Article

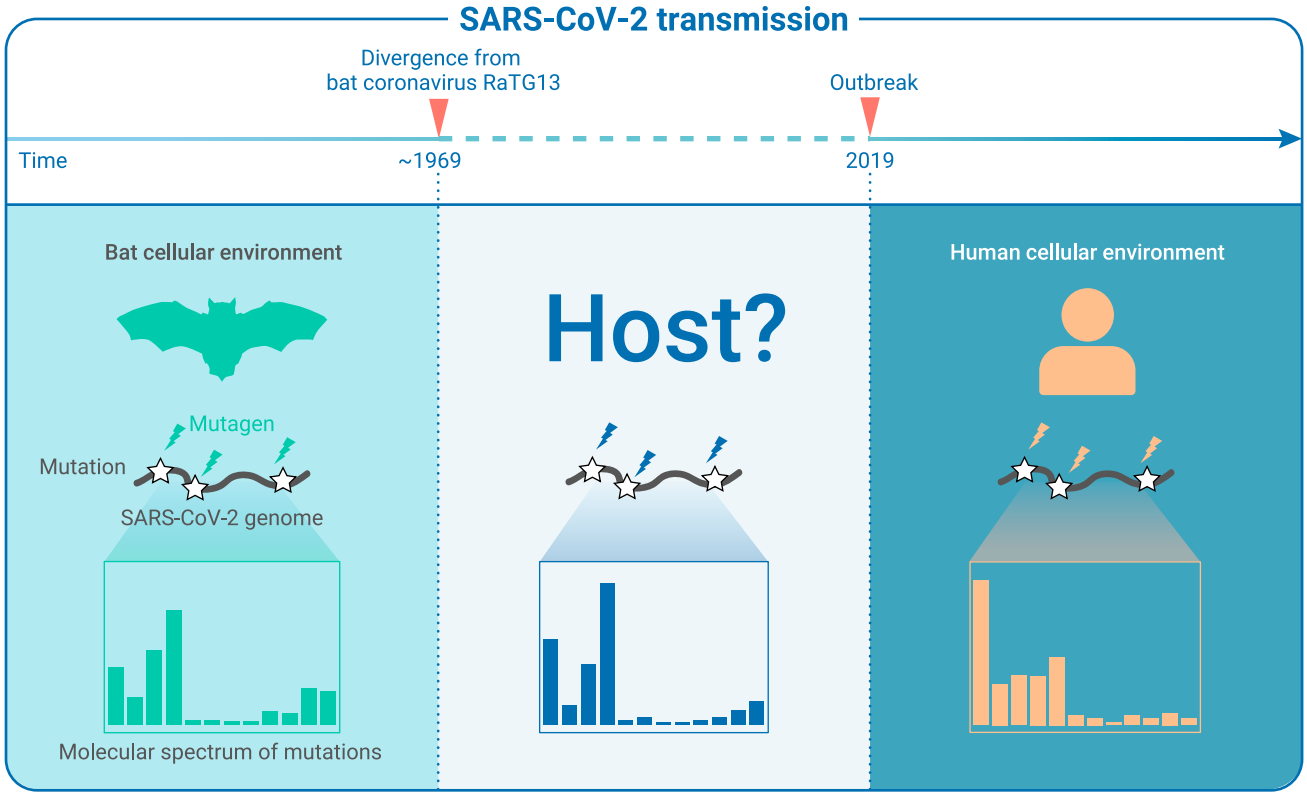
Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process

Ke-Jia Shan,[1,2,3] Changshuo Wei,[1,2,3] Yu Wang,[1,2] Qing Huan,[1,*] and Wenfeng Qian[1,2,*]

*Correspondence: qhuan@genetics.ac.cn (Q.H.); wfqian@genetics.ac.cn (W.Q.)

Received: July 19, 2021; Accepted: August 26, 2021; Published Online: August 12, 2021; https://doi.org/10.1016/j.xinn.2021.100159

Graphical abstract



Public summary

- The asymmetric *de novo* mutations in SARS-CoV-2 are induced by mutagenic mechanisms in the host cellular environment
- *De novo* mutations determine the molecular spectrum of accumulated mutations during SARS-CoV-2 evolution
- Molecular spectra of accumulated mutations in betacoronaviruses cluster according to the host species instead of the phylogenetic relationship
- The mutations accumulated in SARS-CoV-2 prior to its transmission to humans are consistent with an evolutionary process in a bat host

Article

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

Ke-Jia Shan,[1,2,3] Changshuo Wei,[1,2,3] Yu Wang,[1,2] Qing Huan,[1,*] and Wenfeng Qian[1,2,*]

[1]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Innovation Academy for Seed Design, Chinese Academy of Sciences, Beijing 100101, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

[3]These authors contributed equally

*Correspondence: qhuan@genetics.ac.cn (Q.H.); wfqian@genetics.ac.cn (W.Q.)

Received: July 19, 2021; Accepted: August 26, 2021; Published Online: August 12, 2021; https://doi.org/10.1016/j.xinn.2021.100159

Citation: Shan K.-J., Wei C., Wang Y., et al., (2021). Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process. The Innovation **2(4)**, 100159.

The capacity of RNA viruses to adapt to new hosts and rapidly escape the host immune system is largely attributable to *de novo* genetic diversity that emerges through mutations in RNA. Although the molecular spectrum of *de novo* mutations—the relative rates at which various base substitutions occur—are widely recognized as informative toward understanding the evolution of a viral genome, little attention has been paid to the possibility of using molecular spectra to infer the host origins of a virus. Here, we characterize the molecular spectrum of *de novo* mutations for SARS-CoV-2 from transcriptomic data obtained from virus-infected cell lines, enabled by the use of sporadic junctions formed during discontinuous transcription as molecular barcodes. We find that *de novo* mutations are generated in a replication-independent manner, typically on the genomic strand, and highly dependent on mutagenic mechanisms specific to the host cellular environment. *De novo* mutations will then strongly influence the types of base substitutions accumulated during SARS-CoV-2 evolution, in an asymmetric manner favoring specific mutation types. Consequently, similarities between the mutation spectra of SARS-CoV-2 and the bat coronavirus RaTG13, which have accumulated since their divergence strongly suggest that SARS-CoV-2 evolved in a host cellular environment highly similar to that of bats before its zoonotic transfer into humans. Collectively, our findings provide data-driven support for the natural origin of SARS-CoV-2.

Keywords: SARS-CoV-2; molecular spectrum; *de novo* mutations; mutational signature; evolutionary origin; mRNA mutation

INTRODUCTION

Since the first reports of coronavirus disease 2019 (COVID-19), controversies have persisted regarding the origin of its causative agent, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2).[1] While many studies have proposed that a natural origin of SARS-CoV-2 provides the simplest explanation for its emergence,[2–5] counter arguments have speculated that an accidental laboratory escape of an engineered SARS-like coronavirus could not be excluded.[6,7] One factor driving this prolonged controversy is a lack of empirical data that clearly support either possibility. Moreover, the search for related viruses in wild animals that are sufficiently genetically similar to SARS-CoV-2 has not yet shown fruitful results. Since its divergence from RaTG13—the genetically most similar virus identified to date—approximately 50 years ago,[8–11] SARS-CoV-2 had accumulated ~500 mutations before its jump to human hosts (RaTG13 accumulated ~600 mutations meanwhile).

As per the traditional aphorism, "the absence of evidence is not evidence of absence," especially when considering the vast number of unexplored wild animals and the even greater number of viruses they harbor. While the considerable efforts of many research groups to search in nature for a closely related coronavirus may yet provide many insights into the origins of SARS-CoV-2, we instead turned our attention to the ~500 base substitutions that have accumulated in SARS-CoV-2, because, we hypothesized, they could provide us with unprecedented statistical power to test if they accumulated through an evolutionary process that was consistent with those occurring in known, natural coronaviruses.

Virus evolution begins with a *de novo* mutation in its genome, thus providing new variations in the genetic material that is retained or lost under different selection pressures. On the one hand, mutations that confer a fitness advantage or disadvantage will increase or decrease in frequency through natural selection. For example, some mutations may affect transmission efficiency or capacity to escape from the host immune system. On the other hand, neutral mutations, which have little fitness effect, remain unaffected by natural selection, resulting in their genomic accumulation of an equal chance through random genetic drift.

Since *de novo* mutations are the starting point for genetic variation, their molecular spectrum—i.e., the relative rates at which all 12 possible types of base substitutions arise—has been widely recognized as an essential parameter for understanding genome evolution. Since this spectrum of mutation rates can be used to predict sequence changes under neutral processes, this metric can serve as a null model for optimizing phylogenetic reconstructions based on maximum likelihood or for detecting genomic signals of positive selection.[12–18]

In addition to its applications in evolutionary biology, the molecular spectrum has been used to describe somatic mutations accumulated in the genome of cancer cells and to identify etiological agents involved in tumorigenesis. Various mutational processes, such as exposure to mutagens and enzymatic modification, will each generate unique combinations of mutation types, termed "mutational signatures." Therefore, the molecular spectrum can be used to infer the suite of operative mutational processes through which somatic mutations accumulated in the genome of a cancer cell.[19–21] For example, excess C > A or G > T transversions, mainly caused by polycyclic aromatic hydrocarbons, have been identified as a mutational signature for tobacco smoking in the development of lung cancers.[21,22]

Following the same logic, we propose that the molecular spectrum of mutations that accumulated during the evolution of a viral genome may be informative for inferring the ancestral hosts of that virus, because viruses share the same sets of mutagens in the cellular environment as their hosts. However, we realize that this strategy heavily relies on the validity of three assumptions. First, the cellular environment is substantially variable among different hosts such that they can create mutational signatures sufficiently distinct in the viral genome for tracing its transmission history. Second, *de novo* mutations in the viral genome are predominantly introduced through processes

specific to the host cellular environment, rather than through inherently viral mechanisms of mutagenesis. Third, the molecular spectrum of mutations accumulated in the evolution of a given virus is largely determined by *de novo* mutations rather than by natural selection, which in principle could blur any mutational signatures. We realize that the key to testing these assumptions is to characterize the molecular spectrum of *de novo* mutations in SARS-CoV-2, before natural selection has a chance to affect their apparent frequency.

In this study, we first tested each of these three assumptions using a computational strategy specifically developed for detecting *de novo* mutations in SARS-CoV-2 from the transcriptome of virus-infected cell lines. After validating the three assumptions, we constructed a phylogenetic tree for SARS-CoV-2 and related coronaviruses and identified hundreds of mutations that accumulated in its genome before jumping to human hosts. Finally, we investigated whether the accumulation of these mutations was compatible with other viruses in the phylogenetic tree that are reported to have a natural origin. Our data-driven investigation provides transparent and empirical support for the natural origin of SARS-CoV-2.

RESULTS

The rationale for detecting *de novo* mutations in SARS-CoV-2

The identification of *de novo* mutations in SARS-CoV-2 has been technically challenging. For example, the molecular spectrum of *de novo* mutations cannot be inferred from within-individual polymorphisms in samples of bronchoalveolar lavage fluid[23–25] or from mutations that accumulated among patients (i.e., among-patient polymorphisms),[16,26,27] because we are specifically concerned with the extent to which the molecular spectrum of mutations that accumulated during virus evolution reflect the molecular spectrum of *de novo* mutations (refer to the third aforementioned assumption).

For evolutionary genomics studies, *de novo* mutations are ideally detected in newly synthesized virus genomes, before natural selection has a chance to act, for example, using RNA sequencing data from SARS-CoV-2-infected cells. However, *de novo* RNA mutations in SARS-CoV-2 cannot be directly inferred from mismatches between sequencing reads and the reference genome because of the high error rate inherent to high-throughput sequencing (10^{-3} to 10^{-4} errors per nucleotide), which is approximately two orders of magnitude higher than the average *de novo* RNA mutation rate (10^{-5} to 10^{-6} mutations per nucleotide).[28] Furthermore, errors generated during library preparation—that is, in the procedures of reverse transcription and polymerase chain reaction (PCR)-based amplification—will also increase the complexity of identifying bona fide RNA mutations.[29]

Nevertheless, experimental strategies have been designed to detect rare *de novo* mutations in RNA viruses, such as for poliovirus[30] and for Ebola virus.[31] Acevedo et al. developed circular sequencing (CirSeq) in which RNA molecules are first circularized, then serve as the template for rolling circle reverse transcription; bona fide RNA mutations will appear periodically in the resultant complement DNA[30] (Figure 1A, left panel). Different from CirSeq, replicated sequencing (Rep-seq), developed by Gout et al., could also be used to detect *de novo* mutations in RNA viruses, although this method was originally developed for detecting *de novo* mRNA mutations—the differences in sequence between an mRNA and its template DNA. In Rep-seq, each mRNA is barcoded with a unique oligonucleotide and is then reverse transcribed three times. Mismatches repeatedly observed in the sequencing reads that share the same barcode are considered to be bona fide RNA mutations that were extant in the mRNA[29] (Figure 1A, right panel). In addition, a hybrid strategy named accurate RNA consensus sequencing (ARC-seq) was later developed, which uses a rolling circle strategy for multiple times of reverse transcription of an RNA molecule in conjunction with oligonucleotide barcodes for recognizing reads of the same RNA molecule.[32]

Here, we sought to develop a strategy for detecting *de novo* mutations in SARS-CoV-2. SARS-CoV-2 is a positive-sense single-strand RNA virus,[8,33,34] replicating its genome within host cells through two rounds of transcription using an RNA-dependent RNA polymerase (RdRp) encoded in the viral genome: the RdRp first transcribes the positive-sense genomic RNA to generate a few intermediate negative-sense genomic RNAs that can then serve as a template to transcribe several positive-sense RNA genomes. These positive-sense RNA genomes are then packed into individual virions (Figure S1A, left panel). This two-round transcription mechanism is also employed by SARS-CoV-2 to synthesize various positive-sense subgenomes that function as viral mRNAs for the translation of viral proteins (Figure S1A, right panel).

The SARS-CoV-2 subgenomes are nested within the genomic RNA and are produced by discontinuous transcription from positive-sense genomic RNA into intermediate negative-sense subgenomic RNA (i.e., via polymerase jumping, Figure S1B). In addition to canonical junctions generated by the leader-to-body fusion occurring between the leader and one of the eight body transcription-regulating sequences, a huge number of noncanonical fusions can be found at random sites in the viral genome.[35] Most resultant noncanonical junctions are present at a low frequency, likely resulting from sporadic errors in discontinuous transcription.[36] Nevertheless, the negative-sense subgenomic RNA bearing such sporadic junctions can serve as a template for transcription into multiple identical positive-sense subgenomic RNAs.[37]

We realized that sporadic junctions could serve as the molecular barcode for a negative-sense subgenomic RNA, since it is unlikely that two independently synthesized, positive-sense subgenomic RNAs will share identical genomic coordinates of a pair of upstream and downstream junction sites. Therefore, these "junction barcodes" can be used to group sequencing reads into families; each read family will include all sequencing reads derived from the positive-sense subgenomic RNAs that have been transcribed from the same negative-sense subgenomic RNA. Repeated detection of the same mismatch within a read family implies that an RNA mutation was present in the negative-sense subgenomes (Figure 1B). In contrast, errors generated during reverse transcription or sequencing can be excluded, as they will appear randomly (i.e., not at identical sites).

C > U and G > U are over-represented in SARS-CoV-2 *de novo* mutations

Using the junction-barcoding approach (Figures S2 and S3A), we effectively distinguished reverse transcription or sequencing errors from the bona fide RNA mutations (Figure 1C, see supplemental materials and methods for details). From the RNA sequencing data for SARS-CoV-2-infected Vero cells,[35] we identified a total of 197 *de novo* RNA point mutations in the SARS-CoV-2 genome. These mutations could be categorized into 12 distinct types with respect to the positive-sense genomic RNA (Figure 1D, with an example shown in Figure S4). To estimate the rate of each mutation type, we controlled for the nucleotide composition of the viral genome and the potential coverage bias generated during high-throughput sequencing. For this calculation, we estimated the coverage for each site by all read families (similar to RNA mutation calling in Figure S2, but no mismatch was required) and aggregated this coverage according to the nucleotide (A, C, G, or U) in the reference genome. We divided the number of mutations of each type by the total coverage of all sites with the nucleotide in the reference genome, and used this ratio to infer the molecular spectrum of mutations in SARS-CoV-2 in Vero cells (Figure 1E).

Among the 197 RNA mutations we identified in SARS-CoV-2, 122 were transitions (purine to purine or pyrimidine to pyrimidine interchanges) and 75 were transversions (interchanges of a purine to a pyrimidine or vice versa), which significantly deviated from the randomly expected ratio (4:8, $p = 3 \times 10^{-16}$, binomial test, Figure 1D). On average, transitions occurred three times more frequently (1.1×10^{-5} substitutions per site) than transversions (3.6×10^{-6} substitutions per site, Figure 1E), which was possibly attributable to the structural similarity between bases that are substituted in transitions.

In particular, C > U mutations appeared to be the most abundant transition ($p = 6 \times 10^{-5}$, binomial test with probability equal to ¼, Figure 1D). Among the eight types of RNA transversions, G > U mutations occurred much more frequently than the other seven mutation types ($p = 8 \times 10^{-6}$, binomial test with probability equal to ⅛, Figure 1D), reaching frequencies comparable with that of transitions. We then focused on these two major signatures of

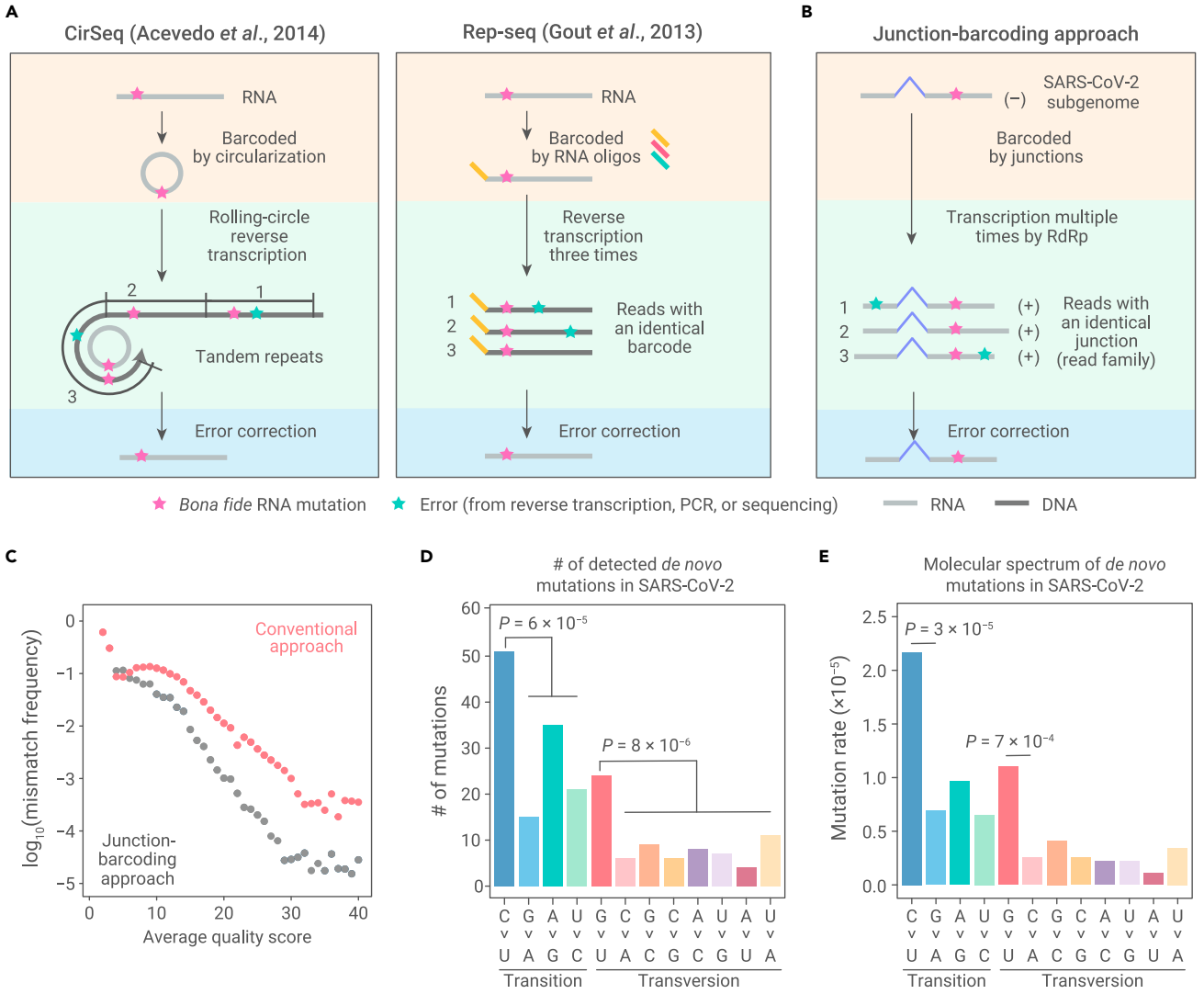


Figure 1. The molecular spectrum of *de novo* SARS-CoV-2 mutations (A) Schematic of two experimental approaches previously developed to detect RNA mutations. Bona fide RNA mutations (magenta stars) should be repeatedly detected, while errors generated during reverse transcription, PCR amplification, or high-throughput sequencing (green stars) should only be occasionally detected. (B) Schematic of our junction-barcoding approach to detect RNA mutations for SARS-CoV-2. The genomic coordinates of a pair of upstream and downstream sites of sporadic junctions can serve as the molecular barcode to group sequencing reads derived from the same negative-sense subgenome into read families. Bona fide RNA mutations should be unanimously detected in a read family. (C) Comparison of overall mismatch frequency between our junction-barcoding approach and the conventional computational approach. (D) The numbers of *de novo* RNA mutations of 12 base-substitution types, with respect to the positive-sense SARS-CoV-2 genome. Two-tailed p values were calculated from binomial tests assuming an equal frequency for each type of base substitutions. (E) The molecular spectrum of *de novo* SARS-CoV-2 mutations. Two-tailed p values were calculated from Fisher's exact tests.

SARS-CoV-2 mutations, i.e., over-representation of G > U and C > U mutations, in subsequent analyses.

The molecular spectrum of *de novo* mutations shapes the spectrum of polymorphisms during the evolution of SARS-CoV-2 in human patients

Given that the *de novo* mutations provide the raw materials for virus evolution (Figure 2A), before the further investigation of the molecular mechanisms underlying the over-representation of G > U and C > U mutations, we first sought to determine the levels at which SARS-CoV-2 evolution in human patients was affected by the molecular spectrum of *de novo* mutations. To this end, we retrieved 34,853 high-quality sequences of SARS-CoV-2 variants isolated from patients worldwide (Table S1) from GISAID (global initiative on sharing all influenza data),[38] and reconstructed the genomic sequence of their last common ancestor.

We then identified genetic differences between each variant and the ancestor and treated those differences observed in at least two patients as among-patient polymorphisms. This process enabled the removal of singletons that were potentially generated by sequencing errors. The molecular spectrum of SARS-CoV-2 among-patient polymorphisms (Figure 2B) highly resembled that of the *de novo* mutations at 4-fold degenerate sites (Figure 2C, r = 0.81, p = 0.001), as well as in the whole genome (Figures 2D and 2E, r = 0.86, p = 3 × 10^{-4}). The signatures indicating over-representation of G > U and C > U among *de novo* mutations were also observed in among-patient polymorphisms. This finding indicated that the molecular spectrum of *de novo* mutations dominated the base-substitution types of polymorphisms in SARS-CoV-2 during its evolution in human patients.

It is worth noting that, despite their apparent similarity, the *de novo* mutation identified in this study is by definition different from the among-patient polymorphisms,[16,26] because the latter has been influenced by natural selection related to the processes of infection, propagation, or release from infected cells.[13,39] Only with the characterization of the molecular spectrum of *de novo* mutations can we thus examine the relative influence of mutation versus selection in driving the genomic evolution of SARS-CoV-2 (Figure 2A). The observation that the molecular spectrum of among-patient polymorphisms largely resembled that of *de novo* mutations indicated that the

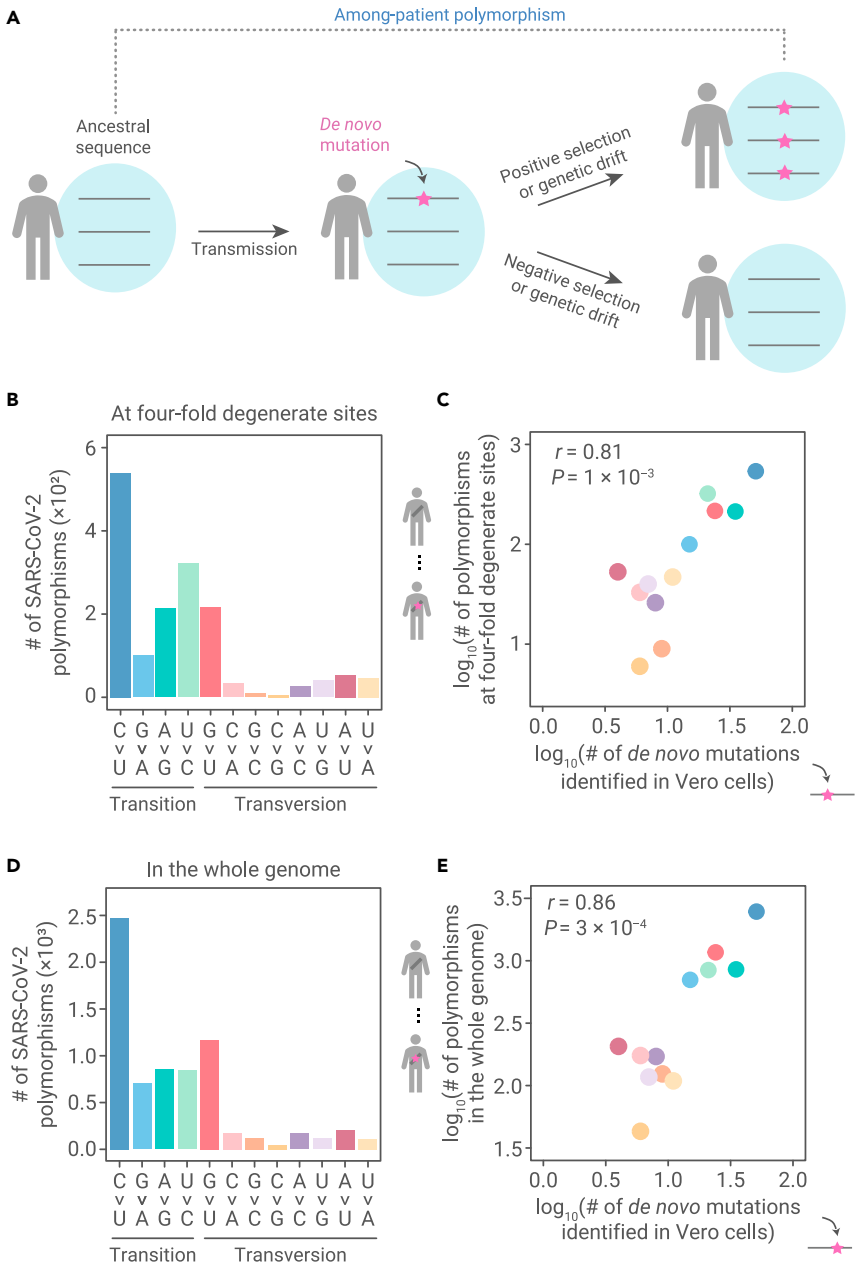


Figure 2. The molecular spectrum of SARS-CoV-2 polymorphisms among patients (A) The emergence of among-patient polymorphisms through the accumulation of *de novo* mutations. The frequency of a *de novo* mutation (the magenta star with an arrow pointing to it) may be increased by positive selection, decreased by negative selection, or changed through genetic drift due to chance events. If a mutation becomes predominant within a patient, it can be detected as an among-patient polymorphism.
(B) The molecular spectrum of among-patient polymorphisms at 4-fold degenerate sites in SARS-CoV-2.
(C) A scatterplot shows the molecular spectrum of *de novo* mutations versus among-patient polymorphisms at 4-fold degenerate sites in SARS-CoV-2. Each dot represents a base-substitution type, colored according to (B). Pearson's correlation coefficient (r) and the corresponding p value are shown.
(D) The molecular spectrum of among-patient polymorphisms in the whole genome of SARS-CoV-2.
(E) Similar to (C), for all polymorphisms.

proportions of deleterious mutations were largely uniform among the 12 types of base substitutions.

Replication-independent asymmetric emergence of mutations on single-strand RNA

Presumably, the over-representation of G > U and C > U mutations could be consequences of transcriptional errors that produce RNAs carrying different sequences from that of the template (i.e., replication-dependent mutations), or these mismatches could result from exposure to environmental mutagens that can induce mutations in the absence of transcriptional machinery (i.e., replication-independent mutations).[40] We realized that these two mechanisms could be distinguished by comparing the frequency of G > U (or C > U) mutations with that of its complement mutation, C > A (or G > A). For example, a mutation observed in the negative-sense genome (e.g., C > A, which will be transcribed into a G > U mutation in the positive-sense genome) that is generated during positive to negative strand transcription should occur at approximately equal frequency in negative to positive strand transcription (leading to a C > A mutation in the positive-sense genome), because the same polymerase performs both functions. Consequently, when the SARS-CoV-2 replication cycle (i.e., two rounds of transcription) is completed, the molecular spectrum of replication-dependent mutations should be "symmetric," meaning that the frequencies of G > U and its complement mutation, C > A, should be similar (Figure 3A, left panel). Alternatively, if a mutation is generated in a replication-independent manner, for example, induced by mutagens specifically in the positive-sense (or negative-sense) single-strand RNA of SARS-CoV-2, then the frequency of complement mutations will not necessarily be symmetrical (Figure 3A, middle panel).

We reasoned that among-patient polymorphisms in SARS-CoV-2 could be used to investigate whether the mutagenic mechanisms underlying the over-representation of some mutation types were replication dependent or independent because polymorphisms were generated by complete replication cycles. The polymorphism data revealed that G > U transversions occurred with greater frequency than C > A ($p = 2 \times 10^{-116}$, Fisher's exact

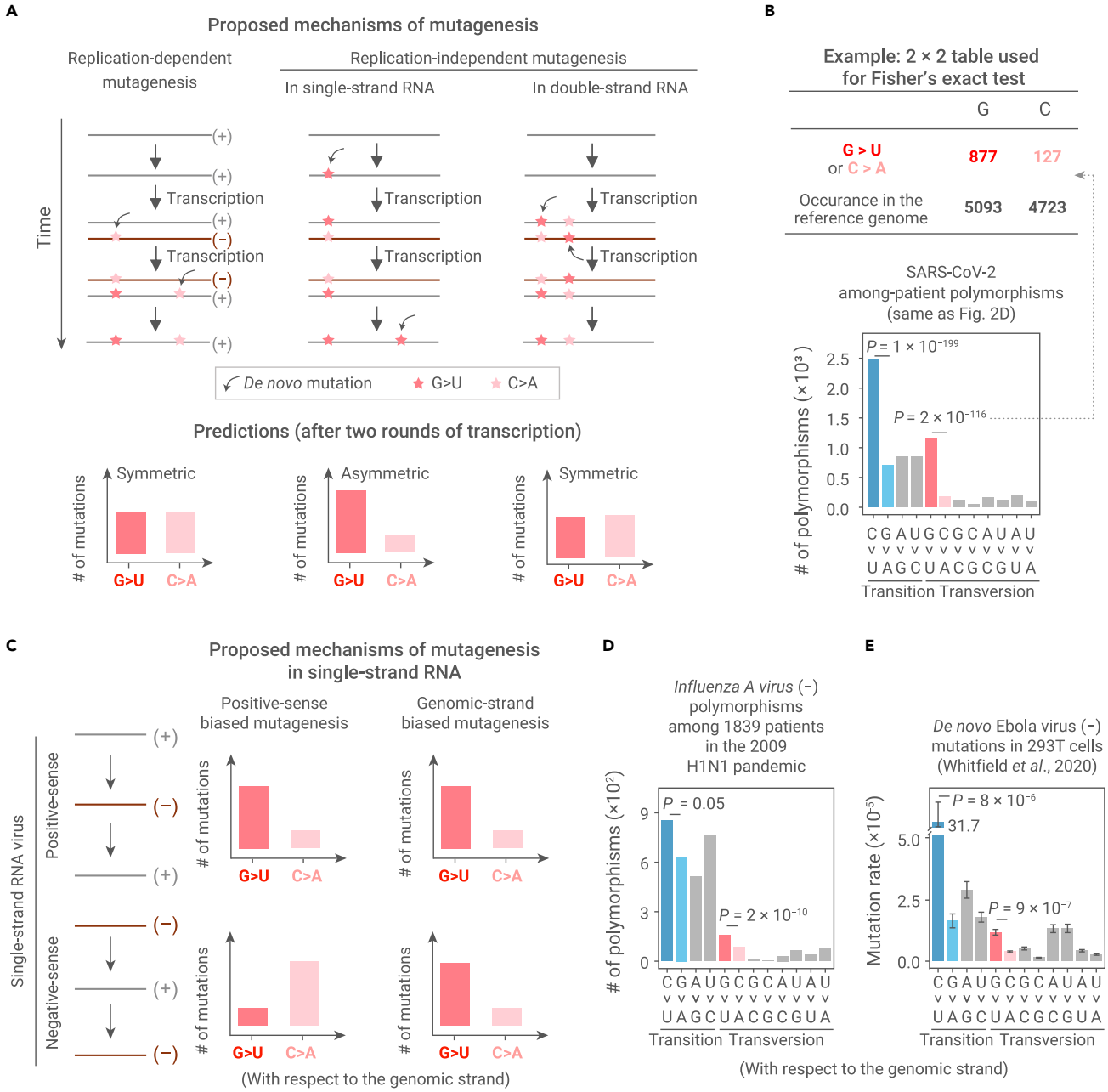


Figure 3. Predictions and observations for various mutagenic mechanisms on the symmetry of mutations (A) Predictions on the symmetry between a pair of complement base-substitution types for three potential mutagenic mechanisms. If *de novo* mutations are introduced during transcription by RdRp (left panel), or by a replication-independent mechanism in double-strand RNAs (right panel), mutations should be symmetric when a replication cycle is completed: a base-substitution type and its complement base-substitution type should arise at the same rate in the viral genome. On the contrary, if *de novo* mutations are introduced by a replication-independent mechanism specific to single-strand RNAs, mutations could be asymmetric (middle panel). (B) The statistical assessment on the symmetry of mutations using Fisher's exact tests. (C) Predictions for two potential mutagenic mechanisms in single-strand RNAs, positive-sense biased versus genomic-strand biased mutagenesis. (D) The molecular spectrum of among-patient polymorphisms in a negative-sense, single-strand RNA virus, *Influenza A virus* (subtype H1N1). Two-tailed p values were calculated from Fisher's exact tests. (E) The molecular spectrum of *de novo* mutations in a negative-sense, single-strand RNA virus, Ebola virus. *De novo* mutations were identified from isolated virions, at which time replication cycles have completed. Error bars represent standard errors (N = 21) of the average mutation rates of each base-substitution type. Two-tailed p values were calculated using the t tests.

test) and that C > U transitions occurred with greater frequency than G > A ($p = 1 \times 10^{-199}$, Figure 3B). This asymmetric distribution indicates that the observed over-representation of G > U and C > U mutations unlikely results from a replication-dependent process.

Presumably, replication-independent mutations can occur in either double- or single-strand RNA. We reasoned that mutations arising in double-strand RNA should also lead to a symmetric molecular spectrum (Figure 3A, right panel). This possibility was excluded by the observation that, among SARS-CoV-2 isolates from human patients, G > U and C > U polymorphisms were distributed in asymmetrically greater numbers (Figure 3B), supporting the likelihood that the mechanism responsible for introducing these mutations involved single-strand RNA. Furthermore, since the negative-sense RNA of SARS-CoV-2 is mainly present in the double-strand RNA (i.e., paired with positive-sense RNA; Figure 3A), we therefore proposed that the observed G > U and C > U mutations were most likely introduced to the single-strand positive-sense RNA.

Article

Asymmetric emergence of *de novo* mutations on the genomic-strand RNA

Thus far, our results indicated that the disproportionate abundance of G > U and C > U mutations in SARS-CoV-2 likely arose in positive-sense single-strand RNAs. There are two possible mechanisms that could account for this outcome. First, the positive-sense RNA is more vulnerable to mutagens—for example, due to destruction of the RNA secondary structure by translating ribosomes, which subsequently exposes the single-strand RNAs to mutagens. Second, the viral genetic information spends the majority of its life cycle as a positive-sense RNA. We thus reasoned that the molecular spectrum of mutations in negative-sense single-strand RNA viruses could be used to investigate which of the two mechanisms underlay the emergence of single-strand RNA mutations, since the negative-sense RNA was predominant in these viruses (Figure 3C).

We first assessed this mechanism by analyzing the genetic polymorphisms among 1,839 confirmed variants (Table S2) of the negative-sense single-strand RNA virus, *Influenza A virus*, collected during the 2009 pandemic.[38] The asymmetric frequency of the among-patient polymorphisms that we observed in SARS-CoV-2, a positive-sense RNA virus (i.e., over-representation of G > U and C > U on the positive strand), was reversed in *Influenza A virus*: G > U and C > U polymorphisms were more abundant in the negative-sense genome ($p = 2 \times 10^{-10}$ and 0.05, respectively, Fisher's exact tests, Figure 3D). This result allowed us to exclude the possibility that a positive-sense-specific mechanism was responsible for the over-representation of G > U and C > U mutations.

Furthermore, the molecular spectrum of mutations in the negative-sense single-strand RNA virus, Ebola (*Zaire ebolavirus*), was previously characterized using CirSeq with virions isolated from 293T cells,[31] which have completed cycles of replication. Asymmetric accumulation of G > U and C > U mutations were observed in the negative-sense genomic RNA of the Ebola virus (Figure 3E), which further supported a replication-independent mutation mechanism on single-strand RNAs that acts on the strand carrying the genetic information.

Asymmetric emergence of *de novo* RNA mutations in host cellular environment

In light of these findings, we next sought to determine whether these replication-independent mutations were introduced to the genomic-strand RNA in the extracellular virion environment, where the viral RNA is protected by the capsid, or if they occurred in the cellular environment following host invasion (Figure 4A). To address this issue, we reasoned that positive-sense single-strand, persistent yeast RNA viruses (e.g., *Saccharomyces 20S RNA narnavirus* and *Saccharomyces 23S RNA narnavirus*) could be used to test if the eukaryotic cellular environment was able to induce G > U or C > U mutations in single-strand RNAs (Figure 4A). These viruses represented a strong experimental model for this question because they persist in yeast cells as naked RNA, without a capsid.[41] Therefore, the introduction of G > U and C > U mutations is dependent on mutagenic mechanisms within the yeast cellular environment, and without which the asymmetric accumulation of mutations will not be detectable (Figure 4A).

Both CirSeq[42] and ARC-seq[32] experiments were previously conducted in budding yeast, although the aims of the previous studies were to identify mutations in endogenous RNAs. We reasoned that some reads might be derived from the persistent yeast RNA viruses, which can be used to calculate the symmetry of mutations in the virus genomes. While the yeast strain used in the CirSeq generated minimal reads from either 20S or 23S RNA narnaviruses, the ARC-seq data contained 43,034 sequencing reads from the 20S RNA narnavirus genome (but no reads from the 23S RNA narnavirus). Among the reads that mapped to the 20S RNA narnavirus genome, we identified a significantly higher abundance of G > U mutations than C > A mutations (46 versus 22, p = 0.003, Fisher's exact test) and more C > U than G > A mutations (157 versus 69, $p = 1 \times 10^{-8}$, Fisher's exact test, Figure 4B). These results indicated that the yeast cellular environment was sufficient to induce the asymmetric emergence of G > U and C > U mutations in a positive-sense single-strand RNA viral genome.

Postulating that mutagens in the cellular environment cannot discriminate endogenous and viral RNAs, we further predicted that G > U and C > U mutations should also be over-represented in yeast endogenous mRNAs. To test this prediction, we characterized the molecular spectrum of mRNA mutations in the same ARC-seq dataset for the budding yeast.[32] The results showed that the molecular spectrum of mutations was highly similar between the 20S RNA narnavirus and that of the yeast-derived mRNAs (r = 0.93, $p = 2 \times 10^{-5}$, Figures 4C and 4D). Furthermore, G > U and C > U mutations occurred more frequently in the endogenous mRNAs than C > A and G > A mutations, respectively ($p < 10^{-2311}$ and 10^{-9126}, respectively, Fisher's exact tests, Figure 4C). A similar molecular spectrum of mRNA mutations was also observed in the CirSeq data for the budding yeast[42] (r = 0.84, $p = 1 \times 10^{-3}$, Figure S5). The similarities between these molecular spectra indicated the possibility that mutations were induced by the same mutagens in the cellular environment of yeast.

Variation among host cells in providing the cellular environment for asymmetric mutations

Although the particular mutagenic mechanisms that caused the observed asymmetric G > U mutations in the cellular environment remain unknown, we hypothesized that reactive oxygen species (ROS) could serve as a strong candidate,[43] particularly considering that oxidative stress is associated with the infection of some respiratory viruses.[44] Specifically, some ROS can oxidize guanine to 8-oxoguanine and thereby induce G > U transversions after an additional round of transcription.[45,46] Alternatively, we also suspected that chemicals with similar property to polycyclic aromatic hydrocarbons, which are also well known to induce G > T somatic mutations in the lung cancer samples among tobacco smokers,[47] could serve as potential cytosolic mutagens of viral RNA. For C > U mutation, potential candidates that could induce its accumulation included RNA-editing activity by cytidine deaminases.[48,49]

Given the broad array of potential mechanisms, we reasoned that, regardless of the exact nature of the mutagens that caused asymmetric accumulation of G > U or C > U RNA mutations in the cellular environment, these mutagens were unlikely to discriminate between RNA and DNA.[50] Consequently, somatic mutations in DNA would also arise,[51–53] particularly in the coding strand, which is exposed to the cellular environment in the single-strand state during transcription[54–56] (illustrated in Figure 5A). Based on this assumption, we investigated the capacity of the cellular environment to generate G > T and C > T somatic mutations in the coding strand of genomic DNA to subsequently infer its capacity to induce G > U and C > U mutations in RNA viruses. To this end, we retrieved the somatic mutations identified for each of the 36 human tissues[55] from publicly available Genotype-Tissue Expression (GTEx) data.[57]

We characterized the molecular spectra of somatic mutations that emerged in these 36 human tissues (Figure S6) and projected them into a two-dimensional space based on the levels of asymmetry in G > T (versus C > A) and C > T (versus G > A) base substitutions (Figure 5B). Among them, 20 tissues, including the lung, showed asymmetry in both G > T and C > T mutations, while 10 tissues showed asymmetry only in G > T mutations. The cellular environments of the remaining six tissues could not induce detectable asymmetry in either G > T or C > T. These results indicated that human tissues exhibit wide-ranging differences in their capacity to induce various types of *de novo* mutations.

To further investigate the cellular environment that likely drove SARS-CoV-2 evolution in human patients, we plotted the asymmetries of G > U and C > U for SARS-CoV-2 among-patient polymorphisms (abbreviated as pSCV2 in the figure) into the two-dimensional space and found that the cellular environment where SARS-CoV-2 propagated in patients was most similar to that of the lung (Figure 5B). This finding is in agreement with numerous reports that showed the airborne transmission[58] of SARS-CoV-2 and supports a cellular environment-dependent genomic evolution of SARS-CoV-2.

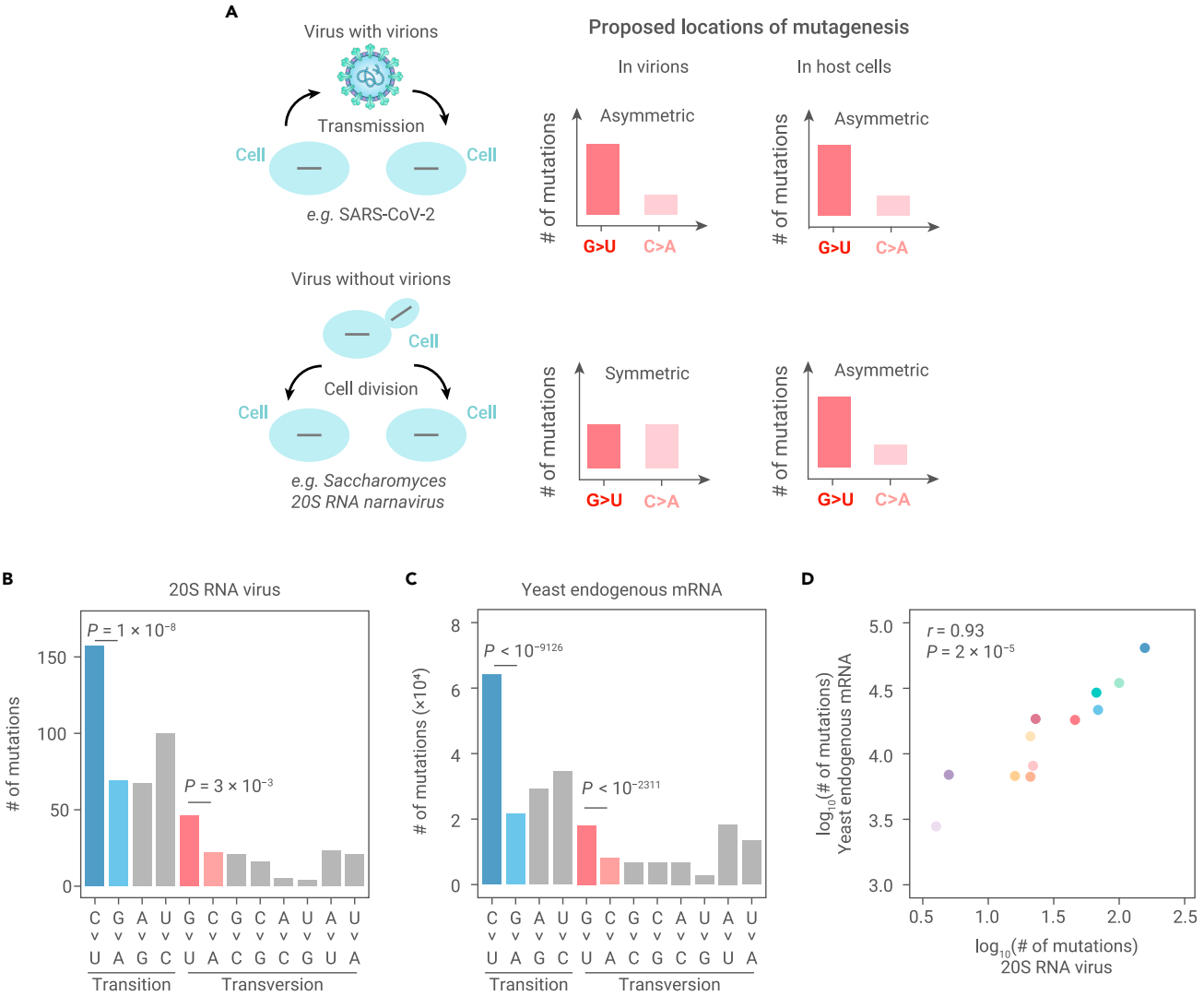


Figure 4. Predictions and observations for mutagenic processes in virions versus in host cells (A) Predictions on the symmetry of mutations for mutagenic processes in virions versus in host cells. (B) The molecular spectrum of *de novo* mutations that we detected in 20S RNA narnavirus from previously published ARC-seq data. Two-tailed p values were calculated from Fisher's exact tests. (C) The molecular spectrum of yeast mRNA mutations that we detected from previously published ARC-seq data. Two-tailed p values were calculated from Fisher's exact tests. (D) A scatterplot shows the molecular spectrum of *de novo* mutations in 20S RNA narnavirus versus in yeast endogenous mRNAs. Each dot represents a base-substitution type, colored according to Figure 1E. Pearson's correlation coefficient (r) and the corresponding p value are shown.

The molecular spectrum of 529 accumulated mutations in SARS-CoV-2 prior to its transmission to humans resembled that of coronaviruses evolved in bats

Thus far, our results showed that the molecular spectrum of mutations that accumulated during SARS-CoV-2 evolution is reflective of the asymmetric emergence of *de novo* mutations (Figure 2) caused by host cellular environments (Figures 3 and 4). Given that different types of base substitutions are disproportionately induced in various cell environments (Figure 5), we reasoned that the ancestral cellular environment where SARS-CoV-2 propagated prior to its transmission to humans could, in principle, be inferred from the mutations in the SARS-CoV-2 genome that accumulated during that period. These mutations could be identified from a phylogenetic tree including SARS-CoV-2 and related coronaviruses.

We built an evolutionary tree (Figure 6A), including the last common ancestor of SARS-CoV-2 isolated from patients and its closely related coronaviruses isolated from *Rhinolophus* bats (RaTG13, RshSTT200, and ZC45) and pangolins (GD-1 and GX-P5L), using Rc-o319 from bats as an outgroup.[8,59–63] We then reconstructed the ancestral sequence for each internal node (N_1–N_5) and determined which mutations accumulated in the evolutionary history represented by each branch in the phylogenetic tree (B_1–B_9, Figure 6A). Based on the parsimony principle we labeled seven branches (B_1–B_4 and B_6–B_8) that represented the evolutionary history exclusively in the cellular environments of bats and two that represented a mixed evolutionary history in bats and pangolins (B_5 and B_9, Figure 6A).

The results showed that the molecular spectra of the nine branches appeared similar (Figure 6A, inset on top of each branch), and were highly correlated with each other (Figure 6B). However, these spectra were only moderately correlated with the spectrum of mutations that accumulated during SARS-CoV-2 evolution in human patients, the spectrum of *de novo* viral mutations in Vero cells, or the spectrum of somatic mutations in the lung (Figure 6B). For example, the asymmetric emergence and accumulation of G > U mutations in the viral genome, which was observed in Vero cells (Figure 1D) and among human patients (Figure 2C), respectively, was no longer detectable in the seven bat-exclusive or two bat-pangolin branches ($p > 0.5$ in all one-sided Fisher's exact tests). This finding is in agreement with previous reports of low production of ROS and high concentrations of endogenous antioxidants in bat cells.[64] These observations indicated that bats (probably

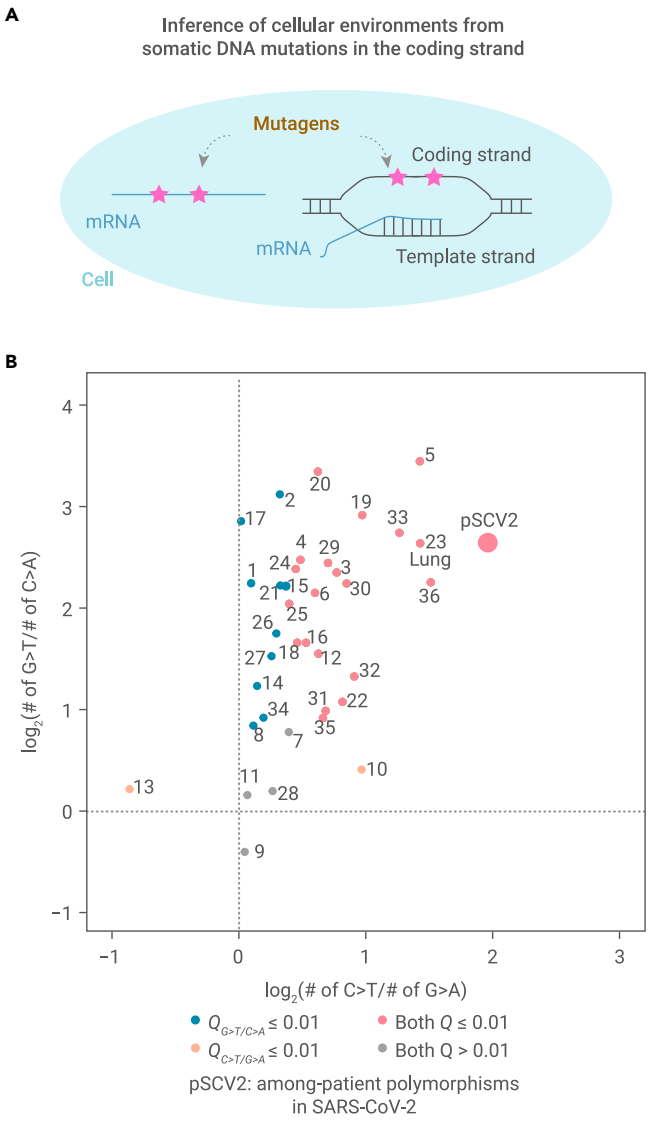


Figure 5. Variation among 36 human tissues in providing the cellular environment for asymmetric mutations in RNA viruses (A) The rationale underlying assessment of cellular environments in generating asymmetric mutations in RNA based on somatic mutations in the coding strand of DNA. (B) A scatterplot shows the asymmetric accumulation of two types of somatic mutations among 36 human tissues. 1, adipose subcutaneous; 2, adipose visceral omentum; 3, adrenal gland; 4, artery aorta; 5, artery coronary; 6, artery tibial; 7, brain caudate basal ganglia; 8, brain cortex; 9, brain frontal cortex BA9; 10, brain hippocampus; 11, brain hypothalamus; 12, brain nucleus accumbens basal ganglia; 13, brain putamen basal ganglia; 14, breast mammary tissue; 15, colon sigmoid; 16, colon transverse; 17, esophagus gastroesophageal junction; 18, esophagus mucosa; 19, esophagus muscularis; 20, heart atrial appendage; 21, heart left ventricle; 22, liver; 23, lung; 24, muscle skeletal; 25, nerve tibial; 26, ovary; 27, pancreas; 28, pituitary; 29, prostate; 30, skin not sun-exposed suprapubic; 31, skin sun-exposed lower leg; 32, small intestine terminal ileum; 33, spleen; 34, stomach; 35, thyroid; 36, whole blood. Odds ratios and two-tailed p values were calculated with Fisher's exact tests. Dots were colored according to the false discovery rates (Q values).

also pangolins) provided a cellular environment for the genomic evolution of RNA viruses that substantially differed from that of humans.

We determined the 529 base substitutions that apparently accumulated in the SARS-CoV-2 genome since its divergence from RaTG13 (represented by branch B_0 in Figure 6A). The molecular spectrum of this branch (Figure 6A, inset in the top left corner) was highly correlated with the bat-related branches (B_1–B_9), but showed a much lower correlation with the spectrum of mutations that accumulated in human patients (pSCV2, Figures 6B and 6C). The patterns held when only fragments of the SARS-CoV-2 genome (e.g., the coding sequence of the spike protein) were used for the comparison (Figure S7). These observations suggested that, after its divergence from RaTG13, SARS-CoV-2 likely evolved in a host cellular environment similar to that of bats prior to its zoonotic transfer into humans.

The apparent similarity in the molecular spectra between the mutations accumulated in branch B_0 and those in the bat-related branches (B_1–B_9) could be either attributable to their propagation in a common cellular environment, or a common inherent mutational bias caused by shared genetic variation in the genes that modulate the relative rates of various *de novo* base substitutions among closely related viruses. To distinguish these two possibilities, we included the genetically more distant betacoronaviruses, SARS-CoV and MERS-CoV and their related viruses, in the analysis (Figure 7A). Bats were reported to be the natural host for both SARS-CoV and MERS-CoV, and zoonotic transfers from their putative intermediate hosts (market civets and dromedary camels) into humans led to SARS and MERS outbreaks in 2003 and 2012, respectively.[65–67]

We constructed phylogenetic trees separately for SARS-CoV-related and MERS-CoV-related viruses, and labeled putative host species for each branch according to the parsimony principle (Figure 7A). If the molecular spectrum was shaped largely by the host cellular environment, we predicted that all branches representative of evolutionary history in the same host species would exhibit similar molecular spectra. On the contrary, if the molecular spectrum was an inherent feature encoded in the virus genome that becomes more distinct as genetic distance increases, we predicted that the lineages of SARS-CoV-2, SARS-CoV, and MERS-CoV would each exhibit their own mutational signatures.

To test these two predictions, we estimated the proportions of each base-substitution type for each branch (Figure 7A). To visualize the similarity across molecular spectra, we performed principal-component analysis, projecting these branches into a two-dimensional space (Figures 7B and 7C). The results showed that 17 branches with a reported evolutionary history exclusive to bats clustered together (Figure 7B), while the viruses from three distinct lineages of betacoronaviruses did not (Figure 7C). This observation indicated that the molecular spectrum of virus genome evolution mainly reflected the cellular environment in which viruses propagated, rather than their phylogenetic relationship. We consequently used the 95% confidence ellipse estimated from these 17 bat-exclusive branches (Figure 7B) to define the borderline of the bat cellular environment in the two-dimensional space.

Branches that represented the host history in camels (B_{18} and B_{19}) fell outside of the 95% confidence ellipse (Figure 7B), indicating that camels had a distinct cellular environment from that of bats. Branches that represented host history entirely in humans (pSCV2), or a host history partly in humans (B_{10} for a SARS patient and pMERS for among-patient polymorphisms in MERS), also fell outside of the 95% confidence ellipse (Figure 7B). Notably, they appeared to cluster together with the spectra of *de novo* mutations detected in SARS-CoV-2 (mSCV2), Ebola virus (mEbola), or poliovirus (mPV), which were cultivated in primate cell lines. Furthermore, the 95% confidence ellipse of these six human-related molecular spectra was not overlapped with that estimated from the spectra of the 17 bat-exclusive branches (Figure 7B), highlighting the potential application of our approach for detecting a jumping event from bats to a new host.

The branch leading to SARS-CoV-2 (B_0) was located within the 95% confidence ellipse defined by the 17 bat-exclusive branches (Figure 7B) and, in particular, was within the 95% confidence ellipse defined by the 13 *Rhinolophus*-exclusive branches. Considering the consistency of this approach in identifying well-established host jumping events (e.g., from bats to camels for MERS-CoV as shown in branches B_{18} and B_{19}, and from bats to humans for SARS-CoV as shown in branch B_{10}, Figure 7B), we concluded that since its divergence from RaTG13, SARS-CoV-2 most likely propagated primarily in a cellular environment highly similar to bats, particularly *Rhinolophus* bats, prior to its zoonotic spillover into humans.

DISCUSSION

The central dogma of molecular biology asserts that the genetic information of cellular organisms is stored in DNA, which must be transcribed into

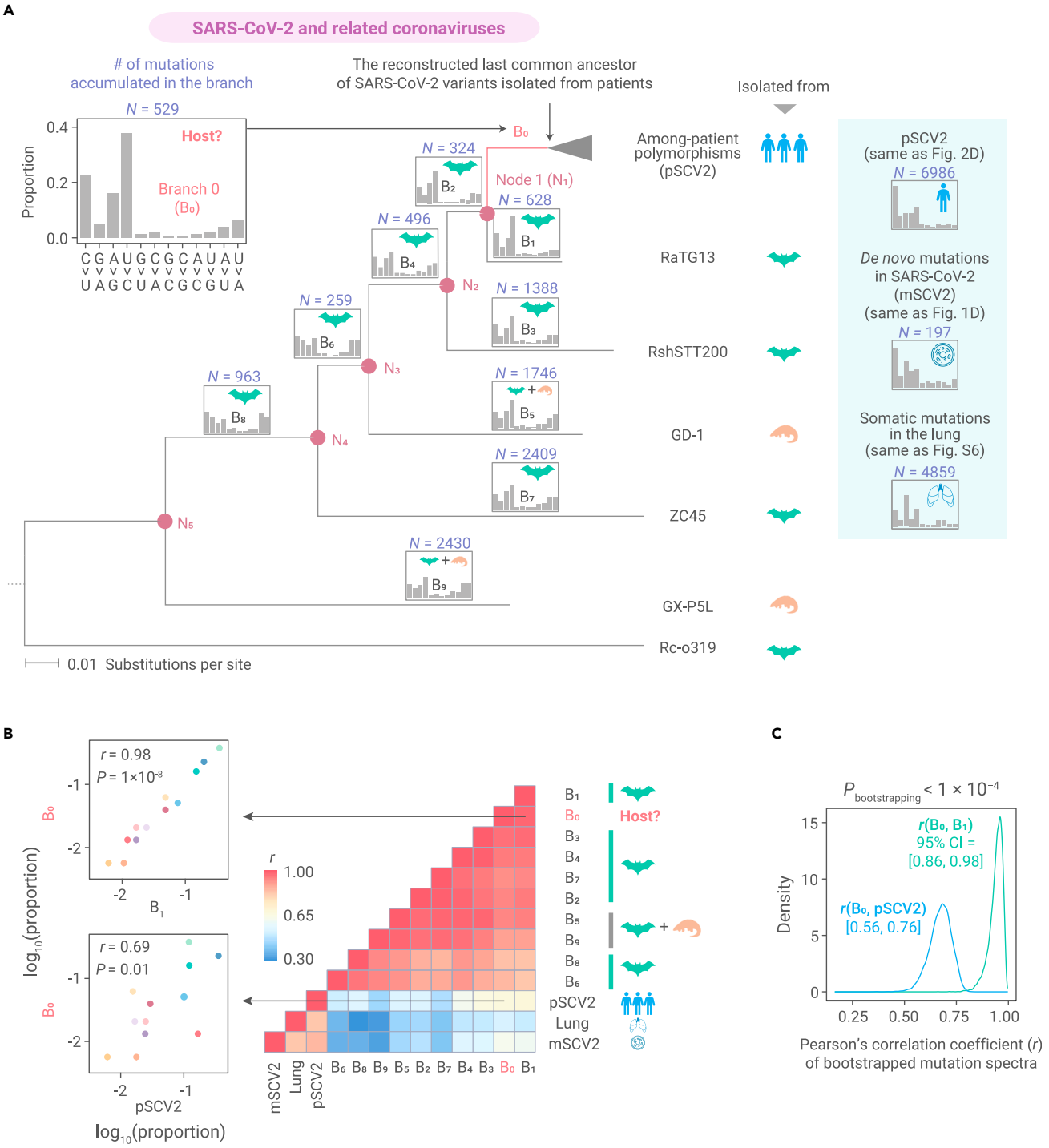


Figure 6. The molecular spectra of mutations accumulated in SARS-CoV-2 and related viruses (A) The maximum likelihood phylogenetic, tree including SARS-CoV-2 and related coronaviruses, using Rc-o319 as an outgroup. Internal nodes are labeled as N_1–N_5, and the icon on the side of a tip indicates the host species from which a SARS-CoV-2-related virus was isolated. The branches are labeled as B_0–B_9, among which the red branch (B_0) represents the evolutionary history in which the host organism is to be determined. The molecular spectrum of accumulated mutations is shown on the top of each branch, and the icon inside shows the inferred host species for the branch according to the parsimony principle. (B) A heatmap shows Pearson's correlation coefficient (r) between a pair of molecular spectra. Two scatterplots are shown to exemplify the similarity in the molecular spectrum. (C) The distribution of r for the bootstrapped mutation spectra. In all 10,000 paired bootstrapped observations, r(B_0, B_1) was greater than r(B_0, pSCV2), meaning that the p value was smaller than 0.0001. Numbers in the brackets represent the 95% confident intervals (CI) of r.

mRNA for transmission of the genetic information into functional proteins. Although the presence of mRNA mutations has been confirmed in a few cellular organisms,[29,32,42,68,69] they affect only a limited number of proteins due to the transient nature of mRNA. However, for RNA viruses whose genomic information is stored in RNA, mutations in RNA can have a long-term influence because such mutations have a chance of being inherited. In this study, we exploit this influence to infer the evolutionary history for RNA viruses, in particular SARS-CoV-2.

It is worth noting that four out of five branches that represented a mixed host history in bats and in a non-human organism (pangolins in B_5 and B_9, civets in B_{11}, camels in B_{20}, and hedgehogs in B_{25}) fell within the 95% confidence ellipse estimated from the 17 bat-exclusive branches (Figure 7B). A

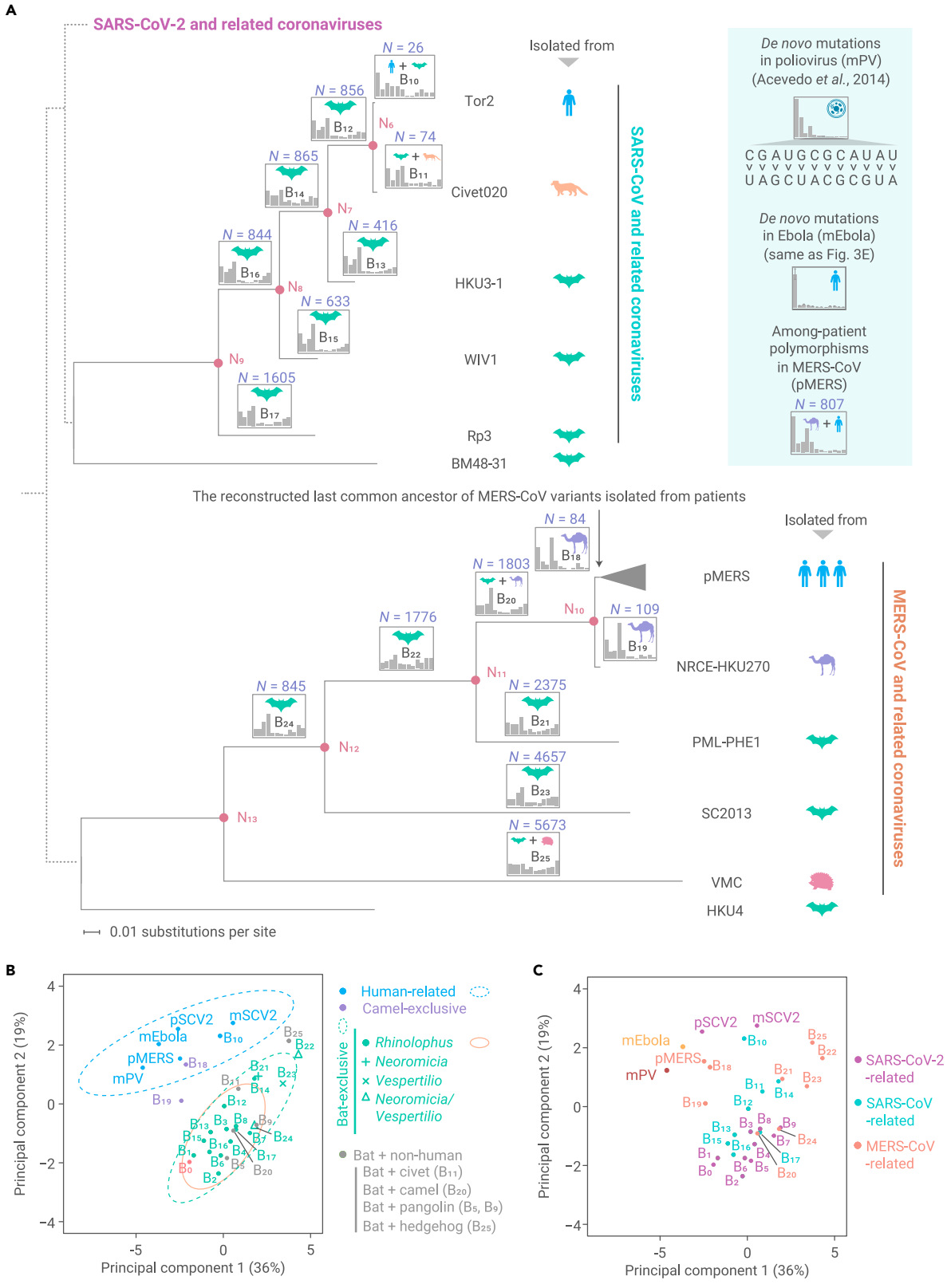


Figure 7. The similarity in mutation spectrum among genetically diverse coronaviruses isolated from various hosts (A) The maximum likelihood phylogenetic trees constructed separately for SARS-CoV-related and MERS-CoV-related viruses, using BM48-31 and HKU4 as outgroups, respectively. The known phylogenetic relationship among SARS-CoV-2-related, SARS-CoV-related, and MERS-CoV-related viruses is depicted by dashed lines, which only reflect the tree topology and give no meaning to branch lengths. (B) The principal-component analysis plot depicts similarity in molecular spectrum. Dots were colored according to the inferred host species. Green, orange, and cyan ellipses represent the 95% confidence intervals for bat, *Rhinolophus* bats, and human cellular environment, respectively. (C) Similar to (B), dots were colored according to the phylogenetic lineage.

plausible explanation is that the majority of the base substitutions present in these four bat-mixed branches were accumulated in bat cellular environments, in light of the unique cellular features reported in bats compared with other mammals, i.e., high concentrations of endogenous antioxidants.[64] In other words, our approach can identify a host jumping event from molecular spectra only if sufficient mutations have accumulated in the new host. In the future, characterization of molecular spectra of RNA mutations in additional species (especially in bats) using CirSeq or Rep-seq will be promising for tracing the transmission route of SARS-CoV-2. Some additional caveats to the conclusions drawn from our results are discussed in the supplemental information.

Although this study focuses on the relative rates among base-substitution types (i.e., molecular spectrum), we estimate an average rate of 1.73×10^{-5} *de novo* mutations per nucleotide in SARS-CoV-2 (Figure 1E). Despite the proofreading mechanism[70] provided by its nonstructural protein 14, SARS-CoV-2 mutations still occur at a rate three to four orders of magnitude higher than DNA mutations,[29] which enables rapid immunological escape, while leaving genomic integrity maintained by natural selection for infectivity.

In addition, our analyses focused on the detection of point mutations. Although not highlighted in the results, we also estimated the molecular spectrum of indels with a similar computational strategy that used junctions as molecular barcodes for the intermediate negative-sense subgenomic RNA (see Figure S2 and supplemental materials and methods). This work identified 2 small insertions and 96 deletions (Figures S3B and S3C), the majority of which were shorter than 6 nucleotides.

Although the significant illness and death caused by the SARS-CoV-2-induced coronavirus disease 2019 pandemic has led to a multitude of studies of this virus, basic understanding is still lacking for several of its key features, such as its origin.[1–7] We show that the mutations accumulated in SARS-CoV-2 prior to its transmission to humans are fully consistent with a natural evolutionary process in a *Rhinolophus* bat host. In addition to this theoretical purport, our methods will also be useful to identify the natural hosts of other RNA viruses and could be potentially applied toward the prevention of future outbreaks.

REFERENCES

1. Rasmussen, A.L. (2021). On the origins of SARS-CoV-2. Nat. Med. ***27***, 9.
2. Andersen, K.G., Rambaut, A., Lipkin, W.I., et al. (2020). The proximal origin of SARS-CoV-2. Nat. Med. ***26***, 450–452.
3. Liu, S.L., Saif, L.J., Weiss, S.R., and Su, L. (2020). No credible evidence supporting claims of the laboratory engineering of SARS-CoV-2. Emerg. Microbes Infect. ***9***, 505–507.
4. Shi, Z.L. (2021). Origins of SARS-CoV-2: focusing on science. Infect. Dis. Immun. ***1***, 3–4.
5. Wu, C.I., Wen, H., Lu, J., et al. (2021). On the origin of SARS-CoV-2—the blind watchmaker argument. Sci. China Life Sci. https://doi.org/10.1007/s11427-021-1972-1.
6. Bloom, J.D., Chan, Y.A., Baric, R.S., et al. (2021). Investigate the origins of COVID-19. Science ***372***, 694.
7. Calisher, C.H., Carroll, D., Colwell, R., et al. (2021). Science, not speculation, is essential to determine how SARS-CoV-2 reached humans. Lancet ***398***, 209–211.
8. 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.
9. Lytras, S., Hughes, J., Martin, D., et al. (2021). Exploring the natural origins of SARS-CoV-2 in the light of recombination. bioRxiv. https://doi.org/10.1101/2021.01.22.427830.
10. Boni, M.F., Lemey, P., Jiang, X., et al. (2020). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. Nat. Microbiol. ***5***, 1408–1417.
11. Kar, S., and Leszczynski, J. (2020). From animal to human: interspecies analysis provides a novel way of ascertaining and fighting COVID-19. Innovation ***1***, 100021.
12. Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. ***24***, 1586–1591.
13. Lynch, M. (2010). Rate, molecular spectrum, and consequences of human mutation. Proc. Natl. Acad. Sci. U S A ***107***, 961–968.
14. Lee, M.B., Dowsett, I.T., Carr, D.T., et al. (2019). Defining the impact of mutation accumulation on replicative lifespan in yeast using cancer-associated mutator phenotypes. Proc. Natl. Acad. Sci. U S A ***116***, 3062–3071.
15. Zhu, Y.O., Siegal, M.L., Hall, D.W., and Petrov, D.A. (2014). Precise estimates of mutation rate and spectrum in yeast. Proc. Natl. Acad. Sci. U S A ***111***, E2310–E2318.
16. De Maio, N., Walker, C.R., Turakhia, Y., et al. (2021). Mutation rates and selection on synonymous mutations in SARS-CoV-2. Genome Biol. Evol. ***13***, evab087.
17. Yang, Z. (1994). Estimating the pattern of nucleotide substitution. J. Mol. Evol. ***39***, 105–111.
18. Siepel, A., and Haussler, D. (2004). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol. Biol. Evol. ***21***, 468–488.
19. Tate, J.G., Bamford, S., Jubb, H.C., et al. (2019). COSMIC: the catalogue of somatic mutations. Cancer Nucleic Acids Res. ***47***, D941–D947.
20. Alexandrov, L.B., Kim, J., Haradhvala, N.J., et al. (2020). The repertoire of mutational signatures in human cancer. Nature ***578***, 94–101.
21. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., et al. (2013). Signatures of mutational processes in human cancer. Nature ***500***, 415–421.
22. Hainaut, P., and Pfeifer, G.P. (2001). Patterns of p53 G–>T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. Carcinogenesis ***22***, 367–374.
23. 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.
24. 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.
25. 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.
26. Panchin, A.Y., and Panchin, Y.V. (2020). Excessive G-U transversions in novel allele variants in SARS-CoV-2 genomes. PeerJ ***8***, e9648.
27. Teng, X., Li, Q., Li, Z., et al. (2020). Compositional variability and mutation spectra of monophyletic SARS-CoV-2 clades. Genomics Proteomics Bioinformatics ***18***, 648–663.
28. Sanjuan, R., Nebot, M.R., Chirico, N., et al. (2010). Viral mutation rates. J. Virol. ***84***, 9733–9748.
29. 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.
30. Acevedo, A., Brodsky, L., and Andino, R. (2014). Mutational and fitness landscapes of an RNA virus revealed through population sequencing. Nature ***505***, 686–690.
31. 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.
32. 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.
33. Wang, C., Horby, P.W., Hayden, F.G., and Gao, G.F. (2020). A novel coronavirus outbreak of global health concern. Lancet ***395***, 470–473.
34. Wu, F., Zhao, S., Yu, B., et al. (2020). A new coronavirus associated with human respiratory disease in China. Nature ***579***, 265–269.
35. Kim, D., Lee, J.Y., Yang, J.S., et al. (2020). The architecture of SARS-CoV-2 transcriptome. Cell ***181***, 914–921.e10.
36. Zhao, Y., Sun, J., Li, Y., et al. (2021). The strand-biased transcription of SARS-CoV-2 and unbalanced inhibition by remdesivir. iScience ***24***, 102857.
37. Sola, I., Almazan, F., Zuniga, S., and Enjuanes, L. (2015). Continuous and discontinuous RNA synthesis in coronaviruses. Annu. Rev. Virol. ***2***, 265–288.
38. Shu, Y., and McCauley, J. (2017). GISAID: global initiative on sharing all influenza data—from vision to reality. Euro Surveill. ***22***, 30494.
39. Wei, C., Chen, Y.M., Chen, Y., and Qian, W. (2021). The missing expression level-evolutionary rate anticorrelation in viruses does not support protein function as a main constraint on sequence evolution. Genome Biol. Evol. ***13***, evab049.
40. Sanjuan, R., and Domingo-Calap, P. (2016). Mechanisms of viral mutation. Cell Mol. Life Sci. ***73***, 4433–4448.
41. Wickner, R.B., Fujimura, T., and Esteban, R. (2013). Viruses and prions of *Saccharomyces cerevisiae*. Adv. Virus Res. ***86***, 1–36.
42. Gout, J.F., Li, W., Fritsch, C., et al. (2017). The landscape of transcription errors in eukaryotic cells. Sci. Adv. ***3***, e1701484.
43. Dai, D.P., Gan, W., Hayakawa, H., et al. (2018). Transcriptional mutagenesis mediated by 8-oxoG induces translational errors in mammalian cells. Proc. Natl. Acad. Sci. U S A ***115***, 4218–4222.
44. Delgado-Roche, L., and Mesta, F. (2020). Oxidative stress as key player in severe acute respiratory syndrome coronavirus (SARS-CoV) infection. Arch. Med. Res. ***51***, 384–387.
45. Kong, Q., and Lin, C.L. (2010). Oxidative damage to RNA: mechanisms, consequences, and diseases. Cell Mol. Life Sci. ***67***, 1817–1829.
46. Li, Z., Wu, J., and Deleo, C.J. (2006). RNA damage and surveillance under oxidative stress. IUBMB Life ***58***, 581–588.
47. Kucab, J.E., Zou, X., Morganella, S., et al. (2019). A compendium of mutational signatures of environmental agents. Cell ***177***, 821–836.e16.
48. Harris, R.S., and Dudley, J.P. (2015). APOBECs and virus restriction. Virology ***479-480***, 131–145.
49. Blanc, V., and Davidson, N.O. (2010). APOBEC-1-mediated RNA editing. Wiley Interdiscip. Rev. Syst. Biol. Med. ***2***, 594–602.

50. Ohno, M., Sakumi, K., Fukumura, R., et al. (2014). 8-Oxoguanine causes spontaneous de novo germline mutations in mice. Sci. Rep. ***4***, 4689.
51. Hofer, T., Badouard, C., Bajak, E., et al. (2005). Hydrogen peroxide causes greater oxidation in cellular RNA than in DNA. Biol. Chem. ***386***, 333–337.
52. Hofer, T., Seo, A.Y., Prudencio, M., and Leeuwenburgh, C. (2006). A method to determine RNA and DNA oxidation simultaneously by HPLC-ECD: greater RNA than DNA oxidation in rat liver after doxorubicin administration. Biol. Chem. ***387***, 103–111.
53. Yan, L.L., and Zaher, H.S. (2019). How do cells cope with RNA damage and its consequences? J. Biol. Chem. ***294***, 15158–15171.
54. Duan, C., Huan, Q., Chen, X., et al. (2018). Reduced intrinsic DNA curvature leads to increased mutation rate. Genome Biol. ***19***, 132.
55. Garcia-Nieto, P.E., Morrison, A.J., and Fraser, H.B. (2019). The somatic mutation landscape of the human body. Genome Biol. ***20***, 298.
56. Haradhvala, N.J., Polak, P., Stojanov, P., et al. (2016). Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. Cell ***164***, 538–549.
57. GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature ***550***, 204–213.
58. 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.
59. Lam, T.T., Jia, N., Zhang, Y.W., et al. (2020). Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature ***583***, 282–285.
60. Xiao, K., Zhai, J., Feng, Y., et al. (2020). Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature ***583***, 286–289.
61. 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.
62. Murakami, S., Kitamura, T., Suzuki, J., et al. (2020). Detection and characterization of bat sarbecovirus phylogenetically related to SARS-CoV-2. Jpn. Emerg Infect. Dis. ***26***, 3025–3029.
63. Hul, V., Delaune, D., Karlsson, E.A., et al. (2021). A novel SARS-CoV-2 related coronavirus in bats from Cambodia. bioRxiv. https://doi.org/10.1101/2021.01.26.428212.
64. Hanadhita, D., Satyaningtijas, A.S., and Agungpriyono, S. (2019). Bats oxidative stress defense. J. Riset Veteriner Indonesia (Journal Indonesian Vet. Research) ***3***. https://doi.org/10.20956/jrvi.v20953i20951.26035.
65. Guan, Y., Zheng, B.J., He, Y.Q., et al. (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science ***302***, 276–278.
66. Alagaili, A.N., Briese, T., Mishra, N., et al. (2014). Middle East respiratory syndrome coronavirus infection in dromedary camels in Saudi Arabia. mBio ***5***, e00884–00814.
67. Cui, J., Li, F., and Shi, Z.L. (2019). Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. ***17***, 181–192.
68. Gordon, A.J., Satory, D., Halliday, J.A., and Herman, C. (2015). Lost in transcription: transient errors in information transfer. Curr. Opin. Microbiol. ***24***, 80–87.
69. Carey, L.B. (2015). RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. eLife ***4***, e09945.
70. V'Kovski, P., Kratzel, A., Steiner, S., et al. (2021). Coronavirus biology and replication: implications for SARS-CoV-2. Nat. Rev. Microbiol. ***19***, 155–170.

ACKNOWLEDGMENTS

We thank Dr. Xionglei He from Sun Yat-sen University, Dr. Lucas Carey from Peking University, and Dr. Taolan Zhao from Institute of Genetics and Developmental Biology CAS for discussion. We acknowledge the authors and laboratories for generating and submitting the sequences to the GISAID Database on which this research is based. The list is detailed in Tables S1 and S2. This work was supported by grants from the National Natural Science Foundation of China (31922014).

AUTHOR CONTRIBUTIONS

W.Q. designed the study. K.-J.S., C.W., Y.W., and Q.H. performed data analyses. K.-J.S., C.W., Q.H., and W.Q. wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

LEAD CONTACT WEBSITE

http://qianlab.genetics.ac.cn/home.html.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xinn.2021.100159.

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

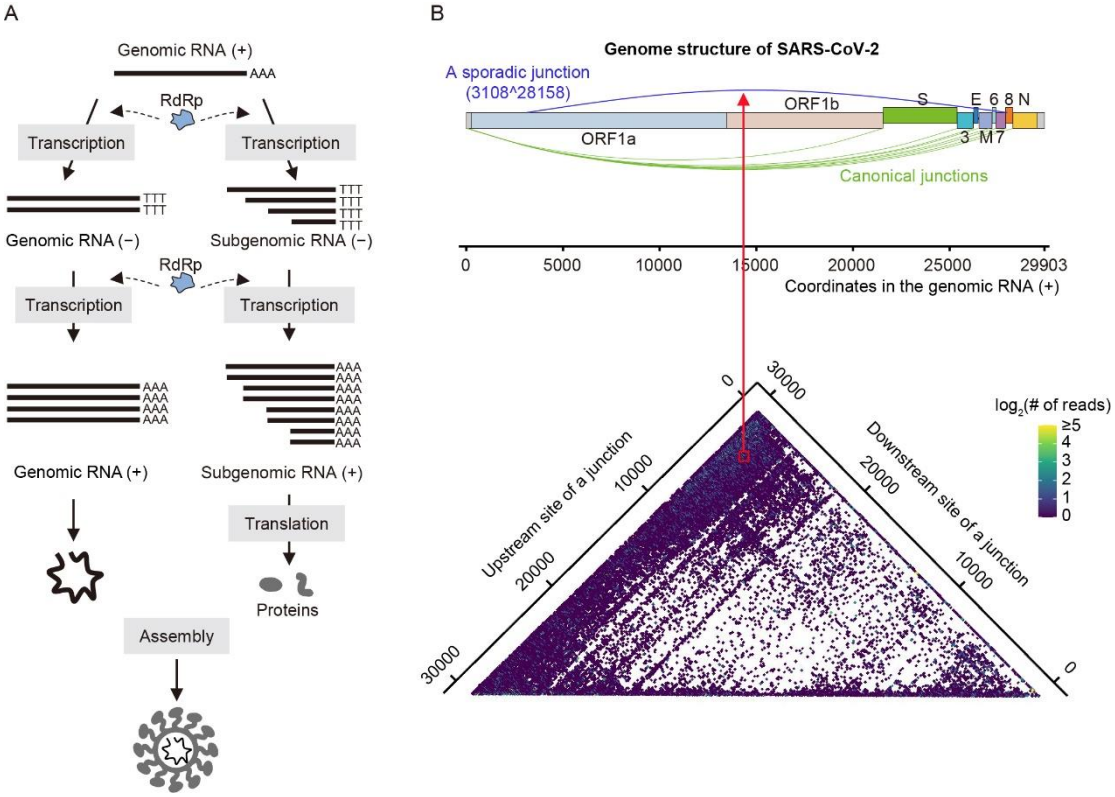


Figure S1. The life cycle of SARS-CoV-2 and its discontinuous transcription.

(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.

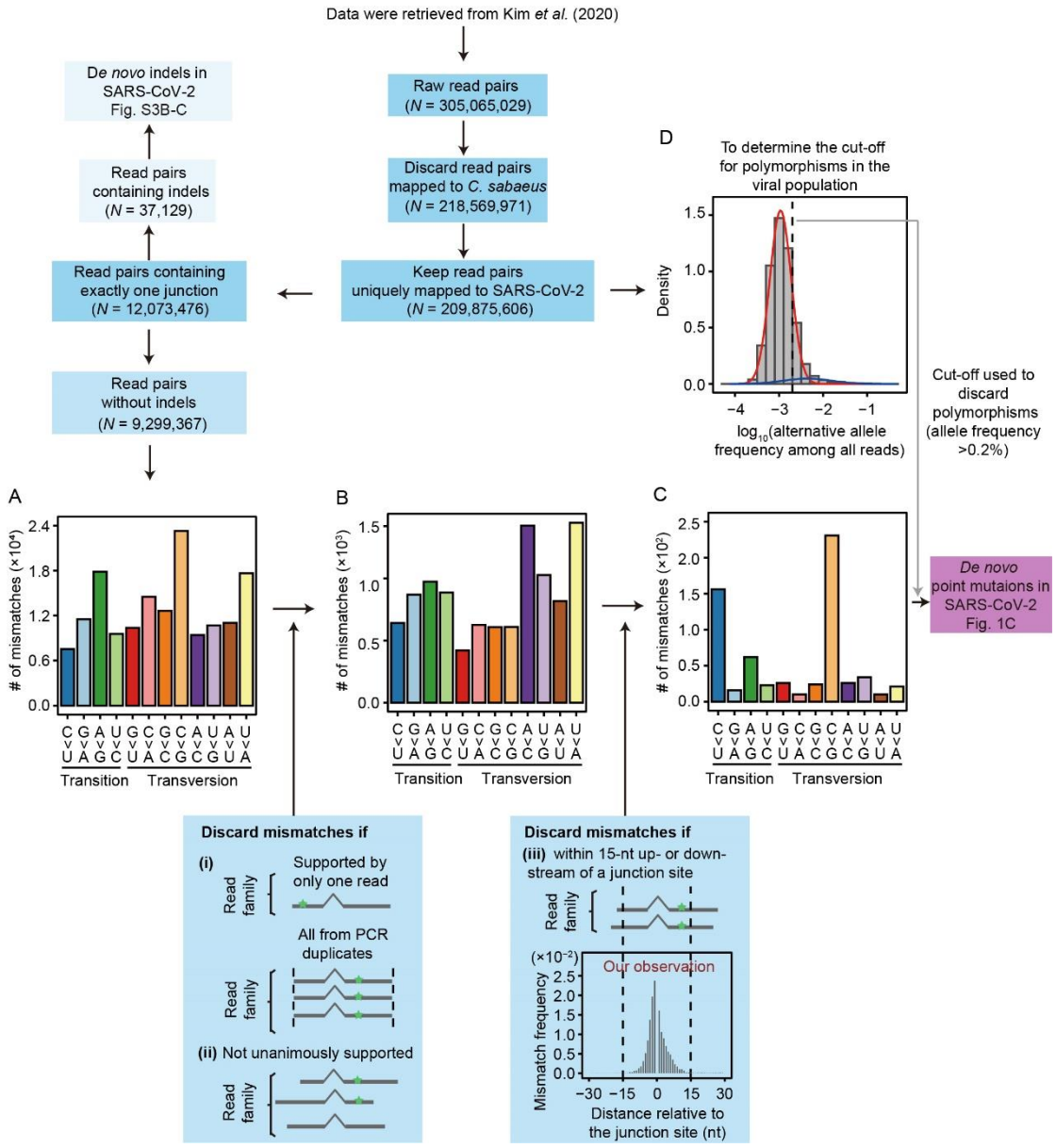


Figure S2. The workflow for the identification of *de novo* RNA mutations in SARS-CoV-2.

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.

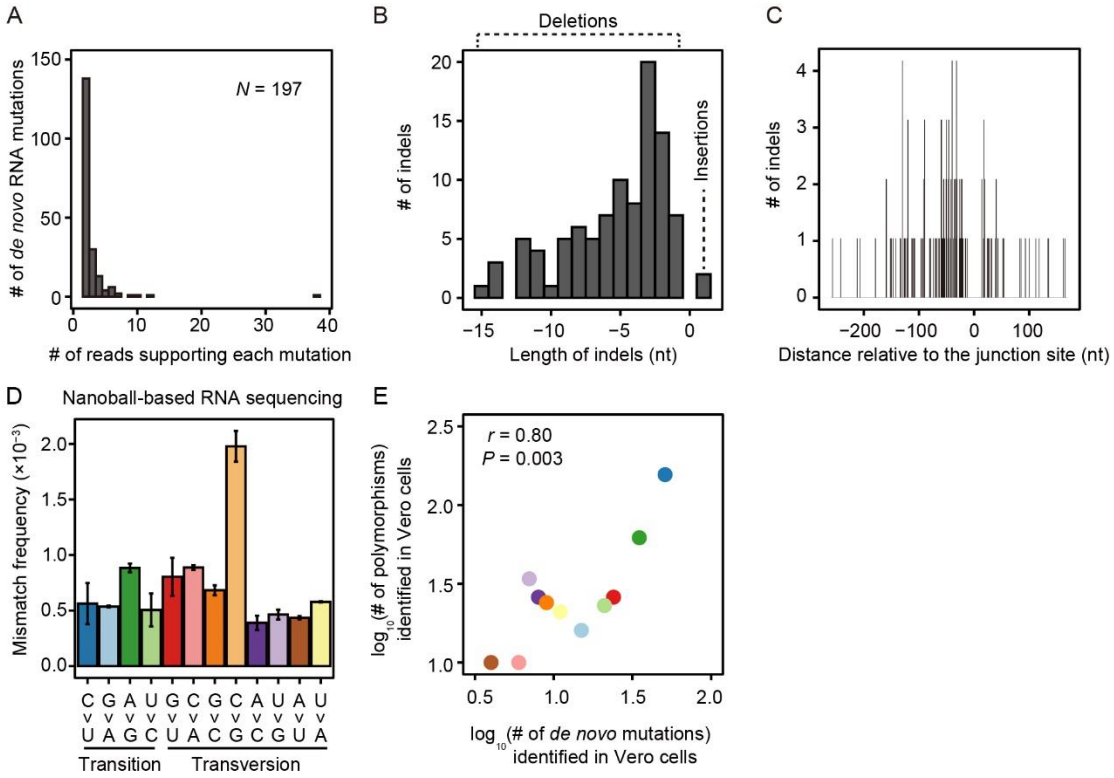


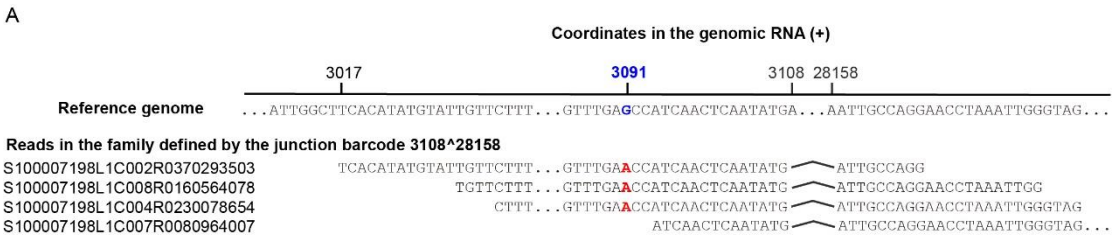
Figure S3. Additional results about *de novo* RNA mutations in SARS-CoV-2.

(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).



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.

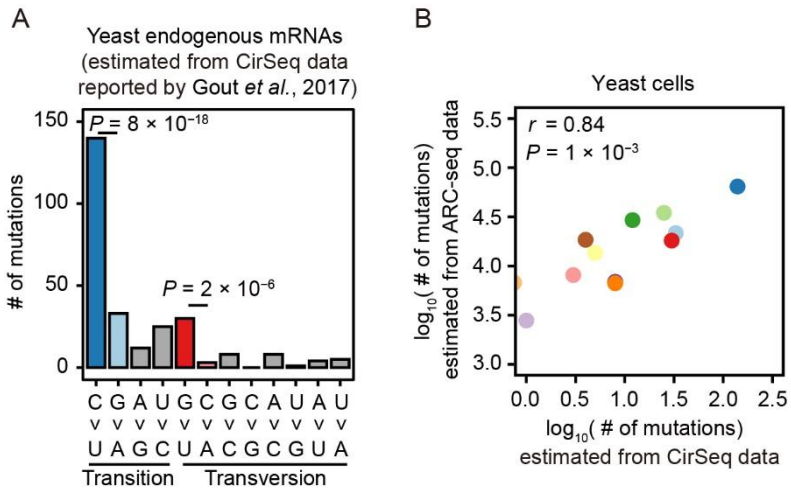


Figure S5. The molecular spectra of mRNA mutations in the budding yeast.

(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.

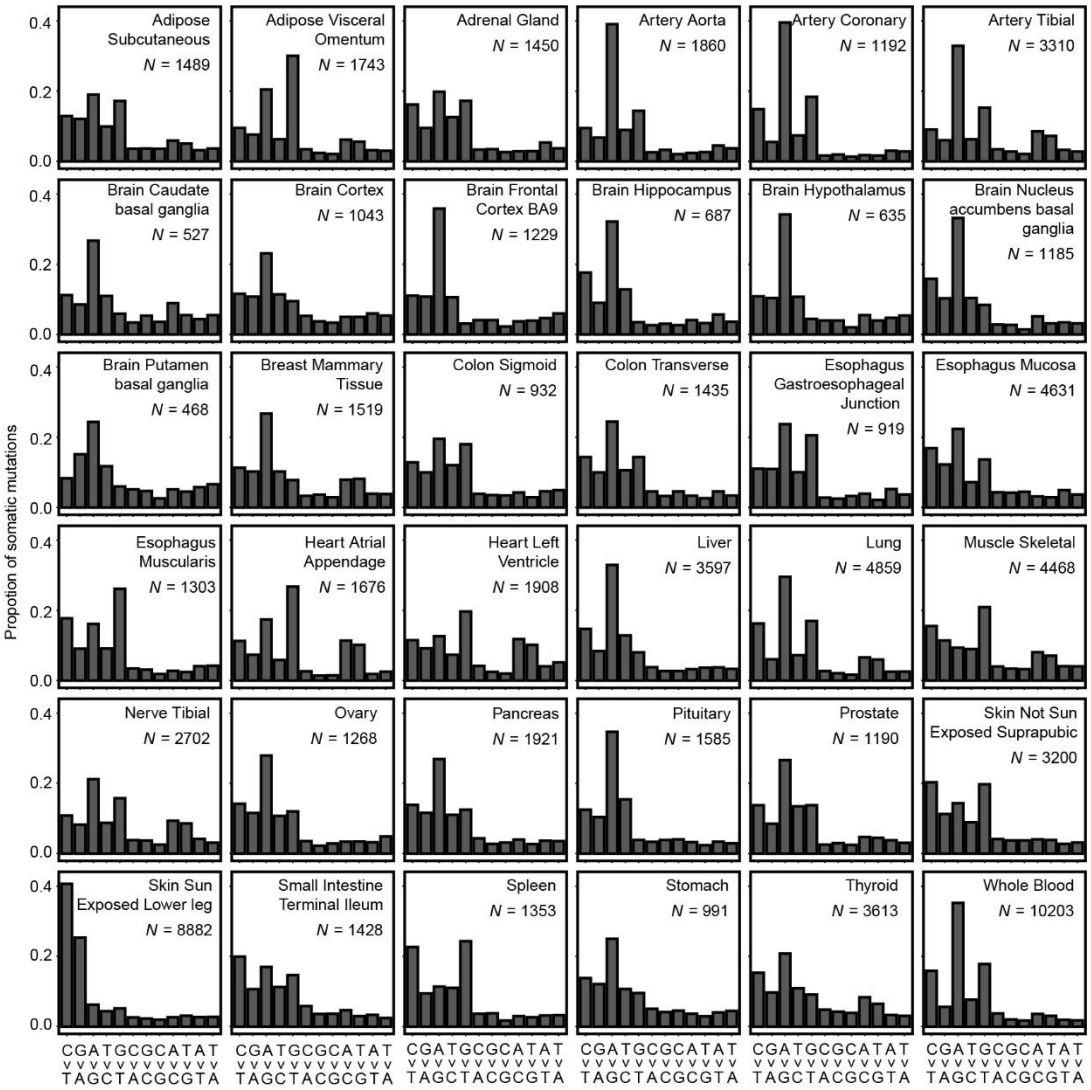


Figure S6. The molecular spectra of somatic mutations in 36 human tissues, polarized according to the coding strand. The number of total somatic mutations detected in a tissue (*N*) is shown in each panel.

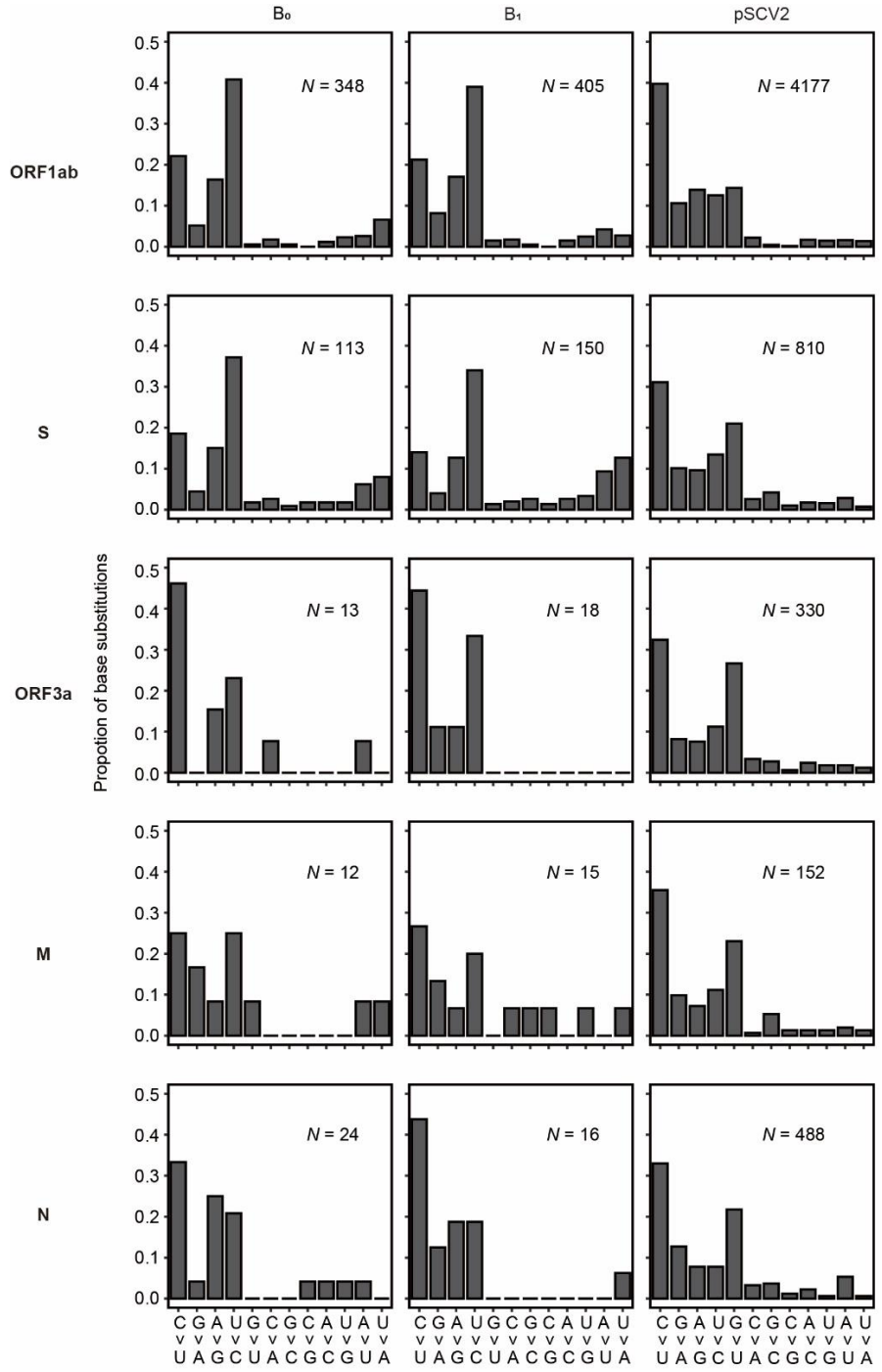


Figure S7. The molecular spectra of mutations accumulated in the branches B_0 and B_1 and among human patients, for five individual SARS-CoV-2 ORFs. The total number of base substitutions detected in a gene (*N*) is shown in each panel. The molecular spectra of the other four ORFs (E, ORF6, ORF7, and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