KSBMB Korean Society for Biochemistry and Molecular Biology

www.nature.com/emm

ARTICLE OPEN

Check for updates

Mutational spectrum of SARS-CoV-2 during the global pandemic

Kijong Yi[1], Su Yeon Kim[1], Thomas Bleazard[2], Taewoo Kim[1], Jeonghwan Youk[1,3] and Young Seok Ju[1,3]✉

Viruses accumulate mutations under the influence of natural selection and host–virus interactions. Through a systematic comparison of 351,525 full viral genome sequences collected during the recent COVID-19 pandemic, we reveal the spectrum of SARS-CoV-2 mutations. Unlike those of other viruses, the mutational spectrum of SARS-CoV-2 exhibits extreme asymmetry, with a much higher rate of C>U than U>C substitutions, as well as a higher rate of G>U than U>G substitutions. This suggests directional genome sequence evolution during transmission. The substantial asymmetry and directionality of the mutational spectrum enable pseudotemporal tracing of SARS-CoV-2 without prior information about the root sequence, collection time, and sampling region. This shows that the viral genome sequences collected in Asia are similar to the original genome sequence. Adjusted estimation of the d*N*/d*S* ratio accounting for the asymmetrical mutational spectrum also shows evidence of negative selection on viral genes, consistent with previous reports. Our findings provide deep insights into the mutational processes in SARS-CoV-2 viral infection and advance the understanding of the history and future evolution of the virus.

Experimental & Molecular Medicine (2021) 53:1229–1237; https://doi.org/10.1038/s12276-021-00658-z

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is spreading rapidly and globally[1,2]. Phylogenetically, SARS-CoV-2 belongs to the subgenus Sarbecovirus, a branch of *Betacoronavirus* in the *Coronaviridae* family[3]. Recent comparative studies proposed bats[4,5] and pangolins[6–8] as possible natural reservoirs of the virus.

Acquisition of new mutations in viral genomes and natural selection acting on the resultant phenotypic diversity are the two constituent processes of viral evolution[9]. Analogous to Darwinian evolution occurring in the origins of species, mutant viral strains gain a driving force when they acquire variations that increase infectivity and survival in the host environment. Therefore, understanding the genome changes of SARS-CoV-2 during the recent outbreak and their proper interpretations are critical for developing preventive, diagnostic, and therapeutic strategies against the virus. Although hundreds of thousands of SARS-CoV-2 genomes have been sequenced[10,11], interpretations of the mutations as a whole have not been conducted.

Mutational signature analyses have been widely applied to cancer genomes to understand the mutational processes operative in human somatic cells[12–14]. Unique patterns of mutations provide deep insights into the mechanistic sources of genomic mutations active in the somatic lineages of cancer cells. These approaches also suggest environmental exposure of the cellular lineage, as seen in transmissible cancers[15]. Here, we apply similar approaches to the catalog of mutations collected from a considerable number of sequenced SARS-CoV-2 genomes ($n = 351{,}525$). Then, using the spectrum, we trace the origin of SARS-CoV-2 spreading in the human population and speculate on the molecular mechanisms that have shaped SARS-CoV-2 evolution in the recent outbreak and subsequent functional impacts of the mutations.

MATERIALS AND METHODS

Data collection and processing

An overview of the steps in our analysis is shown in Supplementary Fig. 1. A set of 351,525 prealigned full-length SARS-CoV-2 genome sequences were downloaded on 1/17/2021 from GISAID (Global Initiative On Sharing All Influenza Data[10]) (Supplementary Table 1). Viral sequences in other orders, families, and genera (e.g., *Alphacoronavirus*), as well as those of several bat coronaviruses, were downloaded from GenBank[10] (Supplementary Table 2). Seven pangolin CoV sequences were downloaded from GISAID with accession numbers EPI_ISL_410538 to EPI_ISL_410544.

The sequences of the viruses other than SARS-CoV-2 were aligned using Kalign 3.2.3[10]. The phylogenetic trees were constructed using FastTree 2.1.11[16], a tool for building an "approximately maximum likelihood" phylogenetic tree with default parameters. All the initial multiple sequence alignments were manually reviewed using Aliview[10]. The trees constructed were also manually reviewed using the 'ape' and 'phangorn' R packages[17,18]. With Wuhan-Hu-1 (NC_045512.2, EPI_ISL_402125)[19] as a temporary root sequence, ancestral alleles in the internal nodes were predicted using the 'ancestral.pml' function with the 'type = MPR' option in the 'phangorn' R package. Mutations were called from the edges of the tree only when the mutated base could be unambiguously assigned to the conserved neighborhood, and the two upstream and two downstream bases from the mutated base had to be identical between the nodes being compared. Multiple base substitutions and short indels (<15 bases) were called in a similar way.

[1]Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Korea. [2]National Institute for Biological Standards and Control, Blanche Lane, South Mimms, Potters Bar, Hertfordshire EN6 3QG, UK. [3]GENOME INSIGHT Inc, Daejeon 34051, Korea. ✉email: ysju@kaist.ac.kr

Received: 30 November 2020 Revised: 29 April 2021 Accepted: 11 May 2021
Published online: 27 August 2021

Visual inspection of highly recurrent mutations

For a better understanding of the recurrent substitutions at position 11,083, we randomly selected and downloaded ~30 Illumina short-read sequencing datasets from the SRA database and reviewed them using the Integrative Genomic Viewer[20].

Reconstruction of the mutational spectrum

For mutational signature analysis, single-base substitutions were classified into 192 subclasses: 4 types of the original bases (4) mutated to the other bases (×3) in the context of immediate upstream (×4) and downstream (×4) bases. In the given set of substitutions, the mutational spectrum was

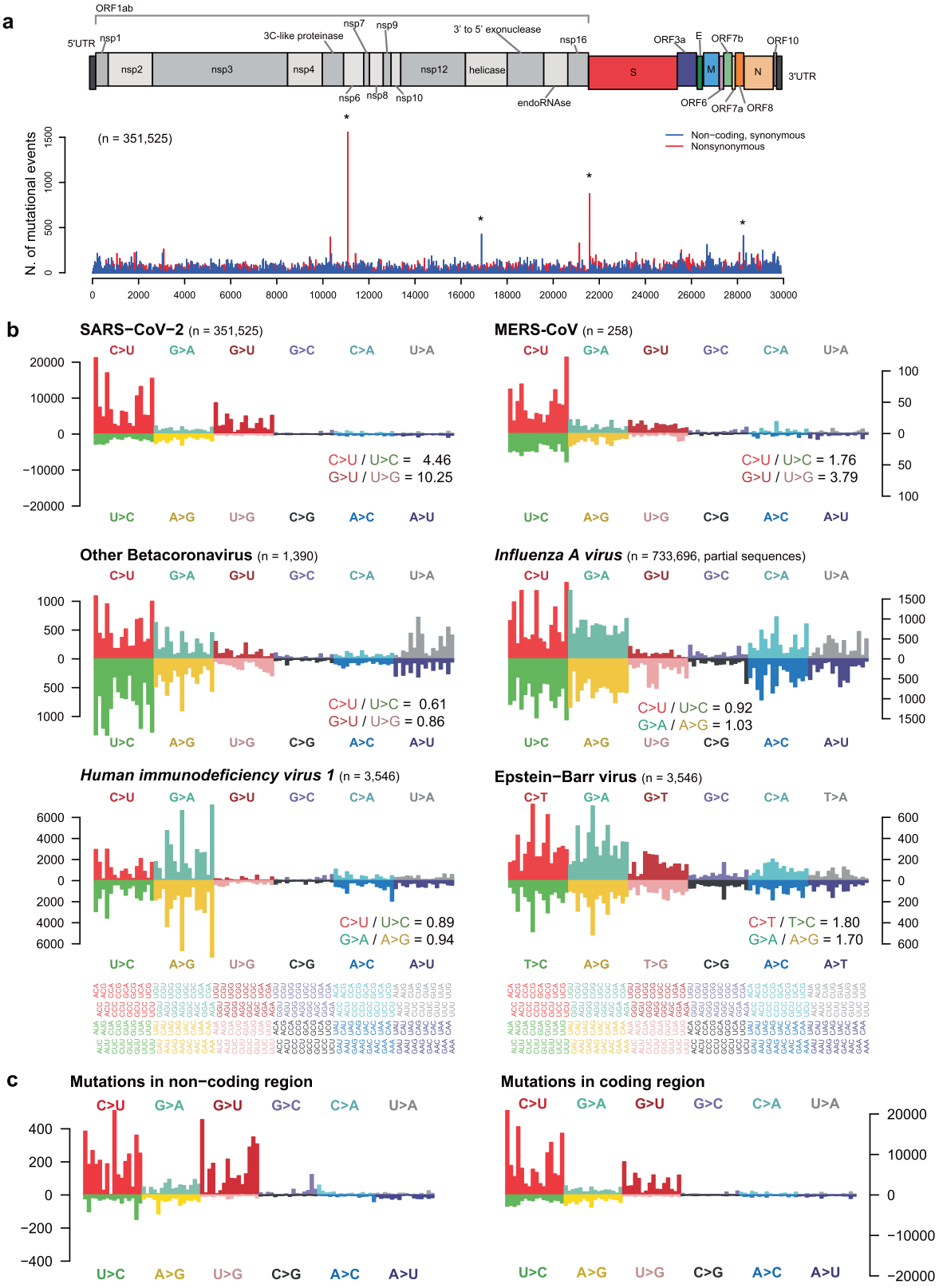

Fig. 1 Mutational signature of SARS-CoV-2. a Distribution of the number of single-base substitutions along the viral genome. Each bar in the lower panel represents the counts for nonsynonymous substitutions (red) and synonymous substitutions in noncoding regions (blue). Except for a few recurrently mutated positions forming peaks, mutations are more or less uniformly distributed along the genome. The four highest peaks marked by asterisks are located in homopolymeric stretches. The highest peak at position 11,083 is caused by recurrent G→U and U→G substitutions in the context of U (5′-UUUUUUUGU-3′) (detailed in Supplementary Fig. 1). **b** Mutational signatures of SARS-CoV-2 and other viruses: MERS-CoV, other betacoronaviruses (including nonhuman hosted ones), *Influenza A virus*, HIV-1, and Epstein-Barr virus. For a given species (indicated in the panel title together with $n =$ sample size), the panel shows the spectrum of observed substitution counts for 192 classes, a combination of changes of the major base (12 scenarios indicated with different colors) together with 4 types of the 5′ immediate upstream base and 4 types of the 3′ immediate downstream. Mutations of SARS-CoV-2 are particularly enriched in five sequence contexts (ACA, ACU, UCA, UCU, and GCU), and the mutational spectrum is asymmetric in terms of Watson-Crick base pairing and directional (i.e., the mutated and substituted bases are not balanced). Specifically, in SARS-CoV-2, C→U is much more frequent than U→C, and similarly, G→U is more frequent than U→G. MERS-CoV also exhibits an asymmetrical mutational spectrum similar to that of SARS-CoV-2. In contrast, *Influenza A virus* and HIV-1 show largely balanced patterns (C→U ≈ U→C and G→A ≈ A→G). Epstein-Barr virus, which is a DNA virus, shows asymmetry in its reversibility (C→T ≉ T→C) but exhibits symmetry in Watson-Crick base pairing (C→T ≈ G→A in a CpG context). **c** Comparison of mutational signatures in mutations of noncoding and coding regions.

defined by the numbers of each class of substitutions. We performed a subsampling analysis, which revealed that ~200 base substitutions are necessary to reconstruct a stable mutational spectrum (Supplementary Fig. 2). Of note, >220 K base substitutions were used for our final spectrum. Mutations on the 2nd–6th terminal branches exclusively shared among patients from the same country were defined as country-specific mutations. Terminal branch mutations were not considered for country-specific mutations because many sequencing errors are enriched on the terminal branches. The same approach was applied by annotating mutations with sampling month instead of country to investigate mutational spectrum changes depending on the virus collection month. The frequency patterns of APOBEC1 and APOBEC3 family enzyme sites were adopted from previous studies[21,22] and represented as base frequency plots using the 'ggseqlogo' R package[23].

Maximum-likelihood estimation of the origin of SARS-CoV-2

Using the mutational spectrum of SARS-CoV-2, we estimated the root by a likelihood method, which computes the likelihood of the root by multiplying conditional probabilities of observing the descendant node base given the ancestral base across all branches in the tree (top-down approach). The conditional probability for each configuration of ancestral and descendant contexts was computed based on the substitution rate estimated from the data using the observed counts for each of the 192 contexts. The estimated substitution rates per class of substitutions were directly adopted from the observed mutational spectrum by scaling the spectral numbers to sum to 1. The likelihood of being the root of each node was calculated using the sum of the logarithmic value of substitution rates of all mutations, assuming the node is the root.

Calculation of dN/dS using the 192-context model

Variations in the methods used to calculate dN/dS come from different assumptions regarding the expected density of neutral mutations[24,25]. We adopted the concept from previous work by Martincorena et al.[26], an application of a 192-context model to the somatic evolution of cancer. We calculated ground neutral mutation rates for one base to another in a certain position by simply dividing the count for each substitution type by the number of corresponding contexts that appeared in the specified region. For a given region in the genome (e.g., *ORF3a*), the expected number of neutral synonymous mutations was calculated by the sum of the probabilities where a base change resulted in a silent mutation. The expected number of nonsynonymous mutations was calculated in a similar way. The dN and dS values were then calculated using the observed and expected numbers of nonsynonymous and silent mutations, respectively. To estimate the confidence interval of d*N*/d*S* for each region, bootstrapping was performed by randomly resampling the observed mutations in that region. Conventional maximum likelihood methods with different models and codon frequencies were tried to evaluate natural selection using IQ-TREE software (v2.1.1)[27].

RESULTS

Detection of SARS-CoV-2 mutations

We used 351,525 prealigned full-length SARS-CoV-2 genome sequences publicly released in the GISAID database as of January 17, 2021[10]. These sequences were collected from 82 countries from December 24, 2019, to January 12, 2021 (Supplementary Table 1). After constructing the phylogenetic tree and estimating internal node sequences ("Materials and Methods"), we compared sequences connected by each edge in the tree. We cataloged base changes, including nucleotide substitutions and short indels. From our efforts, we obtained a list of mutation sites containing 227,639 single-base substitutions, 1070 double- or triple-base substitutions, and 6001 short indels (Fig. 1a, Supplementary Fig. 3a, Supplementary Table 1, and a list of mutations with annotations in Supplementary Table 3). For comparability with other papers, we use the genomic coordinates of the conventional SARS-CoV-2 reference, NC_045512 (known as Wuhan-Hu-1), throughout the manuscript. The mutation sites were more or less uniformly distributed across the entire viral genome (Supplementary Fig. 3b).

Several sites were recurrently mutated on multiple branches, implying many independent mutational events at the same genomic sites if we neglect the small chance of viral genome recombination between different mutants. At first glance, these mutations can be interpreted as evidence of positive selection. However, with careful examination of such mutations, especially for the top four recurrent calls (Fig. 1a), these sites are recurrently mutated in both directions. For example, *ORF1ab* L3606F (11,083 G→U substitution) and its reverse F3606L (11,083 U→G substitution) occurred 1505 times and 49 times, respectively (Supplementary Fig. 4a). This implies something other than evidence of positive selection, such as either replicative or sequencing error-prone sites. Indeed, these recurrent sites were frequently located in the homopolymeric region, and even deep sequencing data of a virus pool from one patient showed reads representing both alleles with various allele frequencies (Supplementary Fig. 4b). Therefore, we speculated that these mutations might not induce obvious functional advantages in the spread of SARS-CoV-2[28].

Mutational signature of SARS-CoV-2

Of the 12 classes of base substitutions, the C→U transition was dominant in our mutation catalog (46.5%). Interestingly, we found substantial asymmetry in base changes. For example, the rate of the C→U transition was much higher than that of its reverse U→C substitution (46.5% vs. 9.4%, respectively). Likewise, the rate of G→U transversion was almost ten times higher than that of U→G substitution (18.2% vs. 1.3%, respectively).

To more deeply explore the mutational spectrum, we considered the sequence context of mutated bases and partitioned the 12 classes of base substitutions into 192 subclasses (i.e., 12 substitution classes × 4 types of the 5′ immediate upstream base × 4 types of the 3′ immediate downstream base; Supplementary Table 4). Under these circumstances, C→U base substitutions were further enriched in five sequence contexts, ACU, ACA, UCA, UCU, and GCU (Fig. 1b). C→U transitions in the five contexts were 7.5 times more frequent than the reverse U→C

transition in the same surrounding bases. These results indicate that the proportion of uracil has increased over time, at least during the early spread of SARS-CoV-2 in the human population. Thus, the pool of viral genomes has been undergoing directional change.

Interestingly, asymmetric mutational spectra have also been found in the spread of other zoonotic betacoronavirus species, such as *Middle East respiratory syndrome coronavirus* (MERS-CoV, Fig. 1b) and SARS-CoV (11 C→U vs. 7 U→C from 11 sequences, data not shown), which are thought to have recently been introduced into the human population. Asymmetry was not found in old viruses in the human population, such as *Influenza A virus* and *Human immunodeficiency virus 1* (HIV-1) (Fig. 1b, Supplementary Fig. 5, and Supplementary Table 4).

Three factors may shape SARS-CoV-2 mutations in infected cells: (1) mistakes in viral genome replication by RNA-dependent RNA polymerase (known as RdRP), (2) active damaging processes in RNA molecules caused by host immune systems[29], and (3) the selective advantage of possessing a certain context due to an underlying mechanism such as codon usage adaptation[30,31]. Of these, the RdRP error cannot explain the asymmetry of the mutational spectrum we observed, in particular the imbalance between substitution classes, because both positive- and negative-sense RNAs are equivalently replicated by the polymerase in infected cells (Supplementary Fig. 6).

Recently, RNA editing was suggested as a mutational process for SARS-CoV-2[29,32]. However, there are some missing links between RNA editing and our mutational spectrum. Although our mutational spectrum is quite similar to the pattern of mutations caused by APOBEC1 ([A/U]p[C→U]p[A/U]), the enzyme is exclusively expressed in the small and large intestine, which is not a predominant target organ for this viral infection. Moreover, APOBEC1-mediated RNA editing requires a mooring sequence downstream of the mutant site (WRAUYANUAU 3-10 bases downstream of the target cytosine site)[21,33]. However, in the mutations we profiled, no such sequence was found even after allowing two bases of mismatch. In addition, the mutational spectrum of APOBEC3A or APOBEC3B (Up[C→U]pN)[34,35] was not very similar to our spectrum, although the enzymes are known to be expressed in the lung (Supplementary Fig. 7a). Indeed, asymmetry in another class of base substitutions (G→U base substitutions, which are 10.2 times more frequent than U→G changes) cannot be explained by currently known RNA-editing enzymes.

We investigated whether the asymmetry of the mutational spectrum is the outcome of selection pressure on infected human cells. Many viruses tend to have codon usage similar to that of their major host due to higher efficiency in protein translation[30,36–40]. However, the asymmetric mutational spectrum overall makes viral codon usage more dissimilar from that of humans (Supplementary Fig. 7b). Among the 20 kinds of amino acids, 15 can have uracil in the third position of the codon. When looking at all the synonymous mutations observed in SARS-CoV-2 by amino acid, in all 15 kinds, the change to codons with uracil in the third position was the most frequent. In the codons for 14 of these 15 amino acids (except arginine), the synonymous mutations most frequently produced codons that are the most used in the virus genome. Moreover, we also observed that the mutation spectrum in the noncoding region was almost identical to that in the coding region, although the mutation rate in the noncoding region was slightly higher (13.93 vs. 9.18 mutations/base in the tree of 351,525 samples) (Fig. 1c and Supplementary Fig. 7c). Overall, the mutational spectrum of SARS-CoV-2 exhibits an asymmetric, currently unknown mutational process that has been operative during transmission in the human population.

Tracing the origin of SARS-CoV-2

Next, we used the asymmetric mutational spectrum to root the SARS-CoV-2 pandemic. Conventionally, the reference genome of SARS-CoV-2, which was sequenced from viral particles isolated from Wuhan in December 2019 (known as Wuhan-Hu-1)[19], is considered a root in viral genome studies[41]. Additionally, previous studies used outgroup sequences, such as bat coronavirus sequences, to find the root of SARS-CoV-2[42,43]. However, due to the immoderate sequence homology among the outgroup sequences, the approaches pointed to incorrect sequences as the origin of SARS-CoV-2.

To tackle this problem, we used the directionality of the mutations to trace the root of SARS-CoV-2 without information about the sampling date. Our asymmetric mutational signature directly suggests that viral sequences harboring a higher proportion of C and G are likely to be more ancestral, that is, more similar to the original genomic sequence. To this end, we employed a likelihood approach to statistically evaluate the root from 351,525 sequences ("Materials and Methods"). We used a fixed substitution rate matrix proportional to the mutational spectrum, based on the observation that the spectrum (1) is robust to the position of the root (Supplementary Fig. 8a, b) and (2) is also uniform across the tree, regardless of the viral clades, countries of patients, or sampling dates (Supplementary Fig. 9a–c).

Our approach clearly suggested a root, which is an ancestral node of four SARS-CoV-2 sequences collected in Wuhan in the earliest period (February 2020, Fig. 2a, b; Supplementary Table 1). The maximum-likelihood node showed three U→C substitutions compared to Wuhan-Hu-1, the typical reference sequence of SARS-CoV-2. The node sequence was 440 times more probable as a root than was Wuhan-Hu-1.

The likelihood for the root of each node exhibits a trend of steadily decreasing over the collection date of the sequences, supporting the validity of our method (Fig. 2c). Such a trend was not clearly seen in the typical outgroup-based analyses (Fig. 2d). Although the sequence identity between SARS-CoV-2 and its closest bat coronavirus (RmYN02) is 96.2%, these viral genomes differ by more than 1,000 mutations. This far outnumbers the number of sequence changes between any two sequenced genomes of SARS-CoV-2 (6.4 substitutions on average). As a consequence, rooting based on the edit distance from the outgroup resulted in arbitrary outcomes[42]. In addition, the overall edit distance of the viral genomes did not significantly change over the collection dates of SARS-CoV-2 (Fig. 2d), highlighting the limitation of the outgroup approach in tracing the origin.

Mutational signature in the natural relatives of SARS-CoV-2

We investigated whether spectrum asymmetry is also present in long-term fixation in close relatives of SARS-CoV-2, such as bat coronaviruses (RaTG13 and RmYN02) and a pangolin coronavirus recently sequenced from Guangdong and Guanxi, China[4–7]. In the divergence of these viruses, the rate of U→C substitutions was higher than that of C→U substitutions (Fig. 3a–i). Similar to the spread among humans, mutations between the six pangolin coronavirus sequences from Guangxi[6] also have more C→U substitutions than U→C substitutions, but the number of mutations is too small for confirmation (19 C→U and 9 U→C among 77 substitutions).

Functional consequences of the SARS-CoV-2 mutations

Of the 227,639 substitutions we identified, 162,221 (58.4%) induce amino acid alterations of the SARS-CoV-2 protein-coding genes. These nonsynonymous mutations were relatively evenly distributed in proportion to the length of the gene and had various ratios of C→U transitions (Supplementary Fig. 10). A large proportion of the amino acid changes follow the preferential C→U substitution in the mutational signature. For example, the most common type of nonsynonymous mutation observed in human SARS-CoV-2 is alanine to valine (Ala→Val, $n = 12{,}202$; 4.6% of all nonsynonymous mutations). Of these, 7252 (59.4%) coincided with the trinucleotide context signature, occurring by

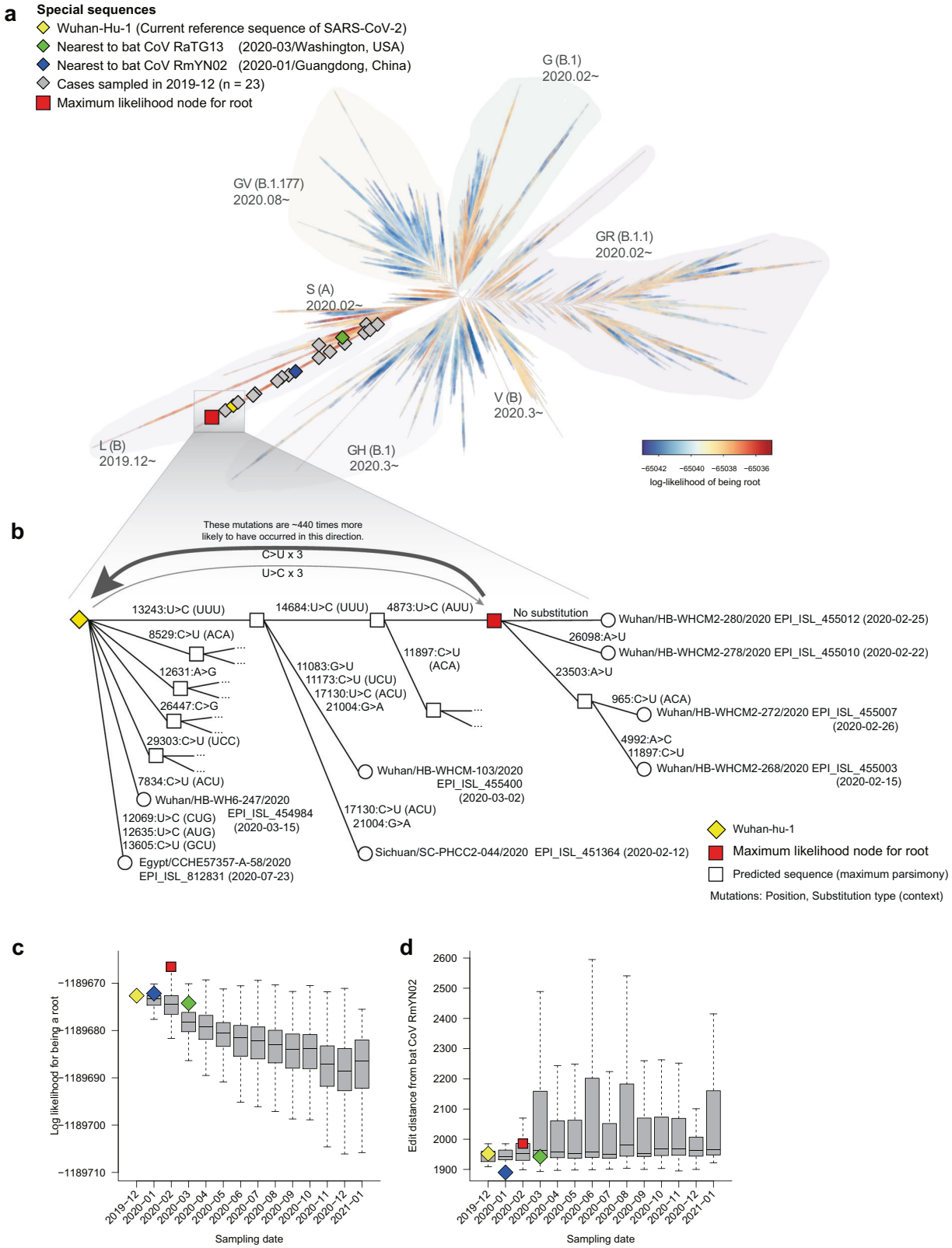

Fig. 2 Tracing the origin of SARS-CoV-2 using its mutational signature. a Phylogenetic tree constructed from 351,525 full-length SARS-CoV-2 genome sequences deposited in the GISAID database. Well-known named clades are indicated with light shades of color and labeled according to widely used nomenclature[75]. Special sequences with prior knowledge: the Wuhan reference sequence (yellow), the sequence nearest to bat CoV RaTG13 (green) or RmYN02 (blue), and sequences sampled in December 2019 (gray) are marked with diamonds, and the presumed root (internal node with the maximum likelihood of being the root) is marked with a red square. The sequence data are constantly being updated on the GISAID website (https://www.gisaid.org/). **b** Detailed phylogenetic structure and associated substitutions between the presumed root (red square) and Wuhan-hu-1. **c** Boxplot showing the log-likelihood of being the root by collection time of the samples. Overall, likelihoods decrease steadily as collection time progresses. **d** Boxplot showing the edit distance from bat CoV RmYN02 by collection time of the samples. Overall, the edit distance does not significantly change over time.

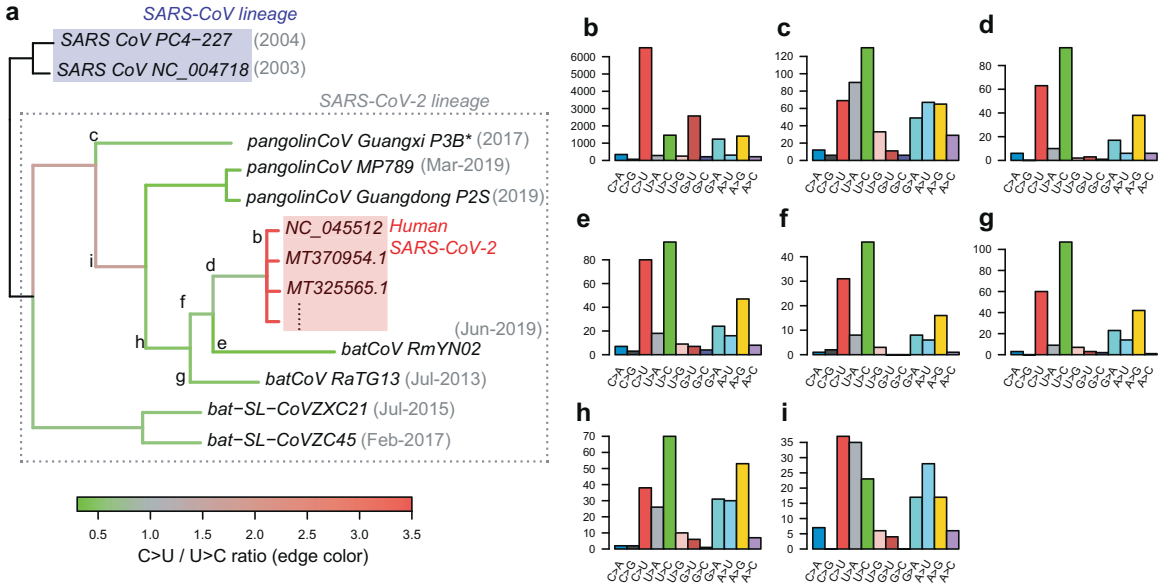

Fig. 3 Changes in C→U preference over the course of SARS-CoV-2 divergence. a Phylogenetic tree of the closest natural relatives of SARS-CoV-2, which was adapted from GISAID. Substitutions between ancestral and descendant sequences were collected in the root-to-leaf direction. The substitution ratio of C→U over U→C is marked by color on each edge (reddish for more C→U and greenish for more U→C). **b**–**i** The substitution spectra for 12 major classes are shown for the spread among humans (**b**) and for each divergence lineage indicated in the middle of branches (**c**–**i**). There are more C→U than U→C changes in human SARS-CoV-2 (**b**), while the reverse pattern is shown in the divergence of bat coronaviruses (**c**–**i**). Although there is uncertainty in the mutation calls for old lineages due to a high degree of mismatch between sequences, one upper branch suggests more C→U than U→C changes (**i**). The status of the ancestral sequence requires the use of information outside the node. The pangolin coronavirus from Guanxi (asterisks) is a consensus sequence of the following six sequences from the same study: P3B, P2V, P5E, P5L, P1E, and P4L. For divergence among these sequences, mutation counts per class and the C→U/U→C ratio are not displayed because their mutation count is not sufficiently large to reliably represent them.

GCU→GUU codon change. Similarly, of the 17,138 threonine to isoleucine (Thr→Ile, 6.4%) changes, 8124 (47.4%) occur by ACU→AUU, congruent with the mutational signature.

The ratio of nonsynonymous to synonymous substitutions, the dN/dS ratio, is a key metric that has frequently been used for inference of natural selection[44]. Currently, the most widely used dN/dS methods are the maximum likelihood method[44] and the counting-based method (e.g., NG86 and YN00[45,46]). These models estimate the ratio of nonsynonymous to synonymous mutations under neutral selection by estimating transition and transversion rate parameters and codon frequencies in given sequences. However, similar to base substitution, when strong asymmetry exists in the mutation spectrum, those codon substitution models may not fully capture the underlying substitutional process. The mutational signature would result in bias in the set of codon changes, and the dN/dS calculation will be skewed unless this is corrected. Here, we calculated the dN/dS ratio, taking into account both the asymmetric mutational signature and codon usage ("Methods"[47]). The adjusted calculation showed that, on average, the overall dN/dS ratio was significantly < 1, indicating that the number of nonsynonymous mutations was lower than that expected under the assumption of neutrality (Fig. 4a). Compared with the widely used Markov chain models, such as the Goldman and Yang model[48] or the Muse and Gaut model[49], our dN/dS method using sequence contexts generally showed similar trends, but some genes showed apparent differences (Fig. 4). The membrane protein (M) had the lowest dN/dS ratio, followed by nonstructural protein 7 (*nsp7*). Seven nonstructural genes, *ORF3a*, *ORF7a*, *ORF7b*, *ORF8*, *ORF9a, ORF9b*, and *ORF10*, by contrast, showed higher dN/dS ratios than other genes, but these values were close to one, suggesting neutral evolution. Collectively, our results indicate that the vast majority of mutations accumulated in the pool of SARS-CoV-2 genomes during transmission would be functionally disadvantageous or neutral.

DISCUSSION

Punctuated viral evolution is a frequently proposed concept in which a virus evolves rapidly over short periods of time followed by long periods of no change[50,51]. Such rapid evolution of new viral strains appears to contradict more clock-like evolution, and this shift is caused by a variety of factors, such as host population changes, migration, and vaccination[50–52]. The SARS-CoV-2 data contain a full spectrum of mutations that were obtained after the virus was transmitted to its new host. However, many researchers use the most widely used general time-reversible (GTR) model to build phylogenetic trees and evaluate mutation rates and selection without considering whether the evolution of SARS-CoV-2 is clock-like or punctuated[53]. Our observation clearly shows that SARS-CoV-2 was introduced to the human population very recently and is not at evolutionary equilibrium.

The C→U and G→U asymmetry in the SARS-CoV-2 and MERS-CoV mutation spectra may be a characteristic of zoonotic RNA viruses recently introduced to human tissues. Previously, linear changes in the base composition over the time of spread were observed in Ebola and influenza viruses[54]. Our results are well aligned with this observation. When zoonotic viruses invade human cells from nonhuman hosts, the human cells are not optimized to provide ideal growth conditions, and the virus base composition is not at equilibrium. In these circumstances, such directional changes in viral genome sequences are possible.

Codon composition is one of the extensively studied genomic properties of many viruses. The codon usage profile of a virus is thought to be passively formed by the mutational pressure exerted on the virus[55–57] or to induce active translational selection itself[30,31]. Because viruses share a tRNA pool with their hosts, codon usage patterns can evolve to become similar to maximize translational efficiency[30], or conversely, when the patterns are too similar, they can become dissimilar to avoid competition[31]. The codon usage profile of SARS-CoV-2 is known to be more similar to that of *Bungarus multicinctus* (snake) or *Rhinolophus sinicus* (bat)

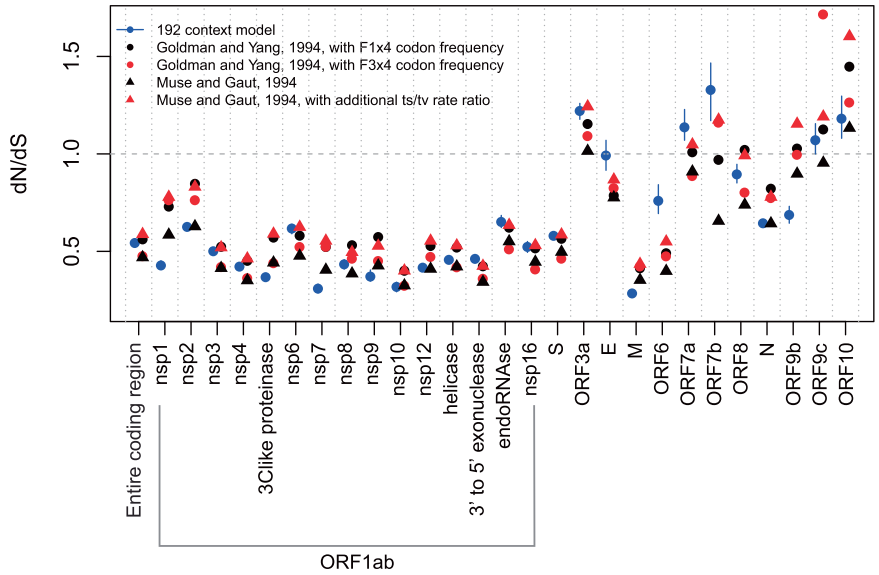

Fig. 4 Assessment of natural selection on SARS-CoV-2 genes. a Across SARS-CoV-2 genes, the ratio of nonsynonymous to synonymous substitutions (dN/dS) was estimated using the 192-context model (Methods). The vertical bar indicates bootstrap variability, reflecting the size of each gene. The dN/dS ratio estimated from the entire coding region is shown as the first tick on the *x*-axis. Circles and triangles indicate the dN/dS ratio calculated based on the Goldman and Yang and the Muse and Gaut models, respectively.

and not similar to that of humans, *Manis javanica* (pangolin)[58], and *Camelus dromedarius* (camel, host of MERS-CoV)[59]. The same mutational spectrum in the noncoding region strongly implies that the mutational spectrum overall is not due to selective pressure on codon usage.

The Markov model-based unrestricted model (UNREST model[60–62]) is a possible method for estimating the root based on the asymmetry of substitutions. However, the UNREST model assumes stationary homogeneous base composition over the time of evolution. This stationarity assumption is frequently reported to be violated in empirical observations, such as continuous decay of the CpG content due to spontaneous cytosine deamination[63,64]. There are Markov-based approaches that allow a nonstationary nonhomogeneous time-irreversible nature[65,66]. However, it would be extremely difficult to apply those methods to our data (351,525 sequences) because of the requirements for large memory and long computing time. Overall, our findings demonstrate the utility of the asymmetric mutational spectrum in searching for the root sequence without imposing information about the sampling date or region[42,43].

Our observation of U→C-dominant mutations in the evolutionary tree of SARS-CoV-2-related viruses suggests differences in the evolutionary characteristics of long-term fixation and those of accelerated (punctuated) evolution. In other words, the genetic diversity of human SARS-CoV-2 has a different mechanistic property from the high divergence of SARS-CoV-2-related viruses of natural origin. Although there is missing information between human SARS-CoV-2 and its natural relatives, making assertions difficult, the predominant C,G→U mutations in human SARS-CoV-2 infections are due to the recent host shift and act as a driving force of the punctuated equilibrium that accelerates viral evolution.

Variations from estimating the dN/dS ratios have been developed to evaluate natural selection on coding sequences. These variations include different ways to estimate nonsynonymous/synonymous mutation counts under the neutral selection assumption. This depends on the belief of how mutations occur and how four bases occupy different codon positions in a given coding sequence[67]. Here, we account for the context of mutations to estimate dN/dS. Our approach has two assumptions. One is that the observed mutation count per context class is sufficiently robust to use the crude count ratio as the actual expected mutation ratio. The other assumption is that the context-specific mutational spectrum is mainly shaped by mutational pressure but not by selection. Further studies aiming to reveal the biochemical mechanism of the C→U- and G→U-dominant mutations in SARS-CoV-2 are required to resolve these assumptions, which will require further validation in functional studies. The context-adjusted estimate of dN/dS is generally in good agreement with previous reports. For example, negative selection is observed for most genes. The diversifying selection on *ORF3* and *ORF8* is more obvious with our context-adjusted method than with conventional models[68,69]. Among the nonstructural protein groups, *Nsp7* showed the lowest dN/dS with the context-adjusted method. It seems to be under the strongest negative selection and thus could be used as a target for antiviral therapeutics[70] or vaccine development[71–74].

In summary, we investigated the mutational signature during the global spread of SARS-CoV-2 using 351,525 genome sequences. A highly asymmetric spectrum was observed, with an exceptionally high proportion of C→U substitutions enriched in a few sequence contexts. We speculate that SARS-CoV-2 is currently in a rapid evolutionary stage prior to reaching static equilibrium. Tracing the outbreak's origin by utilizing the directional substitution pattern indicates that the sequences collected early in Asia are more similar to the original sequence than other sequences are. Despite many nonsynonymous mutations, we observed that the viral genomes were under negative selection.

DATA AVAILABILITY

The phylogenetic tree data are available on our website (https://github.com/ju-lab/SC2_evol_signature).

REFERENCES

1. Arshad Ali, S., Baloch, M., Ahmed, N., Arshad Ali, A. & Iqbal, A. The outbreak of Coronavirus Disease 2019 (COVID-19)—an emerging global health threat. *J. Infect. Public Health* **13**, 644–646 (2020).
2. Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global health concern. *Lancet* **395**, 470–473 (2020).
3. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat. Microbiol.* **5**, 536–544 (2020).

4. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
5. Zhou, H. et al. A novel bat coronavirus closely related to SARS-CoV-2 contains natural insertions at the S1/S2 cleavage site of the spike protein. *Curr. Biol.* **30**, 2196–2203.e3 (2020).
6. Lam, T. T.-Y. et al. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282–285 (2020).
7. Xiao, K. et al. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature* **583**, 286–289 (2020).
8. Zhang, T., Wu, Q. & Zhang, Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr. Biol.* **30**, 1346–1351.e2 (2020).
9. Domingo, E. & Holland, J. J. RNA virus mutations and fitness for survival. *Annu. Rev. Microbiol.* **51**, 151–178 (1997).
10. Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data—from vision to reality. *Eur. Surveill.* **22**, 30494 (2017).
11. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).
12. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
13. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
14. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
15. Baez-Ortega, A. et al. Somatic evolution and global expansion of an ancient transmissible cancer lineage. *Science* **365**, 1–7 (2019).
16. Lassmann, T. Kalign 3: multiple sequence alignment of large data sets. *Bioinformatics* **36**, 1928–1929 (2019).
17. Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).
18. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
19. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
20. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
21. Rosenberg, B. R., Hamilton, C. E., Mwangi, M. M., Dewell, S. & Papavasiliou, F. N. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3′ UTRs. *Nat. Struct. Mol. Biol.* **18**, 230–236 (2011).
22. Asaoka, M., Ishikawa, T., Takabe, K. & Patnaik, S. K. APOBEC3-mediated RNA editing in breast cancer is associated with heightened immune activity and improved survival. *Int. J. Mol. Sci.* **20**, 5621 (2019).
23. Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645–3647 (2017).
24. Greenman, C., Wooster, R., Futreal, P. A., Stratton, M. R. & Easton, D. F. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics* **173**, 2187–2198 (2006).
25. Yang, Z., Ro, S. & Rannala, B. Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics* **165**, 695–705 (2003).
26. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041.e21 (2017).
27. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
28. van Dorp, L. et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Genet. Evol.* **83**, 104351 (2020).
29. Di Giorgio, S., 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**, eabb5813 (2020).
30. Jitobaom, K. et al. Codon usage similarity between viral and some host genes suggests a codon-specific translational regulation. *Heliyon* **6**, e03915 (2020).
31. Chen, F. et al. Dissimilation of synonymous codon usage bias in virus-host coevolution due to translational selection. *Nat. Ecol. Evol.* **4**, 589–600 (2020).
32. Simmonds, P. Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short- and long-term evolutionary trajectories. *mSphere* **5**, e00408–e00420 (2020).
33. Sowden, M., Hamm, J. K. & Smith, H. C. Overexpression of APOBEC-1 results in mooring sequence-dependent promiscuous RNA editing. *J. Biol. Chem.* **271**, 3011–3017 (1996).
34. Seishima, N. et al. Expression and subcellular localisation of AID and APOBEC3 in adenoid and palatine tonsils. *Sci. Rep.* **8**, 918 (2018).
35. Koning, F. A. et al. Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets. *J. Virol.* **83**, 9474–9485 (2009).
36. Wong, E. H. M., Smith, D. K., Rabadan, R., Peiris, M. & Poon, L. L. M. Codon usage bias and the evolution of influenza A viruses. Codon usage biases of influenza virus. *BMC Evol. Biol.* **10**, 253 (2010).
37. Butt, A. M., Nasrullah, I., Qamar, R. & Tong, Y. Evolution of codon usage in Zika virus genomes is host and vector specific. *Emerg. Microbes Infect.* **5**, e107 (2016).
38. B. Miller, J., Hippen, A. A., M. Wright, S., Morris, C. & G. Ridge, P. Human viruses have codon usage biases that match highly expressed proteins in the tissues they infect. *Biomed. Genet. Genomics* **2**, 1–5 (2017).
39. Su, M.-W., Lin, H.-M., Yuan, H. S. & Chu, W.-C. Categorizing host-dependent RNA viruses by principal component analysis of their codon usage preferences. *J. Comput. Biol.* **16**, 1539–1547 (2009).
40. Khandia, R. et al. Analysis of Nipah virus codon usage and adaptation to hosts. *Front. Microbiol.* **10**, 886 (2019).
41. Zhang, Y.-Z. & Holmes, E. C. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell* **181**, 223–227 (2020).
42. Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl Acad. Sci. USA* **117**, 9241–9243 (2020).
43. Mavian, C. et al. Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable. *Proc. Natl Acad. Sci. USA* **117**, 12522–12523 (2020).
44. Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* **15**, 496–503 (2000).
45. Nei, M. & Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).
46. Yang, Z. & Nielsen, R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* **17**, 32–43 (2000).
47. Ju, Y. S. et al. Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *elife* **3**, e02935 (2014).
48. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
49. Muse, S. V. & Gaut, B. S. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* **11**, 715–724 (1994).
50. Nichol, S. T., Rowe, J. E. & Fitch, W. M. Punctuated equilibrium and positive Darwinian evolution in vesicular stomatitis virus. *Proc. Natl Acad. Sci. USA* **90**, 10424–10428 (1993).
51. McCullers, J. A. in *Emerging Infections 10* (eds. Scheld, W. M., Hughes, J. M. & Whitley, R. J.). Ch. 6 (ASM Press, Washington, D.C. 2016).
52. Kerr, P. J. et al. Punctuated evolution of myxoma virus: rapid and disjunct evolution of a recent viral lineage in australia. *J. Virol.* **93**, e01994–18 (2019).
53. Cagliani, R., Forni, D., Clerici, M. & Sironi, M. Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. *Infect. Genet. Evol.* **83**, 104353 (2020).
54. Wada, Y., Wada, K., Iwasaki, Y., Kanaya, S. & Ikemura, T. Directional and reoccurring sequence change in zoonotic RNA virus genomes visualized by time-series word count. *Sci. Rep.* **6**, 36197 (2016).
55. Belalov, I. S. & Lukashev, A. N. Causes and implications of codon usage bias in RNA viruses. *PLoS ONE* **8**, e56642 (2013).
56. Jenkins, G. M. & Holmes, E. C. The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res.* **92**, 1–7 (2003).
57. Khrustalev, V. V., Khrustaleva, T. A., Sharma, N. & Giri, R. Mutational pressure in Zika virus: local ADAR-editing areas associated with pauses in translation and replication. *Front. Cell. Infect. Microbiol.* **7**, 44 (2017).
58. Ji, W., Wang, W., Zhao, X., Zai, J. & Li, X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. *J. Med. Virol.* **92**, 433–440 (2020).
59. Qian, J., Feng, Y. & Li, J. Comments on "Cross-species transmission of the newly identified coronavirus 2019-nCoV". *J. Med. Virol.* **92**, 1437–1439 (2020).
60. Yang, Z. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105–111 (1994).
61. Bettisworth, B. & Stamatakis, A. RootDigger: a root placement program for phylogenetic trees. *BMC Bioinform.* **22.1**, 1–20 (2021).
62. Huelsenbeck, J. P., Bollback, J. P. & Levine, A. M. Inferring the root of a phylogenetic tree. *Syst. Biol.* **51**, 32–43 (2002).
63. Mugal, C. F., Weber, C. C. & Ellegren, H. GC-biased gene conversion links the recombination landscape and demography to genomic base composition: GC-biased gene conversion drives genomic base composition across a wide range of species. *Bioessays* **37**, 1317–1326 (2015).
64. Agashe, D. & Shankar, N. The evolution of bacterial DNA base composition. *J. Exp. Zool. B Mol. Dev. Evol.* **322**, 517–528 (2014).
65. Barry, D. & Hartigan, J. A. Statistical analysis of hominoid molecular evolution. *Stat. Sci.* **2**, 191–207 (1987).
66. Kalaghatgi, P. Phylogeny inference under the general Markov model using MST-backbone. Preprint at https://doi.org/10.1101/2020.06.30.180315 (2020).
67. Yang, Z. *Molecular Evolution: A Statistical Approach*. (Oxford University Press, Oxford, 2014).
68. Velazquez-Salinas, L. et al. Positive selection of ORF3a and ORF8 genes drives the evolution of SARS-CoV-2 during the 2020 COVID-19 pandemic. *Front. Microbiol.* **11**, 550674 (2020).

69. Dearlove, B. et al. A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants. *Proc. Natl Acad. Sci. USA* **117**, 23652–23662 (2020).
70. Ruan, Z. et al. Potential inhibitors targeting RNA-dependent RNA polymerase activity (NSP12) of SARS-CoV-2. Preprint at https://www.preprints.org/manuscript/202003.0024/v1 (2020).
71. Steel, J. et al. Influenza virus vaccine based on the conserved hemagglutinin stalk domain. *MBio* **1**, e00018–10 (2010).
72. Gaschen, B. et al. Diversity considerations in HIV-1 vaccine selection. *Science* **296**, 2354–2360 (2002).
73. Ekiert, D. C. et al. Antibody recognition of a highly conserved influenza virus epitope. *Science* **324**, 246–251 (2009).
74. Staneková, Z. & Varečková, E. Conserved epitopes of influenza A virus inducing protective immunity and their prospects for universal vaccine development. *Virol. J.* **7**, 351 (2010).
75. Rambaut, A. et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407 (2020).

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the researchers from the laboratories responsible for obtaining the specimens or submitting the genomic data and sharing them via GISAID and NCBI GenBank. We thank all the members of Young Seok's laboratory at KAIST and Professors Homin Kim and Gou Young Koh at the Institute for Basic Science (IBS) for productive discussion. This work was supported by KREONET (Korea Research Environment Open NETwork), which is managed and operated by the Korea Institute of Science and Technology Information. This work was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), which was funded by the Ministry of Health & Welfare of Korea (HI16C2387, HI17C1836 to Y.S.J.); Suh Kyungbae Foundation (SUHF-18010082 to Y.S.J.); and National Research Foundation of Korea funded by the Korean government, Ministry of Science and ICT (for Brain Pool Program NRF-2019H1D3A2A02061168 to Y.S.J. and S.Y.K.; Leading Researcher Program NRF-2020R1A3B2078973 to Y.S.J).

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s12276-021-00658-z.

Correspondence and requests for materials should be addressed to Y.S.J.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.