Journal of Molecular Evolution (2023) 91:214–224
https://doi.org/10.1007/s00239-023-10097-1

ORIGINAL ARTICLE

# Evidence Supporting That C-to-U RNA Editing Is the Major Force That Drives SARS-CoV-2 Evolution

**Jinxiang Wang[1] · Liqun Wu[1] · Xiaoxin Pu[1] · Baoyi Liu[1] · Meiling Cao[1]**

Received: 16 November 2022 / Accepted: 3 February 2023 / Published online: 17 February 2023


## Abstract

Mutations of DNA organisms are introduced by replication errors. However, SARS-CoV-2, as an RNA virus, is additionally subjected to rampant RNA editing by hosts. Both resources contributed to SARS-CoV-2 mutation and evolution, but the relative prevalence of the two origins is unknown. We performed comparative genomic analyses at intra-species (world-wide SARS-CoV-2 strains) and inter-species (SARS-CoV-2 and RaTG13 divergence) levels. We made prior predictions of the proportion of each mutation type (nucleotide substitution) under different scenarios and compared the observed versus the expected. C-to-T alteration, representing C-to-U editing, is far more abundant that all other mutation types. Derived allele frequency (DAF) as well as novel mutation rate of C-to-T are the highest in SARS-CoV-2 population, and C-T substitution dominates the divergence sites between SARS-CoV-2 and RaTG13. This is compelling evidence suggesting that C-to-U RNA editing is the major source of SARS-CoV-2 mutation. While replication errors serve as a baseline of novel mutation rate, the C-to-U editing has elevated the mutation rate for orders of magnitudes and accelerates the evolution of the virus.

**Keywords** SARS-CoV-2 · Mutation · C-to-U RNA editing · Allele frequency · Evolution

## Introduction

Evolutionary biology could be studied at two distinct time scales. Micro-evolution studies the population genetics at intra-species level (Alonso-Blanco et al. 2016; Chu and Wei 2021a; Crow 1955; Park 2011; Wei 2020), while macro-evolution studies the phylogenetic tree at inter-species level (Byng et al. 2016; Jiang et al. 2022; Muller 1995; Wang et al. 2022; Yang 1997). These two processes are essentially governed by the same evolutionary rules where novel mutations randomly occur and introduce polymorphisms to a population, and then mutations were subjected to natural selection (and genetic drift) to either be eliminated or be fixed within a species (Chang et al. 2021; Webster and Smith 2004). The divergence between different species derives from the fixation of mutations within population/species (Yu et al. 2021). In a word, adaptation and evolution would not take place without mutation. Mutation is the basis and prerequisite of natural selection and evolution.

Under this global pandemic caused by SARS-CoV-2, understanding the molecular mechanisms underlying its high mutation rate would help us predict the evolutionary trajectory of the virus and control the pandemic. SARS-CoV-2 mutates so fast that the development of vaccines could not catch up with the alteration of viral sequences. New mutants that escape our control are continuously emerging. Thus, there is urgent need to find out why SARS-CoV-2 bears such a high mutation rate and how to potentially slow down this process. Parsing the mutation spectrum is the first step of investigating the molecular details behind high mutation rate. Traditionally, the combination of micro- and macro-evolution usually offers clear signals of mutation spectrum. Particularly, if one uses the sequence of outgroup species (e.g., RaTG13) to infer ancestral state and correct the reference genome of the target species (e.g., SARS-CoV-2), then the direction of mutations within population could be determined and the derived allele frequency (DAF) could be calculated (Fig. 1A).

However, DAF spectrum is not only affected by mutation rates but also shaped by natural selection. For

Handling editor: **Rosa Fregel**.

✉ Meiling Cao
mlc3@163.com

[1] Department of Respiratory Medicine, Qilu Hospital (Qingdao), Cheeloo College of Medicine, Shandong University, Qingdao, Shandong, China

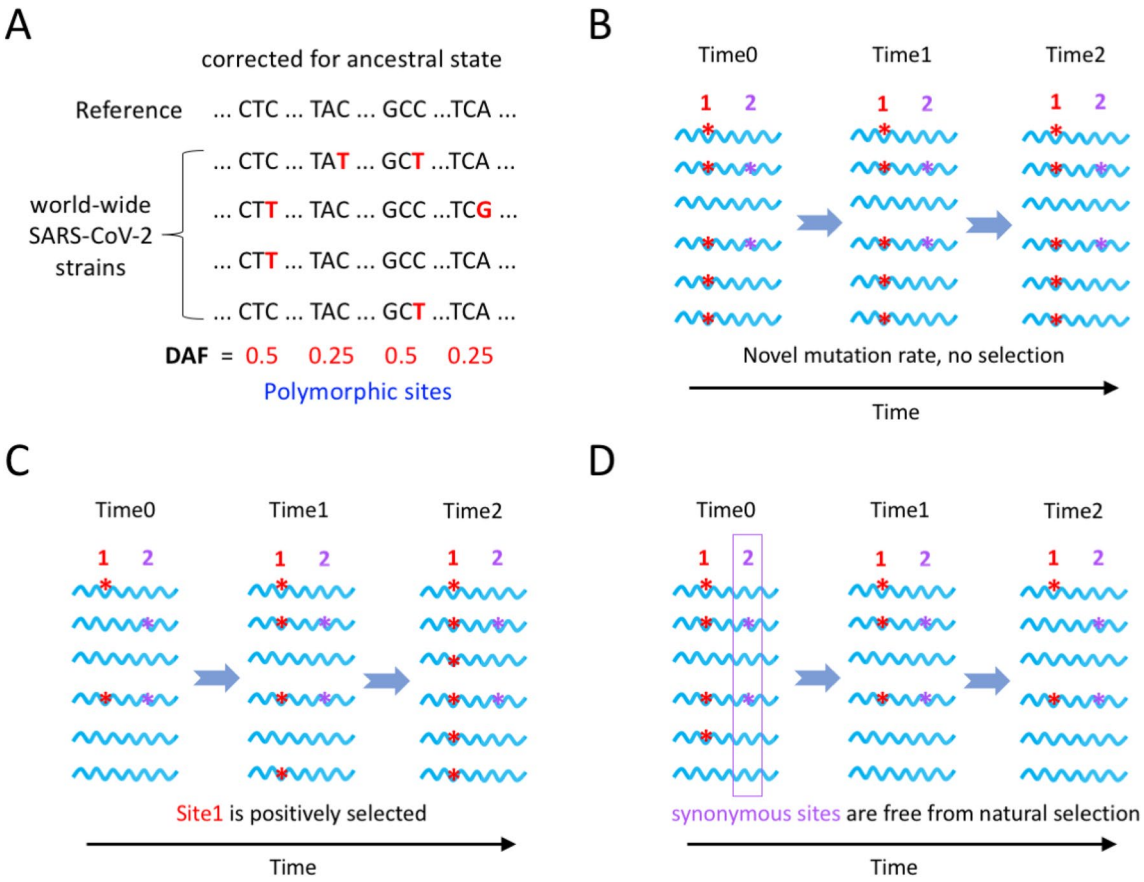


**Fig. 1** Introduction of basic evolutionary concepts. **A** Sequence of outgroup species is used for inference of ancestral state of the target species. Then the direction and derived allele frequency (DAF) of polymorphic sites could be determined. **B** Site1 has higher novel mutation rate than site2, then site1 eventually has higher DAF. **C** Site1 has similar novel mutation rate with site2, but site1 is positively selected and finally has higher DAF than site2. **D** Synonymous mutations are (believed to be) free from natural selection. The final DAF of synonymous sites might mirror the initial mutation rate

example, a mutation site with high DAF could be explained either by (1) this site has high mutation rate (Fig. 1B, site1 has constantly higher DAF than site2), or by (2) this mutation is advantageous and is positively selected (Fig. 1C, site1 eventually has higher DAF than site2). This largely complicates the inference of intrinsic mutation rate. To truly reflect the mutation rate and exclude the impact of natural selection, one may look for neutral sites such as the synonymous mutations (Fig. 1D). Although synonymous mutations are recently believed to affect codon bias and gene expression (Chu and Wei 2021b; Li et al. 2021, 2020a), they are still the best set of neutral sites within the genome, especially when introns are absent. Thus, the final DAF of synonymous sites might faithfully mirror the initial mutation rate (Fig. 1D).

In this study, we will investigate what contributes to the intrinsically high mutation rate in SARS-CoV-2 by looking at the polymorphic and fixed mutation spectrums at neutral (synonymous) sites in the virus genome. As an RNA virus, the novel mutations in SARS-CoV-2 come from both RNA replication errors and RNA editing (editing) by the host cells (Fig. 2A).

Specifically, replication errors are inevitably caused by RNA-dependent RNA polymerase (RDRP) at a low probability, while RNA editing includes ADAR-mediated A-to-I RNA editing and APOBEC-mediated C-to-U RNA editing. Notably, we should clarify that RNA editing would intrinsically occur in endogenous mRNAs in cellular organisms regardless of whether there is virus infection (Liddicoat et al. 2015). However, when RNA viruses invade cells, the editing enzymes would act as a defense system to restrict the virus activity from several aspects (Goila-Gaur and Strebel 2008; Harris 2008; Harris and Dudley 2015; Olson et al. 2018). This means that although the editing enzymes are not produced as a response to the virus infection, they do play a role in fighting against the viruses. This is a well-evolved mechanism that protects the hosts. The editing that occurs in the virus genes is not entirely coincidental.

According to the different source of mutations in SARS-CoV-2, there are three possible outcomes of mutation

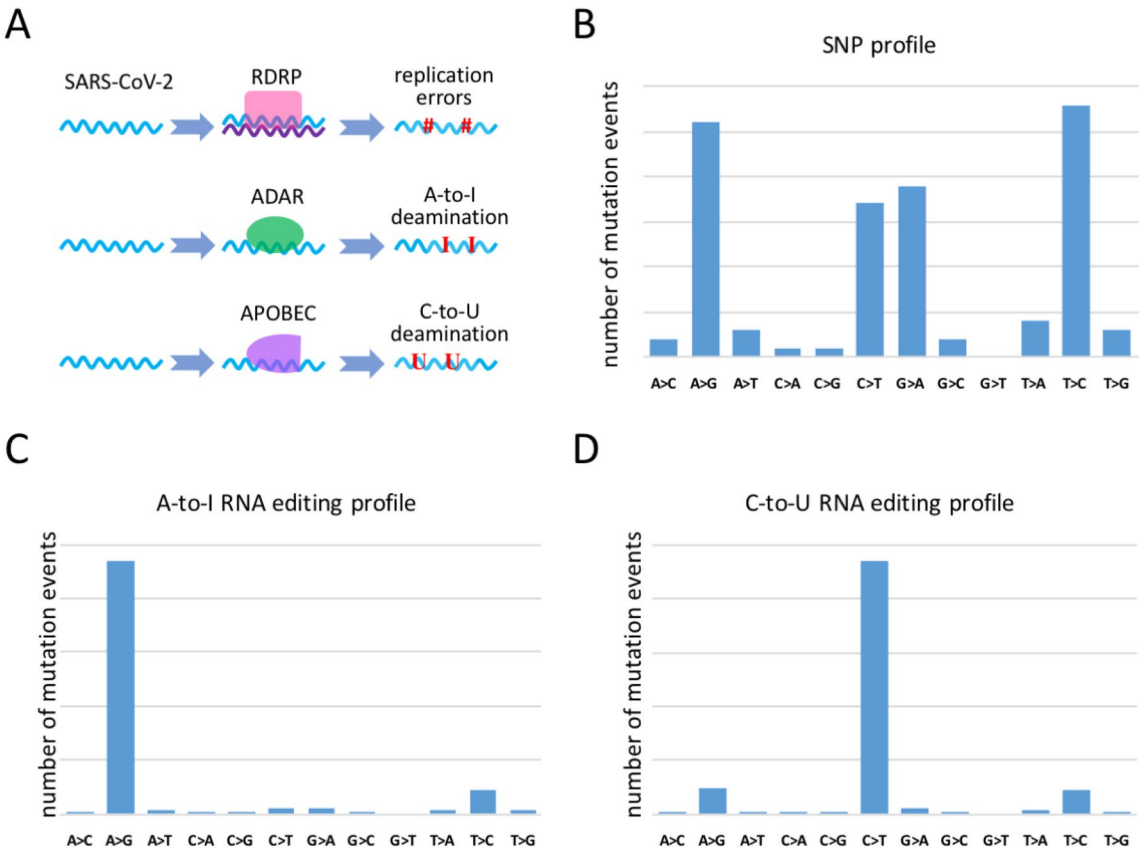


**Fig. 2** Prior prediction of mutation profile based on different mutation sources. **A** The three possible sources of mutations in SARS-CoV-2. RDRP-mediated replication errors, ADAR-mediated A-to-I editing, and APOBEC-mediated C-to-U editing. **B** If replication error is the major mutation source, then the mutation profile should be symmetric. SNP denotes "single nucleotide polymorphism". **C** If A-to-I editing is the major mutation source, then the A-to-G mutations should be dominant. **D** If C-to-U editing is the major mutation source, then the C-to-T mutations should be dominant

profiles when we look at the synonymous sites in SARS-CoV-2. (1) If mutations are mainly contributed by RNA replication errors, then there should be a SNP (single nucleotide polymorphism) profile where the distribution of mutation types is nearly symmetric and that transitions are generally more frequent than transversions (Fig. 2B). (2) If mutations are mainly introduced by A-to-I RNA editing, then the A-to-G mutations should be dominant in the substitution profile just like many previous literatures showed (Li et al. 2014; Liscovitch-Brauer et al. 2017; Porath et al. 2014; Ramaswami et al. 2013) (Fig. 2C). (3) Likewise, if mutations are mainly introduced by C-to-U RNA editing, then the C-to-T mutations should dominate the substitution profile (Chu and Wei 2019; Li et al. 2020b) (Fig. 2D).

Based on these prior predictions, we test the observed mutation profile against the expected ones. We provide strong evidence that suggests that the major mutation source in SARS-CoV-2 is the C-to-U RNA editing rather than A-to-I editing or replication errors. While replication errors serve as a baseline of mutation rate (like in many other organisms), the occurrence of C-to-U RNA editing (editing) is higher than the baseline for orders of magnitudes. Given the fact that the C-to-U editing in SARS-CoV-2 is mediated by host cells, it seems that an effective way to reduce the SARS-CoV-2 mutation rate is to prevent the virus from infecting the hosts, which is, cutting down the transmission of the virus from human to human. Our study gives a nice explanation for the extremely high mutation rate in SARS-CoV-2 at the molecular level, and proposes a potential solution to control the pandemic.

## Methods

### Data Collection

We downloaded the SARS-CoV-2 genome from the NCBI website (https://www.ncbi.nlm.nih.gov/genome/). The RaTG13 sequence was retrieved from previous literature (Li et al. 2020c). The millions of world-wide SARS-CoV-2 sequences were downloaded from GISAID (Shu and McCauley 2017). The outgroup MERS-CoV sequence

was downloaded from NCBI via the following link: https://www.ncbi.nlm.nih.gov/genome/?term=mers-cov. The sequences of DNA viruses were downloaded from NCBI with the following accession numbers: *Mythimna separata* (NC_021246), *Adoxophyes honmai* (NC_021247), *Amsacta moorei* (NC_002520), *Choristoneura biennis* (NC_021248), *Choristoneura rosaceana* (NC_021249).

### Polymorphic Sites and Derived Allele Frequency (DAF)

The millions of world-wide SARS-CoV-2 sequences and their allele frequency (AF) information provided by a previous study (Zhu et al. 2022) were done for the following pipelines. The polymorphic mutations were called against the corrected reference SARS-CoV-2 genome. The reference SARS-CoV-2 sequence was corrected for ancestral state by considering the outgroup RaTG13 sequence. The non-reference nucleotides in the world-wide SARS-CoV-2 sequences were derived alleles. DAF was the number of derived allele counts divided by the total number of sequences. The full list of polymorphic data were provided as Table S1.

### Sequence Alignment and Divergence Analysis

The sequences of SARS-CoV-2 and RaTG13 were aligned with software MUSCLE (Edgar 2004). Since the average similarity between SARS-CoV-2 and RaTG13 is 96%, the alignment was highly robust under different parameters or software. The divergence sites between the two viruses were manually extracted. The ancestral state of each site was inferred from the sequence of other closely related coronaviruses (Wang et al. 2021). The divergent sites between SARS-CoV-2 and RaTG13 were listed in Table S2.

## Results

### Prior Prediction of Mutation Profile Under Different Scenarios

We collected ~10 million world-wide SARS-CoV-2 sequences and the mutation information from GISAID and a previous study (Shu and McCauley 2017; Zhu et al. 2022). As we have stated, there are three mutation sources for SARS-CoV-2 (Fig. 2A). Accordingly, we have prior predictions of the mutation profiles when each mutation source is prevalent. If RDRP-mediated replication errors are the major mutation source, then the mutation profile should be symmetric and show higher substitution rate for transitions than transversions (Fig. 2B). If ADAR-mediated A-to-I editing is the major mutation source, then there should be a peak at the A-to-G substitution (Li et al. 2014; Liscovitch-Brauer et al. 2017; Porath et al. 2014; Ramaswami et al. 2013) (Fig. 2C). If mutations are mainly introduced by APOBEC-mediated C-to-U editing, then C-to-T should be the dominant substitution type (Chu and Wei 2019; Li et al. 2020b) (Fig. 2D). We will compare the observed mutation profile against the expected ones.

### C-to-U Editing Is the Dominant Mutation Type and Has High DAF

To date, almost every nucleotide in the SARS-CoV-2 genome has been hit by mutations. Therefore, displaying the unfiltered mutation sites would be meaningless as it simply represents the nucleotide composition of the SARS-CoV-2 genome. We need to show the mutation profiles under different thresholds. After adjusting the ancestral state of the reference genome, we obtained the derived allele frequency (DAF) of each mutation site in the world-wide SARS-CoV-2 strains (Fig. 1A). We required a mutation site to have DAF > 1E-5. When DAF is between (1E-5, 1E-4), the numbers of SNV (single nucleotide variation) exhibit the classic "SNP profile" where the distribution is symmetric and transitions take place more frequently than transversions (Fig. 3A). When DAF increases, the fraction of C-to-T substitution has remarkably elevated (Fig. 3B). When DAF is higher than 1E-3, C-to-T become the dominant substitution type (Fig. 3C, D). According to our prior prediction of the mutation profile (Fig. 2), our observation suggests that C-to-U editing is prevalent and has generally higher allele frequency than other mutation types.

Notably, although the fraction of C-to-T substitutions increases with the DAF threshold, the exact number of mutation sites has decreased. This pattern is expected because both C-to-T (Fig. 3E) and non-C-to-T mutations (Fig. 3F) follow the classic DAF spectrum where the number of mutations rapidly decrease with DAF. Nevertheless, C-to-T mutations still show significantly higher DAF than non-C-to-T mutations (Fig. 3G).

### Fixation of C-to-U Mutations During the Evolution of SARS-CoV-2

Fixation is the process that the polymorphic mutations (AF < 1) become fixed in the population (AF = 1) by positive selection, genetic drift, or hitchhiking (or other unusual incidents). Studying the fixation of mutations would help us better understand the speciation mechanisms: how the sequences of two species were diverged step by step from a common ancestor. Technically, time-course population data enable us to directly observe the fixation process by looking at the dynamics of AF.

To define newly fixed mutations in SARS-CoV-2 population, we focused on the last 3 time-points of the

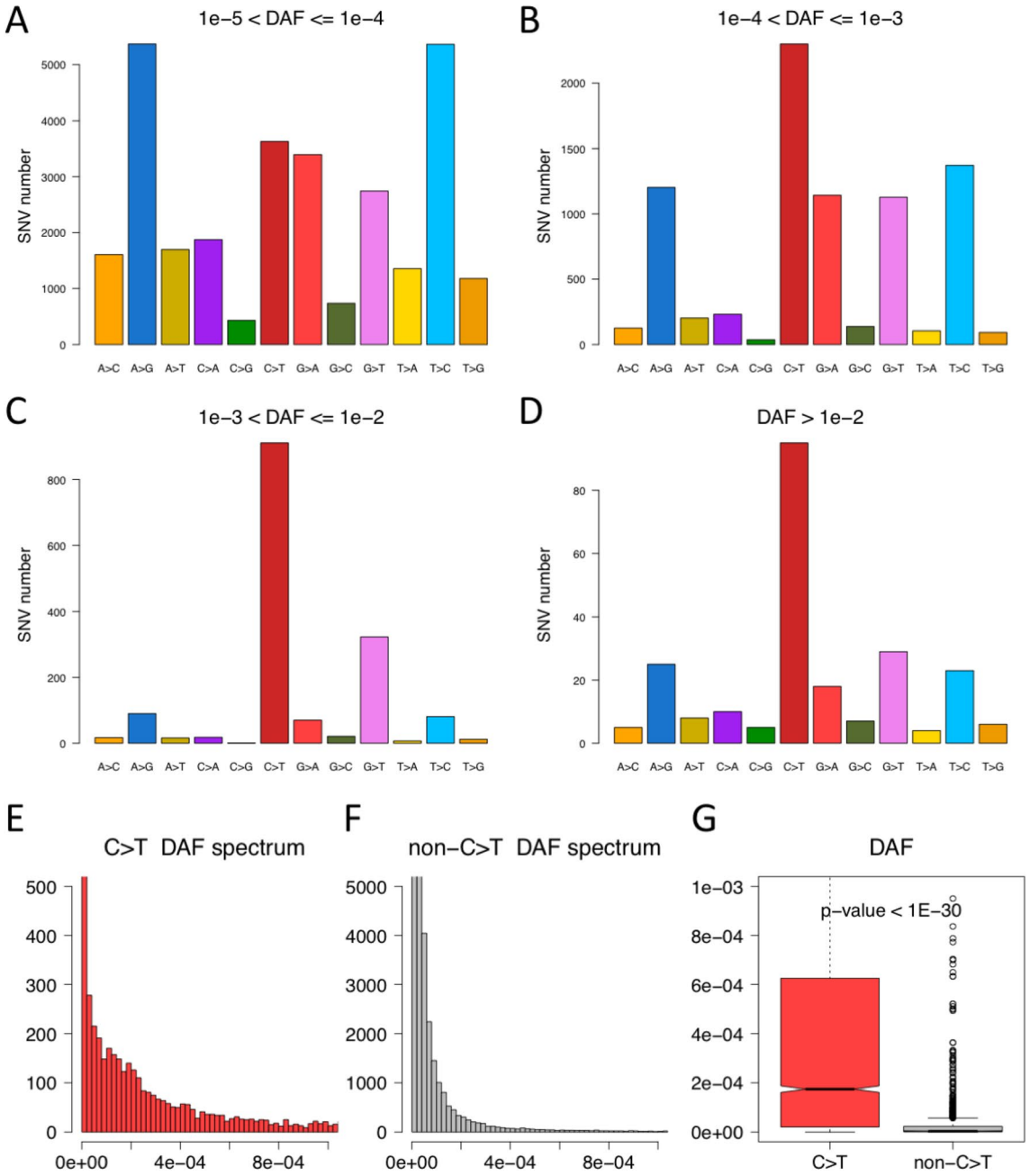


**Fig. 3** The numbers of SNV (single nucleotide variation) in the SARS-CoV-2 genome under different thresholds of DAF (derived allele frequency). **A** When DAF is low, the mutation profile resembles the SNP profile. **B** The fraction of C-to-T substitution increases with DAF. (C-D) When DAF is high, C-to-U editing sites become dominant. **E** DAF spectrum of C-to-T substitutions. **F** DAF spectrum of non-C-to-T mutations. **G** Comparison of C-to-U and non-C-to-U mutations. The difference of DAF was determined by KS test

time-course data provided by the original paper (Zhu et al. 2022). A mutation in the population (against the reference sequence: the WT 2019 strain NC_045512) was either polymorphic (AF < 1) or fixed (AF = 1). We require a fixed mutation to keep AF = 1 for the last 3 time-points: 15 January, 31 January, and 15 February of 2022. Although this criterion is arbitrary, we believe that it meets the definition of newly fixed mutation: (1) this mutation was absent in the ancestor (the reference sequence); (2) now this mutation has maintained AF = 1 for a long-enough time.

Under this criterion, we obtained 726 mutation sites. 403 (55.5%) of the sites were C-to-T(U) mutations and 147 (20.2%) were A-to-G mutations. This proportion was significantly different from a random mutation profile (p-value < 1e-10, Chi-square test). Again, this suggests

that the rampant C-to-U RNA editing promotes the fixation of such mutation.

## C-to-U Editing Has Intrinsically Higher Mutation Rate Than Other Mutation Types

Given the fact that C-to-T mutations, representing the C-to-U RNA editing events, have higher DAF and are more likely to become fixed in the population than all other types of mutations (Fig. 4A), it is still unclear whether the high DAF is caused by intrinsically high mutation rate (Fig. 1B) or by positive selection (Fig. 1C).

In theory, positive selection favors the adaptive mutations regardless of whether the mutation type is C-to-T or non-C-to-T. Any types of mutations might have a chance to be beneficial and increase the fitness of a species. This privilege should not be restricted to C-to-T mutations only. Under this logic, it is highly likely that C-to-T mutations have higher intrinsic mutation rate. The aim of this study is to find out whether the rampant C-to-U RNA editing accounts for the high mutation rate in SARS-CoV-2.

While missense mutations that change protein-sequences are subjected to strong positive and purifying selection, the synonymous mutations are regarded as neutral sites which are free from natural selection. Therefore, synonymous sites could be used as molecular clock to measure the occurrence rate of novel mutations. Here, we display the median and mean DAF of all mutation types with missense and synonymous sites separated (Fig. 4B). There are two clear trends: (1) Synonymous sites have higher DAF than missense sites, and (2) DAF of C-to-T mutation is still the highest among all mutations.

Firstly, it is understandable that missense mutations have lower DAF than synonymous sites as most novel missense mutations are deleterious and rapidly suppressed within the population. Furthermore, these results also demonstrate that even when the effect of natural selection is excluded (by looking at synonymous sites only), C-to-T mutation is still the dominant substitution type in SARS-CoV-2 RNA. These observations are strong evidence suggesting that C-to-U RNA editing indeed contributes a lot to the high mutation rate in the virus. There are no other alternative explanations

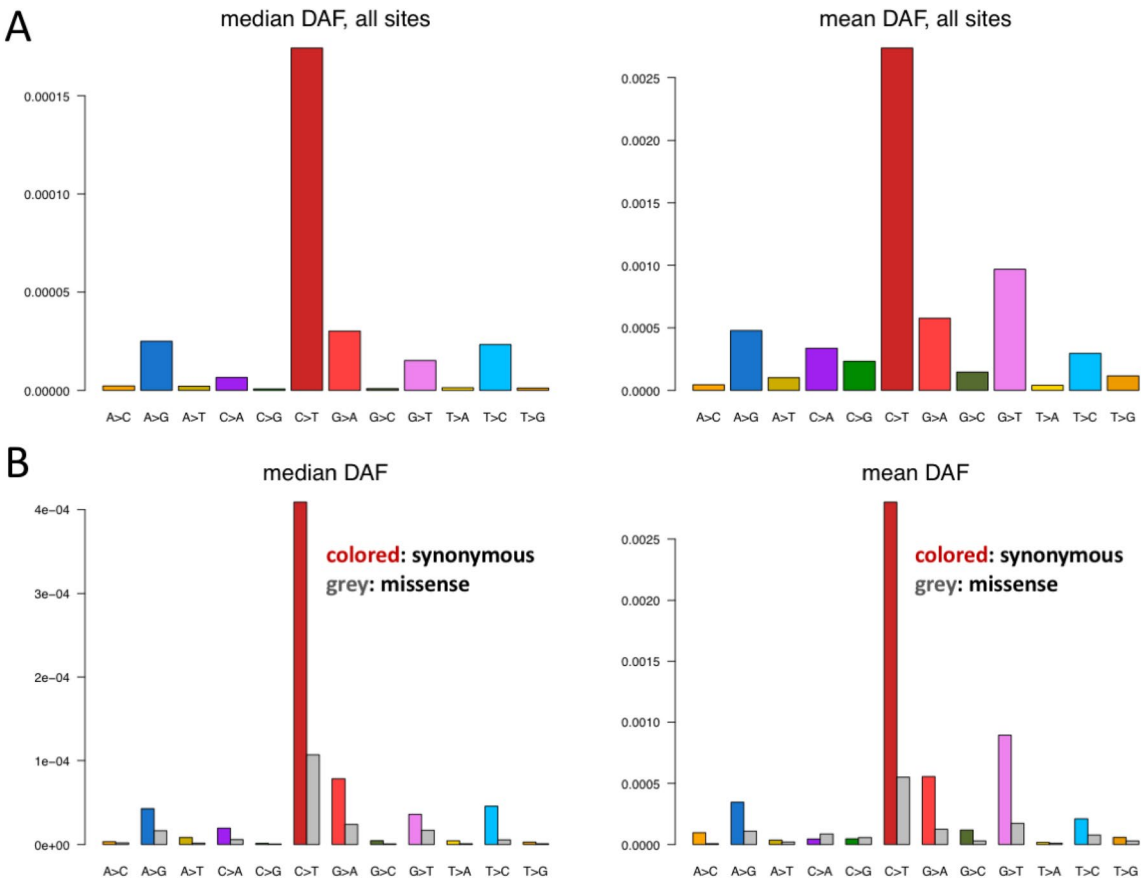


**Fig. 4** The DAF of each mutation type. **A** The median and mean DAF of each mutation type. **B** Synonymous and missense sites are shown separately. Synonymous sites have obviously higher DAF than missense sites

that could account for the extraordinarily prevalent C-to-T substitutions.

## C-to-U Substitution Dominates the Divergence Sites Between SARS-CoV-2 and RaTG13

Polymorphic sites within the population of a species depict the ongoing evolution process (at a smaller time scale), while the final result of selection and evolution is reflected by the divergence (fixed) sites between two species when the "speciation" process is done (Fig. 5A). Polymorphism analysis represents population genetics at micro-evolution level while divergence analysis represents inter-species comparison at macro-evolution level. The divergence sites serve as the genetic basis that accounts for the many different properties and behaviors between two species (Zhang et al. 2021a). Under the assumption that C-to-U RNA editing is prevalent and contributes most to the novel mutations in SARS-CoV-2, we have the following predictions beforehand.

When comparing the sequences of two RNA viruses (e.g., SARS-CoV-2 and RaTG13), both sides might undergo C-to-U editing during the speciation process (Fig. 5B). If no outgroup (or ancestral) information is available, then one could not determine the direction of mutation and could only observe a plenty of non-directional CT substitutions (Fig. 5B). In contrast, if ancestral state is determined by incorporating the outgroup information (like the MERS-CoV sequence), then the direction of mutations could be defined and the mutation profile should show a peak at C-to-T (Fig. 5B).

In the real data, when we looked at the divergence sites between SARS-CoV-2 and RaTG13 (Fig. 6), we indeed observed that: (1) Most substitutions are CT when the direction of mutations is not determined, suggesting that C-to-U editing events have taken place in both lineages during speciation; (2) When the direction of mutation is determined by outgroup MERS-CoV sequences, most substitutions are C-to-T, again representing the prevalent C-to-U RNA editing; (3) Synonymous sites have considerably greater divergence than missense sites (Fig. 6), supporting the classic evolution theory that missense mutations are largely depleted during evolution while synonymous mutations are nearly neutral and have higher chance to be fixed.

In this part, we demonstrated that the inter-species fixed sites between SARS-CoV-2 and RaTG13 are mainly contributed by C-to-U editing. Together with our analyses

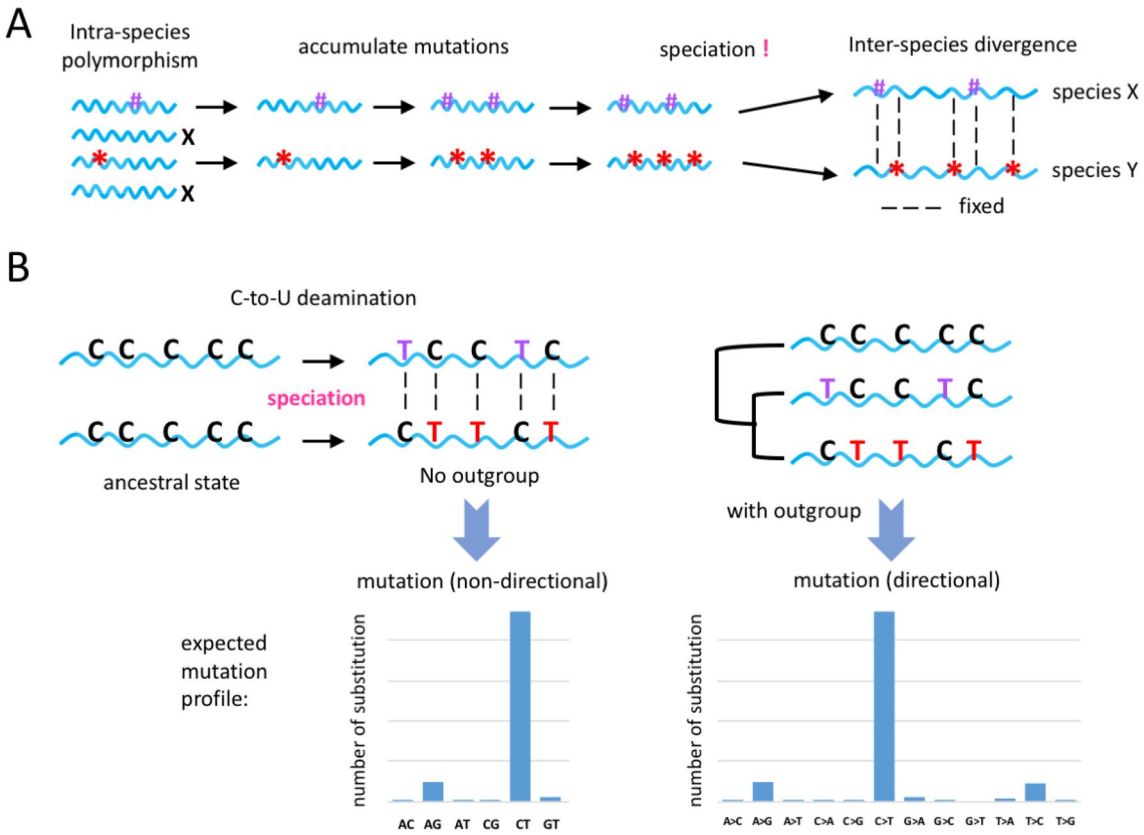


**Fig. 5** The definition of inter-species divergence. **A** How mutations within a population lead to speciation and the sequence divergence. **B** Calculating the numbers of each substitution type when outgroup (MERS-CoV) is absent or present

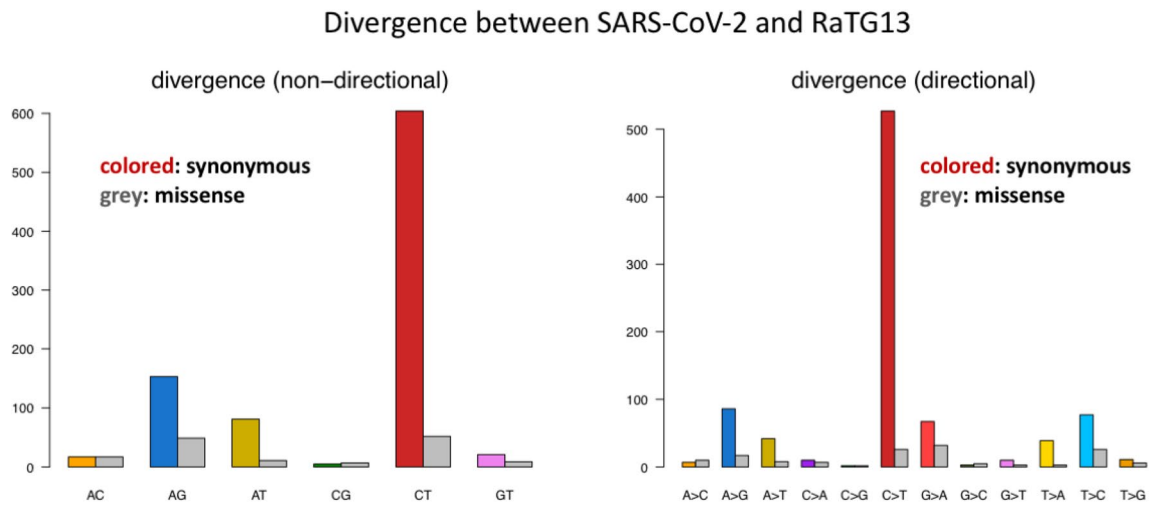


**Fig. 6** The divergence sites between SARS-CoV-2 and RaTG13. Synonymous and missense sites were displayed separately. Non-directional mutations were counted without outgroup information. Directional mutations were inferred based on other coronaviruses as outgroup

on polymorphic sites within world-wide SARS-CoV-2 strains, our results serve as compelling evidence suggesting that C-to-U editing is the major source of mutations in SARS-CoV-2.

### Divergence Between DNA Viruses Does Not Exhibit a Peak for Particular Mutation Types

To test whether the enrichment of C-to-T mutations is unique to RNA viruses, we tested this pattern in DNA viruses. We downloaded the genome sequences of 5 entomopoxviruses (abbreviated as EPV, also known as insect pox viruses) which are typical DNA viruses.

The 5 EPV species and their accession numbers are: *Mythimna separata* (NC_021246), *Adoxophyes honmai* (NC_021247), *Amsacta moorei* (NC_002520), *Choristoneura biennis* (NC_021248), *Choristoneura rosaceana* (NC_021249). The phylogenetic tree was based on the GTR mode of IQtree (Fig. 7A).

Using *A. moo* as an outgroup, we quantified the interspecies mutations from *C. bie* to *C. ros* (Fig. 7B). Clearly, the mutation profile is symmetric, resembling the commonly known SNP profile created by (DNA) replication errors. Similarly, we extracted the mutations from *A. moo* to *C. bie* (Fig. 7C) and from *A. hon* to *A. moo* (Fig. 7D). None of the mutation profiles showed an enrichment of a particular variation type. All profiles were symmetric and the numbers of divergent sites increased with phylogenetic distance. These results serve as negative control to prove that the unusually high C-to-T(U) peak in SARS-CoV-2 is likely caused by C-to-U RNA editing.

## Discussion

There are essential differences between DNA organisms and RNA viruses like SARS-CoV-2. For DNA organisms, their mutations come from DNA replication errors introduced by DNA polymerase. RNA editing in cellular organisms does not affect their genetic materials at all. However, for RNA viruses, the RDRP-mediated RNA replication errors only consist of a minor part of viral mutations. When viral RNAs are deaminated by ADARs or APOBECs, these altered (modified) RNAs are directed used as templates to produce the "offspring" (daughter strand). As a consequence, the nucleotide alterations caused by RNA editing would accumulate linearly at each generation (but with a rate much higher than replication error rate). Given a long-enough time, the editing sites would be overwhelming. This prediction has been nicely verified by the evidence that both polymorphic and fixed mutation sites are dominated by C-to-T substitutions which represent C-to-U editing.

We wholeheartedly agree that finding an APOBEC-deficiency condition is the best way to prove the origin of C-to-U mutations in SARS-CoV-2. For example, it is instructive to study viruses whose hosts do not exhibit such an extensive APOBEC activity as do humans. However, APOBEC is highly conserved in animals so that it is difficult to find such an ideal host without APOBEC. Moreover, we may need some time to collect the "pairing information" between host and viruses. We are willing to do this in our future works. We will plan to take into account the species-specific and even tissue-specific gene

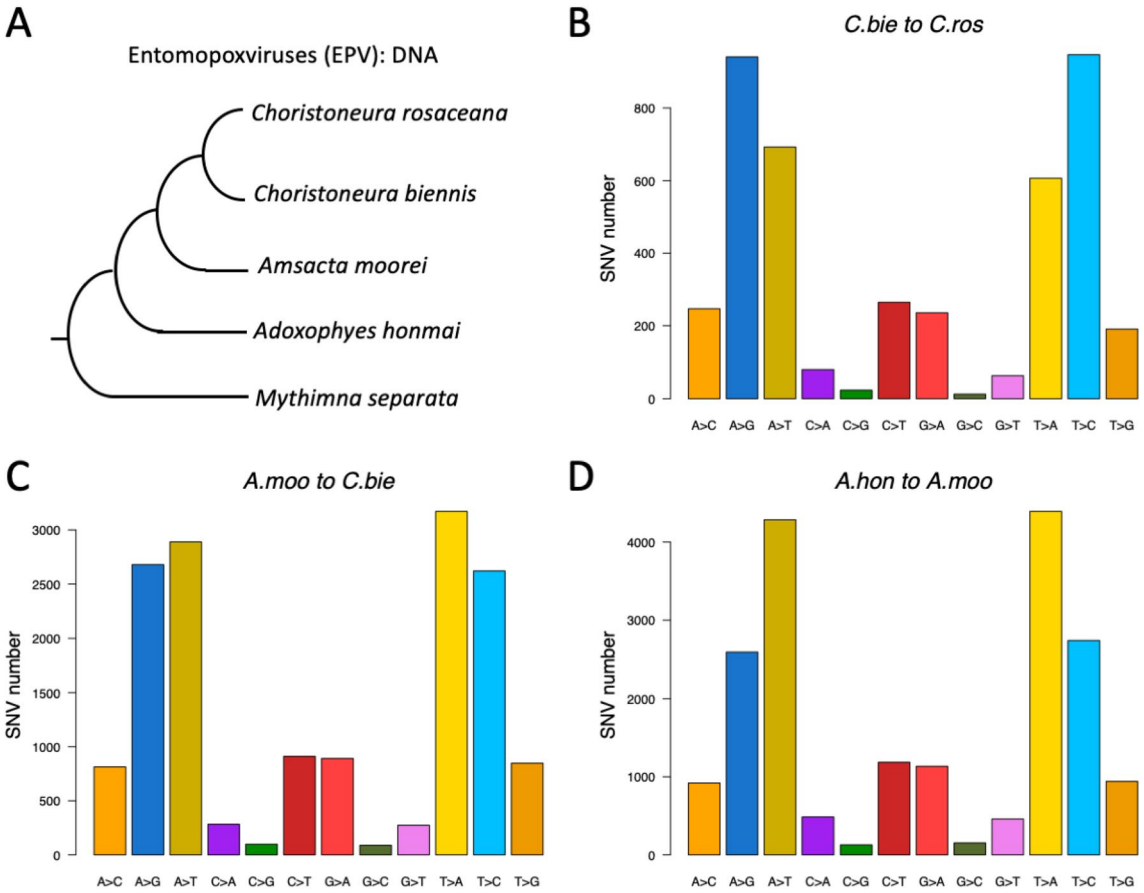


**Fig. 7** The divergence between five entomopoxviruses (EPV). **A** Phylogeny of the five viral sequences. **B, C, D** The directional inter-species divergence between two viral sequences

expression of APOBEC. Some database like GTEx may have such expression data. Higher expression of APOBEC gene is considered to be connected with more prevalent C-to-U editing in a host.

In addition, another concern is how sure are we of the existence of RNA editing events when only a mutation profile is given. Indeed, each virus has its own RDRP, so the patterns of polymerase-induced errors might be different. However, note that the replication errors (SNPs) take place during the DNA or RNA replication process. RDRP should have equal chances to edit the positive strand and the negative strand. No matter the RDRP is prone to which type of mutation, the final mutations called against the reference genome should be symmetric. For example, if an RDRP is prone to C > T mutations during synthesis, then the C > T mutations in negative strand would finally be demonstrated as G > A mutations in the positive strand, leading to a symmetric SNP profile (Cai et al. 2022; Di Giorgio et al. 2020). However, replication only represents a transient (short) time frame during the lifecycle of a virus (Zong et al. 2022). For most time of the lifecycle of SARS-CoV-2, the status should be single-stranded RNA, which could be reached by the editing enzymes, leading to an excess of a particular mutation type. A mutation profile containing abundant RNA editing should be asymmetric. The certainty of RNA editing increases with the enrichment of the desired type of variation. For example, in this field, if over 90% of the variation is A > G, then it would be strong evidence for A-to-I RNA editing (Li et al. 2014; Liscovitch-Brauer et al. 2017; Porath et al. 2014; Ramaswami et al. 2013). No alternative mechanisms could explain the mutation profile.

Notably, RDRP is encoded by virus itself (for replication purpose) while ADARs and APOBECs are encoded by the host genome. Ironically, ADARs and APOBECs are (thought to be) designed for anti-virus purposes (Harris and Dudley 2015; Ward et al. 2011). However, they actually fuel and facilitate the evolution of the virus. This is why researchers believe that SARS-CoV-2 evolution is largely driven by the hosts (Zhang et al. 2021b). Furthermore, there comes a dilemma that should we apply the classic evolutionary theories to the RNA viruses (Li et al. 2020d)? Traditional evolutionary theories and formula are based on a fact that mutations come from DNA replication errors under a relatively constant rate. This basis is invalid for RNA viruses

at all. RNA editing is rampant and largely promiscuous, it is hard to find a uniform formula to describe the "mutation rates" of RNA editing events, let alone describe their evolutionary dynamics. The phenomenon of A-to-I and C-to-U editing has been found for over 30 years in metazoans (Bass and Weintraub 1987, 1988; Gray and Covello 1993; Kudla et al. 1992), not a single literature has invented a formula to describe the "mutation rate" or "occurrence rate" of RNA editing, suggesting that there are fundamental discrepancies between DNA mutation and RNA editing. Nevertheless, one certain thing is that C-to-U editing is the main driving force that accelerates SARS-CoV-2 mutation and evolution. Another implication is that a proper way to slow down SARS-CoV-2 mutation is to cut down the virus transmission. The viral RNAs could not autonomously deaminate without the hosts.

One unexplained puzzle is why C-to-U editing is much more prevalent than A-to-I editing? The answer might be the subcellular localization of ADARs and APOBECs (Martignano et al. 2022; Zong et al. 2022). As previous literatures claimed, APOBECs are the "editors in chief" of cytoplasmic viral editing. Our findings further support this statement. Under this scenario, ADAR-mediated A-to-I events are much less abundant than C-to-U events. Conceivably, accurate identification of A-to-I editing sites in SARS-CoV-2 is challenging because one needs to design a pipeline to exclude all other types of mutations in order to make A-to-G the dominant peak. If one automatically regards all A-to-G mutations as A-to-I events without observing a peak at A-to-G (Di Giorgio et al. 2020), then others might suspect those variations as false positive hits (Martignano et al. 2022; Picardi et al. 2021; Song et al. 2022; Wei 2022; Zong et al. 2022).

Last but not least, in our study, the use of synonymous sites in the virus genome is reasonable since missense sites are subjected to strong purifying selection. Even for the mutations in non-coding regions like 5'UTRs or regulatory sequences, they might undergo selection pressure due to their impact on RNA expression and translation rates (Wang et al. 2021; Zhang et al. 2022). In summary, our study provides compelling evidence suggesting that C-to-U editing is the major source of SARS-CoV-2 mutation. While replication errors serve as a baseline of novel mutation rate, the C-to-U editing mechanism has elevated the mutation rate for orders of magnitudes and fuels the evolution of the virus.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00239-023-10097-1.

**Acknowledgements** We thank the members in our group that have given suggestions to our project. At this SARS-CoV-2 time we should especially thank all the medical workers fighting against SARS-CoV-2.

**Author Contributions** Conceptualization, Supervision and Writing—review and editing: MC; Methodology: BL, XP and LW; Writing—original draft preparation: JW. All authors have read and agreed to the published version of the manuscript.

**Funding** This research received no external funding.

**Data Availability** We downloaded the SARS-CoV-2 genome from the NCBI website (https://www.ncbi.nlm.nih.gov/genome/). The millions of world-wide SARS-CoV-2 sequences were downloaded from GISAID (Shu and McCauley 2017).

## Declarations

**Conflict of interest** The authors declare they have no conflict of interest.

**Ethical Approval** Not applicable.

**Consent to Participate** Not applicable.

**Consent for Publication** Not applicable.

## References

Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, Cao J, Chae E, Dezwaan TM, Ding W et al (2016) 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. Cell 166:481–491

Bass BL, Weintraub H (1987) A developmentally regulated activity that unwinds RNA duplexes. Cell 48:607–613

Bass BL, Weintraub H (1988) An unwinding activity that covalently modifies its double-stranded RNA substrate. Cell 55:1089–1098

Byng JW, Chase MW, Christenhusz MJM, Fay MF, Judd WS, Mabberley DJ, Sennikov AN, Soltis DE, Soltis PS, Stevens PF et al (2016) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc 181:1–20

Cai H, Liu X, Zheng X (2022) RNA editing detection in SARS-CoV-2 transcriptome should be different from traditional SNV identification. J Appl Genet 63:587–594

Chang S, Li J, Li Q, Yu CP, Xie LL, Wang S (2021) Retrieving the deleterious mutations before extinction: genome-wide comparison of shared derived mutations in liver cancer and normal population. Postgrad Med J.

Chu D, Wei L (2019) The chloroplast and mitochondrial C-to-U RNA editing in *Arabidopsis thaliana* shows signals of adaptation. Plant Direct 3:e00169

Chu D, Wei L (2021a) Context-dependent and -independent selection on synonymous mutations revealed by 1,135 genomes of *Arabidopsis thaliana*. BMC Ecol Evol 21:68

Chu D, Wei L (2021b) Direct in vivo observation of the effect of codon usage bias on gene expression in Arabidopsis hybrids. J Plant Physiol 265:153490

Crow JF (1955) General theory of population genetics: synthesis. Cold Spring Harb Symp Quant Biol 20:54–59

Di Giorgio S, Martignano F, Torcia MG, Mattiuz G, Conticello SG (2020) Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci Adv 6:eabb5813

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Goila-Gaur R, Strebel K (2008) HIV-1 Vif, APOBEC, and intrinsic immunity. Retrovirology 5:51

Gray MW, Covello PS (1993) RNA editing in plant-mitochondria and chloroplasts. FASEB J 7:64–71

Harris RS (2008) Enhancing immunity to HIV through APOBEC. Nat Biotechnol 26:1089–1090
Harris RS, Dudley JP (2015) APOBECs and virus restriction. Virology 479–480:131–145
Jiang Y, Ge F, Sun G, Wang H (2022) An evolutionarily conserved mechanism that amplifies the effect of deleterious mutations in osteosarcoma. Mol Genet Genomics 297:373–385
Kudla J, Igloi G, Metzlaff M, Hagemann R, Kossel H (1992) RNA editing in tobacco chloroplasts leads to the formation of a translatable PSBL messenger-RNA by a C to U substitution within the initiation codon. EMBO J 11:1099–1103
Li Q, Wang Z, Lian J, Schiott M, Jin L, Zhang P, Zhang Y, Nygaard S, Peng Z, Zhou Y et al (2014) Caste-specific RNA editomes in the leaf-cutting ant Acromyrmex echinatior. Nat Commun 5:4943
Li Y, Yang X, Wang N, Wang H, Yin B, Yang X, Jiang W (2020a) GC usage of SARS-CoV-2 genes might adapt to the environment of human lung expressed genes. Mol Genet Genomics 295:1537–1546
Li Y, Yang X, Wang N, Wang H, Yin B, Yang X, Jiang W (2020b) Mutation profile of over 4500 SARS-CoV-2 isolations reveals prevalent cytosine-to-uridine deamination on viral RNAs. Future Microbiol 15:1343–1352
Li Y, Yang XN, Wang N, Wang HY, Yin B, Yang XP, Jiang WQ (2020c) The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive RNA modification. Future Virol 15:341–347
Li Y, Yang XN, Wang N, Wang HY, Yin B, Yang XP, Jiang WQ (2020d) Pros and cons of the application of evolutionary theories to the evolution of SARS-CoV-2. Future Virol 15:369–372
Li Q, Li J, Yu CP, Chang S, Xie LL, Wang S (2021) Synonymous mutations that regulate translation speed might play a non-negligible role in liver cancer development. BMC Cancer 21:388
Liddicoat BJ, Piskol R, Chalk AM, Ramaswami G, Higuchi M, Hartner JC, Li JB, Seeburg PH, Walkley CR (2015) RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. Science 349:1115–1120
Liscovitch-Brauer N, Alon S, Porath HT, Elstein B, Unger R, Ziv T, Admon A, Levanon EY, Rosenthal JJC, Eisenberg E (2017) Trade-off between transcriptome plasticity and genome evolution in cephalopods. Cell 169:191–202
Martignano F, Di Giorgio S, Mattiuz G, Conticello SG (2022) Commentary on "poor evidence for host-dependent regular RNA editing in the transcriptome of SARS-CoV-2." J Appl Genet 63:423–428
Muller WE (1995) Molecular phylogeny of Metazoa (animals): monophyletic origin. Naturwissenschaften 82:321–329
Olson ME, Harris RS, Harki DA (2018) APOBEC enzymes as targets for virus and cancer therapy. Cell Chem Biol 25:36–49
Park L (2011) Effective population size of current human population. Genet Res (Camb) 93:105–114
Picardi E, Mansi L, Pesole G (2021) Detection of A-to-I RNA editing in SARS-COV-2. Genes (Basel) 13:41
Porath HT, Carmi S, Levanon EY (2014) A genome-wide map of hyper-edited RNA reveals numerous new sites. Nat Commun. https://doi.org/10.1038/ncomms5726
Ramaswami G, Zhang R, Piskol R, Keegan LP, Deng P, O'Connell MA, Li JB (2013) Identifying RNA editing sites using RNA sequencing data alone. Nat Methods 10:128–132
Shu Y, McCauley J (2017) GISAID: Global initiative on sharing all influenza data—from vision to reality. Euro Surveill. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494
Song Y, He X, Yang W, Wu Y, Cui J, Tang T, Zhang R (2022) Virus-specific editing identification approach reveals the landscape of A-to-I editing and its impacts on SARS-CoV-2 characteristics and evolution. Nucleic Acids Res 50:2509–2521
Wang Y, Gai Y, Li Y, Li C, Li Z, Wang X (2021) SARS-CoV-2 has the advantage of competing the iMet-tRNAs with human hosts to allow efficient translation. Mol Genet Genomics 296:113–118
Wang H, Sun G, Jiang Y (2022) Cost-Efficiency Optimization Serves as a Conserved Mechanism that Promotes Osteosarcoma in Mammals. J Mol Evol 90:139–148
Ward SV, George CX, Welch MJ, Liou LY, Hahm B, Lewicki H, de la Torre JC, Samuel CE, Oldstone MB (2011) RNA editing enzyme adenosine deaminase is a restriction factor for controlling measles virus replication that also is required for embryogenesis. Proc Natl Acad Sci USA 108:331–336
Webster MT, Smith NGC (2004) Fixation biases affecting human SNPs. Trends Genet 20:122–126
Wei L (2020) Selection on synonymous mutations revealed by 1135 genomes of *Arabidopsis thaliana*. Evol Bioinform Online 16:1176934320916794
Wei L (2022) Reconciling the debate on deamination on viral RNA. J Appl Genet 63:583–585
Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556
Yu YY, Li Y, Dong Y, Wang XK, Li CX, Jiang WQ (2021) Natural selection on synonymous mutations in SARS-CoV-2 and the impact on estimating divergence time. Future Virol 16:447–450
Zhang Y, Jin X, Wang H, Miao Y, Yang X, Jiang W, Yin B (2021a) Compelling evidence suggesting the codon usage of SARS-CoV-2 adapts to human after the split from RaTG13. Evol Bioinform Online 17:11769343211052012
Zhang YP, Jiang W, Li Y, Jin XJ, Yang XP, Zhang PR, Jiang WQ, Yin B (2021b) Fast evolution of SARS-CoV-2 driven by deamination systems in hosts. Future Virol 16:587–590
Zhang Y, Jin X, Wang H, Miao Y, Yang X, Jiang W, Yin B (2022) SARS-CoV-2 competes with host mRNAs for efficient translation by maintaining the mutations favorable for translation initiation. J Appl Genet 63:159–167
Zhu L, Wang Q, Zhang W, Hu H, Xu K (2022) Evidence for selection on SARS-CoV-2 RNA translation revealed by the evolutionary dynamics of mutations in UTRs and CDSs. RNA Biol 19:866–876
Zong J, Zhang Y, Guo F, Wang C, Li H, Lin G, Jiang W, Song X, Zhang X, Huang F et al (2022) Poor evidence for host-dependent regular RNA editing in the transcriptome of SARS-CoV-2. J Appl Genet 63:413–421