SCIENCE CHINA
Life Sciences

•INSIGHT•

June 2023 Vol.66 No.6: 1440–1443
https://doi.org/10.1007/s11427-023-2294-5

SARS-CoV-2 shows a much earlier divergence in the world than in the Chinese mainland

Chaoyuan Cheng[1] & Zhibin Zhang[1,2*]

[1]*State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China;*
[2]*CAS Center for Excellence in Biotic Interactions, University of Chinese Academy of Sciences, Beijing 100049, China*

Received January 2, 2023; accepted February 12, 2023; published online March 7, 2023

Citation: Cheng, C., and Zhang, Z. (2023). SARS-CoV-2 shows a much earlier divergence in the world than in the Chinese mainland. Sci China Life Sci 66, 1440–1443. https://doi.org/10.1007/s11427-023-2294-5

Currently, the origin and early transmission pattern of SARS-CoV-2 remain unknown. The Time to the Most Recent Common Ancestor (*TMRCA*) of SARS-CoV-2 has been suggested to be mid-October or November of 2019 (Pekar et al., 2021; van Dorp et al., 2020) by using a small number of early samples which may not cover full lineages of its evolutionary tree. Available methods of estimating *TMRCA* of viruses (Drummond et al., 2003) are often time-consuming and thus constrained by using big data.

By using SARS-CoV-2 variation data of 255,039 genome sequences during the first 13 months since December 24, 2019 from the China National Center for Bioinformation (CNCB), we estimated the divergent time (*DT*) of SARS-CoV-2 of six continents and the world with a Wuhan reference collected on December 31, 2019 (GenBank accession number: MN908947.3) (Figure 1; Figures S1, S2, and Data S1 in Supporting Information). *DT* is a measure of the divergent time of samples from a region with the Wuhan reference based on a cumulative number of nucleotide substitutions (Figure S3 in Supporting Information). *DT* is not the *TMRCA* of a region unless its own reference is used, and *DT* of a region $\leq$*TMRCA* of all regions. The cumulative number of substitutions (M_t) of each sequence at time t and its variance within a time window (V_t^*, variance of M_t) were calculated using the Wuhan reference. First, by using regression models (r*DT* method) between M_t or V_t^*, and sampling time (t), we estimated T_0 (DT=$-T_0/2$), M_0 (M_t on December 31, 2019), substitution rate (a), and substitution variance rate (b) of SARS-CoV-2 in six continents. Second, by using the mean substitution rate, we estimated the i*DT* of each sequence, mean *DT* of all sequences, and some unexpected extreme high-substitution sequences of six continents and the world (i*DT* method).

Daily M_t of all six continents shows a significant linear increase over time (t) (Figure 1H; Table S1 in Supporting Information), indicating the evolution of SARS-CoV-2 follows the prediction of Molecular Clock Theory (Figures S1 and S2 in Supporting Information). The divergent time (r*DT*) was estimated by using regression models: M_t=M_0+a×t and T_0=$-M_0/a$ (rDT=$-T_0/2$) (Table S1 in Supporting Information). The mean r*DT* and M_0 of the six continents are 20.6 days (95%CI: 6.2, 35) and 2.0 (95%CI: 0.6, 3.4), which all are significantly different from zero (P<0.05, t-test), suggesting the divergence should start about 3 weeks in average in the world before December 31, 2019. The ranking order of mean r*DT* (days) is: South America (36.2)>North America (32.6)>Oceania (21.2)>Africa (18.4)>Asia (17.8)>Europe (2.7). r*DT* of the Chinese mainland is 9.1 days. The mean a of six continents is 0.0512 (95%CI: 0.044, 0.058) (i.e., 1.536 substitution per month) with the ranking order of Europe (0.0630)>Oceania (0.0526)>Asia (0.0520)>South America (0.0480)>North America (0.0452)>Africa (0.0462). a of the

*Corresponding author (email: zhangzb@ioz.ac.cn)

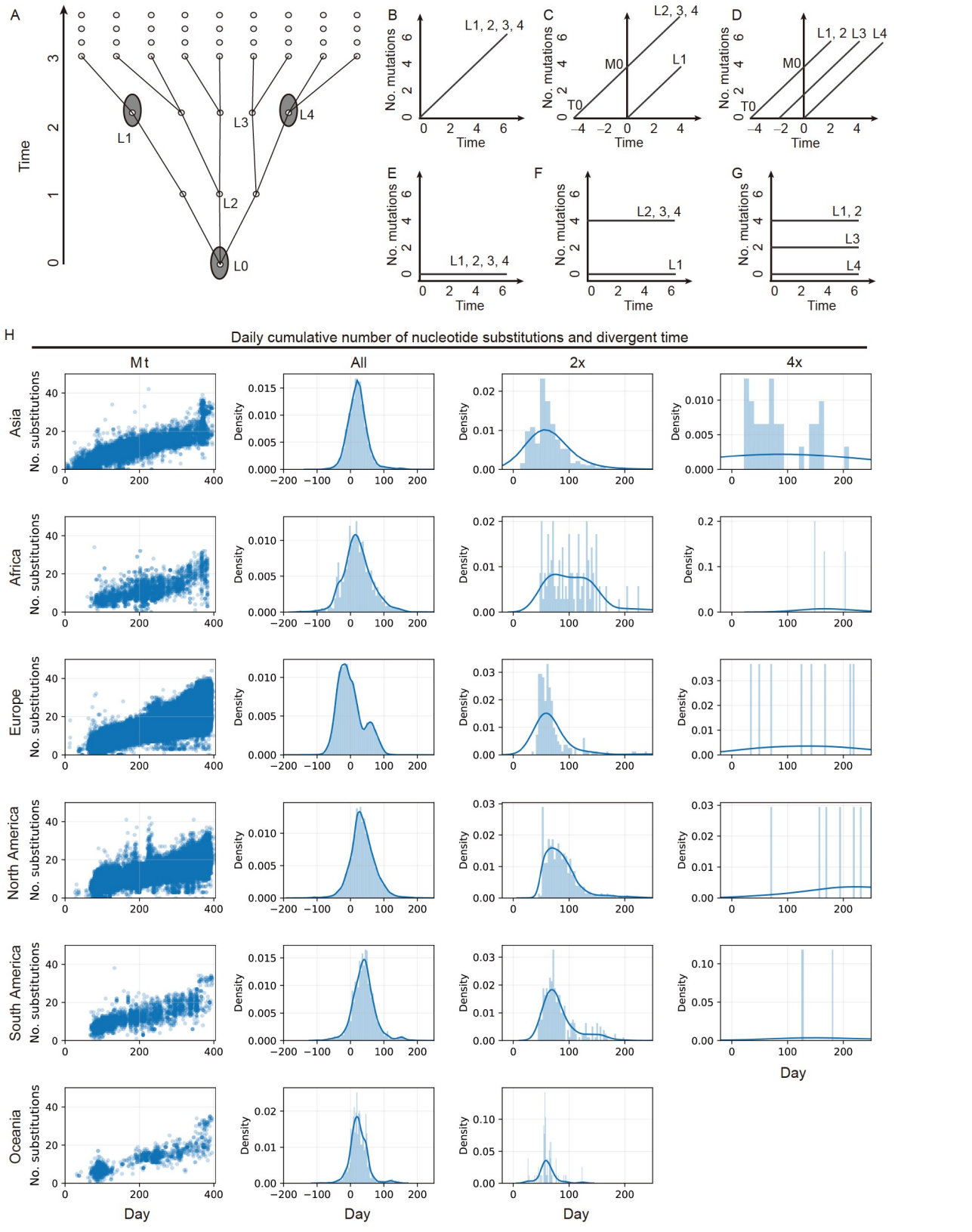

Figure 1 Illustrations of estimating T_0 (DT=$-T_0/2$) and M_0 by using regression models (B, C, D) and the deviated daily cumulative number of mutation (dM_t) from the expected values based on the mean substitution rate (E, F, G) when the reference (grey circles) is located in the ancestor position (B, E) or non-ancestor position (C, D, F, G) in the evolutionary tree (A). H, Daily cumulative number of nucleotide substitutions (M_t) and divergent time (iDT) for SARS-CoV-2 samples of all, M_t>2×eM_t and >4×eM_t of six continents since December 24, 2019 (t=0). eM_t, the expected M_t at day t.

Chinese mainland is 0.0483. M_t^* and V_t^* represent monthly M_t and its variance at month t, respectively (Tables S2 and S3 in Supporting Information). We estimated b by using the regression model: $V_t^* = V_0^* + bt$; V_t^* shows a significant or an approximately significant linear increasing trend with time (Table S4 in Supporting Information), indicating the evolution of SARS-CoV-2 follows the prediction of the Relaxed Molecular Clock theory, not the Strict Molecular Clock theory (Figures S1 and S2 in Supporting Information).

We estimated the daily deviated cumulative number of substitutions (dM_t) of each sequence of the world by using the mean daily substitution rate (0.0512) and the equation: $dM_t=M_t-0.0512\times t$ (t_0=December 24, 2019). The mean dM_t of the world was estimated to be 2.30 (95%CI: 2.28, 2.32; n=255,039), which is significantly different from zero (P<0.01, t-test) (Figure S4 in Supporting Information). Based on dM_t, we calculated the daily i*DT* of each sequence at time t by i*DT*=(dM_t/0.0512)/2+7 (adding 7 days by referring to the 7-day difference between the sampling time of the Wuhan reference and the starting time of samples) (Figure 1H; Table S5 in Supporting Information). Also, i*DT*=(M_t/0.0512−t)/2+7. The mean i*DT* of the world was estimated to be 29.45 days (95%CI: 29.28, 29.63; n=255,039).

The mean i*DT* of the six continents was estimated to be 24.1 (95%CI, 9.7, 34.5) days by using their own substitution rates (Table S5 in Supporting Information), with the ranking order of South America (39.7)>North America (36.0)>Oceania (24.7)>Africa (22.0)>Asia (21.3)>Europe (0.9) (except for Europe, all>9.7 of the Chinese mainland, P<0.001, Table S6 in Supporting Information). *SD* of i*DT*s of the six continents ranges from 26.8 to 44.7 days, with the ranking order of Africa (44.7)>Europe (38.6)>North America (37.3)>South America (34.8)>Asia (34.1)>Oceania (26.8) (all>Chinese mainland, 21.5 days). The results of the i*DT* method are very similar to those of the r*DT* method (Table S1 *vs*. Table S5 in Supporting Information).

Both dM_t and i*DT* show a general and symmetric bell-shaped Kernel density distribution (Figure 1H; Figure S5 in Supporting Information), suggesting the substitution rate varied randomly. The variation (*SD*) may be caused by both lineage diversity and substitution variance rate (Figure 1). However, the bell-shaped distribution of dM_t and i*DT* show a long right-biased tail with a small peak for the world (also for Europe, South America, and Oceania). i*DT* of the first large and second small peaks (divided by 76 days) of the world is 8 and 112.7 days (around September 2019) before December 31, 2019, respectively. Samples of the second small wave were mostly collected in late 2020 (Figure S6 in Supporting Information).

To reveal potential earlier pandemics represented by rare samples, we calculated the i*DT* of sequences with unexpected extreme high daily M_t (Figure 1; Table S5, Figures S4 and S5 in Supporting Information). The expected M_t (eM_t) was calculated by: eM_t=$a\times t$. By referring to b, we defined sequences with M_t>2× or 4×eM_t as the unexpected extreme high-substitution sequences (Figure S6 in Supporting Information). The mean i*DT* for sequences with M_t>2×eM_t and >4×eM_t of the world (a=0.0512) was estimated to be 81.2 days (95%CI: 79.9, 82.6; n=4,467) (around October 2019) and 175.3 days (95%CI: 118.3, 232.4; n=68) (around June 2019); the i*DT* of 16 sequences is larger than 300 days (around March 2019). The mean i*DT* with M_t>2×eM_t of the six continents (using a of each continent) is 81.6 days (95%CI, 68.3, 103.2), with the ranking order of Africa (108.3)>North America (87.8)>South America (83.4)>Europe (77.8)>Asia (71.4)>Oceania (61.1) (all>Chinese mainland=39.71, P<0.001, Tables S5 and S6 in Supporting Information); The mean i*DT* with M_t>4×eM_t is 187.8 days (95%CI: 145.5, 230.2), with the ranking order of South America (192.8)>North America (240.3)>Europe (150.7)>Africa (190.1)>Asia (165.4) (all>Chinese mainland=50.0, P<0.05, Tables S5 and S6 in Supporting Information). Samples of six continents with M_t>4×eM_t were collected within the first 227 days in 22 countries (or regions) from Europe, Asia, North America, South America, and Africa.

Several factors may cause a biased estimation of divergent time (Supplementary Information). Samples of a region could be transported from other regions. The very large variation of the substitution rate of SARS-CoV-2 may result in under- or over-estimation of the divergent time. Big data helps minimize such biased estimations. Estimates of i*DT* around June or March with a small sample size need further investigation.

In summary, a majority of SARS-CoV-2 samples has a divergent time in December 2019, some samples have a divergent time of around October, June, or March; all mean *DT*s of the six continents (except for one estimate of Europe) are significantly larger than those of Chinese mainland. Our study indicates that SARS-CoV-2 shows a much earlier divergence in the world than in the Chinese mainland. It is necessary to search for the origins of SARS-CoV-2 and its natural hosts in a broad time and space scale around the world.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *The work was supported by the Ministry of Science and Technology of the People's Republic of China (2021YFC0863400), and the Institute of Zoology, Chinese Academy of Science (E0517111, E122G611).*

References

Drummond, A., Pybus, O.G., and Rambaut, A. (2003). Inference of viral evolutionary rates from molecular sequences. Adv Parasitol 54, 331–358.

Pekar, J., Worobey, M., Moshiri, N., Scheffler, K., and Wertheim, J.O. (2021). Timing the SARS-CoV-2 index case in Hubei province. Science 372, 412–417.
van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C.S., Boshier, F.A.T., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect Genet Evol 83, 104351.

SUPPORTING INFORMATION

Methods, discussion, and data are attached in Supporting Information. The supporting information is available online at https://doi.org/10.1007/s11427-023-2294-5. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.