The prevalent RNA modification on SARS-CoV-2 RNAs may confound the SNP profile and evolutionary patterns revealed by previous studies

Yan Wang
Qingdao Central Hospital
Yanhong Gai
Qingdao Central Hospital
Yuefan Li
Qingdao Central Hospital
Chunxiao Li
Qingdao Central Hospital
Ziliang Li
Qingdao Central Hospital
Xuekun Wang (✉ jwqwxk@163.com)
Qingdao Central Hospital https://orcid.org/0000-0002-8172-4590

Research

Keywords: SARS-CoV-2, RNA virus, RNA modification, SNP, indistinguishable

Posted Date: July 22nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-41421/v1

Abstract

Background

The recent outbreak of SARS-CoV-2 has caused severe damage to the world. The concomitant papers on the evolutionary patterns of SARS-CoV-2 is continuously emerging. Studies has utilized the publically available RNA-seq data to find out the so-called SNPs in the virus genome and analyzed their selection patterns.

Methods

We downloaded a set of RNA-seq data and performed a well-established but modified variant calling pipeline to allow the identification of multiple clustered mutations.

Results

We found prevalent “putative” but reliably detected A-to-G RNA modifications in the RNA-seq data of SARS-CoV-2 with high signal to noise ratios, presumably caused by the host’s deamination enzymes. Importantly, since SARS-CoV-2 is an RNA virus, it is technically impossible to truly distinguish SNPs and RNA modifications from the RNA-seq data alone.

Conclusions

The technically indistinguishable RNA modifications and SNPs of SARS-CoV-2 have complicated the situation where many researchers intend to unveil the evolutionary patterns behind the mutation spectrum. This is not a problem for DNA organisms but should be seriously considered when we are investigating the RNA viruses.

Background

The outbreak of SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) in the beginning of year 2020 has caused severe damage to China especially the Hubei province [1–3]. Recently, several other countries like America, France, England, Italy, Russia, and Spain are consecutively being hit by this virus. There is urgent need to understand the origin and evolution of SARS-CoV-2 and related coronaviruses [4].

Papers on the evolutionary patterns of SARS-CoV-2 have emerged as rapidly as the outbreak of virus. Several studies downloaded the publically available RNA-seq data and performed “SNP calling”. The SNP distribution or frequency spectrum is usually a super informative inference of selection patterns. The recent study has discussed topics on the origin and continuing evolution of SARS-CoV-2. Although questioned methodologically by other researchers, their attempt to utilize the SNP data called from RNA-seq is unchallenged.

To better illustrate the pipeline of SNP calling, let us take the human genome for instance. The SNP calling process is accomplished by mapping the DNA-seq reads of a sample to the human reference genome, and the reliable mismatch sites found should be a potential SNP (Fig. 1A). If the RNA-seq of the same sample is available, one could find the same nucleotide change from the reference genome to the RNA-seq reads (Fig. 1A), indicating that the mutation takes place at DNA level. However, one could not determine the direction of the mutations without an outgroup (Fig. 1A). In contrast, if a variation site is found in the RNA-seq reads but not the DNA-seq reads, then this is possibly an RNA modification site. For example, the vertebrate adenosine deaminase would change adenosine to guanosine [5], which is interpreted as guanosine in the sequencing data. Thus, the A-to-G variations between RNA and reference genome are presumably caused by the deamination enzyme (Fig. 1A). Unlike the unknown direction of DNA mutations, the direction of A-to-G deamination is very clear even without an outgroup because it takes place at the RNA level (Fig. 1A).

SARS-CoV-2 is a positive strand RNA virus. The so-called reference genome is actually the RNA sequence. Without a DNA template, the mismatches found between the reference and the RNA-seq reads could either be a “SNP” or RNA modification site (Fig. 1B). It is futile to try any filtering criteria on these RNA-seq data because the SNPs and RNA modification sites are technically indistinguishable. Even when multiple outgroup species are available, the reference sequence (RNA) of the outgroup viruses may also undergo the same RNA modification process (by host cells), making it difficult to define the ancestral state and the direction of mutations (Fig. 1B).

In this study, using a well-established mutation finding pipeline (see Fig. 2 and **Methods**), we found prevalent A-to-G RNA modifications in the RNA-seq data of SARS-CoV-2. If other non-A-to-G variations could not be explained by known RNA modification systems and are regarded as SNPs, then we found non-identical codon position preference between the A-to-G and non-A-to-G variations, suggesting that their missense and synonymous profiles are different. Therefore, mixing all the variation sites in the SARS-CoV-2 evolutionary analyses is inaccurate and seems to be a hybrid between traditional SNP analyses and RNA modification analyses.

Results

Variation sites identified by normal mapping pipeline

A recent study (https://www.biorxiv.org/content/10.1101/2020.03.02.973255v2) dealt with similar issues using the same sets of data. We first compared our results with theirs. Here we used ordinary mapping pipeline (not the transform strategy, see **Methods**). Theoretically, the transform strategy was designed to retrieve many clustered variation sites while the normal variant calling pipeline might find fewer sites. Indeed, according to the relevant study, datasets SRR10903401, SRR10903402, SRR11059942, and

SRR11059945 contained 25, 163, 208, and 82 variant sites. Our results showed good overlapping with the study (Fig. 3A) but we have identified slightly more variation sites, possibly due to the different software used or that study performed additional filters. However, for the shared variation sites found by us and others, the correlation of alternative allele frequency is nearly perfect (Fig. 3A). The percentages of different variation types also conform with the previous study (Fig. 3B).

As we have said, the mutation sites found by traditional variant calling pipeline only represent a minority of all possible RNA modification sites in the transcriptome, and may not produce a strikingly high percentage of A > G mutation. In the next sections, we would no longer discuss these sets of variation sites. We would use the transform strategy as introduced in the Methods (Fig. 2) and look at the prevalent base substitution events across the transcriptome.

The prevalent A-to-G variations across SARS-CoV-2 genes

We downloaded the reference and a set of RNA-seq data of SARS-CoV-2. We mapped the RNA-seq reads to the reference sequence with a well-established pipeline (see **Methods** for details). The numbers of (non-unique) mismatch events profiled (Fig. 4A). There are 5310 (59.1%) A-to-G mismatches and 2015 (22.4%) G-to-A mismatches, and the ten other types of mismatches composed only 18.5% of the totally 8989 mismatches (Fig. 4A). The most prevalent A-to-G mismatches could be interpreted as the adenosine-to-inosine deamination conferred by the host cells. For the second-most prevalent G-to-A mismatches, it is possible that the reference sequence (RNA) of SARS-CoV-2 itself suffered from deamination by the host cells. This is also why we worry that the phylogenetic tree constructed by multiple virus sequences could also be skewed by the RNA modification from the hosts. The third-most abundant T-to-C and C-to-T mismatches might represent the cytosine-to-uridine deamination system in the host cells.

We treated the A-to-G as the A-to-G RNA modification sites in the virus sequences. We found that the density of A-to-G modification varied moderately across different genes (Fig. 4B). To technically validate that the mismatches sites are reliably detected rather than artifacts or errors produced from the pipeline, we manually extracted a 150 bp read and aligned it to the reference sequence (Fig. 5). The A-to-G alterations are clearly presented in the reference versus RNA-seq comparison. Pay attention that errors could include the mismatches resulted from mis-alignments or sequencing errors. The validation here is to check the accuracy of the mapping pipeline. The manually examined read told that the sequence alignment is reliable. The control for sequencing errors would be discussed in the following section.

Robustly observed A-to-G variations under different criteria

The nearly nine thousand (non-unique) variation events shown in the above section belong to 4604 unique variation sites. Most of the 4604 unique sites have less than 10 reads supporting the alternative allele (Fig. 6A). There are 2878 (62.5%) unique A-to-G variation sites and 998 (21.7%) unique G-to-A variation sites (Fig. 6A). The signal to noise ratio for A-to-G variations is 1.67. If combined with G-to-A variations, the signal to noise ratio would be as high as 5.32. Among the 2878 unique A-to-G sites, 797 and 2391 sites were found in the two SRR samples, respectively, and their overlapping is 310 sites.

Since these results came from the mapping and variant calling procedures without any filtering criteria, we think it is necessary to see whether the patterns are sensitive to some filtering parameters. We re-did the analysis by requiring mapping quality > 25 and base quality (controlling for sequencing errors) > 35. We found that the number of unique variation sites (3129) slightly declined but the majority (2421, 77.4%) of which are still A-to-G variations (Fig. 6B).

We also wish to prove the reliability of the putative A-to-G modification sites from another angle. The base context of the A-to-G variation sites showed an obvious depletion of G upstream to the putative A-to-G modification sites (Fig. 7A). In contrast, the G-to-A variation sites did not have such a key validation feature (Fig. 7B). This consolidated our assumption that these A-to-G variations are RNA modification sites.

Discussion

The SNPs and RNA modification sites could bear completely different mutation rates and position biases, and might undergo different selection patterns. A mixture of SNPs and RNA modification sites does not tell either of their evolutionary patterns. Unfortunately, these two mutation sources could not be separated from the RNA-seq data from RNA viruses.

For DNA organisms like human, the pair-ended DNA-seq could be either mapped to the positive or negative strand of reference genome. But for single-ended human transcriptome data mapped to the transcriptome sequence, all reads should be theoretically mapped to the positive strand of transcriptome (except when the library itself is on the opposite direction). The same goes for RNA viruses like SARS-CoV-2 for either single-ended or pair-ended RNA-seq library. The authors did not mention the detailed mapping status of the reads. Although it is technically difficult to distinguish the SNPs and RNA modification sites in the viral RNA, at least this unsolvable concern should be stated in the manuscript.

If it is confirmed that the SARS-CoV-2 has been transferred from patient No.1 to patient No.2, then one might consider the RNA-seq from patient No.1 should be the ancestral state. However, the viral RNAs in patient No.1 would also undergo the same A-to-G modification by the hosts. The deamination enzyme only modifies a fraction of the total viral RNAs so that in patient No.1 there is still a mixture of A-version and G-version RNA reads. Technically, one could not know whether this is a polymorphic site in the virus population or it is modified by the host’s enzyme. This uncertainty makes it difficult to define the ancestral state.

There is a less important but unsolved question that we think the G-to-A variation sites could also be the A-to-G modification on RNA of the “reference genome”. However, from the base context of the G-to-A sites, they did not seem to be authentic A-to-G modification sites. This strange pattern remains an open question.

In summary, the optimized algorithms (in numerous software) are only to improve the accuracy of the alignments. Even an alignment is manually verified, we still do not know any A-G mismatches in the alignment should be SNPs or RNA modification sites. We appeal that this issue should be seriously discussed in the studies involving RNA viruses like SARS-CoV-2.

Conclusions

The technically indistinguishable RNA modifications and SNPs of SARS-CoV-2 have complicated the situation where the researchers intend to reveal the evolutionary patterns behind the mutation spectrum. This is not a problem for DNA organisms but should be seriously considered when we are investigating the RNA viruses.

Methods

Data collection

We downloaded the novel coronavirus SARS-CoV-2 genome from the NCBI website (https://www.ncbi.nlm.nih.gov/genome/). The coding sequences were extracted according to the genome annotation. The RNA-seq data were retrieved from NCBI under accession numbers SRR10903401 and SRR10903402. Other two relevant datasets SRR11059942 and SRR11059945 were also downloaded and analyzed.

RNA-seq analyses

We mapped the RNA-seq reads to the CDS of SARS-CoV-2 using BWA mem [6] with default parameters but with a little modification (Fig. 2). Reads with too many mismatches could not be aligned to the reference genome. However, the many clustered mismatches could be RNA modifications. To retrieve more RNA modification sites in clusters like this, we transformed the reference sequence and the RNA-seq reads manually [7]. The transformation could be either of the twelve types of mismatches. The transformed reads mapped to the transformed genome were replaced with the original reads and the unmodified genome. The mismatch sites were extracted from the alignment. We made two versions here, one version is the variation sites without additional filters, another is the mutation sites under the criteria of mapping quality > 25 and base quality > 35. The “transformation followed by re-mapping” workflow is a well-acknowledged pipeline to detect the RNA modification events omitted by traditional mapping procedures [7, 8]. We should emphasize that this pipeline only deals with the reads that could not be mapped by normal procedures. For the majority of reads that could be mapped by normal procedures (which might also contain many RNA modification events), we would discuss their variation profile separately. Using the normal variant calling pipeline (or termed ordinary pipeline), we also required mapping quality > 25 and base quality > 35 to conform with the protocol of a very relevant study posted as preprint recently (https://www.biorxiv.org/content/10.1101/2020.03.02.973255v2). The difference is that we used samtools to find the variants and that study used other software.

Statistical analyses

We used the R language to perform the statistical analyses and graphic work. We also used EXCEL to plot some figures when necessary.

Data Availability

All data used in our study are public data. The accession numbers of the RNA-seq data are SRR10903401, SRR10903402, SRR11059942, and SRR11059945.

Abbreviations

SARS-CoV-2
Severe Acute Respiratory Syndrome Coronavirus 2.
CDS
coding sequence.
RNA-seq
RNA sequencing.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data used in our study are public data. The accession numbers of the RNA-seq data are SRR10903401, SRR10903402, SRR11059942, and SRR11059945.

Competing interest

The authors declare they have no competing interest.

Funding

No funding has supported this manuscript.

Authors' contributions

The corresponding author designed and supervised this research. All authors contributed to writing the manuscript. All authors read and approved the final version.

Acknowledgements

We thank the members in our group that have given suggestions to our project. This is an extremely hard time for the medical staff like us. We managed to finish this article in the part time when most of our colleagues are devoted to saving the patients' lives. We also thank the recent studies on coronavirus which have enlightened us. We thank an anonymous colleague for several novel ideas that have help the formation of this article. We appreciate the other medical staff numbers in the world who are still fighting against SARS-CoV-2 in the front-line of the war-field. Finally, we memorize the victims as well as the people who have sacrificed their lives in this invisible warfare.

References

1. Cowling BJ, Leung GM. **Epidemiological research priorities for public health control of the ongoing global novel coronavirus (2019-nCoV) outbreak**. *Euro Surveill* 2020.
2. Hui DS, Madani EIA, Ntoumi TA, Kock F, Dar R, Ippolito O, McHugh G, Memish TD, Drosten ZA. C et al: The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China. Int J Infect Dis. 2020;91:264–6.
3. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. Lancet. 2020;395(10223):470–3.
4. Edelstein M, Heymann DL. What needs to be done to control the spread of Middle East respiratory syndrome coronavirus? Future Virol. 2015;10(5):497–505.
5. Bass BL, Weintraub H. A Developmentally Regulated Activity That Unwinds Rna Duplexes. Cell. 1987;48(4):607–13.
6. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
7. Porath HT, Carmi S, Levanon EY. **A genome-wide map of hyper-edited RNA reveals numerous new sites**. *Nature Communications* 2014, 5.
8. Porath HT, Knisbacher BA, Eisenberg E, Levanon EY. Massive A-to-I RNA editing is common across the Metazoa and correlates with dsRNA abundance. Genome Biol. 2017;18(1):185.

Figures

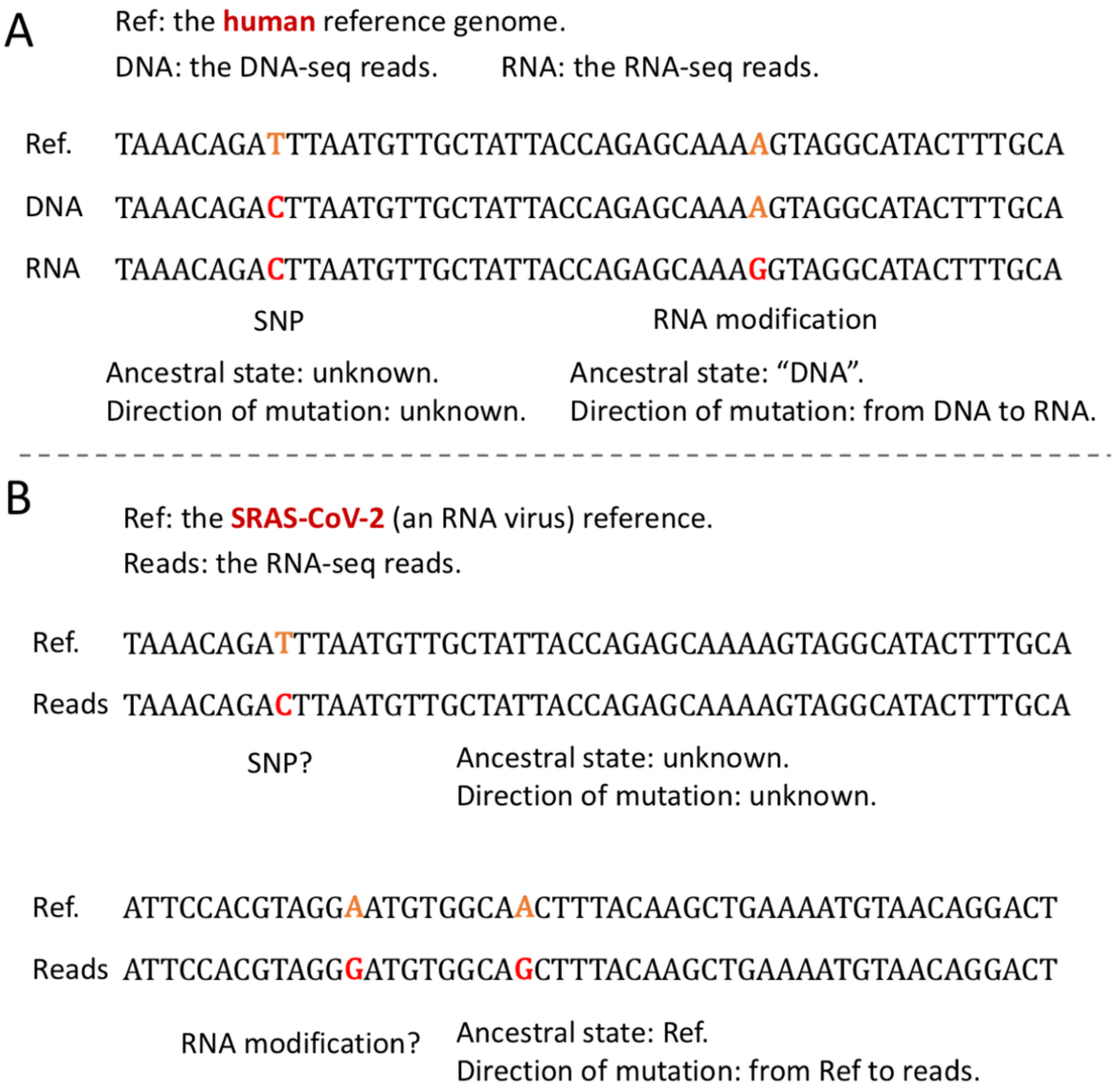
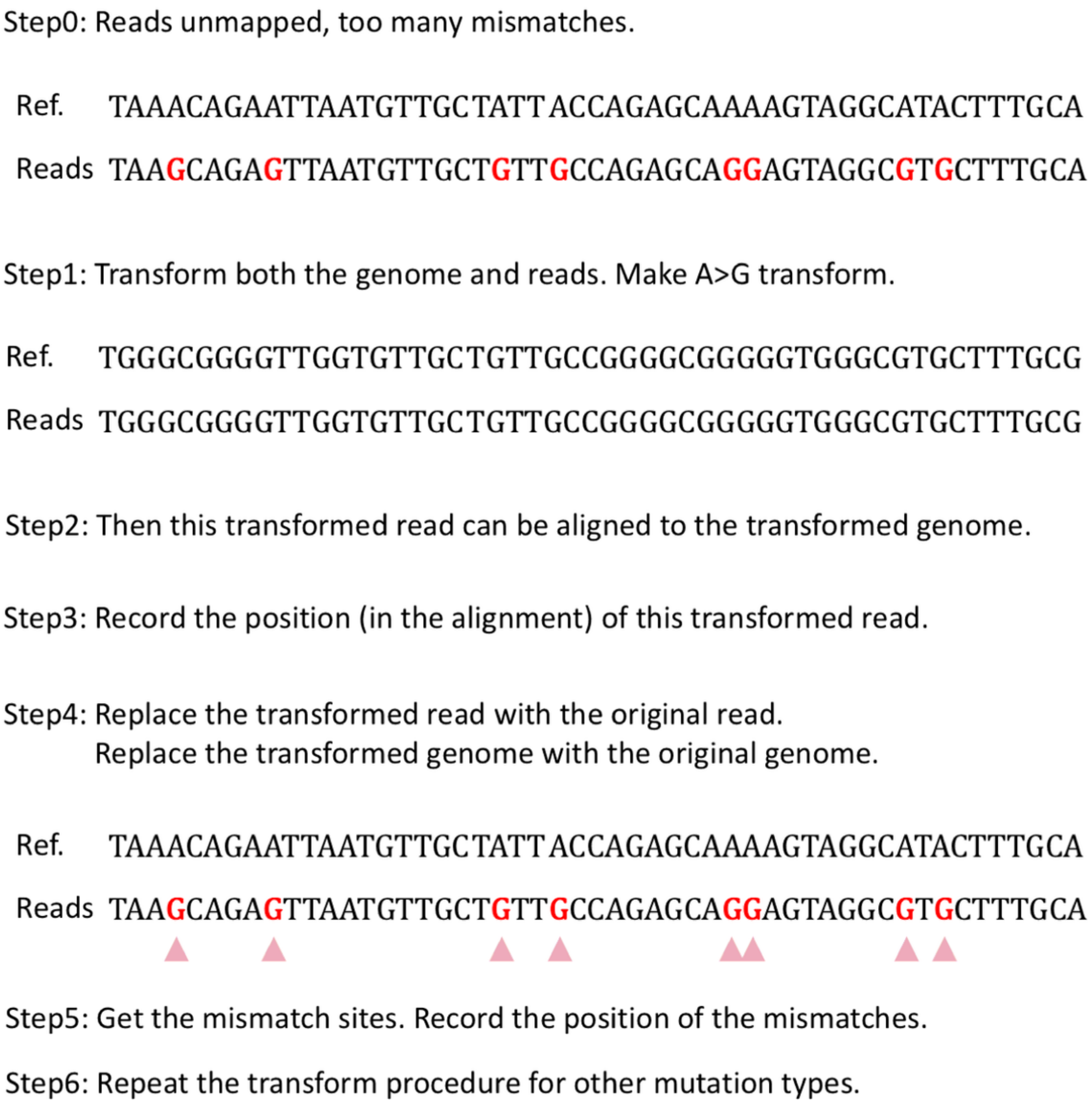

A
Ref: the human reference genome.
DNA: the DNA-seq reads.
RNA: the RNA-seq reads.
Ref. TAAACAGATTTAATGTTGCTATTACCAGAGCAAAAGTAGGCATACTTTGCA
DNA TAAACAGACTTAATGTTGCTATTACCAGAGCAAAAGTAGGCATACTTTGCA
RNA TAAACAGACTTAATGTTGCTATTACCAGAGCAAAGGTAGGCATACTTTGCA
SNP
RNA modification
Ancestral state: unknown.
Direction of mutation: unknown.
Ancestral state: “DNA”.
Direction of mutation: from DNA to RNA.
B
Ref: the SRAS-CoV-2 (an RNA virus) reference.
Reads: the RNA-seq reads.
Ref. TAAACAGATTTAATGTTGCTATTACCAGAGCAAAAGTAGGCATACTTTGCA
Reads TAAACAGACTTAATGTTGCTATTACCAGAGCAAAAGTAGGCATACTTTGCA
SNP?
Ancestral state: unknown.
Direction of mutation: unknown.
Ref. ATTCCACGTAGGAATGTGGCAACTTTACAAGCTGAAAATGTAACAGGACT
Reads ATTCCACGTAGGGATGTGGCAGCTTTACAAGCTGAAAATGTAACAGGACT
RNA modification?
Ancestral state: Ref.
Direction of mutation: from Ref to reads.

Figure 1

An illustration of the relationship between reference genome, DNA-seq reads, and RNA-seq reads. (A) The human reference genome, DNA-seq reads, and RNA-seq reads. (B) The SARS-CoV-2 reference sequence and RNA-seq reads. With the same observation, one could either regard the mismatch as a SNP or treat it as an RNA modification site. These two possibilities are technically indistinguishable. For RNA viruses, any software could only help improve the accuracy of the alignment rather than tell us whether the mismatch is a SNP or RNA modification site.



Step0: Reads unmapped, too many mismatches.

Ref. TAAACAGAATTAATGTTGCTATT ACCAGAGCAAAAGTAGGCATACTTTGCA

Reads TAAGCAGAGTTAATGTTGCTGTTGCCAGAGCAGGAGTAGGCGTGCTTTGCA

Step1: Transform both the genome and reads. Make A>G transform.

Ref. TGGGCGGGGTTGGTGTTGCTGTTGCCGGGGCGGGGGTGGGCGTGCTTTGCG

Reads TGGGCGGGGTTGGTGTTGCTGTTGCCGGGGCGGGGGTGGGCGTGCTTTGCG

Step2: Then this transformed read can be aligned to the transformed genome.

Step3: Record the position (in the alignment) of this transformed read.

Step4: Replace the transformed read with the original read.
Replace the transformed genome with the original genome.

Ref. TAAACAGAATTAATGTTGCTATT ACCAGAGCAAAAGTAGGCATACTTTGCA

Reads TAAGCAGAGTTAATGTTGCTGTTGCCAGAGCAGGAGTAGGCGTGCTTTGCA

Step5: Get the mismatch sites. Record the position of the mismatches.

Step6: Repeat the transform procedure for other mutation types.

Figure 2

The transform strategy of mapping the reads with multiple mismatches.

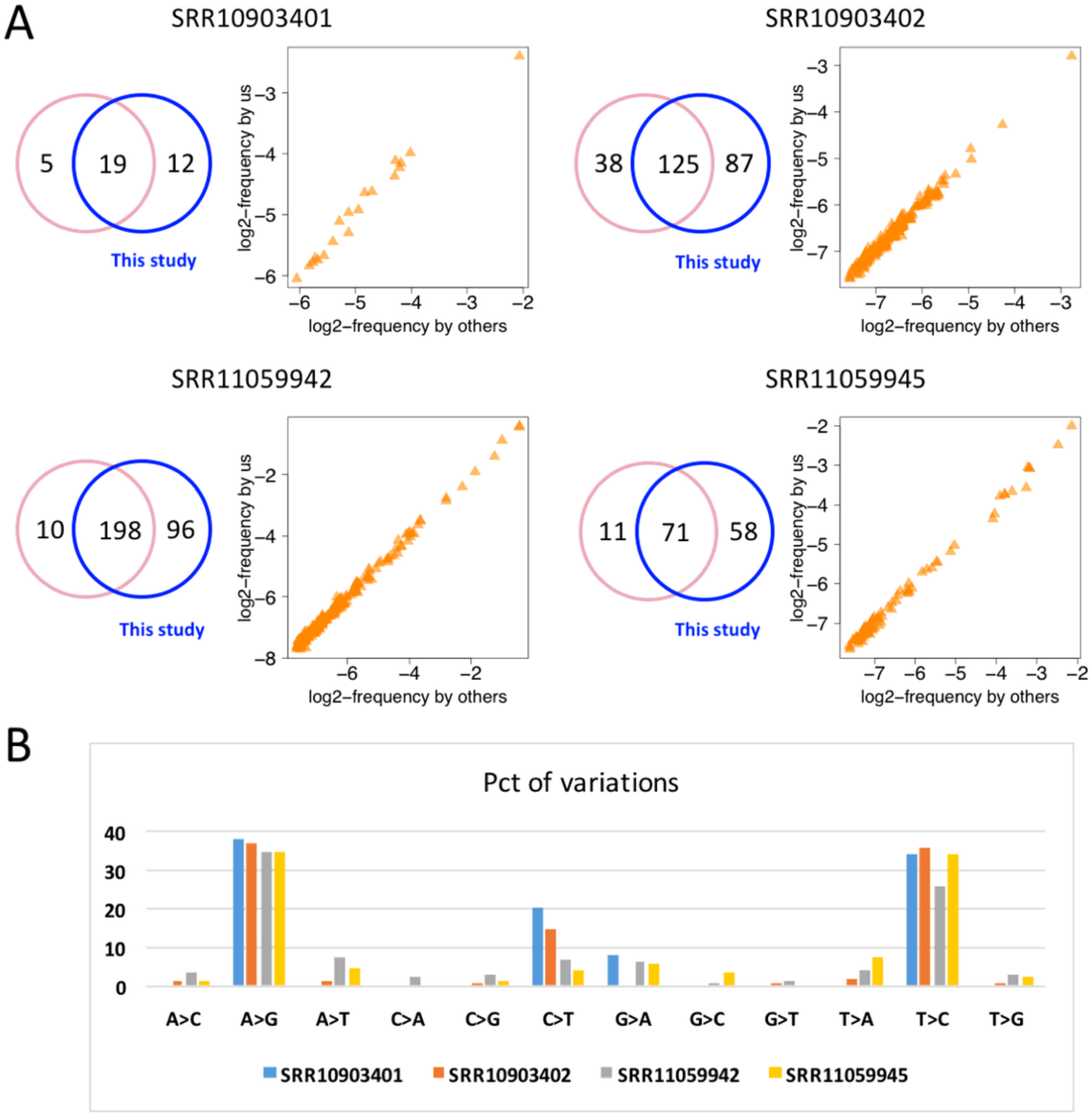


Figure 3

Comparison of results of this study with a recent relevant study. (A) The overlapping of variation sites, and the correlation of frequency (alternative allele count divided by total reads count covering a site) of the shared variation sites. (B) The percentages of each variation type in different samples.

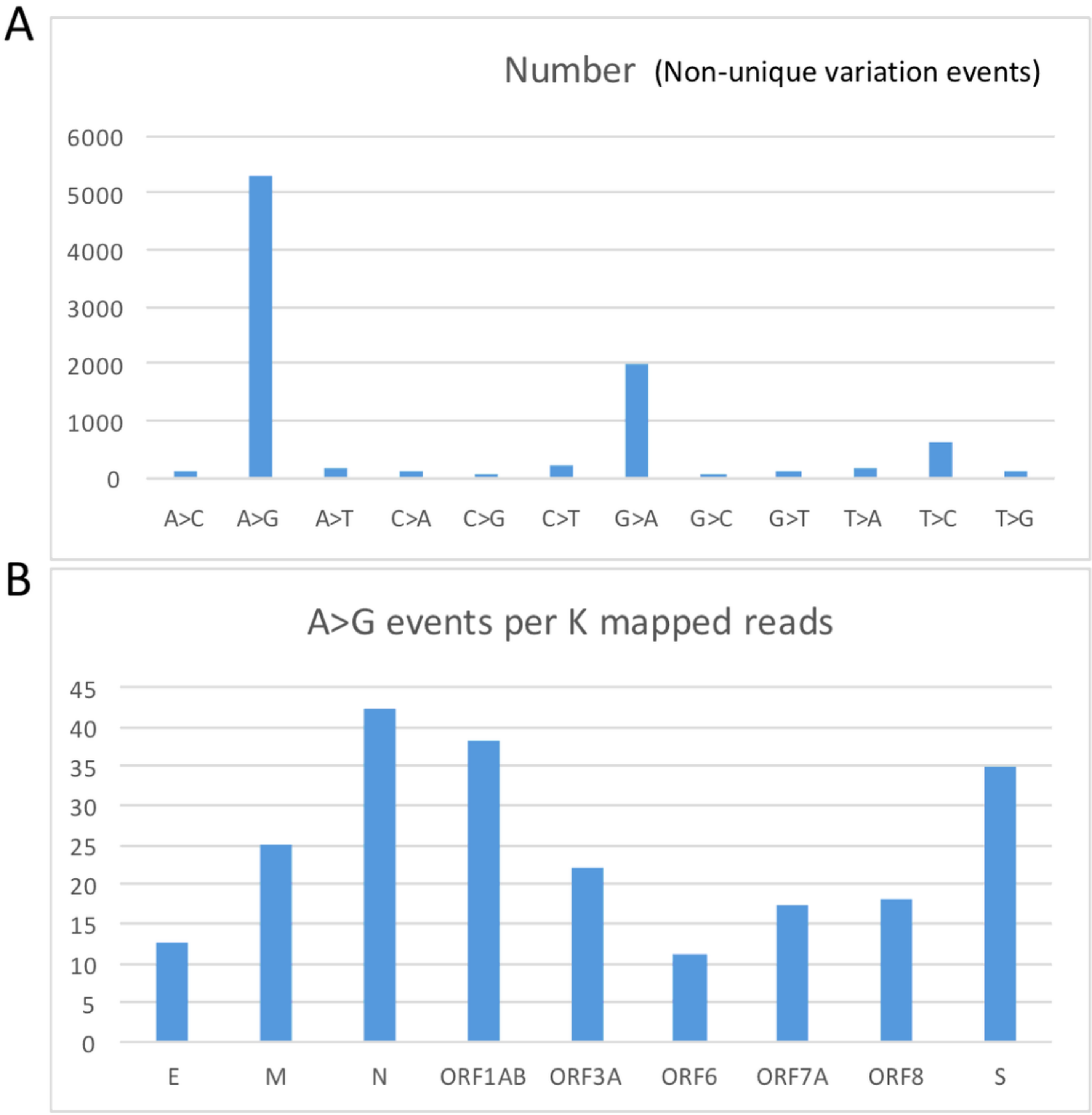
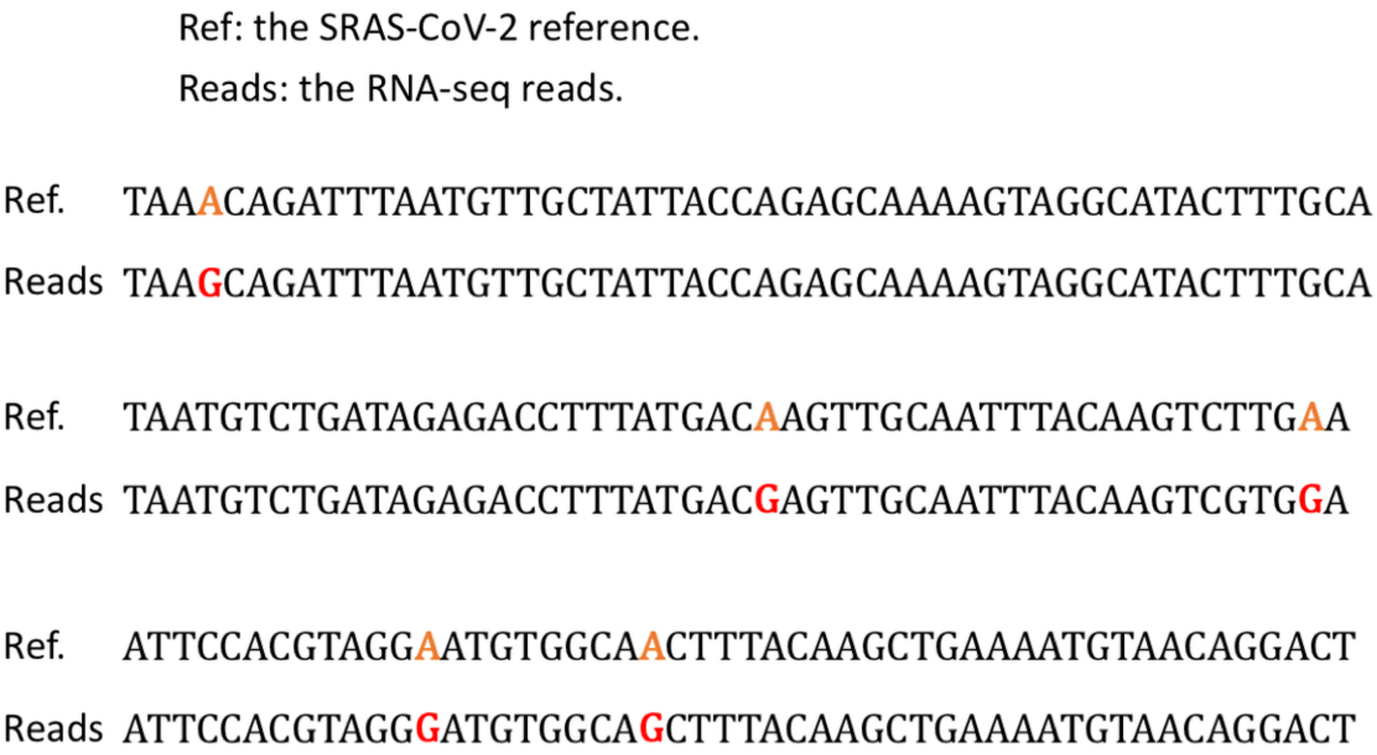


Figure 4

The mismatch profile of a set of RNA-seq data from SARS-CoV-2. (A) Numbers of mismatch events (not unique sites). (B) The normalized number of A-to-G mismatches per gene. There are totally 11 non-redundant genes in the SARS-CoV-2 sequence, and the A-to-G alterations are found in 9 of those genes.



Ref: the SRAS-CoV-2 reference.

Reads: the RNA-seq reads.

```
Ref.   TAAACAGATTTAATGTTGCTATTACCAGAGCAAAAGTAGGCATACTTTGCA
Reads  TAAGCAGATTTAATGTTGCTATTACCAGAGCAAAAGTAGGCATACTTTGCA

Ref.   TAATGTCTGATAGAGACCTTTATGACAAGTTGCAATTTACAAGTCTTGAA
Reads  TAATGTCTGATAGAGACCTTTATGACGAGTTGCAATTTACAAGTCGTGGA

Ref.   ATTCCACGTAGGAATGTGGCAACTTTACAAGCTGAAAATGTAACAGGACT
Reads  ATTCCACGTAGGGATGTGGCAGCTTTACAAGCTGAAAATGTAACAGGACT
```

Figure 5

An example of an alignment between an RNA-seq read and the reference sequence of SARS-CoV-2. The five A-to-G mismatch sites are colored in the figure.

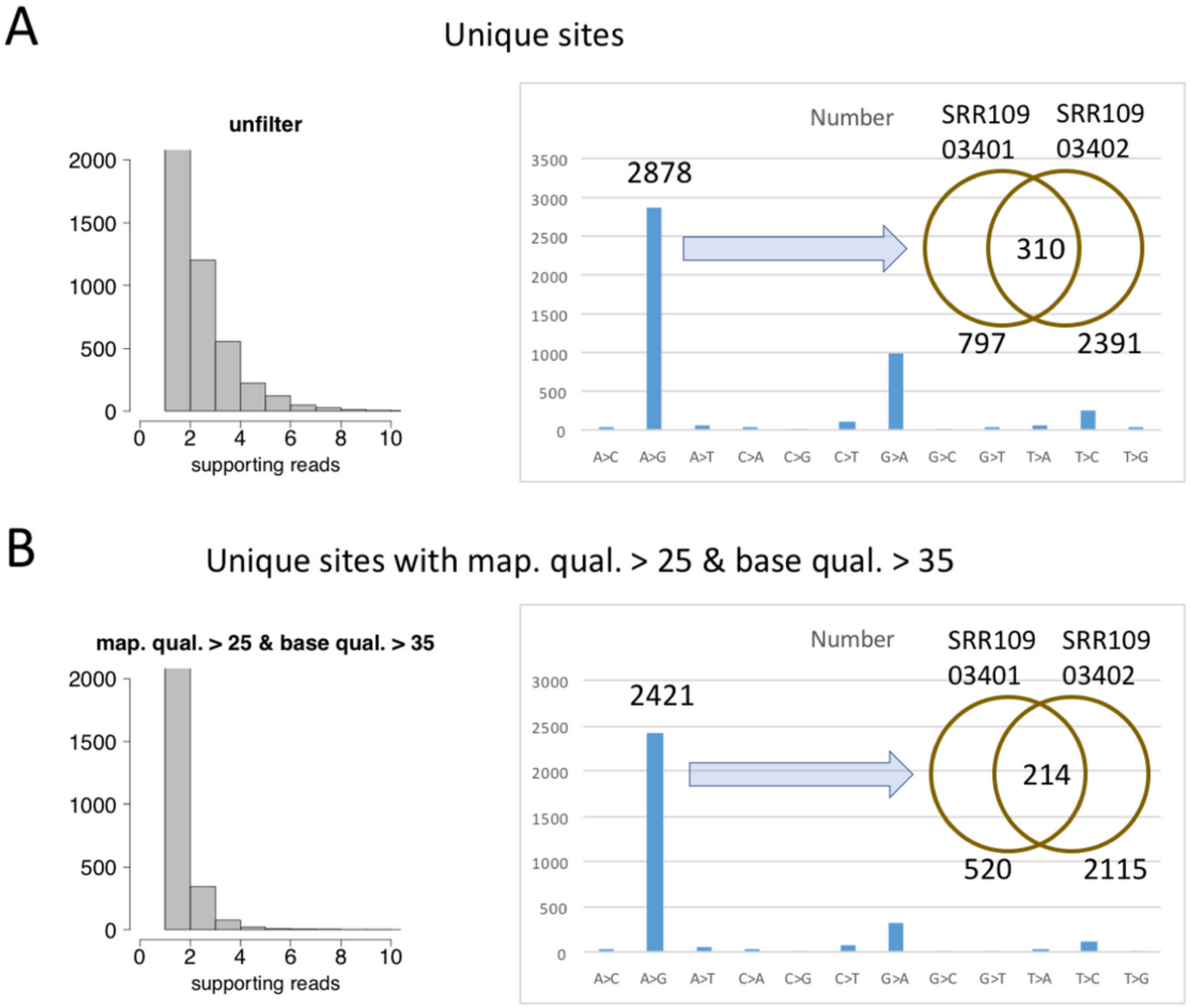


Figure 6

The numbers of unique mutation sites. (A) Distribution of reads count supporting each variation site and the numbers of different mutation types. This is the unfiltered results. (A) Distribution of reads count supporting each variation site and the numbers of different mutation types. This is the results with mapping quality > 25 and base quality > 35.

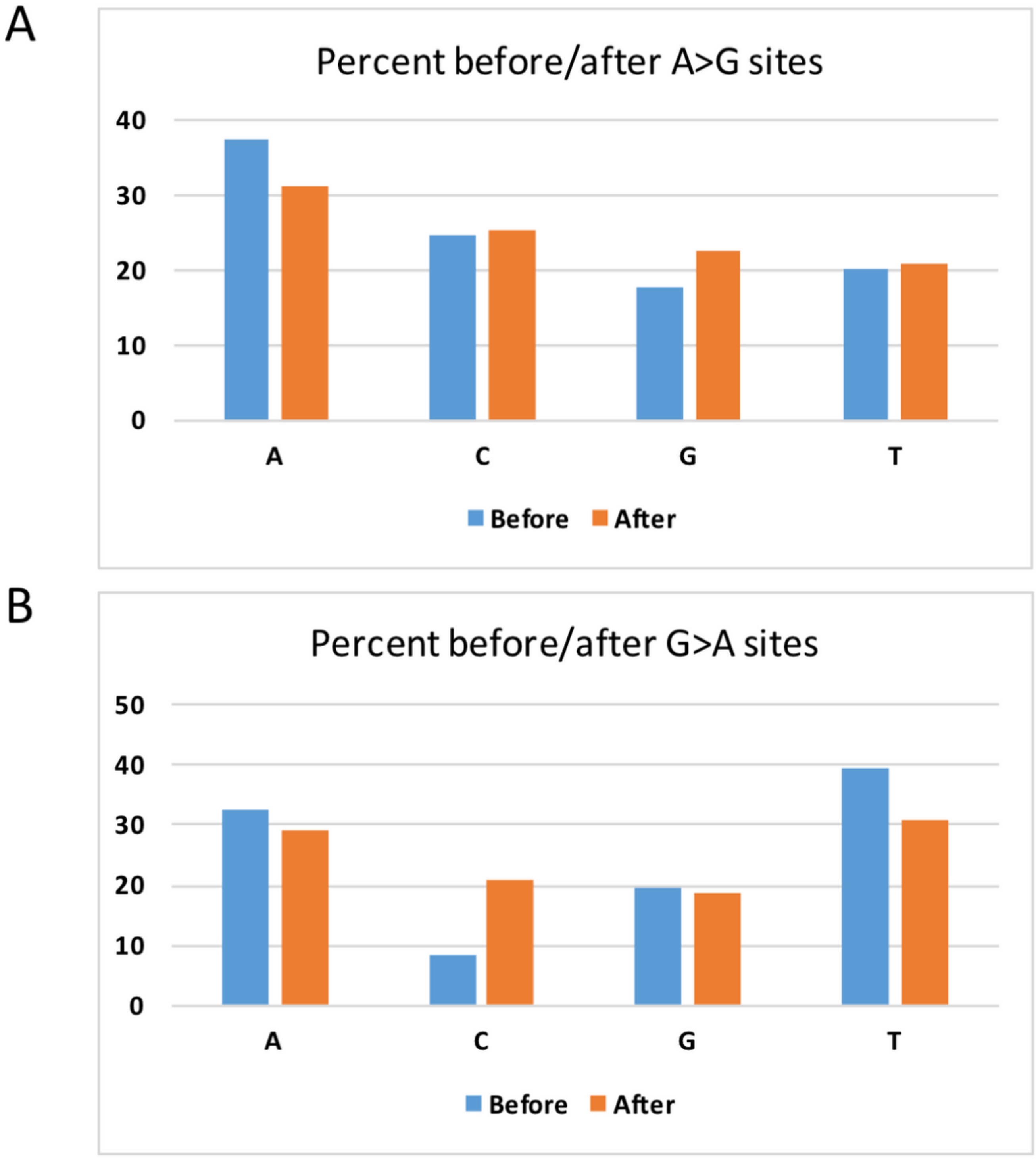


Figure 7

The base context of unique mutation sites. The percentage of A, C, G, and T is provided as bars. (A) Context of A-to-G sites. (B) Context G-to-A sites. “Before” means the five-prime nucleotide. “After” means the three-prime nucleotide.