September 27, 2020

An evolutionary portrait of the progenitor SARS-CoV-2 and its dominant offshoots in COVID-19 pandemic

Sudhir Kumar[1,2,*], Qiqing Tao[1,2], Steven Weaver[1,2], Maxwell Sanderford[1,2], Marcos A. Caraballo-Ortiz[1,2], Sudip Sharma[1,2], Sergei L. K. Pond[1,2,*], and Sayaka Miura[1,2,*]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA.

[2]Department of Biology, Temple University, Philadelphia, PA.

*Co-corresponding author:

Sudhir Kumar (s.kumar@temple.edu)

Sergei Pond (spond@temple.edu)

Sayaka Miura (sayaka.miura@temple.edu)

Abstract:

Severe acute respiratory syndrome coronavirus 2, SARS-CoV-2, was quickly identified as the cause of COVID-19 disease soon after its earliest reports. Global sequencing of thousands of genomes has revealed many common genetic variants, which are the key to unraveling the early evolutionary history of SARS-CoV-2 and tracking its global spread over time. However, our knowledge of fundamental events in the genome evolution and spread of this coronavirus remains grossly incomplete and highly uncertain. A deep understanding of the contemporary evolution of SARS-CoV-2 is urgently needed not only for a retrospective on how, when, and why COVID-19 has emerged and spread, but also for creating remedies through efforts of science, technology, medicine, and public policy. Here, we present the heretofore cryptic mutational history, phylogeny, and dynamics of SARS-CoV-2 from an analysis of tens of thousands of high-quality genomes. The reconstructed mutational progression is highly concordant with the timing of coronavirus sampling dates. It predicts the genome sequence of the progenitor virus whose earliest offspring, without any non-synonymous mutations, were still spreading worldwide months after the report of COVID-19. Over time, mutations of the progenitor gave rise to seven dominant lineages that spread episodically over time, some of which likely arose in Europe and North America after the genesis of the ancestral lineages in China. Mutational barcoding establishes that North American coronaviruses harbor genome signatures different from coronaviruses prevalent in Europe and Asia, which have converged over time. These spatiotemporal patterns continue to evolve as the pandemic progresses and can be viewed live online.

The early evolutionary history and order of mutational events that arose during the pandemic remain unresolved, even months after the initial detection of SARS-CoV-2 as the causal agent of COVID-19 and the acquisition of tens of thousands of genomes[1–8]. Widely-recognized impediments include a limited number of phylogenetically informative variants in genomes, the ubiquity of sequencing errors, and the lack of a closely-related outgroup sequence, all of which have complicated the inference and rooting of the SARS-CoV-2 phylogeny[1–8]. Consequently, the traditional approach to the analysis of viral spread and evolution in which a reliable genome phylogeny is first inferred, and then observed differences among sequences are mapped site-by-site, has not been able to stage the earliest mutational events in the evolution of the novel coronavirus[9,10]. The definition and cataloging of viral lineages are similarly complicated by phylogenetic noise[7].

To circumvent these issues, we applied a mutation order approach (MOA) that does not rely on the inference of phylogeny as an intermediate step in reconstructing the mutational history of genomes[11–13]. MOA is well-suited for analyzing SARS-CoV-2 genomes because its clonal evolution, without evidence for recombination in the early stages of the outbreak, preserves the collinearity of variants in genomes. This feature enables effective use of shared co-occurrence of variants in genomes, as well as the frequencies of individual variants, to reliably infer mutational history, and then genome phylogeny, notwithstanding the presence of sequencing errors and other artifacts[11,14] (see *Methods*).

A mutational history of SARS-CoV-2

We analyzed 29,681 SARS-CoV-2 genomes (herewith referred to as 29KG dataset), each with at least 28,000 bases, sampled between 24 December 2019 and 07 July 2020, representing 97 countries and regions around the world. In the 29KG dataset, 49 single nucleotide variants (SNVs) occur with greater than 1% variant frequency (*vf* > 1%; Extended Data Table 1). We used MOA to reconstruct the history of mutations that gave rise to these common variants, presented in figure 1. The first mutations (μ's, α's, and β's) were all sampled early on in Asia (China) and have the highest frequency in the 29KG dataset (Fig. 1, red pies). In this mutation history, over 95% of the variants showed an extremely high co-occurrence, i.e., each variant was found in the genomic background of the variant preceding it in the graph. The average co-occurrence index, COI, for all 49 variants exceeds 96.9%, indicating a strong signal to infer the mutation history reliably.

The inferred mutation order agreed with the timing of the first sampling for variants in all but two cases (Fig. 1, see *Methods*). This concordance provides independent validation of the reconstructed mutation graph because neither sampling dates nor locations were used in the inference of mutational history. Such independence is the key to avoiding circularity in the subsequent analysis of the origin and spread of new mutants. For example, some early genome samples in China have been used as references in databases and analyses to orient mutational changes[15–18]. This practice assumes that reference genomes are ancestral. The mutation history in figure 1, reconstructed independently of any spatiotemporal information, permits a direct test of this assumption (see below).

56 *The progenitor SARS-CoV-2 genome.* MOA predicts the genome sequence of the most recent common
57 ancestor of all the genomes analyzed, i.e., the progenitor SARS-CoV-2 genome, henceforth proCoV2
58 (available at http://igem.temple.edu/COVID-19). In the proCoV2 genome, there are 170 non-
59 synonymous and 958 synonymous substitutions when compared with the genome of a closely-related
60 coronavirus, RaTG13, found in a *Rhinolophus affinis* bat[19] (Fig. 2a). This amounts to a 96.12% sequence
61 similarity between proCoV2 and RaTG13 sequences.

62 A comparison of the inferred proCoV2 sequence with genomes in the 29KG collection revealed no full
63 matches at the nucleotide level. However, complete amino acid sequence encoded by 120 genomes
64 were identical to that of proCoV2, i.e., these genomes contained only synonymous differences from
65 proCoV2. A majority (80 genomes) of these matches were from coronaviruses sampled in China and
66 other Asian countries. The first sequence was sampled 12 days after the earliest sampled virus whose
67 genome became available on 24 December 2019, referred to as pandemic day 0 (week 0). Multiple
68 matches were found in all sampled continents and detected as late as April 2020 (pandemic day 124)
69 in Europe (Fig. 2c). These spatiotemporal patterns suggest that proCoV2 already possessed the
70 repertoire of protein sequences needed to infect, spread, and persist in the global human population
71 (see also ref.[20]).

72 *Mutations of proCoV2 before the first COVID-19 reports.* In the mutation history graph, the first three
73 synonymous SNVs (μ) of proCoV2 are present individually in more than 98% of the genomes in the 29KG
74 collection. They have almost reached fixation in the global coronavirus population (Fig. 3a). All three
75 SNVs almost always co-occur and were present in all the genomes sequenced in China in December
76 2019 (Fig. 1). The proCoV2 genome containing these three μ variants gave rise to an evolutionary
77 lineage characterized by three additional variants (α), all of which were found together in 90% of the
78 genomes in 29KG and emerged before pandemic day 0. The first two α mutations were synonymous,
79 and the third one was a non-synonymous mutation in ORF8 (Fig. 2a; Extended Data Table 1). Out of a
80 total of six initial mutations of proCoV2 (μ's and α's), the first five were synonymous U→C changes,
81 whose functional significance is being debated[20,21], as they are the most common early nucleotide
82 changes observed in the evolution of SARS-CoV-2[22].

83 The emergence of μ and α variants before the first reports of COVID-19 implies the existence of
84 significant sequence diversity in the ancestral SARS-CoV-2 populations. All the 17 genomes from China
85 sampled in December 2019 carry all three μ and three α variants (Fig. 2d). Interestingly, the six genomes
86 containing μ variants but not α variants were sampled in China and the United States in January 2020
87 (Fig. 2d). Therefore, the earliest sampled genomes (including the designated database references) are
88 not ancestral, and using them as such[23–25] will spuriously predict several reversals or convergent
89 mutations in the early history of SARS-CoV-2 genomes. For example, the root of SARS-CoV-2 phylogeny
90 in the GISAID resource is placed between α_1 and α_3 variants in the mutational history[26]. This placement
91 of root suggests U to be the mutant base and C to be the ancestral base in SARS-CoV-2 at all four
92 positions (three μ's and α_1). This assignment requires four consecutive backward mutations in SARS-

93 CoV-2 or convergent mutations in the bat coronavirus because U is observed at all four positions in
94 RaTG13. This many backward and convergent mutations in significant frequency would be unusual. In
95 fact, the likelihood of mutation history constructed using C as the ancestral base was significantly worse
96 than with U as the ancestral base ($P << 0.01$; see also *Methods*). Our inference is also supported by the
97 observation that a mutant of proCoV2 containing μ variants, but no α variants, was isolated in the
98 United States on pandemic day 59. This means that the root of the SARS-CoV-2 phylogeny is on a lineage
99 before α_1 appeared, which is inconsistent with the previous inference[26].

100 The proCoV2μ genome experienced two additional non-synonymous variants (ν) and became a
101 significant CoV-2 lineage before going extinct a few weeks later (Fig. 1, 3b, and 3f). Therefore, proCoV2
102 genomes with μ variants alone were distributed worldwide and gave rise to new CoV-2 lineages during
103 the pandemic. The discovery of these patterns was possible by mutation order analysis and the use of
104 proCoV2 as a reference, which enables improved molecular evolutionary and phylogenetic analyses of
105 COVID-19.

106 *Mutations after the first reports of COVID-19.* A non-synonymous variant of the Spike protein (p.D614G,
107 β_2) of proCoV2μα genome and two other variants, one synonymous (β_1) and one non-synonymous (β_3)
108 variant of ORF 1ab, co-occur in ~77% of the genomes (β variants). The first two of these variants were
109 detected on pandemic day 31 in China and day 35 in Europe. These two variants are almost always
110 sampled together (Fig. 1). The β_3 mutant was found soon after to complete form the β lineage. It was
111 first sampled in Saudi Arabia on pandemic day 41 and then in Europe on pandemic day 54.

112 Interestingly, β_3 was not observed in Chinese samples until pandemic day 69. By that time, the β lineage
113 had already been established worldwide. As a result of highly uneven and sparse sampling of genomes
114 across the globe during the early phases of the pandemic (Fig. 2e), one cannot exclude the possibility
115 that β variants first arose outside China, possibly the Middle East or Europe. Regardless, the three β
116 variants preceded the coronavirus expansion in Europe two months after the pandemic began (Fig. 3b
117 and 3e). Because of its co-occurrence with the Spike variant p.D614G, the role of the non-synonymous
118 ORF1ab β_3 mutant is likely to become attractive to investigate experimentally, as several studies have
119 argued that the Spike variant increases the infectivity and facilitates the spread of COVID-19, albeit
120 without detectable clinical consequences[27–34].

121 The frequency of β variants has approached fixation in the SARS-CoV-2 population worldwide (Fig. 3B).
122 However, this increase has occurred concomitantly with, or possibly driven by, their descendant
123 lineages (Fig. 3c). One lineage from proCoV2μαβ was founded by three mutations (ε) of the
124 nucleocapsid protein (N) protein (Fig. 1). It was first seen in genomes sampled on pandemic day 54 in
125 Europe. This lineage (proCoV2μαβε) has become dominant in the European region, replacing most
126 other lineages (Fig. 3e). All three ε mutants occur in adjacent codons of the N protein and are non-
127 synonymous changes involving the Arginine residues. The net result of the three ε mutations is the gain
128 of an Arginine, which increases the positive charge of the protein, an essential property for the N
129 protein's nucleic acid binding function that is critical for virus transcription and assembly[35]. Because all

130 three ε variants are always found together and are increasing in frequency in Asia and Europe, they
131 may have hitchhiked with or driven the increase in the frequency of the Spike p.D614G variant.

132 The second evolutionary lineage to emanate from proCoV2μαβ was founded by a mutation (γ) that,
133 along with its descendant genome containing a δ mutation, gave rise to the most common
134 coronaviruses in North America today (Fig. 3f). The γ variant was first detected on day 41 in Saudi Arabia
135 (98 samples). The δ variant was first found in Singapore on pandemic day 54 (25 samples). They both
136 appeared in Europe on pandemic day 59, underwent a limited spread for a few weeks, but did not
137 become common (Fig. 3e). Because neither of these variants was detected in China until later
138 (pandemic day 79), they may have originated in other regions of the world.

139 The presence of the γ and δ variants in and around Europe by pandemic day 59 and their subsequent
140 appearance in eastern North America (first detected on pandemic day 66) is consistent with the
141 suggestion that, early on, coronaviruses from Europe seeded infections in eastern North America[36]. As
142 for the Asian seeding of infections in western North America[37], μ and μα originated in China and
143 dominated in the earliest phases of the pandemic. However, these mutants were replaced a few weeks
144 later by μν genomes that were first isolated on pandemic day 59 in western North America and those
145 found in eastern North America (μαβγδ genomes) (Fig. 3f). In recent months, the spatiotemporal
146 pattern in North America has been led by the expansion of proCoV2μαβγδ genomes (Fig. 3f).

147 Overall, Asian spatiotemporal patterns of coronavirus lineages are more similar to European regions,
148 where μαβ genomes with ε variants have become common (Fig. 3d). These patterns are very different
149 from North American regions where ε remains a minority and μαβ genomes with γ and δ variants
150 dominate (Fig. 3f). Therefore, spatiotemporal patterns have converged between Europe and Asia, both
151 of which have diverged from North America.

152 Molecular phylogeny of SARS-CoV-2 genomes

153 The progression of mutations shown in figure 1 predicts observed and extinct genomes by tracking
154 mutational history from the progenitor sequence. Therefore, it directly transforms into a phylogeny of
155 genomes; each node (leaf and internal) represents a genome type containing all the mutations on the
156 path from that node to the progenitor proCoV2 (Fig. 2a). This phylogeny is a rooted tree of SARS-CoV-
157 2, which has been challenging to infer reliably[1–8]. It shows that all the early novel coronavirus lineages
158 were established by three μ mutations, all of which were first seen in genomes sampled in China. In
159 fact, the next five major mutations ($\alpha_1 - \alpha_3$ and $\beta_1 - \beta_2$) were also detected first in China, indicating that
160 the proCoV2 originated and evolved in China.

161 Genomes containing early mutations spread to other parts of the world, and they gave rise to new
162 lineages much later. For example, a genome with three μ SNVs mutated to establish the ν lineage that
163 likely arose many weeks after the pandemic began (sampled on day 59), spread extensively (4.7% of
164 the genomes in 29KG), and remained restricted to North America. A mutant of the genome containing
165 μ and α_1 SNVs gave rise to a significant offshoot (α_{1a}) more than two months after the pandemic began

166 (pandemic day 63 in France). The genome harboring μ and α SNVs underwent three successive
167 mutations, including Spike p.D614G, to establish the β lineage. This lineage gave rise to two major
168 clades: one founded by γ and δ mutations that became predominant in North America. The other was
169 founded by ε mutations that became predominant in Europe and Asia. Many of these major
170 divergences in the early evolutionary history of SARS-CoV-2 have not been apparent from previous work
171 due to difficulty in rooting and interpreting SARS-CoV-2 phylogenies[1–7].

172 Mutational barcodes for coronavirus genome classification. Each node in the mutation-based
173 phylogeny in figure 2A can be represented by a binary barcode containing the presence/absence of 49
174 variants–a system that provides an intuitive mutational signature approach to classifying genomes (see
175 *Method*s). We mapped genomes from the 29KG dataset to each node based on these signatures and
176 represented high and low genome counts by open and closed circles and triangles (Fig. 2a). A
177 comparison of our mutation-based classification with a widely used phylogeny-based classification[26]
178 revealed many similarities (see Extended Data Figure 1). However, several significant differences
179 between classification schemes arose, likely caused by the use of early coronavirus samples to root
180 SARS-CoV-2 phylogeny and the difficulty in reliably staging the earliest divergence events.

181 Episodic accumulation of variants. The numbers of genomes mapping to many ancestral nodes in the
182 SARS-CoV-2 phylogeny are relatively small (represented by open circles and triangles in Fig. 2a). In
183 particular, the number of genomes with both α_2 and α_3 variants is more than 20-times more than those
184 containing only the α_2 variant (1,116 versus 47 genomes). Similarly, the number of genomes with all
185 three β variants (3,032) exceeds those with predecessor β_1 and β_2 mutants (40 and 9 genomes,
186 respectively). The three ε variants always occur together (5,365 genomes), with other combinations of
187 ε variants found in only two genomes (Extended Data Table 1). Similar trends are observed for many
188 other offshoots, as well. This clustering of mutants suggests an episodic nature of variants spread ($P <$
189 0.01, *Methods*), which may arise from founder effects, positive selection, or both (e.g., ref.[20]). However,
190 such inferences of evolutionary and epidemiological mechanisms are complicated by highly uneven
191 regional and temporal genome sequencing that is unlikely to produce an unbiased representative
192 sample of the actual worldwide population (Fig. 2e).

193 In this phylogeny, the proportion of non-synonymous to synonymous changes (N/S) is 1.9. This ratio is
194 almost 10-times larger than the ratio of 0.18 for the inferred proCoV2 and observed Bat CoV proteins.
195 A McDonald-Kreitman test[38] rejects the similarity of molecular evolutionary patterns observed within
196 the SARS-CoV-2 population (29KG dataset) and between human proCoV2 and the bat coronavirus. It is
197 not prudent to automatically invoke the action of positive selection using such neutrality tests, because
198 synonymous polymorphisms in SARS-CoV-2 genomes are affected by molecular mechanisms (e.g., RNA
199 editing)[21,39] as well as negative selection[21]. Also, even slightly deleterious alleles can become common
200 when there is a population expansion[40]. Furthermore, we cannot assume that all non-synonymous and
201 synonymous differences between human CoV-2 and bat CoV sequences are neutral[20]. Nevertheless,
202 N/S patterns do show that molecular evolutionary patterns observed within SARS-CoV-2 genomes

203 infecting humans are different from those spanning the divergence between the bat RaTG13 and
204 proCoV2 strains.

205 In conclusion, the approach taken here to discover key mutational events, a timeline of their evolution,
206 and spatial distributions of variants and evolutionary lineages will generally be applicable for analyzing
207 pathogens during the early stages of outbreaks. The approach is scalable for even bigger datasets
208 because it does not require more phylogenetically informative variants with an increasing number of
209 samples. In fact, it benefits from bigger datasets as they afford more accurate estimates of individual
210 and co-occurrence frequencies of variants and enable more reliable detection of lower frequency
211 variants. Its application to an extensive collection of SARS-CoV-2 genomes has facilitated the
212 reconstruction of the progenitor viral genome and the identification of mutant lineages, empowering
213 the tracking of distinct SARS-CoV-2 lineages over time and space, improving our understanding of the
214 past, current, and future evolution of SARS-CoV-2 and COVID-19. An initial implementation of a
215 regularly updated SARS-CoV-2 phylogeny and global spatiotemporal patterns utilizing GISAID data is
216 available from http://igem.temple.edu/COVID-19 (an early beta version).

217

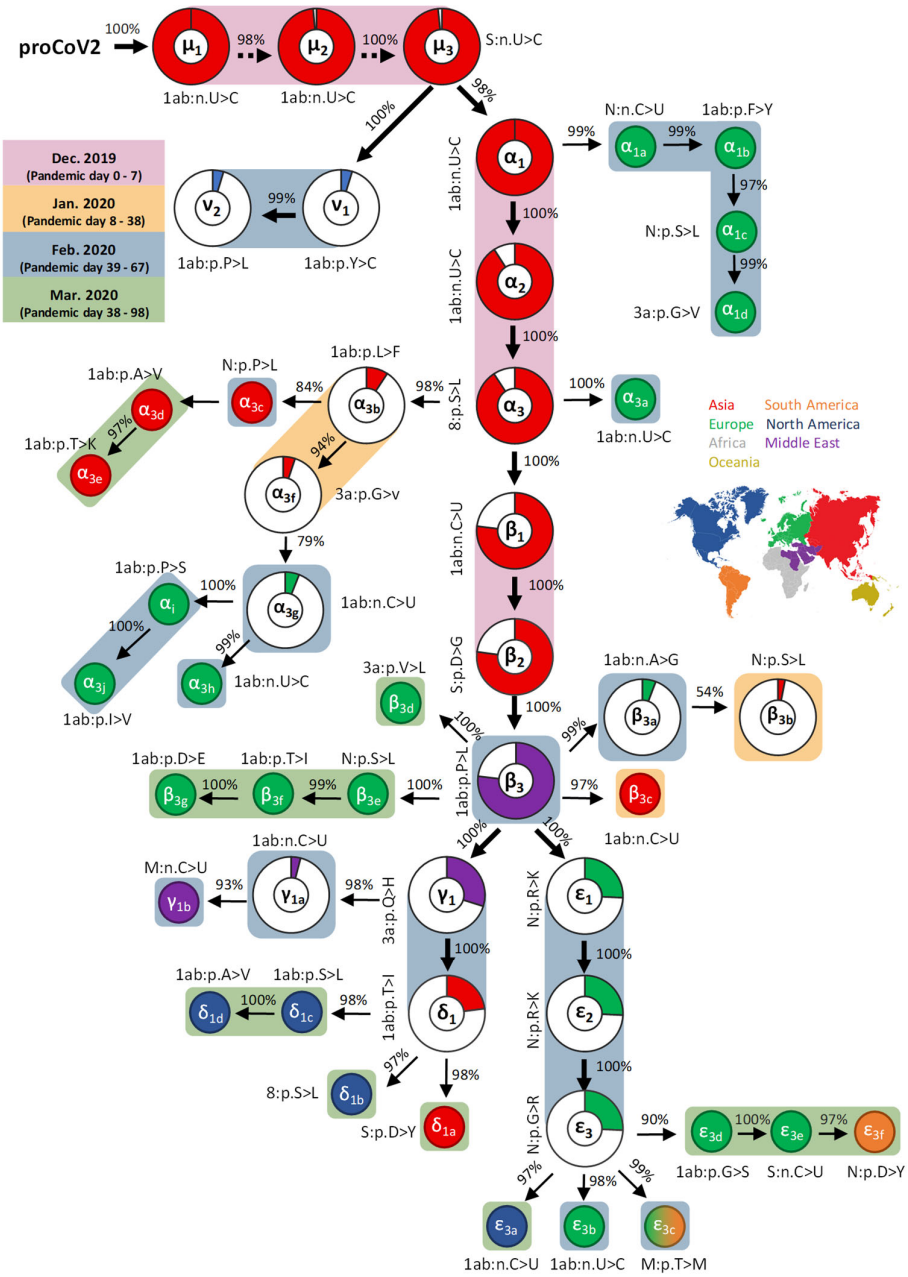

Fig. 1. Mutational history graph of SARS-CoV-2. Thick arrows mark the pathway of widespread variants (frequency, *vf* ≥ 5%), and thin arrows show paths leading to other common mutations (5% > *vf* > 1%). The size of the pie in pie-charts is proportional to variant frequency in the 29KG dataset, with pie-charts shown for variants with *vf* > 3% and pie color based on the region of the world where that mutation was first observed. A circle is used for all other variants, with the filled color corresponding to the earliest sampling region. The co-occurrence index of each mutation and its predecessor mutation is shown next to the arrow connecting them. Base changes (n.) are shown for synonymous mutations, and amino acid changes (p.) are shown for non-synonymous mutations along with the gene/protein names ("ORF" is omitted from gene name abbreviations given in Extended Data Table 1). A rounded rectangular background indicates the earliest month in which a mutation was first found. More details on each mutation are presented in Extended Data Table 1.

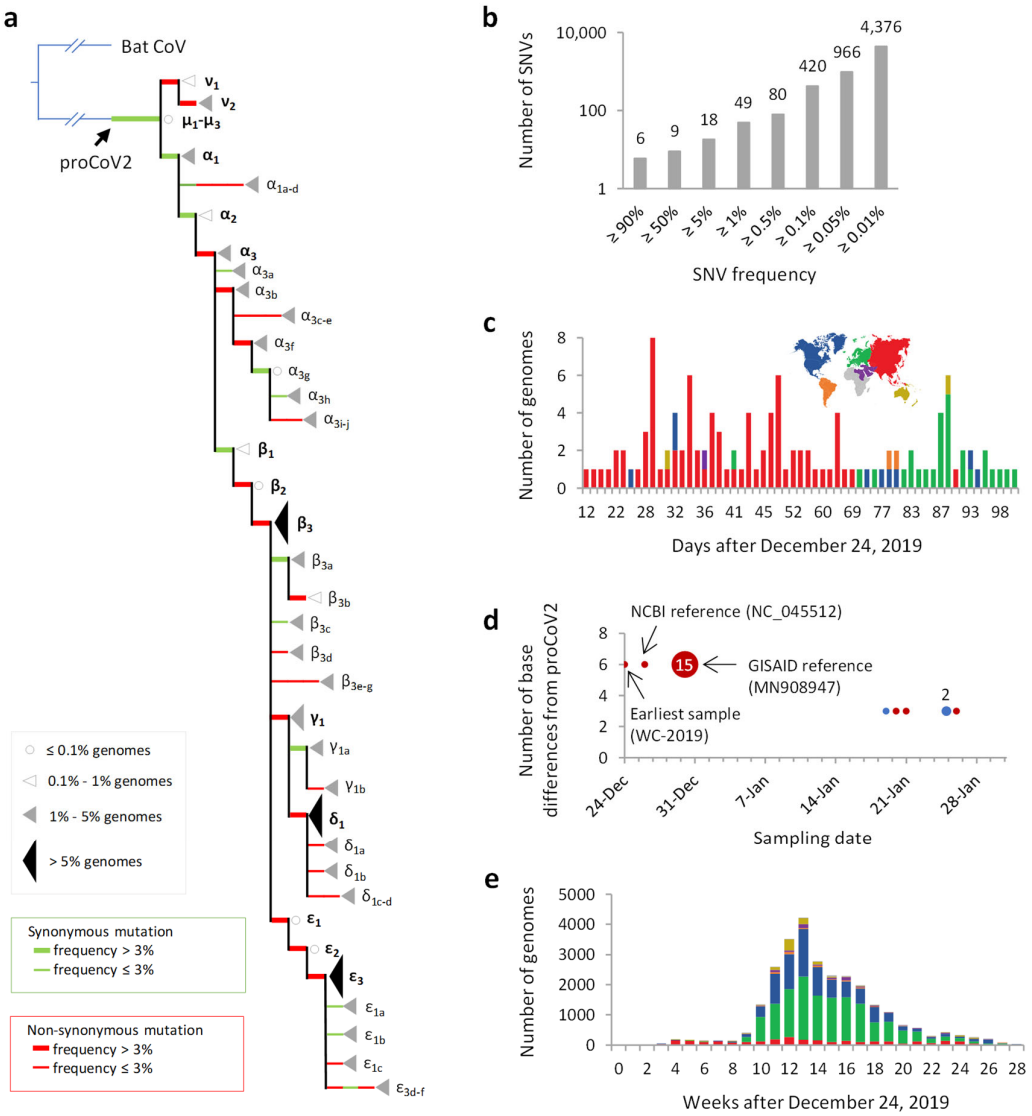

Fig. 2. Evolutionary divergence and history of SARS-CoV-2. (a) A waterfall display of genome phylogeny recapitulating the mutation graph in figure 1. The numbers of genomes mapped to each node are depicted by open circles (very few genomes), open triangles (few genomes), small gray triangles (many genomes), and large black triangles (very many genomes). The tip label is the name of the mutation on the connecting branch. Green and red branches are synonymous and non-synonymous mutations, respectively. Thick branches mark mutations that occur with a frequency greater than 3% in the 29KG data. (b) Cumulative count of single nucleotide variants present in the 29KG genome dataset at different frequencies. (c) Temporal and spatial distribution of strains identical to proCoV2 on amino acid sequence level, i.e., they have only synonymous differences. (d) The number of base differences from proCoV2 for genomes that were sampled in December 2019 and January 2020. The 17 genomes sampled in December 2019 in China (red) have six common SNVs different from proCoV2, whereas six genomes sampled in January 2020 in China (Asia, red) and USA (North America, blue) show only three bases differences. A circle marks the presence of genome, with multiple genomes (2 and 15) sampled on two different days. (e) The number of genomes in the 29KG collection that were isolated weekly during the pandemic. The color scheme used to mark sampling locations is shown in panel c.

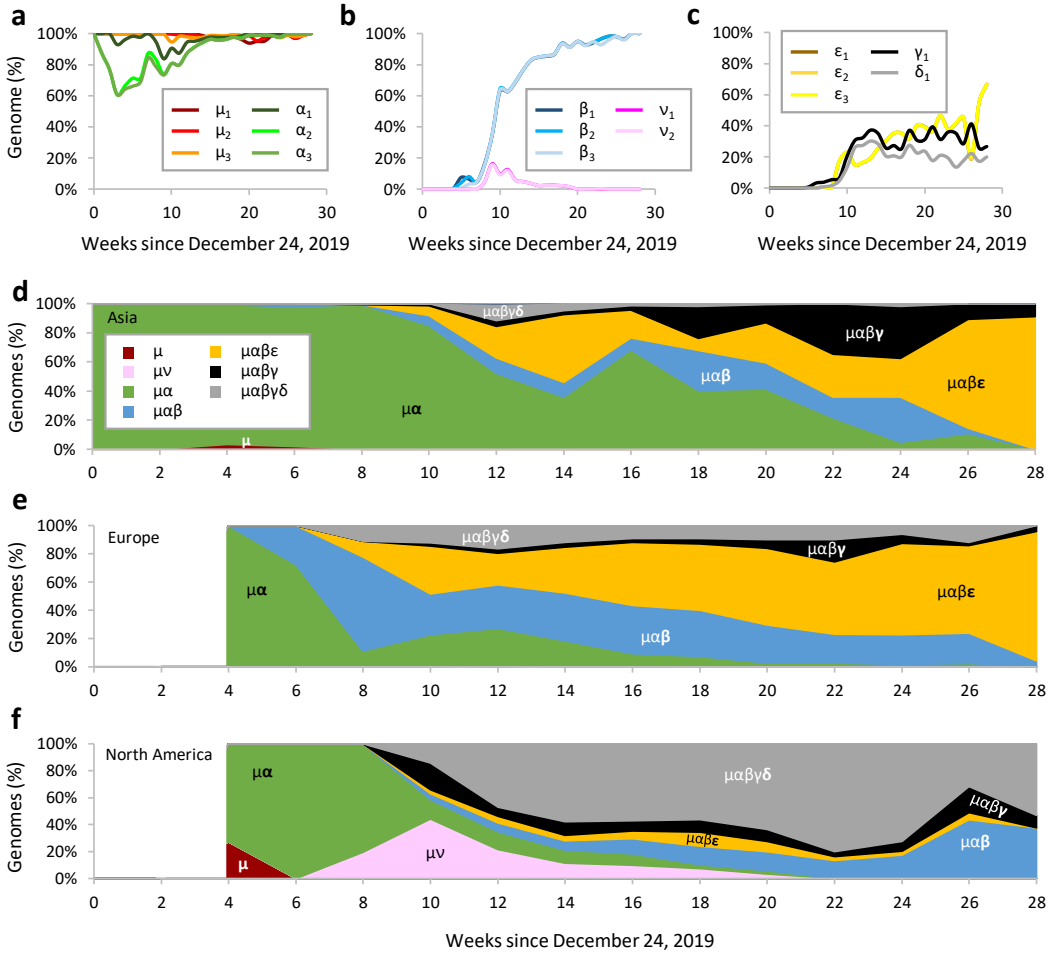

Fig. 3. Spatiotemporal dynamics of SARS-CoV-2 genomes. (a-c) Temporal patterns of frequencies of major variants. The proportion of genomes containing each major mutant is calculated for each week on and after 24 December 2019 and then connected using a smoothed line. (d-f) Spatiotemporal patterns of genomes mapped to lineages containing different combinations of major variants in Asia, Europe, and North America. The number of genomes mapped to mutation lineages (e.g., proCoV2μα contains $\mu_1 - \mu_3$ and $\alpha_1 - \alpha_3$ variants) is counted biweekly to generate this stacked line graph. The area is the proportion of genomes mapped to the corresponding lineage in the two-week time frame. To simplify the display, "proCoV2" is omitted from the lineage names. There was not a sufficient number of genomes sampled from Europe and North America in the first four weeks.

258 ## Methods

259 Genome data acquisition and processing

260 We downloaded SARS-CoV-2 genomes from the GISAID[41] database along with information on sample
261 collection dates and locations. Of all the genomes downloaded, we only retained those with greater
262 than 28,000 bases and marked as originating from human hosts. Each genome was subjected to codon-
263 aware alignment with the NCBI reference genome (accession number NC_045512) and then subdivided
264 into ten regions based on CDS features: ORF1a (including nsp10), ORF1b (starting with nsp12), S, ORF3a,
265 E, M, ORF6, ORF7a, ORF8, N, and ORF10. Gene ORF7b was removed because it was too short for
266 alignment and comparisons. For each region, we scanned and discarded sequences that contained too
267 many ambiguous nucleotides in order to remove data with too many sequencing errors. Thresholds
268 were 0.5% for the S gene, 0.1% for ORF1a and ORF1b genes, and 1% for all other genes. We mapped
269 individual sequences to the NCBI reference genome (NC_045512) using a codon-aware extension to
270 the Smith-Waterman algorithm implemented in HyPhy[42] (https://github.com/veg/hyphy-
271 analyses/tree/master/codon-msa), translated mapped sequence to amino-acids and performed
272 multiple protein sequence alignment with the ----auto settings function of MAFFT (version 7.453)[43].
273 Codon sequences were next mapped onto the amino-aid alignment. The multiple sequence alignment
274 of SARS-CoV-2 genomes was aligned with the sequence of the coronavirus genome of the *Rhinolophus*
275 *affinis* bat (RaTG13)[19] and visually inspected and adjusted in MEGA X[44,45]. Ultimately, the final alignment
276 contained all genomic regions except ORF7b and non-coding regions (5' and 3' UTRs, and intergenic
277 spacers). After these filtering and alignment steps, the multiple sequence alignment contained 29,115
278 sites and 29,681 SARS-CoV-2 genomes, which we refer to as the 29KG dataset.

279 Reference genomes and collection dates

280 We used the dates of viral collections provided by the GISAID database[41] in all our analyses. All genomes
281 were used in the mutation ordering analyses, but genomes with incomplete sampling dates were
282 excluded from the spatiotemporal analyses and derived interpretations. We noted that the earliest
283 sample included in GISAID, (ID: EPI_ISL_402123), was collected on 24 December 2019, although the
284 NCBI website lists its collection date as 23 December 2019 (GenBank ID: MT019529). Therefore, we
285 used the GISAID collection date for the sake of consistency. Regarding the NCBI reference genome
286 (GenBank ID: NC_045512)[46], this sample was collected on 26 December 2019[47]. The collection date for
287 the GISAID reference genome (GenBank ID: MN908947) is listed as 31 December 2019[48], although the
288 patient was reported sick and admitted into the hospital on 26 December 2019 [46]. Thus, in our analyses,
289 we used the GISAID collection date of 31 December 2019 for the GISAID reference genome to be
290 consistent with previous studies.

291 Mutation order analyses (MOA)

292 We used a maximum likelihood method, SCITE [11], and variant co-occurrence analysis for reconstructing
293 the order of mutations corresponding to 49 common variants (frequency > 1%) in the 29KG dataset.

294 MOA has demonstrated high accuracy for analyzing tumor cell genomes that reproduce clonally, have
295 sequencing errors, and exhibit limited sequence divergence[11,12]. In MOA, higher frequency variants are
296 expected to have arisen earlier than low-frequency variants in clonally reproducing populations[11,14].
297 Using the highest frequency variants to anchor the analysis and the shared co-occurrence of variants
298 among genomes to order mutations, while allowing probabilistically for sequencing errors and pooled
299 sequencing of genomes[11], we evaluated and compared the likelihood of various possible mutational
300 histories. MOA is different from traditional phylogenetic approaches where positions are treated
301 independently, i.e., the shared co-occurrence of variants is not directly utilized in the inference
302 procedure. Notably, both traditional phylogenetic and mutation order analyses are expected to
303 produce concordant patterns when sequencing errors and other artifacts are minimized. However,
304 sequencing errors and limited mutational input during the coronavirus history adversely impact
305 traditional methods[1,4,7], as does the fact that the closest coronaviruses useable as outgroups have more
306 than a thousand differences from SARS-CoV-2 genomes that only differ in a handful of bases from each
307 other[1,4].

308 MOA requires a binary matrix of presence/absence of mutant for 29KG. We first designated differences
309 from the bat RaTG13 genome by "1," otherwise a "0" was assigned. This was simply an initial seed for
310 the analysis, as subsequent iterations coded RaTG13 bases as the mutant ("1") for each position
311 individually and selected the coding that produced mutation graphs with the highest maximum
312 likelihood and average shared co-occurrence index (COI) of variants. COI for a given variant (*y*) is the
313 number of genomes that contain *y* and its directly preceding mutation (*x*) divided by the number of
314 genomes that contain *y*. At the genomic position 25563, RaTG13 base A was not present except in one
315 genome in the 29KG dataset, so we assigned the mutant status to the minority base (U; *vf* = 29.8%) and
316 the reference status to the majority base (G), an assumption that was tested in the same way as above.
317 All missing and ambiguous bases were coded to be ignored (missing data) in all the analyses.

318 In SCITE analyses, we first used default parameter settings of false-negative rate (FNR = 0.21545) and
319 false-positive rate (FPR = 0.0000604). We carried out ten independent runs to ensure stability and
320 convergence in analyzing a matrix of 49 × 29861 (SNVs × genomes). The mutation graph with the highest
321 log likelihood (*lnL* = −204,622.9) was used to obtain 29KG collection-specific estimates of FNR and FPR
322 by comparing the observed and predicted sequences based on this mutation graph. These estimated
323 FNR (0.0202) and FPR (0.0491) were very different from the SCITE default parameters, where estimated
324 FNR was much lower, whose use in SCITE produced mutation history graphs with a much higher log-
325 likelihood (*lnL* = −84,134.5). This difference in error rates is unsurprising because we used only common
326 variants (*vf* > 1%), and the 29KG dataset was not obtained from single-cell sequencing, in which the
327 allele dropout during single-cell sequencing elevates FNR, i.e., mutant alleles are not sequenced.

328 We then reversed ancestor/mutant coding for each variant to ensure that the orientation of each
329 mutation was optimal. 49 datasets were subjected to SCITE analyses (FNR = 0.0202 and FPR = 0.0491).
330 The initial assignment of the RaTG13 base to be the ancestor was supported for 47 positions. The log-

likelihood of mutation graphs with those assignments was significantly higher than the alternative ($P << 0.01$ using the AIC protocol in ref.[49]). This means that the Bat CoV base is likely the ancestral base for 95.9% of the variants with a frequency greater than 1%, similar to the pairwise sequence similarity observed between SARS-CoV-2 and RaTG13 CoV genomes (~96%). At one (position 25563) of the remaining two positions, the mutation history with the majority base (G) as the ancestral state received a significantly higher likelihood ($P << 0.01$). At the other position (3037), mutation history graphs from base C received significantly higher likelihood support ($\Delta lnL = 1621.7$) than base U, which indicated that the RaTG13 base (U) was not ancestral. In this case, the co-occurrence index was also much higher for the mutation history graph generated by using base C as the ancestor (91%). Therefore, we re-coded the column for position 3037 and generated a new 49 × 29861 (SNVs × genomes) matrix to conduct a SCITE analysis. It produced a mutation graph with much higher log-likelihood ($lnL = -25{,}979.1$) and lower FNR = 0.00537 and FPR = 0.00193.

We then used the above analysis settings and performed 100 runs of SCITE. Ninety-five mutation history graphs with the best log-likelihood ($lnL = -25{,}979.1$) were found. We chose the mutation graph with the highest average COI (97%) and presented it in figure 1. In this mutation graph, COI for each variant is shown next to the arrow, and an arrow is drawn with a dotted line if it occurred in fewer than a majority of the equally likely graphs. Based on the direction of the mutations in 49 SNVs, we generated the sequence of the progenitor SARS-CoV-2 (proCoV2). We have made available the proCoV2 genome sequence in FastA format at http://igem.temple.edu/COVID-19, which is the same as the NCBI reference genome with base differences (positions 2416, 19524, 23929, 18060, 8782, and 28144) as discussed in the text.

Temporal concordance. Because mutation ordering analysis analyses did not use spatial or temporal information for genomes or mutations, it can be validated by evaluating the concordance of the inferred order of mutations with the timing of their first appearance (*tf*). For a mutation *i*, we compared its tf_i with the tf_j such that *j* is the nearest preceding mutation in the mutation graph for which $tf_i \neq tf_j$. The condition was met for 47 out of 49 mutations (95.9%). Two offshoot mutants of (β_3) were sampled earlier than their predecessors by ten days. COI of one variant (β_{3b}) was low (54%), but the other variant (β_{3c}) showed a very high COI (97%).

Genome phylogeny and classification based on the mutation order

Each node in the mutational history graph predicts an intermediate (ancestral) or a tip sequence, which contains all the mutations from that node to the root of the mutation graph. Then, the phylogeny of these sequences is the same as the topology of the mutation history graph that is drawn as a directional graph anchored on the root node. Every node in the mutation-based phylogeny (Fig. 2a) was encoded in a mutational signature consisting of a binary barcode showing the presence/absence of 49 variants. We mapped each of the 29K genomes to a node in the phylogeny based on the highest sequence similarity at positions containing 49 common SNVs. Mismatches were allowed, as sequencing errors could create them. A small fraction of genomes (1.8%) could not be mapped unambiguously to one

368 node, so they were excluded and investigated in the future. The number of genomes assigned to each
369 node is shown in Extended Data Table 1.

370 We compared our mutational classification with a phylogeny-based classification[26] obtained using the
371 Pangolin service (v2.0.3; https://pangolin.cog-uk.io/). Sequences from our alignment were submitted
372 to the Pangolin website one-by-one, and a clade designation was received. The results are summarized
373 in Extended Data Figure 1. In this table, all phylogenetic-groups with fewer than 20 genomes were
374 excluded. Of the 80 phylogenetic groups shown, 74 are defined primarily by a single mutation-based
375 barcode, as more than 90% of the genomes in those phylogenetic groups share the same barcode. This
376 includes all small and medium-sized phylogenetic groups (up to 488 genomes) and two large groups
377 (A.1 with 1,377 genomes and B.1.2 with 749 genomes). One large group, B.1.1, predominately connects
378 with ε_3 node (79%, 4,832 genomes), but some of its members belong to ε_3 offshoots because they
379 contain respective diagnostic mutations. For group B.1.1.1, two other ε_3 offshoots are mixed up almost
380 equally. Three other large differences between mutational barcoding and phylogeny-based grouping
381 are seen for A, B, B1.1, and B.2 groups. These differences are likely because the location of the root, as
382 well as the earliest branching order of coronavirus lineages, is problematic in phylogeny-based
383 classifications[1,4,6,7]. Overall, our mutation-based classification is more intuitive and can have higher
384 resolution.

385 *Tests for neutral evolution and episodic spread of variants*

386 We tested the null hypothesis of the same molecular evolutionary patterns within the SARS-CoV-2
387 population and between species (i.e., Human SARS-CoV-2 and Bat RaTG13) by using a McDonald-
388 Kreitman test[38]. The numbers of non-synonymous and synonymous polymorphisms with a frequency
389 greater than 1% were 32 and 17, compared with the numbers of non-synonymous and synonymous
390 fixed differences (170 and 958, respectively) inferred between proCoV2 and bat RaTG13 sequences.
391 The McDonald-Kreitman test rejected the null overwhelmingly in a 2×2 contingency table analysis (P
392 $<< 0.01$).

393 We performed non-parametric Wald–Wolfowitz runs-tests[50,51] to examine the temporally episodic
394 spread of common mutants. Only one of the first six mutations was used because they were all sampled
395 first on pandemic day 0. A test using 44 variants rejected the null hypothesis ($P < 0.01$) as did a test in
396 the analysis that was restricted to variants with frequencies greater than 5% ($P < 0.05$). Therefore, the
397 non-random spread of variants is statistically significant.

398 **Data Availability and Code Availability:** Live evolutionary history and spatiotemporal distributions of
399 common variants are available at http://igem.temple.edu/COVID-19 (beta version). All genome
400 sequences and metadata are available publically at GISAID (https://www.gisaid.org/) and predicted
401 proCoV2 sequence is available at http://igem.temple.edu/COVID-19. The other relevant information is
402 provided in the supplementary materials.

References

1. Pipes, L., Wang, H., Huelsenbeck, J. & Nielsen, R. Assessing uncertainty in the rooting of the SARS-CoV-2 phylogeny. *bioRxiv* (2020) doi:10.1101/2020.06.19.160630.
2. Lai, A., Bergna, A., Acciarri, C., Galli, M. & Zehender, G. Early phylogenetic estimate of the effective reproduction number of SARS-CoV-2. *J. Med. Virol.* **92**, 675–679 (2020).
3. Castells, M., Lopez-Tort, F., Colina, R. & Cristina, J. Evidence of Increasing Diversification of Emerging SARS-CoV-2 Strains. *J. Med. Virol.* 1–8 (2020).
4. Mavian, C. *et al.* Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 12522–12523 (2020).
5. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450–452 (2020).
6. Wenzel, J. Origins of SARS-CoV-1 and SARS-CoV-2 are often poorly explored in leading publications. *Cladistics* **36**, 374–379 (2020).
7. Morel, B. *et al.* Phylogenetic analysis of SARS-CoV-2 data is difficult. *bioRxiv* (2020) doi:10.1101/2020.08.05.239046.
8. Gómez-carballa, A., Bello, X., Pardo-seco, J., Martinon-Torres, F. & Salas, A. genome variation worldwide highlights the impact of COVID-19 super-spreaders. *Genome Res.* (2020) doi:10.1101/gr.266221.120.
9. Nei, M. & Kumar, S. *Molecular Evolution and Phylogenetics*. (Oxford University Press, 2002).
10. van Dorp, L. *et al.* Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).
11. Jahn, K., Kuipers, J. & Beerenwinkel, N. Tree inference for single-cell data. *Genome Biol.* **17**, 1–17 (2016).
12. Miura, S. *et al.* Computational enhancement of single-cell sequences for inferring tumor evolution. *Bioinformatics* **34**, i917–i926 (2018).
13. Ross, E. M. & Markowetz, F. OncoNEM: Inferring tumor evolution from single-cell sequencing data. *Genome Biol.* **17**, 1–14 (2016).
14. Kim, K. I. & Simon, R. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics* **15**, (2014).
15. Fauver, J. R. *et al.* Coast-to-Coast Spread of SARS-CoV-2 during the Early Epidemic in the United States. *Cell* **181**, 990-996.e5 (2020).
16. Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.* **7**, 1012–1023 (2020).
17. Dearlove, B. L. *et al.* A SARS-CoV-2 vaccine candidate would likely match all currently circulating strains. *bioRxiv* (2020) doi:10.1101/2020.04.27.064774.
18. Stefanelli, P. *et al.* Whole genome and phylogenetic analysis of two SARSCoV-2 strains isolated in Italy in January and February 2020: Additional clues on multiple introductions and further circulation in Europe. *Eurosurveillance* **25**, 1–5 (2020).
19. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
20. MacLean, O. A. *et al.* Natural selection in the evolution of SARS-CoV-2 in bats, not humans, created a highly capable human pathogen. *bioRxiv* (2020) doi:10.1101/2020.05.28.122366.
21. Rice, A. M. *et al. Evidence for strong mutation bias towards, and selection against, U content in SARS-CoV-2: implications for vaccine design*. *Molecular Biology and Evolution* (2020). doi:10.1093/molbev/msaa188.
22. Matyášek, R. & Kovařík, A. Mutation patterns of human SARS-CoV-2 and bat RaTG13 coronaviruses genomes are strongly biased towards C>U indicating rapid evolution in their hosts. *Genes (Basel)* **11**, 761 (2020).
23. Castillo, A. E. *et al.* Phylogenetic analysis of the first four SARS-CoV-2 cases in Chile. *J. Med. Virol.* **92**, 1562–1566 (2020).
24. Deng, X. *et al.* Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. *Science* **369**, 582–587 (2020).
25. Nextstrain. https://nextstrain.org/coronavirus/SARS-CoV-2.
26. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* (2020) doi:10.1038/s41564-020-0770-5.
27. Islam, O. K. *et al.* Emergence of European and North American mutant variants of SARS-CoV-2

in South-East Asia. *Transbound. Emerg. Dis.* **00**, 1–9 (2020).
28. Chi, X. *et al.* A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **369**, 650–655 (2020).
29. Luan, J., Lu, Y., Jin, X. & Zhang, L. Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection. *Biochem. Biophys. Res. Commun.* **526**, 165–169 (2020).
30. Hoffmann, M., Kleine-Weber, H. & Pöhlmann, S. A Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. *Mol. Cell* **78**, 779-784.e5 (2020).
31. Daniloski, Z. *et al.* The Spike D614G mutation increases SARS-CoV-2 infection of multiple human cell types. *bioRxiv* (2020) doi:10.1101/2020.06.14.151357.
32. Yurkovetskiy, L. *et al.* Structural and Functional Analysis of the D614G SARS-CoV-2 Spike Protein Variant. *bioRxiv* (2020) doi:10.2139/ssrn.3657338.
33. Korber, B. *et al.* Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **182**, 812-827.e19 (2020).
34. Wang, Q. *et al.* Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* **181**, 894-904.e9 (2020).
35. Zeng, W. *et al.* Biochemical characterization of SARS-CoV-2 nucleocapsid protein. *Biochem. Biophys. Res. Commun.* **527**, 618–623 (2020).
36. Gonzalez-Reiche, A. S. *et al.* Introductions and early spread of SARS-CoV-2 in the New York City area. *Science* **369**, 297–301 (2020).
37. Worobey, M. *et al.* The emergence of SARS-CoV-2 in Europe and the US. *bioRxiv* (2020) doi:10.1101/2020.05.21.109322.
38. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652–654 (1991).
39. Giorgio, S. Di, Martignano, F., Torcia, M. G., Mattiuz, G. & Conticello, S. G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* **6**, 1–9 (2020).
40. Casals, F. & Bertranpetit, J. Human genetic variation, shared and private. *Science* **336**, 39–40 (2012).
41. Hadfield, J. *et al.* NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
42. Gianella, S. *et al.* Detection of Minority Resistance during Early HIV-1 Infection: Natural Variation and Spurious Detection rather than Transmission and Evolution of Multiple Viral Variants. *J. Virol.* **85**, 8359–8367 (2011).
43. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
44. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
45. Stecher, G., Tamura, K. & Kumar, S. Molecular evolutionary genetics analysis (MEGA) for macOS. *Mol. Biol. Evol.* **37**, 1237–1239 (2020).
46. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
47. Chiara, M., Horner, D. S., Gissi, C. & Pesole, G. Comparative genomics provides an operational classification system and reveals early emergence and biased spatio-temporal distribution of SARS-CoV-2. *bioRxiv* (2020) doi:10.1101/2020.06.26.172924.
48. Wei, Yulong; Sike, Jordan R.; Aris, Parisa; Xia, X. Coronavirus genomes carry the signatures of their habitats. *bioRxiv* (2020) doi:10.1101/2020.06.13.149591.
49. Pupko, T., Huchon, D., Cao, Y., Okada, N. & Hasegawa, M. Combining multiple data sets in a likelihood analysis: Which models are the best? *Mol. Biol. Evol.* **19**, 2294–2307 (2002).
50. Wald, A. & Wolfowitz, J. On a Test Whether Two Samples are from the Same Population. *Ann. Math. Stat.* **11**, 147–162 (1940).
51. Mateus, A. & Caeiro, F. An R implementation of several randomness tests. *AIP Conf. Proc.* **1618**, 531–534 (2015).

Acknowledgments

We thank all the authors and organizations who have kindly deposited and shared genome data on GISAID (see http://igem.temple.edu/COVID-19 for a list of all the authors). We thank Ananias Escalante, Rob Kulathinal, Li Liu, Jose Barba-Montoya, Antonia Chroni, Ravi Patel, and Caryn Babaian for critical comments. We appreciate the technical support provided by Jared Knoblauch and Glen Stecher. This research was supported by grants from the U.S. National Science Foundation to S.K. (GCR-1934848 and DEB-2034228) and S.P. (DBI-2027196) and from the U.S. National Institutes of Health to S.K. (GM-0126567) and S.P. (AI-134384).

Author Contributions

S.K. conceived the project, designed analyses and visualizations, conducted analyses, and wrote the manuscript. S.M. designed and conducted analyses. S.P., S.W., and M.A.C.O. assembled sequence alignments. M.A.C.O., S.M., S.S., and Q.T. conducted analyses and rendered visualizations. S.P. and S.W. developed interactive visualizations. All authors intellectually contributed by discussing results and patterns, and everyone contributed to writing the manuscript.

Competing Interests

The authors declare that they have no competing interests.

Additional Information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to s.kumar@temple.edu. Reprints and permissions information is available at www.nature.com/reprints.

534 **Extended Data Table 1**. SARS-CoV-2 variants and their molecular types and first timing and location.

Mutant (major)	Mutant (minor)	Gene	Genomic Position	Nucleotide change	Amino acid change	Time (days)	Variant Frequency	Genomes mapped	First location
μ_1		ORF1ab	2416	U>C		0	98.1%	18	China, Asia
μ_2		ORF1ab	19524	U>C		0	98.6%	0	China, Asia
μ_3		S	23929	U>C		0	98.4%	0	China, Asia
α_1		ORF1ab	18060	U>C		0	95.1%	849	China, Asia
	α_{1a}	N	28657	C>U		63	1.3%	2	France, Europe
	α_{1b}	ORF1ab	9477	U>A	F>Y	63	1.2%	8	France, Europe
	α_{1c}	N	28863	C>U	S>L	63	1.2%	0	France, Europe
	α_{1d}	ORF3a	25979	G>U	G>V	63	1.2%	344	France, Europe
α_2		ORF1ab	8782	U>C		0	91.0%	47	China, Asia
α_3		ORF8	28144	C>U	S>L	0	90.8%	1116	China, Asia
	α_{3a}	ORF1ab	1606	U>C		43	1.7%	501	United Kingdom, Europe
	α_{3b}	ORF1ab	11083	G>U	L>F	24	9.2%	377	China, Asia
	α_{3c}	N	28311	C>U	P>L	64	1.9%	3	South Korea, Asia
	α_{3d}	ORF1ab	13730	C>U	A>V	71	1.8%	3	Taiwan/Malaysia, Asia
	α_{3e}	ORF1ab	6312	C>A	T>K	71	1.7%	483	Taiwan/Malaysia, Asia
	α_{3f}	ORF3a	26144	G>U	G>V	28	5.1%	452	China, Asia
	α_{3g}	ORF1ab	14805	C>U		54	6.0%	3	United Kingdom, Europe
	α_{3h}	ORF1ab	17247	U>C		64	2.0%	580	Switzerland, Europe
	α_{3i}	ORF1ab	2558	C>U	P>S	54	1.7%	26	United Kingdom, Europe
	α_{3j}	ORF1ab	2480	A>G	I>V	54	1.6%	462	United Kingdom, Europe
β_1		ORF1ab	3037	C>U		31	77.0%	40	China, Asia
β_2		S	23403	A>G	D>G	31	77.1%	9	China, Asia
β_3		ORF1ab	14408	C>U	P>L	41	76.9%	3032	Saudi Arabia, Middle East
	β_{3a}	ORF1ab	20268	A>G		64	5.7%	1213	Italy, Europe
	β_{3b}	N	28854	C>U	S>L	29	3.1%	34	China, Asia
	β_{3c}	ORF1ab	15324	C>U		29	2.3%	669	China, Asia
	β_{3d}	ORF3a	25429	G>U	V>L	77	1.7%	484	United Kingdom, Europe
	β_{3e}	N	28836	C>U	S>L	74	1.6%	3	Switzerland, Europe
	β_{3f}	ORF1ab	13862	C>U	T>I	74	1.6%	50	Switzerland, Europe
	β_{3g}	ORF1ab	10798	C>A	D>E	86	1.4%	414	United Kingdom, Europe
γ_1		ORF3a	25563	G>U	Q>H	41	29.8%	884	Saudi Arabia, Middle East
	γ_{1a}	ORF1ab	18877	C>U		41	4.0%	757	Saudi Arabia, Middle East
	γ_{1b}	M	26735	C>U		41	1.5%	439	Saudi Arabia, Middle East
δ_1		ORF1ab	1059	C>U	T>I	54	23.0%	5157	Singapore, Asia
	δ_{1a}	S	24368	G>U	D>Y	72	1.3%	389	Singapore, Asia
	δ_{1b}	ORF8	27964	C>U	S>L	76	2.7%	790	USA, North America
	δ_{1c}	ORF1ab	11916	C>U	S>L	72	1.6%	166	USA, North America
	δ_{1d}	ORF1ab	18998	C>U	A>V	72	1.0%	305	USA, North America
ε_1		N	28881	G>A	R>K	54	25.7%	2	United Kingdom, Europe
ε_2		N	28882	G>A	R>K	54	25.7%	2	United Kingdom, Europe
ε_3		N	28883	G>C	G>R	54	25.7%	5365	United Kingdom, Europe
	ε_{3a}	ORF1ab	313	C>U		66	2.1%	608	USA, North America
	ε_{3b}	ORF1ab	19839	U>C		64	1.5%	452	Switzerland, Europe
	ε_{3c}	M	27046	C>U	T>M	69	1.6%	453	Worldwide
	ε_{3d}	ORF1ab	10097	G>A	G>S	69	2.5%	5	Denmark, Europe
	ε_{3e}	S	23731	C>U		69	2.5%	403	Denmark, Europe
	ε_{3f}	N	28580	G>U	D>Y	69	1.2%	353	Chile, South America
ν_1		ORF1ab	17858	A>G	Y>C	59	4.7%	32	USA, North America
ν_2		ORF1ab	17747	C>U	P>L	59	4.7%	1374	USA, North America

535
536 Note.- Genomic locations correspond to those of the NCBI genome (GenBank ID: NC_04551 2). Amino
537 acid changes are shown for non-synonymous variants. Genomes mapped are for nodes in figure 2a.

538

	Mutation-based barcode																																
Phylo-class	α1	α1d	α2	α3	α3a	α3b	α3e	α3f	α3h	α3i	α3j	β1	β3	β3a	β3b	β3c	β3d	β3f	β3g	γ1	γ1a	δ1	δ1a	δ1b	δ1c	δ1d	ε3	ε3a	ε3b	ε3c	ε3e	ε3f	γ1b	ν1	ν2
B.1													2256	515	8	576	204	50		522	588	4410	369	41	165	268									
B.1.1																											4832	558	277	423					
A.1																																		15	1362
B				744	349	126						33		6	22	14																			
B.2				7				367	402																										
B.1.2																								749											
B.1.1.1																															403	342			
A	440		42			9																												16	
B.2.1										26	462																								
B.6							483																												
B.1.5													23	412																					
B.1.93																			414																
A.2		344																																	
B.3				275		11																													
B.1.36																																	252		
A.3	239																																		
B.1.11																	229																		
B.1.22																				205															
B.4				13		197																													
B.1.8													189																						
B.1.113																																	180		
B.1.43																						174													
B.1.1.33																											129								
B.1.1.10																											110								
B.1.26																						110													
B.2.2									107																										
A.5	87																																		
B.1.1.2																													90						
B.1.42																						90													
B.1.9																					90														
B.16					86																														
B.1.75													86																						
B.1.116																						80													
B.1.102														76																					
B.1.13																				67															
B.1.5.2														64																					
B.1.124																					62														
B.1.5.9													61																						
B.1.1.7																											53								
B.2.4								54																											
B.1.80																48																			
B.1.89													49																						
B.1.39																						48													
A.6	41																																		
B.1.107													43																						
B.1.110																				44															
B.1.78													44																						
B.1.5.5														43																					
B.1.1.4																													41						
B.2.7									40																										
B.1.79													39																						
B.1.1.28																											38								
B.1.3																										37									
B.1.37																				37															
B.1.5.10														37																					
B.1.104																	35																		
B.10					35																														
B.1.1.32																												31							
B.1.96													33																						
B.1.1.26																											32								
B.1.41																						32													
B.1.40																						30													
A.4	29																																		
B.1.1.30																											29								
B.1.19													29																						
B.1.1.25																											28								
B.1.12																						27													
B.14				27																															
B.1.1.14																											25								
B.1.1.5																													25						
B.1.101														25																					
B.1.108																						25													
B.1.117																						25													
B.1.23													25																						
B.1.1.17																														24					
B.2.6									22																										
B.1.120																						21													
B.1.29																						21													
B.1.71													21																						
B.1.44																							20												

Extended Data Figure 1. A comparison of mutation-based and phylogeny-based classifications of 29KG genomes. Phylogeny-based classification is obtained by using the Pangolin service (v2.0.3; https://pangolin.cog-uk.io/). The value in each cell is the number of genomes that belong to the corresponding mutation-based and phylogeny-based groups. All phylogenetic-based groups with fewer than 20 genomes are excluded, and cells with fewer than five genomes matching have been left empty to make the comparison more straightforward and to allow for sequencing and estimation errors.