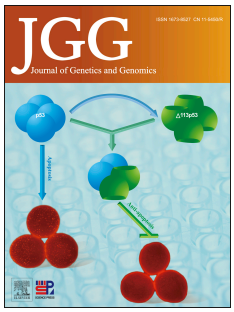
Journal Pre-proof



Pinpointing the animal origins of SARS-CoV-2: a genomic approach

Shilei Zhao, Yali Hou, Xiaolong Zhang, Alice Hughes, Na Liu, Minsheng Peng, Qihui Wang, Yongbiao Xue, Hua Chen

PII: S1673-8527(22)00153-9

DOI: https://doi.org/10.1016/j.jgg.2022.05.002

Reference: JGG 1080

To appear in: *Journal of Genetics and Genomics*

Received Date: 22 April 2022

Revised Date: 11 May 2022

Accepted Date: 12 May 2022

Please cite this article as: Zhao, S., Hou, Y., Zhang, X., Hughes, A., Liu, N., Peng, M., Wang, Q., Xue, Y., Chen, H., Pinpointing the animal origins of SARS-CoV-2: a genomic approach, *Journal of Genetics and Genomics*, https://doi.org/10.1016/j.jgg.2022.05.002.

Pinpointing the animal origins of SARS-CoV-2: a genomic approach

Understanding the origins of zoonoses is critical to developing the means to prevent zoonotic spillover into the future. Like 60% of emerging diseases in humans (Jones et al., 2008), SARS-CoV-2 is likely to have zoonotic origins with possible immediate hosts (Holmes et al., 2021; Wang et al., 2021). Determination of the origin of SARS-CoV-2, *i.e.*, when, where, and how it emerges in humans through possible zoonotic transfer, facilitates prevention of the emergence and establishment of new zoonotic diseases. Yet, so far, whilst many similar CoVs have been detected in wild animals (especially bats), the SARS-CoV-2 progenitor has not been identified, and the closest wild hosts still host viruses estimated to have diverged from SARS-CoV-2 decades ago (Holmes et al., 2021). There have been two controversial speculations on the origin of SARS-CoV-2, the natural origin and the laboratory leak hypotheses (Burki, 2020). In understanding the probability of either theory, we should consider patterns from former epidemics, which all showed immediate spillover from wildlife, or livestock, and the similarity of coronaviruses found in wildlife in the region. Based on these, addressing the gaps which continue to see speculation is important:

1) Despite of the uncertainty of the SARS-CoV-2 progenitor, increasingly high numbers of coronaviruses with relatively high similarity to SARS-CoV-2 genome sequences have been constantly isolated from nature reservoirs of bats or pangolins in Asian countries, including China, Japan, Cambodia, Thailand and Laos (Lam et al., 2020; Xiao et al., 2020; Zhou et al., 2020, 2021; Temmam et al., 2022;). Besides of RaTG13 that shares 96.3% of genome sequences with SARS-CoV-2 (Zhou et al., 2020), another more closely related coronaviruses containing BANAL-20-52 and BANAL-20-103 were isolated from *Rhinolophus malayanus* and *R. pusilus*, respectively, in Vientiane Province in northern Laos, presenting 96.8% sequence identity to SARS-CoV-2 and even only one or two residual divergence in receptor-binding domain

(Temmam et al., 2022). In fact, there is a high diversity of as yet undiscovered CoVs in wildlife across the regions.

2) By comparing the mutation signatures across the currently sampled SARS-CoV-2-related coronaviruses that are heavily shaped by hosts, SARS-CoV-2 genome in infected human cells shows a similar mutation spectrum to a naturally evolved viral genome (RaTG13) in bat cells, highlighting its natural origin (Shan et al., 2021; Deng et al., 2022).

3) The unique genomic signature of SARS-CoV-2, *i.e.*, the furin sites which boot infectivity of SARS-CoV-2 in human, could occur naturally through evolution by analyzing temporally collected viral genomic data (Holmes et al., 2021), and is not unusual in other coronaviruses which have been contacted by humans.

The debate on natural and laboratory leak origin has just subsided (though conspiracy theories continue to distract attention from scientific evidence), recently, two studies (posted as preprints and still under peer review) (Pekar et al., 2022; Worobey et al., 2022) simultaneously take the Huanan Seafood Market in Wuhan as the origin of the COVID-19 pandemic for granted. By performing spatial analyses of the locations of the earliest COVID-19 cases in Wuhan, Worobey et al. (2022) claimed the Market as the spillover site (epicenter) of the pandemic, where the infected animals may introduce virus into humans via wildlife trade. By investigating the genomic diversity of SARS-CoV-2 isolated from early cases, Pekar et al. (2022) modeled the genomic diversity using epidemic simulations, and deduced that there were at least two separate zoonotic events into humans (*i.e.*, lineages A and B) in the Market. These studies have clear limitations:

1) Obfuscating the epidemic outbreak place (P_o) and the origin (*i.e.*, spillover) place (P_s). Worobey et al. (2022) can only indicate that the early cases represented an infection cluster centering around the Market, which is totally different from the P_o. P_o could be far away from P_s based on the lessons learned from viruses like HIV (Ruan et al., 2021; Wang et al., 2021).

2) Overstating conclusions based on limited data and unrealistic simulations. Pekar et al. (2022) compared the phylodynamic patterns of the early-sampled viral sequences to those of epidemic simulations under various scenarios using coalescent process. In the coalescent process of their simulations, they assumed that viruses spread and evolve without population structure, which is inconsistent with viral epidemic processes with extensive clustered infections, founder effects, and sampling bias (Liu et al., 2020). Furthermore, they cannot exclude alternative scenarios that the local outbreak can be caused by asymptomatic infections, infected external traders, travelers or cold-chain transmission.

3) No definitive evidence about what type of animal might have harbored the virus before it spreads to humans. The joint WHO-Chinese study reported they didn't identify immediate host animals related to the Market. Gao et al. (2022) tested 1380 samples from animals and environments related to the Market, finding that 73 environmental samples but no animals themselves were positive for SARS-CoV-2.

Analyzing the genetic diversity of early-stage SARS-CoV-2 genomes may provide limited information on the origin of SARS-CoV-2. It is necessary to investigate worldwide to identify the SARS-CoV-2 progenitor in any of the potential animal reservoirs or intermediate hosts, and as with SARS and MERS to determine the pathway from reservoir hosts to humans.

We make inference on the location of SARS-CoV-2 progenitor based on the information of both sequence similarity and isolated geographical distribution of the currently existing SARS-CoV-2-related sarbecovirus genomes with at least 75% sequence identity compared to SARS-CoV-2. Following the isolation-by-distance theory in population genetics, which posits that genetic differentiation among individuals increases as geographical distances increases (Wright, 1943), we assume that the genomic similarity among SARS-CoV-2-related coronaviruses is generally proportional to their geographical distance. We further assume the spatial decaying rates of the virus genome similarity are in uniform without being affected by

geographical landform, which are additionally allowed to be varied along the x- and y-axis, rotating from the North θ clockwise. Let the sampling location of virus sample i be x_i (longitude), y_i (latitude). Suppose SARS-CoV-2 originates in location (x_0, y_0), which is an unknown parameter to be inferred. We denote sequence similarity between virus i and SARS-CoV-2 as s_i. The virus genomic similarity decays in different rates v and kv along x- and y- axis, rotating θ clockwise from north. Function f(A, B) measures the geographical distance between locations A and B; and $f(\cdot)v$ predicts the expected sequence divergence between SARS-CoV-2 progenitor and the sampled sequences as a function of their geographical locations. We thus defined the loss function as

$$L(\kappa,\theta,v,x_0,y_0)=\sum_{i=1}^{n}\left[f\left(T\begin{bmatrix}x_i\\y_i\end{bmatrix},T\begin{bmatrix}x_0\\y_0\end{bmatrix}\right)v-(1-s_i)\right]^2 \quad (1)$$

where $T=\begin{pmatrix}\kappa & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{pmatrix}$.

The loss function measures the difference between the predicted and the observed sequence similarity to SARS-CoV-2 as a function of geographical locations of the existing SARS-CoV-2 related coronaviruses isolated from bats. By minimizing the loss function using interior point method in the fmincon function in MATLAB, we profile the origin location and spatial distribution of sequence similarity, which reflects the probability of finding the SARS-CoV-2 progenitor. It shows that the highest density area in the fitted f function mainly covers Indochina Peninsula, encompassing Laos, Thailand, and Cambodia (Fig. 1A). The most probable location of the SARS-CoV-2 progenitor was inferred to be centered to 16.0986°N and 104.0617°E, which was located in Thailand in Southeast Asia (blue point, Fig. 1A), though further sequence data from other parts of the region may shift this locality. Since receptor binding domain (RBD) in SARS-CoV-2, the key structure that binds to the ACE2 receptor to enter host cell and determines the host range, shows strong signature of recombination (Temmam et al., 2022), RBD similarity might be the key factor for SARS-CoV-2

progenitor to acquire the capability of efficiently infecting human. Therefore, we also conducted similar analysis for the RBD and non-RBD regions of the samples, respectively. Both the inferred locations are centered to 16.4332/16.2975°N and 103.9537/104.0823°E, very close to the predicted origin location of the whole genome sequences (Fig. 1B). This indicates that recombination events of different virus lineages carrying the backbone sequences and the RBD regions respectively may occur somewhere close to the origin place, contributing to the mosaic genome of SARS-CoV-2 progenitor.

We thus highlight that the most probable origin of SARS-CoV-2 might be pinpointed in Southeast Asia encompassing Laos, Thailand, Cambodia, and neighboring countries. It is important to note that all these countries have found coronaviruses with genomic sequences highly similar to SARS-CoV-2, and given the high diversity of Rhinolophids in forests and caves across the region, further surveys are likely to detect closer relatives. There are multiple lines of evidence supporting this finding. SARS-CoV-2 neutralizing antibodies were detected in Thai cave bats and a pangolin at a wildlife checkpoint in Southern Thailand (Wacharapluesadee et al., 2021). The ecological distribution area of Rhinolophus species, the most likely natural reservoirs of SARS-CoV-2, primarily covers the southern portion of the Eurasian continent, extending from South Laos and Vietnam in Southeast Asia to southern China (Wang et al., 2021; Zhou et al., 2021). We are aware of the fact that the samples of SARS-CoV-2 related coronaviruses are still limited, and most of them are from Southeast Asia and East Asia; furthermore, some of the sequences are from different species (*e.g., R. pusillus*), relying on the assumption that SARS-CoV-2 related viruses are shared among different bat host species locally. Our analysis and conclusions are inevitably affected by limited and potentially biased sample collection, and the assumptions. These preliminary results highly suggest that expanding sample collection of SARS-CoV-2 related coronaviruses will significantly help pinpoint the origin of SARS-CoV-2, and in turn help us trace the path to zoonotic spillover to provide a basis to prevent future emerging infectious diseases.

Acknowledgments

The work was supported by the Key Program of Chinese Academy of Sciences (KJZD-SW-L14) and the National Key R&D Program of China (Grant No. 2021YFC0863400, 2021YFC2301305, and 2020YFC0847000).

References

Burki, T., 2020. The origin of SARS-CoV-2. Lancet Infect. Dis. 20, 1018-1019.

Deng, S., Xing, K., He, X., 2022. Mutation signatures inform the natural host of SARS-CoV-2. Natl. Sci. Rev. 9, nwab220.

Gao, G., Liu, W., Liu, P., Lei, W., Jia, Z., He, X., Liu, L.-L., Shi, W., Tan, Y., Zou, S., et al., 2022. Surveillance of SARS-CoV-2 in the environment and animal samples of the Huanan Seafood Market. https://doi.org/10.21203/rs.3.rs-1370392/v1.

Holmes, E.C., Goldstein, S.A., Rasmussen, A.L., Robertson, D.L., Crits-Christoph, A., Wertheim, J.O., Anthony, S.J., Barclay, W.S., Boni, M.F., Doherty, P.C., et al., 2021. The origins of SARS-CoV-2: A critical review. Cell 184, 4848-4856.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., Daszak, P., 2008. Global trends in emerging infectious diseases. Nature 451, 990-993.

Lam, T.T., Jia, N., Zhang, Y.W., Shum, M.H., Jiang, J.F., Zhu, H.C., Tong, Y.G., Shi, Y.X., Ni, X.B., Liao, Y.S., et al., 2020. Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. Nature 583, 282-285.

Liu, Q., Zhao, S., Shi, C.M., Song, S., Zhu, S., Su, Y., Zhao, W., Li, M., Bao, Y., Xue, Y., et al., 2020. Population Genetics of SARS-CoV-2: Disentangling effects of sampling bias and infection clusters. Genomics Proteomics Bioinformatics 18, 640-647.

Pekar, J.E., Magee, A., Parker, E., Moshiri, N., Izhikevich, K., Havens, J.L., Gangavarapu, K., Malpica Serrano, L.M., Crits-Christoph, A., Matteson, N.L., et al., 2022. SARS-CoV-2 emergence very likely resulted from at least two zoonotic events. https://doi.org/10.5281/zenodo.6291628.

Ruan, Y., Wen, H., He, X., Wu, C.I., 2021. A theoretical exploration of the origin and early evolution of a pandemic. Sci. Bull. 66, 1022-1029.

Shan, K.J., Wei, C., Wang, Y., Huan, Q., Qian, W., 2021. Host-specific asymmetric accumulation of mutation types reveals that the origin of SARS-CoV-2 is consistent with a natural process. Innovation 2, 100159.

Temmam, S., Vongphayloth, K., Baquero, E., Munier, S., Bonomi, M., Regnault, B., Douangboubpha, B., Karami, Y., Chretien, D., Sanamxay, D., et al., 2022. Bat coronaviruses related to SARS-CoV-2 and infectious for human cells. Nature 604, 330-336.

Wacharapluesadee, S., Tan, C.W., Maneeorn, P., Duengkae, P., Zhu, F., Joyjinda, Y., Kaewpom, T., Chia, W.N., Ampoot, W., Lim, B.L., et al., 2021. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. Nat. Commun. 12, 972.

Wang, Q., Chen, H., Shi, Y., Hughes, A.C., Liu, W.J., Jiang, J., Gao, G.F., Xue, Y., Tong, Y., 2021. Tracing the origins of SARS-CoV-2: lessons learned from the past. Cell. Res. 31, 1139-1141.

Worobey, M., Levy, J.I., Serrano, L.M.M., Crits-Christoph, A., Pekar, J.E., Goldstein, S.A., Rasmussen, A.L., Kraemer, M.U.G., Newman, C., Koopmans, M.P.G., et al., 2022. The Huanan market was the epicenter of SARS-CoV-2 emergence. https://doi.org/10.5281/zenodo.6299116.

Wright, S., 1943. Isolation by Distance. Genetics 28, 114-138.

Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J.J., Li, N., Guo, Y., Li, X., Shen, X., et al., 2020. Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. Nature 583, 286-289.

Zhou, H., Ji, J., Chen, X., Bi, Y., Li, J., Wang, Q., Hu, T., Song, H., Zhao, R., Chen, Y., et al., 2021. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. Cell 184, 4380-4391.

Zhou, P., Yang, X.L., Wang, X.G., Hu, B., Zhang, L., Zhang, W., Si, H.R., Zhu, Y., Li, B., Huang, C.L., et al., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579, 270-273.

Shilei Zhao[1], Yali Hou[1], Xiaolong Zhang[1],
Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China
University of Chinese Academy of Sciences, Beijing 100049, China

Alice Hughes,
University of Chinese Academy of Sciences, Beijing 100049, China
Center for Integrative Conservation, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla, Yunnan 666303, China

Na Liu,
National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China

Minsheng Peng,
State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650201, China
KIZ/CUHK Joint Laboratory of Bioresources and Molecular Research in Common Diseases, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650201, China

Qihui Wang,
University of Chinese Academy of Sciences, Beijing 100049, China
CAS Key Laboratory of Pathogen Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

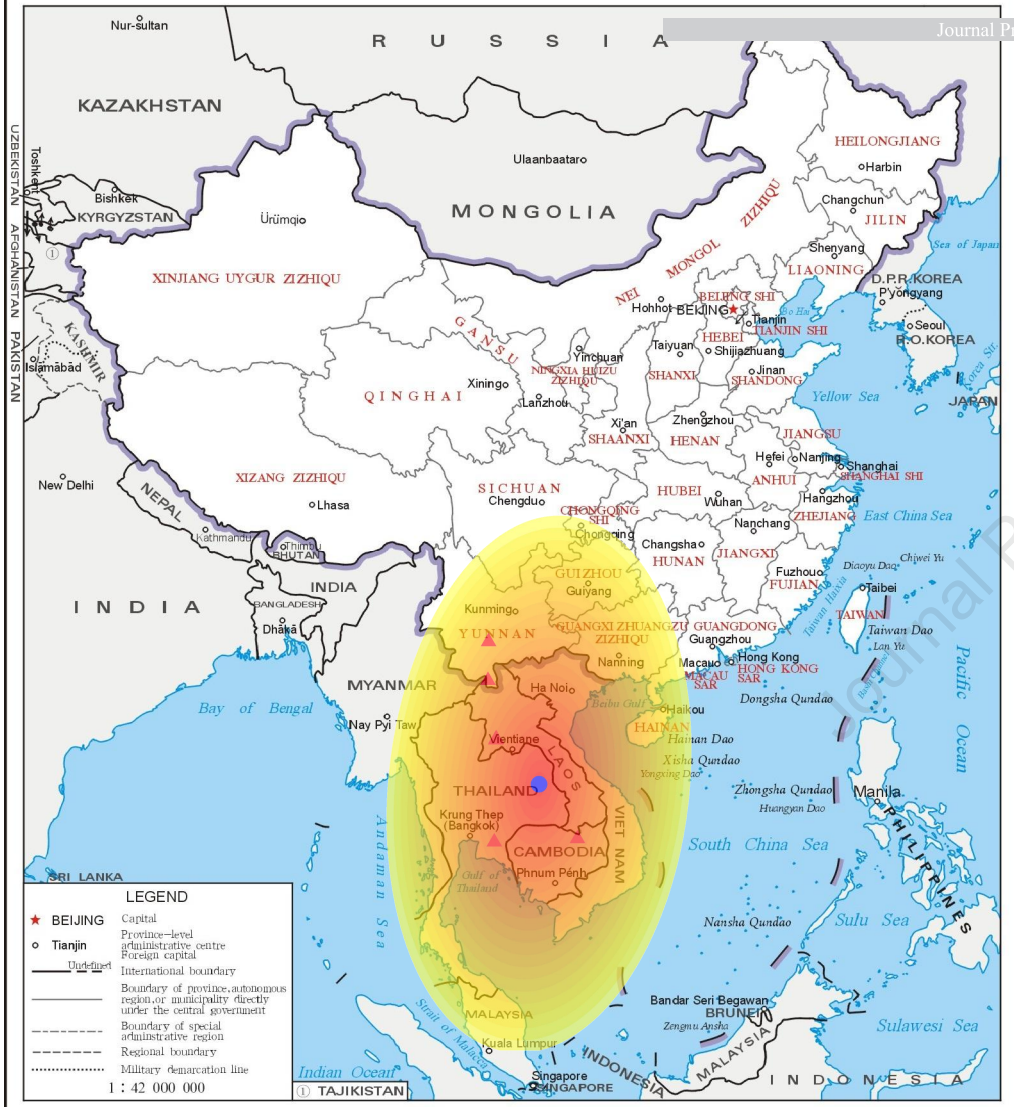
Yongbiao Xue[*],
Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China
University of Chinese Academy of Sciences, Beijing 100049, China

Hua Chen[*]
Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China
University of Chinese Academy of Sciences, Beijing 100049, China
Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, Yunnan 650223, China

[1] These authors contributed equally to this work.

** Corresponding authors.*
Email addresses: ybxue@big.ac.cn (Y. Xue), chenh@big.ac.cn (H. Chen)

Fig. 1. The inferred locations of the SARS-CoV-2 progenitor. **A**: Using the whole genome sequence. The blue dot is the deduced most probable location. **B**: Using receptor binding domain (RBD, hot color map) and non-RBD sequences (cold color map), respectively. The blue dot and green plus are the deduced most probable locations. Genome identity decreases to 78% with distance to the probable locations, covered by the color map. The red triangles denote the viral sampling locations, containing 15 SARS-CoV-2-related sarbecoviruses with at least 75% sequence identity compared to SARS-CoV-2. The viruses consist of RacCS203 in Thailand, RaTG13, PrC31, RmYN02, RsYN04, RmYN05, RpYN06 and RmYN08 in southern border of Yunnan province, China, BANAL-52, BANAL-103, BANAL-116, BANAL-236 and BANAL-247 in Northern Laos, and RshSTT182/200 in Cambodia. The map in the figure is downloaded from http://bzdt.ch.mnr.gov.cn with censorship number of GS(2019)1656.



A

B