Editorial

For reprint orders, please contact: reprints@futuremedicine.com

Natural selection on synonymous mutations in SARS-CoV-2 and the impact on estimating divergence time

Yuanyuan Yu[1], Yan Li[2], Yu Dong[3], Xuekun Wang[2], Chunxiao Li[2] & Wenqing Jiang*[,4]

[1]Department of Anesthesiology, Qingdao Haici Hospital, Qingdao, Shandong, China
[2]Department of Cardiology, Qingdao Center Hospital, Qingdao, Shandong, China
[3]Department of Intervention, Qingdao Center Hospital, Qingdao, Shandong, China
[4]Department of Respiratory Diseases, Qingdao Haici Hospital, Qingdao, Shandong, China
*Author for correspondence: Tel.: +86 0532 8377 7856; qdhospit87@163.com

"We propose the potential approaches to correct the bias and estimate the genuine divergence time between SARS-CoV-2 and other related viruses."

Tweetable abstract: To adapt to human host environment, synonymous mutations in SARS-CoV-2 are shaped by tRNA selection, energy cost and RNA structure.

First draft submitted: 3 April 2021; Accepted for publication: 10 May 2021; Published online: 19 May 2021

Keywords: energy cost • natural selection • RNA structure • SARS-CoV-2 • synonymous • tRNA

Our knowledge of synonymous mutations

Synonymous mutations do not change the amino acid sequences. For a long time, synonymous mutations are regarded as neutral and silent mutations. In evolutionary biology, the estimation of divergence time between species requires a set of mutations with constant accumulation rate. The many deleterious missense mutations would be purged by purifying selection and thus are not suitable for measuring the divergence [1], whereas, synonymous mutations are free from natural selection and could provide a linear function to calculate the divergence time [2]. This notion has been well utilized to investigate the origin and evolution of SARS-CoV-2 by looking at the synonymous divergent sites between SARS-CoV-2 and RaTG13 [3].

However, in recent years, there is growing evidence demonstrating that synonymous mutations are not neutral. Researchers are getting aware that amino acid alterations (caused by missense mutations) are no longer the only feature to be subjected to natural selection. At molecular level, synonymous mutations could affect pre-mRNA splicing [4], tRNA availability [5], guanosine and cytidine (GC) content and codon usage [6], translational speed [7], protein extension [8], cost of energy and resources [9] and RNA structure/stability [10]. All these biological consequences of synonymous mutations may change the fitness and be exposed to natural selection. Beneficial and adaptive synonymous mutations are gradually getting fixed in a species while deleterious synonymous mutations are rapidly eliminated. In other words, the synonymous mutations do not accumulate with constant rate. This fact sheds doubt on the synonymous-based methodology of divergence estimation. Given the essentiality of clarifying the origin and evolution of SARS-CoV-2 under the current COVID-19 pandemic, more meticulous approaches should be considered to estimate the divergence.

In this article, we will first discuss the non-neutral property and the selection patterns on synonymous sites in SARS-CoV-2. We especially focus on the synonymous divergent sites between SARS-CoV-2 and RaTG13 because the latter is usually used as an outgroup sequence of SARS-CoV-2. Based on the many confounding factors disrupting the neutrality of synonymous mutations, we will try to propose a potential solution to the dilemma and wish to roughly estimate the genuine divergence time according to synonymous sites.

SARS-CoV-2 needs to adapt to the host tRNA pool for efficient translation

Although synonymous codons encode the same amino acid, they could have different types of cognate tRNAs that decode the particular codon. The term tRNA adaptation index quantitatively describes the tRNA availability among synonymous codons [5]. Higher tRNA adaptation index correlates with faster translational speed and thus should be advantageous [11]. In humans, the C/G-ending synonymous codons usually have higher tRNA availability than the A/T-ending synonymous codons. For SARS-CoV-2, in order to adapt to the host tRNA pool and achieve higher translation efficiency, the virus should adjust its synonymous codon usage toward more C/G-ending codons which are favored by human.

There are hundreds of synonymous divergent sites between SARS-CoV-2 and RaTG13, mainly contributed by the two longest genes *ORF1AB* (618 sites) and *S* (215 sites) [3]. The synonymous mutations from RaTG13 to SARS-CoV-2 would be advantageous if they switch A/T to C/G, while the synonymous mutations from C/G to A/T would be deleterious. Indeed, the former class of mutations are more frequently observed between RaTG13 and SARS-CoV-2. Since the host of RaTG13 is not human, the distinct synonymous codon divergence between SARS-CoV-2 and RaTG13 simply indicates the selection force acting on synonymous mutations in SARS-CoV-2 that prompts the virus to adapt to the tRNA pool in human hosts.

To date, we know that the synonymous mutations in SARS-CoV-2 are not neutral and are already skewed by natural selection and; therefore, the divergence time based on synonymous sites should be re-evaluated.

Energy cost also constrains the synonymous mutations

We have stressed the importance of considering tRNA availability and translation efficiency when judging the property of synonymous mutations in SARS-CoV-2. However, the cellular system is a balance between cost and efficiency. It is not worthy to slightly enhance the efficiency at extremely high costs. In order to rapidly proliferate, not only the efficiency but also the cost should be taken into account [9].

The biosynthesis of the four basic nucleotides consumes certain numbers of ATP molecules with the order of A > G > C > T [12]. Cheaper nucleotides (less ATP) are favored by natural selection. For synonymous codons, although C/G have the advantage of more efficient translation than A/T, they also cost more ATPs than T. Therefore, the selection pressure on synonymous mutations is a trade-off between energy cost and efficiency.

From the synonymous divergent sites between SARS-CoV-2 and RaTG13, it is hard to parse the selection on ATP cost because both species have the demand to lower the biosynthetic cost. However, from the polymorphic sites in SARS-CoV-2 population (derived mutations could be inferred from the outgroup RaTG13 sequence), one should observe higher derived allele frequencies (an indicator of positive selection) on the synonymous mutations that decrease the energy cost.

Synonymous mutations alter the GC content & RNA structure

RNA structure also affects translation elongation and initiation [13]. C–G base pairs are more stable than A–T base pairs. Therefore, synonymous mutations that increase the GC content are likely to reinforce the RNA structure. It is known that highly structured RNAs are not favorable for efficient translation [14], then the mutations from A/T to C/G would suffer from the disadvantage of this structural effect. Again, the trade-off between RNA structure and tRNA availability should be balanced. Nevertheless, it seems that the structured regions only consist of a small part of the entire transcriptome as there is still the global trend that the synonymous mutations increasing the GC content are selectively advantageous.

Extensive RNA modifications by the hosts

So far, we have discussed several biological features that are constrained by natural selection, making the synonymous mutations non-neutral. It reminds the researchers that the inference of divergence time from synonymous sites should be re-evaluated. However, this statement is based on the assumption that the synonymous mutations occur with constant rate (although, they accumulate with nonconstant rate due to natural selection). Next, we will show that even the occurrence of synonymous mutations does not conform to constant rates.

A broad range of species, from plants to animals, have the adenosine-to-inosine and cytidine-to-uridine deamination mechanisms to diversify their cellular RNAs [15]. When RNA viruses invade hosts, they frequently undergo RNA deamination by the host system, leading to A–G(I) and C–T(U) mismatches between the sequenced virus samples and the ancestral sequence. Intriguingly, the natural mutations in all organisms come from the DNA replication errors (for DNA organisms) or RNA reverse transcription/replication errors (for RNA viruses like

SARS-CoV-2). The natural mutation rate is usually 1E-8 per nucleotide per generation. In contrast, the RNA deamination rate could be higher for orders of magnitudes. Evidence already shows that there is extensive RNA deamination in SARS-CoV-2 sequences [16,17]. More specifically, 87% of the synonymous divergent sites between SARS-CoV-2 and RaTG13 could be explained by RNA deamination rather than natural mutation, leading to overestimated divergence time [3]. This observation suggests that the synonymous variants in SARS-CoV-2 come from two resources with completely different mutation rates (natural replication errors vs RNA deamination), and also warns us that the prevalent deamination events should be considered when estimating the divergence time.

Estimation of the genuine divergence time based on synonymous sites

We have shown that the synonymous mutations in SARS-CoV-2 do not occur under constant rate, and that they also undergo natural selection. Both features hinder the accurate estimation of divergence time based on synonymous sites.

We propose the potential approaches to correct the bias and estimate the genuine divergence time between SARS-CoV-2 and other related viruses.

- Avoid using the synonymous mutations that strongly alter the tRNA availability. For SARS-CoV-2, only keep the synonymous mutations between C and G or between A and T. These mutations are less affected by natural selection on tRNA availability.
- Discard synonymous mutations that dramatically change the ATP cost. As mentioned above, biosynthesis of adenosine costs most ATPs while thymidine costs least ATPs. Therefore, mutations between A and T are no longer considered.
- Discard A–G and C–T mutations that are potentially caused by deamination system. These two types of mutations are already excluded by the first criterion.

By this way, using the observed substitution on the remaining synonymous mutation sites and also considering the natural mutation rate, one could estimate the relatively accurate divergence time between SARS-CoV-2 and other viruses. We hope our summary and ideas could be interesting to the broad virology community as well as evolutionary biologists.

Acknowledgments

We thank the members in our group who have given suggestions to this article.

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

1. Holmes EC. Patterns of intra- and interhost nonsynonymous variation reveal strong purifying selection in dengue virus. *J. Virol.* 77(20), 11296–11298 (2003).
2. Smith JM, Smith NH. Synonymous nucleotide divergence: what is "saturation"? *Genetics* 142(3), 1033–1036 (1996).
3. Li Y, Yang XN, Wang N *et al.* The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive RNA modification. *Future Virol.* 15(6), 341–347 (2020).
4. Supek F, Minana B, Valcarcel J, Gabaldon T, Lehner B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156(6), 1324–1335 (2014).
5. Dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32(17), 5036–5044 (2004).
6. Li Y, Yang XN, Wang N *et al.* GC usage of SARS-CoV-2 genes might adapt to the environment of human lung expressed genes. *Mol. Genet. Genomics* 295(6), 1537–1546 (2020).
7. Hanson G, Coller J. Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* 19(1), 20–30 (2018).
8. Wang N, Wang D. Genome-wide transcriptome and translatome analyses reveal the role of protein extension and domestication in liver cancer oncogenesis. *Mol. Genet. Genomics* doi:10.1007/s00438-021-01766-1 (2021) (Epub ahead of print).

9. Zhao SF, Song SN, Qi Q, Lei W. Cost-efficiency tradeoff is optimized in various cancer types revealed by genome-wide analysis. *Mol. Genet. Genomics* doi:10.1007/s00438-020-01747-w (2021) (Epub ahead of print).
10. Chamary JV, Hurst LD. Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* 6(9), R75 (2005).
11. Sabi R, Tuller T. Modelling the efficiency of codon-tRNA interactions based on codon usage bias. *DNA Res.* 21(5), 511–526 (2014).
12. Chen WH, Lu G, Bork P, Hu S, Lercher MJ. Energy efficiency trade-offs drive nucleotide usage in transcribed regions. *Nat. Commun.* 7, 11334 (2016).
13. Wang Y, Gai Y, Li Y, Li C, Li Z, Wang X. SARS-CoV-2 has the advantage of competing the iMet-tRNAs with human hosts to allow efficient translation. *Mol. Genet. Genomics* doi:10.1007/s00438-020-01731-4 (2020) (Epub ahead of print).
14. Hall MN, Gabay J, Debarbouille M, Schwartz M. A role for mRNA secondary structure in the control of translation initiation. *Nature* 295(5850), 616–618 (1982).
15. Bass BL. RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem.* 71, 817–846 (2002).
16. Li Y, Yang XN, Wang N *et al.* Pros and cons of the application of evolutionary theories to the evolution of SARS-CoV-2. *Future Virol.* 15(6), 369–372 (2020).
17. Li Y, Yang X, Wang N *et al.* Mutation profile of over 4500 SARS-CoV-2 isolations reveals prevalent cytosine-to-uridine deamination on viral RNAs. *Future Microbiol.* 15(14), 1343–1352 (2020).