Should we apply evolutionary theories to the evolution of SARS-CoV-2 ?

Huizhi Feng[1]

1. Zhengzhou University, Zhengzhou, China

Declaration

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. Email to whom: huizhi. Tel: 86-0371-67783235. The corresponding author(s) supervised this article, and is (are) responsible for any questions or contacts.

Abstract

The outbreak of SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) has caused severe damage to the world. With the support of classic evolutionary theories and population genetics principles, many studies on the origin of SARS-CoV-2 have revealed encouraging results but meanwhile are still under debate. We are concerned with the validity of applying classic evolutionary theories and formula to the evolution of RNA viruses. We have raised several factors like the RNA replication feature and the RNA modification systems of the hosts, which might jeopardize the validity of the application of classic methods to analyze the SARS-CoV-2 data.

Keywords: SARS-CoV-2, evolution and origin, RNA virus, application, concerns

Background

The outbreak of SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) has caused severe damage to China, especially the Hubei province [1-3]. Recently, the cases in China are coming under control but the situation in some other countries has become exacerbated. It is urgent to find ways to control its transmission and cure the infected patients.

The emergence of papers on the evolution of SARS-CoV-2 is as fast as the spread of the virus itself. Based on genome sequencing followed by sequence alignment and sequence similarity analysis, researchers have characterized the evolutionary patterns of SARS-CoV-2 and postulated its origins. Theories of bat origin [4-6] and pangolin origin [7, 8] emerged, and even the snake plays an intermediate role in the transmission of SARS-CoV-2 [9]. Many other similar studies and results are not exhaustively listed here at all. In a word, it seems that the continuous change of theories on the evolution of SARS-CoV-2 even exceeds the evolution rate of the virus itself. For example, a recent paper by Tang *et al*. [10] has taken advantage of RNA-seq data. By traditional mutation calling pipeline, the authors discovered two lineages (L and S) of SARS-CoV-2 and claimed that the L type is more virulent than S type based on their frequency among the population. This might be a typical bioinformatic study that informs us of the scenario of SARS-CoV-2 population and evolution. The “snake” paper [9] also emphasized the important fact that viruses adapt to the host expression system.

As far as we know, the classic theories and principles of evolutionary biology, especially population genetics, are based on the central dogma, where DNA should be first transcribed to RNA and then produce proteins, and meanwhile the DNA replicates itself under a certain error rate. The validity of many formula/software is also based on multiple hidden hypotheses.

However, SARS-CoV-2 is a positive-strand RNA virus. It reproduces its own RNA by RNA replication. We doubt the validity of applying classic evolutionary theories and formula to the evolution of RNA viruses. In this article, we would raise and discuss several concerns regarding how the confounding factors would jeopardize the validity of the application of classic methods to analyze the SARS-CoV-2 data. We should emphasize that we neither criticize any studies, nor prove any ideas wrong. Instead, we aim to raise some questions and wish that these concerns could be further discussed by the broad community. We hope our concerns would contribute to the accurate identification of the origins of SARS-CoV-2.

Main text

The problem of RNA modification system in host cells

We have stated that SARS-CoV-2 is an RNA virus. The classic theories, principles and formula of evolutionary biology are based on the central dogma, which assumes the DNA-to-RNA-to-protein chain. The basic notion is that the mutations on DNA could be inherited rather than the modifications on RNA. However, for RNA viruses, their RNA is actually their genetic information. Whether the evolutionary principles could be applied to RNA viruses should be seriously debated. The host cells have multiple RNA modification systems/enzymes. The changes in viral RNA (by host cells) would permanently change its genetic information and be transmitted to the next "generation", which is similar to genetic mutations in higher organisms. Technically, one could not distinguish genetic mutations and RNA modifications from the RNA-sequencing data of RNA viruses. So, what is the point of detecting selection force based on the mutations in the virus population? These mutations maybe randomly created by the host cell's RNA modification systems.

The recent study by Tang *et al.* [10] claimed that the 17% divergence between SARS-CoV-2 and RaTG13 (a bat SARS-related coronavirus) is fourteen times larger than the divergence between human and chimpanzee. The authors concluded that only the neutral evolving sites should be considered rather than all different sites. Let us assume that both SARS-CoV-2 and RaTG13 undergo the RNA modification by host cells, and the modified viral RNA is inheritable, then their sequence (SARS-CoV-2 and RaTG13) could become

quite different within a short time scale. The divergence time is calculated as t = dS/2u. When dS might be largely contributed by the RNA modification system of host cells, this estimation could be inaccurate. In the dN and dS calculation, it is necessary to rule out any mismatch sites that might be produced by RNA modification. They should at least mention why they should or should not consider this factor. And then, the mutation rate "u", how to define "u"? Does "u" include the nucleotide changes conferred by the host's modification enzymes? Therefore, the authors' logic chain is questionable.

The "SNPs" and modified RNAs are technically indistinguishable. The software and algorithms only align the sequences but do not tell you whether the observed mismatch is a "SNP" or RNA modification site. This is a biological problem rather than technical problem and could not be solved by adjusting or improving the alignment parameters or filtering criteria. The sequence similarity could be largely, randomly, and arbitrarily skewed by the hosts' RNA modification systems. The observed divergence (or mismatch sites) may not really reflect the phylogeny of the viruses.

In our opinion, the RNA viruses should obey a different evolutionary theory. So far, the functional experiments are more important and reliable than the pure evolutionary analyses in this case of RNA virus. When traditional evolutionary principles are jeopardized by additional mutation forces, the functional experiments work well as they did in the past.

Problems raised from the RNA replication process

Apart from being modified by the host RNA modification systems, there are other concerns about whether the evolutionary theories could be applied to the RNA viruses like SARS-CoV-2.

For cellular organisms, the DNA mutations are majorly introduced during the DNA replication process. The mutation rate is largely connected with the fidelity of DNA replication. The next step is the natural selection force acting on these mutations, after which the deleterious mutations are purged and those beneficial mutations are maintained. However, RNA viruses either undergo the reverse-transcription process (like HIV) or the RNA replication process (like SARS-CoV-2). For RNA viruses, every newly transcribed RNA molecule is a potential offspring of the original virus. The mismatches introduced during reverse-transcription or RNA replication would be maintained and kept in the offspring. Before applying the evolutionary formula to RNA viruses, one should state whether RNA replication has similar mismatching rates as DNA replication. Intuitively, DNA-DNA pairing (DNA replication), DNA-RNA pairing (transcription) and RNA-RNA pairing (such as RNA replication) should have different mismatching rates. Thus, when applying theories to SARS-CoV-2, should the authors consider the potentially different mutation rates during the reverse-transcription or RNA replication processes? Take the paper by Tang *et al.* [10] for instance. What exactly does mutation rate “u” refer to? Even the problem of technically indistinguishable RNA modification and the “de novo” RNA

mutation is not mentioned by the authors at all, let alone the mutations introduced during the RNA replication process. At least, the authors could briefly introduce the reproduction mode of SARS-CoV-2 rather than “mechanically” apply the formula to an organism which they are not familiar with. For the Tang et al paper https://doi.org/10.1093/nsr/nwaa036, the worse thing is that, given the well-known sexual affairs between the corresponding author Jian Lu and the co-author Yirong Wang, the paper became less reliable as we may guess there are other non-scientific purposes of publishing the paper on National Science Review. It looks strange to see amateurs publishing in the virology field. The mutation rate was not clearly clarified in that paper, but from other papers on virology, we can learn the mechanisms of viral mutation and understand how the host deamination systems would dramatically affect the mutation picture of RNA viruses. The extent of potential overestimation of the mutation number could be roughly inferred.

For HIV, the RNA deamination rate is > 40 folds higher than the reverse-transcription error rate (Rafael Sanjuan and Pilar Domingo-Calap, 2016, DOI 10.1007/s00018-016-2299-6). The RT error rate (for HIV) and the RNA replication error rate (for SARS-CoV-2) are what we commonly understood as the mutation rate. If we use this 40-fold as an estimation, then the number of real mutations in the SARS-CoV-2 populations is 1/40 of the mutations claimed by Tang et al (at least for the mutation sites of potential deamination).

Conclusion

We are concerned with the validity of applying classic evolutionary theories and formula to the evolution of RNA viruses. We have raised several confounding factors like the RNA replication process and the RNA modification systems of the hosts, which might jeopardize the validity of the application of classic methods to analyze the SARS-CoV-2 data. We hope our concerns could be considered by the broad community and could contribute to the accurate identification of the origins of SARS-CoV-2.

Acknowledgements

We thank the members in our group. We thank the recent studies on coronavirus which have enlightened us. We thank the medical workers saving lives of COVID-19 patients. At this SARS-CoV-2 time, we appeal that more studies on experimental biology should be conducted. The evolutionary studies have limited value in this war against SARS-CoV-2 (apart from just getting a publication). Then, after the Hongwei Guo and Can Xie events, the well-known Monday night mating between Jian Lu and Yirong Wang (authors of the Tang et al 2020 paper https://doi.org/10.1093/nsr/nwaa036) further made the related evolutionary studies less reliable. Compared to the bad effects of these affairs, other behaviors like plagiarism seem inconsequential. Having been said this, we admire the experimental biologists working hard to discover the antidote to COVID-19. The medical workers saving other people's lives are really great at this SARS-CoV-2 time. For the scientific field, we thank the people devoted in the experimental development of antidote against SARS-CoV-2.

Author contributions

The corresponding author designed and supervised this research.

References

1. Cowling BJ, Leung GM. Epidemiological research priorities for public health control of the ongoing global novel coronavirus (2019-nCoV) outbreak. *Euro Surveill.* 25(6), 2000110 (2020).
2. Hui DS, E IA, Madani TA *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.* 91, 264-266 (2020).
3. Wang C, Horby PW, Hayden FG, Gao GF. A novel coronavirus outbreak of global health concern. *Lancet.* 395(10223), 470-473 (2020).
4. Li X, Song Y, Wong G, Cui J. Bat origin of a new human coronavirus: there and back again. *Sci. China Life Sci.* 63, 461–462 (2020).
5. Xu X, Chen P, Wang J *et al.* Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. *Sci. China Life Sci.* 63, 457–460 (2020).
6. Benvenuto D, Giovanetti M, Ciccozzi A, Spoto S, Angeletti S, Ciccozzi M. The 2019-new coronavirus epidemic: Evidence for virus evolution. *J. Med. Virol.* doi:10.1002/jmv.25688 (2020).
7. Xiao K, Zhai J, Feng Y *et al.* Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *bioRxiv* doi:10.1101/2020.02.17.951335 2020.2002.2017.951335 (2020).
8. Lam TT-Y, Shum MH-H, Zhu H-C *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv.* doi.org/10.1101/2020.02.17.951335 (2020).
9. Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. *J. Med. Virol.* 92(4), 433-440 (2020).
10. Tang X, Wu C, Li X *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review.* nwaa036 (2020).