated into each constraint tree were added ensuring a diverse set of 50 unique sequences for each reduced alignment. The reference sequence, NC_045512.2, was additionally added to each reduced alignment. PAML[35] 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, nN, nS, tN, and tS were retrieved for each segment being the total number of nonsynonymous and synonymous substitutions as well as these tallies normalized by the respective segment length. 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.

Supplemental Figure Captions

Figure S1. 25th, median, 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)). Here the top 5% of codons with the most nucleotide exchanges in each window are ignored.

Figure S5. Same as Fig. S7 where no codons are excluded. The trends are qualitatively similar indicating outliers do not play an outsized role.

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.

Figure S7. A. The square root of 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 sqrt(2/1.3*x)

reflecting the genome-wide ratio of nonsynonymous to synonymous substitutions, solid lines are sqrt(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. Epistatic network for the tree including mutations with a binomial score above 50 or at least two simultaneous substitutions not restricted to likely positively selected residues. Only nodes of degree 2 or greater are displayed.

Figure S10. 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.

Figures S11. Correlation between sequencing date and tree distance to the root.

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. Clade distributions for each region at two fixed timepoints, mid March to mid April and mid June to mid July, as well as the difference.

Figure S15. Jenson-Shannon divergence between pairs of three randomized regions, where all sequencing locations are randomly assigned to one of the three regions. 25th, 50th, and 75th percentiles shown over 1000 replicates.

Supplemental Tables

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

Table S2. A validation of the genome-wide mutagenic context NCN.

Table S2. All epistatic interactions among states meeting the criteria outlined in the main text for likely positive selection with a binomial score greater than 50 or at least 2 simultaneous substitutions. Each pair is arbitrarily ordered and the numbers of simultaneous, descendant, and independent substitutions are tabulated.

Table S3. 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.

Supplemental References:

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

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

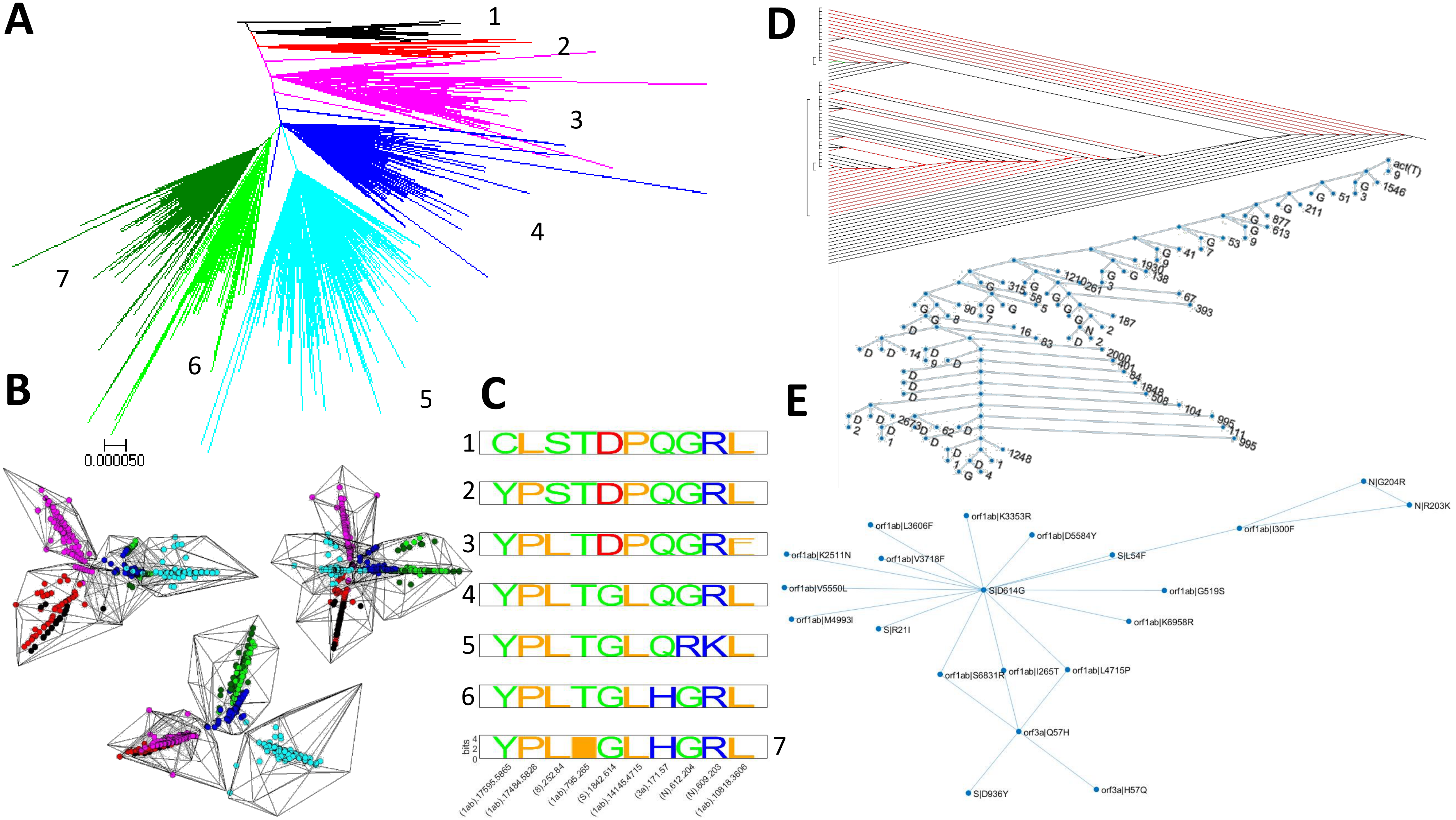
[31] Katoh, Kazutaka, et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research* **30.14**, 3059-3066 (2002)

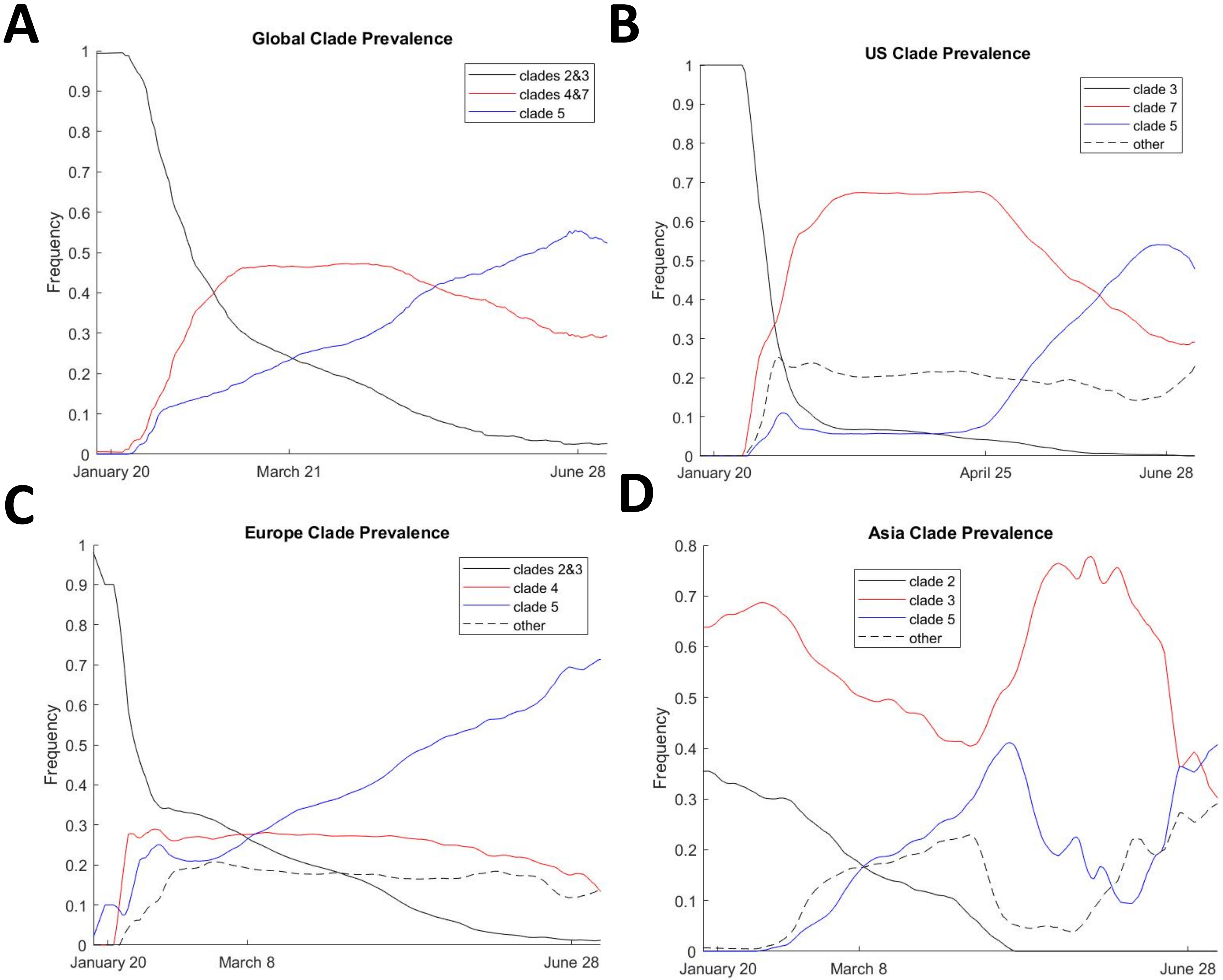
[32] Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **5.3**, e9490 (2010)

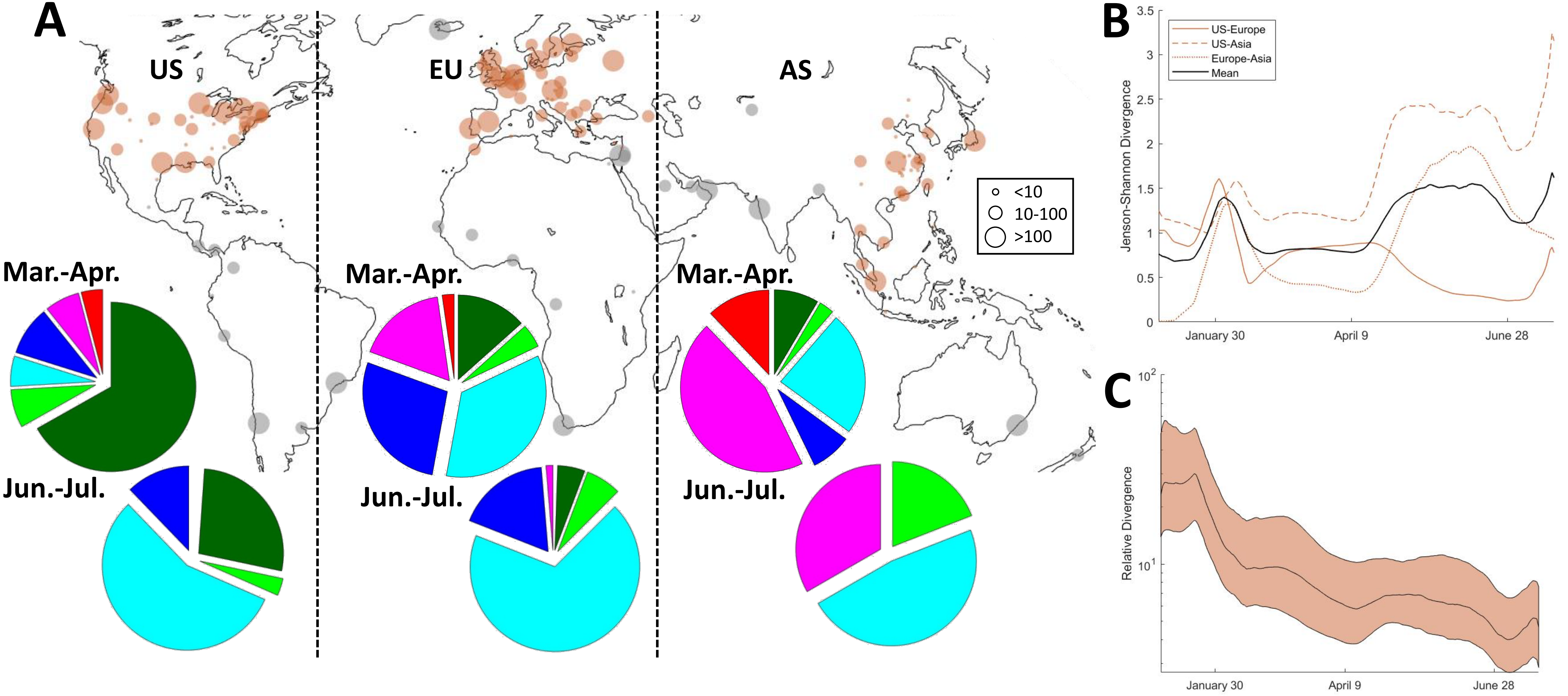
[33] Nguyen, Lam-Tung, et al. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* **32.1**, 268-274 (2015)

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

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


A
1
2
3
4
5
6
7
0.000050
B
C
1 CLSTDPQGRL
2 YPSTDPQGRL
3 YPLTDPQGRF
4 YPLTGLQGRL
5 YPLTGLQRKL
6 YPLTGLHGRL
7 YPL GLHGRL
bits
4
2
0
(1ab).17595.5865
(1ab).17484.5828
(8).252.84
(1ab).795.265
(S).1842.614
(1ab).14145.4715
(3a).171.57
(N).612.204
(N).609.203
(1ab).10818.3606
D
act(T)
9
1546
G
3
51
G
211
G
877
G
613
G
9
53
G
7
41
G
9
1930
138
G
G
3
1210
261
67
393
315
G
58
5
G
G
G
187
2
N
2
D
90
G
7
G
G
G
8
16
83
D
D
D
14
D
9
D
2000
401
84
1848
508
D
D
D
104
995
111
995
D
2
D
D
1
2673
D
62
D
D
1248
D
1
D
1
G
4
E
N|G204R
N|R203K
orf1ab|I300F
orf1ab|K3353R
orf1ab|L3606F
orf1ab|D5584Y
orf1ab|K2511N
orf1ab|V3718F
S|L54F
orf1ab|V5550L
S|D614G
orf1ab|G519S
orf1ab|M4993I
S|R21I
orf1ab|K6958R
orf1ab|S6831R
orf1ab|I265T
orf1ab|L4715P
orf3a|Q57H
S|D936Y
orf3a|H57Q


A
Global Clade Prevalence
Frequency
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
January 20
March 21
June 28
clades 2&3
clades 4&7
clade 5
B
US Clade Prevalence
Frequency
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
January 20
April 25
June 28
clade 3
clade 7
clade 5
other
C
Europe Clade Prevalence
Frequency
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
January 20
March 8
June 28
clades 2&3
clade 4
clade 5
other
D
Asia Clade Prevalence
Frequency
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
January 20
March 8
June 28
clade 2
clade 3
clade 5
other


A
US
EU
AS
Mar.-Apr.
Jun.-Jul.
Mar.-Apr.
Jun.-Jul.
Mar.-Apr.
Jun.-Jul.
<10
10-100
>100
B
US-Europe
US-Asia
Europe-Asia
Mean
Jenson-Shannon Divergence
3.5
3
2.5
2
1.5
1
0.5
0
January 30
April 9
June 28
C
10^2
10^1
Relative Divergence
January 30
April 9
June 28