AMERICAN SOCIETY FOR MICROBIOLOGY | mSphere®

RESEARCH ARTICLE
Clinical Science and Epidemiology

Conserved Genomic Terminals of SARS-CoV-2 as Coevolving Functional Elements and Potential Therapeutic Targets

Agnes P. Chan,[a] Yongwook Choi,[a] Nicholas J. Schork[a,b,c]

[a]The Translational Genomics Research Institute (TGen), Phoenix, Arizona, USA
[b]Department of Population Sciences, The City of Hope National Medical Center, Duarte, California, USA
[c]Department of Molecular and Cell Biology, The City of Hope National Medical Center, Duarte, California, USA

Agnes P. Chan and Yongwook Choi contributed equally to this work. Author order was determined alphabetically.

ABSTRACT Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has infected over 40 million people worldwide, with over 1 million deaths as of October 2020 and with multiple efforts in the development and testing of antiviral drugs and vaccines under way. In order to gain insights into SARS-CoV-2 evolution and drug targets, we investigated how and to what extent the SARS-CoV-2 genome sequence differs from those of other well-characterized human and animal coronavirus genomes, as well as how polymorphic SARS-CoV-2 genomes are generally. We ultimately sought to identify features in the SARS-CoV-2 genome that may contribute to its viral replication, host pathogenicity, and vulnerabilities. Our analyses suggest the presence of unique sequence signatures in the 3′ untranslated region (3′-UTR) of betacoronavirus lineage B, which phylogenetically encompasses SARS-CoV-2 and SARS-CoV as well as multiple groups of bat and animal coronaviruses. In addition, we identified genome-wide patterns of variation across different SARS-CoV-2 strains that likely reflect the effects of selection. Finally, we provide evidence for a possible host-microRNA-mediated interaction between the 3′-UTR and human microRNA hsa-miR-1307-3p based on the results of multiple computational target prediction analyses and an assessment of similar interactions involving the influenza A H1N1 virus. This interaction also suggests a possible survival mechanism, whereby a mutation in the SARS-CoV-2 3′-UTR leads to a weakened host immune response. The potential roles of host microRNAs in SARS-CoV-2 replication and infection and the exploitation of conserved features in the 3′-UTR as therapeutic targets warrant further investigation.

IMPORTANCE The coronavirus disease 2019 (COVID-19) outbreak is having a dramatic global effect on public health and the economy. As of October 2020, SARS-CoV-2 has been detected in over 189 countries, has infected over 40 million people, and is responsible for more than 1 million deaths. The genome of SARS-CoV-2 is small but complex, and its functions and interactions with human host factors are being studied extensively. The significance of our study is that, using extensive SARS-CoV-2 genome analysis techniques, we identified potential interacting human host microRNA targets that share similarity with those of influenza A virus H1N1. Our study results will allow the development of virus-host interaction models that will enhance our understanding of SARS-CoV-2 pathogenesis and motivate the exploitation of both the interacting viral and host factors as therapeutic targets.

KEYWORDS SARS-CoV-2, human microRNA, influenza A H1N1, virus 3′ untranslated region, COVID-19

The coronavirus (CoV) disease 2019 (COVID-19) outbreak is having a dramatic effect not only on public health but also on the global economy. The acute respiratory distress associated with severe acute respiratory syndrome CoV-2 (SARS-CoV-2), the

Citation Chan AP, Choi Y, Schork NJ. 2020. Conserved genomic terminals of SARS-CoV-2 as coevolving functional elements and potential therapeutic targets. mSphere 5:e00754-20. https://doi.org/10.1128/mSphere.00754-20.

Editor Benhur Lee, Icahn School of Medicine at Mount Sinai

Address correspondence to Nicholas J. Schork, nschork@tgen.org.

Received 21 July 2020
Accepted 5 November 2020
Published 25 November 2020

pathogen responsible for COVID-19 illness, was first reported in 2019 (1, 2). As of October 2020, SARS-CoV-2 has been detected in over 189 countries, has infected over 40 million people, and has been responsible for more than 1 million deaths (3; Johns Hopkins Coronavirus Resource Center, https://coronavirus.jhu.edu/map.html). The genome of SARS-CoV-2 is small but complex, encoding structural proteins and regulatory elements whose functions and interactions with host factors have been studied extensively (4, 5). However, many of these studies have, justifiably, focused on one or another aspect of the SARS-CoV-2 genome, such as the structural proteins that it encodes (6), its relationships to other viruses (7), and its diversity across the locations in which people have been infected (8). This leaves room for broader, more integrated approaches for the analysis of the SARS-CoV-2 genome focusing on, e.g., noncoding elements, which may yield insights missed by studies with a singular focus.

The SARS-CoV-2 pathogen is a coronavirus, and CoVs are members of the family *Coronaviridae*. *Coronaviridae* are divided into four genera based on phylogeny: alphaCoV, betaCoV, gammaCoV, and deltaCoV. CoVs have been detected in a diverse group of hosts, from humans, wild mammals (e.g., bats, pangolins, camels, civets), and birds to farm animals and poultry (9, 10). The betaCoVs are further divided into four lineages: A, B, C, and D. SARS-CoV-2 belongs to betaCoV lineage B and shares moderate genetic similarity with two human-pathogenic members, SARS-CoV (lineage B, ~79%) and Middle East respiratory syndrome (MERS) CoV (lineage C, ~50%), which were responsible for outbreaks of severe respiratory diseases in humans in 2002 to 2003 and 2012, respectively (11). Unlike SARS-CoV-2, SARS-CoV, or MERS CoV infection, human infection by other CoVs causes mild, common-cold-like symptoms. For example, the pathogens 229E and NL63, which belong to the alphaCoV, and pathogens OC43 and HKU1, which are within betaCoV lineage A, cause mild symptoms in humans. This suggests that genetic differences between SARS-CoV-2 and related viruses may explain its exceptional infectivity, pathogenicity, and elusiveness to effective vaccine and pharmacological mitigation strategies (12, 13).

Many noncoding elements of the SARS-CoV-2 genome have begun to receive attention as potentially informative with respect to the origins and vulnerabilities of the virus. For example, the genomic terminals of CoVs reflect noncoding 5′ and 3′ untranslated regions (5′- and 3′-UTRs) and encode conserved RNA secondary structures that have unique gene regulatory functions, as reviewed by Yang et al. (14). The UTRs are shared by both genomic and subgenomic RNAs and have been suggested to play important roles in viral replication and transcription. The UTRs can also recruit and interact with a range of host and viral protein factors and may provide long-range RNA-RNA or RNA-protein interactions through circularization of the genome. MicroRNAs (miRNAs) are evolutionarily conserved noncoding RNAs which can repress gene expression posttranscriptionally via partial sequence matches primarily to the 3′-UTRs of the target RNAs. In this light, human miRNAs can target viral RNAs and modulate different stages of the viral replication life cycle, positively or negatively (15). An example of human miRNA providing a positive influence on viral replication can be found in the hepatitis C virus (HCV), in which human-liver-specific miR-122 stabilizes the 5′-UTR of HCV, leading to the promotion of viral replication (16). Antisense oligonucleotides acting as inhibitors of miR-122 have been developed as antiviral drugs to reduce viral loads in patients (17). There are also examples of human miRNAs having the opposite effect. For example, a human miRNA showing a negative influence on viral replication (i.e., a positive effect for the host) has been reported for the influenza A virus (IAV) H1N1. Five human miRNAs that are highly expressed in respiratory epithelial cells targeting multiple gene segments have been shown to have inhibitory effects on IAV replication both *in vitro* and *in vivo* (18).

We pursued a systematic gene-by-gene comparative analysis, assessing sequence conservation in each region and element of the SARS-CoV-2 genome, including the 5′- and 3′-UTRs. We did this to see if conservation and polymorphism analyses could identify novel functional elements worth consideration in vaccine and therapeutic

development. We determined whether each of these regions and elements were broadly conserved across the CoV family or unique to sublineages of CoVs. We also identified mutation hot spots, characterized the likely functional significance of naturally occurring amino acid substitutions, and assessed evidence for coevolving mutations across the genome that may impact the stability of the SARS-CoV-2 genome as a whole. Finally, we identified a unique genomic signature residing in an evolutionarily conserved element in the 3′-UTR which may be involved in host miRNA-mediated interactions and innate immunity response. These findings reveal unique viral and host conserved elements associated with the SARS-CoV-2 genome and warrant further investigation into their possible functional roles during infection as well as potential therapeutic targets.

RESULTS

Conserved sequence features of the coronavirus family. To identify conserved and potentially functional features in the CoV family, *Coronaviridae*, we compared each of the annotated genes and UTR features of the SARS-CoV-2 reference genome (NCBI RefSeq genome accession no. NC_045512.2) against 109 selected CoV family genomes (see Table S1 in the supplemental material). The SARS-CoV-2 reference isolate carries 26 processed peptides and open reading frames (ORFs), as well as two UTRs based on NCBI RefSeq annotation. The CoV family genomes that we studied were collected from four coronavirus genera (alpha, beta, gamma, and delta), including seven human CoVs (SARS-CoV-2, SARS-CoV, MERS, OC43, HKU1, 229E, and NL63), a number of mammalian CoVs (e.g., bats, pigs, pangolins, ferrets, and civets), and avian CoVs (e.g., chicken and fowls). The SARS-CoV-2 sequence features were mapped to the CoV family genome sequences through both nucleotide and amino acid sequence alignments using BLAST (19), independently of any CoV family genome annotation (Fig. 1).

The functional element-based conservation analysis results suggested that the 28 total genomic features (i.e., 26 processed peptides and ORFs plus two UTRs) can be broadly classified into two groups, those that were conserved across all CoV genera (cross-CoV feature group) and those that were conserved only within the betaCoV lineage B (betaCoV lineage B-specific feature group), which includes human SARS-CoV-2 and SARS-CoV, and animal CoVs from bats, pangolins, and civets. The cross-CoV feature group showed moderate levels of protein sequence identity across all genera and included nsp3-10, nsp12-16 (RNA-dependent RNA polymerase, helicase, 3′-to-5′ exonuclease, endoribonuclease, and 2′-*O*-ribose methyltransferase), and the structural proteins spike (S), membrane (M), and nucleocapsid (N) (Fig. 1). The betaCoV lineage B-specific feature group mapped uniquely to betaCoV lineage B, with no sequence similarity detected in other genera at the nucleotide or protein sequence level. The betaCoV lineage B-specific feature group included nonstructural proteins nsp2 and nsp11, accessory proteins ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10, the structural envelope (E) protein, and the 5′- and 3′-UTRs (Fig. 1). Among these, the five most conserved features between SARS-CoV-2 and the betaCoV lineage B isolates in descending order of average nucleotide sequence identity were the 3′-UTR, the E gene, ORF10, the 5′-UTR, and nsp10, with 97.4, 95.1, 93.8, 91.1, and 89.7% sequence identity, respectively (Table S2). A short stretch (~30 nucleotides [nt]) of the SARS-CoV-2 3′-UTR also shared high sequence similarity with specific groups of deltaCoVs (from pigs and birds; 97%) and gammaCoVs (from chicken and fowls; 94%) (see the next section). Taken together, these results showed that the nucleotide sequence of both genomic terminals (3′-UTR and 5′-UTR) are exceptionally conserved and unique within the betaCoV lineage B isolates and therefore suggest that they are of likely functional significance for SARS-CoV-2 replication, life cycle, or sustenance.

Notable signatures in the UTRs of SARS-CoVs and related genomes. To investigate the extent of sequence conservation within the genomic terminals of SARS-CoV-2 and related isolates, we performed a multiple-sequence alignment (MSA) analysis on 620 nearly full-length betaCoV lineage B genomes collected from the NCBI Nucleotide database, which included 361 SARS-CoV-2, 113 SARS-CoV, 75 animal CoV (e.g., bats,

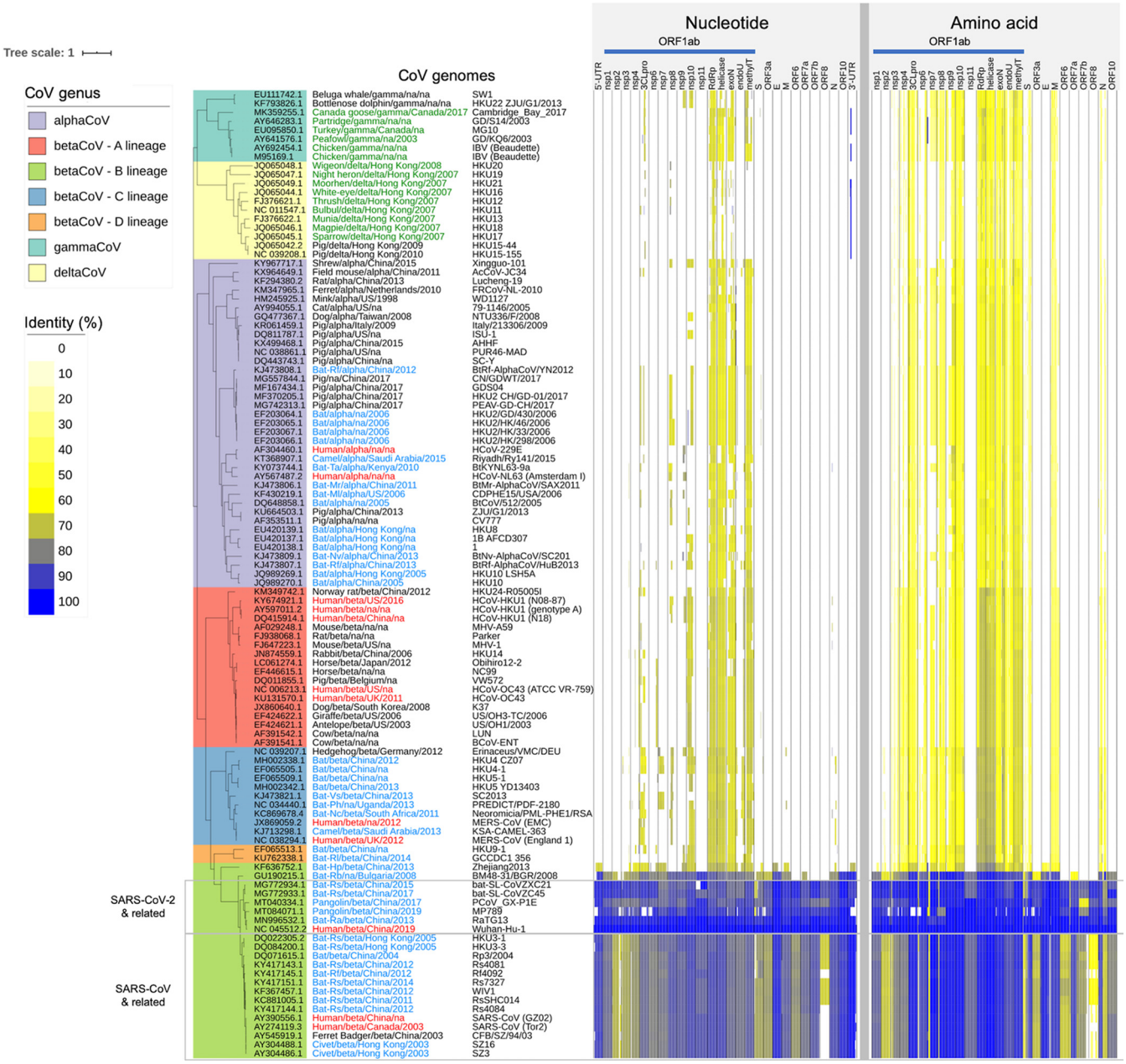


FIG 1 Coronavirus family genome diversity and conserved features. The coronavirus family whole-genome phylogeny, with different genera and sublineages represented, is provided on the left. Each row corresponds to a different coronavirus family member annotated with host, genus, collection location, year, and the isolate name. The CoV names are color coded to indicate host species (red, human; blue, bat, civet, camel; green, bird). The columns on the right correspond to gene products and UTR features along the length of the coronavirus genomes, with each feature normalized to the same column width. The color intensities indicate the degree of nucleotide and amino acid conservation (i.e., sequence identity) with respect to the SARS-CoV-2 reference genome (NCBI RefSeq genome accession no. NC_045512.2).

pangolins, civets), and 71 laboratory isolates (Table S1). The 5′-UTR (SARS-CoV-2, nt 1 to 265) was defined as the 5′ terminus, and both ORF10 and the 3′-UTR together (nt 29558 to 29903) were used for the 3′-terminal analysis. ORF10 was included in the 3′-terminal analysis because ORF10 was a predicted ORF immediately upstream of the 3′-UTR, but no ORF10 expression was detected, as reported in a comprehensive SARS-CoV-2 transcriptome analysis (20). Here, we will refer to the 3′-UTR as a 3′ genomic terminus including both ORF10 and the 3′-UTR, and all genomic coordinates will follow the SARS-CoV-2 reference isolate (NCBI RefSeq genome accession no. NC_045512.2) unless otherwise noted.

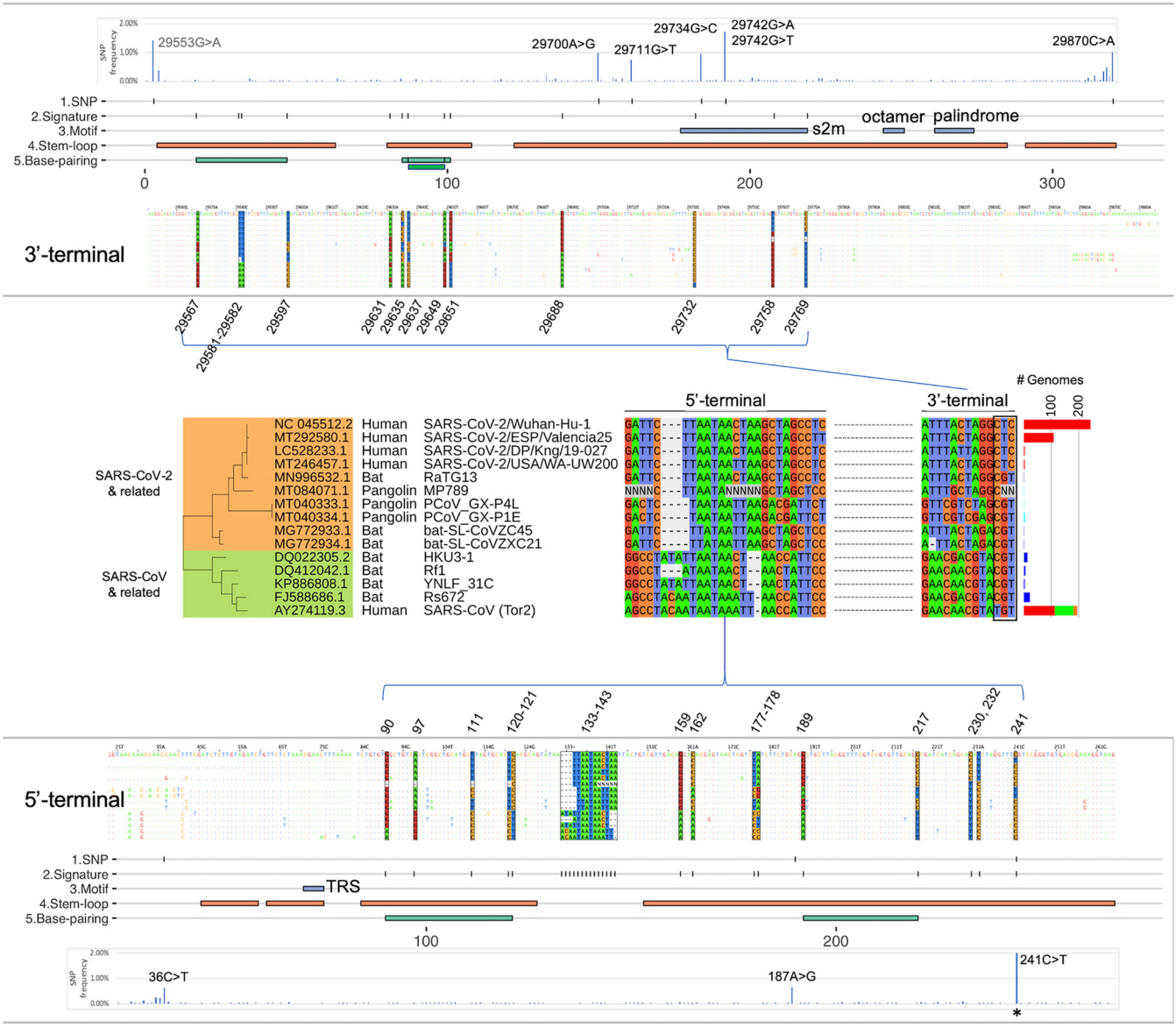


FIG 2 UTR signatures of betaCoV lineage B genomes. Variant positions in the SARS-CoV-2 5′- and 3′-UTRs and their presence in related SARS-CoV genomes (middle section). Base positions are color coded by the four nucleotides and depicted in their genomic locations for the 3′-UTR (top) and 5′-UTR (bottom) sequence coordinates. For each panel, the data tracks are SNV (single-nucleotide polymorphism [SNP]) frequency in SARS-CoV-2 genomes based on 18,599 GISAID genomes analyzed, SNV positions with a >0.5% mutation frequency, UTR signature positions, conserved sequence motifs, predicted stem-loops, and predicted complementary base pairings. The numbers of betaCoV genomes (# Genomes) carrying each unique signature are shown in a bar plot to the right with the following color codes for host species: red, human; blue, bat; green, laboratory; and orange, civet. The 241C>T SNV is indicated with an asterisk (*) and has an observed frequency of 70.2% (outside the frequency scale shown). The 29553G>A SNV is upstream of the 3′ terminus, with no ORF annotation showing a moderately high mutation frequency at 1.42%. s2m, coronavirus 3′ stem-loop II-like motif; TRS, transcription regulatory sequence.

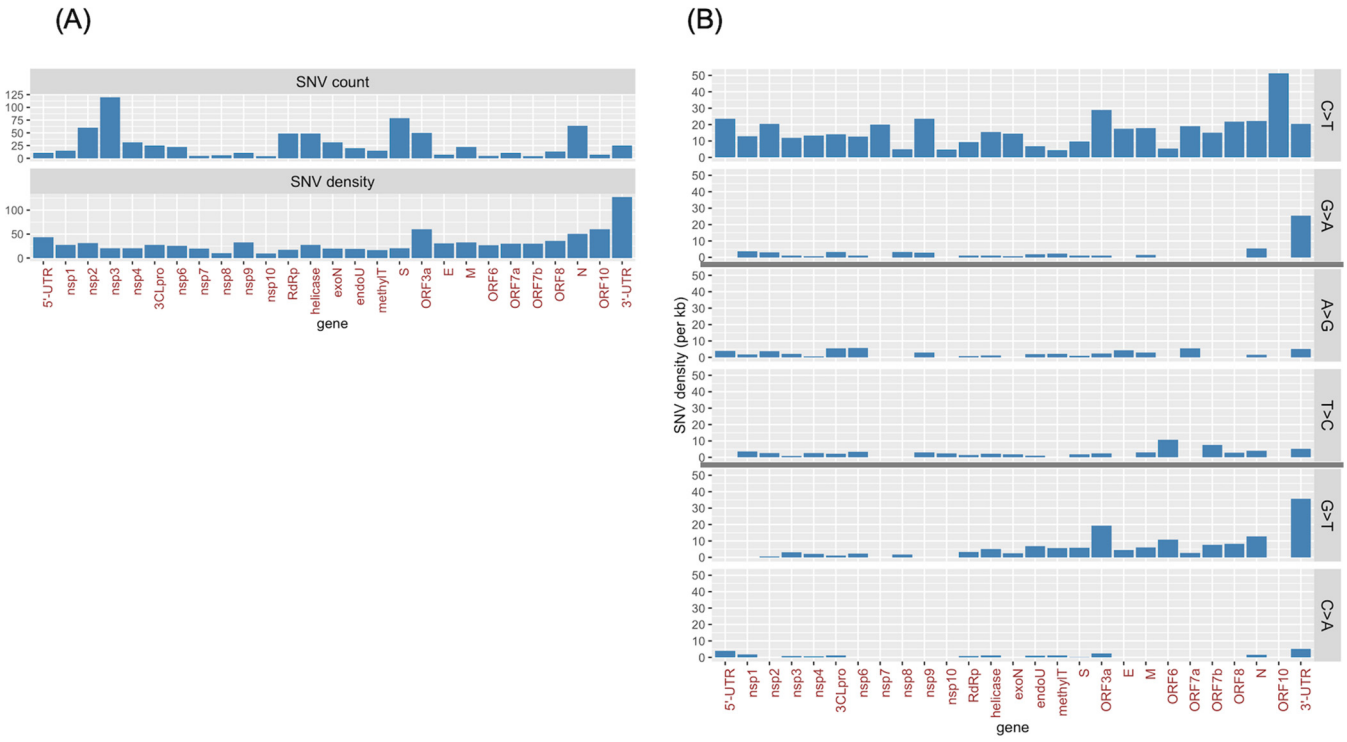
The MSA analysis of the 3′- and 5′-UTRs revealed near-perfect sequence identity of the regions across the betaCoV lineage B genomes. Across the nucleotide positions where most genomes (>99%) have sequence alignments (i.e., ignoring positions near both ends of the genome, where many genomes do not have sequences), 94% of the 3′-UTR positions (234 out of 249) and 84% of the 5′-UTR positions (151 out of 179) exhibited identical nucleotides among 99% of the genomes aligned. Within these conserved regions, a high level of nucleotide diversity was observed at specific positions across the sequence alignments, with 13 and 25 hypervariable positions identified in the 3′- and 5′-UTRs, respectively (Fig. 2). These 38 positions altogether showed distinct nucleotide profiles for subclades of the betaCoV genomes, and we refer to them as the

UTR "signatures." A total of 15 major UTR signatures, as well as their frequency distribution, were identified from the 620 betaCoV genomes (Fig. 2). Based on nucleotide identities, the UTR signatures could be clustered into two distinct groups represented by the SARS-CoV-2 (Wuhan-Hu-1) and SARS-CoV (Tor2) isolates, respectively, which harbored 76% nonidentical nucleotides (29 out of 38 positions at the UTR signature positions). The UTR signature of the SARS-CoV-2 clade was shared by bat CoV isolates (RaTG13, ZC45, and ZXC21) and pangolin CoV isolates (MP789, GX-P4L, and GX-P1E), and that of the SARS-CoV clade was shared by a different group of bat CoVs (HKU3-1, Rf1, YNLF_31C, and Rs672) (Fig. 2).

Overlaying the UTR signatures with predicted RNA secondary structures revealed that a majority of the signature positions (71%; 27 out of 38) were located on stem-loop structures and that 10 positions were involved in complementary base pairings. Interestingly, we noted that the last three positions (nt 29732, 29758, 29769) of the 3′-UTR signature carried distinct nucleotide combinations for each group of the SARS-CoV-2 (CTC), SARS-CoV (TGT), and bat CoV (CGT) isolates (Fig. 2). Notably, these three positions overlapped a conserved RNA motif, S2m (coronavirus 3′ stem-loop II-like motif Rfam RF00164) previously identified in coronavirus and astrovirus (21, 22). In our analysis, the highly conserved S2m RNA element was also detectable using nucleotide searches among avian and animal CoVs belonging to the gamma and delta genera (Fig. 1). In summary, these results show that the 3′- and 5′-UTRs of SARS-CoV-2, SARS-CoV, and bat CoV isolates carry unique signatures involving predicted RNA secondary structures with likely functional and/or regulatory roles.

UTR stability and variant sites within the SARS-CoV-2 genome. To investigate SARS-CoV-2 genomic stability, we analyzed genome-wide nucleotide variants among isolates collected from the ongoing global outbreak. We performed single-nucleotide variant (SNV) discovery by pairwise whole-genome alignments using Nucmer on 18,599 whole-genome sequences available from the GISAID resource (as of 29 May 2020; https://www.gisaid.org) (Fig. S1, Table S3) and a set of stringent filtering criteria to identify high-confidence SNVs (see Materials and Methods). Variant analysis identified 87 variant (SNV) positions, with frequencies of >0.5% (or, equivalently, occurring in at least 93 genomes). Inspection of the UTR signature positions showed that 37 out of 38 positions were relatively stable within SARS-CoV-2 isolates, with variations detected in <0.11% genomes (i.e., 20 isolates or fewer) (Fig. 2). One exception was the variant g.241C>T, which represented one of the signature positions and was originally discovered using 361 SARS-CoV-2 genomes in the betaCoV lineage B analysis above. In this expanded analysis using 18,599 SARS-CoV-2 genomes, the variant g.241C>T was detected at a high prevalence of 70.2%. In addition, six variants were identified at five sites in the 3′-UTR (g.29700A>G, g.29711G>T, g.29734G>C, g.29742G>T, g.29742G>A, g.29870C>A) and three in the 5′-UTR (g.36C>T, g.187A>G, g.241C>T) (Fig. 2). Setting g.241C>T aside, the UTR variants were detected at a low frequency, between 0.62 and 1.05%. A very recent paper by Mishra et al. identified two variant positions corresponding to two found in this analysis in the 5′- and 3′-UTRs, respectively (i.e., g.241C>T, g.29742G>A/T) (23). In our study, all UTR variants were located on predicted stem-loop structures, with the exception of g.36C>T in the 5′-UTR. We note that position 29742 was located within the conserved RNA motif S2m and carried two alternate alleles, making it a triallelic site (Fig. 2; see Discussion). The alternate allele g.29742G>T was observed with a frequency of 1.05%, and the second alternate allele g.29742G>A was observed at a frequency of 0.67%. Based on whole-genome phylogeny analysis, the g.29742G>T and g.29742G>A variants appeared to have arisen in two distinct clades; the g.29742G>T variant was found predominantly in Asia (43% of G>T isolates), and g.29742G>A was almost equally split between Asia and North America (40.0 and 39.5%, respectively, of G>A isolates).

The observed SARS-CoV-2 variants were presumably the result of the evolution of the virus and potential selection pressures on those variants during the pandemic, given their likely functional impact on some aspect of the virus. Imposing a variant



(C)

Gene	Potential SNVs				Observed SNVs			
	Total	Nonsense	Missense	Synonymous	Total	Nonsense	Missense	Synonymous
nsp1	100.0%	4.1%	72.7%	23.2%	100.0%	0.0%	46.7%	53.3%
nsp2	100.0%	4.8%	73.1%	22.1%	100.0%	0.0%	68.3%	31.7%
nsp3	100.0%	4.9%	73.4%	21.7%	100.0%	0.0%	67.5%	32.5%
nsp4	100.0%	4.6%	73.0%	22.5%	100.0%	0.0%	45.2%	54.8%
3CLpro	100.0%	4.4%	73.5%	22.1%	100.0%	0.0%	60.0%	40.0%
nsp6	100.0%	4.8%	73.6%	21.6%	100.0%	0.0%	45.5%	54.5%
nsp7	100.0%	6.2%	71.8%	22.1%	100.0%	0.0%	40.0%	60.0%
nsp8	100.0%	4.3%	74.4%	21.3%	100.0%	0.0%	33.3%	66.7%
nsp9	100.0%	4.7%	71.8%	23.5%	100.0%	0.0%	18.2%	81.8%
nsp10	100.0%	4.2%	73.6%	22.1%	100.0%	0.0%	50.0%	50.0%
nsp11	100.0%	5.1%	70.9%	23.9%	No variants			
RdRp	100.0%	4.9%	74.2%	20.9%	100.0%	0.0%	51.0%	49.0%
helicase	100.0%	5.1%	72.2%	22.8%	100.0%	0.0%	61.2%	38.8%
exoN	100.0%	4.4%	74.5%	21.1%	100.0%	0.0%	48.4%	51.6%
endoU	100.0%	4.8%	74.0%	21.3%	100.0%	0.0%	55.0%	45.0%
MethylT	100.0%	4.6%	74.2%	21.2%	100.0%	0.0%	66.7%	33.3%
S	100.0%	4.5%	73.4%	22.2%	100.0%	0.0%	62.0%	38.0%
ORF3a	100.0%	4.9%	72.8%	22.3%	100.0%	0.0%	68.0%	32.0%
E	100.0%	3.6%	71.0%	25.5%	100.0%	0.0%	42.9%	57.1%
M	100.0%	4.2%	72.4%	23.5%	100.0%	0.0%	36.4%	63.6%
ORF6	100.0%	4.7%	75.8%	19.5%	100.0%	0.0%	60.0%	40.0%
ORF7a	100.0%	4.8%	72.2%	23.0%	100.0%	0.0%	45.5%	54.5%
ORF7b	100.0%	5.4%	74.4%	20.2%	100.0%	0.0%	50.0%	50.0%
ORF8	100.0%	6.0%	73.0%	21.0%	100.0%	7.7%	76.9%	15.4%
N	100.0%	4.5%	72.6%	22.9%	100.0%	0.0%	64.1%	35.9%
ORF10	100.0%	2.9%	76.0%	21.1%	100.0%	0.0%	57.1%	42.9%

FIG 3 SARS-CoV-2 SNV properties. There was a total of 769 SNVs detected at a 0.05% mutation frequency of 18,599 GISAID genomes. (A) SNV counts and density (per kilobase of a feature's length) across genes and UTRs. (B) SNV density is shown by selected base change types: C>T/G>A, A>G/T>C, and G>T/C>A. A full set of SNV distributions across all 12 base change types is shown in Table S4 in the supplemental material. (C) Amino acid mutation bias comparing expected (potential) and observed SNVs for each gene or UTR feature.

frequency threshold of 0.05% or higher (or, equivalently, with the variant occurring in 10 or more genomes) identified 769 SNVs (Table S4). By considering the number of variant positions per kilobase across gene features, we found that both terminal regions (3′-UTR, ORF10, N, and 5′-UTR) and ORF3a harbored the highest number of variant

positions (Fig. 3A). We analyzed two aspects of the 769 SARS-CoV-2 SNVs by classifying them into types of observed base changes (i.e., A>T, A>G, A>C, etc.) and amino acid consequences (i.e., missense, synonymous, and nonsense) across the SARS-CoV-2 genes and UTRs. By assigning SNVs into different base change categories, we observed a predominance of C>T mutations out of all 12 possible base changes. The C>T mutation bias in SARS-CoV-2 has previously been suggested to be associated with human host RNA-editing activities and the subsequent fixation of the edited nucleotides in the viral RNA genome (24). The study by Di Giorgio et al. (24) pointed to C>T/G>A and A>G/T>C variants as base modification outcomes of the human APOBEC and ADAR deaminase family activities, respectively. Results from our gene-by-gene analysis confirmed the study's observations that (i) C>T variants were the most abundant base change across almost all gene features and that (ii) C>T variants were biased toward the positive-sense RNA strand (Fig. 3B). Specifically C>T variants were more abundant than the complementary G>A variants, which would have been the complementary base change if C>T variants were to occur in the negative-sense RNA strand. Importantly, our results further revealed that the two above-mentioned properties did not hold for the 3′-UTR. In the 3′-UTR, we observed that C>T and G>A variants were more or less equally frequent and that G>T instead was the most dominant base change, followed by G>A and C>T. These results may indicate that selection pressure or regulation of the 3′-UTR was different from that of other parts of the genome. In addition, our analyses also detected G>T as the second most prominent base change type when the entire genome was considered. The gene features showing the highest density of G>T mutations were ORF3a, ORF6, the N gene, and the 3′-UTR, all of which were located in the last third of the genome. We determined that the average G>T variant density in the last third of the genome (downstream of ORF1ab) was three times higher than that in the first two-thirds of the genome (entire length of the ORF1ab) (Fig. 3B) (Fisher's exact test, $P = 2.6e{-}09$). In summary, G>T variants are more enriched toward the 3′ end of the genome.

To investigate whether there are any biases in terms of amino acid substitutions (i.e., missense, synonymous, and nonsense), we first determined that if an SNV occurs randomly at any given nucleotide along the genome, the chances that it results in missense, synonymous, and nonsense mutations would be 73, 22, and 5%, respectively. We also determined that such a distribution remained the same across all 26 protein-coding gene features (Fig. 3C). By analyzing the observed proportions of amino acid substitutions of the 769 SNVs, we detected fewer than expected nonsense and missense variants across all genes, with the exception of ORF8. This result likely suggested purifying selection across the protein-coding genes but not on ORF8. Furthermore, we observed that the deviations of the observed proportions from the expected values varied widely across genes (Fig. 3C). In ORF8, for example, the proportions of missense, synonymous, and nonsense variants were 76.9, 15.4, and 7.7%, respectively, which were similar to what we expected. In contrast, for the processed peptide nsp9 (whose putative function is in dimerization and RNA binding), the corresponding proportions were 18.2, 81.8, and 0%, respectively, revealing fewer missense and nonsense variants than expected. These results suggest that there is likely greatly varying selection and evolutionary pressure on individual SARS-CoV-2 genes. In the nonsense amino acid setting, only a single nonsense variant out of the 769 SNVs analyzed was detected. The variant was located in ORF8 (p.Q18*). Previous studies have identified multiple variant forms of ORF8 in SARS-CoV and SARS-CoV-related human and animal isolates (25), including a 29-nt ORF8 deletion variant that had arisen during the late-phase human transmission of SARS-CoV (26). In summary, the characterization of SARS-CoV-2 variants suggests nonrandom selection pressure, may point to undiscovered driving forces of viral genome evolution originating from the hosts or the virus, and may shed light on the identification of mutations with functional or regulatory roles.

Analysis of SARS-CoV-2 variant combinations. We performed linkage disequilibrium (LD) analysis on SNVs from 18,599 GISAID genomes collected in May 2020 using

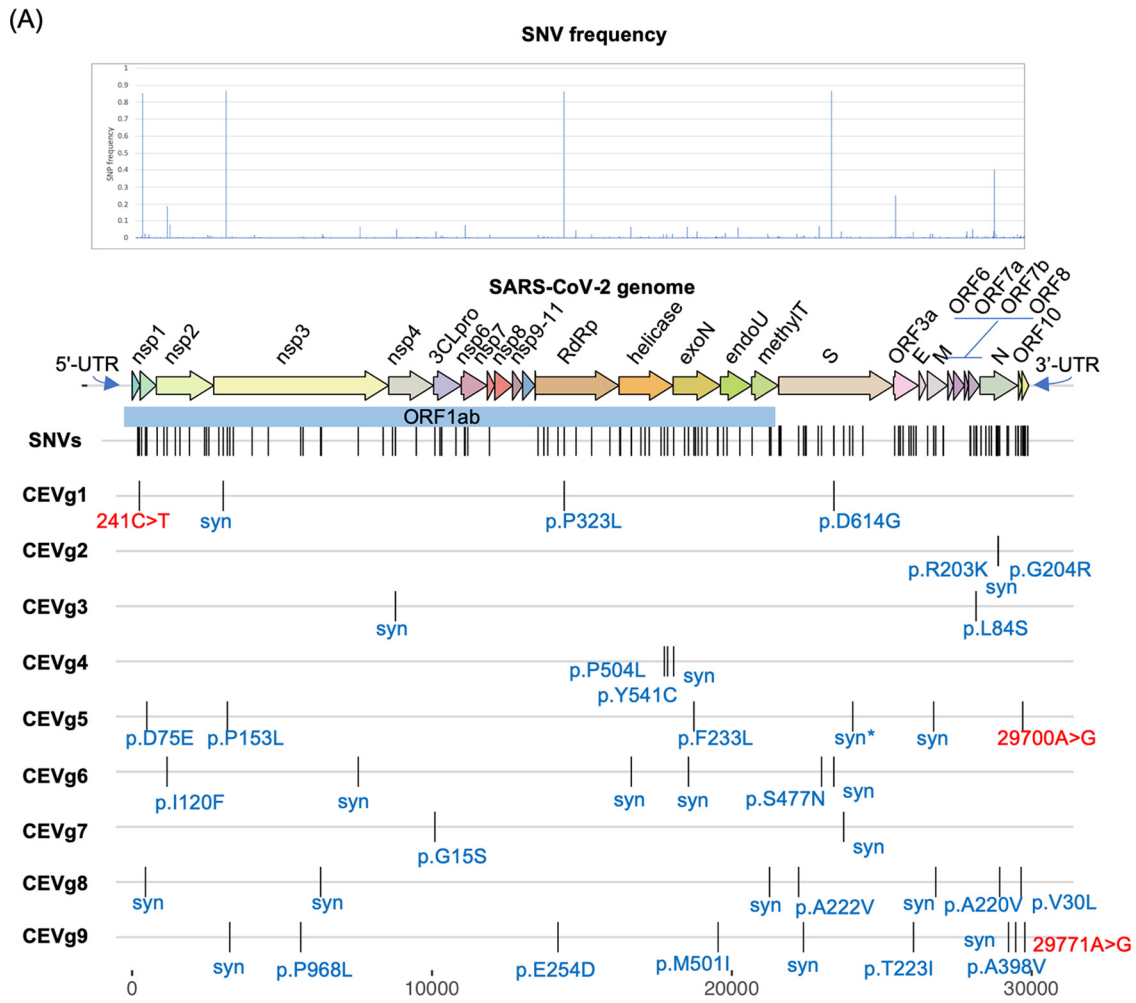


FIG 4 SARS-CoV-2 coevolving SNVs. (A) SNV frequencies are plotted by their positions in the SARS-CoV-2 genome. The relative positions of common SNVs (>0.5%) and 9 representative coevolving variant (CEV) groups and amino acid consequences are shown. (B) Nine representative CEV groups showing different genome frequencies. Three CEV groups involved UTR variants (shown in red). syn, synonymous; RdRP, RNA-dependent RNA polymerase; exoN, 3′-to-5′ exonuclease; endoU, endoRNase; methylT, 2′-*O*-ribose methyltransferase; n.a., not applicable. *, this SNV was associated with other SNVs in CEVg5 in the 29 May 2020 data set but was no longer associated in the 5 October 2020 data set.

Haploview and identified a total of 34 coevolving variant (CEV) groups with a 0.1% or higher genome frequency (Table S3). Notably, we identified two CEV groups that involved the UTRs as well as other gene features, which may motivate testable hypotheses about functional dependencies or interactions of the genomic elements or regions harboring the variants. The first CEV group (CEVg1) was 5′-UTR associated, detected in 69.5% of SARS-CoV-2 genomes, and comprised of four variants that were located in the 5′-UTR (g.241C>T), nsp3 (g.3037C>T, synonymous), the RNA-dependent RNA polymerase (g.14408C>T, p.P323L), and the spike protein (g.23403A>G, p.D614G) (Fig. 4). In terms of geographic distribution by continent, CEVg1 was detected predominantly in South America (88.2%), Africa (86.8%), Europe (79.6%), and North America (66.6%), followed by Oceania (41.6%) and Asia (32.6%) (Fig. S2, Table S4). CEVg1 has shown a dramatic increase from 12.2% to 93.4% between a 3-month period from February to May 2020. The increase of CEVg1 was observed both globally and for each region by continent (Fig. S2). It has been shown that the spike protein D614G mutation, one of the variations implicated in CEVg1, is able to infect human cells more efficiently and therefore enhances transmission (6). Another CEV group (CEVg5) was 3′-

(B)

				As of May 29 (18k genomes)	As of Oct 5 (86k genomes)		
CEV group	Variants	Genic feature	AA change	Genome frequency	Genome frequency	Earliest detection date	Earliest isolate
1	241C>T	5'-UTR	n.a.	69.53%	84.77%	1/25/20	Australia/NSW2153
	3037C>T	nsp3	synonymous				
	14408C>T	RdRp	p.P323L				
	23403A>G	S	p.D614G				
2	28881G>A	N	p.R203K	22.10%	40.20%	2/16/20	England/20168037604
	28882G>A	N	synonymous				
	28883G>C	N	p.G204R				
3	8782C>T	nsp4	synonymous	11.04%	5.03%	12/30/19	Wuhan/IME-WH01
	28144T>C	ORF8	p.L84S				
4	17747C>T	helicase	p.P504L	6.04%	2.24%	2/21/20	USA/WA-S2
	17858A>G	helicase	p.Y541C				
	18060C>T	exoN	synonymous				
5	490T>A	nsp1	p.D75E	0.87%	0.47%	3/3/20	USA/GA_1320
	3177C>T	nsp3	p.P153L				
	18736T>C	exoN	p.F233L				
	24034C>T *	S	synonymous				
	26729T>C	M	synonymous				
	29700A>G	3'-UTR	n.a.				
6	1163A>T	nsp2	p.I120F	n.a.	6.42%	3/19/20	Australia/VIC2165
	7540T>C	nsp3	synonymous				
	16647G>T	helicase	synonymous				
	18555C>T	exoN	synonymous				
	22992G>A	S	p.S477N				
	23401G>A	S	synonymous				
7	10097G>A	3CLpro	p.G15S	n.a.	3.72%	3/2/20	Denmark/SSI-05
	23731C>T	S	synonymous				
8	445T>C	leader	synonymous	n.a.	2.27%	6/20/20	Netherlands/OV-EMC-73
	6286C>T	nsp3	synonymous				
	21255G>C	methylT	synonymous				
	22227C>T	S	p.A222V				
	26801C>G	M	synonymous				
	28932C>T	N	p.A220V				
	29645G>T	ORF10	p.V30L				
9	3256T>C	nsp3	synonymous	n.a.	0.79%	7/28/20	Scotland/QEUH-89C0D4
	5622C>T	nsp3	p.P968L				
	14202G>T	RdRp	p.E254D				
	19542G>T	exoN	p.M501I				
	22388C>T	S	synonymous				
	26060C>T	ORF3a	p.T223I				
	29227G>T	N	synonymous				
	29466C>T	N	p.A398V				
	29771A>G	3'-UTR	n.a.				

FIG 4 (Continued)

UTR associated and detected in 0.9% of the genomes, and it involved six variants that resided in the leader protein or nsp1 (g.490T>A, p.D75E), nsp3 (g.3177C>T, p.P153L), the exonuclease (g.18736T>C, p.F233L), the spike protein (g.24034C>T, synonymous), the membrane protein (g.26729T>C, synonymous), and the 3′-UTR (g.29700A>G) (Fig. 4). CEVg5 was detected in a small proportion of genomes collected in North America (2.4%), Oceania (2.3%), and Europe (0.1%) but not in other regions (Fig. S2, Table S4). CEVg5 remained a minor group in March and April 2020, at 1.2 and 0.53%, respectively.

Three additional CEV groups found in more than 5% of the genomes were identified across gene features among those genomes available as of May 2020 (Fig. 4). The first of these three, CEVg2, was detected entirely within the N protein in 22.1% of the genomes. CEVg2 consisted of three consecutive variants, g.28881G>A, g.28882G>A, and g.28883G>C, which together led to two amino acid substitutions, p.R203K and p.G204R, and the change from one to two positively charged residues. We predicted the functional impact of the two amino acid substitutions (p.R203_G204delinsKR) using

PROVEAN, a prediction tool that we previously developed to determine the likely deleterious impact of amino acid substitutions and indels (i.e., nonsynonymous [Ns] coding variants) on the function of an encoded protein (27). The PROVEAN score of −2.856 suggested a deleterious effect on the protein function as a result of the two amino acid substitutions. These residues were located within a previously identified region (28) referred to as the nucleocapsid linker region (LKR; residues 182 to 247 of SARS-CoV). The LKR was identified as a flexible region joining the N- and C-terminal modular regions and included one of three intrinsically disordered regions found in the N protein; it may be involved in phosphorylation, oligomerization, and N-to-M protein interaction (28). Among the 18,599 SARS-CoV-2 genomes, the N protein also harbored the highest number of SNV counts per gene feature (i.e., 12, including coevolving and single SNVs), of which 8 were found to reside within the LKR. CEVg2 was detected in approximately one-third of the genomes collected in Europe (34.7%) and in South America (28.9%) and was also found in from 3.7 to 14.0% of the genomes in other regions. The prevalence of CEVg2 has increased in Europe (February to May 2020; 31.9 to 58.9%) and South America (February to April 2020; 0 to 36.5%) but has decreased in Asia and Africa (Fig. S2, Table S4).

The second additional CEV group, CEVg3, included two variants located in nsp4 (g.8782C>T, synonymous) and ORF8 (g.28144T>C, p.L84S) and was found in 11.0% of the genomes (Fig. 4). It has previously been reported by other groups (29, 30). CEVg3 showed geographic and temporal profiles different than those described above. CEVg3 appeared predominantly in North America (23.7%), Oceania (18.7%), Asia (17.0%), and other regions and showed a declining trend from 32.3 to 13.4 to 1.3% in January, March, and May, respectively (Fig. S2, Table S4).

The third additional CEV group, CEVg4, consisted of three variants, two in the helicase (g.17747C>T, p.P504L; g.17858A>G, p.Y541C) and one in the exonuclease (g.18060C>T, synonymous), and was detected in 6.0% of genomes (Fig. 4). Both amino acid substitutions in the helicase were predicted to be highly deleterious using PROVEAN (p.P504L score, −8.2; p.Y541C score, −8.9). Most of the genomes harboring CEVg4 SNVs (92%, 1,036 out of 1,124) were detected in North America. The per-month occurrence of CEVg4 decreased from 8.6% in February to 3.3% in April 2020 (Fig. S2, Table S4).

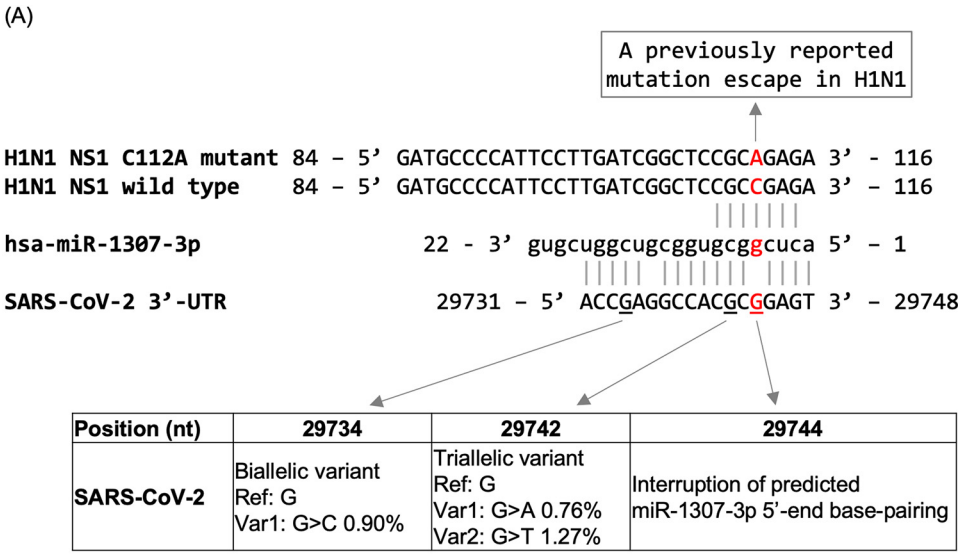
In addition, the processed nsp2 peptide with an unknown function carried the highest number of SNV counts (i.e., 10) after that of nucleocapsid. A moderately prevalent nsp2 mutation was detected in 22.9% of genomes (g.1059C>T, p.T85I), with a predicted deleterious functional outcome (PROVEAN score of −4.09) (Table S4). We also noted that a deletion of three consecutive nucleotides (g.1605_1607delATG), resulting in an amino acid deletion in nsp2 (p.D268del), was predicted to be deleterious (PROVEAN score of −6.370) (Table S4). This deletion of 3 nt, although identified only in a small group of 453 genomes (2.4% global collection), appeared to be highly localized in Europe (95%; 428 out of 453 positive genomes), with only a few occurrences detected in North America (7 genomes) and Oceania (14 genomes). A total of 383 genomes were collected from the following proximal regions: England (124), Netherlands (115), Scotland (102), Northern Ireland (31), and Wales (11). The prevalence of the deletion variant peaked around March in Europe (5.6%) and tapered off in April (2.2%) and May (0.7%) (Fig. S2). In all, our survey of variant positions across 18,599 SARS-CoV-2 genomes collected in May 2020 suggests that coevolving and single variants with likely functional impact on viral fitness or pathogenicity were identified across both the UTRs and functional elements throughout the genome.

In October 2020, over 86,450 high-quality GISAID SARS-CoV-2 genomes became available after our initial analyses were pursued. We have therefore updated our coevolving variant group analysis for the 86,450 genomes during the time that our research was reviewed, which is over four times the size of the first data set of 18,599, analyzed in May 2020 (Table S4, Fig. S3 and S4). A comparison of the frequencies of the CEV groups between the May and October 2020 data sets provided new insights

into the SARS-CoV-2 comutation sites. First, we confirmed the global dominance of CEVg1, which carries the D614G mutation in the spike protein, and observed an increase from 69.53% to 84.77% between May and October 2020. Second, we noted the gradual disappearance (a decrease in genome frequencies) of CEVg3 and CEVg4 around July. Third, we identified two new groups of emerging coevolving mutations (CEVg6 and CEVg8) among other new groups. These two groups showed rapid increases in frequency specifically on only one continent within a short period of time and did not appear on other continents. CEVg6 emerged and increased in Oceania and increased in frequency from 0% in April to 96% in July 2020, whereas CEVg8 in Europe increased in frequency from 0% in June to 36% in September 2020. Interestingly, CEVg6 and CEVg8 each carries a new mutation in the spike protein, S477N and A222V, respectively. The A222V mutation was previously reported in a SARS-CoV-2 strain associated with a confirmed reinfection episode (31).

SARS-CoV-2 UTRs and human miRNAs as potential therapeutic targets. Viral UTRs and human microRNAs have been explored as therapeutic targets in HCV and other viruses because of their essential roles in viral replication and many additional functional phenomena (13). To gain insight into the possible interplay of the SARS-CoV UTRs with host microRNAs in modulating infection pathogenesis, we searched for human miRNAs sharing sequence identity with the UTR sequences of SARS-CoV-2 and SARS-CoV. We used miRNA-specific criteria for BLAST analysis for this purpose (see Materials and Methods) and identified a total of 8 and 7 human microRNAs from the miRBase database (32), including sense and antisense sequences matching the 3′- and 5′-UTRs, respectively (Table S5A). All except one miRNA-matching region (14 out of 15 miRNA regions) were located on predicted stem-loop structures (Fig. S5). Sequence matches to the human miRNAs hsa-miR-1307-3p and hsa-miR-1304-3p were located within the broader conserved RNA motif S2m. In addition to providing BLAST results tuned for miRNA searches, we provide miRNA target prediction results reported from five additional tools, including TargetScan (33), psRNATarget (34), IntaRNA (35), RNA22 (36), and RNAhybrid (37) (Table S5B). These different prediction tools exploit a combination of techniques, from nucleotide sequence-based seed matching and complement matching to structural feature characterization and free energy estimation. For miR-1307-3p, the predicted minimum energy values for RNA-RNA interactions obtained from RNA22, RNAhybrid, and IntaRNA were −31.1, −37.6, and −20.7 kcal/mol, respectively, all below the commonly considered acceptance threshold of −20 kcal/mol (Fig. 5). psRNATarget returned an expectation value (i.e., a penalty for mismatches) of 4, which was below the default and recommended value of 5. TargetScan returned no predictions for miR-1307-3p when considered against the 3′-UTR of SARS-CoV-2, as there is one base mismatch in the middle of the seed region. However, we confirmed that there is a potential interaction between miR-1307-3p and the 3′-UTR by evaluating the target prediction for a 3′-UTR variant (29744G>C). When this base change of interest was introduced at the mismatched position in the wild-type version of the 3′-UTR, a predicted miRNA target of type 7mer-m8 was reported by TargetScan. Furthermore, two recent publications reported results of *in silico* whole-genome scanning of SARS-CoV-2 to identify candidate human miRNA targets (38, 39). Khan et al. (38) applied a combination of three miRNA target prediction tools (IntaRNA, miRanda, psRNATarget) and identified a set of putative miRNAs, including miR-1307-3p for the 3′-UTR. The Khan et al. study provided additional support for a predicted target of human miR-1307-3p in the 3′-UTR of the SARS-CoV-2 genome. Importantly, a previous study of IAV H1N1 provided supporting functional evidence of hsa-miR-1307-3p in mediating antiviral responses and inhibiting viral replication (40). We discuss a possible similar role of human miR-1307-3p in SARS-CoV-2 infection below (see Discussion).

We also examined the endogenous expression of the 15 identified miRNAs using the human miRNA tissue atlas IMOTA (41), which provided categorized miRNA expression levels (i.e., high, medium, low, or not expressed) for 23 human tissues (Table S5C).



(B)

```
SARS-CoV-2 3’-UTR   29730 – 5'    C     G       G    A 3’ – 29749
                                ACCGA GCCACGC GAGU
                                UGGCU CGGUGCG CUCA
hsa-miR-1307-3p       22 – 3' GUGC     G       G      5’ – 1
```

FIG 5 Putative human microRNA miR-1307 interaction with SARS-CoV-2. (A) Predicted base pairings between hsa-miR-1307-3p and the SARS-CoV-2 3′-UTR using blastn and search parameters for miRNAs. The base pairings of miR-1307-3p against the H1N1 NS1 C112A mutant and the H1N1 NS1 wild-type sequences were based on the work of Bavagnoli et al. (40). (B) Predicted miRNA-to-viral RNA interactions based on free energy estimates. RNA22, RNAhybrid, and IntaRNA generated consistent predictions for the RNA-RNA interaction. The prediction output from RNAhybrid is shown.

Among the 8 miRNAs with expression data available, three miRNAs (hsa-miR-1307-3p, hsa-miR-1304-3p, and hsa-miR-15b-5p) were reported to be expressed mostly at medium level in all 23 tissues, including lung, heart, liver, kidney, and small intestine, some of which tissues have been reported to be severely affected during the SARS-CoV-2 infection (42, 43). The expression of miR-1307-3p upon SARS-CoV-2 infection was obtained from the Wyler et al. study (44) using the human lung cell line Calu-3 (GEO accession no. GSE148729). From the raw read count data, we determined the trimmed mean of M (TMM) value-normalized expression levels (45) of miR-1307-3p for mock infection and postinfection to be 362.2 and 485.3 cpm, respectively (Table S5D). The expression level of miR-1307-3p increased slightly by 1.3-fold across 4 to 24 h postinfection compared to that after mock infection. Furthermore, we searched the miRBase database to determine whether the 15 identified human miRNAs were conserved in other organisms. While 6 miRNAs were not detected in other organisms, 9 miRNAs were found in a number of other mammalian species, with the number of organisms ranging from 3 to 25 (Table S5E). The hsa-miR-1307-3p miRNAs, for example, have been found in 12 other mammalian species in various taxonomic orders, such as primates (e.g., orangutan, chimpanzee, baboon, aye-aye), Artiodactyla (e.g., pig, goat, cow), and others (e.g., bat, dog, rabbit, horse, armadillo). SARS-CoV-2 viral sequences have been detected in dogs from households with confirmed human cases, but the dogs remained asymptomatic (46).

In summary, these results suggest that the noncoding UTRs of SARS-CoV-2 are made up of sequences that, based on base pairing, complementarity, and interaction analyses, may interact with microRNAs in humans or other species. Further functional

assays are needed to delineate whether and how microRNAs are involved in the modulation of viral replication and pathogenesis.

DISCUSSION

Our SARS-CoV-2 genome-wide analyses demonstrate that ultraconserved 5′- and 3′-terminal regions of SARS-CoV-2 are shared among betaCoV lineage B genomes, including SARS-CoV and different groups of bat CoVs; however, genome-wide genetic similarity may be as low as ~79%. Notable UTR variant signatures, including complementary base pairing positions with encoded secondary structures, were identified from representative genomes. The high degree of primary sequence conservation of the UTRs identified in this study and the predicted RNA secondary structures reported in two recent studies (47, 48) provide strong evidence for conserved functions of the UTRs in the betaCoV lineage B SARS family of viruses. The likely participation of UTRs in long-distance RNA-RNA and/or RNA-protein interactions involving viral and host factors in the replication of CoVs has been proposed and is consistent with our study results; it therefore deserves greater attention (14).

In addition, our gene-by-gene comparative analysis of the CoV family provided an account of sequence conservation and dissimilarities in both nucleotide and amino acid aspects across each functional unit (processed peptides, ORFs, and UTRs) of the SARS-CoV-2 genome. The CoV family reference genomes were collected from multiple sources, including NCBI RefSeq (49) and previous CoV studies (50, 51), and therefore represent a broad collection of all CoV genera (alpha, beta, gamma, and delta), host species (humans, mammals, and birds), and disease outcomes (human or farm animal outbreaks or mild symptoms). We believe that our genome-wide sequence analysis is complementary to conventional MSA and phylogenetic analyses (e.g., gene tree) (4) or localized window-based analyses (e.g., Simplot) (2), which have been used to assess genome/gene sequence conservation. The cross-CoV conservation data generated in this study will provide the basis for a range of follow-up studies, such as determining the functional significance of highly conserved genes and domains (e.g., the E protein), designing vaccine candidates based on protein or RNA conservation, and developing lineage-specific diagnostic markers for community monitoring and interspecies tracing.

Our analyses also suggest that naturally occurring variants in the SARS-CoV-2 genome sequence were relatively low, with approximately 0.3% of sites exhibiting variations if one imposes a 0.5% or higher mutation frequency threshold. This is consistent with a low mutation rate of the SARS-CoV-2 RNA-dependent RNA polymerase, which likely possesses a proofreading function similar to that of SARS-CoV (52). The observation that the SARS-CoV-2 UTRs harbored higher frequencies of natural variations (3′-UTR, 2.6%; 5′-UTR, 1.2%) than the overall genome-wide mutation rate of 0.3% was likely due to lower evolutionary constraints present in the noncoding UTRs than in genes in the protein-coding regions. A recent report suggesting the influence of human RNA-editing activities on viral genome mutations has provided some explanations for the overall mutation biases that we observed (i.e., the C>T substitution predominance) (24).

Identifying possible therapeutic targets in noncoding regions of a genome has been pursued with other RNA viruses (13), and our investigations suggest possible SARS-CoV-2 UTR interactions with human miRNAs. We used a bioinformatics approach to identify genomic regions sharing strong sequence identity (≥18 nt) to human miRNAs as represented in miRBase (32). Because the mature miRNAs can recognize and bind to a target RNA site through canonical or noncanonical matching positions, our initial analyses used sequence identity as an all-inclusive guiding parameter for the human miRNA screen. We have also attempted to generate predictions from five additional orthogonal miRNA target prediction tools utilizing seed matching, complement matching, structural features, or free energy estimation and included additional supporting evidence for predicted miRNA-virus interactions.

We identified a putative hsa-miR-1307-3p binding site in the 3′-UTR of SARS-CoV-2 with strong sequence identity that exhibits 16 nt of Watson-Crick base pairings out of the first 18 nt of the miRNA (Fig. 5). The putative binding site spanned a conserved RNA motif, S2m, which was also found in the 3′-UTR of subsets of betaCoVs (e.g., SARS-CoV), gammaCoVs (e.g., infectious bronchitis virus from chicken), and deltaCoVs (e.g., birds, pigs). The S2m motif had been previously identified as a conserved element in other CoVs and astrovirus (21, 22). For some of the CoV genomes, due to a lack of high-quality sequences available from the genomic terminals (i.e., nonambiguous bases), the actual frequency or taxonomic distribution of the S2m and other conserved RNA elements present in the UTRs may have been underestimated. Ongoing efforts to collect and whole-genome sequence the repertoire of naturally occurring CoV isolates from wild animals, including bats (53), should help to shed new light on the evolution of CoV functional elements.

Previous studies have associated hsa-miR-1307-3p miRNA with cancer progression as well as lung function. miR-1307 was originally discovered as a novel human miRNA upregulated in Epstein-Barr virus (EBV)-positive nasopharyngeal carcinomas (54) and was also suggested to be associated with the progression of prostate cancer (55). miR-1307 expression has been shown to be dysregulated in newborns with chronic lung disease (56). Importantly, the study by Bavagnoli et al. demonstrated a functional role of miR-1307 in the regulation of viral replication in the influenza A virus H1N1, which was the pathogen responsible for the 2009 H1N1 pandemic (40). Their study predicted sequence complementarity of miR-1307 to H1N1 nonstructural protein 1 (NS1), which functions to limit interferon and proinflammatory responses, thus allowing the virus to evade host innate and adaptive immunity and replicate efficiently in infected cells. The same study also showed that miR-1307 overexpression had regulatory effects on both the virus and host cells. First, miR-1307 overexpression was able to reduce NS1 expression and inhibit wild-type H1N1 replication but had no effects on the NS1 C112A mutant, which carried a nucleotide mismatch to the 5′ region of miR-1307 (Fig. 5). Second, the overexpression of miR-1307 (in a stably transfected lung cell line) was able to induce genes involved in cell proliferation, apoptosis, and the regulation of inflammatory and interferon responses. Taken together, the study concluded that the C112A variant was a viral escape mutation for miR-1307 regulation. Furthermore, the study reported that the C112A mutant was significantly associated with the severe clinical symptom acute respiratory distress syndrome and represented close to one-third of influenza strains that circulated primarily locally in northern Italy during the 2010–2011 influenza season.

In SARS-CoV-2, it is notable that an interruption of base pairings from nt 29744 to the 5th position of the miR-1307-3p sequence coincides with the location of the C112A mutation in H1N1 (Fig. 5). It can be hypothesized that SARS-CoV-2 shares a common host defense mechanism with H1N1, that this mechanism is mediated by host cellular miRNA regulation, and that SARS-CoV-2 carries an allele whose regulation is weakened by human miR-1307 because of the nucleotide mismatch. In support of this hypothesis, our population analysis of SARS-CoV-2 variations identified two nearby mutations at positions 29742 and 29734, which correspond to the 7th and 15th positions of miR-1307, respectively. Mutations that occurred at these two sites may presumably further disrupt the hypothesized base pairings with miR-1307 to escape from binding and inhibition. So far, as of October 2020, the mutations were detected at a low frequency (<1.2%) in the ongoing outbreak. In all, whether SARS-CoV-2 and H1N1 infections have similar host defense mechanisms mediated by host miRNA regulations or whether human population variations of hsa-miR-1307-3p are associated with the severity of clinical symptoms are presently not known and warrant further investigation.

In summary, we utilized a comprehensive genomic analysis approach to assess sequence variations of the SARS-CoV-2 genome with respect to the coronavirus family as well as circulating strains during the current global outbreak collected via the

GISAID repository. We pursued these analyses to gain insights into functional elements within the SARS-CoV-2 viral genome. We identified distinct viral clades sharing coevolving sequence variants and explored emergence and global spread by continent and collection time. We identified possible interactions of the human microRNA miR-1307-3p with the noncoding 3′-UTR of the SARS-CoV-2 genome supported by *in silico* predictions from this study, new analyses from other groups (38, 39), and extensive functional assays that supported a biological role for miR-1307-3p in H1N1 influenza A virus replication (40). Above all, because of the challenges of canonical and noncanonical properties of miRNA binding to targets, we note that important next steps are functional experiments, such as miRNA-virus biochemical interaction assays, mutational analysis, and miRNA overexpression assays to further investigate the biological significance of miR-1307 *in vitro* and *in vivo* during SARS-CoV-2 replication and possibly the regulation of host immune responses. Through this work, we provide evidence for and insights into the possible involvement of miR-1307 in SARS-CoV-2 infection and, consequently, new opportunities for exploring potential targets for antiviral interventions.

MATERIALS AND METHODS

Coronavirus family sequence conservation analysis. The SARS-CoV-2 NCBI RefSeq genome (NC_045512.2) was used as the reference. For gene-by-gene analysis, each sequence of 28 annotated genomic features (ORFs, processed peptides, and UTRs) of SARS-CoV-2 was searched against the 109 representative CoV genomes collected from four genera (alpha, beta, gamma, and delta) (Table S1) using NCBI BLAST+ (blastn and tblastx; v2.9.0), with an E value threshold of 1e–3. The MSA of the 109 CoV family genome sequences was performed using Clustal Omega (v1.2.4) (57). The maximum likelihood phylogeny tree was constructed using RAxML (v8.2.11), with 100 bootstraps under the GTRGAMMA model (58). The tree was visualized using iTOL (59).

SARS-CoV-2 genomic terminal sequences. In the context of this study, the 5′ terminus (nt 1 to 265) corresponded to the annotated 5′-UTR. The 3′ terminus (nt 29558 to 29903), which was also denoted 3′-UTR, corresponded to the annotated ORF10 and 3′-UTR of the SARS-CoV-2 reference genome (NCBI RefSeq genome accession no. NC_045512.2).

Collection of betaCoV lineage B genomes and UTR analysis. A total of 693 betaCoV genome sequences were initially collected from the NCBI Nucleotide database (as of 15 April 2020). Genome sequences were collected using the entire SARS-CoV-2 genome sequence as the query for a blastn search, which required that most of the query sequence length and both UTR regions be aligned sufficiently for sequence comparison (i.e., that at least 85% of the query sequence was covered; an alignment starting from nt 130 or a smaller nucleotide position exists, and an alignment ending at nt 29700 or a higher nucleotide position exists). An MSA was performed on the collected 693 genome sequences, including the SARS-CoV-2 reference genome, using Clustal Omega (v1.2.4). For the 3′- and 5′-UTR regions, variable positions were defined as any positions where 5% or more of the genomes showed nucleotide differences from the reference (excluding ambiguous nucleotides, such as N nucleotides). Positions near either end of the genome (i.e., nucleotides below position 87 or above position 29806) were excluded since over 1% of the genomes do not have aligned sequences and therefore the MSA may not be of high quality. Finally, after the genomes with ambiguous nucleotides in the defined variable positions in UTRs were filtered out, 620 genomes were used as the final genome set for UTR signature analysis. Note that a pangolin CoV (MT084071.1) was included in spite of its having ambiguous nucleotides because it appeared to be one of likely close relatives of SARS-CoV-2 and also carried a unique UTR signature.

Prediction of the UTR secondary structure. RNA secondary structure prediction was performed using the RNAfold Web server (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi) with the default basic option to calculate the minimum free energy (MFE) and partition function. The predicted SARS-CoV-2 5′- and 3′-UTR structures previously reported in reference 4 were used to adjust the prediction.

SARS-CoV-2 variant analysis. A total of 34,217 SARS-CoV-2 genome sequences and their associated metadata were obtained from GISAID (https://www.gisaid.org/) on 29 May 2020. A data sanitization and filtering step was performed, and it included removing gaps (dash and space characters), filtering out genomes from a nonhuman host, and keeping only high-quality genomes (i.e., requiring a genome to be longer than 29 kb and to contain <1% Ns and no other ambiguous nucleotides, such as B and W). Each of the remaining 18,599 high-quality genomes was aligned with the reference genome to identify variants using the nucmer and show-snps functions of the MUMmer package (v3.23) (60). Sequence variants identified within the poly(A) tail or near either end of the sequence (within 10 nt from either end) were ignored. In addition, an MSA of the 18,599 genomes was built using MAFFT (v6.861b), which was used for independent validations of major mutation positions (61). For each sequence variant, the mutation effects on gene products (i.e., genic location and amino acid change, if applicable) were analyzed using in-house scripts. The functional impacts of amino acid substitutions and indels were predicted using PROVEAN (27). Linkage disequilibrium (LD) analysis was performed to identify coevolving variants among SNVs with a frequency of 0.1% or higher using Tagger, implemented in Haploview (v4.2) (62), and using the squared coefficient of correlation (r^2) threshold of 0.8. Non-biallelic sites needed to be excluded from the LD analysis, and a set of 140 genomes with rare mutations on the major mutable sites, causing the sites to become non-biallelic, were also excluded. During the revision of the

manuscript, we repeated the same analyses using an up-to-date (as of 5 October 2020) data set with 135,500 genomes. After the same filtering steps, 86,450 genomes were included for the analyses, and the new findings in the coevolving variants group analysis were also reported.

Protein-coding SNV analysis. Each of the identified protein-coding SNVs was analyzed to determine its amino acid consequence (missense/synonymous/nonsense) using in-house scripts. For the estimation of amino acid consequences under the assumption of random mutations (i.e., to enumerate all potential SNVs given the sequence context of the SARS-CoV-2 genome), all 3 possible SNVs for every nucleotide position on all coding sequences from the start codon to the last codon before the stop codon were included in the analysis.

Identification of putatively interacting human microRNAs. The UTR sequences of SARS-CoV-2 and SARS-CoV were used to search against the miRBase mature RNA sequences (release 22.1) (32) using blastn, with the following parameters set for short sequences: "-penalty -4 -reward 5 -gapopen 25 -gapextend 10 -dust no -soft_masking false." For cross-species conservation analysis of other organisms, we searched the miRBase database with a requirement of 18 or more bases matched with 100% sequence identity. For the additional five miRNA target prediction tools, the results were obtained using the following downloaded scripts or corresponding Web servers with the default parameters: TargetScan, http://www.targetscan.org/vert_72/vert_72_data_download/targetscan_70.zip; psRNATarget, http://plantgrn.noble.org/psRNATarget/analysis?function=3; IntaRNA, http://rna.informatik.uni-freiburg.de/IntaRNA/Input.jsp; RNA22, https://cm.jefferson.edu/rna22/Interactive/; and RNAhybrid, https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid.

Statistical analysis. To test for the significance of the G>T mutation bias toward the 3′ end of the genome, the proportions of G>T mutations out of summed gene lengths were compared between ORF1ab (60 mutations out of 21,326 nt) and the remaining ORFs (66 mutations out of 7,974 nt) using Fisher's exact test implemented in the fisher.test function in the R stats package (v3.6.1).

Data availability. Genome sequence data are available through the NCBI and GISAID.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIF file, 5.7 MB.

FIG S2, TIF file, 18.4 MB.

FIG S3, TIF file, 18.1 MB.

FIG S4, TIF file, 18.4 MB.

FIG S5, TIF file, 10.4 MB.

TABLE S1, XLSX file, 0.05 MB.

TABLE S2, XLSX file, 0.1 MB.

TABLE S3, XLSX file, 13.8 MB.

TABLE S4, XLSX file, 0.5 MB.

TABLE S5, XLSX file, 0.02 MB.

ACKNOWLEDGMENTS

We thank all researchers who have contributed SARS-CoV-2 genome sequences to the GISAID database (https://www.gisaid.org). We also thank Dell, Inc., for making its computing facilities in Austin, TX, available to us. We thank James Lowey and Glen Otero for computational support and Jeff Trent, Paul Keim, Dave Engelthaler, John Altin, Laura Goetz, and the Schork lab for critical feedback.

Aspects of this work were funded in part by NSF grant (FAIN number) 2031819, NIH grants UH2 AG064706, U19 AG023122, U24 AG051129, and U24 AG051129-04S1, Dell, Inc., and the Ivy and Ottosen Foundations.

N.J.S. conceived of the study; N.J.S. and A.P.C. contributed to overall study design; A.P.C., Y.C., and N.J.S. identified the appropriate analytical methods and resources; A.P.C. and Y.C. conducted the analyses; A.P.C., Y.C., and N.J.S. interpreted the results of the analyses; and A.P.C., Y.C., and N.J.S. wrote the paper.

We declare no financial conflicts.

REFERENCES

1. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W, China Novel Coronavirus Investigating and Research Team. 2020. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382:727–733. https://doi.org/10.1056/NEJMoa2001017.
2. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, Si H-R, Zhu Y, Li B, Huang C-L, Chen H-D, Chen J, Luo Y, Guo H, Jiang R-D, Liu M-Q, Chen Y, Shen X-R, Wang X, Zheng X-S, Zhao K, Chen Q-J, Deng F, Liu L-L, Yan B, Zhan F-X, Wang Y-Y, Xiao G-F, Shi Z-L. 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7.
3. Dong E, Du H, Gardner L. 2020. An interactive web-based dashboard to

track COVID-19 in real time. Lancet Infect Dis 20:533–534. https://doi.org/10.1016/S1473-3099(20)30120-1.
4. Chan JF-W, Kok K-H, Zhu Z, Chu H, To KK-W, Yuan S, Yuen K-Y. 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect 9:221–236. https://doi.org/10.1080/22221751.2020.1719902.
5. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, Meng J, Zhu Z, Zhang Z, Wang J, Sheng J, Quan L, Xia Z, Tan W, Cheng G, Jiang T. 2020. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. Cell Host Microbe 27:325–328. https://doi.org/10.1016/j.chom.2020.02.001.
6. Zhang L, Jackson CB, Mou H, Ojha A, Rangarajan ES, Izard T, Farzan M, Choe H. 2020. The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity. bioRxiv https://doi.org/10.1101/2020.06.12.148726.
7. Li X, Giorgi EE, Marichannegowda MH, Foley B, Xiao C, Kong X-P, Chen Y, Gnanakaran S, Korber B, Gao F. 2020. Emergence of SARS-CoV-2 through recombination and strong purifying selection. Sci Adv 6:eabb9153. https://doi.org/10.1126/sciadv.abb9153.
8. Gonzalez-Reiche AS, Hernandez MM, Sullivan MJ, Ciferri B, Alshammary H, Obla A, Fabre S, Kleiner G, Polanco J, Khan Z, Alburquerque B, van de Guchte A, Dutta J, Francoeur N, Melo BS, Oussenko I, Deikus G, Soto J, Sridhar SH, Wang Y-C, Twyman K, Kasarskis A, Altman DR, Smith M, Sebra R, Aberg J, Krammer F, García-Sastre A, Luksza M, Patel G, Paniz-Mondolfi A, Gitman M, Sordillo EM, Simon V, van Bakel H. 2020. Introductions and early spread of SARS-CoV-2 in the New York City area. Science 369:297–301. https://doi.org/10.1126/science.abc1917.
9. Cui J, Li F, Shi Z-L. 2019. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol 17:181–192. https://doi.org/10.1038/s41579-018-0118-9.
10. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, Liu W, Bi Y, Gao GF. 2016. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends Microbiol 24:490–502. https://doi.org/10.1016/j.tim.2016.03.003.
11. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 395:565–574. https://doi.org/10.1016/S0140-6736(20)30251-8.
12. Lim CS, Brown CM. 2017. Know your enemy: successful bioinformatic approaches to predict functional RNA structures in viral RNAs. Front Microbiol 8:2582. https://doi.org/10.3389/fmicb.2017.02582.
13. Angelbello AJ, Chen JL, Childs-Disney JL, Zhang P, Wang Z-F, Disney MD. 2018. Using genome sequence to enable the design of medicines and chemical probes. Chem Rev 118:1599–1663. https://doi.org/10.1021/acs.chemrev.7b00504.
14. Yang D, Leibowitz JL. 2015. The structure and functions of coronavirus genomic 3′ and 5′ ends. Virus Res 206:120–133. https://doi.org/10.1016/j.virusres.2015.02.025.
15. Girardi E, López P, Pfeffer S. 2018. On the importance of host microRNAs during viral infection. Front Genet 9:439. https://doi.org/10.3389/fgene.2018.00439.
16. Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P. 2005. Modulation of hepatitis C virus RNA abundance by a liver-specific microRNA. Science 309:1577–1581. https://doi.org/10.1126/science.1113329.
17. Janssen HLA, Reesink HW, Lawitz EJ, Zeuzem S, Rodriguez-Torres M, Patel K, van der Meer AJ, Patick AK, Chen A, Zhou Y, Persson R, King BD, Kauppinen S, Levin AA, Hodges MR. 2013. Treatment of HCV infection by targeting microRNA. N Engl J Med 368:1685–1694. https://doi.org/10.1056/NEJMoa1209026.
18. Peng S, Wang J, Wei S, Li C, Zhou K, Hu J, Ye X, Yan J, Liu W, Gao GF, Fang M, Meng S. 2018. Endogenous cellular microRNAs mediate antiviral defense against influenza A virus. Mol Ther Nucleic Acids 10:361–375. https://doi.org/10.1016/j.omtn.2017.12.016.
19. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421. https://doi.org/10.1186/1471-2105-10-421.
20. Kim D, Lee J-Y, Yang J-S, Kim JW, Kim VN, Chang H. 2020. The architecture of SARS-CoV-2 transcriptome. Cell 181:914–921.e10. https://doi.org/10.1016/j.cell.2020.04.011.
21. Jonassen CM, Jonassen TO, Grinde B. 1998. A common RNA motif in the 3′ end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus. J Gen Virol 79:715–718. https://doi.org/10.1099/0022-1317-79-4-715.
22. Robertson MP, Igel H, Baertsch R, Haussler D, Ares M, Jr, Scott WG. 2005. The structure of a rigorously conserved RNA element within the SARS virus genome. PLoS Biol 3:e5. https://doi.org/10.1371/journal.pbio.0030005.
23. Mishra A, Pandey AK, Gupta P, Pradhan P, Dhamija S. 2020. Mutation landscape of SARS-CoV-2 reveals three mutually exclusive clusters of leading and trailing single nucleotide substitutions. bioRxiv https://doi.org/10.1101/2020.05.07.082768.
24. Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG. 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv 6:eabb5813. https://doi.org/10.1126/sciadv.abb5813.
25. Wu Z, Yang L, Ren X, Zhang J, Yang F, Zhang S, Jin Q. 2016. ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. J Infect Dis 213:579–583. https://doi.org/10.1093/infdis/jiv476.
26. Oostra M, de Haan CAM, Rottier PJM. 2007. The 29-nucleotide deletion present in human but not in animal severe acute respiratory syndrome coronaviruses disrupts the functional expression of open reading frame 8. J Virol 81:13876–13888. https://doi.org/10.1128/JVI.01631-07.
27. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. 2012. Predicting the functional effect of amino acid substitutions and indels. PLoS One 7:e46688. https://doi.org/10.1371/journal.pone.0046688.
28. Chang C-K, Hou M-H, Chang C-F, Hsiao C-D, Huang T-H. 2014. The SARS coronavirus nucleocapsid protein—forms and functions. Antiviral Res 103:39–50. https://doi.org/10.1016/j.antiviral.2013.12.009.
29. Forster P, Forster L, Renfrew C, Forster M. 2020. Reply to Sánchez-Pacheco et al., Chookajorn, and Mavian et al.: explaining phylogenetic network analysis of SARS-CoV-2 genomes. Proc Natl Acad Sci U S A 117:12524–12525. https://doi.org/10.1073/pnas.2007433117.
30. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, Duan Y, Zhang H, Wang Y, Qian Z, Cui J, Lu J. 2020. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 7:1012–1023. https://doi.org/10.1093/nsr/nwaa036.
31. To KK-W, Hung IF-N, Ip JD, Chu AW-H, Chan W-M, Tam AR, Fong CH-Y, Yuan S, Tsoi H-W, Ng AC-K, Lee LL-Y, Wan P, Tso E, To W-K, Tsang D, Chan K-H, Huang J-D, Kok K-H, Cheng VC-C, Yuen K-Y. 25 August 2020. COVID-19 re-infection by a phylogenetically distinct SARS-coronavirus-2 strain confirmed by whole genome sequencing. Clin Infect Dis https://doi.org/10.1093/cid/ciaa1275.
32. Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. miRBase: from microRNA sequences to function. Nucleic Acids Res 47:D155–D162. https://doi.org/10.1093/nar/gky1141.
33. Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs. Elife 4:e05005. https://doi.org/10.7554/eLife.05005.
34. Dai X, Zhuang Z, Zhao PX. 2018. psRNATarget: a plant small RNA target analysis server (2017 release). Nucleic Acids Res 46:W49–W54. https://doi.org/10.1093/nar/gky316.
35. Mann M, Wright PR, Backofen R. 2017. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. Nucleic Acids Res 45:W435–W439. https://doi.org/10.1093/nar/gkx279.
36. Loher P, Rigoutsos I. 2012. Interactive exploration of RNA22 microRNA target predictions. Bioinformatics 28:3322–3323. https://doi.org/10.1093/bioinformatics/bts615.
37. Krüger J, Rehmsmeier M. 2006. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res 34:W451–W454. https://doi.org/10.1093/nar/gkl243.
38. Khan M-A-K, Sany MRU, Islam MS, Islam A. 2020. Epigenetic regulator miRNA pattern differences among SARS-CoV, SARS-CoV-2, and SARS-CoV-2 world-wide isolates delineated the mystery behind the epic pathogenicity and distinct clinical characteristics of pandemic COVID-19. Front Genet 11:765. https://doi.org/10.3389/fgene.2020.00765.
39. Bartoszewski R, Dabrowski M, Jakiela B, Matalon S, Harrod KS, Sanak M, Collawn JF. 2020. SARS-CoV-2 may regulate cellular responses through depletion of specific host miRNAs. Am J Physiol Lung Cell Mol Physiol 319:L444–L455. https://doi.org/10.1152/ajplung.00252.2020.
40. Bavagnoli L, Campanini G, Forte M, Ceccotti G, Percivalle E, Bione S, Lisa A, Baldanti F, Maga G. 2019. Identification of a novel antiviral micro-RNA targeting the NS1 protein of the H1N1 pandemic human influenza virus and a corresponding viral escape mutation. Antiviral Res 171:104593. https://doi.org/10.1016/j.antiviral.2019.104593.
41. Palmieri V, Backes C, Ludwig N, Fehlmann T, Kern F, Meese E, Keller A. 2018. IMOTA: an interactive multi-omics tissue atlas for the analysis of

human miRNA-target interactions. Nucleic Acids Res 46:D770–D775. https://doi.org/10.1093/nar/gkx701.
42. Shi S, Qin M, Shen B, Cai Y, Liu T, Yang F, Gong W, Liu X, Liang J, Zhao Q, Huang H, Yang B, Huang C. 2020. Association of cardiac injury with mortality in hospitalized patients with COVID-19 in Wuhan, China. JAMA Cardiol 5:802. https://doi.org/10.1001/jamacardio.2020.0950.
43. Su H, Yang M, Wan C, Yi L-X, Tang F, Zhu H-Y, Yi F, Yang H-C, Fogo AB, Nie X, Zhang C. 2020. Renal histopathological analysis of 26 postmortem findings of patients with COVID-19 in China. Kidney Int 98:219–227. https://doi.org/10.1016/j.kint.2020.04.003.
44. Wyler E, Kirstin M, Vedran F, Asija D, Theresa GL, Roberto A, Filippos K, David K, Salah A, Christopher B, Anja R, Ivano L, Andranik I, Tommaso M, Simone DG, Patrick PJ, Alexander MM, Daniela N, Matthias S, Altuna A, Nikolaus R, Christian D, Markus L. 2020. Bulk and single-cell gene expression profiling of SARS-CoV-2 infected human cell lines identifies molecular targets for therapeutic intervention. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
45. Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25. https://doi.org/10.1186/gb-2010-11-3-r25.
46. Sit THC, Brackman CJ, Ip SM, Tam KWS, Law PYT, To EMW, Yu VYT, Sims LD, Tsang DNC, Chu DKW, Perera RAPM, Poon LLM, Peiris M. 2020. Infection of dogs with SARS-CoV-2. Nature 586:776–778. https://doi.org/10.1038/s41586-020-2334-5.
47. Rangan R, Zheludev IN, Hagey RJ, Pham EA, Wayment-Steele HK, Glenn JS, Das R. 2020. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look. RNA 26:937–959. https://doi.org/10.1261/rna.076141.120.
48. Andrews RJ, Peterson JM, Haniff HS, Chen J, Williams C, Grefe M, Disney MD, Moss WN. 2020. An in silico map of the SARS-CoV-2 RNA structurome. bioRxiv https://www.biorxiv.org/content/10.1101/2020.04.17.045161v1.
49. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, et al. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733–D745. https://doi.org/10.1093/nar/gkv1189.
50. Lau SKP, Feng Y, Chen H, Luk HKH, Yang W-H, Li KSM, Zhang Y-Z, Huang Y, Song Z-Z, Chow W-N, Fan RYY, Ahmed SS, Yeung HC, Lam CSF, Cai J-P, Wong SSY, Chan JFW, Yuen K-Y, Zhang H-L, Woo PCY. 2015. Severe acute respiratory syndrome (SARS) coronavirus ORF8 protein is acquired from SARS-related coronavirus from greater horseshoe bats through recombination. J Virol 89:10532–10547. https://doi.org/10.1128/JVI.01048-15.
51. Fu X, Fang B, Liu Y, Cai M, Jun J, Ma J, Bu D, Wang L, Zhou P, Wang H, Zhang G. 2018. Newly emerged porcine enteric alphacoronavirus in southern China: identification, origin and evolutionary history analysis. Infect Genet Evol 62:179–187. https://doi.org/10.1016/j.meegid.2018.04.031.
52. Ogando NS, Ferron F, Decroly E, Canard B, Posthuma CC, Snijder EJ. 2019. The curious case of the nidovirus exoribonuclease: its role in RNA synthesis and replication fidelity. Front Microbiol 10:1813. https://doi.org/10.3389/fmicb.2019.01813.
53. Hu B, Zeng L-P, Yang X-L, Ge X-Y, Zhang W, Li B, Xie J-Z, Shen X-R, Zhang Y-Z, Wang N, Luo D-S, Zheng X-S, Wang M-N, Daszak P, Wang L-F, Cui J, Shi Z-L. 2017. Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog 13:e1006698. https://doi.org/10.1371/journal.ppat.1006698.
54. Zhu JY, Pfuhl T, Motsch N, Barth S, Nicholls J, Grässer F, Meister G. 2009. Identification of novel Epstein-Barr virus microRNA genes from nasopharyngeal carcinomas. J Virol 83:3333–3341. https://doi.org/10.1128/JVI.01689-08.
55. Qiu X, Dou Y. 2017. miR-1307 promotes the proliferation of prostate cancer by targeting FOXO3A. Biomed Pharmacother 88:430–435. https://doi.org/10.1016/j.biopha.2016.11.120.
56. Herrera-Rivero M, Zhang R, Heilmann-Heimbach S, Mueller A, Bagci S, Dresbach T, Schröder L, Holdenrieder S, Reutter HM, Kipfmueller F. 2018. Circulating microRNAs are associated with pulmonary hypertension and development of chronic lung disease in congenital diaphragmatic hernia. Sci Rep 8:10735. https://doi.org/10.1038/s41598-018-29153-8.
57. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75.
58. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.
59. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.
60. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.
61. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010.
62. Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265. https://doi.org/10.1093/bioinformatics/bth457.