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

Dalang Yu[1,7,†], Xiao Yang[1,5,†], Bixia Tang[2,†], Yi-Hsuan Pan[3,†], Jianing Yang[1,7,†], Junwei Zhu[2,†], Guangya Duan[2,7], Zi-Qian Hao[1,7], Hailong Mu[1], Long Dai[1,5], Wangjie Hu[1,7], Mochen Zhang[2,7], Ying Cui[2,7], Tong Jin[2,7], Cuiping Li[2], Lina Ma[2], Language translation team[6], Xiao Su[4], Guo-Qing Zhang[1,7,*], Wenming Zhao[2,7,*], Haipeng Li[1,7,8,*]

[1,‡]National Genomics Data Center, Bio-Med Big Data Center, 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.

[5]Shanghai Shenyou Biotechnology Co. LTD, Shanghai 201315, China.

[6]Beijing Language and Culture University, Beijing 100083, China.

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

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

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

[*]**Corresponding authors:** gqzhang@picb.ac.cn; zhaowm@big.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. With CGB, 178,765 high quality SARS-CoV-2 genomic sequences were analyzed, and 121,522 mutations were identified. In total, 1,041 mutation cold spots were found, suggesting that these spots are key functional elements of SARS-CoV-2 and can be used for detection and vaccine development. CGB revealed 203 accelerated evolutions of SARS-CoV-2, but variants with accelerated evolution were not found to be highly contagious, suggesting that most of these evolutions are neutral. The B.1.1.7 (CGB75056.84017) lineage previously identified in the UK was not found to be significantly accelerated although its adaptive evolution was detected. Moreover, 2,297 strains with a significantly reduced evolutionary rate were identified, including three closely related variants widely spreading in Europe with no mutations in three months. By lineage tracing, a strain dated early March 2020 was determined to be the most recent common ancestor of nine strains collected from six different regions in three continents. This strain was also found to cause the outbreak in Xinfadi, Beijing, China in June 2020. CGB allows visualization and analysis of hundreds of thousands of SARS-CoV-2 genomic sequences. Distributed genome alignments and its effective analysis pipeline ensure timely update of the latest genomic data of SARS-CoV-2. CGB is an efficient platform for the general public to monitor the transmission and evolution of SARS-CoV-2.

Main Text

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1-3] has infected more than 102 million people, and more than 2 million people 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 317,000 SARS-CoV-2 strains have been sequenced (Figure S1) [10-12], 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 with 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, 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 bat coronavirus RaTG13 [15], pangolin coronavirus PCoV-GX-P1E [16], and early SARS-CoV-2 strains collected before Jan 31, 2020 were jointly used to identify ancestral alleles of SARS-CoV-2. Mutations in strains of each branch of the evolutionary tree were indicated according to the principle of parsimony [17]. A highly effective maximum-likelihood method (TreeTime) [18] was used to determine the dates of internal nodes with very minor revisions.

In total, 178,765 high quality SARS-CoV-2 genomic sequences collected globally were analyzed, and 121,522 mutations were identified. With sliding window analysis, 1,041 mutation cold spots were found with a false discovery rate (FDR) corrected *P*-value < 0.01 (Figure S5, Supplemental excel file). The top three cold spots were located in ORF1a encoding nsp3 phosphoesterase (nucleotides 7,394 – 7,419), ORF1b (nucleotides 15,451 – 15,539), and the receptor binding domain of the spike protein (nucleotides 23,128 – 23,184) [3] (FDR corrected *P*-value $\leq 1.84 \times 10^{-12}$). These mutation cold spots may be key functional elements of SARS-CoV-2 and can potentially be used for vaccine development and targets for detection.

The genome-wide mutation rate of coronaviruses has been determined to be $10^{-4} - 10^{-2}$ per nucleotide per year [19]. 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 $\mu = 6.8017 \times 10^{-4}$ per nucleotide per year (95% confidence interval: 5.4262 to 8.2721×10^{-4}). 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) [19] and MERS-CoV (1.12×10^{-3} per nucleotide per year) [20]. It was slightly lower than that determined by other investigators (9.90×10^{-4} per nucleotide per year) [21]. Various mutation rates were found in different regions of the SARS-CoV-2 genome. The mutation rate of each gene is presented in Table S1.

Similar to Nextstrain [9], the pre-analyzed genomic data of SARS-CoV-2 variants on CGB are shared with the general public. The size of distributed alignments is 5,130 Mb for the high-quality 178,765 SARS-CoV-2 genomic sequences. The tree-based data format allows the compression ratio to reach 2,527:1, meaning that the size of compressed data file is as small as 2.03 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 [23], 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 (Figure S10). 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 250,000 viral strains can be visualized even on a smart phone.

CGB detects on-going positive selection based on S-shaped frequency trajectory of a selected allele (Figures S16, S17). It has been shown that the SARS-CoV-2 variant with G614 spike protein has a fitness advantage [24,25]. Our analysis using CGB confirmed this finding even when the G614 frequency was very low ($< 10\%$) (Figure 2). Thus, CGB is an efficient monitoring platform for detecting putative advantageous variants before they become widely spread. As an increase in mutation frequency could be due to sampling bias and epidemiological factors [24], putative advantageous variants should be closely monitored.

Using CGB, we analyzed branch-specific accelerated evolution of SARS-CoV-2 and found that 203 internal branches of the evolutionary tree (FDR corrected $P < 0.05$, Poisson probability, Supplemental excel file) had significantly more mutations. All evolution-accelerated variants were not found to spread significantly faster than other variants during the same period of time, suggesting that these variants are not highly

contagious and that most of the accelerated evolutions are neutral. The majority ($157/203 = 77.3\%$) of these variants even had relatively fewer descendants.

The B.1.1.7 (CGB75056.84017) lineage was recently identified in the UK [26]. Although it has 23 mutations, its evolutionary rate was not significantly accelerated (FDR corrected $P = 0.346$, Poisson probability) as the branch of this lineage spans 7.5 months. Its mutation rate was determined to be 5.5382×10^{-4} per nucleotide per year (see Supplemental materials), slightly lower than that estimated from the entire set of strains. However, the spread of B.1.1.7 variants was significantly faster than other variants collected in mid-September 2020 (FDR corrected $P = 0.0032$). The frequencies of S:S982A and S:D1118H mutations, first found on the B.1.1.7 branch, appeared to be in the early stage of an S-shaped rising (Figure 2).

CGB is also an efficient platform to investigate local and global transmission of COVID-19 (Figure 3). There was an outbreak in Qingdao, China [27] 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 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 controlling COVID-19 pandemic. There was an outbreak in Xinfadi, Beijing, China [28,29]. The sequences of two isolates (Beijing/IVDC-02-06 and 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 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, Poisson probability) because no mutations were detected between March and June 2020.

CGB is a powerful tool for 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 [29] or other reasons. Among the 178,765 SARS-CoV-2 strains, 2,297 strains were found to evolve significantly slowly (FDR corrected $P = 3.45 \times 10^{-4} \sim 0.05$, Poisson probability, Supplemental excel file) and did not mutate in 133 days. In addition, three closely related variants were found to have no mutations in 3 months, and their descendants widely spread in Europe (Figure 4A).

All timely-updated data are freely available at https://bigd.big.ac.cn/ncov/apis/. The desktop standalone version provides the full function of CGB and has a plug-in module for the eGPS software (http://www.egps-software.net/) [30]. Although the web-based CGB is a simplified version (https://www.biosino.org/genbrowser/ and https://bigd.big.ac.cn/genbrowser/) and designed mainly for educational purpose, it provides a convenient way to access the data via a web browser, such as Google Chrome, Firefox, and Safari (Figure 4B, C). The web-based CGB package can be downloaded and reinstalled on any websites. Nine language versions (Chinese, English, German, Japanese, 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 sequence 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[6], Jing Lv[6], Ting Peng[6]
Italian: Ting Zhou[6], Nan Yang[6], Siyi Hou[6]
Portuguese: Huang Li[6], Jingxuan Yan[6], Chenglin Zhu[6], Wenjing Liu[6]
Russian: Yuhong Guan[6], Huanxiao Song[6]
Spanish: Qin Zhou[6], Han Gao[6], Jinglan He[6], Tiantian Li[6], Ruiwen Fei[6], Shumei Zhang[6]
French: Yuyuan Guo[6]

Author contributions

YHP, GQZ, WZ, and HL designed the study; DY, XY, BT, YHP, JY, JZ, GD, ZQH, HM, LD, GQZ, WZ, and HL wrote the code and developed CGB; DY, XY, BT, YHP, JY, JZ, GD, ZQH, WH, XS, GQZ, WZ, and HL acquired, analyzed, and interpreted the data; LM, MZ, YC, GD, TJ and CL integrated and curated the source data; members of the language translation team translated CGB into multiple languages; DY, YHP, JY, JZ, GQZ, WZ, and HL wrote the manuscript. All authors have approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper.
Correspondence and requests for materials should be addressed to G.Q.Z., W.Z., or H.L.

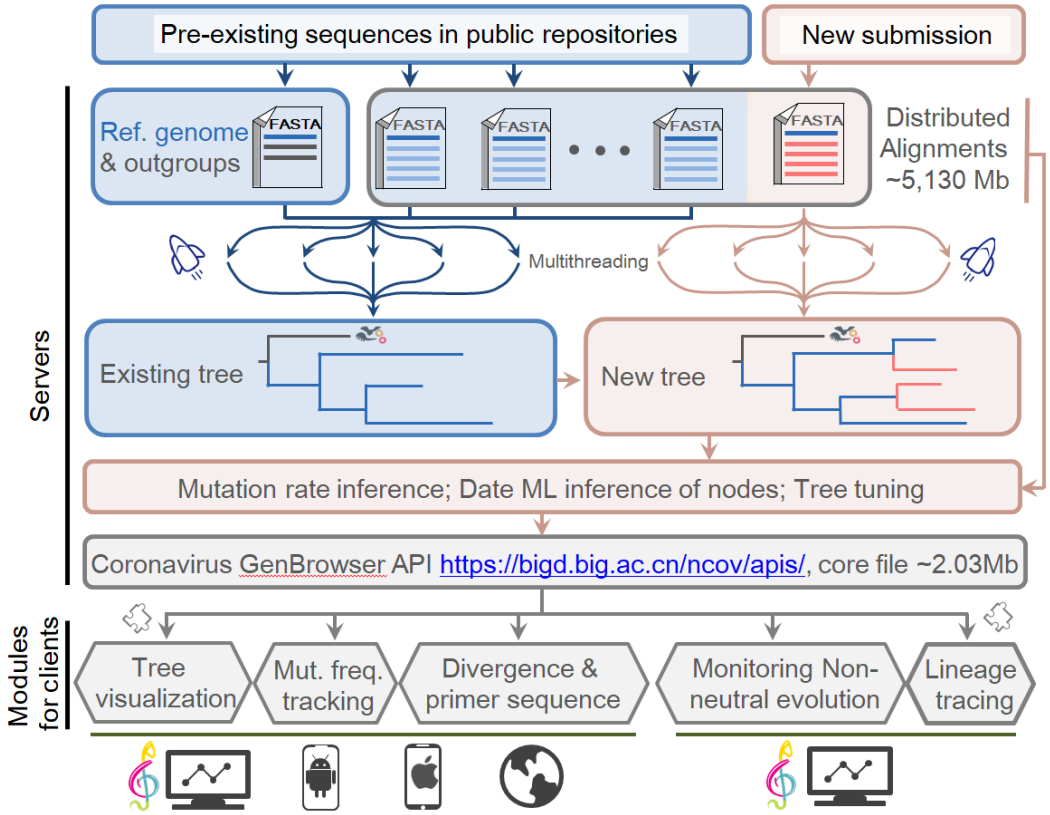

Figure 1. Timely updates and visualization framework of Coronavirus GenBrowser. The pre-analyzed genomic data of SARS-CoV-2 variants can be freely accessed via https://bigd.big.ac.cn/ncov/apis/.

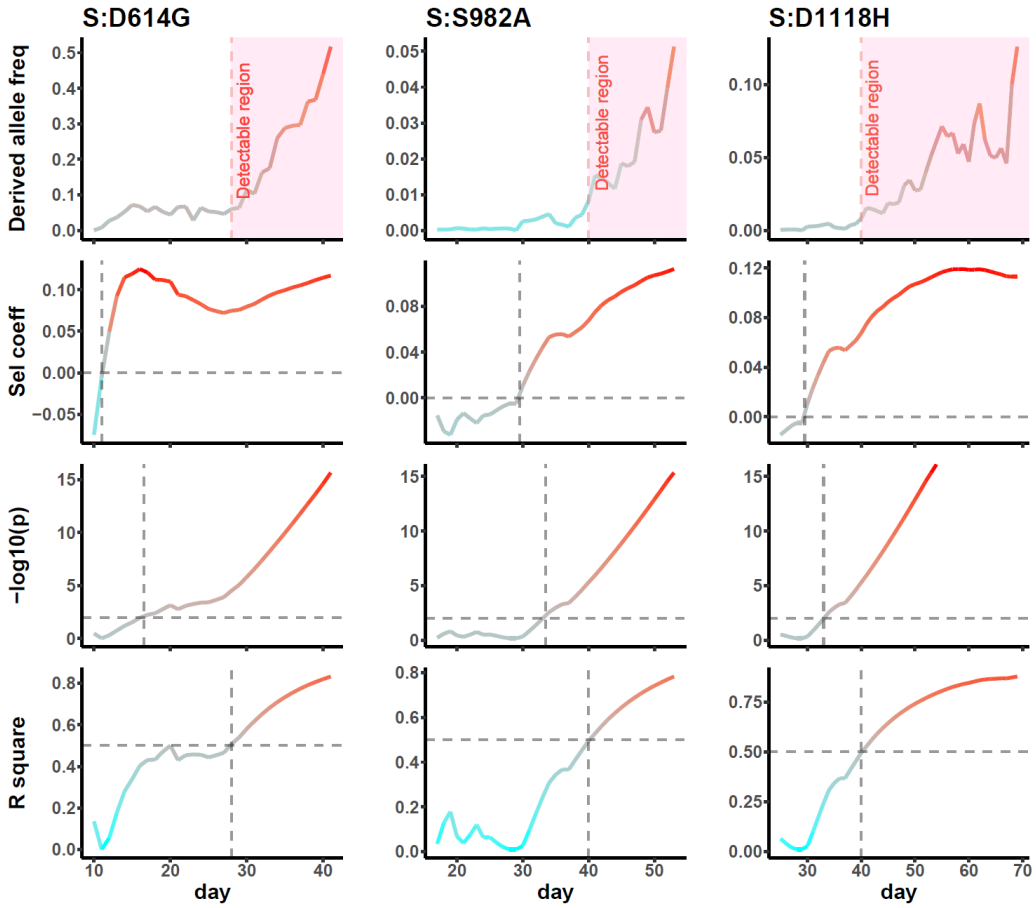

Figure 2. Putative advantageous variants of SARS-CoV-2. The *x*-axis displays number of days since the first appearance of derived allele in the global viral population. Predicted adaptation is marked in pink. Dashed gray crossings denote meaningful top right corners with a positive selection coefficient,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

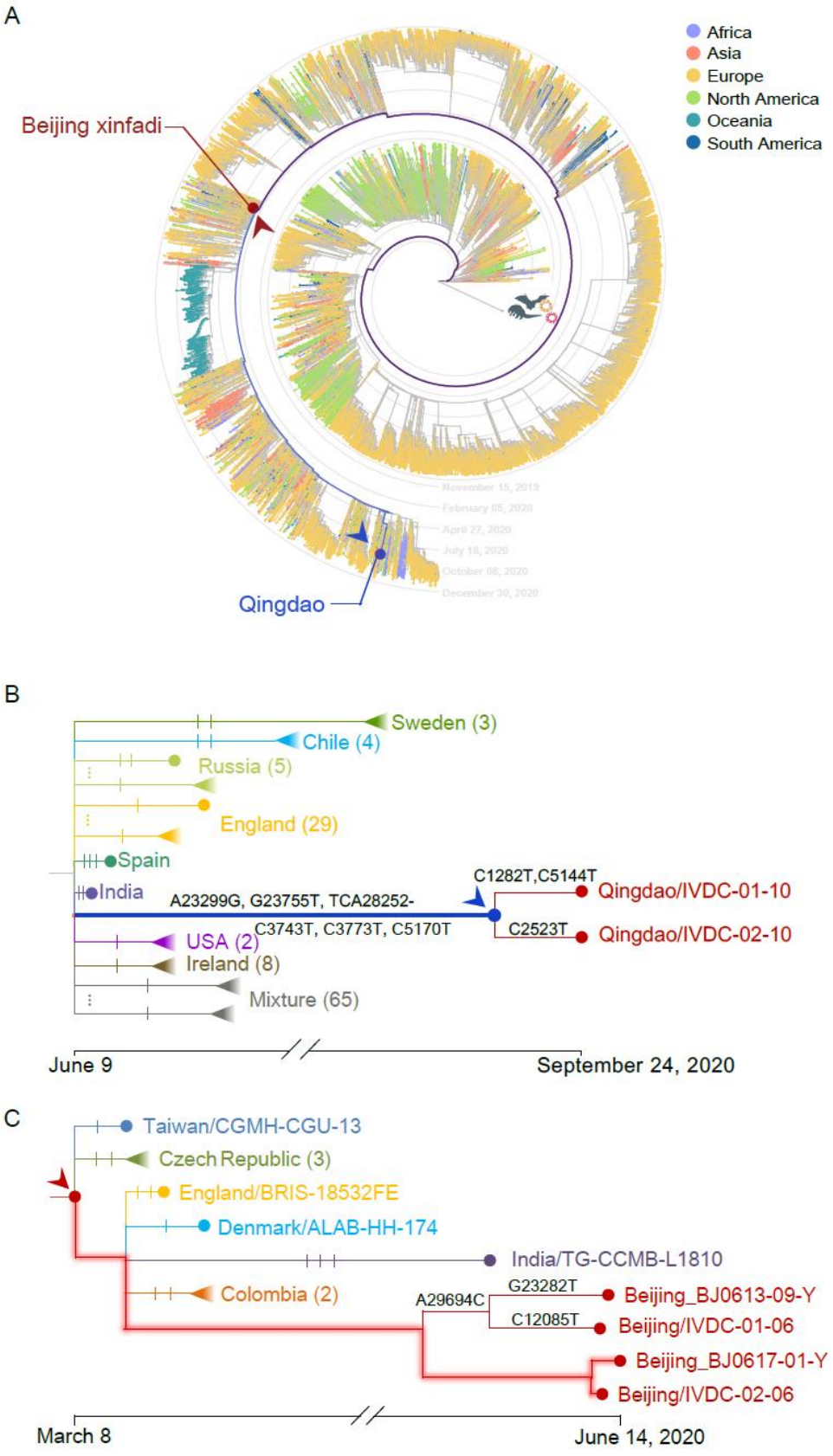

Figure 3. Global and zoom-in views of lineages associated with Qingdao and Beijing outbreaks.

A) The lineages of traced targets are shown in blue and dark red lines. The tree of 178,765 viral strains was used.
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 arrowhead, 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 and an arrowhead. This strain is identical to the two strains (Beijing/IVDC-02-06 and Beijing/BJ0617-01-Y) collected from two Xinfadi cases on June 11 and 14, 2020. The branches with no mutations are highlighted.

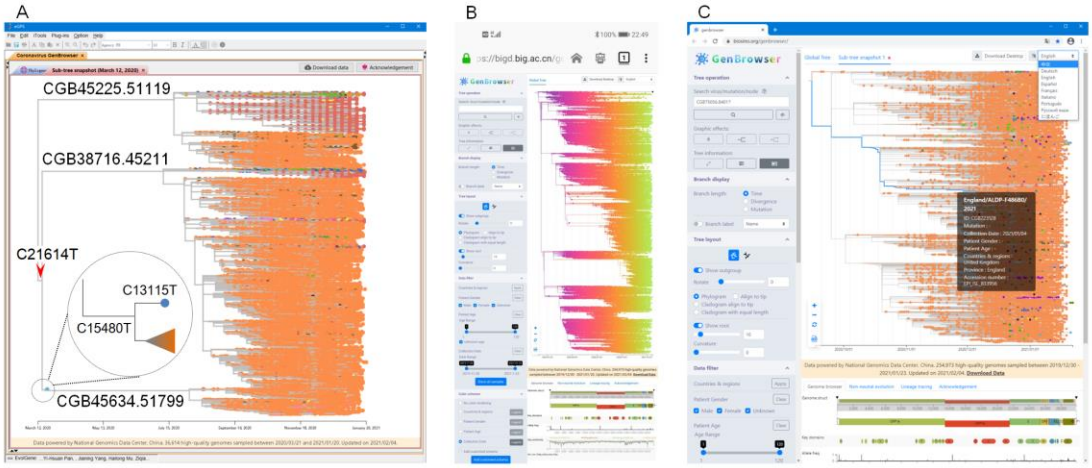

Figure 4. Tree visualization with CGB.

A) Tree visualization of three closely related lineages with the reduced evolutionary rate among 254,973 SARS-CoV-2 genomic sequences with desktop standalone CGB. There were no mutations in strains of three long branches (CGB45225.51119, CGB38716.45211, and CGB45634.51799) with inferred duration of 130, 116, and 138 days, respectively. The numbers of their descendants were 8,738, 25,050, and 2,825, respectively. The three strains spread to 14 European countries (Austria, Belgium, Denmark, Finland, France, Germany, Iceland, Ireland, Lithuania, Luxembourg, Netherlands, Norway, Switzerland, and United Kingdom), and most descendants were from Denmark (in red tip) and UK (in orange tip). These three lineages shared 13 mutations (G204T, C241T, T445C, C3037T, C6286T, C14408T, G21255C, C22227T, A23403G, C26801G, C27944T, C28932T, and G29645T) but had two unique mutations (C21614T and C15480T). The control panel and annotated tracks are hidden.

B) Web-based CGB tree visualization of 254,973 genomes with the Android version of Firefox.

C) Web-based CGB tree visualization of the B.1.1.7 (CGB75056.84017) UK lineage among 254,973 SARS-CoV-2 genomic sequences with the desktop version of Google Chrome.

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* **579**, 265-269 (2020).

4 Boni, M. F. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* **5**, 1408-1417 (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* **584**, 420-424 (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* **579**, 270-273 (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 Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol* **4**, vex042 (2018).

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

20 Cotten, M. *et al.* Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. *Mbio* **5**, e01062-01013 (2014).

21 Nie, Q. *et al.* Phylogenetic and phylodynamic analyses of SARS-CoV-2. *Virus Res* **287**, 198098 (2020).

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

338 23 Flynn, J. A. *et al.* Exploring the coronavirus pandemic with the WashU Virus Genome
339 Browser. *Nat Genet* **52**, 986-1001 (2020).
340 24 Korber, B. *et al.* Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases
341 infectivity of the COVID-19 virus. *Cell* **182**, 812-827 (2020).
342 25 Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature*,
343 doi:10.1038/s41586-020-2895-3 (2020).
344 26 Rambaut, A. *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage
345 in the UK defined by a novel set of spike mutations. *virological.org*,
346 https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-linea
347 ge-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563 (2020).
348 27 Xing, Y., Wong, G. W. K., Ni, W., Hu, X. & Xing, Q. Rapid response to an outbreak in
349 Qingdao, China. *N Engl J Med* **383**, e129 (2020).
350 28 Zhang, Y. *et al.* Genomic characterization of SARS-CoV-2 identified in a reemerging
351 COVID-19 outbreak in Beijing's Xinfadi market in 2020. *Biosaf Health* **2**, 202-205 (2020).
352 29 Pang, X. *et al.* Cold-chain food contamination as the possible origin of Covid-19 resurgence in
353 Beijing. *Natl Sci Rev* **7**, 1861-1864 (2020).
354 30 Yu, D. *et al.* eGPS 1.0: comprehensive software for multi-omic and evolutionary analyses.
355 *Natl Sci Rev* **6**, 867-869 (2019).
356