Mutation profile of over 4,500 SARS-CoV-2 isolations reveals prevalent cytosine-to-uridine deamination on viral RNAs

Wenqing Jiang[1]*

1. Department of Respiratory Diseases, Qingdao Haici Hospital, China

*Contact:

Wenqing Jiang

Email: qdhospit87@163.com

Running head: Deamination on SARS-CoV-2 RNAs

Abstract

Aims: The sequencing data of SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus-2) are rapidly emerging. The mutation profile across SARS-CoV-2 populations is an important inference of the evolution of coronaviruses.

Materials & Methods: With 4,521 lines of SARS-CoV-2, we obtained 3,169 unique point mutation sites in the SARS-CoV-2 genome. We counted the numbers and calculated the MAF (minor allele frequency) of each mutation type.

Results: Nearly half of the point mutations are C-T mismatches and 20% are A-G mismatches. The MAF of C-T and A-G mismatches is significantly higher than MAF of other mutation types.

Conclusions: The excessive C-T mismatches do not resemble the random mutation profile, and are likely to be explained by the cytosine-to-uridine deamination system in hosts. Not only the population analyses in previous studies are questionable, but also the 17% divergence between SARS-CoV-2 and RaTG13 could be erroneous due to the deamination.

Keywords: SARS-CoV-2; 4521 lines; mutations; MAF (minor allele frequency); deamination

Introduction

The outbreak of SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) allows numerous research groups to sequence different samples of SARS-CoV-2 isolated from patients of different countries [1-5]. By mapping these sequences to the reference "genome" of SARS-CoV-2, multiple mutation sites are identified and restored in public databases (https://bigd.big.ac.cn/ncov/variation/statistics?lang=en).

In population genetics, allele frequency refers to the frequency of the mutation across the whole population. A mutation of high allele frequency simply implies a potentially beneficial effect of the mutation [6], and vice versa. In the case of SARS-CoV-2, the mutation profile across SARS-CoV-2 populations (or otherwise termed lines/isolations) is an essential inference of the origin and evolution of coronaviruses. Although the mutation profile itself is important, what is more important is the accurate definition, identification, and interpretation of these mutations. Even a researcher ensures that the bioinformatic pipeline is infallible, if the interpretation of the mutation is biased, then the conclusion could be uncertain and inaccurate.

In the bioinformatic community, although people habitually call it the reference "genome" of SARS-CoV-2 [3, 4], we should note that SARS-CoV-2 is an RNA virus and it does not have a real "genome". It is better called it the reference sequence used for alignment. In contrast, the DNA organisms like plants and animals have a real reference genome as we usually call them [7, 8]. Why should we distinguish RNA viruses and DNA organisms? Technically, the mutation identification pipelines have no difference between

RNA viruses and DNA organisms. However, the interpretation of the identified mutations is largely based on the will of the researcher. Here are some concrete examples.

For DNA organisms like human, the mutation calling process is accomplished by mapping the DNA-seq reads to the human reference genome, and the reliable mismatch sites found should be a potential SNP (Figure 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 (Figure 1A), indicating that the mutation takes place at DNA level. However, one could not determine the direction of the mutations without an outgroup (Figure 1A). In contrast, if a variation site is found in the RNA-seq reads but not in the DNA-seq reads, then this is possibly an RNA modification site. For example, the ADAR [9] and APOBEC [10, 11] enzymes in animals, causing a A-G or C-T mismatch in the RNA-seq data. Thus, the A-G or C-T variations between RNA and reference genome are presumably caused by the deamination enzyme (Figure 1A). Unlike the unknown direction of DNA mutations, the direction of A-G and C-T deamination is very clear because it takes place at the RNA level (Figure 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 (Figure 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 used, 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 (Figure 1B).

Our goal is to show that some mutations identified as "SNPs" in SARS-CoV-2 lines might actually be C-T or A-G deamination sites rather than the commonly defined SNPs (introduced by viral RNA replication errors). A rational way to distinguish these two mutation sources is to see the relative proportion of different mutation types. For commonly defined SNPs, the occurrence of transition should be roughly two times higher than the occurrence of transversion, and the allele frequency among population should not show significant differences. If the mutation profile obviously deviates from that proportion, then we should look for other explanations for the skewed mutation profile.

For the mutations in SARS-CoV-2 lines, the alternative explanation for the disproportionate mutation profile could be the ADAR and APOBEC deamination systems in human hosts, leading to excessive A-G and C-T mismatches in the sequence alignment. In this article, we are going to test if this hypothesized scenario exists.

Materials & Methods

Data download

We downloaded the vcf file (file name "2019-nCoV_total.vcf") of SARS-CoV-2 mutations from China National Center for Bioinformation (CNCB) 2019 Novel Coronavirus

Resource (2019nCoVR). The link is https://bigd.big.ac.cn/ncov/variation/statistics?lang=en (download date: 13 April 2020).

File description

The file contains 4,530 columns and 3,607 non-header lines. Each line represents one variation site. The first nine columns are information fields. Column one to column three tell the genomic position of the variation site. Column four is the reference allele of SARS-CoV-2. Column five is the alternative allele. If more than one alternative allele exists, then column five would be comma-separated. If the variation site is a point mutation site, the reference allele and alternative allele(s) would be single base. If the variation is an Indel, the reference allele or alternative allele(s) would be multiple bases. If the alternative allele is ambiguously sequenced, then the nucleotide would be degenerated symbols like Y, W, etc.

Column six to column eight is empty. Column nine just tell users that the following columns contain genotype information.

From column 10 to column 4,530, there are totally 4,521 columns, each containing the genotype information of a SARS-CoV-2 isolation line. The genotypes belong to one of the alternative alleles in column five.

Mutation and filter

We removed Indels and ambiguously sequenced bases. 3,169 unique point mutation sites are maintained in our analyses. With our own python scripts, we counted the numbers of reference allele and each alternative allele at every point mutation site.

MAF (minor allele frequency)

We defined major allele and minor allele. For each mutation site, major allele is the allele of the highest count. Minor allele is the allele of the second highest count (not the lowest count). Most mutation sites have no more than one alternative allele so that the definition of minor allele only slightly affects the results. MAF (minor allele frequency) is calculated as minor allele count divided by all allele counts at a site.

Mutation profile: permutation (directional) and combination (non-directional)

There are totally twelve possible permutations and six possible combinations of the reference allele and alternative allele. Permutation is directional, where the mutation is defined (by us in this article) as from the major allele to minor allele. Combination is non-directional, where the mutation could be wither from the major allele to minor allele or the opposite way. We recommend using non-directional mutation profile so that six mutation types are available. Because we focus on the C-to-T and A-to-G deamination, both the reference genome and the later RNA-seq data could undergo these deamination events. There is no reason to deem that the direction is from reference to alterative allele. Even if an outgroup is used (the outgroup should also be an RNA virus), one could not rule

out the possibility that the outgroup sequence also undergo C-to-T and A-to-G deamination.

Results

Mutation profile of 4,521 lines of SARS-CoV-2

We downloaded the mutation file of 4,521 lines of SARS-CoV-2 as we described in Materials & Methods. After removing Indels and ambiguously sequenced bases, 3,169 unique point mutation sites are maintained. We investigated these ~3000 mutation sites. For a few sites with multiple alternative alleles, we are only interested in the major allele and minor allele. A basic hypothesis is that the mutations in RNA virus should be mainly introduced by the RNA replication error. As inferred from the DNA replication error profile (in higher organisms), the occurrence of transition should be slightly higher than the occurrence of transversion, and the mutation types should be "symmetric" in the profile.

We counted the occurrence of non-directional mutations between the major allele and the minor allele. Among the six possible combinations (A-C, A-G, A-T, C-G, C-T, G-T) between major allele and minor allele, 1,461 (46.1%) are C-T mismatches and 633 (20.0%) are A-G mismatches (Figure 2).

One may argue why we use non-directional mutations rather than the twelve types of directional mutations. As we mentioned in Introduction section, the mutation (such as cytidine-to-uridine deamination) could either take place in the reads or in the reference "genome" (Figure 1B), particularly for the RNA modification on RNA viruses.

If we simulate the mutation profile by assuming that transitions (A-G, C-T) are twice as likely as transversions (A-C, A-T, C-G, G-T) to take place, then theoretically we would obtain 1/8 A-C, 1/8 A-T, 1/8 C-G, 1/8 G-T, 1/4 A-G and 1/4 C-T. The observed number of C-T mismatches (1,461) is significantly higher than the simulated number (0.25*3169 = 792) using Chi-square test. It means that the number of real C-T SNPs might be overestimated by 1461/792 = 1.84 folds. The excessive C-T mismatches could not be explained by random mutations (introduced by RNA replication errors) alone, which indicates an alternative explanation of cytosine-to-uridine deamination.

The MAF (minor allele frequency) of C-T mismatches is significantly higher

The profile of mutation types exhibits non-random occurrence of C-T mismatches, suggesting potential cytosine-to-uridine deamination. We try to discover other aspects to reflect the difference between C-T and other types of mismatches.

The evolutionary selection strength is usually measured by MAF in population genetics. The allele information of the 4,521 SARS-CoV-2 lines provides a great opportunity for researchers to dig into the MAF spectrum. For the six types of mismatches, we calculated the mean MAF value of them. The MAF values of A-G and C-T are significantly higher than the MAF of the four transition types (Figure 3).

If one deems that all the mismatches are real “SNPs”, then the higher MAF should suggest more advantageous effects conferred by the mutations. However, there is no reason to believe that the A-G and C-T mutations are more advantageous than other

mutation types. Mutations are randomly distributed across the SARS-CoV-2 “genome” and most of them are neutral. Again, a plausible explanation is the A-G and C-T deamination systems in hosts. The occurrence of deamination on RNAs is definitely more prevalent than the RNA replication errors, so that they appear to have higher frequencies in the different SARS-CoV-2 lines. This result supports the possibility of the “deamination origin” of these mutations.

Discussion

Given the same set of mutation sites downloaded from the databases, different researchers might give different interpretations. The population geneticists and evolutionary biologists are willing to believe that all these mutations are SNPs. For virologists, especially those dealing with RNA viruses, they would doubt whether the observed mutations could be explained by other sources such as RNA modification systems in hosts.

Although we claimed that a rational way to distinguish the two mutation sources (RNA modifications or RNA replication errors) is to see the relative proportion of different mutation types, we should emphasize that this is only an inference but not direct evidence. We say that if the mutation profile is skewed, then it might be better explained by RNA modification.

Additionally, the MAF spectrum also rejects the “pure SNP origin” of these mutations because the MAF could differ among different genomic regions (such as neutral compared with non-neutral regions) but should not differ among different mutations types (such as transition compared with transversion). The unexpectedly higher MAF of A-G and C-T mismatches should be better explained by the deamination system. The occurrence of deamination on RNA should be orders of magnitude higher than the RNA replication error rates.

But we still do not have any decisive evidence to prove the “deamination origin” of the C-T mismatches. We think the potentially strong evidence could be the RNA-seq data isolated from different times points after virus infection. For example, if the proportion of C-T mismatches increases with the infection time, then the deamination hypothesis could be validated.

However, for the current medical purposes at this SARS-CoV-2 time, few people would be willing to conduct such an experiment. So, our hypothesis remains speculative. At this stage, we do not have the confidence to claim that the C-T mismatches are certainly caused by deamination, but we are confident to show that the skewed mutation profile could be better explained by deamination compared with the situation if one regards all the mutations as real “SNPs”. If one arbitrarily deems all the mutations to be SNPs and conducts the evolutionary analyses with them, then the results might be questionable.

Conclusion

The excessive C-T and A-G mismatches do not resemble the random mutation profile, and are more likely to be explained by the cytosine-to-uridine and adenosine-to-inosine deamination systems in hosts. It might be necessary to exclude these sites in the evolutionary analyses.

Acknowledgements

We thank all the people who contributed to the world in this SARS-CoV-2 time. We also thank our group members for their support to this work.

Author contributions

The corresponding author designed and supervised this research. All authors contributed to the construction and writing this article.

Summary points

- Mutation data from multiple SARS-CoV-2 lines are now available.
- We downloaded the mutations in ~4,500 lines of SARS-CoV-2.
- Over 3,000 unique point mutation sites in SARS-CoV-2 were maintained.
- C-T mismatches are excessive in the mutation profile.
- The occurrence of C-T mutations is significantly higher than randomness.

- C-T and A-G mismatches have higher MAF than other mutation types.
- C-T mutations are likely caused by cytidine-to-uridine deamination by the host.

References

1. Colson P, Lagier JC, Baudoin JP, Bou Khalil J, La Scola B, Raoult D. Ultrarapid diagnosis, microscope imaging, genome sequencing, and culture isolation of SARS-CoV-2. *Eur. J. Clin. Microbiol. Infect. Dis.* doi:10.1007/s10096-020-03869-w (2020).
2. Lu IN, Muller CP, He FQ. Applying next-generation sequencing to unravel the mutational landscape in viral quasispecies: A mini-review. *Virus Res.* doi:10.1016/j.virusres.2020.197963 197963 (2020).
3. Yadav PD, Potdar VA, Choudhary ML *et al*. Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J Med Res* doi:10.4103/ijmr.IJMR_663_20 (2020).
4. Licastro D, Rajasekharan S, Dal Monego S *et al*. Isolation and full-length genome characterization of SARS-CoV-2 from COVID-19 cases in Northern Italy. *J. Virol.* doi:10.1128/JVI.00543-20 (2020).
5. Caly L, Druce J, Roberts J *et al*. Isolation and rapid sharing of the 2019 novel coronavirus (SARS-CoV-2) from the first patient diagnosed with COVID-19 in Australia. *Med J Aust* doi:10.5694/mja2.50569 (2020).
6. Alonso-Blanco C, Andrade J, Becker C *et al*. 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. *Cell* 166(2), 481-491 (2016).
7. Kaul S, Koo HL, Jenkins J *et al*. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814), 796-815 (2000).
8. Korstanje C. The human genome project: understanding the role of inflammation in disease and disease prevention. *Curr Opin Investig Drugs* 7(11), 964-965 (2006).
9. Zhao Z, Li H, Wu X *et al*. Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol* 4 21 (2004).
10. Gonzales-Van Horn SR, Sarnow P. Making the Mark: The Role of Adenosine Modifications in the Life Cycle of RNA Viruses. *Cell Host Microbe* 21(6), 661-669 (2017).
11. Viehweger A, Krautwurst S, Lamkiewicz K *et al*. Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis. *Genome Res.* 29(9), 1545-1554 (2019).

Figure legends

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.

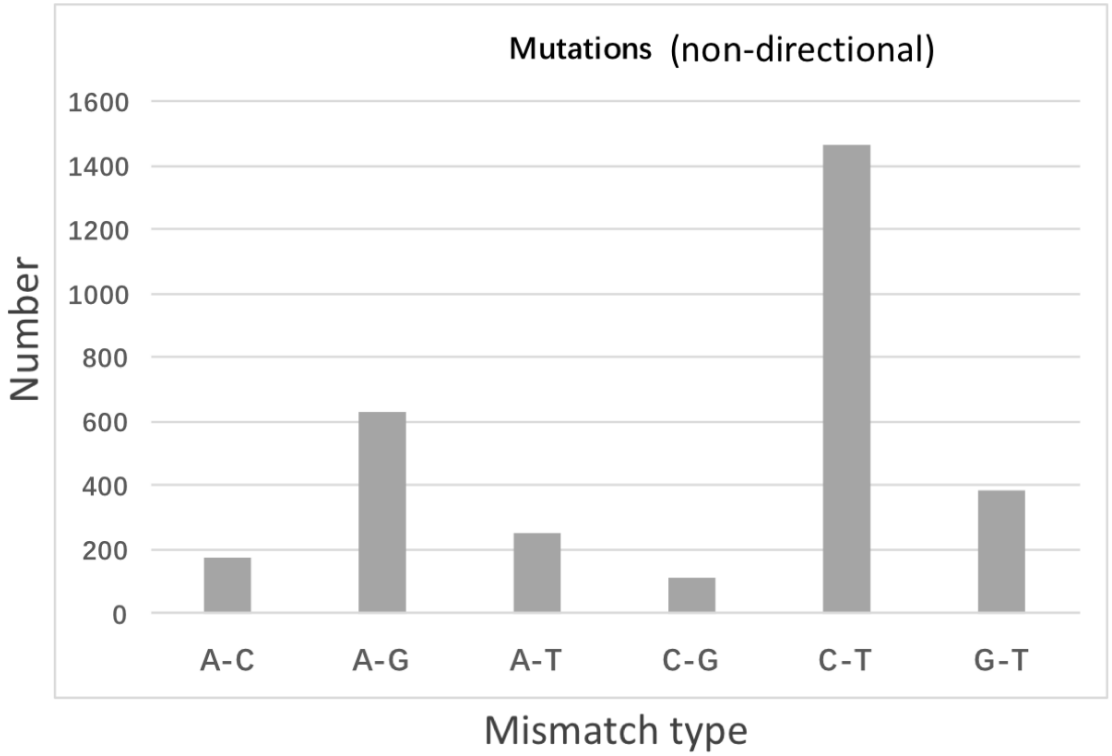

Figure 2. The numbers of mutation types (non-directional) among 4,521 lines of SARS-CoV-2.

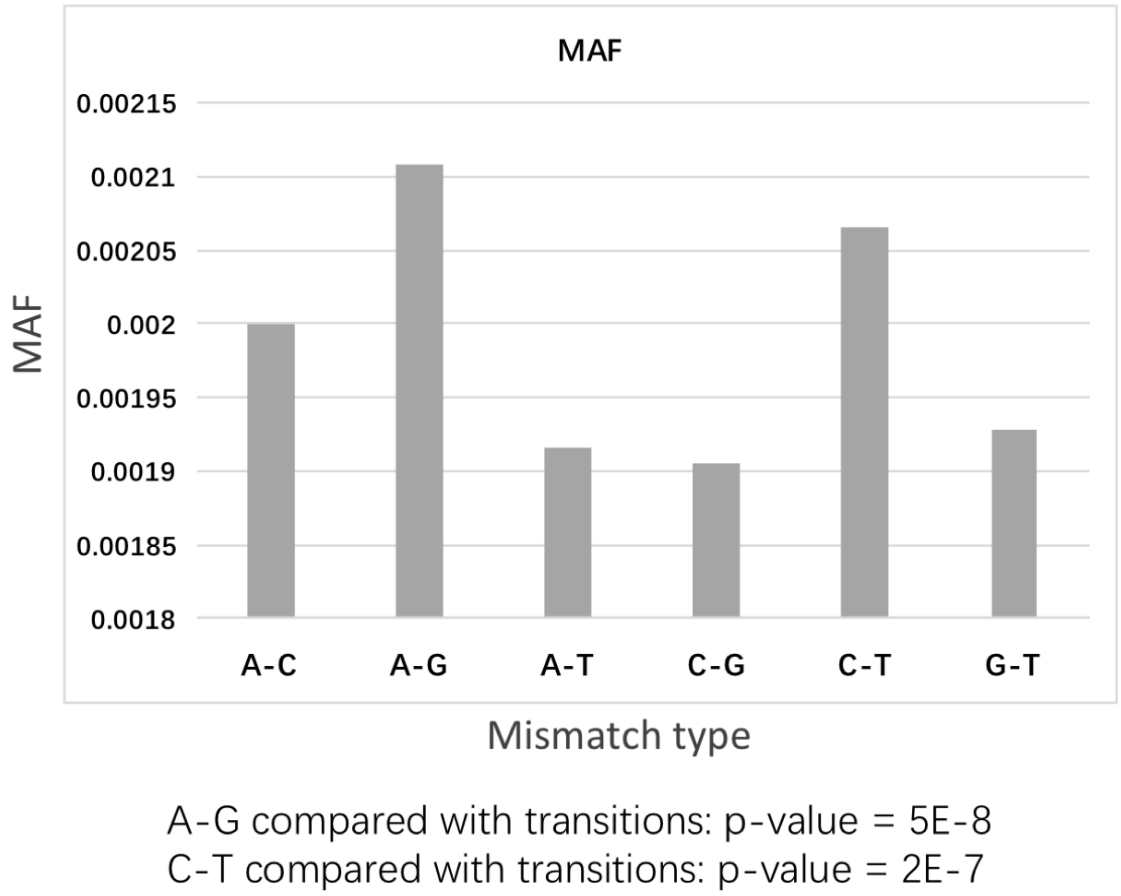

Figure 3. The mean MAF (minor allele frequency) of each mutation type (non-directional) among 4,521 lines of SARS-CoV-2. We used KS tests to calculate the p-values. Transition means A-C, A-T, C-G, and G-T. Transversions are A-G and C-T.