TITLE

Synonymous sites in SARS-CoV-2 genes display trends affecting translational efficiency

AUTHORS

Qingsheng Huang[1]*, Huan Gao[1]*, Lingling Zheng[1], Xiujuan Chen[1], Shuai Huang[1], Huiying Liang[1]†

AFFILIATIONS

[1]Institute of Pediatrics, Guangzhou Women and Children's Medical Centre, Guangzhou Medical University, Guangzhou, 510623, China.

*These authors contributed equally to this work.

†Correspondence to: Huiying Liang (lianghuiying(AT)hotmail.com)

ABSTRACT

A novel coronavirus, SARS-CoV-2, has caused a pandemic of COVID-19. The evolutionary trend of the virus genome may have implications for infection control policy but remains obscure. We introduce an estimation of fold change of translational efficiency based on synonymous variant sites to characterize the adaptation of the virus to hosts. The increased translational efficiency of the M and N genes suggests that the population of SARS-CoV-2 benefits from mutations toward favored codons, while the ORF1ab gene has slightly decreased the translational efficiency. In the coding region of the ORF1ab gene upstream of the -1 frameshift site, the decreasing of the translational efficiency has been weakening parallel to the growth of the epidemic, indicating inhibition of synthesis of RNA-dependent RNA polymerase and promotion of replication of the genome. Such an evolutionary trend suggests that multiple infections increased virulence in the absence of social distancing.

INTRODUCTION

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a previously unknown coronavirus, has caused the pandemic of coronavirus disease 2019 (COVID-19) since December 2019(*1-4*). Coronaviruses are positive-sense single-stranded RNA viruses with envelopes. The reference genome of SARS-CoV-2 (NCBI accession NC_045512) has 29,903 nucleotides (nt) containing ORF1ab (also known as NSP, non-structural protein), S (surface or spike), E (envelope), M (membrane or matrix), N (nucleocapsid), and other seven putative open reading frames (ORFs)(*2, 3*). Comparisons between genomes of SARS-CoV-2 and related viruses have revealed many important insights about the origin of SARS-CoV-2(*5-8*), but there are few studies on how the genome of SARS-CoV-2 will evolve in the initial stage of the pandemic, which may have implications for infection control policy (*9*).

Codon usage of viral genes is subject to selection pressures imposed by the gene translation machinery of hosts (*10, 11*). Spontaneous mutation, fine-tuning of translation kinetics, and evasion of immune recognition together shape the evolutionary trend of codon usage (*12*). Traditional methods based on relative synonymous codon usage (RSCU)(*13, 14*) count the frequencies of codons of virus genes and evaluate the adaptation to the hosts. However, the sparsity of single nucleotide variants (SNVs) in genomes of SARS-CoV-2 precludes the traditional methods, which were designed for comparison between distant species, from extracting the subtle trend. Here, we introduce a method based on fold change of translational efficiency (FCTE), which accumulates the effects of synonymous SNV distributed sparsely in the genome of SARS-CoV-2 and characterize the adaptation to hosts in the epidemic in China.

RESULTS

Density of SNV in the coding regions of SARS-CoV-2

In 12 ORFs of SARS-CoV-2, there are on average only 0.021 nonsynonymous sites and 0.012 synonymous sites per 1000 nt per genome (Fig. 1A). The numbers of synonymous

and nonsynonymous sites in a coding region are both correlated with the length of the coding region (Fig. 1B). Compare to the average mutation density, ORF1ab:A, which is the coding region of the ORF1ab gene upstream of the -1 frameshift site, contains slightly more synonymous sites, while ORF1ab:B, which is downstrem of the -1 frameshift site, contains less synonymous sites.

FCTE of synonymous mutation

FCTE measures the effect of mutation on the translation of the residing gene. We estimated FCTE by the fold change of codon usage frequencies of the codon pair before and after a synonymous mutation (Table S1). The codon usage frequency was calibrated by the codon frequency averaged over a repertoire of genes weighted by their expression levels in the type II alveolar (AT2) cells of lung tissue, which is probably the target cells of SARS-CoV-2(*15*). Two topologies of phylogenetic trees of the genomes of SARS-CoV-2 were examined. The first topology is star-like, in which the central ancestor is the consensus sequence (Fig. 2A). The second topology is a maximum likelihood tree rooted at the earliest collected genome, in which a mutation is defined as a pair of codon states in the parent and child nodes (Fig. S1, Fig. 2B).

FCTE of mutations were grouped by the residing coding regions and the effects of mutations on the translation were evaluated. The Wilcoxon tests yield almost the same p values, pseudomedians, and nonparametric 95% confidence intervals for a coding region considered in different topologies, suggesting that FCTE is insensitive to the topologies of the phylogenetic trees (Fig. 2, C and D). Despite of the ORF6 and ORF7b genes which have no synonymous sites, evolutionary stability of translational efficiency of the S, ORF3a, E, ORF7a, ORF8, and ORF10 genes withstands the Wilcoxon tests. For the M and N genes, the pseudomedians of $\log_2$ FCTE are 0.45 and 0.68, respectively, suggesting that during the evolution in the epidemic, the population of SARS-CoV-2 benefits from mutations toward codons favored by the hosts, which possibly accelerate the translation of these genes. In the ORF1ab gene, synonymous sites have evolved toward unfavored codons, especially in the ORF1ab:A fragment.

Evolutionary trends of ORF1ab

To demonstrate the relationship between $\log_2$ FCTE of mutations and the collection times of the genomes in China, we fitted linear models to mutations in the ORF1ab:A and ORF1ab:B fragments and the full-length ORF1ab gene (Fig. 3). The evolutionary trend of the ORF1ab:A fragment is parallel to the growth of the epidemic (Fig. 3A). The ORF1ab:B fragment shows an evolutionary trend opposite to the ORF1ab:A fragment. A summation of ORF1ab:A and ORF1ab:B, however, obscures the evolutionary trend of the full-length ORF1ab. Examined in context of the maximum likelihood tree rooted at the earliest collected genome, the opposite evolutionary trends of ORF1ab:A and ORF1ab:B are even significant, i.e., the p values of the fitted linear models are both below 0.05 and the correlation coefficients get larger (Fig. 3B).

Estimation of FCTE for average lung tissue

We estimated FCTE by codon usage frequencies calibrated for average lung tissue (Table S1). The values of codon usage frequencies are similar to those for the AT2 cells. The FCTE of genes (Fig. S2, A, B, E, and F) and the evolutionary trend of the two fragments of ORF1ab (Fig. S3, A and B) are similar to previous results calibrated for the AT2 cells, except that the statistical tests against the evolutionary tendency of the N gene and the evolutionary trend of ORF1ab:A become non-significant (Fig. S2, E and F, and Fig. S3B). We also estimated FCTE by codon usage frequencies calibrated for human genes without weighting (*16*). Although the values of codon usage frequencies are similar to those for the AT2 cells and for average lung tissue, the statistical tests against the evolutionary tendency of the M and N gene and the evolutionary trend of ORF1ab:A are non-significant (Fig. S2, G and H, and Fig. S3D).

DISCUSSION

Usage of synonymous codons is non-random and non-uniform, the so-called codon bias, the extend of which varies among genes and species (*17*). Codon bias distinguishes expression levels of genes(*13, 18*). Codon usage of the viral genes might be subject to selection pressures imposed by the gene translation machinery of the host, because the

virus relies on ribosomes and tRNAs of the host cells(*12*). For human, genes encoding tRNA of various codons are scattered throughout all but the Y chromosome. Relative tRNA abundance, as a result of the number of tRNA genes and the expression levels of these tRNA genes, have been proved correlated significantly to the codon usage of highly expressed genes in a tissue-specific manner(*19*). Since we do not have the tRNA abundance data for lung tissues, we estimate the change of translational efficiency by the change of codon usage in expressed genes in the AT2 cells and in average lung tissue. Although such an estimation may be even more suitable for evaluation of adaptation of virus genome, because the codon usage frequency also reflects the stability of interactions between mRNA and anticodon in addition to tRNA abundance(*12*), an direct measurement of tRNA abundance may help to further address various effects of translation kinetics.

FCTE in the M and N genes confirm the adaptation of the genome of SARS-CoV-2 to the target tissue (*12, 20*). Translation products of the M and N genes, both the most abundant structural proteins, interact with the genomic RNA intimately in the viral replication and virion packaging (*21, 22*). By contrast, synonymous sites in the S gene suggest no tendency of translational efficiency, which can be explained by the fact that the S protein plays no roles in the replication and only a passive role in the packaging. Selection pressure detected in the M and N genes may nullify previous studies on the virus evolution based on the nonsynonymous-to-synonymous site ratio (Ka/Ks test), though speculations on the evolution of the S gene still holds.

The ORF1ab gene contains a ribosomal -1 frameshift site (Fig. 1A)(*23*). When the frameshifting is urged, the product is the ORF1ab protein, which can be further processed to produce RNA-dependent RNA polymerase (RdRp). The newly generated RdRp protein starts synthesizing the complementary strand from the 3' end of the genomic RNA and disrupts long distance pairing of RNA secondary structure, preventing the frameshifting in later translation(*24*). So the ribosome will encounter a stop codon soon and the product is the ORF1a protein contains no RdRp. By such a mechanism the virus encourages RdRp to replicate the genomic RNA which is free of hindering ribosomes. Although synonymous site may changing the RNA secondary structure(*23*), we observed

only one synonymous site appeared in the stem-loop regions overlapped with the ORF1ab gene in one genome(*20*). A tendency of mutation to rare codons shown by the negative value of $\log_2$ FCTE of the ORF1ab:A fragment (Fig. 2, C and D) imply that SARS-CoV-2 has slowed down the translation and enhanced the -1 frameshifting, producing redundant RdRp and distracting virus replication. According to the SIR (susceptible-infected-recovered) epidemiological model, such a tendency may be evolutionarily beneficial to the virus when the distraction of virus replication assists immune evasion or reduces the mortality rate of the hosts, but is essentially evolutionarily unfavored if there were multiple infections and competition (*25*). In such cases, the genome will be cheated by peers without rare codons in the ORF1ab:A fragment, which produce few RdRp proteins and replicate their genomic RNA with redundant RdRp produced by the cheatee. Parallel to the growth of the epidemic(*4*) which might increase the possibility of multiple infections, the tendency of mutation toward rare codons in the ORF1ab:A region was weakening in the SARS-CoV-2 epidemic in China (Fig. 3B). Hinged by the fine-tuning of FCTE of the ORF1ab:A region, multiple infections break the balance between RdRp production and genome replication, and may drive the evolution of efficient replication. Because high abundance of virus in an infected host increases the virulence and delays the host curing the infection(*25*), prevention against multiple infection seems to be a precondition for taming the virus, may provide a new theoretical basis for social distancing in addition to reducing the infected population (*26*).

MATERIALS AND METHODS

Codon usage frequency in the AT2 cells and in lung

The codon usage frequency was calibrated by the averaged frequency of codons in a repertoire of genes weighted by the expression levels. SARS-CoV-2 mainly infects lung tissue, and probably targets the type II alveolar (AT2) cells(*15*). A dataset of single-cell transcriptome of human lung tissue (GEO: GSE122960)(*27*) was re-analysis with R package Seurat version 3.1.4(*28*) following a previous study(*15*). Briefly, high-quality cells from eight donor lung biopsies were extracted, which have 200 to 6,000 detected genes and their mitochondrial gene content was <10%, and genes detected in less than

three cells were discarded. The gene-cell matrices from eight donor lung biopsies were integrated by SCT normalization method with 2000 anchor genes. By a shared 20-nearest neighbor graph constructed with the first 30 principle components, the cells were clustered at a resolution of 0.15. The AT2 cluster was identified by comparing the marker genes found by the function FindAllMarkers() to marker genes reported by Reyfman *et al*.(*27*) For each AT2 cell, the frequency of codons in high-quality cells were counted according to coding DNA sequences of GRCh38.84 provided by ENSEMBL(*29*). Codon usage frequency in an AT2 cell was calculated by the frequency of codons averaged among genes weighted by the expression levels. Codon usage frequency in all AT2 cells was averaged and used in subsequent analysis.

In addition, we also calculated codon usage frequency for the average lung tissue. A dataset of the expression level of transcripts measured as TPM in lung tissue was downloaded from GTEx portal (version 8)(*30*). The corresponding transcript sequences were obtained from GENCODE version 26(*31*). Codon usage frequency in a biopsy was calculated by the averaged frequency of codons weighted by the expression levels. Codon usage frequency in lung tissue was averaged over 578 biopsies, of which the autolysis score was 0 or 1, the donor was 20 to 59 years old, and the death Hardy Scale was 1 or 2.

Multiple sequence alignment and phylogenetic analysis

We downloaded 623 sequences from GISAID's EpiCoV™ (http://www.gisaid.org, last visit on 13 March 2020, External Data S1), and obtained the sequence and annotations of the reference genome of SARS-CoV-2 from NCBI (accession: NC_045512) (*1*). A multiple sequence alignment (MSA) of 516 genome sequences (sequences of more than 29000 nt) was built using MUSCLE version 3.8.31(*32*) with default setting. The MSA was trimmed, removing nucleotides preceding the first ORF (ORF1ab) and following the last ORF (ORF10). For quality control, we discarded genomes of which the collection date was ambiguous or the sequence contained any gaps or more than one unresolved nucleotide (symbol "N"), which also included genomes collected from pangolins and bats. The refined MSA consists of 317 sequences.

A maximum likelihood phylogenetic tree of the 317 sequences was built using RAxML-NG version 0.9.0(*33*) with GTR+G substitution model and bootstrapping with MRE-based convergence criterion up to 1000 replicates. The tree was essentially unrooted, and we rooted the tree at the earliest collected sample (EPI_ISL_402123) (Fig. S1).

Synonymous triplet sites

Codon triplets on 12 ORFs and two fragments of ORF1ab, the fragments of ORF1ab:A and ORF1ab:B, totally 14 coding regions were examined (Fig. 1). In the 12 ORFs, 366 SNV triplet sites among 317 genomes distribute in the coding regions uniformly (Fig. 1A). The numbers of nonsynonymous and synonymous sites in a coding region are both correlated with the length of the coding region (Fig. 1B).

In columns of the MSA of a coding region, the frequency of codons was counted. For a synonymous triplet site, the states before and after a mutation were defined by two topologies of the phylogenetic tree of the virus genomes. The first topology was star-like, in which all 317 genomes as nodes were individually connected to the root, which was the consensus sequence of the 317 genomes. A pair of states of a synonymous site before and after the mutation was the consensus codon versus the minority codon, which was the codon in a few virus genomes different from the consensus. Identical minority codons in different genomes were counted separately as pairs of states of the mutation. Minority codons that contained any partial ambiguity symbols were neglected. The star-like topology might blur the evolutionary trend, because it counted descendant of a synonymous mutation several times.

The second topology of the phylogenetic tree of the virus genomes was a maximum likelihood tree rooted at the earliest collected genome (Fig. S1). For a synonymous triplet site, the maximum likelihood ancestor states in internal nodes of the tree were reconstructed using R package phangorn version 2.5.5 (*34*). Pairs of states of a synonymous site before and after the mutation were defined for edges whose two ends had different codons, i.e., the sum of absolute difference of reconstructed probability of states of codons was greater than 1. The ancestral and mutational states of the edge were

the codon of the parent and the child nodes, respectively. The number of descendants of the child node indicates the prosperity of the mutation.

Fold change of translational efficiency (FCTE)

FCTE characterizes the tendency of codon usage bias in coding regions. In this study, we estimated FCTE by the fold change of codon usage frequency before and after a synonymous mutation (Fig. 2, A and B). Synonymous sites of a coding region were grouped by the residing coding regions, and we performed a one-sample exact Wilcoxon test to see if translational efficiency of the coding region was evolutionary stable (Fig. 2, C and D).

We plotted $\log_2$ FCTE of mutations in the ORF1ab gene against collection times of the virus genomes. For displaying the parallel development of FCTE of mutations with epidemic (Fig. 3), we extracted from World Health Organization's report on COVID-19 the epidemic curve of clinical diagnosed cases, which include laboratory confirmed cases, in Wuhan (*4*).

In addition to the FCTE estimated by codon usage frequency of AT2 cells, we also calculated the FCTE by codon usage frequency for average lung tissue and for all human genes without weighting (Fig. S2 and S3).

Statistical Analysis

We implemented scripts for finding synonymous triplet sites and calculating FCTE with R version 3.6.2. To evaluate the evolutionary stability of the translational efficiency of a coding region, synonymous sites of the coding region were grouped by the residing coding regions, and a one-sample exact Wilcoxon test was performed to see if the distribution of the logarithm to base 2 ($\log_2$) of FCTE of a region was symmetric about zero. Along with the Wilcoxon tests, pseudomedians and the nonparametric 95% confidence intervals were reported except for the M gene, of which the small sample size precluded the estimation of a 95% confidence interval (Fig. 2, C and D). To demonstrate the evolutionary trend of translational efficiency of the ORF1ab gene, the mutations were

fitted by linear models. We reported the correlation coefficients (R^2) of the linear models and performed *F*-tests to show the overall significance (Fig. 3).

REFERENCES AND NOTES

1. N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao, W. Tan, I. China Novel Coronavirus, T. Research, A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* **382**, 727-733 (2020).
2. F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.-L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E. C. Holmes, Y.-Z. Zhang, A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265-269 (2020).
3. J. F. Chan, S. Yuan, K. H. Kok, K. K. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C. Yip, R. W. Poon, H. W. Tsoi, S. K. Lo, K. H. Chan, V. K. Poon, W. M. Chan, J. D. Ip, J. P. Cai, V. C. Cheng, H. Chen, C. K. Hui, K. Y. Yuen, A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **395**, 514-523 (2020).
4. WHO-China Joint Mission, *Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19)*, (2020).
5. D. Benvenuto, M. Giovanetti, A. Ciccozzi, S. Spoto, S. Angeletti, M. Ciccozzi, The 2019-new coronavirus epidemic: Evidence for virus evolution. *J. Med. Virol.* **92**, 455-459 (2020).
6. C. Ceraolo, F. M. Giorgi, Genomic variance of the 2019 - nCoV coronavirus. *J. Med. Virol.*, (2020).
7. X. Tang, C. Wu, X. Li, Y. Song, X. Yao, X. Wu, Y. Duan, H. Zhang, Y. Wang, Z. Qian, J. Cui, J. Lu, On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.*, (2020).
8. P. Zhou, X. L. Yang, X. G. Wang, B. Hu, L. Zhang, W. Zhang, H. R. Si, Y. Zhu, B. Li, C. L. Huang, H. D. Chen, J. Chen, Y. Luo, H. Guo, R. D. Jiang, M. Q. Liu, Y. Chen, X. R. Shhotmailen, X. Wang, X. S. Zheng, K. Zhao, Q. J. Chen, F. Deng, L. L. Liu, B. Yan, F. X. Zhan, Y. Y. Wang, G. F. Xiao, Z. L. Shi, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273 (2020).
9. P. Forster, L. Forster, C. Renfrew, M. Forster, Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc. Natl. Acad. Sci. U.S.A.*, (2020).
10. N. Stoletzki, A. Eyre-Walker, Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol. Biol. Evol.* **24**, 374-381 (2007).
11. T. Ikemura, Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* **2**, 13-34 (1985).
12. E. H. Wong, D. K. Smith, R. Rabadan, M. Peiris, L. L. Poon, Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evol. Biol.* **10**, 253 (2010).

13. P. M. Sharp, T. M. Tuohy, K. R. Mosurski, Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Res.* **14**, 5125-5143 (1986).
14. P. M. Sharp, W. H. Li, The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281-1295 (1987).
15. Y. Zhao, Z. Zhao, Y. Wang, Y. Zhou, Y. Ma, W. Zuo, Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. *BioRxiv*, (2020).
16. Y. Nakamura, T. Gojobori, T. Ikemura, Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* **28**, 292 (2000).
17. J. M. Comeron, M. Aguade, An evaluation of measures of synonymous codon usage bias. *J. Mol. Evol.* **47**, 268-274 (1998).
18. I. Henry, P. M. Sharp, Predicting gene expression level from codon usage bias. *Mol. Biol. Evol.* **24**, 10-12 (2007).
19. K. A. Dittmar, J. M. Goodenbour, T. Pan, Tissue-specific differences in human transfer RNA expression. *PLoS Genet.* **2**, e221 (2006).
20. C. Polte, D. Wedemeyer, K. E. Oliver, J. Wagner, M. J. C. Bijvelds, J. Mahoney, H. R. de Jonge, E. J. Sorscher, Z. Ignatova, Assessing cell-specific effects of genetic variations using tRNA microarrays. *BMC Genomics* **20**, 549 (2019).
21. C. K. Chang, M. H. Hou, C. F. Chang, C. D. Hsiao, T. H. Huang, The SARS coronavirus nucleocapsid protein--forms and functions. *Antiviral Res.* **103**, 39-50 (2014).
22. H. Jayaram, H. Fan, B. R. Bowman, A. Ooi, J. Jayaram, E. W. Collisson, J. Lescar, B. V. Prasad, X-ray structures of the N- and C-terminal domains of a coronavirus nucleocapsid protein: implications for nucleocapsid formation. *J. Virol.* **80**, 6612-6620 (2006).
23. O. Namy, S. J. Moran, D. I. Stuart, R. J. Gilbert, I. Brierley, A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. *Nature* **441**, 244-247 (2006).
24. J. F. Atkins, G. Loughran, P. R. Bhatt, A. E. Firth, P. V. Baranov, Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. *Nucleic Acids Res.* **44**, 7007-7078 (2016).
25. C. E. Cressler, L. D. Mc, C. Rozins, V. D. H. J, T. Day, The adaptive evolution of virulence: a review of theoretical predictions and empirical tests. *Parasitology* **143**, 915-930 (2016).
26. S. Cobey, Modeling infectious disease dynamics. *Science*, (2020).
27. P. A. Reyfman, J. M. Walter, N. Joshi, K. R. Anekalla, A. C. McQuattie-Pimentel, S. Chiu, R. Fernandez, M. Akbarpour, C. I. Chen, Z. Ren, R. Verma, H. Abdala-Valencia, K. Nam, M. Chi, S. Han, F. J. Gonzalez-Gonzalez, S. Soberanes, S. Watanabe, K. J. N. Williams, A. S. Flozak, T. T. Nicholson, V. K. Morgan, D. R. Winter, M. Hinchcliff, C. L. Hrusch, R. D. Guzy, C. A. Bonham, A. I. Sperling, R. Bag, R. B. Hamanaka, G. M. Mutlu, A. V. Yeldandi, S. A. Marshall, A. Shilatifard, L. A. N. Amaral, H. Perlman, J. I. Sznajder, A. C. Argento, C. T. Gillespie, J. Dematte, M. Jain, B. D. Singer, K. M. Ridge, A. P. Lam, A. Bharat, S. M. Bhorade, C. J. Gottardi, G. R. S. Budinger, A. V. Misharin, Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *Am. J. Respir. Crit. Care Med.* **199**, 1517-1536 (2019).

28. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, 3rd, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. *Cell* **177**, 1888-1902 e1821 (2019).
29. A. Yates, W. Akanni, M. R. Amode, D. Barrell, K. Billis, D. Carvalho-Silva, C. Cummins, P. Clapham, S. Fitzgerald, L. Gil, C. G. Giron, L. Gordon, T. Hourlier, S. E. Hunt, S. H. Janacek, N. Johnson, T. Juettemann, S. Keenan, I. Lavidas, F. J. Martin, T. Maurel, W. McLaren, D. N. Murphy, R. Nag, M. Nuhn, A. Parker, M. Patricio, M. Pignatelli, M. Rahtz, H. S. Riat, D. Sheppard, K. Taylor, A. Thormann, A. Vullo, S. P. Wilder, A. Zadissa, E. Birney, J. Harrow, M. Muffato, E. Perry, M. Ruffier, G. Spudich, S. J. Trevanion, F. Cunningham, B. L. Aken, D. R. Zerbino, P. Flicek, Ensembl 2016. *Nucleic Acids Res.* **44**, D710-716 (2016).
30. The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on Aug. 27, 2019.
31. A. Frankish, M. Diekhans, A. M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. Garcia Giron, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martinez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M. M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigo, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress, P. Flicek, GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766-D773 (2019).
32. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792-1797 (2004).
33. A. M. Kozlov, D. Darriba, T. Flouri, B. Morel, A. Stamatakis, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics (Oxford, England)* **35**, 4453-4455 (2019).
34. K. P. Schliep, phangorn: phylogenetic analysis in R. *Bioinformatics (Oxford, England)* **27**, 592-593 (2011).

ACKNOWLEDGMENTS

General: We acknowledge the researchers originating and submitting the sequences from GISAID's EpiCoV™ on which this study is based (External Data S1).

Funding: This work was supported by National Key R&D Plan (No. 2018YFC1315400.2, to HL).

374 **Author contributions:** QH and HL conceived the study. QH and HG contributed
375 implementation of the methods and data analysis. LZ, XC, and SH contributed the data
376 collection. QH and HG prepared the manuscript. All authors read and approved the
377 manuscript.

378 **Competing interests:** Authors declare no competing interests.

379

FIGURES AND TABLES

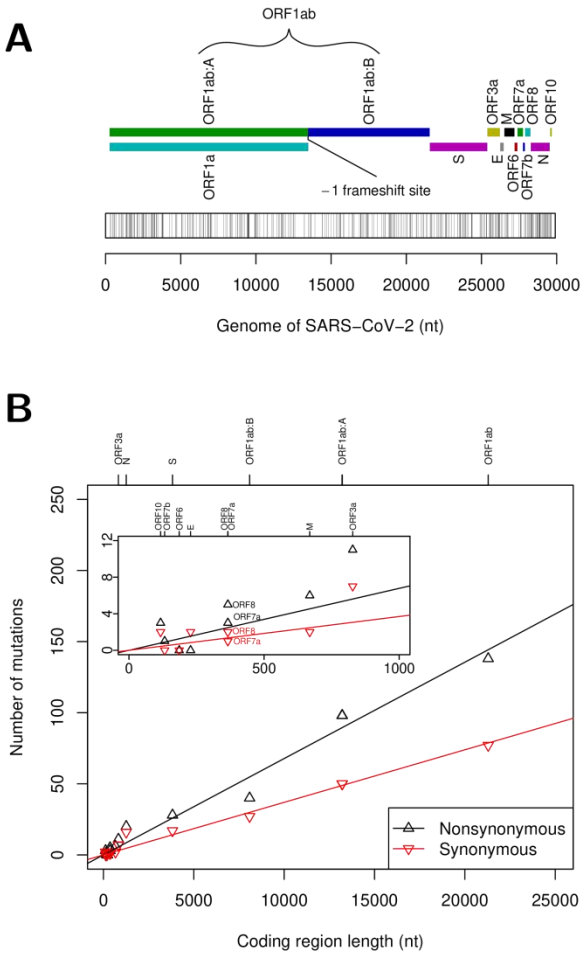
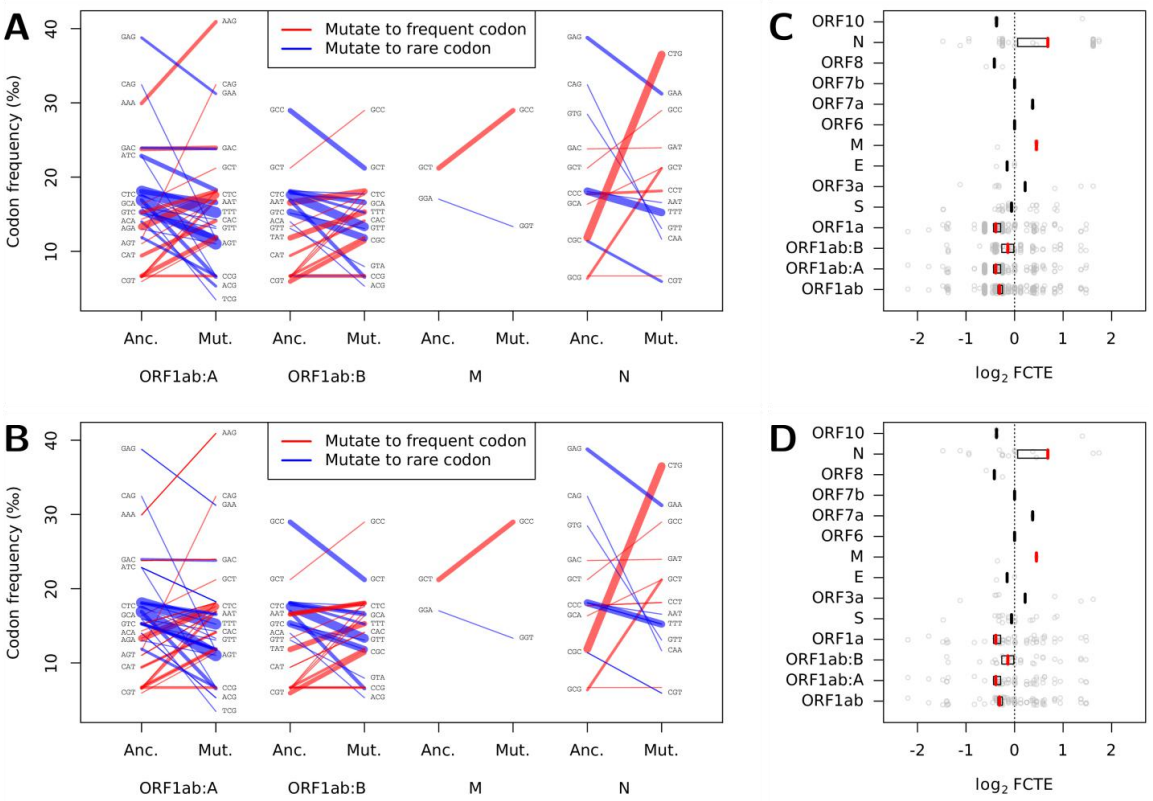


Fig. 1. Variant sites in coding regions of the genome of SARS-CoV-2. (A) Organization of coding regions in the genome and locations of variant triplet sites. **(B)** Numbers of nonsynonymous and synonymous sites in coding regions.



386

387 **Fig. 2. Synonymous sites and fold change of translational efficiency (FCTE). (A)**
388 Pairs of ancestral (Anc.) versus mutational (Mut.) states of synonymous sites for the star-
389 like tree in the ORF1ab:A and ORF1ab:B fragments and the M and N genes. **(B)** Pairs of
390 ancestral (Anc.) versus mutational (Mut.) states of synonymous sites for the maximum
391 likelihood tree rooted at the earliest collected genome in the ORF1ab:A and ORF1ab:B
392 fragments and the M and N genes. The width of a segment is roughly in proportion to the
393 number of the same mutations in different genomes in the star-like tree or to the
394 prosperity of a mutation in the maximum likelihood tree. Some labels of codons are
395 neglected for clarity. **(C)** FCTE for the star-like tree. **(D)** FCTE for the maximum
396 likelihood tree. Gray points indicate $\log_2$ FCTE values of mutations, whose
397 pseudomedians are indicated by vertical bars. For coding regions where the Wilcoxon
398 tests reject the null hypothesis that the $\log_2$ FCTE is zero, the vertical bars are colored red
399 and black boxes are added indicating the nonparametric 95% confidence intervals (except
400 the M gene).

401

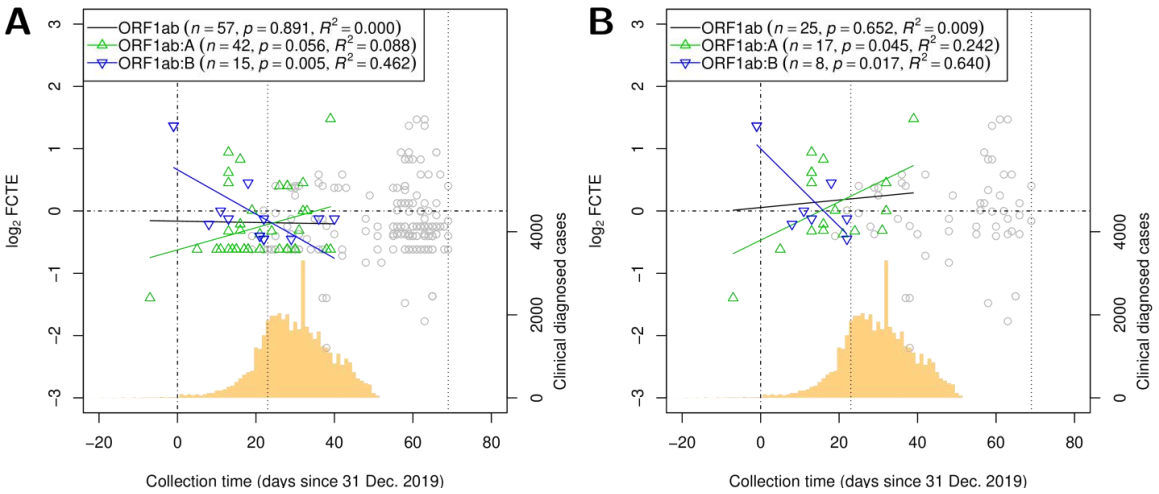
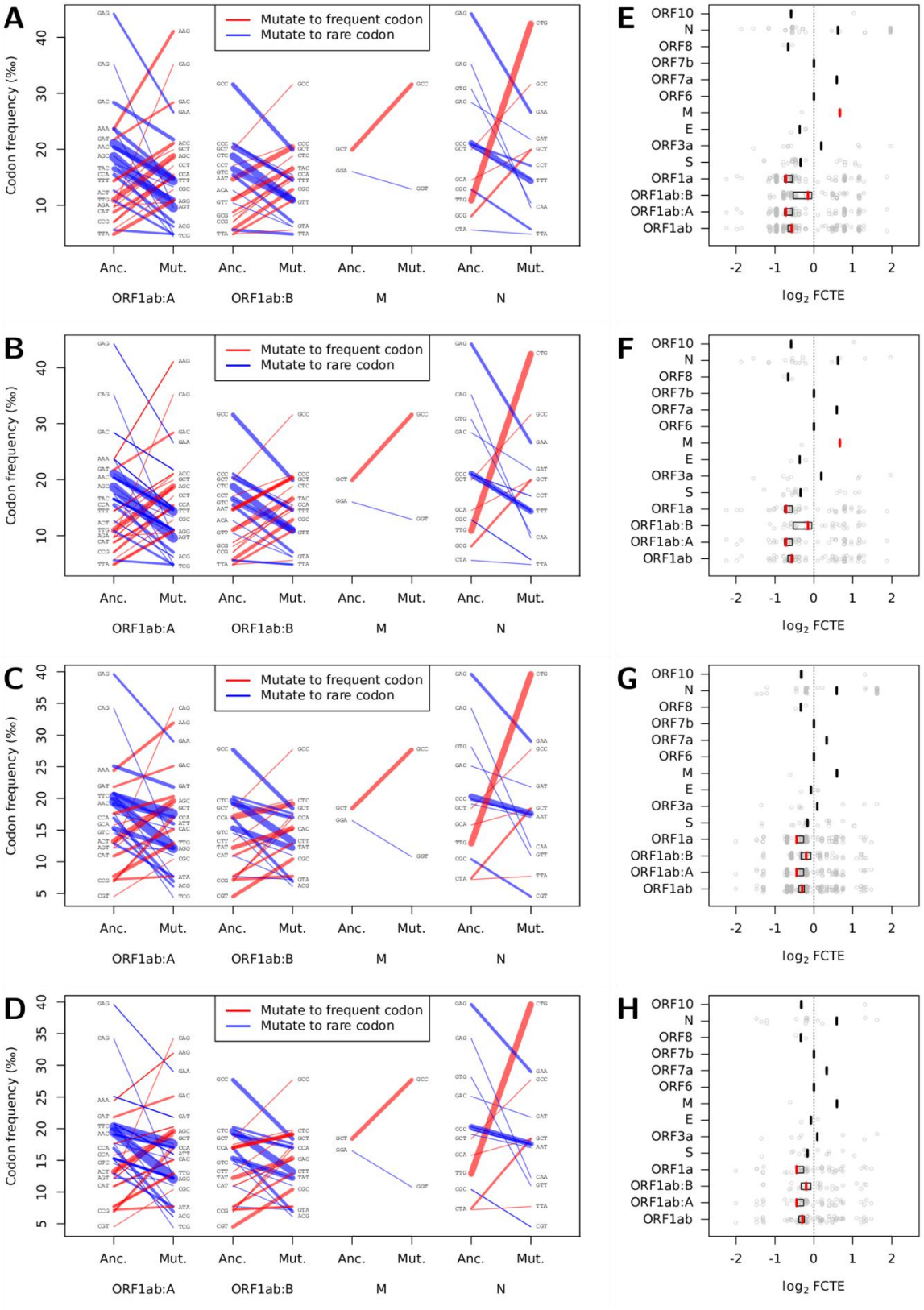


Fig. 3. Evolutionary trends of log_2 fold change of translational efficiency (FCTE) of synonymous sites in the ORF1ab gene. (A) log_2 FCTE for the star-like tree plotted against collection time of the genomes. **(B)** log_2 FCTE for the maximum likelihood tree plotted against collection time of the genomes. Mutation in genomes collected in China are highlighted with green up-pointing triangles for mutations in ORF1ab:A and blue down-pointing triangles for mutations in ORF1ab:B, with fitted lines. Black lines are fitted to both the green and blue triangle. The numbers of mutations (n), p values of the linear models, and correlation coefficients (R2) are shown. Samples collected elsewhere are shown as gray circles. Vertical dotted lines indicate the dates of Wuhan lockdown (23 Jan. 2020) and Italy lockdown (9 March 2020). The backgrounded histogram shows the numbers of clinical diagnosed cases in Wuhan reported by WHO (*4*).

SUPPLEMENTARY MATERIALS

416

Fig. S1. The maximum likelihood tree of 317 virus genomes rooted at the earliest collected virus (EPI_ISL_402123). Accession ID from GISAID's EpiCoV™ database is colored black if the virus was collected before 31 Dec. 2019, red if between 31 Dec. 2019 and 23 Jan. 2020 (the date of Wuhan lockdown), green if between 23 Jan. 2020 and 9 March 2020 (Italy lockdown), and blue if after 9 March 2020.



423

Fig. S2. Synonymous sites and fold change of translational efficiency (FCTE) with gene expression levels calibrated for lung tissue and without tissue specificity. (A) Pairs of ancestral (Anc.) versus mutational (Mut.) states of synonymous sites for the star-like tree in the ORF1ab:A and ORF1ab:B fragments and the M and N genes with gene expression levels calibrated for lung tissue. **(B)** Pairs of ancestral (Anc.) versus mutational (Mut.) states of synonymous sites for the maximum likelihood tree rooted at the earliest collected genome in the ORF1ab:A and ORF1ab:B fragments and the M and N genes with gene expression levels calibrated for lung tissue. **(C)** Pairs of ancestral (Anc.) versus mutational (Mut.) states of synonymous sites for the star-like tree in the ORF1ab:A and ORF1ab:B fragments and the M and N genes without tissue specificity. **(D)** Pairs of ancestral (Anc.) versus mutational (Mut.) states of synonymous sites for the maximum likelihood tree rooted at the earliest collected genome in the ORF1ab:A and ORF1ab:B fragments and the M and N genes without tissue specificity. The width of a segment is roughly in proportion to the number of the same mutations in different genomes in the star-like tree or to the prosperity of a mutation in the maximum likelihood tree. Some labels of codons are neglected for clarity. **(E)** FCTE for the star-like tree with gene expression levels calibrated for lung tissue. **(F)** FCTE for the maximum likelihood tree with gene expression levels calibrated for lung tissue. **(G)** FCTE for the star-like tree without tissue specificity. **(H)** FCTE for the maximum likelihood tree without tissue specificity. Gray points indicate $\log_2$ FCTE values of mutations, whose pseudomedians are indicated by vertical bars. For coding regions where the Wilcoxon tests reject the null hypothesis that the $\log_2$ FCTE is zero, the vertical bars are colored red and black boxes are added indicating the nonparametric 95% confidence intervals (except the M gene).

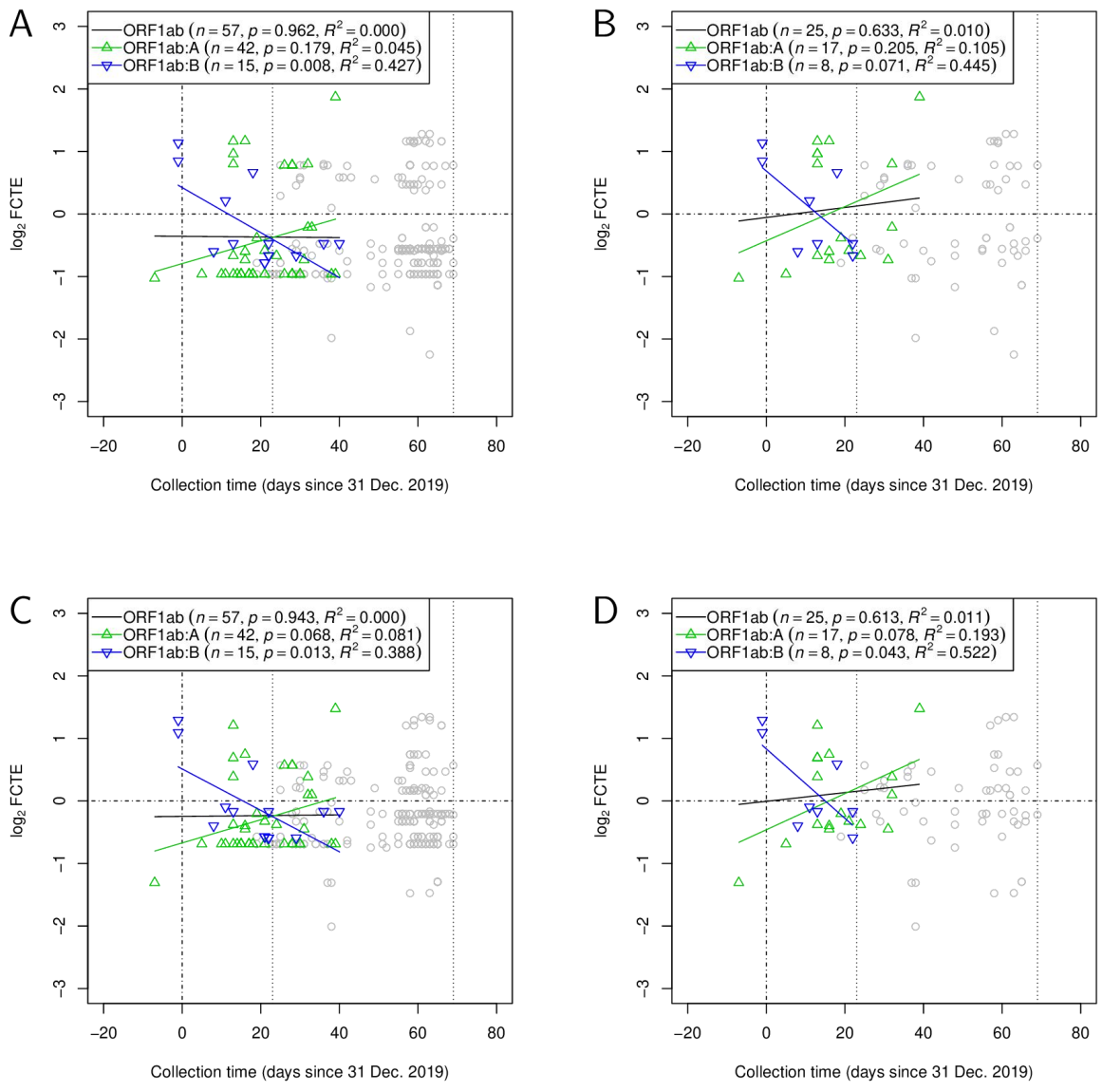


Fig. S3. Evolutionary trends of $\log_2$ fold change of translational efficiency (FCTE) of synonymous sites in the ORF1ab gene with gene expression levels calibrated for lung tissue and without tissue specificity. (A) $\log_2$ FCTE for the star-like tree plotted against collection time of the genomes with gene expression levels calibrated for lung tissue. **(B)** $\log_2$ FCTE for the maximum likelihood tree plotted against collection time of the genomes with gene expression levels calibrated for lung tissue. **(C)** $\log_2$ FCTE for the star-like tree plotted against collection time of the genomes without tissue specificity. **(D)** $\log_2$ FCTE for the maximum likelihood tree plotted against collection time of the genomes without tissue specificity. Mutation in genomes collected in China are highlighted with green up-pointing triangles for mutations in ORF1ab:A and blue down-

459 pointing triangles for mutations in ORF1ab:B, with fitted lines. Black lines are fitted to
460 both the green and blue triangle. The numbers of mutations (n), p values of the linear
461 models, and correlation coefficients (R^2) are shown. Samples collected elsewhere are
462 shown as gray circles. Vertical dotted lines indicate the dates of Wuhan lockdown (23 Jan.
463 2020) and Italy lockdown (9 March 2020).

464

Table S1. Codon usage frequency table used for estimation of fold change of translational efficiency (FCTE).

Codon	Amino acid	AT2 cell		Lung		Human**	
		Frequency (‰)	Fraction* (%)	Frequency (‰)	Fraction (%)	Frequency (‰)	Fraction (%)
TAA	Stop	1.4	37.7	1.1	37.1	1.0	29.4
TAG		0.7	18.6	0.6	19.9	0.8	23.5
TGA		1.6	43.7	1.3	43.0	1.6	47.1
GCT	A	21.2	29.1	19.9	26.9	18.4	26.6
GCC		29.0	39.8	31.6	42.6	27.7	40.0
GCA		16.3	22.4	14.5	19.6	15.8	22.8
GCG		6.3	8.7	8.1	10.9	7.4	10.7
TGT	C	9.3	44.8	8.6	41.4	10.6	45.7
TGC		11.4	55.2	12.2	58.6	12.6	54.3
GAT	D	23.9	50.2	21.8	43.4	21.8	46.5
GAC		23.8	49.8	28.4	56.6	25.1	53.5
GAA	E	31.2	44.6	26.6	37.6	29.0	42.3
GAG		38.8	55.4	44.2	62.4	39.6	57.7
TTT	F	15.2	45.7	14.3	40.5	17.6	46.4
TTC		18.1	54.3	21.1	59.5	20.3	53.6
GGT	G	13.3	19.2	12.9	17.8	10.8	16.4
GGC		24.4	35.2	26.9	37.2	22.2	33.6
GGA		17.0	24.6	16.0	22.1	16.5	25.0
GGG		14.5	20.9	16.5	22.8	16.5	25.0
CAT	H	9.4	40.0	8.8	37.0	10.9	41.9
CAC		14.1	60.0	15.0	63.0	15.1	58.1
ATT	I	18.2	38.2	14.9	34.2	16.0	36.1
ATC		22.8	47.8	23.7	54.3	20.8	47.0
ATA		6.7	14.0	5.0	11.4	7.5	16.9
AAA	K	29.9	42.2	23.6	36.5	24.4	43.3
AAG		40.9	57.8	41.1	63.5	31.9	56.7
TTA	L	6.7	7.2	4.8	5.2	7.7	7.7
TTG		11.9	12.8	10.9	11.6	12.9	12.9
CTT		13.3	14.4	10.9	11.7	13.2	13.2
CTC		17.6	19.0	18.8	20.1	19.6	19.6
CTA		6.7	7.2	5.6	6.0	7.2	7.2
CTG		36.5	39.4	42.4	45.4	39.6	39.5
ATG	M	24.6	100.0	23.3	100.0	22.0	100.0
AAT	N	16.6	47.8	14.7	41.9	17.0	47.1
AAC		18.1	52.2	20.4	58.1	19.1	52.9
CCT	P	18.1	30.6	17.1	28.2	17.5	28.6
CCC		17.7	29.9	20.9	34.6	19.8	32.4
CCA		16.9	28.5	15.4	25.6	16.9	27.7
CCG		6.5	11.0	7.0	11.6	6.9	11.3
CAA	Q	11.6	26.4	9.6	21.5	12.3	26.5
CAG		32.4	73.6	35.2	78.5	34.2	73.5
CGT	R	5.9	10.1	5.7	9.9	4.5	7.9
CGC		11.4	19.3	12.8	22.2	10.4	18.3
CGA		7.2	12.2	6.4	11.1	6.2	10.9
CGG		10.4	17.6	12.3	21.3	11.4	20.1

AGA		13.0	22.1	9.9	17.1	12.2	21.5
AGG		11.0	18.7	10.6	18.4	12.0	21.2
TCT	S	14.3	19.8	13.3	18.3	15.2	18.7
TCC		15.9	22.1	17.4	24.0	17.7	21.8
TCA		10.5	14.5	9.2	12.7	12.2	15.0
TCG		3.5	4.8	4.4	6.1	4.4	5.4
AGT		11.0	15.3	9.6	13.2	12.1	14.9
AGC		16.9	23.5	18.7	25.7	19.5	24.0
ACT	T	14.3	27.7	12.1	23.4	13.1	24.6
ACC		18.1	35.0	20.9	40.3	18.9	35.5
ACA		14.0	27.1	12.7	24.4	15.1	28.4
ACG		5.3	10.3	6.2	12.0	6.1	11.5
GTT	V	13.1	20.2	10.4	16.5	11.0	18.1
GTC		15.1	23.4	15.8	25.0	14.5	23.9
GTA		7.9	12.3	6.2	9.8	7.1	11.7
GTG		28.5	44.0	30.8	48.7	28.1	46.3
TGG	W	10.6	100.0	11.1	100.0	13.2	100.0
TAT	Y	11.8	43.6	11.0	40.0	12.2	44.4
TAC		15.3	56.4	16.6	60.0	15.3	55.6

467 * Fraction of each codon among all those for a given amino acid.

468 ** Downloaded from Codon Usage Database (*16*).

469

Data S1. Sequences from GISAID's EpiCoV™ on which this study is based. The last column "317 genomes" indicates whether a genome was used in constructing the maximum likelihood phylogenetic tree and examining the synonymous triplet sites.