Coronavirus GenBrowser for monitoring adaptive evolution and transmission of SARS-CoV-2

Dalang Yu[1,8,†], Xiao Yang[1,6,†], Bixia Tang[2,†], Yi-Hsuan Pan[3,†], Jianing Yang[1,8,†], Junwei Zhu[2,†], Guangya Duan[2,8], Zi-Qian Hao[1,8], Hailong Mu[1], Long Dai[1,6], Wangjie Hu[1,8], Language translation team[7], Xiao Su[4], Guo-Qing Zhang[1,*], Wenming Zhao[2,8,*], Haipeng Li[1,8,9,*]

[1,‡]National Genomics Data Center, Bio-Med Big Data Center/Laboratory of Evolutionary Genomics, CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China.

[2,‡]National Genomics Data Center, Beijing Institute of Genomics (China National Center for Bioinformation), Chinese Academy of Sciences, Beijing 100101, China.

[3]Key Laboratory of Brain Functional Genomics of Ministry of Education, School of Life Science, East China Normal University, Shanghai 200062, China.

[4]Institut Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai 200031, China.

[6]Shanghai Shenyou Biotechnology Co. LTD

[7]Beijing Language and Culture University.

[8]University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100101, China.

[9]Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China.

[†]These authors contributed equally.
[‡]The two institutes contributed equally.

[*]**Corresponding Authors:** zhaowm@big.ac.cn; gqzhang@picb.ac.cn; lihaipeng@picb.ac.cn

Short title: Coronavirus GenBrowser

Abstract

COVID-19 has widely spread across the world, and much research is being conducted on the causative virus SARS-CoV-2. To help control the infection, we developed the Coronavirus GenBrowser (CGB) to monitor the pandemic. CGB allows visualization and analysis of the latest viral genomic data. Distributed genome alignments and an evolutionary tree built on the existing subtree are implemented for easy and frequent updates. The tree-based data are compressed at a ratio of 2,760:1, enabling fast access and analysis of SARS-CoV-2 variants. CGB can effectively detect adaptive evolution of specific alleles, such as D614G of the spike protein, in their early stage of spreading. By lineage tracing, the most recent common ancestor, dated in early March 2020, of nine strains collected from six different regions in three continents was found to cause the outbreak in Xinfadi, Beijing, China in June 2020. CGB also revealed that the first COVID-19 outbreak in Washington State was caused by multiple introductions of SARS-CoV-2. To encourage data sharing, CGB credits the person who first discovers any SARS-CoV-2 variant. As CGB is developed with eight different languages, it allows the general public in many regions of the world to easily access pre-analyzed results of more than 132,000 SARS-CoV-2 genomes. CGB is an efficient platform to monitor adaptive evolution and transmission of SARS-CoV-2.

Main Text

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1-3] has infected more than 75 million people, and at least 1.6 million people in more than 200 countries have died from COVID-19. Many factors have contributed to the COVID-19 pandemic [4-6], and it has been predicted that the COVID-19 pandemic may last until 2025 [7,8]. The pathogen genomics platform Nextstrain has allowed analysis of genomic sequences of approximately 4,000 strains of SARS-CoV-2 and investigation of its evolution [9]. As more than 210,000 SARS-CoV-2 strains have been sequenced (Figure S1), analysis of all strains has far exceeded the capacity of Nextstrain. New approaches are needed to accomplish this task.

To allow timely analysis of a large number of viral genomes, we first solved the problem that all viral genomes have to be re-aligned when nucleotide sequences of new genomes become available. This is extremely time-consuming. With the distributed alignment system (Figure 1), we dramatically reduced the total time required for the alignment. We also built the evolutionary tree on the existing tree and new genomic data in order to reduce the complexity of tree construction. With these modifications, hundreds of thousands of SARS-CoV-2 genomes can be timely analyzed with data easily shared and visualized on personal computers and smart phones (Figure 1).

For genomic sequence alignments, 132,443 high quality SARS-CoV-2 genomic sequences were obtained from the 2019nCoVR database [10], which is an integrated resource based on CNGBdb, GenBank, GISAID [11,12], GWH [13], and NMDC. The sequences were aligned [14] to that of the reference genome and presented as distributed alignments. Genomic sequences of the bat coronavirus RaTG13 [15], the pangolin coronavirus PCoV-GX-P1E [16], and the early SARS-CoV-2 strains collected before Jan 31, 2020 were jointly used to identify ancestral alleles of SARS-CoV-2. The evolutionary tree was rebuilt based on new data and the existing tree, and mutations in strains of each branch were indicated according to the principle of parsimony [17].

The genome-wide mutation rate of coronaviruses has been determined to be $10^{-4} - 10^{-2}$ per nucleotide per year [18]. As this range of mutation rate is too wide, we decided to estimate more precisely the genome-wide mutation rate (μ) of SARS-CoV-2 in a timely manner and determined that $\mu = 6.753 \times 10^{-4}$ per nucleotide per year (95% confidence interval: 4.581×10^{-4} to 9.253×10^{-4}). This calculation did not require information on demography and the time of appearance of the most recent common ancestor (MRCA) of SARS-CoV-2. The estimated μ was lower than that of other coronaviruses, such as SARS-CoV (0.80 to 2.38×10^{-3} per nucleotide per year) [18] and MERS-CoV (1.12×10^{-3} per nucleotide per year) [19]. It was also lower than that determined by other investigators (1.19 to 1.31×10^{-3} per nucleotide per year) [20]. Various mutation rates were found in different regions of SARS-CoV-2 genome. The mutation rate of each gene is presented in Table S1.

Similar to Nextstrain [9] and the WashU Virus Genome Browser [21], the pre-analyzed genomic variant data on CGB are shared with the general public. The size of distributed alignments is 3,894 Mb for the high-quality 132,443 SARS-CoV-2 genomic sequences collected globally. The tree-based data format allows the compression ratio to reach 2,760:1, meaning that the size of compressed data file is as small as 1.41 Mb. This approach ensures low-latency access to the data and enables fast sharing and re-analysis of a large number of SARS-CoV-2 genomic variants.

To visualize, search, and filter the results of genomic analysis, both desktop standalone and web-based user-interface of CGB were developed. Similar to the UCSC SARS-CoV-2 Genome Browser [22] and the WashU Virus Genome Browser [21], six genomic-coordinate annotated tracks were developed to show genome structure and key domains, allele frequencies, sequence similarity, multi-coronavirus genome alignment, and primer sets for detection of various SARS-CoV-2 strains. To efficiently visualize the results of genomic analysis, movie-making ability was implemented for painting the evolutionary tree, and only elements shown on the screen and visible to the user would be painted. This design makes the visualization process highly efficient, and the tree of more than 132,000 viral strains can be visualized even on a smart phone.

CGB detects on-going positive selection based on frequency trajectory of a selected allele. It has been shown that the spike protein G614 variant has a fitness advantage [23]. Our analysis using CGB confirmed this finding even when the frequency of this mutation was very low ($< 10\%$). Moreover, two previously identified variants (ORF1b:P314L, and N:A220V) [24] and five potentially advantageous variants were also identified even though their frequency was lower than 10% (Figure 2, Table S3). Thus, CGB is an efficient monitoring platform for detecting advantageous variants before they become widely spread (Figure S12).

CGB is also an efficient platform to investigate local and global transmission of COVID-19 (Figure 3). There was a recent outbreak in Qingdao, China [25] after two dock workers were found to have asymptomatic infections on September 24, 2020. CGB lineage tracing revealed that the sequence of a sample collected from the outer packaging of cold-chain products is identical to that of the most recent common ancestor of the two viral strains isolated from the two dock workers (Figure 3B), suggesting that infection of these two individuals was cold-chain related. However, this possibility remains to be determined.

CGB lineage tracing also revealed the difficulty in the control of COVID-19 pandemic. There was a recent outbreak in Xinfadi, Beijing, China [26]. The sequences of two viral isolates (Beijing/IVDC-02-06, Beijing/BJ0617-01-Y), collected from two Xinfadi cases on June 11 and 14, 2020, were found to be identical to the sequence of an ancestral strain (Figure 3C) dated on March 6, 2020 (95% CI: February 28 – March 17, 2020). This ancestral strain was found to spread to Taiwan, India, Czech Republic,

England, Denmark, and Colombia and caused the outbreak in Beijing three months later. These two Xinfadi strains were also found to evolve significantly slowly ($P = 0.0043$ and 0.0051, respectively) because no mutations were detected during the three months.

CGB is a powerful tool for the identification of global and regional routes of virus transmission as it is specially designed to determine whether the mutation rate of a specific strain is lower than the average mutation rate of the entire set of strains. This lineage-specific reduced mutation rate could be due to a long period of dormancy caused by the yet to be confirmed cold-chain preservation [27] or other reasons. Among the 132,443 SARS-CoV-2 strains, 4,597 strains were found to evolve significantly slowly ($P = 2.18 \times 10^{-8} \sim 0.0041$, Supplemental excel file) and did not mutate within at least 100 days. This data showed that CGB can narrow the time period for tracing the transmission of a specific strain.

A study on the sequences of 453 SARS-CoV-2 genomes collected before mid-March 2020 suggested that the first COVID-19 outbreak in Washington State was due to a single introduction [28]. However, results of CGB analysis suggest that the first Washington State outbreak was actually caused by multiple introductions (Figure S14).

All the timely-updated data are freely available at https://bigd.big.ac.cn/ncov/apis/. The free desktop standalone version provides the full function of CGB and has a plug-in module for the eGPS software (http://www.egps-software.net/) [29]. Although the web-based tool is a simplified version of CGB (Figure 4) (https://www.biosino.org/genbrowser/ and https://bigd.big.ac.cn/genbrowser/), it provides a convenient way to access the data via a web browser, such as Google Chrome, Firefox and Safari. The web-based CGB package can be downloaded and reinstalled on any websites. For educational purpose, eight language versions (Chinese, English, German, French, Italian, Portuguese, Russian, and Spanish) are available.

Acknowledgments

We thank Ya-Ping Zhang for providing valuable advices and encouragement, and the researchers who generated and deposited the sequencing data of SARS-CoV-2 in GISAID, GenBank, CNGBdb, GWH, and NMDC, making this study possible. This work was supported by a grant from the National Key Research and Development Project (No. 2020YFC0847000).

Members of the language translation team

German: Ning He[7], Jing Lv[7], Ting Peng[7]

Italian: Ting Zhou[7], Nan Yang[7], Siyi Hou[7]

Portuguese: Huang Li[7], Jingxuan Yan[7], Chenglin Zhu[7], Wenjing Liu[7]

Russian: Yuhong Guan[7], Huanxiao Song[7]
Spanish: Qin Zhou[7], Han Gao[7], Jinglan He[7], Tiantian Li[7], Ruiwen Fei[7], Shumei Zhang[7]
French: Yuyuan Guo[7]

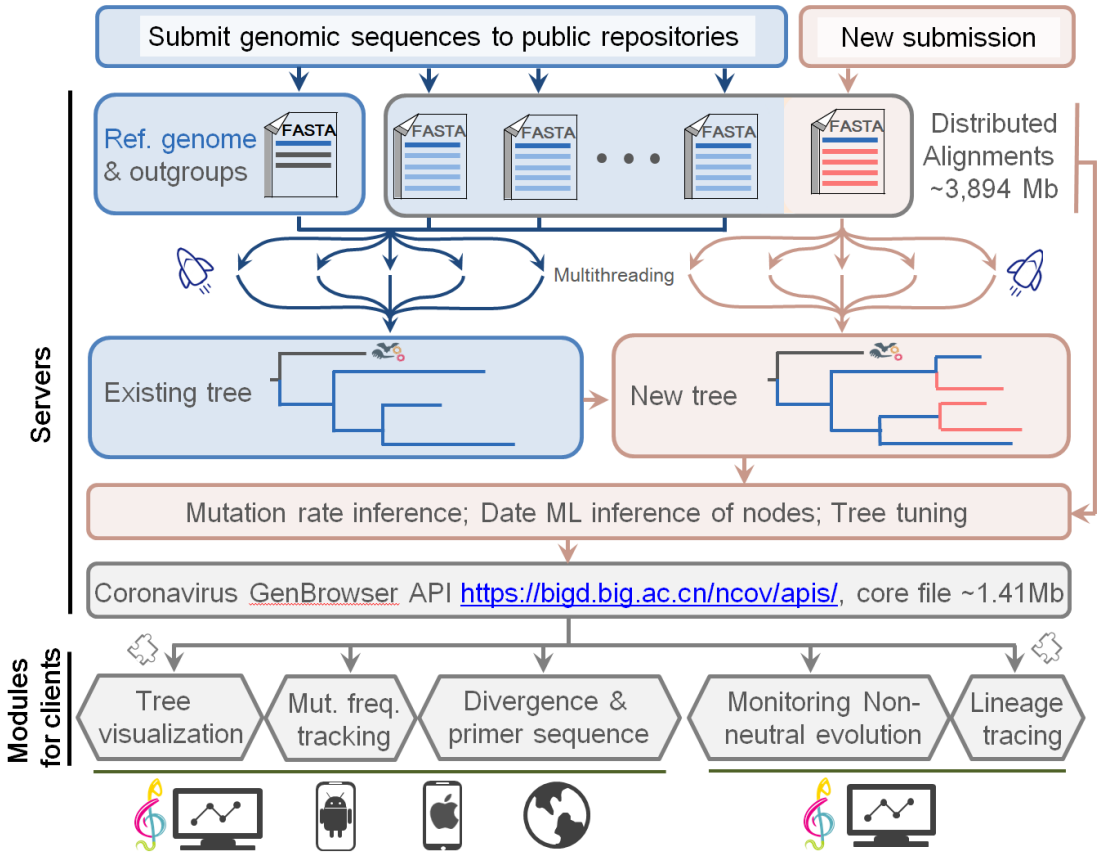

Figure 1. Timely-update and visualization framework of the Coronavirus GenBrowser. The pre-analyzed genomic variant data can be freely accessed via https://bigd.big.ac.cn/ncov/apis/.

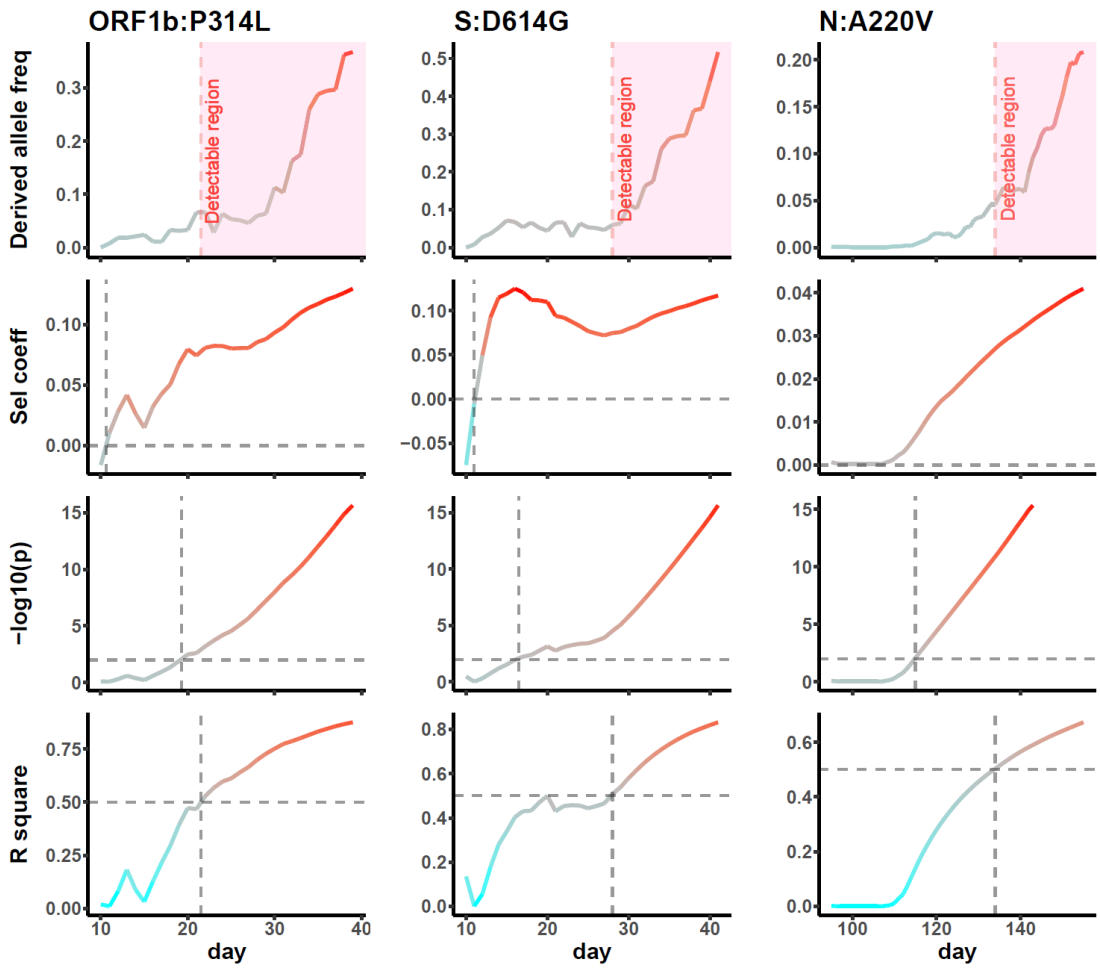

Figure 2. Putative advantageous variants of SARS-CoV-2. The *x*-axis displays number of days since the first appearance of derived allele in the global viral population. Predicted adaptation is marked in pink. Dashed gray crossings denote meaningful top right corners with a positive selection coefficient, $p < 0.01$, and $R^2 > 50\%$.

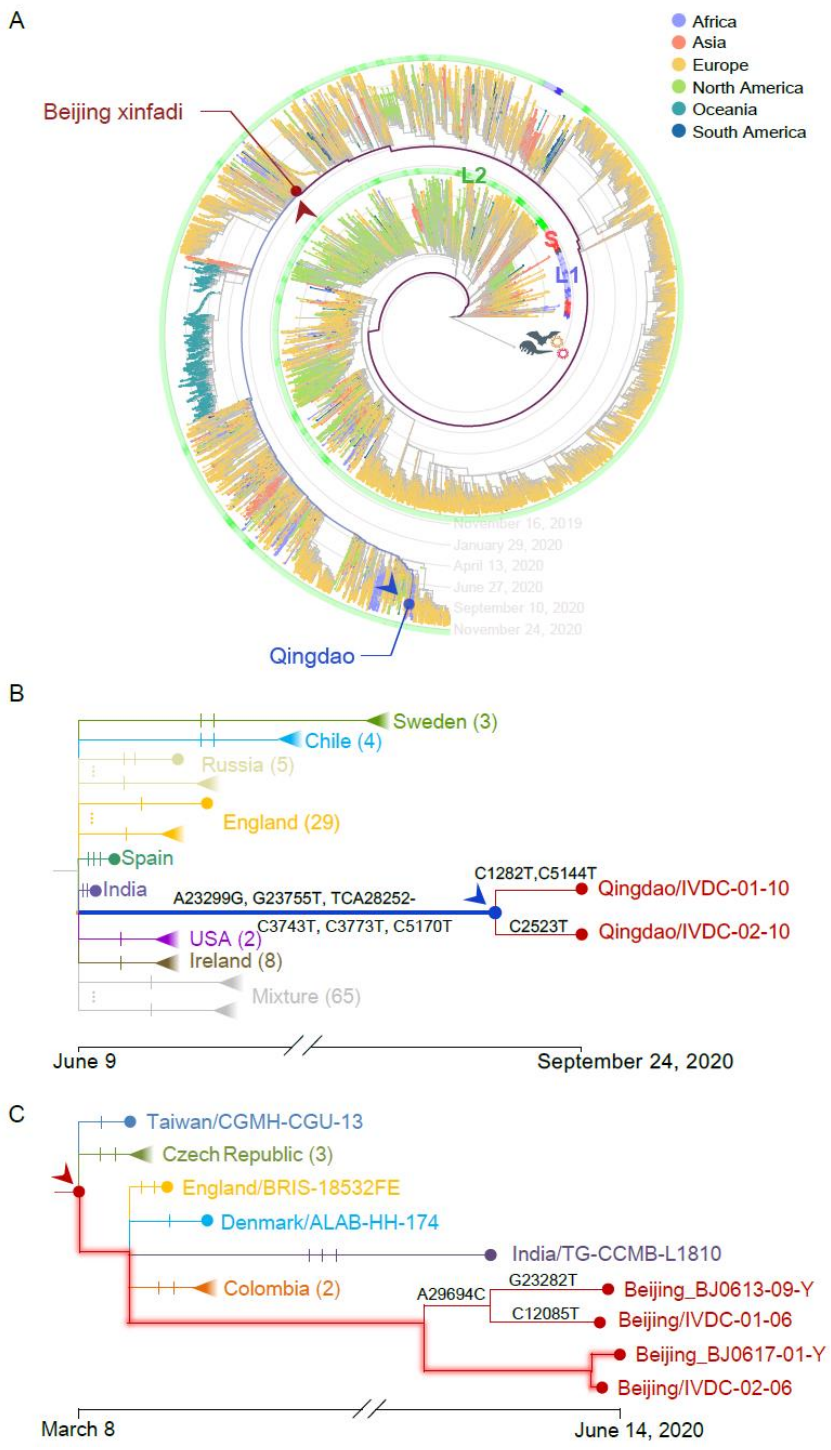

Figure 3. Global and zoomed views of lineages associated with Qingdao and Beijing outbreaks.

A) The lineages of traced targets are shown in blue and dark-red lines. The tree of 132,443 viral strains was used. L/S lineage types [30] are marked with an outside circle.

B) Qingdao/IVDC-01-10 and Qingdao/IVDC-02-10 were the two SARS-CoV-2 strains collected on September 24, 2020 from two dock workers in Qingdao, China. The query strain (env/Qingdao/IVDC-011-10) was found on an outer

packaging of cold-chain products on October 7, 2020 The environmental strain, marked with a blue solid circle with an arrow head, was found to be identical to the most recent common ancestor of the two strains from the two dock workers. Each notch of the branches represents a mutation. Mutations of the Qingdao strains are indicated.

C) The ancestral viral strain found in early March 2020 is marked with a dark-red solid circle with an arrow head. This strain is identical to the two strains (Beijing/IVDC-02-06, Beijing/BJ0617-01-Y) collected from two Xinfadi cases on June 11 and 14, 2020. The branches with no mutations are highlighted.

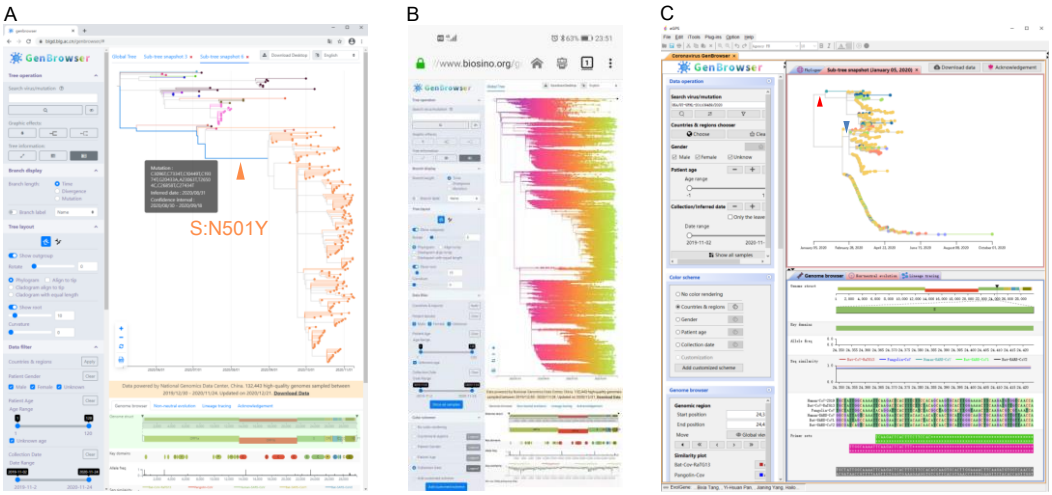

Figure 4. Detection of non-neutral evolution of SARS-CoV-2 and tree visualization with CGB.

A) Web-based CGB tree visualization of an accelerated lineage in the UK, out of 132,443 SARS-CoV-2 genomic sequences, with the desktop version of Google Chrome.

B) Web-based CGB tree visualization of 132,443 genomes with the Android version of Firefox.

C) Tree visualization of a lineage (USA/UT-UPHL-201109489/2020) with the mostly reduced evolutionary rate and its neighbors with desktop standalone CGB. There are only two mutations (A20268G, red arrow head; C15324T, blue arrow head) happened in 966 strains within nearly 9 months.

References

1 Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med* **382**, 727-733 (2020).

2 Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565-574 (2020).

3 Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* (2020).

4 Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nature Microbiology* (2020).

5 Lan, J. *et al.* Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature* **581**, 215-220 (2020).

6 Hao, X. *et al.* Reconstruction of the full transmission dynamics of COVID-19 in Wuhan. *Nature* (2020).

7 Kissler, S. M., Tedijanto, C., Goldstein, E., Grad, Y. H. & Lipsitch, M. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science* **368**, 860-868 (2020).

8 Scudellari, M. The pandemic's future. *Nature* **584**, 22-25 (2020).

9 Hadfield, J. *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121-4123 (2018).

10 Zhao, W.-M. *et al.* The 2019 novel coronavirus resource. *Yi Chuan* **42**, 212-221 (2020).

11 Shu, Y. L. & McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance* **22**, 2-4 (2017).

12 Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall* **1**, 33-46 (2017).

13 Zhang, Z. *et al.* Database resources of the National Genomics Data Center in 2020. *Nucleic Acids Res* **48**, D24-D33 (2020).

14 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**, 772-780 (2013).

15 Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* (2020).

16 Lam, T. T.-Y. *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282-285 (2020).

17 Hartigan, J. A. Minimum mutation fits to a given tree. *Biometrics* **29**, 53-65 (1973).

18 Zhao, Z. M. *et al.* Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol* **4**, 21 (2004).

19 Cotten, M. *et al.* Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. *Mbio* **5** (2014).

20 Li, X. *et al.* Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. *J Med Virol* (2020).

21 Flynn, J. A. *et al.* Exploring the coronavirus pandemic with the WashU Virus Genome Browser. *Nat Genet* **52**, 986-1001 (2020).

22 Fernandes, J. D. *et al.* The UCSC SARS-CoV-2 Genome Browser. *Nat Genet* **52**, 986-991 (2020).

23 Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* **182**, 812-827 (2020).

24 Hodcroft, E. B. *et al.* Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv*, doi:https://doi.org/10.1101/2020.10.25.20219063 (2020).

25 Xing, Y., Wong, G. W. K., Ni, W., Hu, X. & Xing, Q. Rapid response to an outbreak in Qingdao, China. *N Engl J Med*, doi:10.1056/NEJMc2032361 (2020).

26 Zhang, Y. *et al.* Genomic characterization of SARS-CoV-2 identified in a reemerging COVID-19 outbreak in Beijing's Xinfadi market in 2020. *Biosaf Health*, doi:http://dx.doi.org/10.1016/j.bsheal.2020.08.006 (2020).

27 Pang, X. *et al.* Cold-chain food contamination as the possible origin of Covid-19 resurgence in Beijing. *Natl Sci Rev*, doi:10.1093/nsr/nwaa264 (2020).

28 Bedford, T. *et al.* Cryptic transmission of SARS-CoV-2 in Washington state. *Science* **370**, 571-575 (2020).

29 Yu, D. *et al.* eGPS 1.0: comprehensive software for multi-omic and evolutionary analyses. *Natl Sci Rev* **6**, 867-869 (2019).

30 Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *Natl Sci Rev* (2020).