medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

# 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], Cui-Ping Li[2], Lina Ma[2], Language translation team[6], Xiao Su[4], Guoqing Zhang[1,7,*], Wenming Zhao[2,7,*], Haipeng Li[1,7,8,*]

[1,‡]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

**NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.**

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

**Abstract:** A large volume of SARS-CoV-2 genomic data has accumulated. To timely analyze and visualize the exponentially increasing viral genomic sequences, we developed the Coronavirus GenBrowser (CGB) based on the framework of distributed genome alignments and the evolutionary tree built on an existing subtree. All 98,496 internal nodes were named by CGB binary nomenclature. Among the 330,942 high-quality SARS-CoV-2 genomic sequences analyzed, 253,798 mutations and 971 mutation cold spots were identified. This analysis also revealed a strain dated early March 2020 causing an outbreak in Beijing after three months of dormancy. Three prevalent European variants were found to have no mutations in three months. Another strain with S:D614G was found to have 671 identical descendants spreading in six continents between February 2020 and January 2021. Mutation-dormant strains provide evidences of cold-chain related transmission. Results of this study show that CGB is an efficient platform for monitoring the dynamics of SARS-CoV-2 transmission.

**Key words**

Coronavirus GenBrowser; SARS-CoV-2; cold-chain related transmission; mutation-dormant strains

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

**Main Text:** Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (1-3) has infected more than 116 million people, and more than 2.5 million people have died from COVID-19. Several web browsers have been developed to track the COVID-19 pandemic. The browser of the Johns Hopkins Coronavirus Resource Center displays global and regional COVID-19 status, including numbers of confirmed cases, deaths, tests, hospitalizations, and vaccinations (https://coronavirus.jhu.edu/). The COVID-3D browser, developed by the University of Melbourne, shows the three-dimensional structures of various wild-type and mutant proteins of SARS-CoV-2 based on their nucleotide sequences (4). The UCSC SARS-CoV-2 Genome Browser is derived from the well-established UCSC genome-browser for visualization of nucleotide and protein sequences, sequence conservations, potential epitopes for production of antibodies, primers for RT-PCR and sequencing, and many other properties of specific parts of wild-type and variants of SARS-CoV-2 (5). The WashU Virus Genome Browser provides Nextstrain-based phylogenetic-tree view and genomic-coordinate, track-based view of genomic features of viruses (6). It also has many functions similar to those of the UCSC SARS-CoV-2 Genome Browser such as visualization of sequence variations, genomic sites for diagnostic tests, and immunogenic epitopes.

Nextstrain is a browser for real-time tracking of pathogen evolution. It hosts tools for phylodynamic analysis of genomic sequences of pathogens of both endemic and pandemic diseases. Nextstrain has allowed analysis of genomic sequences of approximately 4,000 strains of SARS-CoV-2 and investigation of its evolution (7). As more than one million SARS-CoV-2 genomic sequences have been deposited in public databases, such as National Center for Biotechnology Information (NCBI) GenBank (8) and Global Initiative on Sharing All Influenza Data (GISAID) (9, 10), analysis of these data has far exceeded the capacity of Nextstrain. New approaches are needed to accomplish this task. Therefore, we developed the Coronavirus GenBrowser (CGB).

To allow timely analysis of a large number of SAR-CoV-2 genomic sequences, we first solved the problem that all viral genomic sequences 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 (Figure 1).

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

For genomic sequence alignments, high quality SARS-CoV-2 genomic sequences were obtained from the 2019nCoVR database (11, 12), which is an integrated resource based on GenBank, GISAID (9, 10), China National GeneBank DataBase (CNGBdb) (13), the Genome Warehouse (GWH) (14), and the National Microbiology Data Center (NMDC, https://nmdc.cn/). The sequences were aligned (15) to that of the reference genome (3) and presented as distributed alignments. Genomic sequences of bat coronavirus RaTG13 (16), pangolin coronavirus PCoV-GX-P1E (17), and SARS-CoV-2 strains collected before 31 January, 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 (18). A highly effective maximum-likelihood method (TreeTime) (19) was used to determine the dates of internal nodes with very minor revisions.

CGB is frequently updated. By analyzing 330,942 high quality SARS-CoV-2 genomic sequences, 253,798 (recurrent) mutations were identified. With sliding window analysis, 971 mutation cold spots were found with a false discovery rate (FDR) corrected *P*-value $< 0.01$ (Figure S5, Supplemental excel file). The coldest spot is located in ORF1a, which encodes nsp3 phosphoesterase (nucleotides 7,394 – 7,419) (FDR corrected *P*-value $= 4.28 \times 10^{-29}$). 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 (20). As this range of mutation rate is too wide, we decided to estimate more precisely the genome-wide mutation rate ($\mu$) of SARS-CoV-2 and determined that $\mu = 6.8017 \times 10^{-4}$ per nucleotide per year (95% confidence interval: $5.4262$ to $8.2721 \times 10^{-4}$). The estimated $\mu$ was lower than that of other coronaviruses, such as SARS-CoV (0.80 to $2.38 \times 10^{-3}$ per nucleotide per year) (20) and MERS-CoV ($1.12 \times 10^{-3}$ per nucleotide per year) (21). It was slightly lower than that determined by other investigators ($9.90 \times 10^{-4}$ per nucleotide per year) (22). Various mutation rates were found in different regions of the SARS-CoV-2 genome. The mutation rate of each gene is shown in Table S1.

Similar to Nextstrain (7), the pre-analyzed genomic data of SARS-CoV-2 variants and the associated metadata on CGB are shared with the general public. The size of distributed alignments is 9,498 Mb for the high-quality 330,942 SARS-CoV-2 genomic

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

sequences. The tree-based data format allows the compression ratio to reach 2,499:1, meaning that the size of compressed core file containing pre-analyzed genomic data and associated metadata is as small as 3.80 Mb (Figure 1). 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 (5) and the WashU Virus Genome Browser (6), 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 S11). 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 evolutionary tree of more than one million strains can be visualized simultaneously.

CGB detects on-going positive selection based on S-shaped frequency trajectory of a selected allele (Figures S18, S19). It has been shown that the SARS-CoV-2 variant with G614 spike protein has a fitness advantage (23). Our analysis using CGB confirmed this finding even when the G614 frequency was very low ($< 10\%$) (Figure 2), indicating that CGB can detect putative advantageous variants before they become widely spread. GCB also predicted an increase in the frequency of ORF1a:SGF3675-deletion of the B.1.1.7 linage (24) (Figure 2), suggesting that variants with the deletion may be advantageous. ORF1a:SGF3675- is the last non-synonymous mutation occurred on the B.1.1.7 lineage (Figure S15). As an increase in mutation frequency could be due to sampling bias and epidemiological factors (23), putative advantageous variants should be closely monitored.

Using CGB, we analyzed branch-specific accelerated evolution of SARS-CoV-2 and found that 246 internal branches of the evolutionary tree (FDR corrected $P < 0.05$, 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 as the majority ($168/225 = 74.6\%$) of these variants had relatively fewer descendants. This observation suggests that these variants are not

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

highly contagious. It is likely that these evolution-accelerated variants were evolved neutrally instead of adaptively.

CGB is also an efficient platform to investigate local and global transmission of COVID-19 (Figure 3). There was an outbreak in Qingdao, China (25) after two dock workers were found to have asymptomatic infections on 24 September, 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.

CGB lineage tracing also revealed the difficulty in controlling COVID-19 pandemic. There was an outbreak in Xinfadi, Beijing, China (26, 27). The sequences of two isolates (Beijing/IVDC-02-06 and Beijing/BJ0617-01-Y), collected from two Xinfadi cases on 11 June and 14 June, 2020, were found to be identical to the sequence of the variant (CGB4268.5142) (Figure 3C) dated 6 March, 2020 (95% CI: 28 February – 17 March, 2020). This variant was found to spread to Taiwan, India, Czech Republic, England, Denmark, and Colombia and caused the outbreak in Beijing three months later. It was also detected in the United States (NY-Wadsworth-21013812-01/2020) on 18 December, 2020. These two Xinfadi strains were found to evolve significantly slowly ($P = 0.0043$ and $0.0051$, respectively) because no mutations were detected between March and June 2020, consistent with the hypothesis of cold-chain related transmission (27).

CGB is a powerful tool for identification of global and regional transmission of mutation-dormant strains as it can determine whether the mutation rate of a specific strain is significantly 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 or other reasons. Among the 330,942 SARS-CoV-2 strains, 7,534 strains were found to evolve significantly slowly (FDR corrected $P = 9.68 \times 10^{-9} \sim 0.04$, Supplemental excel file) and did not mutate in 100 days. In addition, three closely related variants were found to have no mutations in 3 months before their descendants were found to widely spread in Europe in late July 2020 (Figure 4A). Another variant with C3037T, A23403G (S:D614G), C14408T (ORF1b:P314L), and C241T, emerged in early December 2019. This variant had 671 identical descendants spreading in six continents between February 2020 and January 2021, indicating that it has no additional mutations in one

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

year (Figure 4B). The most ($582/671 = 86.7\%$) of the descendants were found in Europe and North America. These results (Figures 3 and 4) suggest that the transmission of mutation-dormant variants also plays a major role in COVID-19 pandemic.

All timely-updated data are freely available at https://bigd.big.ac.cn/ncov/apis/. The desktop standalone version (Figure S10A) provides the full function of CGB and has a plug-in module for the eGPS software (http://www.egps-software.net/) (28). 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 S10B, 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.

## References


1. Zhu, N, Zhang, D, Wang, W, *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020; **382**(8): 727-33.
2. Lu, R, Zhao, X, Li, J, *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*. 2020; **395**(10224): 565-74.
3. Wu, F, Zhao, S, Yu, B, *et al.* A new coronavirus associated with human respiratory disease in China. *Nature*. 2020; **579**: 265-9.
4. Portelli, S, Olshansky, M, Rodrigues, CHM, *et al.* Exploring the structural distribution of genetic variation in SARS-CoV-2 with the COVID-3D online resource. *Nat Genet*. 2020; **52**(10): 999-1001.
5. Fernandes, JD, Hinrichs, AS, Clawson, H, *et al.* The UCSC SARS-CoV-2 Genome Browser. *Nat Genet*. 2020; **52**: 986-91.
6. Flynn, JA, Purushotham, D, Choudhary, MNK, *et al.* Exploring the coronavirus pandemic with the WashU Virus Genome Browser. *Nat Genet*. 2020; **52**: 986-1001.
7. Hadfield, J, Megill, C, Bell, SM, *et al.* Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. 2018; **34**(23): 4121-3.
8. Sayers, EW, Beck, J, Bolton, EE, *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*. 2021; **49**: D10-D7.
9. Shu, YL, McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality. *Eurosurveillance*. 2017; **22**(13): 2-4.
10. Elbe, S, Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob Chall*. 2017; **1**(1): 33-46.
11. Zhao, W-M, Song, S-H, Chen, M-L, *et al.* The 2019 novel coronavirus resource. *Hereditas*

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

*(Beijing)*. 2020; **42**(2): 212-21.

12. Gong, Z, Zhu, J-W, Li, C-P, *et al.* An online coronavirus analysis platform from the National Genomics Data Center. *Zool Res*. 2020; **41**(6): 705-8.
13. Chen, F, You, L, Yang, F, *et al.* CNGBdb: China National GeneBank DataBase. *Hereditas (Beijing)*. 2020; **42**(8): 799-809.
14. Zhang, Z, Zhao, W, Xiao, J, *et al.* Database resources of the National Genomics Data Center in 2020. *Nucleic Acids Res*. 2020; **48**(D1): D24-D33.
15. Katoh, K, Standley, DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol*. 2013; **30**(4): 772-80.
16. Zhou, P, Yang, X-L, Wang, X-G, *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020; **579**: 270-3.
17. Lam, TT-Y, Jia, N, Zhang, Y-W, *et al.* Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature*. 2020; **583**(7815): 282-5.
18. Hartigan, JA. Minimum mutation fits to a given tree. *Biometrics*. 1973; **29**(1): 53-65.
19. Sagulenko, P, Puller, V, Neher, RA. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol*. 2018; **4**: vex042.
20. Zhao, ZM, Li, HP, Wu, XZ, *et al.* Moderate mutation rate in the SARS coronavirus genome and its implications. *BMC Evol Biol*. 2004; **4**: 21.
21. Cotten, M, Watson, SJ, Zumla, AI, *et al.* Spread, circulation, and evolution of the Middle East respiratory syndrome coronavirus. *Mbio*. 2014; **5**(1): e01062-13.
22. Nie, Q, Li, X, Chen, W, *et al.* Phylogenetic and phylodynamic analyses of SARS-CoV-2. *Virus Res*. 2020; **287**: 198098.
23. Korber, B, Fischer, WM, Gnanakaran, S, *et al.* Tracking changes in SARS-CoV-2 Spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell*. 2020; **182**(4): 812-27.
24. Rambaut, A, Loman, N, Pybus, O, *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. *virologicalorg*. 2020: https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563.
25. Xing, Y, Wong, GWK, Ni, W, *et al.* Rapid response to an outbreak in Qingdao, China. *N Engl J Med*. 2020; **383**: e129.
26. Zhang, Y, Pan, Y, Zhao, X, *et al.* Genomic characterization of SARS-CoV-2 identified in a reemerging COVID-19 outbreak in Beijing's Xinfadi market in 2020. *Biosaf Health*. 2020; **2**: 202-5.
27. Pang, X, Ren, L, Wu, S, *et al.* Cold-chain food contamination as the possible origin of Covid-19 resurgence in Beijing. *Natl Sci Rev*. 2020; **7**: 1861-4.
28. Yu, D, Dong, L, Yan, F, *et al.* eGPS 1.0: comprehensive software for multi-omic and evolutionary analyses. *Natl Sci Rev*. 2019; **6**: 867-9.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

**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.

**Funding**

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.

**Data and materials availability:** The coronavirus genomic sequences used in this study were obtained from the 2019nCoVR database (11). Timely updated data of genomic sequences of SARS-CoV-2 variants are shared with the general public at https://bigd.big.ac.cn/ncov/apis/. The free software (desktop and web-based versions) can be downloaded from http://www.egps-software.net/.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

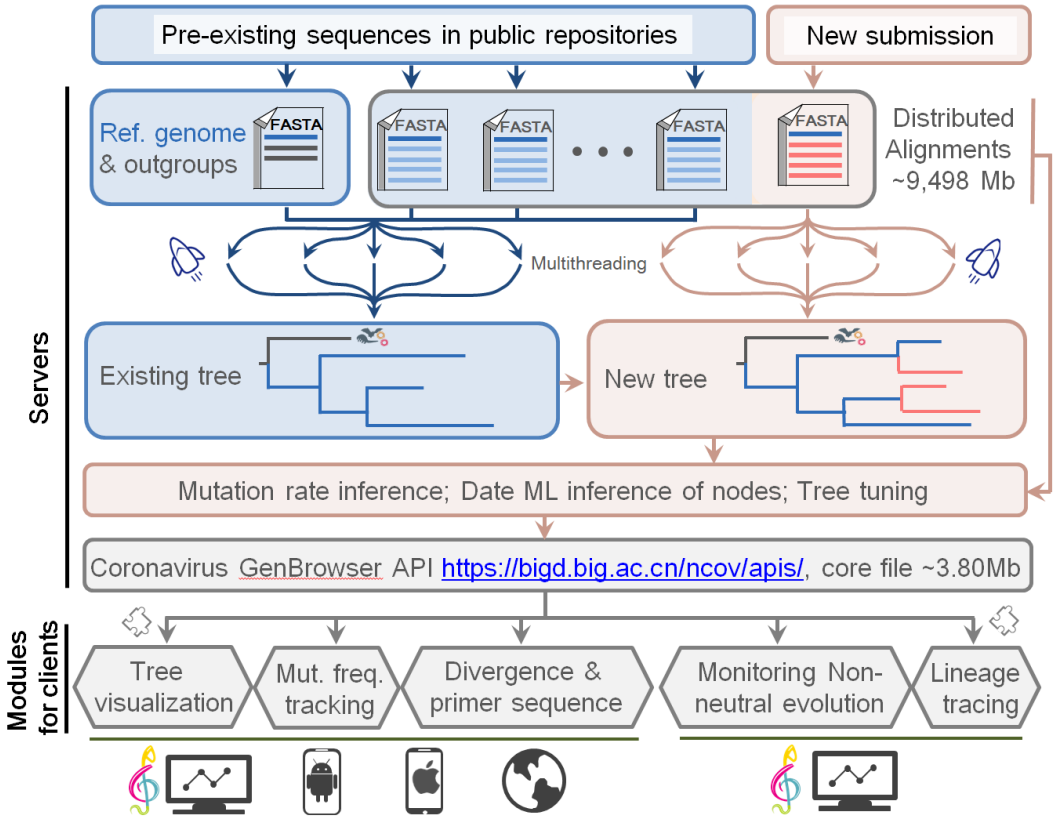


**Figure 1. Timely updates of SARS-CoV-2 genomic data and visualization framework of Coronavirus GenBrowser.** The core file includes the pre-analyzed genomic data of SARS-CoV-2 variants and the associated metadata. All timely-updated data can be freely accessed at https://bigd.big.ac.cn/ncov/apis/.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

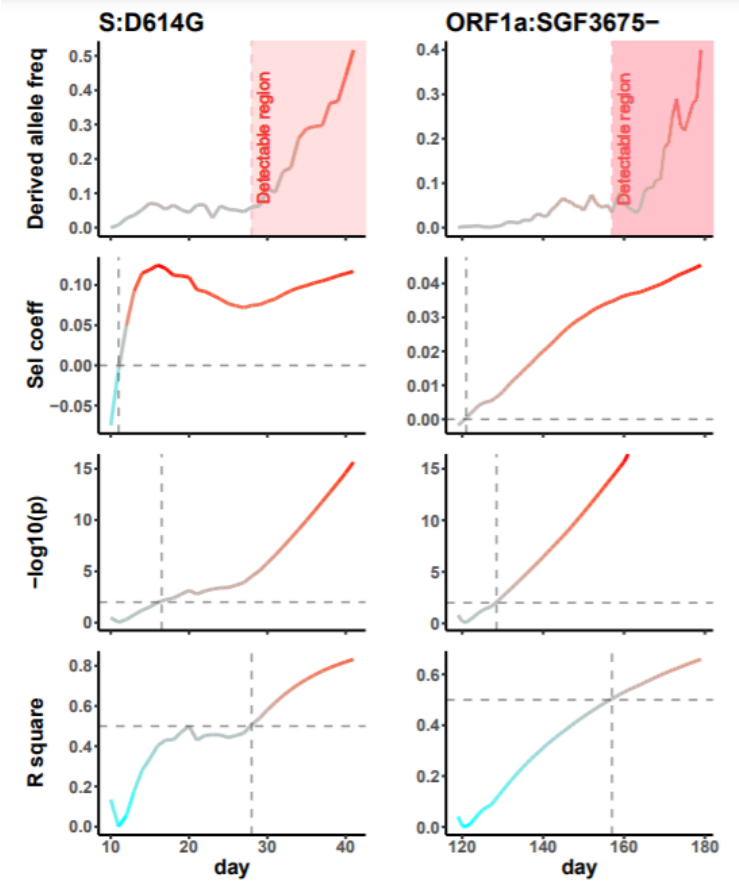


**Figure 2. Putative advantageous variants of SARS-CoV-2.** The *x*-axis displays number of days since the first appearance of derived allele in the global viral population. Predicted adaptation is marked in pink. Dashed gray crossings denote top right corners with a positive selection coefficient, $p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

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

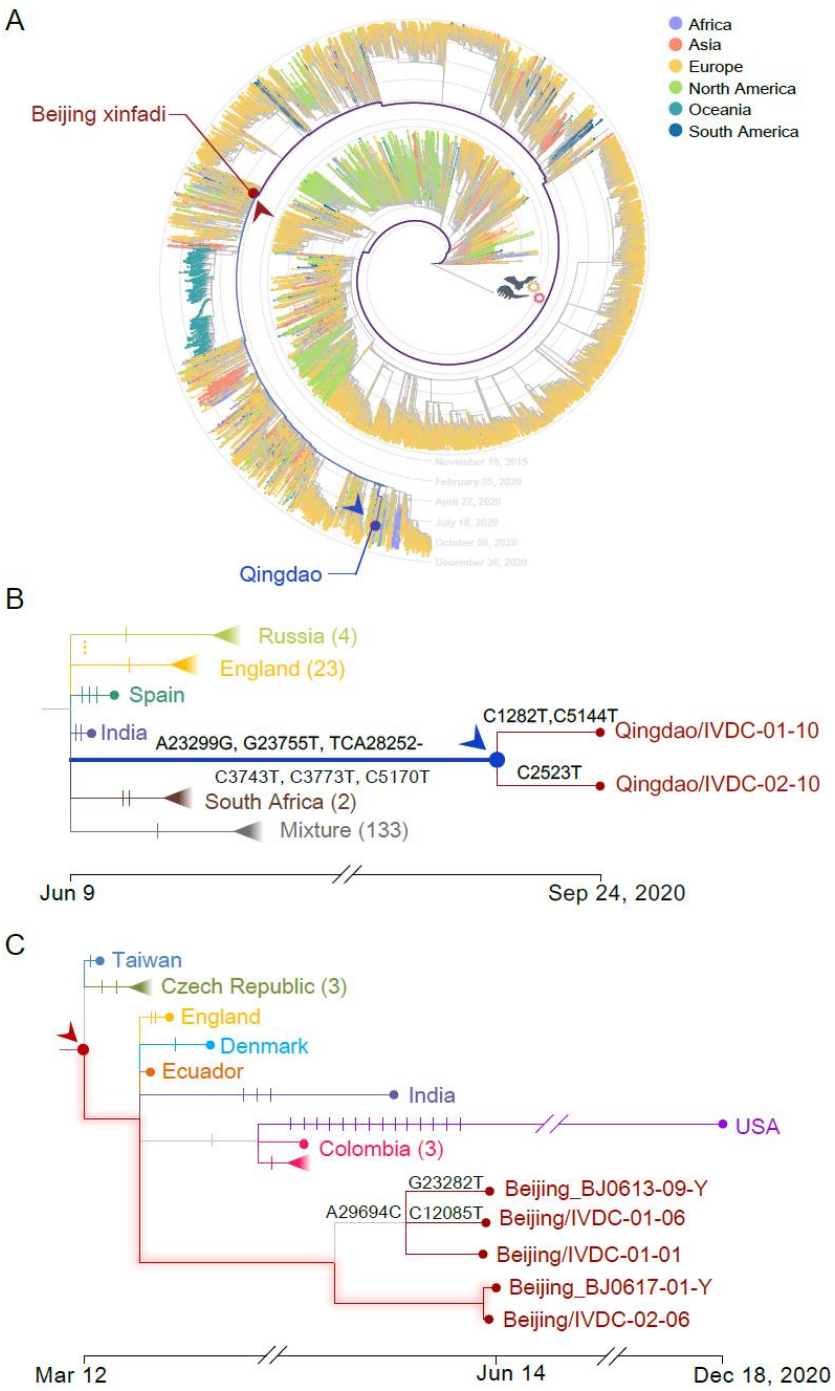


**Figure 3. Global and zoom-in views of lineages associated with Qingdao and Beijing outbreaks.**

A) The lineages of traced targets are shown in blue and dark red lines.

B) Qingdao/IVDC-01-10 and Qingdao/IVDC-02-10 were the two SARS-CoV-2 strains collected on 24 September, 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 7 October, 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

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

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 11 June and 14 June, 2020. The branches with no mutations are highlighted.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

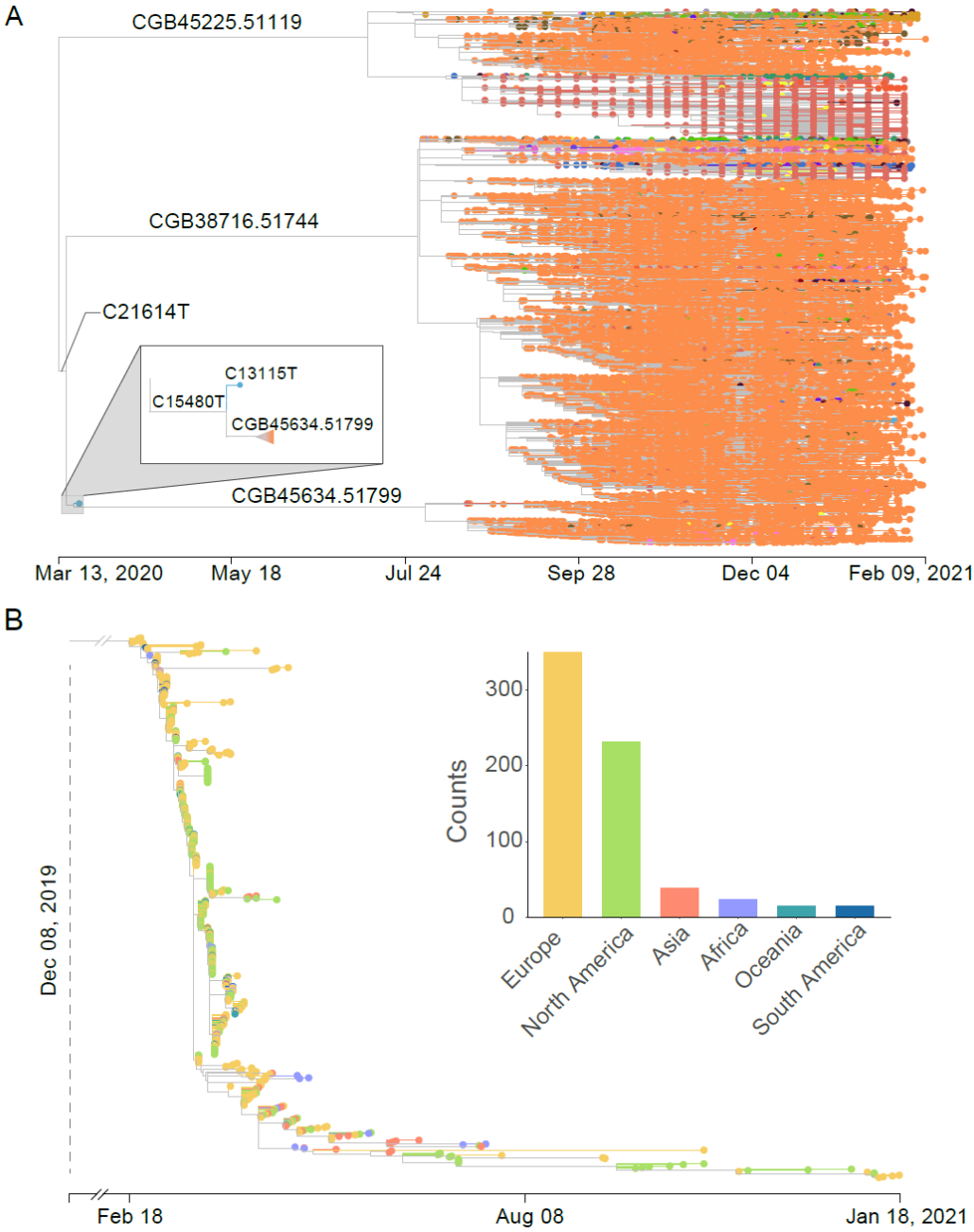


**Figure 4. Transmission of mutation-dormant SARS-CoV-2 variants.**

A) Tree visualization of strains of three related lineages with reduced evolutionary rate among 330,942 SARS-CoV-2 genomic sequences. There were no mutations in the strains of three long branches (CGB45225.51119, CGB38716.51744, and CGB45634.51799) in 118, 112, and 135 days, respectively. The numbers of their descendants were 9,815, 28,126, and 3,015, respectively. The three strains had spread to more than 10 European countries, and most descendants were found in Denmark (red tip) and UK (orange tip). Variants of these three lineages shared 13 mutations (G204T, C241T, T445C, C3037T, C6286T, C14408T, G21255C,

medRxiv preprint doi: https://doi.org/10.1101/2020.12.23.20248612; this version posted April 30, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-ND 4.0 International license .

C22227T, A23403G, C26801G, C27944T, C28932T, and G29645T) but had two unique mutations (C21614T and C15480T).

B) Tree visualization of 671 strains with no mutations between 8 December, 2019 and 18 January, 2021. These strains are the descendants of a variant with four mutations (C3037T, A23403G, C14408T, and C241T) emerged before 8 December, 2019. Among the four mutations, A23403G (S:D614G) and C14408T (ORF1b:P314L) are non-synonymous. The geographic distribution of the descendants is shown in the subpanel. Their siblings with additional mutations (16,442 strains) are not shown in the evolutionary tree. The CGB ID of their MRCA is CGB121.124.