The Innovation, Volume *2*

Supplemental Information

Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process

Ke-Jia Shan, Changshuo Wei, Yu Wang, Qing Huan, and Wenfeng Qian

MATERIALS AND METHODS

Identification of *de novo* RNA mutations in SARS-CoV-2

We identified *de novo* RNA mutations in SARS-CoV-2 from the nanoball-based RNA sequencing data reported in a previous study,[1] which were generated for the Vero cells infected by SARS-CoV-2 BetaCoV/Korea/KCDC03/2020, at a multiplicity of infection (MOI) of 0.05 for 24 h. The 305,065,029 high-throughput sequencing read pairs (2×100-nucleotide) were retrieved from the Open Science Framework under the digital object identifier number 10.17605/OSF.IO/8F6N9. The bioinformatics pipeline can be found in **Figure S2**, and we describe specific parameters below.

Since the Vero cell was isolated from African green monkey kidney,[2] the sequencing read pairs were first mapped to the *Chlorocebus sabaeus* genome (Ensembl: ChlSab1.1) using STAR 2.7.1a[3] under the parameters supplied by Kim et al. (--outFilterMultimapNmax 1000000 --outFilterType BySJout --alignSJoverhangMin 8 --outSJfilterOverhangMin 12 12 12 12 --outSJfilterCountUniqueMin 1 1 1 1 --outSJfilterCountTotalMin 1 1 1 1 --outSJfilterDistToOtherSJmin 0 0 0 0 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --scoreGapNoncan -4 --scoreGapATAC -4 --chimOutType WithinBAM HardClip --chimScoreJunctionNonGTAG 0 --alignSJstitchMismatchNmax -1 -1 -1 -1 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000).

The read pairs mapped to the *C. sabaeus* genome were discarded, and the remaining ones were mapped to SARS-CoV-2 BetaCoV/Korea/KCDC03/2020 genome (GISAID: EPI_ISL_407193) using the same set of parameters as described above. Based on the junction barcode (*i.e.*, a pair of upstream and downstream junction sites), these read pairs were grouped into 269,125 read families, among which the majority were present at a low frequency in the transcriptome (*e.g.*, 264,613 read families each including ≤20 read pairs, 258,356 each including ≤10 read pairs, and 244,294 each including ≤5 read pairs). The read pairs mapped to multiple positions in the SARS-CoV-2 genome were also discarded. The read pairs that contained exactly one junction and were mapped end-to-end along the full length of the read pair were used to identify *de novo* RNA mutations (**Figure S2**).

Single-nucleotide mismatches were detected by samtools mpileup v1.9[4] with the parameters (-d 0 --output-BP --output-QNAME). These mismatched bases were then retained as candidate RNA mutations (**Figure S2**). It is worth noting that the mismatch frequency of C>G was much higher than the others, suggesting a C>G sequencing bias in the nanoball-based high-throughput sequencing (**Figure S3D**).

Among these mismatches, we identified *bona fide* RNA sequence variation extant in the negative-sense subgenomic RNA by applying three criteria (**Figure S2**). First, we discarded any mismatches that were supported by only one sequencing read, as such a mismatch could have been created through errors in high-throughput sequencing, PCR amplification, or reverse transcription. Similarly, mismatches were also discarded if all supportive read pairs appeared likely to be artifacts of PCR amplification or reverse transcription during library preparation, as indicated by identical mapping positions of the 5′- and 3′-ends of the read pair in the reference genome. Second, to be conservative, we kept only mismatches that were unanimously supported by all sequencing reads in a family (**Figure S2**). Third, we observed a greater number of mismatches immediately adjacent to junction sites which decreased in frequency through ~15 nucleotides up- and down-stream of the junction site (**Figure S2**). Most of these mismatches were likely alignment artifacts,[5,6] and therefore, we excluded all mismatches located less than 15 nucleotides away from the junction site.

It is also noteworthy that the detected sequence variations could also result from extant polymorphisms in the viral population used to infect the Vero cells. If a sequence variation was observed in multiple transcripts, we surmise that it was likely derived from a viral polymorphism rather than a *de novo* mutation. To this end, we fitted two normal distributions to the distribution of background mismatch frequency (*i.e.*, among all reads covering a site, whether or not bearing a junction), and found that a cut-off of 0.2% in background mismatch frequency would result in a false discovery rate of ~2% in identification of *de novo* mutations (**Figure S2D**). Therefore, we further discarded mismatches that appeared at >0.2% background frequency.

Note that although we detected mutations present in the negative-sense subgenome, these mutations could arise in the positive-sense genome due to exposure to mutagens after the virus infected a cell.

To estimate the molecular spectrum of *de novo* mutations in SARS-CoV-2, we divided the number of mutations of each of the 12 base-substitution types by the total number (*N*) of the particular nucleotide type (A, C, G, or U) where such mutation type could have arisen. Provided that various regions in the viral genome are presented at different frequencies in the transcriptome (subgenomes), we estimated *N* from the total coverage in the transcriptome for all sites in the reference genome that exhibits the particular nucleotide type. For example, when mutation type G>U is under consideration, we identified all guanines in the reference genome and counted their total coverage estimated from the transcriptome data. Similar to the identification of *de novo* mutations, the coverage here was defined within each read family, only for the sites that were covered by at least two non-duplicate reads, that the nucleotide type was unanimously supported by all reads in the family, and that located at least 15 nucleotides away from the junction site.

A total of 37,129 junction-containing read pairs showed insertions or deletions of a few nucleotides (indels). These read pairs were used to detect *de novo* indels, using the same cut-off as that for the detection of *de novo* point mutations. Indels existing in multiple read families were counted only once.

Comparison of the mismatch frequency between the junction-barcoding approach and the conventional computational approach

We compared the accuracy of our junction-barcoding approach in identifying mutations with the conventional computational approach that treats all mismatches called from the sequencing data as mutations. Specifically, we estimated the overall mismatch frequencies as a function of Phred quality score, as described in a previous study.[7] To make the comparison fair, we used the same read families for the conventional computational approach as those used for the junction-barcoding approach, except that the conventional computational approach treats reads individually while the junction-barcoding approach treats read family as a whole. The 30-nucleotide region centered at the junction sites was similarly discarded for both approaches.

We obtained the base quality scores for each nucleotide on individual reads using samtools mpileup under the parameters (-d 0 --output-BP --output-QNAME -Q 0 -B). For the conventional approach, for each Phred quality score from 1–40, we divided the number of mismatches supported at the confidence level indicated by a particular quality score, by the total number of sites showing this quality score across all individual reads; this ratio was defined as the mismatches frequency. For the junction-barcoding approach, the mismatch frequency was similarly estimated, except that the Phred quality score of each site is the average rounded quality score of all reads covering a particular site in a read family. The mismatch frequencies were not estimated for average Phred quality score of 1, 2, or 3 because less than 10 sites exhibited such quality scores across all read families.

The mismatch frequency estimated using the junction-barcoding approach was lower than the estimate generated by conventional computational approaches for the same dataset (**Figure 1C**). These results indicated that sequencing errors were effectively removed using our strategy (although sometimes two or more independent *de novo* mutations might be treated as one by our algorithm). Moreover, the mismatch frequency was largely stable over a range of sequencing quality scores (**Figure 1C**, from 28 to 40), suggesting that our approach was not heavily dependent on an extremely low sequencing error rate.

Characterization of the molecular spectra of among-patient polymorphisms for three virus species

A total of 34,852 complete genome sequences of SARS-CoV-2 variants were downloaded from GISAID (Global Initiative on Sharing All Influenza Data, https://www.gisaid.org/)[8] on Jun 29, 2020. 1839 complete genome sequences of *Influenza A virus* variants, which were collected during the 2009 H1N1 pandemic, were also downloaded from GISAID. 258 complete genome sequences of MERS-CoV variants isolated from patients were downloaded from NCBI Virus (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/).[9]

We performed MUSCLE v3.8.1551[10] to align each virus variant to its corresponding reference genome and aggregated individual alignments into a single multiple sequence alignment. We reconstructed the sequence of the last common ancestor for each virus species using FastML v3.11 under the default parameters.[11] We compared the sequence of each virus variant with that of the last common ancestor to identify sequence variations. Sequence variations supported by at least two virus variants were considered as polymorphisms, to reduce the possibility of potential sequencing errors being recognized as polymorphisms. Sequence variations that were identified in multiple patients were counted only once to minimize the influence of positive selection.

Characterization of the molecular spectrum of RNA mutations in *Saccharomyces 20S RNA narnavirus* and yeast endogenous mRNAs

We identified the RNA mutations in *Saccharomyces 20S RNA narnavirus* and in endogenous mRNA from the ARC-seq data for the budding yeast.[12] The processed consensus sequences were downloaded from NCBI under the accession number of BioProject PRJNA396053. These sequences were mapped to the yeast genome (Ensembl, R64-1-1) and the *Saccharomyces 20S RNA narnavirus* genome (GenBank: NC_004051.1) using STAR with the default parameters. RNA mutations were detected by samtools mpileup with the parameters as follows: -d 0 --output-BP --output-QNAME -Q 30, and were polarized according to the coding strand, based on the yeast genome annotation (Ensembl, R64-1-1, version 48). The RNA mutations that locate at the genome positions with >1% mismatch frequency were discarded, as they might be caused by polymorphisms among individual yeast cells. Endogenous RNA mutations that located in the mitochondrial genome were discarded.

Endogenous mRNA mutations were also identified in the budding yeast using CirSeq.[13] We downloaded the raw sequences from NCBI under the accession number of BioProject PRJNA430448 and called mRNA mutations using the pipeline provided by the authors.

Retrieval of reported molecular spectra of *de novo* mutations for Ebola virus and poliovirus

The molecular spectrum of *de novo* mutations in Ebola virus was retrieved from a previous study,[14] in which 293T cells were infected by Ebola virus at an MOI of 0.1. The molecular spectrum of *de novo* mutations in poliovirus was retrieved from a previous study,[7] in which HeLa S3 cells were infected by poliovirus at an MOI of 0.1 for 6–8 h. In both studies, CirSeq were applied to identify *de novo* mutations in RNA viruses.

Characterization of molecular spectra of somatic mutations in 36 human tissues

We retrieved the somatic mutation data identified for 36 human tissues from a previous study (Garcia-Nieto et al., 2019). Somatic mutations were polarized according to the coding strand DNA based on the human genome annotation (Ensembl, GRCh37, version 84). Somatic mutations located in the overlapping regions between two genes were discarded, as the DNA strand in which these mutations arose could not be determined. We discarded mutations detected in multiple humans to reduce the interference from the potential standing polymorphisms in the population. We also discarded somatic mutations that existed in multiple tissues of the same human to exclude potential RNA editing events. The frequencies of mutations were normalized by the nucleotide content in the transcribed regions.

Characterization of the molecular spectra of mutations that accumulated in the evolution of SARS-CoV-2, SARS-CoV, and MERS-CoV and their related coronaviruses

We retrieved the genomic sequences of six SARS-CoV-2-related coronaviruses: RaTG13 (GenBank: MN996532.1) isolated from *R. affinis*,[15] RshSTT200 (GISAID: EPI_ISL_852605) from *R. shameli*,[16] ZC45 (GenBank: MG772933.1) from *R. pusillus*,[17] Rc-o319 (GenBank: LC556375.1) from *R. cornutus*,[18] GD-1 (GISAID: EPI_ISL_410721) from *M. javanica*,[19] and GX-P5L (GISAID: EPI_ISL_410540) from *M. javanica*.[20] We retrieved the genomic sequences of six SARS-CoV-related coronaviruses: Tor2 (GenBank: NC_004718.3) isolated from a patient,[21] Civet020 (GenBank: AY572038.1) from *Paguma larvata*,[22] WIV1 (GenBank: KF367457.1) from *R. sinicus*,[23] Rp3 (GenBank: DQ071615.1) from *R. pearsoni*,[24] HKU3-1 (GenBank: DQ022305.2) from *R. sinicus*,[25] and BM48-31 (GenBank: NC_014470.1) from *R. blasii*.[26] We retrieved the genomic sequences of five MERS-CoV-related coronaviruses: NRCE-HKU270 (GenBank: KJ477103.2) isolated from *Camelus dromedarius* in Egypt,[27] PML-PHE1 (GenBank: KC869678.4) from *Neoromicia zuluensis*,[28] SC2013 (GenBank: KJ473821.1) from *Vespertilio superans*,[29] VMC (GenBank: KC545386.1) from *Erinaceus europaeus*,[30] and HKU4 (GenBank: NC_009019.1) from *Tylonycteris pachypus*.[31] We used MUSCLE to separately create multiple alignments for SARS-CoV-2, SARS-CoV, and MERS-CoV. We built the maximum likelihood trees and reconstructed the ancestral sequence for each internal node using FastML under the default parameters.

We labeled putative host species for each branch according to the parsimony principle (**Figures 6A** and **7A**). There are three cases worth noting. First, since multiple MERS-CoV variants were independently transmitted from camels to humans,[32] the viral polymorphisms detected among human patients reflected its evolutionary history in both camels and humans. Second, similar to other previously published phylogenetic trees for SARS-CoV-related viruses,[33] our phylogeny did not show a clear transmission route from civets to humans. Thus, in light of previous work which proposed that bats were likely to serve as the natural reservoirs of coronaviruses,[24,34,35] we reasoned that node N_6 represented a host status in bats. Note that due the paucity of mutations that accumulated in branches B_{10} and B_{11}, we were not able to exclude the possibility that the host of node N_6 was civets, which would better reflect the conventional understanding that civets were the intermediate host for SARS-CoV.[36] Third, given the paucity in full-length sequences of SARS-CoV among patients, we used only one SARS-CoV variant, Tor2,[21] in the analysis, and the branch leading to this variant (B_{10}) represented its mixed evolutionary history in bats and humans.

Bootstrapping

To determine if two correlations in molecular spectra are significantly different, we performed resampling tests using bootstrapping (**Figure 6C**). Specifically, we randomly sampled the identified mutations in each branch 10,000 times with replacement, keeping the number of mutations unchanged. *P* value and the 95% confident intervals (CI) of *r* were estimated based on the 10,000 paired bootstrapped observations.

Principal component analyses

We performed a principal component analysis (prcomp function in *R*) with the proportions of the 12 base-substitution types as the input. We projected molecular spectra into a two-dimensional space according to the first two principal components. We estimated the 95% confidence ellipses (stat_ellipse option in *R*) from the 17 bat-exclusive branches, 13 *Rhinolophus*-exclusive branches, or six human-related spectra (B_{10}, pSCV2, pMERS, mEbola, mSCV2 and mPV), in an effort to define the borderlines of cellular environments for bats, *Rhinolophus* bats, and humans, respectively. Note that the Vero cell, where the *de novo* mutations of SARS-CoV-2 were detected, was isolated from African green monkey, and here is also considered human-related.

Code and data availability

All scripts used to analyze the data and to generate the figures are available at https://github.com/kjshan/SARS-CoV-2-Mutation-Spectrum/ and Zenodo (https://doi.org/10.5281/zenodo.5203190). All data that were used to support the findings of this study are available in the public databases.

SUPPLEMENTAL DISCUSSION

It is noteworthy that in previous studies, differences between mRNA and genomic DNA sequences have been termed “transcription errors”.[13,37] In this study, we show that a proportion of G>U and C>U mutations arise independently of the transcription process, and therefore, we used the term “RNA mutation” instead to clarify the origin of such mutations. This new term echoes previous observations in poliovirus made by Korboukh *et al.*, who found that the mutation rates of C>U and G>U were not significantly affected by a defect in RdRp (H273R) that could significantly increase the mutation rate generated during transcription.[38]

There are some caveats to the conclusions drawn from our results. First, our junction-barcoding approach requires at least two independent mismatches in a sequencing read family to call a mutation. While this requirement has reduced errors associated with high-throughput sequencing by up to four orders of magnitude, from 10^{-4} to 10^{-8} false positives per nucleotide, the false positive rate for detecting mutations is higher than that reported (10^{-12} false positives per nucleotide) for CirSeq.[7] Nevertheless, the rate of 10^{-8} false positives per nucleotide is approximately two orders of magnitude below that of the previously estimated RNA mutation rate,[37,39] indicating that this junction-barcoding approach provides an accurate gauge of the molecular spectrum of *de novo* mutations.

Second, although we discarded the mismatches that appeared at >0.2% background frequency (**Figure S2**) because we suspected that they were extant polymorphisms in the viral population used to infect the Vero cells. Nevertheless, the molecular spectrum of these polymorphisms was highly correlated with that of the *de novo* mutations (Pearson’s correlation coefficient, $r = 0.80$, $P = 3\times10^{-3}$, **Figure S3E**), indicating that the molecular spectrum of *de novo* mutations dominates the base substitution types of within-individual polymorphisms in SARS-CoV-2 during its evolution. Note that the C>G base substitution type were excluded in this analysis due to its high sequencing error rate (**Figure S3D**).

Third, the mutations in SARS-CoV-2 we detected were those in the intermediate negative-sense subgenomes and appeared to be non-heritable. Nevertheless, we reason that they can be used to infer the molecular spectrum of the heritable mutations in the genomic RNA since both genomic and subgenomic RNAs were synthesized by the same polymerase and shared the same cellular environment. Consistent with this hypothesis, the asymmetric emergence of G>U and C>U RNA mutations were observed in the heritable genomes of other single-strand RNA viruses, such as Ebola virus (**Figure 3E**) and 20S RNA narnavirus (**Figure 4B**), with respect to their respective genomic strands.

Fourth, based on the assumption that single-strand RNA and DNA are sensitive to similar (or the same) mutagens, such as ROS, we inferred differences in cellular environments using somatic DNA mutations as a reliable proxy for mutagenesis of the same mechanism in viral RNA (**Figure 5**). These somatic mutations were identified in a previous study[6] from the transcriptomic data collected by the GTEx project.[40] Although some of the identified somatic mutations could, in principle, have resulted from editing or damage specific to single-strand RNA,[6] the cellular environment that can induce RNA editing or damage is exactly what we aimed to investigate initially, because these are the mechanisms that drive the evolution of RNA viruses.

Fifth, we reported that the cellular environment of the lung could induce both G>U and C>U mutations in RNA viruses, using the somatic mutations identified from 345 lung samples collected in the GTEx project. However, the cellular environment can vary among individuals. A previous study[41] identified within-patient polymorphisms among SARS-CoV-2 virions isolated from bronchoalveolar lavage fluids of eight patients.[42,43] Although on average G>U and C>U polymorphisms were more abundant than their respective complement polymorphisms, it was not always the case for each patient. Although the numbers of these within-patient polymorphisms were generally low for statistical tests, this observation suggests variability in the cellular environment among individuals that can influence the accumulation of G>U or C>U mutations.

SUPPLEMENTAL REFERENCES

1. Kim, D., Lee, J.Y., Yang, J.S., et al. (2020). The architecture of SARS-CoV-2 transcriptome. Cell ***181***, 914-921 e910.
2. Rhim, J.S., Schell, K., Creasy, B., and Case, W. (1969). Biological characteristics and viral susceptibility of an African green monkey kidney cell line (Vero). Proc Soc Exp Biol Med ***132***, 670-678.
3. Dobin, A., Davis, C.A., Schlesinger, F., et al. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics ***29***, 15-21.
4. Li, H., Handsaker, B., Wysoker, A., et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics ***25***, 2078-2079.
5. Carey, L.B. (2015). RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. Elife ***4***, e09945.
6. Garcia-Nieto, P.E., Morrison, A.J., and Fraser, H.B. (2019). The somatic mutation landscape of the human body. Genome Biol ***20***, 298.
7. Acevedo, A., Brodsky, L., and Andino, R. (2014). Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature ***505***, 686-690.
8. Shu, Y., and McCauley, J. (2017). GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill ***22***, 30494.
9. Hatcher, E.L., Zhdanov, S.A., Bao, Y., et al. (2017). Virus Variation Resource - improved response to emergent viral outbreaks. Nucleic Acids Res ***45***, D482-D490.
10. Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics ***5***, 113.
11. Ashkenazy, H., Penn, O., Doron-Faigenboim, A., et al. (2012). FastML: a web server for probabilistic reconstruction of ancestral sequences. Nucleic Acids Res ***40***, W580-584.
12. Reid-Bayliss, K.S., and Loeb, L.A. (2017). Accurate RNA consensus sequencing for high-fidelity detection of transcriptional mutagenesis-induced epimutations. Proc Natl Acad Sci U S A ***114***, 9415-9420.
13. Gout, J.F., Li, W., Fritsch, C., et al. (2017). The landscape of transcription errors in eukaryotic cells. Sci Adv ***3***, e1701484.
14. Whitfield, Z.J., Prasad, A.N., Ronk, A.J., et al. (2020). Species-specific evolution of Ebola virus during replication in human and bat cells. Cell Rep ***32***, 108028.
15. Zhou, P., Yang, X.L., Wang, X.G., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature ***579***, 270-273.
16. Hul, V., Delaune, D., Karlsson, E.A., et al. (2021). A novel SARS-CoV-2 related coronavirus in bats from Cambodia. bioRxiv 10.1101/2021.1101.1126.428212.
17. Hu, D., Zhu, C., Ai, L., et al. (2018). Genomic characterization and infectivity

of a novel SARS-like coronavirus in Chinese bats. Emerg Microbes Infect *7*, 154.

18. Murakami, S., Kitamura, T., Suzuki, J., et al. (2020). Detection and characterization of bat sarbecovirus phylogenetically related to SARS-CoV-2, Japan. Emerg Infect Dis ***26***, 3025-3029.
19. Xiao, K., Zhai, J., Feng, Y., et al. (2020). Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature ***583***, 286-289.
20. Lam, T.T., Jia, N., Zhang, Y.W., et al. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature ***583***, 282-285.
21. He, R., Dobie, F., Ballantine, M., et al. (2004). Analysis of multimerization of the SARS coronavirus nucleocapsid protein. Biochem Biophys Res Commun ***316***, 476-483.
22. Wang, M., Yan, M., Xu, H., et al. (2005). SARS-CoV infection in a restaurant from palm civet. Emerg Infect Dis ***11***, 1860-1865.
23. Ge, X.Y., Li, J.L., Yang, X.L., et al. (2013). Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. Nature ***503***, 535-538.
24. Li, W., Shi, Z., Yu, M., et al. (2005). Bats are natural reservoirs of SARS-like coronaviruses. Science ***310***, 676-679.
25. Lau, S.K., Woo, P.C., Li, K.S., et al. (2005). Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. Proc Natl Acad Sci U S A ***102***, 14040-14045.
26. Drexler, J.F., Gloza-Rausch, F., Glende, J., et al. (2010). Genomic characterization of severe acute respiratory syndrome-related coronavirus in European bats and classification of coronaviruses based on partial RNA-dependent RNA polymerase gene sequences. J Virol ***84***, 11336-11349.
27. Chu, D.K., Poon, L.L., Gomaa, M.M., et al. (2014). MERS coronaviruses in dromedary camels, Egypt. Emerg Infect Dis ***20***, 1049-1053.
28. Ithete, N.L., Stoffberg, S., Corman, V.M., et al. (2013). Close relative of human Middle East respiratory syndrome coronavirus in bat, South Africa. Emerg Infect Dis ***19***, 1697-1699.
29. Yang, L., Wu, Z., Ren, X., et al. (2014). MERS-related betacoronavirus in Vespertilio superans bats, China. Emerg Infect Dis ***20***, 1260-1262.
30. Corman, V.M., Kallies, R., Philipps, H., et al. (2014). Characterization of a novel betacoronavirus related to middle East respiratory syndrome coronavirus in European hedgehogs. J Virol ***88***, 717-724.
31. Woo, P.C., Wang, M., Lau, S.K., et al. (2007). Comparative analysis of twelve genomes of three novel group 2c and group 2d coronaviruses reveals unique group and subgroup features. J Virol ***81***, 1574-1585.
32. Zhang, Z., Shen, L., and Gu, X. (2016). Evolutionary dynamics of MERS-CoV: potential recombination, positive selection and transmission. Sci Rep ***6***, 25049.
33. Hu, B., Guo, H., Zhou, P., and Shi, Z.L. (2021). Characteristics of SARS-CoV-2 and COVID-19. Nat Rev Microbiol ***19***, 141-154.
34. Hu, B., Ge, X., Wang, L.F., and Shi, Z. (2015). Bat origin of human

coronaviruses. Virol J ***12***, 221.

35. Hu, B., Zeng, L.P., Yang, X.L., et al. (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.

36. Cui, J., Li, F., and Shi, Z.L. (2019). Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol ***17***, 181-192.

37. Gout, J.F., Thomas, W.K., Smith, Z., et al. (2013). Large-scale detection of in vivo transcription errors. Proc Natl Acad Sci U S A ***110***, 18584-18589.

38. Korboukh, V.K., Lee, C.A., Acevedo, A., et al. (2014). RNA virus population diversity, an optimum for maximal fitness and virulence. J Biol Chem ***289***, 29531-29544.

39. Sanjuan, R., Nebot, M.R., Chirico, N., et al. (2010). Viral mutation rates. J Virol ***84***, 9733-9748.

40. GTEx Consortium, Laboratory, D.A., Coordinating Center -Analysis Working, G., et al. (2017). Genetic effects on gene expression across human tissues. Nature ***550***, 204-213.

41. Di Giorgio, S., Martignano, F., Torcia, M.G., et al. (2020). Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv ***6***, eabb5813.

42. Chen, L., Liu, W., Zhang, Q., et al. (2020). RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect ***9***, 313-319.

43. Shen, Z., Xiao, Y., Kang, L., et al. (2020). Genomic diversity of severe acute respiratory syndrome-coronavirus 2 in patients with coronavirus disease 2019. Clin Infect Dis ***71***, 713-720.

SUPPLEMENTAL FIGURES

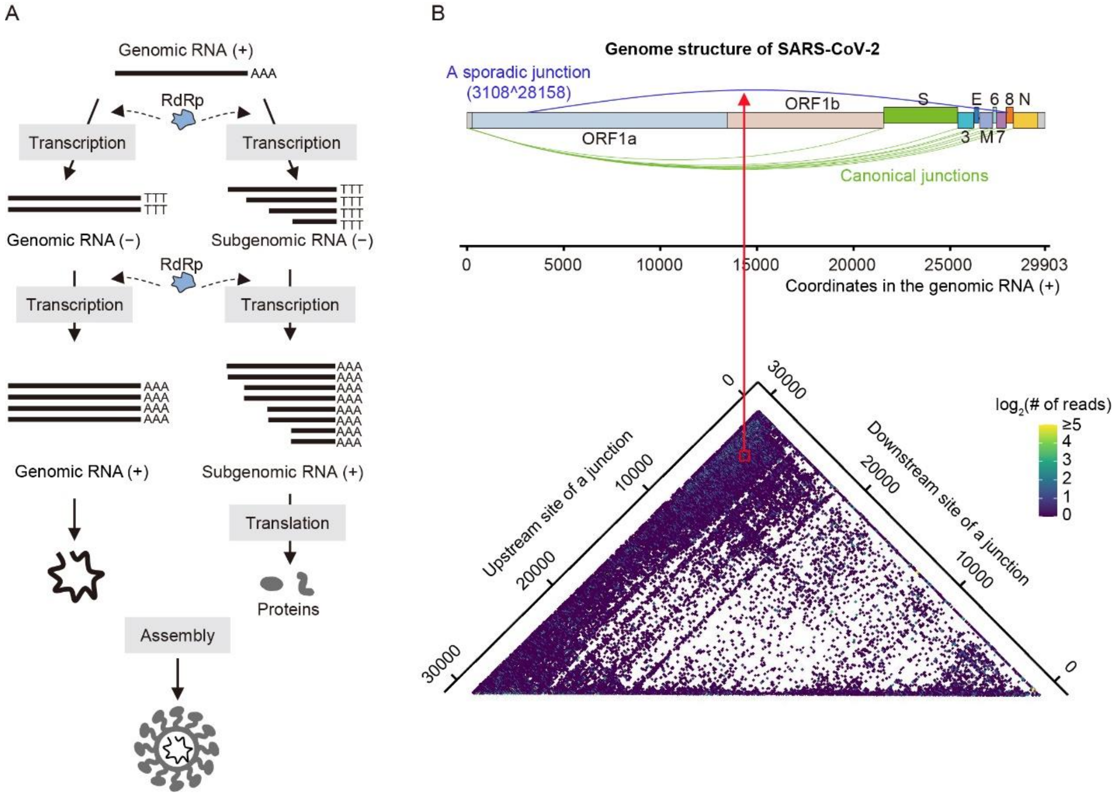
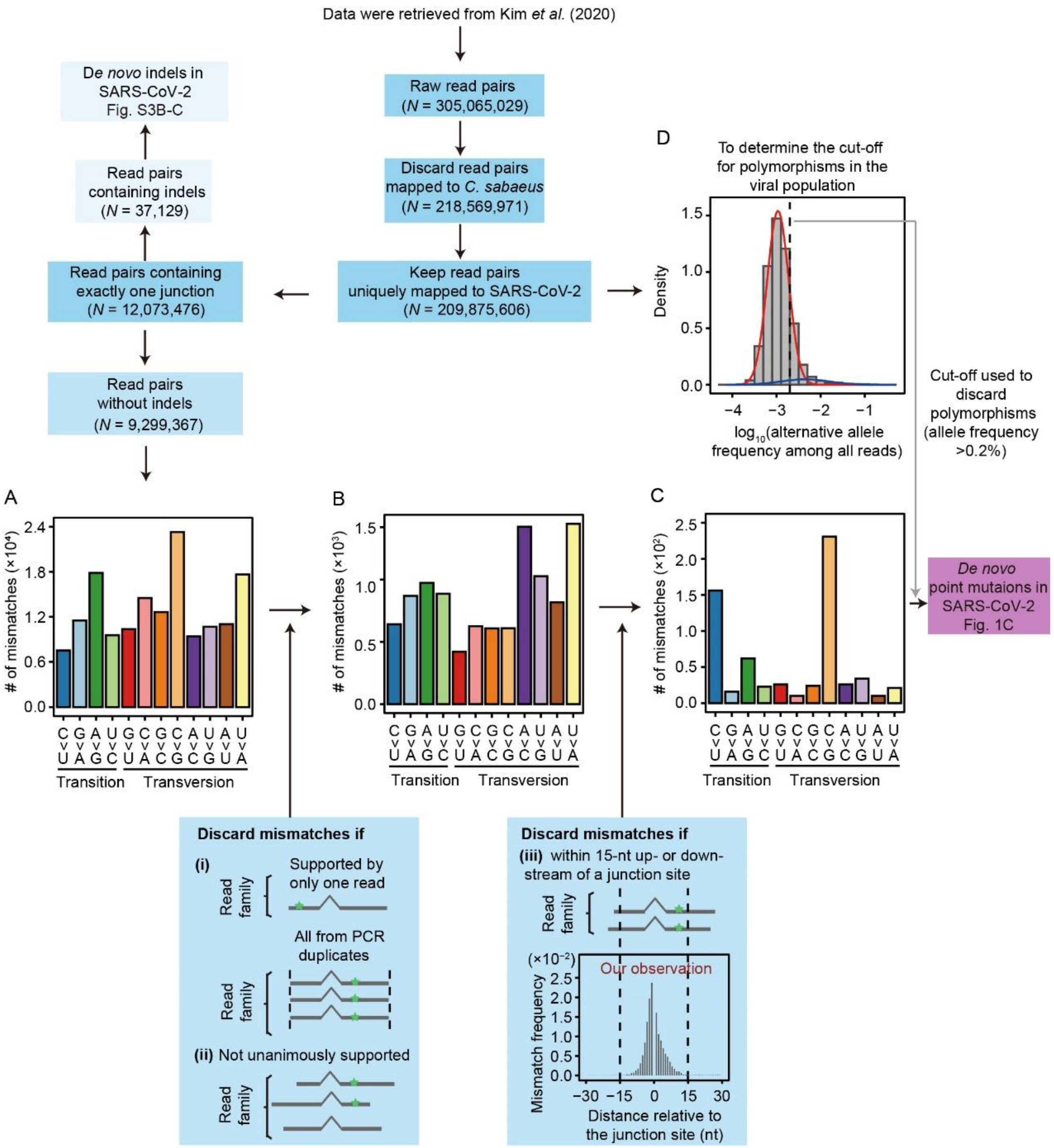


Figure S1. The life cycle of SARS-CoV-2 and its discontinuous transcription.

(A) The life cycle of SARS-CoV-2.

(B) Schematic of junctions generated during discontinuous transcription in SARS-CoV-2. Green curves denote the canonical junctions generated from the leader-to-body fusion, while the blue curve denotes a sporadic junction generated randomly from discontinuous transcription. The color in the heat map shows the number of reads sharing the same pair of upstream and downstream junction sites.



Data were retrieved from Kim et al. (2020)
Raw read pairs (N = 305,065,029)
Discard read pairs mapped to C. sabaeus (N = 218,569,971)
Keep read pairs uniquely mapped to SARS-CoV-2 (N = 209,875,606)
Read pairs containing exactly one junction (N = 12,073,476)
Read pairs containing indels (N = 37,129)
De novo indels in SARS-CoV-2 Fig. S3B-C
Read pairs without indels (N = 9,299,367)
D
To determine the cut-off for polymorphisms in the viral population
Density
1.5
1.0
0.5
0.0
−4
−3
−2
−1
log10(alternative allele frequency among all reads)
Cut-off used to discard polymorphisms (allele frequency >0.2%)
A
# of mismatches (×10⁴)
2.4
1.8
1.2
0.6
0.0
B
# of mismatches (×10³)
1.5
1.0
0.5
0.0
C
# of mismatches (×10²)
2.5
2.0
1.5
1.0
0.5
0.0
C>U G>A A>G U>C G>U C>A G>C C>G A>U U>A A>U U>A
Transition
Transversion
De novo point mutaions in SARS-CoV-2 Fig. 1C
Discard mismatches if
(i)
Supported by only one read
Read family
All from PCR duplicates
Read family
(ii) Not unanimously supported
Read family
Discard mismatches if
(iii) within 15-nt up- or down-stream of a junction site
Read family
(×10⁻²)
Our observation
Mismatch frequency
2.5
2.0
1.5
1.0
0.5
0.0
−30
−15
0
15
30
Distance relative to the junction site (nt)

Figure S2. The workflow for the identification of *de novo* RNA mutations in SARS-CoV-2.

Some intermediate molecular spectra of mismatches are shown in insets (A–C). Inset (D) shows a mixture of two normal distributions that fit the distribution of mismatch frequency, which was estimated from all uniquely mapped reads that covered a site. The red and blue lines indicate two normal distributions, and the black dash line indicates the cut-off frequency (0.2%) used to remove polymorphisms in the viral population in this study.

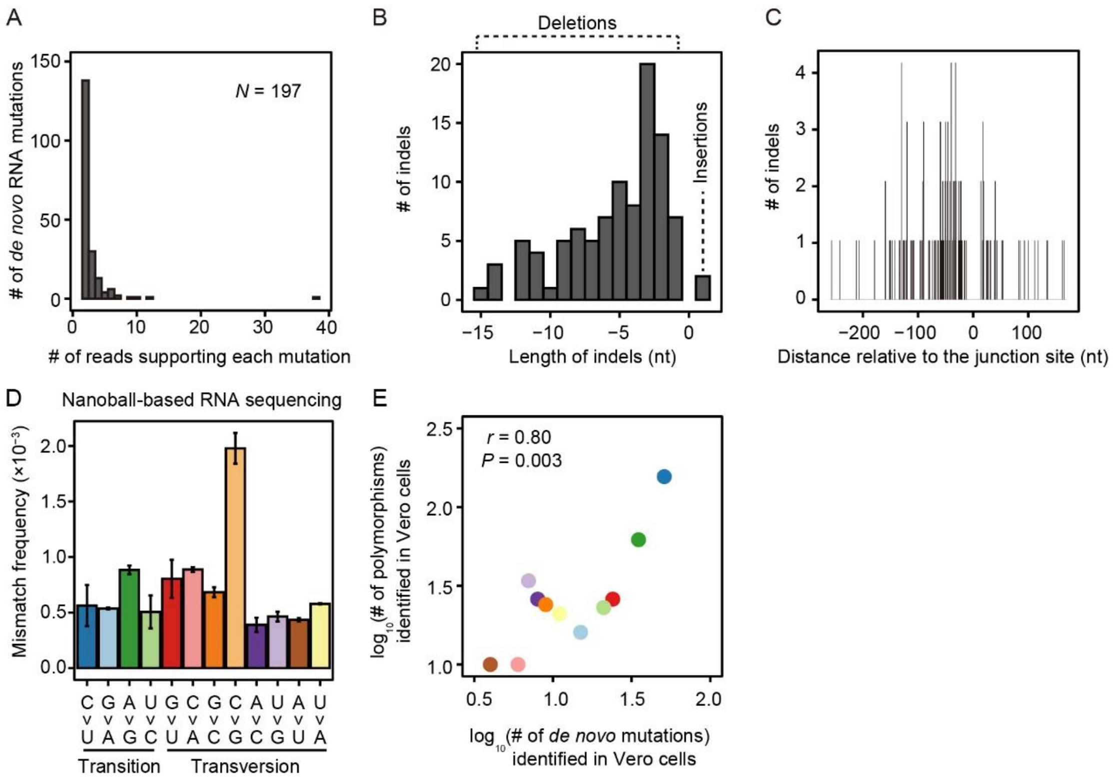


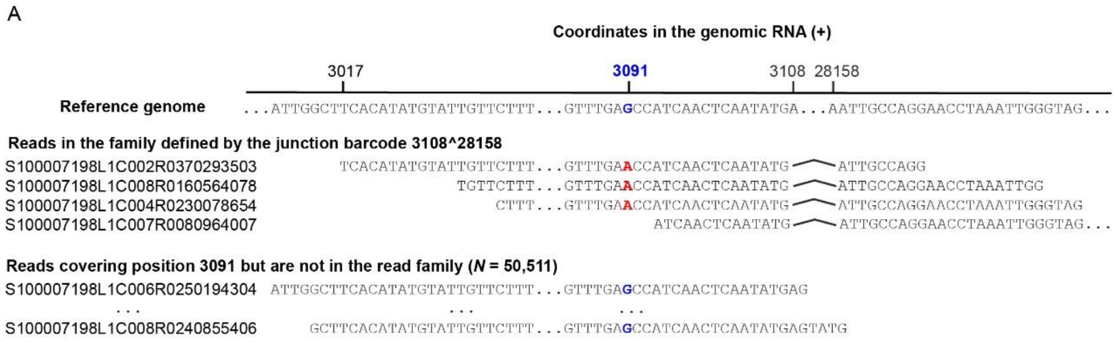
Figure S3. Additional results about *de novo* RNA mutations in SARS-CoV-2.

(A) Histogram shows the distribution of the number of non-duplicated reads that supports each of the 197 *de novo* RNA mutations detected from the transcriptome data in Vero cells.

(B–C) Histograms show the length and position (relative to the junction site) distributions for indels detected in Vero cells.

(D) Histogram shows the mismatch frequency of 12 base-substitution types among 209,875,606 read pairs uniquely mapped to the SARS-CoV-2 genome.

(E) A scatter plot shows the molecular spectrum of *de novo* mutations vs. within-cell-line polymorphisms identified in the SARS-CoV-2 genome. Note that the C>G base substitution was excluded (thereby *N* = 11) because of its higher sequencing error rate as shown in (D).


A
Coordinates in the genomic RNA (+)
3017
3091
3108
28158
Reference genome
...ATTGGCTTCACATATGTATTGTTCTTT...GTTTGAGCCATCAACTCAATATGA...AATTGCCAGGAACCTAAATTGGGTAG...
Reads in the family defined by the junction barcode 3108^28158
S100007198L1C002R0370293503
TCACATATGTATTGTTCTTT...GTTTGAACCATCAACTCAATATG ATTGCCAGG
S100007198L1C008R0160564078
TGTTCTTT...GTTTGAACCATCAACTCAATATG ATTGCCAGGAACCTAAATTGG
S100007198L1C004R0230078654
CTTT...GTTTGAACCATCAACTCAATATG ATTGCCAGGAACCTAAATTGGGTAG
S100007198L1C007R0080964007
ATCAACTCAATATG ATTGCCAGGAACCTAAATTGGGTAG...
Reads covering position 3091 but are not in the read family (N = 50,511)
S100007198L1C006R0250194304
ATTGGCTTCACATATGTATTGTTCTTT...GTTTGAGCCATCAACTCAATATGAG
...
S100007198L1C008R0240855406
GCTTCACATATGTATTGTTCTTT...GTTTGAGCCATCAACTCAATATGAGTATG

B

# of reads covering position "3091"

	Alternative (A)	Reference (G)
Reads barcoded by junction 3108^28158	3	0
All uniquely mapped reads	15	50434

Fisher's exact test: two-tailed P-value = 4×10^{-11}

Figure S4. An example RNA mutation in SARS-CoV-2 identified by our junction-barcoding approach from the transcriptome data published in Kim *et al*. (2020).

(A) All four sequences in a read family that bore junction barcode (upstream:3108^downstream:28158). Two out of 50,511 other sequencing reads that covered position 3091 are also shown; among them, 12 reads showed "A" at position 3091, likely caused by errors generated from reverse transcription, PCR, and sequencing.

(B) A 2 × 2 table summarizes the sequencing information shown in (A). The table shows the enrichment of G>A mismatch in the read family barcoded by junction 3108^28158, at position 3091.

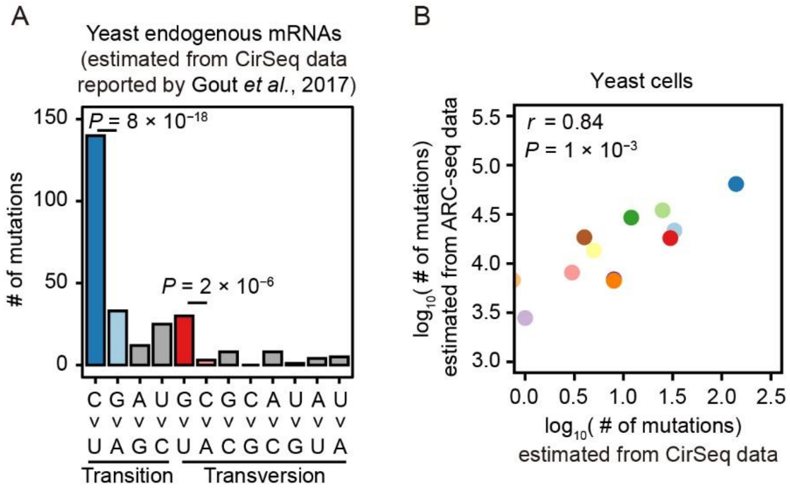


Figure S5. The molecular spectra of mRNA mutations in the budding yeast.

(A) The molecular spectrum of mRNA mutations that we estimated from the CirSeq data for yeast cells. Two-tailed *P*-values were calculated from Fisher's exact tests.

(B) A scatter plot shows the molecular spectra of yeast mRNA mutations estimated from CirSeq vs. ARC-seq. Pearson's correlation coefficient (*r*) and the corresponding *P*-value are shown. Each dot represents a base-substitution type, colored according to **Figure S2C**. Since C>G mutations were not identified in the CirSeq data, we drew it on the *y*-axis.

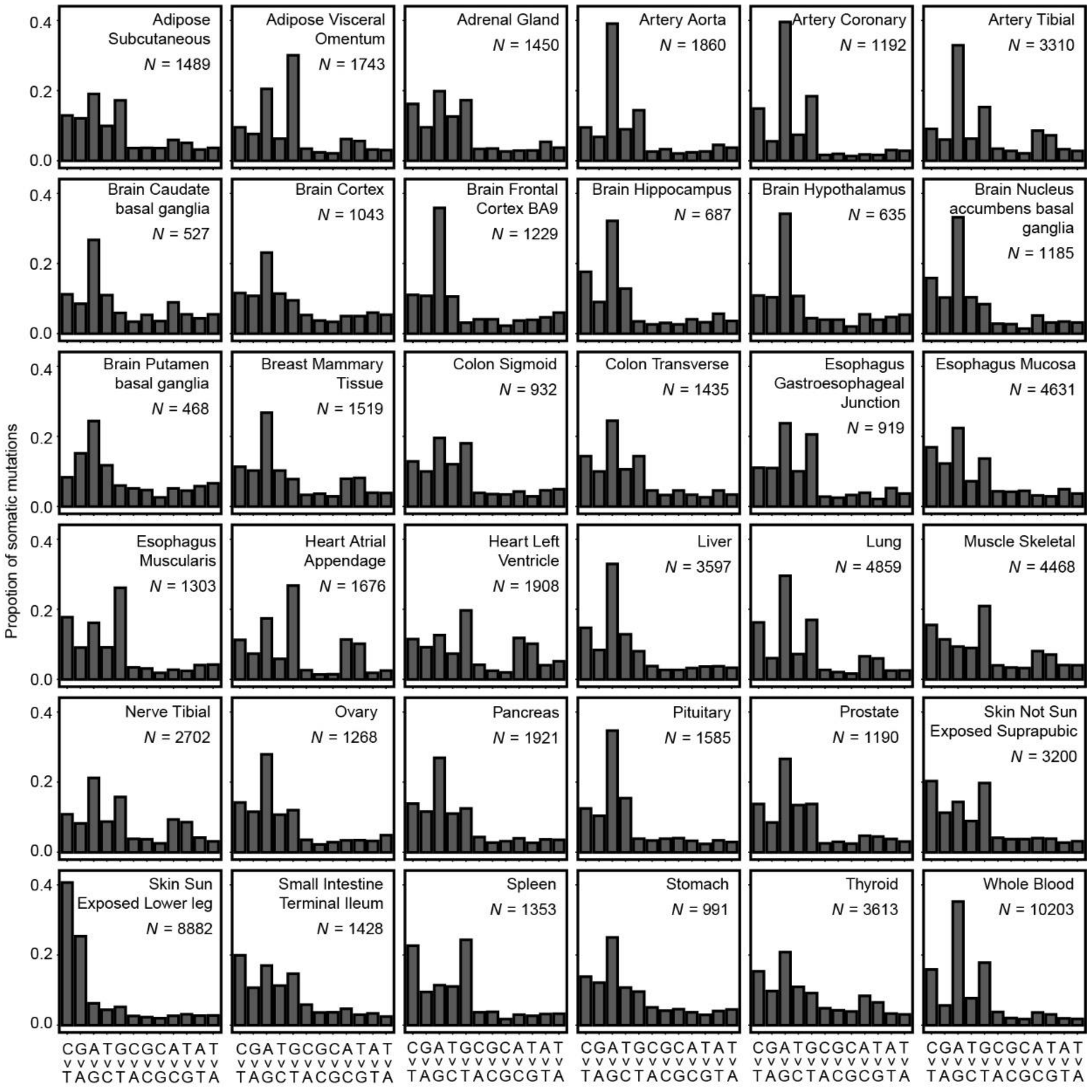


Figure S6. The molecular spectra of somatic mutations in 36 human tissues, polarized according to the coding strand. The number of total somatic mutations detected in a tissue (*N*) is shown in each panel.

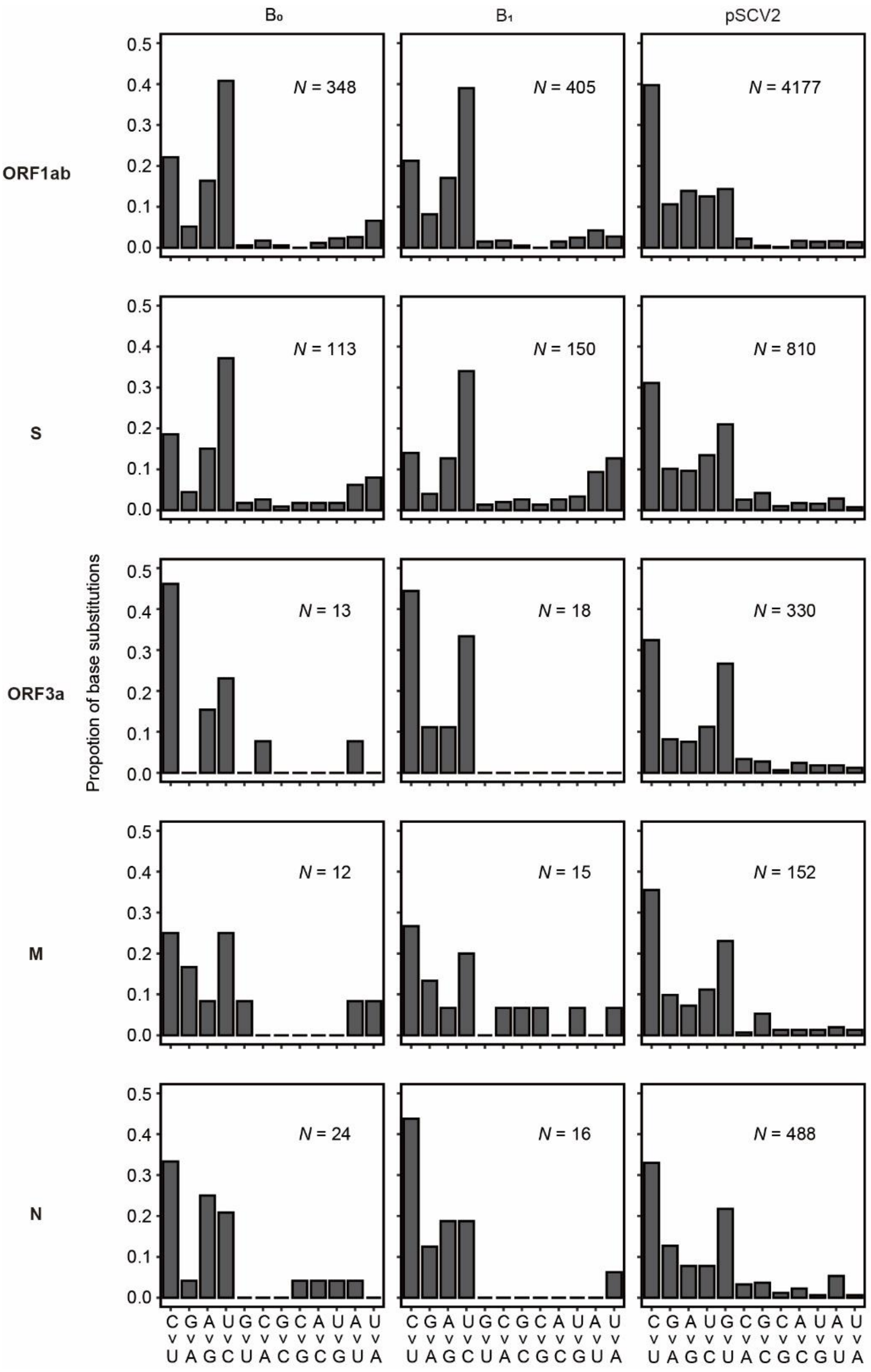


Figure S7. The molecular spectra of mutations accumulated in the branches B_0 and B_1 and among human patients, for five individual SARS-CoV-2 ORFs. The total number of base substitutions detected in a gene (*N*) is shown in each panel. The molecular spectra of the other four ORFs (E, ORF6, ORF7, and ORF8) were not shown because in these ORFs, less than 10 mutations were accumulated in either branch B_0 or B_1.